Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

DETECÇÃO DE FAKE NEWS USANDO OS ALGORITMOS DECISION TREE, SUPPORT VECTOR MACHINE E K-NEAREST NEIGHBORS

DETECTION OF FAKE NEWS USING THE DECISION TREE, SUPPORT VECTOR MACHINE AND K-NEAREST NEIGHBORS ALGORITHMS

DETECCIÓN DE FAKE NEWS UTILIZANDO LOS ALGORITMOS DECISION TREE, SUPPORT VECTOR MACHINE Y K-NEAREST NEIGHBORS

Lucas Monteiro Bastos [1]
Centro Universitário UNDB, São Luís, MA, Brasil.

Rodrigo Monteiro de Lima[2]
Centro Universitário UNDB, São Luís, MA, Brasil.

RESUMO

Fake News é uma informação falsa, normalmente divulgada nas redes sociais e devido ao seu poder de disseminação na internet pode gerar problemas significativos na sociedade. Dessa forma, é necessário que se desenvolvam alternativas que busquem amenizar tais efeitos, como a detecção de Fake News aplicando técnicas e algoritmos de Machine Learning. O objetivo desse trabalho é analisar o uso e eficiência de algoritmos de ML na detecção de notícias falsas. A pesquisa é da modalidade aplicada, visando um estudo descritivo e quantitativo. Os dados foram coletados em conjunto na plataforma Kaggle e a extração de dados utilizando a linguagem Python, o processamento foi realizado pela plataforma Jupyter Notebook.

Palavras-chave: Fake News; Machine Learning; Python; Algoritmos.

ABSTRACT

Fake News is false information, normally disseminated on social networks and due to its power of dissemination on the internet it can gene- rate significant

[1] Aluno de Sistemas de Informação do Centro Universitário UNDB. E-mail: monteiro.bastos1994@gmail.com

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

problems in society. Thus, it is necessary to develop alternatives that seek to mitigate such effects, such as the detection of Fake News through the application of Machine Learning techniques and algorithms. The objective of this work is to analyze the use and efficiency of ML algorithms in detecting fake news. the research is of the applied modality, aiming at a descriptive and quantitative study. The data were collected together in the Kaggle platform and the data extraction using the Python language, the processing was performed by the Jupyter Notebook platform.

Palavras-chave: Fake News; Machine Learning; Python; Algoritmos.

RESUMEN

Las Fake News son informaciones falsas, normalmente difundidas en las redes sociales y que por su poder de difusión en internet pueden generar importantes problemas en la sociedad. Por ello, es necesario desarrollar alternativas que busquen mitigar dichos efectos, como la detección de Fake News mediante la aplicación de técnicas y algoritmos de Machine Learning. El objetivo de este trabajo es analizar el uso y la eficiencia de los algoritmos de ML en la detección de noticias falsas. La investigación es de la modalidad aplicada, visando un estudio descriptivo y cuantitativo. Los datos fueron recolectados en conjunto en la plataforma Kaggle y la extracción de datos utilizando el lenguaje Python, el procesamiento fue realizado por la plataforma Jupyter Notebook.

Palavras-chave: Noticias Falsas; Aprendizaje automático; Python; Algoritmos.

1 INTRODUÇÃO

Os malefícios que uma notícia pode trazer no tocante aos riscos para a sociedade em geral. A divulgação de uma informação adulterada, habitualmente chamada de Fake News, pode interferir negativamente em vários aspectos da sociedade, tanto no âmbito político, da segurança e da saúde. Por esse motivo serão apresentadas algumas práticas que contribuirão para saber a veracidade das informações compartilhadas. Qualquer tipo de informação falsa, dá mais simples até a mais descabida tem um peso, visto que essas informações manipulam as pessoas ao erro, devido ao fato de que a notícia pode conter uma informação falsa rodeada de outras verídicas [DELMAZO e VALENTE, 2020].

A partir de uma experiência acadêmica onde se foi proposta a elaboração de um projeto prático da disciplina de Programação Web para fazer a detecção de Fake News por meio de um algoritmo de Machine Learning, houve interesse em conhecer mais profundamente a tecnologia referida, bem como analisar e

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

aplicá-la em uma pesquisa pessoal, como em um trabalho de conclusão de curso.

Sendo assim, como relevância social e profissional, infere-se que a busca de possíveis soluções para apurar os fatos e exposição do que de fato é real e o que é uma notícia falsa possui o intuito de auxiliar a sociedade como um todo, visto que uma notícia com alta relevância seja ela no aspecto econômico, político ou social pode causar impactos consideráveis ou irreversíveis a um determinado grupo social.

O trabalho realizado é de modalidade aplicada e modelo teórico descritivo e quantitativo, com dados coletados na plataforma Kaggle, posteriormente extraídos com a utilização de algoritmos de Machine Learning e analisados por meio da linguagem Python. Os dados extraídos foram processados e analisados por meio da plataforma interativa Jupyter Notebook

A importância dessa aplicação pode trazer benefícios notavelmente. Como consequência disso verificou-se que a detecção correta dessas informações se tornou um problema complexo e desafiador por diversos motivos. A disseminação de Fake News na maior parte do tempo é espalhada acidentalmente com objetivo de compartilhar tal informação na intenção de alertar os familiares ou de desconstruir um

argumento proposto por alguém [CARVALHO e MATEUS, 2018].

2 REVISÃO DE LITERATURA

Nesta seção serão evidenciados os embasamentos teóricos a respeito da pesquisa. Serão tratados os conceitos e abordagens de Fake News, bibliotecas para análise de dados Pandas, NumPy, Matplotlib, Seaborn e Scikit-learn, linguagem de programação Python e Machine Learning, além de elencar técnicas de algoritmos como Decision Tree, Support Vector Machine e K-Nearest Neighbors.

2.1 Tipos de Machine Learning

Os tipos de algoritmos de ML mais bem empregados são aqueles que podem automatizar processos de tomada de decisão a partir de padrões conhecidos. Neste exemplo, que é conhecido como aprendizado supervisionado,

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

o usuário fornece ao algoritmo alguns pares de entradas e saídas desejadas, e o algoritmo encontra uma maneira de produzir a saída desejada dada uma entrada. Em particular, o algoritmo é

capaz de criar uma saída para uma entrada que nunca foi vista antes, sem qualquer ajuda de um humano [Müller e Guido 2017].

2.2 Aprendizado Supervisionado

Algoritmos de Machine Learning que aprendem a partir de pares de entrada e saída são chamados de supervisionados, porque um "instrutor" fornece supervisão para os algoritmos na forma dos resultados desejados para cada exemplo com o qual eles aprendem [Müller e Guido 2017]. Este modelo de aprendizado é habitualmente utilizado em aplicativos, como os de reconhecimento de rosto e voz, recomendações de produtos ou filmes e previsão de vendas. O Aprendizado Supervisionado pode ser organizado em dois tipos: Regressão e Classificação [Duda et al. 2000].

A Regressão treina e prevê uma resposta de valor contínuo, por exemplo, prevendo preços de imóveis. A Classificação tenta encontrar o rótulo de classe apropriado, como analisar sentimentos positivos e negativos, pessoas do sexo masculino e feminino, tumores benignos e malignos, empréstimos seguros e não seguros, entre outros exemplos [Izbicki e Santos 2020].

2.3 Aprendizado Não Supervisionado

O Aprendizado Não Supervisionado é utilizado para detectar anomalias, outliers, fraude ou equipamento defeituoso, ou para agrupar clientes com comportamentos seme- lhantes para uma campanha de vendas. Diferentemente da Aprendizagem Supervisionada, não há dados rotulados neste modelo [Duda et al., 2000].

Um algoritmo de aprendizagem não supervisionado tenta classificar um conjunto de dados dado em certo número de grupos de maneira eficaz. Algoritmos de aprendizagem não supervisionados são ferramentas extremamente importantes para analisar dados e identificar padrões e tendências. Eles são mais comumente usados para agrupar entradas semelhantes em grupos lógicos. Algoritmos de aprendizagem não

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

supervisionados incluem Kmeans, Random Forests, Hierarchical Clustering e outros.

3 METODOLOGIA

A pesquisa realizada é de modalidade aplicada e modelo teórico descritivo e quantitativo, com dados coletados na plataforma Kaggle, posteriormente extraídos com a utilização de algoritmos de Machine Learning e analisados por meio da linguagem Python. Os dados extraídos foram processados e analisados por meio da plataforma interativa Jupyter Notebook.

Neste artigo abordam-se referenciais teóricos que serão realizados e utilizados para embasamento em sua produção, como a definição de Fake News, as bibliotecas para análise de dados tais como Pandas, NumPy, Matplotlib, Seaborn e Scikit-Learn, conceitos de linguagem de programação Python, Machine Learning, Random Forest, Logistic Regression, Decision Tree, Support Vector Machine e K-Nearest Neighbors.

Nesta seção serão evidenciados os embasamentos teóricos a respeito da pesquisa. Serão tratados os conceitos e abordagens de Fake News, bibliotecas para análise de dados Pandas, NumPy, Matplotlib, Seaborn e Scikit-learn, linguagem de programação Python e Machine Learning, além de elencar técnicas de algoritmos como Random Forest, Logistic Regression, Decision Tree, Support Vector Machine e K-Nearest Neighbors.

4 RESULTADOS E DISCUSSÃO

Nesta seção serão explanados os principais instrumentos de desenvolvimento utilizados, bem como a explicação das etapas de evolução do projeto para a obtenção dos resultados desejados enquanto realiza-se a discussão dos achados.

4.1 Bibliotecas necessárias para o desenvolvimento

No primeiro momento, foram utilizadas as bibliotecas Pandas, NumPy, Matplot, Seaborn e Sklearn, todas necessárias para o desenvolvimento

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

do projeto na plataforma Jupyter Notebook. Essas bibliotecas são fundamentais para sua elaboração, uma vez que seria inviável ou dificultoso utilizar somente Python puro para atingir o objetivo desejado.

4.2 Carregando e analisando o tamanho das bases de dados Fake e True

Nessa etapa foi feita a carga dos dados, que foram extraídos da plataforma Kaggle, onde foram encontrados os respectivos datasets, um deles com artigos falsos e outro com artigos reais, ambos contendo as colunas: título, texto, assunto e data em que o artigo foi publicado.

Usando o método shape é possível visualizar que a base de dados fake contém 23.481 linhas e quatro colunas, e a base true, que contém 21.417 linhas e quatro colunas. Analisando os dados usando o método head que por default com as cinco primeiras linhas do dataframe.

Nessa etapa foi adicionada à label target ou variável alvo, apontando quais bases são fake ou true, foi criada uma quinta coluna em ambas as bases para expor ao algoritmo quais artigos realmente são fakes ou verdadeiros.

4.3 Realizando a concatenação das duas tabelas Fake e True

Anteriormente eram duas tabelas, uma fake e outra true, já neste ponto é feita a união das duas tabelas e com a utilização de uma biblioteca chamada Shuffle que embaralha as informações contidas no dataset. Caso isso não seja feito, o resultado do modelo pode ser enviesado, podendo o mesmo selecionar somente informações fake ou somente informações true, não sendo este um resultado ideal. Agora, pode-se perceber que existe uma tabela única com 44.898 linhas e cinco colunas.

4.4 Limpeza, tratamento e padronização dos dados

Nesta fase, inicia-se a parte de limpeza e tratamento dos dados, onde foi removida a coluna que contém a data, dessa forma, os dados ficam padronizados.

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

Neste ponto também foi removida a coluna Título, utilizando o método drop na coluna “title” dando, desse modo, continuidade no processo de limpeza e tratamento dos dados para tornar o modelo mais assertivo.

4.5 Remoção de informações adicionais

Nesta etapa, foi feita a parte de limpeza e tratamento dos dados para padronizar as informações contidas no dataset com a finalidade de melhorar a assertividade dos modelos. Primeiramente, removeu-se a coluna Data com a aplicabilidade do método drop. Após isso, removeu-se o Título e transformaram-se todas as palavras contidas no artigo em minúsculas, com a utilização de uma função lambda juntamente com o método lower. Em seguida, importou-se uma biblioteca string que retirou as pontuações, manipulando uma função e um loop for para percorrer a coluna texto e retirar as pontuações. Posteriormente, utilizou-se a biblioteca nltk que auxilia na remoção dos stopwords, que são palavras consideradas desnecessárias para a análise dos textos contidos nos artigos, como por exemplo, algumas preposições (de, e, para, as, do, da, com, entre e etc.).

4.6 Análise visual dos tópicos

Iniciando a análise dos dados propriamente dita, foi avaliada cada parte dos dados de forma detalhada.

A biblioteca WordCloud, muito utilizada para fazer análises de maneira analítica acerca de textos ou documentos, com ela é possível saber quais são as palavras usadas com mais frequência no texto e a partir dessa frequência, o tamanho será definido proporcionalmente à repetição das palavras, dessa forma, as palavras maiores são as que mais aparecem no texto e as menores são as que menos aparecem, facilitando, assim, a análise visual com a formação de uma nuvem de palavras.

4.7 Teste e treino dos modelos

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

A técnica utilizada nesta pesquisa foi o Holdout, um subconjunto que fornece a validação de dados de teste, dando uma estimativa final acerca do desempenho de determinado algoritmo utilizado e demonstrando sua eficácia após seu treinamento e validação. Entretanto, apesar de determinar o desempenho final dos modelos, ele não é um fator determinante para tomar decisões sobre qual algoritmo utilizar, seu objetivo é identificar e/ou comparar parâmetros.

A partir desta etapa serão utilizados alguns modelos que foram treinados sob os dados propostos, para então compreender qual a performance e eficiência dos algoritmos. Para tanto, serão empregados cinco modelos de Machine Learning, sendo o primeiro deles o Logistic Regression, modelo de classificação que demonstrou acurácia de 98,92%. Também foi possível comprovar através da matriz de confusão o seguinte resultado:

a) De 4.657 notícias verdadeiras positivas VP, que são as notícias realmente reais, e 67 notícias falsas positivas FP, que estão sendo compartilhadas como verdade, porém são falsas, e 43 notícias falsas negativas FN, que são dadas como falsas, porém são reais e por último obtiveram-se 4.213 notícias falsas positivas VN que de fato são falsas.

O modelo Decision Tree, algoritmo de aprendizado supervisionado empregado principalmente em problemas de classificação, alcançou nesta pesquisa uma acurácia de 99,6%. Aplicando a matriz de confusão foram demonstrados os seguintes resultados:

b) 4.711 notícias verdadeiras positivas VP, que são as notícias realmente reais, e 13 notícias falsas positivas FP, que estão sendo compartilhadas como falsas, porém são reais, e 22 notícias falsas negativas FN, que são notícias dadas como falsas, porém são reais e por último, 4.234 notícias verdadeiras negativas VN, que são as únicas falsas.

O terceiro modelo aplicado, o Random Forest, algoritmo de aprendizagem supervisionada, evidenciou uma acurácia de 98,74%, com afirmação deste dado por meio da matriz de confusão, que conteve os seguintes resultados:

c) 4.675 notícias verdadeiras positivas VP, que são as notícias realmente reais, e 49 notícias falsas positivas FP, que estão sendo compartilhadas como

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

verdadeira, porém são falsas, e 34 notícias falsas negativas FN, que são notícias dadas como falsas, porém são verdadeiras, e, por último, 4.222 notícias verdadeiras negativas VN que são, de fato, falsas.

No modelo Suporte Vector Machine, algoritmo de aprendizagem de máquina que tenta tomar dados de entrada e classificá-los, e que tem sua eficiência baseada na utilização de um conjunto de dados de entrada e saída, alcançou nesta pesquisa uma acurácia de 99,5%. Aplicando a matriz de confusão foram demonstrados os seguintes resultados:

d) De 4.699 notícias verdadeiras positivas VP, que são as notícias realmente reais, e 25 notícias falsas positivas FP, que estão sendo compartilhadas como reais, porém são falsas, e 18 notícias falsas negativas FN, que são dadas como falsas, porém são reais e por último obtiveram-se 4.238 notícias verdadeiras negativas VN, que de fato são falsas.

O último modelo aplicado, o k-Nearest Neighbors (kNN), um dos algoritmos mais utilizados no ML devido ao seu processo de classificação, obteve o resultado de uma acurácia de 60.84%, com afirmação deste dado por meio da matriz de confusão que conteve os seguintes resultados:

e) De 4.709 notícias verdadeiras positivas VP, que são de fato notícias realmente reais, e 15 notícias falsas positivas FP, que estão sendo compartilhadas como verdadeiras, porém são falsas, e 3502 notícias falsas negativas FN, que são dadas como falsas, mas na verdade são verdadeiras e por último obtiveram-se 754 notícias verdadeiras positivas VN, que de fato são falsas.

4.8 Métricas de desempenho

Neste tópico destacam-se os resultados das características em cada modelo utilizado. A acurácia refere-se à quantificação de acertos de previsões possíveis, a precisão determina o quão próximo o modelo conseguiu predizer, a sensibilidade diz respeito ao quanto o modelo foi distintivo nos acertos das predições verdadeiras e a especificidade refere-se à eficiência do modelo em acertar o que de fato é negativo.

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

a) Logistic Regression

Acurácia: VP + VN / VP + FP + FN + VN = 98%

Precisão: VP/VP+FP = 98%

Sensibilidade: VP/VP+FN = 99%

Especificidade: FP/FP+VN = 15%

b) Decision Tree

Acurácia: VP + VN / VP + FP + FN + VN = 98%

Precisão: VP/VP+FP = 98,5 = 99%

Sensibilidade: VP/VP+FN = 99%

Especificidade: FP/FP+VN = 3%

c) Random Forest

Acurácia: VP + VN / VP + FP + FN + VN = 99%

Precisão: VP/VP+FP = 98,5 = 98%

Sensibilidade: VP/VP+FN = 99%

Especificidade: FP/FP+VN = 11%

d) SVM

Acurácia: VP + VN / VP + FP + FN + VN = 99%

Precisão: VP/VP+FP = 99%

Sensibilidade: VP/VP+FN = 99%

Especificidade: FP/FP+VN = 5%

e) KNN

Acurácia: VP + VN / VP + FP + FN + VN = 60%

Precisão: VP/VP+FP = 99%

Sensibilidade: VP/VP+FN = 57%

Especificidade: FP/FP+VN = 19%

5 CONSIDERAÇÕES FINAIS

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

O advento das redes sociais juntamente com a facilitação de seu uso por meio de smartphones, tablets e notebooks trouxe, entre benefícios e malefícios, a crescente propagação das Fake News. As notícias falsas que são divulgadas como fatos podem gerar problemas políticos, de educação e saúde em uma sociedade. Tais problemas requerem soluções que busquem facilitar a compreensão daquilo que possa ser real ou uma inverdade, principalmente na internet, onde as notícias são geradas de forma rápida e compartilhadas indiscriminadamente.

Sendo assim, os fatores supracitados evidenciam a relevância do desenvolvimento de pesquisas e da elaboração de aplicações tecnológicas que auxiliem na solução de problemas ocorrentes do mundo digital como, por exemplo, a disseminação das Fake News.

O Machine Learning, por ser uma ferramenta capaz de extrair conhecimento dos dados produzidos initerruptamente e auxiliar na forma como sua pesquisa é feita, pode contribuir para saber a veracidade das informações compartilhadas na internet. Seu uso está em evolução em diversas áreas, inclusive na detecção e predição de artigos e sua veracidade. O desenvolvimento deste tipo de aplicação pode gerar vantagens de forma considerável, uma vez que os modelos manipulados se apresentaram capazes de detectar corretamente as informações falsas ou verdadeiras.

Este estudo utilizou técnicas de Machine Learning tais como Holdoult e variados algoritmos de classificação para fazer as predições sobre os artigos utilizados na base de dados, mediante às características contidas no conjunto de dados tais como as instâncias e atributos. Foram aplicados cinco modelos, entre os quais, quatro alcançaram acurácia superior a 90% e somente um restringiu-se à margem de 60%, conforme explicitado nos capítulos anteriores.

Os algoritmos de classificação Random Forest e Suport Vector Machine demonstraram o melhor desempenho e o que evidenciou a menor performance foi o KNN, com maior incidência de erros nos resultados de Falsos Negativo, incluindo uma Sensibilidade de 57% e Especificidade de 19%.

Ademais, destaca-se como limitação do trabalho a busca pela base de dados com notícias Fake e True, uma vez que foi dificultoso encontrar uma base

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

padronizada e que contivesse informações concisas. Além disso, é importante ressaltar que durante a pesquisa foi observado que há escassez de aplicações para detecção de notícias falsas, sobretudo no Brasil. Portanto, é necessária a ampliação e prosseguimento de projetos deste âmbito que busquem, de preferência, avaliar outros algoritmos de classificação para fazer predição de artigos e notícias, como Naive Bayes, Boosting, K-means e Gradiente Descendente, aplicar outras técnicas de aprendizado de máquina como a Validação Cruzada, K-fold e Leave-one-4 a fim de mensurar o melhor desempenho, bem como avaliar a performance dos modelos sem uso da variável target, e ainda, acrescer a utilização de bases de dados na língua portuguesa (Brasil).

Assim sendo, a metodologia adotada neste estudo visa à inserção com uma aplicação WEB para fazer detecção de notícias reais ou falsas, classificando-as e, dessa forma, utilizando e comprovando a importância da aplicabilidade de Machine Learning para dirimir a disseminação e os malefícios que as Fake News podem provocar na sociedade em geral.

REFERÊNCIAS

ALBON, Chris. ***Machine Learning with Python Cookbook***. 1ª edição. Estados Unidos: *O'Reilly Media,* Abril de 2018.

AMORIM, Paulo H. J.; MORAES, Thiago F. de; AZEVEDO, Fábio de S.; SILVA, Jorge V. L. da. INVESALIUS: SOFTWARE LIVRE DE IMAGENS MÉDICAS. Nº 4, 2011. **Anais eletrônicos**. Campinas – SP. Disponível em: <WIM_Sessao_1_Artigo_4_Amorim.pdf (ufrn.br)>. Acesso em 25 de março de 2021.

BARRETT, P., HUNTER, J., MILLER, J. T., J.-C. GREENFIELD, Hsu, and P., "matplotlib -- A Portable Python Plotting Package," **ASP Conf. Ser**., vol. 347, no. June, p. 91, 2015.

BEYER, Kevin et al. **When is "nearest neighbor" meaningful?.** International conference on database theory. Springer, Berlin, Heidelberg, 1999. p. 217-235.

CARVALHO, Mariana Freitas Canielo de; MATEUS, Cristielle Andrade. Fake News E Desinformação No Meio Digital: Análise Da Produção Científica Sobre O Tema Na Área De Ciência Da Informação. **V EREBD**, 2018. Universidade Federal de Minas Gerais. Belo Horizonte – MG.

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

COUTINHO, Bernardo**. Modelos de predição SVM.** Medium, 2019. Disponível em: <https://medium.com/turing-talks/turing-talks-12-classifica%C3%A7%C3%A3o-por-svm-f4598094a3f1>. Acesso em: 01 de Outubro de 2020.

DUDA, Richard O.. HART, Peter E. STORK, David G.. **Pattern Classification**. 2ª edição. Wiley-Interscience, 2000.

FOSSENG, S. **Learning Distance Functions in k-Nearest Neighbors**. Dissertação (Mestrado) — Institutt for datateknikk og informasjonsvitenskap, 2013.

GUIMARÃES, Amanda Munari. Estatística: Análise de regressão linear e análise de regressão logística com R. **Medium**, 2019. Disponível em: <https://medium.com/omixdata/estat%C3%ADstica-an%C3%A1lise-deregress%C3%A3o-linear-e-an%C3%A1lise-de-regress%C3%A3o-log%C3%ADsticacom-r-a4be254df106>. Acesso em: 01 de Outubro de 2020.

Introdução ao Python. **DevMedia**. Disponível em: <https://www.devmedia.com.br/guia/python/37024>. Acesso em: 29 de Setembro de 2020.

IZBICKI, Rafael. SANTOS, Tiago Mendonça dos. **Aprendizado de máquina: uma abordagem estatística** [livro eletrônico], - São Carlos, SP, 2020.

RISING, F. O. I.. ODEGUA, O., **“DataSist: APython-based library for easy data analysis, visualization and modeling,”** 2017.

CHANG, Winston. R graphics cookbook: practical recipes for visualizing data. "

O'Reilly Media, Inc.", 2012.