Student: Anton Kucherenko, 2103787

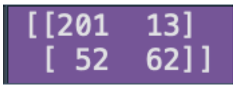
Week project: Week 39, Decision Tree & Random Forest

The titanic data set includes 4 columns and 1312 rows. The written code analyses the data of all the people who died or stayed alive after the crash of the Titanic, by their age, PClass, Gender and at the end gives a prediction whether a person with a certain characteristic would be dead or alive. In addition, it also shows a tree, the values of which are preceded by training a data, and the tree itself depicts a 16-level layout, which clearly shows the criteria by which the program determines whether a person died or survived.

In the given task, I needed to test the data with DecisionTreeClassifier and then with RandomForestClassifier in order to see the difference in the results.

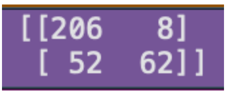
Beforehand, i want to emphasize that Accuracy of RandomForestClassifier is higher, 0.813 – 0,84 compare to DecisionTreeClassifier accuracy rate of which is about 0,787.

DecisionTreeClassifier:



```
[[201  13]
 [ 52  62]]
```

RandomForestClassifier:



```
[[206   8]
 [ 52  62]]
```

(TP(1,1) – true positive, TN(0,0)- true false, FP(1,0) – false positive, FN(0,1) – false negative)

Looking at the confusion matrix of both Classifiers, we can see that TP, TN, FN, FP values varies close to each other but, of course, using RandomForestClassifier would be better decision as, for instance, it has FP value which is twice lower.

Besides Accuracy rate, the program also calculated Precision and Recall.

- "Precision gives the proportion of positive predictions that are actually correct. It takes into account false positives, which are cases that were incorrectly flagged for inclusion."
- "Recall measures the proportion of actual positives that were predicted correctly. It takes into account false negatives, which are cases that should have been flagged for inclusion but weren't."

Therefore, looking at precision value in both cases it is not bad, quite good I would say, the value fluctuates around 0,85.

However, Recall value in both cases is very doubtful, the value fluctuates around 0,55, which tells us of low accuracy in terms of the ratio of TP to FN (≈ 62 people that are actually survived to **≈ 52** of those who are actually dead but the program considered them as survived)

As I mentioned above, the program also visualizes the tree, and I just would say that in our case when we have quite big data it will take a while if you want to find whether a person will be dead or not, just by looking at the tree and moving your finger to the end of the tree through all "if methods". But it seems to be correct.

To determine if I built the program correctly, it tests two people at the end using prediction.
So, as for an example, I created two test subjects (Rose and Jack from Titanic) where it should tell that Rose survived and Jack did not, whether we use RandomForestClassifier or DecisionTreeClassifier, it shows correct result.

Summarization:

In my opinion, the program works fine, perhaps not perfectly, but for some non-100% accurate purposes it is enough. Anyway, here we are dealing with machine learning, the result does not always turn out perfectly, unless you are putting a lot of efforts in it and developing a super intelligent algorithm.