Dissertação apresentada à Pró-Reitoria de Pós-Graduação e Pesquisa do Instituto Tecnológico de Aeronáutica e da Universidade Federal de São Paulo, como parte dos requisitos para obtenção do título de Mestre em Ciências no Programa de Pós-Graduação em Pesquisa Operacional, Área de Engenharia de Produção/Pesquisa Operacional.

Alex Fernandes de Souza

PROTOCOLO DE COLETA DE DADOS PARA PREDIÇÃO DE COVID-19

Dissertação aprovada em sua versão final pelos abaixo assinados:

Prof. Dr. Filipe Alves Neto Verri

Orientador

Profa. Dra. Emília Villani

Pró-Reitora de Pós-Graduação

Campo Montenegro
São José dos Campos, SP - Brasil
2022

Dados Internacionais de Catalogação-na-Publicação (CIP)
Divisão de Informação e Documentação

Fernandes de Souza, Alex
Protocolo de coleta de dados para predição de COVID-19 / Alex Fernandes de Souza.
São José dos Campos, 2022.
61f.

Dissertação de Mestrado – Curso de Pesquisa Operacional. Área de Engenharia de Produção/Pesquisa Operacional – Instituto Tecnológico de Aeronáutica e Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo, 2022. Orientador: Prof. Dr. Filipe Alves Neto Verri.

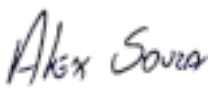
1. Coleta de dados. 2. Covid-19. 3. Aprendizado de máquina. I. Instituto Tecnológico de Aeronáutica. II. Universidade Federal de São Paulo. III. Título.

REFERÊNCIA BIBLIOGRÁFICA

FERNANDES DE SOUZA, Alex. **Protocolo de coleta de dados para predição de COVID-19**. 2022. 61f. Dissertação de Mestrado – Instituto Tecnológico de Aeronáutica e Universidade Federal de São Paulo, São José dos Campos.

CESSÃO DE DIREITOS

NOME DO AUTOR: Alex Fernandes de Souza
TITULO DO TRABALHO: Protocolo de coleta de dados para predição de COVID-19.
TIPO DO TRABALHO/ANO: Dissertação / 2022



Alex Fernandes de Souza
Av. Cidade Jardim, 679
12.233-066 – São José dos Campos–SP

PROTOCOLO DE COLETA DE DADOS PARA PREDIÇÃO DE COVID-19

Alex Fernandes de Souza

Composição e Assinatura da Banca Examinadora:

Profª. Drª.	Ana Carolina Lorena	Presidente	-	Universidade
Prof. Dr.	Filipe Alves Neto Verri	Orientador	-	ITA
Profª. Drª.	Ana Carolina Lorena		-	ITA
Prof. Dr.	Marcos Gonçalves Quiles		-	UNIFESP
Prof. Dr.	Rafael Paes Leme		-	EMAER

ITA/UNIFESP

Dedico este trabalho à minha mãe, Maria José, e ao Alisson Juan que me apoiaram de forma incondicional ao longo desta etapa.

Agradecimentos

Agradeço primeiramente à minha mãe Maria José, que sempre me incentivou a estudar, mesmo apesar das condições adversas de nossa vida. Agradeço pelo seu suporte, carinho e compreensão em todos os momentos. Agradeço também ao professor Filipe Alves Neto Verri por sua orientação presente, se mostrando disponível sempre que necessário, problematizando e levantando questões que fazem com que seus alunos reflitam sobre seu objeto de estudo. Também agradeço por seu lado humano, fazendo das reuniões semanais algo tranquilo, mas sempre muito produtivo. Agradeço a todo corpo docente do ITA e UNIFESP e aos colegas de classe com quem tive maior contato. O Mestrado na Pesquisa Operacional trouxe um desenvolvimento pessoal e profissional imensurável, contribuindo para que novas oportunidades surgissem. Por fim, agradeço também à professora Ana Carolina Lorena. Os primeiros avanços em meu estudo veio em sua disciplina. Além disso, suas ações dentro do programa, e especificamente dentro deste projeto, mostram seu empenho por reunir alunos, profissionais da área e pesquisadores, sempre em prol da ciência e do crescimento dos alunos.

Resumo

A coleta de dados representa um desafio em diversos setores da sociedade. Na pandemia de Covid-19, grandes volumes de dados foram gerados com a finalidade de usá-los em tarefas de aprendizado de máquina (AM) para auxiliar na tomada de decisão. Contudo, a forma como estes dados foram coletados dificulta a elaboração de análises estatísticas e uso em tarefas de diagnóstico e prognóstico. Estas análises demandam conjuntos de dados arrumados, que representam uma forma de conectar a estrutura dos dados à sua semântica. Este estudo propõe um protocolo de coleta de dados a partir do estudo de *datasets* clínicos disponibilizados no Repositório do COVID-19 DataSharing/BR para uso em tarefas de aprendizado de máquina. Foram analisados dados do Laboratório Fleury, que apontam o diagnóstico, e dados do Hospital Sírio-Libanês, que permitem estudar o prognóstico dos casos. Ambos os *datasets* demandaram um extenso pré-processamento e, em seguida, foram arrumados para que pudessem ser utilizados em tarefas de AM. Entre os problemas observados ao longo das etapas de pré-processamento, destacam-se a falta de padronização, informações redundantes, atributos sem relevância, dados ausentes, entre outros. Após o pré-processamento inicial, ambos os conjuntos foram arrumados de modo que tornassem seu uso eficiente. Na sequência, outras tarefas foram realizadas para tornar os dados utilizáveis, eliminado, por exemplo, a extensa quantidade de valores ausentes. Com os dados arrumados, aplicou-se três técnicas preditivas de AM, sendo estes *K-Nearest Neighbor* (KNN), *Support-Vector Machine* e Árvore de decisão. Na tarefa de diagnóstico de Covid-19, a técnica KNN apresentou melhor desempenho com valores de área sob a curva ROC igual a 0.77. Para os dados de prognóstico de Covid-19, os algoritmos KNN e SVM apresentaram os melhores desempenho, ambos com 0.81 da mesma medida. A partir desses resultados, pode-se afirmar que os conjuntos de dados, dentro de uma estrutura arrumada, podem ser utilizados no auxílio ao diagnóstico e prognóstico de Covid-19. Logo, a partir do protocolo de coleta de dados proposto neste estudo, o qual garante a obtenção de dados em formato arrumado, observou-se a redução da necessidade de diversas tarefas de pré-processamento. Assim, o uso dos dados em tarefas de aprendizado de máquina e análises estatísticas é facilitado, potencializando também o manejo eficiente de pacientes e recursos hospitalares. Além disso, este protocolo pode ser utilizado em eventos futuros, facilitando a forma como os dados são coletados e seu uso subsequente.

Abstract

Data collection represents a challenge in various sectors of society. In the Covid-19 pandemic, large volumes of data were generated. Machine learning (ML) techniques use these data to assist in decision making. However, the way data is usually collected makes it difficult to prepare statistical analyzes and use them in diagnostic and prognostic tasks. Tidy datasets eases this tasks, representing a way of connecting the data structure to its semantics. This study proposes a data collection protocol from the study of clinical datasets available in the COVID-19 DataSharing/BR Repository for use in machine learning tasks. We analyze data from the Fleury Laboratory, which point to the diagnosis, and data from the Sírio-Libanês Hospital, which allow the study of the prognosis of cases. Both datasets required extensive pre-processing, and we tidied them up so they could be used in ML tasks. Among the problems observed during the pre-processing stages, we highlight the lack of standardization, redundant information, irrelevant attributes, and missing data. After the initial pre-processing, we arranged both sets to make them efficient to use. Subsequently, other tasks were performed to make the data usable, eliminating, for example, the large amount of missing values. With the data arranged, three predictive models of ML were trained, these being K-Nearest Neighbor (KNN), Support-Vector Machine (SVM) and Decision Tree. In the Covid-19 diagnostic task that used data from the Fleury Laboratory, the KNN technique presented the best performance with values of area under the ROC curve (AUC) equals to 0.77. For the Covid-19 prognostic data using data from Hospital Sírio-Libanês, the KNN and SVM algorithms showed the best performance, both with AUC equals to 0.81. From these results, we conclude that the datasets, in a tidy structure, can be used efficiently to aid in the diagnosis and prognosis of Covid-19. The data collection protocol proposed in this study, which aims to obtain data in a tidy format, avoids several pre-processing tasks. Thus, it facilitates the use of data in machine learning tasks and statistical analysis, and also enhances the efficient management of patients and hospital resources. In addition, this protocol can be used in future events, facilitating the way data is collected and its subsequent use.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

KNN	*K Nearest Neighbor*
RF	*Random Forest*
SVM	*Support Vector Machine*
IA	Inteligência Artificial
AM	Aprendizado de Máquina
ML	*Machine Learning*
ROC	*Receiver Operating Characteristic*
AUC	*Area Under the Curve*
SD	Desvio Padrão
CEP	Código de Endereçamento Postal

Sumário

1 Introdução

As técnicas de aprendizado de máquina (AM) em inteligência artificial (IA) têm sido utilizadas com sucesso em diversas análises descritivas e preditivas. Nesse contexto e de forma geral, compreende-se como IA a capacidade de máquinas executarem tarefas complexas associadas a seres inteligentes. Já AM refere-se à capacidade de algoritmos aprenderem a partir de um conjunto de dados. Neste ponto destacam-se dois tipos de análises: descritiva, que busca descrever as características dos dados, e preditivas que tem como objetivo fazer previsões com base em dados já coletados(FACELI *et al.*, 2011).

Na área médica, sistemas de AM são utilizados na geração de modelos para suporte ao diagnóstico médico (KONONENKO, 2001; OBERMEYER; EMANUEL, 2016) e no prognóstico de pacientes frente à evolução de seu quadro sintomático (KOUROU *et al.*, 2015), por exemplo. Contudo, como tais modelos são obtidos a partir de dados, deve haver também a preocupação quanto à qualidade deles e os registros médicos devem ser cuidadosamente curados antes de seu uso (OBERMEYER; EMANUEL, 2016).

No caso de ambientes hospitalares, uma grande massa de registros dos pacientes admitidos é mantida, desde dados pessoais (nome, endereço, plano de saúde, etc.) até resultados de exames clínicos e o diagnóstico aferido por um ou mais profissionais. Contudo, no Brasil, não há a cultura de uso de tais dados em análises mais sofisticadas, tais como as providas por técnicas de AM. Sendo assim, a qualidade dos dados é bastante deteriorada. Ademais, os profissionais da área de saúde muitas vezes não têm ideia do tipo de informação que pode agregar valor às análises necessárias.

Dessa forma, são necessários diversos procedimentos para tornar esses dados deteriorados em dados trabalháveis. Esses procedimentos visam reduzir o tempo entre coleta e seu uso, seja em tarefas de AM para auxílio no diagnóstico e prognóstico, ou ainda na tomada de decisões gerais em relação ao manejo de paciente em um contexto onde o tempo é um fator crucial. Logo, reduzir as etapas de pré-processamento através de uma coleta apropriada de informações, possibilita a geração de um conjunto de dados organizado, potencializando as operações realizadas nos espaços de saúde, em especial no que diz respeito à Covid-19.

De antemão, é importante trazer dois conceitos fundamentais para este estudo, sendo

estes diagnóstico e prognóstico. Diagnóstico refere-se ao fato de determinar se um sujeito tem ou não uma doença, considerando exames clínicos ou sintomas observados; prognóstico, por sua vez, indica a evolução de uma doença e suas prováveis consequências (SOUSA; RIBEIRO, 2009). Esses termos são extensivamente utilizados ao longo deste texto.

Além disso, a pandemia do Coronavírus mostrou que o mundo necessita de um plano de ação eficaz e ágil para eventos futuros. Sendo assim, a determinação de um formato de coleta de dados para o caso da Covid-19 tem potencial para aplicação também em outros contextos. Este protocolo pode ser utilizado para novas doenças ainda desconhecidas pelo homem, fornecendo diretrizes de como coletar as informações e assim estudar seus padrões, características e aplicação em modelos de AM.

Nesse contexto, este trabalho busca responder o seguinte problema de pesquisa: Como uma coleta de dados em formato arrumado pode contribuir para o uso eficiente dos dados hospitalares no auxílio ao diagnóstico e prognóstico de Covid-19 por meio de aprendizado de máquina?

A hipótese adotada para este problema considera que, a partir de um sistema eficiente de coleta, pode-se gerar conjuntos de dados apropriados para análises. Dessa forma, a obtenção de estatísticas descritivas pode ser facilitada, o uso em tarefas de aprendizado de máquina pode ser mais rápido, bem como a realização de outras investigações quaisquer, sem a necessidade de realizar diversas etapas de pré-processamento para tornar os dados trabalháveis e compreensíveis.

1.1 Objetivo

O objetivo deste projeto de mestrado foi analisar *datasets* de diagnóstico e prognósticos de Covid-19 com a finalidade de propor um método de coleta de dados em formato arrumado que permita a extração de informações estatísticas de forma eficiente e aplicação em modelos de aprendizado de máquina.

Foram estabelecidos os seguintes objetivos específicos:

- Coletar e estudar dados disponibilizados no Repositório do Covid-19 DataSharing/BR;
- Propor uma estruturação dos dados em um formato considerado organizado, *tidy data* de Wickham (2014);
- Aplicar os dados organizados em modelos de aprendizado de máquina;
- Avaliar os modelos gerados e apontar se os dados estão apropriados.

1.2 Motivação

A ciência de dados é uma disciplina que abrange inúmeras áreas. Um dos campos de aplicação é a área da saúde, que adota técnicas diversas para contribuir na tomada de decisões, tanto em nível clínico, quanto em nível de gestão. Dessa forma é relevante compreender os principais desafios que este setor enfrenta no que diz respeito ao uso das abordagens da ciência de dados, especialmente no que diz respeito à coleta e utilização de dados.

Vale ressaltar que a área de ciências de dados visa uma gestão eficiente desses dados com o objetivo de produzir informações e conhecimento que possam dar suporte para a tomada de decisões (PROVOST; FAWCETT, 2013; VEAUX *et al.*, 2017). Entre as principais ações, ilustradas na Figura 1.1, destacam-se as seguintes:

- Projeto e coleta de dados, com atividades de criação, captura, integração e armazenamento;
- Limpeza dos dados, de maneira a aumentar sua qualidade e tratar inconsistências;
- Uso dos dados, por meio de análises, visualizações, produções de modelos, entre outros.

FIGURA 1.1 – Principais tarefas realizadas na área de ciências dos dados.
Fonte: Provost e Fawcett (2013).

Considerando o exposto anteriormente, as bases utilizadas neste estudo, que compreendem o diagnóstico e prognóstico de Covid-19, apresentam inúmeros problemas que impossibilitam o uso adequado dos respectivos dados, entre os quais pode-se destacar a grande quantidade de valores ausentes, presença de *outliers*, formato inadequado, entre outros, justificando, portanto, a realização de um pré-processamento para tornar os dados adequados para serem utilizados em tarefas de aprendizado de máquina.

Outro aspecto relevante diz respeito ao manejo adequado do paciente e uso eficiente dos recursos hospitalares. Uma vez que as técnicas de AM podem contribuir com diagnóstico e prognóstico de pacientes com Covid-19, estas informações podem colaborar na identificação de melhores alternativas para as instituições de saúde, colaborando com um processo de tomada de decisão eficaz.

Nessa linha, estudos corroboram com o fato de que é preciso diversas etapas de pré-processamento para tornar os dados adequados, como limpeza, tratamento de dados ausentes e dados enviesados. Além disso, notou-se que não há uma padronização básica nos atributos selecionados, como nomenclatura e padronização de caracteres. Os trabalhos também indicam a necessidade de tarefas de agrupamento de dados, identificação de atributos mais relevantes e *outliers*, como apontam os trabalhos de Zoabi *et al.* (2021), Kukar *et al.* (2021) e Podder *et al.* (2021), destacando a importância de um procedimento de coleta e organização dos dados bem estabelecido.

1.3 Organização do trabalho

Este trabalho está divido da seguinte forma: o Capítulo 2 traz uma breve revisão sobre aprendizado de máquina. São apresentados conceitos fundamentais com destaque para aprendizado supervisionado e não supervisionado. No Capítulo 3, são abordados conceitos de organização de dados e sua importância para tarefas de AM. Neste tópico aborda-se como estruturar um dataset para permitir o uso eficiente dos dados em análises estatísticas e aplicação em modelos de AM. O Capítulo 4 relata os resultados de estudos voltados para o diagnóstico e prognóstico de Covid-19 usando modelos de AM. Este Capítulo aponta as técnicas aplicadas e descreve aspectos voltados para a coleta e tratamento dos dados. O capítulo 5 descreve a metodologia utilizada neste estudo. Já o Capítulo 6 apresenta os resultados obtidos a partir das análises realizadas com os dados do Laboratório Fleury e com os dados do Hospital Sírio-Libanês. Por fim, o Capítulo 7 traz as conclusões e discussão sobre trabalhos futuros.

2 Aprendizado de Máquina

Este capítulo apresenta conceitos fundamentais da área de aprendizado de máquina, descrevendo principalmente em que consiste o aprendizado de máquina e as diferenças entre aprendizado supervisionado e não supervisionado, que são abordagens clássicas e de grande importância para a área. Nesse contexto, este estudo foca em aprendizado supervisionado, enquanto o aprendizado não supervisionado pode ser aplicado em trabalhos futuros a partir dos resultados obtidos nesse estudo, como exposto nos próximos tópicos.

2.1 Fundamentos do aprendizado de máquina

Um dos conceitos fundamentais quando se fala em Inteligência Artificial é aprendizado de máquina (AM). Pode-se definir inteligência artificial (IA) como o processo de simulação da inteligência humana por meio de máquinas e sistemas computacionais especiais, que incluem aprendizado, raciocínio e autocorreção (EANEFF *et al.*, 2020).

Já o aprendizado de máquina trata-se de subcategoria da IA. Enquanto a IA compreende a criação de máquinas que visam imitar os humanos, o AM ensina as máquinas a aprender a partir de conjuntos de dados, sem a ajuda explícita de humanos. O aprendizado de máquina utiliza algoritmos projetados para aprender ao longo do tempo por meio de parâmetros definidos e sistemas de recompensa, melhorando em tarefas específicas.

Este é um campo de pesquisa vasto, onde as máquinas apresentam habilidades cognitivas como comportamentos de aprendizagem, interação proativa com o meio ambiente, inferência e dedução, visão computacional, reconhecimento de voz, resolução de problemas, representação do conhecimento, percepção e muitos outros. De forma resumida, a IA vê qualquer atividade em que as máquinas imitam comportamentos inteligentes normalmente exibidos por humanos. Além disso, a IA é inspirada em elementos de computação, matemática e estatística (FACELI *et al.*, 2011).

No AM, há três abordagens principais de aprendizado: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço.

O aprendizado supervisionado geralmente inicia-se com um conjunto estabelecido de

dados e alguma compreensão de como esses dados são classificados, como por exemplo, se são fornecidas informações abundantes sobre imagens de animais (cães e gatos, por exemplo), e se foi rotulada cada imagem como um cão ou gato, de modo que o sistema aprenderá a identificar um gato ou um cão em qualquer imagem diferente daquela com a qual foi treinado (BZDOK *et al.*, 2018).

Em relação ao aprendizado não supervisionado, este é mais adequado quando o problema requer conjuntos de dados não rotulados. Por exemplo, são fornecidas muitas informações sobre imagens de gatos e cachorros, mas não é informado ao sistema que eles são gatos ou cachorros. Logo é necessário usar algoritmos que possam "entender" as imagens e agrupá-las corretamente em cães ou gatos (FISHER *et al.*, 2014).

Já a aprendizagem por reforço é um modelo de aprendizagem comportamental onde ocorre o treinamento de modelos de AM para tomar uma sequência de decisões. O modelo utiliza tentativa e erro para encontrar a solução para o problema, de modo que recebe recompensas quando se chega mais próxima da meta, ou é penalizado quando comete erros (MAZYAVKINA *et al.*, 2021).

2.2 Aprendizado supervisionado

Como exposto anteriormente, o aprendizado supervisionado é projetado para encontrar padrões nos dados que correspondem a um rótulo que define o significado dos dados. Por exemplo, pode haver milhões de fotos de animais e incluir uma explicação sobre o que cada animal é e, em seguida, pode-se criar um aplicativo de aprendizado de máquina que distingue um animal de outro (BZDOK *et al.*, 2018). Outro exemplo de aprendizado supervisionado é ilustrado na Figura 2.1.

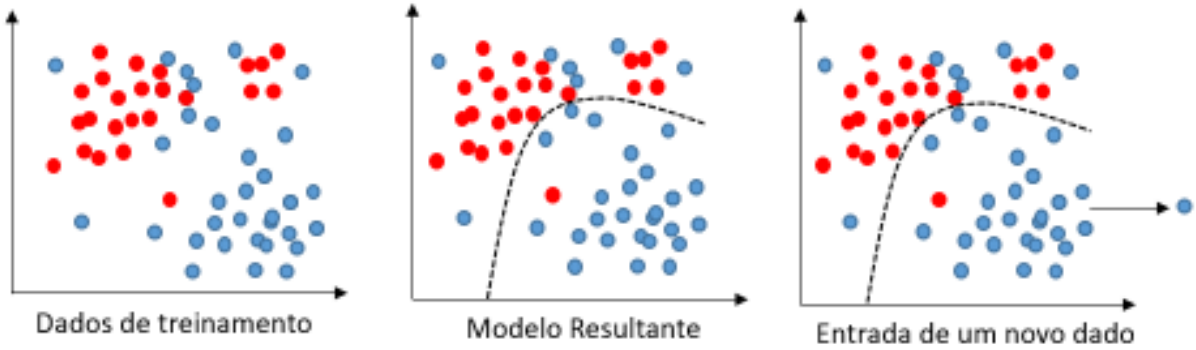


FIGURA 2.1 – Aprendizado supervisionado: cada exemplo de treinamento tem um rótulo. O modelo aprende um limite de decisão e atribui rótulos a novos dados.

Fonte: Adaptado Langs *et al.* (2018).

Segundo Bzdok *et al.* (2018), existem dois tipos principais de problemas de aprendizado de máquina supervisionado, chamados de classificação e regressão.

Quando o rótulo é quantitativo, é uma regressão; e quando os dados vêm de um conjunto finito de valores, é conhecido como classificação. Em essência, o uso de regressão para aprendizado supervisionado ajuda a entender a correlação entre as variáveis (ZHOU, 2018). Um exemplo de aprendizagem supervisionada, usando análise de regressão, é a previsão do clima. A previsão do clima leva em consideração os padrões meteorológicos históricos conhecidos e as condições atuais para fornecer uma previsão sobre o clima. Outro exemplo de aprendizado supervisionado, usando classificação, seria a classificação dos animais (SABA *et al.*, 2017).

Os algoritmos são treinados usando exemplos pré-processados, e neste ponto, o desempenho dos algoritmos é avaliado com dados de validação do próprio conjunto. Vale destacar que pré-processamento consiste em uma série de etapas que tem como finalidade preparar, organizar e estruturar os dados, enquanto validação refere-se ao uso de parte dos dados para avaliar a capacidade de generalização dos modelos. Ocasionalmente, os padrões identificados em um subconjunto de dados não podem ser detectados na população maior destes dados. Se o modelo for capaz de representar apenas os padrões existentes no subconjunto de treinamento, será criado um problema chamado *overfitting* (YING, 2019).

Overfitting significa que o modelo está super ajustado para os dados de treinamento, mas pode não ser aplicável para conjuntos de dados diferentes do treinado (YING, 2019). Para reduzir as chances contra *overfitting*, as validações devem ser realizadas em dados desconhecidos. Usar dados desconhecidos no conjunto de validação pode ajudar a avaliar a precisão do modelo na previsão de resultados.

Os modelos de aprendizagem supervisionada têm ampla aplicabilidade a uma variedade de problemas, incluindo detecção de identidade, soluções de recomendação, reconhecimento de voz ou análise de risco, entre outros (COGSWELL *et al.*, 2015).

É importante apontar que os modelos preditivos, como descreve a Figura 2.2, são voltados para a realização de previsões, que consiste em encontrar uma função, hipótese ou modelos que possa ser utilizado em alguma tarefa de predição (SARKER, 2021). Além disso, Candanedo *et al.* (2018) aponta que o objetivo de um modelo preditivo é minimizar o erro entre o valor real e o valor previsto, considerando todos os possíveis fatores de interferência.

As observações individuais são frequentemente caracterizadas por um conjunto de atributos quantificáveis. Essas propriedades podem ser categóricas, ordinais, valores numéricos inteiros ou reais (BAŞTANLAR; ÖZUYSAL, 2014).

De acordo com Khan *et al.* (2020), ao construir um classificador, deve-se escolher o método de classificação e a amostra de dados a ser tratada. Focando nos dados, e por se tratar de aprendizagem supervisionada, para cada conjunto de características da amostra se deve saber qual é a classe correta correspondente. Para fazer isso, sugere-se escolher

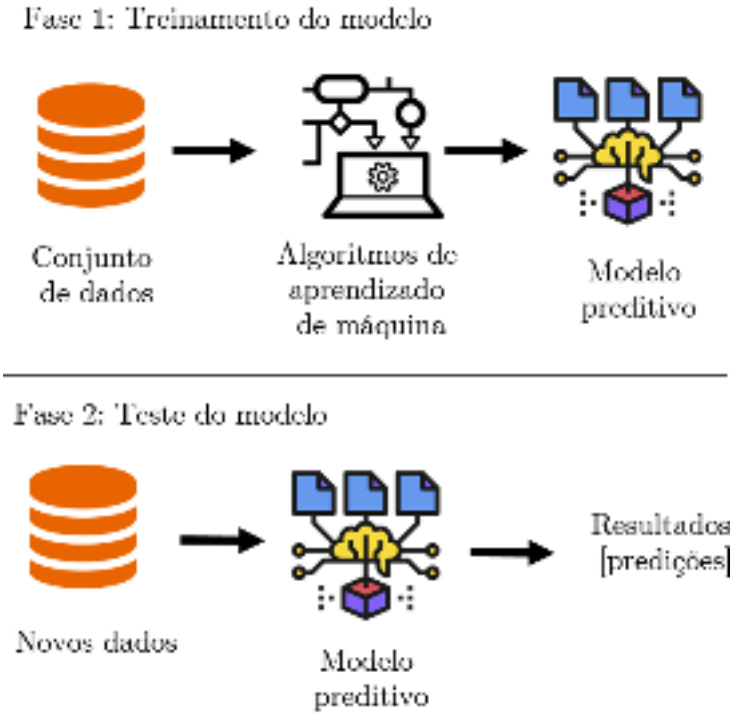


FIGURA 2.2 – Aprendizado preditivo.
Fonte: Adaptado Sarker (2021).

primeiro o método de classificação. Depois é necessário uma amostra de dados, onde todos os valores das classes são conhecidos. Esses dados podem ser divididos em dois conjuntos: treinamento e teste.

O conjunto de treinamento representa a entrada para o algoritmo de treinamento, que resulta em um classificador. Em seguida, o classificador é testado em relação aos dados de teste, onde os valores das classes não são conhecidos. Se o classificador classificar a maioria dos casos corretamente, pode-se supor que ele funcione com precisão também com os novos dados, ou seja, ele generaliza bem. Pelo contrário, se fizer muitas classificações erradas com os dados de teste, pode-se admitir que é o modelo com baixo desempenho (WEI; JR, 2013).

Dessa forma, considerando o exposto ao longo deste tópico, destaca-se que este estudo utiliza aprendizado de máquina supervisionado, uma vez que são utilizados dados rotulados para tratar um problema de classificação, que neste caso, considera o diagnóstico e prognósticos de pacientes no contexto da Covid-19.

Além disso, em trabalhos futuros pode-se aplicar técnicas de aprendizado não supervisionado nos conjuntos de dados gerados a partir deste estudo, uma vez que este é um campo de estudo com grande potencial de analisar os dados de outras formas além daquelas realizadas através de técnicas de AM supervisionado.

2.3 Aprendizado não supervisionado

No aprendizado não supervisionado os dados não são rotulados e os algoritmos trabalham para tentar aprender com estas informações. Para isso, os diferentes algoritmos buscam padrões ou características para agrupar os dados (USAMA *et al.*, 2019). Langs *et al.,* (2018) relata que é utilizado principalmente para redução de dimensionalidade, agrupamento, detecção de anomalias. Um exemplo de de uso desta técnicas pode ser verificado na Figura 2.3.

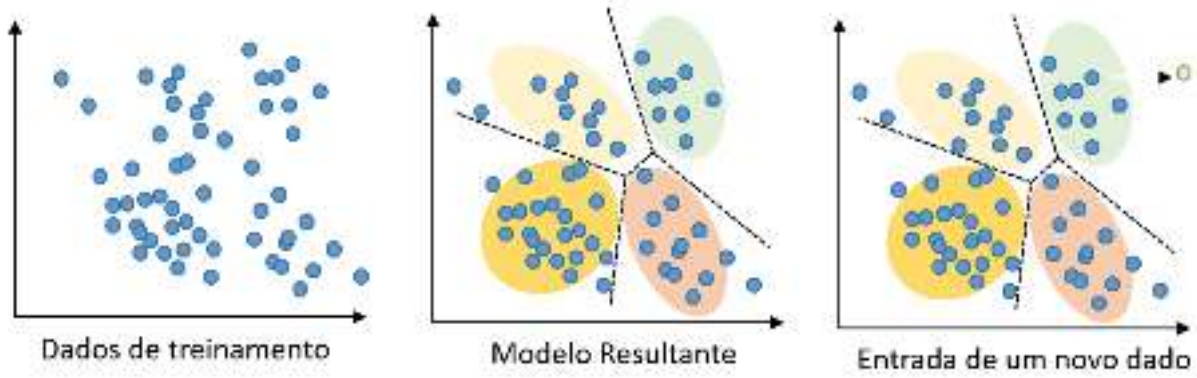


FIGURA 2.3 – Aprendizado não supervisionado: os exemplos de treinamento não possuem rótulos. O modelo identifica a estrutura como agrupamento. Novos dados podem ser atribuídos ao agrupamento.

Fonte: Adaptado Langs *et al.,* (2018).

O modelo de AM não supervisionado não recebe nenhuma informação do ambiente que diga se a saída gerada em resposta a uma determinada entrada está correta ou não. Esses algoritmos devem encontrar as características, regularidades, correlações ou categorias que podem ser estabelecidas entre os dados apresentados em sua entrada (ALSHEIKH *et al.*, 2014).

Existem várias possibilidades quanto à interpretação da saída dessas modelos, que dependem de sua estrutura e do algoritmo de aprendizado utilizado. Em alguns casos, a saída representa o grau de familiaridade ou semelhança entre as informações que estão sendo apresentadas na entrada e as informações que foram mostradas até agora (no passado).

Em outro caso, poderia realizar um agrupamento (*clustering*) ou estabelecimento de categorias, indicando na saída a qual categoria pertence a informação apresentada na entrada, sendo o próprio modelo que deve encontrar as categorias apropriadas a partir das correlações entre as informações apresentado.

2.4 Considerações finais

As técnicas de aprendizado de máquina têm sido utilizadas em diversos segmentos como forma de contribuir com a tomada de decisões em diferentes cenários. A principal característica do aprendizado de máquina é o fato de que estes algoritmos podem "aprender" com os dados, possibilitando prever comportamentos por meio de técnicas preditivas consolidadas.

Exitem ainda diferentes tipos de aprendizado de máquina. Neste estudo foram destacados duas modalidades amplamente utilizadas na literatura científica, sendo estas o aprendizado por não supervisionado e supervisionado.

No aprendizado não supervisionado o algoritmo tenta aprender com dados não rotulados. Neste tipo de aprendizagem, o modelo não consegue etiquetar os objetos, dessa forma, tenta agrupá-los de acordo com suas características. Para isso, são identificados padrões e semelhanças em dados não rotulados.

O aprendizado supervisionado é utilizado quando se tem um problema conhecido e dados rotulados. Dessa forma, o algoritmo aprende de forma interativa com os dados para permitir que informações sejam encontradas sem necessariamente ser programados em onde procurar.

Para os modelos supervisionados, destacam-se os modelos preditivos. Estes modelos permitem que uma previsão seja feita a partir de um modelo treinado com dados devidamente rotulados. Sendo assim, é possível inferir, por exemplo, se uma pessoa tende a ficar doente ou não a partir de um conjunto de dados de novos pacientes.

Também é importante destacar que não há uma técnica melhor que a outra. Cada modelo de aprendizagem tem características singulares e sua aplicação se dá de acordo com o problema tratado e, principalmente, de acordo com os tipos de dados utilizados.

Nessa linha, o formato e tipo dos dados de entrada são essenciais para que os modelos tenham sucessos em suas tarefas. Os dados devem permitir a extração de estatísticas e padrões, além de estarem organizados em configuração que permita seu uso nas tarefa de AM.

3 Pré-processamento de dados

Atualmente, um dos maiores desafios ao utilizar técnicas de aprendizado de máquina supervisionado continua sendo trabalhar com conjuntos de dados brutos. A extensa maioria dos processos de coleta de dados são não supervisionados, gerando ao final informações redundantes, dados com ruídos ou mesmo atributos sem relevância. Portanto, o pré-processamento dos dados consiste em uma etapa importante na extração dos dados, que pode reduzir estes problemas e gerar *datasets* mais aprimorados.

Nesta seção são discutidos conceitos que definem o que é um conjunto de dados e a importância de organizar um respectivo conjunto de formas distintas com o objetivo de tornar as análises mais eficientes.

3.1 Características do pré-processamento de dados

O pré-processamento de dados aborda diversas etapas que tem como finalidade obter um conjunto de dados limpo e que esteja adequado para ser utilizado em análises estatísticas e aprendizado de máquina. Neste tópico são abordados os principais conceitos em pré-processamento de dados.

Conforme aponta Obaid *et al.* (2019), uma das primeiras atividades no pré-processamento trata-se da integração dos dados. A integração refere-se a um processo de fusão de dados de origens distintas e diferentes. Este processo deve ser feito com cautela para evitar a ocorrência de redundância ou levar à inconsistências no dataset final. As operações mais comuns realizadas na integração de dados são: identificar e unificar variáveis e domínios, analisar a correlação de variáveis, detectar conflitos em valores de dados de diferentes fontes.

Outra atividade é a transformação de dados. Nesta fase do pré-processamento, os dados são transformados ou consolidados para que se possa utilizar o conjunto de dados de uma forma mais eficiente. Na etapa de transformação, as tarefas mais comuns são a homogeneização, a discretização, a generalização e a normalização dos dados (KUMARI; KUMAR, 2015).

Como o número de tarefas pode ser extenso dependendo do conjunto de dados utilizados, estas tarefas são realizadas separadamente. Além disso, destaca-se que as tarefas clássicas são aquelas que requerem supervisão humana, por exemplo, geração de relatórios, agregar novos atributos aos existentes e generalização de dados, especialmente em atributos categóricos. Por generalização entende-se a possibilidade de subdividir ou agrupar variáveis em função de suas características.

A limpeza de dados também compreende uma das etapas mais importantes do pré-processamento, conforme aponta Chu *et al.,* (2016). A limpeza de dados é o processo de identificar partes incompletas, incorretas, imprecisas, irrelevantes ou dados ausentes e, em seguida, alterar, substituir ou excluir essa parte conforme necessário. A limpeza de dados é um elemento fundamental no pré-processamento de dados. Este campo de atuação inclui o tratamento de dados ausentes, a detecção de anomalias e dados "sujos" (fragmentos dos dados originais que não fazem sentido), bem como o tratamento de ruídos.

A presença de ruído nos dados é um problema comum que produz diversas consequências negativas em problemas de indução. O ruído é um problema inevitável, afetando os processos de coleta e preparação de dados em aplicações de mineração de dados, onde erros comumente ocorrem. O desempenho dos modelos de aprendizagem construídos em tais circunstâncias dependerá em grande parte da qualidade dos dados de treinamento, mas também da robustez contra ruídos do próprio modelo (ILYAS; CHU, 2019). Portanto, problemas contendo ruído são problemas complexos e soluções bem-sucedidas são muitas vezes difíceis de alcançar sem o uso de técnicas especializadas.

Outra tarefa é a redução dos dados. A redução de dados compreende o conjunto de técnicas que, de uma forma ou de outra, obtêm uma representação reduzida dos dados originais, mantendo a estrutura e integridade essenciais dos dados originais. A redução de dados é uma etapa opcional, porém, os algoritmos utilizados na mineração de dados possuem tempos de execução que dependem de diversos parâmetros e alguns desses parâmetros geralmente são proporcionais ao tamanho do banco de dados de entrada. Se o referido tamanho for excessivo, o funcionamento do algoritmo pode ser proibitivo e, portanto, a tarefa de redução de dados pode se tornar tão crucial quanto a fase de preparação de dados (REDDY *et al.*, 2020). Quanto a outros fatores, como reduzir a complexidade e melhorar a qualidade dos modelos produzidos, o papel da redução de dados é igualmente decisivo.

De forma geral, são diversas as tarefas presentes no pré-processamento de dados. Cada uma delas tem uma finalidade específica que busca resolver algum problema que afeta os conjuntos de dados. Como destacado, essas tarefas, em sua maioria, são supervisionadas, o que demanda força de trabalho profissional e tempo, podendo afetar os projetos em ciências de dados, tanto no aspecto prático quanto econômico.

Além do pré-processamento, destaca-se neste trabalho outra metodologia que envolve a organização dos dados, conhecido como "Tidy data". Esse conceito é abordado no tópico a seguir e indica que um conjunto de dados pode ser organizado de diferentes formas de acordo com a necessidade do projeto.

3.2 Dados arrumados

Em Aprendizado de Máquina, um conceito importante refere-se à organização dos dados. Um problema que a análise de dados compartilha é ter dados de entrada arrumados, bem como saídas das diferentes funções que os analisam, que são igualmente organizadas, facilmente interpretáveis e que podem ser uma entrada organizada para um procedimento subsequente. O termo usado na literatura de Ciência de Dados é *tidy data* (WICKHAM, 2014). Neste trabalho, utilizaremos a tradução livre *dados arrumados*.

Os conjuntos de dados são caracterizados pela forma como estão organizados. O objetivo de se ter dados organizados é fornecer um método lógico que permita manipular esses dados em função do problema a ser resolvido e do algoritmo que vai resolvê-lo.

Nessa linha, Wickham (2014) traz em seu estudo uma análise sobre um processo básico mas pouco discutido na literatura, que é a organização dos dados. Há poucos estudos que abordam os processo de limpeza e preparação de *dataset*, mesmo sendo essa etapa uma das mais demoradas. Para entender o conceito de organização de dados, é necessário compreender alguns conceitos básicos.

A grande parte dos conjuntos de dados possuem uma configuração tabular. Isso significa que estes dados são representados em uma tabela formada por linhas e colunas. No entanto, estes dados podem ser representados de formas distintas, como mostra a Figura 3.1.

Um conjunto de dados é, portanto, uma coleção de valores que podem ser numéricos ou textuais. Cada valor pertence à uma variável (coluna) e a uma observação (linha). A variável refere-se à uma medida com uma unidade específica, por exemplo, comprimento (metros), temperatura, (°C), tempo (s). Cada observação, que são as linhas, é uma unidade daquilo que está medido, como por exemplo, dias ou pessoas.

Nesse contexto, a Figura 3.2 propõe uma reorganização dos dados apresentados na representação típica presente na Figura 3.1 para tornar os valores, variáveis e observações mais claros e coesos. Neste caso, são utilizadas três variáveis: nomes (*name*), tratamento (*treatment*) e resultados (*result*).

Embora seja simples apontar quais são as variáveis e observações para uma tabela específica, é difícil fazer uma definição geral. A proposta de Wickham (2014) de da-

Tabela I

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Representação típica de um conjunto de dados

Tabela II

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Mesmos dados da Tabela I acima, mas em uma estrutura diferente

FIGURA 3.1 – Representações distintas para um mesmo conjunto de dados.
Fonte: Wickham (2014).

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

FIGURA 3.2 – Os mesmos dados apresentados na Figura 3.1, mas com variáveis em colunas e observações em linhas.
Fonte: Wickham (2014).

dos arrumado é justamente fixar um padrão que relacione a estrutura dos dados e seus significados. Esse padrão é simples, segue apenas três regras fundamentais:

1. Cada variável é uma coluna;
2. Cada observação é uma linha;
3. Cada tipo de unidade de observação é uma tabela.

Sendo assim, observa-se que conjuntos de dados podem assumir diferentes organizações que podem ser mais adequadas ou não com base no contexto em que estes dados serão utilizados. Além disso, dependo do conjunto de dados utilizados, organizar estes dados

pode ser uma tarefa árdua, já que alguns *datasets* podem ter dezenas de colunas e milhões de linhas.

Outro desafio encontrado quando se trabalha com conjunto de dados, é que muitas vezes estes conjuntos podem ser *bagunçados* (do inglês, *messy*) (WICKHAM, 2014). Conjuntos de dados reais podem, e muitas vezes ocorrem, violar os três preceitos de dados arrumados em quase todas as formas imagináveis. Embora ocasionalmente se obtenha um conjunto de dados que possibilita analisá-lo imediatamente, essa é a exceção, não a regra. Wickham (2014) descreve então os cinco problemas mais comuns com conjuntos de dados bagunçados, juntamente com o procedimento de arrumá-los:

1. Os cabeçalhos das colunas são valores, não nomes de variáveis;
2. Várias variáveis são armazenadas em uma mesma coluna;
3. As variáveis são armazenadas em linhas e colunas;
4. Vários tipos de unidades observacionais são armazenados na mesma tabela;
5. Uma única unidade observacional é armazenada em várias tabelas.

Em relação ao primeiro problema, um tipo comum de conjunto de dados bagunçados são os dados tabulares projetados para apresentação, em que as variáveis formam as linhas e as colunas, e os cabeçalhos das colunas são valores, não nomes de variáveis. Embora Wickham chame essa configuração de bagunçada, em alguns casos pode ser extremamente útil. Contudo, Wickham (2014) descreve que ao tratar os cabeçalhos como nomes, pode-se arrumar o dataset de forma que em uma única coluna tenha apenas um mesmo tipo de variável (WICKHAM, 2014). Por exemplo, na coluna "Exames" só haverá nomes de exames e na colunas "Valores" haverá apenas os resultados em formato numérico para esses exames. Isso evita que se tenha, por exemplo, inúmeros colunas para vários tipos de exames, por exemplo.

Um exemplo genérico do primeiro problema pode ser um conjunto de dados onde as colunas representam doenças distintas e as linhas indicam os países e o número de casos, como mostra a Tabela A na Figura 3.3. Logo, uma forma de arrumar esses dados de maneira eficiente é apresentada na Tabela B da Figura 3.3. Para o primeiro formato na Figura 3.3 (Tabela A) acima é necessário um pré-processamento mais cuidadoso, enquanto que o segundo formato (Figura 3.3 -Tabela B) permite, entre outras tarefas, combinar *datasets* mais facilmente, bem como possibilita adicionar novas colunas mais amigavelmente. Também é importante destacar que não há um formato certo ou errado, mas sim, um formato mais adequado para uma aplicação específica.

Tabela A

País	Doença 1	Doença 2	Doença 3	Doença 4	Doença 5
Brasil	32	12	50	41	7
Chile	21	9	26	35	24

Tabela B

País	Doenças	N_casos
Brasil	Doença 1	32
Brasil	Doença 2	12
Brasil	Doença 3	50
Brasil	Doença 4	41
Brasil	Doença 5	7
Chile	Doença 1	21
Chile	Doença 2	9
Chile	Doença 3	26
Chile	Doença 4	35
Chile	Doença 5	24

FIGURA 3.3 – Diferentes arranjos para um mesmo conjunto de dados. Tabela A: Dados em um formato convencional; Tabela B: Dados arrumados.

Seguindo na linha do descrito anteriormente, o segundo problema tratado considera várias variáveis armazenadas em uma coluna. Logo, para este tipo de problema sugere-se que cada coluna armazene apenas um tipo de variável. No caso deste estudo, por exemplo, um atributo que refere-se à um exame clínico, deve armazenar somente variáveis numéricas.

O terceiro erro descrito por Wickham (2014) é o fato das variáveis serem armazenadas em linhas e colunas. De forma geral, o autor relata que as colunas devem ser variáveis, enquanto as linhas são as observações. Dessa forma é possível trabalhar com variáveis isoladas ou múltiplas variáveis mais facilmente, indicando as colunas desejadas. Essa padronização ajuda a reduzir, por exemplo, o número de valores ausentes em conjunto de dados.

O próximo erro diz respeito ao uso de diferentes unidades de medidas em uma única tabela. Os conjuntos de dados geralmente envolvem valores coletados em vários níveis, em diferentes tipos de unidades observacionais. Durante a arrumação, cada tipo de unidade observacional deve ser armazenado em sua própria tabela. Isso está intimamente relacionado à ideia de normalização do banco de dados, onde cada fato é expresso em apenas um lugar. Se isso não for feito, é possível que ocorram inconsistências (WICKHAM, 2014).

O quinto erro descrito por Wickham (2014) se refere a uma única unidade observacional estar em várias tabelas. Também é comum encontrar valores de dados sobre um único tipo de unidade observacional espalhados por várias tabelas ou arquivos. Essas tabelas e arquivos geralmente são divididos por outra variável, de modo que cada uma represente um único ano, pessoa ou local. Desde que o formato dos registros individuais seja consistente, este é um problema fácil de corrigir. Uma situação mais complicada ocorre quando a estrutura do conjunto de dados muda ao longo do tempo. Por exemplo, os conjuntos de dados podem conter variáveis diferentes, as mesmas variáveis com nomes diferentes,

formatos de arquivo diferentes ou convenções diferentes para valores ausentes. Isso pode exigir que seja necessário arrumar cada arquivo individualmente e só depois os combiná-los.

3.2.1 Formato dos dados

De acordo com Bonaccorso (2017), em todos os problemas de aprendizagem de máquina, seja supervisionado ou não, haverá um conjunto de dados X definido como um número finito n de vetores reais com m características, dado por

$$X = \{\vec{x_1}, \vec{x_2}..., \vec{x_n}\}, \text{ onde } \vec{x_i} \in \mathbb{R}^m. \tag{3.1}$$

Considerando uma abordagem probabilística, supõe-se que a probabilidade p de cada $\vec{x_i} \in X$ é dada por uma distribuição estatística multivariada D. Esse processo conhecido como um processo de geração de dados. Também é importante apontar que outra condição importante em um conjunto de dados X: espera-se que todas as amostras sejam independentes e distribuídas de forma idêntica. Isso significa que todos os exemplos $\vec{x_i}$ pertencem a mesma distribuição $\vec{x_i} \sim D$. Sendo assim, considerando um subconjunto de k observações, tem-se

$$p\left(\vec{x_1}, \vec{x_2}, ...\vec{x_k}\right) = \prod_{i=1}^{k} p\left(\vec{x_i}\right). \tag{3.2}$$

É importante ressaltar que todas as áreas de AM considera o uso de *datasets* com distribuições bem definidas e o conjunto dos dados reais são amostras extraídas deste.

Anteriormente foi apontado que o conceito de aprendizado de máquina considera entre um agente e uma situação desconhecida. Isso é possível devido à capacidade de aprender uma representação da distribuição e não do próprio conjunto de dados. Portanto, a partir de agora, sempre que um conjunto de dados finito for usado, deve-se sempre considerar lidar com novas amostras que compartilham a mesma distribuição.

3.3 Considerações finais

Este capítulo teve como base principal o artigo chamado “Tidy data”, publicado por Wickham (2014). Os principais conceitos apresentados consideram uma organização dos dados de um formato estruturado, onde as colunas são as variáveis e as linhas são as observações preenchidos por valores definidos.

A estruturação dos dados é uma tarefa importante por ser uma etapa demorada e que converte dados brutos em dados utilizáveis, ou seja, em um formato que permita seu

uso eficiente. É nesta etapa que também se decide quais dados são relevantes e quais podem ser descartados. Isso é feito a partir de um processo de limpeza e compreensão das variáveis que fazem sentido usar no problema investigado.

Também é importante lembrar que dados ideais para uso em tarefas de aprendizado de máquina, que no caso deste estudo consiste em aprendizado supervisionado, devem ter uma distribuição estatística multivariada bem definida que permita a identificação de padrões com maior clareza. Sendo assim, dados com uma grande frequências *outliers*, por exemplo, podem representar um desafio.

Diversos fatores podem incidir sobre a ocorrência de problemas em conjunto de dados, como o método utilizado na coleta e em como estes dados são inseridos nas bases. Diante disso, aponta-se ainda mais para a necessidade de se determinar um protocolo que determine uma forma organizada de coletar estes dados em um formato adequado e permita seu uso de forma eficiente.

4 Aprendizado de máquina aplicado à Covid-19

O uso de ciências de dados e técnicas de AM têm sido cada vez mais comuns no campo da saúde. Neste cenário, as interações entre profissional de saúde e paciente são informadas e apoiadas por grandes volumes de dados originados a partir de interações com pacientes semelhantes. Estes dados são coletados e selecionados com o objetivo de fornecer avaliações e recomendações baseadas em evidências (RAJKOMAR *et al.*, 2019).

Desde o surgimento da pandemia de SARS-Cov-2, inúmeros estudos têm relatado o uso de técnicas que possam contribuir para o combate da pandemia do Coronavírus, entre as quais destacam-se as técnicas de aprendizado de máquina utilizadas no auxílio diagnóstico ou prognóstico de Covid-19 (ALBALLA; AL-TURAIKI, 2021; FERNANDES *et al.*, 2021).

4.1 Aprendizado de máquina aplicado ao diagnóstico de Covid-19

Diagnóstico, em medicina, consiste na análise de exames clínicos, ou de uma condição, com o objetivo de indicar uma conclusão. Está conclusão pode ser uma doença ou condição de saúde (SCHAFFNER, 2021). Em relação ao diagnóstico, Moraes *et al.* (2020) utilizaram técnicas de aprendizado de máquina na predição da Covid-19 utilizando dados de exames clínicos e características como sexo e idade. Na pesquisa foram utilizados dados de 235 pacientes adultos, com uma média de idade de 49 anos, atendidos no Hospital Israelita Albert Einstein em São Paulo em março de 2020.

Foram utilizados cinco tipos de algoritmos de aprendizado de máquina: redes neurais, floresta aleatória (RF), árvore de aumento gradiente, regressão logística (LR) e máquinas de vetores de suporte (SVM). As variáveis mais importantes para a predição foram aquelas relacionadas à contagem de: linfócitos, leucócitos e eosinófilos. Os resultados obtidos pelos autores são apresentados na Figura 4.1.

É possível notar que os algoritmos tiveram um comportamento semelhante e que os

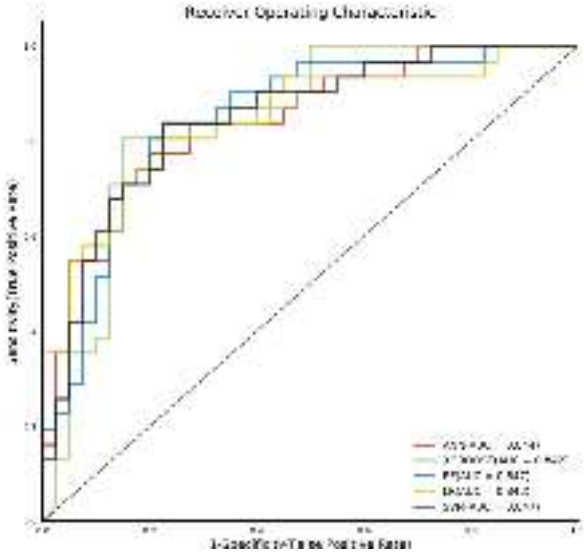


FIGURA 4.1 – Curva ROC (Receiver Operating Characteristic) para o conjunto de teste de cada um dos cinco algoritmos de aprendizado de máquina no estudo de Moraes *et al.*,(2020).

Fonte: Moraes *et al.* (2020).

algoritmos RF e SVM apresentaram os melhores resultados na predição, como 0,677 de sensibilidade e 0,850 de especificidade.

Kukar *et al.* (2021) também realizaram um estudo de predição de Covid-19 usando exames de sangue de rotina. Os dados dos exames foram coletados Departamento de Doenças Infecciosas do Centro Médico Universitário de Ljubljana (UMCL), Eslovênia. Para a construção do conjunto de dados, os autores coletaram informações de 160 pacientes testados positivos para Coronavírus entre março e abril de 2020.

A amostra de pacientes negativos foi gerada a partir de dados de 52.306 pacientes coletados entre 2012 e 2019. Desse total, retirou-se uma amostra de 22.385 pacientes que apresentaram quadro de infecções virais e bacterianas. Para construir o *dataset* final os pacientes foram amostrados para aproximar a proporção de pacientes testados para Covid-19, retirando uma amostra aleatória de 5.333 pacientes com 225 infecções virais e bacterianas diferentes. Logo, para gerar o modelo foram considerados exames de sangue de 160 casos positivos de Covid-19 e 5.333 negativos (KUKAR *et al.*, 2021).

Para o pré-processamento dos dados, Kukar *et al.* (2021) identificaram 117 parâmetros medidos nos pacientes testados para coronavírus. Foram removidos parâmetros medidos em menos de 25% da amostra positiva, além de omitir parâmetros não sanguíneos e parâmetros sanguíneos arteriais. Assim, ao final foram selecionados 35 parâmetros obtidos a partir de exames de sangue. Todos os valores foram centralizados e escalonados com os intervalos de referências. O comportamentos destes dados podem ser verificados na Figura 4.2.

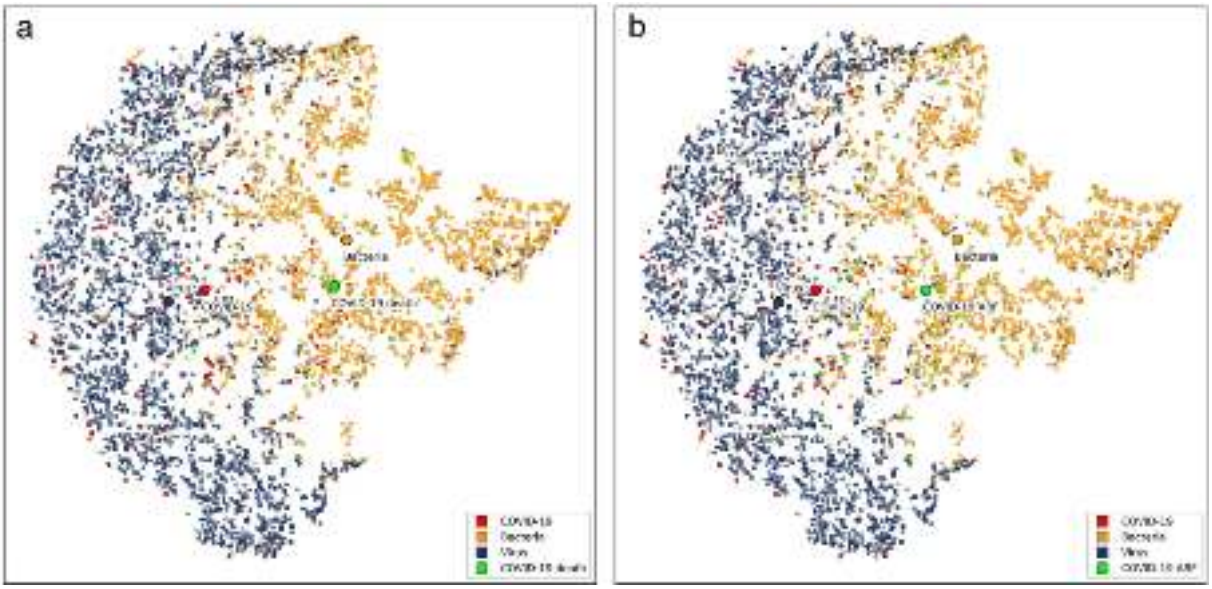


FIGURA 4.2 – Visualização do espaço de parâmetros de bactérias/vírus/Covid-19 com o método t-SNE. Os pontos verdes no painel (a) representam pacientes com Covid-19 que morreram (10 pacientes) e no painel (b) pacientes com Covid-19 diagnosticados com insuficiência respiratória aguda (38 pacientes).

Fonte: Kukar *et al.* (2021).

O modelo preditivo para o diagnóstico de Covid-19 foi gerado usando XGBoost e funcionou de forma satisfatória. Kukar *et al.* (2021) avaliaram sua respectiva abordagem usando teste de validação cruzada estratificada de dez pastas. Os resultados e os intervalos de confiança binomiais correspondentes, calibrados em relação ao ponto ROC operacional foram os seguintes: uma sensibilidade de 81,9% ± 6%, especificidade de 97,9% ± 0,4% e AUC de 0,97.

Seguindo na mesma linha, Zoabi *et al.* (2021) utilizaram aprendizado supervisionado para treinar o modelo com cadastro de 51.831 pessoas, das quais 4.769 foram confirmados com Covid-19. Para o aprendizado Zoabi *et al.* (2021) utilizaram oito variáveis baseadas em sintomas, diferente dos estudos descritos anteriormente: febre, tosse, contato com pessoa infectada, sexo, idade 60+, dor de cabeça, dor de garganta e falta de ar.

Zoabi *et al.* (2021) relatam ainda problemas no dados, como viés e dados ausentes. Os autores identificaram, por exemplo, relatos de cefaleia em 66,2% da amostra, dor de garganta em 92,3% e falta de ar em 92,4% e sintomas com relatos balanceados, que é o caso da tosse com 27,4% e febre com 45,9%. A rotulagem incorreta pode afetar os modelos e levar à uma subnotificação dos dados.

Esse modelo foi capaz de dar a probabilidade de um paciente ser diagnosticado com Covid-19 e prever os resultados do conjunto de testes com alta confiabilidade, como mostra a Figura 4.3 (AUC = 0,90). Contudo, mesmo apesar da eficiência do modelo, os autores relatam a necessidade de dados mais robustos, já que dados auto-relatados de sintomas

estão sempre sujeitos à viés, destacando a necessidade de um modelo de coleta de dados mais eficaz (ZOABI *et al.*, 2021).

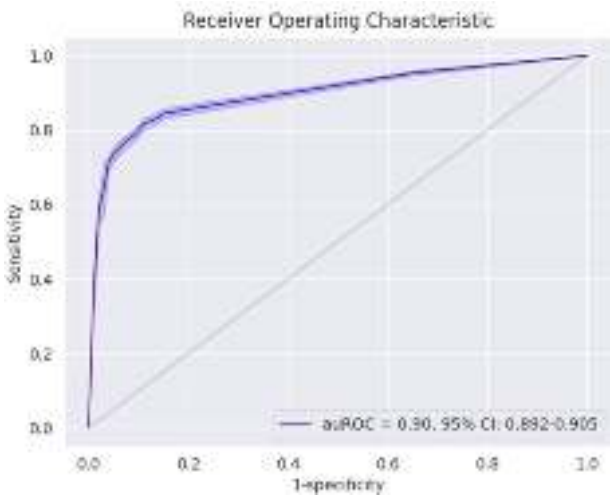


FIGURA 4.3 – Previsão baseada em aprendizado de máquina para o diagnóstico Covid-19 com base nos sintomas.

Fonte: Zoabi *et al.* (2021).

Outro estudo publicado por Podder *et al.* (2021) investigou o diagnóstico de Covid-19 utilizando dados do Hospital Israelita Albert Einstein, no Brasil. Foi utilizado um conjunto de dados contendo 5.644 linhas e 111 colunas que, que segundo os autores, apresenta problemas que que precisaram ser contornados para tornar os dados utilizáveis. Entre os problemas destacados, chamou a atenção a quantidade de valores ausentes na maior parte das colunas, como é possível notar na Figura 4.4.

As barras apresentadas na Figura 4.4 representam a quantidade de valores ausentes no conjunto de dados. É possível notar claramente que a extensa maioria das colunas apresenta mais de 50% dos valores ausentes, que resulta em uma abordagem de exclusão de dados. Os autores determinaram a exclusão de variáveis com mais de 99,8% dos dados faltantes para os pacientes que testaram positivos. Dessa forma, o *dataset* final passou a ser composto por 1.091 linhas e 61 colunas. Foram perdidas, portanto, 50 colunas e 4.553 linhas.

É importante destacar que a exclusão de dados deve ser adotada prioritariamente para informações que não são relevantes para o uso em modelos de AM. Nesse caso, a exclusão se deu em função de valores ausentes, que pode ser um problema ocasionado, por exemplo, a partir do método de coleta desses dados.

Avançando com os resultados apresentados por Podder *et al.* (2021), ou autores relatam que das 61 colunas selecionadas utilizou-se 25 que foram selecionadas a partir de técnicas de seleção de atributos, neste caso por meio da função SelectKBest() da biblioteca scikit-learn. Em seguida foram utilizados cinco algoritmos para a tarefa de diagnóstico

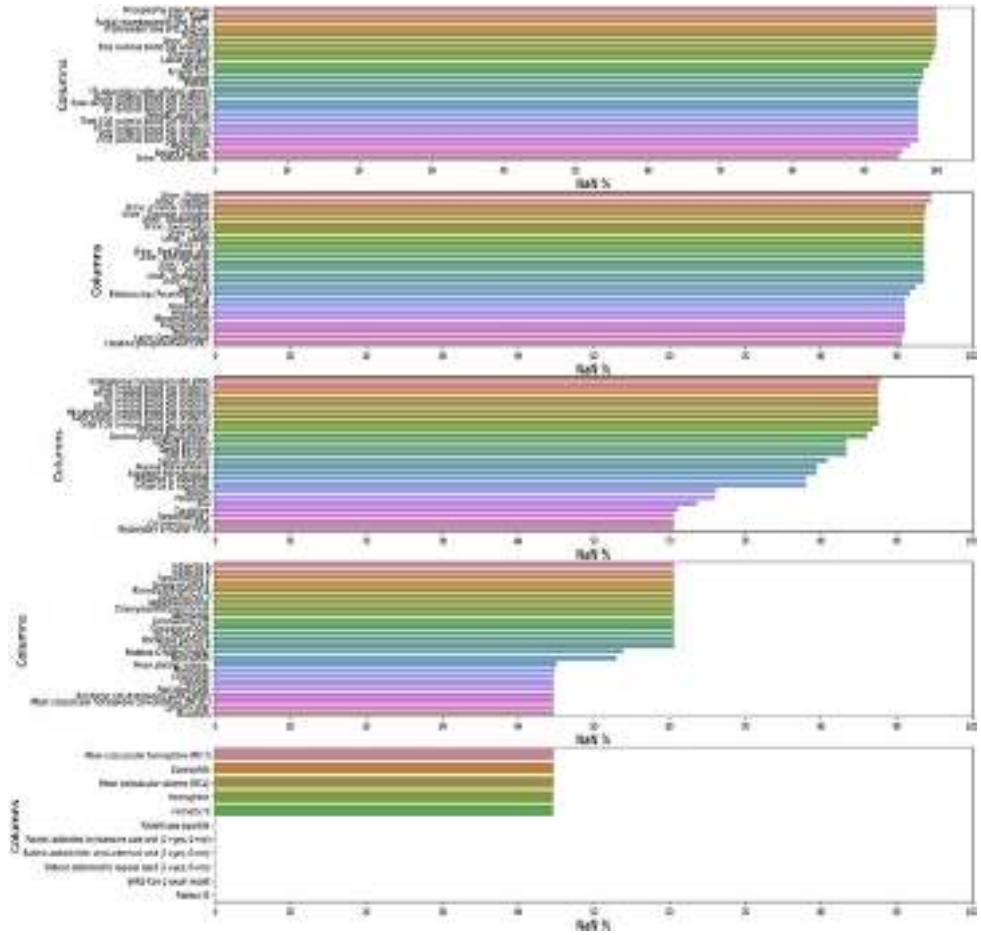


FIGURA 4.4 – Identificação dos valores ausentes no conjunto de dados do Hospital Israelita Albert Einstein.

Fonte: Podder *et al.* (2021).

de Covid-19 através de árvore de decisão (DT), regressão Logística, floresta aleatória e XGBoost. Os resultados apontam que o modelo XGBoost apresentou maiores valores de acurácia (92,67%) enquanto a regressão logística ficou em segundo lugar com acurácia igual a 92,58%, indicando que ambos os modelos podem ser utilizados para prever casos da doença.

4.2 Aprendizado de máquina aplicado ao prognóstico de Covid-19

Além da importância do AM no diagnóstico da doença Sars-Cov-2, destaca-se também o seu potencial na determinação do prognóstico da Covid-19. O prognóstico eficiente permite apontar quais pacientes podem evoluir para um estado mais grave da doença, potencializando um manejo eficaz do paciente e dos recursos disponíveis. Além disso, a observação de fatores que podem agravar a doença pode contribuir na tomada de decisões

pelas instituições de saúde com o objetivo de potencializar o tratamento da Covid-19 e melhor gestão dos recursos hospitalares.

Nessa linha, o estudo de Linssen *et al.* (2020) indica que é possível usar a alteração em dez parâmetros do hemograma completo para prever o agravamento das condições clínicas da Covid-19. Para isso, os autores usaram dados de 982 pacientes adultos, positivos para Covid-19, e realizaram um escore prognóstico para prever durante os primeiros três dias após o atendimento quais pacientes se recuperam sem ventilação mecânica ou se deterioram em um período de duas semanas. A pesquisa obteve uma curva AUC de 0,875.

A pesquisa de Yan *et al.* (2020) mostra o prognóstico da Covid-19 por meio de um modelo para identificar biomarcadores preditivos de mortalidade por doenças, usando dados de 485 pacientes infectados. Foram utilizados neste estudo dados coletados do Hospital Tongji, que recebeu uma maior quantidade de casos graves. Sendo assim, da amostra inicial de 375 pacientes, 174 foram à óbito. Em seguida foram adicionados mais 110 pacientes que foram à óbito ou tiveram alta para uma análise de conjunto de dados externos, como aponta os autores.

As ferramentas de aprendizado de máquina selecionaram três biomarcadores (desidrogenase lática - LDH, linfócitos e proteína C reativa) capazes de prever a mortalidade com mais de 10 dias de antecedência, com precisão de 0,90. Estudos preditivos são importantes, pois podem auxiliar na tomada de decisão sobre quais pacientes devem ser priorizados, visando reduzir a taxa de mortalidade (YAN *et al.*, 2020).

Na mesma linha, Schöning *et al.* (2021), com objetivo de apontar o agravamento da Covid-19, desenvolveram um modelo de pontuação de gravidade da doença (COSA) e compararam com modelos clássicos de aprendizado de máquina. Foram utilizados dados de pacientes suíços que testaram positivo para coronavírus coletados pelo Insel Hospital Group Bern. Os autores destacam ainda que utilizaram duas amostras de pacientes, sendo: uma da primeira onda da doença que compreende o período de 1° de fevereiro a 31 de agosto de 2020 com 198 pacientes e uma segunda amostra de 459 pacientes que foram atendidos entre 1° de setembro e 16 de novembro de 2020.

Foram utilizados dados demográficos, histórico médico e valores laboratoriais obtidos até 3 dias antes ou 1 dia após o teste positivo para prever resultados graves de hospitalização. Em relação aos exames clínicos, os autores selecionaram 20 variáveis correlacionadas positivamente ou negativamente com a Covid-19. Além disso, o conjunto de dados apresentavam 3% de valores ausentes que foram imputados artificialmente usando o algoritmo *k-nearest neighbors* (KNN). Para a realização do prognóstico da doença foram utilizados oito algoritmos que apresentaram desempenho adequado, como mostra a Figura 4.5.

Na Figura 4.5 nota-se um modelo DTI (CART) que refere-se à indução de árvore de

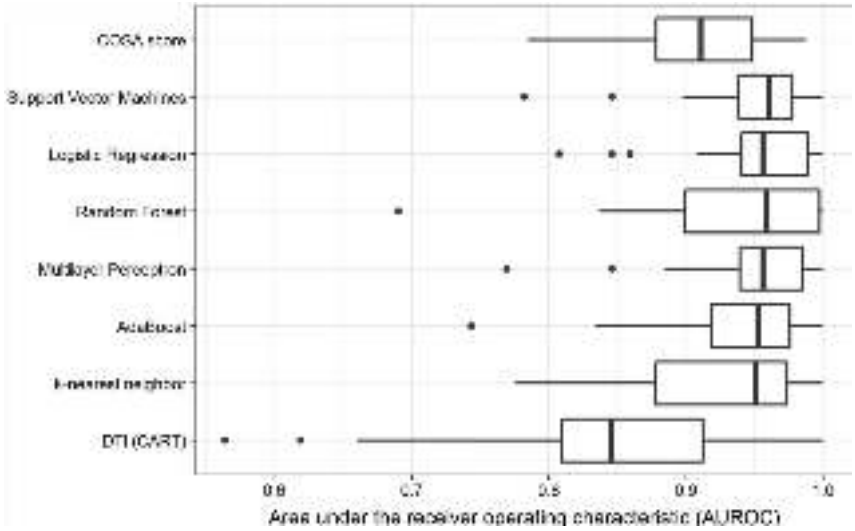


FIGURA 4.5 – Desempenho da pontuação de avaliação de gravidade Covid-19 de estratificação de risco clínico (COSA) e modelos de aprendizado de máquina em pacientes diagnosticados com Sars-Cov-2 no estudo de Schöning *et al.* (2021)

Fonte: Schöning *et al.* (2021)

decisão com árvores de classificação e regressão. Vale destacar que dos oito modelos testados, seis apresentam valor de curva ROC em torno de 0,95, enquanto o pior desempenho foi do algoritmo DTI (CART). Por fim os autores concluem que utilizando parâmetros de fácil obtenção, os modelos preditivos conseguem prever desfechos graves.

O estudo realizado Fernandes *et al.* (2021) teve como objetivo testar se os algoritmos podem generalizar padrões de risco para condições graves para que possam auxiliar no prognóstico de resultados negativos distintos para pacientes com Covid-19. O estudo utilizou uma amostra total de 3.280 pacientes, sendo 1.040 infectados pelo Coronavírus. A amostra de infectados é majoritariamente composta por homens (53,3%). Os autores consideram três estados distintos de gravidade: Admissão na UTI, intubação em ventilação mecânica e óbito. Foram testados cinco algoritmos: redes neurais artificiais, árvores extras, florestas aleatórias, *catboost* e *extreme gradient boosting*.

Os autores relatam que os algoritmos de aprendizado de máquina foram capazes de prever resultados prognósticos negativos com bom desempenho geral para COVID-19, mesmo quando o resultado específico não foi incluído no treinamento dos algoritmos. Todos os modelos apresentaram curva ROC superior a 0,91 (média de 0,92) no conjunto de testes, com alta sensibilidade e especificidade (média de 0,92 e 0,82, respectivamente) (FERNANDES *et al.*, 2021). Os resultados destacam a possibilidade de algoritmos de aprendizado de máquina de alto desempenho serem capazes de prever resultados negativos inespecíficos do Covid-19 usando dados coletados rotineiramente.

4.3 Considerações finais

Assim como na predição do diagnósticos da Covid-19, os algoritmos de aprendizado de máquina se mostraram eficazes também no prognóstico da doença. O uso dessas ferramentas para a indicação de prognóstico visa auxiliar na tomada de decisões pela equipe médica e proporcionar o melhor manejo do paciente e pela minimização dos efeitos da doença.

Alguns desafios podem ser observados nos trabalhos citados, entre os quais destacam-se, por exemplo, a grande quantidade de dados ausentes relatos por Podder *et al.* (2021), que gerou a exclusão de colunas que poderiam ser utilizadas no desenvolvimento dos modelos. Também foi observado que os trabalhos realizaram um pré-processamento dos dados com o objetivo preparar os dataset para uso em AM, visto que estes estavam inadequados para serem utilizados diretamente. Tarefas como limpeza, remoção de dados, estruturação e imputação foram notadas.

É importante frisar que o objetivo desse estudo compreende justamente indicar um método para coletar os dados em um formato arrumado, de modo que elimine, ou reduza, a necessidade de pré-processamento dos dados. Dessa forma, busca-se evitar problemas de grandes volumes de dados faltantes, preenchimento incorreto das informações e inconsistências. Uma vez eliminados estes problemas, diminui-se, por exemplo, a chance de dois atributos iguais estarem em colunas diferentes. A coleta dos dados em um formato arrumado permite seu uso mais imediato pelos especialista em saúde, dispensando diversas etapas no tratamento de dados clínicos dos pacientes. Contudo, ressalta-se que mesmo usando os dados em um formato arrumado, algumas tarefas de pré-processamento ainda serão necessárias

5 Metodologia

Os dados foram coletados a partir do repositório do COVID-19 DataSharing/BR[1]. Este repositório contém cinco conjuntos de dados de instituições de saúde distintas. Foram utilizados nesse trabalhos os dados disponibilizados pelo Laboratório Fleury[2] atualizados em junho de 2021, e também o dados disponibilizados pelo Hospital Sírio-Libanês[3] também atualizados pela última vez em junho de 2021.

O pré-processamento foi feito utilizando a ferramento Colab[4]. Foram utilizadas as bibliotecas NumPy (HARRIS *et al.*, 2020), Pandas (MCKINNEY *et al.*, 2010), Matplotlib (HUNTER, 2007) e Seaborn (WASKOM, 2021). Para o aprendizado de máquina utilizou-se a biblioteca Scikit-learn (PEDREGOSA *et al.*, 2011).

5.1 Pré-Processamentos dos dados para o laboratório Fleury

O pré-processamento foi feito usando a linguagem de programação Python. Foram utilizados os pacotes Pandas, Numpy e Matplotlib. A primeira ação do etapa foi juntar *datasets* com informações distintas, formando ao final um único conjunto de dados mais completo. Para isso, foi feita a união dos conjuntos de dados `EXAMES`, que contém informações sobre os exames, analitos e valores obtidos, e o conjuntos de dados `PACIENTES`, composto por informações como sexo, idade e cidade. A união foi realizada usando o identificador do paciente (`ID_PACIENTE`) presente em ambos os *datasets*, gerando ao final o conjunto de dados de trabalho[5].

Para este *dataset* utilizou-se uma abordagem de filtragem dos pacientes que foram testados para Covid-19. Em seguida foram selecionados os exames mais realizados por estes pacientes. A etapa seguinte foi a estruturação dos dados no formato arrumado de acordo com Wickham (2014). Após a estruturação dos dados, foi feito uma análise das estatísticas descritivas e aplicação do respectivo conjunto em aprendizado de máquina. Um

[1] `https:repositoriodatasharingfapesp.uspdigital.usp.br`
[2] `https://repositoriodatasharingfapesp.uspdigital.usp.br/handle/item/99`
[3] `https://repositoriodatasharingfapesp.uspdigital.usp.br/handle/item/97`
[4] `https://colab.research.google.com/`
[5] `https://drive.google.com/file/d/1Uuq_66cTR3ALo9MP4ciikcWWwuTuaVTT/view?usp=sharing`

resumo da metodologia utilizada para os dados do Laboratório Fleury pode ser observada na Figura 5.1.

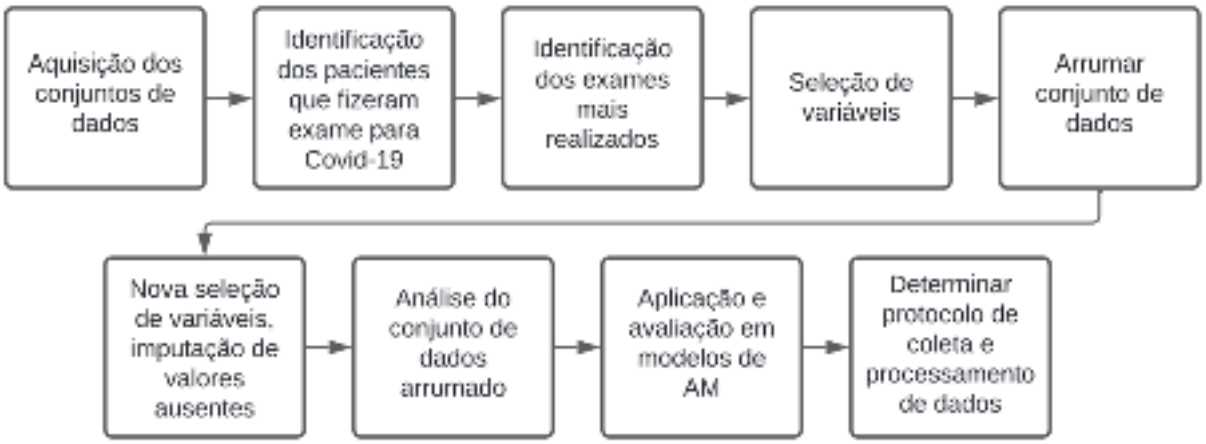


FIGURA 5.1 – Síntese da metodologia utilizada para os conjunto de dados de diagnóstico de Covid-19 do laboratório Fleury.

5.2 Pré-Processamentos dos dados para o Hospital Sírio-Libanês

Neste conjunto de dados são utilizados três *datasets*, sendo estes: `HSL_Exames`, `HSL_Pacientes` e `HSL_Desfecho`. O conjunto de dados `HSL_Exames` contém informações clínicas dos pacientes, como os exames realizados e os analitos referentes a cada exame. O conjunto de dados `HSL_Paciente` apresenta informações como ano de nascimento, gênero e estado, por exemplo. Já o conjunto `HSL_Desfecho` traz informações sobre o estado final do paciente, que varia, de forma geral, entre alta e melhora do quadro clínico ou óbito, bem como a data em que se deu o desfecho do caso, possibilitando apontar o período em que o paciente permaneceu em atendimento. Estes três conjuntos foram unidos usando a função "join" em um único arquivo de trabalho[6]. Este conjunto contém informações sobre sexo, idade e exames clínicos.

Como este *dataset* não demanda poder de processamento elevado, foi possível aplicar algumas padronizações, sendo estas a conversão de todas as letras maiúsculas, remoção dos acentos e remoção dos espaços duplos. A finalidade dessas ações consiste no fato de que um protocolo de coleta deve fornecer as informações padronizadas.

Em seguida, foram identificados os analitos presentes na coluna `DE_ANALITO` e posteriormente separados em 3 grupos de acordo com a frequência. O primeiro conjunto contém

[6] `https://drive.google.com/file/d/1qvp_p3FIvWLAYbrI6dh4ijIO4iVCgtIz/view?usp=sharing`

52 analitos com 18 mil entradas ou mais, o segundo conjunto contém 71 analitos com 8 mil entradas ou mais e o terceiro conjunto contém 167 analitos com mil entradas ou mais.

Neste estudo, foi considerado apenas o conjunto de analitos com 18 mil entradas ou mais, cujo objetivo é reduzir o número de valores ausentes no *dataset* final. Além disso, algumas colunas não apresentavam informações relevantes para a tarefa de prognósticos e foram excluídas, como por exemplo, a coluna contendo o estado em que o paciente foi atendido, a origem do exame e valores de referências, que apresentam relevância no prognóstico do paciente por AM. Ao final dessas etapas obtivemos um novo conjunto de dados de trabalho[7]. Este dados foram arrumados com base no estudo de Wickham (2014). De forma resumida, a metodologia utilizada nos dados de prognóstico de Covid-19 do Hospital Sírio-Libanês pode ser observada na Figura 5.2

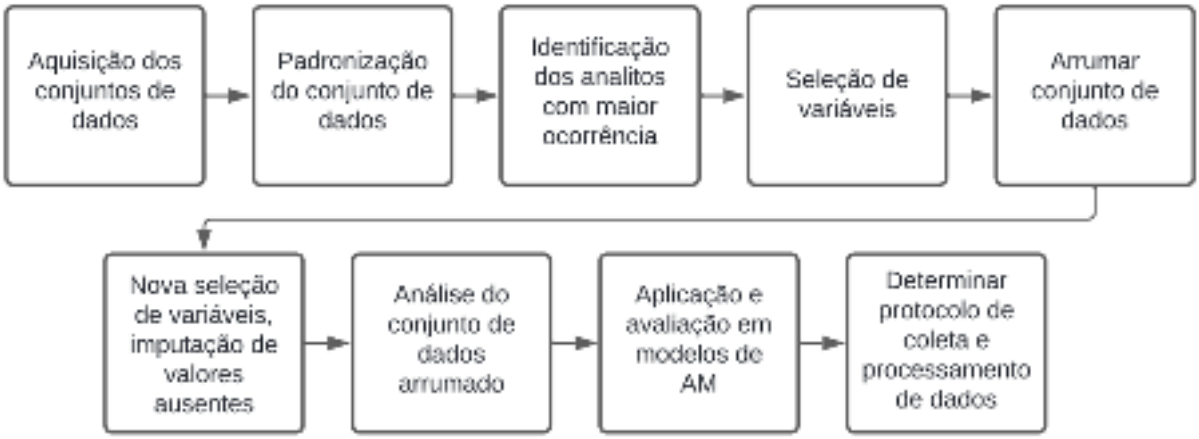


FIGURA 5.2 – Síntese da metodologia utilizada para os conjunto de dados de prognóstico de Covid-19 do Hospital Sírio-Libanês.

5.3 Aprendizado de máquina e indicação do protocolo de coleta de dados

Após o pré-processamento, algumas técnicas de visualização que permitem compreender o comportamento dos dados em ambos os conjuntos utilizados foram utilizadas. Logo após, os conjuntos de dados arrumados foram utilizados em modelos de aprendizado de máquina, que, a partir dos resultados obtidos, indicou-se um protocolo de coleta de dados.

Também foram utilizadas técnicas para determinar a proporção de valores ausentes para a verificação da necessidade de imputação artificial dos dados. Dessa forma, para atributos clínicos como até 5% de valores ausentes, foi feita a imputação utilizando a mediana. Atributos com grandes proporções de valores ausentes foram excluídos na etapa de aprendizado de máquina.

[7] `https://drive.google.com/file/d/1HnETgR35_qv-JdKFixKuCegQWYndX2Tb/view?usp=sharing`

Histogramas foram utilizados para verificar o comportamento dos dados considerando variáveis específicas, como a relação entre idade, diagnóstico e gênero. Na mesma linha, a média e desvio padrão (SD) também se destacam como técnicas de grande importância, já que possibilitam a identificação de atributos com valores de SD elevados, destacando a necessidade de maior atenção para estas informações (IZBICKI; SANTOS, 2020).

O uso de gráficos do tipo *boxplot* também foi dfoi utilizado. Através desse tipo de gráfico, é possível notar a proporção de *outliers* em cada atributo. Esta análise é importante para que, ao se determinar um protocolo de coleta, é necessário inferir o que é ou não um *outlier*, de modo que esses valores afetam diretamente a eficiência do modelo de AM (FACELI *et al.*, 2011).

Outra análise que se mostrou importante foi o uso de gráfico de dispersão comparando os mesmos atributos em classes diferentes de pacientes. Este tipo de gráfico permite observar de forma visual se os resultados dos exames, tanto para pacientes que foram testados como positivos ou negativos para Covid-19, quanto pacientes que tiveram sua condição agravada ou não, estão próximos ou não. Além disso, este tipo de informação permite a indicação de um modelo de AM mais adequado.

Após o estudo do comportamento dos dados, foram utilizados três modelos de aprendizado de máquina, sendo eles: Árvores de decisão com profundidade , KNN *k Nearest Neighbor*) e SVM (*Support Vector Machine*). Em relação aos hiperparâmetros, para a Árvore de decisão foi utilizada a máxima profundidade que a árvore pode ter (`max_depth = none`), para o KNN foi utilizada k=3, e para o SVM foi utilizado o kernel de função de base radial (RBF). Para cada algoritmo considerou-se 70% dos dados para treino e 30% para teste.

Para os resultados dos modelos AM foram utilizados tabelas contendo duas métricas: F1 e AUC. A medida F1 compreende a média harmônica de duas métricas (Precisão e Revocação) sem a necessidade de avaliar cada uma de forma isolada. A medida F1 é

$$\mathrm{F1} = \frac{2 \cdot \mathrm{precisão} \cdot revocacão}{precisão + revocacão}. \tag{5.1}$$

Quanto mais próximo de 1 for o valor de F1, mais o modelo é eficiente. Se este este valor é baixo, significa que a precisão ou *recall* está baixo e o modelo é pouco eficiente (ANZAI, 2012).

A métrica AUC considera a probabilidade de duas previsões serem ranqueadas corretamente, neste caso, a taxa de falsos positivos e a taxa de verdadeiros positivos. A AUC retorna um valor entre 0 e 1. Para valores de AUC mais próximos de 1, melhor a capacidade dos modelos de separar as classes. Já a curva ROC, de forma generalista, tem como finalidade validar um teste. Esta métrica considera a sensibilidade e especificidade

do modelo, que diz que, quanto mais para cima no eixo Y e mais para a esquerda no eixo X, melhor é o modelo (HAJIAN-TILAKI, 2013).

Essas duas métricas devem apresentar um resultado coerente para os modelos gerados e permitem inferir se o conjunto de dados utilizados está adequado para aplicação em aprendizado de máquina.

Com os resultados dos modelos de AM e considerando todo o processamento necessário para utilizar os dados, indicou-se ao final um protocolo de coleta que busca eliminar algumas etapas do pré-processamento e tornar os dados mais adequados para uso.

6 Resultados e discussão

Este capítulo é divido em duas seções. A primeira traz os resultados para as análises dos dados do Laboratório Fleury e a segunda traz as análises para os dados do Hospital Sírio-Libanês.

6.1 Resultados da análise dos dados do Laboratório Fleury

Para a organização dos dados considerou-se a exclusão das colunas que não apresentam informações relevantes para este estudo, sendo estas: [DE_ORIGEM, CD_PAIS, CD_-MUNICIPIO, CD_UF, CD_REPRODUZIDO, DE_VALOR_REFERENCIA]. Em seguida, foi criada uma coluna chamada `DIAGNOSTICO` que identificava os pacientes em POSITIVOS ou NEGATIVOS de acordo com os resultados para os exames de COVID-19. A coluna `AA_NASCIMENTO` foi transformada em uma nova coluna chamada `IDADE`, calculando a idade do paciente de acordo com o ano de nascimento e a data de coleta do dado.

A seleção de pacientes considerou ainda o tipo de exame mais realizado. Essa abordagem visa obter o maior número possível de pacientes com o mesmo tipo exame. Foram identificados a realização de 863 tipos de exames distintos, entre os quais os mais realizados são destacados na Figura 6.1.

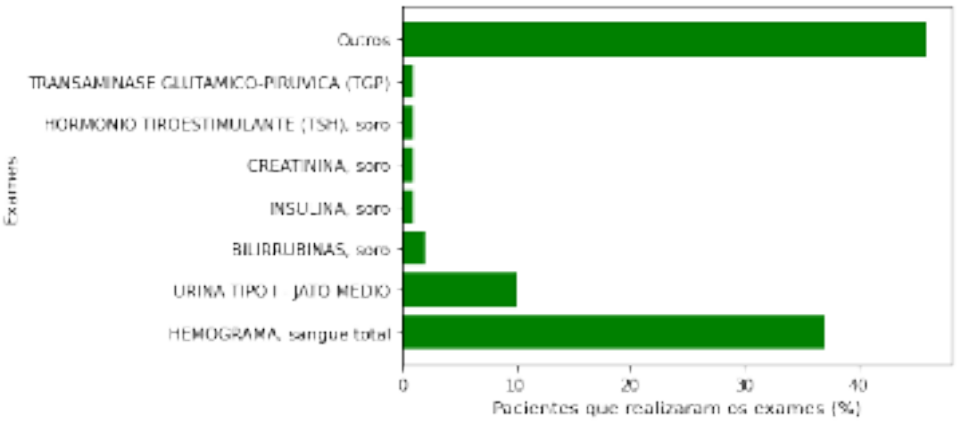


FIGURA 6.1 – Exames realizados para os pacientes que fizeram o teste para COVID-19 - Laboratório Fleury.

Considerando as informações apresentadas na Figura 6.1, é possível notar que aproximadamente 40% dos pacientes fizeram o exame `HEMOGRAMA, sangue total`, enquanto o segundo exame mais realizados foi `URINA TIPO I - JATO MEDIO`. Os outros exames, que agrupa vários tipos distintos de exames, foram automaticamente descartados já que foram realizados em um número expressivamente menor de pacientes.

Em relação ao exame de Urina, este apresentou diferentes resultados não numéricos e distintos que inviabilizaram seu uso sem a consulta à um especialista. Dessa forma, considerou-se apenas o uso do exame de sangue `HEMOGRAMA, sangue total` para compor o dataset utilizado em aprendizado de máquina. Com esta seleção definida, uma verificação foi feita na coluna `AA_NASCIMENTO` e foi observado que 0.56% das linhas apresentavam como valor de entrada a *string* "AAAA", indicando que a informação não fora disponibilizada. Estas linhas foram excluídas. A partir deste refinamento, foi obtido um *dataset* com as seguintes características apresentadas na Tabela 6.1[1].

TABELA 6.1 – Características do conjunto de dados obtido após a seleção de pacientes que realizaram o exame de sangue pelo Laboratório Fleury.

Características do dataset pré-processado	
Linhas	7342823
Linhas duplicadas	0
Colunas (atributos)	9
Atributos categóricos	4
Atributos numéricos	2
Atributos textuais	3

O conjunto de dados descrito na Tabela 6.1 possui um total de 174.640 pacientes. O número de linhas é expressivamente maior que o número de pacientes já que os dados ainda não foram arrumados. Neste conjunto, cada linha considera a identificação do paciente, identificação do atendimento, exame de sangue, analito coletado, data de coleta e outras variáveis. Logo, um único paciente pode estar em diferentes linhas devido aos diferentes analitos coletados, ou ainda, porque há mais de um exame de sangue realizado para um determinado paciente em datas distintas, o que torna o conjunto de dados extenso.

Setenta e oito por cento (78%) dos pacientes da Tabela 6.1 foram diagnosticados como negativo para Covid-19 e 22% tiveram resultados positivos. Em seguida este conjunto foi reestruturado de modo que as colunas `DE_ANALITO` e `CD_UNIDADE` foram associadas e preenchidas com os valores da coluna `DE_RESULTADO`.

Em seguida o conjunto de dados foi arrumado da seguinte forma: As colunas `DE_ANALITO` e `CD_UNIDADE` foram combinadas formando novas variáveis (novas colunas). Estas variáveis foram preenchidas pelos dados da coluna `DE_RESULTADO`. Logo, no conjunto de

[1] `https://drive.google.com/file/d/1Yh2qhZqmX_3xJ1TpFwjKsR9Szryt784C/view?usp=sharing`

dados arrumado, as colunas que trazem informações clínicas compreendem a combinação de outras 3 colunas do antigo conjunto de dados. Essa transformação gerou um dataset[2] com 39 colunas e 365.207 observações (linhas). A relação de colunas e valores ausentes para estes dados podem ser observados na Figura 6.2.

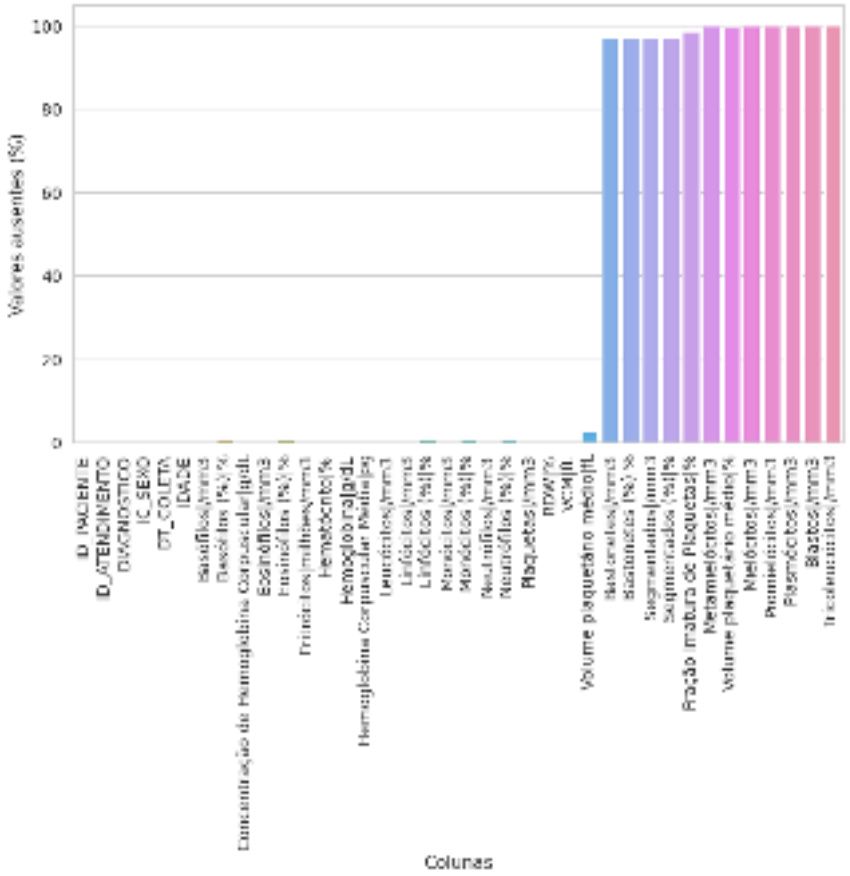


FIGURA 6.2 – Valores ausentes por coluna para o dataset do Laboratório Fleury pré-processado e arrumado.

A partir da Figura 6.2, é possível observar que as 13 colunas deslocadas mais à direita apresentam uma elevada quantidade de valores ausentes, passando dos 90%. Estes dados foram descartados, uma vez que é necessário um estudo a parte para verificar a possibilidade de imputação de dados artificias através de técnicas mais sofisticadas, como KNN (SOWMYA; KAYARVIZHY, 2021), MICE (SLADE; NAYLOR, 2020) ou Método Hot Decks (FOUAD *et al.*, 2021), por exemplo. As demais colunas tiveram seus valores ausentes preenchidos usando o algoritmo KNN com K=3. É importante ressaltar que a imputação dos dados artificiais foi feita apenas para o uso do conjunto de dados em aprendizado de máquina.

Outro aspecto importante consiste em compreender o comportamento da amostra de pacientes que compõe este conjunto de dados. A relação idade *versus* gênero pode ser

[2] `https://drive.google.com/file/d/1hJ-M0d_gYoHAtCxl1FFpfdEjCmIFt2hu/view?usp=sharing`

analisada por meio da Figura 6.3.

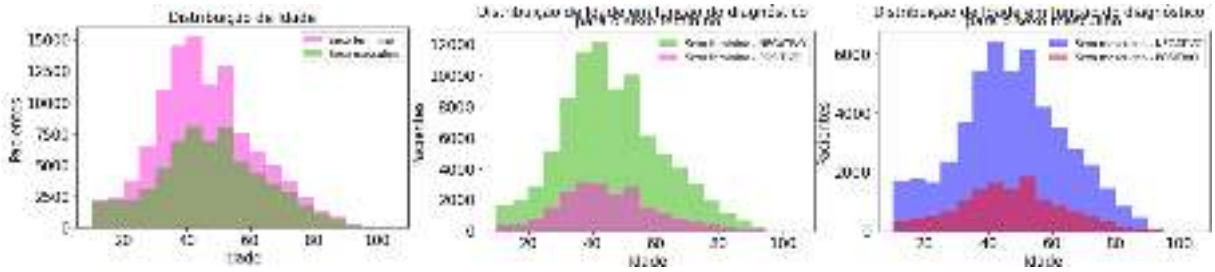


FIGURA 6.3 – Relação de pacientes em função da idade *vs* gênero testados para Covid-19.

É possível notar que da amostra de 174.640 pacientes, a extensa maioria é composta pelo sexo feminino com idade predominante entre 30 e 50 anos. O sexo masculino representa uma parcela inferior da amostra que e tem como característica a idade mais elevada, com a maior parte da amostra com idades entre 40 e 60 anos.

O sexo feminino também é o que apresenta maiores resultados positivos para Covid-19 com mulheres entre os 30 e 50 anos de idade. A parcela masculina que apresenta resultados positivos são de idade mais avançadas, com pico na faixa dos 50 anos de idade.

Observado o comportamentos dos dados em função do sexo dos pacientes, torna-se relevante observar as diferenças entre os resultados dos exames clínicos realizados. Vale lembrar que o exame selecionado neste estudo refere-se ao `HEMOGRAMA, sangue total`. A Tabela 6.2 traz uma análise comparativa entre os pacientes que testaram positivo e negativo para Covid-19, possibilitando a identificação de analitos que se comportam de forma distinta para os diferentes diagnósticos.

A partir da observação dos valores de média e de desvio-padrão, algumas conclusões podem ser inferidas. Seis analitos apresentam desvio-padrão elevado (`Eosinófilos|mm3`, `Leucócitos|mm3`, `Linfócitos (%)|%`, `Neutrófilos|mm3`, `Plaquetas|mm3` e `Monócitos (%)|%`), indicando que os dados não possuem um comportamento uniforme, enquanto outros atributos apresentam uma distribuição mais próxima dos valores de média, como é o caso dos analitos `Eritrócitos|milhõesmm3` e `Basófilos(%)|%`. Essa observação é importante uma vez que não é possível apontar se esses valores de desvio-padrão elevados são erros originados a partir da coleta ou são dados consistentes.

Mesmo os dados sendo originados a partir de um único local, neste caso um laboratório clínico, não se sabe como estes dados são coletados e inseridos na base de dados. Outro fator que corrobora para um questionamento do protocolo de coleta utilizado refere-se à uma grande quantidade de *outliers* em todos os analitos, mesmo naqueles em que os valores estão mais concentrados em torno da média. Os *outliers* podem ser verificados por meio da Figura 6.4.

Um possível tratamento para estes *outliers* seria a exclusão considerando a regra do

TABELA 6.2 – Estatísticas descritivas para o exame "HEMOGRAMA, sangue total" obtidas a partir do conjunto de dados arrumado dos pacientes que fizeram o teste para Covid-19 pelo Laboratório Fleury.

Analitos	Covid-19 Positivo	Covid-19 Negativo
	Média (SD)	
Basófilos|mm3	39.14 (22.75)	41.09 (36.29)
Basófilos (%)|%	0.62 (0.33)	0.64 (0.34)
Concentração de Hemoglobina Corpuscular|g/dL	33.61 (1.10)	33.60 (1.12)
Eosinófilos|mm3	167.93 (153.91)	181.95 (230.83)
Eosinófilos (%)|%	2.65 (2.16)	2.81 (2.35)
Eritrócitos|milhõesmm3	4.64 (0.52)	4.63 (0.54)
Hematócrito|%	40.78 (4.07)	40.74 (4.16)
Hemoglobina|gdL	13.71 (1.48)	13.69 (1.51)
Hemoglobina Corpuscular Média|pg	29.60 (1.97)	29.63 (2.04)
Leucócitos|mm3	6517.48 (2312.38)	6634.52 (3156.97)
Linfócitos|mm3	2138.40 (1245.51)	2164.92 (2247.60)
Linfócitos (%)|%	33.67 (9.61)	33.42 (9.66)
Monócitos|mm3	521.50 (188.08)	526.58 (261.17)
Monócitos (%)|%	8.22 (2.31)	8.14 (2.31)
Neutrófilos|mm3	3650.49 (1690.02)	3719.09 (1797.12)
Neutrófilos (%)|%	54.81 (10.54)	54.96 (10.57)
Plaquetas|mm3	2.56e+05 (6.83e+04)	2.55e+05 (6.85+e04)
RDW|%	13.17 (1.24)	13.19 (1.32)
VCM|fL	88.08 (5.10)	88.18 (5.33)
Volume plaquetário médio|fL	10.63 (0.87)	10.62 (0.88)

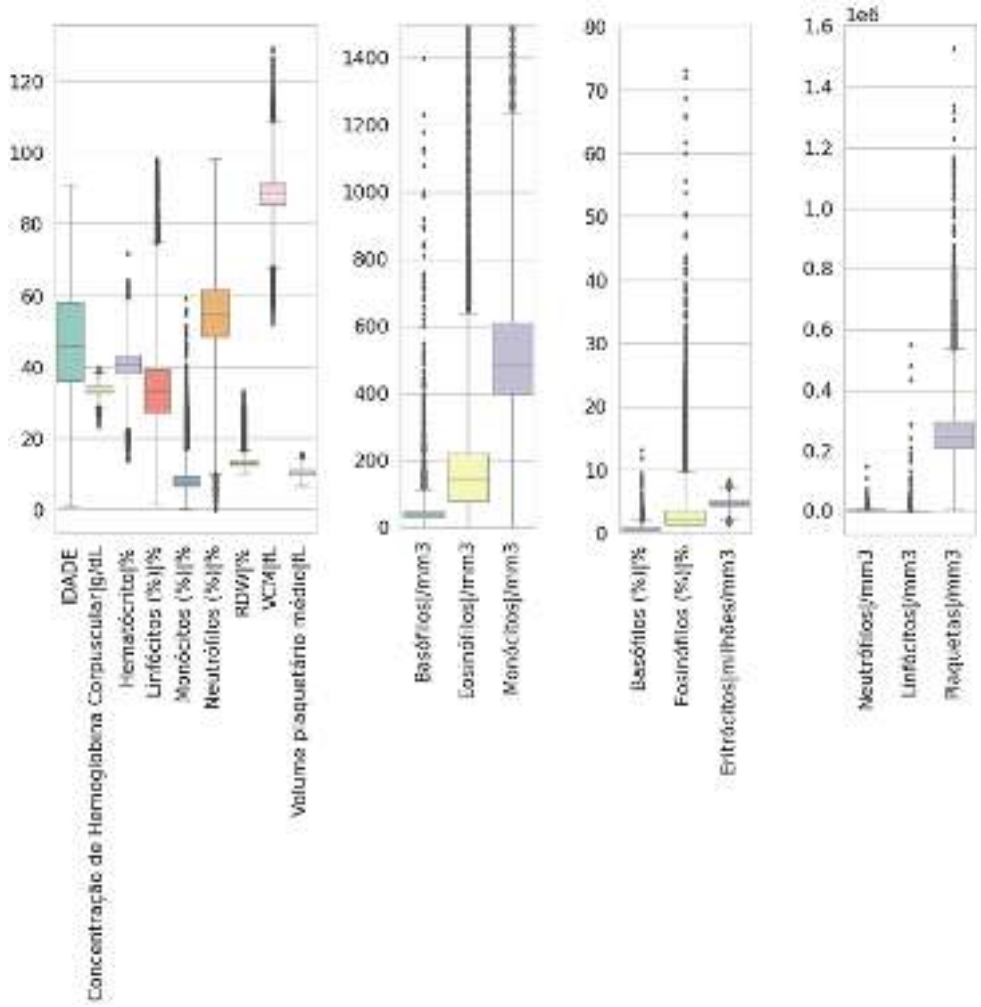


FIGURA 6.4 – *Outliers* para os analitos associados à unidades de medidas.

primeiro e terceiro quartil, contudo, como não há a certeza de que estes dados são ou não uma expressão da realidade dos pacientes analisados, uma vez que não foi possível ter acesso à profissional especialista em saúde, optou-se por mantê-los.

Até o momento pode-se perceber que os dados apresentam alguns problemas que demandam várias etapas de pré-processamento e a necessidade de um especialista para validar algumas informações. Além disso, notou-se a necessidade de uma estruturação para que estes dados possam ser utilizados em aprendizado de máquina para a determinação do diagnóstico dos pacientes, contribuindo com a tomada de decisões nos ambiente de saúde.

Para o uso do *dataset* em aprendizado de máquina, foram selecionados os seguintes algoritmos: KNN, Árvore de Decisão e SVM. O resultados obtidos para as tarefas de aprendizado de máquina são apresentados a seguir.

6.1.1 Aprendizado de máquina para diagnóstico de Covid-19

Como destacado na seção anterior, o conjunto de dados do laboratório Fleury, após as etapas de pré-processamento realizadas, é composto por um total de 365.207 pacientes, o que torna as análises nas tarefas de aprendizado de máquina inviáveis com os recursos computacionais disponíveis. Lembrando que estes recursos se limitam ao uso do da ferramenta Colab com 27.3 gigabytes de memória RAM. Dessa forma, foram selecionados de forma aleatória uma amostra de 3.500 pacientes que foram diagnósticos como negativos para Covid-19 e uma outra amostra de 1.000 pacientes diagnosticados como positivos, respeitando a proporção de cada grupo de pacientes no respectivo conjunto de dados.

Vale lembrar que para este grupo de pacientes foram selecionados atributos obtidos a partir do exame de sangue. Também é importante apontar que um mesmo paciente pode ter os mesmos exames em diferentes datas, nesse caso, considerou-se a última data de coleta do respectivo exame. Com base nisso, foram selecionadas 12 variáveis, que se correlacionam mais positivamente ou negativamente, com o objetivo de verificar se estes atributos se diferenciam em gráfico de dispersão, exposto na Figura 6.5.

É possível notar que não há distinção precisa entre os dois grupos, mostrando que os exames clínicos, de uma forma geral, apresentam valores similares para os atributos obtidos a partir do exame de sangue. Esse comportamento pode resultar em resultados insatisfatórios no aprendizado de máquina.

Para o aprendizado de máquina foram utilizados os algoritmos KNN, SVM com kernel linear e árvore de decisão, utilizando validação cruzada com dez pastas. Foram utilizadas duas métricas de avaliação: AUC e medida F1. Além disso, foi feita a normalização dos dados devido às diferenças de escalas entre os atributos selecionados[3]. Para o KNN foi feita uma normalização de reescala, enquanto para o SVM foi utilizada a padronização. De acordo com Faceli *et al.* (2011), a normalização por reescala consiste em determinar uma escala com um intervalo padrão, geralmente 0 e 1, enquanto a padronização considera que os atributos deverão possuir os mesmos valores para alguma medida de posição e de variação.

Foram obtidos dez resultados para cada métrica. Na sequência, foi calculada a média e valor de desvio padrão para cada algoritmo. Esses valores são apresentados na Tabela 6.3.

Em relação aos resultados apresentados, pode-se observar que o uso dos dados arrumados pode gerar resultados interessantes em tarefas de aprendizado de máquina, uma vez que a técnicas KNN, por exemplo, conseguiu classificar de forma satisfatória os pacientes em positivos ou negativos para Covid-19. Contudo, para que estes dados pudessem

[3] `https://github.com/alexsouza1989/DIAGNOSTICO_COVID`

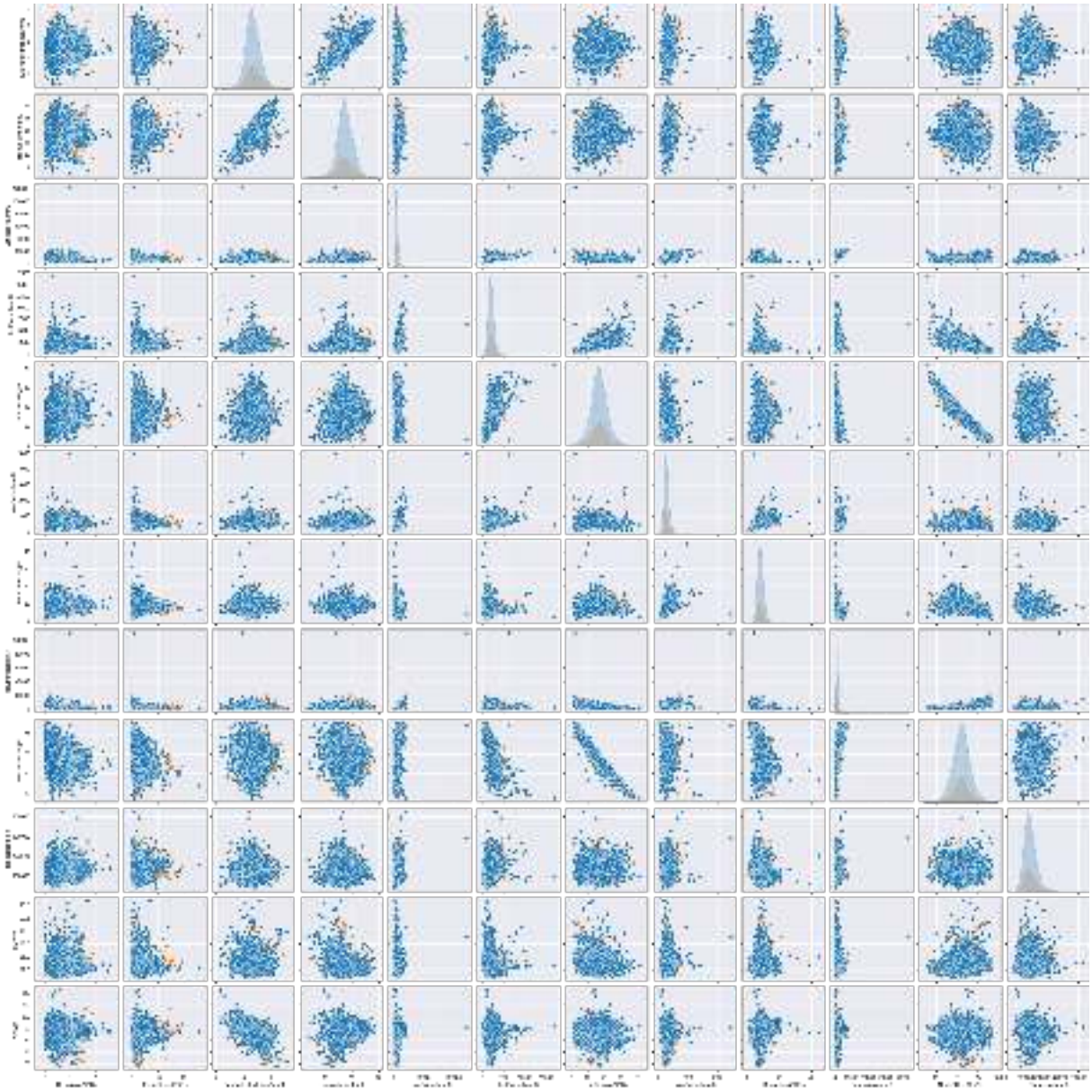


FIGURA 6.5 – Comportamento dos diferentes atributos entre pacientes diagnosticados como positivo para Covid-19 (cor laranja) e pacientes negativos para Covid-19 (cor azul).

TABELA 6.3 – Resultados para a classificação do diagnóstico por aprendizado de máquina para os dados do Laboratório Fleury.

Algoritmo	AUC	Medida F1
KNN	0.77 ± 0.02	0.35 ± 0.07
Decision Tree	0.67 ± 0.04	0.28 ± 0.05
SVM	0.78 ± 0.00	0.13 ± 0.05

ser utilizados na tarefa de diagnóstico por AM, foram necessárias diversas tarefas de pré-processamento, que poderiam ser evitadas ao se coletar essas informações já em um formato adequado.

Outro aspecto importante refere-se a qualidade dos dados, que apresentam inconsistências que também podem ser minimizadas através de uma coleta eficiente, reduzindo o tempo de pré-processamento necessário e permitindo que este conjunto seja utilizado de forma mais rápida pelos especialistas em saúde.

6.2 Resultados da análise dos dados do Hospital Sírio-Libanês

Os dados disponibilizados pelo hospital Sírio-Libanês permitem estudar o avanço da doença nos pacientes. O dataset utilizado nas análises foi formado pela união do seguintes conjuntos de dados: `HSL_Exames_4` (Exames clínicos), `HSL_Desfecho_4` (Condição final do pacientes) e `HSL_Pacientes_4` (Informações dos pacientes). Para esta ação utilizou-se a união dos conjuntos utilizando a variável `ID_PACIENTE`.

A proposta para este conjunto consiste em padronizar as informações, pré-processar os dados e preparar o dataset para aplicação em aprendizado de máquina. Dessa forma, a primeira modificação realizada foi a conversão de todas as letras para maiúsculas, remoção dos acentos e remoção de espaços duplos. Este dataset pode ser consultado online[4].

Com os dados padronizados e utilizando a coluna `DE_ANALITO`, verificou-se quais analitos foram obtidos por meio dos exames clínicos e o número de entrada de cada um. Identificou-se 1083 analitos distintos. Destes, o analito `CREATININA` é que apresenta maior número de entradas com 103471 resultados, enquanto outros analitos distintos possuem apenas uma entrada.

Devido ao grande números de analitos, determinou-se uma seleção arbitrária em três grupos distintos descritos na Tabela 6.4.

TABELA 6.4 – Características dos analitos na base de dados do Hospital Sírio-Libanês.

Número de analitos	Quantidade de valores no conjunto
52	18000 ou mais
71	8000 ou mais
167	1000 ou mais

Considerando a classificação descrita na Tabela 6.4, determinou-se o uso dos 52 analitos mais coletados. Essa seleção visa obter um *dataset*[5] com menor proporção de valores ausentes, visto que ao aumentar o número de analitos, aumenta-se também o números e pacientes que realizam poucos exames, gerando uma grande quantidade de dados faltantes.

[4] `https://drive.google.com/file/d/1xwkwv5zxKXFLGueLlrsHDlYM2MSfV5Yp/view?usp=sharing`
[5] `https://drive.google.com/file/d/1-0Br4ckiwUDuxn3vTA9SOa2DHiE-Jzge/view?usp=sharing`

Após a seleção dos analitos, o *dataset* foi reorganizado em função dos analitos e suas unidades de medidas. Esse método de organização garante que cada valor de analito seja correspondente à apenas uma variável. O novo conjunto de dados apresenta 235061 linhas e 68 colunas[6]. Todas as colunas apresentaram um grande volume de dados ausentes, com exceção das colunas `ID_PACIENTE`, `ID_ATENDIMENTO` e `DT_COLETA`, todas que estão 100% preenchidas. A proporção de ausentes nas colunas pode ser verificada na Figura 6.6.

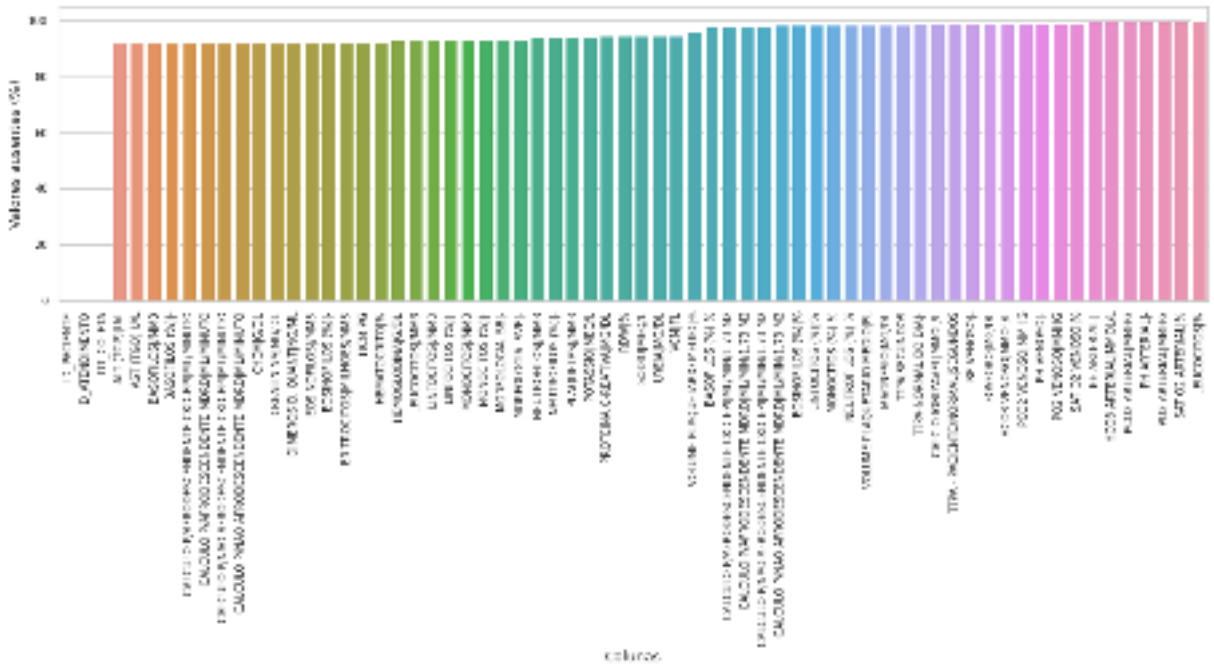


FIGURA 6.6 – Proporção de valores ausentes para o *dataset* arrumado obtido a partir dos dados Hospital Sírio-Libanês.

É possível notar que todas as colunas apresentam mais de 90% de dados ausentes. As últimas colunas à direita da Figura 6.6 apresentam cerca de 100% de dados ausentes. Esse comportamento se dá porque alguns pacientes apresentam a coleta de poucos analitos em uma mesma data e atendimento. Dessa forma, os analitos que não foram coletados nessa mesma data e atendimento ficam sem valores para preencher as demais colunas, gerando uma proporção elevada de valores ausentes.

Ainda sobre a Figura 6.6 e a proporção de valores ausentes, destaca-se que o conjunto de dados do Hospital Sírio-Libanês, no formato original, inviabiliza a realização de análises que permitam extrair alguma informação estatística ou identificação de padrões sem manipular os dados, mostrando que, mesmo com muitos valores ausentes, os dados com a estrutura indicada na Figura 6.6 ainda são melhores que os dados no formato original.

Uma forma de contornar esse problema seria ainda na fase de coleta de dados seria determinar uma lista de analitos que deveriam ser coletados em todos os atendimentos.

[6] `https://drive.google.com/file/d/1HjILaSR3Ug_NtoVnKhveKccYnvR-x0yJ/view?usp=sharing`

Dessa forma, a tendências é que os valores ausentes sejam menores, uma vez que se tem dados de todos as variáveis.

Esse problema também foi observado no estudo de Podder *et al.* (2021), que utilizaram um *dataset* com um grande número de variáveis apresentando mais de 80% de dados ausentes. Dessa forma, assim como no *dataset* para o laboratório Fleury, técnicas de imputação de dados artificias são necessárias, que, neste caso, precisam ser implementadas e avaliadas para que os dados representem a real natureza do cenário. Vale destacar que nesse estudo não foram utilizadas técnicas pra imputar dados artificialmente, sendo necessário um estudo separado para tratar desse problema.

Também foi notado que alguns analitos apresentam nomenclaturas semelhantes e mesma unidade de medida. Estes analitos foram combinados em uma nova coluna, reduzindo de 68 colunas para 62 colunas o *dataset*. Este processamento é descrito na Tabela 6.5.

TABELA 6.5 – Associação de colunas com os mesmos analitos identificados no conjunto de dados estruturado para o Hospital Sírio-Libanês.

Analitos agrupados	Nova coluna
[EOSINOFILOS (%)|] - [EOSINOFILOS (%)|%]	[EOSINOFILOS|%]
[BASOFILOS (%)|-] - [BASOFILOS (%)|(%)]	[BASOFILOS|%]
[LINFOCITOS (%)|-] - [LINFOCITOS (%)|%]	[LINFOCITOS|%]
[MONOCITOS (%)|-] - [MONOCITOS (%)|%]	[MONOCITOS|%]
[NEUTROFILOS (%)|-] - [NEUTROFILOS (%)|%]	[NEUTROFILOS|%]

Outro processamento necessário para o conjunto de dados foi a conversão dos valores de analitos do formato textual para o formato numérico. Essa conversão foi necessária para tornar os dados adequados para tarefas de aprendizado de máquina. Vale ressaltar também que esse tipo de operação é citada pela literatura médica como *capping*, que determina limites superiores e inferiores para alguma variável clínica atípica com a finalidade de tornar esse dado mais adequado para uso (SHESKIN, 2003). Esses analitos são mostrados na Tabela 6.6.

Duas colunas foram excluídas do conjunto de dados: "`MORFOLOGIA. SB|-`" que apresenta 92% de valores ausentes e 242 tipos diferentes de entradas; e "`MORFOLOGIA. SVE|-`" que também possui 92% valores ausentes e 729 tipos de entradas diferentes. Para esses analitos, sugere-se um estudo à parte devido à grande quantidade de informações que precisam ser analisadas por um especialista, por isso decidiu-se descartar deste estudo.

Considerando essas variáveis e os desafios relatados até o momento, existe um sério problema de falta de padronização nas informações médicas inseridas no banco de dados do Hospital Sírio-Libanês. Notou-se também que o preenchimento desses dois analitos não possibilita a identificação de padrões, dificultando o processamento dos dados para

TABELA 6.6 – Conversão de valores no formato textual para formato numérico nos dados do Hospital Sírio-Libanês.

Analitos	*String*	Valor inserido
CALCULO P/AFRODESCENDENTE CKD-EPI|ML/MINUTO	SUPERIOR A 90	91
CALCULO P/AFRODESCENDENTE MDRD|ML/MINUTO	SUPERIOR A 60	61
CALCULO P/NAO AFRODESCENDENTE CKDEPI|ML/MINUTO	SUPERIOR A 90	91
CALCULO P/NAO AFRODESCENDENTE MDRD|ML/MINUTO	SUPERIOR A 60	61
DIMEROS D. QUANT|NG/ML	INFERIOR A 215	214

obtenção de informações utilizáveis.

Assim, o novo conjunto de dados[7] para o hospital Sírio-Libanês, em formato ideal para tarefas de aprendizado de máquina, é composto por 60 variáveis, sendo duas categóricas (identificação do paciente e identificação do atendimento), uma variável temporal (data de coleta) e 57 numéricas (analitos em cada unidade de medida). Vale ressaltar que cada observação é formada pela identificação do paciente (`ID_PACIENTE`), a identificação do atendimento (`ID_ATENDIMENTO`), a data de coleta do respectivo analito (`DT_COLETA`) e valores obtidos a partir das combinações de analitos e unidades de medida.

Além dos erros de padronização reportados, também é possível notar a presença de variáveis com *outliers* extremos, levantando dúvidas sobre a veracidade dos dados coletados e inseridos no sistema. Assim, o conjunto de dados gerado a partir desses processamentos resulta em uma grande quantidade de dados ausentes associados a *outliers*. Além dos dados ausentes, observa-se também o mesmo problema de *outliers* na maior parte das variáveis, como mostra a Figura 6.7.

O tratamento dos *outliers* não foi feito nesse estudo já que é necessário um especialista clínico para identificar em que consiste esses pontos extremos. Isso se dá porque estes dados anômalos podem ser características clínicas dos pacientes ou erros de coleta.

A partir dos resultados apresentados, pode-se observar que os conjuntos de dados gerados para uso em AM possuem características similares, principalmente a grande quantidade de valores ausentes. Esses problemas são gerados ao preparar os *datasets* para uso em tarefas de aprendizado de máquina considerando o formato sugerido por Wickham

[7] `https://drive.google.com/file/d/1--l_P5_xHUgLD5BricnIk7IWeZOOvPcM/view?usp=sharing`

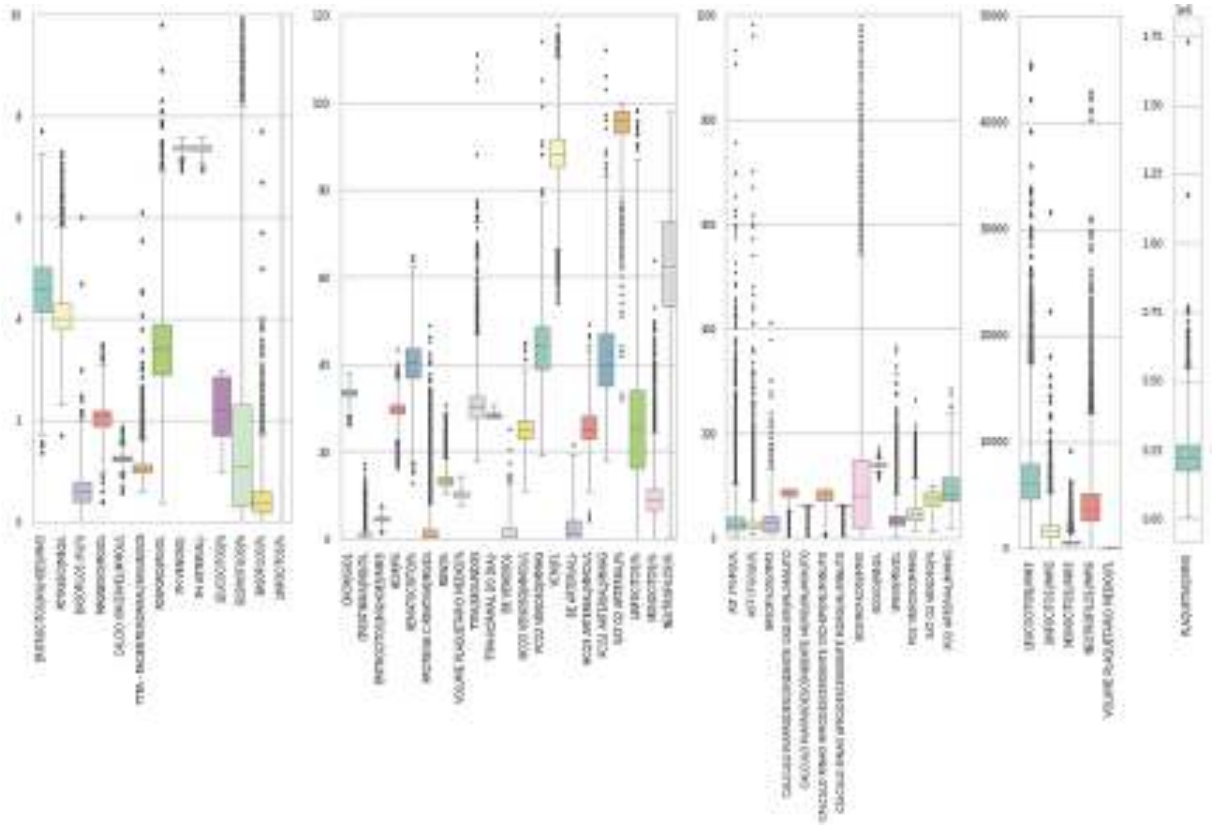


FIGURA 6.7 – Proporção de *outliers* para o *dataset* estruturado obtidos a partir dos dados Hospital Sírio-Libanês.

(2014).

Dessa forma, um protocolo de coleta de dados que resulta em dados devidamente estruturados reduziria a possibilidade de ocorrência desses erros, além de eliminar etapas no pré-processamento. Também é importante lembrar que para este *dataset* foi proposto uma padronização dos dados, como conversão das letras para maiúsculas e retirada de acentos gráficos, permitindo uma leitura mais simples dos dados. Nestes dados estas transformações conversão foram possíveis porque não houve a necessidade de desempenho de máquina elevado, enquanto no *dataset* do Laboratório Fleury, expressivamente maior, não foi possível aplicar essas transformações utilizando a ferramenta Colab.

6.2.1 Aprendizado de máquina para prognóstico de pacientes com Covid-19

Para a elaboração do aprendizado de máquina foram feitas outras etapas de pré-processamento, uma vez que o *dataset* descrito anteriormente apresenta uma grande quantidade de valores ausentes, tornando seu uso inviável. Utilizando, portanto, o conjunto com as transformações de padronização já feitas, selecionou-se pacientes que fizeram o exame de sangue chamado “HEMOGRAMA”, que se trata do exame mais realizado na

base de dados do Hospital Sírio-Libanês.

Como já foi descrito, diversas informações foram perdidas, reduzindo o conjunto de dados utilizados. Os pacientes que não apresentavam data de coleta de exames e/ou data de desfecho foram descartados da base, já que não foi possível inferir sua permanência no hospital.

Para determinar o fator gravidade, optou-se por uma método arbitrário onde são considerados como "casos grave" pacientes que ficaram a partir de 10 dias hospitalizados ou que foram à óbito. Pacientes com um período de hospitalização inferior a este são considerados como "casos não graves".

Também notou-se que alguns pacientes apresentavam exames coletados em mais de uma data, o que coloca em questão qual informação tem maior relevância no contexto médico. Neste estudo, considerou-se a última data em que os exames foram coletados. É importante destacar que um profissional de saúde pode solicitar exames de diferentes datas de acordo com sua necessidade, por exemplo, comparar exames em datas diferentes, situação esta que não se aplica à esta dissertação. Além disso, destaca-se que não foi feita seleção de exames com base nas datas de coleta e data de atendimento, visando assim obter a maior quantidade de dados possíveis.

A partir da seleção do exame, foi feito a organização do *dataset* associando os analitos à suas unidades de medidas, sendo o valor preenchido pela coluna `DE_RESULTADO`. Dessa forma, foi obtido ao final um conjunto de dados 23 colunas e 12.703 linhas, sendo que 20 colunas referem-se aos analitos selecionados, e as demais contendo informações como a idade do paciente e gravidade. Cada linha corresponde a um único paciente. Desse total, 9227 pacientes foram classificados como não graves e 3476 pacientes foram classificados como graves. Neste caso, é importante apontar que esta abordagem reduziu o número de analitos de 63 para 21, contudo, não há a geração de valores ausentes, tornado o conjunto de dados mais adequado para uso em AM.

Ao analisar alguns dos analitos selecionados, observou-se uma separação clara entre a parcela classificada como "GRAVE" e "NÃO GRAVE", como mostra a Figura 6.8.

É possível notar a partir da Figura 6.8 que há uma separação entre os pacientes graves e não graves para os atributos selecionados. Esta característica é importante uma vez que potencializa o desempenho dos modelos de aprendizado de máquina, que pode classificar os pacientes com maior precisão.

Para o aprendizado de máquina foram utilizadas as técnicas KNN, SVM e árvore de decisão[8]. Estas são técnicas extensivamente utilizadas em estudos de diagnóstico e prognóstico de doenças.

[8] `https://github.com/alexsouza1989/COVID-19_PROGNOSIS/blob/main/PRE_PREOCESSAMENTO_HSL.ipynb`

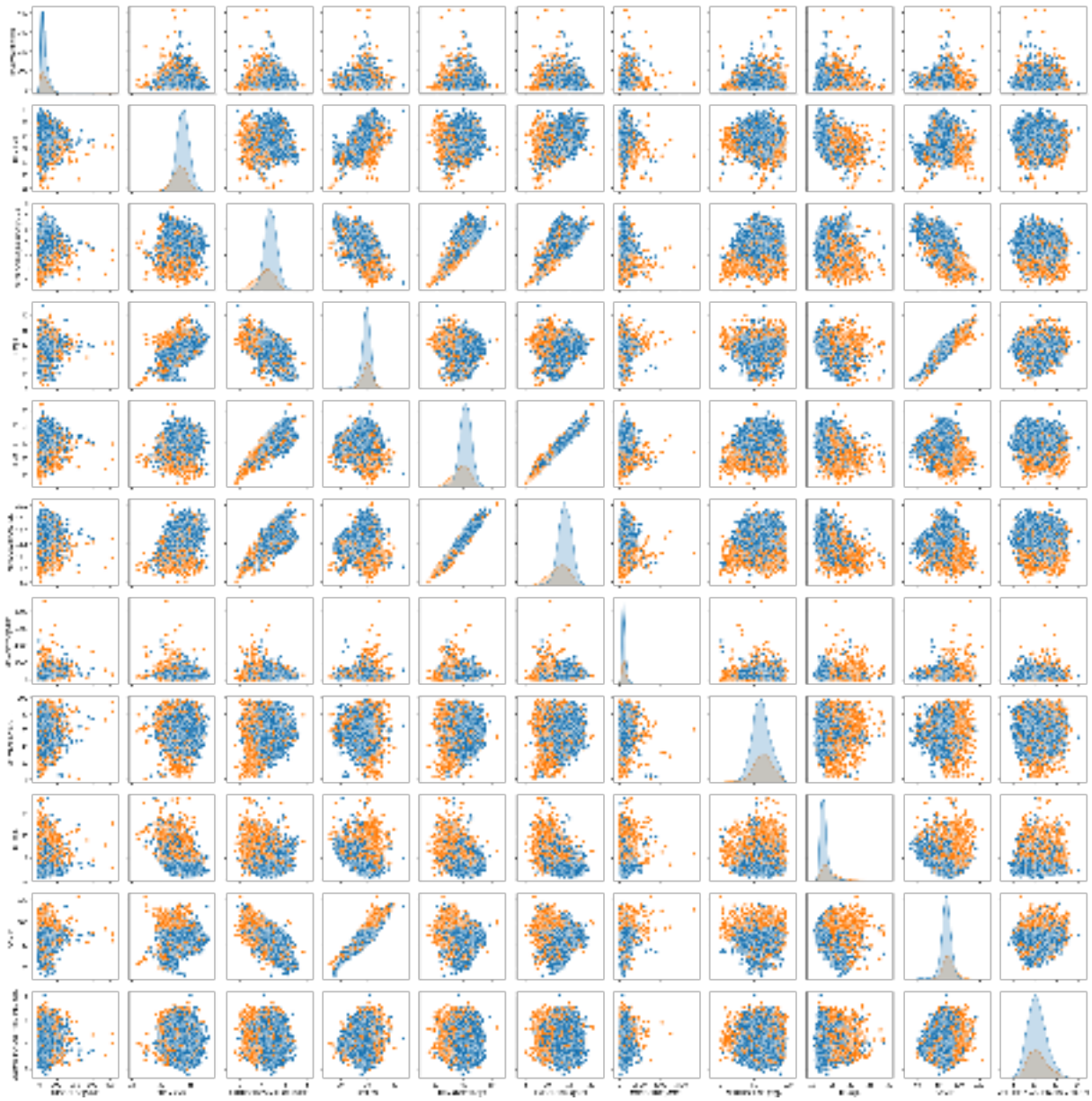


FIGURA 6.8 – Análise comparativa dos analitos para pacientes graves (cor laranja) e não graves (cor azul) para Covid-19 obtidos a partir dos dados Hospital Sírio-Libanês.

Para a geração dos modelos, foi utilizada a técnica de validação cruzada com 10 pastas. Além disso, foi feita a normalização dos dados, já que estes apresentam escalas distintas entre si.

Para avaliar os modelos foram utilizadas duas métricas, AUC e medida F1. Para cada uma das métricas foram obtidos dez resultados, sendo que a médias representa o resultado final apresentado seguido pelo desvio padrão. Os valores de AUC e F1 obtidos podem ser verificados na Tabela 6.7.

Considerando as informações na Tabela 6.7, pode-se afirmar que o KNN teve melhor desempenho quando comparado com os outros dois métodos, tanto para os valores de

TABELA 6.7 – Resultados para o prognóstico por aprendizado de máquina para os dados do Hospital Sírio-Libanês.

Algoritmo	AUC	Medida F1
KNN	0.81 ± 0.02	0.58 ± 0.04
Decision Tree	0.79 ± 0.01	0.51 ± 0.03
SVM	0.81 ± 0.01	0.53 ± 0.05

AUC quanto para os valores de F1. Lembrando que nessas duas métricas, quanto mais próximo de um for a saída, melhor é o desempenho do algoritmo. Dessa forma, a partir dos atributos coletados no exame de sangue, é possível determinar a gravidade de um paciente com razoável grau de precisão. Contudo, o KNN também apresenta valores de desvio padrão mais elevados. Considerando que o resultado exposto na 6.7 é a média dos dez dos valores obtidos para os algoritmos, isso pode indicar que o KNN pode ter alguns valores mais afastados da média, especialmente quando se olha para a medida F1, o que pode indicar a necessidade de se melhorar ainda mais os dados.

Além disso, é importante lembrar que para usar este conjunto de dados foram necessárias diversas etapas de pré-processamento que auxilia no prognóstico clínico de Covid-19 usando técnicas de AM, indicando a importância de se ter dados arrumados para a realização de análises clínicas.

6.3 Protocolo de coleta e transformação de dados

Ao longo dos tópicos anteriores foram descritas diversas atividades de limpeza e transformações dos dados com o objetivo de obter dois conjuntos dados adequados para uso em aprendizado de máquina, sendo um conjunto de dados para diagnóstico e outro para prognóstico. Dessa forma, este tópico tem como objetivo sumarizar todas as operações realizadas em uma ordem específica, de modo que ao final seja gerado um *dataset* com as propriedades necessárias para seu uso. Algumas dessas etapas poderiam ser eliminadas caso os dados fossem coletados de maneira apropriada.

Dessa forma, a Figura 6.9 busca descrever as ações que foram realizadas para que ao final fossem obtidos dois conjuntos de dados arrumados que foram respectivamente utilizados em tarefas de aprendizado de máquina.

Como mostrado na Figura 6.9, foram realizadas tarefas distintas para ambos os conjuntos de dados, bem como tarefas comuns. Algumas dessas tarefas são essenciais para que os dados sejam utilizados, por exemplo, a união de *datasets* que é a primeira tarefa realizada. Esta união combina informações clínicas com informações de características do paciente, como sexo e ano de nascimento.

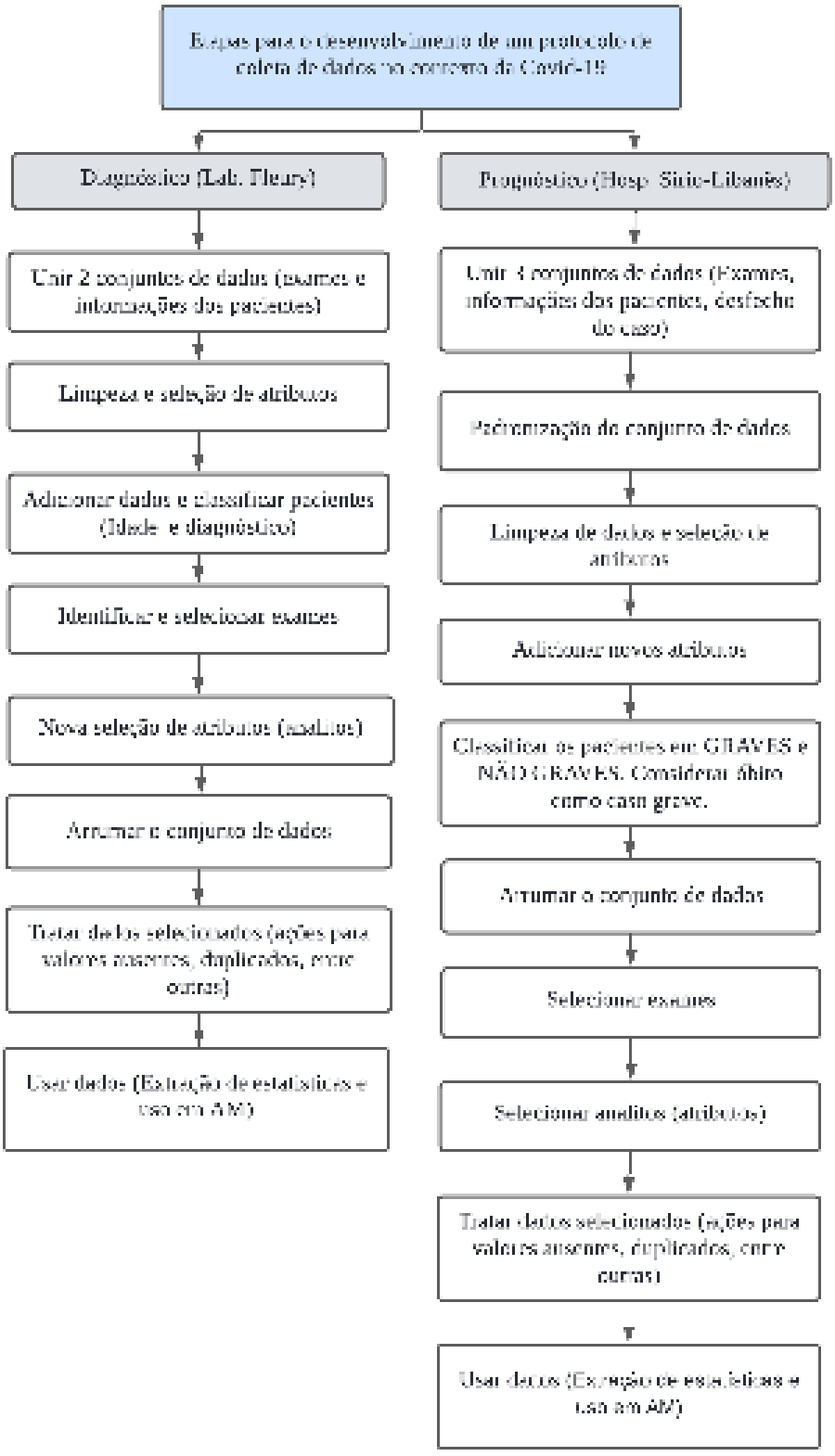


FIGURA 6.9 – Ações realizadas para determinação do protocolo de coleta e processamento de dados no contexto da Covid-19.

Em seguida, foi realizada a limpeza dos dados e seleção de atributos, que consiste na

remoção de informações duplicadas e exclusão de informações irrelevantes para diagnóstico e prognóstico, como Código de Endereçamento Postal (CEP) e origem do exame.

O próximo passo foi adicionar algumas informações que ajudam a compreender o perfil dos pacientes. Para os dados do Laboratório Fleury foi criada uma coluna com diagnóstico positivo e negativo para Covid-19 com base nas informações dos exames coletados. Para o Hospital Sírio-Libanês foi criada uma coluna indicando se o paciente se tratava de um caso grave ou não, considerando seu tempo de permanência no hospital ou se veio à óbito.

A partir desse ponto, todas as operações realizadas buscam obter conjuntos de dados arrumados, sendo possível estabelecer esses procedimentos ao longo do processo de coleta.

Para o Laboratório Fleury adotou-se a seleção dos exames mais realizados e, em seguida, a seleção dos analitos. Após, o dataset foi modificado, de modo que cada analito presente na coluna DE_ANALITO, associado à sua unidade de medida, se formasse uma única coluna preenchida pelos valores da coluna DE_RESULTADO.

Com o *dataset* já arrumado, foi feito o processamento de dados. Esse processamento teve como objetivo compreender as características do conjunto de dados e identificar problemas como *outliers*, dados ausentes e informações sem relevância, buscando obter ao final um conjunto de dados ideal para uso em modelos de AM, tanto para diagnóstico quanto para prognóstico, no caso do Hospital Sírio-Libanês.

Para o hospital Sírio-Libanês foram feitas etapas semelhantes. Foram unidos 3 conjuntos de dados distintos. Depois foi feito um pré-processamento inicial para eliminar informações repetidas e atributos irrelevantes. Na sequência foram selecionados 52 analitos mais frequentes. A partir dessa seleção, o dataset foi arrumado para análises estatísticas e uso em AM.

A partir deste ponto, as operações realizadas visam reduzir problemas gerados durante a coleta, como dados ausentes, informações desnecessárias, entrada de dados errados, entre outros. Vale destacar que o objetivo do estudo foi determinar um protocolo que permite coletar e usar os dados tanto para diagnóstico quanto para prognóstico e que cada conjunto apresenta problemas específicos. Logo, o protocolo apontado visa atender aos problemas de ambas as instituições citadas aqui.

Nesse contexto, o protocolo de coleta estabelecido pode ser observado na Tabela 6.8. Vale lembrar que algumas destas ações podem ser realizadas ao longo do procedimento de coleta para reduzir as tarefas de pré-processamento.

Como mostrado na Tabela 6.8, o primeiro passo para coletar os dados corretamente consiste em uma padronização do formato desses dados. Dessa forma, indica-se que os caracteres estejam padronizados para uso em letras maiúsculas, sem acentos gráficos e sem caracteres especiais como ç ou trema. Para os formatos numéricos, indica-se o uso

TABELA 6.8 – Protocolo para coletar dados clínicos.

Operação	Descrição
Padronização de caracteres	Indica-se que todos os caracteres devem estar em letras maiúsculas, sem acento e sem caracteres especiais (ç, trema, entre outros)
Coletar datas	A data é um dado sem custo, importante e deve ser coletado corretamente. Não é indicado usar DDMMAAA na imputação da informação. Em caso de ausência da data, tratar como ausente.
Combinações atributos	É indicado que ao coletar os dados clínicos, as colunas sejam formadas da seguinte forma: EXAME|ANALITO|UNIDADE_DE_MEDIDA e preenchida com valor numérico do analito em questão. Informações textuais, como observações, devem ser registradas a parte.
Selecionar *features*	Dados clínicos subjetivos devem ser tratados à parte e retirado do conjunto de dados final . Este tipo de dado é onde o profissional coloca sua percepção sobre uma amostra clínica ou sintoma, não sendo possível identificar padrões. Exemplo: Analito MORFOLOGIA.SVE. Não apresenta informações numéricas e diferentes tipos de textos para cada paciente.
Modificar dados	Em casos onde um tipo de é majoritariamente numérico e poucos valores são textos, verificar modificações nos dados. Exemplo: SUPERIOR A 90 alterar para 91.
Coletar analitos e unidades de medidas em campos distintos.	Esse campo visa evitar problemas de nomenclatura que prejudicam as análises, especialmente quando a unidade de medidas compõe o nome do analito. Exemplo real: Linfócitos(%)|%. Exemplo ideal: LINFOCITOS|%.

do ponto para separar as casas decimais, facilitando a leitura desses dados por programas como Python e R.

Como observado ao longo das manipulações dos conjuntos de dados do Hospital Sírio-Libanês e laboratório Fleury descritas na Figura 6.9, observou-se problemas comuns, por exemplo, a grande ocorrência de datas em formato “DDMMAA”, dificultando determinar o período em que o paciente coletou o exame, por quanto tempo este paciente permaneceu no hospital, ou ainda determinar em quanto tempo a partir do atendimento o exame foi coletado. Esses problemas dificultam tanto o uso dos conjuntos de dados em AM como também possíveis análises clínicas realizadas pela equipe médica.

Outro problema com a data consiste na entrada de informações erradas no sistema, gerando, por exemplo, um período de permanência com dias negativos no hospital. Este

problema inviabiliza saber se o erro está na data de atendimento ou na data do desfecho do caso, levantando um ponto de atenção que deve ser resolvido no momento em que esta informação for coletada.

Ainda em relação à data, esta é importante para determinar diferentes características do paciente, sendo um dos dados mais importantes, visto que o mesmo paciente pode realizar os mesmos exames em diferentes datas, possibilitando compreender o avanço ou retrocesso da doença, enfatizando, portanto, a importância dessa informação.

Entre outros problemas, foi observado que é necessário uma associação entre exame, analito e unidade de medida que possibilita reduzir a necessidade de identificar quantos tipos de exames há no conjunto de dados, bem como apontar quantos analitos há em cada exame. Essa tarefa, apesar de parecer simples, demanda grande esforço, visto que, por exemplo, pode haver exames de sangue com nomenclaturas distintas, mas que na prática são os mesmos exames.

Vale ressaltar que a identificação de exames clínicos e analitos com nomes distintos, mas que na prática são os mesmos, demanda profissionais da área médica/clínica e e profissionais da área de ciências de dados para validar essas informações, tornando o processo lento e oneroso.

Também foi observado que alguns atributos são difíceis de serem utilizados. Esses atributos apresentam dados em formato de texto, com informações subjetivas (por exemplo, a cor da urina que é preenchida por um enfermeiro), sendo necessário realizar um tratamento separado.

É indicado também que o sistema de coleta de dados permita inserir o analito e unidade de medida em campos distintos. Dessa forma é possível reduzir redundâncias e eliminar informações irrelevantes, uma vez que não é necessário investigar se dois atributos tratam o mesmo tipo de informação, mas que na prática possuem nomenclaturas distintas.

Outra ação necessária foi a modificação de dados indicado na Tabela 6.8. Alguns analitos podem apresentar majoritariamente dados numéricos e apenas um tipo de dado no formato texto, como é o caso do analito (`CALCULO P AFRODESCENTE CKD-EPI|MLMINUTO`) que apresenta apenas resultados numéricos e entradas com o texto “SUPERIOR A 90”, sendo possível converter esse texto para 91 e usá-lo sem maiores problemas em AM.

De forma geral, o protocolo de coleta de dados estabelecido aqui busca apontar algumas diretrizes fundamentais para coletar estas informações de uma forma mais eficiente, reduzindo redundâncias, descartando informações irrelevantes e transformando dados para que estes sejam utilizáveis. Um esboço de visual de como esta coleta de dados deve ser realizada, comparado com o modelo atual, é apresentado na Figura 6.10

Conforme mostrado na Figura 6.10, o conjunto de dados gerado para diagnóstico e

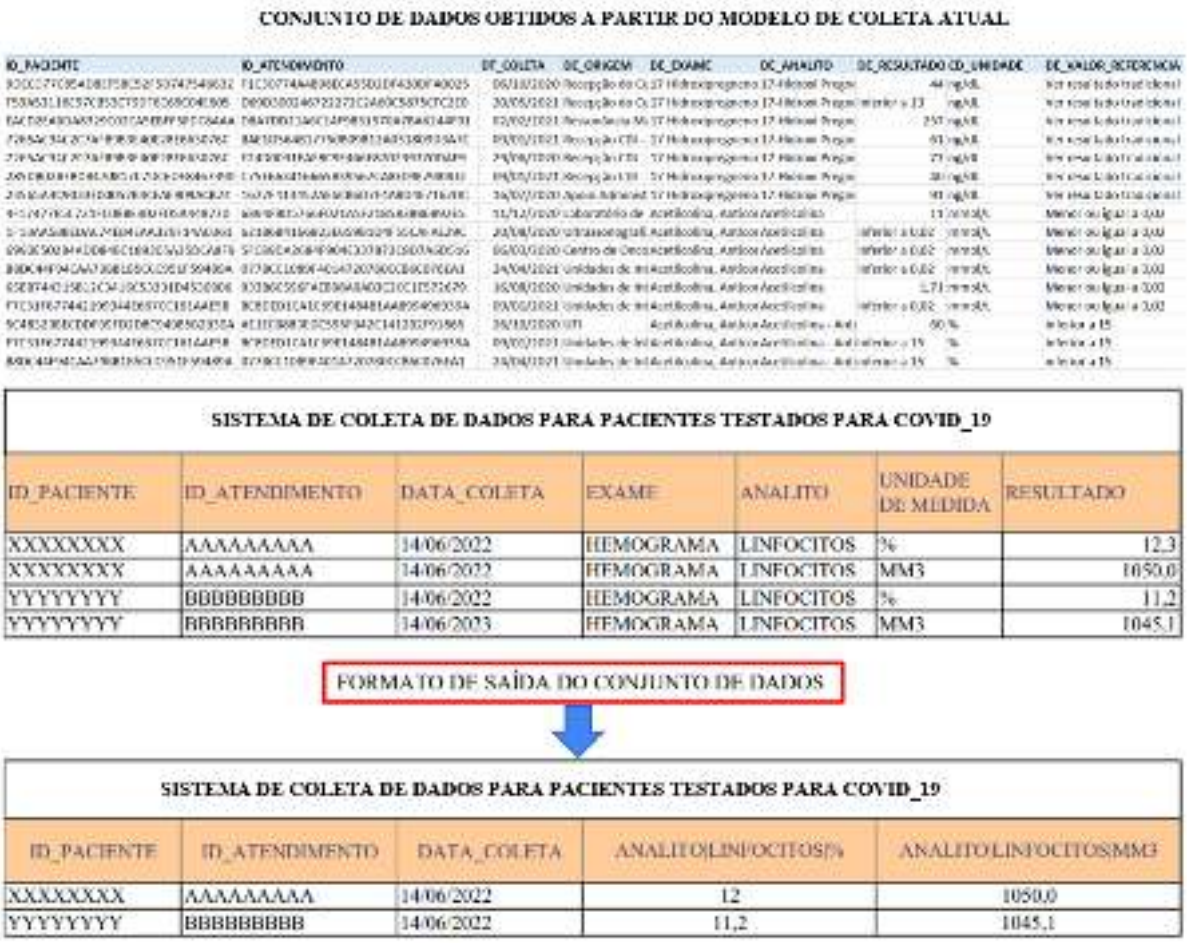


FIGURA 6.10 – Exemplo do formato proposto para coletar os dados de pacientes testados para Covid-19. A parte de cima da figura indica o modelo atual de coleta de dados. A parte inferior da figura mostra como os dados devem ser coletados e o formato do conjunto de dados obtidos a partir de uma coleta de dados adequada.

prognóstico da Covid-19 deve atender aos requisitos do protocolo estabelecido, onde as colunas são formadas por uma combinação de variáveis, reduzindo diversas etapas de pré-processamento. Vale destacar também que, mesmo com este protocolo, há operações que ainda são necessárias, como a combinação de conjuntos de dados contendo informações clínicas e informações sobre as características do paciente. Além disso, para uma coleta de dados voltada para o prognóstico do paciente, deve-se seguir a mesma sugestão indicada na Figura 6.10, contudo, deve-se adicionar os campos de data de atendimento do paciente, data do desfecho e desfecho do caso.

Outro ponto refere-se à variáveis como “data de nascimento”, que devem ser mantidas, uma vez que é possível determinar a idade de forma simples, sem afetar as análises em diferentes contextos e épocas. As outras datas, como data de atendimento e data de coleta também são importantes e devem ser mantidas.

Vale destacar também que esse método de coleta facilita as análises ao se considerar

mais de um exame como fonte de dados. Dessa forma, é possível, por exemplo, eliminar variáveis com maior ocorrência e estudar apenas variáveis com menor ocorrência, com o objetivo de entender a relação entre estes exames menos realizados.

Dessa forma, uma coleta de dados que gere um *dataset* arrumado possibilita o uso desses conjuntos de dados em tarefas de AM e análises estatísticas mais rapidamente, descartando, portanto, etapas de pré-processamento que demandam tempo e mão-de-obra especializada. Este mesmo protocolo pode ainda ser utilizado para coletar informações em eventos futuros, potencializando o uso dessas informações e a tomada de ação rápidas pelas instituições de saúde, contribuindo também para um manejo mais rápido e eficiente dos pacientes. O protocolo proposto visa reduzir também os erros descritos anteriormente, facilitando a semântica das informações contidas no *dataset*, como agrupar um único tipo de variável em uma coluna e indicar um arranjo que facilite a manipulação dos dados.

Nessa linha, o protocolo de coleta proposto, bem como as etapas de pré-processamento realizadas ao longo deste estudo, se aplicam apenas para dados numéricos, o que fez com que dados textuais não fossem considerados. Logo, outra possibilidade de investigação é justamente estudar esses dados textuais na busca de padrões e informações relevantes para auxiliar no combate à doença.

Além disso, os conjuntos de dados referentes à Covid-19 são extensos e contém inúmeros problemas que podem ser investigados. Trata-se de um campo de estudo interessante para a área de ciências de dados e aprendizado de máquina, já que diversas abordagens podem ser utilizadas, destacando-se a possibilidade de uso de aprendizado não supervisionado, especialmente nos conjuntos de dados arrumados, o que pode trazer novas perspectivas para abordar a Covid-19 ou alguma nova doença desconhecida pelo homem.

7 Conclusão

O uso de aprendizado de máquina tem sido cada vez mais estudado no combate a diversos tipos de doenças, especialmente no combate ao Coronavírus. O surgimento da Covid-19 afetou o mundo de diferentes formas, causando centenas de milhares de mortes, levando à medidas de isolamento rigorosas para reduzir o avanço de uma doença sem cura e com tratamento ainda desconhecido no ano de sua origem, em 2020.

Devido à necessidade de medidas urgentes, diversos dados foram utilizados para compreender características clínicas da Covid-19. No contexto brasileiro, os dados disponibilizados apresentam inúmeros desafios que devem ser enfrentados para que estas informações se tornem adequadas ao uso, tanto em tarefas de AM como em análises estatísticas.

Nesse estudo foram utilizados dois conjuntos com finalidades distintas. O primeiro é o conjunto de dados fornecido pelo laboratório Fleury, que tem como finalidade indicar o diagnóstico dos pacientes em relação à Covid-19. O segundo dataset refere-se a dados do Hospital Sírio-Libanês, que tem como objetivo apontar o prognóstico da doença.

Para ambos os conjuntos de dados foram necessárias ações como: união de conjuntos de dados, padronização de caracteres, transformação de dados, limpeza de dados, reorganização dos conjuntos de dados, seleção e eliminação de atributos e eliminação de redundâncias. A partir deste pré-processamento e adequação dos conjuntos, os mesmos foram utilizados para prever o diagnóstico e prognóstico da doença.

Para os dados do Laboratório Fleury, determinou-se basicamente o uso de atributos obtidos a partir do exame de sangue. A partir de todo processamento e arrumação dos dados e seu uso em modelos de aprendizado de máquina, observou-se que o algoritmo KNN obteve melhor desempenho na predição da doença, com AUC de 0.77, indicando que o uso do dataset, no formato sugerido, mostra-se eficientes no auxílio ao diagnóstico.

Para os dados do Hospital Sírio-Libanês, considerou-se a abordagem de seleção dos analitos independente de seus exames, contudo, este método gerou uma grande quantidade de valores ausentes. Este conjunto de dados precisou de uma série de transformações e adequações além daquelas realizadas nos dados do Laboratório Fleury, para então, gerar um dataset com dados utilizáveis. Em seguida, foram selecionados os analitos obtidos a partir do exame de sangue, permitindo assim um conjunto de dados com poucos valores

ausentes. Ao utilizar esse conjunto de dados em tarefas de AM para prognóstico, observou-se que os três algoritmos selecionados apresentaram bons desempenhos, sendo: KNN e SVM com AUC de 0.81, respectivamente e árvore de decisão com AUC = 0.72.

O desempenho superior em relação aos dados de diagnóstico do Laboratório Fleury ocorre devido a uma diferença clara nos resultados dos exames para os pacientes que tiveram seu estado de saúde classificado como grave. O mesmo não ocorre entre pacientes testados como positivo ou negativo para Covid-19, que possuem valores de exames similares, indicando que, mesmo com os dados arrumados, os algoritmos apresentam baixo despenho.

É importante ressaltar que este trabalho não priorizou o estudo da aplicação de aprendizado de máquina para estes problemas. Por exemplo, não foram realizados testes estatísticos para comparar os diferentes modelos treinados. Ainda, o impacto da qualidade dos dados não foi avaliado.

Com base no estudo realizado, notou-se que o uso de conjuntos de dados, em um formato arrumado, permite o uso eficiente dessas informações, ressaltando, portanto, a importância de um protocolo de coleta que gere *datasets* arrumados com o objetivo de potencializar e contribuir para a tomada de decisões médicas, tanto no contexto da Covid-19, ou ainda, em possíveis eventos futuros, onde, aplicando-se o protocolo estabelecido, pode-se otimizar as análises e manejo de pacientes e recursos.

Por fim, vale destacar que os conjuntos de dados utilizados neste estudo permitem uma extensa abordagem de uso de métodos computacionais, e que as técnicas e métodos utilizados aqui representam uma fração dessas possibilidades.

Nessa linha, sugere-se como trabalhos futuros estudar os conjuntos de dados obtidos nesta dissertação sob a perspectiva do aprendizado de máquina não supervisionado, agregando ainda mais no desenvolvimento de análises que podem contribuir para a área médica de uma forma geral. Também sugere-se um estudo para analisar o impacto no desempenho dos modelos de AM caso os dados sejam coletados com erros, com diferentes frações de dados ausentes, presença de *outliers*, entre outros.

7.1 Impactos e indicadores do trabalho

Em 2019, o mundo conheceu uma nova doença que modificou diversas esferas da vida em sociedade. Diante dessa mudança brusca de cenário, diversos estudos e contribuições científicas foram desenvolvidas ao redor do mundo. No que diz respeito a este estudo, sua contribuição consiste em fornecer um método que possibilita a coleta de dados clínicos de uma forma eficiente. De forma mais prática, o protocolo indicado aqui pode ser

aplicado em casos de Covid-19 ou ainda no surgimento de novas doenças, facilitando a semântica desses dados clínicos e possibilitando seu uso de forma mais rápida em tarefas de aprendizado de máquina.

Com o uso desse protocolo de coleta de dados, espera-se contribuir para a tomada de decisões em saúde, com o objetivo de proporcionar o manejo mais adequado dos pacientes e dos recursos hospitalares, contribuindo diretamente para a promoção da saúde da população. Uma vez que é possível identificar padrões de uma doença, ou agravamento da mesma, os profissionais de saúde podem tomar decisões, por exemplo, direcionadas a evitar maiores danos à integridade física e à saúde do paciente considerando também os recursos disponíveis.

No campo econômico, o método de coleta de dados aqui descrito, bem como seu uso em tarefas de aprendizado de máquina, pode otimizar a gestão dos insumos utilizados em hospitais e demais tipos de instituições de saúde. Uma vez que é possível prever padrões de agravamento da doenças, por exemplo, pode-se direcionar mais recursos para esse paciente, e o contrário também pode ocorrer. Sabendo que um paciente não apresenta um padrão ou tendência ao agravamento da doença, pode-se reduzir custos operacionais com este paciente.

Por fim, a facilitação do uso de dados, considerando a coleta em formato arrumado e sua aplicação de uma forma mais ampla, tem uma implicação direta em gastos com saúde, promoção de uma melhor abordagem ao paciente no espaço de saúde e também pode reduzir o tempo de trabalho dos profissionais que lidam com essas informações diariamente.

Referências

ALBALLA, N.; AL-TURAIKI, I. Machine learning approaches in covid-19 diagnosis, mortality, and severity risk prediction: A review. **Informatics in Medicine Unlocked**, Elsevier, v. 24, p. 100564, 2021.

ALSHEIKH, M. A.; LIN, S.; NIYATO, D.; TAN, H.-P. Machine learning in wireless sensor networks: Algorithms, strategies, and applications. **IEEE Communications Surveys & Tutorials**, IEEE, v. 16, n. 4, p. 1996–2018, 2014.

ANZAI, Y. **Pattern recognition and machine learning**. [S.l.]: Elsevier, 2012.

BAŞTANLAR, Y.; ÖZUYSAL, M. Introduction to machine learning. **miRNomics: MicroRNA biology and computational analysis**, Springer, p. 105–128, 2014.

BONACCORSO, G. **Machine learning algorithms**. [S.l.]: Packt Publishing Ltd, 2017.

BZDOK, D.; KRZYWINSKI, M.; ALTMAN, N. Machine learning: supervised methods. **Nature methods**, NIH Public Access, v. 15, n. 1, p. 5, 2018.

CANDANEDO, I. S.; NIEVES, E. H.; GONZÁLEZ, S. R.; MARTÍN, M.; BRIONES, A. G. Machine learning predictive model for industry 4.0. In: SPRINGER. **International Conference on Knowledge Management in Organizations**. [S.l.], 2018. p. 501–510.

CHU, X.; ILYAS, I. F.; KRISHNAN, S.; WANG, J. Data cleaning: Overview and emerging challenges. In: **Proceedings of the 2016 international conference on management of data**. [S.l.: s.n.], 2016. p. 2201–2206.

COGSWELL, M.; AHMED, F.; GIRSHICK, R.; ZITNICK, L.; BATRA, D. Reducing overfitting in deep networks by decorrelating representations. **arXiv preprint arXiv:1511.06068**, 2015.

EANEFF, S.; OBERMEYER, Z.; BUTTE, A. J. The case for algorithmic stewardship for artificial intelligence and machine learning technologies. **Jama**, American Medical Association, v. 324, n. 14, p. 1397–1398, 2020.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. d. L. F. d. Inteligência artificial: uma abordagem de aprendizado de máquina. 2011.

FERNANDES, F. T.; OLIVEIRA, T. A. de; TEIXEIRA, C. E.; BATISTA, A. F. d. M.; COSTA, G. D.; FILHO, A. D. P. C. A multipurpose machine learning approach to predict covid-19 negative prognosis in são paulo, brazil. **Scientific reports**, Nature Publishing Group, v. 11, n. 1, p. 1–7, 2021.

FISHER, D. H.; PAZZANI, M. J.; LANGLEY, P. **Concept formation: Knowledge and experience in unsupervised learning**. [S.l.]: Morgan Kaufmann, 2014.

FOUAD, K. M.; ISMAIL, M. M.; AZAR, A. T.; ARAFA, M. M. Advanced methods for missing values imputation based on similarity learning. **PeerJ Computer Science**, PeerJ Inc., v. 7, p. e619, 2021.

HAJIAN-TILAKI, K. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. **Caspian journal of internal medicine**, Babol University of Medical Sciences, v. 4, n. 2, p. 627, 2013.

HARRIS, C. R.; MILLMAN, K. J.; WALT, S. J. van der; GOMMERS, R.; VIRTANEN, P.; COURNAPEAU, D.; WIESER, E.; TAYLOR, J.; BERG, S.; SMITH, N. J.; KERN, R.; PICUS, M.; HOYER, S.; KERKWIJK, M. H. van; BRETT, M.; HALDANE, A.; RÍO, J. F. del; WIEBE, M.; PETERSON, P.; GÉRARD-MARCHANT, P.; SHEPPARD, K.; REDDY, T.; WECKESSER, W.; ABBASI, H.; GOHLKE, C.; OLIPHANT, T. E. Array programming with NumPy. **Nature**, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <https://doi.org/10.1038/s41586-020-2649-2>.

HUNTER, J. D. Matplotlib: A 2d graphics environment. **Computing in Science & Engineering**, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007.

ILYAS, I. F.; CHU, X. **Data cleaning**. [S.l.]: Morgan & Claypool, 2019.

IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina: uma abordagem estatística**. [S.l.]: Rafael Izbicki, 2020.

KHAN, I.; ZHANG, X.; REHMAN, M.; ALI, R. A literature survey and empirical study of meta-learning for classifier selection. **IEEE Access**, IEEE, v. 8, p. 10262–10281, 2020.

KONONENKO, I. Machine learning for medical diagnosis: history, state of the art and perspective. **Artificial Intelligence in medicine**, Elsevier, v. 23, n. 1, p. 89–109, 2001.

KOUROU, K.; EXARCHOS, T. P.; EXARCHOS, K. P.; KARAMOUZIS, M. V.; FOTIADIS, D. I. Machine learning applications in cancer prognosis and prediction. **Computational and structural biotechnology journal**, Elsevier, v. 13, p. 8–17, 2015.

KUKAR, M.; GUNČAR, G.; VOVKO, T.; PODNAR, S.; ČERNELČ, P.; BRVAR, M.; ZALAZNIK, M.; NOTAR, M.; MOŠKON, S.; NOTAR, M. Covid-19 diagnosis by routine blood tests using machine learning. **Scientific reports**, Nature Publishing Group, v. 11, n. 1, p. 1–9, 2021.

KUMARI, S.; KUMAR, S. A comparative study of various data transformation techniques in data mining. **International Journal of Scientific Engineering and Technology**, Citeseer, v. 4, n. 3, p. 146–148, 2015.

LANGS, G.; RÖHRICH, S.; HOFMANNINGER, J.; PRAYER, F.; PAN, J.; HEROLD, C.; PROSCH, H. Machine learning: from radiomics to discovery and routine. **Der Radiologe**, Springer, v. 58, n. 1, p. 1–6, 2018.

LINSSEN, J.; ERMENS, A.; BERREVOETS, M.; SEGHEZZI, M.; PREVITALI, G.; RUSSCHER, H.; VERBON, A.; GILLIS, J.; RIEDL, J.; JONGH, E. de *et al.* A novel haemocytometric covid-19 prognostic score developed and validated in an observational multicentre european hospital-based study. **Elife**, eLife Sciences Publications Limited, v. 9, p. e63195, 2020.

MAZYAVKINA, N.; SVIRIDOV, S.; IVANOV, S.; BURNAEV, E. Reinforcement learning for combinatorial optimization: A survey. **Computers & Operations Research**, Elsevier, v. 134, p. 105400, 2021.

MCKINNEY, W. *et al.* Data structures for statistical computing in python. In: AUSTIN, TX. **Proceedings of the 9th Python in Science Conference**. [S.l.], 2010. v. 445, p. 51–56.

MORAES, B. A. F. de; MIRAGLIA, J.; DONATO, T.; FILHO, A. Covid-19 diagnosis prediction in emergency care patients: a machine learning approach. **https://www. medrxiv. org/content/medrxiv/early/2020/04/07/2020.04. 04.20052092. full. pdf**, 2020.

OBAID, H. S.; DHEYAB, S. A.; SABRY, S. S. The impact of data pre-processing techniques and dimensionality reduction on the accuracy of machine learning. In: IEEE. **2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)**. [S.l.], 2019. p. 279–283.

OBERMEYER, Z.; EMANUEL, E. J. Predicting the future—big data, machine learning, and clinical medicine. **The New England journal of medicine**, NIH Public Access, v. 375, n. 13, p. 1216, 2016.

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

PODDER, P.; BHARATI, S.; MONDAL, M. R. H.; KOSE, U. 9 - Application of machine learning for the diagnosis of COVID-19. In: KOSE, U.; GUPTA, D.; ALBUQUERQUE, V. H. C. de; KHANNA, A. (Ed.). **Data Science for COVID-19**. Academic Press, 2021. p. 175–194. ISBN 978-0-12-824536-1. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128245361000083>.

PROVOST, F.; FAWCETT, T. Data science and its relationship to big data and data-driven decision making. **Big data**, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 1, n. 1, p. 51–59, 2013.

RAJKOMAR, A.; DEAN, J.; KOHANE, I. Machine learning in medicine. **New England Journal of Medicine**, Mass Medical Soc, v. 380, n. 14, p. 1347–1358, 2019.

REDDY, G. T.; REDDY, M. P. K.; LAKSHMANNA, K.; KALURI, R.; RAJPUT, D. S.; SRIVASTAVA, G.; BAKER, T. Analysis of dimensionality reduction techniques on big data. **IEEE Access**, IEEE, v. 8, p. 54776–54788, 2020.

SABA, T.; REHMAN, A.; ALGHAMDI, J. S. Weather forecasting based on hybrid neural model. **Applied Water Science**, Springer, v. 7, n. 7, p. 3869–3874, 2017.

SARKER, I. H. Machine learning: Algorithms, real-world applications and research directions. **SN Computer Science**, Springer, v. 2, n. 3, p. 1–21, 2021.

SCHAFFNER, K. F. **Logic of discovery and diagnosis in medicine**. [S.l.]: University of California Press, 2021.

SCHÖNING, V.; LIAKONI, E.; BAUMGARTNER, C.; EXADAKTYLOS, A. K.; HAUTZ, W. E.; ATKINSON, A.; HAMMANN, F. Development and validation of a prognostic covid-19 severity assessment (cosa) score and machine learning models for patient triage at a tertiary hospital. **Journal of translational medicine**, Springer, v. 19, n. 1, p. 1–11, 2021.

SHESKIN, D. J. **Handbook of parametric and nonparametric statistical procedures**. [S.l.]: Chapman and Hall/CRC, 2003.

SLADE, E.; NAYLOR, M. G. A fair comparison of tree-based and parametric methods in multiple imputation by chained equations. **Statistics in medicine**, Wiley Online Library, v. 39, n. 8, p. 1156–1166, 2020.

SOUSA, M. R. de; RIBEIRO, A. L. P. Revisão sistemática e meta-análise de estudos de diagnóstico e prognóstico: um tutorial. **Arq. Bras. Cardiol**, v. 92, p. 241–251, 2009.

SOWMYA, V.; KAYARVIZHY, N. An efficient missing data imputation model on numerical data. In: IEEE. **2021 2nd Global Conference for Advancement in Technology (GCAT)**. [S.l.], 2021. p. 1–8.

USAMA, M.; QADIR, J.; RAZA, A.; ARIF, H.; YAU, K.-L. A.; ELKHATIB, Y.; HUSSAIN, A.; AL-FUQAHA, A. Unsupervised machine learning for networking: Techniques, applications and research challenges. **IEEE access**, IEEE, v. 7, p. 65579–65615, 2019.

VEAUX, R. D. D.; AGARWAL, M.; AVERETT, M.; BAUMER, B. S.; BRAY, A.; BRESSOUD, T. C.; BRYANT, L.; CHENG, L. Z.; FRANCIS, A.; GOULD, R. *et al.* Curriculum guidelines for undergraduate programs in data science. **Annual Review of Statistics and Its Application**, Annual Reviews, v. 4, p. 15–30, 2017.

WASKOM, M. L. seaborn: statistical data visualization. **Journal of Open Source Software**, The Open Journal, v. 6, n. 60, p. 3021, 2021. Disponível em: <https://doi.org/10.21105/joss.03021>.

WEI, Q.; JR, R. L. D. The role of balanced training and testing data sets for binary classifiers in bioinformatics. **PloS one**, Public Library of Science San Francisco, USA, v. 8, n. 7, p. e67863, 2013.

WICKHAM, H. Tidy data. **Journal of statistical software**, v. 59, n. 1, p. 1–23, 2014.

YAN, L.; ZHANG, H.-T.; GONCALVES, J.; XIAO, Y.; WANG, M.; GUO, Y.; SUN, C.; TANG, X.; JING, L.; ZHANG, M. *et al.* An interpretable mortality prediction model for covid-19 patients. **Nature machine intelligence**, Nature Publishing Group, v. 2, n. 5, p. 283–288, 2020.

YING, X. An overview of overfitting and its solutions. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2019. v. 1168, n. 2, p. 02.

ZHOU, Z.-H. A brief introduction to weakly supervised learning. **National science review**, Oxford University Press, v. 5, n. 1, p. 44–53, 2018.

ZOABI, Y.; DERI-ROZOV, S.; SHOMRON, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. **npj Digital Medicine**, v. 4, n. 1, 2021.

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. DOCUMENTO Nº	4. Nº DE PÁGINAS
DM	28 de novembro de 2022	DCTA/ITA/DM-125/2022	75

5. TÍTULO E SUBTÍTULO:

Protocolo de coleta de dados para predição de COVID-19

6. AUTOR(ES):

Alex Fernandes de Souza

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):

Instituto Tecnológico de Aeronáutica – ITA / Universidade Federal de São Paulo – UNIFESP

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:

Coleta de dados; Covid-19; Aprendizado de máquina.

9. PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

1. SARS-CoV2 2. Aquisição de dados 3. Aprendizagem (inteligência artificial) 4. Inteligência artificial 5. Computação.

10. APRESENTAÇÃO: (**X**) **Nacional** ( ) **Internacional**

ITA/UNIFESP, São José dos Campos. Curso de Mestrado. Programa de Pós-Graduação em Pesquisa Operacional. Área de Engenharia de Produção/Pesquisa Operacional. Orientador: Prof. Dr. Filipe Verri Alves Neto. Defesa em 05/10/2022. Publicada em 2022.

11. RESUMO:

A coleta de dados representa um desafio em diversos setores da sociedade. Na pandemia de Covid-19, grandes volumes de dados foram gerados com a finalidade de usá-los em tarefas de aprendizado de máquina (AM) para auxiliar na tomada de decisão. Contudo, a forma como estes dados foram coletados dificulta a elaboração de análises estatísticas e uso em tarefas de diagnóstico e prognóstico. Estas análises demandam conjuntos de dados arrumados, que representam uma forma de conectar a estrutura dos dados à sua semântica. Este estudo propõe um protocolo de coleta de dados a partir do estudo de *datasets* clínicos disponibilizados no Repositório do COVID-19 DataSharing/BR para uso em tarefas de aprendizado de máquina. Foram analisados dados do Laboratório Fleury, que apontam o diagnóstico, e dados do Hospital Sírio-Libanês, que permitem estudar o prognóstico dos casos. Ambos os *datasets* demandaram um extenso pré-processamento e, em seguida, foram arrumados para que pudessem ser utilizados em tarefas de AM. Entre os problemas observados ao longo das etapas de pré-processamento, destacam-se a falta de padronização, informações redundantes, atributos sem relevância, dados ausentes, entre outros. Após o pré-processamento inicial, ambos os conjuntos foram arrumados de modo que tornassem seu uso eficiente. Na sequência, outras tarefas foram realizadas para tornar os dados utilizáveis, eliminado, por exemplo, a extensa quantidade de valores ausentes. Com os dados arrumados, aplicou-se três técnicas preditivas de AM, sendo estes *K-Nearest Neighbor* (KNN), *Support-Vector Machine* e Árvore de decisão. Na tarefa de diagnóstico de Covid-19, a técnica KNN apresentou melhor desempenho com valores de área sob a curva ROC igual a 0.77. Para os dados de prognóstico de Covid-19, os algoritmos KNN e SVM apresentaram os melhores desempenho, ambos com 0.81 da mesma medida. A partir desses resultados, pode-se afirmar que os conjuntos de dados, dentro de uma estrutura arrumada, podem ser utilizados no auxílio ao diagnóstico e prognóstico de Covid-19. Logo, a partir do protocolo de coleta de dados proposto neste estudo, o qual garante a obtenção de dados em formato arrumado, observou-se a redução da necessidade de diversas tarefas de pré-processamento. Assim, o uso dos dados em tarefas de aprendizado de máquina e análises estatísticas é facilitado, potencializando também o manejo eficiente de pacientes e recursos hospitalares. Além disso, este protocolo pode ser utilizado em eventos futuros, facilitando a forma como os dados são coletados e seu uso subsequente.

12. GRAU DE SIGILO:

(**X**) **OSTENSIVO** ( ) **RESERVADO** ( ) **SECRETO**