**Data Mining project**

**IS665 – spring 2020**

**Vishal Goyal**

This Data Mining project implemented by selecting a dataset from the available public sources, to implement Logistic Regression a mining algorithm using Rapidminer software.

**Dataset Description:**

Data Source: - <https://www.kaggle.com/jakeshbohaju/brain-tumor>

Brain Tumor:

This is a brain tumor feature dataset including five first-order feature and eight texture feature and four quality assessment parameters with the target level.

Dataset contains:

Total Columns – 19

Total Rows - 1644

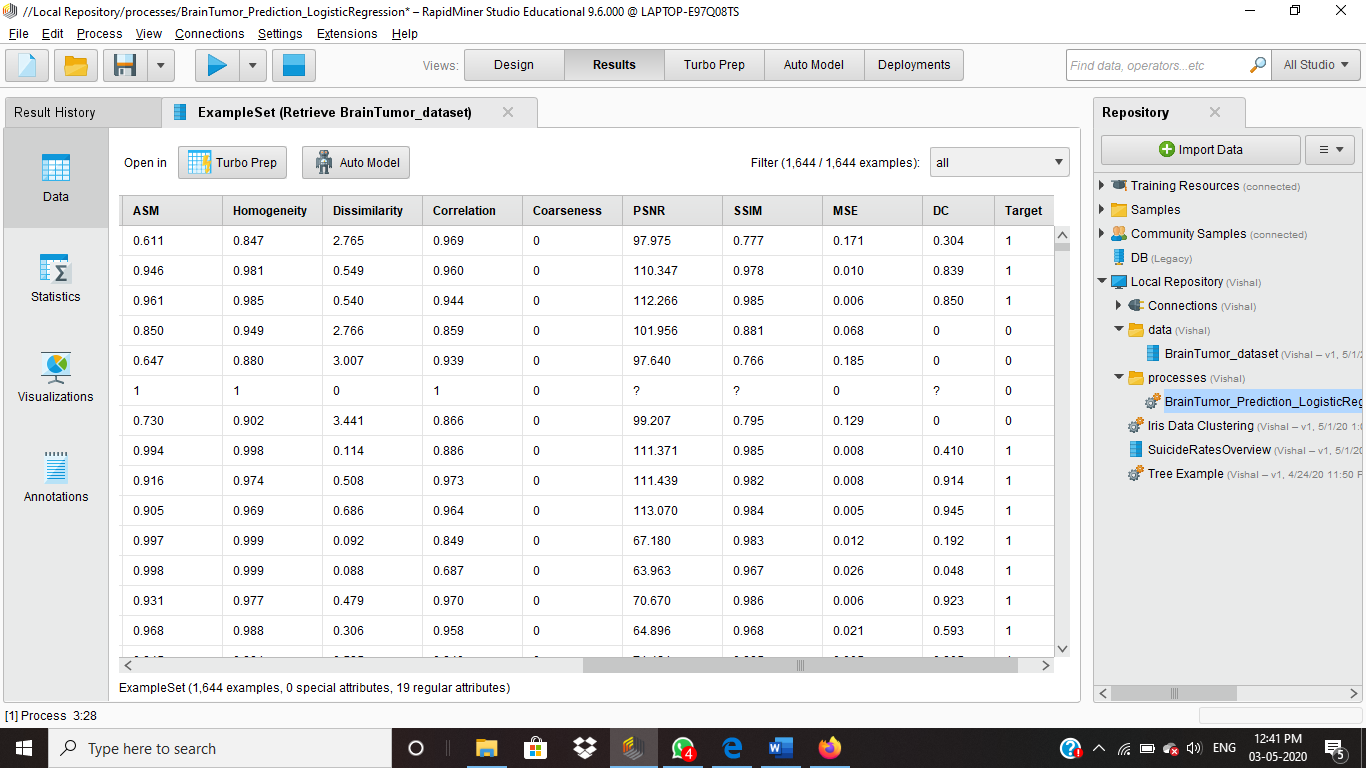
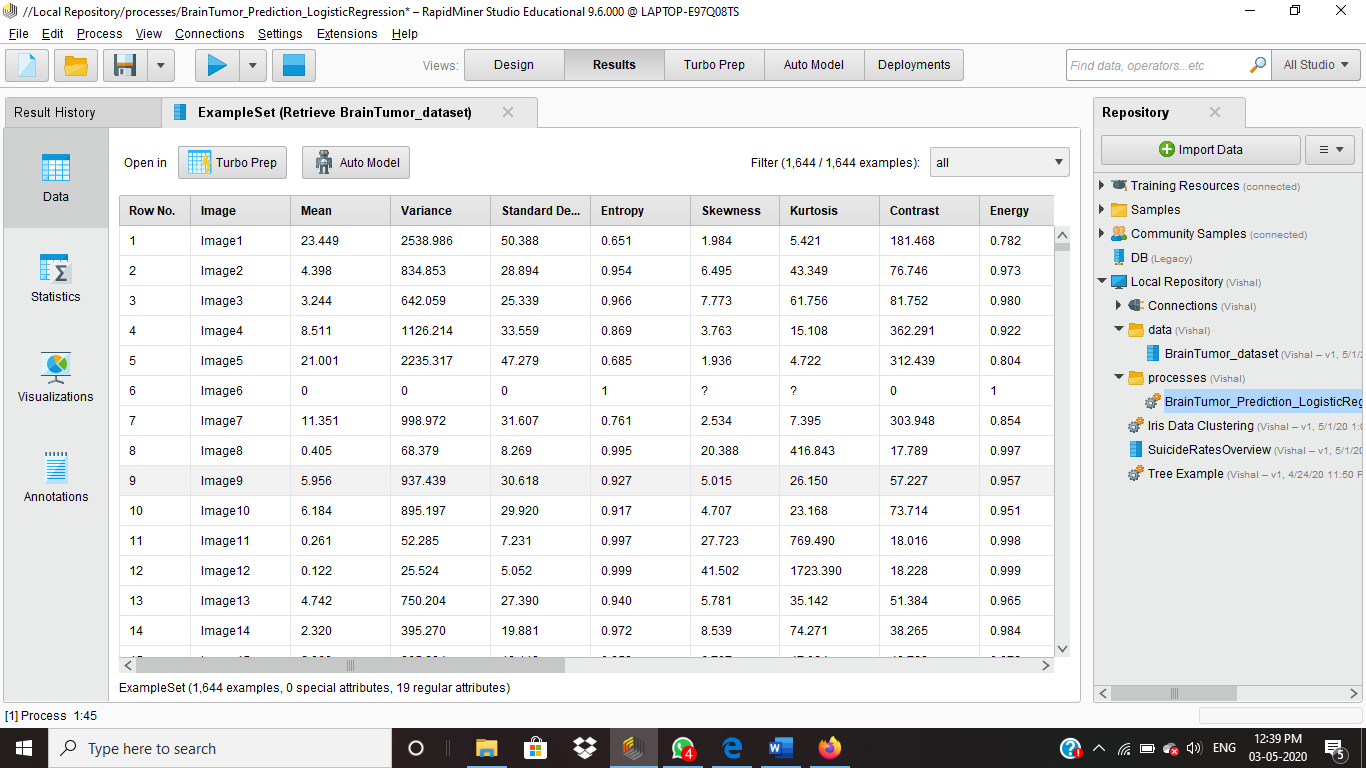
* First Order Features
  + Mean
  + Variance
  + Standard Deviation
  + Skewness
  + Kurtosis
* Second Order Features
  + Contrast
  + Energy
  + ASM (Angular second moment)
  + Entropy
  + Homogeneity
  + Dissimilarity
  + Correlation
  + Coarseness
* Features Assessment Parameter
  + PSNR (Peak signal-to-noise ratio)
  + SSIM (Structured Similarity Index)
  + MSE (Mean Square Error)
  + DC (Dice Coefficient)

Image column defines image name and Target column defines either the image has tumor or not (1 = Tumor, 0 = Non-Tumor)

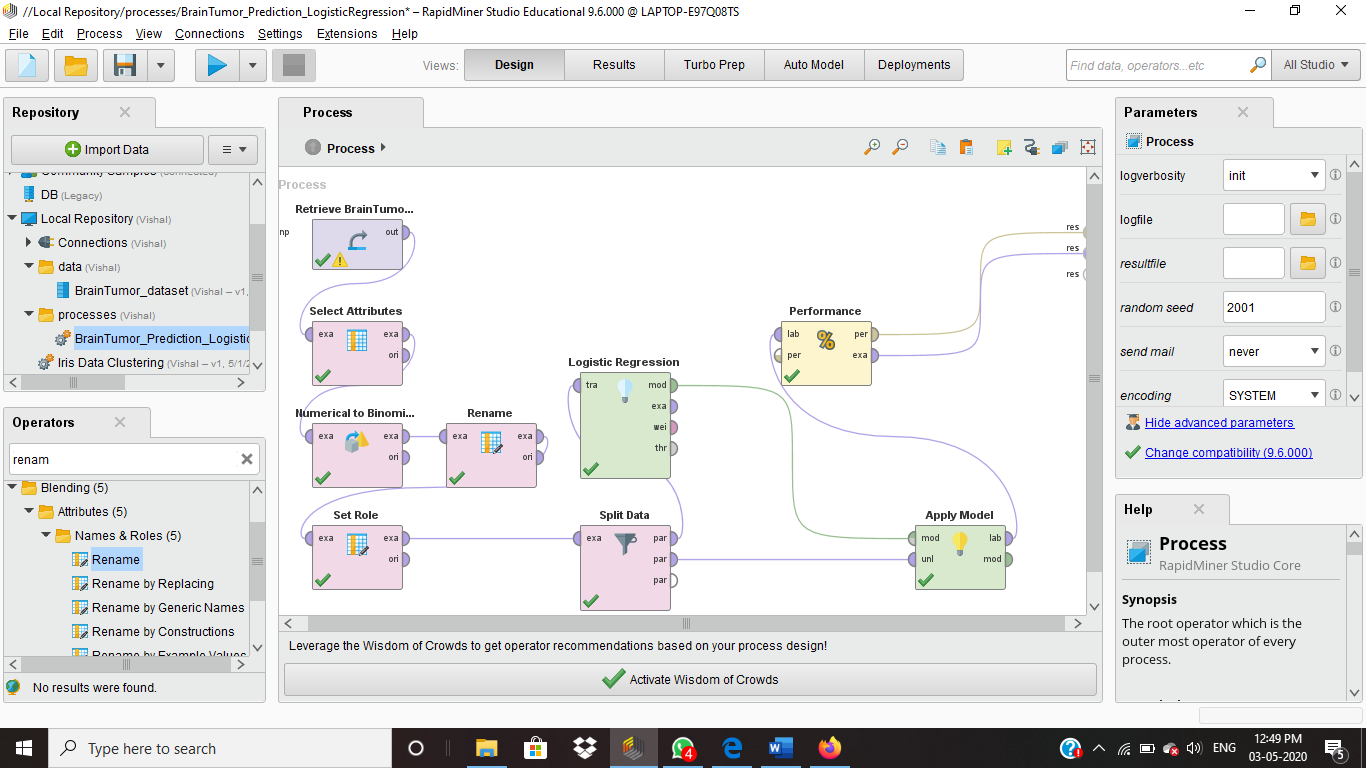
**Error Resolved while Loading Data: -**

As data in column PSNR contained 16 rows with ‘inf’ value, which was replaced by 0 to convert it to real data type.

**Dataset After loading in Rapid Miner: -**



**RapidMiner Model for Logistic Regression: -**



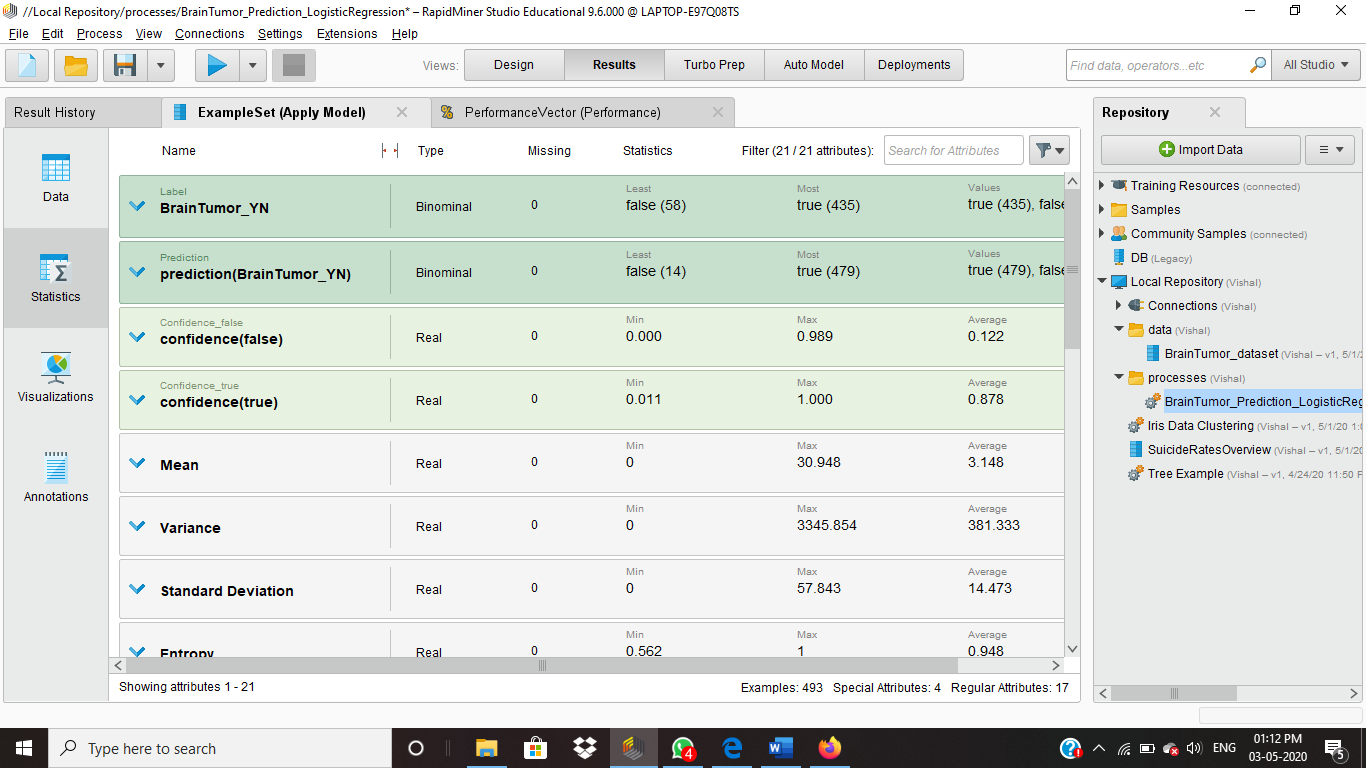
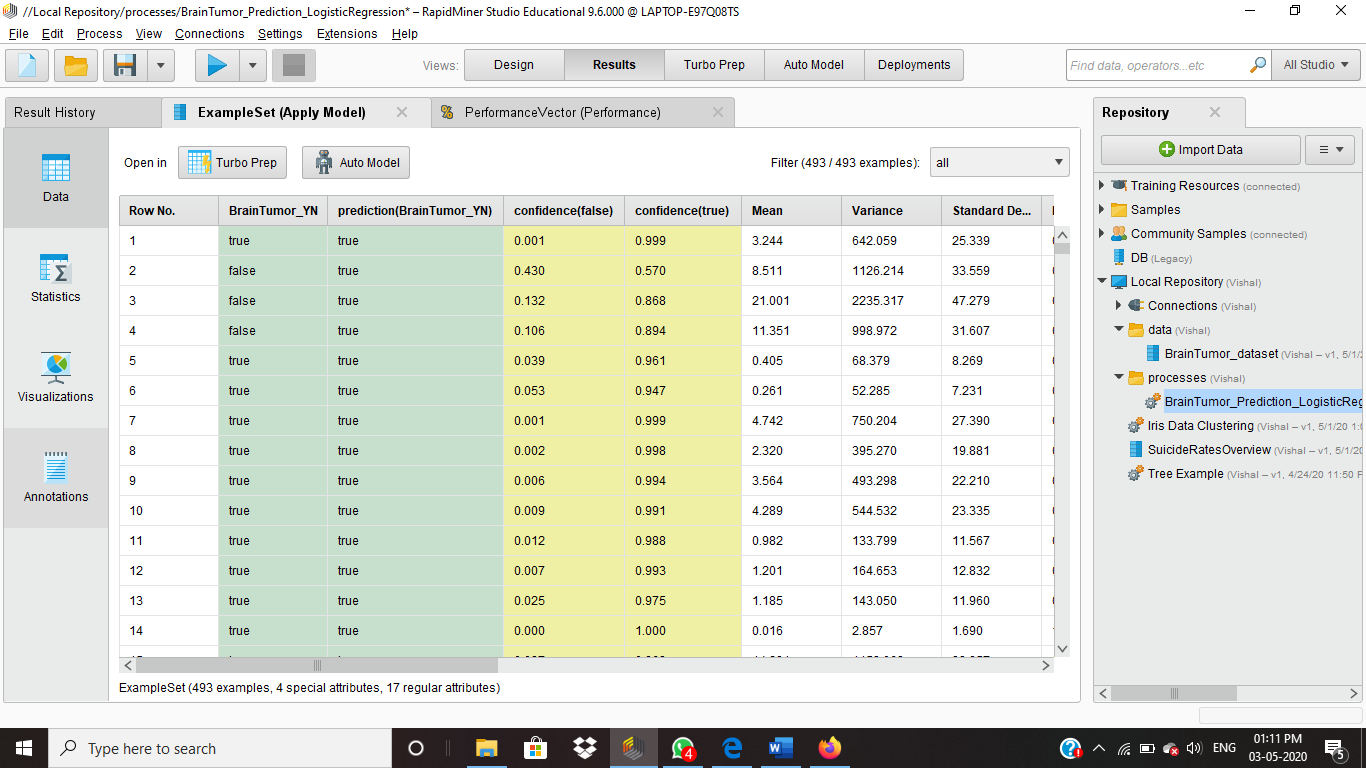
Operators Used: -

* Select Attribute: This operator is used for selecting specific columns for performing the further processing. In this model, ‘IMAGE’ column removed to analyze remaining dataset.
* Numerical to Binominal: This Operator is used to convert numerical columns to binominal. This attribute is used in this model as target values are given in numerical as 0s and 1s. Which will not be considered further as label for dependent value.
* Rename: Target column is renamed as BrainTumor\_YN to make better sense of target value.
* Set Role: This is used to define target value which is required to predict.
* Split Data: Data is split by 7:3 ratio for further processing.
* Logistic Regression: 70% of split data output is given to logistic regression for data processing.
* Apply Model: 70% Processed data and 30% unlabeled data is passed to apply model for output.
* Performance: This Data is given to performance operator to check the performance of output.

**Result Analysis: -**

Result of applied model:

Below figure shows results of Given data, Brain Tumor is present or not and predicted values of tumor is present or not.



Performance of applied model:

Accuracy of Brain tumor not present is shown as 10.34%, but as Accuracy of tumor present is 98.16% it is shown as good model for supervised learning.

If results would have been better accuracy for brain tumor not present then we could have used Threshold attribute to create better accuracy for tumor is present. Here accuracy of Brain tumor is present should be first concern so Threshold attribute is not required.

As shown in below picture total accuracy is given as 87.83% of this model.

