INSTITUTO INFNET

Faculdade de Engenharia e Desenvolvimento de Software

Dados para Machine Learning: Feature Engineering

[25E2_2]

TP1

Aluno: JEAN MICHAEL ESTEVEZ ALVAREZ

E-mail: jean.alvarez@al.infnet.edu.br

Matrícula:

Professor: Ricardo Mesquita

Distrito Federal

Abril, 2025

SUMÁRIO

1 EXERCÍCIOS

Todas as implementações realizadas para avaliar e entender o funcionamento de cada processo estão no Github, link disponível no Capitulo 2

1. **Examine o dataset Palmer Penguins e explique o que são 'features' no contexto deste dataset específico. Discuta como as features influenciam o desempenho de um modelo de Machine Learning:**

R: O dataset Palmer Penguins consiste em medições biológicas de três espécies de pinguins coletadas no arquipélago Palmer, na Antártica.

(fonte: https://www.tensorflow.org/datasets/catalog/penguins?hl=pt-br)

No contexto de Machine Learning, as features (ou características) são as variáveis de entrada do modelo, ou seja, os atributos associados a cada pinguim. Cada indivíduo no conjunto de dados possui um conjunto específico de valores para essas features, e é com base nessas informações que o modelo pode aprender a predizer um valor de saída, como a espécie de um novo pinguim.

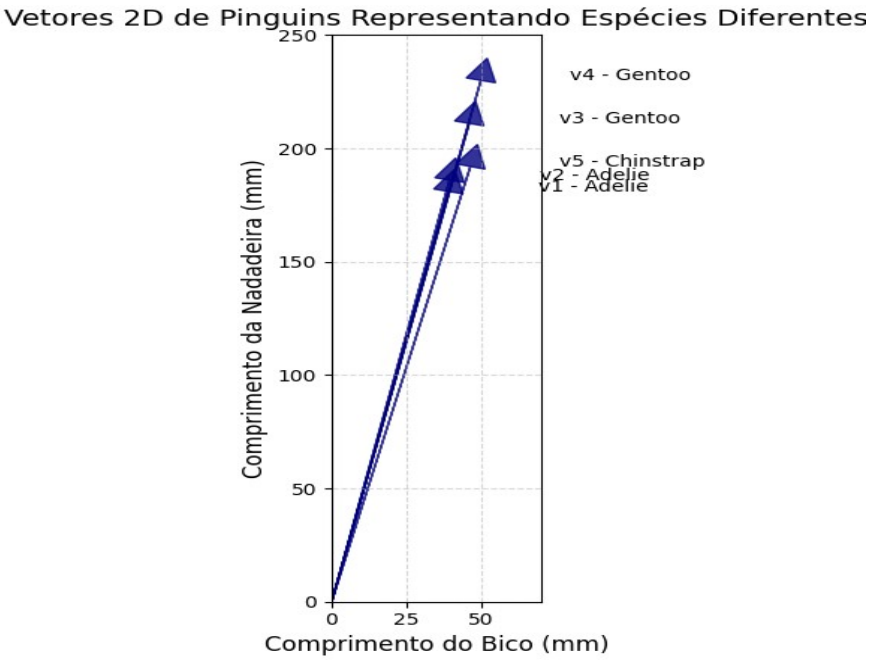
Nem todas as features têm o mesmo impacto na predição. Por exemplo, se o objetivo for classificar a espécie do pinguim, características como o comprimento do bico, profundidade do bico e comprimento das nadadeiras tendem a ser mais relevantes, pois diferenciam bem entre as espécies.

Por outro lado, atributos como o sexo têm menor poder discriminativo nesse caso, pois machos e fêmeas existem em todas as espécies representadas no conjunto.

Assim, entender quais features são mais influentes no desempenho do modelo é essencial para construir algoritmos mais eficientes e interpretáveis.

2. **Com base nos dados do dataset Palmer Penguins, identifique exemplos de escalares, vetores e explique o conceito de espaços em Machine Learning.**

R: No dataset Palmer Penguins, cada medida individual como o comprimento do bico (ex: 45.3 mm) é um escalar — um valor numérico simples. Quando agrupamos várias dessas medições Exemplo : comprimento do bico, nadadeira formamos um vetor de características que representa um pinguim.



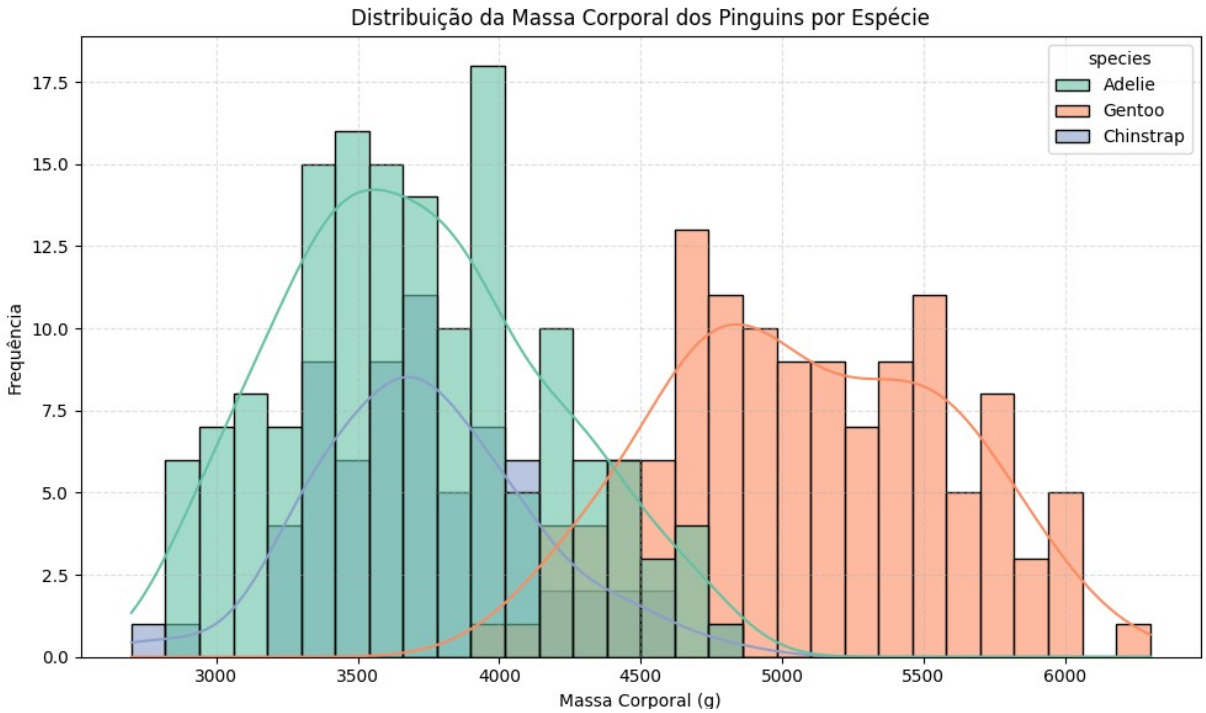
Podemos aumentar a dimensão deste vetor somando outras caracteristicas, como massa, etc. Todos esses vetores juntos compõem um espaço vetorial, onde cada dimensão é uma feature. Nesse espaço, os algoritmos de Machine Learning comparam distâncias e aprendem padrões para classificar, por exemplo, a espécie de um novo pinguim com base na posição do vetor correspondente.

3. **Utilize a técnica de quantização com bins fixos para discretizar uma variável contínua do dataset Palmer Penguins. Explique a razão pela qual você escolheu essa variável e como a discretização pode afetar a análise.**

R: A técnica de quantização com bins fixos (ou uniform binning) é um método de discretização de variáveis contínuas, que consiste em dividir o intervalo dos dados em faixas (bins) de largura constante, transformando valores numéricos contínuos em categorias ordenadas.

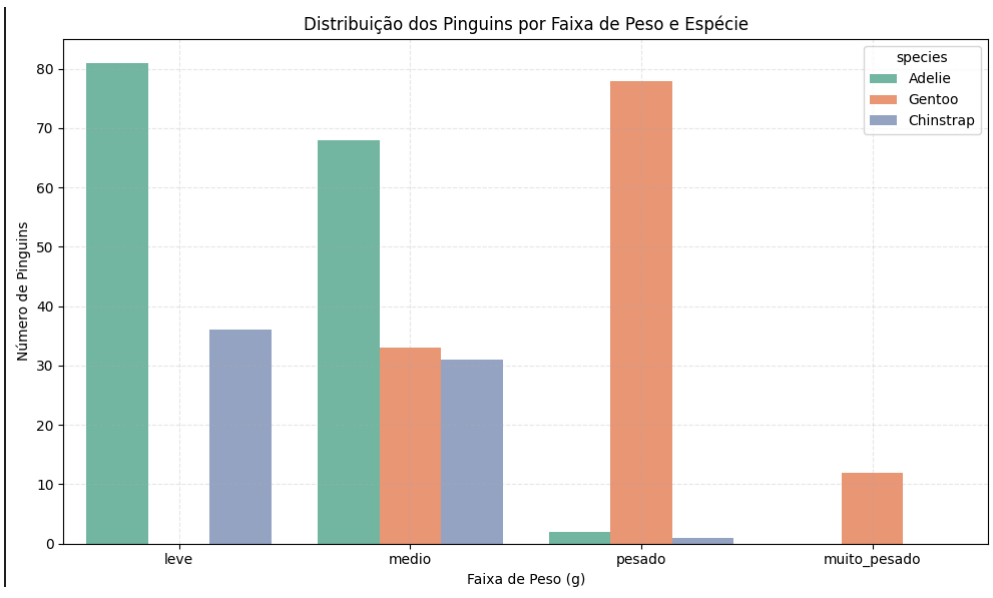
São uteis para simplificar a analise ou visualização de um dado continuo e utilizar algoritmos que lidam melhor com dados categóricos.

Vou realizar este procedimento na feature "**body_mass_g**" porque, é um dado continuo, com poucos vazios de relevancia e facil interpretação, ao plotar o dado consigo ver o seguinte:



A especie “Gentoo” consegue assumir valores de massa bem superiores as outras duas especies, o que ajuda a separar “Gentoo” das demais, no entanto não ajuda na separação entre “Chinstrap” e “Adelie”, podemos discretizar os valores de forma manual ou automática, observando os valores que foram plotados seria interessante uma discretização manual do tipo :

- Bin 1 → Leves (2700g a 3700g)
- Bin 2 → Médios (3700g a 4700g)
- Bin 3 → Pesados (4700g a 5700g)
- Bin 4 → Muito Pesados (> 5700g)



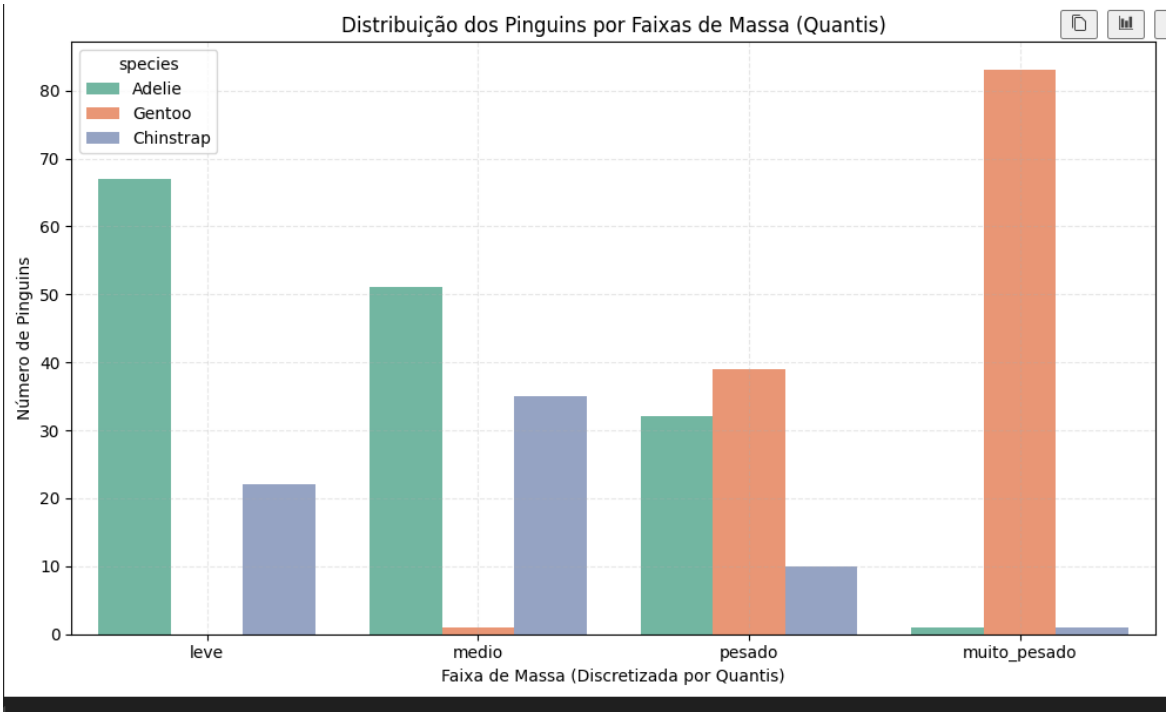
A discretização afeta o modelo de diversas formas, causando por um lado vantagens e por outro limitações, somente avaliando métricas estatísticas antes e depois da discretização para verificar se o impacto foi positivo.

As principais vantagens são simplificação interpretativa, Ajuste a modelos baseados em variáveis categóricas, Facilidade na geração de relatórios e visualizações.

As principais limitações são Perda de granularidade, Introdução de fronteiras artificiais, Redução da capacidade preditiva em modelos sensíveis a variações contínuas e Possibilidade de desbalanceamento entre categorias.

4. **Aplique a técnica de quantização com bins variáveis em uma variável contínua do dataset Palmer Penguins. Compare os resultados com a discretização feita no exercício anterior.**

Aplicando a técnica com bins variáveis chegamos ao seguinte resultado:

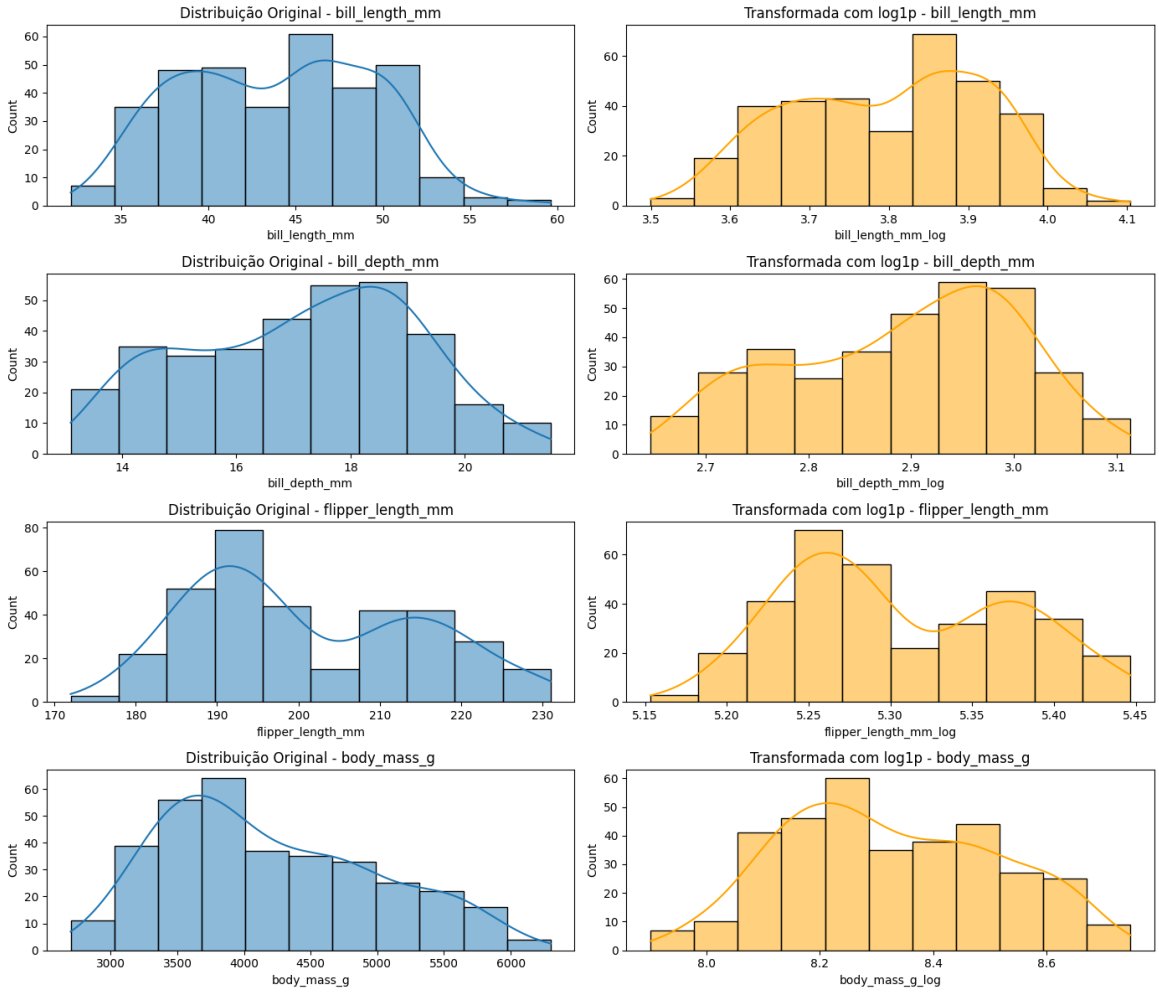


A discretização com bins variáveis (quantis) é vantajosa quando se deseja classes com tamanhos semelhantes, o que pode ser útil para modelos supervisionados, especialmente quando há desequilíbrio entre valores contínuos. No entanto, ela pode reduzir a interpretabilidade física dos grupos, já que os intervalos são definidos com base na ordem dos dados, e não nos seus valores absolutos.

No caso do dataset Palmer Penguins, os bins fixos destacam melhor a espécie Gentoo como grupo de maior massa, enquanto os bins variáveis suavizam essa separação, favorecendo a homogeneidade numérica entre classes.

5. **Aplique a FunctionTransformer do Scikit-Learn em uma variável do dataset Palmer Penguins. Descreva o processo e explique como essa transformação pode ser benéfica.**

R: Para entender o funcionamento da FunctionTransformer, foi utilizado uma transformação log1p com o objetivo de normalizar o dado, aplicando a diferentes features, foi percebido um bom resultado na feature de massa dos pinguins:



Antes da Transformação a Distribuição assimétrica à direita (valores altos são mais esparsos)

Depois: A transformação logarítmica reduziu a cauda longa, centralizou os dados e produziu uma forma mais próxima da distribuição normal.

A transformação foi eficaz como podemos ver, porque body_mass_g tinha grande amplitude e assimetria significativa. O log1p compressa os valores maiores, suavizando a distribuição.

6. **Utilize a PowerTransformer do Scikit-Learn para transformar uma variável do dataset Palmer Penguins. Discuta os benefícios dessa transformação.**

R: A aplicação da transformação PowerTransformer com o método Yeo-Johnson nos dados do dataset Palmer Penguins teve como principal objetivo a redução da assimetria das variáveis contínuas e a aproximação das distribuições empíricas de uma distribuição normal.

Essa transformação é especialmente indicada para dados que apresentam caudas longas ou distribuição assimétrica, pois realiza uma transformação não linear parametrizada que ajusta a curvatura dos dados conforme sua distribuição. No conjunto analisado, a variável body_mass_g foi a que mais se beneficiou da aplicação da Yeo-Johnson. A distribuição original apresentava assimetria à direita, com acúmulo de valores na faixa inferior e dispersão nas faixas superiores.

Após a transformação, a distribuição tornou-se mais simétrica, centrada e próxima da normalidade, o que é desejável para técnicas estatísticas que assumem homocedasticidade e distribuição normal das variáveis independentes.

Já nas variáveis bill_length_mm, bill_depth_mm e flipper_length_mm, a transformação Yeo-Johnson teve impacto limitado. Essas variáveis já apresentavam, em sua maioria, distribuições aproximadamente simétricas e com variação controlada.

Com isso, a transformação não produziu alterações significativas na forma da distribuição, o que é esperado quando os dados já se aproximam de uma distribuição gaussiana.Em síntese, a PowerTransformer com Yeo-Johnson demonstrou ser eficaz em body_mass_g por atuar sobre uma distribuição originalmente assimétrica. Nas demais variáveis, o efeito foi neutro, o que indica que sua aplicação deve ser precedida por uma análise exploratória das distribuições, evitando transformações desnecessárias.

7. **Aplique a normalização Min-Max do Scikit-Learn em uma ou mais variáveis do dataset Palmer Penguins. Explique como essa normalização impacta o modelo de Machine Learning.**

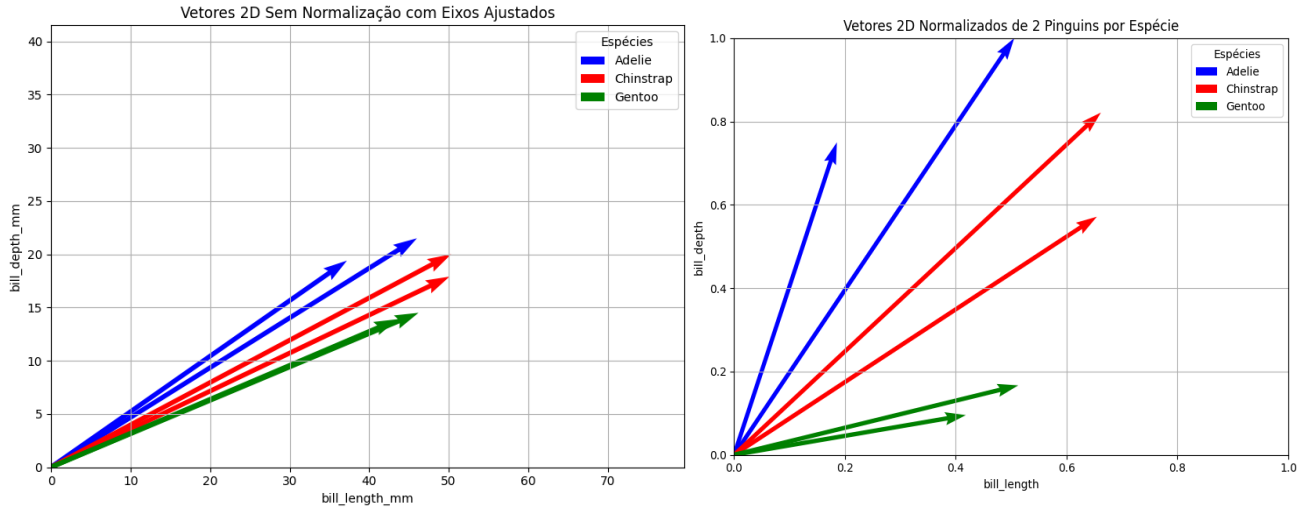
A normalização Min-Max consiste em reescalar os valores de uma variável para um intervalo definido, geralmente entre zero e um. Ao aplicar essa transformação às variáveis

contínuas do dataset Palmer Penguins, ajusta-se cada valor com base nos extremos mínimos e máximos de sua respectiva distribuição. O principal efeito desse procedimento é padronizar a escala dos atributos, permitindo que todos contribuam de maneira equilibrada para algoritmos sensíveis a magnitude, como redes neurais, máquinas de vetor de suporte e métodos baseados em distância, como k-NN.

Ao contrário de transformações que visam modificar a forma da distribuição, como PowerTransformer ou logaritmo, a normalização Min-Max preserva a forma original da distribuição, limitando-se à sua amplitude. Isso significa que, mesmo após a transformação, a distribuição de uma variável continuará assimétrica ou multimodal se já o era antes, mas todos os valores estarão contidos no mesmo intervalo, o que evita que variáveis com maior variância tenham peso desproporcional em modelos de aprendizado supervisionado.

Na prática, a aplicação da normalização Min-Max em variáveis como comprimento do bico, profundidade do bico, comprimento da nadadeira e massa corporal resulta em dados preparados para entrada em modelos preditivos que exigem uniformidade de escala. Essa técnica é particularmente útil quando se deseja integrar múltiplas variáveis com ordens de grandeza distintas em um pipeline de modelagem estatística ou de machine learning.

Podemos plotar os vetores para ver a diferença entre os dados normalizados e não normalizados.

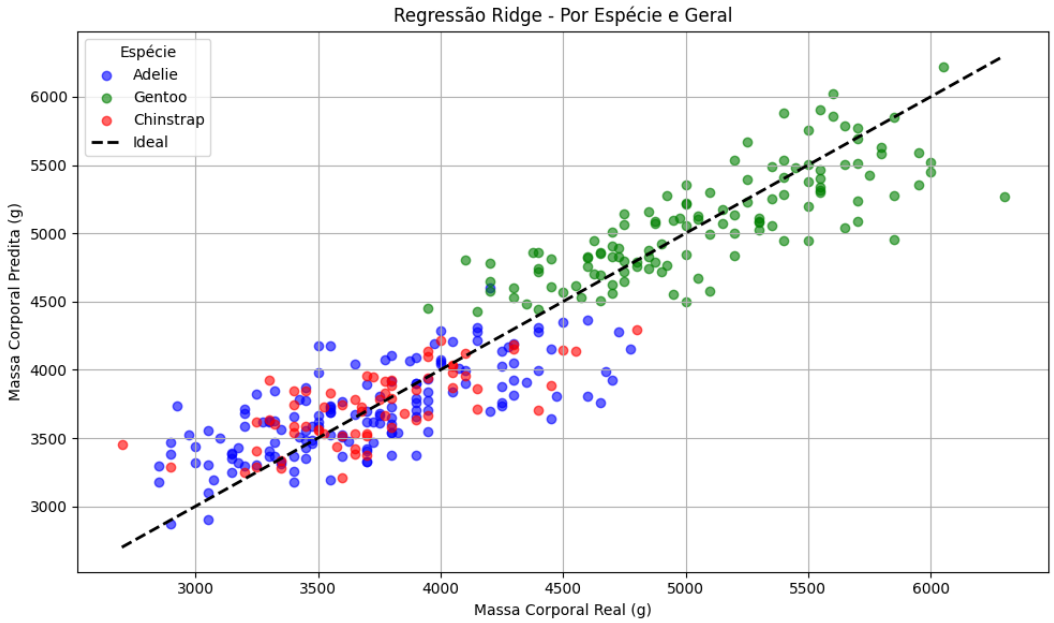


8. Utilize o StandardScaler do Scikit-Learn para normalizar variáveis do dataset Palmer Penguins. Discuta a diferença entre esta técnica e a normalização Min-Max.

A técnica de padronização StandardScaler, disponível no Scikit-Learn, transforma as variáveis contínuas para que cada uma delas tenha média igual a zero e desvio padrão igual a um. Isso é feito subtraindo a média e dividindo pelo desvio padrão de cada variável. No contexto do dataset Palmer Penguins, essa técnica é útil para uniformizar variáveis com escalas diferentes, especialmente antes da aplicação de modelos que assumem distribuição centrada, como regressão linear, análise de componentes principais (PCA), SVM e k-NN.

A principal diferença está no fato de que o StandardScaler preserva a estrutura estatística dos dados, centrando e padronizando cada variável com base em sua média e desvio padrão, o que facilita a convergência de muitos algoritmos. Já o MinMaxScaler apenas reescala os valores para um intervalo fixo, o que é útil para algoritmos que não assumem normalidade, mas que são sensíveis à escala bruta dos dados.

9. Implemente um modelo de regressão linear com regularização norma-L2 utilizando o dataset Palmer Penguins.



10. Utilize a FunctionTransformer do Scikit-Learn para aplicar múltiplas transformações em sequência a uma variável do dataset Palmer Penguins. Por exemplo, aplique uma transformação logarítmica seguida por uma transformação exponencial inversa. Explique os passos realizados e analise como essas transformações impactam a distribuição da variável transformada.

11. **Aplique a PowerTransformer do Scikit-Learn a uma variável do dataset Palmer Penguins. Em seguida, aplique uma transformação alternativa, como a normalização z-score (StandardScaler). Compare as distribuições resultantes das variáveis transformadas usando gráficos e estatísticas descritivas. Discuta as diferenças observadas e as situações em que cada técnica pode ser mais apropriada.**

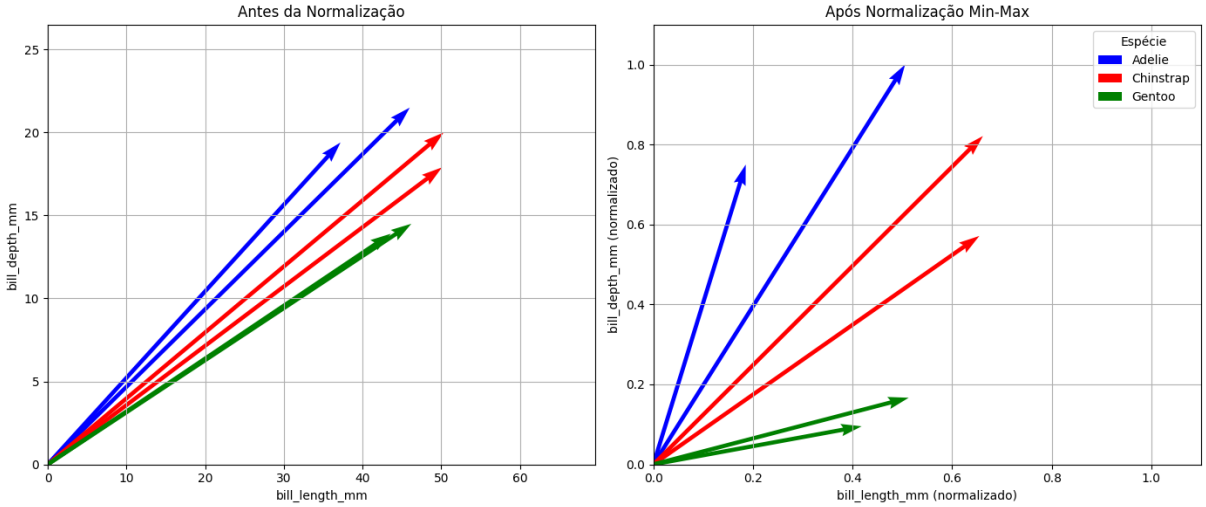
A aplicação sequencial de transformações utilizando o FunctionTransformer do Scikit-Learn sobre a variável body_mass_g do dataset Palmer Penguins permitiu demonstrar os efeitos e a reversibilidade de operações matemáticas comuns em pré-processamento de dados. Inicialmente, foi empregada a transformação logarítmica natural ajustada, por meio da função log1p, a qual suaviza distribuições assimétricas e comprime valores extremos, tornando a variável mais próxima de uma distribuição normal. Essa transformação é frequentemente utilizada em contextos onde há alta variância e cauda longa à direita, como ocorre com variáveis de massa ou renda.

Na sequência, aplicou-se a função inversa expm1, responsável por reverter os valores log-transformados ao seu domínio original. Essa composição de transformações é particularmente útil para validar a estabilidade numérica de pipelines de pré-processamento e garantir que não há distorções acumuladas na manipulação dos dados. Após a aplicação da sequência logarítmica seguida da exponencial inversa, observou-se que a distribuição resultante da variável transformada permaneceu praticamente idêntica à distribuição original. Essa equivalência confirma que, quando utilizadas de maneira simétrica e correta, transformações reversíveis preservam a estrutura dos dados, permitindo experimentações seguras sobre a adequação estatística de variáveis antes da modelagem.

Em um contexto mais amplo, a técnica demonstra a importância de compreender o comportamento funcional das transformações aplicadas, sobretudo quando elas antecedem algoritmos que dependem de suposições sobre a distribuição das variáveis. O uso de pipelines encadeados reforça ainda a prática recomendada de modularizar os processos de preparação de dados, promovendo reprodutibilidade, controle e rastreabilidade durante o desenvolvimento de modelos preditivos.

12. **Aplique a normalização Min-Max do Scikit-Learn a uma variável contínua do dataset Palmer Penguins. Visualize os dados normalizados e compare com os dados originais. Explique como a normalização Min-Max influencia a visualização dos dados e a comparação entre diferentes variáveis.**

Como já foi demostrado anteriormente, A normalização Min-Max reescala cada variável para o intervalo [0,1], preservando as proporções relativas, mas eliminando a influência da magnitude absoluta. Isso torna a comparação entre diferentes variáveis mais justa, sobretudo para algoritmos sensíveis à escala, como redes neurais, k-NN ou regressões com regularização.



2 LINK GITHUB

https://github.com/EstevezCodando/ml_infnet