1. Escolha do Tema e Dataset

temas:

O IMPACTO DA VARIAÇÃO DE PREÇO DOS COMBUSTÍVEIS NA SOCIEDADE.

Nome: daniel oliveira dos santos

Mestrando na UFC-Ceará

Link do dataset principal

<https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-abertos/serie-historica-de-precos-de-combustiveis>

Os dados foram disponibilizados no site do governo.

2. Estrutura do Artigo

## Contextualização do Trabalho

Introdução

Este trabalho tem como principal objetivo de ver relações de preços de combustível do estado do ceara entre município por bandeira, municípios, postos, bairro. Foi pega uma amostra do primeiro semestre de 2022 as colunas foram agrupadas e foi criada uma coluna de classificação que representa, melhor preço, preço, mais barato e o caro. Esta coluna vai ser usada para adicionar em uma nova coluna como complemento de classificação, a coluna produto vai ser usado como objetivo de classificar o combustível como gasolina, gasolina aditivada, etanol, diesel entre outros, para treinamento de inteligência artificial usando machine learning serão usados dois modelos de classificação dos quais se encontra Regressão logística e arvore de descrição, estes modelos serão treinados e suas métricas vão corresponder o melhor resultados, este artigo vai iniciar explicando sobre gasolina, gasolina aditivada e o etanol e cada modelo e o que cada corresponde entre melhor resultado, na sociedade saber qual o posto de gasolina mais próxima do município que é mais barata pode fazer diferença de quem usa diariamente qualquer veículo, este artigo é o inicio para ajudar e desenvolver métodos de classificação de melhores preços e mais acessivo, mais para isso saber diferenciar o combustível é a iniciativa para esse pequeno passo.

## Escolha e sua relevância no campo da inteligência artificial e lógica.

Com uso dos dados coletados poderemos demonstrar e mensurar as variações de preços sobre os custos dos combustíveis para os consumidores. Com isso conseguiremos responder alguns questionamentos: Quanto a variação diária, semanal e mensal da gasolina, gasolina aditivada e etanol. Com isso poderemos construir dashboards, gráficos e relatórios usando machine learning demonstrando os dados e apontando qual as melhores escolhas adotadas pelo consumidor.

## Identificação do Sistema e Componentes em Teste

Descreva os algoritmos, técnicas e ferramentas que serão utilizados no seu projeto.

## Objetivo da Avaliação de Desempenho

O objetivo é avaliar e comparar o desempenho dos dois modelos de classificação na tarefa de prever o combustível de cada posto de combustível com base nas características fornecidas. Isso será feito em coletar os dados sobre os preços da gasolina em diferentes postos de combustível em cada município do Ceará. as informações foram coletadas no site do governo dados aberto para uso de transparência, sobe licenciamento do site. Após a coleta será feito tratamento dos dados, limpeza, organização e transformação dos dados para o treino, com separação em conjuntos de treinamentos e testes. O conjunto de treinamento será usado para treinar o modelo de aprendizado de máquina, enquanto o conjunto de teste será usado para avaliar a precisão do modelo. Após o treinamento, avaliação do desempenho se dará cada modelo utilizando o conjunto de teste. Calculando métricas relevantes, acurácia, precisão, matriz de confusão especificidade recall, f1\_score, acurácia entre outros para entender a precisão do modelo na previsão dos melhores preços de gasolina. As etapas serão:

• Identificar como funciona o preço da gasolina e a variação do preço dos combustíveis em cada município.

• Identificar se a variação de preço ocorre também em postos da mesma marca ou bandeira.

• Mensurar os municípios para abastecer e o melhor posto.

## Métricas Utilizadas

Escolha métricas relevantes para o seu problema. Exemplo: precisão, recall, F1-score, AUC-ROC para algoritmos de classificação.

Estrutura da Matriz de Confusão para Problemas Binários

Para um problema de classificação binária, a matriz de confusão tem a seguinte forma:

|  |  |  |
| --- | --- | --- |
|  | Classe Prevista Positiva | Classe Prevista Negativa |
| Classe Real Positiva | True Positive (TP) | False Negative (FN) |
| Classe Real Negativa | False Positive (FP) | True Negative (TN) |
|  |  |  |

Código látex

A acurácia é a proporção de previsões corretas entre o total de casos avaliados.

\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN}

A matriz de confusão é uma tabela que permite a visualização do desempenho de um algoritmo de classificação. Ela compara as previsões do modelo com os valores reais.

\begin{array}{|c|c|c|}

\hline

& \text{Previsto Positivo} & \text{Previsto Negativo} \\

\hline

\text{Real Positivo} & TP & FN \\

\hline

\text{Real Negativo} & FP & TN \\

\hline

\end{array}

Proporção de verdadeiros positivos entre todos os exemplos classificados como positivos.

\text{Precisão} = \frac{TP}{TP + FP}

Proporção de verdadeiros positivos entre todos os exemplos realmente positivos.

\text{Recall ou sensibilidade} = \frac{TP}{TP + FN}

Média harmônica entre precisão e recall.

\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}

\text{Taxa de Falsos Positivos} = \frac{FP}{FP + TN}

\begin{table}[ht]

\centering

\begin{tabular}{c|ccc}

& \text{Classe A Prevista} & \text{Classe B Prevista} & \text{Classe C Prevista} \\

\hline

\text{Classe A Real} & C\_{AA} & C\_{AB} & C\_{AC} \\

\text{Classe B Real} & C\_{BA} & C\_{BB} & C\_{BC} \\

\text{Classe C Real} & C\_{CA} & C\_{CB} & C\_{CC} \\

\end{tabular}

\caption{Matriz de Confusão Multiclasse}

\end{table}

### Interpretação

* CiiC\_{ii}Cii​ (diagonal principal) representa o número de instâncias corretamente classificadas para a classe iii.
* CijC\_{ij}Cij​ (fora da diagonal principal) representa o número de instâncias da classe iii que foram erroneamente classificadas como classe jjj.

## Detalhes do(s) Workload(s) Utilizado(s)

* **Dataset:** Dados de postos de combustíveis do estado do Ceará.
* **Entradas:** Municípios, bairros, valor de venda e revenda.
* **Saída:** Bandeira do posto (ALESAT, SP, IPIRANGA, VIBRA ENERGIA, RAIZEN, BRANCA, FAN).

O dataset foi feito uma limpeza de filtro para o estado do ceara pois se encontra em todos estados do Brasil, as colunas que serão usadas como entradas serão municípios e bairro do estado do ceara, valor de venda e revenda que representa o posto de combustíveis, status que é uma coluna que representa uma classificação de barato, caro e melhor preço, ela foi criada para organizar mais os dados, e a coluna bandeira que são 6 tipos de bandeiras ALESAT, SP, IPIRANGA, VIBRA ENERGIA, RAIZEN, BRANCA, FAN. E a coluna que será usada como classificação será produto que tem 'GASOLINA', 'ETANOL', 'GASOLINA ADITIVADA', 'DIESEL S10', 'DIESEL', 'GNV'.

A interpretação do status ela se dá por município, vamos supor que tenho um valor para o município de ITAPIPOCA seja R$ 7.12 considerado barato e MARACANAU um valor de R$ 6.77 caro, percebe que há uma diferença de preço que Maracanaú seja barato e Itapipoca seja caro porém o método de classificação se da por aquele município, entre o município de Maracanaú aquele valor é o mais caro, é só imaginar que a tabela seja agrupada por município bairro e revenda.

## Parâmetros/Fatores/Níveis do Sistema e do Workload

Foram escolhidos dois modelos de classificações de multiclasses dos quais se encontra arvore de decisão regressão, logística esses modelos são mais utilizados e atende o requisito para classificações são os mais perto para o objetivo que se encontra no dataset.

#### Parâmetros

* **Entradas:** Município, Revenda, Bairro, Valor Venda, Bandeira, status.
* **Saída:** Produtos que tem os Tipos de combustível.

#### Fatores

* **Modelos de Classificação:** Regressão Logística, Árvore de Decisão.
* **Métricas de Avaliação:** Acurácia, Precisão, Recall, Especificidade F1\_score Matriz de confusão.
* **Regressão Logística:** Configurações padrão e ajustes de hiper parâmetros como regularização. max\_iter quantidade de interações.
* **Árvore de Decisão:**
* min\_samples\_split: O número mínimo de amostras necessárias para dividir um nó interno.
* min\_samples\_leaf: O número mínimo de amostras necessárias para ser um nó folha.
* max\_features: O número máximo de features a serem consideradas para fazer a divisão em cada nó.
* criterion: A função para medir a qualidade da divisão.
* max\_depth: A profundidade máxima da árvore.

Justificativa das escolhas (de ferramentas, métricas, fatores/níveis, workload)

**Ferramentas:**

Os dois modelos foram escolhidos pela sua popularidade e eficácia em problemas de classificação.

**Métricas:**

As métricas selecionadas fornecem uma visão completa do desempenho do modelo, considerando tanto a precisão quanto a capacidade de capturar todos os casos positivos.

**Fatores/Níveis:**

Diferentes configurações dos modelos ajudam a identificar a combinação ideal para melhor desempenho.

## Apresentação dos Resultados

Os resultados serão apresentados em tabelas e figuras que mostram as métricas de desempenho para cada modelo. Comparações visuais (como gráficos de barras) serão usadas para facilitar a interpretação.

Testes Estatísticos e Discussão dos Resultados

Realize testes estatísticos para validar os resultados. Exemplo: teste t, ANOVA, testes de hipótese não paramétricos.

Discuta os resultados, interpretando os dados e destacando achados significativos.

## Ferramentas e Análise

Ferramentas de Desenvolvimento: Python, Jupyter Notebooks, bibliotecas como scikit-learn, TensorFlow, PyTorch, Numpy, Pandas.

Ferramentas de Visualização: Matplotlib, Seaborn, Plotly.

Análise Estatística: Scipy, Statsmodels.

\begin{resumo}

Este trabalho tem como principal objetivo investigar as relações de preços de combustíveis no estado do Ceará entre municípios por bandeira, municípios e postos. Foi utilizada uma amostra do primeiro semestre de 2022, onde as colunas foram agrupadas e foi criada uma coluna de classificação que representa melhor preço, preço mais barato e o caro. Esta coluna será usada como target para treinamento de inteligência artificial usando machine learning. Serão utilizados três modelos de classificação: Regressão Logística, Árvore de Decisão e Máquina de Vetores de Suporte. Este artigo inicia explicando sobre gasolina, gasolina aditivada e etanol, e apresenta os resultados dos modelos, identificando aquele que teve melhor desempenho. A pesquisa busca auxiliar a sociedade na identificação dos postos de gasolina mais econômicos, contribuindo para escolhas mais informadas.

\end{resumo}