PREPARAÇÃO DE DADOS, ENGENHARIA DE RECURSOS E EXPLORAÇÃO DE MODELOS

Projecto: Tella Turismo Nacional – Plataforma Web de Planeamento de Viagens Inteligente

GRUPO: 09

1. Visão Geral

A fase de preparação de dados e engenharia de recursos é essencial para o sucesso do projecto Tella Turismo Nacional. Esta etapa garante a integridade e a utilidade das informações utilizadas nos modelos de Machine Learning, permitindo a extracção de insights relevantes e previsões precisas sobre custos e destinos turísticos. A qualidade dos dados é a base para recomendações personalizadas e estimativas financeiras realistas.

2. Coleta de Dados

Os dados utilizados foram recolhidos de fontes públicas e privadas, incluindo o Ministério da Hotelaria e Turismo (MINHOTUR), Instituto Nacional de Estatística (INE), TripAdvisor, Google Travel e pesquisas de mercado locais. As informações abrangem destinos turísticos, preços médios de transporte, hospedagem, alimentação e lazer. Os dados foram armazenados em formatos CSV e JSON.

3. Limpeza de Dados

A limpeza de dados incluiu o tratamento de valores ausentes, remoção de duplicatas e correcção de inconsistências textuais. Valores discrepantes foram tratados com métodos estatísticos, utilizando a técnica de IQR (Interquartile Range) para detecção e substituição de outliers. As variáveis categóricas foram padronizadas e normalizadas para evitar redundâncias.

4. Análise Exploratória de Dados (EDA)

A análise exploratória envolveu visualizações como histogramas, gráficos de dispersão e nuvens de palavras. Foram identificadas correlações entre variáveis como classificação média, sentimento e tipo de viajante. Os resultados demonstraram tendências de preferências de turistas e padrões de avaliação entre destinos.

5. Engenharia de Recursos

Foram criadas novas features para representar melhor o comportamento dos viajantes e as características dos destinos. Entre as variáveis derivadas estão:

- Índice de Sazonalidade (média de preços por estação);
- Pontuação de Popularidade (com base em avaliações de usuários);
- Factor de Sustentabilidade (classificação de práticas ecológicas dos destinos);
- Custo Total Previsto (função dos custos médios de transporte, alimentação e hospedagem). Essas variáveis enriquecem o modelo e permitem maior precisão nas previsões.

6. Transformação de Dados

Os dados numéricos foram normalizados usando o método Min-Max Scaling, enquanto as variáveis categóricas foram convertidas em vetores binários (one-hot encoding). O dataset final foi dividido em 80% para treino e 20% para teste, garantindo boa representatividade e evitando overfitting. A transformação permitiu alimentar eficientemente os algoritmos de aprendizado.

Exploração de Modelos

1. Selecção de Modelo

Foram testados diversos algoritmos, incluindo Regressão Logística, Random Forest, K-Means e um pouco de Redes Neurais. A escolha final baseou-se na performance em precisão, recall e tempo de execução. O modelo híbrido de recomendação (baseado em conteúdo e colaborativo) apresentou melhor desempenho.

2. Treinamento de Modelo

Os modelos foram treinados com validação cruzada k-fold e ajuste de hiperparâmetros via GridSearchCV. Os dados foram divididos em conjuntos de treino (80%) e teste (20%). A métrica principal de avaliação foi o F1-Score.

3. Avaliação do Modelo

Os resultados demonstraram um F1-Score médio de 0.70 a 0.93 e acurácia de 75% a 88%. A análise de erros indicou boa generalização, com menor precisão em classes com menos amostras. Curvas ROC e matrizes de confusão foram utilizadas para comparar modelos.

4. Implementação de Código

O código foi desenvolvido principalmente em Python, estruturado em módulos para coleta, limpeza, análise e modelagem. Foram utilizadas bibliotecas como Pandas, Scikit-learn, Matplotlib e TensorFlow. Comentários foram incluídos para garantir a compreensão das principais secções.