UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de São José do Rio Preto

INSTITUTO DE BIOCIÊNCIAS, LETRAS E CIÊNCIAS EXATAS
IBILCE / UNESP - CÂMPUS DE SÃO JOSÉ DO RIO PRETO

GUSTAVO HENRIQUE DE OLIVEIRA VILLAR

Redução da Dimensionalidade em Dados da Saúde por meio de combinação de algoritmos de Seleção de Atributos

São José do Rio Preto
2021

GUSTAVO HENRIQUE DE OLIVEIRA VILLAR

Redução da Dimensionalidade em Dados da Saúde por meio de combinação de algoritmos de Seleção de Atributos

Trabalho de Conclusão de Curso apresentado ao Departamento de Ciências de Computação e Estatística do Instituto de Biociências Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Carlos Roberto Valêncio

São José do Rio Preto
2021

V719r

Villar, Gustavo Henrique de Oliveira
Redução da Dimensionalidade em Dados da Saúde por meio de combinação de algoritmos de Seleção de Atributos / Gustavo Henrique de Oliveira Villar. -- São José do Rio Preto, 2021
61 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (Unesp), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto
Orientador: Carlos Roberto Valêncio

1. Data Mining. 2. Machine Learning. 3. Seleção de atributos. 4. Big Data. 5. Banco de Dados. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

GUSTAVO HENRIQUE DE OLIVEIRA VILLAR

Redução da Dimensionalidade em Dados da Saúde por meio de combinação de algoritmos de Seleção de Atributos

Trabalho de Conclusão de Curso apresentado ao Departamento de Ciências de Computação e Estatística do Instituto de Biociências Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Banca examinadora:

Prof. Dr. Carlos Roberto Valêncio
UNESP – São José do Rio Preto
Orientador

Prof. Dr. Adriano Mauro Cansian
UNESP – São José do Rio Preto

Prof. Dr. Valeriano Antunes de Oliveira
UNESP – São José do Rio Preto

São José do Rio Preto
2021

Agradecimentos

Inicialmente agradeço aos meus pais, Fátima e Wellington, por tudo que fizeram por mim, por todo o carinho, apoio, paciência, incentivo e por terem me proporcionado a oportunidade de me dedicar aos estudos desde sempre.

Agradeço ao Grupo de Banco de Dados (GBD) com todos seus membros e ao Prof. Dr. Carlos Roberto Valêncio por ter aceitado ser meu orientador e tornar a realização desse trabalho possível.

Agradeço a todos os meus amigos de São José do Rio Preto, especialmente ao David por sempre ter me ajudado quando possível, ao Douglas Canevarollo por ter me ensinado que muitas vezes o melhor caminho não é aquele que nós queremos, ao Douglas Brandão por me mostrar que não existe uma alternativa melhor ao esforço, ao Igor por me mostrar que muitas vezes o melhor que você pode fazer é ouvir o outro lado, ao Rafael por todas as conversas, conselhos e momentos descontraídos, ao Jardel que mesmo sendo uma amizade recente se provou insubstituível, ao Heinrich por estar presente em diversas madrugadas em claro depois de decisões aleatórias e questionáveis, ao Douglas Honda por sempre estar disposto a me ouvir e oferecer um ombro amigo, ao Pedro Afonso por me entender e me aturar nos mais diversos momentos, ao Gava por ter sido minha grande dupla durante todos os diversos desafios da graduação e da vida e, finalmente, à Vitoria por ter feito eu me lembrar que os sonhos das pessoas não tem fim.

Aos meus amigos de Jales agradeço ao Pedro Henrique, Gabriel Lopes, Felipe Navarro e Luiz por todo o apoio e carinho nos mais diversos momentos. Agradeço, também, ao Renan e ao meu primo Otávio, que mesmo com à distância sempre estiveram lá para me ajudar e motivar.

Não sei o que nos aguarda no futuro, mas foi um prazer estar com todos vocês.

"Eu não sei usar espadas, não sei navegar, também não sei cozinhar e nem mentir, o que eu sei, é que dependo dos meus amigos se quiser continuar vivendo!"

- Monkey D. Luffy (One Piece)

Resumo

Os avanços tecnológicos ocorridos nos últimos anos criaram ferramentas que tornaram possível um grande aumento na geração de dados em diferentes áreas da sociedade, entre elas, a área da saúde. Essa abundância de dados possui um grande potencial de conhecimento a ser extraído, porém, seu grande volume inviabiliza a exploração manual de toda essa capacidade. Neste contexto, é necessário recorrer à aplicação de técnicas automatizadas e bem definidas para extração do conhecimento. Uma dessas técnicas é o processo de *Data Mining*, que pode ser aplicado como uma das etapas da extração de conhecimento para o auxílio do diagnóstico preditivo de doenças a partir da classificação de elementos de um conjunto de dados, metodologia que se mostra promissora na busca em melhorar a qualidade de vida dos pacientes a partir de diagnósticos mais precisos e rápidos em comparação com aqueles sem assistência computacional. Contudo, as características de elevado volume e alta dimensionalidade desses dados geram uma dificuldade a mais em seu processo de exploração pois tornam os métodos mais custosos e menos eficientes, tornando-se assim necessário a aplicação de técnicas de seleção de atributos, que diminuem a quantidade de atributos presentes e permitem uma melhor predição e entendimento da influência de um determinado atributo sobre o resultado final. Sendo assim, este trabalho apresenta uma análise deste cenário a partir da aplicação de diversas técnicas de seleção de características associadas à mais de um algoritmo de classificação com múltiplos conjuntos de dados da área da saúde, sendo eles repositórios que abordam câncer de mama e doenças cardíacas. Os resultados mostram que técnicas de seleção de atributos podem aumentar significativamente a acurácia da classificação desse tipo de informação sem total ou nenhum comprometimento do tempo de execução, havendo casos em que até mesmo diminui-se o custo envolvido, levando a predições mais rápidas e eficazes, o que muitas vezes é crucial dentro do escopo médico.

Palavras-chave: Mineração de Dados (Computação), Extração de Conhecimento, Área da Saúde, Predição de dados, *Big Data*, Seleção de Atributos, Algoritmos de Classificação, KNN, J48

Abstract

Technological advances in the last years allowed the creation of tools capable of collecting a significant amount of data in the different sectors of society, among those sectors is the health industry. This abundance of data has a vast potential of generating knowledge after it has been processed; however, the volume of this data makes it unfeasible to manually explore all this capacity. In this scenario it is necessary the application of well-defined automation techniques that are able to extract knowledge from the data. One of these techniques is the process of Data Mining as a way to predict diagnosis using element classification on a dataset, methodology that has been reaching uplifting results given that it makes for a more accurate and faster diagnostic process when compared to those that do not rely on computer assisted decision making. Nevertheless, the characteristics of health data, such as, elevated volume and high dimensionality create challenges on the process of knowledge discovery making it less efficient and with a higher computational cost, as a way to solve this problem it is possible to apply feature selection algorithms that reduce the number of present attributes and allow us to have a better understanding of how much a single attribute can affect the final result. Therefore, the objective of this work is to create a scientific contribution based on the application of several feature selection methods associated to a couple of classification algorithms on multiple datasets that explore both breast cancer and cardiac diseases. The results show that feature selection techniques can significantly increase the accuracy of the classification of health data with very little or no losses on execution time, given that in most cases it's actually doable do reduce the execution time, leading to faster and more accurate predictions a very important aspect when taking the medical field in consideration.

Keywords: Data Mining (Computing), Knowledge Discovery, Health Data, Data Prediction, Big Data, Feature Selection, Classification Algorithms, KNN, J48

Lista de ilustrações

Figura 1 – Os cinco V's de Big Data....17
Figura 2 - Etapas do processo de KDD....19
Figura 3 – Exemplo de funcionamento do algoritmo KNN....22
Figura 4 – Exemplo de árvore de decisão....24
Figura 5 – Exemplo do funcionamento de um wrapper method....26
Figura 6 – Processos do Presente Trabalho....30
Figura 7 – Gráfico de Barras de Quantidade de Classes....34
Figura 8 – Gráfico de Violino 1....35
Figura 9 – Gráfico de Violino 2....36
Figura 10 – Gráfico de Violino 3....37
Figura 11 – Gráfico de Espalhamento 1....38
Figura 12 – Gráfico de Espalhamento 2....39
Figura 13 – Gráfico de Espalhamento 3....40
Figura 14 – Gráfico Conjunto 1....41
Figura 15 – Gráfico de Calor 1....42
Figura 16 – Gráfico Barras 2....43
Figura 17– Gráfico Violino....43
Figura 18– Gráfico Scatter 2....44
Figura 19 – Gráfico de Calor 2....45
Figura 20 – Formula do Cálculo de Acurácia....49
Figura 21 – Comparação da Acurácia....50
Figura 22 – Comparação de Tempo....50
Figura 23 – Comparação de Acurácia....52
Figura 24 – Comparação de Tempo....53

Lista de tabelas

Tabela 1 – Exemplo de dados não limpos 19
Tabela 2 – Exemplo de dados limpos 20
Tabela 3 – Comparativo entre os trabalhos correlatos 27
Tabela 4 – Estrutura de uma matriz de confusão 44
Tabela 5 – Valores de acurácia obtidos em cada caso (em porcentagem) 45
Tabela 6 – Tempo de execução de cada caso (em segundos) 46
Tabela 7 – Valores de acurácia obtido em cada caso (em porcentagem) 47
Tabela 8 – Tempo de execução de cada caso (em segundos) 47

Lista de abreviaturas e siglas

IoT– *Internet of Things*

KDD – *Knowledge Discovery in Databases*

VPN – *Virtual Private Network*

DM – *Data Mining*

KNN – *K Nearest Neighbor*

HD – *Heart Disease*

BC – *Breast Cancer*

Sumário

1 Introdução 12

1.1 Motivação e escopo 12

1.2 Objetivo 14

1.3 Metodologia 14

1.4 Organização da Monografia 15

2 Revisão bibliográfica 16

2.1 Aumento da disponibilidade de dados 16

2.1.1 Dados Médicos 17

2.2 Extração de conhecimento 18

2.2.1 *Data Clean* 19

2.2.2 *Data Mining* 20

2.3 Algoritmos de Classificação 21

2.3.1 *K Nearest Neighbor* (KNN) 21

2.3.2 Algoritmo J48 23

2.4 Feature Selection 24

2.4.1 Seleção de Atributos por Filtro 25

2.4.2 Seleção de Atributos por embrulho*(wrapper)* 26

2.5 Trabalhos correlatos e estado da arte 27

2.6 Considerações finais 29

3 Desenvolvimento 30

3.1 Seleção dos Conjuntos de Dados 30

3.1.1 *Heart Disease Data set* 30

3.1.2 *Breast Cancer Wisconsin (Diagnostic) Data Set* 31

3.2 Limpeza dos Dados 32

3.3 Visualização e Interpretação dos Dados 33

3.3.1 Análise do Primeiro Conjunto de Dados 33

3.3.2 Análise do Segundo Conjunto de Dados 42

3.4 Seleção de Atributos e Classificação 45

3.4.1 Algoritmos de Seleção por Filtro 46

3.4.2 Algoritmos de Seleção por Embrulho 46

3.5 Considerações Finais 47

4 Avaliação Experimental 48

4.1 Estratégia de Testes 48

4.2 Resultados e Discussões 49
4.2.1 *Breast Cancer Wisconsin (Diagnostic)* 49
4.2.2 *Heart Disease Dataset* 51
4.3 Considerações finais 53
5 Conclusão 55
5.1 Contribuições científicas 55
5.2 Trabalhos Futuros 56
Referências 58

1 Introdução

Com a evolução da tecnologia das últimas décadas foram criadas diversas ferramentas que permitiram a criação, manipulação, controle e armazenamento de dados em diversas áreas de conhecimento da humanidade e o setor da saúde não está isento desta tendência. Segundo Asri et al. (2016) celulares, pacientes, pesquisadores de hospitais e até mesmo o aumento do uso de *Internet of Things* (IoT) estão gerando uma quantidade de dados superior à capacidade atual do mercado de gerar valor a partir dela.

Era previsto que até o ano de 2020 existiriam mais de 12 Zettabytes (ZBs) de dados da área da saúde gerados a partir de diversas fontes como Prontuários Eletrônicos de Pacientes, dispositivos móveis e sensores específicos. Também era estimado que cerca de 50% dos hospitais iriam integrar soluções de *Big Data* em sua metodologia operacional (ASRI et al., 2016). Segundo Kaur et al. (2018) tais dados podem ser utilizados para auxiliar profissionais da saúde em questões, como: cuidados com paciente, formas de tratamento, uso de recursos dos hospitais e até mesmo na predição de doenças em larga escala, como em casos de epidemias e pandemias (MORSE et al., 2012), (WANG et al., 2020) diagnóstico de pacientes individuais (LE et al., 2018) e até mesmo casos de óbitos como mostrado por Santos et al. (2019). Essas abordagens fazem uso de conceitos computacionais cada vez mais difundidos (SENGUPTA et al., 2019) como algoritmos de extração de características, algoritmos de classificação ou regressão (RESEMEIRO; BOLON-CANEDO, 2019) e estratégias de mineração de dados (ISHAQ et al., 2021).

Entretanto, a utilização desses métodos não está livre de empecilhos, em especial a abordagem baseada em classificação, uma vez que a informação fornecida pode incluir redundâncias ou sintomas não relacionados, gerando predições equivocadas e um desperdício de custo computacional, colocando em risco a saúde do paciente e podendo causar um mal aproveitamento dos recursos da instituição responsável (LE et al., 2018).

1.1 Motivação e escopo

A grande quantidade de dados referentes à área da saúde, assim como sua aplicação, está sendo artefato de estudo de diversos trabalhos científicos que buscam conseguir extrair informações úteis a partir de processos de *Data Mining* (DM), uma vez que é possível se obter regras e associações subentendidas por meio de algoritmos de predição que não seriam encontrados utilizando-se somente a capacidade humana (FU et al., 1997). Porém, esses dados

costumam possuir alta dimensionalidade, isto é, grande número de atributos, prejudicando assim o desempenho de muitos algoritmos de classificação pois muitos desses atributos são redundantes, ruídos, ou não agregam valor à predição a ser feita prejudicando sua acurácia e tempo de execução (CIA et al., 2018).

Uma forma de lidar com essa particularidade é a aplicação de algoritmos de seleção de características (do inglês, *feature selection)* (MIAO et al., 2016), que consiste na redução da quantidade de atributos a partir da seleção de um subconjunto relevante para o processo de classificação. Como mostrado por Le et al. (2018), a redução do número de atributos resulta em um processo de classificação com melhor acurácia e com o processamento de uma quantidade menor de dados.

Segundo Remeseiro e Bolon-Canedo (2019) outra vantagem da redução de dimensionalidade é também auxiliar no entendimento da causa da doença, uma vez que menos atributos significam um maior poder de distinção. Este aumento de transparência, facilidade de interpretação e leitura de resultados são características fundamentais para sistemas de predição aplicados em áreas sensíveis, como é o caso do setor médico.

Na literatura também é possível encontrar trabalhos cujo foco não é o algoritmo de seleção de características, mas sim o método de classificação. Entre os métodos mais comuns pode-se citar *BayesNetwork*, J48, *Random Trees* (PARIMALA; PORKODI, 2018) e *K-Nearest Neighbor* (KNN) (WENCHAO; YILIN, 2021). Existem, também, obras que propõe melhorias para suas contrapartes clássicas, como o proposto por Wenchao e Yilin (2021) que busca melhorar o desempenho do método KNN para conjuntos de dados volumosos e com alta dimensionalidade. O algoritmo J48 também se mostra presente em estudos preditivos para casos de diabetes (KAUR; CHHABRA, 2014) e câncer de mama (ORTEGA et al., 2020), porém, seus resultados são superados por outros métodos citados dentro do cenário médico, como mostrado por Parimala e Porkodi (2018).

Sendo assim, é possível afirmar que melhorias em técnicas de predição podem auxiliar a automatizar tarefas de diagnóstico enquanto os profissionais da área ficam livres para funções de maior complexidade, ou que exigem maior interação humana (NGIAM; KHOR, 2019). Além disso, para dados médicos, a necessidade de garantir uma alta acurácia e velocidade são essenciais para que um diagnóstico possa ser feito com tempo hábil para tomada de decisão (NG et al., 2016).

1.2 Objetivo

Tendo em vista a crescente importância da extração de conhecimento útil a partir de dados armazenados utilizando os processos de *Knowledge Discovery in Databases* (KDD) (FAYYAD et al., 1996), este trabalho visa a criação de uma ferramenta que possa amparar as diversas decisões diagnósticas tomadas por profissionais deste domínio.

Levando em consideração a potencialização de algoritmos de classificação como forma de predição de informações convenientes à saúde das pessoas o presente trabalho busca realizar a associação de diferentes métodos de seleção de atributos com algoritmos de classificação baseados em instância e árvores de decisão em conjuntos de dados com características distintas. Deste modo, o ambiente proposto visa oferecer a possibilidade de análise das combinações entre os algoritmos disponíveis a fim de recomendar melhor associação e resultados.

1.3 Metodologia

Para embasar o trabalho, foi realizado um levantamento bibliográfico a respeito dos temas pertinentes e conceitos envolvidos como: *Big Data,* KDD, *Data Mining*, *Data Clean,* algoritmos de classificação, *feature selection* e seus métodos. Também foi necessário realizar pesquisas sobre os diversos conjuntos de dados disponíveis. As fontes para obtenção de tais informações foram os diversos repositórios de divulgação científica disponibilizados pela UNESP através de sua VPN.

Posteriormente, foram definidos os conjuntos de dados (do inglês, *dataset*) viáveis da área da saúde com ênfase em classificação de sintomas para diagnóstico, sendo eles *Breast Cancer Wisconsin (Diagnostic) Data Set* e *Heart Disease Data Set*, ambos disponíveis no repositório online da Universidade da Califórnia Irvine (UCI) (WOLBERG et al., 1995), (JANOSI et al., 1988). Caso necessário, serão realizados processos de limpeza de dados.

Em seguida, serão executados os algoritmos de classificação KNN e J48 para a obtenção de resultados iniciais que serão utilizados como base de *benchmark*. Baseando-se na literatura serão definidos algoritmos de *feature selection* que, aplicados nos conjuntos de dados previamente ao processo de classificação, poderão resultar em uma predição de maior qualidade quando comparados com os resultados obtidos sem a aplicação de *feature selection*.

Sendo assim, a metodologia do desenvolvimento deste trabalho é baseada nas etapas de: levantamento bibliográfico, aplicação dos algoritmos de *feature selection*, aplicação dos métodos de classificação e a realização dos testes necessários.

1.4 Organização da Monografia

Este capítulo possui um teor introdutório no qual foram apresentadas algumas dificuldades na predição de dados na área da saúde além de elaborar a motivação, a justificativa e o objetivo do trabalho.

Os próximos capítulos estão organizados da seguinte forma:

a) Capítulo 2 - Revisão bibliográfica: Conceitos importantes sobre predição de dados, algoritmos de classificação, algoritmos de *feature selection*, trabalhos correlatos e estado da arte;
b) Capítulo 3 – Metodologia e Desenvolvimento do Trabalho: Apresenta de forma completa a metodologia que foi utilizada durante a formulação do trabalho;
c) Capítulo 4 – Avaliação Experimental: exibição e discussão dos experimentos realizados e resultados óbitos.
d) Referências – Lista com todas as publicações cientificas que foram utilizadas para embasar a confecção deste trabalho.

2 Revisão bibliográfica

Neste capítulo são apresentados os conceitos fundamentais pertinentes ao entendimento deste trabalho. Inicialmente, será elaborado sobre o estado atual da geração, acesso e aproveitamento de dados médicos no processo de KDD. Em seguida, serão caracterizados os conceitos de *Data Mining* e as vantagens de sua aplicação em conjuntos de dados. Além disso, será detalhado o funcionamento de algoritmos de classificação e métodos de seleção de características. Finalmente, será realizada uma comparação com os trabalhos correlatos e uma elucidação do estado da arte.

2.1 Aumento da disponibilidade de dados

O avanço da tecnologia serviu como um catalisador para a criação de ferramentas como sensores, celulares, vestimentas tecnológicas (*wearables)* e objetos inteligentes conectados à internet capazes de coletar dados em tempo integral. Este aumento na quantidade de dados gerou a necessidade de melhores processos de administração, controle e processamento de dados para que se possa retirar conhecimento útil e aplicado ao cotidiano das pessoas (ASRI et al., 2016).

Neste cenário, foi cunhado o termo *Big Data* para descrever as características desta quantidade crescente de dados (ANREU-PEREZ et al., 2015), (MURDOCK; DETSKY, 2013). Os principais aspectos desse fenômeno podem ser resumidos em cinco V's conforme ilustrado na Figura 1 e listados abaixo (ASRI et al., 2016) (DEMCHENKO et al., 2014).

- Volume: se refere a grande quantidade de dados;
- Variedade: se refere a como os dados podem ser criados: estruturados, não estruturados, semiestruturados;
- Velocidade: se refere a taxa com que os dados são criados;
- Veracidade: se refere ao quão acurados e corretos são os dados a serem trabalhados;
- Valor: se refere aos benefícios que podem ser obtidos a partir da análise desses dados.

Trabalhar com *Big Data* vem se tornando uma constante em todas áreas do conhecimento a ponto de nos últimos anos a quantidade de publicações referentes a este tópico dobrarem, incluindo dados da área da saúde (ANREU-PEREZ et al., 2015).

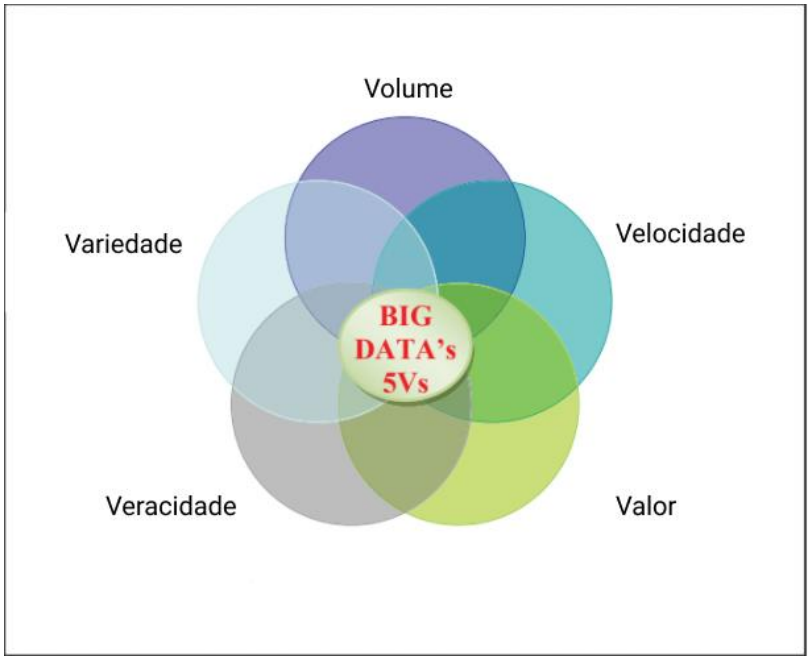
2.1.1 Dados Médicos

Dada a definição apresentada na seção anterior é possível concluir que os dados médicos fazem parte do conceito de *Big Data* uma vez que os dados de saúde utilizados por sistemas estão na ordem de *terabytes* e continuam crescendo à uma taxa de 2.4 *exabytes* ao ano (Volume); a maior parte dos dados são oriundos de prontuários eletrônicos, anotações médicas mas também existem fontes mais modernas e estruturadas como dados clínicos e resultados de procedimentos médicos (Variedade); dados são gerados tão rapidamente que não é possível garantir que são armazenados e atualizados na mesma frequência com que são criados (Velocidade); as diferentes formas de se obter dados podem gerar valores inconstantes (Veracidade) e não acurados e existe grande valor na retirada de informação desses dados (Valor) (ASRI et al., 2016).

Outra característica comum em dados médicos é a alta dimensionalidade. Devido à forma como são gerados é comum que esses dados apresentem muitos atributos sendo que nem todos possuem grande impacto na informação final que pode ser retirada desses dados, dificultando, assim, a extração de conhecimento útil e tornando os processos aplicados mais custosos (FAN; LI, 2006) (FAYYAD et al., 1996).

Sendo assim, é possível inferir que a análise de dados médicos permite uma série de vantagens ao serem explorados, como aprimorar o atendimento ao paciente, melhorar o desempenho de exames e da qualidade de serviço prestada pelos profissionais da área, gerir e planejar o uso de recursos das instituições de saúde (NEW et al., 2018) e a melhoria do potencial diagnóstico por meio da extração de conhecimento e aplicação de técnicas de predição (LE et al., 2018) (NGIAM; KHOR, 2019).

Figura 1 – Os cinco V's de *Big Data*



Fonte: Adaptado de (ASRI et al., 2016).

2.2 Extração de conhecimento

Conforme mencionado na seção anterior, o crescimento da quantidade de dados gerados é superior à capacidade atual de analisar e gerar conhecimento útil a partir dos mesmos. Para mitigar essa discrepância entre dados gerados e capacidade analítica é necessário aplicar técnicas computacionais bem definidas de extração de conhecimento em bases de dados (KDD) (FAYYAD et al., 1996).

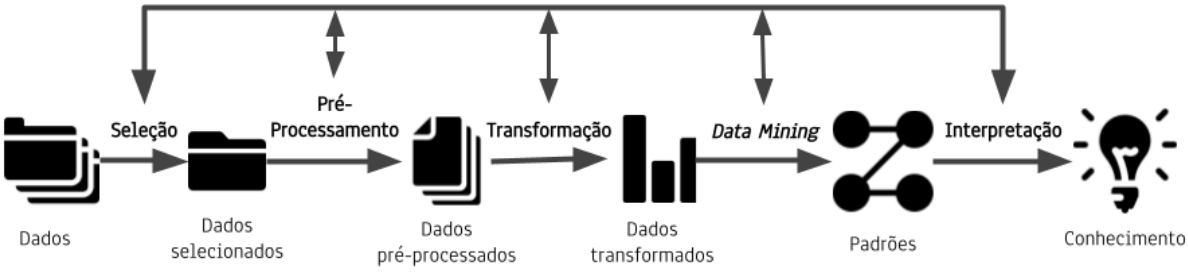
Segundo Fayyad et al. (1996) a definição de KDD pode ser dada por "O processo não trivial de identificar padrões válidos, novos, potencialmente úteis e, por fim, entendíveis nos dados" [1](FAYYAD et al., 1996, tradução nossa). Os dados precisam ser válidos pois é necessário possuir algum grau de certeza sobre as informações que podem estar contidas ali, os padrões precisam ser novos ao usuário e úteis para uma tarefa ou atividade e entendíveis imediatamente ou após sofrer pós processamento. O processo de KDD é ilustrado de forma simplificada na Figura 2 uma vez que os processos não ocorrem de maneira sequencial e podem retornar a uma etapa prévia caso seja necesário (FAYYAD et al., 1996) (BRACHMAN; ANAND, 1994). As etapas são:

- **Seleção**: consiste na escolha de um conjunto de dados ou subconjunto de atributos relevantes no qual ocorrerá o processo de descoberta.
- **Pré-processamento:** inclui processos de *Data Clean* como remoção de ruído e dados discrepantes e qual será a metodologia abordada para lidar com dados faltantes e as consequências dessa perda de informação.
- **Transformação:** é a etapa na qual os dados são transformados para encontrar as características mais relevantes ao conhecimento que se deseja obter, normalmente por meio da redução da dimensionalidade utilizando algoritmos de transformação, como é o caso dos métodos de seleção de características.
- **Data Mining (DM):** é o estágio no qual ocorre o processamento dos dados já transformados, sendo que este processamento é realizado por algoritmos específicos com características distintas dependendo da natureza dos dados a serem processados. Ao final dessa etapa são obtidos os padrões relevantes para a interpretação.
- **Interpretação:** inclui os processos de interpretar e usar conscientemente as informações obtidas, normalmente, com o auxílio de técnicas de visualização e transformando esses resultados em termos entendíveis para os usuários. Finalmente,

[1] "The nontrivial process of identifying valid, novel, potentially usefull, and ultimately understandable patterns in data."

as conclusões obtidas são utilizadas na tomada de decisão ou simplesmente documentadas e reportadas.

Figura 2 – Etapas do processo de KDD



Fonte: Elaborado pelo autor.

2.2.1 *Data Clean*

Conforme mostrado na subseção anterior o processo de *Data Clean* é uma das etapas de destaque do KDD uma vez que melhora a qualidade dos dados a serem trabalhados por meio da detecção e remoção de erros ocorridos durante o processo de geração e armazenamento dos dados. Estes erros podem ser oriundos de operações singulares como: dados repetidos ou nulos, erros humanos e falta de padronização na inserção (conforme mostrado na Tabela 1), ou de dificuldades de integração a partir de múltiplas fontes, as quais, em cada instância, possuem os problemas citados (RAHM; DO, 2000).

Tabela 1 – Exemplo de dados não limpos

ID	Nome	Idade	Temperatura
1	arthur jorge nicolas ferrreira	58	36.33
2	Thiago I.Novaes	042	37,4
3	ELISA ALESSANDRA MOURA	01/03/1994	
4	Thiago I. Novaes		37.4
5	Fernanda Rebeca Assunção	0012	388

Fonte: Elaborado pelo autor.

O processo de limpeza é fundamental no auxílio à tomada de decisão pois dados inconstantes podem levar a conclusões que não são um reflexo da realidade, porém as medidas necessárias para realizar a limpeza de um conjunto de dados não são padronizadas e exigem uma avaliação caso a caso. A Tabela 2 mostra os mesmos dados presentes na Tabela 1 após passarem por um processo de limpeza (RAHM; DO, 2000).

Tabela 2 – Exemplo de dados limpos

ID	Nome	Idade	Temperatura
1	Arthur Jorge Nicolas Ferreira	58	36.8
2	Thiago Iago Novaes	42	37.4
3	Elisa Alessandra Moura	27	36.3
4	Fernanda Rebeca Assunção	12	38.8

Fonte: Elaborado pelo autor.

2.2.2 *Data Mining*

Como elaborado na seção 2.2, *Data Mining* é uma das etapas do processo de KDD, na qual se aplica algoritmos específicos com a finalidade de extrair padrões a partir dos dados pré-processados, uma vez que é necessário o conhecimento do contexto no qual esses dados estão inseridos para garantir um processo de mineração com resultados significativos com valor no mundo real (GOLDSCHMIDT et al., 2015).

Os algoritmos de mineração de dados têm a função de gerar conhecimento inferido a partir de uma composição de técnicas e princípios básicos: o modelo é composto por dois fatores relevantes, a função e a forma de representação do modelo; o critério de preferência e o algoritmo de busca. Não é comumente encontrado na literatura os princípios utilizados em um determinado algoritmo uma vez que eles costumam aparecer de forma mista (FAYYAD et al., 1996).

Segundo Fayyad (1996) as funções modelo podem ser categorizadas de acordo com suas características, sendo que cada uma delas possui suas próprias condições de uso e peculiaridades:

- **Classificação**: categoriza um item discreto em uma de duas categorias pré-existentes
- **Regressão:** categoriza um item contínuo em uma categoria pré-existente.
- **Clusterização:** categoriza um item em uma categoria de diversas classes que são definidas a partir de agrupamentos naturais baseados em métricas de similaridade.
- **Sumarização:** gera uma descrição de um subconjunto de dados.
- **Modelagem de dependência:** descreve dependências significativas entre duas variáveis e pode ser aplicado de forma estruturada e de forma quantitativa.
- **Análise de ligação:** determina relações entre dois campos de um conjunto de dados.

É importante ressaltar que cada tipo de técnica se encaixa melhor para um tipo específico de problema e que, na prática, uma grande parte dos problemas podem ser melhor resolvidos ao se gastar mais tempo em sua formulação do que na otimização de detalhes particulares de um algoritmo específico de mineração de dados.

Um algoritmo de *Data Mining* possui dois objetivos principais: realizar uma predição, uma descrição, ou ainda, uma combinação das duas primeiras possibilidades. Em casos majoritariamente preditivos o objetivo é conseguir a maior acurácia possível na classificação dos dados, enquanto em situações preponderantemente descritivas o maior ganho ocorre devido ao entendimento das relações subjacentes no conjunto de dados. Na prática, é comum que essas duas vertentes coexistam em algum grau (FAYYAD et al., 1996) (GOLDSCHMIDT et al., 2015).

Uma das dificuldades do processo de mineração é conseguir realizar a extração de padrões de forma correta em conjuntos de dados com muitos registros e alta dimensionalidade, pois é possível que o algoritmo encontre padrões que, apesar de existentes, não condizem com o conhecimento real que se pretende explorar. Para lidar com esse tipo de problema é comum a aplicação de algoritmos muito eficientes, paralelização de processos ou então a redução da dimensionalidade dos dados utilizando métodos de transformação como algoritmos de seleção de características (FAYYAD et al., 1996) (LE et al., 2018).

2.3 Algoritmos de Classificação

Os algoritmos de classificação passaram a ser uma importante técnica de mineração de dados com capacidade de aplicação em diversas áreas, cada uma com suas peculiaridades, pois ao possibilitar a classificação de forma correta de um item a partir de suas características (atributos) proeminentes é possível realizar a predição de resultados que, na área da saúde, pode ser utilizado como fator determinante para o diagnóstico correto e realizado de forma rápida(FAYYAD et al., 1996) (SARITAS; YASAR, 2019). Existem diversos algoritmos distintos baseados em múltiplas estratégias, o presente trabalho irá abordar os algoritmos de KNN, baseado em métodos de instâncias e o algoritmo J48, que consiste em uma implementação do algoritmo C4.5 (PARIMALA; PORKODI, 2018), baseado em indução de árvores de decisão (GOLDSCHMIDT et al., 2015).

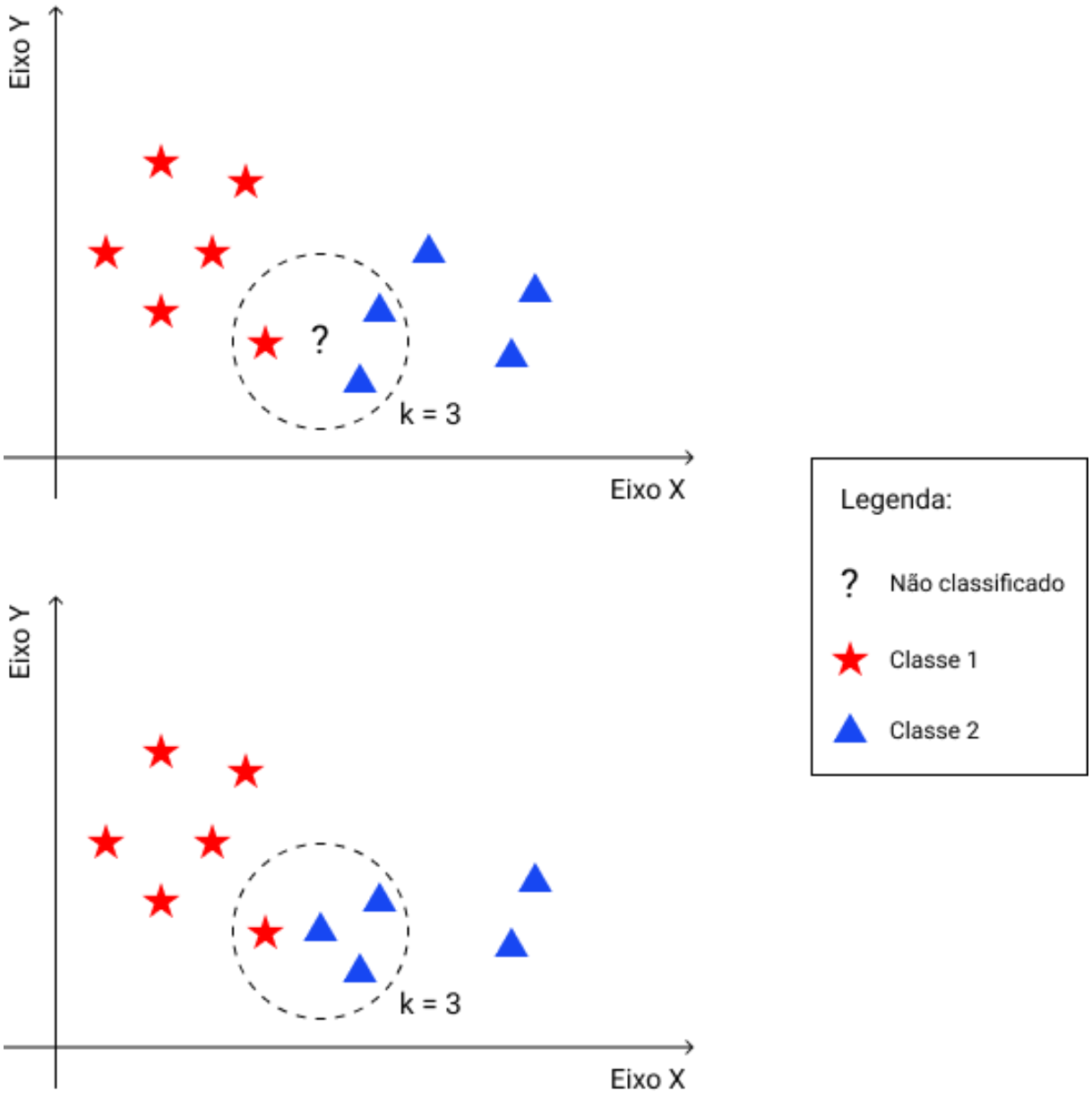
2.3.1 *K Nearest Neighbor* (KNN)

Conforme apresentado anteriormente, o algoritmo KNN é um método baseado em instâncias, ou seja, é um método que considera os registros já existentes no banco de dados para realizar o processamento do novo item. Sua implementação é relativamente simples e não exige um processo de treinamento prévio.

Para cada um dos itens a serem classificados o algoritmo realiza as mesmas etapas, que são: fazer o cálculo da distância do item a ser classificado para cada um dos outros registros já existentes e utilizados como referência utilizando uma métrica de distância; identificação dos k elementos com menor distância até o novo registro; contagem de qual das classes previamente

estabelecidas possuem mais elementos e, finalmente, a classificação do novo registro é feita e em casos supervisionados é realizado o teste de acerto para definir a acurácia do algoritmo, se o algoritmo não estiver sendo aplicado de maneira supervisionada sua execução é encerrada após a classificação. Na Figura 3 é apresentado um exemplo do funcionamento do algoritmo para o valor de k = 3 (GOLDSCHMIDT et al., 2015).

Figura 3 – Exemplo de funcionamento do algoritmo KNN



Fonte: Elaborado pelo autor.

2.3.2 Algoritmo J48

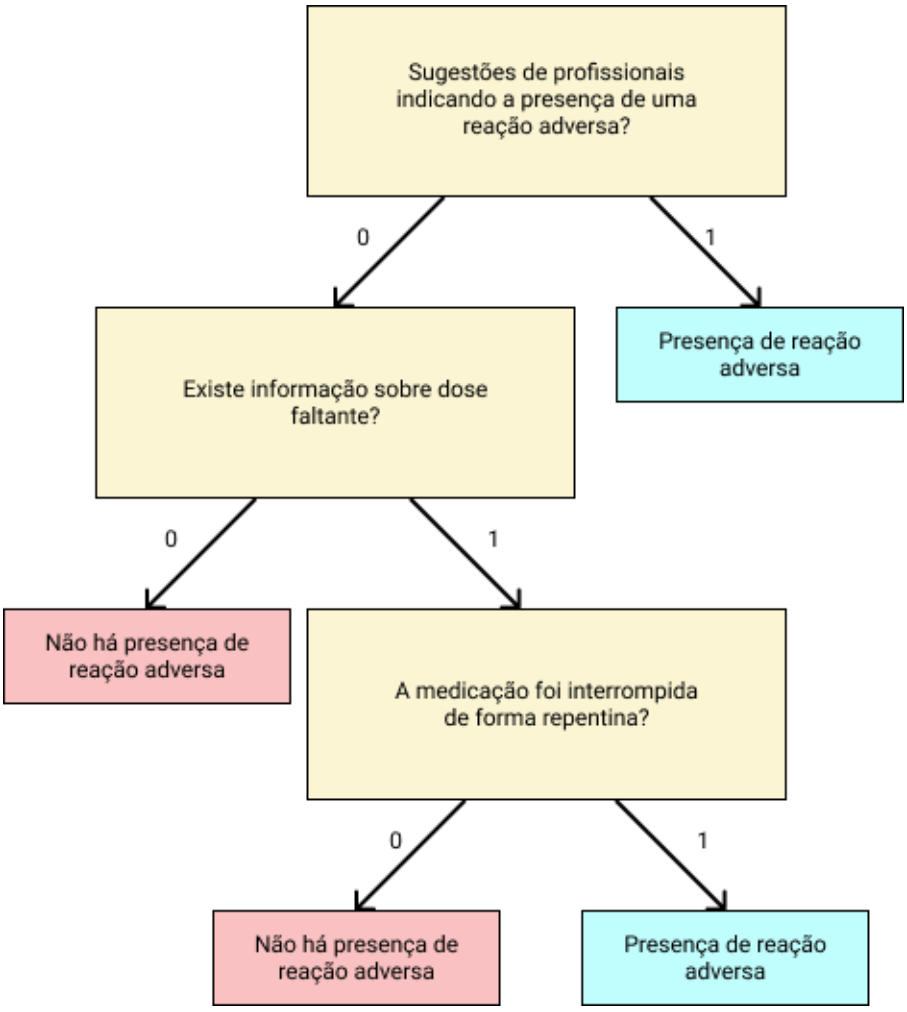
O algoritmo J48 é uma implementação em JAVA do algoritmo C4.5 que determina a classificação baseando-se na minimização da entropia de informação (GLASGOW; KABOLI, 2010) (LE et al., 2018). Esse algoritmo é baseado em métodos de indução de árvores de decisão, ou seja, utiliza-se os dados da base para realizar a tomada de decisão e a construção da árvore, resultando assim, em um processo de classificação (GOLDSCHMIDT et al., 2015).

Uma árvore de decisão é uma estrutura na qual cada nó representa uma decisão que foi tomada com a finalidade de chegar à uma classificação final. Durante os primeiros processos a árvore é composta somente por sua raiz e utiliza de todo o conjunto de dados para realizar a tomada de uma decisão, em sequência, os atributos do problema são divididos em dois ou mais conjuntos e assim sucessivamente até que o conjunto referente a cada nó seja majoritariamente de uma única classe. De forma geral, algoritmos que apresentam essa característica são divididos em duas fases (GOLDSCHMIDT et al., 2015):

- **Fase de construção**: ocorre a divisão do conjunto de dados em uma parte de treinamento e uma parte de testes sendo que o primeiro é separado em diversas subpartições a partir dos valores de cada um dos atributos presentes. Esse processo é executado de forma recursiva até que todos os nós sejam predominantemente de uma única classe.
- **Fase de simplificação**: consiste da fase de poda da árvore, na qual o atributo que foi mais relevante para a tomada de decisão é selecionado e, consequentemente, define em qual categoria o item analisado deve ser classificado.

A Figura 4 apresenta um exemplo de como seria uma possível árvore de decisão para a detecção de ocorrências de reações adversas à medicamentos (do inglês, *Adverse Drug Event*) com duas possibilidades de classificação (GLASGOW; KABOLI, 2010)

Figura 4 – Exemplo de árvore de decisão



Fonte: Adaptado de (GLASGOW; KABOLI, 2010).

2.4 Feature Selection

Cada método de *Data Mining* possui características próprias que geram a necessidade de diferentes formas de pré-processamento de acordo com sua função (GOLDSCHMIDT et al., 2015). No cenário de dados médicos uma das principais estratégias de transformação de dados é a aplicação de algoritmos de seleção de características por possibilitarem a redução da dimensionalidade do *dataset*, fato que pode contribuir para uma mineração mais acurada, menor geração de falsos padrõese menos tempo gasto em processamento que não irá retornar conhecimento útil. (MIAO et al., 2016) (RESEMEIRO; BOLON-CANEDO, 2019).

Algoritmos de *feature selection* são técnicas de pré-processamento que identificam quais são os atributos chave de um determinado problema e os utilizam para gerar um subconjunto de

dados com maior poder discriminativo, melhorando aspectos de legibilidade, facilidade de interpretação e transparência de características, fatores essenciais em um processo de extração de conhecimento de dados médicos (RESEMEIRO; BOLON-CANEDO, 2019).
Baseando-se nas estratégias de busca é possível dividir esses algoritmos em três métodos distintos (MIAO et al., 2016): métodos por filtro, métodos por embrulho e métodos embutidos (tradução nossa). Neste trabalho será dado ênfase aos dois primeiros.

- **Métodos por filtro (do inglês *Filter Methods)*:** são independentes do método de classificação pois focam nas características gerais dos dados, possuindo menor custo computacional e grande poder de generalização. Devido a essa característica, costumam apresentar uma estratégia de execução em duas etapas.
- **Métodos por embrulho (do inglês *Wrapper Methods):*** são dependentes do método de classificação, necessitando que o mesmo realize a avaliação do subconjunto de atributos candidatos. Essas operações os tornam mais custosos que os métodos de filtro, porém costumam obter melhores resultados em termos de acurácia.
- ***Embedded Methods (*Métodos embutidos) (tradução nossa)*:** consistem em métodos intermediários entre os citados previamente, uma vez que necessitam que o processo de seleção seja parte do treinamento do algoritmo de classificação, ocorrendo nessa etapa a busca pelo melhor subconjunto a ser selecionado.

2.4.1 Seleção de Atributos por Filtro

A seleção de atributos por meio de correlação consiste de uma forma de *feature selection* baseada em método de filtro uma vez que não depende do algoritmo de classificação que será utilizado. Os métodos dessa categoria apresentam a vantagem de serem dependentes de características gerais do conjunto de dados e, portanto, apresentam um custo computacional menor (YU; LIU, 2003). Segundo Hall (1999) a eliminação de dados correlatos, redundantes ou que se mostrem irrelevantes auxiliam na redução da complexidade dos algoritmos de classificação, além de permitir que uma quantidade menor de dados gere um modelo com uma acurácia maior.

Existem algoritmos de filtros como o caso do Relief (KIRA; RENDELL, 1992) que por meio da atribuição de pesos consegue eliminar as características mais irrelevantes, porém, esse tipo de algoritmo não realiza a diferenciação entre as características correlatas, gerando problemas ao treinar o modelo com diversas características redundantes. Algoritmos de busca em subgrupos conseguem solucionar o problema de eliminação de características correlatas, porém, a partir de dados experimentais é possível observar que possuem uma complexidade alta, sendo o número

de iterações uma exponencial quadrática em relação ao número de características inicialmente apresentadas pelo conjunto de dados (YU; LIU, 2003).

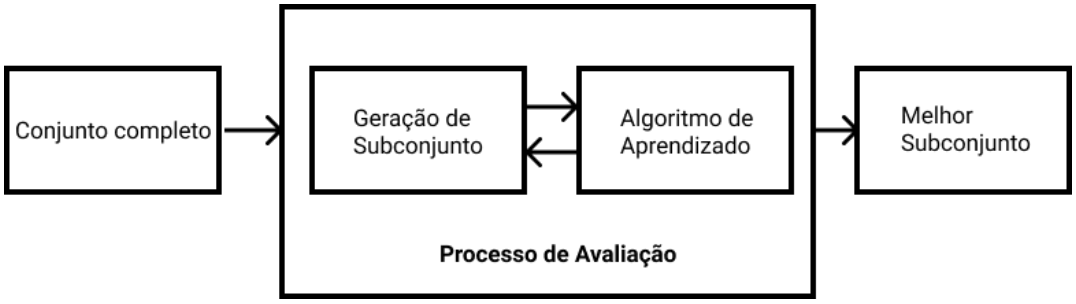
As técnicas apresentadas no parágrafo anterior são exemplos, respectivamente, de abordagens de algoritmos univariados e multivariados. Sendo que os pertencentes ao primeiro grupo realizam o ranqueamento de importância de atributos a partir da análise de um único atributo por vez, sem considerar o relacionamento deste com os demais. Já os algoritmos multivariados dependem de uma estratégia de busca para selecionar um subconjunto de características dentro do conjunto inicial, levando em consideração, o relacionamento entre eles (JOVIĆ et al, 2015).

2.4.2 Seleção de Atributos por embrulho *(wrapper)*

Conforme apresentado por El Aboudi e Benhlima (2016), métodos de seleção de características baseados em *wrapper* utilizam um algoritmo de aprendizado de máquina para avaliar a performance de uma característica candidata a fazer parte do novo subgrupo de atributos. Apesar de usualmente apresentarem bons resultados, apresentam um grande custo computacional que, muitas vezes, os torna proibitivos (MIAO et al., 2016). A estrutura geral de um algoritmo dessa categoria é apresentada na Figura 5.

Dentre os algoritmos dessa categoria pode-se citar o algoritmo de eliminação recursiva de atributos (do inglês, *recursive feature elimination)* (RFE), que apresenta grande eficácia na solução de problemas de *feature selection* por realizar a eliminação de atributos redundantes ou que não importantes a partir de chamadas recursivas (CHEN; JEONG, 2007). Tal algoritmo foi criado como uma tentativa de reduzir a dimensionalidade em dados sobre câncer, onde se trabalhava com uma quantidade baixa de dados, mas que possuíam muitos atributos, levando a problemas de *overfitting* nos modelos treinados.

Figura 5 – Exemplo do funcionamento de um *wrapper method*



Fonte: Adaptado de (EL ABOUDI; BENHLIMA, 2016).

2.5 Trabalhos correlatos e estado da arte

Na literatura é possível encontrar diversas publicações que abordam a crescente quantidade de dados na área da saúde e as possibilidades que foram abertas pelo avanço do poder computacional, principalmente no aspecto de *Data Mining*, KDD e abordagens preditivas com a finalidade de aumentar o apoio a profissionais do setor na tomada de decisão, permitindo resultados mais eficientes e uma melhoria no poderio diagnóstico disponível.

Segundo Xing e Bei (2019) o rápido desenvolvimento tecnológico levou ao desenvolvimento de uma inteligência informacional no setor médico que pode ser aproveitado por meio da aplicação de algoritmos computacionais, como é o caso do KNN, conhecido por sua simplicidade de implementação. Entretanto, este algoritmo de classificação não se mostra tão eficiente ao lidar com os volumes de dados comuns na chamada *Big Data.* Assim, os autores propuseram modificações no algoritmo tradicional com a finalidade de melhorar seu desempenho sob estas circunstâncias a partir da aplicação de pesos em cada uma das possíveis classes, permitindo um processo de clusterização, redução de ruídos e modificações durante a fase de busca por vizinhos. Com isso, obtiveram sucesso na criação de um novo algoritmo de KNN para lidar com grandes quantidades de dados que mantêm a acurácia, porém apresenta melhor desempenho.

Os trabalhos de Solanki et al. (2021) e Ishaq et al. (2021) tratam de outro problema recorrente no processo de extração de conhecimento a partir da classificação de grande quantidade de dados: o desbalanceamento entre as classes. O primeiro aborda a aplicação de métodos de seleção de características bioinspirados e métodos *Wrapper* como técnica para diminuição da dimensionalidade dos dados, ressaltando os principais atributos que mais influenciam a acurácia dos algoritmos de *Machine Learning* e aplica esses algoritmos no *dataset* sobre câncer de mama da Universidade da Califórnia Irvin. Já o segundo, usa uma abordagem similar visando reduzir a quantidade de atributos em um conjunto de dados sobre doenças cardíacas por meio da aplicação de uma técnica de *oversampling* sintético dos dados da classe que possui menor quantidades de representantes (do inglês, *Synthetic Minority Oversampling Technique)* e, posteriormente, aplica diversos algoritmos de *Machine Learning* com a finalidade de comparar a acurácia em ambos os casos. Resultados promissores puderam ser observados em ambos os trabalhos.

Um trabalho correlacionado é o de Le et al. (2018) que também utiliza um conjunto de dados da Universidade da California Irvine, porém, sobre dados relacionados a problemas cardíacos. Como esse conjunto possui uma alta dimensionalidade, com 76 atributos, é proposto a utilização de algoritmos de *feature selection* com a finalidade de reduzir a quantidade de informações a serem processadas e melhorar o desempenho geral do método de classificação, que

no caso, é um *Support Vetor Machine* (SVM). A aplicação da etapa de seleção de características se mostrou auspiciosa quanto ao aumento da acurácia na classificação. De maneira similar Saba et al. (2019) realiza a seleção de características em dados cardíacos, mas propõe um estudo dos efeitos causados por tal prática em outros algoritmos de classificação como árvores de decisão, regressão logística, florestas aleatórias e o método de Naive Bayes.

O trabalho de Pei et al. (2020) também segue uma abordagem de predição diagnóstica, mas com algumas diferenças. O conjunto de dados trata sobre diabetes, e o processo de seleção de características não é realizado a partir de algoritmos e sim por meio do conhecimento prévio sobre os fatores de risco da doença por parte dos autores. Então, foi realizado o processo de classificação utilizando o algoritmo J48 atingindo uma acurácia satisfatória.

Na Tabela 3 é apresentado um comparativo entre os principais trabalhos correlatos apresentados nesta seção e o que é proposto pelo presente trabalho.

Tabela 3 - Comparativo entre os trabalhos correlatos

	(LE et al. 2018)	(SABA et al., 2019)	(PEI et al., 2020)
Aplicação de algoritmos de *feature selection*	✓	✓	✗
Utilização de múltiplos conjuntos de dados	✗	✗	✗
Comparação entre algoritmos de classificação	✗	✓	✗
Aplicação de algoritmos de árvore de decisão	✗	✓	✓
Aplicação de algoritmos baseados em instâncias	✗	✗	✗

Fonte: Elaborado pelo autor.

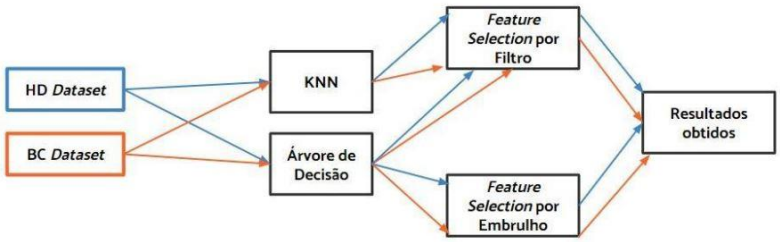
2.6 Considerações finais

Neste capítulo foi apresentada a fundamentação teórica na qual esse trabalho está baseado. Entre os temas estudados estão a relação entre os dados médicos com suas particularidades e *Big Data*, as etapas necessárias do KDD para que seja possível extrair conhecimento de forma eficiente dos dados mencionados, dando maior destaque para os processos de *Data Clean* e *Data Mining*. Em seguida foi discorrido sobre o funcionamento de algoritmos de classificação e de seleção de características. Finalmente, foi feita uma breve discussão sobre o estado da arte e os trabalhos correlatos, sendo possível observar que dentre os trabalhos não houve uma combinação que resultasse em um trabalho que aplicasse métodos de *feature selection* em mais de um conjunto de dados e realizasse a comparação entre múltiplos algoritmos de classificação baseados em árvores de decisão e métodos baseados em instâncias.

3 Desenvolvimento

Conforme mencionado na sessão 1.3, foi conduzido um estudo inicial acerca dos temas relevantes para este trabalho como: processos e etapas do KDD, métodos e técnicas de *feature selection*, algoritmos de classificação e conceitos de *machine learning* para possibilitar a predição de resultados. Neste capítulo é discutida a implementação de todo processo proposto, que está demonstrado na Figura 6, iniciando-se na escolha dos conjuntos de dados até obtenção do resultado da classificação, passando pelas etapas de limpeza, manipulação, seleção de atributos e classificação.

Figura 6 – Processos do Presente Trabalho



Fonte: Elaborado pelo autor.

3.1 Seleção dos Conjuntos de Dados

Ambos os conjuntos de dados utilizados são disponibilizados com fácil acesso na internet, aparecendo de forma recorrente na literatura como apoio para aferir performance de diversos tipos de técnicas computacionais que lidam com dados médicos.

3.1.1 *Heart Disease Data set*

O conjunto de dados intitulado *Heart Disease Dataset* (HD Dataset) pode ser encontrado gratuitamente no repositório da UCI (JANOSI et al., 1988), e contém informações sobre pacientes com doenças cardíacas de quatro localidades: Budapest, Hungria e fornecido por Andras Janosi, Zurique, Suíça e fornecido por William Steinbrunn, na Basiléia, Suiça e fornecido por Matthias Pfisterer e Cleveland, Estados Unidos da América e fornecido por Robert Detrano. Essas informações são compostas por 76 atributos relativos a informações do paciente como idade,

gênero, presença, ou não, de dor no peito, se é, ou não, fumante, valor de pressão sanguínea, histórico familiar, entre outros. A documentação oficial disponibilizada pela universidade está presente no repositório e menciona a existência de um subconjunto de 14 atributos frequentemente referenciados pelos trabalhos feitos utilizando esse conjunto de dados como objeto de estudo. Esses atributos são:

1. Idade
2. Gênero
3. Tipo de dor no peito
4. Pressão Arterial
5. Níveis de Colesterol
6. Nível de açúcar no sangue
7. Resultados de eletrocardiograma
8. Frequência cardíaca máxima na prática de exercícios
9. Dor após atividade física
10. Índice ST (valor presente no eletrocardiograma) durante exercício em relação à situação de repouso
11. Declínio dos valores de ST
12. Quantidade de vasos observados em exame de fluoroscopia
13. Presença ou não de cardiomiopatia causada por thalassemia
14. Diagnóstico (doente ou não)

Este subconjunto de atributos foi utilizado como referência para auxiliar na avaliação da execução dos métodos de seleção de características neste conjunto de dados, uma vez que predições utilizando esse subconjunto se mostram mais acuradas do que classificações utilizando a totalidade dos atributos presentes.

3.1.2 *Breast Cancer Wisconsin (Diagnostic) Data Set*

O conjunto de dados intitulado *Brest Cancer Wiscosin (Diagnostic) Data Set* (BC Dataset) é sobre o tipo de câncer mais comum em mulheres no mundo Ocidental (SARITAS; YASAR, 2019), o câncer de mama, e pode ser encontrado gratuitamente no repositória da UCI (WOLBERG et al., 1995) e contém informações retiradas de biópsias por punção aspirativa utilizando agulha fina em tecido mamário. Os dados são foram obtidos a através da extração de características das imagens obtidas nas biópsias e são divididos em até 32 atributos dependendo da quantidade de núcleos presentes na célula tumoral. Os atributos são:

1. ID
2. Diagnóstico (tumor benigno ou maligno)

3. Dados particulares de cada um dos núcleos, podendo variar entre 1 e 3:
    a. Raio do núcleo
    b. Textura
    c. Valor do perímetro
    d. Área
    e. Suavidade
    f. Compacidade
    g. Concavidade
    h. Número de pontos de concavidade no contorno do núcleo
    i. Simetria
    j. Dimensão fractal

Este conjunto de dados foi igualmente utilizado para referência ao avaliar o desempenho dos classificadores sem e com a aplicação de algoritmos de seleção de características.

3.2 Limpeza dos Dados

Ambos os conjuntos de dados (JANOSI et al., 1988) (WOLBERG et al., 1995) foram obtidos por meio de um arquivo de texto plano, fator que dificultaria o acesso e o uso dos dados de forma direta e consistente, visto que cada atributo teria que ser acessado por um índice numérico pouco semântico e de difícil análise. Além disto, observou-se a falta de valores para alguns atributos em determinadas instâncias, sendo necessária sua eliminação. Para realizar essa e outras manipulações necessárias nos conjuntos de dados foi utilizada a linguagem de programação Python em sua versão 3.8.8 e a biblioteca Pandas, construída em Python, cujas funções atuam na manipulação e análise de dados.

Os arquivos de texto de ambos conjuntos de dados foram abertos e modificados para que cada uma de suas colunas receba o rótulo atribuído a ela pela documentação oficial de cada um dos conjuntos. Após isso, os dados foram processados e salvos de uma forma estruturada em um arquivo de planilha no formato *csv* (do inglês, comma separated values).

A partir desse novo arquivo foi realizada a limpeza dos dados no qual foram removidas as linhas que continham atributos vazios. Para garantir uma maior integridade dos dados, realizou-se uma análise acerca do valor semântico de cada coluna: no primeiro conjunto, referente a casos de câncer, verificou-se que a coluna “id” não possuía valores relevantes para o processo de classificação e que a coluna “Unamed: 32” não apresentava valores consistentes, sendo ambos atributos removidos. Também foi retirada a coluna que contêm o rótulo de cada instância, no caso a coluna “Diagnosis”, porém esta foi salva a parte para ser utilizada posteriormente no processo

de classificação. No segundo conjunto, com dados de doenças cardíacas, eliminou-se a coluna “id” por também não influenciar na classificação e a coluna “target” foi isolada por conter os rótulos dos dados, também sendo usada na classificação.

Com os dados já processados analisou-se mais detalhadamente os valores presentes com o intuito de entender melhor sua distribuição e as quantidades presentes, através de questões como quantidade, média e valores mínimos e máximos. Através dessa análise observou-se que o intervalo de valores referente a cada atributo era desigual e muito extenso, de modo que, para facilitar a atuação dos algoritmos e minimizar a ocorrência de erros, optou-se por padronizar os valores, de modo que os todos os dados foram trabalhados em um intervalo de 0 a 1.

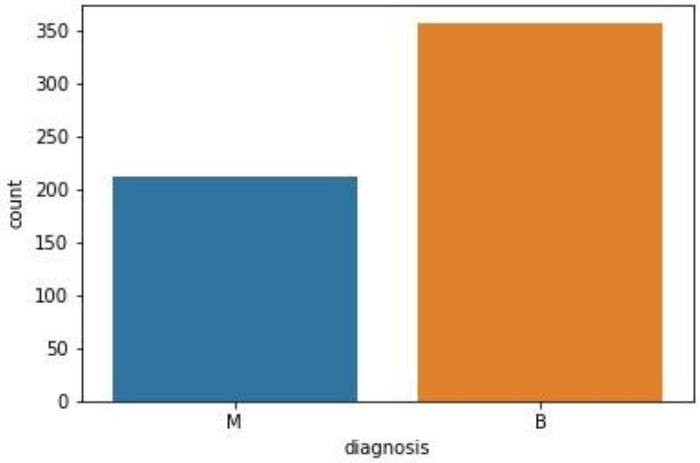
3.3 Visualização e Interpretação dos Dados

Após o processo de limpeza dos dados obteve-se uma lista com os atributos remanescentes de cada conjunto, entretanto ainda não é possível realizar afirmações categóricas sobre o que eles representam e quais informações eles carregam. Portanto, antes de iniciar a aplicação dos algoritmos de *feature selection* e classificação é necessária uma análise exploratória a fim de compreender o valor semântico de cada atributo no rótulo final. As etapas dessa análise foram feitas de forma paralela para ambos conjuntos de dados. Todos os gráficos presentes nesta seção foram construídos por meio de bibliotecas de criação de gráfico e visualização presentes na linguagem Python como *matplotlib* e *seaborn*, ressaltando que os gráficos gerados por meio destas bibliotecas possuem seus atributos em inglês uma vez que são gerados dinamicamente a partir dos conjuntos de dados.

3.3.1 Análise do Primeiro Conjunto de Dados

Inicialmente, realizou-se um balanço sobre a quantidade de elementos em cada uma das classes do conjunto, benigno (B) e maligno (M), como uma forma de verificar se os dados estavam ou não balanceados. Conforme ilustrado na Figura 6, identificou-se que a quantidade de dados da classe com mais elementos corresponde a 62,74% do tamanho do conjunto, confirmando que, estatisticamente, o conjunto está balanceado.

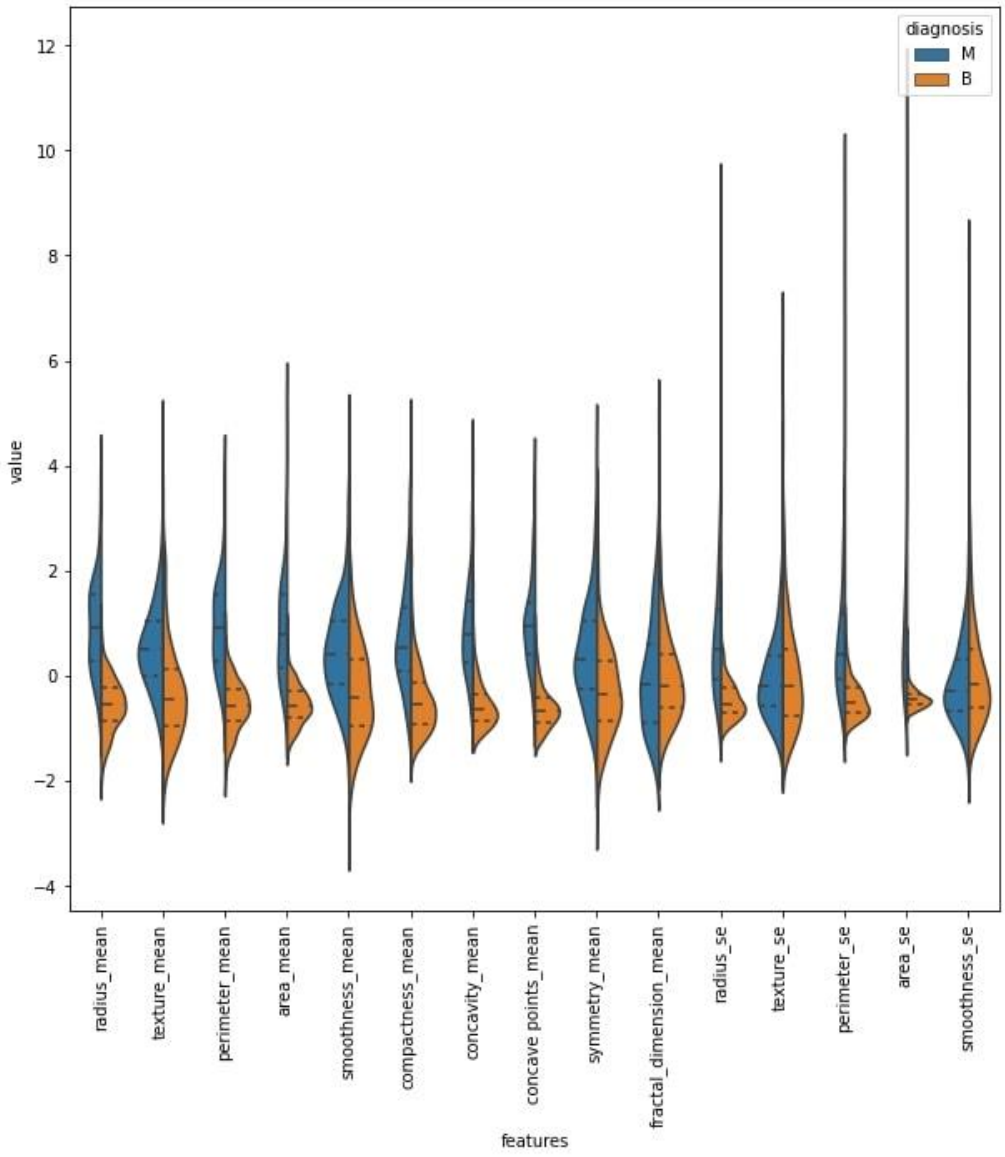
Figura 7 – Gráfico de Barras de Quantidade de Classes



Fonte: Elaborado pelo autor.

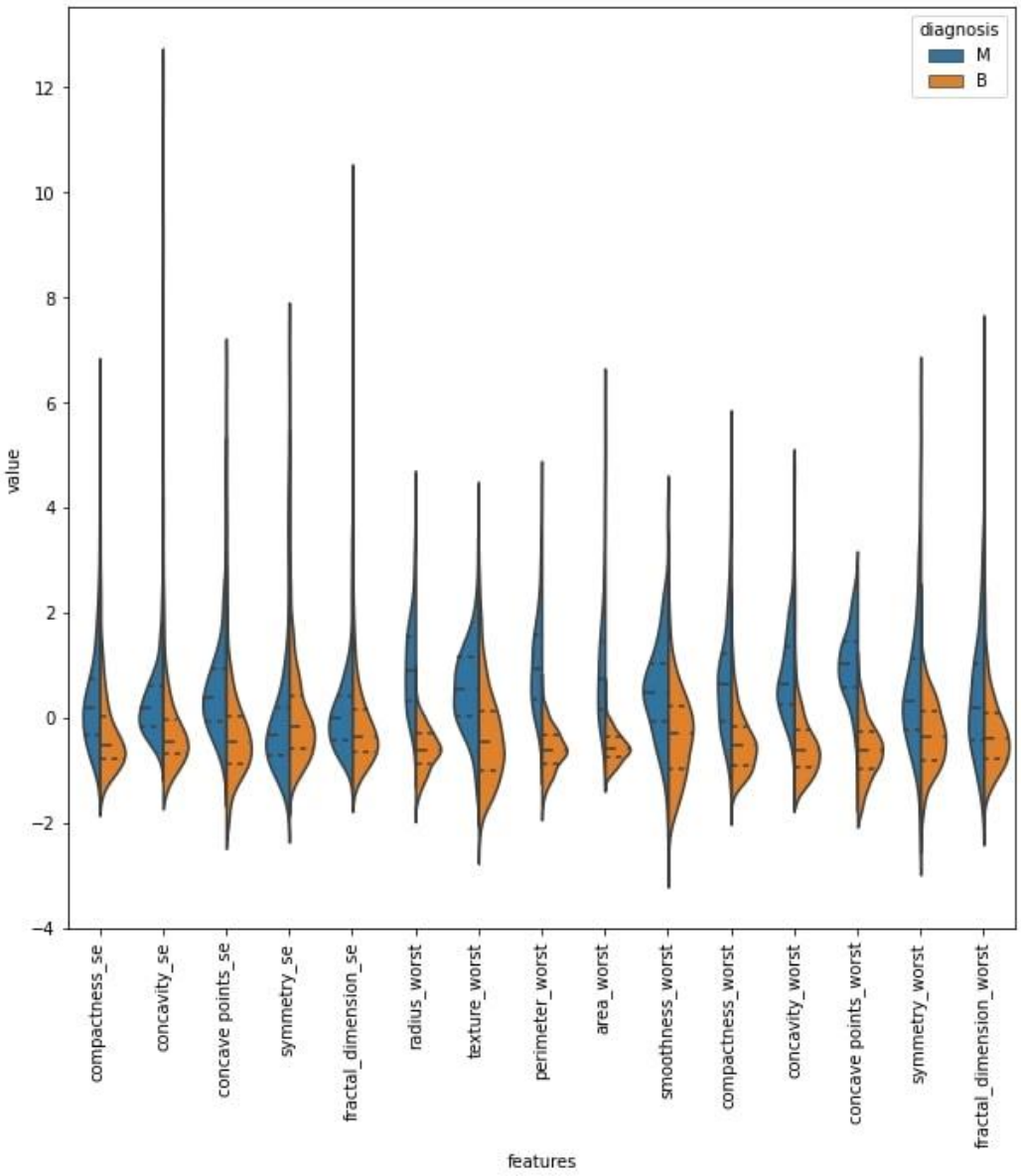
Após isso, foi feito um estudo analítico sobre como os atributo do conjunto de dados estão relacionados com as classes e entre si. Para isso foram gerados gráficos violinos e gráficos de espalhamento, uma vez que estes tipos de gráficos ilustram de maneira mais intuitiva a relação de um atributo com uma classe alvo, assim como possíveis semelhanças entre os atributos. O gráficos de violino consistem de uma forma de visualizar a relação entre as medianas de cada um dos elementos do conjunto de dados em relação a sua classificação, a partir deles é possível observar quando dois ou mais atributos possuem uma curva semelhante, sendo um indicativo que são atributos correlatos, ou se possuem as medianas de ambas as classes muito próximas, sendo um indicativo de que aquele atributo não possui influência significativa na classificação. Já a visualização por espalhamento permite observar uma representação de cada um dos elementos do conjunto de dados junto a sua classificação, permitindo assim, a análise da importância daquele atributo para a classificação, sendo que nos casos que o atributo possui grande importância, é de se esperar que os semelhantes fiquem próximos. Os gráficos obtidos foram repartidos em três gráficos menores apresentados nas Figuras 7, 8 e 9, a fim de facilitar a visualização dos mesmos.

Figura 8 – Gráfico de Violino 1



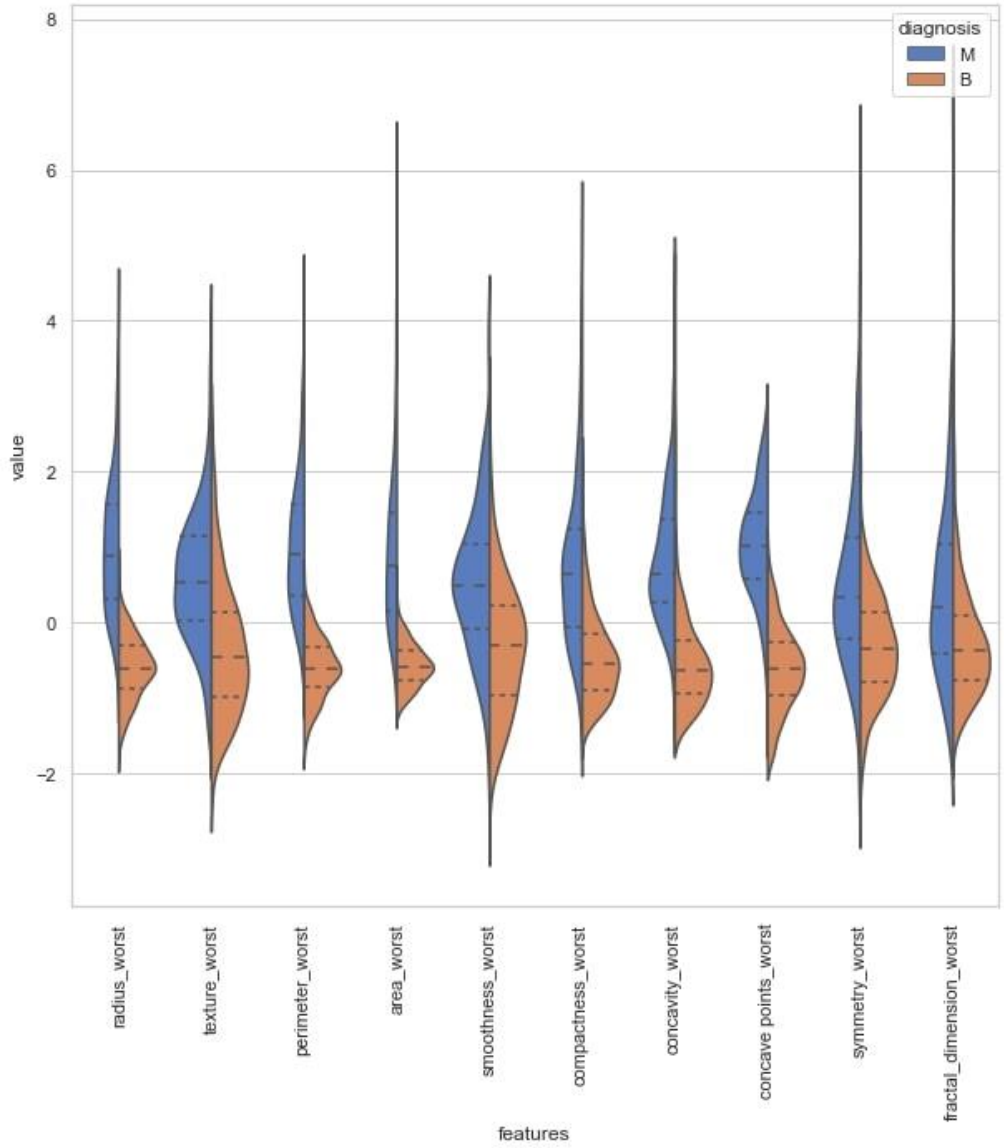
Fonte: Elaborado pelo autor.

Figura 9 – Gráfico de Violino 2



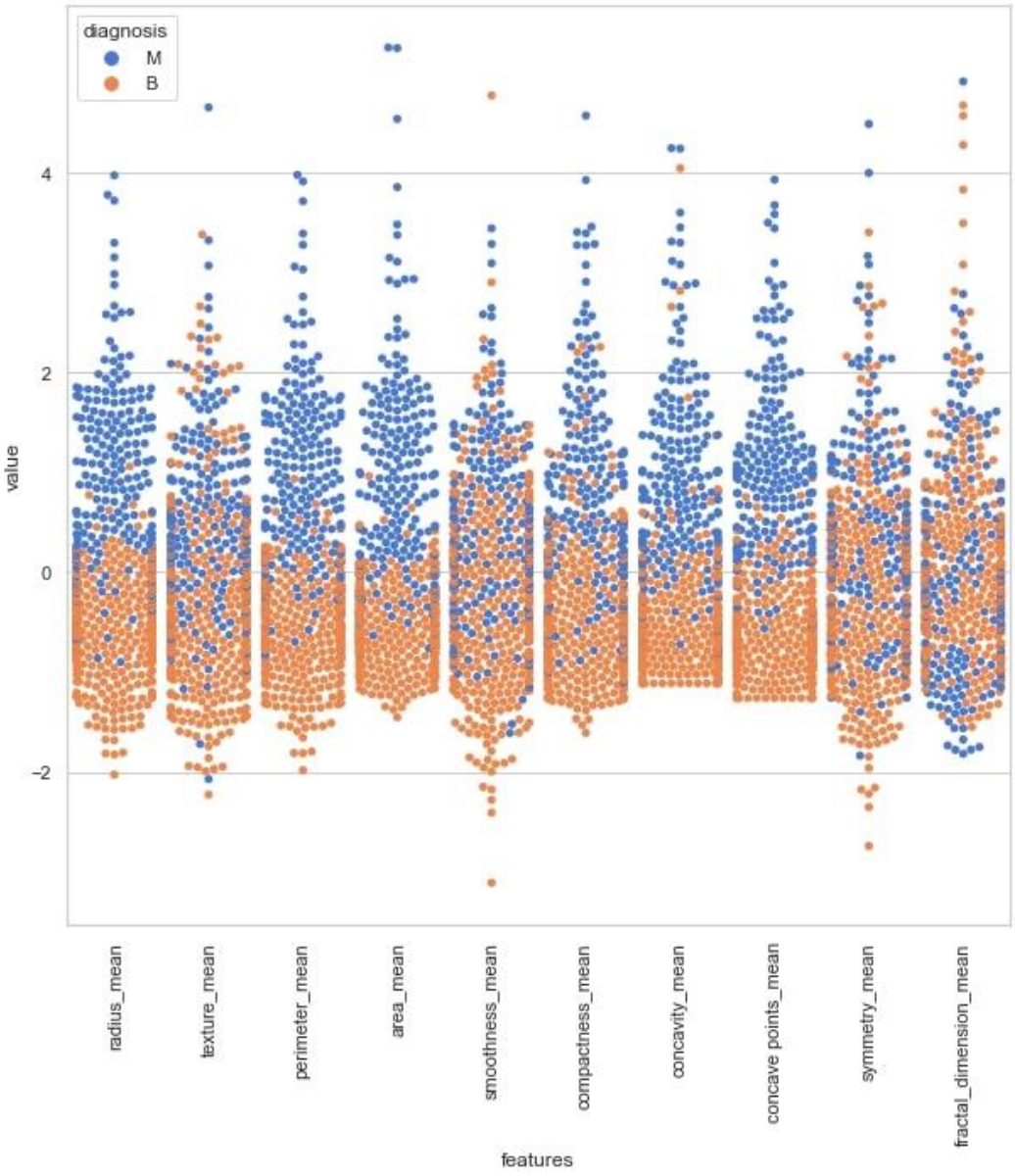
Fonte: Elaborado pelo autor.

Figura 10 – Gráfico de Violino 3



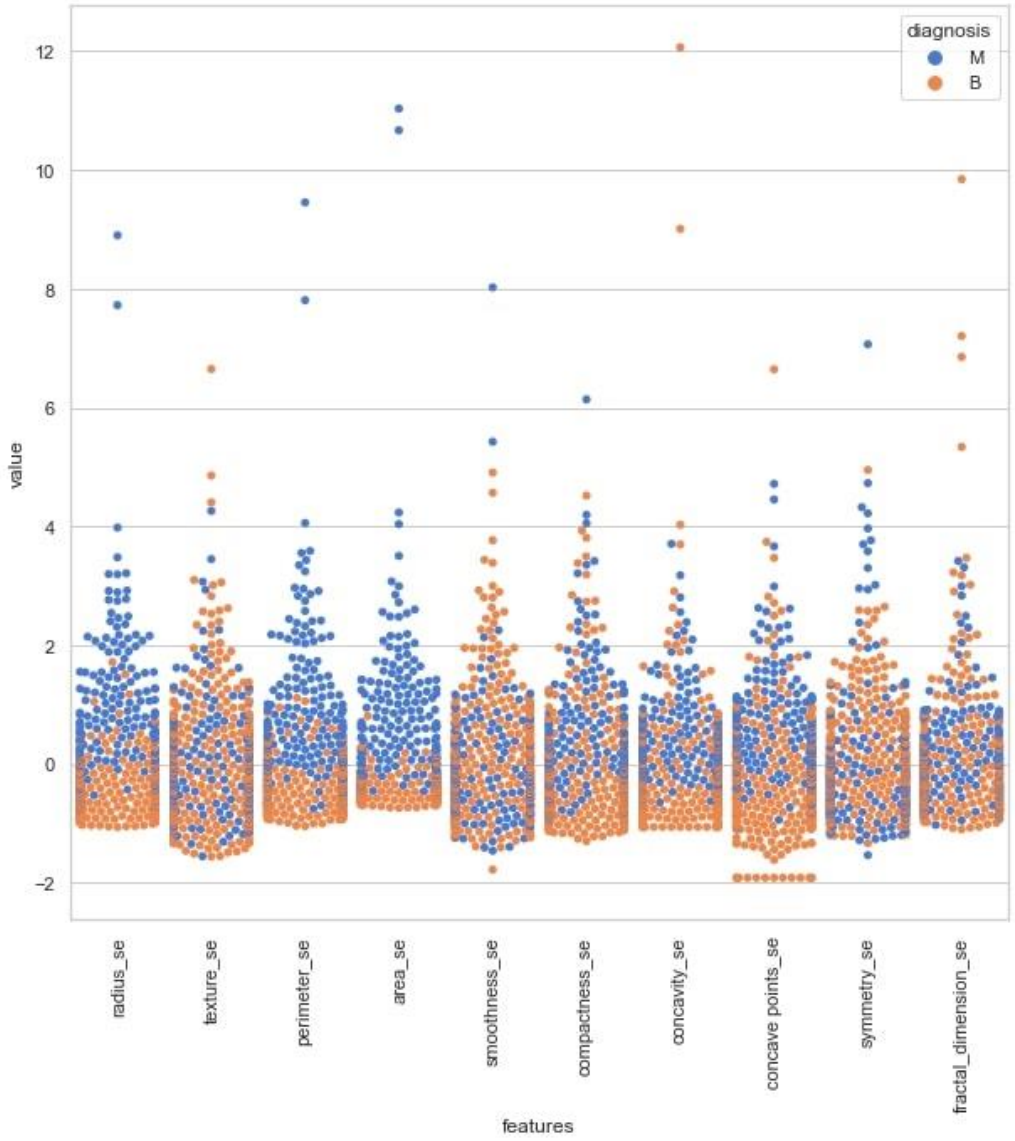
Fonte: Elaborado pelo autor.

Figura 11 – Gráfico de Espalhamento 1



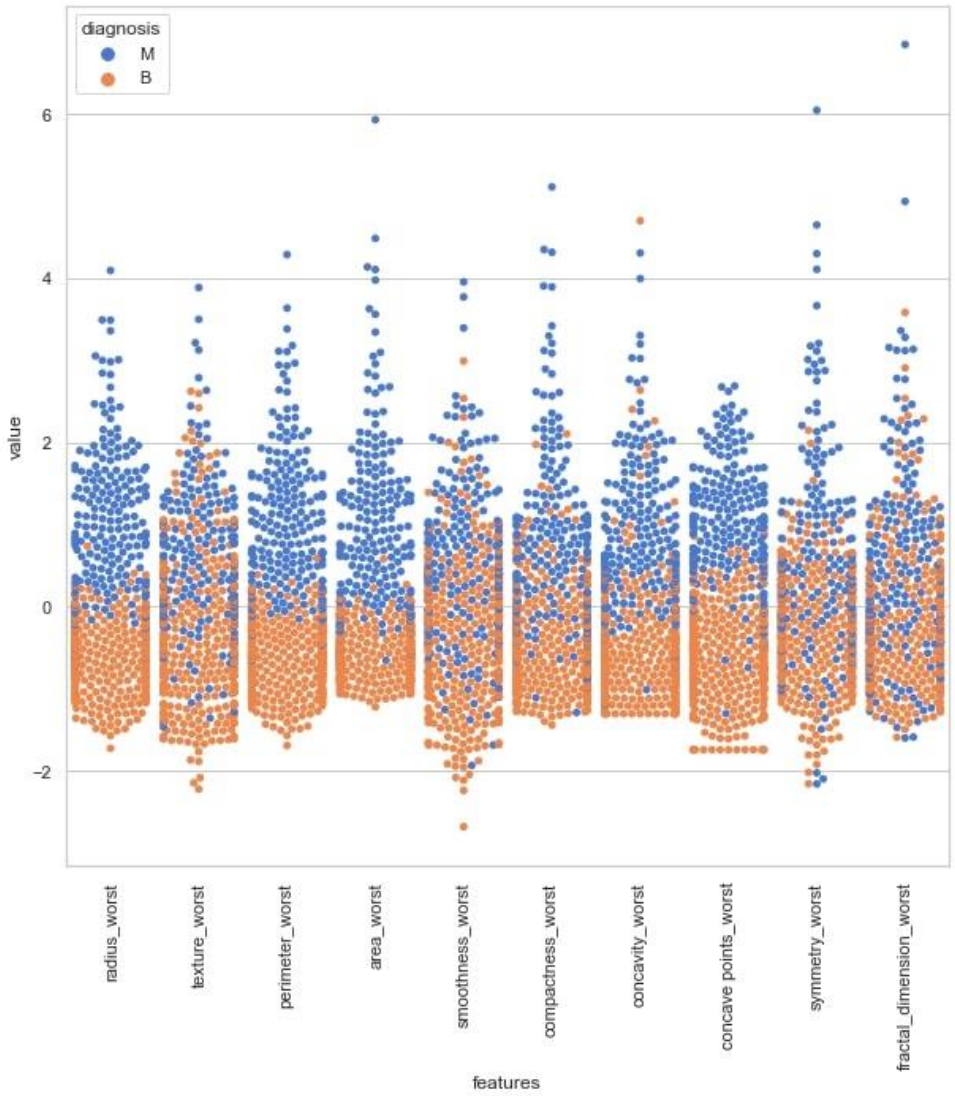
Fonte: Elaborado pelo autor.

Figura 12 – Gráfico de Espalhamento 2



Fonte: Elaborado pelo autor.

Figura 13 – Gráfico de Espalhamento 3



Fonte: Elaborado pelo autor.

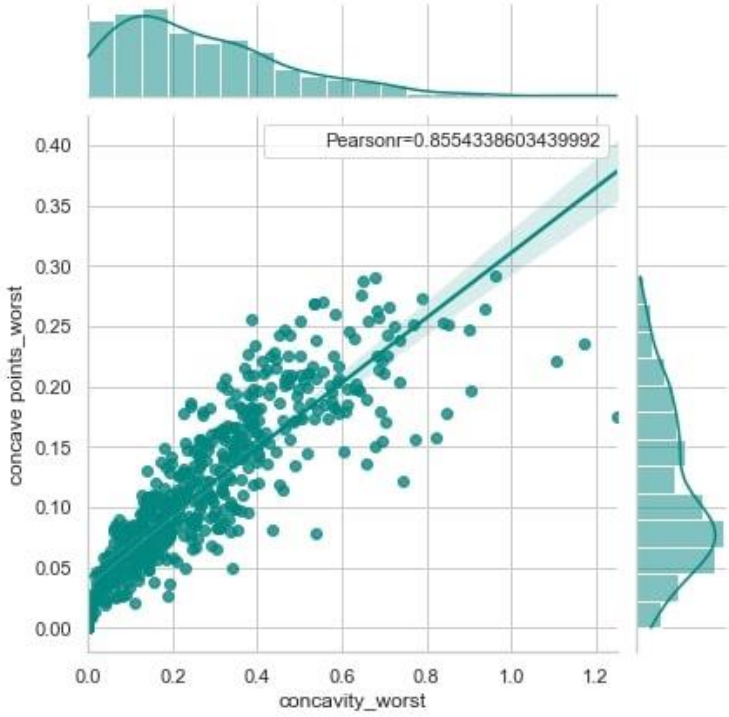
A partir da análise dos gráficos apresentados nas Figuras 7, 8 e 9 foi possível observar a correlação entre os atributos e o peso de cada um na determinação e atribuição dos rótulos. Para melhor aproveitamento dos gráficos violino é preciso calcular a mediana de cada atributo, uma vez que situações onde as medianas de cada classe diferem entre si são indicativos de que aquele atributo influencia significativamente no processo de classificação. Por exemplo, observa-se que na Figura 7 as medianas de cada classe considerando o atributo "*texture_mean"* são distintas enquanto no atributo "*fractal_dimension_mean*" elas são semelhantes, indicando que este não é um atributo relevante para a classificação, enquanto o primeiro tende a ser.

Com os gráficos de dispersão verifica-se a tendência dos pontos em cada um dos atributos, de modo que situações onde os pontos não possuem uma divisão nítida tendem a refletir uma baixa influência do atributo na classificação, enquanto casos em que os pontos apresentam a

divisão entre as classes de modo claro indicam uma maior influência daquela característica na classificação. Por exemplo, a Figura 12 ilustra uma divisão bem definida entre as classes "benigno" e "maligno" no atributo "*area_worst*", indicando que este está no conjunto de atributos com grande relevância no processo de classificação, enquanto "*smoothness_se*" possui pontos muito próximos e embaralhados entre si, levando a concluir que este não possui muito valor semântico.

Em casos em que um atributo se assemelha a outro é necessário estudar a correlação entre eles para que se trabalhe a dimensionalidade dos dados. Para isso, foi criado um gráfico conjunto e calculou-se o coeficiente de correlação (R de Pearson) para cada par de atributos cuja correlação mereça uma maior atenção. Um exemplo dessa situação é apresentado na Figura 13 com a comparação entre os atributos "*concave_points_worst*" e "*concavity_worst*" onde o valor do coeficiente de correlação indica uma semelhança aproximadamente de 85,5%.

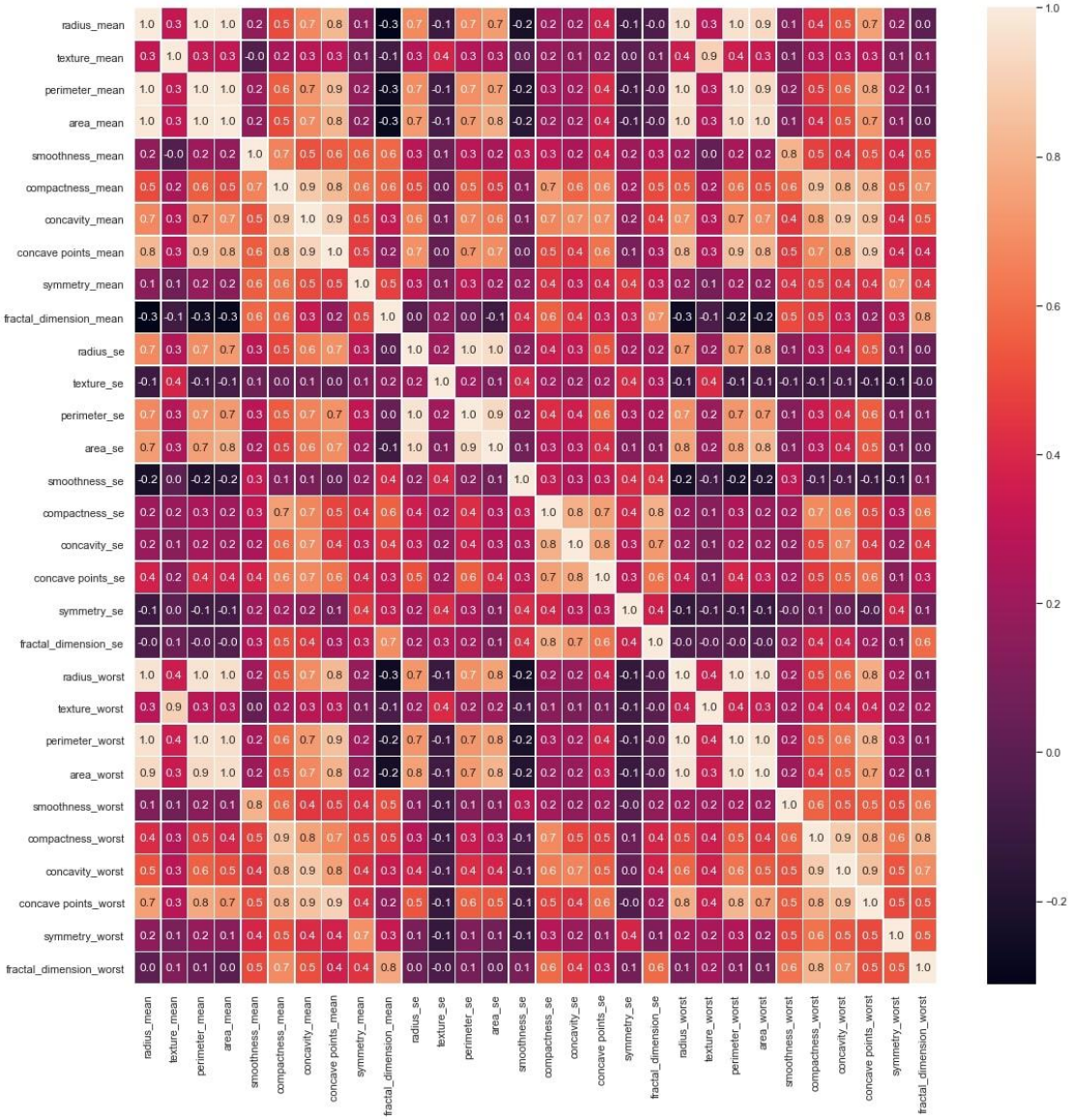
Figura 14 – Gráfico Conjunto 1



Fonte: Elaborado pelo autor.

Assim construiu-se um mapa de calor com o valor do coeficiente de correlação entre todos os atributos, possibilitando uma melhor visualização dos valores. O gráfico é apresentado na figura 14.

Figura 15 – Gráfico de Calor 1

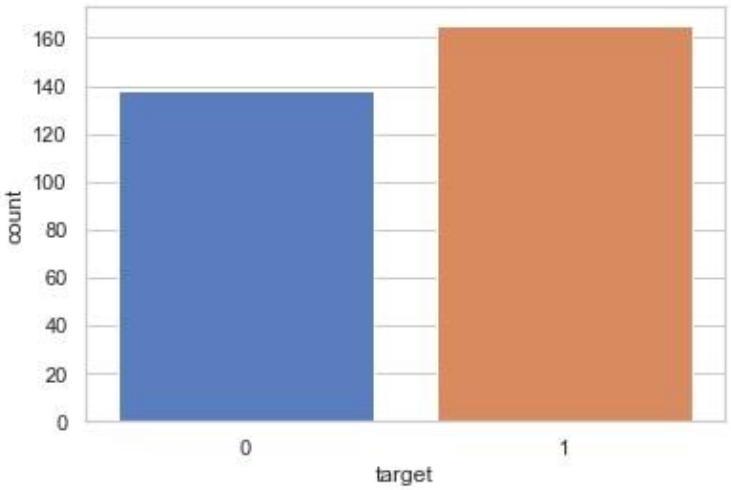


Fonte: Elaborado pelo autor.

3.3.2 Análise do Segundo Conjunto de Dados

O processo explicitado anteriormente foi repetido para o segundo conjunto de dados, porém as classes são representadas por “0” (Ausência de Doença Cardíaca) e “1” (Presença de Doença Cardíaca). Conforme mostrado na Figura 15 a classe predominante representa um total de 54,45% dos dados presentes, permitindo a continuidade da análise sem necessidade de uma etapa de balanceamento.

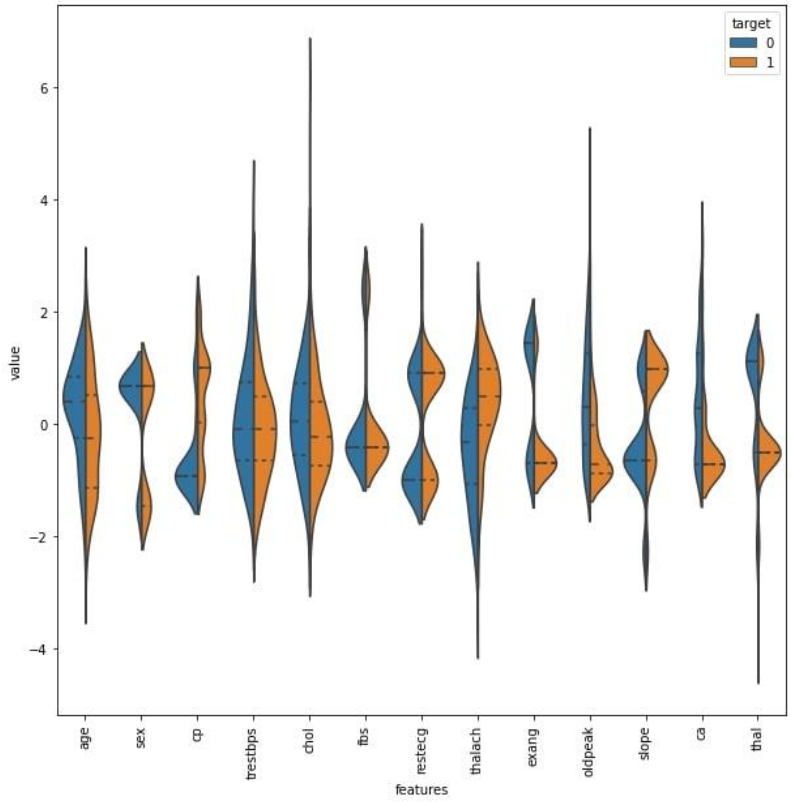
Figura 16 – Gráfico Barras 2



Fonte: Elaborado pelo autor.

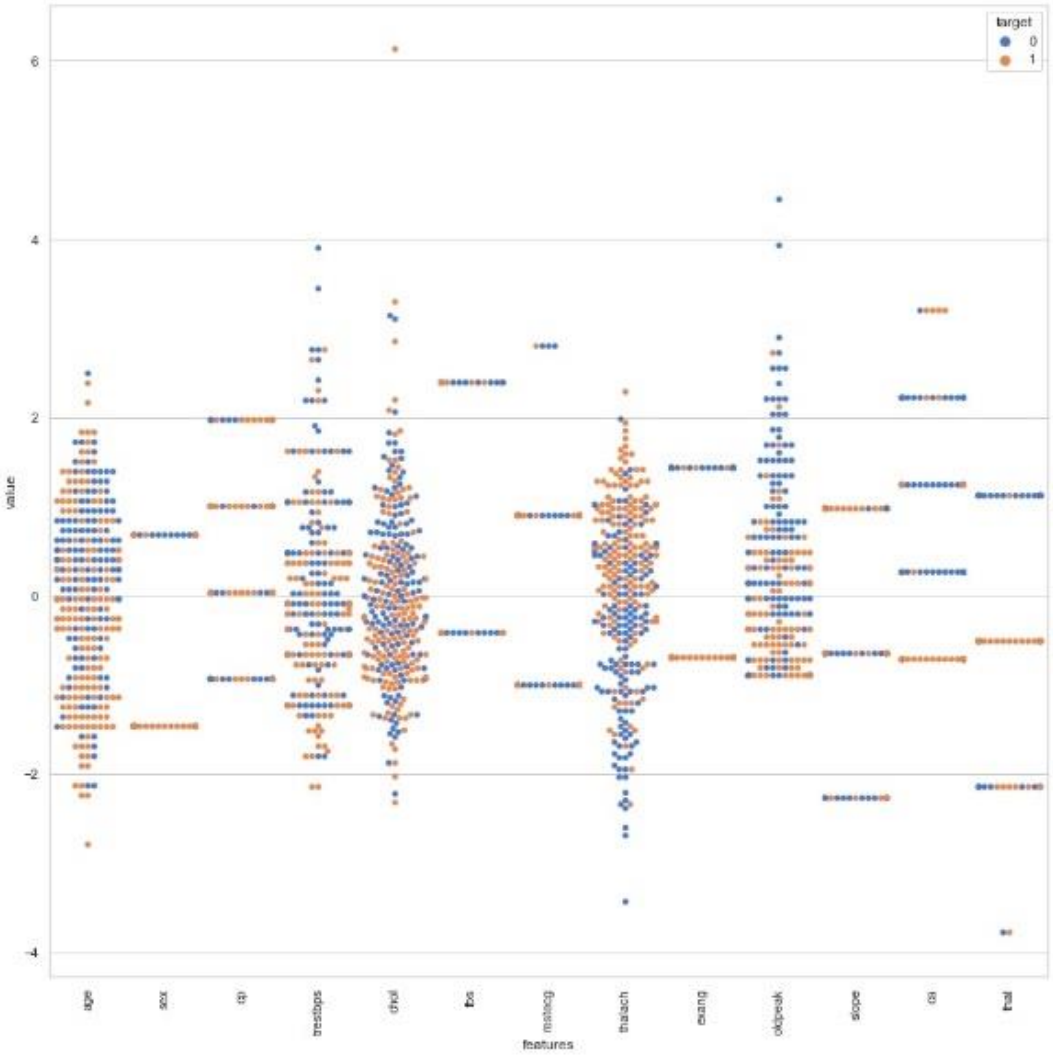
Prosseguiu-se com o estudo analítico sobre como as características do conjunto de dados são correlacionadas. Para isso foram gerados gráficos violinos e gráficos de espalhamento com a intenção de visualizar a relação de uma classe alvo com cada atributo, e as semelhanças entre eles. Neste caso, há um total de 14 atributos no conjunto de dados, não havendo necessidade de dividir o gráfico como na situação anterior.

Figura 17– Gráfico Violino



Fonte: Elaborado pelo autor.

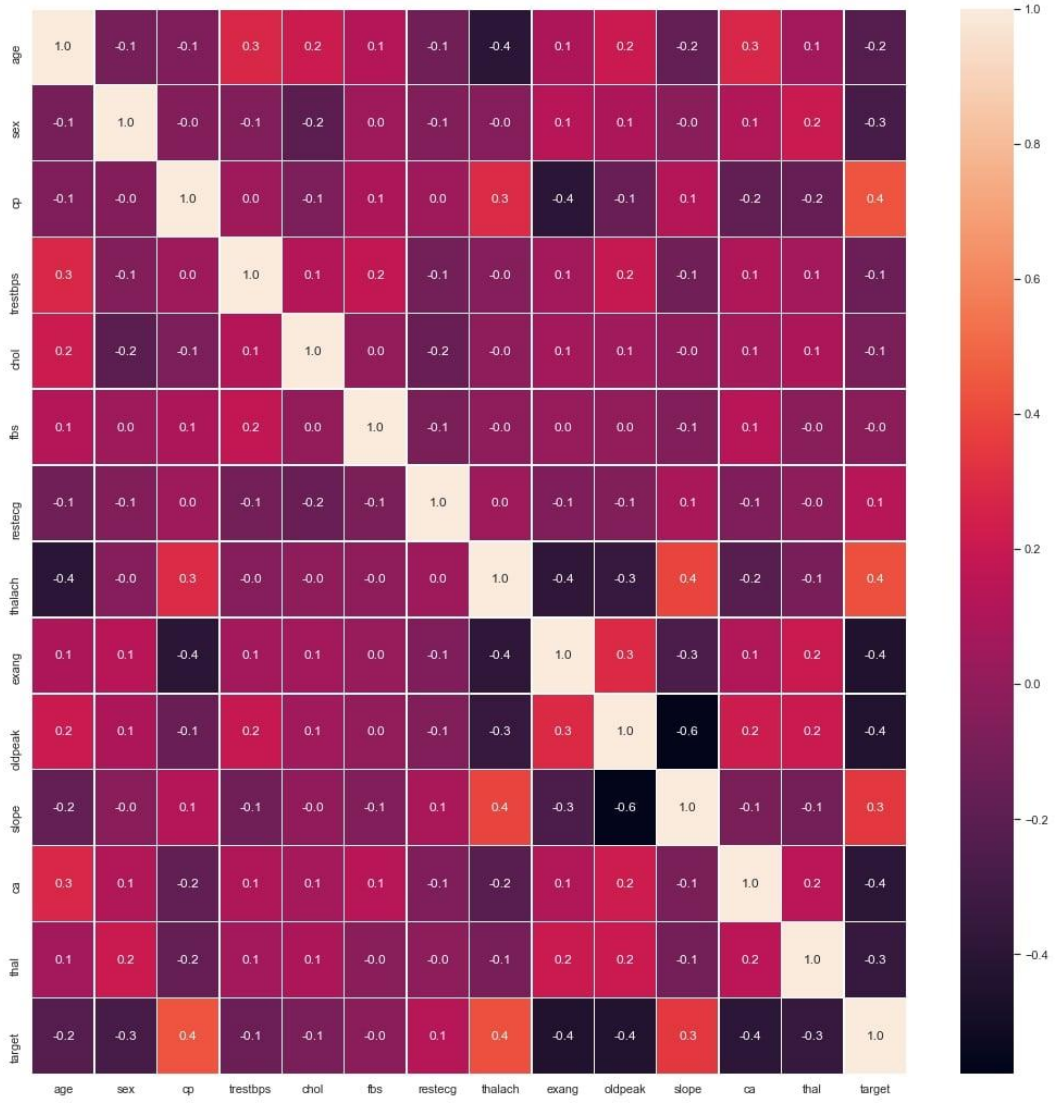
Figura 18– Gráfico Scatter 2



Fonte: Elaborado pelo autor.

Ao observar as Figuras 16 e 17 não é possível afirmar com certo grau de certeza quais atributos são mais relevantes para a definição das classes, refletindo também na visualização das correlações entre os atributos. O mapa de calor também não apresenta informação conclusiva, uma vez que o maior coeficiente de correlação é de apenas 0,4, conforme apresentado na Figura 18. Esse comportamento gera a necessidade de mais atenção ao realizar a seleção dos atributos.

Figura 19 – Gráfico de Calor 2



Fonte: Elaborado pelo autor.

3.4 Seleção de Atributos e Classificação

O processo de seleção de atributos contribui para a eliminação de atributos poucos relevantes, melhorando a qualidade e o custo computacional da classificação, uma vez que este último está relacionado à dimensionalidade dos dados de entrada. Assim, a aplicação de diferentes métodos de *feature selection*, é o ponto central deste trabalho, sendo usados métodos de filtro e métodos de embrulho. Os métodos de filtro aplicados selecionados foram colunas colinares por meio da análise gráfica, método univariado por meio do algoritmo *Select K Best* e *Principle Componente Analysis* (PCA), enquanto os métodos de embrulho escolhidos foram *Recursive Feature Elimation* (RFE) e *Recursive Feature Elimination Cross-Validation* (RFEC).

Para o processo de classificação, foram utilizados o algoritmo K-Neareast Neighbours (KNN) e o algoritmo CART (*Classification and Regression Trees)*, que consiste no algoritmo J48 com suporte para variáveis alvo numéricas. Ambos os algoritmos foram implementados com o suporte da biblioteca de código aberto *scikit-learn* (*sklearn*) que foi desenvolvida voltada para o escopo de aprendizado de máquina. O software *WEKA* também foi utilizado para auxiliar na aplicação dos métodos de seleção de atributos, oferecendo também uma estimativa do subconjunto ideal de atributos para maximizar o resultado das classificações.

Inicialmente foi feita a classificação dos dados de ambos os conjuntos pelos dois algoritmos sem a aplicação de métodos de *feature_selection* de modo a criar resultados de referência para posterior avaliação dos resultados obtidos após a realização das etapas descritas nesta seção.

3.4.1 Algoritmos de Seleção por Filtro

Os algoritmos de seleção por filtro funcionam de forma independente do método de classificação a ser utilizado, sendo possível aplicá-los anteriormente à execução de ambos os algoritmos de classificação nos dois conjuntos de dados sem dificuldades.

Baseando-se as análises gráficas e de correlação apresentadas na sessão 3.3. foram definidas as colunas com valores redundantes e eliminou-se as mesmas do conjunto final que foi utilizado como entrada para a etapa de treinamento dos classificadores. Então os classificadores são executados e sua acurácia e tempo de execução são armazenados.

Primeiro aplicou-se o algoritmo de seleção de atributos univariado *Select K-Best*, presente na biblioteca *sklearn* , que seleciona os atributos com as maiores pontuações baseados em no cálculo estatístico qui-quadrado, seguido pela execução do método não supervisionado *PCA* que realiza combinações lineares de uma matriz de amostras e atributos com o intuito de encontrar quais os atributos que mais impactam a classificação baseando-se em sua variância.

3.4.2 Algoritmos de Seleção por Embrulho

Como explicitado na sessão 2.4.2, este tipo de algoritmo é dependente do método de classificação para ser executado, porém o algoritmo KNN não possui suporte lógico para sua aplicação, baseando-se no cálculo de distância do dado a ser classificado em relação aos dados de treino, de modo que a quantidade de atributos não influencia nesse cálculo. Assim, os métodos dessa categoria foram implementados em ambos os conjuntos de dados, mas atuando somente com o algoritmo CART.

O algoritmo de eliminação recursiva de atributos realiza a adequação de um modelo a partir da remoção dos atributos menos significativos partindo do conjunto inicial de dados, gerando subconjuntos cada vez menores até que a quantidade desejada de atributos seja alcançada.

O funcionamento do algoritmo RFEC é muito semelhante ao citado acima, com a principal diferença sendo a presença de um estimador de validação cruzada que permite uma maior validação de resultados a partir de valores computados em etapas prévias do processo.

3.5 Considerações Finais

Neste capítulo foram apresentados os processos envolvidos no desenvolvimento das associações entre algoritmos de seleção de atributos e de classificação. Foram utilizados os dados presentes nas bases de dados sobre câncer de mama e doenças cardíacas para a execução dos métodos citados. Além disso, realizou-se a implementação dos métodos com o auxílio de bibliotecas já consolidadas e amplamente difundidas.

Em sequência, foram explicitados quais os algoritmos aplicados para o processo de seleção de atributos e classificação. Os métodos tiveram seus funcionamentos elaborados e respectivas limitações evidenciadas.

Portanto, neste trabalho foi desenvolvida uma associação entre técnicas de seleção de atributos e de classificação que permite a obtenção de resultados mais acurados e com menor gasto computacional, possibilitando assim diagnósticos médicos mais rápidos e precisos.

4 Avaliação Experimental

Neste capítulo é apresentada a estratégia de testes adotada, os experimentos realizados e os resultados obtidos pelas combinações de algoritmos proposta no presente trabalho. Ao final os resultados são discutidos.

4.1 Estratégia de Testes

Os experimentos foram realizados a fim de validar as melhorias propostas na sessão 1.2. Para tal, em cada um dos casos, foram avaliados três métricas distintas:

- **Matriz de confusão**: consiste em um recurso que auxilia na compreensão dos resultados de um classificador podendo ter quatro resultados distintos para cada classificação realizada – verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo – onde os dois primeiros refletem a quantidade de classificações corretas. A estrutura de uma matriz de confusão é apresentada na tabela 4. E seus elementos são utilizados no cálculo da acurácia
- **Acurácia:** porcentagem total de classificações corretas e pode ser calculada a partir da soma dos valores de verdadeiro negativo e verdadeiro positivo e da divisão dessa soma pelos quatro valores presentes na matriz de confusão, conforme demostrado na Figura 20.
- **Tempo de execução** de cada uma das combinações.

Para todos os experimentos foram realizadas mil execuções para cada caso a fim de garantir um resultado estatisticamente significante e com a possibilidade de eliminação de outliers.

Tabela 4 – Estrutura de uma matriz de confusão.

Real	Classificado como negativo	Classificado como positivo
Negativo	Verdadeiro negativo	Falso positivo
Positivo	Falso negativo	Verdadeiro positivo

Fonte: Elaborado pelo autor.

Figura 20 – Formula do Cálculo de Acurácia

Fonte: Elaborado pelo autor.

4.2 Resultados e Discussões

A discussão dos resultados será feita considerando primeiramente o conjunto de dados *Breast Cancer Wisconsin (Diagnostic)* e depois o *Heart Disease Dataset*.

4.2.1 *Breast Cancer Wisconsin (Diagnostic)*

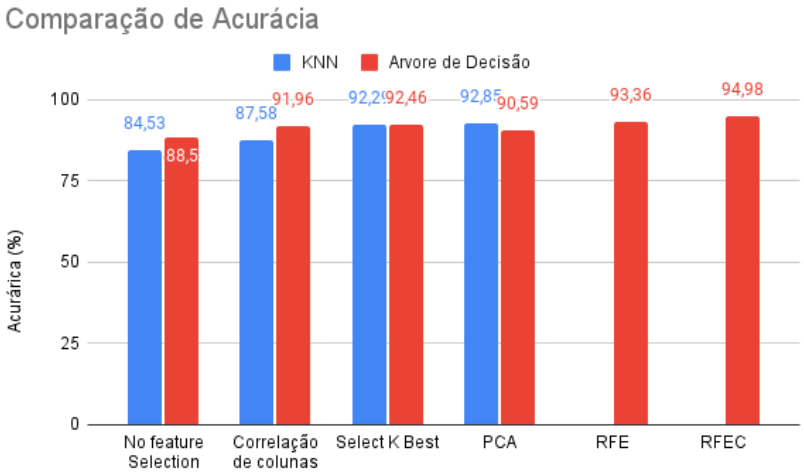
Os resultados de acurácia obtidos estão apresentados na tabela 5 e na figura 19, sendo o valor da acurácia uma média de todas as mil execuções, em porcentagem, com valor arredondando para a segunda casa decimal. Os tempos de execução, em segundos, são apresentados na tabela 6 e na figura 20. Vale relembrar que o algoritmo de classificação KNN não apresenta suporte lógico para a aplicação de métodos de seleção de atributos baseados em embrulho (RFE e RFEC), e, portanto, essa situação não é considerada.

Tabela 5 – Valores de acurácia obtidos em cada caso (em porcentagem).

	KNN	Arvore de Decisão
No feature Selection	84,53	88,50
Correlação de colunas	87,58	91,96
Select K Best	92,29	92,46
PCA	92,85	90,59
RFE	-	93,36
RFEC	-	94,98

Fonte: Elaborado pelo autor.

Figura 21 – Comparação da Acurácia.



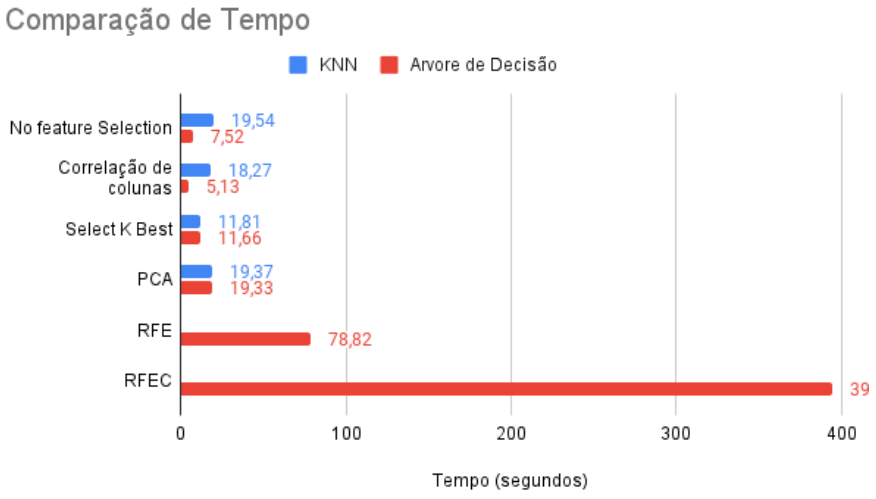
Fonte: Elaborado pelo autor.

Tabela 6 – Tempo de execução de cada caso (em segundos).

	KNN	Arvore de Decisão
No feature Selection	19,54	7,52
Correlação de colunas	18,27	5,13
Select K Best	11,81	11,66
PCA	19,37	19,33
RFE	-	78,82
RFEC	-	393,84

Fonte: Elaborado pelo autor.

Figura 22 – Comparação de Tempo.



Fonte: Elaborado pelo autor.

Ao analisar os resultados apresentados foi observado que a redução da dimensionalidade dos dados ocasionou um aumento relevante na acurácia da classificação ao comparar a classificação sem *feature selection* com os casos onde se aplicou essa técnica. Também é possível perceber que que os métodos de embrulho, quando associados à um algoritmo de classificação adequado, apresentam resultados mais acurados do que os métodos baseados em filtro. Em contrapartida, observa-se um aumento de 5200%, aproximadamente, no tempo de execução da classificação utilizando o algoritmo RFEC em relação à classificação sem nenhuma técnica de seleção de atributos, sem aumento proporcional de acurácia, sendo este de apenas 7.3% aproximadamente. Enquanto isso, no que diz respeito ao KNN, o algoritmo PCA apresenta tempo semelhante ao caso sem *feature selection*, porém com um aumento de aproximadamente 10% na acurácia.

4.2.2 *Heart Disease Dataset*

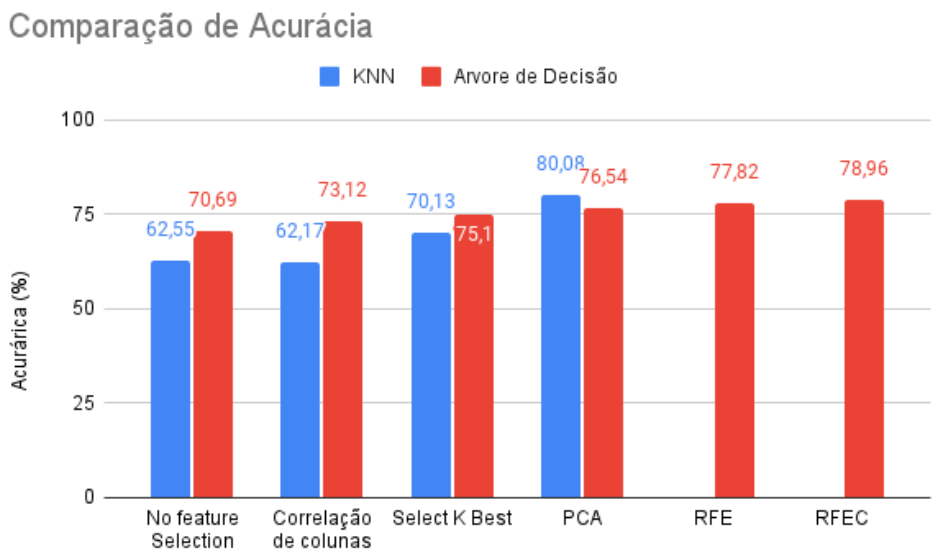
No escopo do conjunto de dados relativos a doenças cardíacas realizou-se os mesmos testes da seção 4.2.1. Na tabela 7 e figura 21 são observados os valores em porcentagem de acurácia obtidos em casa caso de testes, enquanto na tabela 8 e figura 22 são observados os tempos de execução da classificação em segundos.

Tabela 7 – Valores de acurácia obtido em cada caso (em porcentagem).

	KNN	Arvore de Decisão
No feature Selection	62,55	70,69
Correlação de colunas	62,17	73,12
Select K Best	70,13	75,10
PCA	80,08	76,54
RFE	-	77,82
RFEC	-	78,96

Fonte: Elaborado pelo autor.

Figura 23 – Comparação de Acurácia.



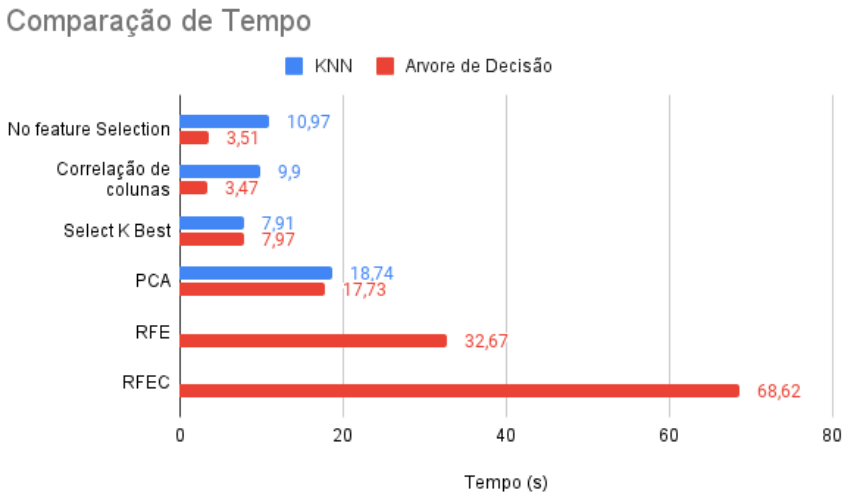
Fonte: Elaborado pelo autor.

Tabela 8 – Tempo de execução de cada caso (em segundos).

	KNN	Arvore de Decisão
No feature Selection	10,97	3,51
Correlação de colunas	9,90	3,47
Select K Best	7,91	7,97
PCA	18,74	17,73
RFE	-	32,67
RFEC	-	68,62

Fonte: Elaborado pelo autor.

Figura 24 – Comparação de Tempo.



Fonte: Elaborado pelo autor.

Assim como no caso anterior, é possível observar que a aplicação de métodos de seleção de atributos causou, de forma geral, um aumento na acurácia da classificação, sendo a única exceção a correlação de colunas para o algoritmo KNN. Tal diferença pode ser sido ocasionada por uma análise errônea de quais colunas, de fato, implicavam em melhoria ou não do desempenho da classificação. Também é possível observar que o método PCA não reproduziu resultados relevantes quando associado ao algoritmo baseado em árvore de decisão. Quanto ao tempo de execução observou-se que não somente os métodos de embrulho levaram à um aumento, como o método de filtro PCA apresentou um maior gasto de recursos.

As sutis diferenças entre os resultados obtidos com cada conjunto de dados são justificadas por fatores como diferente dimensionalidade, qualidade dos dados e até características intrínsecas do escopo de cada um.

4.3 Considerações finais

A realização dos experimentos confirmou a hipótese proposta neste trabalho, uma vez que os resultados obtidos explicitam a melhora da qualidade do processo de predição ao associar algoritmos de classificação com um processo de seleção de características. Ao analisar os valores mais detalhadamente fica evidente que, para os conjuntos de dados selecionados, é possível melhorar a qualidade da predição em até 10% para o primeiro conjunto de dados e até 18% para o segundo, melhorando, de forma significativa, o descobrimento de padrões e qualidade do

resultado obtido sem comprometimento do tempo de execução, inclusive com situações onde este foi reduzido e ainda sim obteve-se uma melhora relevante.

Vale ressaltar, também, que as características particulares de cada conjunto de dados e o escopo de sua aplicação devem ser levadas em consideração ao se definir qual método de *feature selection* será utilizado, já que existem situações nas quais um ganho de acurácia diminuto é mais benéfico quando comparado a um possível aumento do tempo de execução, enquanto em outros casos o aumento da acurácia é almejado independente do tempo de execução.

5 Conclusão

A evolução tecnológica recente permitiu a criação e difusão de uma grande quantidade de dados, que necessitam ser manipulados, armazenados, processados e estudados com o intuito de gerar conhecimento para as mais diversas áreas (ASRI et al., 2016). Para a área da saúde isso não foi diferente, e tais dados podem ser utilizados para auxiliar profissionais da saúde em uma série de questões, entre elas, a predição diagnóstica de quadros patológicos de pacientes, permitindo diagnósticos precoces e com maiores chances de recuperação aos pacientes, causando uma melhoria na qualidade de vida dos pacientes e um impacto positivo na sociedade (LE et al., 2018). Porém, as características desses dados, como a alta dimensionalidade, tendem a prejudicar a performance e a capacidade de classificação correta dos algoritmos de classificação difundidos atualmente (CIA et al., 2018).

Sendo assim, foi proposto por meio deste trabalho, contribuir com o estudo de associações de técnicas que possibilitam a redução da dimensionalidade, como métodos de seleção de atributos, e algoritmos de classificação com propriedades distintas. Para isso, essas diversas combinações foram realizadas e executadas em conjuntos de dados amplamente difundidos na literatura.

Por fim, foram realizados diversos testes com o objetivo de validar a proposta feita pelo trabalho. Para a validação foram realizadas mil execuções para cada um dos casos envolvendo uma combinação entre os seis métodos escolhidos de seleção de atributos e os dois algoritmos de classificação definidos. Por meio da análise da acurácia e tempo de execução foi possível obter conclusões satisfatórias sobre o impacto da seleção de atributos na redução da dimensionalidade e melhoria do processo preditivo. Portanto, é possível afirmar que a associação entre os métodos de classificação e o algoritmo de seleção de atributos PCA foi o que apresentou os melhores ganhos sem gerar um aumento demasiado no tempo de execução total do modelo.

5.1 Contribuições científicas

O estudo da dimensionalidade dos dados vem se tornando um tema recorrente na literatura, porém, existem muitas técnicas de redução de dimensionalidade, algoritmos de classificação e conjuntos de dados com características próprias a serem estudados. Portanto, a contribuição científica deste trabalho é a apresentação de uma análise das características dos

métodos de *feature selection* de filtro e embrulho quando associados a algoritmos de classificação baseado em instâncias e árvores de decisões, utilizando dois conjuntos de dados da área da saúde, mas com características distintas entre si.

Na Tabela 9 são apresentadas as comparações das características dos trabalhos correlatos apresentados durante a discussão teórica na sessão 2.5 e este trabalho. Evidenciando, assim, que somente o presente trabalho estuda e faz uma análise dos aspectos propostos.

Tabela 9 - Comparativo entre os trabalhos correlatos e o trabalho desenvolvido

	(LE et al. 2018)	(SABA et al., 2019)	(PEI et al., 2020)	Este Trabalho
Aplicação de algoritmos de *feature selection*	✓	✓	✗	✓
Utilização de múltiplos conjuntos de dados	✗	✗	✗	✓
Comparação entre algoritmos de classificação	✗	✓	✗	✓
Aplicação de algoritmos de árvore de decisão	✗	✓	✓	✓
Aplicação de algoritmos baseados em instâncias	✗	✗	✗	✓

Fonte: Elaborado pelo autor.

5.2 Trabalhos Futuros

Finalmente, após o desenvolvimento do trabalho, foi possível observar melhorias que podem vir a ser realizadas com a finalidade de contribuir com o estado da arte.

a) Expandir os algoritmos de classificação – realizar implementações de outros algoritmos de classificação e associá-los a técnicas de seleção de atributos como uma forma de conduzir um estudo sobre quais tipos de algoritmos possuem uma relação mais intrínseca com os dados da área de saúde

b) Aplicar outras técnicas de diminuição de dimensionalidade – aplicar outras técnicas de seleção de atributos, ou até mesmo técnicas de extração de atributos

c) Realizar as combinações de técnicas em outros bancos de dados – aplicar os testes propostos em cima de outros conjuntos de dados com características diferentes, como imagens ou em Processamento de Linguagem Natural, ou até mesmo em conjuntos de dados mais generalistas, uma possibilidade seria a utilização dos dados do hospital parceiro do Grupo de Banco de Dados.

Referências

ANDREU-PEREZ, Javier et al. Big data for health. **IEEE journal of biomedical and health informatics**, v. 19, n. 4, p. 1193-1208, 2015.

ASRI, Hiba et al. Big data in healthcare: Challenges and opportunities. In: **2015 International Conference on Cloud Technologies and Applications (CloudTech)**. IEEE, 2015. p. 1-7.

BASHIR, Saba et al. Improving heart disease prediction using feature selection approaches. In: **2019 16th international bhurban conference on applied sciences and technology (IBCAST)**. IEEE, 2019. p. 619-623.

BRACHMAN, Ronald J.; ANAND, Tej. **The process of knowledge discovery in a first sketch**. AAAI Tech. Rep. WS-94-03, 1994.

CAI, Jie et al. Feature selection in machine learning: A new perspective. **Neurocomputing**, v. 300, p. 70-79, 2018.

CHEN, Xue-wen; JEONG, Jong Cheol. Enhanced recursive feature elimination. In: **Sixth International Conference on Machine Learning and Applications (ICMLA 2007)**. IEEE, 2007. p. 429-435.

DEMCHENKO, Yuri; DE LAAT, Cees; MEMBREY, Peter. Defining architecture components of the Big Data Ecosystem. In: **2014 International conference on collaboration technologies and systems (CTS)**. IEEE, 2014. p. 104-112.

EL ABOUDI, Naoual; BENHLIMA, Laila. Review on wrapper feature selection approaches. In: **2016 International Conference on Engineering & MIS (ICEMIS)**. IEEE, 2016. p. 1-5.

FAN, Jianqing; LI, Runze. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. **arXiv preprint math/0602133**, 2006.

FAYYAD, Usama M.; HAUSSLER, David; STOLORZ, Paul E. KDD for Science Data Analysis: Issues and Examples. In: **KDD**. 1996. p. 50-56.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 39, n. 11, p. 27-34, 1996.

FU, Yongjian. Data mining. **IEEE potentials**, v. 16, n. 4, p. 18-20, 1997.

GLASGOW, Justin M.; KABOLI, Peter J. Detecting adverse drug events through data mining. **American journal of health-system pharmacy**, v. 67, n. 4, p. 317-320, 2010.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel; BEZERRA, Eduardo. **Data Mining**. Elsevier Brasil, 2015.

HALL, Mark A. Correlation-based feature selection of discrete and numeric class machine learning. 2000.

HALL, Mark Andrew. Correlation-based feature selection for machine learning. 1999.

ISHAQ, Abid et al. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. **IEEE Access**, v. 9, p. 39707-39716, 2021.

JANOSI, Andras. et al. **UCI Machine Learning Repository,** 1988. Heart Disease Data Set. Disponível em: https://archive.ics.uci.edu/. Acesso em: 17 de julho de 2021

JOVIĆ, Alan; BRKIĆ, Karla; BOGUNOVIĆ, Nikola. A review of feature selection methods with applications. In: **2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)**. Ieee, 2015. p. 1200-1205.

KAUR, Gaganjot; CHHABRA, Amit. Improved J48 classification algorithm for the prediction of diabetes. **International journal of computer applications**, v. 98, n. 22, 2014.

KAUR, Prableen; SHARMA, Manik; MITTAL, Mamta. Big data and machine learning based secure healthcare framework. **Procedia computer science**, v. 132, p. 1049-1059, 2018.

LE, Hung Minh; TRAN, Toan Dinh; VAN TRAN, L. A. N. G. Automatic heart disease prediction using feature selection and data mining technique. **Journal of Computer Science and Cybernetics**, v. 34, n. 1, p. 33-48, 2018.

MIAO, Jianyu; NIU, Lingfeng. A survey on feature selection. **Procedia Computer Science**, v. 91, p. 919-926, 2016.

MORSE, Stephen S. et al. Prediction and prevention of the next pandemic zoonosis. **The Lancet**, v. 380, n. 9857, p. 1956-1965, 2012.

MURDOCH, Travis B.; DETSKY, Allan S. The inevitable application of big data to health care. **Jama**, v. 309, n. 13, p. 1351-1352, 2013.

NEW, J. P.; LEATHER, D.; BAKERLY, N. D.; MCCRAE, J.; GIBSON, J. M. Putting patients in control of data from electronic health records. **BMJ**, v. 360, p. j5554, 2018.

NG, Kenney et al. Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. **Circulation: Cardiovascular Quality and Outcomes**, v. 9, n. 6, p. 649-658, 2016.

NGIAM, Kee Yuan; KHOR, Wei. Big data and machine learning algorithms for health-care delivery. **The Lancet Oncology**, v. 20, n. 5, p. e262-e273, 2019.

ORTEGA, John Heland Jasper C. et al. An analysis of classification of breast cancer dataset using J48 algorithm. **International Journalof Advanced Trends in Computer Science and Engineering**, v. 9, n. 3, 2020.

PARIMALA, C.; PORKODI, R. Classification algorithms in data mining: a survey. **Proceedings of the International Journal of Scientific Research in Computer Science**, v. 3, p. 349-355, 2018.

PEI, Dongmei; YANG, Tengfei; ZHANG, Chengpu. Estimation of Diabetes in a High-Risk Adult Chinese Population Using J48 Decision Tree Model. **Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy**, v. 13, p. 4621, 2020.

RAHM, Erhard; DO, Hong Hai. Data cleaning: Problems and current approaches. **IEEE Data Eng. Bull.**, v. 23, n. 4, p. 3-13, 2000.

REMESEIRO, Beatriz; BOLON-CANEDO, Veronica. A review of feature selection methods in medical applications. **Computers in biology and medicine**, v. 112, p. 103375, 2019.

SANTOS, Hellen Geremias dos et al. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. **Cadernos de Saúde Pública**, v. 35, 2019.

SARITAS, Mucahid Mustafa; YASAR, Ali. Performance analysis of ANN and Naive Bayes classification algorithm for data classification. **International Journal of Intelligent Systems and Applications in Engineering**, v. 7, n. 2, p. 88-91, 2019.

SENGUPTA, Partho P.; SHRESTHA, Sirish. Machine learning for data-driven discovery: the rise and relevance. 2019.

SOLANKI, Yogendra Singh et al. A Hybrid Supervised Machine Learning Classifier System for Breast Cancer Prognosis Using Feature Selection and Data Imbalance Handling Approaches. **Electronics**, v. 10, n. 6, p. 699, 2021.

WANG, Peipei et al. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. **Chaos, Solitons & Fractals**, v. 139, p. 110058, 2020.

WOLBERG, William H. et al. **UCI Machine Learning Repository,** 1995. Breast Cancer Wisconsin (Diagnostic) Data Set. Disponível em: https://archive.ics.uci.edu/. Acesso em: 17 de julho de 2021

XING, Wenchao; BEI, Yilin. Medical health big data classification based on KNN classification algorithm. **IEEE Access**, v. 8, p. 28808-28819, 2019.

YU, Lei; LIU, Huan. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: **Proceedings of the 20th international conference on machine learning (ICML-03)**. 2003. p. 856-863.