****

# DEPARTMENT OF COMPUTER SCIENCE

# MACHINE LEARNING

# CUSTOMER REPURCHASE PREDICTION

# FINAL REPORT

|  |  |
| --- | --- |
| **NAME:** | **SHAHZAIB BUTT (11495)** |
| **SUBMITTED TO:** | **SIR SHARRAF HUSSAIN** |
| **DATE** | **9.MAY.2019** |

**ACKNOWLEDGMENT**

I have taken efforts in this project. However, I am highly indebted to “**Sir Sharraf Hussain”** for her guidance and constant supervision and as well as for providing the necessary information regarding the project and also for her support in completing the research. I would like to express my gratitude towards my parents and members of Iqra University for their kind co-operation and encouragement which helped me in completion of the research project. My thanks and appreciations go to my colleagues in developing the research project and people who willingly helped me out with their abilities.

***INDEX***

* ***Introduction***
* ***Tool***
* ***Features Used***
* ***Machine Learning Techniques***
  + ***Information Based Learning***
    - ***Iterative Dichotomiser 3 (ID3) Algorithm***
      * ***Model***
      * ***Decision Tree View***
      * ***Classified Data***
      * ***Model Accuracy***
* ***Similarity Based Learning***
  + ***K Nearest Neighbor Algorithm***
    - ***Model***
    - ***Classified Data***
    - ***Model Accuracy***
* ***Probability Based Learning***
  + ***Naïve Bayes Algorithm***
    - ***Model***
    - ***Classified Data***
    - ***Model Accuracy***
* ***Error Based Learning***
  + ***Logistic Regression Algorithm***
    - ***Model***
    - ***Predictive Data***
    - ***Model Accuracy***
* ***Comparing All Machine Learning Techniques Result***
* ***Best Performed Machine learning Technique***

**INTRODUCTION:**

In this project I am predicting the business problem through customer dataset. I am also discussing the goals of business problem. The main focus of this project is that, we have data set which have almost thousands above records of different customers. So, I have need to predict the customers repurchase possibilities.

**TOOL:**

KNIME, the Konstanz Information Miner, is a free and open-source data analytics, reporting and integration platform. KNIME integrates various components for machine learning and data mining through its modular data pipelining concept.

**FEATURES USED:**

• **PARTITIONING:**

The input table is split into two partitions (i.e. row-wise), e.g. train and test data. The two partitions are available at the two output ports.

• **COLOR MANAGER:**

Colors can be assigned for either nominal (possible values have to be available) or numeric columns (with lower and upper bounds). If these bounds are not available, a '?' is provided as a minimum and maximum value. The values are then computed during execute. If a column attribute is selected, the color can be changed with the color chooser.

• **COLOR APPENDER:**

Assigns an existing color model to a table. If a color model was configured for a dataset and this color model should be reused, the model out port of the Color Manager should be connected to the model in port of the Color Appender.

• **DECISION TREE LEARNER:**

The decision tree is a classic predictive analytics algorithm to solve binary or multinomial classification problems

• **DECISION TREE PREDICTOR:**

This node uses an existing decision tree (passed in through the model port) to predict the class value for new patterns.

• **SCORER:**

Compares two columns by their attribute value pairs and shows the confusion matrix, i.e. how many rows of which attribute and their classification match. Additionally, it is possible to highlight cells of this matrix to determine the underlying rows.

• **RENAME COLUMN:**

Rename column names or change their types. The dialog allows you to change the name of individual columns by editing the text field or to change the column type by picking one of the possible types in the combo box. Compatible types are those to which the cells in a column can be either safely cast or transformed to. A configuration with a red border indicates that the configured column does not longer exist.

• **K NEAREST NEIGHBOR:**

Classifies a set of test data based on the k Nearest Neighbor algorithm using the training data. The underlying algorithm uses a KD tree and should therefore exhibit reasonable performance. However, this type of classifier is still only suited for a few thousand to ten thousand or so training instances. All (and only) numeric columns and the Euclidean distance are used in this implementation. All other columns (of non-numeric type) in the test data are being forwarded as-is to the output.

• **NORMALIZER (PMMl):**

This node normalizes the values of all (numeric) columns. In the dialog, you can choose the columns you want to work on. PMML Integration in KNIME. These input and output ports are used to pass the PMML documents describing actions taken by the corresponding node. As one of few products, KNIME also supports PMML ensemble models.

• **NAIVE BAYES LEARNER AND PREDICTOR:**

The node creates a Bayesian model from the given training data. It calculates the number of rows per attribute value per class for nominal attributes and the Gaussian distribution for numerical attributes. The created model could be used in the naive Bayes predictor to predict the class membership of unclassified data. The node displays a warning message if any columns are ignored due to unsupported data types. For example, Bit Vector columns are ignored when the PMML compatibility flag is enabled since they are not supported by the PMML standard.

• **LOGISTIC REGRESSION LEARNER AND PREDICTOR:**

Predicts the response using a logistic regression model. The node needs to be connected to a logistic regression node model and some test data. It is only executable if the test data contains the columns that are used by the learner model. This node appends new columns to the input table containing the prediction for each row.

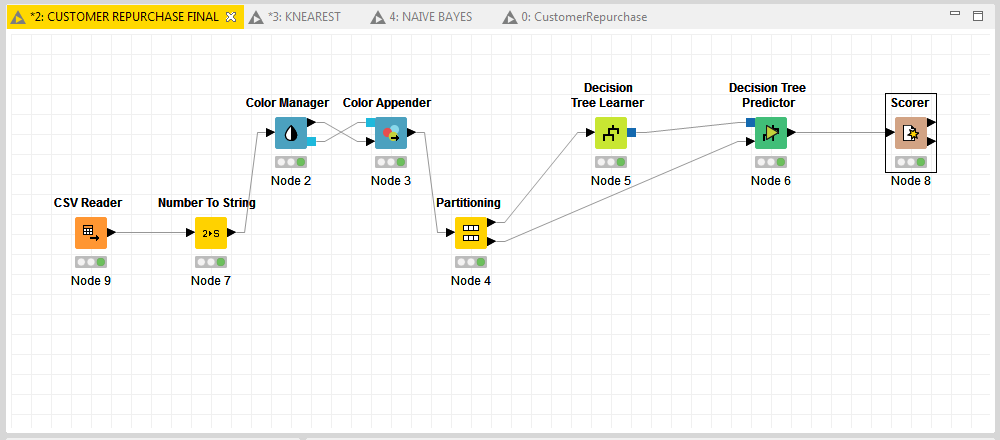
**MACHINE LEARNING TECHNIQUES:**

**INFORMATION BASED LEARNING:**

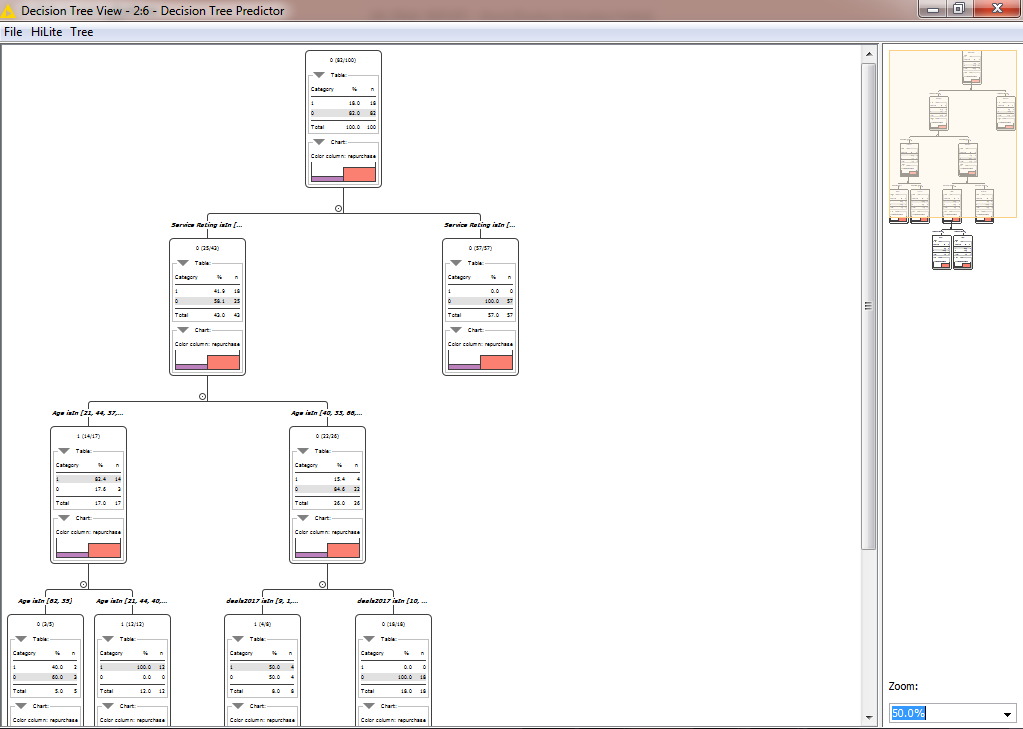
**ITERATIVE DICHOTOMISER 3 (ID3):**

In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan used to generate a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithms, and is typically used in the machine learning and natural language processing domains.

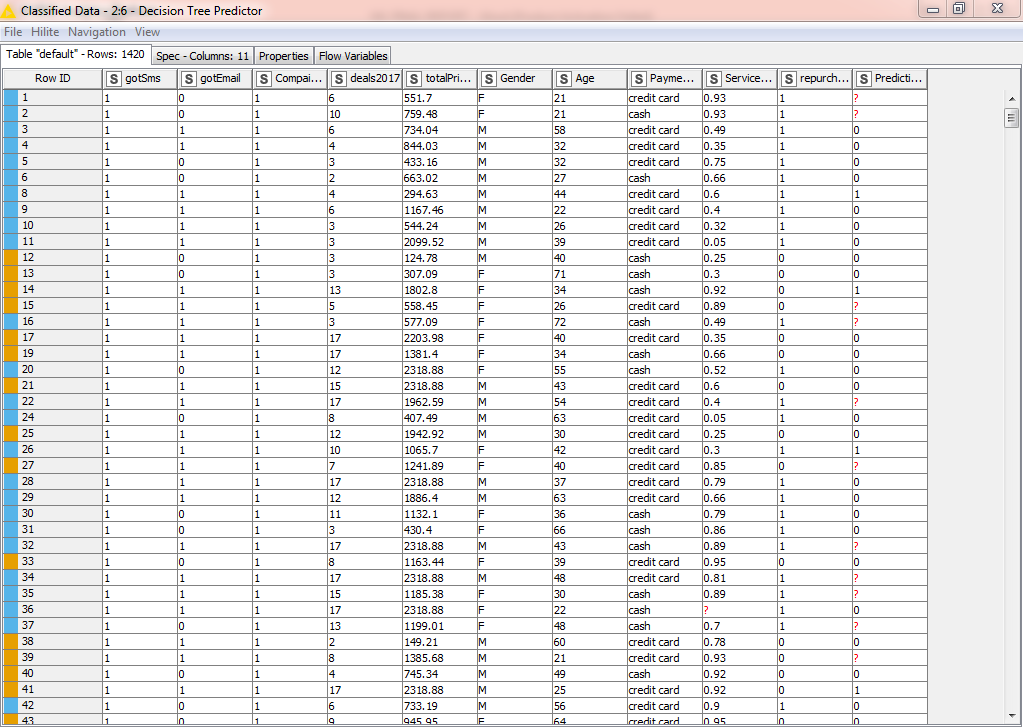
**MODEL:**



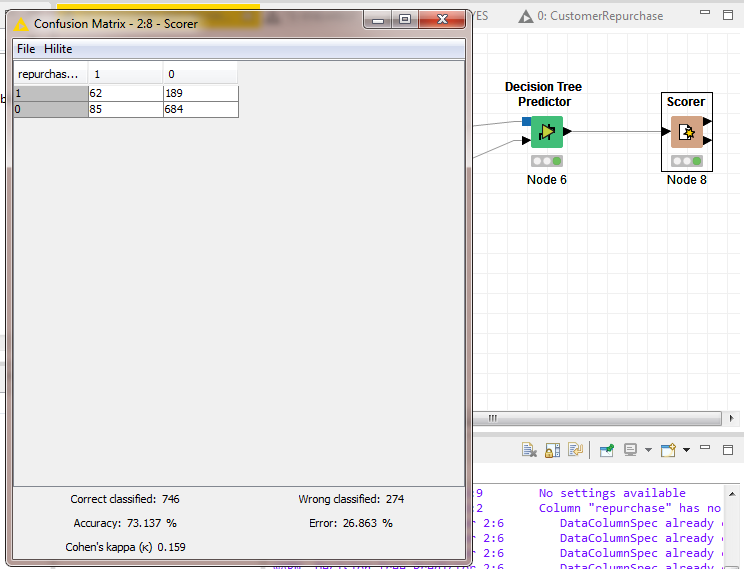
**DECISION TREE VIEW:**



**CLASSIFIED DATA:**



**MODEL ACCURACY:**



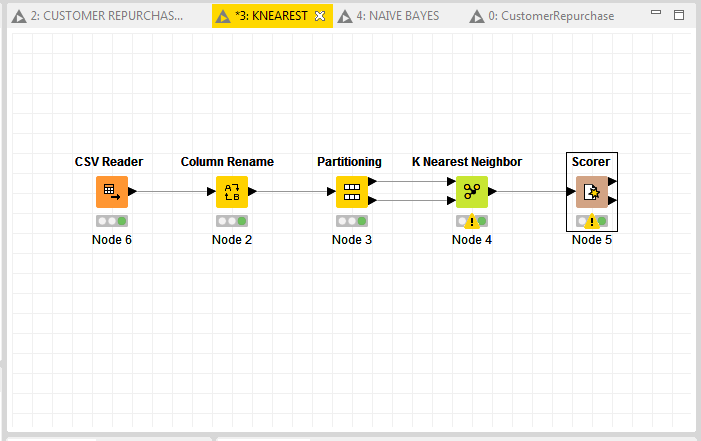
**SIMILARITY BASED LEARNING:**

Similarity based learning is an area of supervised machine learning in artificial intelligence. It is closely related to regression and classification, but the goal is to learn from examples a similarity function that measures how similar or related two objects are.

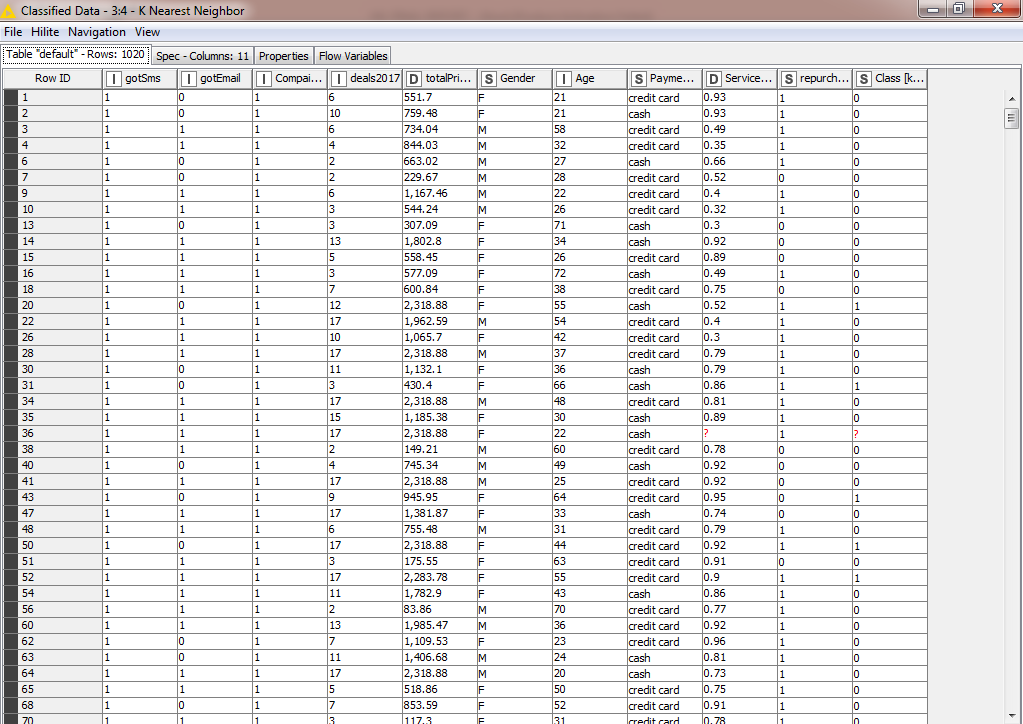
**K NEAREST NEIGHBOR:**

The k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression.

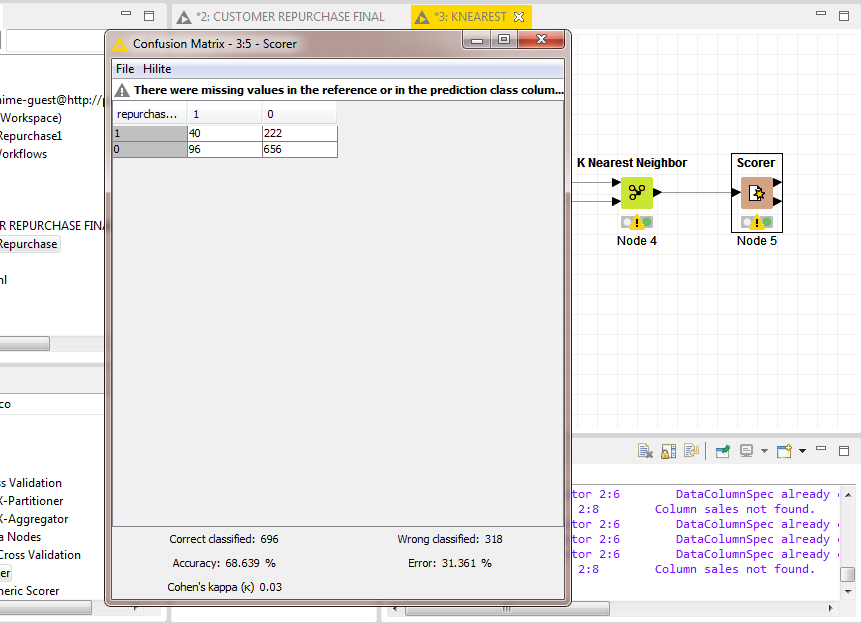
**MODEL:**



**CLASSIFIED DATA:**



**MODEL ACCURACY:**



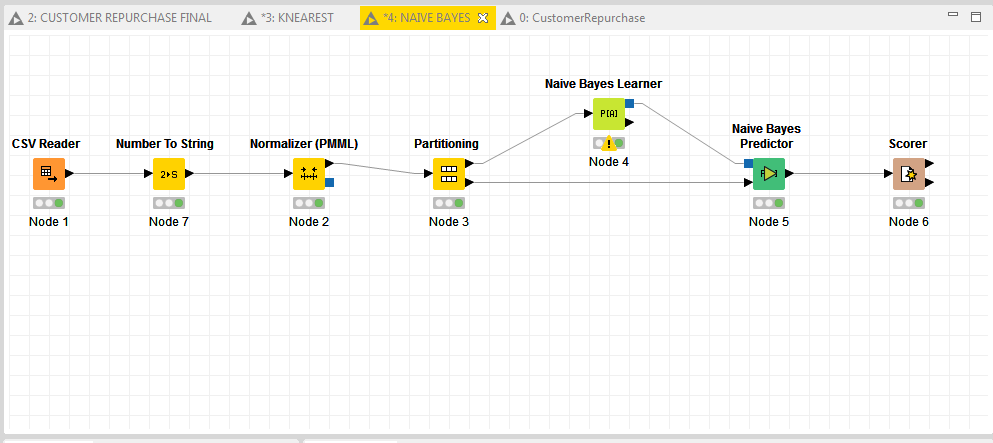
**PROBABILITY BASED LEARNING:**

The objective of a probability-based M.L. model is no different from that of any other kind; it attempts to predict some variable in a data set as a function of other variables.

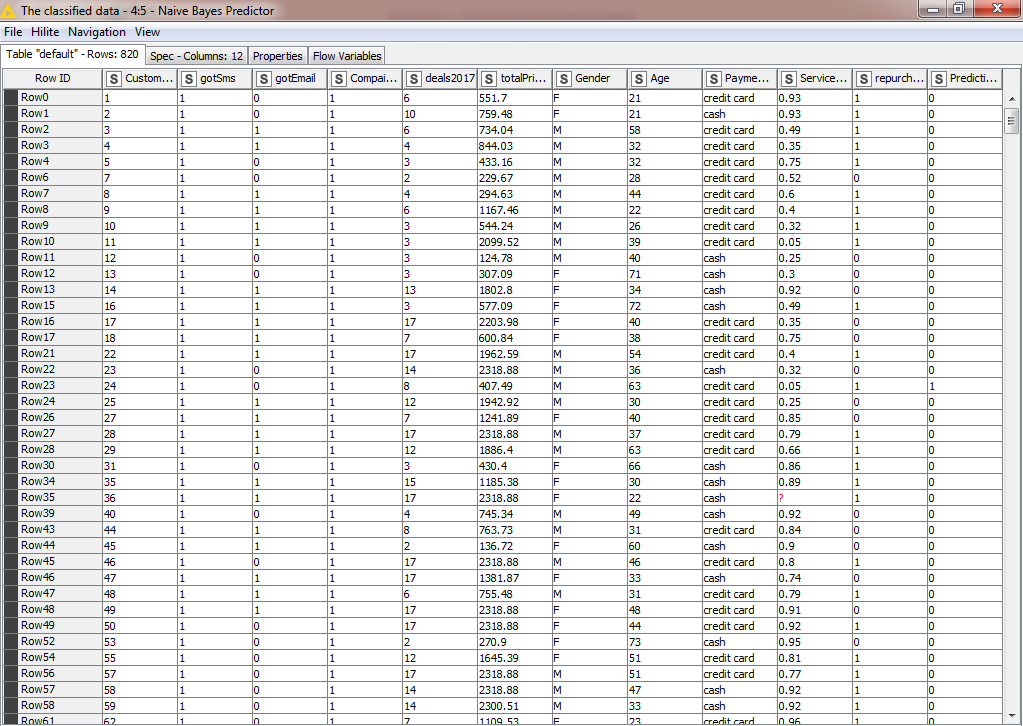
**NAIVE BAYES:**

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong naive independence assumptions between the features.

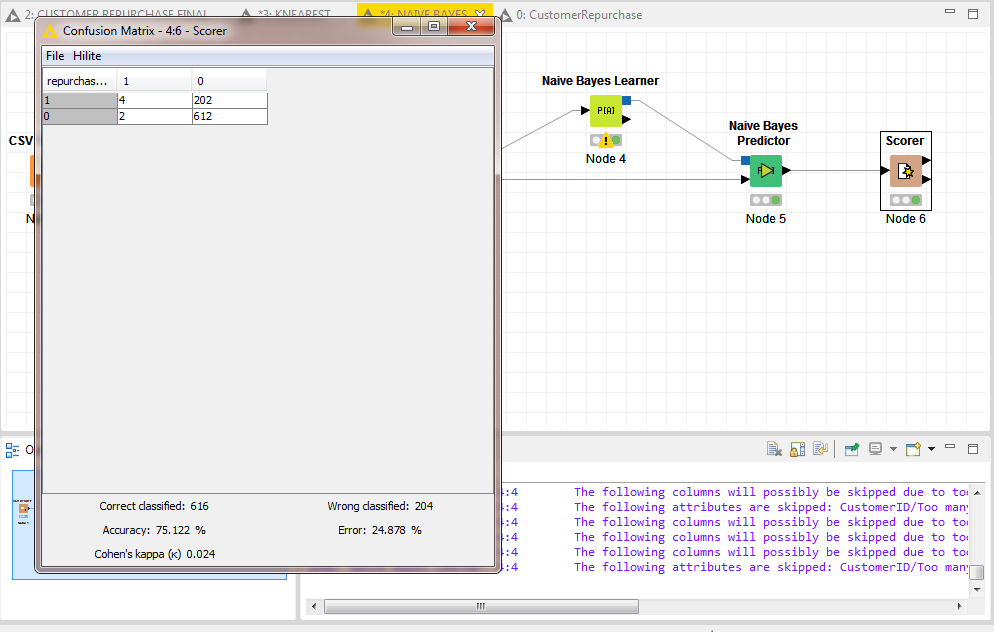
**MODEL:**



**CLASSIFIED DATA:**



**MODEL ACCURACY:**



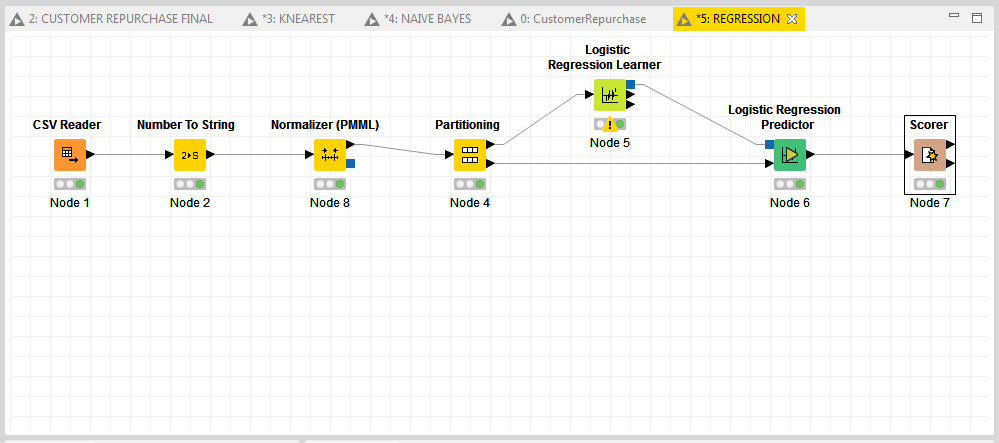
**ERROR BASED LEARNING:**

Error-driven learning. Error-driven learning is a sub-area of machine learning concerned with how an agent ought to take actions in an environment so as to minimize some error feedback. It is a type of reinforcement learning.

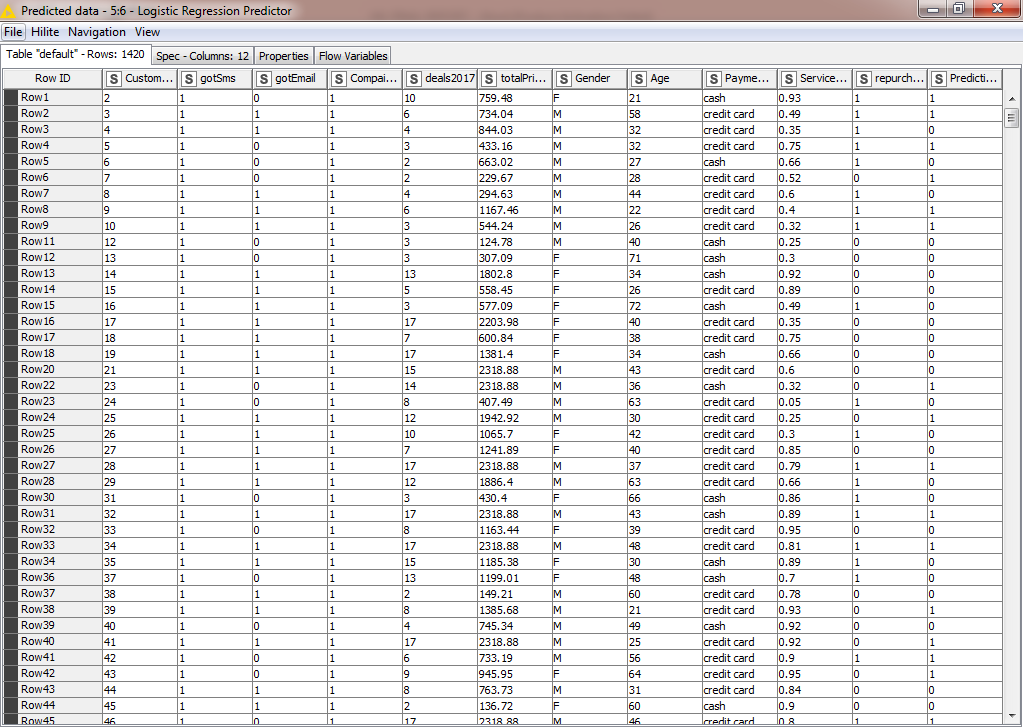
**LOGISTIC REGRESSION:**

Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables.

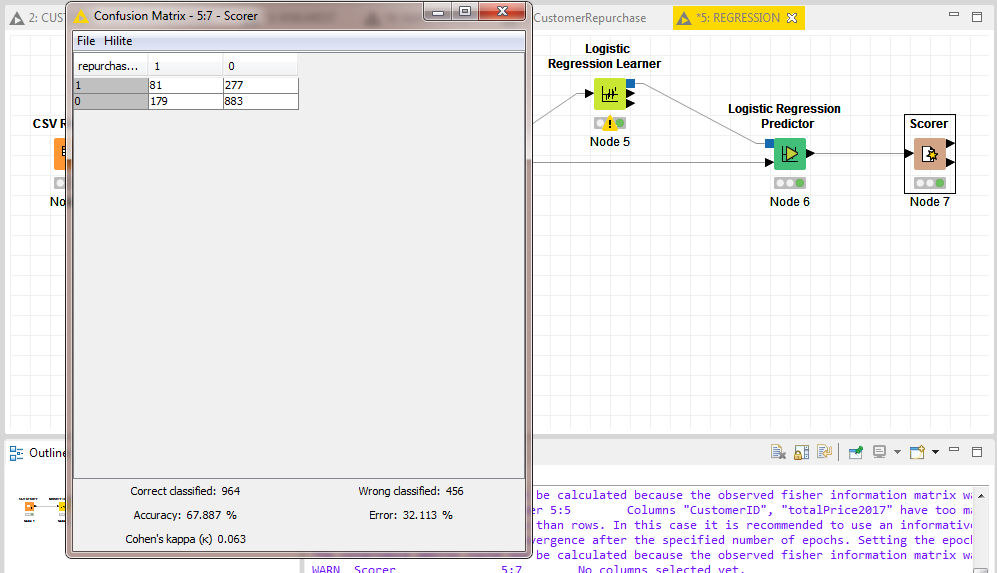
**MODEL:**



**PREDICTIVE DATA:**



**MODEL ACCURACY:**



**COMPARE THE RESULT OF EACH MACHINE LEARNING TECHNIQUE:**

After applying four different algorithms on the model, and the result is collected as,

First we apply Information Based Learning Algorithm i.e. Iterative Dichotomiser 3 (ID3) Algorithm on our dataset & after training the model, the model predicts the 73.137% accuracy. After that we apply Similarity Based Learning Algorithm i.e. K Nearest Neighbor Algorithm on our Customer purchase dataset & after training the model, the model predicts the 68.639% accuracy. Then we apply Probability Based Learning Algorithm i.e. Naive Bayes Algorithm on our customer purchase dataset & after training the model, the model predicts 75.122% accuracy which is the highest accuracy I achieve on this customer purchase dataset. At the end we apply our last Error Based Learning Algorithm i.e. Logistic Regression Algorithm on our same dataset & after train the model predict 67.887% accuracy which is so far the lowest accuracy I found on my customer purchase Dataset.

**WHICH ML TECHNIQUE IS PERFORMING BEST ON MY DATA AND WHY?**

So Far Naive Bayes Algorithm is performing best on my customer purchase data with the model accuracy of 75.122%.

Naive Bayes is easy and fast to predict class of test data set. It also performs well in multi class prediction It perform well in case of categorical input variables compared to numerical variables. For numerical variable, normal distribution is assumed bell curve, which is a strong assumption.