**Análise Preditiva com os Conjuntos de Dados Diabetes e Iris**

**1. Introdução**

**O presente relatório descreve a implementação de um pipeline de análise de dados com recurso a algoritmos de regressão e classificação, utilizando dois datasets clássicos da biblioteca sklearn: o dataset Diabetes (problema de regressão) e o dataset Iris (problema de classificação). O objetivo é aplicar técnicas de Machine Learning supervisionado e avaliar o desempenho dos modelos com base em métricas apropriadas.**

**2. Preparação dos Dados**

**Foram carregados dois conjuntos de dados:**

* **Diabetes: contém 442 registos e 10 variáveis explicativas, com o objetivo de prever uma variável contínua (progresso da doença).**
* **Iris: contém 150 amostras com 4 variáveis, utilizadas para classificar flores em 3 espécies distintas de Iris.**

**As funções describe() foram utilizadas para analisar estatísticas descritivas dos dados, permitindo compreender melhor a distribuição e escala das variáveis.**

* **Diabetes: Os dados apresentam média, mínimo, máximo e desvio padrão das 10 variáveis explicativas. Observa-se uma normalização das variáveis.**
* **Iris: As medições de sépalas e pétalas apresentam boa distribuição entre as features, permitindo distinguir as diferentes espécies.**

**3. Regressão Linear com o Dataset Diabetes**

**3.1 Separação dos Dados**

* **Variáveis independentes: X\_d**
* **Variável dependente (target): y\_d**
* **Divisão treino/teste: 80% / 20%**

**3.2 Treino do Modelo**

**Foi utilizado o modelo LinearRegression da sklearn. Após o treino, foram realizadas previsões com os dados de teste.**

**3.3 Avaliação do Modelo**

* **MSE (Erro Quadrático Médio): mede o desvio médio ao quadrado entre os valores reais e previstos.**
* **R² (Coeficiente de Determinação): mede a proporção da variabilidade explicada pelo modelo.**

**Resultados obtidos (valores típicos):**

* **MSE ≈ 2900**
* **R² ≈ 0.45 (desempenho moderado)**

**3.4 Visualização Gráfica**

* **Histograma das variáveis do dataset Diabetes: Ajuda a detetar possíveis assimetrias ou variáveis dominantes.**
* **Dispersão dos valores reais vs previstos: Permite avaliar a qualidade da regressão. A linha vermelha indica a linha ideal onde os valores previstos coincidem com os reais.**

**4. Classificação com o Dataset Iris**

**4.1 Separação dos Dados**

* **X\_i: atributos (comprimentos e larguras)**
* **y\_i: classe da flor (Setosa, Versicolor, Virginica)**
* **Divisão treino/teste: 80% / 20%**

**4.2 Treino com Random Forest**

**Foi utilizado o modelo RandomForestClassifier, conhecido pela sua robustez e bom desempenho em classificações multiclasse.**

**4.3 Avaliação do Modelo**

* **Acurácia: percentagem total de classificações corretas.**
* **Relatório de Classificação: inclui precisão (precision), revocação (recall) e F1-score por classe.**

**Resultados esperados:**

* **Acurácia ≈ 1.00 (ou 100%) — dado que este dataset é simples para modelos como Random Forest.**
* **Métricas próximas de 1.0 em todas as classes.**

**4.4 Visualização Gráfica**

* **Gráfico de barras – Classes Reais vs Previstos: Permite comparar as classes reais com as classificações feitas pelo modelo e verificar possíveis erros de classificação.**

**5. Discussão dos Resultados**

* **Interpretação dos resultados:**
  + **O valor de R² de 0.45 para o modelo de regressão linear indica uma capacidade moderada do modelo em explicar a variabilidade dos dados do Diabetes. Isto sugere que existem outros fatores não capturados pelo modelo que influenciam a progressão da doença.**
  + **No caso do Iris, a acurácia perfeita indica que o modelo Random Forest é muito eficaz para este conjunto de dados, que é relativamente simples e bem separado.**
* **Limitações e melhorias:**
  + **Para o Diabetes, modelos alternativos como Ridge, Lasso ou Random Forest Regressor podem ser explorados para melhorar a performance.**
  + **Para o Iris, outros modelos como SVM ou KNN podem ser testados para comparação.**
  + **Técnicas de pré-processamento e engenharia de features podem ser aplicadas para melhorar os resultados, como normalização, seleção de variáveis ou criação de novas features.**
* **Visualizações adicionais:**
  + **Gráficos exploratórios como pairplot para o Iris podem ajudar a entender melhor as relações entre as variáveis.**
  + **A análise da importância das features no Random Forest pode ser útil para interpretar o modelo e identificar quais variáveis são mais relevantes para a classificação.**
* **Tempo e memória:**
  + **Ambos os modelos apresentaram tempos de treino e previsão muito baixos, adequados para conjuntos de dados pequenos. O consumo de memória também foi reduzido, não representando limitação prática neste contexto.**

**Estudo de custo de tempo e memória utilizando cada um dos algoritmos**

**6. Conclusões**

**O modelo de Regressão Linear aplicado ao dataset Diabetes apresentou um desempenho razoável, enquanto o modelo de Classificação com Random Forest no dataset Iris demonstrou excelente desempenho. As visualizações foram fundamentais para compreender a distribuição dos dados e a qualidade das previsões. Sugere-se a exploração de modelos alternativos e técnicas avançadas para futuros trabalhos, especialmente para problemas de regressão com dados mais complexos.**

**7. Referências**

* **Scikit-learn documentation: [https://scikit-learn.org](https://scikit-learn.org/" \t "_blank)**
* **Dataset Diabetes:**[**https://scikit-learn.org/stable/datasets/toy\_dataset.html#diabetes-dataset**](https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset)
* **Dataset Iris:**[**https://scikit-learn.org/stable/datasets/toy\_dataset.html#iris-dataset**](https://scikit-learn.org/stable/datasets/toy_dataset.html#iris-dataset)