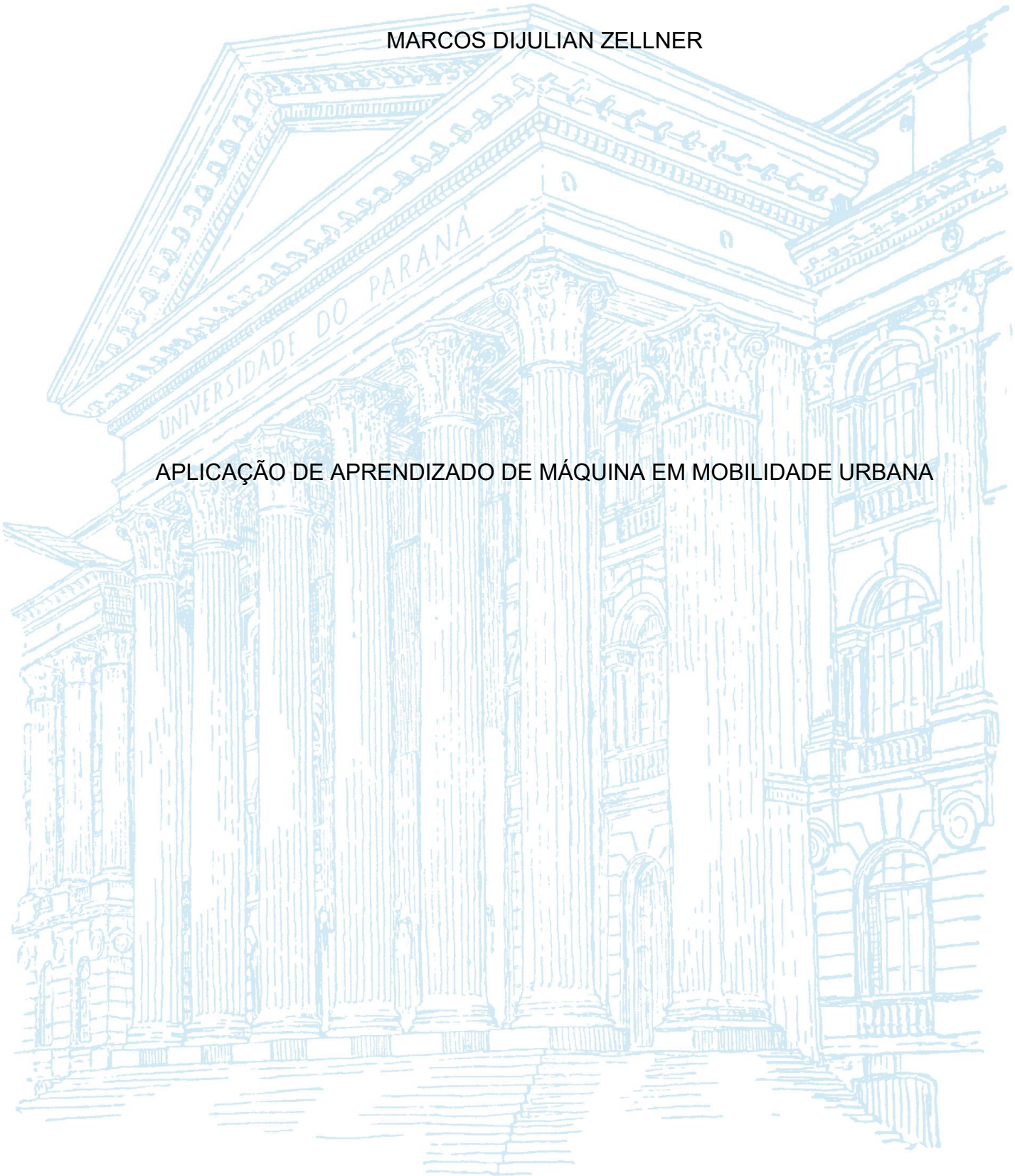
UNIVERSIDADE FERDERAL DO PARANÁ

MARCOS DIJULIAN ZELLNER



APLICAÇÃO DE APRENDIZADO DE MÁQUINA EM MOBILIDADE URBANA

CURITIBA

2024

MARCOS DIJULIAN ZELLNER

APLICAÇÃO DE APRENDIZADO DE MÁQUINA EM MOBILIDADE URBANA

Trabalho de conclusão de curso apresentado como requisito parcial à obtenção do grau de Bacharel em Gestão da Informação no curso de graduação em Gestão da Informação, Departamento de Ciência e Gestão da Informação do Setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná.

Orientadora: Profª. Drª. Denise Fukumi Tsunoda.

CURITIBA

2024

AGRADECIMENTOS

À Professora Dra. Denise Fukumi Tsunoda, minha profunda gratidão por sua paciência nas orientações, pelas valiosas dicas, conselhos, recomendações, pelo apoio emocional e pela confiança depositada em mim. Sua dedicação foi essencial para me motivar na busca por conhecimento, permitindo a conclusão deste estudo e despertando em mim o desejo de continuar explorando esta área.

Agradeço ao meu pai por seu constante incentivo nos estudos, especialmente nos momentos difíceis. Sua presença ao meu lado, oferecendo apoio e encorajamento, foi fundamental para que eu continuasse a perseguir meus sonhos e aspirações.

À minha prima Ana, minha eterna gratidão por tudo que você fez e faz por mim. Seu apoio incondicional e sua disponibilidade em estar presente sempre que preciso são gestos que guardo com muito carinho. Você não apenas me ajudou em momentos importantes, mas também mostrou o verdadeiro significado de família.

À minha vó, que infelizmente não teve a oportunidade de presenciar minha formatura, mas cuja presença e apoio foram essenciais desde o momento em que iniciei o curso de Gestão da Informação. Sua curiosidade constante sobre como eu estava indo e suas palavras de incentivo mostraram o quanto ela acreditava em mim e no meu potencial.

A Luana, meu mais sincero agradecimento por estar ao meu lado durante todo o curso. Seu apoio, carinho e amizade foram fundamentais em cada etapa, tornando todas as matérias mais leves e os desafios mais fáceis de enfrentar. A forma como você esteve presente, compartilhando momentos, nas festas na sua casa, em viagens feitas, ou somente em estar presente na minha vida, tenho muito orgulho de tê-la ao meu lado e conto que essa amizade dure para sempre.

Helena, quero te agradecer imensamente por tudo o que vivemos durante o curso. Passei por momentos de grande dificuldade, e você esteve sempre ao meu lado, me apoiando. Tenha certeza de que o curso foi muito mais fácil e leve com a sua presença.

Quero agradecer ao meu amigo Nicolas, você foi meu primeiro amigo no curso, mas espero que essa amizade dure para sempre, aproveitamos muito dentro e fora do curso com jogatinas, trilhas e viagens.

Obrigado a Ana Clara pela sua amizade, todas as matérias que fizemos juntos foram com certeza mais divertidas.

A todos meus professores pela passagem de conhecimento e de experiência de vida e que contribuíram de alguma forma para minha formação profissional e acadêmica. Vocês são fundamentais na formação de qualquer profissional e por isso meus professores têm o meu eterno agradecimento.

“Todos mudamos, se pensarmos nisso.
Somos diversas pessoas ao longo da vida.
E tudo bem. Isso é bom. Temos de continuar em frente.
Mas sempre devemos lembrar de todos que fomos antes.”
DOCTOR WHO

RESUMO

O estudo utiliza algoritmos de aprendizado de máquina para analisar dados de mobilidade urbana. O objetivo geral deste estudo é propor uma metodologia que otimize o fluxo de tráfego nas grandes cidades, utilizando bases de dados artificiais que simulam cenários urbanos complexos. Para atingir esse objetivo, foi realizada a prospecção de estudos relacionados ao tema de pesquisa, identificando os algoritmos mais adequados ao problema. Em seguida, foi concebida uma base de dados artificial, representando diferentes dimensões da mobilidade urbana, como fluxo de tráfego, uso de transporte público e condições climáticas. Por fim, foram aplicados algoritmos de aprendizado de máquina, incluindo Random Forest, KNN e RNA, para modelar padrões e prever congestionamentos. A coleta de dados foi realizada em bases disponíveis em repositórios, como Kaggle e UCI Repository. Após o pré-processamento, os dados foram categorizados em variáveis temporais, fatores de mobilidade e métricas de desempenho, garantindo a qualidade e a representatividade dos cenários analisados. Os resultados obtidos mostraram que os algoritmos apresentaram acurácia entre 64% e 66%, dependendo da complexidade do modelo. O Random Forest se destacou como o mais robusto, enquanto o KNN apresentou maior sensibilidade a padrões locais. Os desafios incluíram a necessidade de balanceamento de classes e ajustes de hiperparâmetros para melhorar a performance geral. Como proposta de trabalhos futuros, sugere-se a aplicação dessa metodologia em bases de dados reais e a ampliação do estudo para incluir outros fatores de mobilidade urbana, como impactos ambientais e acessibilidade. A replicação da metodologia em cidades com diferentes características também pode fornecer insights valiosos para gestores urbanos.

Palavras-chave: mobilidade urbana, aprendizado de máquina, mineração de dados, redes neurais artificiais.

ABSTRACT

This study uses machine learning algorithms to analyze urban mobility data. The overall aim of this study is to propose a methodology that optimizes traffic flow in large cities, using artificial databases that simulate complex urban scenarios. In order to achieve this objective, a survey of studies related to the research topic was carried out, identifying the algorithms best suited to the problem. Next, an artificial database was designed, representing different dimensions of urban mobility, such as traffic flow, use of public transport and climatic conditions. Finally, machine learning algorithms were applied, including Random Forest, KNN and RNA, to model patterns and predict congestion. Data was collected from databases available in repositories such as Kaggle and UCI Repository. After pre-processing, the data was categorized into time variables, mobility factors and performance metrics, ensuring the quality and representativeness of the scenarios analyzed. The results showed that the algorithms were between 64% and 66% accurate, depending on the complexity of the model. Random Forest stood out as the most robust, while KNN showed greater sensitivity to local patterns. Challenges included the need for class balancing and hyperparameter adjustments to improve overall performance. As a proposal for future work, we suggest applying this methodology to real databases and extending the study to include other urban mobility factors, such as environmental impacts and accessibility. Replicating the methodology in cities with different characteristics could also provide valuable insights for urban managers.

Keywords: urban mobility, machine learning, data mining, artificial neural networks.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS OU SIGLAS

UFPR	- UNIVERSIDADE FEDERAL DO PARANÁ
ML	- MACHINE LEARNING
AP	- APRENDIZADO DE MÁQUINA

SUMÁRIO

1 INTRODUÇÃO

A mobilidade urbana é um dos principais desafios enfrentados, principalmente, pelas grandes cidades. Com o crescimento acelerado das populações urbanas, a demanda por transporte eficiente e sustentável tem aumentado significativamente “O aumento da mobilidade, resultado do incremento dos fluxos de pessoas e bens, tem implicado em impactos negativos sobre o ambiente local e global, sobre a qualidade de vida e sobre o desempenho econômico das cidades” (Costa, 2008, p.7). Congestionamentos, poluição do ar e ineficiência nos sistemas de transporte público são problemas recorrentes que afetam não apenas a qualidade de vida dos cidadãos, mas também a economia e o meio ambiente.

No cenário atual, tecnologias emergentes, como o aprendizado de máquina, oferecem novas possibilidades de abordagem desses desafios. O aprendizado de máquina é um subcampo da inteligência artificial com o potencial de transformar a forma como é realizada a coleta, análise e uso de diversos tipos de dados, inclusive os de mobilidade urbana “a área de Aprendizado de Máquina estuda métodos computacionais para adquirir novos conhecimentos, novas habilidades e novos meios de organizar o conhecimento já existente” (MITCHELL et al., 1990). Quando é pensado nessa análise de dados, aprendizado de máquina é uma área que utiliza algoritmos para processar grandes quantidades de dados, auxiliando na análise e na obtenção de resultados conclusivos.

No contexto da mobilidade urbana, dados sobre o trânsito, padrões de deslocamento e *feedback* dos usuários são coletados e analisados para melhorar a eficiência dos sistemas de transporte. Ao coletar esses dados é possível aplicar os modelos de aprendizado de máquina como de regressão e classificação que podem, por exemplo, ajudar a prever os horários de pico e sugerir rotas alternativas ou melhorias na infraestrutura de transporte, auxiliando na tomada de decisões informadas para otimizar a mobilidade nas cidades “Para além dos benefícios individuais, essas tecnologias de análise de grandes massas de dados se apresentam como ferramentas poderosas no auxílio à tomada de decisão pelos gestores” (Morares, 2021, p.33).

Para a aplicação dos modelos de aprendizado de máquina é necessário possuir um repositório de dados para aplicação de um algoritmo e dele retirar

resultados que geram conhecimento. Este estudo tem por objetivo propor usar uma metodologia de análise de dados em uma base de dados de mobilidade urbana.

1.1 PROBLEMA DE PESQUISA

O aprendizado de máquina teve seu começo na década de 50, com o algoritmo chamado "*perceptron*" esses avanços pavimentaram o caminho para o desenvolvimento subsequente dos algoritmos e técnicas de aprendizado de máquina que utilizamos atualmente. Essa máquina nomeada por Frank Rosenblatt que futuramente se tornou protótipo para as redes neurais artificiais, e seu modelo de aprendizado era semelhante aos modelos de aprendizado animal e humano desenvolvidos na psicologia como dito por Neto (2010) algoritmo *perceptron* é a forma mais simples de rede neural artificial, sendo utilizado como classificador linear, o qual consiste em um único neurônio com pesos sinápticos ajustáveis e bias.

O aprendizado de máquina tem evoluído e se desenvolvido cada vez mais, impulsionado por avanços tecnológicos, maior poder computacional e a disponibilidade de grandes conjuntos de dados. Dados estão por todos os lugares, de fogões a carros, "a virada do milênio testemunhou avanços extraordinários, impulsionados pela disponibilidade de grandes conjuntos de dados e poder computacional exponencialmente maior" (Handaya, 2024, p.3), tudo que fazemos gera uma informação que pode ser útil se aplicarmos a um modelo de aprendizado de máquina, só no ano de 2012 foi produzido e criado muito mais dados do que todos os dados criados e somados dos últimos 5 mil anos anteriores (Barros, 2017). Diante disso, cada vez mais surge a necessidade de construir técnicas para melhorar processos que vivemos no dia a dia.

Um dos grandes desafios contemporâneos é a mobilidade urbana. Com o crescimento das cidades e o aumento das frotas de veículos, sejam eles carros, motos ou bicicletas, há uma necessidade de analisar e otimizar esses dados para melhorar a eficiência dos sistemas de transporte. O tempo gasto no trânsito tem um impacto significativo na qualidade de vida das pessoas e na produtividade das cidades, segundo Araújo (2011) o planejamento da dinâmica da mobilidade no espaço urbano torna-se essencial, pois influencia diretamente tanto a qualidade de vida das pessoas, reduzindo o tempo gasto no trânsito, quanto a produtividade das cidades como um todo. Com isso, a aplicação de técnicas de aprendizado de

máquina para a análise de dados de mobilidade urbana pode oferecer soluções inovadoras para reduzir congestionamentos, melhorar o transporte público e promover um deslocamento mais eficiente, beneficiando tanto os cidadãos quanto os gestores urbanos.

Com base no exposto, busca-se solucionar o principal problema deste estudo: **Como técnicas de aprendizado de máquina podem ser aplicadas para melhorar a mobilidade urbana?**

1.2 OBJETIVOS

Depois de analisar e definir o problema de pesquisa, foi possível estabelecer os objetivos a serem seguidos neste projeto. Nas próximas seções, serão apresentados o objetivo geral e os objetivos específicos deste estudo.

1.2.1 Objetivo geral

Este estudo tem como objetivo principal aplicar soluções baseadas em aprendizado de máquina que contribuam para a gestão eficiente e a otimização da mobilidade urbana.

1.2.2 Objetivos específicos

Para atingir o objetivo geral deste estudo deverão ser atingidos os objetivos específicos a seguir:

a) realizar uma revisão bibliográfica sobre estudos e aplicações de aprendizado de máquina no contexto da mobilidade urbana;
b) treinar modelos de aprendizado de máquina para análise e previsão de padrões de mobilidade urbana;
c) documentar os resultados e apresentar as análises e considerações de forma clara e detalhada com vistas às melhorias decorrentes do estudo.

1.3 JUSTIFICATIVA

O tema aprendizagem de máquina aplicado no levantamento de dados sobre mobilidade urbana ainda é uma área que pode ser explorada “uma vez que os sistemas de indicadores de mobilidade urbana sustentável se constituem em ferramentas ainda pouco exploradas no Brasil” (Costa, 2008, p.61). Foi feito um levantamento na base de Periódicos Capes do governo federal do Brasil referente ao tema “Mobilidade Urbana”. O objetivo dessa busca foi verificar a existência de publicações em periódicos das áreas de informação e tecnologia relacionadas ao tema deste estudo. Os resultados da pesquisa feita em 28/06/2024, no site Periódicos Capes resultaram em 181 títulos relacionados ao assunto.

Conforme os resultados apresentados, em uma análise dos 30 primeiros artigos, foi lido seus resumos e palavras-chaves, não foram encontradas nenhuma relação com o tema desta pesquisa.

No final da pesquisa realizada na base Periódicos Capes, pode-se verificar que existiam dois documentos publicados em periódicos da área da informação e o outro na área da tecnologia. Porém nenhum dos dois abrange a área de Aprendizado de Máquinas focado na otimização da mobilidade urbana. O documento do periódico da área da tecnologia de Ansélmo, Borille e Correia (2022) “Análise de sentimentos sobre o acesso terrestre ao aeroporto utilizando mídias sociais”. É referente ao uso de aprendizado de máquinas, mas com o foco na análise de sentimentos usando a técnica de Naïve Bayes que foi lida no artigo “*A framework with eficient extraction and analysis of Twitter data for evaluating public opinions on transportation services.*” (Qi, Costin & Jia, 2020), onde o termo “Mobilidade Urbana” é utilizado para se referir a ônibus, táxi, trem e veículos privados da região de Guarulhos onde foi feita a pesquisa, o qual não é o foco deste estudo. Para o artigo da área da informação escrito por Lima e Fontgalland (2022) “Mobilidade Urbana Sustentável para Cidades Inteligentes”. foi possível verificar que se trata de um estudo de uma pesquisa documental e social do tema mobilidade urbana, com as leis e diretrizes focadas no tema. Não envolvendo a área da tecnologia.

Fora a pesquisa dessa análise científica que justifica a falta de pesquisa na área de aprendizado de máquinas, é um tema que despertou a curiosidade do autor

por conta de experiências passadas. Esse fato motivou o uso e aprofundamento desse tema de pesquisa.

Com cada vez mais veículos nas ruas das grandes cidades, principalmente no ano de 2023 com as marcas de carros mudando para carros elétricos ou híbridos, ter meios de amenizar o caos do trânsito é uma ótima motivação de pesquisa e ainda conseguir envolver o aprendizado de máquina para isso, pode de certa forma, subsidiar novas pesquisas que aprofundem ainda mais a inteligência artificial.

Para o curso de Gestão da Informação (GI) esse é um tema que permite aplicar os conceitos estudados em todos os núcleos de aprendizado das matérias, como conceitos vistos na ciência da informação como identificação de extração de conhecimentos e ciclo informacional, na administração apresentando pontos que gestores visão ter na concepção de projetos informacionais. E para a tecnologia da informação com aplicação dos estudos de mineração de dados, banco de dados e o uso da linguagem de programação para criação de *insights* úteis para empresas e a sociedade.

1.4 LIMITAÇÕES

Essa pesquisa possui limitações de coleta de dados de mobilidade urbana, como dados de trânsito, transporte público e comportamento dos usuários, que são restritos por questões de privacidade ou propriedade. Para superar isso foi feito um levantamento de bases de dados livres de todo o mundo para juntar e criar uma base que possa servir como objeto de pesquisa.

Outro fator limitador é a variabilidade de cenários urbanos, pois modelos que funcionam bem em uma cidade podem não ser diretamente aplicáveis a outra devido a diferenças nas infraestruturas de transporte, padrões de mobilidade e comportamentos dos usuários.

1.5 ESTRUTURA DO DOCUMENTO

Este documento está estruturado em seções, conforme descrito a seguir. Na seção 2, apresenta-se o embasamento teórico desta pesquisa, abrangendo conceitos aplicados à Gestão da Informação e Mineração de Dados.

A seção 3 detalha a descrição e os procedimentos metodológicos escolhidos, incluindo a metodologia de pesquisa, a população e a amostragem.

Na seção 4, são apresentados os resultados e discussões.

Finalmente, na seção 6, encontram-se as considerações finais deste estudo e sugestões para pesquisas futuras na área.

2 LITERATURA PERTINENTE

Nas seguintes seções serão apresentados conceitos que fundamentam a pesquisa e proposta do estudo. Serão abordados os conceitos de gestão da informação; Mineração de dados; Aprendizagem de máquina e mobilidade urbana com suas definições e importância.

O propósito desta pesquisa foi desenvolvido com o auxílio desses conceitos e ajudaram no desenvolvimento da metodologia para a construção da base de dados usada.

2.1 GESTÃO DA INFORMAÇÃO

A informação tem ganhado cada vez mais importância e destaque com o surgimento de novas tecnologias, o acesso à informação ganhou extrema relevância (Pinheiro, 2002). A informação se tornou o elemento mais valioso na chamada Era da Informação, pois tudo, ou quase tudo nessa sociedade, é impulsionado por ela. As pessoas, as organizações públicas e as privadas estão atreladas à informação e dependem dela para melhor gerenciar seus recursos e, desse modo, oferecer melhores produtos e serviços (Campagnaro, Cervantes, 2011).

A primeira vez que o termo gestão da informação foi mencionado na era moderna foi em 1934, no trabalho de Paul Otlet, Traité de documentation, e um dos principais fatores para GI ser o que é conhecido nos dias de hoje veio deste livro. De fato, muito do que hoje é conhecido modernamente por gerência de recursos informacionais tem suas origens nos trabalhos de Otlet. (Barbosa, 2008). Antes conhecida como documentação, o GI passou a ter um foco no acesso às informações de uma forma mais rápida, e de fácil recuperação. Entre os anos de 1980 e 1990 Rowley (1988) descreveu o papel da gestão da informação como o de organização, planejamento de políticas de informação, desenvolvimento e manutenção de sistemas e serviços, a otimização dos fluxos de informação e o aproveitamento de tecnologias de ponta aos requisitos funcionais dos usuários finais, isso em qualquer organização.

O uso da gestão da informação por conta da sua multidisciplinaridade, é geralmente apresentada no contexto organizacional, McGee e Prusak (1994), para a compreensão da gestão da informação sob a perspectiva de processo é preciso

entendê-la como um conjunto de atividades com aspectos dinâmicos, conectadas logicamente e que cruzam limites funcionais das organizações. Mas além do ambiente organizacional, Páez Urdaneta (1992), GI é um conjunto de elementos e processos vitais dentro da gestão em diferentes dimensões da informação, sendo eles dentro ou fora das organizações. Choo (2003) indicou os processos da GI que são identificação das necessidades de informação, aquisição da informação, organização e armazenamento da informação, desenvolvimento de produtos e serviços de informação, distribuição da informação e o seu uso.

Para um profissional da informação aplicar esses processos existe uma necessidade que apresente conhecimentos multidisciplinares, técnicas de utilização de processos, habilidades gerenciais de controle da informação. Segundo Marchiori (2002), habilidades de análise, condensação, interpretação, representação e estratégias de busca e apresentação/formatação da informação considerando diferentes suportes.

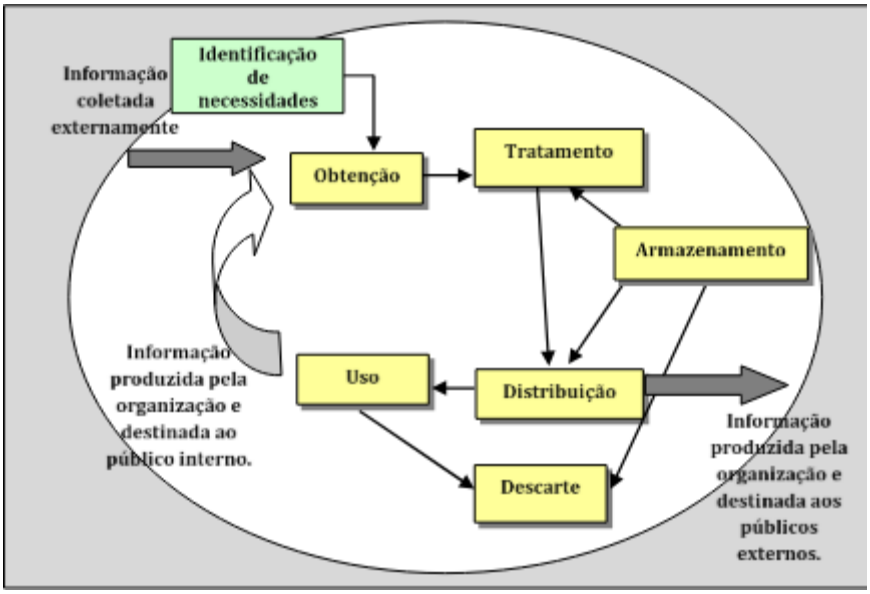
Junto ao que já foi exposto, a gestão da informação engloba o componente de Ciclo de Vida da Informação. A próxima seção irá apresentar a definição e conceitos.

2.1.1 Ciclo de Vida da Informação

O Ciclo de Vida da Informação refere-se às etapas pelas quais a informação passa desde a sua criação até a sua disposição final. Compreende ciclo de vida da informação que é composto pelas seguintes etapas: identificação das necessidades da informação, criação da informação, aquisição da informação, organização da informação, armazenamento da informação, disseminação da informação, distribuição da informação e uso da informação (Aganette, Nonato, 2022).

Compreender esse ciclo é essencial para a gestão eficaz da informação, garantindo que os dados sejam coletados, armazenados, processados, distribuídos e utilizados de maneira eficiente e segura. O ciclo de vida da informação ganhou tal importância na medida em que alimenta e renova o conhecimento organizacional, fornecendo insumos tanto para a tomada de decisões quanto para a realização de atividades-meio e atividades fim (Beal, 2012). Para representar isso foi escolhido a FIGURA 1 que mostra como é o seu ciclo de vida da informação.

FIGURA 1 – Modelo de representação do fluxo da informação



FONTE: Beal (2012, p. 29)

No primeiro momento, a informação é vista sob a perspectiva da coleta da informação realizada fora do ambiente organizacional. Em um segundo e terceiro momentos, respectivamente, a informação produzida pela organização é destinada ao público interno e externo, vislumbrando atender às suas necessidades (Miranda et al., 2019).

2.2 MINERAÇÃO DE DADOS

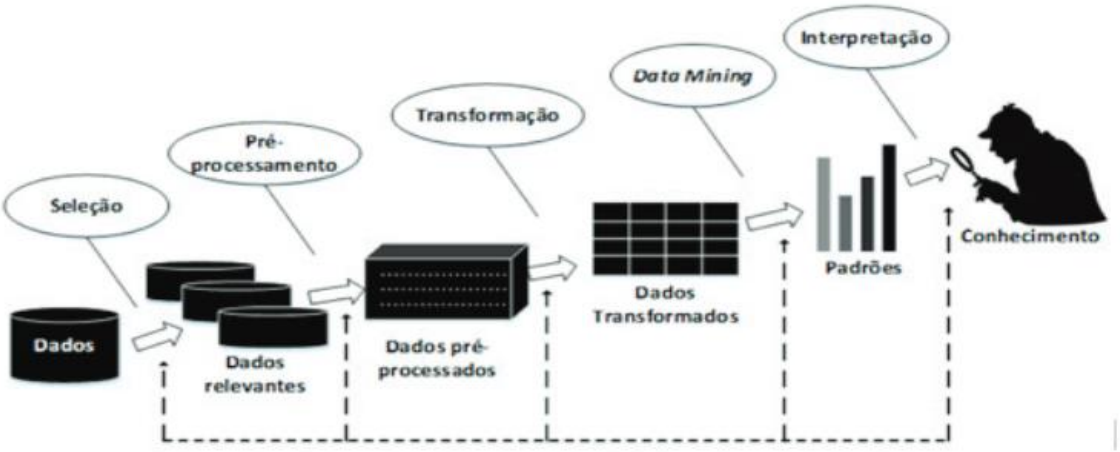
Com cada vez mais equipamentos do dia a dia conectados na internet, o ambiente da *web* se tornou um local de grande armazenamento de dados e informações. A *big data*, que é a área do conhecimento que estuda como tratar, analisar e obter informações a partir de conjuntos de dados muito grandes que serão apresentados nas próximas seções, surgiu dessa grande explosão de dados. Para conseguir se adaptar a essa grande quantidade de dados foi criado e desenvolvido a mineração de dados, pois se a ajuda desse método o homem não conseguiria analisar a aplicar a grande quantidade sem o auxílio de ferramentas computacionais (Goldscmidt; Passos, 2005).

Combinando técnicas de estatística e programação avançada conseguiram desenvolver sistemas que podem efetuar a extração e sumarização automática de informações úteis a partir de grandes bases de dados, o que foi chamado de Mineração de Dados (Quilici-Gonzalez; Zampirolli, 2015). O termo mineração de

dados tem sido usado principalmente por estatísticos, analistas de dados e as comunidades de sistemas de informação de gestão, também popularizando no campo de banco de dados (Fayyad, Piatetky-Shapiro; Smyth, 1996). Junto a isso as novas profissões como cientista de dados usam a mineração de dados como ferramenta de trabalho.

A mineração de dados está fortemente ligada às etapas de KDD (*Knowledge Discovery in Data bases*) sendo uma etapa desse processo. É na etapa de mineração de dados (*Data Mining*) onde acontece a exploração e análise, de forma automática ou semiautomática, de grandes bases de dados com objetivo de descobrir padrões e regras (Fayyad et al., 1996). Como descrito na FIGURA 2.

FIGURA 2 – Etapas do processo KDD



FONTE: Adaptado de Fayyad e outros autores (1996).

Para realizar o processo de mineração dos dados é possível a partir de três etapas:

a) pré-processamento: Um conjunto de dados pode conter diversos tipos de ruídos e/ou imperfeições, como valores incorretos, inconsistentes, duplicados ou ausentes. Frequentemente são utilizadas técnicas de pré-processamento de dados para melhorar a qualidade dos mesmos, essas técnicas podem ser de eliminação ou minimização dos problemas citados (Costa, Bernadili, Viterbo, 2014);
b) exploração dos dados: os dados processados são alimentados por meio de um algoritmo de mineração de dados que irá produzir padrões ou conhecimento;

c) pós-processamento: engloba o tratamento do conhecimento obtido na mineração de dados; as principais funções são elaboração e organização, também podendo incluir simplificação de gráficos, diagramas, ou relatórios demonstrativos (Horn et al., 2020).

Outra definição para mineração de dados é a de Castro e Ferrari (2016) Mineração de Dados por ser uma disciplina interdisciplinar e multidisciplinar envolve conhecimento de áreas como banco de dados, estatística, aprendizagem de máquina, computação de alto desempenho, reconhecimento de padrões, computação natural, visualização de dados, recuperação da informação, processamento de imagens e de sinais, análise espacial de dados, inteligência artificial. Fontes de dados para esse propósito incluem bancos de dados, *data warehouses*, a *web* e outros repositórios de informações/dados, que são transmitidos para o sistema de maneira dinâmica.

2.2.1 Big Data

Com cada vez mais dados surgindo de forma mais desestruturada graças ao avanço da indústria 4.0, as organizações privadas precisaram se adaptar às mudanças. *Big data* é definido pelo grande volume de dados de diversas fontes, estruturados ou não (Arunachalam et al., 2018;Brinch et al., 2018; Félix et al., 2018; Mikalef et al., 2019). A definição de big data abrange muitos elementos: volume, velocidade, veracidade, variedade, verificação e valor (Manyika et al., 2011; Poulovassilis, 2016).

Por conta de sua grande quantidade de dados, a *Big Data* normalmente está associada a técnicas de aprendizado de máquina que requerem essa grande quantidade para melhorar sua performance. Leonelli (2022) define como dados de diferentes tipos e origens que se relacionam, muitas vezes em formato digital e de formas que se prestam ao aprendizado de máquina, de modo a produzir novos procedimentos de análise e conhecimento.

Existem muitas definições de o que é a *Big Data*, Kitchin (2013) elaborou as características para ser considerado *Big Data*:

a) enorme em volume, composto por *terabytes* ou *petabytes* de dados;
b) alta velocidade, sendo gerados praticamente em tempo real;

c) diversos, sendo de natureza estruturada ou não estruturada;
d) exaustivo em escopo, pois busca capturar populações ou sistemas inteiros;
e) refinado em resolução, ou seja, busca ser o mais detalhado possível;
f) de natureza relacional, contendo campos comuns que permitem a junção de diferentes conjuntos de dados.
g) flexível, sendo possível a adição de novos campos e escalável, permitindo a expansão do seu tamanho de maneira rápida.

Outra definição para esse tipo de dados é a de Dutra, Karpinski e Silva Junior (2020) a *big data* não se refere exclusivamente a grandes coleções de dados e às ferramentas e procedimentos para manipulá-los e analisá-los, mas é também uma mudança fundamentada no modo de pensar e pesquisar cientificamente. Normalmente os dados por trás da *big data* estão ligados por cinco “v” volume, velocidade, variedade, veracidade e valor. Esse valor quer dizer que dados tem pouco valor em relação ao volume. No entanto, é possível conseguir um alto valor analisando grandes volumes desses dados (Gandomi; Haider 2015).

O volume diz respeito à quantidade de dados produzidos: diariamente, bilhões de e-mails, mensagens, interações em redes sociais, transferências bancárias, chamadas telefônicas e diversas outras atividades geram uma enorme quantidade de rastros digitais. A velocidade é a comparação dessa produção em relação ao tempo: são produzidos a cada milésimo de segundo. A variedade significa que os dados aparecem em diversas formas e não em padrões únicos: sejam textos, fotos, vídeos, áudios, documentos e entre outros. A veracidade diz respeito à confiabilidade em relação à fonte. E por fim, o valor é o ponto mais importante da *big data*. Não adianta ter acesso a um enorme volume, grande variedade de dados confiáveis e coletados em tempo real, se você não for capaz de gerar valor através deles. É o objetivo-fim de quem acumula e opera *Big Data*: agregar valor (capital) através do conhecimento extraído pela análise.

A aplicação dessas tecnologias não apenas facilita o manejo dos "5 Vs" da *Big Data*, mas também potencializa a capacidade das empresas de inovar e se adaptar rapidamente às mudanças do mercado. O uso estratégico de *Big Data*, portanto, não se limita à capacidade de lidar com grandes volumes de informações, mas envolve uma abordagem integrada e inteligente para extrair valor real e aplicável dos dados disponíveis.

2.2.2 Banco de dados

Para treinar um modelo de aprendizado de máquina eficaz, é geralmente necessária uma grande quantidade de dados armazenados. A qualidade e a quantidade dos dados são fatores cruciais para a performance do modelo. Esses dados são comumente armazenados em bancos de dados estruturados, devido à sua conformidade com os princípios ACID (Atomicidade, Consistência, Isolamento e Durabilidade) (Garcia, Sotto, 2019):

a) atomicidade: em uma transação que envolve duas ou mais partes de informações discretas, a transação será executada completamente ou não será executada. Isso garante que, em caso de falha, não haja estados intermediários, mantendo a integridade dos dados;
b) consistência: a transação cria um estado válido dos dados ou, em caso de falha, retorna todos os dados ao estado anterior ao início da transação. Isso assegura que os dados no banco permaneçam corretos e válidos;
c) isolamento: uma transação em andamento, mas ainda não validada, deve permanecer isolada de qualquer outra operação. Isso significa que as transações concorrentes não interferem umas nas outras, garantindo a integridade e a consistência dos dados durante operações simultâneas;
d) durabilidade: dados validados são registrados pelo sistema de tal forma que, mesmo no caso de uma falha ou reinício do sistema, os dados estejam disponíveis em seu estado correto. Isso é vital para garantir que os dados permaneçam persistentes e não sejam perdidos.

O gerenciamento de dados envolve a definição de estruturas para o armazenamento de informações e fornece mecanismos para a manipulação dessas informações (Silberschatz; Korth; Sudarshan, 2006). Este gerenciamento eficiente é fundamental para a mineração de dados, pois facilita o acesso às informações e simplifica seu uso. A estrutura organizada dos bancos de dados permite estabelecer

relacionamentos entre diferentes registros e aplicações, que podem ser facilmente acessados e combinados para análises mais complexas (Matsumoto, 2006).

Para os autores, um banco de dados que seja útil não só armazena dados de forma organizada, mas também garante a integridade e o acesso rápido à informação. Esse aspecto é particularmente importante em contextos de aprendizado de máquina, onde a eficiência no pré-processamento dos dados pode economizar tempo significativo. Dados bem estruturados permitem que algoritmos de aprendizado de máquina realizem análises mais rápidas e precisas, levando a insights mais profundos e decisões mais informadas.

Além disso, a integração dos princípios ACID nos bancos de dados assegura que a manipulação dos dados seja confiável e robusta. Isso é essencial em ambientes de aprendizado de máquina, onde grandes volumes de dados são processados continuamente, e a consistência e a durabilidade dos dados são cruciais para o treinamento de modelos precisos e confiáveis.

Portanto, a escolha de um banco de dados adequado, que siga os princípios ACID, é um passo fundamental para garantir o sucesso no treinamento de modelos de aprendizado de máquina. A eficiência no armazenamento e no acesso aos dados não só melhora a qualidade do modelo, mas também otimiza o tempo e os recursos necessários para seu desenvolvimento.

2.3 APRENDIZAGEM DE MÁQUINA

Aprendizagem de Máquina, ou aprendizado de máquina (ML), é um campo da inteligência artificial que desenvolve algoritmos e técnicas para que os sistemas computacionais possam aprender a partir de dados, identificar padrões e tomar decisões com o mínimo de intervenção humana. O objetivo é criar modelos que possam generalizar comportamentos a partir de exemplos e realizar tarefas específicas de maneira autônoma. De acordo com Monard e Baranauskas (2003), o aprendizado de máquina é uma área da inteligência artificial que tem como objetivo desenvolver técnicas computacionais sobre o aprendizado, além de possibilitar a criação de sistemas capazes de adquirir conhecimento de maneira automática.

Cada modelo de aprendizado de máquina é único, ou seja, cada um possui suas características próprias, com cada uma com seu treinamento, uma máquina que aprendeu de X maneira, dificilmente vai funcionar em um ambiente Y. A

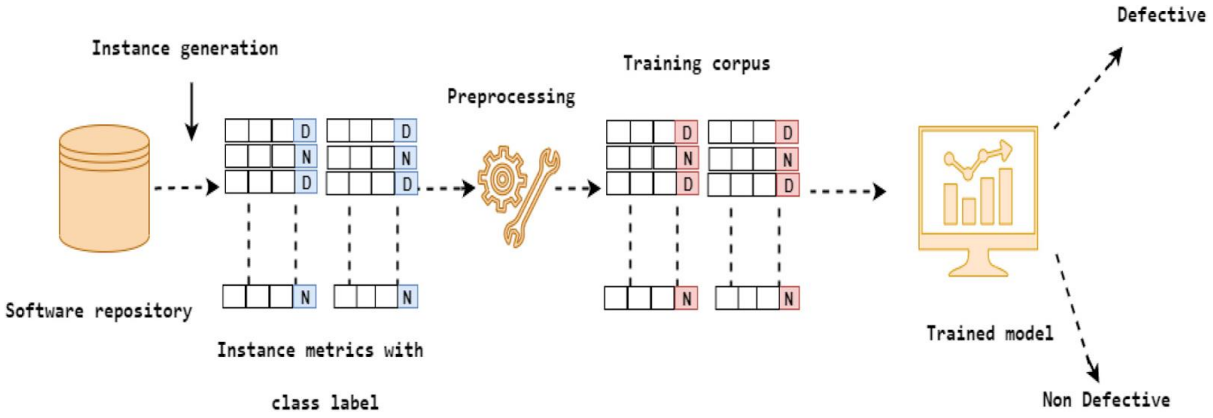
utilização de objetos de aprendizagem neste âmbito tende a aperfeiçoar o estilo do ensino educacional através, por exemplo, de softwares que “acompanhem” o nível de conhecimento do aprendiz, inserindo novos conteúdos à medida que este aprende (Beserra et al., 2014).

FIGURA 3 – Um sistema para processamento de imagens

FONTE: FILHO; NETO, 1999

Conforme na FIGURA 3, o ML é um processo contínuo e várias repetições conforme o modelo conhece mais os dados, cada vez mais vai acertar os seus resultados. Para dar um exemplo melhor de como é o treinamento de um modelo na FIGURA 4, é apresentado um modelo para classificar se é defeituosa ou não.

FIGURA 4 – Basic structure of SDP model



FONTE: PANDEY (2020)

Dois dos principais desafios de um modelo de ML é o de combater o*verfitting* e *underfitting*, O *overfitting* ocorre quando um modelo aprende os detalhes e o ruído nos dados de treinamento a ponto de afetar negativamente o desempenho do modelo em novos dados (dados de teste). Isso significa que o ruído ou as flutuações aleatórias nos dados de treinamento são captados e aprendidos como conceitos pelo modelo. O problema é que esses conceitos aprendidos não se aplicam a novos conjuntos de dados e afetam negativamente a capacidade de generalização dos modelos (Pandey; Rathee; Triapathi, 2020). *Underfitting* ocorre quando um modelo é muito simples para capturar os padrões subjacentes nos dados de treinamento. Isso resulta em baixa precisão tanto nos dados de treinamento quanto nos dados de teste. O *underfitting* geralmente acontece quando o modelo escolhido ou a quantidade de treinamento é insuficiente para representar a complexidade dos dados, um modelo é *underfitting* quando generaliza demais o comportamento do exemplo no registro, ou seja, o modelo permite comportamentos muito diferentes do que foi visto no registro (Fahland, Aalst, 2013).

2.3.1 Tipos de aprendizado

No aprendizado de máquina, são definidos em quatro tipos, nesta seção vão ser definidos cada modelo, o supervisionado, o não supervisionado, o semi-supervisionado e aprendizagem por reforço.

A aprendizagem de máquina supervisionada é um ramo do aprendizado de máquina que utiliza conjuntos de dados rotulados para treinar algoritmos. Esses conjuntos de dados incluem entradas (variáveis independentes) e saídas (variáveis dependentes), permitindo que os modelos aprendam a realizar tarefas específicas. À medida que os dados de entrada são inseridos no modelo, ele ajusta seus pesos até que o ajuste seja adequado, geralmente como parte do processo de validação cruzada (IBM, 2023).

A aprendizagem não supervisionada visa identificar padrões e estruturas nos dados sem a necessidade de um conjunto de treinamento rotulado. Em vez disso, o modelo é alimentado com um conjunto de dados não rotulados e é deixado por conta própria para encontrar padrões e estruturas significativas.

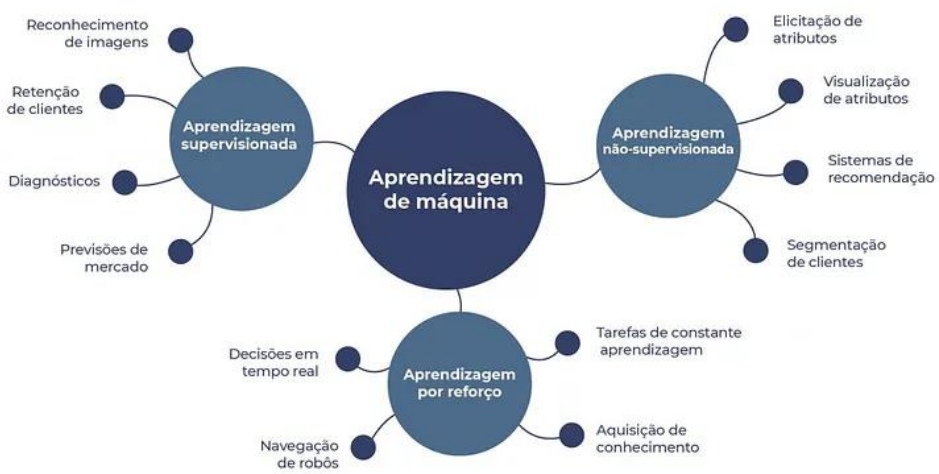
O aprendizado de máquina semi-supervisionado é uma abordagem que combina elementos do aprendizado supervisionado e não supervisionado. Nesse método, o conjunto de dados usado para treinamento inclui tanto exemplos rotulados quanto não rotulados. Em outras palavras, o aprendizado semi-supervisionado é especialmente útil quando há poucos rótulos disponíveis para os dados. Por exemplo, imagine um conjunto de pontos em um gráfico, onde alguns pontos estão rotulados (representados por cores) e outros não. O objetivo é utilizar esses dados parcialmente rotulados para treinar um modelo de inteligência artificial. Essa abordagem é valiosa quando a rotulação completa dos dados é custosa ou inviável. Ela permite que o algoritmo aprenda com os exemplos rotulados e, assim, possa generalizar seu conhecimento para os exemplos não rotulados. Isso ajuda a maximizar o uso de todos os dados disponíveis, melhorando a eficácia do modelo (Almeida, 2023).

O aprendizado por reforço é um tipo de algoritmo de aprendizado de máquina que permite a um agente aprender a tomar decisões em um ambiente desconhecido por meio de tentativa e erro. O agente recebe recompensas ou penalidades por suas ações e, ao longo do tempo, aprende a tomar decisões que maximizem suas recompensas. Em termos simples, no aprendizado por reforço, o sistema de inteligência artificial enfrenta uma situação e usa tentativa e erro para encontrar uma solução para o problema. O objetivo é maximizar a recompensa total. O cientista de dados ou engenheiro de IA define as regras do jogo (política de recompensa), mas o modelo deve descobrir como executar a tarefa para maximizar a recompensa, começando com ações aleatórias e evoluindo para táticas mais

sofisticadas. Essa abordagem é amplamente utilizada em áreas como jogos e robótica e é a técnica por trás do Alpha Go (Data Science Academy, 2022).

Para representar, na FIGURA 5 é mostrado como cada ramo do aprendizado de máquina pode ser usado, para essa pesquisa será focado no aprendizado supervisionado.

FIGURA 5 – Tipos de aprendizado de máquina



FONTE: Maia (2020).

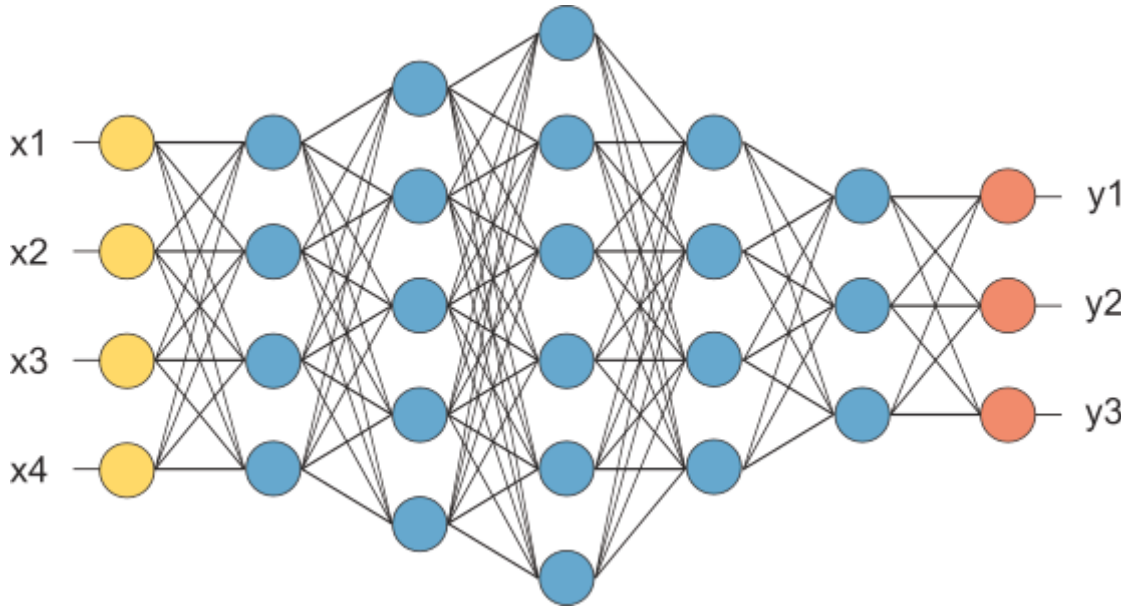
2.3.2 Algoritmos de Aprendizagem de Máquina

Para se aplicar o aprendizado de máquina em uma base de dados, é utilizado os métodos para isso, dentro desta tecnologia existem uma diversidade muito grande, não existe a frase “use esse que é o melhor” cada um desses métodos são usadas para realizar tarefas específicas, nesta secção vão ser apresentados os modelos mais utilizados e suas definições.

Um dos propósitos do ML é tentar fazer que a máquina pense como um humano, a rede neural artificial (RNA) é o que chama mais perto de representar isso, desde o aprendizado a os neurônios, a Rede Neural Artificial (RNA) é uma técnica computacional que constrói modelo matemático inspirado em cérebro humano para reconhecimento de imagens e sons, com capacidade de aprendizado, generalização, associação e abstração, constituído por sistemas paralelos distribuídos em compostos de unidades simples de processamento (Galvão; Marin,

2009). RNA aprende padrões, procura relacionamentos e constrói modelos automaticamente. Na FIGURA 6 apresenta como é o processo de aprendizado, o dado entrado é processado e tem um resultado.

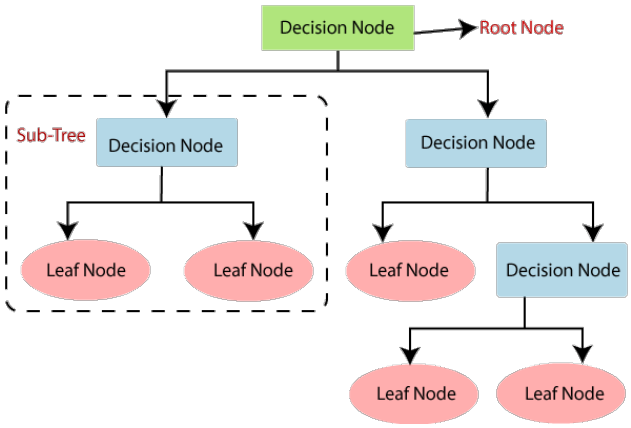
FIGURA 6 – rede neural artificial



FONTE: Coelho (2017).

Outro método muito utilizado como entrada para o ML é a árvore de decisão, muito utilizado no aprendizado de mineração de dados, é um método de aprendizado supervisionado, não paramétrico, usado para classificação e regressão. O objetivo é criar um modelo que preveja o valor de uma variável, aprendendo regras de decisão simples inferidas a partir dos dados (Pedregosa et al., 2011). A árvore de decisão é um modelo representado graficamente por nós e galhos, parecido com uma árvore, mas no sentido invertido. Desse modelo surge o *Random Forest* que um modelo que usa a árvore de decisão, mas gera uma “floresta” que representa um conjunto de muitas árvores de decisão individuais (Chakrabarty et al., 2019). Como representado na FIGURA 7.

FIGURA 7 – Decision Tree Algorithm



FONTE: javatpoint (2016).

O método Naive Bayes é um dos modelos que mais precisam de dados para se ter um sucesso satisfatório pois o algoritmo altamente escalonável, exigindo um alto número de variáveis lineares (preditores) em um problema de aprendizagem. Calculando a probabilidade condicional de cada atributo seguido de uma aplicação contida do teorema de Bayes, para determinar a probabilidade relativa sobre as características dos atributos, no sentido da previsão do resultado (Aggarwal, 2014).

Outro modelo muito utilizado quando se quer tentar encontrar um valor, por exemplo, o valor de um carro com vários dados de cor, ano e cavalos é o de *Logistic Regression*, modelo baseado em regressão logística tem objetivo de criar relação de dependência direta entre a variável de classe e as características, buscando trazer valores compreendidos entre 0 e 1, valores estes que representam a probabilidade de retornar o valor 1 (Andrade et al., 2023).

2.4 MOBILIDADE URBANA

Com cada vez mais movimentação dentro das grandes cidades, o tema da mobilidade urbana vem ganhando mais força, mas o que é essa mobilidade? Mobilidade urbana refere-se à capacidade de movimentação de pessoas e bens dentro de um espaço urbano. Envolve a interação entre diversos modos de transporte, infraestrutura urbana e políticas públicas, visando facilitar o deslocamento eficiente e seguro dentro das cidades. Os problemas de mobilidade urbana estão ligados, entre outros fatores, ao fato de as cidades terem crescido

desordenadamente, sem planejamento de infraestrutura, desde transporte público até o tráfego (Santiago; Peixoto, 2019). De acordo com Marin (2015, p.248), “Com cada vez as cidades crescendo o percurso para ir nos lugares vem aumentando cada vez mais, longas distâncias a serem percorridas para se ter acesso ao trabalho, lazer, comércio, serviços públicos, dentre outros. Deste modo, é necessário se fazer pensar as cidades.”

Uma mobilidade urbana eficiente promove a economia ao reduzir os custos de transporte e aumentar a produtividade. Além disso, tem implicações ambientais significativas, contribuindo para a redução das emissões de gases de efeito estufa e melhorando a saúde pública ao diminuir a poluição do ar.

Outro fator da importância da mobilidade urbana é o que é representado além do tempo de percurso, A qualidade da mobilidade urbana é um importante termômetro que permite compreender o processo de urbanização e as dinâmicas socioespaciais de produção do espaço (Rodrigues, 2022).

> A mobilidade urbana é um atributo das cidades e se refere à facilidade de deslocamentos de pessoas e bens no espaço urbano [...]pensar a mobilidade urbana é, portanto, pensar sobre como se organizam os usos e a ocupação da cidade e a melhor forma de garantir o acesso das pessoas e bens ao que a cidade oferece (locais de emprego, escolas, hospitais, praças e áreas de lazer) não apenas pensar os meios de transporte e o trânsito.(MINISTÉRIO DAS CIDADES, 2005, p.4)

Souza e Belo (2019) argumentam que mobilidade urbana eficiente é fundamental para o crescimento sustentável das cidades e a garantia do exercício pleno dos direitos constitucionais. Principalmente, na questão da inclusão social urbana e a acessibilidade de pessoas com necessidades especiais, pois todos os cidadãos possuem direito de se deslocar de forma adequada e digna.

Os principais desafios para uma boa mobilidade urbana é o congestionamento com tempos de viagem mais longos, aumento do consumo de combustível e estresse para os motoristas. Os fatores que levaram a isso é a baixa qualidade dos transportes coletivos e adoção majoritária do transporte individual resultam em um cenário de congestionamentos, privatização do espaço público (Silveira, 2010).

Outro ponto de relevância é o de infraestrutura das cidades a infraestrutura urbana inadequada pode limitar a eficiência dos sistemas de transporte. Isso inclui a

falta de vias adequadas, ciclovias, calçadas seguras e pontos de acesso ao transporte público. De acordo com Mirando e Cascaes (2013, p.130), “A infraestrutura viária que dá suporte ao transporte a pé nos grandes centros urbanos é constituída por espaços públicos que apoiam a caminhada. Portanto, deve possuir o mesmo nível de qualidade requerido para o transporte motorizado.”

Muito se fala das cidades inteligentes, esse ponto pode servir como base para uma mobilidade urbana mais sustentável, o uso de tecnologias inteligentes pode transformar a mobilidade urbana, tornando-a mais eficiente e sustentável. Em 2006, o *Office of science and technology* fez um relatório onde foi apresentado diretrizes que em relação ao contexto urbano, dentre elas existem duas que ao combinar com o aprendizado de máquina é possível criar um modelo que melhore a mobilidade urbana, são esses:

- sistemas inteligentes: provêm por meio de sensores e data mining, informações para embasar decisões dos cidadãos, planejadores, e prestadores de serviços de transporte.
- infraestrutura inteligente: capaz de processar vastas gamas de informações coletadas por múltiplos sensores, adaptando as dinâmicas das redes de transporte, em tempo real, para proporcionar serviços mais eficientes.

Fica claro que cada vez mais o uso das geotecnologias para a mobilidade urbana pode melhorar o fluxo e trazer uma mobilidade inteligente.

> É perceptível a grande importância das TICs e Geotecnologias no que diz respeito à mobilidade inteligente, todavia, destaca-se que digital não é sinônimo de inteligente. O cerne da mobilidade inteligente está intrinsecamente relacionado ao conceito de mobilidade sustentável, uso do solo, planejamento urbano, e das mudanças culturais da população quanto aos modos de transporte (Menzori; Gonçalves, 2003, p.15)

Portanto, a mobilidade inteligente transcende o simples uso de TICs e Geotecnologias, demandando uma abordagem integrada que contemple a sustentabilidade, o planejamento urbano eficaz e a transformação cultural nos hábitos de transporte. Apesar do potencial das tecnologias digitais em otimizar sistemas de transporte, sua eficácia está condicionada ao alinhamento com estratégias que promovam o uso racional do solo, a redução de impactos ambientais e a adaptação aos desafios específicos das cidades. Assim, a verdadeira

inteligência na mobilidade reside na capacidade de harmonizar avanços tecnológicos com soluções que priorizem o bem-estar coletivo e a preservação ambiental.

3 CARACTERIZAÇÃO E PROCEDIMENTOS METODOLÓGICOS

Neste capítulo, são apresentados a caracterização da pesquisa e os procedimentos metodológicos realizados para alcançar os objetivos propostos, incluindo as etapas de coleta, análise e concepção das bases de dados.

3.1 CARACTERIZAÇÃO DA PESQUISA

Esta pesquisa pode ser classificada como exploratória e descritiva. Segundo Gil (2012), a pesquisa exploratória tem como meta compreender um problema para torná-lo explícito. A pesquisa descritiva expõe as características de uma determinada população. Segundo Malhotra (2001), a pesquisa descritiva “tem como principal objetivo a descrição de algo”, um evento, um fenômeno ou um fato.

Como abordagem, pode ser considerada qualitativa e quantitativa pois a pesquisa se caracteriza como um esforço cuidadoso para a descoberta de novas informações para a verificação e ampliação do conhecimento existente (Godoy, 1995). A abordagem qualitativa se encontra em variados tipos de investigações podendo ser realizada por meio da leitura de livros, obras de referência e artigos de pesquisa na área que está sendo realizada desde o início deste estudo (Godoy, 1995), em relação a quantitativa que os dados são normalmente apresentados em forma numérica e analisados com o emprego de técnicas matemáticas e estatísticas, a exemplo das utilizadas pelos métodos de aprendizado de máquina. Godoy (1995) explica que este tipo de estudo se preocupa com a medição objetiva e a quantificação dos resultados, buscando a precisão, evitando distorções na etapa de análise e interpretação dos dados, garantindo assim uma margem de segurança em relação às inferências obtidas.

3.2 COLETA DE DADOS E ARMAZENAMENTO

A coleta de dados consiste em coletar os dados e selecionar os dados das fontes escolhidas para então poder armazená-los em uma base de dados.

Nessa etapa os dados serão coletados da seguinte forma: os dados serão coletados de várias fontes ao redor do mundo. Isso inclui a identificação de bases de

dados relevantes que contenham informações sobre mobilidade urbana, tais como tráfego, transporte público e condições das vias. Para armazenar esses dados, vai ser estabelecido uma pipeline de pré-processamento dos dados, balanceamento, seleção de colunas e integração à base de dados artificiais final.

3.2.1 Identificação das fontes de dados

Nessa seção será onde será retirada a base que será usada para a utilização final.

Para encontrar a base de dado serão usados os repositórios de dados:

- *Kaggle*: É uma plataforma online para ciência de dados e aprendizado de máquina. Oferece competições com prêmios, *datasets* gratuitos, notebooks Jupyter na nuvem, e cursos de aprendizado de máquina. É uma comunidade colaborativa para cientistas de dados, adquirida pelo Google em 2017.
- UCI *Machine Learning Repository*: a plataforma tem uma coleção de datasets amplamente utilizados na prática de aprendizado de máquina e pesquisa. Fundado em 1987 e hospedado pela Universidade da Califórnia, Irvine, ele serve como um recurso gratuito e valioso para cientistas de dados, pesquisadores e estudantes. A plataforma oferece conjuntos de dados variados que cobrem uma ampla gama de domínios, como biologia, medicina, economia e muitos outros. Cada *dataset* vem com uma descrição detalhada e é frequentemente usado em publicações acadêmicas e competições. Além disso, o repositório facilita a comparação de algoritmos de aprendizado de máquina ao fornecer uma base comum de dados. Ele é amplamente reconhecido e respeitado na comunidade de ciência de dados por sua contribuição contínua à pesquisa e ao desenvolvimento de novas técnicas de análise de dados.
- *Google Dataset Search*: É uma ferramenta de busca especializada para encontrar *datasets* online. Lançada pelo Google, ela permite aos usuários localizar conjuntos de dados públicos disponíveis na web, independentemente do provedor. A ferramenta indexa

datasets de diversas áreas, como governo, educação, pesquisa científica e muito mais. Os usuários podem filtrar resultados por tipo de arquivo, assunto e outras características. É uma ferramenta valiosa para pesquisadores, cientistas de dados e qualquer pessoa que precise de dados para análise. Além disso, promove a transparência e a acessibilidade de dados abertos.

- *Smart Dublin*: é uma iniciativa da cidade de Dublin, na Irlanda, focada em usar tecnologia e inovação para resolver desafios urbanos e melhorar a qualidade de vida. Lançada em 2016, envolve a colaboração entre o governo, empresas, universidades e cidadãos. O programa promove projetos em áreas como mobilidade, sustentabilidade, segurança pública e economia digital. Utiliza dados e tecnologia para criar soluções inteligentes, como sistemas de transporte eficientes e gestão de resíduos. Smart Dublin também incentiva startups e empresas a testar novas tecnologias na cidade, transformando Dublin em um laboratório vivo para inovações urbanas.

3.2.2 Critério de seleção

Nesta seção serão definidos os critérios para a seleção das bases de dados, levando em consideração sua relevância, qualidade e disponibilidade:

Relevância: A prioridade será dada a bases de dados que possuam colunas significativas, capazes de contribuir substancialmente para a construção da base de dados final.

Qualidade: bases de dados com alta integridade, minimizando a presença de dados faltantes, com um limite máximo de 10% de ausências. Isso garantirá a robustez e confiabilidade das informações.

Cobertura temporal: As bases selecionadas devem abranger um período abrangente, compreendido entre os anos de 2015 e 2022, assegurando que os dados estejam atualizados e reflitam as tendências recentes.

Cobertura geográfica: O foco será em dados provenientes de cidades com grande concentração de veículos, como grandes metrópoles e áreas urbanas

densamente povoadas. Isso permitirá uma análise mais abrangente e representativa das condições locais.

Será realizada uma busca sistemática nos repositórios Kaggle, UCI, Google Dataset Search, Smart Dublin com as palavras-chaves definidas como:

- Transporte Público / Public Transport;
- Mobilidade Urbana / Urban Mobility;
- Tráfego Urbano / Urban Traffic;
- Transporte Coletivo / Collective Transport;
- Sistemas de Transporte / Transportation Systems;
- Dados de Tráfego / Traffic Data;
- Horários de Ônibus / Bus Timetables;
- Estações de Metrô / Subway Stations;
- Bicicletas Compartilhadas / Bike Sharing;
- Veículos Autônomos / Autonomous Vehicles;
- Cidades Inteligentes / Smart Cities;
- Dados de GPS / GPS Data;
- Dados de Localização / Location Data.

O processo de filtragem das bases de dados terá foco na mobilidade urbana, considerando critérios como tempo de trajeto, meio de transporte, horário de viagem, entre outros aspectos relevantes.

3.3 PRÉ-PROCESSAMENTO DE DADOS

Nessa seção será definido o processo de limpeza de cada base de dados escolhida, isso inclui a higienização, transformação, normalização e balanceamento dos dados para garantir que estejam em um formato consistente e utilizável.

a) higienização dos dados: nesse processo terá como objetivo o tratamento dos dados ausentes, dados duplicados, e possíveis dados inconsistentes;
b) transformação de dados: a parte de transformação vai ser útil para encontrar possíveis novas colunas de informação e converter aquelas que ajudem o modelo a ser mais assertivo, por exemplo, com colunas de tempo de viagem e distância é possível calcular a

velocidade média, ou criar índices de mobilidade de determinada região, como tempo de trajeto, número de viagens, combinando tudo isso é possível criar um score de região;

c) normalização: outra parte de tratamento de dados que contempla a padronização das escalas e conversão de unidades, como tempo e distância;
d) balanceamento: para lidar com o balanceamento de classes na base de dados utilizada, será aplicada a técnica SMOTE (*Synthetic Minority Over-sampling Technique*), que cria instâncias sintéticas para aumentar o volume das classes minoritárias de forma realista. Essa abordagem melhora a representatividade dos dados e reduz o viés nos modelos de aprendizado de máquina, garantindo predições mais precisas. Contudo, será necessário avaliar o impacto computacional do processo, especialmente em bases maiores ou com alta dimensionalidade.

3.4 SELEÇÃO DA BASE DE DADOS

O objetivo deste estudo é explorar a aplicação de técnicas de aprendizado de máquina para a previsão de volume de tráfego urbano utilizando dados sintéticos. A escolha da base de dados é guiada por critérios que assegurem sua adequação às necessidades do estudo e permitam testar a viabilidade de soluções preditivas em um ambiente controlado.

O uso de dados sintéticos é justificado por suas diversas vantagens, como o maior controle sobre a criação de cenários e a possibilidade de demonstrar aos gestores que dados artificiais, quando projetados para representar a realidade, podem ser eficazmente utilizados para aplicar aprendizado de máquina em melhorias na mobilidade urbana. Além disso, esses dados possibilitam a realização de experimentos que simulam padrões complexos, proporcionando um ambiente ideal para o treinamento e a validação de modelos preditivos.

Para a seleção da base, será considerado um conjunto de critérios que inclui: a presença de variáveis que representem diferentes dimensões da mobilidade urbana (como fluxo de tráfego, uso de transporte público e condições climáticas), estruturação temporal para análise histórica, e a representatividade dos dados,

garantindo cenários realistas. A base deve também apresentar qualidade, com integridade e coerência nas informações.

O objetivo principal é assegurar que a base forneça subsídios suficientes para treinar e testar modelos preditivos, além de permitir a análise das relações entre fatores que influenciam a mobilidade urbana. Assim, será possível demonstrar que dados sintéticos são eficazes para resolver problemas práticos, como a previsão de padrões de tráfego e otimização de sistemas de transporte.

Com base nesses critérios, realiza-se uma pesquisa para identificar para identificar um base sintético existente que atenda às necessidades do estudo. Caso não seja encontrada uma base adequada, será avaliada a possibilidade de desenvolver uma base personalizada para atender aos objetivos propostos.

3.5 TÉCNICAS DE APRENDIZADO DE MÁQUINA

Nesta seção são apresentados os modelos utilizados, bem como os passos seguidos para sua aplicação, incluindo a divisão da base, treinamento, ajuste de hiperparâmetros, análise de resultados e discussão das implicações.

Seleção dos algoritmos de aprendizado de máquina: Conforme discutido no referencial teórico, os modelos selecionados são *Random Forest Classifier*, KNN e RNA.

Divisão da base criada: a base de dados é segmentada em conjuntos de treinamento e teste. O conjunto de treinamento é utilizado para ensinar os modelos a reconhecerem padrões, enquanto o conjunto de teste avalia seu desempenho em dados não vistos anteriormente. Essa abordagem é essencial para evitar o *overfitting* e garantir a generalização dos resultados.

Antes da avaliação, realiza-se um processo de treinamento para identificar os hiperparâmetros mais adequados ao conjunto de dados. Diversos cenários são testados para otimizar o desempenho dos modelos.

Os modelos são treinados utilizando o conjunto de treinamento, ajustando seus parâmetros internos para minimizar os erros de predição. Posteriormente, são avaliados com o conjunto de teste, utilizando métricas como acurácia, precisão, *recall*, F1-score e erro quadrático médio. Os resultados são examinados para identificar padrões, variáveis mais relevantes e possíveis ajustes necessários na base de dados. Ferramentas de visualização auxiliam na identificação de insights

significativos. Usando as ferramentas de visualização de dados será possível encontrar quais variáveis foram mais importantes.

Finalmente, os resultados são discutidos em termos de suas implicações para a melhoria da mobilidade urbana.

3.6 FERRAMENTAS E TECNOLOGIAS UTILIZADAS

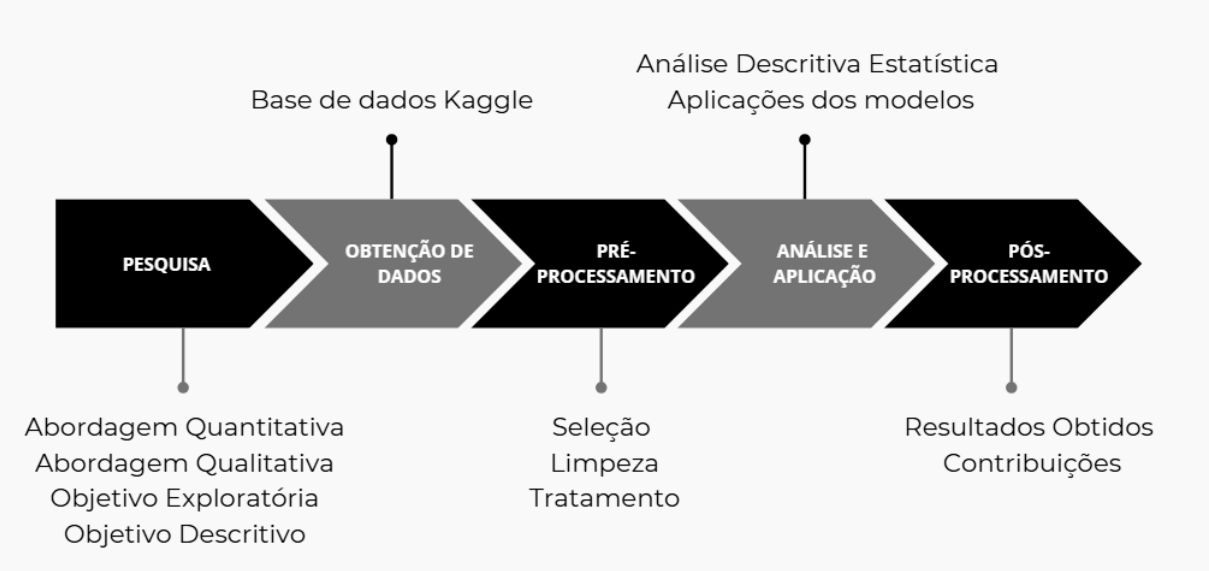
Para essa pesquisa serão empregadas ferramentas específicas para análise de dados, cada uma com um papel fundamental em diferentes etapas do processo. As ferramentas selecionadas incluem:

A ferramenta principal na análise das bases de dados e tratamentos será utilizado o Python pois é a linguagem principal para manipulação de dados, segundo Tiobe (2024), desenvolvimento de modelos de machine learning e visualização de dados. Com o Python é possível ter acesso a uma infinidade de bibliotecas e frameworks com foco no aprendizado de máquina, dentre elas o Pandas, biblioteca fundamental para a manipulação e análise de dados em Python. Facilita a leitura, limpeza e transformação de grandes conjuntos de dados. Num-py que fornece suporte para *arrays* multidimensionais e funções matemáticas de alto desempenho, essenciais para cálculos numéricos em Python. Scikit-learn que vai ser a principal ferramenta pois é uma biblioteca robusta para machine learning em Python, oferecendo ferramentas para classificação, regressão, *clustering* e redução de dimensionalidade. E o PyTorch, é um *framework* de código aberto para machine learning e inteligência artificial. É utilizado principalmente para a construção e treinamento de redes neurais profundas.

Para visualização dos dados vai ser usado o Matplotlib que é a biblioteca de visualização em Python permite a criação de gráficos estáticos, animados e interativos, sendo altamente customizável para diversos tipos de visualização de dados. Em conjunto com o Seaborn que oferece uma interface de alto nível para a criação de gráficos estatísticos atraentes e informativos.

Para representar como será o fluxo dessa pesquisa, foi criado um fluxograma, como pode ser observado na FIGURA 8.

FIGURA 8 – Fluxograma da pesquisa



FONTE: O autor (2024).

Essa figura representa todo o fluxo do início da pesquisa e contempla desde a concepção até os resultados e contribuições.

4 RESULTADOS

Os resultados dessa pesquisa serão apresentados da seguinte forma: apresentação do *dataset* escolhido, pré-processamento, análise exploratória dos dados, aplicação dos modelos, verificação dos resultados.

4.1 CONTEXTUALIZAÇÃO DA BASE DE DADOS

O Urban Mobility Dataset, disponível no Kaggle, foi desenvolvido para representar padrões de mobilidade urbana por meio de dados sintéticos, com o objetivo de oferecer uma base abrangente e controlada para análise. O *dataset* inclui variáveis relacionadas ao uso de transporte público, fluxo de tráfego, programas de compartilhamento de bicicletas, movimento de pedestres, além de fatores contextuais como condições climáticas, eventos e feriados do mundo. Essa diversidade de informações permite a modelagem de cenários complexos, atendendo a demandas como previsão de congestionamento e otimização de sistemas de transporte.

A escolha por dados sintéticos justifica-se por sua capacidade de superar limitações de dados reais, como restrições de custo de obtenção e dificuldade de garantir completude e representatividade. Nesse contexto, o Urban Mobility Dataset foi selecionado devido à sua abordagem sistemática e riqueza de detalhes, que simulam situações urbanas de forma confiável e alinhada aos objetivos do estudo, como prever o volume de tráfego e aprimorar estratégias de planejamento urbano.

Os dados da base estão segmentados da seguinte forma, com 15 (quinze) colunas divididas em variáveis temporais: *timestamp, day of week, holiday*. Fatores de mobilidade: *traffic flow, public transport usage, bike sharing usage*. Fatores contextuais: *weather conditions, temperature, events*. Métricas de desempenho: *public transport delay, road incidents.*

4.2 FORMATAÇÃO DA BASE DE DADOS

Esta seção apresenta os resultados das etapas de tratamento realizadas na base de dados, incluindo a segmentação dos dados, os critérios de seleção das variáveis e a definição do atributo alvo da análise.

Por se tratar de uma base artificial, os dados foram originalmente gerados para o período entre 2022 e 2137, visando cobrir cenários futuros amplos. No entanto, para representar de forma mais realista as condições atuais de mobilidade urbana, foi realizada uma subdivisão da base, restringindo-o ao intervalo de 1º de janeiro de 2023 a 31 de dezembro de 2023.

O atributo alvo selecionado foi a coluna *traffic flow*, originalmente numérica, que foi convertida em categorias para aumentar a precisão de análise e interpretação. A categorização seguiu a seguinte lógica:

Valores entre 0 e 500 foram classificados como "baixo".

Valores de 501 a 1000 foram classificados como "médio".

Valores de 1001 a 2000 foram classificados como "alto".

Valores acima de 2000 foram classificados como "intenso".

Adicionalmente, a coluna *timestamp*, no formato "2023-01-01 00:00:00", foi desmembrada em 1(uma) nova variável: hora. Essa transformação teve como objetivo enriquecer os dados de entrada, proporcionando maior granularidade e contribuindo para melhorar o desempenho dos modelos de aprendizado de máquina.

Para a coluna *public transport usage* foi feita uma engenharia de dados para transformar de numérica para categórica. A categorização seguiu a seguinte lógica:

Valores entre 0 e 200 foram classificados como "baixo".

Valores entre 201 e 350 foram classificados como "médio".

Valores entre 351 e 499 foram classificados como "alto".

O mesmo foi feito para a coluna *bike sharing usage* que seus valores eram distribuídos de maneira semelhante à de transporte público.

Para a coluna de temperatura foi dividida da seguinte forma a sua categorização:

Valores entre -10 e 5 foram classificados como "frio".

Valores entre 6 e 20 foram classificados como "morno".

Valores entre 21 e 30 foram classificados como "quente".

E para a coluna *public transport delay* sua categorização foi feita desta maneira:

Valores entre 0 e 10 foram classificados como "normal".

Valores entre 11 e 20 foram classificados como "médio".

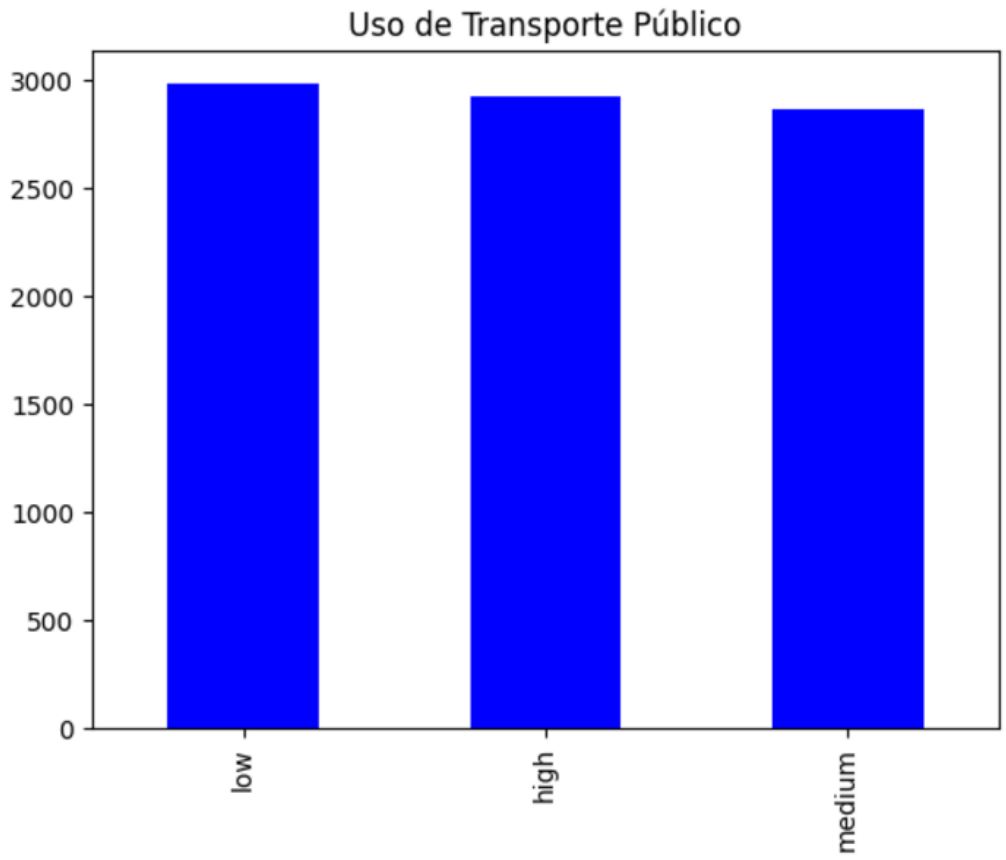
Valores entre 21 e 30 foram classificados como "alto".

4.3 ANÁLISE ESTATÍSTICA DA BASE

Essa seção trará os resultados estatísticos da base usada. Focando em mostrar como estão divididos os dados categóricos e sua relação com o atributo meta.

Após a criação do intervalo, a base de dados resultou em 8.759 linhas e 15 colunas. A primeira coluna analisada foi a de uso do transporte público, categorizada previamente. Os dados foram distribuídos da seguinte forma: 2.983 registros classificados como "baixo", 2.858 como "médio" e 2.918 como "alto". O GRÁFICO 1 apresenta a distribuição total dos dados dessa coluna.

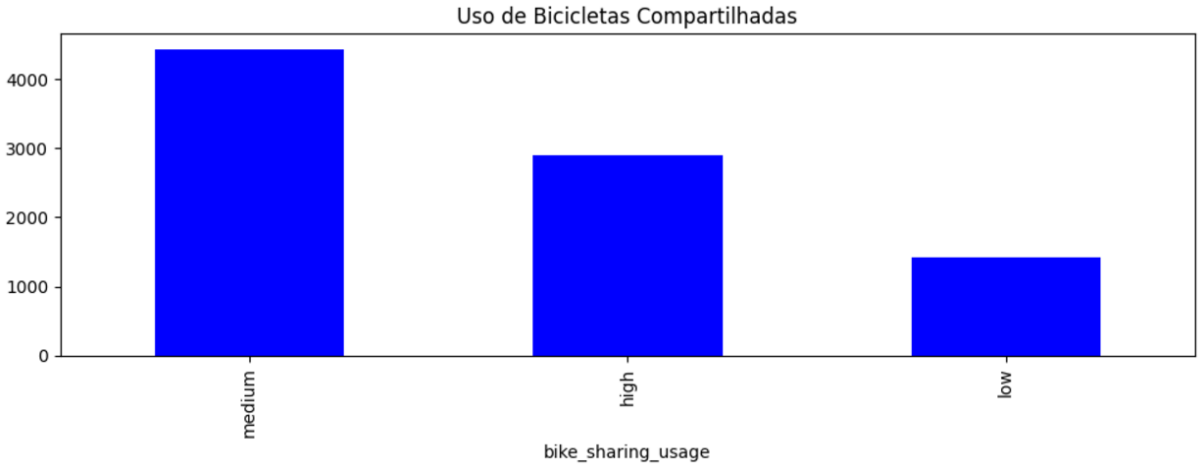
GRÁFICO 1 – Estatísticas dos dados de transporte público



FONTE: Autor (2024)

A coluna relacionada ao uso de bicicletas compartilhadas foi categorizada em três níveis: 1.428 registros classificados como "baixo", 4.438 como "médio" e 2.893 como "alto". Esses resultados estão representados no GRÁFICO 2.

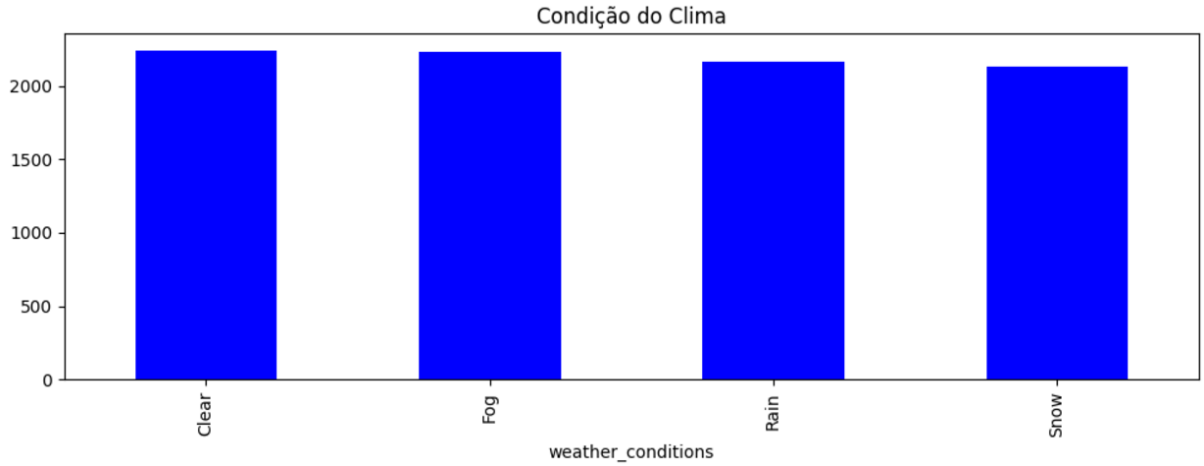
GRÁFICO 2 – Estatísticas dos dados do uso de bicicletas compartilhadas



FONTE: Autor (2024)

A coluna de condição do clima foi a que mais se manteve melhor distribuída entre as variáveis com 2.242 para tempo limpo, 2.229 para neblina, 2.160 para tempo chuvoso e 2.128 para tempo com neve. Sua representação pode ser observada no GRÁFICO 3.

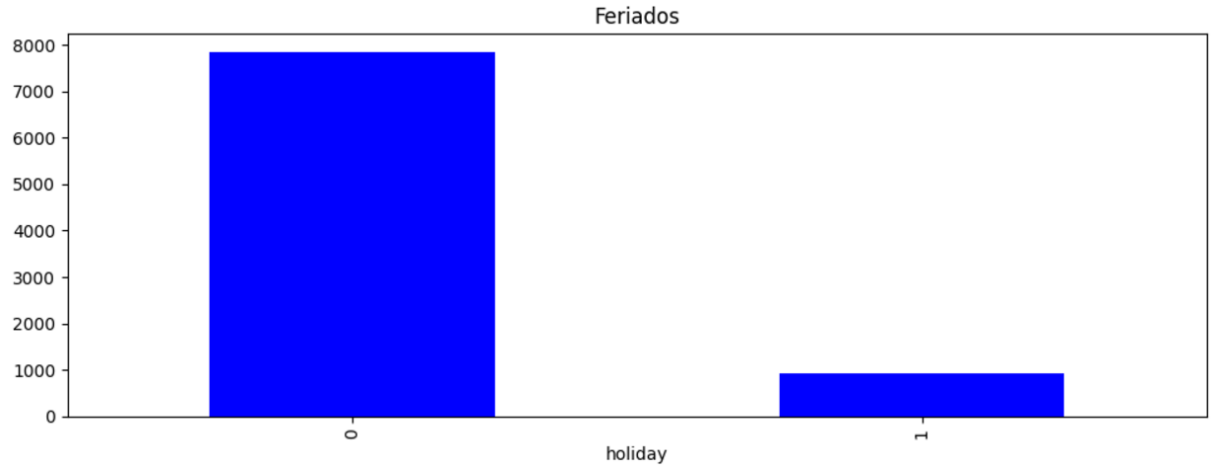
GRÁFICO 3– Estatísticas dos dados do clima



FONTE: Autor (2024)

Para o GRÁFICO 4 os dados de feriados ficaram distribuídos da seguinte forma: 7.839 para “Não é feriado” e 920 para “Sim é feriado”.

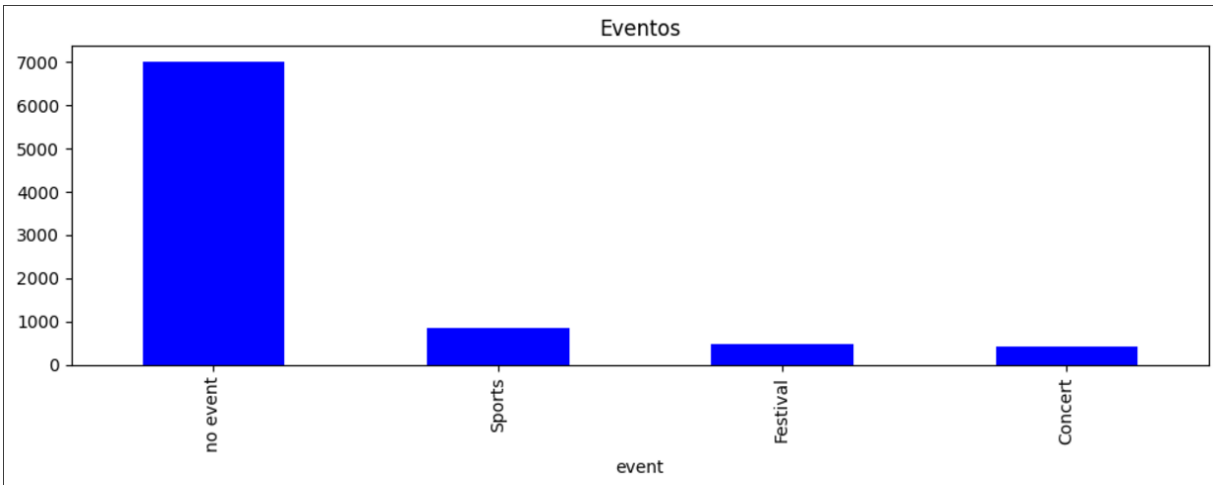
GRÁFICO 4– Estatísticas dos dados de feriados



FONTE: Autor (2024)

Para a coluna de eventos que representa se teve algum tipo de evento na hora os dados ficaram distribuídos da seguinte forma: Sem evento com 7.024, esporte com 861, festival com 465, concerto com 409. Como mostra o GRÁFICO 5.

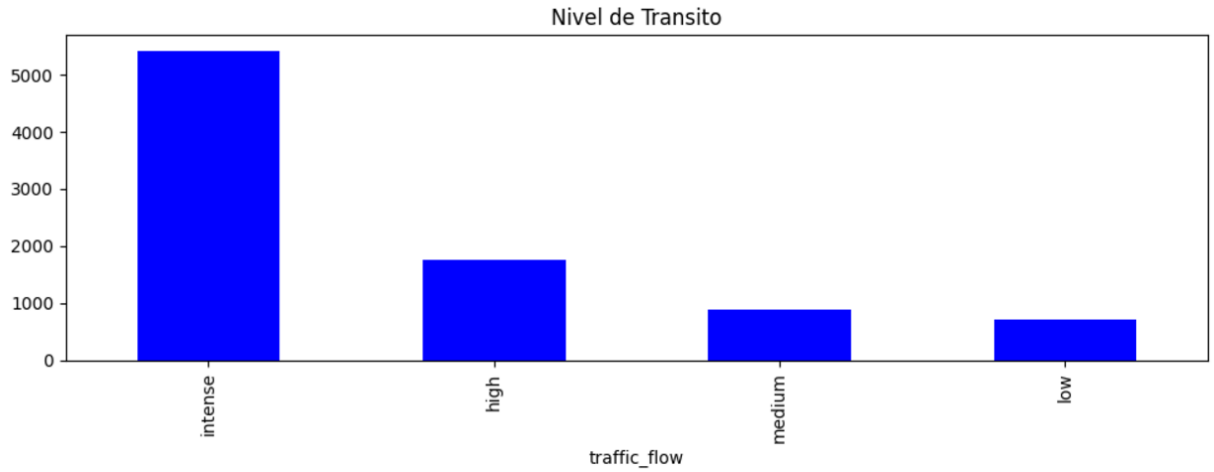
GRÁFICO 5 – Estatísticas dos dados de eventos



FONTE: Autor (2024)

O atributo meta, representado pela coluna *Traffic Flow*, apresentou a maior disparidade entre os dados, indicando a necessidade de balanceamento antes da aplicação de um modelo de aprendizado de máquina. A distribuição foi a seguinte: 707 registros classificados como "baixo", 874 como "médio", 1.756 como "alto" e 5.422 como "intenso". A representação gráfica desses dados pode ser visualizada no GRÁFICO 6.

GRÁFICO 6 – Estatísticas dos dados de níveis de trânsito



FONTE: Autor (2024)

4.4 APLICAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Para a aplicação dos modelos de aprendizado de máquina, os dados foram divididos em 70% para treinamento e 30% para testes. Os algoritmos selecionados para essa análise foram *RandomForestClassifier*, *MLPClassifier* e *KNeighborsClassifier*, devido à sua robustez e adequação a diferentes características de conjuntos de dados.

O treinamento do modelo *RandomForestClassifier* foi realizado utilizando validação cruzada combinada com busca em grade (*GridSearchCV*), o que possibilitou a otimização de hiperparâmetros para maximizar o desempenho do modelo. O grid incluiu combinações para os parâmetros:

- n_estimators: número de estimadores (de 100 a 300);
- max_depth: profundidade máxima das árvores (10 e 20);
- criterion: critérios para divisão dos nós ("gini" e "entropy");
- max_features: número máximo de variáveis consideradas para divisão;
- min_samples_split e min_samples_leaf: quantidade mínima de amostras necessárias para dividir ou formar um nó folha.

Os dados de treinamento foram previamente normalizados (X_treino_normalizados) e acompanhados por suas respectivas classes-alvo (y_treino). A validação cruzada foi realizada com três divisões (cv=3) e o processo foi paralelizado (n_jobs=8) para otimizar a eficiência computacional. Cada

configuração de hiperparâmetros testada foi armazenada em um banco de dados, permitindo uma análise detalhada dos resultados. O melhor conjunto de parâmetros foi selecionado com base na maior acurácia observada durante o treinamento, registrada como best_score_%. O tempo total de execução e o número de combinações avaliadas também foram reportados, refletindo a abrangência e eficiência do processo.

O treinamento do modelo *KNeighborsClassifier*(KNN) também foi conduzido utilizando *GridSearchCV* para otimizar os hiperparâmetros. As combinações testadas incluíram:

- n_neighbors: número de vizinhos (de 3 a 11);
- weights: métodos de ponderação ("uniform" e "distance");
- metric: métricas de distância ("euclidean", "manhattan", "chebyshev" e "minkowski").

Assim como no modelo Random Forest, a validação cruzada com três divisões (cv=3) foi aplicada para reduzir o risco de *overfitting* e garantir uma avaliação robusta. A normalização dos dados foi essencial para evitar interferências na medida de distância. O treinamento foi paralelizado (n_jobs=8) e os resultados detalhados de cada configuração foram armazenados no banco para análise comparativa. O conjunto de hiperparâmetros com maior acurácia (best_score_%) foi selecionado, e o tempo total de treinamento e o número de modelos avaliados foram registrados para comparações.

O treinamento do modelo *MLPClassifier*(RNA) foi estruturado para otimizar hiperparâmetros com o uso de *GridSearchCV*, abrangendo:

- hidden_layer_sizes: configurações de uma a duas camadas ocultas com diferentes números de neurônios;
- activation: funções de ativação ("identity", "logistic", "tanh", "relu");
- solver: otimizadores ("adam", "sgd", "lbfgs");
- alpha: taxas de regularização (10^{-4} a 10^{-1});
- learning_rate: métodos de ajuste da taxa de aprendizado ("constant" e "adaptive").

A validação cruzada (cv=3) foi aplicada, e o processo foi paralelizado (n_jobs=-1) para explorar os recursos computacionais disponíveis. Para garantir a convergência sem comprometer a eficiência, o modelo foi configurado com um limite de 500 iterações. Como nos outros modelos, os resultados de cada combinação

foram também armazenados no banco, permitindo identificar a configuração que maximizou a acurácia (best_score_%). O tempo total e o número de combinações avaliadas foram reportados para medir a eficiência do processo.

4.5 TREINAMENTO DO MODELO

O modelo *RandomForestClassifier* alcançou uma acurácia de 66,73% durante o treinamento, classificando corretamente cerca de dois terços das instâncias no conjunto de validação. Embora não seja uma acurácia extremamente alta, o desempenho pode ser aceitável dependendo da complexidade do problema e dos dados analisados. Esse resultado indica que o modelo identificou padrões relevantes, mas pode enfrentar dificuldades na distinção entre classes devido à variabilidade nos dados.

Os hiperparâmetros otimizados foram:

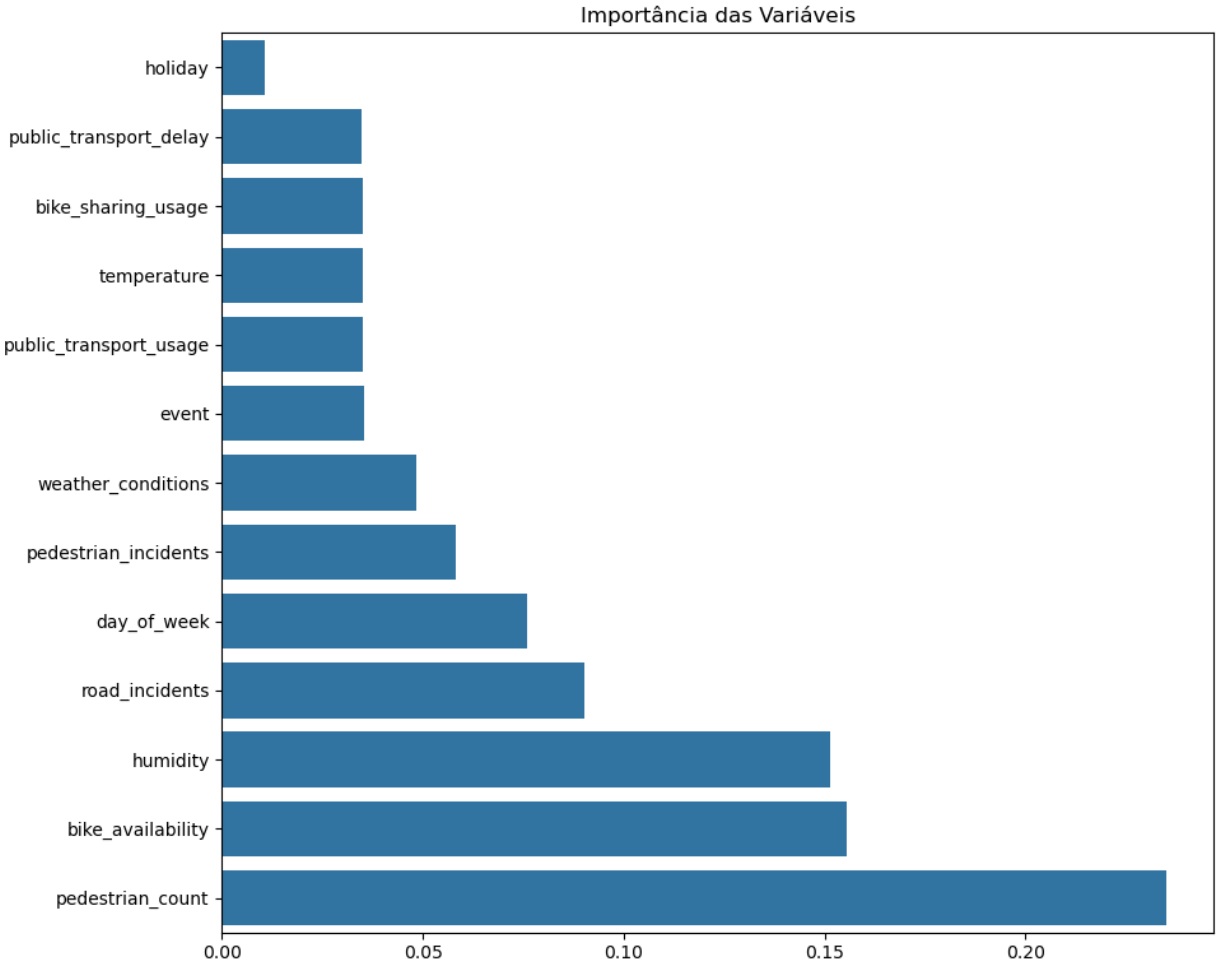
- Critério de impureza: gini
- Profundidade máxima: 20 (max_depth=20)
- Número de árvores no ensemble: 300 (n_estimators=300)
- Número mínimo de amostras por folha: 1 (min_samples_leaf=1)
- Número mínimo de amostras para divisão: 2 (min_samples_split=2)
- Variáveis consideradas por divisão: Todas (max_features=None)

Esses parâmetros configuram um modelo robusto, com alta capacidade de aprendizado. No entanto, o uso de profundidade máxima relativamente alta e a ausência de restrições no número de variáveis aumentam o risco de overfitting.

O treinamento exigiu 6091,81 segundos, devido à extensa busca em grade que avaliou 540 combinações de hiperparâmetros. Apesar da alta demanda computacional, o uso de paralelismo (n_jobs=8) ajudou a reduzir o tempo total. Os resultados mostram que o *RandomForestClassifier* é eficaz para capturar padrões em dados complexos, mas sua performance pode ser melhorada com estratégias como engenharia de features, balanceamento de classes ou uma busca mais refinada por hiperparâmetros.

Após o treinamento rodar é possível verificar o peso de cada variável, mostrando que para o modelo de RandomForestClassifier o atributo com maior peso foi a contagem de pedestres, e a variável com menor peso foi de feriados. Representado no GRÁFICO 7.

GRÁFICO 7 – Importância das variáveis para randomforest



FONTE: Autor (2024)

O modelo *KNeighborsClassifier* alcançou uma acurácia de 64,21% durante o treinamento, indicando que foi capaz de classificar corretamente pouco mais de 60% das instâncias no conjunto de validação cruzada. Esse desempenho sugere que o modelo capturou parte dos padrões nos dados, mas enfrenta desafios como sobreposição de classes ou ruídos.

Os hiperparâmetros otimizados foram:

- Métrica de distância: Manhattan (metric='manhattan')
- Número de vizinhos: 3 (n_neighbors=3)
- Pesos: Baseados na distância (weights='distance')

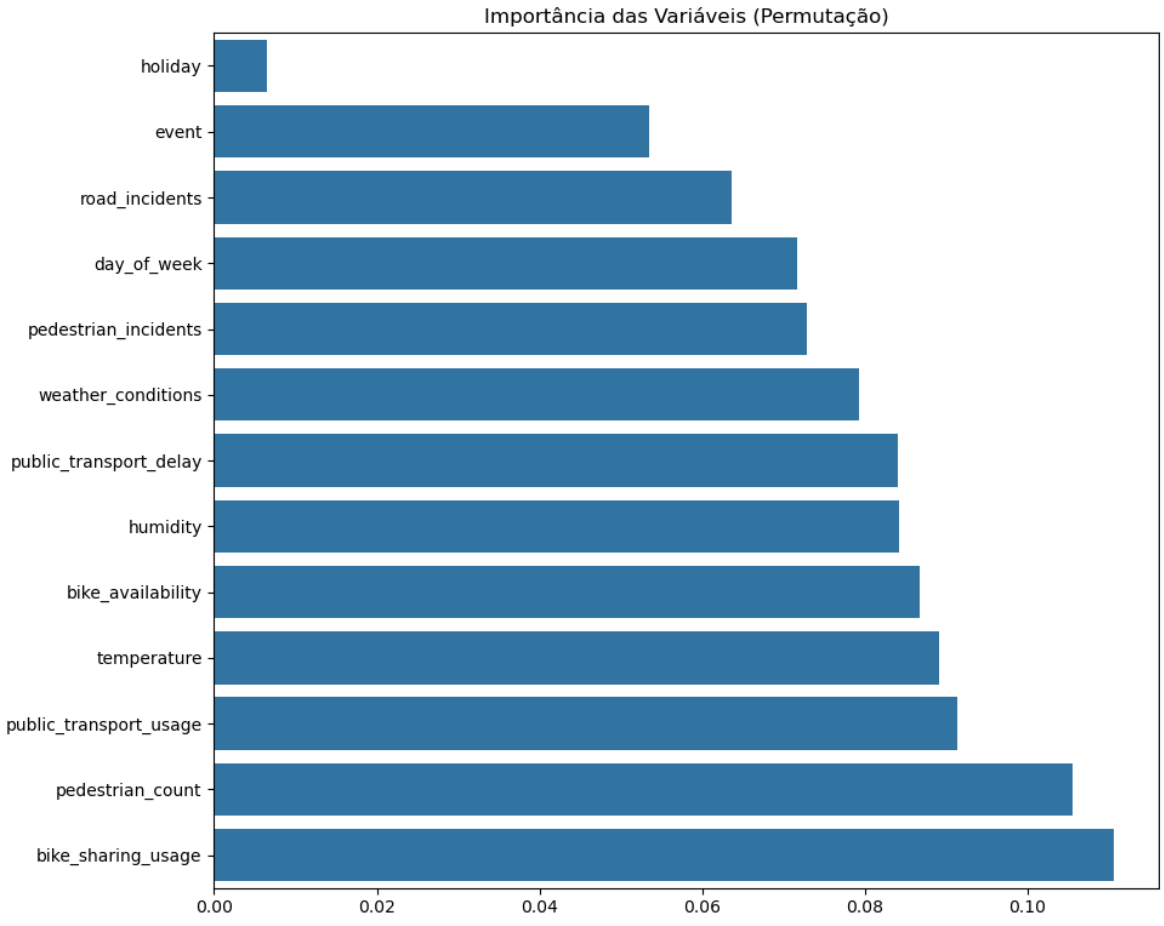
Essa configuração favorece um modelo sensível a padrões locais, com maior influência dos vizinhos mais próximos. O treinamento foi concluído em 290,1 segundos, avaliando 40 combinações de hiperparâmetros. O tempo reduzido reflete

a simplicidade computacional inerente ao KNN, que, combinado ao uso de paralelismo, torna-o uma opção viável em cenários com restrições de tempo.

Apesar da acurácia moderada, o KNN pode ser aprimorado com seleção de características, balanceamento de classes ou ajustes refinados nos hiperparâmetros. Sua simplicidade e baixo custo computacional o tornam uma escolha prática, especialmente em problemas com separações claras entre classes.

Diferente do modelo anterior, o KNN, o atributo mais importante foi o de uso de bicicletas compartilhadas, mas como no modelo de *RandomForest*, a contagem de pedestres esteve muito presente, como visto no GRÁFICO 8.

GRÁFICO 8 – Importância das variáveis para KNN



FONTE: Autor (2024)

O modelo *MLPClassifier* apresentou uma acurácia de 48,48%, indicando dificuldades em capturar padrões significativos nos dados e classificando corretamente menos da metade das instâncias durante a validação cruzada. Esse

desempenho limitado pode ser atribuído ao desequilíbrio entre classes, à complexidade do problema ou à necessidade de ajustes adicionais nos hiperparâmetros e no pré-processamento.

Os hiperparâmetros otimizados foram:

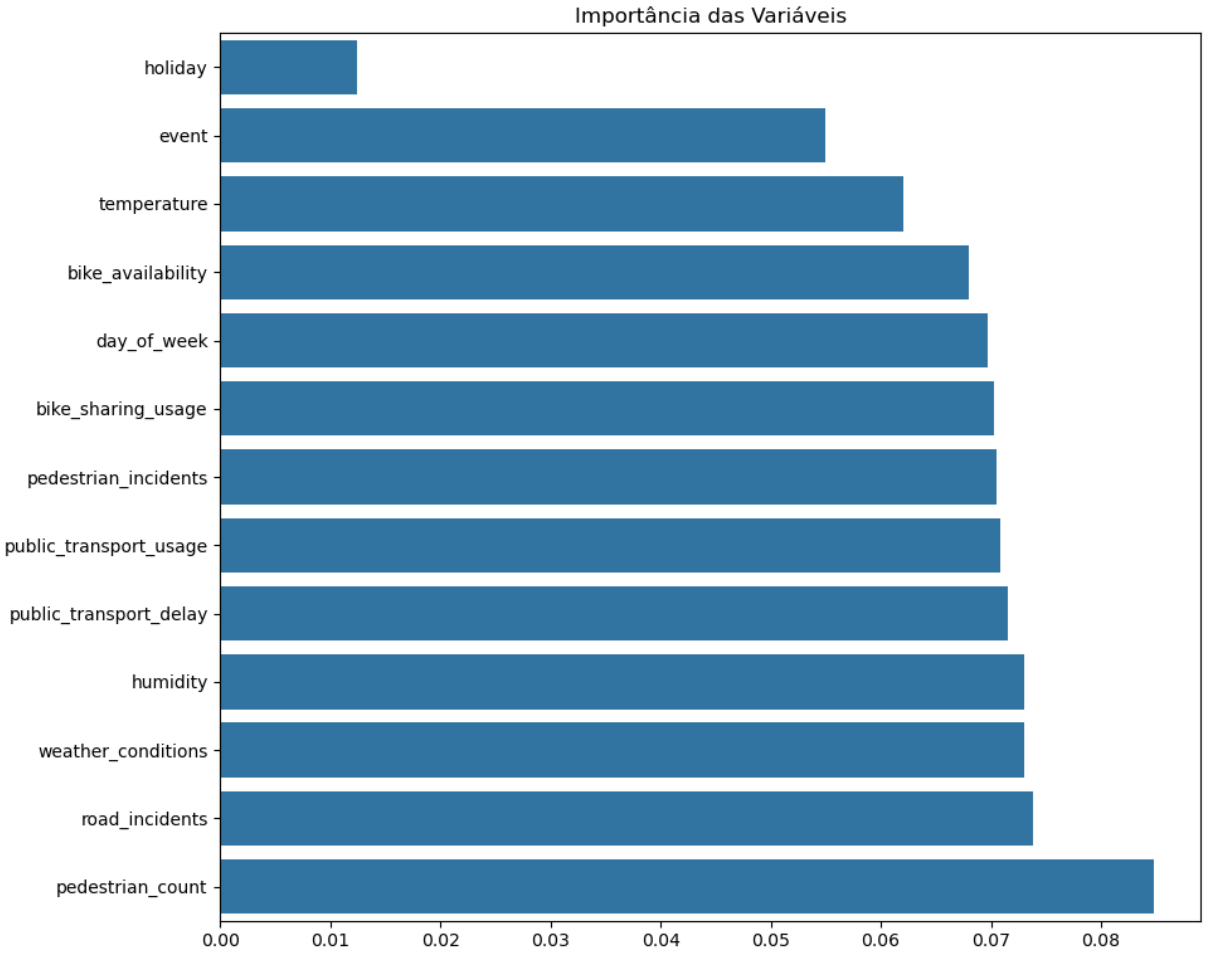
- Função de ativação: ReLU
- Regularização: Alpha = 0,001
- Arquitetura: Duas camadas ocultas com 100 e 50 neurônios
- Otimizador: Adam
- Taxa de aprendizado: Constante

O treinamento exigiu 2783,34 segundos, avaliando 384 combinações de hiperparâmetros. Apesar de ser intensivo, o uso de paralelismo ajudou a mitigar o custo computacional.

Os resultados indicam que o RNA é uma abordagem flexível, capaz de aprender representações complexas, mas dependente da qualidade dos dados e de configurações precisas. A acurácia limitada sugere a necessidade de maior engenharia de features, balanceamento das classes ou ajustes na arquitetura da rede.

E como nos outros modelos já mostrados a variável de maior peso foi a de contagem de pedestre, porém diferente dos outros modelos, nessa cada variável teve quase um peso parecido de importância, representado no GRÁFICO 9, a variável de feriado em todos os três modelos foi o que teve menor destaque, o que pode representar realizar uma troca de variável ou realizar uma engenharia de dados para melhorar seu peso.

GRÁFICO 9 – Importância das variáveis para RNA



FONTE: Autor (2024)

4.6 VERIFICAÇÃO DOS RESULTADOS NA BASE DE TESTE

No conjunto de teste, o modelo *RandomForestClassifier* atingiu uma acurácia de 70,57%, superior a acurácia de treinamento de 66,73%, indicando boa generalização para dados não vistos e ausência de sinais de *overfitting*. Esse desempenho robusto sugere que os padrões aprendidos no treinamento são representativos da distribuição geral dos dados.

As métricas detalhadas mostram um desempenho consistente entre as classes, com médias ponderadas em torno de 70% para precisão, revocação e F1-score:

- Precisão: 70%, indicando que a maioria das predições positivas do modelo está correta;

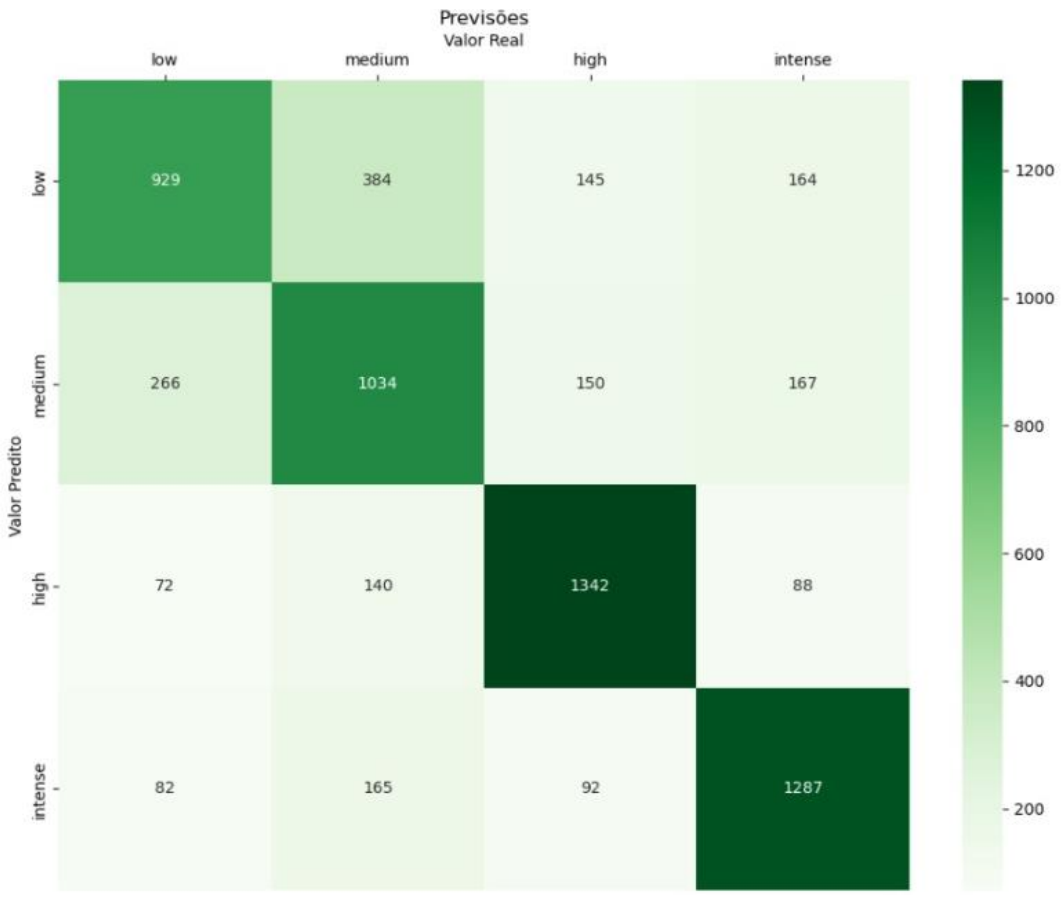
- Revocação: 71%, refletindo boa capacidade de identificar instâncias positivas reais;
- F1-score: 70%, demonstrando um equilíbrio adequado entre precisão e revocação.

A matriz de confusão detalha o desempenho por classe:

- Classe "high": 929 predições corretas de 1622 instâncias reais (revocação de 57%);
- Classe "low": Melhor desempenho, com 1342 acertos de 1642 instâncias reais (revocação de 82%);
- Classe "medium": 1287 acertos de 1626 instâncias reais (revocação de 79%).

Os resultados indicam que o modelo apresenta maior dificuldade em distinguir entre as classes "high" e "intense", enquanto classifica com maior precisão as classes "low" e "medium". Esse comportamento pode ser visualizado no GRÁFICO 10.

GRÁFICO 10 – Matriz de confusão do Random Forest



FONTE: Autor (2024)

O modelo de KNN alcançou uma acurácia de 65,44% no conjunto de teste, levemente superior à acurácia de treinamento de 64,21%, sugerindo boa generalização e ausência de *overfitting*. Embora o desempenho seja menor que o do *RandomForestClassifier*, ele é satisfatório considerando os desafios da classificação multiclasse.

As métricas detalhadas revelam um desempenho equilibrado entre as classes:

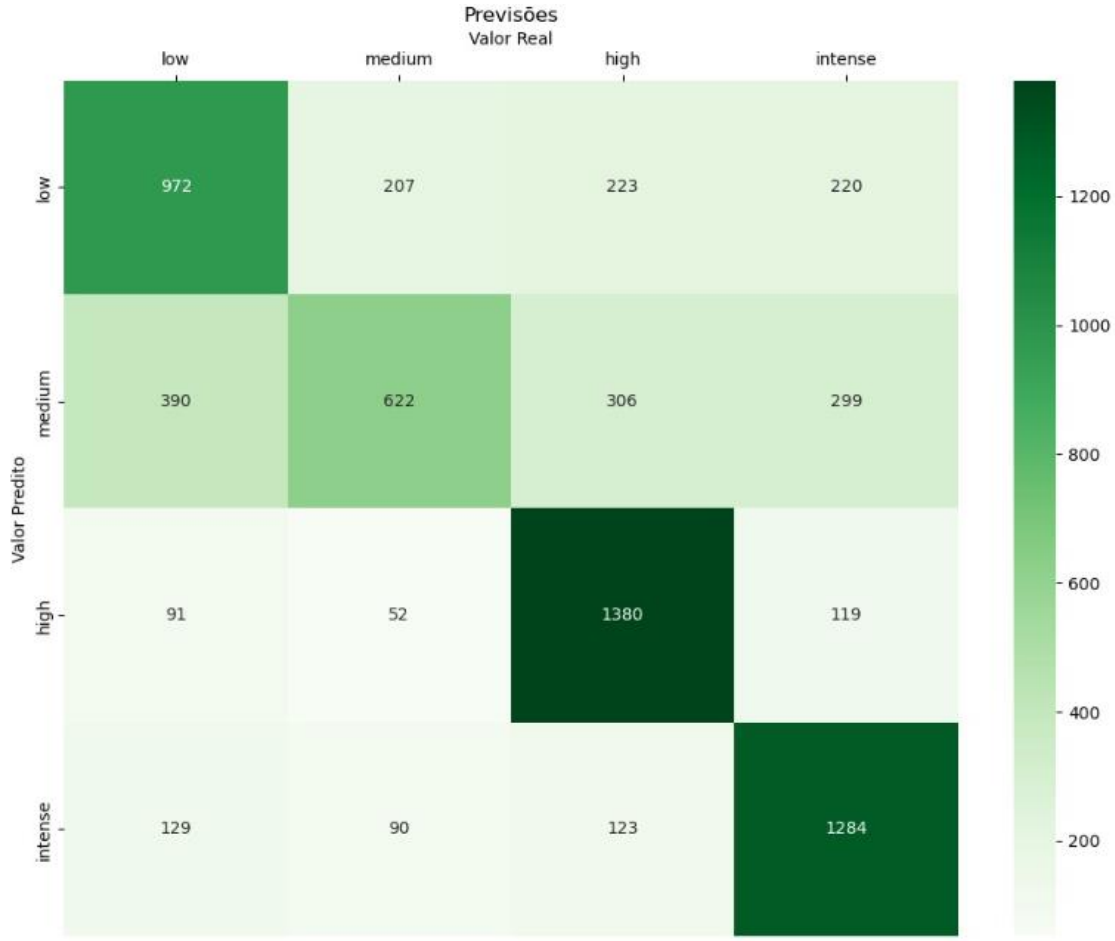
- Precisão: Variou de 61% ("high") a 68% ("low");
- Revocação: Mais alta para a classe "low" (84%) e mais baixa para "intense" (38%);
- F1-score: Melhor desempenho nas classes "low" (75%) e "medium" (72%);

A matriz de confusão mostra os padrões de acertos e erros por classe:

- Classe "high": 972 acertos de 1622 instâncias reais (revocação de 60%);
- Classe "intense": Desempenho mais baixo, com 622 acertos de 1617 instâncias reais (revocação de 38%);
- Classe "low": Melhor desempenho, com 1380 acertos de 1642 instâncias reais (revocação de 84%);
- Classe "medium": 1284 acertos de 1626 instâncias reais (revocação de 79%).

Os dados mostram que o modelo tem dificuldade em diferenciar as classes "high" e "intense", mas apresenta bom desempenho nas classes "low" e "medium", conforme ilustrado no GRÁFICO 11.

GRÁFICO 11 – Matriz de confusão do KNN



FONTE: Autor (2024)

O modelo RNA atingiu uma acurácia de 50,24% no conjunto de teste, classificando corretamente cerca de metade das instâncias. Esse resultado reflete desafios na aprendizagem de padrões claros, possivelmente devido à arquitetura da rede ou à seleção de hiperparâmetros.

As métricas detalhadas destacam áreas de melhoria:

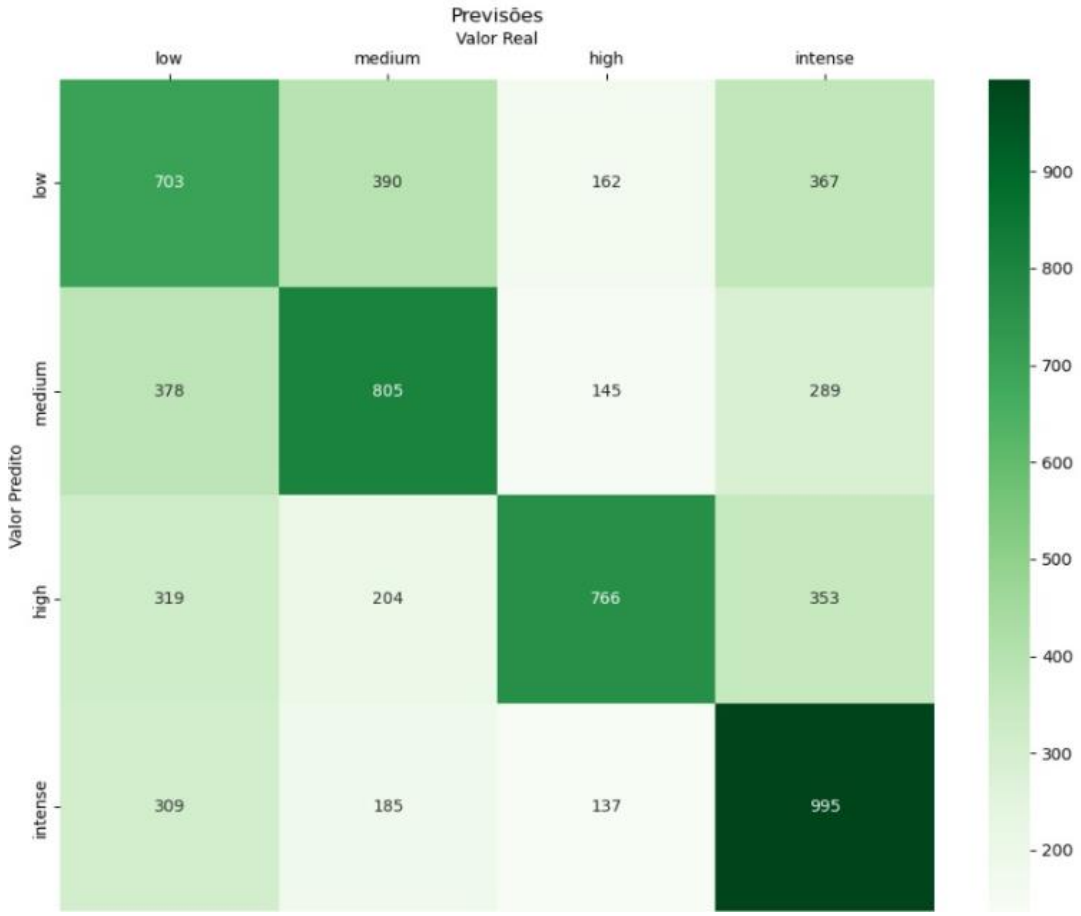
- Precisão: Média ponderada de 51%, indicando que pouco mais da metade das predições positivas estava correta;
- Revocação: Média ponderada de 50%, mostrando que o modelo identificou corretamente metade das instâncias reais;
- F1-score: Média ponderada de 50%, demonstrando equilíbrio limitado entre precisão e revocação.

A matriz de confusão detalha o desempenho por classe:

- Classe "high": 703 acertos de 1622 instâncias reais (revocação de 43%);
- Classe "intense": 805 acertos de 1617 instâncias reais (revocação de 50%);
- Classe "low": 766 acertos de 1642 instâncias reais (revocação de 47%);
- Classe "medium": Melhor desempenho, com 995 acertos de 1626 instâncias reais (revocação de 61%).

Esses resultados sugerem que o modelo enfrenta dificuldades para diferenciar as classes "high", "intense" e "low", mas apresenta desempenho razoável na classificação da classe "medium". Esse comportamento está representado no GRÁFICO 12.

GRÁFICO 12 – Matriz de confusão de RNA



FONTE: Autor (2024)

De forma geral o uso de uma base de dados artificial apresenta vantagens importantes, como controle sobre as distribuições de dados e a possibilidade de criar cenários específicos para avaliar a performance dos modelos. No entanto, esse controle pode limitar a representatividade dos padrões reais, especialmente quando as relações entre as variáveis são mais complexas ou não lineares, como ocorre em muitos problemas do mundo real. Isso pode explicar, em parte, a dificuldade observada nos modelos em distinguir adequadamente classes mais próximas, como "high" e "intense".

O *RandomForestClassifier* destacou-se como o modelo mais robusto, atingindo alta acurácia e equilíbrio entre precisão, revocação e F1-score. Sua capacidade de capturar relações complexas em dados multiclasses é uma grande vantagem, especialmente em cenários onde as classes são bem definidas e os dados são suficientemente representativos. Contudo, a dependência de dados bem balanceados para evitar viés pode ser uma limitação, particularmente em base de dados artificiais onde as classes podem ter distribuições diferentes das observadas em contextos reais.

O KNN apresentou desempenho moderado, com acurácia menor do que o *RandomForestClassifier*, mas ainda consistente. Sua simplicidade é uma vantagem em problemas onde a separação entre as classes é clara, mas a dependência de uma métrica de distância adequada pode tornar o método menos eficaz em *datasets* artificiais com padrões complexos ou com classes que se sobrepõem. Além disso, o custo computacional de KNN pode crescer rapidamente com o aumento do tamanho do base de dados.

Por outro lado, a RNA apresentou o menor desempenho entre os três modelos, o que ressalta a sensibilidade das redes neurais a configurações como arquitetura e hiperparâmetros. Apesar de ser uma ferramenta poderosa, especialmente para padrões complexos e não lineares, a RNA exige uma quantidade significativa de dados e ajustes finos para alcançar resultados competitivos. Em uma base de dados artificial, onde as relações podem não refletir situações do mundo real, a RNA pode ter dificuldades em generalizar padrões, resultando em uma performance aquém do esperado.

De maneira geral, os resultados evidenciam as vantagens do uso de modelos mais interpretáveis, como *RandomForestClassifier*, em base artificiais, enquanto métodos mais sensíveis a parametrizações, como RNA, podem não ser

ideais nesses cenários. A análise destaca a importância de alinhar a escolha do modelo às características do problema e da base, considerando tanto o objetivo da análise quanto às limitações impostas pela natureza artificial dos dados.

Essa pesquisa mostra que por meio de um dataset artificial que combina diferentes dados de trânsito, é possível fazer uma abordagem prática para analisar e prever padrões relacionados ao fluxo de deslocamentos. Com base nas predições realizadas pelos modelos, gestores podem simular cenários e planejar intervenções específicas, como prever o problema de um grande fluxo de trânsito alto em dias de shows ou jogos de futebol. Outro exemplo de como essa pesquisa contribui com o problema de mobilidade urbana, é a promoção de tomadas de decisão baseadas em dados, pois permite ter um embasamento na tomada de decisões.

5 CONSIDERAÇÕES FINAIS

A pesquisa explora o potencial das técnicas de aprendizado de máquina na solução de desafios complexos relacionados à mobilidade urbana. O crescimento acelerado das populações urbanas, juntamente com o aumento da demanda por transporte eficiente e sustentável, exige a adoção de abordagens inovadoras para otimizar o planejamento e a gestão do transporte. Nesse contexto, as técnicas de aprendizado de máquina, com sua capacidade de analisar grandes volumes de dados e identificar padrões ocultos, surgem como uma ferramenta poderosa para enfrentar esses desafios.

Por meio da aplicação de modelos, foi possível identificar padrões significativos e realizar predições que podem contribuir de forma substancial para o planejamento urbano. Os resultados obtidos indicaram que, embora os modelos apresentem níveis variados de acurácia, com alguns superando outros em determinadas tarefas, todos mostraram seu potencial para oferecer soluções práticas e aplicáveis na gestão da mobilidade urbana. A variação nos desempenhos dos modelos sugere que a sua aplicação prática pode ser substancialmente ampliada e melhorada por meio de ajustes no processo de engenharia de dados, como o refinamento das variáveis de entrada e a melhoria da qualidade dos dados. Além disso, o ajuste de hiperparâmetros dos modelos pode levar a uma maior precisão nas predições, especialmente em cenários mais complexos e variáveis.

Outro aspecto relevante abordado neste trabalho foi o uso de dados sintéticos, que se mostrou uma alternativa viável e eficaz para contornar a escassez de dados reais sobre tráfego urbano, que frequentemente estão sujeitos a limitações de acesso ou incompletude. O uso de dados sintéticos permite criar um ambiente controlado e ajustável para experimentação, sem as limitações impostas por dados reais. Isso possibilita a realização de testes em diferentes cenários, com variações nas condições climáticas, eventos urbanos ou modificações na infraestrutura de transporte, o que seria muito mais difícil com dados reais, especialmente em situações de falta de informações históricas.

Este estudo não só demonstrou a aplicabilidade das técnicas de aprendizado de máquina na mobilidade urbana, como também abriu caminho para futuras pesquisas que possam explorar a combinação de dados sintéticos e reais,

bem como a adoção de novas abordagens técnicas e melhorias nos modelos existentes. Isso pode resultar em soluções ainda mais robustas e precisas para a gestão do transporte urbano.

Uma limitação observada no fim da pesquisa foi a falta de dados reais sobre a mobilidade urbana, tanto em Curitiba quanto no Brasil como um todo. Essa carência de bases de dados abrangentes e atualizadas dificulta a análise precisa das dinâmicas de transporte, comprometendo o desenvolvimento de políticas públicas eficientes e a aplicação de modelos preditivos que poderiam auxiliar na melhoria da qualidade de vida e na gestão urbana.

5.1 VERIFICAÇÃO DOS OBJETIVOS PROPOSTOS

A questão de pesquisa que guiou este estudo foi: "Como técnicas de aprendizado de máquina podem ser aplicadas para melhorar a mobilidade urbana?". Com base nessa questão, foram estabelecidos um objetivo geral e três objetivos específicos, cujos resultados são discutidos a seguir, detalhando como cada um foi alcançado.

O objetivo geral deste estudo, que consistiu em investigar e desenvolver um modelo de aprendizado de máquina para otimizar a mobilidade urbana, foi plenamente atingido. Para isso, foi realizada uma seleção criteriosa de uma base de dados sobre mobilidade urbana, incluindo informações sobre fluxo de tráfego, uso de transporte público e padrões de deslocamento. Em seguida, foi aplicado um processo de pré-processamento dos dados, que envolveu limpeza, normalização e transformação, para garantir que estivessem adequados para a aplicação dos modelos de aprendizado de máquina.

Entre os objetivos específicos, o primeiro foi prospectar estudos relacionados ao tema de pesquisa. Este foi alcançado por meio de uma ampla revisão de literatura, que incluiu fontes como o repositório Periódicos CAPES, Kaggle e UCI *Machine Learning Repository*. O levantamento de estudos e dados relevantes possibilitou embasar a pesquisa teoricamente, identificando tendências e metodologias adequadas para a aplicação de aprendizado de máquina em mobilidade urbana.

O segundo objetivo consistiu em propor um modelo de aprendizado de máquina para otimizar a mobilidade urbana, o que foi plenamente alcançado por

meio do desenvolvimento e aplicação de um modelo preditivo utilizando dados sintéticos, representando cenários urbanos de forma controlada. Para isso, foram aplicados algoritmos como Random Forest, Redes Neurais Artificiais e *K-Nearest Neighbors* para prever padrões de trânsito e propor melhorias. O desempenho dos modelos foi avaliado com métricas como acurácia, que variou entre 64% e 66%, permitindo identificar o algoritmo mais robusto para resolver o problema proposto. Além disso, o modelo *RandomForest* se destacou como o mais robusto na identificação de padrões complexos de mobilidade urbana, comprovando a eficácia das técnicas aplicadas. O processo de treinamento envolveu algoritmos supervisionados, ajustados por meio de validação cruzada e busca em grade (*GridSearchCV*), utilizando a base de dados pré-processada, com resultados que validaram o sucesso da abordagem. O terceiro e último objetivo específico foi documentar e apresentar os achados do estudo de forma clara e detalhada. Para alcançar esse objetivo, todos os passos metodológicos foram cuidadosamente registrados, desde a coleta e pré-processamento dos dados até os testes e análise de desempenho dos modelos. A apresentação incluiu gráficos, tabelas e matrizes de confusão, garantindo a clareza dos resultados e sua relevância para o problema abordado.

5.2 TRABALHOS FUTUROS

Dada a complexidade da mobilidade urbana, futuras pesquisas podem expandir este estudo em várias direções:

Aplicação do modelo desenvolvido em dados reais de diferentes cidades para validar a eficácia em cenários práticos.

Exploração de arquiteturas mais avançadas de aprendizado de máquina, como redes neurais profundas, para melhorar a precisão das predições.

Investigação sobre o impacto de políticas públicas e infraestrutura urbana nos padrões de tráfego, utilizando aprendizado de máquina para simular cenários futuros.

5.3 CONTRIBUIÇÕES DA PESQUISA

Este trabalho de conclusão de curso contribuiu para o campo da ciência de dados ao demonstrar como dados sintéticos podem ser usados para treinar modelos de aprendizado de máquina em mobilidade urbana. Além disso, apresentou uma metodologia clara para a coleta, pré-processamento e análise de dados, que pode ser replicada em outras pesquisas.

Do ponto de vista prático, os resultados obtidos oferecem *insights* para gestores urbanos e planejadores, auxiliando na tomada de decisões informadas para reduzir congestionamentos e melhorar a eficiência dos sistemas de transporte.

Por fim, a pesquisa serve como um ponto de partida para futuros trabalhos que busquem integrar inteligência artificial e gestão da informação na construção de cidades mais inteligentes.

REFERÊNCIAS

ALVES BESERRA, C. et al. Aplicação de Técnicas de Aprendizagem de Máquina em Objetos de Aprendizagem baseado em Software: um Mapeamento Sistemático a partir das Publicações do SBIE. **RENOTE**, v. 12, n. 1, 2014.

ANDRADE, M. S. et al. Análise de Métricas de Desempenho sobre a Conjuntura de Intrusões em Redes IEEE 802.11 com Aprendizagem de Máquina no Hospital N.S.C. **Research, Society and Development**, v. 12, n. 4, p. e22512441277, 2023.

ARAUJO, M. et al. transporte público coletivo: discutindo acessibilidade, mobilidade e qualidade de vida. **Psicologia & Sociedade,** v. 23, p. 574-582, 2011 Disponível em: https://www.scielo.br/j/psoc/a/XWXTQXKJ44BtT5Qw7dLWgvF/?format=pdf&lang=pt Acesso em: 19 Dez. 2024.

BARBOSA, R. R. Gestão da informação e do conhecimento: origens, polêmicas e perspectivas. **Informação & informação**, v. 13, n. 1esp, p. 1, 2008.
CAMPAGNARO, E.; CERVANTES, B. N. Hipertexto na coleta caótica da informação nas organizações públicas. **Informação & informação**, v. 16, n. 1, p. 52–71, 2011.

CARVALHO, R. C. DE. Aplicação de mineração de dados em informações oriundas de prontuários de paciente. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, v. 14, n. 1, 2019.

COELHO, M. **Fundamentos de Redes Neurais - Laboratório iMobilis**. Disponível em: <https://www2.decom.ufop.br/imobilis/fundamentos-de-redes-neurais/>. Acesso em: 30 jun. 2024.

COSTA, M. *Um Índice de Mobilidade Urbana Sustentável*. 2008. Tese (Doutorado em Engenharia Civil) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2008.

DAMASCO MENZORI, I.; GONÇALVES, L. M. Sustentável, digital ou inteligente: paradoxo e paradigma das tecnologias na mobilidade urbana. **Engenharia Urbana em Debate**, v. 1, n. 1, p. 230–241, 2023.

DE ALMEIDA, D. S.; DA SILVA, A. L. L. Geoprocessamento com banco de dados NoSQL e MapReduce. **Revista Brasileira de Geomática**, v. 11, n. 2, p. 441, 2023.
Decision Tree Algorithm in Machine Learning - javatpoint. Disponível em: <https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>. Acesso em: 30 jun. 2024.

DEEP LEARNING. Disponível em: <https://www.deeplearningbook.com.br/>. Acesso em: 30 jun. 2024.

FRADKOV, A. L. Early history of machine learning. **IFAC-PapersOnLine**, v. 53, n. 2, p. 1385–1390, 2020.

GALVÃO, N. D.; MARIN, H. DE F. Técnica de mineração de dados: uma revisão da

literatura. **Acta Paulista de Enfermagem**, v. 22, n. 5, p. 686–690, 2009.
GODOY, A. S. Introdução à pesquisa qualitativa e suas possibilidades. **RAE**, v. 35, n. 2, p. 57–63, 1995.

GONÇALVES, Á.; HENRIQUE, C. A internet de todas as coisas e a educação: possibilidades e oportunidades para os processos de ensino e aprendizagem. **Linkscienceplace**, v. 3, n. 3, p. 31–45, 2017.

HANDAYA, A. *Machine Learning com o Google Colab*. **Anais da XIII Semana da Matemática e Educação Matemática** – IFSP/Campus Guarulhos. XIII SEMAT, ISSN: 2965-5056, Guarulhos, 13 a 17 de maio de 2024.

HORA, G. S. et al. Avaliação de ferramentas de mineração de dados: uma abordagem com o modelo tam. **Interfaces Científicas - Exatas e Tecnológicas**, v. 2, n. 3, p. 109–121, 2018.

HORN, B. S. et al. Aplicação de técnicas de mineração de dados como ferramenta para geração de informação: um caso sobre o uso de análise sistemática em periódicos da área de design de moda. **Produto & Produção**, v. 21, n. 3, 2020.

MACHADO, R. P. M.; STREIT, R. E. Gestão da informação em bancos públicos federais: Novos desafios diante da Lei de Acesso à Informação (LAI). **Informação & informação**, v. 23, n. 1, p. 204, 2018.

Machine Learning: o que é aprendizado semi-supervisionado. Disponível em: <https://www.alura.com.br/artigos/machine-learning-aprendizado-semi-supervisionado>. Acesso em: 30 jun. 2024.

MARZINOTTO JUNIOR, F. L. A economia política do Big Data: Um recurso estratégico e de Poder entre oligopólios tecnológicos e vulnerabilidades estatais I the political economy of Big Data: A strategic and power resource among technological oligopolies and state vulnerabilities. **Revista Neiba, Cadernos Argentina Brasil**, v. 10, n. 1, p. e59061, 2021.

MORAES, M. **Tomada de decisão na priorização de pacientes em fila de espera cirúrgica baseada em aprendizado de máquina**. 2021. 102 f. Dissertação (Mestrado Profissional em Gestão para a Competitividade) – Escola de Administração de Empresas de São Paulo, Fundação Getulio Vargas, São Paulo, 2021. Orientador: Prof. Dr. Gonzalo Vecina Neto. Coorientador: Prof. Dr. Marco Antonio Gutierrez.

MIRANDA, M. A. S. DE et al. Ciclo de vida da informação no suporte ao processo de inovação: Uma proposta de modelo interativo. **Gestão & Planejamento**, v. 20, p. 581–599, 2019.

MITCHELL, T.; BUCHANAN, B.; DEJONG, G.; DIETTERICH, T.; ROSENBLOOM, P.; WAIBEL, A. Machine learning. Annual review of computer science, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 4, n. 1, p. 417–433, 1990.

NETO, A; BONINI, C. REDES NEURAIS ARTIFICIAIS: APRESENTAÇÃO E UTILIZAÇÃO DO ALGORITMO PERCEPTRON EM BIOSSISTEMAS. **BioEng,** v.4 n.2, p. 87-95, 2010, Tupã. Disponível em: https://seer.tupa.unesp.br/index.php/BIOENG/article/view/95/95 Acesso em: 19 Dez. 2024.

NONATO, R. DOS S.; AGANETTE, E. C. Gestão da informação: rumo a uma proposta de definição atual e consensual para o termo. **Perspectivas em Ciência da Informação**, v. 27, n. 1, p. 133–159, 2022.

O que é aprendizado supervisionado? Disponível em: <https://www.ibm.com/br-pt/topics/supervised-learning>. Acesso em: 30 jun. 2024.
PANDEY, S. K.; RATHEE, D.; TRIPATHI, A. K. Software defect prediction using K-PCA and various kernel-based extreme learning machine: an empirical study. **IET software**, v. 14, n. 7, p. 768–782, 2020.

PAULA, P. P. DE; DOS SANTOS, C. D. Vulnerabilidade inicial pós-incubação: prevendo a sobrevivência organizacional com aprendizagem de máquina. **Revista Gestão e Desenvolvimento**, v. 21, n. 1, p. 28–50, 2024.

PEIXOTO, L. D. C.; SANTIAGO, M. R. A mobilidade urbana solidária no estatuto da cidade e sua concretização pela via da economia colaborativa. **Revista de Direito da Cidade**, v. 11, n. 2, 2019.

PINHEIRO, L. Gênese da Ciência da Informação ou sinais anunciadores da nova área. In: **O campo da Ciência da Informação: gênese, conexões e especificidades.** João Pessoa, UFPB, P.61-86, 2002.

SALDANHA, R. DE F. A pesquisa científica na era do Big data: cinco maneiras que mostram como o Big data prejudica a ciência, e como podemos salvá-la. **Revista electronica de comunicacao, informacao & inovacao em saude: RECIIS**, v. 16, n. 3, p. 742–745, 2022.

SILVA JÚNIOR, E. M. DA; KARPINSKI, C.; DUTRA, M. L. Conhecimento científico no contexto big data: Reflexões a partir da epistemologia de Popper. **Brazilian Journal of Information Science**, v. 14, n. 4, p. e020017, 2020.

TIOBE. *TIOBE Index*. Disponível em: https://www.tiobe.com/tiobe-index/. Acesso em: 19 dez. 2024.

APÊNDICE 1 – CÓDIGO GITHUB

https://github.com/Togarashi0/TCC-II