Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 7, July 2015, pg.305 – 314

RESEARCH ARTICLE

Comparative Analysis of Data Reduction Model for Credit Scoring

Mahak[1], Pooja Mittal[2]

[1]DCSA, Maharshi Dayanand University, India

[2]DCSA, Maharshi Dayanand University, India

[1] mahak91saini@gmail.com; [2] mpoojamdu@gmail.com

Abstract*— *The development of credit card application should be in proportion with the expectation of terrible credit hazard on the grounds that it doesn't utilize security collateral as guarantee. The use of credit scoring can be utilized to help the credit hazard assay in deciding the customer's eligibility. Data mining has been demonstrated as a significant tool for credit scoring. The objective of this exploration is to design a data mining model for credit scoring in bank keeping in mind the end goal to bolster and enhance the execution of the credit expert job. The proposed model applies filtration, characteristic determination, grouping techniques and the best precision is accomplished by Stratified removal folds filter. Also GainRatioAttribute method has selected the best attributes and bayes net and decision tree have shown equal results under classification techniques.

Keywords*—*Data Mining, Bank, Credit Card, Credit Scoring, Filters, Attribute selection, Classification

I. Introduction

Data mining is a procedure that takes data as input and yields knowledge. Data mining is the way of investigating information from alternate points of view and condensing it into helpful data. Data mining programming is one of the systematic mechanisms to analyse information. One of the earliest and most cited definitions of the data mining procedure, which highlights some of its particular qualities, is given by Fayyad, Piatetsky-Shapiro and Smyth (1996), who characterize it as "the nontrivial procedure of recognizing legitimate, novel, conceivably helpful, and eventually justifiable examples in information." It permits clients to break down information from a wide range of measurements or edges, order it, and condense the connections distinguished. Data mining is a key stride in knowledge discovery in databases (KDD) process that creates helpful examples or models from information. The term KDD and data mining are diverse. KDD alludes to the general procedure of finding valuable information from information. Data mining alludes to find new examples from an abundance of information in databases by concentrating on the calculations to concentrate helpful learning. The KDD procedure is intelligent and iterative, including various steps with numerous choices made by the client.

The KDD procedure is outlined in Figure 1. This procedure incorporates a few stages, comprising of data selection, data treatment, data pre-processing, data mining and interpretation of the results. This procedure is interactive, since there are numerous choices that must be taken by the leader amid the procedure. The phases of KDD procedure are briefly depicted below [15]:

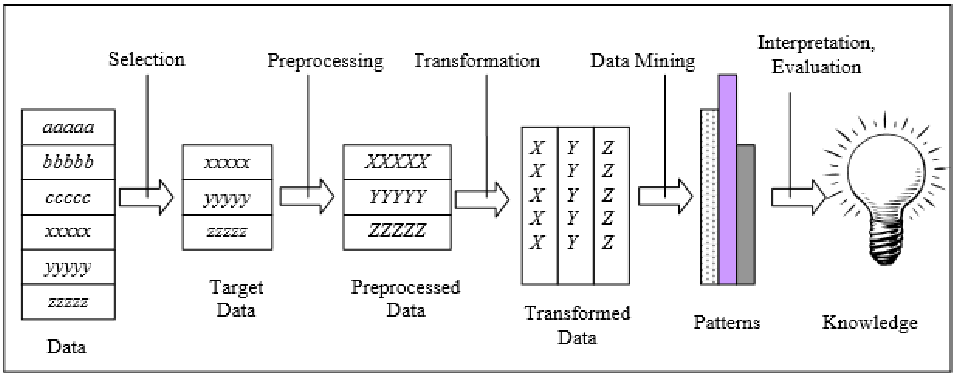


Fig. 1 KDD process

- ***A.*** ***Data selection-*** This stage incorporates the investigation of the application area, and the selection of the data. The domain's study plans to contextualize the venture in the organization's operations, by comprehension the business dialect and characterizing the objectives of the undertaking. In this stage, it is important to assess the base subset of information to be chosen, the relevant attributes and the pertinent time to consider.
- ***B.*** ***Data pre-processing-*** This stage incorporates essential operations, for example, removing noise or outliers, gathering the fundamental data to model or record for noise, choosing methods for taking care of missing information traits, and representing time grouping data and known changes. This stage likewise incorporates issues with respect to the database administration framework, for example, information sorts, diagram, and mapping of missing and obscure qualities.
- ***C.*** ***Data transformation-*** This stage comprises of preparing the information, with a specific end goal to change over the data in the fitting configurations for applying data mining algorithms. The most common transformations are: data standardization, data conglomeration and data discretization. To standardize the information, every value is subtracted from the mean and divided by the standard deviation. A few algorithms just manage quantitative or qualitative information. Thus, it might be important to ruin the information, i.e. map qualitative information to quantitative information, or quantitative information to qualitative information.
- ***D.*** ***Data mining-*** This stage comprises of discovering patterns in a dataset beforehand arranged. A few calculations are assessed with a specific end goal to distinguish the most suitable for a particular assignment. The chosen one is then connected to the appropriate information, keeping in mind the end goal to discover circuitous connections or other interesting examples.
- ***E.*** ***Pattern evaluation-*** This stage comprises of deciphering the discovered patterns and assessing their utility and significance concerning the application area. In this stage it can be reasoned that some significant properties were disregarded in the investigation, accordingly proposing the need to reproduce the procedure with an upgraded arrangement of qualities.

Different commercial ventures have been receiving data mining to their main goal discriminating business procedures to increase game changers and help business develops. A portion of the data mining applications are in sale/marketing, banking/finance, health care and insurance, transportation and medicine.

Data mining is becoming strategically important area for many business organizations including banking sector. It is a process of analyzing the data from various perspectives and summarizing it into valuable information. Data mining assists the banks to look for hidden pattern in a group and discover unknown relationship in the data. Today, customers have so many opinions with regard to where they can choose to do their business. Early data analysis techniques were oriented toward extracting quantitative and statistical data characteristics. These techniques facilitate useful data interpretations for the banking sector to avoid customer attrition. Customer retention is the most important factor to be analyzed in today’s competitive business environment. And also fraud is a significant problem in banking sector. Detecting and preventing fraud is difficult, because fraudsters develop new schemes all the time, and the schemes grow more and more sophisticated to elude easy detection.

This article provides a critique of the concept of Data mining and Credit scoring in organized Banking and Retail industries.

II. LITERATURE SURVEY

As indicated by the most recent RBI insights, there are more than 2 crore credit card clients in India. That development additionally incorporates hazard increase for the bank. The danger is the likelihood of terrible obligations. Furthermore, a credit card does not use security collateral as warranty. In this way, the credit examiner job in figuring out which credit card application to be affirmed or rejected is extremely urgent undertaking. Credit Scoring is utilized to bolster credit hazard investigation [23], [24]. With fitting Credit Scoring models, the bank will have the capacity to assess whether a candidate is plausible to get a Credit card. The reason for this study are to focus the correct data mining framework for credit scoring in Bank so as to enhance the execution and support the credit investigators work.

In the literature [4] investigated the nature and impacts of missing data in credit risk modeling and consider current rare information set on shopper borrowers, which incorporates distinctive percent and dispersions of missing data.

Data mining has been demonstrated as a significant device for banking and retail commercial ventures [5], which distinguish valuable data from expansive size information. While past literary works [6], [7], [19], [20] had connected data mining strategies for credit scoring.

In literature [6] talked about the focal points and use of credit scoring and also the improvement of its model utilizing data mining. Data mining procedures which utilized for acknowledge scoring models, for example, logistic relapse, neural network, and decision tree.

The choice of imperative components or properties that affect the execution of credit scoring model was done in [7]. The procedure of selecting the best components or characteristics utilized four information digging strategies for highlight choice, for example, ReliefF, Correlation-based, and Consistency-based Wrapper calculations for enhancing three parts of credit scoring model execution like effortlessness, velocity and exactness.

Other previous researches looks into component determination for credit scoring model were directed in works [20], [9], [10], [11].

As indicated by [8], client distinguishing proof by a behavioural scoring model is useful qualities of client and encourages showcasing method improvement.

Credit venture investigation turned into the principle concentrate on the financial and banking industries [19]. To enhance exactness, the investigation built by conglomeration of techniques that combined a few delegate algorithms and afterward utilized specific voting approach.

Utilizing retail credit information of banks in Czech Republic, the two credit risk models are built based on light of logistic regression and Classification and Regression Trees (CART) [12].

The exploration [20] utilized logistic relapse, neural networks, C5, naïve bayes updateable, IBK (instance-based learner, k nearest neighbour) and raced incremental keeping in mind the main goal to choose the best classifier which was utilized to enhance the prescient exactness of credit risk of Credit card clients in Malaysian Bank. Also, attribute selection utilizing ID3 calculation was performed to choose subsets of data that has the most noteworthy information gain and gain ratio values.

By breaking down the consequences of past inquires about which utilized data mining for credit scoring, this paper will examine about how to plan a data mining reduction model for credit scoring in the Bank.

III. DATASET USED

A dataset (or information set) is an accumulation of information. Most frequently a dataset relates to the contents of a solitary database table, or a solitary factual information grid, where every segment of the table symbolizes a specific variable, and every row compares to a specific individual from the dataset being referred to. The dataset records values for each of the variables, for example, height and weight of an object, for each member of the dataset. Every value is known as a datum. The dataset may include information for one or more individuals.

The characteristics of the data set used in this research are summarized in following Table I. The dataset is obtained from Credit Approval dataset of UCI Machine Learning Repository. Available records in the dataset are classified into two class labels, '+' and '-'. '+' stands for approval of credit card whereas '-'stands for rejection of credit card. The class label is determined by credit experts' knowledge. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. The dataset is interesting because there is a good mix of attributes -- continuous, nominal with small numbers of values, and nominal with larger numbers of values. There are also a few missing values. [13]

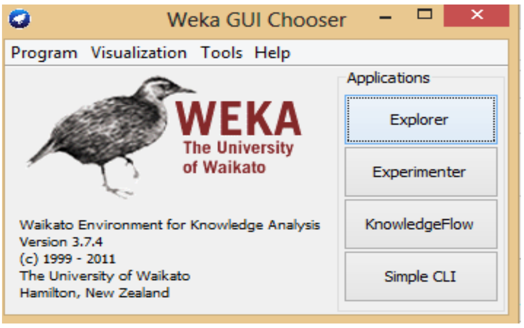
TABLE I

DESCRIPTION OF DATASET

Data Set Characteristics:	Number of Instances:	Area:	Number of Attributes:	Attribute Characteristics:	Associated Tasks:
Multivariate	690	Financial	15	Categorical, Integer, Real	Classification

IV. TOOL

Weka is developed by the University of Waikato (New Zealand) and its first advanced structure is executed in 1997.It is open source implies it is accessible for public use. Weka code is composed in Java language and it contains a GUI for Interacting with information documents and delivering visual results.



Waikato Environment for Knowledge Analysis
Version 3.7.4
(c) 1999 - 2011
The University of Waikato
Hamilton, New Zealand

Fig 2: WEKA Tool

The GUI Chooser comprises of four buttons:

•Explorer: A situation for investigating information with WEKA.
•Experimenter: A situation for performing examinations and directing factual tests between learning plans.
•Knowledge Flow: This environment bolsters basically the same capacities as the Explorer, yet with a drag and drop interface. One favourable position is that it underpins incremental learning.
•Simple CLI: Provides a straightforward command-line interface that permits direct execution of WEKA commands for working frameworks that do not provide their own particular command line interface.

V. EMPIRICAL STUDY

Credit scoring is helpful for credit suppliers, as well as for the credit borrowers. For instance, credit scoring help lessen segregation as the credit scoring model gives a target assay of the viability of the candidate. Furthermore, the credit suppliers concentrate just on data associated with credit risk and maintain a strategic distance from subjectivity of credit investigators. Credit scoring additionally reinforces the rate and consistency of the credit application process and empowers the computerization of the credit application process. Along these lines, human intercession in the credit assessment and expense can be decreased. The use of credit scoring will bolster the monetary organizations to quantify the danger related to provide loan to the candidate in a brief while. Moreover, these organizations can settle on better choices [6].The objective of this research is to outline a data reduction model for credit scoring in bank to support and enhance the execution of the credit analyst job.

Figure 3 demonstrates the proposed model that applies diverse attribute selection and classification techniques. Before building the model, data pre-processing is applied on the dataset. Data pre-processing comprises of data cleaning (fill in missing qualities), data reduction (attribute selection) and data transformation.

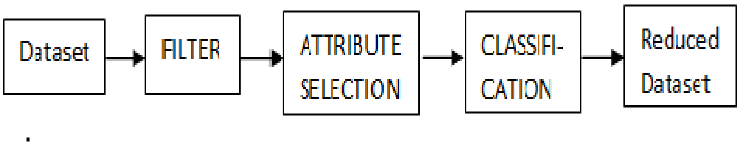


Fig 3: Proposed data reduction model

In the above figure after applying filters, attribute selection is done by using three different methods for improving three aspects of credit scoring model performance like simplicity, speed and accuracy. After that four different classification techniques are used to find out the correctly classified instances. Then on the selected attributes we get above by three methods same classification techniques are applied and in result we find the technique having best correctly classified instances.

A. Filters

We require astounding dataset to deliver the top notch mining results [8]. The elements which influence the data quality, includes precision, completeness, and consistency [8]. Along these lines, we have to apply data pre-processing first before utilizing the principle data mining process. We utilize diverse Supervised and unsupervised channels for this purpose which are resample, stratified removal folds, spread subsample, remove frequent values, remove folds.

B. Attribute Selection

Attribute subset selection is the procedure of recognizing and evacuating however much of the unimportant and excess data as could be expected. Decreasing the dimensionality of the data diminishes the span of the speculation space and permits calculations to work quicker and all the more viably. In some cases accuracy on future classification can be improved; in others, the result is a more compact, easily interpreted representation of the target concept.

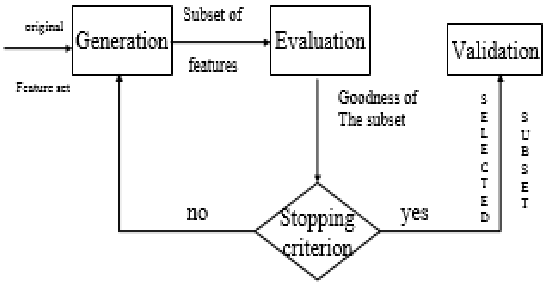


Fig 4: Attribute Selection

Here Stopping criterion refers to a specific condition on the basis of which optimal subset is evaluated from the given dataset [25]. The process of selecting the best features or attributes in this paper uses three data mining methods: CfsSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval.

1) **CfsSubsetEval:** CFS (Correlation-based Feature Selection) [16], [17] is one of the methods that assess subsets of attributes as opposed to individual attribute. At the heart of the calculation is a subset assessment heuristic that considers the value of individual attribute for foreseeing the class alongside the level of relationship between them. The heuristic (Equation 1) assigns high scores to subsets containing attributes that are highly correlated with the class and have low inter-correlation with each other.

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \qquad (1)$$

2) **InfoGainAttributeEval:** This is one of the simplest (and fastest) attribute ranking methods and is often used in text categorization applications [6], [7] where the sheer dimensionality of the data prevent more sophisticated attribute selection techniques. If A is an attribute and C is the class, Equations 2 and 3 give the entropy of the class before and after observing the attribute.

$$H(C) = -\sum_{c \in C} p(c) \log_2 p(c), \qquad (2)$$

$$H(C|A) = -\sum_{a \in A} p(a) \sum_{c \in C} p(c|a) \log_2 p(c|a). \qquad (3)$$

The amount by which the entropy of the class decreases reflects the additional information about the class provided by the attribute and is called information gain [24]

3) **GainRatioAttributeEval:** A decision tree is a basic structure where internal nodes symbolize tests on one or more attributes and external nodes reflect category of results. The information gain measure is utilized to choose the test quality at every node of the decision tree. The information gain measure prefers attributes having enormous value. The simple decision tree induction algorithm ID3 [18] was improved by C4.5 [21, 22]. C4.5 a successor of ID3, uses an expansion of information gain known as gain ratio, which endeavours to defeat this inclination.

C. Classification on Dataset

Classification is the frequently used data mining technique, which utilizes an arrangement of preclassified samples to build up a model that can group the number of records at large [6]. Fundamentally, classification is utilized to arrange every element in a set of data into one of predefined set of classes or groups. Classification techniques utilize numerical methods, for example, decision trees, linear programming, neural network and statistics. In classification, we make the product that can figure out how to order the data items into groups [5]. Credit risk applications are especially appropriate to this kind of assay. The data procedure includes learning and classification. The model uses BayesNet, NaiveBayes, Decision Table, ZeroR techniques to find out the correctly classified instances.

1) **Naïve Bayes:** A naïve (or simple) Bayesian classifier based on Bayes' theorem is a probabilistic factual classifier [23], which the expression "naive" demonstrates conditional autonomy among features or attributes. Its significant point of interest is its speed of utilization on the grounds that it is the easiest algorithm among classification techniques. Thus, it can promptly handle a dataset with numerous characteristics.

2) **Decision Tree:** Ross Quinlan presented a decision tree algorithm (known as Iterative Dichotomiser (ID3) in 1979. Decision tree classifiers build a flowchart-like tree structure in a top down, recursive, partition and-vanquish, way [23]. Utilizing The Attribute Selection Method (ASM), it chooses a parting method (attribute) that best parts the given records into each of the class labels, and afterward selected attributes get to be nodes in a decision tree.

3) **Bayes Net:** Bayesian networks (BNs), also known as belief networks (Or Bayes nets for short), belong to the family of probabilistic graphical models (GMs). These graphical structures are used to represent knowledge about an unknown domain. Specifically, every node in the graph symbolizes an arbitrary variable and the edges between the nodes symbolize probabilistic conditions among the equivalent arbitrary variables. These restrictive conditions in the graph are regularly evaluated by utilizing known statistical and

computational methods. Hence, BNs combine principles from graph theory, probability theory, computer science, and statistics.

4) **ZeroR:** ZeroR is the easiest classification technique which depends on the objective and overlooks all the indicators. It just predicts the dominant part (class). Albeit there is no consistency control in ZeroR, it is valuable for deciding a standard execution as a benchmark for other classification techniques.

D. Classification on selected attributes

The four classification techniques are again applied on the selected attributes to find their correctly classified instances.

VI. RESULTS

The results found during the processing of the model are described in the following sections. Table II shows the result found in each step during the execution and the results in figure 3 shows the significant result.

TABLE II

RESULT OF EACH STEP PERFORMED IN THE PROPOSED MODEL

FILTERS	SELECTED ATTRIBUTES		CLASSIFICATION	
	METHODS USED	**ATTRIBUTES**	**TECHNIQUES USED**	**CORRECTLY CLASSIFIED INSTANCES (%)**
RESAMPLE	CfsSubsetEval	3,8,9,11,15	BayesNet NaiveBayes Decision Table ZeroR	74.49 71.44 75.36 52.31
	GainRatioAttributeEval	9,10,8,11,15	BayesNet NaiveBayes Decision Table ZeroR	100 88.84 100 63.76
	InfoGainAttributeEval	9,11,15,10,8	BayesNet NaiveBayes Decision Table ZeroR	76.37 72.89 76.37 52.31
STRATIFIED REMOVAL FOLDS	CfsSubsetEval	3,7,8,9,10	BayesNet NaiveBayes Decision Table ZeroR	92.75 79.71 73.62 52.31
	GainRatioAttributeEval	9,15,11,10,3	BayesNet NaiveBayes Decision Table ZeroR	100 89.85 100 63.76
	InfoGainAttributeEval	9,11,10,6,15	BayesNet NaiveBayes Decision Table ZeroR	10089.71 100 57.24
REMOVE FREQUENT VALUES	CfsSubsetEval	4,6,8,9,11	BayesNet NaiveBayes Decision Table ZeroR	77.39 72.02 76.37 52.31
	GainRatioAttributeEval	9,10,11,15,8	BayesNet NaiveBayes Decision Table ZeroR	100 87.68 100 57.24

	InfoGainAttributeEval	5,7,6,2,1	BayesNet NaiveBayes Decision Table ZeroR	72.27 72.12 72.86 69.02
REMOVE FOLDS	CfsSubsetEval	7,9,11	BayesNet NaiveBayes Decision Table ZeroR	72.60 67.53 71.01 52.31
	GainRatioAttributeEval	9,11,10,7,6	BayesNet NaiveBayes Decision Table ZeroR	100 91.30 100 47.24
	InfoGainAttributeEval	5,7,6,2,1	BayesNet NaiveBayes Decision Table ZeroR	72.27 72.12 72.86 69.02
SPREAD SUBSAMPLE	CfsSubsetEval	4,6,8,9,11	BayesNet NaiveBayes Decision Table ZeroR	76.16 74.12 75.87 75.87
	GainRatioAttributeEval	9,10,11,15,8	BayesNet NaiveBayes Decision Table ZeroR	100 87.68 100 57.24
	InfoGainAttributeEval	9,11,10,15,8	BayesNet NaiveBayes Decision Table ZeroR	100 87.68 100 57.24

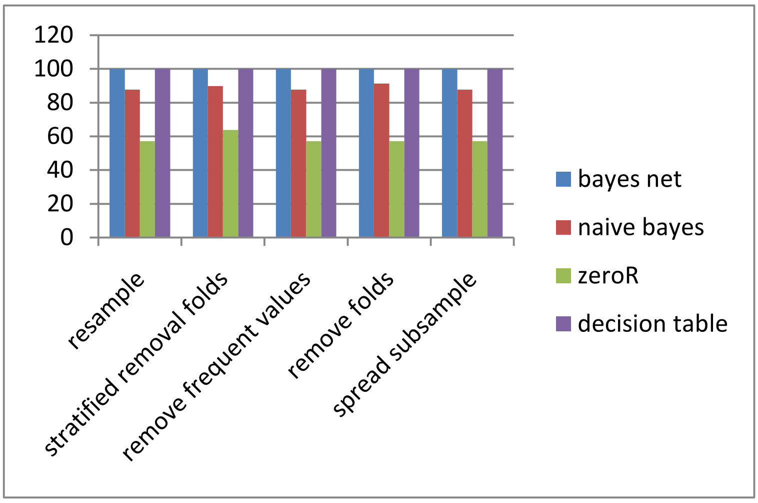


Fig 5: Graph showing significant results

As shown in the table and graph, following three results can be concluded:

1. Stratified removal folds filter has performed well.
2. Among the attribute selection methods, GainRatioAttribute method has selected the best attributes which lead to the significant results.
3. From the four classification techniques used, bayes net and decision tree have shown equal results.

VII. CONCLUSION

We have presented a data mining model which applies filtration, attribute selection, classification methods for credit scoring in credit card application. The best accuracy is achieved by attributes of GainRatioAttribute filtered through stratified removal folds in which naive bayes has 89.85%, bayes net has 100%, ZeroR has 63.76%, decision table has 100% of accuracy. So we can conclude that stratified removal folds have better accuracy than any other filter. In addition, the proposed data mining model able to improve the performance and support the credit analyst's job.

ACKNOWLEDGEMENT

Our thanks to IJCSMC for allowing us to modify templates they had developed. Also I take this opportunity to express my special gratitude to my supervisor Mrs. Pooja Mittal generously making herself available for her most valuable guidance, encouragement. It would get not be possible for me to make this paper without her guidance.

REFERENCES

[1] Aastha Joshi : *A Review: Comparative Study of Various Clustering Techniques in Data Mining* ,Student of Masters of Technology, Department of Computer Science and Engineering Sri Guru Granth Sahib World University, International Journal of Advanced Research in Computer Science and Software Engineering ,Volume 3, Issue 3, March 2013

[2] http://www.uta.edu/faculty/sawasthi/Statistics/stdatmin.html

[3]http://www.ibm.com/developerworks/library/ba-data-mining-techniques/Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. *From data mining to knowledge discovery in databases.* AI Magazine, 17(3):37-54.

[4]R.F. Lopez, "*Effects of missing data in credit risk scoring, A comparative analysis of methods to achieve robustness in the absence of sufficient data*", Journal of the Operational Research Society , Vol. 61, No. 3, 2010, pp. 486 -501.

[5] A.M. Hormozi and S. Giles, "*Data Mining: A Competitive Weapon for Banking and Retail Industries*", Information Systems Management, Spring , Vol. 21, No. 2, ProQuest Research Library, 2004.

[6] G.C. Peng,"*Credit scoring using data mining techniques*", Journal of Singapore Management Review , ISSN: 0129-5977, 2004.

[7] Y. Liu and M. Schumann, "*Data mining feature selection for credit scoring models*", Journal of the Operational Research Society , Vol. 56, 1099–1108 r 2005 Operational Research Society Ltd, 2005.

[8]N.C. Hsieh, "*An integrated data mining and behavioral scoring model for analyzing bank customers*", Expert Systems with Applications , Vol. 27, 2004, pp. 623–63.

[9] J. Li, H. Wei, and W. Hao, "*Weight-selected attribute bagging for credit scoring",* Hindawi Publishing Corporation, Mathematical Problems in Engineering , Volume 2013, doi: 10.1155/2013/379690, 2013.

[10] P. Somol, B. Baesens, P. Pudil, and J. Vanthienen, "*Filter- versus wrapper-based feature selection for credit scoring",* International Journal of Intelligent Systems , Vol. 20, Issue 10, October 2005, doi: 10.1002/int.20103, pp. 985–999.

[11] B. Waad, B.M. Ghazi, and L. Mohamed, "*Rank aggregation for filter feature selection in credit scoring",* International Conference on Control, Engineering, and Information Technology (CEIT'13), Economics, Strategic Management of Business Process , Vol. 1, 2013, pp. 64-68.

[12]E. Kocendaand M. Vojtek, "*Default Predictors in Retail Credit Scoring*: Evidence from Czech Banking Data", William Davidson Institute Working Paper Number 1015 , Electronic copy available at http://ssrn.com/abstract=1912049, 2011.

[13] https://archive.ics.uci.edu/ml/datasets/Credit+Approval

[14] Bharat Chaudhari1, Manan Parik"*A Comparative Study of clustering algorithms Using weka tools*" International Journal of Application or Innovation in Engineering & Management (IJAIEM)

[15] J. Han and M. Kamber, "*Data Mining: Concepts and Techniques*", San Francisco, Morgan Kaufmann Publishers, 2012.

[16] M. A. Hall, Correlation-based feature selection for machine learning, Ph.D. thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1998.

[17] Mark Hall, "*Correlation-based feature selection for discrete and numeric class machine learning*," in Proc. of the 17th International Conference on Machine Learning (ICML2000, 2000).

[18] J.R. Quinlan, Induction of Decision Trees, Machine Learning 1: pp.81-106, Kluwer Academic Publishers, Boston, (1986).

[19]S. Kotsiantis, "*Credit Risk Analysis using Hybrid Data Mining Model"*, Int. Journal Intelligent Systems Technologies and Applications , Vol. 2, No. 4, 2007.

[20] L.K. Sheng and T.Y. Wah, "*A comparative study of data mining techniques in predicting consumers' credit card risk in banks"*, African Journal of Business Management , Vol. 5, No. 20, Available online at http://www.academicjournals.org/AJBM, ISSN 1993-8233, 2011, pp. 8307-8312.

[21] J.R. Quinlan, San Mateo, C4.5 Programs for Machine Learning: Morgan KaufmANN, (1993).

[22] J.R. Quinlan, Bagging, Boosting and C4.5, In Proc. 13th National Conf. Back Propagation Intelligence (AAAI'96), pp. 725-730. Portland, (Aug, 1996).

[23] I. Yoo, P. Alafaireet, M. Marinov, K. Pena- Hernandez, R. Gopidi, J. Chang, and L. Hua, "*Data Mining in Healthcare and Biomedicine: A Survey of the Literature",* J Med Syst , Vol. 36, 2012, pp. 2431–2448.

[24] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA., 1993.

[25] Data Preprocessing for Supervised Leaning, S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas