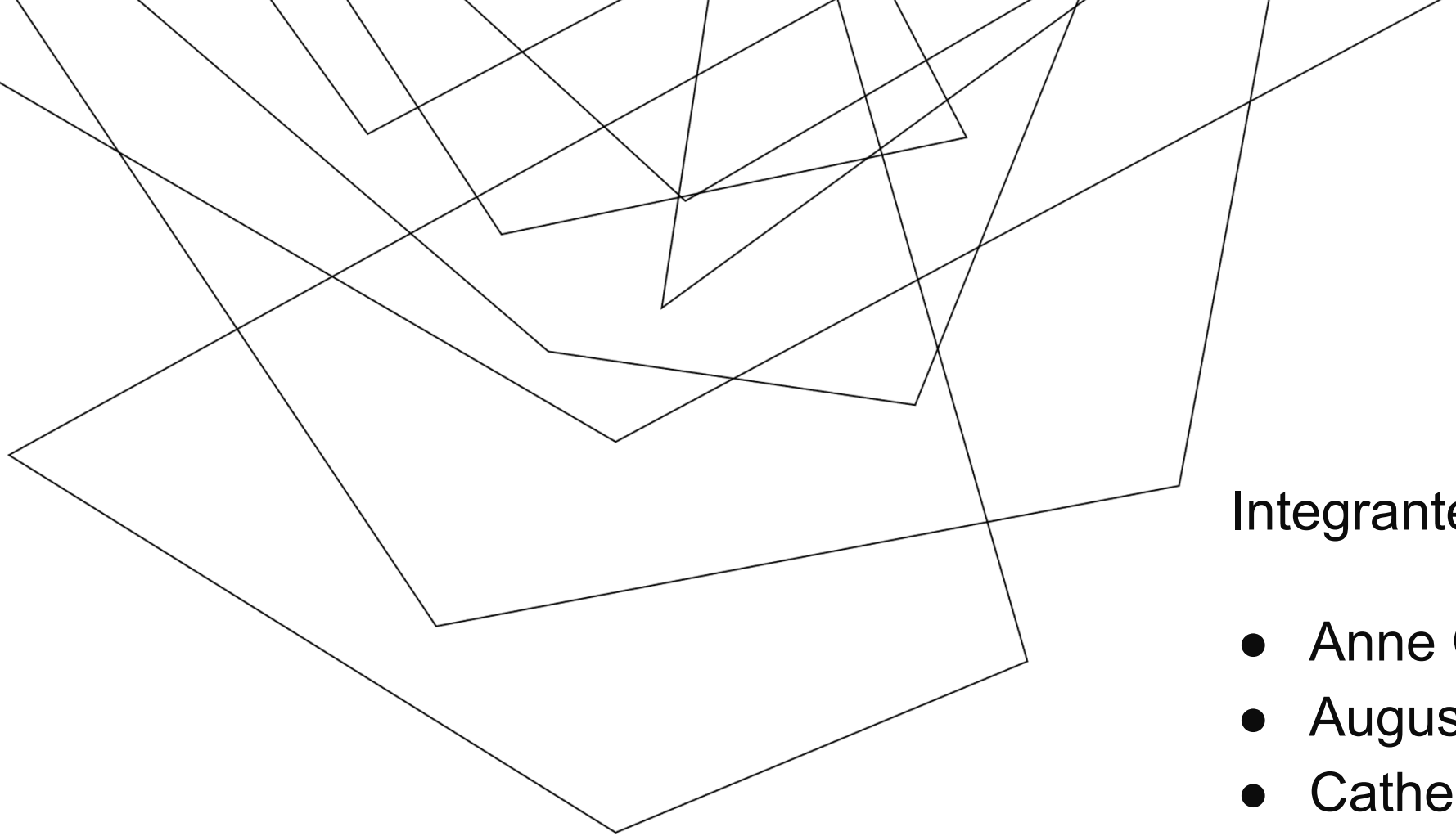
Predição de número de anéis de Abalones

Moluscos de concha achatada com formato arredondado ou ovalado, que possuem uma sequência de poros que varia entre 2 e 9 furos distribuídos ao longo de sua borda externa. Devido a esse formato peculiar, os abalones recebem o apelido de orelhas do mar.



Integrantes:

- Anne Carvalho
- Augusto César
- Catherine Markert
- Hernando Henrique
- João Barboza

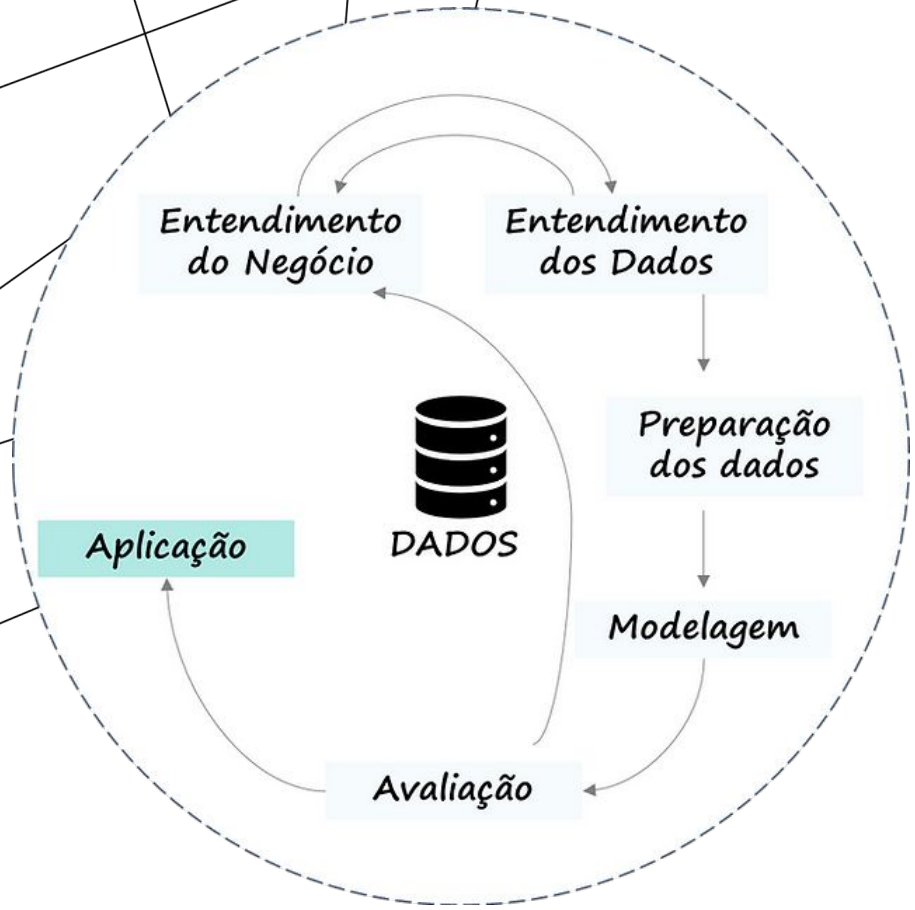
CONTEXTUALIZAÇÃO

A idade do molusco Abalone é definida após cortar a sua casca através do cone, colorindo-a e contando o número de anéis via um microscópio. Tal tarefa demorada e extensa e mata o animal.

OBJETIVO

Predizer a idade dos moluscos abalones por meio de suas características físicas utilizando técnicas de Machine Learning e seguindo a metodologia CRISP-DM.

MÉTODOS



1. **Análise e discussões sobre a atividade proposta**
   - Seguindo a metodologia CRISP-DM.
2. **Análise Exploratória**
   - Exploração dos dados disponibilizados no dataset.
   - Criação de um dicionário de dados.
   - Preparação dos dados.
3. **Técnicas de Machine Learning**
   - Regressão Linear Ordinária.
   - Regressão Ridge.
   - K-Nearest-Neighbors (KNN).
   - Support Vector Machine (SVM).

MÉTODOS

Preditor	Descrição	Tipo	Subtipo
sexo	sexo do molusco	qualitativo	nominal
comprimento	tamanho máximo da concha (mm)	quantitativo	contínuo
diâmetro	medida perpendicular ao comprimento (mm)	qualitativo	contínuo
altura	altura total do molusco com casca (mm)	qualitativo	contínuo
peso total	peso total do molusco (g)	quantitativo	contínuo
peso sem concha	peso do molusco sem casca (g)	quantitativo	contínuo
peso do intestino	peso do intestino após sangrar (g)	quantitativo	contínuo
peso da concha	peso da concha seca (g)	quantitativo	contínuo
anéis	quantidade de anéis, +1.5 é a idade do abalone	quantitativo	discreto

Fig. 01 – Dicionário de Dados

2. Análise Exploratória

Implementamos o dicionário de dados para melhor compreensão das features a serem utilizadas e realizamos algumas perguntas de partida e hipóteses, como:

1. Qual a relação entre as medidas de tendência central?
2. Qual a relação entre as medidas de dispersão?
3. Qual a relação entre sexo, idade, peso e altura?
4. Qual a relação de peso segundo o sexo e a idade?
5. Quais as dimensões do abalone e o seu respectivo sexo?
6. Qual a relação da altura conforme o respectivo sexo?
7. Qual é a matriz de correlação?

RESULTADOS

2. Análise Exploratória

Medidas de Tendência Central e Dispersão

	Comprimento	Diâmetro	Altura	Peso total	Peso sem casca	Peso do intestino	Peso da concha	Anéis
count	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000
mean	0.523992	0.407881	0.139516	0.828742	0.359367	0.180594	0.238831	9.933684
std	0.120093	0.099240	0.041827	0.490389	0.221963	0.109614	0.139203	3.224169
min	0.075000	0.055000	0.000000	0.002000	0.001000	0.000500	0.001500	1.000000
25%	0.450000	0.350000	0.115000	0.441500	0.186000	0.093500	0.130000	8.000000
50%	0.545000	0.425000	0.140000	0.799500	0.336000	0.171000	0.234000	9.000000
75%	0.615000	0.480000	0.165000	1.153000	0.502000	0.253000	0.329000	11.000000
max	0.815000	0.650000	1.130000	2.825500	1.488000	0.760000	1.005000	29.000000

Fig. 02 – Medidas de Tendência Central e Dispersão

RESULTADOS

2. Análise Exploratória

Relações entre as variáveis sexo, altura e anéis.

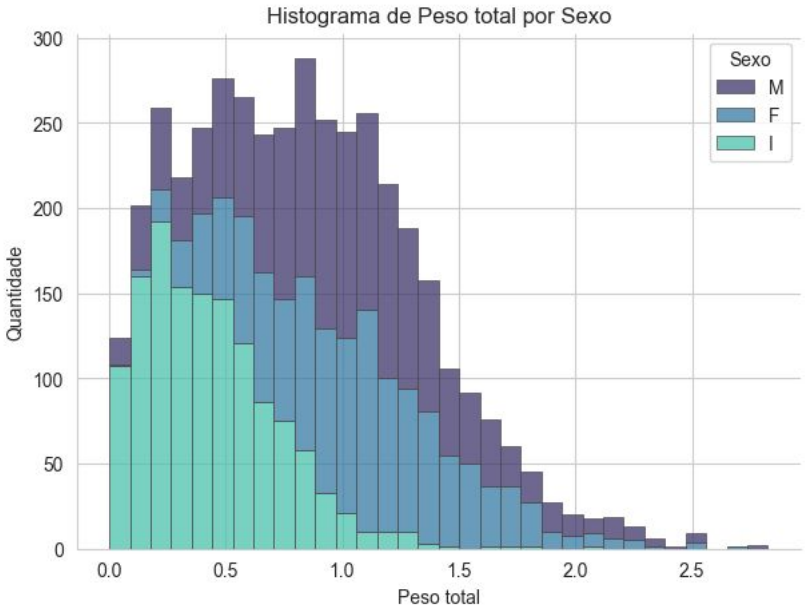


Fig. 03 – Histograma de peso total por sexo

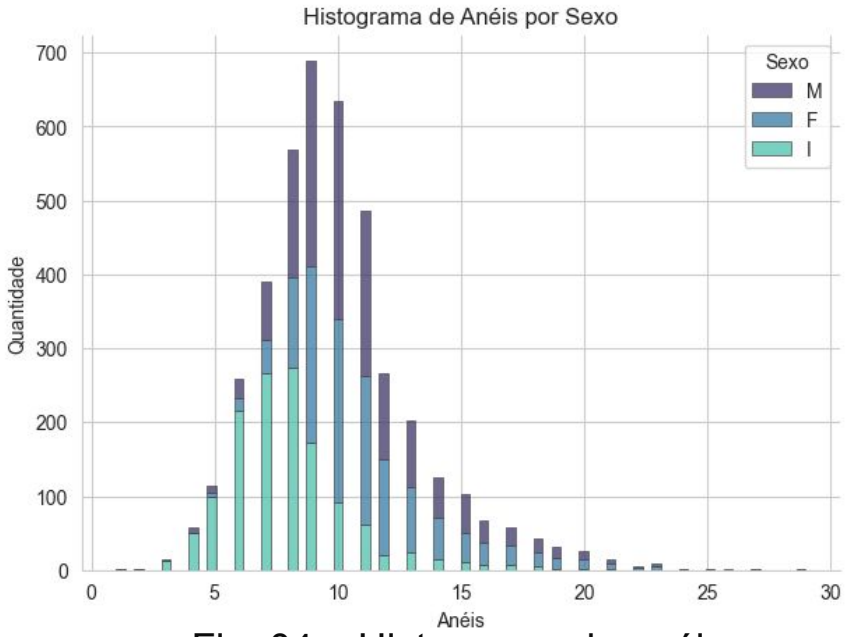


Fig. 04 – Histograma de anéis por sexo

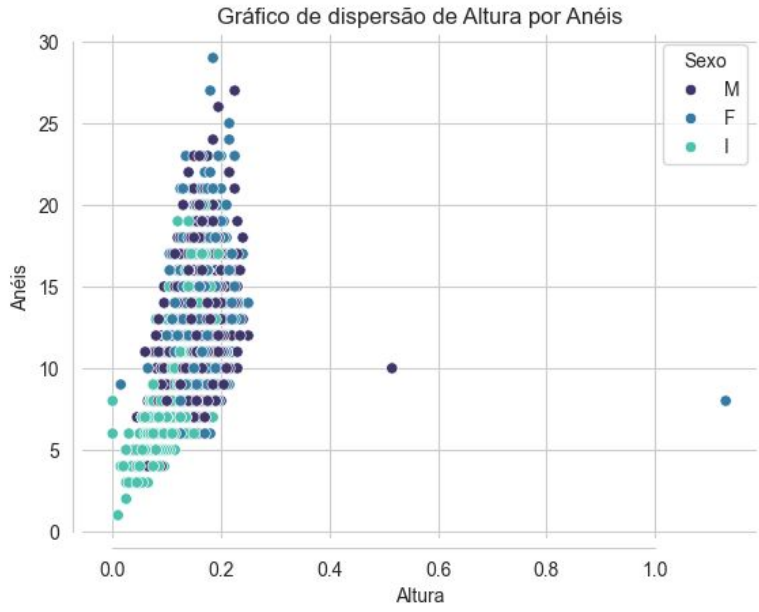


Fig. 05 – Medidas de altura por anéis

RESULTADOS

2. Análise Exploratória

Relações entre as variáveis sexo e peso.

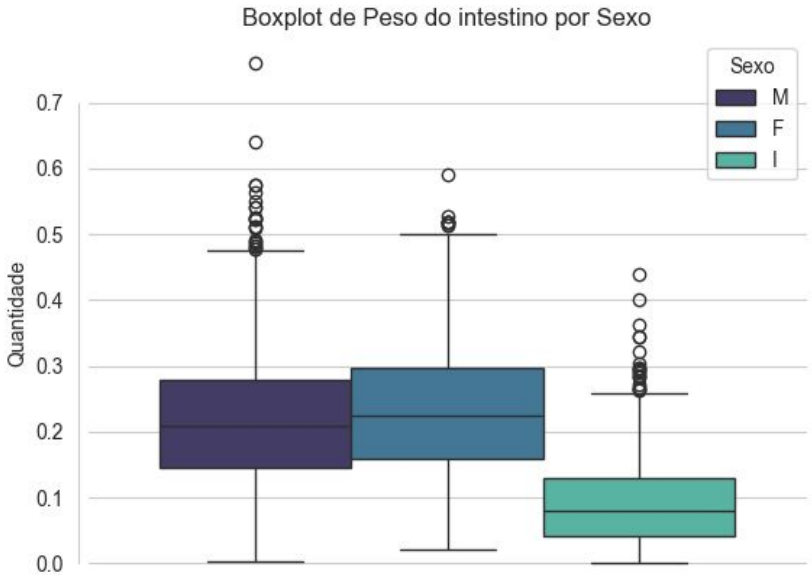


Fig. 06 – Boxplot de peso do intestino por sexo

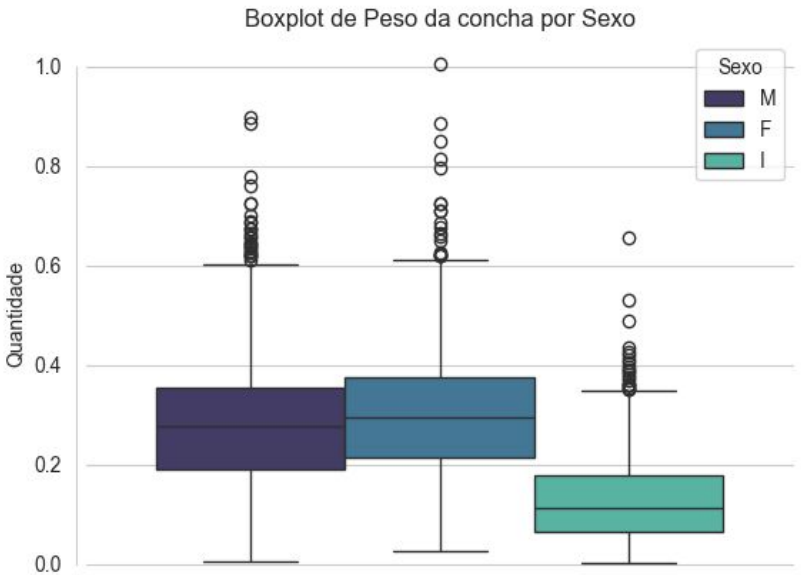


Fig. 07 – Boxplot de peso da concha por sexo

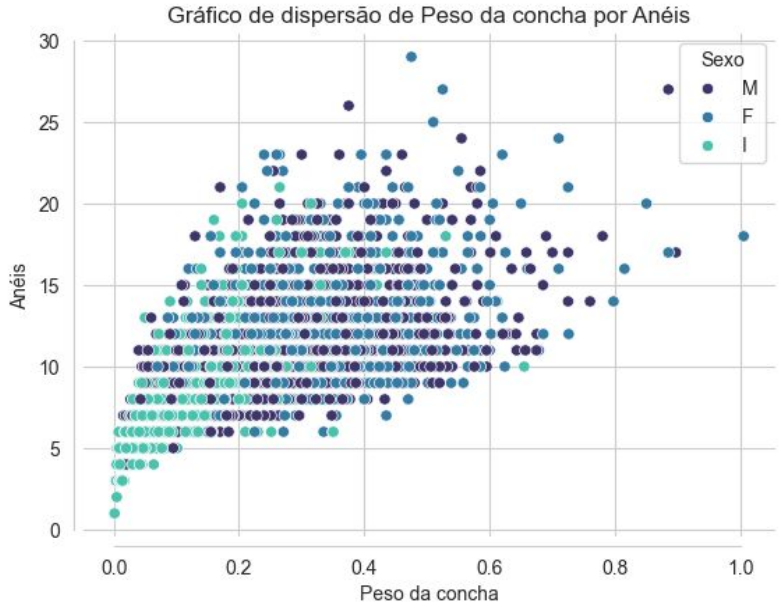


Fig. 08 – Gráfico de dispersão de peso da concha por anéis

RESULTADOS

2. Análise Exploratória

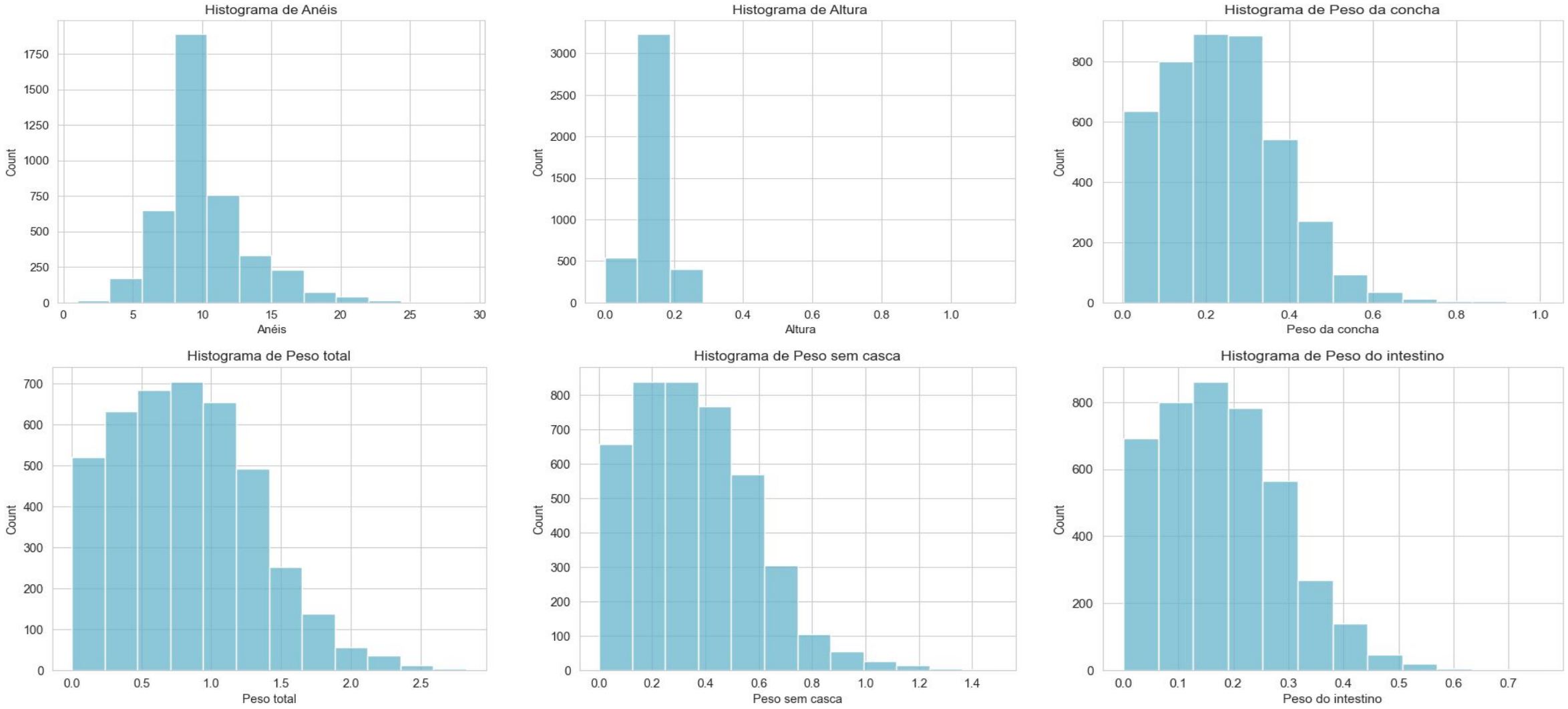


Fig. 09 – Histogramas de anéis, altura e pesos.

RESULTADOS

2. Análise Exploratória

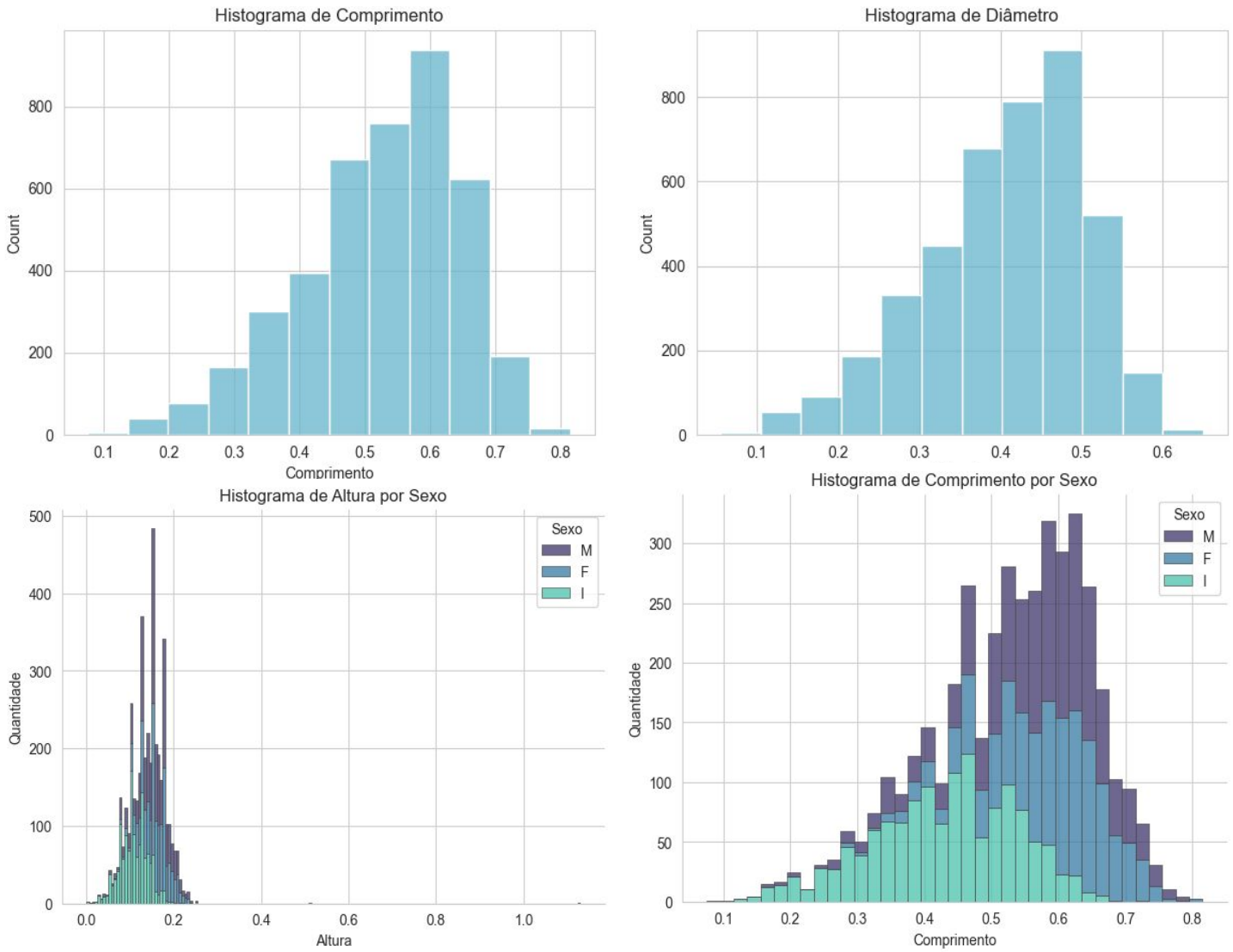
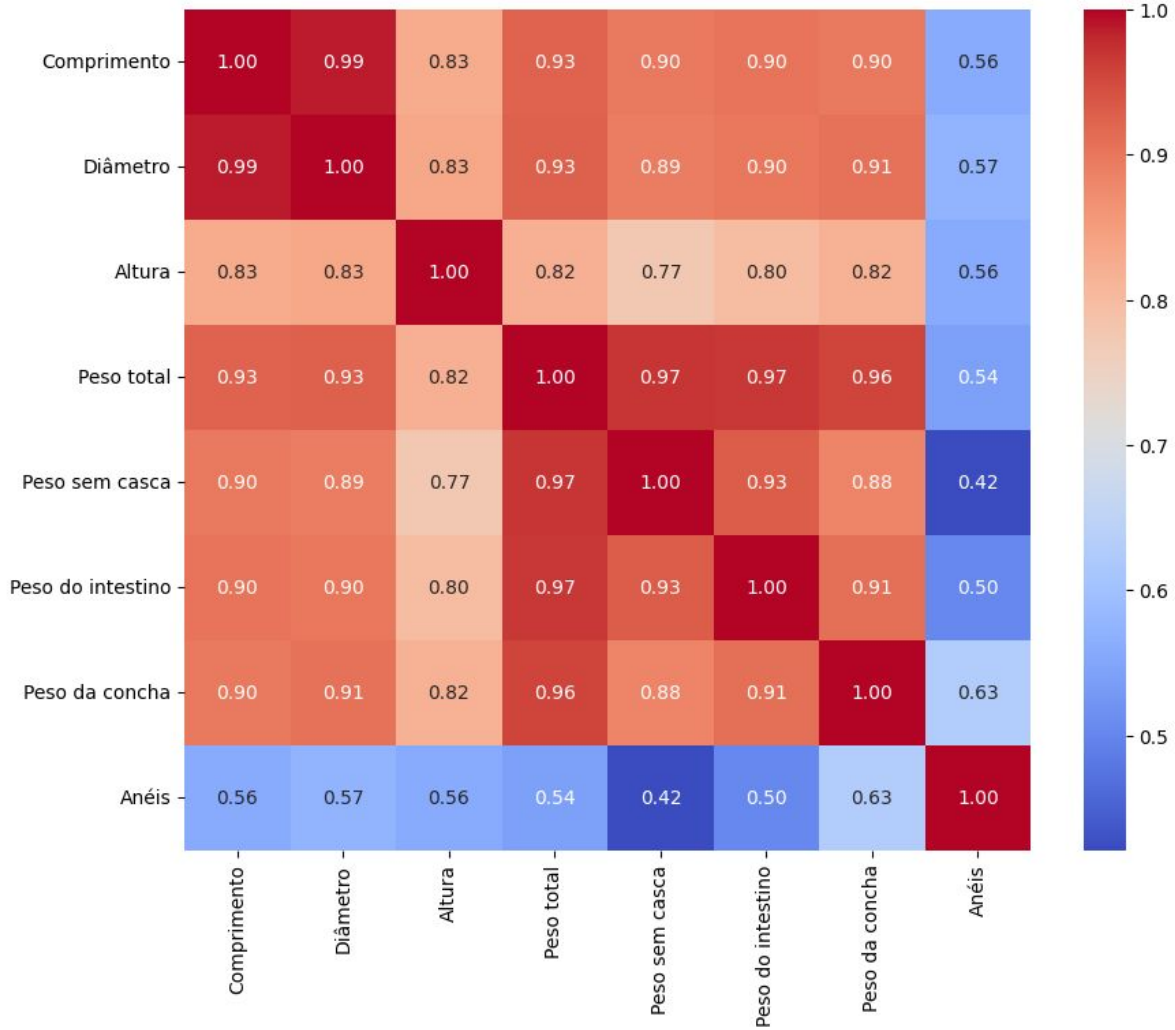


Fig. 10 – Histogramas de comprimento, diâmetro.

RESULTADOS



2. Análise Exploratória

Matriz de Correlação

Percebe-se que os preditores estão muito correlacionados entre si, mas em valores menores com a saída.

Fig. 11 - Matriz de Correlação

MÉTODOS

3. Modelos de Regressão Lineares

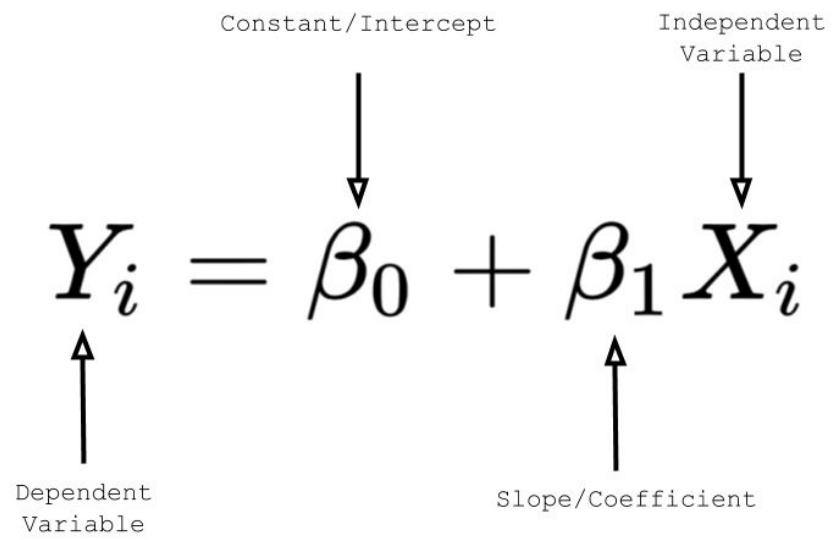


Fig. 12 - Regressão Linear Ordinária

Regressão Linear Ordinária (OLS)

- Pretende-se prever a tendência dos dados traçando uma linha, que representa uma função linear.
- Traçamos uma reta e verificamos a distância entre os pontos dispostos no gráfico e a linha traçada para descobrir a melhor reta.

MÉTODOS

3. Modelos de Regressão Lineares

$$Cost\ (W) = RSS(W) + \ \lambda * (sum\ of\ squares\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i \ - \sum_{j=0}^{M} w_j \, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} w_j^2$$

Fig. 13 - Regressão de Ridge

Regressão de Ridge

- Modelo de penalização, utilizado para evitar overfitting e analisar a multicolinearidade dos dados.
- Coeficiente de penalização lambda, o qual é torna o modelo menos sensível.
- Quanto mais se aumentar o valor de lambda, menos sensível à linha de regressão de Ridge vai se tornar para as abscissas.

MÉTODOS

3. Modelos de Regressão Não-Lineares

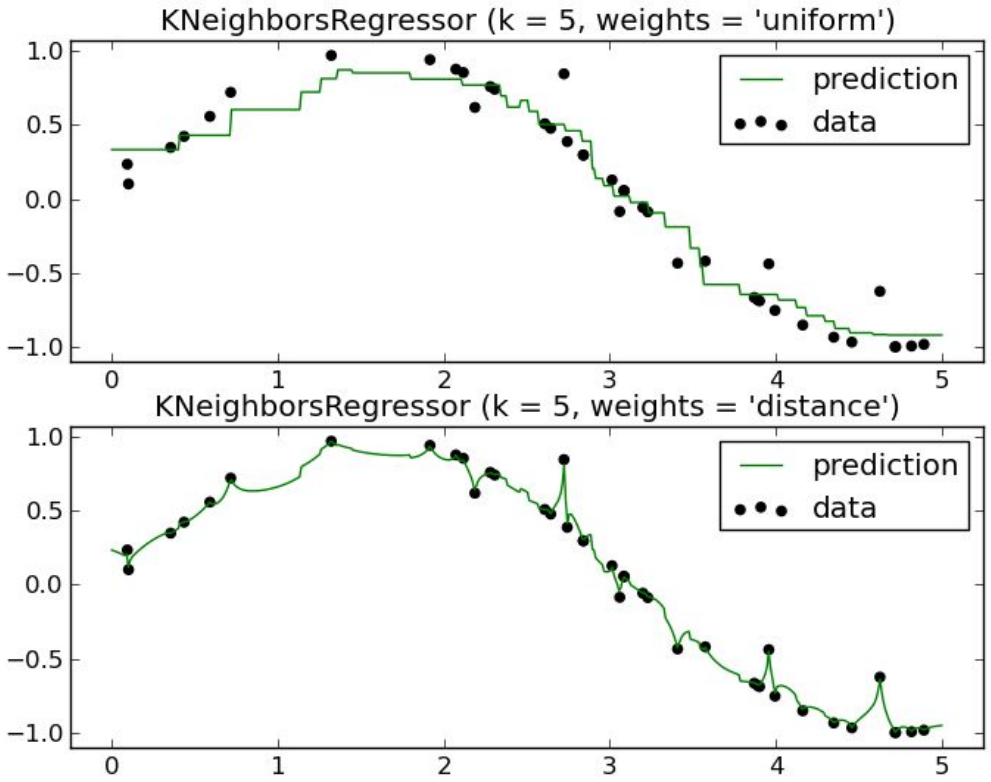


Fig. 14 - KNN para regressão

K-nearest-Neighbors

- Maneira simples de realizar regressão.
- Utiliza métricas de distância para prever a resposta desejada.
- Os k-vizinhos mais próximos definirão o valor de saída predito.

MÉTODOS

3. Modelos de Regressão Não-Lineares

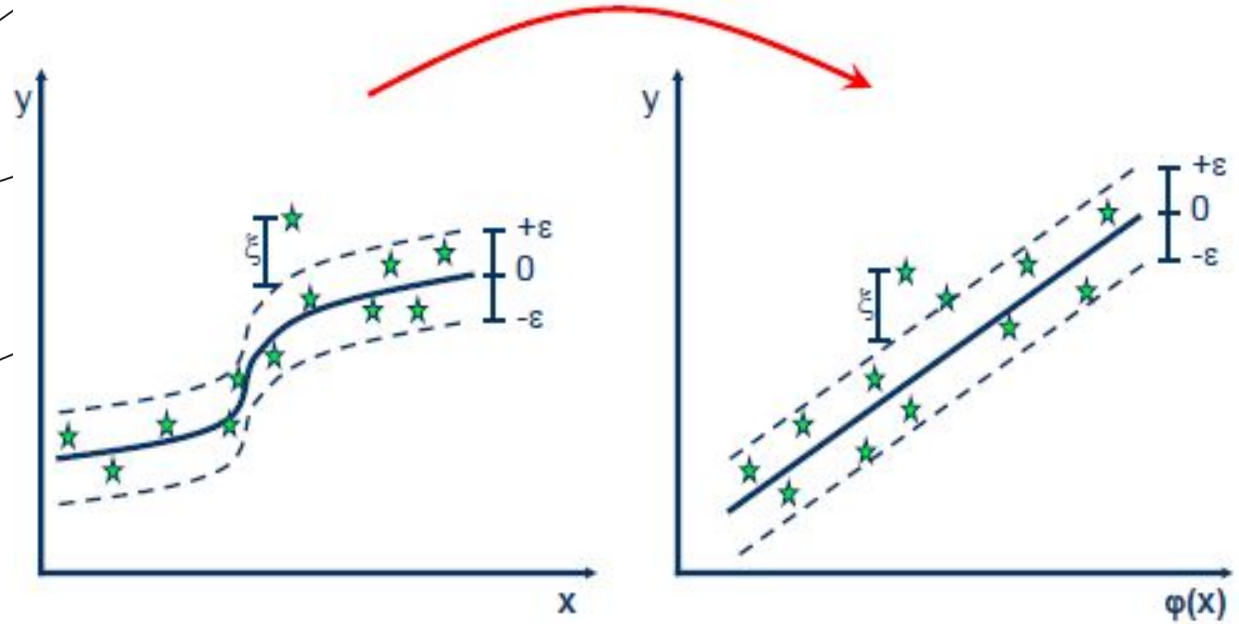


Fig. 15 - SVM para regressão

Support Vector Machine (SVM)

- Encontra o melhor hiperplano que melhor se adequa aos pontos em um espaço contínuo.
- O hiperplano é construído para reduzir o erro e definido a partir dos pontos mais próximos dele, os quais são os vetores de suporte.
- Funções kernel transformam dados para dimensões maiores no espaço das amostras para realizar uma regressão não-linear.

Experimentos

3. Modelos de Regressão

Hiperparâmetros

- **OLS:** Nenhum
- **Regressão Ridge:** Lambda variando de 0.01 a 4.99, no passo 0.01.
- **KNN:** K variando de 1 a 50 (números ímpares), pesos de distância iguais (uniforme) e do inverso da distância (distância)
- **SVM:** Limite de iterações 100000, kernel variando em rbf, polinomial e sigmóide.

Experimentos

3. Modelos de Regressão

Pré-Processamento, Validação Cruzada, Métricas de Avaliação

- Uso de One-Hot-Encoding pra codificar variável categórica.
- Não havia dados faltantes.
- Dados quantitativos normalizados com z-score.
- Grid search para encontrar melhores hiperparâmetros com a validação cruzada k-fold com k=5.
- Resultados obtidos com modelos com os melhores hiperparâmentros e uma validação cruzada de 21 splits, tendo 20% do dataset para teste.
- R^2 e RMSE utilizados como métodos de avaliação.

Resultados

3. Modelos de Regressão

Hiperparâmetros

- **Regressão Ridge:** Lambda igual a 4.99.
- **KNN:** K=15 vizinhos, com peso “distância”.
- **SVM:** Kernel rbf.

```
Best Parameters for OLS: {}
Best Parameters for RR: {'alpha': 4.99}
Best Parameters for KNN: {'n_neighbors': 15, 'weights': 'distance'}
Best Parameters for SVR: {'kernel': 'rbf'}
```

Fig. 16 - Melhores parâmetros dos modelos

Resultados

3. Modelos de Regressão Lineares

	name	OLS	RR
fit_time	mean	0.022639	5.406434
	std	0.002337	0.496836
score_time	mean	0.004975	0.004921
	std	0.002187	0.001235
test_r2	mean	0.520441	0.536488
	std	0.028013	0.021415
test_root_mean_squared_error	mean	2.230271	2.223921
	std	0.089971	0.087914

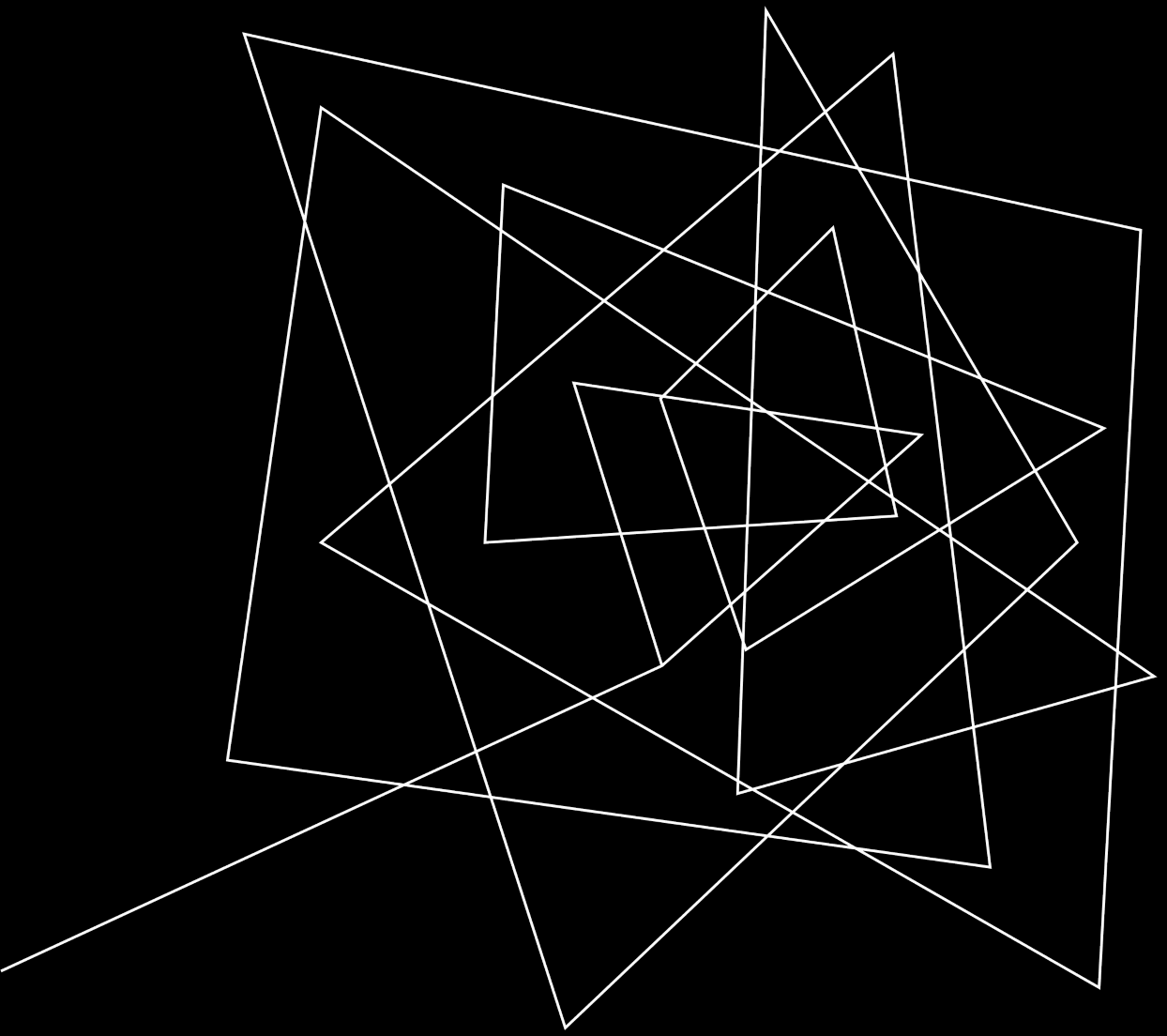
Fig. 17 - Resultados dos modelos lineares

Resultados

3. Modelos de Regressão Não-Lineares

	name	KNN	SVR
fit_time	mean	6.134819	8.197040
	std	1.252128	0.699592
score_time	mean	0.026955	0.280972
	std	0.005692	0.037431
test_r2	mean	0.521831	0.524043
	std	0.017963	0.016654
test_root_mean_squared_error	mean	2.262508	2.226751
	std	0.108292	0.066687

Fig. 18 - Resultados dos modelos não-lineares



CONCLUSÃO

- Resultados muito parecidos.
- Correlação dos preditores entre si, mas não com a saída, podem ter gerado esses resultados.
- Melhor modelo: Regressão Ridge.

OBRIGADO!