Experimental Evaluation of Machine Leaninng Classification Techniques for Blood Bank Dataset

Dr. Pankaj Saxena[1]
[1]R.B.S Management Technical Campus, Agra, India
Dr. A. K. Sharma[2]
[2]MIT, Kota, India
E-mail: drarvindkumarsharma@gmail.com
Ahmad Tasnim Siddiqui[3]
[3]Assistant Professor, Computer Applications, Sherwood College of Professional Management, Lucknow, India
Dr. Kamal Upreti[4], Assistant Professor, Information Technology, Inderprastha Engineering College, Ghaziabad, India

Abstract- Recently, the use of machine learning has risen in popularity among younger scholars. Machine learning is the process of studying massive datasets, processing the data using machine learning techniques, and translating the results into a structure that's easier to understand. Classification has been utilised for many dataset properties in the machine learning process. There have been numerous classifiers employed in the dataset classification process, with the different datasets classified into different classes. Machine learning classification algorithms and types are presented in this paper which is expected to be used to classify datasets. The purpose of this document is to show experimental results on blood bank datasets for various machine learning classification techniques in a practical manner, utilising WEKA.

Keywords– Machine Learning, Classification, WEKA, Blood Bank Dataset

I. INTRODUCTION

This much of data is created and recorded every day in the modern world. Thus, a great deal of data analysis is required, but no analytical tool is available to help with the job. Machine learning is the process of investigating patterns that are obscured in raw data. When we evaluate data using interestingness criteria, we discover hidden patterns. Extracting relevant and critical information from vast amounts of blood bank data to deliver blood machine learning on time is an essential part of the field. Due to the enormous volume of blood donation data, raw datasets cannot be used for mining the data. You need to evaluate them and turn them into something you can use. By applying the machine learning tools and processes, important information is found and pattern identification is made automatic. Discovering patterns in data sets that are difficult to spot with classic statistical methods [2] by using machine learning techniques.

II. CLASSIFICATION TECHNIQUES IN MACHINE LEARNING

Machine learning is made up of a number of approaches used to find knowledge that is interesting and relevant to a problem. Classification, prediction, and grouping are tasks that are included in data mining. Supervised machine learning and unsupervised machine learning are illustrated in figure 1.

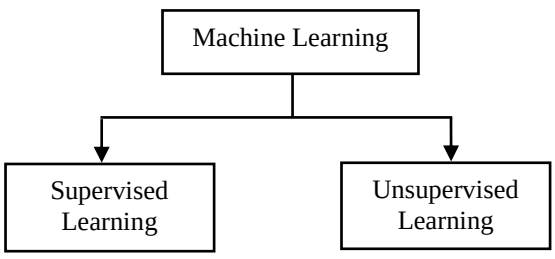

Fig.1. Categories of Machine Learning

Supervised learning is concerned with summarising the general qualities of the data in the database. Unsupervised learning, on the other hand, is utilised to seek for and find patterns that can be used to anticipate future states of the data. Classification is a machine learning approach that categorises collections of items. Classification's main goal is to determine which category applies to each unknown case in the data. Classification techniques are intended for use when there are no previous examples of a particular sample type to aid in classifying an unknown sample. In general, this training set is used to teach the categorization technique. A good example of this is the use of training sets for modifying the parameters of Neural Networks and Support Vector Machine (SVM) in order to tackle a specific classification problem. A supervised learning problem is a classification problem where the output information is a discrete classification, i.e. the potential mutually exclusive classes of the problem are supplied as input attributes and the classification output is assigned one of the classes. Classification is about finding a relationship between the input attributes and the output class. Once the discovery of this relationship is complete, the newly found knowledge will be useful in classifying unknown objects. This classification approach is used to categorise database records based on particular characteristics. Classification is the process of labelling observations to give them a designation that corresponds to one of numerous class labels. Mapping attributes to specific groupings is provided by classification. Once the data has been categorised, it is possible to summarise the characteristics of these groupings.

The following classification approaches are possible:

- Support Vector Machines
- Bayesian Classification
- Decision Trees
- Nearest Neighbour
- Regression
- Genetic Algorithms
- Neural Networks

In this paper, we implement CART tree, neural network, Regression, and SVM, and then assess their relative abilities and eventually arrive at the best predictions.

- **Naive Bayes**

Naive Bayes is a simple classification technique that treats data as vectors of feature values, and therefore it is not an algorithm for training classifiers. Diagnosis includes a list of symptoms, X are features in naive Bayes classifiers, and hence the classifier will predict the presence of disease.

- **Support Vector Machine**

Support vector machine (SVM) is designed to identify the most effective classification approach using training data that can differentiate between the two class instances. The optimal boundary in support vector machine approach is known as a hyperplane. Supporting vectors are placed on the hyperplane, located close to it. A separating hyperplane cannot exist unless the space is linearly separable.

- **CART**

The designers of CART's tree regression and classification algorithm, Breiman, Friedman, Olshen, and Stone, devised it in 1984 [4]. This algorithm is called Surrogate Splitting because it follows an operation method that is often used in surrogate approaches. High-speed deployment is possible with Classification and Regression Tree. Using Classification and Regression Tree, critical links between data points can be found that other analytic methods may not uncover rapidly. Regression trees and classification trees are also referred to as CART. Multivariate decision binary trees are built using Classification and Regression Tree. Every time a CART algorithm is run, the instructional records are divided into two subsets to facilitate subsequent processing. Until the conditions of a stop are set, these divisions will continue. When the impurity parameter value is assigned in CART, it is decided which is the best breaking point or assignment of the value. If the split that yields the greatest advantage is found to result in a level of impurity below the set limit, then that split will not be made. The degree of similarity between target field value and records at a node is referred to as impurity in this visualisation. A node with 100% of the samples placed in one category is known as a "pure" node. The fact that a foreteller field is commonly used in multiple levels of decision trees in the CART algorithm

is impressive. Additionally, this technique is designed to accommodate category and continued forms of foreteller and target fields.

- **C4.5**

The C4.5 method is used to create a decision tree with Classification techniques. C4.5 is an improved version of ID.3. C4.5 generates decision trees by using the concept of information entropy. The C4.5 process uses a post-pruning methodology. The data gain is normalised from the criteria that splits the data. C4.5 constructs an initial tree using a divide and conquer algorithm. This approach performs effectively with data with little to no noise.

- **Regression**

To predict for future samples, we will do regression, which is what statistical methods were originally created for [5]. Linear regression is broken down into two divisions: linear and logistic.

III.LITERATURE REVIEW

This section deals various research works in the area of machine learning classification, based on previous literatures and articles.

In[6], an application for spatial distribution of Blood Bank Information System blood donors is being implemented in order to assist patients in their access to blood from anywhere.

The regional variance in blood donor correlates has been investigated in [7] at rates-a rate research in Canadian urban areas. Canadian blood services, i.e. an organisation to collect and distribute the blood supply around the country, are involved in this activity. In order to analysis service accessibility, accessibility factors are introduced and calculated with a floating 2-step catchment area. This study also uses regression method.

The study article on the administration of blood bank services is being conducted in [8] entitled the "Application of CART Algorithms in blood donor classification." This research uses one of the most prevalent technology for data mining, i.e. classification, and a model determining the conduct of donors is produced using a CART application.

Machine learning tool is utilised in [9] to extract information from PPI systems and developed a PPI search system, called PP Look, which is an excellent tool for 4-tier information extraction, based on a full sentence parsing technique. In this study, researchers proposed a valuable tool, PP Look, which utilises an improved keyword to extract protein-protein interaction data from biomedical literature, the dictionary pattern matching an algorithm. Some visual methods were used as 3D stereoscopic views to conclude PPI.

In [10], a comparison research of the machine learning classification methods in the diagnosis of breast cancer, for example for binary classifiers for health-care datasets is undertaken. This work serves to reveal useful information and also contributes to the improvement of health care services by decision makers. In this study, the experiments provided give doctors and healthcare professionals with a tool which helps them to gain insight into large clinical databases.

In [11], the research paper on 'Blood Transfusion Data Set Interactive Knowledge Discovery' is underway with machine learning experiments that allow health professionals to better manage the Blood Bank facilities.

In [12] the research effort 'Rule Extraction for Blood Donors,' which is named 'Fuzzy Sequential Pattern Mining,' is utilised to extract rules of the Blood Transfusion Data Centre, which predict the future behaviour of the donor.

In the research publication in [13] entitled a comparison of models of Blood donor data collecting that employ decision tree to analyse the classification of Blood donors. In this work the prolonged RVD-based model is compared with DB2K7. In this work it has been observed that in terms of reminder and precision, the RVD classification is superior than DB2K7.

Machine learning techniques are used to examine/analysis the classification of donors and encourage them in developing real-time blood donor management using Dash boards with a blood profile or RVD profile and geo-location data in [14] in the research article Real Time Donors Management with Dash boards based on Data Model[21].

The healthcare applications for the diagnosis of different disorders using the Red Blood Cell count are researched in [15]. The researchers introduced RBC's automatic picture count procedure and utilised certain technologies such as segmentation, equalisation and K-means grouping for image

preprocessing. Certain diseases linked to RBC count disease are also anticipated. In addition, the processing and decision tree leverages the RBC automatic recognition and counting to categorise RBC.

IV. PROPOSED METHODOLOGY

Different classification techniques are used to find the best algorithm for blood donor's dataset based on the 10 fold Cross validation. The flow diagram of the complete proposed work is shown in figure 2.

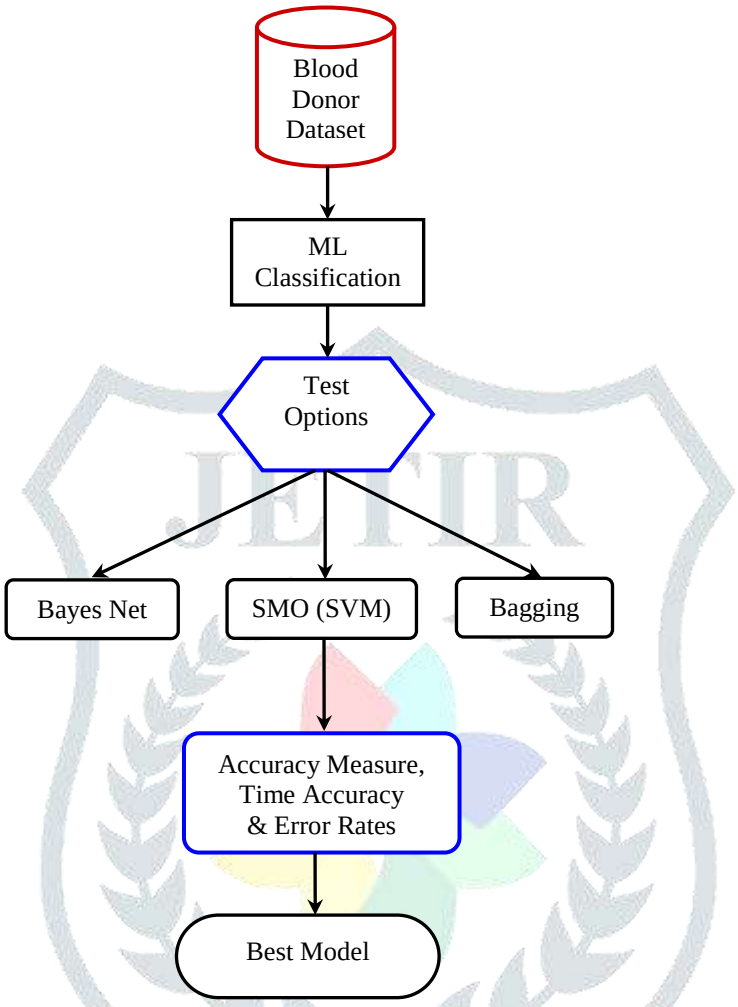

Fig.2. Flow Diagram of Porposed Model

Classification is one of the approaches used in machine learning to tackle problems such as algorithm selection, training division, data testing, and model comparison. The first step is to create a classification model using training data. Pre-classification is required for each data object. After that, the model created in the previous stage is put to the test by labelling data objects in a test dataset with class labels. As a result, the test data is not the same as the training data. As a result, the classification model's accuracy is measured by comparing genuine class labels in the testing set to those assigned by the model [16].

The three distinct classification techniques are employed in this part to identify a superior algorithm for the blood donor dataset. The following are the methods that will be discussed:

1) Bayes Net

Under the conditions of nominal characteristics and no missing values, the Bayes Net learns Bayesian networks. A Bayesian network is a framework that depicts the conditional dependencies between domain variables and can also be used to visualise the probabilistic causal links between domain variables graphically. A directed acyclic graph plus probability tables make up a Bayesian network. The domain variables are represented by the network's nodes, and an arc between two nodes denotes the presence of a causal relationship between them. There is a probability table associated with each node. Although domain variables can be continuous, they are usually discretized for ease of use and efficiency. A Bayesian network is used to infer the probability of a variable given the observation of

other variables, in addition to illustrating the dependencies between domain variables. The Bayesian networks are implemented using WEKA[17].

2) Sequential Minimal Optimization

The Sequential Minimal Optimization (SMO) class implements this type of classifier learning. SMO is a new training algorithm for Support Vector Machines (SVM). Sequential minimum optimization, one of the fastest approaches for learning support vector machinery, is sluggish to converge to a solution, especially when the data is not linearly separable in the space spanned by the nonlinear mapping. Because of the noise in the data, this happens frequently. The values given to two parameters are crucial for both run time and accuracy: the upper constraint on the coefficients values in the equation for the hyperplane, and the degree of the values in the non-linear mapping, both of which are default values set to one. Only through trial and error can the best configuration for a given dataset be discovered [18].

3) Bagging

Bootstrapping and aggregation are combined in the Bagging approach. It is an approach for data categorization and forecasting that is both broad and ensemble. In order to improve classification accuracy in comparison to producing a final output class, an upgraded complicated classifier is built. Cross-validation by discussion sample of examples with replacement qualifies all individual classifiers by a factor of ten. Each sample size and 10 folds Cross-validation are the same [19].

V. EXPERIMENTAL EVALUATION

We have used the popular, open-source machine learning tool Weka (version 3.6.6) for this experimental analysis. The blood bank dataset is used and the performances of a comprehensive set of machine learning classification algorithms (classifiers) are analyzed.

A. Dataset Description

Dataset used in this work is more precise and accurate in order to improve the predictive accuracy of machine learning algorithms. The dataset used in this work has been collected from a Blood Bank of Kota city, the dataset is available in a common separated value (CSV) format. The dataset has **8** attributes and **3010** records. This dataset contains following attributes which are shown in table-1.

TABLE -1 : ATTRIBUTES OF BLOOD DONORS DATASET

S.No.	Attribute	Description
1	BagID	Number of the bags
2	Age	Age of the donors (numeric)
3	Sex	1=Male, 0=Female
4	VBD	Voluntary/Replacement
5	BG	Blood group of the donor
6	Avail	Availability of blood group (Yes/No)
7	Test	Blood is tested Okay or not (Yes/No)
8	Region	Rural/Urban/Town

WEKA implements algorithms for data pre-processing, feature reduction, classification such as Naïve Bayes, Bayesian Network, J48. The performance of the machine learning algorithms for blood donor's data set is analyzed using visualization tools.

B. Implementation of ML Algorithms

WEKA is open source software machine learning tool that implements a large collection of machine leaning algorithms and widely used in machine learning applications [20]. From the above data, csv, arff file have been created. This file was loaded into WEKA explorer. The classify panel enables the user to apply classification algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions or the model itself. There are 16 decision tree algorithms like ID3, J48, ADT etc. implemented in WEKA. The algorithm used for classification is ID3, C4.5 and CART. Under the "Test options", the 10-fold cross-validation is selected as our evaluation approach. Since there is no separate evaluation data set, this is necessary to get a reasonable idea of accuracy of the generated model. The model is generated in the form of decision tree.

C. Implementation in WEKA

Classification is a statistical technique that helps to classify any new client into one of the existing groups. It will create a model on the blood bank data available. And then classifies the new data based on the model that is developed using the test data. Out of the various machine learning techniques provided by the WEKA, classification, clustering, feature selection, data pre-processing, regression and visualization, in this section, the implementation of classification techniques is shown. Classification creates a model based on which a new instance can be classified into the existing classes or determined classes.

For example by creating a decision tree based on past data of blood donors we can determine how likely is a person to donate the blood, his attribute like: bagid, age, sex, vbd, bg, avail, test, region etc. The aim of this paper is to create a decision tree using WEKA, so that we can classify new or unknown donor's samples. The algorithm we are going to implement to classify is WEKA's J4.8 decision tree learner[20].

The steps of classification algorithm in WEKA are as follows:

Step-1:

Create a data file in the format csv. WEKA understands these two formats. We are using data file in csv format i.e. BloodBank.csv

Step-2:

Open the WEKA GUI Interface that is shown in fig.3.

Fig.3. WEKA GUI Interface

Follow the underlying steps to classify:

- Choose explorer in WEKA. Then open the data file BloodBank.csv.
- Go to classify tab.
- Click "choose" and choose J48 algorithm under trees section.

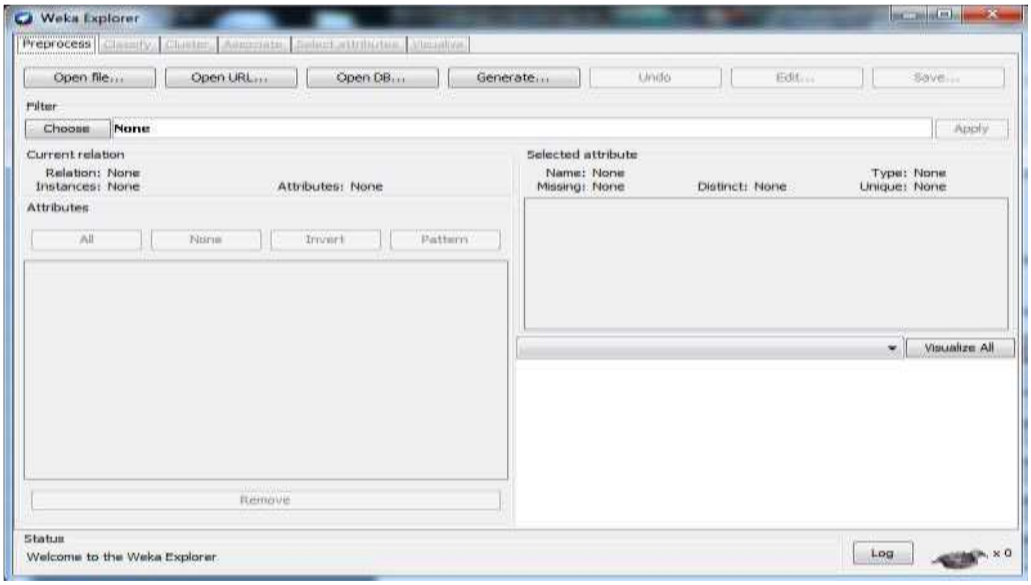

Fig.4: WEKA Explorer Window

Step-3: Loading data into WEKA.
To do that click on the open file button and browse for the bank.csv file. Then it shows all the attributes as shown in the figure below.

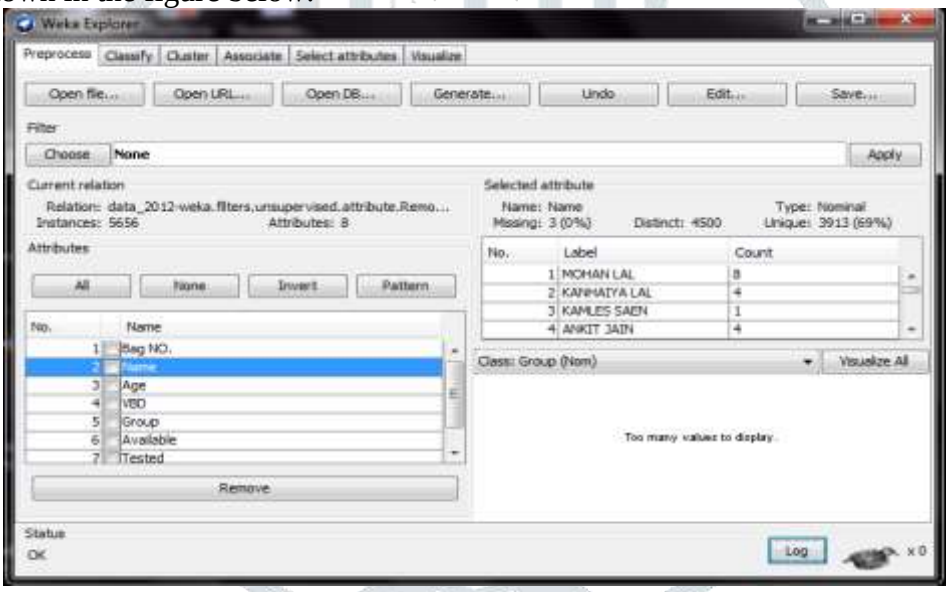

Fig.5: Loading Dataset in WEKA

D. Tools Used
Weka, Tanagra, Orange

E. Parameters Used
Two Accurate rate, Error rate

F. Results Analysis

In this section, we have applied the different machine learning techniques through different three data mining tools on the educational datasets of blood donors and results are shown in the different tables below.

Table-1: Results obtained by Weka

Techniques	Accurate rate	Error rate
MLP	70.18	25.12
NB	69.19	27.14
Logistic	75.20	23.21
J48	76.12	22.20
BN	74.89	24.10

Table-2: Results obtained by Tanagra

Techniques	Accurate rate	Error rate
MLP	69.66	24.13
NB	74.51	23.10
Logistic	75.31	25.12
J48	76.12	22.20
BN	76.45	26.41

Table-3: Results obtained by Orange

Techniques	Accurate rate	Error rate
MLP	69.23	22.12
NB	77.21	23.25
Logistic	78.26	26.14
J48	76.12	22.23
BN	75.12	27.31

The comparison of results of different machine learning classification techniques using three tools for best suitable classification is as follow:

Table-3: Performance comparison of machine learning classification tools

Tools	Techniques	Accuracy rate	Error rate
Weka	SVM	88.22	11.78
Tanagra	C4.5	79.20	21.25
Organge	MLP	78.26	26.14

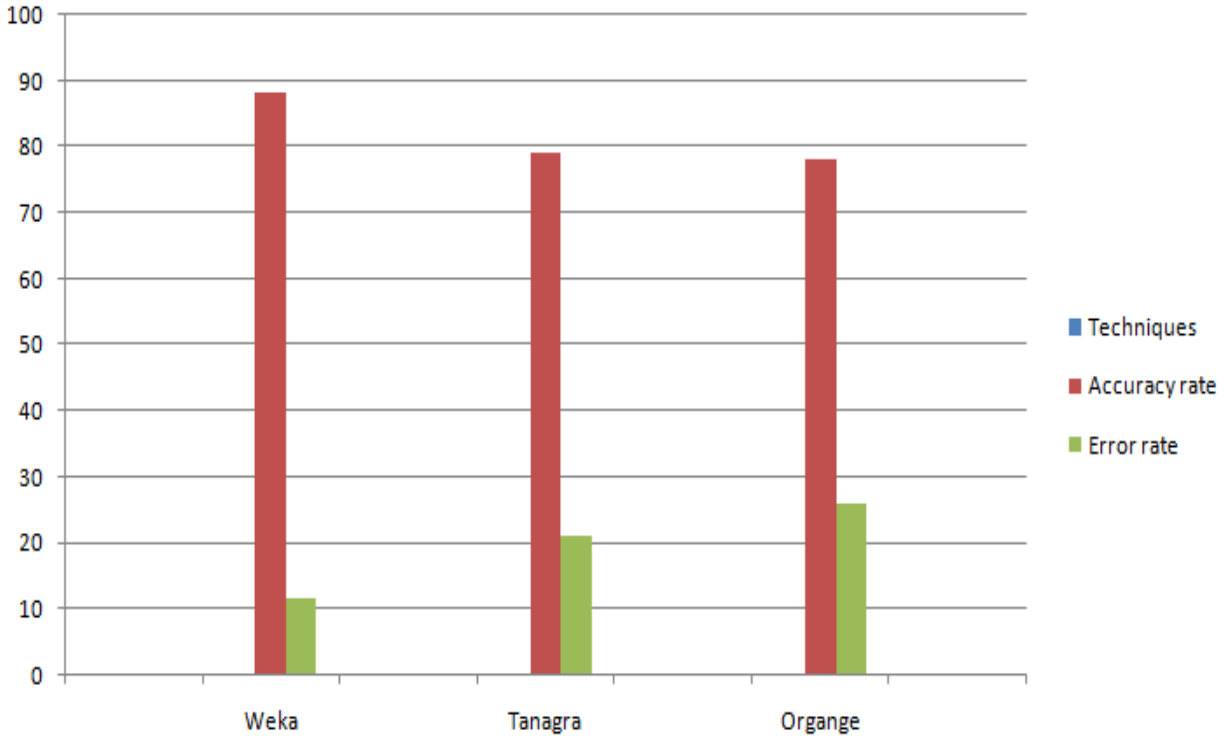

Fig.6: Performance Comparison of Machine Learning Tools & Techniques

VI. CONCLUSION

The machine learning classification techniques which are to be tested such as: Naïve Bayes (NB), SVM, CART, MLP, ID3, KNN, and Decision Trees. The training performance of these techniques has measured according to their accuracy. This research uses three machine learning tools, Weka, Tanagra, and Orange open source machine learning toolkit to apply several machine learning classification approaches on blood donor datasets. This research study seeks to utilise the machine learning technology practically in difficulties associated with blood banks. It allows users, who wait for the last drop in the blood and are about to death, to collect and donate blood. Blood Bank Centre's data were tested. The objective of this study is to establish a machine learning model for extracting knowledge from the classification of blood donors in order to assist clinical choices in blood banks. It employed real-world data from an EDP department of the blood bank centre and the J48 algorithm to develop a classification rule base that can help the blood bank owner make correct judgments more quickly and accurately. Such strategies assist blood bank organisations in decision support. As a scope of future work, the deep learning techniques may be applied on different datasets along with different parameters.

REFERENCES

[1] Ramachandran P, Girija N, Bhuvaneswari T. Classifying blood donors using machine learning techniques. IJCSET; 1 (1): 10-13, 2011.

[2] Ashoori M, NajiMoghaddam V, Alizadeh S, Safi M. Classification and clustering algorithm application for prediction of tablet numbers: case study diabetes Disease. Health Information Management1392; 10(5): 739-749. [In Persian] ,View 2 Jan 2015.
[3] Arvind Sharma, P.C. Gupta, "Predicting the Number of Blood Donors through their Age and Blood Group by using Machine learning Tool," International Journal of Communication and Computer Technologies(ICCT), Vol.1–No.6, Issue: 02 September 2012.
[4] Breiman, Leo, Jerome Friedman, R. Olshen and C. Stone, Classification and Regression Trees. Belmont, California: Wadsworth, 1984.
[5] Andrew W. Moore. "Regression and Classification with Neural Networks". School of Computer Science Carnegie Mellon University, 2001.
[6] 6. B.G. Premasudha et al., An Application to find spatial distribution of Blood Donors from Blood Bank Information" July-December 2009, Vol. II, Issue No.2.
[7] T. Santhanm and Shyam Sunderam: "Application of Cart Algorithm in Blood donor's classification"; Journal of Computer Science"; Journal of computer Science Vol. 6, Issue 5, 2009.
[8] Zhangetal. "PPLook: an automated machine learning tool for protein–protein interaction, BMC Bio-informatics; 2010.
[9] Dr. Varun Kumar, Luxmiverma; "Binary classifiers for Health Care databases"-A comparative study of machine learning classification algorithms in the diagnosis of Breast cancer, IJCST, Vol. I, Issue 2, December 2010.
[10] Vikram Singh and Sapna Nagpal; "Interactive Knowledge discovery in Blood Transfusion Data Set"; VSRD International Journal of Computer Science and Information Technology; Vol. I, Issue 8, 2011.
[11] Wen-ChanLee and Bor-Wen Cheng; "An Intelligent system for improving performance of blood donation" Journal of Quality, Vol. 18, Issue No. II, 2011.
[12] Shyam Sundaram and Santhanam T: "A comparison of Blood donor classification machine learning models", Journal of Theoretical and Applied Information Technology, Vol.30, No.2, 31 August 2011.
[13] Shyam Sundaram and Santhanam T; "Real Time Blood donor management using Dash boards based on Machine learning models" International Journal of Computing issues, Vol.8, Issue5, No.2, September 2011.
[14] Prof. Dr. P.K. Srimani et al. "Outlier machine learning in medical databases by using statistical methods" Vol.4, No.1, January 2012.
[15] Ivana D. Radojevic et al. "Total coliforms and machine learning as a tool in water quality monitoring", African journal of Microbiology Research, Vol. 6(10), 16 March 2012.
[16] Nikita Bhatt, Amit Thakkar, Amit Ganatra, "A Survey & Current Research Challenges in Meta Learning Approaches based on Dataset Characteristics", International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-1, March 2012.
[17] Rashedur M. Rahman, Farhana Afroz, "Comparison of Various Classification Techniques Using Different Machine learning Tools for Diabetes Diagnosis", Journal of Software Engineering and Applications, 2013, 6, 85-97.
[18] Eibe Frank, Ian H. Witten," WEKA Machine Learning Algorithms in Java", Morgan Kaufmann Publishers.
[19] Esteban Alfaro, Matias Gamez, Noelia Garcia, "adabag: An R Package for Classification with Boosting and Bagging", Journal of Statistical Software, August 2013, Vol. 54, Issue 2.
[20] Arvind Sharma, P.C. Gupta, Predicting the Number of Blood Donors through their Age and Blood Group by using Machine learning Tool," International Journal of Communication and Computer Technologies, Vol.01, No.6, Issue: 02 September 2012.