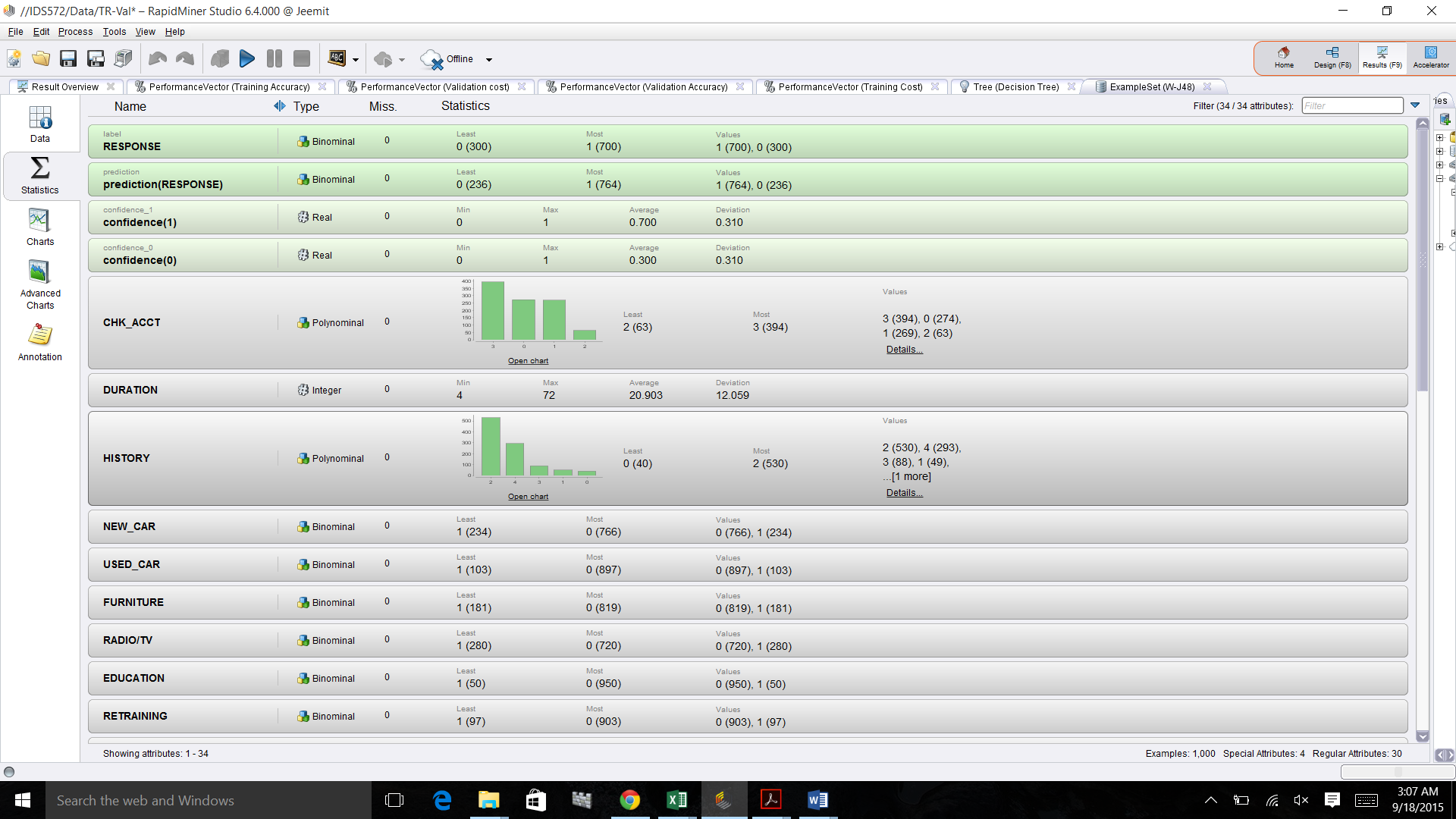
Assignment-1 (Decision Trees)

Made By: - Jitendra Patel

Q1) Proportion of Good to Bad in given data is 70:30 initially. There are 700 cases with response 1 (Good) and 300 Cases with response 0 (Bad).

**Mean, Standard Deviation and related Frequency distribution of the Attributes.**



**Noteworthy Points in Data:-**

Foreign Residents have lesser rejection rate. (App. 5-10%)

If the Age of applicant is greater than 45 than rejection rate is Low.

If the amount is more than 8000 than the rejection rate is almost 50%.

If the duration is more than 40 months than the rejection rate is more than 50%.

**Note:-**

**In Given data of 1000 cases, there are 56 cases in which the purpose of credit is not defined**

**(i.e. neither of New car, used car, furniture, radio/TV, education, retraining)**

**Variables which seems to be important and relevant:-**

**Amount, Age, Duration, History, Check-in account, job, savings account.**

Q2) **Variables used to determine good or bad cases:-**

**Amount, Age, Duration**

With different type of decision tree and by changing the parameters it showed that the initial stages of separation

are those mentioned variables.

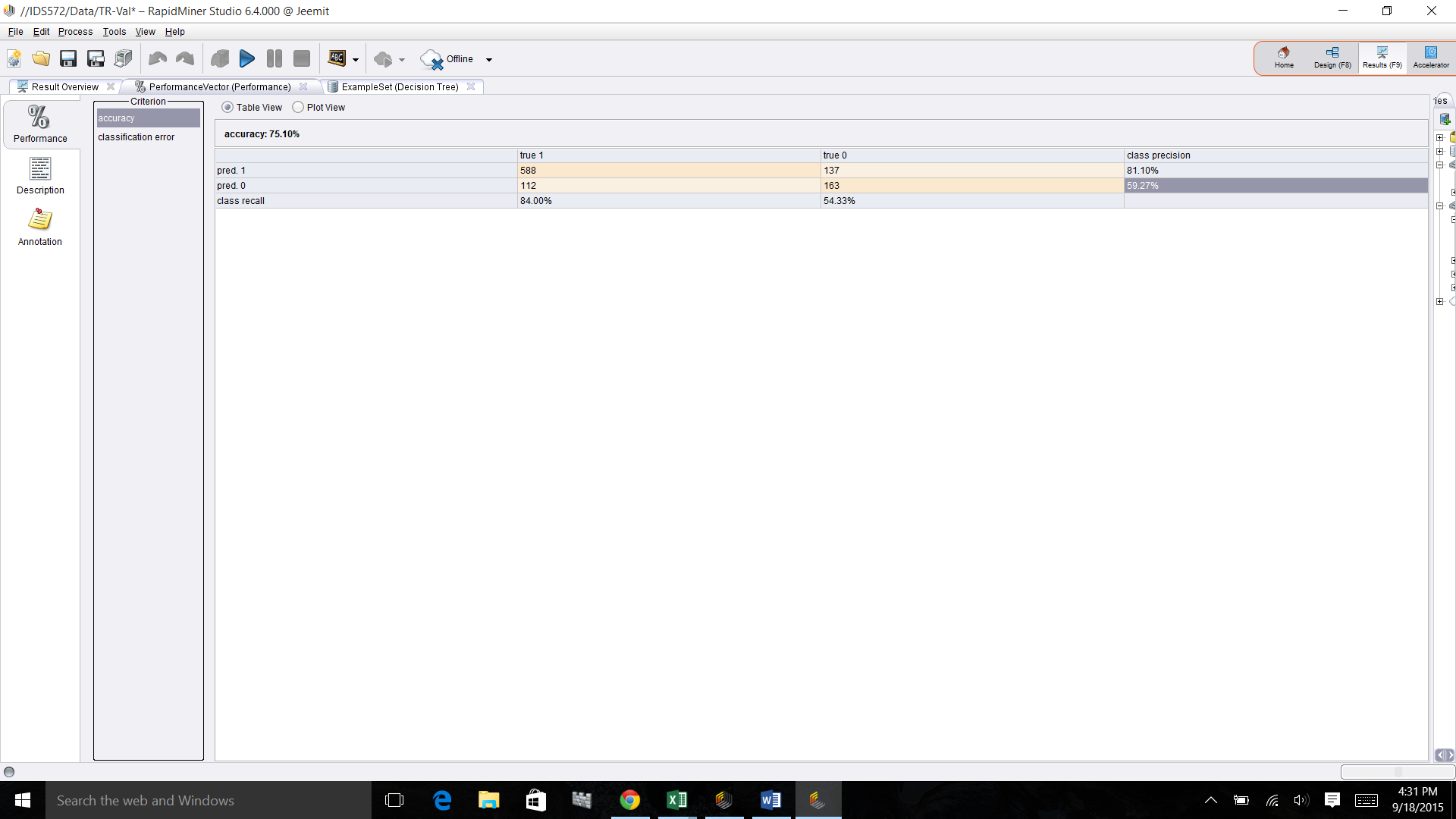
**Accuracy obtained**

Overall Accuracy=75.10%

|  |  |  |  |
| --- | --- | --- | --- |
|  | true 1 | true 0 | class precision |
| pred. 1 | 588 | 137 | 81.10% |
| pred. 0 | 112 | 163 | 59.27% |
| class recall | 84.00% | 54.33% |  |

Accuracy of good cases=585/700=83.57%

Accuracy of bad cases= 163/300=54.33%



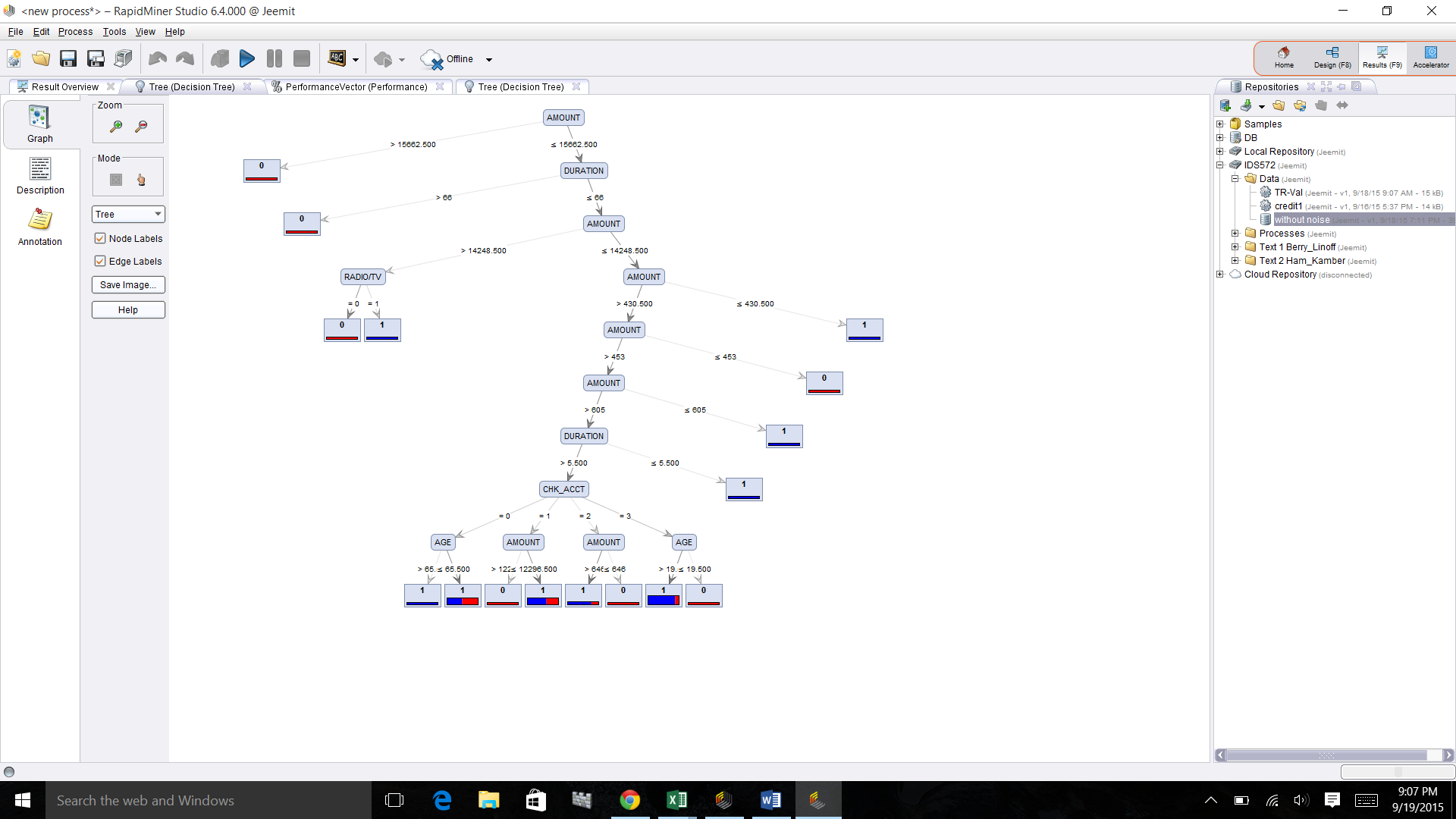
**Robustness and other parameters:-**

The overall accuracy we are getting with this model of Gain Ratio is more than 70% (i.e. more than No model case) and it generates a decision tree that gives the data in a much simplified manner as compared to other 3 criterion (Gini Index, Information Gain, and Accuracy)

**Node Parameters used to get a good model:-**

Maximal Depth of the tree and Gain ratio.

These two parameters give a much more reliable and simple to understand model.



**Q3 a) Results for Data with 50:50 ratio split:-**

**Testing Data Results**

|  |  |  |  |
| --- | --- | --- | --- |
|  | true 1 | true 0 | class precision |
| pred. 1 | 321 | 86 | 78.87% |
| pred. 0 | 30 | 63 | 67.74% |
| class recall | 91.45% | 42.28% |  |

**Validation Data Results**

|  |  |  |  |
| --- | --- | --- | --- |
|  | true 1 | true 0 | class precision |
| pred. 1 | 304 | 89 | 77.35% |
| pred. 0 | 45 | 62 | 57.94% |
| class recall | 87.11% | 41.06% |  |
|  |  |  |  |

**The model is less reliable since the variance is more than 3%**

**Q3b) Data with 80:20 split:-**

**Training Data Results**

|  |  |  |  |
| --- | --- | --- | --- |
|  | true 1 | true 0 | class precision |
| pred. 1 | 403 | 101 | 79.96% |
| pred. 0 | 90 | 106 | 54.08% |
| class recall | 81.74% | 51.21% |  |

**Validation Data Results**

|  |  |  |  |
| --- | --- | --- | --- |
|  | true 1 | true 0 | class precision |
| pred. 1 | 165 | 48 | 77.46% |
| pred. 0 | 42 | 45 | 51.72% |
| class recall | 79.71% | 48.39% |  |

**Data with 70:30 split:-**

**Training Data Results**

|  |  |  |  |
| --- | --- | --- | --- |
|  | true 1 | true 0 | class precision |
| pred. 1 | 463 | 122 | 79.15% |
| pred. 0 | 101 | 114 | 53.02% |
| class recall | 82.09% | 48.31% |  |

**Validation Data Results**

|  |  |  |  |
| --- | --- | --- | --- |
|  | true 1 | true 0 | class precision |
| pred. 1 | 108 | 34 | 76.06% |
| pred. 0 | 28 | 30 | 51.72% |
| class recall | 79.41% | 46.88% |  |

Out of these 3 models ,The model with the 70/30 split gives bettter accuracy and less variance.

It is always advisable to choose such models.

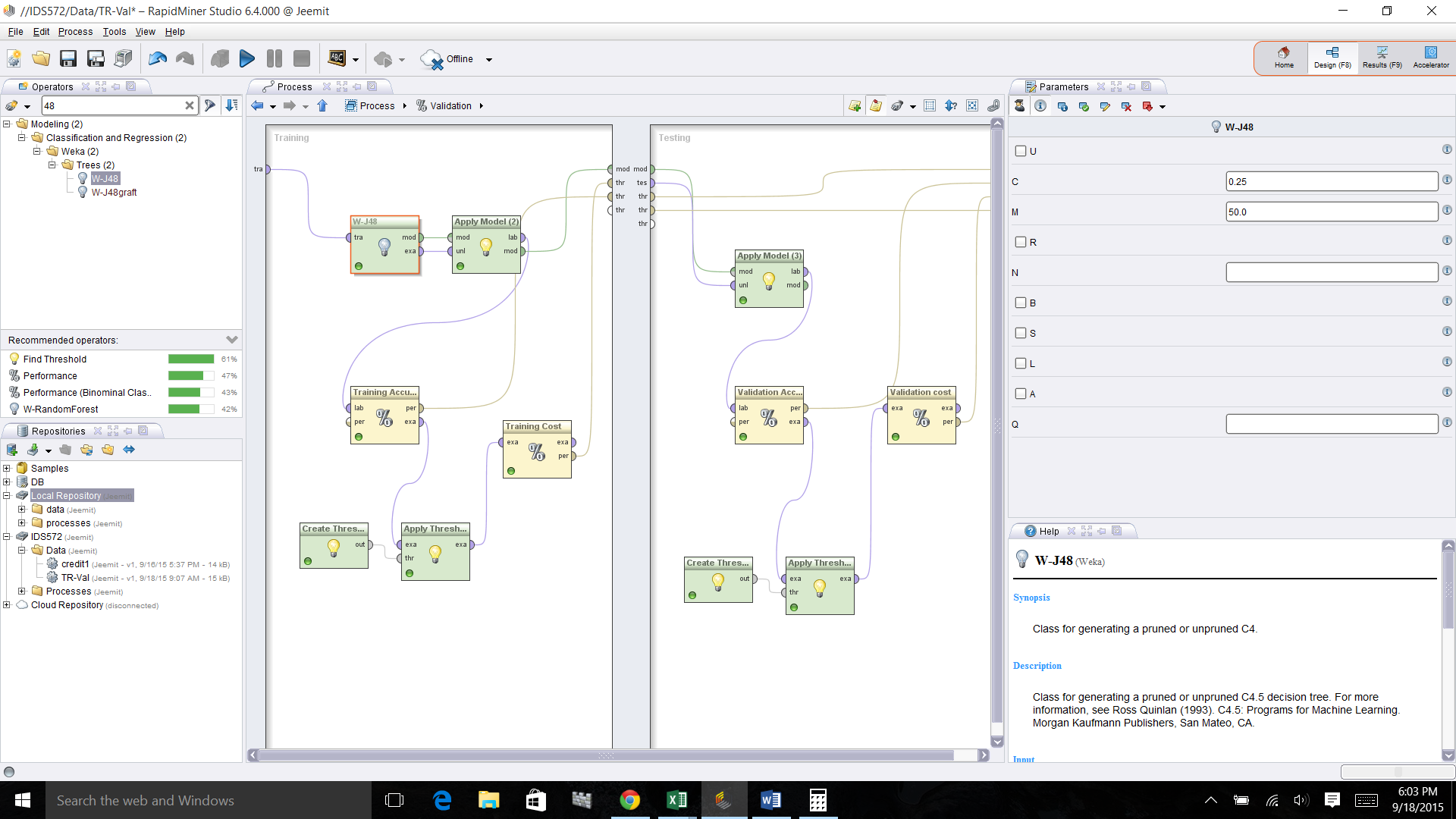
Prunning does not help in the model as it reduces the tree size to 1 level and gives accuracy less than 70%.which is less than the accuracy of No model.

The most important parameter is the tree depth and the accuracy remains the same for different training sets example.

Tree depth value= Set as 10.

**Q3c)**

**W-J48 Model (Optimum model):-**



**Training Results for W-J48 Model (Accuracy=73.14) :-**

|  |  |  |  |
| --- | --- | --- | --- |
|  | true 1 | true 0 | class precision |
| pred. 1 | 414 | 122 | 77.24% |
| pred. 0 | 66 | 98 | 59.76% |
| class recall | 86.25% | 44.55% |  |

**Validation Results for W-J48 Model (Accuracy=73.00):-**

|  |  |  |  |
| --- | --- | --- | --- |
|  | true 1 | true 0 | class precision |
| pred. 1 | 192 | 53 | 78.37% |
| pred. 0 | 28 | 27 | 49.09% |
| class recall | 87.27% | 33.75% |  |

After changing Random seeds and training dataset for the W-J48 model. It seems to be the more stable model

compared to other models as it gives the better accuracy and lesser variance after changing the parameters.

And the upper part of the tree remains the same as there is always a minor change in the accuracy and variance

between training and validation dataset.

**Q4) Misclassification cost:-**

**Cut off Training data Misclassification cost Validation data Misclassification cost**

0.2 187.6 177.7

0.5 64.5 64.3

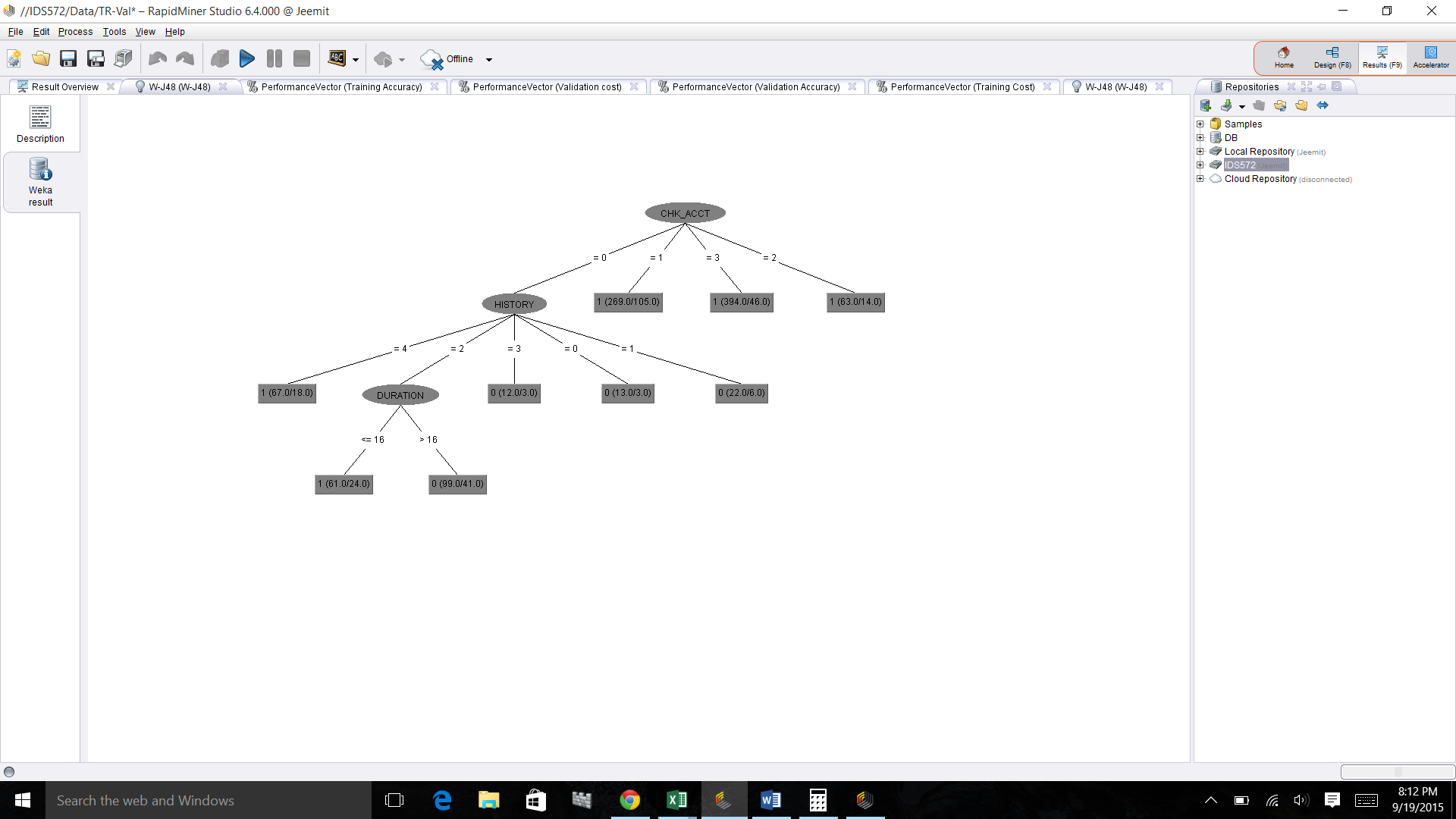
**0.7 31.42 26.667**

The results shows that for classification threshold cut-off as the value of cut off increases the Misclassification

Cost decreases.

**Q5 a) Decision Tree:-**

**Decision Tree**



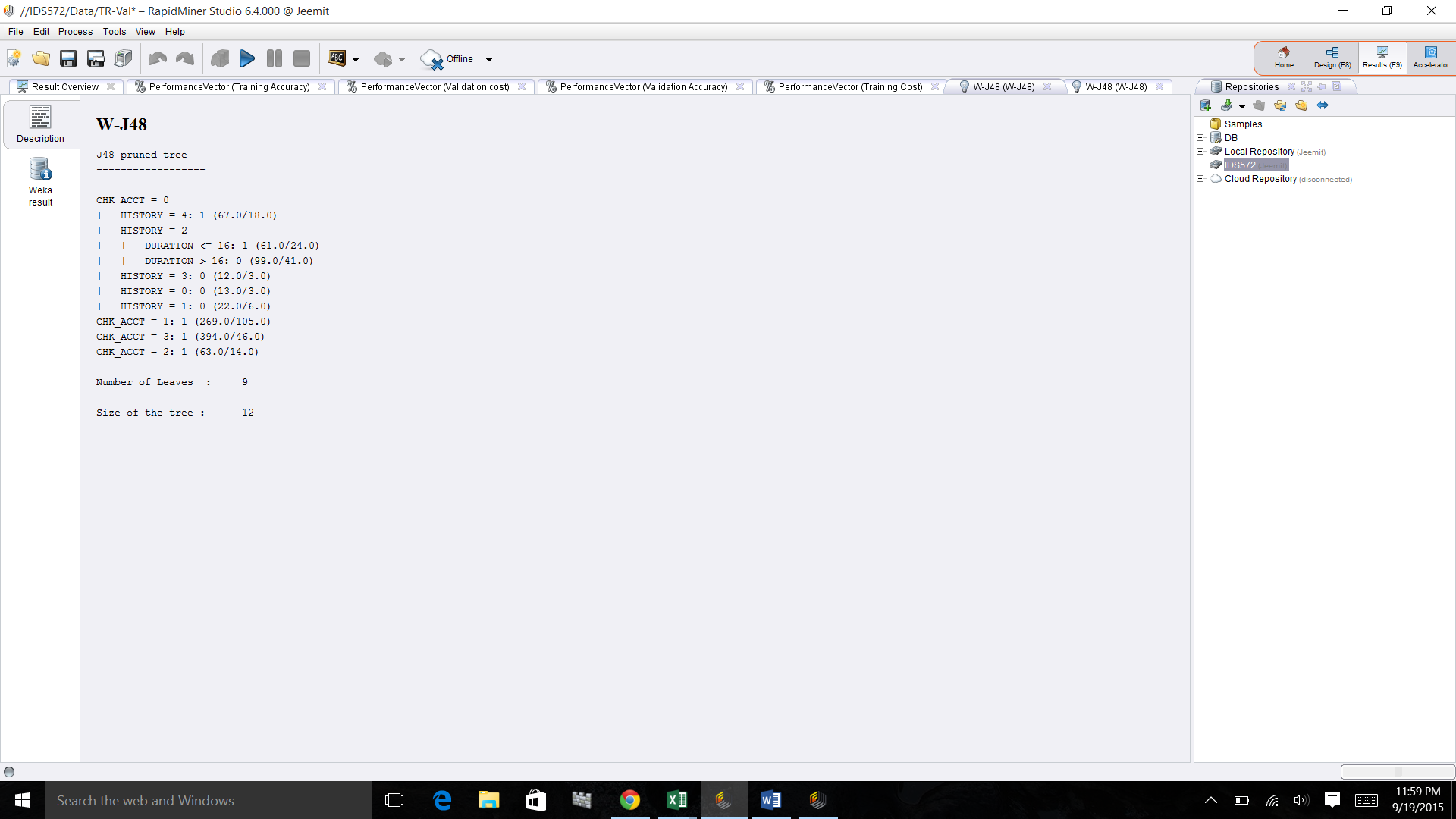
Depth of the tree = 3

Number of Nodes = 12

Variables at the top of the tree:- Checking account, History, Duration.

These variables are not same compared to the above question 2.

**Q5b)**



Pure Leafs:-

Checking Account = 1

Duration >60

**Q5c) Predictions/ Rules:-**

1. If checking account is equal to 1 then probability of good cases are 164/269=60.9%
2. If history is equal to 4 then the probability of good cases are 49/67 =73.14%

Q6) **Profit and Cumulative Profit:-**

**Cost chart for whole data**

The above figure shows the data point for all the cases which are predicted as 1.

Above chart is prepared by comparing the RESPONSE & PREDICTION

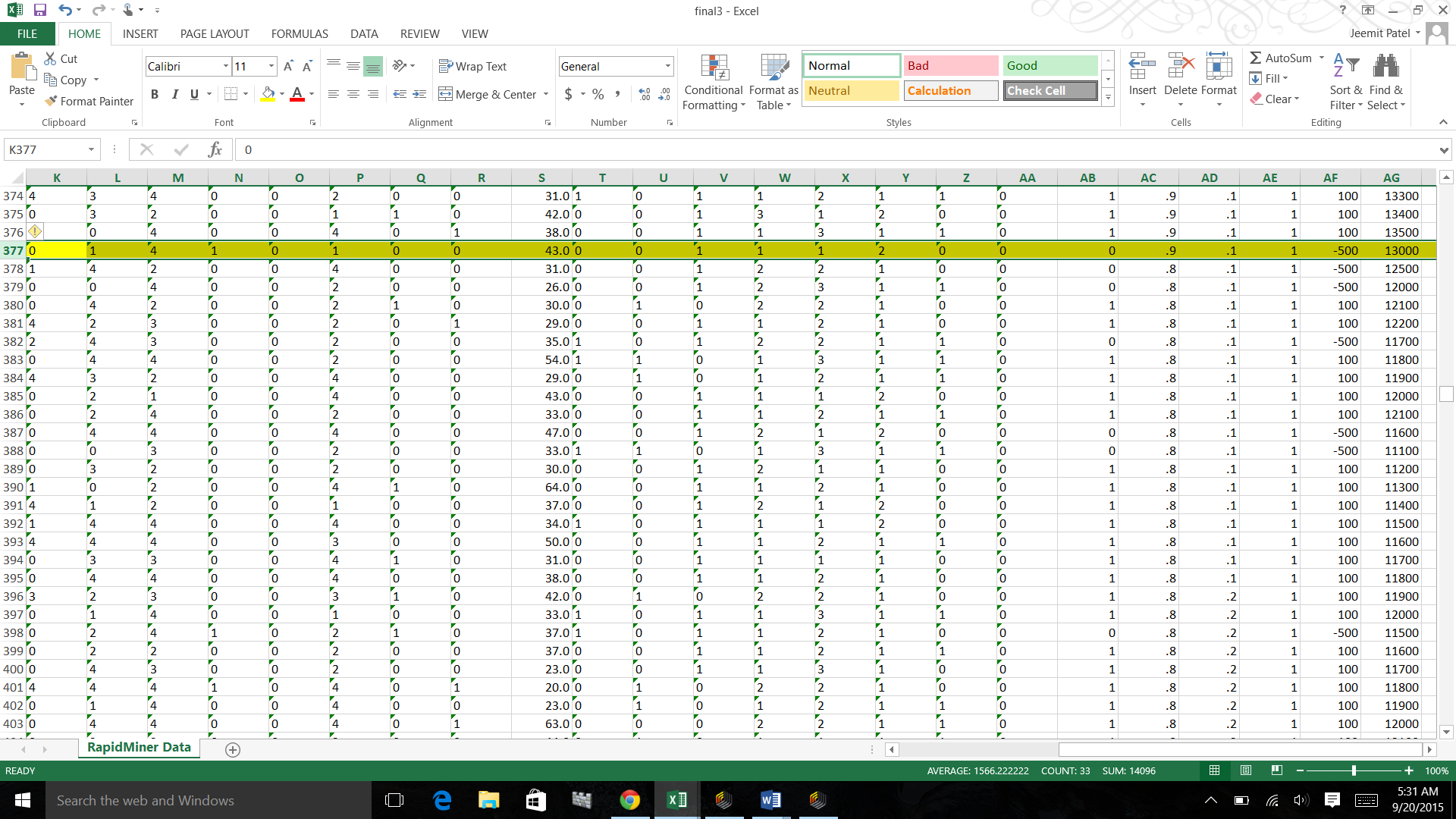
With the threshold of 0.7 (confidence)

The maximum Profit we can gain is $13000 in this case.

**This amount of profit we are attaining at 0.9 confidence.**

**So in this case we should limit our self to the customers attaining 0.9 confidence.**

**This will save us a lot of money and resources.**



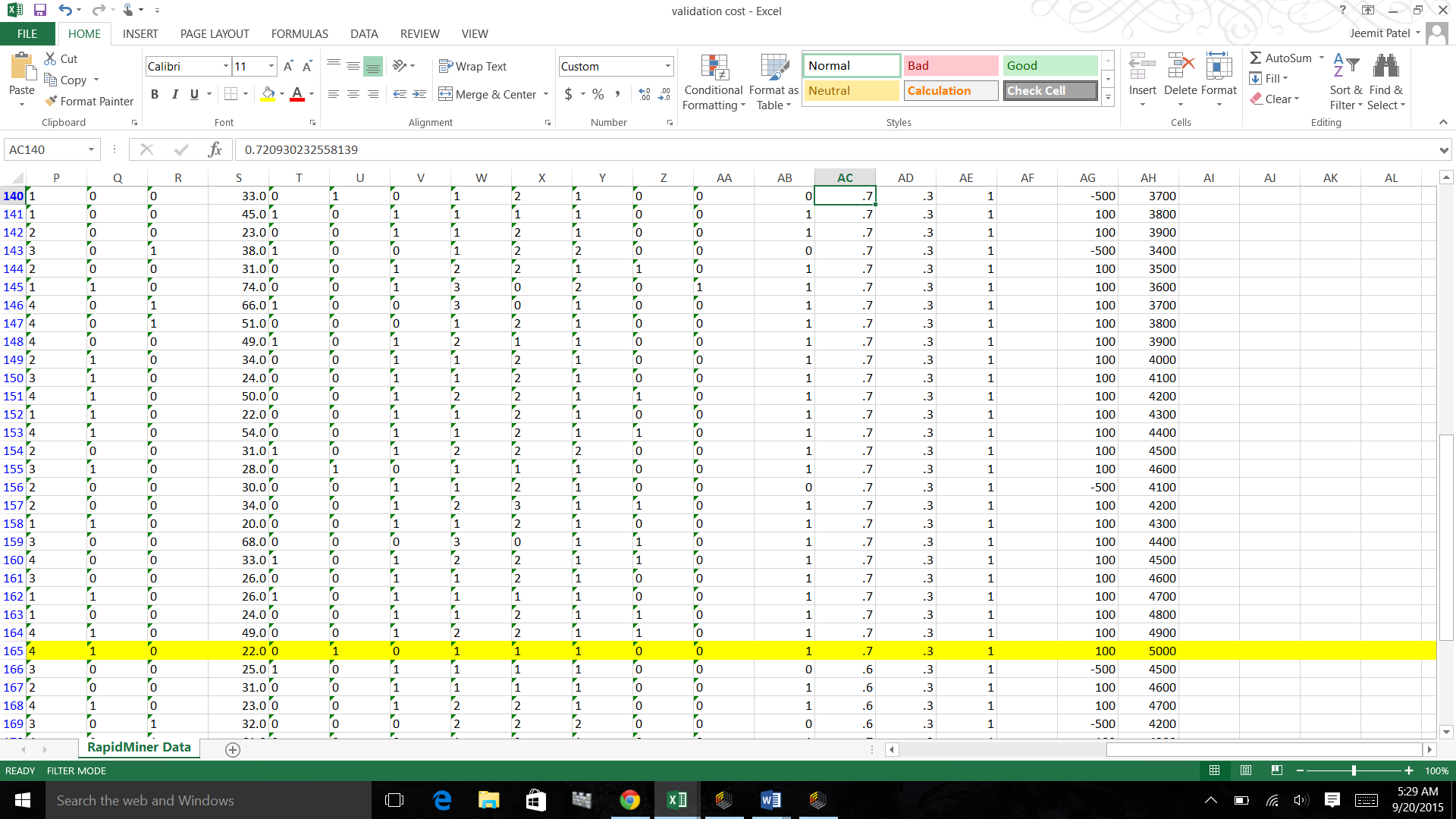
**Cost Chart for Validation Data**

The above figure shows the data point for Validation Set (30%) cases which are predicted as 1.

Above chart is prepared by comparing the RESPONSE & PREDICTION

With the threshold of 0.7 (confidence)

The maximum Profit we can gain is $5000 in this case.



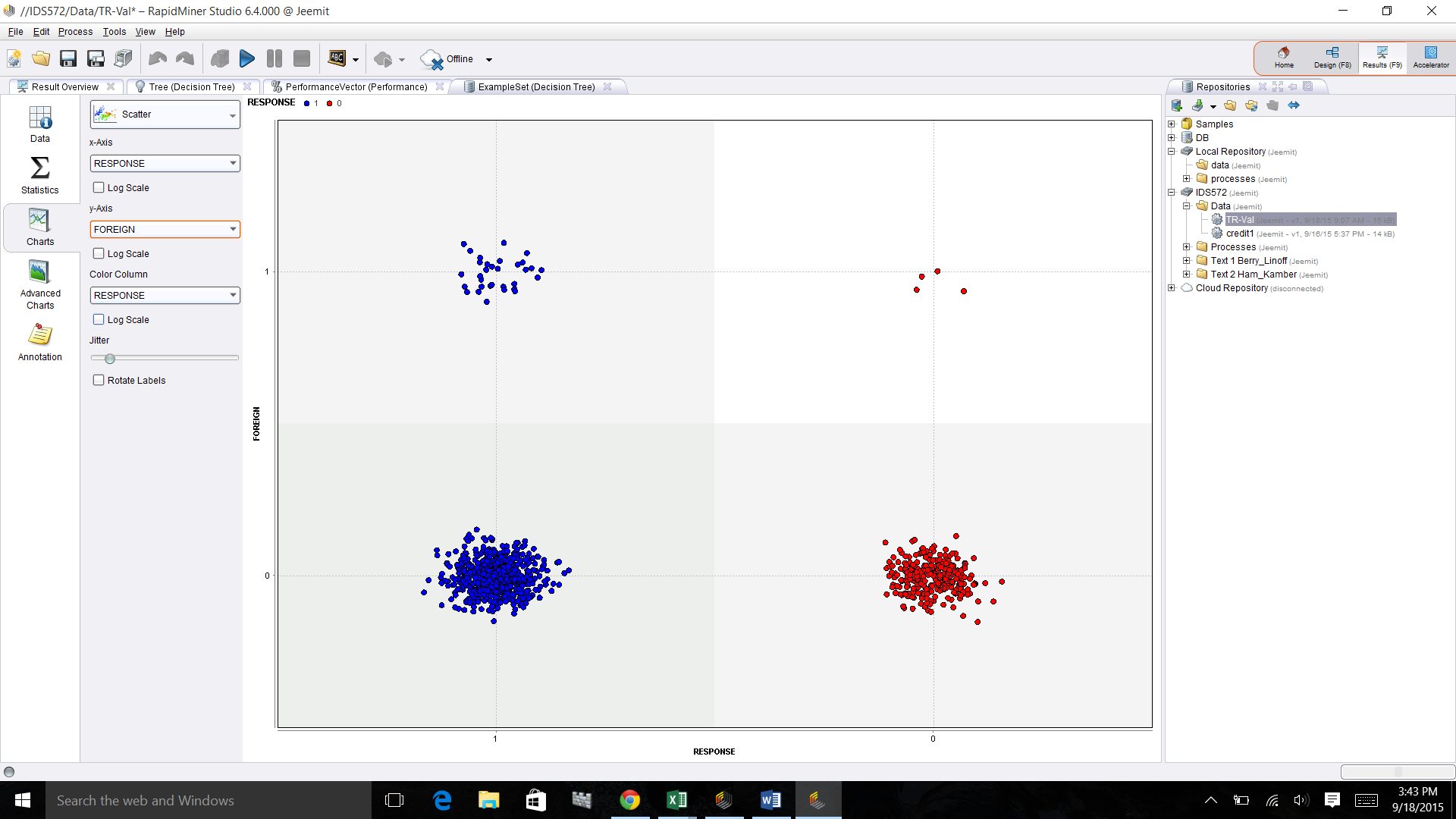
For future cases, prediction parameters should be the cases with the confidence more than 0.7. and 0.9

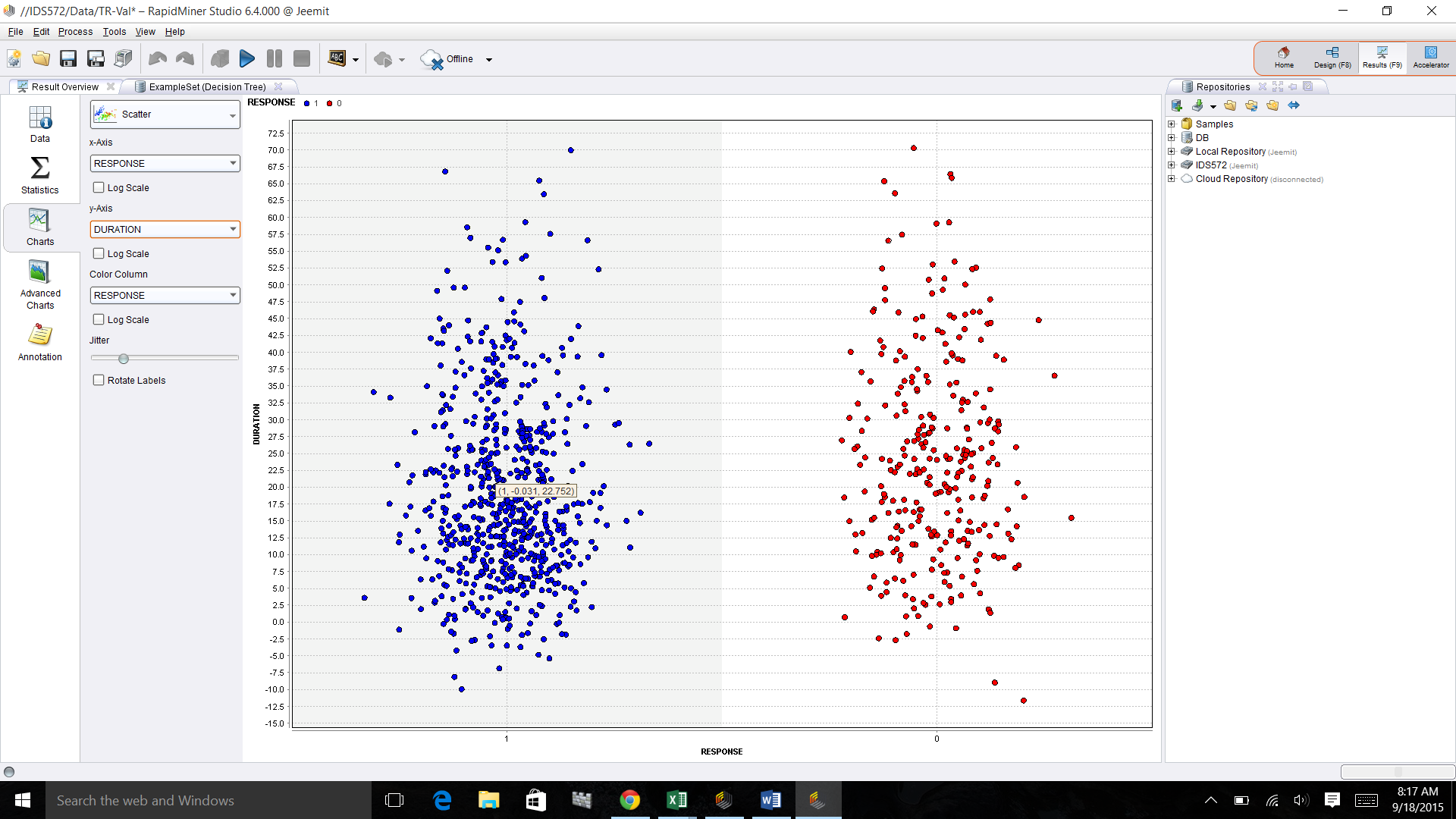
With the above shown model and set parameters as mentioned we would be able to get the maximum cumulative profit.

**APPENDIX**

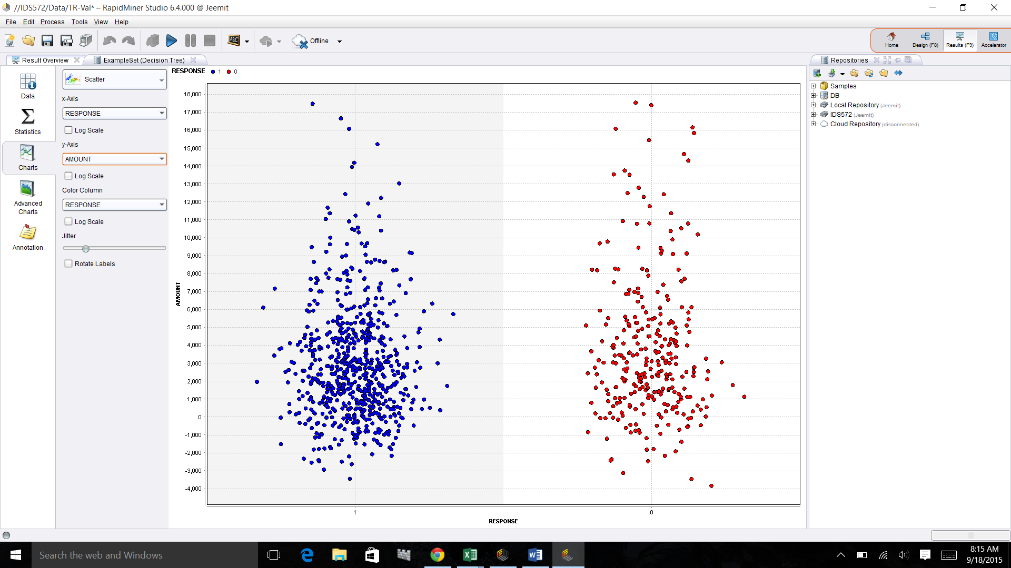
**Noteworthy Points:-**

The Image below shows the that the probability of a foreign oerson getting a good response (1) is higher.

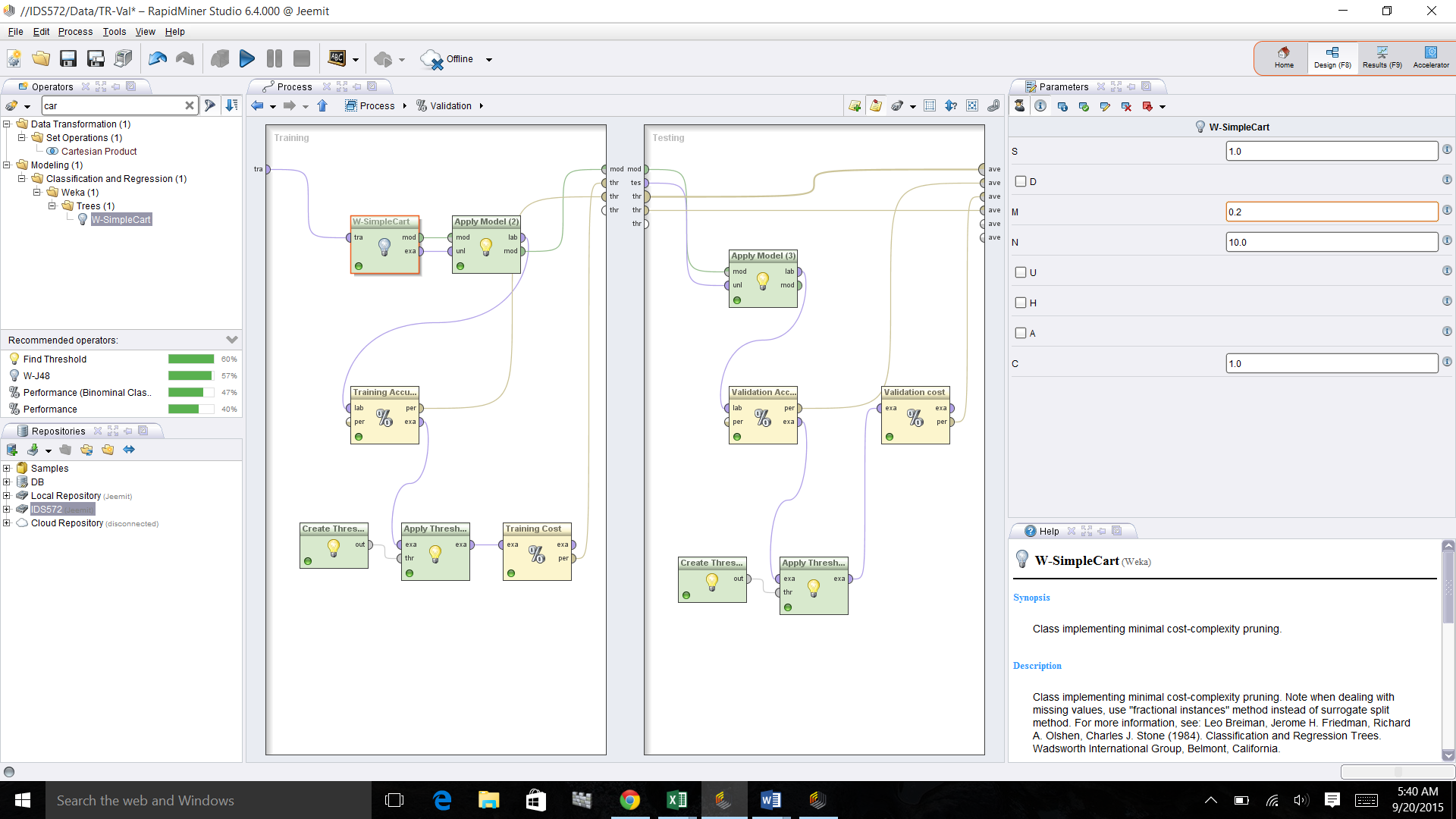


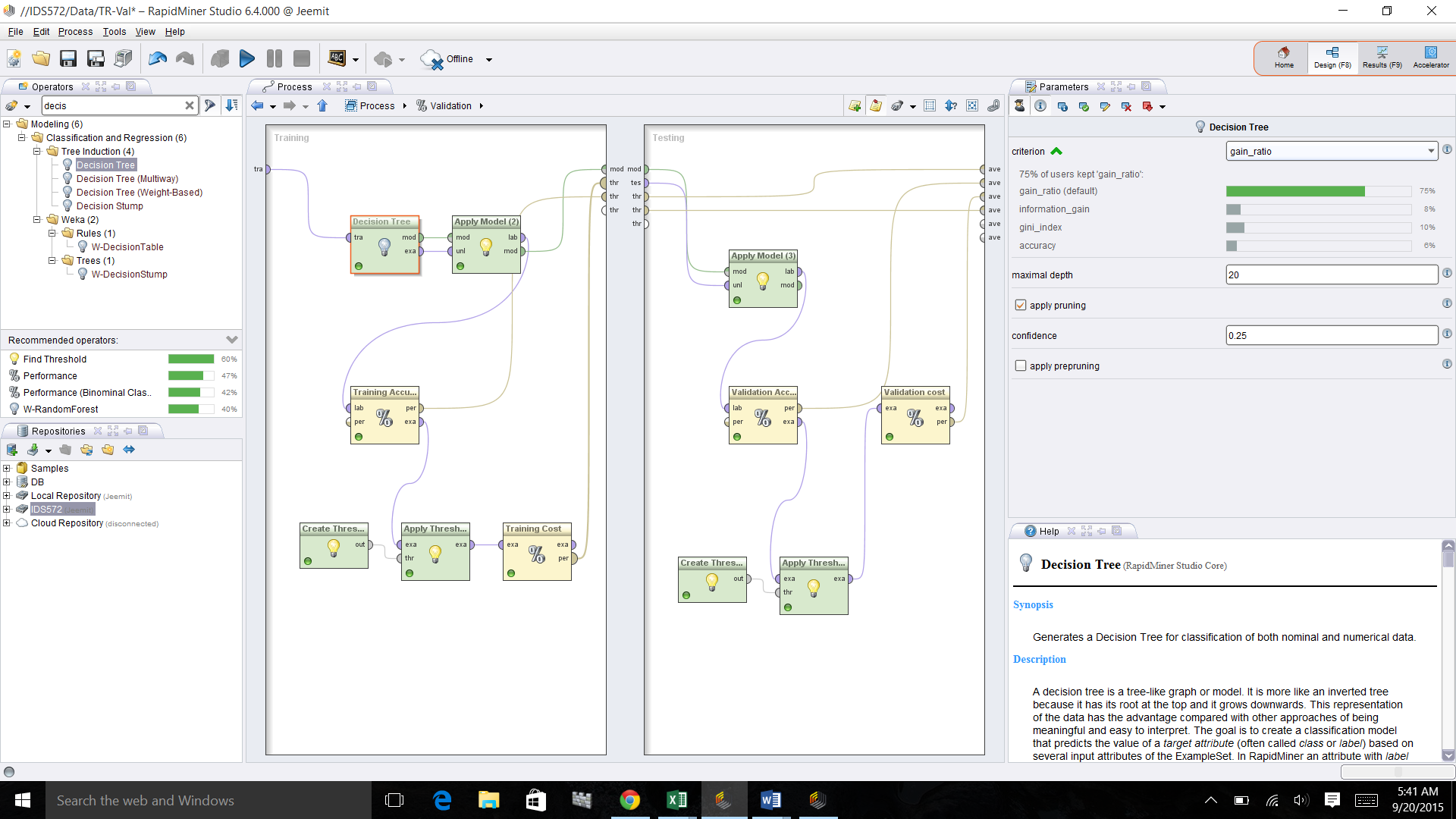
The Image below shows that the probability of a good response (0) will be more if the age of applicant is more than 45..

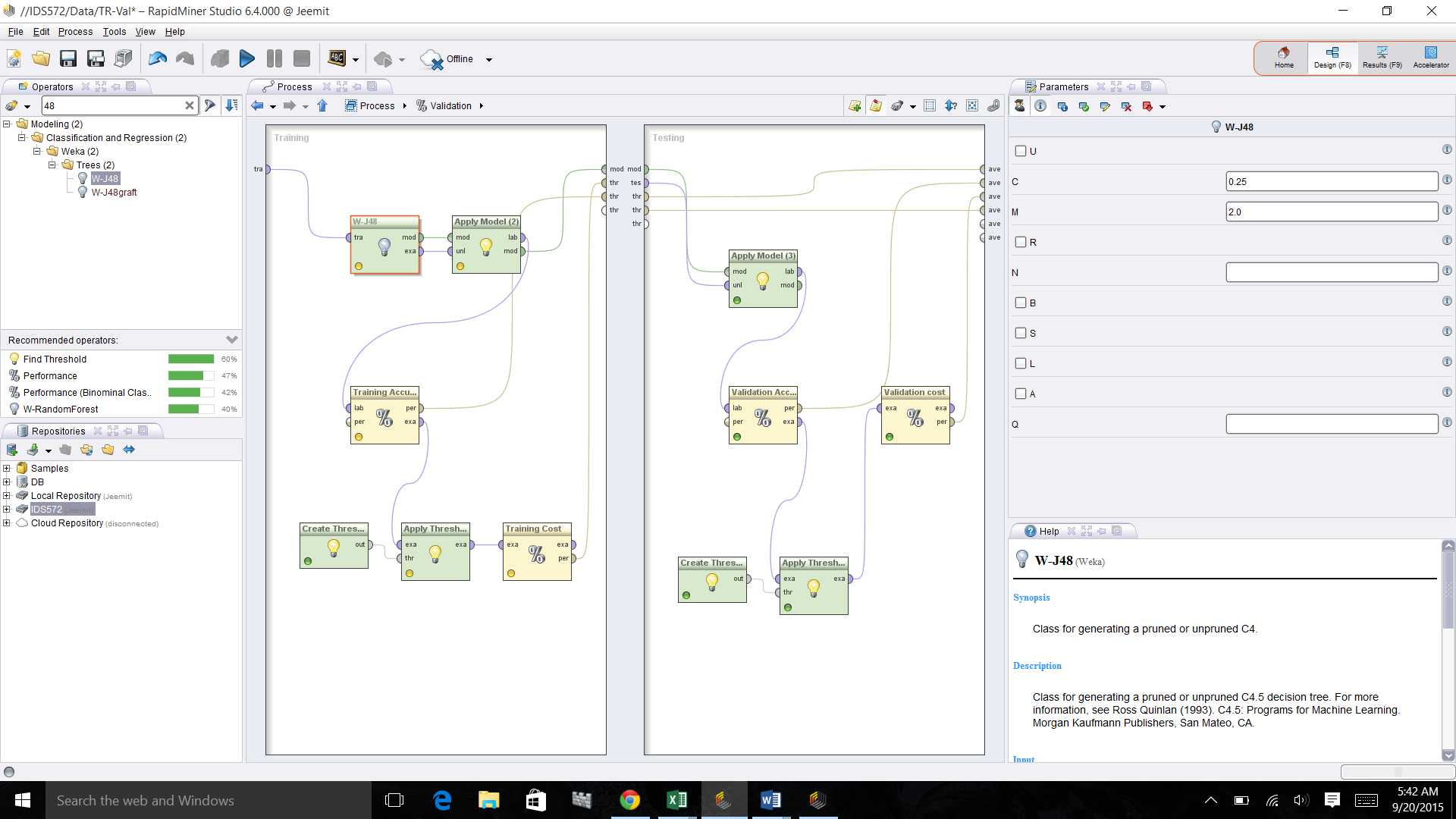
The image shows that if the credit amount is more than 8000 than the probablity becomes 50:50.

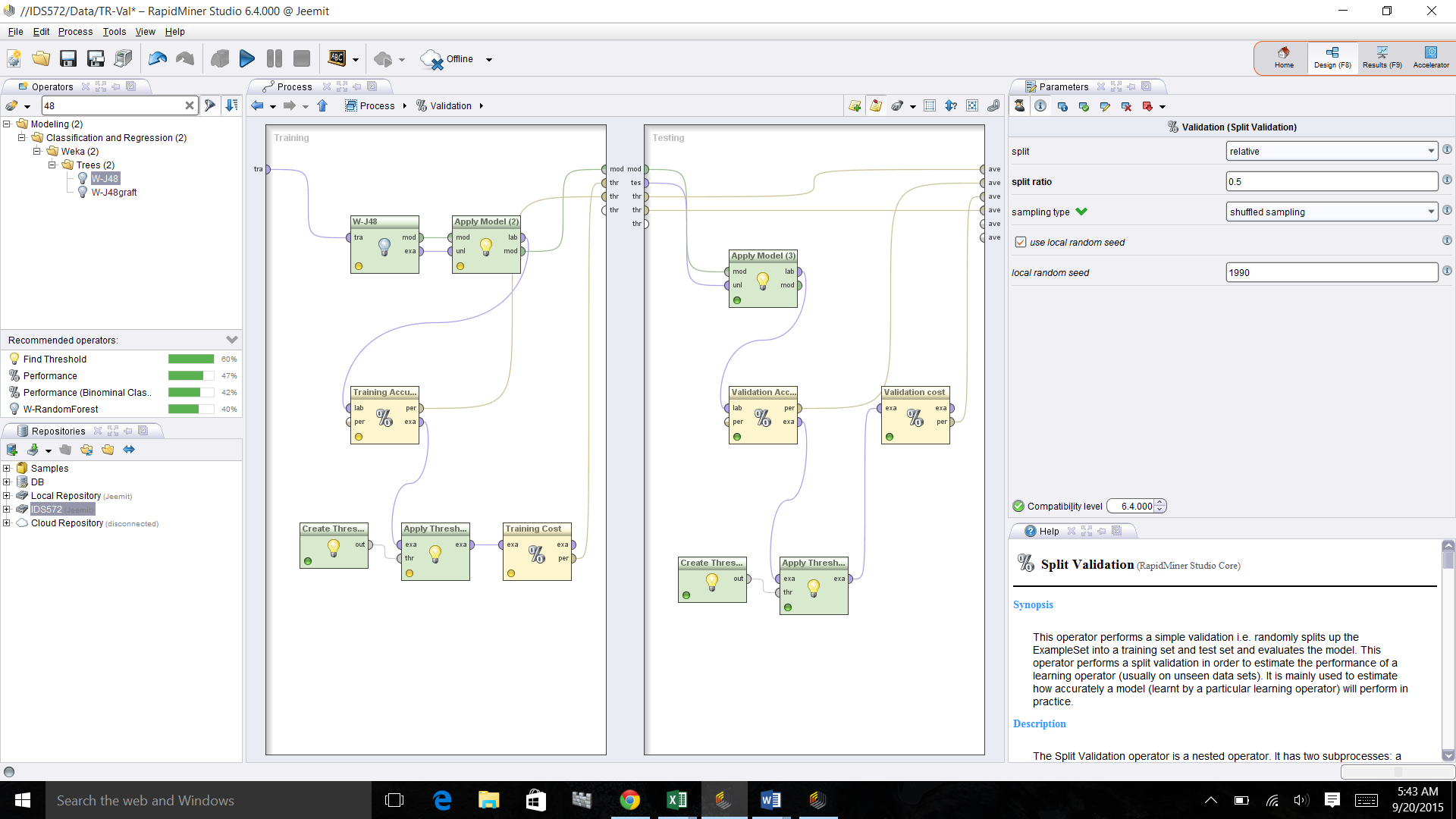


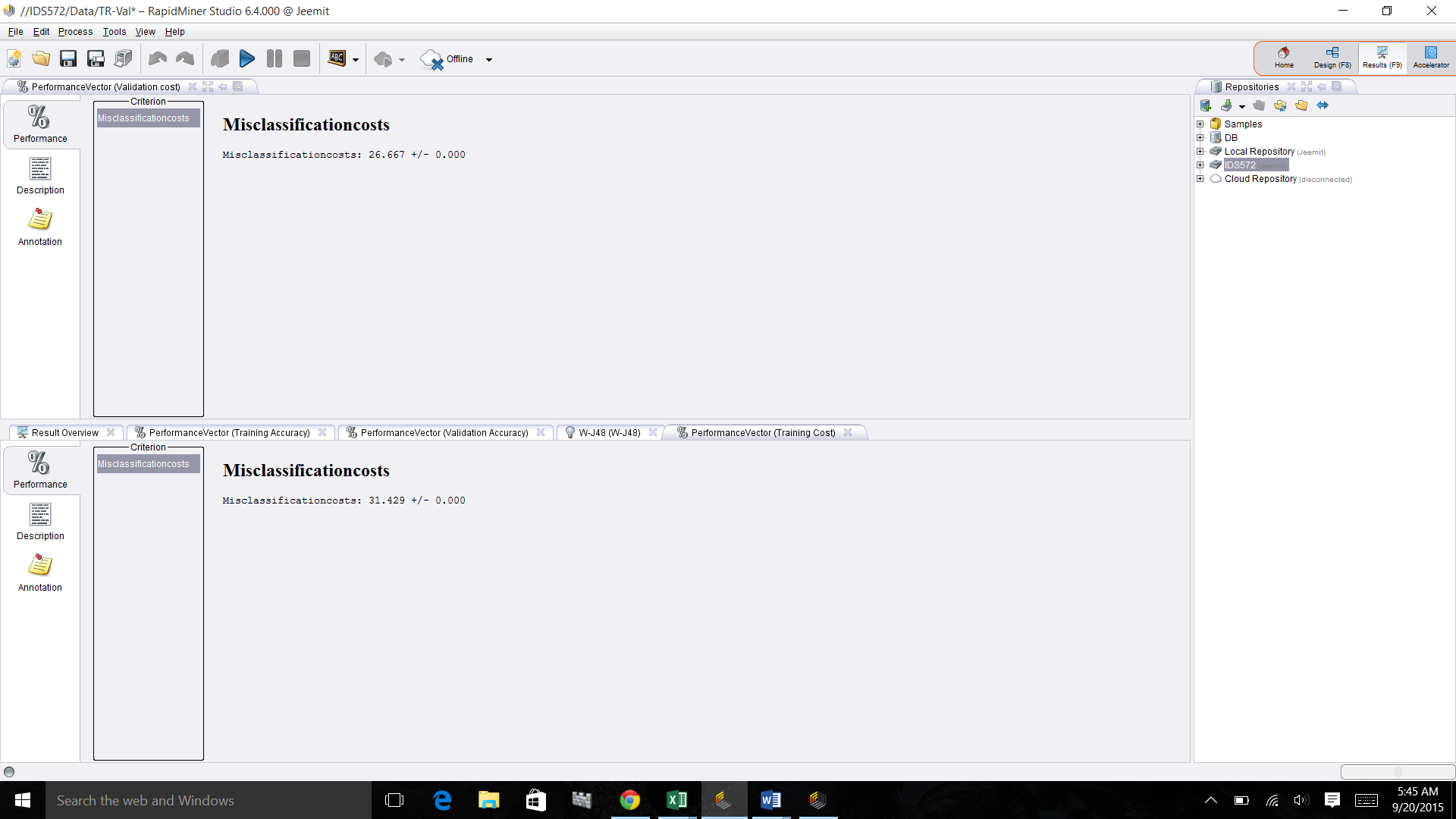
**CART MODEL**

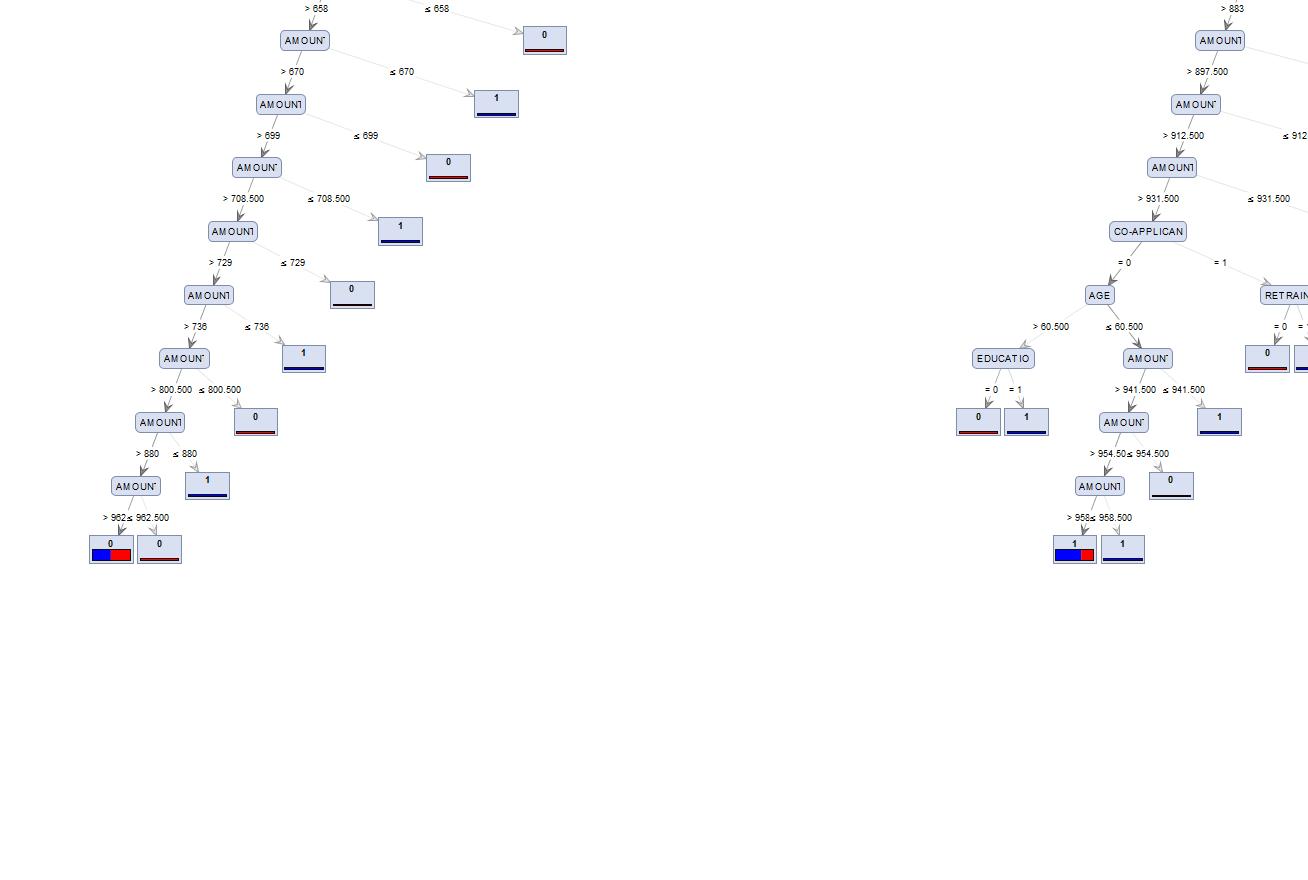
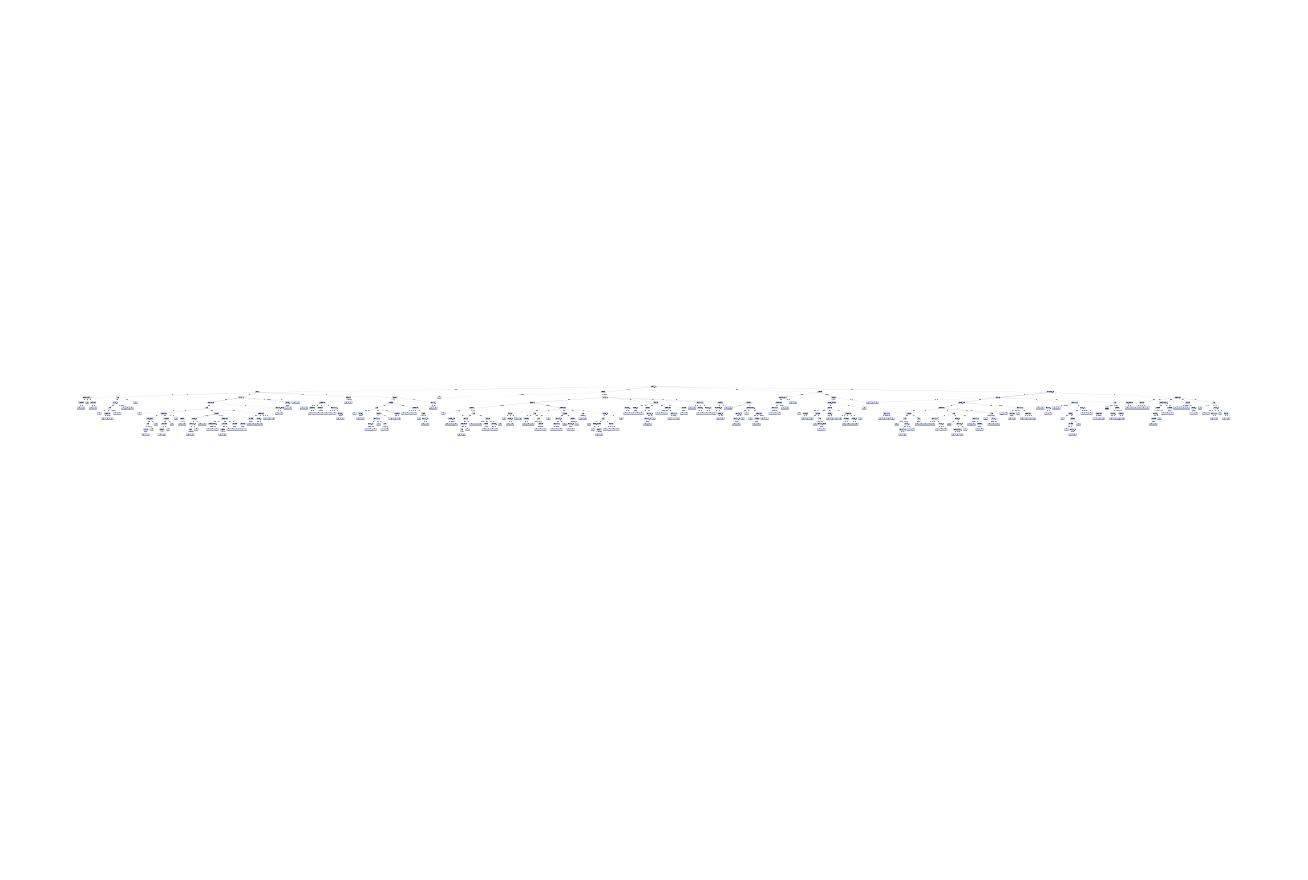


**Decision Tree Model**

W-J48 MODEL:

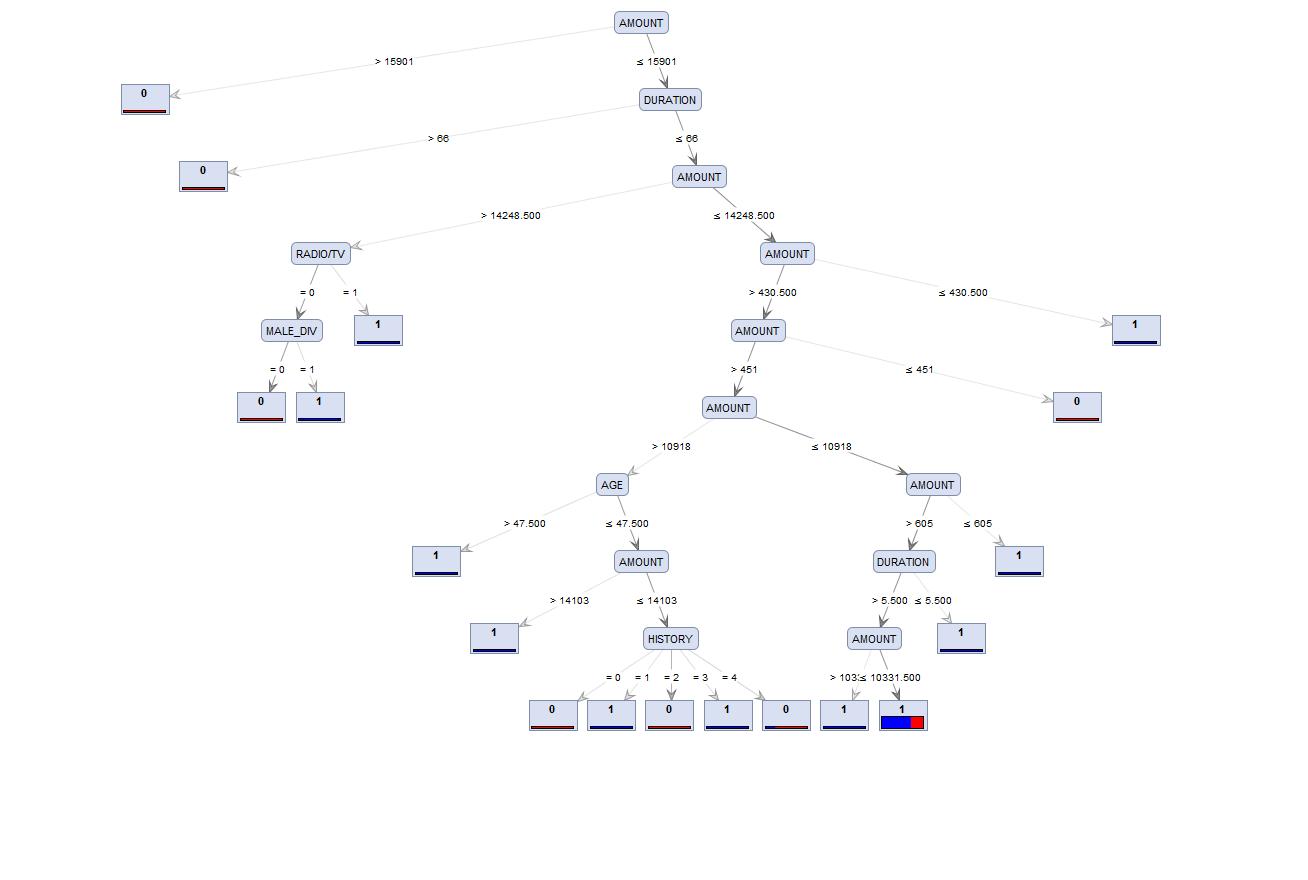
Split parameters on the right corner(70/30 split ratio gives the best accuracy)

Misclasification Cost for both Training and Validation data in proposed W-J48 model is less compared to other models.

Above images are the results of parameters like information gain and accuracy and gini index.

In the above cases the size of the tree is either large or size of the nodes are insignificant.

So these are the criterias which we can not consider while choosing an ideal model



Decision tree in this image is not reliable as the highest number of cases 963 are considered at only one node (right extreme). This makes a model unstable.