International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

A Prediction to Choose Customers in Auto Ancillary Automotive Products using K-Tree-Bayes Model for Improving Business Profits and Retention

K. Shyamala, C.S Padmasini

Abstract: ***Customers are of paramount importance for running business enterprises. A K-Tree-Bayes model, when applied for the purpose of customer retention and business promotion, it retains them and their favorite choices. This model is extended to work in various aspects as and when new customers as well as existing customers provide their wishes and those data will become impertinent to improve the product in all aspects right from the manufacturing to reach to the customers. The model shows reasonable accuracy to predict the changing customer choices towards their desire to buy any automotive as companies are investing heavily on customer prediction thereby providing automotive of their choices to retain them forever.***

Keywords: ***Clustering, Pruning, Classification, Accuracy, Naïve Bayes***

I. INTRODUCTION

Automotive industry is booming World-wide. Automobile vehicle industry in general is very competitive in nature. Attracting customers and bringing them to the show room and retaining their loyalty is a challenge for any automobile company. Customer data set is collected from Manufacturers, Distributors and dealers who play a very important role for customer satisfaction aspects of automotive industry.

Customers are of two types: Direct and Indirect customers. Direct customers are consumers of the final product. Indirect customers are the entities who promote the business in secondary sales like dealers, distributors and retailers .Original Equipment Manufacturers (OEM's) play a very important role as they manufacture vehicles of competitive nature and generate revenues for all their channels. Department such as design (R&D), Engineering, Manufacturing , Marketing and sales team are to be sufficiently nurtured to improve the original equipment's products and secondary sales also. The secondary customers are Distributors, Dealers, Retailers and direct customers.

Manuscript published on November 30, 2019.
* Correspondence Author
Dr K. Shyamala*, Associate Professor PG & Research Department of Computer Science Dr.Ambedkar Government Arts College (Autonomous).
C.S Padmasini, Assistant Professor, Department of Computer Science,M.O.P Vaishnav College for Women(Autonomous)

Both the customers are important to increase business profits. Depending upon the various categories listed below under quality, cost, and delivery (QCD) and logistics. These are the set of determinants to edge over the competition. Automobile customers either original equipment manufacturers or Secondary sales customer both decide the profit of the company. The business with direct customers and indirect customers plays various aspects and roles for making business decisions. Any customer is oriented towards quality, cost and delivery of the product. They don't like to comprise on these aspects.

II. OBJECTIVES

Customers are very important for business. To make them get products and retain them is a challenge**.**

The customer orientation has been classified as follows.

- To identify the choicest Customer group.
- Attract them through various participation in ACMA, SIAM, CII (for existing customers) and public media (for new customers) and also accreditations like TPM, TQM etc.,
- Provide them the products and services as desired by them
- Obtain their suggestions and desires through an appropriate CSI.
- Improve Morale and Loyalty and enhance customer satisfaction through trust worthy practices for longer customer retention.
- Repeat this process for continuous improvements.

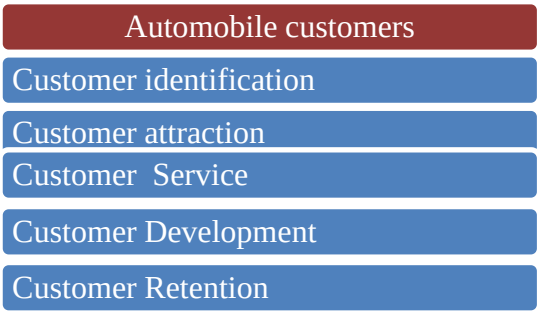

Fig.1 Automobile customer service.

The data set contains major categories such as Quality, Cost, Delivery, Logistics and new product development (R&D).For the satisfaction of the customer, Quality is the key.

Customer will never comprise on quality as the end user will rely on retailers who rely on distributors, dealers and manufactures etc. Hence, The first aspect of customer satisfaction is Quality. Addressing the Quality complaints and the Warranty, PPM levels and exhibiting the system transparently is important. The Second aspect is Cost. The categories in Cost will be value for money and also the payment terms. The third aspect is Delivery. The categories are Reliability, Coping with demand and Dispatch activities. The fourth aspect is with Logistics. The categories are Delivery cost and time, Packaging, Inventory level, overall design and achieving deadline All the above categories play a very important role in the automotive industry right from raw materials to end products including various processes and value added activities to enhance customer satisfaction. Customer feedbacks on the above are also equally important and bound for evaluation.

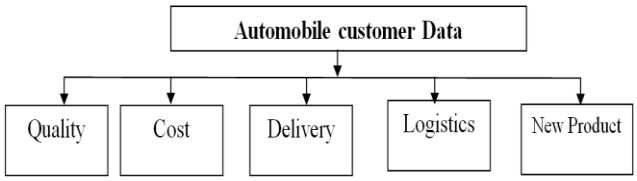

Fig.2. Attributes for Customer requirement

III. RELATED WORK

Ahmed Mohammed Ahmed [6] proposed the concept of pruning and Bayes classification concepts. A detailed explanation of dataset and feature selection using pruning is done. Diagrammatic representation of pre pruning and post pruning concepts are clearly explained.

Yamuna N R [7] proposed difference between the decision tree algorithms such as C4.5 in pruned and un-pruned data. The comparison between sensitivity and specificity of pruned and un pruned data set using C4.5 is implemented. The values had a considerable change.

Kavitha Mittal [8] proposed discovering item set using Association rule mining and reducing the tree using pruning techniques. This paper gave a detailed description about Association rule mining. Prasanna jyothi[9] gave a comparative study on different classification algorithms such as K Nearest neighbor, Naïve Bayes,ID3 etc for the existing Medical data set. Naive Bayes algorithm gives a mathematical solution for classification concept.

A Shameen Fathima[10] proposes about various classification methods such as SVM, Decision trees, Neural network, Rough set etc. The Author has selected the disease based data set. Chikingunia disease is taken for analysis.

Oyelade, O. J [11] proposes system for analyzing students 'results based on cluster analysis and uses standard statistical algorithms to arrange their scores data according to the level of their performance is described.

Huda [12] proposes K means clustering for IRIS data set, Leukemia disease data set. He gave a detailed diagrammatic representation of stages and steps that was carried out in K means clustering. This paper listed out Advantages and disadvantages of clustering in his data set. He concluded that K means can be applied in data mining and pattern recognition.

Keyvan Vahidy Rodpysh [13] proposes applying Data mining in customer relationship management. This paper gave a overall customer relationship management such as customer identification, customer attraction, customer retention and customer development.

Maneesh Singhal [14] proposes the implementation of data mining concepts using WEKA tool. Naïve Bayes classifiers are used as supervised learning algorithm. Data set has co orelated features. The tool gave solution of co related and non- co related data using Naïve Bayes Classifiers. Accuracy, Kappa static, Time , Mean absolute error are the measures to judge the classifiers.

From the related papers the concepts of clustering, pruning, and classification is implemented with various existing and real time data set. The proposed work is developed on these concepts such as Kmeans clustering as class label is not available. Using Decision tree pruning concept is implemented. To develop supervised learning, classification and prediction is implemented. Various strategies were found such as accuracy, Time taken, Kappa statistic, Mean Absolute error, Relative absolute error are compared with existing bench algorithms.

IV. 4. K TREE BAYES MODEL

KTree Bayes model deals with various aspects in data mining to obtain customer retention, customer satisfaction and new customer attraction.This model is based on clustering classification. The data set is extended with new customers.This proposed model supports growth business trends and a solution for customer retention. The raw data is collected from customers such as Manufacturers, distributors and dealers of the Automobile industry. The data set has 23 attributes which determines the growth of the automotive business. This raw data does not have class label. The class label is the first step to be obtained. The label is to find whether the customer is satisfied or not satisfied with products and logistics of the company.

Algorithm 1: K Tree Bayes model
Input: Sample S
Output: Data with Maximum accuracy

1. *Begin*
2. *Compute the cluster using K means partition clustering*
3. *If* $x_i \rightarrow Empty Then\ x_i = HigherRanked(c_i)$
4. *For all Attribute a in S do*
5. *gain <- InformationGain(a,e)*
6. *If gain>maxGain then maxGain<- gain*
7. *split A<-a*
8. *end if; end for*
9. *Do pruning*
10. *Initialize classification*
11. *For each class do classification*
12. *Do prediction*
13. *End for*
14. *End [18]*

4.1 Preprocessing

Data Preprocessing is of much important step in the data processing. There are different categories of preprocessing methods .Data cleaning, Data integration, Data transformation, Data reduction. Data Cleaning is vital as data may be consistent or incomplete data. The incomplete information on vital attributes by customers in the data set will not be considered for evaluation. When few non vital attributes in the data set are missing, then that customer information may be considered. When the data is available at least above 60% of vital information then the data is filled with feasible values. To fill the data the neighbor values can be considered. Bunning method is adopted for filling out missing values from the closest or nearest attributes. After suitable preprocessing techniques applied to the data, the data set becomes a complete one for evaluation.

4.2 Clustering

K means is a traditional clustering method to group similar elements. It is a well-known partitioning method for clustering The customer data set is segregated and grouped into 2 clusters such as satisfied and not satisfied category. This will be considered for scope and improve business plans and growth. K means is a partitioned clustering method. It is a mutually exclusive method where each object belongs to a group[1]. It groups the data based on their closeness to each other according to the Euclidean distance. It takes ky as input parameter and partition a set of n object from ky clusters. The mean value of the object is taken as similarity parameter to form clusters. The cluster mean or center is formed by the random selection of ky object. By comparing most similarity other objects are assigning to the cluster. For each data vector this algorithm calculates the distance between data vector and each cluster centroid [2].K means is calculated using Euclidean distance. Euclidean distance computes the root of square difference between co-ordinates of pair of objects [3] . K means is calculated using Euclidean distance. Euclidean distance computes the root of square difference between co-ordinates of pair of objects [3].

Select ‘c’ cluster centers randomly. Calculate the distance between each data point and cluster centers using Euclidean distance. Data point is assigned to the cluster center whose distance from the cluster center is minimum of all the cluster centers. New cluster center is calculated. The distance between each data point and new obtained cluster centers is recalculated.

If no data point was reassigned then stop, otherwise repeat steps from 3 to 5 [3].The advantages of Kmeans clustering are it is very fast, simple and reliable and efficient. In this clustering specifying the cluster numbers decides the accuracy of data. Based on K means clustering with Euclidean distance metrics the data set is grouped into 2 clusters (satisfied, not satisfied).Based on unsupervised learning the data set is grouped into two class label either satisfied or not satisfied

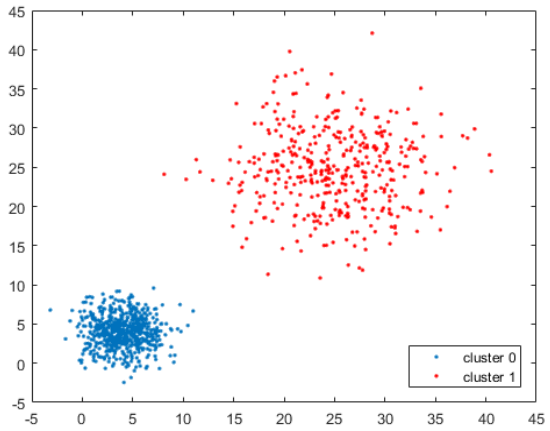

Fig. 3 Cluster group

4.3 Pruning

Pruning is a technique in machine learning that reduces the size of decision trees by removing sections of the tree that provide little power to classify the instances. Pruning reduces the complexity of the final classifier [5].The decision tree generated by the ID3, C4.5 are accurate and efficient, but they often provide very large trees that make them incomprehensible to the experts [4]. When decision tree are formed, many of the tree branches will reflect anomalies in the training data due to outliers and noise. Pruning reduces training set error. Pruning a decision tree is a fundamental step in optimizing the computational efficiency as well as classification accuracy [4]. Pruning reduces the size of tree to avoid unnecessary complexity and ambiguity.

Using J48 algorithm the data set is pruned. The data set shows 1024 records as customers from manufacturers, distributors and dealers. After pruning, the concept of reducing complexity 851 customer data is identified as part of training set. There are two types of pruning, pre pruning and post pruning. Post pruning is also called as Back ward pruning. The first step is to generate decision tree and remove non relevant branches. The advantage of implementing post pruning is to improve classification accuracy as this a good model to follow to predict new set of customers.

There are two main methods of doing pruning techniques. First method is converting the decision tree to set of protocols. Second method is to retain the decision tree and replace into leaf nodes. Pruning reduces the decision tree size to avoid complexity. Pre pruning technique is to avoid branches while building the decision tree. Post pruning technique first builds the tree and reduces the branch which is ambiguous. J48 algorithm shows two pruning methods. First method is known as sub tree replacement and second method is called sub tree raising. Sub tree replacement means nodes in the decision tree may be replaced with leaf and works backwards. In Sub tree raising the nodes are moved upwards in the direction of root. If the customer dataset consists of “n” attributes then to decide which attribute to place at the root is a tedious step?

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

By just choosing randomly any node to be the root cannot give a perfect solution and results in low accuracy. To solve this attribute selection problem, researchers have given different and unique solution. These solutions will calculate values for every attribute. The values are sorted, and attributes are placed in the tree in a sorted order i.e, the attribute with a Maximum value forms the root of the tree. Assuming categorical attributes in information Gain as a criterion, and for gini index, attributes are assumed to be continuous. Choosing information gain as a criterion for pruning[18]. Using J48 algorithm decision tree is calculated and pruning is done to meet the accuracy.

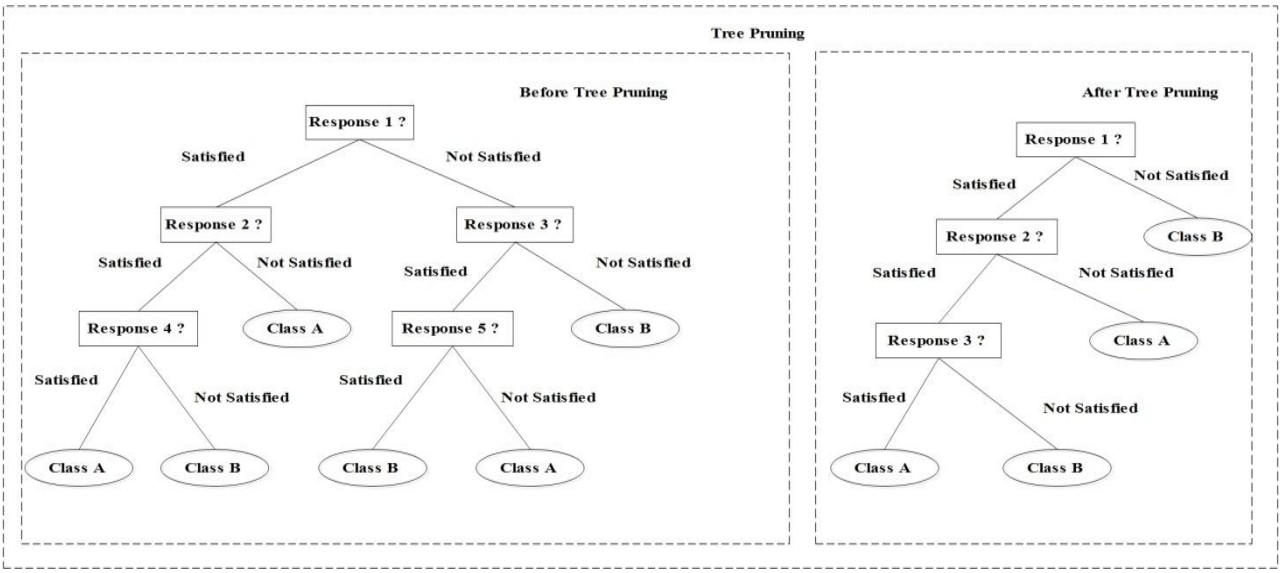

Fig. 4 Before and After Tree pruning

4.4 Classification

Naïve Bayes is a simple technique for constructing classifier models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set [14]. To classify the data and to gauge the data classification techniques are used. Naïve Bayes classification is frequently used classifier. Accuracy of the classifier is calculated to prove the model as best. Response time is also calculated to find which the best in terms of time complexity is. Based on the accuracy percentage models can be selected for classifying the data. Future data can be predicted with the classifier and prediction of data is done

V. RESULTS AND DISCUSSION

The automobile customer data set can be handled with this model to cluster and classify to predict future data on customer orientation. There are various measures to find the best and suitable model. They are Accuracy of the model, Kappa static; Mean absolute error, Root mean Squared Error, Relative Absolute error and time taken for this model for this existing data set.

***5.1 Accuracy*:** The accuracy of a test is its ability to differentiate the data set correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as: [19]

5.2 Kappa statistic: Kappa statistic given in the equation [3] is another measure to find in the statistical aspects of the data set. [20]

$$kappa = \frac{Observed\ Agreement - Expected\ Agreement}{1 - Expected\ Agreement} \quad (2)$$

5.3 Mean Absolute Error: The **Mean Absolute Error** (MAE) in the equation [4] is the average of all absolute errors. The formula [21] is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} xi - x \quad (3)$$

Where: n = the number of errors, Σ = summation symbol |xi – x| = the absolute errors.

$$Accuracy = TP + TN \div TN + TP + FP + FN \ (1)$$

5.4 Root Mean squared error. The R**oot-mean-square error (RMSE)** is a frequently used measure of the differences between values (sample or population values) .Predicted by a model and the values observed. It is defined the square root of the mean square error [22].

Table I: calculation and comparison of Algorithms

Algorithms	Accuracy	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error
Decision Tree	92.5355	89.91	5.1	17.77	13.8907	41.4729
Naïve Bayes	86.3744	81.87	6.79	25.1	18.5004	58.5847
KTreeBayes	99.6098	99.18	0.63	6.58	1.319	13.4458

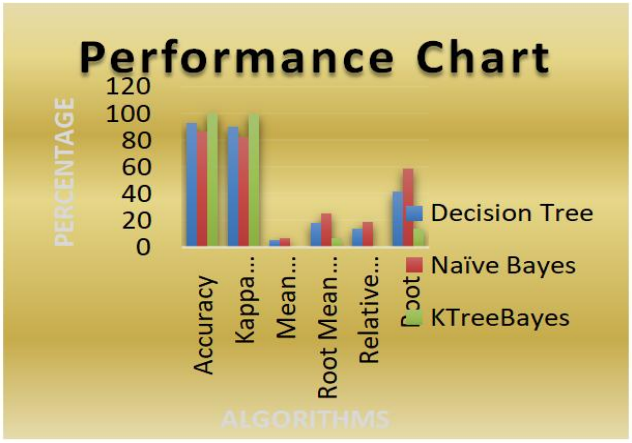

Fig.5 Performance of the model

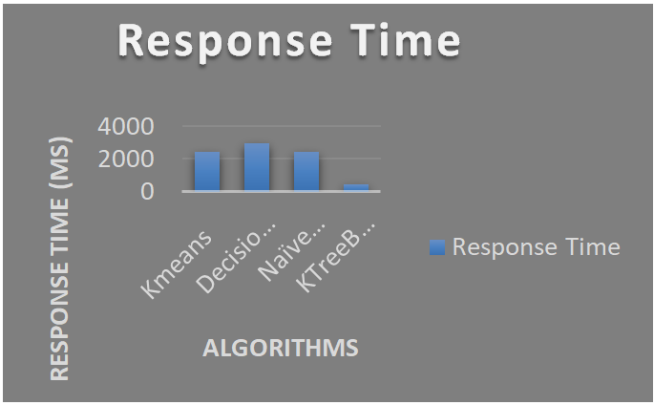

Fig.6 Response time

Table II: Comparison of algorithm based on time

Algorithms	Response Time (in milli seconds)
Kmeans	2408
Decision Tree	2940
Naïve Bayes	2421
KTreeBayes	421

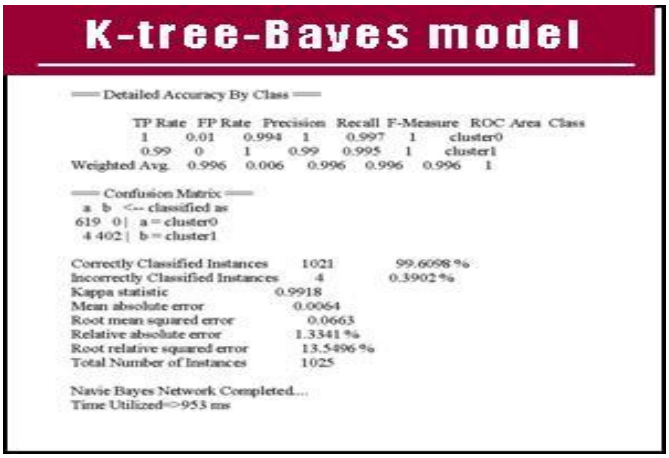

Fig.7 output of the proposed model

VI. CONCLUSION

K-Tree–Bayes model gives a perfect solution for customer data analysis with respect to the complexity in customer preferences, desires and needs based on Quality, Cost, Delivery and logistics etc., and to retain them forever. Compared to all other algorithms, K-Tree Bayes reduces the noise and ambiguity and enable to proceed with absolute clarity in analysis and to provide most appropriate data for decision making. The decisions made out of a solution provided by adopting K-Tree Bayes model algorithms are reliable, true to the nature and the results are provided with absolute timeline than other models. Solutions coming out of applying K-Tree Bayes model has become a perfect solution for Automotive companies for quick decision making and to invite new customers and to retain their existing customers.

REFERENCES

1. Bhaskar Mondal, J. Paul Choudhury, " A Comparative Study on K Means and PAM Algorithm using Physical Characters of Different Varieties of Mango in India", *International Journal of Computer Applications (0975 – 8887) Volume 78 – No.5, September 2013*
2. Preeti Arora1, Dr. Deepali2 , Shipra Varshney," Analysis of K-Means and K-Medoids Algorithm For Big Data", International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015,Nagpur, Procedia Computer Science 78 (2016) 507 – 512
3. Archana Singh, Avantika Yadav, Ajay Rana, " K-means with Three different Distance Metrics", *International Journal of Computer Applications (0975 – 8887) Volume 67– No.10, April 2013*
4. Nikita Patel, Saurabh Upadhyay, Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA, International Journal of Computer Applications (0975 – 8887) Volume 60– No.12, December 2012.
5. https://en.wikipedia.org/wiki/ Decision_tree_pruning.
6. Ahmed Mohamed Ahmed ,Ahmet Rizaner ,Ali Hakan Ulusoy , "A novel decision tree classification based on post-pruning with Bayes minimum risk", https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194168
7. Yamuna N R, Venkatesan P A Comparative Analysis of Decision Tree Methods to Predict Kidney Transplant Survival, International Journal of Advanced Research in Computer Science, Volume 5, No. 3, March-April 2014, ISSN No. 0976-5697
8. Kavita Mittal, Dr.Gaurav Aggarwal, Dr.Prerna Mahajan," A COMPARATIVE STUDY OF ASSOCIATION RULE MINING TECHNIQUES AND PREDICTIVE MINING APPROACHES FOR ASSOCIATION CLASSIFICATION", International Journal of Advanced Research in Computer Science, Volume 8, No. 9, November-December 2017, ISSN No. 0976-5697
9. K Prasanna Jyothi1 , Dr R SivaRanjani2 , Dr Tusar Kanti Mishra3 , S Ranjan Mishra4," A Study of Classification Techniques of Data Mining Techniques in Health Related Research", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 7, July 2017, ISSN(Online): 2320-9801 ISSN (Print): 2320-9798.
10. A.Shameem Fathima1 ,D.Manimegalai2 and Nisar Hundewale ," A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011 ISSN (Online): 1694-0814
11. Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C," Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, o. 1, 2010
12. Huda Hamdan Ali1, Lubna Emad Kadhum, "K- Means Clustering Algorithm Applications in Data Mining and Pattern Recognition", International Journal of Science and Research (IJSR), Index Copernicus Value (2015), ISSN (Online): 2319-7064.
13. Keyvan Vahidy Rodpysh1 , Amir Aghai2 and Meysam Majdi," APPLYING DATA MINING IN CUSTOMER RELATIONSHIP MANAGEMENT", International Journal of Information Technology, Control and Automation (IJITCA) Vol.2, No.3, July 2012
14. https://en.wikipedia.org/wiki/Naive_Bayes_classifier.
15. http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/ Lab4-NaiveBayes.pdf

16. Maneesh Singhal , Ramashankar Sharma ," Optimization of Naïve Bayes Data Mining Classification Algorithm", INTERNATIONAL JOURNAL FOR RES EARCH IN APPL I ED SCIENC E AND ENGINEERING TECHNOLO GY (IJRAS ET), Vol. 2 Issue VIII, August 2014 ISSN: 2321-9653.
17. Pouria Kaviani1 , Mrs. Sunita Dhotre2," Short Survey on Naive Bayes Algorithm", International Journal of Advance Engineering and Research Development Volume 4, Issue 11, November -2017, e-ISSN (O): 2348-4470 p-ISSN (P): 2348-6406
18. C.S Padmasini, Dr K. Shyamala "Novel K-Tree-Bayes based Multimodel Business Risk Prediction for structured and unstructured Automotive Customer data ",International Journal of Pure and Applied Mathematics, Volume 119 No. 15 2018, 1405-1414 , ISSN: 1314-3395 (on-line version).
19. https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC4614595/
20. http://epiville.ccnmtl.columbia.edu/popup/how_to_calculate_kappa.html
21. https://www.statisticshowto.datasciencecentral.com/absolute-error/
22. https://en.wikipedia.org/wiki/Root-mean-square_deviation

AUTHORS PROFILE

Dr. K. Shyamala, Associate Professor in PG Department of Computer Science at Dr. Ambedkar Government Arts College, Chennai. She has published more than 40 Journal publications. She has contributed her valuable knowledge and expertise as Technical Advisory/Programme Committee Member, Session Chair, Reviewer for various Conference and a Member in Board of studies. She is currently serving as the coordinator for Computer Literacy Programme (CLP) in the same college for more than 18 years. She has 28 years of rich teaching and research experience at various levels and domains.She has been a research guide for aspiring M.Phil and Ph.d research scholars.

Mrs. Padmasini C S, currently works as an Assistant Professor at M.O.P Vaishnav College for Women (Autonomous), Chennai. She has done Post graduation in Computer Applications and M.Phil in Computer Science. Presently, she is pursuing research under the guidance of Dr. K Shyamala. Three research papers were published in her specialized domain. She has a rich teaching experience of 15 years at the same college. She has obtained good knowledge on teaching the core concepts and Programming languages. Authored a book for Alagappa university course material on Cryptography and cyber security.