PÓS-GRADUAÇÃO EM BUSINESS ANALYTICS – ISCAP IPP

Análise Eploratória de Dados & Modelos de Machine Learning

1. ENQUADRAMENTO

Considere o *dataset* "bank_data" disponível na pasta do trabalho. O *dataset* em questão contém dados relacionados com campanhas de *marketing* direto de uma instituição bancária portuguesa. O *dataset* contém 20 variáveis explicativas relacionadas com as características dos clientes do banco que foram contactados durante a campanha de *marketing*. A variável a prever classifica se um determinado cliente subscreveu a um depósito a prazo ou não. A descrição de cada variável encontra-se na secção que se segue.

O objetivo deste trabalho é realizar um estudo que incida primeiramente na análise completa dos dados. Para tal, deverá utilizar os conhecimentos adquiridos no módulo de "Análise Exploratória de Dados". Deverão utilizar o *output* deste estudo para selecionar o algoritmo de *Machine Learning* adequado a aplicar neste contexto para prever se um determinado cliente irá subscrever ao depósito a prazo ou não. A justificação para a escolha de um algoritmo em detrimento de outros deve levar em consideração diversos fatores, como interpretabilidade, precisão do modelo, entre outros. Os pontos fortes e fracos de cada modelo aplicado devem ser enumerados. Identifique também as variáveis que impactam mais significativamente na previsão. Para esta segunda parte deverão utilizar os conhecimentos adquiridos no módulo "Modelos de Machine Learning".

Este trabalho está integrado no programa dos módulos "Análise Exploratória de Dados" e "Modelos de *Machine Learning*" e será avaliado conjuntamente pelos docentes responsáveis por cada módulo. Assim, o resultado final refletirá o desempenho dos alunos em ambos os módulos, de acordo com os critérios de avaliação definidos por cada professor.

Como entregáveis, pretendemos os seguintes documentos:

- Relatório que descreva a solução implementada como, também, os resultados obtidos.
- O *Jupyter notebook* que contenha o código utilizado para atingir os resultados obtidos.

O trabalho contém ainda um componente bónus opcional, que no caso de ser entregue, acresce o valor da nota final. Esta componente extra requer os seguintes documentos:

- Relatório em *Power BI* (PBI) com os principais resultados da Análise Exploratória de dados (1ª parte).
- Relatório em *Power BI* (PBI) com os principais resultados do modelo de *Machine Learning* (2ª parte).

2. DESCRIÇÂO DOS DADOS FORNECIDOS

No total, existem 41188 registos e 21 *features*:

Feature	Tipo	Descrição
age	Numérica	Idade do cliente.
job	Categórica	Tipo de emprego.
marital	Categórica	Estado civil.
education	Categórica	Grau de escolaridade.
default	Categórica	Se o cliente já falhou o pagamento do empréstimo.
housing	Categórica	Se tem crédito habitação.
loan	Categórica	Se tem crédito ao consumo.
contact	Categórica	Tipo de comunicação.
month	Categórica	Mês em que o cliente foi contactado durante a campanha de *marketing*.
day_of_week	Categórica	Dia da semana em que o cliente foi contactado durante a campanha de *marketing*.
duration	Numérica	Duração do contacto em segundos.
campaign	Numérica	Número de contactos que o cliente que recebeu durante a campanha de marketing.
pdays	Numérica	Número de dias decorridos desde o último contacto referente a outra campanha de *marketing*.
previous	Numérica	Número de contactos decorridos desde o último contacto referente a outra campanha de *marketing*.
poutcome	Numérica	Resultado da última campanha de marketing (se o cliente aderiu ao produto ou não).
emp.var.rate	Numérica	Taxa de emprego.
cons.price.idx	Numérica	Índice do consumidor.
cons.conf.idx	Numérica	Índice de confiança do consumidor.

euribor3m	Numérica	Taxa Euribor a 3 meses.
nr.employed	Numérica	Número de trabalhadores.
y	Categórica	**Variável a prever:** Se o cliente aderiu ou não ao depósito bancário publicitado durante a campanha de *marketing*.

***Missing values*:** Existem vários valores em falta em algumas *features* categóricas, todos codificados com a etiqueta "unkown". Deverão utilizar os conhecimentos de "Análise Exploratória de Dados" para tratar estes casos.

3. OBJETIVOS PRINCIPAIS DO TRABALHO

- Executar análise exploratória de dados (EDA).
- Processar e enriquecer os dados com novas *features* (*feature engineering*).
- Selecionar os modelos de *Machine Learning* adequados para os dados em questão.
- Treinar e otimizar hiperparâmetros dos modelos selecionados.
- Análise de interpretabilidade dos modelos que foram treinados.
- Utilizar as métricas de performance para comparar os modelos selecionados.
- Selecionar o melhor modelo face ao problema exposto neste trabalho.

4. ENTREGA

O trabalho é elaborado em equipas de 2 elementos. As equipas de trabalho serão acompanhadas pela Prof. Alexandra Oliveira e o Prof. Luís Dias, durante o decorrer das aulas. Cada equipa terá de entregar até ao dia 23 de fevereiro de 2025 um relatório em formato pdf com os seguintes tópicos:

1. Resumo: descrever a motivação, metodologia aplicada e principais *insights* obtidos de uma forma resumida (máximo de 250 palavras);
2. Introdução: onde apresentam a motivação do projeto;
3. Metodologia: onde detalham a metodologia adotada como, também, os resultados obtidos;
4. Desafios e oportunidades: onde justificam o potencial valor acrescentado da aplicação do projeto como, também, as principais dificuldades que foram sendo encontradas durante o desenvolvimento do projeto;
5. Conclusões: onde resumem as conclusões do projeto e apontam sugestões.

O relatório deverá ter no máximo 25 páginas (excluindo os anexos). As equipas terão, também, de entregar o *Jupyer notebook* com os resultados obtidos. Todos os entregáveis do trabalho deverão ser **enviados por *email* para ambos os docentes até ao dia 23 de fevereiro de 2025**. Em caso de incumprimento, cada dia de atraso corresponderá a uma dedução de 2 valores na nota final do trabalho. A avaliação do trabalho inclui a componente do relatório, com um peso de 50% e o *Jupyer notebook*, com um peso de 50%.

A componente bónus do trabalho pode valer até mais 2 valores, onde 50% correspondem ao PBI relativo à Análise Exploratória dos Dados e os recentes 50%, ao PBI relativo aos resultados do Modelo de *Machine Learning*.