MINISTÉRIO DA EDUCAÇÃO
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
GOIANO - CAMPUS RIO VERDE

ISAAC DA SILVA CARDOSO

TÉCNICAS DE OTIMIZAÇÃO E MÉTRICAS DE AVALIAÇÃO APLICADAS A MACHINE LEARNING

RIO VERDE
2022

ISAAC DA SILVA CARDOSO

TÉCNICAS DE OTIMIZAÇÃO E MÉTRICAS DE AVALIAÇÃO APLICADAS A MACHINE LEARNING

Trabalho de Conclusão de Curso apresentado ao Instituto Federal de Educação, Ciência e Tecnologia Goiano - Campus Rio Verde ligado ao Ministério da Educação, como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Área de concentração: Ciência da Computação

Linha de pesquisa: Pré-processamento de dados

Orientador: Heyde Francielle do Carmo Franca
Instituto Federal de Educação, Ciência e Tecnologia Goiano - Campus Rio Verde

RIO VERDE
2022

Sistema desenvolvido pelo ICMC/USP
Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas - Instituto Federal Goiano

Cardoso, Isaac da Silva
C268t Técnicas de Otimização e Métricas de Avaliação Aplicadas a Machine Learning / Isaac da Silva Cardoso; orientadora Heyde Francielle do Carmo França. -- Rio Verde, 2022.
37 p.

TCC (Graduação em Ciência da Computação) -- Instituto Federal Goiano, Campus Rio Verde, 2022.

1. Aprendizado de Máquina. 2. Pré-processamento de dados. 3. Algoritmos. 4. Métricas de Avaliação. I. França, Heyde Francielle do Carmo , orient. II. Título.

Responsável: Johnathan Pereira Alves Diniz - Bibliotecário-Documentalista CRB-1 n°2376

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO
PARA DISPONIBILIZAR PRODUÇÕES TÉCNICO-CIENTÍFICAS NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Com base no disposto na Lei Federal nº 9.610, de 19 de fevereiro de 1998, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano a disponibilizar gratuitamente o documento em formato digital no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

IDENTIFICAÇÃO DA PRODUÇÃO TÉCNICO-CIENTÍFICA

- ☐ Tese (doutorado)
- ☐ Dissertação (mestrado)
- ☐ Monografia (especialização)
- ☑ TCC (graduação)
- ☐ Artigo científico
- ☐ Capítulo de livro
- ☐ Livro
- ☐ Trabalho apresentado em evento
- ☐ Produto técnico e educacional - Tipo:

Nome completo do autor:
Isaac da Silva Cardoso

Matrícula:
2017102201910420

Título do trabalho:
Técnicas de Otimização e Métricas de Avaliação Aplicadas A Machine Learning

RESTRIÇÕES DE ACESSO AO DOCUMENTO

Documento confidencial: ☑ Não ☐ Sim, justifique:

Informe a data que poderá ser disponibilizado no RIIF Goiano: 12 /08 /2022

O documento está sujeito a registro de patente? ☐ Sim ☑ Não

O documento pode vir a ser publicado como livro? ☐ Sim ☑ Não

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O(a) referido(a) autor(a) declara:

- Que o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;

- Que obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autoria, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;

- Que cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Rio Verde — Local
30 /08 /2022 — Data

Isaac da Silva Cardoso

Assinatura do autor e/ou detentor dos direitos autorais

Ciente e de acordo:

Assinatura do(a) orientador(a)

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO

Ata nº 30/2022 - GEPTNM-RV/DE-RV/CMPRV/IFGOIANO

ATA DE DEFESA DE TRABALHO DE CURSO

Ao(s) 12 do mês de agosto de 2022, às 9 horas, reuniu-se a banca examinadora composta pelos docentes:

Heyde Francielle do Carmo França, Douglas Cedrim Oliveira, André da Cunha Ribeiro, para examinar o Trabalho de Curso intitulado “Técnicas de otimização e métricas de avaliação aplicadas a Machine Learning” do(a) estudante Isaac da Silva Cardoso, Matrícula nº 2017102201910420 do Curso de Ciência da Computação do IF Goiano – Campus Rio Verde. A palavra foi concedida ao(a) estudante para a apresentação oral do TC, houve arguição do(a) candidato pelos membros da banca examinadora. Após tal etapa, a banca examinadora decidiu pela APROVAÇÃO do(a) estudante. Ao final da sessão pública de defesa foi lavrada a presente ata que segue assinada pelos membros da Banca Examinadora.

(Assinado Eletronicamente)

Heyde Francielle do Carmo França

Orientador(a)

(Assinado Eletronicamente)

Douglas Cedrim Oliveira

Membro

(Assinado Eletronicamente)

André da Cunha Ribeiro

Membro

ISAAC DA SILVA CARDOSO

TÉCNICAS DE OTIMIZAÇÃO E MÉTRICAS DE AVALIAÇÃO APLICADAS A MACHINE LEARNING

Trabalho de curso DEFENDIDO E APROVADO em 12 de agosto de 2022, pela Banca Examinadora constituída pelos membros:

Dr.André da Cunha Ribeiro
Instituto Federal Goiano

Dr. Douglas Cedrim Oliveira
Instituto Federal Goiano

Heyde Francielle do Carmo França
Orientadora

Rio Verde, GO

2022

"Dedico-me à cor rubra e escarlate como o meu sangue de homem em plena idade e portanto dedico-me a meu sangue. Dedico-me sobretudo aos gnomos, anões, sílfides e ninfas que me habitam a vida."(LISPECTOR, Clarice, 1977)

AGRADECIMENTOS

Agradeço ao apoio, paciência, compreensão e incentivo da minha orientadora Heyde. Sem você esse trabalho não existiria. A todos os meus professores e professoras, especialmente Aline, Cristiane e André. Muito obrigado por tudo o que vocês me ensinaram. À minha mãe, Alec Sânia, por sempre me apoiar e me impulsionar a seguir. Ao meu pai, Carlos e à minha irmã Thaís, por estarem comigo o tempo todo. Amor é quando não se dá nome à identidade das coisas? À Clarice Lispector, que nunca soube ou poderia saber quem sou, mas me ensinou e ensina a viver. Aos meus familiares, por acreditarem em mim. Aos meus colegas, por compartilharem essa jornada ao meu lado. A todos os meus amigos, especialmente Micaele, Lunna, Cauã, Júlia, Marcos e Guilherme, por enfrentarem os momentos mais difíceis ao meu lado e me ajudarem a persistir. As pessoas têm tanto medo das verdades umas das outras, mas vocês nunca tiveram das minhas. Ao meu namorado, Manoel, por me lembrar que eu era capaz. A solução para esse absurdo que se chama "eu existo", a solução é amar um outro ser que, este, nós compreendemos que exista.

Porque há o direito ao grito. Então eu grito.
(LISPECTOR, Clarice, 1977).

RESUMO

CARDOSO, Isaac. Técnicas de Otimização e Métricas de Avaliação Aplicadas a Machine Learning. 2022. 37 f. Trabalho de Conclusão de Curso – Bacharelado em Ciência da Computação, Instituto Federal de Educação, Ciência e Tecnologia Goiano - Campus Rio Verde. Rio Verde, 2022.

Para o aprendizado de máquina, a qualidade dos dados, bem como seu pré-processamento, uma vez que os algoritmos utilizados derivam conhecimento principalmente desses mesmos dados, é decisiva para a qualidade dos resultados. A presença ou ausência de valores desconhecidos e a seleção de dados chave, de forma a não contaminar a base de dados utilizada com informações sem relação com o objetivo, por exemplo, são fatores decisivos em se tratando do mérito citado. Em muitas aplicações é importante pensar em como proceder com as informações disponíveis caso estejam incompletas pois o manuseio de valores desconhecidos deve ser cuidadosamente planejado ou surge o risco de distorção. Este trabalho apresenta uma visão panorâmica do tema bem como a implementação de técnicas de pré-processamento aplicados a uma base que contém dados de Covid 19, demonstrando os processos seguidos e os resultados adquiridos, e mostrando como o pré-processamento de dados pode influenciar no resultado final da implementação de um algoritmo. Aprender a trabalhar com dados desbalanceados é primordial, pois nos diversos campos em que estes se encontram, dificilmente se encontrará uma base perfeitamente balanceada, portanto, entender como realizar um bom pré-processamento é uma solução possível para melhorar os resultados de projetos com aprendizado de máquina.

Palavras-chave: Aprendizado de Máquina. Pré-processamento de dados. Algoritmos. Métricas de Avaliação.

ABSTRACT

CARDOSO, Isaac. Optimization Techniques and Evaluation Metrics Applied to machine learning. 2022. 37 f. Trabalho de Conclusão de Curso – Bacharelado em Ciência da Computação, Instituto Federal de Educação, Ciência e Tecnologia Goiano - Campus Rio Verde. Rio Verde, 2022.

For machine learning, the quality of the data, as well as its pre-processing, since the algorithms used mainly derive knowledge from these same data, is decisive for the quality of the results. The presence or absence of unknown values and the selection of key data, so as not to contaminate the database used with information unrelated to the objective, for example, are decisive factors when it comes to the cited merit. In many applications it is important to think about how to proceed with available information if it is incomplete as the handling of unknown values must be carefully planned or the risk of misstatement arises. This work presents a panoramic view of the subject as well as the implementation of pre-processing techniques applied to a database that contains Covid 19 data, demonstrating the processes followed and the results acquired, and showing how the pre-processing of data can influence the final result of implementing an algorithm. Learning to work with unbalanced data is essential, because in the various fields in which they are found, it is difficult to find a perfectly balanced base, therefore, understanding how to perform a good pre-processing is a possible solution to improve the results of projects with learning from machine.

Keywords: Machine Learning. Data pre-processing. Algorithms.Evaluation Metrics

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
IA	Inteligência Artificial
KNN	K Nearest Neighbours
RF	Random Forest
S	Sensibilidade
R	Recall
P	Precisão
A	Acurácia
FP	Falso Positivo
FN	Falso Negativo
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo
ROC	Receiver Operating Characteristic

SUMÁRIO

1 INTRODUÇÃO

O aprendizado de máquina é uma técnica que permite a construção de programas que melhorem o desempenho através do exemplo. As técnicas de AM (Machine Learning, ou Aprendizado de Máquina), são orientadas por dados, ou seja, aprendem automaticamente com grandes quantidades destes. Os dados gerados por um algoritmo de AM são processados e geram resultados, mas antes da execução propriamente dita, existe uma fase essencial a se passar por: o pré-processamento de dados (Mitchell apud Ludermir (2021)).

Atualmente possuímos à nossa disposição inúmeros bancos dados públicos de diversas areas. Todavia, em sua grande maioria, estes estão não estruturados. A partir da estruturação desses dados, é possível inferir informações e entender comportamentos, podendo-se assim, inclusive, prever as próximas ações. Reconhecer padrões é a habilidade de identificar e analisar semelhanças entre determinados objetos, processando aquele padrão e seccionando-o em diferentes classes (SIMON, 2001). Diferentemente dos seres humanos, as máquinas não conseguem gerar por conta própria informações adicionais associadas a experiências de vida que permitam a compreensão de um conjunto de dados, portanto, o arranjo bem organizado destes facilita e acelera os resultados de algoritmos de AM e melhoram seu desempenho.

As técnicas de Aprendizado de Máquina podem ser aplicadas a diferentes áreas por diferentes motivos, inclusive na área da saúde, como no caso da pandemia de Covid-19 que teve início no ano de 2020, doença causada pelo coronavírus (SARS-CoV-2) e identificada pela primeira vez na China, em dezembro de 2019 (OLIVEIRA et al.,). O interesse em se trabalhar especialmente com esse tipo de dados foi justamente pelo impacto que esta doença teve em todo o mundo durante o século XXI, deixando um rastro de morte e mudanças impactantes na estrutura de vida das pessoas.

As aplicações das técnicas aqui demonstradas sob uma base de Covid 19, permitem perceber que algoritmos deste tipo (AM) possibilitam a obtenção de diagnósticos mais rápidos e precisos, onde o software desenvolvido é capaz de detectar padrões para cada doença presente no dataset que se está analisando. Claro que os bancos de dados são longos e extensos (podendo atingir terabytes em quantidade de informações), o que enfatiza ainda mais a importância das técnicas de pré-processamento. Temos então que, através da própria inteligência artificial pode-se identificar uma doença por meio da análise de dados, o que apoia os processos médicos (SOUZA et al., 2020).

Como dito anteriormente, o desempenho dos algoritmos de AM depende diretamente da qualidade das bases de dados utilizadas, ou seja, a boa formatação e filtragem das informações que serão utilizadas pelos diferentes processos impacta no quão bem os resultados da implementação de um algoritmo serão. As técnicas de pré-processamento são indispensáveis para essa etapa, pois permitem identificar e corrigir erros ou imperfeições

no dataset utilizado, deixando-o o mais bem estruturado possível. Para tanto, estudar, compreender e aplicar as diferentes técnicas de pré-processamento e aferi-las com diferentes métricas de avaliação torna-se um processo orgânico na aplicação dos conceitos acima citados.

A fase de pré-processamento começa assim que os dados são coletados e classificados, de modo a permitir a identificação de dados corrompidos, atributos não relacionados e valores indefinidos, por exemplo. É possível também se estar interessado na análise visual dos dados através de gráficos e afins, de modo a estruturar conclusões e tomar decisões no projeto. Pode-se ainda querer modificar a estrutura dos dados, através, por exemplo, da alteração do grau de granularidade destes. As atividades concluídas na fase de pré-processamento objetivam preparar os dados para que a próxima fase (extração de conhecimento) seja mais efetiva (BATISTA et al., 2003).

Em geral, o pré-processamento de dados é um processo semiautomático, ou seja, depende da capacidade de quem lidera a atividade em identificar os problemas presentes nos dados, além de sua natureza, e de utilizar métodos adequados para a solução destes (BATISTA et al., 2003). Assim, a aplicação de técnicas de pré-processamento de dados pode melhorar a qualidade dos algoritmos de Aprendizado de Máquina, permitindo assim a obtenção de resultados melhores.

A presente pesquisa buscou mostrar o quanto o pré-processamento de dados pode melhorar significativamente a qualidade nos resultados dos algoritmos de Aprendizado de Máquina, entendendo seu impacto. Comparando algumas das diferentes técnicas de pre-processamento aplicadas a uma base de Covid-19, mostrando os resultados obtidos através desses processos utilizando o algoritmo Random Forest como classificador, explicando e exemplificando as diferentes etapas do pré-processamento, explicitando os resultados obtidos nas diferentes métricas de avaliações apresentadas.

O restante deste trabalho está organizado em 6 seções, sendo a segunda uma contextualização teórica do tema, a terceira uma apresentação de alguns trabalhos relacionados,a quarta mostra as ferramentas e métodos empregados, assim como as atividades realizadas durante a execução do projeto, a quinta a apresentação dos resultados obtidos e a sexta as conclusões obtidas a partir desses resultados.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta a base teórica utilizada para a fundamentação do presente trabalho. Está dividida em 4 subseções que discutem conceitos relacionados ao Aprendizado de Máquina, ao pré-processamento de dados e às diferentes métricas de avaliação passíveis de serem utilizadas.

2.1 Inteligência Artificial e Aprendizado de Máquina

Inteligência Artificial (IA) é um termo utilizado em diversos estudos científicos que expressa a associação de computação com inteligência a partir da apresentação do teste de Turing (TURING, 1950). O termo foi desenvolvido em 1956 por John MacCarthy em uma conferência realizada na Universidade de Dartmouth e indicava, para ele, uma ciência e engenharia capazes de criar máquinas e programas de computadores inteligentes.

Somers (2013) alcunham duas ideias fundamentais para a IA, sendo elas a capacidade de aprendizagem e a manifestação do chamado "comportamento inteligente", e descrevem categorias de definições: sistemas que se portam como humanos, que pensam como humanos, que pensam racionalmente e que agem racionalmente.

As definições ao redor da Inteligência Artificial são inúmeras, contudo é consenso que algoritmos de IA necessitam da compreensão de processos humanos de aprendizagem e suas modelagens, o que permite a emulação desses processos em uma máquina computacional.

Pode-se definir, portanto Inteligência Artificial como a habilidade de programas de computadores operarem de modo a se acreditar na imitação dos processos humanos (COLLINS et al., 2021).

Aprendizado de Máquina é uma subárea da Inteligência Artificial que emprega uma grande variedade de técnicas probabilísticas, estatísticas e de otimização que possibilita, como sugerido por seu nome, aos computadores aprenderem e detectarem padrões a partir da análise de dados previamente adquiridos e a geração de novos dados a partir destes, usando as características e habilidades das máquinas de modo a complementar a inteligência humana (SHALEV-SHWARTZ; BEN-DAVID, 2014). Abaixo, na figura 1, temos representado o esquema básico de Aprendizado de Máquina, que mostra o fluxo que se inicia com a entrada de um conjunto de treinamento, seguido pela implementação de algum algoritmo de aprendizado e uma saída dada por um classificador.

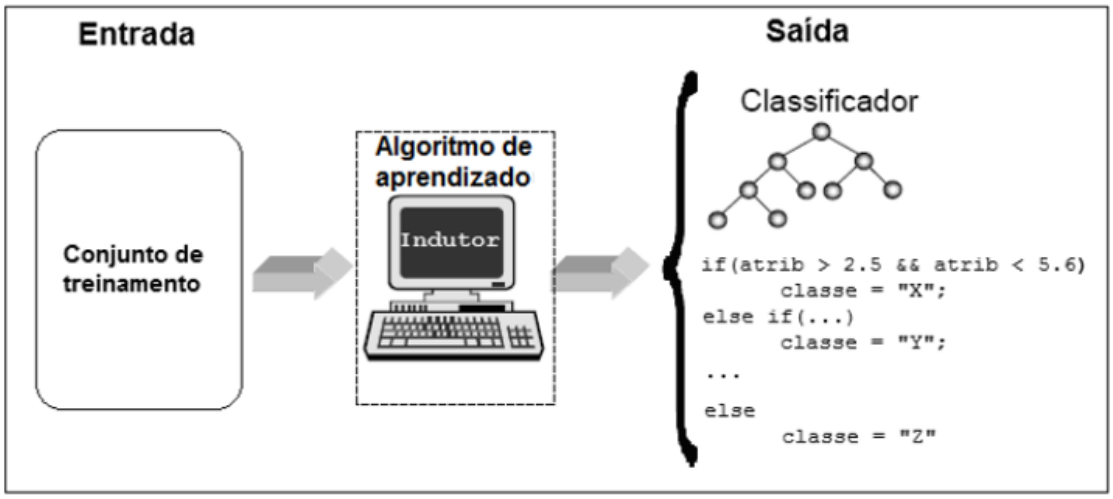

Figura 1 – Esquema básico de AM.Fonte: (MONARD; BARANAUSKAS, 2000)

Os métodos aplicados com AM procuram derivar modelos preditivos utilizando dados atuais e históricos, o que significa que o algoritmo consegue aumentar sua acurácia e precisão à medida que mais iterações aconteçam. Olhando-se um panorama, as ferramentas de AM podem ser classificadas em duas categorias principais (FRIEDMAN; HASTIE; TIBSHIRANI, 2008): aprendizado supervisionado (que utiliza dados classificados, ou seja, rotulados) e não supervisionado (que utiliza dados não classificados, ou seja, não rotulados) (KUHN; JOHNSON et al., 2013).

A análise preditiva é uma técnica que consiste na implementação de algoritmos de modo a entender a como funcionam os dados existentes e gerar regras para a predição. Os algoritmos desse tipo são geralmente utilizados em cenários não supervisionados, onde apenas os preditores estão disponíveis no dataset, ou em cenários com problemas supervisionados, onde, além dos preditores, há uma resposta de interesse, responsável por guiar toda a análise (SANTOS et al., 2019a).

A análise preditiva, portanto, objetiva realizar a estimativa do risco de eventos futuros reincidirem baseada em experiências já vividas, de modo a auxiliar na tomada de decisão atual através da implementação de uma série de algoritmos de Aprendizado de Máquina usados para a compreensão de dados (KUHN; JOHNSON et al., 2013).

Algoritmos de Aprendizado de Máquina trabalham a partir de exemplos que são fornecidos como entrada de modo a representar experiências que permitem adquirir conhecimento sobre um determinado assunto, criando, assim, uma referência. Pensando nos modos a partir dos quais um algoritmo aprende, podemos dividir em duas categorias, classificando-os entre aprendizado supervisionado e não-supervisionado.

No primeiro caso, é dado ao sistema um conjunto de exemplos com uma saída já conhecida, enquanto que no segundo caso os algoritmos inferem que não se conhece

à classe dos exemplos, procurando similaridade entre os atributos fornecidos (SOMERS, 2013). O aprendizado supervisionado geralmente é usado para previsão de eventos. Na aprendizagem supervisionada, o objetivo é inferir uma função ou mapeamento a partir dos dados de treinamento. No processamento dos algoritmos de Aprendizado de Máquina, o conjunto de dados é dividido em teste e treinamento, de modo a treinar o algoritmo com parte dos dados e testar os níveis da aprendizagem com o outro, conforme demonstrado na figura 2 abaixo.

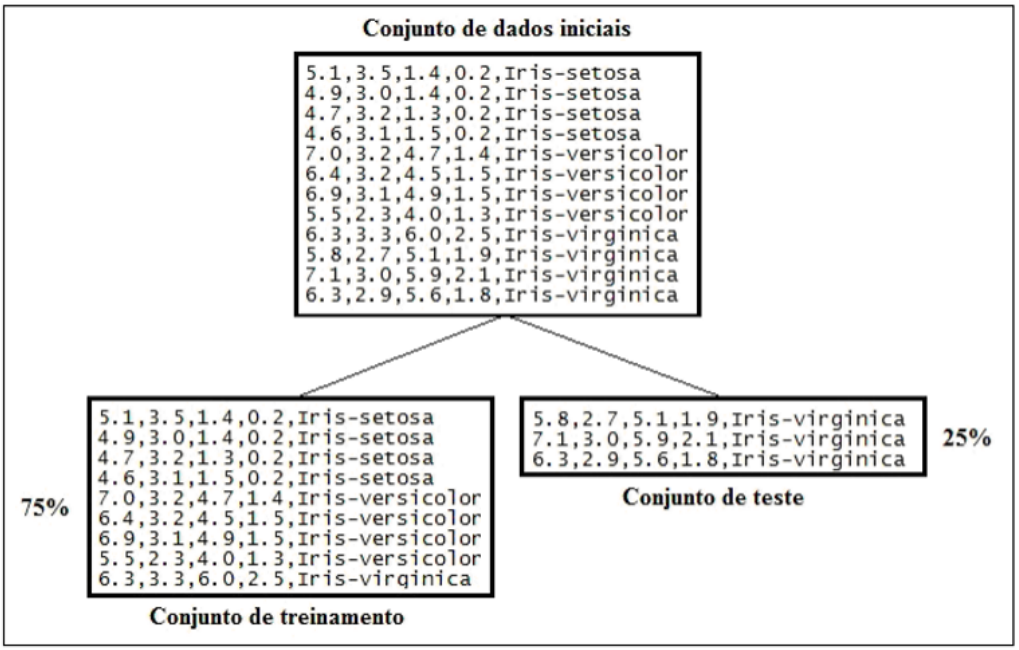

Figura 2 – Conjunto de dados iniciais dividido em conjunto de treinamento e conjunto de teste.Fonte: (REAL; NICOLETTI, 2014a)

2.2 Mineração de Dados

De acordo com (FERNANDEZ, 2003), a mineração de dados automatiza o processo de descoberta de relacionamentos e padrões nos dados e apresenta resultados que podem ser usados em sistemas automatizados de suporte à decisão ou acessados por tomadores de decisão. As técnicas de mineração de dados são derivadas de inteligência artificial (IA) e estatísticas e ajudam a descobrir regras ou padrões, prever tendências e comportamentos futuros ou entender grupos semelhantes(SCHMITT et al., 2005).

Podemos concluir então que a mineração de dados é uma técnica que auxilia imensamente os processos que envolvem AM e consiste em decidir quais algoritmos devem ser aplicados aos dados. Nesta etapa, algoritmos de diferentes áreas do conhecimento podem ser utilizados, como aprendizado de máquina, estatística, redes neurais e bancos de

dados. Se o objetivo desta etapa é criar um modelo preditivo, então decidir qual algoritmo é o ideal para o problema que está sendo analisado não é uma tarefa trivial (REAL; NICOLETTI, 2014b).

Alguns dos algoritmos mais relevantes nessa área são descritos a seguir. O SVM (CORTES; VAPNIK, 1995) trata-se de um algoritmo de aprendizado supervisionado que objetiva classificar e analisar regressivamente os dados, recebendo um dataset como entrada e predizendo, para cada uma das entradas, qual de duas possibilidades, geralmente classes, a entrada faz parte, ou seja, trata-se de um classificador linear binário não probabilístico. Esse algoritmo molda um modelo que atribui novos exemplos a cada uma das categorias. É dito que um modelo SVM representa exemplos como pontos no espaço, mapeando-os de modo que os exemplos de cada uma das categrorias sejam divididos no espaço.

O KNN (K Nearest Neighbours) parte do pressuposto que dados semelhantes estão próximos um dos outros e depende que essa suposição seja real para a utilidade do algoritmo, utilizando-se dessa ideia de proximidade através do cálculo da distância entre os pontos de um gráfico. O algoritmo é executado diversas vezes com diferentes valores atribuídos para K e, eventualmente, escolhe o K que possibilite a redução do número de erros (SILVA; MARREIROS, 2021).

O Random Forest (MISRA; LI; HE, 2019) é um método que treina diferentes árvores de decisão em paralelo, seguido de um processo de agregação, junto ao bootstraping, que indica que inúmeras árvores de decisão individuais são treinadas paralelamente e diferentes subconjuntos do conjunto de dados de treino, o que garante que cada árvore seja única. Eventualmente uma floresta aleatória será formada e um grande número de classificadores formará um classificador forte (MISRA; LI; HE, 2019). Portanto, é um classificador largamente implementado uma vez que supera, na maioria dos casos, outros métodos de classificação quando se trata de precisão (MISRA; LI; HE, 2019), o que motivou a implementação deste no presente trabalho.

2.2.1 Pré-processamento

O pré-processamento é uma etapa de suma importância na análise e seleção dos dados de entrada da rede neural. São vários os problemas que podem ser descobertos nessa fase, como por exemplo: existência de um viés e diferença de escalas nos valores expressados, valores faltantes, anotações inconsistentes, corrompidas ou redundantes e dados desbalanceados. Para solucionar o problema de enviesamento é comum que seja utilizado técnicas de normalização dos dados de entrada. Para o problema de dados faltantes, podem ser usada média, valores mais comuns para aquele atributo, ou até mesmo deletar aquele registro caso nenhuma solução seja plausível. No caso de dados desbalanceamento existem algumas técnicas como oversample e downsample que podem ser utilizadas para minimizar os efeitos no treinamento da rede.

Dados brutos, no geral, não apresentam um bom resultado na performance de

um algoritmo, portanto é preciso levar em consideração características que se relacionam à dimensão de todo o conjunto de dados disponível (RASCHKA; MIRJALILI, 2017). Há inúmeras técnicas de pré-processamento e, de modo geral, estas estão relacionadas à transformação dos dados originais e curadoria desses dados em relação à sua relevância para a pesquisa. Abaixo na figura 3 temos uma representação básica das etapas de pré-processamento de dados.

Figura 3 – Representação de um esquema básico de AM. Fonte: (GONZALEZ, 2021)

Uma boa transformação dos dados impacta diretamente na melhora do desempenho de vários algoritmos. A normalização (redimensionamento das variáveis para uma escala que varie num intervalo delimitado, geralmente 0 e 1) e a padronização (centralização das variáveis na média 0 com desvio padrão 1) são exemplos de abordagens mais usadas na modificação da escala de uma dada variável (RASCHKA; MIRJALILI, 2017)(KUHN; JOHNSON et al., 2013).

A etapa do pré-processamento de dados tem início assim que os dados são coletados e estruturados em um dataset. Há inúmeros motivos pelos quais é importante se utilizar de uma fase de pré-processamento, como identificar e tratar dados corrompidos e valores desconhecidos, aprender mais a respeito dos dados, alterar sua estrutura e afins.

Segundo Han, Pei e Kamber (2011), há três elementos que definem a qualidade dos dados: precisão, completude e consistência. Todavia, comumente um dataset não apresenta dados perfeitos, muitas vezes contendo problemas relacionados a erros humanos ou falhas na obtenção dos dados.

É de conhecimento geral que o pré-processamento depende da identificação e tratamento de inúmeros problemas que podem aparecer, configurando uma tarefa extensa e trabalhosa. Portanto, torna-se indispensável um bom ambiente computacional para processar tais dados.

A maioria das vezes, trabalhamos com dados coletados para um outro fim que não a pesquisa em andamento, o que faz com que inúmeras vezes eles estejam armazenados em formatos divergentes do necessário ou apresentar inconsistências (TAN; STEINBACH; KUMAR, 2009), como ruído (erros randômicos que talvez impliquem na distorção de um valor ou na adição de objetos falsos), outliers (objetos com características diferentes dos outros objetos do mesmo dataset) e valores ausentes (geralmente decorrentes de erros humanos ou de campos cujo preenchimento não foi obrigatório) (WITTEN et al., 2011).

O objetivo principal do pré-processamento é, portanto, extrair de dados não estruturados uma representação estruturada e manipulável por algoritmos que consigam identificar dados mais relevantes e padronizar as informações, de modo a reduzir a dimensionalidade dos dados através, por exemplo, da remoção de dados irrelevantes (REZENDE; MARCACINI; MOURA, 2011)

No caso do presente trabalho, o pré-processamento de dados de Aprendizado de Máquina objetivou modificar dados de entrada de modo a configurá-los em um formato mais aplicável para análises posteriores, o que envolve desde a limpeza destes para possíveis remoção de ruídos, fusão de dados de diferentes fontes, observação e correção de dados faltantes e duplicatas e categorização dos dados mais importantes ao problema (TAN; STEINBACH; KUMAR, 2009).

As técnicas de pré-processamento podem ser, basicamente, divididas em Han, Pei e Kamber (2011) limpeza dos dados, integração dos dados, transformação dos dados e redução dos dados. Importante salientar que, geralmente, as bases iniciam-se com dados

faltantes, especialmente se o dataset não tiver sido montado com o objetivo de se trabalhar com algum algoritmo.

Há inúmeros métodos para lidar com dados faltantes, como a eliminação de linhas (simples porém não tão indicada, uma vez que pode diminuir o desempenho caso haja muitas linhas desse tipo, sendo mais utilizada quando a linha contiver inúmeros valores ausentes) (HAN; PEI; KAMBER, 2011), ignorar valores ausentes durante a análise (TAN; STEINBACH; KUMAR, 2009), remover valores ausentes, que podem ser substituídos de acordo com o contexto, utilizando-se, por exemplo, a média para valores numéricos e a moda para valores categóricos (LAROSE, 2005).

As formas de limpeza de dados constituem uma investigação inicial focada na detecção de registros incompletos ou dados faltantes, incorretos ou duplicados. Para resolver esses problemas, é possível ignorar os registros, completá-los manualmente, substituir por um valor global obtido através da média, moda ou mediana dos valores já apresentados ou utilização do valor mais provável, obtido através de uma regressão ou árvore de decisão.

Quanto à redução da dimensionalidade dos dados, podemos pensar especialmente em bases de dados muito grandes, onde sua redução pode acarretar na diminuição do tempo computacional necessário para processar os dados. Pode ser importante também no caso da obtenção do melhor aproveitamento dos dados, o que facilita a extração de conhecimento. É necessário, contudo, tomar cuidado nessa etapa, pois uma limpeza não cautelosa pode acarretar na perda de informações possivelmente relevantes.

Os conjuntos de dados também podem ter muitos atributos, então o uso da tecnologia de redução de dimensionalidade pode eliminar recursos desnecessários e reduzir ruídos, além de tornar a forma mais clara e facilitar a visualização dos dados (TAN; STEINBACH; KUMAR, 2009). Ao se trabalhar com o pré-processamento de dados é importante salientar que a base final é, em geral, um arquivo completamente diferente do original (CARVALHO et al., 2003).

A seleção de atributos é feita a partir da organização agrupada de uma grande quantidade dados. Os dados que serão utilizados estão armazenados em bases e nem sempre estão estruturados da melhor maneira possível. Juntar todos esses dados em uma única sempre não é uma tarefa fácil, uma vez em que envolve dados de baixo nível, conjuntos de elementos diferentes e afins (CARVALHO et al., 2003). Vemos, então, que a seleção de dados não é uma atividade trivial.

A qualidade dos dados está relacionada de forma direta com o nível de ruído que apresenta, sendo este proveniente de dados alterados, erros de digitação ou transmissão de dados insuficientes. Em datasets (bases de dados) pequenos, pode-se tentar substituir os outliers por valores consistentes ao domínio. Em datasets grandes é possível eliminar os dados com ruídos ou, de acordo com a conveniência, substitui-los ou ignorá-los (CARVALHO et al., 2003).

Dados de baixa qualidade levam a resultados de baixa qualidade. Logo, o pré-

processamento constitui uma técnica diferente que pode ser usada para melhorar a implementação em termos de tempo, custo e qualidade (HAN; PEI; KAMBER, 2011).

Antes de se iniciar uma pesquisa é necessário identificar o problema. Uma vez definido e entendido o problema, é necessário definir os atributos que serão utilizados na análise. Um especialista de domínio pode fornecer a um analista de dados informações sobre os atributos que ele acredita serem mais relevantes para a construção do modelo. No entanto, esse processo pode limitar a originalidade do conhecimento descoberto(REAL; NICOLETTI, 2014b).

Antes de se aplicar quaisquer técnicas estatísticas avançadas, é necessário iniciar um estudo através da análise exploratória dos dados, a partir da qual é possível conhecer as caraterísticas da base, conhecendo-se como os valores estão distribuídos, identificar os valores discrepantes e outliers. Tendo esse conhecimento em mãos consegue-se decidir como tratar os registros.(REAL; NICOLETTI, 2014b)

Sempre que possível, o analista de dados deve adicionar novos atributos e verificar a importância dessas variáveis no conhecimento gerado. Também é importante verificar se esses dados existem no banco de dados da organização ou podem ser encontrados em fontes de dados externas (MARTINS, 2003).Usar apenas um subconjunto de atributos é outra maneira de reduzir a dimensionalidade, uma vez que o conjunto de dados pode conter atributos redundantes ou irrelevantes.

Atributos redundantes replicam as informações apresentadas em um ou mais deles, enquanto os irrelevantes não possuem informações importantes para a tarefa previsível. Embora o bom senso ou o conhecimento de domínio possam ser usados para excluir alguns atributos irrelevantes e redundantes, selecionar o melhor subconjunto de requer uma abordagem sistemática (TAN; STEINBACH; KUMAR, 2009).

O próximo passo é coletar atributos que serão usados para analisar o banco de dados da organização. A coleta de dados é uma operação importante porque os dados podem não estar disponíveis em um formato adequado para uso no pré-processamento. Ou, mesmo quando disponíveis, os dados podem precisar ser rotulados com assistência especializada, conforme relatado por (PROVOST; DANYLUK, 1995), e, se for necessário, aplicar um método de classificação.

Um dos principais problemas da coleta de dados é determinar onde os dados são armazenados no banco de dados. Muitos dos sistemas de gerenciamento de dados operacionais de hoje, usados para armazenar e gerenciar bancos de dados, foram criados há muitos anos, quando as técnicas de engenharia de software não eram bem desenvolvidas e/ou difundidas. Como resultado, muitos desses sistemas são proprietários e mal documentados, tornando a coleta de dados extremamente difícil (BATISTA et al., 2003) .

Após o pré-processamento dos dados, pode ser necessário mudar a forma como os dados são representados para superar as limitações existentes no algoritmo de extração do modelo a ser aplicado. Dessa forma, propriedades com tipos de dados que não podem

ser analisados, geralmente são convertidas em outro atributo com as mesmas informações, mas com um tipo de dados que o algoritmo pode analisar. Por exemplo, um atributo do tipo data pode ser convertido em um atributo de tipo inteiro, representando o número de dias que se passaram desde uma data fixa (REAL; NICOLETTI, 2014b).

A seleção de atributos baseada em correlação avalia o valor de um subconjunto de atributos considerando a previsibilidade individual de cada recurso, bem como o grau de redundância entre eles. Subconjuntos de características com correlação de classe forte e correlação fraca são priorizados para seleção (HALL et al., 2009).

Hall et al. (2009) afirma claramente que correlação é um termo usado em seu sentido geral e não se destina a abordar especificamente a correlação linear clássica. Os autores o utilizam para especificar o grau de dependência ou previsibilidade de uma variável em relação a outra. Inconsistências podem ocorrer quando dados diferentes são representados pelo mesmo rótulo ou quando os mesmos dados são representados por rótulos diferentes. Um exemplo de inconsistência ocorre quando um atributo assume valores diferentes, que na verdade representam a mesma informação.

Existem várias fontes de poluição de dados, como ruídos, que podem ser valores fora do domínio, ausência de valores, inconsistências e afins (CARVALHO; DALLAGASSA, 2014), muito identificados na base utilizada, por exemplo. De certa forma, a poluição pode ser entendida como a presença de dados distorcidos, que não representam os valores verdadeiros. Embora os campos de um banco de dados possam ser incluídos para coletar informações valiosas, esses campos podem ser deixados em branco, incompletos ou simplesmente incorretos (REAL; NICOLETTI, 2014b).

A análise de integridade de dados geralmente envolve a análise dos relacionamentos permitidos entre os atributos. Por exemplo, um funcionário pode possuir vários carros, no entanto, um funcionário não pode ter mais de um número de funções em um determinado sistema. Dessa forma, é possível analisar propriedades em um intervalo de valores válidos (SILVA, 2021b). Um caso especial de verificação de integridade de dados é a identificação de casos extremos. Um caso extremo é um caso em que uma combinação de valores é válida porque os atributos estão dentro da faixa aceitável de valores, porém, combinar os valores dos atributos é muito improvável.

A redundância ocorre quando informações essencialmente idênticas são armazenadas em vários atributos. Um exemplo seria ter atributos na mesma tabela como preço unitário, quantidade comprada e preço total. O maior dano causado pela redundância para a maioria dos algoritmos utilizados na fase MD é o aumento do tempo de processamento. No entanto, alguns métodos são particularmente sensíveis ao número de propriedades e variáveis redundantes que podem afetar seu desempenho. Se o problema de coleta de atributos redundantes não for resolvido durante a fase de coleta de dados, métodos de pré-processamento de dados, conhecidos como métodos de seleção de atributos, podem ser usados para tentar determinar e remover atributos redundantes (SILVA, 2021b).

Um problema comum no pré-processamento de dados é a manipulação de valores indefinidos. Muitas técnicas foram aplicadas, algumas bem simples, como substituir valores desconhecidos pela média ou moda do atributo. No entanto, outras técnicas mais sofisticadas podem ser implementadas e avaliadas experimentalmente. Por exemplo, valores desconhecidos podem ser substituídos por valores previstos usando um algoritmo de aprendizado (MARTINS, 2003).

Conjuntos de dados com classes desbalanceadas são aqueles em que há uma grande diferença entre o número de exemplos pertencentes a cada valor do atributo de classe qualitativa. A maioria dos algoritmos de AM luta para gerar um modelo que classifique corretamente exemplos de classes minoritárias. Uma maneira de resolver esse problema é encontrar uma distribuição de classes que forneça um desempenho de classificador aceitável para a classe minoritária. Abaixo na figura 4 temos um exemplo de classificador de dados.

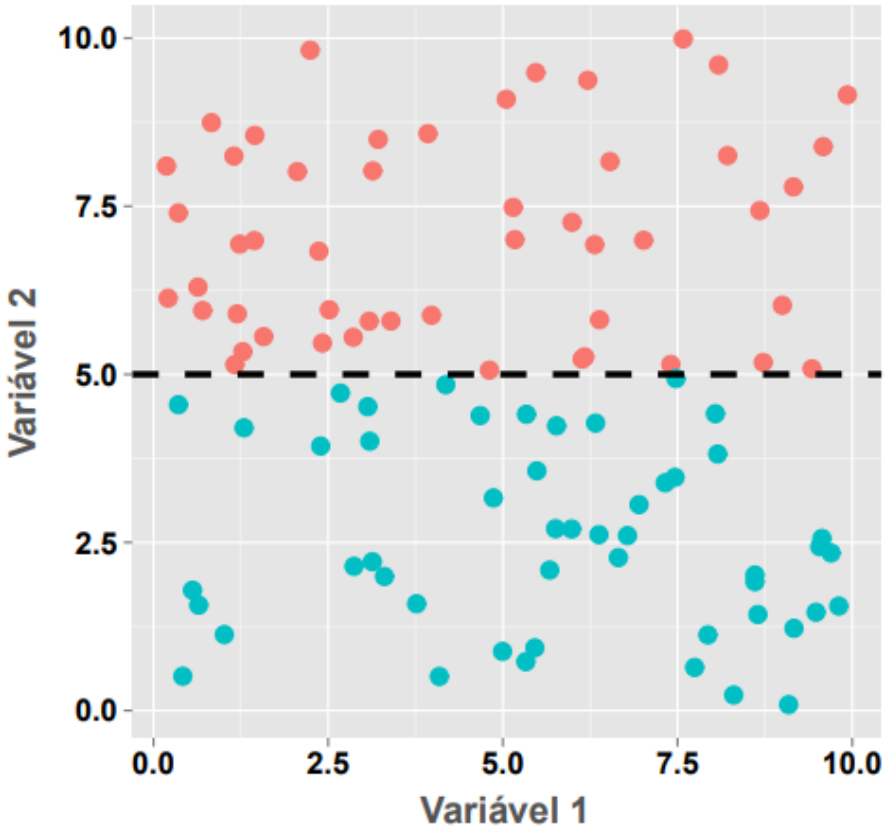

Figura 4 – Exemplo de classificador que separa as variáveis em duas classes diferentes. Fonte:(MAIONE et al., 2020)

Para se trabalhar com uma base de dados, o ideal é que estes estejam balanceados, de modo a permitir uma melhor implementação a partir deles. Existem inúmeras técnicas

de balanceamento de dados e, dentre elas, destacam-se downsample e oversample ou seja, remoção e adição de dados, respectivamente.

> "Dizemos que um conjunto de dados é balanceado quando a quantidade de amostras para todas as classes possíveis é igual ou diferente em apenas uma pequena porcentagem, de maneira que todas as classes estejam igualmente representadas por suas distribuições."(MAIONE et al., 2020)

notar na figura 5 abaixo que, à esquerda temos temos dados balanceados, uma vez que temos a mesma quantidade de valores da classe A e da classe B, em detrimento da figura da direita, que apresenta mais dados da classe A do que da classe B.

Var1	Var2	Var3	Var4	Classe
8.05	1	15.15	97.85	A
12.12	1	30.21	96.84	A
36.1	2	32.68	95.14	A
24.89	1	18.13	98.27	A
9.36	3	12.79	97.39	B
15.42	2	31.43	99.1	B
11.05	3	39.24	96.28	B
10.74	2	10.54	95.99	B

Var1	Var2	Var3	Var4	Classe
8.05	1	15.15	97.85	A
12.12	1	30.21	96.84	A
36.1	2	32.68	95.14	A
24.89	1	18.13	98.27	A
9.36	3	12.79	97.39	B
15.42	2	31.43	99.1	B

Figura 5 – Exemplos de conjunto de dados balanceado esquerda e outro conjunto de dados desbalanceado direita. Fonte: (MAIONE et al., 2020)

A autora prossegue afirmando que o principal problema enfrentado pela aprendizagem com uma base de dados desbalanceada é o viés que modelos de classificação geralmente apresentam, indo, geralmente, em direção aos dados da classe majoritária. Logo, é possível perceber que, os modelos de classificação mais populares, quando treinados, geralmente expressarão um bom desempenho nas classificações das classes majoritárias e um desempenho menor nas classes minoritárias (MAIONE et al., 2020).

2.2.2 Processos aplicados aos atributos

A seleção de atributos trata-se de um problema muito importante em pré-processamento (MATSUBARA; MARTINS; MONARD, 2003). Ele consiste em encontrar um subconjunto de atributos no qual o algoritmo de Aprendizado de Máquina utilizado irá se concentrar.

Muitos algoritmos de AM não funcionam bem com um grande número de atributos, portanto, a seleção de atributos pode melhorar o desempenho desses algoritmos. Com

menos atributos, o conhecimento gerado por algoritmos simbólicos de AM geralmente é mais fácil de entender. Alguns domínios possuem um alto custo para coleta de dados, nestes casos métodos de seleção de atributos podem reduzir o custo da aplicação. Várias abordagens são sugeridas para selecionar um subconjunto de atributos (MARTINS, 2003).

A normalização envolve a conversão de valores de atributos de seu intervalo original para um intervalo específico. Esse tipo de transformação é especialmente útil para métodos que calculam distâncias entre atributos. Por exemplo, um método como os métodos do vizinho mais próximo tende a dar mais importância aos atributos com uma faixa mais ampla de valores. Outros métodos como redes neurais são conhecidos por treinar melhor quando os valores dos atributos são pequenos. No entanto, a normalização não é muito útil para a maioria dos métodos de geração de representações simbólicas, como árvores de decisão e regras de decisão, pois a normalização tende a reduzir a compreensibilidade dos modelos gerados por esses algoritmos.

Muitos algoritmos têm a limitação de trabalhar apenas com atributos quantitativos. No entanto, muitos conjuntos de dados possuem propriedades qualitativas, e para aplicar esses algoritmos é necessário utilizar o método de conversão de um atributo qualitativo em um atributo quantitativos, ou seja, em faixas de valores. Vários métodos para personalizar propriedades foram sugeridos pela comunidade (MARTINS, 2003).

Há várias abordagens para realizar essa transformação dependendo das características e limitações de cada algoritmo. De maneira geral, atributos qualitativos sem ordem inerente, tal como verde, amarelo e vermelho, podem ser mapeados arbitrariamente para números. Contudo, esse mapeamento cria uma ordem nos valores do atributo que não é real. Atributos qualitativos ordenados, como pequeno, médio e grande, podem ser mapeados para valores numéricos para manter a ordem dos valores, como pequeno = 1, médio = 2 e grande = 3 (MARTINS, 2003).

A qualidade de dados é uma preocupação central em aprendizado de máquina e outras áreas de pesquisa relacionadas à descoberta de conhecimento em bancos de dados. Como a maioria dos algoritmos de aprendizado de máquina inferem conhecimento rigorosamente dos dados, a qualidade do conhecimento extraído é amplamente determinada pela qualidade dos dados de entrada (BATISTA et al., 2003). Um problema relacionado à qualidade de dados é a presença de valores desconhecidos, também conhecidos como valores ausentes.

Valores desconhecidos ou ausentes incluem não medir os valores de um atributo em determinadas circunstâncias. Valores desconhecidos podem ter uma série de causas como óbito de paciente, falha de equipamento, respondentes se recusando a responder algumas perguntas, entre outras. Apesar da ocorrência frequente de valores desconhecidos em um conjunto de dados, muitos analistas de dados lidam com valores desconhecidos de forma bastante simples. No entanto, o manuseio de valores desconhecidos deve ser considerado com cuidado, caso contrário, distorções podem ser introduzidas no conhecimento indutivo.

Na maioria dos casos, os atributos no conjunto de dados não são independentes uns dos outros. Dessa forma, valores aproximados podem ser determinados por meio da definição de relacionamentos entre atributos. Imputação é um termo usado para se referir a um procedimento que substitui valores desconhecidos em um conjunto de dados por valores estimados. Essa abordagem permite o tratamento de valores desconhecidos independente do algoritmo de aprendizado utilizado, permitindo ao analista de dados escolher o método mais adequado para tratar os valores desconhecidos para cada conjunto de dados (BATISTA et al., 2003). O método de imputação envolve a substituição dos valores desconhecidos de um atributo pela média dos valores conhecidos do atributo, caso o atributo seja quantitativo; ou modo de valores de atributos conhecidos, se o atributo for qualitativo.

A aleatoriedade de valores desconhecidos é um fator importante a ser considerado ao escolher um método para lidar com valores desconhecidos. Em sua forma mais simples, valores desconhecidos podem ser distribuídos aleatoriamente nos dados.(BATISTA et al., 2003). Em contraste, valores desconhecidos podem ser distribuídos de forma não aleatória. Isso significa que a probabilidade de encontrar um valor desconhecido pode, por exemplo, depender do valor verdadeiro (desconhecido) do valor desconhecido.

Existem vários métodos de lidar com valores desconhecidos disponíveis na literatura. Muitos desses métodos, como a abordagem de substituição de caso, foram desenvolvidos para pesquisa e têm certas limitações se a análise da perspectiva de análise de dados for usada no processo. Outros métodos, como substituir valores desconhecidos pela média ou moda do atributo, são muito simples e devem ser aplicados com cuidado para evitar a introdução de dados severamente distorcidos (REAL; NICOLETTI, 2014b).

Existem duas abordagens mais comumente usadas para excluir dados de valor desconhecido. O primeiro é chamado de business case completo. Este método está disponível na maioria dos programas estatísticos e é o método padrão em muitos programas. Esta abordagem consiste em eliminar qualquer caso com um ou mais valores desconhecidos. A segunda abordagem é chamada de exclusão de caso e/ou atributo. Essa abordagem envolve determinar a extensão de valores desconhecidos em cada instância e atributo e remover instâncias e/ou atributos com grandes quantidades de valores desconhecidos. As duas abordagens, analisando casos completos e descartando casos e/ou atributos, devem ser aplicadas apenas quando os valores desconhecidos são distribuídos aleatoriamente, pois os valores desconhecidos são distribuídos de forma não aleatória possui elementos não aleatórios que podem introduzir distorções aos dados (BATISTA et al., 2003).

A imputação é uma classe de procedimentos que visa substituir valores desconhecidos por valores estimados. Existem várias maneiras de estimar um valor desconhecido. As abordagens mais simples usam estatísticas obtidas de dados como a média ou moda de valores de atributos conhecidos. No entanto, métodos mais sofisticados podem usar relacionamentos entre atributos que podem ser definidos nos dados(REAL; NICOLETTI,

2014b). Conforme descrito anteriormente, os métodos de atribuição substituem valores desconhecidos por valores estimados. Os valores são estimados usando algumas informações extraídas do conjunto de dados. A seguir, são descritos alguns dos métodos de imposição amplamente utilizados na literatura.

O método da Substituição de casos é frequentemente usado em pesquisas de opinião. Um caso de valor indeterminado, por exemplo, uma pessoa não contatável, sendo substituído por outro caso, ou seja, outra pessoa, não é considerado a amostra primária da pesquisa. A imputação pela média ou pela moda é um dos mais utilizados. Envolve a substituição dos valores desconhecidos de um determinado atributo pela média, por propriedades quantitativas ou modais, por atributos qualitativos, ambos computados através do valor observável do atributo. A média é a melhor estimativa do valor de um atributo desconhecido, na ausência de outras informações sobre os dados. Esse procedimento tem a vantagem de ser conservador, pois essa substituição não altera a média geral do atributo. Por outro lado, a variância (dispersão) da variável é reduzida porque a média provavelmente está mais próxima de si mesma do que o valor verdadeiro do atributo desconhecido. Além disso, os relacionamentos entre os atributos também podem ser alterados (BATISTA et al., 2003).

O conhecimento do domínio pode ser usado pelo especialista do domínio para substituir valores desconhecidos por valores estimados a partir da experiência do especialista. Em geral, esse procedimento é seguro quando o especialista está familiarizado com o aplicativo, o conjunto de dados é grande e o número de valores desconhecidos é pequeno. Além disso, o especialista de domínio pode personalizar um atributo quantitativo por exemplo, o atributo de renda pode ser personalizado para classe A, classe B, classe C e classe D para que as pessoas possam prever com confiança. adivinhar qual tipo é sua instância. . Portanto, a variável discreta pode substituir a variável quantitativa na análise, levando em consideração a perda de informação nessa transformação (BATISTA et al., 2003).

Esse método tem a vantagem de ser menos conservador do que apenas excluir instâncias ou atributos, uma vez que busca entender a base e preencher informações baseando-se nos dados já presentes. Conhecendo a aplicação, o especialista do domínio deve ser capaz de estimar os valores com mais precisão do que substituir pela média. No entanto, esta substituição é manual e limitada a um pequeno número de valores desconhecidos. Além disso, as estimativas de especialistas são limitadas ao conhecimento existente dos dados, que de alguma forma podem orientar o conhecimento a ser coletado(BATISTA et al., 2003). Abaixo, na figura 6, temos uma representação do fluxo das técnicas de pré-processamento.

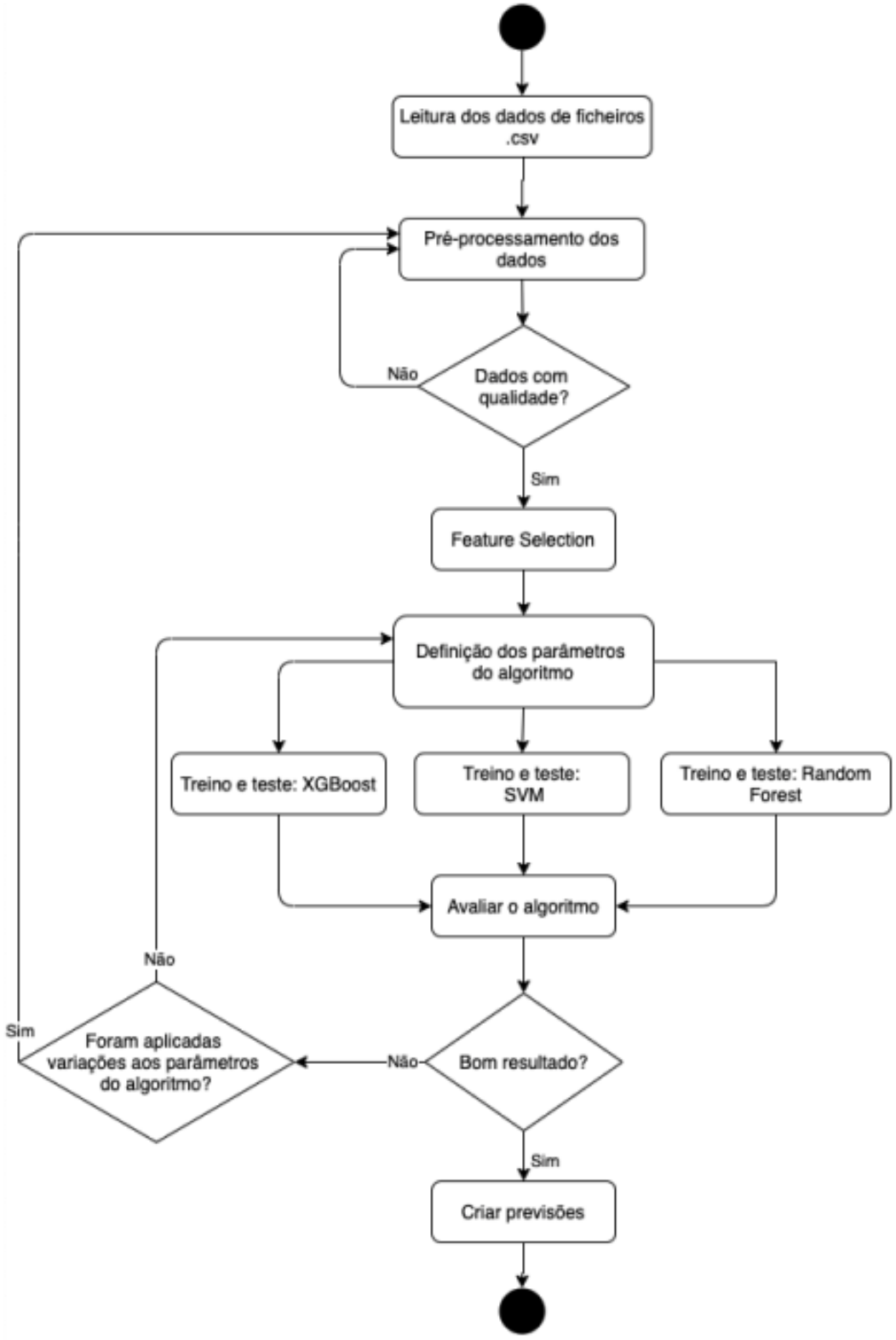

Figura 6 – Representação das técnicas de pré-processamento de dados. Fonte: (SILVA, 2021a)

2.3 Random Forest

No campo do aprendizado de máquina, espera-se que a combinação de resultados de vários classificadores forneça melhor desempenho e maior confiança na decisão do que apenas um classificador. Além disso, de acordo com Sirikulviriya e Sinthupinyo apud (LOPES et al., 2017), dado o mesmo número informações de treinamento, muitos classificadores geralmente superam um único classificador.

Consequentemente, tem havido um interesse considerável no estudo e exploração de métodos de ensemble caracterizados pela geração de muitos classificadores e pela combinação de seus resultados Dietterich apud (LOPES et al., 2017). Como um exemplo clássico de uma abordagem de conjunto, podemos citar: Boosting, bagging e, mais recentemente, Random Forest.

O algoritmo Random Forest (RF) introduzido por (BREIMAN, 2001) é um termo geral para métodos de ensemble usando classificadores do tipo árvore. A RF constrói um grande número de árvores de decisão (Figura 7) a partir de um único subconjunto de dados de treinamento definido. Este treinamento é realizado usando bagging, um meta-algoritmo que melhora a classificação e regressão do modelo com base na estabilidade e precisão da classificação (LOPES et al., 2017).

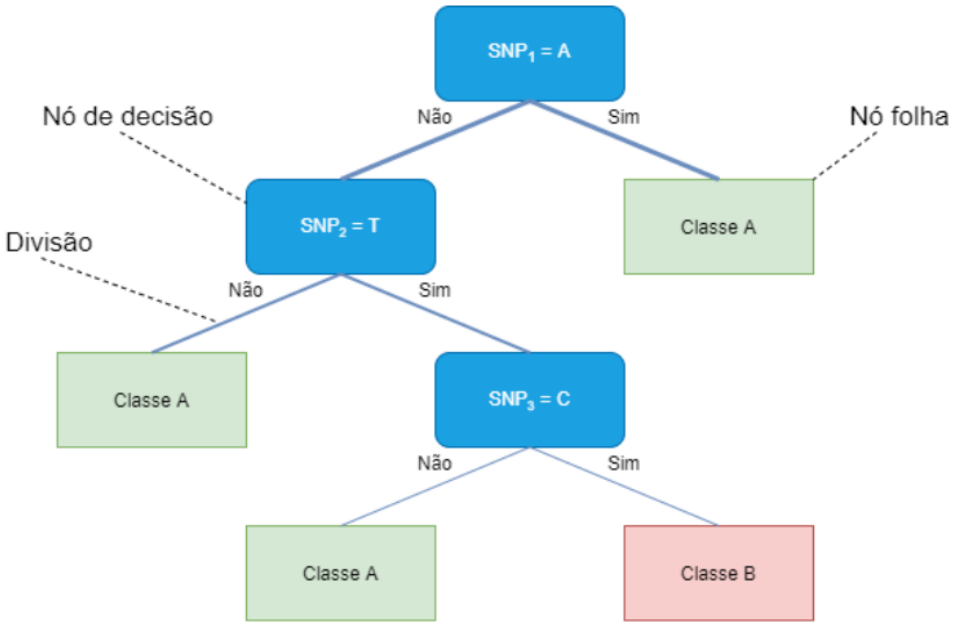

Figura 7 – Exemplo de Árvore de Decisão. Fonte: (FRAJACOMO, 2020)

O Random Forest (Figura 8) se trata, portanto, de um conjunto de Árvores de decisão treinadas a partir de amostras aleatórias do conjunto de dados, onde cada uma

das divisões é feita a partir de subconjuntos aleatórios e com amostras repetidas a fim de criar árvores levemente diferentes. A partir daí, as classificações são feitas pelo "voto"da maioria das árvores existe, o que diminui o subreajuste que ocorreria no caso de se utilizar apenas uma árvore (FRAJACOMO, 2020).

Cada árvore classificadora é identificada como Um componente de previsão. RF constrói suas decisões contando os votos dos componentes preditores em cada classe, então é escolhida a classe vencedora com base no número votos cumulativos (LOPES et al., 2017).

Portanto, todo o algoritmo consiste em duas fases importantes: o período de formação de cada árvore e o período de votação. A primeira etapa consiste em treinar cada árvore de decisão, selecionando um subconjunto de dados do conjunto de dados de treinamento e definindo-o usando uma estratégia de ensacamento aleatório. O segundo seria a aplicação desse treinamento em testes feitos pelo algoritmo.

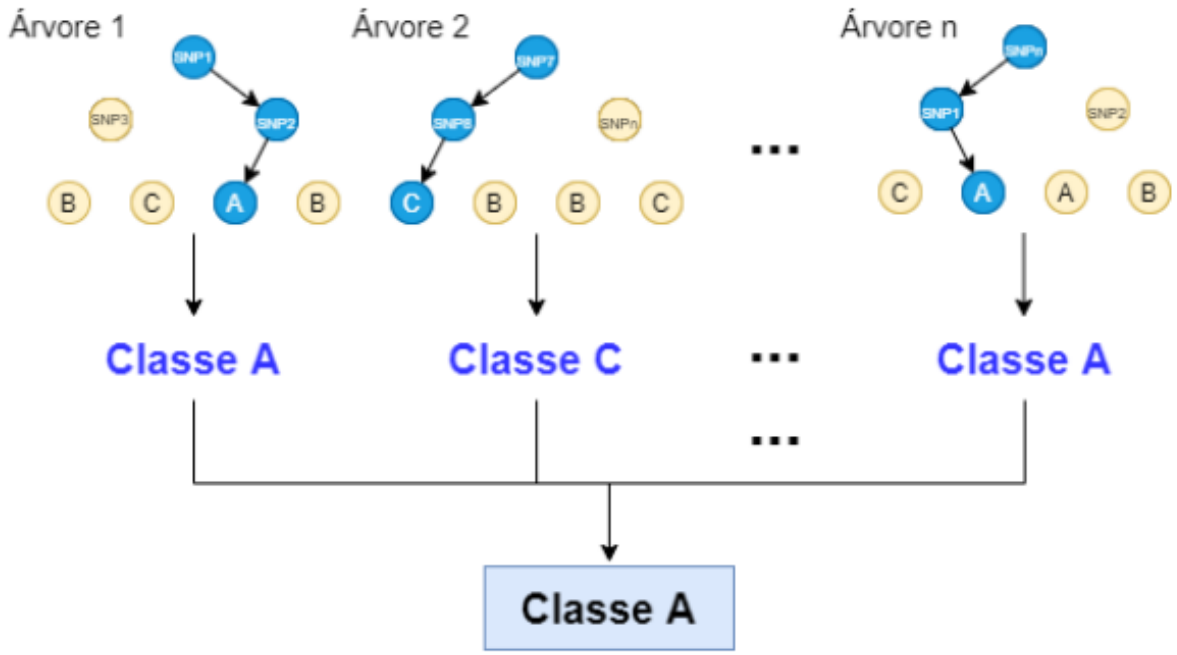

Figura 8 – Ilustração simplificada do processo de classificação Random Forest. Fonte: (FRAJACOMO, 2020)

2.4 Métricas de avaliação

A etapa seguinte consiste em avaliar e interpretar os resultados,onde o analista de dados busca descobrir se o classificador atendeu às expectativas avaliando os resultados em relação a determinadas métricas, como taxa de erro, tempo de CPU e complexidade do modelo. O especialista do domínio verificará a compatibilidade dos resultados com o conhecimento disponível no domínio. (GUIMARÃES; MEIRELES; ALMEIDA, 2019).

Há inúmeras métricas de avaliação para algoritmos de Aprendizado de Máquina, que possibilitam identificar se as técnicas de pré-processamento aplicadas resultam numa melhoria significativa para o projeto ou não. As métricas possibilitam avaliar, quantitativamente, os algoritmos de classificação utilizados nos experimentos. A acurácia pode ser definida como a porcentagem dos exemplos de teste que são classificados de forma correta (SHALEV-SHWARTZ; BEN-DAVID, 2014). Todavia, de acordo com os autores, não é possível considerar, apenas com a acurácia, se um classificador é ou não eficiente.

A sensibilidade (S), também chamada de Recall ou especificidade, é caracterizada como o percentual de instâncias classificadas de modo correto como positivas em meio às outras instâncias da base que são, efetivamente, positivas (SHALEV-SHWARTZ; BEN-DAVID, 2014):

$$S = \frac{VP}{VP + FN} \tag{1}$$

A precisão (P) é descrita como o percentual de instâncias classificadas corretamente como positivas em meio a todas as instâncias também classificadas como positivas (SHALEV-SHWARTZ; BEN-DAVID, 2014):

$$P = \frac{VP}{VP + FP} \tag{2}$$

A acurácia (A) é descrita como a fração de previsões corretas do modelo (SHALEV-SHWARTZ; BEN-DAVID, 2014) e pode ser descrita abaixo:

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{3}$$

Nesses casos, VP indica "Verdadeiro Positivo", representando as instâncias positivas que foram corretamente classificadas como tal; FP (Falso Positivo) representa as instâncias negativas, porém classificadas como positivas e FN (Falso Negativo), ou seja, as instâncias classificadas como negativas quando eram na verdade positivas.

Já a métrica F1 (ou F1-score) é um número localizado entre 0 e 1, que representa a média harmônica entre precisção e recall, através da fórmula descrita abaixo (AGARWAL, 2020):

$$2 * \frac{precision * recall}{precision + recall} \tag{4}$$

Um equilíbrio entre sensibilidade e especificidade pode ser apropriado quando há penalidades diferentes para cada tipo de erro. Neste caso, a curva ROC (Receiver

Operating Characteristic) representa uma ferramenta adequada para avaliar a sensibilidade e especificidade gerada por todos os possíveis pontos de corte para probabilidades previstas, de modo que o desempenho geral do classificador possa ser avaliado pela área abaixo da curva (AUC) ROC, ou seja, quanto maior a AUC (mais próxima de 1), melhor é a performance do modelo (MEURER; TOLLES, 2017).

3 TRABALHOS RELACIONADOS

Por ser uma área em expansão e de grande aplicabilidade, o pré-processamento de dados é amplamente estudado.

O trabalho “Pré-processamento de Dados e Comparação entre Algoritmos Aprendizado de Máquina para a Análise Preditiva de Falhas em Linhas de Produção para o Controle de Qualidade”(SILVA, 2021b) apresenta um estudo de caso com a implementação da técnica da aprendizagem supervisionada, comparando diferentes algoritmos e optando pelo XGBoost, que, no caso apresentado, apresentou os melhores resultados, apesar de o SVM ter sido mais rápido. Eles concluíram que o Random Forest se coloca no “meio-termo”entre os dois outros.

O texto “Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado, de Claudia Aparecida Martins (MARTINS, 2003), realiza uma série de experimentos em busca de um bom classificador, de modo a buscar uma representação com uma dimensão menor para o conjunto de atributos, tentando melhorar ou manter a precisão do classificador. Ela chegou à conclusão de que o algoritmo de aprendizado utilizado sempre depende do conjunto de dados e da tarefa que se quer realizar. O trabalho “Ferramentas de Pré e Pós processamento para Data Mining”(CARVALHO et al., 2003) propõe e descreve uma ferramente que auxilia as etapas de pré-processamento, mostrando a importância dessa etapa e o impacto obtido pela boa execução dela.

Já o artigo Investigação de algoritmos sequenciais de agrupamento com pré-processamento de dados em aprendizado de máquina investiga diferentes algoritmos de aprendizado de máquina não supervisionados assim como a implementação de estratégias de refinamento pós-agrupamento de dados, concluindo que algoritmos sequenciais geralmente produzem bons resultados na maioria dos conjuntos de dados, porém a ordem em que os dados são processados, assim como seus valores, podem influenciar nos agrupamentos formados (REAL; NICOLETTI, 2014b).

(BATISTA et al., 2003) explicita o quanto são importantes os algoritmos de pré-processamento de dados, especialmente nos casos de tratamento de valores desconhecidos, chegando a conclusão de que, embora não haja uma análise matemática que possa predizer se o desempenho de um método é superior aos demais em todos os casos, as análises experimentais e comparativas dos diferentes métodos aplicadas a diferentes bases podem nos dar um referencial de onde partir para aplicar as técnicas de pré-processamento.

O texto ”Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil”((SANTOS et al., 2019b)) traz uma visão diferente de como é possível aplicar Aprendizado de Máquina na saúde, apresentando as etapas relacionadas à implementação de algoritmos de AM em uma b se de dados de idosos sob uma variável de ocorrência de óbito, usando cinco algoritmos para o ajuste de

modelos (incluindo o Random Forest) de modo a tentar auxiliar profissionais de saúde na tomada de melhores decisões.

Diversos outros trabalhos foram e são desenvolvidos na área, uma vez que se trata de um ramo em expansão e de suma importância, especialmente na era da informação, em que o volume de dados adquiridos e armazenados é imenso, e muito há que se desenvolver nos trabalhos que se seguirão.

4 METODOLOGIA

Nesta seção, apresenta-se o material utilizado no desenvolvimento da pesquisa, tais como equipamentos, softwares e sistema operacional, mostrando todas as etapas dos experimentos e fazendo as comparações entre elas, de modo a auxiliar na compreensão da importância da implementação de técnicas de pré-processamento de dados.

4.1 Equipamentos

Para os experimentos foi utilizado o computador DELL (Inspiron 15 3000), processador de 10ª geração de Intel® Core™ i5-1035G1 (cache de 6MB, até 3.6GHz), placa de vídeo integrada Intel® UHD Graphics com memória gráfica compartilhada, memória de 8GB (1x8GB), DDR4, 2666MHz; Expansível até 16GB (2 slots soDIMM, 1 slot livre).

4.2 Softwares

Neste trabalho foram utilizadas técnicas de otimização e métricas de avaliação aplicadas na imputação de dados baseados na média, moda e mediana da base de dados de COVID-19 com o uso da ferramenta Google Colab. A ferramenta possui integração com o Google Drive e neste trabalho utilizamos uma base dados no formato CSV (Valores Separados por Vírgula) salvo na nuvem.

Foi utilizada a linguagem python para programação. O Python é uma linguagem de programação de alto nível - ou linguagem de alto nível - dinâmica, interpretada, modular, multiplataforma e orientada a objetos - uma forma específica de organizar software nela, grosso modo, procedimentos são objetos de classes, permitindo melhor controle de código e mais estabilidade para grandes projetos.

O estudo teve início com a chamada das principais bibliotecas que seriam usadas no trabalho: pandas(utilizada principalmente para análise de dados associada à manipulação de dados tabulares em dataframes), matplotlib (utilizada para geração de gráficos), seaborn (utilizada também para a plotagem de gráficos, através da biblioteca matplotlib), numpy (que oferece um grande conjunto de funções e operações de biblioteca para a execução de cálculos numéricos), XGBClassifier (que provém a criação de árvores paralelas e o trabalho com regressão, classificação e ranqueamento de problemas), train_test_split (utilizada para treinar e testar datasets), accuracy_score (usada para identificar o nível de acurácia) e SimpleImputer (usado para o preenchimento de dados faltantes). Em seguida, o notebook foi configurado (uma vez que os códigos foram implementados dentro do Google Colab) e o dataset original foi importado, estando ele salvo em uma pasta dentro do Google Drive.

O Random Forest, algoritmo utilizado, é um método que treina inúmeras “árvores”de decisão em paralelo ao bootstrapping, seguido de agregação, ou seja, várias árvores

de decisão são treinadas paralelamente gerando vários subconjuntos de dados de treino e garantindo que cada árvore de decisão individual na "floresta"seja única. Trata-se, portanto, de um classificador muito utilizado que, no geral, supera a maioria dos outros em termos de precisão (MISRA; LI; HE, 2019).

5 RESULTADOS E DISCUSSÕES

Para o projeto foi utilizada uma base dados de Covid-19, com 5644 linhas (cada linha representando um paciente) e 93 colunas (cada coluna identificando uma característica, ou seja, um exame, por exemplo), tendo como principais features, ou seja, as características mais relevantes para o modelo, os campos ”idade”, ”diagnóstico”, e campos como ”Hematocrit”e ”Hemoglobin”. Objetivando transformá-la da melhor maneira possível, todas as etapas pertinentes foram implementadas e três experimentos principais foram desenvolvidos.

A base (DATA4U, 2020) foi obtida ainda no começo da pandemia de Covid-19, do hospital Albert Einstein, o que quer dizer que as informações ainda não eram padronizadas. Por, na época, ainda se tratar de uma doença desconhecida, os médicos testavam diferentes exames para tentar identificar a infecção com o vírus, o que fez com que a base fosse desbalanceada, ou seja, apresentava mais dados negativos do que positivos(Figura 9), uma vez que alguns médicos realizavam determinados exames e outros não, ou seja, os pacientes eram plurais e os dados coletados eram diferentes, o que ampliou ainda mais o trabalho e a necessidade de uma etapa de pré-processamento bem desenvolvida.

Na base utilizada, foi possível perceber algumas colunas com campos completamente vazios, as quais foram excluídas, e o principal problema encontrado foram as lacunas de dados em cada uma das linhas do dataset. Para resolver esses problemas, as experiências com preenchimento de moda, meda e mediana foram implementadas. Seguem os resultados. As colunas restantes foram: ID, age, diagnosis, Patientaddmitedtoregularward, Patientaddmitetosemiintensiveunit, Patientaddmitedtointensivecareunit, Hematocrit, Hemoglobin, Platelets, Meanplateletvolume, Hbsaturation, pCO2, Baseexcess, pH, TotalCO2, HCO3, pO2, ArteiralFio2, Phosphor, ctO2, dentre outros.

A análise inicial, como comentado, se iniciou com um dataset com 5644 linhas e 93 colunas. Com uma impressão em tela inicial, pôde-se perceber que a grande maioria das células estava em branco. Através de uma impressão dos tipos presentes na base de dados, foi possível observar inúmeros dados que, visualmente eram numéricos, mas que internamente eram textuais. Uma vez que trabalhar-se-ia com porcentagens e métodos estatísticos, foi preciso modificar o tipo de dados, que eram textuais, para dados numéricos.

Alguns dados textuais (tais como “negative” e “positive”) precisavam ser transformados em números pelos mesmos motivos citados acima. Para isso, uma função de substituição foi implementada, substituindo os dados “negative” por 0, os “positive” por 1, os “not_detected” por 0, os “detected” por 1 e os “not_done” por 0. Foram analisas, para compreensão do dataset, as colunas com mais valores ausentes dentro do dataset.

Após essas primeiras análises, entram as diferenças entre cada experimento. Uma alternativa ao preenchimento dos dados faltantes através de alguma técnica seria a exclusão

destes (anteriormente foram removidas do dataset todas as colunas nulas), mas, de modo a conservar a maioria das features, optou-se por preencher os dados faltantes. A base era desbalanceada, ou seja, havia muito mais dados de um tipo, no caso diagnósticos negativos ("diagnosis" era a variável-alvo), como é possível perceber no gráfico abaixo.

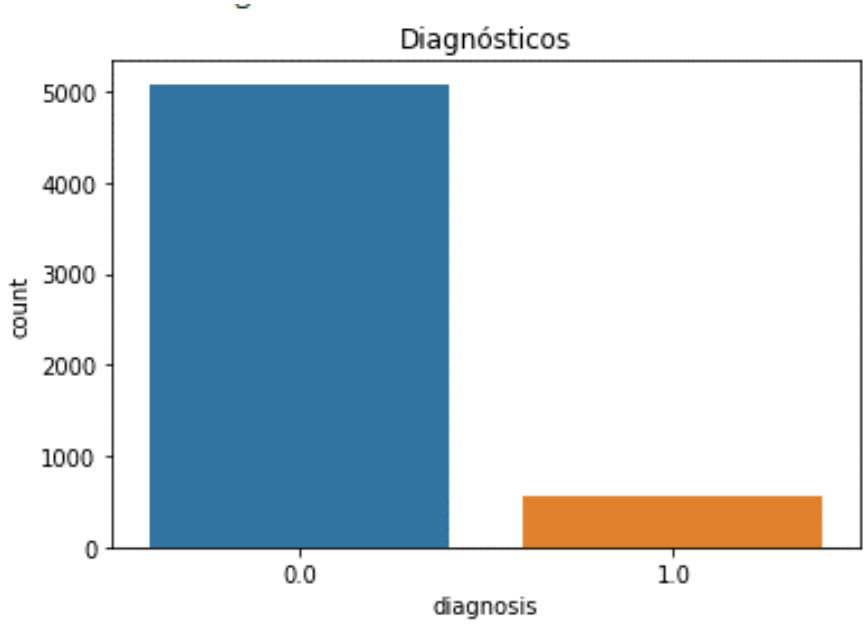

Figura 9 – Porcentagem de diagnósticos. Fonte: próprio autor

Observamos também as correlações entre as features. De acordo com o mapa de calor abaixo (Figura 10), quanto mais vermelha a escala, melhores as correlações positivas, ou seja, é possível notar que uma variável tende a aumentar quando a outra aumenta (o mesmo vale para o inverso desta afirmação). As correlações são moldadas através da correlação de Pearson.

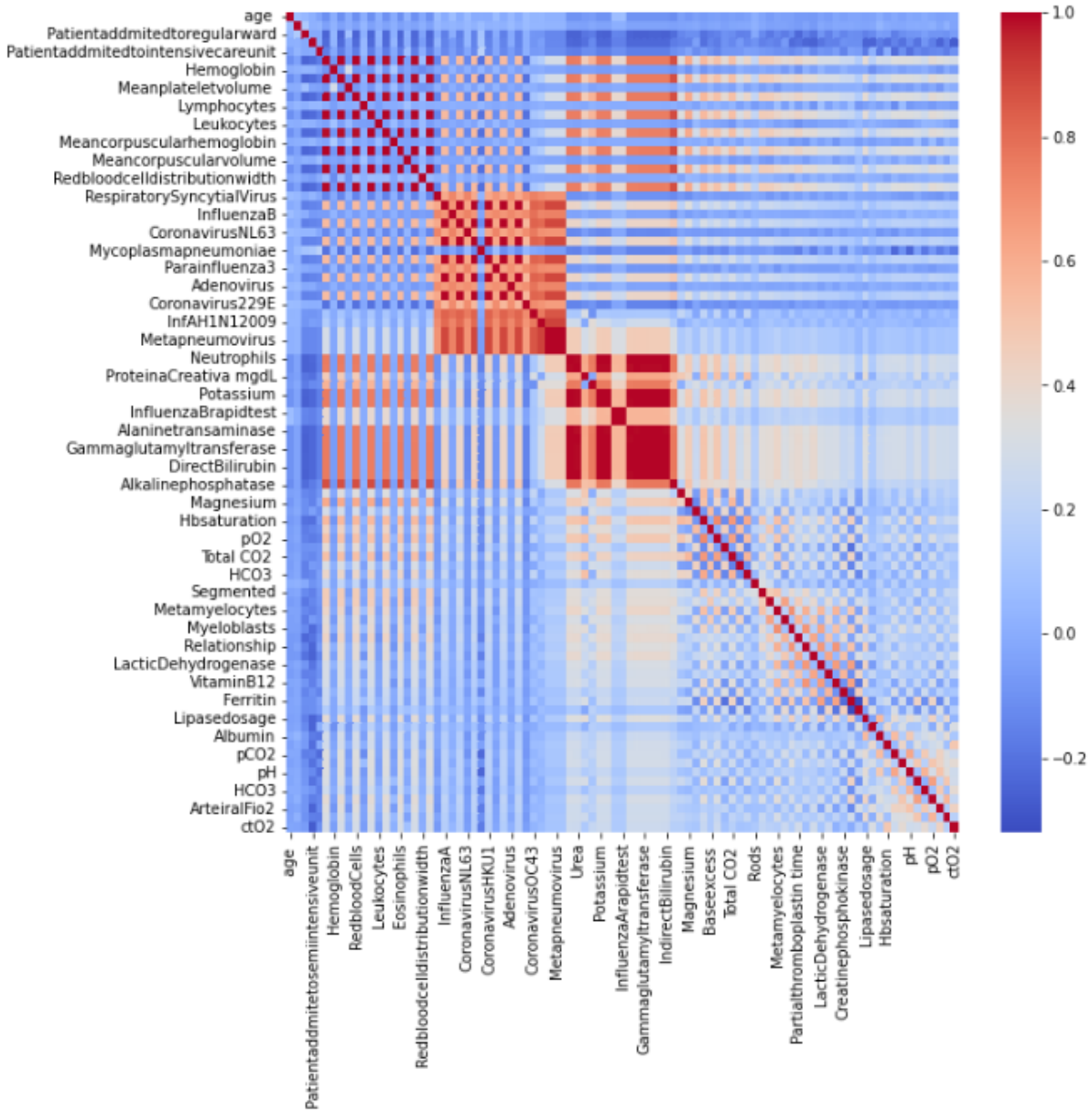

Figura 10 – Mapa de calor, mostrando a correlação entre as diferentes features. Fonte: próprio autor

No primeiro experimento, selecionado randomicamente, foram testadas as métricas de avaliação da base com o preenchimento dos dados faltantes feito através da mediana. O processo se deu através da extração da mediana da coluna em questão e preenchimento de todas as linhas vazias pertencentes a essa coluna com o valor encontrado. Em seguida, para validar a implementação, foi checado se havia valores ausentes naquela coluna. O processo se repetiu em todas as colunas.

Durante o desenvolvimento do trabalho, um dos problemas identificados foi o

nome muito grande de algumas colunas, o que dificultava a codificação. Para facilitar o estudo, estas tiveram seus nomes alterados para identificadores mais simples, o que não altera a implementação dos códigos.

Depois de padronizar os dados numéricos e do preenchimento dos dados, foram feitos os testes para avaliar o dataset. A base então preenchida foi dividida em duas: uma base de treino, que seria usada pelo algoritmo para treinar a base, e uma de teste, que, baseada no treinamento feito, verificaria os resultados (positivo ou negativo para covid-19) e confirmaria no dataset original na coluna "diagnosis". Nesse processo, foram separadas as variáveis independentes da variável-alvo, padronizadas as colunas numéricas e usado o algoritmo Random Forest, a partir do qual o modelo implementado foi instanciado e treinado.

Uma vez concluído o treinamento, o algoritmo foi testado e as métricas de avaliação foram implementadas (foram usadas as métricas acurácia, precisão e f1-score). Os dados recolhidos estão descritos na tabela abaixo.

Tabela 1 – Métricas de avaliação (base não pré-processada)

	precisão	recall	f1-score
0	0,91	0,94	0,93
1	0,26	0,017	0,02
acurácia			0,87

Tabela 2 – Métricas de avaliação (mediana)

[Acurácia] Random Forest:	0,90909090909091		
[Classification Report] Random Forest			
	precisão	recall	f1-score
0	0,91	1,00	0,95
1	0,67	0,01	0,03
acurácia			0,91

Para prosseguir com os testes de modo a tentar identificar o quanto um processo de pré-processamento pode influenciar na utilização de algoritmos de ML, um novo teste foi feito, de modo a comparar os resultados. Foram repetidos todos os procedimentos descrito acima, porém preenchendo os dados faltantes com a moda. Para as mesmas aplicações do Random Forest e com as mesmas métricas de avaliação, tivemos o seguinte resultado:

Tabela 3 – Métricas de avaliação (moda)

[Acurácia] Random Forest:	0,8931523022432113		
[Classification Report] Random Forest			
	precisão	recall	f1-score
0	0,89	1,00	0,94
1	0,80	0,04	0,08
acurácia		0,89	

Para ampliar os testes com a moda, as "árvores" utilizadas no algoritmo foram aumentadas e diminuídas até um teto e uma base. Com até 355 árvores, neste experimento, obteve-se 80% de precisão em casos positivos e 90% de acurácia. Com 50 árvores obteve-se 67% de precisão. A terceira experiência se deu repetindo os passos anteriormente mencionados, mas com o preenchimento dos dados feito através da média dos valores de cada coluna. Colunas como as que representavam idade e disgnóstico estavam todas preenchidas, porém colunas como o "Albumin", "Hbsaturation", "pCO2", "ctO2", "ProteinaCreativa mgdL","Phosphor", "DDimer","ArterialLacticAcid", "Lipasedosage"e "pH", as 10 colunas com mais valores ausentes, tinham mais de 5400 dados faltantes em suas linhas, então o preenchimento dessas informações se deu de forma imprescindível. Os mesmos testes foram feitos e, com 120 "árvores" os seguintes resultados foram obtidos:

Tabela 4 – Métricas de avaliação (média)

[Acurácia] Random Forest:	0,9055489964580874		
[Classification Report] Random Forest			
	precisão	recall	f1-score
0	0,91	1,00	0,95
1	1,00	0,01	0,02
acurácia		0,91	

Devemos sempre nos lembrar que cada métrica de avaliação pode ser utilizada para um fim diferente. A acurácia, por exemplo, é ideal para dados balanceados, porém para os desbalanceados, ela por si só não basta. Podemos ver isso pensando, por exemplo, em uma caixa de emails onde queremos identificar se um email é spam ou não. Se tivermos 100 emails e dentre eles apenas um for spam, e dissermos que todos os emails não são spam, teremos um nível de acerto muito alto mesmo que o modelo não seja eficiente.

A precisão objetiva diminuir os falsos positivos encontrados, enquanto o recall busca diminuir os falsos negativos, o que aumenta o seu valor. Outra métrica de avaliação importante é a Matriz de Confusão. Podemos analisar a seguir a matriz de confusão, que mostra os verdadeiros positivos, falso positivos, verdadeiros negativos e falso negativos, obtida com a implementação do preenchimento de dados com a média:

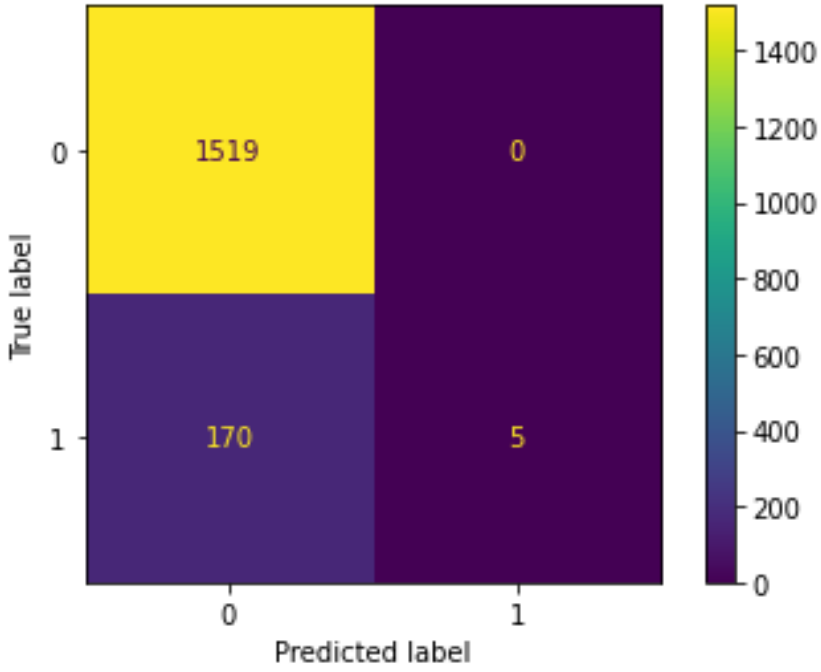

Figura 11 – Matriz de confusão. Fonte: próprio autor

Podemos ver abaixo um comparativo entre média, moda e mediana diante das métricas de avaliação usadas:

Tabela 5 – Tabela comparativa

	Média	Moda	Mediana
Acurácia	0,9	0,89	0,9
Precisão em casos positivos	1	0,8	0,67
Precisão para casos negativos	0,91	0,89	0,91
F1-score para casos positivos	0,02	0,08	0,03
F1-score para casos negativos	0,95	0,94	0,95

Ao final destes testes, para possível utilização dos dados em trabalhos futuros, foram selecionadas as features relevantes para o modelo, ou seja, todas as que não possuíam seus dados completamente nulos, o que poderá melhorar ainda mais o desempenho do algoritmo na etapa de processamento de dados. Foi possível, portanto, que a aplicação de técnicas de pré-processamento realmente influenciam na qualidade dos resultados do processamento de um dado algoritmo, uma vez que os resultados foram diferentes a cada nova experiência feita na mesma base de dados com o mesmo algoritmo, o que quer dizer que, uma vez dedicado um tempo para se trabalhar com o pré-processamento de dados, identificando e aplicando as melhores técnicas, melhores resultados serão obtidos.

6 CONCLUSÃO

A área de Machine Learning é uma área em expansão e o conhecimento dos métodos que permeiam essa ciência é primordial para seu aprimoramento. Pudemos ver ao longo do trabalho diferentes técnicas de pré-processamento de dados, bem como sua importância para a melhoria dos resultados das aplicações de algoritmos de ML, compreendendo as limitações de cada uma e conseguindo avaliar a qualidade do trabalho com esses dados.

O presente estudo apresentou a aplicação de alguns métodos de pré-processamento, e posteriormente a implementação do algoritmo Random Forest, finalizando com a avaliação dessas técnicas através de métricas matemáticas. Os métodos foram empregados em uma base de dados de COVID-19 de modo a identificar, baseado no treinamento realizado com a base de dados, quando um paciente seria dado como positivo ou negativo para a doença. Foram apresentadas as etapas de preenchimento dos dados faltantes e, para esse caso, o preenchimento com a média dos dados trouxe melhores resultados. Através desses resultados foi possível analisar o desempenho do classificador e a implementação da técnicas de pré-processamento.

O pré-processamento de dados é uma das técnicas utilizadas para a melhoria dos resultados de algoritmos de reconhecimento de padrões e classificação, incluindo aprendizado de máquina, portanto, é crucial entender as formas diferentes de realizar o pre-processamento e avaliação dos resultados, comparando-os a situações onde não houve limpeza, integração, transformação e redução da base dados, antes da implementação de um determinado algoritmo.

A utilização de diferentes técnicas de pré-processamento de dados influenciou diretamente nos resultados, uma vez que cada implementação possui sua particularidade e que o resultado não pode ser previsto (em se tratando em melhoria de resultados quando comparadas as diferentes técnicas), o que implica na aplicação de várias técnicas.

O trabalho em questão, portanto, mostrou que o pré-processamento de dados pode influenciar no resultado da implementação de um algoritmo de Machine Learning, uma vez passadas todas as etapas aqui discutidas. Pudemos verificar, por exemplo, diferenças entre os resultados obtidos com o preenchimento de dados faltantes com a média, moda e mediana dos valores presentes, isso tudo implementado na mesma base de dados, mostrando precisão e acurácia diferentes em cada implementação.

Percebemos que a acurácia sozinha não é suficiente para permitir a averiguação dos resultados obtidos, porém vimos também que, combinada a outros métodos, é possível identificar as melhores soluções para se trabalhar com datasets desbalanceados e/ou com dados faltantes.

É importante salientar o porquê dos diferentes resultados quando aplicadas técnicas diferentes a uma mesma base de dados, entendendo que, diferentes técnicas representam

diferentes repercussões nos resultados finais, uma vez que o modo de preenchimento dos dados faltantes é diferente em cada uma das implementações.

As dificuldades encontradas neste trabalho se relacionaram especialmente ao aprendizado da linguagem de programação Pyhthon por parte do autor, assim como a compreensão e implementação das diferentes técnicas aqui apresentadas e ao emprego das diferentes métricas de avaliação dessas implementações, partindo não apenas de uma base teórica mas de um conjunto entre teoria e prática no que permeia o assunto.

6.1 Trabalhos Futuros

Para experimentos e pesquisas futuros sugere-se um estudo mais aprofundado das diferentes técnicas de pré-processamento, comparando os resultados obtidos neste trabalho com os apresentados ao se preencher os dados faltantes com os valores mais prováveis, através de técnicas estatísticas, uma vez que este documento apresenta o que se gerou do preenchimento com a moda, a média e a mediana dos dados na base. Sugerem-se também estudos que tratem de forma diferente os valores discrepantes que foram excluídos, de modo a testar diferentes formas de tratamento que não se resumem à mera exclusão. A comparação do tempo computacional de cada um dos resultados obtidos já implementados em um algoritmo de processamento de dados também seria interessante, mensurando a qualidade das aplicações individualmente e comparando-as. O teste de outras técnicas com a mesma base ou a implementação das mesmas técnicas numa base semelhante, porém com mais dados, de forma a perceber e comparar seus resultados, também é uma sugestão de trabalho futuro, assim como o trabalho com o balanceamento de classes.

Referências

AGARWAL, R. The 5 classification evaluation metrics every data scientist must know. **Towards Data Science**, 2020. Citado na página 21.

BATISTA, G. E. d. A. P. et al. **Pré-processamento de dados em aprendizado de máquina supervisionado**. Tese (Doutorado) — Universidade de São Paulo, 2003. Citado 6 vezes nas páginas 2, 11, 15, 16, 17 e 23.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 19.

CARVALHO, D. R. et al. Ferramenta de pré e pós-processamento para data mining. **SEMINÁRIO DE COMPUTAÇÃO, VII, Blumenau, Brasil**, 2003. Citado 2 vezes nas páginas 10 e 23.

CARVALHO, D. R.; DALLAGASSA, M. R. Mineração de dados: aplicações, ferramentas, tipos de aprendizado e outros subtemas. **AtoZ: novas práticas em informação e conhecimento**, v. 3, n. 2, p. 82–86, 2014. Citado na página 12.

COLLINS, C. et al. Artificial intelligence in information systems research: A systematic literature review and research agenda. **International Journal of Information Management**, Elsevier, v. 60, p. 102383, 2021. Citado na página 3.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, Springer, v. 20, n. 3, p. 273–297, 1995. Citado na página 6.

DATA4U, E. **Diagnosis of COVID-19 and its clinical spectrum**. 2020. Disponível em: <https://www.kaggle.com/datasets/einsteindata4u/covid19>. Acesso em: 08 de agosto 2022. Citado na página 27.

FERNANDEZ, G. **Data Mining Using SAS Applications: A Case Study Approach**. [S.l.]: SAS Publishing, 2003. Citado na página 5.

FRAJACOMO, H. C. Seleção de snps utilizando random forests. Universidade Federal de São Carlos, 2020. Citado 3 vezes nas páginas , 19 e 20.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Sparse inverse covariance estimation with the graphical lasso. **Biostatistics**, Oxford University Press, v. 9, n. 3, p. 432–441, 2008. Citado na página 4.

GONZALEZ, S. **Por que o Machine Learning é um grande aliado da cibersegurança**. 2021. Disponível em: <https://www.welivesecurity.com/br/2021/12/29/por-que-o-machine-learning-e-um-grande-aliado-da-ciberseguranca/>. Acesso em: 08 de agosto 2022. Citado 2 vezes nas páginas e 8.

GUIMARÃES, L. M. S.; MEIRELES, M. R. G.; ALMEIDA, P. E. M. d. Avaliação das etapas de pré-processamento e de treinamento em algoritmos de classificação de textos no contexto da recuperação da informação. **Perspectivas em Ciência da Informação**, SciELO Brasil, v. 24, p. 169–190, 2019. Citado na página 20.

HALL, M. et al. The weka data mining software: an update. **ACM SIGKDD explorations newsletter**, ACM New York, NY, USA, v. 11, n. 1, p. 10–18, 2009. Citado na página 12.

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. [S.l.]: Elsevier, 2011. Citado 3 vezes nas páginas 9, 10 e 11.

KUHN, M.; JOHNSON, K. et al. **Applied predictive modeling**. [S.l.]: Springer, 2013. v. 26. Citado 2 vezes nas páginas 4 e 9.

LOPES, T. et al. Aplicação do algoritmo random forest como classificador de padrões de falhas em rolamentos de motores de indução. **XIII Simpósio Brasileiro de Automação Inteligente, Porto Alegre**, 2017. Citado 2 vezes nas páginas 19 e 20.

LUDERMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. **Estudos Avançados**, SciELO Brasil, v. 35, p. 85–94, 2021. Citado na página 1.

MAIONE, C. et al. Balanceamento de dados com base em oversampling em dados transformados. Universidade Federal de Goiás, 2020. Citado 3 vezes nas páginas , 13 e 14.

MARTINS, C. A. **Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado**. Tese (Doutorado) — Universidade de São Paulo, 2003. Citado 4 vezes nas páginas 11, 13, 15 e 23.

MATSUBARA, E. T.; MARTINS, C. A.; MONARD, M. C. Pretext: Uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. **Techinical Report**, v. 209, n. 4, p. 10–11, 2003. Citado na página 14.

MEURER, W. J.; TOLLES, J. Logistic regression diagnostics: understanding how well a model predicts outcomes. **Jama**, American Medical Association, v. 317, n. 10, p. 1068–1069, 2017. Citado na página 22.

MISRA, S.; LI, H.; HE, J. **Machine learning for subsurface characterization**. [S.l.]: Gulf Professional Publishing, 2019. Citado 2 vezes nas páginas 6 e 26.

MONARD, M. C.; BARANAUSKAS, J. A. Aplicações de inteligência artificial: uma visão geral. **Anais - Repositório USP**, 2000. Citado 2 vezes nas páginas e 4.

OLIVEIRA, W. K. d. et al. Como o brasil pode deter a covid-19. SciELO Public Health. Citado na página 1.

PROVOST, F.; DANYLUK, A. Learning from bad data. In: CITESEER. **Proceedings of the ML-95 Workshop on Applying Machine Learning in Practice**. [S.l.], 1995. Citado na página 11.

RASCHKA, S.; MIRJALILI, V. Python machine learning: Machine learning and deep learning with python. **Scikit-Learn, and TensorFlow. Second edition ed**, v. 10, p. 3175783, 2017. Citado 2 vezes nas páginas 7 e 9.

REAL, E. M.; NICOLETTI, M. d. C. Investigação de algoritmos sequenciais de agrupamento com pré-processamento de dados em aprendizado de máquina. **Faculdade Campo Limpo Paulista. Campo Limpo Paulista**, 2014. Citado 2 vezes nas páginas e 5.

REAL, E. M.; NICOLETTI, M. d. C. Investigação de algoritmos sequenciais de agrupamento com pré-processamento de dados em aprendizado de máquina. **Faculdade Campo Limpo Paulista. Campo Limpo Paulista**, 2014. Citado 6 vezes nas páginas 6, 11, 12, 16, 17 e 23.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. **Revista de Sistema de Informação da FSMA**, Macaé, n. 7, p. 7-21, 2011., 2011. Citado na página 9.

SANTOS, H. G. d. 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**, SciELO Public Health, v. 35, p. e00050818, 2019. Citado na página 4.

SANTOS, H. G. dos et al. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de são paulo, brasil. **Cad. Saúde Pública**, v. 35, n. 7, p. e00050818, 2019. Citado na página 23.

SCHMITT, J. et al. Pré-processamento para a mineração de dados: uso da análise de componentes principais com escalonamento ótimo. ório UFSC, 2005. Citado na página 5.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. **Understanding machine learning: From theory to algorithms**. [S.l.]: Cambridge university press, 2014. Citado 2 vezes nas páginas 3 e 21.

SILVA, D. F. B. F. d. **Pré-processamento de Dados e Comparação entre Algoritmos de Machine Learning para a Análise Preditiva de Falhas em Linhas de Produção para o Controlo**. Tese (Doutorado), 2021. Citado 2 vezes nas páginas e 18.

SILVA, D. F. B. F. d. **Pré-processamento de Dados e Comparação entre Algoritmos de Machine Learning para a Análise Preditiva de Falhas em Linhas de Produção para o Controlo**. Tese (Doutorado), 2021. Citado 2 vezes nas páginas 12 e 23.

SILVA, D. F. B. F. da; MARREIROS, G. **Pré-processamento de Dados e Comparação Entre Algoritmos de Machine Learning Para a Análise Preditiva de Falhas em Linhas de Produção Para O Controlo**. Tese (Doutorado) — Instituto Politecnico do Porto (Portugal), 2021. Citado na página 6.

SIMON, H. Redes neurais–princípios e prática. **Bookman**, 2001. Citado na página 1.

SOMERS, J. The man who would teach machines to think. **The Atlantic**, v. 11, 2013. Citado 2 vezes nas páginas 3 e 5.

SOUZA, E. P. d. et al. Aplicações do deep learning para diagnóstico de doenças e identificação de insetos vetores. **Saúde em Debate**, SciELO Brasil, v. 43, p. 147–154, 2020. Citado na página 1.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao datamining: mineração de dados**. [S.l.]: Ciência Moderna, 2009. Citado 3 vezes nas páginas 9, 10 e 11.

TURING, A. M. Mind. **Mind**, v. 59, n. 236, p. 433–460, 1950. Citado na página 3.

WITTEN, I. et al. Algorithms: the basic methods. **Data Mining: Practical machine learning tools and techniques**, Morgan Kaufmann Boston, p. 86–87, 2011. Citado na página 9.