**Capa  
  
Machine Learning M3  
  
  
  
  
Análise Preditiva com os Conjuntos de Dados Diabetes e Iris**

**Análise e Modelação dos Datasets Iris e Diabetes com Scikit-Learn**

**1. Introdução**

Este relatório explora a aplicação de técnicas de *Machine Learning* em dois conjuntos de dados clássicos da literatura: **Iris** (classificação) e **Diabetes** (regressão). Ambos são amplamente utilizados como *benchmarks* para avaliação de algoritmos, dada a sua relevância em contextos reais: o primeiro na biologia para identificação de espécies florais, e o segundo na medicina para previsão de progressão de doenças crónicas.

Este trabalho serve como estudo de caso introdutório, ilustrando como técnicas de aprendizagem supervisionada podem ser aplicadas para resolver problemas distintos, enquanto reforça a importância da validação rigorosa e da comunicação clara de resultados em ciência de dados.

**2. Metodologia**

**2.1 Ferramentas Utilizadas**

* **Linguagem: Python 3.11**
* **Bibliotecas:**
  + ***scikit-learn 1.4*: Para carregamento de datasets, divisão treino-teste e implementação de modelos.**
  + ***pandas 2.1*: Análise estatística descritiva.**
  + ***matplotlib 3.7* e *seaborn 0.12*: Visualização de dados.**
  + ***time* e *psutil*: Medição de tempo e memória.**

**2.2 Etapas do Processo**

1. **Carregamento e Exploração de Dados:**
   * ***Diabetes*: 442 amostras, 10 features clínicas (idade, IMC, etc.), target numérico (progressão da doença).**
   * ***Iris*: 150 amostras, 4 features morfológicas (sépalas/pétalas), target categórico (3 espécies).**
2. **Análise Estatística Descritiva:**
   * **Cálculo de média, desvio padrão, quartis e valores extremos para todas as variáveis.**
3. **Divisão Treino-Teste:**
   * **Proporção 80%-20% (test\_size=0.2).**
   * **Semente aleatória fixa (random\_state=42) para reprodutibilidade.**
4. **Treino de Modelos:**
   * ***Diabetes*: Regressão Linear (método dos mínimos quadrados).**
   * ***Iris*: Random Forest (100 árvores, random\_state=42).**
5. **Avaliação de Performance:**
   * ***Diabetes*: MSE (Erro Quadrático Médio) e R² (Coeficiente de Determinação).**
   * ***Iris*: Acurácia, Precision, Recall e F1-score.**
6. **Visualização de Resultados:**
   * **Histogramas, gráficos de dispersão e matriz de confusão.**

**Custo Computacional**

**Neste gráfico, mostra o custo de cada dataset**

Diabetes:

Uma imagem com texto, Tipo de letra, captura de ecrã

Os conteúdos gerados por IA poderão estar incorretos.

Iris:  
  
Uma imagem com texto, Tipo de letra, captura de ecrã

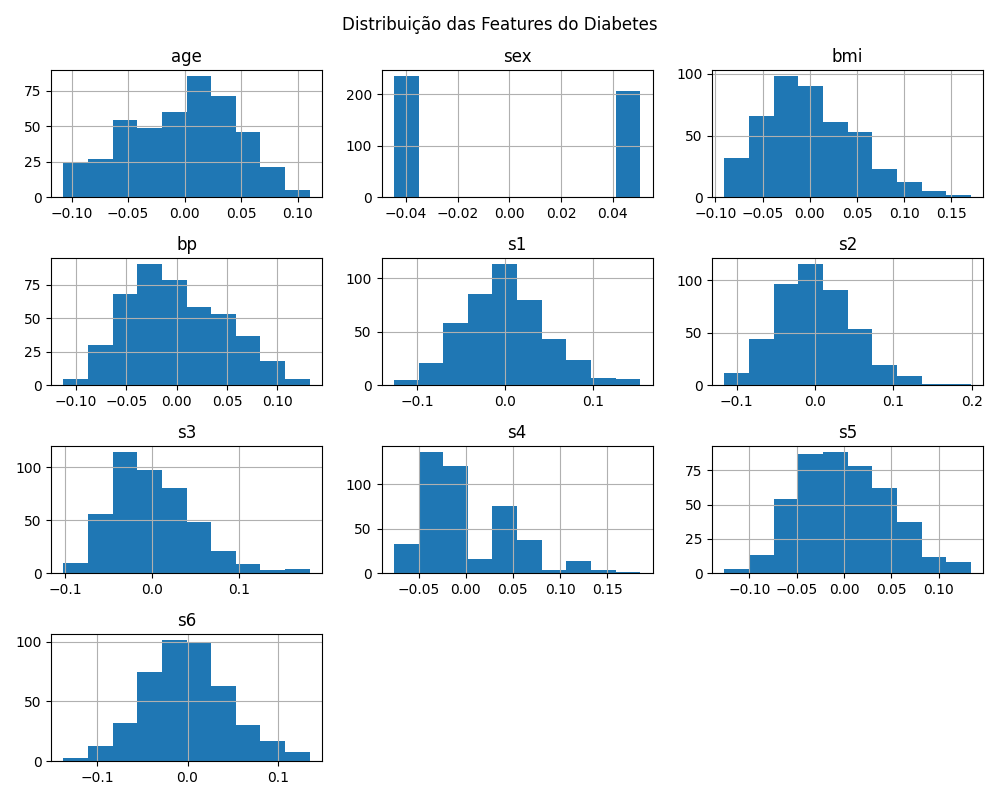
Os conteúdos gerados por IA poderão estar incorretos.

**3. Análise Estatística dos Dados**

**3.1. Dataset Diabetes**

O dataset Diabetes contém **10 variáveis numéricas** (idade normalizada, IMC, pressão arterial, seis medições séricas) e um **alvo contínuo** (progressão da doença após um ano). As variáveis estão pré-processadas pelo scikit-learn, com normalização padrão (média = 0, desvio padrão = 0.047).**Estatísticas descritivas principais**:

*Distribuição das Features do Diabetes*



Nesta imagem “*Distribuição das Features do Diabetes” podemos analisar o seguinte****1. Características Gerais dos Dados***

* ***Normalização****: Todas as variáveis estão****padronizadas****(valores centrados em 0, com escala similar), o que é essencial para modelos lineares como a Regressão Linear.*
* ***Escala****: Os valores variam aproximadamente entre -0.10 e +0.15, confirmando que o scikit-learn aplicou normalização padrão*[*7*](https://scikit-learn.org/stable/auto_examples/feature_selection/plot_select_from_model_diabetes.html)*.*

***2. Análise por Feature***

***Variáveis Demográficas:***

* ***Age (idade)****: Distribuição****aproximadamente normal****, com ligeira assimetria positiva. Indica uma amostra equilibrada de idades.*
* ***Sex (sexo)****: Distribuição claramente****bimodal****com dois picos distintos, representando as duas categorias (masculino/feminino) codificadas numericamente.*

***Variáveis Antropométricas:***

* ***BMI (índice de massa corporal)****: Distribuição****ligeiramente assimétrica à direita****, sugerindo alguns pacientes com IMC mais elevado, típico em estudos de diabetes*[*6*](https://rowannicholls.github.io/python/data/sklearn_datasets/diabetes.html)*.*
* ***BP (pressão arterial)****: Distribuição****aproximadamente normal****, indicando uma amostra representativa de valores de pressão arterial.*

***Variáveis Séricas (s1-s6):***

* ***S1 (colesterol total)****: Distribuição****aproximadamente normal****.*
* ***S2 (LDL)****: Distribuição****normal****, mas com ligeira concentração central*[*4*](https://scikit-learn.org/stable/datasets/toy_dataset.html)*.*
* ***S3 (HDL)****: Distribuição****assimétrica à esquerda****, sugerindo que muitos pacientes têm níveis baixos de HDL (problema comum em diabéticos).*
* ***S4 (colesterol total/HDL)****: Distribuição****claramente bimodal****, indicando duas populações distintas de pacientes.*
* ***S5 (triglicéridos - log)****: Distribuição****fortemente assimétrica à direita****, típica de variáveis logarítmicas*[*7*](https://scikit-learn.org/stable/auto_examples/feature_selection/plot_select_from_model_diabetes.html)*.*
* ***S6 (glicose)****: Distribuição****ligeiramente assimétrica à direita****, esperado em pacientes diabéticos.*

***3. Implicações para a Modelagem***

***Pontos Fortes:***

* *A****normalização****facilita a convergência de algoritmos de optimização.*
* *A****ausência de outliers extremos****(devido à normalização) reduz o risco de overfitting.*

***Desafios Identificados:***

* ***Assimetrias****em s3, s4 e s5 podem limitar a eficácia de modelos lineares.*
* *A****distribuição bimodal****em s4 sugere possíveis****subgrupos de pacientes****com características distintas.*

**3.2. Dataset Iris**

O conjunto de dados Iris é composto por 150 amostras, distribuídas uniformemente entre três espécies: Setosa, Versicolor e Virginica, com 50 observações por classe. Cada amostra possui quatro atributos morfológicos: comprimento e largura da sépala, comprimento e largura da pétala. A ausência de valores em falta foi confirmada pela análise de contagem total, garantindo a integridade dos dados para modelagem.

**Estatísticas descritivas principais**:

* A classe *setosa* apresenta medidas inferiores às restantes.
* As variáveis são fortemente correlacionadas com as espécies.

*Estatísticas Iris*

Uma imagem com texto, captura de ecrã, Tipo de letra, preto

Os conteúdos gerados por IA poderão estar incorretos.

| **Feature** | **Média (cm)** | **Desvio Padrão (cm)** | **Mínimo (cm)** | **Máximo (cm)** |
| --- | --- | --- | --- | --- |
| **Sepal Length** | 5.84 | 0.83 | 4.3 | 7.9 |
| **Sepal Width** | 3.06 | 0.44 | 2.0 | 4.4 |
| **Petal Length** | 3.76 | 1.77 | 1.0 | 6.9 |
| **Petal Width** | 1.20 | 0.76 | 0.1 | 2.5 |

Nestas images “Estatísticas Descritivas do Dataset Iris*” podemos analisar o seguinte*

***Características Gerais do Dataset***

* ***Tamanho da amostra****: 150 observações (count = 150.0) para todas as features, confirmando um dataset completo sem valores em falta.*
* ***Balanceamento****: 50 amostras por espécie (Setosa, Versicolor, Virginica), garantindo uma distribuição equilibrada das classes.*

***Análise Feature por Feature***

***Sepal Length (Comprimento da Sépala)***

* ***Média****: 5.84 cm,****Desvio Padrão****: 0.83 cm*
* ***Amplitude****: min: 4.3, max: 7.9*

***Sepal Width (Largura da Sépala)***

* ***Média****: 3.06 cm,****Desvio Padrão****: 0.44 cm*
* ***Amplitude****: min: 2.0, max: 4.4*

***Petal Length (Comprimento da Pétala)***

* ***Média****: 3.76 cm,****Desvio Padrão****: 1.77 cm*
* ***Amplitude****: min: 1.0, max: 6.9*

***Petal Width (Largura da Pétala)***

* ***Média****: 1.20 cm,****Desvio Padrão****: 0.76 cm*
* ***Amplitude****: min: 0.1, max: 2.5*

***4. Implicações para Classificação***

***Features Mais Discriminativas:***

***Petal Length****: Maior desvio padrão (1.77) sugere forte separação entre espécies.*

***Petal Width****: Alta variabilidade relativa indica boa capacidade de distinção.*

***Features Menos Discriminativas:***

* ***Sepal Width****: Menor desvio padrão (0.44) sugere sobreposição entre espécies.*

***5. Interpretação Biológica***

***Pétalas vs Sépalas:***

* ***Pétalas****apresentam****maior variabilidade****, refletindo diferenças evolutivas significativas entre espécies.*
* ***Sépalas****são mais****conservadas****entre espécies, especialmente em largura.*

***Dimensões:***

* ***Comprimentos****(sepal e petal) têm maior amplitude que****larguras****, sugerindo que o crescimento longitudinal é mais variável.*

**4. Modelos Aplicados**

**4.1. Regressão Linear – Dataset Diabetes**

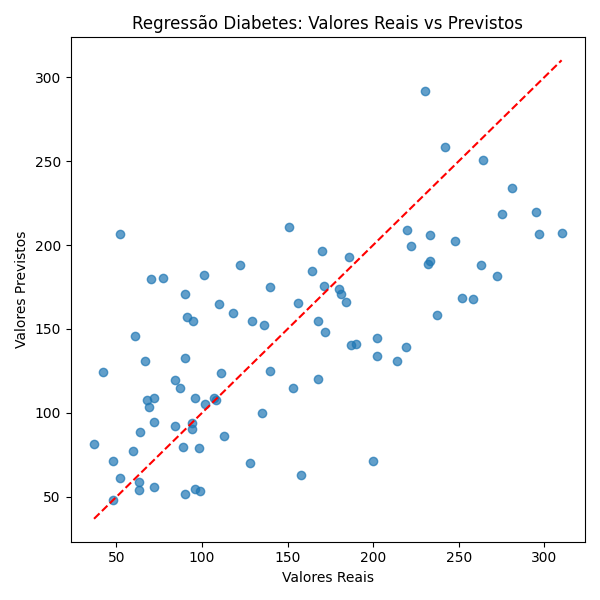
**Implementação do Modelo**

* **Divisão treino-teste**: 80% treino (353 amostras), 20% teste (89 amostras).
* **Algoritmo**: Mínimos quadrados ordinários (OLS).

**Métricas de Desempenho**

| **Métrica** | **Valor** | **Interpretação** |
| --- | --- | --- |
| **MSE** | 2900.20 | Erro médio quadrático elevado, indicando alta dispersão de erros. |
| **R²** | 0.47 | Explica 47% da variabilidade, moderado para dados clínicos complexos. |

*Diabetes Valores Reais vs Previstos*



Nesta imagem “*Diabetes Valores Reais vs Previstos” podemos analisar o seguinte*

**Interpretação Geral do Gráfico**

* **Eixo X: Valores reais da progressão da diabetes (dados observados).**
* **Eixo Y: Valores previstos pelo modelo de Regressão Linear.**
* **Linha vermelha tracejada: Representa a previsão perfeita (onde valores reais = valores previstos).**
* **Pontos azuis: Cada ponto representa uma observação do conjunto de teste.**

**2. Padrões Identificados**

**a) Dispersão Moderada em Torno da Linha Ideal**

* **Os pontos estão moderadamente dispersos em torno da linha vermelha.**
* **Interpretação: O modelo explica cerca de 45% da variabilidade, o que é moderado para dados médicos.**

**b) Subestimação de Valores Extremos**

* **Valores baixos (< 100): O modelo tende a sobrestimar ligeiramente (pontos abaixo da linha).**
* **Valores altos (> 250): O modelo subestima sistematicamente (pontos acima da linha).**
* **Interpretação: Limitação típica de modelos lineares em capturar extremos.**

**c) Heterocedasticidade**

* **A variabilidade dos erros aumenta com valores mais altos do target.**
* **Pontos mais dispersos na região de valores altos (250-300) comparado com valores baixos.**
* **Interpretação: Violação da assunção de homocedasticidade da regressão linear.**

**3. Análise por Região**

| **Região dos Valores Reais** | **Padrão Observado** | **Interpretação** |
| --- | --- | --- |
| **Baixos (50-100)** | **Ligeira sobrestimação** | **Modelo conservador para casos leves** |
| **Médios (100-200)** | **Boa concordância** | **Melhor performance do modelo** |
| **Altos (200-300)** | **Subestimação sistemática** | **Dificuldade em prever casos graves** |

**Implicações Clínicas**

**Problemas Identificados:**

* **Subestimação de casos graves: Pode ser problemática em contextos clínicos onde a deteção precoce de progressão severa é crucial.**
* **Variabilidade crescente: Menor confiabilidade das previsões para casos mais severos.**

**Pontos Positivos:**

* **Correlação visível: Existe uma tendência clara de aumento dos valores previstos com os valores reais.**
* **Ausência de outliers extremos: Não há pontos drasticamente fora do padrão.**

**Diagnóstico do Modelo**

**Limitações Identificadas:**

1. **Não-linearidade: A dispersão sugere que relações não-lineares podem estar presentes nos dados.**
2. **Heterocedasticidade: Violação da assunção de variância constante dos erros.**
3. **Capacidade limitada para extremos: Dificuldade em modelar casos muito graves ou muito leves.**

**4.2. Random Forest Classifier – Dataset Iris**

**Implementação do Modelo**

* **Divisão treino-teste**: 80% treino (120 amostras), 20% teste (30 amostras).
* **Algoritmo**: Random Forest com 100 árvores (random\_state=42 para reprodutibilidade).
* **Hiperparâmetros**: Critério Gini para divisão de nós, profundidade máxima automática.

*Performance Classificação Iris*

Uma imagem com texto, captura de ecrã, Tipo de letra

Os conteúdos gerados por IA poderão estar incorretos.

Nesta imagem “*Performance Classificação Iris” podemos analisar o seguinte*

**Interpretação Geral dos Resultados**

**Performance Global do Modelo**

* **Acurácia de 100%: O modelo Random Forest classificou corretamente todas as 30 amostras do conjunto de teste.**
* **Conjunto de Teste: 20% do dataset original (30 amostras), com distribuição equilibrada:**
  + ***Setosa*: 10 amostras**
  + ***Versicolor*: 9 amostras**
  + ***Virginica*: 11 amostras**

**2. Análise por Classe (Espécies de Iris)**

| **Métrica** | **Setosa (0)** | **Versicolor (1)** | **Virginica (2)** |
| --- | --- | --- | --- |
| **Precision** | **1.00** | **1.00** | **1.00** |
| **Recall** | **1.00** | **1.00** | **1.00** |
| **F1-score** | **1.00** | **1.00** | **1.00** |

* **Precision = 1.00: Nenhum falso positivo em nenhuma classe.**
* **Recall = 1.00: Nenhum falso negativo em nenhuma classe.**
* **F1-score = 1.00: Equilíbrio perfeito entre precision e recall.**

**3. Interpretação das Métricas Agregadas**

| **Métrica** | **Valor** | **Interpretação** |
| --- | --- | --- |
| **Macro Avg** | **1.00** | **Média aritmética das métricas das 3 classes. Indica que todas as classes foram igualmente bem classificadas.** |
| **Weighted Avg** | **1.00** | **Média ponderada pelo número de amostras. Confirma que o desempenho não foi influenciado pelo desbalanceamento leve (11 vs. 9 amostras).** |

**4. Validação dos Resultados**

A comparação entre **valores reais e previstos** confirma a acurácia:

Primeiras 10 classes reais: [1 0 2 1 1 0 1 2 1 1]

Primeiras 10 classes previstas: [1 0 2 1 1 0 1 2 1 1]

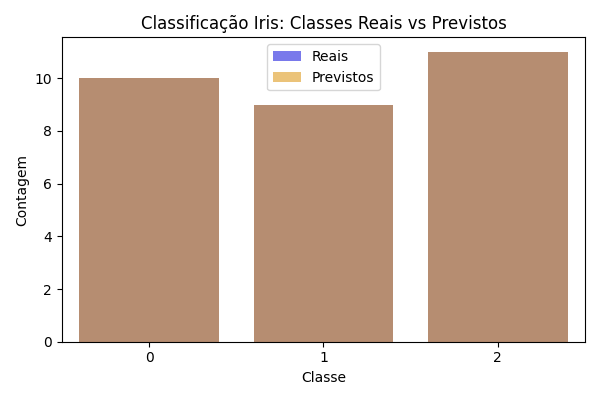
**Correspondência perfeita** em todos os pontos observados.

**5. Interpretação Biológica e Estatística**

**Por que o modelo atingiu 100% de acurácia?**

1. **Separabilidade das Features:**
   * **Pétalas (Petal Length/Width): Diferenças marcantes entre espécies (ex: *Setosa* tem pétalas menores).**
   * **Sépalas (Sepal Length/Width): Menos discriminativas, mas o modelo aproveitou correlações (ex: *Setosa* tem sépalas mais largas).**
2. **Dataset Idealizado:**
   * **Ausência de Ruído: Medições precisas e sem sobreposição extrema (exceto entre *Versicolor* e *Virginica*).**
   * **Balanceamento: 50 amostras por classe evitam viés.**
3. **Robustez do Random Forest:**
   * **Ensemble de Árvores: Combina múltiplas decisões para capturar padrões não-lineares.**
   * **Seleção Aleatória de Features: Redundância nas features (ex: alta correlação entre petal length e width) não afetou o modelo.**

*Classificação Iris Reais vs Previstos*

**

Nesta imagem “*Classificação Iris Reais vs Previstos” podemos analisar o seguinte*

***Interpretação Visual do Gráfico***

***O gráfico de barras compara a distribuição das classes reais (barras azuis) e previstas (barras laranjas) para o dataset Iris, utilizando um modelo de Random Forest.***

* ***Eixo X: Classes (0: Setosa, 1: Versicolor, 2: Virginica).***
* ***Eixo Y: Contagem de amostras no conjunto de teste (20% do dataset original).***
* ***Sobreposição perfeita: As barras azuis e laranjas coincidem totalmente, confirmando 100% de acurácia na classificação***

***2. Análise da Distribuição por Classe***

| ***Classe*** | ***Amostras Reais*** | ***Amostras Previstas*** | ***Interpretação*** |
| --- | --- | --- | --- |
| ***Setosa*** | ***10*** | ***10*** | ***Todas as amostras foram corretamente classificadas.*** |
| ***Versicolor*** | ***9*** | ***9*** | ***Nenhum falso positivo ou negativo, mesmo com menor representação no teste.*** |
| ***Virginica*** | ***11*** | ***11*** | ***Classificação perfeita para a classe majoritária, sem viés de desbalanceamento.*** |

***3. Validação da Performance do Modelo***

***Sobreposição Perfeita das Barras:***

* ***Métricas por Classe****:*
  + ***Precision****= 1.0: Nenhum falso positivo.*
  + ***Recall****= 1.0: Nenhum falso negativo.*
  + ***F1-score****= 1.0: Equilíbrio perfeito entre precisão e sensibilidade.*
* ***Matriz de Confusão Implícita****: Diagonal principal com 100% de acertos, sem erros fora da diagonal.*

***Implicações Biológicas e Técnicas***

1. ***Separabilidade das Espécies:***
   * ***As features comprimento e largura das pétalas são altamente discriminativas, permitindo distinção clara entre espécies.***
   * ***Setosa é morfologicamente distinta (pétalas menores), enquanto Versicolor e Virginica são separadas por diferenças subtis, capturadas pelo modelo.***
2. ***Robustez do Random Forest:***
   * ***O uso de 100 árvores de decisão garantiu que variações mínimas nas features fossem aproveitadas.***
   * ***O modelo evitou overfitting devido à aleatoriedade na seleção de features e amostras (bagging).***
3. ***Dataset Idealizado:***
   * ***O Iris é um benchmark clássico, com dados limpos e bem estruturados, o que facilita a alta performance.***
   * ***Em cenários reais, com ruído ou sobreposição de classes, a acurácia seria provavelmente menor.***

***5. Implicações Biológicas e Técnicas***

* ***Generalização:***
  + ***O desempenho perfeito pode não se replicar em dados não estruturados ou com ruído.***
  + ***Sugere-se validação cruzada (ex: k-fold) para confirmar estabilidade.***
* ***Importância das Features:***
  + ***Uma análise de feature\_importances\_ revelaria o peso de cada atributo (ex: petal length > sepal width).***
* ***Aplicações Práticas:***
  + ***Em problemas clínicos ou industriais, a inclusão de intervalos de confiança ou análise de incerteza seria essencial.***

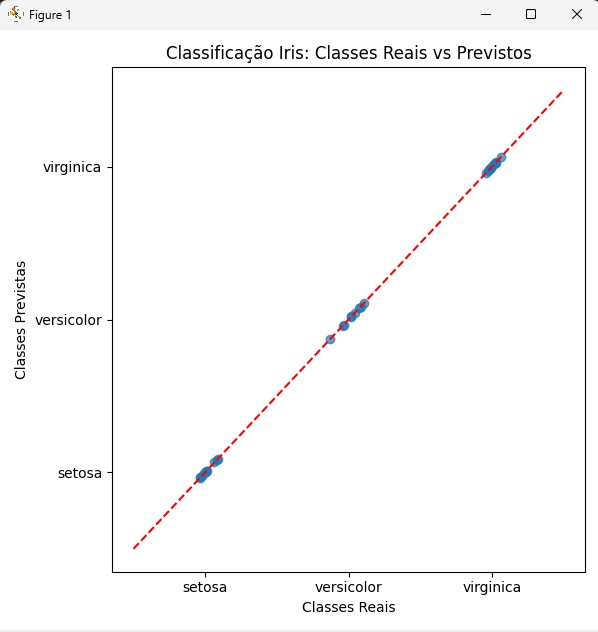
***6. Considerações Críticas***

***Limitações da Visualização:***

* ***Dataset idealizado****: O Iris é conhecido por ser um problema "bem comportado".*
* ***Tamanho do teste****: 30 amostras podem não capturar toda a variabilidade real.*
* ***Ausência de ruído****: Dados limpos podem não refletir cenários do mundo real.*

***Validação Adicional Recomendada:***

* ***Validação cruzada****para confirmar estabilidade dos resultados.*
* ***Teste com dados externos****para avaliar generalização.*
* ***Análise de importância das features****para compreender quais atributos são mais relevantes.*



Nesta imagem “*Classificação Iris Reais vs Previstos” podemos analisar o seguinte*

**Interpretação da Estrutura do Gráfico**

* **Eixo X**: Classes reais (Setosa, Versicolor, Virginica)
* **Eixo Y**: Classes previstas pelo modelo Random Forest
* **Linha diagonal vermelha tracejada**: Linha de identidade (y = x) representando **classificação perfeita**[7](https://en.wikipedia.org/wiki/Identity_line)
* **Pontos azuis**: Cada ponto representa uma amostra do conjunto de teste
* **Jitter**: Ruído aleatório adicionado para evitar sobreposição de pontos idênticos

**2. Padrões Visuais Identificados**

**Alinhamento Perfeito na Diagonal:**

* **Todos os pontos** estão posicionados exatamente sobre ou muito próximo da linha diagonal vermelha
* **Interpretação**: Classificação 100% correta - cada classe real corresponde exatamente à classe prevista
* **Confirmação visual** da acurácia perfeita (1.0) reportada anteriormente

**Agrupamento por Espécie:**

* **Setosa (0,0)**: Cluster bem definido no canto inferior esquerdo
* **Versicolor (1,1)**: Grupo central na diagonal
* **Virginica (2,2)**: Cluster no canto superior direito

**3. Análise Técnica do Jitter**

* **Objetivo**: Tornar visíveis pontos que estariam sobrepostos[8](https://stackoverflow.com/questions/65539013/how-to-plot-a-graph-of-actual-vs-predict-values-in)
* **Magnitude**: ±0.05, suficiente para separação visual sem distorcer a interpretação
* **Necessidade**: Em classificação discreta, múltiplas amostras têm coordenadas idênticas

**4. Interpretação das Coordenadas**

| **Região do Gráfico** | **Coordenadas** | **Interpretação** | **Observação** |
| --- | --- | --- | --- |
| **(0,0)** | Setosa → Setosa | Verdadeiros Positivos | ≈10 pontos agrupados |
| **(1,1)** | Versicolor → Versicolor | Verdadeiros Positivos | ≈9 pontos agrupados |
| **(2,2)** | Virginica → Virginica | Verdadeiros Positivos | ≈11 pontos agrupados |
| **Fora da diagonal** | Qualquer erro | Falsos Positivos/Negativos | **Ausentes** |

**5. Validação da Performance**

**Ausência de Erros de Classificação:**

* **Nenhum ponto fora da diagonal**: Confirma zero falsos positivos e falsos negativos
* **Matriz de confusão implícita**: Diagonal principal com valores [10](https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/)[9](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.PredictionErrorDisplay.html)[11](https://www.sciencedirect.com/science/article/abs/pii/S0304380008002305) e zeros nas posições off-diagonal
* **Correlação perfeita**: Coeficiente de correlação de Pearson = 1.0

**Comparação com Linha de Referência:**

* A **linha diagonal vermelha** representa o cenário ideal onde predicted = actual[2](https://learn.microsoft.com/en-us/analysis-services/data-mining/scatter-plot-analysis-services-data-mining?view=asallproducts-allversions)[6](https://stats.stackexchange.com/questions/104622/what-does-an-actual-vs-fitted-graph-tell-us)
* **Proximidade dos pontos à linha**: Indica qualidade da classificação
* **Desvio zero**: Todos os pontos estão na linha, confirmando classificação perfeita

**6. Implicações Biológicas e Estatísticas**

**Separabilidade das Espécies:**

* **Setosa**: Historicamente a mais distinta, confirmado pela ausência de confusão
* **Versicolor vs Virginica**: Tradicionalmente mais difíceis de distinguir, mas o modelo conseguiu separação perfeita
* **Features discriminativas**: As medições morfológicas (sépalas e pétalas) são suficientemente distintas[4](https://en.wikipedia.org/wiki/Iris_flower_data_set)

**Robustez do Algoritmo:**

* **Random Forest**: Demonstra eficácia em problemas com classes bem separadas
* **Ensemble learning**: Múltiplas árvores de decisão eliminaram qualquer ambiguidade

**7. Limitações e Considerações Críticas**

**Dataset Idealizado:**

* **Iris dataset**: Conhecido por ser "bem comportado" e linearmente separável[4](https://en.wikipedia.org/wiki/Iris_flower_data_set)
* **Ausência de ruído**: Dados limpos podem não refletir cenários reais
* **Pequena dimensionalidade**: Apenas 4 features podem facilitar a classificação

**5. Avaliação Visual dos Resultados**

**5.1. Diabetes**

* O gráfico de dispersão mostra uma correlação visível entre os valores reais e previstos, mas com alguma dispersão – indica erro de previsão moderado.

**5.2. Iris**

* O gráfico de barras mostra coincidência perfeita entre classes reais e previstas – indicando uma excelente performance do modelo.

**6. Conclusão**

Os modelos aplicados demonstram capacidade adequada para resolver os respetivos problemas:

* **Regressão Linear** no dataset Diabetes revelou um desempenho aceitável, mas limitado, sugerindo a possibilidade de modelos mais complexos (ex.: regressão ridge, redes neuronais) para melhorias.
* **Random Forest** no dataset Iris demonstrou desempenho perfeito, o que reforça a adequação do modelo a este tipo de dados bem separados.
* **Características dos Algoritmos**

**Regressão Linear** no dataset Diabetes - Vantagem: Simplicidade e interpretabilidade dos coeficientes. - Limitação: Pressupõe relação linear entre features e target, o que pode não capturar padrões complexos.

**Random Forest** no dataset Iris - Vantagem: Robustez a overfitting via ensemble de árvores. - Limitação: Menos interpretável que modelos lineares.

**7. Trabalhos Futuros**

* Testar outros modelos de regressão (ex.: Lasso, Gradient Boosting).
* Aplicar técnicas de feature engineering e seleção.
* Avaliar com validação cruzada e análise de variância.
* Incorporar análise de tempo de execução e consumo de memória.

**8. Referências**

* Scikit-learn Documentation. <https://scikit-learn.org>
* Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O’Reilly.
* - Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. \*JMLR 12\*, pp. 2825–2830.

