FA084 – Introdução à Mineração de Dados: Conceitos e Aplicações na Agricultura

Prova – 10 de Julho de 2020

PARTE 1: Definição de Conceitos (Valor Total: 2,0)

OBSERVAÇÕES.

- Entre as três questões apresentadas, **ESCOLHA DUAS** para responder.
  Valor de cada questão: 1,0.
- Respostas com **MÁXIMO DE 5 LINHAS**. Fonte: Arial 10. Espaçamento: Simples.

Questão 1.1: Certamente, o ideal é que os resultados de um modelo não apresentem 'Falsos Positivos' nem 'Falsos Negativos'.
Porém, se isso ocorrer, é melhor o modelo apresentar uma quantidade maior de 'Falso Positivo' ou de 'Falso Negativo'?
JUSTIFIQUE sua resposta e **APRESENTE UM EXEMPLO** ilustrativo.

Questão 1.2: O que é 'Balanceamento de Dados'?
Quais as principais abordagens para balancear um conjunto de dados?
Em que momento do processo deve ser feito?
Antes ou depois da criação de conjuntos de treino e teste? Por que?

Questão 1.3: O que são, para que servem e como são criados conjuntos de 'Treino', 'Teste' e 'Validação'?

EXPLICAÇÃO DO CONJUNTO DE DADOS

Considere o conjunto de dados FA084-P1-Dataset-UCR.csv.

Trata-se de um conjunto de dados de produção da Usina Costa Rica, do Grupo ATVOS.

Descrição dos atributos:

safra:	Ano da Safra (Plantio/Colheita): 21213 (2012/2013), 21314 (2013/2014), etc.
codFaz:	Código da Fazenda onde se realizou a colheita
bloco:	Bloco onde se realizou a colheita
talhão:	Talhão onde se realizou a colheita
estagio:	Número de corte da colheita: 1° corte (12m, 15M e 18m), 2° corte, 3° corte, etc
variedade:	Variedade da cana
usina:	Código da Usina
ambProd:	Ambiente de Produção do talhão
tchEst:	Produtividade (TCH – t/ha) ESTIMADA do talhão
tchReal:	Produtividade (TCH – t/ha) REAL do talhão
days:	Intervalo (em dias) entre a última colheita e a colheita anterior

PARTE 2: Exploração do conjunto de dados (Valor Total: 3,0)

Para responder as questões 2.1 a 2.4, as instruções dos itens A, B e C devem ser atendidas

2.A) Mantenha apenas os dados das **QUATRO SAFRAS COM MAIOR NÚMERO DE REGISTROS (safras 2012/2013, 2013/2014, 2014/2015 e 2015/2016)**, **DESCARTANDO OS DEMAIS REGISTROS**.

2.B) Mantenha apenas os dados das **TRÊS VARIEDADES COM MAIOR NÚMERO DE REGISTROS (RB867515, SP813250 e RB855453)**, **DESCARTANDO OS DEMAIS REGISTROS**.

2.C) Com relação ao estágio (número de cortes):

2.C.1) Estágios '12m', '15M' e '18m' são CANA PLANTA (1° corte).
O que os diferencia é que são cana de 12 meses, 15 meses ou 18 meses.
Para que todos esses registros sejam 'cana de primeiro corte', considere todos como estágio '1'.

2.C.2) Mantenha apenas os dados dos registros **DOS CINCO PRIMEIROS CORTES** (cortes de 1 a 5), **DESCARTANDO OS DEMAIS REGISTROS**

Questão 2.1.

0,50 Faça uma análise dos resultados encontrados nos itens 2.1.A e 2.1.B.

2.1.A) Qual a média de **Produtividade Real** da cana para cada estágio de corte?

2.1.B) Qual a média de **Produtividade Real** da cana para cada variedade?

Questão 2.2.

0,50 Faça uma análise dos resultados encontrados nas figuras 2.2.A e 2.2.B.

2.2.A) 'Boxplot' da **Produtividade Real** da cana para cada estágio de corte

2.2.B) 'Boxplot' da **Produtividade Real** da cana para cada variedade

Questão 2.3.

Considere o erro entre a produtividade estimada e a produtividade real (colhida).

0,75 2.3.A) Faça uma análise crítica dos resultados obtidos com as figuras solicitadas:

- Histograma dos erros entre as produtividades estimada (tchEst) e real (tchReal)
  - Geral,
  - Por variedade,
  - Por estágio (número de corte)
- Gráfico 'scatter' comparando tchReal (eixo x) e tchEst (eixo y).

DICA: Embora não seja obrigatório, se fizer esse gráfico separando por cores de acordo com alguns grupos, a análise pode ficar mais rica

0,75 2.3.B) Faça uma análise crítica dos resultados obtidos com os itens a seguir.

- Valores da produtividade estimada e real, quando o erro foi máximo
- Estágio (número do corte) e variedade quando o erro foi máximo
- Produtividade real média, máxima e mínima para esta variedade neste corte.

Questão 2.4.

0,50 Considere o erro entre a produtividade estimada e a produtividade real (colhida).

Faça uma análise crítica dos resultados obtidos com as figuras solicitadas:

- Gráfico 'scatter' estagio (eixo x) e tchErro (eixo y)
- Gráfico 'scatter' variedade (eixo x) e tchErro (eixo y)
- Gráfico 'scatter' safra (eixo x) e tchErro (eixo y)
- Gráfico 'scatter' ambProd (eixo x) e tchErro (eixo y)

PARTE 3: Modelagem e Avaliação dos Resultados (Valor Total: 5,0)

Considere o conjunto de dados FA084-P1-Dataset-UCR.csv COM TODAS AS ALTERAÇÕES REALIZADAS nos itens 2.A, 2.B e 2.C (Parte 2).

Para responder as Questões 3.1 a 3.3, as instruções a seguir devem ser atendidas:

- SEMPRE QUE PERTINENTE, utilizar random_state = 2020
- SEMPRE QUE PERTINENTE, transformar variáveis categóricas em numéricas
  OBS: Você deve escolher que método utilizar: One-Hot-Encode ou Label Encode
- SEMPRE QUE PERTINENTE, normalizar as variáveis numéricas (você escolhe o tipo de normalização)
- SEMPRE OTIMIZAR os hiperparâmetros
  OBS: Você deve escolher como otimizar (GridSearchCV ou RandomizedSearchCV) e quais parâmetros otimizar, indicando a faixa (range) escolhida e o valor ótimo obtido.

FA084 – Introdução à Mineração de Dados: Conceitos e Aplicações na Agricultura

Prova – 10 de Julho de 2020

- SEMPRE CRIAR CONJUNTO DE TREINO e TESTE com PROPORÇÃO 70/30 e ESTRATIFICADO pelo atributo meta.
- SEMPRE AVALIAR o modelo com RMSE e MAE no CONJUNTO DE TESTE.

Questão 3 - Construir modelos de regressão para predizer a Produtividade Real

3.A) ESCOLHA UMA entre as técnicas **KNN**, **Árvore de Decisão** e **Regressão Linear**
- Indicar a técnica escolhida, construir e avaliar o modelo.

3.B) ESCOLHA UMA entre as técnicas '*ensemble'*: **RandomForest**, **GBoost** e **XGBoost**
- Indicar a técnica escolhida, construir e avaliar o modelo.

3.C) ESCOLHA UMA entre as técnicas **SVM** e **Redes Neurais**
- Indicar a técnica escolhida, construir e avaliar o modelo.

- Construir uma tabela (DataFrame) com a estrutura abaixo:

	KNN ou Árvore de Decisão ou Regressão Linear	Random Forest ou GBoost ou XGBoost	SVM ou Redes Neurais	Erro da Estimativa do Conjunto de Dados (tchEst)
RMSE				
MAE				

1,0 3.1) Fazer uma **ANÁLISE CRÍTICA** dos resultados, com base na tabela

2,0 3.2) Construir quatro gráficos (2 linhas x 2 colunas) utilizando as mesmas escalas para os eixos x e y, com os histogramas dos erros dos modelos dos itens 3.A, 3.B e 3.C e dos erros das estimativas do conjunto de dados.

Fazer uma **ANÁLISE CRÍTICA** dos resultados, com base nos gráficos.

2,0 3.3) Construir quatro gráficos 'scatter' (2 linhas x 2 colunas) utilizando as mesmas escalas para os eixos x e y, tendo no **"eixo x"** o **tchReal** e no **"eixo y"**, os erros dos modelos escolhidos acima e os erros das estimativas do conjunto de dados

Fazer uma **ANÁLISE CRÍTICA** dos resultados, com base nos gráficos.