UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

RENATO GOMES BORGES JÚNIOR

Aprendizado de Máquina para Análise de Recaída para Depressão em Pacientes com Transtorno Bipolar

Goiânia
2018

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

1. Identificação do material bibliográfico: **[X] Dissertação** **[ ] Tese**

2. Identificação da Tese ou Dissertação:

Nome completo do autor: Renato Gomes Borges Júnior

Título do trabalho: Aprendizado de Máquina para Análise de Recaída para Depressão em Pacientes com Transtorno Bipolar

3. Informações de acesso ao documento:

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF da tese ou dissertação.

Renato Gomes Borges Júnior
Assinatura do(a) autor(a)[2]

Ciente e de acordo:

Assinatura do(a) orientador(a)[2] Data: 31 / 10 / 2018

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Os dados do documento não serão disponibilizados durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

[2] A assinatura deve ser escaneada.

RENATO GOMES BORGES JÚNIOR

Aprendizado de Máquina para Análise de Recaída para Depressão em Pacientes com Transtorno Bipolar

Dissertação apresentada ao Programa de Pós–Graduação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação.

Orientador: Prof. Dr. Rogerio Lopes Salvini

Goiânia
2018

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Borges Júnior, Renato
Aprendizado de Máquina para Análise de Recaída para Depressão em Pacientes com Transtorno Bipolar [manuscrito] / Renato Borges Júnior. - 2018.
LXXXIII, 83 f.

Orientador: Prof. Dr. Rogerio Salvini.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2018.
Bibliografia. Apêndice.
Inclui algoritmos, lista de figuras, lista de tabelas.

1. transtorno bipolar. 2. depressão. 3. aprendizado de máquina. 4. mineração de dados. I. Salvini, Rogerio, orient. II. Título.

CDU 004

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

UFG

ATA Nº 17/2018

ATA DA SESSÃO DE JULGAMENTO DA DISSERTAÇÃO DE MESTRADO DE RENATO GOMES BORGES JÚNIOR

Aos quatro dias do mês de outubro de dois mil e dezoito, às catorze horas, na sala 150 do Instituto de Informática da Universidade Federal de Goiás, Campus Samambaia, reuniu-se a banca examinadora designada na forma regimental pela Coordenação do Curso para julgar a dissertação de mestrado intitulada **"Aprendizado de máquina para análise de recaída para depressão em pacientes com Transtorno Bipolar"**, apresentada pelo aluno Renato Gomes Borges Júnior como parte dos requisitos necessários à obtenção do grau de Mestre em Ciência da Computação, área de concentração Ciência da Computação. A banca examinadora foi presidida pelo orientador do trabalho de dissertação, Professor Doutor Rogerio Lopes Salvini (INF/UFG), tendo como membros os Professores Doutores Fernando Marques Federson (INF/UFG) e Eduardo José Aguilar Alonso (UNIFAL). Aberta a sessão, o candidato expôs seu trabalho. Em seguida, o aluno foi arguido pelos membros da banca e:

(X) tendo demonstrado suficiência de conhecimento e capacidade de sistematização do tema de sua dissertação, a banca concluiu pela **aprovação** do candidato, sem restrições.

( ) não tendo demonstrado suficiência de conhecimento e capacidade de sistematização do tema de sua dissertação, a banca concluiu pela **reprovação** do candidato.

Os trabalhos foram encerrados às 12:30 horas. Nos termos do Regulamento Geral dos Cursos de Pós-Graduação desta Universidade, lavrou-se a presente ata que, lida e julgada conforme, segue assinada pelos membros da banca examinadora.

Prof. Dr. Rogerio Lopes Salvini _____

Prof. Dr. Fernando Marques Federson _____

Prof. Dr. Eduardo José Aguilar Alonso _____

Renato Gomes Borges Júnior

Graduou-se em Ciência da Computação pela UFG – Universidade Federal de Goiás – tendo cursado 1 ano de intercâmbio na Vrije Universiteit Amsterdam. Atua como Desenvolvedor Java na empresa Oobj Tecnologia da Informação a mais de 3 anos. Durante o Mestrado, na UFG, foi bolsista pela CAPES e desenvolvou um trabalho de análise de recaída para depressão em pacientes com transtorno bipolar utilizando aprendizado de máquina.

À minha família.

Agradecimentos

Agradeço principalmente aos meus pais, Renato e Joelma, pelo apoio e confiança durante toda a minha vida e em especial durante este momento de grande esforço e sacrifício.

Agradeço também ao meu orientador, Prof. Dr. Rogerio Salvini, pela confiança, paciência, incentivo e pelo aprendizado proporcionado. Extendo o agradecimento ao Prof. Dr. Rodrigo Dias, médico psiquiatra que colaborou e auxíliou nesta pesquisa.

Por fim agradeço ao meu irmão Leonardo pelo incentivo. À minha namorada Nattane pelo auxílio, suporte e companhia. Por fim, agradeço a todos os amigos e colegas que me acompanharam durante esta jornada e que me proporcionaram momentos de descontração e alegria.

"Acts of goodness are not always wise and acts of evil are not always foolish, but regardless, we shall always strive to be good."

Martyr Logarius,
Bloodborne.

Resumo

Borges-Júnior, Renato Gomes. **Aprendizado de Máquina para Análise de Recaída para Depressão em Pacientes com Transtorno Bipolar**. Goiânia, 2018. 83p. Dissertação de Mestrado. Instituto de Informática, Universidade Federal de Goiás.

A recaída para depressão em pacientes com Transtorno Afetivo Bipolar (TAB) atinge taxas de 70% de recorrência nos 4 primeiros anos de tratamento e pode causar uma drástica redução na qualidade de vida e levar até o suicídio. O TAB é uma desordem do humor caracterizada por episódios recorrentes de depressão ou mania. Para estudar o transtorno e encontrar tratamentos mais eficientes, o *Systematic Treatment Enhancement Program for Bipolar Disorder* (STEP-BD) foi criado pela Escola de Medicina de Harvard. O STEP-BD é um conjunto de dados composto por informações de 4.360 pacientes com TAB, o qual pode ser considerado atualmente uma das mais completas bases de dados em termos de escopo. Vários estudos foram desenvolvidos para descobrir tratamentos mais eficientes para prevenir recaídas. Porém, a maioria destes estudos usaram apenas métodos clássicos de estatística, principalmente com o objetivo de medir a sua correlação com atributos específicos. Este trabalho apresenta uma análise do uso de algoritmos de aprendizado de máquina para encontrar padrões relacionados a recaída para depressão no TAB com o uso de dados longitudinais providos pelo STEP-BD. Estes dados longitudinais incluem 148 atributos coletados em um total de 50.987 visitas de pacientes espalhadas ao longo de semanas durante anos. Assim, diversos experimentos foram conduzidos neste trabalho e os resultados mostram que os algoritmos obtiveram desempenho limitado. Foi possível perceber que atributos relacionados ao estado de humor de depressão e mania, coletados pelo STEP-BD, não podem ser usados propriamente para predizer recaída para depressão antes de sua ocorrência, sendo apropriados apenas para uso como um indicador que o paciente já se encontra no estado de depressão.

Palavras–chave

transtorno bipolar, recaída para depressão, aprendizado de máquina, mineração de dados

Abstract

Borges-Júnior, Renato Gomes. **Machine Learning to Analyze Depression Relapse in Bipolar Disorder Patients**. Goiânia, 2018. 83p. MSc. Dissertation. Instituto de Informática, Universidade Federal de Goiás.

Depression relapse in patients with Bipolar Disorder (BD) have 70% rate of recurrence in the first 4 years of treatment and may cause a severe loss of quality of life and even lead to suicide. BD is a mood disorder characterized by recurrent episodes of depression or mania. To study the disorder and find more efficient treatments, the Harvard Medical School created the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD). It is a widely used dataset that comprises data of 4,360 patients with BD, which can be considered one of the most complete databases in terms of scope nowadays. Several studies have been developed to discover more efficient treatments to prevent relapses in BD. However, most of them used only classical statistical methods, mainly aimed at measuring its correlation to specific features. This study presents an analysis of the use of machine learning algorithms to discover patterns related to depression relapse in BD with the use of longitudinal data provided by STEP-BD. This longitudinal data includes 148 features collected in 50,987 visits of patients spread across different weeks over the years. Thus, several experiments were conducted and the results show that the algorithms attained limited performance. We concluded that features related to depression and mania mood states, collected by the STEP-BD, cannot be used properly to predict the relapse to depression before it occurs, being suited only as an indicator that the patient is already in the state of depression.

Keywords

bipolar disorder, depression relapse, machine learning, data mining

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Códigos de Programas

CAPÍTULO 1

Introdução

1.1 Transtorno Afetivo Bipolar

O Transtorno Afetivo Bipolar (TAB) é uma desordem do humor que afeta em torno de 2% da população mundial. O paciente manifesta episódios depressivos e de mania que podem causar prejuízos na vida pessoal e profissional e levar a uma baixa qualidade de vida [40, 27].

Um estudo da Organização Mundial de Saúde (OMS) mostrou que aproximadamente 29 milhões de pessoas no mundo apresentam o TAB, o qual ocupa a 12^a posição em relação a causa principal de incapacitação de moderada à grave [52].

Os portadores do TAB podem ciclar entre os estados de depressão e mania/hipomania, além de um estado misto entre estes dois. Apesar da depressão ocorrer com mais frequência no curso da doença, o diagnóstico é determinado principalmente pela identificação do paciente no estado de mania ou hipomania, devido a dificuldade de diferenciação com a depressão unipolar [25].

Episódios de depressão são caracterizados por profunda perda de interesse em atividades, além de sintomas como fatiga, perda ou ganho de peso, distúrbios no sono, atividade psicomotora lenta, sentimentos de inutilidade, culpa excessiva e pensamentos ou ações suicidas.

O estado de mania se caracteriza por auto-estima inflada e desinibição. O paciente pode apresentar necessidade de conversar de forma incessante e a fala pode ser pressionada, mais rápida e mais alta que o usual. Em casos mais graves, a pessoa pode experimentar pensamentos velozes, o que torna difícil expressar ideias de forma coerente.

A hipomania se diferencia por apresentar sintomas de mania menos graves. Já o estado misto se caracteriza pela presença de sintomas de mania e depressão em conjunto por pelo menos 1 semana.

O tratamento do TAB consiste no foco em dois pontos:

1. estabilização de sintomas de mania ou depressão para o estado de eutimia (que um estado onde o paciente não apresenta sintomas da doença) e,

2. na manutenção desse estado, cujo objetivo é prevenir episódios de recaídas.

A complexidade em cada fase do TAB trás um problema, pois tratamentos que aliviam depressão podem causar mania ou hipomania, e tratamentos que reduzem sintomas de mania, podem aumentar a chance de um episódio de depressão [27].

1.2 Motivação

Recaídas para depressão no TAB atingem taxas próximas de 50% em 1 ano e 70% em até 4 anos de tratamento [29]. A recaída para depressão é caracterizada pela mudança no paciente do estado de eutimia (estável) para a depressão, e é o objeto de estudo deste trabalho.

Com o objetivo de tratar os efeitos do TAB, a Escola de Medicina de Harvard criou o *Systematic Treatment Enhancement Program for Bipolar Disorder* (STEP-BD). O STEP-BD é um estudo desenvolvido de 1995 à 2005, financiado pelo National Institute of Mental Health (NIMH), que contou com a colaboração de 4.360 pacientes com TAB [68][1]. Os dados dos pacientes foram coletados no início do programa e a longo prazo por médicos que receberam treinamento especializado [68].

Trabalhos prévios foram realizados para analisar de forma qualitativa e quantitativa, com auxílio principalmente de técnicas estatísticas, os dados do STEP-BD [62, 61, 49]. Não se encontrou na literatura, porém, o uso de técnicas de mineração de dados, ou mais especificamente, algoritmos de aprendizado de máquina para extrair conhecimento dessa base que, dado sua dimensão e escopo, representa hoje a maior fonte de informações clínicas sobre TAB.

1.3 Objetivo

O objetivo deste trabalho é fazer a análise de dados clínicos de pacientes da base de dados STEP-BD com a intenção de verificar a possibilidade de predição de recaídas para depressão. A metodologia utilizada consiste na aplicação do processo de Descoberta de Conhecimento em Bases de Dados, com a utilização de algoritmos de aprendizado de máquina aplicados à base do STEP-BD na fase de Mineração de Dados. Com isso, espera-se com este estudo, fornecer uma analise dos dados que permita encontrar padrões e características para auxiliar na previsão de recaída para depressão.

[1] Para mais informações: https://www.nimh.nih.gov/funding/clinical-research/practical/step-bd/index.shtml

1.4 Estrutura

O restante deste trabalho está organizado da seguinte forma: no Capítulo 2, os conceitos teóricos das ferramentas usadas são detalhados. Em seguida, no Capítulo 3, uma revisão sistemática de literatura é apresentada, que detalha o estado da arte de aprendizado de máquina aplicado ao TAB para análise de recaída para depressão. No Capítulo 4, é apresentado a abordagem metodológica usada nos experimentos desenvolvidos. Os experimentos realizados e os resultados obtidos são exibidos no Capítulo 5. Por fim, os resultados são discutidos no Capítulo 6 e uma conclusão acerca do trabalho como um todo e possíveis trabalhos futuros é apresentada no Capítulo 7.

CAPÍTULO 2

Fundamentação Teórica

Este capítulo destina-se a definição dos conceitos técnicos abordados ao longo da dissertação. Inicialmente, é descrito o processo de Descoberta de Conhecimento em Bases de Dados. Em seguida, o conceito de aprendizado de máquina é apresentado e são detalhados os algoritmos usados neste trabalho, o que inclui algoritmos proposicionais e relacionais. Por fim, é descrito como validar e avaliar o seu desempenho para a tarefa de predição.

2.1 Descoberta de Conhecimento em Bases de Dados

O termo "Descoberta de Conhecimento em Bases de Dados" (em inglês, *Knowledge Discovery in Databases*, KDD) foi inicialmente utilizado na primeira conferência de KDD em 1989 por Piatetsky-Shapiro [23] para enfatizar que o conhecimento é o produto final na análise de dados. KDD, então, é um processo para analisar e encontrar padrões presentes nos dados com o objetivo de obter conhecimento útil para o domínio do problema em questão.

Fayad *et al*. [23] definem o processo de KDD como uma sequência de passos iteráveis. A Figura 2.1 mostra cada passo e o resultado produzido, que serve como entrada para o passo seguinte. É possível retornar às fases anteriores se necessário, por exemplo, caso os dados disponíveis não produzam resultados suficientes, pode-se retornar ao inicio para encontrar novas fontes de dados.

O primeiro passo é entender o domínio do problema, encontrar conhecimentos relevantes e determinar o que se espera como resultado do processo. A partir disto, é feito a seleção do conjunto de dados e as variáveis que serão usadas para realizar a descoberta de conhecimento. Uma forma comum de representar estes dados é por tabelas atributo-valor. Neste caso, cada instância do conjunto de dados é representada por um conjunto de atributos (também chamado de características ou propriedades) que descrevem aquele exemplo no domínio do problema estudado. Cada atributo pode assumir diferentes valores, podendo ser de tipos distintos como um valor inteiro, real, ou categórico (um valor dentre um conjunto finito de categorias ou símbolos).

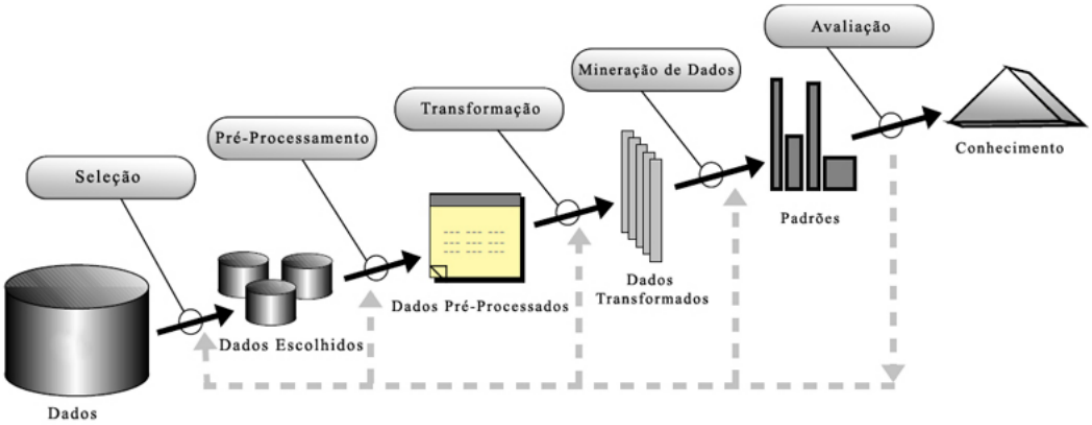

Figura 2.1: *Sequência de passos que compoem o processo de KDD e os produtos resultantes de cada iteração. Retirado de Fayad et al. (1996) [23].*

Em seguida, é feito a limpeza dos dados e o pré-processamento. Esse passo consiste na remoção de ruídos, como dados redundantes ou inconsistentes, e na determinação de uma estratégia para lidar com valores faltantes e *outliers* (dados que fogem do padrão do restante).

Na terceira fase, ocorre a transformação dos dados para o formato que seja possível aplicar os algoritmos de aprendizado de máquina. Além disso, nesta fase pode-se aplicar métodos de redução de dimensionalidade ou seleção de variáveis para reduzir o espaço em questão.

A fase de Mineração de Dados é a que costuma receber mais ênfase na literatura. Nesta fase, os algoritmos de busca de padrões são aplicados. Em alto nível, dois dos objetivos primários da mineração de dados são predição e descrição [64]. Predição envolve o uso de variáveis que descrevem os dados para predizer o valor de outra variável de interesse (comumente chamada de classe). Por exemplo, predizer se determinado paciente, a partir de seus sintomas atuais, terá uma recaída para depressão. Já a descrição tem o objetivo de encontrar padrões interpretáveis pelo ser humano, por exemplo, encontrar quais medicamentos são mais efetivos no tratamento do paciente com TAB.

Para atingir esses objetivos alguns métodos de Mineração de Dados podem ser usados, como: classificação, regressão, *clustering* (agrupamento), associação e sumarização (discutidos em mais detalhes na seção a seguir). Por fim, os padrões minerados são interpretados. A partir daí, o conhecimento é extraído e pode ser usado para tomada de decisões, integrado a outros sistemas e/ou simplesmente documentado.

2.1.1 Pré-processamento

O pré-processamento é uma das partes que compõe o processo de KDD. Em muitos casos, os dados são coletados sem muito controle ou de forma desorganizada, o

que resulta em valores incorretos ou fora de escala, valores faltantes, dados redundantes ou classes desbalanceadas. Essa combinação de fatores pode ser prejudicial para a tarefa de mineração de dados, portanto, diversas técnicas podem ser aplicadas para melhorar a qualidade dos dados.

Balanceamento de Classes

Classes são o conjunto de valores que a variável de saída, usada para predição, pode assumir. Portanto, um conjunto de dados é considerado desbalanceado se as classes usadas para predição não estão representadas igualmente [8]. Nesses casos, o desempenho do modelo criado para predição pode ser impactado de forma negativa. Por convenção, a classe que possui um número maior de exemplos é chamado de classe majoritária, caso contrário, ela é chamada de classe minoritária.

Para balancear os dados, o algoritmo *Synthetic Minority Over-sampling Technique* (SMOTE) [9] cria amostras sintéticas da classe com o menor número de exemplos, chamada de classe minoritária. Primeiramente é escolhido aleatoriamente um número l de exemplos da classe minoritária. Em seguida, para cada um deles (denotados por A_{atual}), um vizinho aleatório, denotado por $A_{vizinho}$, é escolhido dentre os cinco vizinhos mais próximos. Por fim, a nova amostra é criada conforme definido na Equação 2-1. Basicamente, a nova amostra é criada como um ponto no espaço entre A_{atual} e $A_{vizinho}$. O valor de Δ representa um número aleatório entre $[0,1]$.

$$A_{novo} = (A_{atual} - A_{vizinho}) \times \Delta + A_{atual} \tag{2-1}$$

No geral, SMOTE permite obter melhor desempenho dos classificadores, principalmente quando usado em conjunto com a técnica de *under-sampling*, que reduz o número de exemplos da classe majoritária [9].

Valores Faltantes

É possível encontrar bases de dados que apresentam valores faltantes (do inglês, *missing values*) em alguns campos. Em bases médicas, o paciente pode não querer expor alguma informação ou, ainda, a informação pode não existir para algumas pessoas, por exemplo um atributo medicamento pode não ser preenchido se alguns pacientes não estão tomando nenhuma medicação em determinado instante de tempo.

Alguns algoritmos não conseguem trabalhar com valores faltantes e, em muitos casos, é possível melhorar o desempenho dos classificadores com a aplicação de técnicas para preencher os dados que faltam.

Diversas técnicas foram criadas para isso, dentre elas a forma mais simples é preencher os valores que faltam de determinado atributo com a média (ou moda para variáveis categóricas) dos valores já conhecidos daquele atributo.

Seleção de Variáveis

Métodos de seleção de variáveis são usados para reduzir a dimensão dos dados, o que permite diminuir o tempo computacional e aumentar o desempenho dos modelos preditivos. Além disso, a seleção possibilita uma análise das variáveis que possuem maior impacto no desempenho dos classificadores, o que em muitos casos por si só já é uma informação útil [7].

As técnicas de seleção de variáveis podem ser organizadas em paradigmas. Dentre eles, o paradigma de filtros utiliza heurísticas relacionadas às propriedades dos dados. Assim, diferentes heurísticas foram desenvolvidas para encontrar um subconjunto ótimo de variáveis para predição.

Um método recente de seleção de variáveis, aplicado a dados clínicos, é o *RMean* [60]. Ele combina quatro algoritmos de seleção de variáveis baseados no paradigma de filtros para rankear as variáveis em ordem de importância. Os algoritmos Chi2 (*Chi-Squared*) [48] baseado na função estatística chi-quadrado, IG (*Information Gain*) [65] baseado em uma medida de ganho de informação, 1Rule (*One Rule*) [35] baseado em regras como função de avaliação e o *Relief* [14] baseado em medidas de distância. Com o resultado dos quatro algoritmos, *RMean* calcula uma média dos índices da posição de cada atributo no ranking gerado pelos algoritmos para obter o ranking final.

2.2 Aprendizado de Máquina

Aprendizado de máquina é um subcampo da ciência da computação que estuda a capacidade de computadores aprenderem com experiências anteriores. Atualmente, o aprendizado de máquina já é usado para automatizar certas tarefas, como reconhecimento de fala, ou para extração de conhecimento útil em bases de dados comerciais [53].

De modo geral, os algoritmos de aprendizado de máquina podem ser divididos em dois tipos: aprendizado supervisionado e aprendizado não-supervisionado. Os algoritmos supervisionados podem ser separados em algoritmos de classificação ou regressão. Já os não-supervisionados em algoritmos de *clustering*, associação ou sumarização.

No caso supervisionado, os algoritmos recebem variáveis de entrada e o rótulo (ou valor) esperado para a variável de saída. Com isso, eles buscam, em uma fase chamada de treinamento, pela função que melhor se ajuste aos dados. Por outro lado, os algoritmos de aprendizado não-supervisionado não sabem a priori a saída esperada para cada instância. O que eles fazem é buscar formas de descrever os dados, por exemplo,

agrupando as instâncias em determinados conjuntos ou determinando associações entre seus atributos.

Na classificação, deseja-se obter uma função que mapeie uma instância do dado em alguma classe dentro de um conjunto discreto de classes pré-definidas. Já na regressão, deseja-se uma função que mapeie a instância em uma variável do tipo real.

Nos algoritmos não-supervisionados, o método de *clustering* busca dividir o conjunto de dados em diferentes grupos. Em outras palavras, dado um número finito de grupos distintos, o objetivo é agrupar as instâncias em cada grupo, sendo que instâncias no mesmo grupo possuem propriedades semelhantes. Já regras de associação procuram encontrar relacionamentos em comum entre atributos do conjunto de dados, enquanto a sumarização procura descrever os dados de forma compacta, por exemplo por meio da média ou desvio padrão das variáveis [23].

Neste trabalho, foram utilizados algoritmos supervisionados para a análise de padrões de recaída para depressão no TAB. Uma breve descrição destes algoritmos é apresentada nas próximas seções.

2.2.1 Máquinas de Vetores de Suporte

Máquinas de Vetores de Suporte (do inglês, *Support Vector Machine*, SVM) é uma técnica inicialmente idealizada por Cortes & Vapnik [12] que segue a ideia de que os dados podem ser separados entre as classes por meio de uma reta, plano ou hiperplano, de acordo com o seu número de dimensões.

Para isso, as SVM buscam por um hiperplano que apresenta a maior margem de separação entre pontos de diferentes classes no espaço multidimensional, conforme observado na Figura 2.2. Nos casos em que o problema não é linearmente separável, o algoritmo faz um mapeamento das variáveis por meio de uma função *kernel*, que leva as instâncias para um espaço de dimensão superior onde as classes são melhor separáveis linearmente [11].

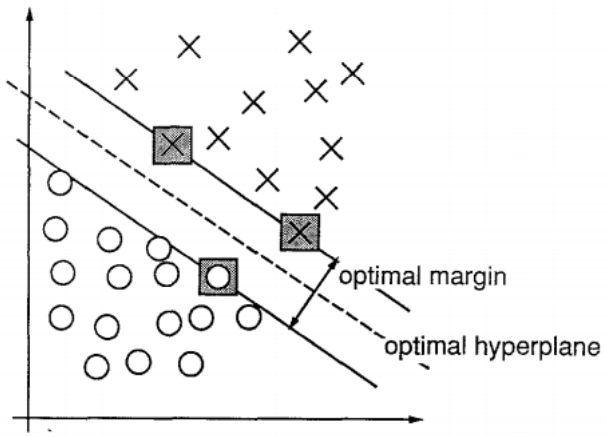

Figura 2.2: *Margem de separação criada pelo SVM para um problema de classificação binária. (Cortes & Vapnik [12])*

2.2.2 Florestas Aleatórias

Florestas Aleatórias (do inglês, *Random Forest*, RF) [4] pertencem a uma família de métodos chamados de *ensembles*. Tais métodos combinam um conjunto de classificadores que no fim agregam os seus resultados para gerar uma predição final. Dois métodos bastante conhecidos são *boosting* e *bagging*. Em *boosting*, os classificadores são construídos sequencialmente e dão peso extra para exemplos classificados incorretamente. Dessa forma, o próximo classificador terá maior chance de classificar corretamente aquele exemplo. Já no método de *bagging*, cada classificador é construído com um subconjunto de exemplos selecionado aleatoriamente do conjunto de dados [46].

RF se baseia no algoritmo de Árvores de Decisão (do inglês, *Decision Trees*, DT), para construir um conjunto de árvores para classificação. Uma DT é construída particionando os dados recursivamente. Em cada iteração, um atributo é escolhido como um nó da árvore a partir de uma função que determina o ganho de informação com a escolha do atributo. Em outras palavras, o atributo com mais informações, que permite melhor separar os exemplos de cada classe, é escolhido. Em seguida, são criadas arestas para cada valor do atributo e o processo é repetido em cada subárvore gerada até atingir um critério de parada. No final, as folhas da árvore representam a escolha de uma das classes [5]. RF é uma extensão do método de *bagging* que introduz um nível adicional de aleatoriedade. Além de usar um subconjunto aleatório de instâncias para a construção das árvores, a divisão de cada nó na geração de uma árvore é feita utilizando também apenas um subconjunto aleatório de atributos [4].

2.2.3 Redes Neurais Artificiais

Redes Neurais Artificiais (do inglês, *Artificial Neural Networks*, ANN) são algoritmos que surgiram inspirados pelo modelo biológico do cérebro humano [53]. De forma análoga, ANNs são compostas por um conjunto de unidades, ou neurônios, interconectadas entre si por arestas, chamadas de sinapses. As sinapses possuem pesos, que são ajustados de forma a minimizar o erro resultante na saída da rede. É neste ajuste em que ocorre o aprendizado.

As ANNs mais simples são compostas por um único neurônio, conhecido como *Perceptron*. Conforme ilustrado na Figura 2.3, o *Perceptron* recebe como entrada um vetor com os valores dos atributos $x = [x_1, x_2, ..., x_n]$ que se ligam ao neurônio pelo conjunto de sinapses com os respectivos pesos $w = [w_1, w_2, ..., w_n]$. Por sua vez, o neurônio computa a saída de acordo com a Equação 2-2 [53]. Para um problema de classificação binária, por exemplo, a saída do perceptron representa uma das duas classes.

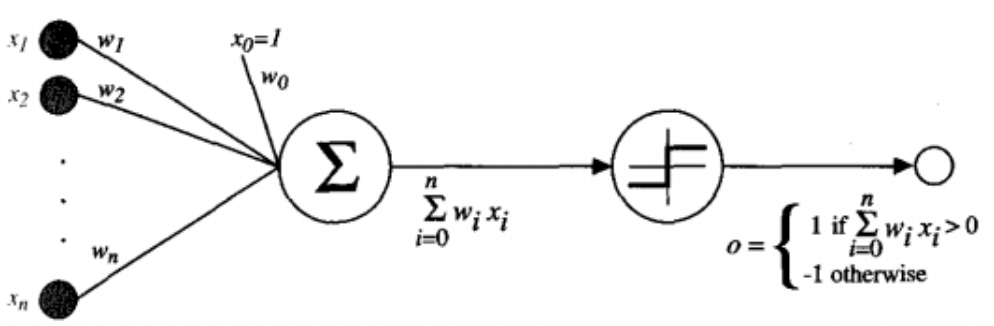

Figura 2.3: *Representação do perceptron. Retirado de Mitchel [53].*

$$o = \begin{cases} 1 & \text{if } \sum_{i=0}^{n} w_i x_i > 0 \\ -1 & otherwise \end{cases} \tag{2-2}$$

O aprendizado consiste então em ajustar o vetor de pesos w e um dos métodos mais utilizados para isso é o gradiente descendente. A partir de um vetor inicial de pesos (que pode ser escolhido aleatoriamente ou a partir de alguma heurística) e de uma função que determina o erro da saída do *Perceptron* para cada instância, os pesos são ajustados iterativamente para minimizar este erro.

Existem ainda, redes com vários neurônios, que são organizados em camadas, chamadas de *Multilayer Perceptron* (MLP) [32]. Este tipo de ANN permite resolver problemas mais complexos de aprendizado em que o conjunto de treinamento não é linearmente separável. No MLP, o aprendizado é feito por meio do algoritmo conhecido como *backpropagation*. Este algoritmo permite propagar o erro observado na saída de

toda a rede para cada sinapse em camadas intermediárias para que o peso seja ajustado na proporção em que ela contribui para o erro global.

2.2.4 Redes Bayesianas

Redes bayesianas são modelos probabilísticos baseados no Teorema de Bayes que representam um conjunto de variáveis e suas dependências condicionais por meio de um grafo acíclico direcionado (do inglês, DAG, *Directed Acyclic Graph*). Cada nó representa um atributo do domínio e arestas entre nós representam dependências condicionais. Para cada nó é possível calcular a sua probabilidade condicional dado os valores dos outros nós, com isso uma rede bayesiana pode ser usada como classificador para obter o valor do nó classe a partir da distribuição de probabilidade dos outros nós [10].

O algoritmo *Naïve Bayes* é uma especialização de redes bayesianas que supõe que todos os atributos são independentes entre si, o que simplifica a construção da rede, pois ela já é estabelecida a priori. Dessa forma, o atributo classe é o pai de todos os nós e mais nenhuma outra conexão é feita. Apesar de fazer uma suposição irreal, este algoritmo apresenta bom desempenho, comparável com outros classificadores mencionados anteriormente [26].

2.3 Aprendizado Relacional

Algoritmos que aprendem a partir de dados em uma única tabela atributo-valor também são chamados de algoritmos proposicionais. No entanto, os dados podem estar representados em múltiplas tabelas ou possuir relações intrínsecas entre si. Algoritmos que trabalham com este tipo de representação procuram por padrões em múltiplas tabelas de uma base de dados relacional [19].

Para que algoritmos proposicionais consigam trabalhar com esse tipo de dados é necessário convertê-lo em uma representação em tabela única. Por exemplo, suponha que uma base de dados contenha informações de pacientes armazenados em duas tabelas. Na primeira tabela são armazenados dados a respeito do paciente, como nome e data de nascimento. Já a segunda tabela contém os sintomas que um paciente pode apresentar em diferentes visitas ao consultório médico. A Tabela 2.1 mostra o esquema destas duas tabelas.

Tabela 2.1: *Exemplo de base de dados com duas tabelas.*

ID	Nome	Data_Nasc
1	João	01/05/1990
2	Antônio	22/11/1992
3	Maria	20/02/1998

(a) *Dados gerais dos pacientes.*

ID_Paciente	Sintoma_1	Sintoma_2
1	A	B
1	A	
1	A	C
2	C	
3	B	
3	B	D

(b) *Dados dos sintomas dos pacientes em cada visita ao médico.*

A Tabela 2.2 mostra a união das duas tabelas em uma. Para isso, foi preciso criar atributos que representassem cada visita do paciente e o seu sintoma nessa visita. Observe que conforme o número de visitas dos pacientes aumentam, novos atributos devem ser criados. Os atributos criados na tabela única passam a ser independentes, e acabam perdendo a relação de ordenação ou temporal das visitas dos pacientes e um mesmo valor de sintoma aparece em colunas separadas (em "atributos diferentes"). Além disso, isso gera uma tabela esparsa, pois nem todos os pacientes possuem o mesmo número de visitas. Portanto, todos esses fatores dificultam a aplicação de algoritmos proposicionais em bases relacionais.

Tabela 2.2: *Exemplo da união das duas bases de dados anteriores em uma só. O atributo $V_i\_S_j$ representa o sintoma j na visita i do paciente.*

ID	Nome	Data_Nasc	V1_S1	V1_S2	V2_S1	V2_S2	V3_S1	V3_S2
1	João	01/05/1990	A	B	A		A	C
2	Antônio	22/11/1992	C					
3	Maria	20/02/1998	B		B	D		

Outro problema é que bases relacionais introduzem maior complexidade para a tarefa de aprendizado, pois o número de padrões encontrados é potencialmente muito maior que o encontrado em uma única tabela. É necessário o uso de algoritmos que consigam suprimir o espaço de busca. Portanto, a exploração deve ser gerenciada de maneira eficaz para permitir encontrar conjuntos de padrões válidos e úteis para o domínio do problema [44].

Algumas técnicas de mineração de dados proposicional têm sido estendidas para permitir a busca por padrões em dados relacionais, como: regras de associação relacional, árvores de decisão relacionais e abordagens baseadas em distância relacional [19].

Um exemplo de dados multi-relacionais são dados longitudinais. Dados clínicos são comumente registrados desta forma. Cada vez que um paciente realiza uma consulta, um novo registro é adicionado à base de dados. Isso trás alguns problemas, pois nem sempre em cada visita o paciente realiza todos os exames, o que introduz valores faltantes. Além disso, as visitas nem sempre ocorrem no mesmo intervalo de tempo [37].

2.3.1 Programação Lógica Indutiva

Assim como grande parte dos algoritmos de mineração de dados vem do campo do aprendizado de máquina, grande parte dos algoritmos de mineração de dados relacional vem do campo da Programação Lógica Indutiva (do inglês, *Inductive Logic Programming*, ILP). A ILP é um campo de estudo que combina aprendizado de máquina e programação lógica. Assim como outros algoritmos de aprendizado de máquina, ILP busca induzir uma hipótese a partir de um conjunto de exemplos, porém com o uso de Lógica de Primeira Ordem (LPO) [24] como forma de representação do conhecimento [56, 19].

LPO, também conhecida por Cálculo de Predicados de Primeira Ordem ou Lógica de Predicados, é um sistema dedutivo formal usado em matemática, filosofia, linguística e computação. Diferentemente das linguagens naturais, LPO usa uma linguagem formal, sem ambiguidades, interpretada por estruturas matemáticas.

O uso de LPO como linguagem de representação em sistemas de aprendizado, permite que relações ou predicados possam ser induzidos. Isso faz com que o espaço dos conceitos passíveis de serem aprendidos seja aumentado. Esses sistemas possuem uma alta expressividade para representar conceitos e a habilidade de representar conhecimento do domínio. Além disso, os sistemas de aprendizado relacional têm a vantagem de expressarem seu conhecimento de uma forma diretamente inteligível aos humanos, característica muito importante quando o objetivo é a extração de conhecimento.

Em LPO, uma cláusula é definida como uma disjunção de literais, $L_1 \vee L_2 \vee \cdots \vee L_n$. Literais são formulas atômicas na forma $P(t_1, \cdots, t_m)$ no qual P é um simbolo de predicado e t_i são seus termos. Os literais podem aparecer na forma positiva ou na forma de negação, como $\neg P(t_1, \cdots, t_m)$ (a negação do literal P). Por sua vez, os termos representam variáveis, constantes ou expressões representadas por funções.

Cláusulas de Horn são formas específicas de cláusulas, com propriedades comumente usadas em programação lógica e ILP. Elas são definidas por cláusulas que contém no máximo um literal positivo. As cláusulas representadas nas Equações 2-3 e 2-4 são

exemplos de cláusulas Horn.

$$P \tag{2-3}$$

$$P \vee \neg Q \vee \neg R \vee \cdots \vee \neg U \tag{2-4}$$

A cláusula exibida na Equação 2-3 é comumente chamada de fato. Além disso, em ILP é comum escrever cláusulas representadas pela Equação 2-4 na forma de implicação, já que são logicamente equivalentes, conforme exibido na Equação 2-5.

$$P \leftarrow Q \wedge R \wedge \cdots \wedge U \tag{2-5}$$

A ILP traz duas grandes vantagens:

- produz classificadores que são de fácil entendimento por especialistas, e
- consegue resolver problemas de aprendizado multi-relacional

Outra vantagem é a possibilidade de expressar conhecimentos específicos de domínio do problema por meio do *background knowledge* (BK), graças ao poder de expressão da LPO como linguagem de representação [3].

Por exemplo, para estabelecer um sentido de relação entre atributos, é possível expressar os sintomas do paciente com TAB em relação a cada visita médica, ou ainda, definir a medicação tomada por um paciente em termos da sua dosagem especifica.

Formalmente, o aprendizado em ILP é definido como: dado o BK B, um conjunto de exemplos positivos E^+ e um conjunto de exemplos negativos E^-, ILP buscará por uma hipótese H tal que, $B \wedge H$ consiga derivar todos os exemplos positivos E^+, e nenhum exemplo negativo E^-. Assim, este conceito pode ser representado como: $B \wedge H \models E^+$.

A hipótese inferida H é definida como uma conjunção de cláusulas de Horn no formato definido pela Equação 2-5. Este conjunto de cláusulas, ou regras, é chamado de teoria. Para classificar um novo exemplo, cada regra da teoria é testada. Se houver correspondência com pelo menos uma delas, o exemplo é classificado como positivo, caso contrário ele é classificado como negativo.

Algumas aplicações de sucesso com ILP incluem: projeto de malha de elementos finitos, usado extensivamente por engenheiros para analisar o estresse aplicado em estruturas físicas [17]; predição da mutagenicidade de compostos químicos [72], e busca de regras que governam a estrutura da proteína [75]. Bratko & Muggleton [3] apresentam em mais detalhes outros casos de sucesso. Além disso, alguns estudos mostram o uso das regras geradas por ILP como atributos de entrada para outros classificadores, de forma a melhorar o desempenho do modelo final [42, 54].

Atualmente existem diversos sistemas que permitem trabalhar com ILP, como: Progol [55], FOIL [66], FORS [41] e Tilde [2]. Neste trabalho, utilizou-se o sistema Aleph (*A Learning Engine For Proposing Hypotheses*) [71] que permite simular diversas funcionalidades de outros sistemas por meio do ajuste de parâmetros, além de ser de uso livre. O único requisito para sua execução é a instalação de um compilador Prolog, podendo ser o Yap na versão 4.1.15 ou maior ou o compilador SWI Prolog na versão 5.1.10 ou maior.

O funcionamento do Aleph segue um processo que pode ser descrito em 4 passos:

1. Seleção: um exemplo positivo é selecionado do conjunto de treinamento para ser generalizado. Se não existir nenhum, o algoritmo termina, caso contrário segue para o próximo passo.
2. Saturação: é construída uma cláusula mais especifica que implique no exemplo selecionado, chamada de cláusula saturada (*bottom clause*).
3. Redução: é feita a busca pela cláusula de "melhor qualidade" que seja mais geral que a *bottom clause*. A qualidade de uma cláusula é determinada por uma função de avaliação.
4. Remoção de cobertura: a cláusula de "melhor qualidade" é adicionada à hipótese (teoria) e todos os exemplos cobertos por ela (redundantes) são removidos. Por fim, retorna ao passo 1.

Um dos passos mais importantes é o de redução, pois o algoritmo que realiza a busca pela melhor cláusula é o que possibilita enumerar cláusulas aceitáveis de forma inteligente e sob diferentes condições parametrizáveis.

2.4 Avaliação de Desempenho

Espera-se que bons algoritmos de aprendizado de máquina consigam gerar modelos que permitam a generalização. Isto é, após o treinamento do modelo com o conjunto de treinamento, espera-se que ele consiga uma alta acurácia para outros exemplos do domínio do problema.

Portanto, avaliar o desempenho desses algoritmos é importante tanto para estimar a sua acurácia, quanto para escolher entre um conjunto de classificadores. Para estimar a acurácia de um classificador, é desejável um método que tenha baixo *bias* (viés) e baixa *variance* (variância) [45].

Um modelo com alto *bias* pode fazer suposições erradas sobre o domínio do problema, de forma que ele não consiga estabelecer boas relações entre os atributos e a classe. Isso leva a uma baixa acurácia no conjunto de treinamento, chamado de *underfitting*.

Variance, por outro lado, acontece quando o modelo se ajusta muito bem ao conjunto de treinamento, porém não consegue generalizar bem para novas instâncias. Isso é chamado de *overfitting*.

Alguns métodos foram criados para estimar a acurácia dos classificadores. No geral, métodos que geram baixo *bias*, podem causar uma alta *variance*. O método *k-fold cross validation* consiste em dividir o conjunto de dados em k subconjuntos, mutuamente exclusivos e de mesmo tamanho. A partir daí, um desses conjuntos é usado para teste e os restantes $k-1$ são usados para treinamento. Esse processo é repetido alternando o conjunto de testes entre os k *folds*. Por fim, o erro médio é calculado entre todos os testes realizados e isso produz o desempenho final.

De acordo com Kohavi [45], com um valor moderado para k (10-20), este método produz baixa *variance* enquanto aumenta o *bias*. Conforme o valor de k diminui, o contrário é percebido.

A avaliação dos classificadores é feita nos conjuntos de teste. Os números de acertos e erros do classificador podem ser organizados numa matriz chamada *matriz de confusão*. A Tabela 2.3 ilustra uma matriz de confusão para um problema de duas classes, ditas *positiva* e *negativa*. Nesta matriz, TP (*True Positive*) representa a quantidade de exemplos que o modelo classificou como sendo da classe positiva e que de fato eram; FN (*False Negative*) são os erros em que o modelo classificou os exemplos como negativos, porém na verdade eles eram positivos; FP (*False Positive*) são os erros em que o modelo classificou os exemplos como positivos, quando eles eram negativos; e, TN (*True Negative*) representam a quantidade de acertos negativos.

Tabela 2.3: *Forma geral de uma matriz de confusão para um problema com duas classes.*

		Valor Predito	
		positivo	negativo
Valor Real	positivo	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	negativo	Falso Positivo (FP)	Verdadeiro Negativo (TN)

A partir destes valores, é possível calcular várias métricas de avaliação. As métricas mais usadas são: acurácia (Eq. 2-6), precisão (Eq. 2-7), sensibilidade (Eq. 2-8), especificidade (Eq. 2-9) e *f-measure* (Eq. 2-10). Estas métricas são importantes para validar diferentes aspectos dos resultados gerados pelos modelos.

A acurácia pode ser entendida como a taxa geral de acertos do modelo, considerando tanto os acertos para exemplos positivos, quanto para negativos. A precisão mede a taxa de acertos apenas entre os exemplos em que o classificador atribuiu como positivos. A sensibilidade olha apenas para as instâncias positivas do conjunto de dados, e mede a taxa de acertos entre estes exemplos. Já a especificidade pode ser considerada o complemento da sensibilidade, pois ela mede a taxa de acertos entre os exemplos negativos do conjunto de dados. O *f-measure* é uma média harmônica entre a precisão e a sensibilidade buscando uma medida de compromisso (*trade-off*) entre estas duas.

$$Acuracia = \frac{TP+TN}{TP+TN+FP+FN} \tag{2-6}$$

$$Precisao = \frac{TP}{TP+FP} \tag{2-7}$$

$$Sensibilidade = \frac{TP}{TP+FN} \tag{2-8}$$

$$Especificidade = \frac{TN}{FP+TN} \tag{2-9}$$

$$F\text{-}measure = 2 * \frac{Precisao * Sensibilidade}{Precisao + Sensibilidade} \tag{2-10}$$

CAPÍTULO 3

Revisão Bibliográfica Sistemática

O objetivo deste capítulo é apresentar uma revisão sistemática sobre características e métodos usados em bases de dados, para predição de recaída para depressão em pacientes com TAB. Espera-se entender o tema e investigar quais métodos são mais usados, quais ainda podem ser usados e os resultados obtidos a partir da sua aplicação. Esta revisão foi feita de acordo com o processo descrito por Kitchenham (2004) [43].

3.1 Materiais e Métodos

Para verificar o estado da arte e aprofundar o conhecimento sobre o tema estudado, as seguintes perguntas foram elaboradas: quais metodologias e bases de dados são usadas para criar modelos preditivos de depressão para recaída em TAB? Quais características ou atributos são incluídos nesses modelos e quais os resultados obtidos?

Os artigos foram buscados em 3 bases: *PubMed*, *IEEEXplore* e *ACM Digital Library*. Apenas o primeiro trouxe resultados em relação a *string* de busca criada para a pesquisa, descrita logo abaixo.

> ("step bd"OR "clinical data"OR "longitudinal data") AND ("bipolar disorder"OR "bipolar affective disorder"OR "bipolar depression") AND (depress* OR "recurrence"OR "depressive relapse"OR "mood episode*"OR "remission"OR "depression recovery") AND ("machine learning"OR "data science"OR "data mining"OR "knowledge discovery"OR predict* OR classificat*)

Foram mantidos artigos que buscavam predizer variáveis relacionadas a recaída para depressão, como: tempo de recuperação e frequência de episódios de depressão, aumento em sintomas de depressão e reduzida qualidade de vida associada ao risco de recaída. Por outro lado, o critério de exclusão consistiu de: artigos que abrangiam outros tipos de transtornos mentais ou, eram sobre TAB, porém não sobre recaída.

A Figura 3.1 ilustra o processo de seleção dos artigos. Foram encontrados 84 artigos no *PubMed*, dentre os quais 46 foram excluídos após leitura do título e do resumo.

Posteriormente, mais 8 artigos foram excluídos após leitura completa do texto, por não terem relação com TAB ou recaída para depressão. Outros 5 artigos foram excluídos por não conseguir obter o texto completo. Pela busca manual de referências bibliográficas, foi incluído mais 2 artigos. Portanto, a pesquisa incluiu 27 artigos no total.

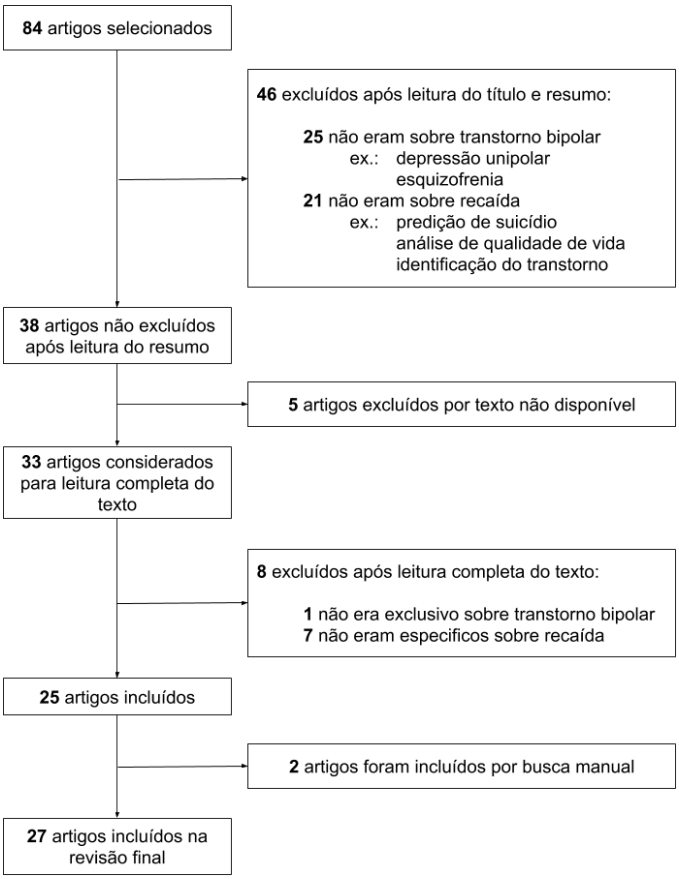

Figura 3.1: *Processo de seleção dos artigos incluídos na revisão*

3.2 Resultados

Os artigos foram separados de acordo com a metodologia usada para criar modelos de predição de recaída. A Tabela 3.1 mostra os artigos categorizados pela metodologia e a base de dados. Os artigos são identificados pelos autores e ano de publicação na segunda coluna, a terceira coluna mostra quais foram as metodologias empregadas no estudo e a ultima coluna identifica a base de dados usada.

Dos 27 estudos, 21 (77,8%) foram feitos com dados do STEP-BD. Em relação à metodologia, observa-se que alguns trabalhos usaram mais de uma técnica. Dentre as técnicas utilizadas, regressão foi a mais comum, relatado em 21 (77,8%) artigos. Análise de sobrevivência foi usada em 9 (33,3%) casos. Em 2 (7,4%) artigos recentes de 2014 e 2016, foram usadas técnicas de aprendizado de máquina: ILP e algoritmos de *clustering*, respectivamente.

Um único trabalho usou apenas testes estatísticos [51], porém todos os outros trabalhos também usaram esses testes para validar a hipótese que se queria observar e estabelecer a relação entre os dados, por isso não foram incluídos na Tabela 3.1.

Tabela 3.1: *Metodologia dos trabalhos selecionados sobre preditores de recaída em pacientes com TAB.*

#	Autor/Ano	Metodologia	Base de Dados
1	Deckersbach *et al*. (2016)	*Clustering*/Análise de Sobrevivência/Regressão	STEP-BD
2	Salvini *et al*. (2015)	ILP	108 pacientes em São Paulo
3	Busch *et al*. (2012)	Regressão	STEP-BD
4	Stange *et al*. (2016)	Análise de Sobrevivência/Regressão Linear	STEP-BD
5	Peters *et al*. (2014)	Análise de Sobrevivência/Regressão	STEP-BD
6	El-Mallakh *et al*. (2015)	Regressão	STEP-BD
7	Hochman *et al*. (2014)	Regressão	43 pacientes em Petah Tikva
8	Zhang *et al*. (2006)	Regressão	STEP-BD
9	Hong *et al*. (2016)	Regressão	3896 pacientes na China
10	Drago *et al*. (2015)	Regressão	STEP-BD
11	Magalhães *et al*. (2012)	Regressão	STEP-BD
12	Jaffe *et al*. (2009)	Regressão	115 pacientes em Belmont
13	Stange *et al*. (2013)	Análise de Sobrevivência/Regressão	STEP-BD
14	Fabbri e Serretti (2016)	Regressão	STEP-BD
15	Ostacher *et al*. (2009)	Análise de Sobrevivência	STEP-BD
16	Marsh *et al*. (2008)	Testes Estatísticos (Wilcoxon, Mann-Whitney, t-test)	STEP-BD
17	Dias *et al*. (2011)	Análise de Sobrevivência	STEP-BD
18	Bernstein *et al*. (2016)	Regressão	STEP-BD
19	Perlis *et al*. (2006)	Análise de Sobrevivência	STEP-BD
20	Waxmonsky *et al*. (2005)	Regressão	STEP-BD
21	Marsh *et al*. (2012)	Regressão	STEP-BD
22	Otto *et al*. (2006)	Análise de Sobrevivência/Regressão	STEP-BD
23	Perlis *et al*. (2009)	Análise de Sobrevivência/Regressão	STEP-BD
24	Gruber *et al*. (2011)	Regressão	STEP-BD
25	Cretu *et al*. (2016)	Regressão	STEP-BD
26	Novis *et al*. (2014)	Regressão	102 pacientes no Rio de Janeiro

Os trabalhos se diferenciam principalmente nas características analisadas. Dis-

túrbio no sono, por exemplo, é uma característica que parece ter relação com a recorrência nos pacientes com TAB, usado inclusive como critério para identificar episódios de mania ou depressão.

Gruber *et al*. (2011) [30] e Cretu *et al*. (2016) [13], por meio de técnicas de regressão, mostraram que a diminuição no tempo total de sono está associado a aumento de sintomas de mania, e uma alta variância no tempo de sono, está associado a aumento de sintomas de mania e depressão. Além disso, em média, pessoas com TAB tem maior prejuízo na eficiência do sono que pessoas saudáveis.

Perlis *et al*. (2006) [62] analisaram variáveis preditoras de recaída em dados sócio-demográficos e histórico clínico dos pacientes com análise de sobrevivência. As variáveis encontradas foram:

- uso de substâncias químicas,
- presença de transtornos de ansiedade,
- transtorno alimentar sofrido alguma vez na vida,
- mais de 20 episódios anteriores de alteração de humor,
- presença de sintomas residuais de depressão.
- proporção de dias com depressão no ultimo ano,

Por meio disso, verifica-se que o histórico anterior do TAB nos pacientes são características fortes para predizer novas recaídas. Em seu trabalho mais recente, Perlis *et al*. (2009) [61] mostraram, por meio de análise de sobrevivência e regressão linear, que quanto antes a doença se manifesta nas pessoas, pior o seu curso.

De forma semelhante, Peters *et al*. (2014) [63] mostraram que o curso prévio do TAB (quantidade de episódios no passado, idade no primeiro episódio e duração dos episódios) são boas variáveis para predizer o tempo até a recuperação em novos episódios de depressão. As técnicas usadas foram análise de sobrevivência e regressão.

Magalhães *et al*. [49] usaram regressão logística para mostrar que pacientes com múltiplas recaídas no passado tem um curso pior da doença, isso inclui: pior funcionamento e qualidade de vida e sintomas mais crônicos e severos.

Períodos de significante flutuação hormonal, principalmente através do ciclo menstrual nas mulheres, podem acarretar em maior risco de alterações de humor. Os trabalhos de Marsh *et al*. (2008, 2012) [51][50] mostraram que existe uma relação entre o período de transição para a menopausa com maior frequência de episódios de depressão. As técnicas usadas foram testes estatísticos para estabelecer a relação entre as variáveis e os dados, e regressão para determinar a frequência dos episódios.

Ainda em relação à mulheres, Dias *et al*. (2011) [16] analisaram o curso do TAB, indicado por episódios de alteração de humor, tempo mais curto até recaídas e elevação de sintomas de depressão e mania, com relação a irritação pré-menstrual nos

pacientes femininos. Eles concluem, após aplicar técnicas de análise de sobrevivência, que elas sofrem com um maior fardo nos sintomas, possivelmente devido a flutuações nos hormônios reprodutivos ao longo do ciclo menstrual.

Infelizmente, ainda não se sabe ao certo a patogenia do TAB. Porém, acredita-se ser um resultado entre a interação de fatores genéticos e ambientais. Os estudos de Fabbri e Serreti (2015) [21] e Drago *et al*. (2015) [18] investigaram a relação entre depressão e a genética dos pacientes com TAB com uso de regressão linear.

Estilos de atribuição podem ser definidos sobre como a pessoa atribui a si mesma a ocorrência de determinados eventos. Um estilo pessimista é a tendência da pessoa de atribuir causas de eventos negativos a motivos internos (por exemplo, "Eu fui demitido porque sou inútil") e a causa de eventos positivos a motivos externos (por exemplo, "Recebi a promoção porque tive sorte"). Stange *et al*. (2013) [73] usaram técnicas de análise de sobrevivência e regressão linear para predizer o tempo de recuperação dos pacientes em relação ao seu estilo de atribuição. A conclusão é que tanto estilos extremamente pessimistas, quanto otimistas, levam a um tempo maior para recuperação.

Outra relação com sintomas de depressão é investigada por Emily *et al*. (2016) [1], que estudaram a percepção de saúde física dos pacientes. De modo geral, indivíduos com uma visão mais negativa de sua saúde física tendem a ter mais sintomas de alteração de humor.

A instabilidade afetiva (instabilidade nos sintomas de depressão e mania) é avaliada como variável para predizer a duração dos episódios no TAB em Stange *et al*. (2016) [74]. Por meio da criação de um modelo de análise de sobrevivência e técnicas de regressão logística, os autores comprovam a hipótese de que instabilidade afetiva prediz maior duração até recuperação.

El-Mallakh *et al*. (2015) [20] demonstraram, com regressão linear, que o uso de medicamentos antidepressivos estava associado a uma taxa três vezes maior de episódios depressivos em pacientes com ciclagem rápida (pacientes que tem um número alto de episódios de alteração de humor em um curto período de tempo, mais especificamente, 4 ou mais episódios em um período de 12 meses).

O TAB muitas vezes vem acompanhado de outras comorbidades como, transtornos de ansiedade e abuso de álcool e outras substâncias químicas. Waxmonsky *et al*. (2005) [78], usaram de regressão linear para criar um modelo consistente com a hipótese, que pacientes fumantes e com TAB tendem a ter maior agitação e irritabilidade durante episódios de alteração de humor que não-fumantes.

Em linha de pesquisa parecida, Ostacher *et al*. (2009) [58], encontraram, com técnicas de análise de sobrevivência, que o uso de substâncias por pacientes de TAB produz uma troca mais rápida entre os estados do TAB (depressão, mania/hipomania e estado misto).

Para verificar a relação do consumo de bebidas alcoólicas com TAB, Jaffe *et al*. (2009) [38] usaram regressão linear em dados de 115 pacientes com TAB do McLean Hospital localizado na cidade de Belmont nos EUA, e mostraram que cada dia de consumo de álcool no mês atual, aumentava as chances do paciente experimentar um episódio depressivo no mês seguinte por cerca de 3,6%.

Transtornos de ansiedade apresentados por pacientes com TAB foram analisados no trabalho de Otto *et al*. (2006) [59]. Eles provaram que quanto maior o número de diferentes transtornos de ansiedade afetam o paciente, maior a chance de recaída. Dentre os transtornos individuais, apenas transtorno de ansiedade social e PTSD (*Post-Traumatic Stress Disorder*) elevam significativamente o risco de recaída nos pacientes que também possuem TAB.

A qualidade de vida é prejudicada nos portadores do TAB, principalmente durante episódios de alteração de humor. Com uso de modelos lineares, Zhang *et al*. (2006) [79] demonstraram que sintomas depressivos estavam fortemente associados com reduzida qualidade de vida.

Todos os trabalhos citados até então tiveram como metodologia o uso de técnicas estatísticas para análise dos dados. Um trabalho que usou uma técnica de aprendizado de máquina foi o de Deckersbach *et al*. (2016) [15]. O algoritmo *k-means* [31] foi usado para criar *clusters* de forma a agrupar pacientes com pouca ou muita recorrência. Em seguida foi utilizada análise de sobrevivência para determinar quais grupos tem maior probabilidade de se recuperar de episódios de recaída com duas formas distintas de tratamento.

O trabalho de Salvini *et al*. (2015) [69] é o único até o momento que usou aprendizado de máquina supervisionado para encontrar padrões de recaída para depressão. Eles aplicaram a ILP em uma base de dados longitudinal com 211 características de informações sócio-demográficas e clínicas de 108 pacientes do Programa de Transtorno Bipolar (PROMAN) do Instituto de Psiquiatria da Universidade Estadual de São Paulo (IPq-HCFMUSP). Foram geradas 6 regras que comprovam resultados encontrados na literatura. O modelo de predição atingiu acurácia de até 85% para identificar se um paciente terá recaída. Esse resultado demonstra o poder da ILP e, consequentemente, aprendizado de máquina relacional para encontrar padrões em dados clínicos de pacientes com TAB.

Poucos trabalhos encontrados usaram bases diferentes do STEP-BD. No geral esses estudos foram feitos com grupos pequenos de pacientes. A única exceção é o trabalho de Hong *et al*. (2016) [36], que usaram dados de 3.896 pacientes de 26 hospitais na China para determinar a relação entre o tempo até o diagnóstico efetivo do TAB (DUBP, *duration of undiagnosed bipolar disorder*) e a frequência de recaídas. O DUBP representa, em outras palavras, a quantidade de tempo que se passou do primeiro episódio de depressão até o efetivo diagnóstico do paciente como bipolar. Foi demonstrado, com

regressão linear, que um tempo maior do primeiro episódio de depressão até o diagnóstico do TAB contribui para um aumento de recaídas.

Hochman *et al*. (2014) [33] analisaram a concentração de fluídos e eletrólitos corporais com relação a episódios de depressão e mania. Os dados são de 43 pacientes do Geha Mental Health Center em Petach Tikva no Israel. Nos resultados foi encontrado níveis mais baixos de sódio, hemoglobina e concentração de albumina e hematócrito em episódios de mania comparado aos de depressão. O estudo foi feito com uso de regressão linear.

Dados de 102 brasileiros com TAB foram examinados por Novis *et al*. (2014) [57], para determinar fatores preditivos de cronicidade. Os resultados estão de acordo com outros estudos internacionais: maior duração e pouca idade no inicio da doença, mais episódios depressivos, sexo feminino e polaridade depressiva no primeiro episódio são fortes indicadores para predizer sintomas depressivos mais intensos.

O trabalho de Librenza-Garcia *et al*. (2017) [47] faz uma revisão do uso de técnicas de ML no estudo do TAB. Embora o seu foco seja em estudos que avaliassem o diagnóstico, eles também incluíram trabalhos relacionados a tratamento e prognóstico. Entre eles, o uso de ML para predizer recaída para depressão foi encontrado apenas em um trabalho [69]. Porém, quatro artigos avaliaram a predição de alterações de humor e o estado afetivo dos pacientes baseado em medidas de: séries de intervalos de batimento cardíaco extraído de eletrocardiogramas e sinais respiratórios [76]; eletrocardiogramas, respirogramas e dados de postura corporal [77]; atributos de variabilidade da taxa de batimento cardíaco em eletrocardiograma [28]; e, atributos de voz coletados em chamadas de voz em *smartphones* [22].

Por fim, o artigo de Busch *et al*. (2012) [6] analisou o uso de bases de dados eletrônicas para predição acurada de alterações de humor em TAB. A ideia principal era verificar se é possível criar modelos preditivos a partir de pequenas bases com informações limitadas (comumente encontradas em arquivos administrativos). Isso permitiria criar formas mais simples e com menos custos para guiar políticas de saúde. A metodologia consistiu em criar dois modelos, um mais detalhado, e outro limitado (com um subconjunto dos dados do STEP-BD). Para estimar a acurácia dos dois modelos foi usado regressão logística. Os resultados mostram que é possível criar modelos preditivos mais simples, pois ambos atingiram acurácia semelhante: 91% e 89% AUC (*Area Under the Curve*, métrica de desempenho usada em classificação), para o modelo detalhado e limitado, respectivamente.

3.3 Discussão

Mediante os estudos apresentados, observa-se que episódios de depressão podem ser causados por diferentes propriedades. Essas propriedades ou características, portanto, podem ser usadas para criar modelos preditivos de recaída. Nota-se que os trabalhos usam essas características de forma individual para identificar a existência de relação com a recaída.

Os principais fatores identificados como forte preditores foram:

- histórico prévio do TAB no paciente,
- distúrbios no sono,
- sexo feminino,
- predisposição genética,
- presença de outras comorbidades,
- estilo de atribuição extremamente otimista ou pessimista,
- percepção negativa de saúde física,
- alteração nos níveis de fluídos e eletrólitos corporais,
- sintomas residuais de depressão.

Estudos que usem esses atributos em conjunto ainda precisam ser desenvolvidos, com isso seria possível identificar quais os mais importantes e desenvolver tratamentos mais efetivos que lidem direto com aquela característica.

Verificamos também a importância das bases de dados para desenvolvimento dos trabalhos. Enquanto o STEP-BD reúne informações de diversas clínicas espalhadas por todo os EUA, outras bases, normalmente criadas a partir de dados de hospitais ou cidades específicas, possuem, em geral, menos instâncias, conforme visto em Salvini *et al*. (2014) [69], Jaffe *et al* (2009) [38], Novis *et al*. (2014) [57] e Hochman *et al*. (2014) [33]. O STEP-BD se consolida como a mais utilizada e por isso, dado sua dimensão e escopo, pode ser considerada como a maior fonte de informações clínicas sobre TAB.

Em relação à metodologia, a maioria dos trabalhos utiliza técnicas estatísticas como regressão, análise de sobrevivência e testes de hipótese. Uma lacuna a ser preenchida é o uso de algoritmos de aprendizado de máquina aplicados aos dados do STEP-BD. Com isso espera-se encontrar novos padrões e conhecimentos que permitam entender melhor como prevenir recaídas.

Com base nos resultados apresentados e na discussão feita, é possível responder as perguntas elicitadas na Seção 4: a principal base de dados usada é o STEP-BD, o principal método usado para criação dos modelos preditivos é regressão e as principais características usadas para gerar esses modelos são as elicitadas nos itens no início desta seção.

Portanto, é possível concluir que existe uma necessidade de explorar bases de dados, como o STEP-BD, com metodologias além de técnicas estatísticas. Com a aplicação do processo de KDD, espera-se descobrir novos conhecimentos, de forma a auxiliar a tomada de decisão médica e tratamentos, a fim trazer uma melhoria efetiva na qualidade de vida das pessoas com TAB.

CAPÍTULO 4

Métodos

Neste capítulo os métodos usados para execução dos experimentos são detalhados. Primeiro, a base de dados do STEP-BD é apresentada em mais detalhes. Segundo, é apresentado o algoritmo utilizado para seleção de amostras da base, ou seja, os pacientes que tiveram recaída para depressão. Terceiro, as variáveis utilizadas são descritas. Por fim, é apresentado uma análise inicial da distribuição das visitas dos pacientes coletadas pelo STEP-BD.

4.1 A Base de Dados STEP-BD

Os dados usados neste projeto são do *Systematic Treatment Enhancement Program for Bipolar Disorder* (STEP-BD) [68] que contém informações e dados clínicos de 4.360 pacientes com diagnóstico de algum transtorno mental dentro do espectro de doenças bipolares. A base do STEP-BD foi formada a partir de dados coletados por meio de diversos formulários, com diferentes avaliações e características dos pacientes. Os dois principais formulários são o ADE e o CMF.

O formulário ADE (*Affective Disorders Evaluation*) inclui informações usadas para estabelecer o diagnóstico do TAB, além de dados específicos como, idade de inicio da doença, estimativa do número de episódios anteriores, resposta a tratamentos passados, traumas, condições médicas, uso de substâncias psicoativas, histórico familiar, histórico menstrual e estado mental.

O CMF (*Clinical Monitoring Form*) contém dados longitudinais de seguimento que, em outras palavras, são registros das visitas médicas do paciente ao longo do tempo. Dentre os dados coletados no CMF estão medições dos sintomas de depressão e mania, uso de substâncias como cafeína, nicotina ou álcool, distúrbios do sono, se o paciente apresenta pensamentos ativos ou passivos de suicídio, presença de ataques de panico, dores de cabeça ou outras doenças significantes, além dos medicamentos em uso e suas dosagens. O CMF inclui ainda uma avaliação do estado do paciente, identificada pelo atributo *clinstat*. Esse atributo pode assumir um conjunto de valores como, "*Recovered*",

"*Recovering*", "*Depression*" ou "*Mania/Hipomania*", que indicam o estado clínico de humor do paciente em determinada visita.

Existem ainda outros formulários específicos que foram utilizados para coleta de informações para a formação da base de dados do STEP-BD, são eles: MINI (*Mini International Neuropsychiatric Interview*) e MINIF (MINI - *Followup*), com informações sobre comorbidades dos pacientes; MH (*Menstrual History*), com informações sobre eventos reprodutivos das mulheres; SQ (*Suicide Questionnaire*), com informações sobre tentativas de suicídio; e DF (*Demographic Form Study Entry*) com informações sócio-demográficas dos pacientes.

Os dados originais encontram-se em arquivos separados de acordo com o formulário de onde foram coletados, mas é possível relacionar estas informações através do código de identificação do paciente. A Tabela 4.1 mostra o número de variáveis e o número de observações de cada um destes arquivos. Alguns arquivos possuem mais de um registro de um paciente devido a mais de uma coleta de dados em períodos distintos.

É importante notar a natureza multi-relacional do STEP-BD. As tabelas estão relacionadas entre si, interligadas pelo *ID* do paciente. Além disso, existe uma relação temporal em algumas tabelas. No CMF, por exemplo, cada linha representa a visita de um paciente ao psiquiatra, sendo que um mesmo paciente pode ter varias visitas ao longo do tempo. Há também a relação entre atributos, como os medicamentos com a sua dosagem. Por isto, o uso de Mineração de Dados multi-relacional torna-se interessante para explorar esse tipo de dado [19].

Tabela 4.1: *Bases de dados original do STEP-BD*

Formulário origem	Número de variáveis	Número de observações
ADE	347	4107
CMF	148	50987
MINI	93	3730
MINIF	49	2262
MH	55	1745
SQ	31	2087
DF	86	6183

4.2 Seleção de Amostras

Foram selecionados pacientes que tiveram um seguimento de até 54 semanas após a primeira avaliação em estado de remissão, que é definido como 8 ou mais semanas consecutivas com dois ou menos sintomas de alteração de humor após uma alteração

aguda. O prazo de 54 semanas foi definido em conjunto com o médico especialista, baseado no trabalho de Perlis *et al.* [61].

A seleção dos pacientes foi feita pelo algoritmo ilustrado no fluxograma da Figura 4.1. Para cada paciente na base de dados, o algoritmo itera nas suas visitas até que ele encontre a primeira visita com o estado de remissão. A partir daí, o estado das visitas seguintes são iterados até a 54^a semana. Se uma visita com o estado de depressão for encontrada neste intervalo, o *ID* do paciente é salvo com a classe "*yes*", que indica uma recaída para depressão. Caso contrário, se o status se mantêm como remissão até o fim da 54^a semana, então o valor "*no*" é atribuido para a classe. Se ocorrer um outro estado que não o de depressão, busca-se por uma próxima visita em que o paciente entre em remissão; se não houver, o próximo paciente da base de dados é selecionado.

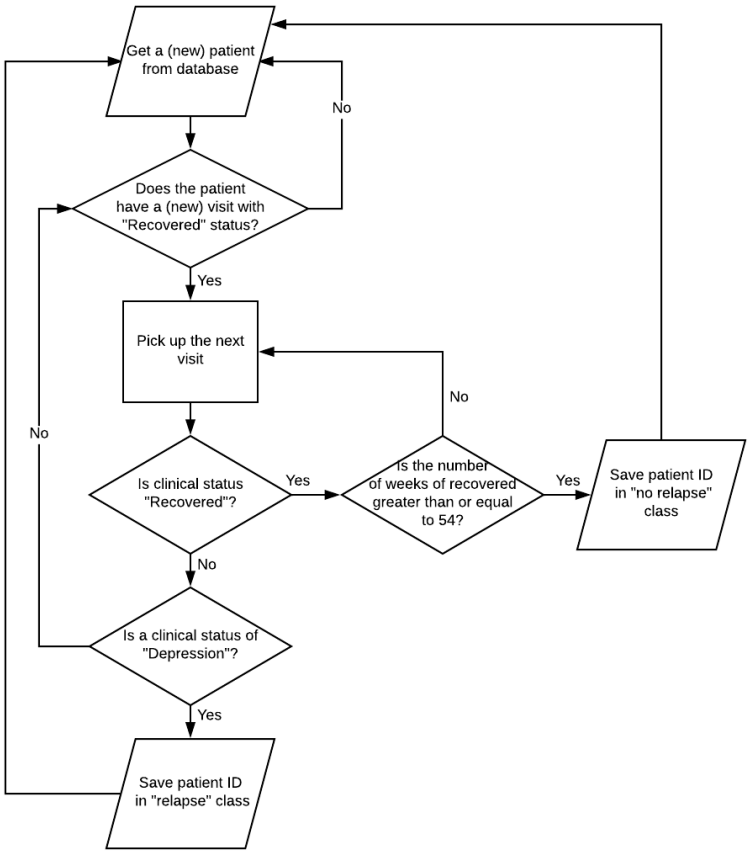

Figura 4.1: *Fluxograma do algoritmo de seleção dos pacientes.*

Após execução do algoritmo, foram selecionados no total $n = 800$ pacientes, $n_{pos} = 507$ que tiveram recaída, classificados na classe positiva, e $n_{neg} = 293$ que não tiveram recaída, classificados na classe negativa. Com base nisto, foram selecionados 3.436 visitas dos pacientes.

4.3 Seleção de Variáveis

Dentre os atributos do CMF, a principio foram incluídos atributos que representassem sintomas de mania/hipomania, quantidades mínimas e máximas de horas de sono por noite, uso de substâncias psicoativas ou álcool, pensamentos de suicídio, percentual da quantidade de dias em estados de depressão, alegria, mania, irritação ou ansiedade e seu nível, e a medicação em uso pelo paciente na respectiva visita, junto com a sua dosagem. Além disso, os atributos *visitid*, *stepid* e *recovered_week* representam o *ID* da visita do paciente, o *ID* do paciente e o número da semana em que aconteceu a visita, respectivamente. A primeira visita do paciente é de quando ele entrou em remissão e então o atributo *recovered_week* $= 0$. Para as demais visitas, este atributo é numerado conforme o número de semanas que se passaram desde a última visita.

Os atributos relacionados aos sintomas de depressão e mania podem assumir os valores numéricos mostrados na Tabela 4.2, que representam categorias que descrevem características e intensidades perceptíveis do sintoma apresentado pelo paciente. Esses valores recebem um sinal positivo ou negativo, que indica a direção do desvio em relação ao considerado normal. Estes atributos podem ainda assumir os valores -4, -5, -6, -7 e -8, que representam "*missing*", "*n/a*", "*unknown*", "*refused*" e "*not on original form*", respectivamente, e foram considerados como valores faltantes.

Tabela 4.2: *Valores atribuídos aos sintomas de acordo com a intensidade observada no paciente.*

0 **Nenhum ou comum**		
Reduzido	**Descrição da intensidade do sintoma**	**Aumentado**
-0,25	*Questionável, fraco, sintomas raros (ocorreram uma vez ou duas), mas sem significância clínica*	+0,25
-0,5	*Sintomas claramente presentes, mas abaixo do limiar de diagnóstico* **Suave**	+0,5
-1	*Claramente presente e preenche critério de diagnóstico* **Moderado**	+1
-1,5	**Acentuado**	+1,5
-2	**Grave**	+2

Os medicamentos estão representados em 12 atributos categóricos ($med1, med2, ..., med12$), onde cada um destes atributos pode armazenar o nome de um de 57 medicamentos possíveis, assim como suas respectivas doses ($dose1, dose2, ..., dose12$), representadas por um número real indicando a dosagem. Isso significa que o paciente poderia estar sob o uso de até 12 medicamentos de uma só vez. No geral, porém, observou-se poucos casos em que os pacientes tomavam mais de 1 ou 2 medicamentos, o que produziria uma tabela esparsa, com vários *missing values*.

Para facilitar a interpretação das regras geradas no Aleph pelo especialista, os valores numéricos dos atributos de sintomas foram substituídos por textos que representassem àquela categoria. Isso foi feito em Prolog e inserido no *background knowledge* por meio de predicados que traduzissem os valores numéricos, exibidos na Tabela 4.2, para o seu significado semântico. O conjunto de atributos usados e a discretização feita pode ser visualizado em mais detalhes na Tabela A.1 do Apêndice A.

4.4 Análise de Visitas

Os dados do CMF foram coletados com base em um estudo naturalístico, em que o médico interfere o menos possível no comportamento do paciente, o que inclui a frequência de visitas, que são feitas de acordo com a necessidade e escolha do paciente. Por isso, as visitas não seguem uma frequência pré-definida e variam de paciente para paciente. A Figura 4.2 mostra a quantidade de visitas que cada paciente teve. De forma visual, observa-se que os pacientes que não tiveram recaída para depressão em geral tiveram um número maior de visitas do que quem teve recaída. Isto acontece, pois foi incluído no estudo apenas as visitas até a recaída dos pacientes positivos, porém é possível que ele tenha mais visitas não adicionadas no espaço de amostras.

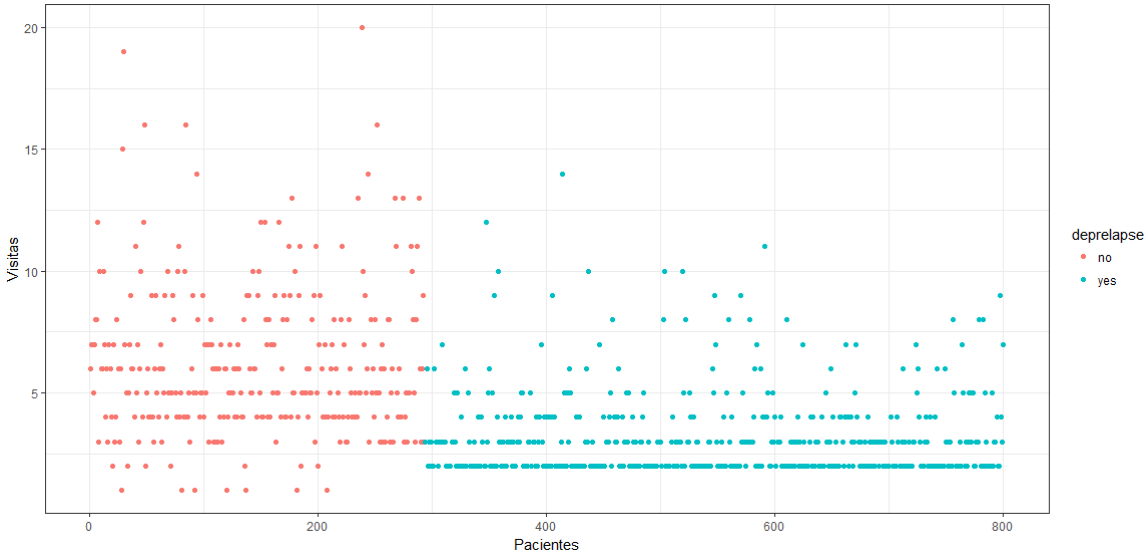

Figura 4.2: *Quantidade de visitas por paciente separados por quem teve recaída ou não.*

Para verificar se realmente há diferença entre as duas populações, primeiramente foi aplicado o teste de Shapiro-Wilk [67] para verificar se os dados das visitas possuem o comportamento de uma distribuição normal de probabilidade. Foi definido um intervalo de confiança de 95% ($\alpha = 0,05$). Para as duas populações, de pacientes que tiveram recaída e os que não tiveram, verificou-se que as visitas não seguem uma distribuição normal ($p = 2,2 \times 10^{-16}$, $deprelapse = yes$; e, $p = 1,11 \times 10^{-12}$, $deprelapse = no$).

A partir disso, foi aplicado o teste não-paramétrico de Wilcoxon [34] que verifica se há diferença entre as duas populações sem assumir que os dados seguem uma distribuição de probabilidade específica. O resultado mostrou que os pacientes de diferentes classes não pertencem à mesma população ($p = 2,2 \times 10^{-16}$), o que possibilita explorar os dados para encontrar padrões que mostrem a diferença entre pacientes que recaíram e não recaíram.

CAPÍTULO 5

Experimentos

Neste capítulo serão apresentados os 15 experimentos que tiveram resultados mais relevantes em relação ao objetivo deste trabalho. Primeiro, a estrutura geral dos experimentos é mostrada. Em seguida, são apresentados os parâmetros usados pelos algoritmos de classificação, e por fim os experimentos são apresentados em detalhes.

5.1 Estrutura dos Experimentos

Na Figura 5.1 é apresentado um esquema de organização destes experimentos. A numeração dos experimentos corresponde a sequência temporal em que foram executados. Além disso, eles foram agrupados em 3 grupos que possuem características semelhantes. Os grupos são definidos pelos conjuntos A, B e C. O Experimento 15 possui particularidades distintas e por isso foi incluido à parte. Adicionalmente, a cor de cada experimento denota o conjunto de atributos que foi usado, por exemplo, nos experimentos com a cor verde os atributos usados foram sintomas de depressão e mania, conforme descrito na legenda da figura. Na Tabela 5.1, é possível visualizar o desempenho geral de cada experimento.

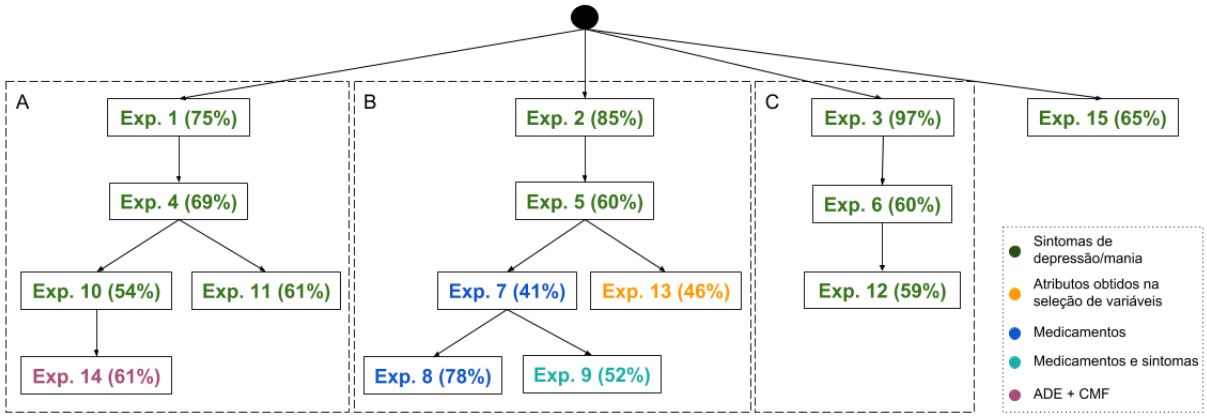

Figura 5.1: *Diagrama de experimentos agrupados pelas semelhanças e na sequência de execução (valores em parênteses representam a acurácia).*

Tabela 5.1: *Desempenho obtido em cada experimento.*

Experimento	Acurácia	Sensibilidade	Precisão	Especificidade
1	75,1% (0,05)	76,7% (0,06)	82,9% (0,05)	72,3% (0,09)
2	84,7% (0,03)	85,0% (0,06)	90,5% (0,03)	84,3% (0,06)
3	96,6% (0,03)	96,6% (0,04)	97,5% (0,03)	96,6% (0,04)
4	69,4% (0,04)	68,3% (0,05)	80,5% (0,05)	71,3% (0,09)
5	60,4% (0,06)	63,3% (0,09)	70,8% (0,04)	55,3% (0,05)
6	60,3% (0,06)	61,8% (0,04)	69,5% (0,08)	58,0% (0,15)
7	40,9% (0,04)	19,1% (0,04)	60,2% (0,10)	78,5% (0,05)
8	77,7% (0,06)	48,7% (0,12)	83,5% (0,12)	94,5% (0,04)
9	52,1% (0,06)	45,8% (0,09)	68,2% (0,05)	63,1% (0,06)
10	53,7% (0,08)	57,5% (0,07)	81,9% (0,06)	34,4% (0,25)
11	61,1% (0,07)	59,1% (0,07)	74,5% (0,07)	64,5% (0,12)
12	59,4% (0,05)	49,1% (0,06)	61,8% (0,08)	69,6% (0,11)
13	46,4% (0,07)	33,5% (0,08)	64,9% (0,11)	68,6% (0,12)
14	65,3% (0,06)	68,6% (0,06)	74,5% (0,06)	59,4% (0,10)

Valores em parênteses representam o desvio padrão de cada métrica entre os *folds*.

No conjunto de experimentos A, foi levado em consideração a semana da

visita dos pacientes. No geral, nesses experimentos buscou-se pacientes que possuíam determinada regularidade ou frequência de visitação.

No conjunto de experimentos B, as visitas foram agrupadas em intervalos. Para isso, as semanas foram discretizadas de forma que cada grupo contenha uma mesma quantidade aproximada de visitas. O resultado desta discretização pode ser observado na Figura 5.2. A semana 0 é a única em que todos os 800 pacientes possuem visitas pois é a semana quando o paciente entra em remissão e por isso ela foi omitida nos gráficos para permitir o ajuste de escala. O gráfico à esquerda mostra a distribuição de semanas que houveram visitas. A semana 4 foi a que houve mais visitas, em que 160 pacientes foram ao médico. As semanas foram agrupadas de forma que cada grupo tivesse uma quantidade próxima deste valor. O gráfico à direita mostra como ficou a distribuição de visitas após o agrupamento.

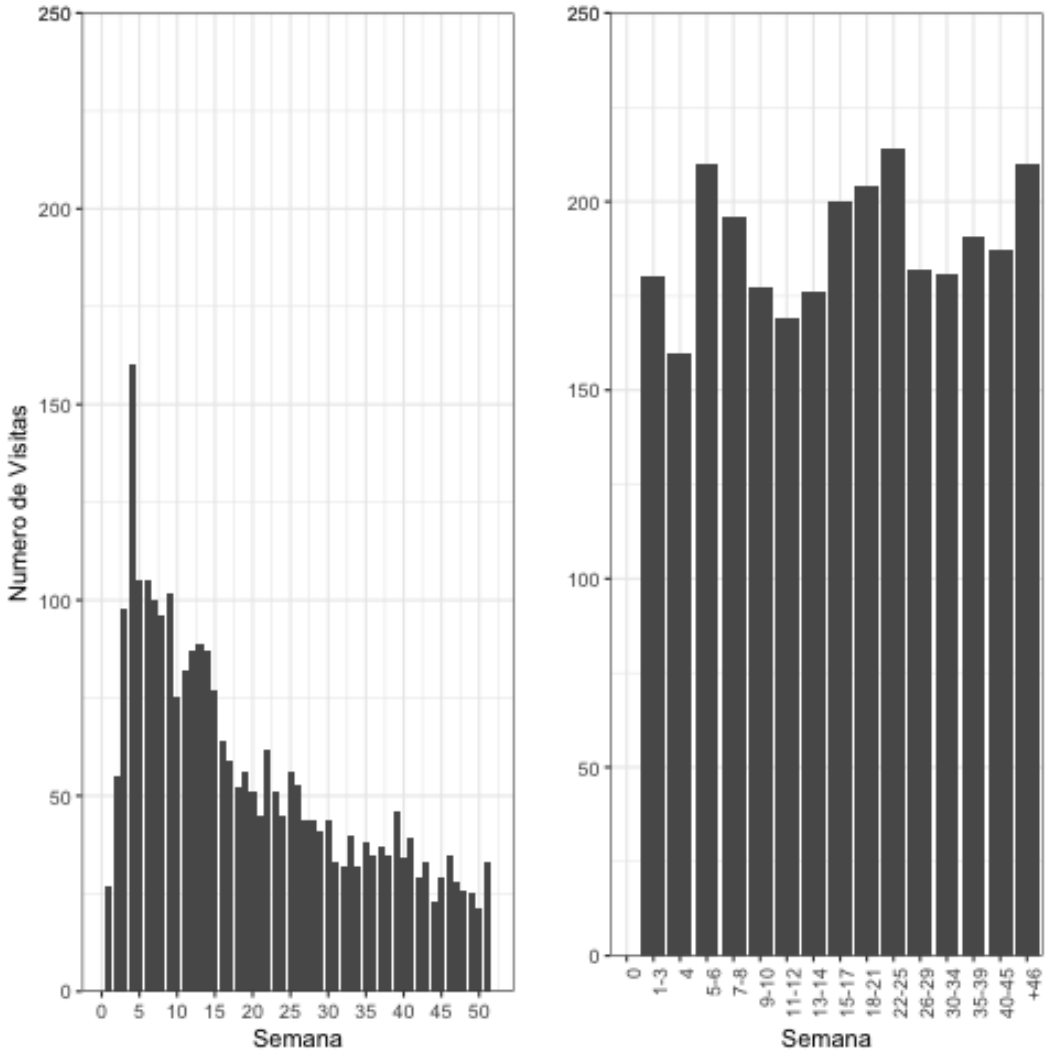

Figura 5.2: *Quantidade de visitas por semana.*

Já no conjunto de experimentos C, foi feito um pareamento entre as visitas dos pacientes positivos (que tiveram recaída) e negativos (que não tiveram recaída). O

processo de pareamento é ilustrado na Figura 5.3. Para cada exemplo positivo, buscou-se dentre o conjunto de exemplos negativos algum que tivesse visitas tanto na semana referente a visita imediatamente anterior à recaída (*pre_relapse_visit*), quanto na visita onde foi estabelecida a recaída (*relapse_visit*). Quando encontrado, o exemplo negativo foi incluído no estudo e a iteração segue para o próximo caso positivo. Nos casos que não houve combinação com nenhum negativo, o exemplo positivo foi removido. Além disso, foram consideradas todas as visitas da semana 0 (*baseline*) as quais todos os pacientes possuem.

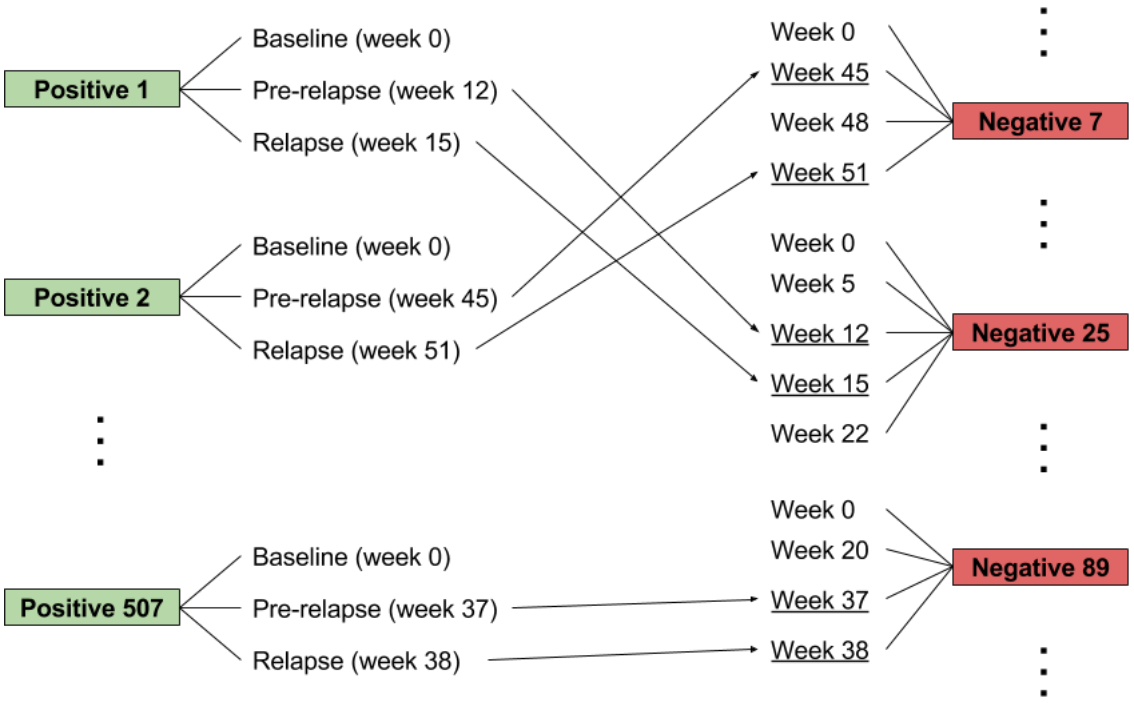

Figura 5.3: *Processo de pareamento dos exemplos positivos e negativos.*

5.2 Parâmetros dos Classificadores

Os experimentos 1 a 14 foram feitos com ILP e o experimento 15 usou os algoritmos proposicionais SVM, RF, MLP e *Naïve Bayes*. Os parâmetros usados no ILP foram $clauselength = 6$ (número máximo de predicados nas cláusulas), $nodes = 1000000$ (número máximo de nós na árvore de busca do sistema Aleph), $noise = 0$ (número máximo de exemplos negativos cobertos pelas cláusulas encontradas) e $minpos = 2$ (número mínimo de exemplos positivos cobertos pelas cláusulas encontradas). Para o SVM usou-se o kernel polinomial. No RF, 100 árvores foram construídas. Já no MLP, o número de camadas ocultas foi igual a $(atributos + classes)/2$ e a taxa de aprendizado $\alpha = 0.3$. O algoritmo *Naïve Bayes* não possui parâmetros a serem configurados.

5.3 Experimento 1 – Todas as Visitas

O experimento 1 consiste no uso de todas as 3.436 visitas dos 800 pacientes selecionados. Apenas os atributos de sintomas de depressão e mania foram usados. Além disso, foi criado o predicado `has_visit/3` que relaciona a visita ao paciente, ou seja, permite ligar os sintomas observados no paciente em cada uma de suas visitas.

A declaração do predicado `has_visit/3` é exibido no Código 5.1. Na sua definição de modo verifica-se que ele recebe como argumentos uma variável de entrada que representa o *ID* do paciente (`+stepid`), uma variável de saída que representa o *ID* da visita (`-visitid`) e uma constante com o número da semana dessa visita (`#recovered_week`). O predicado é montado a partir da conjunção (em Prolog representado por vírgula entre os predicados) dos atributos de *ID* do paciente e a semana da visita.

Código 5.1 `predicado has_visit`

```
1 :- determination(deprelapse/1, has_visit/3).
2 :- modeb(*, has_visit(+stepid, -visitid, #semana)).
3
4 has_visit(Paciente, Visita, Semana) :-
5         stepid(Visita, Paciente),
6         recovered_week(Visita, Semana).
```

Na primeira linha da Tabela 5.1, vemos as métricas de desempenho do modelo gerado por este experimento. Foi alcançado acurácia de 75,1%, sensibilidade de 76,7%, precisão de 82,9% e especificidade de 72,3%.

5.4 Experimento 2 – Visitas Agrupadas

O experimento 2 foi realizado com base no agrupamento mostrado no gráfico a direita da Figura 5.2. Para agrupar as visitas, foi criado um predicado em Prolog e adicionado ao BK. Basicamente, esse predicado verifica qual intervalo a visita pertence e substitui o número da visita no predicado `has_visit/3` pelo valor do intervalo.

Os atributos de sintomas de depressão e mania foram usados. Os resultados obtidos podem ser visualizados na segunda linha da Tabela 5.1. É possível notar uma melhora de quase 10 p.p. para cada métrica em relação ao experimento 1. Isso sugere que quanto maior os intervalos, melhor seria o desempenho. No entanto, isso prejudicaria a criação das regras, pois deseja-se encontrar padrões que indiquem como a evolução dos sintomas leva o paciente a ter recaída.

5.5 Experimento 3 – Visitas Pareadas

No experimento 3 foi feito o pareamento conforme explicado no grupo C. Após esse processo, a base de dados ficou com um número reduzido de instâncias. Foram selecionados 355 exemplos positivos e 265 exemplos negativos (total de 620 pacientes). Portanto, 180 pacientes foram removidos neste estudo.

Foram criados predicados para indicar quais eram as visitas *baseline*, *pre_relapse* e *relapse*. Esses predicados substituiram o predicado `has_visit/3` e permitiam determinar a qual visita o atributo de sintoma pertencia.

A terceira linha da Tabela 5.1 mostra os resultados obtidos neste experimento. Uma melhora significativa foi obtida, com todas as métricas atingindo desempenho superior a 96%.

Por fim, foi feito ainda outro experimento de forma similar. Porém, dessa vez o pareamento foi relaxado para permitir parear visitas com uma semana a mais ou a menos de diferença. Os resultados foram semelhantes e, por simplicidade, foram omitidos da Tabela 5.1.

5.6 Experimentos 4, 5 e 6 – Remoção das Visitas de Recaída

Os experimentos anteriores geraram taxas crescentes de desempenho. Foi levantada então a hipótese de que na ultima visita dos pacientes positivos, eles já apresentavam sintomas claros de depressão, pois é a visita que determina a recaída, logo, considerá-la poderia enviesar o modelo.

A partir disso, o experimento 4 é semelhante ao experimento 1, porém ele não inclui a ultima visita dos pacientes que tiveram recaída, chamada de visita de recaída. Novamente, os atributos de sintomas de depressão e mania foram usados. Observa-se na quarta linha da Tabela 5.1 uma redução considerável em todas as métricas de desempenho, com acurácia abaixo de 70%, o que prova que a visita de recaída poderia de fato enviesar o modelo.

Após perceber isto, a visita de recaída foi retirada dos experimentos a seguir. Portanto, os experimentos 5 e 6 apenas replicam os experimentos 2 e 3, porém sem o uso da visita de recaída. No experimento 2, o agrupamento foi mantido com os mesmos intervalos, enquanto no experimento 3, o pareamento foi feito apenas com os predicados referentes às visitas *baseline* e *pre_relapse*.

Com isso, verificou-se uma redução ainda maior nas taxas de desempenho, com acurácia próximo a 60% para ambos os experimentos, conforme observado nas linhas 5 e 6 da Tabela 5.1.

5.7 Experimentos 7, 8 e 9 – Inclusão de Medicamentos

Neste experimento os atributos de sintomas foram substituídos por atributos de medicamentos. A base do CMF inclui 12 atributos para medicamentos, pois o paciente pode usar mais de um medicamento em um determinada semana. Dessa forma, um mesmo medicamento poderia aparecer em diferentes colunas para diferentes pacientes.

Para trabalhar com esses dados no ILP, foi criado um predicado de medicamento, que especifica quais medicamentos o paciente tomava em determinada visita. Além do nome do medicamento, o CMF possui um atributo que indica a dose ingerida pelo paciente para aquele medicamento. Esse valor foi incluído no predicado de forma discretizada, podendo ser:

- 0-25% da dose máxima
- 26-50% da dose máxima
- 51-75% da dose máxima
- 76-100% da dose máxima

A dose máxima foi definida de acordo com o medicamento específico. Foram feitos experimentos tanto com e sem a dose no predicado. Quando incluída a dose, os resultados foram levemente inferiores. Portanto, a sétima linha da Tabela 5.1 mostra apenas o resultado sem a dose. A acurácia atingiu 40,9% e a sensibilidade 19,1%. Isso demonstra que apenas os atributos de medicamentos tem pouco fator preditivo.

A partir do resultado do experimento anterior, foi levantado a hipótese de que os atributos de medicamentos seriam melhor usados para predizer quando um paciente não terá recaída, pois o objetivo dos medicamentos é justamente previni-la. Isso pode ser explorado pela forma como o algoritmo do Aleph funciona: as melhores regras são as que cobrem mais exemplos positivos e ao mesmo tempo menos exemplos negativos. Por isso, ao inverter a classe de forma que os exemplos positivos passem a ser os pacientes que não tiveram recaída e os exemplos negativos os pacientes que tiveram recaída, esperava-se melhorar o desempenho do modelo.

Os resultados obtidos provam que a hipótese estava correta em partes, eles são exibidos na oitava linha da Tabela 5.1. A acurácia apresenta desempenho de 77,7%, uma melhora de mais de 35 p.p. em relação ao estudo anterior, como esperado. Porém, quando observado a sensibilidade, apenas 48,7% dentre os pacientes positivos foram classificados corretamente (neste caso, os que não tiveram recaída). Isso significa que dos pacientes que realmente eram positivos, aproximadamente apenas a metade foi classificada corretamente.

Por outro lado, o modelo apresenta uma taxa alta de especificidade, de 94,5%. Logo, esse modelo acerta bem quando um paciente terá recaída, que são os casos que

não são cobertos por nenhuma regra. Isso é um resultado importante clinicamente, pois permite ao médico dar atenção especial aos pacientes que realmente precisam.

Já no experimento 9, os atributos de medicamentos foram usados em conjunto com os sintomas de depressão e mania. Conforme observado na nona linha da Tabela 5.1, o desempenho ainda foi baixo, não muito melhor do que rolar uma moeda. Porém, quando comparado ao experimento 7 foi percebido uma pequena melhora, já que a acurácia subiu de 40,9% para 52,1%.

5.8 Experimento 10 – Pacientes com Maior Frequência

No experimento 10 voltou-se ao uso apenas dos sintomas. O objetivo deste experimento foi estruturar os dados longitudinais de forma que se assemelhassem a uma serie temporal, com a expectativa de que isso facilitasse a busca por padrões feitas pelo ILP.

Logo, foram incluídos apenas pacientes que possuíam maior frequência de visitas, com pelo menos uma visita a cada dois meses até sua ultima visita. Esse processo reduziu o tamanho da base, restando 1.218 visitas de 376 pacientes, em que 315 eram positivos e 61 negativos.

Diferente do esperado, esse processo não produziu bons resultados, com acurácia e sensibilidade ainda próximos de 50%, conforme visto na decima linha da Tabela 5.1.

5.9 Experimentos 11 e 12 – Intervalo até a Recaída e Pareamento Completo

Algumas variações foram feitas em experimentos anteriores para obter melhores resultados. O experimento 11 tem como base o experimento 4. Porém, o número da semana da visita no predicado `has_visit/3` foi substituído pela distância da visita atual até a ultima visita. Esperava-se facilitar a busca por padrões, visto que pacientes podem ter o mesmo intervalo de uma determinada visita até a visita de recaída, porém em semanas completamente diferentes. No entanto, quando comparado com o experimento 4, percebe-se uma piora em todas as métricas.

Já o experimento 12 se assemelha ao experimento 6, no qual foi feito o pareamento das visitas dos pacientes positivos e negativos. Desta vez, porém, o pareamento deveria ser completo. Em outras palavras, foram selecionados apenas pacientes positivos que possuíam um correspondente negativo com todas as visitas exatamente na mesma semana.

Como resultado, foram selecionados 293 pacientes positivos e 293 negativos. Novamente os atributos de sintomas foram usados. Foi obtido 59,4% de acurácia, ou seja, resultado semelhante ao observado no experimento 6.

5.10 Experimento 13 – Seleção de Variáveis

Neste experimento tentou-se aplicar algoritmos de seleção de variáveis. O algoritmo Rmean foi usado, que consiste em aplicar diversos algoritmos que ranqueiam as variáveis por ordem de importância e no final calcula a média dos ranks para obter a ordem final de importância.

Como o *Rmean* trabalha com dados em tabelas chave-valor, foram feitos dois testes; primeiro ele foi usado apenas sobre as visitas do baseline e em seguida apenas sobre as ultimas visitas dos pacientes, ou seja, as visitas de recaída. Após obter o rank de importância das variáveis, as dez primeiras variáveis foram selecionadas de forma empírica e usadas para criar o modelo com ILP.

No baseline as variáveis selecionadas foram: *cgi*, *depenerg*, *depinter*, *depconcn*, *gafweek*, *silnwl*, *elvdistr*, *gafmonth*, *moodelev* e *depdist*. Já na ultima visita as variáveis foram: *depinter*, *curenjoy*, *curdepr*, *deprmd*, *cgi*, *depresd*, *lessint*, *depconcn*, *depenerg* e *depse*.

Ambos os resultados foram semelhantes, portanto apenas o do primeiro teste foi exibido na Tabela 5.1, com acurácia e sensibilidade atingindo valores abaixo de 50%.

5.11 Experimento 14 – Inclusão de dados do ADE

Este experimento consistiu no uso de dados do formulário ADE em conjunto com o CMF, que vinha sendo usado unicamente até então. O ADE é uma entrevista clínica padronizada abrangente, que avalia a história psiquiátrica, estado atual, diagnóstico, episódios mais graves, padrão de sintomas de humor, infância, história social e familiar, tratamento anterior e histórico médico. Neste caso, o ADE representou um ponto de partida para comparações do estado do paciente com relação ao CMF que contem as visitas subsequentes.

Os sintomas foram representados como a variação do seu valor no ADE e o seu valor na visita atual do CMF. Para isso, foram criados predicados conforme o exibido no Código 5.2. Este predicado determina a variação do sintoma de depressão do paciente relacionado ao interesse em atividades prazerosas (atributo *depinter*). Foram criados predicados semelhantes para cada sintoma de depressão e mania.

Código 5.2 predicado variacao_depinter

```
1  :- determination(deprelapse/1, variacao_depinter/2).
2  :- modeb(*, variacao_depinter(+stepid, #dif)).
3
4  variacao_depinter(Paciente, Dif) :-
5    has_visit(Paciente, Visita),
6    depinter_ade(Paciente, V1),
7    depinter(Visita, V2),
8    recovered_week(Visita, Semana),
9    Semana < 13,
10   Dif is V2 - V1.
```

Além disso, foram incluídos ainda atributos do ADE, comumente usado pelo especialista para identificar uma possível recaída para depressão, são eles:

- *migraine*: se o paciente apresenta enxaqueca;
- *dm*: se o paciente apresenta diabetes mellitus;
- *thyroid*: distúrbios na tireoide;
- *bpiageon*: idade de início do TAB;
- *bpifamhx*: histórico familiar de transtornos mentais.

Após induzir a nova teoria e valida-la, verificamos uma melhora em relação aos experimentos anteriores que também não incluíram a visita de recaída, com acurácia atingindo 65,3%, sensibilidade 68,6% e precisão 74,5%.

5.12 Experimento 15 – Algoritmos Proposicionais

Como através do uso da ILP não produziu resultados esperados que pudessem ser considerados satisfatórios ou relevantes para uso clínico, com muitos apresentando acurácia próximo ou abaixo de 50%, tentou-se organizar as visitas de forma que fosse possível aplicar algoritmos proposicionais.

As visitas longitudinais do CMF foram usadas da seguinte forma: inicialmente os algoritmos foram aplicados apenas com a visita inicial de remissão (*baseline*), em seguida, foi usado o *baseline* e a última visita antes da recaída, depois, foram aplicadas com o *baseline* e a penúltima visita antes da recaída, depois com a antepenúltima e assim por diante até que não houvessem dados suficientes de visitas dos pacientes.

Dessa forma, foi possível aplicar os algoritmos até a sétima ultima visita, pois a partir desse ponto a quantidade de exemplos ficou bastante reduzida, uma vez que poucos pacientes, após entrar no estado de remissão, chegaram a ter mais de sete visitas antes da

recaída. A Tabela 5.2 mostra a quantidade de exemplos usados na classificação de cada visita. A visita "*Baseline*" representa a primeira visita e "Relapse" a ultima, "Last - 1" é a penúltima, a "Last - 2" é a antepenúltima e assim por diante.

Em alguns casos os dados apresentam um desbalanceamento entre as classes, o que é prejudicial para a tarefa de predição. Assim, foi usado o algoritmo SMOTE [9] para melhor balancear os casos em que a classe majoritária estava desproporcional em relação a classe minoritária.

Tabela 5.2: *Quantidade de exemplos para classificação em cada visita, antes e após o uso do algoritmo SMOTE. Nas visitas Baseline, Ultima e Ultima - 1 não foi aplicado SMOTE, pois o desbalanceamento não era grande entre as classes.*

Visita	**n^o de exemplos**		**n^o de exemplos com SMOTE**	
	Positivos	Negativos	Positivos	Negativos
Baseline	507	293	-	-
Relapse	507	286	-	-
Last - 1	271	279	-	-
Last - 2	137	250	274	250
Last - 3	79	197	158	197
Last - 4	47	143	94	143
Last - 5	32	103	64	103
Last - 6	21	78	42	78

Foram usados os algoritmos RF, NB, MLP e SVM, descritos no Capítulo 2. Estes algoritmos foram executados em cada visita. A Tabela 5.3 mostra os algoritmos que tiveram melhor acurácia em cada visita. Novamente, os dados da visita de recaída (*Relapse*) apresentaram um desempenho bastante superior as demais. É possível notar ainda, um desempenho maior em algumas visitas intermediarias, como *Last - 4* com 86,4% de acurácia.

Esses resultados foram descritos e discutidos em um artigo publicado na *IEEE International Conference on Bioinformatics and Biomedicine* (BIBM'2018), que se encontra no Apêndice B.

Tabela 5.3: *Algoritmos com maior acurácia por visita.*

Visita	Acurácia (%)	Algoritmo
Relapse	99.1	RF
Last - 1	68.3	NB
Last - 2	73.6	MLP
Last - 3	73.8	RF
Last - 4	86.4	RF
Last - 5	77.2	RF
Last - 6	80.8	SVM
Baseline	66.1	RF

CAPÍTULO 6

Discussão

Como relatado no Capítulo 5, diversos experimentos foram realizados para obter regras válidas que pudessem ser interpretadas para se obter padrões relacionados a recaída para depressão em pacientes com TAB. No entanto, conforme mostrado na Tabela 5.1, o baixo desempenho obtido torna as regras pouco confiáveis e impede que elas sejam generalizadas para outros pacientes.

Nos 3 primeiros experimentos, foi possível verificar que manter a visita de recaída nos dados, enviesa os modelos. Ao se observar as regras geradas pelo Aleph, é possível ver que as outras visitas são praticamente ignoradas, como mostrado na regra exibida no Código 6.1. Esta é uma das regras que mais cobriu exemplos positivos. É possível interpreta-la da seguinte forma: o paciente *A* terá uma recaída se ele possui na ultima visita *B* os sintomas *depenerg* com o valor de definitiva falta de energia e *depse* com definitiva perda de interesse ou capacidade de aproveitar atividades prazerosas.

Código 6.1 clausulas experimento 3

```
1 [Rule 1] [Pos cover = 197 Neg cover = 0]
2 deprelapse(A) :-
3    has_relapse_visit(A, B),
4    depenerg(B, 'definite lack of energy'),
5    depse(B, 'definite lost of interest or capacity
6                  to enjoymost things').
```

Verifica-se então nestes modelos, que os pacientes já apresentam sintomas claros de depressão, facilmente identificado pelo psiquiatra, o que torna desnecessário o modelo gerado. Além disso, regras que indicam que um paciente terá recaída se na última visita ele apresenta determinados sintomas elevados de depressão ou mania não são úteis na prática clínica, pois deixa pouco espaço de tempo para o médico realizar um tratamento adequado.

Com o uso do BK do ILP, diferentes abordagens foram tentadas para se obter um resultado melhor. Por exemplo, com o agrupamento das visitas em intervalos, ou o

pareamento de pacientes positivos e negativos como forma de balanceamento dos dados. Porém, nenhuma dessas alternativas trouxe melhorias consistentes.

Por ser um algoritmo combinatorial, usar todos os 148 atributos do CMF seria inviável no ILP. Por isso, em um primeiro momento, trabalhou-se com atributos de sintomas depressão e mania. Em seguida, outros atributos foram usados, o que incluiu testes com medicamentos e doses, e testes com algoritmos de seleção de variáveis que, por sua vez, incluiu variáveis como CGI (*Clinical Global Impressions*) usado para avaliar o estado de funcionamento do paciente de acordo com a visão do médico antes e depois de iniciar uma medicação [70], e o GAF (*The Global Assessment of Functioning*) usado para medir o funcionamento do paciente em três áreas: psicológica, social e ocupacional [39].

Adicionalmente, no experimento 14 foram incluídos atributos do ADE, encontrados na revisão de literatura como forte preditores de recaída [63, 78]. Mais especificamente, foram incluídos atributos relacionados a histórico do TAB no paciente e presença de outras comorbidades, como diabetes, enxaqueca e alterações na tireoide. Esse estudo gerou uma leve melhora no desempenho (acurácia de 65%), o que sugere que dados do baseline podem ser efetivos para identificar uma recaída. Isso é comprovado pelo estudo de Salvini *et al*. (2015) [69], que usou dados socio-demográficos e clínicos do baseline e atingiu acurácia de 85% com ILP.

Já em relação aos atributos do CMF, pode-se dizer que informações relacionadas a recaída para depressão não foram bem capturadas, visto que nos experimentos longitudinais apenas é possível identificar uma recaída com alta taxa de acerto ao se usar a última visita. Por outro lado, ao longo de todas as outras visitas não foi possível identificar alterações significativas nesses atributos que permitissem caracterizar a recaída antes que ela acontecesse. Portanto, é possível afirmar que essas variáveis possuem pouco valor preditivo, sendo úteis apenas para constatar a depressão após já ter ocorrido.

Nos testes com algoritmos proposicionais é possível constatar dois pontos importantes. Primeiro, a necessidade de separar as visitas em conjuntos que permitisse a aplicação destes algoritmos (como *Baseline*, *Last*–1, *Last*–2 e outros), causa a perda de um componente importante para análise de recaída, o tempo.

Observa-se que o tempo até a recaída de dois pacientes diferentes na penúltima visita podem ser completamente diferentes, apesar de estarem agrupados dentro do mesmo conjunto de visitas. Por exemplo, suponha que o paciente A tenha visitas nas semanas 0, 2 e 4 e o paciente B nas semanas 0, 5 e 10. Isso significa que a ultima visita do paciente A ocorre antes da penúltima do paciente B, mas isso não é levado em consideração na escolha dos conjuntos, o que impossibilita determinar uma média de tempo até o paciente ter a recaída.

Em segundo, os resultados obtidos em alguns conjuntos de visitas não trouxeram

bons resultados. Por exemplo no *Baseline* a melhor acurácia obtida foi de 66.1% com RF. Porém, um classificador majoritário, que classificasse todas as instâncias como a classe com maior número de exemplos (sem considerar o uso do SMOTE), teria uma acurácia de 63.4%, ou seja, uma melhora pouco expressiva. A Tabela 6.1 apresenta essa comparação entre os classificadores com o classificador majoritário. Na ultima coluna vemos a diferença de acurácia entre ambos. Note que para a maioria, como o *Baseline*, *Last*–3, Last–5 e *Last*–6, a diferença é pequena. Isso mostra que, assim como o ILP, os algoritmos proposicionais também apresentaram baixo desempenho.

Tabela 6.1: *Comparação do desempenho entre os classificadores criados e um classificador majoritário por visita.*

Visita	Acurácia (%)	Acurácia Majoritário (%)	Δ
Relapse	99,1	63,9	35,2
Last - 1	68,3	50,7	17,6
Last - 2	73,6	64,6	9,0
Last - 3	73,8	71,4	2,4
Last - 4	86,4	75,3	11,1
Last - 5	77,2	76,3	0,9
Last - 6	80,8	78,8	2,0
Baseline	66,1	63,4	2,7

Os dados do STEP-BD consistem em dados clínicos, observados e coletados pelo psiquiatra durante consultas com o paciente, normalmente descritos em forma de escalas, como o CGI e o GAF por exemplo. Dessa forma, essas informações são subjetivas e dependem de interpretação médica, o que é um limitador para o estudo.

A partir disso e dos resultados obtidos neste trabalho, percebe-se a necessidade de identificar atributos que possuam um conteúdo informacional com maior relação a causa de episódios de depressão no TAB. Atualmente já é possível encontrar estudos neste sentido. No trabalho de Faurholt-Jepsen *et al*. (2016) [22] foram usados dados de voz coletados durante ligações, dados de auto-monitoramento em que o usuário responde algumas perguntas diariamente em um aplicativo de celular e dados gerados automaticamente relacionados ao comportamento do usuário, como quantidade de mensagens enviadas por dia. Para classificação foi utilizado o algoritmo *Random Forest* com o objetivo de classificar se o paciente se encontra no estado de depressão ou eutimia. Diferentes modelos foram gerados de acordo com os dados coletados. No modelo que utiliza apenas dados coletados sem interferência do usuário (dados de voz), a acurácia obtida foi de 68%. Já no modelo que inclui dados de auto-monitoramento pelo paciente a acurácia foi de 70%.

É importante notar que neste estudo o tempo até que o paciente entre no estado de depressão não é levado em consideração, o que é um limitador do estudo. Porém, o desempenho obtido sugere uma melhora com uso de diferentes tipos de dados, tanto coletados automaticamente ou com auxílio do usuário. Portanto, é razoável supor que o uso de outras fontes diversas de dados poderia acarretar em um desempenho maior também para predição de recaída para depressão em dados longitudinais.

CAPÍTULO 7

Conclusão

O TAB é uma doença que afeta parte significativa da população e devido ao seu caráter crônico pode causar prejuízos na vida do paciente, inclusive levar ao suicídio. Predizer recaídas para depressão o quanto antes é importante para realizar os tratamentos adequados antes que isto aconteça.

O objetivo deste projeto foi usar técnicas de aprendizado de máquina para encontrar padrões que permitam identificar quando um paciente com TAB terá recaída para depressão, em especial, explorar a ILP para trabalhar com os dados relacionais do STEP-BD.

Foi levantado, na revisão bibliográfica, a oportunidade de aplicação de técnicas de mineração de dados para análise de propriedades relacionadas a recaída para depressão, visto que até o presente momento não foram encontrados trabalhos que explorassem os dados do STEP-BD desta forma.

Os resultados foram aquém do esperado, em relação à acurácia, nos diversos experimentos realizados, e sugerem que somente os atributos de sintomas de depressão e mania, e medicação, contidos na tabela CMF do STEP-BD, não são bons preditores de recaída para depressão. No entanto, outras análises qualitativas dos resultados ainda precisam ser melhor exploradas.

Trabalhos futuros poderiam incluir diferentes tipos de dados, incluindo dados de ressonância magnéticas ou gerados automaticamente, como voz ou movimentação do paciente. Além disso, experimentos mais amplos com dados de outros formulários do STEP-BD poderiam melhorar o desempenho, conforme observado no experimento que relaciona os dados do ADE com o CMF.

Por fim, os experimentos permitiram melhor compreensão dos dados, principalmente para contribuir para o estado da arte do estudo de recaída para depressão no TAB, já que os estudos foram realizados sobre diferentes perspectivas, como o uso de sintomas e de medicamentos, e das diferentes formas de manipulação das visitas como conhecimento de fundo do ILP.

Referências Bibliográficas

[1] BERNSTEIN, E. E.; RABIDEAU, D. J.; GIGLER, M. E.; NIERENBERG, A. A.; DECKERSBACH, T.; SYLVIA, L. G. **Patient perceptions of physical health and bipolar symptoms: The intersection of mental and physical health**. *Journal of affective disorders*, 189:203–206, 2016.

[2] BLOCKEEL, H.; DE RAEDT, L. **Top-down induction of first-order logical decision trees**. *Artificial intelligence*, 101(1-2):285–297, 1998.

[3] BRATKO, I.; MUGGLETON, S. **Applications of inductive logic programming**. *Communications of the ACM*, 38(11):65–70, 1995.

[4] BREIMAN, L. **Random forests**. *Machine learning*, 45(1):5–32, 2001.

[5] BREIMAN, L. **Classification and regression trees**. Routledge, 2017.

[6] BUSCH, A. B.; NEELON, B.; ZELEVINSKY, K.; HE, Y.; NORMAND, S.-L. T. **Accurately predicting bipolar disorder mood outcomes: implications for the use of electronic databases**. *Medical care*, 50(4):311–319, 2012.

[7] CHANDRASHEKAR, G.; SAHIN, F. **A survey on feature selection methods**. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[8] CHAWLA, N. V. **Data mining for imbalanced datasets: An overview**. In: *Data mining and knowledge discovery handbook*, p. 875–886. Springer, 2009.

[9] CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. **Smote: synthetic minority over-sampling technique**. *Journal of artificial intelligence research*, 16:321–357, 2002.

[10] CHENG, J.; GREINER, R. **Comparing bayesian network classifiers**. In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, p. 101–108. Morgan Kaufmann Publishers Inc., 1999.

[11] CHRISTIANINI, N.; SHAWE-TAYLOR, J. **Support vector machines and other kernel-based learning methods**, 2000.

[12] CORTES, C.; VAPNIK, V. **Support-vector networks**. *Machine learning*, 20(3):273–297, 1995.

[13] CRETU, J. B.; CULVER, J. L.; GOFFIN, K. C.; SHAH, S.; KETTER, T. A. **Sleep, residual mood symptoms, and time to relapse in recovered patients with bipolar disorder**. *Journal of affective disorders*, 190:162–166, 2016.

[14] DAS, S. **Filters, wrappers and a boosting-based hybrid for feature selection**. In: *Icml*, volume 1, p. 74–81, 2001.

[15] DECKERSBACH, T.; PETERS, A. T.; SYLVIA, L. G.; GOLD, A. K.; DA SILVA MAGALHAES, P. V.; HENRY, D. B.; FRANK, E.; OTTO, M. W.; BERK, M.; DOUGHERTY, D. D.; OTHERS. **A cluster analytic approach to identifying predictors and moderators of psychosocial treatment for bipolar depression: Results from step-bd**. *Journal of Affective Disorders*, 2016.

[16] DIAS, R. S.; LAFER, B.; RUSSO, C.; DEL DEBBIO, A.; NIERENBERG, A. A.; SACHS, G. S.; JOFFE, H. **Longitudinal follow-up of bipolar disorder in women with premenstrual exacerbation: findings from step-bd**. *American Journal of Psychiatry*, 2011.

[17] DOLSAK, B.; MUGGLETON, S. **The application of inductive logic programming to finite element mesh design**. In: *Inductive logic programming*. Citeseer, 1992.

[18] DRAGO, A.; MONTI, B.; DE RONCHI, D.; SERRETTI, A. **Cry1 variations impacts on the depressive relapse rate in a sample of bipolar patients**. *Psychiatry investigation*, 12(1):118–124, 2015.

[19] DŽEROSKI, S. **Multi-relational data mining: an introduction**. *ACM SIGKDD Explorations Newsletter*, 5(1):1–16, 2003.

[20] EL-MALLAKH, R. S.; VÖHRINGER, P. A.; OSTACHER, M. M.; BALDASSANO, C. F.; HOLTZMAN, N. S.; WHITHAM, E. A.; THOMMI, S. B.; GOODWIN, F. K.; GHAEMI, S. N. **Antidepressants worsen rapid-cycling course in bipolar depression: a step-bd randomized clinical trial**. *Journal of affective disorders*, 184:318–321, 2015.

[21] FABBRI, C.; SERRETTI, A. **Genetics of long-term treatment outcome in bipolar disorder**. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 65:17–24, 2016.

[22] FAURHOLT-JEPSEN, M.; BUSK, J.; FROST, M.; VINBERG, M.; CHRISTENSEN, E.; WINTHER, O.; BARDRAM, J. E.; KESSING, L. **Voice analysis as an objective state marker in bipolar disorder**. *Translational psychiatry*, 6(7):e856, 2016.

[23] FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From data mining to knowledge discovery in databases**. *AI magazine*, 17(3):37, 1996.

[24] FITTING, M. **First-order logic**. In: *First-order logic and automated theorem proving*, p. 97–125. Springer, 1990.

[25] FOR MENTAL HEALTH, N. C. C. **Bipolar disorder: The management of bipolar disorder in adults, children and adolescents, in primary and secondary care**. British Psychological Society, 2006.

[26] FRIEDMAN, N.; GOLDSZMIDT, M. **Building classifiers using bayesian networks**. In: *Proceedings of the national conference on artificial intelligence*, p. 1277–1284, 1996.

[27] GEDDES, J. R.; MIKLOWITZ, D. J. **Treatment of bipolar disorder**. *The Lancet*, 381(9878):1672–1682, 2013.

[28] GENTILI, C.; VALENZA, G.; NARDELLI, M.; LANATÀ, A.; BERTSCHY, G.; WEINER, L.; MAURI, M.; SCILINGO, E. P.; PIETRINI, P. **Longitudinal monitoring of heartbeat dynamics predicts mood changes in bipolar patients: a pilot study**. *Journal of affective disorders*, 209:30–38, 2017.

[29] GOODWIN, F. K.; JAMISON, K. R. **Manic-depressive illness: bipolar disorders and recurrent depression**, volume 1. Oxford University Press, 2007.

[30] GRUBER, J.; MIKLOWITZ, D. J.; HARVEY, A. G.; FRANK, E.; KUPFER, D.; THASE, M. E.; SACHS, G. S.; KETTER, T. A. **Sleep matters: sleep functioning and course of illness in bipolar disorder**. *Journal of affective disorders*, 134(1):416–420, 2011.

[31] HARTIGAN, J. **Clustering algorithms**. *John Willey & Sons ()*, 1975.

[32] HAYKIN, S. S.; HAYKIN, S. S.; HAYKIN, S. S.; HAYKIN, S. S. **Neural networks and learning machines**, volume 3. Pearson Upper Saddle River, NJ, USA:, 2009.

[33] HOCHMAN, E.; WEIZMAN, A.; VALEVSKI, A.; FISCHEL, T.; KRIVOY, A. **Association between bipolar episodes and fluid and electrolyte homeostasis: a retrospective longitudinal study**. *Bipolar disorders*, 16(8):781–789, 2014.

[34] HOLLANDER, M.; WOLFE, D. A. **Nonparametric statistical methods**. 1999.

[35] HOLTE, R. C. **Very simple classification rules perform well on most commonly used datasets**. *Machine learning*, 11(1):63–90, 1993.

[36] HONG, W.; ZHANG, C.; XING, M. J.; WU, Z. G.; WANG, Z. W.; CHEN, J.; YUAN, C. M.; SU, Y. S.; HU, Y. Y.; CAO, L.; OTHERS. **Contribution of long duration of undiagnosed bipolar disorder to high frequency of relapse: A naturalistic study in china**. *Comprehensive Psychiatry*, 70:77–81, 2016.

[37] ICHISE, R.; NUMAO, M. **First-order rule mining by using graphs created from temporal medical data**. In: *Active Mining*, p. 112–125. Springer, 2005.

[38] JAFFEE, W. B.; GRIFFIN, M. L.; GALLOP, R.; MEADE, C. S.; GRAFF, F.; BENDER, R. E.; WEISS, R. D. **Does alcohol use precipitate depression among patients with co-occurring bipolar and substance use disorders?** *The Journal of clinical psychiatry*, 70(2):171, 2009.

[39] JONES, S. H.; THORNICROFT, G.; COFFEY, M.; DUNN, G. **A brief mental health outcome scale-reliability and validity of the global assessment of functioning (gaf).** *The British Journal of Psychiatry*, 166(5):654–659, 1995.

[40] JUDD, L. L.; AKISKAL, H. S.; SCHETTLER, P. J.; ENDICOTT, J.; LEON, A. C.; SOLOMON, D. A.; CORYELL, W.; MASER, J. D.; KELLER, M. B. **Psychosocial disability in the course of bipolar i and ii disorders: a prospective, comparative, longitudinal study**. *Archives of general psychiatry*, 62(12):1322–1330, 2005.

[41] KARALIČ, A.; BRATKO, I. **First order regression**. *Machine Learning*, 26(2-3):147–176, 1997.

[42] KERSTING, K.; DE RAEDT, L. **Towards combining inductive logic programming with bayesian networks**. In: *International Conference on Inductive Logic Programming*, p. 118–131. Springer, 2001.

[43] KITCHENHAM, B. **Procedures for performing systematic reviews**. *Keele, UK, Keele University*, 33(2004):1–26, 2004.

[44] KNOBBE, A.; BLOCKEEL, H.; SIEBES, A.; VAN DER WALLEN, D. **Multi-relational data mining**. 1999.

[45] KOHAVI, R.; OTHERS. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. In: *Ijcai*, volume 14, p. 1137–1145. Stanford, CA, 1995.

[46] LIAW, A.; WIENER, M.; OTHERS. **Classification and regression by randomforest**. *R news*, 2(3):18–22, 2002.

[47] LIBRENZA-GARCIA, D.; KOTZIAN, B. J.; YANG, J.; MWANGI, B.; CAO, B.; LIMA, L. N. P.; BERMUDEZ, M. B.; BOEIRA, M. V.; KAPCZINSKI, F.; PASSOS, I. C. **The impact of machine learning techniques in the study of bipolar disorder: a systematic review**. *Neuroscience & Biobehavioral Reviews*, 80:538–554, 2017.

[48] LIU, H.; SETIONO, R. **Chi2: Feature selection and discretization of numeric attributes**. In: *Tools with artificial intelligence, 1995. proceedings., seventh international conference on*, p. 388–391. IEEE, 1995.

[49] MAGALHÃES, P. V.; DODD, S.; NIERENBERG, A. A.; BERK, M. **Cumulative morbidity and prognostic staging of illness in the systematic treatment enhancement program for bipolar disorder (step-bd)**. *Australian and New Zealand journal of psychiatry*, p. 0004867412460593, 2012.

[50] MARSH, W. K.; KETTER, T. A.; CRAWFORD, S. L.; JOHNSON, J. V.; KROLL-DESROSIERS, A. R.; ROTHSCHILD, A. J. **Progression of female reproductive stages associated with bipolar illness exacerbation**. *Bipolar disorders*, 14(5):515–526, 2012.

[51] MARSH, W. K.; TEMPLETON, A.; KETTER, T. A.; RASGON, N. L. **Increased frequency of depressive episodes during the menopausal transition in women with bipolar disorder: preliminary report**. *Journal of psychiatric research*, 42(3):247–251, 2008.

[52] MATHERS, C.; FAT, D. M.; BOERMA, J. T. **The global burden of disease: 2004 update**. World Health Organization, 2008.

[53] MITCHELL, T. M.; OTHERS. **Machine learning. 1997**. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877, 1997.

[54] MUGGLETON, S. **Bayesian inductive logic programming**. In: *Proceedings of the Seventh Annual Conference on Computational Learning Theory*, COLT '94, p. 3–11, New York, NY, USA, 1994. ACM.

[55] MUGGLETON, S. **Inverse entailment and progol**. *New generation computing*, 13(3):245–286, 1995.

[56] MUGGLETON, S.; DE RAEDT, L. **Inductive logic programming: Theory and methods**. *The Journal of Logic Programming*, 19:629–679, 1994.

[57] NOVIS, F.; CIRILLO, P.; SILVA, R. A. D.; SANTOS, A. L.; SILVEIRA, L. A. S.; CARDOSO, A.; COSCARELLI, P.; NARDI, A. E.; CHENIAUX, E. **The progression**

of 102 brazilian patients with bipolar disorder: outcome of first 12 months of prospective follow-up. *Trends in psychiatry and psychotherapy*, 36(1):16–22, 2014.

[58] OSTACHER, M. J.; PERLIS, R. H.; NIERENBERG, A. A.; CALABRESE, J.; STANGE, J. P.; SALLOUM, I.; WEISS, R. D.; SACHS, G. S. **Impact of substance use disorders on recovery from episodes of depression in bipolar disorder patients: prospective data from the systematic treatment enhancement program for bipolar disorder (step-bd)**. *American Journal of Psychiatry*, 167(3):289–297, 2009.

[59] OTTO, M.; SIMON, N.; WISNIEWSKI, S.; MIKLOWITZ, D.; KOGAN, J.; REILLY-HARRINGTON, N.; FRANK, E.; NIERENBERG, A.; MARANGELL, L.; SAGDUYU, K.; OTHERS. **Prospective 12-month course of bipolar disorder in out-patients with and without comorbid anxiety disorders**. *The British Journal of Psychiatry*, 189(1):20–25, 2006.

[60] PÉREZ, N. P. **Improving variable selection and mammography-based machine learning classifiers for breast cancer cadx**. 2015.

[61] PERLIS, R. H.; DENNEHY, E. B.; MIKLOWITZ, D. J.; DELBELLO, M. P.; OSTACHER, M.; CALABRESE, J. R.; AMETRANO, R. M.; WISNIEWSKI, S. R.; BOWDEN, C. L.; THASE, M. E.; OTHERS. **Retrospective age at onset of bipolar disorder and outcome during two-year follow-up: results from the step-bd study**. *Bipolar disorders*, 11(4):391–400, 2009.

[62] PERLIS, R. H.; OSTACHER, M. J.; PATEL, J. K.; MARANGELL, L. B.; ZHANG, H.; WISNIEWSKI, S. R.; KETTER, T. A.; MIKLOWITZ, D. J.; OTTO, M. W.; GYULAI, L.; OTHERS. **Predictors of recurrence in bipolar disorder: primary outcomes from the systematic treatment enhancement program for bipolar disorder (step-bd)**. *American Journal of Psychiatry*, 163(2):217–224, 2006.

[63] PETERS, A.; SYLVIA, L.; DA SILVA MAGALHAES, P.; MIKLOWITZ, D.; FRANK, E.; OTTO, M.; HANSEN, N.; DOUGHERTY, D.; BERK, M.; NIERENBERG, A.; OTHERS. **Age at onset, course of illness and response to psychotherapy in bipolar disorder: results from the systematic treatment enhancement program for bipolar disorder (step-bd)**. *Psychological medicine*, 44(16):3455–3467, 2014.

[64] PIATETSKY-SHAPIRO, G.; BRACHMAN, R. J.; KHABAZA, T.; KLOESGEN, W.; SIMOUDIS, E. **An overview of issues in developing industrial data mining and knowledge discovery applications.** In: *KDD*, volume 96, p. 89–95, 1996.

[65] PRESS, W. H.; TEUKOLSKY, S. A.; VETTERLING, W. T.; FLANNERY, B. P. **Numerical recipes in c**. *Cambridge University Press*, 1:3, 1988.

[66] QUINLAN, J. R. **Learning logical definitions from relations**. *Machine learning*, 5(3):239–266, 1990.

[67] ROYSTON, J. **An extension of shapiro and wilk's w test for normality to large samples**. *Applied statistics*, p. 115–124, 1982.

[68] SACHS, G. S.; THASE, M. E.; OTTO, M. W.; BAUER, M.; MIKLOWITZ, D.; WISNIEWSKI, S. R.; LAVORI, P.; LEBOWITZ, B.; RUDORFER, M.; FRANK, E.; OTHERS. **Rationale, design, and methods of the systematic treatment enhancement program for bipolar disorder (step-bd)**. *Biological psychiatry*, 53(11):1028–1042, 2003.

[69] SALVINI, R.; DA SILVA, D. R.; LAFER, B.; DUTRA, I. **A multi-relational model for depression relapse in patients with bipolar disorder.** *Studies in health technology and informatics*, 216:741–745, 2015.

[70] SPEARING, M. K.; POST, R. M.; LEVERICH, G. S.; BRANDT, D.; NOLEN, W. **Modification of the clinical global impressions (cgi) scale for use in bipolar illness (bp): the cgi-bp**. *Psychiatry research*, 73(3):159–171, 1997.

[71] SRINIVASAN, A. **The aleph manual, 2001**. *URL http://www.comlab.ox.ac.uk/activities/machinelearning/Aleph*, 2017.

[72] SRINIVASAN, A.; MUGGLETON, S.; KING, R. D.; STERNBERG, M. J. **Mutagenesis: Ilp experiments in a non-determinate biological domain**. In: *Proceedings of the 4th international workshop on inductive logic programming*, volume 237, p. 217–232. Citeseer, 1994.

[73] STANGE, J. P.; SYLVIA, L. G.; DA SILVA MAGALHÃES, P. V.; MIKLOWITZ, D. J.; OTTO, M. W.; FRANK, E.; BERK, M.; NIERENBERG, A. A.; DECKERSBACH, T. **Extreme attributions predict the course of bipolar depression: results from the step-bd randomized controlled trial of psychosocial treatment**. *The Journal of clinical psychiatry*, 74(3):249–255, 2013.

[74] STANGE, J. P.; SYLVIA, L. G.; DA SILVA MAGALHÃES, P. V.; MIKLOWITZ, D. J.; OTTO, M. W.; FRANK, E.; YIM, C.; BERK, M.; DOUGHERTY, D. D.; NIERENBERG, A. A.; OTHERS. **Affective instability and the course of bipolar depression: results from the step-bd randomised controlled trial of psychosocial treatment**. *The British Journal of Psychiatry*, 208(4):352–358, 2016.

[75] TURCOTTE, M.; MUGGLETON, S.; STERNBERG, M. **Application of inductive logic programming to discover rules governing the three-dimensional topology of protein structure**. *Inductive Logic Programming*, p. 53–64, 1998.

[76] VALENZA, G.; GENTILI, C.; LANATÀ, A.; SCILINGO, E. P. **Mood recognition in bipolar patients through the psyche platform: preliminary evaluations and perspectives**. *Artificial intelligence in medicine*, 57(1):49–58, 2013.

[77] VALENZA, G.; NARDELLI, M.; LANATA, A.; GENTILI, C.; BERTSCHY, G.; PARADISO, R.; SCILINGO, E. P. **Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis**. *IEEE Journal of Biomedical and Health Informatics*, 18(5):1625–1635, 2014.

[78] WAXMONSKY, J. A.; THOMAS, M. R.; MIKLOWITZ, D. J.; ALLEN, M. H.; WISNIEWSKI, S. R.; ZHANG, H.; OSTACHER, M. J.; FOSSEY, M. D. **Prevalence and correlates of tobacco use in bipolar disorder: data from the first 2000 participants in the systematic treatment enhancement program**. *General hospital psychiatry*, 27(5):321–328, 2005.

[79] ZHANG, H.; WISNIEWSKI, S. R.; BAUER, M. S.; SACHS, G. S.; THASE, M. E.; INVESTIGATORS, S.-B.; OTHERS. **Comparisons of perceived quality of life across clinical states in bipolar disorder: data from the first 2000 systematic treatment enhancement program for bipolar disorder (step-bd) participants**. *Comprehensive psychiatry*, 47(3):161–168, 2006.

APÊNDICE A

Estatística Descritiva

Tabela A.1: *Estatística descritiva dos atributos usados em relação a quantidade de visitas.*

	n	**n (%)**
deprmd (*rate of depression*)		
0: *not depressed*	1711	62,99
0,25: *questionable, slight or rare dysphoria*	177	6,51
0,5: *clearly but subthreshold dysphoria*	383	14,10
1: *define depressed/dysphoric most of the day*	376	13,84
1,5: *depressed/dysphoric most of the day*	55	2,02
2: *constant/unremitting intense dysphoria*	14	0,51
depslmax (*maximum sleep time*)		
$<$ 5 horas	42	1,36
$[5,7)$ horas	120	6,81
$[7,9)$ horas	1171	37,98
$[9,11)$ horas	1069	34,67
$[11,13)$ horas	394	12,77
$>$ 12 horas	197	6,38
depse (*depression rate of self-esteem*)		
-2: *emotionally constricted*	7	0,20
-1,5: *lost of interest or capacity to enjoy most things*	32	0,93
-1: *definite lost of interest or capacity to enjoy most things*	475	13,86
-0,5: *clearly but subthreshold decreased interest*	475	13,86
-0,25 *not clinically significant decreased interest*	142	4,14
0: *enjoy activity as usual*	2295	66,98
depdist (*depression rate of distractibility*)		
0: *no evidence of distractibility*	2507	73,28

Continua na próxima página

Tabela A.1 – continuação da página anterior

	n	n (%)
0,25: *not clinically occasional subjective or objective distraction*	154	4,50
0,5: *clearly but occasional subjective or objective distraction*	446	13,03
1: *definite decreased ability to complete tasks due to distract*	290	8,47
1,5: *definite great effort to complete tasks due to distractibility*	17	0,49
2: *persistence evidence of distractibility present also at int*	7	0,20
deppma (*depression rate of physical motor activity*)		
0: *no evidence of restlessness*	2926	85,45
0,25: *not clinically significant restlessness*	103	3,00
0,5: *clearly but subthreshold restlessness*	282	8,23
1: *definite restlessness*	104	3,03
1,5: *difficulty remaining still*	7	0,20
2: *unable to sit still*	2	0,05
sipassiv (*passive suicide ideation*)		
No	1232	88,37
Yes	162	11,62
elvsleep (*elevation of sleep time*)		
-2: *definite > 5 hour decrease*	20	0,58
-0,5: *clearly but subthreshold decrease need of sleep*	340	9,90
-0,25 *not clinicaly significant decrease need for sleep*	49	1,42
0: *sleeping normaly*	3025	88,08
elvdistr (*elevation of distractibility*)		
0: *none*	2619	76,22
0,25: *not clinically significant subjective or objective distraction*	134	3,89
0,5: *clearly but subthreshold subjective or objective distraction*	405	11,78
1: *definite distractibility, compromising function*	257	7,47
1,5: *persistent distractibility compromising task*	16	0,46
2: *cant stay on a topic, severe distractibility*	5	0,14
elvhrb (*elevation of high risk behavior*)		
0: *no risk taking*	3289	95,88
0,25: *not clinicaly significant risk taking*	29	0,84
0,5: *clearly but subthreshold risk taking*	87	2,53
1: *definite risky behavior*	24	0,69
1,5: *definite risky behavior outside the normal tolerance*	1	0,02

Continua na próxima página

Tabela A.1 – continuação da página anterior

	n	n (%)
depsleep (*depression rate of sleep*)		
-2: *> 50% sleep decrease*	22	0,64
-1,5: *definite > 25% decrease*	28	0,81
-1: *definite > 1 hour decrease*	267	7,77
-0,5: *clearly but subthreshold sleep decreasing*	368	10,71
-0,25: *not clinically significant sleep decreasing*	107	3,11
0: *sleeping normally*	2093	60,96
0,25: *not clinically significant sleep increasing*	48	1,39
0,5: *any sleep increasing*	200	5,82
1: *definite > 1 hour increase*	249	7,25
1,5: *> 25% sleep increase*	34	0,99
2: *< 50% sleep increase*	17	0,49
depinter (*depression rate of interest*)		
-2: *emotionally constricted*	16	0,46
-1,5: *lost of interest or capacity to enjoy most things*	51	1,48
-1: *definite lost of interest or capacity to enjoy most things*	452	13,16
-0,5: *clearly but subthreshold decreased interest*	438	12,75
-0,25: *not clinically significant decreased interest*	134	3,90
0: *enjoy activity as usual*	2342	68,22
depenerg (*depression rate of energy*)		
-2: *stays in bed, lethargic*	16	0,46
-1,5: *too tired even for pleasant task*	43	1,25
-1: *definite lack of energy*	540	15,73
-0,5: *clearly but subthreshold feeling of tiredness*	599	17,45
-0,25: *not clinically significant tired*	195	5,68
0: *usual energy level*	2038	59,39
depappet (*depression rate of appetite*)		
-2: *< 5 pound, < 50% food intake*	8	0,23
-1,5: *definite decreased consumption*	9	0,26
-1: *definite decreased consumption*	191	5,56
-0,5: *clearly but subthreshold decrease in appetite*	188	5,47
-0,25: *questionable decrease in appetite*	55	1,60
0: *appetite normal*	2427	70,65

Continua na próxima página

Tabela A.1 – continuação da página anterior

	n	n (%)
0,25: *not clinically significant increase in appetite*	77	2,24
0,5: *clearly but subthreshold increase in appetite*	276	8,03
1: *definite increased consumption*	194	5,64
1,5: *definite food craving or seeking in addition to usual meals*	10	0,29
depsi (*depression rate of suicide ideation*)		
0: *no morbid preoccupation*	3057	89,12
0,25: *not clinically significant LNWL, passive SI*	106	3,09
0,5: *clearly but subthreshold LNWL, passive SI, active SI*	182	5,30
1: *definite LNWL, passive SI, active SI*	81	2,36
1,5: *active SI, persistent SI*	3	0,08
2: *active SI, persistent SI, huge to harm self*	1	0,02
siactive (*active suicide ideation*)		
No	1373	98,49
Yes	21	1,50
elvtalk (*elevation of talk*)		
0: *normal rate and quantity of speech*	3036	88,35
0,25: *not clinically significant talkativeness*	92	2,67
0,5: *clearly but subthreshold talkativeness*	238	6,92
1: *definite talkativeness*	68	1,97
1,5: *persistent pressure speech*	1	0,02
2: *incessant talking*	1	0,02
elvgdact (*elevation of group directed activity*)		
0: *none*	3063	89,43
0,25: *not clinically significant starting new projects*	52	1,51
0,5: *clearly but subthreshold new project started*	203	5,92
1: *definite engagement in new projects*	98	2,86
1,5: *multiple creative new projects take > 8 hours/wk*	8	0,23
2: *total work effort dedicated to new projects*	1	0,02
depslmin (*minumum sleep time*)		
< 5 horas	404	13,05
$[5,7)$ horas	992	32,06
$[7,9)$ horas	1373	44,37
$[9,11)$ horas	264	8,53

Continua na próxima página

Tabela A.1 – continuação da página anterior

	n	n (%)
$[11, 13)$ horas	56	1,80
> 12 horas	5	0,16
depguilt (*depression rate of feeling guilty*		
0: *no excessive self blame or guilty preoccupation*	2761	80,66
0,25: *not clinically significant self depreciating thoughts*	96	2,80
0,5: *clearly but subthreshold self depreciating thoughts*	307	8,96
1: *definite self depreciating thoughts*	246	7,18
1,5: *definite self depreciating thoughts*	12	0,35
2: *persistent ideas of guilty rumination*	1	0,02
depconcn (*depression rate of concentration*)		
-2: *clear cognitive impairment at the interview*	12	0,34
-1,5: *unable to function in role*	26	0,75
-1: *definite lack of concentration*	414	12,06
-0,5: *clearly but subthreshold feeling of tiredness*	483	14,08
-0,25: *not clinically significant concentration trouble*	163	4,75
0: *usual concentration level*	2332	67,98
deppmr (*depression rate of physical motor retardation*)		
0: *no evidence of slowing*	2842	82,88
0,25: *not clinically significant slowing*	67	1,95
0,5: *clearly but subthreshold slowing*	275	8,01
1: *definite slowness*	233	6,79
1,5: *slowness observed by others*	9	0,26
2: *impedeting slowness*	3	0,08
silnwl (*suicide ideation – life not worth living*)		
No	1122	80,37
Yes	274	19,62
elvselfe (*elevation in self-esteem*)		
0: *no excessive self esteem/confidence*	2606	75,86
0,25: *not clinically significant exaggerated sense of abilities*	109	3,17
0,5: *clearly but subthreshold exaggerated sense of abilities*	375	10,91
1: *definite inflated sense of abilities*	313	9,11
1,5: *excessive inflated sense of abilities*	28	0,81
2: *grossly excessive ideas sense of abilities*	4	0,11

Continua na próxima página

Tabela A.1 – continuação da página anterior

	n	**n (%)**
elvfoi (*elevation in flight of ideas*)		
0: *none*	3007	87,54
0,25: *not clinically significant degree of thinking fast*	123	3,58
0,5: *clearly but subthreshold degree of thinking fast*	223	6,49
1: *definite rapid train of thoughts*	77	2,24
1,5: *persistent rapid train of thoughts*	4	0,11
2: *speech cannot keep up thoughts*	1	0,02
elvpma (*elevation in physical motor activity*)		
0: *no evidence of restlessness*	2944	86,15
0,25: *not clinically significant restlessness*	96	2,80
0,5: *clearly but subthreshold restlessness*	277	8,10
1: *definite restlessness*	93	2,72
1,5: *difficulty remaining still*	6	0,17
2: *unable to sit still*	1	0,02

APÊNDICE B

Produção Científica

Artigo publicado na *IEEE International Conference on Bioinformatics and Biomedicine* (BIBM'2018).

Forecasting depressive relapse in Bipolar Disorder from clinical data

Renato Borges-Júnior*, Rogerio Salvini*, Andrew A. Nierenberg‡,
Gary S. Sachs‡, Beny Lafer§ and Rodrigo S. Dias§
*Instituto de Informatica, Universidade Federal de Goias, Goiania, GO, Brazil
email: {renatoborges, rogeriosalvini}@inf.ufg.br
‡Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
§Bipolar Disorder Research Program, Department of Psychiatry, University of Sao Paulo Medical School, Sao Paulo, SP, Brazil

***Abstract*—Bipolar disorder (BD) is a mood disorder characterized by recurrent episodes of depression and mania/hypomania. Depressive relapse in BD reach rates close to 50% in 1 year and 70% in up to 4 years of treatment. Several studies have been developed to discover more efficient treatments for BD and prevent relapses. However, most of relapse studies used only statistical methods. We aim to analyze the performance of machine learning algorithms in predicting depressive relapse using only clinical data from patients. Five well-used machine learning algorithms (Support Vector Machines, Random Forests, Naïve Bayes and Multilayer Perceptron) were applied to the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD) dataset of a cohort of 800 patients who became euthymic during the study and were followed up for 1 year: 507 presented a depressive relapse and 293 did not. The algorithms showed reasonable performance in the prediction task, ranging from 61% to 80% in the F-measure. Random Forest algorithm had a higher average of performance (Relapse Group 68%; No Relapse Group 74%), although, the performance between classifiers showed no significant difference. Random Forest analysis demonstrated that the three most important mood symptoms observed were: interest, depression mood and energy. Results show that the machine learning algorithms could be seen as a sensible approach to better support medical decision-making in the BD treatment and prevention of future relapses.**

***Index Terms*—bipolar disorder, mental health, depressive relapse, machine learning, artificial intelligence**

I. Introduction

Bipolar Disorder (BD) is a chronic mental disorder characterized by the presence of mood episodes of depression, mania (elation) or hypomania, and mixed states [1]. Manic states are identified by feelings of high energy, elated mood, and increased activity with rapid thoughts about different things, sleeping problems, irritability and risk taking behavior. Depressive state is identified by feelings of sadness, hopelessness, loss of interest for activities, reduced energy, worthlessness, excessive self-blame and by symptoms like fatigue, loss or gain of weight, sleeping problems, trouble concentrating and suicidal thoughts or actions.

BD affects between 43.5–54.4 million around the globe, with an overall prevalence of 0.7% (0.6%–0.8%), with higher prevalence of 1.2 (1.0-1.4) between 25-29 years old with male to female prevalence of 0.8 (0.5–1.1) [2]. BD is the 54th leading cause of global and 5th among mental and substance use disorders considering the DAILYs (disability adjusted life years) and 16th place as a cause of YLDs (years lived with disability) metrics [2]. During life, between 25%-50% BD patients attempt suicide [1]. Under treatment, in a systematic review BD recurrence was observed in long-term prospective naturalistic studies rates of 55.2% (26.3% per year), 39.3% (21.9% per year) in randomized clinical trials with mood-stabilizers and 60.3% (31.3% per year) with placebo. Depression was the most common mood episode (52.0%) [3]. Thus, predicting relapse as soon as possible is crucial to improve treatments and avoid great losses.

The use of electronic health records and the mining of clinical data provide opportunities to create analytic models to assist in clinical decision-making [4]. Therefore, the mining of clinical data is an important task to recognize patterns and extract useful knowledge of depressive relapse that may potentially lead to new and more effective treatment.

The work of Librenza-Garcia *et al*. [5] thoroughly reviewed the use of Machine Learning (ML) techniques in the study of BD. Although their focus was on studies that assessed diagnosis, they also included studies related to treatment, prognosis and development of data-driven phenotypes. Among these, the use of ML to predict depression relapse was found only in one study.

In addition, we systematic reviewed studies that assessed the use of statistical or ML to predict depressive relapses in BD patients based on clinical data. This review was performed according to the process suggested by Kitchenham [6]. Among the 15 selected studies, linear/logistic regression was the most commonly used, present in 10 (66.7%) articles. Survival analysis was used in 6 (40.0%) works and only 2 studies (13.3%) used ML techniques (some studies used more than one method).

The unsupervised learning algorithm k-means along with hierarchical clustering was used by Deckersbach *et al*. [7] to create clusters (patient groups) based on relapse. Each group was related to the recovery capability from relapse according to two distinct treatments. The Inductive Logic Programming (ILP), a supervised method, was used to discover patterns of depressive relapse by Salvini *et al*. [8] that analyzed 211 socio-demographic characteristics and clinical information of 108 BD patients, reaching up to 85% of accuracy to predict patient relapse.

This work aims to analyze the performances of ML algorithms to predict depressive relapses in the STEP-BD dataset, which can be considered one of the most large and complete dataset with clinical data of BD patients [9]. We hypothesized that they could achieve good performance in the depressive relapse prediction task. As far as we know, this is the first application of ML algorithms to predict depressive relapse in the STEP-BD dataset.

The rest of this paper is organized as follows: in Section II we present the methodology to achieve the goals proposed; in Section III, we discuss the results of this study; and, in Sections IV and V, a final discussion about the study and a conclusion are presented.

II. Materials and Methods

A. Data Set

Data were obtained from the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD), an NIMH large-scale public health initiative to address the impact of treatments on the course of BD [9]. Twenty sites across the U.S.A. were involved, from 1998 to 2007 and evaluated 4,360 BD patients. Data were collected from those patients at the baseline evaluation and during the follow-up period [9]. The Institutional Review Board of each participating STEP-BD site approved the study procedures. Informed consent was obtained from all participants.

The dataset consists of several files, with different evaluations and characteristics of the patients. The CMF (Clinical Monitoring Form) contains longitudinal data including the evaluation of the current mood state, identified by the feature "clinstat". This feature can take on a set of values such as depression, mania, hypomania, mixed-cycling, recovered, recovering, roughening and continuous symptomatic, which indicate the mood state of the patient in the current visit. Only visits in recovered or depression state of each patient were selected. Both states were defined by DSM-IV [10]. In CMF, the recovered clinical state was also defined if, for eight or more weeks, the participant had two or less mood symptoms. The depression state was defined as a presence of five or more depressive symptoms for at least 10 days [9].

B. Participants' Sample Selection

The presence of residual depressive and manic symptoms at recovery was associated with earlier time to depressive and manic relapse. Our patient selection was based on other STEP-BD study conducted by Perlis *et al.* [11], who used a mean follow-up after recovery of 56.2 weeks. Patients' selection was made according to the following algorithm: for each patient in the data set, the algorithm searches the visits until it finds the first visit whose status is recovered. Then, the status of the following visits are verified for the next 54 weeks. If a visit with the status of depression is found, the patient is labeled in the "relapse" class (and considered as a positive example). Otherwise, if the visit has the clinical status of recovered, then the previous step is repeated. After 54 weeks, if the patient kept the recovered status along every visit, then he or she is labeled as "no relapse" (and considered a negative example). Finally, the algorithm moves to the next patient and this procedure is repeated. As a result, this algorithm selected 800 participants, 507 who presented a depressive relapse and 293 who did not.

Given that patients' visits are not recorded as successive equally spaced points in time, applying time series analysis techniques could not be possible. Therefore, the longitudinal visits were used as follows. First, the ML algorithms were applied only to data from the initial visit of recovered (called "Baseline" visit). Next, they were applied to data from the last visit (denoted by "Relapse" visit, which states that the patient had a relapse to depression in this visit), then they were applied to data from second to last visit (denoted by "Last - 1", which can be read as "one visit before the relapse visit"), then to data from third-to-last visit (denoted by "Last - 2"), and so on. We repeat this process up to "Last - 6" (six visits before the relapse visit). It was expected that the earlier the relapse could be predicted the better.

Table I shows the number of observations for each visit. In the cases that the minority class had a large fewer number of observations, the SMOTE [12] algorithm was used to double these samples by creating synthetic ones.

Table I
Number of examples in each visit.

Visit	**n# of observations after SMOTE**	
	positive	negative
Baseline	507	293
Relapse	507	286
Last - 1	271	279
Last - 2	274	250
Last - 3	158	197
Last - 4	94	143
Last - 5	64	103
Last - 6	42	78

C. Variables Selection

We also did a study of the importance of the variables used in this work. The variables used were the mood symptoms for depression (depression, sleep, interest, guilt, self-esteem, energy, concentration/distractibility, appetite, psychomotor retardation/psychomotor agitation, suicidal ideation) and mood elevation (self-esteem/self-confidence, need for sleep, talking, flight of ideas/racing thoughts, distractibility, goal directed activity, psychomotor agitation, high-risk behavior). Accordingly with the STEP-BD's protocol, the symptoms were categorized with scores 0-normal, euthymic state; 0.25-questionable, slight or rare symptom; 0.5-mild, clearly present but subthreshold for DSM-IV; 1-moderate, clearly present and fulfills DSM-IV criteria; 1.5-marked and 2-severe [9].

D. Machine Learning Algorithms

The algorithms used for prediction were classic ML algorithms from the literature. It included, Support Vector Machines (SVM) with a polynomial kernel; Random Forest (RF) with 100 trees built for classification; Naïve Bayes (NB); and, Multilayer Perceptron (MLP) with the number of hidden layers set as $(attributes + classes)/2$ and the learning rate as $\alpha = 0.3$. Logistic Regression has also been used, since it is well known in the medical scientific community.

Implementations of those algorithms are the ones available at the Weka 3.8.0 framework [13]. The Friedman non-parametric test was used to compare the performance between the classifiers, as suggested by Demšar [14], admitting the tendency of $p > 0.05$.

The 10-fold cross validation method has been used to estimate the performance of the predictive models [15]. The confusion matrix of the classifiers, along with the *Precision* (proportion of predicted positives which are actual positive), *Recall* (proportion of actual positives which are predicted positive) and *F-measure* (harmonic mean between precision and recall) metrics, was calculated in each visit. The F-measure was reported, since it unites the recall and precision metrics in a single equation.

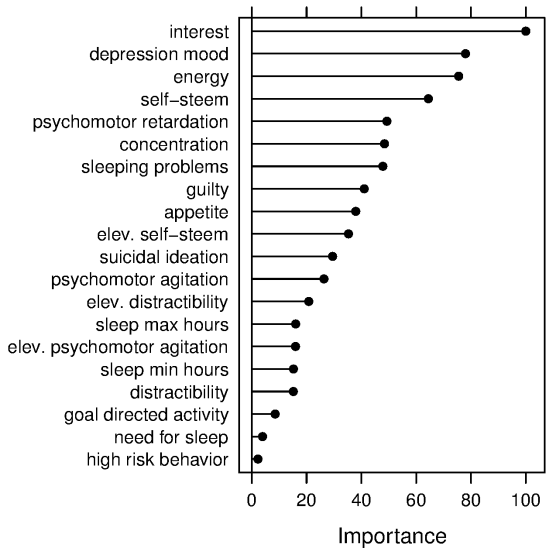

Figure 1. Variable importance measured by the RF algorithm in the "Relapse" visit. This measure is the mean decrease accuracy value that estimate the accuracy if the feature was removed, scaled from 0 to 100.

III. Results

Table II shows the F-measure of the positive (patients that had a relapse) and negative classes for each algorithm in each visit. The last column indicates the F-measure obtained by the classifier for the "Relapse" visit, where patients already had the depressive clinical state. The average and standard deviation were calculated for each classifier from the "Last - 1" up to the "Baseline" visits. Results showed no significant difference between the algorithms. The RF algorithm obtained the best average, with 72.0% for the positive class and 77.3% for negative. The best performance in a single visit was also achieved by the RF with 80.0% F-measure in the "Last - 4" visit.

Figure 1 shows the importance of each symptom according to the best classifier (RF) in the "Relapse" visits. The first and most important feature represents the interest of the patient to enjoy pleasant things. The second feature represents the depression rate, measured by the frequency of days the patient exhibited feelings of dysphoria. The third feature is the level of energy or feeling of weariness after performing any task. Less important features were (1) if the patient exhibited high-risk behaviors, (2) the need for sleep and the ability to function with decreased sleep, and (3) the capacity of the patient to engage in new projects.

IV. Discussion

To our knowledge, there are no previous studies comparing ML algorithms performance to predict depressive relapses in BD patients. The higher performance of the ML algorithms to identify participants who relapsed was observed at the "Relapse" visit. In addition, the algorithms had a tendency to improve their performance, going all the way to the "Last - 6" visit, which opens up the possibility to accurately use intermediate visits to identify future relapses. The overall RF higher performances could be due to it being an ensemble of Decision Trees that works better with the categorical features used in this work.

Considering the higher rates of relapse in BD, more patients were expected to be found in the depressive relapse group, as it had been when it comes to our patients' selection algorithm. We believe that our method of patient selection has been responsible for that since our selection criteria required that all patients had a remission of mood episodes and at least 54 weeks of follow-up after recovery. As a result, we didn't select patients who presented relapse to depression coming from mood episodes such as hypo/mania, mixed and subsyndromic mood states. At the same time, we reduced variables that could influence a higher risk of relapse.

The depressive relapse group showed several clinical features emphasizing greater severity, which are associated with higher risk to relapse. The analysis of the importance of the RF factors in the "Relapse" visit showed cohesion with the relevance criteria observed in the DSM-V [16]. At our analysis the main factors were interest and depression mood. Those two factors are considered the essential factors for the diagnosis of a major depressive episode, which at least one of them was associated with at least four other symptoms. They were followed by other seven depressive symptoms. In BD it is expected, in some cases, to observe mood elevation symptoms during depressive episodes. Hence, our ML approach did also include the presence of hypo/mania mood symptoms to show its relevance on BD recurrence. The two most relevant were increased self-esteem and distractibility.

Table II
F-MEASURE OF THE POSITIVE AND NEGATIVE CLASSES FOR EACH ALGORITHM BY VISIT. BOLD VALUES REPRESENT THE ALGORITHM WITH BEST POSITIVE F-MEASURE IN THAT VISIT AND UNDERLINED VALUES SHOW THE BEST NEGATIVE F-MEASURE.

	Class	Last - 1	Last - 2	Last - 3	Last - 4	Last - 5	Last - 6	Baseline	Avg. (std.)	Relapse
SVM	yes	0.604	0.696	0.615	0.739	0.611	0.729	0.709	0.711 (0.12)	0.991
	no	0.654	0.681	0.703	0.834	0.749	0.852	0.550	0.750 (0.13)	0.984
RF	yes	0.579	0.726	0.669	**0.800**	**0.648**	0.603	**0.744**	**0.720** (0.13)	**0.993**
	no	0.616	0.712	0.783	0.898	0.832	0.859	0.501	0.773 (0.15)	0.988
NB	yes	**0.653**	0.671	**0.675**	0.651	0.615	0.635	0.695	0.698 (0.12)	0.990
	no	0.709	0.684	0.724	0.710	0.755	0.800	0.586	0.743 (0.11)	0.982
MLP	yes	0.613	**0.743**	0.671	0.515	0.602	0.651	0.710	0.687 (0.14)	0.992
	no	0.612	0.729	0.727	0.746	0.736	0.805	0.534	0.734 (0.13)	0.986
LR	yes	0.576	0.638	0.610	0.661	0.630	**0.753**	0.723	0.697 (0.13)	0.987
	no	0.615	0.592	0.634	0.711	0.713	0.865	0.525	0.704 (0.15)	0.977

The mixed characteristics, presence of depressive and mood elevation symptoms at the same time, are associated with a younger age of onset, younger age at hospitalization, more frequent hospitalizations for mixed episodes, more severe symptomatology, mood episode recurrence, higher rates of comorbidity, poorer clinical outcomes, and greater suicide risk. In addition, it is expected that 30-40% patients experience mixed features at some point.

V. CONCLUSION

Our results showed the capability of the intelligent algorithms to predict relapse in earlier visits with a reasonable performance. With more research, ML algorithms could be seen as one option to better support medical decision making and, as a consequence, improve the quality of life of patients with BD. In future work, we will apply methods based on relational learning to build models that consider every visit of the patient at the same time. We speculate that patterns can be identified across the whole course of the patients' BD history leading to better relapse prevention strategies.

ACKNOWLEDGMENT

The authors would like to thank Celso Camilo Júnior for discussions. The first author would like to acknowledge CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for the scholarship.

REFERENCES

[1] F. K. Goodwin and K. R. Jamison, *Manic-depressive illness: bipolar disorders and recurrent depression*. Oxford University Press, 2007, vol. 1.

[2] A. J. Ferrari, E. Stockings, J.-P. Khoo, H. E. Erskine, L. Degenhardt, T. Vos, and H. A. Whiteford, "The prevalence and burden of bipolar disorder: findings from the global burden of disease study 2013," *Bipolar disorders*, vol. 18, no. 5, pp. 440–450, 2016.

[3] G. H. Vázquez, J. N. Holtzman, M. Lolich, T. A. Ketter, and R. J. Baldessarini, "Recurrence rates in bipolar disorder: systematic comparison of long-term prospective, naturalistic studies versus randomized controlled trials," *European Neuropsychopharmacology*, vol. 25, no. 10, pp. 1501–1512, 2015.

[4] B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. P. A. Ioannidis, "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review," *Journal of the American Medical Informatics Association*, vol. 24, no. 1, p. 198, 2017.

[5] D. Librenza-Garcia, B. J. Kotzian, J. Yang, B. Mwangi, B. Cao, L. N. P. Lima, M. B. Bermudez, M. V. Boeira, F. Kapczinski, and I. C. Passos, "The impact of machine learning techniques in the study of bipolar disorder: a systematic review," *Neuroscience & Biobehavioral Reviews*, vol. 80, pp. 538–554, 2017.

[6] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, no. 2004, pp. 1–26, 2004.

[7] T. Deckersbach, A. T. Peters, L. G. Sylvia, A. K. Gold, P. V. da Silva Magalhaes, D. B. Henry, E. Frank, M. W. Otto, M. Berk, D. D. Dougherty *et al.*, "A cluster analytic approach to identifying predictors and moderators of psychosocial treatment for bipolar depression: Results from step-bd," *Journal of Affective Disorders*, 2016.

[8] R. Salvini, R. D. da Silva, B. Lafer, and I. Dutra, "A multi-relational model for depression relapse in patients with bipolar disorder." *Studies in health technology and informatics*, vol. 216, pp. 741–745, 2015.

[9] G. S. Sachs, M. E. Thase, M. W. Otto, M. Bauer, D. Miklowitz, S. R. Wisniewski, P. Lavori, B. Lebowitz, M. Rudorfer, E. Frank *et al.*, "Rationale, design, and methods of the systematic treatment enhancement program for bipolar disorder (step-bd)," *Biological psychiatry*, vol. 53, no. 11, pp. 1028–1042, 2003.

[10] R. H. Perlis, M. J. Ostacher, J. K. Patel, L. B. Marangell, H. Zhang, S. R. Wisniewski, T. A. Ketter, D. J. Miklowitz, M. W. Otto, L. Gyulai *et al.*, "Predictors of recurrence in bipolar disorder: primary outcomes from the systematic treatment enhancement program for bipolar disorder (step-bd)," *American Journal of Psychiatry*, vol. 163, no. 2, pp. 217–224, 2006.

[11] R. H. Perlis, E. B. Dennehy, D. J. Miklowitz, M. P. DelBello, M. Ostacher, J. R. Calabrese, R. M. Ametrano, S. R. Wisniewski, C. L. Bowden, M. E. Thase *et al.*, "Retrospective age at onset of bipolar disorder and outcome during two-year follow-up: results from the step-bd study," *Bipolar disorders*, vol. 11, no. 4, pp. 391–400, 2009.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[14] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, pp. 1–30, 2006.

[15] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2, 1995, pp. 1137–1145.

[16] I. H. Shim, Y. S. Woo, and W.-M. Bahk, "Prevalence rates and clinical implications of bipolar disorder "with mixed features" as defined by dsm-5," *Journal of affective disorders*, vol. 173, pp. 120–125, 2015.