O Impacto da Variação de Preço dos Combustíveis na Sociedade

Daniel Oliveira dos Santos[1]

[1]Universidade Federal do Ceará (UFC) – Ceará, Brazil

daniel.oliveira@ufc.br

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 para classificar o tipo de combustível como etanol, gasolina, gasolina aditivada, diesel, e gás natural. Esta coluna será usada como target para treinamento de inteligência artificial usando machine learning. Serão utilizados dois modelos de classificação para avaliação de desempenho, Regressão Logística e Árvore de Decisão. 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.*

1. 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.

2. Escolha e 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.

3. Identificação do Sistema e Componentes em Teste

Descreveremos os algoritmos, técnicas e ferramentas que serão utilizados no projeto:

- **Algoritmos:** Regressão Logística, Árvore de Decisão.
- **Ferramentas:** Python, Jupyter Notebooks, scikit-learn, Numpy, Pandas.
- **Ferramentas de Visualização:** plotly express, plotly graph-objects.
- **Análise Estatística:** Scipy.

4. 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. Os resultados foram obtidos em graficos primeiro foi agrupados os dados por municipios e calculados como soma, média e mediana dos combustivel como mostra a (Figure 1).

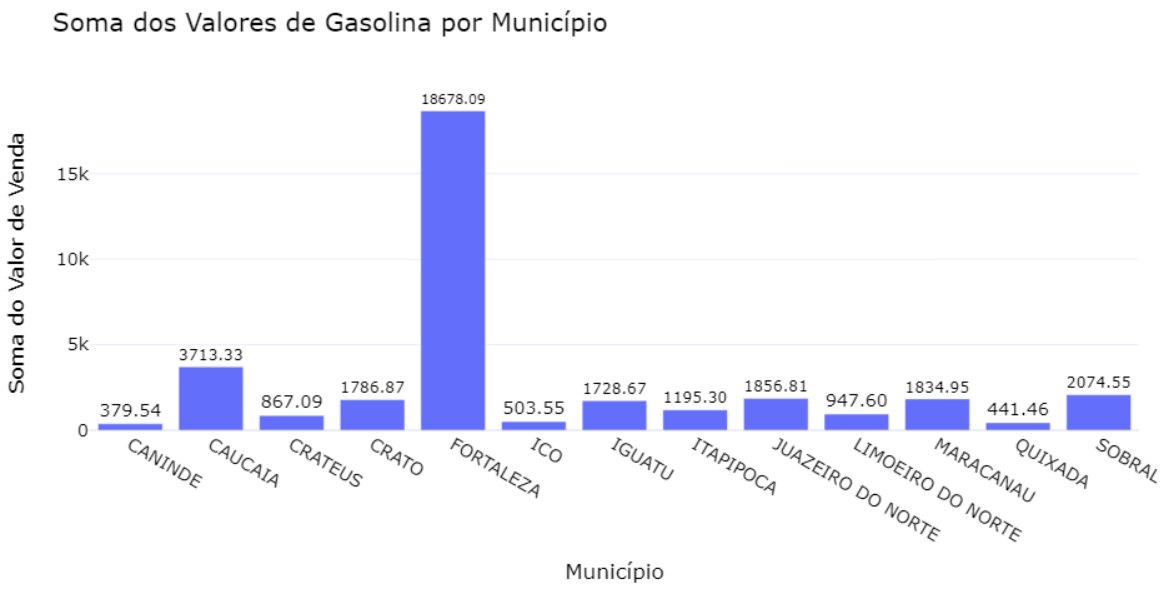

Figura 1. soma total de gasolina por municípios

- Mensurar os municípios para abastecer no melhor posto, com isso foi feito uma analise estastitica primeiro foi calculado a media dos postos por municipios em seguinda investigado a variação dessas média como mostra a (Figure 2).

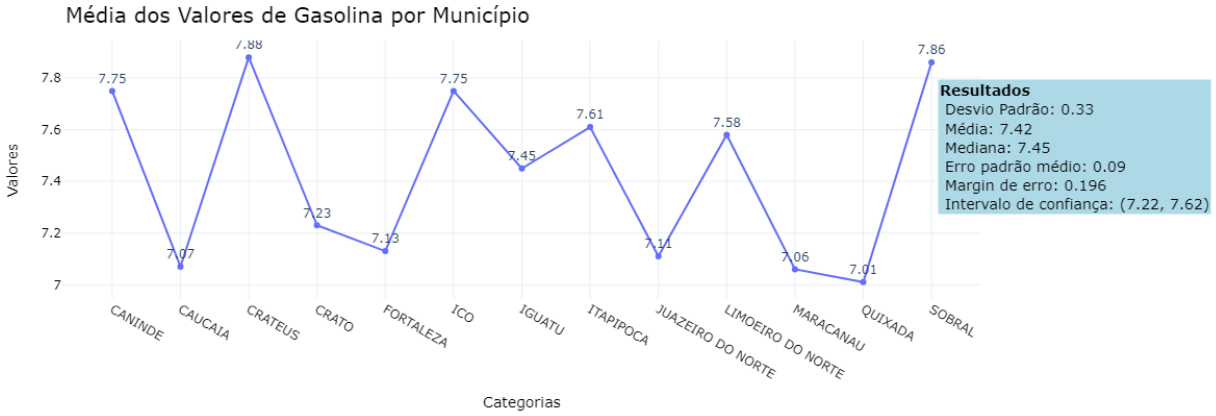

Figura 2. media total de gasolina por municípios

- Identificar se a variação de preço ocorre também em postos da mesma marca ou bandeira, foi realizado a mediana como mostra o (Figure 3).

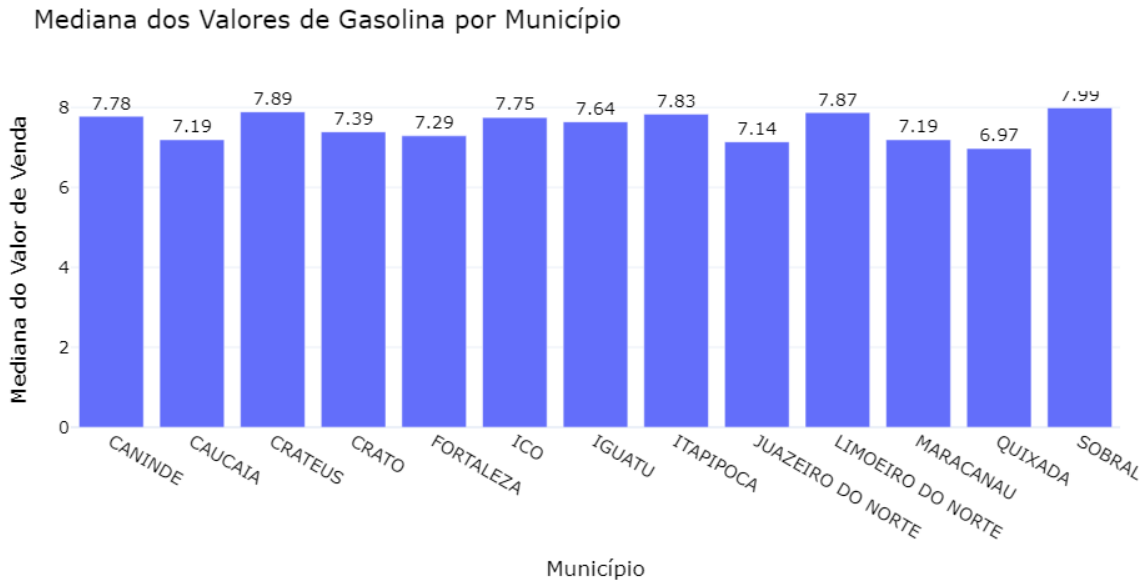

Figura 3. mediana total de gasolina por municípios

5. Métricas Utilizadas

Escolhemos as seguintes métricas relevantes para nosso problema:

- **Acurácia:** Proporção de previsões corretas entre o total de casos avaliados.

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- **Precisão (Precision):** Proporção de verdadeiros positivos entre todos os exemplos classificados como positivos.

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{2}$$

- **Recall (Sensibilidade ou True Positive Rate):** Proporção de verdadeiros positivos entre todos os exemplos realmente positivos.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

- **F1-Score:** Média harmônica entre precisão e recall.

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{4}$$

- **Taxa de Falsos Positivos (False Positive Rate):**

$$\text{Taxa de Falsos Positivos} = \frac{FP}{FP + TN} \tag{5}$$

6. Matriz de Confusão para Problemas Multiclasse

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

	Classe A Prevista	Classe B Prevista	Classe C Prevista
Classe A Real	C_{AA}	C_{AB}	C_{AC}
Classe B Real	C_{BA}	C_{BB}	C_{BC}
Classe C Real	C_{CA}	C_{CB}	C_{CC}

Tabela 1. Matriz de Confusão Multiclasse

Interpretação:

- C_{ii} (diagonal principal) representa o número de instâncias corretamente classificadas para a classe i.
- C_{ij} (fora da diagonal principal) representa o número de instâncias da classe i que foram erroneamente classificadas como classe j.

7. 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'.

Foi feito uma analise analitica sobre os dados, depois dos dados serem tratados, foram agrupados por municipios para saber a soma total, como motra a .

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.

8. 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.

8.1. Parâmetros

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

8.2. Fatores

- **Modelos de Classificação:** Regressão Logística, Árvore de Decisão.
- **Métricas de Avaliação:** : Acurácia, Precisão, Recall, Especificidade F1-score Matriz de confusão.

8.3. Níveis do Sistema

- **Regressão Logística:**
  Configurações padrão e ajustes de hiper-parâmetros como regularização. max-iter quantidade de interações. E os resultados obtido foram como mostra a (Figure 4).

	Acuracia	Precisao	Recall	Especificidade	F1_score	Qtd_real	Qtd_previsto
DIESEL	0.97	0.50	0.35	0.99	0.41	181	126
DIESEL S10	0.76	0.52	0.44	0.87	0.48	1623	1360
ETANOL	0.87	0.70	0.77	0.89	0.73	1575	1744
GASOLINA	0.76	0.51	0.57	0.82	0.54	1632	1796
GASOLINA ADITIVADA	0.79	0.51	0.51	0.86	0.51	1449	1435
GNV	1.00	0.90	0.89	1.00	0.89	129	128

Figura 4. tabela resultado metricas regressão logistica

- **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. E os resultados obtidos do modelo árvore de decisão foram como mostra a (Figure 5).

	Acuracia	Precisao	Recall	Especificidade	F1_score	Qtd_real	Qtd_previsto
DIESEL	0.97	0.00	0.00	1.00	0.00	181	0
DIESEL S10	0.77	0.64	0.12	0.98	0.20	1623	301
ETANOL	0.75	0.49	0.94	0.69	0.64	1575	3042
GASOLINA	0.64	0.39	0.75	0.61	0.51	1632	3163
GASOLINA ADITIVADA	0.79	1.00	0.02	1.00	0.04	1449	35
GNV	0.99	1.00	0.37	1.00	0.54	129	48

Figura 5. tabela resultado metricas árvore de decisão

9. Justificativa das Escolhas

9.1. Ferramentas

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

9.2. 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.

9.3. Fatores/Níveis

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

10. Apresentação dos Resultados

Acurácia da Árvode de decisão:

Foi representada como a média da acurácia de cada classe como mostra a (Figure 6) e como acurácias de cada classe do modelo Árvore de decisão como mostra a (Figure 7).

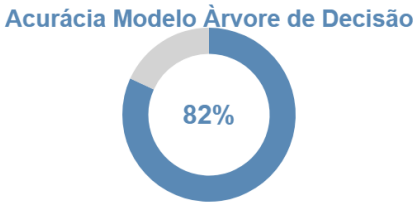

Figura 6. média acurácia do modelo árvore de decisão

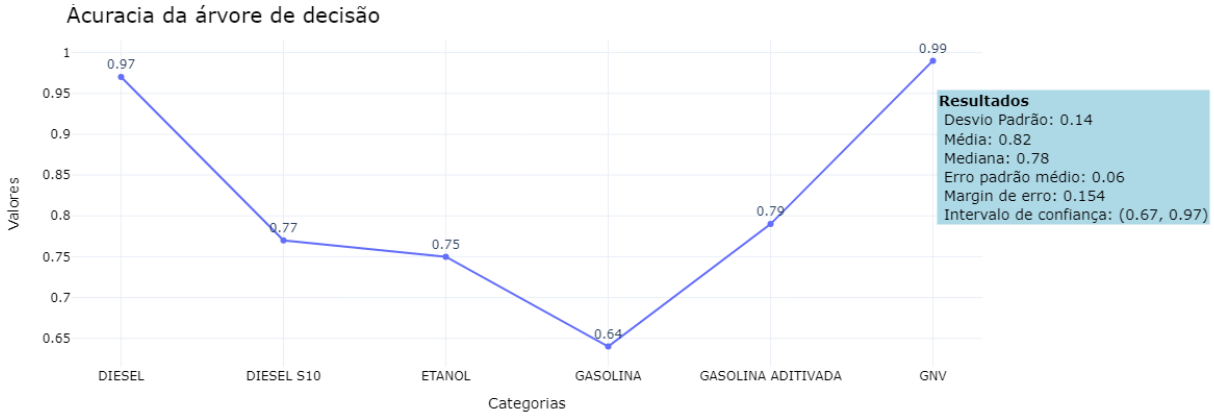

Figura 7. acurácia do modelo árvore de decisão

Acurácia da Regressão Logistica:

Foi representada como a média da acurácia de cada classe como mostra a (Figure 8) e como acurácias de cada classe do modelo regressão logistica como mostra a (Figure 9).

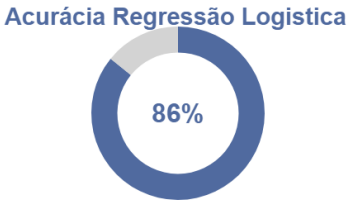

Figura 8. média acurácia do modelo árvore de decisão

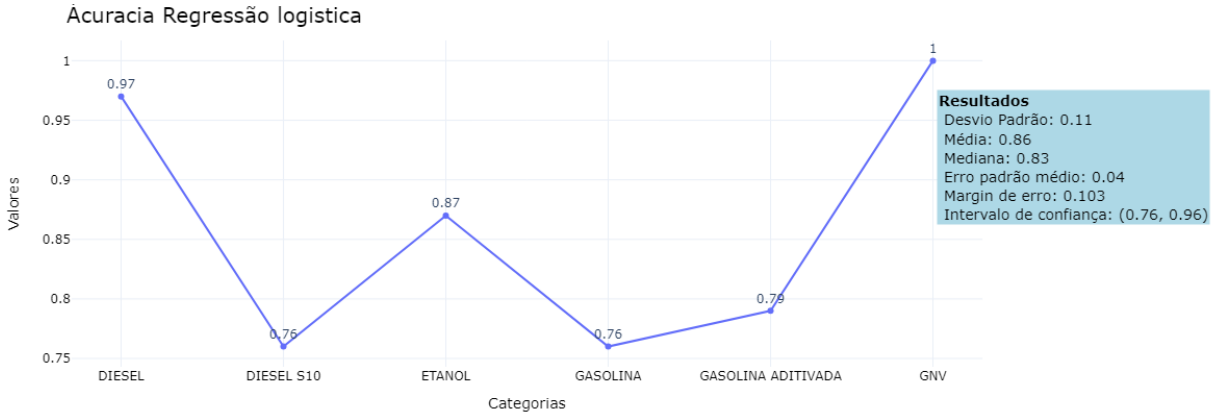

Figura 9. acurácia do modelo árvore de decisão

Matriz de confusão foi escolhido umas das classes que teve o maior desempenho na acurácia como mostra a (Figure 10) e (Figure 11).

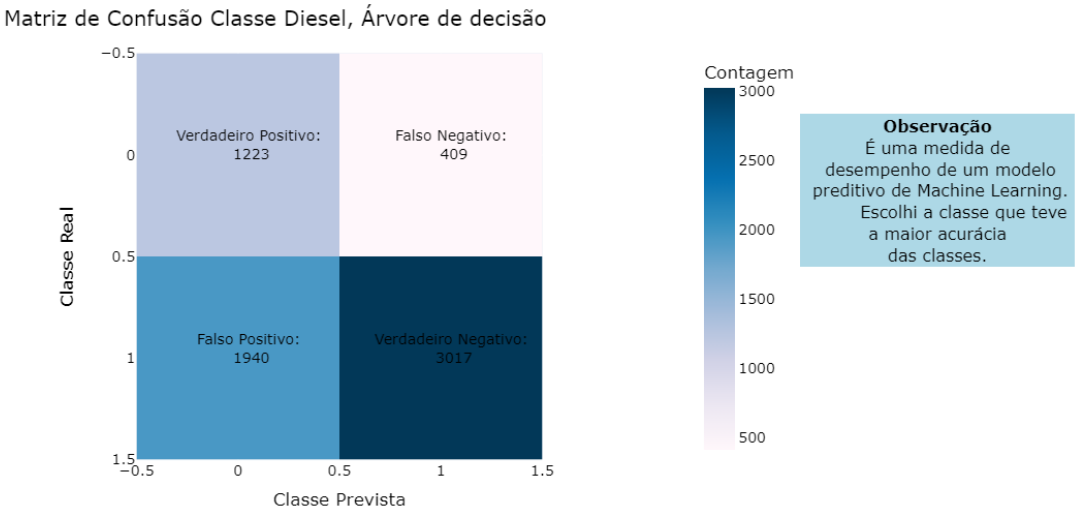

Figura 10. Matriz de confusão da arvore de decisão

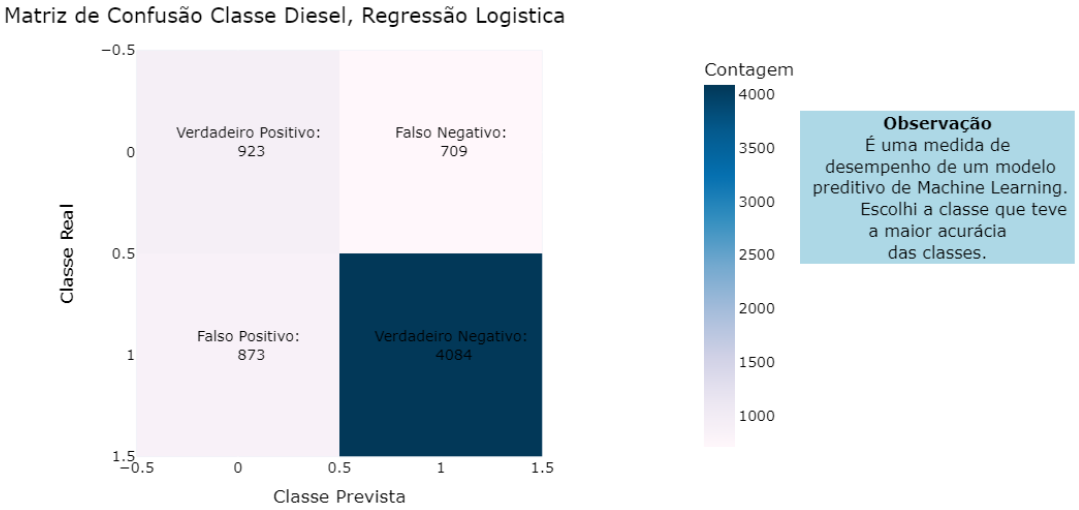

Figura 11. Matriz de confusão da regressão logistica

11. Recursos Adicionais

Para acessar os recursos adicionais utilizados e analisados neste estudo, utilize os links abaixo:

- **Arquivo IPYNB com a análise dos dados:** IPYNB com a análise dos dados
- **CSV com os dados para avaliação:** CSV com os dados para avaliação
- **Pasta com informações adicionais:** Pasta com informações adicionais
- **Link do artigo:** Link do artigo