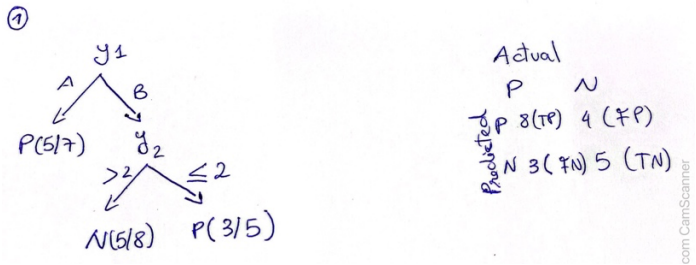
Aprendizagem 2021/22

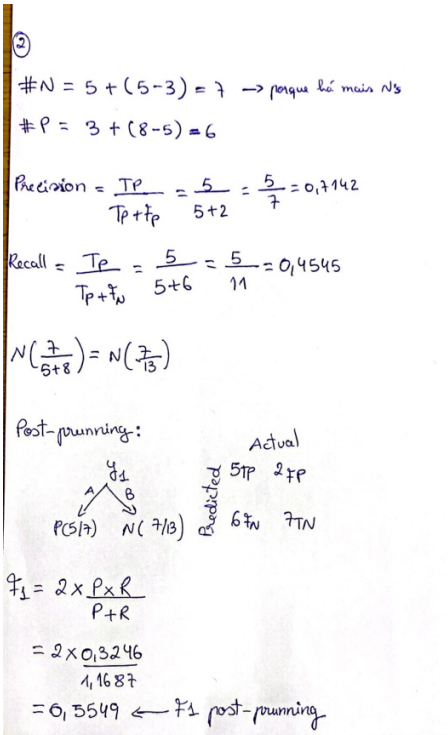
Homework I – Group 044

I. Pen-and-paper

1) Answer 1



2) Answer 2



3) Answer 3

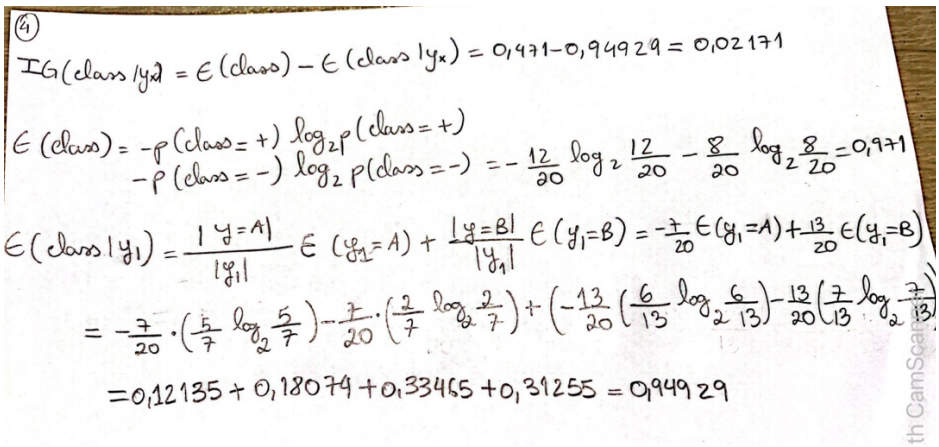
The branches on the left could have a very low probability of happening, so by representing them we would make the tree more prone to overfitting because the model would be very specific.

Another possible reason would be that when we split the left node, the entropy would not improve significantly to justify the split.

Aprendizagem 2021/22

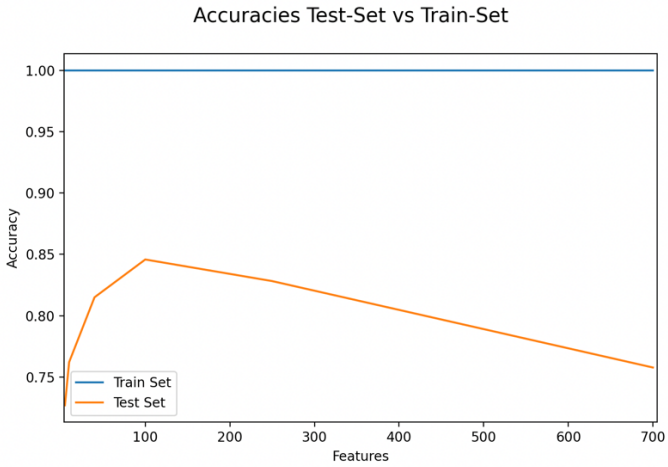
Homework I – Group 044

4) Answer 4



II. Programming and critical analysis

5) Answer 1



6) Answer 2

The accuracy is always 1 in the training set due to the fact that overfitting has occurred, this is confirmed by the testing set which demonstrates that the performance of the created model was not good despite the optimal training accuracy.

III. APPENDIX

```python
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeClassifier
import pandas as pd
from sklearn.feature_selection import SelectKBest
from scipy.io import arff
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import mutual_info_classif


df = pd.DataFrame(arff.loadarff("pd_speech.arff")[0])

X = df.iloc[:,:-1]
y = df.iloc[:,-1].astype(int)

features = [5,10,40,100,250,700]
train = list()
test = list()
Xtrain, Xtest, Ytrain, Ytest = train_test_split(X,y,test_size=0.3,random_state=1)
mutual_inf_classifier = mutual_info_classif(Xtrain,Ytrain)
mutual_inf_classifier = pd.Series(mutual_inf_classifier)
mutual_inf_classifier.index = Xtrain.columns

def scoreFunc(X, y):
    return  mutual_info_classif(X, y, random_state=1)

for feature in features:
    selectFeat=SelectKBest(score_func=scoreFunc,k=feature)
    selectFeat.fit(Xtrain, Ytrain)
    xTrain2 = Xtrain.loc[:,Xtrain.columns[selectFeat.get_support()]]
    xTest2 = Xtest.loc[:,Xtest.columns[selectFeat.get_support()]]
    classifier = DecisionTreeClassifier(random_state=1).fit(xTrain2, Ytrain)
    train.append(classifier.score(xTrain2,Ytrain))
    test.append(classifier.score(xTest2,Ytest))


fig, ax = plt.subplots()
fig.set_size_inches((8, 5))
ax.set_xlim(3,710)
fig.suptitle('Accuracies Testing-Set vs Training-Set', fontsize=15)
plt.plot(features, train, label='Training Set')
plt.plot(features, test, label='Testing Set')
ax.set_xlabel('Features', fontsize=8)
ax.set_ylabel('Accuracy', fontsize=8)
plt.xticks(fontsize=8)
plt.yticks(fontsize=8)
ax.legend(fontsize=8)

plt.show()
```

TÉCNICO LISBOA

Aprendizagem 2021/22

Homework I – Group 044

END