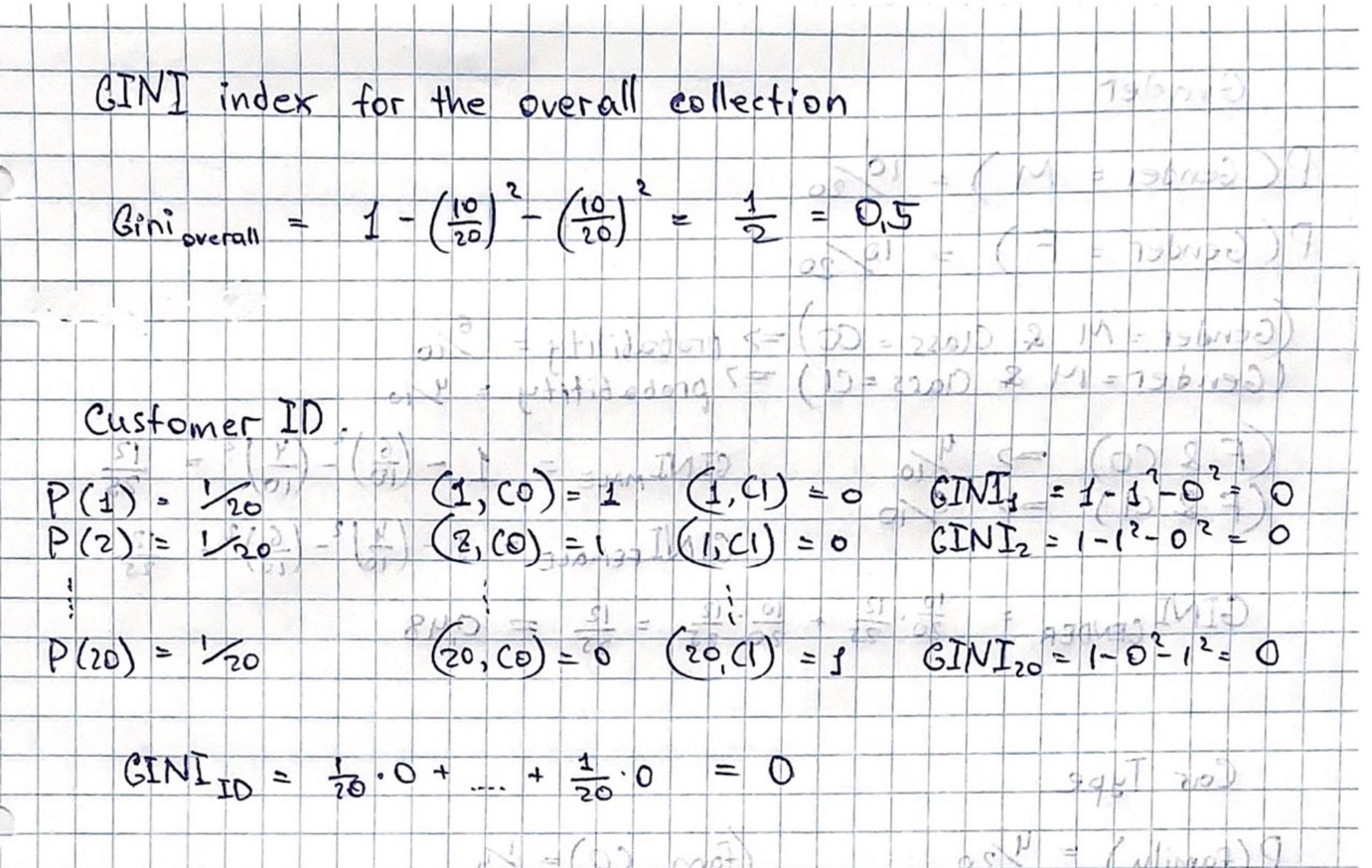
Assignment – Decision Trees

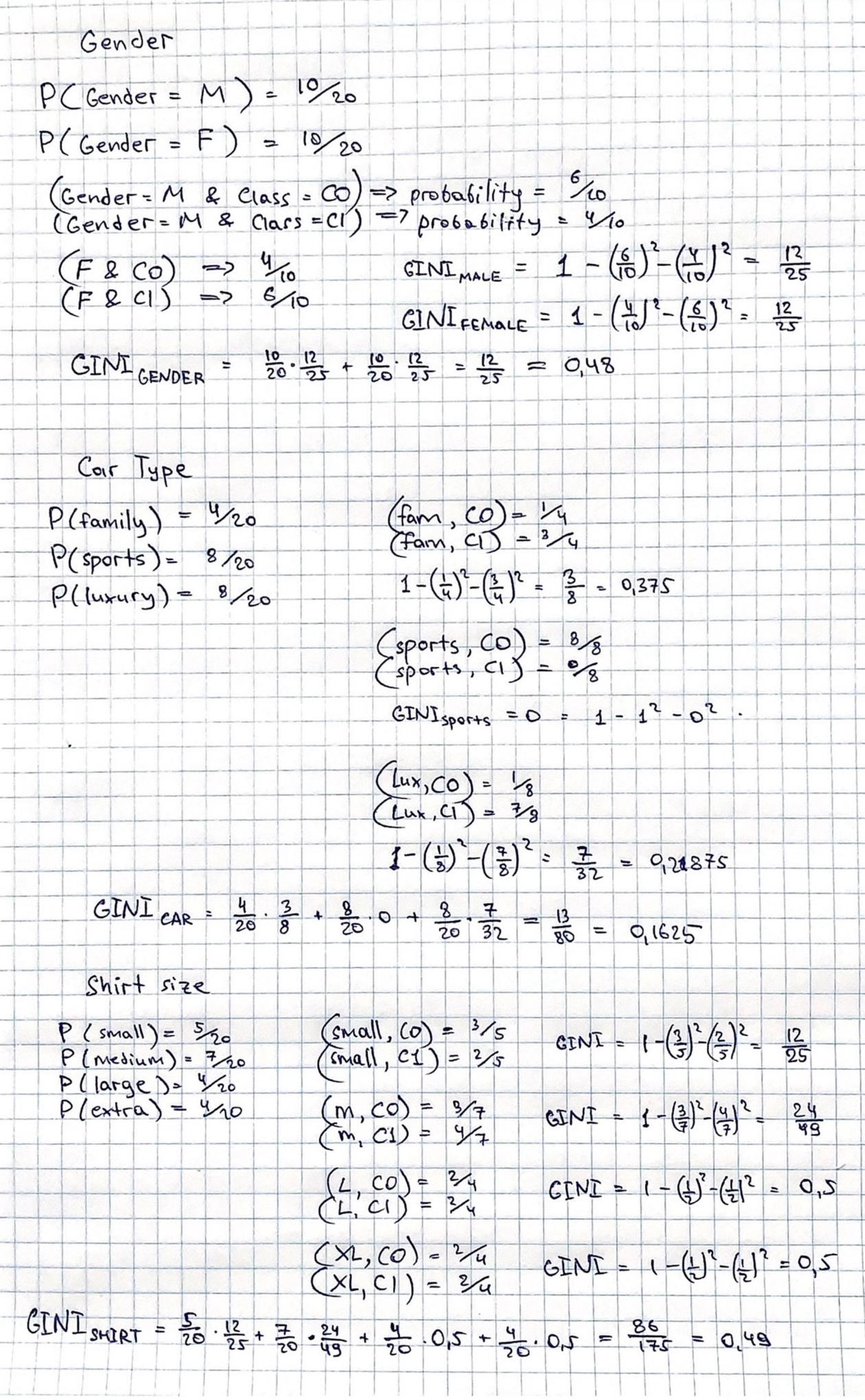
Rogov Sergei

Activity 1

1. b)



1. d) e)



f) Car Type is the best attribute since it has the lowest Gini index (0.1625) from these three attributes. [Gender is 0.48 and Shirt Size is 0.49]

g) Every customer is assigned to a unique ID number, that is why this attribute should not be considered. Customer ID attribute is not predictive.

Activity 2

Wine dataset

Report on accuracy with default parameters:

Accuracy (correct/total) = 0.9074074074074074

Scikit accuracy = 0.9074074074074074

Confusion Matrix

[[18, 1]

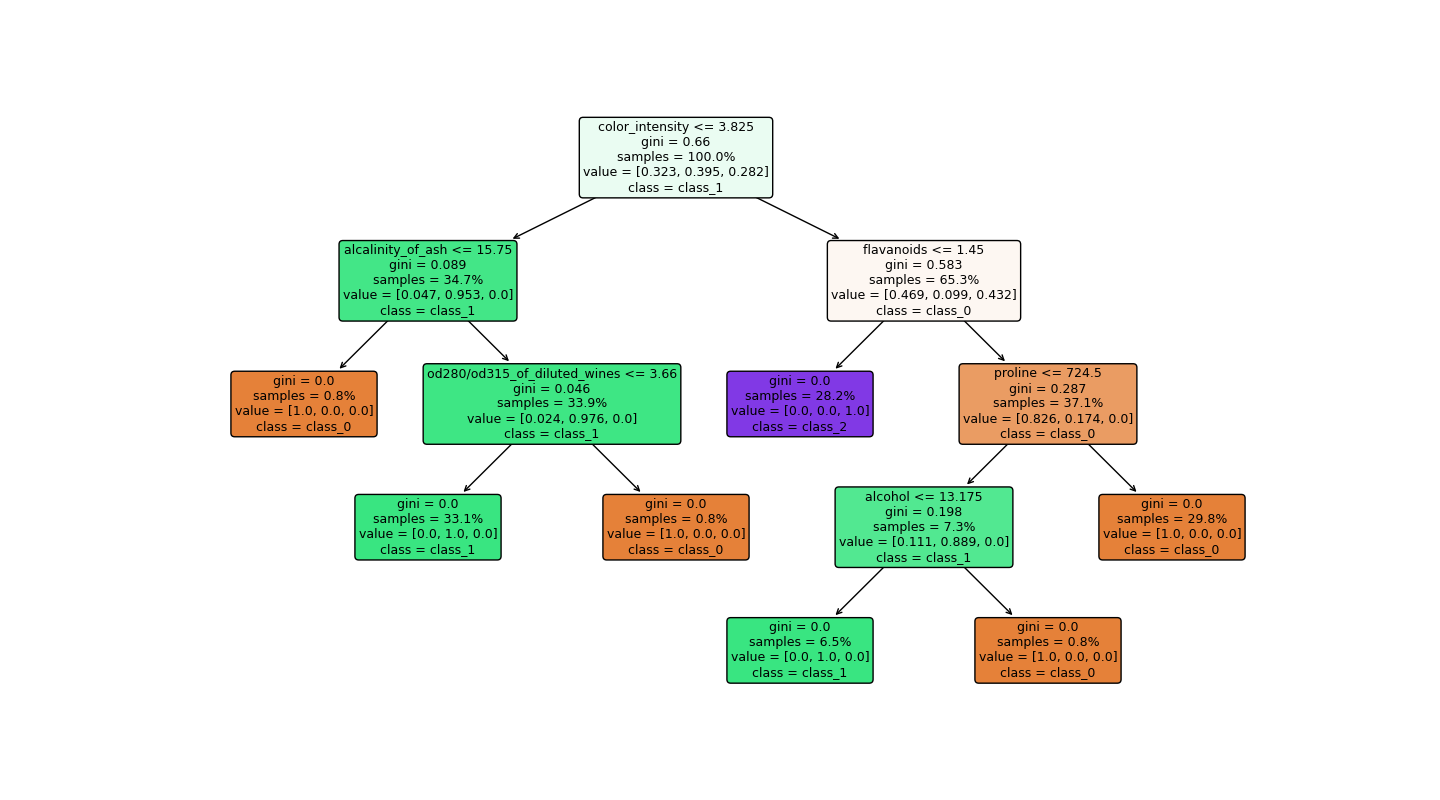
[ 2, 19]]

precision recall f1-score

0 0.86 0.95 0.90

1 0.95 0.86 0.90

2 0.92 0.92 0.92

Decision tree itself: 

IMPORTANCE OF ATTRIBUTES:

flavanoids 0.415881

color\_intensity 0.376051

proline 0.139755

od280/od315\_of\_diluted\_wines 0.023852

alcalinity\_of\_ash 0.022742

alcohol 0.021719

malic\_acid 0.000000

ash 0.000000

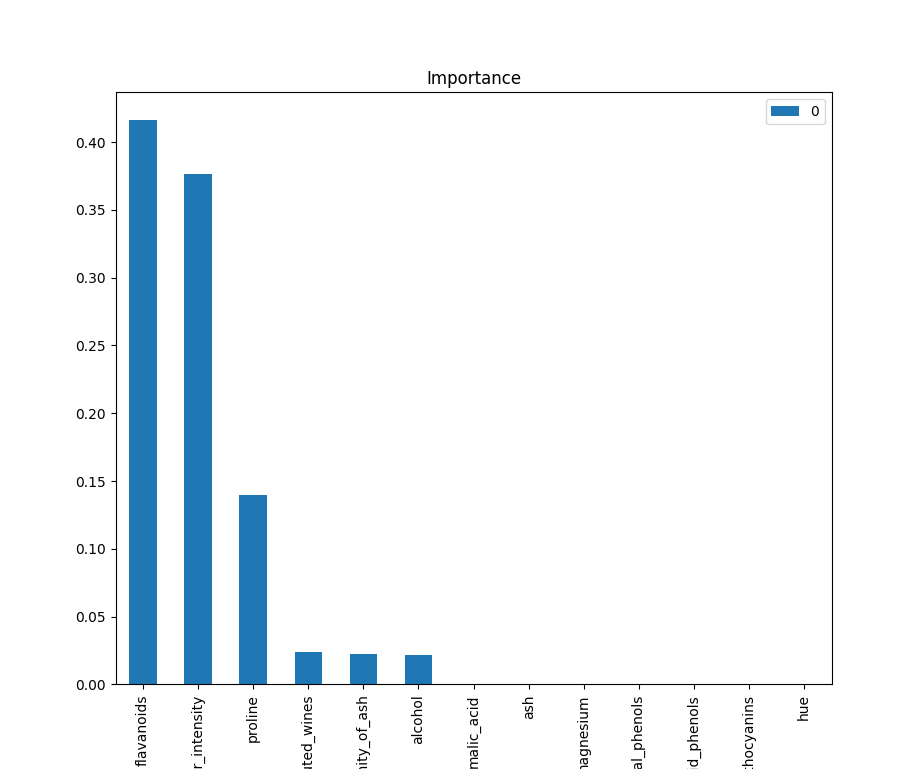
magnesium 0.000000

total\_phenols 0.000000

nonflavanoid\_phenols 0.000000

proanthocyanins 0.000000

hue 0.000000



Experimenting with parameters:

I varied ccp\_alpha parameter which is responsible for pruning the tree from 0 to 0.1 with a step of 0.005. The resulting plot looks like this:

The reason why the accuracy at ccp\_alpha=0 does not equal 0.907 (as with default parameters) is because for splitting dataset into train and split sets I used a built-in train\_test\_split Scikit Learn function. It forms these groups differently every other time, that’s why the model is trained differently and is not exactly the same every time.

