iscte TECNOLOGIAS E ARQUITETURA

Algoritmos para Big Data
Projeto final

Sistema de classificação com aprendizagem automática **10 de abril de 2025**

1 Introdução

Este projeto enquadra-se na área de análise de dados em larga escala, com foco na implementação de uma solução computacional para classificação de dados recorrendo a técnicas de aprendizagem automática. Nesse sentido, a implementação pressupõe a utilização de funcionalidades disponíveis nas plataformas Apache Spark e Apache Kafka.

A realização do projeto será feita por grupos de trabalho constituídos por dois ou três estudantes.

2 Problema

Pretende-se que seja implementada uma solução computacional para classificação de dados, assente na construção de um modelo de aprendizagem supervisionada com base em algoritmos referidos nas aulas.

A escolha do domínio de dados e respetivo conjunto de dados a utilizar, bem a formulação do próprio problema em análise, será da responsabilidade dos autores do trabalho.

2.1 Domínio de dados e algoritmia

A tabela 2.1 apresenta a lista de hipóteses de seleção para o conjunto de dados. O problema de classificação a formular terá como base a seleção de dados realizada, o que condicionará a análise subsequente.

Os algoritmos a utilizar para a construção do modelo têm de fazer parte da plataforma Apache Spark.

Tabela 1: Lista de fontes de informação para seleção de dados.

Dados (#)	URL de acesso à fonte de informação
1	A Molecular Benchmark for Disease and Target Based Machine Learning
2	Acquire Valued Shoppers Challenge
3	Chicago Crime Dataset
4	Flight Status Prediction
5	ACI IoT Network Trafic Dataset 2023
6	Citi Bike Stations
7	MeteoNet North-West France
8	Smart Meters in London

2.2 Tarefas principais

Para alcançar os objetivos propostos, devem ter em atenção os seguintes aspetos, entre outros:

1. Leitura e validação de dados.
2. Análise exploratória de dados, tendo em visto o entendimento dos mesmos.
3. Construção do modelo de aprendizagem automática, no pressuposto que devem construir várias soluções e optar no final por aquela que considere ser a melhor solução.
4. Utilização do modelo criado, em dois contextos de utilização distintos (*model deployment*):
   (a) Em modo *batch*, ou seja, em ambiente análogo ao que foi utilizado aquando da construção do modelo.
   (b) Em modo *streaming*, recorrendo neste caso ao sistema de mensagens Apache Kafka para simular o processo de fluxo de dados.

Durante a construção do modelo, devem ter em consideração mais do que um algoritmo de classificação. Assim, será necessário realizar uma análise comparativa entre as várias soluções, quer em termos de avaliação de métricas, quer em termos de complexidade, sobretudo temporal (tempo de execução do código). Neste contexto, considerem também alterações à dimensão de dados.

3 Implementação

A implementação da solução deve ser modular, ou seja, deve ser composta por mais do que um notebook ou módulo Python. Compete aos autores do trabalho estruturar de forma criteriosa o código implementado. Assume-se que o mesmo é auto-explicativo, contendo comentários com nível de detalhe apropriado.

Por outro lado, chama-se a atenção para os seguintes aspetos, também já referidos ao longo das aulas:

- A escolha do domínio de dados e consequentemente seleção de dados, bem como a formulação do problema em estudo, são da maior importância para o sucesso do projeto como um todo. Estas fases não devem ser menosprezadas, em termos relativos.
- Por questões de produtividade, devem ser considerados dois conjuntos de dados aquando do desenvolvimento da solução. Assim, para além dos dados originais na sua íntegra, deve ser utilizado um conjunto de dados de menor dimensão (sub-conjunto dos anteriores), para o caso de tarefas intensivas e frequentes, inerentes ao próprio processo de desenvolvimento da solução.
- Cada notebook ou módulo Python deverá ser autónomo em termos de dados a utilizar. Ou seja, sugere-se que o código seja estruturado de modo a ler e gravar dados entre etapas distintas do projeto. Por exemplo: a geração de um gráfico ou tabela não deve implicar a realização da simulação/processamento no mesmo instante. Preferencialmente, os dados devem ser importados já processados e a partir de ficheiros.

4 Material a entregar e prazos estabelecidos

O projeto terá duas submissões, uma parcial, intermédia e outra final. Em ambos os casos, a submissão será feita através de um arquivo em formato **zip** (extensão zip e não outra) com os respetivos elementos de avaliação, a ser submetido na plataforma de ensino Moodle. Os *links* a utilizar serão indicados em momento oportuno.

Importante: Qualquer uma das submissões no Moodle não pode conter ficheiros de dados.

4.1 Submissão intermédia

Os elementos de avaliação na submissão intermédia são os seguintes:

- Notebooks e/ou módulos Python relativos às duas tarefas iniciais acima refiradas:
  1. Leitura e validação de dados.
  2. Análise exploratória de dados.
- Um relatório muito sucinto e em formato **pdf** sobre as referidas tarefas, podendo também ser incluído um relatório autónomo sobre qualidade de dados.

O prazo da submissão intermédia é: **23:59 de 17 de maio de 2024**.

4.2 Submissão final

Os elementos de avaliação na submissão final são os seguintes:

- Notebooks e/ou módulos Python que constituem a solução computacional.
- Um relatório final sobre o trabalho realizado, sucinto e em formato **pdf**, com o máximo de oito páginas. O relatório deve:
  - Conter uma descrição do problema em estudo e respetivos dados utilizados.
  - Abordar os aspetos mais relevantes sobre as decisões tomadas.
  - Incluir informação sobre as experiências e testes realizados. Por exemplo, indicação dos tempos de execução, recursos e capacidades de processamento utilizadas.
  - Incluir uma análise sobre os resultados obtidos, não só em termos de desempenho da solução mas, sobretudo, na perspectiva do problema formulado. Ou seja, com base no problema enunciado, indicar quais são as conclusões a retirar após a análise que é feita aos resultados obtidos.
  - Incluir informações adicionais que os autores do trabalho considerem relevantes.

O prazo da submissão final é: **23:59 de 31 de maio de 2024**.

5 Discussão e valorização do trabalho

O trabalho será discutido presencialmente, em local e hora a indicar após submissão final do projeto e de acordo com a disponibilidade dos membros do grupo e dos docentes.

Por fim, relembra-se que as regras de avaliação da unidade curricular relativamente ao projeto estabelecem o seguinte:

- A avaliação do projeto tem uma ponderação de 30% na nota final da unidade curricular.
- A valorização do projeto será repartida pelas duas submissões, com uma ponderação de 30% para a submissão intermédia e 70% para a submissão final.
- O resultado da avaliação do projeto é individual.
- Não havendo submissão parcial, considera-se que a avaliação na unidade curricular será feita através da modalidade de avaliação por exame.