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

GUILHERME ALBERTO SOUSA RIBEIRO

Abordagem de seleção de características baseada em AUC com estimativa de probabilidade combinada a técnica de suavização de La Place

Goiânia
2023

Processo: 23070.050173/2023-45 Documento: 4175818

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) 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 9.610/98, 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.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[ ] Dissertação [X] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Guilherme Alberto Sousa Ribeiro

3. Título do trabalho

A**bordagem de seleção de características baseada em AUC com estimativa de probabilidade combinada a técnica de suavização de La Place**

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

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

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado 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.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Guilherme Alberto Sousa Ribeiro**, **Discente**, em 07/11/2023, às 09:46, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rommel Melgaco Barbosa**, **Professor do Magistério Superior**, em 07/11/2023, às 10:40, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4175818** e o código CRC **AEBC4C66**.

Referência: Processo nº 23070.050173/2023-45 SEI nº 4175818

GUILHERME ALBERTO SOUSA RIBEIRO

Abordagem de seleção de características baseada em AUC com estimativa de probabilidade combinada a técnica de suavização de La Place

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

Área de concentração: Ciência da Computação.
Linha de pesquisa: Sistemas Inteligentes e Aplicações.

Orientador: Prof. Rommel Melgaço Barbosa

Co-Orientadora: Profa. Nattane Luíza da Costa

Goiânia
2023

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

Ribeiro, Guilherme Alberto Sousa
Abordagem de seleção de características baseada em AUC com estimativa de probabilidade combinada a técnica de suavização de La Place. [manuscrito] / Guilherme Alberto Sousa Ribeiro. - 2023.
113, CXIII f.

Orientador: Prof. Dr. Rommel Melgaço Barbosa; co-orientadora Dra. Nattane Luiza da Costa.
Tese (Doutorado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2023.
Bibliografia. Apêndice.
Inclui siglas, algoritmos, lista de figuras, lista de tabelas.

1. Seleção de características. 2. Área abaixo da curva ROC. 3. Aprendizado supervisionado. 4. Classificação. I. Melgaço Barbosa, Rommel , orient. II. Título.

CDU 004

Processo: 23070.050173/2023-45 Documento: 4173887

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE TESE

Ata nº **22** da sessão de Defesa de Tese de **Guilherme Alberto Sousa Ribeiro**, que confere o título de Doutor em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos vinte e oito dias do mês de setembro de dois mil e vinte e três, a partir das onze horas, via sistema de webconferência da RNP, realizou-se a sessão pública de Defesa de Tese intitulada **"Abordagem de seleção de características baseada em AUC com estimativa de probabilidade combinada a técnica de suavização de La Place"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor Rommel Melgaço Barbosa (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professora Doutora Nattane Luíza da Costa (IF Goiano), coorientadora; Professor Doutor Diego de Castro Rodrigues (IFTO), membro titular externo; Professor Doutor Marcio Dias de Lima (IFG), membro titular externo; Professor Doutor Alexandre César Muniz de Oliveira (UFMA), membro titular externo; Professora Doutora Cristhiane Gonçalves (UTFPR), membra titular externa. A realização da banca ocorreu por meio de videoconferência. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Tese, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor Rommel Melgaço Barbosa, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos vinte e oito dias do mês de setembro de dois mil e vinte e três.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Cristhiane Gonçalves**, **Usuário Externo**, em 06/11/2023, às 15:15, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **NATTANE LUÍZA DA COSTA**, **Usuário Externo**, em 06/11/2023, às 15:27, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **DIEGO DE CASTRO RODRIGUES**, **Usuário Externo**, em 06/11/2023, às 15:36, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rommel Melgaco Barbosa**, **Professor do Magistério Superior**, em 06/11/2023, às 15:39, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Guilherme Alberto Sousa Ribeiro**, **Discente**, em 06/11/2023, às 15:43, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Márcio Dias de Lima**, **Usuário Externo**, em 06/11/2023, às 16:51, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **ALEXANDRE CÉSAR MUNIZ DE OLIVEIRA**, **Usuário Externo**, em 07/11/2023, às 07:55, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4173887** e o código CRC **0FB98E91**.

Referência: Processo nº 23070.050173/2023-45 SEI nº 4173887

Guilherme Alberto Sousa Ribeiro

Graduou-se em Ciência da Computação pela Universidade Federal do Maranhão (2013). Possui mestrado em Ciência da Computação pela Universidade Federal do Maranhão (2015). Durante o mestrado, aprofundou seus estudos em mineração de dados para problemas de classificação, desenvolvendo um modelo para classificação de pacientes em específico, utilizando dados relacionados a problemas cardíacos coletados pela Universidade de *Pittsburgh*. Atualmente é Cientista de Dados no Hospital Israelita Albert Einstein que fica na cidade de São Paulo e tem interesse em pesquisas sobre mineração de dados, aprendizado de máquina, seleção de variáveis, visão computacional e aprendizado profundo.

A Deus e a minha família.

Agradecimentos

Agradeço em primeiro lugar a Deus, por ter me dado sabedoria, paciência e perseverança para trilhar todo esse caminho necessário até a defesa do doutorado.

A minha esposa, Fernanda, que foi grande incentivadora e apoiadora durante os anos de realização do doutorado. Sem seu amor, compreensão nos momentos em que estive ausente e conversas nos momentos de desânimo, talvez não teria chegado até aqui. Esse parágrafo não é capaz de expressar o quanto sou grato por ter você ao meu lado e por todo o seu apoio.

A minha avó, Maria da Paz, aos meus pais, João Guilherme e Marilourdes e ao meu avô Sampaio (*in memoriam*) por todo o incentivo, suporte amor e carinho durante toda a minha vida e principalmente nos anos de faculdade e mestrado, que serviram de base para que pudesse estar cursando o doutorado. Ao meu irmão, Lucas, por compreender os momentos de ausência e por tornar divertidos os poucos momentos de lazer que tivemos nesse período. Aos meus filhos de quatro patas, Yoshi e Anne, meus grandes companheiros durante as madrugadas de estudo e momentos de reflexão durante os passeios diários.

Agradeço ao professor Dr. Rommel Melgaço Barbosa por ter aceitado ser o meu orientador, por todos os ensinamentos, confiança e apoio.

Agradeço aos amigos da pós-graduação, Camila, Nattane e Diego, por todas as conversas, conhecimentos e experiências trocadas durante esses últimos anos. Em especial a Camila e a Nattane, pela amizade, parceria e conhecimentos trocados, pois serviram de apoio para conclusão deste trabalho.

Agradeço aos amigos do Hospital Albert Einstein, Márcio, Paulo, Thiago, Luan e Carol, por todo conhecimento e experiências trocadas durante os últimos anos. Todas as conversas, suporte e parceria serviram de apoio para a conclusão deste trabalho.

Agradeço aos professores do Instituto de Informática (INF) da Universidade Federal de Goiás (UFG) por todo ensinamento que foi passado. Todos, sem exceção, contribuíram para a construção do conhecimento que me fez chegar até esse momento.

Agradeço à equipe administrativa do INF, em especial Mariana e Mirian, por toda dedicação e suporte durante todo o meu percurso no programa de doutorado. Vocês são fundamentais para o funcionamento eficiente do programa de pós-graduação.

Agradeço à todos os membros da banca examinadora, pois suas contribuições e correções foram fundamentais para o sucesso deste trabalho.

A todos que contribuíram de forma direta ou indireta para a conclusão desta tese de doutorado, muito obrigado!

Para o sucesso, faça o que ama, trabalhe duro, seja humilde e honesto, saiba ouvir, sonhe grande, dê o seu melhor e nunca desista.

Autor desconhecido,

.

Resumo

Ribeiro, Guilherme A. S.. **Abordagem de seleção de características baseada em AUC com estimativa de probabilidade combinada a técnica de suavização de La Place**. Goiânia, 2023. 113p. Tese de Doutorado. Programa de Pós-Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

A alta dimensionalidade em que muitos dados são dispostos trouxe a necessidade de algoritmos de redução de dimensionalidade, os quais potencializam a performance, reduzem o esforço computacional e simplificam o processamento de dados em aplicações voltadas para as áreas de aprendizagem de máquina ou reconhecimento de padrões. Devido a necessidade e importância de ter uma base de dados reduzida, este trabalho propõe estudo sobre métodos de seleção de características, com ênfase aos métodos que utilizam AUC (*Area Under the ROC curve*). Foram avaliadas as tendências no uso de métodos de seleção de características em geral e para os métodos que usam AUC como estimador, aplicados a dados *microarray*. Em seguida, foi desenvolvido novo algoritmo de seleção de características denominado método de seleção de características baseado em AUC com estimativa de probabilidade e método de suavização de *La PLace* (AUC-EPS). O método proposto calcula o AUC levando em consideração todos os possíveis valores de cada característica, associado a estimativa de probabilidade e ao método de suavização de *La Place* (*smoothing*). Os experimentos foram realizados de forma a comparar a técnica proposta com os algoritmos FAST (*Feature Assessment by Sliding Thresholds*) e ARCO (*AUC and Rank Correlation coefficient Optimization*) a partir do uso de oito conjuntos de dados de relacionadas a expressão genética em *microarrays*, sendo a totalidade de conjuntos utilizada para o experimento de validação cruzada e quatro utilizadas no experimento de *bootstrap*. Os resultados demonstraram que o método proposto colaborou para a melhoria de performance de alguns classificadores e, na maioria casos, atingiu tal objetivo usando um conjunto de características completamente diferente das demais técnicas, sendo algumas dessas características identificadas pelo AUC-EPS determinantes para identificar doenças. O trabalho concluiu que o método proposto, denominado AUC-EPS, seleciona características diferentes dos algoritmos FAST e ARCO, colaborando para a melhoria de desempenho de alguns classificadores e identificando características determinantes para discriminar câncer.

Palavras–chave

Seleção de características, Área abaixo da curva ROC, Aprendizado supervisionado, Classificação.

Abstract

Ribeiro, Guilherme A. S.. **AUC-based feature selection approach with probability estimation combined with LaPlace smoothing technique**. Goiânia, 2023. 113p. Doctor Degree. Programa de Pós-Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

The high dimensionality of many datasets has led to the need for dimensionality reduction algorithms that increase performance, reduce computational effort and simplify data processing in applications focused on machine learning or pattern recognition. Due to the need and importance of reduced data, this paper proposes an investigation of feature selection methods, focusing on methods that use AUC (Area Under the ROC curve). Trends in the use of feature selection methods in general and for methods using AUC as an estimator, applied to microarray data, were evaluated. A new feature selection algorithm, the AUC-based feature selection method with probability estimation and the La PLace smoothing method (AUC-EPS), was then developed. The proposed method calculates the AUC considering all possible values of each feature associated with estimation probability and the La Place smoothing method. Experiments were conducted to compare the proposed technique with the FAST (Feature Assessment by Sliding Thresholds) and ARCO (AUC and Rank Correlation coefficient Optimization) algorithms. Eight datasets related to gene expression in microarrays were used, all of which were used for the cross-validation experiment and four for the bootstrap experiment. The results showed that the proposed method helped improve the performance of some classifiers and in most cases with a completely different set of features than the other techniques, with some of these features identified by AUC-EPS being critical for disease identification. The work concluded that the proposed method, called AUC-EPS, selects features different from the algorithms FAST and ARCO that help to improve the performance of some classifiers and identify features that are crucial for discriminating cancer.

Keywords

Feature selection, Area under the ROC curve, Supervised learning, Classification.

Lista de Siglas

AUC	*Area Under the ROC Curve*
ROC	*Receiver Operating Characteristic Curve*
FAST	*Feature Assessment by Sliding Thresholds*
ARCO	*AUC and Rank Correlation coefficient Optimization*
FROC	*Feature selection based-on ROC-curves*
AUC-EPS	*Area Under the ROC Curve and Estimation Probability and Smoothing*
AVC	*AUC-based Variable Complementarity*
SVM	*Support Vector Machines*
KNN	*K-nearest neighbors*
AD	Árvore de Decisão
FR	Florestas Randômicas
ID3	*Iterative Dichotomiser 3*
C4.5	*Classification and regression tree*
CART	*Classification and Regression Trees*
RF	*Random Forests*
VP	Verdadeiros Positivos
FP	Falsos Positivos
KDD	*Knowledge Discovery in Database*
WoS	*Web of Science*
MTD	Matriz de Termos de Documentos
RFE	*Recursive Feature Elimination*
AMC	Análise de Múltipla Correspondência
LDA	*Latent Dirichlet Allocation*
SVM-RFE	*Support Vector Machines with Recursive Feature Elimination*
PSO	*Particle Swarm Algorithm*
PCA	*Principal Component Analysis*

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

CAPÍTULO 1

Introdução

A seleção de características pode ser utilizada como uma das fases da análise de dados. Ela ajuda a compreender melhor os dados, reduzindo a dimensionalidade e melhorando a performance em relação a classificação. Essa técnica faz parte de uma das etapas mais importantes do pré-processamento dentro da área de *Knowledge Data Discovery* [11]. Ao executar uma cuidadosa seleção de características, é possível melhorar significativamente o desempenho final de um classificador, utilizando um custo computacional mais baixo do que se toda a base de dados fosse utilizada. Além disso, trabalhar com uma base reduzida, sem redundâncias, pode evitar o *overfitting* por parte do classificador e melhorar o poder de generalização do modelo [34, 126].

Métodos de seleção de características podem ser categorizados como supervisionados, semi-supervisionados e não supervisionados. Para o primeiro, as técnicas são aplicadas em bases de dados com instâncias rotuladas, ou seja, que possuem uma classe, para o segundo caso, parte das instâncias estão rotuladas e parte não estão rotuladas e por fim, para a última categoria, os dados não estão rotulados [52, 85, 110, 10]. Dentre a categoria dos supervisionados, os métodos de seleção de características podem ser divididos em três categorias: métodos de *filtering*, *wrapper* e *embbeded*. Os métodos de *filtering* usam os valores das características levando em consideração o rótulo de classe de cada uma das instâncias através do uso de alguns métodos estatísticos [54, 53, 70, 95]. Os métodos *wrapped* trabalham dentro de um espaço de busca e selecionam um subconjunto de características, avaliam a performance do classificador com subconjuntos selecionados usando acurácia e o subconjunto selecionado será aquele que possuir um maior resultado para a métrica selecionada [62, 111, 75]. Por fim, os métodos *embbeded* que são métodos formados através da combinação de métodos de *filtering* e *wrapped* [52, 85, 110, 10].

Entre as técnicas de seleção de características, a *Area Under the Curve* (AUC) é bastante utilizada para avaliar variáveis em bases de dados com características de desbalanceamento [140]. Essa área pode ser definida como a representação da probabilidade em que um método pode classificar um dado corretamente. O valor dessa área varia de 0 a 1. Quanto mais próximo de 1, melhor é o modelo e quanto mais próximo de 0.5, significa que o modelo tem um alto grau de aleatoriedade. Isso é possível devido a *Receiver Ope-*

rating Characteristic Curve (ROC) comparar a performance dos métodos de classificação através de todo o intervalo de distribuição das classes e o erro obtido. Devido essas características a AUC tem sido bastando usada também para avaliar a performance de métodos de aprendizagem de máquina [83]. Alguns dos métodos de seleção de características baseado em AUC são os algoritmos *Feature Assessment by Sliding Thresholds* (FAST) e *AUC and Rank Correlation coefficient Optimization* (ARCO).

O método de avaliação de características FAST utiliza a área abaixo da curva ROC para medir o grau de importância de cada uma das características de um conjunto de dados baseado na classe das instâncias [140]. Por outro lado, o algoritmo ARCO [135] usa o mesmo cálculo de AUC do algoritmo FAST, mas com o diferencial de que o ARCO utiliza o coeficiente de correlação de *Spearman* para medir o grau de redundância entre as características. Uma outra abordagem que utiliza AUC como parte do cálculo de seleção de atributos é o *Feature selection based-on ROC-curves* (FROC) [86]. No FROC, a seleção de características acontece de acordo com aquelas que estiverem com valor entre área existente entre a curva ROC e a linha diagonal que vai do ponto (0,0) ao ponto (1,1) (AUC - 0,5). Uma vez feito isto, as características redundantes são removidas utilizando a análise de *Markov Blanket*.

Grande parte dos métodos que utilizam o valor de AUC para definir o grau de importância de uma característica levam em consideração apenas a quantidade de amostras que pertencem a cada uma das classes do problema, deixando de dar importância para a variação de valores que existe dentro de cada uma delas. Dessa forma, o valor de AUC atribuído a uma característica pode não representar o seu real grau de importância.

A lacuna consiste na falta de estudos que utilizem técnicas de seleção de características baseados em AUC que considerem a variabilidade de valores para cada característica, através da combinação das técnicas de estimativa de probabilidade, cálculo de AUC e do método de suavização de *La Place*. Métodos de seleção de características baseados em AUC geralmente utilizam o particionamento dos dados com base nas classes para calcular o grau de importância de uma determinada característica. No entanto, levar em consideração o grau de variabilidade de cada característica é um desafio. Não foi encontrado na literatura, de forma abrangente, esse contexto específico. Portanto, há necessidade de investigar novas abordagens ou combinação de técnicas que possam aprimorar a seleção de características dentro desse contexto, o que justifica este trabalho.

A originalidade do trabalho reside na proposta de combinar diferentes técnicas (cálculo de AUC, estimativa de probabilidade e método de suavização de *La Place*) para aprimorar a seleção de características em relação aos métodos tradicionais existentes baseados em AUC.

Baseado na justificativa, formula-se a hipótese de que se um algoritmo de seleção de características baseado em AUC, combinado com outras técnicas, pode ser utilizado

como ranqueador dos melhores atributos de um conjunto de dados, então é possível que através da técnica de seleção de características baseada em AUC combinada as técnicas de estimativa de probabilidade e ao método de suavização de *La Place* será possível selecionar as melhores características de um conjunto de dados, levando em consideração o grau de variabilidade de cada uma delas.

Este trabalho propõe um algoritmo de seleção de características nomeado AUC com estimativa de probabilidade e *smoothing*, que seleciona os melhores atributos de um conjunto de dados considerando todos os possíveis valores de cada característica, associado a estimativa de probabilidade e ao método de suavização de LaPlace (*smoothing*). A ideia de utilização dessas técnicas combinadas surgiu baseada no desenvolvimento de modelos de classificação que combinaram essas técnicas e obtiveram boas performances [129, 107]. Além disso, o fato de não ter sido encontrado um trabalho na literatura que utilize essas técnicas combinadas para seleção de características foi uma das motivações do trabalho. Os experimentos são conduzidos através de duas metodologias e os resultados são comparados com os algoritmos ARCO e FAST.

A inovação deste trabalho está na utilização do cálculo de AUC combinado a estimativa de probabilidade e a técnica de suavização de *LaPlace* para selecionar características. A relevância está na novidade do método e o algoritmo que considera o grau de variabilidade de cada característica antes de calcular o grau de importância delas. É um trabalho que não envolve custo, a não ser o esforço de desenvolvimento. Haverá custo reduzido somente em caso de necessidade de implantação em algum ambiente. Este trabalho pode ser aplicado em hospitais, empresas de biotecnologia e dados *microarray*, pois é de grande valia na redução de dimensionalidade e maximização de processos e resultados, facilitando assim a detecção de doenças.

1.1 Objetivos

1.1.1 Objetivo geral

Investigar as tendências no uso de métodos de seleção de características em geral e métodos de seleção de características que usam AUC como estimador, aplicados a dados *microarray* e melhorar a aplicabilidade das técnicas baseadas em AUC em pesquisas científicas.

1.1.2 Objetivos específicos

- Identificar as técnicas de seleção de características mais utilizadas atualmente, com o foco nas técnicas que utilizam AUC como métrica de seleção de atributos;

- Verificar a associação de um algoritmo de seleção de características baseado em AUC, com técnica de estimativa de probabilidade e a técnica de *smoothing*;
- Verificar o desempenho de classificadores, com um subconjunto de características selecionadas pelo método proposto, em bases de dados com alta dimensionalidade e baixa quantidade de instâncias;
- Comparar o modelo proposto com os métodos clássicos de seleção de características baseados em AUC;
- Avaliar e discutir os resultados obtidos e definir os próximos passos com o intuito de explorar mais possibilidades com o método proposto.

1.2 Publicações

Durante o período do curso de doutorado, dois estudos foram publicados em revistas/conferências internacionais bem conceituadas. Um dos trabalhos apresenta o algoritmo desenvolvido durante a tese com alguns resultados parciais, além de comparações com outras abordagens de seleção de características similares, O outro apresenta um estudo de revisão utilizando bibliometria com o objetivo de verificar pontos relevantes na relação da seleção de características com bases de dados *microarray*.

O artigo publicado foi:

1. *Feature Selection Method AUC-Based with Estimation Probability and Smoothing* (2021). *International Conference on Engineering and Emerging Technologies (ICEET), Istanbul, Turkey*, p. 1-8. [106]

Além dessa publicação, o seguinte estudo está submetido e em processo de revisão por pares em revista internacional:

1. *From Bibliometrics to Text Mining: Exploring Feature Selection Methods in Microarray Research* (Submetido na revista Communications in Statistics - Simulation and Computation);

Por fim, o(s) seguinte(s) estudo(s) está(ão) em processo de finalização de escrita para então ser(em) submetido(s):

1. Bibliometria relacionada a seleção de características utilizando AUC como técnica para realização da seleção;
2. Resultados dos testes estatísticos comparando os resultados do método proposto com os demais métodos;

1.3 Organização do trabalho

Este trabalho está organizado da seguinte forma:

- O **Capítulo 1** é o introdutório que busca contextualizar o problema, apresentá-lo e apresentar os objetivos e motivações do trabalho;
- No **Capítulo 2** são apresentados os conceitos fundamentais do trabalho, ou seja, principais conceitos e técnicas referentes a seleção de características e aprendizado supervisionado que fornecem embasamento teórico para o trabalho que está sendo desenvolvido;
- O **Capítulo 3** apresenta os principais trabalhos correlacionados ao tema desenvolvido nesta tese através do uso das técnicas de bibliometria e mineração de texto;
- O **Capítulo 4** apresenta o método proposto para realizar seleção de características utilizando AUC com estimativa de probabilidade e o método de suavização de *La Place* (AUC-EPS).
- O **Capítulo 5** é o capítulo que apresenta todo o ambiente de realização dos experimentos, bases de dados utilizadas e resultados apresentados para todas as metodologias de testes que foram utilizadas;
- O **Capítulo 6** apresenta as conclusões obtidas em relação ao método proposto após a realização dos experimentos iniciais, destacando as oportunidades de melhoria e as limitações do AUC-EPS.

CAPÍTULO 2

Conceitos fundamentais

2.1 Seleção de características

A redução de dimensionalidade é uma técnica utilizada para remoção de características redundantes e irrelevantes de uma base de dados. A partir dessa redução é possível melhorar a performance de acurácia, diminuir a complexidade computacional dos modelos, construir modelos mais generalistas (evita *overfitting*) e diminui a quantidade de armazenamento necessário para trabalhar com os dados [133, 127]. Duas abordagens são utilizadas para cumprir com esse objetivo: a extração de características e a seleção de características [41]. A primeira abordagem consiste na aplicação de uma série de transformações, sejam elas lineares ou não, sobre os dados, com o propósito de transpô-los de um espaço de características original para um novo espaço com uma dimensionalidade reduzida. Por outro lado, a seleção de características reduz diretamente o número original de características através da seleção de um subconjunto delas que possuem informações suficientes para manter a capacidade de representação dos dados.

A seleção de características tem sido uma área de pesquisa bastante ativa em tópicos relacionados a mineração de dados, reconhecimento de padrões e também em estudos estatísticos. De acordo com Liu et al. [79], a principal ideia da seleção de características é escolher um subconjunto de características dentro de uma base de dados, eliminando as características com baixa relevância ou que possuem informações irrelevantes para a predição e também aquelas que possuem características redundantes que são fortemente correlacionadas.

Grande parte dos algoritmos de seleção de características consistem em quatro passos: geração de um subconjunto de características, critério de parada, avaliação do subconjunto e validação dos resultados [33, 124]. A Figura 2.1 apresenta um fluxo desses passos. Em cada passo do processo de busca, um subconjunto de um conjunto de características candidatas é gerado a partir das originais e seus valores gerados e avaliados de acordo com determinado critério. O processo de formação do subconjunto seguirá até que um critério de parada pré-estabelecido tenha sido atingido. No fim, é feita a avaliação deste subconjunto de características selecionadas no conjunto de dados.

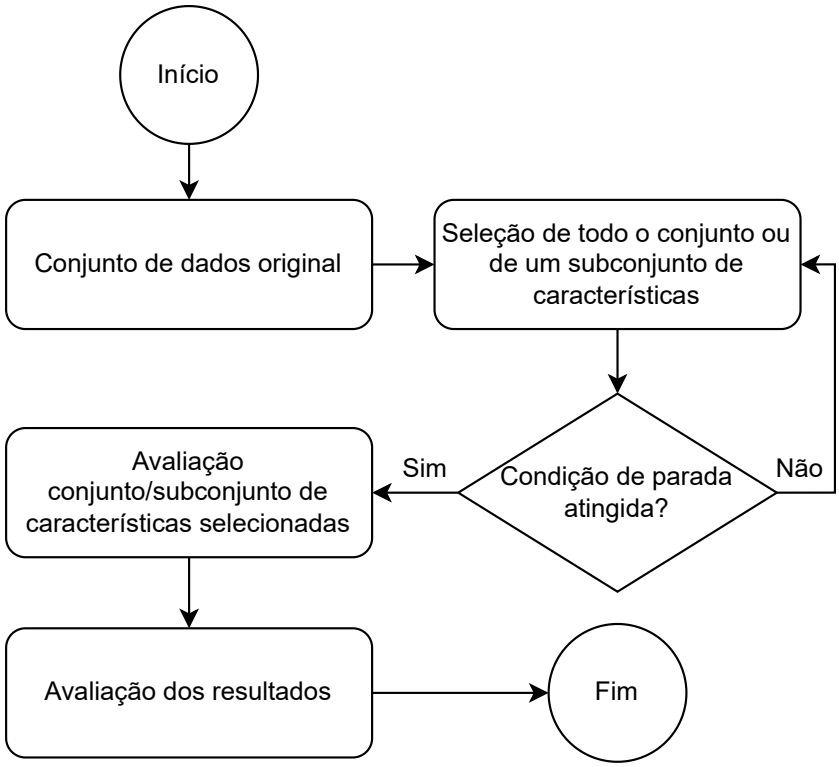

Figura 2.1: Representação do fluxo de um algoritmo de seleção de características.

De forma geral, métodos de seleção de características são divididos em duas categorias: supervisionados e não supervisionados [125]. Nos métodos supervisionados, um conjunto de dados é disponibilizado com as características que representam determinados rótulos de classe, enquanto os métodos não supervisionados não possuem esses rótulos de classe. Devido a existência dos rótulos, os métodos supervisionados possuem uma grande vantagem em relação aos não supervisionados, por conta da dificuldade que os não supervisionados possuem para selecionar características que realmente sejam relevantes e que façam a diferença na classificação [35, 150].

Os métodos de seleção de características podem ser divididos em três grupos: métodos de *filtering*, *wrapper* e *embbeded*. Os métodos de *filtering* avaliam a relevância das características sem o uso de algoritmos de aprendizagem. Nesses métodos, características são avaliadas e ranqueadas utilizando algumas medidas e aquelas que possuem maior ranking são selecionadas [72]. Dentro dessa abordagem, os métodos podem ser classificados como univariados e multivariados. O primeiro método utiliza apenas uma variável para o cálculo de um valor, como por exemplo a informação mútua, o *score* Laplaciano, *score* de *Fisher*, ganho de informação, dentre outros [144, 60, 51, 100]. Os métodos multivariados surgiram devido os métodos univariados ignorarem a dependência existente entre características similares no conjunto final. Os métodos que se enquadram nesse tipo manipulam informações irrelevantes e redundantes para suas estratégias de ranqueamento,

ou seja, métodos de *filtering* multivariados foram propostos, visando a consideração de dependências de características.

A técnica *wrapper* visa treinar um classificador para avaliar um conjunto de características. Nessa técnica um modelo de aprendizagem de máquina é usado para avaliar um subconjunto de características com a intenção de selecionar o conjunto que oferecer um maior valor de acurácia para a tarefa de classificação. Grande parte dos métodos *wrapper* utilizam processos de busca iterativos, no qual cada iteração do modelo de aprendizagem é usada para orientar a população de soluções para a melhor delas. No entanto, essas técnicas sofrem com o elevado custo computacional e a perda do poder de generalização devido essa categoria de algoritmos estar associada a um modelo de aprendizado de máquina [22, 77].

Os métodos *embbeded* são a combinação dos métodos de *filtering* com os métodos de *wrapper*, de forma a tentar tirar as melhores características dos métodos. Tal abordagem se concentra principalmente na obtenção do melhor desempenho através de um algoritmo de aprendizagem (característica *wrapper*) e de uma complexidade baixa (característica *filtering*) [3, 117].

As abordagens embutidas consideram o problema de seleção de características como parte de um problema completo de aprendizagem de máquina. Nesse tipo de abordagem um método de aprendizagem de máquina é utilizado em busca de um subconjunto de características final [151].

2.1.1 Métodos de seleção de características baseados em AUC

Dentre os métodos de seleção de características, existem aqueles que utilizam o AUC como métrica para formar um *ranking* ou escolher um subconjunto de características.

Um dos métodos de seleção de características baseado em AUC mais populares é o algoritmo *Feature Assessment by Sliding Theresholds*, conhecido pela sigla FAST [140]. Esse método simplesmente realiza o cálculo de AUC para cada uma das características da base de dados. Logo depois, é feito um ranqueamento das características de acordo com o valor de AUC obtido por cada uma. Esses valores podem ser calculados através da área que existe abaixo da curva ROC ou através do teste de *Wilcoxon-Mann-Whitney* [141] que tem um valor equivalente a técnica que foi mencionada anteriormente e simplifica o cálculo. A Equação (2-1) mostra como o AUC é calculado no algoritmo FAST:

$$AUC = \frac{\sum_{i=1}^{n_+}(r_i - i)}{n_+ * n_-} = \frac{\sum_{i=1}^{n_+}(r_i) - \frac{n_+ * (n_+ + 1)}{2}}{n_+ * n_-}, \tag{2-1}$$

em que n_+ é a quantidade de instâncias positivas, r_i é o ranking obtido por cada exemplo positivo e n_- é a quantidade de instâncias negativas.

Um outro método que surgiu para solucionar o problema de redundância de características que existe no método FAST foi o *AUC and Rank Correlation coefficient Optimization*(ARCO) [135]. Esse método utiliza o cálculo de AUC associado ao processo de eliminação de características redundantes. A metodologia de exclusão das características redundantes é realizada através do ranking do coeficiente de correlação *Spearman*. Esse coeficiente é calculado de acordo com a Equação (2-2),

$$RCC(v_1, v_2) = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n * (n^2 - 1)}, \tag{2-2}$$

em que v_1 e v_2 são as duas variáveis que são comparadas, d_i é a diferença entre cada ranking instâncias das duas características e n é o número de instâncias. No entanto, esta é apenas uma parte do cálculo, pois para selecionar f características de todo o conjunto de dados, o algoritmo precisa selecionar uma inicialmente, que é aquela que possui o maior valor de AUC. Depois dessa seleção, o algoritmo ARCO irá iterar através de um *loop* entre as características que ainda não foram selecionadas e irá selecionar aquela que tiver o maior valor de $E(v_i)$, que é calculado através da Equação (2-3),

$$E(v_i) = AUC(v_i) - \frac{\sum_{v_j \in S} RCC(v_i, v_j)}{|S|}, \tag{2-3}$$

em que AUC(v_i) é o valor de AUC da característica v_i, S é o subconjunto atual de características já selecionadas e $|S|$ é o tamanho desse subconjunto.

O algoritmo *Feature selection based-on ROC-curves* (FROC) é uma outra abordagem de seleção de características que utiliza AUC [86]. Basicamente, esse método possui dois passos principais: o passo denominado *one-gene-at-a-time* baseado na curva ROC e a filtragem de *Blanket Markov* baseada na curva ROC. No primeiro passo a curva ROC é usada como um critério para descobrir a relevância das características em relação as classes. Ao invés de calcular o valor de AUC diretamente, essa fase terá como valor para representar uma característica a área entre a área abaixo da curva ROC e a reta traçada entre os pontos (0,0) e (1,1) de acordo com a Figura 2.2, adaptada de [6]. O segundo passo remove as características redundantes utilizando usando a filtragem de *Blanket Markov* [6]. Nessa segunda abordagem, o algoritmo FROC usa a área entre a curva ROC e a diagonal para medir o grau de redundância entre duas características. A variável que apresentar a menor área entre a reta diagonal formada pelos pontos (0,0) e (1,1) e a área abaixo da curva é considerada a mais redundante entre as duas características avaliadas e será removida. O grande problema desse método é que o algoritmo *Blanket Markov* não avalia corretamente algumas características. Isso pode levar o algoritmo a remover carac-

terísticas indevidamente, gerando assim uma influência em todo o processo de seleção do conjunto de características e consequentemente no processo de classificação.

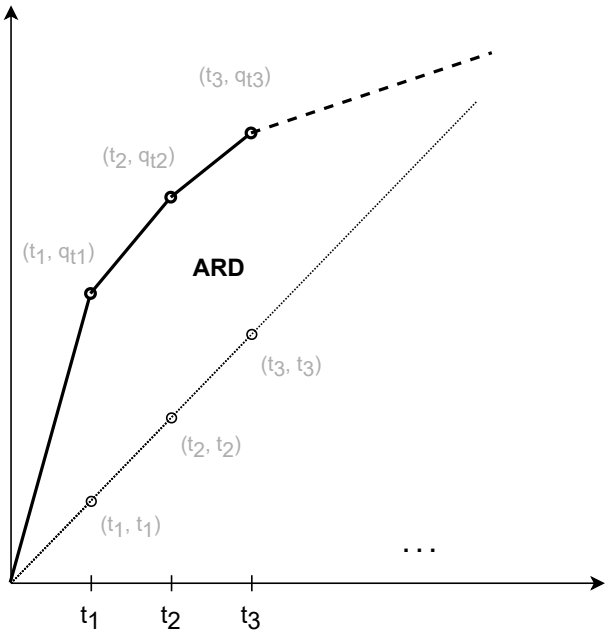

Figura 2.2: Imagem que representa a área (ARD) que é calculada pelo algoritmo FROC para cada característica.

Outro trabalho relacionado a técnicas de seleção de características baseadas em AUC, foi o método baseado em AUC com variável de complementaridade (AVC) [121]. Esse método utiliza a técnica de análise da curva ROC para acessar a relevância das características de acordo com as classes. No entanto, ele explora as informações das instâncias classificadas incorretamente através de classificadores de uma característica baseados na curva ROC para analisar a complementaridade das características. Aparentemente, quando é selecionada uma característica individualmente para observação de dimensão, mais ou menos instâncias podem ser classificadas incorretamente. Tal cálculo de complementaridade associado a seleção de características faz com que este método funcione de forma similar ao algoritmo *RelieF* [55], por exemplo.

No método AVC, foi aplicada uma estratégia eficiente de busca para selecionar o conjunto ótimo de características com os maiores valores de complementaridade. O primeiro passo é selecionar a característica mais significante com o maior valor de AUC no estado inicial. Depois, iterativamente, são selecionadas as características que tem complementaridade máxima de acordo com as características selecionadas na fase inicial. Quando a busca por características ótimas no estado atual são somados os valores das duas características como seus pesos de complementaridade. A ideia é que para cada característica, se existem mais de uma característica com o mesmo valor de

complementaridade para ela, a preferencia será por uma característica com o máximo valor de AUC.

2.2 Aprendizado de máquina

Um conjunto de dados usado na tarefa de aprendizagem consiste em um conjunto de registros (também chamados de exemplo, instância, caso, ou um vetor) que descrevem uma experiência passada, dispostos como uma simples tabela relacional que são descritos por um conjunto de características, cada um com um determinado tamanho e conjunto de valores. Cada registro desse conjunto de dados tem um atributo alvo que pode-se chamar de classe e cada um dos valores existentes nesse atributo podem ser chamados de rótulo de classe [64, 84].

Dado um determinado conjunto de dados, o objetivo do aprendizado é produzir uma função de predição para relacionar os valores dos atributos com as classes. Essa função será utilizada para realizar predições de rótulos de classe para novos dados que serão apresentados a ela. Tal função é chamada de modelo de classificação (modelo preditivo ou simplesmente classificador). Esses modelos podem ser de diversas formas, desde uma árvore de decisão a um modelo de separação por hiperplano como modelos Bayesianos. Essa abordagem de utilizar dados que possuam os rótulos de classe para a tarefa de fazer com que a máquina aprenda é denominada aprendizado supervisionado.

O aprendizado supervisionado tem tido bastante sucesso nas aplicações de *machine learning*. É utilizado em aplicações de quase todos os tipos de domínio. Esse tipo de aprendizado é também chamado de classificação ou aprendizagem indutiva, dentro da área de aprendizagem de máquina. Tal aprendizado é análogo ao aprendizado humano baseado em experiências passadas para ganhar um novo conhecimento com o objetivo de melhorar a habilidade de realizar tarefas do mundo real [29].

O aprendizado supervisionado é o tipo de aprendizado mais comum, em grande parte das aplicações de aprendizagem de máquina [13]. Ele utiliza um elemento supervisor durante o processo de aprendizado, que será responsável por dizer se as previsões feitas por este processo estão corretas. Utilizando árvores de decisão como exemplo, o elemento supervisor se baseia em um conjunto de treinamento, composto por diversas instâncias, que por sua vez possuem rótulos conhecidos por tal elemento. Ao realizar suas previsões, a máquina compara o resultado obtido com o rótulo já conhecido, a fim de verificar se condiz com o resultado desejado, de forma a minimizar o erro desse conjunto de dados [115]. Essa forma de aprendizado é muito utilizada, pois como os rótulos já são conhecidos, o processo de aprendizado é realizado de maneira mais simples. No entanto, a utilização dessa metodologia depende muito do que se deseja fazer com os dados.

Os dados usados para o aprendizado são chamados de dados de treino [103, 130]. Depois do modelo treinado com um modelo de aprendizagem usando esses dados ele é avaliado usando um conjunto de dados de teste para avaliar a performance do modelo. Para realização desta tarefa de avaliação de desempenho, várias métricas têm sido aplicadas na literatura como acurácia, precisão, AUC, dentre outras [63].

2.3 Algoritmos de classificação

Nesta seção são citados alguns algoritmos que podem ser utilizados para realização da tarefa de classificação e que serão utilizados para a realização de alguns experimentos. Serão descritos os algoritmos *Support Vector Machines* (SVM), *k-Nearest Neighbor* (kNN), *Naive Bayes*, Árvore de Decisão (AD) e Florestas Aleatórias (FA) por conta da relação entre simplicidade e performance, ou seja, são métodos fáceis de serem implementados e tem bons resultados para a tarefa de classificação.

2.3.1 Support Vector Machines (SVM)

O algoritmo de Máquinas de Vetor de Suporte (SVM), proposto por [15], é uma das técnicas mais eficazes para a mineração de dados. Esse algoritmo foi criado inicialmente para resolver problemas de classificação, porém, ele também pode ser utilizado para resolver problemas de regressão ou clusterização. As especificações desse algoritmo, bem como suas variações serão apresentadas a seguir.

O SVM busca dividir o conjunto de dados em hiperplanos através da minimização dos erros, de forma que as classes que compõem o conjunto de dados possam ser linearmente separáveis. A ideia do algoritmo é separar estas classes ao máximo. O algoritmo pode ser descrito matematicamente [112] da seguinte forma: dado o conjunto de treinamento

$$\mathrm{T} = \{(\vec{x_1}, y_1), \ldots, (\vec{x_l}, y_l)\} \in (\mathbb{R}^n \times \{-1, 1\})^l$$

$\vec{x_i}$ é a entrada, com os atributos $[\vec{x_i}]_j$ e y_i é a saída do algoritmo (-1 ou 1). Com isso, para um novo valor de entrada $\vec{x} = ([x_1], ..., [x_n])^T$ o algoritmo deve ser capaz de encontrar os y correspondentes.

Durante a etapa de treinamento uma distribuição de probabilidade existirá, de forma que o processo de treinamento resultará em um classificador que seja capaz de aprender um mapeamento $x \rightarrow y$ através dos exemplos de treinamento, de forma que dado um novo exemplo nunca visto pelo algoritmo, ele siga esta mesma distribuição de probabilidade e determine a classe deste novo exemplo. Para que as classes possam ser definidas, o SVM trabalha com a minimização do erro, através do risco empírico, que

é uma média da taxa de erro dos elementos do conjunto de treinamento. Além disso, deve haver um cálculo para restringir o valor do erro durante a execução do algoritmo, relacionando o risco esperado de uma função ao seu risco empírico e um termo de capacidade, denominado limite de erro esperado. Existem diversas maneiras de calcular este limite, algumas delas estão apresentadas em [82].

As classes devem ser separadas através da maximização das margens dos planos, por meio dos pontos mais próximos do hiperplano traçado para separá-los. A Equação (2-4) apresenta como são construídos os hiperplanos de separação, onde w é o vetor de pesos, com mesma dimensão das amostras (x) e b é um escalar.

$$f(x) = wx + b. \tag{2-4}$$

O vetor w é perpendicular ao hiperplano e deve ser maximizados para que a margem separe os dados da melhor forma possível. Essa margem é calculada através da Equação (2-5),

$$m = \frac{2}{||w||}. \tag{2-5}$$

A Figura 2.3 apresenta a separação de classes por meio da aplicação do algoritmo SVM.

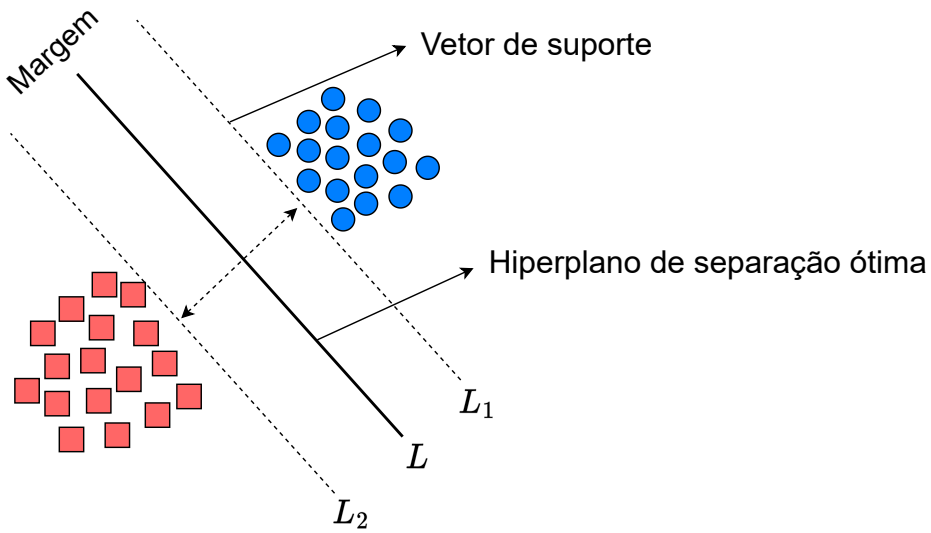

Figura 2.3: Exemplo de aplicação do SVM

Além de todas essas características, o SVM utiliza um *kernel* que é parte crucial desse algoritmo, pois permite transformar os dados de entrada em um formato em que a separação das classes é mais eficaz, mesmo quando as classes não são linearmente separáveis no espaço de características original. O objetivo principal do *kernel* é realizar

transformações para encontrar o hiperplano de separação ótima, ou seja, que maximize a distância (margem) entre as classes. Isso permite ao SVM encontrar um limite de decisão em um espaço de alta dimensão, mesmo quando esse limite não for linear. Existem três tipos mais comuns de *kernel*, como apresentado na Tabela 2.1.

Tabela 2.1: Tipos de *kernel*

Tipos de Kernel	**Função** $K(x_i, x_j)$	**Parâmetros**
Polinomial	$(\delta\ (x_i \cdot x_j) + \kappa)^d$	δ, κ e d
Gaussiano	$\exp\ (\sigma - \|\|x_i - x_j\|\|^2)$	σ
Sigmoidal	$\tanh(\delta\ (x_i \cdot x_j) + \kappa)$	δ e κ

O SVM lida muito bem com conjuntos de dados lineares, sendo possível fazer uma separação entre as classes através de um hiperplano facilmente. Porém, existe a possibilidade do SVM lidar com problemas não lineares transformando-os em problemas linearmente separáveis através um novo espaço de dimensão maior, chamado de espaço de características [82]. Essa proposta está baseada no Teorema de *Cover*. De acordo com este Teorema, dado um conjunto de dados não-linear no espaço de entradas X, ele afirma que X pode ser transformado em um espaço de características $\Im$, no qual esse possui alta possibilidade de que os dados sejam linearmente separáveis [59].

2.3.2 Naive Bayes

Um outro algoritmo bastante utilizado para a classificação é o *Naive Bayes*. Através do Teorema de Bayes, esse algoritmo calcula a probabilidade de um evento acontecer (A) dado que outro ocorra (B), onde o evento ocorrido é a evidência e a probabilidade dele acontecer é a hipótese [138]. A prerrogativa assumida aqui é que as características são independentes, ou seja, a presença de uma determinada característica não afeta a outra. Exemplificando, uma fruta pode ser considerada como maça se for vermelha, redonda e com cerca de 3 polegadas de diâmetro. Mesmo que essas características dependam uma das outras ou da existência de outras, todas essas particularidades contribuem independentemente para a probabilidade de que esta fruta seja uma maçã.

O *Naive Bayes*, apesar de sua simplicidade, é um modelo conhecido por ter desempenho superior que vários métodos de classificação mais sofisticados. O cálculo da probabilidade posterior é realizado através da Equação (2-6)

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \tag{2-6}$$

em que $P(c|d)$ é a probabilidade posterior da classe c, dada uma variável preditora d; $P(c)$ é a probabilidade da classe prioritária; $P(d|c)$ é a probabilidade de uma variável preditora para uma determinada classe; e $P(d)$ é a probabilidade prioritária da variável preditora.

Para realizar o processo de aprendizagem de um algoritmo Naive Bayes os seguintes passos devem ser seguidos:

1. Converter os dados em uma tabela de frequência;
2. Criar uma tabela de probabilidades para encontrar as probabilidades de cada uma das variáveis em relação as classes e também para calcular as probabilidades gerais para da um dos possíveis valores da variável preditora;
3. Utilizar a Equação (2-6) para calcular a probabilidade posterior para cada uma das classes. A classe que obtiver a maior probabilidade posterior é a saída predita pelo modelo.

2.3.3 k-Nearest Neighborn (kNN)

O algoritmo *k-Nearest Neighborn (kNN)* é baseado no princípio que as instâncias dentro de um conjunto de dados possuem proximidades umas com as outras por conta de algumas propriedades existentes nelas [26]. O que define o rótulo de uma instância sem classificação é a proximidade dela com outra instância, que possui um determinado rótulo. A Figura 2.4 apresenta como funciona um algoritmo *kNN*, antes e depois do treinamento. A métrica utilizada busca a menor distância entre as instâncias de treino e a de teste a serem classificadas, dependendo do valor de k, que na Figura 2.4 é igual a 3. Isso faz com que a vizinhança local seja composta apenas por pontos que possuem similaridade com a instância que se deseja classificar, fazendo com que as demais fiquem fora do raio de vizinhança.

Para uma instância de teste, o método kNN, por exemplo, seleciona as k instâncias de treinos mais parecidas e considera tanto em termos de média ou leva em consideração a maioria dos votos dos seus valores alvo. Versões modificadas do kNN têm sido aplicadas com sucesso em conjuntos de dados clínicos, para diagnósticos e extração de conhecimento [47].

Em geral, as instâncias são consideradas como pontos dentro de um espaço de instâncias n-dimensional, onde cada uma dessas dimensões corresponde a uma das características que são utilizadas para descrever uma instância. A similaridade de uma instância de teste com uma de treino é verificada através de uma distância, que é determinada através de um cálculo, utilizando uma métrica. Existem diversas formas de calcular esta proximidade de uma instância a ser rotulada com uma que se deseja rotular. O método mais utilizado é a Distância Euclidiana, que pode ser calculada conforme a Equação (2-7)

$$D(x,y) = \left(\sum_{i=1}^{m} | x_i - y_i |^2 \right)^{\frac{1}{2}}. \tag{2-7}$$

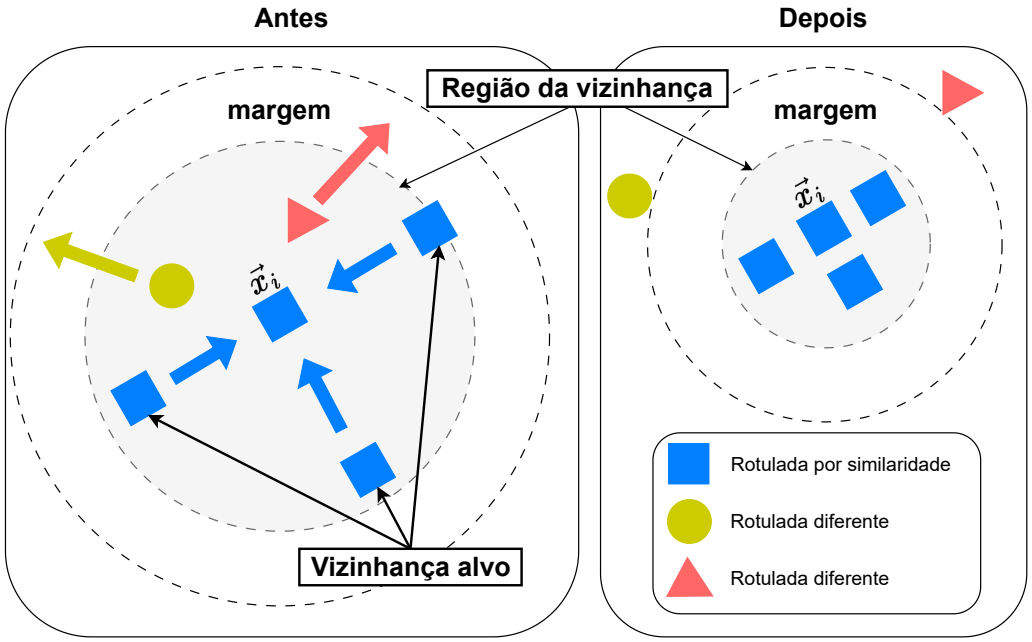

Figura 2.4: Apresentação do funcionamento de um algoritmo kNN. Adaptada de [139].

Os valores de x_i e y_i correspondem a dois pontos, nos quais se deseja encontrar a distância entre eles e m é a quantidade total de pontos que realizam o cálculo descrito pela Equação (2-7).

2.3.4 Árvore de Decisão

As árvores de decisão (AD) são modelos de aprendizado de máquina que podem ser usados para resolver problemas de classificação e regressão [90]. Uma árvore de decisão é uma estrutura hierárquica composta por nós e arestas, onde cada nó representa uma decisão ou um atributo, e as arestas representam as possíveis respostas ou resultados dessas decisões, conforme apresentado na Figura 2.5. O processo de construção de uma árvore de decisão envolve a divisão recursiva do conjunto de dados com base nos atributos para formar subconjuntos mais puros ou homogêneos de dados em relação à variável de destino.

Os algoritmos ID3 (*Iterative Dichotomiser 3*) [96], C4.5 (*Classification and Regression Tree*) [88], CART (*Classification and Regression Trees*) [119] e o algoritmo *Random Forest* [17], são amplamente utilizados para construir árvores de decisão. Eles empregam diferentes critérios de divisão, como ganho de informação [101], índice Gini [31] ou erro quadrático médio [48], para determinar qual atributo deve ser usado para dividir os dados em cada nó da árvore.

As AD oferecem algumas vantagens, como interpretabilidade, facilidade de uso e capacidade de lidar com dados categóricos e numéricos. No entanto, elas podem ser

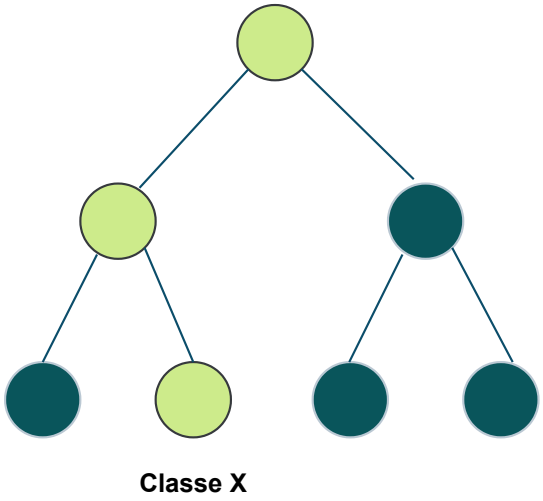

Figura 2.5: Exemplo do algoritmo de árvore de decisão.

suscetíveis a problemas como *overfitting* e instabilidade, especialmente quando a árvore se torna muito profunda ou quando os dados possuem ruído [90].

2.3.5 Florestas Aleatórias

As florestas aleatórias (FA), também conhecidas como *Random Forests* (RF), são um conjunto de árvores de decisão individuais combinadas para realizar tarefas de classificação, regressão e outras [17]. Essa abordagem utiliza a técnica de "*ensemble learning*" (aprendizado em conjunto) [36], que combina as previsões de vários modelos para obter um resultado mais robusto e geralmente mais preciso.

A Figura 2.6 apresenta um exemplo do funcionamento desse algoritmo. Ao contrário de uma única árvore de decisão, em uma floresta randômica, várias árvores são construídas a partir de diferentes subconjuntos de dados de treinamento, criados por amostragem aleatória com reposição (conhecida como *bootstrap*) [74]. Além disso, durante a construção de cada árvore, apenas um subconjunto aleatório de atributos é considerado para fazer as divisões nos nós. Essas duas fontes de aleatoriedade tornam as árvores de decisão individuais diferentes entre si, e a combinação de suas previsões resulta em uma estimativa mais robusta e geralmente mais precisa [80].

As FA oferecem algumas vantagens, como alta precisão, capacidade de lidar com grandes conjuntos de dados e resistência ao *overfitting*. Elas também podem fornecer informações sobre a importância relativa dos atributos na tarefa de previsão. No entanto, a construção de uma floresta randômica pode ser computacionalmente intensiva, especialmente quando o número de árvores e a dimensionalidade dos dados são altos [105].

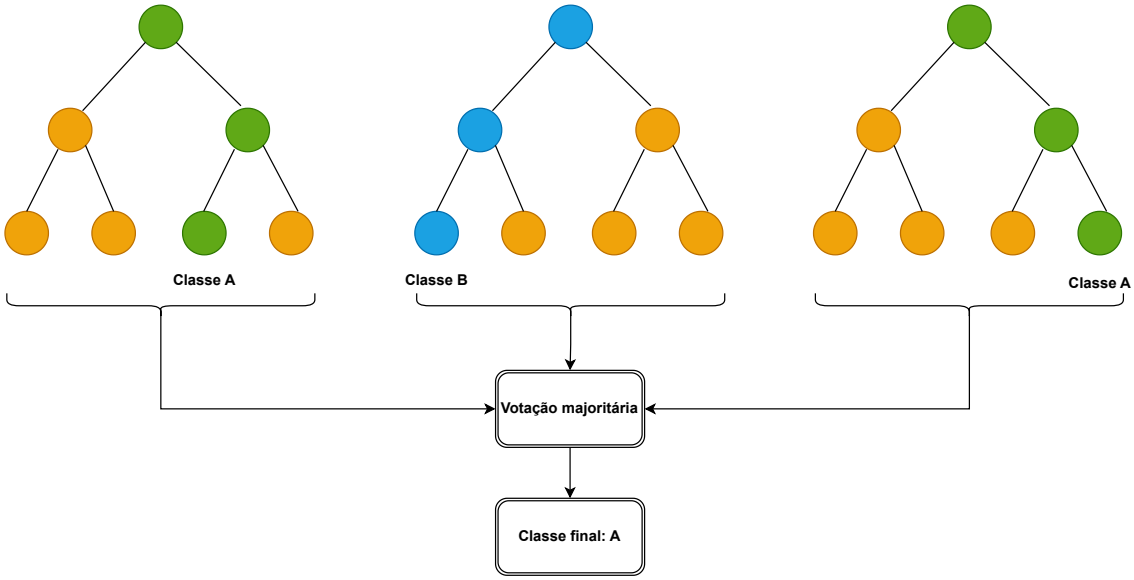

Figura 2.6: Exemplo do algoritmo de florestas aleatórias.

2.4 Avaliação de desempenho dos modelos de classificação

A avaliação de desempenho de um modelo de classificação faz com que o mesmo classifique algumas instâncias diferentes das que foram utilizadas no treino, de forma que ele possa informar a classe a qual pertencem essas instâncias. Existem diversas abordagens para avaliação de um modelo, mas existe duas que são mais utilizadas. Uma é através da divisão do conjunto de dados em treino, validação e teste e a outra é a utilização da validação cruzada com um determinado número de *folds* [58].

Na primeira técnica é feita uma divisão da base de dados de forma experimental. Os valores mais utilizados são 70-20-10, onde 70% dos dados são utilizados para treino, 20% dos dados utilizados para validação e 10% são utilizados para testagem do modelo. Para as etapas de validação e teste, o rótulo original não é apresentado para o modelo de forma que ele possa inferir e consequentemente possa ser medido o desempenho do mesmo, tendo em vista que o rótulo original é conhecido (partindo do pressuposto que está se trabalhando com o aprendizado supervisionado).

No caso da técnica de validação cruzada, esse método divide o conjunto de dados randomicamente em n partições (conhecidos como *folds*), proporcionalmente iguais, de forma que cada uma dessas partes tenha um valor de casos positivos proporcional ao conjunto de dados original. Esta técnica de particionar o conjunto de dados mantendo as proporções originais é conhecida como validação cruzada estratificada. Para realização de experimentos, combina-se $n-1$ partições para treinamento e utiliza-se o conjunto restante

como o conjunto de teste.

A partir das técnicas citadas anteriormente é possível calcular métricas que avaliam o desempenho como precisão, acurácia e AUC, as quais são descritas a seguir.

2.4.1 Métricas para avaliação de performance dos modelos

A curva ROC, que no inglês é conhecida como *Receiver Operating Characteristics* (ROC) é um método gráfico para avaliação, organização e seleção de sistemas de diagnóstico e/ou predição, segundo [97]. Essa técnica tem sido bastante utilizada nas áreas de aprendizagem de máquina e mineração de dados, para avaliação de métodos de classificação [16, 118]. A ideia inicial dessa medida era realizar avaliações sobre a qualidade de transmissão de sinais em um determinado canal [39]. Além dessa aplicação, essa curva também é bastante utilizada para avaliar a capacidade de um indivíduo de diferenciar um estímulo de um não estímulo [50]; dentro da área médica, serve para avaliar quão efetivo foi determinado teste clínico [152, 114]; e em climatologia, avaliando qualitativamente as predições de situações incomuns no clima [91].

Em um conjunto de dados binário uma determinada instância deverá ser rotulada de acordo com o conjunto ***p, n***, em outras palavras, ou com a classe positiva ou com a classe negativa. Porém, essa atribuição é feita de acordo com cálculos probabilísticos sobre o conjunto de treino. O ideal é que os dados sejam classificados corretamente, ou seja, que as instâncias positivas sejam classificadas como positivas e as instâncias negativas sejam classificadas como negativas, pelo algoritmo. Porém, existem casos em que uma instância que deveria ser classificada como positiva é classificada como negativa e vice-versa. A partir dessa observação, [43] verificou quatro possíveis formas de classificar uma instância:

- Caso uma instância positiva seja classificada como positiva, ela deverá ser chamada de instância verdadeira positiva (*true positive*);
- Se a instância positiva for classificada como negativa, ela é deverá ser chamada de instância falsa negativa (*false negative*);
- Caso uma instância negativa seja classificada como negativa, ela deverá ser chamada de instância verdadeira negativa (*true negative*);
- Se a instância negativa for classificada como positiva, ela deverá ser chamada de instância falsa positiva (*false positive*);

Desta forma, é possível montar um modelo conhecido por matriz de confusão, conforme apresentado na Tabela 2.2. A partir dessa Tabela, diversas métricas são derivadas, conforme apresentado nas Equações (2-8) a (2-13).

$$Precisão = \frac{verdadeiros\ positivos}{verdadeiros\ positivos + falsos\ positivos} \tag{2-8}$$

$$Acurácia = \frac{verdadeiros\ positivos + verdadeiros\ negativos}{total\ de\ casos} \tag{2-9}$$

$$Taxa\ de\ VP = \frac{verdadeiros\ positivos}{total\ de\ casos\ positivos} \tag{2-10}$$

$$Taxa\ de\ FP = \frac{falsos\ positivos}{total\ de\ casos\ negativos} \tag{2-11}$$

$$Sensibilidade = \frac{verdadeiros\ positivos}{verdadeiros\ positivos\ +\ falsos\ negativos} \tag{2-12}$$

$$Especificidade = \frac{verdadeiros\ negativos}{falsos\ positivos\ +\ verdadeiros\ negativos} \tag{2-13}$$

A Equação (2-8) e a Equação (2-9) representam duas medidas comumente utilizadas para avaliar desempenho de algoritmos, são elas a precisão e a acurácia. Por sua vez, a Equação (2-10) e a Equação (2-11) representam as taxas de verdadeiros positivos e falsos positivos, respectivamente. Por fim, a Equação (2-12) apresenta a medida de sensibilidade, que avaliar a capacidade do método em identificar corretamente casos classificados como positivo e a Equação (2-13) apresenta a especificidade, avalia a capacidade do método em identificar corretamente casos classificados como negativo.

Tabela 2.2: Representação da matriz de confusão.

		Valor predito	
		Sim	**Não**
Valor Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

As taxas calculadas a partir dos pontos existentes do plano cartesiano apresentado na Figura 2.7 são a probabilidade de classificar uma instância como sendo pertencente a classe positiva ou a classe negativa de forma correta. Existem pontos padrões neste plano que merecem alguns comentários, como os pontos (0,0) e (1,1). O primeiro representa um limiar de classificação em que todas as previsões de um determinado algoritmo são da classe negativa. O segundo ponto representa um limiar de classificação em que todas as previsões do modelo são positivas. Além desses, merecem destaque os pontos (1,0) e (0,1). Enquanto aquele representa um limiar de classificação em que todas as previsões do algoritmo são negativas (classe negativa), este representa um limiar de classificação em que todas as previsões do algoritmo foram realizadas corretamente, ou seja, todos

os verdadeiros positivos foram corretamente detectados e todos os falsos positivos foram corretamente detectados.

Além dos pontos, em específico citados anteriormente, existem características que podem ser inferidas dependendo do posicionamento dos pontos no plano, o que é apresentado na Figura 2.7. De acordo com [97], os modelos que se concentram no canto inferior esquerdo do plano (ponto **a** da Figura 2.7), possuem grande segurança na classificação e consequentemente cometem baixos erros, são conhecidos como conservativos. Os modelos mais próximos do canto superior direito (ponto **b** da Figura 2.7), responsáveis por predizer a classe positiva com maior frequência (porém, com altas taxas de erro) são considerados liberais. Em planos cartesianos *xy* para plotagem da curva ROC, geralmente se utiliza uma linha diagonal partindo da origem até o ponto máximo do plano, dividindo-o em dois triângulos. Os pontos que ficam sob essa linha, representam um desempenho aleatório. Aqueles que ficam acima da linha (no triângulo superior), possuem desempenho melhor que o aleatório.

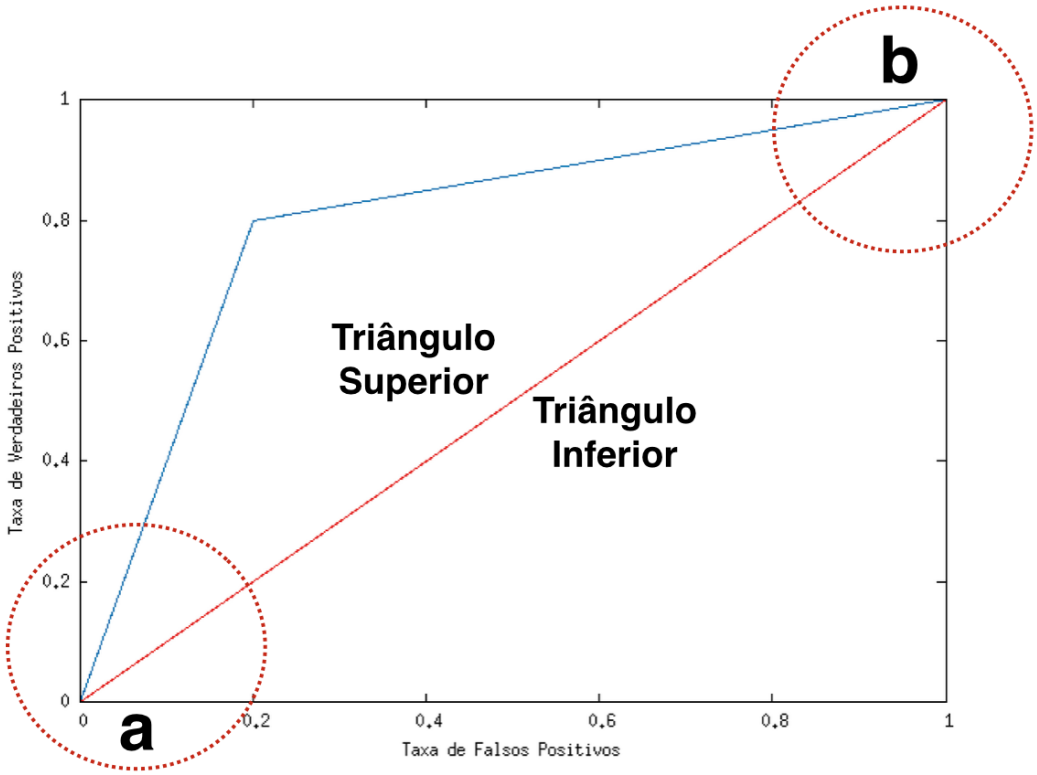

Figura 2.7: Interpretação de um gráfico com curva ROC

A partir de todas essas características e análises sob o plano da curva ROC (Figura 2.7), junto com a ideia de ordenação dos pontos, surge a área abaixo da curva ROC, o AUC (*Area Under the Curve*). O cálculo do AUC é numericamente equivalente a probabilidade de *Wilcoxon*, conforme [57], além de utilizar algumas propriedades do *Gini Index*, de acordo com [56]. Essa medida tem sido bastante utilizada como uma forma de

avaliar o potencial de métodos de aprendizado de máquina. Este aumento do uso do AUC se deve a sua confiabilidade e menor deficiência se comparada com a taxa de erro de classificação [76].

Os gráficos ROC são ferramentas muito úteis para visualização e avaliação de modelos de classificação. Possuem a característica de avaliar o aprendizado de um sistema ordenando exemplos, o que permite uma análise independente do limiar de classificação. Análises realizadas através dessa curva, fornecem uma avaliação mais rica, do que avaliar o modelo de classificação através de uma única medida [97].

2.5 Bases de dados *microarray*

Diversos problemas podem ser solucionados através do uso de algoritmos de aprendizagem. Áreas como finanças, indústria, educação e saúde têm investido bastante em inteligência artificial para automatização e previsibilidade de eventos que possam acontecer [68, 136, 104, 69, 89, 61, 148].

Especificamente na área médica, muitos problemas têm sido solucionados através do uso de aprendizado de máquina, como predição de doenças em imagens e identificação de doenças com base em dados clínicos [137, 134, 71]. Um dos dados que possibilitam a identificação de características importantes para predição de doenças são as bases de dados *microarray* [8].

Esse tipo de dados é utilizado para coletar informações a partir de amostras de tecidos e células com diferenças de expressões genéticas que podem ser úteis para o diagnóstico de doenças ou para distinguir determinados tipos de tumor. A classificação de dados de *microarray* é um desafio grande para pesquisadores da área de aprendizagem de máquina por conta da quantidade elevada de características e a pouca quantidade de amostras. Dessa forma, muitas dessas bases são utilizadas em algoritmos de seleção de características com o intuito de reduzir a complexidade da base e aumentar a acurácia dos algoritmos de classificação. Esse tipo de base de dados foi escolhido devido a característica de possuir uma grande quantidade de atributos e poucas instâncias, o que torna o processo de seleção de características mais desafiador e é uma boa forma de avaliar esse tipo algoritmo.

CAPÍTULO 3

Trabalhos relacionados

Este capítulo apresenta uma visão geral de como a seleção de características é aplicada a base de dados *microarray* por pesquisadores. A partir do levantamento de dados bibliométricos sobre o tema e aplicação da técnica de mineração de texto, foi possível identificar alguns padrões e potenciais problemas a serem explorados ao utilizar seleção de características aplicada a base de dados *microarray*.

3.1 Motivação

Existem estudos que utilizam seleção de características em bases de dados *microarray*, alguns fazem uma revisão de técnicas usadas [14], uso de métodos *ensemble* [146] e também os principais conjuntos de dados utilizados. [7]. Dentro da área médica, existem alguns estudos que foram desenvolvidos utilizando mineração de texto para extrair informações importantes dos dados dos pacientes, como dados de exames, laudos e resultados de tratamentos [123, 155]. Além disso, existem também publicações que empregam o uso de mineração de texto em dados *microarray*, seja para encontrar padrões nos conjuntos de dados como para analisar textualmente conjuntos de artigos completos [87, 102, 92, 42, 18].

No entanto, dentro das pesquisas que foram realizadas para elaboração desta tese, existe uma lacuna de estudos de mineração de texto aplicados a dados bibliométricos relacionados a seleção de características aplicados a conjuntos de dados *microarray* e também de dados bibliométricos relacionados a técnicas de seleção de características que usam AUC como estimador aplicados a conjuntos de dados *microarray*.

Desta forma, este capítulo apresenta duas caracterizações:

1. Utilização de mineração de texto para descobrir tendências em dados bibliométricos relacionados a seleção de características aplicados a dados *microarray* a partir de publicações ocorridas entre 2001 e 2022;
2. Utilização de mineração de texto para descobrir tendências em dados bibliométricos relacionados a técnicas de seleção de características que utilizam AUC como

estimador aplicados a dados *microarray* a partir de publicações ocorridas entre 1976 e 2022.

Em ambos os casos, foi utilizada a análise bibliométrica e de mineração de texto para obtenção de conhecimento sobre os artigos que foram selecionados relacionados aos dois conjuntos de dados bibliométricos coletados.

3.2 Metodologia

Ambos os estudos realizados e que são apresentados neste capítulo foram conduzidos seguindo os seguintes passos:

1. Elaboração da *string* de busca;
2. Seleção dos artigos;
3. Leitura e extração de dados;
4. Download dos metadados dos artigos;
5. Análise bibliométrica;
6. Mineração de texto;
7. Avaliação e discussão dos resultados.

As Seções 3.4 e 3.5 apresentam como foi a condução dos dois estudos de mineração de texto em dados bibliométricos. Dentro de cada Seção são descritas as fontes de dados, como foi realizada a coleta e ferramentas utilizadas tanto para a análise bibliométrica como para a mineração de texto.

3.3 Análise de dados

Ambas as análises de dados que são apresentadas foram realizadas com base em conceitos de diferentes áreas. A principal área que orientou o desenvolvimento das pesquisas foi a mineração de dados, associada ao uso da técnica de análise bibliométrica e mineração de texto.

A descoberta de conhecimento em bases de dados (*Knowledge Discovery in Database* - KDD) é um processo que envolve diversas etapas, incluindo a mineração de dados [30]. Essas etapas consistem na seleção dos dados, pré-processamento, transformação, mineração e avaliação dos resultados. Durante a mineração de dados, o pesquisador pode percorrer várias vezes essas etapas para identificar padrões úteis, com o objetivo de analisar os dados, organizá-los e planejar experimentos.

Neste contexto, são selecionados os artigos, os metadados foram armazenados e algumas outras informações com potencial relevância foram selecionadas para caracterizar as análises relacionadas a seleção de características (com e sem o foco em métodos

baseados em AUC) e dados *microarray*. Os dados foram extraídos a partir do *Web of Science* (WoS) e do *Scopus* e armazenados em arquivos ".bib". Toda a análise dos dados bem como geração de métricas, tabelas e gráficos foi feita através do software R. Na sequência, todo o processo de extração do conhecimento foi realizado através de algumas das técnicas de análise bibliométrica e mineração de texto.

3.3.1 Análise bibliométrica

A bibliometria é uma técnica que permite obter dados quantitativos sobre as publicações ao longo do tempo. A utilização da técnica de bibliometria em tópicos relacionados com mineração de dados e aprendizagem de máquina é bastante comum, uma vez que essa técnica permite a criação de visualizações e resumos das referências para facilitar a análise das publicações mais relevantes sobre um tema ou campo de estudo [38, 1].

Esse tipo de análise pode facilmente detectar temas emergentes em artigos e revistas onde os pesquisadores têm colaborado entre si e ajuda a explorar o conhecimento existente em determinadas áreas existentes na literatura [128, 37]. Isso pode ser feito através das técnicas existentes dentro da análise bibliométrica como por exemplo a análise de tendências, rede de co-ocorrência de palavras chave e análise de múltipla correspondência.

Dessa forma, são utilizados o pacote *bibliometrix* e a interface gráfica *biblioshiny* através da linguagem R e o *software* VOSViewer para realização das análises. Através dessas ferramentas é possível analisar informações relacionadas a quantidade de publicações por ano, universidades com maior quantidade de estudos publicados, estudos com maior quantidade de citações, colaboração entre países, *h-index*, rede de co-ocorrência de palavras chave, dentre outros. No entanto, foram utilizadas somente as técnicas que trazem algum tipo de conhecimento útil para os dois estudos relacionados a dados bibliométricos.

3.3.2 Mineração de texto

A mineração de texto é um processo de análise de dados que envolve a extração de informações úteis e relevantes a partir de grandes conjuntos de dados textuais. O objetivo da mineração de texto é identificar padrões e tendências em dados de texto, permitindo que pesquisadores possam obter *insights* valiosos sobre seus objetos de estudo [25]. Os passos para realização da mineração de texto incluem a organização, pré-processamento, extração de características e a avaliação dos resultados. A mineração de texto tem sido aplicada na análise de dados de redes sociais [131], na avaliação de informações produzidas por ambientes educacionais [45], na análise de dados biomédicos [24, 153], dentre outros.

A mineração de texto realizada nesse estudo foi aplicada nos campos de título, resumo e palavras-chave. Na fase de pré-processamento, foram realizadas as seguintes tarefas: transformação de letras maiúsculas em minúsculas, remoção de espaços em branco extras, remoção de pontuação, remoção de *stopwords* (exemplo: "*or*", "*by*", "*its*", "*how*", "*and*", "*as*", "*who*", dentre outras) e redução das palavras somente aos seus radicais [67]. Após a realização dessas etapas, foi utilizada a análise baseada na distância entre as frequências das palavras.

As informações contidas nos documentos foram dispostas no "*Bag of Words*", que compreende todos os documentos [23]. Para cada análise de texto (título, resumo, palavras-chave), uma matriz de termos de documento (MTD) foi criada. Cada linha representa um documento e cada coluna representa uma palavra, expressão ou frase. Em cada célula é registrada a frequência com que as palavras aparecem nos documentos.

Os dados já pré-processados foram introduzidos em uma função existente no pacote *tm* da linguagem R chamada *findAssocs* [99]. Essa função recebe como parâmetro uma MTD, os termos que deverão ser buscados dentro da matriz e um vetor com o valor limite de correlação (variando de 0 a 1) dos respectivos termos. Feito isso, a função retorna todos os termos encontrados na matriz que possuem valor menor ou igual ao limite de correlação definido para os termos que se deseja relacionar. Para melhor visualizar esses resultados é utilizado o pacote *igraph* da linguagem R [27] para formatar os termos encontrados com os respectivos graus de correlação. Essa representação facilita o entendimento de todo o processo de mineração de texto e simplifica a detecção de termos correlatos.

3.4 Análise de métodos de seleção de características aplicado a bases *microarray* usando dados bibliométricos

Nesta seção são apresentados os detalhes sobre a utilização de mineração de texto para descobrir as tendências da área de seleção de características aplicadas a bases de dados *microarray*.

3.4.1 Coleta dos dados

Foram selecionados os artigos científicos publicados no período entre 2001 e 2022 na base do *Web of Science* (*WoS*). O objetivo da análise desses artigos é utilizar a mineração de texto para descobrir as tendências da área de seleção de características aplicadas a dados *microarray* com base nos dados bibliométricos.

Os artigos foram selecionados através de uma *string* de busca combinando os termos "*feature select*, *variable select* e *microarray*". Essa busca foi realizada nos campos

de título, palavras chave e também nos resumos dos artigos. A Tabela 3.1 mostra como foi realizada a busca dentro do *Web of Science*. O resultado da busca original retornou 1479 artigos. Porém, após remover os estudos que não estavam na língua inglesa e remover os estudos que não foram publicados em revistas ou em anais de congressos, restaram 1448 artigos.

Tabela 3.1: *String* final usada para recuperar os dados a partir do *Web of Science*.

Passos	***String* de busca**	**Resultados**
1	***String* final de busca**: ((TI=(("feature select*"OR "variable select*") AND "microarray")) OR AB=(("feature select*"OR "variable select*") AND "microarray")) OR AK=(("feature select*"OR "variable select*") AND "microarray")	1479
2	Limitar somente aos estudos em inglês	1470
3	Limitar somente artigos de revistas e conferências	1448

3.4.2 Resultados

A base de dados final foi composta por 1448 documentos de 726 fontes diferentes. Dentre esses estudos, existem 3467 autores com um índice de colaboração de 2.39. Esse índice significa que existem 2.39 autores em média em cada artigo. A Figura 3.1 ilustra a distribuição dessas publicações entre 2001 e 2022.

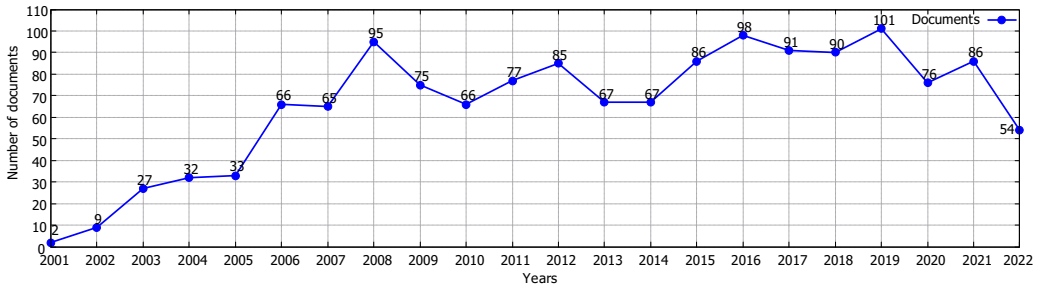

Figura 3.1: Crescimento de publicações de seleção de características e dados *microarray*.

Durante esse tempo a curva tem algumas oscilações. No entanto, entre 2001 e 2008, houve um crescimento de mais de 3000% no número de publicações. A partir de 2009, houveram algumas oscilações que ocasionaram em uma queda no número de publicações durante alguns anos. Houve uma nova tendência de crescimento de 2010 até 2015 em estudos relacionados a seleção de características aplicados a bases *microarray*, mas nada que superasse o que aconteceu no início dos anos 2000. Houve um novo crescimento entre 2014 e 2016, mesmo período em que houve uma popularização da aprendizagem de máquina e da inteligência artificial [93]. Essa tendência não seguiu em 2017 e 2018, com exceção do ano de 2019. De 2020 em diante houve uma grande queda

no número de publicações que pode ter uma relação direta com o cenário pandênmico da COVID-19 em todo o mundo.

A Figura 3.2 apresenta a rede de co-ocorrência de palavras chaves. Essa visualização apresenta os *clusters* encontrados dentro das palavras chaves do conjunto de artigos selecionados. Sobre essa rede, foram detectados três *clusters*: i) um na cor vermelha, ii) outro na cor azul e iii) um verde.

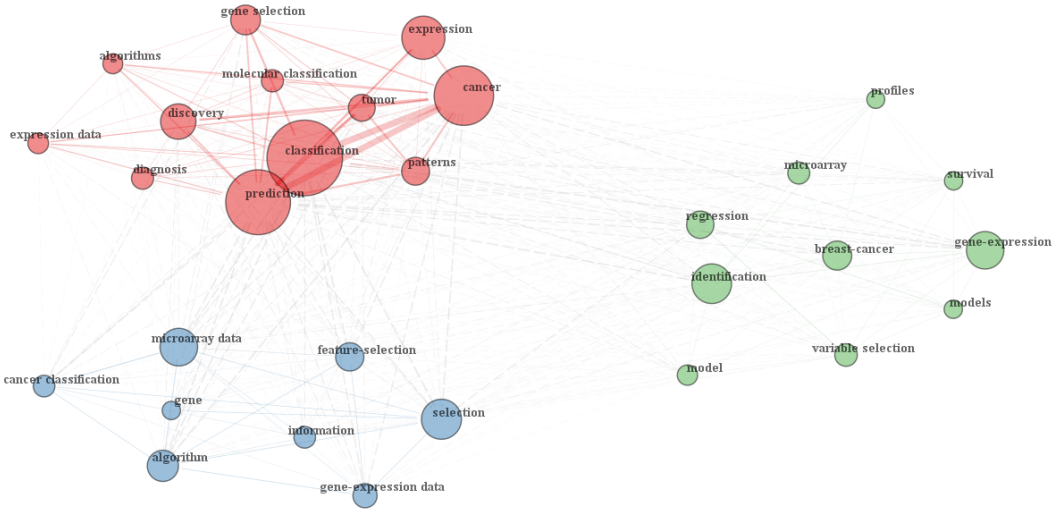

Figura 3.2: Rede de co-ocorrência de palavras-chave.

O *cluster* vermelho conecta tópicos relacionados a dados *microarray*, expressão gênica (*gene selection*) e doenças que podem ser diagnosticadas com esses dados, além de termos como algoritmos (*algorithms*), classificação (*classification*) e predição (*prediction*), que são relacionados a inteligência artificial e aprendizagem de máquina. O *cluster* azul mostra uma forte conexão de termos relacionados a bases de dados *microarray* com algoritmos e seleção de características (*feature selection*), ou seja, dois dos termos usados na *string* de busca. Por fim, o *cluster* verde cita alguns termos menos usuais dentre os existentes na literatura, como seleção de variáveis (*variable selection*) e regressão (*regression*) para resolver problemas em bases de dados *microarray*.

A análise de múltipla correspondência apresentada na Figura 3.3 mostra um detalhamento dos termos presentes dentro de cada um dos *clusters* apresentados na análise de co-ocorrência das palavras chaves. Em vermelho, podem ser vistos os termos mais relacionados a área médica ou biológica (*tumor, cancer, diagnosis, microarray* e outros) e alguns termos relacionados a área de aprendizagem de máquina (*feature selection,*

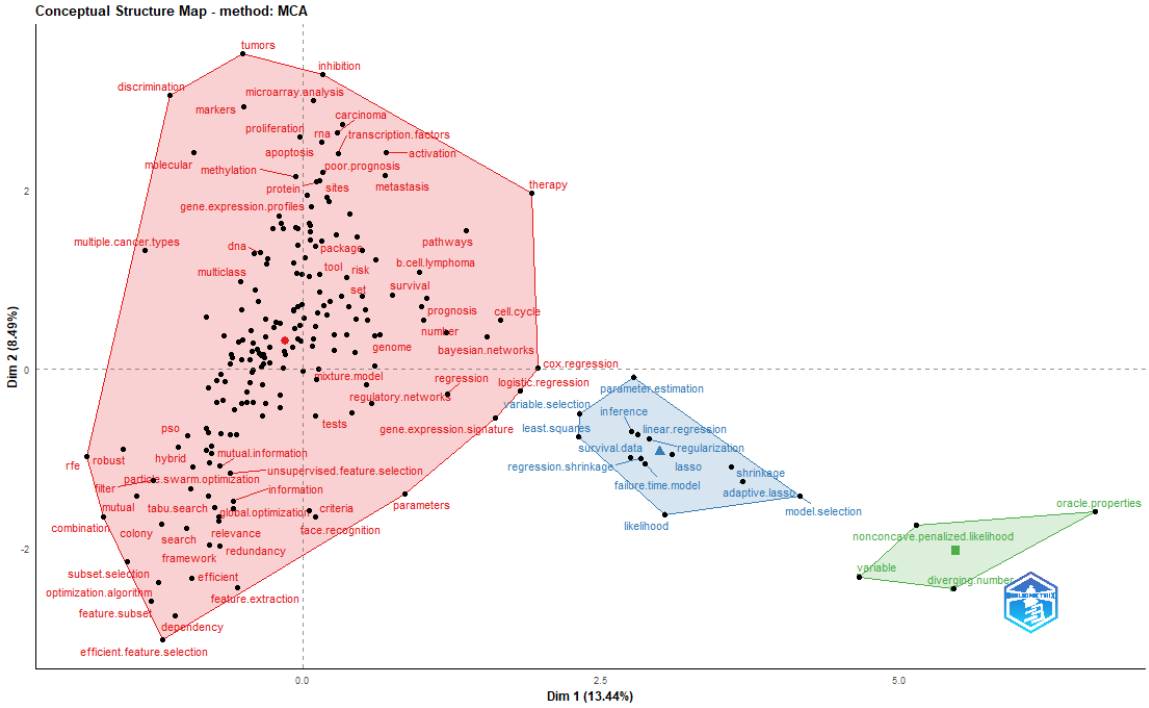

Figura 3.3: Análise de múltipla correspondência baseada nas palavras chaves *plus*.

SVM, RFE e outros). Na região em azul, a segunda maior, termos secundários a área de aprendizagem de máquina (dentro do contexto dos dados utilizados) tem um predomínio, como regressão (*regression*), seleção de variáveis (*variable selection*), dentre outros. Por fim, o *cluster* verde contém tópicos mais desconectados de toda a base extraída e todo o seu subconjunto de termos tem poucas conexões com os demais *clusters*, com exceção do termo "*variable*".

Outra análise foi realizada através da utilização do mapa temático que mostra algumas tendências em tópicos específicos dentro da base de dados analisada. A Figura 3.4 mostra um mapa que foi criado com base em bigramas dos termos para os três *clusters*. O *cluster* vermelho mostra que os termos "*feature select*", "*gene express*" e "*express data*" são associados com os quatro aspectos do mapa: temas emergentes ou em declínio, temas básicos, temas importantes (motores) e também os temas nicho. Esse resultado mostra que o tema de seleção de características em dados de expressão gênica é importante em todos os aspectos do mapa, uma vez que reduz a complexidade desses dados. O *cluster* azul mostra que os termos "*microarray data*", "*gene select*" e "*cancer classif*" foram classificados como temas básicos, o que significa que há uma ampla utilização de dados *microarray* para a detecção de câncer. Por fim, o *cluster* verde apresenta os temas nicho, onde há um relacionamento entre os termos "*variable select*", "*support vector*" e "*vector machin*", que mostra a força do algoritmo SVM para classificação junto com a técnica de seleção de características para redução de dimensionalidade, o que simplifica o processo

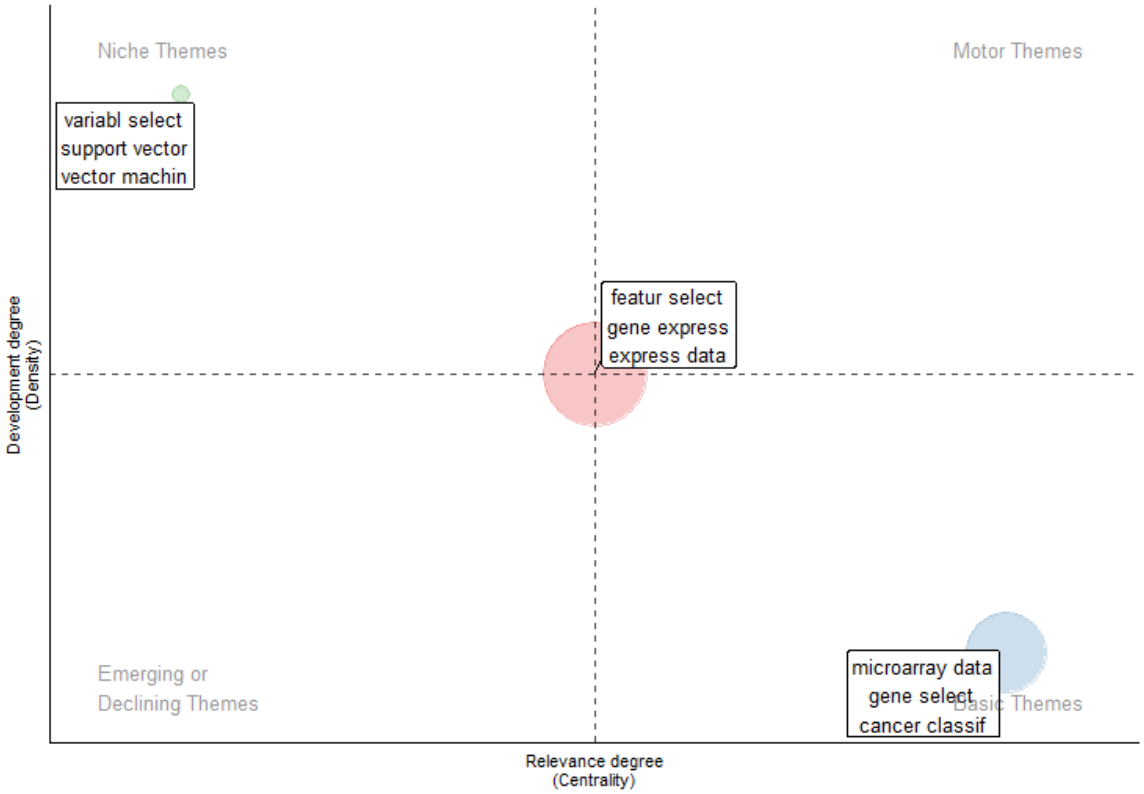

Figura 3.4: Mapa temático usando o título dos artigos com bigramas.

de classificação de bases de dados *microarray*.

3.4.3 Mineração de texto

Para a mineração de texto, foram utilizados termos únicos (unigramas) e compostos (bigramas) para procurar outros que possuam algum grau de correlação. Dois critérios de busca foram definidos para essa análise: buscar por "*cancer*" e "*feature selection*" separadamente e verificar o grau de correlação com outros termos dentro dos *abstracts* dos artigos selecionados. A Figura 3.5 mostra que as principais correlações para o termo "cancer" são com termos da área médica e biológica. Os termos encontrados, "*lung, breast* e *colon*", estão entre as bases de dados *microarray* de câncer mais estudadas. O artigo de revisão publicado por Alhenawi [5] confirma esse resultado encontrado pela mineração de texto.

A Figura 3.6 mostra as principais correlações ao termo composto "*feature selection*", que é um termo relacionado a área de aprendizagem de máquina. É possível inferir que o algoritmo SVM é um dos mais usados para classificar dados *microarray*, confirmando o que foi apresentado no mapa temático. No entanto, existe um alto grau de correlação do algoritmo SVM com o SVM-RFE, que aplica recursivamente a técnica de seleção de características antes do processo de classificação, o que pode levar a conclusão

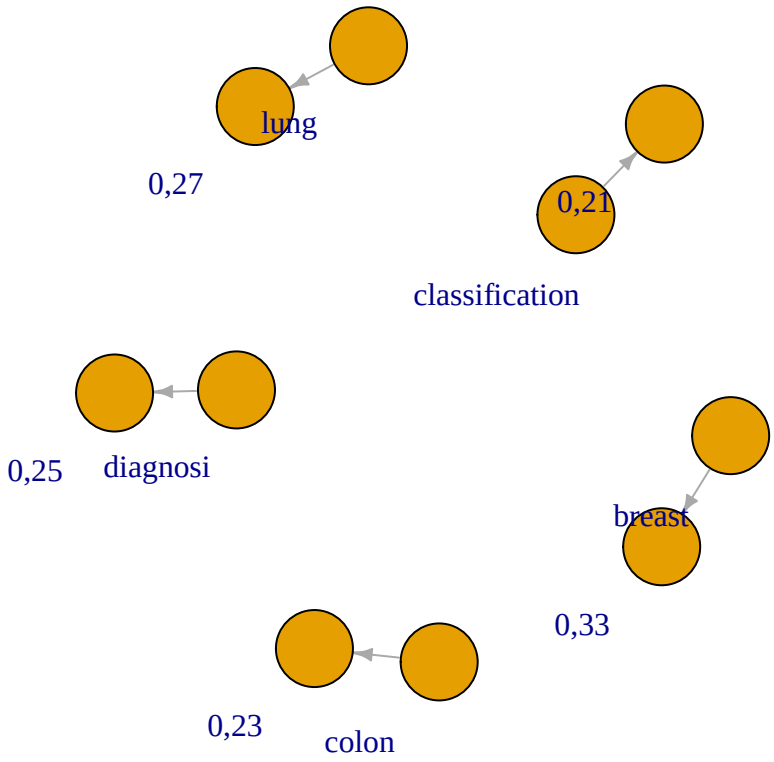

Figura 3.5: Principais correlações com o termo "cancer".

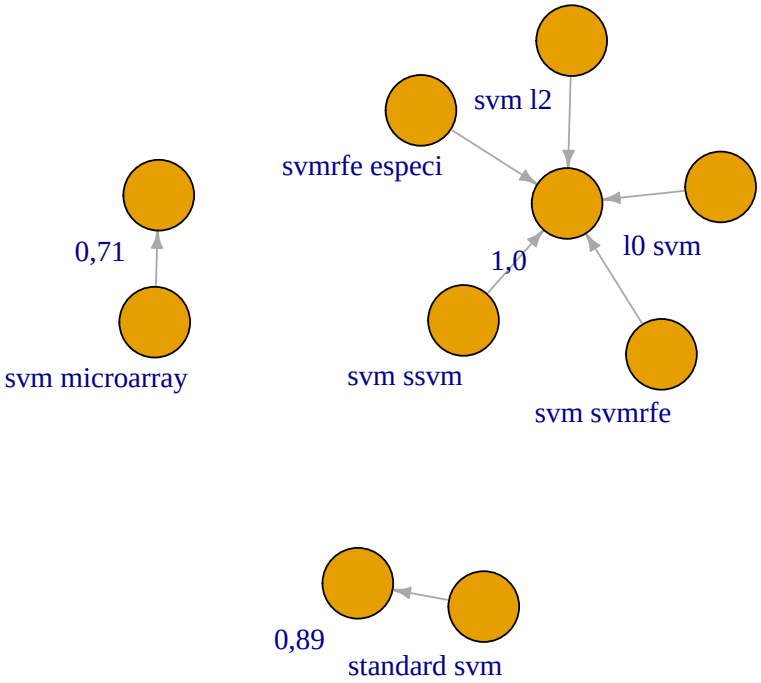

Figura 3.6: Principais correlações com o termo "feature selection".

que a seleção de características é de grande importância para reduzir a complexidade na análise de dados *microarray* [65].

3.5 Análise de métodos de seleção de características baseados em AUC aplicado a bases *microarray* usando dados bibliométricos

Nesta seção são apresentados os detalhes sobre a utilização de mineração de texto para descobrir as tendências da área de seleção de características, que utilizam AUC como estimador, aplicadas a bases de dados *microarray*.

3.5.1 Coleta dos dados

Foram selecionados os artigos científicos publicados no período entre 1976 e 2022 nas bases do *Web of Science* e *Scopus*. O objetivo da análise desses artigos é utilizar a mineração de texto para descobrir as tendências da área de seleção de características, principalmente nas técnicas que usam AUC como estimador, aplicadas a bases *microarray*, com base em dados bibliométricos. Esse estudo servirá como base para a proposta que é apresentada nesta tese de doutorado.

Os artigos foram selecionados através de uma *string* de busca combinando os termos "*auc feature select, auc variable select, auc, based* e *roc feature select*". Essa busca foi realizada nos campos de título, palavras chave e também nos resumos dos artigos. A Tabela 3.2 mostra como foram estruturadas as *strings* de busca tanto para a base do *Web of Science* como para a base do *Scopus*. O resultado da busca original retornou 1985 e 3738 artigos das bases do *WoS* e *Scopus*, respectivamente, totalizando 5723 estudos. Porém, após remover os estudos duplicados, remover os estudos que não estavam na língua inglesa e remover os estudos que não foram publicados em revistas ou em anais de congressos, restaram 3292 artigos.

Deve-se destacar que apesar dos esforços, nem todos os estudos retornados através da *string* de busca utilizam AUC como métrica para selecionar características. Embora tenha sido feita uma tentativa na construção de uma *string* de busca para limitar a seleção de artigos a esse conjunto. É uma limitação dessa análise não poder garantir isso de forma precisa, pois seria necessário analisar artigo por artigo para ter essa garantia, o que não é o objetivo desse estudo.

3.5.2 Resultados

A base de dados final inclui 3292 estudos publicados em 1316 fontes diferentes. Foram coletados todos os estudos existentes até o ano de 2022, ou seja, não foi delimitada uma data de início para a busca. Dessa forma, foram encontrados estudos desde 1976 até 2022. Dentre os 3292 artigos, existem 10998 autores com um índice de colaboração de

Tabela 3.2: *String* final de busca usada para recuperar dados das bases do *Web of Science* e *Scopus*.

Passos	***String* de busca**	**Resultados**
1	***String* de busca final (*WoS*)**: (**TI**=("auc feature select*"OR "auc variable select*"OR ("auc"AND "feature select*"AND "based") OR ("auc"AND "variable select*"AND "based") OR "roc feature select*"OR "roc variable select*"OR ("roc"and "feature select*") OR ("roc"and "variable select*") OR ("roc"AND "feature select*"AND "based") OR ("roc"AND "variable select*"AND "based")) OR **AB** = ("auc feature select*"OR "auc variable select*"OR ("auc"AND "feature select*"AND "based") OR ("auc"AND "variable select*"AND "based") OR "roc feature select*"OR "roc variable select*"OR ("roc"and "feature select*") OR ("roc"and "variable select*") OR ("roc"AND "feature select*"AND "based") OR ("roc"AND "variable select*"AND "based")) OR **AK**=("auc feature select*"OR "auc variable select*"OR ("auc"AND "feature select*"AND "based") OR ("auc"AND "variable select*"AND "based") OR "roc feature select*"OR "roc variable select*"OR ("roc"and "feature select*") OR ("roc"and "variable select*") OR ("roc"AND "feature select*"AND "based") OR ("roc"AND "variable select*"AND "based")))	1985
2	***String* de busca final (*Scopus*)**: TITLE-ABS-KEY ("auc feature select*"OR "auc variable select*"OR ("auc"AND "feature select*"AND "based") OR ("auc"AND "variable select*"AND "based") OR "roc feature select*"OR "roc variable select*"OR ("roc"AND "feature select*") OR ("roc"AND "variable select*") OR ("roc"AND "feature select*"AND "based") OR ("roc"AND "variable select*"AND "based"))	3738
3	Total de estudos nos passos 1 e 2	5723
4	Remoção de estudos duplicados	4157
5	Limitar somente aos estudos em inglês	4059
6	Limitar somente artigos de revistas e conferências	3292

3,34. A Figura 3.7 mostra o número de publicações realizadas entre 1976 e 2022. Apesar da busca ter sido limitada até 2022, foram retornados dois artigos de revista com data de publicação em 2023. Ambos foram considerados para esta análise.

Entre os anos de 1976 e 2011, o número de artigos publicados de acordo com

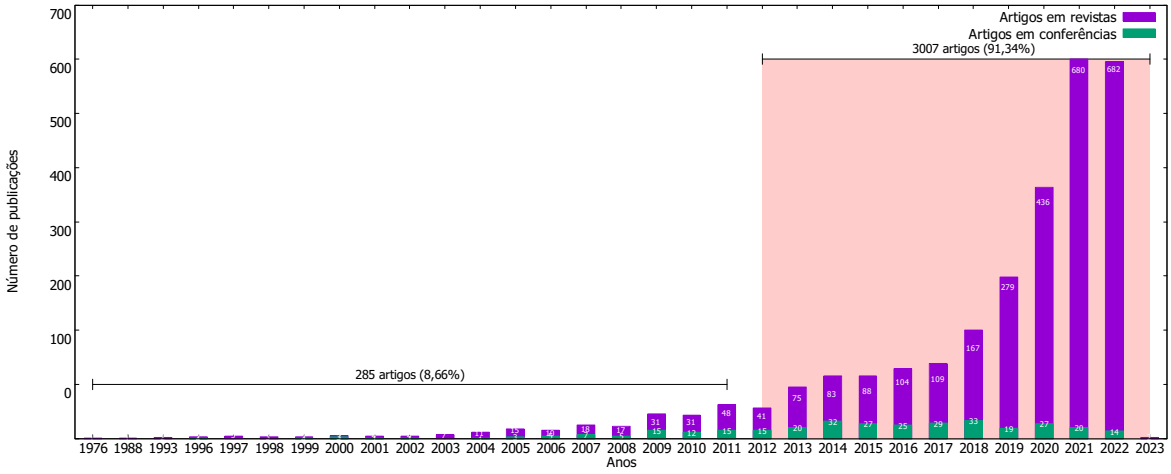

Figura 3.7: Crescimento de publicações relacionadas a métodos de seleção de características que usam AUC em bases *microarray*.

a *string* de busca foi de 285, o que representa 8,66% do total de estudos. De 2012 em diante, existem 3007 estudos, o que representa 91,34% do total de estudos. Na primeira divisão pode-se notar que o número de estudos segue em um padrão baixo até 2008. De 2008 para 2009 há um crescimento de pouco mais que o dobro do número de publicações (109,09%). Depois desse salto, os demais crescimentos foram todos na faixa dos 30%, com exceção do período entre 2019 e 2021, quando o número de estudos cresceu 49% (2018 - 2019), 55,36% (2019 - 2020) e 51,18% (2020 - 2021). Os anos 2021 e 2022 foram os anos com o maior número de publicações, ambos passando de 650 publicações.

A Figura 3.8 apresenta a rede de co-ocorrência de palavras chaves, que torna possível a visualização dos *clusters* que existem na literatura que está sendo analisada. Através dela é possível notar que as palavras mais frequentes no *cluster* vermelho estão relacionadas a métricas de análise de algoritmos, técnicas de validação clínica e alguns tópicos mais relacionados a algoritmos, aprendizagem de máquina e inteligência artificial como valor de predição, extração de características e *random forest*. O *cluster* azul mostra uma ligação forte entre termos mais relacionados com o campo da aprendizagem de máquina, algoritmos (especialmente SVM), curva ROC, AUC e termos relacionados com a predição e diagnóstico.

A Figura 3.9 apresenta análise de múltipla correspondência (AMC), de forma complementar a rede de co-ocorrência de palavras chaves. No *cluster* em vermelho, existem muitos termos relacionados a área médica e biológica (*lung, diabetes* e *pancreatic neoplasms*, além de alguns termos relacionados a aprendizagem de máquina e métricas para avaliação dessas técnicas (SVM, *Ada boost, random forest, artificial neural networks*, ROC, *accuracy* e outros). No *cluster* azul, termos relacionados a aprendizagem de máquina predominam, por exemplo *prediction models, classification algorithms, prediction*

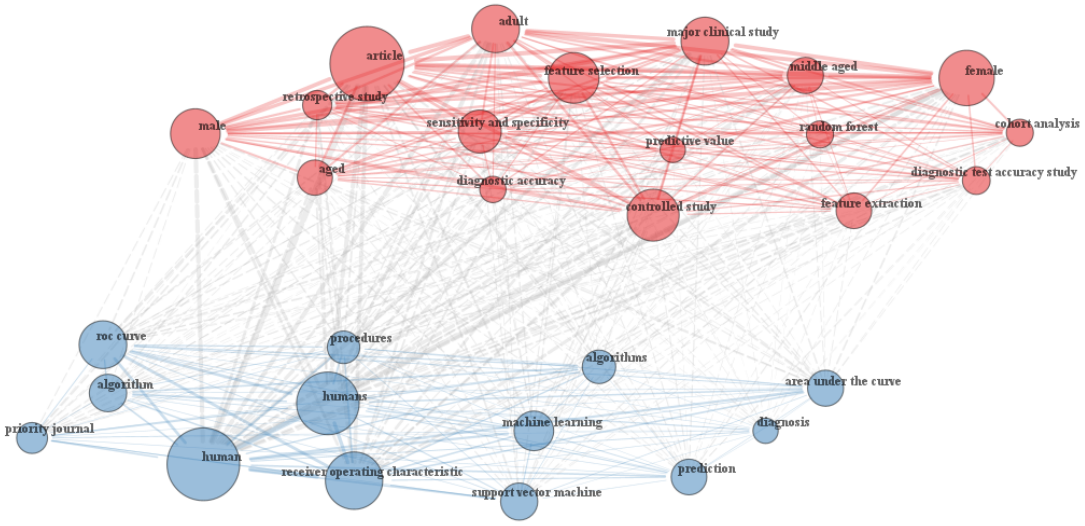

Figura 3.8: Rede de co-ocorrência de palavras chaves.

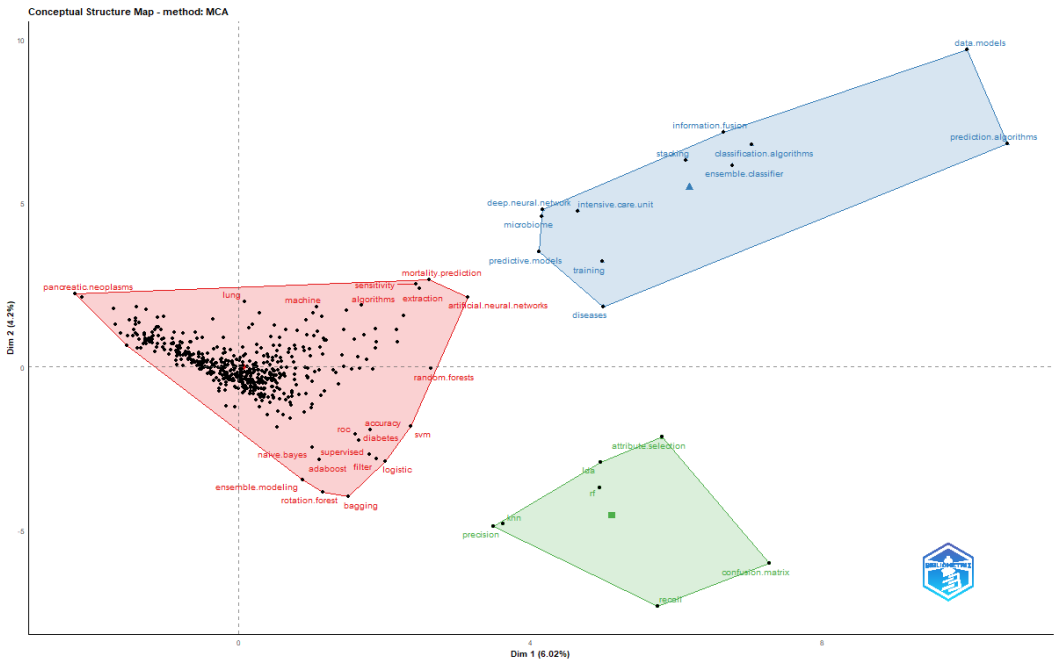

Figura 3.9: Análise de múltipla correspondência baseada nas palavras chaves *plus*.

algorithms e outros. Por fim, diferentemente da rede de co-ocorrência de palavras chave, o AMC mostra um *cluster* verde que é muito menor que todos os citados anteriormente. Ele contém sete termos, quatro relacionados a algoritmos e aprendizagem de máquina (*attribute selection*, LDA, RF e kNN) e os outros três são relacionados a métricas e avaliação

de modelos (*precision, confusion matrix* e *recall*).

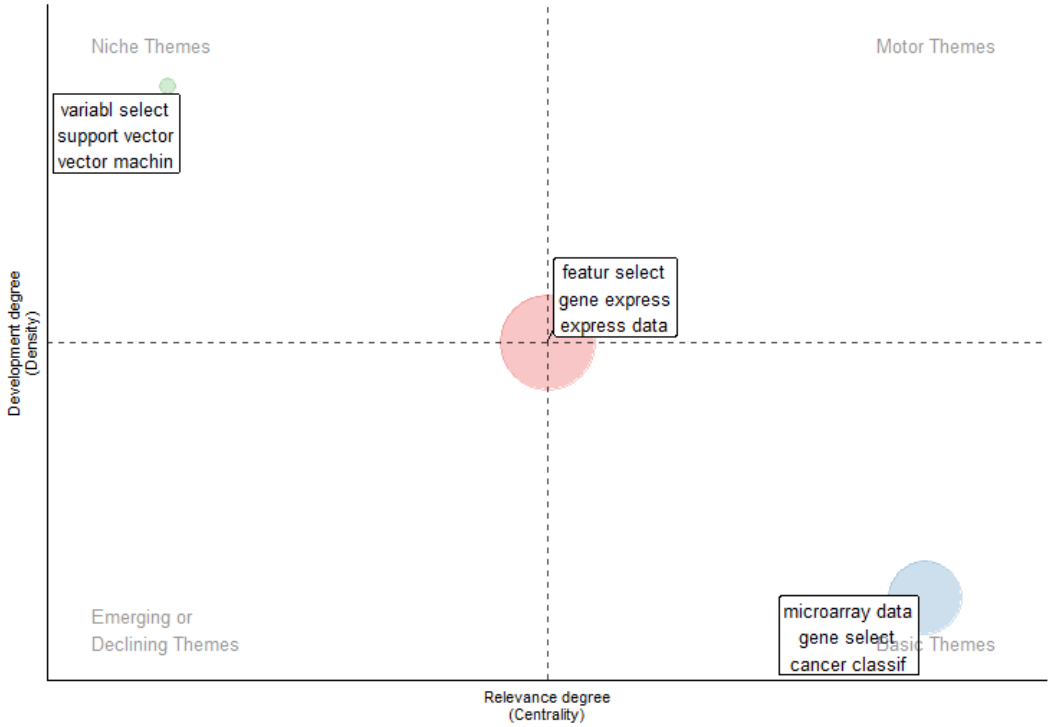

Figura 3.10: Mapa temático usando o título dos papers com bigramas.

Para visualizar melhor temas emergentes que possam existir em uma base bibliométrica, utiliza-se o mapa temático. A Figura 3.10 mostra um mapa temático criado com bigramas que gerou quatro *clusters*. O *cluster* vermelho mostra que os termos "*machine learn, predict model* e *learn algorithm*"possuem forte conexão com aspectos de tópicos emergentes ou em declínio. Isso mostra que termos relacionados a aprendizagem de máquina, modelos de predição e algoritmos pode estar em crescimento ou em declínio quando usado para representar técnicas de seleção de características e AUC. O *cluster* verde contém termos como "*featur select, neural network* e *support vector*", que são também relacionados a temas emergentes ou em declínio, mas dado o tamanho do *cluster* e a proximidade com a região central do mapa, esses termos tem um alto grau de relevância. Além disso, o fato desses termos estarem juntos no mesmo *cluster* mostra que a seleção de características é importante quando se utiliza modelos de aprendizagem de máquina e inteligência artificial como SVM e redes neurais.

Seguindo, o *cluster* azul contém os termos "*breast cancer, radiom analysis* e *radiom model*", que estão posicionados como temas motores e o *cluster* amarelo, que é caracterizado pelos termos "*radiom featur, lung cancer* e *cell carcinoma*". O *cluster* azul mostra um relacionamento entre a detecção de câncer de mama, padrões radiológicos e análise radiográfica. O *cluster* amarelo, por outro lado, mostra o relacionamento entre câncer de pulmão com características radiológicas e câncer de pele. Os relacionamentos

entre esses tipos de câncer com radiografias de tórax em ambos os *clusters* é normal, considerando que esse tipo de exame pode ajudar a detectar lesões indicativa dessas doenças.

3.5.3 Mineração de texto

Da mesma forma que o estudo citado na Seção 3.4, durante a mineração de texto foram utilizados termos únicos (unigramas) e compostos (bigramas) para descobrir outros termos com algum grau de correlação. Três critérios de busca foram definidos para essa análise: *cancer, AUC* e *feature selection*, buscando encontrar algum grau de associação com outros termos dentro dos *abstracts*.

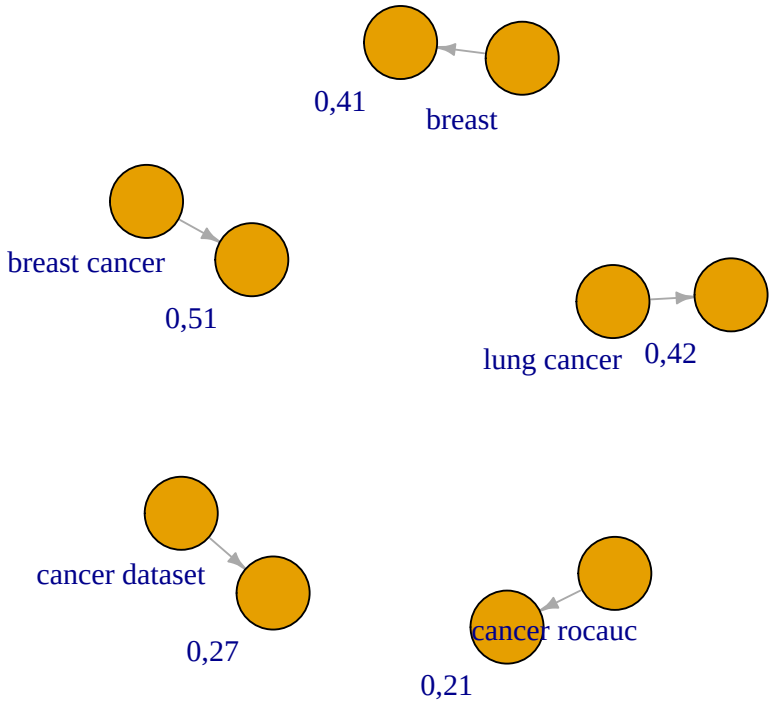

Figura 3.11: Principais correlações com o termo "cancer".

A Figura 3.11 mostra algumas das correlações existentes com o termo "*cancer*". Como pode ser visto, esse termo possui um alto grau de correlação com termos como "*breast cancer, lung cancer* e *cancer dataset*", que são referentes a bases de dados utilizadas para identificar essas doenças, como as bases de dados *microarray*. Outros tópicos com baixo grau de correlação também podem ser destacados, como "*cancer rocauc, pathway featur select, curv rocauc auc*" e "*featur select breast*", por exemplo. Esses termos com baixo grau de correlação comparados com os termos já mencionados destacam possíveis oportunidades para o desenvolvimento de novas pesquisas e soluções.

A Figura 3.12 mostra a importante correlação existente com o termo AUC. Como pode ser visto, há um alto grau de correlação com um sinônimo chamado *curve*

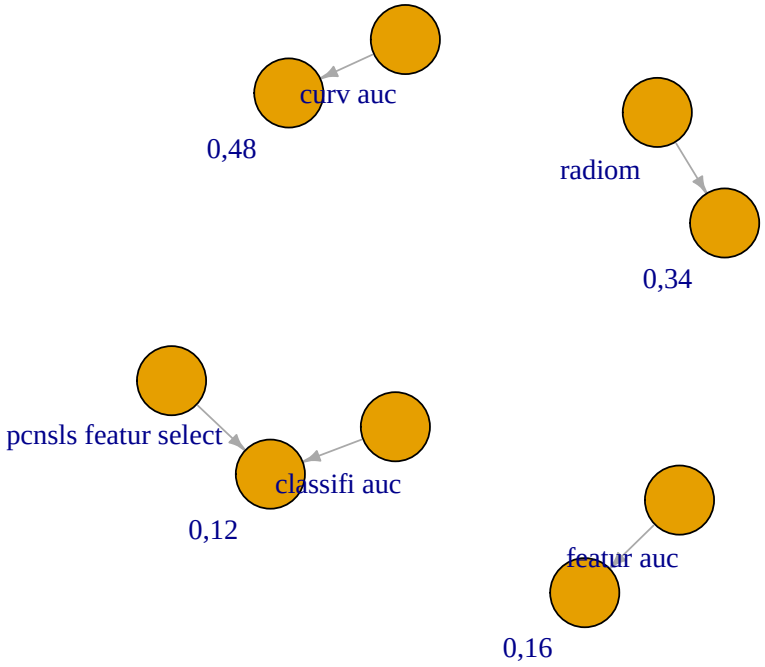

Figura 3.12: Principais correlações do termo "auc".

AUC e também com o termo "*radiom*", que vem da palavra *radiomics* e refere-se a padrões radiológicos. Existem outros termos com um grau de correlação mais baixo mas que podem ser destacados, como os termos "*featur auc*" e "*pcnsls featur select*" que são relacionados a seleção de características usando AUC [122] e no uso da seleção de características em uma base de dados *microarray* chamada *CNS Lymphoma* [98], respectivamente.

A Figura 3.13 apresenta algumas das correlações mais significativas encontradas ao pesquisar pelo termo "*auc select*". Entre os vários termos identificados, aqueles que apresentaram maior correlação (0,38) e foram considerados mais relevantes para o contexto do estudo foram: "*learn classifierssupport*", "*better features*", "*features knowledge*", "*svm gdt*" e "*svm combinatorial*". Os três primeiros termos mencionados estão relacionados ao fato de que o processo de seleção de características visa identificar as características mais relevantes de um conjunto de dados. Esse processo favorece a tarefa de classificação, o que facilita o processo de aprendizado [28, 2].

O termo "*svm gdt*" é uma abreviação de "SVM gradiente" e se refere a trabalhos que realizam análises em bases de dados de microarranjos usando o classificador SVM e suas variações, bem como algoritmos baseados em gradientes descendentes [12, 78]. Por fim, o termo "*svm combinatorial*" apresenta duas possibilidades dentro do conjunto de trabalhos analisados: a utilização do SVM em conjunto com outros classificadores para construir um *ensemble* que realize o processo de aprendizagem [142]; e a utilização do SVM combinado a algum método de seleção de características para identificar os

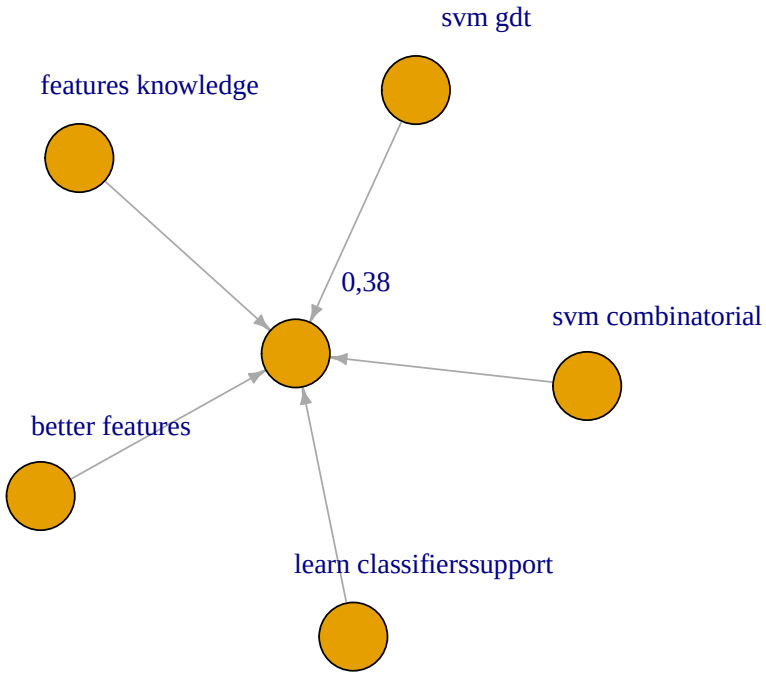

Figura 3.13: Principais correlações do termo "auc select".

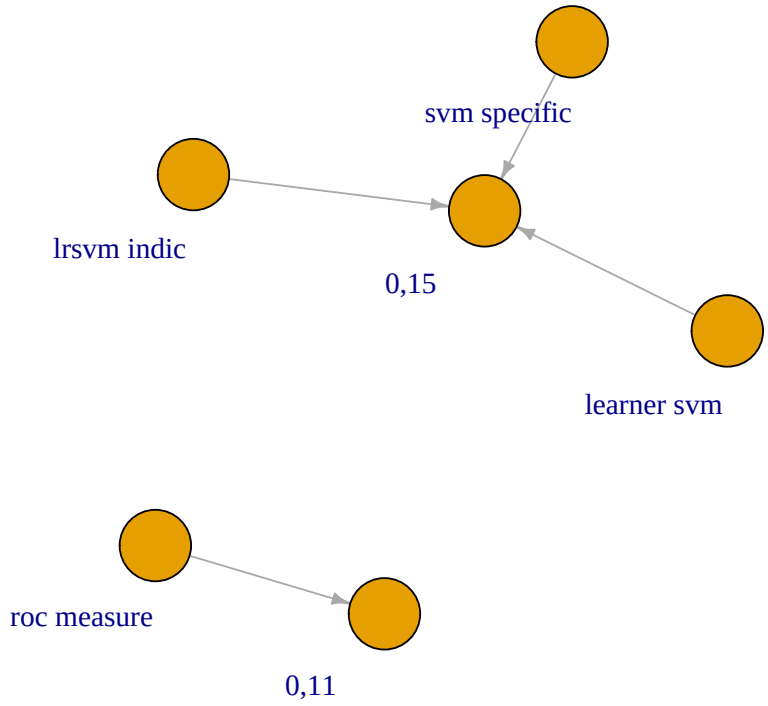

Figura 3.14: Principais correlações do termo "feature selection".

melhores genes e, dessa forma, facilitar a tarefa de classificação das amostras [147].

Por fim, a Figura 3.14 mostra as principais correlações existentes com "*feature selection*". Com base nela, é possível deduzir que o processo de seleção de características é relacionado ao processo de classificação, incluindo o algoritmo SVM, que confirma o que foi representado no mapa temático. A respeito do SVM, o termo "*svm specific*", pode

estar relacionado as variações do algoritmo SVM ou suas especificidades [19, 113]. Um outro termo é "*roc measure*", que está relacionado a seleção de características de duas formas: no uso do AUC como métrica para avaliação de algoritmos [143, 145, 40] e como uma métrica para selecionar os melhores atributos dentro do processo de seleção de características, como é o caso do algoritmo *Feature Assessment by Sliding Theresholds* (FAST) [140] (que não aparece nos dados bibliométricos por ser um trabalho de conferência não-indexado) e o algoritmo *AUC and Rank Correlation coefficient Optimization* (ARCO) [135]. Essa identificação realizada com a mineração de texto coincide com alguns dos resultados apresentados pelo AMC na Figura 3.9.

3.5.4 Análise LDA

Uma outra análise realizada foi a *Latent Dirichlet Allocation* (LDA). Esse método trata cada texto usado como entrada como uma mistura de tópicos e cada um deles como um conjunto de palavras. Com isso, é possível que os textos se sobreponham em termos de conteúdos ao invés de serem separados em grupos distintos [109, 120]. Além disso, a LDA pode ser utilizada como uma análise conjunta com a análise de múltipla correspondência (AMC) [116].

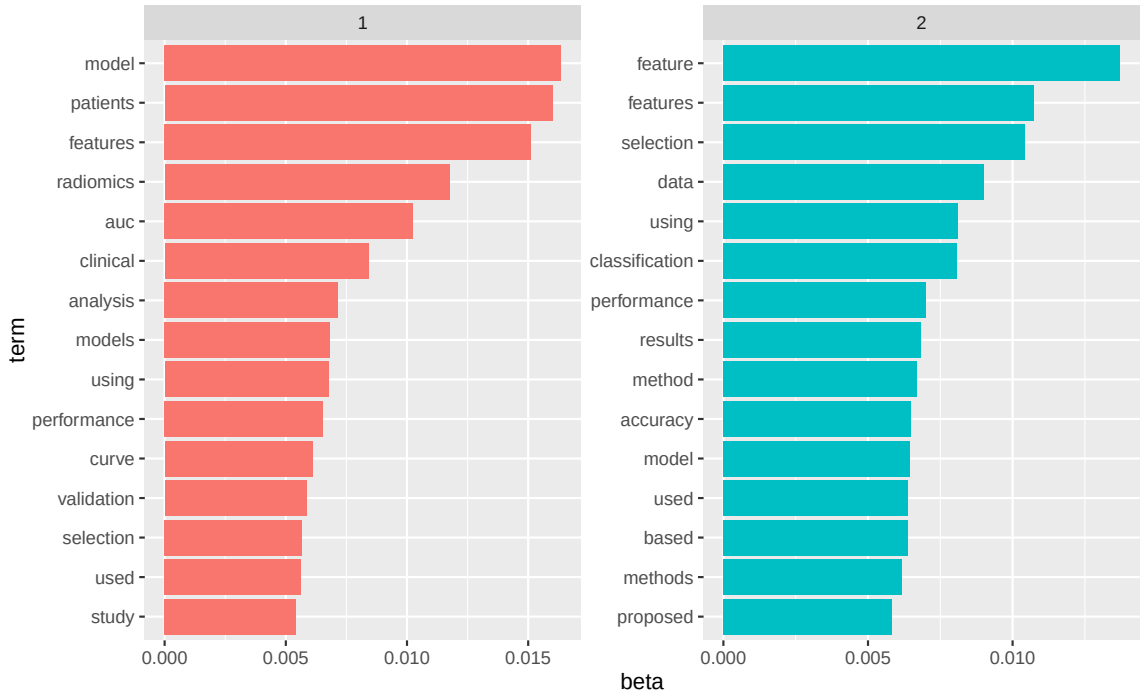

Figura 3.15: Principais termos encontrados nos tópicos definidos pela análise LDA.

A Figura 3.15 mostra os tópicos que foram criados com base na análise textual dos *abstracts*. Ambos os tópicos mostram termos relacionados ao aprendizado de má-

quina, seleção de características e algumas métricas de avaliação de performance. Porém, é possível notar que no tópico em vermelho existem alguns termos relacionados à área médica, como "*clinical*", "*validation*" e "*patients*", enquanto que o tópico em azul possui termos quase que em sua totalidade relacionados ao aprendizado de máquina ("*features*"; "classificação"), métricas de avaliação ("*accuracy*") e termos gerais da área ("*proposed*"; "*model*"; "performance"). Esse resultado da análise de LDA coincide com o resultado do AMC apresentado na Figura 3.9.

Essa procedimento foi realizado somente para a análise dos métodos de seleção de características baseado em AUC devido o foco ser menos genérico que a caracterização apresentada na Seção 3.4.

3.6 Discussão

Os resultados dos estudos apresentados nesse capítulo mostram as tendências existentes na área de seleção de características aplicadas a bases de dados *microarray* utilizando dados bibliométricos e a técnica de mineração de texto. O estudo apresentado na Seção 3.4 mostra através da técnica de mineração de texto e análises bibliométricas, que bases de dados *microarray* relacionadas a câncer têm sido usadas e são um tópico em evidência dentro da área de seleção de características, pois as bases citadas são usadas para diagnóstico de alguns tipos de câncer, como por exemplo câncer de cólon, pulmão e de mama. Uma investigação profunda dos tipos de câncer, a disponibilidade das bases de dados *microarray* e seus achados são necessárias para estabelecer o estado da arte sobre este tema.

Por outro lado, foi encontrado através da mineração de texto e da análise bibliométrica que o algoritmo SVM tem sido bastante utilizado para seleção de características e para a classificação de bases de dados *microarray*. No entanto, mais estudos são necessários para investigar a "família" dos algoritmos SVM e como eles podem ser usados para melhorar as tarefas de classificação e regressão. Apesar da força do uso de seleção de características para redução de dimensionalidade de dados *microarray*, os seletores de características também podem ser usados para identificar os melhores preditores para um diagnóstico de câncer. O SVM-RFE foi uma das técnicas identificadas nos dados bibliométricos durante a fase de mineração de texto. Algoritmos como FAST, ARCO e *Particle Swarm Optimization* (PSO) [4] podem não aparecer nas avaliações bibliométricas ou de mineração de texto por não serem tão proeminentes como o SVM-RFE. No entanto, mais estudos são necessários para investigar a performance e aplicação de diferentes tipos de técnicas de seleção de características a dados *microarray*.

O estudo apresentado na Seção 3.5 tentou ser um pouco mais específico. Utilizando também as técnicas de bibliometria e mineração de texto, a ideia era descobrir

as tendências em técnicas de seleção de características que usam AUC como estimador aplicadas a bases *microarray* usando dados bibliométricos. Da mesma forma que para o primeiro estudo, esse também mostrou que bases de dados *microarray* têm sido um tópico bastante relevante dentro dos métodos de seleção de características que usam AUC. Além disso, também foi identificado uma certa frequência na utilização de algoritmos da "família" do SVM para realização das atividades de classificação e regressão. As observações feitas para o estudo anterior relacionadas ao aprofundamento das pesquisas servem para esta. O destaque do segundo estudo acaba sendo que apesar dos dados bibliométricos terem um foco diferente (técnicas de seleção de características baseadas em AUC aplicadas a dados *microarray*), as conclusões são similares ao estudo anterior.

Em relação a caracterização de alguns estudos relacionados a seleção de características baseados em AUC, foram apresentadas as principais características deles e destacados as principais vantagens e desvantagens de cada um. No entanto, dentro dos dados bibliométricos somente os algoritmos ARCO e AVC foram encontrados, pois o FAST e o FROC foram publicados em conferência e revista não-indexados, respectivamente.

3.7 Considerações finais

Este capítulo apresentou um contexto geral e tendências relacionadas a técnicas de seleção de características em geral e as que usam AUC como estimador aplicadas a bases *microarray*. No primeiro estudo, foram analisados dados de 2001 a 2022 e no segundo o período analisado foi de 1976 a 2022. A aplicação da técnica de mineração de texto mostrou que há uma convergência dentro da área de seleção de características aplicados a dados *microarray*, na qual a maioria dos trabalhos utilizam algoritmos da família SVM aplicados da bases *microarray* relacionadas a câncer.

Estudos que utilizam a medida original do AUC combinada a outras técnicas são minoria dentro dos métodos de seleção de características. Além disso, esses estudos não levam em consideração as possíveis variações de valores que as características podem ter em relação aos rótulos de classe, o que pode levar a um valor errôneo de importância das características.

Baseado nessa lacuna e na possibilidade de utilizar AUC combinado com técnicas que podem aprimorar o cálculo do valor da área abaixo da curva ROC, o próximo capítulo apresenta uma abordagem que combina AUC com a técnica de estimativa de probabilidade e o método de correção de *LaPlace* para realização da tarefa de seleção de características.

CAPÍTULO 4

Proposta

A combinação da técnica de estimativa de probabilidade com o cálculo do AUC podem ser utilizadas para o desenvolvimento de modelos de classificação, e capazes de obter boa performance, conforme ocorrido com o algoritmo de *paths* de decisão [129] e do algoritmo de *paths* de decisão específico para cada paciente [107]. Devido a isso, levantou-se a hipótese de utilizar o AUC combinado a estimativa de probabilidade associado a técnica de suavização de *La Place smoothing* para realizar a tarefa de seleção de características.

4.1 Método proposto

Neste estudo, investiga-se a combinação do cálculo de AUC com a estimativa de probabilidade e o método de suavização de *La Place* (também conhecido como *smoothing*). Ao aplicar essa abordagem para cada valor possível de uma característica, há a possibilidade de aprimorar os resultados dos métodos tradicionais de seleção de características que se baseiam apenas na quantidade total de instâncias por característica em cada classe para calcular sua relevância.

Com tal objetivo, foi proposto o algoritmo de seleção de características baseado em AUC com estimativa de probabilidade e o método de *smoothing*, denominado AUC-EPS. A área abaixo da curva ROC (AUC) é amplamente utilizada para avaliar o desempenho de classificadores supervisionados, principalmente devido a característica de não ter a obrigação de especificar os custos de classificações incorretas. No entanto, esta métrica também pode ser utilizada no processo de seleção de características. Um fluxograma de como funciona o algoritmo AUC-EPS é apresentado na Figura 4.1.

Além da implementação do AUC-EPS, existe uma motivação em analisar o *ranking* das 100 características escolhidas pelo método proposto, com o objetivo de comparar com as 100 características selecionadas pelas abordagens tradicionais utilizadas nos experimentos, verificando quais são as características que foram escolhidas por cada método, quais foram iguais, diferentes e estabelecer alguma relação, com base nessas análises.

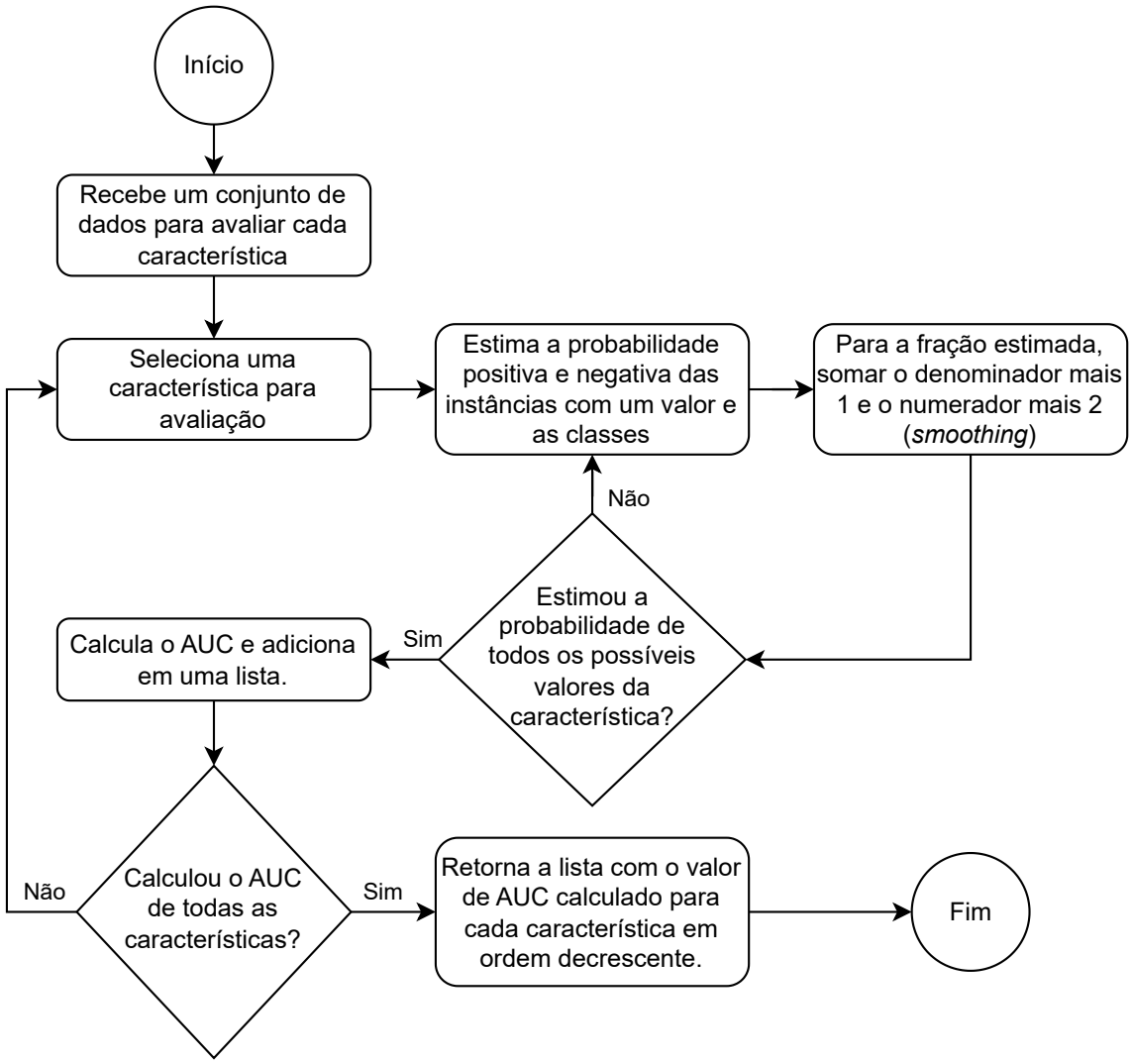

Figura 4.1: Fluxograma que apresenta o funcionamento do algoritmo proposto (AUC-EPS).

O algoritmo desenvolvido recebe como parâmetro o conjunto de dados e uma lista com as *labels* de cada uma das instâncias da base de dados. O primeiro passo é identificar todos os possíveis valores para a primeira característica (f_0) da base de dados. Com a aquisição de todos os possíveis valores para a característica f_0, esse método seleciona o primeiro possível valor dentre o conjunto de valores e na sequência o algoritmo seleciona todas as instâncias que possuem o valor selecionado. O próximo passo será particionar a base de dados em duas: uma partição irá conter todas as instâncias com valor igual ao valor selecionado para a característica f_0 e a outra partição irá conter todas as instâncias que possuem valores diferentes ao valor selecionado para a variável que está sendo analisada. Com base na primeira partição, a estimativa de probabilidade [90] é calculada para as instâncias com o mesmo valor de acordo com a Equação (4-1),

$$E_p = \left(\frac{E_-}{I}, \frac{E_+}{I} \right) \tag{4-1}$$

em que E_- é a quantidade de instâncias com o valor selecionado para a característica f_0 com a classe negativa e E_+ é a quantidade de instâncias com o valor selecionado para a característica f_0 com a classe positiva e I é o tamanho da partição, ou seja, $I = E_- + E_+$.

Para evitar o *overfitting*, antes do valor final do cálculo destas probabilidades o *LaPlace smoothing* [149] é aplicado, o que resulta em somar cada estimativa encontrada em seu numerador ao valor um e no denominador ao valor dois. Esse processo todo será repetido até que todos os possíveis valores da característica f_0 tenham sido selecionados e todas as instâncias tenham seus pares de probabilidades (instância positiva e negativa) calculados.

Com o fim da fase de estimativa de probabilidade, o próximo passo é calcular o AUC da característica f_0. A área abaixo da curva ROC é feita através da função *roc_auc_score* disponível na biblioteca *Python Sklearn* [94]. Os passos da estimativa de probabilidade e cálculo do AUC são repetidos até que todas as características da base de dados $(f_0, f_1, f_2, ..., f_{n-1})$, em que n é a quantidade de características, tenham o valor de AUC calculado e então o *ranking* de variáveis pode ser feito de acordo com o *score* de AUC obtido por cada uma. Um pseudocódigo do algoritmo proposto é apresentado no Algoritmo 1.

A principal diferença do método proposto AUC-EPS para as demais abordagens de seleção de características baseadas em AUC analisadas nesta tese é a tarefa de estimar probabilidades para cada valor que existe em cada uma das características, associada a técnica de *smoothing* de *LaPlace*. Isso faz com que o AUC-EPS leve em consideração o grau de variabilidade das características no momento de calcular o seu grau de importância, diferentemente de outras técnicas que utilizam AUC como métrica de seleção das características, como ARCO e o FAST.

Após a execução do algoritmo e obter o ranking das características de acordo com o resultado de AUC de cada uma delas, as 100 características com os maiores valores são selecionadas para realizar a avaliação do algoritmo proposto e comparar com os métodos ARCO e FAST. São selecionadas as 100 melhores características devido o artigo que apresenta os resultados do ARCO [135] também selecionar a mesma quantidade, o que facilita o processo de comparação do modelo proposto. Além disso, outros trabalhos aplicados a conjuntos de dados *microarray* também selecionaram a mesma quantidade durante o processo de seleção de características, como foi o caso do experimento realizado por Lan [73] ao classificar câncer e usando o classificador SVM, de Wang [132] ao classificar câncer usando o classificador kNN e de Alrefai [9] ao classificar câncer usando métodos do estado da arte como SVM, kNN e árvores de decisão.

Algoritmo 1 Pseudocódigo do algoritmo AUC-EPS.

AUC-EPS(*dataset*, *labels*)
{INPUT: ***dataset*** é o conjunto de dados com suas linhas (instâncias) e colunas (características)
labels é o conjunto de classes de cada uma das instâncias}
{OUTPUT: uma lista ordenada de características baseado no *score* delas.}
$coluna = 0$
while $len(score\_caracteristicas) < len(dataset.shape[1])$ **do**
  $valoresCaracteristica = unique(dataset[coluna])$
  for $i = 1$ **to** $len(valoresCaracteristica)$ **do**
    $subconj = dataset[coluna == valoresCaracteristica[i]]$
    $qtdeInstancias = subconj.shape[0]$
    $proba_{-} = \frac{count(subconj[coluna]==valoresCaracteristica[i])_{neg}+1}{qtdeInstancias+2}$
    $proba_{+} = \frac{count(subconj[coluna]==valoresCaracteristica[i])_{pos}+1}{qtdeInstancias+2}$
    $proba\_predita[subconj, 0] = proba_{-}$
    $proba\_predita[subconj, 1] = proba_{+}$
  end for
  $score\_caracteristicas[coluna] = roc\_auc\_score(labels, proba\_predita[:, 1]$
  $coluna = coluna + 1$
end while
return $score\_caracteristicas.sort(order = desc)$

4.2 Desenho de experimentos

A avaliação do método proposto (AUC-EPS) foi realizada considerando oito bases de dados e três classificadores. O método foi comparado com os algoritmos FAST e ARCO, e para padronizar os experimentos, foi feita uma etapa de pré-processamento dos dados antes de submetê-los ao processo de classificação. As próximas seções descrevem os detalhes de cada etapa.

4.2.1 Conjuntos de dados

As bases de dados utilizadas neste estudo estão apresentadas na Tabela 4.1. A maioria das bases de dados estão relacionadas ao diagnóstico de câncer. Além disso, todas essas bases têm como particularidade um número de características maior que o número de instâncias. Essa discrepância entre o número de características e o número de instâncias torna o processo de seleção de características mais desafiador e consequentemente é

comum avaliar o desempenho do modelo proposto em relação as técnicas do estado da arte [140, 135].

Dos oito conjuntos de dados utilizados, sete foram adquiridos através do repositório de Zexuan[1] [154] e um (*DLBC Lymphoma*) foi adquirido através do repositório de Glaab[2] [49]. Em todos os casos as 100 melhores características foram selecionadas, seguindo o experimento realizado por [135].

Tabela 4.1: Bases de dados *microarray* utilizadas durante os experimentos.

Bases de dados	Características	Instâncias	Número de classes
Colon	2000	62	2
Breast Cancer	24481	97	2
CNS	7129	60	2
Leukemia	7129	72	4
DLBC Lymphoma	2647	77	2
Lung Cancer	12600	203	2
MLL	12582	72	3
Ovarian	15154	253	2

As variáveis contínuas foram discretizadas usando o método baseado em entropia desenvolvido por Fayyad e Irani [44]. No entanto, essa discretização ocorreu somente antes da inserção dos dados nos classificadores. Durante o processo de seleção de características, foram utilizados os dados originais (contínuos) dos conjuntos de dados. A técnica de discretização foi aplicada com o intuito de reduzir a taxa de erro do processo de classificação, gerando assim uma melhor performance para os modelos utilizados [20].

Para o método proposto, a AUC é calculada usando como parâmetro a probabilidade positiva estimada para cada subconjunto dos valores das instâncias do conjunto de dados junto dos respectivos *labels* de cada uma destas instâncias. Valores ausentes (*missing values*) foram atribuídos usando o algoritmo não paramétrico desenvolvido por Caruana [21].

Os conjuntos de dados que apresentam problemas de classificação multi-classe foram submetidas a abordagem de "um contra todos" para serem reduzidos a problemas de classificação binário. Essa abordagem consiste em selecionar a classe positiva como sendo pertencente ao lado "um" e todas as demais classes ficam do lado "todos" [81].

4.2.2 Configuração do experimento

O desempenho dos classificadores foi avaliado usando a técnica de validação cruzada de 10 *folds* e a metodologia *bootstrap* de 100 *folds*. Na primeira abordagem, cada base de dados é dividida randomicamente em 10 conjuntos aproximadamente iguais, de

[1] Disponível em https://csse.szu.edu.cn/staff/zhuzx/Datasets.html
[2] Disponível em http://ico2s.org/datasets/microarray.html

forma que cada um deles tenha uma proporção similar de indivíduos que tenham um rótulo igual a 1, ou seja, uma saída positiva. Para cada um dos algoritmos usados nos experimentos para a tarefa de classificação, foram combinados 9 desses *folds*, que foram usados para treino e o 1 *fold* restante como base de teste após o treino. Esse processo é repetido até que todos os 10 *folds* tenham sido testados. O resultado obtido é a média do resultado dos 10 *folds* testados.

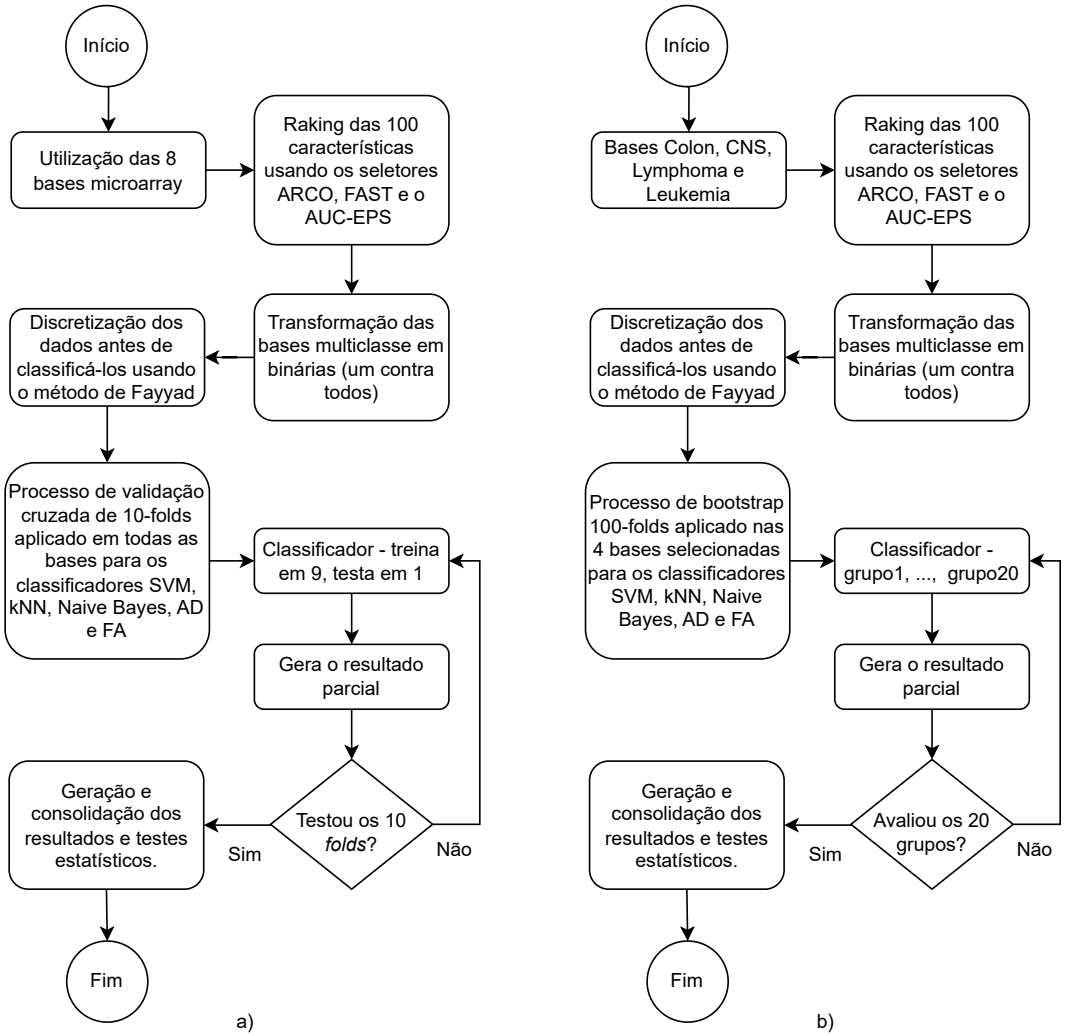

Figura 4.2: Fluxo de realização dos experimentos para as metodologias de validação cruzada (a) e *bootstrap* (b).

Para a segunda abordagem, foram avaliados 20 subconjuntos de características. O subconjunto inicial possuía 5 características e assim que fosse finalizada a avaliação deste subconjunto, o próximo era apresentado ao classificador, com a quantidade de características utilizadas incrementada com mais 5, ou seja, o primeiro subconjunto

possuía 5 características, o segundo 10, o terceiro 15 e assim sucessivamente até completar os 20 subconjuntos. O último possuía as 100 características selecionadas. Para cada um destes subconjuntos de características foi utilizada a divisão do conjunto de dados de forma aleatória em treino (70% das instâncias) e teste (30% restantes) para realização do treinamento e da testagem dos classificadores. Depois dessa avaliação, a média da performance de cada um desses subgrupos de características foi calculado para cada algoritmo.

Os resultados foram apresentados utilizando AUC, sensibilidade, especificidade e acurácia como métricas de avaliação para ambas as abordagens de partição de dados. Foram utilizados os testes estatísticos de *Friedman* e *Nemenyi* para medir a significância estatística entre os métodos para a abordagem de validação cruzada [32]. Além disso, foi utilizado o teste de *Jaccard* para avaliar o subconjunto de características selecionadas por cada um dos algoritmos, tanto o proposto (AUC-EPS) como os demais que foram testados (FAST e ARCO). O Fluxograma apresentado na Figura 4.2 apresenta de forma resumida a metodologia seguida para realização dos experimentos.

CAPÍTULO 5

Resultados

Neste capítulo são apresentados os resultados do algoritmo proposto, com base no desenho do experimento apresentado no Capítulo 4.

5.1 Resultados dos experimentos

As primeiras análises foram realizadas no experimento relacionado a validação cruzada em 10-*folds*. A Tabela 5.1 apresenta o desempenho obtido através da métrica de AUC para os classificadores *Naive Bayes*, SVM, kNN, Árvore de Decisão e Florestas Aleatórias, usando as melhores 100 características selecionadas pelos três algoritmos: ARCO, FAST e o método proposto o AUC-EPS. Para cada base de dados, foi apresentado o resultado médio baseado nas métricas de AUC, sensibilidade, especificidade e acurácia. para metodologia de validação cruzada em 10-*folds* e o respectivo intervalo de confiança em um nível de 0,05 para todos os conjuntos de dados. O teste estatístico de *Friedman* foi usado para verificar o nível de significância entre os algoritmos. Caso seja identificada algum nível de significância, o teste de *Nemenyi* é aplicado de forma complementar. O intuito desse teste é identificar se há algum resultado estatisticamente significativo comparando os resultados em pares.

Não foi encontrada diferença significativa entre o desempenho em nenhum dos classificadores, levando em consideração os resultados obtidos para a métrica de AUC, conforme apresentado na Tabela 5.2.

A Tabela 5.3 apresenta o desempenho obtido pelos classificadores através da métrica de sensibilidade. Ao aplicar o teste de *Friedman*, constatou-se que não há diferença significativa no desempenho dos classificadores, conforme apresentado na Tabela 5.4.

Um outro resultado é apresentado na Tabela 5.5 com o desempenho dos classificadores para a métrica de especificidade. Ao aplicar o teste de *Friedman*, constatou-se que não há diferença significativa no desempenho dos classificadores, conforme apresentado na Tabela 5.6.

Tabela 5.1: Resultados obtidos através da validação cruzada de 10-*folds* para a métrica de AUC com os classificadores **Naive Bayes**, **SVM**, **kNN**, **árvore de decisão** e **florestas aleatórias** depois da seleção das 100 melhores características dos conjuntos de dados descritos na Tabela 4.1.

Alg.	Conjunto de dados	ARCO (%)	FAST (%)	AUC-EPS (%)
Naive Bayes	Colon	93,33 [86,66, 100,00]	90,00 [81,06, 98,93]	91,25 [83,37, 99,13]
	Breast C.	94,13 [88,52, 99,74]	89,26 [83,82, 94,70]	90,57 [87,48, 93,65]
	CNS	90,83 [84,58, 97,08]	88,33 [79,35, 97,30]	83,54 [70,89, 96,20]
	Leukemia	92,08 [85,45, 98,71]	85,33 [71,92, 98,73]	92,50 [85,00, 100,00]
	Lymphoma	92,08 [84,16, 100,00]	87,50 [78,07, 96,92]	92,50 [85,00, 100,00]
	Lung	90,00 [80,00, 100,00]	92,50 [85,00, 100,00]	91,41 [82,83, 100,00]
	MLL	89,19 [78,39, 100,00]	90,00 [80,76, 99,23]	98,50 [97,00, 100,00]
	Ovarian	99,80 [99,60, 100,00]	98,50 [97,00, 100,00]	98,54 [97,08, 100,00]
SVM	Colon	83,75 [67,50, 100,00]	92,91 [85,83, 100,00]	95,83 [91,66, 100,00]
	Breast C.	99,33 [98,66, 100,00]	96,93 [93,86, 100,00]	90,26 [86,46, 94,06]
	CNS	96,66 [93,33, 100,00]	93,96 [88,23, 99,68]	73,54 [51,52, 95,55]
	Leukemia	98,75 [97,50, 100,00]	85,33 [71,92, 98,73]	100,00 [100,00, 100,00]
	Lymphoma	100,00 [100,00, 100,00]	99,16 [98,33, 100,00]	99,16 [98,33, 100,00]
	Lung	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]	97,34 [94,68, 100,00]
	MLL	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]	98,50 [97,00, 100,00]
	Ovarian	100,00 [100,00, 100,00]	99,82 [99,64, 100,00]	99,41 [98,82, 100,00]
kNN	Colon	86,45 [72,91, 100,00]	88,75 [78,90, 98,59]	92,08 [84,16, 100,00]
	Breast C.	97,83 [95,66, 100,00]	92,11 [87,27, 96,96]	89,43 [85,48, 93,37]
	CNS	87,50 [81,13, 93,86]	82,29 [73,73, 90,84]	67,91 [57,49, 78,34]
	Leukemia	95,41 [90,83, 100,00]	100,00 [100,00, 100,00]	98,16 [96,33, 100,00]
	Lymphoma	98,33 [96,66, 100,00]	98,75 [97,50, 100,00]	97,91 [95,83, 100,00]
	Lung	99,72 [99,44, 100,00]	97,22 [94,44, 100,00]	89,19 [78,39, 100,00]
	MLL	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]	93,66 [87,33, 100,00]
	Ovarian	98,85 [97,71, 100,00]	98,18 [96,62, 99,74]	98,44 [96,87, 100,00]
AD	Colon	64,38 [52,82, 75,93]	80,00 [65,89, 94,11]	85,83 [72,92, 98,75]
	Breast C.	79,33 [70,98, 87,69]	73,33 [63,74, 82,92]	85,28 [78,58, 91,99]
	CNS	83,75 [75,33, 92,17]	84,58 [73,21, 95,96]	83,54 [70,89, 96,20]
	Leukemia	97,00 [91,00,100,00]	85,50 [74,76, 96,24]	88,08 [76,87, 99,30]
	Lymphoma	82,17 [73,98, 90,35]	81,50 [71,94, 91,06]	96,67 [90,91, 100,00]
	Lung	93,09 [85,86, 100,00]	91,42 [82,82, 100,00]	87,05 [74,51, 99,58]
	MLL	80,17 [67,69, 92,64]	88,67 [78,95, 98,38]	98,50 [95,11, 100,00]
	Ovarian	93,74 [90,02, 97,46]	92,89 [88,57, 97,20]	92,76 [87,87, 97,65]
FA	Colon	80,63 [64,10, 97,15]	94,17 [88,01, 100,00]	93,96 [85,51, 100,00]
	Breast C.	99,80 [99,35, 100,00]	93,32 [89,48, 97,15]	86,72 [80,06, 93,37]
	CNS	92,92 [84,03, 100,00]	90,21 [79,93, 100,00]	83,54 [70,89, 96,20]
	Leukemia	97,00 [91,00, 100,00]	100,00 [100,00, 100,00]	99,33 [97,82, 100,00]
	Lymphoma	99,17 [97,28, 100,00]	97,08 [93,63, 100,00]	100,00 [100,00, 100,00]
	Lung	99,73 [99,32, 100,00]	99,73 [99,32, 100,00]	93,57 [85,88, 100,00]
	MLL	99,00 [96,74, 100,00]	99,00 [96,74, 100,00]	96,83 [92,05, 100,00]
	Ovarian	99,83 [99,60, 100,00]	99,67 [99,07, 100,00]	99,06 [97,59, 100,00]

Tabela 5.2: Resultado do teste de *Friedman* com base nos resultados da métrica de AUC dos classificadores.

Algoritmo	*p-value*	Rejeita hipótese nula?
Naive Bayes	0,22	Não
SVM	0,14	Não
kNN	0,20	Não
AD	0,68	Não
FA	0,20	Não

Por fim, a Tabela 5.7 apresenta o desempenho dos classificadores para a métrica de acurácia. Ao aplicar o teste de *Friedman* foi encontrada diferença significativa no desempenho do classificador SVM, conforme apresentado na Tabela 5.8. No entanto,

para identificar qual(is) método(s) possui(em) diferença, foi aplicado o teste estatístico de *Nemenyi* e o resultado é apresentado na Tabela 5.9. O teste de *Nemenyi* mostra que para o algoritmo SVM há diferença estatística do algoritmo AUC-EPS em relação ao ARCO e ao FAST.

Com base nos resultados apresentados para o cenário de validação cruzada com 10-*folds*, não houveram diferenças estatísticas para as métricas de AUC, sensibilidade e especificidade. Por outro lado, houve diferença estatística para a métrica de acurácia.

Tabela 5.3: Resultados obtidos através da validação cruzada de 10-*folds* para a métrica de sensibilidade com os classificadores **Naive Bayes**, **SVM**, **kNN**, **árvore de decisão** e **florestas aleatórias** depois da seleção das 100 melhores características dos conjuntos de dados descritos na Tabela 4.1.

Alg.	Conjunto de dados	ARCO (%)	FAST (%)	AUC-EPS (%)
Naive Bayes	Colon	90,00 [72,72, 100,00]	90,47 [78,40, 100,00]	90,47 [67,77, 100,00]
	Breast C.	83,00 [71,42, 94,58]	88,00 [74,18, 100,00]	98,00 [93,48, 100,00]
	CNS	95,00 [83,69, 100,00]	95,00 [83,69, 100,00]	100,00 [100,00, 100,00]
	Leukemia	91,67 [78,79, 100,00]	96,67 [89,13, 100,00]	91,67 [78,79, 100,00]
	Lymphoma	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]	84,33 [74,00, 66,94]
	Lung	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]	92,49 [87,37,97,61]
	MLL	75,00 [56,15, 93,85]	80,00 [61,53, 98,47]	100,00 [100,00, 100,00]
	Ovarian	95,15 [91,25, 99,04]	93,35 [86,90, 99,79]	91,40 [84,75, 98,05]
SVM	Colon	95,00 [87,46, 100,00]	90,00 [80,76, 99,24]	100,00 [100,00, 100,00]
	Breast C.	98,00 [93,48, 100,00]	92,00 [84,61, 99,39]	77,50 [56,76, 98,24]
	CNS	0,00 [0,00, 0,00]	0,00 [0,00, 0,00]	0,00 [0,00, 0,00]
	Leukemia	0,00 [0,00, 0,00]	100,00 [100,00, 100,00]	5,00 [0,00, 16,31]
	Lymphoma	98,33 [94,56, 100,00]	93,33 [81,81, 100,00]	100,00 [100,00, 100,00]
	Lung	99,44 [98,19, 100,00]	99,44 [98,19, 100,00]	100,00 [100,00, 100,00]
	MLL	100,00 [100,00, 100,00]	95,00 [83,69, 100,00]	0,00 [0,00, 0,00]
	Ovarian	96,40 [92,31, 100,00]	96,99 [93,97, 100,00]	98,75 [96,86, 100,00]
kNN	Colon	87,50 [74,85, 100,00]	72,50 [54,72, 90,28]	90,00 [77,50, 100,00]
	Breast C.	98,00 [93,48, 100,00]	83,00 [76,43, 89,57]	83,50 [68,60, 98,40]
	CNS	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]
	Leukemia	96,67 [89,13, 100,00]	100,00 [100,00, 100,00]	96,67 [89,13, 100,00]
	Lymphoma	96,67 [91,64, 100,00]	93,33 [81,81, 100,00]	95,00 [89,24, 100,00]
	Lung	99,44 [98,19, 100,00]	99,44 [98,19, 100,00]	98,92 [97,29, 100,00]
	MLL	100,00 [100,00, 100,00]	95,00 [83,69, 100,00]	80,00 [61,53, 98,47]
	Ovarian	96,40 [92,31, 100,00]	96,36 [93,37, 99,35]	96,88 [92,53, 100,00]
AD	Colon	65,00 [55,76, 74,24]	90,00 [80,76, 99,24]	82,50 [70,43, 94,57]
	Breast C.	78,00 [61,99, 94,01]	68,00 [48,98, 87,02]	87,50 [71,93, 100,00]
	CNS	85,00 [67,72, 100,00]	76,67 [58,72, 94,62]	65,00 [35,55, 94,45]
	Leukemia	96,67 [89,13, 100,00]	75,00 [55,33, 94,67]	76,67 [61,59, 91,75]
	Lymphoma	89,33 [78,33, 100,00]	88,00 [79,80, 96,20]	98,33 [94,56, 100,00]
	Lung	96,17 [91,21, 100,00]	97,84 [95,84, 99,84]	94,09 [90,31, 97,88]
	MLL	70,00 [44,99, 95,01]	85,00 [67,72, 100,00]	100,00 [100,00, 100,00]
	Ovarian	96,36 [92,26, 100,00]	95,77 [91,75, 99,79]	95,11 [90,56, 99,66]
FA	Colon	77,50 [61,84, 93,16]	82,50 [67,78, 97,22]	90,00 [77,50, 100,00]
	Breast C.	91,50 [83,58, 99,42]	83,00 [73,58, 92,42]	83,50 [62,30, 100,00]
	CNS	85,00 [67,72, 100,00]	85,00 [67,72, 100,00]	65,00 [35,55, 94,45]
	Leukemia	96,67 [89,13, 100,00]	96,67 [89,13, 100,00]	90,00 [78,48, 100,00]
	Lymphoma	95,00 [86,95, 100,00]	91,33 [82,68, 99,98]	95,00 [89,24, 100,00]
	Lung	100,00 [100,00, 100,00]	99,44 [98,19, 100,00]	99,47 [98,28, 100,00]
	MLL	90,00 [74,92, 100,00]	80,00 [61,53, 98,47]	100,00 [100,00, 100,00]
	Ovarian	95,81 [90,53, 100,00]	95,77 [92,28, 99,27]	97,50 [95,19, 99,81]

A segunda análise realizada utilizou a metodologia *bootstrap*. O experimento foi realizado utilizando os classificadores *Naive Bayes*, SVM, kNN, Árvore de Decisão e Florestas Aleatórias. As 100 melhores características selecionadas pelo método proposto

Tabela 5.4: Resultado do teste de *Friedman* com base nos resultados da métrica de sensibilidade dos classificadores.

Algoritmo	***p-value***	**Rejeita a hipótese nula?**
Naive Bayes	0,55	Não
SVM	0,61	Não
kNN	0,30	Não
AD	0,88	Não
FA	0,17	Não

Tabela 5.5: Resultados obtidos através da validação cruzada de 10-*folds* para a métrica de especificidade com os classificadores **Naive Bayes**, **SVM**, **kNN**, **árvore de decisão** e **florestas aleatórias** depois da seleção das 100 melhores características dos conjuntos de dados descritos na Tabela 4.1.

Alg.	**Conjunto de dados**	**ARCO (%)**	**FAST (%)**	**AUC-EPS (%)**
Naive Bayes	**Colon**	75,00 [56,15, 93,85]	90,47 [74,77, 100,00]	90,47 [77,59, 100,00]
	Breast C.	98,00 [93,48, 100,00]	82,00 [71,44, 92,56]	54,67 [36,57, 72,76]
	CNS	86,67 [76,46, 96,88]	81,67 [69,16, 94,17]	35,83 [14,98, 56,68]
	Leukemia	73,00 [57,16, 88,84]	100,00 [100,00, 100,00]	91,50 [83,58, 99,42]
	Lymphoma	85,00 [67,72, 100,00]	75,00 [56,15, 93,85]	90,00 [74,92, 100,00]
	Lung	80,00 [54,99, 100,00]	85,00 [67,72, 100,00]	80,00 [54,99, 100,00]
	MLL	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]	94,33 [87,77, 100,89]
	Ovarian	97,78 [94,43, 100,00]	95,56 [90,00, 100,00]	98,89 [96,38, 100,00]
SVM	**Colon**	68,33 [52,84, 67,00]	80,00 [54,99, 100,00]	0,00 [0,00, 0,00]
	Breast C.	94,00 [87,09, 100,00]	84,00 [72,71, 95,29]	84,33 [75,34, 93,33]
	CNS	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]
	Leukemia	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]
	Lymphoma	100,00 [100,00, 100,00]	95,00 [83,69, 100,00]	0,00 [0,00, 0,00]
	Lung	90,00 [74,92, 100,00]	90,00 [74,92, 100,00]	0,00 [0,00, 0,00]
	MLL	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]
	Ovarian	97,78 [94,43, 100,00]	95,56 [90,00, 100,00]	90,11 [82,21, 98,01]
kNN	**Colon**	71,67 [53,85, 89,48]	85,00 [67,72, 100,00]	86,67 [70,97, 100,00]
	Breast C.	98,00 [93,48, 100,00]	86,33 [74,63, 98,04]	76,33 [63,11, 89,56]
	CNS	69,17 [57,91, 80,43]	61,67 [46,74, 76,59]	30,83 [12,39, 49,27]
	Leukemia	94,00 [87,09, 100,00]	100,00 [100,00, 100,00]	91,50 [83,58, 99,42]
	Lymphoma	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]
	Lung	90,00 [74,92, 100,00]	90,00 [74,92, 100,00]	80,00 [54,99, 100,00]
	MLL	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]	92,33 [85,22, 99,45]
	Ovarian	98,89 [96,38, 100,00]	90,00 [82,10, 97,90]	94,45 [87,69, 100,00]
AD	**Colon**	60,00 [37,38, 82,62]	70,00 [39,84, 100,00]	86,67 [70,97, 100,00]
	Breast C.	80,67 [69,55, 91,78]	78,67 [66,6, 90,73]	76,67 [65,83, 87,51]
	CNS	76,67 [58,72, 94,62]	79,17 [62,48, 95,85]	92,50 [83,86, 100,00]
	Leukemia	96,00 [89,97, 100,00]	96,00 [89,97, 100,00]	93,50 [85,95, 100,00]
	Lymphoma	75,00 [56,15, 93,85]	75,00 [56,15, 93,85]	95,00 [83,69, 100,00]
	Lung	90,00 [74,92, 100,00]	85,00 [67,72, 100,00]	80,00 [54,99, 100,00]
	MLL	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]
	Ovarian	91,11 [82,90, 99,32]	90,00 [80,49, 99,52]	88,89 [78,98, 98,80]
FA	**Colon**	75,00 [56,15, 93,85]	81,67 [64,39, 98,94]	73,33 [46,85, 99,81]
	Breast C.	100,00 [100,00, 100,00]	86,33 [76,77, 95,90]	80,33 [70,77, 89,90]
	CNS	91,67 [81,93, 100,00]	86,67 [76,46, 96,88]	92,50 [83,86, 100,00]
	Leukemia	96,00 [89,97, 100,00]	100,00 [100,00, 100,00]	96,00 [89,97, 100,00]
	Lymphoma	95,00 [83,69, 100,00]	90,00 [74,92, 100,00]	90,00 [74,92, 100,00]
	Lung	85,00 [67,72, 100,00]	85,00 [60,86, 100,00]	50,00 [20,80, 79,20]
	MLL	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]	100,00 [100,00, 100,00]
	Ovarian	96,67 [92,83, 100,00]	96,67 [92,83, 100,00]	94,45 [88,83, 100,00]

(AUC-EPS) e outros pelos outros dois métodos de seleção de características (ARCO e FAST) foram utilizadas seguindo a metodologia *bootstrap*. ou seja, o ranking das 100 características foi dividido em 20 subconjuntos de características, começando em 5 e

Tabela 5.6: Resultado do teste de *Friedman* com base nos resultados da métrica de especificidade dos classificadores.

Algoritmo	***p-value***	**Rejeita hipótese nula?**
Naive Bayes	0,87	Não
SVM	0,05	Não
kNN	0,13	Não
AD	1,00	Não
FA	0,07	Não

Tabela 5.7: Resultados obtidos através da validação cruzada de 10-*folds* para a métrica de acurácia com os classificadores **Naive Bayes**, **SVM**, **kNN**, **árvore de decisão** e **florestas aleatórias** depois da seleção das 100 melhores características dos conjuntos de dados descritos na Tabela 4.1.

Alg.	**Conjunto de dados**	**ARCO (%)**	**FAST (%)**	**AUC-EPS (%)**
Naive Bayes	**Colon**	85,00 [69,66, 100,00]	90,47 [82,32, 98,62]	90,47 [78,13, 100,00]
	Breast Cancer	90,85 [85,72, 95,98]	84,83 [79,07, 90,58]	74,85 [65,96, 83,73]
	CNS	89,67 [83,26, 96,07]	86,57 [77,10, 96,04]	58,62 [45,51, 71,72]
	Leukemia	92,92 [87,46, 98,37]	98,75 [95,92, 100,00]	91,73 [83,50, 99,95]
	Lymphoma	96,25 [91,93, 100,00]	93,75 [89,04, 98,46]	85,89 [77,73, 94,05]
	Lung	98,07 [95,67, 100,00]	98,55 [96,87, 100,00]	91,20 [84,59, 97,82]
	MLL	92,86 [87,47, 98,24]	94,28 [89,01, 99,56]	95,89 [91,150, 100,00]
	Ovarian	96,08 [93,83, 98,32]	94,11 [90,38, 97,83]	94,08 [90,00, 98,15]
SVM	**Colon**	85,48 [76,70, 94,25]	86,90 [77,53, 96,28]	64,76 [61,88, 67,63]
	Breast C.	95,87 [92,03, 99,71]	87,76 [81,91, 93,60]	80,80 [72,2, 89,41]
	CNS	65,05 [62,56, 67,54]	65,05 [62,56, 67,54]	65,04 [62,55, 67,53]
	Leukemia	87,92 [77,94, 97,90]	100,00 [100,00, 100,00]	67,20 [62,39, 72,01]
	Lymphoma	98,75 [95,92, 100,00]	93,75 [84,09, 100,00]	75,48 [73,34, 77,61]
	Lung	98,55 [96,87, 100,00]	98,55 [96,87, 100,00]	90,42 [90,32, 90,53]
	MLL	100,00 [100,00, 100,00]	98,57 [95,34, 100,00]	72,14 [71,06, 73,22]
	Ovarian	96,88 [94,34, 99,41]	96,46 [94,39, 98,53]	95,63 [92,2, 99,06]
kNN	**Colon**	82,14 [70,23, 94,06]	77,62 [63,12, 92,12]	88,81 [80,90, 96,72]
	Breast Cancer	97,98 [94,91, 100,00]	84,74 [78,70, 90,78]	79,49 [73,69, 85,29]
	CNS	80,14 [73,05, 87,23]	74,91 [64,77, 85,04]	55,28 [43,67, 66,89]
	Leukemia	92,92 [84,03, 100,00]	100,00 [100,00, 100,00]	92,98 [87,61, 98,34]
	Lymphoma	97,50 [93,73, 100,00]	95,00 [86,36, 100,00]	96,25 [91,93, 100,00]
	Lung	98,55 [96,87, 100,00]	98,55 [96,87, 100,00]	97,09 [94,15, 100,00]
	MLL	100,00 [100,00, 100,00]	98,57 [95,34, 100,00]	88,93 [82,53,95,32]
	Ovarian	97,28 [94,66, 99,90]	94,06 [91,63, 96,49]	96,00 [92,98, 99,02]
AD	**Colon**	90,47 [79,13, 100,00]	90,47 [80,72, 100,00]	90,47 [79,58, 100,00]
	Breast C.	79,27 [70,88, 87,67]	73,46 [64,2, 82,71]	81,70 [74,93, 88,47]
	CNS	79,90 [70,47, 89,34]	78,48 [70,83, 86,12]	81,71 [70,24, 93,19]
	Leukemia	96,25 [90,21, 100,00]	89,23 [80,81, 97,64]	87,56 [78,41, 96,71]
	Lymphoma	85,89 [78,91, 92,88]	84,82 [77,79, 91,85]	97,50 [93,73, 100,00]
	Lung	95,49 [91,00, 99,99]	96,59 [94,27, 98,92]	92,69 [88,63, 96,74]
	MLL	84,46 [75,49, 93,44]	90,18 [81,81, 98,55]	95,89 [91,15, 100,00]
	Ovarian	94,45 [91,42, 97,47]	93,68 [90,35, 97,01]	92,86 [88,85, 96,87]
FA	**Colon**	90,47 [80,49, 100,00]	90,47 [78,55, 100,00]	90,47 [81,14, 99,80]
	Breast C.	95,98 [92,25, 99,71]	84,74 [78,01, 91,46]	81,70 [73,41, 89,98]
	CNS	89,67 [83,26, 96,07]	86,57 [78,94, 94,20]	81,71 [70,24, 93,19]
	Leukemia	96,25 [90,21, 100,00]	98,75 [95,92, 100,00]	93,57 [88,71, 98,43]
	Lymphoma	95,00 [87,46, 100,00]	91,07 [83,65, 98,50]	93,75 [89,04, 98,46]
	Lung	98,55 [96,87, 100,00]	98,07 [95,67, 100,00]	95,59 [93,64, 97,55]
	MLL	94,28 [89,01, 99,56]	92,86 [87,47, 98,24]	95,89 [91,15, 100,00]
	Ovarian	96,09 [92,92, 99,27]	96,09 [94,29, 97,89]	96,40 [94,29, 98,51]

crescendo de 5 em 5 (ex.: 5, 10, 15 e assim por diante) até chegar ao grupo com 100.

Cada subgrupo de cada método de seleção de característica foi submetido a cada

Tabela 5.8: Resultado do teste de *Friedman* com base nos resultados da métrica de acurácia dos classificadores.

Algoritmo	*p-value*	Rejeita hipótese nula?
Naive Bayes	0,13	Não
SVM	0,00	Sim
kNN	0,13	Não
AD	0,65	Não
FA	0,18	Não

Tabela 5.9: Resultado do teste estatístico de *Nemenyi* de acordo com a performance do algoritmo SVM para a métrica de acurácia.

Teste Nemenyi	ARCO	FAST	AUC-EPS
ARCO	1,00	0,85	0,00
FAST	0,85	1,00	0,01
AUC-EPS	0,00	0,01	1,00

um dos classificadores e os resultados foram calculados para as métricas de AUC, sensibilidade, especificidade e acurácia. Para esse experimento, foram utilizados somente quatro conjuntos de dados (Colon, Leukemia, Lymphoma e CNS), repetindo o experimento desenvolvido por [135], que aplicou a metodologia *bootstrap* para 20 grupos de características nesses mesmos conjuntos de dados.

O diagrama de diferença crítica foi utilizado para representar o resultado dos testes estatísticos de *Nemenyi* para o cenário de *bootstrap*. Nesse diagrama, os resultados que estão mais para a esquerda são maiores e quanto mais para a direita os resultados são menores. Além disso, esse diagrama demonstra quais os métodos são estatisticamente iguais, através de uma barra que os conecta.

Os testes estatísticos foram realizados utilizando todos os resultados de todos os classificadores, para todos os conjuntos de dados por seletor de características. Os testes foram realizados dessa forma para aumentar o tamanho das amostras.

A Tabela A.1 no Apêndice A apresenta os resultados dos algoritmos *Naive Bayes*, SVM e kNN para a métrica de AUC. A Tabela A.2 no Apêndice A também apresenta os resultados para a mesma métrica, mas para os algoritmos Árvore de Decisão e Florestas Aleatórias. O teste de *Friedman* foi aplicado e através dele foi encontrada diferença significativa entre o desempenho dos seletores de características (p-*value* = 0,00, $\alpha < 0{,}05$). Para identificar quais métodos possuem diferença, foi utilizado o diagrama de diferença crítica apresentado na Figura 5.1. O diagrama mostra que para a métrica de AUC, há diferença estatística entre os algoritmos AUC-EPS, FAST e ARCO.

A métrica de sensibilidade apresenta os resultados dos algoritmos *Naive Bayes*, SVM e kNN na Tabela A.3 no Apêndice A e dos algoritmos Árvore de Decisão e Florestas

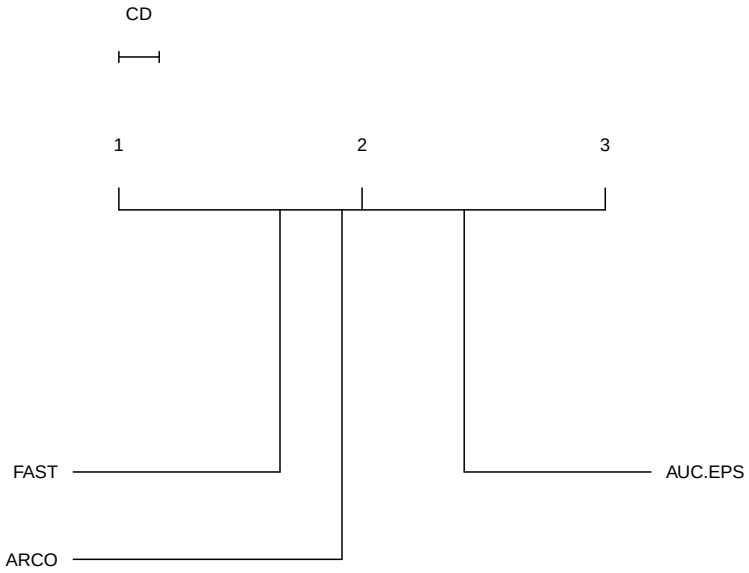

Figura 5.1: Diagrama de diferença crítica para os resultados da métrica de AUC para a validação *bootstrap*.

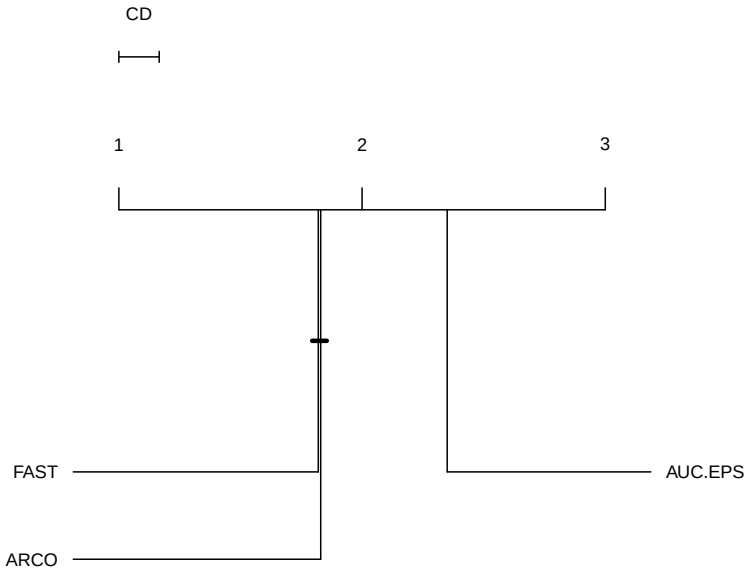

Figura 5.2: Diagrama de diferença crítica para os resultados da métrica de sensibilidade para a validação *bootstrap*.

Aleatórias na Tabela A.4 no Apêndice A. O teste estatístico de *Friedman* foi aplicado e através dele foi encontrada diferença significativa entre o desempenho dos seletores de características (*p-value* = 0,00, $\alpha < 0{,}05$). Com isso, foi utilizado o diagrama de diferença crítica e o resultado é apresentado na Figura 5.2. Esse resultado mostra que há uma

diferença estatística do algoritmo AUC-EPS em relação ao ARCO e ao FAST, porém, não foi identificada diferença estatística entre os métodos ARCO e FAST (*p-value* = 0,90, $\alpha < 0{,}05$).

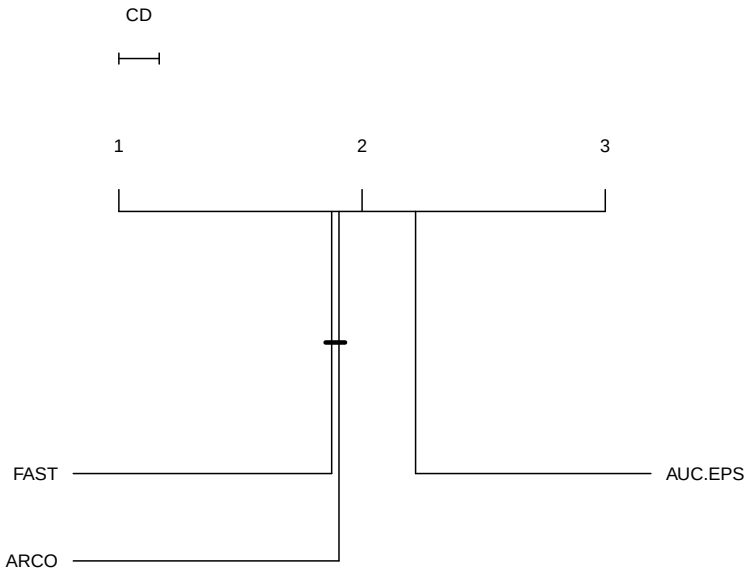

Figura 5.3: Diagrama de diferença crítica para os resultados da métrica de especificidade para a validação *bootstrap*.

A Tabela A.5 e a Tabela A.6, ambas no Apêndice A, apresentam os resultados para a métrica de especificidade. Ao aplicar o teste de *Friedman*, foi possível encontrar diferença significativa entre o desempenho dos seletores de características (*p-value* = 0,00, $\alpha < 0{,}05$). Dessa forma, foi utilizado o diagrama de diferença crítica e o resultado é apresentado na Figura 5.3 Esse resultado mostra que há uma diferença estatística do algoritmo AUC-EPS em relação ao ARCO e ao FAST, no entanto, não foi identificada diferença estatística entre os métodos ARCO e FAST (*p-value* = 0,90, $\alpha < 0{,}05$).

Por fim, a Tabela A.7 e a Tabela A.8, que estão no Apêndice A, apresentam os resultados da metodologia *bootstrap* para a métrica de acurácia. O teste estatístico de *Friedman* foi aplicado e através dele foi encontrada diferença significativa entre o desempenho dos seletores de características (*p-value* = 0,00, $\alpha < 0{,}05$). Para identificar quais métodos possuem diferença entre eles, foi aplicado o diagrama de diferença crítica e o resultado é apresentado na Figura 5.4. O resultado desse teste mostra que para a métrica de AUC, há diferença estatística entre os algoritmos AUC-EPS, FAST e ARCO.

Outra visualização apresentada para os resultados do *bootstrap* foi a utilização do gráfico de caixas (*boxplot*), com o objetivo de mostrar o comportamento de cada classificador utilizando os rankings de cada um dos seletores de características para os conjuntos de dados Colon, Leukemia, Lymphoma e CNS.

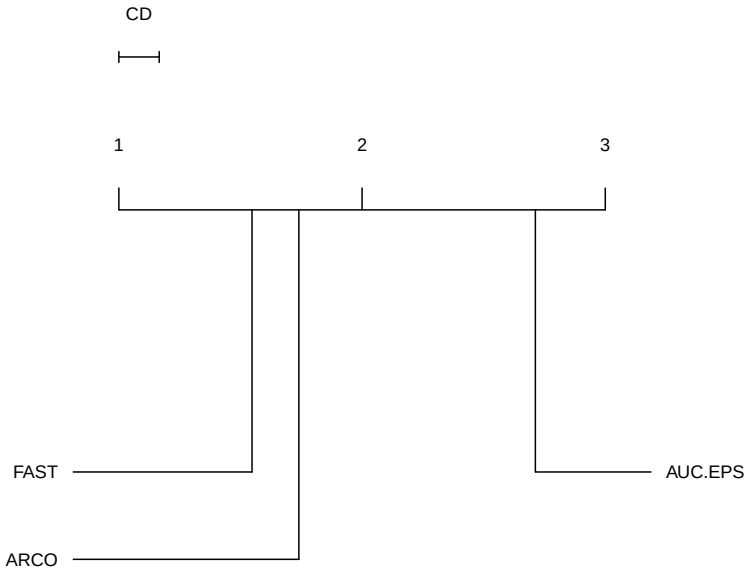

Figura 5.4: Diagrama de diferença crítica para os resultados da métrica de acurácia para a validação *bootstrap*.

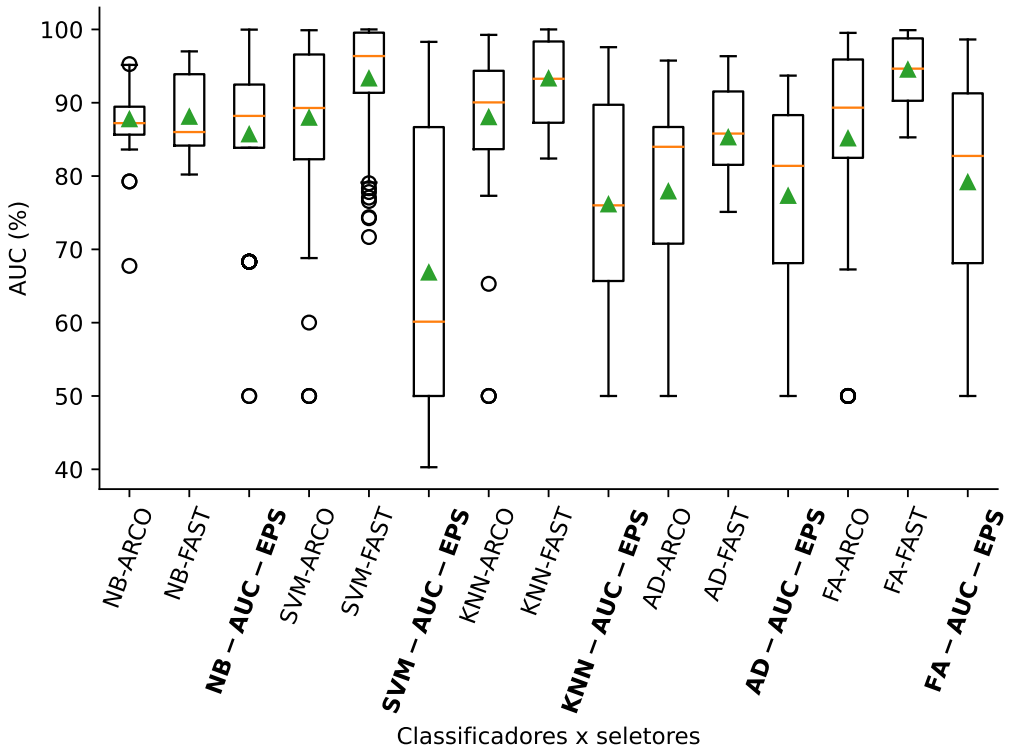

Figura 5.5: *Boxplot* considerando os 5 classificadores para a métrica de AUC.

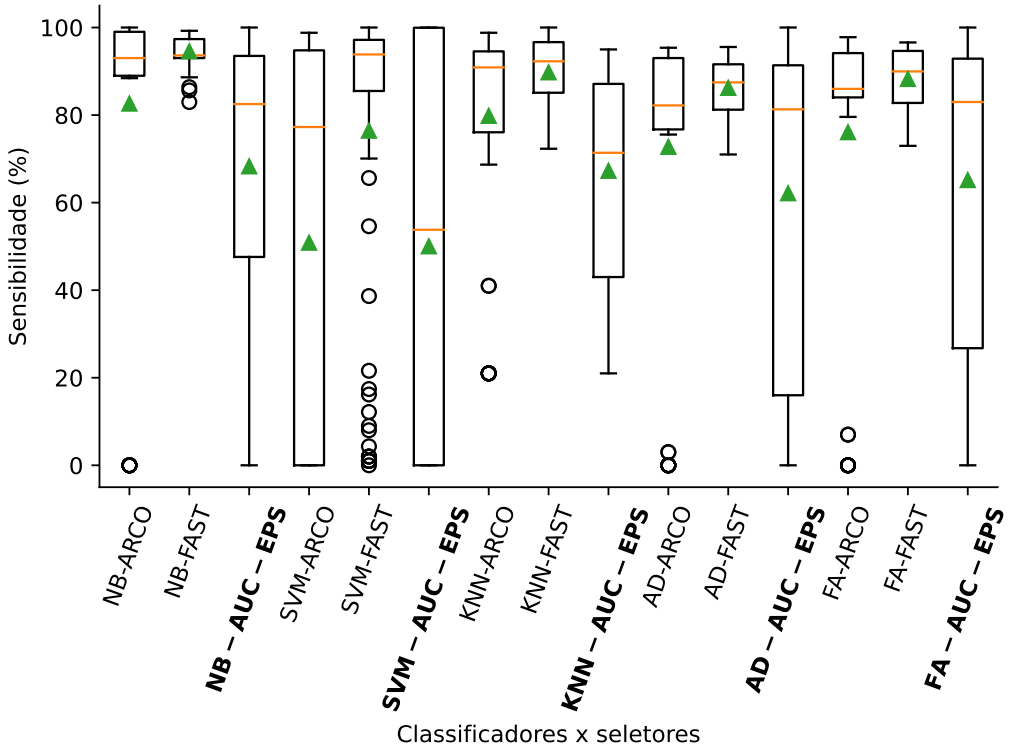

Figura 5.6: *Boxplot* considerando os 5 classificadores para a métrica de sensibilidade.

A Figura 5.5 e a Figura 5.6 apresentam os resultados desse cenário para as métricas de AUC e sensibilidade, respectivamente, enquanto que a Figura 5.7 e a Figura 5.8 apresentam os resultados para as métricas de especificidade e acurácia, respectivamente. Além de apresentarem as medidas normais de um gráfico deste tipo (mediana, quartil, mínimo e máximo), foi destacado o ponto médio (triângulo verde) para o resultado de cada um dos métodos de seleção de características, nos cinco classificadores, considerando os resultados obtidos nas quatro conjuntos de dados utilizadas no experimento do *bootstrap*.

As Figuras A.1-A.4 no Apêndice A fornecem uma perspectiva adicional sobre os resultados derivados da técnica de *bootstrap*. Com base nessas Figuras é possível observar a evolução dos resultados dos classificadores conforme cada subconjunto de características é submetido aos mesmos. Esse gráfico foi construído utilizando os resultados dos cinco classificadores utilizados durante os experimentos e abrange as quatro métricas utilizadas para avaliar o desempenho dos modelos. A Figura A.1 e a Figura A.2 (Apêndice A) apresentam os resultados relacionados as métricas de AUC e sensibilidade, respectivamente, enquanto que a Figura A.3 e a Figura A.4 (Apêndice A) apresentam os resultados relacionados as métricas de especificidade e acurácia, respectivamente.

Elas apresentam os 20 pontos correspondentes a cada um dos quatro conjuntos de dados investigados no experimento de *bootstrap*: Colon, Leukemia, Lymphoma e CNS. A

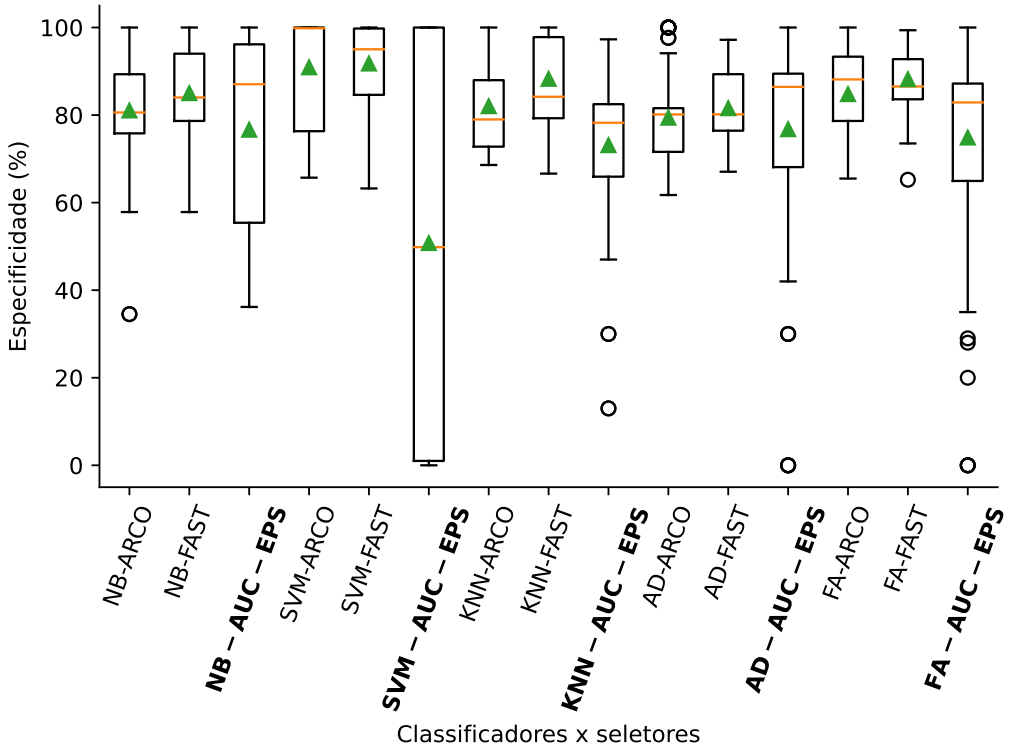

Figura 5.7: *Boxplot* considerando os 5 classificadores para a métrica de especificidade.

disposição desses pontos nos gráficos segue a sequência dos conjuntos de dados, ou seja, do ponto 0 ao 19 correspondem aos resultados do conjunto Colon; do ponto 20 ao 39, ao conjunto Leukemia; do ponto 40 ao 59, ao conjunto Lymphoma; e do ponto 60 ao 79, ao conjunto CNS. Esse conjunto de resultados visa ilustrar o desempenho dos diferentes classificadores com base nas características selecionadas pelos métodos ARCO, AUC-EPS e FAST, conforme a quantidade de características aumenta em intervalos de cinco.

Esses gráficos evidenciam que, embora os resultados do AUC-EPS, em relação ao ARCO e ao FAST, sejam inferiores, ele é capaz de alcançar um bom desempenho a partir do momento em que o número de características passa a ser igual ou superior a 50.

Os resultados demonstram que há certos níveis de similaridade e divergência para cada uma das métricas que foram selecionadas. Com o intuito de verificar a similaridade das características selecionadas por cada método, foi realizada uma verificação se o subconjunto de características selecionadas pelos algoritmos FAST e ARCO é igual ou diferente ao subconjunto do algoritmo AUC-EPS. Para realizar essa avaliação, utilizou-se o cálculo do índice de estabilidade de Jaccard [66]. Este índice é calculado de acordo com a Equação (5-1), em que s e s' são dois subconjuntos de características, o é o número de características que é comum nos dois conjuntos e l é a quantidade de características existentes em ambos os conjuntos,

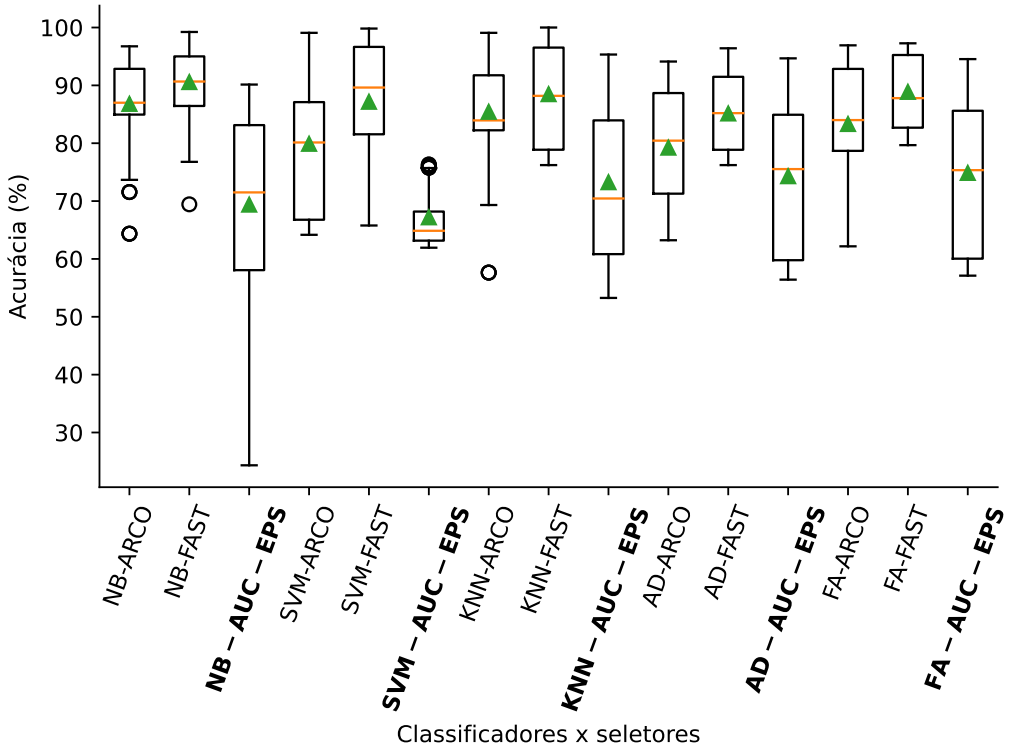

Figura 5.8: *Boxplot* considerando os 5 classificadores para a métrica de acurácia.

$$JI(\mathbf{s},\mathbf{s'}) == \frac{|\mathbf{s} \cap \mathbf{s'}|}{|\mathbf{s} \cup \mathbf{s'}|} = \frac{o}{l}. \tag{5-1}$$

A Tabela 5.10 apresenta o resultado do teste de *Jaccard* para avaliação do subconjunto de características selecionadas pelo algoritmo AUC-EPS comparado com os algoritmos FAST e ARCO. Nela, é possível notar que os valores obtidos foram muito baixos, o que significa dizer que os subconjuntos de características selecionados pelos algoritmos ARCO e FAST é bem diferente dos obtidos pelo algoritmo AUC-EPS. Com isso, não se fez necessária a análise dos rankings, pois como os subconjuntos já são comprovadamente diferentes com base no teste de *Jaccard*, os rankings também serão.

Tabela 5.10: Avaliação da similaridade de subconjuntos de características usando *Jaccard*.

Conjunto de dados	AUC-EPS *vs* FAST	AUC-EPS *vs* ARCO	FAST *vs* ARCO
Colon	0,041	0,036	0,000
Breast Cancer	0,005	0,000	0,156
CNS	0,015	0,010	0,333
Leukemia	0,005	0,000	0,000
Lung	0,010	0,015	0,379
Lymphoma	0,010	0,020	0,226
MLL	0,000	0,015	0,290
Ovarian	0,000	0,000	0,574

5.2 Discussão

Os resultados demonstram que o algoritmo AUC-EPS pode contribuir para selecionar características que reflitam a variação de cada uma delas, quando comparado a outros métodos de seleção de características baseados em AUC, como é o caso do FAST e do ARCO. O algoritmo proposto utiliza AUC combinado à técnica de estimativa de probabilidade ao método de suavização de *La Place*.

Os resultados mostram ainda que o algoritmo AUC-EPS é eficaz para a tarefa de seleção de características em dados *microarray*. Mesmo com variação nos resultados dos classificadores, não existem intervalos que sejam considerados atipicamente diferentes ou discrepantes em relação aos métodos de seleção de características utilizados, na abordagem de validação cruzada utilizando 10-*folds*. Além disso, o fato de possuir desempenho equivalente aos métodos ARCO e FAST e pode ser um atrativo para utilização do AUC-EPS, devido o algoritmo proposto selecionar características completamente diferentes dos demais métodos. No caso da abordagem de *bootstrap*, o AUC-EPS, conforme demonstrado nos diagramas de diferença crítica, possui diferença estatística para os demais métodos de forma que o ARCO e o FAST têm desempenho melhor que o método proposto.

Levando em consideração os resultados obtidos através da validação cruzada utilizando 10-*folds*, a exceção ao que foi citado anteriormente é a métrica de acurácia. Analisando os resultados de cada classificador para cada conjunto de dados, é possível notar que os métodos ARCO e FAST superam, na maioria dos casos, o algoritmo AUC-EPS, apesar de na maioria desses casos ter ocorrido vantagem com proximidade entre os resultados. A exceção é o algoritmo de árvore de decisão, em que o método proposto (AUC-EPS) obteve resultados superiores para quatro conjuntos de dados (Breast C., CNS, Lymphoma e MLL) e equivalência em um conjunto de dados (Colon).

Apesar dos resultados da métrica de acurácia serem superados pelos métodos ARCO e FAST no experimento de validação cruzada, o AUC-EPS teve desempenho

estatisticamente equivalente nas demais métricas. Em cenários clínicos, sensibilidade, especificidade e AUC são métricas de maior importância que a taxa de acerto geral [108]. Com base nas características selecionadas pelo AUC-EPS e nos resultados obtidos para sensibilidade, especificidade e AUC, pode-se afirmar que o método auxilia a identificar a presença ou ausência de determinado fenômeno (doença), o que é clinicamente relevante. Cabe frisar também que os *rankings* do ARCO e do FAST, em geral, são divergentes porém estão mais próximos quando comparados ao AUC-EPS.

O estudo fornece contribuição teórica com relação aos métodos de seleção de características aplicados a bases *microarray* e também a técnicas de seleção de características baseadas em AUC aplicadas às mesmas bases. O classificador SVM e seus derivados, combinados a técnicas de seleção de características e aplicados em conjuntos de dados *microarray*, podem auxiliar na maximização das métricas de avaliação.

Um ponto de destaque é o fato do AUC-EPS considerar a variabilidade de valores existentes dentro de cada característica. Com base nos experimentos realizados, foi possível demonstrar a importância dessa característica do AUC-EPS. Foi observado que a combinação das técnicas de AUC, estimativa de probabilidade e método de suavização de *La Place* podem trazer um resultado mais representativo para cada característica quando comparado a outras abordagens (algoritmos FAST e ARCO). Além disso, foi demonstrado que mesmo o conjunto das características selecionadas, pelo algoritmo proposto neste estudo, sendo diferente dos quais foi comparado, os resultados possuíram desempenho semelhantes ou superiores em parte, quando comparados aos algoritmos FAST e ARCO.

Outro ponto a ser destacado neste trabalho é o fato do AUC-EPS selecionar características bem diferentes do ARCO e do FAST. Utilizando como exemplo o conjunto de dados Ovarian, em que o AUC-EPS obteve *ranking* completamente diferente dos outros dois métodos, algumas das 100 características selecionadas pelo método proposto estão entre as mais importantes para discriminar casos de câncer [46], são elas: MZ25.307419, MZ25.401403, MZ25.49556, MZ25.589892 e MZ25.21361. Desta forma, o algoritmo AUC-EPS, com base na combinação de técnicas já citadas e pelo fato de considerar o grau de variabilidade de cada característica, é capaz de auxiliar na detecção de câncer.

A consideração do grau de variabilidade dos valores de cada característica por meio da estimativa de probabilidade antes de calcular o valor de AUC pode estar restringindo a maximização de algumas métricas. No entanto, essa abordagem traz maior confiabilidade ao grau de importância calculado para as características. Uma possibilidade futura é combinar o método proposto com técnica de eliminação de características redundantes, como é o caso do algoritmo ARCO.

Apesar dos tempos de execução dos métodos de seleção de características utilizados nos experimentos não terem sido considerados, percebeu-se durante os experimentos

que o método AUC-EPS possuiu menor tempo de execução quando comparado ao método ARCO (complexidade: O(kmnlogn)) e maior tempo de execução quando comparado ao método FAST (complexidade: O(mnlogn)). Como sugestão de aprimoramento da metodologia, os resultados relacionados ao tempo de execução poderão ser considerados para efeito de otimização e comparação dos métodos de seleção de características.

CAPÍTULO 6

Conclusão

Este estudo apresenta a teoria dos métodos de seleção de características aplicados a bases de dados *microarray* através de mineração de texto usando dados bibliométricos, junto da proposta de algoritmo de seleção de características que usa AUC combinado com técnicas que proporcionam melhorias ao processo.

O objetivo central desta pesquisa foi avaliar quais as tendências no uso de métodos de seleção de características em geral e métodos de seleção de características que usam AUC como estimador, aplicados a dados *microarray* e melhorar a aplicabilidade das técnicas baseadas em AUC em pesquisas científicas. Apesar de haverem diversas pesquisas sobre esse tema, não foram encontradas até o presente estudo pesquisas que utilizassem AUC combinado as técnicas de estimativa de probabilidade e *smoothing*, aplicado a seleção de características.

Uma das contribuições do trabalho foi o desenvolvimento dos estudos de mineração de texto aplicados a dados bibliométricos. Esses estudos permitiram a caracterização da seleção de características aplicadas a bases de dados *microarray* e a caracterização das técnicas de seleção de características que utilizam AUC como estimador aplicados as mesmas bases de dados.

Foi possível realizar o desenvolvimento do método AUC-EPS, o algoritmo de seleção de características baseado em AUC com estimativa de probabilidade e *smoothing*. A técnica proposta foi avaliada utilizando oito bases de dados médicas, mais especificamente, conjuntos de dados de expressão gênica *microarray*. Os resultados apresentados mostraram que o algoritmo desenvolvido tem resultados comparáveis aos métodos do estado da arte para seleção de características baseados em AUC, de acordo com a comparação realizada através da metodologia de validação cruzada 10-*folds*.

Na metodologia conhecida como *boostrap*, em que foram utilizados quatro conjuntos de dados (Colon, Leukemia, Lymphoma e CNS), foi possível verificar que o método teve performance abaixo dos métodos ARCO e FAST. No entanto, nesse mesmo experimento, foi possível verificar que ao chegar nos subconjuntos que utilizam entre 50-70 características, há melhora nos resultados dos classificadores utilizando o método proposto (AUC-EPS), superando, em alguns momentos, os resultados dos demais méto-

dos. O fato de não possuir bom desempenho com quantidade inferior de características necessita de estudos mais aprofundados que deverão ser realizados futuramente.

O foco deste estudo foi o desenvolvimento do seletor de características combinando as técnicas de AUC, estimativa de probabilidade e método de suavização de *La Place*. Existem outras técnicas que podem ser abordadas e combinadas ao AUC com o objetivo de ressaltar a variabilidade existente dentro de cada característica. No entanto, com base nos experimentos realizados, foi possível demonstrar a importância dessa característica do AUC-EPS utilizando a combinação das técnicas citadas. Além disso, este estudo fornece informações que podem ser utilizadas para realizar melhorias na aplicação de técnica de seleção de características dentro do contexto de AUC a bases de dados *microarray*.

A hipótese desse trabalho foi corroborada através dos resultados apresentados. As evidências e as análises apresentaram que o método AUC-EPS possui desempenho equiparável ou superior quando comparados aos métodos FAST e ARCO para maioria das métricas de avaliação, fornecendo suporte substancial à hipótese formulada inicialmente. Essa confirmação fortalece a compreensão atual e contribui para o corpo de conhecimento existente na área de seleção de características, principalmente no contexto dos métodos baseados em AUC.

Como trabalho futuro, há o planejamento em analisar detalhadamente o subconjunto das 100 características escolhidas pelo método proposto, comparando com as 100 características selecionadas pelas abordagens ARCO e FAST utilizadas nos experimentos, verificando quais são as características que foram escolhidas por cada método, quais foram iguais, diferentes e estabelecer alguma relação com base nestas análises. Tal análise necessitará do apoio de um especialista em expressão gênica. Por esse motivo, essa análise não foi realizada de forma detalhada durante o desenvolvimento deste trabalho. Uma segunda possibilidade, que surgiu com base na revisão realizada sobre os métodos de seleção de características, é que percebeu-se que ao associar algoritmos de seleção de características a técnicas de eliminação de redundâncias é possível refinar o processo de seleção de características de forma que o conjunto selecionado potencialize a performance dos classificadores. Com base nisso, há o objetivo secundário de realizar teste combinando o método AUC-EPS a algumas dessas abordagens e verificar o desempenho.

Referências Bibliográficas

[1] ABD KARIM, S. A.; NOHUDDIN, P. N. **Bibliometric analysis of data mining on medical imaging**. In: *Journal of Physics: Conference Series*, volume 1997, p. 012017. IOP Publishing, 2021.

[2] ADNAN, N.; LEI, C.; RUAN, J. **Robust edge-based biomarker discovery improves prediction of breast cancer metastasis**. *BMC bioinformatics*, 21(14):1–18, 2020.

[3] AKAY, M. F. **Support vector machines combined with feature selection for breast cancer diagnosis**. *Expert systems with applications*, 36(2):3240–3247, 2009.

[4] AL-TASHI, Q.; KADIR, S. J. A.; RAIS, H. M.; MIRJALILI, S.; ALHUSSIAN, H. **Binary optimization using hybrid grey wolf optimization for feature selection**. *IEEE Access*, 7:39496–39508, 2019.

[5] ALHENAWI, E.; AL-SAYYED, R.; HUDAIB, A.; MIRJALILI, S. **Feature selection methods on gene expression microarray data for cancer classification: A systematic review**. *Computers in Biology and Medicine*, 140:105051, 2022.

[6] ALIFERIS, C. F.; TSAMARDINOS, I.; STATNIKOV, A. **Hiton: a novel markov blanket algorithm for optimal variable selection**. In: *AMIA annual symposium proceedings*, volume 2003, p. 21. American Medical Informatics Association, 2003.

[7] ALONSO-BETANZOS, A.; BOLÓN-CANEDO, V.; MORÁN-FERNÁNDEZ, L.; SÁNCHEZ-MAROÑO, N. **A review of microarray datasets: where to find them and specific characteristics**. *Microarray Bioinformatics*, p. 65–85, 2019.

[8] ALONSO-BETANZOS, A.; BOLÓN-CANEDO, V.; MORÁN-FERNÁNDEZ, L.; SÁNCHEZ-MAROÑO, N. **A Review of Microarray Datasets: Where to Find Them and Specific Characteristics**, p. 65–85. Springer New York, New York, NY, 2019.

[9] ALREFAI, N.; IBRAHIM, O. **Optimized feature selection method using particle swarm intelligence with ensemble learning for cancer classification based on**

microarray datasets. *Neural Computing and Applications*, 34(16):13513–13528, 2022.

[10] ANAISSI, A.; KENNEDY, P. J.; GOYAL, M. **Feature selection of imbalanced gene expression microarray data**. In: *2011 12th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, p. 73–78. IEEE, 2011.

[11] ARUN KUMAR, C.; SOORAJ, M.; RAMAKRISHNAN, S. **A comparative performance evaluation of supervised feature selection algorithms on microarray datasets**. *Procedia Computer Science*, 115:209–217, 2017. 7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India.

[12] AUGUSTINE, J.; JEREESH, A. **Blood-based gene-expression biomarkers identification for the non-invasive diagnosis of parkinson's disease using two-layer hybrid feature selection**. *Gene*, 823:146366, 2022.

[13] BASHEER, I. A.; HAJMEER, M. **Artificial neural networks: fundamentals, computing, design, and application**. *Journal of Microbiological Methods*, 43:3–31, 2000.

[14] BOLÓN-CANEDO, V.; SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A.; BENÍTEZ, J.; HERRERA, F. **A review of microarray datasets and applied feature selection methods**. *Information Sciences*, 282:111–135, 2014.

[15] BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. **A training algorithm for optimal margin classifiers**. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, p. 144–152, New York, NY, USA, 1992. Association for Computing Machinery.

[16] BRADLEY, A. P. **The use of the area under the roc curve in the evaluation of machine learning algorithms**. *Pattern Recogn.*, 30(7):1145–1159, July 1997.

[17] BREIMAN, L. **Random forests**. *Machine learning*, 45:5–32, 2001.

[18] BREMER, E. G.; NATARAJAN, J.; ZHANG, Y.; DESESA, C.; HACK, C. J.; DUBITZKY, W. **Text mining of full text articles and creation of a knowledge base for analysis of microarray data**. In: *Knowledge Exploration in Life Science Informatics: International Symposium KELSI 2004, Milan, Italy, November 25-26, 2004. Proceedings*, p. 84–95. Springer, 2004.

[19] BRON, E. E.; SMITS, M.; NIESSEN, W. J.; KLEIN, S. **Feature selection based on the svm weight vector for classification of dementia**. *IEEE Journal of Biomedical and Health Informatics*, 19(5):1617–1626, 2015.

[20] BUTTERWORTH, R.; SIMOVICI, D. A.; SANTOS, G. S.; OHNO-MACHADO, L. **A greedy algorithm for supervised discretization**. *Journal of Biomedical Informatics*, 37(4):285–292, 2004. Biomedical Machine Learning.

[21] CARUANA, R. **A non-parametric em-style algorithm for imputing missing values.** In: *in Proceedings of Artificial Intelligence and Statistics. 2001.*, 2001.

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

[23] CICHOSZ, P. **A case study in text mining of discussion forum posts: classification with bag of words and global vectors**. *International Journal of Applied Mathematics and Computer Science*, 28(4):787–801, 2018.

[24] COHEN, A. M.; HERSH, W. R. **A survey of current work in biomedical text mining**. *Briefings in bioinformatics*, 6(1):57–71, 2005.

[25] COHEN, K. B.; HUNTER, L. **Getting started in text mining**. *PLoS computational biology*, 4(1):e20, 2008.

[26] COVER, T.; HART, P. **Nearest neighbor pattern classification**. *IEEE Trans. Inf. Theor.*, 13(1):21–27, Sept. 2006.

[27] CSARDI, G.; NEPUSZ, T. **The igraph software package for complex network research**. *InterJournal*, Complex Systems:1695, 2006.

[28] CUETO-LÓPEZ, N.; GARCÍA-ORDÁS, M. T.; DÁVILA-BATISTA, V.; MORENO, V.; ARAGONÉS, N.; ALAIZ-RODRÍGUEZ, R. **A comparative study on feature selection for a risk prediction model for colorectal cancer**. *Computer methods and programs in biomedicine*, 177:219–229, 2019.

[29] CUNNINGHAM, P.; CORD, M.; DELANY, S. J. **Supervised learning**. *Machine learning techniques for multimedia: case studies on organization and retrieval*, p. 21–49, 2008.

[30] DA COSTA CÔRTES, S.; PORCARO, R. M.; LIFSCHITZ, S. **Mineração de dados-funcionalidades, técnicas e abordagens**. PUC, 2002.

[31] DANIYA, T.; GEETHA, M.; KUMAR, K. S. **Classification and regression trees with gini index**. *Advances in Mathematics: Scientific Journal*, 9(10):8237–8247, 2020.

[32] DEMSAR, J. **Statistical comparisons of classifiers over multiple data sets**. *J. Mach. Learn. Res.*, 7:1–30, Dec. 2006.

[33] DEVI D, R.; SS, S. **Online feature selection (ofs) with accelerated bat algorithm (aba) and ensemble incremental deep multiple layer perceptron (eidmlp) for big data streams**. 01 2019.

[34] DING, C.; PENG, H. **Minimum redundancy feature selection from microarray gene expression data**. In: *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, p. 523–528, 2003.

[35] DING, D.; YANG, X.; XIA, F.; MA, T.; LIU, H.; TANG, C. **Unsupervised feature selection via adaptive hypergraph regularized latent representation learning**. *Neurocomputing*, 378:79–97, 2020.

[36] DONG, X.; YU, Z.; CAO, W.; SHI, Y.; MA, Q. **A survey on ensemble learning**. *Frontiers of Computer Science*, 14:241–258, 2020.

[37] DONTHU, N.; KUMAR, S.; PANDEY, N.; LIM, W. M. **Research constituents, intellectual structure, and collaboration patterns in journal of international marketing: An analytical retrospective**. *Journal of International Marketing*, 29(2):1–25, 2021.

[38] DOS SANTOS, B. S.; STEINER, M. T. A.; FENERICH, A. T.; LIMA, R. H. P. **Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018**. *Computers & Industrial Engineering*, 138:106120, 2019.

[39] EGAN, J. P. **Signal detection theory and ROC analysis**. Academic Press, New York, NY, 1975.

[40] FAN, L.; SONG, B.; GU, X.; LIANG, Z. **Semi-supervised graph embedding-based feature extraction and adaptive kernel-based classification for computer-aided detection in ct colonography**. In: *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)*, p. 3983–3988, 2012.

[41] FARAHAT, A. K.; GHODSI, A.; KAMEL, M. S. **Efficient greedy feature selection for unsupervised learning**. *Knowledge and Information Systems*, 35:285–310, 2012.

[42] FARO, A.; GIORDANO, D.; SPAMPINATO, C. **Combining literature text mining with microarray data: advances for system biology modeling**. *Briefings in bioinformatics*, 13(1):61–82, 2012.

[43] FAWCETT, T. **An introduction to roc analysis**. *Pattern Recogn. Lett.*, 27(8):861–874, 2006.

[44] FAYYAD, U. M.; IRANI, K. B. **Multi-interval discretization of continuous-valued attributes for classification learning.** In: *IJCAI*, p. 1022–1029, 1993.

[45] FERREIRA-MELLO, R.; ANDRÉ, M.; PINHEIRO, A.; COSTA, E.; ROMERO, C. **Text mining in education**. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1332, 2019.

[46] FOSS, A. **High-dimensional data mining: subspace clustering, outlier detection and applications to classification**. 2010.

[47] GAGLIARDI, F. **Instance-based classifiers applied to medical databases: Diagnosis and knowledge extraction.** *Artificial Intelligence in Medicine*, 52(3):123–139, 2011.

[48] GAYATRI, N.; NICKOLAS, S.; REDDY, A.; REDDY, S.; NICKOLAS, A. **Feature selection using decision tree induction in class level metrics dataset for software defect predictions**. In: *Proceedings of the world congress on engineering and computer science*, volume 1, p. 124–129, 2010.

[49] GLAAB, E.; BACARDIT, J.; GARIBALDI, J. M.; KRASNOGOR, N. **Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data**. *PLoS ONE*, 7(7):e39932, July 2012.

[50] GREEN, D. M.; SWETS, J. A. **Signal Detection Theory and Psychophysics**. Wiley, New York, 1966.

[51] GU, Q.; LI, Z.; HAN, J. **Generalized fisher score for feature selection**. UAI'11, p. 266–273, Arlington, Virginia, USA, 2011. AUAI Press.

[52] GUYON, I.; WESTON, J.; BARNHILL, S.; VAPNIK, V. **Gene selection for cancer classification using support vector machines**. *Machine learning*, 46(1):389–422, 2002.

[53] HALL, M. A.; SMITH, L. A. **Practical feature subset selection for machine learning**. 1998.

[54] HALL, M. A. **Correlation-based feature selection for machine learning**. 1999.

[55] HANCER, E.; XUE, B.; ZHANG, M. **Differential evolution for filter feature selection based on information theory and feature ranking**. *Knowledge-Based Systems*, 140:103–119, 1 2018.

[56] HAND, D. J.; TILL, R. J. **A simple generalisation of the area under the roc curve for multiple class classification problems.** *Machine Learning*, 45(2):171–186, 2001.

[57] HANLEY, J. A.; MCNEIL, B. J. **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** *Radiology*, 143:29–36, 1982.

[58] HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[59] HAYKIN, S. **Neural Networks: A Comprehensive Foundation**. Prentice Hall, 1999.

[60] HE, X.; CAI, D.; NIYOGI, P. **Laplacian score for feature selection**. *Advances in neural information processing systems*, 18, 2005.

[61] HUANG, C.-L.; TSAI, C.-Y. **A hybrid sofm-svr with a filter-based feature selection for stock market forecasting**. *Expert Systems with Applications*, 36(2, Part 1):1529–1539, 2009.

[62] INZA, I.; SIERRA, B.; BLANCO, R.; LARRAÑAGA, P. **Gene selection by sequential search wrapper approaches in microarray cancer class prediction**. *Journal of Intelligent & Fuzzy Systems*, 12(1):25–33, 2002.

[63] JIAO, Y.; DU, P. **Performance measures in evaluating machine learning based bioinformatics predictors for classifications**. *Quantitative Biology*, 4:320–330, 2016.

[64] JORDAN, M. I.; MITCHELL, T. M. **Machine learning: Trends, perspectives, and prospects**. *Science*, 349(6245):255–260, 2015.

[65] KAFRAWY, P. E.; FATHI, H.; QARAAD, M.; KELANY, A. K.; CHEN, X. **An efficient svm-based feature selection model for cancer classification using high-dimensional microarray data**. *IEEE Access*, 9:155353–155369, 2021.

[66] KALOUSIS, A.; PRADOS, J.; HILARIO, M. **Stability of feature selection algorithms: a study on high-dimensional spaces**. *Knowledge and information systems*, 12:95–116, 2007.

[67] KANNAN, S.; GURUSAMY, V.; VIJAYARANI, S.; ILAMATHI, J.; NITHYA, M.; KANNAN, S.; GURUSAMY, V. **Preprocessing techniques for text mining**. *International Journal of Computer Science & Communication Networks*, 5(1):7–16, 2014.

[68] KÖNIGSTORFER, F.; THALMANN, S. **Applications of artificial intelligence in commercial banks–a research agenda for behavioral finance**. *Journal of behavioral and experimental finance*, 27:100352, 2020.

[69] KÖNIGSTORFER, F.; THALMANN, S. **Applications of artificial intelligence in commercial banks–a research agenda for behavioral finance**. *Journal of behavioral and experimental finance*, 27:100352, 2020.

[70] KONONENKO, I. **Estimating attributes: Analysis and extensions of relief**. In: *European conference on machine learning*, p. 171–182. Springer, 1994.

[71] KUMAR, V.; KUMAR, D. **Automatic clustering and feature selection using gravitational search algorithm and its application to microarray data analysis**. *Neural Computing and Applications*, 31(8):3647–3663, 2019.

[72] LABANI, M.; MORADI, P.; AHMADIZAR, F.; JALILI, M. **A novel multivariate filter method for feature selection in text classification problems**. *Engineering Applications of Artificial Intelligence*, 70:25–37, 2018.

[73] LAN, L.; VUCETIC, S. **Improving accuracy of microarray classification by a simple multi-task feature selection filter**. *International journal of data mining and bioinformatics*, 5(2):189–208, 2011.

[74] LEE, T.-H.; ULLAH, A.; WANG, R. **Bootstrap aggregating and random forest**. *Macroeconomic forecasting in the era of big data: Theory and practice*, p. 389–429, 2020.

[75] LI, L.; DARDEN, T. A.; WEINGBERG, C.; LEVINE, A.; PEDERSEN, L. G. **Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method**. *Combinatorial chemistry & high throughput screening*, 4(8):727–739, 2001.

[76] LING, C. X.; HUANG, J.; ZHANG, H. **Auc: A statistically consistent and more discriminating measure than accuracy**. In: *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, p. 519–524, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

[77] LIU, H.; YU, L. **Toward integrating feature selection algorithms for classification and clustering**. *IEEE Transactions on knowledge and data engineering*, 17(4):491–502, 2005.

[78] LIU, Q.; SUNG, A. H.; CHEN, Z.; LIU, J.; HUANG, X.; DENG, Y. **Feature selection and classification of maqc-ii breast cancer and multiple myeloma microarray gene expression data**. *PloS one*, 4(12):e8250, 2009.

[79] LIU, Y.; NIE, F.; GAO, Q.; GAO, X.; HAN, J.; SHAO, L. **Flexible unsupervised feature extraction for image classification**. *Neural Networks*, 115:65–71, 2019.

[80] LIU, Y.; WANG, Y.; ZHANG, J. **New machine learning algorithm: Random forest**. In: *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3*, p. 246–252. Springer, 2012.

[81] LIU, Z.; BENSMAIL, H.; TAN, M. **Efficient feature selection and multiclass classification with integrated instance and model based learning**. *Evolutionary Bioinformatics*, 8:EBO.S9407, 2012. PMID: 22577297.

[82] LORENA, A.; COGNIÇÃO, C.; CARVALHO, A. **Uma introdução às support vector machines**. *Revista de Informática Teórica e Aplicada; Vol. 14, No 2 (2007); 43-67*, 14, 12 2007.

[83] MA, S.; HUANG, J. **Regularized roc method for disease classification and biomarker selection with microarray data**. *Bioinformatics*, 21(24):4356–4362, 2005.

[84] MAHESH, B. **Machine learning algorithms -a review**, 01 2019.

[85] MALDONADO, S.; WEBER, R.; BASAK, J. **Simultaneous feature selection and classification using kernel-penalized support vector machines**. *Information Sciences*, 181(1):115–128, 2011.

[86] MAMITSUKA, H. **Selecting features in microarray classification using roc curves**. *Pattern Recognition*, 39(12):2393–2404, 2006. Bioinformatics.

[87] MANOHARAN, S.; IYYAPPAN, O. R. **A hybrid protocol for finding novel gene targets for various diseases using microarray expression data analysis and text mining**. In: *Biomedical Text Mining*, p. 41–70. Springer, 2022.

[88] MANTAS, C. J.; ABELLAN, J. **Credal-c4. 5: Decision tree based on imprecise probabilities to classify noisy data**. *Expert Systems with Applications*, 41(10):4625–4637, 2014.

[89] MHLANGA, D. **Industry 4.0 in finance: the impact of artificial intelligence (ai) on digital financial inclusion**. *International Journal of Financial Studies*, 8(3):45, 2020.

[90] MITCHELL, T. M. **Machine Learning**. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[91] MYLNE, K. R. **Decision-making from probability forecasts based on forecast value**. 9:307–315, 2002.

[92] NATARAJAN, J. **Text mining perspectives in microarray data mining**. *International Scholarly Research Notices*, 2013, 2013.

[93] NIU, J.; TANG, W.; XU, F.; ZHOU, X.; SONG, Y. **Global research on artificial intelligence from 1990–2014: Spatially-explicit bibliometric analysis**. *ISPRS International Journal of Geo-Information*, 5(5):66, 2016.

[94] PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. **Scikit-learn: Machine learning in Python**. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[95] PENG, H.; LONG, F.; DING, C. **Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy**. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[96] PENG, W.; CHEN, J.; ZHOU, H. **An implementation of id3-decision tree learning algorithm**. *From web. arch. usyd. edu. au/wpeng/DecisionTree2. pdf Retrieved date: May*, 13, 2009.

[97] PRATI, R. C.; BATISTA.; MONARD, M. C. **Curvas ROC para avaliação de classificadores**. *Revista IEEE América Latina*, 2008.

[98] PRIYA, S.; LIU, Y.; WARD, C.; LE, N.; SONI, N.; PILLENAHALLI MAHESHWARAPPA, R.; MONGA, V.; ZHANG, H.; SONKA, M.; BATHLA, G. **Radiomic based machine learning performance for a three class problem in neuro-oncology: Time to test the waters?** *Cancers*, 13:2568, 05 2021.

[99] R CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[100] RAILEANU, L.; STOFFEL, K. **Theoretical comparison between the gini index and information gain criteria**. *Annals of Mathematics and Artificial Intelligence*, 41:77–93, 05 2004.

[101] RAILEANU, L. E.; STOFFEL, K. **Theoretical comparison between the gini index and information gain criteria**. *Annals of Mathematics and Artificial Intelligence*, 41:77–93, 2004.

[102] RAJA, K. **Biomedical Text Mining**. Springer, 2022.

[103] REITERMANOVA, Z.; OTHERS. **Data splitting**. In: *WDS*, volume 10, p. 31–36. Matfyzpress Prague, 2010.

[104] REMESEIRO, B.; BOLON-CANEDO, V. **A review of feature selection methods in medical applications**, 9 2019.

[105] REN, S.; CAO, X.; WEI, Y.; SUN, J. **Global refinement of random forest**. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 723–730, 2015.

[106] RIBEIRO, G.; GONÇALVES, C.; VICTOR DOS SANTOS, P.; MELGAÇO BARBOSA, R. **Feature selection method auc-based with estimation probability and smoothing**. In: *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, p. 1–8, 2021.

[107] RIBEIRO, G. A. S.; DE OLIVEIRA, A. C. M.; FERREIRA, A. L. S.; VISWESWARAN, S.; COOPER, G. F. **Patient-specific modeling of medical data**. In: Perner, P., editor, *Machine Learning and Data Mining in Pattern Recognition*, p. 415–424, Cham, 2015. Springer International Publishing.

[108] SAITO, T.; REHMSMEIER, M. **The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets**. *PloS one*, 10(3):e0118432, 2015.

[109] SAQIB, S. M.; AHMAD, S.; SYED, A. H.; NAEEM, T.; ALOTAIBI, F. M. **Analysis of latent dirichlet allocation and non-negative matrix factorization using latent semantic indexing**. *International Journal of Advanced and Applied Sciences*, 6(10):94–102, 2019.

[110] SHAH, M.; MARCHAND, M.; CORBEIL, J. **Feature selection with conjunctions of decision stumps and learning from microarray data**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):174–186, 2012.

[111] SHARMA, A.; IMOTO, S.; MIYANO, S. **A top-r feature selection algorithm for microarray gene expression data**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(3):754–764, 2011.

[112] SHI, Y.; JIE TIAN, Y.; KOU, G.; PENG, Y.; LI, J. **Optimization based data mining: Theory and applications**. In: *Advanced Information and Knowledge Processing*, 2011.

[113] SHIZHI, Z.; MINGJIN, Z. **Use of SVM-based ensemble feature selection method for gene expression data analysis**. *Statistical Applications in Genetics and Molecular Biology*, 21(1):1–10, January 2022.

[114] SILVA, A.; CARVALHO, P.; GATTASS, M. **Diagnosis of solitary lung nodule using texture e geometry in computerized tomography images: Preliminary results**. 2:75–80, 2004.

[115] SILVA, M. M. **Uma abordagem evolucionária para o aprendizado semi-supervisionado em máquinas de vetores de suporte**. Mestrado, PPGEE, Programa de Pós-Graduação em Engenharia Elétrica, UFMG, 2008.

[116] SIM, Y.; LEE, S. K.; SUTHERLAND, I. **The impact of latent topic valence of online reviews on purchase intention for the accommodation industry**. *Tourism Management Perspectives*, 40:100903, 2021.

[117] SOARES, F.; ANZANELLO, M. J.; FOGLIATTO, F. S.; MARCELO, M. C.; FERRÃO, M. F.; MANFROI, V.; POZEBON, D. **Element selection and concentration analysis for classifying south america wine samples according to the country of origin**. *Computers and electronics in agriculture*, 150:33–40, 2018.

[118] SPACKMAN, K. A. **Signal detection theory: Valuable tools for evaluating inductive learning**. In: *Proceedings of the Sixth International Workshop on Machine Learning*, p. 160–163, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

[119] STEINBERG, D. **Cart: classification and regression trees**. In: *The top ten algorithms in data mining*, p. 193–216. Chapman and HallCRC, 2009.

[120] STERNECK, R.; POLISH, A.; BOWERN, C. **Topic modeling in the voynich manuscript**. *arXiv preprint arXiv:2107.02858*, 2021.

[121] SUN, L.; WANG, J.; WEI, J. **Avc: Selecting discriminative features on basis of auc by maximizing variable complementarity**. *BMC Bioinformatics*, 18, 03 2017.

[122] SUN, L.; WANG, J.; WEI, J. **Avc: Selecting discriminative features on basis of auc by maximizing variable complementarity**. *BMC Bioinformatics*, 18, 03 2017.

[123] SUN, W.; CAI, Z.; LI, Y.; LIU, F.; FANG, S.; WANG, G. **Data processing and text mining technologies on electronic medical records: a review**. *Journal of healthcare engineering*, 2018, 2018.

[124] TADIST, K.; NAJAH, S.; NIKOLOV, N.; MRABTI, F.; ZAHI, A. **Feature selection methods and genomic big data: a systematic review**. *Journal of Big Data*, 6, 08 2019.

[125] TANG, C.; LIU, X.; LI, M.; WANG, P.; CHEN, J.; WANG, L.; LI, W. **Robust unsupervised feature selection via dual self-representation and manifold regularization**. *Knowledge-Based Systems*, 145:109–120, 2018.

[126] TANG, E.; SUGANTHAN, P.; YAO, X. **Gene selection algorithms for microarray data based on least square support vector machine**. *BMC bioinformatics*, 7:95, 02 2006.

[127] TANG, X.; DAI, Y.; XIANG, Y. **Feature selection based on feature interactions with application to text categorization**. *Expert Systems with Applications*, 120:207–216, 2019.

[128] VERMA, S.; GUSTAFSSON, A. **Investigating the emerging covid-19 research trends in the field of business and management: A bibliometric analysis approach**. *Journal of Business Research*, 118:253–261, 2020.

[129] VISWESWARAN, S.; FERREIRA, A.; RIBEIRO, G. A.; OLIVEIRA, A. C.; COOPER, G. F. **Personalized modeling for prediction with decision-path models**. *PLOS ONE*, 10(6):1–15, 06 2015.

[130] VRIGAZOVA, B. **The proportion for splitting data into training and test set for the bootstrap in classification problems**. *Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy*, 12(1):228–242, 2021.

[131] WAKADE, S.; SHEKAR, C.; LISZKA, K. J.; CHAN, C.-C. **Text mining for sentiment analysis of twitter data**. In: *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, p. 1. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2012.

[132] WANG, H.; TAN, L.; NIU, B. **Feature selection for classification of microarray gene expression cancers using bacterial colony optimization with multidimensional population**. *Swarm and Evolutionary Computation*, 48:172–181, 2019.

[133] WANG, H.; ZHANG, Y.; ZHANG, J.; LI, T.; PENG, L. **A factor graph model for unsupervised feature selection**. *Information Sciences*, 480:144–159, 2019.

[134] WANG, J.; ZHAO, P.; HOI, S. C.; JIN, R. **Online feature selection and its applications**. *IEEE Transactions on Knowledge and Data Engineering*, 26(3):698–710, 2014.

[135] WANG, R.; TANG, K. **Feature selection for maximizing the area under the roc curve**. In: *2009 IEEE International Conference on Data Mining Workshops*, p. 400–405. IEEE, 2009.

[136] WANG, Y.; TETKO, I. V.; HALL, M. A.; FRANK, E.; FACIUS, A.; MAYER, K. F.; MEWES, H. W. **Gene selection from microarray data for cancer classification—a machine learning approach**. *Computational biology and chemistry*, 29(1):37–46, 2005.

[137] WANG, Y.; MAKEDON, F. **Application of relief-f feature filtering algorithm to selecting informative genes for cancer classification using microarray data**. In: *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.*, p. 497–498. IEEE, 2004.

[138] WEBB, G. I. **Naïve Bayes**, p. 713–714. Springer US, Boston, MA, 2010.

[139] WEINBERGER, K. Q.; SAUL, L. K. **Distance metric learning for large margin nearest neighbor classification**. *J. Mach. Learn. Res.*, 10:207–244, June 2009.

[140] WEN CHEN, X.; WASIKOWSKI, M. **Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems**. In: *KDD*, p. 124–132, 2008.

[141] WILCOXON, F. **Individual comparisons by ranking methods**. *Biometrics Bulletin*, 1(6):80–83, 1945.

[142] XIONG, Y.; YE, M.; WU, C. **Cancer classification with a cost-sensitive naive bayes stacking ensemble**. *Computational and Mathematical Methods in Medicine*, 2021, 2021.

[143] XU, J.-W.; SUZUKI, K. **Maximal partial auc feature selection in computer-aided detection of hepatocellular carcinoma in contrast-enhanced hepatic ct**. *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, 8315:15–, 02 2012.

[144] XU, Y.; PAPAKONSTANTINOU, Y. **Efficient lca based keyword search in xml data**. In: *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, p. 1007–1010, New York, NY, USA, 2007. Association for Computing Machinery.

[145] YANG, B.; HUANG, X.; QIN, G. **Variable selection in roc curve analysis with focused information criteria**. *Statistics and Its Interface*, 10:229–238, 01 2017.

[146] YANG, P.; HWA YANG, Y.; B ZHOU, B.; Y ZOMAYA, A. **A review of ensemble methods in bioinformatics**. *Current Bioinformatics*, 5(4):296–308, 2010.

[147] YERUKALA SATHIPATI, S.; HO, S.-Y. **Identifying a mirna signature for predicting the stage of breast cancer**. *Scientific reports*, 8(1):16138, 2018.

[148] YUAN, X.; YUAN, J.; JIANG, T.; AIN, Q. U. **Integrated long-term stock selection models based on feature selection and machine learning algorithms for china stock market**. *IEEE Access*, 8:22672–22685, 2020.

[149] ZADROZNY, B.; ELKAN, C. **Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers**. In: *In Proceedings of the Eighteenth International Conference on Machine Learning*, p. 609–616. Morgan Kaufmann, 2001.

[150] ZHANG, H.; ZHANG, R.; NIE, F.; LI, X. **An efficient framework for unsupervised feature selection**. *Neurocomputing*, 366:194–207, 2019.

[151] ZHANG, X.; WU, G.; DONG, Z.; CRAWFORD, C. **Embedded feature-selection support vector machine for driving pattern recognition**. *Journal of the Franklin Institute*, 352(2):669–685, 2015. Special Issue on Control and Estimation of Electrified vehicles.

[152] ZHOU, X.-H.; OBUCHOWSKI, N. A.; MCCLISH, D. K. **Statistical Methods in Diagnostic Medicine**. John Wiley & Sons, Inc., 2011.

[153] ZHU, F.; PATUMCHAROENPOL, P.; ZHANG, C.; YANG, Y.; CHAN, J.; MEECHAI, A.; VONGSANGNAK, W.; SHEN, B. **Biomedical text mining and its applications in cancer research**. *Journal of biomedical informatics*, 46(2):200–211, 2013.

[154] ZHU, Z.; ONG, Y.-S.; DASH, M. **Markov blanket-embedded genetic algorithm for gene selection**. *Pattern Recogn.*, 40(11):3236–3248, Nov. 2007.

[155] ZIA, A.; AZIZ, M.; POPA, I.; KHAN, S. A.; HAMEDANI, A. F.; ASIF, A. R. **Artificial intelligence-based medical data mining**. *Journal of Personalized Medicine*, 12(9), 2022.

APÊNDICE A

Resultados da metodologia *bootstrap*

Tabela A.1: Resultados obtidos (AUC) para os algoritmos ***Naive Bayes***, **SVM** e **kNN** utilizando a metodologia *bootstrap*.

Dados		**Colon**			**Leukemia**			**Lymphoma**			**CNS**	
Alg,	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**
NB	90,04	90,33	86,84	67,75	96,08	99,91	86,65	93,17	50,00	84,00	86,76	68,31
	88,37	86,76	89,45	79,27	95,80	99,93	92,02	93,52	50,00	84,47	88,43	68,31
	87,95	85,05	89,86	79,27	96,81	99,82	89,67	90,97	86,90	84,47	87,60	68,31
	87,87	83,64	89,79	87,78	96,43	99,94	88,16	89,24	86,90	84,78	85,10	68,31
	87,34	83,60	90,04	87,78	96,77	99,95	87,58	88,46	86,90	85,86	85,10	68,31
	87,18	84,13	88,83	87,78	96,68	99,96	91,26	86,86	87,26	84,10	85,10	68,31
	87,32	84,39	88,81	95,27	95,01	99,97	90,52	87,61	87,26	89,39	85,10	68,31
	87,22	84,52	89,09	95,16	95,62	99,95	90,46	86,98	87,26	87,72	85,10	68,31
	87,22	84,10	88,74	95,16	96,08	99,86	89,32	85,96	89,34	86,48	85,10	68,31
	87,22	83,63	88,98	95,16	97,01	99,94	88,20	86,05	87,56	86,05	85,10	68,31
	87,22	83,10	88,74	95,16	96,62	99,86	87,62	85,53	87,55	86,05	85,10	68,31
	87,22	83,36	88,78	95,16	96,69	99,94	86,79	85,79	86,58	86,05	85,10	68,31
	87,22	82,63	88,98	93,90	96,53	99,90	85,54	85,35	86,58	85,97	85,10	68,31
	87,07	82,82	88,77	93,90	96,98	99,89	86,40	85,44	86,58	85,58	85,58	68,31
	86,90	82,58	88,23	93,90	96,39	99,89	85,11	84,16	86,08	85,66	86,55	68,31
	86,90	82,88	88,13	93,90	96,67	99,89	85,11	84,12	88,29	85,66	86,55	68,31
	86,57	81,35	88,18	93,90	96,41	99,85	84,53	83,40	87,14	85,61	86,55	68,31
	86,32	81,02	88,54	93,96	96,53	99,86	84,29	82,37	88,26	85,61	86,55	83,85
	86,32	80,64	88,80	93,96	96,52	99,81	84,29	82,08	88,26	85,53	86,55	83,85
	86,32	80,21	84,64	93,96	96,45	99,87	83,62	81,95	86,75	85,53	86,55	83,85
SVM	90,30	93,58	50,00	50,00	99,93	50,00	79,94	98,52	50,00	72,99	84,57	43,36
	88,08	92,96	50,00	50,00	99,96	50,00	99,25	99,09	50,00	69,32	88,47	42,38
	88,24	94,22	44,48	50,00	99,93	50,00	99,34	99,30	84,09	68,81	88,85	50,71
	90,04	93,60	41,16	60,01	99,99	50,00	99,37	99,44	80,02	83,61	86,36	50,24
	88,39	92,91	53,04	80,68	100,00	50,00	99,84	99,38	81,49	79,81	84,60	42,82
	89,28	92,85	40,28	77,17	100,00	50,00	99,87	99,36	88,46	79,18	83,72	49,26
	88,51	92,77	57,75	84,05	100,00	61,65	99,89	99,33	89,85	81,66	82,84	50,16
	88,56	93,87	50,65	91,54	100,00	59,54	99,88	99,32	90,91	83,03	82,84	41,83
	88,69	93,86	51,33	94,62	100,00	54,62	99,83	99,15	92,90	86,57	80,24	42,03
	88,75	92,90	55,27	93,78	100,00	53,90	99,82	99,25	94,26	83,03	82,70	44,93
	90,38	93,84	57,23	96,58	100,00	66,54	99,85	99,30	96,78	78,02	78,34	50,80
	88,88	93,88	61,80	96,58	100,00	81,65	99,81	99,19	96,72	77,28	77,82	42,45
	89,87	93,07	58,80	93,67	100,00	75,73	99,79	99,14	96,72	83,40	79,02	51,49
	89,46	93,08	85,53	96,59	100,00	74,48	99,78	99,10	96,72	80,98	79,12	41,80
	87,65	94,14	86,73	94,90	100,00	82,50	99,77	99,05	97,48	82,63	74,39	46,07
	88,57	93,13	87,31	96,02	100,00	82,49	99,76	99,01	97,76	81,96	71,68	47,81
	88,52	92,19	82,85	93,03	100,00	86,66	99,76	99,00	97,55	79,97	77,06	41,91
	89,30	93,12	79,27	95,63	100,00	88,50	99,76	98,94	98,30	79,22	81,28	60,95
	89,30	93,24	84,95	95,59	100,00	88,62	99,77	98,85	98,30	82,40	76,60	62,04
	89,30	94,04	78,31	95,80	100,00	92,93	99,77	98,82	97,84	80,93	74,25	60,73
kNN	87,18	89,78	50,00	50,00	98,87	50,00	85,82	95,85	50,00	79,23	82,39	65,68
	85,42	90,30	50,00	50,00	99,78	50,00	95,06	97,12	50,00	79,23	86,60	65,68
	85,59	90,53	64,73	50,00	99,04	50,00	96,07	97,00	83,03	79,23	86,95	65,68
	84,59	89,89	64,73	65,30	100,00	50,00	97,78	97,54	83,03	85,63	86,96	65,68
	83,76	90,69	64,73	77,31	100,00	50,00	99,08	97,55	83,03	85,63	86,96	65,68
	83,70	89,76	64,83	77,31	99,97	50,00	99,24	97,87	85,76	85,63	86,96	65,68
	83,40	89,53	74,25	81,86	100,00	74,35	99,26	97,67	85,76	88,18	86,96	65,68
	83,67	89,94	74,25	84,41	100,00	74,35	99,22	97,92	85,76	88,18	86,96	65,68
	83,67	89,68	75,73	89,41	100,00	74,35	99,17	98,04	87,87	91,54	86,96	65,68
	83,67	89,75	76,01	90,03	100,00	74,35	99,24	98,09	91,09	91,54	86,96	65,68
	83,67	89,29	76,00	90,86	100,00	79,04	99,24	98,15	93,40	91,54	86,96	65,68
	83,67	88,81	76,00	90,86	100,00	85,65	99,20	98,18	95,07	91,54	86,96	65,68
	83,67	88,69	76,52	90,86	100,00	85,65	99,17	98,18	95,07	91,54	86,96	65,68
	84,07	88,10	89,12	90,86	100,00	85,65	99,18	98,19	95,07	91,54	86,46	65,68
	83,69	87,89	90,04	92,18	100,00	88,47	99,12	98,12	95,76	90,93	86,66	65,68
	83,69	88,09	89,93	91,25	99,83	88,47	99,03	98,06	96,54	90,93	86,66	65,68
	83,47	87,60	89,93	92,62	99,80	89,65	99,06	97,96	96,90	90,33	86,66	65,68
	83,36	87,59	89,93	94,29	99,97	90,46	98,96	97,73	97,58	90,05	86,66	77,80
	83,36	87,42	89,94	94,29	99,97	91,71	98,85	97,70	97,58	90,05	86,66	77,80
	83,36	87,37	89,94	94,55	99,91	96,30	98,86	97,58	97,11	90,05	86,66	77,80

Tabela A.2: Resultados obtidos (AUC) para os algoritmos **Árvore de Decisão** e **Florestas Aleatórias** utilizando a metodologia *bootstrap*.

Dados		**Colon**			**Leukemia**			**Lymphoma**			**CNS**	
Alg,	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**
AD	85,21	89,54	50,00	50,00	96,35	50,00	95,75	91,52	50,00	87,31	84,33	68,12
	75,33	80,12	50,00	50,00	94,22	50,00	91,63	90,03	50,00	87,31	85,49	68,12
	75,28	81,31	68,49	50,00	93,89	50,00	83,48	88,06	88,82	87,31	85,66	68,12
	74,96	79,24	68,49	50,00	93,08	50,00	81,11	88,10	88,82	84,28	84,45	68,12
	72,99	78,43	68,49	50,00	92,31	50,00	86,08	87,33	88,82	84,28	84,43	68,12
	70,25	76,63	70,19	50,00	93,15	50,00	86,56	87,34	88,23	84,14	84,16	68,12
	70,79	76,57	80,88	50,00	92,60	79,63	86,93	87,45	88,10	84,32	84,66	68,12
	71,38	77,31	80,88	50,00	93,12	79,63	86,99	86,98	88,26	84,31	84,38	68,12
	71,36	76,80	82,56	50,00	92,77	79,63	86,76	86,62	91,51	84,05	84,00	68,12
	70,27	76,88	79,92	50,00	92,83	79,63	87,26	85,91	91,00	84,24	84,33	68,12
	70,73	76,26	81,16	67,26	93,46	85,46	87,54	87,04	91,58	84,17	84,19	68,12
	70,92	75,30	81,16	67,26	92,56	89,18	87,20	86,31	90,47	84,14	84,27	68,12
	71,03	76,09	81,62	89,33	92,53	89,18	86,97	86,40	90,40	83,81	84,40	68,12
	70,95	76,72	88,26	89,33	91,96	89,18	86,49	86,73	90,55	84,05	81,56	68,12
	70,67	76,07	85,97	89,33	92,55	88,30	86,43	86,28	89,20	83,34	81,68	68,12
	70,82	75,65	86,73	89,33	92,31	88,36	86,32	86,29	92,88	83,23	81,44	68,12
	69,87	75,88	86,61	89,33	92,37	88,66	86,67	86,18	89,23	83,06	81,77	68,12
	70,25	75,55	86,65	92,88	92,61	86,71	86,22	86,08	92,95	83,93	81,76	84,05
	70,35	75,11	86,09	92,88	91,82	85,05	86,07	85,98	93,70	83,50	81,60	84,05
	69,89	75,64	86,48	94,45	91,60	87,30	86,25	85,68	92,15	83,74	81,75	84,05
FA	84,93	90,43	50,00	50,00	99,76	50,00	96,18	96,32	50,00	87,40	85,28	68,12
	82,93	89,92	50,00	50,00	99,83	50,00	98,46	97,23	50,00	87,28	89,20	68,12
	81,84	92,98	68,49	50,00	99,79	50,00	97,32	98,42	88,54	87,37	88,75	68,12
	82,71	90,89	68,49	50,00	99,87	50,00	97,95	98,47	88,30	88,23	89,41	68,12
	82,44	91,10	68,49	50,00	99,78	50,00	99,49	98,37	88,49	87,96	89,81	68,12
	80,99	90,88	70,94	50,00	99,75	50,00	99,53	98,09	89,81	88,56	89,87	68,12
	82,39	91,05	81,26	50,00	99,87	79,63	99,40	98,23	89,43	89,84	89,29	68,12
	82,61	91,61	81,26	50,00	99,77	79,63	99,30	97,94	89,21	90,13	90,38	68,12
	82,87	90,77	83,12	50,00	99,88	79,63	99,36	98,03	92,93	92,12	90,04	68,12
	82,49	91,43	80,96	50,00	99,89	79,63	99,36	97,94	94,98	92,53	89,63	68,12
	82,86	91,16	82,23	67,26	99,81	85,33	99,33	98,13	96,49	92,53	89,68	68,12
	82,11	90,69	82,90	67,26	99,88	89,96	99,25	98,10	96,85	92,25	89,23	68,12
	82,96	90,79	82,59	89,33	99,87	89,76	99,32	97,92	96,13	92,52	89,68	68,12
	82,29	90,27	91,22	89,33	99,80	89,96	99,32	98,21	96,80	93,02	89,61	68,12
	83,31	90,75	91,00	89,33	99,90	91,70	99,39	98,24	96,48	92,28	89,31	68,12
	82,76	90,60	90,85	89,33	99,72	92,26	98,90	98,32	97,68	92,19	89,85	68,12
	82,85	89,84	91,45	89,33	99,82	94,14	99,13	98,08	97,78	92,12	88,89	68,12
	82,47	90,38	90,87	93,13	99,73	93,44	99,12	97,65	98,63	92,30	89,50	84,30
	82,15	90,26	91,23	93,11	99,81	94,52	98,98	97,98	98,56	92,50	90,30	84,29
	82,84	90,52	91,56	95,81	99,80	97,66	99,14	98,09	97,82	92,81	90,36	84,35

Tabela A.3: Resultados obtidos (sensibilidade) para os algoritmos **Árvore de Decisão** e **Florestas Aleatórias** utilizando a metodologia *bootstrap*.

Dados		**Colon**			**Leukemia**			**Lymphoma**			**CNS**	
Alg,	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**
NB	96,61	82,97	0,00	0,00	97,09	0,00	93,79	93,01	0,00	99,02	93,02	100,00
	93,92	91,84	0,00	0,00	98,06	0,00	95,04	92,74	0,00	99,02	93,02	100,00
	92,18	85,63	36,95	0,00	98,22	0,00	95,34	95,04	89,40	99,02	93,02	100,00
	93,16	86,37	36,95	0,00	99,20	0,00	97,11	97,13	89,40	93,02	93,02	100,00
	92,89	88,64	36,95	0,00	99,25	0,00	100,00	97,36	89,40	93,02	93,02	100,00
	88,63	88,63	41,85	0,00	99,07	0,00	100,00	97,36	87,99	93,02	93,02	100,00
	88,47	88,63	41,85	0,00	98,00	26,40	100,00	97,36	87,99	93,02	93,02	100,00
	88,43	88,89	41,85	0,00	97,80	26,40	100,00	97,36	87,99	93,02	93,02	100,00
	88,43	91,58	49,51	0,00	97,59	26,40	100,00	97,36	72,67	93,02	93,02	100,00
	88,43	91,58	51,13	0,00	96,86	26,40	100,00	97,36	76,01	93,02	93,02	100,00
	88,43	92,06	59,24	100,00	96,86	91,35	100,00	98,50	77,04	93,02	93,02	100,00
	88,43	92,06	59,24	100,00	96,86	88,12	100,00	98,07	81,02	93,02	93,02	100,00
	88,43	93,47	60,12	91,35	96,88	88,12	100,00	98,07	81,02	93,02	93,02	100,00
	88,44	93,55	60,12	91,35	96,85	88,12	100,00	98,07	81,02	93,02	93,02	100,00
	89,40	93,63	63,98	91,35	95,85	91,32	100,00	98,07	82,34	93,02	93,02	100,00
	89,40	93,63	63,98	91,35	95,51	91,32	100,00	98,07	82,13	93,02	93,02	100,00
	89,37	93,63	63,98	91,35	94,82	89,73	100,00	98,30	83,12	93,02	93,02	100,00
	88,98	93,63	63,98	96,87	94,82	82,69	100,00	98,30	87,74	93,02	93,02	100,00
	88,98	93,63	63,98	96,87	95,85	86,34	100,00	98,25	87,74	93,02	93,02	100,00
	88,98	93,63	63,98	97,72	95,74	86,96	100,00	98,25	87,86	93,02	93,02	100,00
SVM	92,80	91,57	100,00	0,00	99,81	0,00	94,58	94,77	100,00	61,75	70,07	9,00
	94,05	92,93	100,00	0,00	99,91	0,00	96,38	97,21	100,00	23,65	76,54	0,00
	95,81	92,16	100,00	0,00	98,84	0,00	95,41	97,04	99,10	14,05	65,62	0,00
	94,58	90,77	100,00	0,00	100,00	0,00	96,45	97,02	99,86	32,77	54,61	0,00
	94,01	91,12	100,00	0,00	99,64	0,00	98,04	96,87	99,90	18,90	38,66	0,00
	94,20	90,76	100,00	0,00	99,54	0,00	98,80	96,32	99,81	14,03	21,59	0,00
	94,20	90,92	100,00	0,00	99,37	0,00	98,80	96,48	99,95	15,03	17,42	0,00
	94,39	89,14	100,00	0,00	99,29	0,00	98,80	96,43	100,00	11,35	16,17	0,00
	94,39	88,66	99,72	0,00	99,81	0,00	98,80	96,28	100,00	11,68	12,17	0,00
	94,39	88,49	99,10	0,00	99,81	0,00	98,80	96,04	100,00	8,00	9,00	0,00
	94,48	89,10	99,10	0,00	99,81	0,00	98,80	96,45	100,00	4,00	8,00	0,00
	94,48	89,78	99,10	0,00	99,60	0,00	98,80	96,45	100,00	2,00	4,33	0,00
	94,48	88,82	98,60	0,00	99,18	0,00	98,80	96,18	100,00	0,00	2,00	0,00
	94,18	89,26	99,09	0,00	99,05	0,00	98,80	96,29	100,00	0,00	2,00	0,00
	94,18	88,95	98,61	0,00	99,27	0,00	98,80	95,79	99,95	0,00	2,00	0,00
	94,29	89,11	98,61	0,00	98,50	0,00	98,80	95,42	99,91	0,00	1,00	0,00
	94,08	89,31	98,67	0,00	97,17	0,00	98,80	95,85	99,91	0,00	1,00	0,00
	94,29	89,15	98,67	0,00	98,56	1,52	98,80	95,59	100,00	0,00	1,00	0,00
	94,29	89,82	98,67	0,00	99,48	1,32	98,75	95,66	100,00	0,00	1,00	0,00
	94,29	90,01	98,67	0,00	98,34	3,13	98,75	95,70	99,91	0,00	0,00	0,00
KNN	91,40	88,41	70,00	21,00	100,00	21,00	95,72	94,45	87,00	69,77	72,31	43,00
	90,33	90,35	70,00	21,00	100,00	21,00	95,96	95,42	87,00	69,77	77,56	43,00
	93,86	85,75	69,51	21,00	97,41	21,00	95,52	95,95	89,97	69,77	77,56	43,00
	91,64	86,20	69,51	21,00	100,00	21,00	96,44	95,80	89,97	79,25	85,10	43,00
	92,06	84,41	69,51	21,00	99,91	21,00	96,45	96,18	89,97	79,25	85,10	43,00
	89,96	82,69	69,22	21,00	98,45	21,00	98,53	96,42	90,70	79,25	85,10	43,00
	89,86	81,22	73,77	21,00	98,34	49,40	98,80	96,06	90,70	84,67	85,10	43,00
	90,89	81,35	73,77	21,00	97,77	49,40	97,59	96,06	90,70	84,67	85,10	43,00
	90,89	80,58	72,77	21,00	97,87	49,40	96,82	96,01	90,50	93,42	85,10	43,00
	90,89	80,48	74,34	21,00	98,34	49,40	98,18	95,88	90,96	93,42	85,10	43,00
	90,89	80,74	73,64	41,00	98,12	58,15	98,19	95,07	92,79	93,42	85,10	43,00
	90,89	80,92	73,64	41,00	97,87	69,68	98,47	95,66	92,84	93,42	85,10	43,00
	90,89	79,10	73,58	68,68	97,87	69,68	98,04	95,75	92,84	93,42	85,10	43,00
	90,43	78,31	86,73	68,68	97,87	69,68	98,02	95,15	92,84	93,42	86,27	43,00
	90,32	76,51	87,49	68,68	98,12	74,02	97,87	95,11	93,87	93,85	86,43	43,00
	90,32	76,72	86,63	68,68	97,64	74,02	97,31	94,88	94,90	93,85	86,43	43,00
	90,32	76,44	86,63	68,68	97,72	78,87	97,76	95,25	94,90	94,23	86,43	43,00
	89,91	76,87	86,63	78,17	97,87	78,82	97,31	94,42	94,28	94,23	86,43	64,94
	89,91	77,13	86,63	78,17	98,19	85,99	96,95	94,21	94,28	94,23	86,43	64,94
	89,91	76,76	86,63	88,70	97,65	94,96	96,38	94,25	94,99	94,23	86,43	64,94

Tabela A.4: Resultados obtidos (sensibilidade) para os algoritmos **Árvore de Decisão** e **Florestas Aleatórias** utilizando a metodologia *bootstrap*.

Dados		**Colon**			**Leukemia**			**Lymphoma**			**CNS**	
Alg,	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**
AD	90,65	84,64	100,00	0,00	95,54	0,00	95,38	88,72	100,00	83,92	71,00	16,00
	88,76	84,69	100,00	0,00	91,44	0,00	93,57	92,08	100,00	83,92	72,16	16,00
	85,88	84,87	79,41	0,00	90,97	0,00	94,44	94,78	92,64	83,92	71,99	16,00
	82,57	86,38	79,41	0,00	90,36	0,00	95,37	94,32	92,64	77,58	78,40	16,00
	81,90	85,87	79,41	0,00	89,51	0,00	93,16	95,37	92,64	77,58	78,40	16,00
	76,73	84,16	72,39	0,00	89,93	0,00	92,98	95,55	92,27	77,41	78,40	16,00
	76,98	84,87	88,82	0,00	89,22	38,20	93,31	95,12	91,88	78,40	78,56	16,00
	77,89	84,14	88,82	0,00	90,36	38,20	93,22	94,91	92,34	78,56	78,23	16,00
	77,35	84,17	88,31	0,00	89,71	38,20	93,21	94,60	91,29	82,49	78,40	16,00
	76,19	84,73	80,71	0,00	90,28	38,20	94,68	92,89	88,84	82,21	78,56	16,00
	76,25	83,20	74,14	3,00	91,20	65,63	93,74	92,53	92,77	82,37	78,56	16,00
	76,38	82,12	74,14	3,00	90,31	89,36	93,72	93,27	92,01	82,04	78,40	16,00
	76,95	82,66	71,24	89,53	89,24	89,36	93,91	92,60	92,56	82,21	78,23	16,00
	78,06	84,00	82,81	89,53	88,63	89,36	94,08	92,31	92,17	82,37	78,40	16,00
	76,70	83,02	82,17	89,53	90,08	84,24	94,20	92,54	91,62	82,21	78,23	16,00
	78,08	83,37	82,24	89,53	89,69	84,36	93,54	92,33	94,02	82,04	78,56	16,00
	76,36	83,49	82,24	89,53	89,00	89,39	94,29	92,79	93,63	82,37	78,56	16,00
	76,26	82,51	82,09	91,35	89,88	75,44	93,48	92,25	95,24	82,21	78,40	62,78
	76,34	82,64	81,90	91,35	88,75	82,84	93,42	92,93	95,45	82,21	78,56	62,78
	75,55	84,27	81,99	94,03	88,55	85,37	94,04	92,50	94,97	82,21	78,40	62,78
FA	92,74	88,75	100,00	0,00	95,57	0,00	95,38	93,76	100,00	83,92	72,97	26,00
	91,83	86,78	100,00	0,00	95,41	0,00	96,43	95,65	100,00	83,92	74,51	21,00
	89,21	85,54	80,13	0,00	93,08	0,00	96,49	96,59	93,70	82,92	75,43	24,00
	88,04	86,23	80,59	0,00	93,30	0,00	96,41	96,37	93,44	80,46	81,71	26,00
	87,34	84,60	85,84	0,00	91,90	0,00	97,01	95,83	92,74	79,94	83,15	22,00
	85,24	83,36	77,08	0,00	93,45	0,00	97,59	96,38	93,46	79,58	81,52	26,00
	85,42	84,32	88,82	0,00	92,95	41,82	97,46	96,42	93,71	82,24	83,43	22,00
	86,61	84,83	88,82	0,00	92,72	44,56	97,74	96,06	94,40	81,40	82,37	27,00
	86,49	83,85	86,87	0,00	92,52	40,85	97,79	95,73	93,68	84,71	79,74	27,00
	87,16	83,81	81,39	0,00	91,22	46,96	97,75	95,05	91,66	84,09	81,65	22,00
	86,35	83,88	79,08	7,00	92,68	68,23	97,68	95,10	93,58	85,06	81,35	19,00
	86,35	84,17	78,13	7,00	92,34	92,27	96,85	94,86	93,77	85,03	80,98	18,00
	86,81	82,79	78,22	89,73	93,34	90,01	97,09	94,72	93,33	84,42	82,71	20,00
	85,54	83,85	85,81	89,90	91,98	90,34	97,37	95,56	94,22	84,48	81,44	19,00
	85,82	83,49	88,07	90,35	93,60	87,69	97,40	95,52	94,91	85,09	80,30	20,00
	86,15	83,37	87,06	89,73	91,38	82,68	97,41	94,91	95,65	84,46	81,01	27,00
	85,59	81,78	87,79	87,83	91,20	88,55	97,18	94,69	96,02	84,30	81,57	22,00
	84,99	83,61	87,63	90,47	92,38	78,10	96,65	95,05	95,75	84,42	81,67	62,78
	85,01	82,43	87,73	90,45	92,81	83,29	97,09	94,64	95,63	84,07	81,07	62,78
	85,37	84,25	87,48	93,75	91,53	86,89	96,58	95,16	95,46	84,87	80,48	62,78

Tabela A.5: Resultados obtidos (especificidade) para os algoritmos ***Naive Bayes***, **SVM** e **kNN** utilizando a metodologia *bootstrap*.

Dados		**Colon**			**Leukemia**			**Lymphoma**			**CNS**	
Alg,	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**
NB	70,45	85,44	100,00	100,00	97,72	100,00	94,27	88,53	100,00	57,85	57,85	36,17
	73,62	79,44	100,00	100,00	97,72	100,00	91,78	89,92	100,00	57,85	69,18	36,17
	74,64	92,46	100,00	100,00	97,91	100,00	87,92	87,16	61,81	57,85	70,80	36,17
	76,01	90,03	100,00	100,00	98,25	100,00	85,32	86,30	61,81	64,59	78,66	36,17
	75,98	92,79	100,00	100,00	99,39	100,00	85,11	83,84	61,81	64,59	78,66	36,17
	78,53	87,42	100,00	100,00	99,47	100,00	87,30	78,15	75,07	64,59	78,66	36,17
	78,53	87,42	100,00	100,00	100,00	96,16	86,61	78,14	75,07	78,66	78,66	36,17
	80,37	87,42	100,00	100,00	99,94	96,16	86,61	74,19	75,07	78,66	78,66	36,17
	80,37	86,72	100,00	100,00	99,94	96,16	83,86	73,99	88,15	89,33	78,66	36,17
	80,37	86,72	95,05	100,00	99,87	96,16	83,86	73,99	86,96	89,33	78,66	36,17
	80,37	84,83	95,05	34,53	99,87	69,70	81,21	73,85	89,97	89,33	78,66	36,17
	80,37	84,83	95,05	34,53	99,87	83,20	80,87	73,99	89,74	89,33	78,66	36,17
	80,37	84,83	95,05	80,84	99,87	83,20	78,96	73,65	89,74	89,33	78,66	36,17
	79,54	84,83	95,05	80,84	99,94	83,20	75,26	73,88	89,74	89,33	79,92	36,17
	78,62	84,83	91,52	80,84	99,65	83,11	73,72	73,05	86,96	89,33	83,59	36,17
	78,62	84,19	91,52	80,84	98,50	83,11	72,45	73,05	87,45	89,33	83,59	36,17
	78,01	84,19	91,52	80,84	98,62	85,11	72,11	72,45	86,42	89,33	83,59	36,17
	77,75	81,97	91,52	75,34	98,62	85,11	72,31	69,11	86,42	89,33	83,59	36,17
	77,75	82,60	91,52	75,34	98,62	87,13	72,31	69,11	86,42	89,33	83,59	36,17
	77,75	82,60	91,52	80,82	99,66	91,89	72,31	68,85	86,04	89,33	83,59	36,17
SVM	71,57	63,24	0,00	100,00	97,72	100,00	77,90	87,39	0,00	80,24	85,58	93,81
	69,22	73,91	0,00	100,00	97,72	100,00	98,50	91,57	0,00	91,03	86,77	100,00
	69,87	77,05	0,00	100,00	97,86	100,00	94,97	86,94	4,75	95,34	89,58	100,00
	69,44	75,93	0,00	100,00	98,64	100,00	97,95	93,56	1,00	93,94	91,44	100,00
	69,17	76,64	0,00	100,00	99,08	100,00	100,00	93,00	0,33	96,12	93,02	100,00
	68,54	78,14	0,00	100,00	99,45	100,00	99,86	92,75	2,17	97,71	95,08	100,00
	68,43	78,27	0,00	100,00	99,85	100,00	99,61	92,35	0,33	98,04	96,41	100,00
	68,04	80,98	0,00	100,00	99,58	100,00	100,00	91,94	0,00	98,81	97,43	100,00
	67,91	81,72	0,87	100,00	99,85	100,00	100,00	93,66	0,00	99,32	98,41	100,00
	67,64	81,72	1,87	100,00	99,85	100,00	99,66	94,01	0,00	99,60	99,09	100,00
	67,42	80,78	1,87	100,00	99,85	100,00	100,00	95,33	1,17	99,80	99,21	100,00
	66,96	78,78	1,87	100,00	99,85	100,00	99,53	94,21	0,83	99,87	99,61	100,00
	66,79	79,73	2,47	100,00	99,85	100,00	99,73	94,09	0,50	100,00	99,81	100,00
	66,96	78,78	2,60	100,00	100,00	100,00	100,00	94,15	0,00	100,00	99,81	100,00
	66,59	78,57	4,15	100,00	100,00	100,00	100,00	95,82	1,00	100,00	99,81	100,00
	66,49	78,74	4,15	100,00	99,85	100,00	99,73	96,31	2,67	100,00	99,94	100,00
	66,35	78,33	3,40	100,00	99,77	100,00	99,88	95,73	4,87	100,00	100,00	100,00
	65,90	78,23	2,80	100,00	99,77	100,00	100,00	96,27	5,62	100,00	100,00	100,00
	65,90	78,46	2,80	100,00	99,77	100,00	100,00	94,27	3,87	100,00	100,00	100,00
	65,70	75,92	2,80	100,00	99,77	100,00	100,00	94,97	5,85	100,00	100,00	100,00
KNN	71,80	66,63	30,00	79,00	97,72	79,00	84,88	83,11	13,00	79,23	79,22	72,15
	68,60	76,33	30,00	79,00	98,09	79,00	88,20	93,03	13,00	79,23	82,71	72,15
	72,02	84,29	47,00	79,00	98,42	79,00	80,96	92,82	60,74	79,23	83,03	72,15
	70,97	80,25	47,00	79,00	100,00	79,00	91,77	95,27	60,74	76,56	76,90	72,15
	69,95	83,60	47,00	79,00	100,00	79,00	100,00	95,38	60,74	76,56	76,90	72,15
	69,07	80,23	49,00	79,00	99,93	79,00	100,00	94,33	65,93	76,56	76,90	72,15
	69,07	79,30	60,95	79,00	100,00	82,49	100,00	94,52	65,93	80,25	76,90	72,15
	69,14	79,88	60,95	79,00	100,00	82,49	100,00	95,94	65,93	80,25	76,90	72,15
	69,14	81,63	63,05	79,00	100,00	82,49	100,00	94,76	71,33	81,57	76,90	72,15
	69,14	81,63	61,40	79,00	100,00	82,49	100,00	94,29	80,67	81,57	76,90	72,15
	69,14	82,13	63,05	73,04	100,00	81,96	100,00	96,77	81,69	81,57	76,90	72,15
	69,14	81,94	63,05	73,04	100,00	85,68	100,00	97,07	87,48	81,57	76,90	72,15
	69,14	82,23	65,21	86,49	100,00	85,68	100,00	97,33	87,48	81,57	76,90	72,15
	69,94	82,24	77,49	86,49	100,00	85,68	100,00	97,24	87,48	81,57	75,02	72,15
	70,09	84,07	80,35	86,49	100,00	86,23	100,00	96,46	90,41	78,99	75,29	72,15
	70,09	84,05	80,35	86,49	100,00	86,23	100,00	96,26	92,99	78,99	75,29	72,15
	70,21	83,66	80,35	86,49	100,00	85,89	100,00	96,06	94,30	77,41	75,29	72,15
	69,97	83,77	80,35	87,97	100,00	85,51	100,00	96,40	97,30	76,34	75,29	80,34
	69,97	83,11	80,35	87,97	100,00	86,96	100,00	97,20	97,30	76,34	75,29	80,34
	69,97	83,44	80,35	93,16	100,00	89,59	100,00	98,52	96,75	76,34	75,29	80,34

Tabela A.6: Resultados obtidos (especificidade) para os algoritmos **Árvore de Decisão** e **Florestas Aleatórias** utilizando a metodologia *bootstrap*.

Dados		**Colon**			**Leukemia**			**Lymphoma**			**CNS**	
Alg,	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**
AD	74,85	72,23	0,00	100,00	97,19	100,00	91,81	87,57	0,00	82,24	83,68	89,46
	68,19	74,49	0,00	100,00	96,96	100,00	84,62	85,87	0,00	82,24	86,44	89,46
	68,85	77,81	30,00	100,00	96,79	100,00	72,51	81,33	60,05	82,24	86,93	89,46
	66,15	72,17	30,00	100,00	95,83	100,00	66,89	81,91	60,05	79,29	80,64	89,46
	65,37	71,03	30,00	100,00	95,07	100,00	79,00	79,17	60,05	79,29	80,95	89,46
	62,54	69,17	42,00	100,00	96,31	100,00	79,96	79,00	68,11	79,36	80,57	89,46
	62,87	68,32	53,25	100,00	95,94	87,21	80,48	79,67	68,11	80,86	81,21	89,46
	62,66	70,53	53,25	100,00	95,82	87,21	80,59	78,95	68,11	80,48	80,90	89,46
	63,22	69,46	53,45	100,00	95,74	87,21	80,25	78,51	73,84	80,74	80,51	89,46
	62,15	69,09	56,47	100,00	95,31	87,21	79,78	78,75	93,86	81,35	80,51	89,46
	62,93	69,37	61,46	97,64	95,72	77,37	81,27	81,41	89,05	80,73	80,43	89,46
	63,12	68,53	61,46	97,64	94,79	82,48	80,58	79,21	87,29	81,32	80,87	89,46
	63,36	69,51	65,39	81,00	95,81	82,48	79,92	80,09	87,16	80,39	80,96	89,46
	62,27	69,49	82,54	81,00	95,21	82,48	78,84	81,01	87,36	81,02	77,43	89,46
	63,03	69,14	85,27	81,00	95,01	81,21	78,67	79,98	85,24	80,07	76,74	89,46
	62,28	68,00	86,44	81,00	94,92	81,21	79,03	80,25	90,76	79,93	76,20	89,46
	61,74	68,32	86,14	81,00	95,73	83,37	78,96	79,52	84,65	79,04	76,51	89,46
	62,43	68,63	86,45	86,93	95,28	84,09	78,88	79,85	90,62	80,57	77,04	92,21
	62,58	67,71	85,76	86,93	94,86	83,68	78,71	78,89	91,91	79,93	76,57	92,21
	62,40	67,06	86,37	94,12	94,65	85,17	78,41	78,82	89,36	80,21	77,12	92,21
FA	71,89	65,23	0,00	100,00	97,92	100,00	92,95	83,62	0,00	82,55	83,57	82,57
	66,10	73,52	0,00	100,00	98,43	100,00	85,32	90,44	0,00	83,15	86,60	85,57
	67,69	83,88	29,00	100,00	98,42	100,00	81,87	90,17	59,64	82,84	86,11	83,59
	66,78	78,06	28,00	100,00	98,82	100,00	86,10	89,84	59,68	83,68	84,67	82,57
	67,43	78,89	20,00	100,00	98,68	100,00	94,47	90,82	61,05	84,10	85,58	85,20
	65,94	77,75	35,00	100,00	98,95	100,00	93,56	88,91	64,39	83,64	85,30	82,39
	65,61	78,88	53,25	100,00	99,38	85,04	93,27	89,72	65,14	86,38	85,54	84,60
	66,11	79,73	52,55	100,00	99,04	82,92	94,39	89,01	64,27	85,64	85,48	81,92
	67,53	79,65	54,29	100,00	99,16	85,88	94,39	90,21	70,04	88,39	86,44	81,94
	67,18	79,60	56,62	100,00	99,16	81,80	93,55	89,32	85,60	88,97	85,88	85,02
	67,45	79,96	60,01	95,17	98,81	76,73	94,76	90,51	85,74	89,43	86,33	87,17
	65,94	79,66	60,37	95,00	98,96	83,08	92,58	91,06	87,28	89,07	85,99	88,31
	68,28	81,53	63,25	81,01	99,31	83,13	93,78	90,49	88,26	89,68	84,90	86,18
	67,42	81,85	79,41	80,93	99,18	82,85	92,54	90,01	88,32	89,25	84,87	87,14
	67,79	79,81	79,05	80,91	99,05	81,88	93,17	89,66	87,36	88,20	84,41	86,34
	67,78	81,14	80,47	81,01	98,86	82,59	92,29	90,59	89,22	88,75	84,90	81,89
	66,33	80,34	81,58	81,25	99,38	86,60	92,12	90,28	86,46	88,32	84,90	85,29
	65,50	78,59	79,63	87,43	99,10	88,16	91,55	88,36	91,28	87,90	84,57	93,01
	67,01	79,98	81,04	87,57	99,11	90,87	93,27	88,94	89,90	88,34	85,66	93,00
	67,14	79,16	81,41	94,75	99,14	95,66	91,83	88,99	86,71	88,09	85,64	93,28

Tabela A.7: Resultados obtidos (acurácia) para os algoritmos ***Naive Bayes***, **SVM** e **kNN** utilizando a metodologia *bootstrap*.

Dados		**Colon**			**Leukemia**			**Lymphoma**			**CNS**	
Alg,	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**
NB	87,00	84,16	36,84	64,36	97,32	64,36	93,67	92,00	24,33	71,56	69,44	58,06
	86,42	87,32	36,84	64,36	97,68	64,36	94,17	92,08	24,33	71,56	76,78	58,06
	85,68	88,21	60,21	64,36	97,86	64,36	93,46	92,96	82,88	71,56	77,89	58,06
	86,89	87,68	60,21	75,73	98,50	64,36	94,29	94,33	82,88	73,67	83,11	58,06
	86,74	90,00	60,21	83,64	99,23	64,36	96,29	93,92	82,88	73,67	83,11	58,06
	85,00	88,05	63,37	83,64	99,18	64,36	96,75	92,50	84,88	73,67	83,11	58,06
	84,89	88,05	63,37	83,59	99,09	71,50	96,58	92,62	84,88	83,11	83,11	58,06
	85,58	88,21	63,37	83,59	98,95	71,50	96,58	91,71	84,88	83,11	83,11	58,06
	85,58	89,68	68,26	83,59	98,86	71,50	95,87	91,67	76,54	90,22	83,11	58,06
	85,58	89,68	67,37	84,00	98,55	71,50	95,87	91,67	78,71	90,22	83,11	58,06
	85,58	89,32	72,42	87,00	98,55	77,45	95,21	92,50	80,17	90,22	83,11	58,06
	85,58	89,32	72,42	87,00	98,55	85,05	95,12	92,21	83,13	90,22	83,11	58,06
	85,58	90,26	72,95	87,00	98,59	85,05	94,71	92,12	83,13	90,22	83,11	58,06
	85,26	90,32	72,95	87,00	98,64	85,05	93,83	92,21	83,13	90,22	84,00	58,06
	85,53	90,37	73,89	86,09	98,05	86,18	93,46	91,96	83,33	90,22	86,44	58,06
	85,53	90,11	73,89	88,45	97,18	86,18	93,21	91,96	83,33	90,22	86,44	58,06
	85,32	90,11	73,89	88,00	97,05	86,95	93,12	91,96	83,83	90,22	86,44	58,06
	84,95	89,26	73,89	90,14	97,05	84,27	93,17	91,08	87,33	90,22	86,44	58,06
	84,95	89,53	73,89	90,14	97,36	87,09	93,17	91,04	87,33	90,22	86,44	58,06
	84,95	89,53	73,89	92,77	98,00	90,14	93,17	90,96	87,29	90,22	86,44	58,06
SVM	64,36	98,41	64,36	84,58	80,84	63,16	89,79	92,83	75,67	71,28	80,00	61,94
	64,36	98,45	64,36	84,79	85,68	63,16	96,75	95,79	75,67	64,17	81,78	65,78
	64,36	98,09	64,36	86,21	86,42	63,16	95,25	94,50	75,42	64,50	78,72	65,78
	64,36	99,09	64,36	85,26	84,89	63,16	96,71	96,17	75,67	69,28	75,61	65,78
	70,23	99,23	64,36	84,95	85,37	63,16	98,50	95,92	75,62	66,39	70,39	65,78
	68,09	99,41	64,36	84,79	85,84	63,16	99,04	95,42	75,71	66,50	66,39	65,78
	70,73	99,64	64,36	84,74	85,95	63,16	98,96	95,46	75,67	66,83	66,11	65,78
	71,82	99,41	64,36	84,68	85,79	63,16	99,08	95,29	75,67	66,61	66,61	65,78
	74,73	99,82	64,36	84,63	85,79	63,11	99,08	95,62	75,67	67,22	66,39	65,78
	73,82	99,82	64,36	84,53	85,68	62,74	99,00	95,54	75,67	66,61	66,33	65,78
	76,68	99,82	64,36	84,47	85,63	62,74	99,08	96,12	75,79	66,17	66,28	65,78
	75,05	99,73	64,36	84,26	85,26	62,74	98,96	95,87	75,75	65,89	66,06	65,78
	72,50	99,55	64,36	84,21	84,95	62,53	99,00	95,62	75,71	65,78	65,83	65,78
	70,86	99,59	64,36	84,11	84,89	62,89	99,08	95,62	75,67	65,78	65,83	65,78
	70,64	99,68	64,36	83,95	84,58	62,89	99,08	95,71	75,71	65,78	65,83	65,78
	70,95	99,27	64,36	83,95	84,74	62,89	99,00	95,54	75,83	65,78	65,78	65,78
	71,68	98,77	64,36	83,79	84,68	62,79	99,04	95,67	76,12	65,78	65,83	65,78
	71,64	99,27	64,68	83,68	84,53	62,63	99,08	95,62	76,29	65,78	65,83	65,78
	70,36	99,64	64,64	83,68	84,89	62,63	99,04	95,17	76,08	65,78	65,83	65,78
	74,86	99,18	65,05	83,58	84,00	62,63	99,04	95,37	76,33	65,78	65,78	65,78
KNN	84,00	80,00	53,26	57,64	98,50	57,64	92,50	91,38	68,08	75,56	76,22	60,83
	82,26	85,00	53,26	57,64	98,73	57,64	93,83	94,79	68,08	75,56	80,44	60,83
	85,74	84,95	59,53	57,64	98,00	57,64	91,79	95,08	82,88	75,56	80,61	60,83
	83,84	83,68	59,53	69,32	100,00	57,64	95,21	95,62	82,88	76,67	78,83	60,83
	83,89	83,74	59,53	78,36	99,95	57,64	97,29	95,96	82,88	76,67	78,83	60,83
	82,16	81,58	60,00	78,36	99,32	57,64	98,87	95,83	84,46	76,67	78,83	60,83
	82,05	80,21	68,26	76,23	99,36	70,45	99,08	95,62	84,46	80,94	78,83	60,83
	82,79	80,53	68,26	82,73	99,14	70,45	98,17	96,04	84,46	80,94	78,83	60,83
	82,79	80,68	68,53	86,82	99,18	70,45	97,58	95,62	85,50	85,17	78,83	60,83
	82,79	80,63	68,63	84,86	99,36	70,45	98,62	95,46	88,04	85,17	78,83	60,83
	82,79	81,00	68,74	86,00	99,27	73,09	98,62	95,42	90,00	85,17	78,83	60,83
	82,79	81,05	68,74	86,00	99,18	80,00	98,83	95,92	91,50	85,17	78,83	60,83
	82,79	80,00	69,58	86,00	99,18	80,00	98,50	96,04	91,50	85,17	78,83	60,83
	82,79	79,37	83,11	86,00	99,18	80,00	98,50	95,50	91,50	85,17	78,06	60,83
	82,79	78,95	84,47	86,73	99,27	81,45	98,37	95,25	92,92	83,67	78,33	60,83
	82,79	79,05	83,95	86,09	99,09	81,45	97,96	95,04	94,25	83,67	78,33	60,83
	82,84	78,74	83,95	86,41	99,14	82,95	98,29	95,33	94,58	82,83	78,33	60,83
	82,47	79,00	83,95	89,27	99,18	82,95	97,96	94,75	94,92	82,11	78,33	74,06
	82,47	78,89	83,95	89,27	99,32	86,41	97,71	94,83	94,92	82,11	78,33	74,06
	82,47	78,79	83,95	91,73	99,09	91,32	97,25	95,12	95,33	82,11	78,33	74,06

Tabela A.8: Resultados obtidos (acurácia) para os algoritmos **Árvore de Decisão** e **Florestas Aleatórias** utilizando a metodologia *bootstrap*.

Dados		**Colon**			**Leukemia**			**Lymphoma**			**CNS**	
Alg,	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**	**ARCO**	**FAST**	**AUC EPS**
AD	84,53	79,74	63,16	64,36	96,41	64,36	94,12	88,08	75,67	82,78	79,17	59,78
	81,11	80,95	63,16	64,36	94,68	64,36	91,04	90,46	75,67	82,78	81,33	59,78
	79,42	82,32	56,47	64,36	94,45	64,36	88,92	91,25	84,96	82,78	81,56	59,78
	76,21	81,21	56,47	64,36	93,36	64,36	88,38	91,08	84,96	78,06	79,17	59,78
	75,47	80,42	56,47	64,36	92,68	64,36	89,38	91,21	84,96	78,06	79,44	59,78
	71,16	78,68	56,42	64,36	93,73	64,36	89,62	91,29	86,33	78,06	79,17	59,78
	71,42	78,79	75,53	64,36	93,18	67,50	90,00	91,08	86,04	79,39	79,61	59,78
	71,89	79,00	75,53	64,36	93,50	67,50	89,96	90,75	86,38	79,22	79,28	59,78
	71,74	78,63	75,21	64,36	93,27	67,50	89,87	90,29	87,04	80,89	79,11	59,78
	70,68	78,89	70,95	64,36	93,23	67,50	90,92	89,08	89,96	81,11	79,17	59,78
	70,84	78,05	68,63	63,23	93,68	71,00	90,58	89,63	91,83	80,78	79,11	59,78
	71,00	77,05	68,63	63,23	92,73	84,91	90,42	89,54	90,96	81,06	79,33	59,78
	71,53	77,79	67,95	83,95	93,05	84,91	90,37	89,29	91,33	80,50	79,33	59,78
	71,95	78,42	82,47	83,95	92,55	84,91	90,38	89,42	91,08	81,00	77,00	59,78
	71,32	77,63	83,21	83,95	92,95	81,55	90,33	89,17	90,04	80,33	76,61	59,78
	71,79	77,47	83,74	83,95	92,64	81,59	90,00	89,25	93,12	80,22	76,22	59,78
	70,58	77,68	83,63	83,95	93,00	85,36	90,50	89,50	91,25	79,72	76,44	59,78
	70,79	77,26	83,63	88,59	93,00	80,64	89,88	89,08	94,04	80,61	76,83	81,89
	70,79	77,16	83,21	88,59	92,36	83,09	89,71	89,42	94,67	80,22	76,56	81,89
	70,26	77,84	83,53	94,05	92,05	85,00	90,17	89,08	93,58	80,39	76,78	81,89
FA	84,74	79,79	63,16	64,36	96,95	64,36	94,46	90,96	75,67	82,94	79,67	58,61
	82,32	81,68	63,16	64,36	97,27	64,36	93,38	94,21	75,67	83,44	82,11	59,39
	81,11	84,63	57,11	64,36	96,41	64,36	92,67	95,00	85,67	82,67	82,00	58,61
	80,11	82,89	57,37	64,36	96,68	64,36	93,88	94,67	85,54	81,78	83,06	58,61
	79,74	82,26	58,21	64,36	96,05	64,36	96,21	94,54	85,13	82,11	84,00	59,11
	77,89	81,00	58,11	64,36	96,73	64,36	96,50	94,33	86,46	81,56	83,28	58,33
	78,00	82,05	75,53	64,36	96,77	67,32	96,38	94,71	86,75	84,17	84,06	58,72
	78,68	82,74	75,16	64,36	96,45	67,18	96,92	94,29	87,13	83,72	83,72	58,44
	79,21	82,37	74,63	64,36	96,55	67,59	96,79	94,38	88,00	86,50	83,56	58,50
	79,58	81,79	71,42	64,36	96,09	67,27	96,75	93,50	89,92	86,72	83,78	59,00
	79,26	82,05	71,26	62,18	96,41	71,73	96,88	93,92	91,42	87,50	84,00	59,44
	78,63	81,95	70,79	62,41	96,32	86,45	95,71	93,88	92,12	87,17	83,78	60,06
	79,63	82,21	72,05	84,00	96,91	85,59	96,25	93,67	91,87	87,28	83,56	60,00
	78,68	82,58	83,26	84,00	96,41	85,41	96,17	94,04	92,67	87,06	83,00	59,89
	79,05	81,79	84,47	84,32	96,95	83,41	96,29	94,08	92,92	86,56	82,72	59,89
	79,05	82,11	84,32	84,00	96,00	81,82	96,12	93,75	93,92	86,72	82,94	58,06
	78,32	80,68	85,37	83,32	96,27	87,14	96,00	93,50	93,46	86,39	83,11	58,94
	77,63	81,42	84,42	88,59	96,41	84,50	95,42	93,25	94,54	86,06	82,83	82,44
	78,26	81,11	84,95	88,59	96,73	87,82	96,08	93,17	94,13	86,22	83,44	82,44
	78,53	82,11	85,00	94,41	96,18	92,32	95,29	93,58	93,37	86,33	83,44	82,67

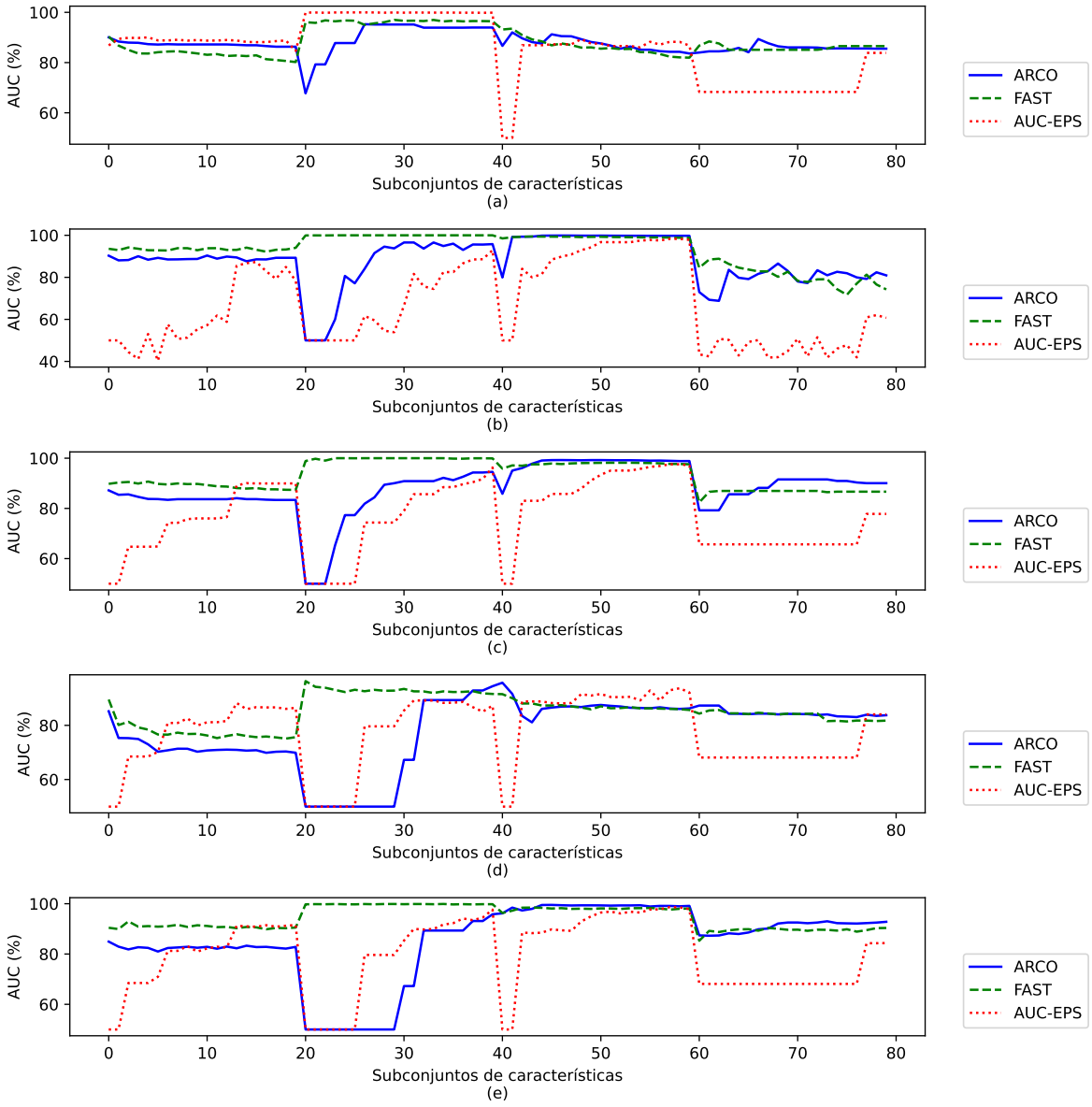

Figura A.1: Resultado da abordagem de *bootstrap* (AUC) utilizando os 20 pontos dos quatro conjuntos de dados do experimento. Cada gráfico corresponde, em ordem, aos seguintes classificadores: Naive Bayes (a), SVM (b), KNN (c), Árvore de Decisão (d) e Florestas Aleatórias (e).

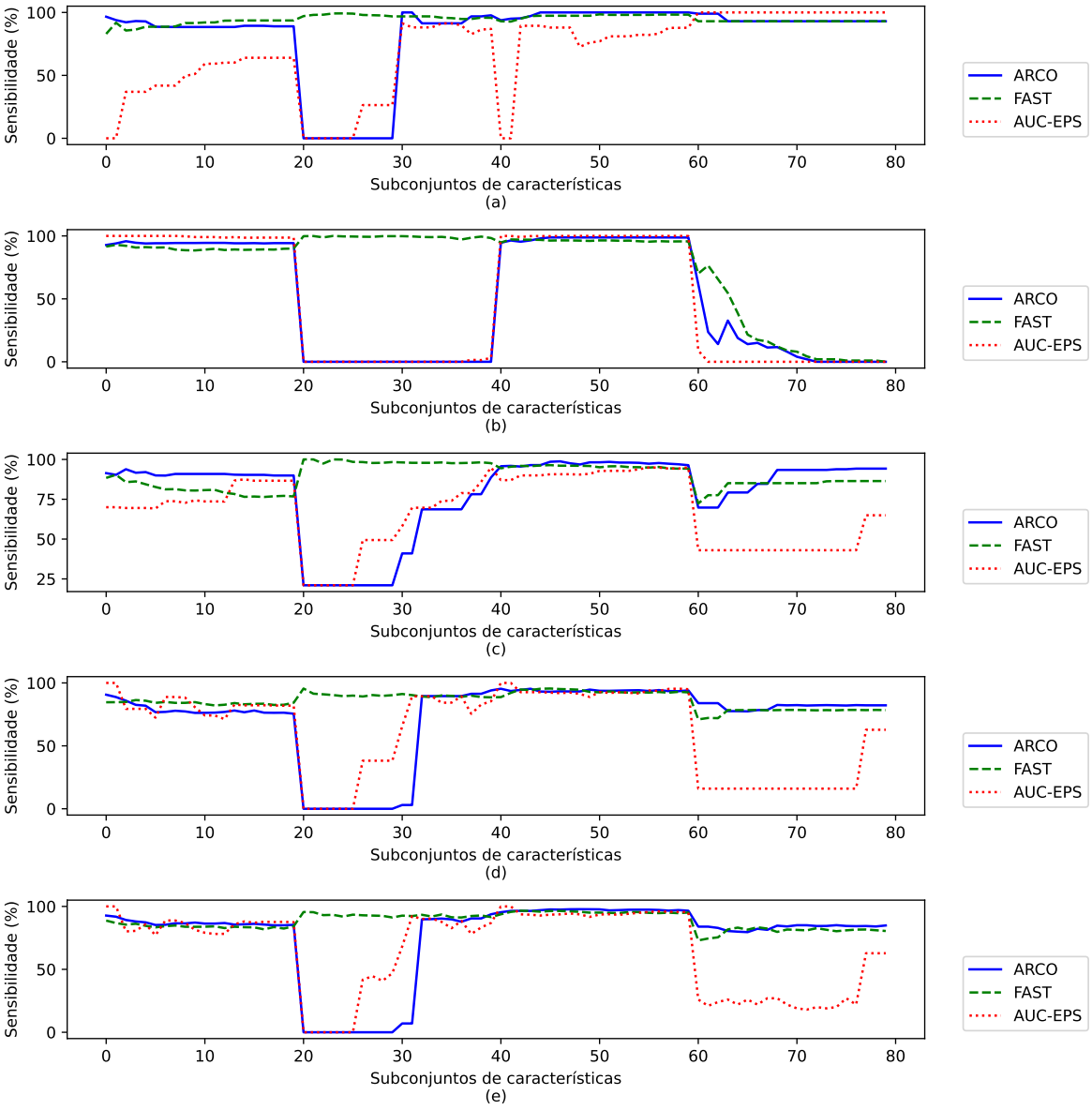

Figura A.2: Resultado da abordagem de *bootstrap* (sensibilidade) utilizando os 20 pontos dos quatro conjuntos de dados do experimento. Cada gráfico corresponde, em ordem, aos seguintes classificadores: Naive Bayes (a), SVM (b), KNN (c), Árvore de Decisão (d) e Florestas Aleatórias (e).

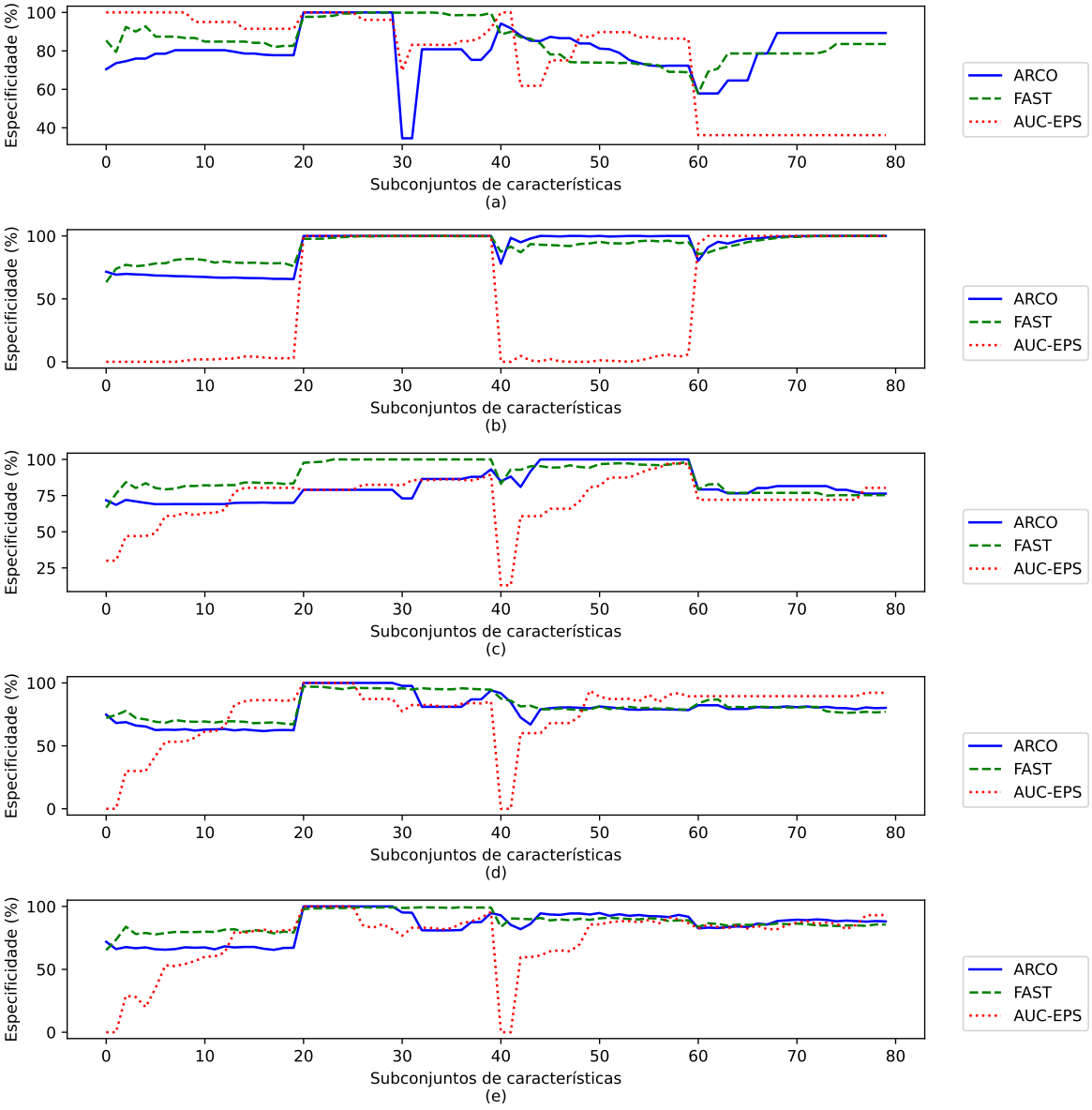

Figura A.3: Resultado da abordagem de *bootstrap* (especificidade) utilizando os 20 pontos dos quatro conjuntos de dados do experimento. Cada gráfico corresponde, em ordem, aos seguintes classificadores: Naive Bayes (a), SVM (b), KNN (c), Árvore de Decisão (d) e Florestas Aleatórias (e).

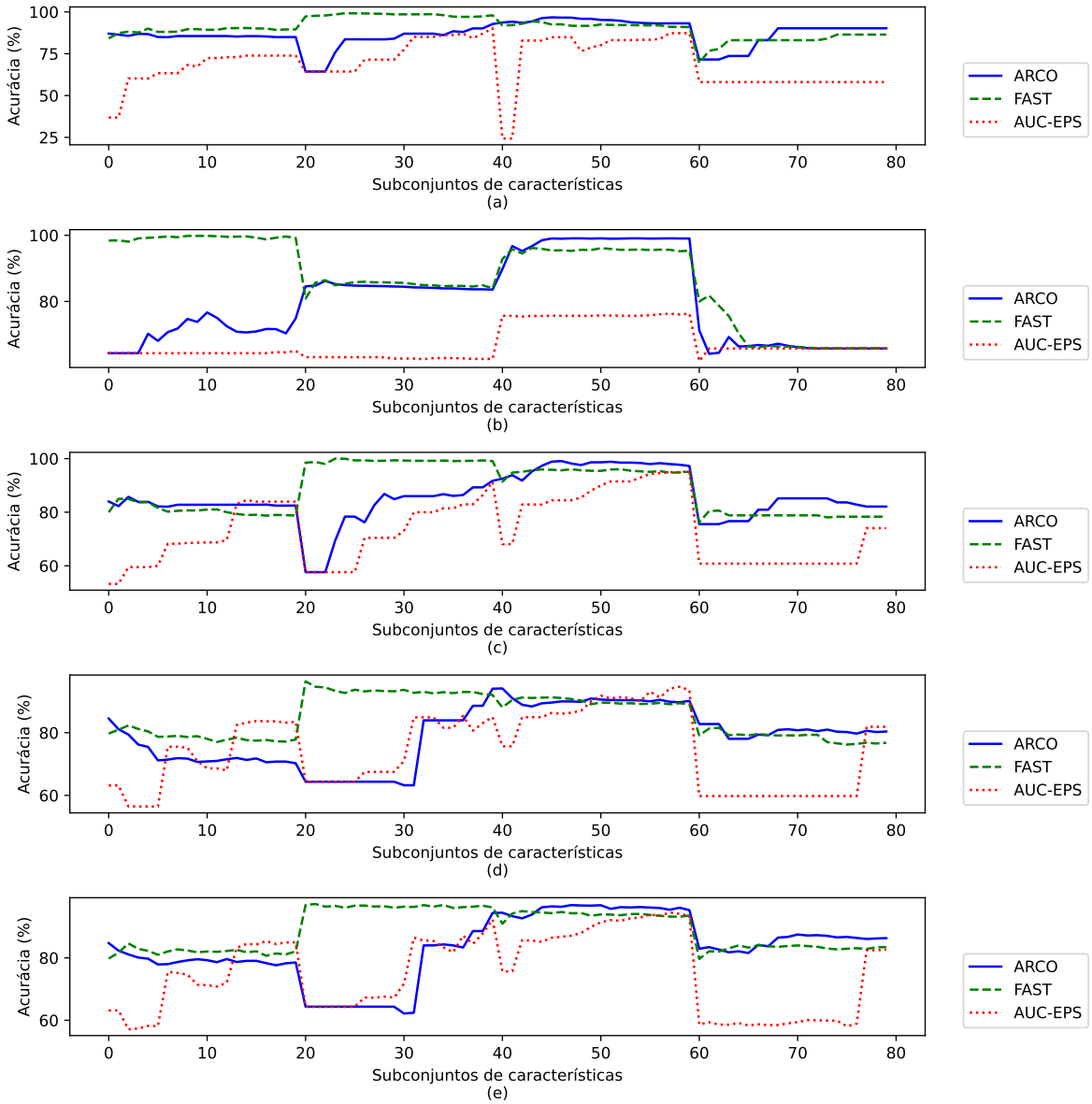

Figura A.4: Resultado da abordagem de *bootstrap* (acurácia) utilizando os 20 pontos dos quatro conjuntos de dados do experimento. Cada gráfico corresponde, em ordem, aos seguintes classificadores: Naive Bayes (a), SVM (b), KNN (c), Árvore de Decisão (d) e Florestas Aleatórias (e).