INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA E GESTÃO DA CADEIA DE ABASTECIMENTO




APLICAÇÃO DE APRENDIZAGEM AUTOMÁTICA À PREVISÃO DE ROTAS BASEADO NO COMPORTAMENTO DO MOTORISTA

LUIS VITOR LISBOA NAVEGA
setembro de 2024

ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

APLICAÇÃO DE *MACHINE LEARNING* À PREVISÃO DE ROTAS BASEADO NO COMPORTAMENTO DO MOTORISTA

Luis Vitor Lisboa Navega

2024

Instituto Superior de Engenharia do Porto

Departamento de Engenharia Mecânica

P.PORTO


P.PORTO

ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

APLICAÇÃO DE *MACHINE LEARNING* À PREVISÃO DE ROTAS BASEADO NO COMPORTAMENTO DO MOTORISTA

Luis Vitor Lisboa Navega

Estudante n.º 1220290

Dissertação apresentada ao Instituto Superior de Engenharia do Porto para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia e Gestão da Cadeia de Abastecimento, realizada sob a orientação do Doutor Carlos Manuel Abreu Gomes Ferreira e coorientação do Doutor António José Galrão Ramos

2024

Instituto Superior de Engenharia do Porto

Departamento de Engenharia Mecânica

P.PORTO


P.PORTO

AGRADECIMENTOS

Ao concluir esta etapa importante da minha jornada académica, não poderia deixar de expressar a minha mais profunda gratidão a todos aqueles que, de alguma forma, contribuíram para a realização deste trabalho.

Em primeiro lugar, gostaria de agradecer ao Instituto Superior de Engenharia do Porto (ISEP) por proporcionar a estrutura e os recursos necessários para que este projeto fosse possível. A oportunidade de aplicar os conhecimentos adquiridos no curso de Engenharia e Gestão da Cadeia de Abastecimento, particularmente na aplicação de Aprendizagem Automática à Previsão de Rotas com Base no Comportamento do Motorista, foi inestimável e desafiadora, permitindo-me explorar o campo de forma profunda e prática.

Agradeço ao meu orientador, Professor Doutor Carlos Manuel Abreu Gomes Ferreira, cuja orientação foi essencial para o sucesso desta pesquisa. A sua disponibilidade, paciência e expertise foram fundamentais para que este trabalho atingisse a qualidade desejada.

Não posso deixar de mencionar o apoio constante dos meus pais, da minha namorada e de toda a minha família, que sempre acreditaram no meu potencial e me apoiaram incondicionalmente em cada etapa desta jornada. A eles, devo toda a força e determinação que me permitiram chegar até aqui.

Aos meus amigos e colegas de curso, cujo incentivo e companheirismo tornaram os desafios mais leves e as conquistas mais significativas, o meu sincero agradecimento. Partilhar esta jornada convosco foi uma experiência enriquecedora e gratificante.

Por fim, dedico este trabalho a todos que, direta ou indiretamente, contribuíram para que eu pudesse alcançar este objetivo. As suas palavras de incentivo, gestos de apoio e confiança depositada em mim foram a base que sustentou este projeto.

A todos, o meu muito obrigado!

RESUMO

Este estudo investiga a aplicação de técnicas de *Machine Learning* na previsão de rotas, com foco no comportamento do motorista, utilizando uma base de dados fornecida por uma empresa de logística. O comportamento dos motoristas é influenciado por vários fatores, como preferências pessoais, experiência com determinadas rotas, condições do tráfego e características das entregas, como a distância, o peso da mercadoria, a urgência e a frequência de visita a um cliente específico. Estas variáveis tornam o planeamento de rotas uma tarefa complexa, mas essencial para a eficiência logística.

O principal objetivo deste trabalho foi desenvolver um sistema preditivo robusto e eficiente que pudesse não apenas prever as rotas mais prováveis com base nas escolhas históricas dos motoristas, mas também otimizar essas rotas com base em múltiplos atributos operacionais. Estes atributos incluem a distância total percorrida, o peso da carga, a capacidade do veículo, e em alguns modelos a frequência de visita ao cliente. A adição da frequência de visita a certos clientes foi particularmente importante para aumentar a precisão das previsões. Esses atributos influenciam diretamente a escolha de uma rota, uma vez que os motoristas tendem a preferir rotas mais conhecidas para clientes frequentes ou ajustar as suas escolhas com base no peso e no volume das entregas, visando minimizar o esforço físico e o tempo de transporte.

O estudo iniciou-se com a organização dos dados em conjuntos de treino e teste, permitindo a aplicação e a validação dos modelos de *Machine Learning*. Diversas técnicas foram exploradas, incluindo modelo de *Osquare* juntamente com técnicas de Regressão Linear, *Random Forest*, Redes Neurais e Support Vector Machine (SVM). A eficácia desses modelos foi avaliada através de métricas como *Kendall*'s Tau, *Accuracy* e *Edit Distance*, permitindo uma análise comparativa dos resultados.

Os resultados indicaram que os modelos de regressão com Redes Neurais, com a adição da frequência como um atributo, se destacaram em termos de precisão e eficiência geral, obtendo uma acurácia média de 0,9603, e um *Kendall* médio de 0,0314. O modelo *Random Forest* também apresentou um bom desempenho, particularmente quando otimizado e com o atributo adicional da frequência de visita, atingindo uma acurácia média de 0,9583 e um *Kendall* médio de 0,333. Em contrapartida, técnicas como a regressão logística e SVM mostraram-se menos eficazes em certos cenários.

Este estudo demonstra o potencial do *Machine Learning* para otimizar processos logísticos, destacando a importância de uma análise detalhada e criteriosa dos dados. As descobertas oferecem contribuições significativas para o campo da logística, mostrando como a integração de técnicas avançadas de aprendizagem automática pode ser aplicada de forma eficaz no contexto industrial, ao mesmo tempo em que aponta caminhos para melhorias futuras.

Palavras-Chave: *Machine Learning, Osquare, Imitation Learning, Kendall, Accuracy, Edit Distance*, Previsão de Rotas.

ABSTRACT

This study investigates the application of *Machine Learning* techniques in route prediction, with a focus on driver behavior, using a dataset provided by a logistics company. Driver behavior is influenced by various factors, such as personal preferences, experience with specific routes, traffic conditions, and delivery characteristics, including distance, cargo weight, urgency, and the frequency of visits to a specific customer. These variables make route planning a complex but essential task for logistical efficiency.

The main objective of this work was to develop a robust and efficient predictive system that could not only forecast the most likely routes based on drivers' historical choices but also optimize these routes based on multiple operational attributes. These attributes include the total distance traveled, load weight, vehicle capacity, and, in some models, customer visit frequency. The addition of visit frequency to certain customers was particularly important for improving prediction *accuracy*. These attributes directly influence route choices, as drivers tend to prefer familiar routes to frequent customers or adjust their choices based on the weight and volume of deliveries to minimize physical effort and transport time.

The study began by organizing the data into training and testing sets, allowing for the application and validation of the *Machine Learning* models. Various techniques were explored, including *Osquare* modeling along with Linear Regression, *Random Forest*, Neural Networks, and Support Vector Machine (SVM) techniques. The effectiveness of these models was evaluated through metrics such as *Kendall*'s Tau, *Accuracy*, and *Edit Distance*, allowing for a comparative analysis of the results.

The results indicated that regression models with Neural Networks, with the addition of frequency as an attribute, stood out in terms of overall precision and efficiency, achieving an *accuracy* of 0.9603 and a *Kendall* of 0.0314. The *Random Forest* model also performed well, particularly when optimized and with the additional frequency attribute, reaching an *accuracy* of 0.9583 and a *Kendall* of 0.333. In contrast, techniques such as logistic regression and SVM were less effective in certain scenarios.

This study demonstrates the potential of *Machine Learning* to optimize logistical processes, highlighting the importance of a detailed and careful data analysis. The findings offer significant contributions to the field of logistics, showing how the integration of advanced *Machine Learning* techniques can be effectively applied in an industrial context while also pointing to avenues for future improvements.

Keywords: *Machine Learning*, *Osquare*, *Imitation Learning*, *Kendall*, *Accuracy*, *Edit Distance*, Route Prediction.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTAS DE SIGLAS E SÍMBOLOS

Lista de Siglas

ANN	*Artificial Neural Network*
CNN	*Convultional Neural Network*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
GPS	*Global Positioning System*
IA	Inteligêcia Artificial
IL	*Imitation Learning*
PCV	Problema do Caixeiro Viajante
SIG	Sistema de Informação Geográfica
SVM	*Support Vector Machine*
ReLU	*Rectified Linear Unit*
RNN	Recurrent Neural Network
VRP	*Vehicle Routing Problem*

1. INTRODUÇÃO

A área de Engenharia e Gestão da Cadeia de Abastecimento tem revelado uma preocupação crescente em desenvolver metodologias avançadas que melhorem a eficiência e a sustentabilidade dos processos operacionais. Esta abordagem implica um progresso, considerando as necessidades do setor para elevar os padrões de excelência e competitividade.

Este capítulo aborda o problema de investigação e as questões que esta dissertação se propõe a responder, explicando o contexto envolvente e a sua relevância para a comunidade científica. Serão detalhados os objetivos da investigação e a questão principal a que se pretende responder. Serão também descritas as opções metodológicas seguidas. Por fim, será apresentada a estrutura global da dissertação.

1.1. Enquadramento e pertinência

O aumento da competitividade no mercado empresarial impulsionou a procura por técnicas e ferramentas tecnológicas mais inovadoras, com o objetivo de otimizar os processos organizacionais. O compromisso com a inovação e a busca por métodos que assegurem um aumento da eficiência operacional oferece uma vantagem competitiva face à concorrência, reduzindo os constrangimentos operacionais. Uma empresa que adota métodos eficazes para a otimização dos seus processos consegue atrair significativamente novos clientes, contribuindo para o aumento da lucratividade da companhia e para a redução de custos desnecessários com recursos adicionais. Assim, é fundamental salientar a importância da aplicação de métodos que visem a otimização dos processos da empresa, independentemente do sector em que são aplicados, focando-se na redução dos custos operacionais e na diminuição da utilização desnecessária de recursos (SILVA, 2014).

O desafio da previsão de rotas com base no comportamento do condutor, segundo Andriotti (2004), refere-se à capacidade de prever trajectos otimizados, tendo em conta as preferências e os padrões de condução dos motoristas. Este problema tem-se tornado cada vez mais relevante para as empresas de logística, especialmente com o aumento das operações de comércio electrónico e a crescente procura por eficiência no transporte. A complexidade de prever rotas eficientes é exacerbada pela variabilidade do comportamento dos condutores e pelas condições dinâmicas das vias urbanas.

De acordo com Ferreira (2011), a análise comportamental dos motoristas é crucial para a otimização de rotas, pois permite a consideração de fatores que vão além dos simples cálculos de distância ou tempo. A capacidade de prever trajetos com base no comportamento dos motoristas pode resultar em significativas economias de custo e melhorias na eficiência operacional. Dada a elevada despesa associada ao transporte e à entrega urbana, otimizar essas rotas é de extrema importância para as empresas de logística.

Segundo Mendonça (2021), a aplicação de algoritmos de *Machine Learning* tem demonstrado um grande potencial na previsão de rotas. A aprendizagem Supervisionada, por exemplo, pode ser utilizada para analisar padrões comportamentais passados e prever comportamentos futuros, ajustando as rotas de acordo com as preferências e os hábitos dos motoristas. Além disso, técnicas avançadas como as Redes Neurais Artificiais (ANN) oferecem capacidades poderosas para lidar com a complexidade destes problemas.

Este estudo adota uma abordagem estruturada com base nas fases do CRISP-DM (*Cross Industry Standard Process for Data Mining*). O processo começa pela compreensão dos objetivos e requisitos do setor logístico, assegurando que a solução esteja alinhada com as necessidades reais. Em seguida, procede-se à análise dos dados disponíveis, como rotas, padrões de condução e variáveis contextuais. Após a limpeza e transformação dos dados, de forma a torná-los adequados para a modelação, são aplicados vários modelos de *Machine Learning*, incluindo Regressão Linear, *Random Forest* e Redes Neurais Artificiais, com ou sem a utilização de atributos derivados de tabelas de frequências de visita aos clientes. A avaliação dos modelos é feita através de métricas como precisão (*accuracy*), correlação de *Kendall* e *Edit Distance*.

A relevância deste problema é reforçada pela crescente procura por soluções eficientes e sustentáveis no sector da logística. A previsão de rotas com base no comportamento dos motoristas não possibilita só a melhora da eficiência operacional, mas também pode contribuir para a redução de custos e para uma gestão mais eficaz dos recursos. Ao aplicar técnicas de *Machine Learning* para este fim, este estudo procura preencher uma lacuna significativa na área, proporcionando uma ferramenta para a previsão de rotas. Desta forma, esta dissertação oferece uma contribuição importante para o sector dos Transportes e Logística, sublinhando a importância da análise comportamental e da modelagem preditiva para enfrentar os desafios atuais do transporte urbano.

1.2. Questão e objetivos de investigação

O principal objetivo desta pesquisa é explorar como a aplicação de técnicas de *Machine Learning* podem melhorar a previsão de rotas baseada no comportamento dos motoristas. Pretende- se identificar e avaliar de que forma diferentes representações baseadas no histórico de comportamentos de motorista impactam a eficiência das rotas sugeridas e o desempenho geral do sistema. A questão central desta investigação é: “Qual modelo de previsão de rotas baseado no comportamento do motorista apresenta melhor desempenho? Além disso, a adição do atributo de frequência de visita impacta positivamente na precisão do modelo?”.

Para atingir o objetivo geral, os seguintes objetivos específicos foram delineados:

- Aprender modelos, utilizando dados históricos, para prever rotas que aproximam o comportamento dos motoristas. Sendo estes modelos construídos utilizando um conjunto de rotas passadas, estes aprendem e codificam conhecimento tácito dos motoristas;
- Aplicar técnicas de *Imitation Learning* para treinar agentes que aprendam a otimizar suas decisões de roteamento com base em feedback contínuo e interação com o ambiente;
- Utilização de modelos de Regressão e Classificação junto ao *Osquare*: Serão utilizados modelos de regressão, como Regressão Linear Simples, *Random Forest* e Redes Neurais,

bem como modelos de classificação, como Regressão Logística, *Random Forest* e SVM (*Support Vector Machine*), para prever e classificar o comportamento dos motoristas e melhorar a precisão das rotas sugeridas e ser analisada relativamente às rotas reais realizada pelo motorista. (Harrison, 2019)

- Aplicação de métricas para avaliar o desempenho: Empregar métricas como *Accuracy, Edit Distance* e *Kendall* para avaliar os modelos e apoiar as análises de resultados e a tomada de decisões.

Através da implementação e validação destes objetivos, a pesquisa visa não apenas aprimorar a eficiência operacional das empresas de logística, mas também contribuir com insights na aplicação de *Machine Learning* ao roteamento de veículos, tendo em conta o comportamento dos motoristas. Um estudo experimental será conduzido, seguindo as orientações do CRISP-DM, para desenvolver, testar e validar modelos preditivos que melhorem a precisão das rotas sugeridas.

1.3. Opções metodológicas

Este capítulo tem como objetivo apresentar as opções metodológicas consideradas para a realização desta dissertação. Conforme mencionado anteriormente, a abordagem de investigação segue uma perspetiva quantitativa, uma vez que se trata de um estudo objetivo, fundamentado em métodos científicos, para prever rotas com base no comportamento dos motoristas, testando os resultados e avaliando-os com métricas em relação as rotas realizadas pelos motoristas. Esta abordagem é suportada por um modelo hipotético-dedutivo, partindo do princípio de que o problema de investigação alcança uma solução objetiva através de um método científico. No que diz respeito às técnicas de recolha de dados, estas centram-se na análise documental, abrangendo desde dissertações de mestrado até artigos científicos relevantes nas áreas de logística e *Machine Learning.* Sendo assim, para estrutura do trabalho foi aplicado a metodologia CRISP – DM, de modo a considerar cinco fases desse método.

No âmbito dos métodos de investigação, este projeto insere-se no paradigma do estudo de caso, caracterizado por uma abordagem intensiva e minuciosa de uma entidade claramente definida. Isso implica uma análise aprofundada de uma situação específica no contexto real. Assim, o trabalho inicia-se com uma base teórica sólida e uma descrição detalhada do problema em análise, visando uma contextualização eficaz e a compreensão do negócio. Em seguida, na fase de compreensão dos dados foram analisados a base de dados para possibilitar o conhecimento do que se tratam os dados, utilizando dados históricos sobre o comportamento dos motoristas, com o objetivo de alcançar os resultados pretendidos.

A terceira etapa envolveu a limpeza e transformação dos dados brutos, tornando-os adequados para a modelagem. Este processo incluiu a normalização dos dados, tratamento de valores ausentes, categorização de variáveis e criação de novas características relevantes. Foram construídas tabelas de frequência para analisar padrões de comportamento dos motoristas, que posteriormente foram utilizadas como atributos adicionais nos modelos preditivos.

Sendo assim, os modelos desenvolvidos foram avaliados utilizando um conjunto de dados de teste independente. Foram empregues várias métricas de desempenho, incluindo precisão

(*accuracy*), correlação de *Kendall* e *Edit Distance*, distância total e tempo total, para comparar a eficácia dos diferentes modelos em relação aos cenários das rotas reais da base de dados. Esta análise comparativa permitiu identificar os modelos que melhor se adequam às necessidades da empresa e apresentavam o melhor desempenho preditivo. Essa metodologia adotada, fundamentada no CRISP-DM, garante uma abordagem rigorosa e sistemática para a aplicação de *Machine Learning* na previsão de rotas com base no comportamento dos motoristas, assegurando a relevância prática e a validade científica dos resultados obtidos

1.4. Inovações e Contribuições

O estudo busca a previsão de rotas com base no comportamento dos motoristas através de modelos de *Machine Learning*. Estes modelos podem oferecer uma grande flexibilidade e conseguem aprender de forma eficaz as relações complexas entre os padrões de condução dos motoristas e as rotas otimizadas desejadas.

Portanto, este trabalho busca automatizar a previsão de rotas e avaliar alguns modelos, considerando as características comportamentais dos motoristas. Além disso, esta investigação abre novas oportunidades para pesquisas futuras e avanços, promovendo uma colaboração mais estreita entre o campo da inteligência artificial e as aplicações práticas na indústria logística. A aplicação de técnicas de *Machine Learning* neste contexto não só aumenta a eficiência operacional, mas também, proporciona uma ferramenta prática e avançada para enfrentar os desafios contemporâneos do transporte urbano.

1.5. Estrutura do Documento

Este documento está organizado em cinco capítulos principais, cada um dos quais visa conduzir o leitor através das etapas lógicas da investigação realizada, desde o contexto inicial até à análise de resultados e conclusões.

No primeiro capítulo, introdução, são apresentados o enquadramento e a relevância do tema, bem como as principais questões e objetivos de investigação. Esta secção também aborda as opções metodológicas adotadas, as inovações e contributos esperados com o trabalho, além de fornecer uma visão geral da estrutura do documento.

O segundo capítulo, revisão bibliográfica, explora o estado da arte nos temas que sustentam o desenvolvimento deste trabalho. Neste tópico, são discutidos conceitos essenciais relacionados com a logística, planeamento e otimização de rotas, previsão de trajectos com base no comportamento do motorista e os principais algoritmos de *Machine Learning* utilizados. São também abordadas as métricas de avaliação empregues para validar os modelos desenvolvidos.

No terceiro capítulo sobre métodos e aplicação, são detalhadas as etapas seguidas ao longo do desenvolvimento do projeto, fundamentada na metodologia CRISP-DM. São explicadas as fases de compreensão do negócio, compreensão dos dados, preparação dos dados, modelação e avaliação, com o intuito de esclarecer o processo de construção dos modelos preditivos e as suas respetivas aplicações.

No quarto capítulo, são apresentados os resultados obtidos, permitindo uma análise quantitativa entre os diversos métodos de aprendizagem automática utilizados.

Por fim, o quinto capítulo resume as conclusões do trabalho e oferece uma perspetiva ampla sobre possíveis desenvolvimentos futuros

2. REVISÃO BIBLIOGRÁFICA

Neste capítulo será realizada uma abordagem referente a conceitos relevantes para compreensão do projeto. Sendo assim, estão inclusas informações sobre importância da logística, planeamento de rotas e otimização, definição sobre *Machine Learning* e seus modelos de regressão e classificação. Além disso, será abordado conceito de *Osquare*, métricas de avaliação e fundamentos matemáticos.

2.1. A Importância da Logística

A logística é frequentemente definida como o conjunto de atividades destinadas a garantir a entrega de mercadorias ao destino correto. No entanto, para além de transportar encomendas de um local para outro, a logística também é responsável por controlar o volume de mercadorias produzidas, monitorizar o stock e planear as rotas de distribuição. A missão da logística é assegurar o transporte de mercadorias nas condições desejadas em termos de tempo, custo e distância. Para que isso aconteça, a empresa deve dispor de uma infraestrutura adequada, com métodos eficientes que reduzam a probabilidade de gargalos durante o processo. Assim, a logística deve ser amplamente estudada pelas organizações, para que se torne um ponto estratégico na eliminação ou mitigação de problemas que surgem no transporte de mercadorias, garantindo um percurso mais otimizado e que gere menos custos operacionais (Nazário, 1999).

Segundo Moura (2006), a logística é concebida como uma estrutura que abrange atividades desde a esfera estratégica até ao nível mais operacional, evidenciando a relevância da administração de armazéns e da circulação de materiais para o desempenho logístico. No nível estratégico, o sector logístico desempenha um papel essencial no progresso empresarial, contribuindo de forma significativa para a obtenção de vantagens ao longo de todo o processo de produção. Neste contexto, tarefas primárias como a gestão de estoques, o processamento de pedidos e o transporte são fundamentais para a execução das funções logísticas. Já as tarefas secundárias, como a manutenção da informação, a armazenagem e o manuseamento de materiais, são classificadas como tarefas de suporte.

Uma sistematização logística com elevada fiabilidade operacional permite que uma empresa localizada numa determinada região consiga vender os seus produtos para outras localidades sem enfrentar grandes conflitos. Os processos logísticos que operam com alta eficiência apresentam uma estrutura económica bem organizada, aumentando as probabilidades de sucesso e reduzindo desperdícios desnecessários devido à falta de informação. Por isso, o setor logístico deve ser estudado e analisado com rigor, aplicando métodos como Inteligência Artificial e *Machine Learning* para otimizar rotas e outros componentes relacionados com a logística (Santos, 2012)

2.2. Planeamento de Rotas

O planeamento de rotas desempenha um papel crucial na logística e no transporte, focando na determinação dos percursos mais eficientes para a entrega de mercadorias. Com o aumento significativo dos pedidos de comércio eletrónico, os desafios associados ao transporte de última milha (o estágio final da entrega ao cliente) tornaram-se ainda mais evidentes. Este processo não visa apenas reduzir os custos operacionais, mas também melhorar a eficiência, a satisfação do cliente e a sustentabilidade ambiental (Silva, 2016).

Historicamente, o planeamento de rotas evoluiu de métodos rudimentares, que dependiam de mapas físicos e da experiência dos motoristas, para sistemas sofisticados que utilizam tecnologias como Sistemas de Informação Geográfica (SIG) e Sistemas de Posicionamento Global (GPS). Estas tecnologias permitem a visualização e análise de dados geográficos, bem como a localização em tempo real dos veículos, facilitando o planeamento dinâmico de rotas. Diversos algoritmos clássicos foram desenvolvidos para abordar o Problema de Roteamento de Veículos (PRV), que tem como objetivo identificar rotas com o custo mínimo. Nas últimas décadas, o número de variantes deste problema cresceu consideravelmente, impulsionado pela crescente demanda nas entregas de última milha. No contexto mais amplo, o planeamento de rotas é uma parte significativa deste conhecido problema de roteamento de veículos, conforme destacado por Toth e Vigo (2014). Cattaruzza et al. (2017) enfatizam que vários estudos sobre PRV incorporam janelas de tempo rígidas, e Savelsbergh e Van Woensel (2016) discutem a aplicabilidade prática desses métodos para aprimorar a eficiência na última milha.

A complexidade do planeamento de rotas é acentuada pela variabilidade no comportamento dos motoristas e pelas condições dinâmicas das vias urbanas. Diversos estudos têm incorporado perfis de disponibilidade do cliente (CAPs) para modelar a presença do cliente ao longo do período de entrega e melhorar as taxas de sucesso. Van Duin et al. (2016) demonstram que a eficiência na entrega está intimamente ligada às características demográficas de uma área, enquanto Florio et al. (2018) introduziram CAPs para modelar a presença do cliente. Ozarık et al. (2021) e Voigt et al. (2021) utilizam CAPs para resolver problemas de roteamento e agendamento para entregas de última milha com a presença incerta do cliente.

Conforme Cunha (2000), o roteamento é um mecanismo que permite não só a determinação do percurso, mas também o sequenciamento das localidades de paradas dispersas geograficamente que um veículo deve seguir. Segundo Novaes (2004), o processo de roteamento de veículos expandiu-se com a necessidade de melhorar o setor operacional empresarial, realizando a distribuição, entrega e coleta de mercadorias com alta eficiência. A aplicação desse conceito permite ao setor realizar operações com custos reduzidos, devido à minimização das distâncias percorridas entre pontos de atendimento, reduzindo gastos com manutenção, pneus, combustível e outros.

No entanto, nem sempre é possível otimizar diretamente em alguns casos, sendo necessário elaborar estratégias com base nos dados disponíveis. Oliveira (2017) identifica as variáveis que podem indicar se uma rota se desviará da sequência planejada e prever o grau de desvio em distância. Observa- se que rotas com um maior número de paragens têm maior probabilidade de desvio. Até o momento, nenhum estudo anterior se concentrou em métodos orientados por dados

para capturar o conhecimento implícito dos motoristas e, com base nesse conhecimento, melhorar o planeamento de rotas na última milha. Neste contexto, o problema em questão está mais intimamente relacionado à otimização inversa orientada por dados, cujo objetivo é derivar uma função objetivo que explique os dados observados.

Assim, pode-se afirmar que um roteamento eficiente envolve a alocação ótima de um conjunto de clientes em percursos, considerando as inúmeras restrições do problema e o objetivo de roteamento. Para alcançar um roteamento e programação de veículos fiáveis, é recomendável combinar as paragens para dias distintos, além de combinar as rotas de recolha com as de entrega. Novaes (2004) ressalta que quanto menor for o número de locais a serem atendidos, mais fácil é definir uma rota fiável. Por outro lado, se o número de destinos for elevado, o nível de dificuldade aumenta, exigindo métodos mais robustos. Para a resolução de alguns problemas envolvendo a roteirização, é possível aplicar conceitos de *Machine Learning* e inteligência artificial para aumentar a eficiência na resolução desses problemas.

2.3. Previsão de Rotas baseada no Comportamento do Motorista

A previsão de rotas baseada no comportamento do motorista é um dos elementos fundamentais na otimização das operações logísticas, especialmente quando se utiliza técnicas de *Machine Learning*. Este processo visa não apenas criar rotas eficientes, mas também adaptar essas rotas ao comportamento específico de cada motorista, reconhecendo que as decisões humanas e os padrões de condução podem variar significativamente (Miranda, 2018).

O primeiro passo para uma previsão eficaz de rotas é compreender que o comportamento do motorista desempenha um papel crucial na definição da melhor trajetória a seguir. O planeamento do motorista, neste contexto, envolve a análise de variáveis como preferências pessoais, histórico de condução, tempo de reação a diferentes cenários e até mesmo fatores emocionais que podem influenciar as decisões durante a condução. Esses dados, uma vez capturados e analisados, são utilizados para modelar o comportamento do motorista, permitindo prever as escolhas de rota que ele ou ela faria em situações específicas (Andriotti, 2004).

Ao desenvolver um modelo de previsão de rotas, é essencial estabelecer uma metodologia que inclua diferentes níveis de abstração, assegurando que tanto a movimentação do veículo quanto o planeamento das rotas sejam abordados de forma integrada. A movimentação refere-se à capacidade do motorista de conduzir e interagir com o ambiente, enquanto o planeamento envolve a escolha da rota que melhor atende aos objetivos definidos, como minimizar o tempo de viagem ou reduzir o consumo de combustível. Neste estudo, a previsão de rotas foi realizada utilizando algoritmos que incorporam *Machine Learning*, permitindo que o sistema aprenda e imite padrões de comportamento dos motoristas com base em dados históricos. A aplicação de técnicas como regressão e modelos de classificação permitiu prever com maior precisão as rotas que os motoristas provavelmente escolheriam, ajustando-se a diferentes condições de tráfego e preferências pessoais (Aruwajoye, 2016).

2.3.1. Método da Matriz

O método da matriz, frequentemente associado às Cadeias de Markov, é uma abordagem matemática utilizada para modelar sistemas dinâmicos nos quais o estado futuro depende exclusivamente do estado atual, e não do caminho que levou a esse estado. No contexto das Cadeias de Markov, o sistema é descrito por um conjunto de estados possíveis, e a transição entre esses estados ocorre com base em probabilidades definidas, organizadas numa matriz de transição. Esta matriz contém as probabilidades de transição de um estado para outro e permite prever o comportamento do sistema ao longo do tempo, a partir de um estado inicial (Ebling, 2012).

Por exemplo, ao modelar o comportamento de um motorista, pode-se usar uma Cadeia de Markov para representar diferentes estados de condução, como "a conduzir numa estrada", "parado no semáforo" ou "a mudar de faixa". A matriz de transição associada indicaria as probabilidades de um motorista passar de um estado para outro em um determinado intervalo de tempo. Aplicando essa matriz repetidamente, é possível simular a evolução do comportamento de condução ao longo do tempo. Uma das principais vantagens do método da matriz em Cadeias de Markov é a sua capacidade de simplificar a análise de sistemas complexos. Ao focar apenas no estado atual e nas transições imediatas, o método evita a necessidade de considerar toda a história passada do sistema, tornando as previsões mais eficientes e viáveis do ponto de vista computacional (Júnior, 2011).

Em suma, o método da matriz, pode oferecer uma ferramenta poderosa para modelar sistemas este analisar a evolução de estados ao longo do tempo. A sua aplicação pode ser importante para entender padrões de comportamento e realizar simulações precisas em diversos campos de estudo. Sendo assim, neste trabalho não foi aplicado a cadeia de Markov, mas uma ideia similar como matrizes baseadas em frequências de visitas para determinar o próximo ponto da rota a ser visitado de modo que se aproxime ao comportamento real do motorista.

2.3.2. Osquare

O método *Osquare Dispatch* é considerado uma abordagem inovadora para o gerenciamento e alocação de novas ordens em serviços de entrega. Neste método, "*M*" novas ordens são despachadas uma a uma. Para cada nova ordem, o sistema avalia a possibilidade de atribuí-la a cada um dos "*N*" entregadores disponíveis. Ao designar uma ordem a um entregador específico, a rota desse entregador é recalculada com base no novo conjunto de ordens. Assim, para cada nova ordem, o sistema analisa todas as rotas possíveis, recalculando a rota prevista com a inclusão da nova ordem e ajustando-a para considerar as novas paradas. A partir da rota prevista, é possível estimar a taxa de atraso do entregador, ou seja, a probabilidade de não cumprir os horários de entrega prometidos. Esta estimativa é crucial, pois uma alta taxa de atraso pode afetar negativamente a satisfação do cliente (Yan Zhang, 2019).

Além da taxa de atraso, o método calcula a distância adicional que o entregador terá de percorrer. Esta distância adicional é obtida subtraindo a distância original da rota do entregador da nova distância que inclui a nova ordem. Após avaliar as rotas de todos os *N* entregadores com a

nova ordem incluída, o sistema decide qual entregador deve receber a nova ordem com base na menor taxa de atraso prevista. Se duas ou mais opções apresentarem a mesma taxa de atraso, a decisão é feita considerando a menor distância adicional percorrida. A taxa de atraso é um dos critérios prioritários no processo de decisão, e o método *Osquare Dispatch* foca na minimização desta taxa ou de alguma outra taxa para garantir que as entregas sejam realizadas dentro dos prazos estipulados, aumentando a fiabilidade do serviço. Apenas quando várias opções têm a mesma taxa de atraso é que a distância adicional da jornada é usada como critério secundário (Genjian, 2019).

O método *Osquare Dispatch* oferece vários benefícios. Em primeiro lugar, melhora a eficiência operacional dos serviços de entrega ao reduzir a taxa de atraso e a distância adicional, resultando em menos tempo gasto em trânsito e menor consumo de combustível, o que reduz os custos operacionais. Em segundo lugar, contribui para a satisfação do cliente, uma vez que a redução da taxa de atraso melhora a fiabilidade e pontualidade das entregas. Entregas pontuais são cruciais para manter a confiança e fidelidade dos clientes, especialmente em setores onde a pontualidade é essencial. Em terceiro lugar, garante a utilização otimizada dos recursos ao distribuir as ordens de forma equilibrada entre os entregadores, evitando sobrecargas (Hao Zhang, 2019).

Em conclusão, o método *Osquare Dispatch* representa um avanço significativo na logística de entrega, oferecendo uma abordagem sistemática e eficiente para o gerenciamento de novas ordens. Ao priorizar a taxa de atraso e considerar a distância adicional da jornada, este método não só otimiza os processos internos, mas também melhora a experiência do cliente, tornando-se uma solução robusta para os desafios contemporâneos na logística de última milha. Nesse estudo foi implementada uma versão personalizada do método *Osquare*, incorporando atributos como distância, peso e capacidade do veículo. Além disso, foram aplicados modelos de *Machine Learning*, permitindo a previsão de novas rotas. Essa adaptação do método *Osquare* com *Machine Learning* aprimorou a precisão da previsão de rotas, integrando padrões de comportamento dos motoristas com base em frequência de visitas

Para exemplificar a aplicação do *Osquare* temos um caso da Tabela 1 utilizando a regressão linear como algoritmo de previsão de rotas. Cada linha da tabela refere-se a um par de origem e destino, com atributos que incluem a distância euclidiana entre os pontos, o peso da carga, a capacidade do veículo, o target real e o target previsto.

O target real é um valor binário que indica se a rota original (extraída da base de dados) inclui aquele destino como o próximo ponto a ser visitado. Quando o valor é "1", significa que aquele destino é o ponto correto da sequência da rota, de acordo com os dados históricos. Já o target é o valor previsto pelo modelo de regressão linear aplicado no *Osquare* desse exemplo, que tenta prever qual será o próximo ponto da rota com base nas variáveis fornecidas.

O modelo realiza uma ordenação dos destinos previstos de acordo com os valores de target, e a linha com o maior valor é destacada em amarelo, representando o próximo ponto a ser visitado na rota. Caso haja múltiplos destinos com a mesma origem e diferentes previsões para o próximo ponto, a escolha segue o maior valor de target. Em cenários onde há repetições, por exemplo, múltiplas origens apontando para o mesmo destino, o próximo maior valor é selecionado para dar continuidade ao processo de construção da rota. Nesse caso temos como rota: “ Armazém - Partida, 21004645, 21002724, 21002723, 21001040, 21001043, 21006376, 21006377, Armazém – Chegada”. Os resultados são: Média da *Accuracy*: 0.931, Desvio Padrão *Accuracy*: 0.0601, Média do

Kendall Tau: 0.1261, Desvio Padrão *Kendall* Tau: 0.0436, Média da *Edit Distance*: 85.8148, Desvio Padrão *Edit Distance*: 36.5423, Média Distância Total: 99.0276 km, Desvio Padrão Distância Total: 51.3259 km, Média Tempo Total: 1.6505 horas e Desvio Padrão Tempo Total: 0.8554 horas.

Tabela 1: *Osquare* com Regressão Linear

Origem	Destino	Distância Euclidiana	Peso	Capacidade	Target Real	Target
Armazém - Partida	21002723	0,9413	593,54	5500	1	-0,018212
Armazém - Partida	21002724	0,9428	583,29	5500	0	-0,018225
Armazém - Partida	21004645	0,9712	165,1	5500	0	-0,016753
Armazém - Partida	21006376	1,021	73,42	5500	0	-0,019378
Armazém - Partida	21006377	1,101	251,13	5500	0	-0,020733
Armazém - Partida	21001043	1,023	133,83	5500	0	-0,020032
Armazém - Partida	21001040	1,027	92,88	5500	0	-0,019953
21004645	21002723	0,1426	593,54	5500	0	0,035712
21004645	21002724	0,1432	583,29	5500	0	0,035752
21004645	21006376	0,2914	73,42	5500	1	0,029888
21004645	21006377	0,2918	251,13	5500	0	0,028421
21004645	21001043	0,2234	133,83	5500	0	0,033985
21004645	21001040	0,2241	92,88	5500	0	0,034274
21002724	21002723	0,001657	593,54	5500	0	0,045231
21002724	21004645	0,143282	165,1	5500	1	0,039147
21002724	21006376	0,43179	73,42	5500	0	0,020413
21002724	21006377	0,432035	251,13	5500	0	0,018954
21002724	21001043	0,365829	133,83	5500	0	0,024376
21002724	21001040	0,366672	92,88	5500	0	0,024651
21002723	21002724	0,001657	583,29	5500	1	0,045314
21002723	21004645	0,142647	165,1	5500	0	0,03919
21002723	21006376	0,430905	73,42	5500	0	0,020473
21002723	21006377	0,431146	251,13	5500	0	0,019014
21002723	21001043	0,365055	133,83	5500	0	0,024428
21002723	21001040	0,365912	92,88	5500	0	0,024702
21001040	21002723	0,365912	593,54	1200	0	0,049071
21001040	21002724	0,366672	583,29	1200	0	0,049103
21001040	21004645	0,224133	165,1	1200	0	0,062121
21001040	21006376	0,073631	73,42	1200	0	0,073026
21001040	21006377	0,074511	251,13	1200	0	0,071524
21001040	21001043	0,003811	133,83	1200	0	0,077249
21001043	21002723	0,365055	593,54	1200	0	0,049129
21001043	21002724	0,365829	583,29	1200	0	0,04916
21001043	21004645	0,223492	165,1	1200	0	0,062164
21001043	21006376	0,072587	73,42	1200	0	0,073096
21001043	21006377	0,073403	251,13	1200	0	0,071599
21001043	21001040	0,003811	92,88	1200	1	0,077582
21006376	21002723	0,430905	593,54	1200	0	0,044684
21006376	21002724	0,43179	583,29	1200	0	0,044707
21006376	21004645	0,291439	165,1	1200	0	0,057577
21006376	21006377	0,001531	251,13	1200	1	0,076451
21006376	21001043	0,072587	133,83	1200	0	0,072606
21006376	21001040	0,073631	92,88	1200	0	0,072868
21006377	Armazém Chegada					

Essa abordagem de ordenação e seleção permite que o modelo siga um fluxo baseado em probabilidades de visita previstas, ajustando-se de acordo com o comportamento dos dados

históricos e priorizando os destinos mais prováveis. A eficiência do algoritmo é avaliada ao comparar o target real com o target previsto, sendo possível identificar o quão precisa foi a previsão em relação ao comportamento real dos motoristas.

2.4. Machine Learning

Atualmente, o campo do Aprendizado de Máquina (*Machine Learning*) está em constante evolução, sendo caracterizado pela interseção entre técnicas estatísticas, inteligência artificial e ciência da computação (Jordan & Mitchell, 2015; Guido & Müller, 2016). O *Machine Learning* é reconhecido como um conjunto de técnicas computacionais que utilizam experiências passadas para melhorar o desempenho ou realizar previsões precisas, através do desenvolvimento de algoritmos de previsão eficientes (Mohri et al., 2014).

A essência do conceito reside na capacidade das máquinas de aprender e desenvolver comportamentos adaptativos sem necessidade de uma programação específica (Attaran & Deb, 2018). Dada a imensa quantidade de dados gerados atualmente, compreendê-los pode ser um desafio significativo. Nesse contexto, o *Machine Learning* aplica técnicas computacionais para criar algoritmos que identificam padrões nos dados, com o objetivo de extrair informações relevantes da complexidade e desorganização dos dados (Kashyap, 2017).

Os métodos de *Machine Learning* já são amplamente utilizados em vários aspectos do nosso quotidiano, como em filtros de spam, motores de busca na web, sistemas de recomendação, publicidade direcionada, detecção de fraudes, entre outros (Pedro, 2012). Além do *Machine Learning* a, existem várias modalidades de aprendizado, incluindo a aprendizagem supervisionada, a aprendizagem não supervisionada e a aprendizagem por reforço. Neste contexto, o presente trabalho explora a definição de Monard e Baranauskas (2003), que descrevem a aprendizagem de máquina como uma área da Inteligência Artificial (IA) dedicada ao desenvolvimento de técnicas computacionais para adquirir conhecimento automaticamente. Os sistemas de aprendizado são programas de computador que tomam decisões com base em experiências anteriores bem-sucedidas.

Figura 1: Hierarquia do *Machine Learning*
Fonte: Vilar, 2017

O *Machine Learning* envolve aprender com os dados para fazer previsões e tomar decisões. As principais categorias incluem a aprendizagem supervisionada, que utiliza dados rotulados; a aprendizagem não supervisionada, que trabalha com dados não rotulados; e a aprendizagem por reforço, que se baseia em feedbacks avaliativos sem sinais supervisionados. A Figura 1 contém exemplos de problemas de aprendizagem supervisionada (representados neste trabalho), que incluem a classificação e a regressão, que lidam com saídas categóricas e numéricas, respetivamente (Li, 2017).

2.4.1. Aprendizagem Supervisionada

A aprendizagem supervisionada, conforme descrito por Ludermir (2021), exige que, para cada exemplo apresentado ao algoritmo de aprendizado, seja fornecida a resposta desejada, ou seja, um rótulo que indica a classe à qual o exemplo pertence. Cada exemplo é representado por um vetor de atributos e o rótulo da classe correspondente. O objetivo é construir um classificador capaz de determinar corretamente a classe de novos exemplos ainda não rotulados. Este método é, de acordo com Ludermir (2021), o mais amplamente utilizado, e consiste em aprender a partir de um conjunto de exemplos rotulados fornecidos por um supervisor externo qualificado.

A aprendizagem supervisionada é frequentemente utilizada quando o objetivo é prever um resultado com base em dados de entrada. Para isso, são utilizados pares de entrada/saída como base para os modelos, que fazem parte do conjunto de dados de treinamento (Guido & Müller, 2016). O propósito é deduzir uma função com base nos dados conhecidos, onde uma variável alvo é prevista com base em um conjunto de variáveis independentes fornecidas (Mohammed et al., n.d.; Kashyap, 2017). Esse processo envolve a divisão do conjunto de dados em conjuntos de teste e de treino. O conjunto de treino contém a variável de saída conhecida, sobre a qual as técnicas de aprendizagem supervisionada são aplicadas para descobrir padrões. Esses padrões são então aplicados ao conjunto de dados de teste para realizar previsões (Batta, 2020).

Na aprendizagem supervisionada, as máquinas são encarregues de descobrir a relação entre os dados de entrada e saída (Attaran & Deb, 2018). As técnicas de aprendizagem podem ser categorizadas como classificação ou regressão, dependendo se as variáveis alvo são categóricas ou contínuas (Alzubi et al., 2018). É importante ressaltar que a aprendizagem supervisionada é essencial para que o sistema generalize suas respostas e atue corretamente em situações não presentes no conjunto de treinamento (Sutton & Barto, 2018).

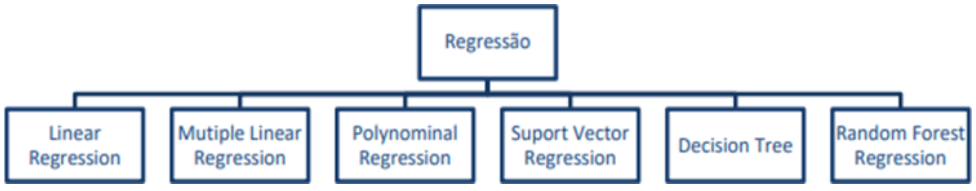

Figura 2: Algoritmos de *Machine Learning* de Regressão
Fonte: Italo (2018)

No contexto da regressão, lidamos com métodos que analisam as relações entre variáveis distintas (Swamynathan, 2017). O principal objetivo desses algoritmos é tratar de problemas que envolvem respostas contínuas ou numéricas (Alzubi et al., 2018). Eles são aplicados em situações como prever a temperatura para o próximo dia (Kashyap, 2017). A Figura 2 contém os principais algoritmos de regressão.

Os métodos de classificação envolvem a análise de respostas que possuem um valor fixo, geralmente conhecido de antemão, como "sim/não", "1/0", entre outros (Alzubi et al., 2018). Esses métodos podem ser classificados como binários, quando lidam com duas classes distintas, ou como multiclasse, quando diferenciam mais de duas classes (Guido & Müller, 2016). O principal objetivo da classificação é determinar a probabilidade de um novo conjunto de dados pertencer a uma classe específica (Swamynathan, 2017). São amplamente utilizados para categorizar elementos, como determinar se uma imagem retrata um ser humano ou uma máquina (Kashyap, 2017). A figura 3 ilustra os principais algoritmos de classificação.

Figura 3: Algoritmos de *Machine Learning* de Classificação
Fonte: Italo (2018)

2.4.2. Aprendizagem Não Supervisionada

A Aprendizagem não supervisionada é uma abordagem em que os algoritmos de *Machine Learning* são aplicados a conjuntos de dados nos quais as informações sobre as classes não são fornecidas previamente (Jung, 2022). Neste tipo de aprendizagem, apenas os dados de entrada são conhecidos, sem informações sobre os resultados desejados para o algoritmo (Guido & Müller, 2016). Assim, os algoritmos devem identificar padrões e estruturas nos dados sem uma variável alvo específica a ser prevista (Kashyap, 2017). A aprendizagem não supervisionada é útil em situações onde não há categorias de dados predefinidas, permitindo que os algoritmos descubram e explorem padrões nos dados (Alzubi et al., 2018).

Dentro desse contexto, existem três modelos principais de aprendizagem não supervisionada: clustering, redução de dimensionalidade e detecção de anomalias. No clustering, os algoritmos agrupam os exemplos com base em suas características semelhantes, formando clusters que representam grupos de dados semelhantes (Ludermir, 2021). Após a formação dos clusters, é necessário analisar e interpretar o significado de cada agrupamento dentro do contexto do problema em questão. Esta abordagem permite uma compreensão mais profunda da estrutura dos dados e das relações entre as diferentes instâncias, mesmo na ausência de uma orientação explícita sobre as categorias.

Os algoritmos de clustering têm como objetivo identificar estruturas nos dados e agrupar conjuntos com base na sua semelhança (Swamynathan, 2017). Uma vez identificados, os diferentes

clusters são rotulados para facilitar a interpretação e análise dos dados (Alzubi et al., 2018). Esses algoritmos são frequentemente utilizados em contextos práticos, como quando uma empresa de marketing deseja segmentar diferentes grupos de clientes em distintos segmentos de mercado (Kashyap, 2017).

No que diz respeito aos algoritmos de redução de dimensionalidade, estes transformam uma representação inicial dos dados em uma representação de menor dimensão, preservando certas características da representação original (Mohri et al., 2014). Este método é comumente utilizado em tarefas de visão por computador, especialmente para processar imagens (Mohri et al., 2014).

Os algoritmos de detecção de anomalias, como o próprio nome indica, são projetados para identificar elementos que se desviam de um padrão estabelecido nos dados (Alzubi et al., 2018). Eles procuram elementos que não se conformam ao comportamento ou padrão esperadocom base no conjunto de dados disponível (Swamynathan, 2017). Esses algoritmos são valiosos em cenários como o setor de cartões de crédito, onde são usados para detectar possíveis fraudes com base nos padrões de transações dos clientes (Alzubi et al., 2018).

O modelo de associação é utilizado para identificar relações frequentes entre itens em grandes conjuntos de dados. Esses algoritmos, como as Regras de Associação, são amplamente empregados em análise de cestas de compras, onde o objetivo é encontrar padrões que indicam a probabilidade de itens serem comprados juntos. Ao descobrir essas associações, as empresas podem tomar decisões mais informadas sobre recomendações de produtos e ofertas promocionais (Ludermir, 2021). A mineração de regras de associação é especialmente útil em contextos de e-commerce, onde a identificação de padrões de compra pode melhorar as estratégias de marketing e aumentar as vendas.

Por fim, os algoritmos de sumarização têm como objetivo criar representações compactas dos dados, permitindo uma descrição condensada e informativa do conjunto de dados. Este tipo de modelo é utilizado para extrair insights chave a partir de grandes volumes de dados, facilitando a compreensão e visualização de padrões complexos. A sumarização é amplamente aplicada em análise de texto, onde grandes documentos são reduzidos a resumos breves que ainda contêm as informações mais relevantes (Swamynathan, 2017). Isso é particularmente útil em áreas como jornalismo, serviços de notícias e sistemas de recomendação de conteúdo

2.4.3. Random Forest

Segundo Rigatti (2017), o *Random Forest* é um algoritmo de *Machine Learning* poderoso e versátil, amplamente utilizado para tarefas tanto de classificação quanto de regressão. Desenvolvido por Leo Breiman em 2001, combina o conceito de "*bagging*" (*bootstrap aggregating*) com uma aleatoriedade adicional no processo de construção das árvores, o que aumenta a robustez e a precisão do modelo. Para compreender o *Random Forest*, é fundamental começar com as árvores de decisão, que são estruturas hierárquicas que dividem iterativamente os dados de entrada em subconjuntos com base em critérios de divisão. Cada nó da árvore representa uma decisão baseada em um atributo, e as folhas da árvore representam as previsões finais. Embora as

árvores de decisão sejam fáceis de interpretar e rápidas de treinar, elas têm a desvantagem de serem suscetíveis ao *overfitting*, especialmente quando são profundas e complexas.

O *Random Forest* utiliza o conceito de *ensemble learning*, que combina os resultados de vários modelos fracos (neste caso, árvores de decisão) para formar um modelo mais forte e robusto. A técnica específica de ensemble utilizada no *Random Forest* é o *bagging*, que envolve a criação de várias versões diferentes do conjunto de dados original através de amostragem com reposição. Segundo Biau (2016), cada árvore de decisão no *Random Forest* é treinada em um desses subconjuntos, e como resultado, cada árvore pode ser ligeiramente diferente, pois os dados de treinamento variam ligeiramente de uma árvore para outra. Após o treinamento, as previsões de todas as árvores são agregadas: no caso de classificação, por meio de votação majoritária, e no caso de regressão, através da média das previsões individuais.

Além da variação introduzida pelo *bagging*, o *Random Forest* adiciona outra camada de aleatoriedade durante o processo de construção das árvores. Em vez de considerar todas as características possíveis para determinar o melhor ponto de divisão em cada nó, o algoritmo seleciona aleatoriamente um subconjunto de características. Esse procedimento impede que as árvores se tornem muito semelhantes entre si e melhora a generalização do modelo, reduzindo o risco de overfitting (Bonissone, 2010).

As principais vantagens do *Random Forest* incluem a sua robustez ao *overfitting*, a sua versatilidade para diferentes tipos de problemas e dados, a sua estabilidade frente a ruídos, a capacidade de fornecer estimativas da importância relativa de cada característica e a sua escalabilidade para grandes conjuntos de dados. No entanto, também apresenta algumas desvantagens, como a necessidade de mais recursos computacionais e memória em comparação com uma única árvore de decisão, além da menor interpretabilidade, dado que a previsão final resulta da combinação de múltiplas árvores. Além disso, a seleção dos hiperparâmetros, como o número de árvores na floresta e a profundidade máxima das árvores, é crucial para o desempenho do modelo (Biau, 2016).

Em suma, o *Random Forest* é uma ferramenta poderosa que oferece uma combinação de alta precisão, robustez e capacidade de lidar com dados complexos. O seu uso é difundido em diversas áreas, refletindo a sua eficácia e adaptabilidade. No entanto, como qualquer modelo, requer uma compreensão profunda dos seus princípios e limitações para garantir uma aplicação correta e eficaz.

2.4.4. Redes Neurais

Conforme Neis (2019), as redes neurais são modelos computacionais inspirados no funcionamento do cérebro humano, projetados para reconhecer padrões complexos e realizar tarefas como classificação, regressão, reconhecimento de imagens, processamento de linguagem natural, entre outras. O conceito fundamental por trás de uma rede neural é um conjunto de unidades interconectadas, chamadas neurônios, organizadas em camadas. Cada neurônio processa informações e as transmite para os neurônios da camada seguinte.

Uma rede neural típica é composta por três tipos principais de camadas: a camada de entrada, as camadas ocultas e a camada de saída. A camada de entrada recebe os dados brutos, como pixels

de uma imagem ou valores numéricos em um conjunto de dados. As camadas ocultas, que podem variar em número, são responsáveis por processar as informações, realizando operações matemáticas para extrair características e aprender padrões. Por fim, a camada de saída gera a previsão ou o resultado, como uma classificação ou uma regressão (Osório, 1999).

Cada neurônio em uma camada está conectado a neurônios da camada seguinte através de pesos, que determinam a importância de cada conexão. Esses pesos são ajustados durante o treinamento da rede, normalmente utilizando um algoritmo chamado *backpropagation* em combinação com métodos de otimização, como o gradiente descendente. Durante o treinamento, a rede neural ajusta iterativamente os pesos para minimizar a diferença entre suas previsões e os valores reais, medidos por uma função de perda. O *backpropagation* calcula o gradiente da função de perda em relação a cada peso, permitindo que o algoritmo faça pequenas atualizações para melhorar a precisão do modelo (França, 2020).

Um conceito importante em redes neurais é o de funções de ativação, que são aplicadas à saída de cada neurônio antes de ser passada para a próxima camada. Segundo Almeida (2019), as funções de ativação introduzem não-linearidades no modelo, permitindo que a rede aprenda e represente relações complexas nos dados. Funções de ativação comuns incluem a sigmoide, a ReLU (Rectified Linear Unit) e a tangente hiperbólica (tanh).

As redes neurais podem ser divididas em diferentes tipos, dependendo da arquitetura e do problema que estão resolvendo. As redes neurais *feedforward*, por exemplo, são as mais básicas, onde os dados fluem em uma única direção, da camada de entrada para a camada de saída. Já as redes neurais convolucionais (CNNs) são amplamente utilizadas em tarefas de visão computacional, como reconhecimento de imagens, e são projetadas para capturar hierarquias de características espaciais em dados bidimensionais, como imagens. As redes neurais recorrentes (RNNs), por outro lado, são usadas em problemas de sequências temporais, como séries temporais e processamento de linguagem natural, devido à sua capacidade de manter informações contextuais através de ciclos internos de retroalimentação (Osório, 1999).

Nos últimos anos, as redes neurais profundas, ou *deep learning*, têm se destacado, compostas por várias camadas ocultas que permitem a aprendizagem de representações de dados em múltiplos níveis de abstração. Esse avanço tem sido fundamental em muitas inovações recentes, como veículos autónomos, tradução automática e diagnósticos médicos assistidos por IA. No entanto, redes neurais também apresentam limitações e desafios. Elas podem ser difíceis de treinar devido à necessidade de grandes quantidades de dados e poder computacional, além de serem suscetíveis ao *overfitting*, especialmente em conjuntos de dados pequenos. A interpretabilidade das redes neurais é outra questão importante, uma vez que os modelos tendem a ser vistos como "caixas-pretas", dificultando a explicação de como chegam a certas decisões (França, 2020).

Portanto, as redes neurais são uma classe poderosa e flexível de modelos de *Machine Learning*, capazes de resolver uma ampla gama de problemas complexos. No entanto, seu uso eficaz requer um entendimento profundo da sua arquitetura, técnicas de treinamento e desafios associados. À medida que a pesquisa em inteligência artificial avança, as redes neurais continuam a evoluir, abrindo novas possibilidades para aplicações inovadoras em diversas áreas

2.4.5. Regressão

A regressão é uma técnica fundamental no campo do *Machine Learning*, especialmente voltada para a análise e modelagem de relações entre variáveis. O principal objetivo da regressão é prever uma variável contínua, conhecida como variável dependente, a partir de uma ou mais variáveis independentes ou preditoras. Diferente de técnicas de classificação, que tentam atribuir rótulos discretos a partir de dados de entrada, a regressão busca prever valores numéricos contínuos, como no caso da porcentagem de frequência que um ponto é visitado a seguir na rota. (Lima, 2022)

A regressão em *Machine Learning* é uma técnica amplamente utilizada para modelar a relação entre uma variável dependente y e uma ou mais variáveis independentes X, sendo fundamental para prever valores contínuos. A forma mais básica dessa abordagem é a regressão linear simples, onde se assume uma relação linear entre as variáveis. A equação da regressão linear simples pode ser expressa como $y = \beta_0 + \beta_1 x + \varepsilon$, onde y é a variável alvo que queremos prever, x é a variável independente, β_0 é o intercepto (o valor de y quando x=0), β_1 é o coeficiente angular (a taxa de mudança de y por unidade de x), e ε é o termo de erro, representando a diferença entre o valor previsto e o valor real (Lima, 2022).

No caso da regressão linear múltipla, a abordagem é expandida para incluir várias variáveis preditoras. A fórmula se torna $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon\ (1)$, onde $x_1, x_2, \ldots, x_n$ são as diferentes variáveis independentes. Cada coeficiente β_i representa a contribuição da respectiva variável x_i para a previsão de y, enquanto o termo de erro ε continua a captar as variações que o modelo não consegue explicar. (Lima, 2022)

Além da regressão linear, também existem técnicas mais avançadas, como a Regressão por Vetor de Suporte (SVR), que utiliza margens para lidar com *outliers*, e a regressão baseada em árvores de decisão, onde o espaço de características é dividido de forma hierárquica, criando previsões baseadas em médias locais dos dados. Outra técnica muito usada é o *Random Forest Regressor*, que constroi uma série de árvores de decisão aleatórias e combina os resultados para aumentar a precisão das previsões (Okamura, 2019).

A principal vantagem dessas abordagens avançadas é a capacidade de lidar com relações não lineares e dados complexos. No entanto, como qualquer modelo de *Machine Learning*, os modelos de regressão precisam ser cuidadosamente ajustados para evitar o problema de *overfitting*, que ocorre quando o modelo se ajusta muito bem aos dados de treinamento, mas não generaliza bem para novos dados (Okamura, 2019).

Por fim, a regressão é uma técnica poderosa que, ao ser combinada com métodos avançados e otimização de parâmetros, pode fornecer previsões precisas e úteis para uma ampla gama de aplicações, desde finanças e economia até previsão de demanda e análise de comportamento de consumidores (Lima, 2022).

2.4.6. Classificação

A classificação é uma técnica central em *Machine Learning*, usada para prever categorias ou rótulos de dados com base em variáveis independentes. Ao contrário da regressão, que visa prever valores contínuos, a classificação lida com variáveis categóricas, como tipos de objetos, grupos de clientes ou resultados binários (como "sim" ou "não"). A ideia central da classificação é encontrar um modelo capaz de separar ou distinguir diferentes classes a partir dos atributos dos dados, permitindo assim que novas observações sejam corretamente rotuladas (Laydner, 2022).

A classificação supervisionada é o tipo mais comum, em que o modelo é treinado usando dados rotulados, ou seja, dados em que a classe correta já foi atribuída a cada exemplo. O objetivo do modelo é aprender a mapear os atributos de entrada para a classe correspondente, generalizando essa relação de maneira que ele possa prever corretamente a classe de novos dados. Entre os algoritmos mais conhecidos de classificação estão a regressão logística, *k-nearest neighbors* (k-NN), máquinas de vetores de suporte (SVM), árvores de decisão, e *Random Forest*, além de métodos baseados em redes neurais, como as redes neurais artificiais e as redes neurais profundas (Teixeira, 2022).

Para Fuhr (2022) a regressão logística é um dos modelos mais simples e amplamente utilizados para classificação binária (dois rótulos possíveis). Ela calcula a probabilidade de uma instância pertencer a uma das classes, utilizando a função sigmoide para mapear uma combinação linear dos atributos de entrada para um valor entre 0 e 1. O modelo pode ser representado pela fórmula:

$$P(x) = \frac{1}{1+e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \quad (2)$$

onde P(x) é a probabilidade de a classe y ser 1, x1,…,xn são os atributos de entrada, e β0,…,βn são os coeficientes ajustáveis do modelo.

Em cenários mais complexos, algoritmos como o *Random Forest* e as redes neurais profundas podem ser aplicados. O *Random Forest* é uma técnica de *ensemble learning* que cria múltiplas árvores de decisão em subconjuntos diferentes dos dados, combinando os resultados de cada árvore para melhorar a precisão e a robustez do modelo. Já as redes neurais profundas, com múltiplas camadas ocultas, são capazes de capturar padrões altamente complexos e não lineares nos dados, o que as torna particularmente úteis em problemas como classificação de imagens e reconhecimento de fala (Teixeira, 2022).

Em resumo, a classificação é uma técnica poderosa e versátil em *Machine Learning*, utilizada para resolver uma ampla variedade de problemas de categorização. A escolha do algoritmo de classificação adequado depende das características dos dados e do tipo de problema, mas, com uma seleção apropriada, os modelos de classificação podem proporcionar previsões precisas e insights valiosos (Laydner, 2022). la (Teixeira, 2022).

2.4.7. Imitation Learning

A aprendizagem por imitação (*Imitation Learning* ou *IL*) é uma técnica inspirada na forma como humanos e animais adquirem habilidades: observando as ações de outros e tentando replicá-las. Este conceito tem ganho cada vez mais relevância no campo da inteligência artificial (IA), uma vez que oferece uma forma eficiente de treinar máquinas para realizar tarefas complexas sem necessidade de programação explícita. O IL distingue-se entre as diversas abordagens de *Machine Learning*, como a aprendizagem supervisionada, não supervisionado e por reforço. Enquanto o aprendizado supervisionado envolve treinar algoritmos com dados rotulados, onde um humano indica se a decisão foi correta ou não, o aprendizado não supervisionado trabalha com dados não rotulados, categorizando e segmentando informações com base em padrões descobertos pelos próprios algoritmos. Já o aprendizado por reforço recompensa decisões que levam a bons resultados, incentivando a maximização da coleta de recompensas (Hussein et al., 2017).

No aprendizado por imitação, o foco está em ensinar máquinas a replicar ações realizadas por um especialista. Neste processo, um agente — que pode ser um robô ou um software — observa um demonstrador humano executando uma tarefa e, a partir dessa observação, aprende a mapear estados para ações. O objetivo é que o agente desenvolva uma política que lhe permita tomar decisões em situações semelhantes às observadas, executando a tarefa de forma eficiente. A política, nesse contexto, é uma função que mapeia os estados do ambiente para as ações que o agente deve realizar, considerando características como posição, velocidade e contexto do entorno. O processo de IL envolve mais do que simplesmente copiar as ações observadas. Embora a imitação seja o ponto de partida, o agente deve ser capaz de adaptar-se a mudanças no ambiente, como variações nas condições climáticas ou obstáculos inesperados, ajustando sua política conforme necessário. Isso requer uma etapa adicional de reotimização, onde o agente refina a política aprendida com base no desempenho ao longo do tempo (Ho & Ermon, 2016).

Uma das grandes vantagens do IL é a redução do "espaço de busca" — ou seja, o número de variáveis e possibilidades que o agente precisa considerar ao procurar a solução ótima para um problema. Em vez de depender de um processo intensivo de tentativa e erro, como ocorre em outras formas de aprendizado, o IL permite que o agente aprenda com exemplos diretos, resultando em uma necessidade menor de poder computacional. Além disso, o IL pode ser particularmente útil em cenários onde não há disponibilidade de grandes volumes de dados para treinamento (Ho & Ermon, 2016). Existem diferentes métodos de implementar o aprendizado por imitação. A clonagem comportamental é uma técnica supervisionada onde o agente tenta replicar o comportamento observado do especialista. O aprendizado direto de políticas é um método iterativo em que o agente pode interagir com o demonstrador durante o treinamento, em vez de apenas observar. Já o aprendizado por reforço inverso combina a observação de demonstrações com uma função de recompensa, treinando o algoritmo para replicar os resultados obtidos pelo especialista.

O IL é especialmente valioso em aplicações onde a explicabilidade da IA é crucial, como na área da saúde ou em recursos humanos, pois permite que os humanos compreendam mais facilmente como a máquina chegou a uma solução. Além disso, essa técnica é promissora para a programação de robôs e dispositivos automatizados que operam em ambientes com um alto grau de liberdade,

como carros voadores ou robôs domésticos, onde a quantidade de variáveis a serem consideradas é significativamente maior (Hussein et al., 2017).

Em resumo, o aprendizado por imitação oferece uma abordagem poderosa e eficiente para ensinar máquinas a realizar uma ampla gama de tarefas, potencialmente avançando o desenvolvimento da IA generalizada, também conhecida como "IA forte", que se assemelha à flexibilidade e capacidade do cérebro humano de se aplicar a diversas tarefas (Osa et al., 2018).

2.5. Métricas

As métricas de *Machine Learning* são ferramentas essenciais para avaliar o desempenho e a eficácia dos modelos preditivos. Estas métricas fornecem informações valiosas sobre a precisão dos modelos e auxiliam na comparação entre diferentes abordagens ou algoritmos. A escolha das métricas adequadas é crucial para garantir que o modelo cumpra os objetivos específicos da tarefa e para identificar áreas que necessitam de melhoria (Silva, 2022).

2.5.1 *Kendall*

A métrica de *Kendall*, também conhecida como *Tau de Kendall*, é uma medida estatística amplamente utilizada para avaliar a concordância entre duas ordens ou classificações. No contexto da previsão de rotas e modelagem do comportamento dos condutores, a métrica de *Kendall* é particularmente importante, pois permite comparar a ordem das rotas previstas por um modelo com a ordem real das rotas escolhidas pelos motoristas (Santos, 2021).

A métrica de *Kendall* baseia-se na análise de concordância e discordância entre pares de elementos. Quando se têm duas classificações de um mesmo conjunto de elementos, a métrica considera todos os pares possíveis de elementos e verifica se as ordens relativas desses pares são as mesmas nas duas classificações. Se a ordem de um par for idêntica em ambas as classificações, esse par é considerado concordante; caso contrário, é considerado discordante.

A fórmula da métrica de *Kendall* é expressa como:

$$\tau = \frac{(\text{Números de pares concordantes}) - (\text{Números de pares discordantes})}{n\,(n-1)/\,2} \tag{3}$$

Onde:

- n é o número total de elementos;
- O denominador, n(n−1)/2, representa o número total de pares possíveis.

O valor de τ varia entre -1 e 1, onde:

- τ=1 indica uma concordância perfeita entre as duas classificações;

- τ=−1 indica uma discordância total;
- τ=0 sugere que não há associação entre as classificações.

No contexto da previsão de rotas baseada no comportamento do motorista, a métrica de *Kendall* é crucial para avaliar a precisão dos modelos preditivos. Por exemplo, quando um modelo de *Machine Learning* prevê a ordem preferencial das rotas que um motorista pode escolher, a métrica de *Kendall* é utilizada para comparar essa previsão com a ordem real das rotas selecionadas pelo motorista em situações reais. Uma alta concordância, indicada por um valor de τ próximo de 1, sugere que o modelo está capturando com precisão as preferências do motorista, enquanto um valor mais baixo indica a necessidade de ajustes no modelo (Silva, 2022).

Além disso, conforme Silva (2022), a métrica de *Kendall* é robusta a pequenos desvios e é particularmente útil em cenários onde a ordem relativa das previsões é mais importante do que os valores absolutos. Isso a torna especialmente adequada para aplicações em que as decisões são baseadas em classificações ou rankings, como no planejamento de rotas, onde o objetivo é prever a sequência de ações de um motorista.

Em suma, a métrica de *Kendall* é uma ferramenta valiosa para aprimorar a compreensão do desempenho de modelos preditivos. Ela permite que pesquisadores e engenheiros avaliem e ajustem seus algoritmos para refletir melhor o comportamento humano, o que é essencial em aplicações que envolvem a interação direta com motoristas e a adaptação às suas preferências individuais (Santos, 2021).

2.5.2 *Accuracy*

A métrica *Accuracy* (ou precisão, em português) é uma das medidas de desempenho mais amplamente utilizadas em modelos de classificação, sendo especialmente relevante em contextos onde se pretende avaliar a proporção de previsões corretas feitas por um modelo em relação ao total de previsões. No contexto da previsão de rotas com base no comportamento do motorista, a *Accuracy* é fundamental para determinar a eficácia de um modelo na previsão precisa das escolhas de rota de um motorista (Gómez, 2020).

A *Accuracy* é calculada como a razão entre o número de previsões corretas e o número total de previsões realizadas. A fórmula é representada por:

$$Accuracy = \frac{Número\ de\ Previsões\ Corretas}{Número\ total\ de\ Previsões} \tag{4}$$

O valor resultante varia de 0 a 1, sendo que:

- Um valor de 1 indica que o modelo acertou todas as previsões, correspondendo a uma *Accuracy* de 100%;
- Um valor de 0 indica que o modelo falhou em todas as previsões.

No contexto da previsão de rotas, a *Accuracy* é utilizada para avaliar a capacidade do modelo em prever corretamente a rota específica escolhida pelo motorista entre várias opções possíveis. Por exemplo, se um modelo prevê que um motorista escolherá a Rota A em um dado cenário, e o motorista de fato escolhe a Rota A, essa previsão é considerada correta e contribui positivamente para a métrica de *Accuracy* (Gómez, 2020).

Uma *Accuracy* elevada indica que o modelo está alinhado com as decisões reais dos motoristas, o que é crucial para o desenvolvimento de sistemas preditivos confiáveis. No entanto, é importante considerar o contexto em que a *Accuracy* é aplicada. Em situações onde as classes (ou rotas, neste caso) são desbalanceadas, ou seja, algumas rotas são escolhidas com muito mais frequência do que outras, a *Accuracy* pode ser uma métrica enganadora. Por exemplo, se a maioria dos motoristas escolhe uma rota específica em 90% das vezes, um modelo que sempre prevê essa rota terá uma *Accuracy* alta, mas pode não ser útil para prever corretamente as escolhas nos 10% restantes, onde outras rotas são preferidas (Silva, 2022).

Portanto, ao usar a *Accuracy* como métrica de avaliação, é essencial considerar a distribuição das escolhas de rotas e, se necessário, complementar a análise com outras métricas, como *Precision*, *Recall* ou *F1-score*, para obter uma visão mais completa do desempenho do modelo. Em suma, a *Accuracy* é uma métrica importante para avaliar a precisão geral de um modelo preditivo no contexto da previsão de rotas, oferecendo uma visão inicial sobre a capacidade do modelo de refletir o comportamento real dos motoristas. Contudo, deve ser usada com cautela, especialmente em cenários onde o equilíbrio entre classes pode influenciar indevidamente a interpretação dos resultados (Silva, 2022).

2.5.3 *Edit Distance*

A métrica *Edit Distance*, também conhecida como distância de edição, é uma ferramenta fundamental na análise das diferenças entre duas sequências de texto, como palavras, frases ou até mesmo rotas (Medina, 2023). Esta métrica calcula o número mínimo de operações necessárias para transformar uma sequência em outra, considerando operações básicas como inserções, eliminações e substituições de caracteres.

O cálculo do *Edit Distance* é baseado no algoritmo de Levenshtein, que avalia a distância entre duas cadeias de caracteres. A distância de edição é definida como a menor quantidade de operações necessárias para converter uma sequência na outra, sendo o cálculo realizado através de uma abordagem dinâmica. O algoritmo utiliza uma matriz onde cada célula representa a distância de edição entre os prefixos das duas sequências. A célula (i, j) da matriz contém o custo mínimo para transformar os primeiros i caracteres da primeira sequência nos primeiros j caracteres da segunda sequência. O valor de cada célula é determinado com base nas seguintes operações: se os caracteres nas posições i e j são iguais, o custo é o mesmo que o valor na célula (i−1, j−1); se forem diferentes, o custo é o menor entre as opções de eliminar, inserir ou substituir um caractere (Marinho, 2023).

A fórmula para calcular a distância de edição é expressa por:

$$(5)\ D(i,j) = min\ (D(i-1,j)+1, D(i,j-1)+1, D(i-1,j-1)+cost) \quad (5)$$

Onde:

- D(i,j): A distância de edição entre os primeiros i caracteres da primeira sequência e os primeiros j caracteres da segunda sequência;
- D(i-1,j) + 1: Custo de deletar um caractere;
- D(i,j-1) + 1: Custo de inserir um caractere;
- D(i-1,j-1) + {*cost*}: Custo de substituir um caractere, onde *cost* é 0 se os caracteres são iguais e 1 se são diferentes.

No contexto da previsão de rotas com base no comportamento do motorista, a métrica *Edit Distance* é particularmente útil para medir a similaridade entre as rotas previstas por um modelo e as rotas efetivamente seguidas pelos motoristas. Ao comparar a sequência de rotas previstas com as rotas reais, a Distância de Edição fornece uma medida quantitativa da precisão do modelo. Por exemplo, se um modelo prevê uma rota que difere ligeiramente da rota real escolhida devido a pequenas variações, o *Edit Distance* ajuda a quantificar o número de ajustes necessários para transformar a previsão do modelo na rota real, permitindo uma avaliação detalhada da eficácia do modelo em capturar as escolhas de rota reais (Marinho, 2023).

Embora o *Edit Distance* forneça uma visão clara das diferenças entre sequências, pode ser sensível a pequenas variações e não capturar bem as semelhanças contextuais mais complexas.

Portanto, é vantajoso combinar *o Edit Distance* com outras métricas, como *Accuracy* ou o Tau de *Kendall*, para obter uma avaliação mais completa do desempenho do modelo de previsão de rotas. Em resumo, a métrica *Edit Distance* é uma ferramenta valiosa para quantificar discrepâncias entre previsões e resultados reais, oferecendo insights importantes para melhorar a precisão dos modelos preditivos (Medina, 2023).

2.6. Trabalhos Relacionados

A aplicação de técnicas de *Machine Learning* na previsão de rotas e na otimização de processos logísticos tem vindo a ser um campo de estudo crescente, refletindo a necessidade crescente de soluções eficientes e adaptativas num cenário logístico complexo. Diversos estudos têm explorado diferentes abordagens e modelos para enfrentar estes desafios, proporcionando uma base sólida sobre a qual este trabalho se apoia e avança.

Diversos investigadores têm abordado a previsão de rotas utilizando técnicas de *Machine Learning*. O trabalho de Alzubi (2018) investigou a utilização de Redes Neurais Artificiais (ANNs) para prever rotas de transporte com base em dados históricos de tráfego e características das rotas. Os resultados demonstraram a eficácia das ANNs em prever rotas com alta precisão, embora o estudo tenha sido limitado à análise de dados de tráfego urbano. Em contraste, Attaran et al. (2018) aplicaram técnicas de *Random Forest* para prever a procura de transporte e otimizar rotas de entrega. A pesquisa mostrou que o modelo *Random Forest* era particularmente eficaz em lidar com

grandes volumes de dados e variáveis complexas, apresentando um bom equilíbrio entre precisão e interpretabilidade. Este estudo está alinhado com os resultados encontrados neste trabalho, onde o modelo *Random Forest* também demonstrou um bom desempenho, especialmente quando otimizado.

A consideração do comportamento do condutor na previsão de rotas é um aspeto crucial e tem recebido atenção significativa. O trabalho de Campos et al. (2022) explora a influência do comportamento do condutor nas escolhas de rotas utilizando *Support Vector Machines* (SVM). Os autores identificaram que, embora os SVMs possam captar padrões comportamentais complexos, a necessidade de ajuste fino dos parâmetros e a interpretação dos resultados representaram desafios significativos. Além disso, Chen et al. (2019) realizaram um estudo que integrou técnicas de *Machine Learning* e análise comportamental para melhorar a previsão de rotas. A pesquisa utilizou uma combinação de algoritmos de regressão e redes neurais para analisar como diferentes fatores comportamentais influenciam as escolhas de rotas dos condutores. O estudo sublinhou a importância da personalização dos modelos para captar adequadamente as nuances do comportamento do condutor.

A escolha das métricas para avaliar a eficácia dos modelos é essencial. Andriotti (2004) propuseram uma abordagem integrada para a avaliação de modelos de previsão de rotas, utilizando métricas como Acurácia, *Kendall's Tau* e *Edit Distance*. A pesquisa destacou a importância de utilizar múltiplas métricas para obter uma visão abrangente do desempenho do modelo, uma abordagem que foi seguida neste trabalho para garantir uma avaliação robusta e detalhada dos modelos aplicados.

Os trabalhos anteriores fornecem uma base sólida para a pesquisa em previsão de rotas, com várias técnicas a demonstrar eficácia em diferentes contextos. No entanto, muitos estudos têm limitações, como a falta de generalização dos modelos para diferentes cenários logísticos ou a necessidade de ajustes específicos para melhorar o desempenho. Este trabalho contribui para a área ao integrar e comparar diversos modelos de *Machine Learning* num contexto logístico específico, utilizando uma base de dados real e abrangente, e oferecendo insights para futuras melhorias na previsão e otimização de rotas.

3. MÉTODOS E APLICAÇÃO

Para este projeto, foi adotada a metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM), representado na Figura 4, um modelo amplamente reconhecido e eficaz para a estruturação e execução de projetos de *Machine Learning*. O CRISP-DM divide o processo de mineração de dados em seis etapas principais: Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implantação. Cada uma dessas etapas foi cuidadosamente adaptada para atender às necessidades específicas do estudo sobre a aplicação de *Machine Learning* na previsão de rotas com base no comportamento dos motoristas.

Na fase de *Business Understanding*, o projeto iniciou-se com a definição clara dos objetivos e requisitos da empresa de logística. A meta era utilizar técnicas avançadas de *Machine Learning* para melhorar a eficiência operacional da empresa, focando na previsão precisa das rotas com base nas informações sobre o comportamento dos motoristas. Esta etapa envolveu identificar os problemas a serem resolvidos e os critérios de sucesso do projeto, garantindo que o estudo estivesse alinhado com as metas da empresa.

Na fase de *Data Understanding*, foram analisados os dados fornecidos pela empresa Compal, armazenados numa base de dados Excel. A análise inicial concentrou-se na avaliação da qualidade dos dados, identificação de padrões e verificação de possíveis problemas, como dados ausentes ou inconsistências. Compreender os dados em profundidade foi crucial para garantir que estivessem adequados para a modelagem subsequente.

Na etapa de *Data Preparation*, os dados foram limpos e transformados para estarem prontos para a modelagem. Isso incluiu o tratamento de dados ausentes, correção de erros e aplicação de técnicas de normalização e transformação. Além disso, os dados foram divididos em conjuntos de treino e teste, preparando-os para a construção e avaliação dos modelos de *Machine Learning*.

Durante a fase de *Modeling*, foram explorados diversos algoritmos de *Machine Learning* para prever as rotas com base no comportamento dos motoristas. Utilizando o *Visual Studio Code* e codificação em *Python*, foram aplicados modelos de regressão, técnicas de classificação e métodos de aprendizado por imitação. Esta etapa envolveu a experimentação com diferentes algoritmos e a definição de parâmetros para identificar o modelo que oferecesse o melhor desempenho em termos de precisão e eficiência.

Na fase de *Evaluation*, os modelos construídos foram avaliados com base em métricas como acurácia, precisão, recall e distância de edição. A análise comparativa dos resultados permitiu identificar o modelo mais eficaz para a previsão das rotas e destacou áreas que poderiam ser aprimoradas. Esta avaliação crítica foi essencial para assegurar que os modelos atingissem os objetivos definidos inicialmente.

Embora a fase de *Deployment* não tenha sido completamente abordada neste estudo, é importante mencionar que ela envolve a implementação do modelo final em um ambiente de produção. Isso inclui a integração do modelo com os sistemas existentes da empresa e a criação de interfaces para facilitar o acesso aos resultados. As decisões sobre a implantação serão tomadas após a conclusão da validação e otimização do modelo.

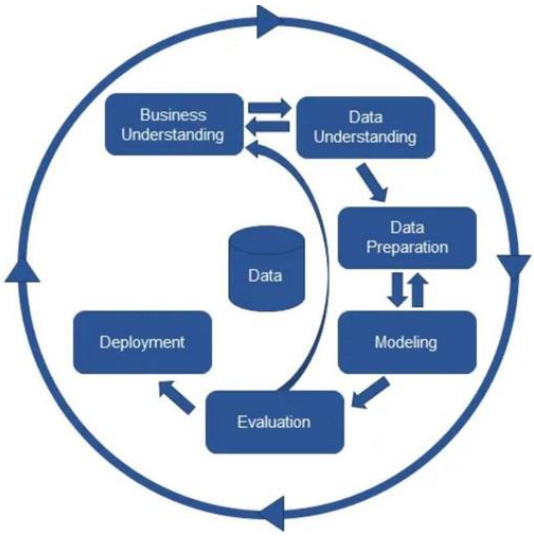

Figura 4: Processo CRISP-DM
Fonte: Gonzales (2019)

Para a execução deste projeto, foi utilizado o *Visual Studio Code* em combinação com a codificação em *Python*. O *Visual Studio Code* foi escolhido devido à sua flexibilidade e capacidade de suportar diversas extensões e ferramentas úteis para o desenvolvimento e análise de algoritmos de *Machine Learning*. Confome Borges (2014), a utilização do *Python* permite a implementação eficiente de técnicas avançadas e a experimentação com diferentes algoritmos, contribuindo significativamente para a realização dos objetivos do projeto. A adoção desta metodologia estruturada e o uso das ferramentas adequadas garantiram uma abordagem organizada e eficaz para a previsão de rotas, alinhando-se com as melhores práticas em *Machine Learning* e análise de dados.

3.1. Business Understanding

No contexto deste estudo, o tópico de *Business Understanding* desempenha um papel crucial na definição do âmbito e dos objetivos do projeto de previsão de rotas com base no comportamento dos motoristas. Este trabalho foi realizado em colaboração com uma empresa de logística, que forneceu um conjunto abrangente de dados operacionais. Estes dados foram essenciais para o desenvolvimento e teste dos modelos de *Machine Learning*, com o objetivo final de imitar o comportamento dos motoristas e gerar previsões de rotas baseadas nas frequências de locais visitados.

O conjunto de dados disponibilizado pela empresa consiste num registo detalhado das rotas realizadas pelos motoristas ao longo de um período significativo. Este registo inclui informações sobre horários de início e término das viagens, data, latitude, longitude. Esses dados foram coletados continuamente através de sistemas de telemetria instalados nos veículos da frota, garantindo uma coleta precisa e em tempo real

A etapa do *Business Understanding* envolveu uma análise profunda das necessidades e desafios enfrentados pela empresa de logística na gestão das suas operações de transporte. Um dos principais problemas identificados foi a variabilidade nas rotas escolhidas pelos motoristas, o que resultava em inconsistências nos tempos de entrega e, consequentemente, na satisfação dos clientes. Além disso, a empresa procurava reduzir os custos operacionais relacionados ao consumo de combustível e manutenção dos veículos, que estavam diretamente associados ao comportamento de condução dos motoristas. Com esses desafios em mente, o objetivo deste projeto foi desenvolver modelos de *Machine Learning* capaz de prever as rotas mais eficientes, considerando as variáveis relacionadas ao comportamento dos motoristas.

A compreensão detalhada do negócio e das operações da empresa de logística permitiu estabelecer requisitos claros para o sistema preditivo, como a necessidade de integrar o modelo com os sistemas de gestão de frota existentes e a capacidade de fornecer recomendações em tempo real para os motoristas. Assim, a etapa de Compreensão do Negócio não só definiu as bases para a coleta e análise dos dados, mas também orientou o desenvolvimento de um modelo que atende diretamente às necessidades da empresa, proporcionando uma ferramenta poderosa para melhorar a eficiência operacional e a competitividade no mercado de logística

3.2. Data Understanding

A base de dados unificada fornecida pela empresa Compal, é composta por 17.817 linhas e 37 colunas, representa uma organização em que cada entrega de material a um cliente é registada numa linha separada. Sendo que o armazém de Coimbra é sempre considerado como o ponto de partida e o ponto de chegada de todas as rotas. Isto resulta em múltiplas entradas para o mesmo cliente quando são entregues diferentes materiais, gerando informações duplicadas que podem impactar a análise, especialmente no que diz respeito à eficiência operacional. As variáveis categóricas presentes na base de dados incluem informações como Data, Matrícula, Tipologia, Nome do Cliente, Morada, Tipo de Visita, entre outras.

Ao analisar os valores únicos dessas variáveis, observa-se que a variável “Data” apresenta 21 valores distintos. A variável Matrícula contém 13 valores únicos, indicando a utilização de diferentes veículos nas operações de entrega, sugerindo uma frota diversificada. A Tipologia dos veículos está dividida em duas categorias: "Pesado" e "Ligeiro", refletindo a classificação dos tipos de veículos. A variável Nome do Cliente apresenta 118 valores únicos, o que pode indicar uma diversidade de clientes ou variações na forma como os mesmos clientes foram registados na base de dados

A análise dos valores faltantes nas colunas revela que algumas colunas essenciais, como Data, Matrícula, Capacidade, Tipologia, Nome, Kg, Morada, Latitude e Longitude, estão completas, garantindo a integridade dos registos críticos. No entanto, outras colunas apresentam quantidades significativas de valores ausentes. Por exemplo, a coluna Referência Cliente tem 273 valores ausentes, enquanto Fornecimento possui 291 valores ausentes. As colunas Receb. Merc e Material também apresentam 311 valores ausentes cada, e a coluna Tamanho/Dimensões contém 1.528 valores ausentes. Esses dados em falta podem comprometer a completude das análises, especialmente se essas colunas forem relevantes para a análise de desempenho ou planeamento

logístico. As estatísticas das colunas numéricas disponíveis na base de dados estão representadas na Tabela 2.

Tabela 2: Estatísticas descritivas das variáveis numéricas

	Capacidade	Kg	Distancia	Qtd.remessa	Peso líquido	Peso bruto
Contagem	17793	17793	17793	17502	17502	17502
Média	4949.05	275.79	5.44	2.59	18.05	22.41
Desvio Padrão	1300.79	502.29	14.41	5.73	55.17	61.53
Minimo	1200	0	0	1	0	0
25%	5000	73,11	0.303	1	3.135	5.287
50%	5500	137.24	1.21	1	7.914	9.912
75%	5500	258.7	3.979	2	12.54	18.298
Máximo	5500	5811.8	141.137	160	16560	1.689.396

A Figura 5, que mostra a variação da capacidade ao longo do tempo, revela uma flutuação significativa na utilização da capacidade média diária dos camiões, particularmente entre os dias 21/03/2022 e 04/04/2022. A análise desses dados revela padrões interessantes que podem estar diretamente relacionados à procura de entregas durante esse período. Nos dias 23/03/2022 e 24/03/2022, observa-se um aumento acentuado na utilização de caminhões da tipologia "Pesado", indicando um volume maior de pedidos a serem entregues. Esse aumento na procura provavelmente exigiu a mobilização de caminhões com maior capacidade de carga, capazes de transportar volumes maiores de mercadorias numa única viagem. Este comportamento sugere picos de atividade logística, possivelmente impulsionados por fatores como promoções, sazonalidade ou aumento temporário na demanda dos clientes.

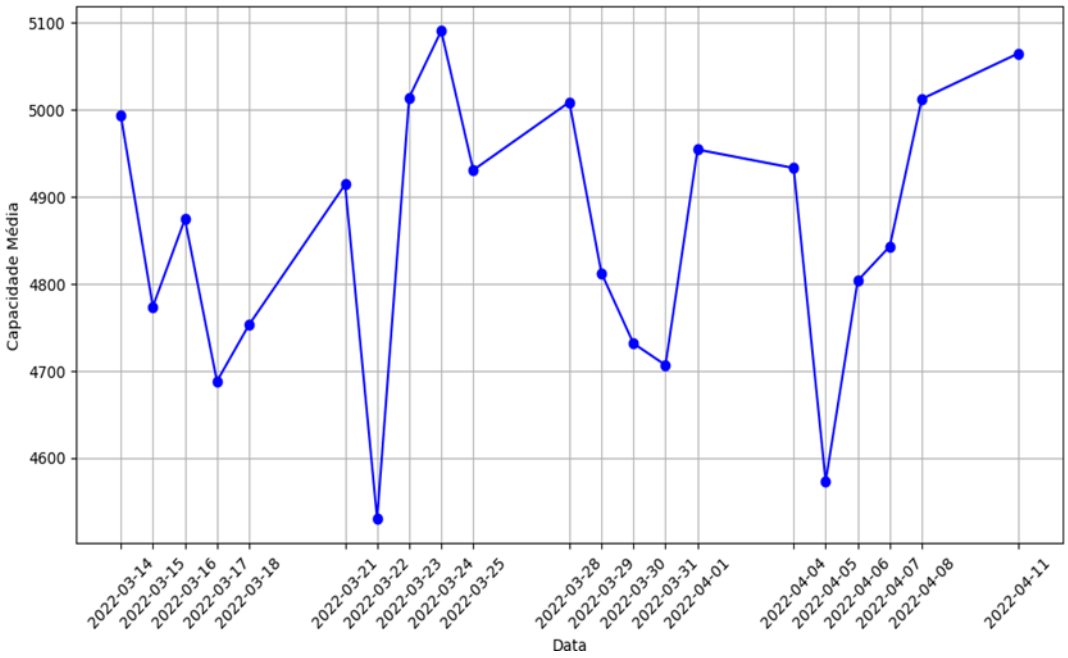

Figura 5: Gráfico da Variação da Capacidade ao Longo do Tempo

Por outro lado, os dias 21/03/2022 e 04/04/2022 apresentam os menores valores de capacidade utilizada, o que pode indicar períodos de menor atividade ou uma redução na quantidade de pedidos a serem atendidos. A menor utilização de camiões "Pesado" nesses dias pode refletir uma necessidade reduzida de capacidade de transporte, com entregas realizadas por veículos de menor porte ou em menor quantidade. Esses períodos de baixa podem ser indicativos de dias com menor volume de pedidos ou de uma logística mais eficiente, onde a capacidade dos caminhões foi otimizada para evitar o uso de veículos maiores. A variação na utilização da capacidade dos caminhões ao longo do tempo reflete, portanto, a dinâmica da operação logística, que se ajusta conforme as demandas flutua. A identificação desses picos e vales na utilização da frota é crucial para o planeamento logístico, permitindo a alocação eficiente de recursos e a previsão de necessidades futuras. Além disso, essa análise pode fornecer insights valiosos para a otimização de rotas e a gestão de frotas, contribuindo para uma operação mais ágil e eficiente

A Figura 6 sobre a sazonalidade das visitas ao longo do tempo revela padrões evidentes na distribuição das atividades de entrega, destacando o dia 30/03/2022 como o que registou o maior número de visitas realizadas. Este pico pode estar associado a uma alta demanda específica ou ao acúmulo de entregas programadas para essa data, sugerindo um dia de maior atividade logística. Ao analisar os padrões semanais, nota-se que os finais de semana apresentam um número significativamente menor de visitas, o que é consistente com a prática comum de muitos setores que reduzem ou suspendem operações durante esses dias. Essa tendência é visível nos dados, com cinco dias consecutivos de entregas seguidos por uma redução ou ausência de visitas durante dois dias, que provavelmente correspondem ao sábado e domingo.

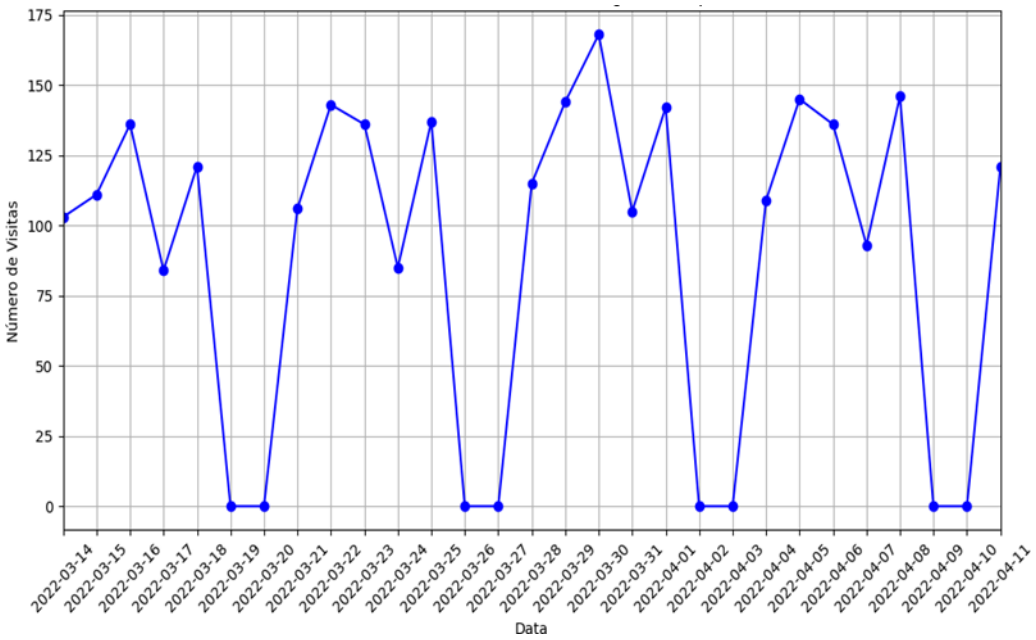

Figura 6: Gráfico da Sazonalidade das Visitas ao Longo do Tempo

Esse comportamento sugere uma sazonalidade semanal, onde o volume de visitas é maior durante os dias úteis, refletindo uma rotina operacional que prioriza as entregas de segunda a sexta-feira. Nos finais de semana, a redução no número de visitas pode indicar um planeamento

estratégico para concentrar as atividades nos dias de maior demanda e disponibilidade, ou pode estar relacionado a acordos comerciais que não exigem entregas durante o fim de semana. A análise dessa sazonalidade é crucial para compreender as variações no fluxo de trabalho e para otimizar a alocação de recursos humanos e logísticos. Compreender esses padrões ajuda a planear melhor as operações, garantindo que a capacidade esteja alinhada com a demanda e que os recursos estejam disponíveis nos dias de maior atividade. Além disso, essa sazonalidade pode ser utilizada para ajustar o planeamento de rotas e para gerir a equipe de forma eficiente, evitando sobrecargas em dias específicos e garantindo que as entregas sejam realizadas de forma consistente e pontual ao longo da semana.

A Figura 7, que mostra a diferença média de tempo entre as entregas previstas e reais por data, revela informações importantes sobre a pontualidade das operações de entrega ao longo do tempo. Além disso, a Tabela 3 facilita a compreensão das diferenças entre os valores. Os valores positivos no gráfico indicam que as entregas ocorreram, em média, após o horário previsto. O dia 30/03/2022 apresenta a maior diferença positiva, com um atraso médio de 0,61 horas. Isso sugere que, nesse dia específico, houve um desvio significativo em relação ao cronograma previsto, possivelmente devido a fatores como congestionamento de trânsito, um número maior de entregas a serem realizadas, ou problemas operacionais que impactaram o cumprimento dos horários programados. Da mesma forma, o dia 08/04/2022 também apresentou uma diferença positiva significativa, com um atraso médio de 0,55 horas, indicando uma situação semelhante de desvios na execução das entregas.

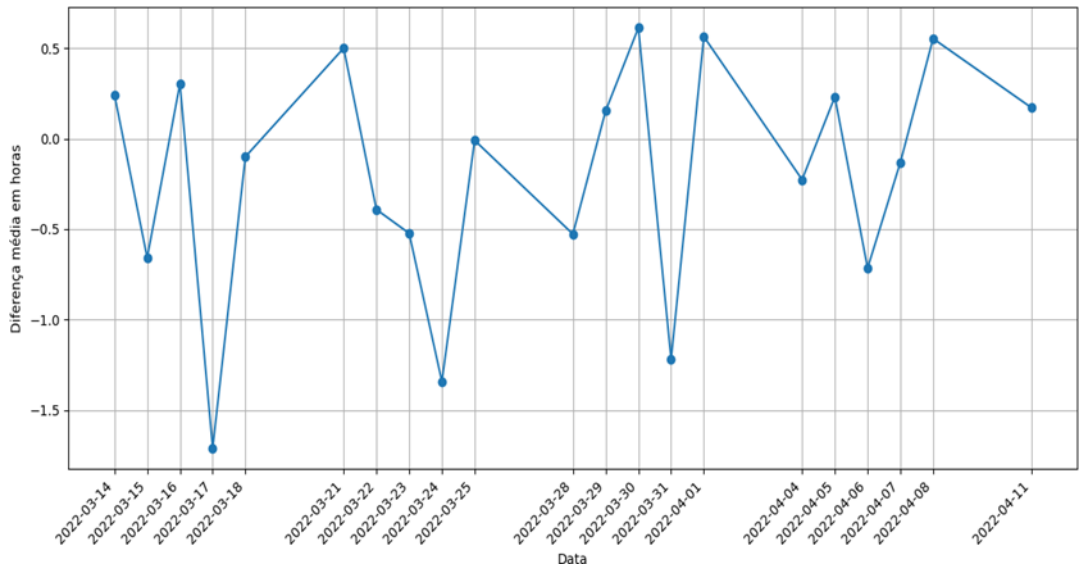

Figura 7: Diferença Média de Tempo entre Entregas Previstas e Reais por Data

Por outro lado, os valores negativos refletem entregas que ocorreram antes do horário previsto, indicando uma antecipação no cumprimento das tarefas. O dia 17/03/2022 registou a maior diferença negativa, com entregas realizadas, em média, 1,71 horas antes do esperado. Este cenário pode sugerir uma eficiência operacional excecional, onde os processos foram realizados de maneira mais rápida do que o planeado, ou pode indicar um planeamento de horários conservador.

De forma semelhante, no dia 24/03/2022, registou-se uma diferença negativa significativa de -1,34 horas, reforçando a ideia de antecipação nas entregas.

Esses resultados destacam a variabilidade na pontualidade das entregas ao longo do período analisado. Os dias com maiores diferenças positivas podem sinalizar a necessidade de ajustar o planeamento ou considerar fatores externos que possam ter impactado as operações. Por outro lado, os dias com diferenças negativas sugerem que, em determinadas circunstâncias, as entregas foram realizadas com maior agilidade do que o previsto, o que pode representar uma oportunidade para otimizar ainda mais o processo de planeamento e execução das entregas.

Tabela 3: Diferença Média em horas entre Entregas Previstas e Reais por Data

Data	Diferença (horas)
2022-03-14	0.24
2022-03-15	-0.66
2022-03-16	0.3
2022-03-17	-1.71
2022-03-18	-0.1
2022-03-21	0.5
2022-03-22	-0.39
2022-03-23	-0.52
2022-03-24	-1.34
2022-03-25	-0.01
2022-03-28	-0.53
2022-03-29	0.15
2022-03-30	0.61
2022-03-31	-1.22
2022-04-01	0.56
2022-04-04	-0.23
2022-04-05	0.23
2022-04-06	-0.72
2022-04-07	-0.13
2022-04-08	0.55
2022-04-11	0.17

A análise das distâncias totais percorridas por dia, apresentada na Figura 8, revela padrões distintos na intensidade das operações de entrega. Nos dias com maiores distâncias, como 11/04/2022, 04/04/2022, 21/03/2022 e 14/03/2022, observou-se uma quilometragem total significativamente elevada. Isso pode indicar um aumento na demanda, uma programação mais intensa de entregas ou a necessidade de atender a áreas mais distantes e mais clientes para esses dias. Esses dias com altos índices de distância sugerem uma carga de trabalho maior e possivelmente um esforço adicional para cumprir as metas de entrega.

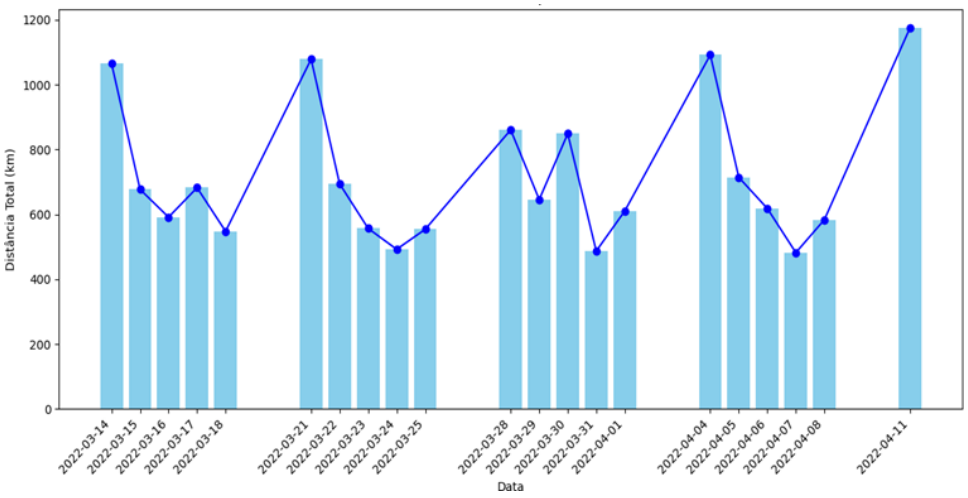

Figura 8: Distância Total por Dia

Por outro lado, os dias com menores distâncias, como 07/04/2022, 31/03/2022 e 24/03/2022, apresentam uma quilometragem reduzida. Isso pode refletir uma menor demanda, uma operação mais eficiente com rotas mais curtas ou uma concentração das entregas em áreas mais próximas. A baixa quilometragem nesses dias pode indicar uma menor necessidade de movimentação, o que pode estar relacionado a uma programação mais eficiente ou a uma redução no volume de entregas

Essas variações na distância total percorrida são cruciais para compreender a dinâmica das operações de entrega. A alta quilometragem pode evidenciar dias de intensa atividade e a necessidade de recursos adicionais, enquanto a baixa quilometragem pode indicar períodos de menor atividade ou uma operação mais eficiente. A análise desses padrões é fundamental para otimizar o planeamento logístico e ajustar as operações de acordo com a demanda e a eficiência pretendidas.

A análise do tempo total de entrega por dia, apresentada na Figura 9, oferece uma perspetiva valiosa sobre a eficiência e a pontualidade das operações. Observando os dados, destaca-se que o dia 30/03/2022 registou o maior tempo total de entrega. Isso indica que, nesse dia específico, as entregas demoraram mais a ser concluídas em comparação com outros dias. Esse aumento no tempo de entrega pode ser atribuído a diversos fatores, como congestionamento de trânsito, problemas operacionais ou um volume elevado de entregas que exigiu mais tempo para ser gerido

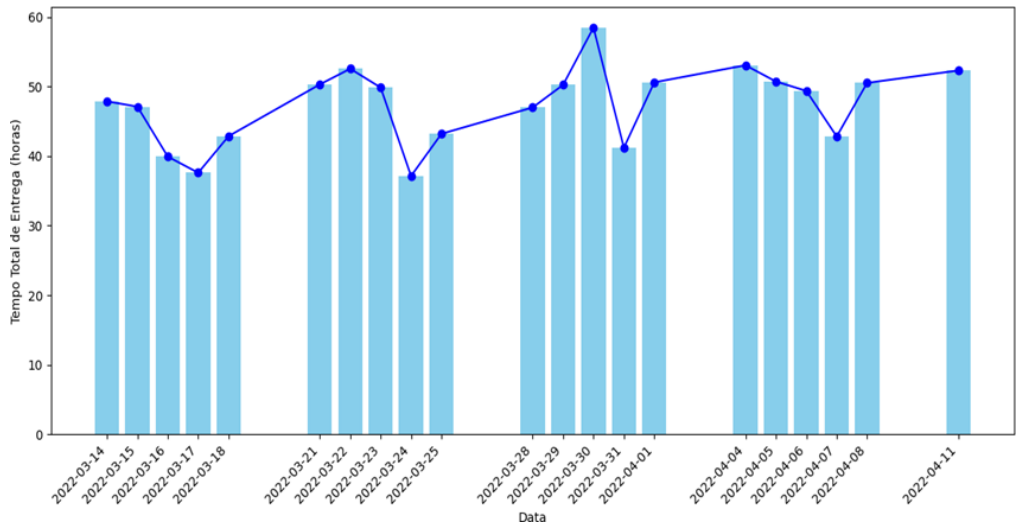

Figura 9: Tempo Total de Entrega por Dia

Em contraste, o dia 24/03/2022 registou o menor tempo total de entrega. Isso sugere que, nesse dia, as entregas foram realizadas de forma mais rápida e eficiente, possivelmente devido a condições mais favoráveis, menor congestionamento ou uma programação de entregas mais otimizada. A redução no tempo total pode refletir uma operação mais fluida e uma execução mais eficaz das tarefas. Essas variações no tempo total de entrega são essenciais para avaliar a eficácia das operações logísticas. Dias com tempos de entrega mais longos podem destacar áreas que necessitam de melhorias, enquanto os dias com tempos mais curtos podem servir como referência para práticas eficientes. Compreender e analisar essas flutuações ajuda a ajustar estratégias e otimizar o processo de entrega, assegurando uma melhor gestão do tempo e uma maior satisfação dos clientes.

A Figura 10, que apresenta a ocupação dos caminhões com espaço menor ou igual a 50%, fornece uma visão clara sobre a eficiência na utilização da capacidade dos veículos. Neste gráfico, os eixos representam as matrículas dos caminhões no eixo X e os pesos na partida, em kg, no eixo Y. As colunas são coloridas de acordo com a capacidade dos camiões: vermelha para veículos com capacidade de 5500 kg, amarela para 5000 kg e verde para 1200 kg. As linhas tracejadas indicam os limites de carga para cada categoria de caminhão, com a linha vermelha marcando o limite para os caminhões de 5500 kg, a linha amarela para os de 5000 kg e a linha verde para os de 1200 kg.

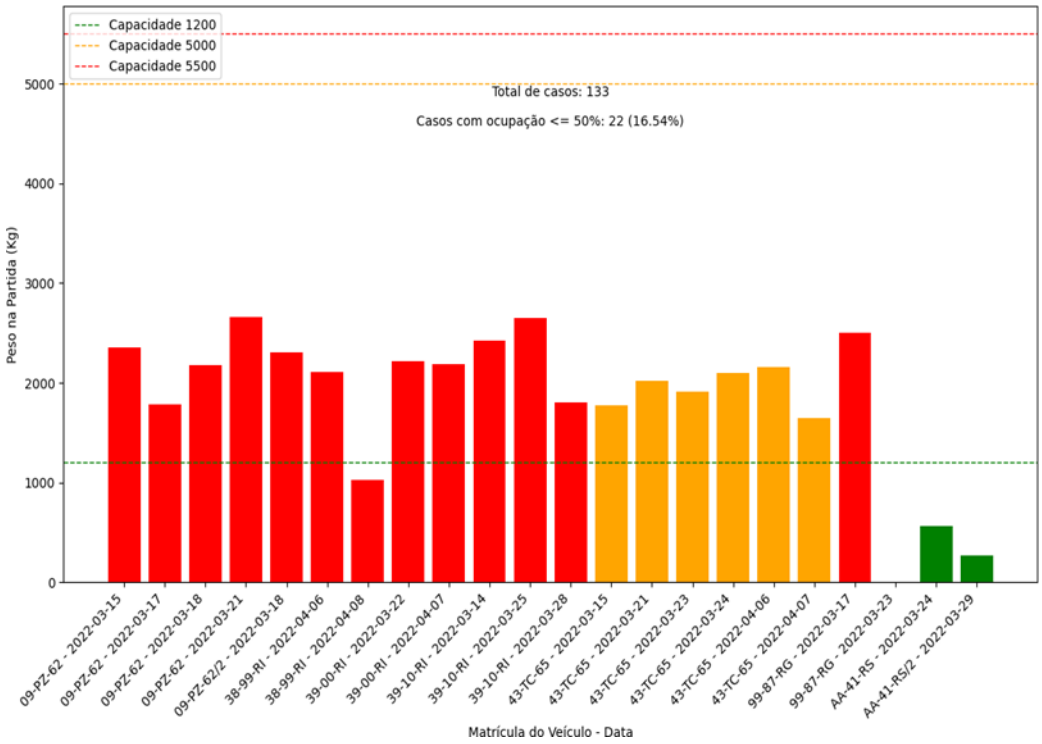

Figura 10: Camiões com Ocupação de Espaço Menor ou Igual a 50%

Observando as colunas vermelhas, que representam camiões com capacidade de 5500 kg, é possível identificar quais veículos estão operando com cargas inferiores à metade da capacidade máxima. Caminhões cujas colunas estão abaixo da linha vermelha estão transportando menos de 2750 kg, o que sugere uma possível subutilização da capacidade desses veículos. Isso pode indicar que esses caminhões poderiam carregar mais por viagem, aumentando assim a eficiência operacional.

As colunas amarelas, que correspondem a camiões com capacidade de 5000 kg, seguem uma lógica semelhante. Veículos com cargas abaixo da linha amarela estão utilizando menos da capacidade disponível, indicando também uma possível subutilização.

Finalmente, as colunas verdes representam caminhões com capacidade de 1200 kg. A análise das cargas abaixo da linha verde ajuda a verificar se esses veículos estão igualmente subutilizados. Esta análise é essencial para compreender a eficiência na utilização dos caminhões e pode auxiliar na identificação de oportunidades para melhorar o planejamento das cargas. Ajustar a alocação dos recursos para maximizar a capacidade de transporte e otimizar os custos operacionais é fundamental para alcançar uma operação mais eficiente.

3.3. Data Preparation

Na fase de preparação dos dados, o objetivo principal foi garantir a integridade e a consistência dos dados, de forma a assegurar análises precisas e fiáveis. O primeiro passo foi a remoção de valores ausentes em várias colunas da base de dados. Inicialmente, foram eliminadas as entradas com dados faltantes significativos em colunas como "Fornecimento", "Sequência real", "Horário

real de entrega", "Receb. merc", "Material", "Nº do material", "Qtd. remessa", entre outras. Para manter a qualidade da análise, essas linhas incompletas foram excluídas, uma vez que poderiam comprometer a precisão dos resultados.

Durante o processo de limpeza, foi também identificado que algumas colunas continham valores ausentes recorrentes para a data de 18/03/2022 e para o veículo com a matrícula 99-87-RG. Esses registros estavam incompletos em relação ao fornecimento, sequência real, horário real de entrega e recebimento de mercadoria. Como esses dados não eram suficientemente robustos para análises subsequentes, foram excluídos da base para evitar a inclusão de informações incertas.

Além disso, foram removidas colunas redundantes que não acrescentavam valor à análise, pois simplesmente eram informações repetidas em relação a primeira coluna como "Fornecimento2" por exemplo era a mesma informação de “Fornecimento 1”, "Un.", "Un.3" e "Tamanho/dimens.". Essas colunas repetitivas ou desnecessárias foram eliminadas para simplificar a base de dados.

Relativamente às unidades de medida, foi efetuado um ajuste para padronizar todas as dimensões para centímetros, uma vez que a base de dados continha medidas expressas em diferentes unidades (milímetros e centímetros). Esta padronização garantiu consistência nas análises que envolvem dimensões físicas. Após a limpeza e ajuste dos dados, a base final passou a incluir as seguintes colunas: 'Data', 'Matrícula', 'Capacidade', 'Tipologia', 'Nome', 'Kg', 'Morada', 'Latitude', 'Longitude', 'Tipo de visita', 'Janela Horária', 'Horário', 'Visita', 'Distância', 'Referência Cliente', 'Transporte', 'Fornecimento', 'Sequência real', 'Horário real de entrega', 'Material', 'Qtd. remessa', 'UM', 'Peso líquido', 'Peso bruto', 'TxtBreveMaterial', 'Altura', 'Largura' e 'Comprimento'.

A análise pós-limpeza revelou que a base de dados contém 135 rotas distintas, realizadas por 13 veículos diferentes e atendendo a 933 clientes únicos, conforme indicado na coluna "Referência Cliente". Com essas informações, a base está agora pronta para análises mais profundas, assegurando que a qualidade dos dados não comprometa os resultados futuros.

3.4. Data Modeling

Neste capítulo, o foco está na aplicação e avaliação de técnicas avançadas de modelagem para a previsão de rotas de entrega, com base em dados sobre o comportamento dos motoristas e padrões históricos de visita. Para alcançar este objetivo, foram exploradas diversas metodologias, destacando-se pela sua capacidade de lidar com a complexidade e variabilidade dos dados logísticos. As abordagens adotadas são as seguintes:

1. Modelação Baseada em Frequência de Visita: A modelagem inicial foi realizada utilizando técnicas de Aprendizagem por Imitação (*Imitation Learning*), que se fundamentam em padrões históricos de visita dos motoristas. Esta abordagem utiliza uma matriz de frequências para prever a próxima visita com base nas transições mais frequentes entre clientes, fornecendo uma base sólida para a otimização das rotas de entrega.
2. *Osquare* com diferentes algoritmos de previsão: Após a modelagem baseada em frequência, foram aplicadas técnicas de regressão e classificação para refinar e avaliar as previsões de rotas. Modelos de regressão linear, *Random Forest*, redes neurais, regressão logística, *Random Forest* para classificação e Máquinas de Vetores de Suporte (SVM) foram

treinados e avaliados com dados de treino e teste. Estas técnicas foram implementadas com o objetivo de identificar o modelo com melhor desempenho na previsão de rotas.

3. *Osquare* com a Incorporação de Atributos de Frequência: Para melhorar ainda mais a precisão dos modelos, foi realizada uma reavaliação com a inclusão de atributos de frequência. A adição desses atributos permitiu a consideração dos padrões históricos de visita, oferecendo uma nova dimensão para a previsão e otimização das rotas. Os modelos foram avaliados com e sem esses atributos, possibilitando uma análise comparativa detalhada.
4. Aplicação de métricas: Para avaliar a qualidade dos modelos de previsão de rotas, foram aplicadas métricas como Acurácia, que mede a proporção de rotas corretamente previstas, e o Coeficiente de *Kendall*, que avalia a correlação entre a ordem das rotas previstas e as reais. Além disso, métricas como *Edit Distance* analisaram as diferenças entre sequências de rotas, enquanto Distância Total e Tempo Total foram utilizadas para avaliar o impacto da otimização na eficiência das rotas, considerando a redução no percurso e no tempo de entrega.

O capítulo também realiza uma comparação entre os modelos desenvolvidos e o algoritmo de otimização de rotas atualmente utilizado pela empresa, que se baseia em um algoritmo clássico de natureza desconhecida. Essa comparação é essencial para avaliar se as técnicas de *Machine Learning* propostas superam ou complementam a abordagem tradicional, oferecendo ganhos em termos de eficiência e precisão na previsão e otimização das rotas de entrega. Os resultados dessa análise comparativa serão detalhados no próximo capítulo, junto às recomendações para a implementação.

A etapa de modelação iniciou-se com a divisão das rotas em conjuntos de treino e teste, um passo crucial para garantir uma avaliação rigorosa dos modelos. O banco de dados, contendo 134 rotas, foi repartido em dois subconjuntos: 80% das rotas (107 rotas) foram alocadas para o conjunto de treino, enquanto 20% (27 rotas) foram destinadas ao conjunto de teste, considerando alguns veículos e algumas datas, sendo que foram feitas entregas somente nos dias úteis. Esta divisão foi essencial para treinar os modelos e avaliar a sua capacidade de generalização em dados não previamente observados. Após a separação, foram elaboradas tabelas detalhadas para ambos os conjuntos, registando a ordem dos clientes, identificações fiscais, datas e matrículas. Com base nesses dados, foram construídas matrizes de frequência para os conjuntos de treino e teste. A matriz de frequência do conjunto de treino indicava a frequência com que cada cliente era visitado e revelava 871 clientes únicos. Da mesma forma, a matriz de frequência do conjunto de teste mostrou 374 clientes únicos. Estas matrizes foram fundamentais para a construção e avaliação dos modelos subsequentes.

Inicialmente, aplicou-se um modelo baseado em frequência do motorista utilizando a técnica de Aprendizagem por Imitação (*Imitation Learning*). Este modelo visava otimizar as rotas de entrega com base nos padrões históricos de visita dos clientes. A matriz de frequências, derivada dos dados de treino, orientou a seleção da melhor rota de entrega. A função de otimização começava com um ponto de partida definido ("Armazém - Partida") e, através de um processo iterativo, selecionava o próximo cliente a ser visitado com base na maior frequência de transição a partir do cliente atual. Este processo continuava até que todos os clientes de interesse fossem incluídos na rota, terminando com o destino final ("Armazém - Chegada").

Seguindo esta etapa, foi aplicada o *Osquare* junto a diversos algortimos de *Machine Learning* de regressão e classificação e além disso considerando atributos como capacidade dos veículos, peso e capacidade, o primeiro entre eles foi o de regressão linear para gerar o target de visita. O modelo de regressão foi treinado para prever um valor binário (1 ou 0), indicando se um cliente seria o próximo ponto a ser visitado. No conjunto de teste, o modelo produziu valores decimais, que foram utilizados para o ranqueamento dos clientes e para prever o próximo ponto de visita

Posteriormente, foi aplicada a técnica de *Random Forest* no *Osquare*. O *Random Forest*, um método de ensemble que utiliza múltiplas árvores de decisão, foi ajustado com os dados de treino e os modelos foram avaliados no conjunto de teste. Esta abordagem seguiu a mesma metodologia da regressão linear, permitindo uma comparação direta entre os modelos quanto à capacidade de previsão das rotas.

Além da Regressão e *Random Forest*, foram exploradas técnicas de redes neurais. Este modelo, conhecido pela sua capacidade de capturar padrões complexos e não lineares, foi treinado com o conjunto de dados de treino e avaliado no conjunto de teste. As redes neurais foram ajustadas para prever os próximos pontos de visita, e os resultados obtidos foram comparados com os das técnicas de regressão.

Na fase de classificação, foram empregues três modelos distintos no *Osquare*: regressão logística, *Random Forest* e *Support Vector Machines* (SVM). A regressão logística foi utilizada para classificar as rotas com base em uma série de eventos binários. O *Random Forest* e o SVM foram ajustados para a tarefa de classificação, utilizando técnicas de ensemble e *kernels*, respetivamente. Cada modelo foi treinado e avaliado com as métricas estabelecidas, proporcionando uma visão detalhada da eficácia de cada abordagem na previsão das rotas de entrega.

Por fim, todos os modelos foram reavaliados com a inclusão dos atributos de frequência. A adição desses atributos, que refletem os padrões históricos de visita, teve como objetivo aprimorar a precisão das previsões. Modelos de regressão linear, *Random Forest*, redes neurais, regressão logística, *Random Forest* para classificação e SVM foram ajustados e avaliados com e sem os atributos de frequência, permitindo uma comparação abrangente das abordagens.

Em resumo, o processo de modelagem envolveu uma análise meticulosa do *Osquare* considerando as frequências de visita e a aplicação de diversas técnicas de aprendizado automático para otimização e previsão das rotas de entrega. A combinação das técnicas de Aprendizagem por Imitação, regressão e classificação, forneceu uma visão abrangente e detalhada das melhores práticas para a previsão e otimização das rotas de entrega.

3.5. Evaluation

A avaliação dos modelos desenvolvidos para a previsão de rotas de entrega é essencial para determinar a sua eficácia e precisão. Este processo foi realizado utilizando diversas métricas que oferecem uma visão detalhada sobre o desempenho dos modelos. As métricas escolhidas para esta análise incluíram o *Kendall Tau*, a *accuracy*, e *Edit Distance*.

Além dessas métricas específicas, foram também avaliadas a Distância Total e o Tempo Total das rotas. A Distância Total refere-se à soma das distâncias percorridas em todas as rotas previstas

pelo modelo, enquanto o Tempo Total considera a soma dos tempos necessários para percorrer essas rotas. Ambas as métricas fornecem informações valiosas sobre a eficiência logística das rotas sugeridas, ajudando a determinar se as previsões do modelo são não apenas corretas, mas também viáveis em termos operacionais.

A análise dessas métricas permite uma avaliação abrangente da qualidade dos modelos de previsão. O *Kendall Tau*, a acurácia e a Distância de Edição fornecem uma visão detalhada sobre a precisão das previsões e a capacidade do modelo em manter a ordem correta dos destinos. Simultaneamente, a avaliação da Distância Total e do Tempo Total oferece uma perspetiva sobre a eficiência das rotas propostas, ajudando a identificar possíveis melhorias na logística de entrega.

Portanto, a avaliação dos modelos envolveu uma análise minuciosa das métricas de desempenho para garantir que as previsões de rotas não apenas refletissem com precisão a sequência correta de destinos, mas também fossem práticas e eficientes em termos de distância e tempo. A combinação dessas métricas permite uma compreensão profunda da eficácia dos modelos e da sua aplicabilidade na previsão das rotas de entrega

4. RESULTADOS E DISCUSSÃO

A busca pela otimização eficiente das rotas de entrega através de modelos preditivos avançados representa um desafio significativo na logística moderna. Este capítulo tem como objetivo explorar e avaliar os resultados obtidos com as diversas abordagens de modelagem aplicadas, visando identificar a técnica mais eficaz para a previsão da rota. Foram implementadas e analisadas várias técnicas de modelagem, cada uma oferecendo uma perspetiva distinta para prever a sequência ideal de destinos. As abordagens exploradas foram:

1. Análise de Matriz de Frequências com Base nos NIFs dos Clientes: Inicialmente, foi realizada uma análise detalhada utilizando uma matriz de frequências, onde se avaliou a frequência de visitas aos NIFs (Número de Identificação Fiscal) dos clientes. Nessa matriz, observou-se a frequência com que os NIFs eram visitados a partir de um ponto específico. O objetivo foi identificar os pontos mais frequentemente visitados com base nas rotas reais, fornecendo uma visão detalhada dos padrões de visitas. Esta análise foi aplicada tanto no conjunto de treino quanto no de teste, permitindo uma avaliação dos padrões recorrentes nas rotas logísticas. A matriz de frequências forneceu uma base sólida para a compreensão do comportamento das rotas reais, sendo posteriormente utilizada como um recurso para alimentar e melhorar os algoritmos.
2. Modelação Baseada em Frequências de Visitas (*Imitation Learning*): Nesta abordagem, utilizou-se o aprendizado por imitação para prever a próxima visita com base na frequência histórica de transições entre clientes. Esta técnica mostrou-se promissora inicialmente ao identificar padrões recorrentes nas rotas de entrega e prever os destinos mais frequentes.
3. Modelação do O-Square com Regressão Linear, Random Forest e Redes Neurais: Utilizou-se o *Osquare* com regressão linear, *Random Forest* e redes neurais para prever se um ponto de destino seria o próximo na rota com base em atributos como distância, peso e capacidade do veículo.
4. Modelação do *Osquare* com modelos de Classificação: Regressão Logística, *Random Forest* e SVM: Foram implementados três modelos para a tarefa de classificação: regressão logística, *Random Forest* e Support Vector Machines (SVM). Cada modelo foi treinado para prever um target probabilístico de um destino ser o próximo na rota.
5. Aplicação das métricas *Kendall*, *Accuracy*, *Edit Distance* para avaliarmos cada modelo e compará-los para analisar qual seria o mais adequado a ser considerado. Além disso, foi avaliado a distância total e o tempo total de cada um dos algoritmos implementados.

Inclusão dos Atributos de Frequência: Após a avaliação inicial dos modelos, foi realizada uma análise adicional com a inclusão de atributos de frequência. Esta abordagem visou melhorar a precisão das previsões ao considerar os padrões históricos de visita em cada modelo, tanto de regressão quanto de classificação. A incorporação desses atributos envolveu a reavaliação dos modelos com dados que incluíam frequências de visitas como uma característica adicional.

A inclusão dos atributos de frequência demonstrou um impacto significativo na melhoria das previsões em alguns algoritmos. Para os modelos de regressão, essa adição ajudou a reduzir a discrepância entre as previsões e os valores reais, proporcionando uma visão mais precisa das rotas. Da mesma forma, para os modelos de classificação, a inclusão dos atributos de frequência contribuiu para um aumento na acurácia das previsões, evidenciando a importância de considerar padrões históricos para otimizar as previsões.

Os resultados indicam que os modelos de redes neurais e *Random Forest*, especialmente quando ajustados com a inclusão dos atributos de frequência, apresentaram o melhor desempenho global. Estes modelos destacaram-se pela sua capacidade de prever as rotas com precisão e eficiência superiores em comparação com os outros métodos avaliados. Embora a abordagem de aprendizado por regressão e os modelos de classificação também tenham mostrado desempenhos notáveis, apresentaram limitações em cenários mais complexos.

Este capítulo fornece uma visão abrangente das técnicas de modelagem aplicadas e dos resultados obtidos, destacando as abordagens mais eficazes para a previsão e otimização das rotas de entrega. A análise detalhada das métricas e a comparação dos modelos oferecem uma compreensão crítica das melhores práticas e das áreas que podem ser aprimoradas, estabelecendo uma base sólida para futuras investigações e melhorias na otimização das rotas logísticas.

4.1. Matriz de Frequências com base nos NIFs dos Clientes

No contexto da previsão e otimização de rotas logísticas, a análise de matriz de frequências com base nos NIFs dos clientes representa uma ferramenta importante para identificar padrões de visita e comportamento das rotas. Esse tipo de matriz é construído com base na frequência histórica das visitas realizadas entre os diferentes pontos de partida e os clientes atendidos, considerando os trajetos reais registados no conjunto de dados. Para explicar a aplicação prática dessa técnica, na tabela 4 há um caso bem simplificado do que foi desenvolvido ao longo do projeto.

A matriz de frequências foi construída utilizando os NIFs dos clientes que receberam mercadorias durante as rotas do conjunto de teste. Nessa análise, foram observados 374 clientes únicos, o que reflete a diversidade e amplitude das entregas realizadas. No exemplo mostrado na tabela 4, foi feito um recorte de algumas transições de rotas entre o armazém de partida e os NIFs de destino, permitindo uma análise focada em um pequeno número de clientes.

A tabela 4 apresentada acima ilustra como a matriz de frequências é organizada, destacando a relação entre o armazém de partida e os destinos subsequentes. Nesse caso, há o armazém de partida, e a frequência indicada em percentagem com que um determinado NIF é visitado. Por exemplo, podemos observar que saindo do armazém de partida relativamente ao cliente com NIF 21003162 tem uma frequência de 3,70% e ao cliente NIF 21002723 7,41%. Esse tipo de análise permite identificar quais são os destinos mais comuns partindo de um ponto específico, ajudando a identificar padrões recorrentes nas rotas.

Além disso, observamos também outros exemplos interessantes nessa pequena amostra. O cliente de NIF 11113134 mostra que há uma distribuição uniforme de 50% das visitas entre os NIFs 11101464 e 21003161, o que sugere que, para esse ponto de partida, essas duas rotas são

igualmente frequentes. Da mesma forma, o NIF 21003162 apresenta uma divisão de 50% nas visitas entre os NIFs 21006078 e 11099778.

Esses padrões observados na matriz de frequências fornecem uma visão clara de como os destinos são conectados e quais pontos são mais frequentemente visitados a partir do armazém de partida ou de algum específico cliente.

No conjunto de teste, a análise de 374 clientes mostrou que muitos padrões de visitas são consistentes com as rotas registadas, mas também houve casos de variabilidade, onde o comportamento não seguiu os padrões frequentes. Essa variação pode ser explicada por uma série de fatores logísticos, como mudanças nas demandas dos clientes, ajustes operacionais e a introdução de novos pontos de entrega.

Portanto, a análise da matriz de frequências não apenas auxilia na visualização dos padrões de visitas, mas também serve como uma base sólida para a modelagem preditiva. Ao incorporar essas frequências nos algoritmos, podemos otimizar as previsões das próximas rotas e melhorar a eficiência das operações logísticas.

Tabela 4: Matriz de Frequência de Visitas

	Armazém - Partida	**21003162**	**21002723**	**11101464**	**21003161**	**21006078**	**11099778**
Armazém - Partida		3,70%	7,41%				
11113134				50%	50%		
21003162						50%	50%

4.2. Modelagem Baseada em Frequências de Visitas (*Imitation Learning*)

A análise dos resultados obtidos para a modelação baseada em frequências de visitas, utilizando a aprendizagem por imitação, revela insights importantes sobre o desempenho do modelo. As métricas de Acurácia, coeficiente de *Kendall* e *Edit Distance* foram utilizadas para avaliar a eficácia e precisão das previsões. A referência para a implementação da avaliação das métricas foi relativamente às rotas reais realizadas pelos motoristas. Os resultados estão disponíveis na Tabela 5.

A Tabela 5 contém dados relevantes para a análise da previsão de rotas com base no comportamento dos condutores, utilizando uma abordagem de *Machine Learning* de *Imitation Learning*, onde é considerada uma imitação da rota real realizada pelo condutor com base em frequência. Neste caso, está a ser medido por métricas como Acurácia, coeficiente de *Kendall*, *Edit Distance*, distância total e tempo total. Esta combinação de métricas permite uma visão detalhada do desempenho do modelo e da eficácia das previsões em diferentes rotas e condições.

Inicialmente, é importante observar a variação de Acurácia entre as diversas entradas da tabela. A Acurácia, que reflete a precisão das previsões, apresenta valores que variam desde 0,071 até 1. Valores de 1 indicam previsões perfeitas, como nos casos das matrículas 99-XR-26 (17/03/2022), 38-99-RI (22/03/2022) e 99-87-RG (24/03/2022). Estes resultados sugerem que,

nessas datas, o comportamento do condutor ou as características da rota foram previstas com alta precisão pelo modelo. Em contrapartida, matrículas como 39-00-RI (25/03/2022) e 43-TC-65 (05/04/2022) mostram Acurácias muito baixas (0,156 e 0,071, respetivamente), indicando dificuldade do modelo em prever corretamente os destinos, possivelmente devido a maior variabilidade no comportamento do condutor ou condições da rota.

Tabela 5: Resultados da Modelação Baseada em Frequências de Visitas

Data	Matricula	*Accuracy*	*Kendall*	*Edit Distance*	Distância Total (km)	Tempo Total (h)
14/03/2022	43-TC-65	0,333333	0,350427	6	125,222	2,087
16/03/2022	39-10-RI	0,166667	0,072464	7	27,671	0,461
17/03/2022	63-SJ-29	0,363636	0,309091	6	104,907	1,748
17/03/2022	99-XR-26	1	1	0	110,887	1,848
18/03/2022	63-SJ-29	0,535714	0,42328	8	134,7	2,244
22/03/2022	38-99-RI	1	1	0	54,828	0,913
22/03/2022	39-00-RI	0,526316	0,309942	5	141,73	2,362
24/03/2022	43-TC-65	0,388889	0,660131	4	111,211	1,853
24/03/2022	99-87-RG	1	1	0	46,312	0,7718
25/03/2022	39-00-RI	0,15625	0,08871	8	116,08	1,9346
25/03/2022	43-TC-65	0,347826	0,612648	11	30,3707	0,5061
28/03/2022	99-87-RG	0,153846	0,179487	9	261,962	4,366
28/03/2022	39-00-RI	0,259259	0,373219	16	128,088	2,13479
30/03/2022	99-87-RG	1	1	0	68,070	1,1345
01/04//2022	38-99-RI	0,0952381	0,152381	6	116,998	1,9499
04/04//2022	09-PZ-62	0,368421	0,22807	7	139,538	2,3256
04/04/2022	39-10-RI	0,26087	0,486166	11	174,626	2,9104
05/04/2022	39-00-RI	0,173913	0,272727	11	187,391	3,123
05/04/2022	39-10-RI	0,56	0,48	8	24,8184	0,4136
05/04/2022	43-TC-65	0,0714286	0,201058	12	179,914	2,99857
06/04/2022	AA-41-RS	0,684211	0,836257	4	71,2639	1,18773
07/04/2022	09-PZ-62	0,636364	0,781818	4	32,5989	0,5433
07/04/2022	39-00-RI	0,904762	0,87619	2	109,407	1,823
11/04/2022	09-PZ-62	0,5	0,555556	6	105,592	1,759
11/04/2022	39-10-RI	0,75	0,798942	6	116,455	1,940
11/04/2022	43-TC-65	0,571429	0,767196	9	179,733	2,995
11/04/2022	AA-41-RS	1	1	0	238,228	3,970

O coeficiente de *Kendall* segue uma tendência semelhante à da Acurácia. Valores próximos de 1, como os encontrados para as mesmas matrículas e datas com Acurácia, indicam forte correlação entre as previsões e os resultados observados. No entanto, valores mais baixos, como 0,072 para a matrícula 39-10-RI (16/03/2022) e 0,088 para 39-00-RI (25/03/2022), sugerem uma menor correspondência entre os padrões de visitas previstos e reais, o que pode estar relacionado com variações imprevisíveis no comportamento dos condutores.

O *Edit Distance* também reflecte o grau de precisão do modelo. Um valor de 0, como visto nas previsões perfeitas, indica que a sequência de visitas prevista é idêntica à sequência real. Já valores

mais altos, como 16 para a matrícula 39-00-RI (28/03/2022), indicam um desvio significativo entre o que foi previsto e o que ocorreu, sugerindo possíveis inconsistências no comportamento do condutor ou factores externos que influenciaram a rota.

Os dados de distância total e tempo total complementam esta análise, fornecendo insights sobre a eficiência e duração das rotas. Por exemplo, a matrícula AA-41-RS (11/04/2022) apresenta uma distância total de 238,228 km e um tempo total de 3,97 horas, o que, combinado com uma Acurácia e coeficiente de *Kendall* perfeitos (1), sugere que o modelo conseguiu prever uma rota de longa distância com alta precisão. Em contrapartida, rotas mais curtas, como a da matrícula 99-87-RG (24/03/2022), com 46,312 km e 0,77 horas, também foram previstas com alta exactidão (Acurácia e *Kendall* de 1). Isto sugere que o modelo consegue lidar bem tanto com rotas curtas quanto longas, desde que os padrões de visita dos condutores sejam previsíveis.

No entanto, as rotas com baixa Acurácia e alta distância de edição, como a da matrícula 43-TC-65 (05/04/2022), com uma Acurácia de 0,071 e uma distância de edição de 12, indicam que a previsão falhou significativamente, apesar de a distância total (179,914 km) e o tempo total (2,99 horas) não serem tão discrepantes. Isto pode indicar que, embora o modelo tenha dificuldade em prever a sequência exata das visitas, ele ainda consegue estimar razoavelmente o tempo e a distância totais.

De um modo geral, a Tabela 6 evidencia que o modelo de *Machine Learning* de *Imitation Learning* tem um desempenho variável, com bons resultados em cenários de rotas mais previsíveis e desafios quando há maior variabilidade no comportamento dos condutores. Estes resultados fornecem uma base sólida para discussões futuras sobre ajustes e melhorias no modelo, como o refinamento dos parâmetros ou a inclusão de novos factores que possam influenciar o comportamento dos condutores e, consequentemente, melhorar as previsões das rotas. Assim, as tabelas com todos os detalhes de cada algoritmo de *Machine Learning* aplicado neste projecto serão apresentadas na secção de Anexo A.

Tabela 6: Resultados das médias das métricas e os desvios padrões

	Accuracy	***Kendall***	***Edit Distance***	**Distância Total (km)**	**Tempo Total (h)**
Média	0,51	0,55	6,15	116,24	1,94
Desvio	0,31	0,32	4,17	61,5	1,03

Em termos de Acurácia, a média observada foi de 0,51, com um desvio padrão de 0,31. Isto indica que o modelo, em média, acerta cerca de 51% das previsões. Embora a acurácia média seja ligeiramente superior à linha de base de 0,5, sugerindo um desempenho melhor do que o aleatório, a ampla variação na acurácia sugere que o modelo pode não ser consistente em todas as situações. Esta variação pode indicar que o modelo tem dificuldade em generalizar de forma uniforme em diferentes contextos ou conjuntos de dados.

O coeficiente de *Kendall*, que mede a concordância entre as previsões do modelo e a verdade conhecida, apresentou uma média de 0,55 e um desvio padrão de 0,32. Este resultado reflecte uma concordância moderada nas previsões do modelo, mas a alta variabilidade sugere que há instâncias

em que o modelo está mais ou menos alinhado com a ordem correta, mostrando que a sua capacidade de ordenar corretamente as instâncias pode ser inconsistente.

O *Edit Distance* médio foi de 6,15, com um desvio padrão de 4,17. O *Edit Distance* mede o número mínimo de operações necessárias para transformar uma sequência prevista pelo modelo na sequência real. A média relativamente alta indica que, em média, o modelo precisa de realizar cerca de 6 operações para alinhar as suas previsões com a verdade conhecida. Esta diferença substancial entre as previsões e a realidade pode impactar a precisão geral das previsões, sugerindo que o modelo pode ter dificuldades em captar com precisão as sequências esperadas.

Além disso, foram calculadas duas métricas adicionais: distância total e tempo total. A média da distância total foi de 116,24 km, com um desvio padrão de 61,50 km. Esta alta variabilidade indica que as distâncias totais previstas pelo modelo podem variar significativamente, reflectindo tanto a incerteza nas previsões como a diversidade nas rotas analisadas. Uma distância total média elevada sugere que, em alguns casos, o modelo pode estar a prever rotas mais longas do que o necessário, o que pode impactar a eficiência operacional e a Acurácia da previsão.

O tempo total médio foi de 1,94 horas, com um desvio padrão de 1,03 horas. A variabilidade neste indicador é considerável, indicando que o tempo previsto para completar as rotas pode variar bastante. Esta diferença pode afectar a precisão das previsões temporais e a capacidade do modelo de fornecer estimativas fiáveis para o planeamento e execução das rotas. A discrepância significativa entre o tempo previsto e o tempo real pode sugerir que o modelo está a lutar para captar as nuances temporais associadas às rotas.

Estes resultados destacam tanto os pontos fortes como as limitações da modelação baseada em frequências de visita. O modelo demonstra uma capacidade moderada de previsão, mas com uma variabilidade significativa no desempenho, o que pode ser atribuído a factores como a complexidade dos dados ou a diversidade das características das instâncias analisadas. A alta variabilidade na distância total e no tempo total sugere que há áreas para melhorias adicionais. Para melhorar o desempenho geral, pode ser necessário ajustar o modelo ou explorar técnicas adicionais para reduzir a variação e melhorar a precisão das previsões.

4.3. Modelos de Regressão

Para a aplicação do modelo *Osquare*, foram seleccionados três algoritmos de regressão: regressão linear, *Random Forest* e redes neuronais. A metodologia começou com a configuração do modelo para operar com os atributos de origem, destino, distância, peso e capacidade, escolhidos pela sua relevância na definição da rota e previsão do próximo ponto de visita. Os dados foram divididos em conjuntos de treino e teste para permitir uma avaliação robusta do desempenho dos modelos.

No conjunto de treino, os algoritmos foram utilizados para ajustar modelos preditivos. A regressão linear forneceu uma abordagem directa baseada numa combinação linear dos atributos seleccionados. O *Random Forest* aplicou uma técnica baseada em árvores de decisão para capturar relações não lineares e interacções complexas, enquanto as redes neuronais foram empregues para explorar padrões complexos e não lineares, aproveitando a sua capacidade de modelação avançada.

Após o treino e ajuste dos hiperparâmetros, os modelos foram aplicados ao conjunto de teste para prever o target de cada ponto na rota. O target real é binário e indica se o ponto é o próximo local a ser visitado (1) ou não (0), sendo o caminho real que o motorista faz. O modelo *Osquare* gerou previsões contínuas do target com base nos atributos, que foram então utilizadas para classificar e hierarquizar os pontos da rota. O ponto com o valor previsto mais alto foi identificado como o próximo local a ser visitado, com o segundo maior valor seleccionado em caso de empate.

Para analisar se teria impacto no desempenho dos modelos, foram aplicados ajustes de hiperparâmetros em cada algoritmo. Este processo visou otimizar os parâmetros dos modelos para obter melhores resultados nas métricas de Acurácia, coeficiente de *Kendall*, distância de edição, distância total e tempo total. Os ajustes de hiperparâmetros ajudaram a refinar a capacidade dos modelos de prever com maior precisão o próximo ponto da rota, considerando tanto a eficiência preditiva quanto a adequação prática das previsões.

A análise dos resultados envolveu a comparação das previsões com os valores reais, considerando as métricas de Acurácia, coeficiente de *Kendall* e distância de edição. Além disso, foram calculadas a distância total e o tempo total para avaliar a eficácia das previsões em termos de logística prática.

A aplicação do modelo *Osquare* com os algoritmos de regressão forneceu uma visão abrangente sobre a capacidade do modelo em prever a sequência de visitas. A combinação das técnicas de modelação e a aplicação cuidadosa dos atributos, aliadas ao ajuste dos parâmetros, permitiram uma análise robusta da eficiência e precisão do modelo, destacando tanto as áreas de sucesso como as oportunidades de melhorias.

4.3.1. Regressão Linear Simples

Neste tópico, explora-se a modelação de rotas logísticas utilizando regressão linear, com o objetivo de otimizar as rotas e avaliar a eficácia do modelo, sendo que, para este caso, o melhor resultado está na Tabela 5. O processo é dividido em várias etapas principais: construção das tabelas de rotas, preparação dos dados, treino do modelo e avaliação do desempenho.

Inicialmente, o processo começa com a construção das tabelas de rotas a partir dos dados fornecidos. Utilizando a função *construir_tabela_rotas*, criam-se tabelas que descrevem todas as possíveis rotas entre os clientes. Para cada ordem de clientes, a função gera pares de origem e destino. A distância entre esses pontos é calculada com a função *calcular_distancia*, que utiliza a métrica euclidiana para medir a distância espacial com base nas coordenadas de latitude e longitude. Esta métrica é fundamental para avaliar a proximidade entre os pontos e, consequentemente, a eficiência da rota.

Além da distância, são recolhidos dados adicionais, como o peso da carga e a capacidade do veículo. A coluna Target Real é definida para indicar se o destino é o próximo na sequência correta da rota. Um valor de 1 na coluna Target Real significa que o destino deve seguir imediatamente após a origem na rota, enquanto um valor de 0 indica o contrário. Estes dados são organizados em tabelas, cada uma representando uma rota e contendo todas as combinações possíveis de origem e destino.

Após a construção das tabelas de rotas, o próximo passo é preparar os dados para o treino do modelo. As tabelas de rotas de treino são combinadas num único DataFrame. A partir deste DataFrame, extraem-se as características das rotas, que incluem distância, peso e capacidade. Estas características são, então, normalizadas utilizando o StandardScaler, o que garante que todas as variáveis estejam na mesma escala. Esta normalização é crucial para que o modelo de regressão linear possa processar os dados de forma eficiente e precisa.

Com os dados preparados, treinamos o modelo de regressão linear. Este modelo é ajustado para prever a coluna Target Real com base nas características das rotas. O objetivo do treino é fazer com que o modelo aprenda a prever correctamente se um par de origem e destino deve seguir na rota. Após o treino, avaliamos o desempenho do modelo utilizando um conjunto de dados de teste, o mesmo foi realizado para todos os algoritmos aplicados neste trabalho. No entanto, nesta etapa, adicionamos as previsões do modelo à coluna Target. As características das rotas de teste são extraídas e normalizadas usando o mesmo StandardScaler aplicado aos dados de treino. O modelo de regressão linear é, então, utilizado para prever os valores da coluna Target para os dados de teste.

Tabela 7: Resultados do Modelo de Regressão Linear Simples

	Accuracy	***Kendall***	***Edit Distance***	**Distância Total (km)**	**Tempo Total (h)**
Média	0,9350	0.1263	85,8889	99,03	1,6505
Desvio	0.0601	0.0437	36,6155	51,3268	0,8554

Os resultados das métricas de desempenho na Tabela 7 apresentam uma acurácia média de 0,9350, com um desvio padrão de 0,0601, indicando que o modelo tem um alto nível de precisão nas suas previsões. No entanto, o coeficiente de *Kendall* Tau tem uma média de 0,1263 e um desvio padrão de 0,0437, sugerindo uma concordância relativamente baixa na ordem das rotas previstas em comparação com as rotas reais. O *Edit Distance* médio é de 85,8889, com um desvio padrão de 36,6155, o que indica uma diferença considerável entre as sequências previstas e as reais.

A distância total média percorrida é de 99,0300 km, com um desvio padrão de 51,3268 km, enquanto o tempo total médio é de 1,6505 horas, com um desvio padrão de 0,8554 horas. Esses resultados destacam a eficácia do modelo de regressão linear na previsão das rotas, mas também apontam para áreas de melhoria, especialmente na concordância da ordem e na precisão das sequências previstas. A análise detalhada das métricas permite uma compreensão mais profunda do desempenho do modelo e fornece direções para ajustes futuros e melhorias na modelação das rotas logísticas.

4.3.2. Regressão com *Random Forest*

A introdução de ajustes nos hiperparâmetros do algoritmo de *Random Forest* resultou em melhorias notáveis na performance do modelo para a previsão e otimização de rotas. Durante o processo de ajuste, foram testados diferentes números de árvores no intervalo de 50, 100, 150 e

200. Observou-se que o modelo apresentou os melhores resultados com 200 árvores, uma profundidade máxima igual a 10, um número mínimo de amostras por nó interno igual a 5 e um número mínimo de amostras nas folhas igual a 4. Esses ajustes proporcionaram um desempenho mais robusto, especialmente na captura de variáveis complexas e no balanceamento entre precisão e generalização, mostrando-se a configuração mais indicada para a otimização das rotas analisadas.

Tabela 8: Resultados do Modelo de Regressão com *Random Forest*

	Accuracy	***Kendall***	***Edit Distance***	**Distância Total (km)**	**Tempo Total (h)**
Média	0,9297	0,1422	84,7407	112,57	1,88
Desvio	0,06	0,0526	34,3194	58,40	0,97

Confrome observado na Tabela 8, a *accuracy* média foi de 0,9297, com um desvio padrão de 0,06. Esses valores indicam que o tem uma taxa média de acerto de 93% das previsões. O desvio padrão baixo de 0,06 sugere que o modelo é consistente nas suas previsões, mantendo um nível elevado de precisão em diferentes cenários. Isso demonstra que o *Random Forest* conseguiu captar bem os padrões subjacentes nos dados, resultando em previsões bastante precisas.

A métrica de *Edit Distance*, com uma média de 84,7407, também indica o número médio de edições (inserções, deleções ou substituições) necessárias para alinhar a rota prevista à rota real. Esse valor relativamente alto sugere que, embora o modelo tenha uma alta acurácia em prever o próximo ponto, a sequência exata das rotas pode divergir significativamente, especialmente em rotas mais longas ou com maior variabilidade.

Em termos de eficiência operacional, os resultados para a distância total e o tempo total são também positivos. A distância total média de 112,57 km e o tempo total médio de 1,88 horas indicam que o modelo foi capaz de manter um equilíbrio adequado entre a precisão das previsões e a otimização das rotas, considerando trajetos realistas que não desviam significativamente da distância e do tempo observados nas rotas reais.

O desvio padrão de cada métrica também oferece insights importantes. O desvio padrão da acurácia, de 0,06, mostra uma baixa variação nos resultados, sugerindo que o modelo é consistente em suas previsões, independentemente da complexidade das rotas. No entanto, o desvio padrão de 0,0526 para o coeficiente de *Kendall* e de 34,3194 para a *Edit Distance* refletem uma variação maior em termos de ordem e alinhamento das rotas previstas, o que pode ser uma área de melhoria futura. Além disso, o desvio padrão para a distância total (58,40 km) e o tempo total (0,97 horas) indica uma certa variabilidade nas rotas previstas, o que pode ser atribuído a diferentes comprimentos e complexidades das rotas analisadas, mas sem impactar drasticamente a eficiência do modelo.

De forma geral, os resultados indicam que o modelo de *Random Forest* ajustado, com 200 árvores e outros parâmetros otimizados, apresentou uma performance sólida em termos de acurácia e eficiência nas rotas, mas ainda há espaço para ajustes finos em relação à sequência exata das previsões

4.3.3. Regressão com *Redes Neurais*

A análise dos resultados obtidos com a aplicação do algoritmo de redes neurais, juntamente com a manipulação do número de neurônios, a regularização por meio de *dropout* e o ajuste na taxa de aprendizado, revela nuances importantes sobre o desempenho do modelo. A regularização com *dropout* ajudou a minimizar o *overfitting*, permitindo que o modelo lidasse melhor com os dados de teste e evitasse super ajustes aos dados de treino. O ajuste na taxa de aprendizado também foi essencial para alcançar convergência de maneira mais eficiente, resultando em uma melhoria na velocidade de treinamento sem comprometer a precisão das previsões.

Tabela 9: Resultados do Modelo de Regressão com Redes Neurais

	Accuracy	***Kendall***	***Edit Distance***	**Distância Total (km)**	**Tempo Total (h)**
Média	0,9329	-0,1253	85,6667	109,5437	2,19
Desvio	0,06	0,0954	31,3475	57,4170	1,15

A *accuracy* média de 0,9329, conforme representado na Tabela 9, reflete uma excelente performance global do modelo na previsão de rotas, indicando que o modelo é altamente eficaz em prever o próximo ponto da rota na maioria das situações. Este valor demonstra que o modelo de redes neurais conseguiu aprender os padrões subjacentes de forma consistente, oferecendo previsões precisas. No entanto, o desvio padrão de 0,06 revela que, em alguns casos, o modelo pode apresentar flutuações, embora essas variações sejam relativamente pequenas. Essas oscilações podem estar associadas a dados com padrões menos consistentes ou a situações em que o modelo enfrenta maior complexidade nas rotas.

Por outro lado, o coeficiente de *Kendall* médio de -0,1253, em termos de ordenação de prioridades de rotas, sugere que o modelo pode ter dificuldades em identificar corretamente a sequência exata dos destinos, levando a uma inversão de posições ou a um desempenho inconsistente na ordenação. Embora a precisão seja alta, o valor de *Kendall* sugere que, em termos de ranking das rotas, o modelo não captura tão bem as relações entre os diferentes destinos. O desvio padrão de 0,0954 reforça essa observação, mostrando que o modelo apresenta uma certa variabilidade na capacidade de gerar ordens consistentes.

A distância de edição (*Edit Distance*) média de 85,6667 destaca o esforço do modelo em alinhar as suas previsões com as rotas reais, ou seja, a quantidade de operações necessárias para transformar a rota prevista na rota real. Esse valor, juntamente com o desvio de 31,3475, indica que, embora o modelo acerte a maioria das previsões, há espaço para melhorias na otimização da sequência dos pontos, dado que as mudanças necessárias para ajustar as previsões ao resultado correto ainda são consideráveis.

Em relação às métricas de distância total e tempo total, observa-se uma média de 109,5437 km e 2,19 horas, respetivamente. Esses valores indicam que o modelo de redes neurais está a fazer previsões de rotas que, em termos de eficiência logística, são bastante adequadas. No entanto, o desvio padrão de 57,4170 km e 1,15 horas sugere que, embora o modelo consiga manter previsões

próximas da realidade na maior parte dos casos, há uma variabilidade significativa. Isso pode ser resultado de rotas com padrões complexos ou com destinos mais dispersos, onde o modelo enfrenta mais dificuldades em manter a consistência.

Em suma, os resultados mostram que o modelo de redes neurais ajustado com manipulação de hiperparâmetros conseguiu obter um desempenho globalmente robusto, especialmente em termos de *accuracy*. No entanto, a análise das métricas de *Kendall* e *Edit Distance* sugere que o modelo poderia ser aprimorado na ordenação e na sequência das rotas, especialmente em cenários mais complexos. A variabilidade observada nas métricas de distância e tempo totais reforça a necessidade de ajustes adicionais para melhorar a eficiência geral das rotas previstas.

4.4. Modelos de Classificação

Para a aplicação do modelo *Osquare* com algoritmos de classificação, foram selecionadas técnicas capazes de trabalhar com um target binário, indicando se um ponto é o próximo destino na rota (1) ou não (0). Esse processo de classificação permitiu transformar o problema de previsão de rotas numa tarefa de decisão binária, onde a previsão correta implica identificar qual ponto de entrega será o próximo a ser visitado pelo motorista.

Os algoritmos de classificação utilizados foram treinados para prever a variável target com base nos atributos de origem, destino, distância, peso e capacidade. A modelagem começou com a configuração dos dados, separando-os em conjuntos de treino e teste para garantir uma avaliação robusta dos modelos. No conjunto de treino, os modelos foram ajustados para aprender os padrões que indicam se um ponto seria o próximo destino com base no histórico de visitas e nas características da rota, sempre considerando também o target da rota real que foi realizada pelo motorista.

Uma característica importante deste processo é que, em casos onde múltiplos pontos têm previsão de 1 (indicando que podem ser o próximo destino), a escolha final é feita com base no critério de menor distância entre a origem atual e o destino previsto. Esse procedimento visa garantir que a rota prevista não só seja precisa em termos de sequência de visitas, mas também otimizada em termos de distância percorrida, maximizando a eficiência logística.

Os algoritmos de classificação, como Regressão Logística, *Random Forest* e Support Vector Machines (SVM), foram configurados para maximizar a capacidade de identificar corretamente o próximo ponto da rota. Além disso, foram aplicados ajustes de hiperparâmetros para otimizar o desempenho dos modelos, visando aumentar a precisão das previsões. Esses ajustes incluíram parâmetros como profundidade máxima, número de árvores no caso do *Random Forest*, e parâmetros de regularização nos modelos de SVM e Regressão Logística.

A análise dos resultados dos modelos de classificação focou em métricas como precisão, coeficiente de *Kendall* e distância de edição, além das métricas logísticas de distância total percorrida e tempo total da rota. A aplicação do modelo *Osquare* com algoritmos de classificação mostrou-se eficaz na previsão da sequência de visitas, proporcionando uma análise robusta sobre a eficiência das rotas e a precisão dos modelos num contexto logístico real.

Em suma, o uso de técnicas de classificação no modelo *Osquare*, aliado a critérios de desempate baseados em distância, permitiu a construção de um sistema eficiente para prever a sequência de pontos de entrega, destacando a importância de um processo de otimização orientado tanto pela precisão das previsões quanto pela eficiência operacional.

4.4.1. Classificação com Regressão Logística

A aplicação da regressão logística com ajuste de hiperparâmetros, utilizando Searchgrid, gerou uma série de resultados que oferecem insights importantes sobre a eficácia do modelo. As métricas obtidas, apesar de apresentarem melhorias subtis, continuam a reflectir algumas das características e limitações persistentes deste método, tanto no desempenho geral como na capacidade de previsão detalhada. Os resultados do modelo de regressão logística com GridSearch estão apresentados na Tabela 10.

Tabela 10: Resultados do Modelo de Classificação com Regressão logística

	Accuracy	***Kendall* Tau**	**Edit Distance**	**Distância Total (km)**	**Tempo Total (h)**
Média	0,5646	-0,0466	1562,7407	114,3936	2,29
Desvio Padrão	0,1796	0,0986	1369,2886	56,0874	1,12

A análise dos resultados obtidos com o modelo de classificação baseado em regressão logística revela uma performance mista, destacando tanto pontos positivos quanto áreas a serem aprimoradas. A *accuracy* média de 56,46% indica que o modelo conseguiu prever corretamente pouco mais da metade dos pontos de destino na rota. Embora isso demonstre uma eficácia moderada, esse valor sugere que a regressão logística, apesar de ser uma abordagem direta e eficaz em cenários de menor complexidade, pode ter dificuldades para capturar as nuances de cenários mais complexos, onde interações não lineares entre os atributos são mais evidentes.

O coeficiente de *Kendall* Tau, com um valor médio de -0,0466, mostra um desempenho abaixo do esperado. Um coeficiente negativo indica que o modelo, em média, prevê a sequência de visitas de maneira inversa à ordem real, sugerindo que ele não está capturando adequadamente a relação entre os atributos e a ordem dos pontos na rota. Esse é um sinal claro de que a abordagem linear da regressão logística pode não ser suficiente para lidar com a complexidade dos padrões de rota presentes nos dados.

A distância de edição média (*Edit Distance*), com um valor elevado de 1562,7407, reforça a ideia de que o modelo tem dificuldades em prever corretamente a sequência dos pontos de visita, resultando em grandes discrepâncias em relação à rota real. Esse valor expressa quantas mudanças seriam necessárias para transformar a rota prevista na rota real, e a magnitude elevada indica que o modelo não está capturando a ordem correta de forma eficaz.

Por outro lado, a distância total percorrida e o tempo total apresentam valores mais razoáveis, com uma média de 114,3936 km e 2,29 horas, respectivamente. Esses resultados indicam que,

embora o modelo possa não prever a sequência exata de visitas, ele ainda consegue estimar uma rota com distância e tempo relativamente próximos dos valores reais, o que pode ser útil em determinados cenários práticos. No entanto, o desvio padrão elevado, especialmente na distância de edição e na acurácia, revela uma inconsistência considerável nas previsões, o que indica que o modelo pode funcionar bem em alguns casos, mas falha em outros.

Em resumo, os resultados mostram que a regressão logística, embora útil para capturar padrões lineares e mais simples, apresenta limitações significativas em prever corretamente a sequência de visitas nas rotas. Isso é evidenciado pela baixa *accuracy* e o coeficiente de *Kendall* Tau negativo, além da elevada distância de edição. Esses resultados sugerem que, para cenários mais complexos, seria necessário explorar modelos mais robustos que possam lidar melhor com a variabilidade e a complexidade dos dados

4.4.2. Classificação com *Random Forest*

O algoritmo *Random Forest* foi utilizado para a tarefa de classificação, e os resultados obtidos fornecem uma visão abrangente da eficácia e eficiência do modelo. Na tabela 11, a métrica de *accuracy*, que indica a proporção de previsões corretas feitas pelo modelo, apresentou uma média de 0,8930. Esse valor sugere que o modelo é bastante eficaz na sua capacidade de classificar corretamente os dados, com uma taxa de acerto de aproximadamente 89,30%. O desvio padrão da precisão foi de 0,0754, indicando que, apesar da alta precisão média, há uma variação moderada na performance do modelo entre diferentes conjuntos de dados. Essa consistência na performance é um indicativo positivo da robustez do algoritmo *Random Forest*.

Tabela 11: Resultados do Modelo de Classificação com *Random Forest*

	Accuracy	***Kendall* Tau**	**Edit Distance**	**Distância Total (km)**	**Tempo Total (h)**
Média	0,8930	-0,0664	91,4815	70,5132	1,41
Desvio Padrão	0,0754	0,0592	34,0831	53,2003	1,06

Por outro lado, o coeficiente de *Kendall* Tau foi de -0,0664, com um desvio padrão de 0,0592. O valor negativo e relativamente próximo de zero sugere uma fraca correlação entre as previsões do modelo e a classificação real dos dados. Portanto, embora o modelo seja eficaz em termos de *accuracy*, a sua performance em termos de ordenação e previsibilidade precisa de melhorias.

O *Edit Distance* uma média de 91,4815, com um desvio padrão de 34,0831. O valor médio relativamente alto sugere que, em muitos casos, o modelo faz previsões que requerem uma quantidade significativa de alterações para corresponder às classificações reais. O desvio padrão indica que há uma variação considerável nessa distância entre diferentes casos, o que pode refletir a dificuldade do modelo em adaptar as suas previsões às variabilidades dos dados reais.

Em relação à distância total e tempo total, os resultados médios foram de 70,5132 km e 1,41 horas, respetivamente. O desvio padrão associado a essas métricas foi de 53,2003 km para a

distância e 1,06 horas para o tempo. A distância total e o tempo total representam, respetivamente, o comprimento total das rotas e o tempo total necessário para a execução dessas rotas conforme previstas pelo modelo. A grande variação nos valores pode indicar que o modelo apresenta inconsistências na sua capacidade de prever a distância e o tempo de forma precisa, o que pode afetar a eficiência operacional na prática.

Por tanto, o modelo *Random Forest* demonstrou uma alta precisão na classificação, mas apresentou limitações significativas em termos de ordenação das previsões e consistência na distância e tempo previstos. A análise dos resultados sugere que, embora o modelo seja eficiente em termos de precisão geral, são necessárias melhorias para otimizar a sua performance nas métricas de *Kendall* Tau, distância de edição e nas previsões de distância total e tempo total

4.4.3. Classificação com Máquinas de Vetores de Suporte

O modelo de classificação utilizando Máquinas de Vetores de Suporte (SVM) produziu métricas que destacam várias características importantes sobre a sua performance. Na tabela 12, a métrica de precisão para o modelo SVM apresentou uma média de 0,2744, com um desvio padrão de 0,1735. Este resultado sugere uma taxa de acerto relativamente baixa, indicando que o modelo tem dificuldades em classificar corretamente as rotas. A elevada variabilidade na precisão, refletida pelo desvio padrão, sugere que o modelo pode ser instável e menos confiável, apresentando uma performance que varia consideravelmente em diferentes execuções.

Tabela 12: Resultados do Modelo de Classificação com SVM

	Accuracy	***Kendall* Tau**	**Edit Distance**	**Distância Total (km)**	**Tempo Total (h)**
Média	0,2744	-0,0765	171,1852	89,9985	1,80
Desvio Padrão	0,1735	0,0940	62,7169	47,2436	0,94

O coeficiente de *Kendall* Tau médio foi de -0,0765, com um desvio padrão de 0,0940. Este valor negativo e relativamente próximo de zero indica uma baixa capacidade do modelo em capturar a correlação de ordem entre as previsões e as classificações reais. A falta de concordância na ordenação das rotas sugere que o modelo enfrenta desafios em prever a sequência correta das rotas, o que é crucial para aplicações que dependem da ordenação precisa dos eventos.

A distância de edição (*Edit Distance*) média foi de 171,1852, com um desvio padrão de 62,7169. Esse valor elevado indica que as previsões de sequência de rotas feitas pelo modelo SVM estão bastante distantes das sequências reais. A grande distância de edição sugere que o modelo tem dificuldades significativas em prever a sequência correta das rotas, o que pode afetar a eficácia do modelo em situações práticas. A alta variabilidade observada, refletida pelo desvio padrão, indica que as previsões são não apenas imprecisas, mas também inconsistentes.

Quanto à distância total, o modelo SVM obteve uma média de 89,9985 km, com um desvio padrão de 47,2436 km. Esse valor sugere que, embora a distância total média prevista pelo modelo possa parecer razoável, há uma grande flutuação nas previsões individuais. A elevada variabilidade

nas previsões de distância total aponta para uma incerteza considerável, indicando que o modelo pode não ser confiável na previsão precisa das distâncias.

Finalmente, o tempo total médio foi de 1,80 horas, com um desvio padrão de 0,94 horas. Este valor sugere uma previsão de tempo médio que, apesar de ser relativamente próxima do esperado, apresenta uma variabilidade que pode afetar a consistência das previsões. O desvio padrão indica que o tempo total previsto pelo modelo pode variar significativamente, o que pode impactar a confiabilidade geral das previsões de tempo.

Em resumo, o modelo SVM apresenta uma baixa taxa de acerto e uma variabilidade considerável nas métricas de desempenho. A dificuldade em prever a ordem correta das rotas, bem como a grande distância de edição e a variabilidade nas previsões de distância e tempo, destacam desafios significativos na capacidade do modelo de classificar e prever rotas com precisão. Essas características sugerem que o modelo pode precisar de ajustes ou de exploração de abordagens alternativas para melhorar a sua eficácia e consistência

4.5. Modelos de Regressão com adição da frequência de visita

Neste tópico, exploraremos como a inclusão da frequência de visita foi integrada nos modelos de regressão, analisando o seu impacto nas métricas de desempenho, como a precisão das previsões de distância e tempo. Investigaremos como esta variável adicional pode influenciar a capacidade dos modelos de captar padrões complexos e melhorar a eficácia geral das previsões de rotas. A análise incluirá a avaliação das métricas de desempenho antes e após a inclusão da frequência de visita, fornecendo insights sobre a contribuição desta variável para a modelagem e otimização logística

4.5.1. Regressão Linear Simples com Frequência

A aplicação da Regressão Linear Simples, incorporando a variável de frequência e com ajustes de hiperparâmetros considerando o GridSearch, proporcionou resultados que destacam uma performance robusta em diversas métricas de avaliação. Na tabela 13 a métrica de *Accuracy* para o modelo de Regressão Linear Simples com a variável de frequência apresentou uma média de 0,9559, com um desvio padrão de 0,0347. Esta elevada média indica uma alta taxa de acerto, sugerindo que o modelo é altamente eficaz na previsão correta das rotas. O baixo desvio padrão indica uma alta consistência na performance do modelo, refletindo uma capacidade estável e confiável para realizar previsões precisas.

O coeficiente de *Kendall* Tau médio foi de 0,3304, com um desvio padrão de 0,1012. Este valor positivo sugere uma correlação moderada entre as previsões e a ordem real das rotas, o que é um indicativo de que o modelo é relativamente eficaz em capturar a sequência correta dos eventos. A variabilidade observada, embora presente, não compromete a capacidade do modelo de ordenar corretamente as previsões em relação às classificações reais, demonstrando uma melhoria em relação a abordagens que não utilizam a frequência.

Tabela 13: Resultados do Modelo de Regressão Linear Simples com Frequência

	Accuracy	***Kendall* Tau**	**Edit Distance**	**Distância Total (km)**	**Tempo Total (h)**
Média	0,9559	0,3304	73,2222	127,4612	2,1244
Desvio Padrão	0,0347	0,1012	37,6891	63,6018	1,06

A média do *Edit Distance* foi de 73,2222, com um desvio padrão de 37,6891. O *Edit Distance* relativamente baixa indica que as previsões do modelo estão mais próximas das sequências reais comparadas a outros métodos. A menor distância média sugere que o modelo é mais preciso na previsão da sequência correta das rotas. A variabilidade, embora significativa, ainda é menor em comparação com a observada em abordagens anteriores, apontando para uma melhoria na consistência das previsões.

Em relação a distância total, o modelo apresentou uma média de 127,4612 km, com um desvio padrão de 63,6018 km. Este valor sugere que, embora a previsão da distância total seja razoável, ainda há uma variação considerável nas previsões individuais. O desvio padrão elevado indica que as previsões de distância total têm uma flutuação significativa, o que pode afetar a precisão em cenários práticos onde a previsão exata da distância é crucial.

O Tempo Total médio foi de 2,1244 horas, com um desvio padrão de 1,06 horas. O valor médio sugere uma previsão de tempo que é relativamente adequada, mas com uma variação que pode impactar a consistência das previsões. A alta variabilidade reflete a necessidade de mais ajustes ou refinamentos no modelo para melhorar a precisão nas previsões de tempo total.

Em resumo, o modelo de Regressão Linear Simples com a adição da variável de frequência demonstrou um desempenho robusto com alta acurácia e uma correlação moderada na ordenação das previsões. A inclusão da frequência como variável no modelo proporcionou uma performance geral aprimorada, destacando sua importância na modelagem e previsão de rotas

4.5.2. Regressão *Random Forest* com Frequência

A aplicação do modelo de Regressão *Random Forest*, incorporando o atributo de frequência, gerou resultados que evidenciam um desempenho robusto e consistente em várias métricas de avaliação. A métrica de *Accuracy* para o modelo *Random Forest* com o atributo de frequência apresentou uma média de 0,9583, com um desvio padrão de 0,0314. A baixa variabilidade na *Accuracy*, refletida pelo desvio padrão, evidencia uma performance estável e confiável do modelo, com pouca variação na precisão entre diferentes execuções. Os resultados do modelo de regressão com *Random Forest* e com a frequência de visita pode ser visualizado na Tabela 14.

O coeficiente de *Kendall* Tau médio foi de 0,3333, com um desvio padrão de 0,1019. O valor positivo e relativamente alto indica uma boa correlação entre as previsões e a ordem real das rotas. Isso sugere que o modelo *Random Forest* é eficaz em capturar a sequência correta dos eventos, com uma capacidade notável de ordenar previsões de forma adequada. A variabilidade na métrica é modesta, indicando que o modelo mantém uma boa consistência na ordenação das previsões

Tabela 14: Resultados do Modelo *Random Forest* com frequência

	Accuracy	***Kendall* Tau**	**Edit Distance**	**Distância Total (km)**	**Tempo Total (h)**
Média	0,9583	0,3333	72,1481	112,42	1,87
Desvio Padrão	0,0314	0,1019	39,6547	55,98	0,93

O *Edit Distance* médio foi de 72,1481, com um desvio padrão de 39,6547. Esse valor relativamente baixo indica que as previsões do modelo estão próximas das sequências reais das rotas. A distância média reduzida sugere que o modelo *Random Forest* é eficaz na previsão da sequência correta das rotas, com uma melhora em relação à precisão observada em abordagens anteriores. A variabilidade na distância de edição, embora presente, é menor, o que reflete uma maior consistência nas previsões.

Para a Distância Total, o modelo apresentou uma média de 112,42 km, com um desvio padrão de 55,98 km. Esse valor médio sugere que o modelo prevê distâncias de forma razoavelmente precisa, embora a variação nas previsões individuais ainda seja significativa. A alta variabilidade nas previsões de distância total pode indicar que, apesar da média ser adequada, há flutuações consideráveis nas previsões, o que pode impactar a precisão em situações práticas onde a exatidão na previsão de distância é crucial.

O Tempo Total médio foi de 1,87 horas, com um desvio padrão de 0,93 horas. O valor médio sugere uma previsão de tempo total que é relativamente interessante, com um desvio padrão que reflete uma variabilidade considerável. A eficiência na previsão do tempo é positiva, mas a variabilidade indica que há espaço para melhorias na consistência das previsões de tempo total.

Em resumo, o modelo *Random Forest* com a adição do atributo de frequência demonstrou um desempenho altamente eficaz com alta acurácia e boa correlação na ordenação das previsões. As métricas de Distância de Edição e Tempo Total indicam uma boa precisão nas previsões, embora haja uma variabilidade notável nas previsões de distância e tempo. Sendo assim, a adição do atributo de frequência contribuiu significativamente para a melhoria geral do desempenho do modelo, destacando a importância desta variável na modelagem e previsão de rotas

4.5.3. Regressão Redes Neurais com Frequência

A aplicação da regressão utilizando Redes Neurais com hiperparâmetros ajustados de neurônios e considerando o atributo de frequência foi avaliada para verificar se esses ajustes impactariam as métricas do modelo. Os resultados do modelo de redes neurais com frequência e com hiperparâmetros estão apresentados na Tabela 15.

A métrica de *Accuracy* para o modelo de Redes Neurais com a variável de frequência apresentou uma média de 0,9603, com um desvio padrão de 0,0309. Esse elevado valor médio demonstra uma excelente taxa de acerto, refletindo a capacidade do modelo em prever corretamente as rotas com alta precisão. O baixo desvio padrão indica uma performance

consistente e confiável, com pouca variação entre as execuções do modelo, evidenciando uma alta estabilidade nas previsões

Tabela 15: Resultados do Modelo de Redes Neurais com frequência

	Accuracy	*Kendall* Tau	Edit Distance	Distância Total (km)	Tempo Total (h)
Média	0,9603	0,0314	63,1481	108,4389	2,17
Desvio Padrão	0,0309	0,0344	36,9331	54,6837	1,09

O coeficiente de *Kendall* Tau médio foi de 0,0314, com um desvio padrão de 0,0344. Este valor próximo de zero sugere que a correlação entre as previsões e a ordem real das rotas é relativamente fraca. A pequena variabilidade na métrica reflete uma consistência na dificuldade do modelo em capturar a ordem correta, mesmo com alta precisão na previsão.

A média do *Edit Distance* foi de 63,1481, com um desvio padrão de 36,9331. Este valor relativamente baixo indica que o modelo de Redes Neurais está bastante próximo das sequências reais das rotas, com um desempenho favorável na previsão da sequência correta. A variabilidade observada, refletida pelo desvio padrão, sugere uma consistência relativamente boa nas previsões, embora a distância média ainda possa indicar áreas para aprimoramento na precisão da sequência.

Em relação à Distância Total, o modelo apresentou uma média de 108,4389 km, com um desvio padrão de 54,6837 km. A distância total média é razoavelmente baixa, sugerindo que o modelo tem uma boa performance na previsão das distâncias. No entanto, a alta variabilidade nas previsões de distância total indica que há flutuações consideráveis nas previsões individuais, o que pode afetar a precisão prática, especialmente em contextos onde a exatidão da distância é crítica.

O Tempo Total médio foi de 2,17 horas, com um desvio padrão de 1,09 horas. O valor médio sugere que o modelo prevê o tempo total de forma adequada, mas a variabilidade significativa reflete uma falta de consistência nas previsões. Embora a média do tempo total esteja dentro de uma faixa aceitável, a alta variabilidade pode impactar a confiança nas previsões de tempo, especialmente em cenários onde a precisão é crucial para o planejamento.

Em resumo, o modelo de Redes Neurais com a inclusão do atributo de frequência demonstrou uma performance geral muito boa, com alta acurácia e uma baixa distância de edição média. No entanto, a correlação fraca na ordenação das previsões e a variabilidade nas previsões de distância e tempo destacam áreas onde o modelo ainda pode ser aprimorado. A alta precisão geral do modelo é positiva, mas a dificuldade em capturar a ordem correta das rotas e a variabilidade nas previsões indicam a necessidade de ajustes adicionais para melhorar a consistência e a precisão geral do modelo

4.6. Modelos de Classificação com adição da Frequência

Serão analisados os modelos de classificação com a adição do atributo de frequência de visita.

4.6.1. Classificação *Random Forest* com Frequência

A transição para os modelos de classificação marca um novo estágio na análise, focando em como esses modelos se comportam na categorização dos dados. As métricas agregadas obtidas para esses modelos revelam uma visão detalhada de seu desempenho. Os resultados do modelo de classificação com *Random Forest* e adição do atributo de frequência estão apresentados na Tabela 16.

Tabela 16: Resultados do Classificador *Random Forest* com frequência

	Accuracy	***Kendall* Tau**	**Edit Distance**	**Distância Total (Km)**	**Tempo Total (h)**
Média	0,9422	0,0301	89,2593	116,7887	2,34
Desvio Padrão	0,0338	0,025	40,9965	58,9150	1,18

O *Accuracy* Médio foi de 0.9422, com um desvio padrão de 0.0338. Isso indica que, em média, o modelo conseguiu classificar corretamente cerca de 94% dos dados, com uma variação relativamente baixa entre diferentes execuções, demonstrando consistência.

A métrica *Kendall* Tau Médio apresentou um valor de 0.0301, com um desvio padrão de 0.0250. Esse resultado mostra que, embora haja alguma concordância na ordem prevista em relação à ordem real, essa concordância ainda é relativamente baixa, o que sugere que a ordenação dos dados por esse modelo pode não ser tão precisa quanto seria desejável.

Quanto ao *Edit Distance* médio, o valor foi de 89.2593, com um desvio padrão de 40.9965. Isso reflete o número médio de operações necessárias para transformar a sequência prevista na sequência correta. O valor relativamente alto, junto com a significativa variabilidade, indica que as previsões do modelo podem frequentemente diferir consideravelmente das sequências reais.

A Distância Total Média foi de 116.7887 km, com um desvio padrão de 58.9150 km. Esse resultado sugere que as rotas previstas pelo modelo tendem a ser, em média, de uma extensão significativa, com uma considerável variação de uma execução para outra.

Por fim, o Tempo Total Médio foi de 2.34 horas, com um desvio padrão de 1.18 horas, indicando que as rotas previstas levam, em média, um pouco mais de duas horas para serem completadas, com uma variabilidade notável.

Esses resultados sugerem que, enquanto o modelo de classificação aplicado consegue manter uma boa taxa de acerto geral, ainda há desafios em termos de precisão na ordem das previsões e na otimização das rotas, o que é refletido nas distâncias e tempos calculados. A variabilidade nas métricas sugere que os resultados podem ser inconsistentes, apontando para a necessidade de ajustes adicionais ou de considerar outras abordagens para melhorar a estabilidade e a eficiência das previsões.

4.6.2. Classificação Regressão Logística com Frequência

Ao aplicar o modelo de regressão logística para classificação, os resultados obtidos fornecem uma visão clara do seu desempenho em diferentes métricas, revelando tanto seus pontos fortes quanto áreas de potencial melhoria. Os resultados do modelo de classificação com regressão logística e adição do atributo de frequência estão apresentados na Tabela 17.

Tabela 17: Resultados do Classificador Regressão Logística com frequência

	Accuracy	***Kendall* Tau**	**Edit Distance**	**Distância Total (Km)**	**Tempo Total (h)**
Média	0,9567	0,0783	111,6296	107,5889	2,15
Desvio Padrão	0,0325	0,0306	63,241	56,3078	1,13

O *Accuracy* Médio foi de 0.9567, com um desvio padrão de 0.0325. Esse alto valor de acurácia indica que o modelo conseguiu classificar corretamente uma alta proporção dos dados, com uma variação relativamente baixa entre diferentes execuções, sugerindo que o modelo é consistente em sua capacidade de prever corretamente as classes dos dados.

O *Kendall* Tau Médio apresentou um valor de 0.0783, com um desvio padrão de 0.0306. Esse resultado, embora melhor do que no modelo anterior, ainda indica uma concordância modesta na ordem prevista em relação à ordem real das rotas. A leve melhora em relação ao modelo anterior sugere que a regressão logística é um pouco mais eficaz em capturar a ordem correta das rotas, mas ainda há espaço para melhorias.

A média do *Edit Distance* foi de 111.6296, com um desvio padrão de 63.2410. Esse valor relativamente alto, junto com a significativa variabilidade, indica que o modelo pode, em algumas situações, gerar sequências de rotas que diferem bastante das sequências reais, exigindo muitas operações para alinhá-las corretamente.

Quanto à Distância Total Média, o resultado foi de 107.5889 km, com um desvio padrão de 56.3078 km. Isso sugere que, em média, as rotas previstas pelo modelo são razoavelmente eficientes em termos de extensão, com uma variabilidade moderada, indicando que o modelo pode ocasionalmente prever rotas com distâncias bem diferentes.

O Tempo Total Médio foi de 2.15 horas, com um desvio padrão de 1.13 horas. Esse tempo médio de conclusão das rotas é consistente com as distâncias previstas, e a variação indica que o modelo pode gerar tempos de rota diferentes entre as execuções.

Em resumo, o modelo de regressão logística para classificação demonstrou uma boa capacidade de acurácia e uma ligeira melhora na ordenação das rotas, em comparação com outros modelos. No entanto, a alta variabilidade nas distâncias de edição e na distância total sugere que ainda há desafios na precisão das previsões de sequência e na otimização das rotas, o que poderia ser abordado com ajustes adicionais ou a consideração de outros métodos para melhorar a eficiência e a consistência das previsões.

4.6.3. Classificação SVM com Frequência

Ao aplicar o modelo de classificação SVM (*Support Vector Machine*) com ajustes para considerar a frequência das visitas dos clientes, os resultados mostraram um desempenho robusto, especialmente em termos de acurácia, embora com algumas variações nas demais métricas. A média de *accuracy* foi de 0.9569, com um desvio padrão de 0.0320, indicando que o modelo foi altamente eficaz na classificação correta das rotas. A variação relativamente baixa sugere que o desempenho do SVM foi consistentemente bom em diferentes execuções, o que é um sinal positivo da estabilidade do modelo. Os resultados do modelo de classificação com SVM e adição do atributo de frequência estão apresentados na Tabela 18.

Tabela 18: Resultados do Classificador SVM com frequência

	Accuracy	***Kendall Tau***	**Edit Distance**	**Distância Total (km)**	**Tempo Total (h)**
Média	0,9568	0,0807	119,9529	79,4281	1,59
Desvio Padrão	0,0320	0,0299	54,6761	48,2537	0,97

No entanto, ao observar o *Kendall*'s *Tau*, que mede a concordância entre a ordem prevista e a ordem real das rotas, o valor médio foi de 0.0807, com um desvio padrão de 0.0299. Embora positivo, o valor de *Kendall*'s *Tau* ainda indica que o modelo teve dificuldades em capturar perfeitamente a ordem das rotas, o que pode afetar a eficiência prática das rotas geradas.

O *Edit Distance*, que mede o quão diferente a sequência prevista está da sequência real, teve uma média de 119.9259, com um desvio padrão de 54.6761. Esse valor relativamente elevado sugere que as rotas previstas pelo modelo ainda apresentam diferenças significativas em relação à sequência ideal.

Em termos de distância total percorrida pelas rotas previstas, a média foi de 79.4281 km, com um desvio padrão de 48.2537 km. Este resultado é interessante, pois, apesar das dificuldades na ordenação das rotas, o SVM conseguiu prever rotas que, em média, eram mais curtas em comparação a alguns outros modelos testados. Essa eficiência pode ser um fator positivo para aplicações práticas, onde a minimização da distância total é crucial.

Finalmente, o tempo total médio para completar as rotas previstas foi de 1.59 horas, com um desvio padrão de 0.97 horas. Esse resultado indica que, além de prever rotas relativamente curtas, o modelo também foi eficiente em termos de tempo, o que é um aspeto importante em contextos onde o tempo de operação é crítico.

Por fim, o modelo SVM apresentou uma alta acurácia e mostrou eficiência em prever rotas curtas e rápidas, embora ainda enfrente desafios na ordenação exata das rotas, como indicado pelo *Kendall*'s *Tau* e pela distância de edição. Esses resultados sugerem que, enquanto o SVM é promissor, pode haver espaço para melhorias na otimização da sequência das rotas para obter um desempenho ainda melhor.

4.7. Comparação entre os Modelos

A análise dos resultados dos diferentes algoritmos de *Machine Learning* aplicados à previsão de rotas com base no comportamento do motorista revela tendências interessantes e possibilita uma compreensão da eficiência de cada método. Considerando os modelos de regressão e classificação, tanto com quanto sem a adição do atributo de frequência de visita, é possível avaliar seu desempenho através de métricas como *accuracy*, *Kendall Tau*, *Edit Distance*, distância total percorrida e tempo total estimado. A Tabela 19 apresenta a comparação dos resultados dos modelos.

Tabela 19: Comparação dos Resultados dos Modelos

		Accuracy	***Kendall Tau***	***Edit Distance***	**Distância Total (km)**	**Tempo Total (h)**
Regressão Linear	Média	0,9350	0,1263	85,8889	99,03	1,6505
	Desvio Padrão	0,0601	0,0437	36,6155	51,3268	0,8554
Random Forest	Média	0,9297	0,1422	84,7407	112,57	1,88
	Desvio Padrão	0,06	0,0526	34,3194	58,4	0,97
Redes Neurais	Média	0,9329	-0,1253	85,6667	109,5437	2,19
	Desvio Padrão	0,06	0,0954	31,3475	57,417	1,15
Classificação – Regressão Logística	Média	0,5646	-0,0466	1562,74	114,3936	2,29
	Desvio Padrão	0,1796	0,0986	1369,28	56,0874	1,12
Classificação – *Random Forest*	Média	0,893	-0,0664	91,4815	70,5132	1,41
	Desvio Padrão	0,0754	0,0592	34,0831	53,2003	1,06
Classificação - *Support Vector Machine* (SVM)	Média	0,2744	-0,0765	171,1852	89,9985	1,8
	Desvio Padrão	0,1735	0,094	62,7169	47,2436	0,94
Regressão Linear com Frequência	Média	0,9559	0,3304	73,2222	127,4612	2,1244
	Desvio Padrão	0,0347	0,1012	37,6891	63,6018	1,06
Random Forest com frequência	Média	0,9583	0,3333	72,1481	112,42	1,87
	Desvio Padrão	0,0314	0,1019	39,6547	55,98	0,93
Redes Neurais com frequência	Média	0,9603	0,0314	63,1481	108,4389	2,17
	Desvio Padrão	0,0309	0,0344	36,9331	54,6837	1,09
Classificador *Random Forest* com frequência	Média	0,9422	0,0301	89,2593	116,7887	2,34
	Desvio Padrão	0,0338	0,025	40,9965	58,915	1,18
Classificador Regressão Logística com frequência	Média	0,9567	0,0783	111,6296	107,5889	2,15
	Desvio Padrão	0,0325	0,0306	63,241	56,3078	1,13
Classificador SVM com frequência	Média	0,9568	0,0807	119,9529	79,4281	1,59
	Desvio Padrão	0,032	0,0299	54,6761	48,2537	0,97

O modelo de regressão linear, sem o atributo de frequência, demonstrou uma precisão média elevada, com *accuracy* de 93,50%. No entanto, a baixa correlação representada pelo *Kendall Tau* (12,63%) sugere que, apesar da alta precisão, o modelo teve dificuldades em ordenar correctamente as rotas previstas em relação às reais. O *Edit Distance* média foi de 85,88, o que reflecte uma distância moderada entre as sequências previstas e reais. A distância total percorrida e o tempo total também foram estimados de forma razoável, com média de 99,03 km e 1,65 horas, respectivamente. Quando o atributo de frequência foi adicionado, a *accuracy* aumentou para 95,59%, e o *Kendall Tau* também melhorou, chegando a 33,04%. A *Edit Distance* reduziu para 73,22, sugerindo uma melhoria na previsão da sequência correta das rotas, enquanto a distância total aumentou para 127,46 km e o tempo para 2,12 horas, indicando que o modelo passou a prever rotas um pouco mais longas, mas de forma mais precisa e consistente.

O modelo de *Random Forest* apresentou resultados semelhantes. Sem a inclusão da frequência, a *accuracy* foi de 92,97%, com um *Kendall Tau* de 14,22%. A *Edit Distance* foi ligeiramente menor (84,74), e a distância total percorrida foi de 112,57 km, com um tempo total de 1,88 horas. Com a adição do atributo de frequência, a *accuracy* subiu para 95,83%, e o *Kendall Tau* teve uma ligeira melhora para 33,33%. A *Edit Distance* caiu novamente para 72,14, o que representa uma maior proximidade das previsões em relação à sequência real das rotas. Embora a distância total tenha caído para 112,42 km e o tempo total para 1,87 horas, esses resultados mostram que a adição da frequência beneficiou o modelo, especialmente em termos de ordenação correta das rotas e precisão geral.

As redes neurais, por sua vez, também apresentaram desempenho elevado, mas com comportamentos diferentes em relação às outras abordagens. Sem o atributo de frequência, o modelo teve uma *accuracy* de 93,29%, mas com um *Kendall Tau* negativo (-12,53%), indicando uma pior capacidade de ordenar corretamente as rotas. A *Edit Distance* foi de 85,66, bastante próxima dos outros modelos, enquanto a distância total foi de 109,54 km e o tempo total de 2,19 horas. No entanto, com a adição da frequência, a *accuracy* subiu para 96,03%, uma das mais altas entre todos os modelos, e o *Kendall Tau* melhorou para 3,14%, ainda muito abaixo dos modelos de regressão linear e *Random Forest*. A *Edit Distance* caiu significativamente para 63,14, o menor valor entre os modelos testados, sugerindo que as redes neurais com o atributo de frequência são altamente eficazes em prever a sequência correta de rotas. A distância total foi de 108,43 km e o tempo total de 2,17 horas, valores que mostram uma boa precisão nas previsões.

Os modelos de classificação, no entanto, apresentaram maiores variações. A regressão logística sem a frequência teve uma *accuracy* muito baixa (56,46%) e um *Kendall Tau* negativo (-4,66%), evidenciando dificuldades em ordenar corretamente as rotas. Além disso, a *Edit Distance* foi extremamente alta (1562,74), sugerindo que o modelo falhou em prever sequências próximas das reais. Com a adição do atributo de frequência, no entanto, a *accuracy* subiu para 95,67%, e o *Kendall Tau* para 7,83%, com uma queda significativa na *Edit Distance* para 111,62, embora ainda seja maior do que a observada em outros modelos. A distância total foi de 107,58 km e o tempo total de 2,15 horas.

O SVM apresentou os piores resultados sem a frequência, com *accuracy* de 27,44% e *Kendall Tau* de -7,65%, o que indica uma fraca capacidade de previsão e ordenação. A *Edit Distance* de 171,18 e as médias de distância total (89,99 km) e tempo total (1,80 horas) refletem o baixo desempenho do modelo. Entretanto, com a inclusão da frequência, a *accuracy* melhorou

drasticamente para 95,68%, e o *Kendall Tau* subiu para 8,07%. A *Edit Distance* ainda foi relativamente alta (119,95), mas o modelo foi capaz de prever distâncias menores (79,42 km) com um tempo total de 1,59 horas, sugerindo que o atributo de frequência foi decisivo para o aumento de desempenho.

Por fim, ao analisarmos de forma mais holística os resultados de todos os algoritmos, fica evidente que a adição do atributo de frequência de visita foi essencial para a melhoria significativa da precisão, especialmente nos modelos de regressão e classificação. Os algoritmos de regressão linear e *Random Forest* mostraram-se particularmente eficazes, tanto com quanto sem o atributo de frequência, mantendo alta acurácia e boas previsões de sequência. As redes neurais, embora com desempenho elevado, mostraram maior variação na capacidade de ordenar corretamente as rotas. Já os modelos de classificação, especialmente o SVM, demonstraram uma enorme dependência da inclusão da frequência para melhorar seu desempenho.

Portanto, a principal conclusão que podemos tirar deste trabalho é que a escolha do algoritmo de *Machine Learning*, combinada com o uso de atributos relevantes como a frequência de visitas, tem um impacto substancial no desempenho da previsão de rotas. Em termos gerais, os modelos baseados em regressão, especialmente o *Random Forest* e as redes neurais, com a inclusão da frequência, apresentaram os melhores resultados globais, sendo mais adequados para esse tipo de tarefa. A variabilidade nas métricas como *Kendall Tau* e *Edit Distance* sugere que ainda há espaço para melhorias na ordenação correta das rotas e na consistência das previsões.

5. CONCLUSÃO

A conclusão final deste trabalho reflete a abrangente análise dos resultados obtidos com a aplicação de diferentes algoritmos de aprendizagem de máquina para a previsão de rotas baseada no comportamento do motorista. Foram utilizados diversos métodos de regressão e classificação, incluindo regressão linear, *Random Forest*, redes neurais, regressão logística e SVM, com e sem a adição do atributo de frequência de visita. A análise foi conduzida com base em métricas fundamentais, como *accuracy*, *Kendall Tau*, *Edit Distance*, distância total percorrida e tempo total de percurso.

Ao longo do estudo, verificou-se que a regressão linear e o *Random Forest*, tanto em suas versões simples quanto com o atributo de frequência, apresentaram desempenho consistentemente elevado. A *accuracy* dessas abordagens variou em torno de 93% sem a frequência, e aumentou significativamente para aproximadamente 96% com a sua inclusão. Esses resultados indicam que esses modelos têm uma excelente capacidade de prever as rotas de maneira precisa, principalmente quando o histórico de visita é levado em conta. A adição da frequência também melhorou as métricas de ordenação (*Kendall Tau*) e a redução da *Edit Distance*, sugerindo que o modelo passou a capturar de maneira mais eficaz a sequência correta das rotas.

As redes neurais também demonstraram alta *accuracy*, atingindo uma média de 96,03% com a inclusão da frequência. No entanto, a análise da métrica de *Kendall Tau* revelou uma dificuldade maior em ordenar as rotas de forma eficiente, com valores bem abaixo dos observados nos modelos de regressão linear e *Random Forest*. A *Edit Distance* das redes neurais foi a mais baixa entre todos os modelos, o que indica que, apesar da dificuldade em ordenar, o modelo ainda conseguiu prever sequências de rotas próximas das reais. A distância total e o tempo de percurso previstos também foram bastante precisos, especialmente com o uso do atributo de frequência.

Por outro lado, os modelos de classificação, particularmente o SVM e a regressão logística, apresentaram maiores desafios. Sem o atributo de frequência, esses modelos exibiram baixa precisão e alta variabilidade nas previsões, com *accuracy* e *Kendall Tau* muito inferiores. O SVM, em especial, apresentou o pior desempenho, com *accuracy* de apenas 27,44% sem a frequência. No entanto, ao incluir o atributo de frequência, ambos os modelos mostraram uma melhora significativa na *accuracy*, superando 95%, o que demonstra a importância crítica desse atributo na melhoria do desempenho. Ainda assim, a *Edit Distance* permaneceu elevada para ambos os modelos, sugerindo que, embora a precisão geral tenha melhorado, a capacidade de prever a sequência correta de rotas ainda precisa ser aprimorada.

Com base nesses resultados, é possível concluir que a inclusão do atributo de frequência de visita é um fator determinante para o sucesso dos modelos, particularmente nas tarefas de previsão de rotas. Essa variável permitiu que os modelos captassem de maneira mais precisa os padrões históricos de comportamento dos motoristas, refletindo-se em um aumento significativo de precisão e uma melhor ordenação das rotas. Dentre todos os algoritmos, os modelos de regressão, especialmente o *Random Forest* com frequência, mostraram-se os mais robustos e adequados para essa tarefa, equilibrando alta *accuracy*, boa capacidade de ordenação (*Kendall Tau*) e baixas distâncias de edição.

Por fim, este trabalho evidencia a importância da escolha cuidadosa dos modelos de aprendizagem de máquina e dos atributos utilizados na previsão de rotas. A análise comparativa mostrou que os modelos de regressão, com a adição de variáveis relevantes como a frequência de visita aos clientes, são os mais indicados para tarefas de previsão com base em grandes conjuntos de dados. A capacidade de prever rotas com alta precisão e consistência oferece um grande potencial de aplicação em sistemas logísticos e de transporte, contribuindo para a otimização de recursos e melhoria da eficiência operacional. Assim, conclui-se que os métodos aplicados neste estudo podem ser aprimorados e expandidos, mas já oferecem um ponto de partida sólido para futuras implementações em ambientes de previsão de rotas baseados em comportamento.

5.1. Limitações e investigação futura

Os resultados obtidos neste trabalho abriram várias possibilidades para futuras pesquisas e aperfeiçoamentos no campo da previsão de rotas baseada no comportamento do condutor, especialmente em contextos logísticos e de transporte. Embora os modelos de aprendizagem automática aplicados tenham mostrado um bom desempenho em geral, há diversos aspetos que podem ser explorados para melhorar ainda mais a precisão e a eficiência dos algoritmos, além de expandir o âmbito da investigação para novos desafios e aplicações.

Um dos principais pontos de melhoria está relacionado com a otimização dos hiperparâmetros dos modelos. Embora tenham sido aplicados ajustes básicos de hiperparâmetros, uma abordagem mais sistemática e aprofundada, como o uso de técnicas de Bayesian Optimization ou Grid Search combinadas com validação cruzada, poderia levar a uma performance superior dos algoritmos. Em especial, modelos como redes neuronais podem beneficiar de ajustes mais sofisticados na sua arquitetura, como a seleção de camadas, neurónios, funções de ativação e taxas de aprendizagem.

Outro caminho promissor é a inclusão de novas variáveis preditivas. Embora o atributo de frequência de visita tenha mostrado um impacto significativo na melhoria do desempenho dos modelos, a incorporação de outros fatores relevantes, como condições de trânsito em tempo real, clima, ou variáveis sazonais, poderia aumentar ainda mais a capacidade preditiva dos modelos. Estes dados adicionais poderiam fornecer um panorama mais completo do comportamento dos condutores, permitindo previsões mais precisas e adaptáveis a cenários complexos e dinâmicos.

REFERÊNCIAS BIBLIOGRÁFICAS

Alves, Cláudio Manuel Martins, and José Manuel Valério de Carvalho. "Planeamento de rotas num sistema de recolha de desperdícios de madeira." (2004).

Alves, Lucas de Oliveira. "Roteirização de veículos na empresa SOBEBE-a aplicação de um método de roteirização de veículos a fim de aprimorar os processos logísticos da empresa SOBEBE no Plano Piloto."

Alzubi, Jafar, Anand Nayyar, and Akshi Kumar. "*Machine Learning* from theory to algorithms: an overview." Journal of physics: conference series. Vol. 1142. IOP Publishing, 2018.

Andriotti, Gustavo Kuhn. "Modelagem de motoristas e cenários de escolha de rota em simulações de tráfego veicular urbano." (2004).

Attaran, Mohsen, and Promita Deb. "*Machine Learning*: the new'big thing'for competitive advantage." International Journalof Knowledge Engineering and Data Mining 5.4 (2018): 277-305.

Attaran, Mohsen, and Promita Deb. "*Machine Learning*: the new'big thing'for competitive

Bonissone, Piero, et al. "A fuzzy random forest." International Journal of Approximate

Campos, Weslley Pina, Renata Mirella Farina, and Fabiana Florian. "Inteligência Artificial: *Machine Learning* na Gestão Empresarial." RECIMA21-Revista Científica Multidisciplinar-ISSN 2675-6218 3.6 (2022): e361617-e361617.

Cattaruzza, Diego, et al. "Vehicle routing problems for city logistics." EURO Journal on Transportation and Logistics 6.1 (2017): 51-79.

Chen, Zhiqin, and Hao Zhang. "Learning implicit fields for generative shape modeling." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition . 2019.

Chunqin, Xia, Liu Genjian, and Zhu Zhenhua. "Exploration and practice of training mode for innovative talents of electronic information based on competition [J]." Laboratory research and exploration 38.12 (2019): 173-177.

Cortes, Corinna, and Mehryar Mohri. "Domain adaptation and sample bias correction theory and algorithm for regression." Theoretical Computer Science 519 (2014): 103-126.

da Costa, Luciangela Galletti, and Rogério Valle. "Logística reversa: importância, fatores para a aplicação e contexto brasileiro." Anais III Simpósio de Excelência em Gestão e Tecnologia– SEGeT, Resende, Rio de Janeiro (2006).

da Rocha Miranda, Sara Filipa. O Planeamento e a Gestão de Rotas Como Meio de Satisfação. MS thesis. Instituto Politecnico do Porto (Portugal), 2017.

Dai, Hong-Ning, Zibin Zheng, and Yan Zhang. "Blockchain for Internet of Things: A survey." IEEE internet of things journal 6.5 (2019): 8076-8094.

de Almeida, Carlos Caetano. Identificação e classificação de imagens usando rede neural convolucional e" e; *Machine Learning*" e;: implementação em sistema embarcado. Diss. [sn], 2019

DE OLIVEIRA, JOSUÉ JONATHAN BORGES. "Análise Comparativa de Modelos de *Machine Learning* para Sugestão de Inspeções em Clientes de Distribuidoras de Energia Elétrica."

dos Santos, Lucas di Paula Gama, Wendell Ramon Barbosa Machado, and Pedro Vieira Souza Santos. "Aplicação do método de Clarke e Wright na resolução de problemas de roteirização: um estudo de caso." Revista Gestão Industrial 15.3 (2019).

dos-Reis, Marcelo, Fabiano Costa Teixeira, and Humberto T. Marques-Neto. "Utilizando o Modo de Dirigir do Motorista de Veículo Elétrico para o Planejamento e Roteirização de Viagem." Anais do VII Workshop de Computação Urbana. SBC, 2023.

Ebling, Angelo Augusto, et al. "Acuracidade da distribuição diamétrica entre métodos de projeção em Floresta Ombrófila Mista." Ciência Rural 42 (2012): 1020-1026.

Enomoto, Leandro Minoru, and Renato da Silva Lima. "Análise da distribuição física e roteirização em um atacadista." Production 17 (2007): 94-108.

FARIAS, Isaac da Silva. "Ciência de Dados no mercado de crédito: estratégias para mitigação de riscos e otimização de decisões com modelagem preditiva." (2023).

Fernandes, António José Silva. Aplicação de métodos heurísticos no planeamento de rotas: o caso da Tecniwood-Soluções. MS thesis. Universidade do Minho (Portugal), 2012.

Fernando, Erick, Setiawan Assegaff, and AH Hetty Rohayani. "Trends information technology in E-agriculture: A systematic literature review." 2016 3rd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE) . IEEE, 2016.

Ferreira, Ricardo Pinto. "Combinação de técnicas da inteligência artificial para previsão do comportamento do tráfego veicular urbano na cidade de São Paulo." (2011).

Florio, A., Paulo Da Costa, and S. S. Özarik. "A *Machine Learning* framework for last-mile delivery optimization." Technical Proceeding of the 2021 Last Mile Routing Research Challenge (2021).

França, Carolyne G., et al. "Análise comparativa de modelos de previsão de geração de energia eólica baseados em *Machine Learning*." Revista de Sistemas e Computação-RSC 9.2 (2020).

Franco, Pedro Henrique. "Análise de roteirização das entregas de uma empresa do ramo alimentício em Joinville." (2022).

Fuhr, Gabriel Tobias. "Uso de *Machine Learning* para a classificação do crédito de empresas por meio de demonstrativos financeiros." (2022).

Gómez, Carlos Alejandro Ramírez. "Aplicación del *Machine Learning* en agricultura de precisión." Revista Cintex 25.2 (2020): 14-27.

Ho, Jonathan, and Stefano Ermon. "Generative adversarial *Imitation Learning*." Advances in neural information processing systems 29 (2016).

Hussein, Ahmed, et al. "*Imitation Learning*: A survey of learning methods." ACM Computing Surveys (CSUR) 50.2 (2017): 1-35.

Júnior, Divaldo Portilho Fernandes, and Valdivino Vargas Júnior. "Conceitos e simulação de Cadeias de Markov." (2011).

Laydner, Marcos Silva. "Automação da avaliação de aprendizagem de *Machine Learning* para classificação de imagens no ensino fundamental." (2022).

LIMA, João Henrique Martins. Aplicação de *Machine Learning* para apostas esportivas: uso de regressão logística, SVM, árvore de decisão e Naive Bayes. BS thesis. 2022.

Marinho, Mayara Chew. "Estratégias computacionais baseadas em similaridade de textos e visualização exploratória para a identificação de inconsistências em notas fiscais eletrônicas." (2023).

Masioli, Luan Eiriz. "Otimização de rotas de compras em supermercados." (2023).

Medina, Pablo Blanco. Applications of scene text spotting to the darknet and industry 4.0. Diss. Universidad de León, 2023.

MENDONÇA JÚNIOR, Francisco Ferreira de. "Gerenciamento de recursos computacionais em redes fog veiculares: cenários, limites e alocação de tarefas baseada em leilões." (2021).

Miranda, Anderson Louzada, Rodrigo Duarte Soliani, and César Gomes de Freitas. "Otimização de rotas de entregas: um estudo de caso em uma empresa do setor alimentício." Revista Conexão na Amazônia 2.1 (2021): 152-169.

Miranda, Sara Filipa da Rocha. O planeamento e a gestão de rotas como meio de satisfação. Diss. 2018.

Monard, Maria Carolina, and José Augusto Baranauskas. "Conceitos sobre aprendizado de máquina." Sistemas inteligentes-Fundamentos e aplicações 1.1 (2003): 32.

Moura, Benjamim. Logística: conceitos e tendências. Centro Atlantico, 2006. Paura, Glávio Leal. "Fundamentos da logística." (2016).

Müller, Andreas C., and Sarah Guido. Introduction to *Machine Learning* with Python: a guide for data scientists. " O'Reilly Media, Inc.", 2016.

Nazário, Paulo. "A importância de sistemas de informação para a competitividade logística." Revista Tecnologística, São Paulo, ano 5 (1999): 31.

Neis, Lucas Ribeiro. *Machine Learning* na área do petróleo: implementação de redes neurais para aprendizado de distribuições de probabilidade. MS thesis. Florianópolis, SC., 2019.

Okamura, Dalton Akio. "Análise de algoritmos de regressão aplicados a mercado financeiro." (2019): 46-f.

Okugawa, Yoshinaga, et al. "Metastasis-associated long non-coding RNA drives gastric cancer development and promotes peritoneal metastasis." Carcinogenesis 35.12 (2014): 2731-2739.

Oliveira, Hugo Filipe Martins. Desenvolvimento de um modelo de otimização para planeamento de rotas de companhias aéreas. Diss. 2017.

Osa, Takayuki, et al. "An algorithmic perspective on *Imitation Learning*." Foundations and Trends® in Robotics 7.1-2 (2018): 1-179.

Osa, Takayuki, Jan Peters, and Gerhard Neumann. "Hierarchical reinforcement learning of multiple grasping strategies with human instructions." Advanced Robotics 32.18 (2018): 955-968.

Osório, Fernando S. "Redes Neurais Artificiais: Do aprendizado Natural ao Aprendizado Artificial." Tutorial–I Fórum de Inteligência Artificial/Ulbra (1999).

Procianoy, Guilherme Silveira, et al. "Impacto da pandemia do COVID-19 na vacinação de crianças de até um ano de idade: um estudo ecológico." Ciencia & saude coletiva 27 (2022): 969-978.

Reasoning 51.7 (2010): 729-747.

Rigatti, Steven J. "*Random Forest*." Journal of Insurance Medicine 47.1 (2017): 31-39. Biau, Gérard, and Erwan Scornet. "A *Random Forest* guided tour." Test 25 (2016): 197-227.

Rover, Vinicius. "Implementação de uma interface gráfica para uso de algoritmos de aprendizado de máquina." (2024).

Santos, Jaqueline Guimarães. "A logística reversa como ferramenta para a sustentabilidade: um estudo sobre a importância das cooperativas de reciclagem na gestão dos resíduos sólidos urbanos." Revista Reuna 17.2 (2012): 81-96.

SANTOS, Yan Antonino Costa dos. Verificação de assinaturas manuscritas através de análise de redes complexas. MS thesis. Universidade Federal de Pernambuco, 2021.

Savelsbergh, Martin, and Tom Van Woensel. "50th anniversary invited article—city logistics: Challenges and opportunities." Transportation science 50.2 (2016): 579-590.

Shibao, Fábio Ytoshi, Roberto Giro Moori, and MR dos Santos. "A logística reversa e a sustentabilidade empresarial." Seminários em administração 13 (2010): 1-17.

Silva, Daniel Henrique Cordeiro, and Elisa Maria do Nascimento Timo. "*Machine Learning* aplicado à atenção domiciliar para predição de condição de óbito." Research, Society and Development 11.14 (2022): e230111436078-e230111436078.

Silva, João Carlos Lopes da. Planeamento de rotas de distribuição. Diss. 2016.

Silva, Maurício José da. "RouteSpray: um algoritmo de roteamento de múltiplas cópias baseado em rotas de trânsito." (2014).

Sousa, Andreia Patrícia Ferreira. "Um Algoritmo Genético Para O Planeamento De Rotas Com Considerações Ambientais." (2015).

Sutton, Richard S. "Reinforcement learning: an introduction." A Bradford Book (2018).

Swamynathan, Manohar. Mastering *Machine Learning* with python in six steps: A practical implementation guide to predictive data analytics using python . Manohar Swamynathan, 2017.

Teixeira, Pedro Miguel Bento. "Classificação automática de termogramas do pé diabético usando técnicas de *Machine Learning*." (2021).

Tormen, Andréia Fátima, Gustavo Pansera, and Moacir Kripka. "Otimização das rotas para veículos de manutenção do sistema de iluminação pública na cidade de Passo Fundo (RS)." Exacta 16.3 (2018): 89-101.

Toth, Paolo, and Daniele Vigo, eds. Vehicle routing: problems, methods, and applications. Society for industrial and applied mathematics, 2014.

Van Duin, David, and David L. Paterson. "Multidrug-resistant bacteria in the community: trends and lessons learned." Infectious disease clinics 30.2 (2016): 377-390.

Woroniuk, Clare, et al. "Time series analysis of rail freight services by the private sector in Europe." Transport policy 25 (2013): 81-93.

Wu, Jia, et al. "Hyperparameter optimization for *Machine Learning* models based on Bayesian optimization." Journal of Electronic Science and Technology 17.1 (2019): 26-40

ANEXO A

Tabelas do *Osquare* dos modelos de Regressão

Nos anexos, são apresentados os resultados detalhados dos diferentes algoritmos de regressão aplicados, com base nas métricas avaliadas ao longo de várias datas e veículos. Estes dados incluem as métricas de *accuracy*, coeficiente de *Kendall*, *Edit Distance*, distância total (em quilómetros) e tempo total (em horas), permitindo uma visão abrangente do desempenho de cada modelo.

A tabela relativa à Regressão Linear mostra uma variação significativa nas métricas entre diferentes datas e veículos. A *accuracy* apresenta valores elevados, sugerindo que o modelo foi eficaz na previsão das rotas, com uma média próxima de 0,95. No entanto, o coeficiente de *Kendall*, que mede a correlação entre a ordem dos destinos previstos e os destinos reais, oscila entre valores positivos e negativos, o que indica inconsistências na ordenação das rotas. A *Edit Distance*, que avalia o número de operações necessárias para alinhar a rota prevista com a rota real, também varia de forma considerável, assim como as métricas de distância total e tempo total.

Para o modelo de *Random Forest*, observamos também valores consistentes de *accuracy*, com médias geralmente acima de 0,90, refletindo uma boa capacidade preditiva. O coeficiente de *Kendall* apresenta uma variação menor em relação à regressão linear, sugerindo que o modelo consegue capturar melhor as relações entre os destinos, embora ainda haja espaço para melhorias. A *Edit Distance* mantém-se em níveis moderados, com algumas exceções, e as variações nas métricas de distância total e tempo total revelam que o modelo enfrenta desafios em algumas situações de maior complexidade.

No caso da Regressão com Redes Neurais, a *accuracy* também é elevada, mas o coeficiente de *Kendall* apresenta uma tendência negativa, o que indica que o modelo enfrenta dificuldades em ordenar corretamente os destinos previstos. Esta discrepância é evidente, especialmente quando comparada com os outros modelos. A *Edit Distance* apresenta valores elevados em algumas instâncias, sugerindo que o modelo necessita de ajustes adicionais para melhorar a correspondência entre as rotas previstas e as reais. As métricas de distância total e tempo total também apresentam variabilidade significativa, o que pode ser um reflexo da complexidade dos dados e da natureza não linear das redes neurais.

Em resumo, os anexos fornecem uma visão detalhada do desempenho de cada algoritmo ao longo de diferentes situações, revelando tanto os pontos fortes quanto as áreas que necessitam de melhorias. Estes dados complementam as análises realizadas no corpo principal da tese e são essenciais para uma compreensão completa das capacidades preditivas dos modelos implementados

A.1 Regressão Linear

Tabela A. 1: Resultados detalhados do modelo de Regressão Linear

Data	Matricula	*Accuracy*	*Kendall*	*Edit Distance*	Distância Total (km)	Tempo Total (h)
14/03/2022	43-TC-65	0,9615	0,1482	121	117,8	1,9633
16/03/2022	39-10-RI	0,9583	0,0586	105	26,4867	0,4414
17/03/2022	63-SJ-29	0,9091	0,1214	44	83,3693	1,3895
17/03/2022	99-XR-26	0,8889	0,1619	27	108,1032	1,8017
18/03/2022	63-SJ-29	0,9643	0,1506	131	102,2766	1,7046
22/03/2022	38-99-RI	0,6667	0,0945	17	54,8289	0,9138
22/03/2022	39-00-RI	0,9444	0,11	73	93,0834	1,5514
24/03/2022	43-TC-65	0,9444	0,1256	67	49,0717	0,8179
24/03/2022	99-87-RG	0,9688	0,0997	156	108,8097	1,8135
25/03/2022	39-00-RI	0,9565	0,0493	104	31,6342	0,5272
25/03/2022	43-TC-65	0,9231	0,0997	40	239,9821	3,9997
28/03/2022	99-87-RG	0,9630	0,0997	40	239,9821	3,9997
28/03/2022	39-00-RI	0,9630	0,1361	121	92,6943	1,5449
30/03/2022	99-87-RG	0,8333	0,249	25	68,0909	1,1348
01/04//2022	38-99-RI	0,9524	0,1616	104	108,003	1,8
04/04//2022	09-PZ-62	0,9474	0,1457	79	135,5084	2,2585
04/04/2022	39-10-RI	0,9565	0,1614	94	104,0741	1,7346
05/04/2022	39-00-RI	0,9565	0,0883	102	120,5108	2,009
05/04/2022	39-10-RI	0,96	0,0621	104	27,484	0,4581
05/04/2022	43-TC-65	0,9643	0,1309	136	128,08	2,13
06/04/2022	AA-41-RS	0,9474	0,0801	72	69,4162	1,156
07/04/2022	09-PZ-62	0,909	0,0683	58	34,865	0,5811
07/04/2022	39-00-RI	0,9524	0,1726	93	114,2429	1,9
11/04/2022	09-PZ-62	0,9444	0,1673	67	103,22	1,72
11/04/2022	39-10-RI	0,9643	0,1549	122	70,2025	1,17
11/04/2022	43-TC-65	0,9643	0,1535	123	112,47	1,8745
11/04/2022	AA-41-RS	0,8889	0,1479	41	241,2269	4,02

A.2 Regressão *Random Forest*

Tabela A. 2: Resultados detalhados do modelo de Regressão *Random Forest*

Data	Matricula	*Accuracy*	*Kendall*	*Edit Distance*	Distância Total (km)	Tempo Total (h)
14/03/2022	43-TC-65	0,93	0,1844	102	116,34	1,94
16/03/2022	39-10-RI	0,9583	0,1325	111	25,03	0,42
17/03/2022	63-SJ-29	0,9091	0,0336	42	105,54	1,76
17/03/2022	99-XR-26	0,8889	0,1463	38	155,43	2,59
18/03/2022	63-SJ-29	0,9643	0,1672	122	113,24	1,89
22/03/2022	38-99-RI	0,6667	0,2835	17	54,83	0,91
22/03/2022	39-00-RI	0,9474	0,0664	85	160,54	2,68
24/03/2022	43-TC-65	0,944	0,1431	78	115,24	1,92
24/03/2022	99-87-RG	0,9444	0,1619	65	60,75	1,01
25/03/2022	39-00-RI	0,9688	0,1337	152	110,81	1,85
25/03/2022	43-TC-65	0,9565	0,0658	101	36,1	0,6
28/03/2022	99-87-RG	0,9231	0,1929	56	259,77	4,33
28/03/2022	39-00-RI	0,9630	0,1447	122	127,03	2,12
30/03/2022	99-87-RG	0,8333	0,1383	33	68,09	1,13
01/04//2022	38-99-RI	0,9524	0,1545	99	103,04	1,72
04/04//2022	09-PZ-62	0,9474	0,1814	78	147,61	2,46
04/04/2022	39-10-RI	0,9565	0,1571	101	128,77	2,15
05/04/2022	39-00-RI	0,9565	0,0838	102	150,2	2,5
05/04/2022	39-10-RI	0,96	0,1168	80	19,88	0,33
05/04/2022	43-TC-65	0,9643	0,1503	105	128,17	2,14
06/04/2022	AA-41-RS	0,9474	0,1111	86	71,81	1,2
07/04/2022	09-PZ-62	0,9091	0,0437	54	34,99	0,58
07/04/2022	39-00-RI	0,9524	0,1394	94	107,04	1,78
11/04/2022	09-PZ-62	0,9444	0,2217	85	105,16	1,75
11/04/2022	39-10-RI	0,9643	0,1388	143	90,19	1,5
11/04/2022	43-TC-65	0,9643	0,1537	116	205,67	3,43
11/04/2022	AA-41-RS	0,8889	0,192	21	238,01	3,97

A.3 Regressão Redes Neurais

Tabela A. 3: Resultados detalhados do modelo de Regressão Redes Neurais

Data	Matricula	*Accuracy*	*Kendall*	*Edit Distance*	Distância Total (km)	Tempo Total (h)
14/03/2022	43-TC-65	0,9630	-0,0714	111	112,6020	2,25
16/03/2022	39-10-RI	0,9583	-0,08	94	21,2680	0,43
17/03/2022	63-SJ-29	0,9091	-0,1668	54	103,3393	2,07
17/03/2022	99-XR-26	0,8889	-0,2003	48	111,8492	2,24
18/03/2022	63-SJ-29	0,9643	-0,0690	120	147,6717	2,95
22/03/2022	38-99-RI	0,6667	-0,5333	17	54,8289	1,1
22/03/2022	39-00-RI	0,9474	-0,1	79	122,14	2,44
24/03/2022	43-TC-65	0,9444	-0,1053	76	100,2385	2
24/03/2022	99-87-RG	0,9444	-0,1053	76	46,50	0,93
25/03/2022	39-00-RI	0,9688	-0,0606	142	105,3691	2,11
25/03/2022	43-TC-65	0,9565	-0,0833	90	28,3543	0,57
28/03/2022	99-87-RG	0,9231	-0,1429	48	241,007	4,82
28/03/2022	39-00-RI	0,963	-0,0714	130	104,5538	2,09
30/03/2022	99-87-RG	0,833	-0,2874	33	68,09	1,36
01/04//2022	38-99-RI	0,9524	-0,0909	107	142,279	2,85
04/04//2022	09-PZ-62	0,9474	-0,1	92	154,1069	3,08
04/04/2022	39-10-RI	0,9565	-0,083	100	128,142	2,56
05/04/2022	39-00-RI	0,9565	-0,083	96	141,51	2,83
05/04/2022	39-10-RI	0,96	-0,076	107	16,89	0,34
05/04/2022	43-TC-65	0,9643	-0,069	123	126,33	2,53
06/04/2022	AA-41-RS	0,9474	-0,1	65	69,745	1,39
07/04/2022	09-PZ-62	0,909	-0,1668	55	27,415	0,55
07/04/2022	39-00-RI	0,9524	-0,0909	85	163,0312	3,26
11/04/2022	09-PZ-62	0,9444	-0,1053	89	124,21	2,48
11/04/2022	39-10-RI	0,964	-0,06	120	81,9361	1,64
11/04/2022	43-TC-65	0,9643	-0,06	114	172,76	3,46
11/04/2022	AA-41-RS	0,8889	-0,2	42	241,478	4,83

ANEXO B

Tabelas do *Osquare* dos modelos de Classificação

Nos anexos relacionados aos modelos de classificação, os resultados obtidos com diferentes algoritmos foram organizados em tabelas, considerando as métricas de desempenho como *accuracy*, coeficiente de *Kendall*, *Edit Distance*, Distância Total (km) e Tempo Total (h). Essas métricas foram recolhidas para várias datas e veículos específicos, permitindo uma análise detalhada do desempenho de cada modelo em contextos reais.

No caso da Regressão Logística, os resultados demonstram uma variação significativa nas métricas, com a *accuracy* a oscilar entre valores relativamente baixos, como 0,2264 registado em 16 de março de 2022 para o veículo de matrícula 39-10-RI, e valores mais elevados como 0,7905 registado a 4 de abril de 2022 para o veículo 39-10-RI. O coeficiente de *Kendall* manteve-se consistentemente baixo, com a maioria dos valores sendo negativos, o que indica uma baixa correlação entre o comportamento previsto e o real. Por exemplo, o menor valor de *Kendall*, -0,0784, foi registado em 6 de abril de 2022 para o veículo AA-41-RS. O *Edit Distance*, que mede o número de operações necessárias para transformar uma sequência na outra, apresentou resultados como 3289 a 14 de março de 2022 e 3059 a 25 de março de 2022, o que reflete a complexidade das previsões feitas pelo modelo. A Distância Total (km) e o Tempo Total (h) variaram também consideravelmente, com a distância total a variar entre 37,37 km e 226,85 km, enquanto o tempo total variou de 0,53 h a 4,54 h.

O modelo de *Random Forest*, por outro lado, apresentou desempenhos superiores em termos de *accuracy*, com valores acima de 0,9 em várias ocasiões. A 14 de março de 2022, o veículo 43-TC-65 atingiu uma *accuracy* de 0,9534, e a 5 de abril de 2022, o veículo 39-00-RI obteve uma *accuracy* de 0,9317. Apesar desses resultados positivos, o coeficiente de *Kendall* manteve-se geralmente negativo, indicando uma baixa concordância ordinal, embora com menor intensidade em relação à Regressão Logística. Por exemplo, o valor de *Kendall* para o veículo 99-87-RG, em 30 de março de 2022, foi de -0,0821. O *Edit Distance* para *Random Forest* foi consideravelmente menor em comparação com a Regressão Logística, refletindo previsões mais precisas. As Distâncias Totais (km) variaram entre 7,08 km e 200,09 km, e o Tempo Total (h) também registou variações, com valores como 0,14 h e 4 h.

Por fim, o modelo SVM (*Support Vector Machine*) apresentou resultados de *accuracy* mais baixos, com valores como 0,1815 a 14 de março de 2022 para o veículo 43-TC-65, e 0,7222 registados em 11 de abril de 2022 para o veículo AA-41-RS. Tal como nos outros modelos, o coeficiente de *Kendall* foi negativo em várias iterações, com exceção de alguns poucos casos em que se aproximou de zero, refletindo a baixa correlação ordinal. O *Edit Distance* teve oscilações significativas, sendo um dos maiores valores de 242 registado a 18 de março de 2022 para o veículo 63-SJ-29. A Distância Total (km) e o Tempo Total (h) mostraram padrões semelhantes aos outros modelos, com variações consideráveis, como os 228,46 km percorridos a 11 de abril de 2022, e tempos que variaram entre 0,33 h e 4,57 h. Estes resultados evidenciam que, apesar das diferenças entre os modelos, o *Random Forest* mostrou-se mais robusto em termos de *accuracy*, enquanto os demais modelos, como a Regressão Logística e o SVM, apresentaram maior variabilidade no desempenho das previsões.

B.1 Classificação Regressão Logística

Tabela B. 1: Resultados detalhados do modelo de Classificação Regressão Logística

Data	Matrícula	*Accuracy*	*Kendall*	*Edit Distance*	Distância Total (km)	Tempo Total (h)
14/03/2022	43-TC-65	0,3615	-0,0296	3289	168,634	3,37
16/03/2022	39-10-RI	0,2264	-0,0524	3503	40,07	0,8
17/03/2022	63-SJ-29	0,6	-0,0651	302	86,429	1,73
17/03/2022	99-XR-26	0,5833	-0,0313	225	110,8873	2,22
18/03/2022	63-SJ-29	0,6429	-0,0005	2081	142,9399	2,86
22/03/2022	38-99-RI	0,5	-0,533	8	54,828	1,1
22/03/2022	39-00-RI	0,6637	-0,0185	825	144,08	2,88
24/03/2022	43-TC-65	0,5686	-0,0156	1009	97,374	1,95
24/03/2022	99-87-RG	0,5621	-0,0294	1031	51,642	1,03
25/03/2022	39-00-RI	0,3438	-0,0285	5129	121,278	2,43
25/03/2022	43-TC-65	0,2708	-0,0498	3059	37,3768	0,75
28/03/2022	99-87-RG	0,6603	-0,0619	334	226,8538	4,54
28/03/2022	39-00-RI	0,6553	-0,0068	1825	151,9829	3,04
30/03/2022	99-87-RG	0,7667	-0,0115	45	68,0706	1,36
01/04//2022	38-99-RI	0,7571	-0,0075	709	126,0256	2,52
04/04//2022	09-PZ-62	0,7485	-0,0238	565	139,3185	2,79
04/04/2022	39-10-RI	0,7905	0,0071	747	132,1646	2,64
05/04/2022	39-00-RI	0,5237	-0,0210	1849	128,6463	2,57
05/04/2022	39-10-RI	0,3767	-0,0353	3067	26,3169	0,53
05/04/2022	43-TC-65	0,3783	-0,0249	3729	191,2965	3,83
06/04/2022	AA-41-RS	0,1579	-0,0784	2432	71,6506	1,43
07/04/2022	09-PZ-62	0,6545	-0,050	289	31,2399	0,62
07/04/2022	39-00-RI	0,7143	0,0001	882	123,5928	2,47
11/04/2022	09-PZ-62	0,7908	0,0036	480	110,6932	2,21
11/04/2022	39-10-RI	0,7222	0,0048	1631	87,847	1,76
11/04/2022	43-TC-65	0,4894	-0,0108	3057	179,1443	3,58
11/04/2022	AA-41-RS	0,7361	-0,0829	92	238,2283	4,76

B.2 Classificação *Random Forest*

Tabela B. 2: Resultados detalhados do modelo de Classificação *Random Forest*

Data	Matrícula	*Accuracy*	*Kendall*	*Edit Distance*	Distância Total (km)	Tempo Total (h)
14/03/2022	43-TC-65	0,9534	-0,0230	82	92,431	1,85
16/03/2022	39-10-RI	0,9239	-0,0733	66	7,727	0,15
17/03/2022	63-SJ-29	0,8636	-0,1668	106	85,8335	1,72
17/03/2022	99-XR-26	0,8194	-0,1594	90	115,22	2,3
18/03/2022	63-SJ-29	0,9511	-0,039	186	106,1138	2,12
22/03/2022	38-99-RI	0,8333	0,0001	42	54,8289	1,1
22/03/2022	39-00-RI	0,9281	-0,0692	82	79,5593	1,59
24/03/2022	43-TC-65	0,9281	-0,0692	114	14,6550	0,29
24/03/2022	99-87-RG	0,9597	-0,0337	122	76,225	1,52
25/03/2022	39-00-RI	0,5771	-0,0266	154	7,081	0,14
25/03/2022	43-TC-65	0,8269	-0,0208	98	139,7042	3,87
28/03/2022	99-87-RG	0,9302	-0,0451	106	71,6717	1,43
28/03/2022	39-00-RI	0,8	-0,2874	66	68,0706	1,36
30/03/2022	99-87-RG	0,9262	-0,0821	74	26,8783	0,54
01/04//2022	38-99-RI	0,9211	-0,0469	98	73,1828	1,46
04/04//2022	09-PZ-62	0,9269	-0,0256	58	58,661	1,17
04/04/2022	39-10-RI	0,9289	-0,0614	74	107,0198	2,14
05/04/2022	39-00-RI	0,9317	-0,0522	90	10,557	0,21
05/04/2022	39-10-RI	0,9497	-0,0243	162	146,9653	2,94
05/04/2022	43-TC-65	0,9	-0,0895	66	7,4434	0,15
06/04/2022	AA-41-RS	0,8545	-0,1358	58	8,2075	0,16
07/04/2022	09-PZ-62	0,9333	-0,0732	74	60,6987	1,21
07/04/2022	39-00-RI	0,9085	-0,024	98	33,2105	0,66
11/04/2022	09-PZ-62	0,9114	-0,0308	66	9,3124	0,19
11/04/2022	39-10-RI	0,9365	-0,0541	58	62,8627	1,26
11/04/2022	43-TC-65	0,9311	-0,0288	65	9,213	0,23
11/04/2022	AA-41-RS	0,875	-0,0156	58	200,0954	4

B.3 Classificação SVM

Tabela B. 3: Resultados detalhados do modelo de Classificação SVM

Data	Matrícula	*Accuracy*	*Kendall*	*Edit Distance*	Distância Total (km)	Tempo Total (h)
14/03/2022	43-TC-65	0,1815	- 0,0534	226	112,1646	2,24
16/03/2022	39-10-RI	0,0833	-0,073	210	20,5066	0,41
17/03/2022	63-SJ-29	0,4091	-0,0587	106	80,9046	1,62
17/03/2022	99-XR-26	0,5556	-0,0196	90	108,0944	2,16
18/03/2022	63-SJ-29	0,1349	-0,0548	242	85,1569	1,7
22/03/2022	38-99-RI	0,5	-0,5333	42	54,8289	1,1
22/03/2022	39-00-RI	0,2456	-0,0627	170	125,34	2,51
24/03/2022	43-TC-65	0,2549	-0,0648	162	92,5978	1,85
24/03/2022	99-87-RG	0,1144	-0,0922	162	46,4081	0,93
25/03/2022	39-00-RI	0,1502	-0,0464	274	108,4035	2,17
25/03/2022	43-TC-65	0,087	-0,0758	202	28,6963	0,57
28/03/2022	99-87-RG	0,6154	-0,0134	90	189,7327	3,79
28/03/2022	39-00-RI	0,2194	-0,0452	234	86,3838	1,73
30/03/2022	99-87-RG	0,4667	-0,1356	66	68,0909	1,36
01/04//2022	38-99-RI	0,1667	-0,0682	186	81,8188	1,64
04/04//2022	09-PZ-62	0,3392	-0,0481	162	118,1934	2,36
04/04/2022	39-10-RI	0,1779	-0,0607	202	100,4664	2,01
05/04/2022	39-00-RI	0,2036	-0,0574	202	124,5362	2,49
05/04/2022	39-10-RI	0,08	-0,0705	218	16,6792	0,33
05/04/2022	43-TC-65	0,1693	-0,0509	242	119,6505	2,39
06/04/2022	AA-41-RS	0,1901	-0,0720	170	68,8977	1,38
07/04/2022	09-PZ-62	0,1818	-0,1334	106	27,3920	0,55
07/04/2022	39-00-RI	0,2952	-0,0496	186	114,3199	2,29
11/04/2022	09-PZ-62	0,4935	-0,0127	98	43,9130	0,88
11/04/2022	39-10-RI	0,1257	-0,0561	242	67,845	1,36
11/04/2022	43-TC-65	0,246	-0,0408	242	110,4519	2,21
11/04/2022	AA-41-RS	0,7222	-0,0164	90	228,4670	4,57

ANEXO C

Tabelas do *Osquare* dos modelos de Regressão com frequência

Nos anexos relacionados à análise dos modelos de regressão com base na frequência de visita, foram compilados resultados para os algoritmos de Regressão Linear, *Random Forest* e Redes Neurais. Estes modelos foram avaliados através das métricas de *accuracy*, coeficiente de *Kendall*, *Edit Distance*, Distância Total (km) e Tempo Total (h), em diferentes datas e veículos.

Na Regressão Linear com frequência, o modelo apresentou altos valores de *accuracy* em várias ocasiões, com um desempenho particularmente forte a 25 de março de 2022, onde o veículo 39-00-RI atingiu uma *accuracy* de 0,9859, e a 18 de março de 2022, com o veículo 63-SJ-29 a registar um valor de 0,9828. O coeficiente de *Kendall* variou entre 0,2449 e 0,7303, indicando uma correlação moderada entre o comportamento previsto e o real. O *Edit Distance*, que mede a dissimilaridade entre as sequências, apresentou valores como 17 a 22 de março de 2022, para o veículo 39-00-RI, e 134 a 18 de março de 2022, para o veículo 63-SJ-29. A Distância Total (km) variou significativamente, com valores como 31,6586 km a 16 de março de 2022 e 224,1416 km a 18 de março de 2022, enquanto o Tempo Total (h) variou de 0,4714 h a 5 de abril de 2022 para o veículo 39-10-RI, até 3,7357 h a 18 de março de 2022 para o veículo 63-SJ-29.

O modelo de *Random Forest*, ao ser aplicado com a frequência de visita, também apresentou *accuracy* consistentemente alta, como a 22 de março de 2022, onde o veículo 39-00-RI atingiu 0,9708, e a 25 de março de 2022, com o veículo 43-TC-65 alcançando 0,9664. O coeficiente de *Kendall* teve variação moderada, com o valor mais alto de 0,7303 a 22 de março de 2022 para o veículo 38-99-RI. O *Edit Distance* foi, em geral, menor comparado com a regressão linear, sugerindo uma maior precisão na previsão das sequências, com valores como 17 registados várias vezes. A Distância Total (km) oscilou entre 22,71 km a 5 de abril de 2022 para o veículo 39-00-RI e 238,23 km a 11 de abril de 2022 para o veículo AA-41-RS. O Tempo Total (h) também variou, registando-se o valor mais baixo de 0,38 h a 5 de abril de 2022 e o mais alto de 4,02 h a 28 de março de 2022 para o veículo 99-87-RG.

Por fim, o modelo de Redes Neurais apresentou uma *accuracy* ligeiramente inferior aos outros modelos, mas ainda assim com desempenhos sólidos, como a 25 de março de 2022, onde o veículo 39-00-RI registou 0,9899, e a 11 de abril de 2022, onde o veículo 39-10-RI atingiu 0,9828. O coeficiente de *Kendall* foi geralmente mais baixo, refletindo uma menor correlação entre os comportamentos previstos e reais, com valores como -0,0033 a 7 de abril de 2022 e 0,0657 a 11 de abril de 2022. O *Edit Distance* variou consideravelmente, com valores como 150 registados a 25 de março de 2022 e 17 a 22 de março de 2022. A Distância Total (km) variou entre 23,2997 km a 5 de abril de 2022 e 238,2279 km a 11 de abril de 2022 para o veículo AA-41-RS, e o Tempo Total (h) oscilou de 0,47 h a 5 de abril de 2022 para 4,82 h a 28 de março de 2022.

Em resumo, entre os três modelos, o *Random Forest* apresentou uma melhor *accuracy* de forma mais consistente, enquanto a Regressão Linear demonstrou uma maior variabilidade nos resultados, e as Redes Neurais tiveram desempenhos satisfatórios, mas com maior instabilidade nas previsões

C.1 Regressão Linear com Frequência

Tabela C. 1: Resultados detalhados do modelo de Regressão Linear com frequência

Data	Matrícula	*Accuracy*	*Kendall*	*Edit Distance*	Distância Total (km)	Tempo Total (h)
14/03/2022	43-TC-65	0,9769	0,2649	66	157,4	2,623
16/03/2022	39-10-RI	0,9638	0,2759	98	31,6586	0,5276
17/03/2022	63-SJ-29	0,9455	0,3988	51	102,485	1,7081
17/03/2022	99-XR-26	0,9167	0,4304	17	110,8873	1,8481
18/03/2022	63-SJ-29	0,9828	0,2601	134	224,1416	3,7357
22/03/2022	38-99-RI	0,8333	0,7303	17	54,8289	0,9138
22/03/2022	39-00-RI	0,9737	0,3115	70	198,7024	3,3117
24/03/2022	43-TC-65	0,9673	0,3173	33	107,7643	1,7961
24/03/2022	99-87-RG	0,9641	0,3219	17	46,3123	0,7719
25/03/2022	39-00-RI	0,9859	0,2449	130	122,8904	2,0482
25/03/2022	43-TC-65	0,9664	0,2801	111	35,3947	0,5899
28/03/2022	99-87-RG	0,9551	0,3562	48	259,6214	4,327
28/03/2022	39-00-RI	0,9658	0,2593	118	160,885	2,681
30/03/2022	99-87-RG	0,8667	0,5121	17	68,0706	1,1345
01/04//2022	38-99-RI	0,969	0,2959	65	114,3481	1,9058
04/04//2022	09-PZ-62	0,9708	0,3113	87	132,2289	2,2038
04/04/2022	39-10-RI	0,9723	0,2837	105	163,6979	2,7283
05/04/2022	39-00-RI	0,9783	0,2824	107	180,4970	3
05/04/2022	39-10-RI	0,9667	0,27	106	28,2816	0,4714
05/04/2022	43-TC-65	0,9828	0,2596	121	174,146	2,902
06/04/2022	AA-41-RS	0,9503	0,3078	91	100,3767	1,6729
07/04/2022	09-PZ-62	0,9273	0,3931	33	41,4399	0,6907
07/04/2022	39-00-RI	0,9667	0,293	84	141,2616	2,3544
11/04/2022	09-PZ-62	0,9739	0,3186	54	106,9138	1,7819
11/04/2022	39-10-RI	0,9815	0,2592	98	153,409	2,5568
11/04/2022	43-TC-65	0,9762	0,2564	83	192,43	3,207
11/04/2022	AA-41-RS	0,9306	0,4271	17	238,2283	3,9705

C.2 Regressão *Random Forest* com Frequência

Tabela C. 2: Resultados detalhados do modelo de Regressão *Random Forest* com frequência

Data	Matrícula	*Accuracy*	*Kendall*	*Edit Distance*	Distância Total (km)	Tempo Total (h)
14/03/2022	43-TC-65	0,9662	0,2751	80	143,21	2,39
16/03/2022	39-10-RI	0,9674	0,2755	98	32,71	0,55
17/03/2022	63-SJ-29	0,9455	0,4016	55	110,73	1,85
17/03/2022	99-XR-26	0,9583	0,4588	17	110,89	1,85
18/03/2022	63-SJ-29	0,9828	0,2612	102	156,3	2,61
22/03/2022	38-99-RI	0,8333	0,7303	17	54,83	0,91
22/03/2022	39-00-RI	0,9708	0,3133	72	175,89	2,93
24/03/2022	43-TC-65	0,9641	0,3176	33	108,8	1,81
24/03/2022	99-87-RG	0,9837	0,3230	17	46,31	0,77
25/03/2022	39-00-RI	0,9849	0,2449	146	126,18	2,1
25/03/2022	43-TC-65	0,9664	0,2776	113	30,45	0,51
28/03/2022	99-87-RG	0,9231	0,3499	48	241,03	4,02
28/03/2022	39-00-RI	0,9	0,5088	17	68,07	1,13
30/03/2022	99-87-RG	0,969	0,2986	88	124,52	2,08
01/04//2022	38-99-RI	0,9737	0,313	85	125,13	2,09
04/04//2022	09-PZ-62	0,9783	0,288	47	101,08	1,68
04/04/2022	39-10-RI	0,9684	0,2823	103	164,68	2,74
05/04/2022	39-00-RI	0,9633	0,2688	53	22,71	0,38
05/04/2022	39-10-RI	0,9802	0,2629	130	163,91	2,73
05/04/2022	43-TC-65	0,9708	0,3095	66	68,2	1,14
06/04/2022	AA-41-RS	0,9364	0,3903	41	39,86	0,66
07/04/2022	09-PZ-62	0,9643	0,2945	71	133,54	2,23
07/04/2022	39-00-RI	0,9739	0,3225	54	106,91	1,78
11/04/2022	09-PZ-62	0,9828	0,2622	114	95,95	1,6
11/04/2022	39-10-RI	0,9696	0,2678	140	136,23	2,27
11/04/2022	43-TC-65	0,9456	0,312	84	142,26	2,25
11/04/2022	AA-41-RS	0,9306	0,4426	17	238,23	3,97

C.3 Regressão Redes Neurais com Frequência

Tabela C. 3: Resultados detalhados do modelo de Regressão Redes Neurais com frequência

Data	Matrícula	*Accuracy*	*Kendall*	*Edit Distance*	Distância Total (km)	Tempo Total (h)
14/03/2022	43-TC-65	0,9754	0,0025	76	137,6097	2,75
16/03/2022	39-10-RI	0,971	0,0294	122	25,3129	0,51
17/03/2022	63-SJ-29	0,9727	0,1274	25	76,9116	1,54
17/03/2022	99-XR-26	0,9444	0,0462	17	110,8871	2,22
18/03/2022	63-SJ-29	0,9854	0,0326	130	123,9312	2,48
22/03/2022	38-99-RI	0,8333	0	17	54,8289	1,1
22/03/2022	39-00-RI	0,9737	0,0215	76	174,9441	3,5
24/03/2022	43-TC-65	0,9706	0,0645	33	107,7641	2,16
24/03/2022	99-87-RG	0,9739	0,03	17	46,3122	0,93
25/03/2022	39-00-RI	0,9899	0,0367	150	112,9285	2,26
25/03/2022	43-TC-65	0,9644	0,0133	106	28,7624	0,58
28/03/2022	99-87-RG	0,9423	0,0625	48	241,0291	4,82
28/03/2022	39-00-RI	0,9658	-0,065	66	102,9386	2,06
30/03/2022	99-87-RG	0,9	0,0437	17	68,07	1,36
01/04//2022	38-99-RI	0,9643	-0,0011	66	106,8733	2,14
04/04//2022	09-PZ-62	0,9708	0,0104	30	130,4556	2,61
04/04/2022	39-10-RI	0,9763	0,0364	88	123,135	2,46
05/04/2022	39-00-RI	0,9763	0,0364	99	164,209	3,28
05/04/2022	39-10-RI	0,9683	-0,0196	67	23,2997	0,47
05/04/2022	43-TC-65	0,9815	0,042	129	162,7118	3,25
06/04/2022	AA-41-RS	0,9649	0,0413	48	72,8149	1,46
07/04/2022	09-PZ-62	0,9364	-0,0033	33	41,4399	0,83
07/04/2022	39-00-RI	0,9643	0,0421	85	124,4167	2,49
11/04/2022	09-PZ-62	0,9771	0,0657	33	103,2723	2,07
11/04/2022	39-10-RI	0,9828	0,0324	95	98,302	1,97
11/04/2022	43-TC-65	0,9735	0,0167	76	133,6796	2,67
11/04/2022	AA-41-RS	0,9306	0,0423	17	238,2279	4,76

ANEXO D

Tabelas do *Osquare* dos modelos de Classificação com frequência

As tabelas apresentadas mostram os resultados da classificação utilizando três diferentes algoritmos (*Random Forest*, Regressão Logística e SVM), aplicados com base na frequência de visitas para as rotas. Para cada algoritmo, foram calculadas várias métricas, incluindo *accuracy*, coeficiente de *Kendall*, *Edit Distance*, distância total (km) e tempo total (h), permitindo uma análise detalhada do desempenho de cada modelo.

No caso do *Random Forest* com frequência, a *accuracy* variou entre 0,8333 e 0,9837, indicando um bom desempenho na previsão correta das rotas. No entanto, o coeficiente de *Kendall* apresentou valores consistentemente baixos, sugerindo que o modelo teve dificuldade em prever a ordem correta das visitas. A *Edit Distance* variou consideravelmente, com valores entre 8 e 180, o que reflete as discrepâncias entre as rotas previstas e as rotas reais. Por fim, a distância total e o tempo total mostram que, como esperado, as rotas mais longas exigem mais tempo de viagem, o que também influencia as demais métricas de desempenho.

Na Regressão Logística com frequência, a *accuracy* foi ligeiramente mais consistente, variando entre 0,8333 e 0,9879, o que indica que este modelo teve um bom desempenho na previsão das rotas. O coeficiente de *Kendall* foi mais elevado do que no *Random Forest*, atingindo até 0,154 em alguns casos, sugerindo uma melhor correlação entre a ordem prevista e a ordem real das rotas. A *Edit Distance* também apresentou uma variação considerável, com picos em dias específicos, como 28/03/2022, onde atingiu o valor de 150. Os valores de distância total e tempo total mostraram um padrão semelhante ao do *Random Forest*, reforçando a relação entre o tempo de viagem e a extensão da rota.

O SVM com frequência apresentou resultados semelhantes aos outros dois modelos em termos de *accuracy*, com valores entre 0,8333 e 0,9869. No entanto, o coeficiente de *Kendall* foi superior em alguns momentos, como no dia 28/03/2022, quando atingiu o valor de 0,9, indicando uma melhor previsão da ordem das rotas em determinados cenários. A *Edit Distance* variou entre 21 e 242, com algumas discrepâncias notáveis em dias como 05/04/2022 e 11/04/2022. A distância total e o tempo total seguiram o mesmo padrão de variação dos outros modelos, confirmando a correlação entre o tempo de viagem e a extensão das rotas previstas.

Em resumo, os três algoritmos apresentaram desempenhos satisfatórios em termos de *accuracy*, mas ao analisar métricas mais detalhadas, como o coeficiente de *Kendall* e a *Edit Distance*, é possível observar diferenças significativas. Enquanto a Regressão Logística e o SVM tiveram melhor desempenho na ordenação das visitas, o *Random Forest* mostrou uma maior variabilidade nas suas previsões. Essas análises são essenciais para determinar qual o modelo mais adequado para a previsão de rotas baseadas na frequência de visitas e no comportamento do motorista

D.1 Classificação *Random Forest* com Frequência

Tabela D. 1: Resultados detalhados do modelo de Classificação *Random Forest* com frequência

Data	Matrícula	*Accuracy*	*Kendall*	*Edit Distance*	Distância Total (km)	Tempo Total (h)
14/03/2022	43-TC-65	0,9446	0	156	137,1483	2,74
16/03/2022	39-10-RI	0,9583	0,0214	95	27,1711	0,54
17/03/2022	63-SJ-29	0,9273	0,0254	37	103,0516	2,06
17/03/2022	99-XR-26	0,9028	0,0767	54	110,8873	2,22
18/03/2022	63-SJ-29	0,9775	0,0318	80	179,7319	3,59
22/03/2022	38-99-RI	0,8333	0	8	54,8289	1,1
22/03/2022	39-00-RI	0,9620	0,0305	64	139,9236	2,8
24/03/2022	43-TC-65	0,951	0,0267	64	109,6581	2,19
24/03/2022	99-87-RG	0,9837	0,0786	45	46,3123	0,93
25/03/2022	39-00-RI	0,9778	0,0167	85	133,404	2,67
25/03/2022	43-TC-65	0,9348	0,0242	180	31,2868	0,63
28/03/2022	99-87-RG	0,8718	0,0462	136	228,0299	4,56
28/03/2022	39-00-RI	0,9459	-0,0078	164	143,7165	2,87
30/03/2022	99-87-RG	0,9	0,0437	38	68,0706	1,36
01/04//2022	38-99-RI	0,9643	0,0249	75	108,5577	2,17
04/04//2022	09-PZ-62	0,9444	0,0073	103	146,3707	2,93
04/04/2022	39-10-RI	0,9625	0,0203	75	142,8226	2,86
05/04/2022	39-00-RI	0,9605	0,0415	109	134,6713	2,69
05/04/2022	39-10-RI	0,9517	-0,0023	135	29,2749	0,59
05/04/2022	43-TC-65	0,9683	0,0260	135	190,5249	3,81
06/04/2022	AA-41-RS	0,9444	0,0177	104	72,7839	1,46
07/04/2022	09-PZ-62	0,9182	0,0234	54	30,6849	0,61
07/04/2022	39-00-RI	0,9571	0,0325	70	124,3075	2,49
11/04/2022	09-PZ-62	0,9477	0,0715	134	96,125	1,92
11/04/2022	39-10-RI	0,9722	0,0458	137	130,9372	2,62
11/04/2022	43-TC-65	0,9524	-0,0046	136	199,392	3,99
11/04/2022	AA-41-RS	0,9028	0,0767	52	238,2283	4,76

D.2 Classificação Regressão Logística com Frequência

Tabela D. 2: Resultados detalhados do modelo de Classificação Regressão Logística com frequência

Data	Matrícula	*Accuracy*	*Kendall*	*Edit Distance*	Distância Total (km)	Tempo Total (h)
14/03/2022	43-TC-65	0,9538	0,0616	267	128,801	2,58
16/03/2022	39-10-RI	0,9746	0,0628	120	30,1151	0,6
17/03/2022	63-SJ-29	0,9636	0,1244	41	77,901	1,56
17/03/2022	99-XR-26	0,9444	0,1338	50	110,8873	2,22
18/03/2022	63-SJ-29	0,9584	0,0620	88	122,889	2,46
22/03/2022	38-99-RI	0,8333	0	8	54,828	1,1
22/03/2022	39-00-RI	0,9708	0,0837	90	139,9323	2,8
24/03/2022	43-TC-65	0,9673	0,0867	80	107,1818	2,14
24/03/2022	99-87-RG	0,9869	0,0908	41	46,3123	0,93
25/03/2022	39-00-RI	0,9879	0,0554	89	118,1158	2,36
25/03/2022	43-TC-65	0,9466	0,0535	179	30,448	0,61
28/03/2022	99-87-RG	0,8846	0,0882	150	259,6214	5,19
28/03/2022	39-00-RI	0,9630	0,0609	234	105,7069	2,11
30/03/2022	99-87-RG	0,9333	0,154	21	68,0706	1,36
01/04//2022	38-99-RI	0,9738	0,0691	83	100,2898	2,01
04/04//2022	09-PZ-62	0,9678	0,0831	100	129,0167	2,58
04/04/2022	39-10-RI	0,9763	0,0651	88	96,4547	1,93
05/04/2022	39-00-RI	0,9625	0,0698	153	155,418	3,11
05/04/2022	39-10-RI	0,97	0,0662	156	27,012	0,54
05/04/2022	43-TC-65	0,9788	0,0611	141	175,174	3,5
06/04/2022	AA-41-RS	0,9737	0,0739	91	71,6506	1,43
07/04/2022	09-PZ-62	0,9545	0,1214	58	35,35	0,71
07/04/2022	39-00-RI	0,9667	0,0762	112	122,7549	2,46
11/04/2022	09-PZ-62	0,9575	0,0734	111	92,19	1,84
11/04/2022	39-10-RI	0,9735	0,0604	170	112,312	2,25
11/04/2022	43-TC-65	0,963	0,0542	225	148,2303	2,96
11/04/2022	AA-41-RS	0,9167	0,1228	68	238,2283	4,76

D.3 Classificação SVM com Frequência

Tabela D. 3: Resultados detalhados do modelo de Classificação SVM com frequência

Data	Matrícula	*Accuracy*	*Kendall*	*Edit Distance*	Distância Total (km)	Tempo Total (h)
14/03/2022	43-TC-65	0,9508	0,0611	90	75,2267	1,5
16/03/2022	39-10-RI	0,9746	0,0693	74	9,7908	0,2
17/03/2022	63-SJ-29	0,9636	0,1244	106	85,7504	1,72
17/03/2022	99-XR-26	0,9444	0,1338	90	110,8873	2,22
18/03/2022	63-SJ-29	0,9841	0,0619	218	87,3729	1,75
22/03/2022	38-99-RI	0,8333	0	42	54,8289	1,1
22/03/2022	39-00-RI	0,9708	0,0837	106	122,7011	2,44
24/03/2022	43-TC-65	0,9673	0,0867	138	97,6799	1,95
24/03/2022	99-87-RG	0,9869	0,0908	138	39,1179	0,78
25/03/2022	39-00-RI	0,9879	0,0554	154	81,2279	1,62
25/03/2022	43-TC-65	0,9506	0,0679	74	26,5161	0,53
28/03/2022	99-87-RG	0,891	0,9	74	189,9124	3,8
28/03/2022	39-00-RI	0,963	0,0609	218	82,2521	1,65
30/03/2022	99-87-RG	0,9333	0,154	66	68,07	1,36
01/04//2022	38-99-RI	0,9762	0,0779	58	21,8	0,44
04/04//2022	09-PZ-62	0,9678	0,0831	154	126,6137	2,53
04/04/2022	39-10-RI	0,9783	0,0725	202	98,8211	1,98
05/04/2022	39-00-RI	0,9625	0,0698	186	125,1097	2,5
05/04/2022	39-10-RI	0,97	0,0662	130	11,10	0,22
05/04/2022	43-TC-65	0,9788	0,0611	114	108,0813	2,16
06/04/2022	AA-41-RS	0,9708	0,0837	122	9,1224	0,18
07/04/2022	09-PZ-62	0,9545	0,1214	82	28,6255	0,57
07/04/2022	39-00-RI	0,9667	0,0762	106	94,5819	1,89
11/04/2022	09-PZ-62	0,9542	0,0840	146	85,2473	1,7
11/04/2022	39-10-RI	0,9735	0,0604	242	91,1959	1,82
11/04/2022	43-TC-65	0,9643	0,0591	50	21,4975	0,43
11/04/2022	AA-41-RS	0,9167	0,1228	58	191,4256	3,83