Decision Trees and Ensemble Methods

Exercise 1
This exercise aims at using Ensemble Methods on the Titanic dataset using the package `scikit-learn`.

1. Data importation and preparation

   (a) Copy the files `train_titanic.csv` and `test_titanic.csv` from the website on the course

   (b) Import the data

   (c) Describe the dataset `train_titanic.csv`. Remove the features `‘Name’` `‘Ticket’`, `‘Embarked’`,`’Cabin’` and `‘PassengerId’`

   (d) Change gender into 0 (‘male’) ou 1 (‘female’) and convert the dataframe corresponding to the train set `train_df` into a numpy array :
   ```
   train_df['Sex'].replace('male',0, inplace=True)
   train_df['Sex'].replace('female',1, inplace=True)
   ```

   (e) Impute missing values on the train set and test set using the function `fillna` of the library `pandas`

   (f) Transform the two dataframes corresponding to the train and test set into numpy arrays
   ```
   train=train_df.to_numpy()
   ```

2. Analysis of this dataset using Decision Trees

   (a) Import `sklearn` and the library `tree` of `sklearn` appropriate for decision tree

   (b) Fit a decision tree on the train set

   (c) Performance of this model on the test set?

3. Analysis of this dataset using Bagging, Random Forest and Boosting

   (a) Import `BaggingClassifier`, `RandomForestClassifier` and `AdaBoostClassifier` from the library `sklearn.ensemble`

   (b) Compare the performance of these methods.

Exercise 2

We shall work on a dataset allowing to predict if a company is in financial distress or not. This data set deals with the financial distress prediction for a sample of companies and can be downloaded at `https://www.kaggle.com/shebrahimi/financial-distress`

The column of this dataset are as follows

- First column: Company represents sample companies.
- Second column: Time shows different time periods that data belongs to. Time series length varies between 1 to 14 for each company.
- Third column: The target variable is denoted by "Financial Distress" if it is greater than -0.50 the company should be considered as healthy (0). Otherwise, it would be regarded as financially distressed (1).
- Fourth column to the last column: The features denoted by x_1 to x_{83}, are some financial and non-financial characteristics of the sampled companies. These features belong to the previous time period, which should be used to predict whether the company will be financially distressed or not (classification). Feature x_{80} is a categorical variable.

For example, company 1 is financially distressed at time 4 but company 2 is still healthy at time 14.

1. Download the dataset on the website and import it.
2. Isolate the column of labels and transform this vector into 0/1 labels
3. Analyse now the data using Decision trees, Random Forest and GBT
4. What is the importance of each variable? One may see `https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html`