Assessment of Probability Defaults Using K-Means Based Multinomial Logistic Regression

G. Arutjothi and C. Senthamarai

***Abstract*—Classification analysis is a key and easy tool in machine learning and prediction. Because of the large amount of data and the need to convert this data into useful information and knowledge, machine learning has gotten a lot of attention in the information industry and also in society because of the large amount of data and the issues that come with it. In this paper, a K-Means based Multinomial Logistic Regression (MLR) prediction algorithm is used for evaluating the performance of Probability Defaults (PD), and suggestions are made to improve financial status. The necessary information about the members of PD has been collected from the UCI machine learning repository. The parameters are chosen for the study using the feature selection method. The research goal is to find default risk probabilities and they are assessed by accuracy, RMSE (Root Mean Squared Error), error rate, and time. K-means based Multinomial Logistic Regression (MLR) significantly outperforms other classifier models. Assessment of PD will have an impact on the financial industry.**

***Index Terms*—K-means, MLR, Probability Defaults (PD), default classification, classifier models.**

I. Introduction

Data mining is considered the process of extracting useful information from large data sets [1]. It is also used to find consistent patterns that are used to develop businesses, evaluate web-based educational programs, in computer science, chemistry, engineering, and medicine, and is used in all domains where a large amount of data is available [2]. Data is the most important component of data analytics, machine learning, and artificial intelligence. We can't train any model without data, and all current research and automation will be for data [3]. Data mining consists of classification, clustering, and association rule mining. Classification is a main function of the data mining process. Many classification techniques are available nowadays. Techniques are decision tree, support vector machine, neural network, *k*-nearest neighbor, logistic regression, etc.

Machine learning allows machines to behave and learn in the same way that humans do, while also allowing them to develop their learning abilities through data, inputs in the form of real-world interactions, and observations [4]. Machine learning is a field that is very important in computer science and statistics. It is a process that automatically uses analytical model building without the intervention of a human. It is one of the parts of the data mining process. The overall goal of machine learning is to extract patterns from large amounts of data and convert these patterns into understandable ones for further use. There are many machine learning classifiers available for different application purposes. However, these classifier algorithms will not work for all kinds of data and problems. This study will produce a clustering-based classification model for probability default prediction.

Clustering is one of the major tools used by data miners. It allows us to group entities based on their similarity. This is done based on the measurements of the distance between each entity. Cluster analysis leads to classification, structure description, new insights, and eventually exploration for the researcher. Clustering has been utilized in a variety of domains, including web mining, image processing, machine learning, artificial intelligence, pattern recognition, social network analysis, bioinformatics, geography, geology, genetics, psychology, sociology, consumer behavior analysis, marketing, and more [5].

In this research, we will concentrate on the machine learning algorithms that are used to produce the predictions. Algorithms are used to create different models for different problems. In the past several years, much work has focused on developing PD models to provide loans to enterprises. We observe several existing models and their working strategies to achieve our objective of finding the best K-means based classification model. This paper attempts to find the best prediction algorithm based on the evaluation metrics.

This paper is organised as follows: Section II provides a brief overview of the probability of default (PD). Section III gives a description of the algorithms that we use in this paper. Section IV discusses the proposed model. Section V gives the results and discussion. Section VI provides the conclusion.

II. Background of PD

A. Literature Review

The scoring of borrowers' creditworthiness is one of the most important problems to be addressed in the banking industry. PD is defined as the risk that borrowers will fail to meet their credit obligations. The credit scoring system is used to predict the PD and to reduce illegal activities [6]. These credit scoring systems are used to make decisions based on information about the borrowers. In order to make credit decisions, lenders want to minimise the risk of default on each lending decision and realise a return that compensates for the risk [7].

In general, the banking industry's success and failure is based on their ability to predict PD. If the credit amount is not collected properly, the bank will go into a loss. So, bank

Manuscript received August 20, 2021; revised November 11, 2021.

G. Arutjothi and C. Senthamarai are with Department of Computer Applications, Govt. Arts College (Autonomous), Salem-7, TamilNadu, India (e-mail: garutjothi@gmail.com, senthamaraiksrct@gmail.com).

DOI: 10.7763/IJCTE.2022.V14.1314

their profit is correlated to their PD. Predicting PD is a crucial challenge, and it is a complex task to manage and evaluate [3].

There are a large number of quantitative methods to estimate the creditworthiness of loan applicants and to evaluate the probabilities of default (PD). [8] This paper uses multinomial logistic regression with correlation-based feature selection for forecasting. They find logistic regression gives high results. SVM classifier models are powerful learning systems which are suitable for default classification and the estimation of probabilities of default (PD) [9]. Statistical models give good performance measures for credit risk evaluations [10]. Quantitative methods are common in banks' credit risk estimation. The paper [11] mainly focuses on machine learning models for small banks with large financial datasets. And also, they focus only on credit defaults, not on credit risk. They used the Classification and Regression Tree (CART) algorithm only.

In this research [12], logistic regression and decision tree algorithms are used for forecasting the PD. Logistic regression gives a little bit higher results than c4.5. [13] This paper employs three approaches to determine consumer delinquency using data from six different banks: the C4.5 decision tree, logistic regression, and random forest. In this research [14], the work for making loan decisions is made by using the CART decision tree. They compare their results with those of the k-Nearest Neighbor Classifier (K-NN) and the ANN, but find that CART-based default prediction outperforms other techniques. [15] This work proposed a new credit scoring model that is based on the hybrid feature selection method and the C4.5 classifier. This relief-based hybrid system not only has a strong mathematical basis, but also has higher accuracy and effectiveness.

This work uses a combination of ANFIS (Adaptive Network based Fuzzy Inference System), Fuzzy Clustering, and Fuzzy System Algorithm Based Dynamic Model [16]. This dynamic model works well in Iran's banking sector as well. This model replaces the static model. They compare their results with different bank datasets. [17] This paper develops a binary classifier for the prediction of default probability based on machine learning techniques. They find tree-based models are more stable than multilayer neural networks.

From the above study, it is clear that there are many classification techniques available for forecasting. But no one fits all types of datasets. This paper uses logistic regression because this method gives the best results.

III. Models and Methods

A. Logistic Regression

Regression is a statistical method, and it is used for many problems. The process of regression work is correlation and strength between dependent variables and independent variables [8]. Algorithms are used to make machine learning relevant to the current situation. Linear regression, logistic regression, ridge regression, lasso regression, polynomial regression, and Bayesian linear regression are all examples of regression models used in machine learning.

When the dependent variables are discrete, logistic regression is utilized. For example, if 0,1 or true or false, means that the target variables have only two values, A logistic function is to measure the relationship between a target variable and the independent variables. Statistical methods such as regression can be used to model the prediction of continuous values. The aim of regression analysis is to find the best model for explaining the relationship between the output and input data. In general, regression analysis establishes the relationship between the dependent (response) variable Y and one or more independent variables (inputs, regressors, or descriptive variables) X_1, X_2,..., X_n. [18]. Logistic regression is a type of regression model that is used to classify dependent variables into two classes [18]. There are two reasons to use regression analysis:

For prediction purposes, computing the output measurement from input data is inexpensive. Before predicting new unknown input data, input training data is used to classify input data. There are several different types of logistic regression.

1) Logistic regression with binary variables
2) Multinomial logistic regression

B. Multinomial Logistic Regression

Multinomial logistic regression is best suited for large numbers of variables [8]. A target variable can have three or more possible types which are not ordered and is called multinomial logistic regression (i.e., types have no quantitative significance) like "A" vs "B" vs "C". This research works fully on this method. Binary logistic regression can predict only binary output, while multinomial logistic regression can deal with one of K-possible outcomes, where K can be target classes.

$$\text{Input } D1 = i(x1, x2, \ldots xn) \tag{1}$$

$$\text{Target } S = t(D1),\ Pr(yi = k) \tag{2}$$

C. K-means Clustering

The K-means clustering method divides data objects into one-level partitions [19]. We started by selecting K initial centroids, where K is a user-specified parameter indicating the desired number of clusters. The nearest centroid is assigned to each point, and each cluster is a collection of points assigned to a centroid. Based on the points assigned to the cluster, the centroid of each cluster is modified. We repeat the assignment and update the steps until no points change clusters or the centroid remains constant. A collection of n vectors X_j, $_j$ = 1, 2,..., n, must be divided into c groups G_i, I = 1, 2,..., c.

Before convergence, the K-means algorithm will perform the following four steps:

Step1: Find the coordinates of the centroid. Gi, I = 1,2,...,c.
Step2: Measure the distance between each object and the find the centroids (Euclidean distance)
Step3: Sort the objects into groups based on their minimum distance
Step4: Repeat step2, 3 until instance are stable

Clustering is a technique for extracting commonalities and, by dividing them into groups, we can get patterns from large data sets. Clustering is commonly used when the data sets are unlabeled, and unsupervised learning is considered to be the most important problem [20]. This research work uses k-Means clustering to reduce the dataset. It is important to our research because a large dataset will reduce the classifier accuracy. The K-means classifier finds a solution that is competitive with the optimal k-means solutions.

D. Standardization

Standardization is a preprocessing method that is used to transform the original dataset into standard data. [21] paper to standardise the dataset using the *z*-score, min-max, and decimal scoring methods of the *K*-means clustering algorithm. They discovered that the *z*-score-based *K*-means cluster produces the best results when compared to other scalable methods. Preprocessing is an important step in machine learning because scalable input data can produce better results than regular datasets. In this research work, the standaredScaller method is used because it is suitable for different scaled input data. Rescaling the value distribution with the mean of observed values equal to 0 and the standard deviation equal to 1 is known as standardizing. Centering is the process of subtracting the mean value from the data, and scaling is the process of dividing by the standard deviation. As a result, the procedure is also referred to as "center scaling." The mean and standard deviation are used to estimate the more robust dataset.

The following formula can be used to calculate a standard caller.

$$y = (x - mean)/\text{standard\_deviation}$$

where the mean and standard deviation are calculated using the formula:

$$\text{mean} = \text{sum}(x) \,/\, \text{count}(x)$$

$$\text{standard\_deviation} = \sqrt{(\text{sum}(\,(x\text{- mean})\text{^}2\,) \,/\, \text{count}(x))}$$

IV. Proposed Model

In this paper, we mainly focus on finding the best prediction model. This proposed work uses K-means clustering with a Multinomial Logistic Regression classifier to evaluate the PD. Managing and analysing the financial data is more difficult because the volume of the data is huge. A model is designed using machine learning techniques in order to make a good decision. Regression techniques are also compared to find the best model.

A. K-MeansMLR Model

The proposed model is shown in Fig. 1.

Fig. 1 shows the proposed architecture. The whole dataset is taken from financial organisations and analysed to find useful information. This is a difficult or critical job in the banking industry. The proposed work finds the PD and makes the decision on whether the loan can be approved or rejected for the new potential loan applicant. This model works as two different sections. One is data preprocessing using k-means and standardisation methods for data customization. The second part uses MLR as the base classifier, which is used to find the probabilities of default.

.

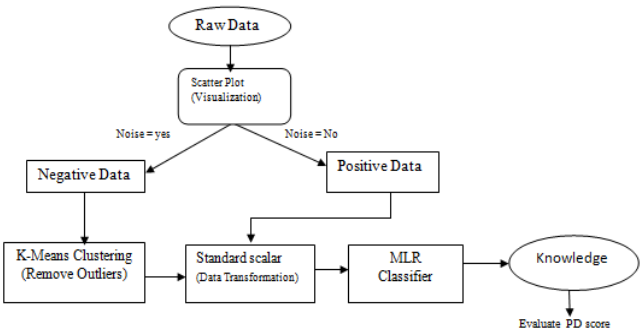

Fig. 1. A proposed model.

B. K-meansMLR Algorithm

K-Means with MLR Based Probability Default prediction Model algorithm steps are given below:

Step 1: Input Raw Dataset D. D={ $x_1, x_2, \ldots x_n$ }

Step 2: Check D=outlier, if yes then goes to step 3 otherwise step4

i) Find outliers using scatterplot method

ii) Relationship between variables

$$r = \frac{1}{N-1} \sum \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{y}}{S_y} \right)$$

r=1 (Strong), r=-1(Weak), r=0 (very poor)

Step 3: Remove the outliers using K-Means cluster

i) Choose the number of clusters k in D

ii) Select k random points from the D as centroids G_i

iii) Assign all the points to the closest cluster centroid

iv) Recompute the centroids of newly formed clusters as d

v) if old_centroids=news_centriods then goto step 4:

vi) Repeat steps iii and iv

Step 4: Normalize the dataset using Standard scalar, D into D_1

i) $D_1 = (D - mean) \,/\, standard\_deviation$

ii) Mean and standard deviation is calculated by

mean = sum (D) / count (D)

standard_deviation = $\sqrt{(sum((D - mean)^2) / count(D))}$

Step 6: Apply multinomial logistic regression to D1

i) $D_1 = f(x_1, x_2, \ldots x_n)$

ii) $S = f(D_1)$, $Pr(y_i=k)$

iii) $Pr(y_i=k \mid x_i ; \beta_1, \beta_2, \ldots \beta_m)$

iv) $$Pr(y_i=k) = \frac{\exp(\beta_{0k} + x_i \beta_k)}{\sum_{j=1}^{m} \exp(\beta_{0k} + x_i \beta_k)}$$

Step7: Validate the cross entropy of the model.

Step 8: Calculate metrics for predicted model.

Step 9: Identifying the probability defaults Score (PD).

where k=1,2,...m of the target class value. Y_i is the probability of the k_{th} class value. β_k is the row vector of regression coefficients of D_i for the k^{th} category of S. This model will work in an effective manner.

V. Result and Discussion

In this paper, credit data from Australian bank clients (available at the UCI repository) is used [22]. The data set contains records on over 30,000 customers, with each record containing 25 features. The classification problem is to find

the PD of a bank customer. The Python software is used to develop the model and evaluate the model. The classification metrics are used to find RMSE (Root Mean Squared Error), error rate, and time. Some important attributes are analysed using exploratory data analysis techniques. The below figures are represented in the original dataset-based analysis shown in Fig. 2, Fig. 3.

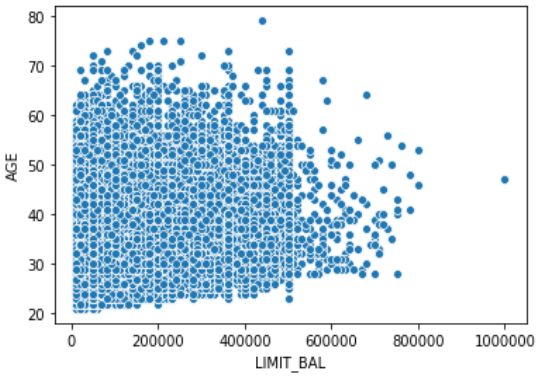

Fig. 2. Age group vs limit balance.

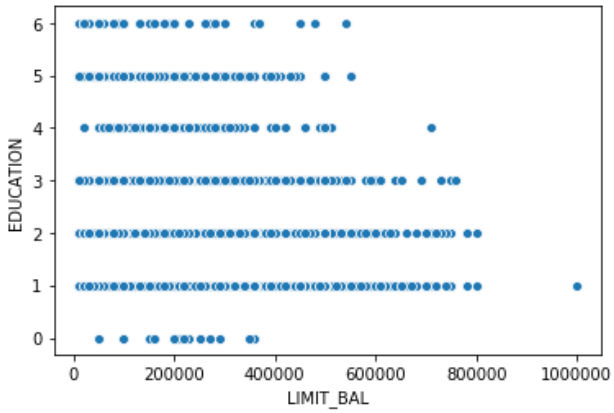

Fig. 3. Education vs Limit balance.

This work was done by Scatter Plot method-based Exploratory Data Analysis. A scatter plot is mainly used for multivariable datasets. Nowadays, the scatter plot method is mainly used for finding outliers. The plot is created for any two variables and finds the outliers. Based on this analysis, we have found the dataset has outliers. The K-means clustering method was used in this study to remove outliers.

A. Performance Metrics

Below is a list of performance indicators that were used to evaluate the proposed models' results. As shown below, the accuracy measure is primarily used to determine the total classifier outcome of the prediction process:

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+PN) \quad (3)$$

Here,

TP – True Positive
TN - True Negative
FP - False Positive
FN – False Negative.

The error is expressed as an error rate if the goal values are categorical. The error rate is calculated by the following formula:

$$\text{Error Rate} = (FP+FN)/(TP+TN+FP+FN) \quad (4)$$

The values range from 0.0, which is the best error rate, to 1.0, which is the worst. This research work used only these two metrics for evaluating the model.

B. Results

The K-means with the MLR credit scoring model was successful in classifying default and non-default loans. Hence, the lender can reduce the risk of investment failure by selecting profitable borrowers after processing the loan applications through this model. This model correctly classified the default and no-default, and it was 97% accurate in the test dataset. Table I presents the classification results of the proposed model. This work is also done with other classifiers, such as K-Nearest Neighbor (K-NN), Logistic Regression (LR), and MLR. K-meansMLR outperforms other classifiers. Table I shows the results for classifier models. The dataset can be classified as 80% of training data and 20% of testing data. Table I shows the results for classifier models.

TABLE I: CLASSIFIER MODEL RESULTS

Classifier Model	Data Splitting 80%: 20%			
	Accuracy	RMSE	Error Rate	Time in Seconds
K-NN	81	0.43	0.76	3.93
LR	82	0.42	0.78	3.21
MLR	82	0.42	0.78	2.11
K-MLR	97	0.67	0.45	6.57

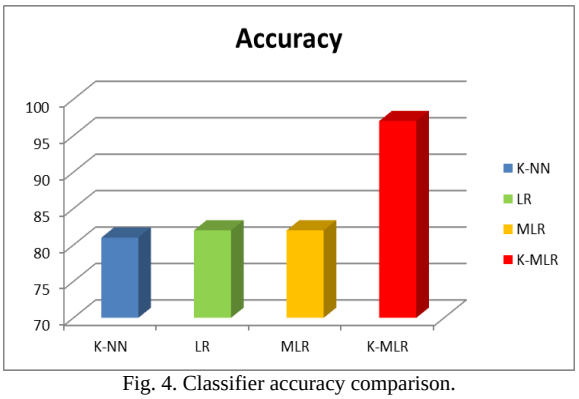

Fig. 4. Classifier accuracy comparison.

Fig. 4 shows the comparison of classifier accuracy. The graph is drawn with the classification algorithms on the *x*-axis and the percentage of classification accuracy on the *y*-axis.

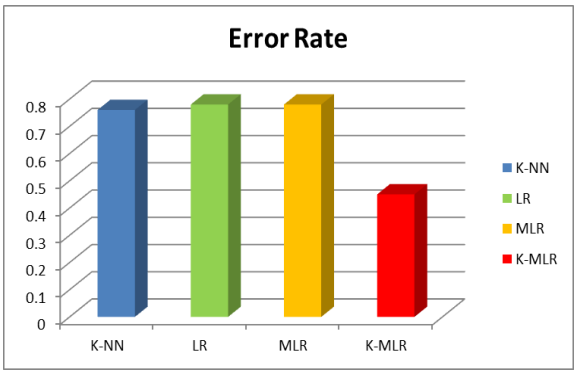

Fig. 5. Comparison of classifier error rate.

The K-means with MLR technique shows the highest percentage of classification accuracy of the other classifiers. Fig. 5. shows the comparison of classifier error rates. The graph is drawn with the classification algorithms on the *x*-axis and the percentage error rate on the *y*-axis.

The K-Means clustering with a multinomial logistic regression based forecasting model gives the highest accuracy (3). The K-meansMLR achieved 97%, which is higher than the other classifier models and also reduces the error rate (4). Comparisons are made between the classifier models. Machine learning techniques are used to develop the best forecasting models. The K-meansMLR-based forecasting system provides higher accuracy than other classifiers. The data is used to develop the K-meansMLR PD model with splitting criteria. Fig. 4 shows the maximum accuracy of training data. This proposed model presented in this study can be effectively used by loan lenders to predict the loan applicant. Lenders can use this model to predict the PD of the loan applicant.

VI. Conclusion

In this paper, we have proposed a model to identify the probability of default of a credit applicant in an effective manner. The proposed model, K-meansMLR, shows a 97% accuracy rate in classifying training data using Python. Furthermore, a comparison study has been conducted with different classifier models. The K-means clustering-based MLR system gives the highest accuracy and also the lowest error rate compared to other classifiers. This model can be used to forecast any type of prediction problem. The current paper only focuses on accuracy and error rate. The result suggests K-meansMLR is best suited for large datasets, but this model requires high processing time compared to other models. However, this study considers only the K-meansMLR-based Probability Default model. Hence, the future research is to enhance the K-meansMLR for all types of datasets and compare it to other machine learning classifiers with optimization techniques.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

The research scholar author, G. Arutjothi proposed the idea, conducted the research, and wrote the paper. The assistant professor author, C. Senthamarai, reviewed the article of whole research and approved this research.

References

[1] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, Cambridge, MA.: MIT Press, 2001.

[2] Y. L. Chen, J. M. Chen, and C. W. Tung, "A data mining approach for retail 14knowledge discovery with consideration of the effect of shelf-space adjacency on sales," *Decision Support Systems*, vol. 42, pp. 1503-1520, 2007.

[3] M. Gupta. (2020). ML introduction and data machine learning. [Online]. Available: https://www.geeksforgeeks.org/ml-introduction-data-machine-learning/

[4] D. Faggella. (2019). What is machine learning? Emerj - Artificial intelligence research and insight. *Emerj*. [Online]. Available: https://emerj.com/ai-glossary-terms/what-is-machine-learning/

[5] Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review", *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, 1999.

[6] Devi, C. R. Durga, and R. M. Chezian, "A relative evaluation of the performance of ensemble learning in credit scoring," in *Proc. IEEE International Conference on Advances in Computer Applications (ICACA)*, IEEE, 2016.

[7] A. Byanjankar, M. Heikkilä, and J. Mezei, "Predicting credit risk in peer-to-peer lending: A neural network approach," in *Proc. 2015 IEEE Symposium Series on Computational Intelligence*, IEEE, 2015.

[8] S.-H. Moon and Y.-H. Kim, "An improved forecast of precipitation type using correlation-based feature selection and multinomial logistic regression," *Atmospheric Research*, vol. 240, 2020.

[9] J.-He. Trustorff, P. M. Konrad, and J. Leker, "Credit risk prediction using support vector machines," *Rev. Quant. Finan. Acc.*, vol. 36, pp. 565–581, 2011.

[10] L. Sun, "A re-evaluation of auditors opinions versus statistical models in bankruptcy prediction," *Rev. Quantitate Finance Account*, vol. 28, pp. 55–78, 2007.

[11] E. A. Khandani, A. J. Kim, and A. W. Lo, "Consumer credit-risk models via machine-learning algorithms," *Journal of Banking and Finance*, vol. 34, pp. 2767–2787, 2010.

[12] G. Nie *et al.*, "Credit card churn forecasting by logistic regression and decision tree," *Expert Systems with Applications*, vol. 38, no. 12, pp. 15273-15285, 2011.

[13] B. Florentin, Q. Chen, B. Clark, S. Das, A. W. Lo, and A. Siddique, "Risk and risk management in the credit card industry," *Journal of Banking and Finance*, vol. 72, pp. 18–39, 2016.

[14] G. Jorge and P. Tamayo, "Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications," *Computational Economics*, vol. 15, pp. 107–143, 2000.

[15] Arutjothi and Senthamarai, "Credit risk evaluation using hybrid feature selection method," *Software Engineering and Technology*, vol. 9, no. 2, pp. 23-26, 2017.

[16] S. Moradi and F. M. Rafiei, "A dynamic credit risk assessment model with data mining techniques: Evidence from Iranian banks," *Financ. Innov.*, vol. 5, p. 15, 2019.

[17] P Addo, D. Guegan, and B. Hassani, "Credit risk analysis using machine and deep learning models," *Risks*, vol. 6, p. 38, 2018.

[18] M. Sustersic, D. Mramor, and J. Zupan, "Consumer credit scoring models with limited data," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4736-4744, 2009.

[19] P.-N. Tan, M. Steinbach, and V. Kumar, "Text book on introduction to data mining," *Pearson Education*, pp. 496-497.

[20] K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. A. Kurgan, *Data Mining: A Knowledge Discovery Approach*, New York: Springer, 2007.

[21] M. B. Ismail and D. Usman, "Standardization and its effects on K-means clustering algorithm," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 6, no. 17, pp. 3299-3303, 2013.

[22] University of Massachusetts Amherst. [Online]. Available: http://mlr.cs.umass.edu/ml/datasets.html

G. Arutjothi received her both B. Sc and M.sc., computer science from Government Arts College, Dharmapuri Tamil Nadu, India. Now she is pursuing her Ph.D Computer Science in PG and Research, Department of Computer Science at Government Arts College (Autonomous), Salem -07, Tamil Nadu, India. Her research interest includes Data Mining, Big Data Analytics and Machine Learning. She Presented five papers in International and National level conferences. She published five papers at various international journals

C. Senthamarai received her MCA degree from Madras University in 1991. She completed her Ph.D degree in Computer Science from Periyar University in 2008. She served as a HOD of Computer applications at KSR College of Technology, Nammakkal. At present she is working as an Assistant Professor in Computer Applications in the Department of Computer Applications, Government Arts College (Autonomous), Salem-07, Tamil Nadu, India. She has published 20 papers in international/ national journals. Her research focus is on Cloud Computing, Big Data and Data Mining.