A black and white logo

Description automatically generated

**Mestrado em Engenharia Informática**

Algoritmos para Big Data

17 de maio de 2025

Previsão de Atrasos de Voos

Henrique Niza (131898)

Paulo Francisco Pinto (128962)

Rute Roque (128919)

* **Formulação do Problema**

O presente projeto tem como objetivo responder a um conjunto de questões práticas relacionadas com o desempenho dos voos comerciais nos Estados Unidos. Com base no conjunto de dados “Flight Delay Dataset 2018–2022”, foram delineadas as seguintes perguntas orientadoras:

* É possível prever quais voos irão ser cancelados ou sofrer atrasos?
* É possível prever o tempo de atraso de um voo?
* Como se comparam as diferentes companhias aéreas entre si?

Estas questões definem o âmbito do problema de classificação supervisionada, onde o objetivo principal é construir um modelo capaz de prever se um voo será **atrasado (ou não)** com base em atributos conhecidos antes da partida (por exemplo, data, hora, aeroporto de origem, companhia aérea, entre outros).

Adicionalmente, procura-se analisar tendências nos dados históricos, identificar padrões de atraso por companhia aérea e construir modelos com utilidade prática tanto em contexto académico como industrial, nomeadamente para planeamento de operações e gestão de expectativas dos passageiros.

* **Descrição do Conjunto de Dados**

O conjunto de dados utilizado neste projeto intitula-se “Flight Delay Dataset 2018–2022”, disponível publicamente na plataforma Kaggle (<https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022>). Este dataset reúne informações detalhadas sobre voos comerciais nos Estados Unidos entre 2018 e 2022, incluindo horários programados e reais, atrasos, cancelamentos, distâncias, e dados operacionais das companhias aéreas.

Para conveniência e eficiência computacional, foram utilizadas as versões combinadas anuais dos dados (Combined\_Flights\_YYYY.csv), onde colunas com valores nulos foram previamente filtradas. A versão original dos dados mensais (Flights\_YYYY\_M.csv) também está disponível, contendo mais de 100 atributos por voo.

Entre os campos mais relevantes para a tarefa de classificação (voo atrasado vs não atrasado) destacam-se:

* **FlightDate**: data do voo (formato yyyymmdd);
* **DepDelayMinutes** / **ArrDelayMinutes**: atraso à partida/chegada em minutos;
* **DepDel15** / **ArrDel15**: atraso superior a 15 minutos (1 = sim, 0 = não);
* **Cancelled**: voo cancelado (1 = sim, 0 = não);
* **Origin** e **Dest**: aeroportos de origem e destino;
* **Marketing\_Airline\_Network**: código da companhia aérea responsável;
* **CRSDepTime** / **CRSArrTime**: hora programada de partida/chegada.

Este dataset é amplamente utilizado em contextos académicos por ser ideal para tarefas de análise exploratória, classificação binária e previsão de atrasos, com uma estrutura realista e rica em atributos.

* **Tarefas Realizadas**

Durante a fase inicial do projeto foram realizadas as seguintes tarefas, conforme definido nos critérios de avaliação da submissão intercalar:

* **Leitura e Validação de Dados**:
  + Importação dos ficheiros Combined\_Flights\_YYYY.csv no ambiente Spark;
  + Verificação do schema dos dados e consistência dos tipos;
  + Contagem de linhas e detecção de valores nulos ou inconsistências;
  + Remoção de colunas com informação totalmente ausente ou redundante;
  + Amostragem de um subconjunto (10% a 30%) dos dados para desenvCololvimento.
* **Análise Exploratória de Dados (EDA)**:
  + Cálculo de estatísticas descritivas (médias, desvios, percentis);
  + Análise de distribuição de atrasos por companhia, por hora e por aeroporto;
  + Visualizações com plotly.express para comparar atrasos por dia da semana, por mês e por origem;
  + Cálculo de correlações entre variáveis numéricas (e.g. atraso vs distância);
  + Identificação de padrões sazonais ou operacionais.
* **Relatório de Auditoria de Qualidade de Dados**

Durante a análise dos dados foram identificados os seguintes aspetos relacionados com a qualidade da informação:

* **Valores em falta**: Algumas colunas como Tail\_Number, CancellationCode ou DivX... apresentam um nível elevado de ausências e foram descartadas ou ignoradas para a fase inicial do modelo;
* **Valores inconsistentes**: Alguns registos contêm DepDelayMinutes negativos combinados com DepDel15 = 1, o que indicia erro no preenchimento;
* **Colunas pouco informativas**: Atributos derivados de aeroportos divergentes (DivXAirport) foram mantidos apenas para futuras extensões do modelo;
* **Desequilíbrio da variável alvo**: A maioria dos voos não estão atrasados, o que constitui um caso típico de class imbalance a ser tratado nas fases seguintes do projeto.

Medidas adotadas incluíram limpeza parcial, tipagem adequada, e exclusão de colunas com >90% de nulls.