COMP8270
Programming for Artificial Intelligence

Class 11

Classification with Python

1. Load the following dataset:

   https://archive.ics.uci.edu/ml/machine-learning-databases/00225/Indian%20Liver%20Patient%20Dataset%20(ILPD).csv

   into a Pandas' DataFrame. Use the following column names:

```
# to make the column names nice
columns = ["Age",
           "Gender",
           "TB Total Bilirubin",
           "DB Direct Bilirubin",
           "Alkphos Alkaline Phosphotase",
           "Sgpt Alamine Aminotransferase",
           "Sgot Aspartate Aminotransferase",
           "TP Total Protiens",
           "ALB Albumin",
           "A/G Ratio Albumin / Globulin Ratio",
           "Class"]
```

2. Build a `DecisionTreeClassifier` using the DataFrame above. You might need to investigate whether to apply any pre-processing step or not. Remember that the 'Class' attributes should not be used as a predictor attribute.

3. As we discussed in the lecture, to reliably evaluate the quality of a classifier we need to divide the data into training and testing. Using the `train_test_split` function, create and evaluate a DecisionTreeClassifier. Repeat the evaluation 10 times using different `random_state` values and calculate the average performance.

4. We can use the same data to model a different classification problem. For example, try to predict the 'Gender' based on the different set of predictors. Train a classifier to predict the 'Gender' value, measuring its performance.

5. Repeat Task 3, this time using a `RandomForestClassifier`. Compare the results obtained by the RandomForest against the DecisionTree.