**TIA(PT)**

**Previsão do cancro da mama**

**Objetivo:**

O objetivo deste projeto é detetar o cancro da mama a partir de um conjunto de dados.

**Descrição do Dataset:**

O cancro da mama é o cancro mais diagnosticado em mulheres, representando 1 em cada 4 casos anuais de cancro em todo o mundo. Em 2020, a incidência na região da OMS na Europa foi estimada em 576.300, e na UE-27 foi de 355.500. Aproximadamente 21% dos casos de cancro da mama na Europa ocorrem em mulheres com menos de 50 anos, enquanto 35% ocorrem entre os 50 e os 64 anos, e os restantes casos afetam mulheres acima dessa faixa etária. Uma em cada 11 mulheres na UE-27 desenvolverá cancro da mama antes dos 74 anos.

Neste projeto, utilizei o Conjunto de Dados Público de Cancro da Mama de Wisconsin (WBCD) do repositório de *Machine Learning* da UCI, que fornece informações muito sólidas e úteis com o único propósito de construir modelos para prever se o cancro é benigno ou maligno. O conjunto de dados com o qual trabalhei possui 32 atributos (o atributo 'ID' é omitido) e 568 instâncias.

**Descrição do Projeto:**

O projeto seguiu várias etapas, incluindo:

**Exploração de dados:** Visualização dos dados e do tamanho do conjunto de dados , análise do tipo de dados existentes no conjunto de dados e transformação da coluna “diagnosis” em 0 ou 1, representação da distribuição do cancro(maligno/benigno), estudo da relação entre os dados através de um *pairplot* da biblioteca *seaborn* .Também foi realizada uma análise da correlação dos dados usando uma matriz de correlação e um *heatmap* da biblioteca *seaborn*.

**Pré-processamento:** Eliminação da variável “Id”, identificação e eliminação de valores em falta.

**Modelos:** Os dados foram divididos em conjuntos X\_train, X\_test, y\_train e y\_test, reservando 45% dos dados para teste. Verificou-se a necessidade de escalar os dados usando o *StandardScaler* da biblioteca *sklearn*. Posteriormente, foram utilizados oito modelos: *Neural Network*, *KNeighborsClassifier, GaussianNB, SVC linear, SVC rbf, Logistic Regression, Decision Tree* e *Random Forest Classifier*.

**Resultados:** Entre os modelos avaliados, o *SVC rbf* obteve os melhores resultados, alcançando uma *accuracy* de 0.98.

**TIA(EN)**

**Breast Cancer Prediction**

**Objective:**

The aim of this project is to predict breast cancer from a dataset.

**Dataset Description:**

Breast cancer is the most diagnosed cancer in women, accounting for 1 in 4 annual cancer cases worldwide. In 2020, the incidence in the WHO Europe region was estimated at 576,300, and in the EU-27, it was 355,500. Approximately 21% of breast cancer cases in Europe occur in women under 50 years old, while 35% occur between 50 and 64 years old, and the remaining cases affect women above this age group. One in 11 women in the EU-27 will develop breast cancer before the age of 74.

In this project, I used the Wisconsin Breast Cancer Dataset (WBCD) from the UCI Machine Learning repository, which provides very solid and useful information for the sole purpose of building models to predict whether cancer is benign or malignant. The dataset I worked with has 32 attributes (the 'ID' attribute is omitted) and 568 instances.

**Project Description:**

The project followed several steps, including:

**Data Exploration:** Visualizing the data and dataset size, analyzing the types of data in the dataset, transforming the 'diagnosis' column into 0 or 1, representing the distribution of cancer (malignant/benign), studying the relationship between the data using a *pairplot* from the *seaborn* library. An analysis of data correlation was also performed using a correlation matrix and a heatmap from the seaborn library.

**Pre-processing:** Removing the 'Id' variable, identifying and treating missing values.

**Models:** The data was split into X\_train, X\_test, y\_train, and y\_test sets, keeping 45% of the data for testing. It was discovered the need to scale the data using StandardScaler from the sklearn library Subsequently, eight models were used: Neural Network, KNeighborsClassifier, GaussianNB, linear SVC, rbf SVC, Logistic Regression, Decision Tree, and Random Forest Classifier.

**Results:** Among the evaluated models, rbf SVC achieved the best results, reaching an accuracy of 0.98.