HOMEWORK #3

* Alexander Shender: 328626114
* Samuel Panzieri: 336239462

## Question 1.

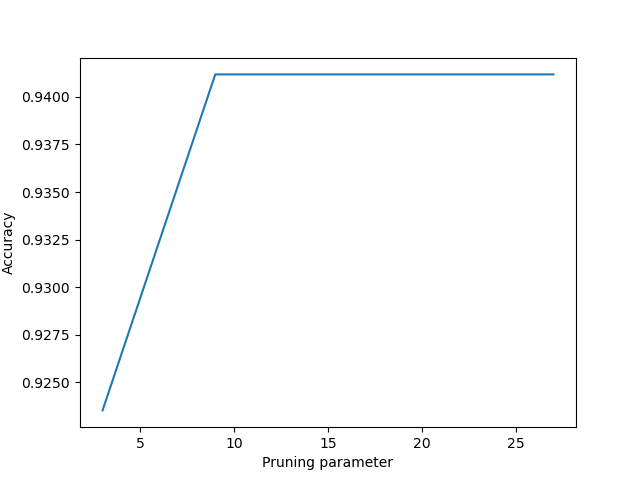
**FILL**

## Question 2.

The accuracy that was received is 0.9235

## Question 3.

The pruning was implemented, and the following graph was obtained for difference x parameters:



Conclusions:

1. We can see that certain pruning does improve the accuracy of the decision tree, and reduces the over-fitting
2. Although it wasn’t required to display, if we increase the pruning parameter even further, the accuracy will eventually start decreasing

## Question 4.

**FILL**

## Question 5.

**FILL**

## Question 6.

The code was implemented, and the accuracy received for Epsilon-Decision-Tree with depth limit of 9 is 0.8647. It is worth noting that if the default value for the case where the classification list (list of classification of leaf nodes that were reached) is 0 instead of 1, the accuracy will increase to 0.89, because most of the samples from the test set have classification False.

More conclusions:

* This algorithm doesn’t account for the number of examples in each leaf node, giving each node similar weight when the final classification is calculated. This may be the reason for the poor performance

## Question 7.

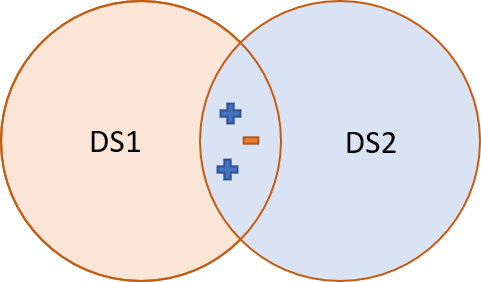
This statement is **FALSE**.

The proof:

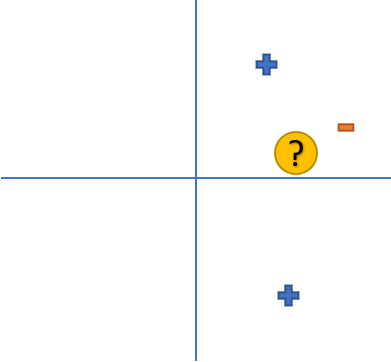
Let’s suppose we have 3 examples in our sets:

* Examples 1, 2, 3 appear in BOTH and

Thus, they appear in both of the sets: and ( once in each of those sets).

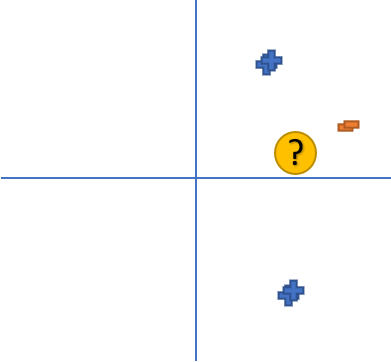


Let’s say for simplicity it’s 2D. The new test example is marked with “?” sign and we use the 3NN to classify it:



Using the 3NN method, it will be classified as TRUE for both : and

But when we classify it with the group, we receive the following:



(In this example the points are in the SAME location, but are slightly moved to show that they appear twice).

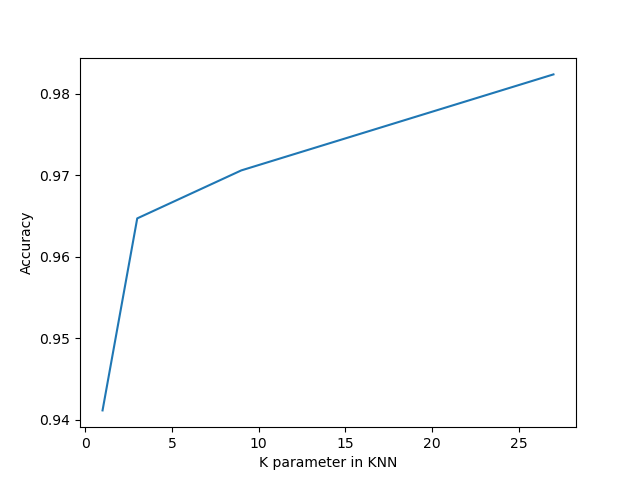
Now, the 3 closest points include two FALSE classifications, which means, the 3NN classification will be FALSE. Thus, this proves the statement is wrong.

## Question 8.

The code was implemented, and the result for k=9 is 0.9706

## Question 9.

The result:



Conclusions:

1. We can see that increasing the K parameter the accuracy increases as well. Taking more neighbors into account helps for the classification (until certain level)
2. We can see again that we haven’t reached the ‘peak’ of the accuracy. Increasing the K values even more will eventually decrease the accuracy, since more distant samples will be taken into account

## Question 10.

**FILL**

## Question 11.

The code is implemented. The accuracy that was reached is 0.9411.

Conclusions:

* We can see improvement over the simple Epsilon-Decision-Tree. Indeed, now the size of the samples at each node are taken into account
* The regular KNN classifier with appropriating K parameter still gives better result

## Notes:

* The required code files are submitted, but also the file ID\_total.py. This file includes the classifiers for questions 2,3,6,11.
* To save time, the trained tree classifiers were saved as “.pkl” files and are loaded upon need to classify new data. Using the parameter “FORCE\_NEW” as 1 will allow the algorithm to load the classifiers that were generated before.
* In files DT.py, DT\_epsilon.py, KNN\_epsilon.py : all the classes and functions are GENERIC and defined in DT\_total.py