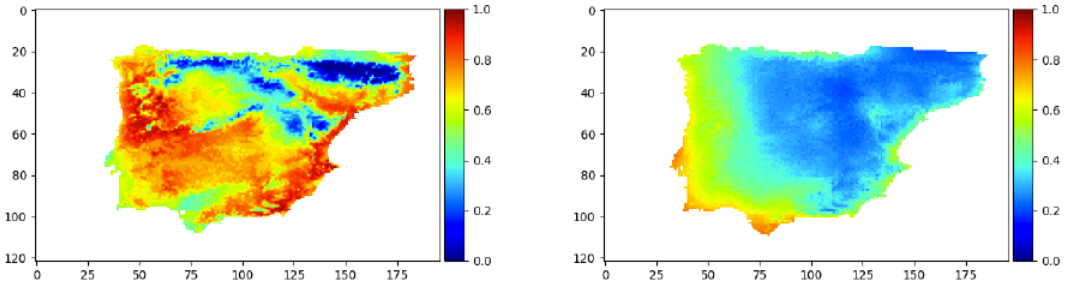
Trabalho Prático Individual, ed. 2022/23

1 - Objetivos

- Utilizar algoritmos de aprendizagem automática para prever o mapa da distribuição da abelha europeia (*apis mellifera honeybee*) na Península Ibérica.
- Avaliar o desempenho dos algoritmos.
- Construir o mapa de adequação da espécie.

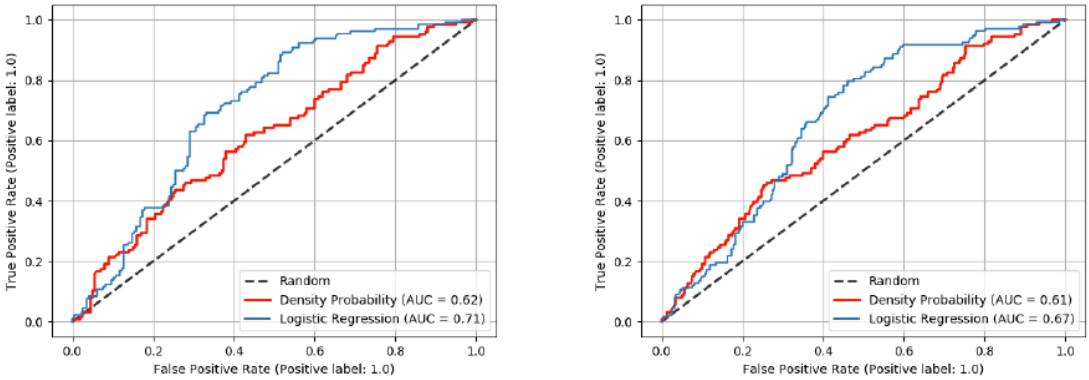
2 - Descrição Geral

O presente trabalho tem como principal objetivo comparar algoritmos de aprendizagem supervisionada, para descrever a relação existente entre os locais de ocorrência da *apis melífera honeybee* na Península Ibérica e um conjunto de variáveis climáticas que determinam o comportamento da espécie.
O resultado será um mapa de adequação que pretende indicar as zonas mais (ou menos) propensas para a proliferação da espécie. Trata-se de um modelo da distribuição da espécie.



Para o treino dos algoritmos, é disponibilizado um *dataset* composto pelas variáveis climáticas (preditores) e os dados de ocorrência da espécie (resposta alvo). Cada observação no *dataset* corresponde a um local no mapa onde pode ser observada (ou não) a presença da espécie (1 -> presença; 0 -> pseudo-ausência).
Desse *dataset*, deve ser selecionada uma amostra contendo **todos** os locais onde a espécie foi observada (resposta = 1) **e um número a determinar,** considerado adequado, de locais onde se crê que a espécie esteja ausente (resposta = 0). Note que as respostas 0 representam pseudo-ausências o que não significa que a espécie não pode viver e prosperar num determinado local mas apenas que aí não foram registadas observações. O número de exemplos negativos que resolver utilizar e a forma de escolha dos locais de amostragem é assim um primeiro problema que será determinante para o sucesso dos seus modelos de ocupação geográfica.

Esta amostra deve posteriormente ser dividida em conjunto de treino e conjunto de teste (a utilização de validação cruzada será valorizada).
Com o objetivo de avaliar o desempenho do algoritmo, a curva ROC e a área sob a curva ROC (AUC) deverão ser calculadas.



No final, opcionalmente, o mapa de adequação da espécie poderá ser construído com o melhor modelo que conseguir sintetizar. Nesse caso o valor da predição (contínuo entre 0 e 1; trata-se de um problema de regressão) terá de ser calculado para cada um dos pontos do mapa.

3 - Dados

- Dados de ocorrência da *apis melífera honeybee* na Peninsula Ibérica (*presence-only data*).
- Conjunto de variáveis ambientais (*mntcm, mxtwm, rfseas, tann*) que descrevem o comportamento da espécie.

4 – Algoritmos sugeridos

- Vizinhos mais próximos (*Nearest Neighbours* como exemplo de Classificadores não paramétricos).
- Máquinas de vetores de suporte (SVM – *Support Vector Machines*).
- Árvores de decisão (*Decision Trees*).
- Floresta Aleatória (*Random Forest* como exemplo de Combinação de classificadores).
- Redes neuronais (*Multi-Layer Perceptron*).

5 - Métricas de Desempenho

- Curva ROC.
- AUC.

6 – Entregas

- Ficheiro com o código fonte.
- PDF com:
  - mapa de adequação da espécie (opcional);

- descrição da metodologia (e cardinalidade) da amostragem escolhida para a construção do conjunto de treino;
- gráfico com a curva ROC e o valor da AUC para cada um dos métodos e valores dos respetivos hiper-parâmetros;
- breve discussão dos resultados.

7 – Dataset

A tabela abaixo descreve a estrutura do ficheiro fornecido. Os valores da latitude e longitude **não** serão utilizados pelo algoritmo de aprendizagem. Serão apenas úteis para a construção do mapa de adequação resultante que irá colocar em cada coordenada o valor da predição produzidos pelos algoritmos (em função das 4 variáveis ambientais).

Portanto, X será uma matriz onde cada linha representa um ponto no mapa (um local). Os valores das 4 variáveis ambientais estão dispostos nas primeiras 4 colunas do ficheiro. A última coluna, y, representa a presença (1) ou hipotética ausência (0) da espécie.

As variáveis consideradas pertinentes para a distribuição desta espécie são: a temperatura média anual (*tann*), a sazonalidade da chuva (*rfseas*), a temperatura máxima do mês mais quente (*mxtwm*) e a temperatura mínima do mês mais frio (*mntcm*).

Os dados não estão normalizados.

mntcm	**mxtwm**	**rfseas**	**tann**	**latitude**	**longitude**	**y**
60	*23*	*36*	*14*	*16*	*52*	*1*
...	...	...	...	...	...	...
6	*2*	*3*	*1*	*16*	*51*	*0*

8 – Material de Apoio

- Scikit-learn:
  - https://scikit-learn.org/
- Numpy
  - https://numpy.org/
- Matplotlib
  - https://matplotlib.org/
- Rasterio
  - https://rasterio.readthedocs.io/en/latest/
- SciPy
  - https://www.scipy.org/
- Jupyter notebook
  - https://jupyter.org/