Módulo | Análise de Dados: Aprendizado de Máquina, Classificação

Caderno de **Exercícios**
Professor André Perez

Tópicos

1. Classificação;
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 a espécie do penguin (**species**) 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(25)
```

1.1. Análise exploratória

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 numéricos por espécie:

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

  grafico = sns.pairplot(
      data=penguim.drop(['sex', 'island'], axis=1),
      hue="species",
      palette="pastel"
  )
```

Comentário: ?

- Sexo por espécie:

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

  grafico = sns.countplot(
      data=penguim,
      x='sex',
      hue="species",
      palette="pastel"
  )
```

Comentário: ?

- Ilha por espécie:

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

  grafico = sns.countplot(
      data=penguim,
      x='island',
      hue="species",
      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 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.

> **Nota**: Por definição, árvores de decisão **não precisam** da transformação de atributos categóricos em numéricos. Contudo, por **limitação** do pacote Python Scikit Learn, devemos conduzir esta etapa. Mais informações neste link.

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

2.3. Limpeza

Descarte as colunas categóricas originais e mantenha a variável resposta na primeira coluna do dataframe.

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

2.4. 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 2.4
```

3. Modelagem

3.1. Treino

Treine um modelo de **árvore de decisão** com os **dados de treino** (2/3). Gere o gráfico da árvore do modelo treinado e responda: quantas **folhas** a árvore treinada possui?

Resposta: ?

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

3.2. Avaliação

a. Matriz de Confusão

Calcule e visualize a **matriz de confusão** para o modelo de **árvore de decisão** treinado com os **dados de teste** (1/3). Comente os resultados.

Comentário: ?

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

b. Acurácia

Calcule a **acurácia** para o modelo de **árvore de decisão** treinado com os **dados de teste** (1/3).

> **Nota:** Como referência, eu consegui uma acurácia de approx. 96% (sua acurácia pode não ser igual).

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

4. Predição

4.1. Novo penguim

Qual a espécie de um penguim com as seguintes características:

island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
Biscoe	38.2	18.1	185.0	3950.0	Male

> **Atenção:** Lembre-se de pre-processar os atributos assim como nos exercício 2.2. A ordem dos atributos importa, deve ser a mesma usada na modelagem.

> **Nota:** Como referência eu obtive **adelie** como espécie predita (a sua predição pode não ser igual).

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