Terceira Prova - Introdução à Ciência de Dados

16 de setembro de 2025

A questão 1 deve ser entregue em um arquivo `txt`. A questão 2 deve ser entregue em um arquivo `R`. O arquivo deve conter o código e os comentários utilizados para chegar às respostase. A prova deve ser entregue para o email `pedrofranklin@ufu.br` e o título do email deve ser `P3 - <seu nome> - FAMAT31308`. Códigos podem ser consultados; você pode usar a internet apenas para consultar documentação de funções. Não é permitido consultar colegas ou qualquer outro tipo de material. A prova deve ser feita individualmente.

Questão 1

Você está ajudando a desenvolver um sistema para classificar e-mails em duas categorias: spam e não-spam. O sistema precisa equilibrar bem os erros. Contudo, um tipo de erro é considerado particularmente grave: quando um e-mail que não é spam é classificado como spam (falso positivo), pois isso pode levar o usuário a perder mensagens importantes de trabalho, pessoais ou financeiras (por exemplo: uma convocação para entrevista, um boleto, um comprovante de matrícula). Assim, minimizar falsos positivos é prioridade.

Foram testados dois modelos, e as matrizes de confusão nos dados de teste encontram-se a seguir. As colunas apresentam as previsões do modelo. Considere que a classe spam é a positiva.

Modelo Floresta Aleatória

	Prev. Spam	Prev. Não-Spam
Real Spam	85	15
Real Não-Spam	30	170

Modelo SVM

	Prev. Spam	Prev. Não-Spam
Real Spam	70	30
Real Não-Spam	20	180

Com base no problema descrito (em que classificar um e-mail legítimo como spam é o erro mais grave), escolha qual dos dois modelos você considera mais adequado.

Questão 2 — Classificação com Floresta Aleatória

Objetivo: construir um classificador por **Floresta Aleatória** para distinguir objetos do telescópio Kepler entre **CONFIRMED** e **NOT_CONFIRMED**.

Você receberá dois data frames:

- `train_data`: conjunto de treino
- `test_data`: conjunto de teste

A coluna-alvo é `binary_class` (fator com níveis `CONFIRMED` e `NOT_CONFIRMED`). As demais colunas são preditoras numéricas/categóricas relacionadas ao trânsito do exoplaneta e às propriedades da estrela.

Dicionário de variáveis (preditoras)

Sinal e geometria do trânsito

- `koi_period` — período orbital estimado (dias).
- `koi_time0bk` — época de referência do trânsito (BKJD).
- `koi_duration` — duração do trânsito (horas).
- `koi_depth` — profundidade do trânsito (ppm).
- `koi_model_snr` — razão sinal-ruído do modelo de trânsito.
- `koi_impact` — parâmetro de impacto do trânsito.
- `koi_ror` — razão dos raios planeta/estrela (Rp/R).
- `koi_srho` — densidade estelar inferida pelo ajuste de trânsito (g/cm³).
- `koi_num_transits` — número de trânsitos utilizados no ajuste.

Flags de triagem de falsos positivos (vetting)

- `koi_fpflag_nt` — "não parece trânsito" (0/1).
- `koi_fpflag_ss` — indícios de eclipse estelar/estelar-estelar (0/1).
- `koi_fpflag_co` — deslocamento de centróide/contaminação espacial (0/1).
- `koi_fpflag_ec` — coincidência de efemérides/contaminação (0/1).

Parâmetros da estrela

- `koi_steff` — temperatura efetiva (K).
- `koi_slogg` — gravidade superficial (log g, cgs).
- `koi_srad` — raio estelar (em raios solares).
- `koi_smet` — metalicidade [Fe/H] (dex).

Alvo

- `binary_class` — rótulo binário: `CONFIRMED` ou `NOT_CONFIRMED`.

Tarefas

1. **Preparação**

   a) Verifique tipos de dados (numéricos/fatores) e níveis de `binary_class`.

   b) Descreva rapidamente a distribuição da classe no treino.

2. **Treinamento**

   a) Ajuste uma **Floresta Aleatória** usando `train_data` (defina `CONFIRMED` como classe positiva).

   b) Documente os principais hiperparâmetros utilizados (ex.: número de árvores, `mtry`).

3. **Avaliação no teste**

   a) Gere a **matriz de confusão** no `test_data`.

   b) Calcule **acurácia**, **precisão**, **sensibilidade** para a classe positiva (`CONFIRMED`).

   c) Interprete a matriz de confusão no contexto: o que significa um falso positivo (classificar `NOT_CONFIRMED` como `CONFIRMED`) e um falso negativo?

4. **Importância de variáveis e visualização (1,0 pt)**
   Comente as variáveis mais relevantes.

5. **Resumo final (0,5 pt)**
   Em algumas linhas, sintetize: desempenho do modelo (métricas), variáveis determinantes, principais erros (padrões de FP/FN) e gráficos que acharem interessantes para comunicar os resultados.