INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA

ÁQUILA SAMUEL AZEVEDO DIAS

MODELO PREDITIVO PARA INDICADOR DE QUALIDADE NO SETOR DE TRANSPORTE POR APLICATIVO.

CAMPINA GRANDE
NOVEMBRO, 2023

CAMPINA GRANDE
NOVEMBRO, 2023

D541m Dias, Áquila Samuel Azevedo
Modelo preditivo para indicador de qualidade no setor de transporte por aplicativo / Áquila Samuel Azevedo Dias. - Campina Grande, 2023.
32f. : il.

Trabalho de Conclusão de Curso (Curso Superior de Engenharia de Computação) - Instituto Federal da Paraíba, 2023.
Orientador: Prof. Dr. Elmano Ramalho Cavalcanti

1. Processamento de dados 2. Gestão de qualidade 3. Aprendizado de máquina I. Cavalcanti, Elmano Ramalho II. Título.

CDU 004.4:658.6

ÁQUILA SAMUEL AZEVEDO DIAS

MODELO PREDITIVO PARA INDICADOR DE QUALIDADE NO SETOR DE TRANSPORTE POR APLICATIVO.

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Engenharia de Computação na Área do Conhecimento de Ciências Exatas e Engenharias do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, em Campina Grande – PB.

Orientador: Prof. Dr. Elmano Ramalho Cavalcanti

CAMPINA GRANDE
NOVEMBRO, 2023

AGRADECIMENTOS

Agradeço, primeiramente a Deus, por sempre me orientar nos momentos de decisão, acalmar nos momentos de tensão e encorajar nos momentos em que o desânimo quis bater à porta da minha mente e do meu coração. Como disse o salmista no Salmo 94:19 “Quando o medo já me dominava no íntimo, o teu consolo trouxe alívio à minha alma.” Com a ajuda de Deus, cheguei até aqui. Tenho plena convicção que sem a graça, a misericórdia e a ajuda de Deus, eu não conseguiria.

A minha família, principalmente aos meus pais: João Ednaldo de Medeiros Dias, Maria da Conceição Azevedo Dias, e ao meu irmão: Aleff Azevedo Dias, que são as minhas bases e meus alicerces. Sempre me deram apoio e suporte necessário para alcançar esse tão sonhado objetivo.

Ao meu orientador, professor Elmano Ramalho, que topou esse desafio tão importante para mim. Sempre com paciência me orientando, mesmo na sua correria diária. Muito obrigado, professor.

CAMPINA GRANDE
NOVEMBRO, 2023

“Confie ao Senhor tudo que você faz, e seus planos serão bem-sucedidos.”

– Provérbios 16:3

CAMPINA GRANDE
NOVEMBRO, 2023

RESUMO

O presente trabalho explora a aplicação do modelo de aprendizado de máquina, SVM, utilizando uma classe específica de regressão, SVR, na previsão das futuras avaliações de qualidade de uma empresa de transporte por aplicativo. Utilizando um conjunto de dados das avaliações de qualidade, dentro do período de seis meses, o projeto emprega aprendizado de máquina supervisionado para treinar o algoritmo. Obtendo quase 100% de eficácia na explicação da variabilidade dos dados nesse contexto, impulsiona a expansão do algoritmo para outras áreas da empresa, visando a melhoria na eficiência da análise dos dados nas demais áreas.

Palavras-chaves: SVM. SVR. Avaliações de Qualidade. Transporte por Aplicativo. Aprendizado de Máquina. Análise de Dados.

CAMPINA GRANDE
NOVEMBRO, 2023

ABSTRACT

The present work explores the application of the machine learning model, SVM, using a specific class of regression, SVR, in predicting future quality assessments of a ride-hailing company. Using a dataset of quality assessments over a six-month period, the project employs supervised machine learning to train the algorithm. Obtaining almost 100% effectiveness in explaining data variability in this context, it drives the expansion of the algorithm to other areas of the company, aiming to improve the efficiency of data analysis in other areas.

Keywords: SVM. SVR. Quality Assessments. Transport by Application. Machine Learning. Data analysis.

CAMPINA GRANDE
NOVEMBRO, 2023

LISTA DE FIGURAS

CAMPINA GRANDE
NOVEMBRO, 2023

LISTA DE TABELAS

CAMPINA GRANDE
NOVEMBRO, 2023

Sumário

CAMPINA GRANDE
NOVEMBRO, 2023

1 INTRODUÇÃO

No cenário dinâmico dos serviços de transporte por aplicativo, entender e prever as avaliações de qualidade é fundamental para as empresas aprimorarem a satisfação do cliente e a eficiência operacional. No contexto corporativo, em todos os setores, principalmente no setor de transporte por aplicativo, na qual trataremos neste trabalho, empresas que fazem uso de inteligência artificial na análise de dados detém uma vantagem competitiva gigantesca no mercado de trabalho. Com a Inteligência Artificial (IA), é possível perceber padrões ocultos nos dados, podendo ser no perfil do cliente, operacional, comportamental etc.

Diante disso, observamos a necessidade de implementar o uso de aprendizagem de máquina para análise de dados internos, inicialmente nas notas de qualidade. Tendo em vista a forma que a empresa atua nas avaliações de qualidade, que é corrigir falhas pontuadas nos atendimentos somente após a conclusão das avaliações, ou seja: só consegue atuar com base em dados históricos. Foi perceptível a perca que a empresa sofria, sendo financeira ou quantitativa. Com o uso de aprendizagem de máquina foi possível reverter essa atuação partindo de corretiva para preventiva.

Apresentamos ainda, o aspecto prático do projeto, explicando a metodologia empregada no desenvolvimento do modelo de predição. Partindo da natureza quantitativa, utilizamos um conjunto de dados de seis meses de avaliações, fornecidos pela empresa na qual foi realizado o projeto para aprendizado de máquina supervisionado. A escolha do RVS é justificada de acordo com a sua eficiência no coeficiente de determinação para a base de dados, na qual foi aplicada, conforme apresentado na subseção 3.1.4. Na seção sobre metodologia, nos detivemos a esclarecer o processo de avaliação, realizado por monitores que escutam e avaliam os atendimentos realizados ao cliente, também o armazenamento de dados, extração, pré-processamento e a execução do algoritmo, para obter predições quantitativas de avaliação de qualidade.

A discussão se estende para as vantagens do RVS em lidar com relações complexas e sua flexibilidade por meio de hiper parâmetros. Reconhecemos, através desse trabalho, o papel crucial do pré-processamento de dados na obtenção dos resultados desejados. Dessa forma, para validação do modelo, são utilizadas métricas de avaliação de desempenho, como Erro Médio Quadrático (MSE) e Coeficiente de Determinação (R^2).

A seção de resultados fornece percepções sobre a aplicação do modelo de previsão, em uma fase piloto dentro de uma empresa de transporte por aplicativo. O desempenho satisfatório do algoritmo, com predições obtendo percentual significativo de assertividade, sustenta seu papel proativo na abordagem de questões de avaliação de qualidade.

Por fim, destacamos a aplicação bem-sucedida de aprendizado de máquina, especificamente RVS, na previsão de avaliações de qualidade para uma empresa de transporte por aplicativo. Os objetivos do projeto de intervenção proativa e medidas preventivas para melhorar as avaliações de qualidade foram alcançados com alto nível de precisão, atendendo a uma demanda crucial dentro da empresa.

1.1 OBJETIVO GERAL

O trabalho visa compreender razoavelmente o funcionamento das avaliações de qualidade no setor de transporte por aplicativo, e, com isso, realizar predições das notas por meio do modelo de aprendizagem de máquina, gerando a oportunidade de reversão da nota de qualidade até o final do ciclo.

1.2 OBJETIVOS ESPECÍFICOS

Para possibilitar a concretização do objetivo principal desse trabalho, citado anteriormente, foi necessário definir algumas etapas:

- Compreensão das definições teóricas que norteiam este trabalho;
    - Inteligência Artificial;
    - Aprendizagem de máquina;
    - Máquina de Vetor de Suporte.
- Conhecer o funcionamento das avaliações de nota de qualidade;
- Apresentar uso do algoritmo e sua eficácia.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção, abordaremos as definições de conceitos teóricos que norteiam e direcionam este trabalho. Experiência do Cliente, Métodos de Avaliação, Inteligência Artificial (IA), Aprendizado de Máquina (AM), Máquina de Vetor de Suporte (MVS), e métodos como Regressão de Vetores de Suporte (RVS), com destaque para sua aplicabilidade em problemas complexos e não lineares, resistência a outliers e controle de flexibilidade.

Trataremos também sobre a importância da divisão entre conjuntos de treinamento e teste, fornecendo uma medida realista do desempenho do modelo em novos dados. Também são apresentadas métricas de avaliação, como Erro Médio Quadrático (MSE) e Coeficiente de Determinação (R^2), fundamentais para a análise de desempenho de modelos de regressão.

2.1 EXPERIÊNCIA DO CLIENTE

Para Mehta, Murphy e Steinman (2020) o termo Experiência do Cliente designa especificamente a avaliação e a gestão da experiência total do cliente, ao longo do seu ciclo de vida.

Experiência do Cliente, portanto, refere-se ao conjunto de percepções e impressões que um cliente adquire em relação a uma empresa após realizar interações com ela. Essa experiência abrange desde o primeiro contato até a conclusão de uma transação ou serviço, moldando a visão e a satisfação do cliente em relação à marca.

2.1.1 MÉTODOS DE AVALIAÇÃO

Existem diversos métodos e métricas que possibilitam avaliar a experiência do cliente com uma determinada empresa. Por exemplo: Net Promoter Score (NPS) é uma métrica utilizada para medir a satisfação e lealdade dos clientes em relação a uma empresa ou produto (REICHHELD, 2003). A pontuação do NPS é obtida por meio de uma pergunta simples feita aos clientes: "Em uma escala de 0 a 10, o quanto você recomendaria nossa empresa a um amigo?".

Com base nas respostas, os clientes são categorizados em três grupos:

- Promotores (Pontuação de 9 a 10);
- Neutros (Pontuação de 7 a 8);
- Detratores (Pontuação de 0 a 6).

No presente trabalho, utilizamos outra métrica de avaliação: nota de qualidade. É uma métrica que avalia como está a qualidade do atendimento fornecido para com o cliente e o desempenho individual de cada operador no atendimento. Essa nota é gerada com base em itens, podemos dividir em 3 principais grupos que são considerados pela empresa como essenciais para obter a excelência do atendimento: conhecimento do produto/serviço, empatia para com o cliente e resolutividade. Dentro desses grupos, cada item tem um peso específico.

2.2 INTELIGÊNCIA ARTIFICIAL

Uma definição clássica de IA foi proposta por John McCarthy, um dos pioneiros no campo, que a descreve como "o ramo da ciência da computação que se ocupa do comportamento inteligente, ou seja, o comportamento que seria considerado inteligente se observado em humanos" (MCCARTHY, 2007, p. 1). Essa abordagem ressalta a aspiração da IA em replicar a inteligência humana em sistemas computacionais. A IA busca desenvolver algoritmos e modelos capazes de simular processos de aprendizado, raciocínio, percepção e tomada de decisão, características atribuídas à inteligência humana.

No contexto da aprendizagem de máquina, uma subárea crucial da IA, Tom Mitchell oferece uma definição mais específica: "Um programa de computador é dito aprender de experiência, e em relação a alguma tarefa T e alguma medida de desempenho P, se seu desempenho em T, medido por P, melhorar com a experiência E" (MITCHELL, 1997, p. 2). Essa definição enfatiza a capacidade dos sistemas de IA de aprimorar seu desempenho ao longo do tempo por meio da exposição a dados e experiências.

A evolução da IA ao longo do tempo tem sido marcada por avanços significativos, desde a concepção do termo "Inteligência Artificial" em 1956 até os progressos recentes impulsionados pelo aumento da capacidade computacional e

o acesso a grandes conjuntos de dados (MCCARTHY,1955, p.1). A IA encontra aplicação em diversas áreas, como reconhecimento de padrões, processamento de linguagem natural, visão computacional, diagnóstico médico, sistemas de recomendação e jogos, entre outros. O desenvolvimento contínuo da IA levanta questões éticas e sociais, incluindo considerações sobre privacidade, viés algorítmico e impactos no mercado de trabalho.

2.3 APRENDIZAGEM DE MÁQUINA

O aprendizado de máquina aborda a questão de como construir computadores que melhorem automaticamente por meio da experiência (MITCHELL, 2017). Os modelos de aprendizagem de máquina desempenham um papel crucial na capacidade de sistemas computacionais realizar tarefas variadas, desde reconhecimento de padrões até predições, que abordaremos neste trabalho.

No contexto da aprendizagem de máquina, é importante compreender os diferentes tipos de abordagens que os modelos podem adotar. Destacam-se três categorias principais: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço.

Exemplificando de modo básico, aprendizado supervisionado, os modelos são treinados com dados rotulados, ou seja, as saídas desejadas já são conhecidas. Nesse sentido, os modelos aprendem a mapear as entradas para as saídas corretas, sendo particularmente eficazes para tarefas de classificação e regressão. Isso permite que o modelo aprenda a fazer previsões precisas para novos dados. "O aprendizado supervisionado caracteriza-se pelo treinamento dos modelos com dados rotulados, onde as saídas desejadas já são conhecidas" (RUSSEL; NORVIG, 2010, p. 725).

Os modelos categorizados de aprendizado não supervisionado, exploram padrões nos dados sem rótulos. "os modelos exploram padrões nos dados sem a presença de rótulos, possibilitando a identificação de estruturas intrínsecas nos dados" (HASTIE; TIBSHIRANI; FRIEDMAN, 2009, p. 485). Isso é útil para descobrir grupos de avaliações similares sem a necessidade de categorias predefinidas.

Por fim, nos modelos categorizados de aprendizado por reforço, aprendem por tentativa e erro em um ambiente interativo. "os modelos aprendem por tentativa e erro em um ambiente interativo, recebendo recompensas ou penalidades com base nas ações executadas" (SUTTON; BARTO, 2018, p. 9). Eles recebem recompensas ou penalidades com base nas ações realizadas e ajustam seu comportamento para maximizar as recompensas.

2.4 MÁQUINA DE VETOR DE SUPORTE

A Máquina de Vetor de Suporte é um tipo específico de algoritmo de aprendizagem de máquina que pode ser aplicado tanto em tarefas de aprendizado supervisionado de classificação e regressão. A MVS tem como principal característica a capacidade de lidar eficazmente com problemas de separação de classes, buscando encontrar um hiperplano que melhor separa as classes de dados, maximizando a margem entre elas. A intuição por trás dessa abordagem é que, ao maximizar a distância entre as margens, o modelo tenha uma boa capacidade de generalizar bem para dados não vistos. (SÁNCHEZ, 2003)

As MVS operam em um contexto de aprendizado supervisionado, buscando encontrar um hiperplano que maximize a margem entre diferentes classes de dados. Tal abordagem é resumida por Cortes (1995), que define MVS como "método de aprendizado supervisionado que se baseia na ideia de encontrar o hiperplano de separação que maximize a margem entre as classes" (CORTES, 1995, p. 408).

A capacidade das MVS em lidar com dados não linearmente separáveis é um dos seus pontos fortes, sendo possibilitada pelo uso de truques de kernel. Como elucidado por Vapnik (2000), "os truques de kernel permitem que as MVS operem eficientemente em espaços de características de alta dimensão, convertendo implicitamente o problema não linear em um espaço de características superior" (VAPNIK, 2000, p. 220).

A versatilidade das MVS as tornam aplicáveis em uma variedade de cenários, desde problemas de classificação de texto até diagnósticos médicos. Em suas palavras, Cristianini e Shawe-Taylor enfatizam que "as MVS têm sido amplamente utilizadas em muitas áreas, incluindo reconhecimento de padrões, biologia computacional, processamento de linguagem natural e visão computacional" (CRISTIANINI; SHAWE-TAYLOR, 2000, p. 7).

Para apresentar uma impressão clara de sua aplicabilidade, seguem alguns exemplos de aplicações práticas do MVS: (SÁNCHEZ, 2003).

- **Classificação de Texto:**

Podem ser usadas para classificar documentos de texto em categorias, como spam ou não spam, categorias de notícias, sentimentos (positivo, negativo, neutro), entre outros.

- **Reconhecimento de Imagem:**

Em visão computacional, são úteis para tarefas de reconhecimento de imagem, como classificação de objetos, detecção de faces e reconhecimento de caracteres.

- **Biomedicina:**

Podem ser aplicadas na classificação de proteínas, diagnóstico médico, previsão de doenças e análise de expressão genética.

- **Sistemas de Recomendação:**

Podem ser empregadas em sistemas de recomendação para prever as preferências dos usuários com base em seu histórico de comportamento.

- **Diagnóstico Médico:**

Podem auxiliar no diagnóstico médico, ajudando a classificar pacientes em diferentes categorias com base em dados clínicos e de exames.

Podemos observar que o MVS é aplicável em diversos setores. Bastante importante destacar que a escolha do algoritmo para uso em análise de dados deve-se considerar a natureza dos dados e do problema específico que está sendo abordado. É importante considerar fatores como: dimensionalidade dos dados, quantidade de dados disponíveis e características do problema ao escolher o MVS ou qualquer outro algoritmo de aprendizado de máquina.

2.4.1 REGRESSÃO DE VETORES DE SUPORTE:

Dentro do MVS temos diversos métodos, e para o trabalho em questão foi utilizado o Regressão de Vetores de Suporte (RVS). É baseado na análise de regressão, que tem como objetivo gerar o número máximo de vetores de suporte com valores pequenos de erros a fim de separar os dados da melhor forma possível (CARMELO, 2017).

2.4.2 VANTAGENS DO RVS:

- Pode ser eficaz quando há relações complexas e não lineares entre as variáveis de entrada e saída;
- Ele pode lidar com dados de alta dimensionalidade e é resistente a outliers, desde que seja ajustado adequadamente;
- O RVS permite controlar a flexibilidade do modelo por meio de hiper parâmetros, como o kernel escolhido e os parâmetros de regularização.

2.5 TREINAMENTO E TESTE

A divisão de treinamento e teste é um procedimento no qual um conjunto de dados é particionado em duas partes distintas: um conjunto de treinamento usado para treinar o modelo e um conjunto de testes utilizado para avaliar a capacidade do modelo de generalizar para dados não vistos. Essa prática é essencial para avaliar o desempenho do modelo em situações do mundo real, proporcionando uma estimativa mais precisa de sua eficácia. (HASTIE, c. 7.2, 2009)

A proporção comum de 80% para treinamento e 20% para teste é uma prática comum geralmente aceita, embora as proporções possam variar dependendo do tamanho do conjunto de dados e da natureza específica do problema.

A técnica de divisão entre conjuntos de treinamento e teste é importante no aprendizado de máquina para avaliar o desempenho de um modelo. (MULLER, 2016). Essa afirmativa é fundamentada por algumas razões, que são elas:

- **Avaliação do Desempenho Genuíno:**

Dividir o conjunto de dados em conjuntos de treinamento e teste permite avaliar o desempenho do modelo em dados não vistos, ou seja, em instâncias que o modelo não encontrou durante o treinamento. Isso fornece uma medida mais realista do quão bem o modelo generaliza para novos dados.

- **Prevenção de Overfitting:**

Overfitting é um caso no qual um modelo de aprendizado de máquina se ajusta muito bem aos dados de treinamento, mas falha em generalizar adequadamente para novos dados ou dados não vistos. (HASTIE, c. 7, 2009)

A divisão entre conjuntos de treinamento e teste é crucial para identificar e evitar o overfitting. Se todo o conjunto de dados fosse usado para treinar o modelo, ele poderia se ajustar excessivamente aos padrões específicos dos dados de treinamento e não generalizar bem para novos dados. O conjunto de teste atua como um indicador de como o modelo se comporta em situações não observadas.

- **Ajuste de Parâmetros e Seleção de Modelos:**

Ao avaliar o desempenho do modelo no conjunto de testes, os ajustes de parâmetros e a seleção de modelos podem ser realizados com base em seu desempenho em dados não vistos. Isso ajuda a escolher modelos que tenham uma boa capacidade de generalização.

2.6 ERRO MÉDIO QUADRÁTICO (MSE):

Extremamente importante a realização de avaliação de desempenho do algoritmo de regressão. A avaliação deve ser feita com base em métricas relevantes para o problema, como o erro médio quadrático (MSE), o coeficiente de determinação (R^2) e outros indicadores específicos da tarefa.

MSE é uma métrica que mede a média dos quadrados dos erros ou resíduos, entre os valores previstos pelo modelo e os valores reais (observados) do conjunto de dados de teste.

Quanto menor o valor do MSE, melhor o desempenho do modelo. Isso significa que o modelo está produzindo previsões mais próximas dos valores reais.

O MSE é calculado usando a fórmula:

$$\mathbf{MSE = \Sigma(yi - \hat{y}i)^2 / n}$$

yi representa os valores reais.

ŷi representa os valores previstos pelo modelo.

n é o número total de observações.

2.7 COEFICIENTE DE DETERMINAÇÃO (R^2):

O coeficiente de determinação, frequentemente denotado como R^2, é uma métrica que fornece uma medida da variabilidade dos valores dependentes (variável de resposta) que é explicada pelo modelo.

O R^2 varia de 0 a 1, onde:

$R^2 = 0$ significa que o modelo não explica nenhuma variabilidade nos dados e é essencialmente inútil.

$R^2 = 1$ significa que o modelo explica perfeitamente toda a variabilidade nos dados.

O R^2 é calculado usando a fórmula:

$$\mathbf{R^2 = 1 - (SSE / SST)}$$

SSE (*Sum of Squared Errors*) é a soma dos quadrados dos erros residuais do modelo.

SST (*Total Sum of Squares*) é a soma total dos quadrados da diferença entre os valores reais e a média dos valores reais.

2.8 JUPYTER NOTEBOOK

É uma ferramenta computacional que se destaca como um ambiente interativo amplamente utilizado para desenvolvimento de projetos relacionados à análise de dados, programação, visualização e aprendizado de máquina. Essa plataforma proporciona uma integração eficiente entre código, texto explicativo, equações matemáticas e visualizações, consolidando-se como uma peça fundamental no ecossistema da ciência de dados.

O funcionamento do Jupyter Notebook baseia-se na criação de documentos que contêm "células", unidades individuais que podem conter tanto código executável quanto elementos textuais formatados usando a linguagem Markdown. Cada célula pode também incluir equações matemáticas em LaTeX, proporcionando uma ampla flexibilidade na criação de relatórios interativos e documentação de projetos.

O código Python é frequentemente utilizado nas células, permitindo a execução e visualização imediata dos resultados. Essa abordagem facilita a análise exploratória de dados e a experimentação rápida com algoritmos de aprendizado de máquina. Como ressaltado por Kluyver,"o Jupyter Notebook é uma ferramenta poderosa para criar e compartilhar documentos que contêm código, equações, visualizações e texto explicativo" (KLUYVER et al., 2016, p. 87).

Importante destacar que tem uma variedade de aplicações, sendo amplamente empregado em projetos de análise de dados, desenvolvimento de algoritmos de aprendizado de máquina, ensino e colaboração científica. Sua capacidade de integrar código e explicação em um único documento facilita a comunicação efetiva de resultados e descobertas. No contexto da ciência de dados, VanderPlas destaca que "o Jupyter é uma ferramenta cada vez mais popular no campo, fornecendo uma interface flexível para análise interativa e exploração de dados" (VANDERPLAS, 2016, p. 5).

2.9 NUMPY E PANDAS

No panorama da ciência de dados e análise estatística em Python, as bibliotecas numpy e pandas desempenham papéis distintos, mas, complementares, pois contribuem de maneira crucial para a manipulação e tratamento de dados complexos. Essas bibliotecas formam a espinha dorsal do ambiente computacional de Python, enriquecendo o escopo de ferramentas disponíveis para profissionais e pesquisadores em suas explorações analíticas.

O numpy, abreviação de *Numerical Python,* é uma biblioteca que oferece suporte a arrays multidimensionais e funções matemáticas de alto desempenho. Ele constitui a base para muitas outras bibliotecas em Python, permitindo operações eficientes em grandes conjuntos de dados. Segundo VanderPlas, "o numpy fornece as estruturas fundamentais necessárias para operações eficientes em arrays multidimensionais" (VANDERPLAS, 2016, p. 58). Essa eficiência é crucial para cálculos complexos e manipulação de dados em projetos de ciência de dados e aprendizado de máquina.

Para realizar operações numéricas avançadas em arrays multidimensionais, o numpy é essencial. Além do mais, ele serve para facilitar cálculos matemáticos complexos, manipulação eficiente de dados numéricos e operações em larga escala, sendo a base para muitas outras bibliotecas em Python.

A biblioteca pandas é reconhecida por suas estruturas de dados poderosas, como o DataFrame, que facilita a manipulação e análise de tabelas de dados. McKinney destaca que "o Pandas fornece estruturas de dados flexíveis e de alto desempenho, especialmente o DataFrame, que é ideal para a representação e manipulação de dados tabulares" (MCKINNEY, 2018, p. 14). Sua capacidade de lidar com dados heterogêneos e realizar operações complexas simplifica tarefas comuns na análise exploratória de dados.

Pandas oferece estruturas de dados tabulares, como o DataFrame, para manipulação e análise de dados. Além do mais, ele permite a leitura, limpeza, transformação e análise de dados tabulares de forma eficiente, sendo especialmente útil em tarefas de análise exploratória de dados.

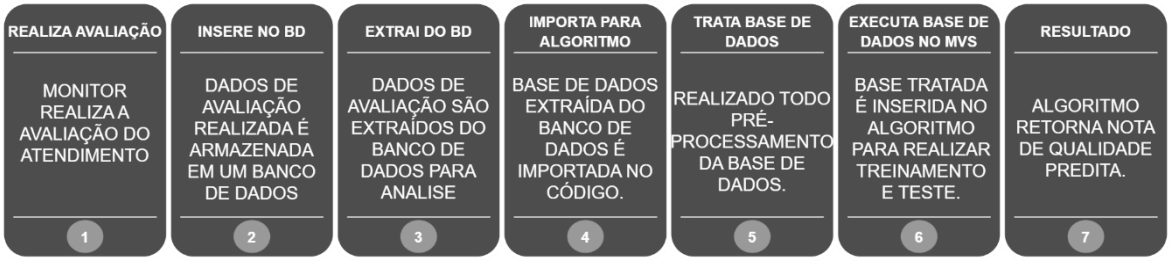
3 METODOLOGIA

Após compreender o que o projeto precisa para ser realizado, as ferramentas, funcionamento e para que são utilizadas, partiremos para como foi conduzido o desenvolvimento do projeto.

É necessário destacar que esse projeto é puramente quantitativo, ou seja, não serão expostas análises ou percepções qualitativas em detrimento do estudo realizado. Conforme descrito anteriormente, o objetivo do projeto é utilizar uma base de dados de avaliações de nota de qualidade, processar e tratar essa base, treinar o modelo preditivo, e obter a nota futura.

O Projeto funciona através do seguinte fluxo:

Figura 1 - FLUXO DE FUNCIONAMENTO



3.1 BASE DE DADOS

Diariamente, o time de monitores é o responsável por escutar e realizar a avaliação dos atendimentos fornecidos pela empresa, onde, de forma imparcial, verificam-se as oportunidades, atribuindo no final a nota de qualidade daquele atendimento. Essas avaliações são armazenadas em uma tabela no banco de dados para aproveitamento e atuação operacional posteriormente. A nota atribuída pode ser de 0 a 100, tendo como meta 80. Qualquer nota abaixo de 80 está gerando impacto negativo para a empresa, seja quantitativamente ou financeiramente. Caso a nota esteja abaixo da meta, é possível atuar nas falhas pontuadas pelo avaliador apresentar novas orientações aos operadores relacionadas a falha pontuada, tirando dúvidas, e corrigindo as formas de atendimento para que outros clientes não sejam impactados por essa experiência negativa com a empresa.

Importante destacar a necessidade destas avaliações para acompanhamento, acolhimento e resolutividade da necessidade do cliente para com a empresa, bem como identificar oportunidades e pontos que precisam ser corrigidos processualmente, na regra do negócio e no atendimento humanizado que a empresa fornece ao cliente final.

As avaliações armazenadas alimentam um relatório de acompanhamento diário das notas de qualidade. Nesse relatório contém a nota quantitativa e as oportunidades que impactam diretamente a nota de qualidade. O algoritmo extrai essa base de dados com 6 meses de histórico dessas notas, em seguida é necessário realizar um tratamento nessa base, pois existem muitas colunas qualitativas que não serão necessárias para o nosso resultado, conforme veremos a seguir. Por questões de segurança da informação, e por conter dados sensíveis, não será possível expor a base completa, mas descreveremos como foi realizado a importação e tratamento dela, para só então, obtermos o resultado esperado.

3.1.1 IMPORTAÇÃO DE BIBLIOTECAS

A importação das bibliotecas Numpy, Pandas e Matplotlib é fundamental para potencializar a capacidade de análise de dados e visualização em projetos desenvolvidos em Python. Cada uma dessas bibliotecas desempenha papéis específicos, oferecendo conjuntos robustos de ferramentas que atendem a diferentes aspectos do ciclo de vida de análise de dados.

3.1.2 IMPORTAÇÃO DA BASE

Para carregar a base de dados no código, foi utilizado o parâmetro **pd.read_csv** do Pandas. É uma ferramenta muito útil e bastante utilizada atualmente, que tem como funcionalidade: ler e carregar dados de arquivos CSV (*Comma-Separated Values*) em um DataFrame, que é uma estrutura de dados tabular bidimensional oferecida pelo Pandas. Essa função simplifica significativamente o processo de importação de conjuntos de dados tabulares

armazenados em arquivos CSV, que são um formato comum para armazenar dados estruturados.

A nossa base de dados foi fornecida pela empresa de transporte por aplicativo na qual estamos realizando o projeto, conta com 27 variáveis e 1.271.970 amostras para análise.

3.1.3 TRATANDO A BASE

Tendo em vista que queremos obter um resultado quantitativo, é necessário que a base de dados seja ajustada apenas com as colunas que serão úteis para o algoritmo verificar recorrência e padrões, para definir sua métrica e predizer uma nota de acordo com os dados históricos obtidos. É importante ressaltar que, caso o objetivo final seja um resultado quantitativo, todas as colunas precisam ter valores atribuídos. Considero que o processo de tratamento de dados é o ponto mais importante de um projeto que envolve predição de dados.

Para que se obtenha um resultado eficaz e desejado, é necessário entender o resultado que deseja obter e estruturar a base de dados de forma que torne possível o algoritmo ler, interpretar e predizer conforme desejado. Após tratar a base, retirar algumas variáveis que não serão necessárias para a nossa análise e utilizar técnicas para converter as variáveis categóricas em um formato numérico para que o MVS possa interpretá-la de maneira apropriada. A base é inserida no algoritmo propriamente dito, para realização do treinamento da base e posteriormente será retornado o valor da nota quantitativa prevista pelo MVS.

Nas avaliações realizadas pelos monitores existem vários itens que possuem diversos pesos. Como essa informação não está explícita na base de dados, e para o MVS compreender e treinar corretamente a base, foi necessário realizar um pivotamento na base e atribuir manualmente todos os pesos de todos os itens que são avaliados pelos monitores diariamente, conforme apresent*ado no exemplo abaixo:*

“df_concatenado.loc[df_concatenado['A imagem contem mais de uma pessoa'] == 'NC', 'A imagem contem mais de uma pessoa'] = 25;”

Na linha de código acima, temos um dataframe definido como **df_concatenado**, e nele contém toda base de dados que está sendo tratada e inclusive as colunas dos itens onde os pesos estão sendo atribuídos.

Através do método **loc** é possível acessar e modificar dados no dataframe com base em uma condição. Ele verifica se os valores na coluna chamada: 'A imagem contém mais de uma pessoa' são iguais a 'NC' e retorna verdadeiro para as linhas que atendem a essa condição e falso para as demais.

['A imagem contém mais de uma pessoa'] = 25: Para as linhas onde a condição seja verdadeira (ou seja, onde 'A imagem contém mais de uma pessoa' é igual a 'NC'), o código atribui o peso 25 a essas células.

3.1.3.1 UTILIZAÇÃO DO ONE-HOT ENCODER

É uma técnica de pré-processamento utilizada em aprendizagem de máquina para lidar com variáveis categóricas. Quando trabalhamos com algoritmos de aprendizado de máquina, muitos deles exigem que os dados de entrada estejam no formato numérico. No entanto, variáveis categóricas, que representam categorias e não valores numéricos, precisam ser convertidas para que o modelo possa interpretá-las corretamente.

O One-Hot Encoder resolve esse problema transformando cada valor único de uma variável categórica em uma nova coluna binária (0 ou 1) e mantendo essas colunas para representar a presença ou ausência da categoria original. A tabela 1 exemplifica com mais clareza na compreensão dessa técnica:

Tabela 1 - USO DO ONE-HOT ENCODER

ANTES

CANAL
VOZ
CHAT
TICKET

DEPOIS

VOZ	CHAT	TICKET
1.0	0.0	0.0
0.0	1.0	0.0
0.0	0.0	1.0

3.1.4 REALIZANDO TREINAMENTO E TESTE

Após realizar a importação das bibliotecas, importação e tratamento da base para o objetivo final, chamamos no código o algoritmo responsável pela predição da nota de qualidade, onde se realiza o treinamento dos dados e o teste.

Foi utilizado o modelo Máquina de Vetores de Suporte, para investigar e prever avaliações de qualidade. Esse modelo permite que um sistema aprenda a partir dos dados históricos sem ser explicitamente programado para realizar tarefas específicas.

Existem muitos outros algoritmos de regressão, como regressão linear, regressão polinomial, árvores de decisão, redes neurais, regressão Ridge, regressão Lasso, entre outros. (HASTIE, T., 2009.)

Nesse caso, realizamos testes com outros algoritmos de regressão já citados anteriormente, porém, ao realizarmos as avaliações de desempenho baseadas no coeficiente de determinação, dos que realizamos o teste, fizemos o uso do que obteve o melhor resultado. O algoritmo de árvore de decisão teve uma excelente performance, porém, como observamos no comparativo da tabela 2, o MVS obteve uma melhor performance para essa base de dados.

Por este motivo, o MVS foi escolhido para utilização neste projeto.

Tabela 2 - COMPARAÇÃO DE DESEMPENHO DOS ALGORITMOS

MODELO	R^2
MVS	0,9759
ARVORE DE DECISÃO	0,9568
REGRESSÃO LINEAR	0,40004

3.1.5 TESTE DE PREDIÇÃO

Após a realização de todo tratamento e pré-processamento da base, será realizado o teste de predição dos dados com o modelo definido, onde será validado se o resultado está conforme o esperado. Conforme descreve a figura 2 abaixo:

Figura 2 - REALIZANDO TESTE DE PREDIÇÃO

```
# Criar um array booleano para filtrar as previsões com base na coluna
filtro = df_encoded['DRIVER | Written Channels | v.1 (11/22)'] == 1.0

# Calcular a média das previsões filtradas
media_previsoes = np.mean(y_pred[filtro])

print(media_previsoes)
```

```
93.72837719540054
```

3.1.6 MÉTRICAS DE VALIDAÇÃO

Conforme apresentado nas seções 2.6 e 2.7, neste trabalho foram utilizadas duas métricas de validação do modelo: MSE e R^2. O MSE mede o quão próximo as previsões do modelo estão dos valores reais, enquanto o R^2 fornece uma medida da qualidade geral do modelo, indicando quanto da variabilidade dos dados é explicada pelo modelo. Ambas as métricas são importantes na avaliação de modelos de regressão, e a escolha de uma métrica específica depende dos objetivos e das características do problema que está sendo abordado.

Após a realização da predição da nota de qualidade, calculamos estes coeficientes para validar o quão bem o modelo se ajusta nesta base de dados, e observamos que o coeficiente de determinação desse modelo é de 98%, conforme apresentamos na figura 3. Isso indica que o modelo explica aproximadamente 98% da variabilidade nos dados, sugerindo um ajuste muito bom do modelo aos dados observados.

Figura 3 - MÉTRICAS DE VALIDAÇÃO DO MODELO

```
from sklearn.metrics import mean_squared_error, r2_score

# Calculando o erro médio quadrático (MSE)
mse = mean_squared_error(y, y_pred)

# Calculando o coeficiente de determinação (R²)
r2 = r2_score(y, y_pred)

print("Erro médio quadrático (MSE):", mse)
print("Coeficiente de determinação (R²):", r2)
```

```
Erro médio quadrático (MSE): 25.619942647545887
Coeficiente de determinação (R²): 0.9759865605515399
```

4 RESULTADOS

Após entendermos o objetivo do projeto, e acompanharmos como foi desenvolvido, agora vamos verificar os resultados. Conforme falado anteriormente, o projeto é aplicado em uma empresa de transporte por aplicativo. Dessa forma, o algoritmo está em atuação piloto na empresa, onde os dados são atualizados e retroalimentados diariamente e semanalmente para que o algoritmo possa atualizar as predições de acordo com as bases de dados atualizadas.

A tabela 3 apresenta o comparativo da nota predita pelo algoritmo com a nota real praticada por cada ficha de avaliação e o percentual de assertividade. Realizam Com isso, conseguimos verificar que as predições geradas pelo algoritmo são confiáveis, e o setor operacional da empresa poderá atuar preventivamente com mais eficácia de acordo com o que está sendo gerado pelo MVS.

Tabela 3 - COMPARATIVO NOTA PREDITA x NOTA REAL

PLANILHA	NOTA PREDITA	NOTA REAL	%
PAYMENTS | PAY | PHONE | v.2 (02/23)	74,5	75,0	99%
PAYMENTS | PAY | Written Channels | v.1 (11/22)	78,6	77,3	102%
PAYMENTS | CONTA | PHONE | v.1 (11/22)	78,3	82,7	95%
PAYMENTS | CONTA | Written Channels | v.1 (11/22	82,6	83,5	99%
DRIVER | PHONE | v.1 (11/22)	83,2	86,3	96%
PAYMENTS | PAY | LIVE CHAT | V.1 (11/22)	86,2	87,0	99%
SAFETY | T2 | 4.0 | VOZ | v.2 (08/22)	92,3	91,9	100%
RIDER | Written Channels | v.1 (11/22)	91,9	92,1	100%
RIDER | LIVE CHAT | v.1 (11/22)	92,1	92,6	99%
DRIVER | LIVE CHAT | v.1 (11/22)	91,5	92,8	99%
DRIVER | Written Channels | v.1 (11/22)	93,7	94,7	99%
ANTI FRAUDE | TICKET | v.1 (05/21)	99,6	100,0	100%

Nesse contexto, realizando o acompanhamento diário na empresa, consideramos que o resultado do projeto é satisfatório pela maneira como foi conduzido e com as ferramentas que foram escolhidas para atuação.

Conforme mostrado anteriormente, nas avaliações realizadas o coeficiente de determinação se aproxima de 1, conseguindo explicar com quase 100% de eficácia a variabilidade dos dados.

Este é um case de sucesso dentro da empresa, tendo em vista que nunca havia utilizado algoritmo de aprendizagem de máquina para auxiliar na análise dos dados. Com este projeto está sendo possível além de solucionar uma das maiores necessidades da empresa em questão de evolução da nota de qualidade, também apresentar a aprendizagem de máquina e os seus benefícios na análise de dados da empresa.

Após concluirmos esse projeto piloto, acreditamos que outras áreas da empresa disponibilizarão suas bases de dados para realizarmos análise preditiva. Portanto, pretendemos aprimorar o algoritmo para abranger a atuação dele em outras possíveis áreas. Dessa forma, conseguiremos auxiliar positivamente em outras vertentes com algoritmos de predição.

5 CONCLUSÃO

Primordialmente, tínhamos uma problemática na empresa de transporte: conseguir atuar previamente em relação às notas de qualidade. Diante deste cenário, foi proposto o projeto supracitado, que fazendo o uso de algoritmo de predição, considerando a base de dados histórica, conseguimos predizer a nota de qualidade com 98% de precisão, tornando possível a atuação preventiva em cima da nota de qualidade.

Portanto, com este projeto foram alcançados os objetivos iniciais, de forma que a atuação passou a ser preventiva evitando notas abaixo da meta e obtendo ganho financeiro. Pois com a redução de notas negativas, a nota de qualidade consequentemente sobe, e a empresa tem ganho financeiro com o projeto realizado.

A nossa maior limitação neste trabalho foi com relação às políticas da empresa, no quesito de compartilhamento de dados e uso de inteligência artificial para a análise de dados.

Como é algo inovador dentro da empresa, foi necessário definir limites de uso e compartilhamento de informações junto com a equipe jurídica, inclusive para a documentação deste trabalho. Deixo como sugestão para próximos trabalhos, a utilização de modelos de predição em outros indicadores dentro deste setor, de forma que consigam encontrar a melhor performance operacional com o mix de indicadores.

REFERÊNCIAS

CARMELO, J. A**. Support Vector Regression utilizando Algoritmos de Otimização Mono-objetivo e Multiobjetivo,** 2017.

CORTES, C. Vapnik-Chervonenkis **Bounds for Semi-infinite Loss Functions. Neural Computation**, v. 7, n. 2, p. 407-417, 1995.

CRISTIANINI, N.; SHAWE-TAYLOR, J. **An Introduction to Support Vector Machines and Other Kernel-based Learning Methods**. Cambridge University Press, 2000.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. Springer, 2009.

KLUYVER, T. et al. **Jupyter Notebooks – a publishing format for reproducible computational workflows.** In: FIFTH INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE. Procedia Computer Science, v. 87, p. 87-90, 2016.

MITCHELL, T. M. **Machine Learning**. McGraw Hill, 1997.

MITCHELL, T. **Machine learning: Trends, perspectives, and prospects.** Science, 2015.

MCCARTHY, J. **What is Artificial Intelligence?** Stanford University. 2007. Disponível em: <http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html\>. Acesso em: 10 out. 2023.

MCCARTHY, John; MINSKY, Marvin; ROCHESTER, Nathan; SHANNON, Claude. **A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence**. 1955. Dartmouth College, Hanover, NH, Estados Unidos.

MCKINNEY, W. **Data Structures for Statistical Computing in Python.** In: VAN DER WALT, S.; MILLMAN, J. (Eds.). Proceedings of the 9th Python in Science Conference, 2010, p. 51-56.

MÜLLER, Andreas C.; GUIDO, Sarah. **Introduction to Machine Learning with Python: A Guide for Data Scientists**. 2016.

RUSSEL, S.; NORVIG, **P. Artificial Intelligence: A Modern Approach**. 3rd ed. Prentice Hall, 2010.

REICHHELD, F. F. **The One Number You Need to Grow.** Harvard Business Review, 81(12), p. 46-54, 2003.

SÁNCHEZ A, V David. **Advanced support vector machines and kernel methods**. 2003.

VANDERPLAS, J. T. **The Python Data Science Handbook: Essential Tools for Working with Data**. O'Reilly Media, 2016.

VAPNIK, V. N. **The Nature of Statistical Learning Theory**. Springer, 2000.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
Campus Campina Grande
R. Tranqüilino Coelho Lemos, 671, Dinamérica, CEP 58432-300, Campina Grande (PB)
CNPJ: 10.783.898/0003-37 - Telefone: (83) 2102.6200

Documento Digitalizado Ostensivo (Público)

MODELO PREDITIVO PARA INDICADOR DE QUALIDADE NO SETOR DE TRANSPORTE POR APLICATIVO

Assunto:	MODELO PREDITIVO PARA INDICADOR DE QUALIDADE NO SETOR DE TRANSPORTE POR APLICATIVO
Assinado por:	Aquila Samuel
Tipo do Documento:	Dissertação
Situação:	Finalizado
Nível de Acesso:	Ostensivo (Público)
Tipo do Conferência:	Cópia Simples

Documento assinado eletronicamente por:

- **Áquila Samuel Azevedo Dias, ALUNO (201711250001) DE BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO - CAMPINA GRANDE**, em 31/01/2024 11:36:37.

Este documento foi armazenado no SUAP em 31/01/2024. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 1067640
Código de Autenticação: 2746fe157b