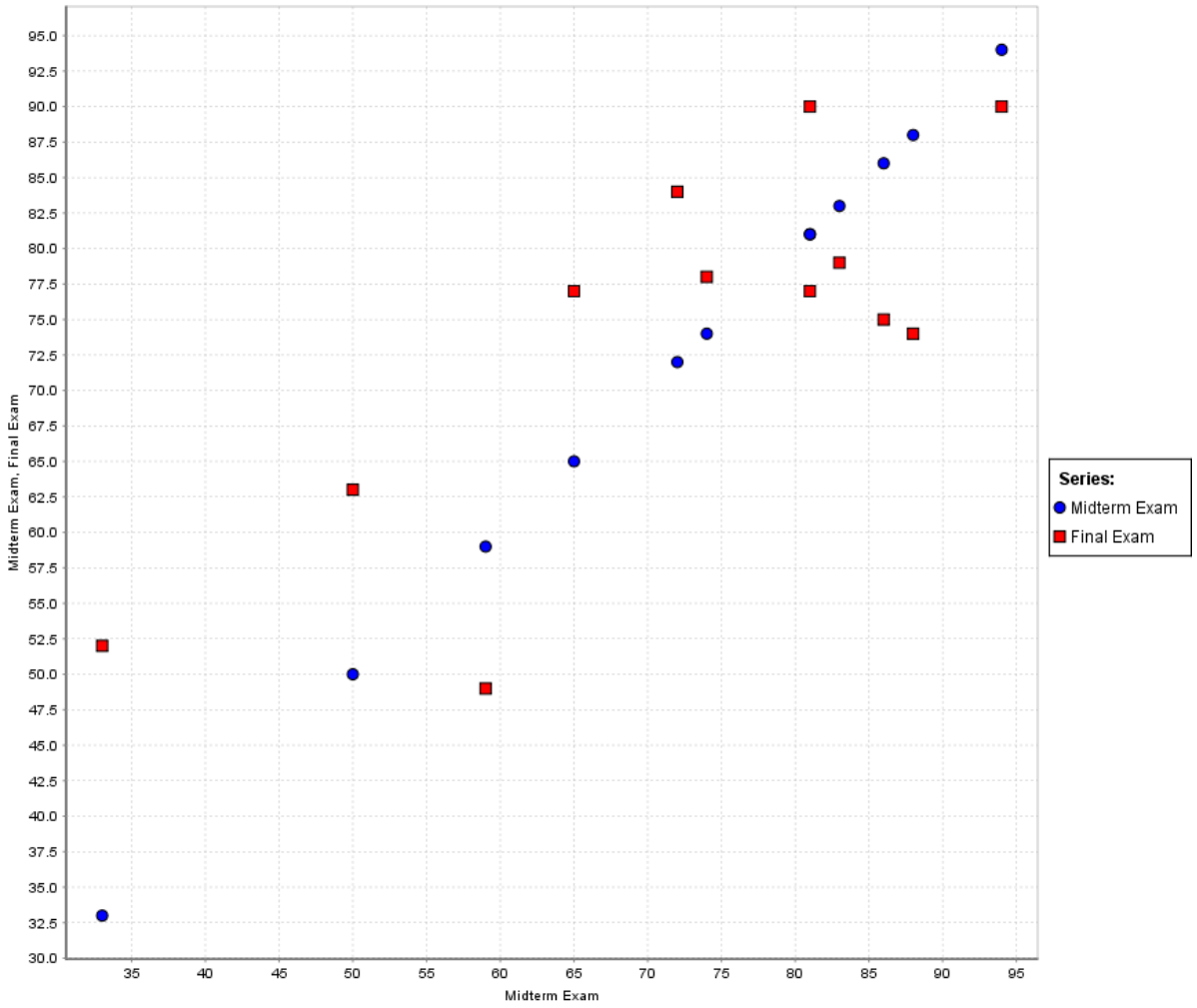
**COMP40370 – Data Mining – Practical 3**

**Student: Felipe Guth  Student Id: 14210231**

**Question 1.**

1.1 -




Although, some cases seems to be linear, when all instances of data are taken into account it is not possible to affirm that the variables Midterm Exam and Final Exam have a linear relationship between them.

1.2 –

a) The rapid miner Linear Regression operator was applied instead of applying the W-SimpleLinearRegression operator (not contained on actual version of rapid miner for default).

# LinearRegression

```
0.582 * Midterm Exam
 + 32.028
```

b) 0.582 * 86 + 32.02 = **82.072**

**Question 2**

2.1-

# PolynomialRegression

```
  1.030 * MCQ1 ^ 1.000
+ 0.719 * MCQ2 ^ 1.000
- 55.839
```

I haven't received my MCQS results yet, but assuming values of 85 and 80 the predicted final mark result is:

1.030 * 85 ^ 1.000 + 0.719 * 80 ^ 1.00 – 55.839 = **89.231**

2.2-

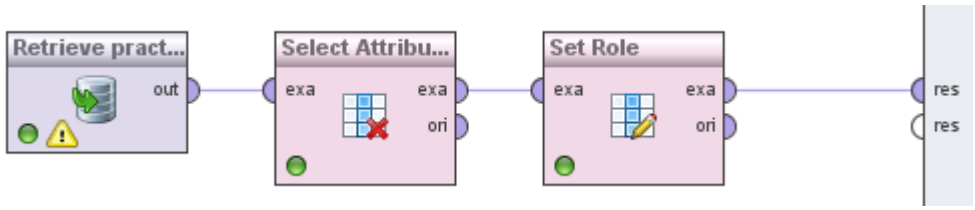
# PolynomialRegression

```
  0.498 * MCQ1 ^ 1.000
+ 0.042 * MCQ2 ^ 1.000
+ 29.913
```

The equation resulted in the question 2.2 is different to the equation generated in 2.1 given that the use of random data was used. The random data introduces unobserved random data that adds noise to the linear regression, thus, generating a slightly different equation.

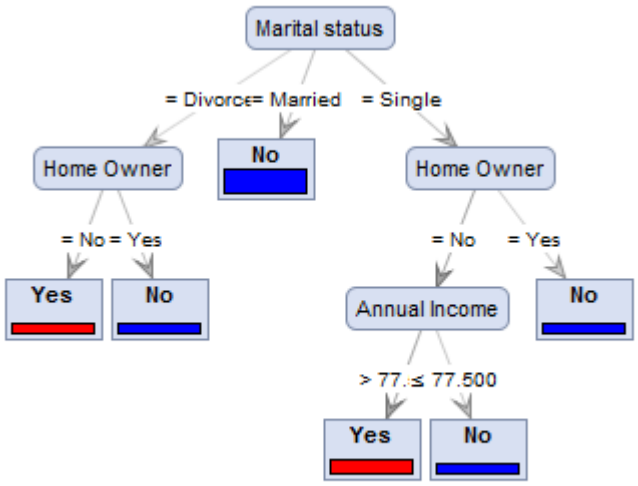
**Question 3 (CROSS VALIDATION = 5)**

3.1 Filter TID attribute.




3.2 Decision tree with Information Gain.

The decision tree took as decision attributes firstly marital status and secondly home owner. The tree generated pure leaf nodes whereas there is only examples of a particular class in each node.




The model was evaluated in 5 fold cross validation scheme. The overall accuracy was of 60% with standard deviation of +/- 20%. Following are showed the confusion matrix, precision and recall statistics, generated by rapid miner.

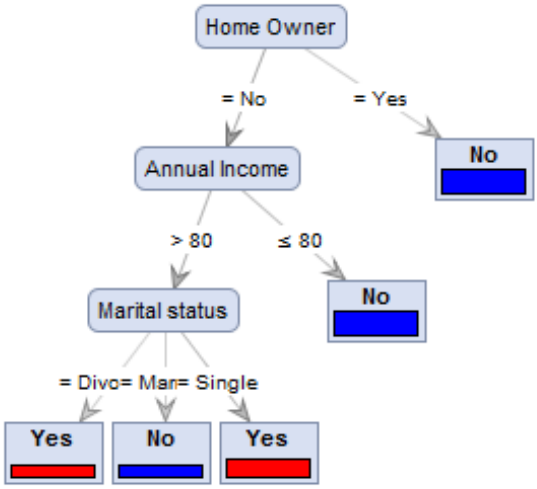
Table View  Plot View

**accuracy: 60.00% +/- 20.00% (mikro: 60.00%)**

| | true No | true Yes | class precision |
|---|---|---|---|
| pred. No | 4 | 1 | 80.00% |
| pred. Yes | 3 | 2 | 40.00% |
| class recall | 57.14% | 66.67% | |

3.3 – Decision tree with gain ratio.

Using the gain ratio method, the same attributes were selected as 3.2, but in a different order to separate the leaves of the tree. Again, pure nodes were produced.




Using a scheme of 5 fold cross validation the overall accuracy was of 70% with a standard deviation of +/- 24.49%. The classification results as in the previous case, are not optimal. Apparently, there is no enough data to train the tree in a proper way, which leads to a faulty classifier with poor generalization capability. The rapid miner statistics and the confusion matrix are showed bellow.

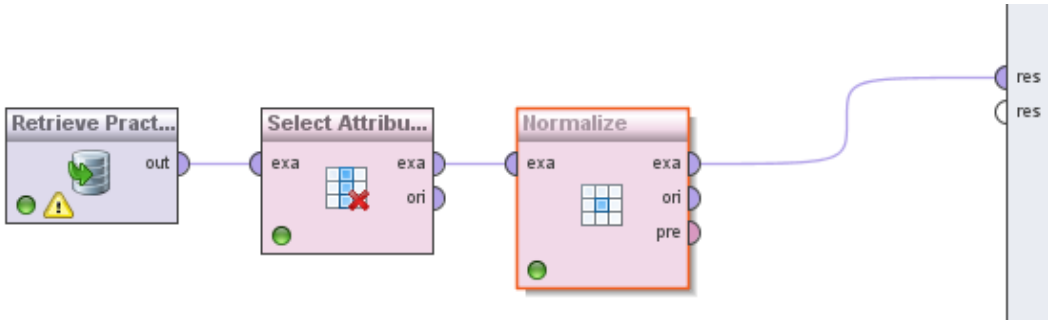
Table View  Plot View

**accuracy: 70.00% +/- 24.49% (mikro: 70.00%)**

| | true No | true Yes | class precision |
|---|---|---|---|
| pred. No | 5 | 1 | 83.33% |
| pred. Yes | 2 | 2 | 50.00% |
| class recall | 71.43% | 66.67% | |

**Question 4 – CROSS VALIDATION 10**

4.1 – Selecting attributes and normalizing numerical features [0-1].




Data transformation.

4.2 –Decision Tree, Gini-index (Default Parameters)

Using the default parameters as shown next:

| Decision Tree | |
|---|---|
| criterion | gini_index |
| maximal depth | 20 |
| ☑ apply pruning | |
| confidence | 0.25 |
| ☑ apply prepruning | |
| minimal gain | 0.1 |
| minimal leaf size | 2 |
| *minimal size for split* | 4 |
| *number of prepruning alternatives* | 3 |

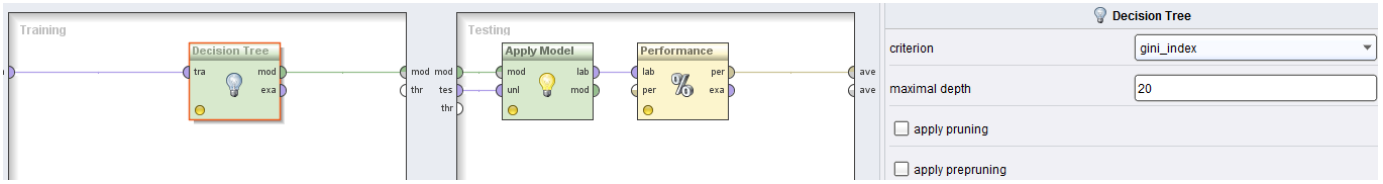
The obtained results clearly show that the tree is biased to the “False” class that had a precision of 85% and recall of 99% whereas the “True” class had a precision of 46 % and recall of just 2.69%. The overall accuracy is of 85.45%. The results were obtained using a 10 fold cross validation scheme. The statistics of rapid miner and the confusion matrix are shown in the next image.

Table View  Plot View

**accuracy: 85.45% +/- 0.26% (mikro: 85.45%)**

| | true False. | true True. | class precision |
|---|---|---|---|
| pred. False. | 2835 | 470 | 85.78% |
| pred. True. | 15 | 13 | 46.43% |
| class recall | 99.47% | 2.69% | |

4.3 – Decision Tree, Gini-Index (no pruning)




Using the gini-index with no pruning the results were improved. Although, the overall accuracy was lessen, the precision and recall results of the class “True” improved. This shows a less biased classifier compared to 4.2. The confusion matrix, showed below, depicts the improved results in the prediction of the class “True” while compared to the previous question. The results were obtained with a 10 fold cross validation training and testing.

Table View  Plot View

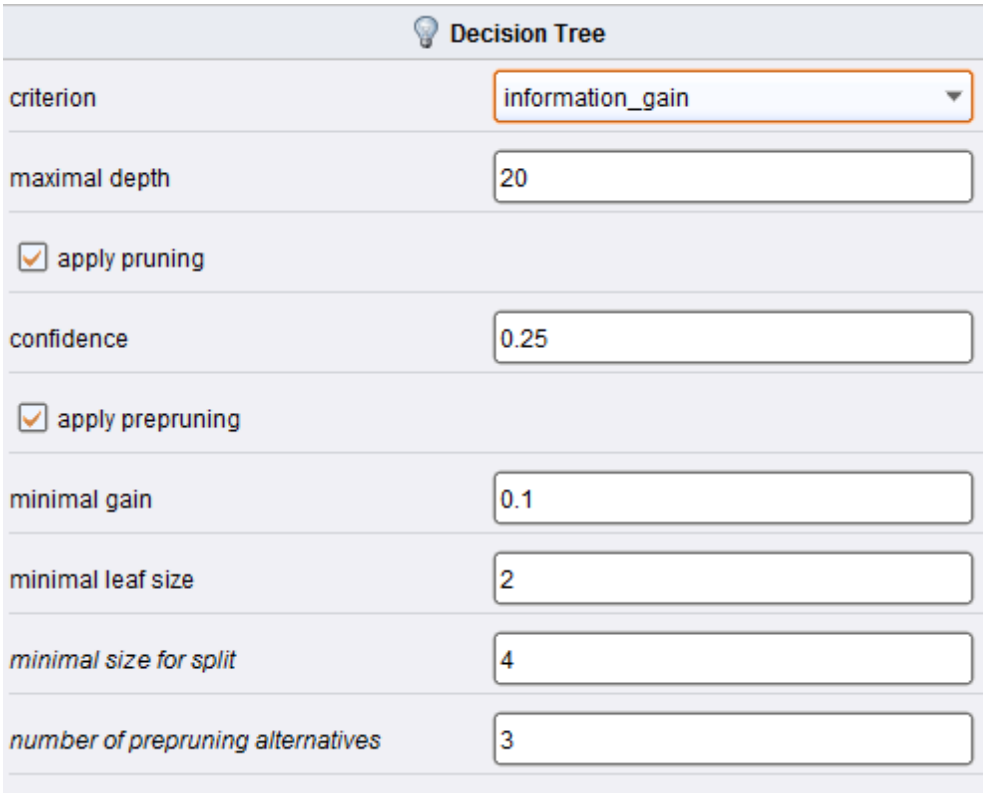
**accuracy: 81.19% +/- 1.41% (mikro: 81.19%)**

| | true False. | true True. | class precision |
|---|---|---|---|
| pred. False. | 2607 | 384 | 87.16% |
| pred. True. | 243 | 99 | 28.95% |
| class recall | 91.47% | 20.50% | |

4.4 –

Decision tree with information gain and default parameters.

Using the default parameters as shown next:




The results were the following.

Table View  Plot View

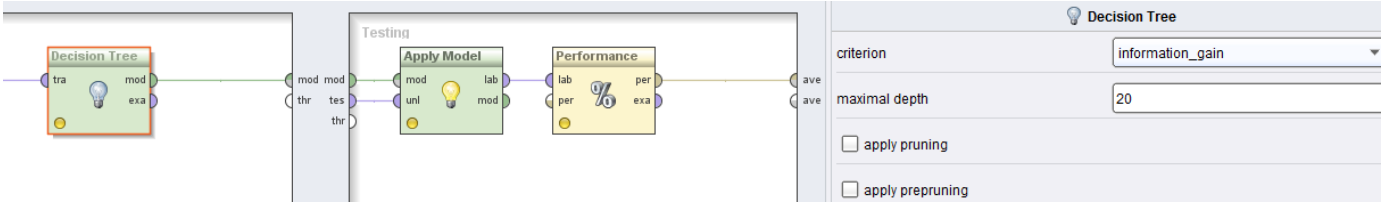
**accuracy: 85.51% +/- 0.12% (mikro: 85.51%)**

| | true False. | true True. | class precision |
|---|---|---|---|
| pred. False. | 2850 | 483 | 85.51% |
| pred. True. | 0 | 0 | 0.00% |
| class recall | 100.00% | 0.00% | |

Compare the classification results with the results of Question 4.2, it can be noticed a high overall accuracy of 85.51%, similar to the overall accuracy of 85.45% of 4.2. Again, the decision tree is totally biased through the class “False” with a recall of 100% and prediction of 85.51% compared to 99.47% and 85.78% of question 4.2. For the class “True”, the recall and precision were even worst compared to the ones obtained on the question 4.2, with 0% recall and 0% prediction. The reason of this is due to the fact that the dataset contains a majority of the “False” class. The 10 fold cross validation method was use to evaluate the classifiers in both cases.

Select no pruning, discuss the classification results.

No pruning:



The results obtained with no pruning can be consider better, even though the overall accuracy decreased the class prediction and recall of the class “True” were much better than previously. The recall and precision of class “True” are respectively 19.88% and 29.27%, much better compared to 0% obtained with pruning. The class “False” had a recall and precision of 91.86% and 87.12%. The overall accuracy of the decision tree classifier was of 81.43%.

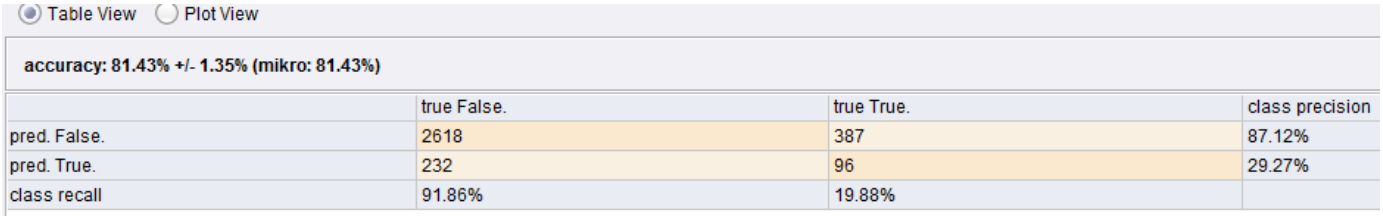


Table View Plot View

**accuracy: 81.43% +/- 1.35% (mikro: 81.43%)**

| | true False. | true True. | class precision |
|---|---|---|---|
| pred. False. | 2618 | 387 | 87.12% |
| pred. True. | 232 | 96 | 29.27% |
| class recall | 91.86% | 19.88% | |