AKCIT
CENTRO DE COMPETÊNCIA EMBRAPII EM TECNOLOGIAS IMERSIVAS

Gestão de Dados de Saúde

Autoria:
Felipe Ferré

Organizadores:
Renata Dutra Braga
Juliana Pereira de Souza-Zinader
Taciana Novo Kudo
Sheila Mara Pedrosa
Arlindo Rodrigues Galvão Filho

Cegraf UFG

Gestão de Dados de Saúde

Autoria:
Felipe Ferré

Organizadores:
Renata Dutra Braga
Juliana Pereira de Souza-Zinader
Taciana Novo Kudo
Sheila Mara Pedrosa
Arlindo Rodrigues Galvão Filho

Cegraf UFG

2024

Revisão Técnica
Juliana Pereira de Souza-Zinader

Revisão Editorial
Ana Laura de Sene Amâncio Zara Brisolla

Capa
Iuri Vaz Miranda

Editoração Eletrônica
Luma Wanderley de Oliveira
Layane Grazielle Souza Dias

https://doi.org/10.5216/FER.ges.ebook.978-85-495-1042-6/2024

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Ferré, Felipe
Gestão de dados de saúde [livro eletrônico] / Felipe Ferré ; organização Renata Dutra Braga...[et al.]. -- 1. ed. -- Goiânia, GO : Cegraf UFG, 2024.
PDF

Outros organizadores: Juliana Pereira de Souza-Zinader, Taciana Novo Kudo, Sheila Mara Pedrosa, Arlindo Rodrigues Galvão Filho.
ISBN 978-85-495-1042-6

1. Ciência da computação 2. Dados - Análise 3. Dados - Estruturas (Ciência da computação) 4. Gestão de saúde 5. Sistema de gestão de dados I. Braga, Renata Dutra. II. Souza-Zinader, Juliana Pereira de. III. Kudo, Taciana Novo. IV. Pedrosa, Sheila Mara. V. Galvão Filho, Arlindo Rodrigues.

24-246182 CDD-005.73

Índices para catálogo sistemático:

1. Dados : Estruturas : Processamento de dados 005.73

Aline Graziele Benitez - Bibliotecária - CRB-1/3129

Gestão de Dados de Saúde

Instituições responsáveis

Universidade Federal de Goiás (UFG)

Centro de Competência Embrapii em Tecnologias Imersivas, denominado AKCIT (Advanced Knowledge Center for Immersive Technologies)

Centro de Excelência em Inteligência Artificial (CEIA)

Instituições financiadoras

Empresa Brasileira de Pesquisa e Inovação Industrial (Embrapii)

Governo do Estado de Goiás

Empresas parceiras do AKCIT

Apoio

Universidade Federal de Goiás (UFG)

Pró-Reitoria de Pesquisa e Inovação (PRPI-UFG)

Instituto de Informática (INF-UFG)

Lista de Abreviaturas e Siglas

1NF	Primeira Forma Normal
2D	2 Dimensões
2NF	Segunda Forma Normal
3D	3 Dimensões
AIH	Autorização de Internação Hospitalar
ANPD	Autoridade Nacional de Proteção de Dados
ANS	Agência Nacional de Saúde Suplementar
API	*Application Programming Interfaces* **-** Interface de Programação de Aplicações
APS	Atenção Primária à Saúde
AR	*Augmented Reality* - Realidade Aumentada
ARIMA	*Autoregressive Integrated Moving Average* - Modelo Autorregressivo Integrado de Médias Móveis
ASYM	*Asymmetry* - Assimetria
ATC	*Anatomical Therapeutic Chemical Code* - Classificação Anatômica, Terapêutica e Química
AUC	*Area Under the Curve* - Área Sob a Curva
B	*Benign* - *Benigno*
BI	*Business Intelligence* - Inteligência de Negócios
BNAFAR	Base Nacional de Dados de Ações e Serviços da Assistência Farmacêutica no Sistema Único de Saúde
BoW	*Bag of Words* - Saco de Palavras

BPS	Banco de Preços em Saúde
BREATHE	Registro Brasileiro de Insuficiência Cardíaca
CADSUS	Sistema de Gestão do Conselho Nacional de Saúde
CALC	Calcificação
CATMAT	Catálogo de Materiais e Serviços
CBM	*Concept-Based Method* - Método Baseado em Conceitos
Ccates	Centro Colaborador do Sistema Único de Saúde para Avaliação de Tecnologias e Excelência em Saúde
CEAF	Componente Especializado da Assistência Farmacêutica
CEIS	Complexo Econômico Industrial da Saúde
CID-10	Classificação Internacional de Doenças
CIDACS	Centro de Integração de Dados e Conhecimento para Saúde
CIIA-Saúde	Centro de Inovação em Inteligência Artificial para a Saúde
CIO	*Chief Information Officer* - Diretor de Informações
CIRC	Massas bem definidas/circunscritas
CMED	Câmara de Regulação do Mercado de Medicamentos
CMIO	*Chief Medical Information Officer* - Diretor de Informações Médicas
CMM	Modelo de Maturidade das Capacidades
CMMI	*Capability Maturity Model Integration* - Integração do Modelo de Maturidade de Capacidades
CNES	Cadastro Nacional de Estabelecimentos de Saúde
CNNs	Redes Neurais Convolucionais
CNS	Cartão Nacional de Saúde
CONASEMS	Conselho Nacional de Secretarias Municipais de Saúde

CONASS	Conselho Nacional de Secretários de Saúde
CONITEC	Comissão Nacional de Incorporação de Tecnologias no Sistema Único de Saúde
CPF	Certidão de Pessoa Física
CRUD	*Create, Read, Update, Delete* - Criar, Ler, Atualizar, Deletar
CT	*Clinical Terms* - Termos Clínicos
D	*Dense-glandular* - Glandular Denso
DBA	*Database Administrator* - Administrador de Banco de Dados
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise* - Agrupamento Espacial Baseado em Densidade de Aplicações com Ruído
DCNT	Doenças Crônicas Não Transmissíveis
DeCS	Descritores em Ciências da Saúde
DICOM	*Digital Imaging and Communications in Medicine* - Imagens Digitais e Comunicações em Medicina
DLP	*Data Loss Prevention* - Prevenção de Perda de Dados
DMP	*Data Management Plan* - Plano de Gerenciamento de Dados
DNA	*Deoxyribonucleic Acid* - Ácido Desoxirribonucleico
DSM-5	*Diagnostic and Statistical Manual of Mental Disorders 5* - Manual Diagnóstico e Estatístico de Transtornos Mentais 5
DW	*Data Warehouse* - Armazém de Dados
EHR	*Electronic Health Records* - Registros Eletrônicos de Saúde
ELSA	Estudo Longitudinal da Saúde do Adulto
EMRAM	*Electronic Medical Record Adoption Model* - Modelo de Adoção de Registros Médicos Eletrônicos
ESB	*Enterprise Service Bus* - Barramento de Serviço Empresarial
ETL	*Extract, Transform, Load* - Extrair, Transformar, Carregar

F	*Fatty* - Gorduroso
FHIR	*Fast Healthcare Interoperability Resources* - Recursos de Interoperabilidade Rápida para Cuidados de Saúde
Fiocruz	Fundação Oswaldo Cruz
FN	Falso Negativo
FP	Falso Positivo
FTP	*File Transfer Protocol* - Protocolo de Transferência de Arquivos
G	*Fatty-glandular* - *Gorduroso-Glandular*
GANs	*Generative Adversarial Networks* - Redes Adversárias Generativas
GDPR	*General Data Protection Regulation* - Regulamento Geral de Proteção de Dados
GO	*Gene Ontology* - Ontologia de Genes
HIMSS	*Healthcare Information and Management Systems Society* - Sociedade de Sistemas de Informação e Gestão em Saúde
HIPAA	*Health Insurance Portability and Accountability Act* - Lei de Portabilidade e Responsabilidade de Seguro Saúde
HL7	*Health Level 7* - Nível de Saúde 7
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IDM+P	Identificação, Definição e Modelagem de Produtos
IDP	*Intelligent Document Processing* - Processamento Inteligente de Documentos
IDS	*Intrusion Detection System* - Sistema de Detecção de Intrusão
IMC	Índice de Massa Corporal
INMSD	Índice Nacional de Maturidade em Saúde Digital
IoT	*Internet of Things* - Internet das Coisas

IPDLN	*International Data Likage Network* - Rede Internacional de Vinculação de Dados
IPS	*Intrusion Prevention System* - Sistema de Prevenção de Intrusão
IQR	*Interquartile Range* - Intervalo Interquartil
JSON	*JavaScript Object Notation* - Notação de Objetos JavaScript
KDD	*Knowledge Discovery in Databases* - Descoberta de Conhecimento em Bancos de Dados
KEGG	*Kyoto Encyclopedia of Genes and Genomes* - Enciclopédia de Genes e Genomas de Quioto
LAI	Lei de Acesso à Informação
LDA	*Linear Discriminant Analysis* - Análise Discriminante Linear
LGPD	Lei Geral de Proteção de Dados Pessoais
LOINC	*Logical Observation Identifiers Names and Codes* - Identificadores Lógicos de Observação, Nomes e Códigos
LSTM	*Long Short-Term Memory* - Memória de Longo e Curto Prazo
M	*Malignant* - Maligna
MAD	Metodologia de Administração de Dados
MAE	*Mean Absolute Error* - Erro Absoluto Médio
MCL	*Markov Clustering Algorithm* - Algoritmo de Agrupamento de Markov
MDSA	Ministério do Desenvolvimento Social e Combate à Fome
MIAS	*Mini Mammographic Database* - Mini Banco de Dados Mamográfico
MISC	*Miscellaneous* - Outras massas mal definidas
ML	*Machine Learning* - Aprendizado de Máquina
MRI	*Magnetic Resonance Imaging* - Ressonância Magnética
MS	Ministério da Saúde

MSE	*Mean Squared Error* - Erro Quadrático Médio
NATS	Núcleos de Avaliação de Tecnologias em Saúde
NDEx	*Network Data Exchange* - Troca de Dados em Rede
NER	*Named Entity Recognition* - Reconhecimento de Entidades Nomeadas
NORM	Normal
noSQL	*Non-Relational* - Não Relacional
NTP	*Network Time Protocol* - Protocolo de Tempo em Rede
OBM	Ontologia Brasileira de Medicamentos
OCR	*Optical Character Recognition* - Reconhecimento Óptico de Caracteres
OLAP	*Online Analytical Processing* - Processamento Analítico *Online*
OLTP	*Online Transaction Processing* - Processamento de Transações *Online*
OPM	*Orthopedic Prosthetic Material* - Meios Auxiliares de Locomoção
OSF	*Open Science Framework* - Estrutura de Ciência Aberta
PACS	*Picture Archiving and Communication System* - Sistema de Arquivamento e Comunicação de Imagens
PBM	*Phrase-Based Method* - Método Baseado em Frases
PCA	*Principal Component Analysis* - Análise de Componentes Principais
PDI	*Pentaho Data Integration* - Integração de Dados Pentaho
PEP	Prontuário Eletrônico do Paciente
PLN	Processamento de Linguagem Natural
PMAQ-SUS	Programa de Melhoria do Acesso e da Qualidade do Sistema Único de Saúde
PNAD Contínua	Pesquisa Nacional por Amostra de Domicílios Contínua

PNAUM	Pesquisa Nacional sobre o Acesso, Utilização e Promoção do Uso Racional de Medicamentos no Brasil
PO	*Product Owner* - Proprietário do Produto
PTM	*Pattern Taxonomy Method* - Método de Taxonomia de Padrões
RAS	Redes de Atenção à Saúde
RENAME	Relação Nacional de Medicamentos Essenciais
RENASES	Relação Nacional de Ações e Serviços de Saúde
RENIP	Registro Nacional de Implantes de Próteses Mamárias
RES	Registros Eletrônicos de Saúde
REST	*Representational State Transfer* - Transferência de Estado Representacional
Ripsa	Rede Interagencial de Informações para a Saúde
RNDS	Rede Nacional de Dados em Saúde
RNNs	Redes Neurais Recorrentes
ROC	*Receiver Operating Characteristic* - Curva Característica de Operação do Receptor
RTMG	Rede de Telessaúde de Minas Gerais
SAGE	Sala de Apoio à Gestão Estratégica
SBIS	Sociedade Brasileira de Informática em Saúde
SEI	*Software Engineering Institute* - Instituto de Engenharia de *Software*
SEIDIGI	Secretaria de Informação e Saúde Digital
SES	Secretarias Estaduais de Saúde
SGBDs	Sistemas Gerenciadores de Banco de Dados
SI-PNI	Programa Nacional de Imunizações
SIA	Sistema de Informações Ambulatoriais do SUS

SIASI	Sistema de Informação de Atenção à Saúde Indígena
SIB	Sistema de Informações de Beneficiários
SIDRA	Sistema IBGE de Recuperação Automática
SIEM	*Security Information Event Management* - Gerenciamento de Eventos e Informações de Segurança
SIGTAP	Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e Órteses, Próteses e Materiais do Sistema Único de Saúde
SIH	Sistema de Informações Hospitalares do Sistema Único de Saúde
SIM	Sistema de Informações de Mortalidade
Sinan	Sistema de Informação de Agravos de Notificação
Sinasc	Sistema de Informação sobre Nascidos Vivos
SIS	Sistemas de Informação em Saúde
SISAB	Sistema de Informação em Saúde para a Atenção Básica
SISAGUA	Sistema de Informação de Vigilância da Qualidade da Água para Consumo Humano
Siscan	Sistema de Informações sobre Câncer
SISPNCD	Sistema de Informação do Programa Nacional de Controle da Dengue
Sisvan	Sistema de Vigilância Alimentar e Nutricional
Sisvep	Sistema de Informação de Vigilância de Populações Expostas a Contaminantes Químicos
SNOMED	*Systematized Nomenclature of Medicine* - Nomenclatura Sistemática da Medicina
SNOMED-CT	*Systematized Nomenclature of Medicine - Clinical Terms* - Nomenclatura Sistemática da Medicina - Termos Clínicos
SNVS	Sistema Nacional de Vigilância em Saúde

SOAP	*Simple Object Access Protocol* - Protocolo de Acesso a Objetos Simples
SPIC	*Spiculated Masses* - Massas Espiculadas
SQL	*Structured Query Language* - Linguagem de Consulta Estruturada
SRAG	Síndrome Respiratória Aguda Grave
SSIS	*Server Integration Services* - Serviços de Integração de Servidor
STT-MG	Sistema de Tele-eletrocardiografia de Minas Gerais
SUS	Sistema Único de Saúde
SVM	*Support Vectors Machine* - Máquinas de Vetores de Suporte
SVS	Secretaria de Vigilância à Saúde
TC	Tomografias Computadorizadas
TI	Tecnologia da Informação
TIC	Tecnologias de Informação e Comunicação
TICS	Tecnologia da Informação e Comunicação em Saúde
TIS	Tecnologia da Informação em Saúde
TISS	Troca de Informação de Saúde Suplementar
TUSS	Terminologia Unificada da Saúde Suplementar
UBS	Unidade Básica de Saúde
UF	Unidade da Federação
UFBA	Universidade Federal da Bahia
UFG	Universidade Federal de Goiás
UFMG	Universidade Federal de Minas Gerais
UI	Interface de Usuário

UMLS	*Unified Medical Language System* - Sistema Unificado de Linguagem Médica
UNB	Universidade de Brasília
UX	*User Experience* - Experiência do Usuário
Visat	Vigilância em Saúde do Trabalhador
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos
VPN	*Virtual Private Network* - Rede Privada Virtual
VR	Realidade Virtual
WYSIWYG	*What You See Is What You Get* - O Que Você Vê é o Que Você Obtém
XML	*Extensible Markup Language* - Linguagem de Marcação Estendida
XP	*Extreme Programming* - Programação Extrema

Listas de Figuras e Tabelas

Sumário

Apresentação

Prezado(a) Participante,

Seja bem-vindo(a) ao Microcurso **Gestão de Dados de Saúde**!

O Microcurso faz parte da Coleção Formação e Capacitação do Centro de Competências Imersivas, uma parceria entre a Embrapii e a Universidade Federal de Goiás (UFG).

A prática do profissional da saúde está cada vez mais associada ao uso de sistemas de informação. Estima-se que o conhecimento da saúde dobra a cada 2,5 meses (Densen, 2011). Da mesma forma, o volume de dados coletados e armazenados cresce exponencialmente, mas a capacidade humana em analisar e tomar decisão informada não amplia na mesma medida.

A tecnologia não substitui o profissional da saúde, mas profissionais que a instrumentalizam podem substituir aqueles que não lidam bem com novos métodos mediados por informação. Surge, assim, não apenas o informata em saúde, aquele que é capaz de lidar com ambas as áreas do conhecimento, mas, na mesma toada, o gestor de dados em saúde, capaz de organizar grandes volumes de dados e gerar valor informacional, reorientando o sistema de saúde para profissionais e usuários atuarem de forma mais inteligente e preditiva.

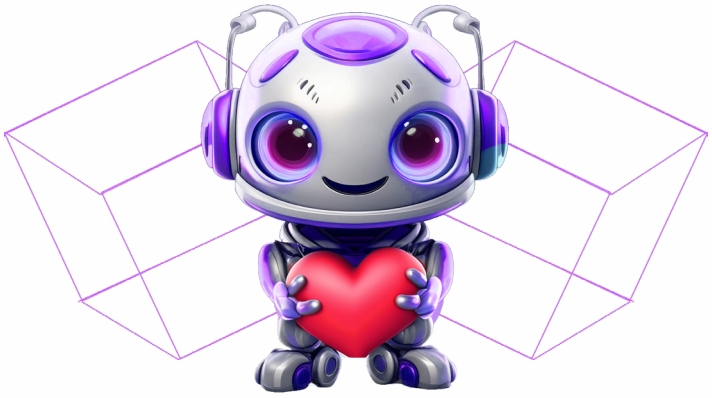
A Tecnologia da Informação (TI) é um instrumento e fonte de novos conhecimentos científicos da área da saúde e sua falha implica em prejuízos à saúde individual e coletiva, sendo tanto atividade meio para as demais, quanto atividade fim. A informática em saúde ou saúde digital é um marco estratégico sensível ao País, frente ao contexto global do desenvolvimento do Complexo Econômico-Industrial da Saúde, cuja produção científica ocorre em modelos colaborativos públicos e competitivos de mercado.

Ao longo de décadas a Informática em Saúde se tornou uma área destacada do conhecimento científico diante da sofisticação e da necessidade de integração entre equipamentos, organizações e processos de trabalho informatizados. Por essa razão, a separação entre profissional da saúde e profissional de TI se torna cada vez mais tênue no perfil do gestor de dados em saúde, o qual ocupa postos estratégicos no organograma institucional.

O programa do Microcurso abrange marcos conceituais básicos de como manter dados e apoiar a decisão, cuja gestão viabiliza informar e comunicar diferentes atores e comunidades. Os conhecimentos são aplicados tanto no nível de Sistemas de Saúde Nacionais quanto em estabelecimentos operados pela lógica de mercado. Com o Microcurso você terá elementos para trabalhar em equipes e, até mesmo, liderar. O gestor não necessariamente é aquele que constrói linhas de código-fonte e algoritmos. Porém, o conhecimento de noções básicas de bancos de dados e tecnologias inteligentes, bem como do vocabulário, é necessário para a integração com a equipe de TI. Ao conhecer os componentes básicos da gestão de dados em saúde você terá elementos para escolher sua área de atuação, evoluir na carreira e se especializar cada vez mais.

Ao final deste Microcurso, você será capaz de:

- Aplicar, em seu quotidiano, o vocabulário básico do meio de gestão de dados e se comunicar com os diversos perfis profissionais de TI e da saúde;
- Modelar e extrair dados das diversas fontes segundo a natureza institucional e os tipos de sistemas de informação;
- Atuar em equipes de desenvolvimento de *software* e apoio à tomada de decisão ao conhecer as técnicas gerais de produção e análise de dados;
- Garantir qualidade, disponibilidade e reprodutibilidade dos processos mediados por sistemas de informação;
- Conhecer técnicas inteligentes para processamento de dados em saúde a partir de Sistemas Gerenciadores de Bancos de Dados estruturados e em linguagem natural, como textos, áudios, imagens e vídeos.



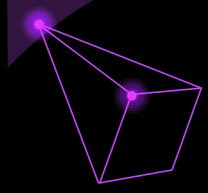
Desejamos um excelente estudo!!!

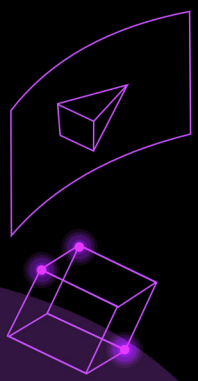
AKCIT
CENTRO DE COMPETÊNCIA EMBRAPII EM TECNOLOGIAS IMERSIVAS

Gestão de Dados de Saúde

Unidade I

Introdução à Gestão de Dados de Saúde





Unidade I: Introdução à Gestão de Dados de Saúde

1.1 Introdução - Gestão e Governança de Dados em Saúde

Nas últimas décadas, um novo profissional vem se estabelecendo, tanto na gestão estatal quanto em estabelecimentos de saúde de maior porte, o “Diretor de TI” ou “Diretor de Informação”. Algumas organizações também utilizam o termo “Diretor de TI”. Essas denominações referem-se ao executivo responsável pela gestão das tecnologias da informação numa instituição. O termo correlato em inglês é *Chief Information Officer* (CIO), sendo um executivo de alto escalão responsável pela implementação, gestão e utilização das tecnologias da informação e informática. As atribuições do CIO incluem:

- Atuação na alta direção da instituição: assegurar que os processos derivados para aplicar decisões estratégicas atinjam os objetivos pretendidos mediados pela informação;
- Manutenção da infraestrutura e infoestrutura: assegurar que estações e servidores operem prevenindo a indisponibilidade de serviços, garantia contra perda ou vazamento de informações e integração entre as diferentes plataformas adotadas na instituição e com parceiros;
- Monitoramento e avaliação: apoiar a renovação do parque tecnológico, incorporar novos métodos e realizar pesquisa, desenvolvimento e inovação;
- Gestão de recursos de TI: gerenciar contratos, prestadores de serviço e força de trabalho;
- Apoio ao letramento digital: estabelecer sistemas de treinamento, educação continuada e educação permanente, mantendo plataformas de gestão de conhecimento com enciclopédias eletrônicas institucionais, bibliotecas de código-fonte, robôs conversacionais (*chatbot*) e sistemas de apoio ao usuário das plataformas de *software*.

Na condução de uma instituição, os níveis operacional, tático e estratégico desempenham papéis interligados e mediados pela governança da informação operada pelo gestor de dados. No nível operacional, as atividades diárias e rotineiras são gerenciadas, assegurando que os processos sejam executados conforme processos e monitorados por indicadores de qualidade pré-definidos. No nível tático, a ênfase

está em traduzir as diretrizes estratégicas em planos específicos e ações coordenadas, visando à implementação de iniciativas que melhoram o desempenho organizacional no médio prazo. No nível estratégico, as decisões são tomadas com uma perspectiva de longo prazo, envolvendo a definição de metas, a formulação de políticas e a alocação de recursos para garantir a sustentabilidade e o crescimento. Esse nível requer uma visão ampla e integrada do ambiente externo e interno, antecipando mudanças e oportunidades que podem influenciar o futuro da organização e está ligada ao maior nível organizacional.

A figura do CIO vem se tornando imprescindível na estrutura de governança institucional. Diversos países estabeleceram políticas de saúde digital, originalmente chamada de informática médica, informática em saúde ou também e-Saúde. No Brasil foram publicados instrumentos de governança norteadores para a Saúde Digital - Política Nacional de Informação e Informática em Saúde (Brasil, 2021) e Estratégia Saúde Digital para o Brasil 2020-2028 (Brasil, 2020). Em 2023, o Ministério da Saúde cria a figura do Secretário de Saúde Digital ao colocar no mais elevado patamar do organograma a área com a Secretaria de Informação e Saúde Digital (SEIDIGI). Outras Secretarias Estaduais de Saúde (SES), como a de Goiás e da Bahia, também vêm alçando o setor ao grau estratégico, antes apenas situado no nível operacional.

No setor privado, a figura do CIO também é importante para o sucesso da área, podemos destacar inclusive o papel do CIO em estabelecimentos de saúde no qual possui departamentos de TI, como por exemplo, hospitais, operadoras de saúde, empresas de medicina diagnóstica, *homecare*, secretarias de saúde e atenção primária. Esses dois últimos fazendo referência ao setor público.

Com os avanços na Saúde Digital no mundo, em especial após o advento da covid-19, destaca-se também a evolução da figura do CIO para o CMIO (do inglês *Chief Medical Information Officer*). Esse cargo pode ser traduzido como líder ou diretor de informática médica. Embora este cargo já exista há muitos anos nos Estados Unidos, por exemplo, no Brasil, a função já existe em alguns hospitais privados (incluindo o departamento de Informática Médica). E, agora, após a pandemia e com a transformação digital em prol da inovação, tecnologias emergentes e imersivas, nota-se uma migração para empresas de TI, que prestam serviços a esses *players*.

Dessa forma, com as atribuições e apoio dos cargos mencionados anteriormente, cabe à gestão de dados em saúde orientar a organização na adoção de planos explícitos de carreira, de gestão de dados e de maturidade institucional, as organizações podem melhorar a qualidade de seus produtos, aumentar a satisfação do usuário, a eficiência operacional ao gerenciar projetos de forma mais eficaz, resultando na melhor alocação de recursos.

1.2 Conceitos Básicos de Dados de Saúde

A presente seção fornece um vocabulário que abrange a diversidade de perfis dentro de uma mesma equipe. É fundamental que a equipe detenha uma linguagem com conceitos comuns. Com a crescente especialização e trabalho multiprofissional, é habitual o convívio de profissionais clínicos, epidemiologistas, tecnologistas de saúde, economistas da saúde e sanitaristas com bioestatísticos, cientistas da computação, cientistas da informação, biblioteconomistas, administradores de bancos de dados, analistas de dados, cientistas de dados e até pessoas com perfis criativos ou ligados às ciências humanas, como profissionais do direito sanitário, antropólogos da saúde, linguistas, cientistas políticos, cientistas sociais, *designers* e profissionais especializados em comportamento, comunicação e influenciadores de redes sociais.

Um conjunto de dados armazenados, organizados e mantidos de modo a recuperar eficazmente as memórias no momento em que são necessárias é um **repositório** ou **banco de dados**. Os repositórios são organizados a partir de uma linguagem. O **modelo** é uma abstração, uma simplificação do mundo real, a qual define a linguagem assertiva do que deve e o que não deve ser armazenado e recuperado posteriormente para assegurar os objetivos pretendidos. O contexto é comumente chamado de **negócio**, ao situarmos dado ambiente institucional. Na TI, diz-se **regra de negócio** no ato da transposição do conhecimento que está sendo informatizado.

O método para lidar com dados e metadados apresenta especificidades na área da saúde, as quais serão abordadas no presente Microcurso, onde se diferencia a gestão de dados e metadados. O **metadado** contém as informações que viabilizam interpretar corretamente o dado e assegurar o significado negocial pretendido para dado significante. A qualidade da comunicação obtida com a recuperação de dados a partir do repositório é proporcional à gestão do dado e do metadado que deve ser explícita e comum entre os interlocutores. Assim como uma linguagem apresenta o vocabulário, o conjunto de metadados delimita, minimamente, o tipo de dados, o atributo com o respectivo nome e rótulo, a descrição, o domínio e as restrições. Os metadados do conjunto de dados constituem o **dicionário de dados**.

Tipo é uma classe à qual o dado pertence, usualmente sendo número, texto, data ou imagem, com variações. Por exemplo, a idade em anos pode ser definida como um número inteiro. A data de nascimento pode ser definida como a junção do dia, mês e ano, no formato AAAA-MM-DD, por exemplo, 2005-10-28 para 28 de outubro de 2005. Dentro de cada tipo existem variações que serão exploradas adiante.

Atributo é a característica do objeto definida no repositório por dada comunidade no contexto da abstração do banco de dados. O tipo é usualmente definido pela

comunidade que mantém a linguagem do banco de dados, o atributo é definido pela comunidade que mantém o repositório específico. O atributo tem um nome, geralmente no formato pré-definido pela linguagem do banco de dados e um rótulo, geralmente o significado curto do nome. Exemplos de atributos com rótulos, respectivamente são, “sg_sexo_usuario_sus” para “Sigla do sexo do usuário do SUS”, “ds_logradouro” para “Logradouro com número e complemento”, co_diagnostico para “Código do diagnóstico”, “dt_nascimento” para “data de nascimento”, “nm_mae” para “nome da mãe”, “no_procedimento” para “nome do procedimento realizado”, “dt_internacao” para “data de internação”. Em estatística, o atributo comumente é chamado de **variável**.

A **descrição** do atributo estabelece, enquanto verbete, um comentário, anotação, apontamento e informações pertinentes à aplicação no contexto. Por exemplo, quanto ao município de residência presente no atributo “codmunres” do atestado de óbito, a descrição apresenta uma instrução: “Código do município de residência”. Em caso de óbito fetal, considerar o município de residência da mãe.

Domínio é o conjunto de valores que um atributo pode assumir. Por exemplo, o atributo “sexo” pode assumir {“feminino”, “masculino” e “intersexo”}; “data de nascimento” pode assumir valores entre “1900-01-01 e a data atual”.

Restrições são características que o atributo deve cumprir para ser armazenado de forma válida. Por exemplo, um campo texto restrito a 256 caracteres, ou um procedimento que apenas pode ser realizado em maiores de 18 anos.

Um **objeto** pode ser descrito por um ou mais atributos. Se os atributos estiverem ordenados sequencialmente, o objeto pode ser descrito na forma de um vetor, também chamado de **tupla** ou **registro**. Por exemplo, pessoa1 = {“Joana Freitas”, “2010-01-20”, “F”} cujos atributos são, respectivamente, {“nm_pessoa”, “dt_nascimento”, “sg_sexo”}. Um conjunto de vetores organizados em tuplas, onde cada linha é um objeto pertencente à mesma classe, perfaz uma matriz, também chamada **tabela**.

A **tabela** é uma apresentação em formato de **dados tabulares**, uma espécie de armazenamento onde os atributos são separados por um **caractere de tabulação**, isto é, um marcador de separação dos atributos e objetos. Os marcadores de separação de objetos usualmente são vírgula “,” ou “;”. Para separar objetos utiliza-se uma quebra de linha, geralmente “\n” ou “\r”. A tabela pode apresentar um cabeçalho em sua primeira linha ou em arquivo à parte. Dados organizados em tabelas, com atributos respectivos às colunas e cujos domínios e validade são devidamente mantidos, chamam-se **dados estruturados**”.

As tabelas expressam classes (objetos de um mesmo conjunto) e podem se relacionar. O **relacionamento** entre duas tabelas modifica, restringe ou amplia o signifi-

cado de dado objeto. Por exemplo, uma tabela da classe “pessoa” pode se relacionar com outra tabela chamada “diagnostico_pessoa” com atributos {co_pessoa, no_doenca, dt_diagnostico}. Assim, se em dado contexto uma pessoa puder apresentar mais de um diagnóstico, as tuplas relativas à pessoa1 com gripe ou tuberculose podem ser expressas como diagnostico1 = {1, “tuberculose”, 2022-05-01} e diagnostico2 = {1, “gripe”, 2023-06-18}.

O **grau do relacionamento**, também chamado de **cardinalidade**, é uma restrição da quantidade de relações que os objetos envolvidos podem assumir. As cardinalidades são zero “0”, um “1” ou muitos “N”. A notação de cardinalidade usualmente utiliza dois pontos “:” respectiva à ordem em que se denomina as tabelas. Por exemplo, 0:1, lê-se nenhum ou um; 1:N lê-se um ou muitos. “0” significa que o preenchimento não é obrigatório. No caso da relação “pessoa” - “diagnóstico_pessoa”, pode-se modelar para uma cardinalidade 0:N, onde nem todas as pessoas registradas contém diagnósticos.

Existem **dados não estruturados**, assim chamados por serem apresentados em linguagem natural e não estarem organizados com atributos constritos em domínios. Uma legislação, receita médica manuscrita ou digitada cursivamente, raio-X, ressonância ou a foto de um ferimento são dados não estruturados. A categorização e organização de metadados que viabilizem a recuperação eficaz de dados em linguagem natural, constituem os **dados semiestruturados**.

O dicionário de dados contendo atributos, descrição, domínio e restrições são orientados pelas **regras de negócio** obtidas a partir do levantamento de requisitos. Existem requisitos funcionais e não funcionais. Os **requisitos funcionais** são relativos ao que deve ser feito pelo sistema de informação. Os **requisitos não funcionais** são qualidades do sistema de informação, como especificações de segurança, desempenho, usabilidade, transparência, etc.. Os dados são produzidos para determinado contexto, mas podem ser utilizados em negócios diferentes. Por essa razão, ontologias são utilizadas para ampliar o uso para diferentes instituições, comunicando, inclusive, países com idiomas distintos.

O **dado primário** é produzido e utilizado para a finalidade definida pelo autor. O **dado secundário** é aquele utilizado fora do contexto da produção e por outros atores. Por exemplo, considere uma tabela produzida em uma Unidade Básica de Saúde (UBS), onde cada tupla contém dados ambulatoriais de um atendimento, cujos atributos são codigo do paciente, código do procedimento realizado, quantidade do procedimento realizado, valor total do procedimento realizado, data de atendimento, código do profissional atendente e código da ocupação do profissional atendente.

Os dados ambulatoriais são primários para o estabelecimento onde foi realizado o registro, bem como para o ente que irá realizar o pagamento ou ressarcimento, seja um ente público, por exemplo, o município, seja a operadora do seguro privado de saúde. Uma análise possível é monitorar o mesmo indivíduo em atendimentos diferentes, a qual usará o código e as datas para avaliar a reincidência ou a persistência no mesmo tratamento. Se os entes envolvidos na produção e sob o mesmo propósito de uso estiverem utilizando as informações, ainda que derivadas, se trata de **uso primário**, como o exemplo acima que descreveu o caso administrativo. Se um pesquisador utilizar para fazer asserções epidemiológicas, comparando, ainda, com outras fontes, ele estará fazendo **uso secundário** do conjunto de dados. Mais adiante, serão colocadas implicações éticas no uso de dados primários e secundários, contextualizando direitos pessoais e o direito de acesso à informação, conciliando demandas individuais e necessidades coletivas de transparência e prestação de contas e responsabilização (*accountability*).

Finalmente, é necessário conceituar dados transacionais e dados analíticos. A produção de dados é geralmente realizada com auxílio de aplicações de **persistência**, isto é, armazenam e mantêm os dados de forma consistente com as regras do negócio, em **bancos de dados transacionais**. Cada transação é um conjunto de operações que garantem a consistência do registro, contemplando características dos atributos, restrições de domínio e de relacionamentos. **Bancos de dados analíticos** são voltados para a detecção de padrões ao longo do conjunto de dados, sobretudo de forma agregada, sendo insumo de relatórios, painéis (*dashboards*) e ferramentas de inteligência de negócios (*Business Intelligence* - BI).

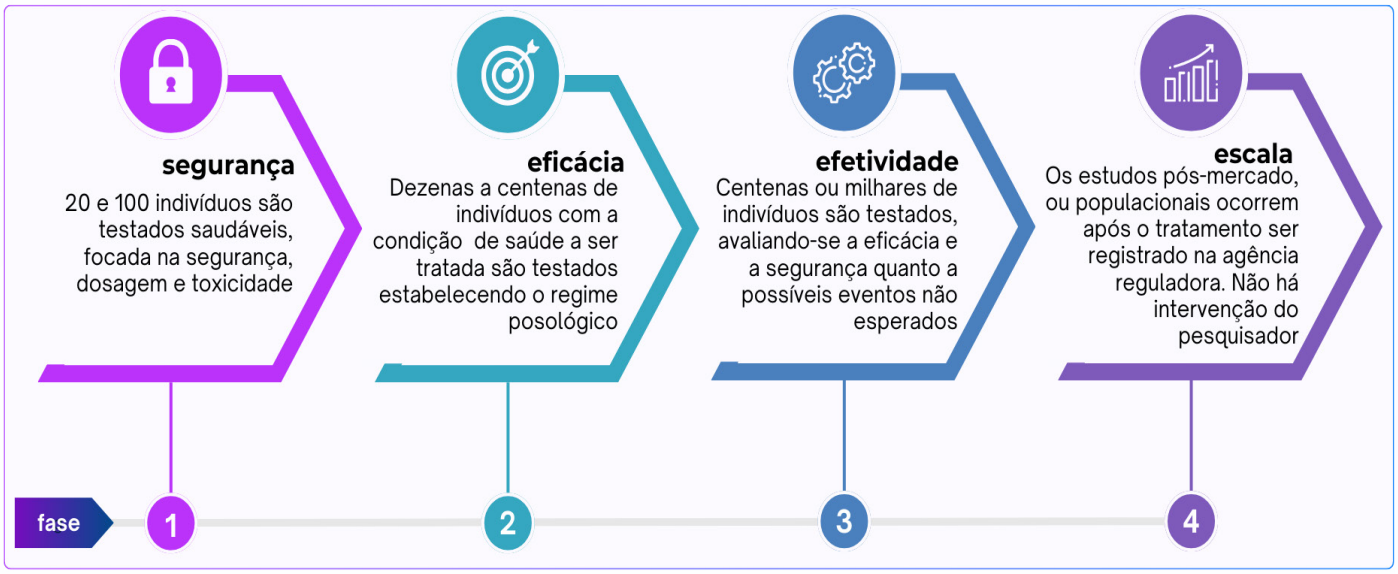
Ao longo do Microcurso, mais informações acerca das modalidades “produção” e “analíticos” serão vistas. Entretanto, é importante salientar que o banco de dados de produção é otimizado para consulta a dados individuais, por exemplo, o prontuário, o qual apresenta dezenas ou até centenas de registros e atributos. O banco de dados analítico é otimizado para estabelecer consultas em grande quantidade de dados (*big data*), geralmente com agregações que tecem relações entre atributos para avaliações estatísticas ou de aprendizado de máquina (*Machine Learning* - ML) voltadas para a gestão ou epidemiologia.

1.3 Quais São as Fontes de Dados Tabulares, Imagens e Textos na Área da Saúde?

Na saúde, os dados são produzidos para a pesquisa, gestão e cuidado. Na pesquisa, existe a modalidade clínica e a populacional, também chamada de epidemiológica ou observacional.

A **pesquisa clínica** faz testes em humanos e apresenta quatro fases. Na fase 1, são avaliados os efeitos em pequenos grupos, geralmente entre 20 e 100 indivíduos saudáveis, focada na segurança, dosagem e toxicidade. Na fase 2, um grupo maior é testado, contemplando indivíduos com a condição de saúde alvo pela terapêutica. Na fase 3, são tratados centenas ou milhares de indivíduos, avaliando-se a eficácia e a segurança quanto a possíveis eventos não esperados. Na fase 4, também chamada pós-mercado, o tratamento é registrado na agência reguladora e decorrem estudos populacionais (Figura 1).

Figura 1 - Dados de estudos clínicos



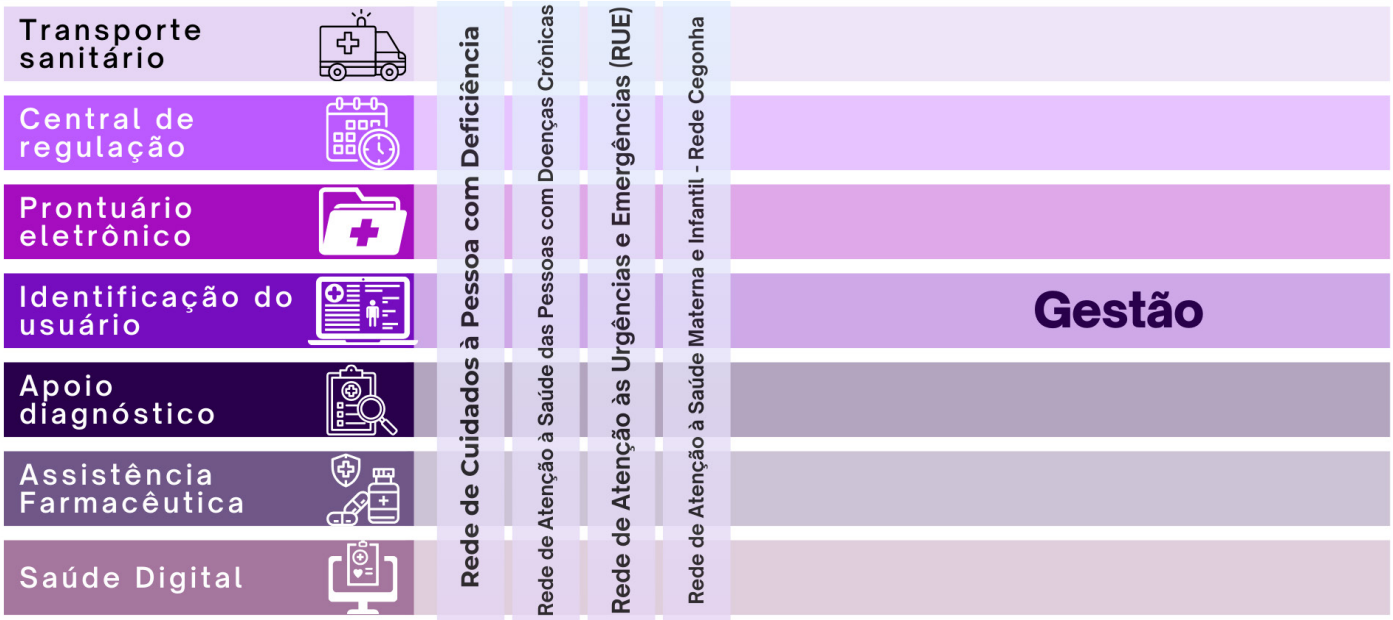
Fonte: autoria própria.

Estudos observacionais ocorrem quando não há intervenção do pesquisador no conjunto de indivíduos avaliado, onde exposição e efeitos são avaliados em função do tempo, seja simultaneamente em estudos transversais, seja tomando-se precedentes e consequentes em estudos de caso-controle ou coortes (Lima-Costa; Barreto, 2003), também chamados de estudos longitudinais.

Figura 2 - Níveis de evidência

Fonte: Sackett *et al.* (2000).

Figura 3 - A estrutura operacional das redes de atenção à saúde

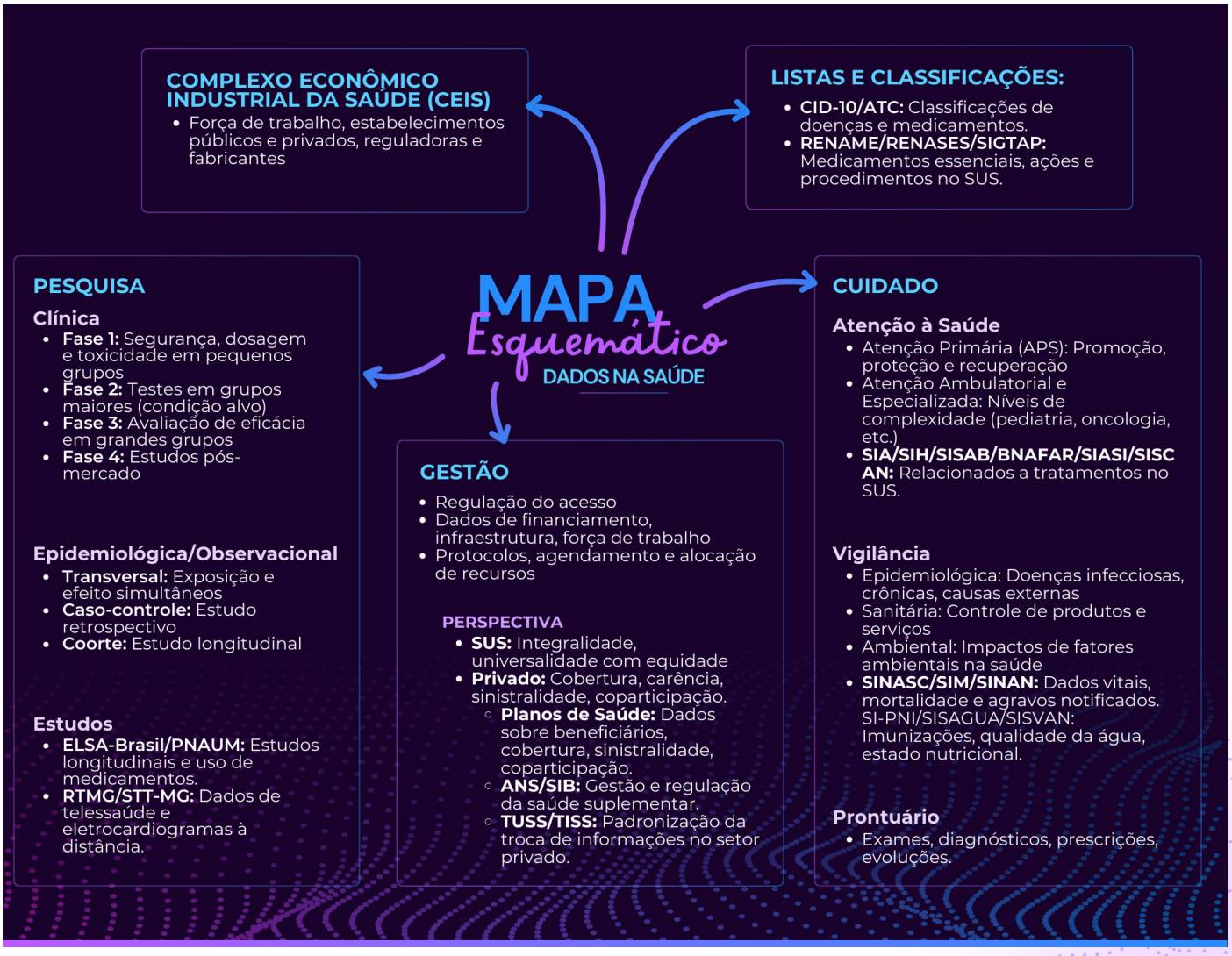


Fonte: adaptado de Vilaça (2011).

Os dados de **gestão** são originados de entidades públicas ou privadas, sobretudo para a **regulação** do atendimento, isto é, disponibilização do acesso mediante protocolos ou segmentação econômica das Ações e Serviços de Saúde (ASP) imediata ou agendada. Ambos os segmentos produzem dados de financiamento, normativos, infraestrutura, força de trabalho, regulação do acesso com agendamento e alocação de recursos mediante critérios estabelecidos em protocolos e diretrizes, autorização

de atendimento e notificações. Na iniciativa privada existe o acesso particular com desembolso direto ou ressarcido a partir do seguro saúde, cuja regulação acrescenta a carência, isto é, tempo para o primeiro atendimento de dada condição de saúde e procedimento visando afastar condições pré-existentes. Existe, ainda, o expediente do reembolso ou coparticipação mediante o tipo do produto comercializado, o qual prevê a lista de atendimentos básicos do plano de saúde. Os vocabulários são diferentes também. Em sistemas universais, como o brasileiro, existe a **integralidade**, isto é, todas as condições de saúde são contempladas e a **universalidade** com **equidade**, uma vez que não existe restrição para quem deve ser atendido e os que precisam mais, devem receber maior prioridade das ações, sendo que denomina-se a todos como usuários do SUS; no plano de saúde são denominados **beneficiários** ou vidas, segmentados conforme a restrição de acesso aos produtos, chamada **cobertura**. As condições de saúde tipificadas em sistemas públicos se tornam **sinistralidade** no léxico do setor privado, visto que cada uma abrange um componente que vista a sustentabilidade econômica do empreendimento.

Figura 4 - Mapa mental sobre os classificadores de aprendizado de máquina



Fonte: autoria própria.

O conjunto de entes públicos e privados, força de trabalho, estabelecimentos de saúde, agências reguladoras, fabricantes de insumos e produtores de conhecimento constitui o Complexo Econômico Industrial da Saúde (CEIS), objeto de estudo da área do saber denominada Economia da Saúde (Figura 4).

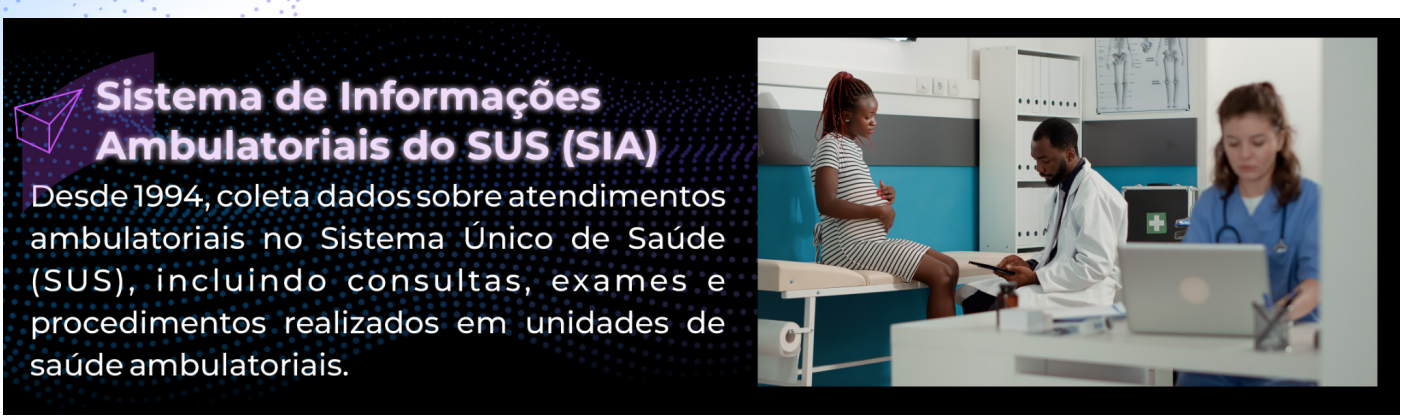
O **cuidado**, tanto do ponto de vista clínico ou enquanto medida de saúde pública, é segmentado em atenção à saúde e vigilância. Existe a **Atenção Primária à Saúde** (APS), focada na promoção, proteção da saúde e recuperação integral num contexto multiprofissional e abrange a complexidade do modo de vida do território e da saúde da família. A **Atenção Especializada** abrange outros níveis média e alta complexidade, porém, com atendimento direcionado segundo a área biomédica como pediatria, oncologia, neurologia, psiquiatria, nefrologia, ortopedia, farmácia clínica, odontologia, obstetrícia, cardiologia, ginecologia, oftalmologia entre outras especialidades (Brasil, 2010b). A ferramenta, por excelência, da atenção à saúde é o prontuário.

O prontuário contém exames, laudos, diagnósticos, relatos de condições de saúde, prescrições e evoluções. As ações de vigilância são denominadas epidemiológica, sanitária ou ambiental, bem como são segmentadas por subgrupo populacional, por exemplo, laboral ou de saúde do trabalhador, de nascidos vivos e de causas externas de óbito. A **vigilância epidemiológica** trabalha com doenças infecciosas, doenças crônicas não transmissíveis (DCNT) e causas externas, como acidentes, agravos à saúde em consequência de violência, afogamento, queimaduras, envenenamento e quedas. A **vigilância sanitária** regulamenta, controla e fiscaliza produtos e serviços que envolvam risco à saúde pública (Brasil, 1999). A **vigilância ambiental** é dedicada a estudar e reduzir impactos à saúde humana de fatores ambientais naturais ou artificiais.

Podemos dividir, também, os dados públicos segundo a origem e finalidade, sendo fundamentalmente dados de produção ou padrões, os quais incluem, listas, classificações ou ontologias. Nas seções seguintes, iremos abordar mais conceitos ao aplicar alguns dos itens listados abaixo.

Um profissional da saúde atualizado conhece as principais fontes de dados e listas correlatos aos sistemas de informação de seu país. As fontes de acesso público estão assinaladas com o ano de início, sendo algumas destas utilizadas no presente curso. Veja, a seguir, alguns exemplos (Figuras 5-9).

Figura 5 - Gestão da atenção à saúde



Fonte: autoria própria.

Figura 6 - Vigilância em saúde

Fonte: autoria própria.

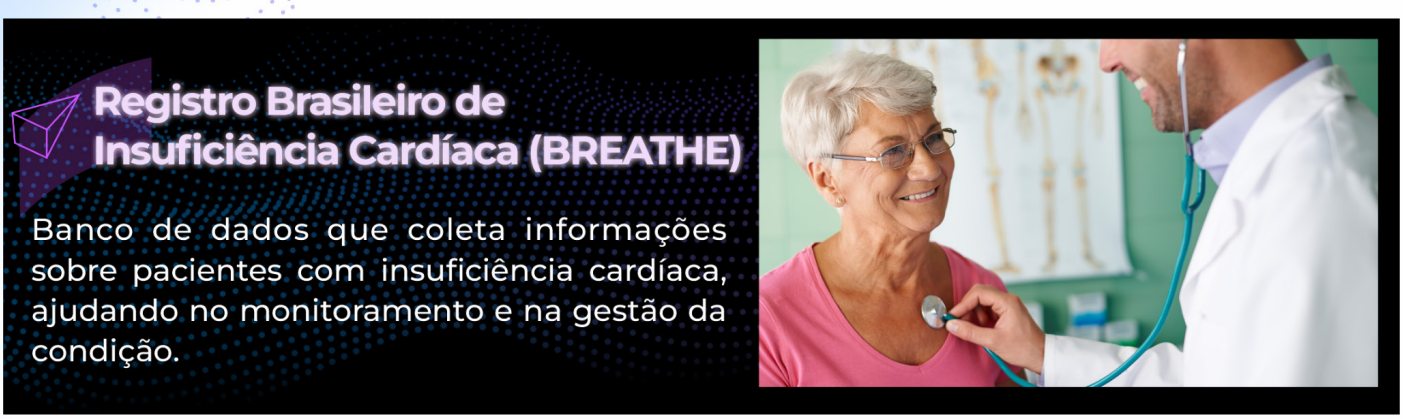
Figura 7 - Sistemas de gestão

Fonte: autoria própria.

Figura 8 - Condições de saúde

Fonte: autoria própria.

Figura 9 - Outras fontes de dados



Fonte: autoria própria.

1.4 Saiba Mais - Atividade de Leitura Opcional

1.4.1 Cargos e Funções na Gestão da Tecnologia da Informação em Saúde

Os principais cargos e funções na gestão da TI em saúde são: **desenvolvedores de *software*,** responsável por escrever e manter o código-fonte das aplicações, garantindo as funcionalidades previstas. Eles trabalham em estreita colaboração com os **analistas de requisitos**, que identificam e documentam as necessidades dos usuários e do negócio, transformando essas necessidades em especificações técnicas claras e detalhadas para orientar o desenvolvimento do *software*.

Os **administradores de banco de dados** (*Database Administrator* - DBA) gerenciam e mantêm os sistemas de banco de dados, assegurando a integridade, disponibilidade e segurança dos dados. Junto a eles, os **analistas de BI** trabalham na coleta, análise e interpretação dos dados, gerando visões para a tomada de decisões estratégicas.

Os **engenheiros de qualidade**, também conhecidos como *testers*, são responsáveis por garantir que o *software* atenda aos padrões de qualidade definidos, realizando testes rigorosos e identificando bugs ou falhas no sistema. Eles colaboram frequentemente com os **arquitetos de *software***, que projetam a estrutura do sistema no maior nível de abstração, definindo os padrões e tecnologias a serem utilizados para garantir a escalabilidade, performance e robustez do *software*.

Os **gestores de projetos** coordenam as atividades da equipe, garantindo que os prazos e orçamentos sejam respeitados e que a comunicação entre os membros da equipe e os *stakeholders* seja eficaz. Eles utilizam metodologias de gerenciamento de projetos para monitorar o progresso e resolver quaisquer impedimentos que possam surgir durante o desenvolvimento.

Os ***designers*** **de Interface de Usuário** (UI) e de **Experiência do Usuário** (UX) são responsáveis por criar interfaces intuitivas e agradáveis, garantindo que os usuários tenham uma experiência positiva e eficiente ao interagir com o *software*. Enquanto os designers de UI focam na aparência visual do aplicativo, os designers de UX concentram-se na usabilidade e na jornada do usuário.

Finalmente, os **analistas de segurança** garantem que o *software* e os dados estejam protegidos contra ameaças e vulnerabilidades, implementando medidas de segurança eficazes e realizando auditorias regulares para identificar e corrigir possíveis brechas.

No contexto de um **escritório de projetos**, os gestores de projetos são apoiados por uma equipe diversificada de profissionais. Entre esses integrantes, podemos encontrar **planejadores de projetos**, que são responsáveis por definir cronogramas detalhados e recursos necessários; analistas de projetos, que monitoram o progresso e a performance dos projetos; **controladores de custos**, que gerenciam orçamentos e asseguram a eficiência financeira; e **coordenadores de comunicação**, que facilitam a troca de informações entre as partes interessadas. Além disso, há especialistas em metodologias de gerenciamento de projetos que garantem a aplicação correta das práticas e padrões adotados pela organização. Este conjunto de profissionais trabalha de forma coesa para garantir que os projetos sejam executados dentro do prazo, do orçamento e com a qualidade esperada.

Os métodos ágeis, como Scrum, Kanban e XP (*Extreme Programming*), são abordagens iterativas e incrementais para o gerenciamento de projetos, particularmente populares no desenvolvimento de *software*. Eles se concentram em entregas frequentes e de alta qualidade, com ciclos de trabalho curtos denominados *sprints* ou iterações. A metodologia ágil promove a colaboração estreita entre equipes multifuncionais e os clientes, permitindo ajustes rápidos às mudanças de requisitos. Priorizando a comunicação direta, a transparência e a flexibilidade, os métodos ágeis ajudam as equipes a responderem eficientemente a incertezas e a fornecer valor contínuo aos usuários finais. Eles também enfatizam a melhoria contínua por meio de revisões regulares e *feedback* constante, buscando sempre otimizar processos e resultados.

A título de exemplo, a **metodologia Scrum**, uma das abordagens ágeis mais utilizadas, aplica os conceitos de *Product Backlog*, *Release Backlog* e *Sprint Backlog* para a organização e priorização do trabalho. O ***Product Backlog*** é uma lista dinâmica e priorizada das funcionalidades e melhorias desejadas para o produto, servindo como um guia para o desenvolvimento contínuo. O ***Release Backlog*** é uma seleção de itens do *Product Backlog* que foram priorizados para serem trabalhados em uma versão específica do produto. Dentro do ***Sprint Backlog***, a equipe define as tarefas específicas que serão realizadas durante o sprint, um ciclo de trabalho que geralmente dura de duas a quatro semanas. O ***Product Owner* (PO)** é o responsável por gerenciar o *Product Backlog* e assegurar que o produto entregue atenda às necessidades dos stakeholders. O ***Scrum Master*** atua como um facilitador, garantindo que a equipe siga as práticas Scrum e ajudando a remover impedimentos que possam surgir. A equipe de desenvolvimento, composta por diversos profissionais com habilidades complementares, é responsável por transformar os itens do *Sprint Backlog* em funcionalidades incrementais, entregando valor contínuo ao produto e, por consequência, aos seus usuários.

1.4.2 Noções de Semiologia e Semiótica

A semiologia e a semiótica são termos que designam o estudo dos signos, mas podem ser utilizados em diferentes contextos. A semiologia, conforme proposta pelo linguista suíço Ferdinand de Saussure (1857-1913), refere-se ao estudo dos sinais dentro dos sistemas de comunicação e significação, com enfoque especial na linguagem. No campo da saúde, no entanto, a semiologia designa o estudo dos sinais e sintomas das doenças. A semiótica, ciência que também estuda os signos, abrange não apenas os sistemas linguísticos, mas também os sistemas de signos não linguísticos, e investiga os processos de significação tanto na cultura quanto na natureza. Enquanto a semiótica é a ciência geral dos signos, abrangendo também os signos presentes na natureza não humana, a semiologia é uma ciência voltada ao estudo dos fenômenos humanos, indo além da linguística para investigar fenômenos translinguísticos, como textos e códigos culturais.

Os conceitos colocados a seguir são, apesar de abstratos, utilizados quotidianamente no manejo de informações em saúde. Existem processos de extração de conhecimento (do inglês, *knowledge-Discovery in Databases* - KDD) cujo caminho entre a informação, desinformação e comunicação, requer um conhecimento teórico que viabilize compreender os processo gerais de como a inteligência humana pode ser ampliada com a inteligência computacional em processos cada vez mais automatizados.

Ao realizar uma primeira leitura não se preocupe caso não fixe determinado significado. Haverá oportunidade para aplicar e estender a compreensão do vocabulário-base ao longo do curso, sobretudo com os exercícios. Fundamentalmente, o curso se baseará em dados do SUS, o qual contempla agências, entidades públicas e privadas, bem como a saúde suplementar regulada pela ANS.

O **dado** é um fato da realidade. Quando situamos o dado num contexto obtemos uma **informação**. Por exemplo, para o dado de temperatura "38°C", é necessário apurar outros dados, como "trata-se de um adulto" e "sob condições normais de repouso", a relação entre dados constitui a informação "febre". **Conhecimento** é a capacidade de armazenar e recuperar um conjunto de informações, distinguir quando e como aplicar numa tomada de decisão e orientar a construção de novos conhecimentos. Continuando o exemplo acima, conhecimento é saber quando se deve aplicar um antipirético. **Inteligência** é a capacidade de extrapolar a tomada de decisão para além dos dados coletados, inclusive, orientando novas coletas de dados e decisões adaptadas a cada situação. Assim, mediante o conhecimento de diversos casos semelhantes, pode-se prognosticar uma infecção ou ainda avaliar se está havendo um surto. O conhecimento e a inteligência estão atrelados à **memória**, uma vez que se trata de organizar conteúdo com base em interações com o ambiente e coordenar ações futuras, reorientando nova coleta e armazenamento de dados e informação.

Autonomia é a capacidade de adaptar o comportamento ao interagir com o ambiente. Autômatos operam mediante estados pré-definidos e decisões lógicas. **Autômatos** são modelos matemáticos que representam sistemas com entradas e saídas configuradas, capazes de transitar entre diferentes estados conforme regras específicas. Máquinas de estados são uma implementação prática desses autômatos, usadas para modelar o comportamento de sistemas dinâmicos que respondem a eventos ou condições. A lógica combinacional, por sua vez, trata de circuitos cujas saídas são determinadas diretamente pelas combinações de suas entradas, sem depender de estados anteriores. Quando aplicamos esses conceitos ao campo da aprendizagem de máquina e Inteligência Artificial (IA), vemos um paralelismo onde modelos de aprendizagem supervisionada ou não supervisionada atuam como autômatos, processando dados de entrada para aprender padrões e tomar decisões. A máquina de estados se reflete nos algoritmos de aprendizado que ajustam seus parâmetros por meio de iterações, enquanto a lógica combinacional encontra seu equivalente em redes neurais artificiais, onde as combinações de pesos e ativações determinam as saídas do modelo. O acoplamento entre conceitos clássicos e avançados permite a criação de sistemas inteligentes capazes de realizar tarefas complexas com eficiência e precisão e será visto na Unidade 4.

As definições acima servem tanto para o mundo natural, quanto para o artificial, sendo desejável tecer analogias com seu campo de trabalho para o informacional. Uma célula armazena dados na forma de bases nitrogenadas e o disco rígido do computador na forma de sinais elétricos. Nesse nível, podemos dizer que ribossomos e processadores são os interpretadores dos dados.

Na saúde, uma grande dificuldade é assegurar a semântica, isto é, a interpretação mediante vocabulários distintos para o mesmo objeto ou denominações semelhantes para entes distintos, o que chamamos, respectivamente, de termos unívocos e equívocos ou degenerados. O **sinal** é o dado registrado em fase de interpretação, o qual é composto por signo, significado e significante, elementos fundamentais da informação e, consequentemente, da comunicação. O **signo** é o átomo da linguagem, o código a ser decomposto a partir do elemento perceptível, o **significante**, e interpretado, o **significado**. Por exemplo, o significante "F", no contexto de dados cadastrais, pode significar "sexo feminino". O significante "G" tem significado de "guanina" quando nos referimos ao ácido desoxirribonucleico (DNA). No caso do processo celular, o significante será a própria molécula de guanina, com seus átomos de carbono, nitrogênio e oxigênio devidamente organizados numa conformação molecular.

Uma **linguagem** é formada por átomos de sinais encadeados e relacionados de modo a viabilizar a comunicação entre diferentes indivíduos. Uma **ontologia** é um arcabouço que viabiliza a comunicação entre diferentes linguagens e, portanto, indivíduos e instituições, conciliando diferentes significantes e significados para o mesmo objeto conforme o contexto. A comunicação entre diferentes aplicações informatizadas é conhecida como **interoperabilidade**. A **integração** é um estágio anterior, menos sofisticado, onde as soluções não trabalham conjuntamente, mas existe uma tradução dos termos e transposição das informações armazenadas por uma ou ambas as partes detentoras dos dados.

Um conjunto de indivíduos que compartilham e mantém a mesma linguagem é uma **comunidade**. É habitual usarmos o termo de comunidade científica ou comunidade de *software* com indivíduos que usam as mesmas terminologias técnicas e instrumentais. A linguagem oferta a capacidade de abstração para a tradução de objetos em sinais que podem ser compartilhados e mutuamente entendidos no ato da **comunicação**.

Os repositórios são, portanto, mantidos com regras derivadas de **modelos**, cujas simplificações do mundo levam à alimentação de parâmetros de entrada que derivam respostas, ou saídas, esperadas em situações reais. Reiterando, os modelos são abstrações do mundo real mantidas por comunidades. Uma comunidade menor pode utilizar uma linguagem mais ampla para estabelecer vocabulário próprio

como extensão de uma linguagem mais difundida. Por exemplo, o português é uma linguagem comum de uma ampla comunidade, mas existem extensões da linguagem que perfazem o vocabulário da saúde, o qual pode apresentar sub vocabulários conforme a especialidade, cada uma aplicando seu modelo para tomada de decisão, frequentemente nominando diferentemente o mesmo objeto.

A **informática** concilia tanto o conhecimento das ciências da informação, que incluem a **semiótica**, conforme descrito nos parágrafos anteriores, quanto a computação, isto é, a ciência que utiliza a matemática para processar dados. A **informática em saúde** reúne a computação e a **semiologia**. Durante a graduação, o estudante se debruça na base do conhecimento da saúde e desenvolve uma inteligência semiológica acerca dos sinais do processo saúde e doença para tecer prognósticos, realizar a terapêutica, cuidar de populações e, consequentemente, construir políticas públicas.

A informática em saúde se apropria de certos vocabulários da informática, sendo fundamental conhecer o léxico aqui abordado para haver comunicação assertiva entre diferentes perfis profissionais com a equipe que lida com Tecnologia da Informação em Saúde (TIS), ou ainda, Tecnologia da Informação e Comunicação em Saúde (TICS).

1.4.3 Fontes de Dados e Métodos de Pesquisa

Nesta Unidade, foram abordadas as principais fontes de dados em saúde. A seguir serão exemplificadas outras fontes de dados. Uma fonte de dados de saúde advém de sistemas embarcados, presentes em dispositivos numa convergência da engenharia eletrônica com a informática, integrando *hardware* e *software* para desempenhar funções específicas. Esses sistemas são projetados para operar com recursos limitados, tanto em termos de processamento quanto de memória, exigindo uma otimização minuciosa para garantir desempenho e eficiência. No campo da saúde estão presentes em dispositivos como monitores cardíacos implantáveis que rastreiam e transmitem dados vitais para profissionais de saúde em tempo real, melhorando significativamente o gerenciamento de condições crônicas. Esses sistemas também são encontrados em máquinas de ressonância magnética e tomografia computadorizada, onde controlam os processos complexos de aquisição e processamento de imagens, garantindo diagnósticos precisos. Além disso, bombas de infusão inteligentes utilizam sistemas embarcados para administrar medicamentos de forma controlada e segura, ajustando as dosagens de acordo com parâmetros predefinidos. Assim, tecnologia embarcada pode ofertar maior precisão e segurança no tratamento de pacientes.

A Indústria 4.0 na saúde é caracterizada pela integração de tecnologias avançadas que transformam os cuidados profissionais e a gestão da saúde. Por exemplo, a Internet das Coisas (IoT) conecta dispositivos e sistemas, permitindo a coleta e análise contínua de dados dos pacientes; a IA, que potencializa diagnósticos precisos e tratamentos personalizados com algoritmos avançados e aprendizado de máquina; e a Realidade Aumentada (AR) e Realidade Virtual (VR), que aprimoram a formação médica e a realização de procedimentos complexos em cirurgias e telessaúde. Além disso, a impressão em 3 Dimensões (3D) possibilita a criação de próteses personalizadas e modelos anatômicos precisos para planejamentos cirúrgicos. A robótica, por sua vez, apresenta avanços em cirurgias com procedimentos minimamente invasivos e assistidos por robôs. As tecnologias, interconectadas por meio de redes de alta velocidade e apoiadas pela análise de big data, estão moldando um novo paradigma de cuidado à saúde personalizada e populacional.

Vestíveis podem ser considerados sistemas embarcados, pois são dispositivos eletrônicos projetados para serem usados no corpo e possuem hardware e software integrados que executam funções específicas. Esses dispositivos, como smartwatches, monitores de fitness e sensores de saúde, coletam dados biométricos em tempo real, como frequência cardíaca, níveis de atividade física e padrões de sono. Os vestíveis processam e analisam esses dados localmente ou os transmitem para outros dispositivos e aplicativos para posterior análise, contribuindo para a medicina preventiva e o monitoramento contínuo da saúde, transformando como as pessoas interagem com suas próprias informações de saúde.

Terminologias, ontologias e padrões

- **HL7 FHIR:** *Fast Healthcare Interoperability Resources* (FHIR), um padrão de mensageria para troca de informações em saúde assegurando a semântica, sob o conjunto de normas internacionais *Health Level* 7 (HL7). Desenvolvido pela organização HL7, o FHIR combina as melhores características dos padrões anteriores com tecnologias *web* modernas, como RESTful APIs e JSON *(JavaScript Object Notation*)/XML (*Extensible Markup Language)*. Ele permite que dados clínicos sejam acessados e compartilhados de maneira mais eficiente e segura, promovendo uma melhor integração entre sistemas de Registros Eletrônicos de Saúde (RES), aplicativos móveis e outras plataformas de saúde digital. O FHIR é altamente modular, permitindo que desenvolvedores utilizem apenas os componentes necessários para suas aplicações específicas.

- A **Ontologia Brasileira de Medicamentos (OBM)** é uma estrutura formal que visa padronizar e integrar informações sobre medicamentos no Brasil, facilitando a interoperabilidade entre Sistemas de Informação em Saúde (SIS). Baseada no modelo IDM+P (Identificação, Definição e Modelagem de Produtos), a OBM contempla elementos essenciais como a Identificação única de cada

medicamento (ID), definições claras e padronizadas de seus componentes e características (Definição), e a modelagem detalhada dos produtos farmacêuticos, incluindo formas farmacêuticas, dosagens e apresentações (Modelagem de Produtos). Essa ontologia viabiliza o rastreamento, controle de qualidade e acesso a informações críticas, apoiando tanto a prática clínica, a pesquisa e a gestão farmacêutica.

» ***openEHR***, para gestão, armazenamento, recuperação e troca de RES (EHR, do inglês, *Electronic Health Records*). A *openEHR* é uma abordagem aberta e padronizada para a gestão de RES, focada na separação entre dados clínicos e aplicações. Desenvolvido pela Fundação *openEHR*, este padrão fornece um modelo de dados robusto e flexível que permite a criação e armazenamento de registros de saúde interoperáveis. Os dados são estruturados em "arquetipos" e "templates," que são definíveis e extensíveis, permitindo que as informações de saúde sejam armazenadas de maneira granular e reutilizável. Esta abordagem facilita a troca de informações clínicas entre diferentes sistemas e organizações, promovendo a continuidade do cuidado e a pesquisa médica.

» **SNOMED CT**, fornece termos clínicos (*Clinical Terms* - CT) com códigos, sinônimos e definições. Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) é a terminologia clínica mais abrangente e precisa do mundo, utilizada para codificar os dados de saúde de forma estruturada. Desenvolvido pela Systematized Nomenclature of Medicine (SNOMED) International, SNOMED CT oferece um vocabulário padronizado que cobre doenças, sintomas, procedimentos, medicamentos e outros aspectos relevantes da assistência médica. A terminologia favorece a interoperabilidade dos sistemas de saúde, permitindo que informações clínicas sejam registradas, recuperadas e analisadas de forma consistente e precisa. SNOMED CT é amplamente utilizada para suporte à decisão clínica, interoperabilidade de RES e pesquisa em saúde.

» **LOINC** (*Logical Observation Identifiers Names and Codes*), que padroniza códigos para testes laboratoriais e observações clínicas;

» **UMLS** (*Unified Medical Language System* - Sistema Unificado de Linguagem Médica);

» **KEGG** (*Kyoto Encyclopedia of Genes and Genomes* - Enciclopédia de Genes e Genomas de Quioto) são ontologias incluindo mapas metabólicos, hierarquias, funcionalidades, reações etc;

» ***Gene Ontology* (GO)**, ou Ontologia Genética, para representação de genes e atributos de produtos genéticos de todas as espécies, inclusive as humanas;

» **DICOM** (*Digital Imaging and Communications in Medicine*) para a gestão, armazenamento, impressão e transmissão de informações em imagens de saúde, visando a integração de equipamentos de imagem e sistemas de arquivamento e comunicação de imagens (do inglês, *Picture Archiving and Communication System* - PACS).

Leia mais sobre os tipos de pesquisa (Fontelles *et al*., 2009; Lakatos, 1983; Pozobon; De David, 2019):

» FONTELLES, M. J. et al.. Metodologia da pesquisa científica: diretrizes para a elaboração de um protocolo de pesquisa. Revista paraense de medicina, v. 23, n. 3, p. 1–8, 2009.

» LAKATOS, E. M. Metodologia científica. [S.l.]: Atlas, 1983.

» Conheça as definições de dados e bancos de dados para fins de proteção de direitos individuais (Brasil, 2018a, 2024).

» Banco de Dados: Conjunto estruturado de dados relativos a pessoas, que pode ser armazenado em formato eletrônico ou físico.

» Dados: Unidades de descrição das variáveis que compõem um banco de dados, podendo ser representadas por palavras, números, símbolos, imagens, entre outros.

» Dado Anonimizado: Dado que não pode ser associado a um titular específico devido ao uso de técnicas de anonimização.

» Dado Pessoal: Informação relacionada a uma pessoa identificada ou identificável.

» Dado Pessoal Sensível: Dado pessoal que revela origem racial ou étnica, convicções religiosas, opiniões políticas, filiação sindical, dados sobre saúde ou vida sexual, dados genéticos ou biométricos.

» Dado Identificado: Informação que pode ser vinculada diretamente à identidade de um indivíduo.

Vamos destacar, aqui, a pesquisa com dados observacionais. Os dados administrativos do SUS são fontes de várias publicações, incluindo dados abertos.

Os dados sensíveis, por exemplo, contendo nome, nome da mãe, endereço, data de nascimento, entre outras informações sociodemográficas são corriqueiramente usados em instituições de pesquisa em saúde coletiva. Descubra como são reunidos

os registros de sistemas de informação diferentes do mesmo indivíduo. A técnica de vinculação de registros de origens distintas a um dado indivíduo chama-se pareamento determinístico-probabilístico (*record linkage*).

No Brasil, diversos institutos estão afiliados à *International Data Likage Network* ("IPDLN *network*", 2024). Ao longo das últimas décadas receberam dados identificados, isto é, com nome, nome da mãe, endereço, número da identidade, etc e vincularam os dados ao longo de sistemas de informação diferentes. Conheça alguns destaques de estudos e outras fontes de dados.

» Centro de Integração de Dados e Conhecimento para Saúde (CIDACS) da Fundação Oswaldo Cruz (Fiocruz). A "Plataforma de estudos e avaliações contínuas dos determinantes sociais e efeitos do Programa Bolsa Família e outros Programas de Proteção Social sobre a saúde – Coorte de 100 milhões de brasileiros" é uma iniciativa para investigar como os determinantes sociais e as políticas públicas afetam a saúde da população brasileira. Este projeto integra dados de diversos programas sociais com informações de sistemas de saúde, abrangendo mortalidade, nascimento, doenças infecciosas e crônicas, entre outros desfechos. Realizada através de uma cooperação técnica entre instituições como a Fiocruz, Universidade de Brasília (UnB), Universidade Federal da Bahia (UFBA) e o Ministério do Desenvolvimento Social e Combate à Fome (MDSA), a coorte busca gerar conhecimento que possa subsidiar a tomada de decisões em políticas sociais, especialmente aquelas voltadas para a redução da pobreza e desigualdades. A pesquisa envolve estudos sobre desigualdades sociais, impacto das políticas públicas na saúde materno-infantil, doenças infecciosas e violência, além de desenvolver metodologias para avaliação do impacto dessas políticas. Leia mais em (Barreto *et al.*, 2022; "Coorte de 100 Milhões de Brasileiros", [s.d.]; Nery *et al.*, 2019).

» O Centro Colaborador do SUS para Avaliação de Tecnologias e Excelência em Saúde (Ccates) da Universidade Federal de Minas Gerais (UFMG) se dedica a promover a qualidade na prescrição, dispensação e uso de medicamentos, procedimentos e outras tecnologias em saúde, através do desenvolvimento de estudos e avaliações que auxiliam na tomada de decisões em saúde. Suas atividades incluem avaliações de tecnologias em saúde, estudos em economia da saúde, epidemiologia e farmacovigilância, pesquisas sobre a efetividade clínica comparativa (pesquisa no mundo real), projetos de desenvolvimento para a formação de recursos humanos, elaboração de laudos e notas técnicas, produção de boletins informativos, desenvolvimento, avaliação e implementação de sistemas informatizados para RES focados no usuário e visando a eficácia e excelência clínica, monitoramento de horizonte tecnológico, vinculação de registros administrativos de saúde de forma determinística-

probabilística, e disseminação de evidências em saúde por meio de visitas acadêmicas detalhadas. Leia mais em "*Activities*" (CCATES, 2024) e Junior *et al.* (2018).

» O Centro de Inovação em Inteligência Artificial para a Saúde (CI-IA-Saúde) dedica-se a desenvolver tecnologias avançadas para a prevenção e tratamento de doenças, com um enfoque especial na vinculação de dados. Utilizando bases de dados nacionais, dispositivos vestíveis, sensores virtuais, e dados genéticos, o centro integra informações heterogêneas de fontes como redes sociais, centros clínicos, mobilidade e censos para complementar os dados do Sistema de Informação de Saúde (SIS). Por meio de modelos de IA, o CI-IA-Saúde identifica indivíduos e populações em risco de doenças crônicas, permitindo a implementação de políticas públicas eficazes e recomendações personalizadas para melhorar a saúde e a qualidade de vida. Leia mais em ("CI-IA Saúde – Centro de Inovação em Inteligência Artificial para a Saúde", (CI-IA Saúde, 2024; Ribeiro *et al.*, 2019).

» Veja mais sobre as fontes de informação conforme o Sistema de Informação em Saúde existentes no Brasil (Coelho Neto; Chioro, 2021; Leal *et al.*, 2021a).

Retomando a Lei Geral de Proteção de Dados Pessoais (LGPD), novas conformidades devem ser atendidas pelo pesquisador, o que representa uma mudança no cenário de pesquisa que exige a figura do gestor de dados para seu cumprimento. Veja destaques a seguir:

» Exigência de Medidas Rigorosas de Segurança: A resolução impõe a adoção de medidas de segurança para garantir a confidencialidade e integridade dos dados (Art. 4°, IV; Art. 9°, I);

» Necessidade de Controle Rastreável de Acesso: Acesso aos bancos de dados deve ser restrito, controlado e rastreável, o que pode requerer infraestrutura tecnológica avançada e procedimentos rigorosos (Art. 4°, V).

» Anonimização e Pseudonimização: anonimização de Dados: Dados pessoais identificadores devem ser removidos quando os dados forem depositados em bancos de dados de acesso público ou restrito (Art. 7°). Este processo pode ser tecnicamente desafiador e caro. Justificativa para Não Anonimização: Se a anonimização irreversível não for possível, isso deve ser justificado no protocolo de pesquisa, ponderando riscos e benefícios (Art. 23, parágrafo único).

» Consentimento Informado: Necessidade de Consentimento Prévio: Para a inclusão e utilização de dados dos participantes, é necessário obter consentimento prévio, exceto se haja dispensa justificada pelo Sistema CEP/Conep (Art. 19, §1°-§5°).

- Reobtenção de Consentimento: Caso os dados não tenham sido autorizados para uso futuro, será necessário obter novo consentimento dos participantes para novas pesquisas (Art. 25).
- Transferência de Dados: Autorização para Transferência: Transferências de dados identificadores a terceiros devem estar previstas no protocolo de pesquisa e justificadas, e devem ser realizadas pelo Controlador do Banco de Dados com meios seguros que permitam rastreabilidade (Art. 11).
- Formalização de Transferências: A transferência de dados para terceiros deve ser formalizada por meio de um Termo de Transferência de Informações de bancos de dados (Art. 11, §1°).
- Documentação e Procedimentos Adicionais:
  - Protocolos Detalhados: Protocolos de pesquisa devem incluir descrição detalhada dos dados, mecanismos de segurança, estratégias de acesso e armazenamento, e critérios para compartilhamento e transferência de dados (Art. 18).
  - Termos de Compromisso e Anuência: Utilização de bancos de dados já constituídos requer a apresentação de vários documentos, incluindo Termos de Compromisso de Uso de Dados e Termos de Anuência Institucional (Art. 26).
- Direitos dos Participantes:
  - Direito de Acesso e Retificação: Os participantes têm direito de acessar, retificar, atualizar e solicitar a retirada parcial ou total de suas informações (Art. 15 e Art. 16).
  - Indenização por Danos: participantes têm direito a indenização caso haja uso indevido ou quebra de segurança/confidencialidade dos dados (Art. 17).
- Complemento com a LGPD:
  - A LGPD reforça muitos dos princípios e requisitos estabelecidos pela nova resolução, aumentando ainda mais as dificuldades para os pesquisadores.
  - Bases legais para o tratamento de dados e consentimento, visto que a obtenção e gestão de consentimento é um requisito essencial tanto na LGPD quanto na resolução do Conselho Nacional de Saúde.
  - Justificativa legal para pesquisadores, pois precisam assegurar que têm uma base legal clara para o tratamento dos dados, como consentimento ou execução de políticas públicas (Art. 7° da LGPD).
  - Direitos dos titulares dos dados como direitos de acesso, retificação e eliminação, uma vez que a LGPD garante aos titulares o direito de acesso,

correção e eliminação de seus dados, o que é reforçado pela resolução (Art. 18 da LGPD).

- Segurança da informação e medidas técnicas e administrativas. A LGPD exige a adoção de medidas técnicas e administrativas para proteger os dados pessoais contra acessos não autorizados e situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou difusão dos dados (Art. 46 da LGPD).
- Relatórios de impacto à proteção de dados com documentação detalhada, pois, para certas operações de tratamento, a LGPD pode requerer a elaboração de Relatórios de Impacto à Proteção de Dados, que descrevem os processos de tratamento de dados pessoais e as medidas de segurança adotadas.
- Penalidades e sanções por descumprimento, pois a LGPD estabelece sanções rigorosas para o descumprimento das suas disposições, incluindo multas que podem chegar a 2% do faturamento da empresa, limitadas a R$ 50 milhões por infração.

A nova resolução do Conselho Nacional de Saúde, complementada pela LGPD, impõe um conjunto rigoroso de requisitos para a proteção de dados em pesquisas, criando desafios significativos para os pesquisadores em termos de obtenção de consentimento, anonimização de dados, medidas de segurança, e cumprimento de direitos dos titulares. A conformidade com essas normas exige uma abordagem robusta e bem planejada para o tratamento de dados pessoais, incluindo investimento em tecnologia, treinamento e desenvolvimento de políticas internas adequadas.

AKCIT
CENTRO DE COMPETÊNCIA EMBRAPII EM TECNOLOGIAS IMERSIVAS

Gestão de Dados de Saúde

Unidade II
Ciclo de Vida dos Dados de Saúde

Unidade II: Ciclo de Vida dos Dados de Saúde

2.1 Etapas do Ciclo de Vida de um Dado de Saúde: Coleta, Processamento, Armazenamento, Análise e Disseminação

Assim como o idioma faz parte do que a sociedade é por ser uma característica unitiva, os sistemas de informação estão imbricados com o sistema de saúde. Diagnósticos, procedimentos, estabelecimentos, territórios, cifras, trabalhadores, usuários e gestores são elementos mantidos e regulamentados pelo sistema de saúde. Ao lidar com a complexa inter-relação desses elementos, espera-se avaliar partes e totalidades, sujeitos e objetos, multidimensões em multivisões, não apenas interdisciplinarmente, mas superar fronteiras das disciplinas de forma transdisciplinar, como ocorre com a área do conhecimento chamada Informática em Saúde. Assim, surge a figura do informática em saúde (Figura 10).

Figura 10 - O informata em saúde

Fonte: Brasil (2020d).

Sistema de saúde é definido, nos Descritores em Ciências da Saúde (DeCS), como uma

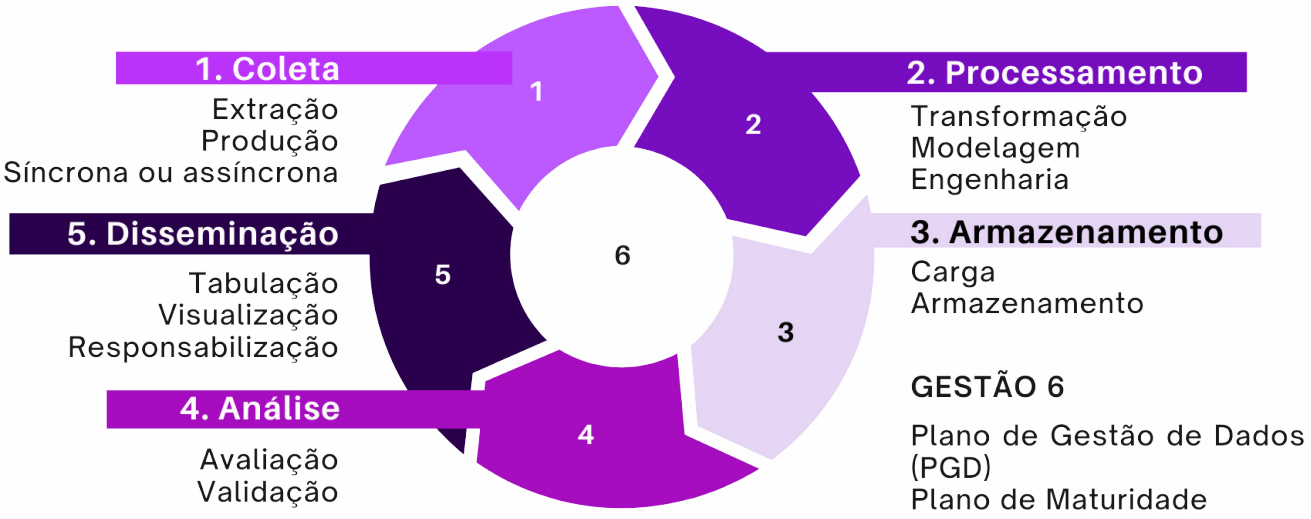
> rede de serviços cujo objetivo é proporcionar um ótimo nível de saúde às pessoas, proteger dos riscos de adoecer, satisfazer as necessidades individuais de saúde e distribuir de forma equitativa o nível de saúde. As funções dos sistemas de saúde compreendem a prestação de serviços, o financiamento, a geração de recursos, a supervisão e a regulação (Brasil, 2022a).

Os **SIS** são parte do processo atrelado aos serviços junto às pessoas e instituições envolvidas - usuários, trabalhadores, gestores e prestadores:

> Compreende-se como SIS a integração e articulação do processo de coleta, armazenamento e processamento dos dados, com o propósito de transformar dados em informação e, consequentemente, apoiar a tomada de decisão nos serviços de saúde, não estando obrigatoriamente atrelados à incorporação das tecnologias da informação, apesar de hoje se apoiarem fortemente na informática (Leandro; Rezende; Pinto, 2020, p. 61).

A coleta de dados em saúde envolve sistemas de informação, mas em muitos casos ainda depende de digitadores que inserem informações manuscritas ou impressas no sistema. A comunicação entre ferramentas pode ser síncrona ou assíncrona, sendo desejável que ocorra em "quase tempo real" por limitações tecnológicas. Após a **coleta**, ocorre o **processamento** dos dados para anotação semântica e codificados conforme padrões institucionais, viabilizando sua análise e geração de relatórios. O **armazenamento** ocorre geralmente em Sistemas Gerenciadores de Banco de Dados (SGBDs). A **análise** gera valor de uso para os dados. A **disseminação** permite o acesso a extratos, promovendo o uso em outros contextos (Figura 11).

Figura 11 - Ciclo de vida do dado em saúde



Fonte: autoria própria.

A **coleta**, formalmente chamada de produção ou extração de dados, ocorre diretamente intermediada por sistema de informação. Entretanto, ainda em 2024 existe a figura do digitador, onde as prescrições, notificações, exames, são entregues manuscritos ou impressos no momento do contato assistencial, sendo digitados no sistema de informação do profissional ou do estabelecimento que deseja dar segmento ao atendimento. A comunicação entre ferramentas pode ser **síncrona**, isto é, simultânea ou em tempo real; ou **assíncrona** ao ser realizada em tempos diferentes entre o emissor e o receptor da informação. A rigor a sincronicidade é desejável, mas diz-se frequentemente que ocorre em *quase tempo real* (*near real time*) por restrições tecnológicas de **latência**, isto é, o tempo do envio e o da recepção do dado.

O **processamento** dos dados ocorre após a coleta de modo a proceder com a anotação semântica, isto é, codificação conforme o padrão adotado na instituição. Deste modo, os dados classificados podem ser transpostos para relatórios ou agregações para fins analíticos. Nas seções seguintes trabalharemos com modelagem de entidades e relacionamentos, exemplificando atividades de processamento.

O **armazenamento** é a etapa onde são persistidos os dados, cuja guarda e indexação viabilizam a recuperação no momento oportuno. Usualmente são utilizados SGBDs, os quais asseguram as regras do negócio logo no registro das informações. Infelizmente ainda é frequente o uso de planilhas de cálculo para a organização de conjuntos de dados. Tal procedimento é contrário às Boas Práticas para o SGBDs, uma vez que não perpassa pela gestão de metadados, não garantido automaticamente o tipo do dado, o domínio do atributo e as regras de negócio. Isso também ocorre com conjuntos de dados mantidos por ferramentas frequentemente usadas para fins matemáticos e estatísticos.

A **análise** requer que dados transacionais sejam processados para se tornarem dados otimizados para a análise. Existem ferramentas de gerenciamento de repositórios ou armazém de dados (DW - *Data Warehouse*), bem como formatos que facilitam o carregamento em *softwares* estatísticos, como o Projeto R (extensão RData).

A **disseminação** consiste na política de permitir o acesso a extratos do banco de dados a indivíduos e instituições que não participam da administração do repositório, o que amplia o uso da informação para outros contextos. No Brasil, existem diversas estratégias de disseminação, desde o acesso direto a microdados, bem como a existência de centenas de tabuladores que resultam em produtos em formato de tabela, onde o usuário seleciona os atributos desejados para linhas, colunas e valores a serem calculados (Brasil, 2009a). A disseminação pode produzir dados de acesso restrito ou dados abertos. Segundo o guia de Elaboração de Plano de Dados Abertos, os **dados abertos** devem ser “completos, primários, atuais, acessíveis, processáveis por máquina, acesso não discriminatório, os formatos não proprietários e livres de licenças” (Brasil, 2017a; Zorzal; Rodrigues, 2016).

Um tipo de disseminação é a de dados abertos. A gestão desses dados é crucial para promover sua reutilização, especialmente no contexto da ciência aberta. Com a produção de conhecimento científico enquanto atividade global, é preconizada a gestão “FAIR” de dados, isto é, “Encontráveis (*Findable*), Acessíveis (*Accessible*), Interoperáveis (*Interoperable*) e Reutilizáveis (*Reusable*)”, especialmente quando as pesquisas são financiadas com recursos públicos (Almeida, 2023).

É importante salientar que o acesso à informação é um direito. A **Lei de Acesso à Informação** (LAI, Lei nº 12.527/2011) regulamenta o direito constitucional de aces-

so às informações públicas, promovendo a transparência e a prestação de contas. A Lei estabelece o princípio da publicidade, onde o acesso à informação é a regra, e o sigilo é a exceção; a máxima divulgação visto que os órgãos públicos devem promover a transparência ativa, divulgando informações de interesse coletivo; transparência passiva, pois é direito de qualquer pessoa solicitar e receber informações dos órgãos públicos quando não praticam transparência ativa; gratuidade ao acesso à informação pública, salvo custo de reprodução de documentos; e proteção de dados pessoais, onde é previsto que a LAI respeita a proteção de dados pessoais, que não podem ser divulgados sem consentimento, salvo em situações de interesse público ou determinadas por lei.

2.2 Coleta de Dados de Saúde: Métodos e Ferramentas

A melhor prática é a transposição automatizada de dados mantidos em SGBD para os formatos otimizados das ferramentas estatísticas. As três etapas anteriores, sobretudo no processamento para fins analíticos, são chamadas de Extração, Transformação e Carga (do inglês, *Extract, Transform, Load*) ou, simplesmente, **ETL**.

Seja para a finalidade de pesquisa, gestão ou cuidado, os dados são consolidados por aplicações transacionais ou por operações de ETL. A transposição pode ser realizada com operações utilizando apenas a linguagem de programação ou com ferramentas de ETL com recursos visuais e de clicar e arrastar.

As operações de ETL podem ser utilizadas tanto para enriquecer um banco de dados de produção, o que é conhecido como integração ou interoperabilidade, quanto com ingestão de dados na formação de repositórios com soluções de DW ou lago de dados, quando contempla dados não estruturados. A ETL pode ocorrer com codificação em linguagens de programação, por exemplo, Apache NiFi, Python/Panda e *Structured Query Language* (SQL); ou com ferramentas WYSIWYG, que permitem aos usuários criar e editar conteúdo visualmente, sem a necessidade de codificação (pronunciada wiz-ee-wig, do inglês, *what you see is what you get*) como *Talend, Microsoft* SQL *Server Integration Services* (SSIS), Informatica *PowerCenter* ou o ecossistema *Pentaho Data Integration* (PDI). O uso de artifícios de programação apresenta maior flexibilidade e capacidade de personalização e versionamento, menores custos, melhor desempenho, e mais força de trabalho disponível. Ferramentas WYSIWYG são mais fáceis para iniciantes e não técnicos, bem como apresentam mais recursos para manter e identificar problemas.

A coleta de dados em formato eletrônico com auxílio de aplicações, isto é, *softwares* com interface de digitação dos dados, ocorre a partir de sistemas de informação desenvolvidos em camadas. A **camada de apresentação** é a de interface com o usu-

ário, a **camada de lógica de negócio (ou camada de aplicação)** mantém as regras de negócio e na **camada de dados** (ou camada de persistência) ocorre a persistência e a gestão com auxílio de SGBD. No vocabulário de tecnologistas da informação, as camadas são chamadas, respectivamente, de *front-end*, camada lógica e *back-end*. A modelagem e desenvolvimento de *software* requer boas práticas com o seguimento de diretrizes arquiteturais preconizadas na saúde e na gestão pública (Brasil, 2017b, 2018b).

Na camada de interação com o usuário, *front-end*, frequentemente ambientada em navegador web ou aplicativo móvel, as informações são processadas e convertidas em formato de página web. As linguagens de programação que se destacam no desenvolvimento web incluem HTML, CSS, PHP, Python, JavaScript, Java, Ruby e Swift.

A **acessibilidade** é uma característica fundamental na camada de interação com o usuário, garantindo que pessoas com diversas habilidades possam utilizar a interface sem barreiras. Isso envolve o uso de práticas como contraste de cores adequado, textos alternativos para imagens, navegação por teclado, e compatibilidade com leitores de tela. A acessibilidade não apenas cumpre requisitos legais, mas também amplia o alcance do produto a uma base de usuários mais diversificada, promovendo uma experiência inclusiva e equitativa.

A **consistência** assegurando que a interface mantenha um padrão visual e funcional em todas as suas partes. Isso significa que botões, ícones, tipografia e fluxos de navegação devem seguir um design coeso, facilitando o reconhecimento e a previsibilidade para o usuário. A consistência ajuda a reduzir a carga cognitiva, permitindo que os usuários aprendam rapidamente a utilizar a interface e sintam-se confortáveis ao navegar por diferentes seções do produto.

A **performance** garante uma experiência de usuário eficiente e agradável. Interfaces que carregam rapidamente e respondem prontamente às ações dos usuários contribuem para uma experiência fluida e sem frustrações. A otimização de recursos, a minimização de tempo de carregamento e a eficiência na renderização dos elementos visuais são aspectos chave que impactam diretamente na satisfação e retenção do usuário.

A **usabilidade** é uma característica central, focada em tornar a interface intuitiva e fácil de usar. Isso envolve a criação de um layout claro, com elementos bem organizados e caminhos de navegação lógicos. A usabilidade também se beneficia de testes frequentes com usuários reais para identificar e corrigir pontos de fricção. Uma interface usável reduz a necessidade de suporte e treinamento, tornando o produto mais acessível para todos os tipos de usuários.

A **segurança** é uma preocupação crescente no design de interfaces de usuário, especialmente em aplicações que lidam com dados sensíveis. Implementar práticas de design seguro, como autenticação robusta, proteção contra ataques de injeção e criptografia de dados, proteção da privacidade e integridade dos dados dos usuários. Interfaces que transmitem confiança com práticas de segurança visíveis e notificações claras sobre privacidade ajudam a construir uma relação de confiança com os usuários.

Seja desenvolvendo um *front-end*, seja utilizando um formulário, caso a coleta seja recorrente, a boa prática é implementar um back-end e realizar a gestão de dados com SGBD. O SGBD pode ser do tipo **relacional**, dedicado a dados estruturados, por exemplo, MySQL, Oracle, Microsoft SQL Server, PostgreSQL e Firebird; ou **Não Relacional** (noSQL) para dados semi-estruturados, como MongoDB, Redis, Cassandra. Em nossos exercícios faremos demonstrações utilizando PostgreSQL, ilustrando a linguagem de consulta estruturada SQL.

Um sistema de informação realiza **operações** conhecidas ludicamente como CRUD, acrônimo das quatro operações básicas de Criar (*Create*), Ler (Read), Atualizar (*Update*) e Remover (*Deletar*). Uma boa prática é organizar as operações segundo a camada de desenvolvimento utilizando as ferramentas e linguagens de programação apropriadas. O conteúdo sobre gestão formal de dados é dado ao longo do curso nas abordagens relacionadas ao SGBD.

A **integração** refere-se ao processo de combinar diferentes sistemas e aplicações de forma que funcionem juntos como uma única unidade coerente. **Interoperabilidade** é a capacidade de sistemas e organizações de diferentes tipos e naturezas trabalharem juntos (interoperar), sem dependência tecnológica, permitindo o uso compartilhado de informações. A integração ocorre, usualmente, com recursos como *Enterprise Service Bus* (ESB), onde sistemas diferentes conectados via barramento de serviço ou Interface de Programação de Aplicação, API (do inglês, *Application Programming Interfaces*), para comunicação entre *softwares*.

Os instrumentos mais comuns para a interoperabilidade em saúde adotam o padrão HL7 FHIR ou Web Services, onde serviços web com padrões como SOAP (*Simple Object Access Protocol*) ou REST (*Representational State Transfer*) para comunicação entre sistemas diferentes. Os formatos mais difundidos são JSON e XML. SOAP está fortemente ligado ao XML como formato de mensagem, enquanto JSON é comumente usado com FHIR no contexto REST para facilitar a interoperabilidade em sistemas de saúde. Exemplos de processamento de dados em JSON serão ilustrados nas atividades práticas.

2.3 Modelagem Relacional em um Cenário de Saúde (Prontuário Eletrônico)

A modelagem relacional em um sistema de prontuário eletrônico organiza dados em tabelas relacionadas, representando entidades como pacientes (Figura 12), atendimentos e exames (Figura 13), profissionais da saúde (Figura 14). Essas entidades são interconectadas por relacionamentos que mostram como os dados estão vinculados.

Figura 12 - Tabela Paciente

id_paciente (primary key)	nome	data_nascimento	endereço
1	Ana	1980-01-01	Rua A, 123
2	João	1975-05-12	Rua B, 456

Fonte: autoria própria.

Figura 13 - Tabela Atendimento

id_atendimento (primary key)	data	id_paciente	id_profissional
1	2024-07-01	1	1
2	2024-07-02	2	2

Fonte: autoria própria.

Figura 14 - Tabela Profissional

id_profissional (Primary Key)	Nome	Data_Nascimento	Endereço
1	Dr. Silva	1968-03-22	Rua C, 789
2	Dr. Pereira	1972-11-10	Rua D, 101

Fonte: autoria própria.

Veja conteúdo adicional no material sobre SQL e bancos de dados no capítulo “Modelagem e gestão de banco de dados com SQL e integração com o R” presente no livro “Avaliação de impacto das políticas de saúde: um guia para o SUS” (Ferré, 2023b).

Considere as entidades e atributos:

- **paciente**: id_paciente, nome, data_nascimento, endereço;
- **profissional**: id_profissional, nome, especialidade;
- **atendimento**: id_atendimento, data, id_paciente, id_profissional; e
- **exame**: id_exame, tipo_exame, resultado, id_atendimento.

Considere os relacionamentos Um-para-Muitos (1:N):

- Um paciente pode ter muitos atendimentos.
- Um profissional pode realizar muitos atendimentos.
- Um atendimento pode ter muitos exames.

Uma chave primária (*Primary Key*) é um campo ou um conjunto de campos em uma tabela de banco de dados relacional que serve para identificar univocamente cada registro desta tabela. As principais características de uma chave primária são:

- **Unicidade:** Cada valor na chave primária deve ser único. Isso significa que não podem existir dois registros na tabela com o mesmo valor de chave primária.
- **Não-nulidade:** Os campos que compõem a chave primária não podem conter valores nulos. Cada registro deve ter um valor válido para a chave primária.
- **Imutabilidade:** Idealmente, o valor de uma chave primária não deve mudar nem ser reaproveitado para outro objeto. Uma chave primária é usada para identificar registros de forma consistente ao longo do tempo.
- **Minimalidade:** A chave primária deve conter apenas os campos necessários para garantir a unicidade. Não deve incluir colunas extras além das que são necessárias para a identificação exclusiva de registros. Usualmente são números inteiros e sequenciais.
- **Relacionamentos:** A chave primária de uma tabela pode ser referenciada como chave estrangeira (*Foreign Key*) em outra tabela, estabelecendo relacionamentos entre as tabelas.
- **Integridade Referencial:** Assegura que cada registro possa ser referenciado de maneira única e consistente, contribuindo para a integridade dos dados no banco de dados.

A organização de bancos de dados requer o redesenho das tabelas, também conhecido como normalização. A **normalização** é um processo utilizado na modelagem de bancos de dados para organizar os dados de maneira eficiente, eliminando redundâncias e inconsistências. Este processo envolve a divisão de uma base de dados em tabelas menores e a definição de relacionamentos entre elas, seguindo uma série de regras ou formas normais (*normal forms*). A normalização visa garantir que cada tabela armazene dados relacionados a uma única entidade ou conceito, reduzindo a duplicidade de dados e facilitando a manutenção e atualização do banco de dados. Ao aplicar a normalização, melhora-se a integridade dos dados, minimiza-se o espaço de armazenamento necessário e otimiza-se o desempenho das consultas, resultando em uma estrutura de banco de dados mais robusta e eficiente.

A **Primeira Forma Normal (1NF)** estabelece que cada tabela em um banco de dados deve ter colunas com valores atômicos, ou seja, indivisíveis, e cada campo deve conter um único valor. Além disso, as entradas em uma coluna devem ser do mesmo tipo de dado, e cada coluna deve conter uma única categoria de dados (Figuras 15-17).

Figura 15 - Tabela Antes (não 1NF)

id_paciente	nome	telefones
1	Ana	99999-1234, 99999-5678
2	João	88888-4321, 88888-8765

Fonte: autoria própria.

Figura 16 - Tabela Depois (1NF): Paciente

id_paciente	nome
1	Ana
2	João

Fonte: autoria própria.

Figura 17 - Tabela Depois (1NF): Telefone

id_paciente	nome	telefone
1	Ana	99999-1234
1	Ana	99999-5678
2	João	88888-4321
2	João	88888-8765

Fonte: autoria própria.

A **Segunda Forma Normal (2NF)** se baseia na 1NF e elimina dependências parciais. Isso significa que os atributos não-chave devem depender da chave primária inteira e não apenas de uma parte dela (isso é especialmente relevante em tabelas com chaves compostas).

Note que, no exemplo abaixo, a data de atendimento é atributo do atendimento e nome do paciente, não devendo permanecer de forma redundante numa tabela de atendimento, uma vez que a chave primária de paciente é suficiente para indicar quem foi atendido (Figuras 18-20).

Figura 18 - Tabela Exemplo: Antes (1NF, mas não 2NF)

id_atendimento	data	id_paciente	nome_paciente
1	2024-07-01	1	Ana
2	2024-07-02	2	João
3	2024-07-03	1	Ana
4	2024-07-04	2	João

Fonte: autoria própria.

Figura 19 - Tabela Depois (2NF): Atendimento

id_atendimento	data	id_paciente
1	2024-07-01	1
2	2024-07-02	2
3	2024-07-03	1
4	2024-07-04	2

Fonte: autoria própria.

Figura 20 - Tabela Depois (2NF): Paciente

ID_Paciente	Nome
1	Ana
2	João

Fonte: autoria própria.

A Terceira Forma Normal (3NF) se baseia na 2NF e elimina dependências transitivas. Isso significa que atributos não-chave não devem depender de outros atributos não-chave. Todos os atributos não-chave devem depender diretamente da chave primária. Nesse momento, são dirimidos relacionamentos muitos para muitos (N:M). Por exemplo, mais de um profissional pode realizar procedimentos no mesmo atendimento, em outras palavras, um paciente pode ser atendido por mais de um profissional. Ou seja, o relacionamento entre paciente e profissional é de muitos para muitos (Figuras 21-24).

Figura 21 - Tabela Atendimento Antes (2NF, mas não 3NF)

id_atendimento	id_profissional	data	id_paciente	nome_paciente	endereco_paciente
1	1	2024-07-01	1	Ana Luz	Rua A, 123
1	2	2024-07-01	1	Ana Luz	Rua A, 123
2	2	2024-07-02	2	João Teixeira	Rua B, 456
3	1	2024-07-03	3	Maria Santos	Rua C, 789
4	6	2024-07-04	4	Carlos Vieira	Rua D, 101

Fonte: autoria própria.

Figura 22 - Tabela depois (3NF)

id_procedimento	id_atendimento	data	id_paciente	id_profissional
1	1	2024-07-01	1	1
2	1	2024-07-01	1	2
3	2	2024-07-02	2	2
4	3	2024-07-03	3	6
5	4	2024-07-04	4	2
6	5	2024-07-05	1	2
7	6	2024-07-06	2	1

Fonte: autoria própria.

Figura 23 - Tabela paciente

id_paciente	nome	endereco
1	Ana Luz	Rua A, 123
2	João Teixeira	Rua B, 456
3	Maria Santos	Rua C, 789
4	Carlos Vieira	Rua D, 101
5	Teresa Silvestre	Rua E, 202
6	Rafael	Rua F, 303

Fonte: autoria própria.

Figura 24 - Tabela Profissional

id_profissional	nome	especialidade
1	Pedro Souza	Neurologia
2	Paulo Santos	Anestesia
3	Ana Silva	Ortopedia
4	João Pereira	Dermatologia
5	Luis Mendonça	Pediatria
6	Carla Cabral	Cardiologia

Fonte: autoria própria.

A 1NF assegura que os valores nas colunas sejam atômicos e não repetidos. A 2NF elimina dependências parciais, garantindo que todos os atributos não-chave dependam da chave primária inteira. A 3NF elimina dependências transitivas, garantindo que todos os atributos não-chave dependam diretamente da chave primária e não de outros atributos não-chave. A normalização 3NF é uma técnica de organização de tabelas em um banco de dados onde todos os atributos são funcionalmente dependentes da chave primária e não existem dependências transitivas, garantindo que os dados sejam únicos e minimize redundâncias.

2.4 Processamento e Armazenamento de Dados: Técnicas e Tecnologias

O processamento de dados estruturados é mais comum com linguagem SQL e o de dados semi-estruturados com linguagem noSQL.

2.4.1 Armazenamento de Dados Estruturados com Linguagem SQL

SQL é uma linguagem de programação padrão utilizada para gerenciar e manipular bancos de dados relacionais. Ela permite realizar operações apelidadas informalmente de CRUD (Create, Read, Update, Delete), que são as quatro funções básicas de armazenamento persistente.

- Create: INSERT INTO - Insere novos registros.
- Read: SELECT - Recupera dados de tabelas.
- Update: UPDATE - Modifica registros existentes.
- Delete: DELETE FROM - Remove registros.

Vamos detalhar cada uma dessas operações:

Create (Criar):

A operação Create é utilizada para inserir novos registros em uma tabela de banco de dados. Em SQL, isso é feito com a instrução INSERT INTO.

```
insert
   into
   pacientes (nome,
   email,
   idade)
values ('João Silva',
'joao.silva@example.com',
30);
```

Essa instrução insere um novo registro na tabela pacientes com os valores fornecidos.

Read (Ler):

A operação Read é utilizada para recuperar dados de uma tabela. Em SQL, isso é feito com a instrução SELECT.

```
select
   nome,
   email,
   idade
from
   pacientes
where
   idade > 25;
```

Essa instrução seleciona e retorna os campos nome, email e idade da tabela pacientes onde a idade é maior que 25.

Update (Atualizar):

A operação Update é utilizada para modificar registros existentes em uma tabela. Em SQL, isso é feito com a instrução UPDATE.

```
update
   pacientes
set
   email = 'joao.silva@provedor.com.br'
where
   nome = 'João Silva';
```

Essa instrução atualiza o campo email do paciente com o nome 'João Silva' para um novo endereço de email.

Delete (Excluir):

A operação Delete é utilizada para remover registros de uma tabela. Em SQL, isso é feito com a instrução DELETE FROM.

```
DELETE FROM pacientes
WHERE nome = 'João Silva';
```

Essa instrução remove todos os registros da tabela pacientes onde o nome é 'João Silva'.

2.4.2 Armazenamento de Dados Não Relacionais com Linguagem noSQL

Linguagens NoSQL são usadas para gerenciar bancos de dados não relacionais projetados para armazenar, recuperar e gerenciar dados sem a necessidade de um esquema tabular tradicional. Elas oferecem flexibilidade para trabalhar com diferentes modelos de dados, como documentos, pares chave-valor, colunas largas e grafos. A título de ilustração, vamos usar a mesma analogia que fizemos acima para SQL com as operações de CRUD. Vamos explorar como essas operações funcionam no contexto de um banco de dados NoSQL. Veja o resumo das operações e os exemplos a seguir.

- Create: insertOne (ou insertMany) - Insere novos documentos.
- Read: find - Recupera documentos de coleções.
- Update: updateOne (ou updateMany) - Modifica documentos existentes.
- Delete: deleteOne (ou deleteMany) - Remove documentos.

Create (Criar):

A operação Create insere novos registros ou documentos em uma coleção ou tabela. A sintaxe pode variar dependendo do tipo de banco de dados NoSQL.

Aqui está um exemplo usando MongoDB, um banco de dados orientado a documentos. Para inserir novos documentos na coleção pacientes no MongoDB, você pode usar o método insertOne para inserir um único documento ou o método insertMany para inserir vários documentos de uma vez. Aqui estão os exemplos de como fazer isso:

Inserindo um único documento com insertOne:

```
#javascript

db.pacientes.insertOne({
  "_id": 205,
  "nome_paciente": "Carlos",
  "idade": 30,
  "sexo": "Masculino",
  "historico_saude": ["Hipertensão"]
});
```

Inserindo múltiplos documentos com insertMany, onde cada documento é representado como um objeto em um arranjo (*array*):

```
#javascript

db.pacientes.insertMany([
  {
    "_id": 206,
```

continua

```
    "nome_paciente": "Luísa",
    "idade": 28,
    "sexo": "Feminino",
    "historico_saude": ["Asma"]
  },
  {
    "_id": 207,
    "nome_paciente": "Pedro",
    "idade": 45,
    "sexo": "Masculino",
    "historico_saude": ["Diabetes", "Hipertensão"]
  },
  {
    "_id": 208,
    "nome_paciente": "Mariana",
    "idade": 35,
    "sexo": "Feminino",
    "historico_saude": ["Enxaqueca"]
  }
]);
```

O Método insertMany: Este método insere vários documentos de uma vez na coleção especificada. Os campos _id são opcionais. Se não forem fornecidos, o MongoDB gerará automaticamente um valor único.

A estrutura dos documentos pode variar conforme as necessidades do seu banco de dados. No exemplo acima, além dos campos _id e nome_paciente, foram adicionados idade, sexo, e historico_saude como campos adicionais.

Esses comandos devem ser executados no terminal (*shell*) do MongoDB ou em um cliente MongoDB que suporte comandos de inserção, como o MongoDB Compass ou uma biblioteca (*driver*) MongoDB em uma linguagem de programação (por exemplo, Mongoose para Node.js).

Read (Ler). A operação Read recupera dados de uma coleção ou tabela. Novamente, a sintaxe varia com o tipo de banco de dados NoSQL. Aqui está um exemplo em MongoDB:

```
#javascript

db.pacientes.find({ idade: { $gt: 25 } });
```

Essa instrução seleciona e retorna todos os documentos da coleção pacientes onde a idade é maior que 25.

Update (Atualizar). A operação *Update* modifica documentos existentes em uma coleção ou tabela. A sintaxe varia conforme o banco de dados NoSQL. Veja um exemplo em MongoDB:

```
#javascript

db.pacientes.updateOne(
  { nome: "João Silva" },
  { $set: { email: "joao.silva@provedor.com.br" } }
);
```

Essa instrução atualiza o campo email do documento onde o nome é 'João Silva' para um novo endereço de email.

Delete (Excluir). A operação Delete remove documentos de uma coleção ou tabela. A sintaxe varia conforme o banco de dados NoSQL. Aqui está um exemplo em MongoDB:

```
#javascript

db.pacientes.deleteOne({ nome: "João Silva" });
```

A instrução remove o documento da coleção pacientes onde o nome é 'João Silva'.

Exemplos em Outros Bancos de Dados NoSQL:

- **Redis (Chave-Valor)**:
  - Create: SET key value
  - Read: GET key
  - Update: SET key value (mesma operação que Create)
  - Delete: DEL key
- **Cassandra (Colunas Largas)**:
  - Create: INSERT INTO table (columns) VALUES (values);
  - Read: SELECT * FROM table WHERE condition;
  - Update: UPDATE table SET column = value WHERE condition;
  - Delete: DELETE FROM table WHERE condition;
- **Neo4j (Grafos)**:
  - Create: CREATE (n:Label {properties});
  - Read: MATCH (n:Label) WHERE condition RETURN n;
  - Update: MATCH (n:Label) WHERE condition SET n.property = value;
  - Delete: MATCH (n:Label) WHERE condition DELETE n;

NoSQL oferece uma variedade de modelos de dados e sintaxes que são mais flexíveis e escaláveis para certos tipos de aplicações comparado aos bancos de dados relacionais tradicionais.

2.4.3 Cruzamento de Dados com Linguagem no SQL

O cruzamento ou junção de dados é realizado com a operação JOIN.

- O INNER JOIN retorna apenas as linhas que têm correspondência em ambas as tabelas envolvidas na junção. Isso significa que ele combina registros da primeira tabela com registros da segunda tabela onde a condição de junção especificada é satisfeita. Se não houver correspondência entre as tabelas, as linhas não serão incluídas no resultado final.
- O LEFT JOIN (ou LEFT OUTER JOIN) retorna todas as linhas da tabela à esquerda e as linhas correspondentes da tabela à direita. Se não houver correspondência na tabela à direita, o resultado ainda incluirá todas as linhas da tabela à esquerda, mas com valores NULL para as colunas da tabela à direita onde não houve correspondência.

- O RIGHT JOIN (ou RIGHT OUTER JOIN) é o oposto do LEFT JOIN. Ele retorna todas as linhas da tabela à direita e as linhas correspondentes da tabela à esquerda. Se não houver correspondência na tabela à esquerda, o resultado ainda incluirá todas as linhas da tabela à direita, mas com valores NULL para as colunas da tabela à esquerda onde não houve correspondência.
- O FULL JOIN (ou FULL OUTER JOIN) combina os resultados do LEFT JOIN e do RIGHT JOIN. Ele retorna todas as linhas quando há uma correspondência em uma das tabelas. Se não houver correspondência em uma das tabelas, os resultados incluirão valores NULL para as colunas da tabela sem correspondência.
- O CROSS JOIN retorna o **produto cartesiano** das duas tabelas, isto é, a combinação de cada linha da primeira tabela com cada linha da segunda tabela. Isso resulta em um conjunto de linhas que é o número total de linhas na primeira tabela multiplicado pelo número total de linhas na segunda tabela. Essa operação geralmente não utiliza uma condição de junção e pode resultar em muitos registros se as tabelas forem grandes.

Os exemplos tornarão mais claras as definições acima. Considere as ilustrações a seguir (Figuras 25-27).

Figura 25 - Tabela profissional

id_profissional	nome_profissional
1	Dr. Silva
2	Dr. Souza
3	Dr. Lima

Fonte: autoria própria.

Figura 26 - Tabela prescrições

id_prescricao	id_profissional	id_paciente	medicamento	data_prescricao
101	1	201	Paracetamol	2023-01-15
102	1	202	Amoxicilina	2023-02-20
103	2	203	Ibuprofeno	2023-03-10

Fonte: autoria própria.

Figura 27 - Tabela pacientes

id_paciente	nome_paciente
201	João
202	Maria
204	Ana

Fonte: autoria própria.

INNER JOIN retorna apenas as linhas que têm correspondência nas duas tabelas (Figura 28):

```
#sql

select
   profissional.nome _ profissional,
   pacientes.nome _ paciente,
   prescricoes.medicamento,
   prescricoes.data _ prescricao
from
   prescricoes
inner join profissional on
   prescricoes.id _ profissional = profissional.id _ profissional
inner join pacientes on
   prescricoes.id _ paciente = pacientes.id _ paciente;
```

Figura 28 - Tabela após *INNER JOIN*

nome_profissional	nome_paciente	medicamento	data_prescricao
Dr. Silva	João	Paracetamol	2023-01-15
Dr. Silva	Maria	Amoxicilina	2023-02-20

Fonte: autoria própria.

LEFT JOIN (ou LEFT OUTER JOIN) retorna todas as linhas da tabela à esquerda (prescricoes), e as correspondências da tabela à direita (pacientes). Se não houver correspondência, os resultados da tabela à direita serão NULL (Figura 29).

```
#sql

select
   profissional.nome _ profissional,
   pacientes.nome _ paciente,
   prescricoes.medicamento,
   prescricoes.data _ prescricao
from
   prescricoes
left join profissional on
   prescricoes.id _ profissional = profissional.id _ profissional
left join pacientes on
   prescricoes.id _ paciente = pacientes.id _ paciente;
```

Figura 29 - Tabela após LEFT JOIN

nome_profissional	nome_paciente	medicamento	data_prescricao
Dr. Silva	João	Paracetamol	2023-01-15
Dr. Silva	Maria	Amoxicilina	2023-02-20
Dr. Souza	NULL	Ibuprofeno	2023-03-10

Fonte: autoria própria.

RIGHT JOIN (ou RIGHT OUTER JOIN) retorna todas as linhas da tabela à direita (pacientes), e as correspondências da tabela à esquerda (prescricoes). Se não houver correspondência, os resultados da tabela à esquerda serão NULL (Figura 30).

```
#sql

select
   profissional.nome _ profissional,
   pacientes.nome _ paciente,
   prescricoes.medicamento,
   prescricoes.data _ prescricao
from
   prescricoes
right join profissional on
   prescricoes.id _ profissional = mediprofissionalcos.id _ profissional
right join pacientes on
   prescricoes.id _ paciente = pacientes.id _ paciente;
```

Figura 30 - Tabela após RIGHT JOIN

nome_profissional	nome_paciente	medicamento	data_prescricao
Dr. Silva	João	Paracetamol	2023-01-15
Dr. Silva	Maria	Amoxicilina	2023-02-20
NULL	Ana	NULL	NULL

Fonte: autoria própria.

FULL JOIN (ou FULL OUTER JOIN) retorna todas as linhas quando há uma correspondência em uma das tabelas. Se não houver correspondência, os resultados das tabelas esquerda ou direita serão NULL (Figura 31).

```
#sql

select
   profissional.nome _ profissional,
   pacientes.nome _ paciente,
   prescricoes.medicamento,
   prescricoes.data _ prescricao
```

continua

```
from
  prescricoes
full outer join profissional on
  prescricoes.id_profissional = profissional.id_profissional
full outer join pacientes on
  prescricoes.id_paciente = pacientes.id_paciente;
```

Figura 31 - Tabela após *FULL JOIN*

nome_profissional	nome_paciente	medicamento	data_prescricao
Dr. Silva	João	Paracetamol	2023-01-15
Dr. Silva	Maria	Amoxicilina	2023-02-20
Dr. Souza	NULL	Ibuprofeno	2023-03-10
NULL	Ana	NULL	NULL

Fonte: autoria própria.

CROSS JOIN retorna o produto cartesiano das duas tabelas, ou seja, combina cada linha da primeira tabela com cada linha da segunda tabela (Figura 32).

```
#sql

select
  profissional.nome_profissional,
  pacientes.nome_paciente
from
  profissional
cross join pacientes;
```

Figura 32 - Tabela após CROSS JOIN

nome_profissional	nome_paciente
Dr. Silva	João
Dr. Silva	Maria
Dr. Silva	Ana
Dr. Souza	João
Dr. Souza	Maria
Dr. Souza	Ana
Dr. Lima	João
Dr. Lima	Maria
Dr. Lima	Ana

Fonte: autoria própria.

2.4.4 Junção de Dados com Linguagem no SQL

Para realizar joins no MongoDB, utilizamos o pipeline de agregação com a etapa $lookup. O MongoDB não suporta diretamente o FULL JOIN. Isso geralmente requer a combinação de LEFT JOIN e RIGHT JOIN com processamento adicional no aplicativo para mesclar os resultados.

Um pipeline de agregação no MongoDB é uma sequência de etapas de processamento de dados que transforma documentos de uma coleção em um formato desejado. Cada etapa no pipeline é uma operação que aplica alguma forma de transformação ou filtro nos dados, como agrupamento, projeção, filtragem, etc. O pipeline de agregação é uma ferramenta poderosa para realizar consultas complexas e análises nos dados.

A etapa $lookup é uma dessas operações de agregação e é usada para realizar junções entre coleções. É equivalente a um JOIN em SQL. A operação $lookup permite combinar dados de diferentes coleções em um único conjunto de resultados com base em um campo comum.

A operação $lookup tem os seguintes parâmetros principais:

- from: O nome da coleção a ser combinada.
- localField: O campo da coleção de entrada que será usado para a junção.

» foreignField: O campo da coleção referenciada que será usado para a junção.

» as: O nome do novo campo que conterá os documentos combinados da coleção

Vamos criar exemplos similares aos que fizemos em SQL, considerando as mesmas tabelas profissional, prescricoes e pacientes.

Coleção profissional:

```
#json

[
  { "_id": 1, "nome_profissional": "Dr. Silva" },
  { "_id": 2, "nome_profissional": "Dr. Souza" },
  { "_id": 3, "nome_profissional": "Dr. Lima" }
]
```

Coleção prescricoes:

```
#json

[
  { "_id": 101, "id_profissional": 1, "id_paciente": 201, "medicamento": "Paracetamol", "data_prescricao": "2023-01-15" },
  { "_id": 102, "id_profissional": 1, "id_paciente": 202, "medicamento": "Amoxicilina", "data_prescricao": "2023-02-20" },
  { "_id": 103, "id_profissional": 2, "id_paciente": 203, "medicamento": "Ibuprofeno", "data_prescricao": "2023-03-10" }
]
```

Coleção pacientes:

```
#json
[
  { "_id": 201, "nome_paciente": "João" },
  { "_id": 202, "nome_paciente": "Maria" },
  { "_id": 204, "nome_paciente": "Ana" }
]
```

INNER JOIN retorna apenas as linhas que têm correspondência nas duas tabelas.

```
#javascript

db.prescricoes.aggregate([
  {
    $lookup: {
      from: "profissional",
      localField: "id_profissional",
      foreignField: "_id",
      as: "profissional_info"
    }
  },
  { $unwind: "$profissional_info" },
  {
    $lookup: {
      from: "pacientes",
      localField: "id_paciente",
      foreignField: "_id",
      as: "paciente_info"
    }
  },
  { $unwind: "$paciente_info" },
  {
    $project: {
      nome_profissional: "$profissional_info.nome_profissional",
      nome_paciente: "$paciente_info.nome_paciente",
      medicamento: 1,
      data_prescricao: 1
    }
  }
])
```

LEFT JOIN (MongoDB's $lookup realiza nativamente o LEFT JOIN) retorna todas as linhas da tabela à esquerda (prescricoes), e as correspondências da tabela à direita (pacientes). Se não houver correspondência, os resultados da tabela à direita serão NULL.

```
#javascript

db.prescricoes.aggregate([
  {
    $lookup: {
      from: "profissional",
      localField: "id_profissional",
      foreignField: "_id",
      as: "profissional_info"
    }
  },
  { $unwind: { path: "$profissional_info",
preserveNullAndEmptyArrays: true } },
  {
    $lookup: {
      from: "pacientes",
      localField: "id_paciente",
      foreignField: "_id",
      as: "paciente_info"
    }
  },
  { $unwind: { path: "$paciente_info",
preserveNullAndEmptyArrays: true } },
  {
    $project: {
      nome_profissional: "$profissional_info.nome_profissional",
      nome_paciente: "$paciente_info.nome_paciente",
```

continua

```
      medicamento: 1,
      data _ prescricao: 1
    }
  }
])
```

O MongoDB não suporta diretamente o RIGHT JOIN. Podemos simular isso invertendo as coleções de left para right. Aqui está um exemplo usando a coleção profissional como ponto de partida.

```
#javascript

db.profissional.aggregate([
  {
    $lookup: {
      from: "prescricoes",
      localField: " _ id",
      foreignField: "id _ profissional",
      as: "prescricao _ info"
    }
  },
  { $unwind: { path: "$prescricao _ info",
preserveNullAndEmptyArrays: true } },
  {
    $lookup: {
      from: "pacientes",
      localField: "prescricao _ info.id _ paciente",
      foreignField: " _ id",
      as: "paciente _ info"
    }
  },
  { $unwind: { path: "$paciente _ info",
preserveNullAndEmptyArrays: true } },
  {
```

continua

```
    $project: {
      nome_profissional: "$nome_profissional",
      nome_paciente: "$paciente_info.nome_paciente",
      medicamento: "$prescricao_info.medicamento",
      data_prescricao: "$prescricao_info.data_prescricao"
    }
  }
])
```

O MongoDB também não suporta diretamente o CROSS JOIN. Isso pode ser simulado usando $lookup em combinação com $unwind e $lookup novamente.

```
#javascript

db.profissional.aggregate([
  { $unwind: "$_id" },
  {
    $lookup: {
      from: "pacientes",
      pipeline: [{ $project: { nome_paciente: 1 } }],
      as: "paciente_info"
    }
  },
  { $unwind: "$paciente_info" },
  {
    $project: {
      nome_profissional: "$nome_profissional",
      nome_paciente: "$paciente_info.nome_paciente"
    }
  }
])
```

2.4.5 Relacionamentos e Integridade Referencial de Dados Estruturados com SQL

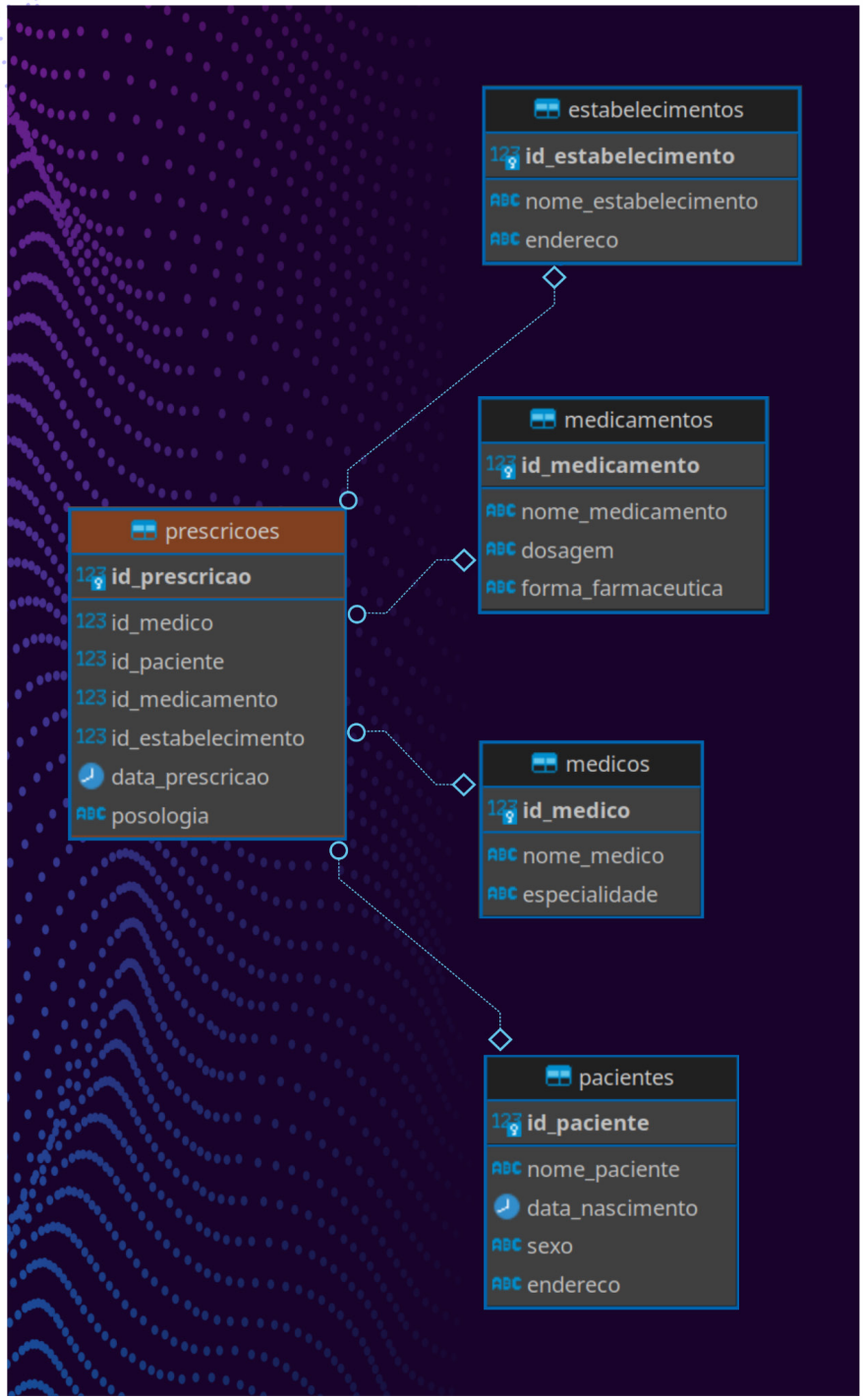
Para criar um modelo relacional em SQL que está na 3NF e garante integridade referencial, vamos considerar as seguintes tabelas: pacientes, prescricoes, profissional, medicamentos, e estabelecimentos.

Lembrando que na 1NF cada campo contém apenas um valor atômico e cada registro é único. Na 2NF a tabela está na 1NF e todos os campos não-chave são totalmente dependentes da chave primária. Na 3NF a tabela está na 2NF e todos os campos não-chave são independentes entre si (ou seja, não existe dependência transitiva).

Cada tabela deve possuir uma **chave primária** (id_paciente, id_profissional, id_medicamento, id_estabelecimento, id_prescricao), garantindo a unicidade dos registros. As **chaves estrangeiras** (id_profissional, id_paciente, id_medicamento, id_estabelecimento) na tabela prescricoes garantem a integridade referencial entre as tabelas. Isso significa que cada valor de chave estrangeira deve corresponder a um valor existente na tabela referenciada. As tabelas medicos, medicamentos, e estabelecimentos são tabelas de domínio que garantem que os dados inseridos na tabela prescricoes estejam limitados a valores válidos e consistentes.

Veja abaixo o modelo relacional e o código-fonte ilustrativo (Figura 33).

Figura 33 - Exemplo de modelo relacional



Fonte: autoria própria.

Tabela pacientes

```
create table pacientes (
    id _ paciente INT primary key,
    nome _ paciente VARCHAR(100) not null,
    data _ nascimento DATE,
    sexo CHAR(1),
    endereco VARCHAR(255)
);
```

Tabela médicos

```
create table medicos (
    id_profissional INT primary key,
    nome_profissional VARCHAR(100) not null,
    especialidade VARCHAR(50)
);
```

Tabela medicamentos

```
create table medicamentos (
    id_medicamento INT primary key,
    nome_medicamento VARCHAR(100) not null,
    dosagem VARCHAR(50),
    forma_farmaceutica VARCHAR(50)
);
```

Tabela estabelecimentos

```
create table estabelecimentos (
    id_estabelecimento INT primary key,
    nome_estabelecimento VARCHAR(100) not null,
    endereco VARCHAR(255)
 );
```

Tabela prescrições

```
create table prescricoes (
    id_prescricao INT primary key,
    id_profissional INT,
    id_paciente INT,
    id_medicamento INT,
    id_estabelecimento INT,
    data_prescricao DATE,
    posologia VARCHAR(255),
    foreign key (id_profissional) references medicos(id_profissional),
    foreign key (id_paciente) references pacientes(id_paciente),
    foreign key (id_medicamento) references
```

continua

```
    medicamentos(id_medicamento),
        foreign key (id_estabelecimento) references
    estabelecimentos(id_estabelecimento)
    );
```

Com esse modelo, garantimos que os dados estão normalizados, minimizando redundâncias e garantindo a integridade dos dados com o uso de chaves estrangeiras e tabelas de domínio.

2.5 Análise e Disseminação de Dados: Ferramentas e Melhores Práticas

A crescente digitalização de documentos clínicos e administrativos promovem a expansão das ferramentas analíticas de apoio à tomada de decisão. As ferramentas podem apoiar diretamente na produção de dados, emitindo alertas inteligentes baseados em regras ou algoritmos, ou apoiando decisões direcionadas por dados (*data driven decision making*).

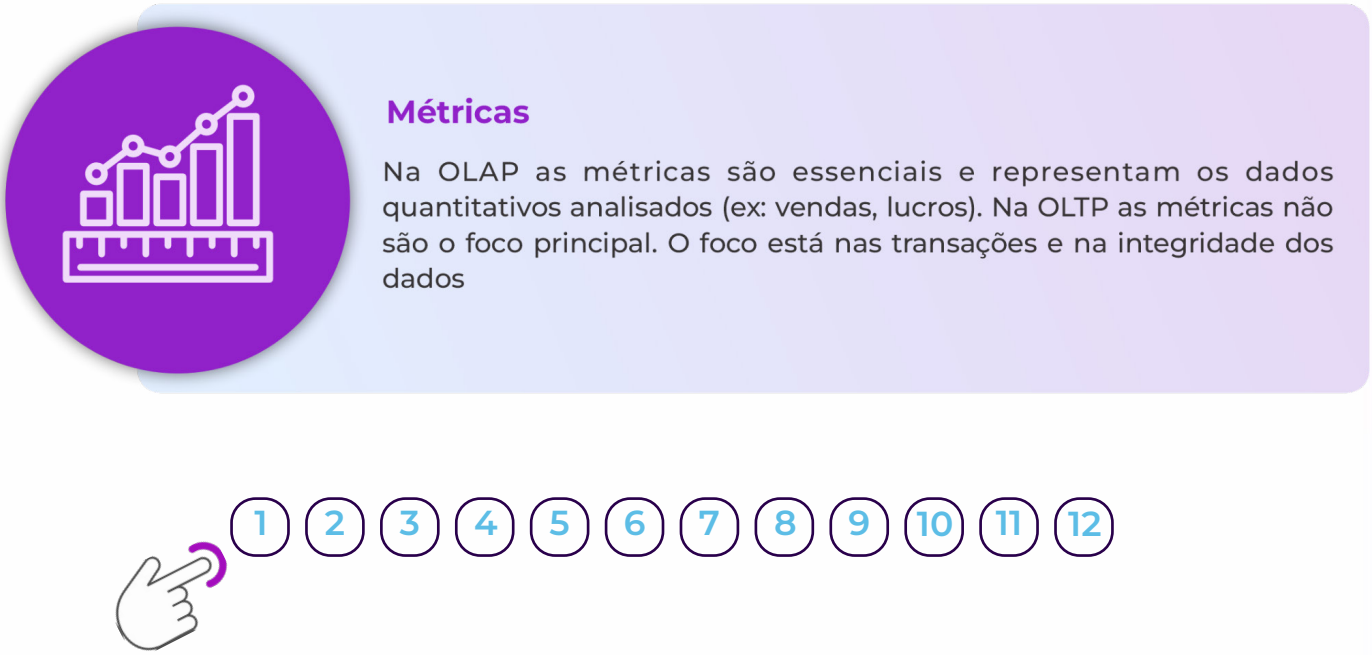
Nas políticas públicas existe o conceito de análise *ex ante* e avaliação *ex post*. A análise *ex ante* produz estimativas e projeções, enquanto a avaliação *ex post* é realizada com base em dados do passado. Entretanto, não adotaremos a distinção e utilizaremos “análise” para ambas as situações, pois estamos trabalhando do ponto de vista das soluções informatizadas.

A análise pode ocorrer em dados granulares ou agregados. O dado granular está no grão mais primitivo do registro. Por exemplo, um extrato do prontuário eletrônico, onde cada linha representa um atendimento e contem {identificador do paciente, sexo, idade na data do atendimento, data de atendimento, procedimento, quantidade do procedimento, diagnóstico primário e valor total do atendimento}. Dados agregados contém métricas que quantificam os atributos em dimensões. No exemplo acima, uma possível agregação é {procedimento, diagnóstico, ano de atendimento, faixa etária [“0 a 60 anos” ou “61 anos ou mais”], soma da quantidade de procedimento, soma do valor de procedimento, contagem de pacientes distintos, contagem de pacientes distintos do sexo feminino}. Agregações serão trabalhadas nas atividades práticas com R e SQL. Por ora, note dois aspectos no exemplo acima: i) nas operações de soma e de contagem de frequência, o número de linhas será igual ou inferior ao número de registros originais no menor grão; ii) a idade foi separada em linhas (dimensões) e o sexo em coluna à parte (métrica).

2.5.1 Modelagem Analítica

A modelagem de Processamento de Transações Online - OLTP (*Online Transaction Processing*) é utilizada para gerenciar e facilitar operações diárias em sistemas transacionais, como prontuários eletrônicos, sistemas de agendamento de consultas ou autorização de procedimentos, controle de estoque e logística, entre outros. Diferentemente da modelagem de Processamento Analítico Online - OLAP (*Online Analytical Processing*), focada na análise de dados históricos e agregados, a modelagem OLTP é otimizada para muitas transações de forma rápida e eficiente. A modelagem OLAP é uma abordagem usada para facilitar a análise rápida e eficiente de grandes volumes de dados em repositórios conhecidos como DW. Os principais conceitos e termos associados a essa modelagem, são (Figura 34):

Figura 34 - Principais conceitos e termos associados a modelagem de processamento de transações online



Fonte: autoria própria.

Modelos *Star Schema* vs. ***Snowflake Schema***. *Star Schema* (Esquema Estrela): Nesse modelo, a tabela fato está no centro e é diretamente conectada às tabelas dimensão. É chamado de estrela devido ao formato que se assemelha a uma estrela. Este modelo é mais simples e fácil de entender. *Snowflake Schema* (Esquema Floco de Neve): Neste modelo, as tabelas dimensões são normalizadas, o que significa que

as dimensões podem ser divididas em tabelas adicionais. Isso resulta em um esquema mais complexo, semelhante a um floco de neve, mas pode economizar espaço e melhorar a integridade dos dados. Na OLTP a normalização é alta (3NF ou superior) para minimizar redundância e maximizar a integridade dos dados. Na OLAP é usado principalmente em *Star Schema* (menos normalizado) ou *Snowflake Schema* (mais normalizado) para otimizar a performance de consultas analíticas.

Para otimizar o modelo relacional para consultas OLAP, é comum desnormalizar as tabelas visando o desempenho de consultas e simplificação da estrutura de dados. A desnormalização envolve a combinação de tabelas para reduzir o número de *joins* necessários durante a execução de consultas, o que é especialmente útil para análise de dados. Veja a tabela desnormalizada única que inclui todas as informações relevantes dos pacientes, prescrições, médicos, medicamentos e estabelecimentos.

```
CREATE TABLE fato_prescricoes (
    id_prescricao INT PRIMARY KEY,
    id_profissional INT,
    nome_profissional VARCHAR(100),
    especialidade_profissional VARCHAR(50),
    id_paciente INT,
    nome_paciente VARCHAR(100),
    data_nascimento_paciente DATE,
    sexo_paciente CHAR(1),
    endereco_paciente VARCHAR(255),
    id_medicamento INT,
    nome_medicamento VARCHAR(100),
    dosagem_medicamento VARCHAR(50),
    forma_farmaceutica_medicamento VARCHAR(50),
    id_estabelecimento INT,
    nome_estabelecimento VARCHAR(100),
    endereco_estabelecimento VARCHAR(255),
    data_prescricao DATE,
    posologia VARCHAR(255)
);
```

No Modelo Desnormalizado ocorre a redução de Joins. Ao combinar todas as informações em uma única tabela, eliminamos a necessidade de realizar múltiplos joins durante a execução de consultas, melhorando significativamente o desempenho das consultas. Observe os campos incluídos:

- id_prescricao: Chave primária da tabela, identificando exclusivamente cada prescrição.
- id_profissional, nome_profissional, especialidade_profissional: Informações do médico que fez a prescrição.
- id_paciente, nome_paciente, data_nascimento_paciente, sexo_paciente, endereco_paciente: Informações do paciente que recebeu a prescrição.
- id_medicamento, nome_medicamento, dosagem_medicamento, forma_farmaceutica_medicamento: Detalhes do medicamento prescrito.
- id_estabelecimento, nome_estabelecimento, endereco_estabelecimento: Informações do estabelecimento onde a prescrição foi realizada.
- data_prescricao, posologia: Detalhes da prescrição, incluindo a data e a posologia recomendada.

Existem vários benefícios na modelagem OLAP. Desempenho de consulta melhorado, visto que as consultas OLAP envolvem geralmente a agregação de grandes volumes de dados. A desnormalização reduz a necessidade de joins complexos, melhorando o desempenho das consultas. Simplicidade das consultas, uma vez que com as informações relevantes em uma única tabela, as consultas se tornam mais simples e fáceis de escrever, facilitando a extração de visões dos dados. Eficiência na Agregação de Dados, pois operações de agregação, como SUM, AVG, COUNT, etc., são mais eficientes em uma tabela desnormalizada, onde todas as informações necessárias estão disponíveis diretamente.

Embora a desnormalização melhore o desempenho de consultas analíticas, ela também aumenta a redundância de dados e pode levar a inconsistências se não for gerenciada adequadamente. Portanto, é importante balancear os benefícios de desempenho com os desafios de manutenção de dados, especialmente em sistemas onde as operações de leitura são muito mais frequentes do que as operações de escrita.

Para realizar consultas em um modelo *Star Schema*, como no caso da tabela desnormalizada fato_prescricoes, podemos usar funções de agregação para resumir dados por diferentes dimensões. Neste caso, vamos usar as informações de endereço para agregar dados por município, Unidade da Federação (UF) e Região do Brasil.

```
select
   SUBSTRING_INDEX(endereco_paciente,
   ',',
   1) as municipio,
   COUNT(id_prescricao) as total_prescricoes
from    fato_prescricoes
group by municipio
order by total_prescricoes desc;
```

Neste exemplo, estamos extraindo o município a partir do campo endereco_paciente e contando o número total de prescrições por município. A função SUBSTRING_INDEX é utilizada para obter o nome do município antes da primeira vírgula no campo endereco_paciente. Veja mais operações de agregação para frequência com COUNT e média com AVG.

```
SELECT
   SUBSTRING_INDEX(endereco_paciente, ',', 1) AS municipio,
   COUNT(id_prescricao) AS total_prescricoes,
   COUNT(DISTINCT id_paciente) AS total_pacientes,
   COUNT(DISTINCT id_profissional) AS total_profissionals,
   AVG(DATEDIFF(CURDATE(), data_nascimento_
paciente)/365) AS idade_media_pacientes
FROM
   fato_prescricoes
GROUP BY
   municipio
ORDER BY
   total_prescricoes DESC;
```

Agora veja o exemplo de Consulta para contagem de prescrições por UF.

```
select
   SUBSTRING_INDEX(SUBSTRING_INDEX(endereco_paciente,
   ',', 2),
   ',', -1) as uf,
```

continua

```
    COUNT(id_prescricao) as total_prescricoes
from    fato_prescricoes
group by uf
order by total_prescricoes desc;
```

```
select
    SUBSTRING_INDEX(SUBSTRING_INDEX(endereco_paciente,
    ',',
    2),
    ',',
    -1) as uf,
    COUNT(id_prescricao) as total_prescricoes,
    COUNT(distinct id_paciente) as total_pacientes,
    COUNT(distinct id_profissional) as total_profissionals,
    AVG(DATEDIFF(CURDATE(), data_nascimento_
paciente)/ 365) as idade_media_pacientes
from    fato_prescricoes
group by uf
order by total_prescricoes desc;
```

Aqui, estamos extraindo a UF a partir do campo endereco_paciente e contando o número de prescrições por UF. A função SUBSTRING_INDEX é usada duas vezes: a primeira para obter os dois primeiros componentes do endereço (até a vírgula que separa a cidade da UF), e a segunda para extrair a UF.

2.5.2 Extraindo Dados Abertos do Sistema Único de Saúde

No SUS existe uma estratégia de disseminação de dados estruturados muito bem sucedida chamada TabNet/TabWin e Sala de apoio à Gestão Estratégica (SAGE) (Brasil, 2008d, 2009b, 2011, 2021c, 2021f; Ministério da Saúde, 2013; Silva, Norberto Peçanha, 2009). Outras tecnologias foram desenvolvidas, porém, ainda não conquistaram o letramento digital para se tornarem hegemônicas por estarem dispersas em diversas em estratégia fragmentária de disseminação, obrigando aos usuários a procurarem informações em lugares dispersos e sem padrão metodológico.

No segundo grupo, existem ferramentas com tecnologias atualizadas, porém com menos conjuntos de dados. Destacam-se no acesso a microdados o opendataSUS

(Brasil, 2019a), no bojo dos dados abertos em dados.gov.br (Brasil, 2012b), e soluções analíticas dispersas em várias plataformas, como o LocalizaSUS (Brasil, 2021d), SISAB (Brasil, 2013) e portal do Fundo Nacional de Saúde (Brasi, 2019b, 2022c) e a PNAD Contínua (Brasil, 2014).

Vamos destacar as principais características da estratégia TabNet/TabWin, por ser a principal do SUS.

Dados granulares coletáveis por humanos e por máquinas. O usuário consegue baixar os microdados, isto é, dados no menor grão, com uma ferramenta cujas interações apontam diretamente para a estrutura de pastas e arquivos do diretório sob a tecnologia de Protocolo de Transferência de Arquivos - FTP (File Transfer Protocol), por exemplo, ftp://ftp.datasus.gov.br/. Note que, apesar de ser um endereço eletrônico, ele não é para utilizar no navegador web, mas num navegador de arquivos (Brasil, 2008e). Veja um exemplo para download de dados hospitalares na Figura 35.

Figura 35 - *Download* manual de dados de Autorização de Internação Hospitalar (AIH) com a ferramenta interativa (acima) ou via navegador de arquivos (abaixo)

Fonte: autoria própria. Fonte dos dados: Brasil (2008e).

Note que a AIH reduzida tem arquivos com prefixo “RD”, sucedidos da sigla do estado PB Paraíba, ano, com dois dígitos, e mês.

Microdados conectados a tabuladores. Tabulação de dados com ferramenta web TabNet (Brasil, 2008b), para tabulações predefinidas, e *standalone*, isto é, com ferramenta para tabulações personalizadas TabWin (Brasil, 2008a) que pode ser baixada e instalada no computador local. Para ilustrar, tente tabular dados hospitalares por município do estado de Alagoas, referentes a abril de 2014 acessando o passo a passo abaixo. O resultado é mostrado na Figura 36 e Tabela 1.

1. https://datasus.saude.gov.br/informacoes-de-saude-tabnet/
2. Assistência à Saúde
3. Produção Hospitalar (SIH/SUS)
4. Dados Consolidados AIH (RD), por local de internação, a partir de 2008.

Microdados conectados a indicadores e painéis com transparência de cálculo e reprodutibilidade. A Ripsa tem como fruto os Indicadores e Dados Básicos - Brasil (Brasil, 2012a). Com as fichas descritas conforme explicado acima, tabuladores que exemplificam a extração e com o acesso direto aos microdados, qualquer um pode reproduzir as informações e ampliar o uso mediante o contexto.

Entretanto, embora ainda seja a estratégia mais difundida para disseminação de dados de saúde no Brasil, foi adotado o formato de arquivo DBC, o qual é um DBF compactado cuja ferramenta de descompactação oficial funciona apenas em ambiente Windows, incomum em servidores cuja distribuição predominante é linux. O DBF é um arquivo de banco de dados dBase, formato introduzido em 1983 e cujo uso foi mitigado a partir de 2000. Atualmente, o formato universal preconizado é CSV, arquivo tabulado com separador de vírgula “,” ou ponto e vírgula “;” e a ferramenta preferível para coleta automatizada é a API. As dificuldades com os formatos obsole-

tos levaram a soluções da comunidade, as quais serão exploradas nas atividades práticas (Ferré *et al.*, 2020; Petruzalek, 2016; Saldanha; Bastos; Barcellos, 2019). A solução **opendatasus** visa corrigir esse problema.

Figura 36 - Tabulador *web* de dados de procedimentos hospitalares do Sistema Único de Saúde, por local de internação

Fonte: Brasil (2018c).

Tabela 1 - Procedimentos hospitalares do Sistema Único de Saúde - por local de internação - Alagoas. Autorização de Internação Hospitalar (AIH) aprovada, valor total (R$), dias de permanência e óbitos, segundo município. Período: abr/2024

Município	AIH aprovadas	Valor total	Dias de permanência	Óbitos
270030 Arapiraca	2.235	3.945.746,36	11.410	89
270040 Atalaia	15	6.682,74	26	1
270070 Batalha	106	41.522,05	166	
270140 Campo Alegre	47	21.035,64	258	3
270210 Colônia Leopoldina	1	451,4		
270230 Coruripe	1.026	1.564.070,89	5.291	24
270400 Junqueiro	35	19.307,50	174	4
270430 Maceió	4.156	6.284.941,39	21.103	110
270630 Palmeira dos Índios	511	572.812,18	1.528	19
270690 Pilar	217	179.430,03	656	3
270730 Porto Calvo	20	6.818,00	60	
270760 Quebrangulo	9	3.106,89	16	1
270800 Santana do Ipanema	697	759.634,60	2.360	23
270840 São Jose da Tapera	23	9.827,17	83	
270860 São Miguel dos Campos	558	526.957,15	1.830	7
270915 Teotônio Vilela	131	43.199,26	630	2
270920 Traipu	5	2.217,00	5	
270940 Viçosa	30	12.560,37	107	3
Total	**9.822**	**14.000.320,62**	**45.703**	**289**

Fonte: Brasil (2018c).

2.6 Saiba Mais - Atividade de Leitura Opcional

2.6.1 Gestão de Dados e Maturidade de Processos Informatizados

Além da gestão da equipe e dos métodos de desenvolvimento e manutenção de soluções, o gestor de dados em saúde é responsável por definir instrumentos como o Plano de Gestão de Dados, do inglês, *Data Management Plan* (DMP) e da avaliação da maturidade documentacional dos processos de uma organização. Tais instrumentos devem ser apoiados e subscritos pela alta direção da organização.

Um **DMP** é um documento que descreve como os dados serão geridos ao longo do ciclo de vida de um projeto, contendo perfis de acesso, rastreabilidade, segurança física e lógica dos dados. Esse plano deve incluir informações sobre como os dados

serão coletados, documentados, armazenados, compartilhados, preservados, processados e analisados. O plano contempla a exclusão de registros, a preservação a longo prazo dos dados relevantes e a disponibilidade para reuso por outros pesquisadores após a conclusão do projeto. Um DMP bem elaborado pode garantir que os resultados da pesquisa fiquem acessíveis e disponíveis, aumentando o valor do trabalho e permitindo que outros pesquisadores o utilizem.

Ao redigir um DMP, é recomendável começar verificando as expectativas dos atores internos e parceiros sobre o que deve ser abrangido, bem como especificações de transparência e prestação de contas. Alguns pontos a serem considerados ao escrever um DMP incluem: o uso de material protegido por direitos autorais ou licenciado para uso, ou distribuição; restrições ao compartilhamento de dados relacionadas a patentes ou licenciamento de tecnologia; publicação científica em periódico que exige a inclusão dos dados subjacentes nos artigos e divulgação de dados abertos ou midiáticos. Além disso, deve-se considerar a possibilidade de restrições sobre os dados, a permissão para reuso, redistribuição ou criação de novas ferramentas, serviços, conjuntos de dados ou produtos, a autorização para uso comercial, sob qual licença os dados estarão disponíveis e se as informações de citação e metadados serão publicamente acessíveis.

O DMP pode incluir também normas e padrões de nomenclaturas, sistemas de gestão de dados homologados, padrões arquiteturais de *software* etc. Um exemplo interessante de gestão de dados é a **Metodologia de Administração de Dados (MAD)** do MS (Brasil, 2017b). A MAD também apresenta um conjunto de regulamentações e descrições que estabelecem padrões e procedimentos a serem seguidos na Administração de Dados, em alinhamento com a Governança de Dados, cujo objetivo é promover uma evolução contínua do modelo, abordando aspectos como o padrão de nomenclatura para objetos de banco de dados e o processo de modelagem de dados para a gestão corporativa e de indicadores da instituição. Além disso, a metodologia inclui diretrizes sobre a privacidade e o tratamento de dados, especialmente no que se refere à pseudonimização e criação de usuários, e apresenta diversos artefatos e guias para apoiar a implementação e documentação do projeto. O padrão de nomenclatura é um exemplo de como fornecedores de *software* podem assegurar a transferência de tecnologia e, consequentemente, a manutenção por equipes diferentes das que idealizaram.

A seguir, apresenta-se alguns modelos com propósito de avaliar processos informatizados de instituições públicas a privadas, inclusive com aplicações específicas para o domínio Saúde:

O **Modelo de Maturidade das Capacidades (CMM)** é um recurso que auxilia as organizações a aprimorar seus processos de desenvolvimento de *software*, identificando áreas que necessitam de melhorias e delineando um caminho para a evolução contínua dos processos. Modelos do tipo trabalham a cultura da qualidade na instituição. Sem especificação não há controle, sem controle não há qualidade, sem qualidade a produção pode parar e as atividades organizacionais podem ser interrompidas.

O *Capability Maturity Model Integration* (CMMI), significa Integração do Modelo de Capacidade e Maturidade, é uma aplicação do CMM e um modelo voltado para a melhoria de processos em empresas, setores, organizações e equipes, fornecendo elementos essenciais para processos eficazes. Este modelo pode ser empregado para orientar a melhoria de processos em projetos, divisões ou em toda a organização, avaliando a maturidade dos processos e oferecendo diretrizes para aprimorá-los e obter produtos de maior qualidade. Além disso, o CMMI serve como um modelo de gerenciamento de riscos, permitindo medir a capacidade da organização em lidar com riscos. Criado pelo SEI (*Software Engineering Institute*) da Universidade Carnegie Mellon, o modelo estabelece cinco níveis de maturidade: Nível 1 – Inicial, Nível 2 – Gerenciado, Nível 3 – Definido, Nível 4 – Gerenciado Quantitativamente e Nível 5 – Otimização. Cada um desses níveis reflete um estágio mais organizado e maduro do processo.

A maturidade de *software* refere-se ao grau de desenvolvimento e refinamento que um *software* ou sistema alcançou ao longo de seu ciclo de vida. Isso envolve processos bem definidos, estáveis e reprodutíveis, com garantia da qualidade das funcionalidades. Um *software* maduro é caracterizado por uma baixa taxa de defeitos, alta confiabilidade, boa performance e uma capacidade de adaptação às mudanças e atualizações. Esse nível de maturidade é geralmente alcançado por meio de práticas de engenharia de *software* sólidas, testes rigorosos, *feedback* contínuo dos usuários e melhoria contínua dos processos de desenvolvimento.

Ferramentas como o CMM, citado anteriormente, são frequentemente utilizadas para avaliar e orientar o progresso na maturidade do *software*, ajudando organizações a alcançar um desenvolvimento mais previsível e controlado.

A maturidade HIMSS (*Healthcare Information and Management Systems Society*) é um modelo que avalia a aplicação de tecnologias de informação em instituições hospitalares, visando aprimorar a qualidade e a eficiência dos serviços de saúde. O modelo, denominado EMRAM (*Electronic Medical Record Adoption Model*), é composto por oito estágios, que vão desde a falta de sistemas de TI (estágio 0) até a plena otimização e a capacidade de operação completamente digitalizada (estágio 7). Os

hospitais *paperless*, ou sem papel, representam o nível mais elevado deste modelo, onde todos os processos clínicos e administrativos estão digitalizados, possibilitando uma integração total e em tempo real das informações de saúde. Isso resulta em uma coordenação de cuidados mais eficaz, com maior precisão e segurança dos dados, além de melhorias significativas na eficiência operacional e na experiência do paciente. Alcançar um alto nível de maturidade HIMSS é um indicativo de excelência na gestão de informações e na prestação de cuidados de saúde apoiados por tecnologia.

O SUS; em ação tripartite do MS, Conselho Nacional de Secretários de Saúde (CONASS) e Conselho Nacional de Secretarias Municipais de Saúde (CONASEMS); estabeleceu o Índice Nacional de Maturidade em Saúde Digital (INMSD) ("MS lança ferramenta que mede o nível de maturidade em saúde digital nas regiões do país", 2024), uma ferramenta destinada a medir a maturidade digital nas regiões do país e a promover a sustentabilidade das ASP. O INMSD avalia estados, municípios e o Distrito Federal. As respostas a um questionário, abrangendo gestão, governança, infraestrutura e segurança, fornecem os dados para calcular o índice em uma escala de 0 a 1. A portaria também institui um Comitê Consultivo para aperfeiçoar e acompanhar a aplicação do índice. O Programa SUS Digital objetiva apoiar a resolutividade das Redes de Atenção à Saúde (RAS) por meio da transformação digital nas macrorregiões de saúde.

O INMSD contempla as dimensões:

- gestão e governança em saúde digital: liderança e articulação, privacidade e confidencialidade, financiamento, política e planejamento;
- formação e desenvolvimento profissional: parceria com instituições de ensino e pesquisa, formação contínua em saúde digital, interdisciplinaridade e abrangência na formação em saúde digital e equipe de TIC e saúde digital;
- sistemas e plataformas de interoperabilidade: registro eletrônico em saúde, sistemas nacionais em saúde, adoção à interoperabilidade, gestão e governança de dados e tecnologias de informação e gestão e governança dos sistemas de informação e bases de dados;
- telessaúde e serviços digitais: gestão de serviços em telessaúde, estratégia de apoio à jornada do paciente, inovação em plataformas para telessaúde, uso de videoconferência síncrona (ao vivo) e monitoramento remoto de pacientes (telemonitoramento);
- infoestrutura: padrões de terminologias clínicas, acesso à informação, ações de comunicação e informação, informação e gestão do conhecimento e combate à desinformação;

» monitoramento, avaliação e disseminação de informações estratégicas: geração e uso de indicadores para avaliação do impacto das tecnologias digitais, disseminação de informações estratégicas e instrumentos de planejamento;

» infraestrutura e segurança: conectividade, segurança da informação, *datacenter* e capacidade de armazenamento em nuvem, estrutura física e capacidade de equipamentos e arquitetura.

2.6.2 Boas Práticas em Gestão de Dados

Na área da saúde é frequente a coleta com ferramentas de visualização direta sem a edição de código-fonte, conhecidas como WYSIWYG, por exemplo, o formulário do Google (*Google Forms*). Ainda assim, é recomendável prosseguir com as análises não utilizando planilhas de cálculo para a transposição de dados, por exemplo, Microsoft Excel ou Planilhas do Google (*Google Sheets*), uma vez que as operações (tabelas dinâmicas e cruzamentos com funções "lookup" ou "proc") não são consistidas, controladas e documentadas, o que dificulta a acurácia e a reprodutibilidade.

A melhor prática ao usar planilha é conectar à ferramenta de análise, tanto para fins matemáticos e estatísticos, como Stata, SPSS/PSPP, MatLab/SciLab, projeto R e Python; quanto para inteligência de negócios, como Apache Superset, Pentaho Community Edition, Metabase Open Source Edition, Google LockerStudio, Microsoft Power BI, QlikView, Tableau, stack ELK (Elasticsearch, Logstash e Kibana) ou SAS Business Analytics.

A separação de ambientes de desenvolvimento é uma boa prática no ciclo de vida do desenvolvimento de *software* e gestão dos dados. Ter ambientes distintos para desenvolvimento, teste, homologação e produção permite um controle mais rigoroso sobre as mudanças no código e garante a qualidade e estabilidade do *software*. Aqui estão algumas boas práticas para a separação de ambientes.

Ambiente de Desenvolvimento: Este ambiente é utilizado pelos desenvolvedores para implementar novas funcionalidades e corrigir bugs. Frequentemente atualizado com commits de novas features e melhorias. Configuração flexível para facilitar o trabalho dos desenvolvedores.

Ambiente de Teste (Testing): Ambiente onde testes automatizados (unitários, integração e regressão) são executados. Utilizado para validar que novas mudanças no código não quebram funcionalidades existentes. Deve espelhar o ambiente de produção o mais próximo possível em termos de configuração.

Ambiente de Homologação (Staging/UAT): Ambiente onde as funcionalidades são verificadas pelos *stakeholders* (clientes, equipe de qualidade). Usado para aceitar ou

rejeitar as novas funcionalidades antes de serem promovidas para produção. Espelha o ambiente de produção e contém dados quase reais ou anonimizados.

Ambiente de Produção (*Production*): Ambiente final onde o *software* é utilizado pelos usuários finais. Deve ser altamente estável e seguro, com controle rigoroso de mudanças. Configuração e dados reais.

Na gestão de ambientes contamos com a poderosa ferramenta Git, um sistema de controle de versão distribuído, amplamente utilizado para gerenciar o código-fonte disponível em plataformas de controle de versão como GitHub, GitLab e Bitbucket. Ele permite que múltiplos desenvolvedores trabalhem simultaneamente em um projeto, rastreando mudanças, colaborando de maneira eficiente e revertendo para versões anteriores do código, se necessário. Ao armazenar o histórico de alterações, o Git facilita a manutenção e a continuidade do desenvolvimento, mesmo em equipes grandes e complexas.

Boas Práticas com Git referem-se a um conjunto de diretrizes e convenções que ajudam a manter a estrutura, a qualidade e a organização do código durante o uso do Git. O Git envolve diversos elementos, citados a seguir. A estrutura de Branches, Branch main (ou master) contém o código de produção estável. Somente código testado e aprovado deve ser subido (*merge*) na branch. O Branch develop contém a última versão do código que está em desenvolvimento e testado pelos desenvolvedores. Base para novas funcionalidades e correções. As Feature Branches são criadas a partir de develop para implementar novas funcionalidades. Nomeadas de forma descritiva, por exemplo, *feature*/login. As Release Branches são criadas a partir de develop quando uma versão está pronta para homologação, assim, permite correções de bugs menores e preparação de release notes sem interromper o desenvolvimento contínuo em develop. Após aprovação, ocorre o *merge* em *main* e *develop*. O "*merge* em *main* e *develop*" refere-se ao processo de integrar as alterações de uma *branch* de desenvolvimento (*develop*) na *branch* principal do projeto (*main*), de modo a atualizar a versão estável com as novas funcionalidades e correções desenvolvidas. As *Hotfix Branches* são criadas a partir de main para corrigir problemas críticos encontrados em produção. Após a correção, a *branch* é "mergeada" de volta em *main* e *develop*. Veja a seguir um resumo das boas práticas na gestão de código-fonte. Conheça as práticas principais para gestão de código-fonte:

- Utilização de Branches: Estruturar o fluxo de trabalho (*workflow*) utilizando diferentes *branches* (ramificações) para desenvolvimento, produção, novas funcionalidades, correções de bugs, etc., como main, develop, feature, release e hotfix.

- Commits Granulares: Fazer *commits* pequenos e significativos, que representem mudanças específicas e bem definidas. Isso facilita a revisão e o entendimento do histórico.
- Mensagens de Commit Claras: Escrever mensagens de commit descritivas e concisas que expliquem o que foi alterado e por quê. Isso ajuda na comunicação com outros desenvolvedores e na documentação do histórico do projeto.
- Revisão de Código: Adotar práticas de revisão de código mediante *pull requests*, garantindo que as alterações sejam discutidas e validadas antes de serem integradas. *Pull Requests* (PRs) são uma das principais ferramentas de colaboração. Um *pull request* é uma solicitação para que as alterações feitas em uma branch (geralmente uma *branch* de *feature* ou *hotfix*) sejam revisadas e integradas (ou "mergeadas") em outra branch, frequentemente a branch principal como a main ou develop. Após finalizar as alterações em uma *branch*, o desenvolvedor cria um *pull request*. Nesse processo, ele descreve geralmente as modificações realizadas e o motivo para a solicitação. Outros desenvolvedores podem visualizar as alterações (*diffs*) propostas no pull request. Eles podem deixar comentários, fazer perguntas, ou sugerir melhorias e revisão do código com a identificação de problemas antes da integração.
- Sincronização Frequente: Realizar *pull* e *push* com frequência para manter o repositório local atualizado e evitar conflitos. No Git, o comando push é utilizado para enviar as alterações locais de um repositório para um repositório remoto. Isso geralmente ocorre após um ou mais *commits* terem sido feitos no repositório local, que incluem alterações a arquivos, novas funcionalidades ou correções de bugs. Sincronização: O *git push* serve para sincronizar as alterações que foram feitas localmente com o repositório remoto. Isso permite que outros colaboradores tenham acesso às atualizações mais recentes.
- Manutenção de um Repositório Limpo: Evitar *branches* desnecessárias e realizar limpezas periódicas no repositório para garantir que a estrutura permaneça organizada.
- Documentação: Manter uma documentação clara sobre o uso do repositório, convenções de branch e fluxos de trabalho, para que todos os membros da equipe estejam alinhados.

Boas práticas científicas implicam na reprodutibilidade por agentes externos à pesquisa. Algoritmos e conjuntos de dados se tornam produtos científicos com status de publicação, sendo desejável a publicação em repositórios de acesso livre, por exemplo:

- Dryad Digital Repository (Andersen, 2021),
- figshare (Singh, 2011),
- Harvard Dataverse Network ("Harvard dataverse", [s.d.]),

- Kaggle (“Find Open Datasets and machine learning Projects”, [s.d.]),
- Network Data Exchange (NDEx)(“NDEx WebApp”, [s.d.]),
- Open Science Framework (OSF, [s.d.]),
- Swedish National Data Service (University Of Gothenburg, 2020) e
- Zenodo (Sicilia; García-Barriocanal; Sánchez-Alonso, 2017).

Devemos também efetuar boas práticas de disseminação de dados. A disseminação de dados deve ser amigável ao usuário comum (*user friendly*) e amigável à automação (*computer friendly*) (Riede *et al.*, 2010). O formato de dados deve ser projetado para ser amigável ao usuário, facilitando a entrada e a compreensão dos dados sem a necessidade de conhecimento técnico avançado. Ao mesmo tempo, deve ser amigável para a automação, permitindo que as aplicações processem, analisem e visualizem os dados de forma eficiente. Isso significa que os dados devem ser organizados de maneira lógica, com metadados claros e bem definidos que possibilitem a automação e integração com outras ferramentas e sistemas de *software*. Dessa forma, tanto o usuário comum quanto os sistemas automatizados passam interagir.

A leitura deve ser fácil e documentada. Os dados devem ser disponibilizados de maneira que o usuário comum consiga ler e entender facilmente e, ao mesmo tempo, ser processáveis por ferramentas amplamente disponíveis e sem barreiras de letramento digital. Deve ser adotado um formato de dados à prova de falhas e robusto contra interpretações erradas por um analisador. Como se espera que as especificações do formato evoluam ao longo do tempo, a compatibilidade retroativa deve sempre ser mantida. A documentação deve ser explícita e deve incluir seções que permitam ao usuário compreender a origem e estar apto a reproduzir os dados em outros contextos, de forma que nenhuma outra fonte seja necessária para entender a origem dos dados. Esse padrão implica que os arquivos de dados sejam pesquisáveis e conjuntos de dados individuais possam ser rastreados por consultas baseadas em palavras-chave ou semântica. Recomenda-se o uso de ferramentas de enciclopédia eletrônica, gestão do conhecimento e marcações de texto, como *wiki*, *git*, *Latex* e *markdown*.

A disseminação deve ser flexível, mas estruturada. O formato de dados deve ser flexível o suficiente para permitir que o usuário organize e classifique os dados de uma maneira intuitiva e conveniente, sem comprometer a estrutura e a legibilidade. A estrutura deve permitir que os arquivos de dados ainda possam ser processados com pacotes de *software* comuns de análise e visualização, facilitando o processamento automatizado de dados de diferentes fontes e séries de medições. Isso impli-

ca que as especificações de formato e sintaxe sejam amplamente desacopladas, permitindo que dados anotados possam perpassar por vocabulários e idiomas distintos.

A disseminação de dados abertos deve ser indexável e acessível a mecanismos de busca. Os conjuntos de dados devem ser organizados numa coleção e recuperados a partir de consultas simples. Além disso, a coleção deve ser catalogada não apenas de acordo com itens bibliográficos ou palavras-chave, mas também com quantidades físicas de volumetria, ofertando informações que subsidiem o dimensionamento da capacidade computacional necessária ao processamento.

2.6.3 Rede Interagencial de Informações para a Saúde (Ripsa)

A curadoria de dados e metadados do SUS apresenta na Ripsa boas práticas que vale a pena conhecer melhor. A Ripsa é uma iniciativa colaborativa e integrada, instituída em 1996 pelo MS em parceria com a Organização Pan-Americana da Saúde, com o objetivo de gerar, analisar e disseminar informações relevantes para as intervenções em saúde pública no Brasil.

Composta por diversas instituições governamentais e não governamentais, a Ripsa trabalha de forma consensual na produção de dados e indicadores sobre as condições de saúde e seus determinantes, visando subsidiar políticas e ações públicas. Atualmente, conta com 44 instituições parceiras que contribuem com sua expertise para a construção coletiva de conhecimento. Promovendo a sinergia entre diferentes atores, a Ripsa se destaca como referência nacional na produção de indicadores de saúde e promove aprimoramento na gestão da informação em saúde, apoiando os processos decisórios do SUS (RIPSA).

Conheça as publicações e navegue pelo livro verde - Indicadores básicos para a saúde no Brasil: conceitos e aplicações (2ª edição - 2008) (OPAS, 2008).

As agregações mais recorrentes na saúde são usadas para gerar os parâmetros de indicadores. No Brasil, foi estabelecida a Rede Interagencial de Informação para a Saúde (Ripsa), a qual congrega dezenas instituições acadêmicas e estatais para estabelecer métodos, conjuntos de dados de indicadores básicos de saúde (Brasil, 2022d). A Ripsa mantém centenas de indicadores categorizados a seguir (OPAS, 2008).

Determinantes de saúde:

A) Demográficos, os quais apresentam fatores relacionados à dinâmica populacional no território;

B) Socioeconômicos, os quais trabalham com o perfil econômico e social da população residente;

C) Situação de saúde (vigilância epidemiológica):

D) Mortalidade, onde são trabalhadas as causas de óbito quanto à ocorrência e distribuição;

E) Morbidade, denotam a ocorrência e distribuição de doenças e agravos à saúde;

F) Estrutura, financiamento e atenção à saúde (conformação da Rede de Atenção à Saúde):

G) Recursos, relativos à força de trabalho, instalações de saúde, habilitações e investimentos.

H) Cobertura, onde é avaliado o acesso e a utilização dos serviços públicos e privados.

Na metodologia Ripsa, cada indicador deve ser documentado, compondo uma ficha contendo as informações (Brasil, 2023a; CONASS, 2023):

» Denominação e código - Nome do indicador (título) e código Ripsa. Por exemplo, “Mortalidade por causas externas (C.9.1)”;

» Conceituação - definições e descrição. Por exemplo, “Número de óbitos por causas externas (acidentes e violência), por 100 mil habitantes, na população residente em determinado espaço geográfico, no ano considerado”;

» Interpretação - como o indicador deve ser lido situacionalmente. Por exemplo, “Reflete aspectos culturais e de desenvolvimento socioeconômico, com o concurso de fatores de risco específicos para cada tipo de acidente ou violência”;

» Usos - Informações úteis para análise e adoção na instituição. Por exemplo, “Contribuir na avaliação dos níveis de saúde e de desenvolvimento socioeconômico da população”;

» Limitações - fatores externos ou subjetivos que podem afetar os resultados. Por exemplo, “Apresenta restrição de uso sempre que ocorra elevada proporção de óbitos sem assistência médica ou por causas mal definidas”;

» Fontes - origem e instituição responsável pela produção dos dados. Por exemplo, “Ministério da Saúde. Secretaria de Vigilância à Saúde (SVS): SIM e base demográfica do IBGE”;

» Método de cálculo - fórmula e parâmetros, usualmente numeradores e denominadores. Por exemplo, [Óbitos por causas externas segundo grupos de

CID-10]÷[população residente]×100.000;

» Categorias sugeridas para análise - granularidade, métricas e dimensões. Por exemplo: “Grupos de causas (CID-10) - Acidentes de transporte (V01-V99), Suicídios X60-X84, Homicídios, incluídas as intervenções legais (X85-Y09) e (Y35-Y36), Causas de intenção indeterminada (Y10-Y34) e Demais causas externas (demais V01-Y98)”;

» Dados estatísticos e comentários - exemplos, tabelas e valores. Por exemplo, mortalidade por causas externas em 2020: 146.038.

2.6.4 Criando uma Função SQL para Extração de Dados

Para este exemplo, vamos assumir que existe uma função chamada get_regiao que, dado o Estado (UF), retorna a Região do Brasil a que pertence (Norte, Nordeste, Centro-Oeste, Sudeste, Sul).

```
select
   get_regiao(SUBSTRING_INDEX(SUBSTRING_INDEX(endereco_paciente,
   ',',
   2),
   ',',
   -1)) as regiao,
   COUNT(id_prescricao) as total_prescricoes,
   COUNT(distinct id_paciente) as total_pacientes,
   COUNT(distinct id_profissional) as total_profissionals,
   AVG(DATEDIFF(CURDATE(), data_nascimento_
paciente)/ 365) as idade_media_pacientes
from
   fato_prescricoes
group by
   regiao
order by
   total_prescricoes desc;
```

Agora veja outro exemplo de consulta com a contagem de prescrições por região do Brasil. Para agrupar por região do Brasil, podemos assumir que a UF está mapeada para uma região específica. Por exemplo:

» Norte: AM, PA, AC, RR, RO, AP, TO

» Nordeste: MA, PI, CE, RN, PB, PE, AL, SE, BA

» Centro-Oeste: MT, MS, GO, DF

» Sudeste: SP, RJ, ES, MG

» Sul: PR, SC, RS

Uma maneira simplificada de fazer isso é utilizando CASE para mapear as UFs para suas respectivas regiões.

```
SELECT
    CASE
        WHEN uf IN ('AM', 'PA', 'AC', 'RR', 'RO', 'AP', 'TO') THEN 'Norte'
        WHEN uf IN ('MA', 'PI', 'CE', 'RN', 'PB', 'PE',
                    'AL', 'SE', 'BA') THEN 'Nordeste'
        WHEN uf IN ('MT', 'MS', 'GO', 'DF') THEN 'Centro-Oeste'
        WHEN uf IN ('SP', 'RJ', 'ES', 'MG') THEN 'Sudeste'
        WHEN uf IN ('PR', 'SC', 'RS') THEN 'Sul'
    END AS regiao,
    COUNT(id_prescricao) AS total_prescricoes
FROM
    fato_prescricoes
GROUP BY
    regiao
ORDER BY
    total_prescricoes DESC;
```

Neste exemplo, usamos uma CASE statement para determinar a região com base na UF extraída do endereço do paciente. Em seguida, contamos o número total de prescrições por região.

Note que na agregação por Município, é utilizada a função SUBSTRING_INDEX para extrair o município do campo endereco_paciente. O *script*, ainda, conta o total de prescrições, total de pacientes únicos, total de médicos únicos e calcula a idade média dos pacientes. Na agregação por UF (Estado), ocorre a extração do estado (UF) a partir do campo endereco_paciente usando a função SUBSTRING_INDEX. Também

é contado o total de prescrições, total de pacientes únicos, total de médicos únicos e calcula a idade média dos pacientes. Na agregação por Região do Brasil é criada a função get_regiao para determinar a região do Brasil a partir do Estado (UF) extraído do campo endereco_paciente. Também conta o total de prescrições, total de pacientes únicos, total de médicos únicos e calcula a idade média dos pacientes.

Observe as funções de agregação COUNT, DISTINCT, e AVG usadas para obter estatísticas agregadas. Observe também as funções de ordenação ORDER BY, onde as consultas são ordenadas pelo total de prescrições em ordem decrescente para destacar os locais com maior número de prescrições.

Para criar a função get_regiao no MySQL, você pode usar a linguagem SQL para definir uma função armazenada que mapeia os Estados (UF) para suas respectivas regiões no Brasil. As cláusulas SQL são:

- DELIMITER: No MySQL, DELIMITER é usado para mudar o delimitador padrão (;) para outro, como //, para definir a função corretamente sem conflito com os pontos e vírgulas internos.
- CREATE FUNCTION: Declara a criação de uma nova função chamada get_regiao que aceita um parâmetro uf do tipo CHAR(2).
- RETURNS VARCHAR(20): Define o tipo de dado de retorno da função como VARCHAR(20).
- BEGIN ... END: Delimita o corpo da função.
- DECLARE regiao VARCHAR(20): Declara uma variável local regiao do tipo VARCHAR(20) que será usada para armazenar a região correspondente ao estado.
- CASE: Um bloco CASE é usado para definir a lógica de mapeamento dos estados para suas respectivas regiões. Cada estado é comparado e, se houver correspondência, a variável regiao é definida com o nome da região.
- RETURN regiao: A função retorna o valor da variável regiao.

```
DELIMITER //
CREATE FUNCTION get _ regiao(uf CHAR(2))
RETURNS VARCHAR(20)
DETERMINISTIC
BEGIN
   DECLARE regiao VARCHAR(20);
```

```
    CASE uf
        WHEN 'AC' THEN SET regiao = 'Norte';
        WHEN 'AP' THEN SET regiao = 'Norte';
        WHEN 'AM' THEN SET regiao = 'Norte';
        WHEN 'PA' THEN SET regiao = 'Norte';
        WHEN 'RO' THEN SET regiao = 'Norte';
        WHEN 'RR' THEN SET regiao = 'Norte';
        WHEN 'TO' THEN SET regiao = 'Norte';
        WHEN 'AL' THEN SET regiao = 'Nordeste';
        WHEN 'BA' THEN SET regiao = 'Nordeste';
        WHEN 'CE' THEN SET regiao = 'Nordeste';
        WHEN 'MA' THEN SET regiao = 'Nordeste';
        WHEN 'PB' THEN SET regiao = 'Nordeste';
        WHEN 'PE' THEN SET regiao = 'Nordeste';
        WHEN 'PI' THEN SET regiao = 'Nordeste';
        WHEN 'RN' THEN SET regiao = 'Nordeste';
        WHEN 'SE' THEN SET regiao = 'Nordeste';
        WHEN 'DF' THEN SET regiao = 'Centro-Oeste';
        WHEN 'GO' THEN SET regiao = 'Centro-Oeste';
        WHEN 'MT' THEN SET regiao = 'Centro-Oeste';
        WHEN 'MS' THEN SET regiao = 'Centro-Oeste';
        WHEN 'ES' THEN SET regiao = 'Sudeste';
        WHEN 'MG' THEN SET regiao = 'Sudeste';
        WHEN 'RJ' THEN SET regiao = 'Sudeste';
        WHEN 'SP' THEN SET regiao = 'Sudeste';
        WHEN 'PR' THEN SET regiao = 'Sul';
        WHEN 'RS' THEN SET regiao = 'Sul';
        WHEN 'SC' THEN SET regiao = 'Sul';
        ELSE SET regiao = 'Desconhecido';
    END CASE;

    RETURN regiao;
END //
DELIMITER ;
```

2.6.5 Linguagem SQL

Para quem deseja aprofundar em linguagens de consultas a bancos de dados e de programação existem os seguintes livros:

- ELMASRI, R.; NAVATHE, S.B.. Sistemas de banco de dados. 6. ed. [S.l.]: Pearson Addison Wesley, 2011 (Elmasri; Navathe, 2011).
- ALCOFORADO, L. F.. Utilizando A Linguagem R: Conceitos, manipulação, visualização, modelagem e elaboração de relatórios. [S.l.]: Alta Books, 2021 (Alcoforado, 2021).
- Avaliação de impacto das políticas de saúde: um guia para o SUS. [S.l.]: Ministério da Saúde, 2023b (Bonat, 2023; Ferré, 2023b; Saldanha; Pedroso; Magalhães, 2023).

CENTRO DE COMPETÊNCIA EMBRAPII EM TECNOLOGIAS IMERSIVAS

Gestão de Dados de Saúde

Unidade III

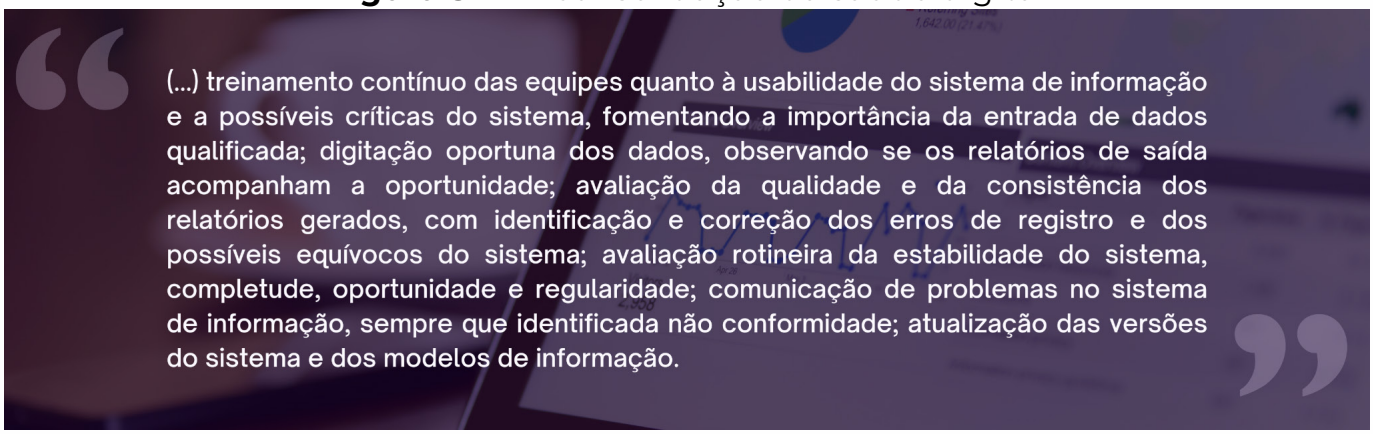
Qualidade de Dados em Saúde

Unidade III: Qualidade de Dados em Saúde

3.1 Importância da Qualidade e Integridade de Dados de Saúde: Segurança do Paciente, Eficiência da Assistência e Pesquisa Clínica

O ciclo de vida do dado em saúde espelha o ciclo da gestão e do cuidado. Assim como a RAS vem se estruturando ao longo das últimas décadas, estamos atravessando uma transformação digital que implica em governança de dados em rede. A consolidação da saúde digital acrescenta novos parâmetros aos cuidados rotineiros ilustrados com o excerto a seguir do Guia de Vigilância em Saúde (2022), referente ao monitoramento de doses de vacinas aplicadas (Figura 37).

Figura 37 - A consolidação da saúde digital



Fonte: Brasil (2022b).

No caso de uso acima, podemos observar diversos processos que podem impactar diretamente na qualidade do serviço prestado e, portanto, na saúde de indivíduos e populações.

A qualidade e integridade dos dados de saúde são fundamentais para o cuidado. Na dimensão da segurança do paciente, a arte do cuidar implica em causar o maior ganho, reduzindo, a um mínimo aceitável, o risco de dano evitável, associado ao cuidado de saúde. **Dados de alta qualidade garantem a promoção, proteção e recuperação da saúde a partir de diagnósticos e intervenções mais precisas, evitando tratamentos inadequados ou desnecessários, prevenindo erros (iatrogenia).** No

contexto da Continuidade do Cuidado, informações completas e precisas permitem uma melhor coordenação entre diferentes profissionais e instituições de saúde, reduzindo o risco de erros de comunicação.

A cultura da qualidade do dado faz parte da cultura da segurança do paciente. A identificação e análise de incidentes nasce de dados completos e precisos, com transparência e aprendizado institucional. A cultura da qualidade incentiva o registro sem implicar em culpabilização individual, mas em aprimoramento de processos de trabalho. Registros de incidentes bem documentados promovem a transparência e o aprendizado organizacional, permitindo a implementação de medidas corretivas eficazes.

Os eventos adversos resultam de efeitos inesperados e negativos que ocorrem durante o tratamento. Esses eventos são identificados tanto clínica quanto epidemiologicamente e são divulgados em compêndios e nas bulas durante o processo regulatório junto às instituições de saúde. Os tratamentos são registrados e autorizados para uso populacional apenas quando considerados suficientemente seguros, com os possíveis malefícios sendo controlados nas circunstâncias clínicas de aplicação. Tratamentos que poderiam ser benéficos em um determinado caso clínico podem ser evitados se houver contraindicações relacionadas a potenciais eventos de segurança. A detecção de eventos adversos começa no desenvolvimento dos fármacos, especialmente nas fases iniciais de testes em indivíduos saudáveis, e continua no monitoramento pós-mercado. Um desafio para a adoção clínica de novas tecnologias de saúde, que são mais eficazes e capazes de erradicar doenças específicas, é a falha no entendimento sistêmico da dinâmica celular, resultando em gastos exorbitantes na busca por alvos terapêuticos seguros. Além da indústria e da academia, grupos de proteção ao consumidor e usuários de medicamentos, juntamente com órgãos competentes de gestão da saúde, estão fortemente interessados em identificar reações adversas a fármacos (Page *et al*., 2012).

Desde o desenvolvimento até a utilização, a segurança no uso de medicamentos depende da disponibilidade das melhores evidências, tanto para usuários quanto para profissionais de saúde. Com a expansão do arsenal terapêutico, torna-se essencial aplicar técnicas que hierarquizem o grau de evidência, orientando as decisões de saúde com base em informações fundamentadas (Carvalho; Moreira; Magalhães, 2013), usualmente realizada por centros colaboradores como o Cochrane (Cochrane, 2022) ou o Oxford (CEBM, 2020) ou por Núcleos de Avaliação de Tecnologias de Saúde (NATS) sob a coordenação da Comissão Nacional de Incorporação de Tecnologias no SUS (Conitec), no Brasil (Brasil, 2010a).

Além das informações obtidas por meio de experimentos *in vitro*, em animais (*in vivo*) ou em populações (*in populo*), há um crescente interesse na simulação de ambientes biológicos in silico, utilizando modelos computacionais de aprendizado de máquina. Dessa maneira, surge uma nova possibilidade que os profissionais de saúde precisam entender como uma ferramenta, similar ao uso de estetoscópios, tomógrafos ou exames sanguíneos.

A aplicação da IA vem crescendo na detecção de eventos adversos, sendo relatada em diversas revisões de modelos para segurança do paciente (Rácz *et al.*, 2021; Syrowatka *et al.*, 2022; Wu *et al.*, 2022; Zhao *et al.*, 2022), bem como em aplicações mais específicas de farmacovigilância (Ball; Dal Pan, 2022; Edrees *et al.*, 2022).

A estratégia para implantar a segurança do paciente deve englobar a criação e o suporte à implementação informatizada de protocolos, guias e manuais de prática clínica e segurança do paciente, com sistemas de apoio à tomada de decisão. Além disso, é essencial promover processos de formação contínua e permanente com foco em letramento digital, estabelecer metas e indicadores de segurança nos processos e serviços, realizar campanhas de comunicação sobre segurança do paciente, incluindo redes sociais e mecanismos para combater a desinformação. Também é importante a vigilância e o monitoramento informatizado de incidentes, bem como a promoção de uma cultura de cibersegurança.

Na cultura da qualidade da assistência e segurança do paciente está incluída a cultura de segurança de dados (Natividade *et al.*, 2023). Existem normas e recomendações internacionais, como a ISO/IEC 27002 e a ABNT NBR ISO/IEC-17799 com códigos de prática para controles de segurança da informação, onde são elencadas medidas e políticas para organização da segurança da informação; segurança dos recursos humanos; gestão de ativos; controle de acesso; criptografia; segurança física e ambiental; operações de segurança; segurança das comunicações; aquisição, desenvolvimento e manutenção de sistemas; relacionamentos com fornecedores; gerenciamento de incidentes de segurança da informação; aspectos de segurança da informação da gestão de continuidade de negócios e práticas de conformidade (ABNT, 2001).

3.1.1 Aspectos Regulatórios

Existem aspectos regulatórios e legais, como a *Health Insurance Portability and Accountability Act* (HIPAA) nos EUA ou a LGPD no Brasil. A LGPD (Lei nº 13.709/2018) estabelece normas para a coleta, processamento, armazenamento e compartilhamento de dados pessoais. Os pontos-chave incluem (Figura 38):

Figura 38 - Pontos-chaves para aspectos regulatórios

Fonte: autoria própria.

Quanto aos atores, podemos nos separar em pessoas físicas e instituições. No primeiro grupo temos:

» Titular: Pessoa natural a quem se referem os dados pessoais e possui o direito de consentir com o uso de seus dados e solicitar informações sobre o tratamento desses dados.

- Controlador: Pessoa natural ou jurídica, pública ou privada, responsável por tomar decisões sobre o tratamento dos dados pessoais. Deve garantir a conformidade com as normas de proteção de dados e responder às solicitações dos titulares.
- Operador: Pessoa natural ou jurídica, pública ou privada, que realiza o tratamento de dados pessoais em nome do controlador. Responsável por seguir as instruções do controlador e garantir a segurança dos dados.
- Encarregado: Pessoa indicada pelo controlador e operador para atuar como canal de comunicação entre o controlador, os titulares dos dados e a Autoridade Nacional de Proteção de Dados (ANPD). Facilita a comunicação e garante a conformidade com a legislação de proteção de dados.
- Pesquisador Responsável: Pessoa que conduz a pesquisa e utiliza os dados pessoais. Compromete-se a manter a confidencialidade e a segurança dos dados e a usá-los apenas para a finalidade declarada.

Quanto aos atores institucionais e suas responsabilidades, elencamos:

- Órgão de Pesquisa: Entidade pública ou privada sem fins lucrativos que realiza pesquisas. Deve garantir que a pesquisa esteja em conformidade com as normas éticas e de proteção de dados.
- Controlador Conjunto: Dois ou mais controladores que determinam conjuntamente as finalidades e os elementos essenciais do tratamento de dados pessoais. Devem estabelecer um acordo para definir as responsabilidades de cada parte.
- ANPD: Órgão da administração pública responsável por zelar, implementar e fiscalizar o cumprimento da LGPD. Deve-se assegurar que as práticas de tratamento de dados estejam em conformidade com a lei.

Dentre outras definições, para fins de proteção de dados pessoais e éticos estipulados na Resolução CNS nº 738/2024 a qual dispõe sobre uso de bancos de dados com finalidade de pesquisa científica envolvendo seres humanos, são considerados:

- Banco de Dados: Conjunto estruturado de dados pessoais estabelecido em suporte eletrônico ou físico. Deve ser gerido de forma a garantir a segurança e a confidencialidade dos dados armazenados.
- Termo de Acordo Institucional: Documento formal pelo qual as instituições se comprometem com a operacionalização, compartilhamento e uso dos dados. Estabelece critérios para a partilha e destinação dos dados em caso de dissolução do acordo.

- Termo de Anuência Institucional: Documento de anuência à realização da pesquisa na instituição, descrevendo as atividades a serem desenvolvidas. Deve ser emitido pelo dirigente institucional ou pessoa por ele delegada.
- Termo de Compromisso de Uso de Dados: Declaração formal em que o pesquisador responsável e sua equipe se comprometem com o sigilo e a confidencialidade dos dados. Define o uso dos dados para a finalidade prevista na pesquisa.
- Termo de Transferência de Informações: Documento pelo qual o(s) pesquisador(es) transfere(m) e recebe(m) dados e informações de bancos já constituídos. Assumem a responsabilidade pela guarda, utilização e garantia do respeito ao sigilo e à privacidade.

O gestor de dados deve garantir os seguintes procedimentos ou instrumentos:

- Anonimização: Uso de técnicas para tornar os dados impossíveis de associar a um indivíduo específico, direta ou indiretamente.
- Pseudonimização: Tratamento dos dados de modo que eles não possam ser atribuídos a um titular específico sem o uso de informações adicionais mantidas separadamente e sob controle seguro.
- Autodeterminação Informativa: Direito do indivíduo de controlar e proteger seus próprios dados pessoais.
- Integridade de Pesquisa: Compromisso com valores éticos e boas práticas na condução de pesquisas.
- Termo de Acordo Institucional: Documento formal que estabelece o compromisso entre instituições para operacionalização, compartilhamento e uso de dados.
- Termo de Anuência Institucional: Documento que autoriza a realização de pesquisa em uma instituição, detalhando as atividades a serem desenvolvidas.
- Termo de Compromisso de Uso de Dados:
  - Declaração formal onde o pesquisador e sua equipe comprometem-se com o sigilo e a confidencialidade dos dados.
  - Termo de Transferência de Informações: Documento que formaliza a transferência e recepção de dados entre pesquisadores, garantindo a responsabilidade pela guarda e utilização dos dados.

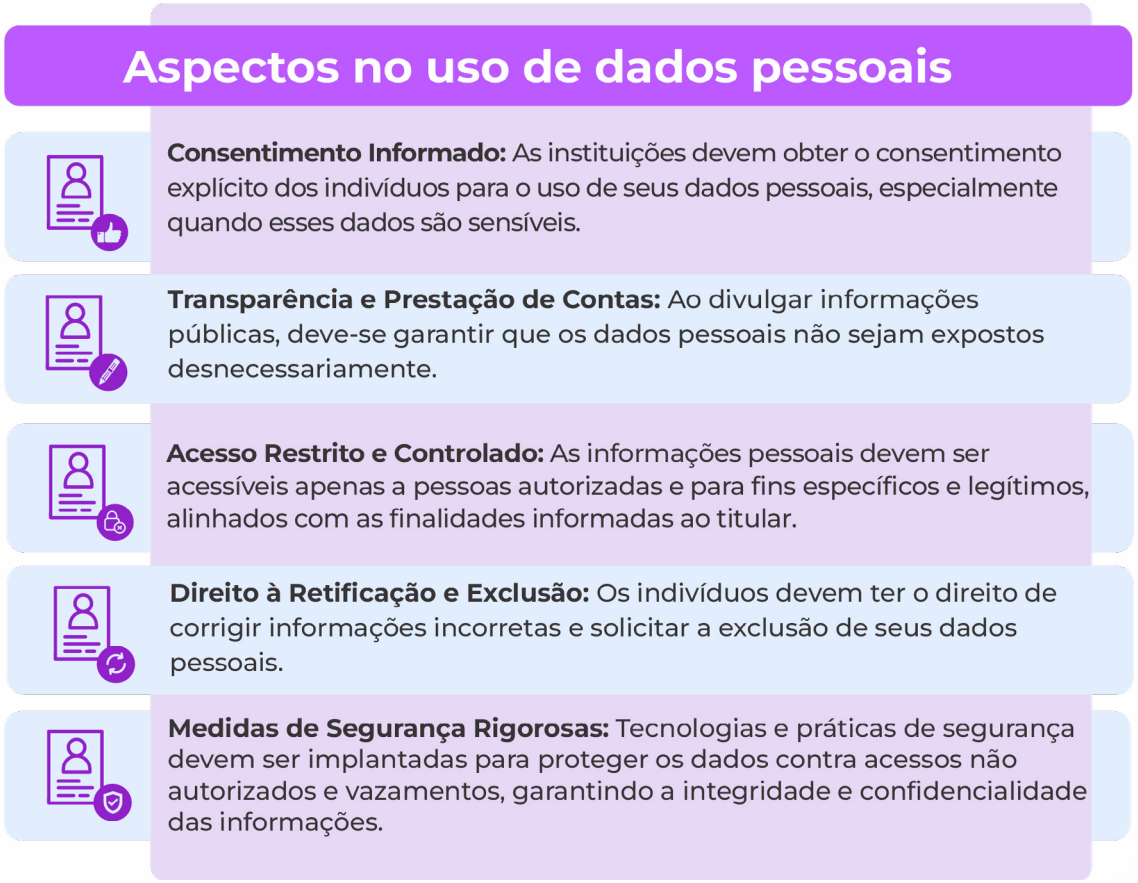
A conciliação entre a proteção de dados pessoais e o direito à transparência envolve equilibrar o direito individual à privacidade com o direito social à informação.

O direito individual à privacidade protege a integridade pessoal e a dignidade dos indivíduos, assegura que os dados pessoais sejam utilizados de forma ética e segura e garante que os titulares tenham controle sobre suas informações. O direito social à transparência e prestação de contas facilita a participação cidadã e o controle social e assegura que a sociedade tenha acesso a informações de interesse público. Na gestão de dados alguns elementos devem ser considerados.

Dados pessoais e dados sensíveis são amplamente empregados em pesquisas científicas e de saúde pública. Com o aumento da produção e utilização de dados, apoiados por tecnologias avançadas capazes de processar, correlacionar e visualizar grandes volumes de informação em diversos setores da sociedade, a regulamentação torna-se vital. Essa regulamentação é necessária para enfrentar os desafios éticos, legais e sociais, assegurando a privacidade, a confidencialidade, a segurança da informação e os direitos humanos em países democráticos. A forma de como esses dados são tratados pode variar, sendo vistos como mercadorias ou bens públicos, dependendo dos atores e dos interesses envolvidos (Almeida, 2023).

No uso de dados pessoais destacam-se diversos aspectos (Figura 39):

Figura 39 - Aspectos no uso de dados pessoais



Fonte: autoria própria.

O formulário de livre esclarecido é um documento necessário em pesquisas que envolvem seres humanos. Ele serve para informar os participantes sobre a natureza do estudo, seus objetivos, procedimentos, riscos, benefícios e a garantia de confidencialidade. O propósito deste formulário é assegurar que os participantes possam tomar uma decisão informada sobre sua participação na pesquisa.

O formulário deve ser redigido de maneira clara e acessível, evitando jargões técnicos que possam dificultar a compreensão. Os participantes devem ser informados sobre sua liberdade de aceitar ou recusar a participação, bem como sobre o direito de desistir a qualquer momento, sem que isso acarrete penalizações ou perda de benefícios.

Além de garantir transparência, o formulário de livre esclarecido também deve detalhar quem são os responsáveis pelo estudo e como os dados obtidos serão utilizados, seguindo as diretrizes éticas e legais que regem a pesquisa. Alguns formulários podem incluir informações sobre a aprovação ética do estudo, mostrando que a pesquisa foi revisada por um comitê de ética em pesquisa.

Por fim, é importante que o formulário seja assinado pelo participante, como uma confirmação de que ele compreendeu as informações apresentadas e consente em participar do estudo. Deste modo, são assegurados os direitos e o bem-estar dos participantes, promovendo a ética e a integridade na pesquisa.

3.2 Diretrizes e Princípios de Qualidade de Dados: Integridade, Precisão, Completude, Consistência, Confiabilidade, Oportunidade, Acessibilidade e Utilidade

As diretrizes e princípios de qualidade de dados garantem que as informações utilizadas nas organizações sejam confiáveis e contribuam para processos reprodutíveis. Existem diversas dimensões de qualidade (Dutra; Barbosa, 2017), vamos abordar abaixo as mais frequentes.

A integridade se refere à precisão e à coerência dos dados, assegurando que não haja perda ou corrupção de informações. A **Integridade** mede se alguma informação essencial está faltando em seus dados, assegurando a credibilidade perante o público-alvo. A **completude** ou suficiência consiste na satisfatoriedade da informação fornecida para o fim a que se propõe, ou seja, se a informação é suficiente ou insuficiente. Completude diz respeito à presença de todos os elementos necessários,

enquanto a consistência assegura que os dados não apresentem contradições internas ou externas. Refere-se, também, à disponibilidade de todos os dados necessários. Um conjunto de dados não será considerado completo se houver campos obrigatórios ausentes.

Consistência significa que os dados obedecem a regras e formatos predefinidos em diferentes plataformas e sistemas. Consistência refere-se à validade dos relacionamentos entre entidades e registros de dados, como referências cruzadas entre tabelas em um banco de dados relacional. A singularidade, exclusividade ou **univocidade** verifica se todos os registros do seu conjunto de dados são distintos e não contêm duplicatas.

A confiabilidade está relacionada à origem dos dados e à certeza de que eles foram coletados de maneira adequada. **Confiabilidade** é o grau em que os dados seguem padrões especificados, convenções e regras de negócios. Por exemplo, um número de Certidão de Pessoa Física (CPF) ou Cartão Nacional de Saúde (CNS) deve estar em formato válido e consistir com a Receita Federal ou o Sistema de Gestão do CNS (CADSUS).

Abrangência, cobertura ou alcance indica a capacidade de compreender uma vasta gama de tópicos. As métricas de integridade para medir a proporção de dados ausentes em um conjunto de dados, incluem a porcentagem de valor faltante e taxa de conclusão calculada com a porcentagem de registros com todos os campos obrigatórios preenchidos. Métricas de consistência aferem se dados obedecem a regras e formatos predefinidos, por exemplo, taxa de padronização obtida pela porcentagem de pontos de dados em conformidade com um formato específico; taxa de discrepância (*outlier*), porcentagem de pontos de dados que se desviam significativamente da norma ou de valores esperados; taxa de registros duplicados obtida com a porcentagem de registros que são cópias idênticas de outros.

A inconsistência nos dados geralmente surge devido a diferentes formatos, unidades de medida ou convenções de nomenclatura entre registros. As principais causas incluem múltiplas fontes e métodos distintos de produção de dados, mudanças nos métodos de coleta de dados ou processos de negócios em evolução. Dados inconsistentes resultam em dificuldades na integração de dados e comprometem a confiabilidade das análises. Além dessas questões, o excesso de dados também pode levar a problemas de qualidade. Esse fenômeno, frequentemente denominado sobrecarga de dados, ou mal da dimensionalidade (Altman; Krzywinski, 2018; Zaki; Meira, 2020a), ocorre quando há um volume enorme de informações para processar tanto em número de registros, quanto em atributos. Isso pode sobrecarregar recursos, retardar a análise e aumentar a probabilidade de erros.

A precisão envolve a veracidade dos dados, que devem refletir a realidade que representam. A **precisão**; em um contexto mais amplo, também chamada exatidão, acurácia ou correção; refere-se à informação livre de erro ou engano, conformidade à verdade ou a um padrão. A precisão é a capacidade dos dados em refletir o mundo real que representam. Refere-se ao grau em que os dados representam a realidade ou a verdade. A precisão pode ser difícil de medir, pois requer um ponto de referência verdadeiro. A **validade** verifica se os valores dos dados estão dentro de intervalos aceitáveis e aderem às restrições definidas. **Concisão** ou objetividade é a propriedade da informação de apresentar um conteúdo de modo reduzido, atendo-se ao essencial. A precisão dos conjuntos de dados apresenta métricas como taxa de erro expressa pela porcentagem de pontos de dados incorretos; taxa de correspondência calculada com a porcentagem de pontos de dados que correspondem a uma fonte de verdade conhecida; e erro médio absoluto obtida pela diferença média entre os pontos de dados e seus valores verdadeiros.

A oportunidade enfatiza a necessidade de que os dados sejam atualizados e estejam disponíveis no momento certo, e a acessibilidade refere-se à facilidade com que os usuários podem acessar e utilizar os dados. **Oportunidade**, atualidade ou atualização é o grau de quão recente e disponível está a informação. Avalia se a coleta e uso da informação ocorrem no momento certo. Refere-se à relevância dos dados no tempo. A oportunidade pode variar dependendo do contexto; por exemplo, dados de exames de um mês atrás podem ser considerados atuais para uma intervenção ou devem ser refeitos mediante uma rápida progressão da condição de saúde. O dado oportuno requer **acessibilidade,** bem como rastreabilidade da origem, autoria e localizabilidade ao deter a capacidade de localizar-se o ente representado pelo registro da informação quando necessário. Métricas incluem calcular a idade dos dados, isto é, o tempo médio decorrido desde que os dados foram capturados ou atualizados; latência ou tempo necessário para que os dados estejam disponíveis após sua geração e a taxa de mudança ou volatilidade com a porcentagem de pontos de dados que refletem as informações mais recentes. Dados acessíveis permitem que os profissionais de saúde tenham as informações necessárias no momento do atendimento e oferecem empoderamento ao paciente, melhorando o autocuidado e a corresponsabilidade por sua própria saúde. Recomenda-se monitorar a taxa de acesso negado, isto é, o número de tentativas de acesso negadas devido a problemas de permissão ou segurança; e o tempo de resposta de acesso ou o tempo médio necessário para acessar e recuperar dados do sistema.

A utilidade destaca a importância de que os dados não apenas existam, mas que também sejam relevantes e aplicáveis para atender às necessidades da organização e de seus tomadores de decisão (*stakeholders*). A **utilidade,** relevância ou importân-

cia é a propriedade que identifica o valor, o interesse ou a implicação da informação para o fim a que se propõe. A relevância clínica ocorre quando os dados são diretamente aplicáveis no contexto clínico. Na pesquisa, os dados devem ser disponibilizados de forma que eventos e desfechos possam ser mensurados ao longo do tempo. Por exemplo, os Dados Abertos de Hospitalização no SUS (AIH) não permitem contar o número de usuários nem a reincidência no tratamento, visto que não apresentam um identificador pseudonimizado, ao contrário dos dados abertos de medicamentos do CEAF disponíveis no SIA, os quais são úteis para avaliar a persistência no tratamento. É importante mensurar a taxa de satisfação dos usuários ou a proporção de usuários satisfeitos com a qualidade e relevância dos dados fornecidos; bem como estabelecer um índice de aplicabilidade com avaliação qualitativa da utilidade dos dados para a prática clínica, pesquisa e gestão.

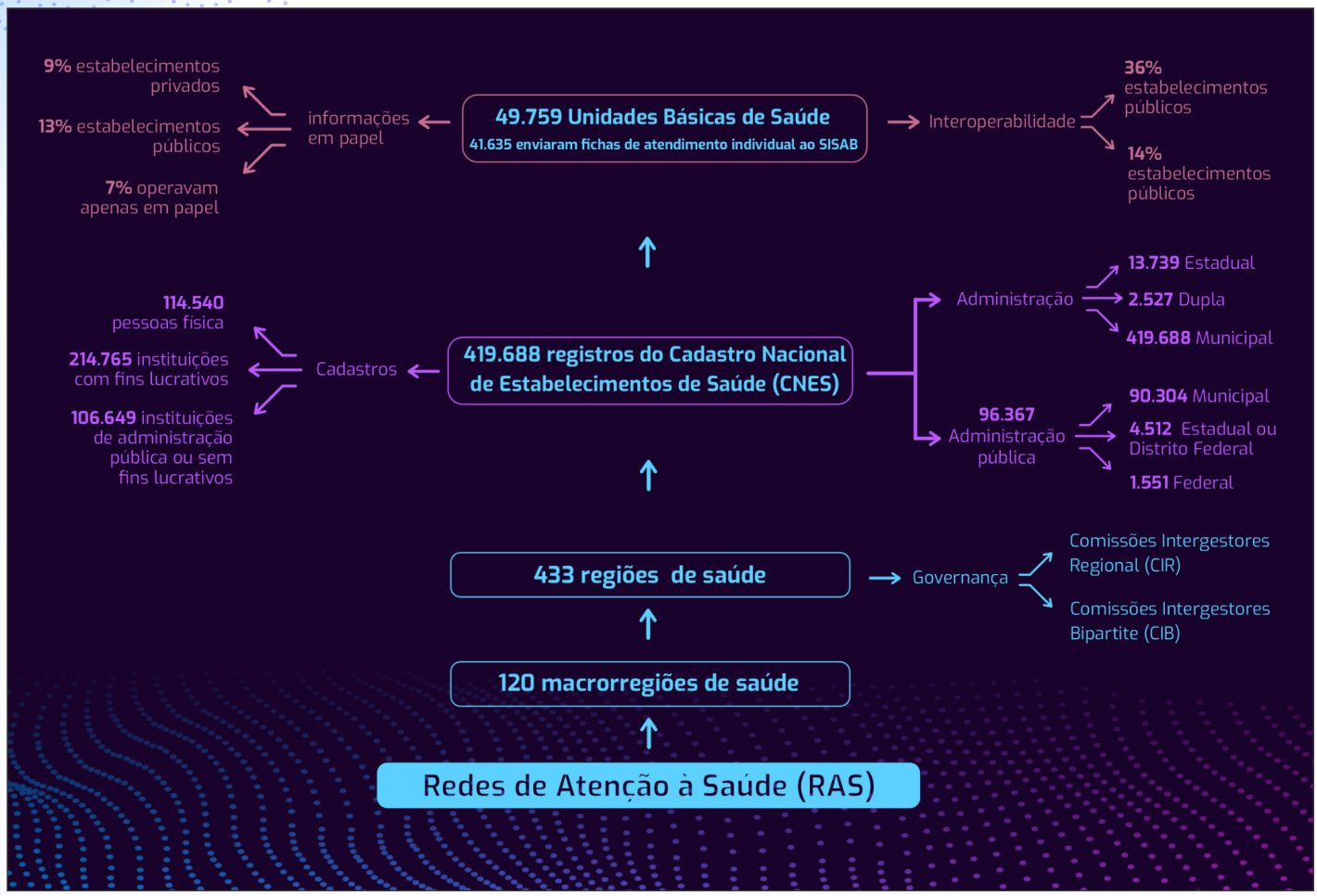
Layout, concepção gráfica, design ou **aparência** avalia se aspectos como cores, letras, tamanhos e estrutura física; e contribuem ou interferem na leitura ou apresentação da informação.

3.3 Indicadores de Saúde Digital

Indicadores de qualidade de dados de saúde são métricas utilizadas para avaliar a eficácia dos sistemas de informação em manter dados precisos, completos, atualizados, unívocos e úteis. Esses indicadores ajudam a identificar áreas que necessitam de melhoria e a garantem que os dados atendam aos requisitos essenciais para a prática clínica, pesquisa, administração e outras atividades no setor de saúde.

A pesquisa TIC Saúde oferece uma visão abrangente sobre a infraestrutura de TI e comunicação (TIC) nos estabelecimentos de saúde, destacando seu impacto na qualidade e integridade dos dados. Realizada desde 2013, investiga a adoção de TIC, com ênfase em médicos e enfermeiros, abordando áreas como gestão de TI, RES, telessaúde, e novas tecnologias, além de identificar barreiras para sua implantação. Apoiada por órgãos governamentais e especialistas, a pesquisa segue metodologias internacionais e utiliza dados do CNES para garantir resultados de qualidade (Brasil, 2023b). A título de exemplo, vamos destacar alguns indicadores (Figura 40).

Figura 40 - Infográfico: como orquestrar a informatização da saúde do Brasil?



Fonte: Brasil, 2024f; 2024g; 2024h; CETIC 2024a, 2024b.

- Dimensão A - Infraestrutura de TIC e Gestão de TI:
  - A1 - Estabelecimentos de Saúde que Utilizaram Computadores nos Últimos 12 Meses, como pressuposto da digitalização dos dados de saúde, permitindo a coleta, armazenamento e análise eficiente dos dados;
  - A2 - Estabelecimentos de Saúde que Utilizaram Internet nos Últimos 12 Meses, uma vez que o acesso à internet facilita a troca de informações em tempo real, a telessaúde e o acesso a sistemas de saúde integrados, essenciais para a manutenção de dados precisos e atualizados;
  - A3 - Estabelecimentos de Saúde com Acesso à Internet, por Tipo de Conexão (como banda larga ou fibra óptica), os quais influenciam a velocidade e a confiabilidade da transmissão de dados, impactando diretamente a eficiência e integridade das informações de saúde;
  - A5 - Estabelecimentos de Saúde com Internet, por Faixa de Velocidade Máxima para *Download* da Principal Conexão, onde velocidades de conexão mais altas permitem transferências de dados mais rápidas e acesso a sistemas complexos, garantindo que informações cruciais estejam disponíveis quando necessário.

- A6 - Estabelecimentos de Saúde que Possuem Departamento ou Área de TI, pois a existência de um departamento de TI dedicado assegura a gestão e manutenção adequadas das infraestruturas de TIC, contribuindo para a qualidade e segurança dos dados de saúde. A tendência é elevar as áreas de TIS no organograma, conforme observado no poder executivo com o estabelecimento da SEIDIGI, MS, e nas SES de Goiás e da Bahia (CONASS; Oliveira, 2024).
- A8 - Estabelecimentos de Saúde, por Existência de Documento que Define uma Política de Segurança da Informação, visando proteger dados sensíveis de pacientes, evitando vazamento e acessos não autorizados.

- Dimensão B - Registro Eletrônico em Saúde e Troca de Informações:
  - B0 - Estabelecimentos de Saúde, por Existência de Sistema Eletrônico para Registro das Informações dos Pacientes, visto que RES e o Prontuário Eletrônico do Paciente (PEP) garantem que os dados dos pacientes sejam registrados de maneira consistente e acessível, facilitando a continuidade do cuidado e a troca de informações.
  - B1 - Estabelecimentos de Saúde, por Forma de Manutenção das Informações Clínicas e Cadastrais nos Prontuários dos Pacientes versus papel, pois afeta a facilidade de acesso e a precisão das informações, com prontuários eletrônicos oferecendo maior integridade e atualização dos dados.
  - B2 - Estabelecimentos de Saúde, por Tipo de Dado sobre o Paciente Disponível Eletronicamente, uma vez que a disponibilidade de diversos tipos de dados (histórico médico, resultados de exames, etc.) em formato eletrônico melhora a precisão diagnóstica e a personalização dos tratamentos.
- Dimensão C - Serviços Oferecidos ao Paciente e Telessaúde:
  - C1 - Estabelecimentos de Saúde, por Serviços Oferecidos ao Paciente via Internet, tais como serviços online, como agendamento de consultas e acesso a resultados de exames, melhoram a conveniência para os pacientes e a eficiência operacional dos estabelecimentos de saúde.
  - C2 - Estabelecimentos de Saúde, por Serviços de Telessaúde Disponíveis, pois a telessaúde amplia o acesso ao atendimento médico, especialmente em áreas remotas, garantindo que mais pacientes recebam cuidados necessários.
- Dimensão D - Novas Tecnologias, onde é avaliada a capacidade de adoção do que está em tendência o mercado:
  - D1 - Estabelecimentos de Saúde que Utilizaram Serviços em Nuvem, pois a tecnologia de armazenamento de dados oferece escalabilidade,

segurança e acessibilidade, contribuindo para a integridade dos dados de saúde.

- D2 - Estabelecimentos de Saúde que Fazem Análises de Big Data para a identificação de tendências e padrões em grandes volumes de dados de saúde, apoiando a tomada de decisões baseadas em evidências.
- D6 - Estabelecimentos de Saúde que utilizam tecnologia de IA, por tipo, ante à tendência da IA na melhoria da precisão diagnóstica, prever desfechos clínicos e otimizar processos administrativos, ante ao progresso da qualidade e integridade dos dados.

3.4 Saiba Mais - Atividade de Leitura Opcional

3.4.1 Ferramentas de Segurança de Sistemas Computadorizados

Os quesitos de segurança da informação descritos a seguir fazem parte dos cuidados com dados pessoais e sensíveis. Existem cuidados a serem tomados em organizações, como o uso de ferramentas de antivírus; anti-spam para e-mails não solicitados; firewall com barreiras a ataques de rede e controle de acesso externo e interno; sistema de detecção de intrusos IDS (do inglês, *Intrusion Detection System*) e IPS (do inglês, *Intrusion Prevention System*), com ferramentas de monitoramento, bloqueio e resposta a incidentes; uso de redes privadas virtuais - VPN (*Virtual Private Network*); configuração de proxy para requisições de acesso a páginas da web com autenticação e autorização do usuário, filtro de conteúdo e acesso a sítios listados de fontes confiáveis; auditorias externas por parte de empresas especializadas e órgãos reguladores, garantindo que as instituições de saúde estejam em conformidade com as normas; entre outros.

A organização deve estabelecer ainda, medidas de proteção contra **vazamento da informação** (*Data Loss Prevention* - DLP), como Política PCI/DSS para identificação de números de cartão de crédito; *Finger Print*, a qual é uma configuração que permite a identificação (total ou parcial) de informação contida em arquivos pré-selecionados pela organização; classificação da Informação para identificação de conteúdo de acordo com a política de classificação da informação da empresa (sigilosa/setorial/interna); CPF/CNPJ com identificação de possível vazamento da carteira de pacientes/fornecedores, de acordo com a quantidade de CPFs/CNPJs identificados; Palavra chave que viabilizam o monitoramento de palavras definidas pelas áreas de negócio.

Hardening, também conhecido como blindagem de estações de trabalho ou servidores, é um processo de mapeamento das ameaças, mitigação dos riscos e execução das atividades corretivas, com foco na infraestrutura e objetivo principal de torná-la preparada para enfrentar tentativas de ataque. O processo inclui remover ou desabilitar nomes ou logins de usuários que não estejam mais em uso, além de serviços desnecessários. Outras providências que um processo de *hardening* pode incluir limitação do *software* instalado àquele a que se destina a função desejada do sistema; aplicar e manter os patches atualizados, tanto de sistema operacional quanto de aplicações; revisar e modificar as permissões dos sistemas de arquivos, em especial no que diz respeito à escrita e execução; reforçar a segurança do login, impondo uma política de senhas fortes. Deve ser aplicado na instalação de uma estação de trabalho ou servidor. Além disso, ele deve ser aplicado logo após qualquer alteração desses ambientes.

A gestão de *patches* também deve ser considerada. *Patch* é um programa criado para atualização ou correção de um programa. *Patch* de segurança é utilizado para evitar a exploração de uma vulnerabilidade de um sistema operacional, aplicativo ou programa. A **Gestão de *Patches*** consiste em mapear os sistemas e suas respectivas versões, para que os patches, principalmente os de segurança, sejam aplicados o mais rápido possível. O teste do patch é realizado sempre em um ambiente de homologação apartado da do ambiente de produção. Deve-se priorizar a aplicação nos servidores de borda, mais expostos à internet. Antes da aplicação de um patch, há necessidade da realização de um backup do sistema. Deve-se utilizar o processo de Gestão de Mudanças para a aplicação de patches.

Na **gestão de vulnerabilidades** o objetivo é reduzir os riscos decorrentes da exploração de vulnerabilidades técnicas conhecidas por meio de um acompanhamento periódico, que inclui publicações das empresas fornecedoras de *software*. Isso envolve a avaliação da exposição da organização a essas vulnerabilidades e a adoção de medidas apropriadas para lidar com os riscos associados, além de configurar o ambiente utilizando as mesmas diretrizes da Gestão de Patches.

A **gestão de conformidade** é um sistema de gerenciamento online que identifica alterações no ambiente que não estão em conformidade com a Política de Segurança configurada, incluindo exemplos como a alteração da quantidade de caracteres da senha do usuário, a desativação da monitoração, a não aplicação do último patch de segurança, a não atualização da política de antivírus e o armazenamento de informações sigilosas sem criptografia. Além disso, permite a integração dos alertas com o sistema de monitoração da empresa e oferece recursos de rastreamento, relatórios e correções automáticas.

As evidências (***logs***) referem-se a uma funcionalidade do sistema que possibilita a rastreabilidade do ambiente e deve ser previamente configurada pelo administrador, permitindo a identificação de elementos como endereço de origem e destino, serviço ou ação realizada, usuário e data/hora, sendo crucial a sincronização de horários dos logs (*Network Time Protocol* [NTP] para sincronizar relógios numa rede) de modo a garantir a correta rastreabilidade de eventos. A segregação de funções e a utilização de um concentrador de logs asseguram a integridade das informações armazenadas. Além disso, o *Security Information Event Management* (SIEM) é uma solução avançada que gerencia eventos e informações de segurança, dispondo também da funcionalidade de abertura de chamados.

3.4.2 Medidas de Segurança Contempladas na Rede Nacional de Dados em Saúde

Existem medidas de proteção, previstas na Rede Nacional de Dados em Saúde (RNDS), incluindo aspectos de governança de informação, definição de papéis e responsabilidades, categorização dos dados, segurança da RNDS, gestão de continuidades, tratamento de dados, consentimento, pesudonimização, transparência e direitos do titular (Brasil, 2020a, c).

O acesso aos dados da RNDS estão restritos ao usuário do SUS e aos profissionais de saúde no contexto de atendimento. O titular do dado acessa via Aplicativo Conecte SUS cidadão, mediante autenticação realizada por meio de acesso *gov.br* dos serviços públicos digitais. O profissional de saúde acessa os dados do cidadão via Conecte SUS Profissional, mediante autenticação do Certificação Digital ICP-Brasil de instalações de Prontuário Eletrônico habilitado para o estabelecimento de saúde.

A estratégia de consentimento do usuário aos respectivos dados aplica os conceitos de *opt-in* e *opt-out*. *Opt-in* é o consentimento explícito do usuário para o tratamento de seus dados pessoais, assegurando que os dados só sejam coletados e usados após a aprovação consciente do titular. *Opt-out* é o direito do titular de retirar seu consentimento ou de solicitar a exclusão de seus dados pessoais, limitado às restrições legais de fins de vigilância. No SUS a regra prevalente é o *opt-out*. Na iniciativa privada, visando evitar conflitos de interesse, o padrão é o *opt-in*.

Ainda, são previstos recursos de anonimização e pseudonimização para proteger dados pessoais, dificultando a identificação dos indivíduos, mas viabilizando ações de pesquisa, transparência e prestação de contas.

3.4.3 Definições para Fins de Proteção de Dados Pessoais

Conheça as definições de dados e banco de dados para fins de proteção de dados pessoais.

- Banco de Dados: Conjunto estruturado de dados relativos a pessoas, que pode ser armazenado em formato eletrônico ou físico.
- Dados: Unidades de descrição das variáveis que compõem um banco de dados, podendo ser representadas por palavras, números, símbolos, imagens, entre outros.
- Dado Anonimizado: Dado que não pode ser associado a um titular específico devido ao uso de técnicas de anonimização.
- Dado Pessoal: Informação relacionada a uma pessoa identificada ou identificável.
- Dado Pessoal Sensível: Dado pessoal que revela origem racial ou étnica, convicções religiosas, opiniões políticas, filiação sindical, dados sobre saúde ou vida sexual, ou dados genéticos ou biométricos.
- Dado Identificador: Informação que pode ser vinculada diretamente à identidade de um indivíduo.
- Informação agregada: Dados combinados de um grupo de indivíduos, impossíveis de detalhar no nível individual.

AKCIT
CENTRO DE COMPETÊNCIA EMBRAPII EM TECNOLOGIAS IMERSIVAS

Gestão de
Dados de Saúde

Unidade IV

Técnicas Inteligentes para Processamento de Dados de Saúde

Unidade IV: Técnicas Inteligentes para Processamento de Dados de Saúde

4.1 Introdução a Técnicas Inteligentes para Processamento de Dados

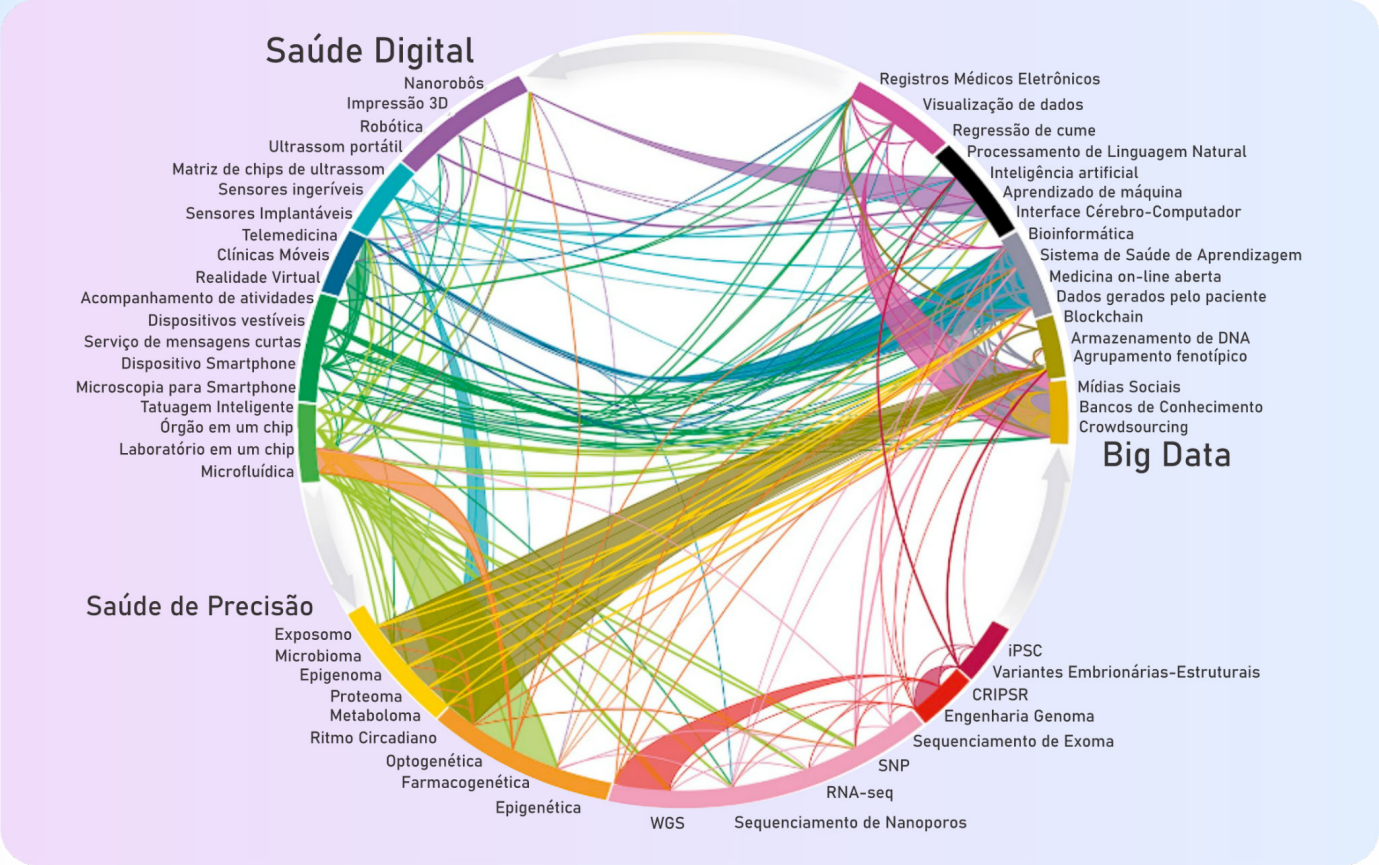
A tecnologia avança com maior velocidade do que nossa capacidade de absorvê-las e aplicá-las eticamente no mundo, ampliando as formas de inclusão, sofisticação e aprimoramento do tecido social com as oportunidades que a inovação tecnológica traz e os desafios de exclusão e geração de iniquidades. O presente curso é um exemplo de uso da tecnologia em transformação. Ao disponibilizar conteúdo em meios digitais, utilizar ambientes interativos de programação em comunidade, como a Google Colab, e incentivar a programação usando IA generativa, estamos promovendo o uso inteligente de tecnologias.

Numa busca rápida na literatura ou em buscadores da web é possível identificar a efervescência de técnicas inteligentes para processamento de dados. A inteligência é uma característica humana atribuída a máquinas capazes de lidar com os mesmos contextos complexos. A informática em saúde é, por si só, inteligente e inovadora, pois funde duas áreas do conhecimento, resultando em algo totalmente novo. Segundo a Sociedade Brasileira de Informática em Saúde (SBIS):

> A Informática em Saúde é a área do conhecimento que trata da aplicação de conceitos e Tecnologias de Informação e Comunicação (TIC) para a melhoria e transformação de sistemas, serviços e processos de Saúde. O termo e-Saúde (tradução do inglês – *eHealth*) pode ser entendido dentro da mesma definição de informática em saúde e tem sido um termo bastante utilizado recentemente (SBIS, 2019).

Veja, na Figura 41, quantas áreas de fronteira tecnológica estão no escopo da informática em saúde.

Figura 41 - Infográfico de inovações e desenvolvimentos emergentes em saúde digital, *big data* e saúde de precisão e suas intraconexões e interconexões



Fonte: Bhavnani *et al.* (2017).

Cada avanço tecnológico não implica em descontinuidade do anterior. Assim, as planilhas de cálculo, cujo *boom* ocorreu nos anos 1990, a popularização de bancos de dados estruturados nos anos 2000, a inserção de ferramentas de inteligência de negócios nos anos 2010 e a popularização da IA e IoT nos anos 2020 não implicou no desuso das tecnologias citadas. Ainda hoje, o profissional de saúde que domina planilhas de cálculo apresenta destaque nas instituições que trabalha.

Dados tabulares referem-se a informações organizadas em tabelas, como RES, resultados de exames laboratoriais, dados demográficos e históricos de tratamento. Técnicas inteligentes podem ser aplicadas para extrair valor desses dados.

Técnicas de aprendizado de Máquina ML apoiam na previsão de resultados clínicos, como a predição de dado paciente desenvolver uma condição específica ou responder a um tratamento; identificam padrões anômalos que podem indicar fraudes em seguros de saúde; classificação de pacientes em grupos com base em características similares, permitindo tratamentos personalizados. A descoberta de padrões ocultos em grandes conjuntos de dados, conhecida como mineração de dados, apoia a otimização de recursos de saúde com a identificação de padrões de utilização de recursos para melhorar a alocação e gestão.

Doenças crônicas são monitoradas ao incluir a análise de dados longitudinais, isto é, coletados ao longo do tempo para identificar tendências e mudanças. A análise de dados de pacientes com doenças crônicas viabiliza identificar mudanças no estado de saúde e ajustar tratamentos.

Devido aos altos custos envolvidos no desenvolvimento de medicamentos e nas despesas hospitalares resultantes do uso ineficaz ou prejudicial, é importante adotar novas maneiras de prever eventos não detectados (Burt *et al.*, 2022). Para adotar os algoritmos mais avançados na área de ciências da computação, é necessário ter uma infraestrutura que integre informações do prontuário eletrônico e permita sua disponibilidade para fins acadêmicos (Ferré, 2021). A Estratégia de Saúde Digital da Política Nacional de Informação e Informática (PNIIS) no Brasil prevê um ambiente inovador com amplo potencial para integrar dados a fim de apoiar decisões imediatas na área (Brasil, 2022e; Pires, 2020). Particularmente no contexto do SUS e de países com sistemas de saúde pública orientados pela saúde universal, espera-se que iniciativas que incorporem código-fonte aberto e dados adotem métodos de IA para prever eventos adversos (Ferré *et al.*, 2020).

A modelagem do problema é a principal diferença entre as abordagens de IA e os métodos estatísticos tradicionais. Enquanto o método estatístico se baseia na inferência estatística com variáveis previamente definidas em relação à hipótese, geralmente relacionada à equivalência ou divergência dos objetos comparados, na IA são admitidas relações não explicadas pelo paradigma do conhecimento aplicado como espúrias para apoiar achados biológicos.

A descoberta de conhecimento em bancos de dados, também conhecida como mineração de dados (KDD), é uma técnica utilizada na IA. Diferentemente dos métodos estatísticos tradicionais, a seleção prévia dos atributos não é recomendada no KDD para evitar problemas com dimensionalidade. Em vez disso, as dependências são tratadas e os fatores dependentes são detectados durante o processo. Além disso, nas avaliações adversas do evento biológico ocorre a necessidade da exclusividade das hipóteses classificatórias ao invés somente uma possibilidade explicativa ser imperativa - esses modelos alternativos rompem com paradigmas antigos que limitavam apenas um tipo possível de solução explanadora para eventos biológicos específicos (Zaki; Meira, 2020b), (Altman; Krzywinski, 2018; Verleysen; François, 2005), (Ribeiro; Garcia; Dos Santos, 2022). Assim como a estrutura de tomada de decisão em saúde envolve múltiplos atores, a criação do Logo inaugura uma nova qualidade de métodos capazes de lidar com estruturas complexas. Isso se deve à associação entre habilidades humanas e computacionais para detectar padrões sem a necessidade simplificá-los ou decompô-los (Arruda; Lopes; Koerich, 2015; Morin, 2015).

Identificar a causalidade dos eventos adversos não é fácil e requer treinamento dos profissionais de saúde, bem como observação persistente dos fenômenos em grandes populações. Porém, mesmo para doenças raras ou novos medicamentos, os métodos computacionais podem fornecer suporte à tomada de decisão, combinando diversas fontes de dados cujo significado da aplicação é verificado *a posteriori* com o detentor do domínio do conhecimento, também conhecido como especialista *ad hoc*. Diversos repositórios são usados para avaliação de eventos adversos como a Classificação Anatômica, Química e Terapêutica ATC/OMS (World Health Organization, 2022), mapas metabólicos, estruturas moleculares e mapas genômicos dispostos em ontologias ("Gene Ontology resource", [s.d.]; Wixon; Kell, 2000) e bancos de dados de medicamentos e interações medicamentosas como o drugs.com (Cerner Multum, 2010), Medscape Multi-Drug Interaction Checker ("Drug interactions checker - medscape drug reference database", [s.d.]), RxList (WEBMD, 2004), DrugBank ("DrugBank Online", [s.d.]) e DrugComb (Zheng *et al.*, 2021).

Há outros compostos usados de forma análoga, mas cujo uso digital é limitado porque não estão disponíveis para consumo direto por algoritmos. Além disso, na Assistência Farmacêutica, a título de exemplo, existem diversas outras bases de dados com grande potencial a ser explorado em estudos sobre o uso de medicamentos (LEAL *et al.*, 2021b). Veja, na Tabela 2, exemplos de trabalhos de IA aplicados a reações adversas.

Existem outras soluções inovadoras que favorecem o ecossistema de técnicas inteligentes em saúde, a exemplo da blockchain e de soluções de interoperabilidade de prontuários eletrônicos com terminologias e ontologias, como vem sendo desenvolvido a partir do domínio BR-Core da RNDS (Brasil, 2021a; Saeed; Mohammed; Al-Nahari, 2021). Outro conjunto de técnicas emergentes, agremiando diversas tecnologias bem conhecidas, é o Processamento Inteligente de Documentos (*Intelligent Document Processing* - IDP) para organizar informações não estruturadas e semiestruturadas em dados utilizáveis, tais como arquivos texto, e-mails, imagens e PDFs. IDP aplica tarefas de automação, capazes de capturar, extrair e processar dados de vários formatos usando tecnologias de IA (Arora *et al.*, 2024; Ling; Gao; Wang, 2020; Mandvikar, 2023).

Tabela 2 - Inteligência artificial aplicada a reações adversas relacionadas a medicamentos

Referência	Periódico	Técnica	Resultado
Kurosaki; Uesawa, 2022	*The Journal of Toxicological Sciences*	*LightGBM, XGBoost, Random Forest, Neural Network e Support Vector Machine SVM*	O modelo preditivo apontou que agentes antivirais da hepatite C, como paritaprevir, pibrentasvir e glecaprevir, e imunossupressores, como tacrolimus, sirolimus e ciclosporina, que são supostamente associado à hepatocarcinogenicidade, cujos achados apresentam corroboração da literatura.
Galati *et al.*, 2022	*International Journal of Molecular Sciences*	*Neural Network*	Plataforma VenomPred (http://www.mmvsl.it/wp/venompred/), uma ferramenta web disponível gratuitamente e de fácil acesso mesmo para usuários não especialistas verificarem a toxicidade potencial de pequenas moléculas.
Kang; Kang, 2021	*Molecules (Basel, Switzerland)*	*Neural Network, Rotating Forest e Ensemble Learning, PCA e MCA*	O modelo consegue prever lesão hepática induzida por drogas com potencialidade para uso nos estágios iniciais de desenvolvimento de medicamentos, garantindo a segurança humana.
Abuhelwa *et al.*, 2022	*Cancer Chemotherapy and Pharmacology*	*Random Forest*	Foi construída ferramenta para predição clínica definida por sexo feminino, hemoglobina alta e bilirrubina baixa com alta discriminação para prever a síndrome mão-pé (SHF), o qual é um efeito adverso grave associado à terapia com sorafenibe.
Létinier *et al.*, 2021	*Clinical Pharmacology and Therapeutics*	*Boosting Trees (LGBM)*	O modelo foi capaz de prever Reação Adversa a Medicamentos (RAM) a partir de textos com dados da vida real, anotados por especialistas.
Martin *et al.*, 2022	*Drug Safety*	*Light Gradient Boosted Machine (LGBM)*	O modelo foi capaz de prever Reação Adversa a Medicamentos (RAM) a partir de textos de registros de pacientes e da terminologia médica *Medical Dictionary for Regulatory Activities* (MedDRA).
Jain; Raj; Mishra, 2021	*BMC Bioinformatics*	*Natural language processing NLP e neural network*	A ferramenta foi capaz de prever reações adversas a medicamentos em mono e em terapia combinada a partir de bases de dados relacionais de textos biomédicos.
Feng; Zhang, 2022	*Molecules (Basel, Switzerland)*	*Neural network*	O modelo é capaz de detectar interações medicamentosas potenciais conhecidas e desconhecidas e auxilia na elucidação de mecanismos de ação subjacentes aos pares de fármacos.
He Et Al., 2022	*BMC Bioinformatics*	*Neural network*	O modelo reúne informações de redes moleculares de fármacos, sequências SMILES para previsão de interações medicamentosas.
Zhang; Lu; Zang, 2022	*BMC Bioinformatics*	*Neural network*	Previsão da interação entre pares de fármacos drogas interagem e tipificação da interação.
Masumshah; Aghdam; Eslahchi, 2021	*BMC Bioinformatics*	*Neural network, PCA*	Prevê efeitos colaterais da polifarmácia onde cada fármaco é representado por um vetor de características baseado em mono efeitos colaterais e interações fármaco-proteína.

Fonte: Ferré (2023a).

A seguir, mostraremos técnicas inteligentes de processamento de dados, ilustrando etapas do aprendizado de máquina e tipos de IA a processamento de texto e imagens.

4.1.1 Aprendizado de Máquina

Um sistema computadorizado autômato restrito a regras pré-programadas deve ser modificado por humanos para se adaptar. Entretanto, existem técnicas adaptativas que podem fazer com que as máquinas tenham respostas diferentes conforme o contexto, a exemplo das inteligências artificiais generativas, como chatGPT, character.ai, Bard, Poe, QuillBot, MidJourney, entre outras.

As etapas de processamento de dados, extração, engenharia, processamento e análise, apresentam contribuições de técnicas desenvolvidas ao longo das últimas décadas, conhecidas como cibernética, descoberta de conhecimento em banco de dados KDD, mineração de dados e, atualmente, IA dentro da indústria 4.0.

Aprendizado de máquina é um campo da IA que se concentra no desenvolvimento de algoritmos que permitem que os computadores aprendam a partir de dados e façam previsões ou decisões sem serem explicitamente programados para tarefas específicas. O processo do aprendizado de máquina envolve várias técnicas de pré-processamento e modelagem dos dados, a escolha e aplicação de algoritmos apropriados e, finalmente, a validação dos modelos para garantir sua eficácia e precisão. As tarefas do aprendizado de máquina são descritivas, preditivas e prescritivas.

O **aprendizado de máquina não supervisionado** trabalha com dados não rotulados, buscando identificar padrões ou estruturas latentes. Algoritmos comuns nessa categoria incluem a análise de agrupamento (*clustering*) e a redução de dimensionalidade. O objetivo é descobrir agrupamentos naturais ou diminuir a complexidade dos dados sem qualquer supervisão externa. Como não há saídas conhecidas para comparar, a avaliação da qualidade desses modelos é mais desafiadora e pode envolver a análise da coesão dos agrupamentos ou a variância explicada na redução de dimensionalidade. Por outro lado, algoritmos não supervisionados trabalham com dados sem rótulos e se concentram em identificar padrões ou agrupamentos inerentes. Exemplos incluem K-Means, Agrupamento Hierárquico, PCA e Redes Neurais Autoencoders. Esses algoritmos são usados para tarefas como agrupamento, redução de dimensionalidade e detecção de anomalias.

O **aprendizado de máquina supervisionado** envolve a construção de modelos a partir de um conjunto de dados rotulados, onde cada exemplo de entrada é associado a uma saída desejada. O objetivo é treinar o modelo para que ele possa prever a

saída correta para novos dados, baseando-se nesses exemplos. Exemplos comuns de algoritmos supervisionados incluem a regressão linear, as Máquinas de Vetores de Suporte (SVM), redes neurais e árvores de decisão. Durante o treino, o modelo ajusta seus parâmetros para minimizar a diferença entre suas previsões e as saídas reais nos dados de treino. Depois, na fase de teste, o desempenho do modelo é avaliado em um conjunto de dados separado, que o modelo não observou durante o treino, para verificar sua capacidade de generalização.

Em outras palavras, algoritmos supervisionados são treinados usando um conjunto de dados que inclui tanto as características (variáveis independentes) quanto os rótulos (variáveis dependentes). Esses algoritmos aprendem a mapear entradas para saídas e são avaliados com base na sua capacidade de prever corretamente os rótulos em novos dados. As etapas de treino e teste no aprendizado supervisionado começam com a divisão do conjunto de dados em subconjuntos (*subsets*) de treino e teste. O modelo é treinado usando o conjunto de treino, onde ajusta seus parâmetros para otimizar seu desempenho com base nas entradas e saídas fornecidas. Após o treino, o modelo é testado no conjunto de teste, onde suas previsões são comparadas com as saídas reais para avaliar a precisão e a capacidade de generalização. No aprendizado não supervisionado, o conjunto de dados é geralmente usado na íntegra para identificar padrões durante o treino, e a avaliação pode envolver métricas internas ou a aplicação do modelo a novos dados para verificar a consistência dos padrões descobertos.

O **pré-processamento** de dados é uma etapa do aprendizado de máquina, pois a qualidade dos dados de entrada impacta diretamente no desempenho do modelo. Primeiramente, a **seleção de atributos** envolve identificar e escolher as características mais relevantes que influenciam a variável de interesse. Métodos como análise de correlação e Análise de Componentes Principais (PCA) são frequentemente usados para este propósito. Em seguida, a normalização é aplicada para transpor os dados para uma faixa específica, geralmente entre [0, 1] ou [-1, 1], utilizando técnicas como Min-Max Scaling ou Z-score Normalization. Na **limpeza** dos dados ocorre imputação de valores ausentes, remoção de discrepâncias (*outliers*) e dados duplicados, garantindo que o conjunto de dados esteja livre de inconsistências e pronto para o treinamento do modelo.

Quanto à modelagem dos dados é relevante acrescentar a modelagem em grafos e as estratégias dos modelos de kernel. Dados em formato de grafo, que representam relações complexas, demandam modelagem matemática de redes de nós e arestas. Essas técnicas exploram a conectividade e a estrutura dos dados, permitindo que o modelo entenda não apenas as propriedades dos nós, mas também as interações entre eles. Métodos de kernel lidam com dados que não podem ser separáveis

linearmente ao transformarem dados em um espaço de alta dimensionalidade, onde se tornam linearmente separáveis, permitindo a aplicação de algoritmos como SVM.

Os quatro principais grupos de técnicas de aprendizado de máquina são mineração de padrões frequentes; agrupamento; classificação e regressão.

Mineração de padrões frequentes é uma técnica utilizada para identificar padrões que se repetem em conjuntos de dados. Essas técnicas são aplicáveis em diversas áreas, como análise de co-ocorrências, detecção de fraudes e recomendações, permitindo a extração de informações não triviais a partir dos dados disponíveis.

Pattern and Rule Assessment (avaliação de padrões e regras) consiste na avaliação da qualidade dos padrões e regras identificados durante a mineração. Nesse processo, métricas como suporte e confiança são utilizadas para validar os resultados, garantindo que os padrões encontrados sejam relevantes e tenham significado dentro do contexto dos dados analisados.

Itemset Mining (mineração de conjuntos de itens) identifica conjuntos de itens que aparecem juntos em transações. Essa técnica utiliza algoritmos como Apriori ou *FP-Growth* para calcular a frequência de *itemsets* (conjuntos de itens) e determinar quais combinações de itens são mais comuns num conjunto de dados. O resultado é uma lista de conjuntos de itens frequentes que podem ser utilizados para análise adicional. Um exemplo é a utilização do algoritmo Apriori para identificar padrões frequentes numa prescrição, encontrando, por exemplo, que “dipirona” e “diclofenaco” são frequentemente comprados juntos numa farmácia comercial.

Summarizing Itemsets (resumo de conjuntos de itens) refere-se à apresentação dos resultados da mineração de *itemsets* de forma concisa e informativa. O propósito é relatar os conjuntos de itens frequentes juntamente com medidas quantitativas como suporte, confiança e lift (aumento), proporcionando uma visão clara das associações encontradas entre os itens e facilitando a comunicação dos resultados. Suporte (ou *Support*) é uma métrica que indica a frequência com que um conjunto de itens aparece em um conjunto de dados. É calculado como a proporção de transações que contêm o conjunto de itens em relação ao total de transações no banco de dados. O suporte é utilizado para identificar os conjuntos de itens mais frequentes. Confiança (ou Confidence) é uma métrica que mede a robustez de uma regra de associação. Ela expressa a probabilidade de que um item está presente em uma transação, dado que um conjunto de itens foi observado. É calculada como a proporção do suporte do conjunto de itens e o suporte do antecedente da regra de associação.

Confiança reflete a força da associação entre os itens, indicando até que ponto a presença de itens no conjunto **A** leva à presença de itens no conjunto **B**. Utilização de técnicas como a descrição de padrões onde, por exemplo, um resumo pode indicar que 70% dos clientes que compram produtos orgânicos também compram produtos cosméticos ecológicos.

Sequence Mining (mineração de sequências) analisa sequências de eventos ou transações ao longo do tempo, identificando padrões que ocorrem em uma ordem específica. Essa técnica é aplicada em áreas como análise de comportamento do consumidor e previsão de séries temporais, ajudando a entender como os eventos estão interligados em uma sequência temporal. Algoritmos como GSP (Generalized Sequential Pattern) que identificam sequências de comportamentos do usuário em navegação na web, como a sequência de páginas visitadas.

Na ***Graph Pattern Mining*** (mineração de padrões em grafo) os nós representam entidades e as arestas refletem as relações entre elas. Essa técnica é útil em contextos como redes sociais e bioinformática, onde as interações complexas têm papel importante na estrutura dos dados analisados. Algoritmo para detectar padrões recorrentes em redes sociais, como comunidades de usuários que interagem entre si com base em suas conexões e interações.

O **agrupamento** representa uma técnica utilizada para agrupar dados em conjuntos, onde os itens dentro de cada conjunto são mais semelhantes entre si do que aqueles de outros conjuntos. Esse processo é útil em diversas aplicações, como segmentação de mercado e organização de grandes volumes de dados.

Representative-based Clustering (agrupamento baseado em representação) é uma técnica que utiliza representações centrais dos grupos para identificar a semelhança entre os dados. Algoritmos como *K-means* são exemplos dessa abordagem, onde os dados são agrupados em torno de centróides, minimizando a distância entre os pontos e suas respectivas representações centrais. Essa técnica é frequentemente utilizada em grandes conjuntos de dados devido à sua eficiência na organização dos mesmos.

Hierarchical Clustering (agregação hierárquica) organiza os dados em uma estrutura de árvore, permitindo a visualização da relação entre os grupos de forma hierárquica. Usando métodos aglomerativos ou divisivos, essa técnica cria um dendrograma que facilita a identificação de agrupamentos em múltiplos níveis de granularidade. Essa abordagem é útil quando se deseja explorar a estrutura dos dados em diferentes escalas. Agregação hierárquica para criar uma árvore de dendrograma, que pode ser usada, por exemplo, em estudos de biodiversidade para agrupar espécies similares. Um dendrograma é uma representação gráfica que ilustra a dis-

posição e a hierarquia de dados em um agrupamento, ou seja, ele mostra como os dados são organizados em clusters e como esses clusters estão relacionados entre si em diferentes níveis de similaridade. O eixo vertical do dendrograma representa a distância ou similaridade entre os pontos de dados, ou clusters. Quanto mais alto estiver um ponto no eixo vertical, mais distantes ou dessemelhantes eles são. O eixo horizontal do dendrograma representa usualmente os dados ou os agrupamentos. Os ramos do dendrograma conectam os pontos de dados que são agrupados juntos. Os ramos do dendocrama, a cada junção ou ramificação, indicam a fusão de clusters. A altura da junção representa a distância entre os clusters que estão sendo combinados. É possível "cortar" o dendrograma a uma altura específica para definir quantos clusters desejamos formar a partir dos dados originais.

Density-based Clustering (agrupamento baseado em densidade) é uma técnica que identifica grupos com base na densidade de pontos em uma determinada região do espaço de dados. Algoritmos como DBSCAN (*Density-Based Spatial Clustering of Applications with Noise* ou agrupamento espacial baseado em densidade de aplicações com ruído) são exemplos que ajudam a detectar agrupamentos que não dependam de uma forma pré-definida, permitindo encontrar grupos e também identifiquem discrepâncias ou ruídos. DBSCAN é frequentemente utilizado em geolocalização, onde grupos de pontos densos representam áreas urbanas.

Spectral and Graph Clustering (agrupamento espectral e em grafo) utiliza informações sobre a estrutura do grafo para agrupar dados. Essa técnica se baseia nas propriedades espectrais de uma matriz de similaridade e é útil em situações em que os dados podem ser representados como nós em um grafo. Essa abordagem é eficaz na detecção de estruturas complexas em conjuntos de dados interconectados. Uma matriz de similaridade é uma estrutura de dados que representa o grau de similaridade ou proximidade entre pares de elementos em um conjunto. Os valores na matriz de similaridade variam geralmente entre 0 e 1, onde valores mais altos indicam maior similaridade e valores mais baixos indicam menor similaridade. Em algumas aplicações, a matriz pode ser preenchida com distâncias, onde valores mais baixos indicam maior similaridade. A diagonal principal da matriz, que representa a similaridade de cada elemento consigo mesmo, geralmente tem todos os valores iguais a 1 (ou o valor máximo na escala de similaridade utilizada). O *Markov Clustering Algorithm* (MCL), algoritmo de clustering de Markov, é um método iterativo que combina etapas de expansão e inflação de matrizes. A expansão da matriz corresponde a elevar a matriz de transição a potências sucessivas, resultando em passeios aleatórios de maior comprimento. Por outro lado, a inflação da matriz aumenta a probabilidade de transições de maior probabilidade e reduz as de menor probabilidade. Como os nós no mesmo cluster tendem a ter pesos mais altos e, consequentemente, maiores

probabilidades de transição entre eles, o operador de inflação torna mais provável permanecer dentro do cluster, limitando assim a extensão do passeio aleatório. A técnica de agrupamento espectral é utilizada para segmentar imagens, onde cada *pixel* é tratado como um nó em um gráfico.

Clustering Validation (validação de agrupamento) é o processo de avaliação e verificação da qualidade dos agrupamentos obtidos a partir das técnicas de agrupamento. Métodos como o índice de Silhouette e a análise de variância entre e dentro dos grupos são utilizados para medir a coerência dos agrupamentos. A validação é importante para garantir que os resultados sejam interpretáveis e os grupos formados representem adequadamente a estrutura dos dados. Medidas de validação como o índice de Silhueta e o método do cotovelo (elbow method) para determinar a qualidade do agrupamento na segmentação de uma classificação pré-existente de pacientes (como grupos de risco definidos por especialistas). Isso ajuda a verificar se os clusters encontrados estão consoante as expectativas clínicas.

Probabilistic Classification (classificação probabilística) refere-se a métodos que atribuem probabilidades a cada classe com base nas características dos dados. Uma abordagem comum é o modelo *Naive Bayes*, que assume que as características são independentes entre si, permitindo calcular rapidamente a probabilidade de classes a partir de um conjunto de dados de treinamento. Essa técnica é frequentemente aplicada em tarefas como filtragem de spam e categorização de textos. Classificador Naive Bayes para categorizar e-mails como spam ou não spam com base em probabilidades condicionalmente independentes.

Decision Tree Classifier (classificador de árvore de decisão) utiliza uma estrutura de árvore para representar decisões e suas consequências. Cada nó da árvore representa uma característica, enquanto as folhas representam as classes finais. Essa técnica é amplamente utilizada em tarefas de classificação devido à sua interpretabilidade, contrariamente a redes neurais que não costumam oferecer rastreabilidade explicativa. A construção da árvore envolve a divisão dos dados com base em critérios de pureza, como o ganho de informação. O ganho de informação é um conceito utilizado em algoritmos de árvores de decisão, como o ID3, para construir a árvore que melhor expressa o conjunto de dados e mede a eficácia de um atributo em classificar os dados. Para entender o conceito de ganho de informação é importante conhecer a entropia, que é uma medida da incerteza ou impureza em um conjunto de dados. O ganho de informação é a diferença entre a entropia do conjunto original e a entropia média dos subconjuntos resultantes após a divisão dos dados por um atributo. Uma entropia alta indica que os dados estão misturados e, portanto, as classes estão mais incertas. O atributo com o maior ganho de informação é escolhido como o nó da árvore, pois significa que ele traz a maior redução na incerteza em relação à clas-

sificação dos dados. O processo se repete de forma recursiva para os subconjuntos criados até atingir um critério de parada (como uma profundidade máxima da árvore ou a pureza dos nós). Uma entropia baixa indica que os dados estão mais homogêneos, ou seja, as classes são mais certas. Uma árvore de decisão pode ser usada para prever a classificação de usuários do sistema de saúde, com base em fatores sociodemográficos como renda, idade, escolaridade e histórico de vacinação.

Linear Discriminant Analysis (LDA - análise discriminante linear) é uma técnica que projeta dados em um espaço de menor dimensão para maximizar a separação entre classes. A LDA busca encontrar uma combinação linear de características que melhor separa diferentes classes, sendo frequentemente aplicada em problemas de reconhecimento de padrões e classificação. A abordagem assume que as classes possuem distribuições normais com covariâncias iguais.Por exemplo, o uso de variáveis como idade, nível de colesterol, pressão arterial, Índice de Massa Corporal (IMC), nível de glicose no sangue e frequência cardíaca podem ser usadas para prever a ocorrência de doenças cardiovasculares. Neste exemplo, os dados dos pacientes são agrupados nas duas classes: presença ou ausência de doença cardíaca. Os dados devem ser bem separados nas classes e ter a suposição de normalidade. A LDA é aplicada aos dados para encontrar uma combinação linear das variáveis preditoras que maximiza a separação entre as duas classes. O modelo gerado cria uma função discriminante. O resultado da LDA pode ser um gráfico que mostra a projeção dos dados nos componentes discriminantes. O modelo fornece uma linha (ou hiperplano em dimensões superiores) que separa os grupos de “Doença Cardíaca” e “Sem Doença Cardíaca”. Os novos pacientes podem ser inseridos nesse modelo, e o valor resultante é utilizado para classificar o paciente em uma das duas categorias. Os profissionais da saúde podem então usar essa classificação, combinada com outros testes, para tomar decisões clínicas e na descoberta de quais variáveis (como colesterol ou pressão arterial) são mais discriminatórias para a presença de doenças cardíacas.

SVM são algoritmos de classificação que utilizam a ideia de encontrar um hiperplano que separa diferentes classes com a maior margem possível. A SVM é eficaz em conjuntos de dados com dimensões elevadas e pode ser aplicada em problemas não lineares através da utilização de kernels, o que possibilita transformar os dados em um espaço dimensional mais alto, facilitando a separação. Um hiperplano é uma dimensão a menos do que o espaço em que está localizado. Por exemplo, em um espaço 2 Dimensões (2D), um hiperplano é uma linha; em um espaço 3D, é um plano; em um espaço de n dimensões, um hiperplano tem dimensão (n-1). O objetivo do SVM é encontrar o hiperplano que melhor separa as classes de dados. O hiperplano divide o espaço em duas regiões — uma para cada classe.

O SVM busca o hiperplano que maximiza a margem entre as classes. A margem é a distância entre o hiperplano e os pontos mais próximos de qualquer classe (chamados de vetores de suporte). O SVM tenta maximizar essa margem, o que ajuda na generalização do modelo. Novos pontos de dados podem ser classificados com base em qual lado do hiperplano eles se encontram. Se um ponto estiver acima do hiperplano, ele é classificado em uma classe; se estiver abaixo, em outra classe. No espaço n-dimensional, quando os dados não são linearmente separáveis no espaço original, o SVM pode utilizar técnicas de mapeamento (*kernel trick*) para transformar os dados em um espaço n-dimensional maior, onde se pode encontrar um hiperplano que separe melhor as classes. Essa técnica é utilizada em diversas áreas, incluindo reconhecimento de imagem e bioinformática. Um exemplo significativo é a aplicação de SVM na detecção precoce do câncer, onde dados clínicos e imagens médicas (como mamografias) são analisados para classificar tumores como benignos ou malignos.

Classification Assessment (métricas de avaliação da classificação) consiste na avaliação do desempenho dos modelos de classificação por meio de métricas específicas. Ferramentas como matriz de confusão, acurácia, precisão e recall são utilizadas para medir a eficácia das classificações realizadas pelos modelos.

A matriz de confusão é uma tabela que permite visualizar o desempenho de um modelo de classificação em relação a classes reais e previstas. Ela contém quatro componentes principais. Verdadeiros Positivos (VP) são casos que foram corretamente classificados como positivos. Verdadeiros Negativos (VN) são casos que foram corretamente classificados como negativos. Falsos Positivos (FP) são casos que foram incorretamente classificados como positivos (também chamados de “erro tipo I”). Falsos Negativos (FN) são casos que foram incorretamente classificados como negativos (também chamados de “erro tipo II”). A partir da matriz de confusão, podemos derivar as métricas que qualificam o modelo e fazer a escolha do mais assertivo.

- A **acurácia** indica a proporção total de previsões corretas em relação ao total de previsões realizadas. Acurácia = (VP + VN) ÷ (VP + VN + FP + FN).
- A **sensibilidade**, também conhecida como taxa de verdadeiro positivo, revocação (*recall*), mede a capacidade do modelo de detectar casos positivos. É calculada como a razão entre o número de VP e a soma dos VP e falsos negativos (FN). Sensibilidade = (VP) ÷ (VP + FN).
- A **especificidade** mede a capacidade do modelo de identificar corretamente os negativos. É importante para avaliar a eficácia do modelo em não rotular falsamente casos que são, na verdade, negativos. Especificidade = (VN) ÷ (VN + FP).

» A **precisão** (ou *Positive Predictive Value*) mede a proporção de VP em relação ao total de casos que o modelo previu como positivos e reflete a confiança do modelo em suas previsões positivas. Precisão = (VP) ÷ (VP + FP).

» O ***F1-score*** é a média harmônica da precisão e da sensibilidade, oferecendo uma única métrica que captura ambas. É útil em contextos onde há uma necessidade de balancear *false positives* e *false negatives*. F1 = 2 × (Precisão × Sensibilidade) ÷ (Precisão + Sensibilidade).

» A **Curva ROC** (*Receiver Operating Characteristic*) é uma ferramenta gráfica que ilustra a capacidade de um modelo de classificação em diferenciar entre classes. Ela plota a taxa de VP (sensibilidade) contra a taxa de falsos positivos (1 - especificidade) em vários limiares de decisão. A AUC (*Area Under the Curve*) quantifica o desempenho do modelo ao calcular a área sob a curva ROC. Um valor de AUC próximo a 1 indica um modelo com excelente capacidade de classificação, enquanto um valor em torno de 0,5 sugere um desempenho semelhante ao aleatório.

» ***Kappa*** (ou *Kappa* de Cohen) é uma estatística utilizada para avaliar a concordância entre dois classificadores ou entre um classificador e um conjunto de rótulos de referência (*ground truth*). Ela é especialmente útil em problemas de classificação, pois vai além da simples acurácia, ajudando a entender como um modelo se comporta, especialmente em conjuntos de dados desbalanceados. A estatística *Kappa* mede a concordância entre classificadores, ajustando a acurácia levando em conta a possibilidade de que a concordância tenha ocorrido por acaso. O valor de *Kappa* varia de -1 a 1, onde *Kappa* = 1 ocorre a concordância perfeita entre os classificadores; *Kappa* = 0 significa concordância igual à que seria esperada por acaso (nenhuma concordância) e *Kappa* < 0 mostra concordância abaixo do esperado por acaso (indica que o modelo está classificando pior do que seria esperado). O coeficiente *Kappa* , ou *Kappa* de Cohen, é uma métrica que mede a concordância entre as previsões de um modelo e os valores reais, corrigida pelo acaso, sendo menos sensível ao desbalanceamento de classes do que a acurácia, que simplesmente calcula a proporção de previsões corretas sem considerar a distribuição das classes. *Kappa* = (Po - Pe) ÷ (1 - Pe), onde **Po** é a proporção de acordo observado (a acurácia do modelo) e **Pe** é a proporção de acordo esperado (calculado com base nas frequências das classes).

O aprendizado de máquina com técnicas de regressão é uma abordagem utilizada para modelar e prever relações entre variáveis contínuas. Os métodos incluem regressão linear, regressão polinomial e regressão de *Ridge*, entre outros, os quais aprendem e mapeiam as entradas (*features*) para uma variável de saída (*target*) por meio da identificação de padrões e relações nos dados. A regressão linear simples,

por exemplo, busca encontrar uma linha reta que melhor se ajusta aos pontos de dados, minimizando a soma dos erros quadráticos entre as previsões e os valores reais. A versatilidade da regressão a torna uma ferramenta valiosa em diversas aplicações, desde previsões econômicas até análise de tendências em ciências sociais e biológicas.

Overfitting (sobreajuste) ocorre quando um modelo se ajusta excessivamente aos dados de treinamento, capturando não apenas os padrões verdadeiros, mas também o ruído e as flutuações nos dados. Quando isso acontece, o modelo apresenta um desempenho excepcionalmente bom nas amostras de treinamento, mas sua capacidade de generalização para novos dados - ou seja, seu desempenho em um conjunto de teste ou em situações do mundo real - deteriora-se drasticamente.

O *overfitting* é frequentemente causado por modelos excessivamente complexos, com muitos parâmetros em relação ao número de observações disponíveis. Por exemplo, uma rede neural com muitas camadas e neurônios pode se ajustar tão bem aos dados de treinamento que termina aprendendo características específicas que não se aplicam a outras situações. Para mitigar o *overfitting*, pode-se usar técnicas como validação cruzada, regularização (como Lasso e Ridge, onde acrescenta uma penalização à função de custo que o modelo busca minimizar, restringindo a complexidade do modelo ao penalizar certos comportamentos, como a magnitude dos coeficientes dos parâmetros), simplificação do modelo, ou aumentar o conjunto de dados de treinamento. O objetivo é garantir que o modelo seja capaz de capturar as tendências gerais dos dados, em vez de se tornar uma representação precisa dos dados de treinamento.

Linear Regression (regressão linear) é uma técnica que modela a relação entre uma variável dependente e uma ou mais variáveis independentes, assumindo que essa relação é linear. O objetivo é encontrar a melhor linha reta que minimize a soma dos erros quadráticos entre os valores previstos e os valores reais. Essa técnica é amplamente utilizada em econometria e ciências sociais para realizar previsões e inferências sobre dados contínuos. Uso de regressão linear para prever vendas com base em investimento em publicidade, modelando a relação entre as duas variáveis.

Logistic Regression (regressão logística) é uma técnica que aplica a função logística para modelar a probabilidade de uma classe binária. Embora seu nome inclua “regressão”, a *logistic regression* é utilizada para problemas de classificação e não para predições contínuas. A técnica quantifica a relação entre uma variável depen-

dente categórica e variáveis independentes, permitindo a estimativa da probabilidade de um evento ocorrer. Regressão logística pode prever a probabilidade de um paciente ter uma doença com base em características como idade, IMC e pressão arterial.

Neural Networks (redes neurais) consistem em camadas de nós interconectados que processam informações por meio de funções matemáticas. As redes neurais são capazes de aprender padrões complexos a partir dos dados de entrada, ajustando os pesos das conexões durante o treinamento. Essa técnica é aplicada em diversas áreas, incluindo Processamento de Linguagem Natural (PLN), reconhecimento de imagens e sistemas de recomendação.

Deep Learning (aprendizado profundo) é uma subárea do aprendizado de máquina que utiliza redes neurais profundas, compostas por várias camadas ocultas, permitindo a modelagem de representações de alto nível dos dados, onde ocorre o aprendizado e extração de características ou padrões complexos e abstratos a partir dos dados brutos. Essa técnica tem se mostrado eficaz em tarefas que envolvem grandes volumes de dados e complexidade, como reconhecimento de fala e geração de imagens. Redes neurais profundas consistem em várias camadas de neurônios, onde cada camada transforma a entrada de maneira sucessiva. Camadas iniciais podem aprender características mais simples, como bordas em imagens, enquanto camadas mais profundas podem combinar essas características para identificar formas, objetos e, eventualmente, conceitos mais complexos, como rostos ou cenas.

Ao contrário dos algoritmos tradicionais de aprendizado de máquina, que frequentemente exigem engenharia de recursos manual para extrair características relevantes dos dados, as redes neurais profundas podem automaticamente aprender as representações mais relevantes para o problema. Em muitos casos, as representações apreendidas pelas redes podem ser organizadas hierarquicamente. Por exemplo, em PLN, palavras podem ser combinadas para formar frases e, em níveis superiores, essas frases podem formar significados contextuais mais complexos. Similarmente, em visão computacional, características de baixo nível (como *pixel*) são combinadas em características de nível superior (como formas e, eventualmente, objetos inteiros). Modelos que aprendem representações de alto nível tendem a generalizar melhor em novos dados, pois capturam padrões subjacentes que são aplicáveis a situações não vistas. Modelos de rede neural profunda, como *Long Short-Term Memory* (LSTM), para prever séries temporais, como a demandas ao longo do tempo.

Regression Evaluation (avaliação da regressão) refere-se ao processo de medir o desempenho de modelos de regressão com métricas, visando a identificação de como o modelo se ajusta aos dados e a comparação entre diferentes abordagens de

modelagem. A análise pode incluir a avaliação do Erro Quadrático Médio (MSE), Erro Absoluto Médio (MAE) e R^2 (Coeficiente de Correlação), que indicam a precisão das previsões.

Vale a pena ressaltar o papel pré-tratamento e da modelagem de dados com a normalização. No contexto de bancos de dados, a normalização refere-se ao processo de estruturar os dados para reduzir a redundância e melhorar a integridade, organizando-os em tabelas relacionadas por meio de chaves primárias e estrangeiras, geralmente seguindo formas normais. Em aprendizado de máquina, a normalização se refere à transformação dos dados de entrada para que suas escalas sejam consistentes, visando melhorar a desempenho de algoritmos que utilizam medidas de distância ou que são sensíveis à escala. Enquanto a normalização em bancos de dados foca na eficiência e integridade da estrutura dos dados, a normalização no aprendizado de máquina visa otimizar o processo de aprendizado e previsão, assegurando que as variáveis contribuem de maneira equilibrada para o modelo.

Muitos algoritmos de aprendizado de máquina, como regressão logística, k-vizinhos mais próximos, SVM e redes neurais, são sensíveis à escala dos dados. Se as variáveis de entrada tiveram escalas muito diferentes, isso pode levar a problemas de convergência e a resultados não satisfatórios. Quando variáveis possuem escalas várias ordens de magnitude diferentes, as variáveis com maior escala podem dominar o processo de aprendizado, tornando difícil para o algoritmo aprender a importância relativa das variáveis de menor escala.

Vamos destacar três métodos de normalização. *Min-Max Scaling* (escalonamento Min-Max) reescala as características para um intervalo específico, geralmente [0, 1]. *Z-Score Normalization* (Normalização Z-Score, padronização) transforma os dados para que eles tenham uma média de 0 e um desvio padrão de 1. *Robust Scaling* (escalonamento robusto) é útil em casos onde há *outliers*. Em vez de usar a média e o desvio padrão, ele usa a mediana e o Intervalo Interquartil (IQR).

A validação é a etapa final e igualmente importante no ciclo de aprendizado de máquina. Ela assegura que o modelo treinado seja generalizável e eficaz em dados não vistos. Validação particionada, como *K-fold Cross-Validation*, envolve dividir o conjunto de dados em K partes e usar K-1 partes para treinamento e a parte restante para validação, repetindo o processo K vezes, obtendo-se uma estimativa robusta do desempenho do modelo. Outra técnica é o *Bootstrap*, que envolve gerar várias amostras aleatórias com reposição do conjunto de dados original, treinando e validando o modelo em cada amostra. As amostras são usadas para estimar a variabilidade do modelo e sua capacidade de generalização.

4.1.2 Inteligência Artificial

IA e aprendizado de máquina são conceitos inter-relacionados, mas distintos. A IA refere-se ao campo mais amplo da ciência da computação que busca criar sistemas capazes de realizar tarefas que normalmente requerem inteligência humana, como raciocínio, compreensão de linguagem natural, percepção visual e tomada de decisão. Dentro desse domínio da IA está o aprendizado de máquina, porém, com foco no desenvolvimento de algoritmos e técnicas que permitem que os sistemas aprendam a partir de dados, melhorando seu desempenho com o tempo sem serem explicitamente programados.

No contexto do processamento de texto, uma das abordagens mais difundidas é o **PLN**. O PLN permite que computadores compreendam, interpretam e mimetizam a linguagem humana. Incluindo a análise de sentimentos, extração de entidades, tradução automática, sumarização de textos e resposta a perguntas. Técnicas como a análise sintática e semântica, juntamente com modelos de aprendizado profundo, como Redes Neurais Recorrentes (RNNs) e transformadores, são fundamentais para o PLN. Os **transformadores** utilizam um mecanismo de atenção, que permite que o modelo foque em diferentes partes de uma sequência de entrada ao gerar uma saída, em vez de processar a informação de forma sequencial, permitindo que transformadores lidem com longas sequências de texto de maneira mais eficaz do que arquiteturas anteriores, como as RNNs e LSTMs.

O aprendizado é utilizado no processamento de texto através do uso de ***embeddings*** de palavras, que são representações vetoriais de palavras que capturam suas relações semânticas. Modelos como Word2Vec e GloVe foram pioneiros nessa área, permitindo que palavras com significados semelhantes tivessem representações vetoriais próximas. Essa representação densa e contínua de palavras facilita a tarefa de encontrar similaridades semânticas e realizar análises mais sofisticadas em textos.

Quando se trata do processamento de imagens, as **Redes Neurais Convolucionais (CNNs)** se destacam, pois são projetadas para reconhecer padrões e características em imagens, tornando-as eficazes para tarefas como reconhecimento de objetos, detecção de rostos, segmentação de imagens e diagnóstico médico. Essas redes utilizam camadas convolucionais que aplicam filtros a pequenas regiões da imagem, capturando características como bordas, texturas e formas. À medida que as imagens passam por múltiplas camadas convolucionais, a rede aprende a identificar padrões cada vez mais complexos e abstratos.

O **aprendizado profundo** em imagens também se beneficia de técnicas como a transferência de estilo neural, onde as características de estilo de uma imagem podem ser transferidas para outra, criando obras de arte digitais que combinam o conteúdo de uma imagem com o estilo de outra. Além disso, Redes Adversárias Generativas (GANs) têm uso na geração de imagens realistas a partir de entradas aleatórias, assim como na melhora de imagens de baixa resolução ou na criação de novas imagens a partir de descrições textuais.

A combinação de IA para texto e imagens abre novas possibilidades, como a descrição automática de imagens, onde modelos de PLN são usados para gerar legendas descritivas para imagens. Esse tipo de aplicação é particularmente útil para a acessibilidade, permitindo que pessoas com deficiências visuais obtenham descrições detalhadas de imagens e vídeos. Além disso, o reconhecimento de texto em imagens (Reconhecimento Ótico de Caracteres - OCR) utiliza técnicas de IA para converter texto manuscrito ou impresso em dados editáveis, facilitando a digitalização e arquivamento de documentos.

4.2 Extração de Informações a Partir de Dados Tabulares

A extração de informações a partir de textos clínicos transforma dados não estruturados em conhecimento útil e não trivial. Este processo envolve o uso de técnicas de aprendizado de máquina, terminologias e padrões que garantem a precisão, consistência e interoperabilidade dos dados extraídos.

Vamos mostrar um exemplo do algoritmo k-means, utilizando os dados da tabela fato_prescricoes, passo a passo:

- Extrair os dados da tabela SQL usando Python com biblioteca para conectar ao banco de dados e extrair os dados.
- Pré-processar os dados com limpeza e transformação dos dados conforme necessário.
- Executar o algoritmo K-means usando a biblioteca *scikit-learn*.
- Analisar os resultados ao visualizar e interpretar os *clusters* formados.

Vamos popular a tabela fato_prescricao com dados fictícios.

```python
# python

import sqlite3
import pandas as pd

# Conectar ao banco de dados SQLite (ou criar um novo banco)
conn = sqlite3.connect(':memory:')  # Usa um banco
de dados em memória para simplicidade
cursor = conn.cursor()

# Criação da tabela fato _ prescricoes
cursor.execute('''
CREATE TABLE fato _ prescricoes (
    id _ prescricao INTEGER PRIMARY KEY,
    id _ profissional INTEGER,
    id _ paciente INTEGER,
    id _ medicamento INTEGER,
    id _ estabelecimento INTEGER,
    data _ nascimento _ paciente TEXT,
    sexo _ paciente TEXT,
    endereco _ paciente TEXT,
    data _ prescricao TEXT,
    posologia TEXT
)
''')

# Inserir dados de exemplo
prescricoes = [
    (1, 1, 1, 1, 1, '1980-01-01', 'M', 'Rua A, Municipio
A, Estado A', '2024-01-01', '1x ao dia'),
    (2, 2, 2, 2, 2, '1990-02-02', 'F', 'Rua B, Municipio
B, Estado B', '2024-02-01', '2x ao dia'),
    (3, 3, 3, 3, 3, '2000-03-03', 'M', 'Rua C, Municipio
C, Estado C', '2024-03-01', '3x ao dia'),
    (4, 4, 4, 4, 4, '1985-04-04', 'F', 'Rua D, Municipio
```

continua

```
D, Estado D', '2024-04-01', '4x ao dia'),
    (5, 5, 5, 5, 5, '1995-05-05', 'M', 'Rua E, Municipio
E, Estado E', '2024-05-01', '1x ao dia'),
    # Adicione mais dados conforme necessário
]

cursor.executemany('''
INSERT INTO fato_prescricoes (id_prescricao, id_profissional,
id_paciente, id_medicamento, id_estabelecimento,
                              data_nascimento_paciente, sexo_
paciente, endereco_paciente, data_prescricao, posologia)
VALUES (?, ?, ?, ?, ?, ?, ?, ?, ?, ?)
''', prescricoes)

conn.commit()
```

Passo 1: Extrair os dados da Tabela SQL.

```
# python

# Consulta SQL para extrair os dados
query = '''
SELECT
    id_prescricao,
    id_profissional,
    id_paciente,
    id_medicamento,
    id_estabelecimento,
    data_nascimento_paciente,
    sexo_paciente,
    endereco_paciente,
    data_prescricao
FROM fato_prescricoes
'''
```

continua

```
# Extrair os dados para um DataFrame
df = pd.read_sql(query, con=conn)

# Exibir os dados extraídos
print(df)
```

Passo 2: Pré-processar os dados.

Para o K-means, precisamos garantir que os dados estejam em um formato numérico ao converter as variáveis categóricas em variáveis *dummy* (*one-hot encoding*) e normalizar os dados. Assim, uma coluna com *strings* (cadeias) é transformada em várias colunas, onde cada coluna representa uma categoria possível.

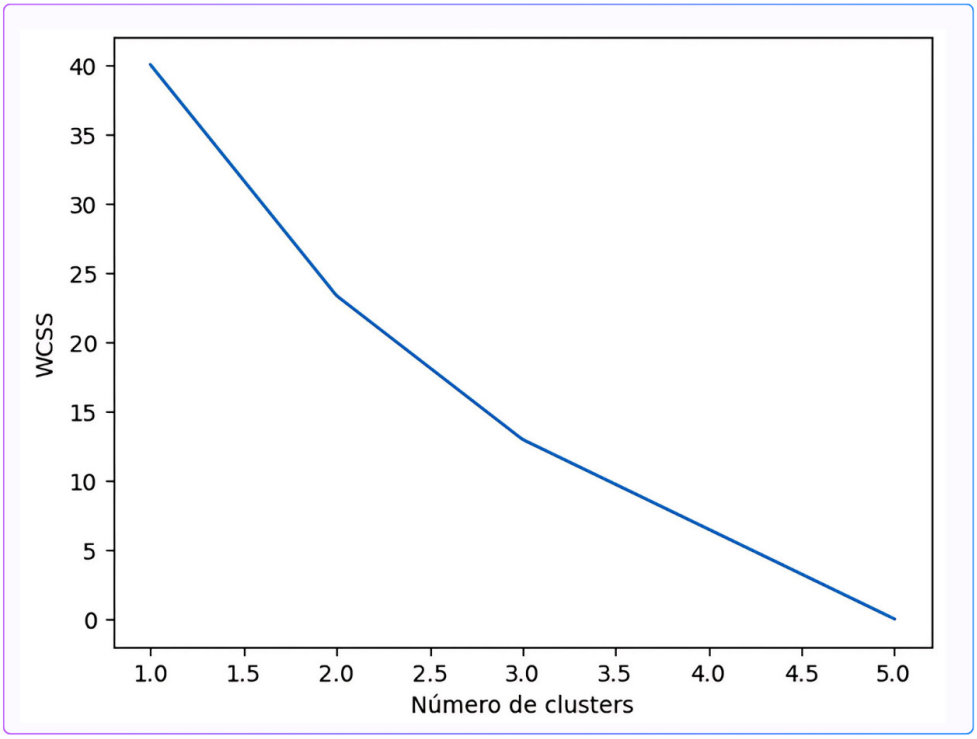
```
# python

# Convertendo variáveis categóricas em dummy
df['idade_paciente'] = (pd.to_datetime('today') - pd.to_
datetime(df['data_nascimento_paciente'])).dt.days / 365.25
df = pd.get_dummies(df, columns=['sexo_
paciente', 'endereco_paciente'])

# Remover colunas não numéricas ou irrelevantes
df.drop(columns=['id_prescricao', 'id_profissional', 'id_
paciente', 'id_medicamento', 'id_estabelecimento', 'data_
prescricao', 'data_nascimento_paciente'], inplace=True)

# Normalizar os dados
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
df_scaled = scaler.fit_transform(df)
```

Passo 3: executar o algoritmo K-means.

Vamos usar a biblioteca *scikit-learn*, biblioteca de código aberto para a linguagem de Python, para aplicar o algoritmo K-means e determinar os *clusters* (Figura 42).

```
# python

from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# Determinando o número ideal de clusters usando o método Elbow
wcss = []  # Within-cluster sum of squares
for i in range(1, 6): # Change the upper
limit to 6 (number of samples + 1)
    kmeans = KMeans(n _ clusters=i, init='k-means++',
max _ iter=300, n _ init=10, random _ state=0)
    kmeans.fit(df _ scaled)
    wcss.append(kmeans.inertia _ )

# Plotando o método Elbow
plt.plot(range(1, 6), wcss) # Change the x-axis range to match the loop
plt.title('Método Elbow')
plt.xlabel('Número de clusters')
plt.ylabel('WCSS')
plt.show()

# Aplicando K-means com o número ideal de clusters
n _ clusters = 3  # Supondo que 3 é o número
ideal após análise do gráfico Elbow
kmeans = KMeans(n _ clusters=n _ clusters, init='k-
means++', max _ iter=300, n _ init=10, random _ state=0)
clusters = kmeans.fit _ predict(df _ scaled)

# Adicionando os clusters ao DataFrame original
df['cluster'] = clusters
```

Figura 42 - Método *Elbow*



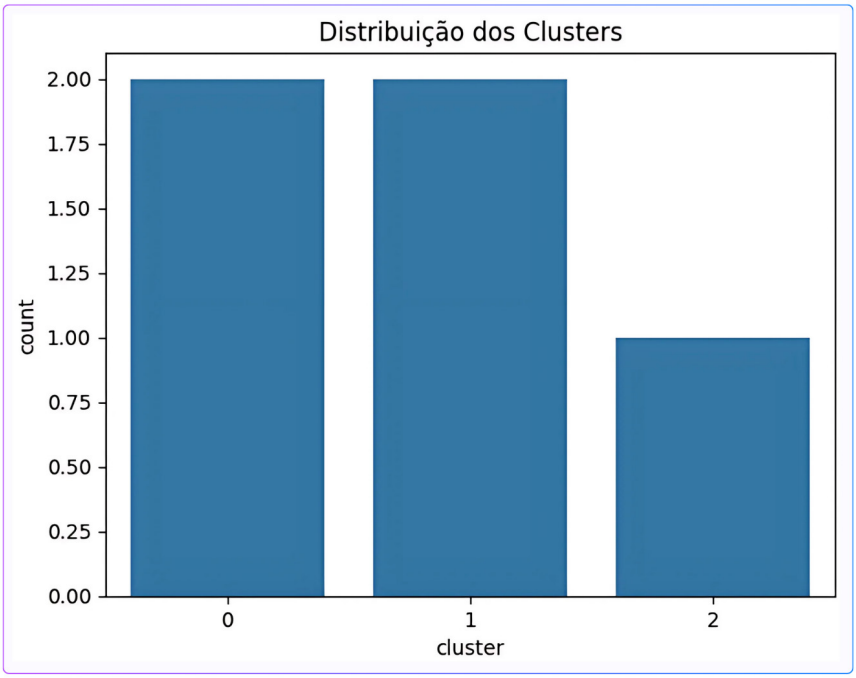
Fonte: autoria própria.

Passo 4: Analisar os resultados.

Agora que temos os *clusters* (Figuras 43-44), podemos analisar os resultados.

```
# python

# Visualizando os clusters
import seaborn as sns
# Plotando a distribuição dos clusters
sns.countplot(x='cluster', data=df)
plt.title('Distribuição dos Clusters')
plt.show()
# Analisando as características de cada cluster
cluster _ analysis = df.groupby('cluster').mean()
print(cluster _ analysis)
```

Figura 43 - Distribuição dos *clusters*

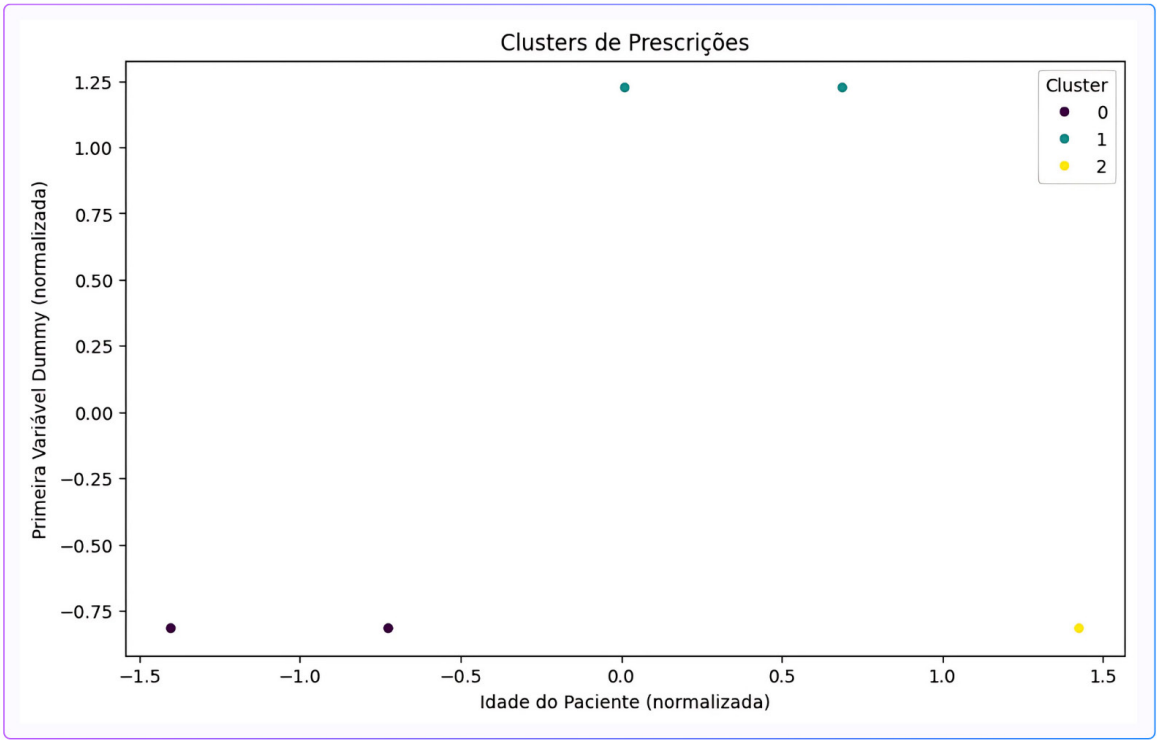


Fonte: autoria própria.

```
# python

# Visualizando os clusters
import seaborn as sns

# Selecionar duas variáveis para visualização (por
exemplo, idade _ paciente e uma das variáveis dummy)
plt.figure(figsize=(10, 6))
sns.scatterplot(x=df _ scaled[:, 0], y=df _ scaled[:,
1], hue=df['cluster'], palette='viridis')
plt.title('Clusters de Prescrições')
plt.xlabel('Idade do Paciente (normalizada)')
plt.ylabel('Primeira Variável Dummy (normalizada)')
plt.legend(title='Cluster')
plt.show()

# Outra opção para visualização mais clara se houver mais dimensões
from sklearn.decomposition import PCA

# Reduzir os dados a 2 dimensões para plotagem
```

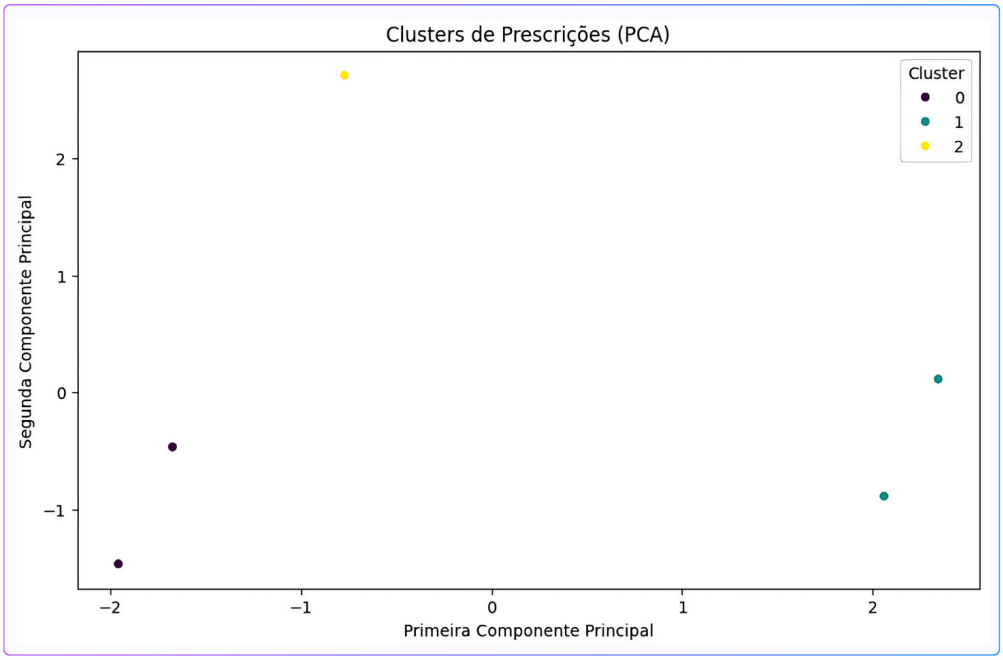
```
pca = PCA(n _ components=2)
principal _ components = pca.fit _ transform(df _ scaled)

# DataFrame com as componentes principais e os clusters
df _ pca = pd.DataFrame(data=principal _
components, columns=['PC1', 'PC2'])
df _ pca['cluster'] = df['cluster']

# Plotagem
plt.figure(figsize=(10, 6))
sns.scatterplot(x='PC1', y='PC2', hue='cluster',
data=df _ pca, palette='viridis')
plt.title('Clusters de Prescrições (PCA)')
plt.xlabel('Primeira Componente Principal')
plt.ylabel('Segunda Componente Principal')
plt.legend(title='Cluster')
plt.show()
```

Figura 44 - *Clusters* de prescrições



Fonte: autoria própria.

Figura 45 - *Clusters* de prescrições (PCA)



Fonte: autoria própria.

4.3 Extração de Informações a Partir de Textos Clínicos

Uma das primeiras etapas no PLN é a formação de um vocabulário ou dicionário a partir de um corpus, que é um conjunto extenso de textos. Um *corpus* (ou corpora no plural) é uma coleção de documentos utilizada para treinar modelos de NLP. A partir desse corpus, cria-se um vocabulário constituído pelas palavras ou *tokens* únicos presentes nos textos. No desenvolvimento de modelos de NLP, diversos métodos podem ser utilizados para representar e analisar textos. Um desses métodos é o *Concept-Based Method* (CBM), que utiliza conceitos em vez de palavras isoladas para representar textos, ajudando a captar o significado relacionado aos textos. Outro método é o *Phrase-Based Method* (PBM), que se concentra na identificação e utilização de frases em lugar de palavras individuais. As frases, por capturarem contextos maiores do que palavras isoladas, podem oferecer uma representação mais completa e precisa dos textos.

O *Pattern Taxonomy Method* (PTM) organiza padrões de texto em uma taxonomia, auxiliando na identificação e categorização de padrões frequentes e relevantes nos textos, o que é útil em tarefas de classificação e extração de informações.

Vamos exemplificar as tarefas de PLN com um *corpus* de casos clínicos fictícios, onde cada linha refere-se ao registro de um paciente.

```
# python
corpus = [
    "Paciente masculino de 45 anos com histórico de
hipertensão. Relata dor no peito há 3 dias.",
    "Paciente feminina de 30 anos, sem comorbidades.
Apresenta febre alta e tosse persistente.",
    "Paciente masculino de 60 anos com diabetes tipo 2.
Queixa-se de visão turva e fadiga constante.",
    "Paciente feminina de 25 anos, atleta. Sofreu
entorse no tornozelo durante treino.",
    "Paciente masculino de 50 anos, fumante. Apresenta
dificuldade para respirar e tosse crônica."
]
```

A **tokenização** é o processo que divide o texto em unidades menores chamadas *tokens*, que podem ser palavras, frases ou outros elementos. A tokenização facilita o processamento subsequente ao permitir que os textos sejam tratados em partes menores e mais manejáveis.

```
# python Tokenização
from nltk.tokenize import word_tokenize
import nltk

# Download the 'punkt' resource for tokenization
nltk.download('punkt')

# Tokenização de cada documento no corpus
tokens = [word_tokenize(doc.lower()) for doc in corpus]
print(tokens)
```

[['paciente', 'masculino', 'de', '45', 'anos', 'com', 'histórico', 'de', 'hipertensão', '.', 'relata', 'dor', 'no', 'peito', 'há', '3', 'dias', '.'],

['paciente', 'feminina', 'de', '30', 'anos', ',', 'sem', 'comorbidades', '.', 'apresenta', 'febre', 'alta', 'e', 'tosse', 'persistente', '.'],

['paciente', 'masculino', 'de', '60', 'anos', 'com', 'diabetes', 'tipo', '2', '.', 'queixa-se', 'de',

‘visão’, ‘turva’, ‘e’, ‘fadiga’, ‘constante’, ‘.’],

[‘paciente’, ‘feminina’, ‘de’, ‘25’, ‘anos’, ‘,’, ‘atleta’, ‘.’, ‘sofreu’, ‘entorse’, ‘no’, ‘tornozelo’, ‘durante’, ‘treino’, ‘.’],

[‘paciente’, ‘masculino’, ‘de’, ‘50’, ‘anos’, ‘,’, ‘fumante’, ‘.’, ‘apresenta’, ‘dificuldade’, ‘para’, ‘respirar’, ‘e’, ‘tosse’, ‘crônica’, ‘.’]]

Stemming e lematização são técnicas para reduzir palavras às suas formas-base ou radicais. O ***stemming*** remove sufixos para obter a raiz da palavra, enquanto a **lematização** considera o contexto e reduz as palavras à sua forma base. Ambas as técnicas são usadas para normalizar os textos e diminuir a dimensionalidade dos dados.

```
# python Stemming
from nltk.stem import PorterStemmer


stemmer = PorterStemmer()


# Aplicando stemming a cada token
stemmed _ tokens = [[stemmer.stem(token) for
token in doc] for doc in tokens]
print(stemmed _ tokens)
```

[[‘paci’, ‘masculino’, ‘de’, ‘45’, ‘ano’, ‘com’, ‘históric’, ‘de’, ‘hipertens’, ‘.’, ‘relata’, ‘dor’, ‘no’, ‘peito’, ‘há’, ‘3’, ‘dia’, ‘.’],

[‘paci’, ‘feminina’, ‘de’, ‘30’, ‘ano’, ‘,’, ‘sem’, ‘comorbid’, ‘.’, ‘apresenta’, ‘febr’, ‘alta’, ‘e’, ‘toss’, ‘persist’, ‘.’],

[‘paci’, ‘masculino’, ‘de’, ‘60’, ‘ano’, ‘com’, ‘diabet’, ‘tipo’, ‘2’, ‘.’, ‘queixa-s’, ‘de’, ‘visão’, ‘turv’, ‘e’, ‘fadig’, ‘constant’, ‘.’],

[‘paci’, ‘feminina’, ‘de’, ‘25’, ‘ano’, ‘,’, ‘atlet’, ‘.’, ‘sofreu’, ‘entors’, ‘no’, ‘tornozel’, ‘durant’, ‘treino’, ‘.’],

[‘paci’, ‘masculino’, ‘de’, ‘50’, ‘ano’, ‘,’, ‘fumant’, ‘.’, ‘apresenta’, ‘dificuldad’, ‘para’, ‘respira’, ‘e’, ‘toss’, ‘crônic’, ‘.’]]

```
# python Lematização
from nltk.stem import WordNetLemmatizer
import nltk

# Download the 'wordnet' resource for lemmatization
nltk.download('wordnet')

lemmatizer = WordNetLemmatizer()

# Aplicando lematização a cada token
lemmatized _ tokens = [[lemmatizer.lemmatize(token)
for token in doc] for doc in tokens]
print(lemmatized _ tokens)
```

[[‘paciente’, ‘masculino’, ‘de’, ‘45’, ‘ano’, ‘com’, ‘histórico’, ‘de’, ‘hipertensão’, ‘.’, ‘relata’, ‘dor’, ‘no’, ‘peito’, ‘há’, ‘3’, ‘dia’, ‘.’],

[‘paciente’, ‘feminina’, ‘de’, ‘30’, ‘ano’, ‘,’, ‘sem’, ‘comorbidade’, ‘.’, ‘apresenta’, ‘febre’, ‘alta’, ‘e’, ‘tosse’, ‘persistente’, ‘.’],

[‘paciente’, ‘masculino’, ‘de’, ‘60’, ‘ano’, ‘com’, ‘diabetes’, ‘tipo’, ‘2’, ‘.’, ‘queixa-se’, ‘de’, ‘visão’, ‘turva’, ‘e’, ‘fadiga’, ‘constante’, ‘.’],

[‘paciente’, ‘feminina’, ‘de’, ‘25’, ‘ano’, ‘,’, ‘atleta’, ‘.’, ‘sofreu’, ‘entorse’, ‘no’, ‘tornozelo’, ‘durante’, ‘treino’, ‘.’],

[‘paciente’, ‘masculino’, ‘de’, ‘50’, ‘ano’, ‘,’, ‘fumante’, ‘.’, ‘apresenta’, ‘dificuldade’, ‘para’, ‘respirar’, ‘e’, ‘tosse’, ‘crônica’, ‘.’]]

Stopwords são palavras comuns, como artigos e preposições, frequentemente eliminadas dos textos durante o processamento, pois geralmente não trazem informações relevantes para a análise.

```
!pip install nltk
import nltk
nltk.download('stopwords')
```

continua

```
# python Stopwords
from nltk.corpus import stopwords
import nltk

# Download the 'stopwords' resource for Portuguese stopwords
nltk.download('stopwords')

# Obter a lista de stopwords em português
stop _ words = set(stopwords.words('portuguese'))

# Remover stopwords de cada documento
filtered _ tokens = [[token for token in doc if token
not in stop _ words] for doc in tokens]
print(filtered _ tokens)
```

[['paciente', 'masculino', '45', 'anos', 'histórico', 'hipertensão', '.', 'relata', 'dor', 'peito', 'há', '3', 'dias', '.'],

['paciente', 'feminina', '30', 'anos', ',', 'comorbidades', '.', 'apresenta', 'febre', 'alta', 'tosse', 'persistente', '.'],

['paciente', 'masculino', '60', 'anos', 'diabetes', 'tipo', '2', '.', 'queixa-se', 'visão', 'turva', 'fadiga', 'constante', '.'],

['paciente', 'feminina', '25', 'anos', ',', 'atleta', '.', 'sofreu', 'entorse', 'tornozelo', 'treino', '.'],

['paciente', 'masculino', '50', 'anos', ',', 'fumante', '.', 'apresenta', 'dificuldade', 'respirar', 'tosse', 'crônica', '.']]

Vamos fazer um exemplo de análise de semântica. Para realizar a extração de entidades e uma análise semântica usando spaCy, primeiro é necessário configurar o ambiente e carregar um modelo de linguagem. A seguir, ocorre a extração de entidades nomeadas (Named Entity Recognition - NER) e um exemplo de similaridade semântica. A configuração do ambiente requer a instalação do *spaCy* e o modelo de linguagem.

```
!pip install spacy
!python -m spacy download pt _ core _ news _ sm
```

A extração de Entidades usando *spaCy* em português precede a extração de entidades nomeadas dos casos clínicos fictícios.

```
#python

import spacy

# Carregar o modelo de linguagem em português
nlp = spacy.load('pt _ core _ news _ sm')

# Corpus com casos clínicos fictícios
corpus = [
    "Paciente masculino de 45 anos com histórico de
hipertensão. Relata dor no peito há 3 dias.",
    "Paciente feminina de 30 anos, sem comorbidades.
Apresenta febre alta e tosse persistente.",
    "Paciente masculino de 60 anos com diabetes tipo 2.
Queixa-se de visão turva e fadiga constante.",
    "Paciente feminina de 25 anos, atleta. Sofreu
entorse no tornozelo durante treino.",
    "Paciente masculino de 50 anos, fumante. Apresenta
dificuldade para respirar e tosse crônica."
]

# Analisar cada documento no corpus
for doc in corpus:
    doc _ nlp = nlp(doc)
    print(f"Texto: {doc}")
    for ent in doc _ nlp.ents:
        print(f" - Entidade: {ent.text}, Tipo: {ent.label _ }")
    print("\n")
```

Na análise semântica (similaridade semântica), calcularemos a similaridade entre frases ou documentos usando vetores de palavras gerados pelo modelo *spaCy*.

```
#python

# Exemplo de frases para comparação
frase1 = "Paciente apresenta dor no peito e dificuldade para respirar."
frase2 = "Paciente relata dificuldade respiratória e dor torácica."

# Processar as frases com o modelo spaCy
doc1 = nlp(frase1)
doc2 = nlp(frase2)

# Calcular a similaridade semântica entre as duas frases
similaridade = doc1.similarity(doc2)

print(f"Similaridade entre as frases: {similaridade:.2f}")
# Similaridade entre as frases: 0.64
```

Para calcular a similaridade semântica entre todas as combinações de itens do corpus de casos clínicos fictícios, podemos usar o produto cartesiano, isto é, a combinação de todos contra todos, no caso, será calculada a similaridade entre cada par de documentos para exibir os resultados (Figura 46). A qualidade do modelo é avaliada se, a matriz de similaridade expressam casos clínicos mais semelhantes, quanto mais próximos de 1 for o valor. No exemplo em questão, é necessário ampliar o conteúdo de cada caso clínico e semiestruturar os dados, pré-agrupando por especialidade médica, por exemplo.

```
from itertools import product # Import the
product function from itertools
import pandas as pd

# Processar os documentos com o modelo spaCy
docs = [nlp(doc) for doc in corpus]

# Calcular a similaridade entre todos os pares de documentos
similaridades = []
for (i, doc1), (j, doc2) in product(enumerate(docs), repeat=2):
```

continua

```
    similaridade = doc1.similarity(doc2)
    similaridades.append((i, j, similaridade))

# Exibir as similaridades
for i, j, similaridade in similaridades:
    print(f"Similaridade entre documento {i+1} e
documento {j+1}: {similaridade:.2f}")

# Criar uma matriz de similaridade
similarity_matrix = pd.DataFrame(index=[f'Doc{i+1}'
for i in range(len(docs))],
                                 columns=[f'Doc{j+1}'
for j in range(len(docs))])

# Preencher a matriz com os valores de similaridade
for i, j, similaridade in similaridades:
    similarity_matrix.iloc[i, j] = similaridade

print(similarity_matrix)
```

Figura 46 - Exemplo de similaridade semântica

Similaridade	Doc1	Doc2	Doc3	Doc4	Doc5
Doc1	1,0000	0,7924	0,8922	0,8628	0,8053
Doc2	0,7924	1,0000	0,8358	0,8342	0,8982
Doc3	0,8922	0,8358	1,0000	0,8088	0,8936
Doc4	0,8628	0,8342	0,8088	1,0000	0,8262
Doc5	0,8053	0,8982	0,8936	0,8262	1,0000

Fonte: autoria própria.

O modelo *Bag of Words* (BoW) é uma abordagem para representar textos, tratando um texto como uma coleção de palavras e ignorando a ordem e a estrutura gramatical, mas mantendo a frequência das palavras. Apesar de sua simplicidade, o BoW pode ser eficaz para várias tarefas de NLP, como classificação de textos. Vamos

pré-processar o texto para criar o modelo BoW. O pré-processamento inclui a tokenização e a remoção de stopwords. Vamos criar o modelo BoW usando a biblioteca *scikit-learn*.

```
from sklearn.feature _ extraction.text import CountVectorizer
import pandas as pd

# Inicializar o CountVectorizer
stop _ words _ pt = ['de', 'a', 'o', 'que', 'e', 'do', 'da', 'em',
'um', 'para', ...]  # Add more stop words as needed
vectorizer = CountVectorizer(stop _ words=stop _ words _ pt)

# Ajustar o modelo e transformar o corpus
X = vectorizer.fit _ transform(corpus)

# Obter o vocabulário
vocabulario = vectorizer.get _ feature _ names _ out()

# Criar um DataFrame com os resultados
df _ bow = pd.DataFrame(X.toarray(), columns=vocabulario)

print(df _ bow)
```

O código acima irá gerar uma matriz onde cada linha representa um documento e cada coluna representa um termo do vocabulário. Cada valor na matriz indica a frequência do termo no documento correspondente.

Aqui está um exemplo de como a matriz de *Bag of Words* pode ser exibida (Figura 46).

Figura 47 - Exemplo de matriz de *bag of words*

	Doc1	Doc2	Doc3	Doc4	Doc5
anos	1	1	1	1	1
atleta	0	0	0	1	0
com	1	1	1	1	1
dificuldade	0	0	0	0	1
dor	1	0	0	0	0
durante	0	0	0	1	0
e	1	1	1	1	1
entorse	0	0	0	1	0
feminina	0	1	0	0	0
febre	0	1	0	0	0
histórico	1	0	0	0	0
hipertensão	1	0	0	0	0
no	1	0	0	0	0
peito	1	0	0	0	0
paciente	1	1	1	1	1
para	0	0	0	0	1
respiratória	0	0	0	0	1
tosse	1	1	0	0	1
tipo	0	0	1	0	0
visão	0	0	1	0	0
anos	0	1	1	0	0
constante	0	1	1	0	1

Fonte: autoria própria.

O Reconhecimento de Entidades Nomeadas NER é uma técnica de NLP que identifica e classifica palavras ou frases em categorias predefinidas, como nomes de pessoas, organizações, localizações, datas e termos médicos. Usada para a identificação de entidades de saúde específicas, como nomes de medicamentos, condições de saúde e procedimentos clínicos (Gomes *et al.*, 2019; Silva; Pollettini; Pazin Filho, 2023).

A análise de sentimento e opinião é outra técnica de análise de textos para determinar a atitude emocional do escritor em relação a um tópico específico, realizando a avaliação do retorno de pacientes sobre tratamentos e serviços de saúde(De Araujo *et al.*, 2012).

Na pesquisa clínica epidemiológica a maior contribuição é a capacidade de coleta e análise de grandes volumes de informações clínicas para identificar padrões e tendências em estudos de coorte, análise de eficácia de medicamentos e monitoramento de surtos de doenças. Outra contribuição é a automação de processos administrativos, reduzindo erros com a codificação de diagnósticos, faturamento e apoio às tarefas logísticas, incluindo otimização espacial. Finalmente, espera-se que haja melhoria da UX do sistema de saúde, uma vez que os dados extraídos podem ser usados para personalizar e melhorar a experiência do paciente, fornecendo informações mais relevantes e oportunas, com personalização de planos terapêuticos e estreitamento da comunicação com a equipe de saúde.

4.4 Aplicação de Técnicas Inteligentes para Análise de Imagens

A análise de imagens de saúde inclui a interpretação de raios-X, tomografias, Ressonâncias Magnéticas (MRI), ultrassons e outras modalidades de imagem médica. Técnicas inteligentes, especialmente aquelas baseadas em aprendizado profundo (*deep learning*), são amplamente utilizadas. Técnicas que utilizam dados rotulados (supervisionado) ou não rotulados (não supervisionado) para treinar modelos de aprendizado são úteis para treinamento de modelos diagnósticos e agrupamento imagens não rotuladas para identificar novos padrões ou subtipos de doenças.

As redes neurais vem se destacando no processamento de imagens. As CNNs, ou rede neural profunda vem sendo adotadas para detecção de anomalias, identificação de tumores, fraturas, lesões e outras anomalias em imagens médicas; classificação de imagens e tipos de doenças com base em imagens anotadas (cujo diagnóstico é conhecido), como diferentes tipos de câncer; e segmentação de imagens para delimitação de áreas de interesse em uma imagem, como o contorno de um tumor.

O algoritmo recebe como *input* imagens de mamografias e, após processamento, gera como output a probabilidade de presença de tumores malignos. A validação desse modelo é realizada comparando os resultados do algoritmo com diagnósticos confirmados por biópsias, utilizando métricas como a área sob a curva ROC (AUC-ROC) para avaliar a performance do modelo.

Outro caso de aplicação é na detecção de retinopatia diabética a partir de imagens de retina. Utilizando CNNs, o sistema recebe imagens fundoscópicas e classifica a gravidade da retinopatia em diferentes níveis. A saída do modelo é uma classificação que indica a presença e a severidade da doença. A validação é feita através da comparação com diagnósticos fornecidos por oftalmologistas especialistas, usando métricas como acurácia, sensibilidade (*recall*) e especificidade.

Na análise de Tomografias Computadorizadas (TC) para a detecção de nódulos pulmonares, algoritmos de deep learning, especificamente CNNs, são aplicados para identificar e segmentar nódulos em imagens de alta resolução. As imagens tomográficas são usadas como input, e o output consiste na localização e classificação dos nódulos. A validação é realizada comparando os resultados do algoritmo com anotações feitas por radiologistas, utilizando métricas como precisão, sensibilidade, e a métrica Dice para avaliação de segmentação.

Na detecção de lesões cerebrais em MRI, técnicas de aprendizado profundo, incluindo CNNs, são empregadas para segmentar e identificar áreas de lesão. As imagens de ressonância magnética são utilizadas como input e o output é a segmentação das áreas afetadas. A validação é realizada através da comparação com segmentações manuais feitas por especialistas, utilizando métricas como a acurácia, a *Dice coefficient* e a precisão, para garantir a confiabilidade dos resultados.

Outra técnica é a ***Transfer Learning*** (Transferência de Aprendizado), onde um modelo treinado em uma tarefa é ajustado (*fine-tuned*) para outra tarefa similar. Assim, ocorre a adaptação de Modelos Pré-Treinados em grandes conjuntos de dados, como o *ImageNet* (Fei-Fei *et al*., 2015), e ajustá-los para tarefas específicas de saúde, como a detecção de pneumonia em raios-X.

Vamos trabalhar num exemplo com radiografias de mama *Mini Mammographic Database* (MIAS) (Mader, 2019; Suckling, 1994). A base de dados original da MIAS (digitalizada com resolução de 50 microns por pixel) foi reduzida para 200 microns por pixel e ajustada/padronizada para que todas as imagens tenham 1024 pixels x 1024 pixels. A base é anotada, isto é, as imagens foram classificadas por especialistas. Ela contém o número de referência da base de dados MIAS; um carácter do tecido de fundo, sendo F - *Fatty* (gorduroso), G - *Fatty-glandular* (gorduroso-glandular), D - *Dense-glandular* (denso-glandular); classe de anormalidade presente, sendo CALC - Calcificação, CIRC - Massas bem definidas/circunscritas, SPIC - Massas espiculadas,

MISC - Outras massas mal definidas, ARCH - Distorção arquitetônica, ASYM - Assimetria e NORM - Normal; e, finalmente, a gravidade da anormalidade, sendo B - Benigna e M - Maligna. Adicionalmente apresenta coordenadas (x,y) da imagem indicando o centro da anomalia e raio aproximado (em pixels) de um círculo que engloba a anomalia.

A lista está organizada em pares de filmes, onde cada par representa as mamografias esquerda (números pares de arquivo) e direita (números ímpares de arquivo) de um único paciente. Quando há calcificações presentes, as localizações dos centros e os raios se aplicam a aglomerados, em vez de calcificações individuais. O sistema de coordenadas tem origem no canto inferior esquerdo. Em alguns casos, as calcificações estão distribuídas por toda a imagem, em vez de concentradas. Nesses casos, as localizações dos centros e os raios são inadequados e foram omitidos.

Passo 1: Bibliotecas.

```
# Instala a biblioteca TensorFlow, necessária para o
processamento e treinamento de modelos de deep learning.
!pip install tensorflow


# Importa bibliotecas para processamento de dados,
visualização e manipulação de arquivos.
import numpy as np # Álgebra linear
import pandas as pd # Processamento de dados,
leitura de arquivos CSV, etc.
import matplotlib.pyplot as plt # Exibição e renderização de figuras


# Importa bibliotecas para manipulação de arquivos
de imagem e outros arquivos de entrada/saída.
from skimage.io import imread
import os
from glob import glob
import h5py
# Necessário em Jupyter para exibir gráficos inline.
%matplotlib inline
```

Passo 2: Pré-processamento.

```
# Converte o arquivo HDF5 em um DataFrame e
uma pasta cheia de arquivos TIFF.
base _ h5 = '/content/all _ mias _ scans.h5' # Caminho do arquivo HDF5
tif _ dir = '/content/tiffs' # Diretório onde
os arquivos TIFF serão salvos
os.makedirs(tif _ dir, exist _ ok=True) # Cria o diretório se não existir

# Abre o arquivo HDF5 e converte seu conteúdo
em um DataFrame do Pandas.
with h5py.File(base _ h5, 'r') as f:
    mammo _ df = pd.DataFrame(
        {k: v[:] if len(v.shape) == 1 else [sub _ v for sub _ v in v]
         for k, v in f.items()}
    )

# Decodifica colunas que estão em formato bytes.
for k in mammo _ df.columns:
    if isinstance(mammo _ df[k].values[0], bytes):
        mammo _ df[k] = mammo _ df[k].map(lambda x: x.decode())

# Salva os dados em disco como arquivos TIFF.
from skimage.io import imsave
def to _ path(c _ row):
    out _ path = os.path.join(tif _ dir, '%s.tif' % c _ row['REFNUM'])
    imsave(out _ path, c _ row['scan'])
    return out _ path

mammo _ df['scan'] = mammo _ df.apply(to _ path, axis=1)
mammo _ df.sample(5) # Exibe uma amostra dos dados
```

Figura 48 - Tabela de Resultado do código-fonte compilado

G	CLASS	RADIUS	REFNUM	SEVERITY	X	Y	path	scan	
87	F	NORM	NaN	mdb087	nan	NaN	NaN	mdb087.pgm	tiffs/mdb087.tif
102	D	ASYM	38.0	mdb102	M	415.0	460.0	mdb102.pgm	tiffs/mdb102.tif
72	G	ASYM	28.0	mdb072	M	266.0	517.0	mdb072.pgm	tiffs/mdb072.tif
78	F	NORM	NaN	mdb078	nan	NaN	NaN	mdb078.pgm	tiffs/mdb078.tif
64	D	NORM	NaN	mdb064	nan	NaN	NaN	mdb064.pgm	tiffs/mdb064.tif

Fonte: autoria própria.

Passo 3: Divisão dos dados em treinamento e teste.

```
# Examina as distribuições
# Mostra como os dados estão distribuídos e
por que precisamos balanceá-los.
from sklearn.preprocessing import LabelEncoder
from tensorflow.keras.utils import to _ categorical

class _ enc = LabelEncoder()
mammo _ df['CLASS _ ID'] = class _ enc.fit _ transform(mammo _ df['CLASS'])
mammo _ df['CLASS _ VEC'] = mammo _ df['CLASS _ ID'].map(lambda x:
to _ categorical(x, num _ classes=len(class _ enc.classes _ )))
mammo _ df[['CLASS _ ID', 'RADIUS', 'SEVERITY']].hist(figsize=(10, 5))

# Divide os dados em treino e validação.
from sklearn.model _ selection import train _ test _ split
raw _ train _ df, valid _ df = train _ test _ split(mammo _ df,
                                    test _ size=0.25,
                                    random _ state=2018,
                                    stratify=mammo _
df[['CLASS _ ID', 'SEVERITY']])
print('train', raw _ train _ df.shape[0], 'validation', valid _ df.shape[0])
raw _ train _ df.sample(1)
```

Passo 4: balanceamento da distribuição no conjunto de treino.

```
train _ df = raw _ train _ df.groupby(['CLASS', 'SEVERITY']).apply(lambda
x: x.sample(100, replace=True)).reset _ index(drop=True)
print('New Data Size:', train _ df.shape[0], 'Old
Size:', raw _ train _ df.shape[0])
train _ df[['CLASS _ ID', 'RADIUS']].hist(figsize=(10, 5))

# Criação de um gerador de dados para aumentar e normalizar as imagens.
from tensorflow.keras.preprocessing.image import ImageDataGenerator
from tensorflow.keras.applications.imagenet _
utils import preprocess _ input

IMG _ SIZE = (192, 192) # Tamanho da imagem
core _ idg = ImageDataGenerator(samplewise _ center=False,
                                samplewise _ std _ normalization=False,
                                horizontal _ flip=True,
                                vertical _ flip=False,
                                height _ shift _ range=0.15,
                                width _ shift _ range=0.15,
                                rotation _ range=5,
                                shear _ range=0.01,
                                fill _ mode='nearest',
                                zoom _ range=0.2,
                                preprocessing _ function=preprocess _ input)
```

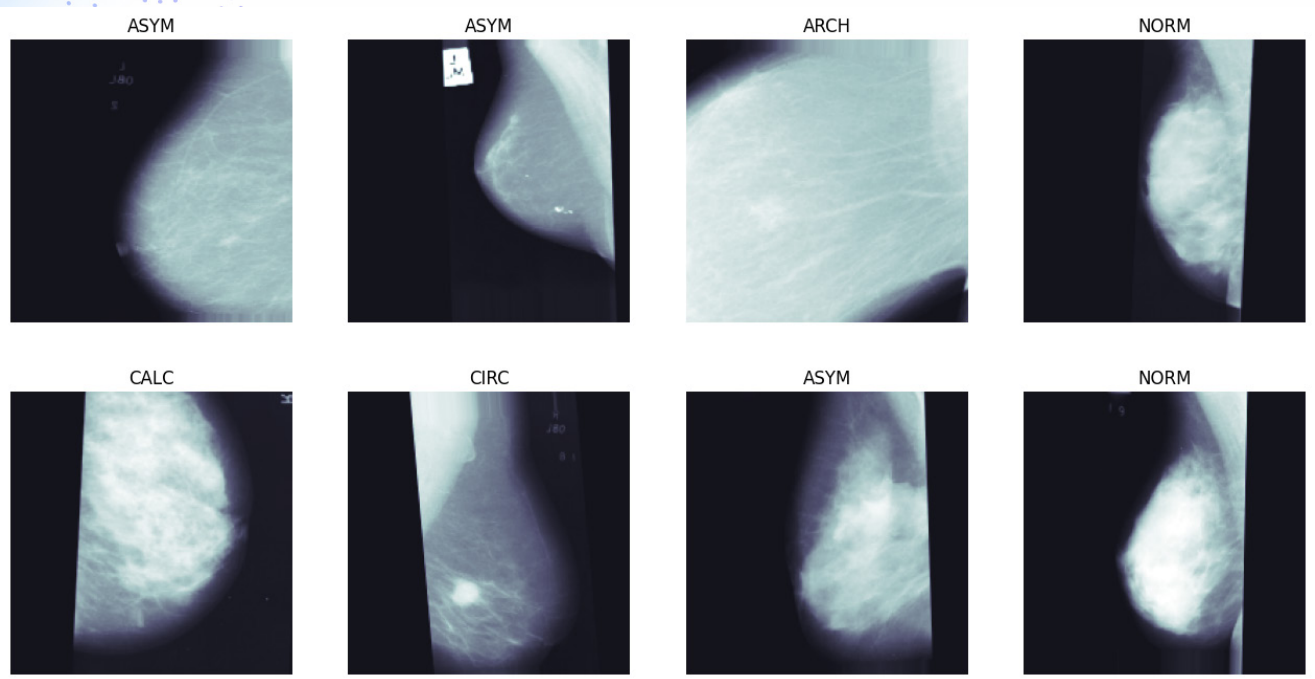
Visualização de um lote de treino.

```
# Função para gerar dados a partir de um DataFrame.
def flow _ from _ dataframe(img _ data _ gen, in _
df, path _ col, y _ col, **dflow _ args):
    df _ gen = img _ data _ gen.flow _ from _
dataframe(in _ df, x _ col=path _ col,
                                            y _ col=y _ col,
                                            class _ mode='raw',
                                            **dflow _ args)
    return df _ gen
```

continua

```
# Cria geradores de dados para treino e validação.
train _ gen = flow _ from _ dataframe(core _ idg, train _ df,
                                      path _ col='scan',
                                      y _ col='CLASS _ ID',
                                      target _ size=IMG _ SIZE,
                                      color _ mode='rgb',
                                      batch _ size=32)

valid _ gen = flow _ from _ dataframe(core _ idg, valid _ df,
                                      path _ col='scan',
                                      y _ col='CLASS _ ID',
                                      target _ size=IMG _ SIZE,
                                      color _ mode='rgb',
                                      batch _ size=256) #
Podemos usar lotes maiores para avaliação

# Cria um conjunto de dados fixo para avaliar o algoritmo.
test _ X, test _ Y = next(flow _ from _ dataframe(core _ idg,
                                                  valid _ df,
                                                  path _ col='scan',
                                                  y _ col='CLASS _ ID',
                                                  target _ size=IMG _ SIZE,
                                                  color _ mode='rgb',
                                                  batch _ size=1024))

# Gera um lote de dados de treino para visualização.
t _ x, t _ y = next(train _ gen)
fig, m _ axs = plt.subplots(2, 4, figsize=(16, 8))
for (c _ x, c _ y, c _ ax) in zip(t _ x, t _ y, m _ axs.flatten()):
    c _ ax.imshow(c _ x[:,:,0], cmap='bone', vmin=-127, vmax=127)
    c _ ax.set _ title('%s' % (class _ enc.classes _ [c _ y]))
    c _ ax.axis('off')
```

Figura 49 - Geradores de dados de imagem para treinamento e avaliação de modelos de aprendizado de máquina



Fonte: autoria própria.

Passo 5: modelo de atenção.

O *Global Average Pooling* (ou agrupamento médio global) é uma técnica comum em CNNs, usada como uma camada de agrupamento. Na técnica de Camadas de *Pooling* em CNNs é calculada a média de todos os valores de um mapa de características. *Global Average Pooling* é uma operação de amostragem projetada para substituir camadas totalmente conectadas em CNNs clássicas. A ideia é gerar um mapa de características para cada categoria correspondente da tarefa de classificação na última camada mlpconv.

Em essência, a *Global Average Pooling* funciona da seguinte forma: para cada mapa de características gerado pela camada convolucional é calculada a média de todos os valores desse mapa. Em seguida, esses valores médios são utilizados como entrada para as próximas camadas da rede neural. Uma das vantagens da *Global Average Pooling* é a redução do número de parâmetros na rede e a evitar o *overfitting*, pois o número de parâmetros é menor do que em outras técnicas de pooling. Além disso, o *Global Average Pooling* mantém a informação espacial das características originais, o que pode ser útil em tarefas de segmentação e classificação de imagens.

A ideia principal é que a técnica de *Global Average Pooling* considera todas as regiões da mesma forma, mesmo que algumas sejam mais importantes do que outras. Para lidar com essa questão, foi desenvolvido um mecanismo de atenção que decide

quais *pixels* devem ser considerados antes de realizar o *pooling*, e então redimensiona os resultados com base na quantidade de *pixels*. Esse método pode ser comparado a uma forma de pooling em que a média é calculada conforme a importância de cada *pixel*.

```
# bibliotecas
from tensorflow.keras.applications.vgg16 import VGG16
from tensorflow.keras.layers import GlobalAveragePooling2D,
Dense, Dropout, Input, Conv2D, multiply, Lambda
from tensorflow.keras.models import Model

# Definindo a entrada do modelo
in _ lay = Input(t _ x.shape[1:])

# Carregando o modelo VGG16 pré-treinado sem a
camada superior e com pesos da ImageNet
base _ pretrained _ model = VGG16(input _ shape=t _ x.
shape[1:], include _ top=False, weights='imagenet')
base _ pretrained _ model.trainable = False  #
Congelando as camadas do modelo pré-treinado

# Obtendo a profundidade das características
da saída do modelo pré-treinado
pt _ depth = base _ pretrained _ model.output.shape[-1]

# Extraindo características da entrada usando o modelo pré-treinado
pt _ features = base _ pretrained _ model(in _ lay)

# Normalizando as características
from tensorflow.keras.layers import BatchNormalization
bn _ features = BatchNormalization()(pt _ features)

# Aqui fazemos um mecanismo de atenção para
ligar e desligar pixels no GAP
attn _ layer = Conv2D(64, kernel _ size=(1, 1),
padding='same', activation='relu')(bn _ features)
```

continua

```
attn _ layer = Conv2D(16, kernel _ size=(1, 1),
padding='same', activation='relu')(attn _ layer)
attn _ layer = Conv2D(1, kernel _ size=(1, 1),
padding='valid', activation='sigmoid')(attn _ layer)
```

```
# Expande a atenção para todos os canais
up _ c2 = Conv2D(pt _ depth, kernel _ size=(1, 1),
padding='same', activation='linear', use _ bias=False)
up _ c2.build(attn _ layer.shape)  # Construindo a
camada para definir a forma de entrada
up _ c2.set _ weights([up _ c2 _ w])  # Definindo os pesos
up _ c2.trainable = False  # Congelando a camada

# Aplicando a camada de atenção
attn _ layer = up _ c2(attn _ layer)

# Multiplicando as características pela atenção
mask _ features = multiply([attn _ layer, bn _ features])

# Aplicando a Global Average Pooling 2D
gap _ features = GlobalAveragePooling2D()(mask _ features)
gap _ mask = GlobalAveragePooling2D()(attn _ layer)

# Para considerar os valores ausentes do modelo de atenção
gap = Lambda(lambda x: x[0] / x[1], name='RescaleGAP')
([gap _ features, gap _ mask])
gap _ dr = Dropout(0.5)(gap)
dr _ steps = Dropout(0.25)(Dense(128, activation='elu')(gap _ dr))
out _ layer = Dense(len(class _ enc.classes _ ),
activation='softmax')(dr _ steps)

# Criando o modelo
mammo _ model = Model(inputs=[in _ lay], outputs=[out _ layer])

# Compilando o modelo
```

continua

```
mammo _ model.compile(optimizer='adam', loss='sparse _ categorical _
crossentropy', metrics=['sparse _ categorical _ accuracy'])


# Resumo do modelo
mammo _ model.summary()
```

A tabela gerada pelo comando *mammo_model.summary()* fornece um resumo detalhado da arquitetura do modelo de rede neural definido no *script* anterior.

Tabela 3 - Comando *mammo_model.summary*

Layer (type)	Output Shape	Param #	Connected to
input_layer_4 (InputLayer)	(None, 192, 192, 3)	0	-
vgg16 (Functional)	(None, 6, 6, 512)	14,714,688	input_layer_4[0][0]
batch_normalization_1 (BatchNormalization)	(None, 6, 6, 512)	2,048	vgg16[0][0]
conv2d_4 (Conv2D)	(None, 6, 6, 64)	32,832	batch_normalization_1[0][0]
conv2d_5 (Conv2D)	(None, 6, 6, 16)	1,04	conv2d_4[0][0]
conv2d_6 (Conv2D)	(None, 6, 6, 1)	17	conv2d_5[0][0]
conv2d_8 (Conv2D)	(None, 6, 6, 512)	512	conv2d_6[0][0]
multiply (Multiply)	(None, 6, 6, 512)	0	conv2d_8[0][0],batch_normalization_1[0][0]
global_average_pooling2d (GlobalAveragePooling2D)	(None, 512)	0	multiply[0][0]
global_average_pooling2d (GlobalAveragePooling2D)	(None, 512)	0	conv2d_8[0][0]
RescaleGAP (Lambda)	(None, 512)	0	global_average_pooling2d[0][0],global_average_pooling2d[1][0]
dropout (Dropout)	(None, 512)	0	RescaleGAP[0][0]
dense (Dense)	(None, 128)	65,664	dropout[0][0]
dropout_1 (Dropout)	(None, 128)	0	dense[0][0]
dense_1 (Dense)	(None, 7)	903	dropout_1[0][0]

Fonte: autoria própria.

Descrição das colunas:

- *Layer* (type): Nome e tipo de camada na rede neural. Por exemplo, InputLayer, Conv2D, BatchNormalization, Dense, etc.
- *Output Shape*: A forma da saída de cada camada. Ela descreve como são os dados que saem de cada camada. Por exemplo, (None, 192, 192, 3) significa que a camada produz uma saída de imagens de 192x192 pixels com 3 canais (cores RGB).

- Param #: O número de parâmetros treináveis em cada camada. Parâmetros são pesos e vieses que a rede ajusta durante o treinamento. Por exemplo, a camada conv2d_4 tem 32.832 parâmetros.
- *Connected to*: Indica as conexões entre as camadas, mostrando de onde a camada está recebendo suas entradas. Por exemplo, batch_normalization_1[0][0] indica que a camada conv2d_4 recebe sua entrada da camada batch_normalization_1.

Descrição das camadas:

- *Input Layer*: A camada de entrada define a forma dos dados que entram na rede. No caso, imagens de 192x192 pixels com 3 canais.
- VGG16 (Functional): Uma rede pré-treinada VGG16 usada como base, sem as camadas de topo (classificação), apenas como extrator de características. A saída dessa camada é um tensor com forma (None, 6, 6, 512).
- *BatchNormalization*: Normaliza a saída da camada anterior para acelerar o treinamento e melhorar a estabilidade do modelo.
- Conv2D: Camadas convolucionais que aplicam filtros à imagem de entrada para extrair características. A sequência de camadas conv2d_4, conv2d_5 e conv2d_6 forma uma atenção (*attention*) que destaca partes importantes da imagem.
- *Multiply*: Multiplica os valores da camada de atenção pelos valores normalizados para destacar as partes mais relevantes da imagem.
- *GlobalAveragePooling2D*: Reduz as dimensões espaciais da imagem (6x6) para uma dimensão, calculando a média dos valores, resultando em um vetor de 512 características.
- *Lambda* (RescaleGAP): Redimensiona os valores da saída da camada de pooling para ajustar os valores que foram afetados pela camada de atenção.
- *Dropout*: Camadas que ajudam a *prevenir overfitting* ao desativar aleatoriamente uma fração das unidades durante o treinamento.
- *Dense*: Camadas totalmente conectadas que combinam as características extraídas em classes finais. A última camada dense_1 possui uma função de ativação softmax para classificação multi-classe (7 classes no total).

A rede neural começa com a entrada das imagens, passa por uma rede VGG16 pré-treinada para extrair características, aplica uma camada de normalização, e então utiliza uma série de camadas convolucionais para criar um mecanismo de atenção que realça regiões importantes da imagem. As características destacadas são

então reduzidas a um vetor de características, redimensionadas e passadas por algumas camadas densas para produzir a classificação final. O modelo foi usado para tarefas como a classificação de imagens de mamografias em diferentes categorias de anomalias, levando em consideração a importância relativa de diferentes regiões da imagem.

```
# Importando callbacks do Keras
from keras.callbacks import ModelCheckpoint,
ReduceLROnPlateau, EarlyStopping

# Definindo o caminho para salvar os pesos do melhor modelo
weight _ path = "{} _ weights.best.weights.h5".format('mammo _ result')

# Definindo os callbacks
checkpoint = ModelCheckpoint(weight _ path, monitor='val _ loss',
verbose=1, save _ best _ only=True, mode='min', save _ weights _ only=True)
reduceLROnPlat = ReduceLROnPlateau(monitor='val _ loss', factor=0.8,
patience=10, verbose=1, mode='auto', min _ lr=0.0001)
early = EarlyStopping(monitor="val _ loss", mode="min", patience=5)

# Lista de callbacks
callbacks _ list = [checkpoint, early, reduceLROnPlat]

# Treinando o modelo
mammo _ model.fit(train _ gen, steps _ per _ epoch=35, validation _
data=(test _ X, test _ Y), epochs=5, callbacks=callbacks _ list)

# Carregando a melhor versão do modelo
mammo _ model.load _ weights(weight _ path)
```

Vamos analisar o processo de treinamento do modelo conforme indicado pela saída do comando mammo_model.fit:

- *train_gen*: O gerador de dados de treinamento.
- steps_per_epoch = 35: Número de lotes de treinamento a serem executados em cada época.

- *validation_data* = (test_X, test_Y): Conjunto de dados de validação usado para avaliar a performance do modelo após cada época.
- *epochs* = 5: Número de épocas (iterações sobre todo o conjunto de dados de treinamento).
- *callbacks* = callbacks_list: Lista de *callbacks* usados durante o treinamento, incluindo salvamento de melhores pesos, redução da taxa de aprendizado e parada antecipada.

Veja um resultado.

```
Epoch 1: val _ loss improved from inf to 2.89962, saving
model to mammo _ result _ weights.best.weights.h5
35/35 ------------------- 566s 16s/step - loss: 1.1880 - sparse _
categorical _ accuracy: 0.5955 - val _ loss: 2.5701 - val _ sparse _
categorical _ accuracy: 0.1928 - learning _ rate: 0.0010
<keras.src.callbacks.history.History at 0x7b54441f5f90>
```

- loss: 1.1880: Perda no conjunto de treinamento.
- sparse_categorical_accuracy: 0.5955: Acurácia no conjunto de treinamento.
- val_loss: 2.5701: Perda no conjunto de validação (melhorou).
- val_sparse_categorical_accuracy: 0.1928: Acurácia no conjunto de validação.

O modelo foi salvo com os melhores pesos, uma vez que *val_loss* melhorou. A perda no treinamento: diminuiu constantemente de 1.9869 para 1.1880. A acurácia no treinamento melhorou de 20.45% para 59.55%. A perda na validação melhorou de 2.8996 para 2.5701, embora tenha aumentado em alguns pontos. A acurácia na validação melhorou de 10.84% para 19.28%, embora a melhoria tenha sido modesta.

O treinamento parece melhorar a performance do modelo em termos de perda e acurácia tanto no conjunto de treinamento quanto no de validação. A acurácia no conjunto de validação é significativamente menor do que no conjunto de treinamento, indicando possível *overfitting*. *Callbacks* como *ModelCheckpoint*, *ReduceLROnPlateau* e *EarlyStopping* ajudam a ajustar e salvar o melhor modelo durante o processo de treinamento.

Passo 6: Avaliação e exibição do modelo de atenção.

```
import tensorflow as tf  # Importar TensorFlow

# Obter a camada de atenção, já que é a única
com uma única dimensão de saída
for attn _ layer in mammo _ model.layers[1:]:  #
Iterar a partir da segunda camada
   if isinstance(attn _ layer, tf.keras.layers.
Conv2D):  # Verificar se a camada é do tipo Conv2D
       c _ shape = attn _ layer.output.shape  #
Usar output.shape para camadas Conv2D
       if len(c _ shape) == 4:  # Verificar se
a forma da saída tem 4 dimensões
           if c _ shape[-1] == 1:  # Verificar se a última dimensão é 1
               print(attn _ layer)  # Imprimir
a camada de atenção encontrada
               break  # Parar a busca após
encontrar a camada de atenção

import tensorflow.keras.backend as K  # Importar
o backend do tensorflow.keras

# Selecionar 6 índices aleatórios dos dados de teste
rand _ idx = np.random.choice(range(len(test _ X)), size=6)

# Para modelos funcionais, acessar diretamente
as camadas de entrada e saída
attn _ func = tf.keras.Model(inputs=mammo _ model.
inputs,  # Usar model.inputs para modelos funcionais
                            outputs=attn _ layer.
output  # Usar layer.output diretamente
                            )

# Criar subplots para mostrar as imagens e mapas de atenção
fig, m _ axs = plt.subplots(len(rand _ idx),
2, figsize=(8, 4*len(rand _ idx)))
```

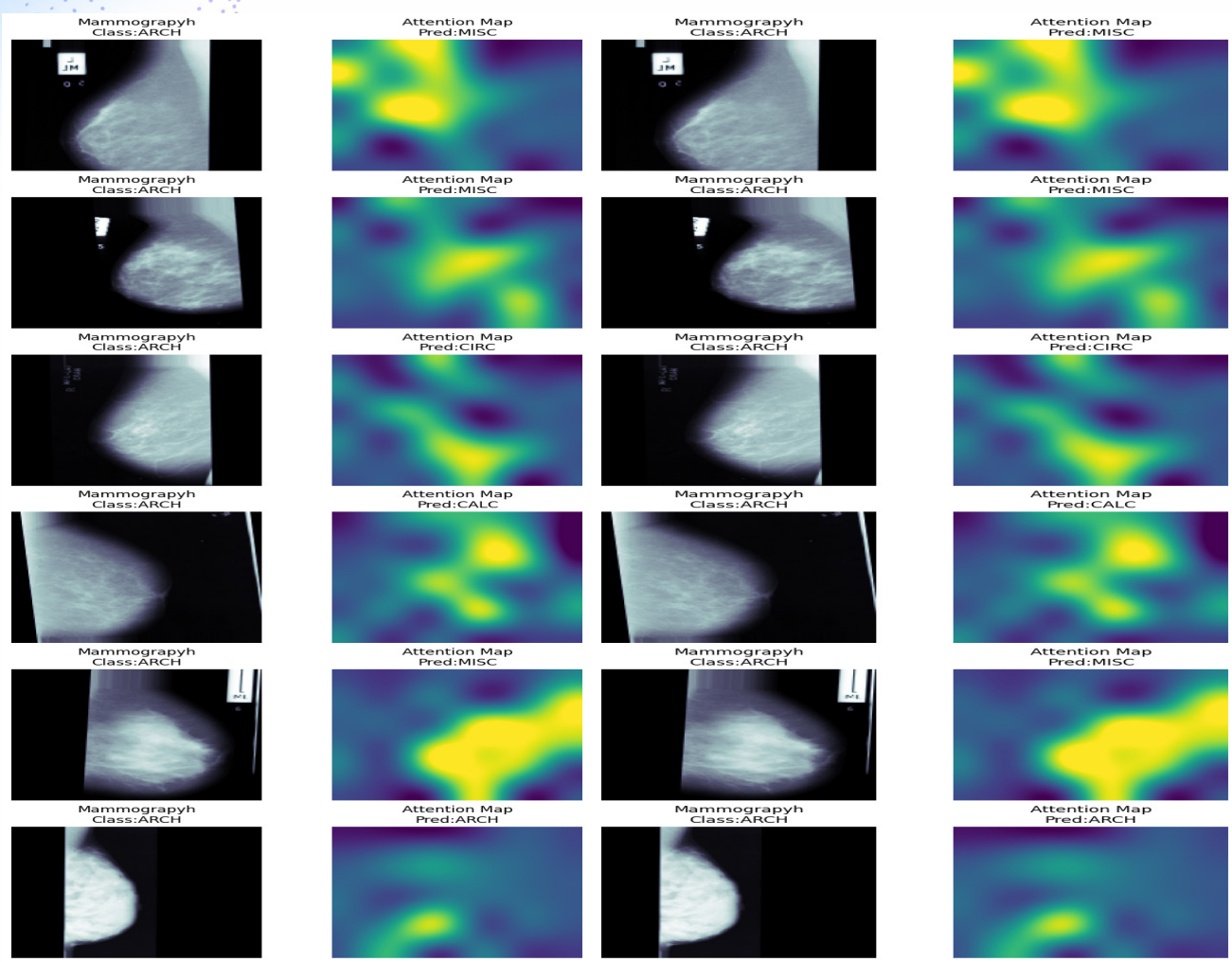
continua

```
[c _ ax.axis('off') for c _ ax in m _ axs.flatten()]  #
Desligar os eixos para todos os subplots


# Iterar sobre os índices aleatórios e os subplots
for c _ idx, (img _ ax, attn _ ax) in zip(rand _ idx, m _ axs):
    cur _ img = test _ X[c _ idx:(c _ idx+1)]  # Selecionar a imagem atual
    attn _ img = attn _ func(cur _ img)  # Gerar o
mapa de atenção para a imagem atual
    img _ ax.imshow(cur _ img[0, :, :, 0], cmap='bone')  #
Mostrar a imagem em escala de cinza
    attn _ ax.imshow(attn _ img[0, :, :, 0],
cmap='viridis',  # Mostrar o mapa de atenção
                     vmin=0, vmax=1,
                     interpolation='lanczos')
    real _ label = class _ enc.classes _ [np.argmax(test _ Y[c _
idx])]  # Obter o rótulo real da imagem
    img _ ax.set _ title('Mamografia\nClasse:%s' % (real _
label))  # Definir o título com o rótulo real
    pred _ confidence = class _ enc.classes _ [np.argmax(mammo _ model.
predict(cur _ img)[0], -1)]  # Prever a classe da imagem atual
    attn _ ax.set _ title('Mapa de Atenção\nPredição:%s' %
(pred _ confidence))  # Definir o título com a predição
fig.savefig('attention _ map.png', dpi=300)  #
Salvar a figura com os mapas de atenção
```

A identificação da camada de atenção foi obtida com a iteração sobre as camadas do modelo mammo_model para encontrar a camada de atenção (Conv2D com uma única dimensão de saída). Na seleção de imagens aleatórias foram obtidos 6 índices dos dados de teste test_X. Na criação do modelo de função de atenção foi criado um modelo funcional *attn_func* que tem como entrada a mesma entrada do modelo mammo_model e como saída a camada de atenção encontrada.

Na visualização dos mapas de atenção criamos subplots para mostrar as imagens de mamografia e seus respectivos mapas de atenção. A iteração sobre os índices aleatórios exibe a imagem de mamografia e o mapa de atenção correspondente. Os títulos das imagens correspondem a classe real e a predição do modelo.

Figura 50 - Imagens de mamografia e seus respectivos mapas de atenção



Fonte: autoria própria.

Passo 7: validação.

```
# Fazer previsões usando o modelo treinado
pred _ Y = mammo _ model.predict(test _ X, batch _ size=32, verbose=True)

# Converter as previsões para as classes categóricas
pred _ Y _ cat = np.argmax(pred _ Y, -1)

# Definir as classes reais dos dados de teste
test _ Y _ cat = test _ Y

# Importar funções para avaliação do modelo
from sklearn.metrics import classification _ report, confusion _ matrix
import matplotlib.pyplot as plt
```

continua

```
# Calcular e mostrar a matriz de confusão
plt.matshow(confusion _ matrix(test _ Y _ cat, pred _ Y _ cat))
plt.title('Matriz de Confusão')
plt.colorbar()
plt.xlabel('Predição')
plt.ylabel('Classe Real')
plt.show()

# Mostrar o relatório de classificação
print(classification _ report(test _ Y _ cat, pred _ Y _
cat, target _ names=class _ enc.classes _ ))
```

A previsão com o modelo foi obtida com *pred_Y = mammo_model.predict(test_X, batch_size=32, verbose=True)*, o qual faz previsões nas amostras de teste test_X usando o modelo mammo_model. *batch_size=32* define o tamanho do lote de previsões e *verbose=True* fornece informações sobre o progresso.

A conversão das previsões para classes foi obtida com *pred_Y_cat = np.argmax(pred_Y, -1)*, pois converte as previsões do modelo em classes categóricas. *np.argmax* é usado para obter o índice da classe com a maior probabilidade em cada previsão.

A definição de classes reais ocorreu com *test_Y_cat = test_Y*: atribui as classes reais dos dados de teste à variável *test_Y_cat*.

Para a exibição da matriz de confusão foi utilizado *plt.matshow(confusion_matrix(test_Y_cat, pred_Y_cat))*, o qual calcula e exibe a matriz de confusão. A matriz mostra a comparação entre as classes reais e as classes previstas pelo modelo. *print(classification_report(test_Y_cat, pred_Y_cat, target_names=class_enc.classes_))* calcula e imprime o relatório de classificação, que inclui métricas como precisão (precision), *recall*, f1-score e o suporte para cada classe. *target_names* é usado para rotular as classes no relatório.

A tabela mostra as métricas de desempenho do modelo para cada classe. *Precision* (precisão) mostra a proporção de previsões positivas corretas sobre o total de previsões positivas feitas pelo modelo. *Recall* (revocação) mostra a proporção de VP sobre o total de VP e FN. F1-*score* mostra a média harmônica da precisão e *recall*, que fornece um balanço entre as duas métricas. *Support* (Suporte) mostra o número de amostras verdadeiras de cada classe.

Figura 51 - Tabela da métricas de desempenho do modelo

class	precision	recall	f1-score	support
ARCH	0.15	0.40	0.22	5
ASYM	0.12	0.33	0.18	3
CALC	0.38	0.62	0.48	8
CIRC	0.23	0.50	0.32	6
MISC	0.08	0.50	0.14	4
NORM	1.00	0.02	0.04	52
SPIC	0.20	0.40	0.27	5
accuracy		0.19		83
macro avg	0.31	0.40	0.23	83
weighted avg	0.71	0.19	0.14	83

Fonte: autoria própria.

A matriz de confusão representa visualmente a quantidade de previsões corretas e incorretas do modelo para cada classe. No relatório de classificação, a *accuracy* (acurácia) mostra o percentual de previsões corretas no total de amostras; a *Macro Average* mostra a média das métricas para cada classe, tratadas igualmente; a *Weighted Average* retorna a média das métricas ponderadas pelo número de amostras em cada classe.

Quem ajudou mais o modelo: a categoria CALC, com o maior *recall* e um f1-*score* relativamente alto, foi a que ajudou mais o modelo. Quem prejudicou mais o modelo: A categoria NORM prejudicou significativamente o modelo devido ao seu alto suporte e ao desequilíbrio extremo entre precisão e recall, resultando no f1-*score* mais baixo. Além disso, categorias como MISC e ASYM também tiveram um impacto negativo significativo devido à baixa precisão e f1-*score*.

A categoria CALC - calcificação tem a melhor pontuação de *recall* (0,62), indicando que o modelo consegue identificar bem as calcificações quando elas estão presentes. No entanto, a precisão (0,38) é mais baixa, o que sugere que o modelo pode

estar classificando erroneamente alguns exemplos que não são calcificações como calcificações. Apesar disso, o f1-score (0,48) é relativamente alto comparado às outras categorias, mostrando um bom equilíbrio entre precisão e recall.

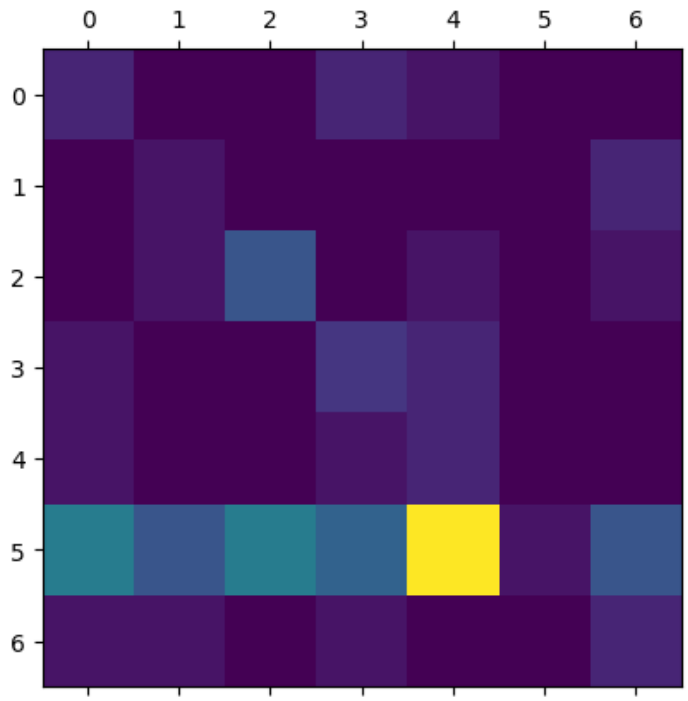
A categoria CIRC tem um recall razoável (0,50), indicando que o modelo pode detectar massas bem definidas/circunscritas com certa eficácia. Contudo, a precisão é baixa (0,23), indicando que muitas das classificações como CIRC podem estar erradas. O f1-score (0,32) reflete essa dificuldade em equilibrar precisão e recall.

A categoria NORM apresenta uma precisão alta (1,00), mas um *recall* muito baixo (0,02). Isso sugere que o modelo está classificando corretamente muitos exemplos como normais, mas falha em identificar adequadamente a grande maioria dos casos de normalidade. O f1-score (0,04) é o mais baixo entre todas as categorias, refletindo um desequilíbrio.

Na tabela de resultados do modelo, as métricas *accuracy*, *macro avg* e *weighted avg* são utilizadas para avaliar o desempenho geral do modelo. *Accuracy* (Acurácia) é a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões feitas. Em outras palavras, é a razão entre o número de previsões corretas e o número total de amostras. A acurácia é 0,19 significa que o modelo acertou 19% das previsões feitas no conjunto de teste. *Macro Average* (Média Macro) é a média das métricas de desempenho (precisão, *recall*, f1-*score*) calculadas para cada classe individualmente, sem levar em consideração o número de amostras em cada classe. É calculada somando as métricas de todas as classes e dividindo pelo número total de classes. A média macro para precisão é 0,31, para *recall* é 0,40, e para f1-*score* é 0,23. Isso significa que, ao calcular a média dos desempenhos das diferentes classes, cada classe é tratada igualmente, independentemente do seu suporte. *Weighted Average* (Média Ponderada) é a média das métricas de desempenho (precisão, *recall*, f1-*score*) onde cada classe contribui para a média de acordo com seu suporte (número de amostras reais). Portanto, classes com mais amostras têm um impacto maior na média ponderada. A média ponderada para precisão é 0,71, para *recall* é 0,19, e para f1-*score* é 0,14. Isso reflete o desempenho do modelo considerando o desequilíbrio no número de amostras de cada classe. As classes com maior suporte têm um impacto mais significativo na média ponderada.

Os resultados indicam que o modelo tem dificuldades em classificar corretamente a maioria das classes, com precisão e *recall* baixos para várias categorias, especialmente para a classe "*NORM*". Isso sugere que pode ser necessário melhorar o modelo ou os dados para obter melhores resultados.

Figura 52 - Desempenho do modelo por classe



Fonte: autoria própria.

O mapa de calor resultante da função confusion_matrix e o relatório de classificação fornecem informações detalhadas sobre o desempenho do modelo em termos de suas previsões para diferentes classes. Linhas (eixo y) representam as classes reais (o que a amostra realmente é). Colunas (eixo x) representam as classes previstas (o que o modelo previu). A diagonal principal mostra o número de previsões corretas para cada classe. Idealmente, esses valores devem ser altos, indicando que o modelo está fazendo boas previsões para essas classes. As áreas adjacentes à diagonal mostram, os erros de classificação, onde o modelo previu uma classe diferente da classe real. Valores mais altos fora da diagonal indicam um número maior de previsões incorretas.

As cores em um mapa de calor da matriz de confusão ajudam a visualizar a intensidade dos valores em diferentes células da matriz. Roxo (ou roxo escuro) representa valores altos na matriz de confusão. Em um contexto de matriz de confusão, isso normalmente indica um alto número de previsões corretas para uma determinada classe (alta contagem de VP) ou um alto número de erros para uma classe específica. Azul (ou azul claro) representa valores baixos. Isso pode significar que há poucas previsões para uma classe específica ou poucos erros associados a essa classe. Azul pode indicar valores mais baixos de contagens ou erros. Amarelo (ou amarelo claro) pode representar valores intermediários entre baixo e alto. Em uma matriz de confusão, pode indicar uma quantidade média de previsões ou erros para uma classe. Tons escuros indicam valores mais altos. Em um mapa de calor da matriz de confusão, sig-

nifica que a célula tem um número elevado de amostras, previsões corretas ou erros para uma determinada classe. Tons claros indicam valores mais baixos, o que sugere que a célula tem um número menor de amostras ou previsões incorretas para uma classe específica.

4.5 Ferramentas de *Software* para Processamento Inteligente de Dados de Saúde: *Python, R, TensorFlow, PyTorch*

Existe uma variedade de bibliotecas e *frameworks* para análise de dados, aprendizado de máquina e processamento de imagens (Figura 53).

Python é uma linguagem de programação versátil e amplamente utilizada, especialmente em ciência de dados e aprendizado de máquina, devido à sua simplicidade e vasta coleção de bibliotecas:

- **Pandas** para manipulação e análise de dados tabulares. Fornece estruturas de dados rápidas, flexíveis e expressivas.
- **NumPy** para computação numérica com suporte para arranjos multidimensionais e uma ampla coleção de funções matemáticas. Útil no processamento de grandes conjuntos de dados, operações matemáticas e científicas.
- **Scikit-learn** para aprendizado de máquina que inclui uma variedade de algoritmos de classificação, regressão e *clustering*. Útil para a construção de modelos preditivos, diagnósticos, segmentação de pacientes e análise de sobrevivência.
- **Matplotlib** e **Seaborn** são bibliotecas para visualização de dados, permitindo a criação de gráficos e plots detalhados. Apoia a visualização de tendências de saúde, gráficos de progressão de doenças e análise de dados demográficos.

R é uma linguagem de programação e ambiente de *software* para computação estatística e gráficos, popular entre estatísticos e cientistas de dados. Os principais Pacotes são:

- **dplyr** e **tidyr** para manipulação e limpeza de dados, transformação e preparação de dados, análise estatística e modelagem.
- **ggplot2** para visualização de dados, criar gráficos complexos, o qual apoia a visualização de dados de estudos clínicos, gráficos de distribuição de doenças e visualização de resultados de pesquisa.
- **caret** que integra várias técnicas de aprendizado de máquina para construção e avaliação de modelos preditivos de dados de saúde.

Figura 53 - Tabela comparativa das principais ferramentas de *software* para o processamento inteligente de dados de saúde, com foco em *Python, R, TensorFlow* e *PyTorch*

Ferramenta	Linguagem	Aplicações principais	Bibliotecas/ Ferramentas destacadas	Pontos fortes	Uso na saúde
Python	Versátil, de uso geral	Análise de dados, aprendizado de máquina, processamento de imagens	• Pandas: Manipulação e análise de dados tabulares • NumPy: Computação numérica • Scikit-learn: Modelagem preditiva e algoritmos de ML • Matplotlib/Seaborn: Visualização de dados	Simplicidade, ampla coleção de bibliotecas, comunidade e xtensa	• Diagnóstico de pacientes • Segmentação de pacientes • Análise de sobrevida • Tendências de saúde
R	Voltada à estatística e ciência de dados	Estatística, modelagem de dados, visualização	• dplyr/tidyr: Manipulação e limpeza de dados • ggplot2: Visualização avançada • Caret: Integração de técnicas de ML	Ferramenta líder em análises estatísticas, visualizações poderosas	• Estudos clínicos • Análise epidemiológica • Modelagem de dados de saúde pública
TensorFlow	Framework de ML (Python)	Redes neurais, deep learning, processamento de imagens	• TensorFlow Core: Construção de redes neurais • TensorFlow Lite: ML em dispositivos móveis • TFX: Produção de pipelines de ML	Suporte para produção em larga escala, flexibilidade em redes neurais profundas	• Classificação de doenças a partir de imagens • Detecção de anomalias em exames • Previsão de desfechos clínicos
PyTorch	Framework de ML (Python)	Redes neurais, deep learning, NLP, processamento de imagens	• TorchVision: Processamento de imagens • TorchScript: Produção e otimização de modelos	Interface dinâmica e intuitiva, flexibilidade em pesquisa	• Extração de informações de textos clínicos • Processamento de imagens médicas • Modelos NLP para análise de prontuários

Fonte: autoria própria.

O ***TensorFlow*** é um *framework* de código aberto para aprendizado de máquina e redes neurais desenvolvido pelo Google®. Útil no desenvolvimento de modelos de *deep learning* com a criação e treinamento de redes neurais profundas e detecção de anomalias em imagens médicas, classificação de doenças a partir de dados de imagem e previsão de desfechos clínicos. O *TensorFlow Lite* é a versão leve do *TensorFlow* para dispositivos móveis e embarcados e apoia a implementação de mo-

delos de aprendizado de máquina em dispositivos móveis para monitoramento de saúde e diagnósticos portáteis. O *TensorFlow Extended* (TFX) é uma plataforma para produção de modelos de aprendizado de máquina e apoia a implementação de pipelines de aprendizado de máquina em larga escala para análise de dados de saúde.

PyTorch é um *framework* de aprendizado de máquina de código aberto desenvolvido pelo *Facebook*, conhecido por sua flexibilidade e facilidade de uso. Útil para desenvolvimento de modelos de *deep learning* com a criação e treinamento de redes neurais com uma interface dinâmica. Na análise de imagens ocorre o desenvolvimento de modelos de NLP para extração de informações de textos clínicos. *TorchVision* é uma biblioteca complementar que fornece ferramentas para o processamento de imagens com pré-processamento de imagens médicas, como normalização, transformação e aumentação de dados. *TorchScript* permite a serialização e otimização de modelos *PyTorch* para produção e implementação de modelos de aprendizado de máquina em sistemas de produção de saúde.

4.6 Saiba Mais - Atividade de Leitura Opcional

4.6.1 *Blockchain*

A tecnologia *blockchain*, originalmente desenvolvida como a base para criptomoedas como o Bitcoin, tem se mostrado uma ferramenta promissora para uma variedade de aplicações além do setor financeiro, incluindo a gestão de RES, também conhecido como EHR, e o PEP. A aplicação de *blockchain* ao PES envolve desafios associados à segurança, privacidade, integridade e interoperabilidade dos dados de saúde.

Blockchain é, por definição, descentralizado (federado) ou distribuído. Em vez de armazenar dados em um servidor centralizado, a *blockchain* distribui cópias do registro pelos nós da rede. Isso significa que não há apenas um ponto de falha, tornando os dados mais resistentes a ataques cibernéticos. Cada nó na rede possui uma cópia completa da cadeia de blocos, o que aumenta a segurança e a confiabilidade dos dados armazenados.

Outro aspecto é a imutabilidade da *blockchain*. Uma vez que um bloco de dados é registrado e adicionado à cadeia, ele não pode ser alterado ou apagado sem que isso seja evidente para todos os participantes da rede. Essa característica garante a integridade dos registros de saúde, fornecendo um histórico auditável e confiável de todas as interações com os dados do paciente.

A transparência é uma característica inerente à *blockchain*, sendo um insumo para assegurar a privacidade. Na *blockchain*, cada transação ou interação com o registro pode ser visível para todos os participantes da rede. No entanto, com a aplicação de criptografia avançada e técnicas de anonimização, é possível proteger a identidade dos pacientes e garantir que apenas partes autorizadas possam acessar informações sensíveis. Isso pode ser conseguido utilizando chaves privadas e públicas para controlar o acesso aos dados.

Interoperabilidade é outro benefício nativo do uso de *blockchain* em prontuários eletrônicos. Os sistemas de saúde frequentemente enfrentam dificuldades na troca de informações entre diferentes plataformas e prestadores de serviços devido a formatos de dados incompatíveis e falta de padrões comuns. A *blockchain* pode atuar como uma camada de interoperabilidade, permitindo que diferentes sistemas acessem e compartilhem dados de maneira segura. Com o uso de contratos inteligentes – programas autoexecutáveis que operam na *blockchain* – é possível automatizar processos de consentimento e autorização para o compartilhamento de dados entre diferentes entidades de saúde.

A tecnologia *blockchain* traz recursos de rastreabilidade dos dados. Cada interação com um prontuário eletrônico é registrada em um bloco, criando um trilho auditável de acessos e alterações. Isso não só ajuda a prevenir fraudes e manipulações, mas também facilita a conformidade com regulamentações de proteção de dados, como a *General Data Protection Regulation* (GDPR) na Europa ou a LGPD no Brasil, que exigem transparência e responsabilidade na gestão de informações pessoais.

Em termos de aplicação prática, a *blockchain* pode ser integrada aos prontuários existentes via APIs e outras interfaces de *software*. Isso permite que os benefícios da *blockchain* sejam aproveitados sem a necessidade de substituir completamente as infraestruturas de TI atuais. Além disso, instituições de saúde podem adotar uma abordagem gradual, começando com pilotos e provas de conceito antes de uma implementação em larga escala.

A aplicação de *blockchain* a prontuários eletrônicos de saúde traz inovações em segurança, privacidade, integridade e interoperabilidade dos dados de saúde. A descentralização, imutabilidade, transparência equilibrada com privacidade, interoperabilidade e rastreabilidade são componentes que tornam a *blockchain* uma tecnologia promissora substituir a atual estratégia fragmentadora em que os registros eletrônicos ficam majoritariamente distribuídos nos estabelecimentos e as estratégias de centralização não retornam os dados após tratados aos entes federativos e estabelecimentos.

4.6.2 *Blockchain* na Rede Nacional de Dados em Saúde

Embora o uso da *blockchain* esteja incerto no SUS no cenário atual, a RNDS foi estruturada para implantar nacionalmente a tecnologia (Brasil, 2020b). A RNDS buscou promover a interoperabilidade dos prontuários por meio de uma estrutura de *blockchain* compartilhada entre os estados brasileiros.

A *blockchain* na RNDS armazenaria a história de interações entre pacientes e agentes de saúde, garantindo a imutabilidade e verificação do conteúdo dos registros. Foi previsto que sempre que um paciente consulta um profissional de saúde, este pode acessar os dados do paciente mediante autorização, ou automaticamente em casos de emergência. Contratos inteligentes foram usados para gerir informações digitais e assegurar as regras de negócio, garantindo que apenas profissionais autorizados tenham acesso, conforme a LGPD.

Utilizando o padrão FHIR, a infraestrutura definida pelo Ministério da Saúde previu, adicionalmente, utilizar *data analytics* e IA num lago de dados (Brasil, 2021b). Foi estabelecida a existência de APIs abertas para que *softwares* de saúde integrem-se à RNDS, enquanto aplicativos desenvolvidos pelo Ministério da Saúde facilitam o acesso direto por cidadãos e profissionais de saúde.

A RNDS adotou uma arquitetura permissionada, onde apenas o MS, secretarias de saúde e futuros participantes privados têm permissão para participar da *blockchain,* com responsabilidade tripartite do SUS pela governança da rede. Testes indicam que a solução pode suportar até 1.800 transações por segundo, o suficiente para atender à população usuária do SUS. Essa adoção tem o potencial de melhorar o atendimento ao paciente, reduzir custos operacionais e garantir a integridade e segurança dos dados clínicos, impossibilitando a alteração de registros passados e prevenindo fraudes e falhas.

4.6.3 Diferença entre Estatística e Aprendizado de Máquina

É importante salientar a diferença metodológica entre estatística e aprendizado de máquina, onde a primeira parte de hipóteses e conjunto controlado de variáveis, enquanto a segunda é exploratória e voltada para a automação. Vamos usar como exemplo a técnica ARIMA (*Autoregressive Integrated Moving Average*) é um modelo utilizado para análise e previsão de séries temporais. ARIMA lida sequências temporais para modelar e prever valores futuros baseando-se em valores passados da mesma série.

Os princípios são autorregressivos por usar valores passados da série para prever o valor atual, integrado visto que envolve a diferenciação da série para torná-la estacionária (ou seja, remover tendências) e média móvel ao usar erros passados de previsões na modelagem.

Apesar da ARIMA incluir conceitos como previsão e modelagem, se baseia em pressupostos estatísticos, sendo considerado um modelo estatístico clássico, ao invés de um método típico de ML voltado para o aprendizado de relações complexas e padrões a partir de grandes conjuntos de dados, sem a necessidade de muitas suposições sobre a forma da relação entre variáveis. Por outro lado, ARIMA requer que a série temporal seja estacionária e certas suposições sobre a distribuição dos erros sejam atendidas. Embora a ARIMA não seja um método de ML, pode ser adotada em conjunto com técnicas de ML. Por exemplo, as previsões da ARIMA podem servir como variáveis de entrada para modelos de aprendizagem de máquina mais complexos, ou pode ser aprimorado utilizando técnicas de ML para modelagem dos resíduos da ARIMA.

4.6.4 pySUS

Conheça ferramentas para utilizar os dados do SUS.

O PySUS é uma coleção de códigos auxiliares para baixar e analisar dados do DataSUS (Brasil). A documentação inclui fontes de dados como o CNES, Sinan, Sinasc, SIM, SIA e SIH, além de tutoriais para pré-processamento de dados do DataSUS, informações sobre dengue, zika, chikungunya, dados do Instituto Brasileiro de Geografia e Estatística (IBGE), entre outros. Esse projeto, desenvolvido por Flávio Codeco Coelho, é construído com Sphinx usando um tema fornecido pelo *Read the Docs* (“Pysus”, [s.d.]).

O pacote “microdatasus” para o R disponibiliza funções para baixar e pré-processar os arquivos de microdados do DataSUS no formato DBC. Ele atribui e trata os rótulos e formatos das variáveis durante o pré-processamento dos dados. O uso do pacote envolve principalmente duas funções: uma para baixar os dados e outra para o pré-processamento. Os SIS suportados incluem SIM, SINASC, SIH, CNES, SIA, SINAN-DENGUE, SINAN-CHIKUNGUNYA, SINAN-ZIKA e SINAN-MALARIA. É importante destacar que o desenvolvimento desse pacote teve a colaboração do pacote read.dbc, criado por Daniela Petruzalek. Leia mais em: Saldanha, Bastos e Barcellos (2019) e Saldanha, Pedroso e Magalhães (2023).

4.6.5 Salas de Situação em Saúde

Você conhece o conceito de **Salas de Situação em Saúde**, atualmente conhecidas como salas de comando ou centros de inteligência? A Sala de Situação em Saúde é um espaço físico ou virtual onde uma equipe técnica analisa sistematicamente informações para caracterizar a saúde de uma população específica. As salas funcionam como centros de inteligência em saúde, promovendo uma visão intersetorial e integrada. Entre suas principais funções estão o planejamento e a avaliação de ações, definição de políticas de saúde, vigilância sanitária e a resposta a emergências, como surtos e desastres naturais. A coleta de dados relevantes abrange não apenas aspectos epidemiológicos, mas também informações socioeconômicas e demográficas, permitindo uma compreensão mais clara da realidade dos serviços de saúde e das necessidades da população.

As salas de situação apoiam a tomada de decisões e planejamento em saúde no nível municipal, estadual e nacional, integrando informações para diagnósticos dinâmicos que atendem às especificidades de cada local. A criação de Salas de Situação possibilita uma resposta a situações emergenciais, promovendo a transparência nas ações de saúde e contribuindo para a melhoria do Sistema de Saúde. Existem diversas iniciativas que apoiam a tomada de decisão, oferecendo apoio por meio de tutoriais e ferramentas que subsidiam o planejamento e a avaliação de ações em saúde, garantindo um compromisso com um sistema de saúde universal e equânime. Conheça algumas delas (Deininger *et al*., 2014; Ferré *et al.*, 2020; Moya; Risi Junior; Martinello, 2010).

4.6.6 Aspectos Regulatórios em Curso e Impactos em Pesquisa

O gestor de dados deve estar atento ao cenário normativo e conhecer as ações em curso, como o Projeto de Lei sobre a RNDS e sua federalização (Brasil, 2023c); e a discussão da judicialização da saúde no Supremo Tribunal Federal, ao longo de 2023 e 2024. Em curso encontra-se, também, a regulamentação da IA, com o Projeto de Lei nº 2338, de 2023.

Conheça alguns aspectos que podem impactar nas pesquisas nacionais na área:

- Custos e infraestrutura:
  - **Recursos limitados**: Pesquisadores em instituições públicas ou sem financiamento privado podem enfrentar dificuldades para adquirir e manter a infraestrutura tecnológica necessária para cumprir as exigências legais e de segurança dos dados.

- **Investimento em segurança**: A necessidade de implementar medidas rigorosas de segurança e controle de acesso pode representar um custo significativo para instituições com orçamentos limitados.

- Capacitação e conhecimento técnico:
  - **Necessidade de especialização** para manipulação segura de dados e o uso de IA exigem conhecimentos especializados. Instituições públicas e sem financiamento podem ter dificuldades em contratar ou treinar pessoal qualificado.
  - **Anonimização e pseudonimização, desafios técnicos e custos associados**: Anonimizar ou pseudonimizar dados populacionais de saúde é um processo tecnicamente complexo e custoso, o que pode ser um desafio significativo para pesquisadores sem financiamento privado.
  - **Cumprimento das normas, burocracia e procedimentos**: O cumprimento das normas estabelecidas pela nova resolução e pelo Projeto de Lei nº 2338 pode aumentar a carga burocrática e administrativa, exigindo tempo e recursos que poderiam ser direcionados para a pesquisa em si.
  - **Transferência e compartilhamento de dados, restrições e procedimentos**: As exigências para transferência de dados, incluindo autorização, formalização e medidas de segurança, podem dificultar a colaboração entre instituições e pesquisadores, especialmente em contextos multicêntricos.
  - **Direitos dos participantes e gestão de consentimento**: A obtenção, gestão e possível reobtenção de consentimento dos participantes para o uso de seus dados pode ser um processo complexo e oneroso, especialmente para estudos de larga escala com populações vulneráveis.

- Definições acrescentadas pelo Projeto de Lei nº 2338, de 2023:
  - **Sistema de IA (Art. I):** Sistema computacional autônomo que utiliza aprendizagem de máquina e/ou lógica e representação do conhecimento para produzir previsões, recomendações ou decisões.
  - **Fornecedor de sistema de IA (Art. II):** Pessoa natural ou jurídica que desenvolve um sistema de IA, visando sua colocação no mercado ou aplicação em serviço, sob seu próprio nome ou marca.
  - **Operador de sistema de IA (Art. III):** Pessoa natural ou jurídica que utiliza um sistema de IA em seu nome ou benefício, exceto quando utilizado em atividade pessoal de caráter não profissional.
  - **Agentes de IA (Art. IV):** Refere-se tanto aos fornecedores quanto aos operadores de sistemas de IA.

- **Autoridade competente (Art. V):** Órgão ou entidade da Administração Pública Federal responsável por fiscalizar o cumprimento da lei.
- **Discriminação (Art. VI):** Qualquer distinção, exclusão, restrição ou preferência que anule ou restrinja o reconhecimento ou exercício de direitos, baseada em características pessoais como origem, raça, gênero, orientação sexual, entre outras.
- **Discriminação indireta (Art. VII):** Discriminação resultante de normativas ou práticas aparentemente neutras que causam desvantagem para grupos específicos, a menos que justificadas de forma razoável e legítima.
- **Mineração de textos e dados (Art. VIII):** Processo de extração e análise de grandes quantidades de dados ou conteúdo textual para identificar padrões e correlações relevantes para sistemas de IA.

- Impacto e dificuldades adicionais:
  - **Transparência e informações e direito à informação prévia (Art. 5º, I)**: Pesquisadores devem garantir que as pessoas afetadas por sistemas de IA sejam informadas previamente sobre suas interações com esses sistemas, o que pode aumentar a complexidade na gestão de dados e na comunicação com os participantes.
  - **Responsabilidade, conformidade e responsabilidade legal**: os pesquisadores, especialmente em instituições públicas, precisarão garantir a conformidade com a legislação tanto em relação à proteção de dados (LGPD) quanto ao uso de IA, adicionando uma camada extra de responsabilidade e potencialmente exigindo novos processos e documentações.
  - **Prevenção de discriminação para evitar discriminação direta e indireta**: sistemas de IA utilizados em pesquisa devem ser projetados e operados de forma a evitar discriminação direta e indireta, o que requer uma análise cuidadosa e contínua dos algoritmos e dados utilizados.
- Documentação e transparência para sistemas de alto risco (Art. 20):
  - **Documentação técnica**: manter documentação detalhada do funcionamento do sistema e das decisões tomadas durante seu desenvolvimento e uso (Art. 20, I).
  - **Registro automático**: utilizar ferramentas de registro automático para avaliar a acurácia e robustez do sistema e identificar vieses potenciais (Art. 20, II).
  - **Testes de confiabilidade**: realizar testes para avaliar a confiabilidade do sistema, incluindo robustez, acurácia, precisão e cobertura (Art. 20, III).

- Gestão de dados e vieses:
  - **Avaliação de vieses**: avaliar dados com medidas de controle de vieses cognitivos humanos e evitar a geração de vieses sociais estruturais (Art. 20, IV).
  - **Diversidade da equipe**: assegurar que a equipe responsável pela concepção e desenvolvimento do sistema seja inclusiva e diversa (Art. 20, IV).
- Explicabilidade dos resultados:
  - **Disponibilização de informações**: Implementar medidas para explicar os resultados dos sistemas de IA e fornecer informações adequadas sobre o funcionamento do modelo (Art. 20, V).
- Implementação de programas de governança (Art. 30, § 2º):
  - **Comprometimento com Normas e Boas Práticas:** Demonstrar comprometimento em adotar processos e políticas internas que assegurem o cumprimento de normas e boas práticas relativas à não maleficência e proporcionalidade entre métodos e finalidades dos sistemas de IA.
  - **Adaptação à Estrutura e Operações**: O programa deve ser adaptado à estrutura, escala, volume de operações e potencial danoso do agente.
  - **Transparência e Participação**: Estabelecer uma relação de confiança com as pessoas afetadas, atuando de forma transparente e assegurando mecanismos de participação.
  - **Supervisão Interna e Externa**: Integrar a governança geral da organização com mecanismos de supervisão internos e externos.
  - **Planos de Resposta**: Contar com planos de resposta para reversão de possíveis resultados prejudiciais dos sistemas de IA.
  - **Atualização Contínua**: Manter o programa atualizado com base em informações obtidas de monitoramento contínuo e avaliações periódicas.

O Projeto de Lei nº 2338, de 2023, juntamente com a nova resolução do Conselho Nacional de Saúde referente ao uso de bancos de dados e a LGPD, estabelecem um conjunto de requisitos para a proteção de dados e o uso de IA em pesquisas. Para pesquisadores públicos e sem financiamento privado, as principais dificuldades incluem a necessidade de investimento em infraestrutura e segurança, a capacitação técnica, a gestão de consentimento e a conformidade com normas complexas.

A adição de definições específicas para sistemas de IA e os direitos associados aumentam a responsabilidade para esses pesquisadores, exigindo uma abordagem integrada e bem planejada para garantir a conformidade e a proteção dos dados dos participantes.

4.6.7 Tecnologias em Nuvem e Outras Ferramentas

O gestor de dados deve conhecer as formas de desenvolvimento a partir das ciências de arquitetura e engenharia de *software*. Atualmente, as metodologias ágeis de entregas rápidas e segmentadas, abordadas anteriormente, são acompanhadas por outras tendências no desenvolvimento de *software* que é a adoção de microsserviços em oposição às soluções monolíticas, as quais desenvolviam toda a aplicação e a entrevam o mais completa possível.

A arquitetura de **microsserviços** constitui-se de pequenos serviços independentes que se comunicam entre si. Cada microsserviço é responsável por uma funcionalidade específica da aplicação e pode ser desenvolvido, implantado e escalado de forma independente, promovendo agilidade e resiliência, uma vez que o não funcionamento de um microsserviço não implica necessariamente na interrupção da aplicação. Assim, é viabilizada a adoção de tecnologias diversificadas e a integração contínua, permitindo que equipes distintas trabalhem em diferentes partes do sistema de forma paralela. Além disso, a modularidade intrínseca dos microsserviços contribui para uma manutenção mais simplificada e uma maior capacidade de resposta às mudanças.

Acima falamos sobre gestão de dados que não são necessariamente produzidos em larga escala. Porém, existem soluções para *Big Data*, usualmente quando são processados acima de bilhões de registros e centenas de atributos, por exemplo, Hadoop, Apache Spark, Apache Storm, RapidMiner, Greenplum e Tableau Hyper (para ingestão rápida de dados). Existem também soluções em nuvem(“Top 6 cloud data warehouse solutions in 2024 [compared]”, [s.d.]):

- ***Azure Synapse Analytics*** para armazenamento de dados empresariais. Azure Synapse Analytics é ideal para integrar dados de centenas de fontes em várias divisões e subsidiárias da empresa, permitindo consultas analíticas serem realizadas em segundos. Os relatórios em todos os níveis de gestão, desde executivos até diretores, gerentes e supervisores, são protegidos com um controle de acesso a dados detalhados.
- ***Amazon Redshift*** para armazenamento de grandes volumes de dados. Amazon Redshift permite consultas SQL em exabytes de dados estruturados, semi-estruturados e não estruturados no armazém de dados, em armazenamentos

de dados operacionais e em um data lake, com a possibilidade de agregar dados com serviços de análise de big data e aprendizado de máquina.

- ***Google BigQuery*** para armazenamento de grandes volumes de dados com consultas infrequentes. *BigQuery* permite um armazenamento em escala de exabytes de forma econômica, com tabelas que podem ter até 10.000 colunas. É mais eficaz quando as principais consultas analíticas filtram dados conforme o particionamento ou clusterização, ou requerem a varredura de todo o conjunto de dados.
- ***Azure SQL Database*** para armazenamento de dados de porte médio. O banco de dados Azure SQL é adequado para cenários de armazenamento de dados com volumes de até 8 TB e muitos usuários ativos (as solicitações simultâneas podem chegar a até 6.400, com até 30.000 sessões simultâneas).
- ***Snowflake*** para armazenamento de dados independente de nuvem. Oferecido como *Software* como Serviço, o Snowflake permite que as empresas aloquem simultaneamente recursos computacionais de diferentes fornecedores de nuvem (AWS, Azure, GCP) no mesmo banco de dados para carregar e consultar dados sem impactar o desempenho do armazém de dados.
- ***Azure Cosmos DB*** para armazenamento de dados operacionais (processamento híbrido transacional/analítico). Azure Cosmos DB e *Azure Synapse Analytics* permitem que equipes empresariais executem consultas rápidas e econômicas sem ETL em grandes conjuntos de dados operacionais em tempo real, sem copiar dados e sem impactar o desempenho das cargas de trabalho transacionais da empresa.

A tradução do conhecimento é fundamental para comunicar o que se pretende informar com dados. Conheça relatos sobre painéis interativos e visualização de dados (Ferré *et al*., 2021; Schrarstzhaupt *et al*., 2024).

Existem técnicas para coleta de dados, sobretudo quando não são usadas boas práticas na disseminação, a despeito dos dados estarem disponíveis em páginas da internet com acesso livre. *Web scraping* é um processo de extrair dados de sites da *web*. Existem diferentes tipos e técnicas de *web scraping*, cada uma adequada para situações específicas.

- *Scraping Manual*: O processo de copiar e colar dados manualmente de um site. Útil para tarefas muito simples ou de uma única vez, onde a automação não é viável.
- *Web Scraping* Automatizado: Uso de scripts ou ferramentas automatizadas para extrair dados de sites. As ferramentas mais comuns são Python (com bibliotecas como Beautiful Soup, Scrapy, Selenium, Puppeteer e Octoparse.

- Análise de HTML é necessária para extrair dados diretamente do HTML de uma página da web. As ferramentas que se destacam são Beautiful Soup (Python), Cheerio (JavaScript). Assim, o objetivo é parsear um documento HTML para extrair títulos de artigos, links, textos, etc. Parse é o processo de analisar e interpretar uma sequência de dados, como texto ou código, para extrair informações estruturadas. Em programação, envolve decompor entradas de dados em componentes significativos, permitindo que um sistema compreenda e manipule esses dados de forma adequada.

- A automação de navegador é um recurso para interagir com a web como um usuário real. As ferramentas disponíveis são Selenium e Puppeteer. Úteis para sites que requerem *login*, navegação por várias páginas, ou interações complexas como cliques em botões.

- É comum sítios fornecerem APIs oficiais para acessar dados de forma estruturada e eficiente. Exemplo: Twitter API, Google Maps API. A extração é mais estável e menos propensa a mudanças frequentes do que o HTML scraping. Parsing de XML e JSON são conhecimentos necessários para extração de dados de APIs ou arquivos que retornam dados em formato XML ou JSON. As ferramentas usuais são Requests e json (*Python*), Axios (JavaScript).

- Scraping de Sites Dinâmicos para extrair dados de sites que carregam conteúdo dinamicamente via JavaScript com ferramentas como Selenium, Puppeteer, Beautiful Soup (com *requests-html*). Deve-se executar JavaScript para carregar a página totalmente antes de extrair os dados.

Verifique os termos de serviço do site para garantir que o scraping não viola suas políticas. Evite sobrecarregar o servidor do site. Use atrasos entre requisições e considere técnicas de caching. Sites podem usar medidas como CAPTCHAs, bloqueio de IPs, ou mudanças frequentes no layout para impedir scraping. Ferramentas como Selenium podem ajudar a contornar algumas dessas medidas. Certifique-se de que o *scraping* esteja dentro dos limites legais, respeitando direitos autorais e privacidade de dados. *Web scraping* frequentemente requer manutenção contínua, pois mudanças no *layout* do site podem quebrar os scripts. Leia mais em: Morais *et al.* (2021).

Ferramentas úteis para quem deseja criar conteúdo digital de forma intuitiva, sem a necessidade de conhecimentos avançados em programação (WYSIWYG):

- *Adobe Dreamweaver*: Um editor de HTML e CSS que oferece uma interface visual e suporte para desenvolvimento responsivo.

- *WordPress*: Com seu editor de blocos, permite que os usuários criem páginas e postagens visualmente, com a opção de editar o código HTML.

- *Wix*: Uma plataforma de criação de *sites* que permite arrastar e soltar elementos para construir páginas da *web*.

- *Squarespace*: Oferece uma interface intuitiva para criar sites com *design* elegante, sem necessidade de programação.
- *Webflow*: Combina *design* visual com a capacidade de exportar código limpo, ideal para designers que desejam controle sobre o *layout*.
- *Froont*: Uma ferramenta de design responsivo que permite criar *layouts* de *sites* visualmente, com foco em *designers*.
- *TinyMCE*: Um editor de texto WYSIWYG que pode ser integrado em aplicativos *web*, permitindo edição de texto rica.
- *CKEditor*: Outro editor de texto WYSIWYG que oferece uma interface amigável para edição de conteúdo em aplicações *web*.

AKCIT
CENTRO DE COMPETÊNCIA EMBRAPII EM TECNOLOGIAS IMERSIVAS

Governança em Saúde Digital

Unidade V
Encerramento

Unidade V: Encerramento

Gestão de dados é uma disciplina que proporciona instrumental para administrar informações, assegurar a reprodutibilidade de processos e a comunicação. A gestão de dados envolve a identificação, armazenamento, acesso, compilação, proteção e uso de dados para fortalecer a estratégia organizacional. Dados são sequências de símbolos quantificados, enquanto informações são a interpretação desses dados. A governança de dados, sendo o objetivo da gestão, garante decisões assertivas utilizando tecnologias avançadas.

Implantar uma gestão de dados e informações pode reduzir problemas de comunicação entre usuário, profissional e o sistema de saúde, garantindo que as informações certas cheguem às pessoas certas no momento certo. Como resultado de uma boa gestão ocorre a coordenação dos cuidados ao paciente, evitando duplicidade de dados ou coletas desnecessárias e reduzindo erros iatrogênicos. Além disso, contribui para a segurança das informações, estimulando políticas públicas informadas por evidência e protegendo contra possíveis conflitos de interesse.

Acesso, qualidade, integração, visualização, acessibilidade, governança e disseminação são componentes da gestão de dados. Profissionais chave nesta área incluem CIO, cientista de dados, administrador de banco de dados, analista de dados e gestor de dados estratégicos ou encarregado de dados. A adoção de um *software* de gestão de dados facilita a captura, armazenamento, organização e análise de grandes volumes de dados.

Implementar uma gestão de dados envolve integrar esta cultura na instituição, realizar diagnósticos iniciais, definir responsabilidades, traçar estratégias, garantir a integração entre setores, fornecer treinamentos e monitorar resultados.

Referências

Associação Brasileira de Normas Técnicas - ABNT. **NBR ISO/IEC 17799 - Tecnologia da informação - Código de prática para a gestão da segurança da informação**. Associação Brasileira de Normas Técnicas, 2001.

ABUHELWA, A. Y. *et al.*. A clinical scoring tool validated with machine learning for predicting severe hand-foot syndrome from sorafenib in hepatocellular carcinoma. **Cancer Chemotherapy and Pharmacology**, v. 89, n. 4, p. 479–485, abr. 2022. Disponível em: <http://dx.doi.org/10.1007/s00280-022-04411-9>. Acesso em: 22 out. 2024.

Centro Colaborador do SUS Avaliação Tecnológica e Excelência em Saúde (CCATES). **Activities**. Disponível em: <https://www.ccates.org.br/en/activities/>. Acesso em: 31 jul. 2024.

ALCOFORADO, L. F.. **Utilizando a linguagem R: conceitos, manipulação, visualização, modelagem e elaboração de relatórios**. [S.l.]: Alta Books, 2021.

ALMEIDA, B.. Um debate sobre dados pessoais e dados pessoais sensíveis para pesquisa científica e para pesquisa em saúde pública a partir da Lei Geral de Proteção de Dados Pessoais. In: ARAGÃO, E. S. *et al.* (Org.). **Avaliação de impacto das políticas de saúde um guia para o SUS**. Brasília: Ministério da Saúde, 2023. Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/avaliacao_impacto_politicas_saude_guia_sus.pdf>. Acesso em: 22 out. 2024.

ALTMAN, Naomi; KRZYWINSKI, Martin. The Curse(s) of Dimensionality. **Nature Methods**, v. 15, n. 6, p. 399–400, jun. 2018. Disponível em: <http://dx.doi.org/10.1038/s41592-018-0019-x>.

ANDERSEN, H. C.. **The Dryad**. [S.l.]: Lindhardt og Ringhof, 2021. Disponível em: <https://play.google.com/store/books/details?id=SR8iEAAAQBAJ>. Acesso em: 22 out. 2024.

ARORA, S. *et al.*. Digitization of Health Insurance Documents for The Cashless Claim Settlement Using Intelligent Document Management System. **Procedia computer science**, v. 235, p. 1319–1331, 1 jan. 2024. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050924008019>. Acesso em: 22 out. 2024.

ARRUDA, C.; LOPES, S. G. R.; KOERICH, M.. Redes de atenção à saúde sob a luz da teoria da complexidade. **Escola Anna Nery Revista de Enfermagem**, 2015. Disponível em: <https://pesquisa.bvsalud.org/saudepublica/resource/pt/lil-741495>. Acesso em: 22 out. 2024.

BALL, R.; DAL PAN, G.. “Artificial Intelligence” for Pharmacovigilance: Ready for Prime Time? **Drug Safety: An International Journal of Medical Toxicology and Drug Experience**, v. 45, n. 5, p. 429–438, maio 2022. Disponível em: <http://dx.doi.org/10.1007/s40264-022-01157-4>. Acesso em: 22 out. 2024.

BARRETO, M. L. *et al*.. Cohort Profile: The 100 Million Brazilian Cohort. **International Journal of Epidemiology**, v. 51, n. 2, p. e27–e38, maio 2022. Disponível em: <http://dx.doi.org/10.1093/ije/dyab213>. Acesso em: 22 out. 2024.

BHAVNANI, S. P. *et al*.. 2017 Roadmap for Innovation-ACC Health Policy Statement on Healthcare Transformation in the Era of Digital Health, Big Data, and Precision Health: A Report of the American College of Cardiology Task Force on Health Policy Statements and Systems of Care. **Journal of the American College of Cardiology**, v. 70, n. 21, p. 2696–2718, 28 nov. 2017. Disponível em: <https://doi.org/10.1016/j.jacc.2017.10.018>. Acesso em: 22 out. 2024.

BONAT, W.H.. Manipulando dados no R. In: ARAGÃO, E. S. *et al*. (Org.). **Avaliação de impacto das políticas de saúde: um guia para o SUS**. Brasília: Ministério da Saúde, 2023.

BRASIL. **Ações para a adequação da RNDS à Lei Geral de Proteção de Dados ConecteSUS**. Brasília: Ministério da Saúde, 2020a. Disponível em: <https://www.gov.br/saude/pt-br/composicao/seidigi/publicacoes/adquecao-da-rnds-a-lgpd.pdf>.

BRASIL. *Sistema de Informações e Gestão da Assistência à Saúde: Demais Macrorregiões*. Ministério da Saúde, 2024f. Disponível em: https://infoms.saude.gov.br/extensions/SEIDIGI_DEMAS_MACRORREGIOES/SEIDIGI_DEMAS_MACRORREGIOES.html. Acesso em: 01 out. 2024.

BRASIL. **A experiência brasileira em sistemas de informação em saúde: Produção e disseminação de informações sobre saúde no Brasil**. Brasília: Editora MS, 2009a. (Textos Básicos de Saúde).

BRASIL. **Aplicações blockchain no setor público do Brasil - apêndice 1**. [S.l.]: Tribunal de Contas da União (TCU), 2020b. Disponível em: <https://portal.tcu.gov.br/data/files/58/02/CE/5E/C4854710A7AE4547E18818A8/Blockchain_apendice1.pdf>. Acesso em: 4 ago. 2024. (Conteúdo relacionado ao Acórdão 1.613/2020-TCU-Plenário, sob relatoria do Ministro Aroldo Cedraz).

BRASIL. **BR-Core - SIMPLIFIER.NET**. Disponível em: <https://simplifier.net/br-core>. Acesso em: 8 jul. 2024a.

BRASIL. **Cadastro Nacional dos Estabelecimentos de Saúde do Brasil - CNES**. Disponível em: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?cnes/cnv/estabbr.def>. Acesso em: 24 dez. 2023b.

BRASIL. CNES - Cadastro Nacional de Estabelecimentos de Saúde: Estatísticas por Abrangência. Ministério da Saúde, 2024g. Disponível em: http://tabnet.datasus.gov.br/cgi/deftohtm.exe?cnes/cnv/estabbr.def. Acesso em: 01 out. 2024.

BRASIL. **CONITEC**. Disponível em: <https://www.gov.br/conitec/pt-br>. Acesso em: 16 out. 2022a.

BRASIL. **Data Mining and Machine Learning: Fundamental Concepts and Algorithms**. [S.l.]: Cambridge University Press, 2020b. Disponível em: <https://play.google.com/store/books/details?id=oafDDwAAQBAJ>. Acesso em: 22 out. 2024.

BRASIL. Data Sources for Drug Utilization Research in Brazil-DUR-BRA Study. **Frontiers in Pharmacology**, v. 12, p. 789872, 2021b. Disponível em: <http://dx.doi.org/10.3389/fphar.2021.789872>. Acesso em: 22 out. 2024.

BRASIL. **DeCS/MeSH Descritores em Ciências da Saúde**. Disponível em: <https://decs.bvsalud.org/>. Acesso em: 28 out. 2022a.

BRASIL. **DIANA - Evolução e fortalecimento da Rede Nacional de Dados em Saúde com ênfase em ciência de dados**. Disponível em: <https://www.proadi-sus.org.br/projeto/evolucao-e-fortalecimento-da-rede-nacional-de-dados-em-saude-com-enfase-em-ciencia-de-dados>. Acesso em: 5 jan. 2024b.

BRASIL. **Doença de Chagas Aguda - Casos confirmados notificados no Sistema de Informação de Agravos de Notificação - SINAN**. Disponível em: <http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sinannet/cnv/Chagasbr.def>. Acesso em: 31 mar. 2022c.

BRASIL. **Elaboração de Plano de Dados Abertos**. [S.l.]: Enap Escola Nacional de Administração Pública, 2017a. Disponível em: <https://repositorio.enap.gov.br/bitstream/1/3152/1/M%C3%B3dulo%201%20-%20Conceitos%20de%20Dados%20Abertos.pdf>. Acesso em: 22 out. 2024.

BRASIL. **Estratégia de Consentimento Rede Nacional de Dados em Saúde**. 2023c. Disponível em: <https://www.gov.br/saude/pt-br/composicao/seidigi/publicacoes/>. Acesso em: 11 jan. 2023.

BRASIL. **Estratégia de Saúde Digital para o Brasil 2020-2028**. Brasília: Ministério da Saúde, 2020d. Disponível: https://bvsms.saude.gov.br/bvs/publicacoes/estrategia_saude_digital_Brasil.pdf. Acesso em: 22 out. 2024.

BRASIL. **Ferramenta de tabulação TabWin**. 2024a. Disponível em: <http://siab.datasus.gov.br/DATASUS/index.php?area=060805&item=3>. Acesso em: 4 jan. 2024.

BRASIL. **Guia de Vigilância em Saúde**. 5ª. ed. Brasília: Ministério da Saúde, 2022b. Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/guia_vigilancia_saude_5ed_rev_atual.pdf>. Acesso em: 22 out. 2024.

BRASIL. **Indicadores e Dados Básicos - IDB**. 2024a. Disponível em: <https://datasus.saude.gov.br/acesso-a-informacao/indicadores-e-dados-basicos/>. Acesso em: 7 jul. 2024.

BRASIL. Infoestrutura para apoio à decisão estratégica no SUS. In: SANTOS, A.O.; LOPES, L.T. (Org.). **Reflexões e futuro**. Coleção covid-19. Brasília: CONASS, 2021. v. 6. p. 114–127.

BRASIL. **Informações de Saúde**. 2024b. Disponível em: <https://datasus.saude.gov.br/informacoes-de-saude-tabnet/>. Acesso em: 7 jul. 2024.

BRASIL. Inteligência artificial e reações adversas relacionadas a medicamentos. In: ULHOA, C.; *et al.*(Org.). **Segurança do paciente em serviços de saúde: uma prioridade com múltiplas dimensões**. Conass documenta. Cadernos de informação técnica e memória do Conass. Brasília: Conass, 2023a. v. 46.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). **Diário Oficial da União**, 2018a. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Acesso em: 22 out. 2024.

BRASIL. **Lei nº 9.782, de 26 de janeiro de 1999. Define o Sistema Nacional de Vigilância Sanitária, cria a Agência Nacional de Vigilância Sanitária, e dá outras providências**. [S.l.]: Presidência da República, 1999. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/l9782.htm>. Acesso em: 22 out. 2024.

BRASIL. **LocalizaSUS - plataforma de dados estratégicos de saúde**. 2024d. Disponível em: <https://www.gov.br/saude/pt-br/composicao/seidigi/demas>. Acesso em: 7 jul. 2024.

BRASIL. **Metodologia de Desenvolvimento de Sofware - Diretriz Arquitetural**. Brasília: Ministério da Saúde, 2018b. Disponível em: <https://datasus.saude.gov.br/wp-content/uploads/2020/04/DATASUS_DiretrizArquitetural.pdf>. Acesso em: 22 out. 2024.

BRASIL. Modelagem e gestão de banco de dados com SQL e integração com o R. In: ARAGÃO, E. S. *et al.* (Org.). **Avaliação de impacto das políticas de saúde: um guia para o SUS**. Brasília: Ministério da Saúde, 2023b. Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/avaliacao_impacto_politicas_saude_guia_sus.pdf>. Acesso em: 22 out. 2024.

BRASIL. **Modelo da ficha de qualificação de Indicadores (FQI)**. 2024a. Disponível em: <https://www.ripsa.org.br/indicadores/fqi/>. Acesso em: 7 jul. 2024.

BRASIL. **Mortalidade**. Disponível em: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obt10uf.def>. 2022. Acesso em: 28 out. 2022.

BRASIL. **Norma de padronização de nomenclatura**. 2022b. Disponível em: <https://datasus.saude.gov.br/mad-norma-de-padronizacao-de-nomenclatura/>. Acesso em: 28 out. 2022.

BRASIL. **OpenDataSUS**. 2024a. Disponível em: <https://opendatasus.saude.gov.br/>. Acesso em: 7 jul. 2024.

BRASIL. **Painéis de Informações do Fundo Nacional de Saúde**. 2022. Disponível em: <https://painelms.saude.gov.br/extensions/Portal_Paineis/Portal_Paineis.html>. Acesso em: 28 out. 2022c.

BRASIL. **Pesquisa de Tecnologia da Informação e Comunicação TIC Saúde**. 2024b. Disponível em: <https://cetic.br/pt/pesquisa/saude/indicadores/>. Acesso em: 7 jul. 2024.

BRASIL. **Portal de Dados Abertos**. 2024b.Disponível em: <https://dados.gov.br/home>. Acesso em: 7 jul. 2024.

BRASIL. **Portal FNS**. 2024. Disponível em: <https://portalfns.saude.gov.br/>. Acesso em: 7 jul. 2024.

BRASIL. Portaria GM/MS Nº 1.434, de 28 de maio de 2020. Institui o Programa Conecte SUS e altera a Portaria de Consolidação nº 1/GM/MS, de 28 de setembro de 2017, para instituir a Rede Nacional de Dados em Saúde e dispor sobre a adoção de padrões de interoperabilidade em saúde. **Ministério da Saúde, Diário Oficial da União**, 2020. Disponível em: <https://bvsms.saude.gov.br/bvs/saudelegis/gm/2020/prt1434_01_06_2020_rep.html>. Acesso em: 22 out. 2024.

BRASIL. **Portaria GM/MS nº 545, de 16 de março de 2022. Dispõe sobre a Rede Interagencial de Informações para a Saúde (RIPSA)**. Brasília: Diário Oficial da União. Ministério da Saúde, 2022d. Disponível em: <https://bvsms.saude.gov.br/bvs/saudelegis/gm/2022/prt0545_21_03_2022.html#:~:text=Disp%C3%B5e%20sobre%20a%20Rede%20Interagencial,do%20par%C3%A1grafo%20%C3%BAnico%20do%20art.>. Acesso em: 22 out. 2024.

BRASIL. Portaria GM/MS nº. 859, de 20 de maio de 2015. Institui a Política Nacional de Informação e Informática em Saúde (PNIIS). **Diário Oficial da União**, 2015. Disponível em: <https://bvsms.saude.gov.br/bvs/saudelegis/gm/2015/prt0589_20_05_2015.html>. Acesso em: 22 out. 2024.

BRASIL. **Portaria Nº 221, de 17 de abril de 2008. Publica a Lista Brasileira de Internações por Condições Sensíveis à Atenção Primária**. [S.l.]: Ministério da Saúde, 2008c. Disponível em: <https://bvsms.saude.gov.br/bvs/saudelegis/sas/2008/prt0221_17_04_2008.html>. Acesso em: 22 out. 2024.

BRASIL. **Portaria Nº 4.279, de 30 de dezembro de 2010. Estabelece diretrizes para a organização da Rede de Atenção à Saúde no âmbito do Sistema Único de Saúde (SUS)**. [S.l.]: Ministério da Saúde, 2010b. Disponível em: <https://bvsms.saude.gov.br/bvs/saudelegis/gm/2010/prt4279_30_12_2010.html>. Acesso em: 22 out. 2024.

BRASIL. **Portaria SAES/MS Nº 234, de 30 de julho de 2022. Institui o Modelo de Informação Registro de Atendimento Clínico (RAC)**. 2022e. Disponível em: <https://bvsms.saude.gov.br/bvs/saudelegis/saes/2022/prt0234_20_07_2022.html>. Acesso em: 22 out. 2024.

BRASIL. **Procedimentos hospitalares do SUS por local de internação**. 2024. Disponível em: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sih/cnv/qial.def>. Acesso em: 7 jul. 2024.

BRASIL. **Relatório de Validação - Sistema de Informação em Saúde para a Atenção Básica (SISAB)**. Ministério da Saúde, 2024h. Disponível em: https://sisab.saude.gov.br/paginas/acessoRestrito/relatorio/federal/envio/RelValidacao.xhtml. Acesso em: 09 set. 2024.

BRASIL. Resolução Nº 659, de 26 de julho de 2021. Dispõe sobre a Política Nacional de Informação e Informática em Saúde (PNIIS). 2021e. Disponível em: <https://bvsms.saude.gov.br/bvs/saudelegis/cns/2022/res0659_15_06_2022.html>. Acesso em: 22 out. 2024.

BRASIL. **Resolução nº 738, de 01 de fevereiro de 2024.** Dispõe sobre uso de bancos de dados com finalidade de pesquisa científica envolvendo seres humanos. 2024. Disponível em: <https://conselho.saude.gov.br/resolucoes-cns/3316-resolucao-n-738-de-01-de-fevereiro-de-2024>. Acesso em: 6 ago. 2024.

BRASIL. Sala de Situação aberta com dados administrativos para gestão de Protocolos Clínicos e Diretrizes Terapêuticas de tecnologias providas pelo SUS. 15 set. 2020, [S.l.]: SBC, 15 set. 2020. p. 392–403. Disponível em: <https://sol.sbc.org.br/index.php/sbcas/article/view/11530>. Acesso em: 16 out. 2020.

BRASIL. **Sistema de Informação em Saúde para a Atenção Básica (SISAB)**. 2024. Disponível em: <https://sisab.saude.gov.br/>. Acesso em: 7 jul. 2024.

BRASIL. **Sistema IBGE de Recuperação Automática. Sidra - Banco de Tabelas Estatísticas**. 2024. Disponível em: <https://sidra.ibge.gov.br/home/pnadcm>. Acesso em: 7 jul. 2024.

BRASIL. **Substitutivo ao Projeto de Lei nº. 5.875, de 2013.** Dispõe sobre a Rede Nacional de Dados em Saúde -- RNDS, a Plataforma Conecte SUS, o Cadastro Nacional de Pessoas para a Saúde -- CadSUS e dá outras providências. 2023c. Disponível em: <https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2363812&filename=Tramitacao-PL%205875/2013>. Acesso em: 24 dez. 2023.

BRASIL. **TabNet - Informações de saúde. Estudo de Estimativas populacionais por município, sexo e idade - 2000-2021**. 2023f. Disponível em: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/cnv/popsvsbr.def>. Acesso em: 22 out. 2023.

BRASIL. **TabNet - Informações de saúde. Produção Ambulatorial do SUS por local de atendimento**. 2022d. Disponível em: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sia/cnv/qauf.def>. Acesso em: 28 out. 2022.

BRASIL. **Transferência de Arquivos – DATASUS**. 2024e. Disponível em: <https://datasus.saude.gov.br/transferencia-de-arquivos/>. Acesso em: 7 jul. 2024.

BURT, T. *et al.*. Strategic, Feasibility, Economic, and Cultural Aspects of Phase 0 Approaches: Is It Time to Change the Drug Development Process in Order to Increase Productivity? **Clinical and Translational Science**, v. 15, n. 6, p. 1355–1379, jun. 2022. Disponível em: <http://dx.doi.org/10.1111/cts.13269>. Acesso em: 22 out. 2024.

CARVALHO, W. S.; MOREIRA, A. M.; MAGALHÃES, S. M. S..Eventos adversos a medicamentos. **Acurcio FA, organizador. Medicamentos: políticas, assistência farmacêutica, farmacoepidemiologia e farmacoeconomia. Belo Horizonte: COOPMED**, p. 147–178, 2013. Disponível em: <https://www.rbfhss.org.br/sbrafh/article/view/219>. Acesso em: 22 out. 2024.

CEBM. **The Oxford Centre for Evidence-Based Medicine (CEBM)**. Disponível em: <https://www.cebm.net/>. Acesso em: 16 out. 2022.

CERNER MULTUM. **Drugs.com**. Disponível em: <http://drugs.com>. Acesso em: 15 out. 2022.

CENTRO REGIONAL DE ESTUDOS PARA O DESENVOLVIMENTO DA SOCIEDADE DA INFORMAÇÃO (CETIC.br). **Pesquisa TIC Saúde 2023: Estabelecimentos - B1**. Disponível em: https://cetic.br/pt/tics/saude/2023/estabelecimentos/B1/. Acesso em: 22 out. 2024.

CENTRO REGIONAL DE ESTUDOS PARA O DESENVOLVIMENTO DA SOCIEDADE DA INFORMAÇÃO (CETIC.br). **Pesquisa TIC Saúde 2023: Estabelecimentos - B9**. Disponível em: https://cetic.br/pt/tics/saude/2023/estabelecimentos/B9/. Acesso em: 22 out. 2024.

CI-IA Saúde – Centro de Inovação em Inteligência Artificial para a Saúde. 2024. Disponível em: <https://ciia-saude.dcc.ufmg.br/>. Acesso em: 31 jul. 2024.

COCHRANE. **Cochrane Library**. Disponível em: <https://www.cochranelibrary.com/central>. Acesso em: 16 out. 2022.

COELHO NETO, Giliate Cardoso; CHIORO, Arthur. Afinal, quantos Sistemas de Informação em Saúde de base nacional existem no Brasil? **Cad. Saúde Pública**, v. 37, n. 7, p. e00182119, jul. 2021.

CONASS. **Mortalidade por causas externas**. Disponível em: <https://wiki.conass.org.br/index.php?title=Mortalidade_por_causas_externas>. Acesso em: 7 jul. 2024.

CONASS; OLIVEIRA, L.. **Secretarias mostram os benefícios da Saúde Digital para a população e gestores do SUS**. Disponível em: <https://www.conass.org.br/secretarias-mostram-os-beneficios-da-saude-digital-para-a-populacao-e-gestores-do-sus/>. Acesso em: 8 jul. 2024.

Coorte de 100 Milhões de Brasileiros. Disponível em: <https://cidacs.bahia.fiocruz.br/plataforma/coorte-de-100-milhoes-de-brasileiros/>. Acesso em: 31 jul. 2024.

DA SAÚDE, Ministério. **Sage: Sala de apoio à gestão estratégica**. 2013. Brasília: Ministério da Saúde Brasília (BR). Disponível em: <http://sage.saude.gov.br/>. Acesso em: 31 jul. 2024.

LEANDRO, B. B. S; REZENDE, F. A. V.; PINTO, J. M. C.. **Informações e registros em saúde e seus usos no SUS**. Brasília: SciELO - Editora Fiocruz, 2020. Disponível em: <https://play.google.com/store/books/details?id=vYLrDwAAQBAJ>. Acesso em: 22 out. 2024.

ARAUJO, G. D. *et al*. Análise de sentimentos sobre temas de saúde em mídia social. **Journal of health informatics in developing countries**, v. 4, n. 3, 25 set. 2012. Disponível em: <https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/195>. Acesso em: 8 jul. 2024.

DEININGER, L. S. C. *et al*. A sala de situação da dengue como ferramenta de gestão em saúde. **Saúde em Debate**, v. 38, p. 50–56, mar. 2014. Disponível em: <https://www.scielosp.org/article/sdeb/2014.v38n100/50-56/pt/>. Acesso em: 12 jul. 2019.

DENSEN, P. Challenges and Opportunities Facing Medical Education. **Transactions of the American Clinical and Climatological Association**, v. 122, p. 48, 2011. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116346/>. Acesso em: 7 set. 2024.

Drug Interactions Checker - Medscape Drug Reference Database. Disponível em: <https://reference.medscape.com/drug-interactionchecker>. Acesso em: 15 out. 2022.

DrugBank Online. Disponível em: <https://go.drugbank.com/>. Acesso em: 15 out. 2022.

DUTRA, F. G.; BARBOSA, R. R.. Modelos e critérios para avaliação da qualidade de fontes de informação: uma revisão sistemática de literatura. v. 27, n. 2, 25 ago. 2017. Disponível em: <https://periodicos.ufpb.br/index.php/ies/article/view/32676>. Acesso em: 9 jul. 2024.

EDREES, H *et al.*. Intelligent Telehealth in Pharmacovigilance: A Future Perspective. **Drug Safety: An International Journal of Medical Toxicology and Drug Experience**, v. 45, n. 5, p. 449–458, maio 2022. Disponível em: <http://dx.doi.org/10.1007/s40264-022-01172-5>. Acesso em: 22 out. 2024.

ELMASRI, R.; NAVATHE, S.B.. **Sistemas de banco de dados**. 6. ed. [S.l.]: Pearson Addison Wesley, 2011.

FEI-FEI, L. *et al.*. **ImageNet**. Disponível em: <https://www.image-net.org/>. Acesso em: 9 jul. 2024.

FENG, Y-H.; ZHANG, S-W.. Prediction of Drug-Drug Interaction Using an Attention-Based Graph Neural Network on Drug Molecular Graphs. **Molecules** , v. 27, n. 9, 7 maio 2022. Disponível em: <http://dx.doi.org/10.3390/molecules27093004>. Acesso em: 22 out. 2024.

FERRÉ, F. *et al.*. **13 - O papel tripartite na divulgação de casos e óbitos por Covid-19 e a atuação do Conass**. **Covid-19 no Brasil: cenários epidemiológicos e vigilância em saúde**. [Brasília: s.n.]. , 2021. Disponível em: <https://books.scielo.org/id/zx6p9/pdf/freitas-9786557081211-16.pdf>. Acesso em: 22 out. 2024.

Find Open Datasets and Machine Learning Projects. Disponível em: <https://www.kaggle.com/datasets>. Acesso em: 8 jul. 2024.

FONTELLES, M. J *et al.*. Metodologia da pesquisa científica: diretrizes para a elaboração de um protocolo de pesquisa. **Revista paraense de medicina**, v. 23, n. 3, p. 1–8, 2009. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/3049277/mod_resource/content/1/DIRETRIZES%20PARA%20A%20ELABORA%C3%87%C3%83O%20DE%20UM%20PROJ%20PESQUISA.pdf>. Acesso em: 22 out. 2024.

GALATI, S *et al.*. VenomPred: A Machine Learning Based Platform for Molecular Toxicity Predictions. **International Journal of Molecular Sciences**, v. 23, n. 4, 14 fev. 2022. Disponível em: <http://dx.doi.org/10.3390/ijms23042105>. Acesso em: 22 out. 2024.

Gene Ontology Resource. Disponível em: <http://geneontology.org/>. Acesso em: 15 out. 2022.

GOMES, D. C *et al*.. Uso de ferramentas computacionais como auxílio ao método de mapeamento cruzado entre terminologias clínicas. **Texto & Contexto - Enfermagem**, v. 28, p. e20170187, 14 fev. 2019. Disponível em: <https://www.scielo.br/j/tce/a/rhgrcS7CbHhQqJKsVGhHNzk/?format=html&lang=pt>. Acesso em: 8 jul. 2024.

Harvard Dataverse. Disponível em: <https://dataverse.harvard.edu/>. Acesso em: 8 jul. 2024.

HE, C *et al*.. Multi-Type Feature Fusion Based on Graph Neural Network for Drug-Drug Interaction Prediction. **BMC Bioinformatics**, v. 23, n. 1, p. 224, 10 jun. 2022. Disponível em: <http://dx.doi.org/10.1186/s12859-022-04763-2>. Acesso em: 8 jul. 2024.

IPDLN Network. Disponível em: <https://ipdln.org/network/>. Acesso em: 31 jul. 2024.

JAIN, H.; RAJ, N.; MISHRA, S.. A Sui Generis QA Approach Using RoBERTa for Adverse Drug Event Identification. **BMC Bioinformatics**, v. 22, n. Suppl 11, p. 330, 21 out. 2021. Disponível em: <http://dx.doi.org/10.1186/s12859-021-04249-7>. Acesso em: 8 jul. 2024.

JUNIOR, A. A. G *et al*.. Building the National Database of Health Centred on the Individual: Administrative and Epidemiological Record Linkage-Brazil, 2000-2015. **International Journal of Population Data Science**, v. 3, n. 1, 2018. Disponível em: <https://ijpds.org/article/view/446>. Acesso em: 8 jul. 2024.

KANG, M-G; KANG, N. S.. Predictive Model for Drug-Induced Liver Injury Using Deep Neural Networks Based on Substructure Space. **Molecules** , v. 26, n. 24, 13 dez. 2021. Disponível em: <http://dx.doi.org/10.3390/molecules26247548>. Acesso em: 8 jul. 2024.

KUROSAKI, K.; UESAWA, Y.. Development of *in Silico* Prediction Models for Drug-Induced Liver Malignant Tumors Based on the Activity of Molecular Initiating Events: Biologically Interpretable Features. **The Journal of Toxicological Sciences**, v. 47, n. 3, p. 89–98, 2022. Disponível em: <http://dx.doi.org/10.2131/jts.47.89>. Acesso em: 8 jul. 2024.

LAKATOS, E. M.. **Metodologia científica**. Brasília: Atlas, 1983. Disponível em: <https://play.google.com/store/books/details?id=yhwnwQEACAAJ>. Acesso em: 8 jul. 2024.

LEAL, L. F *et al*. Data Sources for Drug Utilization Research in Brazil-DUR-BRA Study. **Frontiers in Pharmacology**, v. 12, p. 789872, 2021a. Disponível em: <https://books.scielo.org/id/xhr84>. Acesso em: 8 jul. 2024.

LÉTINIER, L *et al*. Artificial Intelligence for Unstructured Healthcare Data: Application to Coding of Patient Reporting of Adverse Drug Reactions. **Clinical Pharmacology and Therapeutics**, v. 110, n. 2, p. 392–400, ago. 2021. Disponível em: <http://dx.doi.org/10.1002/cpt.2266>. Acesso em: 8 jul. 2024.

LIMA-COSTA, M. F.; BARRETO, S. M.. Tipos de estudos epidemiológicos: conceitos básicos e aplicações na área do envelhecimento. **Epidemiologia e Serviços de Saúde**, v. 12, n. 4, p. 189–201, 2003. Disponível em: <http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S1679-http://scielo.iec.pa.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742003000400003&lng=pt&nrm=iso>. Acesso em: 30 jun. 2024.

LING, X.; GAO, M.; WANG, D.. Intelligent document processing based on RPA and machine learning. Instituto de Engenheiros Eletricistas e Eletrônicos (IEEE). 2020. p. 1349–1353. Disponível em: <http://dx.doi.org/10.1109/CAC51589.2020.9326579>. Acesso em: 8 jul. 2024.

MADER, K. S.. **Pretrained-VGG16 for Mammography Classification**. Disponível em: <https://www.kaggle.com/code/kmader/pretrained-vgg16-for-mammography-classification/input>. Acesso em: 5 ago. 2024.

MANDVIKAR, S.. Augmenting intelligent document processing (IDP) workflows with contemporary large language models (LLMs). **International journal of computer trends and technology**, 30 out. 2023. Disponível em: <https://www.researchgate.net/profile/Shreekant-Mandvikar/publication/375487356_Augmenting_Intelligent_Document_Processing_IDP_Workflows_with_Contemporary_Large_Language_Models_LLMs/links/654bbb443fa26f66f4e74d0b/Augmenting-Intelligent-Document-Processing-IDP-Workflows-with-Contemporary-Large-Language-Models-LLMs.pdf>. Acesso em: 8 jul. 2024.

MARTIN, G. L *et al.*. Validation of Artificial Intelligence to Support the Automatic Coding of Patient Adverse Drug Reaction Reports, Using Nationwide Pharmacovigilance Data. **Drug Safety: An International Journal of Medical Toxicology and Drug Experience**, v. 45, n. 5, p. 535–548, maio 2022. Disponível em: <http://dx.doi.org/10.1007/s40264-022-01153-8>. Acesso em: 8 jul. 2024.

MASUMSHAH, R.; AGHDAM, R.; ESLAHCHI, C.. A Neural Network-Based Method for Polypharmacy Side Effects Prediction. **BMC Bioinformatics**, v. 22, n. 1, p. 385, 24 jul. 2021. Disponível em: <http://dx.doi.org/10.1186/s12859-021-04298-y>. Acesso em: 8 jul. 2024.

Ministério da Saúde lança ferramenta que mede o nível de maturidade em saúde digital nas regiões do país. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/noticias/2024/maio/ministerio-da-saude-lanca-ferramenta-que-mede-o-nivel-de-maturidade-em-saude-digital-nas-regioes-do-pais>. Acesso em: 9 jul. 2024.

MORAIS, J. H. A *et al.*. Rtabnetsp: An R Package for Retrieving São Paulo State Health Status Indicators, Brazil. **Epidemiologia e Serviços de Saúde : Revista do Sistema** Único **de Saúde do Brasil**, v. 30, n. 1, p. e2020576, Jan-Dec 2021. Disponível em: <http://dx.doi.org/10.1590/S1679-4974202100020>. Acesso em: 8 jul. 2024.

MORIN, E.. **Introdução ao Pensamento Complexo**. Porto Alegre: Sulina, 2015. p. 120

MOYA, J.; RISI JUNIOR, J. B.; MARTINELLO, A.. **Salas de situação em saúde: compartilhando as experiências do Brasil**. Brasília, DF: Ministério de Salud [Brasil]; Organización Panamericana de la Salud, 2010. Disponível em:<https://bvsms.saude.gov.br/bvs/publicacoes/sala_situacao_saude_2010.pdf>. Acesso em: 8 jul. 2024.

NATIVIDADE, M *et al*.. Segurança no uso dos dados sensíveis para pesquisa em saúde: Repositório de dados. In: ARAGÃO, Erika Santos de et al. (Org.). **Avaliação de impacto das políticas de saúde: um guia para o SUS**. Brasília: Ministério da Saúde, 2023. Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/avaliacao_impacto_politicas_saude_guia_sus.pdf>. Acesso em: 8 jul. 2024.

NDEx WebApp. Disponível em: <https://www.ndexbio.org/index.html#/>. Acesso em: 8 jul. 2024.

NERY, J. S *et al*.. **Socioeconomic determinants of leprosy new case detection in the 100 Million Brazilian Cohort: a population-based linkage study**. **The Lancet Global Health**. 2019. Disponível em: <http://dx.doi.org/10.1016/s2214-109x(19)30260-8>. Acesso em: 8 jul. 2024.

OPAS. **Indicadores básicos para a saúde no Brasil: conceitos e aplicações**. Brasília: Organização Pan Americana da Saúde, 2008. Disponível em: <https://market.android.com/details?id=book-p9eiQwAACAAJ>. Acesso em: 8 jul. 2024.

OSF. Disponível em: <https://osf.io/>. Acesso em: 8 jul. 2024.

PAGE, D *et al*.. Identifying Adverse Drug Events by Relational Learning. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 2012, p. 790–793, jul. 2012. Disponível em: <https://www.ncbi.nlm.nih.gov/pubmed/24955289>. Acesso em: 8 jul. 2024.

PETRUZALEK, D.. **read.dbc: Read Data Stored in DBC (Compressed DBF) Files**. Disponível em: <https://cran.r-project.org/web/packages/read.dbc/index.html>. Acesso em: 28 out. 2022.

PIRES, Frank James da Silva. **O data lake da RNDS**. Brasília: XVII Congresso Brasileiro de Informática em Saúde (CBIS) da Sociedade Brasileira de Informática em Saúde (SBIS), 2020. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/saude-digital/material-de-apoio/CBIS2020_Datalake_RNDS.pdf>. Acesso em: 8 jul. 2024.

POZOBON, R. O.; DE DAVID, C. S.. Pesquisa em Comunicação: desenho metodológico construído a partir da teoria da Argumentação. **Fronteiras - estudos midiáticos**, v. 21, n. 2, 10 set. 2019. Disponível em: <https://repositorio.enap.gov.br/handle/1/3330>.

Pysus. Disponível em: <https://pypi.org/project/PySUS/>. Acesso em: 5 nov. 2023.

RÁCZ, A *et al*.. Machine Learning Models for Classification Tasks Related to Drug Safety. **Molecular Diversity**, v. 25, n. 3, p. 1409–1424, ago. 2021. Disponível em: <http://dx.doi.org/10.1007/s11030-021-10239-x>. Acesso em: 8 jul. 2024.

RIBEIRO, A. L. P *et al*.. Tele-Electrocardiography and Bigdata: The CODE (Clinical Outcomes in Digital Electrocardiography) Study. **Journal of Electrocardiology**, v. 57S, p. S75–S78, 7 set. 2019. Disponível em: <http://dx.doi.org/10.1016/j.jelectrocard.2019.09.008>. Acesso em: 8 jul. 2024.

RIBEIRO, L. A. P. A.; GARCIA, A. C. B.; DOS SANTOS, P. S. M.. Dependency Factors in Evidence Theory: An Analysis in an Information Fusion Scenario Applied in Adverse Drug Reactions. **Sensors** , v. 22, n. 6, 16 mar. 2022. Disponível em: <http://dx.doi.org/10.3390/s22062310>. Acesso em: 8 jul. 2024.

RIEDE, M *et al*. On the communication of scientific data: The Full-Metadata Format. **Computer physics communications**, v. 181, n. 3, p. 651–662, 1 mar. 2010. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010465509003890>. Acesso em: 8 jul. 2024.

RIPSA. Disponível em: <https://www.ripsa.org.br/>. Acesso em: 31 jul. 2024.

SACKETT, D. L. *et al*.. **Evidence-based medicine: how to practice and teach EBM**. 2nd ed. Edinburgh: Churchill Livingstone, 2000.

SAEED, F.; MOHAMMED, F.; AL-NAHARI, A.. **Innovative Systems for Intelligent Health Informatics: Data Science, Health Informatics, Intelligent Systems, Smart Computing**. Springer Nature, 2021. Disponível em: <https://play.google.com/store/books/details?id=POEsEAAAQBAJ>. Acesso em: 8 jul. 2024.

SALDANHA, R. F.; BASTOS, R. R.; BARCELLOS, C.. Microdatasus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). **Cadernos de Saúde Pública**, v. 35, p. e00032419, set. 2019. Disponível em: <https://www.scielo.br/j/csp/a/gdJXqcrW5PPDHX8rwPDYL7F/>. Acesso em: 29 fev. 2024. Acesso em: 8 jul. 2024.

SALDANHA, R. F.; PEDROSO, M.M.; MAGALHÃES, M. A. F. M. Acesso aos dados agregados e microdados do SUS. In: ARAGÃO, Erika Santos de et al. (Org.). **Avaliação de impacto das políticas de saúde um guia para o SUS**. Brasília: Ministério da Saúde, 2023. . Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/avaliacao_impacto_politicas_saude_guia_sus.pdf>. Acesso em: 8 jul. 2024.

SBIS. **O que é Informática em Saúde**. Disponível em: <http://sbis.org.br/o-que-e-informatica-em-saude/>. Acesso em: 8 jul. 2024.

SCHRARSTZHAUPT, I. N *et al.*. Paineis de monitoramento interativos da pandemia de COVID-19 no mundo com o uso de dados abertos antecipando ondas da doença no Brasil. **Revista brasileira de epidemiologia = Brazilian journal of epidemiology**, v. 27, p. e240004, 5 fev. 2024. Disponível em: <https://www.scielosp.org/article/rbepid/2024.v27/e240004/pt/>. Acesso em: 14 jul. 2024.

SICILIA, M.-A.; GARCÍA-BARRIOCANAL, E.; SÁNCHEZ-ALONSO, S.. Community Curation in Open Dataset Repositories: Insights from Zenodo. **Procedia computer science**, v. 106, p. 54–60, 1 jan. 2017. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050917302776>. Acesso em: 8 jul. 2024.

SILVA, N. P.. **A utilização dos programas TABWIN e TABNET como ferramentas de apoio** à disseminação **das informações em saúde**. 2009. Fundação Oswaldo Cruz. Escola Nacional de Saúde Pública Sergio Arouca, Rio de Janeiro, RJ, Brasil, 2009. Disponível em: <https://www.arca.fiocruz.br/handle/icict/2300>. Acesso em: 29 fev. 2024.

SILVA, R. P.; POLLETTINI, J. T.; PAZIN FILHO, A.. Unsupervised natural language processing in the identification of patients with suspected COVID-19 infection. **Cadernos de Saúde Pública**, v. 39, n. 11, p. e00243722, 4 dez. 2023. Disponível em: <http://dx.doi.org/10.1590/0102-311XPT243722>. Acesso em: 8 jul. 2024.

SINGH, J.. FigShare. **Journal of Pharmacology & Pharmacotherapeutics**, v. 2, n. 2, p. 138–139, abr. 2011. Disponível em: <http://dx.doi.org/10.4103/0976-500X.81919>. Acesso em: 8 jul. 2024.

SUCKLING, J.. The mammographic images analysis society digital mammogram database. 1994, **Exerpta Medica International Congress**. 1994. p. 375–378. Disponível em: <https://cir.nii.ac.jp/crid/1572261551129161728>. Acesso em: 8 jul. 2024.

SYROWATKA, A *et al.*. Key Use Cases for Artificial Intelligence to Reduce the Frequency of Adverse Drug Events: A Scoping Review. **The Lancet. Digital Health**, v. 4, n. 2, p. e137–e148, fev. 2022. Disponível em: <http://dx.doi.org/10.1016/S2589-7500(21)00229-6>. Acesso em: 8 jul. 2024.

Top 6 Cloud Data Warehouse Solutions in 2024 [compared]. Disponível em: <https://www.scnsoft.com/data/data-warehouse/cloud>. Acesso em: 9 jul. 2024.

UNIVERSITY OF GOTHENBURG. **Swedish National Data Service**. Disponível em: <https://snd.se/en>. Acesso em: 8 jul. 2024.

VILAÇA, E. M.. **As redes de atenção à saúde**. Brasília: Organização Pan-Americana da Saúde, 2011. 549. Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/redes_de_atencao_saude.pdf>. Acesso em: 22 out. 2024.

VERLEYSEN, M.; FRANÇOIS, D.. The Curse of Dimensionality in Data Mining and Time Series Prediction. 2005, **Springer Berlin Heidelberg**, 2005. p. 758–770. Disponível em: <http://dx.doi.org/10.1007/11494669_93>. Acesso em: 22 out. 2024.

WEBMD. **RxList - The Internet Drug Index for Prescription Drug Information, Interactions, and Side Effects**. Disponível em: <https://www.rxlist.com/>. Acesso em: 15 out. 2022.

WIXON, J.; KELL, D.. The Kyoto Encyclopedia of Genes and Genomes--KEGG. **Yeast** , v. 17, n. 1, p. 48–55, abr. 2000. Disponível em: <http://dx.doi.org/10.1002/(SICI)1097-0061(200004)17:1<48::AID-YEA2>3.0.CO;2-H>. Acesso em: 22 out. 2024.

WORLD HEALTH ORGANIZATION. **Guidelines for ATC classification and DDD assignment 2022**. 25. ed. WHO Collaborating Centre for Drug Statistics Methodology, 2022. Disponível em: <https://www.whocc.no/filearchive/publications/2022_guidelines_web.pdf>.

WU, L *et al.*. Machine Learning Methods, Databases and Tools for Drug Combination Prediction. **Briefings in Bioinformatics**, v. 23, n. 1, 17 jan. 2022. Disponível em: <http://dx.doi.org/10.1093/bib/bbab355>. Acesso em: 8 jul. 2024.

ZAKI, M. J.; MEIRA JUNIOR, W.. **Data Mining and Machine Learning: Fundamental Concepts and Algorithms**. Cambridge University Press, 2020a. Disponível em: <https://dataminingbook.info/book_html/>. Acesso em: 8 jul. 2024.

ZHANG, C.; LU, Y.; ZANG, T.. CNN-DDI: A Learning-Based Method for Predicting Drug-Drug Interactions Using Convolution Neural Networks. **BMC Bioinformatics**, v. 23, n. Suppl 1, p. 88, 7 mar. 2022. Disponível em: <http://dx.doi.org/10.1186/s12859-022-04612-2>. Acesso em: 8 jul. 2024.

ZHAO, Y *et al.*. Machine Learning in Causal Inference: Application in Pharmacovigilance. **Drug Safety: An International Journal of Medical Toxicology and Drug Experience**, v. 45, n. 5, p. 459–476, maio 2022. Disponível em: <http://dx.doi.org/10.1007/s40264-022-01155-6>. Acesso em: 8 jul. 2024.

ZHENG, S *et al.*. DrugComb Update: A More Comprehensive Drug Sensitivity Data Repository and Analysis Portal. **Nucleic Acids Research**, v. 49, n. W1, p. W174–W184, 2 jul. 2021. Disponível em: <http://dx.doi.org/10.1093/nar/gkab438>. Acesso em: 8 jul. 2024.

ZORZAL, L.; RODRIGUES, G. M.. **Transparência dos relatórios de gestão das universidades federais à luz dos princípios de dados abertos**. 8 jan. 2016. Disponível em: <https://www.redalyc.org/journal/161/16144489001/>. Acesso em: 22 out. 2024.

AKCIT

CENTRO DE COMPETÊNCIA EMBRAPII
EM TECNOLOGIAS IMERSIVAS

SOBRE O E-BOOK

Tipografia: Montserrat

Publicação: Cegraf UFG
Câmpus Samambaia, Goiânia - Goiás. Brasil. CEP 74690-900
Fone: (62) 3521-1358
https://cegraf.ufg.br