Universidade de Fortaleza

T326 - Ciência dos Dados
Projeto 2 - Aprendizado de Máquina: Olist Professor: Caio Ponte Turma: 16/17

A. Objetivo do Trabalho

O objetivo desta atividade final é permitir que os alunos demonstrem suas habilidades aplicando um *pipeline* de Aprendizado de Máquina (AM) em um problema real, utilizando o conjunto de dados da empresa Olist. A tarefa consiste em **prever quanto tempo levará para um pedido chegar à casa do cliente**, a partir das características do pedido, extraídas das tabelas disponíveis no *dataset*.

Para isso, os alunos devem:

- Realizar a extração e integração de *features* relevantes a partir das diversas tabelas do *dataset*, criando um conjunto de dados apropriado para regressão.
- Implementar todas as etapas de um *pipeline* de aprendizado de máquina, desde a definição do problema até a avaliação final do modelo, incluindo a Análise Exploratória de Dados (EDA) e o pré-processamento de dados.

B. Entrega

- **Prazo:** disponível no AVA.
- **Equipe:** até 5 alunos.
- **Formato:** O trabalho deve ser apresentado em formato de vídeo, com duração entre 10 e 15 minutos. Os alunos devem gravar um vídeo apresentando seu Notebook e explicando as principais etapas do projeto (que serão descritas abaixo). Todos os membros do grupo devem apresentar.
- **Envio:** O vídeo deve ser disponibilizado via link público ou não listado do YouTube e enviado pelo AVA.

C. Estrutura do Trabalho

O trabalho deverá ser desenvolvido em um Notebook Python, contemplando todas as etapas de um projeto de aprendizado de máquina. A seguir, estão listadas as etapas que devem **obrigatoriamente** ser abordadas. Cada etapa representa uma parte essencial do processo e deve ser detalhada para demonstrar a aplicação prática das técnicas de AM.

1. Definição do Problema

Explique o objetivo do trabalho no contexto da logística da Olist, destacando a importância de prever o tempo de entrega de pedidos. Apresente a estrutura do dataset e como as tabelas se relacionam.

Esta etapa inicial é fundamental para contextualizar o problema que será resolvido com técnicas de aprendizado de máquina. O aluno deve explicar o cenário em que o problema ocorre, a relevância dele, e fornecer informações sobre os dados utilizados.

Requisitos:

- ↗ Apresentar o contexto da empresa Olist.
- ↗ Descrever claramente o problema de regressão: o que representa cada amostra? Qual é a variável alvo?
- ↗ Identificar as principais tabelas utilizadas e justificar a escolha.
- ↗ Apontar o desafio real representado por esse problema e seu impacto.
- ★ Todos os itens devem ser contemplados!

2. Pré-processamento dos Dados

O pré-processamento dos dados é a fase onde os dados são preparados para os modelos de AM. É essencial para garantir que os dados estejam em um formato adequado e sem inconsistências que possam prejudicar o desempenho dos modelos.

Requisitos:

- ↗ Engenharia de atributos: criação das *features* para alimentar o(s) modelo(s)
- ↗ Filtragem e limpeza dos dados: remover ruídos e inconsistências.
- ↗ Tratamento de valores faltantes.
  - Métodos Pandas, SimpleImputer, KNNImputer etc
- ↗ Normalização/Padronização: adequar as escalas das variáveis
  - StandardScaler, MinMaxScaler, RobustScaler, Log etc
- ↗ Tratamento de variáveis categóricas: aplicar técnicas de codificação
  - OneHotEncoder, OrdinalEncoder, TargetEncoder etc
- ★ Pelos menos 4 itens devem ser contemplados!

3. Análise Exploratória dos Dados (EDA)

Realize uma EDA sobre o dataset construído, buscando entender suas características e padrões.

Requisitos:

- ↗ Distribuição das variáveis: gráficos e estatísticas descritivas.

- ↗ Identificação de outliers: evidenciar valores atípicos.
- ↗ Matriz de correlações: verificar a relação entre as variáveis.
- ↗ Estatísticas descritivas: exibir médias, medianas, desvios padrões etc.
- ↗ Mapas geográficos: análise espacial através dos valores de latitude e longitude.
- ★ Pelos menos 4 itens devem ser contemplados!

4. Treinamento do Modelo

Os dados processados são utilizados para treinar diferentes modelos de aprendizado de máquina. O objetivo é construir modelos que possam fazer previsões e generalizar baseados nos dados fornecidos, com ajuste de parâmetros e técnicas de treinamento/validação.

Requisitos:

- ↗ Separar dados em treino e teste ou usar validação cruzada.
- ↗ Implementar pelo menos três modelos de aprendizado de máquina, dentre:
  - Modelos Lineares e variações: Regressão Linear, Ridge, Lasso etc;
  - Modelos de Árvores de Decisão;
  - Modelos Ensembles: Árvores de Decisão, Random Forest, Gradient; Boosting, Stacking etc;
  - Modelos baseados em instâncias: KNN
- ↗ Ajuste de hiperparâmetros: apresentar o processo de otimização e escolha dos melhores parâmetros para cada modelo
  - GridSearchCV ou RandomizedSearchCV
- ★ Todos os itens devem ser contemplados!

5. Avaliação do Modelo e Conclusão

A avaliação envolve medir seu desempenho utilizando métricas apropriadas (para regressão) e comparar diferentes modelos para escolher o que melhor resolve o problema. Essa fase é crucial para verificar a eficácia e adequação das soluções propostas.

Requisitos:

- ↗ Métricas: usar RMSE, MAE, R2 ou outras adequadas à regressão.
- ↗ Benchmark dos modelos: comparar o desempenho entre os modelos treinados, justificando a escolha do modelo final.
- ↗ Discussão sobre possíveis melhorias ou ajustes futuros, baseado nas métricas observadas.
- ↗ Conclusão: Apresentar uma conclusão objetiva sobre o modelo escolhido, discutindo se ele atende às expectativas iniciais e quais são as implicações práticas de sua utilização no contexto do problema.
- ★ Todos os itens devem ser contemplados!