**International Journal of Computing and Applications** | **ISSN: 0973-5704**
Volume 13, Number 1, (January-June 2018), pp 131-137
© Serials Publications, New Delhi (India)

# PREDICTION OF AUTOMOTIVE CUSTOMER CHOICE THROUGH CLASSIFICATION TECHNIQUES

**K. Shyamala[1] and C.S. Padmasini[2]**

*[1]Associate Professor, Dr. Ambedkar Government Arts College, Chennai*
*[2]Research Scholar, Dr. Ambedkar Government Arts College, Chennai*

***Abstract:*** This paper deals with applying data mining classification techniques in Customer Relationship Management. CRM is the leading approach for retaining customers and creating new customers. This technique applies to identify the potential Customer and their preferences and choices of buying a new car from a range of models and varieties available in the market. Data mining techniques can be powerfully applied with inter-disciplinary research which brings out patterns and knowledge from a huge volume of data that helps the organization in efficient decision making. The research starts from identifying the customers who have the interest to buy new cars and this can be done by a scientific research of large data, which is collected from various sources. Based on the data collected, Classification Algorithm which has the highest accuracy compared with other algorithms is used for predicting customers and their choices. Hence this technique is a successful approach in customer identification. By using classification and prediction concept, car manufacturers' can predict a customer who steps in with their car dreams and the exact model the customer is going to choose from a variety of cars.

***Keywords:*** Classification, CRM, Data mining, Prediction.

## 1. INTRODUCTION

The ideology of choosing Automotive Industry is due to consistent increase in car sales in India year after year given by SIAM reports [17] (Society of Indian Automobile manufacturers) and the changing attributes and factors those are becoming the deciding factors for customers to buy a particular brand and model. Application of data mining techniques on these attributes provides strategic methods to understand the changing mentality of car customers in India. Adopting scientific methods to predict the mindsets of these vulnerable group called "Customers" and classify them has becomes necessary for car manufacturing industries. A fundamental distinction among data mining techniques is between supervised methods and unsupervised methods. "Supervised learning" algorithms are those used in classification and prediction. We must have data available in which the value of the outcome of interest (e.g. purchase or no purchase) is known. These "training data" are the data from which the classification or prediction algorithm "learns," or is "trained," about the relationship between predictor variables and the outcome variable. Once the algorithm was found with best results using the training data, it is then applied to another sample of data (the "validation data") where the outcome is known, to see how well it does in comparison to other models[9]

Rapid Miner is the world-wide leading open-source data mining solution due to the combination of its leading-edge technologies and its functional range. [8] It is a software platform developed by the company of the same name that provides an integrated environment for machine learning, datamining, text mining, predictive analytics and business analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including data preparation, results visualization, validation and optimization[10]. Java programming language is used for code. It provides a GUI to design and execute analytical workflows. It represents a completely new approach to an application design. The advantages of rapid miner are it is freely available as open source software, but it is provided under a commercial license suitable for closed-source commercial applications development. The cost of development with Rapid miner is relatively low when compared with other data mining solutions, as it is an open source software, it is especially suitable for research purposes, intuitive, well-arranged and user friendly graphical user interface, access to many databases and possibility to read many file formats, modular and well-arranged code, large number of built-in operators for data mining containing more than 250 learning algorithms [11]. This software is used for the training and test data which is collected from the cars lovers.

## 2. RELATED WORKS

An organization has to change and compete at par with the new strategies and techniques to retain its customers and to create new customers. Organizations are shifting as per the market trends and requirements. Transforming the business focus of an organization means “shifting the organization into a customer based paradigm [7].

Trilok chan sharma[1] The Author in his paper gave an idea about different types of classifications and a comparison between the same. Different algorithms were tested and measuring the accuracy of test data.

EWT Ngai [5] This paper presents a comprehensive review of literature related to application of data mining techniques in CRM published in academic journals. This paper deals about the various classification of CRM management.

Sukhmeet Kumar[2] This paper gave a new dimension for applying Naïve Bayes theorem to find monthly, yearly sales of car manufacturing industry. The prediction results are compared with the actual and real world values in order to validate the results obtained using naïve bayes algorithm.

Based on these related work I have analyzed my data with classification algorithms and found the future cars what the customer will buy.

## 3. PROPOSED WORK

Classification is supervised learning and also known as class prediction or discriminate analysis. Generally, classification is a process of learning-from-examples. Given a set of pre classified examples, the classifier learns to assign an unseen test case to one of the classes.

Customer relationship management in Automobile industry is very popular in finding the potential customers to know market trends. Customer details are useful in finding the future new data. For example customer information (12 inputs) such as customer name, age, occupation, salary, income, gender, family types are collected using questionnaire and information orally. These inputs are recorded and consolidated. More than 600 records were collected from various customers.

**Table 1**
**Sample Attributes for Data Set**

| *S.No.* | *Attributes* | *Description* |
|---|---|---|
| 1 | Business/salaried | Categorical value |
| 2 | Gender | Categorical value |
| 3 | Urban/rural area | Categorical value |
| 4 | Fuel type | Categorical value |
| 5 | Family type | Categorical value |
| 6 | Income | Numerical value |
| 7 | Self or driver driven | Categorical value |
| 8 | Choice of the car | Budget/high end/mid range |

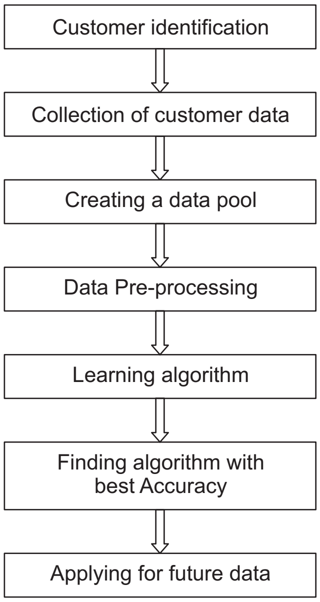


**Figure 1: Proposed framework**

Proposed methodology has following steps:

1. Identifying the right customer
2. Collecting the relevant data for the customer
3. Creating a training set of data
4. Applying pre processing techniques
5. Applying classification algorithm
6. Finding Algorithm with best accuracy
7. Applying for test data (prediction).

Data available for mining is raw data. Data may be in different formats as it comes from different sources, it may consist of noisy data, irrelevant attributes, missing data etc. Data needs to be pre processed before applying any kind of data mining algorithm which is done using following steps [14]:

**Data Integration:** If the data to be mined comes from several different sources data needs to be integrated which involves removing inconsistencies in names of attributes or attribute value names between data sets of different sources.

**Data Cleaning:** This step may involve detecting and correcting errors in the data, filling in missing values, etc. [15].

**Discretization:** When the data mining algorithm cannot cope with continuous attributes, discretization needs to be applied. This step consists of transforming a continuous attribute into a categorical attribute, taking only a few discrete values. Discretization often improves the comprehensibility of the discovered knowledge [13].

**Attribute Selection:** Not all attributes are relevant so for selecting a subset of attributes relevant for mining, among all original attributes, attribute selection is required.

In Classification, training examples are used to learn a model that can classify the data samples into known classes. The Classification process involves following steps:

(a) Create training data set.

(b) Identify class attribute and classes.

(c) Identify useful attributes for classification (Relevance analysis).

(d) Learn a model using training examples in Training set.

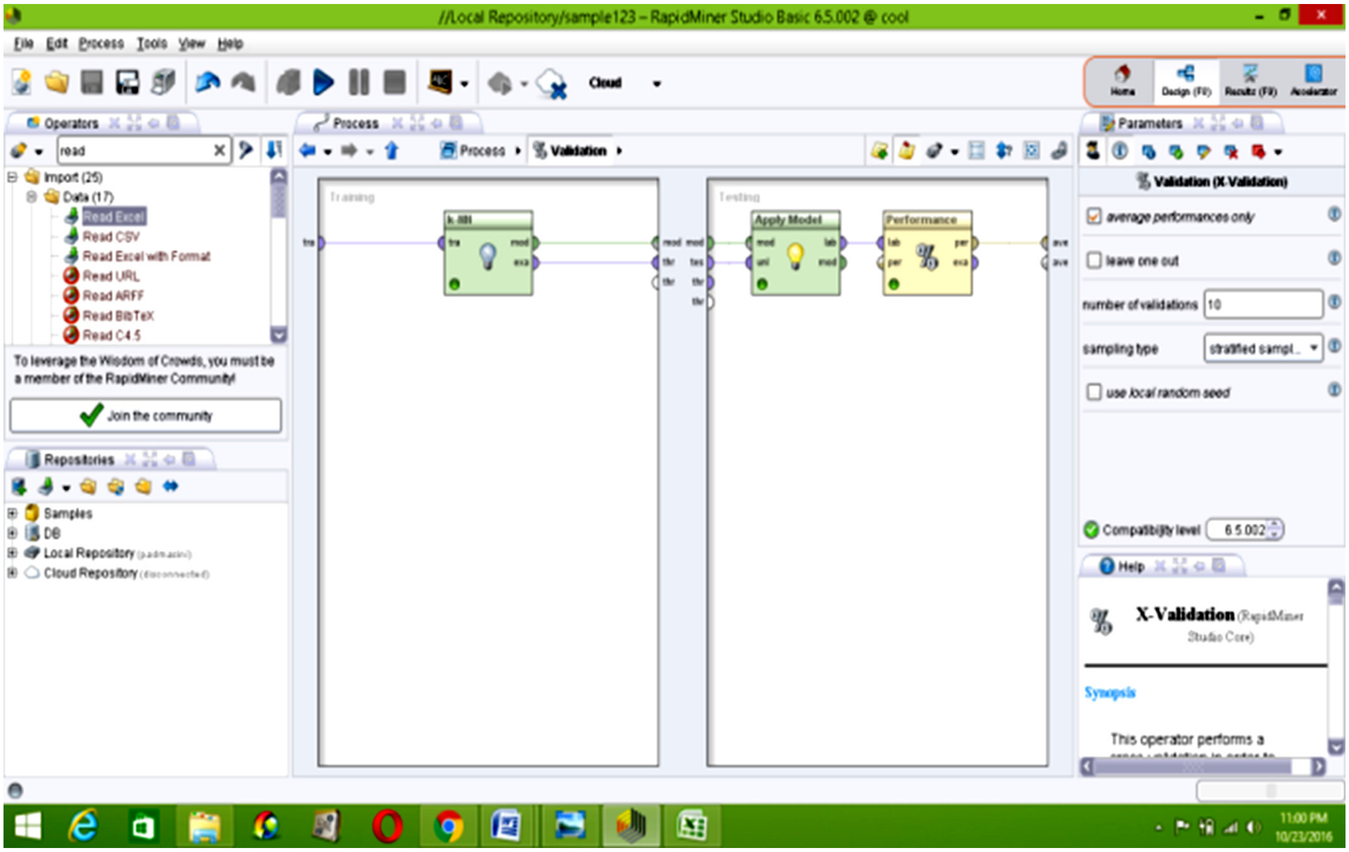
(e) Use the model to classify the unknown data samples.[1]

There are different types of classification such as ID3 algorithm, Bayesian classification, Neural networks, Support vector Machines, Classification based on Associations [3].I have taken ID3 algorithm for my test data because I got accuracy as 100%.Based on the validation, a new value or called as class label will be predicted.

Nearly 600 inputs were collected from various sources such as employees working in different offices and business class people. The data were collected from various sources for varied inputs. Nearly 12 attributes were collected from customers and the validation is done based on different classification algorithm. ID3, Decision tree, KNN, Naïve Bayesian algorithms were tried and performance evaluation is calculated for the training data. Test data is taken and Class label values are predicted based on the algorithm which has highest accuracy. There are various procedures to follow to know the customer who were interested to buy four wheeler vehicles.

Questionnaires were prepared and got the inputs from interested customers. Customers belonging to different age, different area, different income, and different gender were taken into consideration. This input is a base data that any companies or manufactures, retailers or dealers can use this data and predict the customer choice.

The customer preference of either budget cars, mid range cars or high end cars is predicted using different algorithms. Any company such as Ashok Leyland, TATA motors, Volkswagen, Toyota, Fiat, General motors, Honda cars, Mahindra and Mahindra, Maurti Suzuki can analyze the customer behavior and customer choice and decide the production for the profit of the next year, based on available raw data.


//Local Repository/sample123 – RapidMiner Studio Basic 6.5.002 @ cool
File Edit Process Tools View Help
Cloud
Home Design (F8) Results (F9) Accelerator
Operators
read
Import (25)
Data (17)
Read Excel
Read CSV
Read Excel with Format
Read URL
Read ARFF
Read BibTeX
Read C4.5
Join the community
Repositories
Samples
DB
Local Repository (padmasini)
Cloud Repository (disconnected)
Process
Validation
Training
k-NN
Testing
Apply Model
Performance
Parameters
Validation (X-Validation)
Help

To leverage the Wisdom of Crowds, you must be a member of the RapidMiner Community!

☑ average performances only

□ leave one out

number of validations 10

sampling type stratified sampl...

□ use local random seed

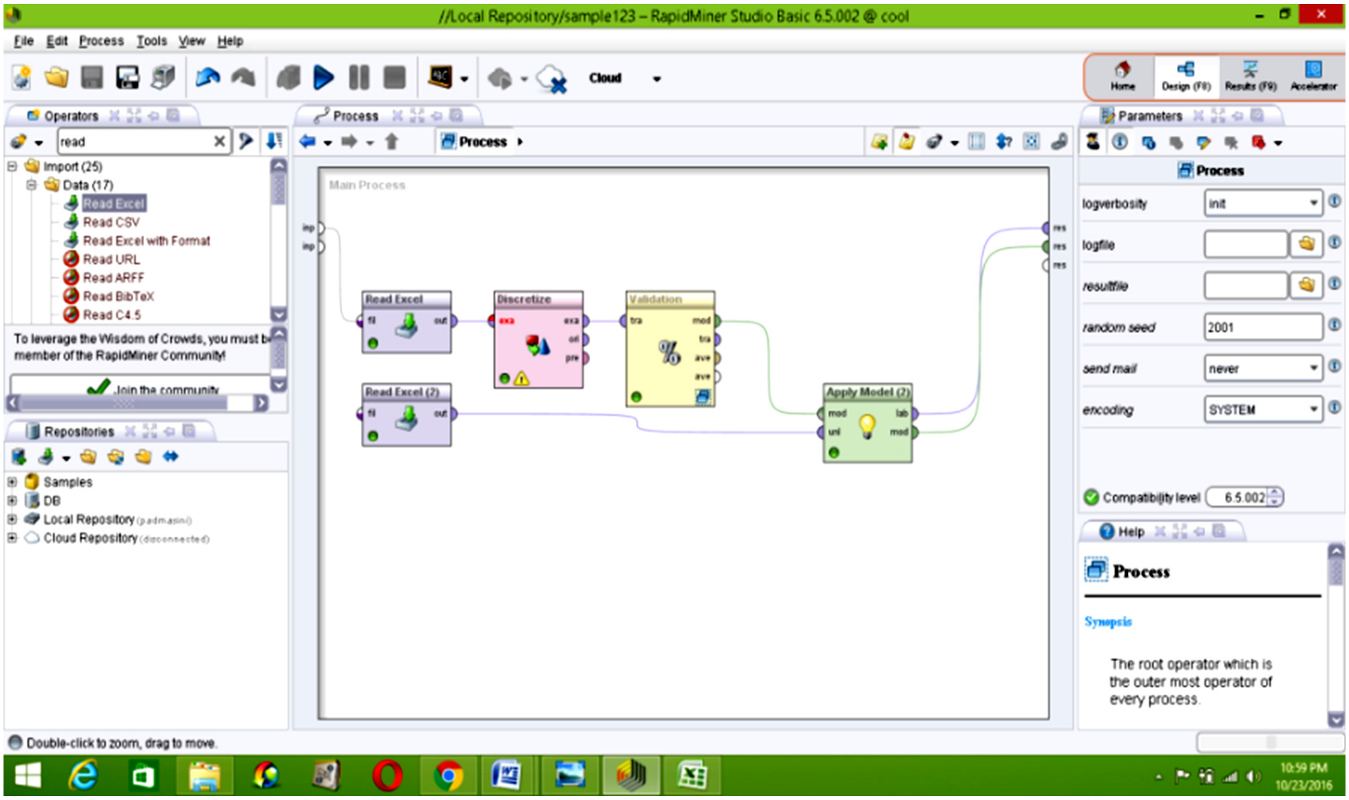
Compatibility level 6.5.002

X-Validation (RapidMiner Studio Core)

Synopsis

This operator performs a

Figure 2: Apply model for performance evaluation



To leverage the Wisdom of Crowds, you must be member of the RapidMiner Community!

logverbosity init

logfile

resultfile

random seed 2001

send mail never

encoding SYSTEM

Compatibility level 6.5.002

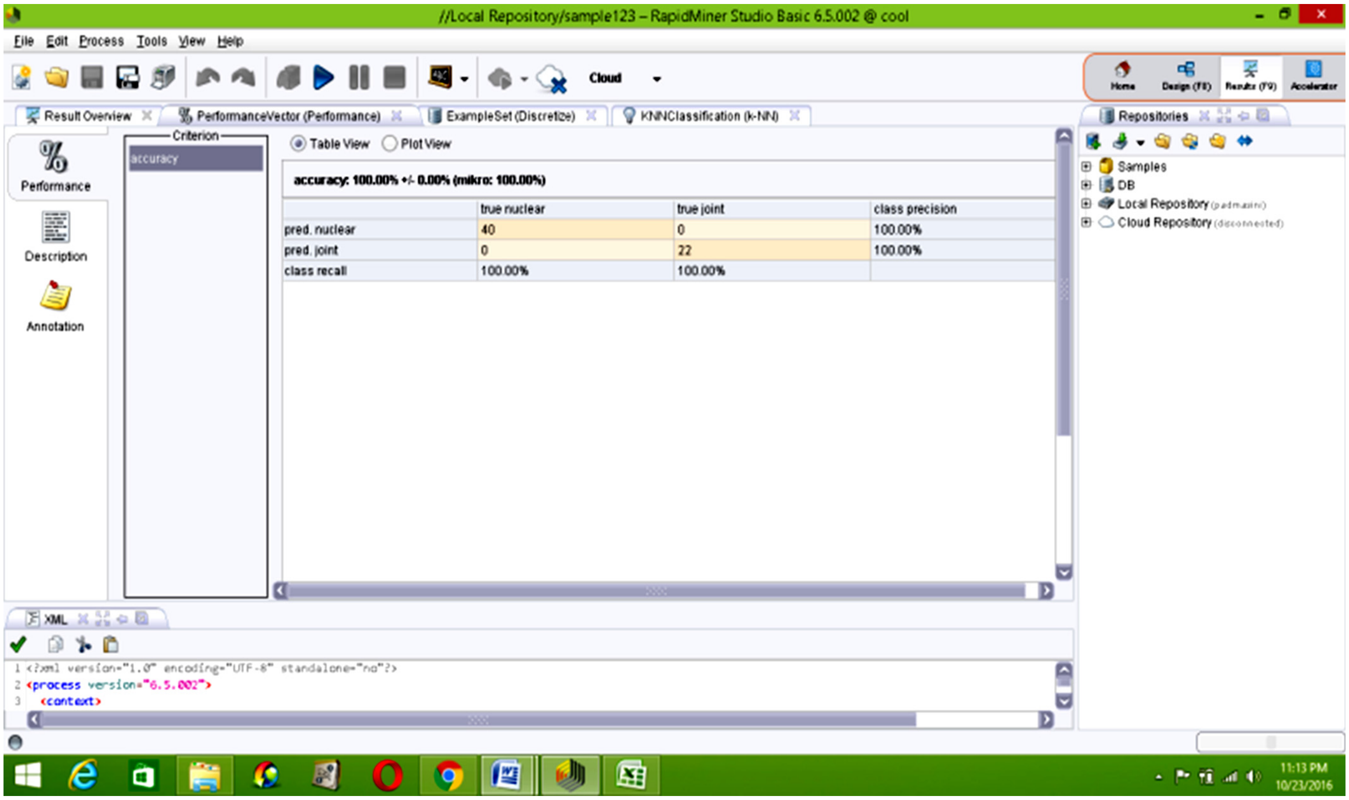
Process

Synopsis

The root operator which is the outer most operator of every process.

Double-click to zoom, drag to move.

Figure 3: Validation of Model



accuracy: 100.00% +/- 0.00% (mikro: 100.00%)

| | true nuclear | true joint | class precision |
|---|---|---|---|
| pred. nuclear | 40 | 0 | 100.00% |
| pred. joint | 0 | 22 | 100.00% |
| class recall | 100.00% | 100.00% | |

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="6.5.002">
  <context>
```

Figure 4: Performance results

## 4. RESULTS AND DISCUSSIONS

### 4.1. Performance Evaluation

There are various parameters on the basis of which we can evaluate the performance of the classifiers such as TP rate, FP rate, Precision, Recall F-Measure and ROC area [18].

The **Accuracy** of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

**Accuracy** is the proportion of the total number of predictions that were correct. It is determined using:

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp}$$

where, TP rate = positives correctly classified/total positives, FP rate = negatives incorrectly classified/ total negatives.

**Precision** is the proportion of the predicted positive cases that were correct, as calculated.

$$\text{Precision} = \frac{tp}{tp + fp}$$

**Recall or Sensitivity or True Positive Rate (TPR):** It is the proportion of positive cases that were correctly identified, as calculated.

$$\text{Recall} = \frac{tp}{tp + fn}$$

Table 2
Comparison of various classification algorithm

| *Algorithm* | *TP Rate* | *FP Rate* | *Precision* | *Recall* | *Accuracy* |
|---|---|---|---|---|---|
| KNN algorithm | 0.986 | 0.026 | 0.987 | 0.986 | 98% |
| Decision tree | 0.945 | 0.086 | 0.946 | 0.949 | 94% |
| Naive Bayesian algorithm | 0.957 | 0.041 | 0.957 | 0.958 | 95% |
| ID3 Algorithm | 1 | 0 | 1 | 1 | 100% |

### 4.2. Classification Algorithm with Best Accuracy

The ID3 algorithm is used to build a decision tree, given a set of non-categorical attributes C1, C2, ..., Cn, the categorical attribute C, and a training set T of records.

Function ID3 (R: a set of non-categorical attributes,

```
Begin
If S is empty, return a single node with value Failure;
If S consists of records all with the same value for the categorical attribute, return a single node with that value;
If R is empty, then return a single node with as value the most frequent of the values of the categorical attribute that are found in records of S;
Let D be the attribute with largest Gain (D, S) among attributes in R;
Let $\{dj \mid j = 1, 2, \ldots, m\}$ be the values of attribute D;
Let $\{Sj \mid j = 1, 2, \ldots, m\}$ be the subsets of S consisting respectively of records with value dj for attribute D;
Return a tree with root labeled D and arcs labeled d1, d2, .., dm going respectively to the trees
ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), ..., ID3(R-{D}, C, Sm);
End;
```

Entropy is the measure of the amount of uncertainty in the data set S.

$$\text{Entropy (S)} = p(\text{I})\log 2\, p(\text{I})$$

Gain (S, A) is information gain of example set S on attribute A is defined as

Gain (S, A)

$$= \text{Entropy(S)} - \left(\frac{|Sy|}{|S|} \times \text{Entropy}(Sy)\right)$$

$S_v$ = subset of S for which attribute A has value $v$

$|S_v|$ = number of elements in $S_v$

$|S|$ = number of elements in S.

Using questionnaire nearly 600 customers were approached and accurate data was consolidated from 300 customers. The training set is considered and performance evaluation is calculated based on different classification algorithm. The highest accuracy was ID3 algorithm. Based on this algorithm with the training data model has been learnt. Test data was collected with various inputs from the customer and predicting the new class values with the learnt algorithm.

Training data collects rules to find the results for test data. For example to decide that customer will purchase Budget car, The customer will be any gender, business people target this type of Budget car. Mid range will be mostly purchased by salaried class, who uses petrol cars and it is used by urban and rural people. High end cars will be purchased by High salaried class and business class people. With these results the test data results are predicted.

Based on the learning model test data values were predicted whether the customer has an option of budget cars, mid range cars, high end cars. These results will be useful the company manufactures dealers and retailers who can judge and come to conclusion to know their profit and it will be useful to find the right customers and the right product.

## 5. CONCLUSION

Customer Relationship Management is a powerful tool and customer is very important for any organization to increase their sales and to get lion share of business in the market. Car industry is a large automotive domain where Indian customers have a lot of interest in buying new cars of their choices. The Financial standard of people has been elevated and it is a prestigious matter for Indian customer to possess branded cars. Global car manufactures are keenly interested in Indian Car Market and likes to invest more in India by understanding customer tastes. In this paper, customer data has been collected and data mining tool Rapid miner was applied. Various Classification algorithms are tested for accuracy. For this analysis, the data set ID3 algorithm gave 100% accuracy and this model was built using training set and it is the model for future test data. This will be useful for the car manufacturers to decide on the changing taste of Indian car customers and predicts their future car production.

## *References*

[1] Trilok Chand Sharma, Manoj Jain, "Weka approach for comparative study of classification algorithms", International Journal of Advanced research in computer and communication Engineering, Volume 2, Issue 4, April 2013.

[2] Sukhmeet Kumar, Kran jyoti, "Predicting the future of car Manufacturing industry using Naïve Bayese Classifier", International Journal for Science and Emerging Technologies with Latest Trends, 2012.

[3] Nazneentarannum S.H. Rizvi, "A systematic overview on Data mining concepts and Techniques",

International Journal of Research in computer & Information Technology (IJRCIT) Volume 1 Special issue 1, 2016.

[4] Divya Tomar and Sonali Aggarwal "A survey on Data mining Approaches for Health care", International Journal of Bio science and Bio Technology.

[5] E W T Ngai, LiXiu, "Application of data mining techniques in customer relationship management: A literature review and classification", Expert Systems with Application, Elservier, 2009.

[6] Wang Ke Nan et. al., (2008), "Apply of data mining techniques in CRM" published in the changhun Taxatin college Research project and IEEE.

[7] Archana Singh (2013), "Mining customer data in an automobile industry using clustering Techniques", International Journal of emerging Technologies in computational and Applied sciences.

[8] https://web.fe.up.pt/~ec/files_1011/week%2006%20-%20lab.pdf

[9] https://mineracaodedados.files.wordpress.com/2012/07/data-mining-in-excel.pdf

[10] https://en.wikipedia.org/wiki/RapidMiner

[11] https://www.researchgate.net/profile/Jan_Karasek/publication/260727350_Rapidminer_image_processing_extension_A_platform_for_collaborative_research/links/004635321843b48bfb000000.pdf

[12] B. Pfahringer, "Supervised and unsupervised discretization of continuous features", *Proc. 12th Int. Conf. Machine Learning*, 1995, pp. 456-463.

[13] J. Catlett, "On changing continuous attributes into ordered discrete attributes", *In Y. Kodratoff (ed), Machine Learning—EWSL-91*, Springer-Verlag, New York,1991, pp 164-178.

[14] D. Pyle, *Data preparation for data mining*, 1st Vol., Morgan Kaufmann publisher, San Francisco, 1999.

[15] I. Guyon, N. Matic and V. Vapnik, "Discovering informative patterns and data cleaning", In*: Fayyad UM, Piatetsky-Shapiro G, Smyth P and Uthurusamy R. (ed) Advances in knowledge discovery and data mining*, AAAI/MIT Press, California, 1996, pp. 181- 203.

[16] Classification and feature selection techniques in data mining, Sunita Behroal, *itender Arora* International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 6, August – 2012, ISSN: 2278-0181.

[17] https://www.siamindia.com

[18] A Novel adaptive feature selector for supervised classification, S. Sasikala, S. Appavu alias Balamurugan, S. Geetha. Information processing Letters, August 2016.