Módulo | Análise de Dados: Aprendizado de Máquina, Regressão

Caderno de **Exercícios**
Professor André Perez

Tópicos

1. Regressão;
2. Dados;
3. Treino;
4. Avaliação;
5. Predição.

Exercícios

1. Pinguins

Neste exercício, vamos utilizar uma base de dados com informações sobre penguins. A idéia é prever o peso do penguin (**body_mass_g**) baseado em suas características físicas e geográficas (variáveis preditivas).

In [ ]:
```
import sklearn
import numpy as np
import pandas as pd
import seaborn as sns
```

In [ ]:
```
penguim = sns.load_dataset('penguins')
```

In [ ]:
```
penguim.head()
```

1.1. Analise exploratoria

Utilize os gráficos abaixo para entender melhor a relação entre os atributos e variável resposta da base de dados. Comente o que observou em cada gráfico.

- Atributos por sexo:

In [ ]:
```
with sns.axes_style('whitegrid'):

  grafico = sns.pairplot(data=penguim, hue="sex", palette="pastel")
```

Comentário: ?

- Atributos por espécie:

In [ ]:
```
with sns.axes_style('whitegrid'):

  grafico = sns.pairplot(data=penguim, hue="species", palette="pastel")
```

Comentário: ?

- Atributos por ilha:

In [ ]:
```
with sns.axes_style('whitegrid'):

  grafico = sns.pairplot(data=penguim, hue="island", palette="pastel")
```

Comentário: ?

2. Dados

2.1. Valores nulos

A base de dados possui valores faltantes, utilize os conceitos da aula para trata-los.

In [ ]:
```
# resposta da questão 2.1
```

2.2. Variáveis numéricas

Identifique as variáveis numéricas e crie uma nova coluna **padronizando** seus valores. A nova coluna deve ter o mesmo nome da coluna original acrescidade de "*_std*".

> **Nota**: Você não deve tratar a variável resposta.

In [ ]:
```
# resposta da questão 2.2
```

2.3. Variáveis categóricas

Identifique as variáveis categóricas nominais e ordinais, crie uma nova coluna aplicando a técnica correta de conversão a seus valores. A nova coluna deve ter o mesmo nome da coluna original acrescidade de "*_nom*" ou "*_ord*".

> **Nota**: Você não deve tratar a variável resposta.

In [ ]:
```
# resposta da questão 2.3
```

2.4. Limpeza

Descarte as colunas originais e mantenha apenas a variável resposta e as variáveis preditivas com o sufixo *_std*", *_nom*" e "*_ord*".

In [ ]:
```
# resposta da questão 2.4
```

2.5. Treino/Teste

Separe a base de dados em treino e teste utilizando uma proporção de 2/3 para treino e 1/3 para testes.

In [ ]:
```
# resposta da questão 2.5
```

3. Modelagem

3.1. Treino

Treine um modelo de **regressão linear** com os **dados de treino** (2/3).

In [ ]:
```
# resposta da questão 3.1
```

3.2. Avaliação

Calcule o **RMSE** para o modelo de **regressão linear** treinado com os **dados de teste** (1/3).

> **Nota:** Como referência, eu consegui um RMSE de approx. 296g, 7% da média do peso, uma performance razoável (seu RMSE pode não ser igual).

In [ ]:
```
# resposta da questão 3.2
```

4. Predição

4.1. Novo penguim

Qual o peso de um penguim com as seguintes características:

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	sex
Adelie	Biscoe	38.2	18.1	185.0	Male

> **Atenção:** Lembre-se de pre-processar os atributos assim como nos exercício 2.2 e 2.3

> **Nota:** Como referência eu obtive um peso predito de 3786.16g (a sua predição pode não ser igual).

In [ ]:
```
# resposta da questão 4.1
```
