Fernando José dos Santos – 23025299

Emilly Mickeli Depine da Silva – 23025480

Renan Teixeira Pinheiro – 23025274

Gustavo Henrique Santos Araujo – 23025397

1. Justificativa para a Escolha da Técnica de Modelagem (Random Forest Regressor)

A tarefa de prever o preço de uma viagem, que é um valor numérico contínuo, é um problema de regressão em Machine Learning. Para este projeto, a escolha do `RandomForestRegressor` (Regresso Floresta Aleatória) se justifica por diversas vantagens:

- Desempenho Robusto: Trata-se de um método de "ensemble learning" (aprendizado de conjunto) que constrói múltiplas árvores de decisão (100 no modelo implementado) durante o treinamento e combina suas previsões. Isso geralmente resulta em maior precisão e estabilidade em comparação com árvores de decisão individuais.
- Tratamento de Relações Complexas: É eficaz em capturar relações não lineares e interações entre as features (como distância, tempo, `ProductID`, `Dia_Numero`) e a variável alvo (preço), sem a necessidade de transformações complexas prévias dos dados.
- Versatilidade com Dados: Lida bem com diferentes tipos de features (numéricas e categóricas, estas últimas após codificação apropriada).

2. Construção e Treinamento do Modelo

O modelo foi desenvolvido a partir de um conjunto de dados de viagens (`Resumo_v2.csv`), com o objetivo de prever o `price`. O processo de construção envolveu:

Pré-processamento dos Dados:

- Remoção de registros com `price` ausente.
- Conversão de features categóricas, como `ProductID`, para representação numérica através de `LabelEncoder`. A feature `Dia` foi transformada em `Dia_Numero` por meio de um mapeamento específico (ex: Segunda-feira = 0).

Engenharia de Features: Foram criadas novas features para enriquecer a informação disponível para o modelo:

- Localização: As coordenadas de strings foram convertidas para valores numéricos de latitude e longitude (`Lat1`, `Lon1`, `Lat2`, `Lon2`), e a partir delas, calculadas as diferenças `Delta_Lat` e `Delta_Lon`.

- Tempo: Horários de início e fim (`H_Inicio`, `H_Fim`) foram convertidos para minutos desde a meia-noite (`H_Inicio_Min`, `H_Fim_Min`), e a `Duração_Viagem` foi calculada.

Seleção de Features: Utilizou-se um conjunto definido de features, incluindo `ProductID`, `Dia_Numero`, `Tempo_Viagem_Min`, `Distância_km`, `Delta_Lat`, `Delta_Lon`, `H_Inicio_Min`, `H_Fim_Min` e `Duração_Viagem`.

Divisão dos Dados: O dataset foi particionado em conjuntos de treino (80%) e teste (20%) utilizando `train_test_split` com `random_state=42` para garantir a reprodutibilidade dos resultados.

Treinamento: O `RandomForestRegressor`, configurado com `n_estimators=100` e `random_state=42`, foi treinado utilizando o conjunto de dados de treino.

3. Avaliação do Modelo e Acurácia

O desempenho do modelo treinado foi avaliado no conjunto de teste (dados não utilizados no treinamento) usando as seguintes métricas:

- Erro Médio Absoluto (MAE): Esta métrica representa a média da diferença absoluta entre os valores previstos e os valores reais.
- Resultado: O modelo apresentou um MAE de R$ 4,83 (assumindo a moeda como Reais). Isso indica que, em média, as previsões de preço do modelo desviam R$ 4,83 do preço real da viagem no conjunto de teste.
- Coeficiente de Determinação (R^2 Score): Mede a proporção da variância na variável alvo (preço) que é explicável pelas features do modelo.

Resultado: O R^2 Score foi de 0.94.

Interpretação: Um valor de 0.94 é considerado excelente, indicando que 94% da variabilidade nos preços das viagens são explicados pelas features utilizadas. Isso sugere um bom ajuste do modelo e uma forte capacidade preditiva em dados novos.

Conclusão da Avaliação:

O modelo `RandomForestRegressor` demonstrou um desempenho robusto e alta acurácia no conjunto de teste. O R^2 elevado indica que as features selecionadas e a engenharia realizada foram eficazes, e o MAE fornece uma medida tangível do erro de previsão em unidades monetárias.