UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Mariana Lovato dos Santos

ALGORITMOS DE APRENDIZADO DE MÁQUINA SUPERVISIONADOS APLICADOS EM TRANSPORTES: COMPARATIVO COM MODELO LOGIT MULTINOMIAL PARA ESCOLHA MODAL

Porto Alegre
2023

MARIANA LOVATO DOS SANTOS

ALGORITMOS DE APRENDIZADO DE MÁQUINA SUPERVISIONADOS APLICADOS EM TRANSPORTES: COMPARATIVO COM MODELO LOGIT MULTINOMIAL PARA ESCOLHA MODAL

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul, como requisito parcial à obtenção do título de Mestre em Engenharia de Produção, na modalidade Acadêmica, na área de concentração em Sistemas de Transportes

Orientador: Prof. Ana Margarita Larranaga Uriarte, Dra.

Porto Alegre

2023

MARIANA LOVATO DOS SANTOS

ALGORITMOS DE APRENDIZADO DE MÁQUINA SUPERVISIONADOS APLICADOS EM TRANSPORTES: COMPARATIVO COM MODELO LOGIT MULTINOMIAL PARA ESCOLHA MODAL

Essa dissertação foi julgada adequada como pré-requisito para a obtenção do título de Mestre em Engenharia de Produção na modalidade Acadêmica e aprovada em sua forma final pela Professora Orientadora e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

Porto Alegre, 05 de abril de 2023

Prof. Ana Margarita Larranaga Uriarte
Dra. pelo PPGEP/UFRGS
Orientadora

Prof. Alejandro Germán Frank
Dr. pelo PPGEP/UFRGS
Coordenador do PPGEP/UFRGS

BANCA EXAMINADORA

Prof. Cira Souza Pitombo (USP)
Dra. pela Escola de Engenharia de São Carlos/USP

Prof. Fernando Dutra Michel (UFRGS)
Dr. pelo PPGEP/UFRGS

Luiz Afonso dos Santos Senna (AGERGS)
PhD. pela University of Leeds

LISTA DE FIGURAS

Dedico este trabalho a meu pai Nilson (*in memorium*)
por me ensinar sobre amor e leveza,
e por sempre incentivar
minha busca por conhecimento

AGRADECIMENTOS

Agradeço a Prof. Ana Margarita Larranaga Uriarte, minha orientadora, pelos ensinamentos e incentivos durante a realização deste trabalho.

Agradeço a Maria Cristina Molina Ladeira, pela dedicação, e apoio durante a realização deste trabalho e na minha carreira profissional.

Agradeço ao Prof. Fernando Dutra Michel pela inspiração e ensinamentos na área de transportes, incentivando-me a seguir essa área durante a graduação e o mestrado.

Agradeço a Felipe Lobo de Souza e a Prof. Cira Souza Pitombo por sua valiosa contribuição na elaboração do segundo artigo, com os algoritmos de Aprendizado de Máquina utilizados.

Agradeço à equipe do Laboratório de Sistemas e Transportes (LASTRAN), pelo aprendizado durante o período de bolsa de iniciação científica e pela disponibilização dos dados para a realização deste trabalho.

Agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo fomento à pesquisa científica.

Agradeço a minha família, em especial aos meus pais, por se fazerem presentes em todas as fases de minha vida, pelo amor, compreensão e incentivo constante a educação.

Agradeço a Alexandra, minha irmã, e ao Francisco, meu primo-irmão, pelo amor incondicional, cumplicidade e presença em todos os momentos da minha vida.

Agradeço aos meus amigos – da vida e aos colegas de profissão –, responsáveis pelo equilíbrio e harmonização entre as expectativas e realidades da vida, por se fazerem presentes e pelos momentos especiais compartilhados ao longo destes anos.

Ou átomos, buracos negros, anãs brancas, quasares e protozoários. E diria, com aquele ar levemente pedante: "Quem só acredita no visível tem um mundo muito pequeno. Os dragões não cabem nesses pequenos mundos de paredes invioláveis para o que não é visível".

Caio Fernando Abreu

RESUMO

Como o planejamento no transporte urbano desempenha um papel essencial para o desenvolvimento sustentável dos sistemas de transporte, torna-se evidente a necessidade de explorar novas técnicas de análise para aprimorar a eficiência e a eficácia. Em particular, o uso de técnicas de *Machine Learning* (Aprendizado de Máquina) tem se mostrado promissor para lidar com os desafios complexos relacionados ao planejamento do transporte urbano. A incorporação desses algoritmos pode melhorar a capacidade de análise de dados e fornecer diretrizes para a tomada de decisões. Dado esse contexto, a presente dissertação foi dividida em dois artigos que tem por objetivos: (i) desenvolvimento de uma revisão sistemática da literatura para analisar de forma quantitativa os estudos existentes sobre planejamento de transporte urbano com modelos de *Machine Learning*, identificar os principais temas abordados, quais são as aplicações e como podem auxiliar na otimização dos sistemas de transporte urbano (ii) comparar modelos tradicionais de escolha discreta com algoritmos de Aprendizado de Máquina, a fim de analisar a previsão da escolha modal, utilizando dados provenientes de uma pesquisa de Preferência Declarada (PD) realizada em Porto Alegre em 2019. Os resultados obtidos na revisão sistemática indicam que os métodos de Aprendizado de Máquina estão em crescente utilização no planejamento de transportes. Dentre os métodos analisados, os modelos de previsão de demanda de tráfego e de transporte público se destacaram como os mais empregados na literatura. Além desses, outros métodos, como reconhecimento de sinais de trânsito, detecção de semáforos, classificação de veículos, detecção de pedestres, planejamento de tempo de viagem e de itinerário e comparativos entre algoritmos diferentes também foram frequentemente utilizados. Os resultados do estudo comparativo indicam que o modelo de Logit Multinomial (MLM) apresentou uma acurácia preditiva significativamente maior em comparação com os outros modelos de Aprendizado de Máquina testados. A taxa de acerto do MLM foi de 52,03%, seguida pelo método de Floresta Aleatória (FA) com 41,79%, e as Redes Neurais Artificiais (RNAs) com 40,94%. Esses resultados podem ser explicados pelo fato de que a base de dados utilizada na análise continha poucas observações para os modos de transporte Lotação e Táxi.

Palavras-chave: escolha modal, Transporte Flexível, Modelo Logit Multinomial, Árvore de Decisão, Floresta Aleatória, Redes Neurais Artificiais

ABSTRACT

As urban transportation planning plays an essential role in the sustainable development of transportation systems, there is a clear need to explore new analysis techniques to improve the efficiency and effectiveness of planning. In particular, the use of Machine Learning (ML) techniques has shown promise in dealing with complex challenges related to urban transportation planning. The incorporation of these algorithms can significantly improve data analysis capabilities and provide guidelines for decision-making. Given this context, this dissertation is divided into two articles that aim to: (i) develop a systematic literature review to quantitatively analyze existing studies on urban transportation planning with Machine Learning models, identify the main themes addressed, what are the applications, and how they can assist in optimizing urban transportation systems, (ii) compare traditional discrete choice models with Machine Learning algorithms to analyze modal choice prediction using data from a Stated Preference (SP) survey conducted in Porto Alegre in 2019. The results of the systematic review indicate that Machine Learning methods are increasingly being used in transportation planning. Among the methods analyzed, traffic and public transport demand prediction models stood out as the most frequently used in the literature. Additionally, other methods such as traffic sign recognition, traffic signal detection, vehicle classification, pedestrian detection, travel time and itinerary planning, and comparative studies between different algorithms were also frequently used. The results of the comparative study indicate that the Multinomial Logit Model (MLM) presented significantly higher predictive accuracy compared to other Machine Learning models tested. The MLM accuracy rate was 52.03%, followed by the Random Forest (RF) method with 41.79%, and the Artificial Neural Networks (ANNs) with 40.94%. These results may be explained by the fact that the database used in the analysis contained few observations for the Lotação and Taxi transportation modes.

Keywords: modal choice, Flexible Transport, Multinomial Logit Model, Decision Tree, Random Forest, Artificial Neural Networks.

LISTA DE TABELAS

LISTA DE SIGLAS

AD – Árvore de Decisão

API – *Application Programming Interface*

CART – *Classification And Regression Tree*

DNNs – *Deep Neural Networks*

FA – Floresta Aleatória

GPS – *Global Positioning System*

ITS – *Intelligent Transportation System*

LASTRAN – Laboratório de Sistema de Transportes

LSTM – *Long Short-Term Memory*

ML – *Machine Learning*

MLM – Modelo Logit Multinomial

OD – Origem e Destino

RNAs – Redes Neurais Artificiais

RUM – *Random Utility Maximization*

SIT – Sistema Integrado de Transporte

TFL – *Transporto of London*

PD – Preferência Declarada

PPGEP – Programa de Pós Graduação da Engenharia de Produção

UFRGS – Universidade Federal do Rio Grande do Sul

USP – Universidade de São Paulo

SUMÁRIO

1 INTRODUÇÃO

A mobilidade urbana sustentável é essencial para o desenvolvimento da economia nacional. Grande volume de veículos particulares, congestionamentos, atrasos indesejáveis, acidentes, poluição do ar e sonora são alguns dos desafios enfrentados para alcançar os objetivos de mobilidade urbana sustentável de médio e longo prazo (Tamim Kashifi *et al.*, 2022).

O modelo de desenvolvimento urbano brasileiro prioriza investimentos em infraestrutura para o deslocamento de veículos particulares, como carros e motocicletas, em detrimento de modos ativos, como transporte público, a pé e por bicicleta (ITDP, 2022). Esse cenário se torna ainda mais crítico com o orçamento de R$ 6,05 bilhões previsto no Projeto de Lei Orçamentária Anual de 2023 (PLOA 2023) para o Ministério de Infraestrutura, o que representa uma diminuição média anual de 10% na última década.

Entre os principais componentes do planejamento de transporte urbano está a análise da escolha do modo de transporte. Os resultados da análise de escolha modal têm inúmeras aplicações, como previsão da demanda por viagens, elaboração de políticas e uma melhor compreensão das variáveis causais (Chang *et al.*, 2019).

Os modelos de escolha discreta baseados na teoria da Maximização da Utilidade Aleatória (RUM) são usados para analisar e modelar o comportamento de escolha modal. A literatura mostra que a escolha do modo de transporte é influenciada por muitos fatores tais como estilo de vida pessoal, profissão e ambiente social (Li *et al.*, 2019).

As formulações clássicas assumem que as utilidades são lineares, aditivas e incluem tanto características individuais como atributos alternativos (Cirillo e Xu, 2011). O modelo de escolha discreta mais utilizado é o Logit Multinomial (MLM) (Mcfadden, 1973), com uma estrutura matemática que facilita a estimação de parâmetros. Dessa forma, tem sido amplamente utilizado para a modelação de escolhas discretas na análise do comportamento de viagens.

Os algoritmos de *Machine Learning* (ML) se encontram em crescente desenvolvimento e aplicação na área de planejamento de transportes. Entre os métodos de Aprendizado de Máquina utilizados atualmente, destacam-se os de previsão demanda de tráfego e de transporte público, de reconhecimento de sinais de trânsito, de detecção de semáforos, de classificação de veículos,

de detecção de pedestres, de planejamento de tempo de viagem e de itinerários e comparativos de previsão entre algoritmos diferentes (Zhu et al., 2019).

Dado esse contexto, o emprego de novas tecnologias como a utilização de algoritmos de ML no planejamento do transporte urbano através de análises de previsão de modo de transporte para realizar o dimensionamento otimizado, apresentam elevado potencial de redução de custos e aumento da eficiência dos sistemas de transporte. O processo de descoberta de conhecimento a partir de dados compreende diversas etapas, que abrangem desde a organização e a limpeza dos dados até a validação e o emprego da informação produzida no processo de tomada de decisões (Peng *et al.*, 2022).

Nesse sentido, esse trabalho se propõe a alinhar as demandas atuais do planejamento de transportes ao analisar de forma quantitativa os estudos existentes, identificar os principais temas abordados e como eles podem auxiliar na otimização dos sistemas de transporte urbano. Além de propor o emprego de algoritmos de *Machine Learning* para comparar com modelos tradicionais de escolha discreta, a fim de analisar a previsão das preferências individuais por diferentes modos de transporte.

Esta dissertação é dividida em cinco partes: i) Introdução; ii) Diretrizes da Pesquisa; iii) Artigo 1: Planejamento do Transporte Urbano através de *Big Data* E *Machine Learning*: uma Revisão Sistemática da Literatura; iv) Artigo 2: Análise de Previsão de Escolha Modal: Comparativo entre Modelos Logit e Algoritmos Supervisionados de Aprendizado De Máquina; e v) Considerações Finais da Dissertação.

2 DIRETRIZES DA PESQUISA

As diretrizes para desenvolvimento do trabalho estão subdivididas em objetivos da pesquisa, justificativa, limitações e delineamento, os quais serão descritos nos próximos itens.

2.1 OBJETIVOS DA PESQUISA

Os objetivos da pesquisa estão classificados em geral e específicos e são descritos a seguir.

2.2.1 Objetivo Geral

O objetivo geral desta dissertação é identificar as aplicações de algoritmos de Machine Learning no planejamento de transporte urbano, bem como compará-los com os modelos de escolha discreta tradicionalmente utilizados na análise da escolha modal para avaliar a implantação de um novo modo de transporte.

2.2.2 Objetivos Específicos

Os objetivos específicos do trabalho são:

a) desenvolver uma revisão sistemática da literatura para analisar de forma quantitativa os estudos existentes sobre planejamento de transporte urbano com algoritmos de *Machine Learning*;

b) identificar os principais temas abordados, quais são as aplicações e como podem auxiliar no aprimoramento dos sistemas de transporte urbano;

c) comparar modelos tradicionais de escolha discreta com algoritmos de Machine Learning, a fim de analisar a previsão da escolha modal.

2.3 JUSTIFICATIVA

O trabalho tem por justificativa a importância de analisar questões sobre algoritmos de ML e sua potencial utilização no planejamento do transporte urbano. A realização de estudos comparativos entre modelos tradicionalmente aplicados e outros algoritmos de *Machine Learning* para a modelagem de escolha modal em transportes, podem auxiliar na identificação de qual algoritmo apresenta melhor desempenho em diferentes cenários e, consequentemente, qual é mais adequado para ser utilizado em projetos de planejamento urbano de transportes.

Cabe salientar que, a realização de estudos comparativos também contribui para o avanço do conhecimento na área, permitindo que novas técnicas sejam desenvolvidas e aprimoradas. Dessa forma, é possível encontrar soluções mais eficientes e sustentáveis para o transporte urbano, contribuindo para a melhoria da qualidade de vida da população e para o desenvolvimento das cidades.

2.4 LIMITAÇÕES

Este estudo limita-se em relação à disponibilidade de dados relativos à pesquisa de preferência declarada realizada de forma presencial com complemento *online* na cidade de Porto Alegre (RS, Brasil) em 2019. Os respondentes considerados aptos a participar foram aqueles que residem em Porto Alegre e/ou Região Metropolitana, com deslocamentos principais para a capital. Sendo assim, o estudo está limitado a essa população e localização geográfica.

2.5 DELINEAMENTO

Esta dissertação está dividida em cinco capítulos. O primeiro capítulo desta dissertação constituiu-se em uma **parte introdutória**, com uma introdução geral sobre os temas abordados. No segundo capítulo foram descritas as **diretrizes da pesquisa**, os objetivos gerais e específicos, as justificativas e as limitações da pesquisa. O terceiro capítulo é baseado no primeiro artigo que compõe esta dissertação, que consiste em uma **revisão sistemática da literatura** sobre o planejamento do transporte urbano através de técnicas de *Machine Learning*. O artigo no qual o capítulo se baseou foi aceito e publicado nos Anais do 36º Congresso da ANPET que foi realizado entre os dias 8 e 12 de novembro de 2022, em Fortaleza/CE, como requisito do regimento de mestrado acadêmico do PPGEP/UFRGS.

O quarto capítulo é baseado no segundo artigo, que realiza uma **análise comparativa** entre modelos de Logit e Aprendizado de Máquina para previsão de escolha modal. Este capítulo é dividido em três partes, que descrevem as técnicas utilizadas, os procedimentos metodológicos adotados e a análise dos resultados obtidos. Por fim, no quinto capítulo, são apresentadas as **considerações finais da dissertação**, que incluem uma discussão sobre os resultados alcançados, as limitações do estudo e possíveis sugestões de aperfeiçoamento para estudos futuros.

3 ARTIGO 1: PLANEJAMENTO DO TRANSPORTE URBANO ATRAVÉS DE BIG DATA E MACHINE LEARNING: UMA REVISÃO SISTEMÁTICA DA LITERATURA

RESUMO

A fim de analisar de forma quantitativa os estudos existentes sobre planejamento de transporte urbano com modelos de *Machine Learning* (ML) através de *Big Data* e identificar os principais temas abordados, analisar quais são as aplicações e como poderam auxiliar na otimização dos sistemas de transporte urbano foi desenvolvida uma revisão sistemática da literatura. Para a sua elaboração foi utilizado o *software* stArt, que divide a revisão em três etapas: planejamento, execução e sumarização. Quanto à análise dentre os artigos identificados na fase de seleção (2690), 67% são mais recentes que 2019, e entre os artigos aceitos na fase de extração (47), 84%. Na fase de seleção e extração, dos 47 artigos selecionados, 18 artigos (41%) foram publicados em 2020 e 15 artigos (34%) são ainda mais recentes, de 2021. Os resultados mostraram que o tema mais recorrente identificado foi previsão de demanda, seguido de previsão de tráfego.

Palavras-chave: Machine Learning, Big Data, Revisão Sistemática da Literatura, Modelos de Previsão, Planejamento de Transportes

ABSTRACT

In order to analyze quantitatively the existing studies on urban transportation planning with Machine Learning models through Big Data and identify the main topics addressed, analyze which are the applications and how they can assist in the optimization of urban transportation systems a systematic literature review was developed. For its elaboration, the stArt software was used, which divides the review into three stages: planning, execution, and summarization. As for the analysis among the articles identified in the selection phase (2690), 67% are more recent than 2019, and among the articles accepted in the extraction phase (47), 84%. In the selection and extraction phase, of the 47 articles selected, 18 articles (41%) were published in 2020 and 15 articles (34%) are even more recent, from 2021. The results showed that the recurrent theme identified was demand forecasting, followed by traffic forecasting.

Keywords: Machine Learning, Big Data, Systematic Literature Review, Prediction Models, Transportation Planning

3.1 INTRODUÇÃO

A partir do aumento substancial de automóveis em circulação nos centros urbanos, os problemas de tráfego tornaram-se recorrentes, o que leva a perda de recursos e tempo, bem como geram externalidades negativas na economia e no desempenho dos transportes (Wang *et al.*, 2016). Nesse sentido, o transporte público é um segmento indispensável do tráfego urbano, pois além de ser a artéria para assegurar a operacionalidade da vida e da produção na cidade, tem importância na melhoria de sua funcionalidade (Wang e Qing-dao-er-ji, 2018).

Para adaptar o gradual desenvolvimento das cidades com o crescente tráfego de veículos, torna-se necessário implementar tecnologias avançadas nos sistemas de transporte urbano, no controle de serviços e na fabricação de veículos, de modo a favorecer a integração modal e um sistema de transporte inteligente (*intelligent transportation system* – ITS) (Liu, 2018). As tecnologias incorporadas pelo ITS incluem: sensores eletrônicos, transmissão de dados, controle de operação, dentre outros. Esses dados são obtidos a partir de diversas fontes, tais como cartões de bilhetagem eletrônica, GPS, sensores, câmeras de vídeo e mídias sociais (Zhu *et al.*, 2019).

Devido ao aumento da celeridade da informação, ao longo das últimas décadas e ao avanço das cidades inteligentes (*smart cities)*, o volume de dados gerado dos sistemas de transporte cresceu de forma exponencial (Wang *et al.*, 2016). Nesse contexto, *Big Data* abrange a área do conhecimento que visa tratar, analisar e obter informações a partir desses conjuntos muito grandes de dados para serem analisados por sistemas tradicionais (Chen *et al.*, 2014). Já a inteligência analítica (*analytics*) é um campo abrangente e multidimensional que utiliza técnicas matemáticas, estatísticas, de modelagem preditiva e Aprendizado de Máquina para identificar padrões e conhecimentos significativos em *Big Data* (Hussein *et al.*, 2018).

No âmbito dos transportes, o aproveitamento de *Big Data* sob a noção dos ITS está em pauta, pois a utilização de dados de sensores eletrônicos, comunicações e GPS possibilita um futuro com sistemas mais eficientes (Kinra *et al.*, 2020). A literatura tem indicado que à medida que o número de algoritmos de ML aumenta, e as ferramentas computacionais tornam-se capazes de lidar com modelos sofisticados em conjunto com a diversidade de dados, há uma tendência crescente em combinar diferentes algoritmos para mitigar as desvantagens e resolver problemas complexos, agregando acurácia aos resultados e possivelmente num tempo de processamento menor (Kaffash *et al.*, 2021).

Tendo isso em vista, o presente artigo tem como questões quais são as aplicações de ML no planejamento do transporte urbano a partir de *Big Data* e como essas aplicações podem auxiliar na otimização dos sistemas de transporte urbano.

O planejamento da operação de transportes necessita um processamento integrado dos dados existentes. Mesmo que condições ideais para o gerenciamento sejam ainda consideradas de difícil consecução, é possível e desejável haver estudos para implementação de sistemas operacionais modernos (Senna, 2014). Sendo assim, o artigo tem por justificativa a importância de analisar massivos volumes de dados (*Big Data*) de ITS e a potencial influência de sua utilização no planejamento e na eficiência do transporte urbano através da comparação de diferentes modelos de ML.

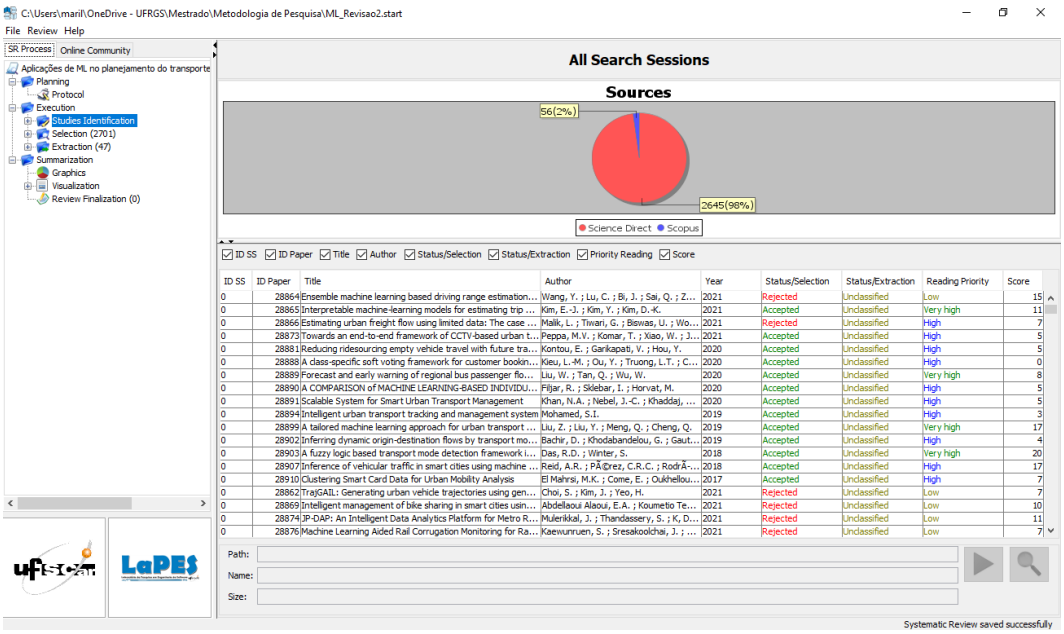
Este artigo foi organizado em cinco seções, sendo essas: i) Introdução; ii) Procedimentos metodológicos; iii) Descrição da literatura e contextualização dos algoritmos de ML com *Big Data* e comparação com modelos tradicionais; iv) Análise e discussão dos resultados; e v) Considerações finais, com sugestões para pesquisas futuras.

3.2 PROCEDIMENTOS METODOLÓGICOS

As revisões da literatura possuem um papel importante na pesquisa acadêmica, fornecendo uma visão geral, síntese e avaliação crítica de pesquisas anteriores, identificando ou construindo questões e problemas de pesquisa promissores (LePine e King, 2010). Neste estudo, a revisão segue conceitos que se diferem de uma revisão tradicional, por seguir uma série de diretrizes estritas, para garantir que a influência da subjetividade seja minimizada, que o processo de revisão seja replicável e que a revisão seja a mais compreensível possível (Haddaway *et al.*, 2015).

Para a elaboração da revisão foi utilizado o *software* stArt desenvolvido para dar suporte as revisões sistemáticas da literatura, acelerando o processo (Fabbri *et al.*, 2016). A ferramenta foi concebida de forma a seguir os conceitos apresentados por Tranfield *et al*. (2003) e Kitchenham e Charters (2007), que dividem a revisão sistemática em três etapas: planejamento, execução e sumarização, conforme apresentado na Figura 1.

Figura 1 - *Software* stArt para apoio a revisões sistemáticas



(fonte: Layout Software stArt)

Nas subseções a seguir cada uma dessas etapas é descrita detalhadamente, com enfoque para a utilização da ferramenta stArt.

3.2.1 Planejamento

A etapa de planejamento, desenvolvida com auxílio da ferramenta stArt, envolve a construção de um protocolo baseado nas diretrizes descritas por Kitchenham e Charters (2007), que apresenta como principais elementos a definição de questões de pesquisa, palavras-chave, bases de dados e critérios de seleção utilizados. A partir dos objetivos e contexto atuais apresentados na seção introdutória foram construídas as seguintes questões de pesquisa:

1. Quais são as aplicações de ML no planejamento do transporte urbano a partir de *Big Data*?

2. Como essas aplicações podem auxiliar na otimização dos sistemas de transporte urbano em comparativos com modelos estatísticos tradicionais?

Com o objetivo de responder as questões de pesquisa foram escolhidos termos que incorporassem estudos relativos a *Machine Learning*, *Big Data* e sistemas de transporte (*transport*), especificamente em relação ao tratamento de dados e utilização de algoritmos que

atendam o objetivo. Essas palavras chaves foram traduzidas no algoritmo de pesquisa: (Transport AND Urban AND (Machine Learning OR Big Data)).

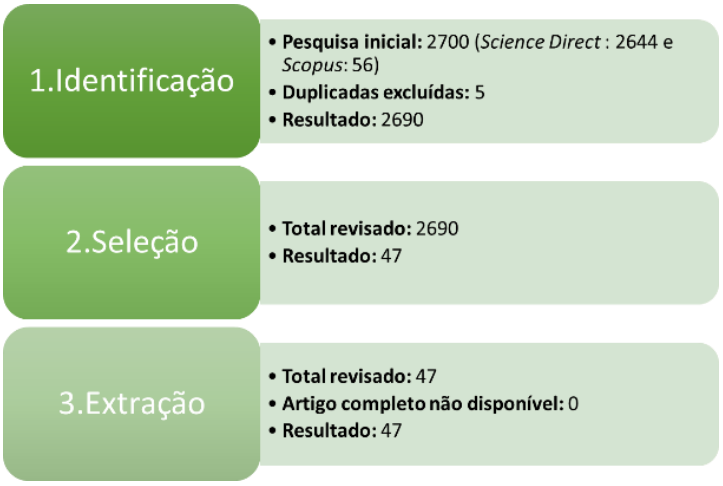
Para a elaboração dessa revisão, o algoritmo de pesquisa apresentado foi utilizado nas bases de dados eletrônicos Scopus e Science Direct, tendo em vista a disponibilidade de acesso e a abrangência de editoras que essas bases cobrem. A compatibilidade dessas bases com a ferramenta stArt também foi um fator decisivo para a escolha das mesmas. Com base nas questões de pesquisa construídas anteriormente, foram definidos os seguintes critérios de seleção:

1. O estudo apresenta alguma metodologia de ML em planejamento de transportes;
2. O estudo foi publicado no período de 2017 - 2021;
3. O estudo foi publicado em língua inglesa;
4. O estudo está em formato de artigo completo.

3.2.2 Execução

Na etapa de execução ocorre a revisão propriamente dita, dividindo-se em três fases: identificação, seleção e extração. A fase de identificação envolve a aplicação do algoritmo de pesquisa nas bases selecionadas, a consolidação dos resultados na ferramenta stArt e a remoção de artigos duplicados. Na fase seguinte são considerados os critérios de seleção na análise do título e *abstract* dos estudos identificados, para a seleção dos artigos relevantes, e posterior extração dos estudos por meio da leitura completa dos mesmos, conforme Figura 2.

Figura 2 - Fases relativas à etapa execução



(fonte: elaborado pela autora)

3.2.3 Sumarização

A última etapa da revisão envolve a sumarização e apresentação dos resultados obtidos. Inicialmente será desenvolvida uma análise descritiva da área de pesquisa, apresentando os principais autores, localização e metodologias apresentadas. Em seguida, será realizada uma descrição aprofundada dos resultados, bem como os desafios e oportunidades dessa área de pesquisa.

3.4 DESCRIÇÃO DA LITERATURA

A literatura internacional tem mostrado que a inteligência analítica de *Big Data* em ITS pode efetivamente melhorar a mobilidade urbana, visto que possibilita aos gestores de transportes uma visão ampla para avaliar interativamente desempenhos e resultados na tomada de decisões, bem como aprimorar o processo de planejamento (Wang *et al.*, 2016). A aplicabilidade dos algoritmos de ML é ampla, incluindo, porém não se limitando ao reconhecimento de sinais, detecção de objetos, previsão do fluxo de tráfego, planejamento de tempos de viagens, planejamento de rota de viagem e segurança viária (Kaffash *et al.*, 2021).

Entre os exemplos existentes na literatura, Antunes *et al.* (2019) utilizaram a tecnologias de *Big Data* (*Apache Spark*) para processar e analisar dados de operadores de transporte em Lisboa, a fim de identificar padrões de mobilidade urbana. Em Londres, o *Transport of London* (TfL) também utiliza plataformas de inteligência analítica nos dados provenientes de: (i) bilhetagem

eletrônica (*Oyster card*), (ii) GPS de cerca de 9.200 ônibus, (iii) 6.000 semáforos e (iv) 1.400 câmeras para auxiliar no planejamento de transporte urbano (Beeharry *et al.*, 2017).

Os algoritmos de ML podem ser classificados principalmente por técnicas paramétricas e não paramétricas (Yao *et al.*, 2017). Entre a variedade de estatísticas, os algoritmos paramétricos incluem os modelos de regressão (Rice e Van Zwet, 2004), autorregressivo integrado de médias móveis sazonal (SARIMA), autorregressivo integrado de médias móveis (ARIMA) (Fu *et al.* 2016), algoritmos de Filtro de Kalman (Zhou *et al.*, 2018), abordagens não paramétricas, como Redes Neurais Artificiais (RNA) (Khodayari *et al.* ,2012), algoritmos K-Vizinhos Mais Próximos (KNN) (Yu *et al.*, 2016), Máquina de Suporte Vetorial (MVS) (Sun *et al.*, 2015), algoritmos genéticos (Lopez-Garcia *et al.*, 2015) e *Deep Learning* (DL) (Koesdwiady *et al.*, 2016). Nos últimos anos, o uso combinado desses algoritmos para resolver problemas de ITS tornou-se mais prevalente (Zhao *et al.*, 2017).

A previsão da demanda é um tema emergente no âmbito dos transportes devido à sua importância no planejamento, controle e gestão. Algoritmos como ARIMA e métodos recentes de Aprendizado Profundo, como *Long Short-Term Memory* (LSTM) são geralmente eficazes para a análise preditiva de variáveis em séries temporais (Kieu *et al.*, 2020). Cabe salientar sobre os desafios que existem em aberto sobre a previsão de tráfego de veículos baseada em modelos de ML e com implementação no mundo real. Destaca-se ainda que o consumo de tempo desprendido para o treinamento de modelos de DL, bastante utilizado em predições é relativamente grande quando comparado a modelos paramétricos como ARIMA e SARIMA (Boukerche e Wang, 2020).

No Reino Unido, uma solução heurística que integra simulação, ML e otimização dentro de uma estrutura interativa foi utilizada para fornecer aproximações rápidas de tempos de travessia de estradas em horários diferentes do dia (Bayliss, 2021). Já em Pequim, para investigar a tendência de conjunto que se beneficia de diferentes modelos e propor uma abordagem com um conjunto de percepção de condições de tráfego foram aplicados algoritmos nos dados da rede de detectores de tráfego. O objetivo foi de capturar padrões espaciais embutidos no fluxo de tráfego e encontrar resultados experimentais capazes de auxiliar na melhora do desempenho da previsão do fluxo de tráfego (Chen *et al.*, 2020).

Um caso de estudo em Shenzhen, na China foi conduzido utilizando o algoritmo de ML com Máquinas de Suporte Vetorial Regressiva (MSVR) e com Filtro de Kalman (K-MSVR) para

criar um algoritmo de previsão automatizada do tempo de chegadas de ônibus. Ao ajustar os parâmetros do Filtro de Kalman, os resultados preditivos foram mais adequados para evitar a frequência de ajustes desnecessários na tabela horária em comparação com o método tradicional. A readequação dos horários pode alocar de forma mais efetiva a oferta de viagens e demonstrou desempenho superior nos indicadores de flutuação de progresso, da taxa de desempenho de intervalos e de cumprimento do cronograma proposto (Zhang *et al.*, 2021).

Através de *Big Data* proveniente de telefonia celular foram calculados os tempos de viagem entre as zonas de transporte estabelecidas usando a *Application Programming Interface* (API) do *Google* e gerando a matriz OD de viagens com os registros de telefonia. Os cenários foram utilizados para analisar a acessibilidade dinâmica e a influência dos componentes. Com a análise de *cluster* foi feita a caracterização das zonas de transporte de acordo com a acessibilidade em cenários e horários. Os resultados indicaram que essas novas fontes de dados de geolocalização apresentam um potencial considerável para uso em estudos de acessibilidade, dado que produzem informações mais precisas e realistas do que as análises estáticas ou parcialmente dinâmicas. Sendo assim, as informações consolidadas puderam auxiliar os tomadores de decisões a respeito de questões sobre transportes e uso do solo (Liu *et al.*, 2020).

Estudos recentes têm mostrado que algoritmos de ML com métodos não paramétricos são capazes de alcançar maior precisão preditiva que técnicas estatísticas tradicionais com fundamentação teórica matematicamente comprovada. Dessa forma, tornam-se necessários estudos comparativos que abordem as lacunas existentes entre esses modelos. Na Califórnia foi desenvolvido um estudo comparativo entre sete modelos (Bayes Ingênuo, Árvores de Classificação e Regressão, Aprimoramento, Agregação de *Bootstrap*, Floresta Aleatória, Máquina de Vetores de Suporte e Redes Neurais Artificiais) de ML e dois modelos Logit (MLM e Logit misto), com objetivo de analisar as principais diferenças no desenvolvimento, avaliação e interpretação comportamental da escolha por modo de viagem (Zhao *et al.*, 2020).

Já na Holanda, através de um conjunto de dados amostrado resultante de 69.918 viagens individuais e um total de 230.608 viagens diárias foi desenvolvido um estudo comparativo entre a performance de previsão com modelos de ML também para análise de escolha modal. Inicialmente, apresentou-se uma comparação abrangente de sete classificadores de ML e se avaliou sistematicamente os classificadores usando técnicas estritas de validação de modelo e estatísticas de teste. Além das características individuais e domiciliares, consideraram-se as

características do ambiente construído e natural, bem como as condições meteorológicas para a construção do modelo. Por fim, investigou-se detalhadamente a importância de cada variável para cada modelo classificador e para cada modo de viagem (Hagenauer e Helbich, 2017).

Dois conjuntos de dados suíços provenientes de bases com tamanho e natureza distintos foram utilizados para comparar a performance entre algoritmos de *Random Utility Maximization* (RUM), de ML e *Deep Neural Networks* (DNNs), visto que estudos que utilizam um único conjunto de dados podem apresentar estimativas de variância tendenciosas devido as dependências na amostra extraída do conjunto de dados (García-García *et al.*, 2022).

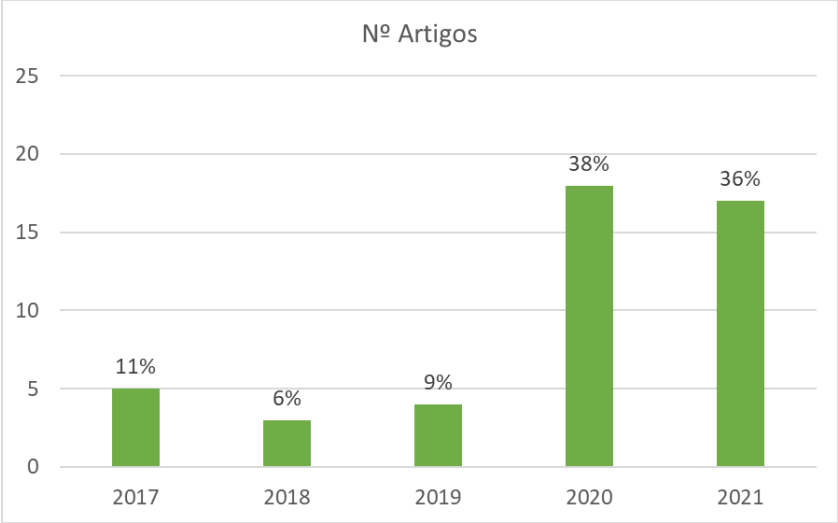
No contexto brasileiro, a partir dos dados de Fortaleza rodados no modelo de programação *MapReduce* com ML foram identificadas regiões de maior demanda na cidade, servindo como indicador para o planejamento das linhas e alocação das paradas de ônibus, além de padrões e relações dos atrasos na tabela horária (Wang *et al.*, 2016). Dada a necessidade de melhoria da qualidade dos serviços de transporte coletivo, um estudo desenvolvido com método de abordagem por avaliação multicritério através dos dados do Sistema Integrado de Transporte de Florianópolis (SIT) foi possível identificar os fatores objetivos e subjetivos que determinam a opinião dos passageiros sobre os serviços e compreender como eles avaliam o sistema, possibilitando análises para melhorias (Barbosa *et al.*, 2017).

3.5 ANÁLISE E DISCUSSÃO DOS RESULTADOS

A presença do assunto "*Big Data* e *Machine Learning* nos sistemas de transportes" na literatura acadêmica teve um aumento considerável nos últimos anos, principalmente com o avanço exponencial da tecnologia e pelo surgimento de novos algoritmos voltados a solução de problemas de demanda por transportes. Entre os artigos identificados na fase de seleção (2690), 67% são mais recentes que 2019, e entre os artigos aceitos na fase de extração (47), 84%.

Na fase de seleção e extração, de acordo com a Figura 3 dos 47 artigos selecionados, 18 artigos (38%) foram publicados em 2020 e 17 artigos (36%) são ainda mais recentes, de 2021. Esses resultados se devem ao fato do filtro de pesquisa de publicações, que exigia publicações datadas dos últimos cinco anos, aplicado com o objetivo de agregar contemporaneidade ao estudo, uma vez que os assuntos estudados possuem celeridade de atualizações.

Figura 3 - Número de artigos por ano de publicação



(fonte: elaborado pela autora)

Tendo em vista que o escopo dessa revisão são estudos que utilizam Aprendizado de Máquina com *Big Data* provenientes de ITS para o planejamento do transporte urbano, a Tabela 1 apresenta, de forma quantitativa, a presença desses termos citados em estudos desse tipo de literatura. Os termos "*Machine Learning"* e "*public transportation*" são os mais citados, o que traduz sua relevância no planejamento de transportes.

Tabela 1 - Quantidade de estudos em que cada termo foi citado

Termo	**Quantidade de estudos**
Machine learning (ML)	16
Public transport	12
Urban transportation	11
Inteligent Transportation Systems (ITS)	8
Big data	8
Deep learning (DL)	7

(fonte: elaborado pela autora)

Quanto à disposição da frequência de palavras-chave dos artigos analisados foi desenvolvida a Figura 4, caracterizada por uma Nuvem de Palavras, a qual sintetiza os termos por recorrência de aparição nos artigos. Ao analisar o resultado, verifica-se que os termos *Machine Learning, urban transportation, Big Data e public transport* possuem notável relevância entre os artigos analisados.

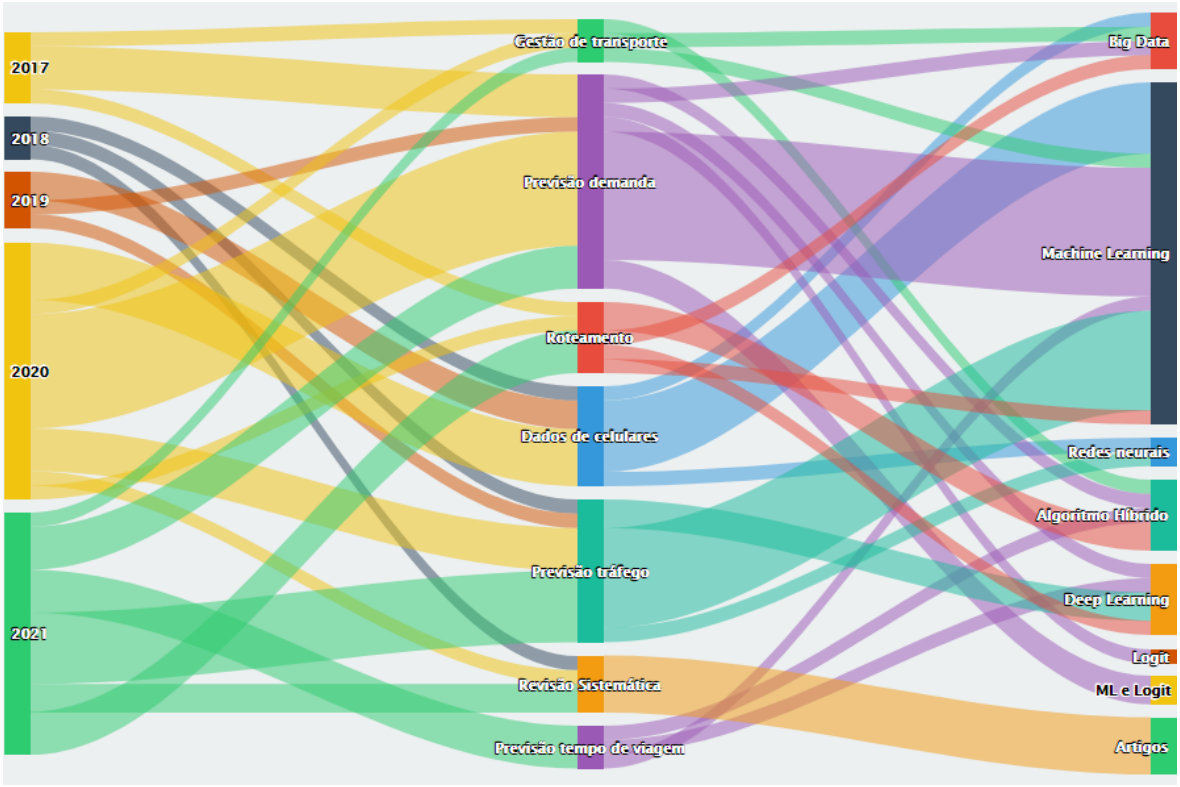
Figura 4 - Nuvem de Palavras das palavras-chave

(fonte: elaborada pela autora)

A explicação para isso pode vir a ser o fato de as técnicas de ML estarem em crescente desenvolvimento e aplicação, inclusive nas áreas de planejamento de transportes. Entre os métodos de aprendizado de máquina utilizados nos artigos analisados, destacam-se os modelos de previsão de tráfego, de demanda por transporte público, reconhecimento de sinais de trânsito, detecção de semáforos, classificação de veículos, detecção de pedestres, planejamento de tempo de viagem e de itinerários e comparativos de previsão entre os modelos.

A Figura 5 resume os resultados descritos na seção anterior, por meio da classificação dos métodos e de acordo com o tema que tenha sido realizado nos estudos em questão, através do Diagrama de *Sankey*. Além da classificação, os outros elementos que foram comuns entre estudos foram identificados pelos métodos utilizados.

Figura 5 - Classificação dos métodos por tema e ano segundo os estudos identificados



(fonte: elaborada pela autora)

Dentre os estudos analisados, o tema mais recorrente identificado foi previsão de demanda com 15 artigos, que corresponde a 32% dos artigos selecionados na fase da extração. A respeito dos métodos utilizados nesse tema destacam-se modelos de ML, modelos Logit, incluindo comparações entre modelos, além de análise de *cluster*, algoritmos híbridos e DL. Entre os assuntos de demanda presentes nos artigos analisados estão previsão de demanda de passageiros de transporte coletivo, de ônibus escolares, de *ridesoursing* e artigos com previsão de modo de viagem.

Na sequência, encontram-se os estudos sobre previsão de tráfego, que englobam 21% dos artigos (10) e possuem métodos de aplicabilidade com predominância de DL e ML, além de Redes Neurais Artificiais. Sobre previsão de tempo de viagem alguns algoritmos de ML utilizados foram MVS, Filtros de Kalman, regressões e LSTM. Essas previsões possuem considerável importância, pois servem de embasamento para o planejamento das operações de transportes.

Quanto aos temas de roteamento, os assuntos englobam otimização de rotas para aplicativos de transporte, itinerários de ônibus urbanos e escolares. Para as questões de roteamento foram

aplicados modelos de solução heurística, algoritmo genético com dados georreferenciados e algoritmos híbridos. Sobre a utilização de dados de celulares e GPS, os artigos analisados apresentam modelos de classificação de mobilidade individual e trajetórias, através do processamento de grande volume de dados.

Ainda foram classificados artigos contendo temas de simulação de tráfego com solução heurística e que englobam noções gerais de gestão de transporte com métodos menos tradicionais de ML. Entre os estudos selecionados foram identificadas outras revisões sistemáticas da literatura, incluindo artigos com relação entre *Big Data* e ITS, sobre busca de soluções de problemas de congestionamento com modelos modernos e previsão de tráfego. Sobre comparativos entre modelos de ML, modelos Logit, redes neurais e outros modelos estatísticos tradicionais foram analisados artigos com temas de previsão de modelos de escolha de modo de viagem.

3.6 CONSIDERAÇÕES FINAIS

O presente artigo abrangeu a revisão sistemática da literatura, cujo tema disserta sobre a as aplicações de aprendizado de máquina a partir de *Big Data* no planejamento do transporte urbano e de como essas aplicações podem auxiliar na otimização dos sistemas de transporte urbano. Em virtude de as técnicas de ML estarem em crescente desenvolvimento e aplicação, inclusive nas áreas de planejamento de transportes.

Para o desenvolvimento foi utilizado o *software* stArt criado para dar suporte as revisões sistemáticas da literatura e o procedimento foi dividido em três etapas: planejamento, execução e sumarização. A partir da definição das questões de pesquisa foram escolhidos termos que incorporassem estudos relativos a *Machine Learning*, *Big Data* e sistemas de transporte (*transport*), especificamente em relação ao tratamento de dados e uso de algoritmos que atendam o objetivo. As palavras chaves foram traduzidas no algoritmo de pesquisa: (Transport AND Urban AND (Machine Learning OR Big Data)). O algoritmo de pesquisa apresentado foi utilizado nas bases Scopus e Science Direct, tendo em vista a disponibilidade de acesso e a abrangência de editoras que essas bases cobrem.

Dentre os artigos identificados na fase de seleção (2690), 67% e entre os artigos aceitos na fase de extração (47), 84% são mais recentes que 2019. Dos 47 artigos selecionados na fase de

seleção e extração, 18 artigos (41%) foram publicados em 2020 e 15 artigos (34%) são ainda mais recentes, de 2021.

Sobre os métodos de aprendizado de máquina mais utilizados nos artigos analisados, destacaram-se os modelos de previsão de demanda de tráfego, de demanda por transporte público, de reconhecimento de sinais de trânsito, de detecção de semáforos, de classificação de veículos, de detecção de pedestres, de planejamento de tempo de viagem e de itinerário.

REFERÊNCIAS

ANTUNES, H., FIGUEIRAS, P., COSTA, R., TEIXEIRA, J., E JARDIM-GONCALVES, R. Analysing Public Transport data through the use of Big Data tecnhologies for urban mobility. **2019 International Young Engineers Forum (YEF-ECE)** (p. 40–45). Apresentado em 2019 International Young Engineers Forum (YEF-ECE), IEEE, Costa da Caparica, Portugal, 2019. Disponível em: doi:10.1109/YEF-ECE.2019.8740816

BARBOSA, S. B., FERREIRA, M. G. G., NICKEL, E. M., CRUZ, J. A., FORCELLINI, F. A., GARCIA, J., E GUERRA, J. B. S. O. DE A. Multi-criteria analysis model to evaluate transport systems: An application in Florianópolis, Brazil. **Transportation Research Part A: Policy and Practice**, 96, 1–13, 2017. Disponível em: doi:10.1016/j.tra.2016.11.019

BAYLISS, C. Machine learning based simulation optimisation for urban routing problems. **Applied Soft Computing**, 105, 107269, 2021. Disponível em: doi:10.1016/j.asoc.2021.107269

BEEHARRY, Y., FOWDUR, T. P., HURBUNGS, V., BASSOO, V., E RAMNARAIN-SEETOHUL, V. Analysing Transportation Data with Open Source Big Data Analytic Tools. **Indonesian Journal of Electrical Engineering and Informatics (IJEEI)**, 5(2), 174–184, 2017. Disponível em: doi:10.11591/ijeei.v5i2.297

BOUKERCHE, A., E WANG, J. A performance modeling and analysis of a novel vehicular traffic flow prediction system using a hybrid machine learning-based model. **Ad Hoc Networks**, 106, 102224, 2020. Disponível em: doi:10.1016/j.adhoc.2020.102224

CHEN, Y., LV, Y., YE, P., E ZHU, F. Traffic-Condition-Awareness Ensemble Learning for Traffic Flow Prediction. **IFAC-PapersOnLine**, 53(5), 582–587, 2020. Disponível em: doi:10.1016/j.ifacol.2021.04.146

CHEN, M., MAO, S., E LIU, Y. (2014) Big Data: A Survey. Mobile Networks and Applications, 19(2), 171–209, 2014. Disponível em: doi:10.1007/s11036-013-0489-0

FABBRI, S., SILVA, C., HERNANDES, E., OCTAVIANO, F., DI THOMMAZO, A., E BELGAMO, A. Improvements in the StArt Tool to Better Support the Systematic Review Process. **Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering** (p. 21:1–21:5). ACM, New York, NY, USA, 2016. Disponível em: doi:10.1145/2915970.2916013

FU, R., ZHANG, Z., LI, L. Using LSTM and GRU neural network methods for traffic flow prediction. In: **2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)**. IEEE, pp. 324–328, 2016. Disponível em: doi:10.1109/YAC.2016.7804912

GARCÍA-GARCÍA, J. C., GARCÍA-RÓDENAS, R., LÓPEZ-GÓMEZ, J. A., E MARTÍN-BAOS, J. Á. A comparative study of machine learning, deep neural networks and random utility maximization models for travel mode choice modelling. **Transportation Research Procedia**, 62, 374–382, 2022. Disponível em: doi:10.1016/j.trpro.2022.02.047

HADDAWAY, N. R., WOODCOCK, P., MACURA, B., E COLLINS, A. Making literature reviews more reliable through application of lessons from systematic reviews: Making Literature Reviews More Reliable. **Conservation Biology**, 29(6), 1596–1605, 2015. Disponível em: doi:10.1111/cobi.12541

HAGENAUER, J., E HELBICH, M. A comparative study of machine learning classifiers for modeling travel mode choice. **Expert Systems with Applications**, 78, 273–282, 2017. Disponível em: doi:10.1016/j.eswa.2017.01.057

HUSSEIN, W. N., KAMARUDIN, L. M., HUSSAIN, H. N., ZAKARIA, A., BADLISHAH AHMED, R., E ZAHRI, N. A. H. The Prospect of Internet of Things and Big Data Analytics in Transportation System. **Journal of Physics: Conference Series**, 1018, 012013, 2018. Disponível em: doi:10.1088/1742-6596/1018/1/012013

KAFFASH, S., NGUYEN, A. T., E ZHU, J. Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis. **International Journal of Production Economics**, 231, 107868, 2021. Disponível em: doi:10.1016/j.ijpe.2020.107868

KIEU, L.-M., OU, Y., TRUONG, L. T., E CAI, C. A class-specific soft voting framework for customer booking prediction in on-demand transport. **Transportation Research Part C: Emerging Technologies**, 114, 377–390, 2020. Disponível em: doi:10.1016/j.trc.2020.02.010

KITCHENHAM, B. AND CHARTERS, S. Guidelines for Performing Systematic Literature Reviews in Software Engineering. **Technical Report. Keele University and University of Durham**, version 2.3, 2007.

KINRA, A., BEHESHTI-KASHI, S., BUCH, R., NIELSEN, T. A. S., E PEREIRA, F. Examining the potential of textual big data analytics for public policy decision-making: A case study with driverless cars in Denmark. **Transport Policy**, 98, 68–78, 2020. Disponível em: doi:10.1016/j.tranpol.2020.05.026

KHODAYARI, A., GHAFFARI, A., KAZEMI, R., BRAUNSTINGL, R. A modified car-following model based on a neural network model of the human driver effects, **IEEE Transactions on Systems. Man, and Cybernetics-Part A: Systems and Humans** 42 (6), 1440–1449, 2012. Disponível em: doi:10.33142/sca.v2i5.813

KOESDWIADY, A., SOUA, R., KARRAY, F. Improving traffic flow prediction with weather information in connected cars: a deep learning approach. **IEEE Trans. Veh. Technol.** 65 (12), 9508–9517, 2016. Disponível em: doi:10.1109/TVT.2016.2585575

LEPINE, J. A., E KING, A. W. (EDS). Editors'comments: Developing Novel Theoretical Insight from Reviews of Existing Theory and Research. **Academy of Management Review**, 35(4), 506–509, 2010. Disponível em: doi:10.5465/amr.35.4.zok506

LIU, Y. Big Data Technology and Its Analysis of Application in Urban Intelligent Transportation System. **2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)** (p. 17–19), IEEE, Xiamen, China, 2018. Disponível em: doi:10.1109/ICITBS.2018.00012

LIU, W., TAN, Q., E WU, W. Forecast and Early Warning of Regional Bus Passenger Flow Based on Machine Learning. **Mathematical Problems in Engineering, 2020,** 1–11, 2020. Disponível em: doi:10.1155/2020/6625435

LIU, Z., LIU, Y., MENG, Q., E CHENG, Q. A tailored machine learning approach for urban transport network flow estimation. **Transportation Research Part C: Emerging Technologies,** 108, 130–150, 2019. Disponível em: doi:10.1016/j.trc.2019.09.006

LOPEZ-GARCIA, P., ONIEVA, E., OSABA, E., MASEGOSA, A.D., PERALLOS, A. A hybrid method for short-term traffic congestion forecasting using genetic algorithms and cross entropy. **IEEE Trans. Intell. Transport.** Syst. 17 (2), 557–569, 2015. Disponível em: doi: 10.1109/TITS.2015.2491365

RICE, J., E VAN ZWET, E. A Simple and Effective Method for Predicting Travel Times on Freeways. **IEEE Transactions on Intelligent Transportation Systems,** 5(3), 200–207, 2004. doi:10.1109/TITS.2004.833765

SENNA, L. A. DOS S. Economia e planejamento dos transportes, 2014. Obtido de http://www.sciencedirect.com/science/book/9788535277364

SUN, Y., LENG, B., GUAN, W. A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system. **Neurocomputing** 166, 109–121, 2015. Disponível em: doi: 10.1016/j.neucom.2015.03.085

TRANFIELD, D., DENYER, D., E SMART, P. Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. ***British Journal of Management,*** 14(3), 207–222, 2003. Disponível em: doi:10.1111/1467-8551.00375

WANG, X., E QING-DAO-ER-JI, R. Application of optimized genetic algorithm based on big data in bus dynamic scheduling. **Cluster Computing**, 2018. Disponível em: doi:10.1007/s10586-018-2625-x

WANG, Y., RAM, S., CURRIM, F., DANTAS, E., E SABOIA, L. A. A big data approach for smart transportation management on bus network. **2016 IEEE International Smart Cities Conference (ISC2),** (p. 1–6), IEEE, Trento, Italy, 2016. Disponível em: doi:10.1109/ISC2.2016.7580839

YAO, B., CHEN, C., CAO, Q., JIN, L., ZHANG, M., ZHU, H., E YU, B. Short-Term Traffic Speed Prediction for an Urban Corridor: Short-term traffic speed prediction for an urban corridor. **Computer-Aided Civil and Infrastructure Engineering**, 32(2), 154–169, 2017. Disponível em: doi:10.1111/mice.12221

YU, B., SONG, X., GUAN, F., YANG, Z., YAO, B. k-Nearest neighbor model for multiple-time-step prediction of short-term traffic condition. **J. Transport. Eng.** 142 (6), 2016. Disponível em: doi:10.1061/(ASCE)TE.1943-5436.0000816

ZHANG, X., YAN, M., XIE, B., YANG, H., E MA, H. An automatic real-time bus schedule redesign method based on bus arrival time prediction. **Advanced Engineering Informatics**, 48, 2021. Disponível em: doi:10.1016/j.aei.2021.101295

ZHAO, X., YAN, X., YU, A., E VAN HENTENRYCK, P. Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. **Travel Behaviour and Society**, 20, 22–35, 2020. Disponível em: doi:10.1016/j.tbs.2020.02.003

ZHAO, Z., CHEN, W., WU, X., CHEN, P.C., LIU, J. LSTM network: a deep learning approach for short-term traffic forecast. **IET Intell. Transp. Syst.** 11 (2), 68–75, 2017. doi: 10.1049/iet-its.2016.0208

ZHOU, Z., YU, H., XU, C., ZHANG, Y., MUMTAZ, S., E RODRIGUEZ, J. Dependable Content Distribution in D2D-Based Cooperative Vehicular Networks: A Big Data-Integrated Coalition Game Approach. **IEEE Transactions on Intelligent Transportation Systems**, 19(3), 953–964, 2018. Disponível em: doi:10.1109/TITS.2017.2771519

ZHU, L., YU, F. R., WANG, Y., NING, B., E TANG, T. Big Data Analytics in Intelligent Transportation Systems: A Survey. **IEEE Transactions on Intelligent Transportation Systems,** 20(1), 383–398, 2019. Disponível em: doi:10.1109/TITS.2018.2815678

4 ARTIGO 2: ANÁLISE DE PREVISÃO DE ESCOLHA MODAL: COMPARATIVO ENTRE MODELOS LOGIT E ALGORITMOS SUPERVISIONADOS DE APRENDIZADO DE MÁQUINA

RESUMO

O objetivo deste estudo é comparar modelos tradicionais de escolha discreta com algoritmos de Aprendizado de Máquina *(Machine Learning*), a fim de analisar a previsão da escolha modal. Para isso, foram estimados modelos Logit Multinomiais (MLM), comumente utilizados na determinação da demanda por transportes. Em seguida, os resultados foram comparados a três algoritmos de Aprendizado de Máquina: Árvore de Decisão (AD), Floresta Aleatória (FA) e Redes Neurais Artificiais (RNAs). Os dados utilizados são provenientes de uma pesquisa de Preferência Declarada (PD), realizada em Porto Alegre (RS, Brasil) em 2019. A pesquisa analisou a escolha em relação a seis modos de transporte, sendo cinco modos existentes e um modo de transporte público sob demanda, hipotético. Foram utilizados projetos experimentais eficientes com 7 atributos (relativos a custos, tempos, confiabilidade e clima). Os resultados indicaram que o modelo Logit teve melhor desempenho, com uma acurácia de 52,03%, seguido pelo Floresta Aleatória com 41,79% e pelo Redes Neurais Artificiais com 40,94%. Esse resultado pode ser devido às escolhas dos indivíduos pesquisados, apresentando baixa participação para os modos Lotação e Táxi. Para trabalhos futuros, recomenda-se comparar outros conjuntos de dados, comparar com outras estruturas de modelos de escolha discreta, como modelos mistos por exemplo, e novos algoritmos de Aprendizado de Máquina.

Palavras-chave: escolha modal, Modelo Logit Multinomial, Redes Neurais Artificiais, Árvore de Decisão, Floresta Aleatória

ABSTRACT

The objective of this study is to compare traditional discrete choice models with Machine Learning algorithms in order to analyze modal choice prediction. To this end, Multinomial Logit models (MNL), commonly used in determining transportation demand, were estimated. Then, the results were compared to three Machine Learning algorithms: Classification And Regression Tree (CART), Random Forest (RF) and Artificial Neural Networks (ANNs). The data used comes from a Stated Preference (DP) survey conducted in Porto Alegre (RS, Brazil) in 2019. The survey analyzed the choice regarding six transportation modes, being five existing

modes and one on-demand, hypothetical public transportation mode. Efficient experimental designs with 7 attributes (regarding costs, times, reliability, and climate) were used. The results indicated that the Logit model performed best, with an accuracy of 52.03%, followed by the Random Forest with 41.79% and the Artificial Neural Networks with 40.94%. This result may be due to the choices of the individuals surveyed, presenting low participation for the Lotation and Taxi modes. For future work, it is recommended to compare other data sets, compare with other discrete choice model structures, such as mixed models for example, and new Machine Learning algorithms.

Keywords: model choice, Multinomial Logit Model, Artificial Neural Networks, Decision Tree, Random Forest

4.1 INTRODUÇÃO

O transporte urbano é uma questão que afeta diretamente áreas da sociedade como educação, economia e saúde. Impasses no trânsito impactam negativamente nos deslocamentos, além da falta de vagas de estacionamento, do excesso de veículos individuais, da poluição e da diminuição da segurança, que são alguns problemas que as cidades enfrentam (Serin *et al.*, 2022).

Um dos componentes chave do planejamento de transporte urbano é a análise da escolha do modo de viagem, considerando fatores como tempo de viagem, custo, conforto e segurança. Esses fatores podem ser influenciados por variáveis como distância percorrida, horário, perfil do usuário e características do modo. A análise da escolha modal tem inúmeras aplicações, como previsão de demanda por viagens, formulação de políticas de transporte e auxílio na compreensão das variáveis causais (Tamim Kashifi *et al.*, 2022).

O desenvolvimento de novas tecnologias e meios de transporte com sistemas modernizados de processamento de dados apresentam um aumento no potencial de mudança na forma como a população se movimenta nos centros urbanos (UN-Habitat, 2022). Dessa forma, torna-se fundamental que os órgãos de planejamento dos transportes se atentem a tais mudanças e busquem as ferramentas adequadas para oferecer serviços mais eficientes (Diallo *et al.*, 2022).

A previsão das preferências individuais por modo de transporte e as mudanças induzidas no comportamento de viagem são fundamentais para o planejamento. Tradicionalmente, a

pesquisa de comportamento de viagem tem sido apoiada por modelos de escolha discreta, principalmente da família Logit, como o modelo Logit Multinomial (MLM).

O Aprendizado de Máquina se tornou difundido em muitos campos, com interesse na sua aplicação para modelar o comportamento de escolha do modo de viagem. Algoritmos populares não são paramétricos, como Árvore de Decisão (AD), Floresta Aleatória (FA) e Redes Neurais Artificiais (RNAs). Enquanto os modelos Logit pressupõem um certo tipo de estrutura dos dados por comportamento e estatísticas, os de ML "permitem que os dados falem por si", sendo mais flexíveis e podendo aumentar a capacidade preditiva (Zhao *et al.*, 2020).

Entre as desvantagens dos algoritmos de Aprendizado de Máquina a serem consideradas está a dependência de dados de treinamento para aprender e tomar decisões, pois se os dados de treinamento forem tendenciosos, incompletos ou imprecisos, o algoritmo pode produzir resultados incorretos ou enviesados. Outra desvantagem é a complexidade, que dificulta a interpretação, podendo impedir a detecção de erros ou vieses nos resultados. Ainda há falta de transparência ("caixa preta"), que engloba a falta de entendimento de como os algoritmos tomam decisões (Behrooz e Hayeri, 2022).

O objetivo deste estudo é comparar modelos tradicionais de escolha discreta com algoritmos de Aprendizado de Máquina, a fim de analisar a previsão da escolha modal, utilizando dados provenientes de uma pesquisa de Preferência Declarada (PD) realizada em Porto Alegre em 2019. Modelos Logit Multinomiais foram estimados, os quais são tradicionalmente utilizados na determinação da demanda por transportes. Os resultados obtidos foram comparados com três algoritmos de Aprendizado de Máquina: Árvore de Decisão (AD), Floresta Aleatória (FA) e Redes Neurais Artificiais (RNAs).

4.2 TÉCNICAS UTILIZADAS

As subseções a seguir descrevem as técnicas utilizadas nesse artigo.

4.2.1 Modelo de escolha discreta

O Modelo Logit Multinomial (McFadden, 1973) é o modelo de escolha discreta tradicionalmente mais utilizado, com amplo emprego no contexto de escolha do modo de viagem. A ideia básica consiste que o comportamento possui uma busca de maximização de

utilidade, um indivíduo n decide selecionar uma opção de um conjunto de alternativas discretas, avaliando seus atributos associados Xj a cada alternativa, com o objetivo de maximizar a utilidade (Lee *et al.*, 2018).

$$U_{nj} = \alpha_j + \beta_j X_{nj} + \epsilon_{nj} \quad (1)$$

Como exposto na Equação (1), U_{nj} é uma função de variáveis preditivas que determine a escolha do j modo de viagem pelo n indivíduo, αj pode ser visto como intercepto para a j^a alternativa; enquanto $\boldsymbol{\beta} j$ é um vetor de parâmetros do modelo (coeficientes), Xnj é um vetor de observável características (variáveis independentes), e ϵnj representa o componente não observado para o usuário específico e o respectivo modo de viagem. Além disso, o modelo MLM clássico assume que os ϵnj são independentes e identicamente distribuídos seguindo uma distribuição de Gumbel. Assim, a probabilidade de escolher alternativa i para individual n é

$$P_{ni} = \frac{e^{\alpha_i + \beta_i X_{ni}}}{\sum_{j=1}^{J} e^{\alpha_j + \beta_j X_{nj}}} \quad (2)$$

Assim, por causa da suposição independente e identicamente distribuída dos termos de erro e dado o vetor de coeficientes $\boldsymbol{\beta}$, a densidade conjunta de todas as realizações de indivíduos e escolhas, para o MLM método, pode ser descrito pela seguinte função de verossimilhança:

$$\begin{aligned} \mathbf{L}(\boldsymbol{\beta}) &= \prod_{n=1}^{N} \prod_{i=1}^{J} \left[\frac{e^{\alpha_i + \boldsymbol{\beta}_i X_{ni}}}{\sum_{j=1}^{J} e^{\alpha_j + \boldsymbol{\beta}_j X_{nj}}} \right]^{y_{ni}} \\ &= \prod_{n=1}^{N} \prod_{i=1}^{J} (P_{ni}^{y_{ni}}), \end{aligned} \quad (3)$$

Onde yni é igual a um se o indivíduo n escolher alternativa i e 0 caso contrário. Uma prática comum é tomar o logaritmo natural de a Eq. (3) para simplificar a matemática e os cálculos, sendo o resultado equação conhecida como a função de log-verossimilhança.

$$\mathbf{LL}(\boldsymbol{\beta}) = \sum_{n=1}^{N} \sum_{i=1}^{J} y_{ni} \log(P_{ni}) \quad (4)$$

Então, para estimar a probabilidade de os dados observados seguirem a forma funcional proposta, o método da máxima verossimilhança é usado para calcular o vetor $\boldsymbol{\beta}^*$ = arg max$\boldsymbol{\beta}$

LL($\boldsymbol{\beta}$) que maximiza a densidade conjunta das amostras. Finalmente, substituindo os valores estimados de $\boldsymbol{\beta}$ em Eq. (2), é possível prever o modo de viagem de um indivíduo, conhecendo apenas os valores das características observadas (Akiva & Lerman, 1985; Trem, 2009). Neste artigo o modelo MLM foi estimado utilizando o pacote Apollo R (Hess & Palma, 2019) por meio da Algoritmo Broyden-Fletcher-Goldfarb-Shanno (BFGS).

4.2.2 Algoritmo de Árvore de Decisão - AD

Árvore de Decisão (AD) é um algoritmo de Aprendizado de Máquina supervisionado que é utilizado para classificação e para regressão. Dessa forma, pode ser usado para prever categorias discretas e para prever valores numéricos. A estrutura do AD é composta por árvore e nós que se relacionam entre si por uma hierarquia. Existe o nó-raiz, que contém o conjunto completo de dados, e os nós-folha que são os resultados finais. A raiz é um dos atributos da base de dados e o nó-folha é a classe ou o valor que será gerado como resposta (Tamim Kashifi *et al.*, 2022; Breiman *et al.*, 2017).

A análise do AD identifica grupos homogêneos de acordo com a variável dependente, interpreta os relacionamentos entre eles e prediz eventos futuros. Sendo assim, origina-se uma sequência de decisões, das quais ocorrem sucessivas divisões em um conjunto de dados até que o mesmo seja representado por diversas classes de observações que correspondem aos nós gerados. Nas situações em que nenhuma outra divisão dos dados é possível, os subconjuntos finais são denominados nós terminais (Breiman *et al.*, 1984).

Hastie *et al.* (2001) define que a AD constrói um modelo de previsão a partir de um conjunto de regras de decisão simples. A estrutura da árvore é construída a partir de um conjunto de dados de treinamento, dividindo recursivamente o espaço de entrada em regiões menores. Em cada divisão, a árvore escolhe a variável e o ponto de corte que melhor separa as classes ou prevê a saída numérica. O resultado final é um modelo hierárquico de regras de decisão, que é usado para a classificação ou regressão de novos pontos de dados. Entre os algoritmos de AD, os principais utilizados pela literatura são: CHAID (Kass, 1980), CART (Breiman *et al.*, 1984) e C4.5 (Quilan, 1983).

Nesse artigo, o algoritmo CART foi utilizado com o pacote *rpart* no R, dado que o modelo Floresta Aleatória também utilizado para comparação nesse artigo também está baseado no método CART (Khalilia *et al.*, 2011). O algoritmo CART consiste na repartição binária dos

dados e permite dos dados e permite a utilização de variáveis dependentes categóricas ou numéricas. Dessa forma, os “nós pais” são divididos em 2 nós filhos a cada camada de segmentação. A partir do algoritmo CART, é possível identificar as variáveis independentes que foram consideradas mais importantes na divisão dos dados, em cada camada, de modo a gerar sub-amostras mais homogêneas em cada nó, em relação à variável dependente (Breiman *et al.*, 1984).

4.2.3 Algoritmo de Floresta Aleatória

Floresta Aleatória é um algoritmo de Aprendizado de Máquina baseado em conjunto que é composto por ‘n’ coleções de árvores de decisão não correlacionadas. É construído a partir da ideia de agregação *bootstrap*, que é um método de re-amostragem com reposição para reduzir a heterogeneidade (Hastie *et al.*, 2009). O FA usa várias árvores para calcular a média (regressão) ou calcular os votos da maioria (classificação) nos nós terminais ao fazer uma previsão. Construídos a partir da ideia de Árvores de Decisão, resultaram em melhorias significativas na precisão da previsão em comparação com uma única árvore crescendo 'n' número de árvores; cada árvore no conjunto de treinamento é amostrada aleatoriamente sem reposição (Kirasich *et al.*, 2018).

O Floresta Aleatória é um tipo de árvore de classificação e regressão (CART) e, também um tipo de algoritmo de aprendizado conjunto. Considerando o problema de *overfitting*, Breiman (2001) propôs o modelo FA que combina os resultados de múltiplas árvores (floresta) sem um aumento significativo na complexidade computacional. O algoritmo de FA é construído em um vetor aleatório do espaço de características de dados.

O método é usado tanto para regressões quanto para classificações. No método de árvore de decisão padrão, o objetivo é encontrar a melhor divisão entre todas as variáveis ao criar um nó na árvore. Mas na Floresta Aleatória, amostra de *bootstrap* diferente escolhida aleatoriamente dos dados neste método. Com essa estratégia, o desempenho do algoritmo pode vir a ser melhor do que outros, como Máquinas de Vetores de Suporte e Redes Neurais Artificiais. O algoritmo de FA pode aumentar a precisão da regressão sem aumentar muito a complexidade computacional. Além disso, devido à seleção aleatória, pode reduzir o *overfitting*. A FA tem sido usada para previsão de séries temporais em muitas áreas, como padronização, saúde, mudanças climáticas, ciências ambientais, energia e ciências sociais. (Serin *et al.*, 2022).

4.2.4 Algoritmo de Redes Neurais Artificiais

Redes Neurais Artificiais representam uma família de algoritmos de Aprendizado de Máquina que tenta simular as funções do cérebro humano em um sistema de computador de forma simplificada. Sendo composto por uma rede de neurônios artificiais interconectados, que processam informações através de suas conexões e fornecem uma saída como resposta (Bayliss, 2021; Rumelhart *et al*., 1986). O termo "rede neural" nasceu na tentativa de encontrar uma representação matemática do processamento da informação em sistemas biológicos e a aprendizagem de uma rede neural envolve a adaptação dos pesos da rede a partir de estímulos fornecidos por um conjunto de dados (Bishop, 2006).

A vantagem das Redes Neurais Artificiais em relação a outros algoritmos de Aprendizado de Máquina é sua capacidade de capturar relações complexas entre as variáveis de entrada e saída, mesmo que não estejam explicitamente definidas. A rede é treinada com um conjunto de dados de entrada e saída, de forma que os pesos das conexões entre os neurônios sejam ajustados de modo a minimizar o erro entre a saída prevista e a saída real, através de algoritmos de otimização (Hillel *et al.*, 2021; Goodfellow *et al.*, 2016).

O algoritmo de RNAs tem sido aplicado em diversos problemas relacionados a transporte, incluindo a escolha modal. A sua aplicação na escolha modal envolve o treinamento de uma Rede Neural Artificial com dados de treinamento, incluindo informações sobre as características da viagem e o modo de transporte escolhido pelo usuário. A partir desses dados, a RNAs é capaz de aprender as relações entre as variáveis e prever o modo de transporte mais provável para uma determinada viagem (García-García *et al.*, 2022).

A aplicação de Redes Neurais Artificiais para a escolha modal tem o potencial de melhorar a eficiência do sistema de transporte, reduzir o congestionamento nas vias e oferecer aos usuários um serviço mais personalizado e eficiente. No entanto, assim como em qualquer modelo de aprendizado de máquina, a qualidade dos resultados depende da qualidade dos dados de treinamento e da seleção adequada das variáveis de entrada. Apesar da eficácia em muitas aplicações, as RNAs também apresentam limitações, tais como a necessidade de grande quantidade de dados para treinamento e a dificuldade de interpretação dos resultados (Zhao *et al.*, 2020).

4.3 PROCEDIMENTOS METODOLÓGICOS

Os procedimentos propostos foram realizados em 5 etapas: (i) Elaboração da pesquisa de Preferência Declarada, (ii) Coleta de dados, (iii) Calibração e Treinamento – Logit Multinomial, Árvore de Decisão, Floresta Aleatória e Redes Neurais Artificiais, (iv) Validação e teste dos modelos estimados, e (v) Comparação dos resultados dos modelos aplicados. Para classificar a importância das variáveis nos modelos foi utilizada toda a base. Já para estimar e testar, a base de dados foi dividida de maneira aleatória em uma amostra de treinamento (80%) e teste (20%) e rodada a mesma em todos os modelos.

4.3.1 Elaboração e coleta de dados de Preferência Declarada

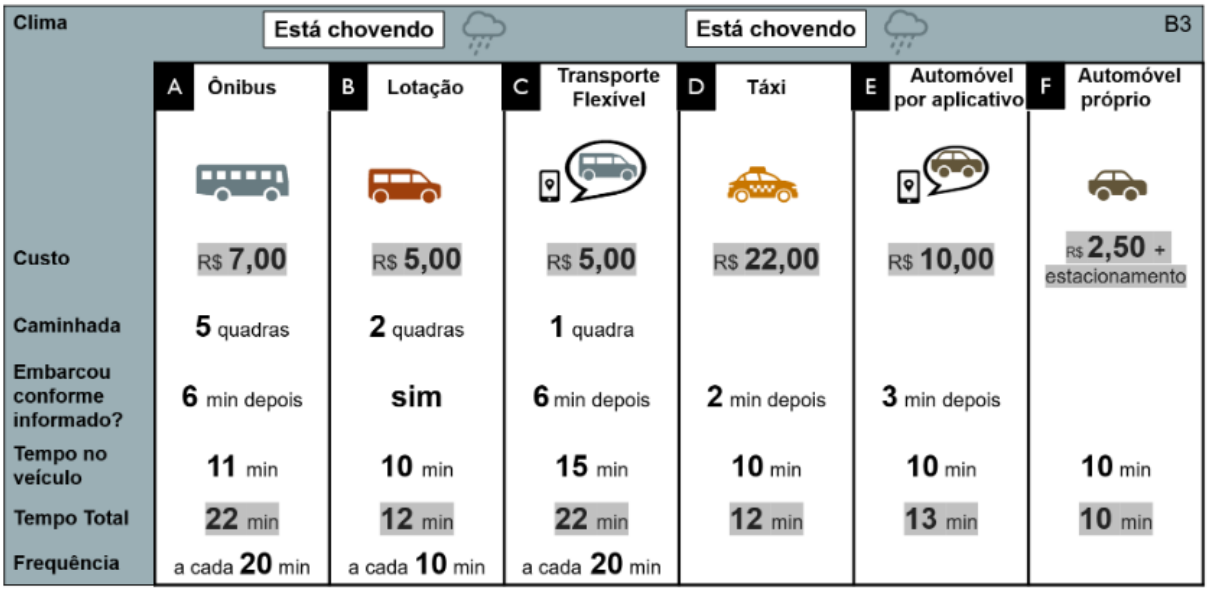
A base de dados utilizada neste estudo foi resultado de uma pesquisa de Preferência Declarada realizada de maneira presencial e com complemento *online*. cujo público-alvo foram habitantes da cidade de Porto Alegre. A pesquisa foi aplicada no período entre outubro de 2019 e janeiro de 2020.

A pesquisa foi estruturada com o objetivo de analisar a escolha dos indivíduos em relação a seis modos de transporte (sendo cinco modos existentes e um modo de transporte público sob demanda, não existente atualmente na cidade). As alternativas consideradas foram: (i) ônibus regular, o qual opera com itinerários fixos e diferentes modelos de veículos a depender da linha e da operadora; (ii) lotação, serviço que complementa o ônibus regular, caracterizado pela operação seletiva, com itinerário fixo, com parada variável e micro-ônibus na totalidade equipados com ar-condicionado; (iii) táxi, o qual é um serviço regulado, (iv) automóvel por aplicativo; (v) automóvel particular e (vi) transporte flexível, definido como um novo um sistema de transporte sob demanda, com rotas flexíveis/dinâmicas, operado por micro-ônibus (15 lugares), com pontos de embarque e desembarque próximos às localizações de origem e destino fornecidas pelo usuário, equipados com ar-condicionado e *wi-fi*.

O projeto experimental foi elaborado com 7 atributos: (i) Custo – custo da viagem; (ii) Caminhada – distância de acesso ao modo; (iii) Tempo no veículo; (iv) Tempo total; (v) Frequência – tempo entre as viagens dos modos coletivos, atributos estes tradicionalmente empregados em estudos de transporte (Ortúzar e Willumnsen, 2011). Também foram incluídos os atributos (vi) Chuva, de forma a apresentar as condições climáticas, que apontam ser

importantes em estudos de escolha modal, principalmente na utilização de transporte por aplicativo (Rodrigues *et al.*, 2019; Frei *et al.*, 2017) e (vii) Confiabilidade, baseado no estudo de Frei *et al.* (2017), o qual informa se houve atraso na oferta do serviço. A Figura 6 mostra um exemplo da situação de escolha apresentada aos respondentes.

Figura 6 - Exemplo de cartão utilizado na pesquisa



(fonte: LASTRAN)

A estrutura utilizada foi um desenho eficiente bayesiano (Rose e Bliemer, 2009) e implementado em NGene (Choice Metrics, 2013). O desenho eficiente foi utilizado para minimizar os erros padrão das estimativas de parâmetros. A medida de eficiência utilizada foi o D-erro Bayesiano, buscando obter os menores valores possíveis, obtendo um valor de 0.059 (Rose e Bliemer, 2009; Choice Metrics, 2013). Os valores iniciais adotados para o projeto experimental foram provenientes do estudo de Frei *et al.* (2017) e de um trabalho prévio realizado na mesma cidade (Sassi, 2017).

A seleção do bloco de cartões a serem respondidos era proveniente do deslocamento mais frequente no último mês. Sendo assim, os conjuntos de problemas de escolha foram divididos em blocos de acordo com as informações do deslocamento descrito. A partir dos dados de origem e destino fornecidos foi calculada a distância percorrida no deslocamento padrão dos respondentes, através do horário de saída se identificou o deslocamento foi realizado no horário pico ou fora dele. Por fim, definiu-se oito blocos de cartões classificados de acordo com o horário e a distância percorrida, conforme a Tabela 2.

O projeto experimental foi elaborado com 9 situações de escolha, separadas nos 8 blocos de forma a fornecer realismo ao experimento. Os blocos representam diferentes combinações de distância (até 4 km, de 4 a 8km, de 8 a 12 km, maior que 12 km) e período da viagem (pico e fora do pico) que poderiam ser experimentados usualmente pelos usuários (Tabela 2).

Tabela 2 - Bloco de cartões classificados

Bloco	Hora Pico	Distância (km)
A	Sim	< 4
B	Não	< 4
C	Sim	4 a 8
D	Não	4 a 8
E	Sim	8 a 12
F	Não	8 a 12
G	Sim	> 12
H	Não	> 12

(fonte: elaborado pela autora)

A categorização de distância foi baseada na análise dos histogramas de uma pesquisa prévia realizada na cidade e na classificação das linhas de ônibus para a cidade (curta, média e longa distância). De acordo com o horário, origem e destino da viagem, reportados pelo entrevistado, para sua viagem mais frequente (seção 1 do questionário de pesquisa), foi calculada a distância percorrida, e identificado se a viagem foi realizada no horário de pico ou fora dele. Dessa forma, o entrevistado foi conduzido para o jogo que representava, de forma mais adequada, o seu deslocamento mais frequente. Uma pesquisa piloto foi realizada para verificar o entendimento do questionário e identificar melhorias no processo de pesquisa.

4.3.2 Calibração e Treinamento

4.3.2.1 Modelo Logit Multinomial

Modelos MLM foram estimados para analisar o processo de decisão em relação ao modo de transporte, especificamente na utilização de um modo de transporte público flexível. Foram utilizadas funções de utilidade lineares nos parâmetros, utilizadas usualmente na literatura, considerando unicamente os atributos da pesquisa de PD. A estimação do modelo foi realizada utilizando o pacote Apollo R (Hess e Palma, 2019).

4.3.2.2 Algoritmo de Árvore de Decisão

Nesse artigo, o algoritmo CART foi utilizado com o pacote *rpart* no R, dado que o algoritmo Floresta Aleatória, também utilizado para comparação nesse artigo, também está baseado no método CART (Khalilia et al., 2011).

Para a obtenção da importância das variáveis, de acordo com Breiman (2001) existem duas maneiras de medir a importância dos atributos para utilização: (i) Importância Baseada no Erro e (ii) Aprimoramento. O primeiro mede o aumento do erro ao se permutarem os valores do atributo de interesse. O segundo se baseia na soma dos decréscimos do Índice de Gini em todos os nós rotulados pelo atributo. Essas medidas podem ser utilizadas para indicar os atributos mais importantes para o modelo (Guyon e Elisseeff, 2003). Nesse artigo foi utilizado aprimoramento pelo Índice Gini, pois a variável dependente é categórica.

4.3.2.3 Algoritmo de Floresta Aleatória

O algoritmo Floresta Aleatória foi desenvolvido com 1500 árvores. As variáveis mais importantes para o modelo são aquelas que apresentam maior média de decréscimo de índice Gini (aumento de homogeneidade nos nós filhos) - aprimoramento.

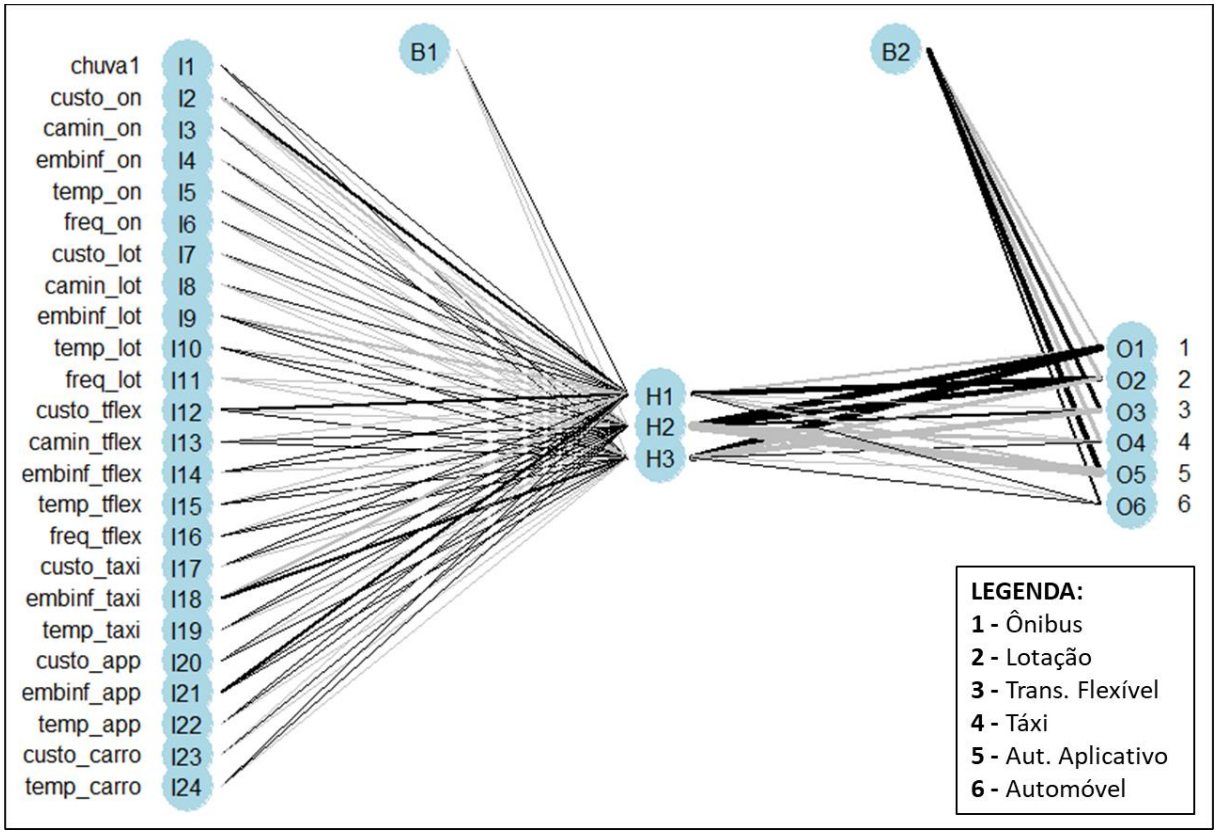
O pacote R *randomForest* utilizado fornece como medida de importância a diminuição média do Gini (DMG) para classificar e selecionar variáveis. O DMG é a soma de todas as diminuições na impureza do Gini devido a uma variável específica (quando essa variável é usada para formar uma divisão na FA), normalizada pelo número de árvores (Calle e Urrea, 2011).

4.3.2.4 Algoritmo de Redes Neurais Artificiais

Para o modelo de Redes Neurais Artificiais a importância das variáveis é calculada através do método proposto por Geyerey *et. al.* (2003), com a aplicação de combinações dos valores absolutos dos pesos. A Figura 7 esquematiza a estrutura com uma única camada oculta de 3 neurônios utilizada no modelo. Os pesos da conexão foram treinados por retropropagação com uma constante de decaimento de peso de 0,4. O método numérico para plotagem requer que os pesos de entrada estejam em uma ordem específica dada a estrutura da rede. Uma estrutura de argumento adicional também foi necessária para listar o número de nós nas camadas de entrada, oculta e de saída. Foi utilizada a sintaxe de plotagem de I, H, O e B para entrada (*input*), oculta

(*hidden*), saída (*output*) e polarização (bias) para indicar conexões ponderadas entre camadas, a ordem de peso correta para o vetor.

Figura 7 – Estrutura do algoritmo de RNAs



(fonte: elaborado pela autora)

4.3.2.5 Validação, teste e comparação dos modelos estimados

Os modelos foram aplicados à amostra de teste (20% da amostra total) e calculadas as taxas de acertos para cada modo e modelo. Os resultados foram comparados utilizando o percentual de acerto global na amostra de teste e o coeficiente *kappa* (Cohen, 1960) que é usualmente utilizado para avaliar o nível de concordância entre os conjuntos de dados.

4.4 ANÁLISE DOS RESULTADOS

Os resultados dos modelos utilizados nesse artigo estão subdivididos e serão descritos nos próximos itens.

4.4.1 Modelo Logit Multinomial

Os resultados dos parâmetros estimados para o modelo MLM estão na Tabela 3. Para cada variável é apresentada a estimativa obtida juntamente com o resultado do teste-t de significância estatística.

Tabela 3 – Resultados do MLM

Atributo	Estimado	Valor-t
const_on	0	NA
const_lot	-0.63919	-6.49209
const_tflex	-0.36213	-7.38726
const_taxi	-2.06150	-11.46198
const_app	-0.37233	-4.751892
const_carro	0.29174	3.67807
custo	-0.10170	-20.72737
temp_cam	-0.18296	-20.30564
temp_emb	-0.03734	-12.57395
temp_vei	-0.01007	-6.08807
freq	-0.07201	-8.68245
chuva	0.45161	9.18657

(fonte: elaborado pela autora)

De acordo com os resultados obtidos, todas os parâmetros do modelo MLM apresentaram significância estatística.

4.4.2 Algoritmo de Árvore de Decisão

O resultado do ranqueamento do Índice Gini está apresentado na Tabela 4. As variáveis classificadas como mais importantes para o AD foram *custo_app*, *custo_tflex* e *custo_on*, respectivamente.

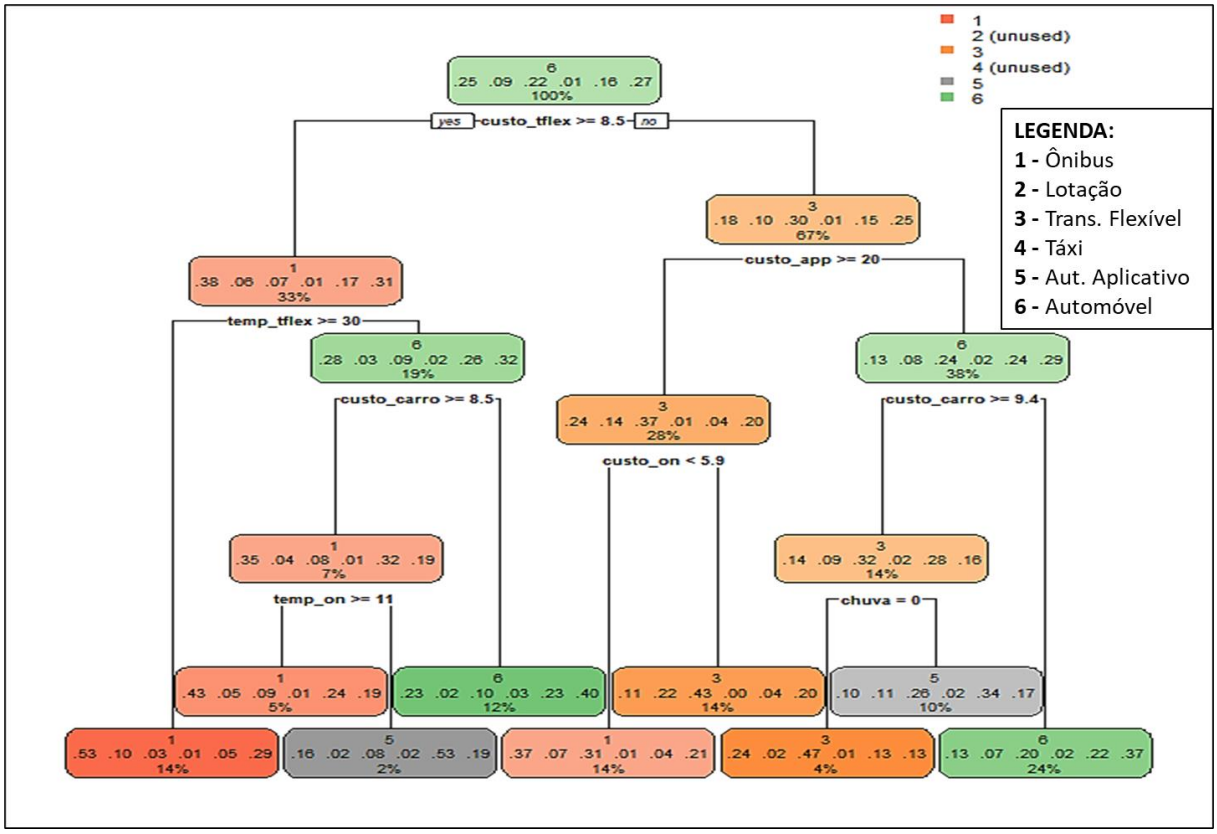
Tabela 4 - Ranqueamento dos atributos da AD

Atributo	Posição
custo_app	1
custo_tflex	2
custo_on	3
temp_on	4
temp_tflex	5
camin_lot	6
embinf_taxi	7
temp_lot	8
embinf_lot	9
custo_taxi	10
chuva	11
embinf_on	12
custo_carro	13
camin_on	14
embinf_tflex	15
freq_on	16
custo_lot	17
freq_lot	18
camin_tflex	19
freq_tflex	20
temp_taxi	21
embinf_app	22
temp_app	23
temp_carro	24

(fonte: elaborado pela autora)

A Figura 8 apresenta esquematicamente a árvore da etapa de treinamento obtida pelo algoritmo. A árvore se constituiu de 4 níveis de profundidade e 9 nós terminais.

Figura 8 - Visualização esquemática da Árvore de Decisão

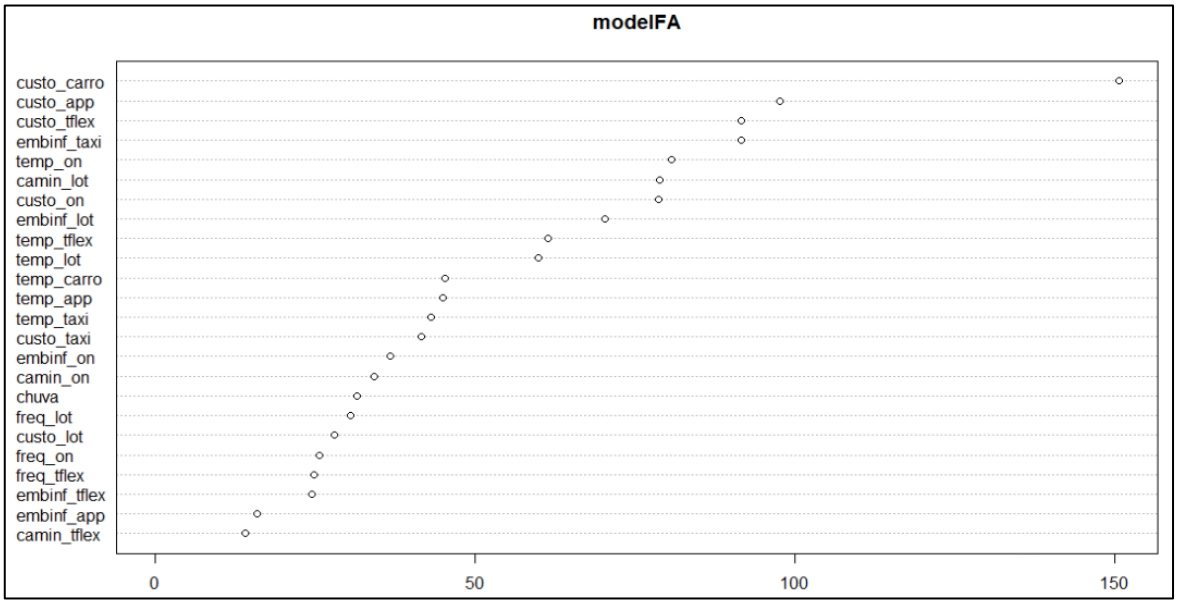


(fonte: elaborado pela autora)

4.4.3 Algoritmo de Floresta Aleatória

A classificação de importância das variáveis para o algoritmo FA se encontra exposta na Figura 9. A variável menos importante apresenta o menor valor de redução média dos valores de índice Gini. De acordo com o resultado do modelo, a escolha modal está relacionada aos três fatores mais importantes: *custo_carro*, *custo_app* e *custo_tflex*. Os resultados foram semelhantes aos encontrados para a árvore isolada (CART). O resultado da importância das variáveis independentes para o FA foi determinado pela diminuição média do Gini (DMG).

Figura 9 - Importância das variáveis independentes do algoritmo de FA



(fonte: elaborado pela autora)

4.4.4 Algoritmo de Redes Neurais Artificiais

O modelo de RNAs mostrou como os três atributos mais importantes: *custo_tflex*, *camin_lot* e *embinf_taxi*. Os resultados são consideravelmente distintos dos anteriores, no entanto, cabe salientar que o método recomendado para esse modelo é diferente dos anteriores. A Tabela 5 o ranqueamento das variáveis.

Tabela 5 – Ranqueamento dos atributos da RNAs

Atributo	Posição
custo_tflex	1
camin_lot	2
embinf_taxi	3
embinf_tflex	4
embinf_lot	5
custo_on	6
freq_on	7
freq_lot	8
temp_carro	9
temp_tflex	10
temp_app	11
temp_taxi	12
custo_app	13
custo_carro	14
chuva1	15
embinf_app	16
camin_tflex	17
embinf_on	18
custo_lot	19
camin_on	20
temp_lot	21
custo_taxi	22
temp_on	23
freq_tflex	24

(fonte: elaborado pela autora)

4.4.5 Validação e teste dos modelos estimados

A Tabela 6 apresenta os resultados da taxa de acertos em percentual de acordo com cada modo para os diferentes modelos utilizados. No modelo MNL a maior taxa de acerto foi a do Ônibus, seguido pelo Automóvel e pelo Transporte Flexível. O modo Táxi não teve previsão nos modelos comparados, uma vez que possui menos de 5% de observações no banco de dados da pesquisa. No modelo AD, os maiores percentuais de acertos foram do Ônibus e do Transporte Flexível, com 42,89% e 42,67%, respectivamente. Para o FA, as melhores taxas de acertos foram novamente para o Ônibus e o Transporte Flexível. Por fim, para o RNAs, verifica-se que o Ônibus segue sendo o modo de melhor previsão, seguido pelo Transporte Flexível e o Automóveis por Aplicativo.

Tabela 6 – Consolidação dos resultados da taxa de acerto (%) por modo e por modelo

Modo de transporte	**1**: Ônibus	**2**: Lotação	**3**:Trans.Flexível	**4**: Táxi	**5**: Aplicativo	**6**: Automóvel
MLM	54,87	33,33	48,95	0	46,35	53,98
AD	42,89	0	42,67	0	33,71	36,07
FA	45,36	33,61	42,28	0	41,32	38,55
RNAs	45,38	0	43,66	0	34,07	37,53

(fonte: elaborado pela autora)

4.4.6 Comparativo entre os Resultados dos Modelos

A Tabela 7 apresenta os resultados dos modelos Logit Multinomial, Árvore de Decisão, Floresta Aleatória e Redes Neurais Artificiais. Os resultados informam o percentual de acerto global na amostra de teste (20%) e o coeficiente *kappa* (Cohen, 1960).

Tabela 7 – Comparação entre os resultados globais dos modelos

Modelo	% Acerto	Coeficiente kappa
Logit Multinomial (MLM)	52.03	0.369
Árvore de Decisão (AD)	39.44	0.201
Floresta Aleatória (FA)	41.79	0.246
Redes Neurais (RNAs)	40.94	0.221

(fonte: elaborado pela autora)

O modelo MLM apresentou a maior acurácia (52,03%), seguido pelo FA (41,79%), RNAs (40,94%) e por último o AD (39,44%) com a menor taxa de acertos. Dessa forma, o MLM que é tradicionalmente utilizado para previsão de escolha modal, demonstrou o melhor desempenho entre os modelos. Cabe salientar que os algoritmos de Aprendizado de Máquina em conjuntos de dados menores apresentam um problema, pois o 'poder' em reconhecer padrões é proporcional ao tamanho do conjunto de dados, ou seja, quanto menor o conjunto de dados, menos poderosos e menos precisos são os algoritmos (Kokol *et al.*, 2022). Nesse sentido, há a necessidade de utilizar bancos de dados mais ricos contendo e informações heterogêneas para reproduzir um conjunto de comportamentos coerentes e a base utilizada possui déficit de observações nos modos Lotação (2) e Táxi (4).

4.5 CONSIDERAÇÕES FINAIS

O presente artigo se propôs a realizar uma análise comparativa entre o Modelo Logit Multinomial, tradicionalmente utilizado, com os algoritmos de Aprendizagem de Máquina: Árvore de Decisão, Floresta Aleatória e Redes Neurais Artificiais para previsão de escolha modal através de dados provenientes de uma pesquisa de Preferência Declarada realizada em Porto Alegre no ano de 2019.

A elaboração do referencial teórico baseou-se em pesquisas que empregaram os modelos analisados em aplicações semelhantes na área de transportes. Esse processo se justifica pois o aumento da precisão na modelagem da escolha do modo de viagem é crucial para o planejamento, permitindo que os gestores de políticas possam prever a demanda de viagens e compreender os fatores subjacentes.

Para o modelo de escolha discreta, as probabilidades de escolha são expressas em uma fórmula matemática relativamente simples e interpretável. No entanto, a construção do modelo requer suposições estatísticas rigorosas, que assumem que todas as fontes de variabilidade e correlação entre as alternativas podem ser capturadas por uma combinação linear de covariáveis. Por outro lado, os modelos de Aprendizado de Máquina se encontram em crescente desenvolvimento e aplicação na área de planejamento de transportes. Dessa forma, combinou-se o emprego de novas tecnologias com a utilização desses algoritmos para comparação na previsão de demanda por modo de transporte.

Os resultados mostraram que o MLM apresentou uma acurácia preditiva maior em comparação com os modelos de aprendizado de máquina testados, com taxa de acerto de 52,03%, seguido pelo FA com 41,79% e o RNAs com 40,94%. A explicação para esse desempenho superior do MLM pode ser o fato de que a base de dados utilizada na PD apresentou poucas observações para os modais Lotação e Táxi. Por ser mais simples e exigir menos dados de treinamento em comparação aos demais, as previsões foram mais precisas para os modais com menos observações e maior acurácia no desempenho geral.

Sendo assim, os resultados indicam que o Logit Multinomial segue como uma opção viável e eficiente para prever a escolha modal em estudos com baixa amostragem para modais específicos. No entanto, é importante destacar que a escolha do modelo de previsão adequado dependerá dos requisitos específicos do problema e das características dos dados.

Para trabalhos futuros, aconselha-se adequar o tipo da base de dados aos modelos aplicados. Recomenda-se ainda realizar comparações com modelos híbridos com novos algoritmos de Aprendizado de Máquina e com banco de dados com maior homogeneidade de respostas para todos os modos ou suprimindo da análise os que não possuírem volume relevante de observações.

REFERÊNCIAS

BAYLISS, C. Machine learning based simulation optimisation for urban routing problems. **Applied Soft Computing**, 105, 107269, 2021. doi:https://doi.org/10.1016/j.asoc.2021.107269

BEHROOZ, H., e HAYERI, Y. M. (2022) Machine Learning Applications in Surface Transportation Systems: A Literature Review. **Applied Sciences,** 12(18), 9156. doi:10.3390/app12189156

BISHOP, M. Pattern recognition and machine learning. **Information science and statistics,** Springer, New York, 2006.

BREIMAN, Leo. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., E STONE, C. J. (2017) Classification And Regression Trees. (1º ed). **Routledge**. doi:10.1201/9781315139470

DIALLO, A; O., LOZENGUEZ, G., DONIEC, A. Estimation of minority modes of transportation based on machine learning approach. **Procedia Computer Science**, v. 201, p. 265–272, 2022. doi:10.1016/j.procs.2022.03.036

FREI, C., HYLAND, M., E MAHMASSANI, H. S. Flexing service schedules: Assessing the potential for demand-adaptive hybrid transit via a stated preference approach. **Transportation Research Part C: Emerging Technologies, 76**, 71–89, 2017. doi:10.1016/j.trc.2016.12.017

GARCÍA-GARCÍA, J. C., GARCÍA-RÓDENAS, R., LÓPEZ-GÓMEZ, J. A., MARTÍN-BAOS, J. A. A comparative study of machine learning, deep neural networks and random utility maximization models for travel mode choice modelling. **Transportation Research Procedia**, v. 62, p. 374–382, 2022. doi:10.1016/j.trpro.2022.02.047

GOODFELLOW, I., BENGIO, Y., E COURVILLE, A. (2016) Deep learning. **The MIT Press**, Cambridge, Massachusetts.

GUYON, I., ELISSEEFF, A. An Introduction to Variable and Feature Selection. **The Journal of Machine Learning Research, 3**, 1157-1182, 2003.

HASTIE, T., FRIEDMAN, J., E TIBSHIRANI, R. (2001) *The Elements of Statistical Learning*. **Springer New York**, New York, NY. doi:10.1007/978-0-387-21606-5

HESS, S., PALMA, D. Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. **Journal of Choice Modelling**, v. 32, p. 100170, 2019. doi:10.1016/j.jocm.2019.100170

HILLEL, T., BIERLAIRE, M., ELSHAFIE, M. Z.E.B., JIN, Y. A systematic review of machine learning classification methodologies for modelling passenger mode choice. **Journal of Choice Modelling**, v. 38, p. 100221, 2021. doi:10.1016/j.jocm.2020.100221

KASS, G. V. An Exploratory Technique for Investigating Large Quantities of Categorical Data. **Applied Statistics, 29**, 119, 1980.

KHALILIA, M., CHAKRABORTY, S., POPESCU, M. Predicting disease risks from highly imbalanced data using random forest. **BMC Medical Informatics and Decision Making**, v. 11, n. 1, p. 51, 2011. doi:10.1186/1472-6947-11-51

KIRASICH, K., SMITH, T., SADLER, B. Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets, **SMU Data Science Review:** Vol. 1: No. 3, Article 9, 2018. Disponível em: https://scholar.smu.edu/datasciencereview/vol1/iss3/9

KOKOL, P., KOKOL, M., ZAGORANSKI, S. Machine learning on small size samples: A synthetic knowledge synthesis. Science Progress, v. 105, n. 1, p. 003685042110297, 2022. Disponível em: http://journals.sagepub.com/doi/10.1177/00368504211029777

LI, P., WU, W., PEI, X. A separate modeling approach for short-term bus passenger flow prediction based on behavioral patterns: A hybrid decision tree method. **Physica A: Statistical Mechanics and its Applications**, p. 128567, 2023. doi:10.1016/j.physa.2023.128567

MCFADDEN, D. Frontiers in econometrics. Economic theory and mathematical economics. **Academic Press, New York,** 1974.

OLAYODE, I. O., TARTIBU, L. K., OKWU, M. O. Prediction and modeling of traffic flow of human-driven vehicles at a signalized road intersection using artificial neural network model: A South African road transportation system scenario. **Transportation Engineering**, v. 6, p. 100095, 2021. doi:10.1016/j.treng.2021.100095

ORTÚZAR, J. D., WILLUMSEN, L. G. Modelling Transport, **Wiley**, ed. 1, 2011. Disponível em: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119993308

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, v. 1, n. 1, p. 81–106, 1986.

ROSE, J. M., BLIEMER, M. C. J. Constructing Efficient Stated Choice Experimental Designs. **Transport Reviews**, v. 29, n. 5, p. 587–617, 2009.

RUMELHART, D. E., HINTON, G. E., E WILLIAMS, R. J. (1986) Learning representations by back-propagating errors. **Nature**, 323(6088), 533–536. doi:10.1038/323533a0

SERIN, F., ALISAN, Y., ERTURKLER, M. Predicting bus travel time using machine learning methods with three-layer architecture. **Measurement**, [*s. l.*], v. 198, p. 111403, 2022. doi:10.1016/j.measurement.2022.111403

TAMIM KASHIFI, M., JAMAL, A., SAMIM KASHEFI, M., ALMOSHAOGEH, M., MASIUR RAHMAN, S. Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. **Travel Behaviour and Society**, v. 29, p. 279–296, 2022. doi:10.1016/j.tbs.2022.07.003

TRAIN, K. **Discrete choice methods with simulation**. 2nd ed. Cambridge ; New York: Cambridge University Press, 2009. Disponível em: https://www.cambridge.org/core/product/identifier/9780511805271/type/book

UN-HABITAT. **Urbanization and development: emerging futures**. Nairobi, Kenya: UN-Habitat, 2016. (World cities report, 2016).

WANG, X., QING-DAO-ER-JI, R. Application of optimized genetic algorithm based on big data in bus dynamic scheduling. **Cluster Computing**, 2018. doi:10.1007/s10586-018-2625-x

ZHAO, X., YAN, X., YU, A. Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. **Travel Behaviour and Society**, v. 20, p. 22–35, 2020. doi:10.1016/j.tbs.2020.02.003

5 CONSIDERAÇÕES FINAIS DA DISSERTAÇÃO

Essa dissertação se propôs a preencher as lacunas de conhecimento das demandas atuais do planejamento de transporte urbano visando o desenvolvimento sustentável dos sistemas de transporte e a necessidade de explorar diferentes técnicas de análise para aprimorar a eficiência e a eficácia. Nesse contexto, a incorporação de algoritmos de *Machine Learning* tem se mostrado promissora para lidar com os desafios complexos relacionados ao planejamento do transporte urbano.

A dissertação foi dividida em dois artigos que visam analisar o uso de algoritmos de *Machine Learning* no planejamento de transporte urbano. O primeiro artigo consistiu em uma revisão sistemática da literatura para analisar de forma quantitativa os estudos existentes sobre o tema, identificando as principais aplicações e como elas podem auxiliar na otimização dos sistemas de transporte urbano.

Os resultados da revisão sistemática indicaram que os métodos de Aprendizado de Máquina estão em crescente utilização no planejamento de transportes. Os modelos de previsão de demanda de tráfego e de transporte público se destacaram como os mais empregados na literatura, além de outros métodos como reconhecimento de sinais de trânsito, detecção de semáforos, classificação de veículos, detecção de pedestres, planejamento de tempo de viagem e de itinerário e comparativos entre algoritmos diferentes.

O segundo artigo comparou modelos tradicionais de escolha discreta com algoritmos de Aprendizado de Máquina para analisar a previsão da escolha modal, a partir de dados de uma pesquisa de Preferência Declara realizada em Porto Alegre em 2019. Os resultados do estudo comparativo indicaram que o modelo de Logit Multinomial (MLM) apresentou uma acurácia preditiva significativamente maior em comparação com os outros modelos de Aprendizado de Máquina testados. A taxa de acerto do MLM foi de 52,03%, seguida pelo método de Floresta Aleatória (FA) com 41,79%, e as Redes Neurais Artificiais (RNAs) com 40,94%. Uma possível explicação para o desempenho superior do MLM pode ser o fato de que a base de dados utilizada na pesquisa continha poucas observações para os modais Lotação e Táxi.

Dessa forma, os resultados mostram que os modelos de escolha discreta seguem como uma opção viável e eficiente para prever a escolha modal em estudos com baixa amostragem para modais específicos. Cabe salientar a importância de estudos comparativos para auxiliar na

escolha do modelo de previsão adequado de acordo com os requisitos específicos do problema e das características dos dados.

Com base no conhecimento adquirido durante o desenvolvimento dessa dissertação, recomenda-se a utilização de uma variedade de bancos de dados na análise da relação entre eles, avaliando a adequação dos modelos aplicados. O campo do Aprendizado de Máquina tem sido caracterizado por uma constante expansão, com o surgimento de novos algoritmos. Por essa razão, sugere-se que novas análises comparativas sejam realizadas, agregando modelos híbridos, para que se possa aproveitar as vantagens de cada um deles. Essas ações podem aprimorar as análises e contribuir para o desenvolvimento da área de Aprendizado de Máquina, permitindo que sejam realizadas previsões mais precisas e confiáveis.

Por fim, a incorporação de algoritmos de *Machine Learning* no planejamento de transporte urbano pode apresentar um elevado potencial de redução de custos e aumento da eficiência dos sistemas de transporte. Através da análise de previsão do modo de transporte, é possível realizar um dimensionamento otimizado dos recursos, como veículos e rotas, para atender a demanda de transporte de forma mais eficiente e econômica dando embasamento para tomada de decisão mais eficiente e baseada em dados, resultando em possíveis melhores soluções para problemas de transporte.

REFERÊNCIAS

CHANG, X., WU, J., LIU, H. Travel mode choice: a data fusion model using machine learning methods and evidence from travel diary survey data. **Transportmetrica A: Transport Science**, v. 15, n. 2, p. 1587–1612, 2019. Disponível em: https://www.tandfonline.com/doi/full/10.1080/23249935.2019.1620380

DIALLO, A; O., LOZENGUEZ, G., DONIEC, A. Estimation of minority modes of transportation based on machine learning approach. **Procedia Computer Science**, v. 201, p. 265–272, 2022. doi:10.1016/j.procs.2022.03.036

ITDP Transporte de média e alta capacidade. **Instituto de Políticas de Transporte e Desenvolvimento**, Rio de Janeiro, RJ, 2022.

KOKOL, P., KOKOL, M., ZAGORANSKI, S. Machine learning on small size samples: A synthetic knowledge synthesis. Science Progress, v. 105, n. 1, p. 003685042110297, 2022. Disponível em: http://journals.sagepub.com/doi/10.1177/00368504211029777

LI, P., WU, W., PEI, X. A separate modeling approach for short-term bus passenger flow prediction based on behavioral patterns: A hybrid decision tree method. **Physica A: Statistical Mechanics and its Applications**, p. 128567, 2023. doi:10.1016/j.physa.2023.128567

MCFADDEN, D. Frontiers in econometrics. Economic theory and mathematical economics. **Academic Press, New York,** 1974.

PENG, J., CHEN, L., ZHANG, B. Transportation planning for sustainable supply chain network using big data technology. Information Sciences, v. 609, p. 781–798, 2022. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0020025522008015

SERIN, F., ALISAN, Y., ERTURKLER, M. Predicting bus travel time using machine learning methods with three-layer architecture. Measurement, v. 198, p. 111403, 2022. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0263224122006364

TAMIM KASHIFI, M., JAMAL, A., SAMIM KASHEFI, M., ALMOSHAOGEH, M., MASIUR RAHMAN, S. Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. **Travel Behaviour and Society**, v. 29, p. 279–296, 2022. doi:10.1016/j.tbs.2022.07.003

ZHAO, X., YAN, X., YU, A. Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. **Travel Behaviour and Society**, v. 20, p. 22–35, 2020. doi:10.1016/j.tbs.2020.02.003

ZHU, L., YU, F. R., WANG, Y., NING, B., E TANG, T. Big Data Analytics in Intelligent Transportation Systems: A Survey. **IEEE Transactions on Intelligent Transportation Systems,** 20(1), 383–398, 2019. Disponível em: doi:10.1109/TITS.2018.2815678