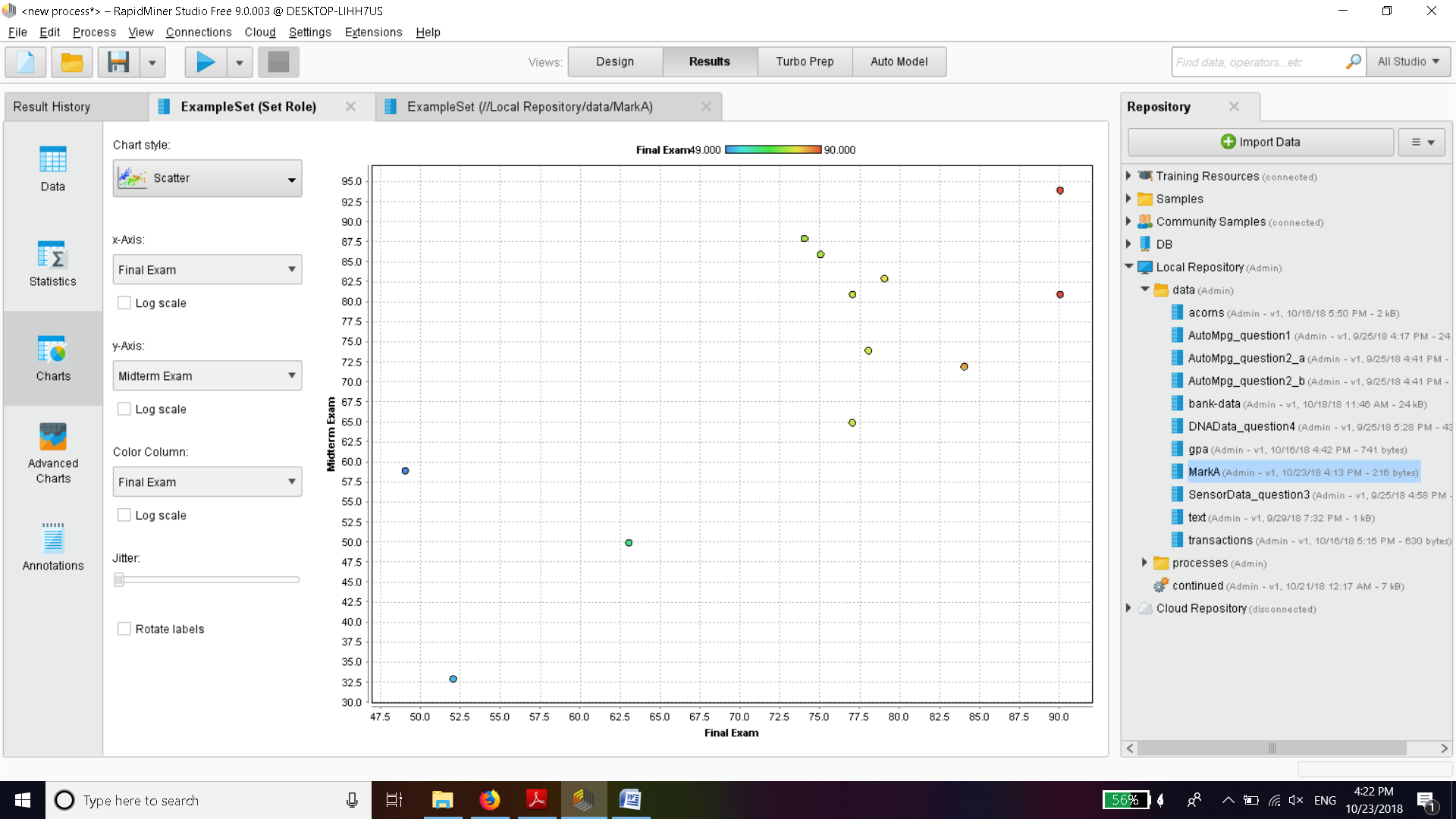
Q1)

a)The attributes do seem to have a positive relationship.



b)The linear relationship determined by Wsimplelinearregression is

=0.58 \* Midterm Exam + 32.03

ii)Plugging values of midterm exam as 86 we get

Final term=(0.58\*86)+32.03

=81.91

Q2)

a)Using polynomial regression with default parameters final exam is given by

= 1.030 \* MCQ1 ^ 1.000

+ 0.719 \* MCQ2 ^ 1.000

- 55.839

b) With seed value of final exam is=

0.498 \* MCQ1 ^ 1.000

+ 0.042 \* MCQ2 ^ 1.000

+ 29.913

After analysis of data the equation generated by seed is more accurate as compared to the equation without seed.

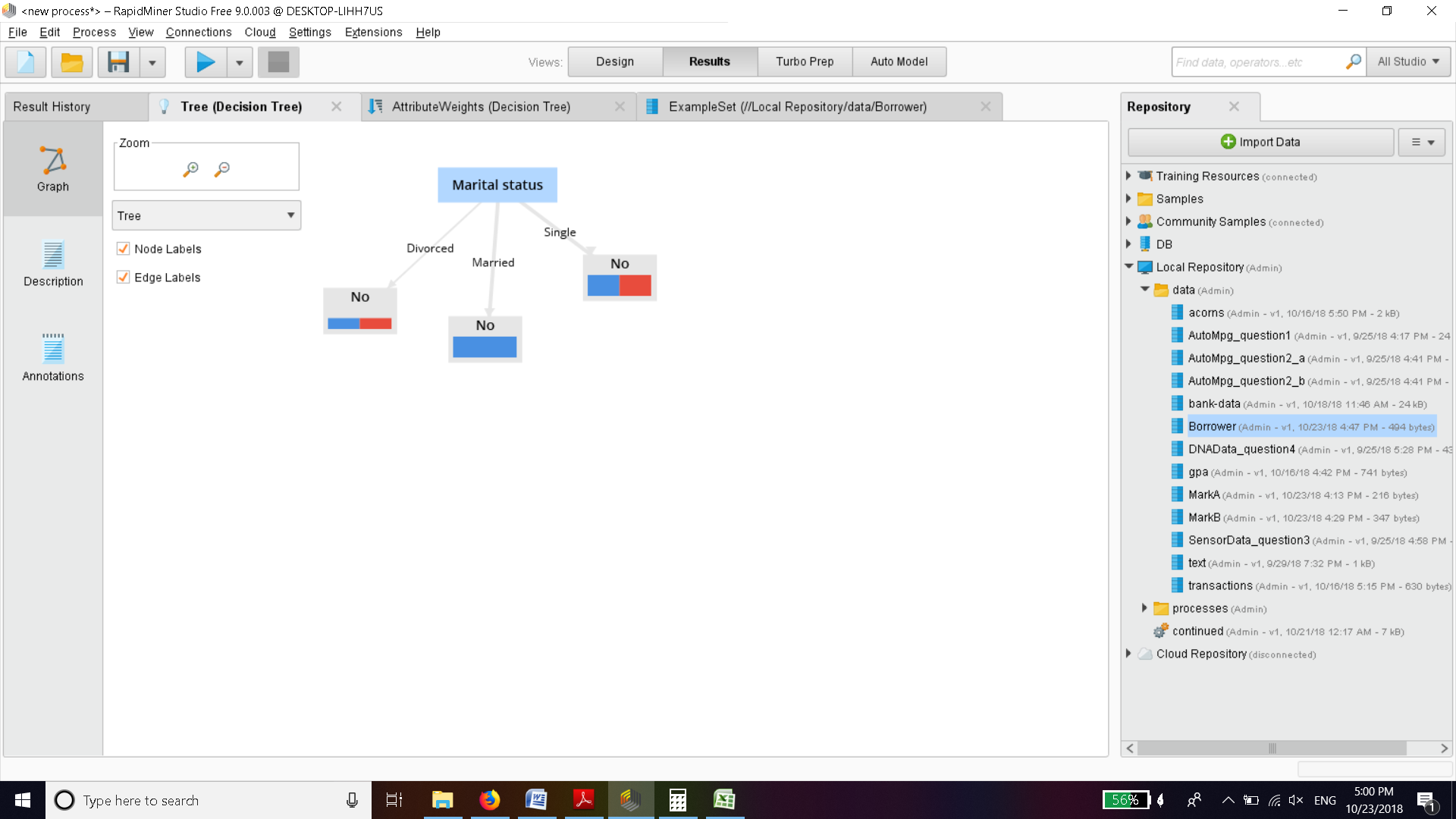
The sum of differences of actual result and obtained is =-48.05 for first equation,

while it is 9.548 for the seed equation.

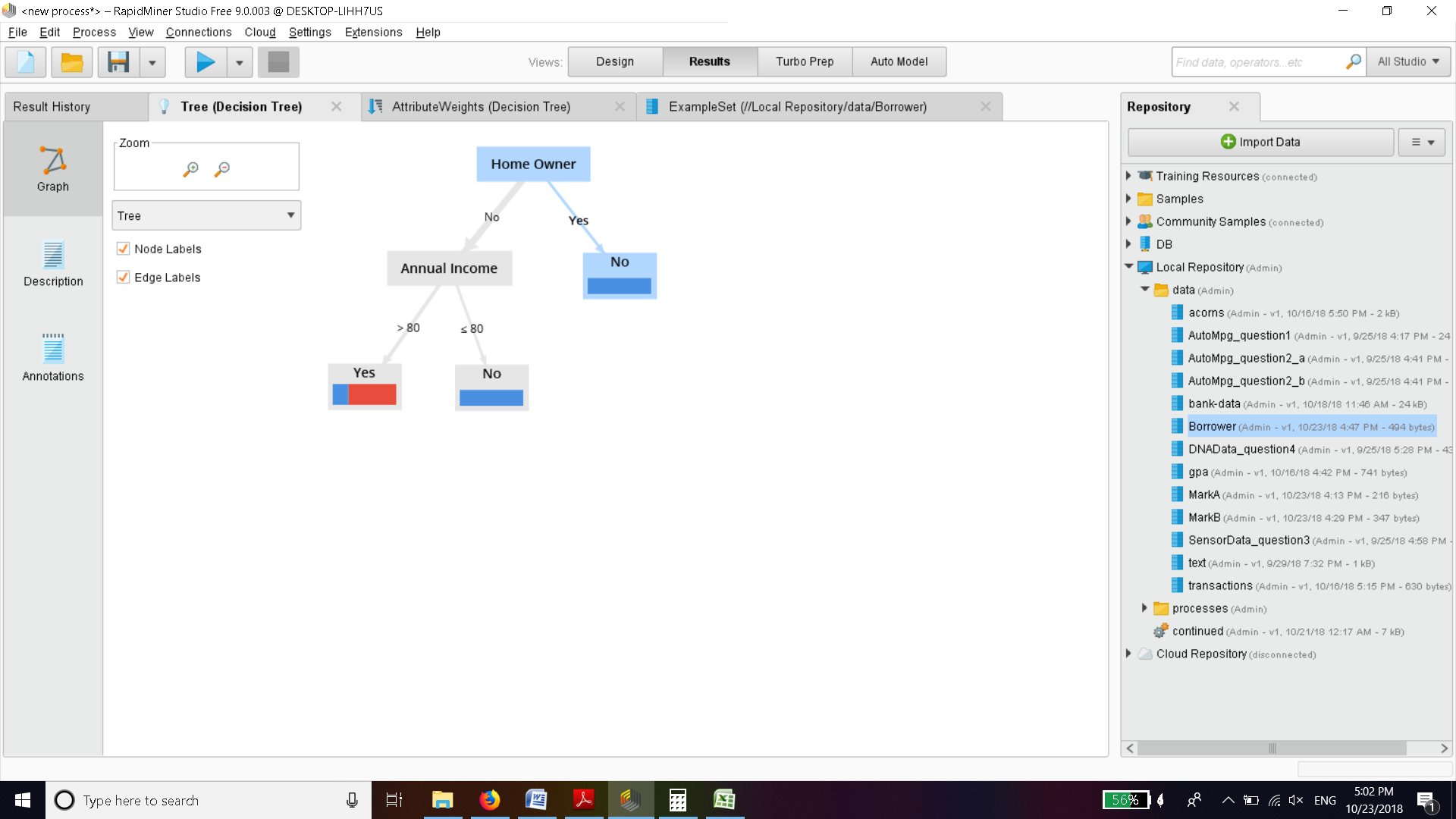
Q3)

a) Removing Tid

b)Generating decision tree with information gain we get



c)Using gain ratio we get decision tree



This method gives a better tree with far better classification of labels into yes or no.

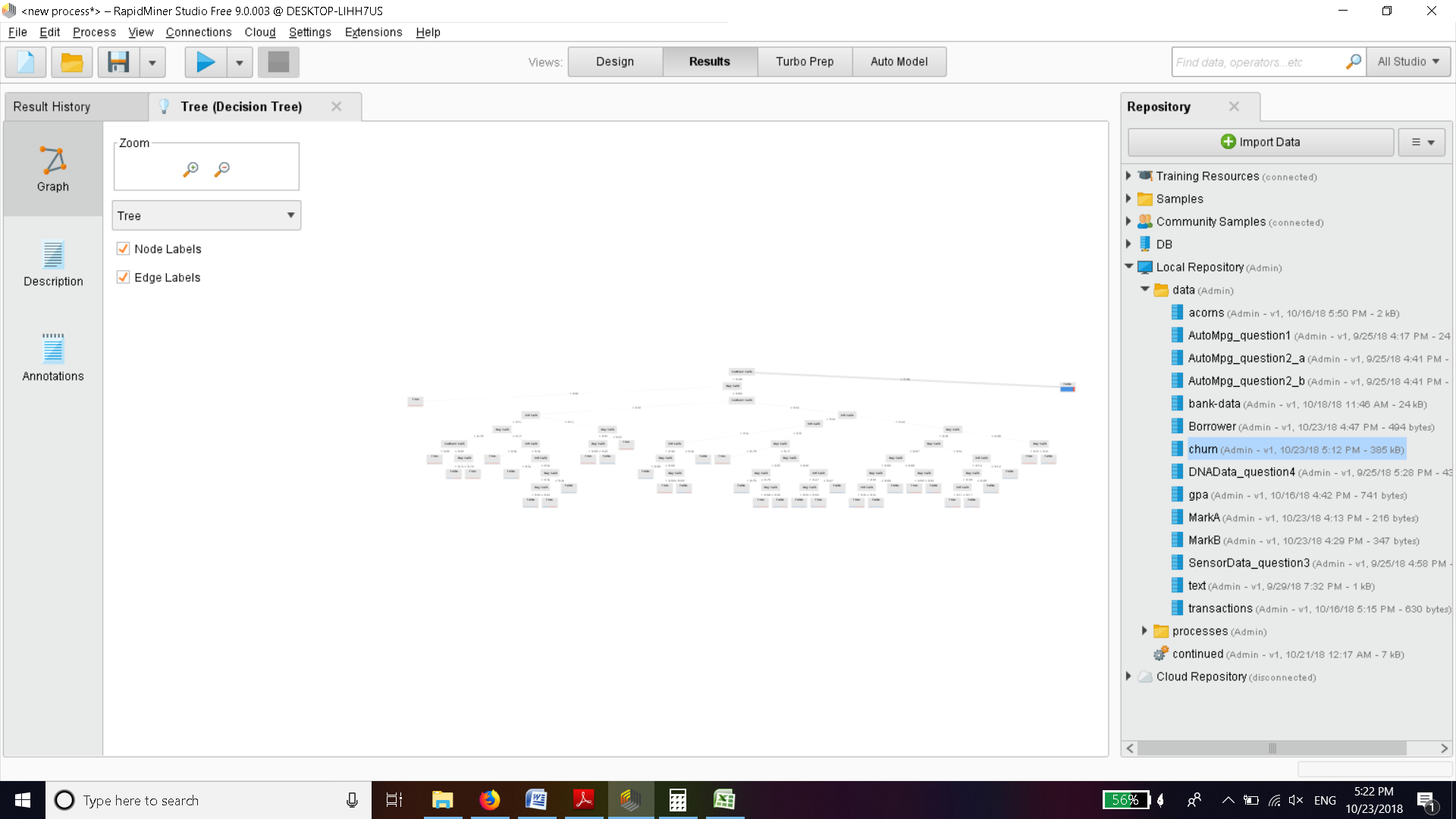
The earlier method had poor information gain because of single split into three classes.

With two splits based on gain ratio we get far better classification and accuracy for these results.

Q4)

a)removing attributes and assigning labels

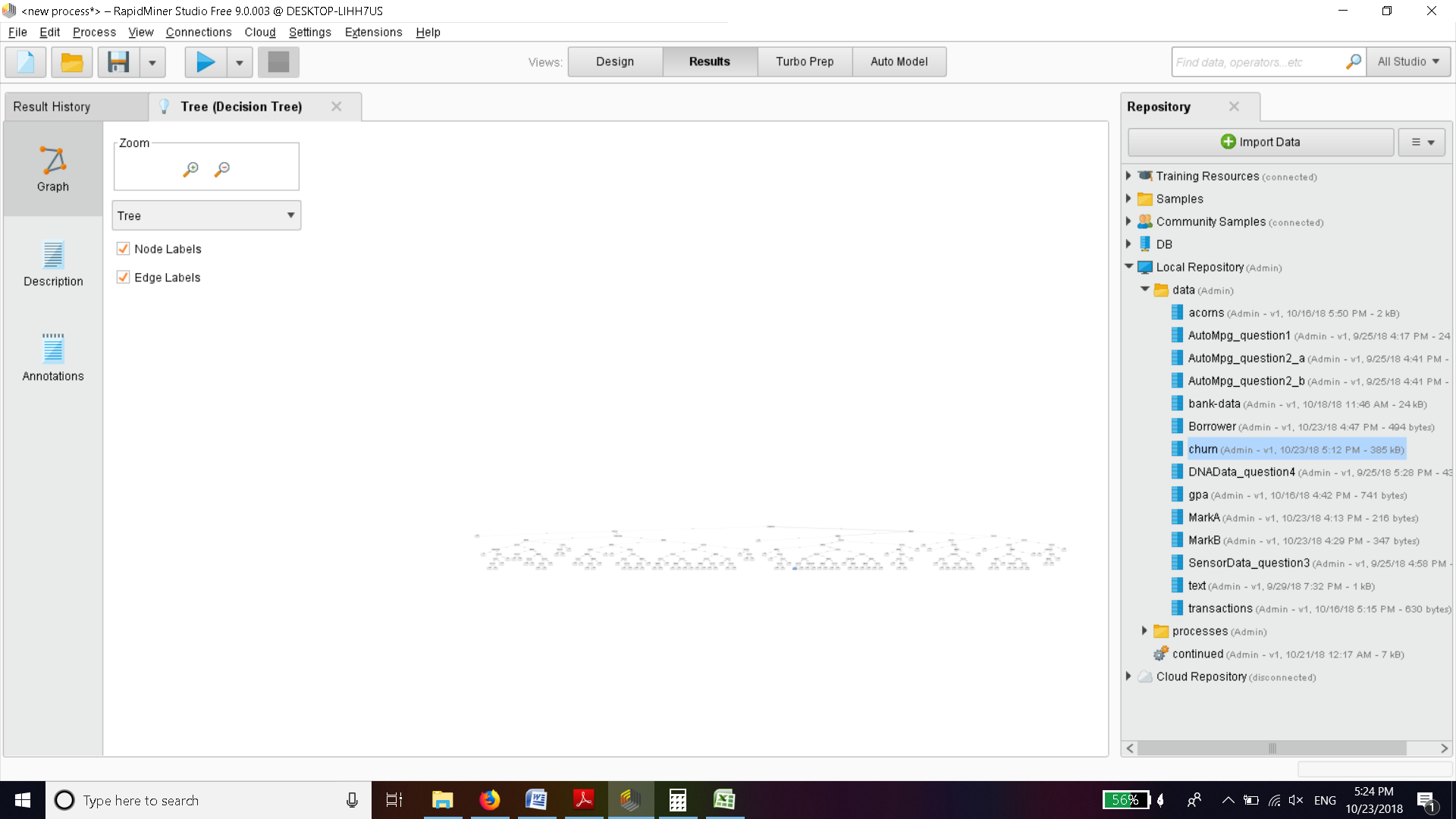
b)gini index of tree with default parameters no pruning.



b)Decision tree with gini index and default parameters results in

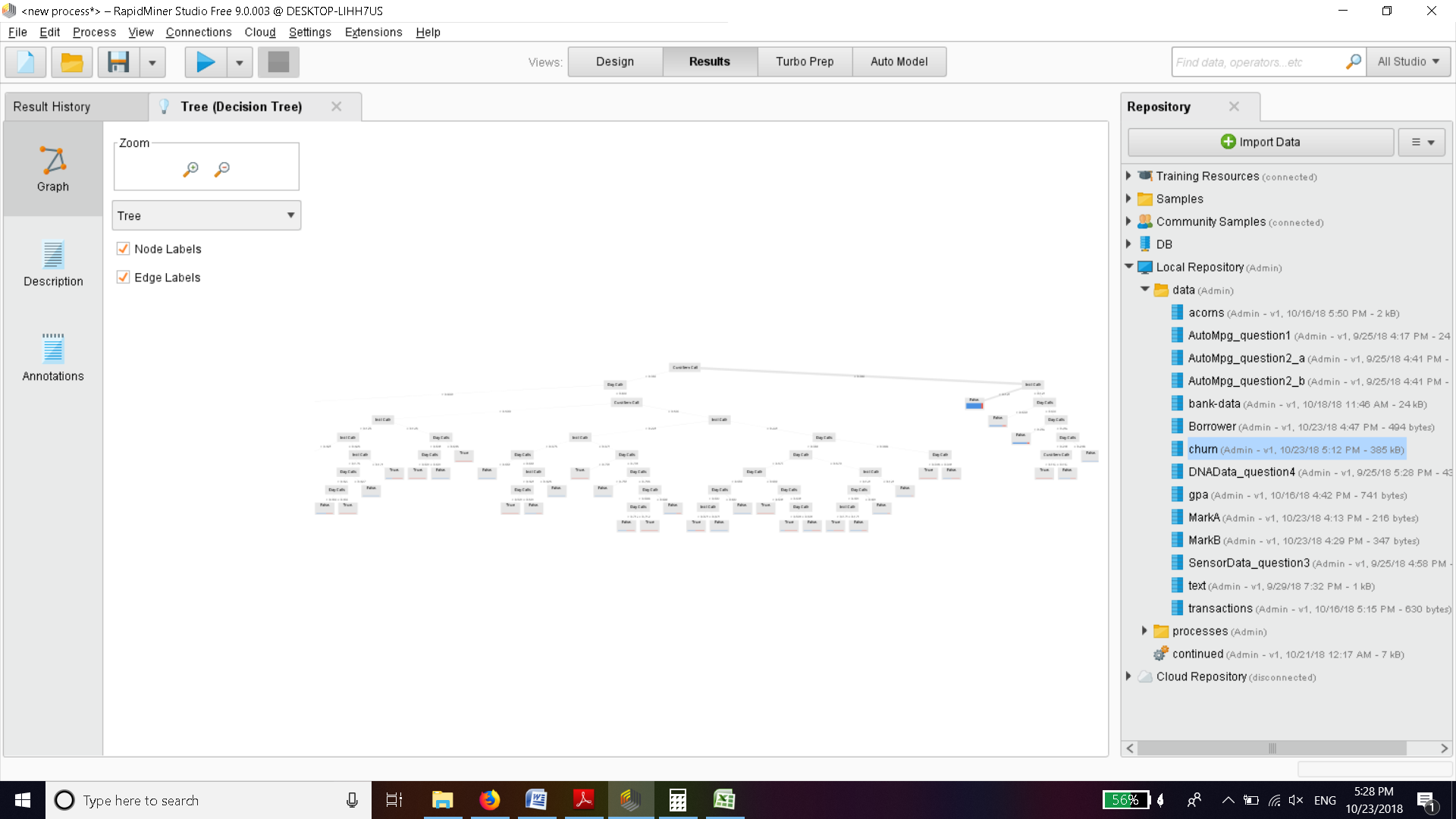
This tree is large.

c)Without pruning results in a very big tree that cant even be seen properly

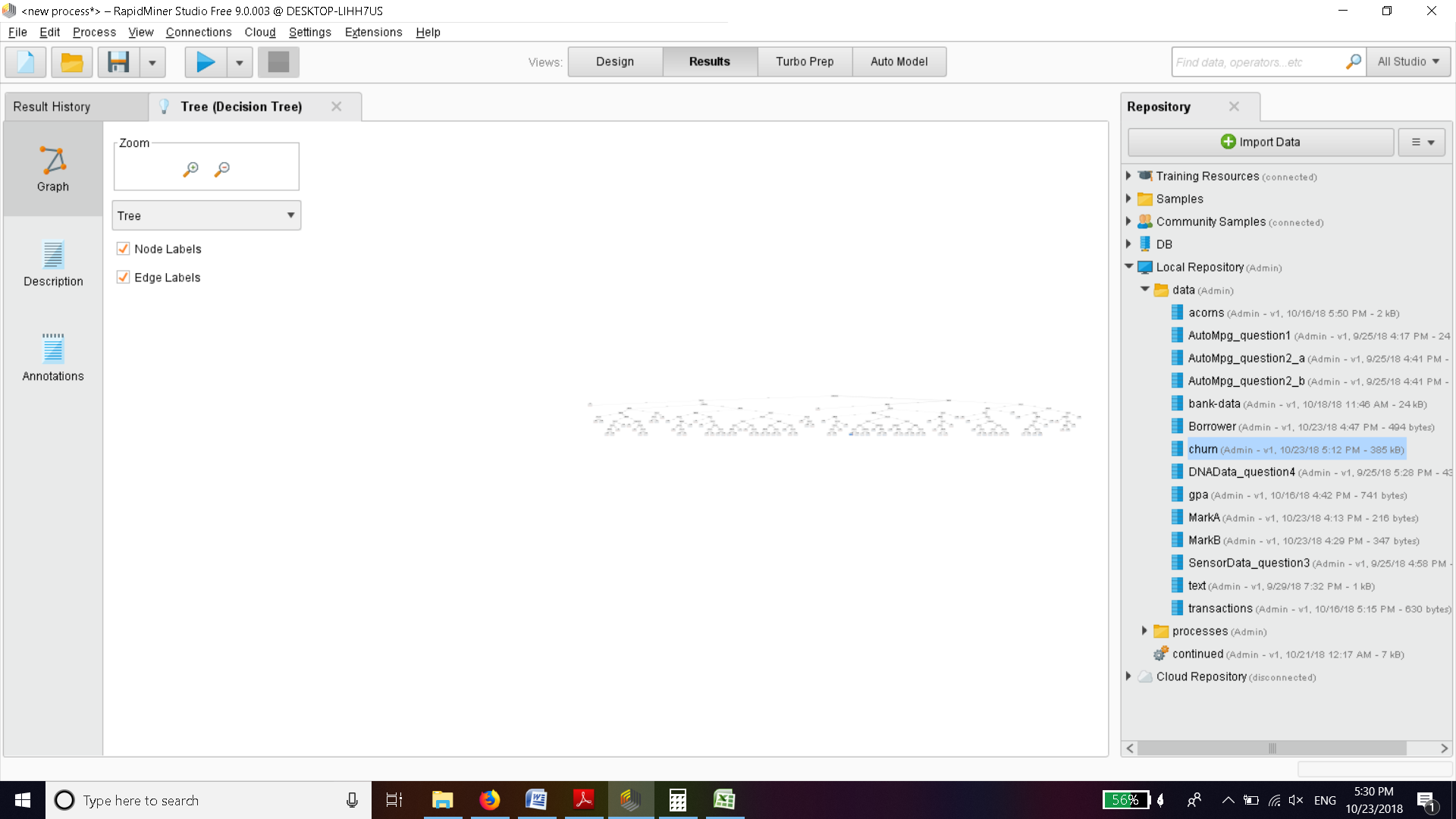


This results in a extremely huge tree with atleast 50 leaf nodes. Any useful analysis is not possible in this method without pruning.

d)

With information gain and pruning results in a good sized tree where analysis is possible quickly

This tree is slightly bigger than (b) but useful analysis is still possible on this tree. The tree also has slightly more depth as compared to gini index.But overall we can say the results are almost the same. In theory there is only 2 % difference between these . [1]However information gain involves intensive logarithmic operations so it might be slower. In general information gain is used for exploratory analysis while gini index is used to reduce misclassification.



References:

[1]<https://www.unine.ch/files/live/sites/imi/files/shared/documents/papers/Gini_index_fulltext.pdf>

“Theoretical comparison between the Gini Index and Information Gain criteria ”