CEFET-MG
Centro Federal de Educação Tecnológica de Minas Gerais
Programa de Pós-graduação em Modelagem Matemática e Computacional

Uso de *Big Data* no planejamento do transporte público: uma aplicação de técnicas de mineração de dados na inferência do motivo da viagem do passageiro

Mirian Greiner de Oliveira Pinheiro

Orientador: Prof. Dr. Gray Farias Moita
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Prof. Dr. Renato Guimarães Ribeiro
Centro Federal de Educação Tecnológica de Minas Gerais

Belo Horizonte
Outubro de 2023

Mirian Greiner de Oliveira Pinheiro

Uso de *Big Data* no planejamento do transporte público: uma aplicação de técnicas de mineração de dados na inferência do motivo da viagem do passageiro

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional.

Linha de pesquisa: Sistemas Inteligentes.

Orientador: Prof. Dr. Gray Farias Moita
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Prof. Dr. Renato Guimarães Ribeiro
Centro Federal de Educação Tecnológica de Minas Gerais

Centro Federal de Educação Tecnológica de Minas Gerais
Programa de Pós-graduação em Modelagem Matemática e Computacional
Belo Horizonte
Outubro de 2023

À minha própria determinação e perseverança, que me trouxeram até aqui e aos meus familiares e amigos pelo apoio e incentivo quando me faltou fé em mim mesma.

Agradecimentos

Concluir um mestrado após dois anos de intenso trabalho e mudanças em minha vida foi uma jornada desafiadora, porém incrivelmente enriquecedora para o meu crescimento pessoal e profissional. Durante esse período, enfrentei uma série de desafios e por isso, olhar para trás e ver que consegui chegar até aqui me enche de orgulho pois sei que precisei abdicar de muitas coisas e me desdobrar em tantas para equilibrar todos os meus pratos durante este tempo.

Nesse momento, quero expressar minha sincera gratidão a todos que estiveram ao meu lado. O apoio e a compreensão de vocês foram fundamentais para que eu chegasse até aqui. Em especial:

A Deus, por ter me dado a coragem e a força necessárias para superar desafios, especialmente por me capacitar a realizar um mestrado em uma área que inicialmente estava fora da minha zona de conforto.

Ao meu companheiro de vida Rafael, que esteve ao meu lado, oferecendo suporte incondicional, compreensão, amor e generosidade, tornando possível essa jornada. A você que sempre acreditou em mim, até mais do que eu mesma, o meu sincero obrigada. Dividir a vida com você é um dádiva.

Aos meus pais Alex e Valéria, que sempre me incentivaram, acreditando no meu potencial e me motivaram a persistir, mesmo nos momentos mais desafiadores.

Aos meus avós, que sempre demonstraram preocupação e carinho e torceram pelo meu sucesso.

Ao meu fiel companheiro de quatro patas, Heimdall, que esteve ao meu lado em trantas madrugadas, trazendo alegria e conforto quando mais precisei.

A todos os amigos que pacientemente me ouviram falar sobre o mestrado ao longo de dois anos, compartilhando as minhas alegrias e desafios. Especialmente à minha amiga Bianca, obrigada pela parceria e por entender a minha ausência neste tempo.

Ao meu amigo André, cuja ajuda e apoio foram inestimáveis no desenvolvimento deste trabalho de mestrado. Sua contribuição foi essencial para o meu sucesso.

Aos meus orienteadores Gray e Renato que me proporcionaram o suporte e a orientação necessários para concluir essa jornada acadêmica.

A Secretaria de Estado de Infraestrutura e Mobilidade de Minas Gerais (SEINFRA) e a Empresa de consultoria e engenharia Systra Brasil, pelo fornecimento de bases de dados para esta dissertação. Agradeço profundamente por sua contribuição para o meu projeto de mestrado.

Centro Federal de Educação Tecnológica (CEFET) por oferecer um ensino público de qualidade e a todos os professores que estiveram presentes nesta jornada, por seu compromisso com a educação de excelência e pela oportunidade de aprendizado que me proporcionaram.

"Por vezes sentimos que aquilo que fazemos não é senão uma gota de água no mar. Mas o mar seria menor se lhe faltasse uma gota" (Madre Tereza de Calcutá)

Resumo

A matriz origem e destino (OD) é uma ferramenta crucial no planejamento dos sistemas de transporte, pois fornece informações sobre os fluxos de viagens entre diferentes áreas da cidade e permite entender as demandas de mobilidade das pessoas e seus padrões de deslocamento. A coleta de dados utilizando o tradicional método das pesquisas domiciliares, é cara, trabalhosa e pouco frequente, sendo realizada a cada década em muitos casos. Para superar tais desafios, dados de cartões inteligentes passaram a ser utilizados no processo de obtenção das matrizes OD pois são coletados de forma contínua e em larga escala, sendo portanto uma fonte de dados mais abrangente e atualizada. Apesar das vantagens, os dados de cartões inteligentes também têm suas limitações, sendo a principal delas a falta de informação sobre atributos importantes da viagem, como por exemplo o motivo do deslocamento. Para suprir essa carência, técnicas de aprendizado de máquina podem ser utilizadas para inferir atributos não disponíveis nesta base de dados, sendo este o principal objetivo deste estudo. Neste trabalho, será apresentada uma aplicação da mineração de dados na previsão do motivo de viagem dos passageiros do transporte público da Região Metropolitana de Belo Horizonte, Minas Gerais, Brasil (RMBH). Para isto, diferentes algoritmos, técnicas de balanceamento de classes, hiperparâmetros e métodos de inferência foram testados e avaliados. Ao final, o método de inferência dividido em dois modelos sequenciais, um para classificar viagens primárias e outro para classificar viagens secundárias, ambos treinados com o algoritmo Random Forest e a técnica de balanceamento de classes Random Oversampling foi selecionado como o mais adequado. Ressalta-se, que o modelo treinado para classificar os motivos primários da viagem apresentou bons resultados, com acurácia superior a 90% e pode ser utilizado para prever com precisão os propósitos de viagem em dados de cartões inteligentes. Em relação aos motivos secundários, concluiu-se que o modelo precisa ser aprimorado (acurácia de 47%), sobretudo em relação aos dados utilizados para seu treinamento.

Palavras-chave: Inferência do propósito de viagem, Técnicas de aprendizado de máquina, Dados de cartões inteligentes

Abstract

The origin and destination matrix (OD) is a crucial tool in the planning of transportation systems, providing information about travel flows between different areas of the city and facilitating an understanding of people's mobility demands and patterns of movement. The conventional method of household surveys for data collection is expensive, labor-intensive, and infrequent, often conducted once a decade. To overcome these challenges, smart card data has become integral in obtaining OD matrices as it is collected continuously and on a large scale, offering a more comprehensive and updated data source. Despite its advantages, smart card data has limitations, notably the lack of information on important trip attributes such as the trip purpose. To address this gap, machine learning techniques can be employed to infer attributes not available in this database, which is the main focus of this study.This work presents an application of data mining to predict the trip purpose among public transport passengers in the Metropolitan Region of Belo Horizonte, Minas Gerais, Brazil (RMBH). Various algorithms, class balancing techniques, hyperparameters, and inference methods were tested and evaluated. In the end, the chosen inference method, involving two sequential models – one for classifying primary trips and the other for classifying secondary trips – both trained with the Random Forest algorithm and the Random Oversampling class balancing technique, was deemed the most suitable. It is worth noting that the model trained to classify primary reasons for travel yielded good results, with an accuracy exceeding 90%, making it a precise tool for predicting trip purposes in smart card data. Regarding secondary purposes, it was concluded that the model needs improvement (47% accuracy), particularly concerning the data used for its training.

Keywords: Trip purpose inference. Machine learning techniques. Smart card data.

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Lista de Abreviaturas e Siglas

RMBH	Região Metropolitana de Belo Horizonte
OD	Origem-Destino
ITS	*Intelligent Transport Systems*
GPS	*Global Position System*
ANPR	*Automatic Number Plate Recognition*
CDR	*Call Detail Records*
RF	*Random Forest*
SVM	*Support Vector Machine*
BC	*Bagging Classifier*
XGB	*eXtreme Gradient Boosting*
MCC	*Matthews Correlation Coeficient*
ROS	*Random Oversampling*
SMOTE	*Synthetic Minority Over-Sampling Technique*
ADASYN	*Adaptative Synthetic Sampling*
SBE	Sistema de Bilhetagem Eletrônica

Sumário

1 Introdução

Este capítulo apresenta uma introdução ao problema abordado neste estudo. Na Seção 1.1 há uma breve contextualização da relevância deste trabalho e dos motivos que levaram ao seu desenvolvimento. Na Seção 1.2 são apresentados os objetivos gerais e específicos. Na Seção 1.3 há uma breve discussão das justificativas cabíveis em relação ao problema abordado e às bases de dados utilizadas. Na Seção 1.4 são apresentadas as principais contribuições deste trabalho em comparação aos estudos similares encontrados na literatura. Por fim, na Seção 1.5 é apresentada a organização do trabalho com um breve detalhamento dos capítulos subsequentes.

1.1 Motivação e Relevância

O transporte público é fundamental para garantir um futuro sustentável, especialmente em grandes áreas metropolitanas com populações crescentes, pois ele promove a democratização da mobilidade e um uso mais racional do solo urbano, contribuindo para a mitigação de problemas gerados pelo uso excessivo do modo de transporte individual como congestionamentos, acidentes de trânsito e poluição ambiental (FERRAZ, 2004; HENSHER; GOLOB, 2008).

A oferta de um transporte público de qualidade é essencial para incentivar seu uso e a adequação dessa oferta inicia-se com a coleta de dados para identificar a demanda atual por deslocamento. É essa informação que possibilita planejar um novo sistema de transportes, reestruturar sistemas existentes e fazer melhorias operacionais (GUERRA, 2014; BRUTON, 1979).

De acordo com Ortuzar e Willumsen (2011), as matrizes Origem e Destino (OD) são essenciais para descrever a demanda por transportes em uma determinada área. Estas matrizes são utilizadas para quantificar e sintetizar o deslocamento de pessoas e bens no espaço, fornecendo informações sobre o número de viagens realizadas entre uma área de origem e uma área de destino por um determinado período de tempo (CÁCERES; WIDEBERG; BENITEZ, 2008; GUERRA, 2014).

Durante décadas, planejadores e pesquisadores de transporte utilizaram métodos diretos de coleta de dados para identificar a demanda por transportes e obter as matrizes OD. Os métodos diretos são aqueles baseados em pesquisas amostrais realizadas por meio de observação, entrevistas ou formulários de autopreenchimento. Este tipo de levantamento de dados possui diversas vantagens, pois abrange todos os modos de transporte, possuem informações detalhadas sobre as viagens e informações socioeconômicas dos entrevistados. No entanto, apesar das vantagens e de terem se provado eficazes no planejamento de transportes, a obtenção de dados por estes métodos é dispendiosa e por isso tais pesquisas são pouco frequentes e aplicadas a uma amostra pequena, tanto em termos de observações quanto em termos de cobertura espaço-temporal (GUERRA, 2014; BAGCHI; WHITE, 2005; CHEN et al., 2016). Além disso, estas

pesquisas correspondem às características sociodemográficas e de viagem de um período de tempo específico, pois ainda é difícil coletar dados do dia a dia por períodos contínuos e a longo prazo utilizando métodos diretos devido ao custo, carga de processamento, precisão e proteção da privacidade dos entrevistados (DEVILLAINE; MUNIZAGA; TREPANIER, 2012; KUSAKABE; ASAKURA, 2014).

Com o passar do tempo e com o avanço da tecnologia aplicada ao setor de transportes, fontes de *Big Data* tornaram-se novas possibilidades para estudar os padrões de mobilidade urbana e determinar os locais de origem e destino dos deslocamentos (GARCÍA-ALBERTOS et al., 2019). Dados de telefonia móvel, de mídias sociais ou aqueles pré-coletados e armazenados pelos Sistemas Inteligentes de Transporte (ITS - do inglês *Intelligent Transport Systems*), como os dados de rastreabilidade de veículos por Sistemas de Posicionamento Global (GPS - do inglês *Global Position Systems*), Câmeras de Reconhecimento Automático de Placas (ANPR - do inglês *Automatic Number Plate Recognition*), cartões inteligentes do sistema de transporte público, detectores de loop indutivo, dentre outros, estão sendo utilizados por diversos pesquisadores como soluções aos problemas trazidos pelas metodologias tradicionais de coleta de dados (HUSSAIN; BHASKAR; CHING, 2021). No entanto, embora essas fontes de *Big Data* sejam valiosas para analisar os padrões de viagem das pessoas de forma contínua, a longo prazo e a um baixo custo, estes dados carecem de algumas informações importantes sobre a viagem e sobre os indivíduos que estão se deslocando, sendo esta uma desvantagem destas fontes de dados em comparação com os levantamentos tradicionais.

Uma das informações coletadas pelas pesquisas domiciliares que não está disponível em fontes de *Big Data* é o motivo de viagem do passageiro. Para planejadores das cidades, saber o motivo pelo qual as pessoas se deslocam proporciona um maior entendimento sobre a mobilidade urbana e sobre o fluxo de pessoas, permitindo-os decidir como as atividades devem ser distribuídas no espaço urbano e como a rede de transporte público deve ser planejada para atender aos objetivos de viagem mais diversos, não limitados às viagens obrigatórias (WANG et al., 2017). Além disso, conhecer este atributo é útil também para que estabelecimentos comerciais e planejadores de negócios e serviços entendam o comportamento de seus consumidores e de onde eles vêm. Este argumento baseia-se no pressuposto de utilidade, advindo da economia, que considera que um deslocamento ocorre apenas porque o benefício obtido ao acessar a atividade no destino supera os custos envolvidos em realizar a viagem. Nesse contexto, é importante entender como os viajantes escolhem seus destinos de viagem, especialmente para realizar atividades que são flexíveis no espaço e no tempo (WANG et al., 2017; ASLAM et al., 2020; ASLAM et al., 2021).

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo nortedor deste estudo consiste na aplicação de técnicas de mineração de dados na inferência de informações de viagem ausentes em fontes de *Big Data*, contribuindo assim com a melhoria do *input* dos modelos de demanda de viagens e com o aprimoramento do processo de tomada de decisão dos planejadores e gestores do serviço de transporte público.

1.2.2 Objetivos Específicos

- Testar três diferentes métodos de inferência, para estimar o motivo das viagens primárias e secundárias dos passageiros em dados de cartões inteligentes em um estudo de caso para a Região Metropolitana de Belo Horizonte (RMBH), sendo eles:
  - **Método 1**: modelo único de classificação multiclasse treinado para rotular um registro de viagem em uma das sete diferentes classes (residência, trabalho, escola, saúde, lazer, negócios ou outros). Neste método, apenas variáveis de viagem serão utilizadas no conjunto de preditores;
  - **Método 2**: dois modelos sequenciais de classificação, sendo o primeiro treinado para classificar um registro de viagem nos motivos (residência, trabalho, escola ou secundário) e o segundo utilizado para reclassificar somente as viagens classificadas como 'Secundário' pelo Modelo 1 nos motivos de saúde, lazer, negócios ou outros. Neste método, apenas variáveis de viagem serão utilizadas no conjunto de preditores;
  - **Método 3**: dois modelos sequenciais de classificação, conforme descrito anteriormente, que contarão com variáveis de viagem e variáveis espaciais de Pontos de Interesse (POIs) no conjunto de preditores.
- Testar, além do método de inferência, diferentes algoritmos, hiperparâmetros e técnicas de balanceamento de classes para escolher os procecimentos mais adequados para o problema em questão;
- Aplicar o modelo escolhido, ao final de todo o processo de avaliação, nos dados de cartões inteligentes para enriquecer esta base com o artibuto de motivo de viagem inferido.

1.3 Justificativa

Em relação ao problema de pesquisa abordado neste estudo, cabem duas justificativas: a primeira delas relacionadas ao tipo de informação que será inferida e a segunda relacionada à base de dados escolhida para ser enriquecida. Optou-se por inferir o motivo de viagem em função do transporte ser caracterizado como uma demanda derivada que não tem fim em si

mesmo mas sim na utilidade da atividade que será realizada no destino. Sob essa perspectiva, o alcance do objetivo proposto neste estudo contribuirá com a melhoria da interpretabilidade dos dados de cartões inteligentes, qualificando-os com uma informação que até o momento é coletada apenas por pesquisas por entrevista. Isso possibilitará acompanhar longitudinalmente a demanda de transporte público por motivo de viagem, melhorar as previsões de demanda, garantir o planejamento integrado entre o transporte público e o uso do solo, fornecer um serviço de transporte público de qualidade e, consequentemente, atrair mais usuários para o serviço. Em relação à base de dados escolhida para ser enriquecida, optou-se pela base dados de cartões inteligentes, pois esta é a fonte de dados coletados de forma passiva e contínua mais facilmentee disponível para planejadores e gestores do serviço de transporte público das cidades utilizarem no processo de planejamento.

1.4 Principais contribuições

A primeira contribuição relevante deste estudo é a aplicação da metodologia de inferência do propósito de viagem no contexto de uma cidade brasileira. Durante a revisão de literatura, não foram encontrados estudos nacionais que discutem esta prática e que considerem as particularidades dos dados de um sistema de coleta automática de tarifas do Brasil, bem como os padrões de viagem dos passageiros de transporte público do país.

Destaca-se, ainda, como uma segunda contribuição, que este estudo compara diversos procedimentos que diferem entre si pelas variáveis preditoras, algoritmos, técnicas de balanceamento e métodos de aplicação utilizados. Também não foi encontrado na literatura um estudo que compare tantos cenários, demonstrando as diferenças e os ganhos resultantes de modificações no processo de modelagem.

1.5 Organização do Trabalho

O presente estudo está organizado em capítulos, incluindo o presente. No Capítulo 2 são apresentados alguns dos principais conceitos necessários ao entendimento deste trabalho, abordando uma discussão sobre a importância das matrizes OD no planejamento de transportes, definições, vantagens e desvantagens de diferentes formas de estimar matrizes OD e uma descrição dos principais atributos de caracterização de viagens. No Capítulo 3 as definições teóricas e matemáticas que norteiam o aprendizado de máquina são apresentadas, dando especial destaque às técnicas de pré-processamento, algoritmos, métricas de avaliação e técnicas de balanceamento de classes utilizados neste estudo. No Capítulo 4 é apresentado o estado da arte em relação ao problema de pesquisa abordado neste estudo, que consiste na previsão de motivos de viagem de passageiros do transporte público. Neste capítulo são discutidos os estudos mais recentes sobre este tema apresentando as metodologias e resultados obtidos por cada um dos autores. No Capítulo 5 o fluxo de trabalho aplicado neste estudo para prever o motivo de viagem dos

passageiros de transporte público é apresentado detalhadamente com o auxílio de um estudo de caso para a RMBH, um aglomerado de 34 municípios situados no estado de Minas Gerais, Brasil. No Capítulo 6, os resultados são apresentados e discutidos. No Capítulo 7 uma amostra de uma base de dados de cartões inteligentes é enriquecida com o motivo de viagem dos passageiros, utilizando a metodologia proposta neste estudo. Por fim, no Capítulo 8, são apresentados os tópicos conclusivos do trabalho, além de suas limitações e indicações de trabalhos futuros.

2 Matrizes OD no contexto do planejamento de transportes: conceitos e métodos

Este capítulo traz conceitos fundamentais e relevantes sobre a estimativa das matrizes OD, componente essencial do planejamento de transportes. A primeira seção apresenta uma descrição geral do que são matrizes OD e sua importância. A segunda seção descreve diferentes métodos de coleta de dados para estimativa destas matrizes. Por fim, na terceira seção são detalhados os atributos principais de caracterização das viagens que idealmente devem estar disponíveis nas fontes de dados utilizadas na produção de matrizes OD para garantir a precisão e eficiência do planejamento da rede de transporte.

2.1 Importância das matrizes OD

Por muitos anos, o planejamento de transportes esteve focado no curto prazo, com decisões que se baseavam na experiência profissional dos planejadores ou em pequenos conjuntos de dados amostrais, coletados por metodologias diretas. Estes longos períodos de planejamento fraco e limitado, trouxeram à tona a percepção de que os velhos problemas de transporte não desaparecem com o passar do tempo, ao contrário, eles ressurgem de forma mais vigorosa, ampla, complexa e difícil de lidar (ORTUZAR; WILLUMSEN, 2011).

Congestionamentos, poluição, exclusão socioespacial, acidentes e insuficiência de infraestrutura viária são alguns dos desafios causados por um sistema de transporte mal planejado. No entanto, o avanço da tecnologia da informação, a disponibilidade de grandes conjuntos de dados e a expansão do poder computacional, incentivaram um novo cenário de planejamento de transportes, cujas soluções técnicas são mais confiáveis (ORTUZAR; WILLUMSEN, 2011).

Em linhas gerais, o objetivo do planejamento de transportes é satisfazer uma demanda por deslocamento altamente qualitativa, dinâmica e diferenciada, seja pela finalidade da viagem, pelo momento do dia em que ela é realizada, pelo modo de transporte utilizado ou pelas características espaciais do deslocamento. Nesse contexto, a estrutura básica de desenvolvimento das soluções de transporte tem como ponto de partida a coleta de dados para identificar e caracterizar a demanda atual pelo serviço (ORTUZAR; WILLUMSEN, 2011; CAMPOS, 2013; CÁCERES; WIDEBERG; BENITEZ, 2008; GUERRA, 2014; BRUTON, 1979).

Para caracterizar a demanda por transporte, inicialmente é importante entender o conceito de viagem. Uma viagem pode ser definida como o movimento feito por um passageiro, que parte de um ponto de origem em direção a um ponto de destino em um movimento único ou dividido por trechos, provocado por uma causa específica e viabilizado por um ou mais modos de

transporte (ALSGER et al., 2018). As informações das inúmeras viagens realizadas pelas pessoas são sintetizadas na matriz OD, uma matriz bidimensional que fornece o fluxo de passageiros entre qualquer par de origem e destino em uma rede de transporte (ORTUZAR; WILLUMSEN, 2011; CAMPOS, 2013; CÁCERES; WIDEBERG; BENITEZ, 2008; GUERRA, 2014; BRUTON, 1979). A partir destes fluxos, é possível analisar a capacidade do sistema de transporte nos horários de pico e fora pico, os padrões de viagem e o comportamento dos passageiros. Por isto, as matrizes OD são um elemento chave do modelo clássico de planejamento de transportes, seja para estruturar um novo sistema de transportes, reformular sistemas existentes ou fazer melhorias operacionais (KANG; ATAEIAN; AMIRIPOUR, 2020; ALSGER et al., 2018).

2.2 Métodos de coleta de dados para estimativa de matrizes OD

Desenvolver uma rede de transporte otimizada requer informações de viagem de alta qualidade. Guiados pela crescente pressão de melhoria do planejamento dos sistemas de transporte, os métodos de coleta de dados de viagem evoluíram consideravelmente ao longo dos anos e o acesso a informações de viagem em tempo real está cada vez mais comum em todo o mundo, possibilitando melhorar significativamente a precisão das matrizes OD. A seguir, serão discutidos os métodos mais comuns de coleta de dados para estimar estas matrizes.

2.2.1 Métodos tradicionais de coleta de dados

Durante décadas, planejadores e pesquisadores de transporte utilizaram métodos diretos de coleta de dados para identificar a demanda por transportes e obter as matrizes OD. Duas categorias principais separam os métodos diretos de coleta de dados, sendo elas as pesquisas domiciliares e as contagens de tráfego (ALSGER et al., 2018; GUERRA, 2014; BAGCHI; WHITE, 2005; CHEN et al., 2016).

2.2.1.1 Pesquisas Domiciliares

Dentre as fontes tradicionais de dados comumente utilizadas para produzir matrizes OD, as pesquisas domiciliares são as mais completas. Nelas, o entrevistado responde a questionários de comportamento de viagem, fornecendo informações detalhadas sobre todos os deslocamentos realizados no dia anterior à execução das entrevistas, devendo este ser um dia típico que represente o padrão de deslocamento de rotina destes indivíduos. Além disso, no mesmo questionário são incluídas perguntas sobre as características econômicas e sociais do respondente, como renda, idade, escolaridade, posse de veículos, dentre outras, tornando os dados provenientes deste tipo de levantamento ainda mais ricos (CAMPOS, 2013).

Em relação ao método de coleta destes dados, as primeiras pesquisas domiciliares foram realizadas por meio de entrevistas em papel e lápis (PAPI). Com o passar do tempo, outras formas coleta surgiram, como entrevistas por telefone, internet ou alguma combinação entre elas. Além

disso, em relação à forma de condução das entrevistas, as pesquisas domiciliares podem contar com um entrevistador ou serem auto-administradas, sendo que esta última opção, em geral, está associada a maiores problemas de subnotificação ou à não conclusão do inquérito, visto que o entrevistador não está presente para levar o entrevistado a pensar com mais cuidado nas respostas ou à finalizar o preenchimento do questionário (STOPHER; FITZGERALD; XU, 2007)

Dentre as vantagens das pesquisas domiciliares, está a abrangência de todos os modos de transporte e a coleta de informações detalhadas sobre os deslocamentos e sobre os entrevistados. No entanto, apesar das vantagens e de terem se provado eficazes no planejamento de transportes por um período, um dos principais problemas deste tipo de dado é que a capacidade das pessoas de relatar com precisão as informações solicitadas interfere diretamente em sua qualidade e confiabilidade (STOPHER; GREAVES, 2007).

Diversos estudos utilizaram a coleta de dados por GPS para validar os resultados de pesquisas domiciliares. No estudo de Stopher, FitzGerald e Xu (2007), os participantes receberam um equipamento GPS que deveriam carregar consigo durante o período de estudo. Além disso, uma pesquisa domiciliar foi realizada com os mesmos indivíduos. Em relação às respostas dadas na pesquisa, Stopher, FitzGerald e Xu (2007) concluíram que os horários de início e término das viagens foram frequentemente arredondados pelos entrevistados para 5 ou 10 minutos mais próximos, no entanto, os horários relatados combinaram bem com os que foram coletados pelo GPS. Por outro lado, tanto o tempo de viagem quanto a distância percorrida foram altamente superestimados pela maioria dos entrevistados. Em relação à subnotificação de deslocamentos, observou-se que as principais viagens omitidas pelas pessoas foram as mais curtas ou aquelas cujo tempo de permanência no destino era pequeno. De forma geral, este estudo demonstrou a dificuldade das pessoas em fornecer relatórios precisos das informações acerca de suas viagens, sendo este problema agravado para algumas características específicas do deslocamento como a duração e distância percorrida.

Juntamente à imprecisão dos dados coletados, o alto custo envolvido nas pesquisas domiciliares aparece como outra importante desvantagem deste método, pois, para diminuir os gastos financeiros e de recursos humanos que tornam tais pesquisas extremanente dispendiosas, a solução comumente utilizada é diminuir substancialmente a frequência de realização dos levantamentos e a amostra pesquisada tanto em termos de observações quanto em termos de cobertura espaço-temporal (GUERRA, 2014; BAGCHI; WHITE, 2005; CHEN et al., 2016).

Por fim, outra desvantagem das pesquisas domiciliares é que os dados coletados correspondem às características sociodemográficas e de viagem de um período de tempo específico, porque é difícil levantar dados do dia a dia por períodos contínuos e a longo prazo, utilizando métodos de entrevista, devido ao custo, carga de processamento, precisão e proteção da privacidade dos entrevistados (DEVILLAINE; MUNIZAGA; TREPANIER, 2012; KUSAKABE; ASAKURA, 2014). Assim, estes métodos convencionais baseados em pesquisa não captam as rápidas mudanças nos padrões de viagem dos indivíduos e não são eficazes para suportar o dinamismo da demanda por deslocamento da cidade.

2.2.1.2 Contagens de Tráfego

As contagens de tráfego, que podem ser realizadas de forma automática ou manual, registram o número de veículos que passam por um ponto ou entram em um cruzamento (ALSGER et al., 2018). Existem diversas metodologias de estimativa de matrizes OD a partir de contagens de tráfego, cujo princípio fundamental baseia-se na combinação de uma matriz de viagem e um padrão de escolha de rotas, fornecendo assim informações diretas sobre a soma de todos os pares OD que utilizam as ligações pesquisadas (ORTUZAR; WILLUMSEN, 2011).

A etapa principal na estimativa de matrizes OD utilizando dados de contagens de tráfego é a identificação dos caminhos percorridos durante as viagens desde a origem até o destino. Define-se a variável Pij como sendo a proporção de viagens que saem da zona i e chegam na zona j viajando por meio de um determinado caminho. O fluxo de viagem nesse caminho consiste na soma das contribuições de todas as viagens entre zonas que passam por este trajeto (ORTUZAR; WILLUMSEN, 2011).

Dentre as principais vantagens deste método de recolha de dados estão o baixo custo e a possibilidade de coletar dados de forma automática e sem a necessidade de interação com o usuário, sendo, portanto, um método mais conveniente do que as pesquisas domiciliares. No entanto, assim como as pesquisas domiciliares, as contagens de tráfego, apesar de úteis para estimar matrizes OD, também sofrem com questões que limitam e afetam a exatidão dos dados recolhidos, como erros de contagem, contagem assíncrona, disponibilidade de dados em um número suficiente de ligações de rede e problemas meteorológicos. Além disso, apesar destes levantamentos serem de baixo custo, os períodos de contagem são curtos e são necessários dados adicionais, como, por exemplo, uma matriz OD inicial, para estimar o par origem-destino dos movimentos coletados (ORTUZAR; WILLUMSEN, 2011; ALSGER et al., 2018).

2.2.2 Métodos contemporâneos de coleta de dados

Os métodos contemporâneos de coleta de dados têm evoluído significativamente com os avanços tecnológicos aplicados ao setor de transportes. As principais diferenças destes métodos em relação aos métodos tradicionais incluem a utilização de tecnologias mais avançadas, automação de processos e integração de fontes de dados diversas, tais como: dados de telefonia móvel, mídias sociais ou de Sistemas Inteligentes de Transporte (ITSs). Estes últimos incluem dados de rastreabilidade de veículos por GPS, câmeras de ANPR, cartões inteligentes do sistema de transporte público, detectores de loop indutivo, dentre outros. A utilização destes dados na elaboração de matrizes OD permite uma análise mais granular e detalhada dos padrões de viagem. Além disso, esses dados reduzem a necessidade de intervenção humana e possibilitam uma análise mais eficiente e escalável (HUSSAIN; BHASKAR; CHING, 2021).

2.2.2.1 Cartões Inteligentes

Desde os anos 2000, quando houve a implementação de sistemas de cobrança automática de tarifa no transporte público coletivo, os dados cartões inteligentes têm sido foco de diversos estudos que buscam analisar seu potencial para rastrear as atividades e padrões de viagem dos passageiros a longo prazo, já que estes cartões possuem um identificador único. Grande parte destes estudos concentram-se na utilização destes dados na estimativa de matrizes OD que podem ser obtidas de forma ágil e com custo inferior se comparada às matrizes OD produzidas utilizando dados coletados por metodologias tradicionais (BARRY et al., 2002; BAGCHI; WHITE, 2005; ZHAO; RAHBEE; WILSON, 2007; TREPANIER; TRANCHANT; CHAPLEAU, 2007; PELLETIER; MORENCY, 2011; MUNIZAGA; PALMA, 2012; KUHLMAN, 2015; CHEN et al., 2016; HUSSAIN; BHASKAR; CHING, 2021).

Segundo Alsger et al. (2018), a principal metodologia de obtenção de matrizes OD a partir de dados de cartões inteligentes é a de encadeamento de viagens. Alguns pressupostos básicos são considerados por esta metodologia, sendo eles:

- um cartão é ligado a um único usuário;
- um passageiro não andará por longas distâncias para chegar ao seu próximo ponto de embarque;
- um passageiro não usará outro modo de transporte no intervalo de dois embarques consecutivos;
- um passageiro precisará fazer pelo menos duas validações do cartão para que a origem e destino da viagem sejam identificadas.

A partir dos pressupostos estabelecidos, na metodologia de encadeamento uma lista de viagens dos passageiros é criada e considera-se como destino de cada viagem o local onde o passageiro embarcou na viagem subsequente (ALSGER et al., 2018).

Dentre as vantagens da utilização de dados de cartões inteligentes na estimativa de matrizes OD, estão a maior cobertura espaço temporal e a possibilidade de coletar dados de forma contínua, constante e passiva, com interação mínima do usuário e com dimensões que jamais poderiam ser obtidas pelas pesquisas domiciliares (GARCÍA-ALBERTOS et al., 2019; BAGCHI; WHITE, 2005). Por outro lado, têm-se a desvantagem de que as informações sobre o destino do deslocamento, motivo de viagem e características socioeconômicas e demográficas dos usuários não são diretamente coletadas, sendo necessário aplicar algoritmos e técnicas de inferência para deduzi-las. Isso acontece pois, ao contrário das pesquisas domiciliares, os cartões inteligentes não possuem como função principal o fornecimento de dados para o planejamento de transportes, mas sim a função de facilitar o processo de pagamento da tarifa e trazer melhorias operacionais ao serviço (PELLETIER; MORENCY, 2011; BAGCHI; WHITE, 2005).

Outra desvantagem significativa dos dados dos cartões inteligentes é que eles contemplam apenas os deslocamentos feitos por transporte público coletivo e, por isso, não possuem

completude suficiente para subsidiarem integralmente o planejamento de redes de transporte multimodais (EGU; BONNEL, 2020; PELLETIER; MORENCY, 2011).

2.2.2.2 Telefonia Móvel

Com a popularização dos aparelhos celulares, que têm se tornado uma das necessidades da vida diária, dados coletados por estes dispositivos surgiram como uma nova possibilidade para identificar padrões de viagem de passageiros independente do modo de transporte. Assim como no caso dos cartões inteligentes, esses dados podem ser coletados de forma passiva, contínua, com interação mínima do usuário, em alto volume e a um baixo custo (WANG et al., 2010; WAG; BROEK; MULINAZZI, 2013; WU; YANG; JING, 2016; KYAING; LWIN; SEKIMOTO, 2020).

Existem dois tipos de dados de telefonia móvel que podem ser utilizados no estudo de padrões de viagem e estimativa de matrizes OD: os dados de localização por GPS e os dados de *Call Detail Records* (CDR). Os dados de GPS são mais precisos e captam informações sobre localização, tempo, velocidade e direção do movimento com uma frequência de um registro a cada dois segundos. Já os dados de CDR consistem em registros de computador produzidos por uma central telefônica contendo detalhes de uma chamada que passou por ela, por isto, este tipo de dado é coletado sempre que uma chamada é realizada. Dados de CDR contém informações sobre o número do chamador e do chamado, hora de início, duração, tipo de chamada (ex: voz, SMS), resultado da chamada (ex: atendida, ocupado), identificador da central, dentre outras (WANG et al., 2010; WAG; BROEK; MULINAZZI, 2013).

Comparativamente aos dados de GPS, os dados de CDR são mais baratos e podem ser coletados de maneira mais conveniente para o usuário, pois um dos problemas da coleta dos dados de GPS dos celulares é o alto consumo de bateria dos aparelhos quando esta funcionalidade está ativada. Isso faz com que muitos usuários escolham manter o GPS de seus aparelhos telefônicos desligados. Por outro lado, os dados de CDR sempre são coletados, independente do propósito, o que os torna uma fonte de dados secundária promissora para ser utilizada na estimativa de matrizes OD (KYAING; LWIN; SEKIMOTO, 2020).

Os dados de CDR utilizados para estimar as matrizes OD devem conter, no mínimo, um identificador único por aparelho telefônico (podendo este ser um ID criptografado) e as informações de localização e horário em que a torre de sinal do celular recebeu os registros de chamada ou mensagem de texto. A partir destes dados, outros atributos são calculados, dentre eles a diferença de tempo, distância e velocidade entre dois registros consecutivos, o estado de movimento do celular e o tempo de permanência nas paradas (WAG; BROEK; MULINAZZI, 2013). A partir das informações de estado de movimento e tempo de permanência, são determinadas as viagens, identificadas sempre que um usuário move-se para outra região após um longo tempo de permanência na região de origem (WAG; BROEK; MULINAZZI, 2013). Terminada a etapa de determinação das viagens, um banco de dados é criado contendo o número sequencial da viagem, o código das áreas de origem e destino, o tempo de permanência no destino, o tempo de viagem e os horários de saída da origem e chegada ao destino (WAG;

BROEK; MULINAZZI, 2013).

As matrizes OD estimadas a partir de dados de dados de telefonia móvel, semelhante às matrizes OD geradas a partir de dados de cartões inteligentes, carecem de informações importantes sobre a viagem, que precisam ser inferidas. Dentre estas informações estão o motivo de viagem, informações socioeconômicos dos usuários e o modo de transporte utilizado, sendo este último atributo o foco da maioria dos estudos que envolvem estes dados (HUANG; CHENG; WEIBEL, 2019).

Dentre as desvantagens de se utilizar dados de CDR para estimar matrizes OD, estudos recentes apontaram a precisão espacial grosseira, o que dificulta a inferência de modos de transporte, a redução da resolução espacial em áreas menos urbanizadas devido à redução do número de antenas, o tamanho de amostra reduzido em função da obtenção de dados de uma operadora específica, o fato dos dados serem coletados a partir de eventos, sendo registrados apenas quando o indivíduo usa ativamente seu aparelho telefônico e, uma das mais importantes limitações, a falta de dados de verdade para validação das metodologias de inferência. No contexto do aprendizado de máquina, o termo "dados de verdade", refere-se a um conjunto de dados corretamente rotulados, que servem como referência para avaliar o desempenho de um modelo de inferência. Por isto, ao utilizar dados de CDR para inferir atributos de viagem como modo de transporte ou motivos de viagem, a falta de uma base corretamente rotulada com estes atributos apresenta-se como uma importante desvantagem que pode afetar a interpretabilidade e a confiabilidade dos modelos treinados, dada a dificuldade de validação das inferências feitas (HUANG; CHENG; WEIBEL, 2019; LANDMARK et al., 2021).

2.3 Descrição de atributos das matrizes OD

Na Seção 2.2 foram apresentadas diferentes metodologias de coleta de dados que podem ser utilizados para criar matrizes OD, bem como suas vantagens e desvantagens em relação às formas de obtenção, custos, carga de processamento, interação do usuário, resolução espacial e temporal, tamanho da amostra, dentre outras. No entanto, uma comparação importante a ser feita entre estas diferentes fontes de dados é relacionada aos atributos de viagem disponíveis em cada uma delas.

Diversos atributos podem ser utilizados para caracterizar um deslocamento, dentre os quais se pode citar o seu propósito, a hora do dia em que acontece, o modo de transporte utilizado e as características socioeconômicas dos indivíduos que estão se deslocando (BRUTON, 1979). Nesse contexto, o Quadro 1 apresenta a comparação da disponibilidade destes atributos em cada um dos conjuntos de dados descritos na Seção 2.2.

Quadro 1 – Atributos de viagem comumente disponíveis por fonte de dados

Atributo	Pesquisas Domiciliares	Contagem de Tráfego	Cartões Inteligentes	Telefonia Móvel
Motivo de viagem	✓	X	X	X
Modo de transporte	✓	✓	✓	X
Atributos socioeconômicos	✓	X	X	X
Atributos temporais	✓	✓	✓	✓

Fonte: Autoria própria

Os dados das pesquisas domiciliares são os únicos que, em geral, possuem todos os atributos descritos no Quadro 1 (STOPHER et al., 2006). Isto acontece porque estas pesquisas são projetadas exclusivamente para coletar dados de viagem e, por isso, todas as informações necessárias para caracterizar uma viagem fazem parte do escopo do levantamento de dados. No entanto, as desvantagens deste método de coleta, já discutidas na Subsubseção 2.2.1.1, em alguns momentos superam a vantagem da completude das informações.

Por outro lado, os dados de cartões inteligentes e de telefonia móvel, por não terem como finalidade o estudo de padrões de viagem e a estimativa de matrizes OD, carecem de informações importantes sobre os deslocamentos, as quais precisam ser inferidas (ZANNAT; CHOUDHURY, 2019). No caso dos dados de cartões inteligentes, é necessário inferir o motivo de viagem e os atributos socioeconômicos dos passageiros, e no caso dos dados de telefonia móvel adiciona-se a necessidade de inferência do modo de transporte.

Diante do exposto, esta seção tem por objetivo conceituar cada um dos atributos apresentados no Quadro 1 e apresentar estudos que propõe métodos de inferência destes atributos, caso eles não estejam disponíveis nas bases de dados.

2.3.1 Motivo de Viagem

Uma das principais premissas do pensamento convencional do transporte é a de que viajar é uma demanda derivada da atividade a ser acessada no destino. Este argumento está alinhado ao ponto de vista utilitarista da racionalidade econômica de que a viagem ocorre apenas porque os benefícios obtidos no destino superam os custos financeiros e de tempo para que o indivíduo chegue até lá (BANISTER, 2018).

Na prática, entender como os indivíduos acessam as atividades distribuídas no espaço urbano e como eles se comportam durante suas viagens é essencial para que os tomadores de decisão, planejadores e órgãos governamentais gerenciem a distribuição dos recursos urbanos, comerciais e de transporte. Além disso, em caso de emergências no âmbito da saúde pública, a exemplo do que foi recentemente observado com a disseminação da COVID-19, o acompanhameno longitudinal da demanda de viagens por motivo pode contribuir para a otimização das ações de contingência (ASLAM et al., 2021).

Em geral, os motivos de viagem são categorizados em: residência, trabalho, escola, saúde, lazer, negócios, e outras viagens. As três primeiras, são conhecidas como viagens primárias ou

obrigatórias e as demais são denominadas viagens discricionárias ou secundárias (ORTUZAR; WILLUMSEN, 2011).

Dada a ausência do motivo da viagem em bases de dados como a de telefonia móvel e a de cartões inteligentes, diversos estudos que propõe metodologias de inferência deste atributo surgiram na literatura. Estes estudos são divididos em dois grupos, de acordo com o método de inferência utilizado que pode ser baseado em regras ou em modelos (HUSSAIN; BHASKAR; CHING, 2021).

Em geral, os métodos baseados em regras utilizam heurísticas sustentadas pelo conhecimento de domínio dos pesquisadores para determinar categoricamente os motivos de viagem (ZHAO; KOUTSOPOULOS; ZHAO, 2020; FAROQUI; MESBAH, 2021; HOSSAIN; HABIB, 2021). A maioria dos estudos que empregaram este método ((CHU; CHAPLEAU, 2008; DEVILLAINE; MUNIZAGA; TREPANIER, 2012; ZOU et al., 2018; ASLAM et al., 2021)), utilizaram o horário de início da viagem e a duração da atividade, ou o tempo de permanência no destino, para inferir o propósito mais provável de uma viagem. Por exemplo: se o horário de partida da viagem é de manhã, com tempo de permanência no destino de cerca de 8h, é provável que esta seja uma viagem à trabalho. De maneira análoga, se a próxima viagem deste passageiro inicia-se à noite, com período de permanência no destino superior a 8h, é provável que o objetivo desta segunda viagem seja de retorno à residência (ZOU et al., 2018).

Os algoritmos baseados em regras, apesar de servirem como base para métodos mais complexos, são frequentemente criticados pela falta de generalização, que dificulta a replicação do método em outros contextos (HOSSAIN; HABIB, 2021). Diante disso, os métodos baseados em modelos surgiram como uma possível solução para este problema. Métodos baseados em modelos usam algoritmos de aprendizado de máquina como Árvore de Decisão (LEE; HICKMAN, 2014; ALI; LEE, 2017; ALSGER et al., 2018), *Random Forests* (KIM; KIM; KIM, 2021; ZHU, 2021), Redes Neurais Artificiais (ASLAM et al., 2021), Redes Neurais Bayesianas (KUSAKABE; ASAKURA, 2014; KUSAKABE; ASAKURA,), além de técnicas de clusterização (LEE; HICKMAN, 2014; LU; KHANI; HAN, 2018; YANG et al., 2019; FAROQUI; MESBAH, 2021; PIERONI et al., 2021) para inferir o propósito de viagem.

Em geral, os estudos de inferência do propósito de viagem são focados nas atividades primárias (DEVILLAINE; MUNIZAGA; TREPANIER, 2012; LEE; HICKMAN, 2014; HAN; SOHN, 2016; ZOU et al., 2018; LU; KHANI; HAN, 2018; KIM; KIM; KIM, 2021). Poucos (KUSAKABE; ASAKURA, 2014; ASLAM et al., 2021; FAROQUI; MESBAH; KIM, 2020) trazem informações sobre os motivos secundários de viagem.

2.3.2 Modo de Transporte

Deslocamentos podem ser realizados de diferentes modos: a pé, com veículo próprio, de ônibus, dentre outros. No que diz respeito à propriedade do veículo e sua liberdade de uso, os modos de transporte são classificados em transporte privado ou público. No transporte privado ou individual, o veículo utilizado pertence, mesmo que temporariamente, à pessoa que está dirigindo,

há completa liberdade de escolha de rotas, horário e número de passageiros transportados, o deslocamento é porta a porta e o número de passageiros transportados é pequeno. Já no transporte público, o veículo pertence à outra pessoa ou empresa, as rotas e horários são fixos e as viagens não são porta a porta, havendo necessidade de completá-las para chegar ao destino. Além disso, este é um modo de transporte de massa, utilizado por muitas pessoas simultaneamente (FERRAZ, 2004).

A escolha do indivíduo por determinado modo de transporte é influenciada por diversos fatores, tais como distância a ser percorrida, hora de início e propósito da viagem, características socioeconômicas e individuais da pessoa que está se deslocando, custo, acessibilidade, conforto, dentre outros (BRUTON, 1979).

Conhecer o modo de transporte com o qual os indivíduos se deslocam contribui significativamente para que os planejadores do serviço entendam a distribuição de viagem entre os modos disponíveis e alavanquem a mobilidade das pessoas com base em diferentes modos de transporte, prevendo a oferta de serviços de mobilidade de forma otimizada em áreas metropolitanas (ASGARI; CLEMENCON, 2018).

Dentre as fontes de dados apresentadas no Quadro 1, a de telefonia móvel é a que possui menos informações coletadas e a única que não contempla a informação do modo de transporte (ZANNAT; CHOUDHURY, 2019). Para inferir este atributo, estudos que trabalharam com dados de GPS utilizaram a taxa de mudança de direção, a frequência de paradas, a velocidade e a aceleração do movimento para identificar o modo de transporte dos segmentos de viagens. Tais estudos consideram que a velocidade e a aceleração com que um indivíduo se desloca estão dentro de uma faixa para cada modo de transporte. Além disso, a taxa de mudança de direção auxilia na diferenciação entre transporte motorizado e não motorizado, pois é muito mais fácil mudar a direção do deslocamento quando a viagem é feita por modos não motorizados como a pé e de bicicleta. Por fim, o número de paradas podem ser indicadores importantes para a diferenciação dos modos de transporte motorizados entre público e privado, já que os deslocamentos por transporte público tendem a ter muito mais paradas em comparação ao transporte individual.

Em relação ao método de inferência, existem estudos que utilizaram abordagens baseadas em regras, lógica *fuzzy* e métodos de classificação supervisionada (redes neurais e *Random Forests*) para inferir tal atributo, sendo as classificações com *Random Forests* as mais utilizadas e as que apresentaram, segundo alguns estudos, melhor desempenho para detectar o modo de transporte em dados de GPS (WANG et al., 2010; REDDY et al., 2010; QUEIXO et al., 2019)

Já os estudos que buscaram inferir o modo de transporte em dados de CDR, utilizaram variáveis como a duração da viagem, a proximidade da rede de transporte e a velocidade da viagem para inferir modos de transporte, geralmente agregados em grupos mais gerais, como transporte público e transporte privado. Em relação ao método, a maioria destes estudos, que ainda são poucos na literatura, utilizaram abordagens baseadas em regras (WU; YANG; JING, 2016), cujos limites de decisão são estabelecidos por especialistas. Alguns utilizaram algoritmos de clusterização, como o K-Means, para agrupar os dados com base na duração ou na velocidade

de viagem (WANG et al., 2010; HUI, 2017). Em relação à avaliação, os métodos encontrados na literatura não apresentam de forma clara os resultados de validação do modo de transporte inferido. A falta de avaliação deve-se, principalmente, à falta de dados rotulados ou de referência (QUEIXO et al., 2019).

2.3.3 Características socioeconômicas e demográficas dos indivíduos

Características socioeconômicas e demográficas, como renda, idade, densidade residencial dentre outras, também influenciam no comportamento de viagem dos indivíduos. Em 1980, pesquisadores começaram a investigar viagens sob a perspectiva da ciência do comportamento. O objetivo era avaliar se as circunstâncias sociais em que os indivíduos viviam influenciavam nas oportunidades ou restrições de acesso a determinadas atividades, impactando assim em sua forma de deslocamento. Os resultados destes estudos evidenciaram que as características socioeconômicas dos indivíduos podem interferir no modo de transporte escolhido, no horário de realização das viagens, na quantidade de movimentos empreendidos, e nos seus motivos de viagem. Indivíduos idosos, por exemplo, geralmente iniciam suas viagens em horários tardios e permanecem no destino por períodos mais curtos e irregulares. Já os indivíduos mais jovens e estudantes, iniciam suas viagens pela manhã, juntamente com indivíduos que trabalham, ou no início da tarde. Estas pessoas possuem características de viagem mais bem definidas, sobretudo em relação ao tempo de permanência no destino (ORTUZAR; WILLUMSEN, 2011).

Indivíduos com maior renda, por exemplo, podem escolher usar outro modo de transporte em face do transporte público, escolha que em geral não é uma possibilidade para os usuários cativos do transporte público, ou seja, usuários que não têm oportunidade de escolher qual modo de transporte utilizarão em suas viagens devido às restrições financeiras (BRUTON, 1979).

Nesse contexto, inferir os atributos socioeconômicos em dados secundários utilizados para estimar matrizes OD contribui com a compreensão do comportamento das pessoas na rede de transporte e como cada grupo social interage com o espaço urbano (FAROQUI; MESBAH; KIM, 2020; KIM; KIM; SOHN, 2022). Em função da crescente utilização de *Big Data* na criação de soluções de transporte, diversos estudos vêm surgindo com o objetivo de extrair informações socioeconômicas em dados anônimos.

Em um estudo recente, Kim, Kim e Sohn (2022) propõe um método de estimativa de atributos sociodemográficos nos dados de cartões inteligentes usando uma rede generativa adversarial condicional, capaz de imputar as informações dos indivíduos nestes dados imitando dados de uma pesquisa de viagem domiciliar. Outros estudos utilizaram algoritmos de aprendizado de máquina como Análise de Discriminante Linear, Máquinas de Vetores de Suporte (SVM's - do inglês *Support Vector Machines*), *perceptrons* multicamadas e árvores de decisão, para enriquecer dados de CDR de telefonia móvel com atributos como gênero, idade e nível educacional (ZHAO; PAWLAK; SIVAKUMAR, 2022).

2.3.4 Hora do dia

Outro atributo caracterizador de viagens é a hora do dia em que o deslocamento ocorre. Em relação à esta característica, as viagens são divididas em dois grupos: viagens realizadas no período de pico e viagens realizadas no período fora de pico. No geral, a proporção de viagens que ocorre em cada um destes períodos varia de acordo com o motivo de viagem. O período de pico, normalmente ocorre em dois momentos do dia: o período de pico da manhã e o período de pico da noite (ORTUZAR; WILLUMSEN, 2011). Em geral, grande parte do volume de viagens que ocorre no pico da manhã são motivadas por atividades obrigatórias como trabalho e escola e o período de pico da noite é composto, majoritariamente, por viagens de residência. Nota-se, portanto, que o horário de início das viagens é fortemente influenciado pelo horário de início das atividades ofertadas no espaço urbano, sobretudo para as viagens obrigatórias, refletindo assim em padrões de viagem por horários do dia (ORTUZAR; WILLUMSEN, 2011).

O atributo de hora do dia em que ocorre o deslocamento é o único presente em todas as fontes de dados apresentadas no Quadro 1. Nesse contexto, essa é uma informação muito utilizada também para inferir outros atributos como o motivo da viagem e o modo de transporte, quando estes não estão disponíveis.

2.4 Resumo do capítulo

A oferta de um sistema de transportes de qualidade inicia-se com a coleta de dados sobre a demanda por viagens. A informação sobre o fluxo de viagens realizadas entre cada par de localidades em um determinado período de tempo pode ser sintetizada em uma matriz bidimensional denominada Matriz Origem e Destino (OD).

Para a elaboração das matrizes OD, diferentes fontes de dados podem ser utilizadas, tais como pesquisas domiciliares, transações de cartões inteligentes e dados de telefonia móvel. A principal diferença entre matrizes OD estimadas por cada uma destas fontes de dados está na disponibilidade de informações coletadas sobre os atributos caracterizadores de uma viagem.

Os atributos básicos que qualificam uma viagem são o seu propósito, a hora do dia em que acontece, o modo de transporte utilizado e as características socioeconômicas dos indivíduos que estão se deslocando. Nesse contexto, uma vantagem substancial das pesquisas domiciliares é sua completude visto que todos estes atributos estão disponíveis nesta fonte de dados. Por outro lado, desvantagens como o alto custo financeiro e de recursos humanos inerentes a este tipo de coleta fazem com que metodologias alternativas de coleta de dados para a obtenção de matrizes OD ganhem força.

Alta cobertura espaço temporal, coleta contínua, constante e passiva e a mínima necessidade de interação com o usuário são algumas das vantagens dos métodos alternativos para coletar dados que podem ser utilizados na obtenção de matrizes OD, como dados de cartões inteligentes e de telefonia móvel. Por outro lado, a principal desvantagem destas fontes de dados é a ausência de atributos importantes que caracterizam uma viagem e, para isto, a mineração de

dados, assunto tratado no próximo capítulo, pode ser empregada para viabilizar o preenchimento desta lacuna e o enriquecimento destas bases com informações não coletadas.

3 Mineração de dados: definições e conceitos

Neste capítulo, é apresentada uma visão detalhada do processo de mineração de dados, dividido em cinco subcapítulos essenciais. O capítulo se inicia com uma visão geral e abrangente do processo. Em seguida, há uma análise da compreensão do problema e dos dados disponíveis, seguida por uma abordagem das técnicas de pré-processamento e transformação de dados. Posteriormente, é abordada a modelagem de dados, onde são aplicados algoritmos e técnicas para extrair informações úteis dos dados preparados. Por fim, são explorados métodos para avaliar os modelos gerados e interpretar os resultados obtidos. Esses subcapítulos juntos formam uma estrutura sólida para a compreensão e aplicação da mineração de dados ao longo deste trabalho.

3.1 Iniciando na Mineração de Dados: uma visão abrangente do processo

O valor dos dados armazenados está intrínsecamente ligado à capacidade de se extrair conhecimento a partir deles, ou seja, informação útil que sirva para apoio à tomada de decisão, e/ou para a exploração e entendimento do fenômeno gerador dos dados (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). O processo utilizado para descobrir padrões ou tendências úteis nos dados é denominado mineração de dados e ele envolve várias etapas, conforme apresentado na Figura 1.

Figura 1 – Etapas envolvidas no processo de mineração de dados

***A ordem pode variar, de acordo com o problema a ser resolvido**

Fonte: Adaptado de Goldschmidt, Passos e Bezerra (2015) pela autora.

Atualmente, a quantidade e a complexidade dos conjuntos de dados disponíveis representam um desafio para os métodos tradicionais de processamento e análise. A explosão de dados gerados em alto volume, velocidade e variedade, fenômeno conhecido como *Big Data*, requer abordagens automatizadas capazes de detectar padrões nos dados disponíveis e, em seguida, utilizá-los para prever dados futuros, ou para dar suporte à tomada de decisão em condições de incerteza, essas abordagens são conhecidas como métodos de aprendizado de máquina (GOLDSCHMIDT; PASSOS; BEZERRA, 2015; ALPAYDIM, 2014).

Os métodos de aprendizado de máquina utilizados na mineração de dados são divididos em dois tipos principais: aprendizado de máquina supervisionado e aprendizado de máquina não supervisionado (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

No aprendizado supervisionado, o objetivo é mapear um *output*, conhecido como variável

target a partir de um ou mais *inputs* denominado(s) variável(is) preditora(s). Neste tipo de aprendizado de máquina, uma base de dados rotulada é utilizada para treinamento dos algoritmos, ela é composta pelas n variáveis preditoras e pela variável target. É a partir desta base que o algoritmo encontra padrões nas variáveis preditoras que geram determinado resultado de variável *target*. Quando a variável *target* da base de dados rotulada é categórica, o problema é conhecido como classificação e quando é numérica, o problema é conhecido como regressão (HAN; KAMBER; PEI, 2012).

No aprendizado não supervisionado, a variável *target* não está disponível e apenas as variáveis preditoras são fornecidas aos algoritmos. Neste caso, o objetivo é encontrar padrões desconhecidos nos dados e, por esta razão, este tipo de aprendizado também é denominado descoberta de conhecimento. Ao contrário da aprendizagem supervisionada, onde é possível comparar o resultado obtido pelo modelo com o valor real da variável *target*, na aprendizagem não supervisionada não existe uma métrica de erro óbvia, visto que inicialmente não se sabe o que está sendo buscado (MURPHY, 2012; BURKOV, 2019).

3.2 Entendendo o problema e os dados

A primeira etapa do processo de mineração de dados é a compreensão clara do problema que deseja-se resolver. Neste estudo, será abordado um problema típico de aprendizagem supervisionada, notadamente de classificação multiclasse. Isto porque o objetivo do trabalho é inferir o motivo de viagem dos passageiros do transporte público - variável *target* categórica - em uma base de dados de cartões inteligentes, utilizando uma base de dados rotulada para treinamento dos algoritmos.

A classificação é uma técnica da aprendizagem supervisionada cujo objetivo é aprender um mapeamento de entradas x para as saídas y, onde $y \in \{1,...,C\}$, sendo C o número de classes, ou valores possíveis para a variável *target*. Se $C = 2$, têm-se um problema de classificação binária, se $C > 2$ têm-se um problema de classificação multiclasse. Dadas as definições, um problema de classificação pode ser formulado matematicamente como $\hat{y} = f(x)$ (O símbolo $\hat{}$ denota uma estimativa) e a aprendizagem consiste em estimar a função f, com base em um conjunto de dados rotulados, que posteriormente será aplicada a novas entradas com o objetivo de fazer previsões. Isto é denominado generalização (MURPHY, 2012).

Após compreender o problema que deseja-se resolver, o passo seguinte consiste no levantamento das bases de dados necessárias e no entendimento destes dados. Para isto, gráficos e estatísticas descritivas são comumente aplicadas para identificar padrões iniciais nos dados. A análise descritiva é uma etapa fundamental no processo de modelagem usando algoritmos de aprendizado de máquina, pois ela permite identificar valores extremos, tendências, padrões, distribuição dos dados e a relação entre as variáveis, o que pode ser usado para selecionar os melhores algoritmos para o problema em questão (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Para um problema de classificação multiclasse, as análises descritivas explicadas a seguir são imprescindíveis para garantir a qualidade dos dados e o sucesso da modelagem:

3.2.1 Análise de distribuição de classes e técnicas de balanceamento

A análise de distribuição de classes consiste em checar se todas as classes possíveis para a variável *target* possuem um número razoável de exemplos pois o desbalanceamento de classes pode afetar negativamente o desempenho do modelo.

Quando os algoritmos de classificação são aplicados a dados desbalanceados, as regras de indução que descrevem os conceitos minoritários são menores e mais fracas do que as que descrevem os conceitos majoritários. Todos os dados são naturalmente desbalanceados, isto porque qualquer conjunto de dados que apresente uma distribuição desigual entre suas classes pode ser considerado desequilibrado. No entanto, na prática, apenas os conjuntos de dados que possuem desequilíbrios significativos e, em alguns casos, extremos são considerados desbalanceados. A isto é dado o nome de desequilíbrio entre classes (HE; GARCIA, 2009; HE; BAI; GARCIA, 2013).

Existem três tipos comuns de desequilíbrio: o desequilíbrio intrínseco que resulta da natureza do espaço de dados, como, por exemplo, em aplicações biomédicas onde o número de pacientes saudáveis excede o número de pacientes com alguma doença; o desequilíbrio extrínseco, que não está diretamente relacionado à natureza do espaço de dados mas sim à fatores variáveis como, por exemplo, tempo e armazenamento; e, o desequilíbrio relativo, onde a classe minoritária não é necessariamente rara por si só, mas é rara em relação à classe majoritária. Desequilíbrios relativos são frequentes em problemas do mundo real e, em geral, são o foco de muitos esforços de descoberta de conhecimento e pesquisa de engenharia de dados (HE; GARCIA, 2009).

Para resolver o problema do desbalanceamento de classes, métodos de amostragem podem ser aplicados para modificar o conjunto de dados de forma a tornar sua distribuição balanceada. Isso ajudará na precisão do classificador. Quando a técnica aplicada é de sobreamostragem dá-se o nome de *Oversampling* e quando é de subamostragem, *Undersampling*. Cada um destes métodos introduz seu próprio conjunto de consequências no aprendizado. No caso da subamostragem, o problema principal é que ao remover exemplos da classe majoritária, o classificador pode perder conceitos importantes relativos a esta classe. Já no caso do *Oversampling*, ao anexar dados replicados ao conjunto de dados original algumas regras podem se tornar muito específicas e levar o modelo ao *overfitting*[1]. Quando isto acontece, a precisão de treinamento é alta, mas o desempenho da classificação nos dados de teste é significativamente menor (HE; GARCIA, 2009; HE; BAI; GARCIA, 2013).

[1] O termo *overfitting* é utilizado para descrever um modelo com baixo erro de treinamento e alto erro de teste. Ou seja, é um modelo com excesso de ajuste aos dados de treinamento e com falta de generalização

3.2.1.1 *Random Undersampling and Oversampling*

Uma das técnicas mais simples de balanceamento de classes é a de amostragem aleatória. Esta técnica pode ser utilizada tanto para a subamostragem, e neste caso é denominada de *Random Undersampling*, quanto para sobreamostragem, neste âmbito denominada *Random Oversampling*. Em ambos casos, a amostragem aleatória possui uma abordagem simples de balanceamento de classes que consiste remover amostras aleatórias da classe majoritária ou replicar instâncias aleatórias da classe minoritária até que o equilíbrio seja atingido (HE; GARCIA, 2009).

Por se tratar de uma técnica simples, nenhuma análise é feita durante o processo de amostragem aleatória independente, se ele é de sobre ou subamostragem, e isto está associado a dois principais problemas: na técnica de *Random Oversampling* as amostras selecionadas para serem replicadas podem já estar próximas umas das outras, o que pode levar a um risco maior de sobrevalorização de determinadas regiões do espaço de amostragem e um maior risco de *overfitting*, e na técnica de *Random Undersampling* as amostras selecionadas para serem removidas podem ser mais informativas do que as amostras que permanecerão no conjunto de dados (HE; GARCIA, 2009).

3.2.1.2 *Informed Undersampling*

Para resolver o principal problema da técnica de *Random Undersampling*, que é a perda de informação decorrente do processo de remoção aleatória de instâncias de dados, foi criada a *Informed Undersampling*. A principal ideia por trás desta técnica é selecionar estrategicamente quais instâncias da classe majoritária devem ser mantidas e quais devem ser removidas, de forma a preservar informações relevantes para a classificação correta. Isso é feito levando em consideração características, padrões ou propriedades das instâncias, a fim de evitar a perda de informações importantes (HE; GARCIA, 2009).

3.2.1.3 SMOTE e ADASYN

Dado o principal problema da técnica de *Random Oversampling*, que é o aumento do risco de *overfitting* em função da replicação aleatória das instâncias de dados da classe minoritária, outras técnicas podem ser aplicadas para mitigação deste ponto, como é o caso das técnicas SMOTE *(Synthetic Minority Over-sampling Technique)* e ADASYN *(Adaptive Synthetic Sampling)* (HE; GARCIA, 2009).

O algoritmo SMOTE trata o aumento da classe minoritária a partir da geração de amostras sintéticas com base nas amostras existentes. Uma das vantagens desta técnica sobre a técnica de *Random Oversampling* é que ela ajuda a aumentar a representação da classe minoritária introduzindo diversidade nos dados sintéticos, ajudando assim a evitar o *overfitting*. No entanto, quando as instâncias de diferentes classes estão localizadas em regiões semelhantes do espaço de recursos, a geração de amostras sintéticas pela técnica SMOTE pode introduzir erros na região de fronteira entre as classes, levando a um desempenho de classificação inferior e trazendo mais

ruídos ao modelo. Nesse contexto, surge a técnica ADASYN como uma extensão do algoritmo SMOTE que ajusta a taxa de geração de amostras sintéticas com base na densidade local das amostras, ou seja, esta é uma técnica que enfoca a geração de amostras nas regiões de fronteira, onde a separação entre as classes é menos clara (HE; GARCIA, 2009).

3.2.2 Análise de distribuição das variáveis preditoras

Antes de selecionar quais algoritmos serão treinados na etapa de modelagem, descrita na Seção 3.4, deve-se checar se os algoritmos pré-selecionados fazem alguma suposição sobre os dados de entrada. Muitos algoritmos de aprendizado de máquina fazem suposições sobre a distribuição dos dados de entrada e sobre a quantidade de parâmetros que precisam ser estimados a partir dos dados, esses algoritmos têm o nome de algoritmos paramétricos (IZBICKI; SANTOS, 2020).

Algoritmos paramétricos fazem suposições específicas sobre a distribuição subjacente dos dados. Eles assumem que os dados seguem uma forma de distribuição predefinida e por isto estes algoritmos são menos flexíveis e podem não se ajustar bem a dados que não se encaixam nas suposições da distribuição. Exemplos de algoritmos paramétricos incluem Regressão Linear, Regressão Logística e *Naive Bayes* (IZBICKI; SANTOS, 2020).

Por outro lado, algoritmos não paramétricos não fazem suposições específicas sobre a distribuição dos dados. Eles permitem que os dados determinem a estrutura do modelo e por isto estes algoritmos são mais flexíveis e podem capturar relações complexas nos dados, tornando-os úteis para dados não lineares ou com distribuições desconhecidas. Além disso, como não fazem suposições rígidas sobre a distribuição dos dados, os algoritmos não paramétricos são menos sensíveis a suposições errôneas. Exemplos de algoritmos não paramétricos incluem Árvores de Decisão, *Random Forest* e Máquinas de Vetores de Suporte com kernel não linear (IZBICKI; SANTOS, 2020).

Para escolher entre algoritmos paramétricos ou não paramétricos, a análise da distribuição das variáveis é uma etapa imprescindível. Para isto, técnicas como gráfico de box-plot, tabela de frequência, histogramas ou gráficos de densidade podem ser utilizados, dependendo do tipo de dado que se deseja analisar a distribuição (numérico ou categórico) (BRUCE; BRUCE; GEDECK, 2020).

Ainda em relação à distribuição das variáveis, além das técnicas descritas acima, é comum que testes estatísticos sejam aplicados principalmente nos casos em que se deseja validar se determinada distribuição segue os padrões de uma distribuição normal. Uma observação é normal quando seu comportamento segue uma distribuição de Gauss (em forma de sino) e possui um determinado valor de média μ e variância σ.

Um teste estatístico comumente utilizado para verificar a presença de normalidade nos dados é o teste de *Shapiro–Wilk*. Este teste analisa os dados observados e calcula o nível de simetria e curtose (formato da curva) para depois calcular a diferença em relação a uma distribuição gaussiana, obtendo o valor p a partir da soma dos quadrados das discrepâncias

(BRUCE; BRUCE; GEDECK, 2020).

No teste de Shapiro-Wilk duas hipóteses são definidas:

- **Hipótese Nula** (H_0): A amostra segue uma distribuição normal
- **Hipótese Alternativa** (H_1): A amostra não segue uma distribuição normal

Após a definição das hipóteses, uma estatística de teste (W) é calculada com base nos desvios dos dados em relação a uma distribuição normal esperada. Em um teste de *Shapiro-Wilk*, a estatística de teste (W) indica quão bem os dados se ajustam a uma distribuição normal. Quanto mais próximo de 1 for o valor de W, mais próximo os dados estão de uma distribuição normal. (PINHEIRO et al., 2009)

Em seguida, o valor p, que corresponde à probabilidade de obter uma estatística de teste igual ou superior à que foi observada é calculado ou encontrado pela comparação da estatística de teste W calculada com uma tabela de valores críticos (PINHEIRO et al., 2009).

Para interpretar o valor de p obtido, este valor é comparado a um nível de significância predefinido (geralmente 0,05 ou outro valor escolhido). O nível de significância é a probabilidade máxima aceitável de cometer um erro do Tipo I, também conhecido como "falso positivo". Um erro do Tipo I ocorre quando erroneamente rejeita-se a hipótese nula quando ela é verdadeira. Se o valor p for menor que o nível de significância, a hipótese nula (H_0) é rejeitada, indicando que os dados não seguem uma distribuição normal. Se o valor p for maior que o nível de significância, a hipótese nula não é rejeitada, sugerindo que não há evidência suficiente para concluir que os dados não são normalmente distribuídos (PINHEIRO et al., 2009).

3.2.3 Análise de correlação ou associação entre variáveis preditoras

Outra análise imprescindível que deve ser feita antes da etapa de modelagem é a de correlação e/ou associação entre as variáveis preditoras e entre as variáveis preditoras e a variável *target*. Variáveis preditoras fortemente correlacionadas ou associadas tendem a indicar multicolinearidade e isso resulta em um "excesso"de informações redundantes no modelo (BRUCE; BRUCE; GEDECK, 2020). Por outro lado, variáveis preditoras fortemente correlacionadas ou associadas com a variável *target* podem indicar que estas são as variáveis mais importantes para que um modelo faça previsões corretas.

A correlação é utilizada quando as duas variáveis analisadas são numéricas e a associação é utilizada quando pelo menos uma delas é categórica. Diz-se que duas variáveis numéricas X e Y são positivamente correlacionadas quando valores altos de X implicam em valores altos de Y, e valores baixos de X implicam em valores baixos de Y. Se valores altos de X implicam valores baixos de Y, e vice-versa, diz-se que as variáveis são correlacionadas negativamente (BRUCE; BRUCE; GEDECK, 2020). Nesse contexto, o coeficiente de correlação varia de +1 a -1, onde valores próximos a +1 indicam que as variáveis estão correlacionadas positivamente, valores

próximos a -1 indicam que as variáveis estão correlacionadas negativamente e valores próximos a 0 indicam que as variáveis não possuem correlação (BRUCE; BRUCE; GEDECK, 2020).

De acordo com Bruce, Bruce e Gedeck (2020), existem três principais métodos para se determinar a correlação entre duas variáveis numéricas:

- Coeficiente de Correlação de Pearson (r): Também conhecido como correlação linear, é usado para medir a relação linear entre duas variáveis contínuas. É amplamente utilizado quando se deseja avaliar a força e a direção da relação linear entre duas variáveis.
- Coeficiente de Correlação de Spearman (ρ): É usado quando as variáveis são ordinais ou quando a relação entre as variáveis é monotônica, mas não necessariamente linear. Ele é menos sensível a valores extremos em comparação com o coeficiente de Pearson e é útil quando os dados não atendem aos pressupostos da normalidade.
- Coeficiente de Correlação de Kendall (τ): Também é usado para variáveis ordinais e mede a concordância entre as classificações de duas variáveis. O coeficiente de Kendall é robusto contra valores extremos e é apropriado quando a relação entre as variáveis é monotônica, mas não necessariamente linear.

Quando as duas variáveis são categóricas, a análise de associação é empregada e as tabelas de contingência são uma forma útil de realizar esta análise. Estas tabelas mostram a frequência de ocorrência de combinações de categorias para duas ou mais variáveis e a análise visual das tabelas de contingência pode revelar padrões de associação. Testes estatísticos, como o Teste de Qui-quadrado, podem ser aplicados para determinar se a associação observada é estatisticamente significativa (BRUCE; BRUCE; GEDECK, 2020).

Por fim, quando uma variável é numérica e a outra é categórica, boxplots podem ser utilizados como uma maneira simples de comparar visualmente as distribuições da variável numérica agrupada de acordo com a variável categórica (BRUCE; BRUCE; GEDECK, 2020). Outra forma de se fazer esta comparação é utilizando testes estatísticos, como é o caso do teste de *Kruskal-Wallis*. Este é um teste estatístico não paramétrico usado para avaliar se há diferenças significativas entre as medianas das variáveis numéricas entre as categorias da variável categórica. Ele é uma alternativa ao teste de análise de variância (ANOVA) quando os pressupostos da ANOVA, como a normalidade dos dados e a homogeneidade das variâncias, não são atendidos (SIEGEL, 1956).

No teste de *Kruskal-Wallis* duas hipóteses são definidas:

- **Hipótese Nula** (H_0): Não há diferenças significativas nas medianas entre os grupos.
- **Hipótese Alternativa** (H_1): Pelo menos um grupo difere significativamente dos outros em relação à variável dependente.

Após a definição das hipóteses, a estatística do teste H é calculada. Sob a hipótese nula, a estatística de teste segue uma distribuição qui-quadrado X^2, uma distribuição de probabilidade

amplamente usada em estatística. Esta distribuição é caracterizada por um parâmetro chamado "graus de liberdade", que afeta a forma da distribuição. No teste de *Kruskal-Wallis*, a distribuição de qui-quadrado possui $(k-1)$ graus de liberdade onde k é o número de categorias distintas da variável categórica. Os graus de liberdade indicam quantos grupos estão sendo comparados e seu valor afeta diretamente na forma da distribuição qui-quadrado. Quanto maior o número de graus de liberdade, mais próximo a distribuição qui-quadrado se assemelha a uma curva normal. Ao final do teste, o valor p é calculado usando a distribuição qui-quadrado. A comparação do valor p com o nível de significância escolhido leva à decisão de rejeitar ou não a hipótese nula, indicando se há diferenças significativas nas medianas entre os grupos analisados (SIEGEL, 1956).

Ao finalizar o teste de *Kruskal-Wallis*, caso a Hipótese Nula H_0 seja rejeitada, o teste *post-hoc* de comparações múltiplas entre pares, denominado teste de Dunn, pode ser aplicado para averiguar quais as medianas são significativamente distintas. Neste teste, as distribuições são comparadas duas a duas e caso o p-valor seja menor que o nível de significância escolhido, conclui-se que há uma diferença estatisticamente significativa entre as medianas para o par analisado (DINNO, 2015).

3.3 Pré-processamento e Transformação dos dados

A etapa de pré-processamento de dados compreende todas as funções relacionadas com a captação, a organização e o tratamento dos dados com o objetivo de prepará-los para serem utilizados pelos algoritmos na etapa de modelagem. As principais funções de pré-processamento e transformação são: seleção de dados, limpeza de dados, codificação de dados, enriquecimento de dados, normalização de dados, engenharia de atributos e a partição do conjunto de dados em subconjuntos de treino e teste (GOLDSCHMIDT; PASSOS; BEZERRA, 2015)

A seleção de dados pode ter dois enfoques distintos: a seleção de atributos e a seleção de registros. A seleção de atributos compreende a identificação do subconjunto de variáveis disponíveis que são relevantes para o processo de modelagem e que, portanto, devem ser efetivamente consideradas. Já a seleção de registros pode ser aplicada, por exemplo, quando o volume de dados é muito grande para ser manipulado. Neste caso, um subconjunto de registros é selecionado para ser utilizado na etapa de modelagem (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

A limpeza de dados abrange qualquer tratamento realizado sobre os dados selecionados para assegurar sua qualidade, completude, veracidade e integridade. Informações ausentes, errôneas ou inconsistentes são tratadas durante a limpeza para que não comprometam a qualidade dos modelos (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Determinados algoritmos de aprendizado de máquina requerem que os dados tenham um tipo específico, alguns deles só trabalham com valores numéricos, por exemplo. Nestes casos, os dados devem ser codificados para que possam ser utilizados como entrada nestes algoritmos. Existem diversas técnicas de codificação de variáveis categóricas para transformá-las em variáveis

numéricas, no entanto, duas delas são frequentemente utilizadas: a técnica de *One-Hot-Encoding* e a técnica de *Label Encoding*. *One-Hot-Encoding* consiste em um método de representação binária de classes ideal para ser utilizado quando a variável a ser codificada é categórica nominal. Nesta técnica, cada nível da variável categórica é codificado em uma coluna separada, que será preenchida com o valor 1 caso a observação pertença àquele nível ou com 0 caso contrário, criando assim uma matriz esparsa. Já a técnica de *Label Encoding* é um tipo de codificação útil quando a ordem da variável categórica é relevante no processo de modelagem, pois neste método as categorias são transformadas em valores numéricos ordinais, onde a ordem dos rótulos atribuídos reflete a ordem das categorias (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

O enriquecimento de dados consiste em agregar mais informações a cada registro do conjunto de dados para que o modelo tenha mais elementos para identificar um padrão assertivo. Em geral, o processo de enriquecimento é realizado ao incorporar aos dados originais informações advindas de outros sistemas ou bases de dados (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Outra etapa importante do pré-processamento é a normalização de dados. Normalizar variáveis significa ajustar valores para uma escala comum, para evitar que uma variável de escala maior possua um peso superior na análise em relação a outras variáveis com escalas menores. Isso acontece pois muitos algoritmos de aprendizado de máquina utilizam cálculo de distância entre as amostras para realizar as previsões e, por isso, a escala das variáveis pode interferir no desempenho destes algoritmos (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

A etapa do pré-processamento denominada engenharia de atributos consiste em criar novas variáveis mais informativas e relevantes a partir de variáveis pré-existentes (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Por fim, a última etapa do pré-processamento consiste na partição da base de dados em pelo menos dois subconjuntos: um conjunto de treinamento e um conjunto de testes. O conjunto de treinamento será utilizado na construção do modelo de aprendizado de máquina, uma vez construídos, a qualidade destes modelos precisa ser avaliada e, para que esta avaliação seja isenta, os dados utilizados nesta etapa não devem ser os mesmos que foram utilizados na etapa de treinamento, sendo justamente esta a utilidade do conjunto de dados de teste. Segundo Goldschmidt, Passos e Bezerra (2015), existem diversas técnicas de partição do conjunto de dados, dentre elas:

- **Holdout**: Nesta técnica os dados são divididos aleatoriamente utilizando uma percentagem p fixa para treinamento e $(1 - p)$ para teste.
- **Validação cruzada com K-conjuntos**: Este método consiste em dividir aleatoriamente o conjunto de dados em K subconjuntos disjuntos com aproximadamente o mesmo número de elementos. Neste processo, cada um dos K subconjuntos é utilizado como conjunto de teste e os $(K - 1)$ demais, são utilizados no treinamento do modelo. Este processo é repetido por K vezes.
- **Validação cruzada com K-conjuntos estratificada**: esta é uma técnica aplicada a

problemas de classificação. A diferença entre esta técnica e a técnica descrita anteriormente é que ao gerar os subconjuntos disjuntos a proporção de exemplos em cada uma das classes é considerada durante a amostragem.

- **Leave-One-Out**: Este é um caso particular da validação cruzada em que K é igual à quantidade de registros, portanto em cada um dos K conjuntos existe apenas um elemento. Este é um método computacionalmente intensivo e por isto é aplicado apenas em conjuntos de dados pequenos.

- ***Bootstrap***: Neste método, o conjunto de treinamento é gerado a partir de N sorteios aleatórios com reposição a partir do conjunto de dados original. O conjunto de teste é composto pelos registros do conjunto de dados original que não foram sorteados.

3.4 Modelagem de Dados

No aprendizado de máquina as análises são feitas a partir do uso de algoritmos, que nada mais são do que um conjunto de instruções sequenciais e lógicas aplicadas para resolver um problema específico (MURPHY, 2012; HASTIE; TIBSHIRANI; FRIEDMAN, 2008; ALPAYDIM, 2014).

3.4.1 Descrição dos algoritmos

As subseções seguintes apresentam as descrições detalhadas e as formulações matemáticas de dois algoritmos e de duas técnicas, ou métodos, *ensemble* que serão utilizados neste estudo. O Método *ensemble*, também conhecido como aprendizado de conjunto, consiste em uma técnica que combina pontos fortes de um conjunto de modelos básicos mais simples com o objetivo de produzir um modelo preditivo forte (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). Um dos algoritmos mais utilizados no aprendizado de máquina e que será utilizado neste estudo para compor os métodos *ensemble Random Forest* (Subsubseção 3.4.1.2), *Bagging Classifier* (Subsubseção 3.4.1.3) e *XGBoost* (Subsubseção 3.4.1.4) é o de árvore de decisão, que será descrito de forma detalhada na Subsubseção 3.4.1.1. A Subsubseção 3.4.1.5 descreve os conceitos e definições envolvidos no algoritmo de Máquinas de Vetores de Suporte que também será testado para o problema de classificação multiclasse deste estudo.

3.4.1.1 Árvores de Decisão

Árvores de decisão são grafos acíclicos que podem ser usados para tomar decisões tanto para problemas de classificação quanto para problemas de regressão. Em cada nó de ramificação do grafo, uma variável preditora específica j é examinada. Em problemas de classificação, se o valor da variável estiver abaixo de um limite específico, o ramo esquerdo será seguido, caso contrário o ramo direito será seguido. À medida que o nó folha é alcançado, é tomada uma

decisão sobre a classe à qual o exemplo pertence, por este motivo, uma árvore de decisão pode ser aprendida a partir dos dados (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; BISHOP, 2006; ALPAYDIM, 2014).

Uma característica do algoritmo árvore de decisão é que ele possui uma abordagem gulosa (ou seja, sem retrocesso) em sua construção. Isso significa que as árvores são construídas de uma forma recursiva, de cima para baixo, dividindo e conquistando (HAN; KAMBER; PEI, 2012). Um algoritmo básico de árvore de decisão de classificação está apresentado na Figura 2.

Figura 2 – Algoritmo básico para induzir uma árvore de decisão a partir de instâncias de dados

Algoritmo: Gerar árvore de decisão a partir dos dados de treinamento

Entrada:

- Instâncias do conjunto de treinamento (D) e seus rótulos de classe associados;
- lista de variáveis preditoras, que descrevem as instâncias de dados
- Método de seleção de atributos

Método:

(1) criar um nó N;
(2) **se** as instâncias de dados em D forem todas da mesma classe C, **então**
(3) retorne N como um nó folha rotulado com a classe C;
(4) **senão** aplique o **método de seleção de variáveis preditoras** para **encontrar** o "melhor"critério de divisão;
(5) rotule o nó N com o critério de divisão
(6) **para cada** resultado i do critério de divisão
(7) particione as instâncias de dados e crie subárvores para cada partição
(8) remova a variável preditora de divisão da lista
(9) deixe Di ser o conjunto de instâncias de dados em D que satisfaçam o resultado de uma partição
(10) **se** Di ou a lista de variáveis é vazia, **então**
(11) anexe um nó folha rotulado com a classe majoritária em D ao nó N
(12) **caso contrário,** anexe o nó retornado por gerar árvore de decisão (Di, lista de variáveis) ao nó N
(13) fim do loop
(14) retorne N

Saída: uma árvore de decisão

Fonte: Adaptado de Han, Kamber e Pei (2012) pela autora.

Na Figura 2, o algoritmo é iniciado com três parâmetros: o conjunto de treinamento (D) com suas variáveis preditoras e variável *target* associada, a lista de variáveis preditoras que descrevem o conjunto de dados D e o método de seleção de atributos que corresponde ao procedimento para determinar qual o melhor critério de divisão das instâncias de dados em classes individuais (HAN; KAMBER; PEI, 2012).

A árvore começa com um único nó N onde D representa todo o conjunto de treinamento. Se todos os registros do conjunto D pertencerem à mesma classe C, então o nó N se torna uma folha e os registros de dados são rotulados com a classe C. Caso contrário, o algoritmo chama o método de seleção de variáveis para determinar o melhor critério de divisão. O método de seleção de variáveis é uma heurística para selecionar a variável preditora que melhor separa um determinado conjunto de dados D em classes individuais, por este motivo as medidas de seleção de variáveis também são conhecidas como regras de divisão (HAN; KAMBER; PEI, 2012).

Existem várias formulações do algoritmo de árvore de decisão, que diferem entre si

principalmente pela medida de seleção de variáveis que utilizam. Algumas formulações mais comuns são a ID3, a C4.5 e a CART, cujas medidas de seleção de variáveis são o ganho de informação, a taxa de ganho e o índice de Gini, respectivamente (HAN; KAMBER; PEI, 2012).

O algoritmo de Árvore de Decisão ID3 (*Iterative Dichotomizer 3*) utiliza um dos mais antigos critérios de seleção de variáveis, denominado ganho de informação. O ganho de informação seleciona como melhor variável de partição aquela que minimiza a entropia dos subconjuntos de dados (QUINLAN, 1986).

A entropia representa a impureza dos dados e, com ela, é possível medir a quantidade de informação necessária para classificar um registro em um determinado nó. A entropia é máxima quando em um nó, todas as classes têm igual prioridade na classificação de um registro e mínima quando existe apenas uma classe possível para se fazer a classificação (QUINLAN, 1986; HAN; KAMBER; PEI, 2012). A formulação matemática para o cálculo da entropia "E" em um conjunto de dados "D", representada por "$E(D)$" é dada por:

$$E(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{1}$$

Onde:

- D é uma partição dos dados, um conjunto de treinamento contendo variáveis preditoras e a variável *target* associada.
- m corresponde à quantidade de valores distintos da classe C
- $C_{i,D}$ são o conjunto de registros da classe C_i $\forall\{i = 1...m\}$ em D.
- $|D|$ e $|C_{i,D}|$ correspondem ao número de registros em D e $C_{i,D}$ respectivamente.
- p_i é a probabilidade não nula de uma instância de dados arbitrária em D pertencer à determinada classe i e pode ser formulada por:

$$p_i = \frac{|C_{i,D}|}{|D|} \tag{2}$$

Com base na medida de entropia, o ganho de informação resultante da adoção de uma partição baseada em uma determinada variável A com v valores distintos $\{a_1, a_2,..., a_v\}$ é dado por:

$$E_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * E(D_j) \tag{3}$$

O termo $\frac{|D_j|}{|D|}$ atua como o peso da j-ésima partição. $E(D)$ é a entropia do conjunto de dados D antes da divisão e $E_A(D)$ é a entropia do conjunto de dados D após o particionamento pelo atributo A. Quanto menor for a entropia resultante, maior será a pureza das partições. O ganho de informação GI é definido, portanto, como a diferença entre a entropia original $E(D)$ e a entropia após o particionamento $E(A)$:

$$GI(D, A) = E(D) - E_A(D) \tag{4}$$

Um dos problemas da medida de ganho de informação, utilizada no algoritmo ID3 é que ela tende a preferir a seleção de variáveis com um grande número de valores pois, neste cenário, cada partição tende a ser mais pura e, portanto, o ganho de informação obtido seria maximizado. No entanto, este tipo de divisão pode não ser a mais útil em um problema de classificação. Diante disso, foi desenvolvido o algoritmo C4.5, que é uma extensão do algoritmo ID3. Este algoritmo introduz o ganho de informação normalizado, conhecido como razão de ganho. Isso leva em consideração o número de valores possíveis das variáveis, permitindo uma comparação mais justa entre diferentes variáveis preditoras (QUINLAN, 1986; HAN; KAMBER; PEI, 2012).

O ganho de razão aplica uma normalização ao ganho de informação usando um valor de "informação de divisão" que pode ser formulada como:

$$SI_A(D) = -\sum_{j=1}^{v} \frac{|D_i|}{|D|} * log_2\left(\frac{|D_j|}{|D|}\right) \tag{5}$$

Este valor representa a informação potencial $SI_A(D)$ gerada pela divisão do conjunto de dados de treinamento D em v partições, correspondendo aos v resultados de um teste na variável A. Para cada resultado, são considerados o número de instâncias de dados que possuem esse resultado em relação ao número total de instâncias em D. A razão de ganho GR é então definida por:

$$GR(A) = \frac{GI(D, A)}{SI_A(D)} \tag{6}$$

A variável preditora com a maior razão de ganho é selecionada como variável de divisão. Dada a formulação apresentada na Equação (6), à medida que a informação da divisão se aproxima de 0, a razão torna-se instável. Por isso, é adicionada uma restrição, em que o ganho de informação do teste selecionado deve ser grande, pelo menos tão grande quanto o ganho médio de todos os testes realizados.

Embora o algoritmo C4.5 tenha superado algumas limitações do algoritmo ID3, em 1984 um grupo de estatísticos (L. Breiman, J. Friedman, R. Olshen e C. Stone) desenvolveram o algoritmo CART (*Classification and Regression Trees*) que, além de ser um algoritmo versátil que pode ser usado tanto para classificação quanto para regressão, é mais eficiente computacionalmente do que o ID3 e o C4.5, sobretudo para grandes conjuntos de dados (HAN; KAMBER; PEI, 2012).

A eficiência computacional do algoritmo CART deriva do uso do índice de Gini como medida de impureza. Essa medida quantifica a probabilidade de um elemento escolhido aleatoriamente, ser incorretamente classificado, considerando a distribuição das classes presentes na partição de dados, os resultados do índice variam de 0 a 1 (HAN; KAMBER; PEI, 2012).

O cálculo do índice de Gini inicia-se com o cálculo, para cada classe, ou categoria possível em um conjunto de dados, da probabilidade p_i de um elemento aleatório pertencer a essa classe. Isso pode ser feito dividindo o número de elementos pertencentes à classe pelo tamanho total do conjunto de dados:

$$p_i = \frac{|C_{i,D}|}{|D|} \tag{7}$$

Para cada classe em seguida é calculado o quadrado da probabilidade p_i denotada na Equação (7). Ao calcular o quadrado, amplifica-se a importância das classes mais dominantes. Isso ocorre porque, ao elevar ao quadrado, os valores próximos de 1 se aproximam ainda mais de 1, enquanto os valores próximos de 0 ficam ainda mais próximos de 0. Essa amplificação das classes dominantes no cálculo do índice de Gini significa que elas terão um peso maior na medida de impureza. Isso pode ser útil para realçar a importância de classes majoritárias em um conjunto de dados desbalanceado, por exemplo.

Em seguida, calcula-se o somatório dos quadrados de todas as probabilidades. Esta soma tem por objetivo calcular a medida geral de impureza para um determinado ponto de divisão em uma variável. Ao somar os quadrados, realiza-se a soma ponderada das impurezas individuais de cada classe. Quanto maior o valor resultante da soma dos quadrados, maior será a impureza total do ponto de divisão. Por fim, subtrai-se o valor obtido de 1, para obter o índice de Gini. A impureza medida pelo índice de Gini é uma medida de quão misturadas estão as classes em um determinado ponto de divisão. Quanto mais misturadas as classes, maior será a impureza. Porém, ao subtrair o resultado de 1, calcula-se a medida de pureza, ou seja, calcula-se o quão pura está a divisão em relação às classes. Um valor próximo de 1 indica uma divisão mais pura, onde a maioria dos elementos pertence a uma mesma classe. Por outro lado, um valor próximo de 0 indica uma divisão mais impura, onde os elementos estão mais igualmente distribuídos entre as classes. Portanto, a formulação geral o índice de Gini é dada por:

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \tag{8}$$

Por fim, a variável escolhida como o ponto de divisão é aquela que resulta no menor valor de índice após a divisão dos dados. A ideia é escolher a variável que melhor separa as classes nos subconjuntos.

Independente do algoritmo de árvore de decisão utilizado, o processo de selecionar a variável preditora, rotular o nó N com o critério de divisão e criar ramos a partir do nó N para cada um dos resultados do critério de divisão é recursivo e só para quando:

1. Todos os dados da partição D (representados no nó N) pertencem à mesma classe;
2. Não há variáveis restantes para novos particionamentos; ou
3. Não existem dados para determinado ramo.

Caso o critério de parada seja o 2 ou o 3, descritos acima, é utilizada votação por maioria para classificar os registros de dados com a classe majoritária em D.

Apesar de serem amplamente usados para resolver problemas de classificação e regressão, os algoritmos de árvore de decisão apresentam algumas limitações como tendência ao *overfitting* e sensibilidade aos dados de treinamento. Como forma de mitigar estas limitações, os métodos *ensemble* podem ser aplicados. Métodos *ensemble*, ou aprendizado de conjunto, é um paradigma de aprendizado que, ao invés de tentar aprender um modelo superpreciso, concentra-se em treinar um grande número de modelos de baixa precisão e, em seguida, combinar as previsões dadas pelos modelos fracos para obter-se um metamodelo de alta precisão (BURKOV, 2019).

O algoritmo de aprendizado fraco usado com mais frequência em métodos *ensemble* é a árvore de decisão. As árvores obtidas são rasas e não particularmente precisas, mas a ideia por trás do aprendizado de conjunto é que se as árvores não forem idênticas e cada árvore for pelo menos um pouco melhor do que a adivinhação aleatória, pode-se obter alta precisão combinando um grande número dessas árvores. Para obter a previsão para a entrada x, as precisões de cada modelo fraco são combinadas usando algum tipo de votação ponderada. A forma de ponderação dos votos depende do algoritmo (BURKOV, 2019).

3.4.1.2 *Random Forest Classifier*

O método *ensemble Random Forest* é uma extensão do algoritmo de árvore de decisão que combina a previsão de várias árvores individuais para obter resultados mais precisos e robustos. A ideia por trás deste algoritmo é contruir uma "floresta"de árvores de decisão, onde cada árvore é treinada em uma amostra aleatória, independente e com a mesma distribuição do conjunto de dados de treinamento. Essa amostragem geralmente é feita com reposição (*bootstrap*, já definido na Seção 3.3), o que significa que cada árvore pode ter amostras repetidas. A amostragem aleatória é realizada para introduzir variação e diversidade nas árvores (BREIMAN, 2001; BURKOV, 2019).

Além da amostragem de dados, o *Random Forest* seleciona um subconjunto aleatório de variáveis preditoras em cada divisão da árvore. Essa seleção aleatória garante que cada árvore

considere apenas um conjunto limitado de variáveis preditoras, reduzindo a correlação entre as árvores e evitando que algumas variáveis dominem a predição, reduzindo o risco de *overfitting* (BREIMAN, 2001; BURKOV, 2019).

Após o treinamento de modelos de classificação utilizando o algoritmo *Random Forest*, as previsões de todas as árvores individuais são combinadas e a classe mais frequente entre as árvores é escolhida como a previsão final (BURKOV, 2019).

3.4.1.3 *Bagging Classifier*

O algoritmo *Bagging Classifier* ou Classificador por *Bootstrap Aggregating* é uma técnica *ensemble* que combina múltiplos classificadores para obter uma previsão mais precisa. Neste algoritmo, o conjunto de dados original é amostrado várias vezes, com reposição, para criar várias amostras de treinamento. Cada amostra é denominada *bootstrap*. Em seguida, para problemas de classificação, em cada *bootstrap* um classificador é treinado utilizando um algoritmo de aprendizado de máquina, como árvores de decisão, SVM ou regressão logística. Ressalta-se que cada classificador é treinado em uma amostra diferente, o que introduz uma variação nos modelos. Em seguida, os resultados dos classificadores individuais são combinados para obter uma previsão final. Isso pode ser feito por votação (onde a classe mais frequente é selecionada) ou por média (BREIMAN, 1996).

No *Bagging Classifier* diferentes amostras de treinamento são criadas e utilizadas para treinar modelos diversificados, reduzindo o viés e a variância do modelo final e resultando em uma previsão mais robusta e geralmente mais precisa. Assim como no caso do *Random Forest*, o *Bagging Classifier* não altera diretamente a matemática inerente aos algoritmos de aprendizado de máquina utilizados, ele apenas utiliza técnicas de amostragem e combinação dos resultados para melhorar o desempenho e a estabilidade do modelo final. Portanto, a principal formulação matemática no *Bagging Classifier* também está relacionada à combinação de previsões dos modelos individiais. Uma das abordagens comuns é a votação por maioria, em que cada modelo possui um voto de mesmo peso e a classe prevista final para um exemplo é aquela que recebe mais votos dos modelos individuais. Outra abordagem comum para a escolha da classe final é a média das probabilidades de classe previstas pelos modelos individuais. Cada modelo fornecerá uma distribuição de probabilidade sobre as classes e a previsão será obtida calculando a média dessas probabilidades (BREIMAN, 1996).

No caso de empate, ou seja, quando duas ou mais classes recebem o mesmo número de votos, podem ser aplicadas diferentes estratégias para determinar a classe final: seleção aleatória de uma das classes empatadas, atribuição de prioridade para certas classes em caso de empate e, caso o empate ocorra a classe com prioridade será escolhida ou desempate adicional, isso pode envolver o uso de informações adicionais ou recursos como características específicas do exemplo de entrada ou a utilização de um modelo adicional para fazer a escolha final.

Ainda que diversos algoritmos possam ser utilizados no *Bagging Classifier*, a árvore de decisão é o mais comum, isso devido à sua capacidade de lidar com diferentes tipos de dados e

ao seu tempo de treinamento que costuma ser relativamente menor.

3.4.1.4 XGBoost

Uma alternativa ao classificador *Bagging*, apresentado na Subsubseção 3.4.1.3, é a técnica de *Boosting* que, ao invés de agregar previsões de diferentes modelos, constrói uma sequência de classificadores onde cada classificador é treinado para corrigir os erros do modelo anterior. Diferentemente do *Bagging classifier*, que pode utilizar qualquer algoritmo de classificação como base, na técnica de *Boosting* em geral os classificadores base são as árvores de decisão (CHEIN; GUESTRIN, 2016).

Uma técnica específica de *Boosting* que utiliza o gradiente da função de perda para ajustar os modelos subsequentes ao invés de ajustar apenas os resíduos, é denominada *Gradient Boosting*. Esta técnica utiliza a descida do gradiente para otimizar diretamente a função de perda, permitindo que o modelo se concentre em áreas onde há mais erros e melhore continuamente suas previsões. O *eXtreme Gradient Boosting* (*XGBoost*) é uma implementação específica do *Gradient Boosting* que se destaca pela eficiência e desempenho aprimorados. Desenvolvido em 2016, o *XGBoost* possui várias melhorias em relação ao *Gradient Boosting* tradicional como o uso de uma função de perda regularizada, manipulação eficiente de dados ausentes, seleção automática de recursos e paralelização para acelerar o treinamento (CHEIN; GUESTRIN, 2016).

Apesar de todas as vantagens que o *XGBoost* oferece, frente ao tradicional *Gradient Boosting*, o fator mais importante para o seu sucesso é a escalabilidade advinda de diversas otimizações algorítmicas feitas neste modelo que permite, inclusive, que ele seja utilizado em ambientes distribuídos, tornando-o eficiente mesmo em cenários de limitações de memória (CHEIN; GUESTRIN, 2016).

Para um conjunto de dados D com n exemplos e m variáveis preditoras, $D = (x_i, y_i)$ $(|D| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R})$, um modelo ensemble de árvore de decisão usa K funções aditivas para prever um *output*. Tais funções podem ser descritas como:

$$\hat{y_i} = \phi(x_i) = \sum_{i}^{K} f_k(x_i), \quad f_k \in F \tag{9}$$

onde $F = \{f(x) = W_{q(x)}\}$ $(q : \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T)$ é o espaço de algoritmos de árvores de regressão (CART), q é a estrurura de cada árvore, T é o número de folhas da árvore e f_k corresponde a uma estrutura de árvore independente q com pesos de folha w. Ao contrário das árvores de decisão, cada árvore de regressão contém uma pontuação contínua em cada uma das folhas e essa pontuação é representada por w_i.

Para aprender o conjunto de funções utilizadas no modelo, é necessário minimizar o objetivo regularizado dado por $L(\varphi)$.

$$L(\varphi) = \sum_{i=1} l(\hat{y_i}, y_i) + \sum_{k} \Omega(f_k) \tag{10}$$

onde $\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2$

Na Equação (10), l é uma função de perda diferenciável que mede a diferença entre a previsão $\hat{y}_i$ e o observado y_i. O termo Ω penaliza a complexidade do modelo (ou seja, as funções da árvore de regressão). O termo de regularização adicional ajuda a suavizar os pesos finais aprendidos para evitar o ajuste excessivo. O objetivo regularizado tenderá, então, a escolher um modelo que empregue funções simples e preditivas. Quando o parâmetro de regularização é definido como zero, o algoritmo volta a ser o tradicional Gradient Boosting

Se $\hat{y}_i{}^{(t)}$ é a previsão da i-ésima instância na t-ésima iteração, o fator f_t será adicionado para minimizar o seguinte objetivo:

$$L^t = \sum_{i=1}^{n} [l(y_i, \hat{y}_i{}^{(t-1)} + f_t(x_i))] + \Omega(f_t) \tag{11}$$

Isto significa que será adicionado na equação o f_t que melhora mais o modelo apresentado na Equação (10). Para otimizar rapidamente o objetivo, a aproximação de segunda ordem pode ser utilizada, cuja equação é dada por:

$$L^{(t)} \approx \sum_{i=1}^{n} [l(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{12}$$

onde, $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}$ e $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ são estatísticas de gradiente de primeira e segunda ordem sobre a função de perda. Removendo-se os termos constantes, obtém-se o seguinte objetivo simplificado:

$$L^t = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{13}$$

Definindo-se I_j como um conjunto de instâncias da folha j, a Equação (13) pode ser reescrita expandido-se Ω da seguinte forma:

$$\tilde{L}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \tag{14}$$

Para uma estrutura fixa de $q(x)$ o peso ótimo w_j^* da folha j é dado por:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{15}$$

o valor ótimo correspondente, é dado por:

$$\tilde{L}^{(}t)(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{16}$$

A Equação (16) pode ser usada como função de pontuação para medir a qualidade de uma estrutura de árvore q, semelhante ao que é feito com a medida de impureza nos algoritmos de árvores de decisão, exceto que esta é uma função derivada para uma gama mais ampla de funções objetivo.

3.4.1.5 Máquinas de Vetores de Suporte

O algoritmo SVM, assim como o algoritmo de árvore de decisão, também é muito utilizado em problemas de classificação, no entanto, a principal diferença entre eles está na abordagem utilizada para a construção do modelo. Enquanto as árvores de decisão criam estruturas hierárquicas de decisão com base nas características dos dados, o SVM busca encontrar um hiperplano de separação entre as classes.

Para o algoritmo SVM, cada vetor de variáveis preditoras é visto como um ponto em um espaço de alta dimensão. O objetivo do algoritmo consiste em encontrar um hiperplano neste espaço n-dimensional que separe os pontos de dados em suas classes potenciais. O número de dimensões do espaço depende da quantidade de variáveis preditoras presentes no conjunto de dados e o valor de cada variável preditora de um registro representa o valor de uma determinada coordenada para seu respectivo ponto de dados (MURPHY, 2012; CORTES; VAPNIK, 1995).

Um hiperplano é um subespaço com uma dimensão a menos do que o espaço dimensional sobre o qual os pontos de dados estão plotados e, a equação do hiperplano, também conhecido como fronteira de decisão, é dada por:

$$w^T x + w_0 = 0 \tag{17}$$

onde, w é um vetor de pesos que determina a orientação do hiperplano, x é o vetor de variáveis preditoras e w_0 é o termo de viés (*bias*) que controla o deslocamento do hiperplano em relação à origem (MURPHY, 2012; CORTES; VAPNIK, 1995).

Para que os pontos de dados sejam classificados, o algoritmo SVM utiliza uma função de decisão para determinar a classe de um exemplo com base no hiperplano de separação. Esta função pode ser descrita como:

$$f(x) = sign(w^T x + w_0) \tag{18}$$

Nesta equação, $f(x)$ representa a classe prevista para o exemplo de entrada x. A função de decisão calcula o valor do hiperplano ponderado pelo vetor de pesos w e adiciona o termo de viés w_0. O resultado é passado pela função $sign$, que retorna 1 se o valor for maior que zero e -1 caso contrário. Dessa forma, a função de decisão atribui uma classe positiva (1) ou negativa (-1) ao exemplo de entrada com base em sua posição em relação ao hiperplano de separação. A função de decisão também pode ser utilizada para obter o *score* de cada ponto de dados, que corresponde à sua distância até o hiperplano de separação e, portanto, quanto maior o valor absoluto do *score*, mais longe o ponto de dados estará da fronteira de separação (MURPHY, 2012; HAN; KAMBER; PEI, 2012; CORTES; VAPNIK, 1995).

Inicialmente, o algoritmo SVM foi concebido para problemas de classificação binária e, por isso, a função de decisão apresentada acima assume valores binários (1 ou -1) para classificar os pontos de dados. No entanto, é possível utilizar este algoritmo também para problemas de classificação multiclasse com algumas modificações na abordagem da função de decisão, como as estratégias *one-vs-all* ou *one-vs-one* (CORTES; VAPNIK, 1995).

Na estratégia *one-vs-all*, um classificador SVM é treinado para cada classe em relação às demais classes. Ou seja, para um problema com N classes, são treinados N classificadores SVM. Durante a fase de teste, o classificador que retorna o maior valor de função de decisão é selecionado como a classe prevista. Na estratégia *one-vs-one*, um classificador SVM é treinado para cada par de classes possíveis. Se houver N classes, serão treinados $N * (N - 1)/2$ classificadores SVM. Durante a fase de teste, cada classificador vota em sua classe correspondente e a classe que recebe o maior número de votos é selecionada como a classe prevista. Além dessas duas estratégias, também podem ser aplicadas as estratégias hierárquica e de conjunto (*ensemble*). Na primeira, as classes são organizadas em uma estrutura hierárquica e um classificador SVM é treinado para cada nível, onde cada classificador é responsável por discriminar entre um subconjunto de classes. Durante a fase de teste, a árvore hierárquica é percorrida para determinar a classe prevista. Na estratégia *ensemble*, vários classificadores SVM são treinados, cada um usando uma estratégia de classificação multiclasse diferente. Os resultados dos classificadores individuais são combinados de alguma forma (por exemplo, votação majoritária) para obter a classe prevista final (CORTES; VAPNIK, 1995).

Para o algoritmo SVM, as previsões dependem apenas de um subconjunto dos dados de treino conhecidos como vetores de suporte. A distância entre estes vetores e o hiperplano de separação é conhecida como margem, e é dada pela fórmula:

$$M = \frac{2}{||w||} \tag{19}$$

onde $||w||$ é a norma euclidiana do vetor de pesos w.

O objetivo do algoritmo SVM é encontrar um hiperplano de separação ótimo entre as classes de um problema de classificação. O hiperplano ótimo é aquele que maximiza a margem de separação entre as classes, ou seja, é o hiperplano que tem a maior distância para os pontos mais próximos de cada classe, chamados de vetores de suporte (MURPHY, 2012; HAN; KAMBER; PEI, 2012; CORTES; VAPNIK, 1995).

O problema de otimização do SVM pode ser formulado como um problema de programação quadrática que envolve minimizar uma função objetivo sujeita a restrições. A Função Objetivo busca minimizar a magnitude de pesos do hiperplano, que representa a margem de separação. As restrições impõem que todos os pontos de treinamento estejam corretamente classificados, ou fiquem do lado correto do hiperplano. Matematicamente, este problema de otimização pode ser escrito como:

$$\begin{aligned} &\text{Minimizar:} \quad \frac{1}{2}||w||^2 \\ &\text{Sujeito a:} \quad y_i * (w^T * x_i + w_0) \geq 1 \quad \forall i \end{aligned} \tag{20}$$

Nesta formulação, w é o vetor de pesos do hiperplano, w_0 é o viés (*bias*), x_i é um vetor de variáveis preditoras de um ponto de treinamento, y_i é o rótulo da classe do ponto de treinamento (+1 ou -1) e $||w||^2$ é a norma euclidiana ao quadrado do vetor de pesos. A solução desta formulação, fornece os valores ótimos de w e w_0 que geram o hiperplano de separação que maximiza a margem entre as classes (MURPHY, 2012; HAN; KAMBER; PEI, 2012; CORTES; VAPNIK, 1995).

Em muitos problemas reais, os dados não podem ser separados linearmente em um espaço de baixa dimensão. Para estes casos, o SVM utiliza uma abordagem chamada mapeamento não linear, que mapeia os dados de entrada para um espaço de maior dimensão, onde eles podem se tornar linearmente separáveis. Para isso, são utilizadas funções chamadas de kernels, que calculam o produto interno entre os vetores de características dos dados de entrada no espaço de maior dimensão, sem a necessidade de realizar explicitamente o mapeamento. Desta forma, o custo computacional não torna-se uma restrição para aplicar o SVM em espaços de alta dimensão.

Existem diferentes tipos de kernels comumente utilizados no SVM, dentre os quais pode-se citar o kernel linear, o polinomial e o gaussiano (também chamado de RBF - *Radial Basis Function*). Cada tipo de kernel possui suas próprias características. O kernel linear é o mais simples dos três e é utilizado quando os dados podem ser separados linearmente, sendo ideal para problemas em que existe uma relação linear entre as classes. O kernel polinomial permite mapear os dados para um espaço de dimensão maior utilizando funções polinomiais. Sendo assim, este kernel introduz a não linearidade ao elevar os produtos internos dos dados de entrada a uma potência. O grau do polinômio controla o nível de não linearidade do mapeamento e, quanto maior o grau, maior a capacidade de modelagem não linear do kernel polinomial. Por fim, o kernel gaussiano utiliza a função de base radial para mapear os dados em um espaço de maior

dimensão. Ele calcula a similaridade entre os vetores de características dos dados de entrada usando a função exponencial de distância euclidiana. Este é um kernel altamente flexível e capaz de modelar relações complexas entre os dados e é especialmente adequado para problemas em que os dados não possuem uma estrutura linear clara, ou quando a distribuição dos dados é desconhecida. No entanto, é importante ajustar o parâmetro de largura do kernel *sigma* para evitar o *overfitting* (BURKOV, 2019).

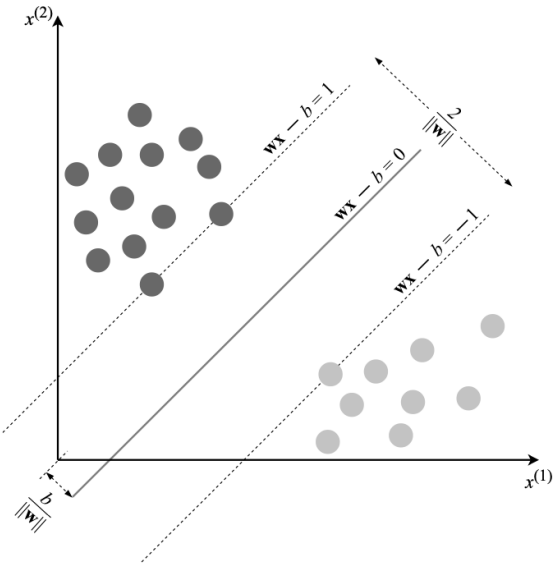
Além da utilização de kernels, quando os dados não são linearmente separáveis, a abordagem de margens suaves, ou *soft margins*, pode ser utilizada. Neste cenário, ao invés de buscar uma separação perfeita dos dados, o SVM permite uma certa quantidade de erros de classificação, permitindo que exemplos fiquem dentro da região de margem ou até mesmo do lado errado da fronteira de decisão. Isso é feito por meio de uma penalidade nos termos de erro na função objetivo do SVM. A escolha do parâmetro de penalidade (C) controla o *trade-off* entre a maximização da margem e a minimização dos erros de classificação. Valores mais altos de C indicam uma margem mais rígida e uma penalidade mais alta para erros, o que pode levar a um modelo com menor viés e maior variância. Valores mais baixos de C, por outro lado, permitem uma margem mais flexível e aceitam mais erros de classificação, resultando em um modelo com maior viés e menor variância (BURKOV, 2019).

$$\begin{aligned} &\text{Minimizar:} \quad \frac{1}{2}||w||^2 + C * \sum_i \xi_i \\ &\text{Sujeito a:} \quad y_i * (w^T * x_i + w_0) \geq 1 - \xi_i \forall i \end{aligned} \tag{21}$$

A Equação (21) apresenta a função objetivo do SVM, já demonstrada na Equação (20), agora com a inserção do parâmetro C e dos ξ_i que correspondem às variáveis de folga que permitem que alguns exemplos fiquem dentro da margem ou do lado errado do hiperplano (BURKOV, 2019).

A Figura 3 apresenta um exemplo gráfico de um modelo SVM para um vetor de recursos bidimensional.

Figura 3 – Exemplo de um modelo SVM para vetores de recursos bidimensionais



Fonte: (BURKOV, 2019).

3.4.2 Otimização de Hiperparâmetros

Hiperparâmetros são variáveis de configuração específicas para cada algoritmo que devem ser definidas antes do processo de treinamento do modelo. O desempenho que um classificador possui no processo de reconhecimento dos padrões em dados rotulados e, posteriormente, na produção de uma função que mapeie pontos de dados do espaço de recursos para o espaço de rótulos, é diretamente afetado por estes hiperparametros, por esse motivo a tarefa de encontrar os valores ideais para eles é um passo crucial para obter bons resultados para os classificadores (REIF; SHAFAIT; DENGEL, 2012).

Existem diferentes estratégias que podem ser aplicadas na otimização hiperparamétrica, dentre as quais estão o *Grid Search* e o *Randomize Search*. No *Grid Search*, cada hiperparâmetro é delimitado em torno de um intervalo de busca e o algoritmo realiza uma busca exaustiva de todas as combinações possíveis de hiperparâmetros até encontrar a combinação mais adequada. O problema deste método reside no crescimento exponencial de treinamentos necessários a medida que o número de hiperparâmetros cresce, o que torna-o oneroso principalmente para modelos com um número considerável de hiperparâmetros. O *Randomize Search*, por sua vez, não realiza uma busca exaustiva mas uma busca amostral de valores candidatos para cada hiperparâmetro (REIF; SHAFAIT; DENGEL, 2012).

O Quadro 2 apresenta, para o método *ensemble Random Forests*, os hiperparâmetros que são frequentemente discutidos na literatura em relação à otimização. Além destes, outros hiperparâmetros deste algoritmo com definições e valores *default* podem ser encontrados em (PEDREGOSA et al., 2011).

Quadro 2 – Hiperparâmetros frequentemente otimizados no método *ensemble Random Forest* segundo a literatura

Hiperparâmetro	Descrição	Valor *default*	Valores possíveis
n_estimators	Número de árvores que compõe o algoritmo de floresta aleatória	100	valores positivos inteiros
criterion	Função que mensura a qualidade da divisão do nó na arvore	Gini	gini, entropy, log_loss
max_depth	Profundidade máxima da árvore. Se nenhum valor for inputado, os nós serão expandidos até que todas as folhas sejam puras	None	valores positivos inteiros
min_samples_split	Número mínimo de amostras necessárias para se dividir um nó inteiro	2	valores positivos inteiros
min_samples_leaf	Número mínimo de amostras necessárias para estar em um nó para que ele seja caracterizado como um nó folha	1	valores positivos inteiros
max_features	É o número de variáveis preditoras a serem consideradas ao procurar a melhor divisão	sqrt	sqrt, log2, None

Fonte: Adaptado de Pedregosa et al. (2011) pela autora.

O Quadro 3 apresenta, para o método *ensemble Bagging Classifier*, os hiperparâmetros que são frequentemente discutidos na literatura em relação à otimização. Além destes, outros hiperparâmetros deste algoritmo com definições e valores *default* podem ser encontrados em (PEDREGOSA et al., 2011).

Quadro 3 – Hiperparâmetros frequentemente otimizados no método *ensemble Bagging Classifier* segundo a literatura

Hiperparâmetro	Descrição	Valor *default*	Valores possíveis
estimator	Classificador base que será utilizado	Classificador de árvores de Decisão	outros algoritmos
max_samples	Número de amostras a serem extraídas de X para treinar o classificador base	1	valores positivos inteiros
max_features	Número de variáveis preditoras que serão extraídas de X para treinar o classificador base	1	valores positivos inteiros
bootstrap	hiperparâmetro que indica se as amostras serão colhidas com reposição	True	True ou False
bootstrap_features	Hiperparâmetro que indica se as variáveis serão selecionadas com reposição	False	True ou False
max_features	É o número de variáveis preditoras a serem consideradas ao procurar a melhor divisão	sqrt	sqrt, log2, None

Fonte: Adaptado de Pedregosa et al. (2011) pela autora.

O Quadro 4 apresenta, para o método *ensemble XGBoost*, os hiperparâmetros que são frequentemente discutidos na literatura em relação à otimização. Além destes, outros hiperparâmetros deste algoritmo com definições e valores *default* podem ser encontrados em (DEVELOPERS, 2023).

Quadro 4 – Hiperparâmetros frequentemente otimizados no método *ensemble XGBoost* segundo a literatura

Hiperparâmetro	Descrição	Valor *default*	Valores possíveis
n_estimator	Quantidade de classificadores de árvore de decisão utilizados	100	valores positivos inteiros
learning_rate	Controla o tamanho do passo no qual o otimizador faz atualizações nos pesos. Um valor menor indica atualizações mais lentas, mas mais precisas.	1	valores positivos reais
max_features	Número de variáveis preditoras que serão extraídas de X para treinar o classificador base	1	valores positivos reais
max_depth	Profundidade máxima de cada árvore	6	valores positivos inteiros
subsample	Controla a fração de observações que devem ser usadas para cada árvore. Um valor menor resulta em modelos menores e menos complexos, ajudando a evitar o *overfitting*	1	valores positivos inteiros
colsample_bytree	É o número de variáveis preditoras a serem consideradas ao procurar a melhor divisão	1	valores positivos inteiros

Fonte: Adaptado de developers (2023) pela autora.

Por fim, o Quadro 5 apresenta, para o algoritmo SVM, os hiperparâmetros que são frequentemente discutidos na literatura em relação à otimização. Além destes, outros hiperparâmetros deste algoritmo com definições e valores *default* podem ser encontrados em (PEDREGOSA et al., 2011).

Quadro 5 – Hiperparâmetros frequentemente otimizados no algoritmo SVM segundo a literatura

Hiperparâmetro	Descrição	Valor *default*	Valores possíveis
C	Controla a tolerância aos erros de classificação, um C baixo torna a superfície de decisão suave enquanto um C alto visa classificar todos os exemplos de treinamento corretamente	1	valores positivos reais
kernel	Especifica o tipo de função de kernel a ser usada no modelo SVM. Este parâmetro irá determinar como os dados de entrada serão transformados para um espaço de características diferente onde é mais fácil separar as classes	RBF	linear, polinomial, RBF, sigmóide
gamma	Define quanta influência um único exemplo de treinamento tem. É utilizado apenas no SVM não linear	scale	valores positivos reais

Fonte: Adaptado de Pedregosa et al. (2011) pela autora.

3.5 Avaliação dos modelos e interpretação dos resultados

A avaliação de desempenho é uma etapa fundamental no processo de mineração de dados para entender o quão preciso é o modelo treinado. Em problemas de classificação multiclasse, é necessário utilizar medidas de avaliação adequadas que capturam a qualidade das predições de forma abrangente e precisa, pois a classificação multiclasse apresenta desafios únicos em comparação com problemas binários. As medidas de avaliação multiclasse permitem entender

o desempenho de um algoritmo em termos de exatidão, precisão, capacidade discriminativa e outras métricas relevantes que fornecem uma visão mais completa do desempenho do algoritmo, permitindo a comparação e seleção adequada de modelos treinados (HAN; KAMBER; PEI, 2012).

Algumas métricas comuns para a avaliação de desempenho dos modelos em problemas de classificação multiclasse são: acurácia, *precision*, *recall*, matriz de confusão e *Matthews Correlation Coeficient* (MCC), sendo esta última uma medida especialmente importante quando há um desequilíbrio entre as classes (HAN; KAMBER; PEI, 2012; BALDI et al., 2000).

Segundo Han, Kamber e Pei (2012), muitas medidas de avaliação de modelos de classificação são baseadas em quatro termos:

- Verdadeiros Positivos - referem-se às instâncias de dados positivas que foram rotuladas corretamente pelo classificador (VP).
- Verdadeiro Negativo - são a instâncias negativas que foram rotuladas corretamente pelo classificador (VN).
- Falso Positivo são as amostras negativas que foram classificadas erroneamente (FP).
- Falso Negativo são as amostras positivas que foram classificadas erroneamente (FN).

A matriz de confusão é uma ferramenta que permite analisar o desempenho do modelo utilizando estes quatro termos de forma resumida. Na matriz de confusão, a diagonal principal é composta por VP e VN, ou seja, pelas previsões corretas do modelo. Os termos FP e FN informam os erros do modelo. A quantidade total de instâncias reais positivas presentes no conjunto de dados é representada por P e as instâncias negativas por N, enquanto a quantidade total de instâncias classificadas como positivas pelo modelo são representadas por P' e as negativas por N'.

Quadro 6 – Matriz de confusão

		Previsão do modelo		
		Positivo	**Negativo**	**Total**
Dados reais	**Positivo**	VP	FN	P
	Negativo	FP	VN	N
	Total	P'	N'	P+N

Fonte: Adaptado de Burkov (2019) pela autora.

A acurácia representa a taxa de acerto global do modelo e mede a proporção de instâncias classificadas corretamente em relação ao total de instâncias no conjunto de dados (HAN; KAMBER; PEI, 2012; BURKOV, 2019). Matematicamente, a acurácia pode ser medida por:

$$acuracia\ (M) = \frac{VP + VN}{P + N} \tag{22}$$

Partindo do conceito de acurácia, deriva-se o conceito de taxa de erro ou taxa de classificação incorreta de um classificador M que é dada por $1 - M$ do modelo M (HAN; KAMBER; PEI, 2012; BURKOV, 2019). Matematicamente, a taxa de erro é dada por:

$$taxa\ de\ erro = \frac{FP + FN}{P + N} \tag{23}$$

No contexto de uma classificação multiclasse, a acurácia pode ser vista como uma importante métrica de visão geral do modelo. No entanto, ela pode ser enganosa caso seja considerada isoladamente, especialmente em casos onde há um desbalanceamento de classes, pois se houver uma classe dominante no conjunto de dados, o modelo poderá simplesmente classificar todas as instâncias segundo essa classe majoritária e ainda assim atingir uma acurácia satisfatória e, consequentemente, uma taxa de erro baixa. Portanto, em problemas de classificação multiclasse, outras métricas devem ser consideradas em conjunto com a acurácia para garantir uma avaliação correta do modelo (HAN; KAMBER; PEI, 2012; BURKOV, 2019; BALDI et al., 2000).

As medidas de sensibilidade e especificidade são amplamente utilizadas na classificação multiclasse, em conjunto com a acurácia, para avaliar o quão bem o classificador pode reconhecer as instâncias positivas e quão bem ele pode reconhecer as instâncias negativas, respectivamente (HAN; KAMBER; PEI, 2012; BURKOV, 2019). Matematicamente, as medidas de sensibilidade e especificidade são dadas por:

$$sensibilidade = \frac{VP}{P} \tag{24}$$

$$especificidade = \frac{VN}{N} \tag{25}$$

As medidas de *precision* e *recall* também são amplamente utilizadas na avaliação de modelos de classificação. *precision* mede a exatidão do modelo, ou seja, qual a porcentagem de instâncias rotuladas como positivas que realmente são positivas. Por outro lado, *recall* mede a completude, ou seja, qual a porcentagem de instâncias positivas que foram rotuladas como positivas (mesmo que sensibilidade ou taxa de verdadeiro positivo) (HAN; KAMBER; PEI, 2012; BURKOV, 2019).

$$precision = \frac{VP}{VP + FP} \tag{26}$$

$$recall = \frac{VP}{VP + FN} = \frac{VP}{P} \tag{27}$$

Conforme afirmam Han, Kamber e Pei (2012), uma medida de combinação das métricas de *precision* e *recall* que também é muito comum em avaliação de modelos de classificação

multiclasse é denominada $F1_{Score}$. Essa medida e corresponde à média harmônica das métricas de *precision* e *recall*, cuja formulação matemática é dada por:

$$F1_{score} = \frac{2 * precision * recall}{precision + recall} \tag{28}$$

Uma outra medida de avaliação de desempenho que leva em consideração a matriz de confusão completa é o *Matthews Correlation Coeficient* (MCC). O MCC é muito relevante em problemas de classificação multiclasse, pois fornece uma medida de desempenho do modelo levando em conta tanto as previsões corretas quanto as incorretas para todas as classes. Este coeficiente varia de -1 a 1, onde 1 indica uma previsão perfeita do modelo, 0 indica uma previsão aleatória e -1 indica uma previsão totalmente invertida. Assim, quanto mais próximo de 1 os valores do MCC estiverem, melhor é o desempenho do classificador. Uma das principais vantagens deste coeficiente é que ele leva em consideração o desbalanceamento entre classes, sendo capaz de fornecer uma medida mais confiável do desempenho do modelo, especialmente quando as classes possuem tamanhos distintos (BALDI et al., 2000). Matematicamente, o MCC é dado por:

$$MCC = \frac{(VP * VN) - (FP * FN)}{\sqrt{((VP + FP) * (VP + FN) * (VN + FP) * (VN + FN))}} \tag{29}$$

3.6 Resumo do capítulo

A mineração de dados é um processo, utilizado para descobrir padrões ou tendências úteis nos dados, composto por seis etapas:

- **Entendimento do problema e da(s) base(s) de dado(s)**: esta é a primeira e uma das mais importantes etapas no processo de mineração de dados, pois é nela que a abordagem mais adequada para resolver o problema será selecionada. Caso o objetivo seja mapear um *output* conhecido a partir de um ou mais *inputs* utilizando uma base de dados rotulada para treinamento dos algoritmos de aprendizado de máquina, a abordagem a ser utilizada é a supervisionada. Por outro lado, caso o objetivo seja encontrar padrões desconhecidos nos dados e a base de dados rotulada não esteja disponível, a estratégia a ser utilizada é a não supervisionada. Nesse contexto, é importante avaliar o problema bem como as bases de dados disponíveis para, então, escolher qual destas abordagens será utilizada. Além disso, análises descritivas devem ser realizadas nesta etapa para entender a distribuição dos dados e se existem problemas como alta correlação ou desbalanceamento entre variáveis. Os problemas encontrados durante a execução destas análises, deverão ser solucionados nas etapas seguintes, de pré-processamento e transformação dos dados.

- **Pré-processamento e transformação dos dados**: nas etapas de pré-processamento e transformação os dados são manipulados e convertidos em formas que rendam melhores

resultados para os algoritmos. Nesta etapa são realizadas conversões de dados para o formato numérico, remoção ou inferência de valores ausentes, cálculos entre variáveis, remoção de duplicatas, dentre outras operações que se façam pertinentes. A quantidade de tarefas realizadas na preparação dos dados faz com que mais da metade do trabalho esteja concentrado nesta etapa.

- **Modelagem dos dados**: esta é a etapa em que as técnicas de mineração de dados serão efetivamente aplicadas. Após entender o problema e preparar os dados, os algoritmos, que nada mais são do que um conjunto de instruções sequenciais e lógicas, serão aplicados para chegar ao objetivo pretendido. Estes algoritmos poderão ser aplicados de forma isolada ou em conjunto (o que é conhecido como método *ensemble*). Em geral, métodos *ensemble* são mais robustos e produzem resultados melhores se comparado à aplicação de algoritmos individuais, pois ao serem combinados, os algoritmos tendem a tornar-se mais fortes, podendo inclusive corrigir erros uns dos outros. Além disso, a combinação de diversos modelos nos métodos *ensemble* faz com que diferentes aspectos dos dados sejam capturados e isso ajuda a reduzir o *overfitting*. Por outro lado, os métodos de conjunto em geral são mais complexos, seus resultados são mais difíceis de serem interpretados, podem exigir mais recursos computacionais e ter um tempo de treinamento mais longo. Nesse contexto, a escolha entre a utilização de um algoritmo isolado ou de um método *ensemble* deve levar em consideração este *trade-off*.

- **Interpretação e avaliação dos resultados**: o objetivo desta etapa é estimar os resultados alcançados na etapa de modelagem para garantir que eles sejam válidos e confiáveis, e que os padrões encontrados nos dados sejam verdadeiros e não apenas anomalias ou idiossincrasias. Para isto, diversas medidas de avaliação são empregadas como acurácia, precisão, taxa de erro, dentre outras.

- **Implantação do modelo escolhido**: por fim, na etapa de implantação, o modelo que apresentar os melhores resultados na etapa anterior será colocado em uso real e fará as predições para as quais foi treinado, solucionando o problema inicial.

Cada uma destas etapas pode apresentar particularidades, dependendo do problema que se deseja resolver. Nesse contexto, o capítulo seguinte apresenta o estado da arte da utilização do processo de mineração de dados para resolver o problema da inferência do motivo de viagem dos passageiros, um atributo importante das matrizes OD.

4 Inferência do motivo de viagem: uma revisão da literatura

Este capítulo inicia-se com uma contextualização da importância de se conhecer o motivo de viagem das pessoas e quais são as principais abordagens utilizadas por outros pesquisadores para inferir este atributo de viagem quando ele não está disponível nos dados. Em seguida, dois subcapítulos são utilizados para explorar de forma mais detalhada os principais estudos que empregam metodologias não supervisionadas e supervisionadas, respectivamente, para inferir o motivo de viagem em dados secundários.

4.1 Contextualização da inferência do motivo de viagem na literatura

O motivo de uma viagem explica o objetivo pelo qual uma pessoa se move de um ponto para outro. Conforme descrito na Subseção 2.3.1, do ponto de vista utilitarista da racionalidade econômica, viajar é uma demanda derivada da atividade a ser desempenhada no destino e por isso um deslocamento ocorre, apenas, quando quando os benefícios obtidos no destino superam os custos que o indivíduo têm para chegar até lá (BANISTER, 2018; PEZOA et al., 2023).

Tradicionalmente, informações associadas ao comportamento de viagem, incluindo o motivo, são obtidas por metodologias ativas como as pesquisas de viagens, detalhadas na Subsubseção 2.2.1.1. Todavia, desvantagens como o alto custo de levantamento de dados e a baixa frequência de realização destas pesquisas fez crescer o interesse pela utilização de dados passivos, como os descritos na Subseção 2.2.2, no processo de planejamento de transportes. A disponibilidade destes dados, que são coletados usando sensores e tecnologias muitas vezes já disponíveis, permitiu o desenvolvimento de diversas abordagens de detecção de informações ausentes, incluindo o propósito da viagem (PEZOA et al., 2023).

Os estudos que usam dados passivos para detectar o propósito de viagem diferem, principalmente, pela abordagem metodológica usada, que pode ser supervisionada ou não supervisionada. Métodos de aprendizagem não supervisionada agrupam viagens com características semelhantes e, em seguida, atribuem um propósito de viagem por meio de um critério especializado ou por uma regra de decisão. A vantagem deste método é que a variável de propósito de viagem não é necessária para treinar o modelo. Entretanto, para que os resultados sejam satisfatórios, é preciso que haja um conhecimento especializado prévio para atribuir o propósito de viagem correto a cada um dos clusters. Isso pode se tornar especialmente difícil para viagens cujo propósito é uma atividade secundária como a de lazer, saúde, compras dentre outras. Metodologias baseadas no aprendizado supervisionado, por outro lado, utilizam dados históricos para treinar um modelo que

será aplicado posteriormente em outro conjunto de dados. Nessa abordagem, o modelo encontrará os padrões existentes nos dados sem a necessidade de conhecimento prévio de especialistas. Contudo, como já descrito, para que esta metodologia possa ser aplicada, dados de viagem rotulados com o propósito do deslocamento devem ser coletados previamente (PEZOA et al., 2023).

4.2 Estudos que utilizam metodologias não supervisionadas

Devillaine, Munizaga e Trepanier (2012) propuseram uma metodologia para uso de dados automatizados de cobrança de tarifa com cartões inteligentes para analisar os padrões de atividade dos usuários de transporte público das cidades de Gatineau (Canadá) e Santiago (Chile). Neste estudo, um conjunto de regras foram aplicadas para identificar o objetivo da viagem em uma das seguintes categorias: Trabalho, Escola, Residência ou Outros, sendo a última uma categoria que abrange diversas finalidades como, por exemplo, lazer, compras e saúde. Para o estudo de caso de Santiago, foram utilizados dados de transação da rede, e para Gatineau, além destes, foram utilizados também dados disponíveis sobre o uso da terra.

Goulet-Langlois, Koutsopoulos e Zhao (2016) utilizaram em seu estudo dados de transações de transporte público da cidade de Londres (Reino Unido) para identificar grupos de usuários com uma estrutura de sequência de atividades semelhante. Neste estudo, cada passageiro foi representado por uma sequência longitudinal de atividades realizadas. O algoritmo de clusterização *KMeans++* foi utilizado para agrupar os indivíduos, gerando 11 clusters com características de sequência de viagem distintas. Estes clusters foram, em seguida, agrupados em 4 padrões de atividades, sendo eles:

- grupos de dias úteis: cuja característica principal é a de padrões de viagem claros nos dias de semana e uma quantidade de viagem reduzida aos finais de semana;
- grupos com viagens ligadas ao domicílio: cuja característica principal é um longo período de tempo passado em áreas residenciais;
- grupos de padrões de atividade complexos: caracterizado por viagens em quase todos os dias do período de análise, dias úteis e fins-de-semana com uma sequência descontínua de viagens e um elevado número de atividades realizadas por dia, bem como maior diversidade nos locais de destino;
- cluster de padrões interrompidos: caracterizado por estruturas de viagem que variam significativamente entre os indivíduos em cada uma das semanas de análise.

Após agrupar os indivíduos a partir dos padrões de viagem, informações sociodemográficas disponíveis para uma pequena amostra de indivíduos foram utilizadas para analisar a associação entre estes padrões e atributos sociodemográficos. A análise revelou que existem relações

significativas entre os atributos sociodemográficos dos usuários e o padrão de atividades identificados pela metodologia do estudo.

Yang et al. (2019) utilizou dados de viagem provenientes de cartões inteligentes e Pontos de Interesse obtidos de dados de mídias sociais para um estudo de caso em Shenzhen (China), cujo objetivo foi examinar as variações temporais na mobilidade e identificar diferentes grupos de usuários com propósitos de viagem distintos. Os grupos de viajantes foram identificados com base em seus tempos e custos de viagem. Os Pontos de Interesse foram utilizados para caracterizar o entorno das estações de metrô em seis tipos de uso da terra. O propósito de viagem dos indivíduos, foi então inferido a partir do tempo de viagem vinculado ao uso da terra nas zonas de origem e destino usando uma medida de taxa de mudança de uso da terra.

Por fim, Faroqui e Mesbah (2021) propuseram um método de inferência do motivo de viagem a partir de atributos temporais disponíveis nos dados de cartões inteligentes e dados de uma pesquisa de mobilidade do sudeste de Queensland (Austrália). Neste estudo, cada passageiro da pesquisa de mobilidade foi representado como uma sequência de viagens e assim o algoritmo de clusterização aprendeu a relação entre os atributos temporais e o motivo de viagem. Em seguida, os clusters descobertos foram utilizados para inferir o propósito de viagem das transações com cartão inteligente, alocando cada passageiro no cluster mais próximo. Uma das descobertas deste estudo é que utilizar a sequência de viagem foi mais significativo no desempenho do modelo do que considerar variáveis de uso da terra, pois cada propósito de viagem foi inferido não apenas com base nos atributos de uma viagem mas sim nos atributos de outras viagens feitas pelo indivíduo durante o dia. Neste estudo foi utilizado o agrupamento hierárquico aglomerativo e os propósitos de viagem inferidos foram trabalho, residência, escola, lazer e compras.

4.3 Estudos que utilizam metodologias supervisionadas

Dentre os estudos que aplicam metodologias supervisionadas para a inferência do propósito de viagem, Alsger et al. (2018) utilizaram informações sobre transações de cartões inteligentes, dados de uso da terra, pesquisa de viagem domiciliar, o modelo estratégico de transporte do sudeste de Queensland (Austrália) e dados da especificação geral de alimentação de trânsito (GTFS) para treinar um modelo de escolha probabilística que classifica as viagens em cinco propósitos: trabalho, educação, compras, casa e recreação. Os resultados deste estudo demonstraram uma melhoria na inferência após a aplicação de atributos temporais junto de atributos espaciais, com aumento da acurácia. No entanto, diferentes propósitos de viagem apresentaram diferentes sensibilidades aos atributos espaciais e temporais aplicados, sendo as viagens de casa e trabalho aquelas que apresentaram os melhores resultados de inferência, com sensibilidade de 92% e 96%, respectivamente, já para viagens secundárias a sensibilidade foi significativamente menor, em torno de 46%.

Medina (2018) apresentou um método para identificar os padrões semanais temporais de atividades primárias realizadas por usuários de transporte público em Singapura. Neste estudo,

dados de uma pesquisa de viagens domiciliares foram usados para treinar dois modelos logit de escolha discreta: um para trabalhadores, e outro para estudantes. Utilizando o horário de início e duração de uma atividade, os modelos foram treinados para inferir o motivo da atividade em casa, trabalho, estudo ou outros. Na segunda etapa do estudo, atividades consistentes foram extraídas de dados de cartões inteligentes registrados durante uma semana. Para estes dados foram aplicados os modelos de escolha discreta estimados e, por fim, o algoritmo de agrupamento DBSCAN foi aplicado para reconhecer os comportamentos de viagem mais comuns no conjunto de dados. Os resultados do estudo demonstraram que trabalhadores de 5 dias úteis são o grupo mais representativo de usuários de transporte público.

Kim, Kim e Kim (2021) estimaram o motivo da viagem para usuários do transporte público da Coréia do Sul utilizando informações de cartões inteligentes, uma pesquisa de viagem, além de dados do censo populacional e de infraestrutura nacional de transporte. Neste estudo, os autores compararam o desempenho de quatro modelos: *Random Forest*, *Gradient Boosting Machine*, *Naive Bayes* e *Multinomial Logit*, todos eles tendo seus hiperparâmetros ajustados com base em um *Grid Search* com validação cruzada quádrupla, para classificar uma viagem em quatro categorias: trabalho, negócios, lazer e retorno para casa. Os resultados deste estudo demonstraram que entre os vários modelos utilizados o desempenho do *Random Forest* foi o melhor. A avaliação dos modelos foi baseada nas medidas de especificidade, sensibilidade e precisão balanceada. Os resultados do modelo *Random Forest* para cada um dos motivos de viagem estão apresentados na Tabela 1.

Tabela 1 – Resultados obtidos por Kim, Kim e Kim (2021)

	Random Forest		
	Especificidade	Sensibilidade	Precisão Balanceada
Trabalho	84%	98%	91%
Negócios	99%	35%	67%
Lazer	96%	79%	88%
Retorno à residência	99%	73%	86%

Fonte: Adaptado de Kim, Kim e Kim (2021) pela autora

A partir dos resultados, nota-se que a sensibilidade de viagens a negócios foi muito menor do que as outras finalidades e, segundo os autores, isto pode ter ocorrido pois as viagens por este propósito eram as menos frequentes no conjunto de dados, representando apenas 3%, e suas características temporais eram semelhantes às de deslocamento para lazer. Kim, Kim e Kim (2021) destacam ainda, que a amostragem estratificada foi tentada para resolver o desbalanceamento de classes, mas que não foi observada melhoria do desempenho geral do modelo.

As variáveis utilizadas no estudo de Kim, Kim e Kim (2021) foram a duração da atividade, a sequência da viagem em um dia, o horário de início da viagem, o horário de chegada ao destino, o tempo de viagem e variáveis espaciais de uso do solo no local de destino (dividido em residencial, comercial e outros). Outras variáveis foram utilizadas como a densidade populacional no destino, o número de trabalhadores e pessoas idosas no destino, o número de paradas de

ônibus e de metrô no destino e a idade média dos residentes do local de destino. A análise de importância das variáveis realizada neste estudo demonstrou ainda, que, atributos temporais foram os mais importantes na identificação do propósito de viagem. No geral, a variável de sequência da viagem, duração da atividade e horário de início do deslocamento foram as mais significativas para a classificação correta dos motivos de viagem. Além disso, o fato de o *Random Forest* ter superado os outros modelos de referência indica que os dados de pesquisas de viagem têm estruturas complexas, como a relação não linear entre recursos, resultados e interações entre recursos.

Aslam et al. (2021) propõem uma estrutura baseada em rede neural para classificar os propósitos de viagem dos passageiros de transporte público de Londres usando dados de cartão inteligente, dados de pesquisa de viagem e dados de Pontos de Interesse coletados usando a API do Twitter e a API de localização Foursquare. Os propósitos de viagem foram classificados em duas categorias, sendo elas primárias (trabalho e casa) e secundárias (entretenimento, refeição, compras, entrega/recolha de crianças e atividades de trabalho de meio período). A estrutura para predição dos motivos de viagem foi denominada ActivityNET e foi dividida em duas fases: a primeira fase do estudo foi concentrada na extração de atividades espaço-temporais de dados de cartões inteligentes e na combinação dessas atividades com pontos de interesse (POIs) usando um algoritmo de consolidação. A segunda fase do estudo utilizou recursos de entrada para prever os propósitos de viagem a partir de uma abordagem baseada em redes neurais artificiais. Devido ao processamento necessário de grandes conjuntos de dados, o framework PySpark foi utilizado para viabilizar a execução da metodologia proposta. Os resultados da pontuação F1-*score* obtidos demonstraram que as atividades primárias apresentaram valores significativamente maiores em comparação às atividades secundárias, sendo os valores para as atividades primárias de trabalho e residência superiores a 90%. Para as atividades secundárias, a melhor precisão de inferência é obtida para a entrega/recolha de crianças com uma pontuação F1-*score* de 89% e atividades de trabalho secundário com 85%. As atividades secundárias restantes apresentaram resultados semelhantes de 78% para entretenimento, 75% para refeição e 68% para compras. Destaca-se ainda que, neste estudo diversas técnicas de balanceamento de classes foram testadas, todas elas de sobreamostragem, e a técnica de sobreamostragem aleatória foi a que apresentou melhor desempenho.

Por fim, em um dos estudos mais recentes sobre a inferência do motivo de viagem, Pezoa et al. (2023) estimaram os propósitos de viagem por transporte público em Santiago (Chile) utilizando uma metodologia de aprendizado de máquina supevisionado, especificamente o algoritmo *XGBoost*. Nesse estudo, diversas fontes de dados foram utilizadas como a pesquisa origem e destino, dados de uso da terra vindos da Receita Federal, dados passivos associados às transações feitas com cartões inteligentes e dados de prioridade social da região gerados pelo Ministério do Desenvolvimento Social e da Família. O modelo proposto nesse estudo é uma concatenação de três modelos *XGBoost*. O primeiro deles foi treinado para discretizar as viagens em motivos primários e secundários. O segundo modelo foi utilizado para refinar as

atividades primárias em Trabalho, Residência e Escola. Por fim, o terceiro modelo foi utilizado para caracterizar as atividades secundárias em saúde, lazer e outros. As variáveis utilizadas nesse estudo foram o número de estágios de uma viagem, o tempo de viagem, a duração da atividade, o horário de início da viagem, o horário de desembarque, a distância entre o ponto de embarque e desembarque em metros, duas variáveis booleanas que indicam se a viagem foi a primeira ou última viagem do dia, o índice de prioridade social associado às zonas de embarque e desembarque da viagem, os percentuais de uso da terra do tipo Y dentro de um raio de 500m dos pontos de embarque e desembarque, a área de uso da terra do tipo Y dentro de um raio de 500m dos pontos de embarque e desembarque e o número de pontos de interesse do tipo X dentro de um raio de 500m dos pontos de embarque e desembarque. A pontuação F1-*score* para o primeiro estágio da modelagem, treinado para discretizar as viagens em motivos primários e secundários, foi de 88%, o segundo modelo utilizado para refinar as atividades primárias em Trabalho, Residência e Escola apresentou resultado de 86% e, finalmente, o terceiro modelo utilizado para caracterizar as atividades secundárias em saúde, lazer e outros apresentou pontuação F1-*score* de 72%. Os resultados por motivo de viagem não foram apresentados.

4.4 Resumo do capítulo

A inferência do motivo de viagem em bases de dados secundários utilizadas para estimar matrizes OD vem sendo amplamente discutida na literatura desde meados de 2012 até os dias atuais, com estudos recentemente publicados.

A principal diferença entre os estudos encontrados na literatura é a abordagem metodológica utilizada no processo de inferência. Alguns estudos utilizaram uma abordagem não supervisionada para agrupar viagens com características semelhantes e, em seguida, atribuir o propósito a partir de um critério especializado ou por uma regra de decisão. Quatro estudos com esta abordagem foram analisados, todos eles utilizaram dados de cartões inteligentes e em alguns casos, outras fontes de dados complementares foram utilizadas como Pesquisas de Mobilidade e localização de Pontos de Interesse. Os estudos selecionados desta categoria são de 2012, 2016, 2019 e 2021 e as aplicações foram para as cidades de Gatineau (Canadá) e Santiago (Chile); Shenzen (China); Londres (Reino Unido); e Queensland (Austrália), respectivamente.

Uma outra categoria de estudos utilizou dados históricos para treinar um modelo com o objetivo de encontrar padrões nos dados, sem a necessidade de um conhecimento prévio de especialistas, e em seguida inferir qual o propósito das viagens. Quatro estudos com esta abordagem foram analisados e a principal característica em comum entre eles é que todos utilizaram pesquisas domiciliares para treinamento dos algoritmos, além de outras fontes complementares de dados como a localização de Pontos de Interesse, dados de censo populacional e dados socioeconômicos. Os estudos selecionados desta categoria são de 2018, 2021 e 2023. Já as aplicações destes estudos foram, respectivamente, em Queensland (Austrália), Coreia do Sul, Londres e Santiago (Chile)

A quantidade de estudos científicos com o objetivo de inferir o motivo de viagem dos

passageiros no transporte público demonstra a importância deste problema de pesquisa. Para contribuir com a literatura existente, o capítulo seguinte apresenta a metodologia utilizada neste trabalho para inferir o motivo de viagem dos passageiros em uma fonte de dados secundários, notadamente a base de dados de cartões inteligentes em um estudo de caso para a RMBH.

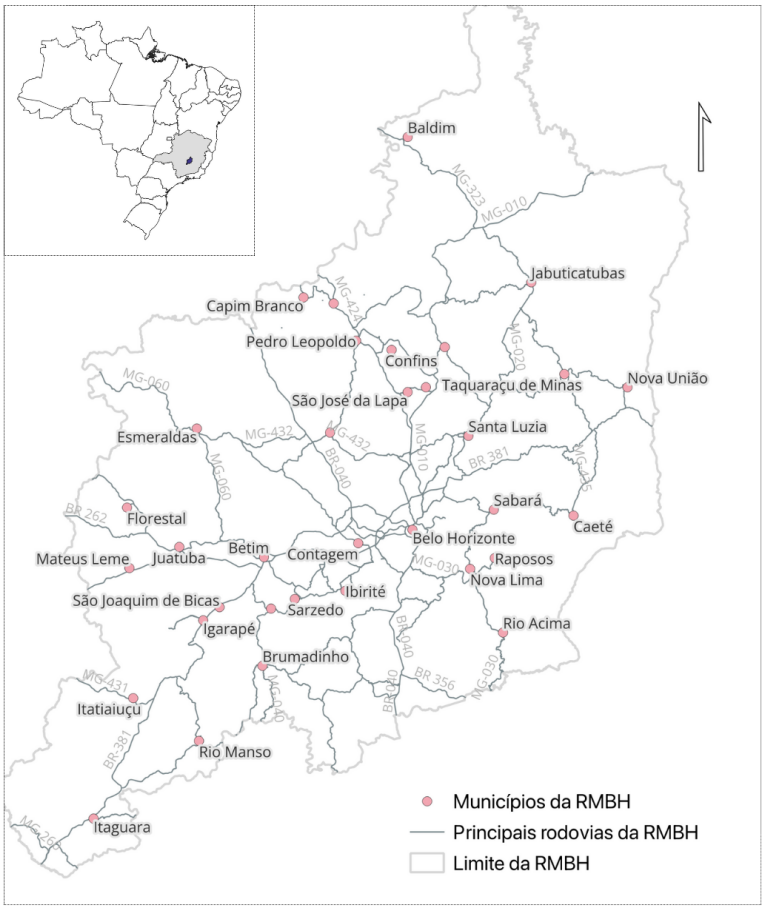
5 Metodologia

Neste capítulo, é delineada a base metodológica que orienta a condução desta pesquisa. Inicialmente, serão apresentados detalhes fundamentais sobre a área de estudo e sobre o fluxo de trabalho adotado, fornecendo uma visão abrangente das etapas sequenciais envolvidas na execução do estudo. Em seguida, serão discutidas as fontes de dados utilizadas e as etapas de pré-processamento, limpeza, transformação e modelagem dos dados para se chegar ao objetivo pretendido. Este capítulo serve como um guia detalhado do método empregado, oferecendo uma compreensão clara dos procedimentos e abordagens adotadas na condução desta pesquisa.

5.1 Descrição da área de estudo e do fluxo de trabalho

O objetivo geral deste estudo consiste na aplicação de abordagens emergentes de mineração de dados para inferir informações ausentes em fontes de *Big Data* utilizadas no planejamento de transportes. Especificamente, pretende-se rotular os registros de viagem de uma base de dados de cartões inteligentes com o motivo de viagem do passageiro em um estudo de caso para a Região Metropolitana de Belo Horizonte (RMBH), um aglomerado de 34 municípios do estado de Minas Gerais (Figura 4).

Figura 4 – Representação gráfica da RMBH com seus municípios e principais rodovias



Fonte: Autoria própria

Inicialmente, duas fontes de dados são coletadas: Os dados da pesquisa domiciliar realizada em 2012 na RMBH (OD_2012) e os dados de estabelecimentos do Cadastro Nacional de Pessoa Jurídica da Receita Federal do Brasil (POIs). Em seguida, ambas bases de dados passam pelas etapas de pré-processamento, limpeza e transformação utilizando técnicas cujos conceitos estão detalhadamente descritos nas seções Seção 3.2 e Seção 3.3.

Na etapa de modelagem dos dados, três diferentes métodos de inferência são implementados. Tais métodos diferem entre si sobretudo pelo conjunto de variáveis preditoras utilizadas na etapa de treinamento, pela quantidade de classes disponíveis na variável *target* e pela quantidade de modelos treinados.

No Método 1, um modelo único de classificação multiclasse é treinado para rotular os registro de viagem em uma das sete classes: "Residência", "Trabalho", "Escola", "Saúde", "Lazer", "Negócios"e "Outros". Neste método, as variáveis preditoras são da base de dados OD_2012. Para selecionar a combinação de algoritmo e técnica de balanceamento de classes mais adequado ao contexto do problema de pesquisa, quatro algoritmos conhecidos, *Random Forest*, *Bagging Classifier*, SVM e *XGBoost*, e três técnicas de balanceamento de classes, *Random Oversampling*, SMOTE e ADASYN, são testados. Uma busca em grade é realizada para encontrar

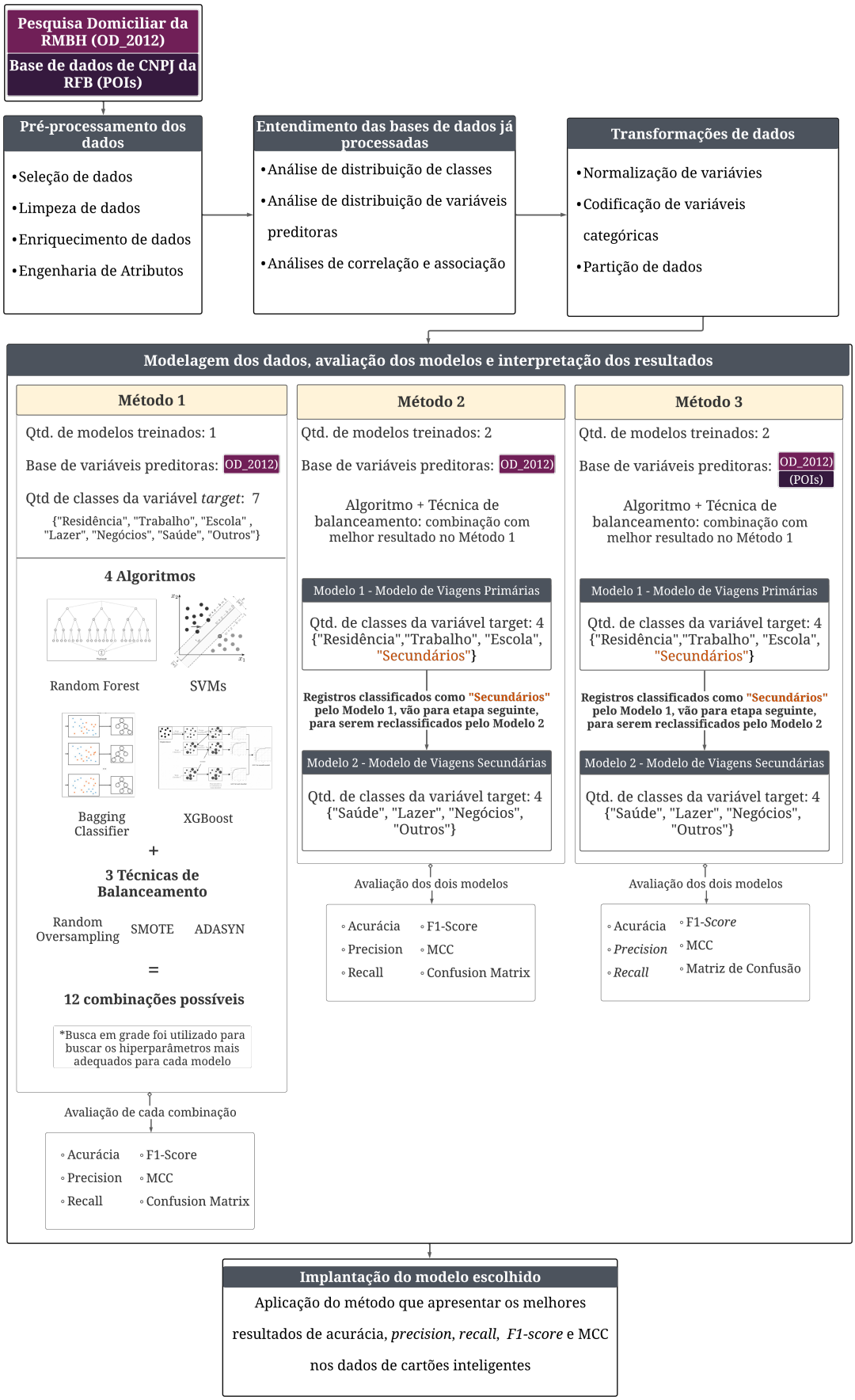
os hiperparâmetros mais adequados para cada algoritmo e o desempenho de cada combinação de algorimo e técnica de balanceamento é mensurado a partir das métricas de acurácia, *precision*, *recall*, *F1-score*, MCC e matriz de confusão. Ao final, a combinação que apresentar os melhores resultados para as métricas de avaliação é selecionado para aplicação nos métodos 2 e 3.

O Método 2 é composto por duas etapas sequenciais. Na primeira, um modelo é treinado para classificar os registros de viagem em uma das 4 classes: "Residência", "Trabalho", "Escola" e "Secundário". Na segunda etapa, outro modelo é treinado com o objetivo de reclassificar/detalhar somente os registros classificados pelo Modelo 1 como "Secundário" em quatro novas classes: "Saúde", "Lazer", "Negócios" ou "Outros". As variáveis preditoras utilizadas no Método 2 também são da OD_2012 e a combinação de algoritmo e técnica de balanceamento utilizada é a que têm o melhor resultado nas avaliações feitas ao final da implementação do Método 1.

Por fim, o Método 3 avalia os possíveis ganhos de se incorporar variáveis espaciais de Pontos de Interesse (POIs), como preditoras no treinamento dos dois modelos do Método 2, sendo esta a única diferença entre os Métodos 3 e 2.

Os dois datasets, OD_2012 e POIs, que são utilizados para treinamento dos modelos neste trabalho, podem ser acessados no repositório do *GitHub* através do *link*: <https://github.com/MirianGreiner/Trip_Purpose_Inference.git> e o fluxograma da metodologia utilizada neste estudo está apresentado na Figura 5.

Figura 5 – Fluxo de trabalho



Fonte: Autoria própria

5.2 Descrição das fontes de dados

5.2.1 Pesquisa Domiciliar de Viagens (OD_2012)

A Pesquisa de viagem domiciliar é realizada na RMBH a cada dez anos. O objetivo desta pesquisa é capturar os padrões de comportamento de viagem da população bem como as características socioeconômicas dos indivíduos.

A primeira pesquisa domiciliar foi realizada na RMBH em 1972 e a edição mais recente deste tipo de coleta é de 2012, os resultados e relatórios completos deste levantamento podem ser acessados em RMBH (2023). Nesta última edição, o levantamento foi feito em domicílios situados em todos os 34 municípios que compõem a Região Metropolitana de Belo Horizonte, com uma amostra de 59.344 indivíduos, 27.910 domicílios e 140.540 viagens, das quais 26.325 (aproximadamente 20%) foram feitas por transporte público coletivo, por ônibus ou metrô. As entrevistas ocorreram presencialmente entre os meses de agosto e novembro de 2012 e as perguntas se referiam sempre aos trajetos feitos no dia anterior pelos entrevistados e pelos demais moradores do domicílio. Além disso, a referência das entrevistas sempre deveria ser um dia útil da semana, exceto para dias com possibilidade de recesso. Por isso, a pesquisa domiciliar foi realizada de terça-feira a sábado, excetuando feriados e recessos, coletando dados de deslocamentos ocorridos entre segunda-feira e sexta-feira.

Embora já tenham se passado 10 anos desde a última edição, a Pesquisa domiciliar realizada em 2012 ainda é a mais recente feita por métodos diretos de levantamento de dados. Em 2019 e 2021 dados de telefonia móvel (do tipo CDR) foram utilizados para quantificar a movimentação de passageiros entre pares OD na RMBH e assim estimar as matrizes OD. Destaca-se que, por se basearem em dados de telefonia móvel, as matrizes OD de 2019 e 2021 também não possuem o atributo de propósito de viagem coletado, e sim inferido, estando eles restritos à "Domicílio", "Trabalho/Estudo"e "Outros". Por este motivo, para garantir que a base de "dados de verdade"esteja corretamente rotulada e possa ser utilizada para avaliar o desempenho dos modelos treinados, optou-se por utilizar uma pesquisa domiciliar mais antiga para treinar os modelos de classificação neste estudo.

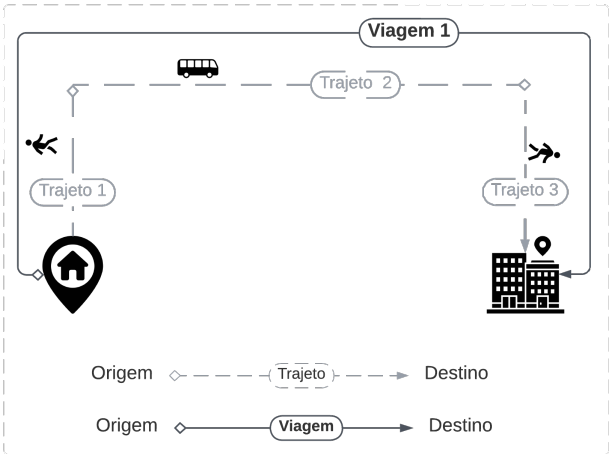
Em relação às informações sobre a mobilidade urbana, os dados coletados na OD_2012, foram divididos em duas tabelas, sendo elas:

- **Tabela de trajetos**: tabela que contém a informação de todos os trajetos de uma viagem. Um trajeto de viagem corresponde a um deslocamento que parte de um ponto de origem até um ponto de destino temporário, onde se realizará uma troca de modo de transporte (Figura 6).

- **Tabela de viagens**: tabela que contém a informação completa de uma viagem. Entende-se por viagem um movimento gerado por uma pessoa em função de um motivo específico qualquer, com a utilização de um ou mais modos de transporte, e que pode ser composta por um ou mais trechos. Portanto, o registro da tabela de viagens contempla apenas uma

origem, um destino, um motivo de viagem e um modo de transporte, definido como o modo de transporte principal ao considerar todos os trajetos feitos dentro da viagem analisada.

Figura 6 – Ilustração de uma viagem composta por três trechos



Fonte: Autoria própria

5.2.2 Estabelecimentos - Cadastro Nacional da Pessoa Jurídica (POIs)

O Cadastro Nacional da Pessoa Jurídica é um banco de dados abertos gerenciado pela Secretaria Especial da Receita Federal do Brasil (RFB) que contém informações cadastrais de todas as empresas registradas no país. Os dados completos podem ser acessados em RFB (2023). A periodicidade de atualização dos dados é mensal e, para este estudo, foram filtradas apenas as empresas que estavam ativas no ano de 2012, ano de referência da pesquisa domiciliar (OD_2012). A base de dados do CNPJ é composta por quatro tabelas principais:

- Empresas: tabela que contém os dados cadastrais da matriz da empresa;
- Estabelecimentos: dados analíticos das empresas por unidade/estabelecimento como telefones, endereço e atividade principal desenvolvida pela empresa;
- Sócios: tabela que contém os dados cadastrais dos sócios das empresas, como nome, faixa etária e classificação;
- Simples: tabela que traz informações adicionais para estabelecimentos classificados como micro e pequenas empresas;

e cinco tabelas complementares, no formato código-descrição:

- Países
- Municípios
- Qualificação dos sócios

- Natureza Jurídica
- CNAE

Neste estudo, a base de dados do CNPJ será utilizada para quantificar os Pontos de Interesse (POIs) por tipo de atividade econômica situados no local de destino de cada viagem da OD_2012. Um POI é um ponto específico no mapa, caracterizado por sua latitude e longitude, que é acessado pelas pessoas em função de um interesse potencial, podendo ser uma atração turística, um hospital, uma escola, ou um estabelecimento de qualquer outra categoria.

Dado o objetivo de utilização da base de dados do CNPJ, neste estudo apenas a tabela de estabelecimentos será utilizada, em conjunto com as tabelas complementares de países, municípios e CNAE. Juntas, estas tabelas contém todas as informações necessárias para quantificar os POIs por tipo de atividade disponíveis no local de destino das viagens.

5.3 Pré-Processamento dos dados

A etapa do pré-processamento dos dados é responsável por organizar e preparar os dados para a aplicação dos algoritmos de aprendizado de máquina e é composta por várias sub-etapas que variam de acordo com o problema de estudo.

Todas as sub-etapas do pré-processamento de dados realizadas neste estudo estão descritas a seguir, e foram implementadas em linguagem Python.

5.3.1 Seleção de dados

A primeira sub-etapa do pré-processamento de dados realizada neste estudo foi a seleção das variáveis relevantes e informativas dos dois conjuntos de dados que serão utilizados no processo de modelagem. O Quadro 7 apresenta, de forma resumida, todos os atributos disponíveis na tabela de viagens da pesquisa domiciliar (OD_2012).

Quadro 7 – Variáveis disponíveis no *dataset* OD_2012

Descrição do atributo	**Código**
Número sequencial atribuído a cada um dos 140.540 registros	ID_Entrevista
Identificação do domicílio pesquisado	ID_Domicilio
Identificação do indivíduo entrevistado dentro de cada domicílio	ID_Individuo
Identificação sequencial das viagens de cada um dos indivíduos pesquisados	ID_Viagem
Hora de início da viagem	Horario_Inicio
Hora de fim da viagem	Horario_Fim
Diferença entre o horário inicial e final da viagem	Tempo_Deslocamento
Faixa horária de realização do deslocamento	Faixa_Horaria
Área Homogênea de origem do deslocamento	AH_Origem
Área Homogênea de destino do deslocamento	AH_Destino
Campo de origem do deslocamento	Campo_Origem
Campo de destino do deslocamento	Campo_Destino
Unidade de Macro Mobilidade de origem do deslocamento	UMM_Origem
Unidade de Macro Mobilidade de destino do deslocamento	UMM_Destino
Município de origem do deslocamento	Municipio_Origem
Município de destino do deslocamento	Municipio_Destino
Eixo de concentricidade de origem do deslocamento	EC_Origem
Eixo de concentricidade de destino do deslocamento	EC_Destino
Modo de transporte principal, ou seja, o modo mais utilizado considerando todos os trajetos feitos dentro da viagem	Modo_Principal
Modos Agrupados, ou seja, o agrupamento dos diferentes modos de transporte em coletivo, individual, não motorizado ou outros	Modo_Agrupado
Descrição do meio de transporte utilizado, como, por exemplo, ônibus, metrô, etc.	Meio_Transporte
Renda do indivíduo	Renda
Escolaridade do indivíduo	Escolaridade
Idade do indivíduo	Idade
Sexo do indivíduo	Sexo
Fator utilizado para expandir a amostra coletada	Fator_Expansao
Atividade que estava sendo realizada pelo passageiro quando o deslocamento iniciou	Motivo_Origem
Atividade que será realizada pelo passageiro ao chegar no destino de seu deslocamento	Motivo_Destino

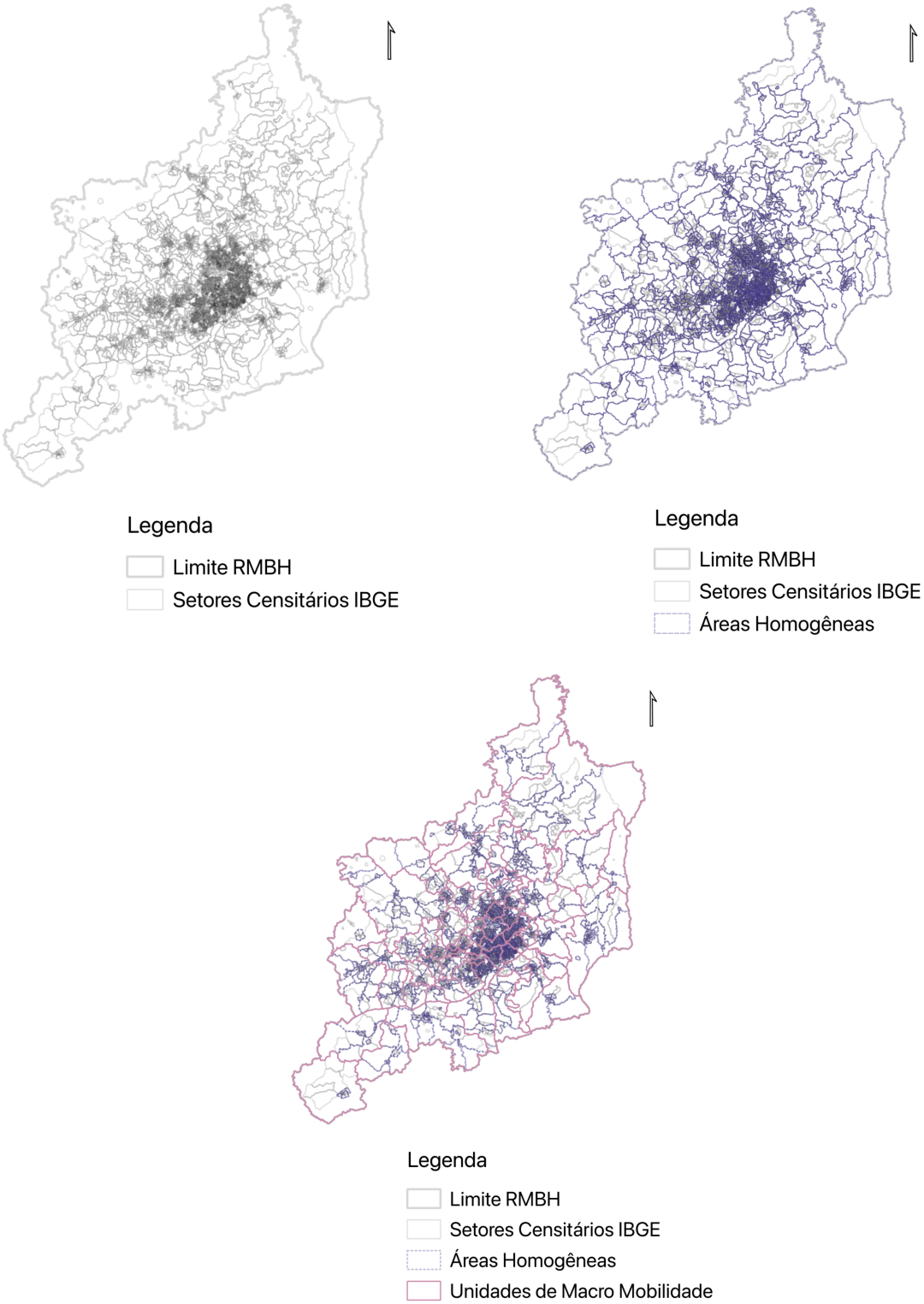
Fonte: Autoria própria

Algumas variáveis apresentadas na Quadro 7 trazem informações semelhantes com variação apenas na escala de representação dos dados, como é o caso das variáveis: Área Homogênea, Campo, Unidade de Macro-Mobilidade, Eixo de Concentricidade e Município.

As Áreas Homogêneas (AHs) são unidades de coleta de dados formadas por agregação de Setores Censitários do IBGE. Estas áreas possuem homogeneidade de características físicas (como delimitações de bacias hidrográficas, topografia e declividade), urbanísticas (como tipologia do uso do solo para residências, comércio, indústria, lazer), de conformação e hierarquização do sistema rodo-ferroviário, de padrão socioeconômico, de histórico de ocupação e de conformação de redes de transporte coletivo.

Os Campos são considerados o primeiro nível de agregação das AHs e podem ser vistos como territórios que possuem significados e identidades próprias. As Unidades de Macro Mobilidade (UMMs) também consistem em agregações diretas de AHs e foram criadas para expressar os principais macros valores e indicadores de mobilidade gerados a partir de dados coletados na Pesquisa OD 2012. A Figura 7 demonstra, graficamente, os setores censitários, a agregação de setores censitários em áreas homogêneas e a agregação das áreas homogêneas em unidades de macro mobilidade.

Figura 7 – Representação gráfica dos Setores Censitários, Áreas Homogêneas e Unidades de Macro Mobilidade



Fonte: Autoria própria

Os eixos de concentricidade correspondem à articulação das Unidades de Macro Mobilidade com os principais eixos rodoviários que atravessam a RMBH e a localização destas áreas em relação ao centro da metrópole. Por fim, os municípios constituem as unidades autônomas de

menor hierarquia dentro da organização político-administrativa do Brasil.

Dada a definição de cada uma das variáveis descritas acima, optou-se por seguir adiante apenas com as variáveis de AH de origem (AH_Origem) e de destino (AH_Destino) do deslocamento por serem a menor unidade espacial que possibilite analisar o padrão de deslocamento dos indivíduos. Ressalta-se que estas variáveis serão utilizadas apenas como chave para a junção dos dados de viagens com os dados de Pontos de Interesse e, portanto, não serão incluídas no conjunto de variáveis preditoras.

As variáveis que tratam sobre características socioeconômicas dos indivíduos, como "Renda", "Escolaridade", "Idade"e "Sexo", também foram removidas, pois, em geral, matrizes OD feitas a partir de dados secundários não possuem atributos socioeconômicos dos passageiros. Nesse contexto, considerando que este estudo propõe um método para inferir o motivo de viagem dos passageiros em matrizes OD feitas com dados de cartões inteligentes, e dada a premissa de que o *dataset* de previsão precisa ter a mesma estrutura do *dataset* de treinamento do modelo, estas variáveis foram desconsideradas.

As variáveis “Modo_Principal”, “Modo_Agrupado” e “Meio_Transporte” também não irão compor o modelo e servirão apenas para filtragem de dados, pois como os dados no *dataset* de previsão contemplam apenas as viagens de transporte público coletivo, foram utilizadas apenas as viagens feitas por este modo de transporte, incluindo viagens feitas por ônibus coletivo e metrô, na base de dados da OD_2012, para treinamento do modelo.

Por fim, a variável de "ID_Entrevista"também foi desconsiderada por não conter nenhuma informação significativa para o processo de modelagem. Finalizada esta etapa de seleção das variáveis, restaram na tabela de dados de viagem os atributos descritos na Quadro 8.

Quadro 8 – Variáveis do *dataset* OD_2012 relevantes para serem mantidas no conjunto de dados

Descrição do atributo	**Código**
Identificação do domicílio pesquisado	ID_Domicilio
Identificação do indivíduo entrevistado dentro de cada domicílio	ID_Individuo
Identificação sequencial das viagens de cada um dos indivíduos pesquisados	ID_Viagem
Hora de início da viagem	Horario_Inicio
Hora de fim da viagem	Horario_Fim
Diferença entre o horário inicial e final da viagem	Tempo_Deslocamento
Área Homogênea de origem do deslocamento	AH_Origem
Área Homogênea de destino do deslocamento	AH_Destino
Modos Agrupados, ou seja, o agrupamento dos diferentes modos de transporte em coletivo, individual, não motorizado ou outros	Modo_Agrupado
Descrição do meio de transporte utilizado, como por exemplo ônibus, metrô, etc.	Meio_Transporte
Atividade que estava sendo realizada pelo passageiro quando o deslocamento iniciou	Motivo_Origem
Atividade que será realizada pelo passageiro ao chegar no destino de seu deslocamento	Motivo_Destino

Fonte: Autoria própria

Em relação ao *dataset* POIs, a Quadro 9 apresenta, de forma resumida, todos os atributos disponíveis.

Quadro 9 – Variáveis disponível no *dataset* POIs

Descrição do atributo	**Código**
Número base de inscrição no CNPJ (oito primeiros dígitos do CNPJ)	CNPJ_Basico
Número de inscrição do estabelecimento no CNPJ (nono até o décimo segundo dígito do CNPJ)	CNPJ_Ordem
Dígito verificador do número de inscrição no cnpj (dois últimos dígitos do CNPJ)	CNPJ_Digito
Identificação do tipo de estabelecimento (matriz ou filial)	Matriz_Filial
Situação cadastral da empresa (nula, ativa, suspensa, inapta ou baixada)	Situacao_Cadastral
Data do evento da situação cadastral	Data_Situacao_Cadastral
Código do motivo da situação cadastral da empresa	Motivo_Situacao_Cadastral
Nome da cidade no exterior	Nome_Cidade_Exterior
Código do país	Pais
Data do início das atividades	Inicio_Atividades
Código da atividade econômica principal do estabelecimento	CNAE
Código das atividades econômicas secundárias do estabelecimento	CNAE_Secundaria
Descrição do tipo de logradouro	Tipo_Logradouro
Nome do logradouro onde se localiza o estabelecimento	Logradouro
Número onde se localiza o estabelecimento	NUMERO
Complemento para o endereço de localização do estabelecimento	Complemento
Bairro onde se localiza o estabelecimento	Bairro
Código de endereçamento postal referente ao logradouro no qual o estabelecimento está localizado	CEP
Sigla na unidade da federação em que se encontra o estabelecimento	UF
Código do município de jurisdição onde se encontra o estabelecimento	Municipio
DDD do telefone principal do estabelecimento	DDD1
Número de telefone principal do estabelecimento	Telefone1
DDD do telefone secundário do estabelecimento	DDD2
Número de telefone secundário do estabelecimento	Telefone2
DDD do fax do estabelecimento	DD_FAX
Número do fax do estabelecimento	FAX
Email do estabelecimento	Email
Situação especial do estabelecimento	Situacao_Especial
Data que o estabelecimento entrou em situação especial	Data_Situacao_Especial

Fonte: Autoria própria

Como o objetivo ao utilizar esta base é calcular o número de POIs por tipo de atividade

econômica disponíveis em cada AH de destino das viagens, inicialmente foram filtrados apenas os estabelecimentos localizados na RMBH e que estavam ativos em 2012, época de coleta de dados da edição da pesquisa domiciliar utilizada neste estudo. Em seguida, foram selecionadas as variáveis que compõe o número de CNPJ do estabelecimento, as variáveis que compõe o endereço e o código da atividade econômica principal, conforme apresentado na Quadro 10.

Quadro 10 – Variáveis do *dataset* POIs relevantes para serem mantidas no conjunto de dados *dataset* POIs

Descrição do atributo	**Código**
Número base de inscrição no CNPJ (oito primeiros dígitos do CNPJ)	CNPJ_Basico
Número de inscrição do estabelecimento no CNPJ (nono até o décimo segundo dígito do CNPJ)	CNPJ_Ordem
Dígito verificador do número de inscrição no cnpj (dois últimos dígitos do CNPJ)	CNPJ_Digito
Código da atividade econômica principal do estabelecimento	CNAE
Descrição do tipo de logradouro	Tipo_Logradouro
Nome do logradouro onde se localiza o estabelecimento	Logradouro
Bairro onde se localiza o estabelecimento	Bairro

Fonte: Autoria própria

O detalhamento dos códigos CNAE que foram considerados em cada categoria de POIs, bem como as descrições de cada atividade, estão apresentadas no Quadro 11.

Quadro 11 – CNAEs considerados por categoria POI

Categoria POI	**Descrição**	**Código(s) CNAE(s)**
Saúde	atividades de apoio à gestão de saúde	8660700
	atividades de atendimento hospitalar	8610101, 8610102
	atividades de assistência psicossocial e à saúde a portadores de distúrbios psiquicos deficiência mental e dependência química	8720401, 8720499
	atividades de atenção ambulatorial executadas por médicos e odontológos	8630599, 8630503, 8630502, 8630501, 8630504, 8630506, 8630507
Negócios	bancos comerciais	6421200
	caixas econômicas	6423900
	bancos múltiplos, com carteira comercial	6422100
	crédito cooperativo	6424703, 6424701, 6424704, 6424702
	atividades de serviços prestados principalmente às empresas não especificadas anteriormente	8299799, 8299706, 8299707, 8299702, 8299701, 8299704, 8299703, 8299705
Lazer	atividades de exibição cinematográfica	5914600
	artes cênicas, espetáculos e atividades complementares	9001902, 9001906, 9001999, 9001901, 9001904, 9001903, 9001905
	atividades de jardins botânicos, zoológicos, parques nacionais, reservas ecológicas e áreas de proteção ambiental	9103100
	atividades de recreação e lazer não especificadas anteriormente	9329899, 9329804, 9329801, 9329803, 9329802
	atividades de museus e de exploração, restauração artística e conservação de lugares e prédios históricos e atrações similares	9102301, 9102302
	parques de diversão e parques temáticos	9321200
Ensino	educação infantil - creche	8511200
	educação infantil - pré-escola	8512100
	ensino fundamental	8513900
	ensino médio	8520100
	educação superior - graduação e pós-graduação	8532500
	educação superior - pós-graduação e extensão	8533300
	educação profissional de nível tecnológico	8542200
	educação profissional de nível técnico	8541400
	ensino de idiomas	8593700
	atividades de apoio à educação e atividades de ensino não especificadas anteriormente	8599699, 8599603, 8599604, 8599605, 8599601, 8550301, 8550302, 8599602

Fonte: Autoria própria

5.3.2 Engenharia de Atributos e Limpeza de dados

Após a seleção das variáveis, a segunda etapa de processamento de dados aplicada foi a engenharia de atributos, que consiste em um processo de extração dos recursos relevantes e informativos dos dados brutos para alimentar os modelos de aprendizado de máquina.

Para a base de dados da pesquisa de viagem, a primeira manipulação aplicada foi a criação de uma variável denominada "Chave", que corresponde à identificação única de um indivíduo. A “Chave” é composta pelas variáveis “ID_Domicilio” e “ID_Individuo”, conforme apresentado no Quadro 12. Após a criação da "Chave", os registros do conjunto de dados foram ordenados para identificar todos os deslocamentos feitos pela mesma pessoa. Além disso, a variável “Qtd_Viagens_Individuo”, que corresponde à contagem de viagens feitas por cada pessoa, foi criada e incorporada à base de dados da OD_2012.

Quadro 12 – Exemplo de criação da variável ’chave’

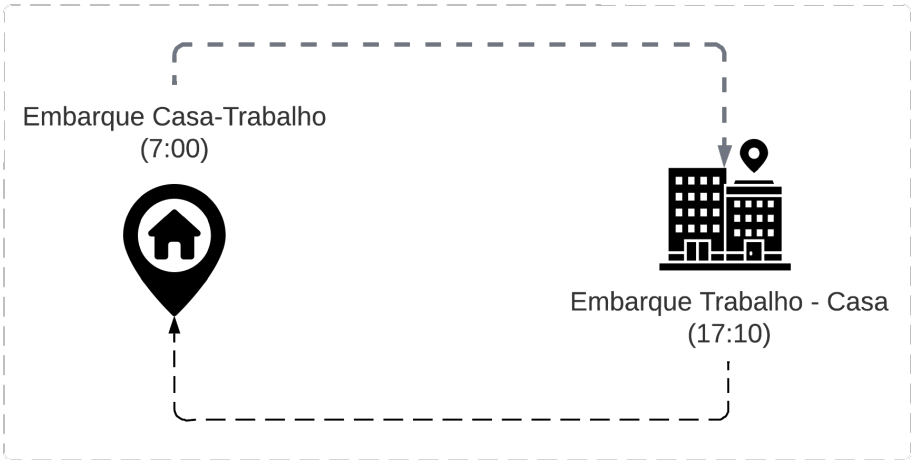
ID_Domicilio	**ID_Individuo**	**Chave**	**ID_Viagem**
100	1	100_1	1
100	1	100_1	2
200	1	200_1	1
200	1	200_1	2
200	1	200_1	3
200	2	200_2	1
200	2	200_2	2

Fonte: Autoria própria

A segunda manipulação feita nos dados consistiu na criação da variável denominada "Saida_Destino", que corresponde, para cada registro de viagem, ao horário do embarque subsequente realizado pelo mesmo passageiro. A Figura 8 ilustra as duas viagens do passageiro 100_1 que inicia seu primeiro deslocamento do dia às 7:00, saindo de casa e indo até o seu local de trabalho, com um tempo de deslocamento de 50 minutos e, ao final do dia, às 17:10, realiza o seu segundo embarque, saindo do seu local de trabalho e retornando à sua residência. O horário do segundo embarque deste passageiro corresponde à variável Saida_Destino.

Após identificar o horário em que o indivíduo saiu do destino, a terceira transformação realizada nos dados foi o cálculo do "Tempo_Entre_Embarques", sendo esta uma proxy da duração da atividade. Para o exemplo apresentado na Figura 8, o "Tempo_Entre_Embarques"foi de 10 horas e 10 minutos.

Figura 8 – Exemplo do cálculo do "Tempo_Entre_Embarques"



Fonte: Autoria própria

Para calcular o horário que o indivíduo saiu do destino e, consequentemente, o "Tempo_Entre_Embarques", duas premissas devem ser atendidas: o indivíduo precisa ter feito pelo menos duas viagens por transporte público coletivo no dia e o motivo de destino de cada viagem precisa corresponder ao motivo de origem da viagem subsequente. Casos onde pelo menos uma destas premissas não é atendida foram removidos da base, o que gerou uma perda de aproximadamente 19,6% dos dados. A maior parte dos registros perdidos correspondem a casos em que o passageiro utilizou outro modo de transporte em sua cadeia de viagem. O *dataset* final permaneceu com 21.160 registros.

Após calcular o "Tempo_Entre_Embarques", uma análise do padrão deste tempo por motivo de viagem foi realizada. A Figura 9 mostra as viagens produzidas e atraídas por motivo. A produção de viagens refere-se ao total de viagens originadas em uma determinada região, como uma área residencial, local de trabalho, escola, comércio ou qualquer outra localização.

Por outro lado, a atração de viagens refere-se à quantidade de viagens que são direcionadas para um destino específico, como um centro comercial, escola, parque de diversões, hospital, entre outros. É o oposto da produção de viagens.

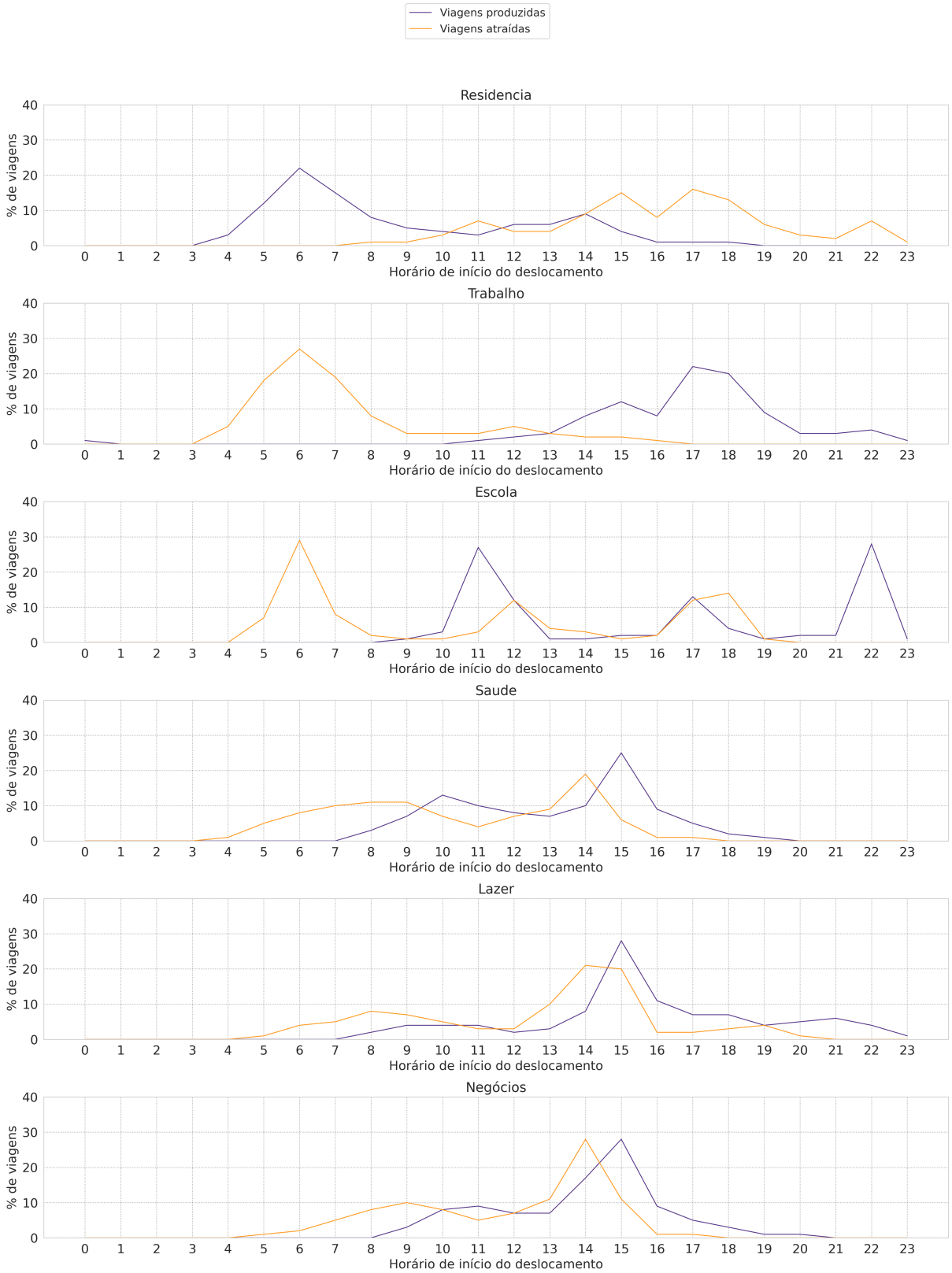
Pela análise da Figura 9, nota-se que viagens que têm o local de residência como ponto de origem acontecem, principalmente, nas primeiras horas do dia. Em outras palavras, áreas residênciais tendem a produzir um grande número de viagens pela manhã. Por outro lado, viagens atraídas para estas áreas ocorrem predominantemente no final da tarde. Este padrão resulta em uma duração (tempo de permanência em casa) de aproximadamente 13 horas, que pode ser estimado pela diferença, em horas, entre o pico da linha de viagens atraídas e o pico da linha de viagens produzidas apresentado na figura.

Para o trabalho, nota-se um comportamento oposto, em que a atração de viagens acontece no período da manhã e a produção no período da tarde, resultando em um padrão de duração de aproximadamente 11 horas. Para o motivo escola, nota-se que o gráfico apresenta três picos tanto para a produção quanto para a atração de viagens, estes três picos estão associados aos períodos

escolares da manhã, tarde e noite, respectivamente, com um padrão de duração de 5 horas.

Para o motivo de viagem secundário de saúde, nota-se que a duração é de aproximadamente 1 hora e as viagens acontecem no meio da manhã ou no meio da tarde. Já para as viagens a lazer, a duração também é curta mas a maioria dos deslocamentos por este motivo ocorrem no período da tarde e noite.

Figura 9 – Padrão de atração, produção e duração por tipo de atividade



Fonte: Autoria própria

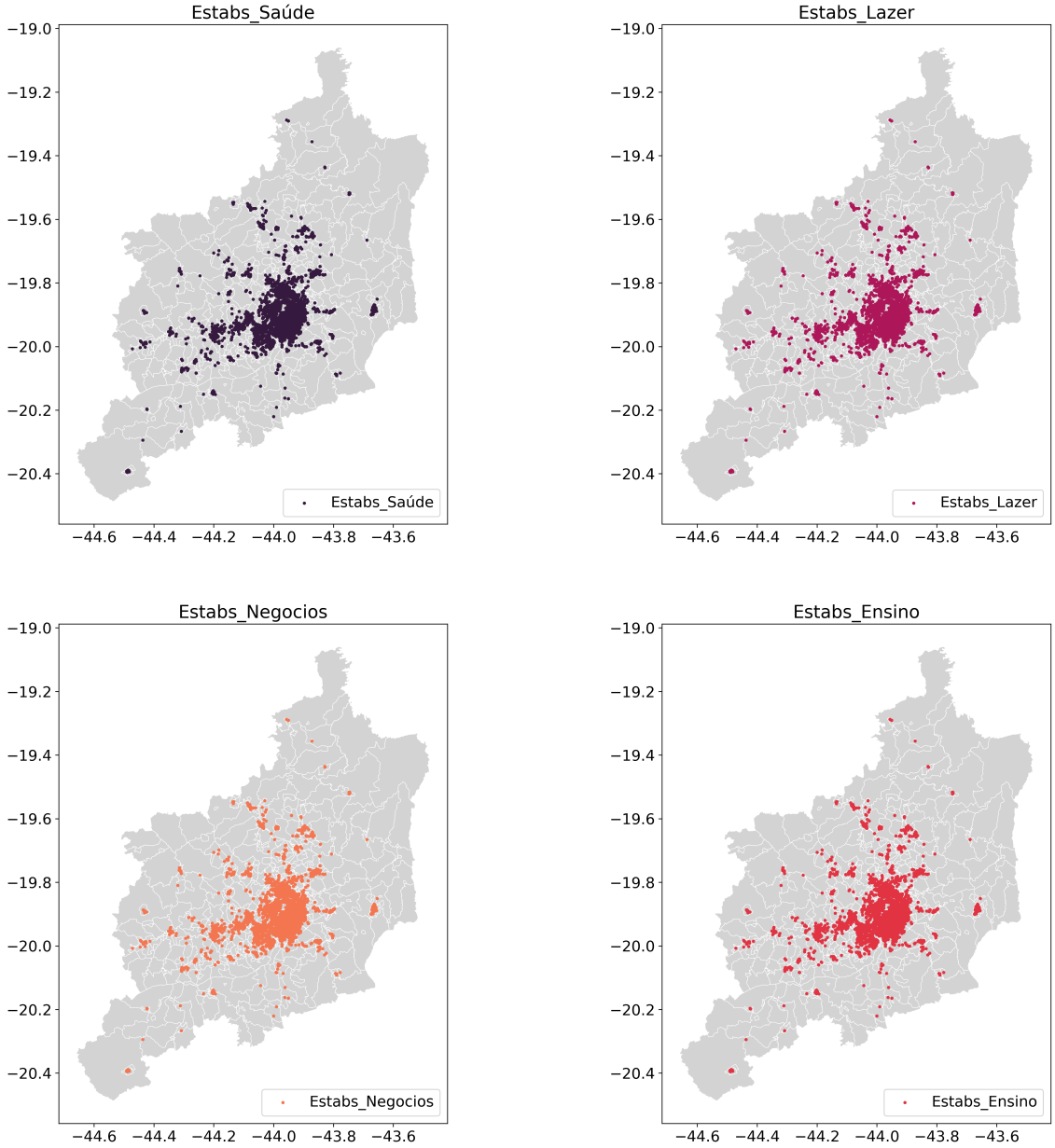
Para a base de dados de CNPJ, a primeira manipulação feita foi a concatenação das partes que compõe o CNPJ do estabelecimento, sendo elas o “CNPJ_Base”, o “CNPJ_Ordem” e o

‘CNPJ_Digito”.

Em seguida, após filtrar apenas os estabelecimentos com os códigos CNAE de interesse (Quadro 11), uma coluna de endereço foi criada a partir da concatenação das variáveis “Tipo_Logradouro”, “Logradouro” e “Bairro”. Esta coluna foi utilizada para geocodificar os endereços obtendo a latitude e longitude correspondente com o auxílio da biblioteca *open source* Nominatim do Python, módulo que permite a busca e geocodificação de endereços usando o serviço Nominatim do *OpenStreetMap*. Os detalhes da API Nominatim estão disponíveis em OPENSTREETMAP (2023)

De posse dos dados de estabelecimentos geocodificados, a biblioteca Geopandas, do Python, foi utilizada para unir os dados espaciais de localização dos estabelecimentos e de delimitação das AHs da RMBH. Assim, foi possível identificar não apenas a localização geográfica de cada estabelecimento, mas também a qual AH ele pertence. Os detalhes da biblioteca Geopandas estão disponíveis em Jordahl et al. (2020). O arquivo geográfico em formato shapefile (.shp) da delimitação das AH da RMBH está disponível no repositório do *GitHub*, junto às bases de dados OD_2012 e POIs no *link*: <https://github.com/MirianGreiner/Trip_Purpose_Inference.git>. A Figura 10 apresenta de forma gráfica o resultado da geolocalização dos estabelecimentos na RMBH por tipo de atividade.

Figura 10 – Distribuição dos Pontos de Interesse no território da RMBH



Fonte: Autoria própria

Os estabelecimentos cujos valores de latitude e longitude retornados no processo de geocodificação estavam fora dos limites esperados para a RMBH ou estabelecimentos que não foram encontrados pela biblioteca Nominatim foram removidos da base de dados.

Ao final, foram contabilizados, por AH, a quantidade de estabelecimentos para cada um dos tipos de atividade, criando assim as variáveis: Estabs_Lazer, Estabs_Negocios, Estabs_Saude, Estabs_Ensino.

5.3.3 Entendimento das bases de dados pós processamento

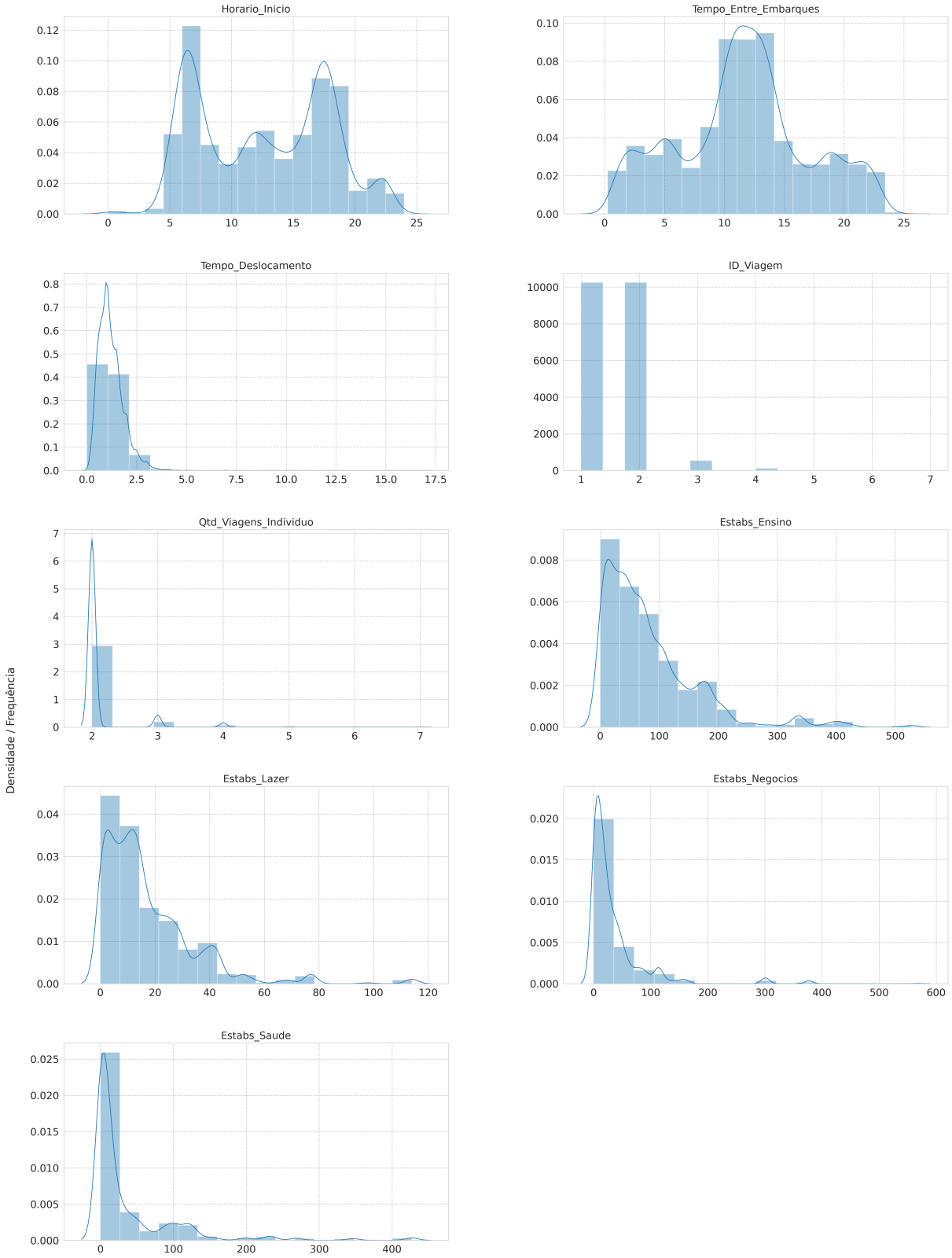
A análise descritiva dos dados é uma etapa fundamental no processo de modelagem usando algoritmos de aprendizado de máquina, pois ela permite identificar valores extremos,

tendências e padrões, além da distribuição dos dados e relação entre as variáveis, o que pode ser usado para selecionar os melhores algoritmos para o problema em questão.

Após a seleção de variáveis e engenharia de atributos, a base de dados da OD_2012 permaneceu com as variáveis: “Horario_Inicio”, “Tempo_Entre_Embarques”, “Tempo_Deslocamento”, “ID_Viagem”, “Qtd_Viagens_Individuo” e “Motivo_Destino”. A base de dados de POIS permaneceu com as variáveis: “Estabs_Lazer”, “Estabs_Negocios”, “Estabs_Saude” e "Estabs_Ensino". Inicialmente, foi feita uma análise da distribuição para estas variáveis. Saber a distribuição das variáveis é fundamental para escolher os testes estatísticos apropriados a serem aplicados, interpretar corretamente os resultados obtidos e escolher os algoritmos adequados para a etapa de modelagem, pois muitos modelos de aprendizado de máquina são sensíveis à distribuição dos dados, conforme detalhado na Subseção 3.2.2.

A Figura 11 apresenta a distribuição das variáveis utilizando histogramas e curvas de densidade para as variáveis numéricas e uma contagem de valores únicos para a variável categórica. Conforme apresentado, nenhuma das variáveis seguem uma distribuição normal, o que foi comprovado também pelo teste de Shapiro-Wilk realizado para testar a normalidade, no caso das variáveis numéricas.

Figura 11 – Distribuição das variáveis

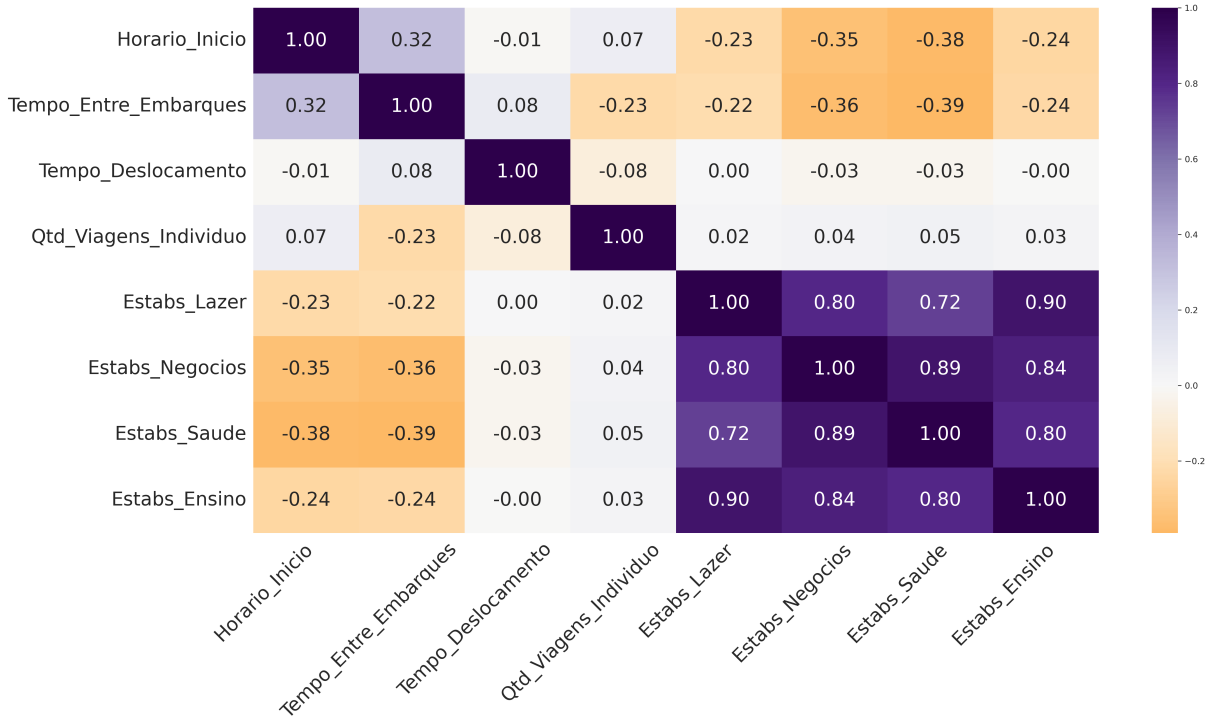


Fonte: Autoria própria

Após analisar a distribuição das variáveis preditoras, foi feita uma análise de correlação para as variáveis numéricas e de associação para a variável categórica com o objetivo de identificar se há multicolinearidade entre elas. A importância da análise de correlação e demais

conceitos pertinentes ao assunto estão descritos detalhadamente na Subseção 3.2.3. Para a análise de correlação utilizou-se o método de *Spearman*, que é uma medida de correlação não paramétrica que avalia a relação monotônica entre duas variáveis. Este coeficiente é adequado quando deseja-se detectar associação de natureza não linear entre variáveis ou quando as variáveis não possuem uma distribuição normal, como é o caso deste estudo. O resultado da análise de correlação está apresentado no *heatmap* da Figura 12, onde os valores indicam a força e a direção da relação entre as variáveis.

Figura 12 – Análise de correlação para variáveis numéricas



Fonte: Autoria própria

Pelos resultados, nota-se que as variáveis que representam a quantidade de POIs por AH apresentam alta correlação entre si. Isso pode ser justificado pela alta concentração de POIs nas áreas centrais da RMBH, independente do tipo de estabelecimento, conforme apresentado na Figura 10.

Para a variável preditora categórica de “ID_Viagem”, foi feita uma análise de associação com as demais variáveis preditoras numéricas. Para isso, foi utilizado o teste não paramétrico de *Kruskal-Wallis*, detalhadamente descrito na Subseção 3.2.3. Em resumo, este teste avalia se as distribuições das variáveis numéricas diferem significativamente entre os níveis da variável categórica. Os resultados obtidos sugerem que a variável preditora “ID_Viagem” está associada a cada uma das demais variáveis preditoras (“Horario_Inicio”, “Tempo_Entre_Embarques”, “Tempo_Deslocamento”, “Qtd_Viagens_Individuo”, “Estabs_Lazer”, “Estabs_Negocios”, “Estabs_Saude”, “Estabs_Ensino”), pois em todos os casos obteve-se um p-valor menor que o nível de significância escolhido (0,05). Os resultados indicam, ainda, que as variáveis preditoras numéricas

que estão mais fortemente associadas com a variável preditora categórica de “ID_Viagem” são: “Horario_Inicio”, “Tempo_Entre_Embarques” e “Qtd_Viagens_Individuo”, que apresentaram os maiores valores de estatística de teste.

O teste de *Kruskall-Wallis* também foi utilizado para avaliar a associação entre as variáveis preditoras numéricas e a variável *target* que é categórica. Os resultados obtidos sugerem que há associação entre a variável de *target* e cada uma das variáveis preditoras numéricas, pois em todos os casos obteve-se um p-valor menor que o nível de significância escolhido (0,05). As variáveis preditoras numéricas que estão mais fortemente associadas com a variável *target* são: “Horario_Inicio” e “Tempo_Entre_Embarques”, que apresentaram os maiores valores de estatística de teste. Este resultado, indica que estas podem ser as variáveis mais significativas para o treinamento do modelo de inferência. As variáveis “Estabs_Negocios” e “Estabs_Saude”, também apresentaram valores elevados de estatística, o que indica que elas também serão variáveis importantes no modelo de inferência.

Para avaliar a associação entre a variável preditora categórica de “ID_Viagem” e a variável *target*, utilizou-se o teste do qui-quadrado. Este é um teste estatístico não paramétrico que determina se existe uma associação estatisticamente significativa entre duas variáveis categóricas assumindo a hipótese nula de independência entre elas. A descrição detalhada pode ser encontrada em Subseção 3.2.3. O resultado do teste de qui-quadrado indicou que há uma associação significativa entre as variáveis analisadas, com p-valor menor do que o nível de significância escolhido (0,05).

5.3.4 Normalização de variáveis

Normalizar variáveis significa ajustar valores para uma escala comum, conforme descrito detalhadamente na Subseção 5.3.4. Para normalizar as variáveis preditoras numéricas, utilizou-se a ferramenta MinMaxScaler do Python, que transforma os valores de uma coluna em um novo valor dentro de um intervalo que varia de 0 a 1. Este processo é realizado pela aplicação da seguinte fórmula:

$$x\_scaled = \frac{x - min(x)}{max(x) - min(x)} \tag{30}$$

onde x é o valor original e x_scaled é o valor transformado na nova escala e os valores $min(x)$ e $max(x)$ são os valores mínimo e máximo da coluna original, respectivamente.

5.3.5 Codificação de variáveis categóricas

Assim como a normalização é uma etapa crucial no processamento dos dados antes que eles possam ser apresentados a um algoritmo de aprendizado de máquina, a codificação de variáveis categóricas também é fundamental neste processo.

O processo de codificação consiste em transformar categorias de uma variável categórica em valores numéricos que podem ser interpretados e processados pelos modelos, conforme descrito na Seção 3.3. Ainda que os algoritmos utilizados neste estudo não requeiram, especificamente, a codificação da variável *target*, optou-se por realizar mais esta etapa de transformação de dados para garantir maior precisão nos resultados obtidos. Para codificar a variável target, é utilizada a técnica *One-hot Encoding*.

5.4 Métodos de inferência do motivo de viagem

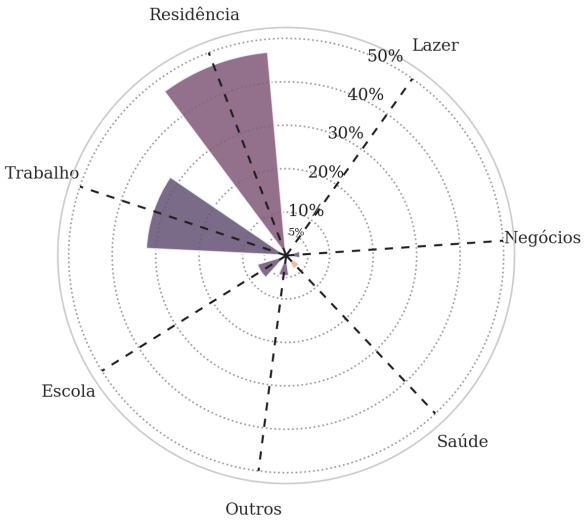
Após finalizar a etapa de processamento, as duas bases de dados (OD_2012 e POIS) estão prontas para serem utilizadas pelos modelos de classificação. Nesta etapa, três diferentes métodos de inferência são testados e, ao final, o que apresentar os melhores resultados é escolhido para aplicação na base de dados de cartões inteligentes para inferir os motivos de viagem dos passageiros.

5.4.1 Método 1

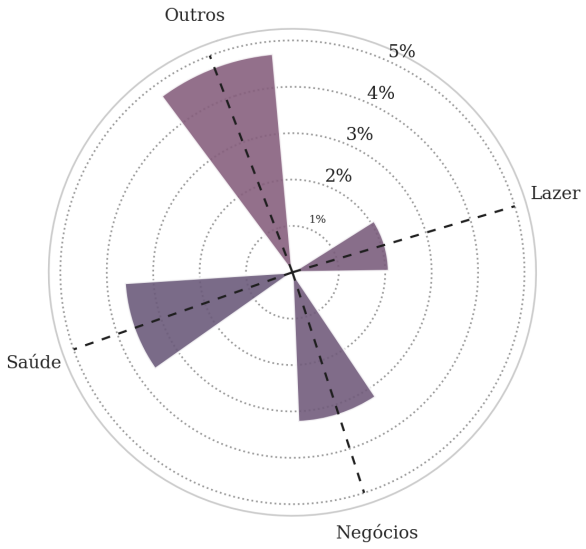
No Método 1, um modelo único de classificação multiclasse é treinado. Neste método, apenas variáveis de viagem foram utilizadas no conjunto de preditores e o motivo de viagem dividido em sete diferentes classes foi utilizado como variável *target*. Inicialmente, uma análise do percentual de registros de viagem por classe foi realizada, pois sabe-se que muitos algoritmos de classificação assumem que há um equilíbrio entre as classes da variável target. Quando o desbalanceamento entre classes é significativo, o modelo pode encontrar dificuldades para identificar padrões em classes minoritárias, levando a um desempenho pior.

Conforme apresentado na Figura 13, a base de dados OD_2012 está fortemente desbalanceada: 47,08% dos registros correspondem ao motivo Residência, 32,27% ao motivo trabalho, 6,98% ao motivo escola, 4,73% a outros motivos de viagem, 3,62% ao motivo saúde, 3,23% ao motivo negócios, e, 2,08% ao motivo lazer.

Figura 13 – Percentual e viagens por motivo de destino



((a)) Geral



((b)) Viagens secundárias

Fonte: Autoria própria

O forte desbalanceamento de classes para a variável *target* possui alguns motivos. O primeiro deles está relacionado com a forma com que os dados da pesquisa domiciliar foram coletados, pois os entrevistados foram questionados sobre suas viagens realizadas em um dia útil típico. Em dias úteis típicos, as viagens obrigatórias são as mais comuns e, por este motivo, atividades secundárias ficam sub-representadas.

Outro ponto a ser considerado é que, para determinados motivos de viagem, os indivíduos entrevistados pela OD_2012 optaram por outros modos de transporte em detrimento do transporte coletivo. Conforme apresentado na Tabela 2, de todos os registros de viagem coletados pela pesquisa domiciliar (utilizando a base de dados completa e sem nenhum filtro), 3,45% são viagens a lazer e, para este motivo de viagem os modos "automóvel particular" e "caminhada ou bicicleta" foram os mais utilizados. Um comportamento semelhante é observado nas viagens

escolares, que representam 16% do total de viagens. Contudo, os principais modos utilizados nestes deslocamentos foram o transporte escolar ou o transporte não motorizado (caminhada ou bicicleta).

Conforme já destacado anteriormente, apenas os registros de viagem da Pesquisa OD feitas por ônibus ou metrô serão utilizados neste estudo para treinamento dos modelos, assim como foi feito por Pezoa et al. (2023). Isto, porque o objetivo é treinar um algoritmo que possa prever, com precisão, o motivo de viagem dos passageiros de transporte público utilizando dados de cartões inteligentes. Nesse contexto, ao filtrar apenas as viagens feitas por ônibus ou metrô na base de dados da OD_2012, o desbalanceamento de classes torna-se ainda mais intenso.

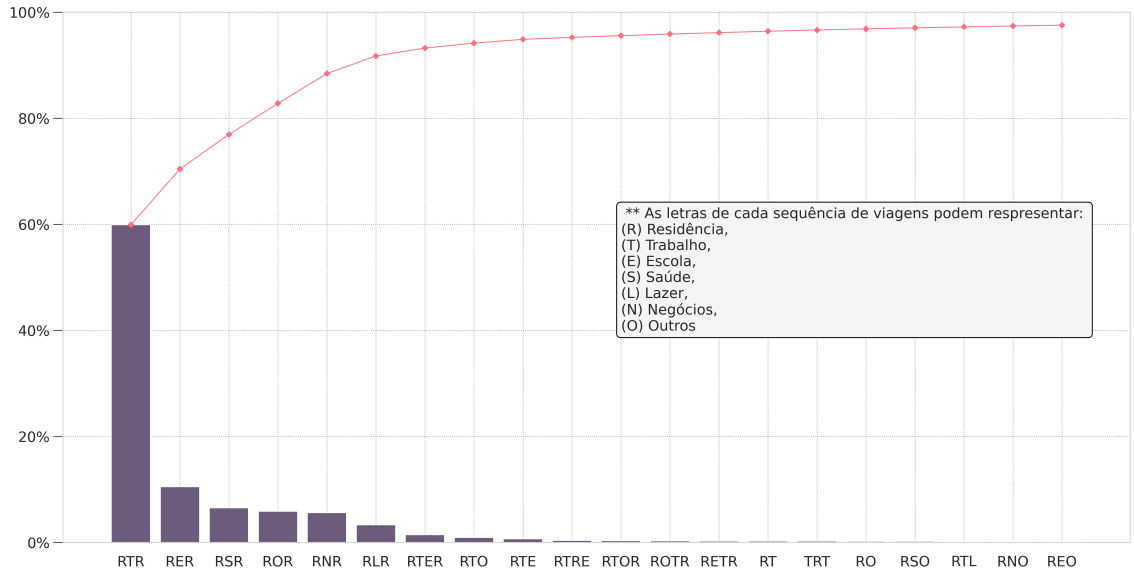
Tabela 2 – Distribuição das viagens da Pesquisa Origem e Destino RMBH (2012) por motivo de viagem e por modo de transporte

Motivo de Viagem	**Modo de Transporte**				Viagens por motivo (total da linha)
	Transporte Escolar/Outros	automóvel particular	caminhada /bicicleta	ônibus /metrô	
Escola	3,74%	2,58%	8,19%	1,52%	16,03%
Lazer	0,09%	1,41%	1,46%	0,49%	3,45%
Negócios	0,04%	0,70%	0,59%	0,68%	2,01%
Outros	0,10%	2,90%	3,84%	1,06%	7,90%
Residência	5,33%	14,42%	17,59%	9,31%	46,65%
Saúde	0,15%	0,74%	0,55%	0,74%	2,18%
Trabalho	1,35%	8,55%	5,69%	6,18%	21,78%
Viagens por modo (total da coluna)	10,81%	31,29%	37,92%	19,99%	100,00%

Fonte: Autoria própria

Por fim, outro motivo que contribui para o desbalanceamento de classes da variável target é a característica *home-based* das viagens feitas por transporte público coletivo na Região Metropolitana de Belo Horizonte. Viagens *home-based* são aquelas que começam ou terminam em casa, conforme apresentado na Figura 14. Por isso, o motivo "Residência"está presente em aproximadamente 50% dos deslocamentos. Pela análise da Figura 14, nota-se que a maior parte das viagens possuem um padrão de início e término na residência, com a realização de apenas uma atividade no intervalo, sendo o trabalho a atividade mais comumente realizada, seguido de escola, saúde, outros, negócios e lazer.

Figura 14 – Percentual de viagens por padrão



Fonte: Autoria própria

A técnica da *Hold-Out* descrita na Seção 3.3 foi aplicada ao conjunto de dados para subdividi-lo em dois: subconjunto de dados de treino e subconjunto de dados de teste, na proporção 80-20, e para evitar que o modelo fique enviesado em função do desbalanceamento de classes descrito acima, três técnicas comuns de balanceamento de classes foram testadas, todas elas de sobreamostragem (*Oversampling*) das classes minoritárias no subconjunto de dados de treino: a *Random Oversampling* sem reposição, a SMOTE *(Synthetic Minority Over-sampling Technique)* e a ADASYN *(Adaptive Synthetic Sampling)*.

Em relação ao algoritmo, para garantir que o melhor modelo seja escolhido para o problema em questão, quatro classificadores foram testados. Por se tratar de um problema de classificação multiclasse com uma base de dados desbalanceada e que contará com variávies não normais, foram escolhidos os algoritmos *Random Forest*, *Suport Vector Machines* (SVM), *Bagging Classifier* e *XGBoost*.

Para selecionar os valores mais adequados de hiperparâmetros para cada algoritmo foi utilizada uma abordagem da busca em grade, pois esta técnica tem a vantagem de encontrar soluções mais precisas, ainda que apresente um maior tempo de computação maior. Uma descrição detalhada sobre o método de busca em grade e sobre alguns hiperparâmetros importantes para cada algoritmo pode ser encontrada em Subseção 3.4.2.

A grade de hiperparâmetros considerada para cada algoritmo está apresentada na Tabela 3 e o *range* de valores utilizados foi baseado intervalos comumente utilizados na literatura.

Tabela 3 – Grade de hiperparâmetros testada para cada algoritmo

Random Forest	
Parametro	Grade de valores
n_*estimators*	[100, 500, 1000],
criterion	[gini, entropy]
max_*depth*	[5, 10]
min_*samples*_leaf	[1,2]
max_*features*	[sqrt, log2]

Support Vector Machine	
Parametro	Grade de valores
C	[0.1, 1, 10]
Kernel	[rbf]
Gamma	[0.1, 1, 10]

XGBoost	
Parametro	Grade de valores
n_*estimators*	[100,500]
learning_rate	[0.1, 0.01]
max_*depth*	[3,5,7]
subsample	[0.8, 1.0]

Bagging Classifier	
Parametro	Grade de valores
n_*estimators*	[50, 100, 200]
max_*samples*	[0.5, 0.7, 1]
max_*features*	[0.5, 0.7, 1]
bootstrap	[True, False]
bootstrap_features	[True, False]

Fonte: Autoria própria

O desempenho dos modelos para cada combinação de hiperparâmetros foi avaliado usando validação cruzada com a estratégia StratifiedKFold, que é especialmente útil para lidar com dados desbalanceados. A técnica de validação cruzada divide os dados do conjunto em k partições (ou dobras) e o modelo é treinado k vezes, cada vez usando k-1 dobras para treinar o modelo e a dobra restante para testar. Ao utilizar a estratégia StratifiedKFold, o processo de divisão dos folds é feito de forma estratificada, o que significa que a proporção de cada classe é mantida em cada fold. Isso ajuda a garantir que cada fold seja representativo da distribuição geral no conjunto de dados. Devido à sua robustez, a técnica de validação cruzada ajuda a evitar a dependência da divisão aleatória inicial e o desempenho do modelo é avaliado pela média dos resultados obtidos em cada iteração. Neste estudo foi utilizado k=5 e, portanto, em cada iteração 4 partes foram usadas para treino e uma parte para teste de cada um dos modelos implementados. O valor de k escolhido para utilização neste estudo, também foi baseado nos valores comumente utilizados na literatura.

Ao final do processo de busca em grade com validação cruzada, foram encontrados os hiperparâmetros mais adequados para cada algoritmo e, em seguida, cada classificador foi treinado novamente utilizando estes hiperparâmetros e diferentes técnicas de balanceamento: *Random Oversampling*, SMOTE e ADASYN.

Cada combinação de algoritmo + técnica de balanceamento foi treinada e testada por 30 vezes e em cada iteração os resultados das métricas de acurácia, MCC, *precision*, *recall* e *F1-score* foram armazenados em um dataset para posteriormente serem utilizados no teste estatístico não paramétrico de *Kruskal-Wallis*, com o objetivo de avaliar se há diferenças estatisticamente significativas entre as médias para as 12 combinações possíveis, conforme demonstrado no Quadro 13.

Quadro 13 – Combinações testadas no Método 1. As combinações foram geradas pela utilização de 4 diferentes algoritmos e 3 técnicas de balanceamento

Algoritmos	**Técnicas de Balanceamento**		
	***Random Oversampling* (ROS)**	**SMOTE**	**ADASYN**
***Random Forest* (RF)**	RF_ROS	RF_SMOTE	RF_ADASYN
SVM	SVM_ROS	SVM_SMOTE	SVM_ADASYN
***Bagging Classifier* (BC)**	BC_ROS	BC_SMOTE	BC_ADASYN
XGBoost (XGB)	XGB_ROS	XGB_SMOTE	XGB_ADASYN

Fonte: Autoria própria

Após esta validação, caso o resultado do teste seja positivo, é escolhida a combinação de algoritmo e técnica de balancemento de classes que apresentar o melhor resultado. Caso não existam diferenças com significância estatística entre os resultados, é escolhida a combinação cuja implementação for mais simples e tiver o menor tempo de treinamento.

5.4.2 Método 2 e Método 3

Após a implementação do teste estatístico e avaliação dos resultados, a combinação de algoritmo e técnica de balanceamento de classes com o melhor resultado é utilizada no Método 2 para treinar dois modelos de classificação.

No primeiro modelo, os mesmos dados de viagem são utilizados como variáveis preditoras, no entanto, a variável target será composta por apenas quatro classes. Isto porque será feita uma agregação dos motivos secundários de saúde, lazer, negócios e outros em apenas uma classe denominada "Secundário".

O segundo modelo é treinado apenas com os dados de viagens secundárias, portanto, o *dataset* de treinamento é filtrado para conter apenas os registros destas viagens e, consequentemente, a variável *target* deste modelo é composta pelas classes de “lazer”, “negócios”, “saúde” e “outros”.

Ao final, ambos modelos são avaliados utilizando as métricas já descritas de *precision*, *recall*, F1-*score*, acurácia e coeficiente de correlação de Matthews.

A única diferença do Método 3 em relação ao Método 2 é que as variáveis espaciais de "Estabs_Saude", "Estabs_Lazer", "Estabs_Negocios" e "Estabs_Ensino", da base de dados POIs, são incorporadas no conjunto de variáveis preditoras e os dois modelos do Método 2 são novamente treinados. Os resultados também são avaliados utilizando as métricas de *precision*, *recall*, *F1-score*, acurácia e coeficiente de correlação de Matthews, conforme esquematizado na Figura 5.

5.5 Resumo do capítulo

O objetivo geral deste estudo consiste na aplicação de abordagens emergentes de mineração de dados para inferir informações ausentes em fontes de *Big Data* utilizadas no planejamento de transporte, notadamente o motivo de viagem do passageiro. O estudo de caso escolhido para aplicação da metodologia foi a Região Metropolitana de Belo Horizonte (RMBH).

Optou-se por utilizar uma abordagem supervisionada de aprendizado de máquina e, portanto, a base de dados rotulada considerada foi a da última pesquisa domiciliar de viagens feita na RMBH em 2012. Na etapa de modelagem dos dados, três diferentes métodos de inferência foram implementados, sendo que as principais diferenças entre eles foram o conjunto de variáveis preditoras utilizadas e a quantidade de classes disponíveis na variável *target*, que corresponde ao motivo de viagem.

No Método 1, um modelo único de classificação multiclasse foi treinado para rotular os registros de viagem em uma das sete classes: "Residência", "Trabalho", "Escola", "Saúde", "Lazer", "Negócios" e "Outros'. Neste método, utilizou-se como variáveis preditoras o horário de início da viagem, o tempo entre embarques, a quantidade de viagens feitas pelo indivíduo e o número sequencial da viagem. Como este foi o primeiro método implementado, para escolher o melhor cenário de modelagem foram testados quatro algoritmos e três técnicas de balanceamento de classes, o que deu origem a 12 combinações. Além disso, uma busca em grade foi realizada para encontrar os hiperparâmetros mais adequados para cada combinação.

No Método 2, o objetivo foi validar se a divisão do modelo único de classificação do Método 1 em duas etapas gera melhores resultados. Para isto, a melhor combinação de algoritmo, técnica de balanceamento e hiperparâmetros encontrados no Método 1 foi aplicada para prever o motivo de viagem em duas etapas sequenciais. Na primeira, o modelo foi treinado para classificar os registros de viagem em quatro classses, sendo elas: "Residência", "Trabalho", "Escola" e "secundários". Na segunda, os registros classificados como "secundários" na primeira etapa foram reclassificados nos motivos de "Saúde", "Lazer", "Negócios" e "Outros'. As variáveis preditorias utilizadas no Método 2 foram as mesmas do Método 1.

Por fim, no Método 3, foi avaliada a importância de se considerar variáveis espaciais no conjunto de preditoras para obter melhores resultados na previsão do motivo de viagem. Para isso, todos os passos do Método 2 foram replicados, no entanto, no conjunto de variáveis preditoras foram incluídas quatro variáveis espaciais que representam o número de Pontos de Interesse

de saúde, lazer, negócios e ensino disponíveis na área homogênea de destino do deslocamento. Estas quantidades foram obtidas a partir do uso da base de dados de CNPJ da Receita Federal Brasileira. Cada estabelecimento foi geolocalizado utilizando-se a biblioteca Nominatim do Python e posteriormente atribuído a uma área homogênea, viabilizando assim a contagem de POIs por tipo de serviço e por área homogênea.

Após a implementação de cada método, métricas comuns de avaliação de modelos de classificação multiclasse foram utilizadas para definir qual o método que traz os melhores resultados. Esta discussão está apresentada no capítulo seguinte.

6 Apresentação e discussão dos resultados

Neste capítulo, são apresentados os resultados obtidos a partir da aplicação dos três métodos de inferência do motivo de viagem apresentados no Capítulo 3. O objetivo principal é avaliar o desempenho de diferentes abordagens de inferência do motivo de viagem, visando selecionar o método mais adequado para implantação prática, proporcionando insights valiosos para aplicações futuras no campo do planejamento de transportes e mobilidade urbana. Para isso, são utilizadas métricas de avaliação como *precision*, *recall*, F1-*score*, entre outras. Para a comparação dos resultados, são utilizados testes estatísticos como o de *Kruskal-Wallis* e o teste de Dunn.

6.1 Método 1

Conforme descrito na Subseção 5.4.1, no Método 1 uma busca em grade foi utilizada para definição dos hiperparâmetros mais adequados para cada algoritmo treinado. Os hiperparâmetros encontrados estão apresentados na Tabela 4.

Tabela 4 – Hiperparâmetros selecionados pela busca em grade para cada algoritmo

Random Forest	
Parametro	valor
n_*estimators*	1000,
criterion	*entropy*
max_*depth*	10
min_*samples_leaf*	1
max_*features*	sqrt

Support Vector Machine	
Parametro	Grade de valores
C	10
Kernel	rbf
Gamma	10

XGBoost	
Parametro	Grade de valores
n_*estimators*	500
learning_rate	0.1
max_*depth*	7
subsample	0.8

Bagging Classifier	
Parametro	Grade de valores
n_*estimators*	200
max_*samples*	1
max_*features*	1
bootstrap	True
bootstrap_features	False

Fonte: Autoria própria

Em seguida, cada classificador foi treinado novamente com 30 iterações, utilizando os hiperparâmetros descritos na Tabela 4 e diferentes técnicas de balanceamento de classes, com

o objetivo de rotular os registros de viagem em uma das sete classes possíveis (“Residência”, “Trabalho”, “Escola”, “Lazer”, “Negócios”, “Saúde” e “Outros”). Como preditoras, foram usadas apenas as variáveis de viagem: “Horario_Inicio”, “Tempo_Entre_Embarques”, “ID_Viagem”, “Tempo_Deslocamento”, “Qtd_Viagens_Individuo”.

Na classificação multiclasse, algumas métricas como acurácia e MCC são calculadas de forma global enquanto outras métricas como as de *precision*, *recall* e F1-*score* são calculadas individualmente por classe. Para obter um valor que represente o desempenho geral do modelo, para estas métricas, é comum calcular a média, que pode ser macro, micro ou *weighted*). A média macro é obtida tratando todas as classes igualmente, independente de quantas instâncias cada classe tenha. Esta forma de calcular a média é útil quando deseja-se dar igual importância a todas as classes. A média micro é uma forma de agregação útil quando a análise do desempenho global do modelo é mais importante do que a análise por classe. Por fim, a média *weighted* é calculada levando-se em consideração o número de amostras em cada classe. Isso significa que classes com mais amostras têm mais impacto na métrica. Essa abordagem é útil quando deseja-se considerar as proporções de cada classe ao avaliar o desempenho do modelo como um todo.

Como neste estudo todas as classes possuem igual importância, optou-se por utilizar a média macro para avaliação dos modelos em relação às métricas de *precision*, *recall* e F1-*score*.

A Tabela 5 apresenta a médias dos resultados obtidos nas 30 iterações para as métricas de acurácia, MCC, *macro avg precision*, *macro avg recall* e *macro avg F1-score* para cada um dos quatro algoritmos implementados.

Tabela 5 – Média dos resultados das medidas de avaliação para cada conjunto de algoritmo + técnica de balanceamento

	Random forest		
	Random Oversampling	**SMOTE**	**ADASYN**
Acurácia	0,838	0,838	0,835
MCC	0,758	0,757	0,753
macro avg precision	0,537	0,533	0,524
macro avg recall	0,542	0,542	0,537
macro avg F1-score	0,527	0,527	0,521

	SVM		
	Random Oversampling	**SMOTE**	**ADASYN**
Acurácia	0,835	0,834	0,833
MCC	0,754	0,752	0,751
macro avg precision	0,530	0,531	0,523
macro avg recall	0,545	0,545	0,542
macro avg F1-score	0,521	0,521	0,516

	XGBoost		
	Random Oversampling	**SMOTE**	**ADASYN**
Acurácia	0,828	0,834	0,834
MCC	0,742	0,749	0,750
macro avg precision	0,518	0,522	0,521
macro avg recall	0,524	0,518	0,518
macro avg F1-score	0,520	0,519	0,519

	Bagging Classifier		
	Random Oversampling	**SMOTE**	**ADASYN**
Acurácia	0,810	0,800	0,787
MCC	0,717	0,704	0,686
macro avg precision	0,494	0,482	0,470
macro avg recall	0,506	0,496	0,491
macro avg F1-score	0,500	0,487	0,479

Fonte: Autoria própria

Conforme apresentado na Tabela 5, a implementação de 4 diferentes algoritmos e 3 técnicas de balanceamento de classes deu origem a 12 combinações possíveis e, para avaliar se pelo menos uma delas tem distribuição estatisticamente diferente das outras para cada uma das métricas de avaliação consideradas (acurácia, MCC, *precision*, *recall* e F1-*score*), testes de *Kruskal-Wallis* foram realizados. Este teste foi escolhido por ser uma alternativa não paramétrica ao teste ANOVA, que não faz suposição sobre a distribuição normal dos dados. Além disso, este teste permite a comparação de múltiplos grupos em relação a uma ou mais métricas de avaliação. Para cada uma das métricas de avaliação, foi implementado um teste de *Kruskal-Wallis* e, como resultado, observou-se que em todos os testes houve evidências de que pelo menos uma das combinações teve uma distribuição diferente, já que o resultado do p-valor foi inferior ao nível de significância de 0,05. Uma descrição detalhada do teste de *Kruskal-Wallis* pode ser encontrada em Subseção 3.2.3.

Como o resultado do teste de *Kruskal-Wallis* foi estatisticamente significativo, uma análise *post-hoc* foi necessária, para identificar quais pares de combinações tiveram desempenho

significativamente diferente, pois apenas o teste de *Kruskal-Wallis* não fornece esta informação. Para isto, foi utilizado o teste de Dunn, que é apropriado para comparações múltiplas, além de fornecer resultados detalhados sobre quais pares de combinações foram estatisticamente diferentes, permitindo uma interpretação mais precisa dos algoritmos. Uma descrição detalhada do teste de Dunn pode ser encontrada em Subseção 3.2.3.

Ao analisar o desempenho dos algoritmos pela métrica da acurácia, observa-se, pela análise da Tabela 6, que o algoritmo *Bagging Classifier*, independente da técnica de balanceamento utilizada em conjunto, apresentou desempenho estatisticamente diferente de quase todas as demais combinações. Além disso, o algoritmo *Random Forest* apresentou diferenças em relação aos algoritmos SVM e *XGBoost*. Na Tabela 6 foram apresentados apenas os resultados de p-valor menores do que o nível de significância escolhido para o teste de Dunn, que foi de 5% (0,05).

Tabela 6 – Resultados do teste de Dunn para a medida de acurácia onde os valores correspondem ao p-valor do teste estatístico entre pares

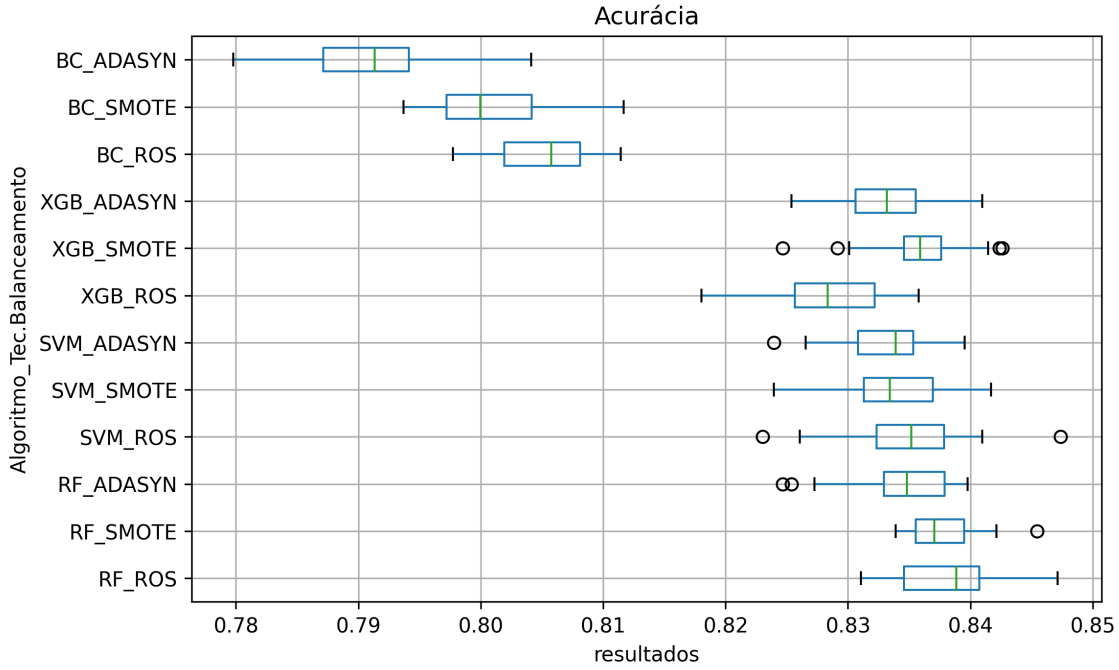
	RF_ROS	**RF_SMOTE**	**RF_ADASYN**	**SVM_ROS**	**SVM_SMOTE**	**SVM_ADASYN**	**XGB_ROS**	**XGB_SMOTE**	**XGB_ADASYN**	**BC_ROS**	**BC_SMOTE**	**BC_ADASYN**
RF_ROS						0,025	0,000			0,000	0,000	0,000
RF_SMOTE						0,047	0,000			0,000	0,000	0,000
RF_ADASYN							0,029			0,000	0,000	0,000
SVM_ROS							0,027			0,000	0,000	0,000
SVM_SMOTE										0,000	0,000	0,000
SVM_ADASYN	0,025	0,047								0,000	0,000	0,000
XGB_ROS	0,000	0,000	0,029	0,027				0,001				0,001
XGB_SMOTE							0,001			0,000	0,000	0,000
XGB_ADASYN	0,046									0,000	0,000	0,000
BC_ROS	0,000	0,000	0,000	0,000	0,000	0,000		0,000	0,000			
BC_SMOTE	0,000	0,000	0,000	0,000	0,000	0,000		0,000	0,000			
BC_ADASYN	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,000	0,000			

Fonte: Autoria própria

A Figura 15 apresenta, para cada combinação de algoritmo e técnica de balanceamento, um boxplot com os dados da medida de acurácia. Conforme já descrito anteriormente, o algoritmo *Bagging Classifier* apresentou distribuição da medida de acurácia estatisticamente diferente dos demais algoritmos, o que pode ser comprovado pela análise da Figura 15. Pela análise de boxplot, observa-se que este algoritmo apresentou desempenho significativamente pior para a medida de acurácia, independentemente da técnica de balanceamento utilizada em conjunto. Para as demais combinações, em geral, os valores de acurácia permaneceram no intervalo entre 0,83 e 0,84 e para algumas combinações foram observados outliers.

A acurácia representa a proporção de predições corretas em relação ao total de predições feitas pelo modelo. Nesse contexto, ao comparar a mediana, nota-se que a combinação do algoritmo *Random Forest* com a técnica de balanceamento *Random Oversampling* foi a que apresentou o melhor desempenho para acurácia e, consequentemente, maior proporção de predições corretas.

Figura 15 – Desempenho por combinação para a medida de acurácia



Fonte: Autoria própria

Ao analisar o desempenho dos algoritmos pela métrica de MCC, observa-se, pela análise da Tabela 7, que o algoritmo *Bagging Classifier* permanece com desempenho estatisticamente diferente de quase todas as demais combinações. Além disso, o algoritmo *Random Forest* também apresentou diferenças estatisticamente significativas em relação ao algoritmo *XGBoost* (aplicado em conjunto com as técnicas ROS e ADASYN).

Tabela 7 – Resultados do teste de Dunn para a medida de *Matthews Correlation Coeficient* (MCC) onde os valores correspondem ao p-valor do teste estatístico entre pares

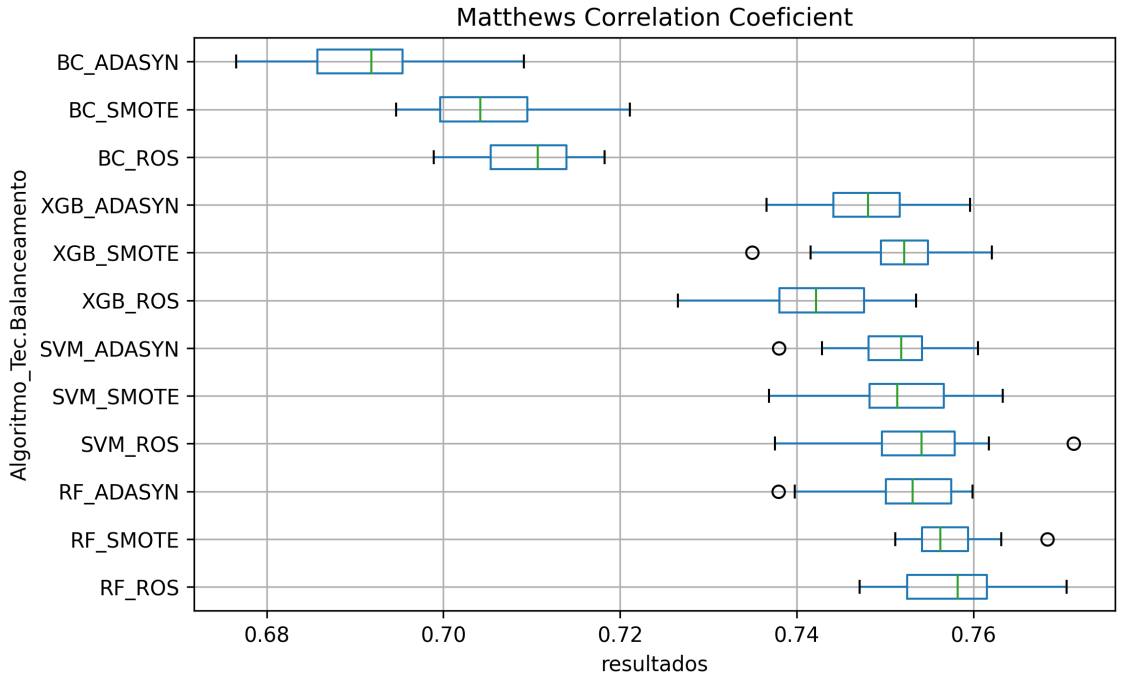
	RF_ROS	**RF_SMOTE**	**RF_ADASYN**	**SVM_ROS**	**SVM_SMOTE**	**SVM_ADASYN**	**XGB_ROS**	**XGB_SMOTE**	**XGB_ADASYN**	**BC_ROS**	**BC_SMOTE**	**BC_ADASYN**
RF_ROS							0,000		0,001	0,000	0,000	0,000
RF_SMOTE							0,000		0,001	0,000	0,000	0,000
RF_ADASYN							0,008			0,000	0,000	0,000
SVM_ROS							0,002			0,000	0,000	0,000
SVM_SMOTE										0,000	0,000	0,000
SVM_ADASYN										0,000	0,000	0,000
XGB_ROS	0,000	0,000	0,008	0,002				0,042				0,001
XGB_SMOTE							0,042			0,000	0,000	0,000
XGB_ADASYN	0,001	0,001								0,004	0,000	0,000
BC_ROS	0,000	0,000	0,000	0,000	0,000	0,000		0,000	0,004			
BC_SMOTE	0,000	0,000	0,000	0,000	0,000	0,000		0,000	0,000			
BC_ADASYN	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,000	0,000			

Fonte: Autoria própria

O coeficiente de correlação de Matthews é uma medida que leva em consideração tanto os verdadeiros quanto os falsos positivos e negativos. Seu valor varia de -1 a 1, onde 1 indica

uma predição perfeita, 0 uma predição aleatória e -1 uma predição inversa. Pela análise da Figura 16, nota-se que também para a métrica de MCC o algoritmo *Bagging Classifier* apresentou os piores resultados, independentemente da técnica de balanceamento aplicada em conjunto. Os demais algoritmos apresentaram valores que variam entre 0,74 e 0,76 para esta métrica. Novamente, o algoritmo *Random Forest* implementado em conjunto com a técnica de *Random Oversampling* apresentou a melhor mediana, indicando o melhor desempenho para a medida de MCC e, consequentemente, um desempenho melhor na classificação.

Figura 16 – Desempenho por combinação para a medida de MCC



Fonte: Autoria própria

Para a medida de *precision*, além do algoritmo *Bagging Classifier* ter apresentado desempenho estatisticamente diferente de todas as demais combinações, nota-se, pela análise da Tabela 8, que também foram observadas diferenças entre os algoritmos *Random Forest* e SVM e o algoritmo *XGBoost* em algumas combinações.

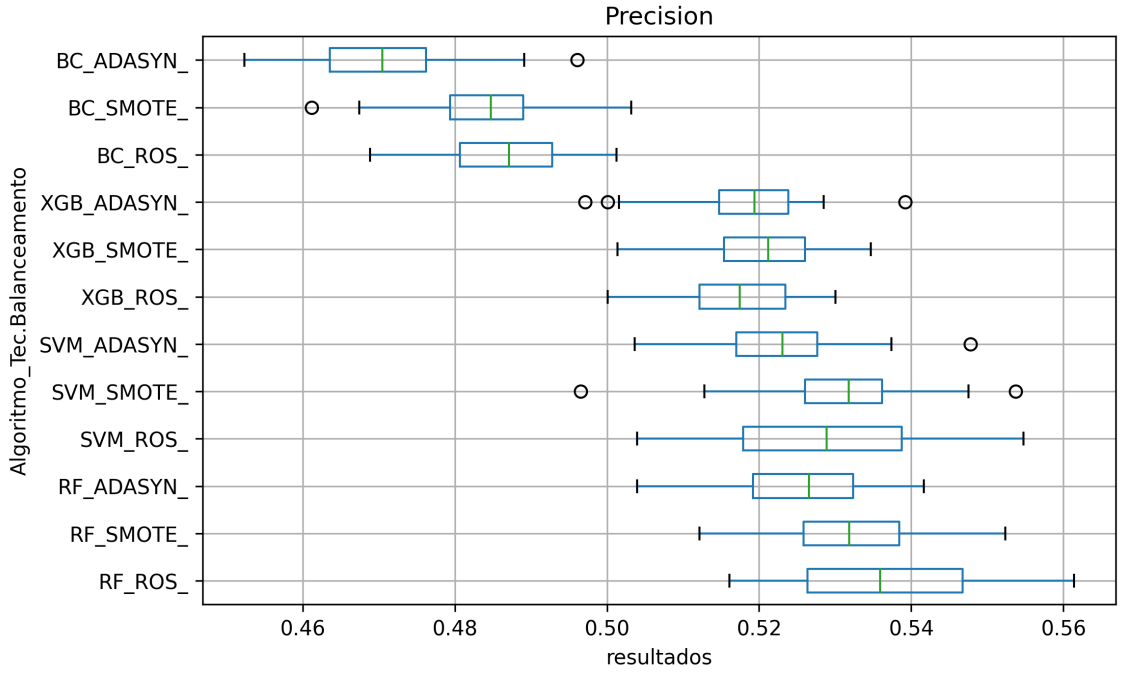
A Figura 17 apresenta a distribuição da macro-*precision* para cada combinação de algoritmo e técnica de hiperparâmetros. Conforme demonstrado, para a métrica de *precision* houve mais variabilidade nos resultados para diferentes combinações do que para as medidas de acurácia e MCC. Para esta métrica, o algoritmo *Bagging Classifier* permaneceu com o pior desempenho. Outro ponto a ser destacado é que a forma dos boxplots, mais alongados, indica menos consistência nos resultados, ou seja, maior variabilidade. Ainda assim, o conjunto do algoritmo *Random Forest* e a técnica de balanceamento *Random Oversampling* apresentou mediana mais alta, assim como foi observado nas métricas de acurácia e MCC, o que indica que esta combinação possui maior proporção de instâncias corretamente classificadas como positivas em relação ao total de instâncias classificadas como positivas.

Tabela 8 – Resultados do teste de Dunn para a medida de *precision* onde os valores correspondem ao p-valor do teste estatístico entre pares

	RF_ROS	**RF_SMOTE**	**RF_ADASYN**	**SVM_ROS**	**SVM_SMOTE**	**SVM_ADASYN**	**XGB_ROS**	**XGB_SMOTE**	**XGB_ADASYN**	**BC_ROS**	**BC_SMOTE**	**BC_ADASYN**
RF_ROS							0,001	0,011	0,001	0,000	0,000	0,000
RF_SMOTE							0,007		0,007	0,000	0,000	0,000
RF_ADASYN										0,000	0,000	0,000
SVM_ROS										0,000	0,000	0,000
SVM_SMOTE							0,028		0,028	0,000	0,000	0,000
SVM_ADASYN										0,000	0,000	0,000
XGB_ROS	0,001	0,007			0,028					0,003	0,001	0,000
XGB_SMOTE	0,011									0,000	0,000	0,000
XGB_ADASYN	0,001	0,007			0,028					0,003	0,001	0,000
BC_ROS	0,000	0,000	0,000	0,000	0,000	0,000	0,003	0,000	0,003			
BC_SMOTE	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,000	0,001			
BC_ADASYN	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000			

Fonte: Autoria própria

Figura 17 – Desempenho por combinação para a medida de *precision*



Fonte: Autoria própria

Para a medida de *recall*, a análise da Tabela 9 indica que, além das diferenças observadas entre o algoritmo *Bagging Classifier* e os demais algoritmos, para esta medida houve diferenças estatisticamente significativas entre os algoritmos *Random Forest* e SVM e o algoritmo XGBoost.

Tabela 9 – Resultados do teste de Dunn para a medida de *recall* onde os valores correspondem ao p-valor do teste estatístico entre pares

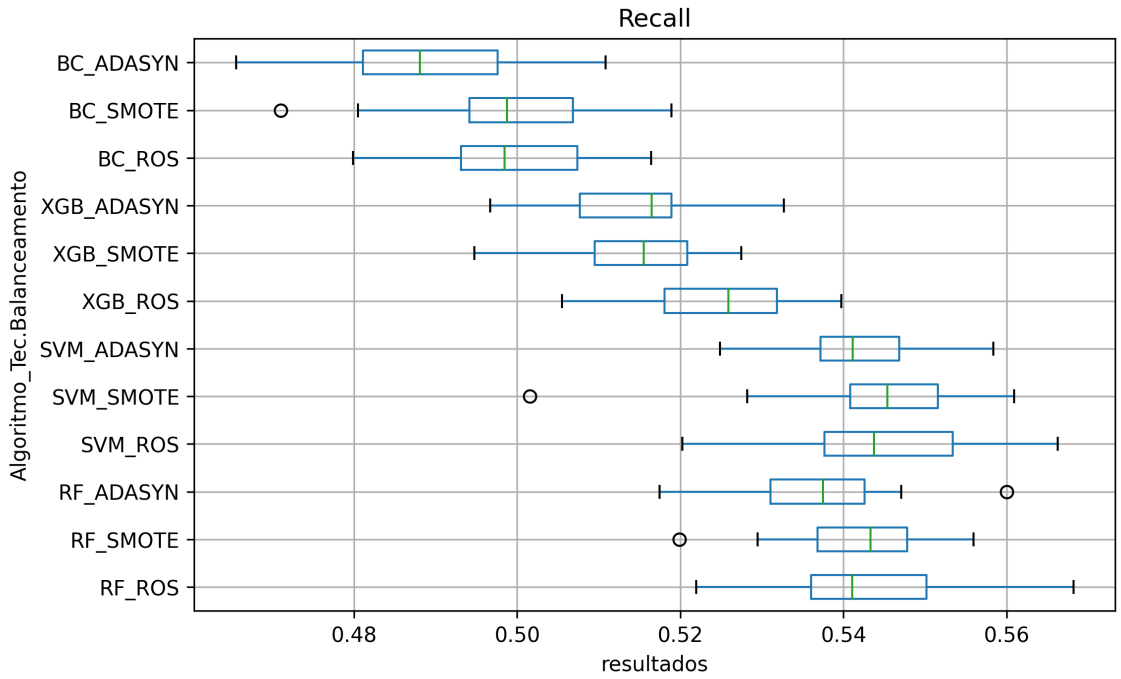
	RF_ROS	**RF_SMOTE**	**RF_ADASYN**	**SVM_ROS**	**SVM_SMOTE**	**SVM_ADASYN**	**XGB_ROS**	**XGB_SMOTE**	**XGB_ADASYN**	**BC_ROS**	**BC_SMOTE**	**BC_ADASYN**
RF_ROS							0,017	0,000	0,000	0,000	0,000	0,000
RF_SMOTE							0,010	0,000	0,000	0,000	0,000	0,000
RF_ADASYN								0,001	0,001	0,000	0,000	0,000
SVM_ROS							0,002	0,000	0,000	0,000	0,000	0,000
SVM_SMOTE							0,001	0,000	0,000	0,000	0,000	0,000
SVM_ADASYN							0,016	0,000	0,000	0,000	0,000	0,000
XGB_ROS	0,017	0,010		0,002	0,001	0,016				0,009	0,001	0,000
XGB_SMOTE	0,000	0,000	0,001	0,000	0,000	0,000						
XGB_ADASYN	0,000	0,000	0,001	0,000	0,000	0,000						
BC_ROS	0,000	0,000	0,000	0,000	0,000	0,000	0,009	0,000	0,003			
BC_SMOTE	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,000	0,001			
BC_ADASYN	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000			

Fonte: Autoria própria

O *recall* mede a proporção de instâncias positivas reais que foram corretamente identificadas, ou seja, é a proporção de verdadeiros positivos (VP) em relação à soma de verdadeiros positivos (VP) e falsos negativos (FN). Em um problema de classificação multiclasse, um falso negativo (FN) ocorre quando o modelo não classifica uma instância como pertencente a uma determinada classe quando ela realmente pertence a esta classe. Assim como no caso da medida de *precision*, a medida *recall* é calculada individualmente para cada classe e, para agrupar todas elas, utiliza-se a média que pode ser macro, micro ou *weighted*.

A Figura 18 apresenta a distribuição da macro-*recall* para cada combinação de algoritmo e técnica de hiperparâmetros. Conforme demonstrado, também foi observada uma variabildiade maior nos resultados, tanto entre as combinações quanto na distribuição dos dados para cada conjunto de algoritmo e técnica de balanceamento, o que pode ser confirmado pelo alongamento dos boxplots. Apesar do algoritmo *Bagging Classifier* ter apresentado o pior desempenho também para a métrica de *recall*, nota-se que para esta medida de avaliação os resultados do *Bagging Classifier* não estão tão distantes dos resultados alcançados pelos demais algoritmos. Para esta métrica, o conjunto de algoritmo e técnica de hiperparâmetros que apresentou maior mediana foi SVM com SMOTE.

Figura 18 – Desempenho por combinação para a medida de *recall*



Fonte: Autoria própria

Por fim, para a medida de F1-*score*, além do algoritmo *Bagging Classifier* ter apresentado desempenho estatisticamente diferente de todas as demais combinações, houve diferenças também entre os os algoritmos *Random Forest* e SVM e *Random Forest* e *XGBoost* em alguns cenários, conforme apresentado na Tabela 10.

Tabela 10 – Resultados do teste de Dunn para a medida de F1-*score* onde os valores correspondem ao p-valor do teste estatístico entre pares

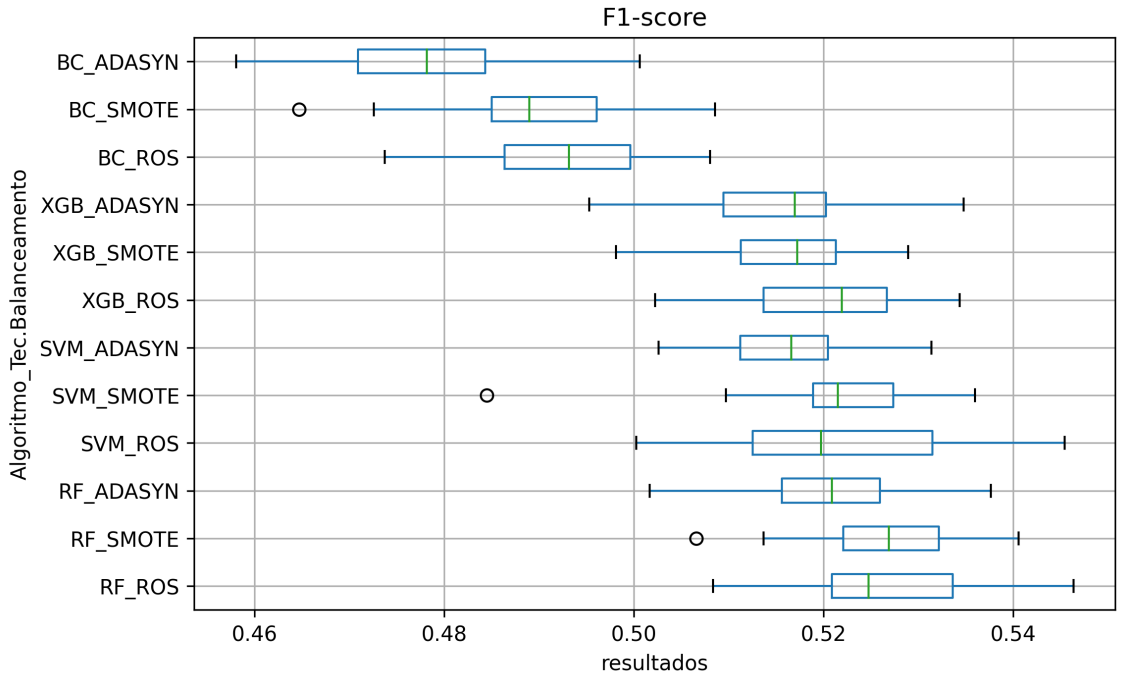
	RF_ROS	**RF_SMOTE**	**RF_ADASYN**	**SVM_ROS**	**SVM_SMOTE**	**SVM_ADASYN**	**XGB_ROS**	**XGB_SMOTE**	**XGB_ADASYN**	**BC_ROS**	**BC_SMOTE**	**BC_ADASYN**
RF_ROS									0,039	0,000	0,000	0,000
RF_SMOTE						0,017			0,011	0,000	0,000	0,000
RF_ADASYN										0,000	0,000	0,000
SVM_ROS										0,000	0,000	0,000
SVM_SMOTE										0,000	0,000	0,000
SVM_ADASYN		0,017								0,001	0,000	0,000
XGB_ROS										0,000	0,000	0,000
XGB_SMOTE										0,000	0,000	0,000
XGB_ADASYN	0,039	0,011								0,001	0,000	0,000
BC_ROS	0,000	0,000	0,000	0,000	0,000	0,001	0,000	0,000	0,001			
BC_SMOTE	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000			
BC_ADASYN	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000			

Fonte: Autoria própria

Os resultados das medidas de *precision* e *recall* combinados, formam a métrica F1-*score*, que também é calculada individualmente por classe e agrupada utilizando a média que pode ser macro, micro ou *weighted*. Em relação à distribuição dos dados, observa-se, pela análise

de boxplot, uma variabilidade maior nos resultados da métrica de F1-*score*, assim como foi observado para as métricas anteriores de *precision* e *recall*. Para esta métrica, o algoritmo *Random Forest* em conjunto com as técnicas de balanceamento SMOTE e ROS apresentaram, respectivamente, as maiores medianas.

Figura 19 – Desempenho por combinação para a medida de F1-*score*



Fonte: Autoria própria

Após analisar o desempenho dos diferentes conjuntos de algoritmos e técnicas de balanceamento de classes sob a ótica das métricas de avaliação de modelos de classificação, observou-se que o algoritmo *Random Forest* em conjunto com a técnica de balanceamento *Random Oversampling* apresentou o maior valor de mediana, indicando que esta combinação foi a melhor para as métricas de acurácia, MCC e *precision*. Para a medida de F1-*score*, embora esta combinação não tenha sido a melhor, ela aparece em segundo lugar com a maior mediana. Diante destes resultados, esta combinação foi selecionada para ser utilizada na implementação dos métodos modelagem com o objetivo de inferir o motivo de viagem dos passageiros do transporte público da RMBH.

Todos os resultados apresentados até o momento basearam-se em um agrupamento das medidas de avaliação utilizando a média macro. No entanto, em um problema de classificação multiclasse, é extremamente relevante analisar as métricas de *precision*, *recall* e F1-*score* individualmente para cada classe, sobretudo quando o resultado agregado está aquém do esperado. Uma ferramenta comumente empregada para analisar o desempenho do modelo, tanto no contexto geral, quanto individualmente para cada classe é o Relatório de Classificação. Após concluir que o modelo treinado utilizando a combinação do algoritmo *Random Forest* e técnica de balanceamento *Random Oversampling* apresentou o melhor desempenho no Método 1, o

Relatório de Classificação deste modelo foi gerado e, os resultados gerais (Tabela 11) detalhados por classe (Tabela 12) estão apresentados.

Tabela 11 – Relatório de Classificação - Geral - Método 1

	precision	***recall***	**F1-*score***	**Acurácia**	**MCC**	**Tamanho Amostral***
Média Macro	53%	54%	52%	84%	78%	4232
Média ponderada (weighted)	84%	84%	84%	-	-	4232

*Número de exemplos reais (amostras) no conjunto de teste.

Fonte: Autoria própria

Tabela 12 – Relatório de Classificação (detalhamento por classe) - Método 1

Classes da variável target	***precision***	***recall***	***F1-score***	**Tamanho Amostral***
Escola	67%	72%	69%	295
Lazer	13%	17%	15%	88
Negócios	26%	43%	33%	137
Outros	39%	16%	23%	200
Residência	96%	98%	97%	1993
Saúde	37%	40%	38%	153
Trabalho	94%	89%	91%	1366

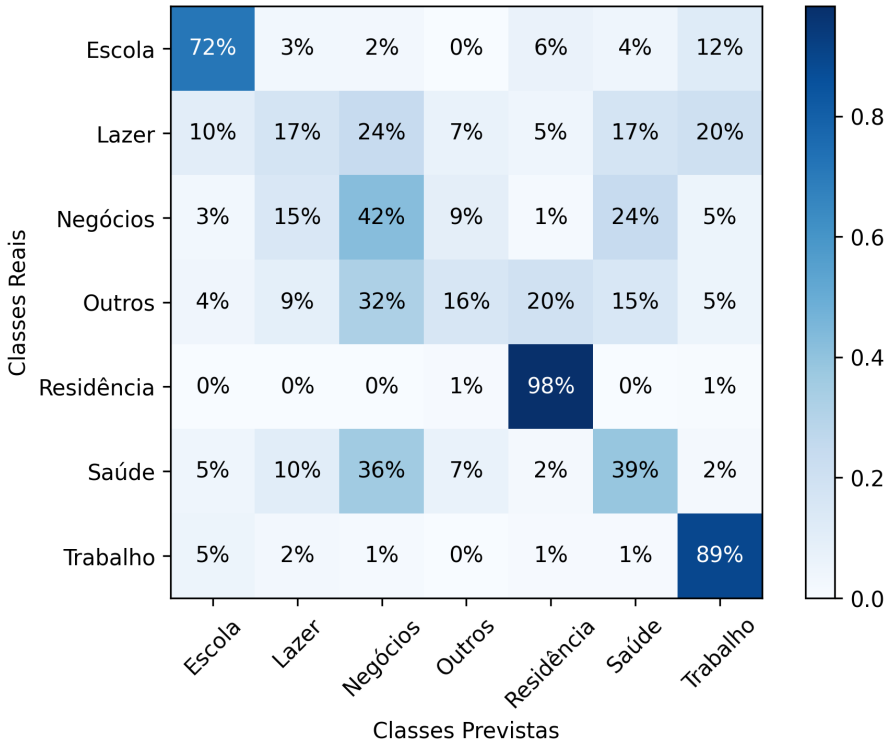
*Número de exemplos reais (amostras) pertencentes a uma classe específica no conjunto de teste.

Fonte: Autoria própria

Inicialmente, nota-se, pela análise da Tabela 11, que há uma diferença substancial entre a *macro average* e a *weighted average* para as medidas de *precision*, *recall* e F1-*score*, isto porque além das classes de motivos primários de “Residência”, “ Trabalho” e “ Escola” terem apresentado os melhores resultados para estas métricas, elas também são as classes que possuem o maior volume de dados (tamanho amostral), conforme apresentado na Tabela 12. Nesse contexto, como a *weighted average* leva em consideração o número de amostras em cada classe como um peso para a média ponderada, seu resultado é significativamente melhor do que o observado para a *macro average*, que consiste em uma média simples. Como neste problema de classificação todas as classes possuem igual importância, optou-se por utilizar a macro-*average* das medidas de *precision*, *recall* e F1-*score* como métrica de avaliação. De forma geral, observa-se, pela análise da Tabela 11, que o modelo apresentou acurácia de 84% e MCC de 78%, que podem ser considerados altos. A macro-*recall* indica que o modelo capturou corretamente 54% das amostras positivas.

Os resultados do desempenho do modelo detalhado por classe (Tabela 12) demonstram que o modelo possui um bom desempenho ao prever os motivos primários de viagem, sobretudo os motivos de “Residência” e “Trabalho”. No entanto, nota-se uma queda significativa no desempenho deste modelo ao prever os motivos de viagem secundários, em especial o motivo de “Lazer”, que apresentou os piores resultados. Este padrão também está apresentado graficamente na Figura 20, que apresenta a matriz de confusão do modelo treinado no Método 1. A matriz de confusão é uma tabela que mostra o desempenho do modelo de classificação, detalhando as previsões corretas e incorretas para cada classe.

Figura 20 – Matriz de Confusão - Método 1



Fonte: Autoria própria

O algoritmo *Random Forest*, selecionado como mais adequado ao problema que está sendo abordado nesta pesquisa, é um algoritmo de aprendizado de máquina capaz de fornecer uma medida de importância das variáveis ao final do processo de treinamento. A importância das variáveis em um modelo de *Random Forest* é uma métrica que indica o quanto cada variável contribui para a capacidade do modelo de fazer previsões precisas. Isso é fundamental em muitos cenários de análise de dados e modelagem preditiva.

Para o modelo do Método 1, a variável "Tempo_Entre_Embarques" foi a que mais se destacou, com 47% de importância, seguido da variável "Horario_Inicio" com 25%, "ID_Viagem" com 15%, "Tempo_Deslocamento" com 10% e "Qtd_Viagens_Individuo" com apenas 2% da importância.

6.2 Método 2

Conforme descrito na Subseção 5.4.2, este método utiliza a combinação de algoritmo e técnica de balanceamento de classes selecionado no Método 1 e o mesmo conjunto de variáveis preditoras para treinar dois modelos distintos e sequenciais que diferem entre si pelas classes da variável *target*.

No primeiro modelo, a variável *target* foi composta por apenas quatro classes, sendo elas: "Residência", "Trabalho", "Escola" e "Secundário". Esta última classe foi utilizada para

agregar todos os motivos secundários de “Saúde”, “Lazer”, “Negócios” e “Outros” em uma única categoria. Este modelo foi denominado “Modelo de Viagens Primárias”.

No segundo modelo, foram utilizados apenas os registros de viagens secundárias para o treinamento. Para isto, o *dataset* de treinamento foi filtrado para conter apenas os registros destas viagens e, consequentemente, a variável *target* deste modelo foi composta pelas classes de “Lazer”, “Negócios”, “Saúde” e “Outros”. O objetivo deste modelo foi reclassificar as viagens classificadas pelo Modelo 1 como “Secundárias” em seus respectivos propósitos secundários de viagem. Este modelo foi denominado “Modelo de Viagens Secundárias”.

O Modelo de Viagens Primárias apresentou resultados muito bons, com acurácia de 92% , MCC de 87%, macro-*precision* de 86%, macro-*recall* de 84% e macro-F1-*score* de 85%, conforme apresentado na Tabela 13. As métricas de avaliação deste modelo por classe estão apresentadas na Tabela 14 e, ao comparar estes resultados com os resultados da Tabela 12 obtidos pelo modelo único treinado no Método 1, nota-se que não há diferenças significativas nos resultados obtidos por ambos modelos na classificação das viagens de “Residência”, “Trabalho” e “Escola”.

O principal ganho trazido pelo Modelo de Viagens Primárias do Método 2 em relação ao modelo único treinado no Método 1 é a agregação das viagens secundárias em uma única classe “Secundário” que apresentou métricas de *precision*, *recall* e F1-*score* muito superiores se comparado aos resultados obtidos de forma desagregada para cada uma das classes de “Lazer”, “Negócios”, “Saúde” e “Outros” no Método 1. A Tabela 15 traz os resultados da métrica de F1-*score* destes dois modelos para facilitar a comparação do desempenho de cada um deles.

Tabela 13 – Relatório de Classificação - Geral - Modelo Viagens Primárias - Método 2

	precision	***recall***	**F1-*score***	**Acurácia**	**MCC**	**Tamanho Amostral***
Média Macro	86%	84%	85%	92%	87%	4232
Média ponderada (weighted)	91%	91%	91%	-	-	4232

*Número de exemplos reais (amostras) no conjunto de teste.

Fonte: Autoria própria

Tabela 14 – Relatório de Classificação (detalhamento por classe) - Modelo Viagens Primárias - Método 2

Classes da variável target	***precision***	***recall***	***F1-score***	**Tamanho Amostral***
Escola	65%	73%	69%	295
Residência	96%	99%	98%	1993
Secundário	85%	78%	82%	578
Trabalho	94%	91%	92%	1366

*Número de exemplos reais (amostras) pertencentes a uma classe específica no conjunto de teste.

Fonte: Autoria própria

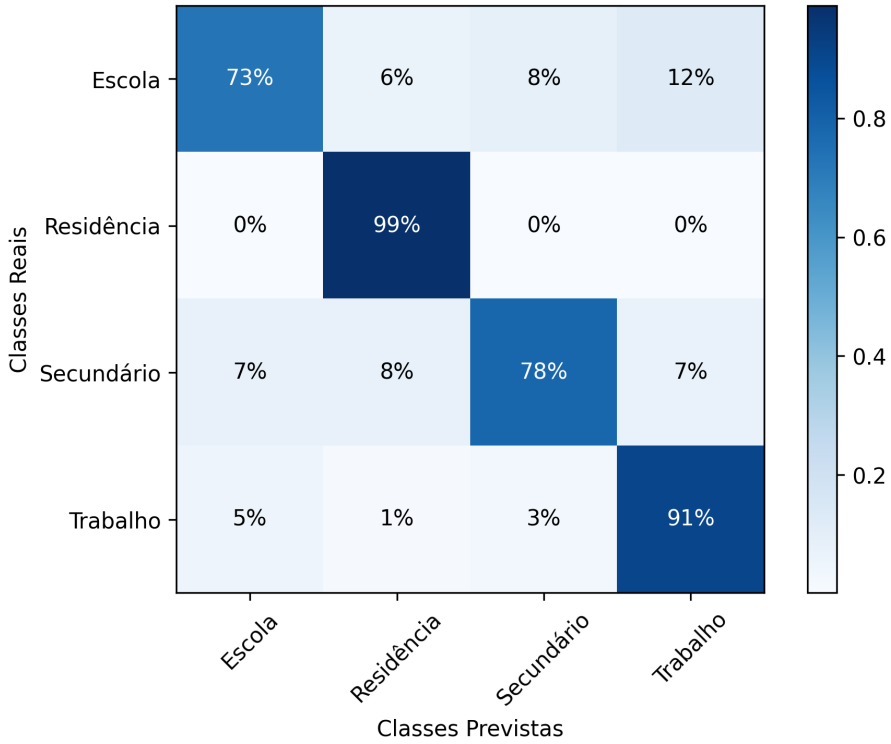
Tabela 15 – Comparação dos resultados da métrica F1-*score* entre o modelo treinado no Método 1 e o Modelo de Viagens Primárias treinado no Método 2

Classes Método 1	**Classes Modelo Viagens Secundárias Método 2**	**F1-*score***	
		Modelo único - Método 1 Tabela 12	**Modelo Viagens Primárias - Método 2 Tabela 14**
Residência	**Residência**	97%	98%
Trabalho	**Trabalho**	91%	92%
Escola	**Escola**	69%	69%
Lazer	**Secundário**	15%	82%
Negócios		33%	
Outros		23%	
Saúde		38%	

Fonte: Autoria própria

Conforme apresentado na Figura 21, a principal confusão do Modelo de Viagens Primárias do Método 2 foi classificar viagens cujo motivo real era "Escola" como "Trabalho" (12%) ou como "Secundário"(8%). Isso justifica desempenho mais baixo do modelo na classificação de viagens escolares, conforme apresentado na Tabela 14.

Figura 21 – Matriz de Confusão Modelo viagens primárias - Método 2



Fonte: Autoria própria

Para este modelo, a variável "Tempo_Entre_Embarques", foi consideravelmente mais importante (46%) do que as demais, sendo duas vezes mais importante do que a variável de "Horario_Inicio" (21%), que aparece em segundo lugar. Conforme já descrito anteriormente na Subseção 2.3.1, as viagens primárias ou obrigatórias são viagens que ocorrem regularmente e que possuem um comportamento temporal bem definido, o que justifica estas variáveis serem as mais importantes no modelo.

Em relação ao modelo treinado para classificar as viagens secundárias no Método 2, apesar do resultado geral permanecer ruim, com acurácia de 40%, MCC de 21%, macro-*precision* de 43%, macro-*recall* de 42% e macro-F1-*score* de 40%, conforme apresentado na Tabela 16, o desempenho individual por classe apresentou melhorias em relação ao que foi observado utilizando um modelo único de classificação no Método 1. Isto pode ser observado ao comparar os resultados do modelo do Método 1 para as classes de "Lazer", "Saúde", "Negócios" e "Outros" apresentados na Tabela 12 e os resultados do Modelo de Viagens Secundárias do Método 2 mostrados na Tabela 17. A Tabela 18 traz os resultados da métrica de F1-*score* destes dois modelos para facilitar a comparação do desempenho de cada um deles na classificação dos motivos secundários de viagem.

Tabela 16 – Relatório de Classificação - Geral - Modelo Viagens Secundárias - Método 2

	precision	***recall***	**F1-*score***	**Acurácia**	**MCC**	**Tamanho Amostral***
Média Macro	43%	42%	40%	40%	21%	578
Média ponderada (weighted)	45%	40%	40%	-	-	578

*Número de exemplos reais (amostras) no conjunto de teste.

Fonte: Autoria própria

Tabela 17 – Relatório de Classificação (detalhamento por classe) - Modelo Viagens Secundárias - Método 2

Classes da variável target	***precision***	***recall***	***F1-score***	**Tamanho Amostral***
Lazer	39%	47%	43%	88
Negócios	29%	46%	36%	137
Outros	60%	29%	39%	200
Saúde	43%	44%	43%	153

*Número de exemplos reais (amostras) pertencentes a uma classe específica no conjunto de teste.

Fonte: Autoria própria

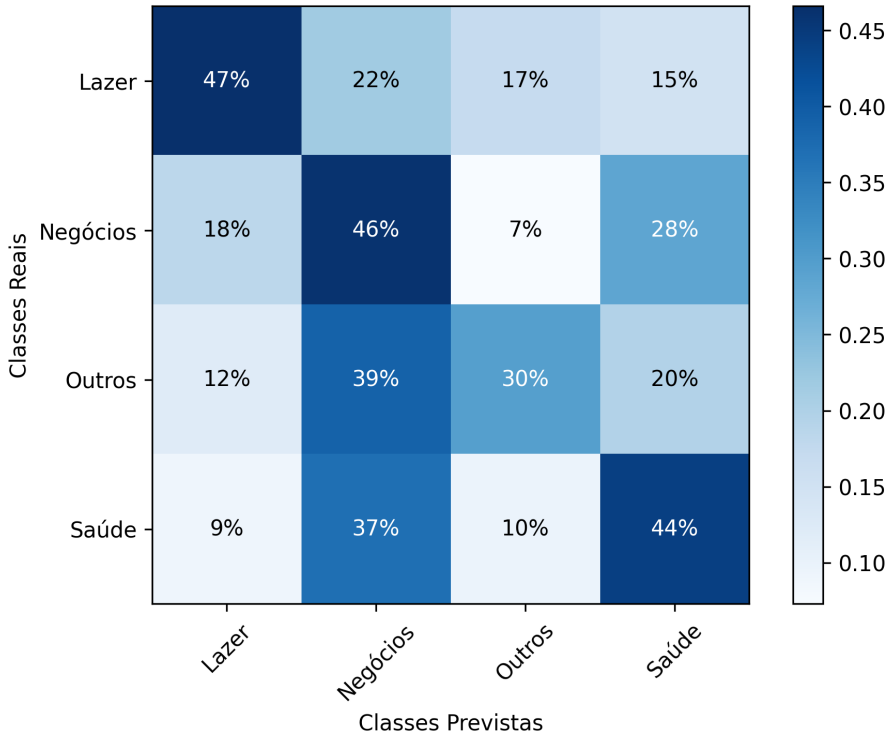
Tabela 18 – Comparação dos resultados da métrica F1-*score* para as classes de "Lazer", "Negócios", "Outros" e "Saúde" entre o modelo treinado no Método 1 e o Modelo de Viagens Secundárias treinado no Método 2

	F1-*score*	
	Modelo único - Método 1 Tabela 12	**Modelo Viagens Secundárias - Método 2 Tabela 17**
Lazer	15%	28%
Negócios	33%	36%
Outros	23%	39%
Saúde	38%	43%

Fonte: Autoria própria

Conforme apresentado na Figura 22, o modelo de classificação de viagens secundárias comete muitos erros para todas as classes em função de sua imprecisão. No entanto, nota-se que a classe "Lazer" foi a que apresentou o melhor desempenho, tendo 47% de acertos.

Figura 22 – Matriz de Confusão Modelo viagens secundárias - Método 2



Fonte: Autoria própria

Em relação à importância das variáveis para o Modelo de Viagens Secundárias, as variáveis de "Horário_Inicio" e de "Tempo_Entre_Embarques" foram as mais importantes, representando, cada uma delas, 34% e 33% da importância, respectivamente. A variável "Tempo_Deslocamento" aparece com 22% e as variáveis "Qtd_Viagens_Individuo" e "ID_Viagem" possuem magnitude de 4% e 5% na importância, respectivamente. Uma diferença relevante deste modelo em relação ao que foi observado no Modelo de Viagens Primárias é que a segunda e terceira variável na ordem de importância são tão relevantes quanto a primeira.

6.3 Método 3

Com o objetivo de tornar as classificações dos motivos de viagem ainda mais precisas, sobretudo para os motivos secundários, variáveis espaciais foram incluídas no conjunto de preditores dos dois modelos treinados no Método 2, sendo esta a única diferença entre o Método 3 e o Método 2. Sabe-se que as viagens secundárias possuem um padrão temporal pouco definido e, por isso, espera-se que as variáveis espaciais tragam ganhos especialmente para este modelo.

As variáveis espaciais consideradas neste estudo foram a quantidade de estabelecimentos de lazer, saúde, negócios e ensino disponíveis na AH de destino do deslocamento. Nomeadamente, utilizou-se as variáveis "Estabs_Lazer", "Estabs_Saude", "Estabs_Negocios" e "Estabs_Ensino"

Ao incluir as variáveis espaciais no conjunto de preditoras, o Modelo de Viagens Primárias do Método 3 apresentou acurácia de 92%, MCC de 87%, macro-*precision* de 85%, macro-*recall* de

86% e macro-F1-*score* de 85%, conforme apresentado na Tabela 19. Em relação ao desempenho individual por classe, apresentado na Tabela 20, o Modelo de Viagens Primárias do Método 3 não apresentou diferenças significativas em relação aos resultados obtidos pelo Modelo de Viagens Primárias do Método 2, conforme pode ser observado pela comparação dos resultados da métrica de F1-*score* apresentado na Tabela 21.

Tabela 19 – Relatório de Classificação - Geral - Modelo Viagens Primárias - Método 3

	precision	***recall***	**F1-*score***	**Acurácia**	**MCC**	**Tamanho Amostral***
Média Macro	85%	86%	85%	92%	87%	4232
Média ponderada (weighted)	92%	92%	92%	-	-	4232

*Número de exemplos reais (amostras) no conjunto de teste.

Fonte: Autoria própria

Tabela 20 – Relatório de Classificação (detalhamento por classe) - Modelo Viagens Primárias - Método 3

Classes da variável target	***precision***	***recall***	***F1-score***	**Tamanho Amostral***
Escola	63%	75%	69%	295
Residência	96%	99%	98%	1993
Secundário	84%	78%	81%	578
Trabalho	94%	90%	92%	1366

*Número de exemplos reais (amostras) pertencentes a uma classe específica no conjunto de teste.

Fonte: Autoria própria

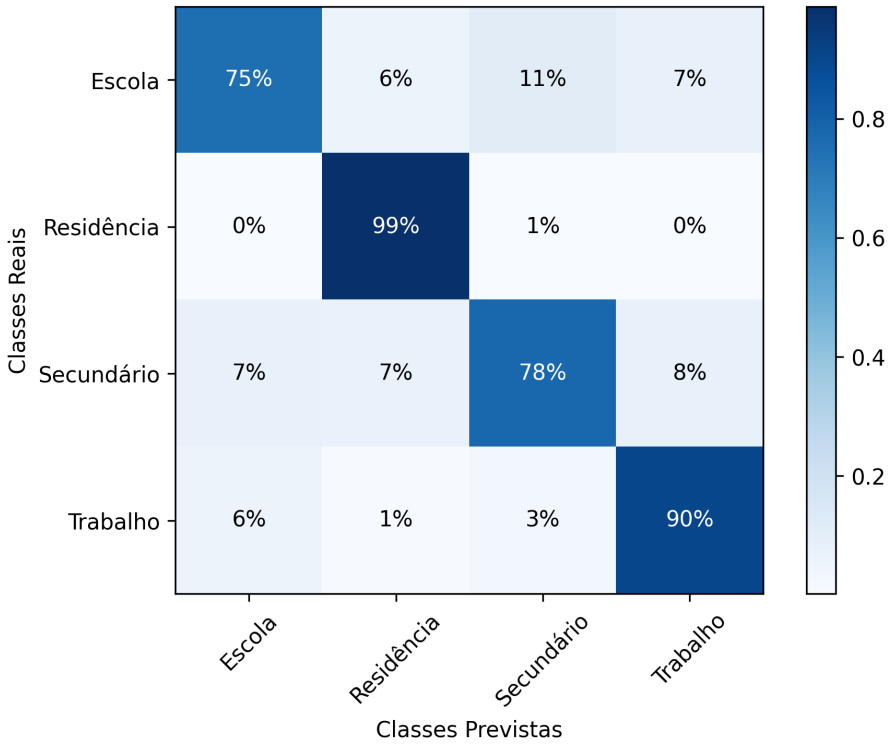
Tabela 21 – Comparação dos resultados da métrica F1-*score* entre o modelo treinado no Método 1, o Modelo de Viagens Primárias treinado no Método 2 e o Modelo de Viagens Primárias treinado no Método 3

Classes Método 1	**Classes Modelo Viagens Secundárias Método 2**	**F1-*score***		
		Modelo único Método 1 Tabela 12	**Modelo Viagens Primárias Método 2 Tabela 14**	**Modelo Viagens Primárias Método 3 Tabela 20**
Residência	**Residência**	97%	98%	98%
Trabalho	**Trabalho**	91%	92%	92%
Escola	**Escola**	69%	69%	69%
Lazer	**Secundário**	15%	82%	81%
Negócios		33%		
Outros		23%		
Saúde		38%		

Fonte: Autoria própria

Conforme apresentado na Figura 23, a inclusão de variáveis espaciais no conjunto de preditores contribuiu para a diminuição da confusão do modelo entre os motivos escola e trabalho que passou de 12% para 7%. No entanto, a confusão de viagens escolares com viagens por motivos secundários aumentou de 8% para 11% e devido a estas compensações, não houve melhorias no resultado geral do modelo.

Figura 23 – Matriz de Confusão Modelo viagens primárias - Método 3



Fonte: Autoria própria

Em relação à importância das variáveis, a variável "Tempo_Entre_Embarques" foi a mais importante (41%), seguido da variável "ID_Viagem" (com 23%) e da variável "Horario_Inicio" (com 17%). As demais variáveis apresentaram pouca importância, com menos de 5% cada.

A contribuição da inclusão de variáveis espaciais no conjunto de preditores foi mais expressiva para o Modelo de Viagens Secundárias do que para o Modelo de Viagens Primárias. A acurácia do Modelo de Viagens Secundárias passou de 40% para 47%, o MCC foi de 21% para 31%, macro-*precision* de 43% para 50%, macro-*recall* de 42% para 49% e macro-F1-*score* de 40% para 48%. Estes percentuais vieram da comparação entre o desempenho geral do Modelo Secundário de Viagens do Método 2, apresentado na Tabela 16, e o desempenho geral do Modelo Secundário de Viagens do Método 3, apresentado na Tabela 22.

Ao comparar o desempenho por classe do Modelo de Viagens Secundárias do Método 2, apresentado na Tabela 17, com o desempenho por classe do Modelo de Viagens Secundárias do Método 3, apresentado na Tabela 23, nota-se melhorias significativas nas métricas de avaliação, sobretudo para as classes "Lazer", "Negócios" e "Outros". Uma comparação dos resultados da métrica F1-*score* para as classes de "Lazer", "Negócios", "Outros" e "Saúde" entre o modelo treinado no Método 1 e os Modelos de Viagens Secundárias treinados nos métodos 2 e 3 está apresentada na Tabela 24.

Tabela 22 – Relatório de Classificação - Geral - Modelo Viagens Secundárias - Método 3

	precision	***recall***	**F1-*score***	**Acurácia**	**MCC**	**Tamanho Amostral***
Média Macro	50%	49	48%	47%	31%	578
Média ponderada (weighted)	52%	47%	48%	-	-	578

*Número de exemplos reais (amostras) no conjunto de teste.

Fonte: Autoria própria

Tabela 23 – Relatório de Classificação (detalhamento por classe) - Modelo Viagens Secundárias - Método 3

Classes da variável target	***precision***	***recall***	***F1-score***	**Tamanho Amostral***
Lazer	41%	52%	46%	88
Negócios	35%	58%	44%	137
Outros	62%	39%	48%	200
Saúde	60%	47%	52%	153

*Número de exemplos reais (amostras) pertencentes a uma classe específica no conjunto de teste.

Fonte: Autoria própria

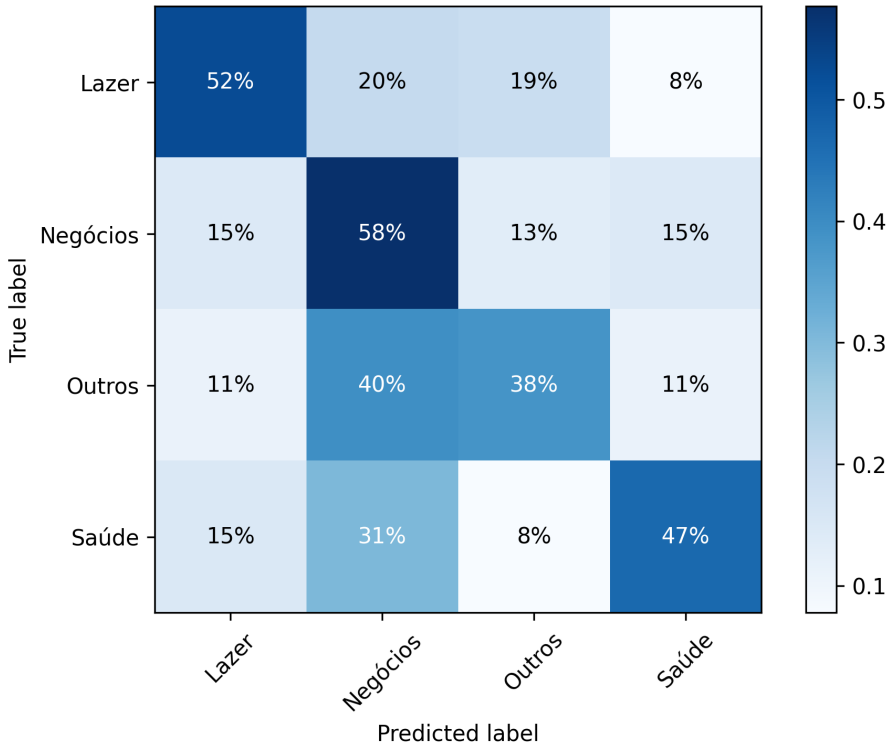
Tabela 24 – Comparação dos resultados da métrica F1-*score* para as classes de “Lazer”, “Negócios”, “Outros” e “Saúde” entre o modelo treinado no Método 1, o Modelo de Viagens Secundárias treinado no Método 2 e o Modelo de Viagens Secundárias treinado no Método 3

	F1-*score*		
	Modelo único - Método 1 Tabela 12	**Modelo Viagens Secundárias - Método 2** Tabela 17	**Modelo Viagens Secundárias - Método 3** Tabela 23
Lazer	15%	28%	46%
Negócios	33%	36%	44%
Outros	23%	39%	48%
Saúde	38%	43%	52%

Fonte: Autoria própria

Apesar do Modelo Viagens Secundárias ter apresentado melhoria com a inclusão de variáveis espaciais, ele ainda comete muitos erros, conforme apresentado na Figura 24, sendo que as principais confusões do modelo estão relacionadas à classificação de algumas viagens como “Negócios”, quando, elas pertenciam a outra classe.

Figura 24 – Matriz de Confusão Modelo Viagens Secundárias - Método 3



Fonte: Autoria própria

Em relação à importância das variáveis, as variáveis "Tempo_Entre_Embarques" e "Horario_Inicio" permaneceram como as mais importantes. Entretanto, a importância destas variáveis caiu de aproximadamente 34% (valor obtido no Método 2) para aproximadamente 20% (valor obtido no Método 3), pois, ao incluir as variáveis espaciais no modelo, parte da importância foi redirecionada para estas novas variáveis. Dentre as variáveis espaciais, a quantidade de estabelecimentos de saúde ("Estabs_Saude") foi a que mais contribuiu, com 15% da importância. A quantidade de estabelecimentos de negócios ("Estabs_Negocios") aparece com 10% e a quantidade de estabelecimentos de lazer ("Estabs_Lazer") e ensino ("Estabs_Ensino") aparecem com 8% e 9%, respectivamente.

6.4 Discussão dos Resultados

Conforme apresentado neste capítulo, diferentes algoritmos, técnicas de balanceamento de classes, hiperparâmetros e métodos de modelagem foram testados neste estudo. Em relação ao algoritmo utilizado, o *Random Forest* foi o que apresentou a melhor performance, junto com a técnica de balanceamento *Random Oversampling*. Este resultado está condizente com a literatura, em especial com o estudo de Kim, Kim e Kim (2021) que também comparou o desempenho de diversos modelos para classificar os motivos de viagem nas categorias de "Trabalho", "Negócios", "Lazer" e "Retorno para casa" e observou que o *Random Forest* apresentou o melhor desempenho.

Em relação à construção do modelo, observou-se que a divisão do processo de classificação em duas etapas, como foi feito nos métodos 2 e 3, trouxe melhorias significativas aos resultados. No Modelo de Viagens Primárias, o algoritmo *Random Forest* foi aplicado para classificar os registros de viagem nos motivos primários de "Residência", "Trabalho", "Escola" e "Secundários", sendo esta última categoria utilizada para agrupar as viagens não obrigatórias. Já o Modelo de Viagens Secundárias foi utilizado para reclassificar apenas os registros que foram classificados como "Secundários" pelo Modelo de Viagens Primárias, detalhando-os nas classes de "Saúde", "Lazer", "Negócios" e "Outros". Esta técnica de divisão do processo de classificação em duas etapas também foi aplicada por Pezoa et al. (2023) que utilizou, ao todo, três modelos, sendo o primeiro deles utilizado para discretizar as viagens em motivos primários e secundários, o segundo para classificar os motivos primários e o terceiro para classificar os motivos secundários.

Assim como foi observado em diversos estudos, dentre os quais se pode citar Alsger et al. (2018) e Aslam et al. (2021), o desempenho do modelo na classificação de motivos primários de viagem foi significativamente melhor em comparação com os resultados obtidos na classificação de motivos secundários, e a utilização de variáveis espaciais, como a quantidade de pontos de interesse por tipo de atividade no local de destino da viagem, contribuiu para melhorar significativamente os resultados da classificação de motivos secundários.

Em relação às variáveis espaciais utilizadas no modelo, destaca-se que, em seu estudo, Pezoa et al. (2023) também utilizou dados advindos da Receita Federal para obter informações de uso da terra. Os resultados obtidos por Pezoa et al. (2023) demonstraram-se satisfatórios tanto na classificação de motivos primários, com pontuação F1-*score* de 86%, quanto na classificação de motivos secundários, com pontuação F1-*score* de 72%. Diante disso, dois pontos devem ser ressaltados: o primeiro deles é que no presente estudo a informação da localização exata do ponto de desembarque do passageiro não está disponível e, por consequência, não é possível calcular o número de pontos de interesse ou o percentual de uso da terra por tipo de atividade em um raio próximo do local de desembarque dos passageiros, como foi feito por Pezoa et al. (2023). As informações disponibilizadas pela pesquisa OD_2012 possuem, como menor unidade espacial, as Áreas Homogêneas (AH) e, por isto, é possível saber apenas em qual AH ocorreram os embarques e desembarques e não a localização exata dos mesmos.

Outro ponto a ser destacado, em relação à distribuição espacial dos pontos de interesse, é que a RMBH possui uma característica marcante de concentração de atividades no município sede de Belo Horizonte, que atrai investimentos, empresas e instituições, tornando-se um polo econômico significativo. A concentração de atividades em Belo Horizonte pode ser atribuída a diversos fatores como sua localização geográfica central, sua infraestrutura desenvolvida, a oferta de mão de obra qualificada, a qualidade de vida e a facilidade de acesso a serviços e recursos. Dada a importância das variáveis espaciais na classificação correta de motivos de viagem secundários, a característica monocêntrica da RMBH introduz mais um desafio à identificação precisa destes motivos de viagem, pois os pontos de interesse tendem a estar concentrados no mesmo local, independentemente da atividade econômica do estabelecimento, conforme demonstrado na

Figura 10.

O baixo volume amostral de viagens secundárias e/ou a baixa utilização do modo de transporte coletivo para alguns motivos de viagem, conforme apresentado na Tabela 2, também são desafios importantes que precisam ser superados para garantir melhores resultados para os modelos de classificação apresentados neste estudo, porque este cenário interferiu diretamente na capacidade do modelo de identificar de forma precisa os padrões de viagem destes motivos, acarretando em uma situação de *underfitting*[1]. Diante disso, ressalta-se a importância da coleta de dados representativos, relevantes e de alta qualidade sobre todos os motivos de viagem, sobretudo dos motivos secundários.

O desempenho de um modelo de aprendizado de máquina está diretamente relacionado à qualidade dos dados com os quais ele foi treinado, ou seja, não importa quão complexo ou avançado seja o algoritmo utilizado se os dados de treinamento não representam adequadamente a diversidade dos motivos de viagem ou se uma classe for super-representada em detrimento de outras, como foi observado neste estudo. Ainda que técnicas de pré-processamento como o balanceamento de classes tenham sido aplicadas, as restrições impostas pela qualidade dos dados de treinamento sobressaíram-se interferindo na qualidade do Modelo de Viagens Secundárias treinados nos métodos 2 e 3.

6.5 Resumo do capítulo

Após a implementação dos três métodos e a avaliação dos modelos treinados, concluiu-se que o Método 3 foi o que apresentou os melhores resultados. De forma resumida, as principais conclusões foram:

- Das 12 possíveis combinações de algoritmo e técnica de balanceamento, a combinação *Random Forest* com *Random Oversampling* foi a que apresentou o melhor desempenho.
- A utilização de dois modelos, um para classificar as viagens em primárias e outro para reclassificar as viagens secundárias, apresentou um resultado melhor do que a utilização de um modelo único de classificação. No entanto, apesar da utilização de dois modelos ter apresentado resultados melhores, observou-se que o modelo de classificação de viagens primárias foi o único que apresentou resultados bons o suficiente para viabilizar sua implementação. O desempenho do modelo de viagens secundárias foi comprometido, sobretudo pela qualidade dos dados utilizados no treinamento, já que os dados da pesquisa domiciliar de viagens correspondem a um dia útil típico, onde os prevalescem os deslocamentos primários.
- A utilização de dados de localização de Pontos de Interesse junto aos dados da pesquisa domiciliar trouxe melhorias na classificação do motivo de viagem, sobretudo para as viagens

[1] *Underfitting* ocorre quando o modelo não consegue aprender o suficiente sobre os dados e não consegue encontrar padrões, o que leva-o a ter um desempenho ruim

secundárias. Por este motivo, após analisar os três métodos apresentados, concluiu-se que o Método 3 foi o que apresentou os melhores resultados.

7 Implementação do modelo

A última etapa do processo de mineração de dados é a implantação do modelo escolhido, ou seja, do modelo que apresentou os melhores resultados na etapa de avaliação, em sistemas de produção para fazer previsões em tempo real, conforme apresentado na Figura 1. Nesse contexto, este capítulo apresenta a implantação do melhor modelo de inferência dos motivos de viagem em uma matriz origem e destino estimada com dados de cartões inteligentes com o objetivo de enriquecê-la com um atributo inicialmente não coletado.

7.1 Inferindo o motivo de viagem na matriz OD estimada a partir de dados de cartões inteligentes

Consoante ao que foi detalhadamente descrito no Capítulo 6, o Método 3 composto por dois modelos de classificação, um para classificar as viagens em motivos primários (denominado Modelo de Viagens Primárias) e um para classificar as viagens em motivos secundárias (denominado Modelo de Viagens Secundárias), ambos treinados utilizando o algoritmo *Random Forest*, a técnica de balanceamento *Random Oversampling* e um conjunto de variáveis preditoras composto por variáveis de viagem e variáveis espaciais, foi o que apresentou os melhores resultados.

Apesar do Método 3 ser composto por dois modelos, apenas o Modelo de Viagens Primárias apresentou desempenho satisfatório para seguir para a etapa de implantação, isto porque, o Modelo de Viagens Secundárias apresentou valores muito baixos para as métricas de acurácia, MCC, *precision*, *recall* e F1-*score*. Por este motivo, nesta etapa será apresentada a utilização apenas do Modelo de Viagens Primárias para prever o motivo de viagem dos passageiros em uma matriz de viagens estimada a partir de dados de cartões inteligentes.

Para que a etapa de implantação do modelo pudesse ser viabilizada neste estudo, uma matriz origem e destino estimada a partir de dados de cartões inteligentes (denominada Matriz OD de bilhetagem eletrônica) foi disponibilizada pela Secretaria de Estado de Infraestrutura e Mobilidade - SEINFRA, por meio da Empresa de Consultoria e Engenharia SYSTRA Brasil.

A Matriz OD de bilhetagem eletrônica foi feita utilizando dados de validação de cartões inteligentes no transporte público em um dia típico de novembro de 2019 nos sistemas metropolitano e municipais de Belo Horizonte, Betim, Contagem, Santa Luzia e Ibirité. Esta base de dados, contempla as seguintes variáveis: local de origem e destino da viagem, código de identificação do usuário, identificação sequencial da viagem, horário de início e fim da viagem e tempo de deslocamento. Os mesmos procedimentos aplicados à base de dados OD_2012 na etapa de pré-processamento e transformação, detalhadamente descritos na Seção 5.3, foram executados na base de dados da Matriz OD de bilhetagem eletrônica, isto porque uma das premissas da

aplicação de um modelo de aprendizado de máquina em uma base de dados desconhecida é que esta base tenha o mesmo formato da base de dados utilizada para treinar o modelo.

A primeira análise feita após a aplicação do Modelo de Viagens Primárias na Matriz OD de bilhetagem eletrônica foi a análise da proporção de viagens por motivo. Os resultados e a comparação com o observado na matriz OD_2012 estão apresentados na Tabela 25 e, pela análise da tabela, nota-se que a proporção de viagens para cada motivo nas duas bases de dados é muito semelhante.

Tabela 25 – Comparação do percentual de viagens por motivo entre a matriz OD_2012 e a matriz OD de bilhetagem eletrônica

	Fonte de Dados	
Motivo de Viagem	**Pesquisa Domiciliar (OD_2012)**	**Base de dados de cartões inteligentes**
Escola	8,6%	4,4%
Residência	33,7%	34,2%
Trabalho	40,9%	46,8%
Motivos Secundários	16,6%	14,5%

Fonte: Autoria própria

Conforme descrito na Capítulo 6, as variáveis de horário de início da viagem e de tempo entre embarques, foram as mais importantes em todos os modelos, mas sobretudo no Modelo de Viagens Primárias. Por este motivo, optou-se por comparar o comportamento da distribuição do horário de início da viagem e do tempo entre embarques das duas matrizes OD.

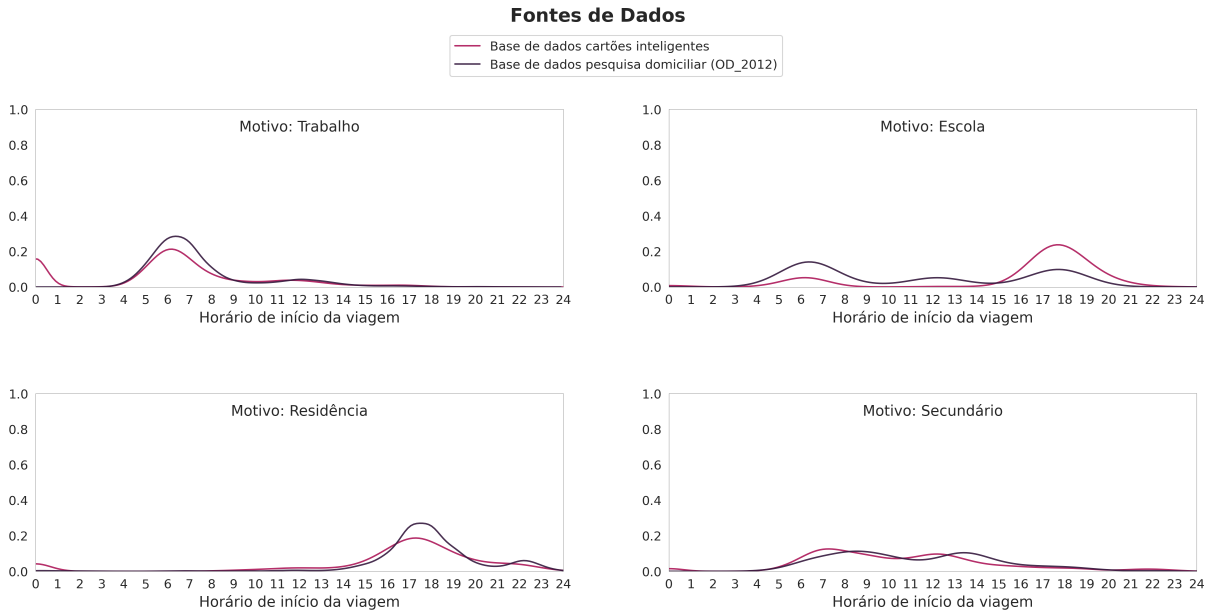
Conforme apresentado na Figura 25, a distribuição do horário de início das viagens é muito semelhante entre as duas matrizes. Para o motivo "Trabalho" nota-se um pico de viagens com horário de início entre 5h e 7h da manhã e uma diminuição substancial das viagens a trabalho com horário de início após as 7h. Uma pequena diferença encontrada entre as duas matrizes é que na matriz OD de bilhetagem foi observado um pico local de viagens a trabalho com horário de início à 00:00.

Para as viagens escolares, apesar do comportamento ser semelhante entre as duas matrizes, tendo cada uma três picos de viagem com horários de início que correspondem aos turnos escolares da manhã, tarde e noite respectivamente, observa-se que na matriz OD_2012 o pico da manhã é o mais expressivo, enquanto na matriz OD de bilhetagem eletrônica o maior pico é o da tarde, com horário de início entre 16 e 18h.

Viagens com motivo "Residência" também apresentaram comportamentos muito semelhantes entre as duas matrizes em relação ao horário de início. Para ambas matrizes, a maior parte das viagens com destino à residência iniciam-se no final da tarde, após as 16h.

Por fim, para os motivos secundários, o comportamento entre as matrizes também foi muito semelhante. Em ambas, as viagens secundárias são mais distribuídas com horários de início que variam entre 6h e 14h, não apresentando picos bem definidos.

Figura 25 – Matriz de Confusão Modelo Viagens Secundárias - Método 3



Fonte: Autoria própria

O tempo entre embarques, neste estudo, foi utilizado como uma *proxy* da duração da atividade no destino, conforme descrito detalhadamente na Subseção 5.3.2. Portanto, analisar o tempo entre embarques significa analisar quanto tempo os indivíduos gastam em cada uma das suas atividades, por exemplo: por quanto tempo os indivíduos permanecem em casa, para o motivo "Residência", ou quanto tempo eles passam trabalhando, estudando ou realizando atividades secundárias, para os motivos "Trabalho", "Escola" e "Secundários", respectivamente.

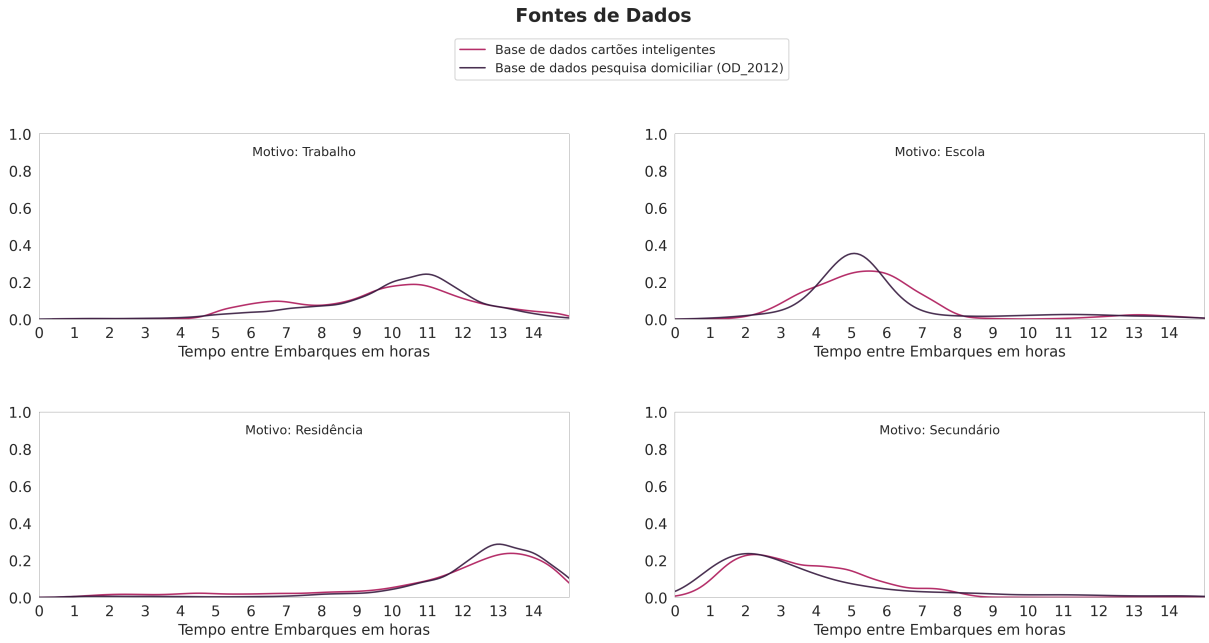
Em relação ao tempo entre embarques, o comportamento das duas matrizes também foi semelhante, conforme observado na Figura 26. Para o motivo "Trabalho", o tempo entre embarques foi de 10 a 12 horas para a maioria das viagens. Observa-se que na matriz OD de bilhetagem eletrônica há um pico local de viagens com duração da atividade de trabalho de 6 horas. Esse pico local foi menor na matriz OD_2012.

As viagens escolares foram as que apresentaram a maior diferença entre as duas matrizes. Observa-se que para a matriz OD de bilhetagem, a duração da atividade "Escola" foi mais distribuída, variando entre 3h e 7h de tempo entre embarques. Já na base de dados da pesquisa domiciliar, em geral, o tempo entre embarques para a atividade "Escola" foi de 4 a 6 horas.

Para as viagens cujo motivo de destino é "Residência", ambas matrizes apresentaram comportamentos muito similares para o tempo entre embarques, indicando que o tempo de permanência em casa varia de 10h a 12h de forma consistente.

Por fim, para as atividades secundárias, nas duas matrizes observou-se um tempo entre embarques de poucas horas, sendo que na matriz de bilhetagem eletrônica houveram algumas viagens com um tempo entre embarques maior que 4 horas.

Figura 26 – Matriz de Confusão Modelo Viagens Secundárias - Método 3



Fonte: Autoria própria

7.2 Resumo do capítulo

Conforme descrito no capítulo anterior, o Método 3 foi o que apresentou os melhores resultados, no entanto, apenas o modelo de viagens primárias alcançou métricas boas o suficientes para seguir para a etapa de implementação. Nesse contexto, este modelo foi aplicado a uma matriz estimada a partir de dados de cartões inteligentes, denominada matriz OD de bilhetagem eletrônica, para enriquecimento.

Após a aplicação do modelo aos dados da matriz OD de bilhetagem, foram feitas três análises comparativas entre esta matriz e a matriz OD_2012, criada a partir dos dados da pesquisa domiciliar de viagem: a proporção de viagens e a distribuição do horário de início e do tempo entre embarque das viagens para cada um dos motivos. De forma geral, observou-se que as duas matrizes apresentaram comportamentos semelhantes nas análises realizadas.

8 Tópicos Conclusivos

Neste capítulo, são apresentadas as considerações finais deste estudo. Ao longo desta dissertação, foi sugerido um método para inferir o motivo das viagens dos passageiros do transporte público coletivo a partir de dados de viagem da OD_2012 e espaciais de POIs. Essa abordagem fornece uma base sólida para prever o propósito da viagem dos passageiros em outros conjuntos de dados coletados de forma passiva, como, por exemplo, os de cartões inteligentes, conforme apresentado no Capítulo 7, podendo ser utilizada para enriquecer este tipo de dado e, assim, reduzir a dependência de pesquisas de viagem por entrevista no processo de planejamento de transportes. Nesse contexto, as considerações gerais abordarão a relevância dos resultados alcançados, suas implicações práticas e possíveis limitações do estudo. Além disso, serão discutidas oportunidades para futuras pesquisas como forma de proporcionar uma visão abrangente das contribuições desse trabalho e inspirar novas iniciativas de aplicação da mineração de dados na área da mobilidade urbana e planejamento de transportes.

8.1 Considerações Gerais

O modelo proposto neste estudo utiliza duas bases de dados para inferir o propósito da viagem, a saber: a Pesquisa domiciliar de Origem e Destino da RMBH e a base de dados de cadastro de CNPJ da Receita Federal do Brasil. Especialmente sobre a última base de dados, destaca-se que sua utilização na resolução de problemas da mobilidade urbana é extremamente valiosa e promissora, isto porque a utilização de dados de diferentes fontes adiciona contextos mais amplos aos dados de mobilidade, como o que foi observado neste estudo.

Os dados da OD_2012 possuíam informações das Áreas Homogêneas (AH) de origem e destino do deslocamento dos entrevistados. A partir disso, utilizando dados de CNPJ foi possível combinar a geolocalização dos endereços das empresas e as informações sobre suas atividades econômicas, possibilitando assim, segmentar as AH em diferentes tipos de interesse com base nas atividades predominantes em cada região. A utilização conjunta de dados de viagem com os dados disponibilizados pela Receita Federal pode ajudar a entender como diferentes setores contribuem para os padrões de viagem e quais áreas atraem mais viagem em função de motivos específicos. Além disso, em um aspecto mais amplo, os dados da Receita Federal podem ser usados para validar hipóteses sobre como diferentes setores afetam a mobilidade, permitindo que as autoridades de transporte e urbanismo tomem decisões mais informadas ao planejar rotas, horários de operação, infraestrutura de transporte e localização de POIs. Nesse contexto, a utilização destes dados no âmbito de uma cidade brasileira é a primeira contribuição deste estudo, sendo esta uma prática encontrada também nas literaturas mais recentes, conforme apresentado em Pezoa et al. (2023).

Após a proposição, calibração e validação de diferentes modelos de inferência do propósito

de viagem dos passageiros de transporte público, conclui-se que os motivos de "Residência", "Trabalho"e "Escola"podem ser classificados com alto nível de precisão, já os motivos secundários apresentam bons resultados quando são classificados todos em conjunto em uma única categoria denominada "Secundários". Ao tentar detalhar a classe de motivos secundários nos motivos de "Lazer", "Saúde", "Negócios"e "Outros", o modelo proposto não apresentou bons resultados, tendo apresentado uma acurácia máxima de 47% (Tabela 22). Destaca-se, em relação a isto, que na maioria dos casos as entrevistas das pesquisas domiciliares de viagem são feitas em dias úteis típicos, onde prevalecem os deslocamentos por motivos primários e, por esta razão, os dados gerados são significativamente desbalanceados. Este desbalanceamento comprometeu significativamente o desempenho do Modelo de Viagens Secundárias, isto porque, dado o tamanho amostral reduzido para este tipo de viagem, o modelo não conseguiu aprender e sintetizar o padrão deste tipo de deslocamento.

As viagens de caráter não obrigatório, em geral, são vivenciadas pelo indivíduo em seu tempo livre, com maior variabilidade espacial e sem regularidade temporal. Nesse contexto, embora estas viagens aconteçam em menor volume, há que se dar a devida importância para este tipo de deslocamento, porque para além do trabalho e da habitação, atividades de lazer, saúde e cultura, por exemplo, representam direitos sociais que devem ser acessados por todos. Por este motivo, ainda que o modelo deste trabalho não tenha obtido resultados satisfatórios para classificar as viagens em motivos secundários, ressalta-se a importância da continuidade dos estudos para se obter modelos melhores. Saber o motivo de viagem e consequentemente as características de cada tipo de deslocamento, em especial para os motivos secundários, é o primeiro passo para garantir a oferta de um transporte público de qualidade, que garanta o acesso igualitário da população aos serviços ofertados pela cidade.

Em relação à aplicação do modelo, a matriz OD estimada a partir de dados de cartões inteligentes foi escolhida como objeto de estudo. Isto, porque os dados de cartões inteligentes são coletados de forma passiva e contínua e estão prontamente disponíveis para planejadores e gestores do serviço de transporte público das cidades utilizarem no processo de planejamento. No entanto, qualquer fonte de dados que tenha informações sobre o horário de início da viagem e sobre o tempo entre embarques pode utilizada nesta etapa de aplicação, porque estas foram as variáveis de viagem mais importantes para o modelo. Ainda que o tempo entre embarques, que representa uma proxy da duração da atividade no destino, não esteja disponível na base de dados, ele pode ser calculado a partir do horário de início de todas as viagens de um mesmo usuário, assim como foi feito neste trabalho na etapa de engenharia de atributos.

Ainda sobre a aplicação do modelo proposto neste estudo em outras bases de dados, cabe dizer que as últimas matrizes Origem-Destino feitas para a RMBH, em 2019 e 2021, foram estimadas a partir de dados de telefonia móvel e, apesar do motivo de viagem já ter sido inferido nesta base de dados, as categorias são menos detalhadas do que as que foram propostas no presente estudo. Na OD de telefonia móvel, os motivos de trabalho-estudo foram inferidos em conjunto e, portanto, a separação destes motivos no método proposto neste estudo já representa

um ganho de informação.

Embora o método de inferência dos motivos de viagem proposto neste estudo gere previsões mais detalhadas, atualmente ele não pode ser aplicado aos dados da matriz OD por telefonia móvel, isto porque, além da base de dados desta matriz não contemplar o tempo entre embarques, cada registro de viagem na tabela corresponde a uma agregação de pelo menos 10 usuários. Tal agregação foi realizada para dificultar que um usuário pudesse ser identificado, no entanto, esta prática inviabiliza o cálculo do tempo entre embarques, já que esse cálculo depende da identificação da sequência de viagens realizadas pelo indivíduo, conforme exemplificado na Figura 8. Mais detalhes sobre a metodologia de obtenção da matriz OD da RMBH utilizando dados de telefonia móvel pode ser encontrada em RMBH (2023).

Ressalta-se que para a aplicação da metodologia apresentada neste estudo em outras fontes de dados, visando inferir os motivos de viagem das pessoas, a base de dados a ser enriquecida precisará conter, minimamente: um código identificador único do usuário, o horário de início e o local de destino de cada viagem, isto porque as demais variávies utilizadas no modelo podem ser extraídas com o auxílio de cálculos, como é o caso do tempo entre embarques e da quantidade de viagens por indivíduo.

De forma geral, entende-se que o objetivo nortedor deste estudo, de aplicação de técnicas de mineração de dados na inferência de informações de viagem ausentes em fontes de Big Data, foi alcançado. Salienta-se que este tipo de estudo tem contribuições recentes na literatura, com estudos que datam desde 2012 até os dias atuais, como é o caso do estudo recente de Pezoa et al. (2023), entretanto não foram encontrados, na revisão bibliográfica, estudos de caso sobre a aplicação do aprendizado de máquina na inferência de atributos ausentes de viagem para cidades brasileiras.

8.2 Trabalhos Futuros

Como o objetivo deste estudo foi inferir o motivo de viagem de passageiros do transporte público, apenas os registros de viagem feitas por este modo de transporte foram utilizados para treinamento do modelo, assim como foi feito por Pezoa et al. (2023). Esta decisão acarretou em uma queda significativa no volume de registros na base de dados rotulada, sobretudo para as classes de motivos secundários de viagem, pois muitas viagens não obrigatórias são feitas por outros modos de transporte, em especial o modo a pé, conforme apresentado na Tabela 2. Nesse contexto, uma abordagem relevante para trabalhos futuros consiste na avaliação do desempenho do modelo ao ser treinado com todos os registros de viagem independente do modo de transporte.

Outro desdobramento relevante deste estudo consiste na aplicação de uma abordagem não supervisionada para clusterização das viagens de acordo com seu padrão, e, posteriormente, inferir os propósitos de viagem. A base de dados rotulada utilizada neste estudo provém de uma pesquisa de campo realizada em 2012 e as desvantagens deste tipo de levantamento de dados, amplamente discutidas neste estudo, podem prejudicar a continuidade deste método de coleta.

Diante disso, faz-se necessário avaliar outras possibilidades de enriquecimento das fontes de *Big Data* com atributos de viagem ausentes, como é o caso do motivo de viagem, preferencialmente com métodos que não dependam de levantamentos externos.

Entretanto, ainda que o aprendizado não supervisionado seja útil para reduzir a dependência de fontes de dados rotulados, sabe-se que estas fontes são extremamente importantes para produzir dados de verdade e possibilitar a validação de metodologias de inferência.

No entanto, dada a alta precisão da inferência dos motivos de viagem primários, devido ao padrão bem definido deste tipo de viagem, os levantamentos das pesquisas domiciliares ou outros levantamentos de dados sobre mobilidade realizados futuramente podem dar mais ênfase na coleta de dados sobre as viagens secundárias, para que futuramente seja possível estimar modelos mais precisos de inferência em relação a este tipo de viagem, utilizando dados mais relevantes e representativos.

Durante o processo de modelagem realizado neste estudo, foi possível concluir que o processo metodológico, a disponibilidade de dados, as características da rede de transporte público e a distribuição de atividades no espaço urbano diferem significativamente dos estudos de caso encontrados na literatura e descritos detalhadamente na Capítulo 4, ainda que o objetivo final seja semelhante: inferir o propósito da viagem.

o mesmo de outros estudos : inferir o motivo de viagem dos passageiros, o processo metodológico, a disponibilidade dos dados, as características da rede de transporte público e a distribuição de atividades no espaço urbano diferem significativamente entre os estudos de caso. A exemplo disso, tem-se a característica da RMBH de concentração das atividades no município sede, Belo Horizonte, que impactou negativamente na contribuição dada pela inclusão de dados espaciais no modelo de classificação, que poderia ter sido maior em um cenário de distribuição mais equilibrada das atividades no espaço urbano.

Ainda sobre a utilização de dados espaciais, embora o modelo proposto mostre uma precisão relativamente alta na inferência do propósito da viagem, entende-se que os resultados podem ser melhorados aumentando o nível de detalhe dos dados de uso da terra. Neste estudo foram consideradas apenas as quantidades de estabelecimentos por tipo de atividade (representatividade absoluta). No entanto, outras formas de representação da informação espacial devem ser testadas como a representatividade relativa (percentual de estabelecimentos por tipo de atividade). Além disso, a ponderação dos estabelecimentos por tamanho pode ser uma forma de superar o desafio da concentração de atividades para a RMBH. Outros dados disponíveis na base de dados da Receita Federal, como o porte do estabelecimento, podem ser utilizados para esta ponderação, de maneira que grandes estabelecimentos, como, por exemplo, um grande hospital, com alta atratividade de passageiros, tenha maior importância na análise do que unidades locais de atendimento à saúde.

Por fim, ainda que o modelo de inferência de motivos de viagem secundários não tenha apresentado um desempenho satisfatório, ressalta-se a importância da continuidade deste estudo na tentativa de obter melhores resultados para estes casos. Atualmente, o planejamento de transportes na Região Metropolitana de Belo Horizonte é pautado pelas viagens obrigatórias.

Esta afirmação é confirmada pelo modo como os dados são coletados nas pesquisas domiciliares, que são realizadas em dias úteis típicos. Uma das características das viagens obrigatórias é que diversos fatores, além da disponibilidade de transporte, são considerados para a escolha do destino, que, na verdade, corresponde a uma decisão de médio/longo prazo acerca do local de residência, trabalho ou estudo. Já para as viagens secundárias, a disponibilidade de um transporte de qualidade, que atenda às necessidades da população para além das demandas rotineiras, pode ser um fator decisivo para o desenvolvimento urbano, para a promoção da acessibilidade e para fomentar o surgimento de outras centralidades. Isto porque para as categorias de Lazer, Saúde, Comércio, Negócios e Outros, as pessoas têm maior flexibilidade na escolha do destino.

Planejar um sistema de transporte público eficiente significa garantir que a população tenha boas condições de acesso tanto aos seus destinos mais visitados, como seu local de trabalho, estudos e residência, quanto a outras áreas da cidade, garantindo acessibilidade e qualidade também nas viagens a lazer ou para a promoção da saúde, por exemplo. Ressalta-se aqui que tal conclusão é especialmente relevante para pessoas de baixa renda, que dependem exclusivamente do transporte público para sua locomoção e para pessoas com idade mais avançada, cujas viagens por motivos secundários são até mais frequentes do que as viagens por motivos primários. Uma análise feita a partir dos dados da OD 2012 indica que, independentemente da faixa etária, aproximadamente 50% das viagens tem como motivo de destino a residência. Isto se deve ao comportamento predominantemente *home-based* das viagens, conforme apresentado na Figura 14. Os outros 50% de viagens variam substancialmente de acordo com a idade do indivíduo: para pessoas de 0 a 19 anos 40% são viagens escolares. Quando a faixa de idade aumenta para 20 a 59 anos, 30% das viagens são pelo motivo trabalho e 10% são por outros motivos. Para indivíduos acima de 60 anos, viagens pelo motivo saúde correspondem a aproximadamente 12% dos deslocamentos, o mesmo percentual é atribuído para viagens a lazer e outros motivos correspondem a aproximadamente 20% dos deslocamentos. Estes resultados reforçam a importância do entendimento do padrão de viagem secundário, pois só assim será possível compreender e acompanhar longitudinalmente o deslocamento da população a medida que ela vai envelhecendo.

Referências

ALI, A.; LEE, S. Destination and activity estimation. In: **Public Transport Planning with Smart Card Data**. [S.l.]: Taylor Francys Group, 2017. p. 37–53. Citado na página 28.

ALPAYDIM, E. **Introduction to Machine Learning**. 3. ed. [S.l.]: The MIT Press, 2014. Citado 3 vezes nas páginas 34, 43 e 44.

ALSGER, A. et al. Public transport trip purpose inference using smart card fare data. **Transportation Research Part C: Emerging Technologies**, v. 87, p. 123–137, 2018. Citado 6 vezes nas páginas 21, 23, 24, 28, 65 e 123.

ASGARI, F.; CLEMENCON, S. Transport mode detection when fine-grained and coarse-grained data meet. **3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE)**, p. 301–307, 2018. Citado na página 29.

ASLAM, N. S. et al. Activitynet: Neural networks to predict public transport trip purposes from individual smart card data and pois. **Geo-spatial Information Science**, v. 24, p. 711–721, 2021. Citado 5 vezes nas páginas 16, 27, 28, 67 e 123.

ASLAM, N. S. et al. Semantic enrichment of secondary activities using smart card data and point of interests: a case study in london. **Annals of GIS**, v. 27, n. 1, p. 29–41, 2020. Citado na página 16.

BAGCHI, M.; WHITE, P. R. The potential of public transport smart card data. **Transport Policy**, v. 12, n. 5, p. 464–474, 2005. Citado 4 vezes nas páginas 15, 21, 22 e 24.

BALDI, P. et al. Assessing the accuracy of prediction algorithms for classification: An overview. **Bioinformatics**, v. 16, p. 412–424, 2000. Citado 3 vezes nas páginas 59, 60 e 61.

BANISTER, D. **Inequality in transport**. 1. ed. [S.l.]: Marcham, Oxfordshire : Alexandrine Press, 2018. Citado 2 vezes nas páginas 27 e 63.

BARRY, J. J. et al. Origin and destination estimation in new york city with automated fare system data. **Transportation Research Record: Journal of the Transportation Research Board**, v. 1817, n. 1, p. 183–187, 2002. Citado na página 24.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. 1. ed. [S.l.]: Springer, 2006. Citado na página 44.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, p. 123–140, 1996. Citado na página 49.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, p. 5–32, 2001. Citado 2 vezes nas páginas 48 e 49.

BRUCE, P.; BRUCE, A.; GEDECK, P. **Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python**. 1. ed. [S.l.]: O'Reilly, 2020. Citado 3 vezes nas páginas 38, 39 e 40.

BRUTON, M. J. **Introdução ao planejamento dos transportes**. 1. ed. [S.l.]: Interciência, 1979. Citado 6 vezes nas páginas 15, 20, 21, 26, 29 e 30.

BURKOV, A. **The Hundred-Page Machine Learning Book**. 1. ed. [S.l.]: Andriy Burkov, 2019. Citado 7 vezes nas páginas 35, 48, 49, 55, 56, 59 e 60.

CÁCERES, N.; WIDEBERG, J.; BENITEZ, F. G. Review of traffic data estimations extracted from cellular networks. **IET Intelligent Transport Systems**, v. 2, n. 3, p. 179–192, 2008. Citado 3 vezes nas páginas 15, 20 e 21.

CAMPOS, V. B. G. **Planejamento de Transportes Conceitos e Modelos**. 1. ed. [S.l.]: Interciência, 2013. Citado 2 vezes nas páginas 20 e 21.

CHEIN, C.; GUESTRIN, C. Xgboost: A scalable tree boosting system. **22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, p. 785–794, 2016. Citado na página 50.

CHEN, C. et al. The promises of big data and small data for travel behavior (aka human mobility) analysis. **Transportation Research Part C: Emerging Technologies**, v. 68, p. 285–299, 2016. Citado 4 vezes nas páginas 15, 21, 22 e 24.

CHU, K. K. A.; CHAPLEAU, R. Enriching archived smart card transaction data for transit demand modeling. **Transportation Research Record**, v. 2063, p. 63–72, 2008. Citado na página 28.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, p. 273–297, 1995. Citado 3 vezes nas páginas 52, 53 e 54.

DEVELOPERS, X. **XGBoost parameter documentation**. 2023. Disponível em: <https://-/xgboost.readthedocs.io/en/latest/parameter.html>. Citado 2 vezes nas páginas 57 e 58.

DEVILLAINE, F.; MUNIZAGA, M. A.; TREPANIER, M. Detection of activities of public transport users by analyzing smart card data. **Transportation Research Record: Journal of the Transportation Research Board**, v. 2276, n. 1, p. 48–55, 2012. Citado 4 vezes nas páginas 16, 22, 28 e 64.

DINNO, A. Nonparametric pairwise multiple comparisons in independent groups using dunn's test. **Stata**, 2015. Citado na página 41.

EGU, O.; BONNEL, P. How comparable are origin-destination matrices estimated from automatic fare collection, origin-destination surveys and household travel survey? an empirical investigation in lyon. **Transportation Research Part A: Policy and Practice**, v. 138, p. 267–282, 2020. Citado na página 25.

FAROQUI, H.; MESBAH, M. Inferring trip purpose by clustering sequences of smart card records. **Transportation Research Part C: Emerging Technologies**, v. 127, p. 103–131, 2021. Citado 2 vezes nas páginas 28 e 65.

FAROQUI, H.; MESBAH, M.; KIM, J. Investigating the correlation between activity similarity and trip similarity of public transit passengers using smart card data. **Transportation Research Procedia**, v. 48, n. 4, p. 2621–2637, 2020. Citado 2 vezes nas páginas 28 e 30.

FERRAZ, I. G. E. T. A. C. C. P. **Transporte Público Urbano**. 2. ed. [S.l.]: Rima, 2004. Citado 2 vezes nas páginas 15 e 29.

GARCÍA-ALBERTOS, P. et al. Exploring the potential of mobile phone records and online route planners for dynamic accessibility analysis. **Transportation Research Part A: Policy and Practice**, v. 125, p. 294–07, 2019. Citado 2 vezes nas páginas 16 e 24.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data minning: conceitos, tecnicas, algoritmos, orientações e aplicações**. 2. ed. [S.l.]: Elsevier, 2015. Citado 5 vezes nas páginas 33, 34, 35, 41 e 42.

GOULET-LANGLOIS, G.; KOUTSOPOULOS, H. N.; ZHAO, J. Inferring patterns in the multi-week activity sequences of public transport users. **Transportation Research Part C: Emerging Technologies**, v. 64, p. 1–16, 2016. Citado na página 64.

GUERRA, A. L. Estimativa de matriz origem/destino utilizando dados do sistema debilhetagem eletrônica: proposta metodológica. **TRANSPORTES**, v. 22, n. 3, p. 26–38, 2014. Citado 4 vezes nas páginas 15, 20, 21 e 22.

HAN, G.; SOHN, K. Activity imputation for trip-chains elicited from smart-card data using a continuous hidden markov model. **Transportation Research Part B: Methodological**, v. 83, p. 121–135, 2016. Citado na página 28.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3. ed. [S.l.]: Elsevier, 2012. Citado 9 vezes nas páginas 35, 44, 45, 46, 47, 53, 54, 59 e 60.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference and Prediction**. 2. ed. [S.l.]: Springer, 2008. Citado 2 vezes nas páginas 43 e 44.

HE, H.; BAI, Y.; GARCIA, E. A. **Imbalanced Learning: Foundations, Algorithms, and Aplications**. [S.l.]: IEEE Press, 2013. Citado na página 36.

HE, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, p. 1263–1284, 2009. Citado 3 vezes nas páginas 36, 37 e 38.

HENSHER, D. A.; GOLOB, T. F. Bus rapid transit systems: a comparative assessment. **Transportation**, v. 35, p. 501–518, 2008. Citado na página 15.

HOSSAIN sanjana; HABIB, K. N. Inferring the purposes of using ride-hailing services through data fusion of trip trajectories, secondary travel surveys, and land use data. **Transportation Research Record: Journal of the Transportation Research Board**, v. 2675, n. 9, p. 558–573, 2021. Citado na página 28.

HUANG, H.; CHENG, Y.; WEIBEL, R. Transport mode detection based on mobile phone network data: A systematic review. **Transportation Research Part C: Emerging Technologies**, v. 101, p. 297–312, 2019. Citado na página 26.

HUI, T. Y. **Investigating the use of anonymous cellular data for intercity travel patterns in Alberta**. Dissertação (Mestrado) — University of Alberta, 2017. Citado na página 30.

HUSSAIN, E.; BHASKAR, A.; CHING, E. Transit od matrix estimation using smartcard data: Recent developments and future research challenges. **Transportation Research Part C**, v. 125, 2021. Citado 4 vezes nas páginas 16, 23, 24 e 28.

IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina: uma abordagem estatística**. 1. ed. [S.l.: s.n.], 2020. Citado na página 38.

JORDAHL, K. et al. **geopandas/geopandas: v0.8.1**. Zenodo, 2020. Disponível em: <https://doi-.org/10.5281/zenodo.3946761>. Citado na página 89.

KANG, M. J.; ATAEIAN, S.; AMIRIPOUR, S. M. M. A procedure for public transit od matrix generation using smart card transaction data. **Public Transport**, v. 13, p. 81–100, 2020. Citado na página 21.

KIM, E. J.; KIM, D. K.; SOHN, K. Imputing qualitative attributes for trip chains extracted from smart card data using a conditional generative adversarial network. **Transportation Research Part C: Emerging Technologies**, v. 137, 2022. Citado na página 30.

KIM, E. J.; KIM, Y.; KIM, D. K. Interpretable machine-learning models for estimating trip purpose in smart card data. In: **Proceedings of the Institution of Civil Engineers: Municipal Engineer**. [S.l.: s.n.], 2021. p. 108–117. Citado 4 vezes nas páginas 8, 28, 66 e 122.

KUHLMAN, W. **The construction of purpose specific OD matrices using public transport smart card data**. Dissertação (Mestrado) — Delft University of Technology, 2015. Citado na página 24.

KUSAKABE, T.; ASAKURA, Y. Combination of smart card data with person trip survey data. In: **Public Transport Planning with Smart Card Data**. [S.l.: s.n.]. Citado na página 28.

KUSAKABE, T.; ASAKURA, Y. Behavioural data mining of transit smart card data: A data fusion approach. **Transportation Research Part C: Emerging Technologies**, v. 46, p. 179–191, 2014. Citado 3 vezes nas páginas 16, 22 e 28.

KYAING; LWIN, K. K.; SEKIMOTO, Y. Identification of various transport modes and rail transit behaviors from mobile cdr data: A case of yangon city. **Asian Transport Studies**, v. 6, 2020. Citado na página 25.

LANDMARK, A. D. et al. Mobile phone data in transportation research: methods for benchmarking against other data sources. **Transportation**, v. 101, p. 2883–2905, 2021. Citado na página 26.

LEE, S. G.; HICKMAN, M. Trip purpose inference using automated fare collection data. **Public Transport**, v. 6, p. 1–20, 2014. Citado na página 28.

LU, K.; KHANI, A.; HAN, B. A trip purpose-based data-driven alighting station choice model using transit smart card data. **Complexity**, v. 2018, 2018. Citado na página 28.

MEDINA, S. A. O. Inferring weekly primary activity patterns using public transport smart card data and a household travel survey. **Travel Behaviour and Society**, v. 12, p. 91–101, 2018. Citado na página 65.

MUNIZAGA, M. A.; PALMA, C. Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile. **Transportation Research Part C: Emerging Technologies**, v. 24, p. 9–18, 2012. Citado na página 24.

MURPHY, K. P. **Machine Learning: a probabilistic perspective**. 1. ed. [S.l.]: The MIT Press, 2012. Citado 5 vezes nas páginas 35, 43, 52, 53 e 54.

OPENSTREETMAP. **Open-source geocoding with OpenStreetMap data**. 2023. Disponível em: <https://nominatim.org>. Citado na página 89.

ORTUZAR, J. D. D.; WILLUMSEN, L. G. **Modelling Transport**. 4. ed. [S.l.]: John Wiley and Sons, 2011. Citado 7 vezes nas páginas 15, 20, 21, 23, 28, 30 e 31.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citado 3 vezes nas páginas 56, 57 e 58.

PELLETIER, M. P.; MORENCY, C. Smart card data use in public transit: A literature review. **Transportation Research Part C: Emerging Technologies**, v. 19, n. 4, p. 557–568, 2011. Citado 2 vezes nas páginas 24 e 25.

PEZOA, R. et al. Estimation of trip purposes in public transport during the covid-19 pandemic: The case of santiago, chile. **Journal of transport geography**, v. 109, 2023. Citado 7 vezes nas páginas 63, 64, 67, 97, 123, 130 e 132.

PIERONI, C. et al. Big data for big issues: Revealing travel patterns of low-income population based on smart card data mining in a global south unequal city. **Journal of Transport Geography**, v. 96, 2021. Citado na página 28.

PINHEIRO, J. I. D. et al. **Estatística Básica: a arte de trabalhar com dados**. 1. ed. [S.l.]: Elsevier, 2009. Citado na página 39.

QUEIXO, K. et al. Inferring fine-grained transport modes from mobile phone cellular signaling data. **Computers, Environment and Urban Systems**, 2019. Citado 2 vezes nas páginas 29 e 30.

QUINLAN, J. R. "induction of decision trees". In: **"Machine Learning"**. [S.l.: s.n.], 1986. p. 81–106. Citado 2 vezes nas páginas 45 e 46.

REDDY, S. et al. Using mobile phones to determinetransportation modes. **ACM Trans. Sensor Netw**, v. 6, p. 1–27, 2010. Citado na página 29.

REIF, M.; SHAFAIT, F.; DENGEL, A. Meta-learning for evolutionary parameter optimization of classifiers. **Machine Learning**, v. 87, p. 357–380, 2012. Citado na página 56.

RFB, S. E. da Receita Federal do B. **Cadastro Nacional da Pessoa Jurídica - CNPJ**. 2023. Disponível em: <https://dados.gov.br/dados/conjuntos-dados/cadastro-nacional-da-pessoa-juridica—cnpj>. Citado na página 75.

RMBH, A. de Desenvolvimento da Região Metropolitana de B. H. **Pesquisa Origem e Destino 2012**. 2023. Disponível em: <http://www.agenciarmbh.mg.gov.br/pesquisa-od/>. Citado 2 vezes nas páginas 74 e 132.

SIEGEL, S. **NonParamec Statistics: for the behavioral sciences**. 1. ed. [S.l.]: McGRAW-HILL BOOK COMPANY, INC, 1956. Citado 2 vezes nas páginas 40 e 41.

STOPHER, P. R.; FITZGERALD, C.; XU, M. Assessing the accuracy of the sydney household travel survey with gps. **Transportation**, v. 34, p. 723–741, 2007. Citado na página 22.

STOPHER, P. R.; GREAVES, S. P. Household travel surveys: where are we going? **Transport Res. Part A: Policy Practice**, v. 41, p. 367–381, 2007. Citado na página 22.

STOPHER, P. R. et al. Household travel surveys: Proposed standards and guidelines. **Travel Survey Methods, Emerald Group Publishing Limited, Leeds**, p. 19–74, 2006. Citado na página 27.

TREPANIER, M.; TRANCHANT, N.; CHAPLEAU, R. Individual trip destination estimation in a transit smart card automated fare collection system. **Journal of Intelligent Transportation Systems**, v. 11, n. 1, p. 1–14, 2007. Citado na página 24.

WAG, M.-H.; BROEK, S. D. S. ad N. V.; MULINAZZI, T. Estimating dynamic origin-destination data and travel demand using cell phone network data. **International Journal of Intelligent Transportation Systems Research**, v. 11, p. 76 – 86, 2013. Citado 2 vezes nas páginas 25 e 26.

WANG, H. et al. Trasportation mode inference from anonymized and aggregated mobile phone call detail records. **13th International IEEE Conference on Intelligent Transportation Systems**, p. 318–323, 2010. Citado 3 vezes nas páginas 25, 29 e 30.

WANG, Y. et al. Using metro smart card data to model location choice of after-work activities: An application to shanghai. **Journal of transport geography**, v. 63, p. 40–47, 2017. Citado na página 16.

WU, L.; YANG, B.; JING, P. Travel model detection based on gps raw data collected by smartphones: a systematic review of the existing methodologies. **Information**, v. 7, 2016. Citado 2 vezes nas páginas 25 e 29.

YANG, Y. et al. Who, where, why and when? using smart card and social media data to understand urban mobility. **ISPRS International Journal of Geo-Information**, v. 8, 2019. Citado 2 vezes nas páginas 28 e 65.

ZANNAT, K. E.; CHOUDHURY, C. F. Emerging big data sources for public transport planning: A systematic review on current state of art and future research directions. **J Indian Inst Sci**, v. 99, p. 601—-619, 2019. Citado 2 vezes nas páginas 27 e 29.

ZHAO, J.; RAHBEE, A.; WILSON, N. H. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. **Computer-Aided Civil and Infrastructure Engineering**, v. 22, n. 5, p. 376–387, 2007. Citado na página 24.

ZHAO, Y.; PAWLAK, J.; SIVAKUMAR, A. Theory for socio-demographic enrichment performance using the inverse discrete choice modelling approach. **Transportation Research Part B: Methodological**, v. 155, p. 101–134, 2022. Citado na página 30.

ZHAO, Z.; KOUTSOPOULOS, H. N.; ZHAO, J. Discovering latent activity patterns from transit smart card data: a spatiotemporal topic model. **Transportation Research Part C: Emerging Technologies**, v. 116, 2020. Citado na página 28.

ZHU, Y. Inference of activity patterns from urban sensing data using conditional random fields. **Environment and Planning B: Urban Analytics and City Science**, v. 49, n. 2, p. 549–565, 2021. Citado na página 28.

ZOU, Q. et al. Detecting home location and trip purpose for cardholders by minning smart card transaction data in beijing subway. **Transportation**, v. 45, p. 919–944, 2018. Citado na página 28.