LISTA 4 – Pré-processamento: tratar os dados (normalização e redução de dimensionalidade)

Exercício 1 – Cartão de Crédito

Disponibilizei um arquivo chamado **CC_DATASET.csv**:

- CUST_ID
- BALANCE
- BALANCE_FREQUENCY
- PURCHASES
- ONEOFF_PURCHASES
- INSTALLMENTS_PURCHASES
- CASH_ADVANCE
- PURCHASES_FREQUENCY
- ONEOFF_PURCHASES_FREQUENCY
- PURCHASES_INSTALLMENTS_FREQUENCY
- CASH_ADVANCE_FREQUENCY
- CASH_ADVANCE_TRX
- PURCHASES_TRX
- CREDIT_LIMIT
- PAYMENTS
- MINIMUM_PAYMENTS
- PRC_FULL_PAYMENT
- TENURE

Fonte dos dados: kaggle.com/arjunbhasin2013/ccdata

Há propositalmente algumas dificuldades nos dados:

- Aparentemente as ausências já estão preenchidas com zeros, porém, nem todas...
- Há alguns poucos valores extremos espalhados pelas variáveis do dataset.

Antes de mexer nos dados, escolha algumas variáveis e experimente calcular valores médios. Tome nota dos valores que resultam.

Pré-processamento dos dados

1. Importe os dados;
2. Tratamento dos outliers:
    a. Aqui você é quem vai escolher o método de tratamento (excluir registros? sobrescrever com zeros? sobrescrever com a média ou mediana?). Analise os dados, e sob o seu julgamento, escolha a melhor forma de tratar os outliers.
    b. Ao final, explique porque você tratou usando o método x, y, z...
3. Valores ausentes:
    a. Nesse contexto (cartão de crédito) as ausências devem ser legítimas, ou seja, não existe informação mesmo. Portanto, trate as ausências com zero.
4. Normalização:
    a. O dataset é praticamente todo composto por variáveis numéricas. Aplique a normalização *Standard Scaler* sobre essas variáveis numéricas.

Modelagem

1. Aplique K-means sobre os dados separando em 2 clusters (k=2) e grave o resultado no df original;
2. Aplique K-means novamente, porém, agora k=3 (essa é uma nova clusterização e não deve sobrescrever a clusterização anterior). Grave esse resultado de clusterização no df original também.

Análise de dados:

1. Escolha duas métricas e plote um gráfico de dispersão, separando por cor os pontos da clusterização k=2.
2. Repita o gráfico da questão anterior para k3. Visualmente qual a clusterização ficou melhor?
3. Quantos casos ficaram enquadrados em cada cluster em cada uma das clusterizações?
4. Qual a média e a mediana de valor de compras, de cada cluster em cada uma das clusterizações?
5. Usando a clusterização k=3, qual a mediana de limite de crédito de cada grupo?

Exercício 2 – Diagnóstico de Câncer

Disponibilizei um dataset clássico: *Breast Cancer Wisconsin*. Nele há informações de diagnósticos de câncer de mama (arquivo **DIAG_CANCER.csv**).

O arquivo possui um layout adequado para tarefas de classificação, contendo uma variável de desfecho (*"diagnosis"*) que informa se o câncer foi diagnosticado como benigno (B) ou maligno (M). As demais variáveis são medidas discretas/contínuas que explicam o desfecho (*"diagnosis"*). Abaixo a lista completa de variáveis:

- id
- diagnosis
- radius_mean
- texture_mean
- perimeter_mean
- area_mean
- smoothness_mean
- compactness_mean
- concavity_mean
- concave points_mean
- symmetry_mean
- fractal_dimension_mean
- radius_se
- texture_se
- perimeter_se
- area_se
- smoothness_se
- compactness_se
- concavity_se
- concave points_se
- symmetry_se
- fractal_dimension_se
- radius_worst
- texture_worst
- perimeter_worst
- area_worst
- smoothness_worst
- compactness_worst

- concavity_worst
- concave points_worst
- symmetry_worst
- fractal_dimension_worst
- Unnamed: 32

Fonte dos dados: kaggle.com/uciml/breast-cancer-wisconsin-data

Pré-processamento dos dados

1. Importe os dados;
2. Tratamento dos outliers:
    a. Temos outliers apenas na extremidade superior da variável "radius_mean". Trate esses valores extremos substituindo-os pela mediana da respectiva variável.
3. Valores ausentes:
    a. A última variável do dataset (Unnamed: 32) só possui valores ausentes. Exclua essa variável.
4. Normalização:
    a. Com exceção dos campos *id* e *diagnosis* o dataset só possui variáveis numéricas. Aplique a normalização *Standard Scaler* sobre essas variáveis numéricas.
5. Redução de Dimensionalidade:
    a. Aplique PCA sobre os dados normalizados para obter 5 variáveis (PCA_1, PCA_2, PCA_3, PCA_4 e PCA_5).
    b. Crie um novo df com as variáveis resultantes do PCA (5) e adicione nele a variável de desfecho (*diagnosis*).

Modelagem

1. Crie um modelo de classificação (árvore aleatória, regressão logística, a sua escolha...), aplique sobre os dados originais (dataframe completo, com as 30 variáveis explicativas). Tome nota da acurácia do modelo e os resultados de matriz de confusão.
2. Repita o processo (modelagem), só que dessa vez treine e teste o modelo usando o df PCA que você criou anteriormente (5 variáveis explicativas + variável de desfecho). Tome nota da acurácia do modelo e os resultados de matriz de confusão.
3. Nesse exercício reduzimos a complexidade computacional na ordem de 6x (de 30 variáveis explicativas para apenas 5). Na sua avaliação observando os resultados de treino/teste, de ambos os modelos, perdemos acurácia? Se sim, quanto?