BRIEF REPORT: DECISION TREE, NAÏVE BAYES CLASSIFIER

Motivation:

- To know about RapidMiner tool
- Get used to different operators as well as some classifiers like; Naive bayes, Decision tree and Bayes classifier.
- Learn about model training, Pre-processing, tokenization, punctuation.

Dataset for TASK- I : IMDB ratings and Movie review on more than 49000 movies.

The tasks we have performed as can be seen from the images shown below.

1. **Data Pre-processing** :

    - **Pre-processing: It** refers to the preparing (cleansing and organizing) the raw data to make it suitable for building and training a ML model.

    - **Tokenization:** It is a process to split a string or text into a list of tokens.

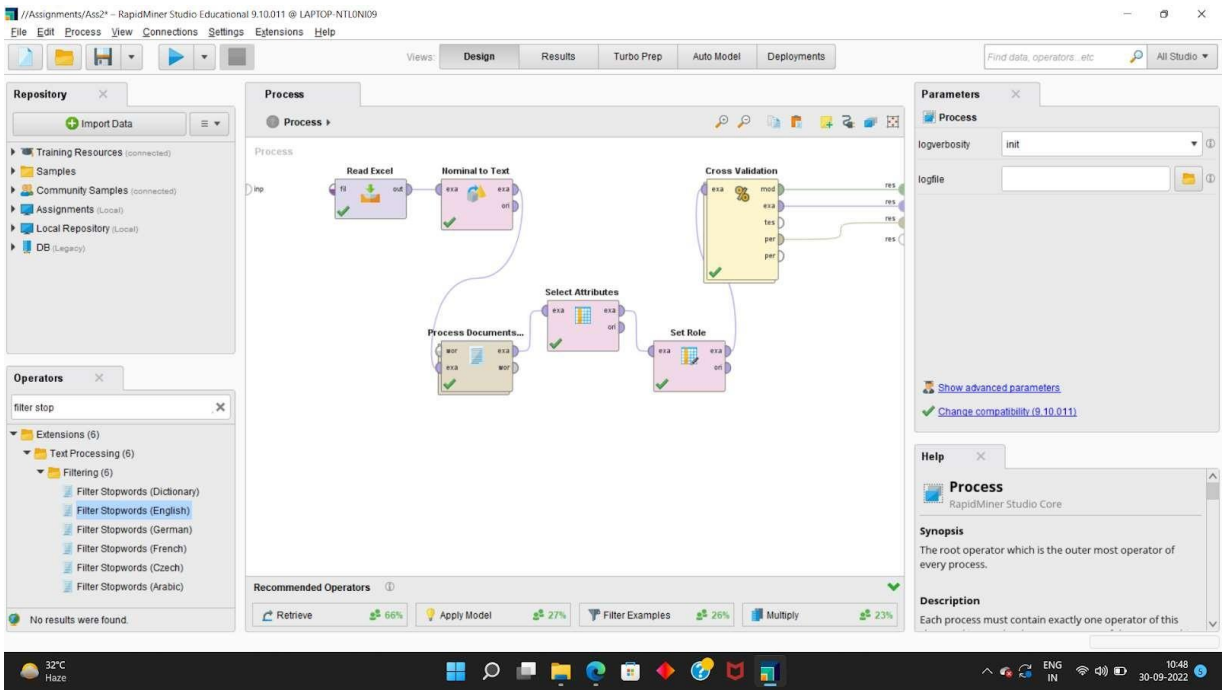
    - **Punctuation Removing: In** this process we replace certain words.

for example
I’m = I am
Let’s = Let us
I've = I Have

- **Frequency Vectorization** : In ML vectorization is a step to extract the feature. The idea is to get some distinct feature out of the text for the model to train on by converting text to numerical vectors.

- **Decision Tree Vs Naive bayes classifier:**

  Decision tree is a discriminative model whereas naive bayes is a generative model. Decision tree is more easy and flexible. In decision tree pruning may neglect some values and can lead the accuracy for a toss.

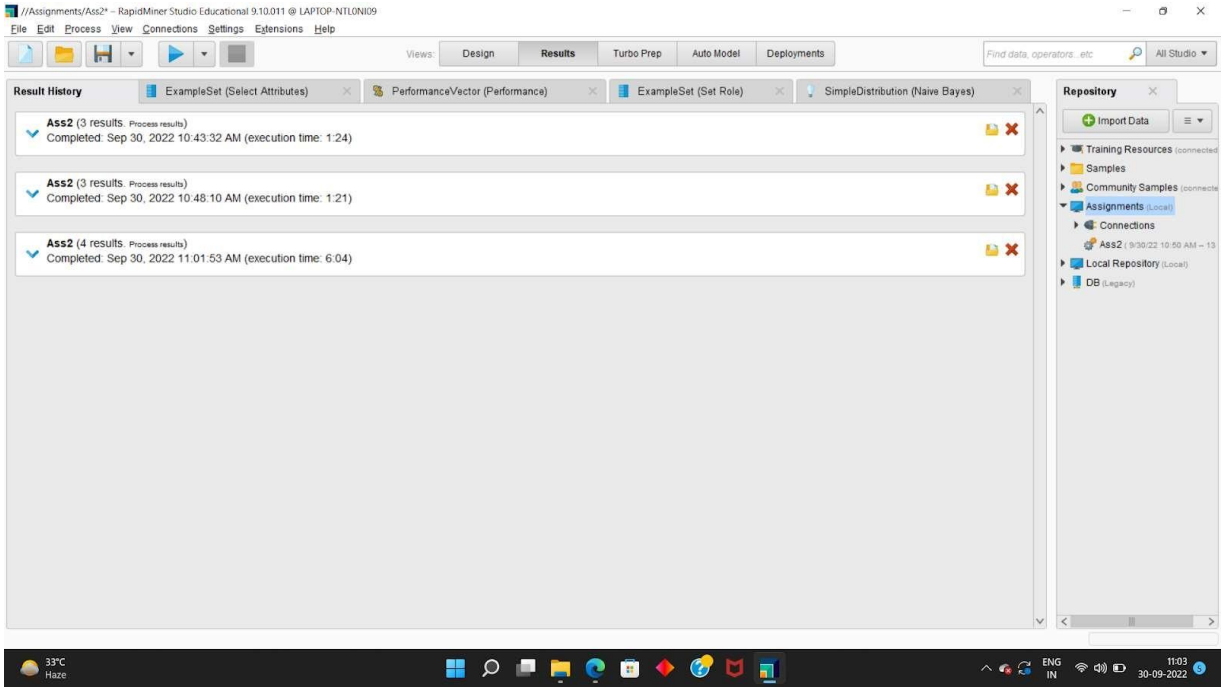


In the following task we reduced the number of examples because we have face some error regarding the entire data analysis which was approximately 49500. So we performed the task over 5000 examples.

It helped us to reduce the processing time and to resolve the error.

Using the filter operator we are able to distinguish between Positive and Negative review.

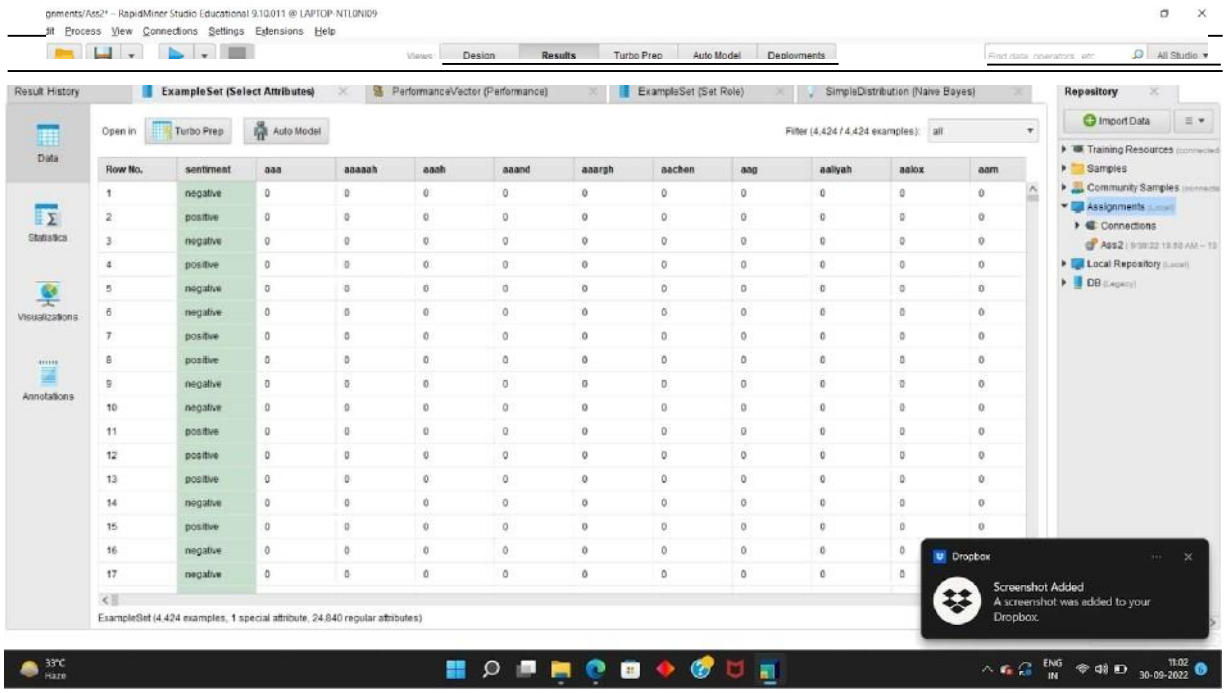
The processing time for about 49000 examples was around 6.5 minutes and error occurred when applying the Set-Role operator.

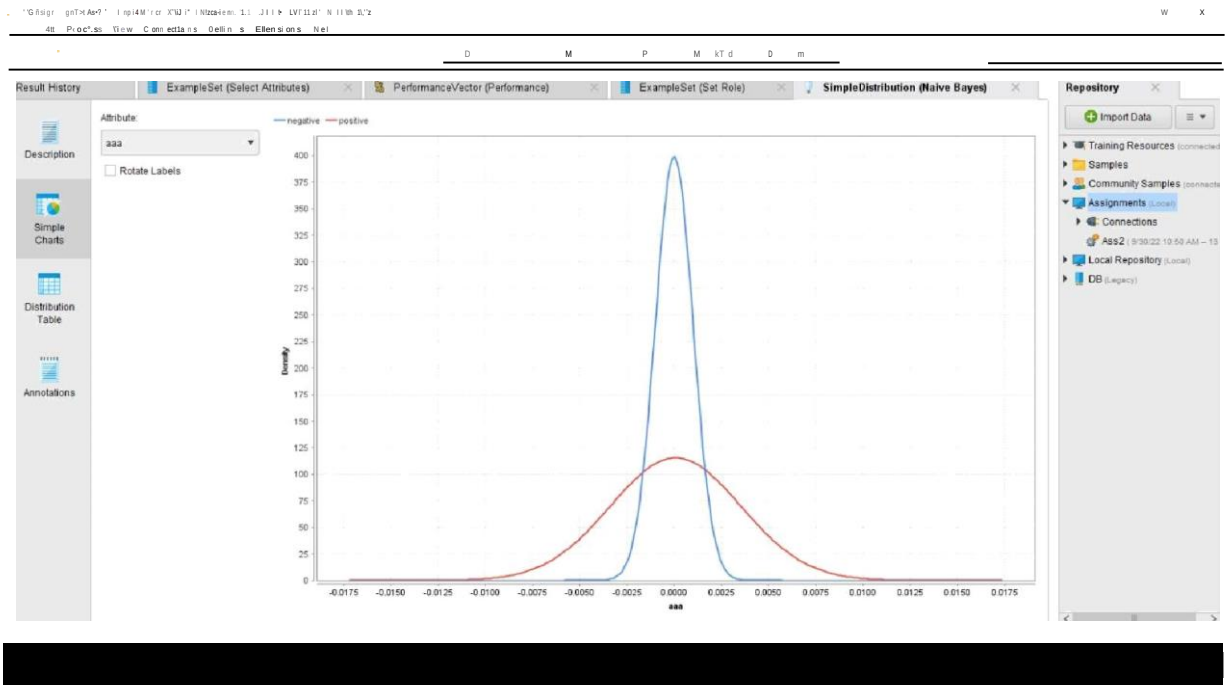


Without applying the filter operator the task performed smoothly.


//Assignments/Ass2 – RapidMiner Studio Educational 9.10.011 @ LAPTOP-NTL0NI09
File Edit Process View Connections Settings Extensions Help
Views:
Design
Results
Turbo Prep
Auto Model
Deployments
Find data, operators...etc
All Studio

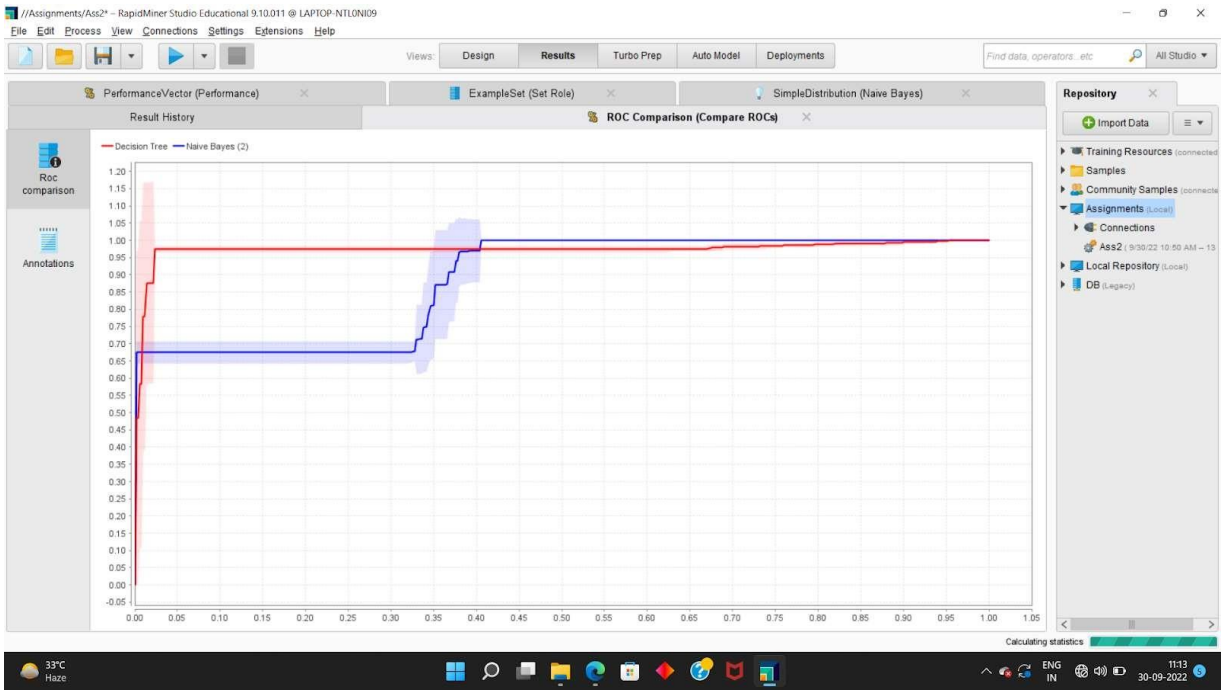

//Assignments/Ass2* – RapidMiner Studio Educational 9.10.011 @ LAPTOP-NTL0NI09
File Edit Process View Connections Settings Extensions Help
Views:
Design
Results
Turbo Prep
Auto Model
Deployments
Find data, operators...etc
All Studio

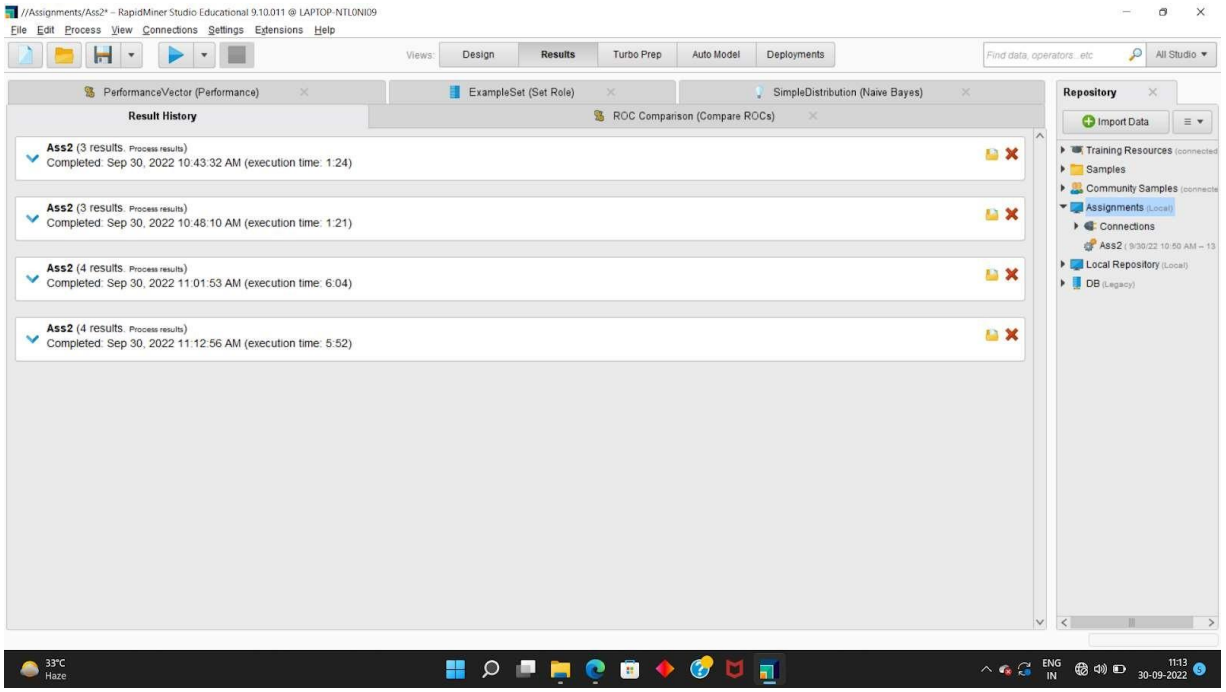




Here as seen from the figure curve of Naive bayes, the classifier distinguishes between positive and negative attributes.
Standard deviation for positive attributes(Red) is greater than negative one(Blue).

Comparing Performances:

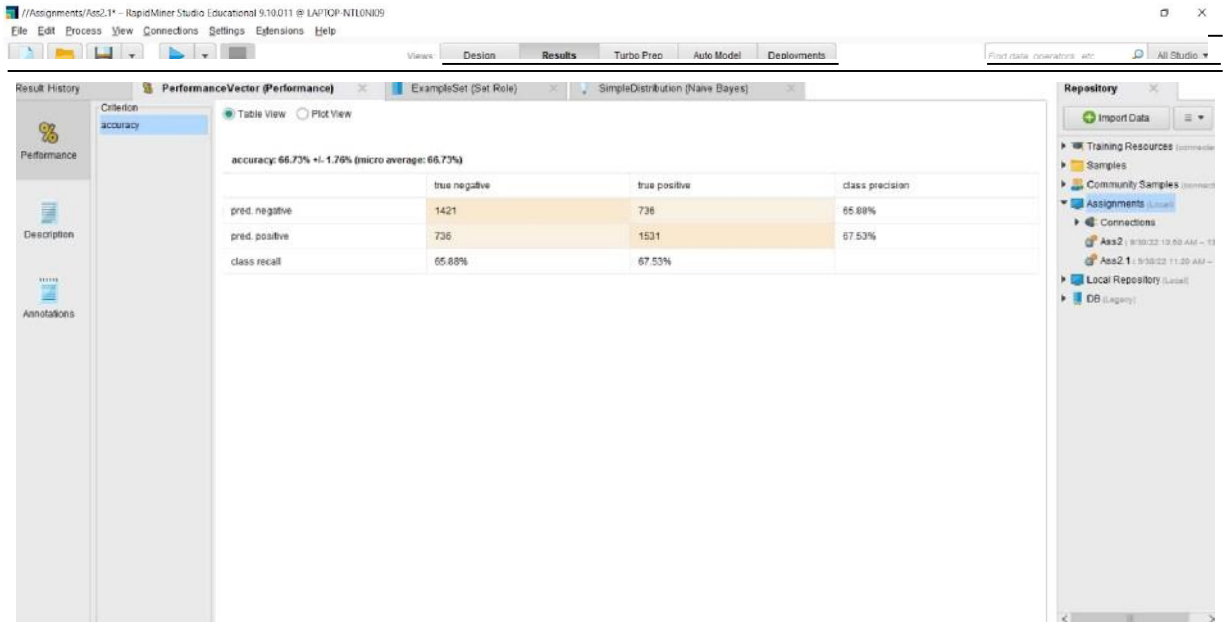




ROC is shown for decision trees and Naive Bayes classifiers. As from the ROC we can predict that Naive classifier maps more false positives. Whereas decision trees predict good for true positives. So, Decision tree is a good classifier as compared to naive bayes classifier w.r.t performance.

Model trained on corpus with stop words being removed.

Model trained on corpus without stop words being removed.



Although there is not much difference between above two images, but still the best result is shown by the model when stop words were being removed.

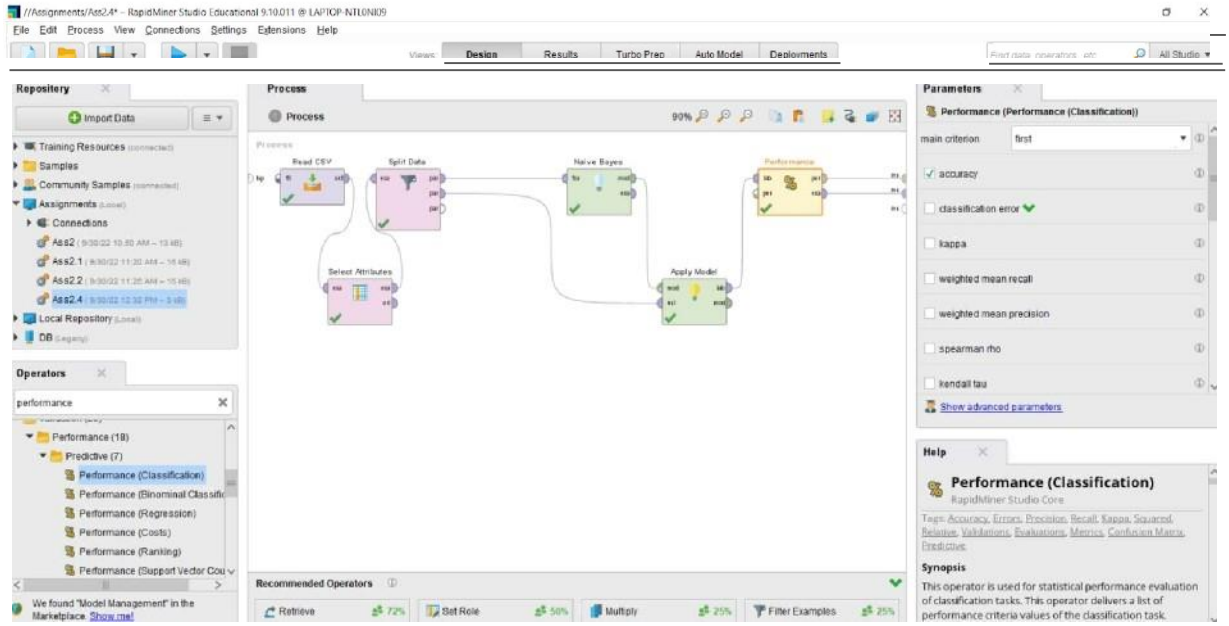
Accuracy: 66.73%(Note: we have taken less number of examples because more examples were taking more processing time)
Recall: 65.88%
Precision: 65.88%

Task 2: Banknote **authentication**

The dataset which we received was in document format(.txt). We converted it in spreadsheet format(.csv)
Some modifications were done like addition of attributes heading of each column.
The attributes were,
1st attribute- Variance
2nd attribute- Skewness
3rd attribute- Curtosis
4th attribute- Entropy

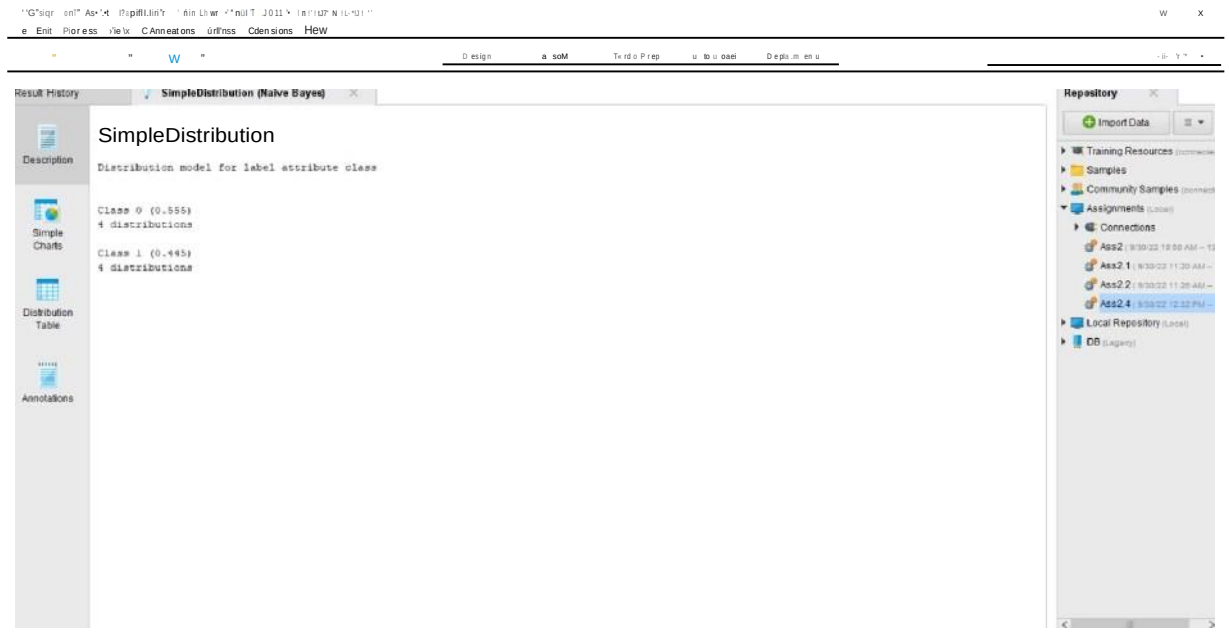
The Sth column was the class. In Rapidminer we converted the class attribute as ‘label’ as there were only two values ‘0’ and ‘1’

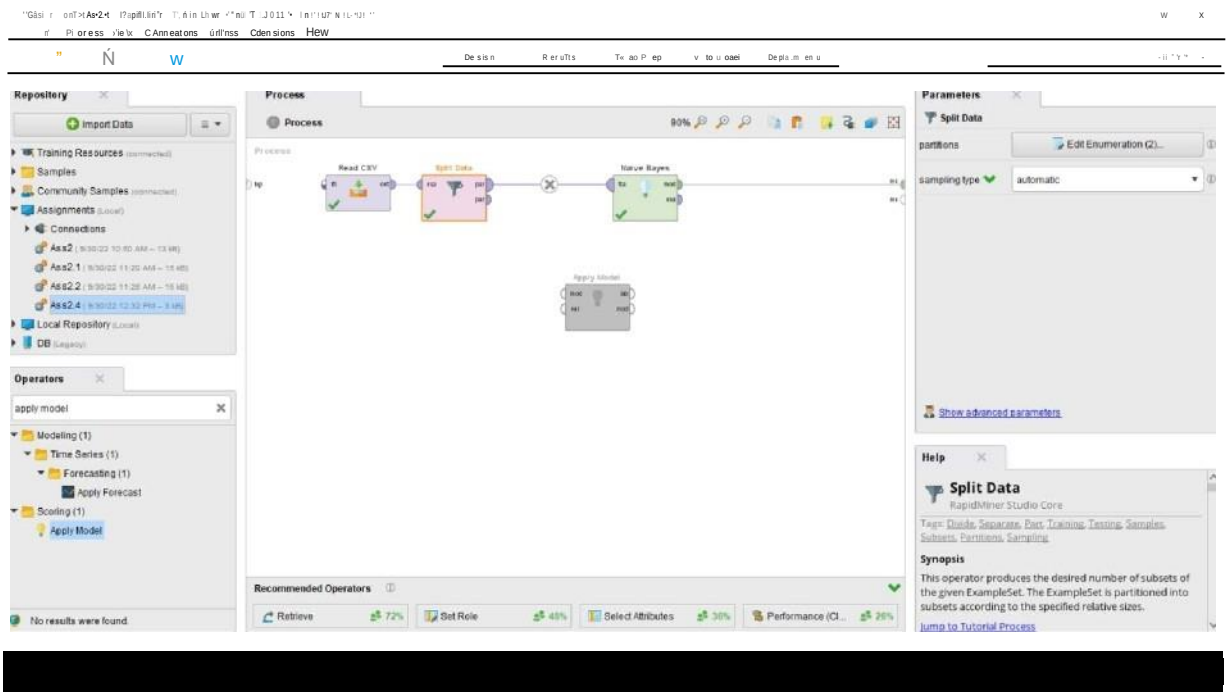
The following image will show the set of operators we use.



1. Model Training:

   We trained the model through a naive bayes classifier. The following image shows the distribution of results with a set of operators.

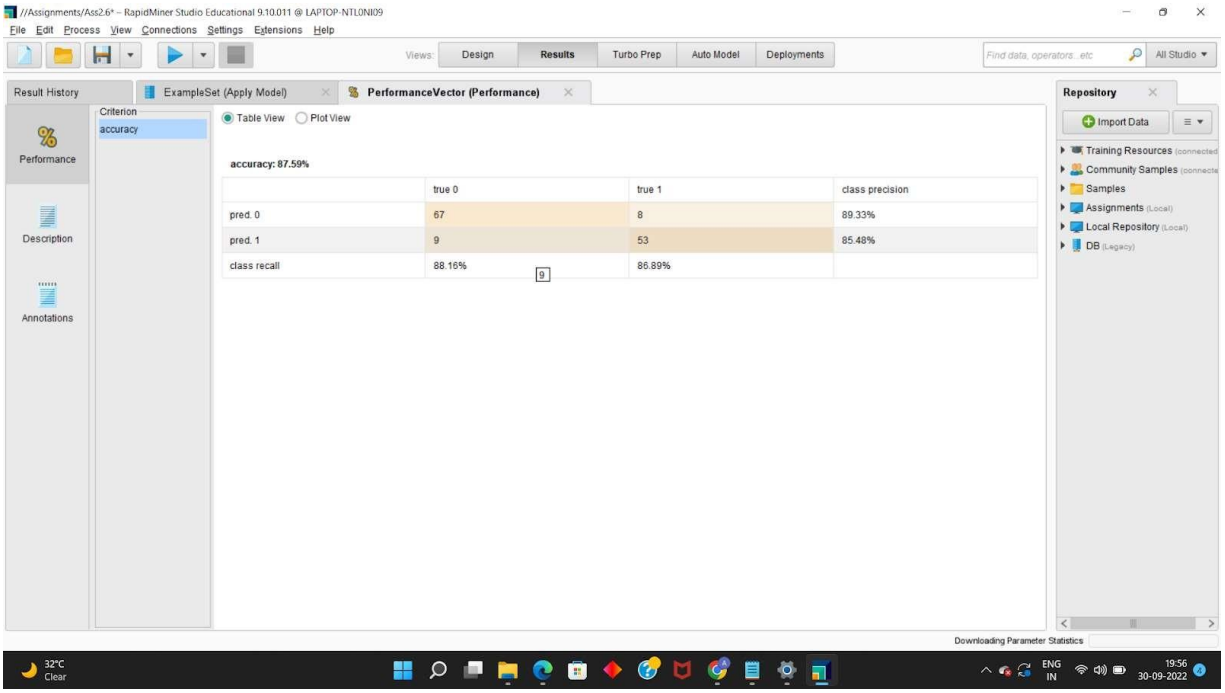




We applied 4 splits

1. 0.6 and 0.4
2. 0.7 and 0.3
3. 0.8 and 0.2
4. 0.9 and 0.1

The best results were obtained for 90% and 10% splitting, The accuracy was maximum song all 4. The accuracy was around 87%. The other three splits were showing accuracy around 84%.

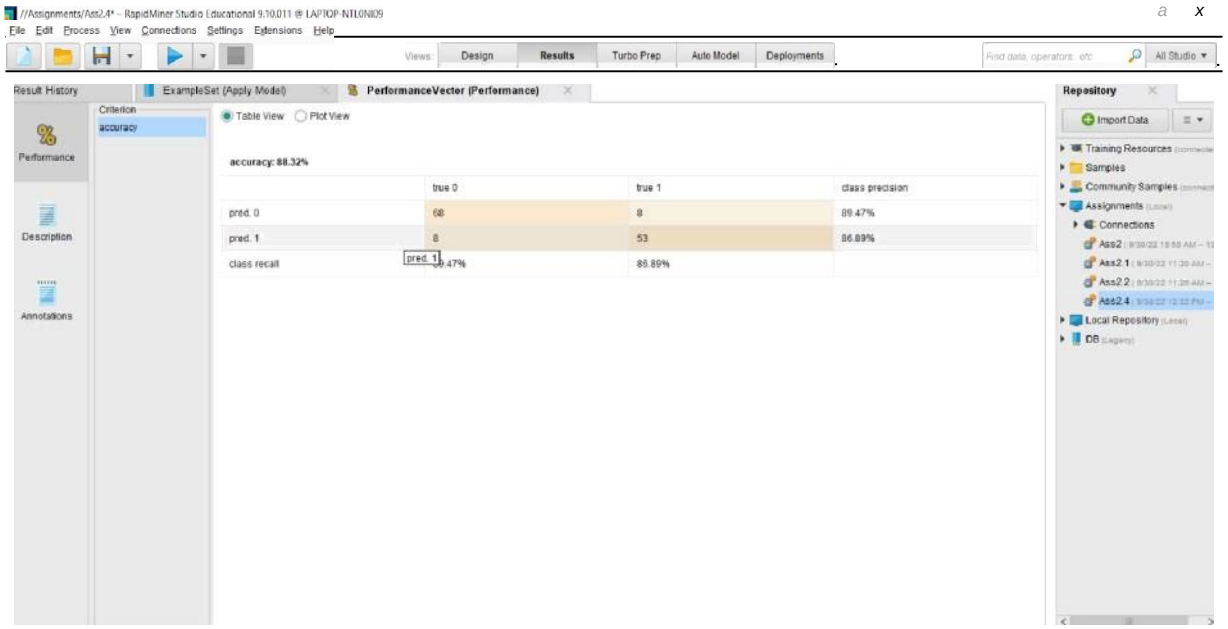


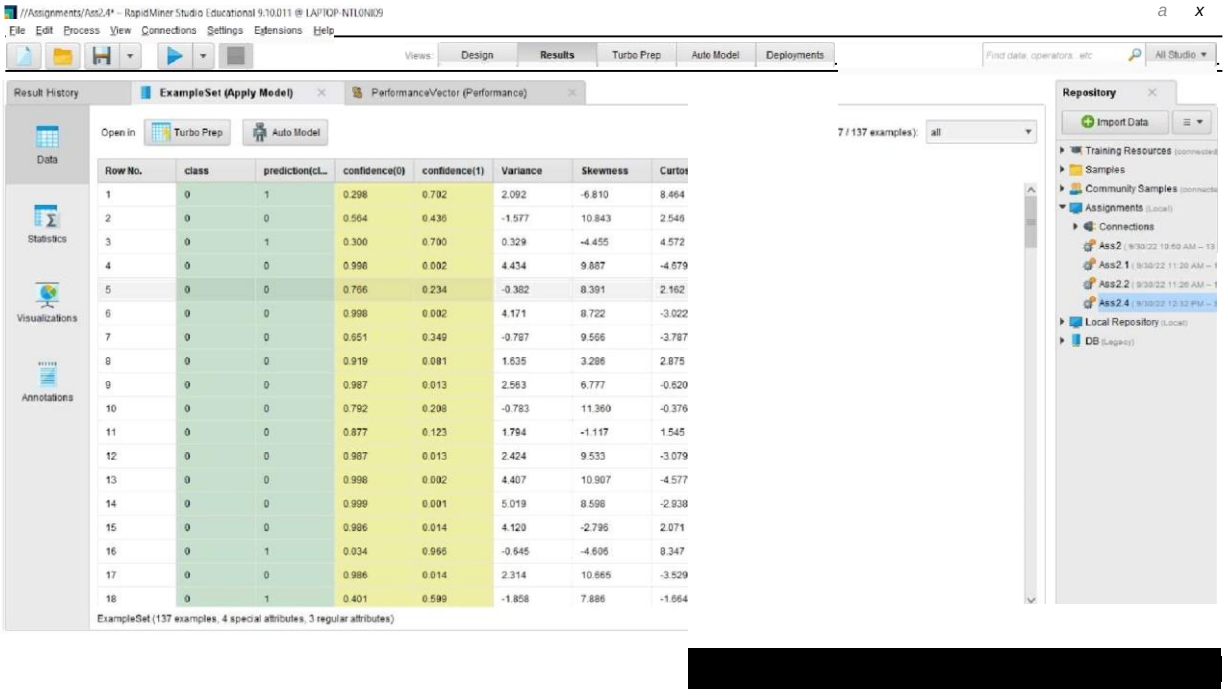
We carried further predictions with 90% and 10% split data.

To select the attributes, we ’Select attributes' operator. We changed the attribute filter to ‘subtype’ and then did the following activities.

1. **Training with 1st three predictors.(Variance,Skewness,Curtosis)**

   The following shows the result

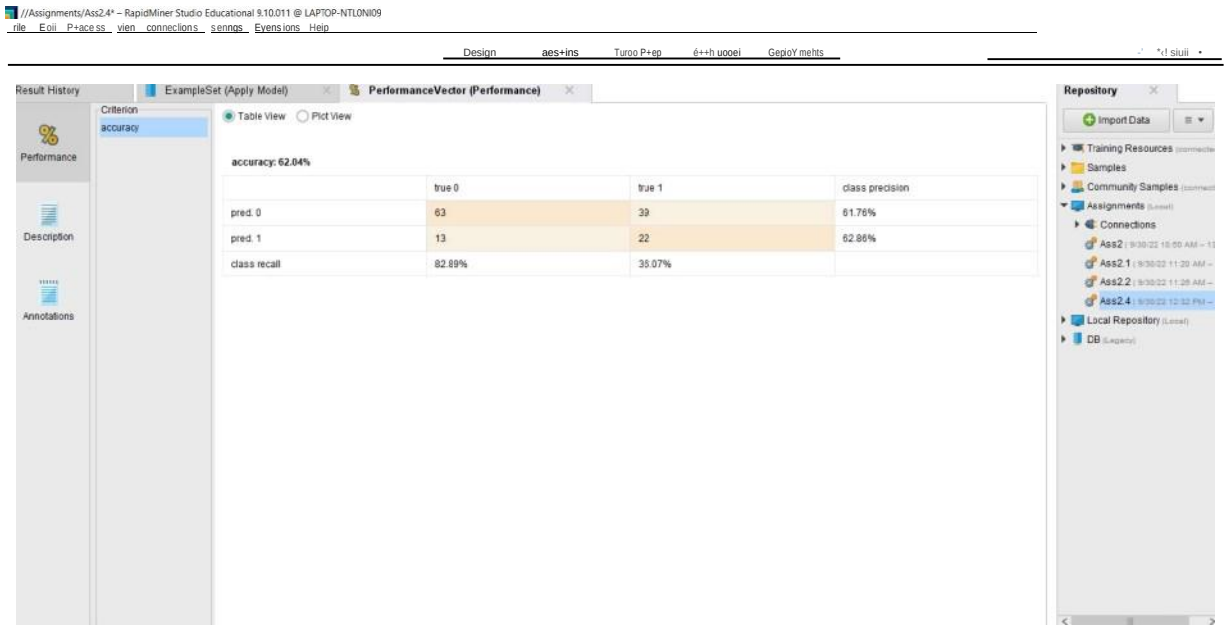




The accuracy for 1st three predictors is 88.32%

2. Training with last 3 predictors (Skewness, Curtosis, Entropy):

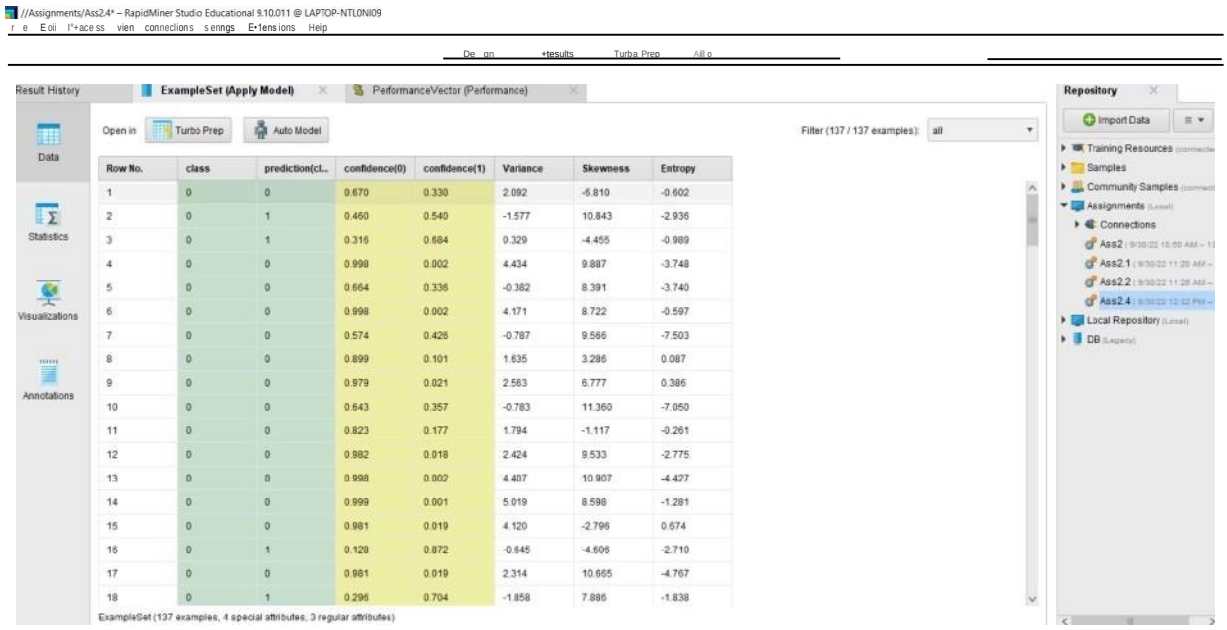
The following image shows the result

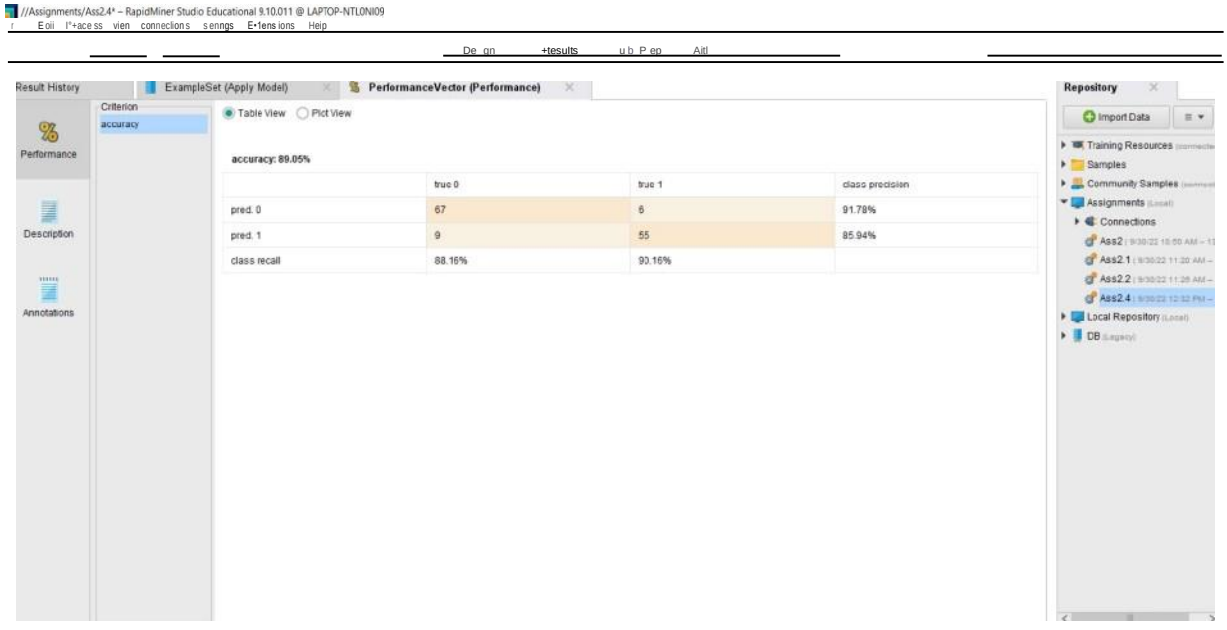


The accuracy for last 3 predictors is 62.04%

Training with 1 2 and 4 predictors (Variance, Skewness, Entropy):

The following image shows the result

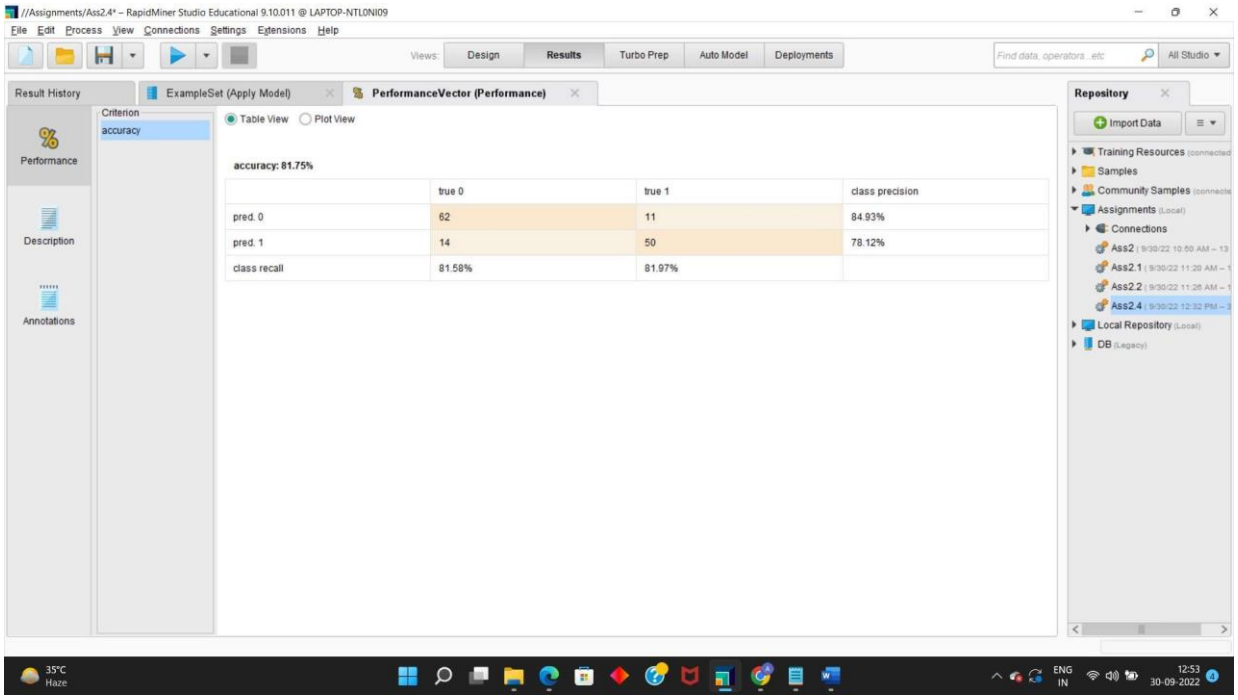


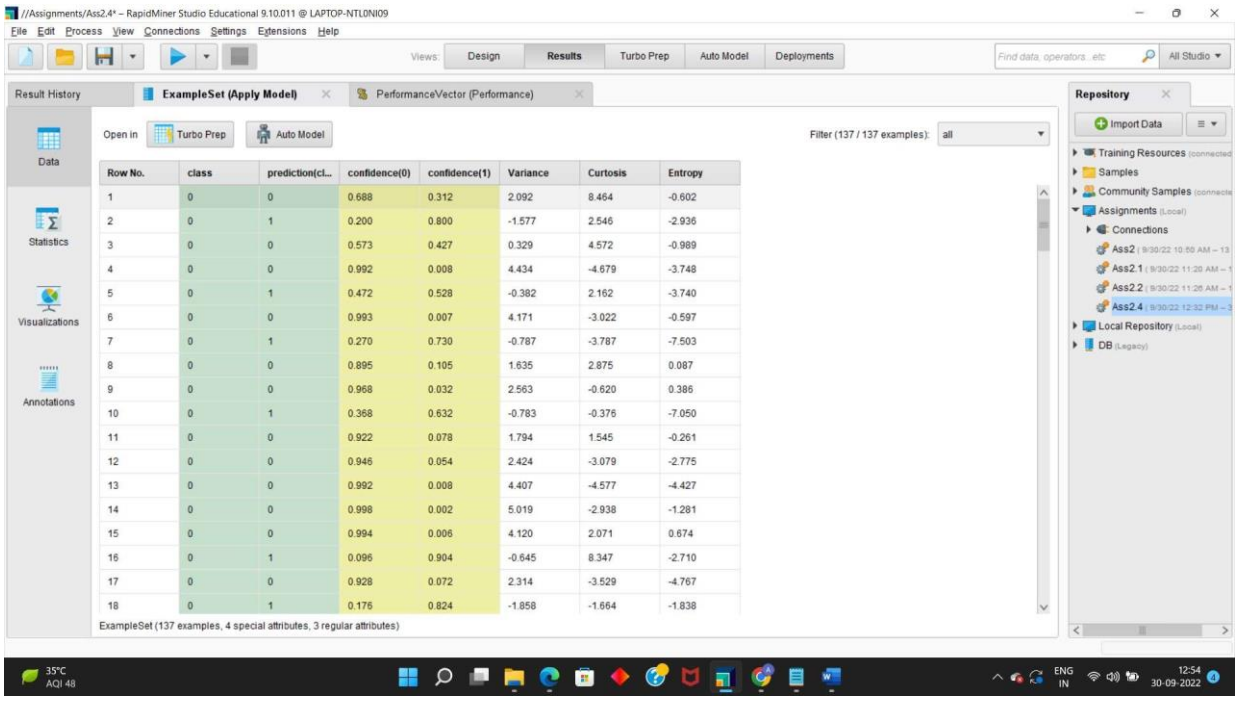


The accuracy for 1*t,2nd and 4th predictors is 89.05%

Training with 1°'3 and 4 predictors (Variance, Curtosis, Entropy):

The following image shows the result



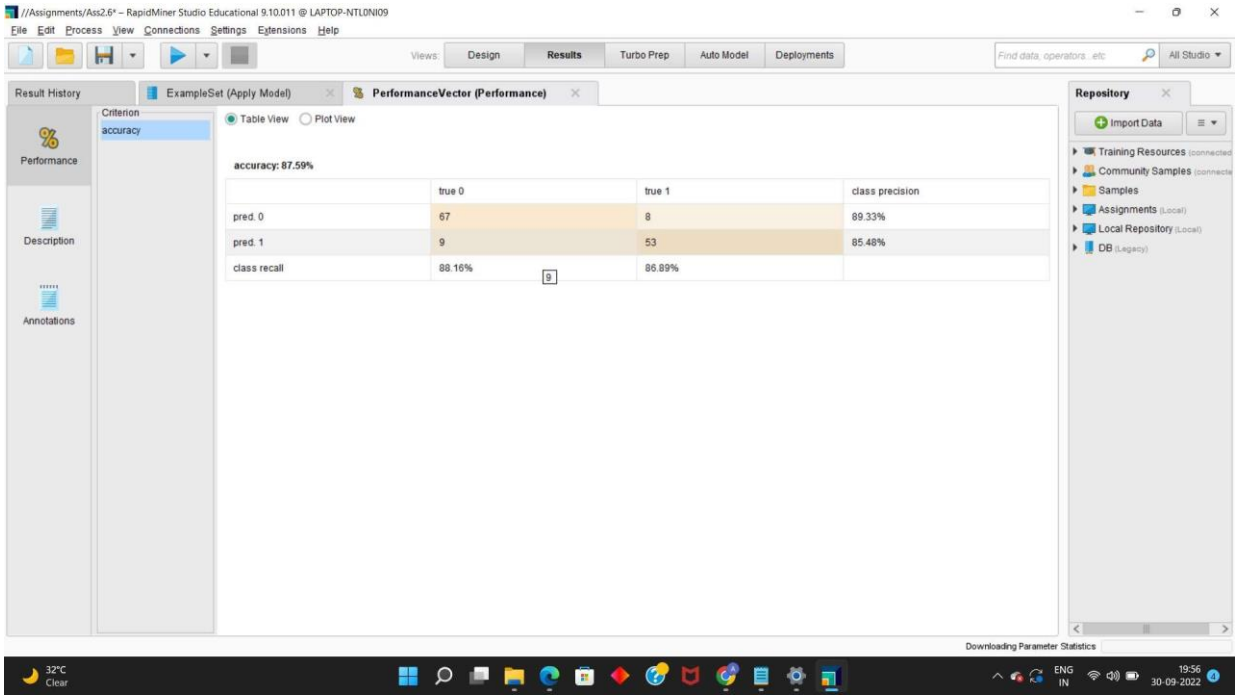


The accuracy is 81.75%

3.Comparing Performances:

From above set of models, the best model is 3rd model having predictors 1st, 2nd and 4th predictors. It shows the maximum accuracy of 89.04%. Not only that,

but if we compare with Naive Bayes classifier(having all four attributes), the 3rd model is still the best because naive bayes classifier has accuracy of 87.58%



The above image shows the result of naive bayes classifier

Therefore the accuracy,recall,precision for best model is:

Accuracy: 89.05%

Recall: 88.16%

Precision:91.78%

The result(confusion matrix) has already been shown for all the variants of model.