Part A

Part B

Question 1

from sklearn.model\_selection import train\_test\_split

import pandas as pd

# Read the data into a pandas dataframe

data = pd.read\_csv('pima-indians-diabetes.data.csv',delimiter=',')

#Remove the target variables for x. Y only has the target variables

x = data.drop("Class variable (0 or 1)", axis=1)

y = data["Class variable (0 or 1)"]

#Split it. Training has 70% of the records. Test has 30%

X\_train, X\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.3)

#Combine the target and features together

train = pd.concat([X\_train, y\_train], axis=1)

test = pd.concat([X\_test, y\_test], axis=1)

#save the two sets to csv files

train.to\_csv("train.csv")

test.to\_csv("test.csv")

The csv files were download and split. And then saved to two csv files. This is done to keep the data consistent, and therefore ours results comparable. Code shown above

Step 1

|  |  |
| --- | --- |
| max\_depth | Test accuracy |
| 2 | 0.7186147186147186 |
| 4 | 0.70995670995671 |
| 6 | 0.7229437229437229 |
| 8 | 0.7012987012987013 |
| 10 | 0.6926406926406926 |
| 12 | 0.683982683982684 |
| 14 | 0.696969696969697 |
| 16 | 0.6796536796536796 |
| 18 | 0.696969696969697 |

Step 2