# Trabalho de Mineração de Dados - N2

## Parte 1 — Análise de Anomalias / Outliers com Clustering (2,0 pontos)

Objetivo: Verificar se existe um grupo incomum de passageiros (outliers) no dataset sfo\_2018\_data\_file\_final\_Weightedv2.  
  
Passos realizados:  
1. Carregamento e limpeza do dataset com remoção de valores ausentes.  
2. Seleção das variáveis: NETPRO, Q20Age, Q21Gender, Q22Income, Q23FLY, Q5TIMESFLOWN, Q6LONGUSE.  
3. Normalização dos dados com StandardScaler.  
4. Aplicação do método do cotovelo para definição do número ideal de clusters (K = 3).  
5. Agrupamento com K-Means.  
6. Identificação do cluster incomum e análise do perfil médio.  
  
Resultado:  
- Cluster incomum identificado: Cluster 1  
- Tamanho: 6,19% dos passageiros  
- Perfil médio do Cluster 1:  
 - NETPRO: 9.75  
 - Idade (Q20Age): 0.73  
 - Gênero (Q21Gender): 0.23  
 - Renda (Q22Income): 0.11  
 - Frequência de voo (Q23FLY): 0.18  
 - Experiência de voo (Q5TIMESFLOWN): 2.35  
 - Tempo de uso do aeroporto (Q6LONGUSE): 2.54  
  
Ferramenta utilizada: Python (pandas, sklearn, matplotlib, seaborn)

## Parte 2 — Regras de Associação em Problema Não-Comercial (2,0 pontos)

Problema: Encontrar padrões de comportamento entre usuários do aeroporto com base em hábitos e satisfação.  
  
Dados utilizados:  
- Dataset: sfo\_2018\_data\_file\_final\_Weightedv2.xlsx  
- Variáveis: NETPRO, Q6LONGUSE, Q23FLY  
- Pré-processamento com binning de variáveis contínuas  
  
Modelagem:  
1. Seleção das colunas relevantes  
2. Discretização (binning) das variáveis numéricas  
3. Conversão para formato transacional (one-hot encoding)  
4. Aplicação do algoritmo FP-Growth com suporte mínimo de 20%  
  
Regras geradas (exemplos):  
1. Se Q6LONGUSE então NETPRO — suporte: 96,69%, confiança: 99,27%, lift: 1,00  
2. Se NETPRO então Q6LONGUSE — suporte: 96,69%, confiança: 97,52%, lift: 1,00  
3. Se NETPRO então Q23FLY — suporte: 92,27%, confiança: 93,07%, lift: 1,00  
4. Se Q23FLY então NETPRO — suporte: 92,27%, confiança: 99,20%, lift: 1,00  
  
Ferramenta utilizada: Python (mlxtend, pandas)

## Parte 3 — Regressão Logística como Alternativa à Linear (2,0 pontos)

Problema: Prever se um passageiro está satisfeito (NETPRO >= 9) com base em variáveis demográficas e comportamentais.  
  
Dados utilizados:  
- Dataset: sfo\_2018\_data\_file\_final\_Weightedv2.xlsx  
- Variáveis preditoras: Q20Age, Q21Gender, Q22Income, Q23FLY, Q5TIMESFLOWN, Q6LONGUSE  
- Variável alvo: Satisfação binária (1 se NETPRO >= 9, senão 0)  
  
Modelagem:  
1. Criação da variável alvo binária  
2. Separação entre treino e teste (75%/25%)  
3. Normalização dos dados  
4. Treinamento com Logistic Regression  
5. Avaliação com métricas: acurácia, f1-score, matriz de confusão, ROC AUC  
  
Resultados:  
- Acurácia: 55%  
- ROC AUC: 0.55  
- Matriz de confusão:  
 [[179 173]  
 [144 207]]  
- Coeficientes:  
 - Q20Age: -0.07  
 - Q21Gender: 0.02  
 - Q22Income: -0.16  
 - Q23FLY: -0.16  
 - Q5TIMESFLOWN: 0.02  
 - Q6LONGUSE: 0.23  
  
Ferramenta utilizada: Python (pandas, scikit-learn)