Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Abordagem Baseada em Aprendizagem de Máquina para Identificar Sinais Comportamento de Depressão na Rede Social Twitter Utilizando Conteúdos das Postagens e Atividades

Luan Mendes Gonçalves Freitas

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. Marcelo Ladeira

Coorientador
Prof. Dr. Marcos Fagundes Caetano

Brasília
2022

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Abordagem Baseada em Aprendizagem de Máquina para Identificar Sinais Comportamento de Depressão na Rede Social Twitter Utilizando Conteúdos das Postagens e Atividades

Luan Mendes Gonçalves Freitas

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. Marcelo Ladeira (Orientador)
CIC/UnB

Prof. Dr. Marcos Fagundes Caetano
CIC/UnB

Dr. Rodrigo Bonifácio de Almeida
CIC/UnB

Prof. Dr. Marcelo Grandi Mandelli
Coordenador do Bacharelado em Ciência da Computação

Brasília, 06 de outubro de 2022

Dedicatória

Dedico essa monografia a todas as pessoas que fizeram parte da minha jornada ao longo desses anos de graduação na UnB. Dedico a minha mãe Gláucia Tabosa Mendes Freitas que que sempre esteve ao meu lado nos momentos difíceis. Dedico ao meu pai Wanderley Gonçalves Freitas que me incentivou a estudar e buscar conhecimento. Sem eles não seria possível chegar aqui. Também quero dedicar a minha família que me apoiaram e rezaram por mim. Obrigado por todo amor que vocês me deram. Sem ela essa graduação não seria possível. Por fim, dedico essa monografia ao meu avô Francisco Teixeira Mendes *(in memoriam)* que foi exemplo de pessoa para mim e também dedico à minha tia Anastásia Tabosa Mendes *(in memoriam)* que foi uma mãe para mim e para todos da minha família.

Agradecimentos

Em primeiro lugar, a Deus, que fez com que meus objetivos fossem alcançados, durante todos os meus anos de estudos. Aos meus familiares, especialmente a meu pai Wanderley, que tanto lutou pela minha educação e nunca me deixou perder a fé na minha capacidade de superação, pela força e amor incondicional e por todo o suporte e apoio em toda minha trajetória acadêmica. À minha mãe Gláucia, que encheu o meu coração de amor e esperança, minha gratidão por seus cuidados diários para comigo, por suas palavras de incentivo e otimismo. À minha tia, Josirene, professora de Letras e especialista em Linguagens, por todo o apoio e dedicação para que meu trabalho se concretizasse. Enfim, a todos os meus familiares que de uma forma ou de outra contribuíram para o meu sucesso. Aos professores orientadores, Prof. Dr. Marcelo Ladeira e Prof. Dr. Marcos Caetano, por terem sido meus orientadores nesse projeto. Agradeço pela confiança e incansável dedicação. Meus sinceros agradecimentos, a esses professores por todos os conselhos, pela ajuda e pela paciência com a qual ajudaram meu aprendizado. À todos os mestres do curso de Ciência da Computação, Marcelo Ladeira, Marcos Fagundes Caetano, Cláudia Nalon, Díbio Leandro Borges, Edna Dias Canedo, Eduardo Adilo Pelinson Alchieri, Fernanda Lima, Flavio Leonardo Cavalcanti Moura, Genaina Nunes Rodrigues, Marcelo Grandi Mandelli e Marcus Vinicius Lamar por enxergarem em mim muito mais do que eu imaginava ser capaz de fazer. Manifesto aqui, minha gratidão eterna por me fazerem acreditar em meu potencial e compartilhar comigo sabedoria e experiência. Aos meus colegas de curso, com quem convivi intensamente durante os últimos anos, pelo companheirismo e pela troca de experiências que me permitiram crescer não somente como pessoa, mas também como formando. À Universidade de Brasília, por me proporcionar um ambiente criativo e amigável para os estudos. Sou grato a cada membro do corpo docente, à direção e administração desta instituição.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

Um desafio, nos dias de hoje, é o de detectar e compreender sinais de transtornos depressivos em postagens de textos nas redes sociais. O projeto é baseado nos trabalhos dos pesquisadores De Choudhury et al. e Coppersmith et al. para criar um modelo para ser incorporado em uma ferramenta que seja capaz de detectar sinais de comportamento depressivo dos usuários a partir da análise das suas postagens no Twitter. São construídas duas bases de tweets, em português e de forma anômia, uma da pré-pandemia (01/01/2018 a 31/12/2019), com N=71.232 usuários e uma na pandemia (01/01/2020 a 31/12/2021), com N=70.370 usuários. Essas bases contêm usuários declarados depressivos e usuários não declarados depressivos (controle). São consideradas as seguintes questões de pesquisa: análise de cinco novos atributos na performance dos modelos e a mobilidade de usuários entre as classes depressivo e não-depressivo após a pandemia. As bases de dados são compostas por dez atributos propostos por De Choudhury et al. e os cinco novos atributos. São induzidos modelos de aprendizagem de máquina (classificadores) que são: Regressão Logística (modelo Baseline; obs.: não conta como modelo de aplicação, serve apenas para avaliar o desempenho dos outros modelos), Análise Discriminante Linear, Árvore de Decisão, Floresta Randômica, Gradient Boosting, K-ésimo Vizinho mais Próximo, Perceptron Multicamadas, Máquina de Vetores de Suporte, Naive Bayes, Bagging, Boosting, Votação Hard e Votação Soft, com intuito de identificar qual melhor modelo para identificar sinais de padrão de comportamento de depressão em postagens na rede social Twitter. Os modelos induzidos alcançam desempenhos superiores a performance de modelos propostos por De Choudhury para tweets em língua inglesa, de acordo nossa literatura, com f1-score médio de 80%. Dessa maneira, esperamos capacitar os usuários a entender melhor seus sinais e orientá-los a buscar assistência profissional sempre que necessário.

Palavras-chave: Twitter, Aprendizagem de Maquina Supervisionada, Mineração de Dados, Depressão, Não Depressão, Classificadores, Ensemble, Nuvem de Palavras

Abstract

A challenge nowadays is to detect and understand signs of depressive disorders in text posts on social networks. In Brazil, the research develops computational models that are able to detect and understand signs of disorders in text posts. These models show promising results, in addition to a variety of possible exploration and research cases. The project is based on the work of researchers De Choudhury et al. and Coppersmith et al.. In this way, we collected data from posts and user activities on the social network Twitter, extracted characteristics that we defined in our project and built two databases of users on Twitter in two different periods: a database of posts and activities from 2018 to 2019, before the outbreak of the COVID-19 pandemic, and a database of posts and activities from 2020 to 2021, during the COVID-19 pandemic. These two databases are induced in 13 supervised machine learning models, in order to identify the best model to be incorporated into a tool that is capable of identifying signs of depressive and non-depressive behavior patterns in posts on the social network Twitter. The supervised learning models are Logistic Regression (Baseline model; obs.: it does not count as an application model, it only serves to evaluate the performance of other models), Linear Discriminant Analysis, Decision Tree, Random Forest, Gradient Boosting, K-th Nearest Neighbor, Multilayer Perceptron, Support Vector Machine, Naive Bayes, Bagging, Boosting, Hard Voting and Soft Voting. To achieve these goals, we extract traits by measuring writing patterns (e.g. language styles, emojis, oriental characters (Japanese, Chinese and Korean) and depressive terms and anti-depressant medications) and Twitter activity history (e.g. number of tweets, likes and comments) in user posts on the social network Twitter. The resulting models successfully distinguish between depressive and non-depressive classes, with performance results comparable to the results of our literature, with an average f1-score of 80%. In this way, we hope to empower users to better understand their signals and guide them to seek professional assistance whenever necessary.

Keywords: Twitter, Supervised Machine Learning, Data Mining, Depression, Non-Depression, Classifiers, Ensembles, Word Cloud

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ANEW Affective Norms for English Words.

API Application Programming Interface.

CIC Ciência da Computação.

CSV Comma-Separated Values.

DM Data Mining.

DT Decision Tree.

EDA Exploratory Data Analysis.

GB Gradient Boosting.

KNN k-Nearest Neighbors.

LDA Linear Discriminant Analysis.

LGPD Lei Geral de Proteção de Dados Pessoais.

LIWC Linguistic Inquiry and Word Count.

LR Logistic Regression.

ML Machine Learning.

MLP Multilayer Perceptron.

NB Naive Bayes.

NLP Natural Language Processing.

OMS Organização Mundial da Saúde.

PCA Principal Component Analysis.

RBF Radial Basis Function.

RF Random Forest.

SVM Support Vector Machine.

UnB Universidade de Brasília.

Capítulo 1

Introdução

As doenças mentais, entre elas a depressão, tornou-se uma das principais causas de desequilíbrio emocional em todo o mundo [6]. Esse transtorno mental traz prejuízos para a qualidade de vida de um indivíduo, nos campos familiar, estudantil, trabalhista e nas relações interpessoais; e se não tratado, pode inclusive desencadear o surgimento de outras enfermidades, culminando em consequências drásticas e imprevisíveis, decorrentes do agravamento das fragilidades emocionais na pessoa acometida por essa doença [7].

Segundo a Organização Mundial da Saúde (OMS), 300 milhões de pessoas sofrem de depressão ou algum tipo de problema mental em determinada fase de sua vida [8]. Dentre as doenças mentais existentes, tais como, esquizofrenia, síndrome do pânico, distúrbio de ansiedade generalizada [9] entre outras, a depressão é a quarta principal doença de origem psiquiátrica mais incapacitante em todo o mundo.

Em geral, a dor que vem com a depressão não afeta apenas aqueles que sofrem desse transtorno, mas ela afeta também todos aqueles que estiverem ao seu redor; portanto, diante do exposto, essa doença é um problema grave, com que todos nós devemos nos preocupar. Como identificar sinais de comportamento depressivo ou não depressivo nas mídias sociais *Twitter* usando aprendizagem de máquina?

Nos dias de hoje, as pessoas estão usando as plataformas de mídia social, como *Twitter*, *Facebook*, *Google+*, entre outras redes, para compartilhar pensamentos e opiniões diários com seus amigos, familiares ou conhecidos [10]. A análise das mensagens divulgadas em seus perfis sociais pode servir de base para obter evidências sobre os seus comportamentos.

Usando-se do conhecimento e técnicas de Mineração de Dados (em inglês: *Data Mining* DM) [11, 12] e de Processamento de Linguagem Natural (em inglês: *Natural Language Processing* NLP) [13] juntamente com a literatura de pesquisa em psicologia, psiquiatria, neurociência e sociolinguística sobre saúde humana, tornou-se possível prever possíveis sinais de padrões de comportamento de depressão [14] e transtorno de estresse pós-traumático [15]. A relação que muitos têm com seus perfis sociais pode levar a com-

preensão sobre o comportamento humano. Torna-se possível prever possíveis padrões de depressão [14] e transtorno de estresse pós-traumático [15], entre outras doenças mentais.

Pesquisadores, por fortes alegações sobre o poder da mineração de dados na saúde mental, continuamente adicionam uma infinidade de métodos que podem ajudar a identificar e a prever doenças mentais, com desempenho satisfatório e explicável. Por exemplo, os modelos ocultos de Markov [16] e *Word Shift Graphs* [17] foram empregados com sucesso para se tornar mais sustentável nas decisões dos modelos [18]. Nesse contexto de transtorno mental de depressão De Choudhury et al. [14] propôs um método para detecção de sinais de comportamento depressivo, baseado em atributos construídos a partir dos conteúdos de postagens e das atividades dos usuários do *Twitter*, tornando possível classificá-los como tendo evidência de serem depressivos ou não depressivos.

1.1 Motivação

As plataformas de mídia social, tais como o *Twitter*, são utilizadas diariamente para compartilhar pensamentos e opiniões. Como consequência, surge uma enorme quantidade de dados armazenados na web. Com o intuito de identificar padrão de comportamento depressivo ou não-depressivo em um indivíduo, tomou-se como referência os métodos e estratégias dos pesquisadores Choudhury [14] e Coppersmith [15] no apoio na identificação de padrões significativos nas postagens e atividade na rede social *Twitter*, utilizando-se para isso dos modelos dos algoritmos classificadores de aprendizagem de máquina supervisionados [2].

Nesse projeto foi formulado três questões de pesquisa que serão respondidas ao longo da pesquisa:

- **Questão 1** - Como identificar sinais de comportamento depressivo em postagens e atividades de usuários na rede social *Twitter*, usando conceito e técnicas de aprendizagem de máquina?
- **Questão 2** - Como identificar sinais de comportamento depressivo através de um grafo social de seguidores na rede social *Twitter*?
- **Questão 3** - Como saber se há evidências de aumento de pessoas com sintomas de depressão na pandemia?

1.2 Objetivos

Este projeto foi elaborado para fins de pesquisa sem intenções lucrativas, com o propósito de aplicar conhecimento adquirido ao longo do curso Ciência da Computação (CIC) na Universidade de Brasília (UnB). Os objetivos gerais e específicos serão explicados nas seções a seguir.

1.2.1 Objetivo Geral

Esse projeto visa detectar sinais padrão de comportamento depressivo e não-depressivo em um usuário no *Twitter* [19], conforme os trabalhos dos pesquisadores De Choudhury et al. [14] e Coppersmith et al. [15], utilizando-se métodos de Mineração de Dados para construção de modelos de aprendizagem de máquina supervisionado [2, 13].

De Choudhury et al. [14] e Coppersmith et al. [15] utilizaram *tweets* em inglês. As contribuições dessa pesquisa são analisar *tweets* em português, avaliar a mobilidade entre os grupos depressivos e não depressivos antes da pandemia COVID-19 (2018 a 2019) e durante a pandemia (2020 a 2021), e avaliar cinco novos possíveis atributos. Os possíveis atributos considerados são a quantidade nos *tweets* de: caracteres orientais (japonês, chinês e coreano), *emojis*, links, mídia (fotos, vídeos e *gifs*), e curtidas.

1.2.2 Objetivos Específicos

Para alcançar objetivo especificado, o projeto foi subdividido em 3 objetivos que deverão ser cumpridos ao longo da pesquisa:

1. Aplicar métodos e técnicas de Mineração de Dados para a construção de 2 bases de dados de *tweets* em português para 2 períodos: pré-pandemia COVID-19 (2018-2019) e pandemia COVID-19 (2020-2021);
2. Utilizar métodos de aprendizagem de máquina supervisionados para gerar um modelo que será incorporado a uma ferramenta de análise de sentimentos capaz de identificar evidências de possíveis sinais de padrões de comportamento depressivo e não depressivo;
3. Realizar uma análise comparativa dos resultados obtidos com os resultados da pesquisadora De Choudhury et al. [14], se melhora ou não as predições dos algoritmos.

1.3 Estrutura do Documento

Para fins de compreensão, apresenta-se, a seguir, a estrutura da dissertação de conclusão de curso.

- O **Capítulo 2**, fornece uma visão geral dos conceitos teóricos e a pesquisa no *Twitter*, mineração de dados e aprendizagem de máquina supervisionada.
- O **Capítulo 3**, apresenta a metodologia utilizada, a partir de coleta de dados, da extração de características, do cálculo de vetores de características, da construção da base de dados, dos modelos de aprendizado de máquina e das métricas de avaliação.
- O **Capítulo 4**, apresenta os resultados da exploração de dados nos *tweets* coletados nos 2 períodos pré-pandemia (2018-2019) e pandemia (2020-2021), visualização dos termos mais comum nos 2 períodos através da nuvem de palavras e a performance dos modelos supervisionados nos dados de validação e de teste das 2 bases de dados.
- O **Capítulo 5**, trata das ameaças à validade, que podem ocorrer nessa área de pesquisa.
- O **Capítulo 6**, apresenta as conclusões da monografia, ideias para futuros trabalhos no intuito de estender as bases de dados, melhorar os algoritmos e descobertas de conhecimentos.

No próximo capítulo, serão aprofundados todos os conceitos que cercam essa pesquisa. A maioria dos conceitos, foram usados como ferramentas de apoio de esclarecimento às pessoas que trabalham nas áreas de Ciência e Tecnologia. Para as pessoas que não são dessas áreas técnicas, pede-se um pouco de esforço para compreender ao máximo possível os conceitos, mas se não for possível sugiro que passe para o capítulo seguinte, sem prejuízo do entendimento essencial proposto nessa pesquisa.

Capítulo 2

Referencial Teórico

Neste capítulo, será discutida a literatura de trabalhos relacionados, que serviram de apoio para a elaboração desse projeto na compreensão de como a aprendizagem de máquina pode ajudar a detectar padrões de depressão nas mídias sociais. Em seguida, apresenta-se uma visão geral de definições e de exemplos das técnicas utilizadas no projeto.

Na *Seção* 2.1 apresentam-se trabalhos relacionados ao projeto que serviram de base para o desenvolvimento desse trabalho. Na 2.2 descreve-se o principal local de busca e coleta de dados da rede social Twitter; seu funcionamento e estrutura de dados, juntamente com sua API oficial, descrita na *Seção* 2.2.1 que trata do detalhamento de suas funcionalidades. Na *Seção* 2.3 explanam-se os conceitos de ciência de dados e onde podem ser aplicados. Na *Seção* 2.3 relaciona-se o que é e como são empregadas as técnicas de mineração de dados.

Na *Seção* 2.5 explicam-se os conceitos de análise exploratória de dados e como usá-los. Na *Seção* 2.6 exemplifica-se sobre base de dados com seus tipos existenciais. Na *Seção* 2.7 exemplificam-se os métodos essenciais para como tratar e como dividir a base de dados, antes de aplicar em algoritmos de aprendizagem de máquina. Na *Seção* 2.8 apresentam-se os conceitos e técnicas, de forma geral, de aprendizagem de máquina, com seus principais tipos. Na *Seção* 2.9 explica-se quais algoritmos foram escolhidos com seus conceitos e técnicas para aplicar no projeto.

Na *Seção* 2.10 descrevem-se as métricas de avaliação para computar a performance dos modelos de algoritmos de aprendizagem de máquina e na *Seção* 2.11 ilustram-se dois casos em que um modelo de aprendizagem de máquina pode (não deve) resultar e como tratá-los. Na *Seção* 2.12 especificam-se as técnicas de dimensionalidade dos dados em base de dados para melhorar as predições dos modelos de aprendizagem de máquina. Na *Seção* 2.13 apresenta-se a legislação brasileira *LGPD* (Lei Geral de Proteção de Dados Pessoais) constituída de normas referentes às atividades de tratamento de dados pessoais, as quais devem ser seguidas para proteger dados e informações das pessoas.

2.1 Trabalhos Relacionados

Desde o início dos anos 2000, tem havido esforços crescentes em alavancar o poder da tecnologia para ajudar na compreensão e na prevenção de doenças mentais. Muito tem sido realizado nas últimas duas décadas, a partir da análise computadorizada de textos escritos, os quais revelaram pistas preditivas sobre tendências neuróticas e transtornos psiquiátricos [20], pistas essas que apoiam a alegação de que o processamento negativo (cognitivo) tende a resolver problemas ambíguos detectados na informação verbal e que pode predizer a depressão subsequente [21]. No entanto, o boom das mídias sociais, no início de 2010, trouxe com ele um fluxo cada vez maior de dados, que permitiu aos pesquisadores obter mais *insights* de sinais correlacionados à depressão e a outras doenças mentais. Por meio do uso de dados com técnicas de Mineração de Dados (DM), a pesquisa mostra que os padrões de comportamento podem ser combinados com eventos do mundo real [22], e que sinais sintomáticos de transtorno depressivo maior podem ser observados a partir de atualizações de status no *Facebook* [23].

Quando se trata de *Twitter*, Park et al. [24] encontra evidências de que as pessoas postam sobre sua depressão e seu tratamento na plataforma, e De Choudhury et al. [14] informa classificadores induzidos (precisão = 0,742, recall = 0,629, F1 = 0,681) para estimar o risco de depressão, antes de seu início, medindo atributos comportamentais relacionados à engajamento, emoção, estilos linguísticos, rede social e menções a medicamentos antidepressivos. Coppersmith et al. [15] propõe heurísticas para automatizar partes da construção do conjunto de dados, que rendeu, apenas para a depressão, um conjunto de dados muito maior do que o que havia sido alcançado anteriormente, além de ampliar o escopo para outras doenças.

Resnik et al. [25] aplica uma variedade de Modelos de Tópicos Supervisionados nos dados set que criou Coppersmith [15] e obteve resultados promissores (AUC = 0,860) ao classificar grupos depressivos versus não depressivos. A pesquisa seguinte questiona os métodos empregados por De Choudhury [14] e Coppersmith [15] e argumenta que métodos são necessários para apoiar alegações mais fortes que os dados do Twitter realmente permitem detectar a previsão de depressão, conforme argumentado por Reece et al. [17]. Interpretabilidade de modelos é mais um forte argumento feito por Reece, e gráficos de deslocamento de palavras juntas com o Modelo oculto de Markov [16] são empregados (precisão = 0,852, recall = 0,518, F1 = 0,644) como uma alternativa que parece identificar signos não totalmente capturados pela *LIWC* (Linguistic Inquiry and Word Count) [26] ou *ANEW* (Affective Norms for English Words) [27]. Dito tudo isso, é bastante improvável que as conclusões de Reece que considera inválidos os métodos de De Choudhury [14] e Coppersmith [15] estejam abertas ao debate. Por outro lado, deve ser visto a importância da interação dos métodos, de modo a produzir reivindicações mais robustas acerca da

análise da captura da natureza da mente humana por meio da mídia social.

O forte argumento deste trabalho científico é o de direcionar a pesquisa para a previsão de chance de detectar casos de doença mental em conjunto de dados em análise, além de possibilitar uma ferramenta do mundo real que permita ajudar médicos, psicólogos e profissionais de saúde a entenderem melhor seus pacientes e os padrões que eles compartilham. Em suma, com o presente trabalho, espera-se dar os primeiros passos em direção ao objetivo de constatar aqueles que precisam agora e não mais tarde, e atrair mais pesquisas para este desafio em nosso país de origem, mostrando que, embora as pessoas e culturas diferem em inúmeras maneiras, as demandas e exigências da vida contemporânea são as mesmas, como gestão de tempo, de produtividade e das emoções, de forma cada vez mais absoluta, o que pode acarretar o definhamento da saúde emocional somado a níveis de estresse que causam o adoecimento do ser humano, revelados por comportamentos que evidenciam sofrimento emocional intenso, como quadros de depressão; comportamentos afastados e instáveis; mudanças drásticas de humor sem motivos aparentes; desesperança, avisos verbais que implicam uma busca inconsciente de ajuda. Nesse sentido alavancar a tecnologia na identificação de casos que necessitem de acompanhamento específico em favor do bem-estar e da saúde mental dessas pessoas é realmente apenas uma questão de coragem e mentes abertas na busca de desmistificar a saúde mental para as pessoas nos ambientes de trabalho, doméstico e social; contribuindo nesse contexto para os encaminhamentos decorrentes dessa pesquisa, o que provocará a redução do preconceito que cerca a doença depressão, além de contribuir na detecção dos diagnósticos para conscientização de tratamentos específicos, o que motivará um impacto direto na produtividade e na saúde física, mental e social das pessoas.

von Sperling [11, 12] extrai atributos comportamentais relacionados ao engajamento, emoção, e estilos linguísticos a partir de *tweets* em português para criar uma base de dados de usuários do Twitter que foram diagnosticados com depressão. De Choudhury et al. [14] constrói classificadores (precisão = 0,742, recall = 0,629, F1 = 0,681) para identificar usuários com possíveis sinais de depressão, mensurando atributos comportamentais relacionados ao engajamento, emoção, estilos linguísticos, rede social, e menções a medicamentos antidepressivos. Coppersmith et al. [15] propõe heurísticas para automatizar partes da construção do conjunto de dados para a depressão, resultando em um conjunto de dados muito maior do que o utilizado por De Choudhury et al., além de ampliar o escopo para outros transtornos.

2.2 Twitter

A Rede Social *Twitter* atualmente é uma das principais redes sociais, na qual seus usuários compartilham e trocam mensagens entre si. As mensagens que alimentam a rede social é limitada em 280 caracteres, dos quais 20 deles são reservados para o usuário originário e os outros 140 caracteres são destinados à mensagem de texto. Essas mensagens, dentro da rede social, recebem o nome de *tweet* e nelas podem existir mensagens de texto puro, imagens, links para endereços externos e vídeos curtos. Os vídeos e imagens postados são apresentados na linha de tempo da página principal do usuário logado, conforme pode ser visto na Figura 2.1, onde se pode verificar a estrutura que um *tweet* possui, apresentando os textos, os links externos e imagens.

Figura 2.1: Exemplo de um Tweet no Twitter[1]

O Twitter possui o sistema de seguidor, onde se tem como objetivo um usuário acompanhar as publicações dos outros usuários, mesmo que esses usuários não acompanhem as publicações deste usuário. Estrutura esta que não exige que os dois lados de uma conexão precisem ser efetuados para que a troca de informações entre eles possa existir. Ele possui também o sistema de listas, onde pode-se criar uma lista de pessoas que se deseja acompanhar, sem ter a necessidade de criar uma conexão de "seguidor"com este usuário. Além disso, existem as palavras-chave, que são chamadas de *hashtags*, onde as mensagens com

[1] `https://developer.twitter.com/`

essas *hashtags* possuem a facilidade de serem encontradas através de buscas. As *hashtags* mais utilizadas formam o *ranking* que é chamado de *Trend Topics*.

Na página inicial do usuário são exibidas as mensagens em tempo real, assim como as mensagens mais próximas do horário de acesso, seguindo uma classificação decrescente do horário de postagem. Dentre essas mensagens estão as publicações do próprio usuário e as publicações dos usuários que são seguidos pelo mesmo. Essas mensagens podem ser repassadas para os próprios seguidores, através de *retweets*, assim como podem ser marcadas com estrela, podendo ser visualizadas facilmente mais tarde. A rede social também tem disponível uma página de busca, que facilita localizar usuários e/ou assuntos específicos, que utilizam de palavras-chave e *hashtags* para agilizar a localização das informações.

2.2.1 Twitter API

A Interface de Programação de Aplicações (em inglês: *Application Programming Interface* (API)) do Twitter pode ser usada para recuperar e analisar dados do Twitter de forma programática, bem como criar um ambiente para a conversa no Twitter.

Ao longo dos anos, a *API* do Twitter v1 cresceu agregando níveis adicionais de acesso para que desenvolvedores e pesquisadores acadêmicos pudessem dimensionar seu acesso para aprimorar e pesquisar a conversa pública. Existem dois modelos da *API* do Twitter v1:

- **Standard v1.1**: As *APIs* padrão gratuitas são ótimas para começar, testar uma integração, validar um conceito ou criar soluções que complementam o que você pode criar com produtos *premium* e corporativos. Os exemplos, incluem postar conteúdo no *Twitter* e recuperar dados semelhantes ao que está no twitter.com e no aplicativo móvel do *Twitter*.
- **Premium v1.1**: As *APIs Premium* oferecem acesso escalável aos dados do Twitter para quem deseja crescer, experimentar e inovar. As *APIs Premium* dão aos desenvolvedores acesso rápido a recursos de qualidade empresarial, como funcionalidade de pesquisa aprimorada e atividade de conta baseada em *webhook*, para continuar construindo e aumentando o uso ao longo do tempo. As *APIs Premium* incluem contratos mensais flexíveis e níveis de acesso escalonados com base no número de solicitações. As *APIs Premium também incluem acesso limitado em um sandbox gratuito.*

Recentemente, foi lançada a *API* do *Twitter* v2. A API do *Twitter* v2 inclui uma base moderna, recursos novos e avançados e integração rápida ao acesso Essential. A *API* do *Twitter* v2 representa a maior atualização da *API* do *Twitter* desde 2012. Com ela, vem

uma série de recursos novos e avançados, além de acesso rápido e gratuito à *API*. Alguns dos recursos disponíveis com a v2 incluem o seguinte:

- Novos *endpoints*;
- Objetos de dados novos e mais detalhados;
- Novos parâmetros para ajudá-lo a recuperar apenas os objetos e campos que você deseja;
- Métricas avançadas;
- Filtrar e identificar quais *tweets* contêm tópicos diferentes;
- Filtrar e identificar quais *tweets* pertencem a um tópico de resposta;
- Um nível de acesso específico para pesquisadores acadêmicos;
- Filtragem de spam de alta confiança;
- *URLs* encurtados são totalmente desenrolados para facilitar a análise de URL;
- Objetos de resposta *JSON* simplificados removendo campos obsoletos e modernizando rótulos;
- Funcionalidade de recuperação e redundância para nossos *endpoints* de *streaming*;
- Retorno de 100% dos *tweets* públicos e disponíveis correspondentes nas consultas de pesquisa;
- "Regras"de *streaming* para que você possa fazer alterações sem perder conexões;
- Linguagem de consulta mais expressiva para fluxo e pesquisa filtrados;
- Especificação *OpenAPI* para criar novas bibliotecas e acompanhar as alterações de forma mais transparente;
- Suporte de *API* para novos recursos e *endpoints* mais rapidamente, à medida que a plataforma evolui para atender às necessidades de desenvolvedores, pesquisadores, empresas e pessoas que usam o *Twitter*.

Nas operações de busca e coleta de objetos nas *APIs* v1 e v2 do *Twitter* existem operadores de busca que podem ser usados para objetivos específicos, conforme as Tabelas 2.1 à 2.9. Esses operadores também podem trabalhar na *Web*, *Mobile* e *Tweetdeck*. Há alguma sobreposição, mas em grande parte desses operadores pode não funcionar para *APIs* v1 e v2 do *Twitter*.

Tabela 2.1: Operadores de Pesquisa Avançada no Twitter Para Conteúdo no Tweet [3]

Classe	Operador	Encontra Tweets...
Conteúdo do Tweet	nasa esa (nasa esa)	Contendo "nasa"e "esa". Os espaços são AND implícitos. Os colchetes podem ser usados para agrupar palavras individuais se estiver usando outros operadores.
	nasa OR esa	Ou "nasa"ou "esa". OR deve estar em letras maiúsculas.
	"state of the art"	A frase completa "state of the art". Também corresponderá ao "start of the art". Use também aspas para evitar a correção ortográfica.
	"this is the * time this week"	Uma frase completa com um curinga. * não funciona fora de uma frase entre aspas ou sem espaços.
	+radiooooo"	Forçar um termo a ser incluído como está. Útil para evitar a correção ortográfica.
	-love -"live laugh love"	- é usado para excluir "love". Também se aplica a frases entre aspas e outros operadores.
	#tgif	Uma hashtag
	$TWTR	Um cashtag, como hashtags, mas para símbolos de ações
	What ?	Os pontos de interrogação são combinados
	:) OR :(	Alguns emoticons são combinados, positivos :) :-) :P :D ou negativos :-(:(
	emoji	As pesquisas de emoji também são correspondidas. Geralmente precisa de outro operador para trabalhar.
	url:google.com	urls são tokenizados e combinados, funciona muito bem para subdomínios e domínios, não tão bem para urls longas, depende da url. Os ids do Youtube funcionam bem. Funciona para urls encurtados e canônicos, por exemplo: gu.com encurtador para theguardian.com. Ao pesquisar domínios com hífens, você deve substituir o hífen por um sublinhado (como url:t_mobile.com), mas os sublinhados _ também são tokenizados e podem não corresponder
	lang:en	Pesquise tweets no idioma especificado, nem sempre preciso.

Tabela 2.2: Operadores de Pesquisa Avançada no Twitter Para Usuários [3]

Classe	**Operador**	**Encontra Tweets...**
Usuário	from:user	Enviado por um determinado @username, por exemplo "dogs from:NASA"
	to:user	Respondendo a um @username específico
	@user	Mencionar um @username específico. Combine com -from:username para obter apenas menções
	list:108534289 list:user/list-slug	Tweets de membros desta lista pública. Use o ID da lista da API ou com URLs como https://twitter.com/i/lists/4143216. O slug de lista é para URLs de lista antigos como http://twitter.com/nasa/lists/astronauts. Não pode ser negado, então você não pode pesquisar por "not on list".
	filter:verified	De usuários verificados
	filter:follows	Somente de contas que você segue
	filter:social filter:trusted	Somente a partir de uma rede de contas expandida por algoritmos com base em seus próprios seguidores e atividades. Funciona em resultados "Top"e não em "Latest"

Tabela 2.3: Operadores de Pesquisa Avançada no Twitter Para Geolocalização [3]

Classe	**Operador**	**Encontra Tweets...**
Geolocalização	near:city	Georreferenciado neste lugar. Também suporta frases, por exemplo: near:"The Hague"
	near:me	Perto de onde o twitter pensa que você está
	within:radius	Dentro do raio específico do operador "perto", para aplicar um limite. Pode usar km ou mi, por exemplo incêndio perto de:san-francisco dentro de:10km
	geocode:lat,long,radius	Por exemplo, para obter tweets a 10km da sede do twitter, use geocode:37.7764685,-122.4172004,10km
	place:96683cc9126741d1	Pesquisar tweets por ID de objeto de local, por exemplo: ID de local dos EUA é 96683cc9126741d1

Tabela 2.4: Operadores de Pesquisa Avançada no Twitter Para Tempo [3]

Classe	Operador	Encontra Tweets...
Tempo	since:yyyy-mm-dd	Em ou após (inclusive) uma data especificada
	until:yyyy-mm-dd	Antes (NOT inclusive) de uma data especificada. Combine com um operador "since"para datas entre.
	since_time:1142974200	Em ou após um timestamp unix especificado em segundos. Combine com o operador "until"para datas entre. Talvez mais fácil de usar do que since_id abaixo.
	until_time:1142974215	Antes de um timestamp unix especificado em segundos. Combine com um operador "since"para datas entre. Talvez mais fácil de usar do que max_id abaixo.
	since_id:tweet_id	Após (NOT inclusive) um ID de floco de neve especificado
	max_id:tweet_id	Em ou antes (inclusive) de um ID de floco de neve especificado
	within_time:2d within_time:3h within_time:5m within_time:30s	Pesquisar no último número de dias, horas, minutos ou segundos

Tabela 2.5: Operadores de Pesquisa Avançada no Twitter Para Tipo de Tweet [3]

Classe	Operador	Encontra Tweets...
Tipo de Tweet	filter:nativeretweets	Apenas retuítes criados usando o botão de retuíte. Funciona bem combinado com from: para mostrar apenas retuítes.
	include:nativeretweets	Retweets nativos são excluídos por padrão. Isso os mostra. Ao contrário do filter:, que mostra apenas retuítes, inclui retuítes além de outros tweets
	filter:retweets	Retweets de estilo antigo ("RT") + tweets citados.
	filter:replies	Tweet é uma resposta a outro Tweet. Bom para encontrar conversas ou tópicos se você adicionar ou remover to:user
	conversation_id:tweet_id	Tweets que fazem parte de uma conversa (respostas diretas e outras respostas)
	filter:quote	Conter Tweets de Citação
	quoted_tweet_id:tweet_id	Pesquisar citações de um tweet específico
	quoted_user_id:user_id	Pesquise todas as cotações de um usuário específico
	card_name:poll2choice_text_only card_name:poll3choice_text_only card_name:poll4choice_text_only card_name:poll2choice_image card_name:poll3choice_image card_name:poll4choice_image	Tweets contendo enquetes. Para enquetes contendo 2, 3, 4 opções ou enquetes de imagem.

Tabela 2.6: Operadores de Pesquisa Avançada no Twitter Para Engajamento [3]

Classe	**Operador**	**Encontra Tweets...**
Engajamento	filter:has_engagement	Tem algum engajamento (respostas, curtidas, retuítes). Pode ser negado para encontrar tweets sem engajamento.
	min_retweets:5	Um número mínimo de retuítes. As contagens parecem ser aproximadas para valores maiores (1000+).
	min_faves:10	Um número mínimo de curtidas
	min_replies:100	Um número mínimo de respostas
	-min_retweets:500	Um número máximo de retuítes
	-min_faves:500	Um número máximo de curtidas
	-min_replies:100	Um número máximo de respostas

Tabela 2.7: Operadores de Pesquisa Avançada no Twitter Para Mídia [3]

Classe	**Operador**	**Encontra Tweets...**
Mídia	filter:media	Todos os tipos de mídia.
	filter:twimg	Imagens nativas do Twitter (links pic.twitter.com)
	filter:images	Todas as imagens.
	filter:videos	Todos os tipos de vídeo, incluindo vídeo nativo do Twitter e fontes externas, como Youtube.
	filter:periscope	Periscópios
	filter:native_video	Todos os tipos de vídeo de propriedade do Twitter (vídeo nativo, vine, periscope)
	filter:vine	Vinhas (RIP)
	filter:consumer_video	Apenas vídeo nativo do Twitter
	filter:pro_video	Apenas vídeo profissional do Twitter (Amplify)
	filter:spaces	Apenas espaços do Twitter

Tabela 2.8: Operadores de Pesquisa Avançada no Twitter Para Mais Filtros [3]

Classe	**Operador**	**Encontra Tweets...**
Mais Filtros	filter:links	Contendo apenas algum URL, inclui mídia, use -filter:media para urls que não são mídia
	filter:mentions	Contendo qualquer tipo de @menções
	filter:news	Contendo link para uma notícia. Combine com um operador de lista restringir ainda mais o conjunto de usuários.
	filter:safe	Excluindo conteúdo NSFW. Exclui o conteúdo que os usuários marcaram como "Potencialmente sensível". Nem sempre garante os resultados do SFW.
	filter:hashtags	Apenas Tweets com Hashtags.

Tabela 2.9: Operadores de Pesquisa Avançada no Twitter Para Aplicativo Específico [3]

Classe	Operador	Encontra Tweets...
Aplicativo Específico	source:client_name	Enviado de um cliente específico, por exemplo source:tweetdeck ex: twitter_ads não funciona sozinho, mas funciona com outro operador.
	card_domain:pscp.tv	Corresponde ao nome de domínio em um cartão do Twitter. Principalmente equivalente a url: operador.
	card_url:pscp.tv	Corresponde ao nome de domínio em um cartão, mas com resultados diferentes para card_domain.
	card_name:audio	Tweets com um Player Card (Links para fontes de áudio, Spotify, Soundcloud etc.)
	card_name:animated_gif	Tweets com GIFs
	card_name:player	Tweets com um cartão de jogador
	card_name:app card_name:promo_image_app	Tweets com links para um App Card promo_app não funciona, promo_image_app é para um link de aplicativo com uma imagem grande, geralmente postado em anúncios.
	card_name:summary	Apenas cartões de resumo de imagem pequena
	card_name:summary_large_image	Apenas cartões de imagem grandes
	card_name:promo_website	Maior que summary_large_image, geralmente postado por meio de anúncios
	card_name:promo_image_convo card_name:promo_video_convo	Encontra anúncios de conversa
	card_name:3260518932:moment	Encontra cartas de Momentos. 3260518932 é o ID do usuário do @TwitterMoments, mas a pesquisa encontra momentos para todos, não para esse usuário específico.

2.3 Ciência de Dados

A Ciência de Dados (em inglês: *Data Science*) é um campo de estudo que trata de obter conhecimento e informação, a partir de dados estruturados ou não, utilizando métodos e modelos computacionais e estéticos, visando auxiliar na tomada de decisões de instituições [28]. Ainda que o nome ciência de dados vincula intensamente a ideia do estudo de banco de dados, ciência da computação, na área deste campo de estudo também abrange a necessidade de habilidades não matemáticas; as fundamentais habilidades destacadas pelo

um breve estudo de caso; por exemplo, incluem habilidades de comunicação, habilidades de análise de dados e habilidades de raciocínio [29].

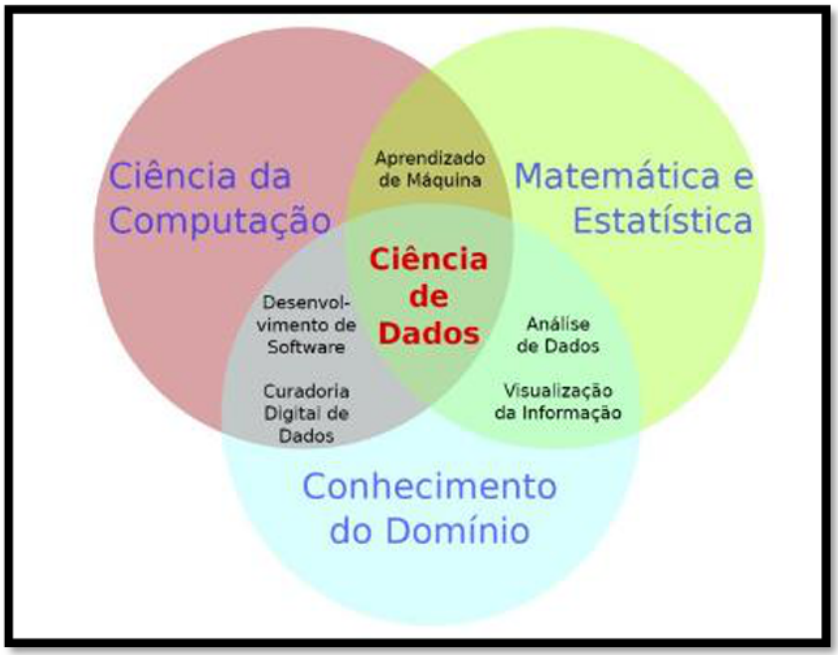
Dentre os objetivos dessa área pode-se destacar o auxílio, na tomada de decisão baseada em dados, para isto são necessárias diversas competências. Uma analogia para melhor entendimento do que é ciência de dados seria visualizá-la como uma "caixa de ferramentas", de onde se procura por habilidades de modelagem de dados, para manipulação dos dados não estruturados e em larga quantidade. O aprendizado de máquina, para utilizar a verificação de padrões e realizar abstrações profundas; a matemática e estatística, para desenvolver modelos e distribuições sólidas; e a análise de dados, para conseguir identificar tendências, tanto de mercado como de comportamento, para assim auxiliar na tomada de decisão.

Nos domínios do conhecimento da ciência de dados, conforme abordado por Rautenberg e Carmo [30], são ressaltadas habilidades na área da ciência da computação, tendo como princípio fundamental o armazenamento de dados e o processamento para computadores. Nessa situação, os ambientes computacionais são ferramentas essenciais para a implementação do aprendizado de máquina para acarretar a curadoria digital e para estabelecer interfaces de visualização da informação. É importante saber como utilizar tecnologias: como acessar e como converter os dados na possível abstração de uma informação útil.

Conhecimento na área de Matemática e Estatística é necessário para desenvolver atividades de Análise de Dados, ou seja, os profissionais que atuam na ciência de dados devem conhecer bem a forma com que os Algoritmos de Aprendizagem de Máquina funcionam, assim também como interpretar os resultados destes algoritmos estatisticamente. Percebe-se como a interpretação é facilitada de forma interdisciplinar, na visualização e compreensão da informação.

O conhecimento no que se refere ao domínio do problema, deve ser amplamente utilizado na abordagem do processo de Tomada de Decisão, para se conseguir obter o devido sucesso das possíveis soluções de Ciência de Dados. Dessa forma, as soluções de Ciência de Dados são designadas para que sejam formuladas hipóteses e informações que se adquiram ao longo do processo de decisão. Como mostrado na Figura 2.2, no que se refere à interdisciplinaridade na Ciência de Dados.

Figura 2.2: Interdisciplinaridade da Ciência de Dados



2.4 Mineração de Dados

A Mineração de Dados (em inglês: *Data Mining* DM) é o processo de extrair informação válida, previamente desconhecida, a partir de grandes bases de dados [31]. Os dois principais objetivos em Mineração de Dados tendem a ser a predição e a descrição de dados. A predição envolve o uso de um conjunto de variáveis, para prever os variáveis desconhecidos ou de uma outra variável objetiva. E a descrição dos dados consiste na descoberta de padrões que os descrevam.

Em Mineração de Dados, existem dois tipos de aprendizagem: a supervisionada e a não supervisionada. A primeira, diz respeito às tarefas de mineração em que se tem um conjunto de treinamento com os dados rotulados e a partir desses dados é gerado um modelo classificador que, em um segundo momento, será usado para classificar um novo registro não rotulado, essa aprendizagem é a preditiva. Já na aprendizagem não-supervisionada, ocorre uma descrição do conjunto de dados de modo a obter algum conhecimento oculto.

A Seguir há uma simples explicação de algumas das tarefas de mineração:

- **Classificação**: técnica de mineração de dados supervisionada preditiva, ou seja, a partir de dados rotulados se gera um classificador que pode ser uma árvore de decisão, *Naive Bayes* etc. Exemplos de algoritmos: *J48* e *ID3* (para árvores de decisão).

- **Agrupamento (Clusterização)**: os objetos são agrupados de modo que a semelhança seja máxima dentro de cada *cluster* e mínima entre instâncias de *clusters* diferentes. Exemplos de algoritmos: *k-means*, *isodata*, *fuzzy C-Médias* (lógica nebulosa).
- **Regras de Associação**: assim como no agrupamento, possuem algoritmos com aprendizagem não supervisionada. Essa tarefa de mineração busca regras sobre relações e coocorrências em bases de dados, por exemplo, regras do tipo: se X ocorre na base de dados, então Y também ocorre (com alguma relação à X). Esse método é muito usado para verificar associações em tabelas de transações. Exemplo de algoritmo: *Apriori*.

2.5 Análise Exploratória de Dados

A Análise Exploratória de Dados (em inglês: *Exploratory Data Analysis* (EDA)) é o processo de investigar o conjunto de dados para descobrir padrões e anomalias (outliers) e formar hipóteses com base em nossa compreensão do conjunto de dados [32]. A Análise Exploratória de Dados envolve a geração de estatísticas resumidas para dados numéricos no conjunto de dados e a criação de várias representações gráficas para entender melhor os dados.

2.6 Base de Dados (Dataset)

Uma Base de Dados ou *Dataset* é um conjunto ou coleção de dados. Esse conjunto é normalmente apresentado em um padrão tabular. Cada coluna descreve uma variável específica. E cada linha corresponde a um determinado membro do conjunto de dados, conforme a pergunta dada. Isso faz parte do gerenciamento de dados. Os conjuntos de dados descrevem valores para cada variável em quantidades desconhecidas, como altura, peso, temperatura, volume, etc. de um objeto ou valores de números aleatórios. Os valores nesse conjunto são conhecidos como *datum*. O conjunto de dados consiste em dados de um ou mais membros correspondentes a cada linha. Nesse trabalho científico, se aprenderá a definição do conjunto de dados, diferentes tipos de conjuntos de dados, propriedades e assim por diante, com muitos exemplos resolvidos.

Em Estatística, tem-se diferentes tipos de conjuntos de dados disponíveis para diferentes tipos de informação. Eles são, a saber:

- **Conjuntos de dados numéricos**: o conjunto de dados numéricos é um conjunto de dados, em que os dados são expressos em números em vez de linguagem natural.

Os dados numéricos, às vezes, são chamados de dados quantitativos. O conjunto de todos os dados quantitativos/dados numéricos é chamado de conjunto de dados numéricos. Os dados numéricos estão sempre na forma de números, de modo que se pode realizar operações aritméticas sobre eles, como peso e altura de uma pessoa, contagem de RBC em um relatório médico ou número de páginas presentes em um livro;

- **Conjuntos de dados bivariados**: um conjunto de dados que tem duas variáveis é chamado de conjunto de dados bivariados. Ele lida com a relação entre as duas variáveis. O conjunto de dados bivariado, geralmente, contém dois tipos de dados relacionados, por exemplo para se encontrar a pontuação percentual e a idade dos alunos de uma turma, tem-se que *score* e idade podem ser considerados como duas variáveis; em outro exemplo tem-se: a venda de sorvete versus a temperatura naquele dia. Nesse caso, as duas variáveis usadas são sorvete e temperatura (Observação: caso você tenha apenas um conjunto de dados, diga-se, temperatura, ele é chamado de conjunto de dados univariado);
- **Conjuntos de dados multivariados**: representa um conjunto de dados com várias variáveis. Quando o conjunto de dados contém três ou mais de três tipos de dados (variáveis), o conjunto de dados é chamado de conjunto de dados multivariados. Em outras palavras, o conjunto de dados multivariados consiste em medidas individuais que são adquiridas em função de três ou mais de três variáveis, por exemplo, se tiver que medir o comprimento, largura, altura, volume de uma caixa retangular, tem-se o uso de múltiplas variáveis, para que se possa distinguir entre essas entidades;
- **Conjuntos de dados categóricos**: conjuntos de dados categóricos representam recursos ou características de uma pessoa ou objeto. O conjunto de dados categóricos consiste em uma variável categórica também chamada de variável qualitativa, que pode assumir exatamente dois valores. Por isso, é denominado como conjunto com uma variável dicotômica. Dados/variáveis categóricos com mais de dois valores possíveis são chamados de variáveis politômicas. As variáveis qualitativas/categóricas são frequentemente consideradas variáveis politômicas, a menos que especificado de outra forma, por exemplo gênero de uma pessoa (masculino ou feminino) ou estado civil (casado/solteiro);
- **Conjuntos de dados de correlação**: o conjunto de valores que demonstram algum relacionamento entre si indica conjuntos de dados de correlação. Aqui os valores são encontrados como dependentes uns dos outros. Geralmente, a correlação é definida como uma relação estatística entre duas entidades/variáveis. Em alguns cenários, pode ser necessário prever a correlação entre as coisas. É essencial entender como a

correlação funciona. A correlação é classificada em três tipos: correlação positiva: duas variáveis se movem na mesma direção (ambas estão para cima ou ambas ou para baixo); correlação negativa: duas variáveis se movem em direções opostas. (uma variável está em alta e outra está em baixo e vice-versa); nenhuma ou zero correlação – nenhuma relação entre duas variáveis. Por exemplo, uma pessoa alta é considerada mais pesada que uma pessoa baixa. Então aqui as variáveis de peso e altura são dependentes umas das outras.

Num conjunto de dados podemos calcular a média, a mediana e a moda, juntamente com o intervalo e esses são os principais tópicos em Estatística. Em outras palavras, calcular a média, a mediana e a moda dos conjuntos de dados são os três métodos para trabalhar com eles. No entanto, antes de se poder calcular essas três medidas do conjunto de dados, deve-se primeiro preparar nosso conjunto de dados reescrevendo-o em ordem crescente do menor para o maior.

A *média* de um conjunto de dados é a *média* de todas as observações presentes na tabela. É a razão entre a *soma das observações* e o *número total de elementos* presentes no conjunto de dados. A fórmula da *média* é dada por:

$$Media = \frac{Soma\_Observacoes}{Numero\_Total\_Elementos\_Conjunto\_Dados} \tag{2.1}$$

A mediana de um conjunto de dados é o valor central da coleção de dados quando organizados em ordem crescente e decrescente. Moda de um conjunto de dados é a variável ou o número ou o valor o qual é repetido o número máximo de vezes no conjunto. O *intervalo* de um conjunto de dados é a diferença entre o *valor máximo* e o *valor mínimo*.

$$Intervalo = Valor\_Maximo - Valor\_Minimo \tag{2.2}$$

Neste projeto será usada a base de dados do tipo numérico, pois vão extrair dados estatísticos dos textos e atividades de *tweets* coletados de usuários candidatos da rede social Twitter.

2.7 Divisão dos Dados na Base de Dados

Na aprendizagem de máquina, uma tarefa comum é o estudo e construção de algoritmos que podem ser apreendidos para se fazer previsões sobre os dados [33]. Esses algoritmos funcionam fazendo previsões ou decisões baseadas em dados, por meio da construção de um modelo matemático a partir dos dados de entrada [34].

Os dados usados para construir o modelo final, geralmente, vêm de vários conjuntos de dados. Em particular, três conjuntos de dados são comumente usados em diferentes estágios da criação do modelo:

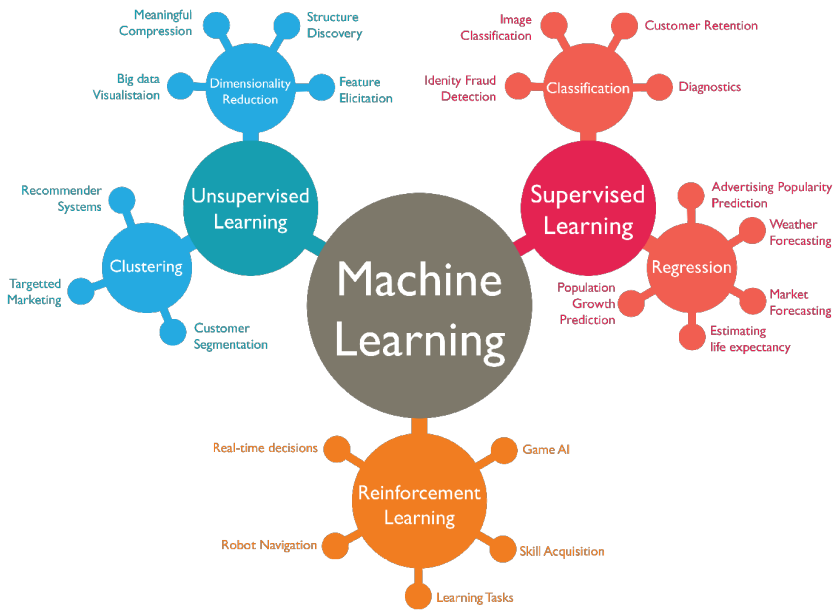
- **Conjunto de Treinamento**: É o conjunto de dados que é usado para treinar e fazer o modelo aprender os recursos/padrões ocultos nos dados. Em cada época, os mesmos dados de treinamento são alimentados repetidamente na arquitetura da rede neural e o modelo continua aprendendo os recursos dos dados [35]. O conjunto de treinamento deve ter um conjunto diversificado de entradas para que o modelo seja treinado em todos os cenários e possa prever qualquer amostra de dados não vista que possa aparecer no futuro [36].
- **Conjunto de Validação**: O conjunto de validação é um conjunto de dados, separado do conjunto de treinamento, que é usado para validar o desempenho do nosso modelo durante o treinamento. Esse processo de validação fornece informações que nos ajudam a ajustar os hiperparâmetros e as configurações do modelo de acordo [35]. É como um crítico nos dizendo se o treinamento está indo na direção certa ou não. O modelo é treinado no conjunto de treinamento e, simultaneamente, a avaliação do modelo é realizada no conjunto de validação após cada época. A ideia principal de dividir o conjunto de dados em um conjunto de validação é evitar que nosso modelo seja superajustado, ou seja, o modelo se torna realmente bom em classificar as amostras no conjunto de treinamento, mas não pode generalizar e fazer classificações precisas nos dados que não viu antes [36].
- **Conjunto de Teste**: O conjunto de teste é um conjunto separado de dados usado para testar o modelo após a conclusão do treinamento [36]. Ele fornece uma métrica de desempenho do modelo final imparcial em termos de exatidão, precisão, etc. Para simplificar, ele responde à pergunta "Qual é o desempenho do modelo? " [35].

2.8 Aprendizagem de Máquina

Ao longo dos anos, várias pessoas criaram definições sobre a Aprendizagem de Máquina (em inglês: *Machine Learning* (ML)). Um grande cientista da computação americano, chamado Tom Michael Mitchell [37] define a aprendizagem de máquina como um programa de computador aprende a partir de uma experiência E com respeito a uma tarefa T é uma métrica de performance P, se a sua performance em T, medida por P, melhora com a experiência E. Aurélien Géron [2] esclarece a aprendizagem de máquina como a ciência (e a arte) de se programar computadores para que eles aprendam a partir dos dados; Arthur Samuel [38] explica que é a aprendizagem de máquina na área de estudo

que dá aos computadores a habilidade de aprender sem interferência humana. Kaplan e Haenlein [39–42] declaram que a capacidade de um sistema em interpretar corretamente dados externos é aprender a partir desses dados e utilizar essas aprendizagens para atingir objetivos e tarefas específicas através de adaptação flexível. A Figura 2.3 mostra os principais algoritmos de aprendizagem de máquina mais comuns, suas técnicas e onde se aplicam:

Figura 2.3: Diagrama da Aprendizagem de Máquina[2]



- **Aprendizagem Supervisionada**: É usado para classificar dados não vistos em categorias estabelecidas e prever tendências e mudanças futuras como um modelo preditivo. Um modelo desenvolvido por meio de aprendizado de máquina supervisionado aprenderá a reconhecer objetos e as características que os classificam. Ao aprender padrões entre dados de entrada e saída, os modelos de aprendizado de máquina supervisionados podem prever resultados de dados novos e não vistos. O aprendizado de máquina supervisionado é frequentemente usado para: Classificação de diferentes tipos de arquivos, como imagens, documentos ou palavras escritas. Previsão de tendências e resultados futuros por meio de padrões de aprendizado em dados de treinamento [43].
- **Aprendizagem não Supervisionada**: È o treinamento de modelos em dados de treinamento brutos e não rotulados. É frequentemente usado para identificar padrões e tendências em conjuntos de dados brutos ou agrupar dados semelhantes em um número específico de grupos. Também é frequentemente uma abordagem usada

[2] `https://www.isaziconsulting.co.za/machinelearning.html`

na fase exploratória inicial para entender melhor os conjuntos de dados. Por outro lado, o aprendizado de máquina supervisionado pode consumir muitos recursos devido à necessidade de dados rotulados. O aprendizado de máquina não supervisionado é usado principalmente para: Conjuntos de dados de cluster em semelhanças entre recursos ou dados de segmento. Entenda a relação entre diferentes pontos de dados, como recomendações de música automatizadas [43].

- **Aprendizagem Semi-Supervisionada**: É uma combinação de aprendizado supervisionado e não supervisionado. Ele usa uma pequena quantidade de dados rotulados e uma grande quantidade de dados não rotulados, o que fornece os benefícios do aprendizado supervisionado e não supervisionado, evitando os desafios de encontrar uma grande quantidade de dados rotulados. Isso significa que você pode treinar um modelo para rotular dados sem precisar usar tantos dados de treinamento rotulados. [44]
- **Aprendizagem por Reforço**: É um método de treinamento de aprendizado de máquina baseado em recompensar comportamentos desejados e/ou punir comportamentos indesejados. Em geral, um agente de aprendizado por reforço é capaz de perceber e interpretar seu ambiente, realizar ações e aprender por tentativa e erro. Este método atribui valores positivos às ações desejadas para estimular o agente e valores negativos a comportamentos indesejados. Isso programa o agente para buscar recompensas globais máximas e de longo prazo para alcançar uma solução ideal. Com o tempo, o agente aprende a evitar o negativo e a buscar o positivo. Esse método de aprendizado foi adotado em inteligência artificial como forma de direcionar o aprendizado de máquina não supervisionado por meio de recompensas e penalidades [45].

2.9 Algoritmos de Aprendizagem Supervisionado

Nesse projeto o tipo de aprendizagem de máquina que mais se adapta aos objetivos do propósito é a Aprendizagem de Máquina Supervisionada. Os algoritmos supervisionados serão esclarecidos e exemplificados com seus conceitos, técnicas e hipóteses, nas subseções a seguir.

2.9.1 Análise Discriminante Linear (LDA)

Análise Discriminante Linear (em inglês: *Linear Discriminant Analysis* (LDA)) é uma técnica de estatística multivariada, utilizada para discriminar e classificar objetos [46]. De acordo com Khatree e Naik [47], a análise discriminante estuda a separação de objetos

de uma população em classes, onde a discriminação é a primeira etapa a qual procura por características capazes de serem utilizadas para alocar objetos em diferentes grupos pré-definidos. Com efeito, as mesmas regras que servem para alocar objetos podem ser utilizadas para separar objetos [46]. Como resultado, a tarefa de discriminação, visando posterior classificação consiste em obter funções matemáticas capazes de classificar uma instância X em uma das várias populações, com base em medidas de um número p de características com o objetivo de minimizar a probabilidade de classificação errônea. Em resumo, o problema consiste em obter-se uma combinação de características observadas, as quais apresentem o maior poder de discriminação entre as populações. À vista disso, a essa combinação dá-se o nome de função discriminante [46].

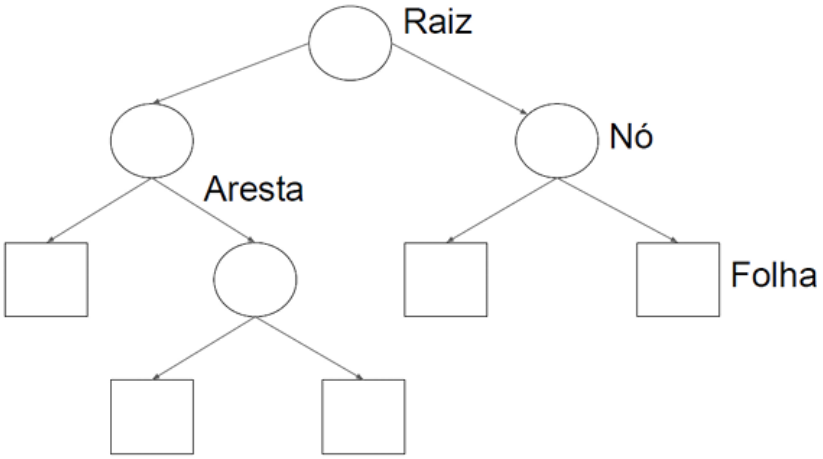
2.9.2 Árvore de Decisão (DT)

Árvore de Decisão (em inglês: *Decision Tree* (DT)) compreende uma série de decisões lógicas, semelhantes a um fluxograma, com "nós"de decisão indicando uma decisão a ser tomada em um atributo. Já os ramos, indicam as escolhas de cada decisão [48].

A construção de uma árvore de classificação binária começa pela identificação da variável independente que melhor segure a amostra em grupos distintos em relação a variável dependente [49]. Estes modelos utilizam a estratégia de dividir para conquistar: um problema complexo é decomposto em subproblemas mais simples e recursivamente esta técnica é aplicada a cada subproblema [50]. Para o caso da análise de crédito, a primeira divisão ou corte, distingue entre bons e maus pagadores; o segundo corte, identifica a variável que diferencia bons e maus pagadores; o terceiro corte, estabelece outra variável que diferencia o comportamento da variável anterior, e assim sucessivamente [1].

Conforme ilustrado na Figura 2.4, em um exemplo de uma Árvore de Decisão é possível visualizar a representação de uma árvore de decisão, onde o primeiro "nó"da árvore é chamado de raiz. Denomina-se de “nó” cada teste de atributo que é feito na árvore, e a sua resposta será uma decisão binária entre “sim” ou “não”, a qual denomina-se de “folha”. A ligação entre "nós"ou entre um "nó"e uma folha é chamada de “aresta”.

Figura 2.4: Exemplo de Árvore de Decisão [1]



Conforme Silva [51], o critério utilizado para realizar as partições é o da utilidade do atributo para a classificação em questão. Posteriormente, aplica-se, por este critério, um determinado ganho de informação a cada atributo. A escolha do atributo teste para o corrente "nó"é aquele que possui o maior ganho de informação e, a partir desta aplicação, inicia-se um novo processo de partição.

Para os casos em que a árvore é usada para classificação, os critérios de partição mais conhecidos são baseados na Entropia e Índice Gini [52]. Segundo Tsai et al. [53], ao selecionar-se um caso aleatório de um conjunto S de casos e estabelecer que ele pertence a alguma classe C_j, a probabilidade de que uma amostra arbitrária pertence à classe C_j é estimada na Equação 2.3:

$$P_i = freq(\frac{C_i S}{|S|}) \tag{2.3}$$

onde $|S|$ é o número de amostras no conjunto S e assim, as informações que transmitem são $-log_{2p_i}$ bits

E para uma dada distribuição discreta de probabilidade $P = p_1, p_2, ..., p_n$, a informação transmitida por esta distribuição, também chamada de entropia de "P", é conhecida como:

$$Info(P) = \sum_{i=1}^{n} -p_i log_2 p_i \tag{2.4}$$

Se a partição de um conjunto de amostras T é feita com base no valor de um atributo não-categórico X em conjuntos de $T_1, T_2, ..., T_m$, então, a informação necessária para identificar a classe de um elemento de T passa a ser a média ponderada da informação necessária para identificar a classe de um elemento de T_i; ou seja, a média ponderada de $Info(T_i)$, é definida na Equação 2.5.

$$Info(P) = \sum_{i=1}^{n} -p_i log_2 p_i \tag{2.5}$$

Dessa forma, pode-se definir o ganho de informação $Gain(X, T)$ como definido na Equação 2.6:

$$Gain(X, T) = Info(T) - Info(X, T) \tag{2.6}$$

E essa função representa a diferença entre a informação necessária para identificar um elemento de T e a informação necessária para identificar um elemento de T, após o valor do atributo X ter sido avaliado. Dessa forma, $Gain(X, T)$ é o ganho de informação devido ao atributo X [53].

Por fim, conforme Silva [51], pode-se dizer também que a construção de uma Árvore de Decisão baseia-se em três objetivos: diminuição da entropia, consistência em relação ao conjunto de dados e um número pequeno de "nós".

2.9.3 Floresta Randômica (RF)

Floresta Randômica (em inglês: *Random Forest* (RF)) pode ser definido como uma evolução do modelo de *Bagging* [54], e também faz parte dos classificadores ensemble, que criam diferentes Árvores de Decisão, as quais serão usadas posteriormente na classificação de um novo exemplo, por meio do voto majoritário.

Esse método, segundo Lantz [48], baseia-se em um conjunto de Árvores de Decisão, que combina versatilidade e potência em uma abordagem de Aprendizagem de Máquina única. O método utiliza apenas uma parte das variáveis independentes disponíveis no conjunto de dados, e as mesmas são selecionadas de forma aleatória para a construção de cada árvore de decisão.

Conforme Breiman [54], pode-se definir a Floresta Randômica como um classificador que consiste em uma coleção de árvores classificadoras estruturadas $h(x, \theta n), n = 1, ..., k$ onde os θn são os vetores aleatórios independentes e identicamente distribuídos e, a partir dos dados de entrada x, cada árvore lança um único voto para a classe mais popular. Tratando-se de crédito, devido às diferentes sub-amostras geradas tanto no Floresta Randômica, quanto no *Bagging*, uma das formas de classificar o novo cliente é pela identificação da maioria das classificações obtidas em cada uma das árvores.

Como já mencionado, Oshiro [55] aponta que a Floresta Randômica aplica o mesmo método que o *Bagging* para produzir amostras *bootstraps* de conjuntos de treinamento para cada árvore de decisão. A única diferença entre os métodos, é que no Floresta Randômica as m co-variáveis em cada "nó"das árvores, são selecionadas de forma aleatória e o valor de m é fixo para todos os "nós".

Ainda, Breiman [54] justifica o uso de *Bagging* em Floresta Randômica pelas seguintes razões: o uso do *Bagging* aparenta melhorar o desempenho, quando atributos aleatórios são tomados e também, esta técnica deve ser usada para fornecer estimativas do erro de generalização do conjunto combinado de árvores. Desse jeito, por gerar diferentes Árvores de Decisão, a aleatorização implica numa correlação baixa entre as mesmas, o que diminui o erro de classificação em ambos os algoritmos.

2.9.4 Gradient Boosting (GB)

Gradient Boosting (em inglês: *Gradient Boosting* (GB)) é uma abordagem baseada em gradiente para apreender um classificador impulsionador incrementalmente e aproximar uma função $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ com base em uma combinação linear de aprendizes fracos $h : \mathbb{R}^n \longrightarrow \mathbb{R}$, definido na Equação 2.7:

$$f(x) = \sum_{j=1}^{M} \alpha_i h_j(x; \theta_j) \tag{2.7}$$

2.9.5 K-ésimo Vizinho mais Próximo (KNN)

K-ésimo Vizinho mais Próximo (em inglês: *K-Nearest Neighbors* (KNN)) é um classificador baseado na analogia, onde o conjunto de treinamento é formado por vetores de n-dimensões, sendo que cada elemento desse conjunto representa um ponto no espaço n-dimensional [56].

De forma a classificar um elemento o qual não pertence ao conjunto de treinamento, o classificador *KNN* procura K elementos no conjunto de treinamento, sendo que estes K elementos devem estar próximos do elemento desconhecido. Em vista disso, verifica-se quais são as classes dos K-vizinhos, sendo que a classe mais frequente é atribuída ao elemento desconhecido. A métrica mais comum para determinação dos K-vizinhos mais próximos é a distância euclidiana, definida pela Equação 2.8 [56].

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_1 - y_1)^2 + ... + (x_n - y_n)^2} \tag{2.8}$$

2.9.6 Perceptron Multicamadas (MLP)

Perceptrons Multicamadas (em inglês: *Multilayer Perceptron* (MLP)), de acordo com Braga et al. [57] só pode resolver problemas linearmente separáveis. Para que fosse possível solucionar esse problema, foi preciso fazer uma implementação de camadas intermediárias (camadas ocultas) entre as camadas de entrada e saída.

Deve-se atentar ao fato de que a utilização de duas camadas intermediárias permite a aproximação de qualquer função. A utilização de uma camada intermediária é o bastante para se resolver problemas em que suas funções são contínuas em todo o domínio, assim, para casos em que haja descontinuidade nas funções a serem aproximadas, faz-se necessário o acréscimo de uma camada intermediária. A utilização de mais de duas camadas diminuirá o desempenho da rede [51].

Segundo Calôba et al. [58] os neurônios da MLP possuem função de ativação não linear, onde geralmente é usada a função Sigmoide. De acordo com Silva [51], geralmente são usadas funções de ativação lineares na camada de saída da rede, e nas camadas intermediárias são usadas funções não lineares, como a função Sigmoide.

2.9.7 Máquina de Vetores de Suporte (SVM)

Máquina de Vetores de Suporte (em inglês: *Support Vector Machine* (SVM) se baseia na construção de diversos hiperplanos (superfície de decisão) utilizados para separar diferentes amostras de forma a se encontrar o hiperplano "ótimo", o qual representa uma determinada instância. Este classificador realiza uma representação de amostras, como pontos no espaço, mapeadas de forma que os exemplos pertencentes a classes distintas apresentam uma divisão espacial (margem) bem definida [59]. O hiperplano pode ser definido através da Equação 2.9:

$$W^T x + b = x \in R^d \tag{2.9}$$

De forma a se encontrar os hiperplanos com margem máxima, Cortes e Vapnik [60] sugerem a utilização de *kernels*, os quais modificam a função que define a margem "ótima"de separação das classes no hiperplano. Destarte, define-se o "kernel linear"através da 2.10:

$$K(x_i, x_j) = x_i^T x_j \tag{2.10}$$

Considerando-se duas amostras com vetores x_i e x_j. Outro kernel existente é a Função Base Radial (em inglês: *Radial Basis Function* (RBF)), a qual apresenta uma simplificação significativa nas computações necessárias para busca do hiperplano e margem "ótimos", sendo o mesmo definido pela 2.11 [60]:

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2} \tag{2.11}$$

2.9.8 Naive Bayes (NB)

Naive Bayes (em inglês: *Naive Bayes* (NB)) de classificação são um conjunto de algoritmos de aprendizado supervisionado o qual se baseia na aplicação do Teorema de Bayes com a hipótese *Naive* (ingênua) de independência entre cada par de características [59]. Assim sendo, dada uma classe y e m, o vetor de características x, o Teorema de Bayes será definido na Equação 2.12:

$$P(y|y, x_1, ..., x_n) = \frac{P(y)P(x_1, ..., x_n|y)}{P(x_1, ..., x_n)} \tag{2.12}$$

Através da hipótese *Naive* define-se a Equação 2.13 [61]

$$P(x_i|y, x_1, ...x_{i-1}, x_{i+1}, x_n) = P(x_i|y) \tag{2.13}$$

A principal diferença entre os diferentes classificadores do método *Naive Bayes* é relacionada com as hipóteses feitas sobre as distribuições de $P(x_i|y)$. O classificador *Naive* Bayes Gaussiano considera como gaussiana a distribuição, enquanto que o classificador *Naive Bayes Multinomial* considera que os dados são distribuídos multinomialmente [62].

2.9.9 Regressão Logística (LR)

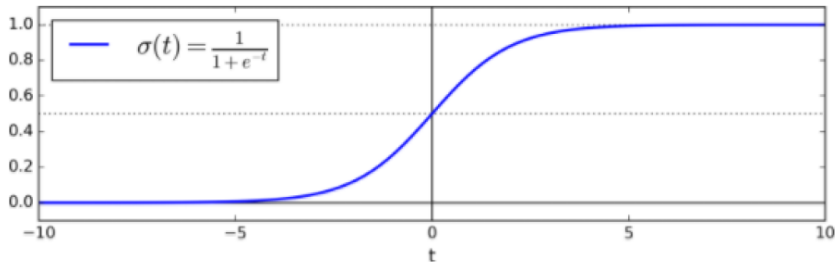
Regressão Logística (em inglês: *Logistic Regression* (LR)) é comumente utilizada para estimar a probabilidade de uma instância pertencer a uma classe em particular (ex.: a probabilidade de um email ser "spam"). Em uma classificação binária, caso essa probabilidade seja maior que 50%, o modelo classifica a instância como pertencente à classe em questão (chamada de classe positiva 1). Caso contrário, o modelo classifica a instância como pertencente à classe negativa (0) [2].

Para tanto, a regressão logística calcula uma soma ponderada das características e distribui a saída seguindo uma distribuição logística, percorrendo uma curva do tipo função Sigmoide, definida na Equação 2.14 e demonstrada na Figura 2.5. A saída calculada pela regressão logística é uma estimativa de probabilidade definida pela Equação 2.15 [2]:

$$\sigma(t) = \frac{1}{1 + e^{-t}} \tag{2.14}$$

$$p = h_\theta(x) = \sigma(\theta^T * x) \tag{2.15}$$

Figura 2.5: Função Logística [2]



Na regressão logística, o objetivo do treinamento é definir o vetor θ de forma que o modelo estime probabilidades elevadas para o caso de instâncias positivas ($y = 1$, ou seja, se a instância pertence à classe verdadeira) e probabilidades baixas para instâncias negativas ($y = 0$, ou seja, se a instância pertence à classe falsa). Para tanto, a regressão logística utiliza uma função de custo, mostrada na Equação 2.16 e na Equação 2.17:

$$c(\theta) = -log(p), \text{se y} = 1 \tag{2.16}$$

$$c(\theta) = -log(1-p), \text{se y} = 0 \tag{2.17}$$

onde p é calculado através da Equação 2.15.

A função de custo cresce de forma expressiva quando t se aproxima de "0", fazendo com que o custo seja alto quando o modelo estima uma probabilidade próxima de "0"para uma instância positiva, sendo que o mesmo ocorre quando o modelo, erroneamente, estima uma probabilidade alta (t próximo de 1, ou seja, $\theta^T x$ x próximo de 1) para uma instância negativa. O custo sobre todo o conjunto de dados de treinamento é a média do custo sobre todas as instâncias de treinamento, definido pela Equação 2.18:

$$J(\theta) = -\frac{1}{m}\sum_{m}^{i=1}[y^{(n)}log(p^{(i)}) + (1-y^{(i)})log(1-p^{(i)})] \tag{2.18}$$

onde y (classe prevista) assume o valor de 0 se $p < 0.5$ e 1 se $p > 0.5$.

Deve-se notar que não existe uma forma fechada de se calcular a Equação 2.18. Porém, como a mesma é convexa, a utilização de um algoritmo de otimização garante o encontro do máximo global. A derivação parcial da Equação 2.18 com relação a θ_j é dada pela Equação 2.19:

$$\frac{\delta}{\delta\theta_j}J(\theta) = \frac{1}{m}\sum_{m}^{i=1}(\sigma(\theta^T * x^i) - y^{(i)})x_j^{(i)} \tag{2.19}$$

Analisando a Equação 2.19 é possível observar que para cada instância será calculado o erro da previsão, multiplicando-o pelo valor da característica j, calculando então a média de todas as instâncias de treinamento [2].

2.9.10 Bagging

Bagging é considerado um método de classificadores ensemble que, segundo Breiman [54], são classificadores treinados de forma independente por diferentes conjuntos de treinamento, por meio de um método de inicialização. Assim, várias Árvores de Decisão são criadas de forma aleatória, para posteriormente serem combinadas. A formação dessas árvores é feita por amostragem *bootstrap*: a partir do conjunto de treinamento inicial, os subconjuntos são criados de forma aleatória, com reposição.

Cada subconjunto gerado possui o mesmo tamanho (número de exemplos) do conjunto original. Considerando um conjunto de treinamento T com n exemplos, Tk é uma amostra bootstrap do conjunto de treinamento a partir de T com reposição, contendo n exemplos. Cada subconjunto Tk é usado para treinar um classificador diferente ($hk(x)$), e a estratégia de combinação dos classificadores é o voto majoritário [55]. Essa estratégia é simples, mas pode reduzir a variância do classificador final, quando combinado com as estratégias de geração de bases de aprendizagem de máquina [63].

Segundo Acuna e Rojas [64], Breiman [54] e Freund e Schapire [65] (1996), *Bagging* é muito eficaz quando os classificadores utilizados possuem um comportamento instável (Árvores de Decisão, por exemplo).

Um classificador é conhecido como instável, quando pequenas mudanças no conjunto de treinamento podem causar grandes mudanças no classificador gerado. Quando isso ocorre, um único classificador instável não é capaz de oferecer uma resposta confiável; ao contrário de um conjunto de classificadores, uma vez que um classificador composto pode ter maior chance de acerto [66].

2.9.11 Boosting

Boosting foi proposto inicialmente por Schapire [67], é semelhante ao *Bagging*, pois cada classificador é construído com base num conjunto de treinamento diferente. Conforme Lantz [48], os conjuntos de dados re-amostrados em *Boosting* são construídos com o intuito de gerar aprendizados complementares e a importância do voto é ponderada seguindo o desempenho de cada modelo.

Pela capacidade de aumentar o desempenho de um limiar arbitrário com a adição dos modelos de aprendizados mais fracos, o *Boosting* é considerado uma das descobertas mais significativas em aprendizagem de máquina [48]. Porém, como em todo algoritmo

de Aprendizagem de Máquina, dificuldades de implementação e/ou execução aparecem. Logo, com o intuito de contornar esta situação, o Adaboost surgiu. O mesmo foi apresentado pela primeira vez por Freund e Schapire [65] e tem originado crescente número de pesquisas e aplicações em várias áreas.

Segundo Tsai et al. [53], para treinar o k-ésimo classificador como um "classificador fraco", são utilizados para treiná-lo n conjuntos de amostras da base de treinamento ($n < m$) entre S. Após, o classificador treinado é avaliado por S, a fim de identificar situações de treinamento que foram classificadas erroneamente, então a árvore $k + 1$ é treinada por um conjunto de treinamento modificado, o que reforça a importância de exemplos classificados incorretamente. Além disso, segundo os autores citados acima, o procedimento de amostragem é repetido até que k amostras de treinamento sejam criadas para a construção da k-ésima árvore e, portanto, a decisão final é baseada na votação ponderada dos classificadores individuais.

Após a proposta original do *Adaboost* ter sido apresentada aos cientistas e pesquisadores, muitas variações e extensões deste algoritmo foram sugeridas e desenvolvidas como alternativas, para fornecerem melhores resultados para problemas específicos [68].

2.9.12 Voting Classifier

Voting Classifier é um tipo homogêneo e heterogêneo de Aprendizagem Ensemble, ou seja, os classificadores base podem ser do mesmo tipo ou de tipos diferentes. Como mencionado anteriormente, este tipo de ensemble também funciona como uma extensão do *Bagging* (por exemplo, *Random Forest*).

A arquitetura de um *Voting Classifier* é composta por um número n de modelos de ML, cujas previsões são avaliadas de duas maneiras diferentes: *hard* e *soft*. No modo difícil, a previsão vencedora é aquela com "mais votos".

Por outro lado, o *Voting Classifier*, no modo *soft*, considera as probabilidades lançadas por cada modelo de Aprendizagem de Máquina, essas probabilidades serão ponderadas e médias, e consequentemente a classe vencedora será aquela com maior probabilidade ponderada e média.

2.10 Métricas de Avaliação dos Algoritmos Supervisionados

As métricas de avaliação são usadas para avaliar o desempenho de modelos de classificação em aprendizado de máquina. O desempenho é avaliado através de métodos de avaliação

que irão comparar as previsões obtidas por um modelo, com os valores reais da base de dados [69]. Segue abaixo os métodos de avaliação:

- **Matrizes de Confusão**: É uma tabela que contém o número de predições corretas e incorretas por classe do modelo [2]. As predições são definidas como: **Verdadeiro Positivo *(True Positive TP)***: ocorre quando no conjunto real, a classe que se busca foi prevista corretamente; **Falso Positivo *(False Positive FP)***: ocorre quando no conjunto real, a classe que se busca prever, foi prevista incorretamente. **Falso Verdadeiro *(True Negative TN)***: ocorre quando no conjunto real, a classe que não se busca prever, foi prevista corretamente. **Falso Negativo (*False Negative FN)***: ocorre quando no conjunto real, a classe que não se busca prever, foi prevista incorretamente A Tabela 2.10 mostra a estrutura da matriz de confusão mais vista em projetos:

Tabela 2.10: Estrutura de uma Matriz de Confusão

		Predito	
		Positivo	Negativo
Real	Positivo	TP	FN
	Negativo	FP	TN

- **Precisão**: Mede a taxa de instâncias corretas, considerando apenas o que foi classificado como positivo. A precisão é calculada, obtendo da quantidade de verdadeiros positivos *(TP)* por relação da quantidade de positivos *(TP + FP)*:

$$Precisao = \frac{TP}{TP + FP} \tag{2.20}$$

- **Recall (Sensibilidade)**: Determina a proporção de predições positivas detectadas corretamente pelo classificador. O Recall é calculado obtendo a quantidade de verdadeiros positivos *(TP)* por relação da quantidade de predições verdadeiros reais *(TP + FN)*:

$$Recall = \frac{TP}{TP + FN} \tag{2.21}$$

- **F1-score (Taxa F1)**: Calcula a média harmônica entre Precisão e Recall, uma maneira simples de comparar dois classificadores. O F1-score é calculado na seguinte forma:

$$F1 - score = \frac{2 * Precisao * Recall}{Precisao + Recall} = \frac{2TP}{2TP + FP + FN} \tag{2.22}$$

- **Acurácia**: Calcula a medida mais geral, pois calcula o número de predições corretas como um todo: A Acurácia é calculada obtendo a quantidade de verdadeiros positivos *(TP)*, mais a quantidade de verdadeiros positivos *(FN)* por relação da soma de todas as predições *(TP + FP + TN + FN)*:

$$Acuracia = \frac{TP + FN}{TP + FP + TN + FN} \tag{2.23}$$

- **Acurácia Balanceada**: A acurácia balanceada, ao contrário da acurácia, não é influenciada pelo desbalanceamento das classes, porque os cálculos ocorrem em cima da taxa de *verdadeiros positivos* e de *verdadeiros negativos*. Logo, é possível conseguir chegar a um valor mais correto em relação aos acertos do modelo, em relação às classes:

$$AcuraciaBalanceada = \frac{1}{2}\left(\frac{VP}{VP + FN} + \frac{VN}{VN + FP}\right) \tag{2.24}$$

2.11 Underfitting e Overfitting

Existem dois pontos a que um classificador pode resultar: *Underfitting* e *Overfitting*. Um classificador pode resultar em *overfitting*, quando é induzido; é possıvel que ele seja muito específico para o conjunto de treinamento, "i.e"; apresentando um alto grau de precisão para os exemplos de treinamento, e uma alta taxa de erro para o conjunto de teste. Nesse caso, pode-se dizer que o classificador “decorou” os dados, não conseguindo generalizar o conceito.

Um classificador pode resultar em *underfitting*, quando, provavelmente o conjunto de treinamento seja composto por exemplos pouco representativos, ou quando o usuário predefina classes muito pequenas (e.g., um alto fator de poda para uma árvore de decisão) ou uma combinação de ambos os casos. Nesse caso, pode-se dizer que o classificador não conseguiu abstrair o conceito, apresentando, baixa performance, no conjunto de treinamento e no conjunto de testes.

2.12 Redução de Dimensionalidade

Em problemas de classificação de aprendizado de máquina, geralmente há muitos fatores com base nos quais a classificação final é feita. Esses fatores são basicamente variáveis chamadas características. Quanto maior o número de recursos, mais difícil fica visualizar o conjunto de treinamento e depois trabalhar nele. Às vezes, a maioria desses recursos está correlacionada e, portanto, redundante. É aqui que os algoritmos de redução de

dimensionalidade entram em ação. A redução de dimensionalidade é o processo de redução do número de variáveis aleatórias em consideração, por meio da obtenção de um conjunto de variáveis principais. Ele pode ser dividido em seleção de recursos e extração de atributos.

2.12.1 Análise de Componentes Principais (PCA)

Análise de Componentes Principais (em inglês: *Principal Component Analysis* (PCA)) é uma técnica de análise multivariada, cujo objetivo é extrair as informações mais importantes de um conjunto de dados e diminuir o tamanho desse conjunto, mantendo as informações principais [70].

Esse processo é feito por meio da criação de novas variáveis, chamadas de componentes principais, construídas através de combinações lineares das variáveis iniciais. Numa interpretação geométrica, um componente principal é um eixo ao longo do qual há o máximo de variância; quanto maior a dispersão de dados, ao longo desse eixo, maior a quantidade de informações que ele contém [71].

Nesse processo, são criadas tantos componentes principais quanto existem variáveis, mas as combinações lineares são feitas de modo que as informações contidas nos dados iniciais não sejam distribuídas de forma homogênea nos componentes principais, pelo contrário, o *PCA* busca concentrar a maior quantidade de informações possível na primeira variável criada, a maior quantidade de informações restantes, na segunda variável, e assim sucessivamente. [71].

Ainda que o *PCA* resulte em uma quantidade de componentes principais, igual à quantidade de componentes iniciais, essa desigualdade acentuada na concentração de informações possibilita o descarte de um número considerável de componentes, sem que haja perda significativa das informações do conjunto de dados de entrada, resultando na redução de dimensionalidade do *dataset*.

2.13 Lei Geral de Proteção de Dados Pessoais (LGPD)

A Lei Geral de Proteção de Dados Pessoais (LGPD), Lei n° 13.709/2018, é a lei brasileira a qual tem o intuito de regular o tratamento dos dados pessoais do cidadão [5], a lei é baseada na legislação denominada GDPR da União Europeia. A sociedade de hoje em dia vive a revolução por conta da tecnologia e da informação, sendo cercada por dispositivos e máquinas que captam e distribuem os dados concebidos pelas pessoas por todo lugar. Devido ao grande fluxo de dados e informações pessoais dos cidadãos, governos de todo o mundo decidiram aprovar leis para proteger as informações de seus cidadãos. Desde 2018, o Brasil se tornou um dos países a reconhecer a importância da proteção de dados.

No Artigo 2º da lei, é estabelecido que a disciplina da proteção de dados pessoais tem como fundamentos, o respeito à privacidade, à inviolabilidade da intimidade do cidadão, à liberdade de se expressar, comunicar, opinar e de se informar, o direito de exercer a liberdade sobre as ações referente aos seus dados, além dos dignos direitos de livre iniciativa e concorrência, de desenvolvimento econômico, tecnológico e à inovação. No Artigo 6º, são listados os princípios que norteiam o tratamento de dados, precedidos, acima de tudo como determina no caput, pela boa-fé. A Tabela 2.11 apresenta os princípios da *LGPD* para tratamento de dados pessoais que são seguidos no trabalho.

Tabela 2.11: LGPD Artigo 6º As atividades de Tratamento de Dados Pessoais [4, 5]

#	**Princípio**	**Definição**
I	Finalidade	Realização do tratamento para propósitos legítimos, específicos, explícitos e informados ao titular, sem possibilidade de tratamento posterior de forma incompatível com essas finalidades.
II	Adequação	Compatibilidade do tratamento com as finalidades informadas ao titular, de acordo com o contexto do tratamento.
III	Necessidade	Limitação do tratamento ao mínimo necessário para a realização de suas finalidades, com abrangência dos dados pertinentes, proporcionais e não excessivos em relação às finalidades do tratamento de dados.
IV	Livre Acesso	Garantia, aos titulares, de consulta facilitada e gratuita sobre a forma e a duração do tratamento, bem como sobre a integralidade de seus dados pessoais.
V	Qualidade de Dados	Garantia, aos titulares, de exatidão, clareza, relevância e atualização dos dados, de acordo com a necessidade e para o cumprimento da finalidade de seu tratamento.
VI	Transparência	Garantia, aos titulares, de informações claras, precisas e facilmente acessíveis sobre a realização do tratamento e os respectivos agentes de tratamento, observados os segredos comercial e industrial.
VII	Segurança	Utilização de medidas técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou difusão.
VIII	Prevenção	Adoção de medidas para prevenir a ocorrência de danos em virtude do tratamento de dados pessoais.
IV	Não Discriminação	Impossibilidade de realização do tratamento para fins discriminatórios ilícitos ou abusivos
X	Responsabilização e Prestação de Contas	Demonstração, pelo agente, da adoção de medidas eficazes e capazes de comprovar a observância e o cumprimento das normas de proteção de dados pessoais e, inclusive, da eficácia dessas medidas.

2.14 Considerações Finais

A partir de trabalhos relacionados, foi possível compreender diversos conceitos técnicos de aprendizagem de máquina na área de saúde na doença depressão, esse conceitos e técnicas foram utilizados nesta pesquisa, desde a coleta e pré-processamento de dados, até melhores modelos de classificação para cada caso e diversas maneiras para avaliações de performance. De toda forma, a maior parte dos trabalhos relacionados explora as várias técnicas de aprendizagem de máquina, usando conjuntos de dados elaborados de candidatos voluntários de pesquisa. Esse projeto tem como objetivo principal criar um modelo de aprendizagem supervisionado, adequado para identificar possíveis padrões e sinais de comportamento depressivo em usuários na rede social Twitter, sendo assim, também é base de uma contribuição para estudos de trabalhos futuros. No próximo capítulo, será detalhado o pipeline elaborado e seguido, explicando-se cada uma de suas etapas na coleta de dados, extração de características, construção da base de dados e aplicação dos algoritmos de máquina utilizados.

Capítulo 3

Metodologia

O projeto dessa pesquisa propõe usar os conceitos de mineração de dados, que é um processo de aplicação de técnica estatística, inteligência artificial e métodos de aprendizagem de máquina supervisionados, para descoberta de padrões e regras significativas; com propósito de transformar dados em informações úteis, novas e compreensíveis que sejam capazes de predizer se um usuário da conta do *Twitter* possui sinais de comportamento *depressivo e não depressivo*.

Na Seção 3.1 apresenta-se o *pipeline* de forma simples e de forma detalhada com suas etapas resumidas. Na Seção 3.2 relata-se o processo de coleta de dados nos períodos selecionados para o estudo de caso, também especificam-se as classes categóricas dos dados de usuários candidatos da rede social *Twitter*. Na Seção 3.3 define-se o *framework* que se utilizou para buscar e coletar dados na rede social *Twitter* com as palavras-chave de buscas que foram escritas para se procurar no *Twitter*. Na Seção 3.4 escrevem-se as técnicas utilizadas para tratar e preparar os dados coletados no *Twitter* para extração de características dos usuários candidatos.

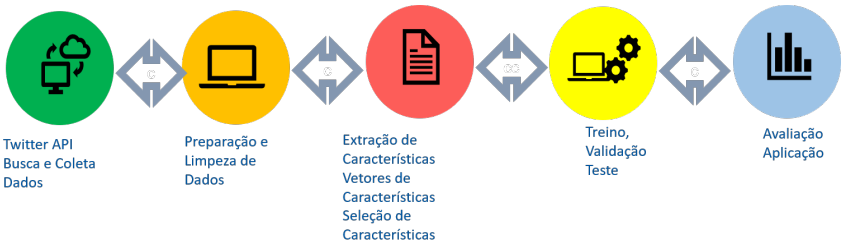
Na Seção 3.5 apresentam-se as 15 séries de características de postagens de textos e atividades que serão aplicadas nos modelos de aprendizagem de máquina para predizer se um usuário possui sinais de padrões de depressão no *Twitter*. Na Seção 3.7 descrevem-se os 4 vetores estatísticos que sumarizam os dados extraídos das características em um só valor para serem aplicados nos algoritmos de aprendizagem de máquina. Na Seção 3.8 explica-se o processo de criação da base de dados, que é usada nos algoritmos de aprendizagem de máquina. Na Seção 3.9 explica-se como se define os hiperparâmetros dos algoritmos de aprendizagem de máquina nesse projeto para predizer usuários na rede social *Twitter* que possuem sinais de padrões de comportamento depressivo e não depressivo.

3.1 Pipeline

A preparação de dados, a extração de características e as tarefas de classificação são realizadas usando os algoritmos implementados no pacote *Scikit-Learn* da linguagem de programação *Python* [10]. Os classificadores são treinados com a base de dados de treinamento, empregando-se a técnica de validação cruzada de *10-folds* para se evitar o *overfitting* (sobreajuste) e, em seguida, é usado um conjunto de teste para se avaliar a desempenho dos modelos através da métrica f1-score [72, 73].

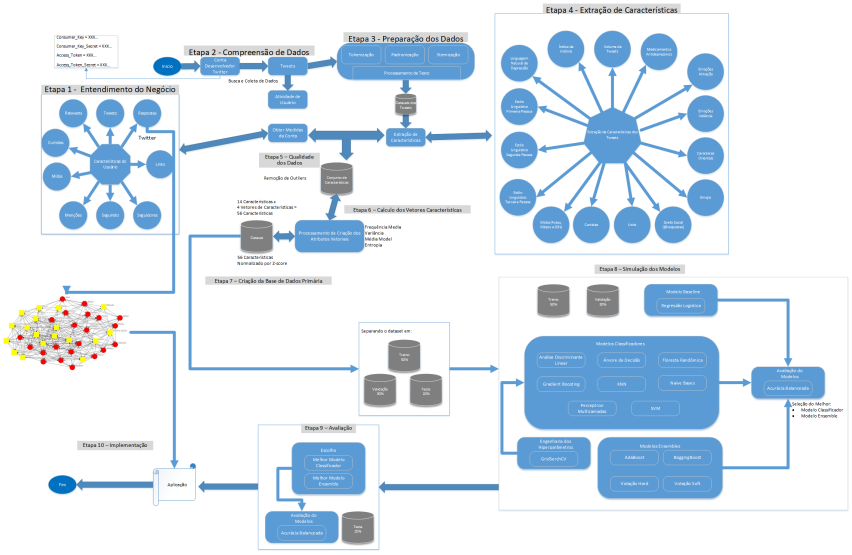
A visão da metodologia do pipeline do projeto foi decomposta em duas visões. A visão geral, ilustrada na Figura 3.1 que tem como propósito auxiliar o entendimento sequencial das atividades realizadas no processo de mineração dos dados.

Figura 3.1: Metodologia do Projeto com Visão Geral



Por outro lado, a visão detalhada da metodologia do pipeline é composta de 10 etapas, conforme ilustrado na Figura 3.2 que tem como objetivo a compreensão detalhada das atividades de coleta de dados, extração das características, bases de dados e algoritmos de aprendizagem de máquina. A metodologia é baseada na modelagem CRISP-DM [74]. A metodologia pode vista com mais detalhes nas modelagens no Apêndice B.

Figura 3.2: Metodologia do Projeto com Visão Detalhada



A metodologia do projeto tem como objetivo o intuito de identificar o algoritmo de aprendizagem de máquina supervisionado, seus pesos ou calibragens no nível de assertividade e de precisão de um modelo do classificador, o qual será incorporado em uma ferramenta para prever sinais de padrões de comportamento depressivo ou não depressivo em um usuário no *Twitter*.

- **Etapa 1 Entendimento do Negócio**: Tem como finalidade a compreensão da estrutura dos dados, o *framework API Twitter*, para usuários cadastrados na conta de desenvolvedor do *Twitter* e um *framework* desenvolvido por terceiros, que será utilizado para desenvolvimento do projeto de pesquisa.
- **Etapa 2 Compreensão dos Dados**: Consiste em manejar os serviços disponíveis da *API do Twitter e do framework* para realizar busca e coleta dos *tweets* dos usuários, com informações da conta, postagens de textos, fotos, vídeos e *links* feitas por dia e seguidores que se interagiram. A coleção de *tweets* de cada usuário coletado será armazenada e salva em um *dataset* individual.
- **Etapa 3 Preparação dos Dados**: Nesta etapa, aplicam-se algoritmos em linguagem *Python* na realização de 2 tarefas: a tarefa do pré-processamento em que é selecionada uma quantidade de *datasets*; que são combinados em um *dataset*; não são todos combinados em um só *dataset*, pois o arquivo ficaria muito grande e a tarefa de processamento de texto em que é feita uma cópia dos textos nos *datasets* e para cada texto são realizados os processos de *Tokenização*, que consistem na identificação e listagem de *tokens* (palavras, espaço em branco, sinais e números); e padronização que padronizam os *tokens* em padrões; como risadas, caracteres não latinos, palavras em minúsculas e remoção de *urls*, menções, *hastags* e *stemização* que é uma alteração de palavras para a forma raiz da palavra.
- **Etapa 4 Extração das Características**: constitui-se na extração dos dados quantitativos dos atributos. No projeto foram definidos 15 séries de atributos [11, 12, 14, 15]: **(1)** Volume de Tweets; **(2)** Índice de Insônia; Estilo Linguístico: **(3)** Pronomes na 1ª Pessoa, **(4)**, Pronomes na 2ª Pessoa e **(5)** Pronomes na 3ª Pessoa; emoções [27, 75]: **(6)** Valência (media do estado emocional) e **(7)** Ativação (media do reação emocional); **(8)** Termos Depressivos [75]; **(9)** Grafo Social, número de seguidores que interagem nos tweets através de respostas; **(10)** Medicamentos Antidepressivos [76]; **(11)** Caracteres Orientais [77]; **(12)** Emojis [72, 73]; **(13)** Frequência de Links; **(14)** Mídia (Fotos, Vídeos e Gifs); **(15)** Número de Curtidas. Cada série de atributo é salvo em *dataset* individual.

- **Etapa 5 Qualidade dos Dados**: essa etapa tem como propósito identificar usuários que estão dentro dos critérios de exclusão, ou seja, compõe-se de usuários que têm menos de 30 *tweets* na conta de *Twitter* ou mais de 300 *tweets* postados em um dia. Esses usuários identificados são classificados como *outliers* [78] e são eliminados na base de dados.

- **Etapa 6 Sumarização dos Vetores Características**: consiste na tarefa de sumarizar cada uma das 15 séries de atributos em 4 métricas vetoriais: frequência média, variância, média móvel e entropia de cada usuário resultando em 60 atributos, com seus respectivos rótulos de classe (depressão ou controle) [11, 12, 14, 15].

- **Etapa 7 Criação da Base de Dados Primária**: consiste na criação das 2 bases de dados de 2 períodos que serão inseridas nos algoritmos supervisionados para geração de modelos; as bases de dados geradas serão normalizadas pela técnica de normalização *Z-score* [79]. Por sigilo aos usuários, as bases de dados contém os dados de forma anônima.

- **Etapa 8 Simulação dos Modelos**: concretiza-se a partir do uso das bases de dados criados na etapa anterior, divididos os dados entre treino (70%), validação (15%) e teste (15%). Os conjuntos de treino e validação são induzidos nos modelos classificadores e ensembles junto com a técnica de validação cruzada de 10 -folds. Os hiperparâmetros dos modelos são definidos pela engenharia de parâmetros *GridSearchCV* que selecionam os melhores hiperparâmetros de acordo com a métrica definida (que foi a métrica f1-score). O melhor modelo de classificação e de ensemble, com melhores resultados nas métricas de *precisão, recall, f1-score e acurácia* são testados e avaliados na próxima etapa com o conjunto de teste, ou seja, com os dados desconhecidos.

- **Etapa 9 Avaliação dos Modelos**: essa etapa consiste em avaliar os 2 melhores modelos (um de classificação e um de ensemble) que tiveram melhores resultados na etapa anterior, no qual serão induzidos o conjunto de teste, para analisar qual modelo supervisionado e adaptar ao problema do projeto, avaliando desempenho dos modelos, que são as métricas de *precisão, recall, f1-score, acurácia e matriz de confusão.*

- **Etapa 10 Implementação**: encerra-se na escolha do modelo que obteve boa desempenho nas simulações, tendo como critério de escolha o menor tempo de processamento e o maior nível de acurácia, o qual será incorporado na ferramenta de análise de tweets para identificar sinais de padrões de depressão.

A linguagem utilizada para o desenvolvimento dos algoritmos foi a linguagem Python. Bibliotecas pprincipais utilizadas foram:

- **Snscrape** para buscar e coletar de tweets em dois períodos: pré-pandemia de 01 de Janeiro de 2018 a 31 de Dezembro de 2019 e pandemia de 01 de Janeiro de 2020 a 31 de Dezembro de 2021;
- **Sklearn** para aplicação e geração de modelos de aprendizagem de máquina;
- **Pandas** para tratamentos e manipulação de arquivos e séries temporais;
- **Numpy** e **scipy** para cálculos para medidas estatísticas;
- **Re** para aplicação de expressões regulares;
- **Nltk** para tratamento de texto;
- **Wordcloud** para geração de nuvem de palavras;
- **Matplotlib** para plotar gráficos;

Os metadados estruturados de um tweet utilizados no projeto foram:

- **created_at**: data e hora criada do tweet;
- **id**: identificação do tweet;
- **full_text**: conteúdo do tweet. Para maiorias dos atributos esse metadado é utilizado para extraição dos atributos;
- **id_screen_name**: identificação do usuário autor do tweet;
- **screen_name**: nome de usuário do autor do tweet;
- **favorite_count**: número de curtidas em tweet. Esse metadado é utilizado para extração do atributo número de curtidas;
- **media**: url da foto, vídeo ou gif no tweet. Esse metadado é utilizado para extração do atributo mídia;
- **links**: links de páginas encontradas no tweet. Esse metadado é utilizado para extração do atributo frequência de links;
- **reply_count**: número de respostas de usuários em um tweets. Esse metadado é utilizado para extração do para o atributo grafo social.

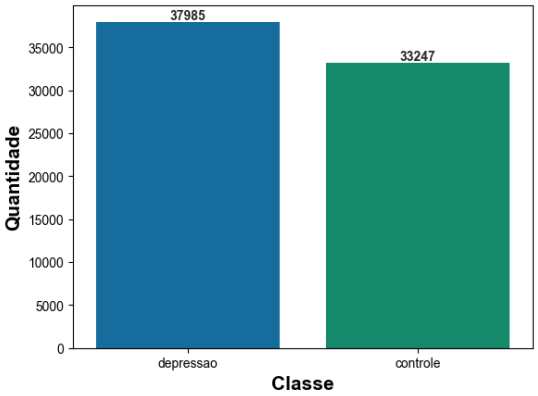
Os metadados restantes podem ser utilizados para trabalhos futuros.

3.2 Coleta de Dados

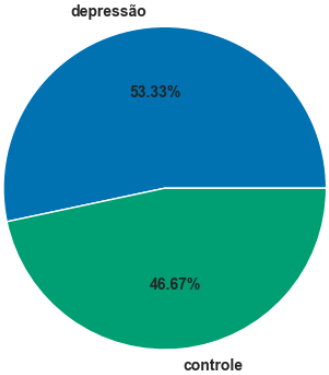
Devido à falta de dados rotulados, direcionados à depressão no Brasil, obteve-se uma oportunidade de direcionar a própria coleta de dados dessa pesquisa especificamente para esse fim. Foram realizadas duas coleta de dados de *tweets* públicos postados em dois períodos distintos; uma coleta foi realizada de 01 de Janeiro de 2018 a 31 de Dezembro de 2019, um período antes do início da pandemia COVID-19 no Brasil, a segunda coleta foi realizada de 01 Janeiro de 2020 a 31 Dezembro de 2021, época da pandemia COVID-19. Os dados coletados foram para ambas as classes: *classe depressiva e classe controle (não depressiva)*; para o período pré-pandemia foram coletados um total de 71.232 usuários (depressivos = 37.985, controle = 33.247) e para o período pandemia foram coletados um total de 70.370 usuários (depressivos = 36.801, controle = 33.569). Ambas as bases de dados não possuem classes desbalanceadas, as classes estariam desbalanceadas se uma classe possuísse mais de 20% de dados que a outra.

Figura 3.3: Análise Exploratória entre as Base de Dados

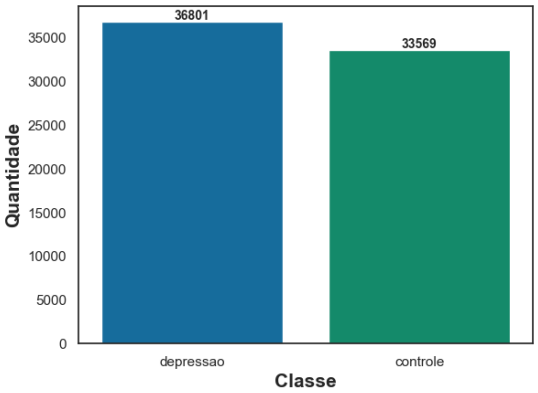
(a) Distribuição entre as Classes (Pré-Pandemia)



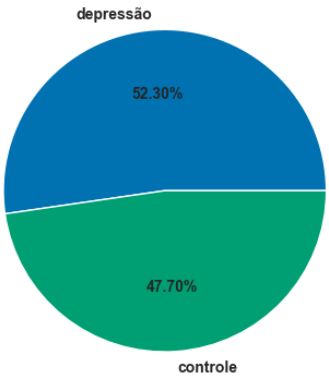
(b) Balanceamento entre as Classes (Pré-Pandemia)



(c) Distribuição entre as Classes (Pandemia)



(d) Balanceamento entre as Classes (Pandemia)

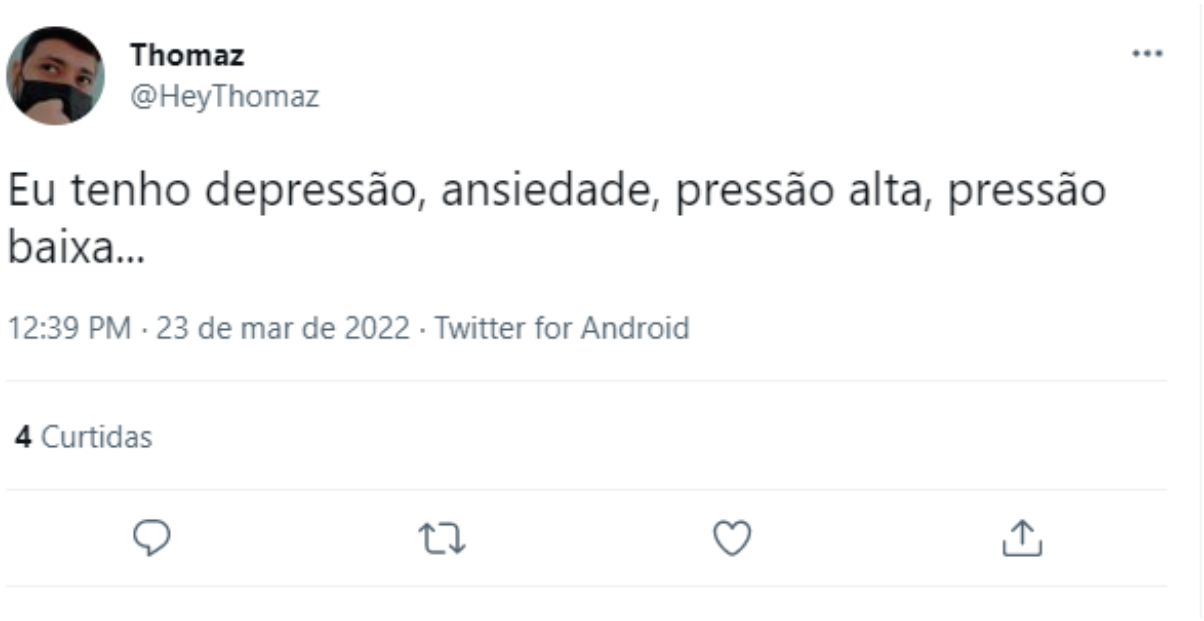


3.2.1 Classe Depressiva

Foram procurados usuários que se declaram publicamente depressivos. Buscou-se na rede social *Twitter* postagens públicas em português que continham declaração de depressão (por exemplo, *"tenho depressão"* ou *"sou depressivo"*). A razão pela qual as pessoas declaram publicamente suas alegações na rede social é possivelmente a busca de apoio da comunidade; para explicar alguns dos seus comportamentos para com os seus conhecidos, ou para lutar contra o estigma da doença mental. Alguns usuários podem fazer tais declarações de forma verdadeira ou falsa, embora a motivação por trás de tais comportamentos deixam ao escopo deste trabalho.

No entanto, presume-se que a maioria dessas declarações são de fato verdadeiras e estatisticamente superaram as falsas. Em seguida, para cada usuário foram coletados todos seus *tweets* públicos postados em um período. Usuários, com menos de 30 mensagens no total ou mais de 300 mensagens em um único dia, são identificados como *outliers* (dados que estão fora do padrão esperado, ou seja, anomalia) e são excluídos na base de dados; aos usuários restantes são consideradas observações positivas. Esse procedimento de remoção de *outliers* tem como propósito remover contas de usuário supostamente spam e contas de marketing e ter pelo menos amostras suficientes para permitir algumas análises - embora o último, seja uma regra geral para estatísticas e não métodos rigorosos. A Figura 3.4 ilustra um exemplo de um *tweet* postado no *Twitter*[1], encontrado na busca de *tweets* de depressão.

Figura 3.4: Exemplo de Tweet de Depressão no Twitter



[1] `https://twitter.com/`

3.2.2 Classe Controle

Para poder validar os dados e induzir modelos de aprendizagem de máquina supervisionado que diferenciem supostamente comportamentos depressivos de sinais não depressivos, consultou-se o *Twitter* para postagens em português, que não contenham declarações de depressão. A seguir, para cada usuário foram coletados todos seus *tweets* postados em um período. Logo depois, realizou-se o mesmo procedimento de remoção de *outliers* que foi especificado na Seção 3.2.1. Para se evitar erros de indução, nos modelos de aprendizagem de máquina supervisionado verificou-se se existem usuários em ambas as classes; se encontrados foram removidos os usuários nas duas classes. A Figura 3.5, ilustra um exemplo de um *tweet* postado no *Twitter*, encontrado na busca de *tweets* de controle.

Figura 3.5: Exemplo de Tweet de Controle no Twitter

3.3 Etapa 2 - Compreensão dos Dados

Nesta etapa foi utilizado o *framework Snscrape* [3], ao invés da *API do Twitter*, o *Snscrape* é um raspador para Serviços de Redes Sociais (*SNS*), ele raspa dados como perfis de usuários, *hashtags* ou pesquisas e retorna os itens descobertos. Esse *framework*, ao contrário da *API do Twitter Tweepy* [80], tem a vantagem de não ter as limitações da conta do desenvolvedor do Twitter, ou seja, podem-se coletar muitos *tweets* em qualquer intervalo de tempo. Com isso, foram coletados todos os *tweets* do usuário no intervalo do período configurado.

Para realizar a busca de candidatos foram utilizados as seguintes strings de busca para a classe depressão:

- "estava com depressão"
- "fui diagnosticada com depressão"
- "fui diagnosticado depressiva"
- "sou deprê"
- "estava com deprê"
- "fui diagnosticada com deprê"
- "fui diagnosticado depressivo"
- "tenho depressão"
- "estou com depressão"
- "fui diagnosticada depressiva"
- "fui diagnosticado deprê"
- "tenho deprê"
- "estou com deprê"
- "fui diagnosticada depressivo"
- "minha depressão"
- "tinha depressão"
- "fui depressiva"
- "fui diagnosticada deprê"
- "minha deprê"
- "tinha deprê"
- "fui depressivo"
- "fui diagnosticado com depressão"
- "sou depressiva"
- "tô depressiva"
- "fui deprê"
- "fui diagnosticado com deprê"
- "sou depressivo"
- "tô depressivo"

Na busca de candidatos para a classe controle utilizaram-se as mesmas palavras-chave, mas com operador de negação '-' na filtragem de busca, ou seja, textos que não contêm a frase exata, por exemplo: *-"eu tenho depressão"*, conforme descrito na Tabela 2.1 na Seção 2.2.1 de operadores de busca.

Armazenaram-se os resultados de busca em arquivo *Comma-Separated Values* CSV por classe. Logo depois, verificou-se se existe interseção dos dois resultados, ou seja, se existe os mesmos usuários em ambos os arquivos se encontrados, são removidos esses usuários nos arquivos.

Com os arquivos de resultados de busca de pontos, realizaram-se a extração de *tweets* usando o "id"(identificação) da conta de usuário do candidato; atribuindo-se nos parâmetros de busca o intervalo de tempo, extraíram-se todos os *tweets* do usuário. Os *tweets* extraídos são armazenados em um arquivo *CSV* e mantidos em máquina local, usada pelo autor que executou o processo de extração do *tweets*, sem risco de vazamento de informações do usuário na internet.

3.4 Etapa 3 - Preparação de Dados

Assim que *tweets* foram extraídos na rede social *Twitter*, elas são pré-processadas para transformar o texto de entrada em um modelo padronizado, procedendo da seguinte forma:

- **Tokenização**: identificação e listagem de **tokens** (palavras, espaço em branco, sinais e números);
- **Padronização**: padronizar os **tokens** em padrões como risadas, caracteres não latinos, palavras em minúsculas e remoção de *urls*, menções, *hashtags*, limpeza de *URLs* (começando com *"http://" ou "https: //"*), *tags* ("@usuário");
- **Stemização**: Convenção de palavras para a forma raiz da palavra.

Logo após o procedimento, é deixado uma cópia do texto original para extração de características de caracteres orientais, *emojis* e estilo linguístico que são detalhados na Seção 3.5.

3.5 Etapa 4 - Extração das Características

Nesta seção apresentaremos as características que são extraídas dos usuários candidatos na rede social *Twitter*, pois essas características apresentam o estado emocional do usuário na interação com outros usuários na rede social *Twitter*. Foram definidos as seguintes características:

1. **Volume de Tweets**: Quantidade de *tweets* por dia. A literatura sobre depressão indica que os usuários que apresentam sinais de depressão tendem a ser muito ativos nas redes sociais o dia todo.
2. **Índice de Insônia**: Relação de *tweets* postados durante o período da *noite* (21:00-6:00) pela quantidade de *tweets* postados durante o período do *dia* (6:01-20:59). A mesma literatura de volume de *tweets*, mas durante o turno da *noite*.
3. **Estilo Linguístico na 1ª Pessoa**: Quantidade de palavras na 1ª pessoa do caso reto no singular e plural, pois o uso de palavras funcionais é conhecido por fornecer uma maneira não reativa de explorar processos sociais e de personalidade [81, 82].
4. **Estilo Linguístico na 2ª Pessoa**: Quantidade de palavras na 2ª pessoa do caso reto no singular, pois o uso de palavras funcionais é conhecido por fornecer uma maneira não reativa de explorar processos sociais e de personalidade [81, 82].

5. **Estilo Linguístico na 3ª Pessoa**: Quantidade de palavras na 3ª pessoa do caso reto do singular e plural por dia, pois o uso de palavras funcionais é conhecido por fornecer uma maneira não reativa de explorar processos sociais e de personalidade [81, 82].

6. **Valência de Emoções**: Cálculo da média de variância (estado emocional) por dia, usando-se a base de dados ANEW-BR [27,75]. Indivíduos com depressão geralmente em estado emocional negativo, apresentam sinais de depressão e caso contrário, apresentam-se em estado positivo.

7. **Ativação de Emoções**: Cálculo da média de ativação (reação emocional) por dia, usando-se a base de dados ANEW-BR [27,75]. Indivíduos com depressão geralmente em estado emocional negativo, apresentam sinais de depressão e caso contrário, apresentam-se em estado positivo.

8. **Termos Depressivos**: Cálculo da Média das palavras com valência menor que 4 (palavras negativas como ansiedade, insônia, nervosismos, irritação, etc.) por dia usando-se a base de dados ANEW-BR [27, 75]. Indivíduos com depressão tendem a usar esses termos com frequência nas suas postagens.

9. **Grafo Social**: Quantidade de respostas de seguidores respondendo um *tweet* por dia. A pesquisidora De Choudhury [14] indica que usuários depressivos com poucos dos seguidores interagindo nos *tweets*, possuem maiores sinais de depressão.

10. **Medicamentos Antidepressivos**: Usando-se uma base de dados de nomes e gírias de medicamentos coletados na internet calcula-se a frequência dos nomes e das gírias de medicamentos por dia. Indivíduos com depressão tendem a usar esses nomes em seus *tweets*, possivelmente para receber ajuda sobre seus efeitos ou podem fazer parte de um grupo de usuários que frequentemente usam esses nomes.

11. **Caracteres Orientais**: Quantidade de caracteres no intervalo *unicode* japonês, chinês e coreano por dia. A literatura sobre depressão indica que indivíduos com depressão são isolados, ficando no mundo fictício oriental [77].

12. **Emojis**: Quantidade de *emojis* no intervalo *unicode de emojis* por dia. Indivíduos depressivos ao invés de se expressar com palavras nos *tweets*, podem usar *emojis* no lugar, por facilidade de escrita [72, 73] por exemplo *emoji* :> representa nervoso, que expressa raiva.

13. **Links**: Quantidade de *links* nos *tweets* por dia. Indivíduos depressivos que focam em interesse específico (por exemplo carros) compartilham muitos *links* do assunto no *Twitter* e podem apresentar sinais de padrões de depressão [77].

14. **Mídia**: Quantidade de fotos, vídeos e *gifs* nos *tweets* por dia. Indivíduos depressivos tentam expressar seus sentimentos ou interesses com fotos, vídeos ou *gifs* nos seus *tweets* e podem apresentar sinais de padrões de depressão [77].

15. **Curtidas**: Quantidade de curtidas nos *tweets* por dia. Indivíduos depressivos com pouca curtida nos seus *tweets*, mostram não estarem conseguindo chamar atenção do seus seguidores e podem apresentar sinais de padrões de comportamento de depressão [77].

3.6 Etapa 5 - Qualidade dos Dados

No final do processo de extração de características, usando-se a série de atributo 'volume', a qual possui a quantidade de *tweets* postados de cada usuário, identificam-se *outliers* (anomalia) dos usuários que possuem mais de 300 postagens de *tweets* em um dia, ou seja, é quando o robô fica postando textos aleatórios todo dia. Os usuários que possuem menos de 30 *tweets* na conta do Twitter são usuários inativos. Esses usuários são removidos das séries de atributos.

3.7 Etapa 6 - Cálculo dos Vetores de Características

Para cada conjunto de medidas comportamentais, obtêm-se medidas diárias por usuário, para se construir a série temporal por medida, por usuário ao longo do período, a fim de serem utilizadas nos algoritmos de algoritmos de aprendizagem supervisionado.

Assim, para cada uma das 15 séries de atributos temporais é sumarizadas em 4 métricas estatísticas para cada usuário: frequência média, variância, média móvel ponderada e entropia.

Dada uma série temporal $X_i(0), X_i(1), ..., X_i(t), ..., X_i(N)$ para o i^{th} atributo, os atributos são calculados da seguinte forma:

- **Frequência Média** μ_i: Uma medida do sinal da série temporal de um atributo durante todo o período de análise:

$$\frac{1}{N}\sum_{t=0}^{N} X_i\left(t\right) \tag{3.1}$$

  Onde o N é a quantidade de atributos e o X_i é valor no atributo, durante todo o período.

- **Variância**: A média da frequência média do quadrado dos desvios médio do sinal da série temporal de um atributo ao longo de todo o período:

$$\frac{1}{N}\sum_{t=0}^{N}\left(X_i\left(t\right)-\mu_i\right)^2 \tag{3.2}$$

  Onde N é a quantidade de atributos, X_i é valor no atributo, durante todo o período e μ_i é frequência média.

- **Média Móvel Ponderada**: Tendência relativa de um sinal de série temporal, em comparação com um período fixo anterior. Dada a série temporal acima, e um período de duração de M (= 7) dias, a média é dada como:

$$\frac{1}{N}\sum_{t=0}^{N}\left(X_i\left(t\right)-\left(\frac{1}{\left(t-M\right)}\right)\sum{}_{\left(M\leq k\leq t-1\right)}X_i\left(k\right)\right) \tag{3.3}$$

  Onde N é a quantidade de atributos e $X_i(t)$ é valor no atributo, durante todo o período e $X_i(k)$ é valor no atributo nos últimos 7 dias.

- **Entropia**: A medida de incerteza em um sinal de série temporal. Para a série temporal é calculada como:

$$-\sum_{t=0}^{N}X_i\left(t\right)\log\left(X_i\left(t\right)\right) \tag{3.4}$$

  Onde N é a quantidade de atributos e X_i é valor no atributo, durante todo o período.

Para cada atributo extraído obtém-se a série temporal diária do atributo para cada usuário. Assim, para cada série temporal do atributo são aplicadas 4 métricas estatísticas para geração de um único registro por usuário (Frequência Média, Variância, Média Móvel Ponderada e Entropia). Tendo como consequência a geração de 60 (15x4) atributos para cada usuário com seu rótulo de classe.

3.8 Etapa 7 - Criação das Bases Dados

O primeiro passo, antes de criar as 2 bases de dados é rotular cada observação (registro) de cada candidato na tabela pelo rótulo de classe definido. Em seguida, as duas tabelas das classes depressão e controle são mescladas em uma tabela. Por último, é realizado o método de normalização z-score, padronizando-se observações (linhas), o que significa representá-las com pontos de desvio padrão da média da variável (colunas).

Seja t a observação para um determinado usuário, a padronização do j^{th} valor é calculado pela Equação 3.5.

$$\bigcup_{j=0}^{63} \frac{X_j(t) - \mu_j}{\sigma_j} \tag{3.5}$$

Assim, cada usuário é representado por um vetor de características padronizado de 64 itens com média zero e variância unitária. A importância da normalização de dados vem com a garantia de que muitos modelos de aprendizagem de máquina possam processar com escalas distintas de variáveis.

Com a base de dados criada e pronta para treinar, testar e avaliar modelos de aprendizagem de máquina, antes de usar a base, deve-se enfatizar a divisão dos dados em conjuntos de treinamento, validação e teste, antes de realizar e aplicar os modelos de aprendizagem de máquina supervisionados, de modo a evitar a injeção de viés entre os conjuntos. Por recomendação, deve-se a base em *treino 70% dos dados, validação 15% e teste 15%*, ou talvez, *treino 80%, validação 10% e teste 10%*; o conjunto de treino tem que possuir mais de 60% da base de dados, para se evitar possíveis erros de performance nos modelos de Aprendizagem de Máquina.

Esses cuidados devem ser tomados para aumentar a probabilidade que os modelos podem generalizar para novos dados, em vez de memorizar esses dados, esse caso é comumente chamado de *"overfitting"*, como explicado na Seção 2.11. A seguir são explorados quais processos são colocados em ordem para se evitar *overfitting* e injeção de viés ao apreender esses modelos.

3.9 Etapa 8 - Simulação dos Modelos - Aprendizagem de Máquina

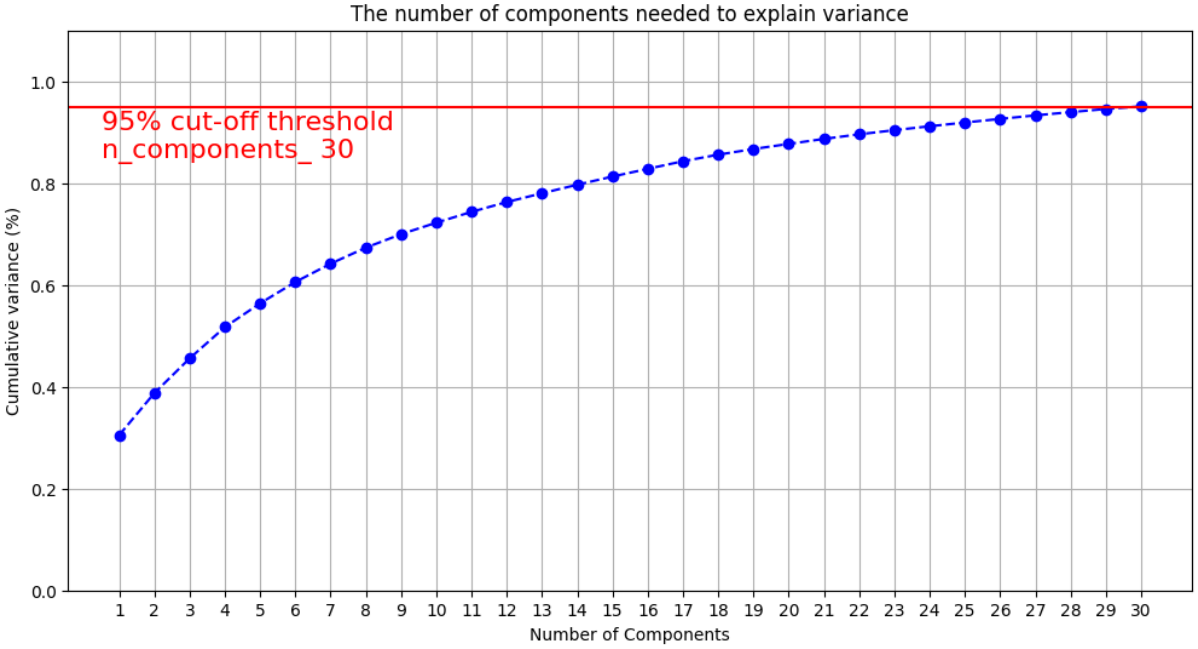
Nesta seção, será explicado como se usará a base de dados gerada na etapa anterior, nos modelos de classificadores de aprendizagem de máquina supervisionado para discriminar usuários da rede social Twitter, entre *depressivos e não depressivos*.

3.9.1 Redução de Dimensionalidade

Um dos muitos métodos para se evitar *overfitting* é reduzir a dimensionalidade dos dados, desse modo foi aplicada a técnica de Análise de Componentes Principais (PCA), como explicado na Seção 2.12.1, para se reduzir a dimensionalidade e capturar 95% da variação nas bases de dados que possuem 60 componentes das 15 séries de atributos calculados

pelos 4 vetores de características o *PCA* identificou 30 componentes principais, como destacado na Figura 3.6.

Figura 3.6: Número de Componentes Necessários para aplicar a soma variância



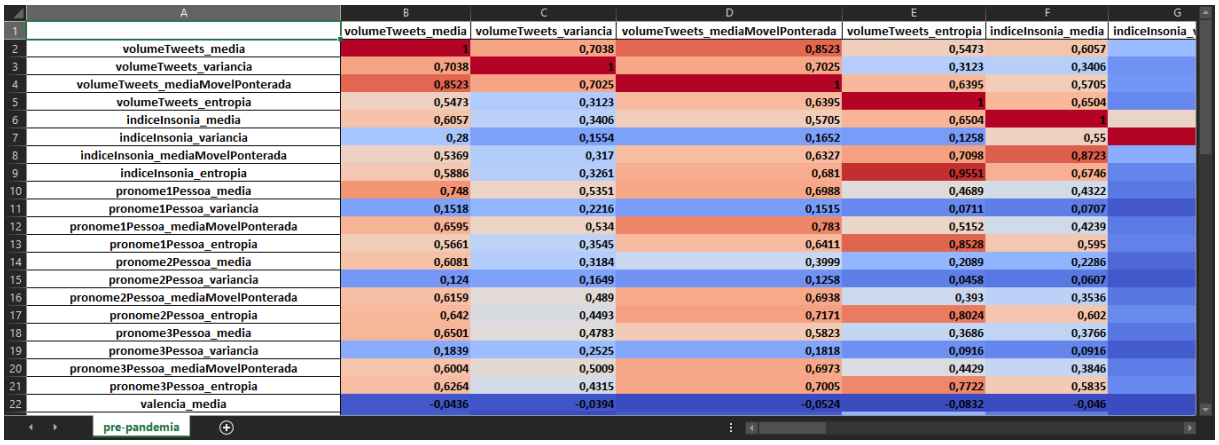
A técnica *PCA* é uma boa estratégia para seleção de atributos, mas é importante destacar que o *PCA* é um algoritmo de aprendizado não supervisionado para ajustar melhor o eixo, para variáveis com base em seus autovetores. No entanto, uma outra boa técnica para selecionar atributos é a Matriz de Correlação de *Person*. As Tabelas 3.7a e 3.7b apresentam uma parte da matriz de correlação das características das bases de dados pré-pandemia e pandemia.

Os valores no espaço que se encontram na cor vermelha, fora da diagonal principal, indicam que o par de características possui uma correlação positiva e quanto mais próximo de "+1"mais correlação eles têm entre si, ou seja, eles estão mais próximos um do outro. Em contrapartida, os valores no espaço da cor azul indicam que o par de características possui uma correlação negativa e quanto mais próximo de -1"mais correlação negativa eles têm, neste caso eles estão muito distantes um do outro.

Nos valores na diagonal principal estão todos os elementos que são iguais a "1", visto que cada variável é totalmente correlacionada com ela mesma, esse valores não contam para a análise de correlação entre as características. Analisando-se a matriz de correlação da Tabela 3.7a, observa-se que a correlação entre os atributos volumeTweets_entropia e indiceInsonia_entropia estão muito próximo de "+1", ou seja, elas possuem uma correlação forte e pode-se excluir um desse atributos e deixar o outro para os algoritmos classificadores. As matrizes de correlação completas podem ser vistas no Apêndice A.

Figura 3.7: Matrizes de Correlação das Bases de Dados Pré-Pandemia e Pandemia

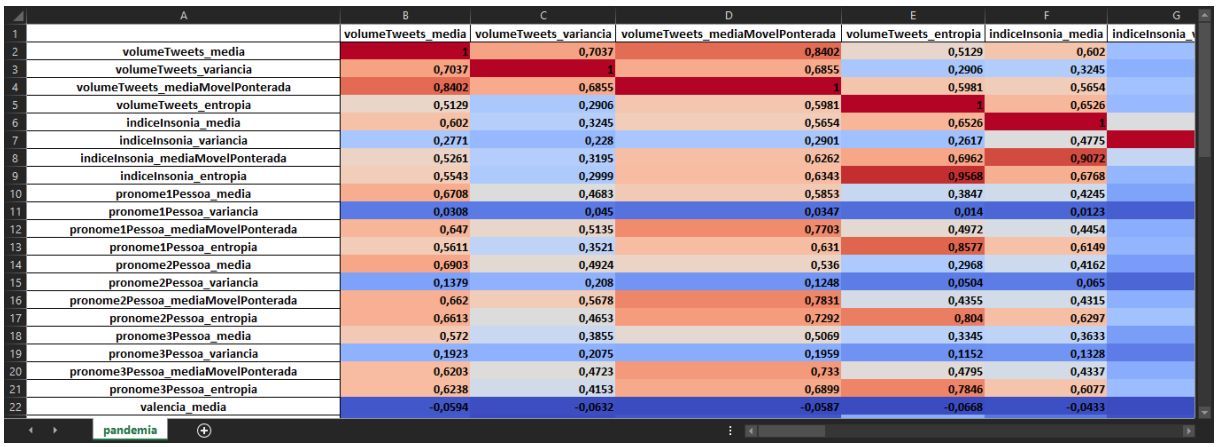
(a) Matriz de Correção da Base de Dados Período Pré-Pandemia



		volumeTweets_media	volumeTweets_variancia	volumeTweets_mediaMovelPonterada	volumeTweets_entropia	indiceInsonia_media	indiceInsonia_
2	volumeTweets_media	1	0,7038	0,8523	0,5473	0,6057	
3	volumeTweets_variancia	0,7038	1	0,7025	0,3123	0,3406	
4	volumeTweets_mediaMovelPonterada	0,8523	0,7025	1	0,6395	0,5705	
5	volumeTweets_entropia	0,5473	0,3123	0,6395	1	0,6504	
6	indiceInsonia_media	0,6057	0,3406	0,5705	0,6504	1	
7	indiceInsonia_variancia	0,28	0,1554	0,1652	0,1258	0,55	
8	indiceInsonia_mediaMovelPonterada	0,5369	0,317	0,6327	0,7098	0,8723	
9	indiceInsonia_entropia	0,5886	0,3261	0,681	0,9551	0,6746	
10	pronome1Pessoa_media	0,748	0,5351	0,6988	0,4689	0,4322	
11	pronome1Pessoa_variancia	0,1518	0,2216	0,1515	0,0711	0,0707	
12	pronome1Pessoa_mediaMovelPonterada	0,6595	0,534	0,783	0,5152	0,4239	
13	pronome1Pessoa_entropia	0,5661	0,3545	0,6411	0,8528	0,595	
14	pronome2Pessoa_media	0,6081	0,3184	0,3999	0,2089	0,2286	
15	pronome2Pessoa_variancia	0,124	0,1649	0,1258	0,0458	0,0607	
16	pronome2Pessoa_mediaMovelPonterada	0,6159	0,489	0,6938	0,393	0,3536	
17	pronome2Pessoa_entropia	0,642	0,4493	0,7171	0,8024	0,602	
18	pronome3Pessoa_media	0,6501	0,4783	0,5823	0,3686	0,3766	
19	pronome3Pessoa_variancia	0,1839	0,2525	0,1818	0,0916	0,0916	
20	pronome3Pessoa_mediaMovelPonterada	0,6004	0,5009	0,6973	0,4429	0,3846	
21	pronome3Pessoa_entropia	0,6264	0,4315	0,7005	0,7722	0,5835	
22	valencia_media	-0,0436	-0,0394	-0,0524	-0,0832	-0,046	

pre-pandemia

(b) Matriz de Correção da Base de Dados Período Pandemia



		volumeTweets_media	volumeTweets_variancia	volumeTweets_mediaMovelPonterada	volumeTweets_entropia	indiceInsonia_media	indiceInsonia_
2	volumeTweets_media	1	0,7037	0,8402	0,5129	0,602	
3	volumeTweets_variancia	0,7037	1	0,6855	0,2906	0,3245	
4	volumeTweets_mediaMovelPonterada	0,8402	0,6855	1	0,5981	0,5654	
5	volumeTweets_entropia	0,5129	0,2906	0,5981	1	0,6526	
6	indiceInsonia_media	0,602	0,3245	0,5654	0,6526	1	
7	indiceInsonia_variancia	0,2771	0,228	0,2901	0,2617	0,4775	
8	indiceInsonia_mediaMovelPonterada	0,5261	0,3195	0,6262	0,6962	0,9072	
9	indiceInsonia_entropia	0,5543	0,2999	0,6343	0,9568	0,6768	
10	pronome1Pessoa_media	0,6708	0,4683	0,5853	0,3847	0,4245	
11	pronome1Pessoa_variancia	0,0308	0,045	0,0347	0,014	0,0123	
12	pronome1Pessoa_mediaMovelPonterada	0,647	0,5135	0,7703	0,4972	0,4454	
13	pronome1Pessoa_entropia	0,5611	0,3521	0,631	0,8577	0,6149	
14	pronome2Pessoa_media	0,6903	0,4924	0,536	0,2968	0,4162	
15	pronome2Pessoa_variancia	0,1379	0,208	0,1248	0,0504	0,065	
16	pronome2Pessoa_mediaMovelPonterada	0,662	0,5678	0,7831	0,4355	0,4315	
17	pronome2Pessoa_entropia	0,6613	0,4653	0,7292	0,804	0,6297	
18	pronome3Pessoa_media	0,572	0,3855	0,5069	0,3345	0,3633	
19	pronome3Pessoa_variancia	0,1923	0,2075	0,1959	0,1152	0,1328	
20	pronome3Pessoa_mediaMovelPonterada	0,6203	0,4723	0,733	0,4795	0,4337	
21	pronome3Pessoa_entropia	0,6238	0,4153	0,6899	0,7846	0,6077	
22	valencia_media	-0,0594	-0,0632	-0,0587	-0,0668	-0,0433	

pandemia

A estratégia de se usar a matriz de correlação de *Person* foi para se visualizar a importância de cada característica e se as bases de dados geradas estão adequadas para o projeto, antes de aplicá-las nos modelos de algoritmos de aprendizagem de máquina supervisionada. Primeiramente, foi analisado a fim de verificar o quanto uma característica estava fortemente correlacionada com a característica alvo (volumeTweets_media), além de identificar se haviam outras características fortemente correlacionadas. No final das análises nas Tabelas 3.7a e 3.7b das duas bases de dados, foram selecionadas 30 características das bases de dados, conforme apresentado na Tabela 3.1.

Tabela 3.1: Características Selecionadas via Análise Matriz de Correlação de Person

volumeTweets_media	volumeTweets_mediaMovelPonterada
indiceInsonia_variancia	indiceInsonia_mediaMovelPonterada
pronome1Pessoa_media	pronome1Pessoa_variancia
pronome2Pessoa_mediaMovelPonterada	pronome2Pessoa_entropia
pronome3Pessoa_media	pronome3Pessoa_mediaMovelPonterada
valencia_mediaMovelPonterada	valencia_entropia
ativacao_mediaMovelPonterada	ativacao_entropia
termosDepressivos_variancia	termosDepressivos_mediaMovelPonterada
grafoSocial_variancia	grafoSocial_mediaMovelPonterada
medicamentosAntiDepressivo_media	medicamentosAntiDepressivo_mediaMovelPonterada
caracteresOrientais_variancia	caracteresOrientais_mediaMovelPonterada
emojis_variancia	emojis_entropia
links_mediaMovelPonterada	links_entropia
midia_variancia	midia_mediaMovelPonterada
curtidas_media	curtidas_mediaMovelPonterada

3.9.2 Dividindo os Dados

Como explicado na seção 2.7, sobre as divisão dos dados, a qual ajuda na construção de um modelo de aprendizagem supervisionado eficiente, sem erros de *overfitting* nesse projeto, foi dividida a nossa base de dados de forma aleatória em: *conjunto de treinamento (70%), conjuntos de validação (15%) e conjunto teste (15%).* Essa partição no conjunto de dados, tem por finalidade garantir a isenção na análise, de tal forma que os dados de testes não sejam os mesmos utilizados durante a aplicação dos conjuntos de treinamento e de validação, pois eles têm a finalidade de criar um bom modelo de aprendizagem de máquina. Os dados de teste serão aplicados no modelo de aprendizagem de máquina obtido através do treinamento.

3.9.3 Modelos de Aprendizagem de Máquina

Tendo realizado a dimensionalidade e divisão dos dados, é hora de se discutir os algoritmos de aprendizagem de máquina supervisionado. Aqui, são apresentados especificidades dos classificadores e os hiperparâmetros adotados, para se ter uma melhor superfície, no entendimento de como os algoritmos funcionam. Como nota técnica, todos os modelos de classificadores são treinados com o pacote *Scikit-Learn do Python.*

Foram aplicados os conjuntos de treino e de validação no modelo escolhido como *baseline*, um modelo *baseline* tem como propósito avaliar se outros modelos estão tendo

bom desempenho no projeto em questão. Para modelos *não-baseline* foram utilizados 15 modelos de aprendizagem supervisionados com seguintes hiperparâmetros:

1. **Regressão Logística (RT) (Baseline)**: no classificador as probabilidades que descrevem os possíveis resultados de uma única tentativa são modeladas usando-se uma função logística. Esse algoritmo foi como *baseline* e os hiperparâmetros foram definidos com valores *default* da classe do algoritmo.

2. **Análise Discriminante Linear (LDA)**: Um classificador com um limite de decisão linear, gerado ajustando-se densidades condicionais de classe aos dados e usando-se a regra de Bayes. Os hiperparâmetros utilizados no algoritmo são:

   - **solver (solução)**: solução para otimização de peso - Valor setado: 0;
   - **shrinkage**: parâmetro de contração - Valor setado: 'lsqr';
   - **tol**: limiar absoluto para que um valor singular de "X"seja considerado significativo - Valor setado: 1e-6.

3. **Arvore de Decisão (DT)**: Um classificador que preveja o valor de uma variável de destino apreendendo regras de decisão simples, inferidas a partir dos recursos de dados. Os hiperparâmetros utilizados no algoritmo são:

   - **max_depth (máximo de profundidade)**: a profundidade máxima da árvore. Se None (Nenhum) valor, os "nós"são expandidos até que todas as folhas sejam puras ou até que todas as folhas contenham menos de um número mínimo de amostras - Valor definido: 9.

4. **k-vizinhos mais Próximos (KNN)**: a classificação baseada em vizinhos é um tipo de aprendizado baseado em instância ou aprendizado não generalizante. A classificação é calculada a partir de uma votação majoritária simples dos vizinhos mais próximos de cada ponto. Os hiperparâmetros utilizados no algoritmo são:

   - **n_neighbors (número de vizinhos)**: número de vizinhos - Valor definido: 9;

5. **Naive Bayes (NB) Bernoulli**: classificador *Naive Bayes* para modelos multivariados de Bernoulli. Os hiperparâmetros utilizados no algoritmo são:

   - **alpha**: parâmetro de suavização aditiva (*Laplace/Lidstone*) - Valor definido: 1e-09.

6. **Perceptron Multi-Camadas (MLP)**: o classificador otimiza a função log-loss usando *LBFGS* ou gradiente descendente estocástico. Os hiperparâmetros utilizados no algoritmo são:

    - **hidden_layer_sizes (tamanhos de camadas ocultas)**: representa número de neurônios na *i-ésima* camada oculta - Valor setado: (10,30,10);
    - **max_iter (número de epochs (épocas))**: número máximo de interações. A solução intera até a convergência (determinada por *'tol'*) ou este número de interações para solucionadores estocásticos (*'sgd', 'adam'*); observe que isso determina o número de épocas (quantas vezes cada ponto de dados será usado), não o número de etapas de gradiente - Valor definido: 100.

7. **Máquina de Vetores de Suporte (SVM) Linear**: a implementação é baseada em liblinear, para mais flexibilidade na escolha de penalidades e funções de perda e deve escalar melhor para um grande número de amostras. Os hiperparâmetros utilizados no algoritmo são:

    - **C**: Parâmetro de regularização. A força da regularização é inversamente proporcional a *"C"*. Deve ser estritamente positiva. A penalidade é uma penalidade de 12 ao quadrado - Valor definido: 1;

8. **Floreta Randômica (RF)**: um classificador que ajusta vários classificadores de árvore de decisão em várias subamostras do conjunto de dados e usa a média para melhorar a precisão preditiva e controlar o ajuste excessivo. Os hiperparâmetros utilizados no algoritmo são:

    - **n_estimators (número de estimadores)**: O número de árvores na floresta - Valor definido: .
    - **max_depth (máximo de profundidade)**: A profundidade máxima da árvore. Se None (Nenhum) valor, os "nós"são expandidos até que todas as folhas sejam puras ou até que todas as folhas contenham menos de um número mínimo de amostras - Valor definido: .

9. **Gradient Boosting (GB)**: um classificador que constrói um modelo aditivo de forma progressiva; ele permite a otimização de funções de perda diferenciáveis arbitrárias. Os hiperparâmetros utilizados no algoritmo são:

    - **n_estimators (número de estimadores)**: o número de árvores na floresta - Valor definido: .

10. **Bagging**: é um meta-estimador de conjunto que ajusta os classificadores base em subconjuntos aleatórios do conjunto de dados original e, em seguida, agrega suas previsões individuais (por votação ou por média) para formar uma previsão final. Os hiperparâmetros utilizados no algoritmo são:

    - **base_estimator**: O estimador base a partir do qual o conjunto impulsionado é construído.
    - **n_estimators**: O número de estimadores de base no conjunto.

11. **Boosting**: é um meta-estimador que começa ajustando um classificador no conjunto de dados original e, em seguida, ajusta cópias adicionais do classificador no mesmo conjunto de dados, mas onde os pesos de instâncias classificadas incorretamente são ajustados de modo que os classificadores subsequentes se concentrem mais em casos difíceis. Os hiperparâmetros utilizados no algoritmo são:

    - **base_estimator**: O estimador base a partir do qual o conjunto impulsionado é construído.
    - **n_estimators**: O número de estimadores de base no conjunto.

12. **Votação (Soft e Hard)**: é um método de combinar vários classificadores de aprendizado de máquina conceitualmente diferentes e usar uma votação majoritária ou as probabilidades previstas médias (voto suave - soft) para prever os rótulos de classe. Os hiperparâmetros utilizados no algoritmo são:

    - **estimators**: lista de modelos de classificadores a serem combinados.
    - **voting**: Se 'hard', usa rótulos de classe previstos para votação da regra da majoritária. Caso contrário, se 'soft', prevê o rótulo da classe com base nas somas das probabilidades previstas.

3.10 Considerações Finais

Neste capítulo foi discutida toda a metodologia desenvolvida neste projeto etapa a etapa. Considere-se uma boa prática esses métodos e resultados, concluindo-se que projetos futuros podem explorar e expandir esse trabalho sem quaisquer dificuldades. O próximo capítulo será apresentado à luz dos resultados, analisando-se os resultados de busca das palavras-chave de autodeclarações de situação de depressão, visualizando-se termos mais frequentes através da nuvem de palavras e da análise do desempenho dos modelos de classificadores e aprendizagem de máquina.

Os dados e códigos utilizados neste projeto encontram-se no Github[2]. Para executar os códigos para construir as bases de dados, siga o está especificado no README ou siga os métodos das etapas numeradas no script metodologia.py.

[2] `https://github.com/luanfreitas5/UnBSense`

Capítulo 4

Resultados

Este capítulo apresenta todos os resultados dessa pesquisa do projeto, resultados obtidos da análise exploratória de dados nos arquivos de busca de candidatos e dos *tweets* coletados nas classes *depressivas* e *controle* nos períodos pré-pandemia e pandemia. Na Seção 4.1 foram discutidos os resultados da exploração de dados busca de candidatos, efetuada através das palavras-chave. Na Seção 4.2 mostra-se a visualização de dados dos termos mais frequentes nos *tweets* nas classes depressiva e controle através da técnica nuvem de palavras. Na Seção 4.3 apresentam-se os resultados dos experimentos dos modelos de aprendizagem de máquina supervisionados com aplicação dos dados de treinamento e validação. Na Seção 4.4 encerra-se com os resultados finais, expondo-se os melhores modelos que se destacaram nas simulações de treino e de validação, aplicando-se somente os dados de teste e apenas uma só vez.

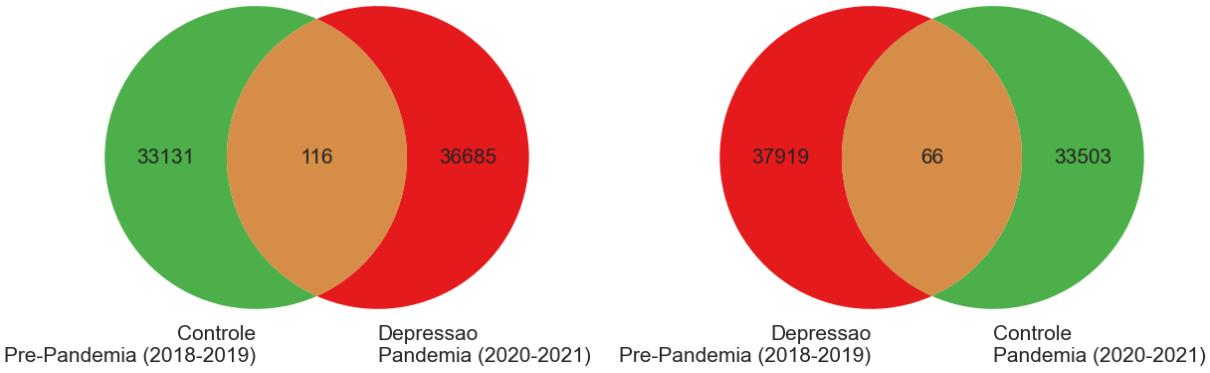
4.1 Exploração de Dados

Explorando-se as bases de dados geradas foi encontrada uma interseção entre as bases: Como mostrado na Figura 4.1a, no período pré-pandemia foram encontrados 116 usuários que não eram depressivos, mas que tornaram-se depressivos no período da pandemia. A COVID-19 no Brasil pode ter sido o gatilho para despertar depressão internas nas pessoas. Nos não reunimos informações suficientes para comprovar essa teoria. Por outro lado, na Figura 4.1b, mostra que no período pandemia 66 usuários não depressivos, mas que no período da pré-pandemia se encontravam depressivos. De acordo com a literatura sobre psicologia humana, psiquiatria, neurociência e sociolinguística sobre saúde humana, vários fatores não relacionados a COVID-19 podem causar depressão nas pessoas, como muito consumo de álcool e outros.

Figura 4.1: Interseções entre as Base de Dados

(a) Interseção Controle Pré-Pandemia Depressão Pandemia

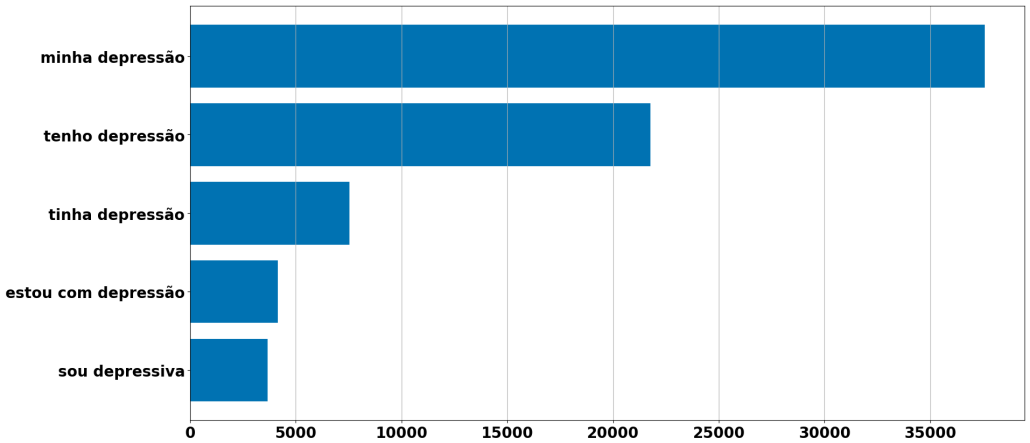
(b) Interseção Depressão Pré-Pandemia Controle Pandemia



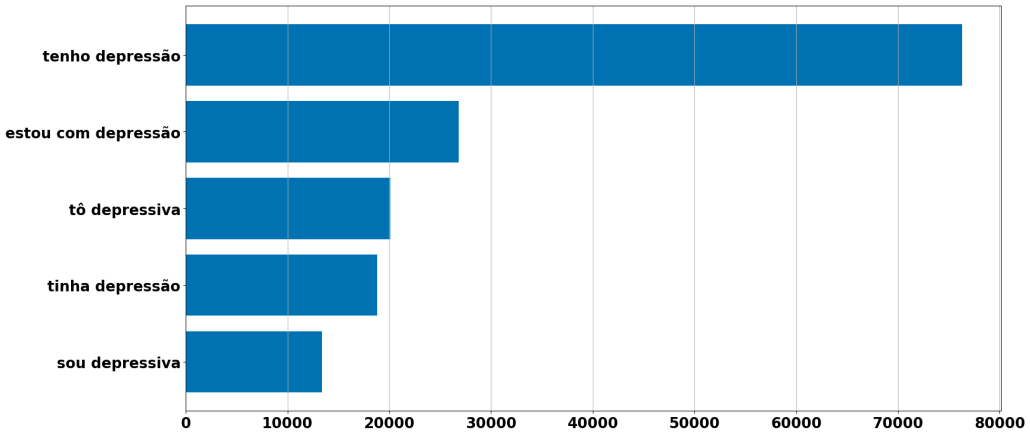
Na Figura 4.2 apresentam-se o Top 5 das palavras-chaves, frequentes encontradas nos resultados de busca e filtragem de usuários que se autodeclararam depressivos nos 2 períodos: pré-pandemia (2018-2019) e pandemia (2020-2021). No Gráfico 4.2a mostra-se que no período pré-pandemia (2018-2019) o quantitativo de declarações depressivas da palavra-chave "minha depressão"com $\sim$ 37.000 ocorrências e "tenho depressão"com $\sim$ 21.000 ocorrências. No Gráfico 4.2b mostra-se que no período da pandemia (2020-2021), houve muitos relatos de usuários no *Twitter* se autodeclarando depressivos, tendo mais do dobro do quantitativo de frequência de autodeclarações depressivas da pré-pandemia. Esse aumento de autodeclarações pode ter sido resultado dos aumentos de casos de COVID-19 no Brasil, obrigando as pessoas a se isolarem em suas casas e com isso levando-as à depressão, mas não conseguimos obter informações suficientes para comprovar.

Figura 4.2: Top 5 - Frequência de Declarações de Depressão

(a) Frequência de Autodeclarações Depressão no Período Pré-Pandemia



(b) Frequência de Autodeclarações de Depressão no Período Pandemia



4.2 Nuvens Palavras

Nesta seção apresentam-se as nuvens de palavras geradas de termos mais frequentes nos *tweets* em cada classe de cada período, também foi gerada a nuvem de palavras de todos os *tweets* da classes depressão e controle juntas; cada nuvem possui um gráfico de barras e de pizza no top de 10 palavras mais usadas. As nuvens de palavras ajudam a identificar quais são os termos mais comuns que as pessoas escrevem em seus textos, tirando as palavras consideradas *stopwords* que são palavras que não tem nenhuma relevância para análise, por exemplo *"as"*, *"e"*, *"os"*, *"de"*, *"para"*, *"com"*, *"sem"*, *"foi"*, etc., essas palavras devem ser excluídas dos textos, antes do início do procedimento executado. As nuvem de palavras podem ajudar futuramente a identificar novos usuários depressivos na rede social Twitter, buscando-se *tweets* através desses termos.

4.2.1 Nuvem Palavras de Depressão Pré-Pandemia

Na Figura 4.3a, ilustra-se a nuvem de palavras de todos *tweets* postados por usuários candidatos de depressão no período pré-pandemia; através do Gráfico 4.3b, analisa-se que os termos mais predominantes usados pelos usuários depressivos em seus *tweets* no período pré-pandemia são *"vida"*, com mais de 800.000 ocorrências, *"deus"* com uma taxa próxima de 800.000 ocorrências, *"mano"* com uma taxa próxima de 700.000 ocorrências e *"mundo"* no total de 600.000 ocorrências.

Figura 4.3: Análise de Tweets de Usuários de Depressão - Pré-Pandemia

(a) Nuvem de Palavra de Tweets de Usuários de Depressão - Pré-Pandemia

(b) Top 5 de Palavras de Tweets de Usuários de Depressão - Pré-Pandemia

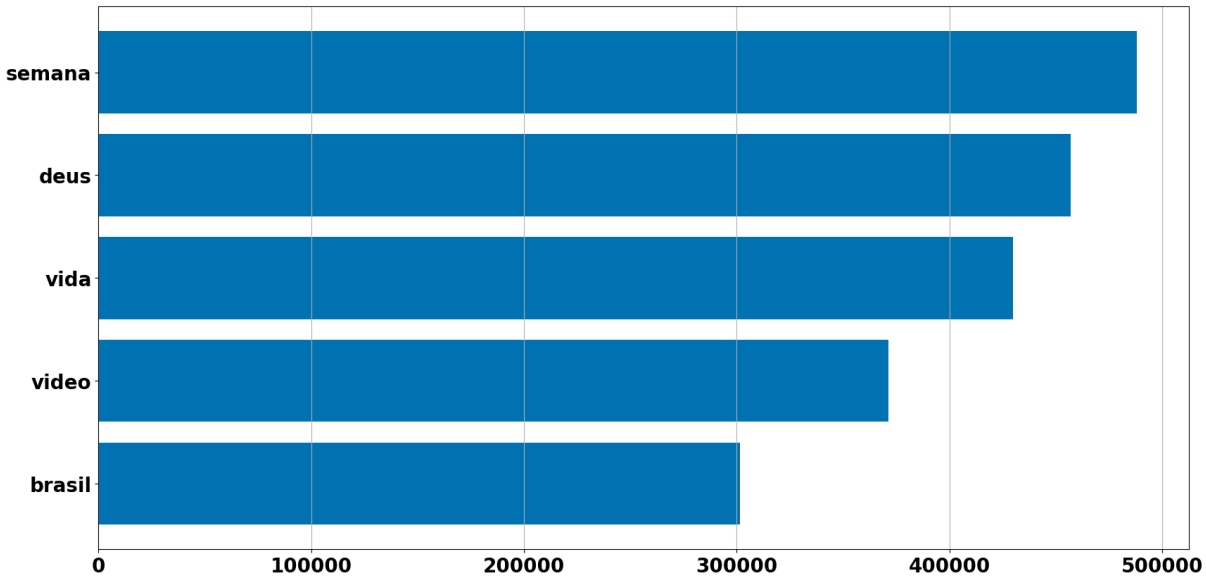
4.2.2 Nuvem Palavras de Controle Pré-Pandemia

Na Figura 4.4a, mostra-se a nuvem de palavras de todos os *tweets* postados por usuários candidatos de controle no período pré-pandemia; através do Gráfico 4.4b, analisa-se que os termos mais predominantes usados pelos usuários de controle em seus *tweets* no período pré-pandemia são *"semana"* com uma taxa próxima de 500.000 ocorrências, *"deus"* com uma taxa de próxima de 450.000 ocorrências, *"vida"* com mais de 400.000 ocorrências e *"video"* com mais de 300.000 ocorrências.

Figura 4.4: Análise de Tweets de Usuários de Controle - Pré-Pandemia

(a) Nuvem de Palavras de Tweets de Usuários de Controle - Pré-Pandemia

(b) Top 5 de Palavras de Tweets de Usuários de Controle - Pré-Pandemia



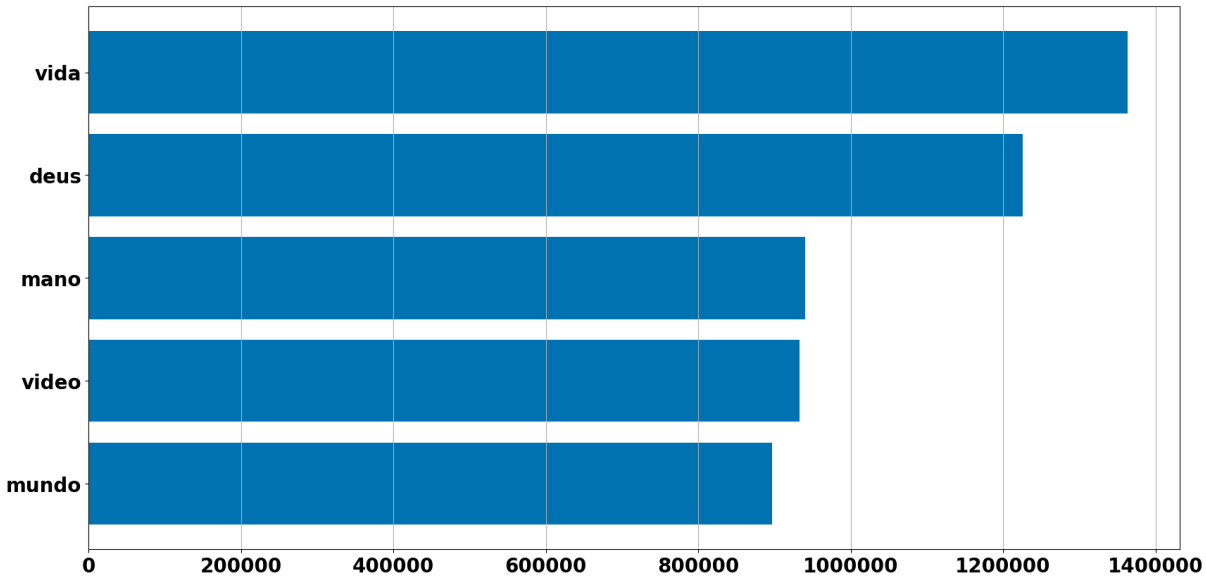
4.2.3 Nuvem Palavras de Depressão e Controle Pré-Pandemia

Na Figura 4.5a, apresenta-se a nuvem de palavras de todos *tweets* postados por usuários candidatos das classes de depressão e de controle no período pré-pandemia; através do Gráfico 4.5b, analisa-se que os termos mais predominantes usados pelos usuários em seus *tweets* no período pré-pandemia são *"vida"* com uma taxa próxima de 1.400.000 ocorrências, *"deus"* com mais de 1.200.000 ocorrências, *"mano"* e *"video"* com taxas próximas de 900.000 ocorrências.

Figura 4.5: Análise de Tweets de Usuários de Depressão e Controle - Pré-Pandemia

(a) Nuvem de Palavras de Tweets de Depressão e Controle - Pré-Pandemia

(b) Top 5 de Palavras de Tweets de Depressão e Controle - Pré-Pandemia



4.2.4 Nuvem Palavras de Depressão Pandemia

Na Figura 4.6a, ilustra-se a nuvem de palavras de todos os *tweets* postados por usuários candidatos de depressão no período pandemia; através do Gráfico 4.6b, analisa-se que os termos mais predominantes usados pelos usuários depressivos em seus *tweets* no período pandemia são *"deus"* com mais de 700.000 ocorrências, *"vida"* com mais de 600.000 ocorrências, *"mano"* com uma taxa aproximadamente 450.000 ocorrências e *"mundo"* com uma taxa próxima ao termo *"mano"*.

Figura 4.6: Análise de Tweets de Usuários de Depressão - Pandemia

(a) Nuvem de Palavras de Tweets de Usuários de Depressão - Pandemia

(b) Top 5 de Palavras de Tweets de Usuários de Depressão - Pandemia

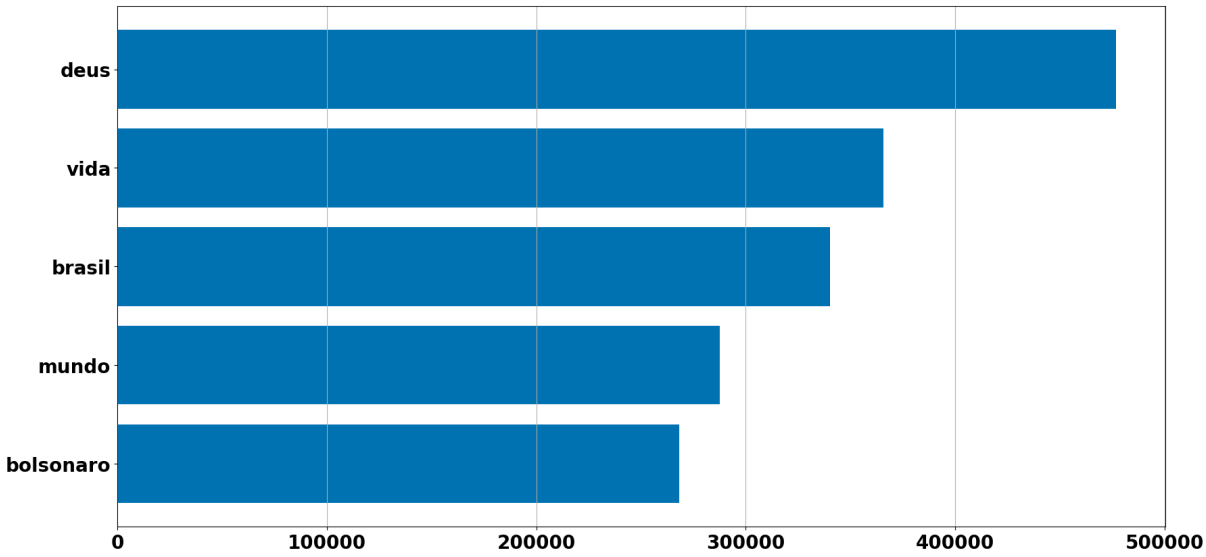
4.2.5 Nuvem Palavras de Controle Pandemia

Na Figura 4.7a, mostra-se a nuvem de palavras de todos os *tweets* postados por usuários candidatos da classe de controle no período pandemia; através do Gráfico 4.7b foi analisado que os termos mais predominantes usados pelos usuários de controle em seus *tweets* no período pré-pandemia são *"deus"* com uma taxa próxima de 500.000 ocorrências, *"vida"* com mais de 350.000 ocorrências, *"brasil"* com mais de 300.000 ocorrências e *"mundo"* com uma taxa próxima a 300.000 ocorrências.

Figura 4.7: Análise de Tweets de Usuários de Controle - Pandemia

(a) Nuvem de Palavras de Tweets de Usuários de Controle - Pandemia

(b) Top 5 de Palavras de Tweets de Usuários de Controle - Pandemia



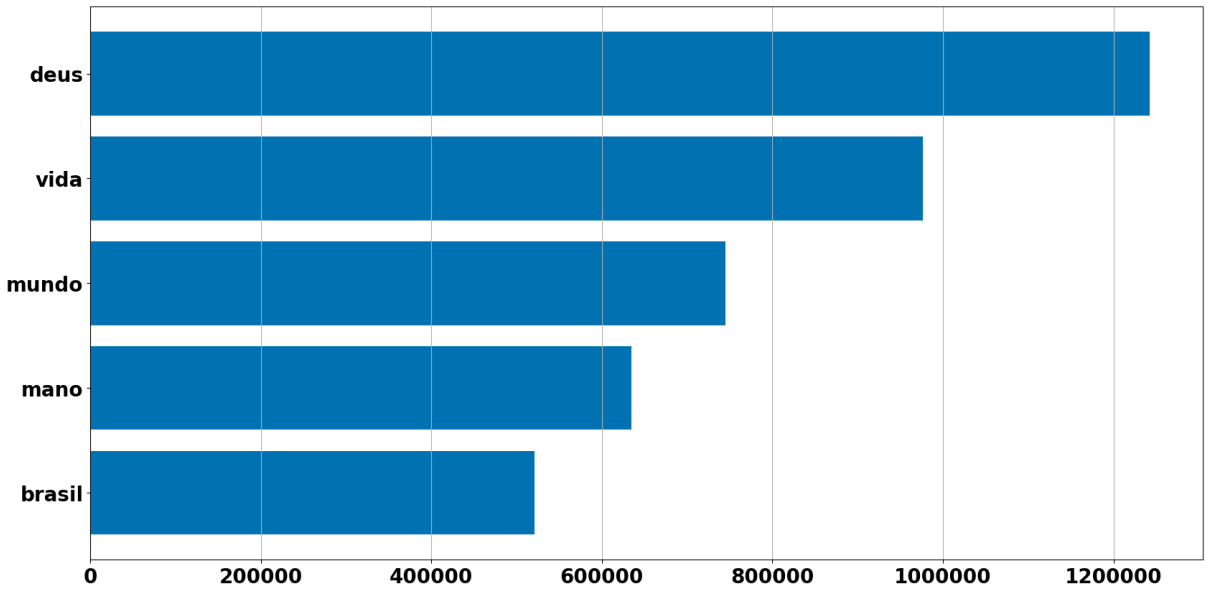
4.2.6 Nuvem Palavras de Depressão e Controle Pandemia

Na 4.8a, apresenta-se a nuvem de palavras de todos *tweets*, os postados por usuários candidatos da classe de depressão e de controle no período pandemia; através do Gráfico 4.8b foi analisado que os termos mais predominantes usados pelos usuários em seus *tweets* no período pandemia são *"deus"* com mais 1.200.000 ocorrências, *"vida"* com uma taxa próxima de 1.000.000 ocorrências e *"mundo"* com taxa próxima de 800.000 ocorrências.

Figura 4.8: Análise de Tweets de Usuários de Depressão e Controle - Pandemia

(a) Nuvem de Palavras de Tweets de Depressão e Controle - Pandemia

(b) Top 5 de Palavras de Tweets de Depressão e Controle - Pandemia



Nas nuvens de palavras podem ter ainda palavras irrelevantes como gírias, que é

costume de muitos brasileiros nas rede sociais hoje em dia ou ainda *stopwords* que é uma lista enorme de palavras vazias no português brasileiro, mas a técnica de nuvem de palavras pode ser útil para futuras análises de textos de *tweets* de usuários depressivos na rede social *Twitter* e também de outras rede sociais com *Facebook*.

4.3 Etapa 8 - Simulação dos Modelos - Treino e Validação

Nesta seção serão apresentados os resultados dos experimentos realizados com os 2 melhores modelos, um de classificação e um de ensemble pré-selecionados na avaliação dos 13 modelos. A análise dos resultados tem 2 objetivos: primeiro objetivo, verificar se os atributos da pesquisadora De Choudhury [14] *(volume de tweets, índice de insônia, estilo linguístico, emoções, termos depressivos, grafo social e medicamentos antidepressivos)* com adição de novos atributos *(caracteres orientais, emojis, mídia, frequência de links e número de curtidas)* se melhoram ou não as predições dos algoritmos de aprendizagem de máquina; segundo objetivo, encontrar os melhores algoritmos de aprendizagem de máquina para predizer se um usuário do *Twitter* possui sinais de padrões de comportamento de depressão ou não.

A Tabela 4.1 apresenta o desempenho dos 13 modelos induzidos com validação cruzada 10-folds e seus hiperparâmetros configurados via GridSearchCV nas bases pré-pandemia (2018-2019) e pandemia (2018-2019), somente com os 10 atributos da De Choudhury [14], utilizando as métricas de precisão (P), recall (R), f1-Score (F1) e acurácia (ACC) para avaliar e comparar os modelos, tendo como métrica alvo o f1-Score.

Tabela 4.1: Pré-Seleção dos Modelos

Pré-Pandemia					**Pandemia**				
Classificadores					**Classificadores**				
Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-Score**	**Modelo**	**Acurácia**	**Precisão**	**Recall**	**F1-Score**
Perceptron Multicamadas (MLP)	**76.1**	**75.5**	**87.4**	**81.0**	**Perceptron Multicamadas (MLP)**	**74.5**	**74.5**	**87.4**	**80.4**
Árvore de Decisão (DT)	74.7	74.1	86.9	80.0	**Árvore de Decisão (DT)**	72.9	72.9	87.2	79.4
KNN	70.5	70.6	84.7	77.0	**SVM**	65.9	65.9	89.2	75.8
SVM	68.3	68.9	83.1	75.3	**KNN**	68.0	69.3	83.5	75.7
Análise Discriminante Linear (LDA)	66.9	67.3	84.2	74.8	**Análise Discriminante Linear (LDA)**	65.0	64.9	90.7	75.6
Regressão Logística (LR) (Baseline)	58.3	58.3	100.0	73.7	**Regressão Logística (LR) (Baseline)**	59.9	59.9	100.0	74.9
Naive Bayes (NB)	59.1	71.3	49.9	58.7	**Naive Bayes (NB)**	57.8	70.9	50.3	58.9
Ensemble					**Ensemble**				
Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-Score**	**Modelo**	**Acurácia**	**Precisão**	**Recall**	**F1-Score**
Gradient Boosting (GB)	**77.2**	**75.5**	**90.1**	**82.2**	**Gradient Boosting (GB)**	**75.5**	**74.4**	**90.0**	**81.5**
Boosting	77.3	76.2	88.8	82.0	**Boosting**	75.6	75.0	88.8	81.3
Floresta Randômica (RF)	76.6	74.8	90.1	81.8	**Floresta Randômica (RF)**	74.8	73.5	90.6	81.2
Bagging	76.5	75.9	87.6	81.3	**Bagging**	75.1	75.2	87.3	80.8
Votação Soft	76.1	75.8	86.7	80.9	**Votação Hard**	74.5	74.4	87.7	80.5
Votação Hard	75.7	75.7	85.8	80.5	**Votação Soft**	74.1	74.4	86.4	80.0

Os modelos Gradient Boosting (GB) e Perceptron Multicamadas (MLP) foram se destacam melhor entre os modelos e serão usados nos experimentos para analisar quais

dos 5 novos atributos (caracteres orientais, emojis, mídias, frequência de links e número de curtidas) tem efeito significativos na interação com os 10 atributos da pesquisadora De Choudhury [14] (volume de tweets, índice de insônia, estilo linguístico de 1ª, 2ª e 3ª pessoa, emoções de valência e ativação, termos depressivos, grafo social e medicamentos antidepressivos).

A Tabela 4.2, apresenta o desempenho dos modelos Gradient Boosting (GB) e Perceptron Multicamadas (MLP) induzidos nos dados de validação das 2 bases de dados pré-pandemia (2018-2019) e pandemia (2020-2021). Os experimentos são realizados induzindo um modelo com os 10 atributos da pesquisadora De Choudhury [14] interagindo com cada um dos 5 novos atributos, se um atributo melhorar o desempenho do modelo no resultado do experimento anterior ele é atribuído junto com os 10 atributos da pesquisadora no experimento seguinte e excluído na lista de interações, repetindo até resultado aceitável.

Tabela 4.2: Experimentos de Validação (Pré-Pandemia e Pandemia)

Pré-Pandemia (2018-2019)																								
Modelo	**Atributos De Choudhury**				**Atributos De Choudhury + Caracteres Orientais**				**Atributos De Choudhury + Emojis**				**Atributos De Choudhury + Links**				**Atributos De Choudhury + Midia**				**Atributos De Choudhury + Curtidas**			
	ACC	**P**	**R**	**F1**	**ACC**	**P**	**R**	**F1**	**ACC**	**P**	**R**	**F1**	**ACC**	**P**	**R**	**F1**	**ACC**	**P**	**R**	**F1**	**ACC**	**P**	**R**	**F1**
GB	77.2	75.5	90.1	**82.2**	77.2	75.5	90.2	**82.2**	77.3	75.9	89.5	**82.1**	77.0	75.4	89.9	**82.0**	77.3	75.6	90.1	**82.2**	77.1	75.4	90.0	**82.1**
MLP	76.1	75.5	87.4	**81.0**	76.4	76.0	87.0	**81.1**	76.3	75.9	86.9	**81.0**	76.3	75.7	87.5	**81.2**	76.1	75.6	87.2	**81.0**	76.2	75.6	87.5	**81.1**
Pandemia (2020-2021)																								
Modelo	**Atributos De Choudhury**				**Atributos De Choudhury + Caracteres Orientais**				**Atributos De Choudhury + Emojis**				**Atributos De Choudhury + Links**				**Atributos De Choudhury + Midia**				**Atributos De Choudhury + Curtidas**			
	ACC	**P**	**R**	**F1**	**ACC**	**P**	**R**	**F1**	**ACC**	**P**	**R**	**F1**	**ACC**	**P**	**R**	**F1**	**ACC**	**P**	**R**	**F1**	**ACC**	**P**	**R**	**F1**
GB	75.5	74.4	90.0	**81.5**	75.5	74.4	90.0	**81.5**	76.0	74.8	90.3	**81.8**	75.6	74.5	90.0	**81.5**	75.4	74.5	89.7	**81.4**	75.6	74.5	90	**81.5**
MLP	74.5	74.5	87.4	**80.4**	74.2	74.0	87.8	**80.3**	75.1	75.1	87.5	**80.8**	74.6	74.6	87.4	**80.5**	74.5	74.2	88.1	**80.6**	74.5	74.2	88	**80.6**

Nenhum dos novos atributos teve um alto desempenho significativo com os atributos do Choudhury [14] nos modelos Gradient Boosting (GB) e Perceptron Multicamadas (MLP), tendo somente pouquíssimo desempenho em relação ao f1-score aos modelos do experimento básico de De Choudhury [14]. Então devemos complementar que, os 5 novos atributos (caracteres orientais, emojis, mídias, frequência de links e número de curtidas) não são úteis para construção de um modelo promissor capaz de detectar sinais de padrão de comportamento depressivo.

4.3.1 Analise dos Resultados

Analisando-se as tabelas, percebeu-se que a adição dos novos atributos *(caracteres orientais, emojis, mídias, links e curtidas)* não tiveram alto desempenho significativo com os 10 atributos do Choudhury [14]. Logo, a incrementação desses atributos não são úteis para detectar sinais de padrões de comportamento depressivos e não depressivos nos usuários do Twitter.

Conclui-se, então, que os modelos Gradient Boosting (GB) e Perceptron Multicamadas (MLP), apresentaram excelente desempenho com os dados de validação das bases de dados pré-pandemia e pandemia. Esses modelos *Ensemble* podem ser mais adequados para detectar se um usuário da rede social *Twitter* possui sinais de padrão de comportamento depressivo ou não, mas ainda será aplicado com os conjuntos de testes das duas bases de dados que serão apresentados na próxima seção.

4.4 Etapa 9 - Avaliação dos modelos - Teste

Nesta seção serão apresentados os resultados dos experimentos de testes (somente com os atributos da De Choudhury [14]) e o f1-score médio dos experimentos dos modelos Gradient Boosting (GB) e Perceptron Multicamadas (MLP) das bases de dados pré-pandemia (2018-2019) e pandemia (2020-2021). Esses experimentos têm como propósito avaliar se os modelos são adequados para detecção de sinais de padrões de comportamento depressivos, quando os modelos receberem novos dados.

Analisando a Tabela 4.3, nota-se que os resultados estão próximos aos resultados dos experimentos realizados com treino e validação, ou seja, estão aptos para serem utilizados para detectar sinais de padrões de comportamento de depressão, quando estes recebem novos dados (*tweets*) de novos candidatos. Os modelos não estariam adequados se os resultados de teste ultrapassassem os resultados de treino e de validação.

Os resultados dos modelos nas bases de dados pré-pandemia e pandemia foram promissores. O modelo Gradient Boosting na base pré-pandemia (2018-2019) teve desempenho f1-score de 81.6% e na base pandemia (2020-2021) teve desempenho f1-score de 79.9%. O modelo Perceptron Multicamadas na base pré-pandemia (2018-2019) teve desempenho f1-score de 80.6% e na base pandemia (2020-2021) teve desempenho f1-score de 79.4%. Esses resultados obtidos são ótimos para o propósito geral do projeto que é detectar sinais de padrões de comportamento de depressão em usuários na rede social Twitter.

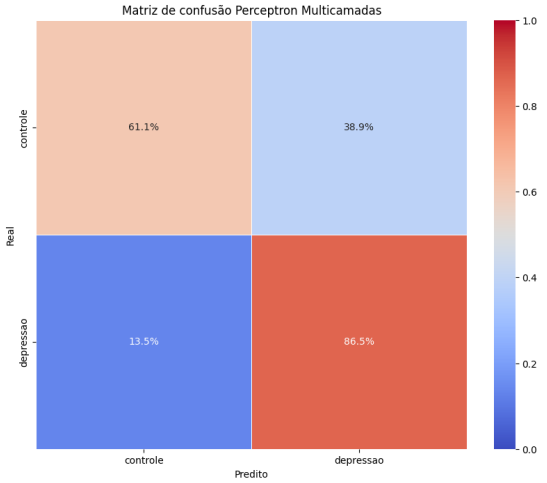
Tabela 4.3: Experimentos de Teste (Pré-Pandemia e Pandemia)

Modelo	Pré-Pandemia (2018-2019)				Pandemia (2020-2021)				Média F1-Score
	Experimento - 10 Atributos De Choudhury				Experimento - 10 Atributos De Choudhury				Experimento -
	Acurácia	Precisão	Recall	F1-Score	Acurácia	Precisão	Recall	F1-Score	10 Atributos De Choudhury
Gradient Boosting (GB)	76.8	75.5	88.9	**81.6**	74.2	73.2	87.8	**79.9**	**80.75**
Perceptron Multicamadas (MLP)	75.8	75.4	86.5	**80.6**	74.2	74.2	85.4	**79.4**	**80.0**

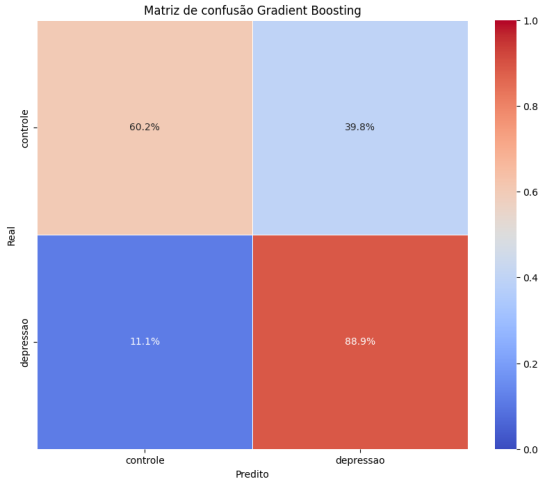
A Figura 4.9, apresenta as matrizes de confusões dos melhores modelos executados, com conjuntos de teste das bases de dados pré-pandemia e pandemia, analisando as matrizes nota-se que a classe dos modelos que mais predizem os usuários candidatos do *Twitter* é depressão: com taxa de acertos entre 88% e 89% e com taxa de erro entre 12% e 11%; em compartida, na *classe controle*, os modelos obtiveram resultados equilibrados com taxa de acertos entre 61% e 57% e com taxa de erro entre 39% e 43%. Essas resultados dos modelos que confundiram dados (palavras) nos textos dos *tweets* da *classe de depressão*, com os dados da *classe de controle*.

Figura 4.9: Matriz de confusões das Base de Dados Pré-Pandemia e Pandemia

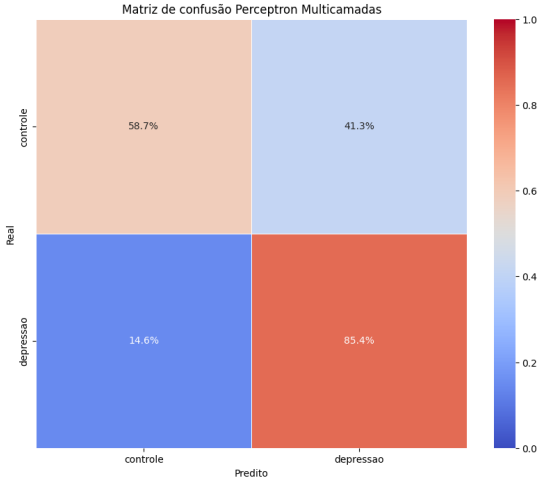
(a) MLP - Pré-Pandemia



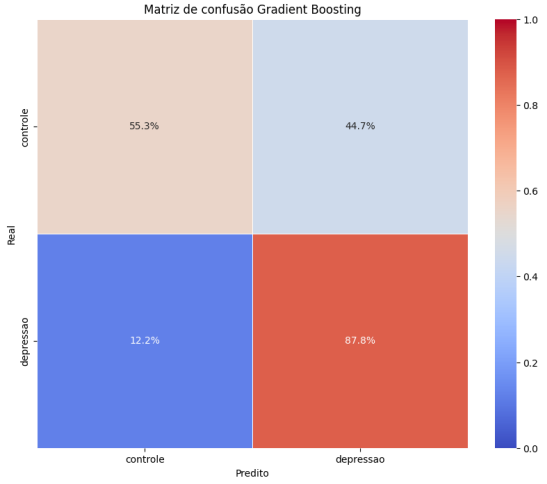
(b) Gradient Boosting - Pré-Pandemia



(c) MLP - Pandemia



(d) Gradient Boosting - Pandemia



4.4.1 Analise dos Resultados

Conclui-se que os modelos Gradient Boosting com f1-score médio de 80.75% e Perceptron Multicamadas com f1-score médio de 80.0% nos experimentos de teste, obtiveram desem-

penho promissores em relação a pesquisadora De Choudhury [14] que obteve desempenho f1-score médio de de 68.0%, de acordo com a literatura. Logo, a incrementação desses novos atributos podem ser úteis para melhoramento de modelos de aprendizagem de máquina no objetivo na detecção de possíveis sinais de padrão de comportamento depressivos ou não depressivos de novos usuários do *Twitter* [2].

4.5 Considerações Finais

Neste capítulo, foram apresentados os resultados de busca pelas palavras-chave de autodeclarações de depressão, análise de frequência de palavras por nuvem de palavras e análise de estatística de desempenho de tarefas de classificação por meio de modelos de aprendizado de máquina. No próximo capítulo, serão discutidas as implicações da mineração de dados para a sociedade, aspectos da ética que devem ser considerados por pesquisadores e empresas, como também conceitos de privacidade e de segurança.

Capítulo 5

Ameaças de Validade

Nesta seção, serão discutidas as limitações e ameaças à validade deste trabalho e como podem surgir indícios que foram parcialmente mitigados durante a realização dessa pesquisa. Foi detalhado os principais riscos que podem ocorrer de exposição de dados pessoais na plataforma *Twitter* e também será explicado quais considerações recomendáveis para o uso de dados de mídias sociais.

5.1 Riscos

No Capítulo 2, apresenta-se como os dados de mídia social são usados para uma ampla gama de estudos sobre a natureza da mente humana. No entanto, os dados de mídia social podem ser usados para outros fins além de saúde mental, podem ser usados também para pesquisas de *marketing*.

Não só pesquisadores de universidades e de instruções estão interessados na aplicação de mineração de dados na área de saúde, como também outras áreas, como por exemplo empresas que podem aproveitar dados de mídia social para seus próprios propósitos, investindo milhões de dinheiro para aprimorar sua capacidade de prever as necessidades futuras de uma pessoa, com base em seus conteúdos de mídia social, dados de dispositivos inteligentes e histórico de atividades, bem como informações mais amplas de técnicas de perfil (por exemplo, foto de perfil, localização, data de nascimento, etc) com o objetivo de personalizar e melhorar seus programas e aplicativos.

Várias preocupações foram levadas em consideração sobre as implicações de empregar mineração de dados em dados de mídias sociais por defensores da privacidade, pesquisadores, formuladores de políticas e pacientes em geral.

No Capítulo 3, discutiu-se a propriedade pública dos dados coletados, o cuidado tomado para não entrar em contato com nenhum usuário da plataforma, além de anonimizar e proteger os dados. A legislação atual, sobre proteção de dados de novas tecnologias e

de princípios éticos, permanece relativamente pouco discutida nas áreas de ciência da computação em todo o Brasil.

Durante a pesquisa, foi aprendido que o pesquisador Conway [83] recomendou uma taxonomia de 10 considerações especificamente relevantes para o uso de dados de mídia social em pesquisa que são:

1. **Privacidade**:

    - Conceito de privacidade: público vs privado; Fluidez no conceito de privacidade; Diferenças geracionais no conceito de privacidade [84, 85]; Efeito panóptico;
    - Confidencialidade: ligação de dados; confidencialidade; direito/desejo de anonimato.
    - Condições médicas estigmatizadas.
    - Política de privacidade do *Twitter*.
    - O *Twitter* é acessível publicamente por padrão.
    - Confiabilidade dos dados pessoais fornecidos pelo usuário.
    - Interpretação de dados descontextualizados do *Twitter* como totalmente representativos de usuários que são de fato multifacetados.
    - Revelação não intencional de informações pessoais.
    - Responsabilidade pessoal dos usuários do *Twitter*.
    - Os usuários do *Twitter* não têm expectativa de privacidade.
    - Identificação do estado de saúde mental dos usuários ou traços de personalidade para: identificar aqueles que necessitam de tratamento; colocação no emprego; marketing direcionado; design da interface do sistema (por exemplo, introvertidos preferem dados apresentados de uma certa maneira); aplicação da lei (por exemplo, identificação de psicopatas).
    - Monitoramento em nível populacional versus diagnóstico individual.
    - Potencial de discriminação com base no estado de saúde obtido das mídias sociais.
    - Perigo de rotular incorretamente um usuário como sofrendo de um problema de saúde específico.
    - Rastreabilidade dos dados do *Twitter*.
    - Grupos de Amigos e Familiares no *Twitter*.

2. **Consentimento Informado**:

- Os usuários do Twitter são participantes de pesquisas inconscientes ou relutantes.
- O consentimento informado é difícil (ou impossível) de obter (ou não é necessário) para o trabalho em grande escala no *Twitter*.

3. **Teoria Ética**:

   - Dificuldades na aplicação de teorias éticas atuais para pesquisas em massa no Twitter.
   - Teorias éticas: deontologia; utilitarismo; feminismo; comunitarismo; aplicação da "*regra de ouro*"; ética ágil/situacional; a teoria da justiça de *Rawls*.

4. **Regularização**:

   - Direitos dos cidadãos de comunicar e compartilhar informações;
   - Crença do pesquisador de que a supervisão regulatória não é necessária ao usar dados do *Twitter*;
   - Discussão de CEP/comitês de ética, em geral;
   - Legislação de proteção de dados;
   - Códigos de conduta profissional;
   - Necessidade de controle regulatório, geralmente;
   - Regulamento de privacidade por país.

5. **Pesquisa Tradicional vs Pesquisa no Twitter**:

   - Apomediação;
   - Escala de pesquisa baseada no *Twitter*;
   - Maior distância entre pesquisador e participantes;
   - *Status* ambíguo dos participantes;
   - Aumento no poder do pesquisador.

6. **Informações Geográficas**:

   - Rastreamento da localização física;
   - Granularidade geográfica apropriada

7. **Pesquisador à Espreita**;

8. **Valor Econômico das Informações Pessoais**;
9. **Excepcionalismo Médico**;
10. **Benefício da Identificação de Condições Médicas Socialmente Nocivas**.

Cada uma dessas considerações visa mitigar um ou mais riscos que a pesquisa possa incorrer. No entanto, é difícil prospectar se a indústria irá aderir a quaisquer considerações feitas exclusivamente pela academia. Nos últimos anos, houve muitos casos de privacidade do usuário sendo violada por grandes e pequenas empresas de tecnologia. Por exemplo, um caso em 2014, em que foi descoberto que entre os 600 aplicativos móveis de saúde mais usados disponíveis para *Android* e *iOS*, apenas 183 (30.5%) tinham políticas de privacidade, dos quais dois terços (66.1%) não faziam menção ao aplicativo em suas políticas [86].

Um outro caso semelhante que ocorreu em 2012 em que a Rede Social *Facebook* recebeu críticas significativas em relação ao seu estudo encoberto de "*contágio emocional*" que teve como objetivo prospectar como alterar o conteúdo da linha do tempo dos usuários, podendo induzir a estados emocionais distintos e isso envolveu aproximadamente 600.000 de seus usuários, dos quais não foi obtido o consentimento da pesquisa e a quem nenhuma informação do estudo foi fornecida e que não puderam se retirar do estudo [87].

Este é um exemplo de como a falta de transparência e dados centralizados, aliados a práticas de pesquisa que não levam em conta quaisquer considerações éticas, podem levar ao abuso da privacidade dos usuários de mídia social e, de fato, causar danos. Na mesma linha, será explorado a seguir como um paradigma tecnológico alternativo tem se mostrado útil em salvaguardar a privacidade e os direitos dos indivíduos.

5.2 Considerações Finais

No próximo capítulo será finalizada essa monografia de conclusão de curso, encerrando-se esse projeto junto às contribuições que esse propósito beneficiou à ciência e às pessoas, assim como também as ideias propostas para trabalhos futuros que podem melhorar esse projeto, alcançando novas descobertas nos dados de usuários nas redes sociais; aperfeiçoando os modelos existentes de algoritmos de aprendizagem de máquina supervisionados e outros métodos que não foram utilizados nesse projeto, mas que seriam muito úteis para criar mais aplicações de identificação de sinais de padrões de comportamento depressivo em usuários das redes sociais.

Capítulo 6

Conclusão

Neste projeto demonstrou-se o potencial de usar a rede social *Twitter* como uma ferramenta para medir e detectar sinais de padrões de comportamento depressivo em postagens e atividades dos usuários no Twitter, tendo como base os trabalhos de De Choudhury et al. [14] e Coppersmith et al. [15]. Usando-se o *framework Snscrape* para filtrar e coletar *tweets* em português de usuários do Twitter (de forma anônima) que se autodeclararam depressivos e aplicando-se métodos e técnicas de Mineração de Dados construíram-se 2 bases de dados de 2 períodos compostos de 10 atributos da De Choudhury et al. [14] (volume de *tweets*, índice de insônia, estilo linguístico de 1ª, 2ª e 3ª pessoa, emoções de valência e ativação, termos depressivos, grafo social e medicamentos antidepressivos) de 5 novos atributos (caracteres orientais, *emojis*, mídia, frequência de links e número de curtidas): o período pré-pandemia (01/01/2018 a 31/12/2019) foram coletados um total de 71.232 usuários (depressivos = 37.985, controle = 33.247) e pandemia (01/01/2020 a 31/12/2021) período pandemia foram coletados um total de 70.370 usuários (depressivos = 36.801, controle = 33.569).

Foi feita uma análise comparativa nas duas bases de dados para verificar se houve aumento ou diminuição dos casos de depressão por consequência da pandemia *COVID-19* no Brasil. Realizamos uma análise exploratória quantitativa dos usuários depressivos nas 2 bases de dados, observamos que no período da pandemia o número de relatos de usuários se autodeclarando depressivo (ex. "eu tenho depressão") é mais que dobro do número de relatos de usuários no período da pré-pandemia. A *COVID-19* pode ter sido a causa de depressão nas pessoas. Não reunimos informações suficientes para comprovar se foi a *COVID-19* ou não.

Na compreensão e exploração dos dados nos *tweets* coletados, através do ponto de vista da pré-pandemia encontramos 116 usuários que não estavam depressivos antes do início da *COVID-19*, mas que se tornaram depressivos na pandemia. Não reunimos informações suficientes para comprovar se foi a *COVID-19* ou não. Por outro lado, do ponto de vista

da pandemia encontramos 66 usuários que já eram depressivos antes da existência da pandemia *COVID-19*, isso pode ter ocorrido de acordo com a literatura sobre psicologia humana, psiquiatria, neurociência e sociolinguística sobre saúde humana que mapearam vários fatores que podem desencadear um quadro depressivo.

Na simulação de teste o modelo *Gradient Boosting (GB)* obteve *f1-score* médio 81.0% e o modelo Perceptron Multicamadas obteve *f1-score* médio de 80.2%. Esses resultados demonstram que a adição de novos aumentou a qualidade da eficiência das predições em comparação com os modelos da pesquisadora De Choudhury et al. [14] que obteve uma média f1-score de 68.0%. Então esses dois modelos são ótimos para serem incorporados numa ferramenta de análise de sentimentos com o intuito de auxiliar as pessoas que necessitam de ajuda na detecção de possíveis sinais de padrões relevantes de comportamento depressivo ou não depressivo.

Entre os passos futuros, podem-se acrescentar novos atributos correlacionados a aspectos sociais, comportamentais e outros tipos de transtornos mentais (ansiedade, trauma pós-parto, suicídio, etc.) de outras redes sociais nas bases de dados. Assim, sugere-se a aplicação de outros algoritmos como por exemplo, aprendizagem de redes neurais *RNN (Recurrent Neural Network)* ou *LSTM (Long short term memory)* ou aprendizagem profunda *Transformers* para aumentar o desempenho da predição de sinais de padrão comportamental, tendo como o resultado suporte para tomada de decisão para governantes e gestores públicos com intuito de criar novas políticas públicas para amparar a população que tem esses transtornos mentais.

6.1 Contribuição

A principal contribuição deste trabalho foi mostrar a possibilidade de aplicar métodos de coleta de dados, extração de características e aplicação de aprendizagem de máquina para identificar sinais de padrões correlacionados à depressão nas mídias sociais no Brasil. Espera-se que esse trabalho possa ajudar outros pesquisadores e estudantes a orientá-los a aplicar a ciência no combate aos problemas de saúde mental, pois a doença mental é um desafio crescente na sociedade.

Atribui-se uma contribuição ao Otto Von Sperling [11,12] por desenvolver a versão inicial do pipeline do projeto para os métodos de coleta de dados, extração de características e modelos de aprendizagem de máquina que foram aperfeiçoados durante o desenvolvimento do projeto. Também deve-se dar créditos aos pesquisadores De Choudhury [14] e Coppersmith [15] por seus conceitos e hipóteses na compreensão dos atributos da rede social Twitter e como aplicá-los nos algoritmos de aprendizagem de máquina.

Outros pesquisadores que pode-se levar em consideração por ajudarem na extraídas de características de conteúdo léxico são os pesquisadores Kristensen et al. [75] por sua base de dados *ANEW-Br* que ajudou a extrair léxicos depressivos e calcular variância e ativação de emoções em textos; aos pesquisadores Selinger e Hartzog [87] por suas hipóteses de como os caracteres orientais nas postagens de textos de usuários na rede social *Twitter* tem efeitos nas suas vidas sociais e aos pesquisadores Araújo et al. [72, 73] por mostrarem que os usuários na rede social Twitter não só expressam seus sentimentos e opiniões com palavras, mas também com *emojis*. Esses pesquisadores foram considerados significativos e de grande ajuda na classificação de sinais de padrões de comportamento depressivos nesse estudo.

Por último, a contribuição mais importante desse projeto é o desenvolvimento de um modelo de aprendizagem de máquina para dar aos usuários uma visão de seus possíveis sinais de comportamento de depressão. Há muito a se ganhar em ajudar as pessoas a buscarem auxílio, individualmente e em conjunto. Com excesso de depressão na sociedade é preciso ter tanto aqueles que têm muita coragem para combater a depressão como também aqueles que buscam a solução. Se esse trabalho puder algum dia ajudar pessoas ou futuros pesquisadores no combate à depressão na saúde mental, ficar-se-á muito gratificado em poder ter contribuído para o avanço do *tratamento depressivo* com base em pesquisas científicas no campo da tecnologia da informação.

6.2 Trabalhos Futuros

Apresentou-se neste trabalho um modelo de aprendizagem de máquina com desempenho e performance consideráveis, mas apesar disso, ainda pode-se melhorar o modelo ou construir um outro modelo melhor, tendo como base os parâmetros utilizados nesta pesquisa, pois o acréscimo futuro não descarta estudos e contribuições consolidados. A primeira sugestão para trabalhos futuros é a de se realizar mais estudos sobre depressão e outras doenças mentais no Brasil, de forma a se melhorar ainda mais o entendimento de como identificar sinais de padrões de *comportamento depressivos* nas pessoas.

Segue-se, que outra maneira interessante de estender esse trabalho é o de se estudar para encontrar outros atributos da rede social Twitter que não foram aplicados nesse trabalho (por exemplo, *hashtags*) os quais possam ajudar a melhorar a performance dos modelos de classificadores de aprendizagem de máquina supervisionados para alcançar resultados mais robustos, com intuito de identificar sinais de padrões de comportamento de depressão. Uma outra ideia seria categorizar os *tweets* depressivos por tipos de depressão ou sintomas de depressão, por exemplo transtorno bipolar, fadiga ou outro tipo de sintoma.

Neste trabalho foram usados dados estatísticos extraídos e calculados das características observadas nas postagens de texto e atividades dos usuários na rede social *Twitter*. Para construção da base de dados, pode-se usar uma outra forma para se construir uma base de dados para se aplicar nos algoritmos de aprendizagem de máquina, utilizando-se da técnica de Processamento de Linguagem Natural *TF-IDF (Term Frequency — Inverse Document Frequency)* que calcula a importância de cada palavra existente em cada documento de frase, assim pode-se ficar sabendo quais palavras são mais relevantes para se correlacionar à depressão, a fim de identificar possíveis sinais de padrões de comportamentos depressivos em usuários na rede social Twitter.

Além dos classificadores de aprendizagem de máquina supervisionados que foram utilizados neste trabalho, pode-se aplicar ainda os métodos de aprendizagem de máquina *RNN (Recurrent Neural Network - em português "redes neurais recorrentes")* como *LSTM (Long short- term memory - em português "memória de curto prazo longa")*, pois eles têm a capacidade de manter as representações de observações muito mais distantes no passado do que outras *RNNs*, podendo ser usado para rastrear os sinais de padrões de comportamento depressivo do usuário, percorrendo-lhe em uma atividade do passado, para que ele aprenda a prevê futuras atividades, porém ao se usar desse método corre-se o risco de se perder informações passadas, prejudicando a eficiência do modelo algoritmo.

Uma última sugestão para se obter mais eficiência e resultado mais preciso no reconhecimento da linguagem de comportamento depressivo, seria o de aplicar o método de aprendizagem de máquina profunda *Transformers* que se utiliza da técnica de Processamento de Linguagem Natural (*NLP*) *TF-IDF*, que ao contrário dos métodos *RNN* que processam palavra por palavra nas frases; no *Transformers* as frases são processadas como um todo, ou seja, as frases são interpretadas calculando-se as pontuações de similaridade entre palavras em uma frase, sem risco de se perder informações passadas necessárias [88]. Uma observação sobre os métodos *Transformers* é a de que no caso de aplicação desse método faz-se necessário, visto que eles são muito complexos no entendimento, um curso para o entendimento sobre os conceitos e as técnicas a serem aplicados.

6.3 Considerações Finais

Diante da explanação dos fatos e da socialização das ideias aqui fundamentadas, chega-se ao final desta monografia. Enfatiza-se que foi uma imensa oportunidade de poder aprender e contribuir por uma causa que excede em muito as complexidades do individualismo. Acredita-se que a doença mental depressão ainda é vista como o problema de uma mente em particular, talvez devido à genética ou a uma educação abusiva, mas contida e vivenciada por indivíduos específicos.

No entanto, foi esclarecido nesta pesquisa que essa ideia não poderia estar mais longe da verdade. A sociedade em todos os seus segmentos carrega o fardo da depressão, embora esse quadro real seja um fenômeno velado pelas demandas da vida contemporânea, a depressão é uma doença "silenciosa"e cercada de tabus, o que dificulta as notificações dos casos pela OMS e agora agravados pelo contexto pós pandemia que reforça a necessidade de alerta para essa ameaça da doença depressão que pode ocorrer em episódios leves, moderados ou graves, o que faz com que se comprometa o bem-estar físico, mental e social das pessoas. Extravia-se quando se deixa de se importar, pois há uma perda irreparável quando alguém decide acabar com a própria vida. Não há estereótipo, nenhum grupo-alvo, não há cor, raça, credo ou gênero favorito. A depressão é um problema de todos nós, e cabe a nós, e somente a nós, resolvê-la.

No futuro, espera-se ver muito mais pesquisas em saúde mental e mineração de dados em nosso país de origem. Lamentavelmente, muitas vezes a tecnologia é tida como incerta ou empregada de forma dissimulada, no entanto inferindo-se pelos significados expostos nesse trabalho, com a utilização da ferramenta de aprendizagem de máquina como fonte de descoberta de casos suspeitos de comportamento depressivo, revelados pelos recursos expressivos ou compositivos do texto na Plataforma *Twitter*, amplia-se a importância da reformulação dessa opinião sobre a necessidade da integração cada vez mais acentuada da ciência com a tecnologia, a qual assume um papel fundamental de suporte à medicina ao possibilitar e facilitar com suas técnicas e processos a produção de resultados confiáveis ao desenvolvimento e aplicações da ciência.

Nesse contexto, a tecnologia torna-se ativa e criativa ao utilizar-se de uma diversidade de operações que propiciam resgatar a importância da pesquisa aliada à tecnologia para desvelar um número expressivo de casos de depressão, como também de outros problemas de saúde mental, principalmente após o cenário da pandemia da COVID-19 no Brasil.

Uma das principais forças motrizes da inovação em modelagem e previsão de comportamento – a maximização do lucro – geralmente não está alinhada com o bem-estar do público em geral. Apesar do desvio, o interesse público vem mudando para um discurso mais aberto (e igualmente difícil) sobre privacidade e direitos individuais. No âmbito dos direitos individuais, defende-se que a saúde mental e os cuidados assistenciais são intrínsecos no fortalecimento da saúde que compreende o bem-estar físico, mental e social das pessoas para se ter uma jornada plena pela vida.

Referências

[1] Reis Filho, José: *Sistema inteligente baseado em árvore de decisão, para apoio ao combate às perdas comerciais na distribuição de energia elétrica.* julho 2006. https://repositorio.ufu.br/handle/123456789/14493, acesso em 2022-03-15, Accepted: 2016-06-22T18:38:46Z Publisher: Universidade Federal de Uberlândia. x, 25, 26

[2] Géron, Aurélien: *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow.* Alta Books, 2ª edição, 2019. x, 2, 3, 22, 30, 31, 32, 34, 74

[3] JustAnotherArchivist: *Snscrape*, março 2022. https://github.com/JustAnotherArchivist/snscrape, acesso em 2022-03-11, original-date: 2018-09-09T20:16:31Z. xii, 11, 12, 13, 14, 15, 16, 47

[4] Carvalho, Luiz Paulo, Jonice Oliveira e Claudia Cappelli: *Pesquisas em Análise de Redes Sociais e LGPD, análises e recomendações.* Em *Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, páginas 73–84. SBC, junho 2020. https://sol.sbc.org.br/index.php/brasnam/article/view/11164, acesso em 2022-04-06, ISSN: 2595-6094. xii, 38

[5] Brasil, Presidência da República Secretaria Geral Subchefia para Assuntos Jurídicos: *Lei Geral de Proteção de Dados Pessoais (LGPD) (LEI Nº 13.709, DE 14 DE AGOSTO DE 2018)*, agosto 2018. http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm, acesso em 2022-04-02. xii, 36, 38

[6] Corbanezi, Elton: *Transtornos Depressivos e Capitalismo Contemporâneo.* Caderno CRH, 31:335–353, agosto 2018, ISSN 0103-4979, 1983-8239. http://www.scielo.br/j/ccrh/a/rkPjhVztHdwQ5Rp4WwcPv7x/?format=html, acesso em 2022-03-07, Publisher: Universidade Federal da Bahia - Faculdade de Filosofia e Ciências Humanas - Centro de Recursos Humanos. 1

[7] Baroni, Daiana Paula Milani, Rômulo Fabiano Silva Vargas e Sandra Noemi Caponi: *Diagnóstico como nome próprio.* Psicologia & Sociedade, 22:70–77, abril 2010, ISSN 1807-0310. http://www.scielo.br/j/psoc/a/HRqmhn6MFr57zsfP78QNQKz/?lang=pt, acesso em 2022-03-08, Publisher: Associação Brasileira de Psicologia Social. 1

[8] Organization, World Health: *Preventing suicide: A global imperative.* World Health Organization, 2014. 1

[9] Saxena, S., M. K. Funk e D. Chisholm: *Comprehensive mental health action plan 2013–2020*. EMHJ-Eastern Mediterranean Health Journal, 21(7):461–463, 2015. https://www.who.int/publications-detail-redirect/9789241506021, Publisher: World Health Organization, Regional Office for the Eastern Mediterranean. 1

[10] Hawn, Carleen: *Take Two Aspirin And Tweet Me In The Morning: How Twitter, Facebook, And Other Social Media Are Reshaping Health Care*. Health Affairs, 28(2):361–368, março 2009, ISSN 0278-2715. https://www.healthaffairs.org/doi/abs/10.1377/hlthaff.28.2.361, acesso em 2022-03-08, Publisher: Health Affairs. 1, 41

[11] Sperling, Otto von e Marcelo Ladeira: *Mining Twitter Data for Signs of Depression in Brazil*. Em *Anais do Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, páginas 25–32. SBC, outubro 2019. https://sol.sbc.org.br/index.php/kdmile/article/view/8785, acesso em 2022-03-08, ISSN: 2763-8944. 1, 7, 42, 43, 80

[12] Sperling, Otto von: *UnB Sense : a web application to probe for signs of depression from user profiles on social media*. dezembro 2019. https://bdm.unb.br/handle/10483/26502, acesso em 2022-03-07, Accepted: 2021-02-01T16:59:26Z. 1, 7, 42, 43, 80

[13] Dos Santos, Fernando Leandro e Marcelo Ladeira: *The Role of Text Pre-processing in Opinion Mining on a Social Media Language Dataset*. Em *2014 Brazilian Conference on Intelligent Systems*, páginas 50–54, outubro 2014. 1, 3

[14] De Choudhury, Munmun, Michael Gamon, Scott Counts e Eric Horvitz: *Predicting Depression via Social Media*. Em Kiciman, Emre, Nicole B. Ellison, Bernie Hogan, Paul Resnick e Ian Soboroff (editores): *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013*. The AAAI Press, 2013. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6124, acesso em 2022-03-07. 1, 2, 3, 6, 7, 42, 43, 50, 70, 71, 72, 74, 79, 80

[15] Coppersmith, Glen, Mark Dredze e Craig Harman: *Quantifying Mental Health Signals in Twitter*. Em *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, páginas 51–60, Baltimore, Maryland, USA, junho 2014. Association for Computational Linguistics. https://aclanthology.org/W14-3207, acesso em 2022-03-08. 1, 2, 3, 6, 7, 42, 43, 79, 80

[16] Ghahramani, Zoubin: *An introduction to hidden markov models and bayesian networks*. International Journal of Pattern Recognition and Artificial Intelligence, 15(01):9–42, fevereiro 2001, ISSN 0218-0014. https://www.worldscientific.com/doi/abs/10.1142/S0218001401000836, acesso em 2022-03-08, Publisher: World Scientific Publishing Co. 2, 6

[17] Reagan, Andrew J., Christopher M. Danforth, Brian Tivnan, Jake Ryland Williams e Peter Sheridan Dodds: *Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs*. EPJ Data Science, 6(1):1–21, dezembro 2017, ISSN 2193-1127. https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-017-0121-9, acesso em 2022-03-08, Number: 1 Publisher: SpringerOpen. 2, 6

[18] Reece, Andrew G., Andrew J. Reagan, Katharina L. M. Lix, Peter Sheridan Dodds, Christopher M. Danforth e Ellen J. Langer: *Forecasting the onset and course of mental illness with Twitter data*. arXiv:1608.07740 [physics], agosto 2016. http://arxiv.org/abs/1608.07740, acesso em 2022-03-08, arXiv: 1608.07740. 2

[19] Nakov, Preslav, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani e Veselin Stoyanov: *SemEval-2016 Task 4: Sentiment Analysis in Twitter*. Em *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, páginas 1–18, San Diego, California, junho 2016. Association for Computational Linguistics. https://aclanthology.org/S16-1001, acesso em 2022-03-08. 3

[20] Rude, Stephanie, Eva Maria Gortner e James Pennebaker: *Language use of depressed and depression-vulnerable college students*. Cognition and Emotion, 18(8):1121–1133, 2004, ISSN 0269-9931. https://doi.org/10.1080/02699930441000030, acesso em 2022-03-08, Publisher: Routledge _eprint: https://doi.org/10.1080/02699930441000030. 6

[21] Williams, Keith L. e Renee V. Galliher: *Predicting Depression and Self–Esteem from Social Connectedness, Support, and Competence*. Journal of Social and Clinical Psychology, 25(8):855–874, outubro 2006, ISSN 0736-7236. https://guilfordjournals.com/doi/abs/10.1521/jscp.2006.25.8.855, acesso em 2022-03-08, Publisher: Guilford Publications Inc. 6

[22] Bollen, Johan, Alberto Pepe e Huina Mao: *Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena*. arXiv:0911.1583 [cs], 2011. http://arxiv.org/abs/0911.1583, acesso em 2022-03-08, arXiv: 0911.1583. 6

[23] Moreno, Megan A, Lauren A Jelenchick, Katie G Egan, Elizabeth Cox, Henry Young, Kerry E Gannon e Tara Becker: *Feeling Bad on Facebook: Depression disclosures by college students on a Social Networking Site*. Depression and anxiety, 28(6):447–455, junho 2011, ISSN 1091-4269. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3110617/, acesso em 2022-03-08. 6

[24] Park, Minsu, Chiyoung Cha e Meeyoung Cha: *Depressive moods of users portrayed in Twitter*. Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining, SIGKDD 2012, páginas 1–8, 2012. 6

[25] Resnik, Philip, William Armstrong, Leonardo Claudino e Thang Nguyen: *The University of Maryland CLPsych 2015 Shared Task System*. Em *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, páginas 54–60, Denver, Colorado, junho 2015. Association

for Computational Linguistics. https://aclanthology.org/W15-1207, acesso em 2022-03-08. 6

[26] Pennebaker, James W., Matthias R. Mehl e Kate G. Niederhoffer: *Psychological Aspects of Natural Language Use: Our Words, Our Selves.* Annual Review of Psychology, 54(1):547–577, 2003. https://doi.org/10.1146/annurev.psych.54.101601.145041, acesso em 2022-03-08, _eprint: https://doi.org/10.1146/annurev.psych.54.101601.145041. 6

[27] Bradley, Margaret M. e Peter J. Lang: *Affective norms for English words (ANEW): Instruction manual and affective ratings.* Relatório Técnico, Technical report C-1, the center for research in psychophysiology ..., 1999. 6, 42, 50

[28] CCM: *Ciência de dados: o que é, como funciona e qual importância - Blog da CCM - O melhor conteúdo para profissionais de TI.* https://blog.ccmtecnologia.com.br/post/ciencia-de-dados-o-que-e-como-funciona-qual-importancia, acesso em 2022-03-18. 16

[29] Saltz, Jeffrey S. e Jeffrey M. Stanton: *An Introduction to Data Science.* SAGE Publications, agosto 2017, ISBN 978-1-5063-7754-4. Google-Books-ID: dLsyDwAAQBAJ. 17

[30] Rautenberg, Sandro e Paulo Ricardo Viviurka do Carmo: *Big data e ciência de dados: complementariedade conceitual no processo de tomada de decisão.* Brazilian Journal of Information Science: research trends, 13(1):56–67, março 2019. https://revistas.marilia.unesp.br/index.php/bjis/article/view/8315. 17

[31] Han, Jiawei, Jian Pei e Micheline Kamber: *Data Mining: Concepts and Techniques.* Elsevier, junho 2011, ISBN 978-0-12-381480-7. Google-Books-ID: pQws07tdpjoC. 18

[32] O'Neil, Cathy e Rachel Schutt: *Doing Data Science: Straight Talk from the Frontline.* "O'Reilly Media, Inc.", outubro 2013, ISBN 978-1-4493-6389-5. Google-Books-ID: vcVKAQAAQBAJ. 19

[33] Kohavi, Ron e Foster Provost: *Glossary of Terms.* Machine Learning, 30(2):271–274, fevereiro 1998, ISSN 1573-0565. https://doi.org/10.1023/A:1017181826899, acesso em 2022-03-16. 21

[34] Bishop, Christopher M e Nasser M Nasrabadi: *Pattern recognition and machine learning*, volume 4. Springer, 2006, ISBN 1613-9011. https://link.springer.com/book/9780387310732. 21

[35] James, Gareth, Daniela Witten, Trevor Hastie e Robert Tibshirani: *An Introduction to Statistical Learning.* Springer Texts in Statistics. Springer, New York, NY, 2013, ISBN 978-1-4614-7138-7. https://doi.org/10.1007/978-1-4614-7138-7, acesso em 2022-03-16. 22

[36] Ripley, Brian D.: *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, 1996, ISBN 978-0-521-71770-0. https://www.cambridge.org/core/books/pattern-recognition-and-neural-networks/4E038249C9BAA06C8F4EE6F044D09C5C, acesso em 2022-03-16. 22

[37] Mitchell, Thomas M.: *Machine Learning*. McGraw-Hill, Inc., USA, 1ª edição, 1997, ISBN 978-0-07-042807-2. 22

[38] Samuel, A. L.: *Some Studies in Machine Learning Using the Game of Checkers*. IBM Journal of Research and Development, 3(3):210–229, julho 1959, ISSN 0018-8646. Conference Name: IBM Journal of Research and Development. 22

[39] Haenlein, Michael e Andreas Kaplan: *A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence*. California Management Review, 61(4):5–14, 2019, ISSN 0008-1256. `https://doi.org/10.1177/0008125619864925`, acesso em 2022-03-08, Publisher: SAGE Publications Inc. 23

[40] Kaplan, Andreas e Michael Haenlein: *Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence*. Business Horizons, 62(1):15–25, janeiro 2019, ISSN 0007-6813. `https://www.sciencedirect.com/science/article/pii/S0007681318301393`, acesso em 2022-03-08. 23

[41] Haenlein, Michael, Ertan Anadol, Tyler Farnsworth, Harry Hugo, Jess Hunichen e Diana Welte: *Navigating the New Era of Influencer Marketing: How to be Successful on Instagram, TikTok, & Co*. California Management Review, 63(1):5–25, novembro 2020, ISSN 0008-1256. `https://doi.org/10.1177/0008125620958166`, acesso em 2022-03-08, Publisher: SAGE Publications Inc. 23

[42] Haenlein, Michael, Ming Hui Huang e Andreas Kaplan: *Guest Editorial: Business Ethics in the Era of Artificial Intelligence*. Journal of Business Ethics, fevereiro 2022, ISSN 1573-0697. `https://doi.org/10.1007/s10551-022-05060-x`, acesso em 2022-03-08. 23

[43] Russell, Stuart J. e Peter Norvig: *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson, 2020, ISBN 978-0-13-461099-3. `http://aima.cs.berkeley.edu/`, acesso em 2022-03-08. 23, 24

[44] Sutton, Richard S e Andrew G Barto: *Reinforcement learning: An introduction*. MIT press, 2018. 24

[45] Kolesnikov, Alexander, Xiaohua Zhai e Lucas Beyer: *Revisiting Self-Supervised Visual Representation Learning*. arXiv:1901.09005 [cs], janeiro 2019. `http://arxiv.org/abs/1901.09005`, acesso em 2022-03-08, arXiv: 1901.09005. 24

[46] Varella, CARLOS ALBERTO ALVES: *Análise multivariada aplicada as ciências agrárias: análise de componentes principais*. Universidade Federal Rural do Rio de Janeiro–UFRRJ. Seropédica–RJ, 2008. 24, 25

[47] Khatree, Ravinda e Dayanand N. Naik: *Applied Multivariate Statistics with SAS Software*. SAS Publishing, 1997, ISBN 978-1-55544-239-2. 24

[48] Lantz, Brett: *Machine Learning with R*. Packt Publishing, 2013, ISBN 978-1-78216-214-8. 25, 27, 32

[49] Zekic-Susac, M., N. Sarlija e M. Bensic: *Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models*. Em *26th International Conference on Information Technology Interfaces, 2004.*, páginas 265–270 Vol.1, junho 2004. 25

[50] Gama, João, Ricardo Rocha e Pedro Medas: *Accurate decision trees for mining high-speed data streams*. Em *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, páginas 523–528, New York, NY, USA, 2003. Association for Computing Machinery, ISBN 978-1-58113-737-8. https://doi.org/10.1145/956750.956813, acesso em 2022-03-15. 25

[51] Silva, Gleidson Leite da: *Desenvolvimento de rede neural de multicamadas para predição de parâmetros de soldagem*. março 2018. http://repositorio.ufersa.edu.br/handle/prefix/2913, acesso em 2022-03-09, Accepted: 2019-12-05T12:13:12Z Publisher: Universidade Federal Rural do Semi-Árido. 26, 27, 29

[52] Onoda, M: *Estudo sobre um algoritmo de árvores de decisão acoplado a um sistema de banco de dados relacional. 2001. 110p.* PhD Thesis, Dissertação (Mestrado)-Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2001. 26

[53] Tsai, Chih Fong, Yu Feng Hsu e David C. Yen: *A comparative study of classifier ensembles for bankruptcy prediction*. Applied Soft Computing, 24:977–984, novembro 2014, ISSN 1568-4946. https://www.sciencedirect.com/science/article/pii/S1568494614004128, acesso em 2022-03-15. 26, 27, 33

[54] Breiman, Leo: *Random Forests*. Machine Learning, 45(1):5–32, outubro 2001, ISSN 1573-0565. https://doi.org/10.1023/A:1010933404324, acesso em 2022-03-08. 27, 28, 32

[55] Oshiro, Thais Mayumi: *Uma abordagem para a construção de uma única árvore a partir de uma Random Forest para classificação de bases de expressão gênica*. text, Universidade de São Paulo, agosto 2013. http://www.teses.usp.br/teses/disponiveis/95/95131/tde-15102013-183234/, acesso em 2022-03-15. 27, 32

[56] Maria Oliveira Da Silva, Luiza: *Uma Aplicação de Árvores de Decisão, Redes Neurais e KNN para a Identificação de Modelos ARMA Não-Sazonais e Sazonais*. Doutor em Ciências em Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brazil, setembro 2005. http://www.maxwell.vrac.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=7587@1, acesso em 2022-03-08. 28

[57] Braga, Antonio de Pádua, Teresa Bernarda Ludermir e André Carlos Ponce de Leon Ferreira Carvalho: *Redes neurais artificiais: teoria e aplicações*. 2007. https://repositorio.usp.br/single.php?_id=001618274, acesso em 2022-03-09. 28

[58] Calôba, Guilherme Marques, Luiz Pereira Calôba e Eduardo Saliby: *Cooperação entre redes neurais artificiais e técnicas 'clássicas' para previsão de demanda de uma série de vendas de cerveja na Austrália*. Pesquisa Operacional, 22:345–358, julho 2002, ISSN 0101-7438, 1678-5142. http://www.scielo.br/j/pope/a/GvmZQVNPKXxtMBBpkGsSjyH/?lang=pt, acesso em 2022-03-09, Publisher: Sociedade Brasileira de Pesquisa Operacional. 29

[59] Cristianini, Nello e John Shawe-Taylor: *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, 2000, ISBN 978-0-521-78019-3. https://www.cambridge.org/core/books/an-introduction-to-support-vector-machines-and-other-kernelbased-learning-methods/A6A6F4084056A4B23F88648DDBFDD6FC, acesso em 2022-03-08. 29, 30

[60] Cortes, Corinna e Vladimir Vapnik: *Support-vector networks*. Machine Learning, 20(3):273–297, setembro 1995, ISSN 1573-0565. https://doi.org/10.1007/BF00994018, acesso em 2022-03-08. 29

[61] Zhang, Harry: *The optimality of naive Bayes*. Aa, 1(2):3, 2004. 30

[62] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot e Édouard Duchesnay: *Scikit-learn: Machine Learning in Python*. arXiv:1201.0490 [cs], junho 2018. http://arxiv.org/abs/1201.0490, acesso em 2022-03-09, arXiv: 1201.0490. 30

[63] Wang, Gang, Jinxing Hao, Jian Ma e Hongbing Jiang: *A comparative assessment of ensemble learning for credit scoring*. Expert Systems with Applications, 38(1):223–230, janeiro 2011, ISSN 0957-4174. https://www.sciencedirect.com/science/article/pii/S095741741000552X, acesso em 2022-03-15. 32

[64] Acuna, Edgar: *Bagging Classifiers Based on Kernel Density Estimators*. https://www.academia.edu/21150903/Bagging_Classifiers_Based_on_Kernel_Density_Estimators, acesso em 2022-03-15. 32

[65] Freund, Yoav e Robert E. Schapire: *Experiments with a new boosting algorithm*. Em *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ICML'96, páginas 148–156, San Francisco, CA, USA, julho 1996. Morgan Kaufmann Publishers Inc., ISBN 978-1-55860-419-3. 32, 33

[66] Lopes, Lucelene e Caroline Riella: *Aprendizagem de Máquina através de Combinação de Classificadores em Bases de Dados da Área da Saúde*. 32

[67] Schapire, Robert E.: *The strength of weak learnability*. Machine Learning, 5(2):197–227, junho 1990, ISSN 1573-0565. https://doi.org/10.1007/BF00116037, acesso em 2022-03-15. 32

[68] Chaves, Bruno Butilhão: *Estudo do algoritmo AdaBoost de aprendizagem de máquina aplicado a sensores e sistemas embarcados*. text, Universidade de São Paulo, dezembro 2011. http://www.teses.usp.br/teses/disponiveis/3/3152/tde-12062012-163740/, acesso em 2022-03-15. 33

[69] De Castro, Leandro e Daniel Ferrari: *Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações*. maio 2016, ISBN 978-85-472-0098-5. 34

[70] Abdi, Hervé e Lynne J. Williams: *Principal component analysis*. WIREs Computational Statistics, 2(4):433–459, 2010, ISSN 1939-0068. https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.101, acesso em 2022-03-20, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.101. 36

[71] Jaadi, Zakaria: *A step-by-step explanation of Principal Component Analysis (PCA)*. Retrieved June, 7:2021, 2021. 36

[72] Araújo, Matheus, Pollyanna Gonçalves, Fabrício Benevenuto e M. Cha: *Métodos para análise de sentimentos no twitter*. Em *Proceedings of the 19th Brazilian symposium on Multimedia and the Web (WebMedia'13)*, página 19. sn, 2013. 41, 42, 50, 81

[73] Araújo, Matheus, Fabrício Benevenuto e Filipe Ribeiro: *Métodos para análise de sentimentos em mídias sociais*. Sociedade Brasileira de Computação, 2015. 41, 42, 50, 81

[74] Schröer, Christoph, Felix Kruse e Jorge Marx Gómez: *A Systematic Literature Review on Applying CRISP-DM Process Model*. Procedia Computer Science, 181:526–534, janeiro 2021, ISSN 1877-0509. https://www.sciencedirect.com/science/article/pii/S1877050921002416, acesso em 2022-09-30. 41

[75] Kristensen, Christian Haag, Carlos Falcão de Azevedo Gomes, Alice Reuwsaat Justo e Karin Vieira: *Normas brasileiras para o Affective Norms for English Words*. Trends in Psychiatry and Psychotherapy, 33:135–146, 2011, ISSN 2237-6089, 2238-0019. http://www.scielo.br/j/trends/a/dHp9B8gqyKfpMTdygF7cT7b/?lang=pt, acesso em 2022-03-08, Publisher: Associação de Psiquiatria do Rio Grande do Sul. 42, 50, 81

[76] Ramirez-Esparza, Nairan, Cindy Chung, Ewa Kacewic e James Pennebaker: *The Psychology of Word Use in Depression Forums in English and in Spanish: Testing Two Text Analytic Approaches*. Proceedings of the International AAAI Conference on Web and Social Media, 2(1):102–108, 2008, ISSN 2334-0770. https://ojs.aaai.org/index.php/ICWSM/article/view/18623, acesso em 2022-03-08, Number: 1. 42

[77] Preoţiuc-Pietro, Daniel, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H. Andrew Schwartz e Lyle Ungar: *The role of personality, age, and gender in tweeting about mental illness*. Em *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, páginas 21–30, Denver, Colorado, junho 2015. Association for Computational Linguistics. https://aclanthology.org/W15-1203, acesso em 2022-03-08. 42, 50, 51

[78] Ghosh, Aniruddha, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden e Antonio Reyes: *SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter*. Em *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, páginas 470–478, Denver, Colorado, junho 2015. Association for Computational Linguistics. https://aclanthology.org/S15-2080, acesso em 2022-03-08. 43

[79] Liu, Yifan, Zengchang Qin, Pengyu Li e Tao Wan: *Stock Volatility Prediction Using Recurrent Neural Networks with Sentiment Analysis*. Em Benferhat, Salem, Karim Tabia e Moonis Ali (editores): *Advances in Artificial Intelligence: From Theory to Practice*, Lecture Notes in Computer Science, páginas 192–201, Cham, 2017. Springer International Publishing, ISBN 978-3-319-60042-0. 43

[80] Roesslein, Joshua: *Tweepy: Twitter for Python!*, março 2022. `https://github.com/tweepy/tweepy`, acesso em 2022-03-12, original-date: 2009-07-06T04:15:34Z. 47

[81] Carvalho, Flavio, Rafael Guimarães Rodrigues, Lilian Ferrari e Gustavo Paiva Guedes: *A dictionary of pronouns for Brazilian Portuguese*. Em *Congresso Internacional de Informática Educativa (TISE 2018)*, Brasília, Brazil, novembro 2018. 49, 50

[82] Rude, Stephanie S., Carmen R. Valdez, Susan Odom e Arshia Ebrahimi: *Negative Cognitive Biases Predict Subsequent Depression*. Cognitive Therapy and Research, 27(4):415–429, agosto 2003, ISSN 1573-2819. `https://doi.org/10.1023/A:1025472413805`, acesso em 2022-04-15. 49, 50

[83] Conway, Mike: *Ethical Issues in Using Twitter for Public Health Surveillance and Research: Developing a Taxonomy of Ethical Concepts From the Research Literature*. Journal of Medical Internet Research, 16(12):e3617, dezembro 2014. `https://www.jmir.org/2014/12/e290`, acesso em 2022-03-14, Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. 76

[84] McKee, Rebecca: *Ethical issues in using social media for health and health care research*. Health Policy, 110(2):298–301, maio 2013, ISSN 0168-8510. `https://www.sciencedirect.com/science/article/pii/S0168851013000468`, acesso em 2022-07-07. 76

[85] O'Connor, Dan: *The Apomediated World: Regulating Research When Social Media Has Changed Research*. Journal of Law, Medicine & Ethics, 41(2):470–483, 2013, ISSN 1073-1105, 1748-720X. `https://www.cambridge.org/core/journals/journal-of-law-medicine-and-ethics/article/abs/apomediated-world-regulating-research-when-social-media-has-changed-research/6C76FC3156F6BBC67315F0AAB77DA8CA`, acesso em 2022-07-07, Publisher: Cambridge University Press. 76

[86] Sunyaev, Ali, Tobias Dehling, Patrick L Taylor e Kenneth D Mandl: *Availability and quality of mobile health app privacy policies*. Journal of the American Medical Informatics Association, 22(e1):e28–e33, abril 2015, ISSN 1067-5027. `https://doi.org/10.1136/amiajnl-2013-002605`, acesso em 2022-03-14. 78

[87] Selinger, Evan e Woodrow Hartzog: *Facebook's emotional contagion study and the ethical problem of co-opted identity in mediated environments where users lack control*. Research Ethics, 12(1):35–43, janeiro 2016, ISSN 1747-0161. `https://doi.org/`

10.1177/1747016115579531, acesso em 2022-03-14, Publisher: SAGE Publications Ltd. 78, 81

[88] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser e Illia Polosukhin: *Attention Is All You Need*. junho 2017. https://arxiv.org/abs/1706.03762v5, acesso em 2022-03-30. 82

Apêndice A

Matrizes de Correlação

Figura A.1: Matriz de Correção da Base de Dados Período Pré-Pandemia (Completa)

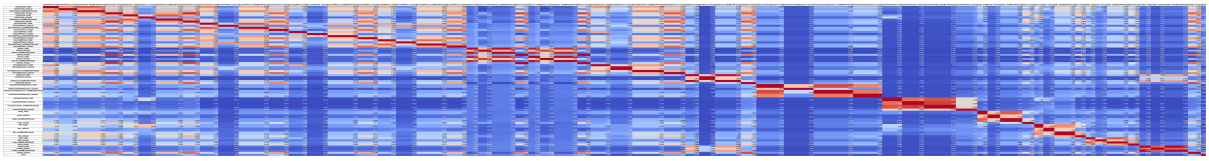
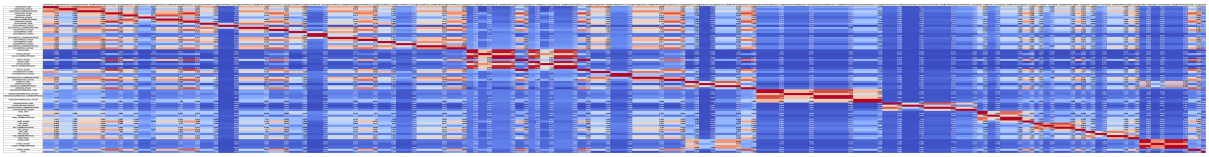


Figura A.2: Matriz de Correção da Base de Dados Período Pandemia (Completa)



Apêndice B

Modelagem da Metodologia

Figura B.1: Metodologia Modelagem - Parte 1

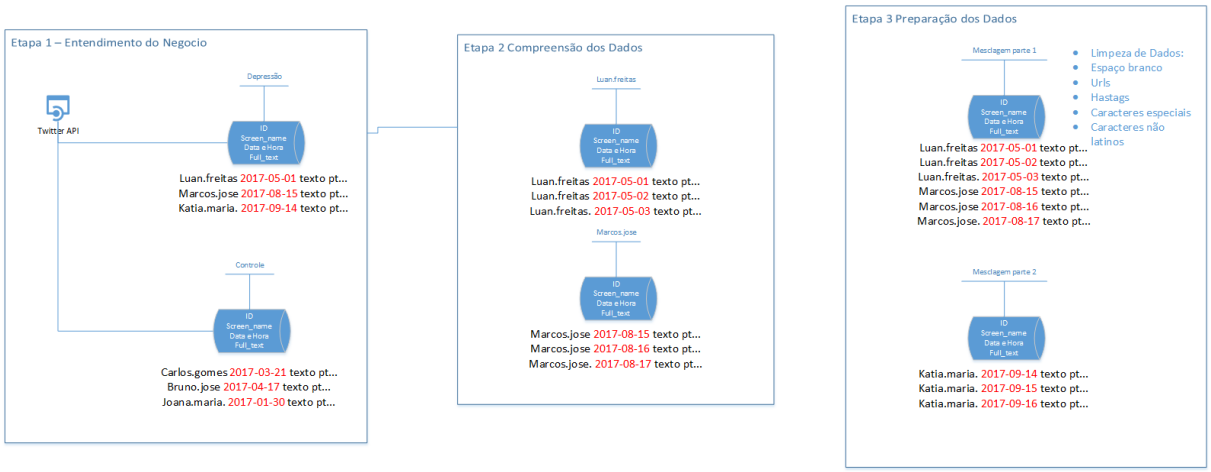


Figura B.2: Metodologia Modelagem - Parte 2

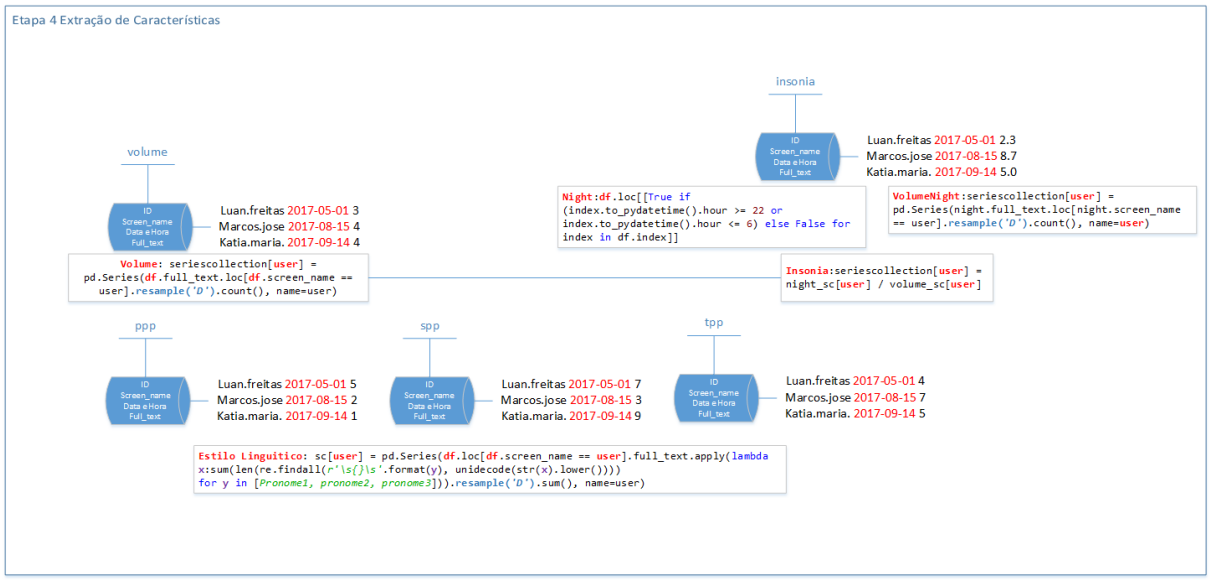


Figura B.3: Metodologia Modelagem - Parte 3

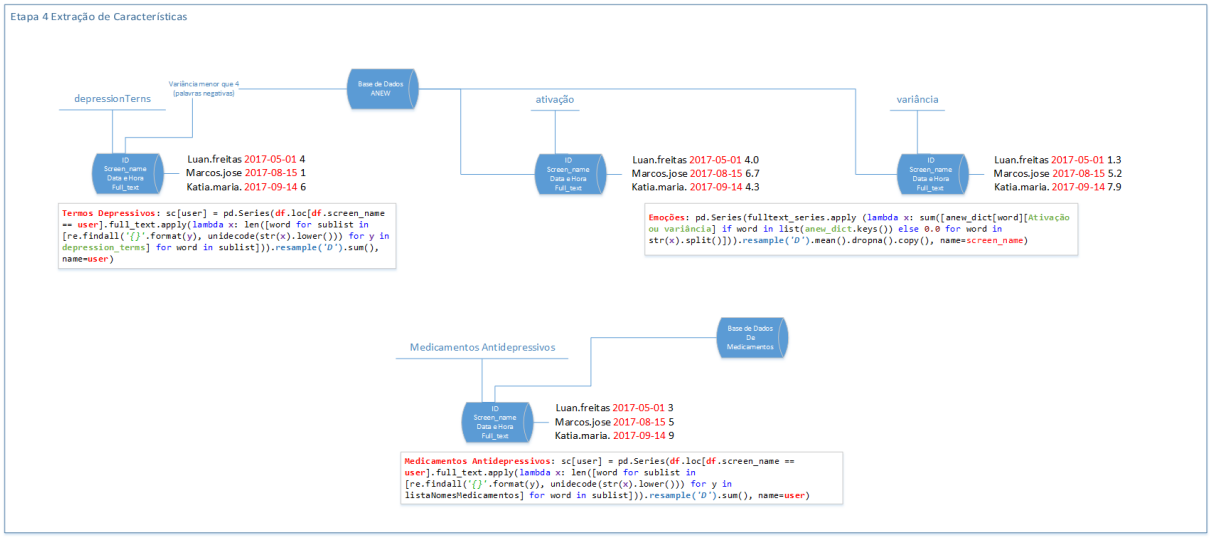


Figura B.4: Metodologia Modelagem - Parte 4

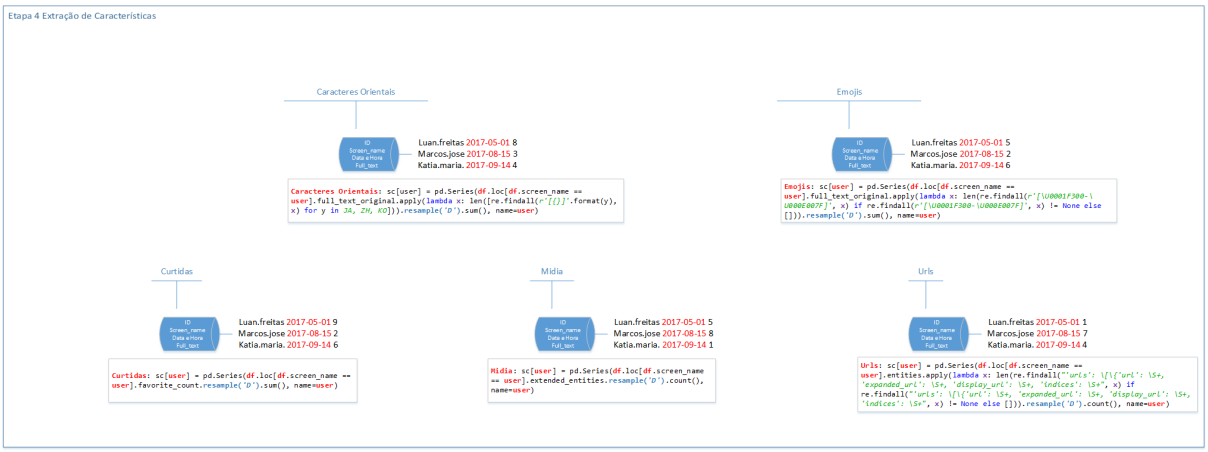


Figura B.5: Metodologia Modelagem - Parte 5

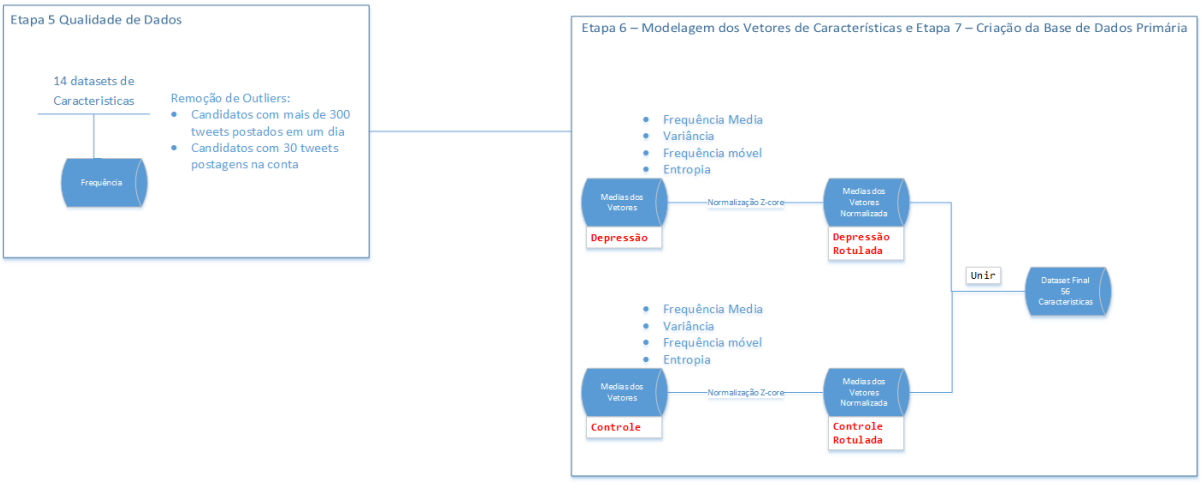


Figura B.6: Metodologia Modelagem - Parte 6

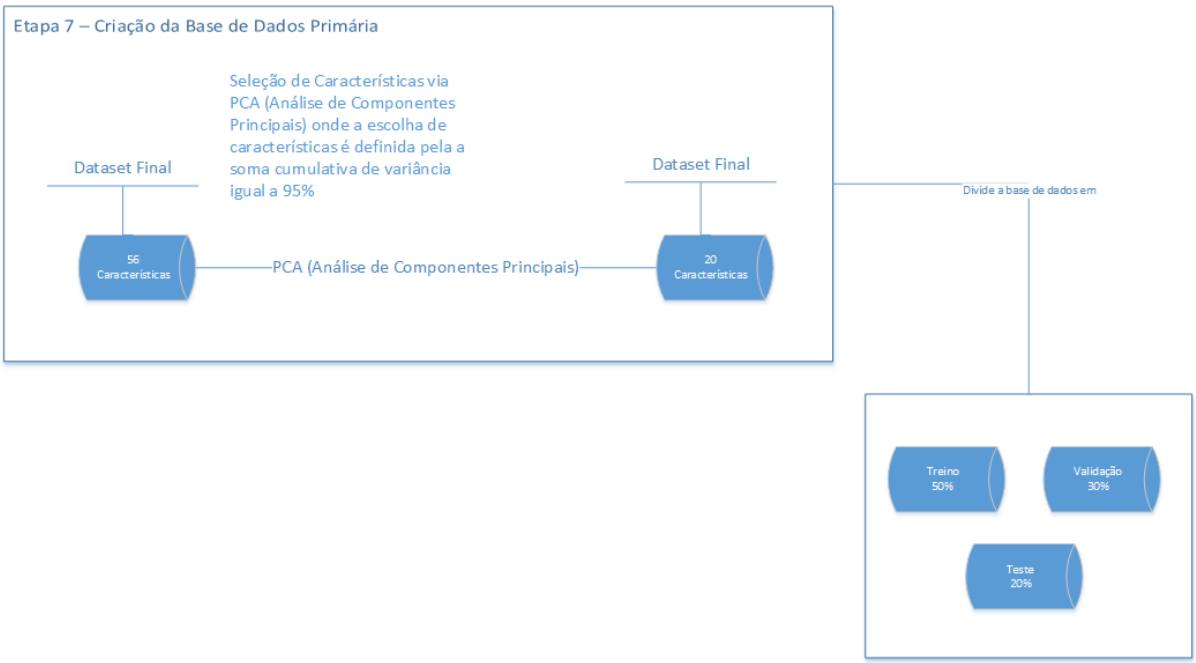


Figura B.7: Metodologia Modelagem - Parte 7

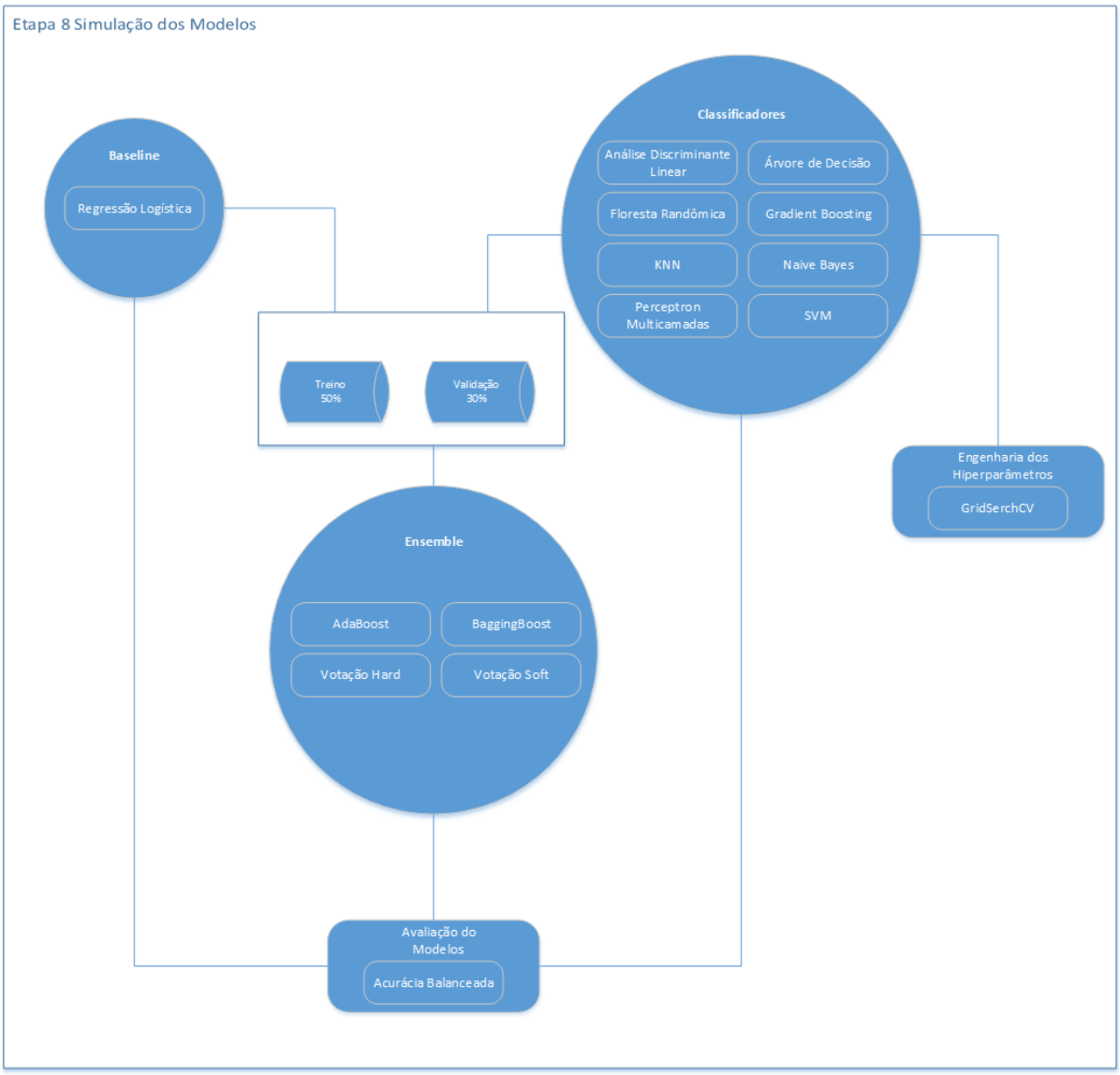


Figura B.8: Metodologia Modelagem - Parte 8