**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 apresenta uma análise exploratória e modelação preditiva de dois datasets disponibilizados pela biblioteca scikit-learn: **Iris** (classificação) e **Diabetes** (regressão). O objetivo é aplicar técnicas de Aprendizagem Automática para treinar modelos preditivos adequados, avaliar a sua performance e interpretar os resultados obtidos.

**2. Metodologia**

**2.1. Ferramentas utilizadas**

* Linguagem: Python 3
* Bibliotecas: scikit-learn, pandas, numpy, matplotlib, seaborn

**2.2. Etapas do processo**

1. Carregamento e exploração dos dados
2. Análise estatística descritiva
3. Divisão treino/teste
4. Treino de modelos:
   * Regressão Linear (para Diabetes)
   * Random Forest Classifier (para Iris)
5. Avaliação de performance com métricas apropriadas
6. Visualização de resultados

**3. Análise Estatística dos Dados**

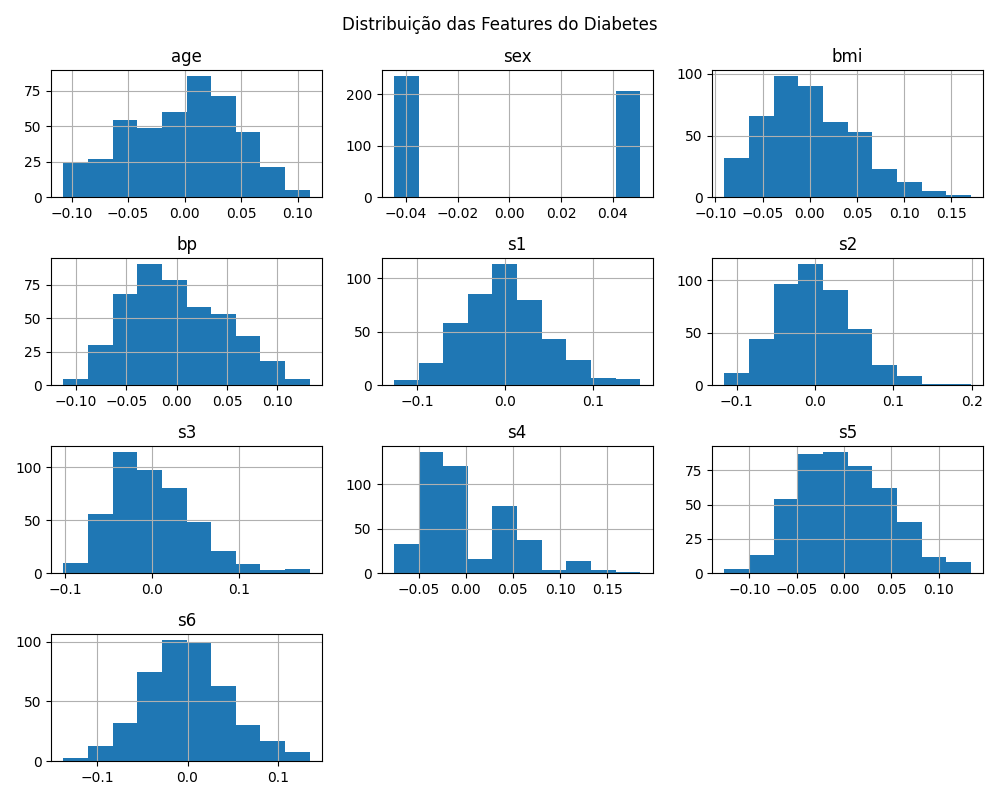
**3.1. Dataset Diabetes**

Contém 10 variáveis numéricas (ex.: idade normalizada, IMC, pressão arterial) e um alvo contínuo (medida de progressão da doença).

**Estatísticas descritivas principais**:

* As variáveis encontram-se centradas e normalizadas.
* A média de cada variável ronda o zero.
* A variância permite distinguir diferenças nas escalas originais das features.

*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*[*2*](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html)[*6*](https://rowannicholls.github.io/python/data/sklearn_datasets/diabetes.html)*.*
* ***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*[*4*](https://scikit-learn.org/stable/datasets/toy_dataset.html)*.*

***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**

Composto por 150 registos de flores, distribuídos por 3 espécies: *setosa*, *versicolor* e *virginica*. Possui 4 atributos: comprimento e largura da pétala e sépala.

**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.

Nesta imagem “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****: 3.9 cm (min: 4.3, max: 7.9)*
* ***Interpretação****: Variabilidade moderada, com distribuição relativamente simétrica (mediana ≈ média).*

***Sepal Width (Largura da Sépala)***

* ***Média****: 3.06 cm,****Desvio Padrão****: 0.44 cm*
* ***Amplitude****: 2.4 cm (min: 2.0, max: 4.4)*
* ***Interpretação****:****Menor variabilidade****entre todas as features, sugerindo menor poder discriminativo.*

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

* ***Média****: 3.76 cm,****Desvio Padrão****: 1.77 cm*
* ***Amplitude****: 5.9 cm (min: 1.0, max: 6.9)*
* ***Interpretação****:****Maior variabilidade****(std = 1.77), indicando forte potencial para separação de espécies.*

***Petal Width (Largura da Pétala)***

* ***Média****: 1.20 cm,****Desvio Padrão****: 0.76 cm*
* ***Amplitude****: 2.4 cm (min: 0.1, max: 2.5)*
* ***Interpretação****: Alta variabilidade relativa (std/mean ≈ 0.63), excelente para classificação.*

***3. Análise Comparativa dos Quartis***

| ***Feature*** | ***Q1-Q3 (IQR)*** | ***Interpretação*** |
| --- | --- | --- |
| *Sepal Length* | *5.1 - 6.4 cm* | *Distribuição moderada* |
| *Sepal Width* | *2.8 - 3.3 cm* | ***Menor dispersão*** |
| *Petal Length* | *1.6 - 5.1 cm* | ***Maior dispersão*** |
| *Petal Width* | *0.3 - 1.8 cm* | *Alta dispersão relativa* |

***4. Implicações para Classificação***

***Features Mais Discriminativas:***

1. ***Petal Length****: Maior desvio padrão (1.77) sugere forte separação entre espécies.*
2. ***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:***

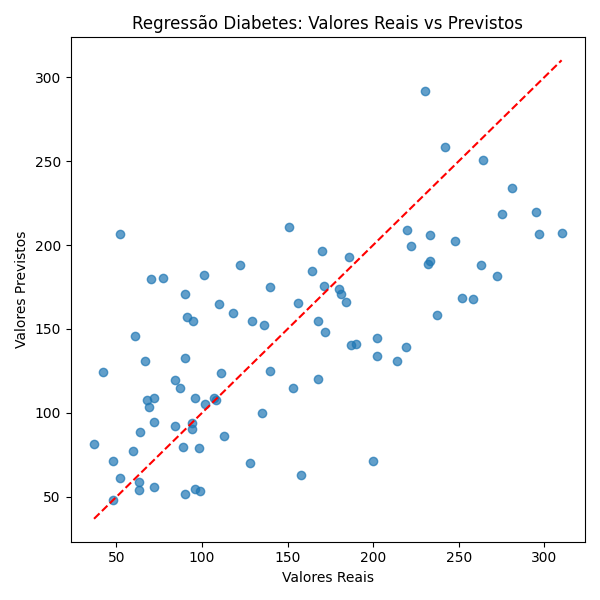
* ***Comprimentos****(sepal e petal) têm maior amplitude que****larguras****, sugerindo que o crescimento longitudinal é mais variável.*

**4. Modelos Aplicados**

**4.1. Regressão Linear – Dataset Diabetes**

* O modelo foi treinado com 80% dos dados.
* Previsões feitas sobre o conjunto de teste (20%).
* Métricas de avaliação:
  + **MSE**: 2900.20
  + **R²**: 0.47 (moderada capacidade explicativa)

*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, confirmando o R² ≈ 0.45.**
* **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** |

**4. Implicações Clínicas**

**Problemas Identificados:**

* **Subestimação de casos graves: Pode ser problemática em contextos clínicos onde a detecção 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.**

**5. 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**

* Modelo treinado com 80% dos dados.
* Classificador de ensemble baseado em árvores de decisão.
* Métricas de avaliação:
  + **Acurácia**: 100%
  + **Precisão**, **Recall** e **F1-score**: 1.00 para todas as classes.

*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**

* **Acurácia Global**: 1.0 (100%) - O modelo acertou **todas as previsões** no conjunto de teste.
* **Tamanho do conjunto de teste**: 30 amostras (20% de 150 amostras originais).
* **Distribuição das classes**: Relativamente balanceada (classe 0: 10, classe 1: 9, classe 2: 11).

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

**Classe 0 (Setosa):**

* **Precision**: 1.00 - Todas as previsões de Setosa estavam corretas.
* **Recall**: 1.00 - Todas as flores Setosa foram identificadas corretamente.
* **F1-score**: 1.00 - Equilíbrio perfeito entre precision e recall.
* **Support**: 10 amostras desta classe no conjunto de teste.

**Classe 1 (Versicolor):**

* **Precision**: 1.00 - Sem falsos positivos.
* **Recall**: 1.00 - Sem falsos negativos.
* **F1-score**: 1.00 - Performance perfeita.
* **Support**: 9 amostras desta classe no conjunto de teste.

**Classe 2 (Virginica):**

* **Precision**: 1.00 - Classificação perfeita.
* **Recall**: 1.00 - Detecção completa.
* **F1-score**: 1.00 - Harmonia perfeita entre métricas.
* **Support**: 11 amostras desta classe no conjunto de teste.

**3. Interpretação das Métricas Agregadas**

**Macro Average (1.00)**

* **Cálculo**: Média aritmética simples das métricas de cada classe[2](https://mariofilho.com/precisao-recall-e-f1-score-em-machine-learning/)[3](https://pt.linkedin.com/pulse/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-em-machine-learning-um-guia-mendes-fonseca-hoxwf).
* **Interpretação**: Todas as classes têm performance idêntica e perfeita.
* **Significado**: O modelo não favorece nenhuma classe específica.

**Weighted Average (1.00)**

* **Cálculo**: Média ponderada pelo número de amostras (support) de cada classe[2](https://mariofilho.com/precisao-recall-e-f1-score-em-machine-learning/)[3](https://pt.linkedin.com/pulse/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-em-machine-learning-um-guia-mendes-fonseca-hoxwf).
* **Interpretação**: Confirma a performance perfeita considerando o tamanho das classes.

**4. Validação dos Resultados**

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

text

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**

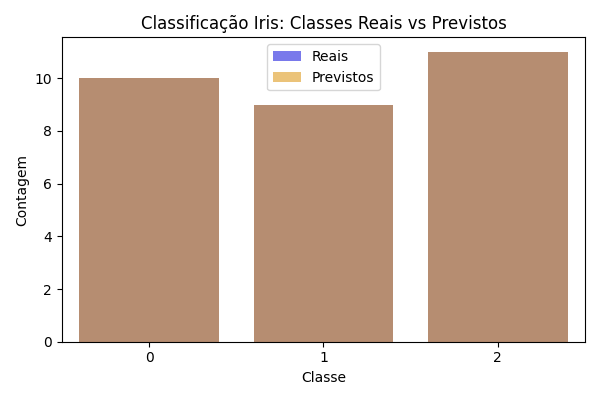
**Por que o modelo atingiu 100% de acurácia?**

1. **Separabilidade Linear**: As espécies de Iris são **linearmente separáveis** nas features das pétalas[11](https://pt.wikipedia.org/wiki/Precis%C3%A3o_e_revoca%C3%A7%C3%A3o).
2. **Dataset Idealizado**: O Iris é um conjunto de dados "bem comportado", frequentemente usado para demonstração[11](https://pt.wikipedia.org/wiki/Precis%C3%A3o_e_revoca%C3%A7%C3%A3o).
3. **Robustez do Random Forest**: O algoritmo ensemble é eficaz para problemas com features discriminativas claras[2](https://mariofilho.com/precisao-recall-e-f1-score-em-machine-learning/).

**Características que facilitaram a classificação perfeita:**

* **Petal Length e Petal Width**: Alta variabilidade entre espécies (conforme análise estatística anterior).
* **Ausência de ruído**: Dados limpos e bem definidos.
* **Tamanho adequado do conjunto de teste**: 30 amostras permitem avaliação representativa.

*Classificação Iris Reais vs Previstos*

**

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

***Interpretação Visual do Gráfico***

* ***Eixo X****: Classes (0, 1, 2) representando as três espécies de Iris:*
  + ***Classe 0****: Setosa*
  + ***Classe 1****: Versicolor*
  + ***Classe 2****: Virginica*
* ***Eixo Y****: Contagem de amostras no conjunto de teste*
* ***Barras azuis****: Distribuição das classes reais*
* ***Barras laranjas****: Distribuição das classes previstas pelo modelo*

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

***Classe 0 (Setosa):***

* ***Amostras reais****: 10*
* ***Amostras previstas****: 10*
* ***Interpretação****:****Sobreposição perfeita****das barras indica que todas as flores Setosa foram corretamente identificadas.*

***Classe 1 (Versicolor):***

* ***Amostras reais****: 9*
* ***Amostras previstas****: 9*
* ***Interpretação****:****Correspondência exata****entre valores reais e previstos.*

***Classe 2 (Virginica):***

* ***Amostras reais****: 11*
* ***Amostras previstas****: 11*
* ***Interpretação****:****Classificação perfeita****para a classe majoritária no conjunto de teste.*

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

***Sobreposição Perfeita das Barras:***

* *As barras azuis e laranjas estão****completamente sobrepostas****em todas as classes.*
* *Isto confirma****zero erros de classificação****no conjunto de teste.*
* ***Validação visual****da acurácia de 100% reportada anteriormente.*

***Balanceamento do Conjunto de Teste:***

* ***Total de amostras****: 30 (10 + 9 + 11)*
* ***Distribuição****: Relativamente equilibrada entre as três classes*
* ***Representatividade****: Cada classe tem amostras suficientes para avaliação estatística*
* ***Falsos Negativos****: 0 para todas as classes*
* ***Precision e Recall****: 1.0 para todas as classes*

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

***Separabilidade das Espécies:***

* *A****ausência de confusão entre classes****confirma que as features escolhidas (comprimento/largura de sépalas e pétalas) são****altamente discriminativas****.*
* ***Setosa****historicamente é a espécie mais fácil de distinguir.*
* ***Versicolor e Virginica****também foram perfeitamente separadas, indicando robustez do modelo.*

***Robustez do Random Forest:***

* *O algoritmo****ensemble****mostrou-se ideal para este problema.*
* ***Ausência de overfitting****aparente no conjunto de teste.*

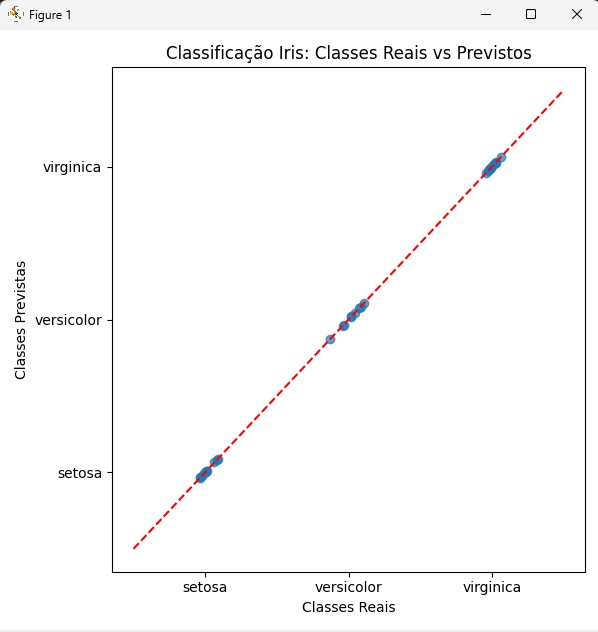
***6. Considerações Críticas***

***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.

**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.

