COMP40370 Practical 4

Ryan Jennings 19205824

Question 1: Simple Linear Regression

1)

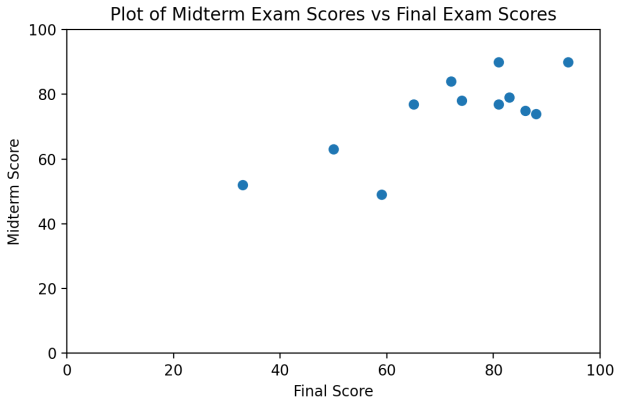

Figure 1. Plot of Midterm Exam Scores vs Final Exam Scores

The two attributes when plotted seem to have a linear relationship. While the number of data points (n) is still quite low to be certain we can definitely see an increasing positive trend with no outliers in the plot. The students who scored highly on the midterm were more likely to score well on the final exam. This can be further seen with a line of best fit plotted through the data which we can see in the next question.

2) A model can be simply generated with `sklearn.linear_model`'s `LinearRegression`, given our dataset is read into the variable `df_1`, with `lin_reg = LinearRegression().fit(df_1['midterm'].values.reshape(-1, 1), df_1['final'])`
The attribute 'midterm' needs to be reshaped to fit the expected training data for our model.
From our model we can then find the coefficient and intercept to describe the model
`lin_reg.coef_` is set to `array([0.58160008])` and
`lin_reg.intercept_` is set to `32.027861081551706`
Using these values we can create the line of best fit we talked about in Q1.1 and see the positive relationship

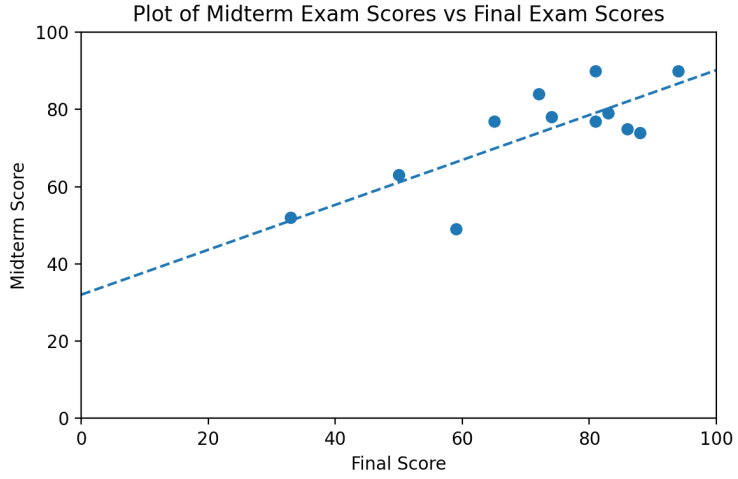

Figure 3. Line of best fit through scatter plot from Figure 1

3) If we plug in the midterm score of 86 into our model with the line `lin_reg.predict([[86]])`
We get back the value of `82.04546774` that the student will receive for their final exam score.

Question 2: Classification with Decision Tree

1) Just as in the previous labs. To remove an attribute from our dataframe we just need to run `del df_2['TID']`
2) To keep to page length please see the explanation of the code in the run.py file.

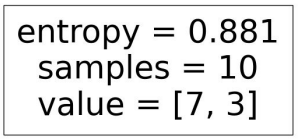

Figure 3 shows the one leaf tree due to the minimum impurity decrease being 0.5.
Our one leaf tree shows us the entropy, the fact that we have 10 values and that 7 of them can be split into 0.0 (No to DefaultedBorrower) and 3 into 1.0 (Yes to DefaultedBorrower)
3) This time for question 2.3 we switched to a minimum purity decrease of 0.1

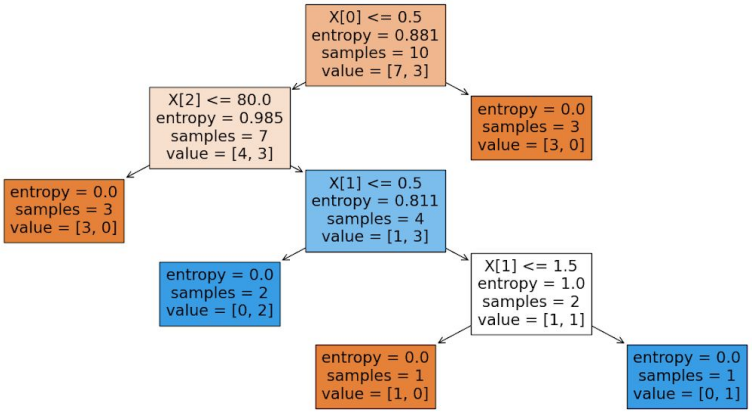

Figure 4. Shows how we have much more leaves for our tree.
We can see now that there are actually decisions seen by X[0] <= 0.5 etc.
We are able to split this data into more segments to classify them.
e.g. on the far left 3 people didn't default and they do not own a home and they have an annual income of less than 80.
4) The second tree had more leaves but that was a symptom of our low number of sample data points rather than the high threshold for impurity. With more values we could get a better look at the data and not unnecessarily split the data too much as seen in Figure 4 where if it was scaled up, there would be far too many leaves. Still it gives us good information on determining whether somebody is able to pay their loan based on their current status. Interestingly as mentioned in 2.3 that poorer individuals with lower income and no house ownership are less likely to default.