# **Relatório do Trabalho Prático 3**

# **Resumo**

Este trabalho prático teve como objetivo aplicar técnicas de Machine Learning a dois conjuntos de dados disponibilizados pelo scikit-learn: o conjunto Iris (para classificação) e o conjunto Diabetes (para regressão).

Foram treinados modelos distintos para cada conjunto, com o intuito de identificar padrões nos dados e avaliar o desempenho dos modelos em termos de eficácia, tempo de execução e uso de memória.

1. **Descrição dos Conjuntos de Dados**

Neste trabalho prático, foram abordados dois problemas distintos de aprendizagem supervisionada, utilizando conjuntos de dados de referência fornecidos pela biblioteca *scikit-learn*. Cada conjunto de dados foi associado a um tipo específico de tarefa: classificação no caso do dataset Iris e regressão no caso do dataset Diabetes.

* **Iris**: O dataset Iris é amplamente utilizado em contextos de aprendizagem supervisionada para a resolução de problemas de classificação multiclasse. Contém 150 amostras de flores da planta *Iris*, divididas em três espécies distintas: *setosa*, *versicolor* e *virginica*.  
  Cada amostra é descrita por quatro atributos morfológicos: comprimento e largura da sépala, e comprimento e largura da pétala, medidos em centímetros.

O objetivo consiste em treinar um modelo capaz de prever corretamente a espécie da flor com base nessas características. A simplicidade e separabilidade dos dados tornam este conjunto ideal para testar algoritmos de classificação.

* **Diabetes**: O dataset Diabetes é um conjunto de dados clínicos amplamente utilizado para tarefas de regressão. Cada instância representa informações biométricas e clínicas de um paciente, e está associada a uma variável alvo contínua, relacionada com a progressão da diabetes.

Este dataset inclui atributos como:

* + Colesterol total
  + Índice de massa corporal (IMC)
  + Pressão arterial média
  + Idade
  + Tempo desde o diagnóstico (ltg)

1. **Descrição de cada um dos Algoritmos (modelos) Utilizados** 
   1. **Decision Tree**

É um método de aprendizagem supervisionada usado para gerar árvores de decisão, bastante popular em tarefas de **classificação**. Este algoritmo baseia-se na **Entropia** e no **Ganho de Informação** para determinar o atributo mais relevante em cada etapa da construção da árvore.

**Funcionamento:**

* + - **Entropia** mede a incerteza ou impureza de um conjunto de dados.
    - O **Ganho de Informação** calcula a redução dessa incerteza ao dividir os dados com base num atributo.
    - A árvore cresce recursivamente, escolhendo em cada nó o atributo com maior ganho informativo.
    - A árvore para de crescer quando todos os exemplos são da mesma classe ou quando não há mais atributos a considerar.
  1. **SVR – Support Vector Regression**

O **SVR** (*Support Vector Regressor*) é uma variação das **Support Vector Machines (SVM)** aplicada a problemas de **regressão**. O seu objetivo é ajustar uma função que tenha no máximo um determinado erro (ε), sendo o mais “plana” possível.

**Características:**

* Utiliza **vetores de suporte**, os pontos mais próximos da margem da função.
* Pode usar **funções kernel** para transformar os dados e aplicar uma regressão não-linear (ex. *RBF*).
* Maximiza a margem de tolerância ao erro, promovendo boa generalização.

1. **Discussão das principais características de cada um dos algoritmos.** 
   1. **Decision Tree**

O algoritmo ID3 constrói árvores de decisão com base no conceito de **ganho de informação**, selecionando, a cada divisão, o atributo que mais reduz a incerteza (medida através da **entropia**) no conjunto de dados. Entre as suas principais características destacam-se:

* Interpretação simples: as árvores geradas podem ser facilmente visualizadas e compreendidas, sendo úteis para explicar decisões.
* Boa performance com dados bem estruturados: funciona eficazmente quando as classes são separáveis com base em regras lógicas simples.
* Propensão ao sobreajustamento (overfitting): sem técnicas como poda, o ID3 pode adaptar-se demasiado aos dados de treino.
* Critério baseado em entropia: toma decisões informadas, ainda que ligeiramente mais custosas computacionalmente que o critério de Gini.
  1. **SVR – Support Vector Regression**

O SVR é um algoritmo de regressão baseado no conceito de **margem máxima** das SVMs (Support Vector Machines). Em vez de prever classes, ajusta uma função contínua que procura manter os desvios das previsões dentro de uma margem aceitável. As suas características mais relevantes são:

* **Capacidade de modelar relações complexas**: ao utilizar *kernels* (como o RBF), consegue capturar padrões não lineares nos dados.Pode usar **funções kernel** para transformar os dados e aplicar uma regressão não-linear (ex. *RBF*).
* **Robustez à variância**: o modelo procura generalizar, considerando apenas os dados mais críticos (vetores de suporte).
* **Requer normalização dos dados**: para um desempenho adequado, as variáveis devem estar na mesma escala.
* **Ideal para problemas com variabilidade gradual:** como é o caso de progressões médicas, como a diabetes.

1. **Estudo de Custo de Tempo e de Memória**

Nesta secção analisam-se os tempos de execução e o consumo de memória registados durante o treino e a fase de previsão para ambos os algoritmos: ID3 (classificação) e SVR (regressão).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algoritmo | Tempo de Treino (s) | Tempo de Previsão (s) | Pico Memória Treino (MB) | Pico Memória Previsão (MB) |
| ID3 | **0.0060** | |  | | --- | |  |  |  | | --- | | **0.0010** | | |  | | --- | |  |  |  | | --- | | **0.03** | | **0.01** |
| SVR | |  | | --- | |  |  |  | | --- | | **0.0050** | | **0.0030** | |  | | --- | |  |  |  | | --- | | **0.09** | | |  | | --- | |  |  |  | | --- | | **0.02** | |

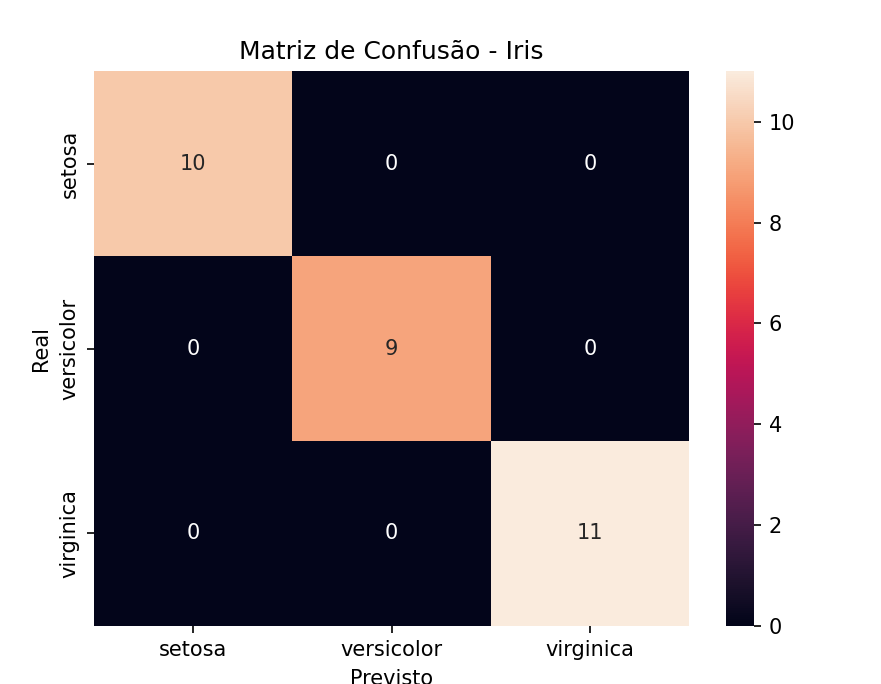
1. **Estudo da performance dos modelos**

A performance dos modelos foi avaliada através de métricas adequadas a cada tarefa.  
Para classificação (ID3), utilizaram-se métricas como **acurácia**, **precisão**, **recall** e **F1-score**.  
Para regressão (SVR), recorreu-se ao **erro quadrático médio (MSE)** e ao **coeficiente de determinação (R²)**.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Algoritmo | Acurácia / R² | Precisão (média) | |  | | --- | | **Recall (média)** |  |  | | --- | |  | | F1-score (média) | MSE |
| ID3 | 1.00 | |  | | --- | |  |  |  | | --- | | 1.00 | | 1.00 | 1.00 | - |
| SVR | 0.18 | - | - | - | |  |  | | --- | --- | |  | 4333.29 | |

1. **Discussão dos Resultados**

* **Tabela de Tempo e Memória:**
  + O **ID3 (Decision Tree)** apresenta uma **execução extremamente rápida**, tanto no treino como na previsão (0.0060s e 0.0010s), com **baixo consumo de memória** (0.03 MB e 0.01 MB). Isso confirma a sua **eficiência computacional**, sobretudo em datasets pequenos e bem estruturados como o Iris.
  + O **SVR (Regressão)** tem um desempenho computacional semelhante em tempo de execução, embora **requeira mais memória**, especialmente durante o treino (0.09 MB). Isto deve-se ao maior custo algorítmico do SVR, que depende de operações com kernels e vetores de suporte.
* **Tabela de Performance:**
  + O ID3 alcançou **valores perfeitos (1.00)** para todas as métricas de classificação, incluindo acurácia, precisão, recall e F1-score, evidenciando um **modelo idealmente ajustado ao problema**.
  + Já o SVR obteve um **R² de apenas 0.18**, o que indica que o modelo **só explica 18% da variância nos dados**. Apesar disso, o **MSE (4333.29)** não é extremo, sugerindo que os erros não são sistematicamente grandes, mas sim que o modelo **não capta bem a tendência geral**.



Esta matriz de confusão demonstra que o modelo ID3 classificou **corretamente todas as amostras** do conjunto de teste. As diagonais têm valores 10, 9 e 11, correspondentes respetivamente às classes *setosa*, *versicolor* e *virginica*, sem qualquer erro de classificação.  
Isto confirma a separabilidade clara entre as classes do dataset Iris, e que o ID3 foi **capaz de aprender perfeitamente** os padrões nos dados de treino.

O gráfico mostra a dispersão entre os **valores reais** (eixo X) e os **valores previstos** (eixo Y) pelo modelo SVR.  
Embora exista alguma **tendência crescente**, nota-se que **muitos pontos se afastam da linha ideal (diagonal)**, indicando **previsões com erro considerável** em várias regiões do conjunto de teste.  
Esta visualização apoia o valor de R² baixo (0.18), evidenciando que o modelo **não conseguiu capturar eficazmente a relação entre variáveis**.Uma imagem com texto, diagrama, file, captura de ecrã

Os conteúdos gerados por IA poderão estar incorretos.

1. **Conclusões**

Neste trabalho usámos dois modelos de Machine Learning: uma **árvore de decisão (ID3)** para classificar flores e um **SVR** para prever valores ligados à diabetes.O modelo ID3 teve **ótimos resultados**, acertando **todas as classificações** no dataset Iris, e usou **pouco tempo e memória**.O modelo SVR teve **resultados mais fracos** no dataset Diabetes, com previsões pouco precisas. Mesmo assim, foi rápido e leve em termos de recursos.No geral, aprendemos que **cada modelo tem pontos fortes e fracos**, e que a escolha depende do tipo de dados e problema.