

**Tech Challenge II**

**IA para Devs**

**Daniela Cruz de Malta - RM353365**

**Gabriela Maciel Godoi - RM355125**

**Lucas Sutelo - RM353721**

**Sumário**

[**Caso de estudo 3**](#_heading=h.x7t0bmscopan)

1. **Introdução**

**1.** Descrição do problema  [**3**](#_heading=h.gjdgxs)

**2. Metodologia**

**1.** Inicialização dos Dados

**2.** Funções Auxiliares

**3.** Algoritmo Genético

4. Visualização da Rota [**9**](#_heading=h.ja3r6f16fnz5)

**3. Metodologia**

**3. Implementação**

**4.Resultado** [**12**](#_heading=h.cog3d54gbwe0)

**5.** [**Conclusão**](#_heading=h.c1br9swxhfw)

**6. Referências** [**13**](#_heading=h.c1br9swxhfw)

[**Anexo 14**](#_heading=h.3qzsze2tbhmh)

**Caso de estudo**

**Objetivo:**

Encontrar a rota de menor distância que permita a um técnico visitar todos os condomínios uma vez e retornar ao ponto de partida, utilizando um algoritmo genético para otimização da rota.

**Introdução**

1. Descrição do problema

O grupo decidiu utilizar um problema de rotas de uma empresa de elevadores, onde o técnico é responsável por visitar uma quantidade de 15 condomínios. O objetivo é encontrar o melhor resultado, ou seja, a rota com menor distância total que permita ao técnico retornar ao seu ponto inicial após visitar todos os locais.

**Metodologia**

1. Inicialização dos Dados:

* Definimos um conjunto de coordenadas de latitude e longitude para os edifícios (condomínios).
* Caso o número de edifícios fornecidos fosse menor que o necessário, coordenadas aleatórias adicionais seriam geradas.

2. Funções Auxiliares:

* haversine: Calcula a distância entre dois pontos geográficos usando a fórmula Haversine.
* calcular\_aptidao: Calcula a aptidão de uma rota, que é a distância total percorrida.

3. Algoritmo Genético:

* **Geração da População Inicial:** Utilizando gerar\_populacao para criar uma população inicial de rotas aleatórias.
* **Avaliação da Aptidão:** Avaliando a aptidão de cada indivíduo na população.
* **Seleção:** Ordenando a população com base na aptidão.
* **Elitismo:** Mantendo a melhor solução de cada geração.
* **Crossover e Mutação:** Aplicando crossover e mutação para criar novos indivíduos.
* **Prevenção de Divisão por Zero:** Adicionando uma pequena constante para evitar divisão por zero nas probabilidades de seleção.
* **Normalização das Probabilidades:** Normalizando as probabilidades de seleção para garantir uma distribuição correta.

4. Visualização da Rota:

* Criando um mapa interativo usando Folium para visualizar a melhor rota encontrada.
* Salvando o mapa em um arquivo HTML.

ser *outliers*. Além disso foram utilizados alguns gráficos de histograma que mostram como está a distribuição dos dados

A variável idade (ou Age) não possui *outliers* e é uma variável bem simétrica, sendo que existe uma grande concentração nos dados abaixo dos 20 anos.

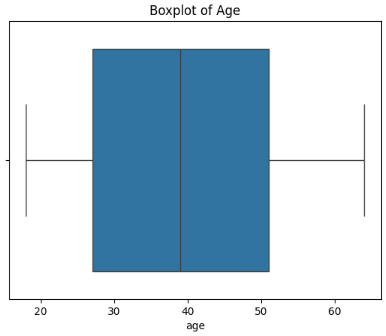


Imagem 3: *Boxplot* da variável idade

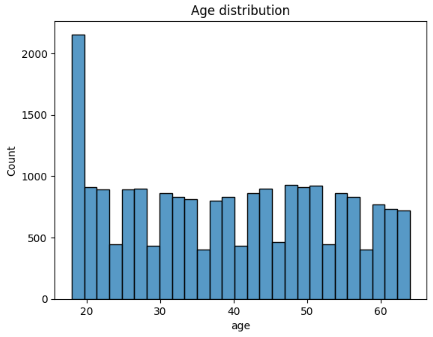


Imagem 4: Distribuição da variável idade

O IMC (ou *bmi*) é uma variável menos simétrica, se verificarmos nas tabelas de IMC vemos que a maior parte das pessoas nessa base de dados está acima do peso.

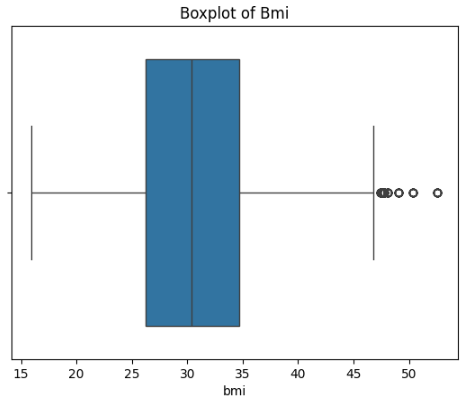


Imagem 5: *Boxplot* da variável IMC

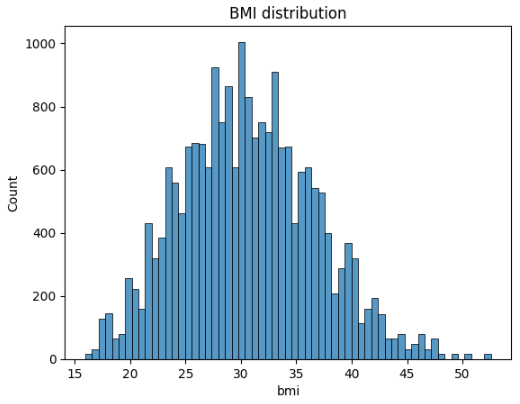


Imagem 6: Distribuição da variável IMC

A quantidade de não fumantes é maior que a quantidade de fumantes nessa base de dados, sendo que isso não varia com o gênero.

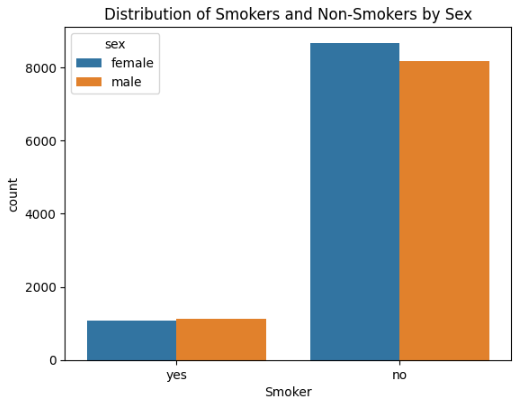


Imagem 7: Quantidade de fumantes

Verificando a variável de custos, vemos que alguns dos custos estão bem acima da média, mas é algo que pode ser explicado pelas demais variáveis.

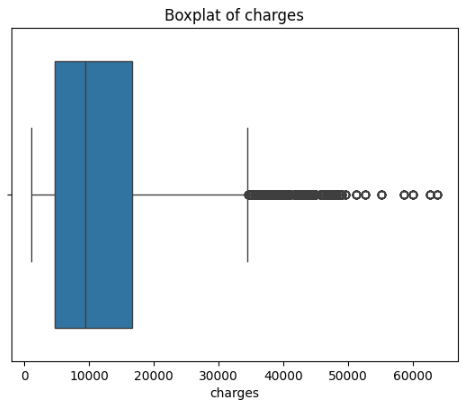


Imagem 8: *Boxplot* da variável custo

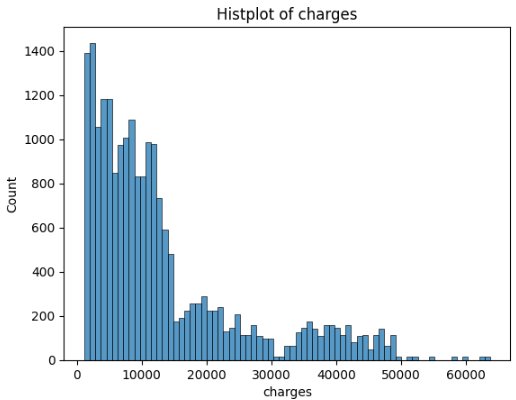


Imagem 9: Distribuição da variável custo

**Pré-processamento de dados**

O tratamento de dados realizado nesta base foi a aplicação do Label Encoder para converter as variáveis não numéricas e numéricas e a normalização para melhorar os resultados do modelo.

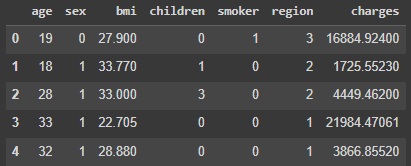


Imagem 10: Base de dados após a aplicação do Label Encoder

Com todos os dados numéricos foi possível realizar a correlação entre as variáveis. Veja que a variável mais relacionada com o custo é a de fumante, o que pode nos mostrar que a pessoa ser fumante pode ter uma relação com custos maiores no plano de saúde. Outras variáveis que se correlacionam com o custo são o IMC e a idade, mas com menos força que o tabagismo.



Imagem 11: Correlação entre as variáveis

Agora sim, podemos separar as bases de treino e teste para partir para aplicação do modelo de regressão múltipla. A proporção de 20% para dados de teste foi usada, sendo que a base de treino ficou com 14295 linhas e a base de teste ficou com 7041 linhas.

Somente após a separação das bases podemos aplicar a normalização (de modo que os dados não fiquem enviesados). Os resultados dessa normalização podem ser vistos no exemplo com os dados de teste nas imagens a seguir.

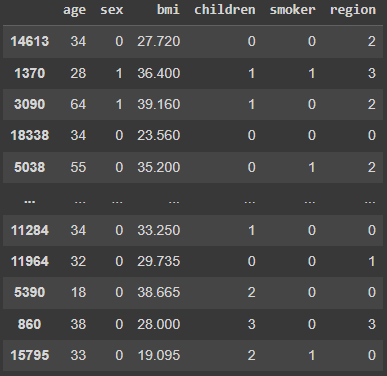


Imagem 12: Dados de teste antes da normalização



Imagem 13: Dados de teste após a normalização

**Modelo de regressão múltipla**

Aplicamos o modelo de regressão múltipla aos dados tratados. O modelo de regressão simples utiliza a função de primeiro grau (y = ax + b) para encontrar a melhor reta para a base de dados, ou seja, a reta que gera o menor erro versus os dados reais. No modelo de regressão múltipla se tenta encontrar a equação de reta, porém com mais coeficientes a\_n e b\_n, cada um para uma variável da base de dados.

Veja a seguir os resultados obtidos e conclusões sobre esses resultados.

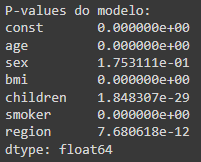


Imagem 14: Resultado dos P-values.

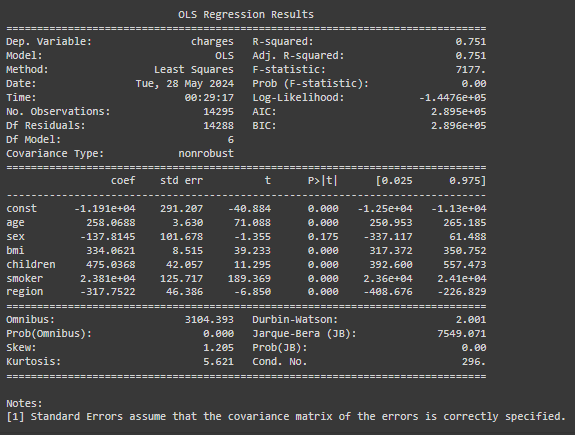


Imagem 15: Resultado sumarizado

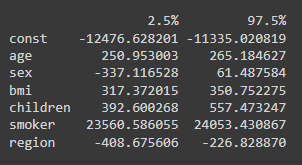


Imagem 16: Intervalos de confiança do modelo



Imagem 17: Erros do modelo

A principal métrica que podemos usar para avaliar a performance do nosso modelo é o R-score que ficou em 0.751, o que pode ser um bom resultado tendo em vista a quantidade de dados de treino e que não foi feito nenhum tratamento adicional nesses dados. Uma possível melhoria futura seria complementar e melhorar a base de dados para que o resultado seja acima de 0.8.

Nos intervalos de confiança podemos verificar o quanto o custo aumenta ou diminui de acordo com as demais variáveis. Veja, por exemplo, que a variável fumante aumenta muito o custo (intervalo mínimo de 23560 dólares), já a variável idade tem uma influência menor (intervalo máximo de 265 dólares).

Nos erros podemos verificar que o modelo tem um erro médio de 4169 dólares e que a raiz do erro quadrático é 6026 dólares. Uma melhoria para esse modelo seria diminuir a diferença entre esses erros, pois uma menor diferença significa ter uma menor influência dos *outliers* da base de teste.

**Conclusão**

Com este Tech Challenge o grupo conseguiu colocar em prática os assuntos abordados durante a primeira fase da pós, que foi bem focada em modelos de Machine Learning, que é a base para os modelos de inteligência artificial.

O modelo de regressão é amplamente utilizado em diversos setores e casos. Pode ser usado, por exemplo, para fazer previsões de vendas, encontrar possíveis valores de imóveis, entre outros.

Verificamos os passos necessários para criar um modelo e, principalmente, como avaliar os resultados do modelo para que possa ser usado para fazer novas previsões.

**Anexo**

Código utilizado para complementar a base de dados.

import pandas as pd

import numpy as np

# Carregar o conjunto de dados recém-carregado

new\_file\_path = "custos\_medicos.csv"

new\_df = pd.read\_csv(new\_file\_path)

# Gerar dados adicionais

num\_new\_rows = 10000

np.random.seed(42)

# Gerar novos dados

add\_data = {

'age': np.random.randint(18, 80, num\_new\_rows),

'sex': np.random.choice(['male', 'female'], num\_new\_rows),

'bmi': np.random.uniform(15, 40, num\_new\_rows),

'children': np.random.randint(0, 5, num\_new\_rows),

'smoker': np.random.choice(['yes', 'no'], num\_new\_rows),

'region': np.random.choice(['north', 'south', 'east', 'west', 'southwest'], num\_new\_rows),

'charges': np.random.uniform(1000, 50000, num\_new\_rows)

}

# Criar DataFrame com os novos dados

add\_df = pd.DataFrame(add\_data)

# Anexar os novos dados ao conjunto de dados existente

combined\_df = pd.concat([new\_df, add\_data], ignore\_index=True)

# Salvar o conjunto de dados atualizado em CSV

combined\_file = 'path\_to\_your\_file/medical\_costs\_updated\_with\_additional.csv'

combined\_df.to\_csv(combined\_file, index=False)