ESTIN

Machine Learning

S4 2023-2024

Lab (Random Forest)

Exercise:

Consider the **breast cancer dataset** (from **sklearn**).

1- Split the dataset into training and test data with 30% as test data and **random_state**=0.
2- Create and train **RF model**. (Use RandomForestClassifier from **sklearn.ensemble**).
3- What are the hyperparameters for a random forest?
4- Compute the score for both training and testing
5- Generate confusion matrix for testing data.
6- Choose the **best hyperparameters** for this classifier. **list_max_depth**=[3, 5, 7], **n_estimators**= list(range (10, 200, 10))
7- One way to interpret the model is to see the **importance of each feature**. Type the following code:

```
Importance= pd.DataFrame ({'Importance':RFmodel.feature_importances_*100},
index=BreastData.feature_names)
Importance.sort_values (by='Importance', axis=0, ascending=True).plot(kind='barh',
color='r')
plt.xlabel('feature importance')
plt.gca().legend_=None
```

   - What are **the most important features**?
8- Do the most important features change if you choose a different number of trees in a random forest?