Comparative Analysis of Different Credit Card Fraud Detection Datasets using Data Mining Tools and Techniques.

[1]Sheetal, [2]Dr.K.L.Bansal,

[1]M.Tech Student, [2]Professor,
[1]Department of Computer Science,
[1]Himachal Pradesh University, Shimla, India

Abstract: With growing development in technology and improvement in communication channels, there are a large number of online transactions that take place each and every day which is paid by credit card is targeted by fraudulent activities. Such activities could be identified; automatically they would be very helpful in fraud detection. This classification technique is used to classify data of different kinds which predicts the class labels for the new data. In this paper, a comparative analysis of these classification techniques such as J48 which is a type of decision tree classifier and naïve Bayes classifier is used. Three open source data mining tools: Orange, Weka, and Rapid miner are used to predict and classify the customer's credit card dataset as Good/Bad to know which technique and tool work better in predicting fraud values with the highest accuracy. Parameters to be used are accuracy and error rate. A credit card fraud detection data set is sourced from the kaggle.com, and OpenMl. The result shows that the Orange tool with the Naïve Bayes algorithm shows the highest accuracy and lowest error rate for all three datasets.

Keywords **–Data mining, classification, credit card fraud datasets, data mining tools.**

I. INTRODUCTION

In this era of digital age and with the improvement in computer technology, many organizations usually gather large volumes of data from operational activities and after which are left to waste in data repositories. Data mining is all about the analysis of that large amount of data usually found in data repositories in many organizations. Data mining is the process of extracting unknown and predictive information from a large amount of data by Knowledge discovery in databases (KDD) process. The Knowledge Discovery in databases (KDD) process in the data mining methods is used for extracting patterns from data. In this step of the KDD process, various methods are applied to extract data patterns. Data mining can handle different kinds of data ranging from ordinary text and numeric data to image and voice data. Analysis and prediction are also a part of the data mining process which is used to extract models with different data classes and to predict future models using extracted models by analyzing it. With the help of data mining, such data can now be mined using different mining methods such as clustering, classification, association, and detection methods in order to unravel hidden information that can help in the improved decision-making process. In this paper, data mining classification techniques like a decision tree and naïve Bayes classifier are used to extract the data Patterns on various credit card fraud detection datasets using Rapid Miner, Weka and Orange tool. The Performance is analyzed based on the parameters such as Accuracy and Error Rate. The dataset includes Default Payment of credit card client in Taiwan of 2005, German credit dataset and Abstract data set for credit card fraud detection

Section I of this paper presents an introduction, Section ii presents a literature survey, section iii objectives, and problem statement, Section IV presents a methodology, Section v results and discussion and Section VI presents conclusion and future scope.

II. LITERATURE SURVEY

Priyanka kumara and smita prava Mishra [2019] presented "Analysis of credit card fraud detection using fusion classifiers". In this paper they analyzed some ensemble classifiers such as Bagging, Random forest, classification via Regression, voting and compared them with some effective single classifiers like K-NN, naïve Bayes, SVM, RBF classifiers, MLP, Decision Tree. The evaluation of these algorithms is carried out through three different datasets and treated with SMOTE, to deal with the class imbalance problem. The comparison is based on some evaluation metrics like accuracy, precision, true positive, true positive rate or recall, and false positive rate. They conclude that there is no single classifier in data mining that can perform better than the ensemble classifiers. The classification via Regression ensemble classification technique performs well on both German Data with accuracy of 95.21% and Australian data with accuracy of 91.17% [1].

Sahil Dhan khad, Emad A.Mohammed[2018] presented "Supervised Machine Learning algorithms for credit card fraudulent transaction detection: A comparative study. In this paper, they apply many supervised machine learning algorithms to detect credit card fraudulent transactions using a real-world dataset. Furthermore, they employ these algorithms to implement a super classifier using ensemble learning methods. They identify the most important variables that may lead to higher accuracy in credit card fraudulent transaction detection. Additionally, they compare and discuss the performance of various supervised machine learning algorithms that exist in literature against the super classifier that they implement in this paper. Overall results show that stacking classifier which is used LR as meta classifier is most promising for predicting fraud transaction in the dataset, followed by the random forest and XGB classifier [2].

Shiyang Xuan et.al[2018] presented " Random Forest for credit card fraud detection". In this paper, two kinds of random forests are used to train the behavior features of normal and abnormal transactions. They make a comparison of the two random forests which are different in their base classifiers, and analyze their performance on credit card fraud detection. The data used in experiments come from an e-commerce company in china. A real –life B2C dataset on credit card transactions is used. The algorithm of random forest itself should be improved. For example, the voting mechanism assumes that each o base classifiers have equal weight, but some of them may be more important than others. Therefore, they also try to make some improvement for this algorithm[3].

Guan Jun Liu et.al [2018] presented "A new credit card fraud detecting method Based on Behavior certificate". In this paper they propose a new credit card fraud detection system (FDS) based on behavior certificate (BC) which reflects card holder's transaction habits. In this method, the correlation between behavior features and some special cares such as festival and we kind are considered into BC. First, they extract a set of behavior features from each card holder's transaction records. Then they construct her/his BC based on these behavior features, finally, they compute the risk degree for each card holder's incoming transaction based on her /his BC. If the degree is higher than a threshold, it is considered as a fraud. Result shows that accuracy of their method is above 90 percent over various input datasets with different fraud rate. Comparative experiments reveal that their method is better than the FDS with support vector machines[4].

Anita Jog, Anjali A. Chandvale[2018] presented "implementation of credit card Fraud Detection System with Concept drifts adaptation". In this paper, the developed algorithm detects credit card fraud. Prediction of any algorithm is based on certain attribute like customer's buying behavior, a network of merchants that customer usually deals with, the location of the transaction, amount of transaction,etc.But these attribute changes over time. So, the algorithm model needs to be updated periodically to reduce this kind of errors. The proposed and developed system filters 80% fraudulent transactions and acts as a support system for the society at a large[5].

Archana Gahlaut et.al [2017] presented "Prediction analysis of risky credit using Data mining classification models'. In this paper, they look whether data mining techniques are useful to predict and classify the customer's credit score (good/bad) to overcome the future risks giving loans to clients who cannot repay. They used Decision Tree; support Vector Machine, Adaptive Boosting model, Linear Regression, Random Forest and neural Network are used to build predictive models.Result found that the best algorithm for risky credit classification is Random Forest algorithm [6].

III. Problem statement and objectives

3.1 Problem statement:

In today's world, most people use credit cards and debit cards for executing online banking transactions rather than gaining to the bank. This is due to easily available modern resources like laptops, phones, and tablets. Another reason for online credit and debit card transactions is gaining the popularity of online shopping trends. Online shopping allows customers to view more items in a short time on a single screen and can also compare the prices of the same item from different vendors. During online transactions, personal data like account number, password, date of birth, credit card details, etc. are revealed over the network. This results in financial fraud. Due to lack of proper knowledge or unawareness before using the digital card in online transactions. Card fraud begins either with the theft of the physical card or people may save their card details on fraud websites during the online transaction or sometimes people share their card details with unauthorized persons which result in financial fraud. The most important aspect of fraud detection is to correctly identify fraudulent activity during the transactions. Since the fraudulent transaction are very few as compared to the legitimate transactions. The detection of fraud is a difficult task and there is no such ideal system that accurately predicts fraudulent transactions. The e-commerce system is used by both the authorized as well as unauthorized person and there is no such system that identifies the difference between them.

3.2 Objective:

The main objective of our study is to know which technique and tool work better in predicting fraud values with the highest accuracy on different credit card fraud dataset.

3.2.1 Study of following classification techniques in data mining:

- Decision Tree(J48)
- Naïve Bayes

3.2.2 Comparative analysis of classification techniques on the basis of following parameters:

- Accuracy
- Error Rate

3.2.3 Study of following data mining tools:

- Rapid Miner
- Weka
- Orange

IV. METHODOLOGY:

The methodology of our study consists of three preparatory steps:
4.1 The selection of classification algorithm to evaluate.
4.2 The selection of Data Mining tools to test.
4.3 The selection of parameters.

4.1 Classification:

Classification is the most common data mining technique. Classification is the act of looking for a model that describes a class label in such a way that such a model can be used to predict an unknown class label. So, classification is usually used to predict an unknown class label. This Paper works on two of the methods i.e. Naïve Bayes and Decision tree.

4.1.1 Naïve Bayes

The Naïve Bayes classifier technique is based upon the Bayesian theorem and practically used when the dimensionality of the input is high. Naïve means "simple" and this classification method is based on Bayes rule. It is a simple probabilistic classifier. The Bayesian classifier is capable of calculating the most possible output based on input. Naïve Bayes is a supervised learning method used in large datasets and in complex situations. It is well scalable and based on occurrence data. A Naïve Bayes classifier assumes that the presence or absence of any particular feature or attribute is unrelated to the presence or absence of any other feature or attribute when the class variable is given e.g. a fruit may be considered as Orange if it is orange in color and has a round shape. Even if these features depend upon each other or on the existence of other features of the class, a naïve Bayes classifier considers all these properties as independently contribute to the probability that the fruit is orange.

Algorithm work as follow:
There are two types of probability:
•Posterior probability [P(c/x)]
•Prior probability [p(x)]
Bayes theorem provides a way to calculate a posterior probability p(c/x) from p(c), p(x) and p(x/c). This algorithm considers the effect of the value of predictor (x) on the value of other predictors.
According to Bayes theorem:

$$P(c/x) = \frac{(P(x/c)*P(c))}{P(x)}$$

•P(c/x) is the posterior probability of class (target) of the given predictor (attribute) class.
•P(c) is known as a class prior to probability
•P(x) is known as predictor prior probability
•P(x/c) is the likelihood, which is the probability of predictor of a given class [7].

4.1.2 Decision Tree:

The decision tree is the most powerful and popular method for classification and prediction. A decision tree is a tree-like structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. The topmost node in the tree is called the root node. When a tuple T, is given for which the related class label is not known, the attribute class label is not known, and the attribute values of the tuple are tested alongside the decision tree. A path is outlined starting from the root node to a leaf node that holds the prediction of the handy because the construction of a Decision tree classifier does not involve any prior domain knowledge. It can efficiently handle high dimensional data [7].

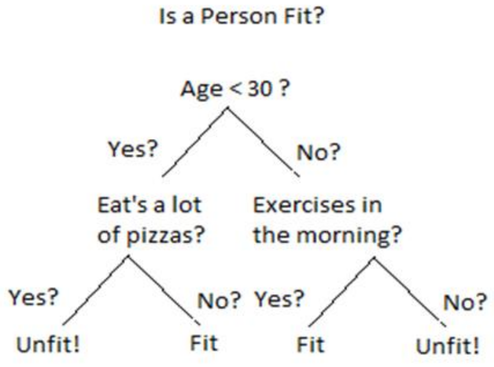

Fig1: A simple Decision Tree [9]

4.2 Tools:

The selection of the tools to test was done according to the three best data mining tools. All have a graphical user interface (GUI). We selected to study only those that an analyst is able to use.

4.2.1 Weka

In the presented paper, the experimentation has been done by using the Weka tool version 3.8.3 version. Weka is the abbreviation of the Waikato environment of knowledge analysis. It is an open source and reliable tool for data mining techniques. It can be freely downloaded from this website address http://www.cs.waikato.ac.nl[9]. It accepts its data in the Arff file format. It is used for several applications such as classification, clustering, feature selection, regression and association, and standard data mining problems.

4.2.2 Rapid Miner

In this paper Rapid Miner 9.3.0 Version used. Rapid Miner is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, text mining, and predictive analytics. Data mining provides machine learning procedures including data loading and transformation (ETL) data preprocessing and visualization, modeling evaluation, and development. RapidMiner is written in a Java programming language. The Rapid Miner studio free edition, which is limited to 1 logical processor and 10,000 data row, is available under the AGPL license. It can be freely downloaded from this website address https://my.rapidminer.com/nexus/account/index.html#downloads[10]. It accepts its data in CSV file format. It is used for several applications such as classification, clustering feature selection, regression and association, and standard data mining problems.

4.2.3 Orange

Orange is an open source tool. Orange is developed at the Bioinformatics Laboratory at the faculty of computer and information science, the University Of Ljubljana, Slovenia along with open source community. It can be freely downloaded from https://orange.biolab/si/download/.orange[11] is a collection of python based modules that sit over the core library of C++ objects and routines that handles machine learning and data mining algorithms. It is an open source tool. It accepts its data file format in CSV and Arff both. It is used for several data mining techniques such as classification, clustering, regression, and association.

4.3 Parameters:

Classification algorithms are usually assessed using the confusion matrix. Fig.2 illustrates the confusion matrix. The columns are the class prediction, while the rows are the actual class.
TN: denotes the number of correctly classified negative examples (True Negative).
FP: denotes the number of misclassified negative examples predicted as positives (False positive).

FN: denotes the number of positive examples that are misclassified as negative (False Negative).
TP: denotes the number of correctly classified positive examples (True Positive).

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Fig.2 confusion Matrix

4.3.1 Accuracy:

It is the overall performance of the classifiers. It shows a relative number of correctly classified instances or in other words percentage of correctly classified instances.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

4.3.2 Error Rate:

Error rates refer to the frequency of errors occurred, defined as "the ratio of the total number of data units in error to the total number of data unit transmitted." As the error rate increases, the data transmission reliability decreases. [8]

Error rate = 1-accuracy

Or

$$= \frac{FP+FN}{TP+FP+TN+FN}$$

V. RESULTS AND DISCUSSION

5.1 DATASET

The dataset used for analysis Purpose were downloaded from the UCI Machine learning repository(Kaggle.com), and Openml website. We choose three different credit card fraud datasets of the different countries of different attributes having different sizes. Table 1 shows their characterization: the name by which they are known online, the variable data type with nominal and numerical, the number of instances, the number of attributes and the size of the dataset. The ds1 dataset shows result values in 1and 0 form, ds2 dataset shows result class value in good or bad credit form and for ds3 result class values shows in N and Y form. Which shows that N and 0 is a good credit value whereas Y and 1 show bad credit values.

Table 1 Dataset Description

Dataset Name	Abbreviation	Year	Source	Instances	Attribute	size
Default Payments of Credit Card Clients in Taiwan from 2005 (UCI_Credit_card.csv)	DS1	2005	www.Kaggle.com[12]	10,000	24	1190kb
German Credit data (credit-g)	DS2	1994	UCI _1994 www.openml.org[13]	1000	21	148kb
Abstract data set for Credit card fraud detection. (creditcardcsvpresent.csv)	DS3	2018	www.kaggle.com[14]	3075	12	157kb

5.2 Accuracy

Rapid Miner, weak and orange tool are used for the analysis purpose on the basis of evaluation parameters accuracy and error rate for three datasets using two technique Naïve Bayes and decision tree. The results are shown in the graph to observe the performance of algorithms in different datasets with different sizes.

5.2.1 Comparison of accuracy using Naïve Bayes for da1, ds2and ds3

In fig.3 we evaluate the accuracy of different datasets i.e. Ds1, Ds2 and Ds3 on Rapid Miner, weka and orange tool for Naïve Bayes. In the case of Ds1, the orange tool gives more accuracy when we used a Naïve Bayes algorithm as compared to the Rapid Miner tool and the weka tool. In the case of Ds2, the orange tool gives more accuracy when we used a Naïve Bayes algorithm as compared to the Rapid Miner tool and the weka tool. In the case of Ds3, the orange tool and Rapid Miner give more accuracy when we used the Naïve Bayes algorithm as compared to the weka tool. The overall result for the naïve Bayes algorithm shows that the orange tool gives maximum Accuracy for all three datasets. This is 76.36% for Ds1, 75.6% for DS2 and 95.15 for Ds3.

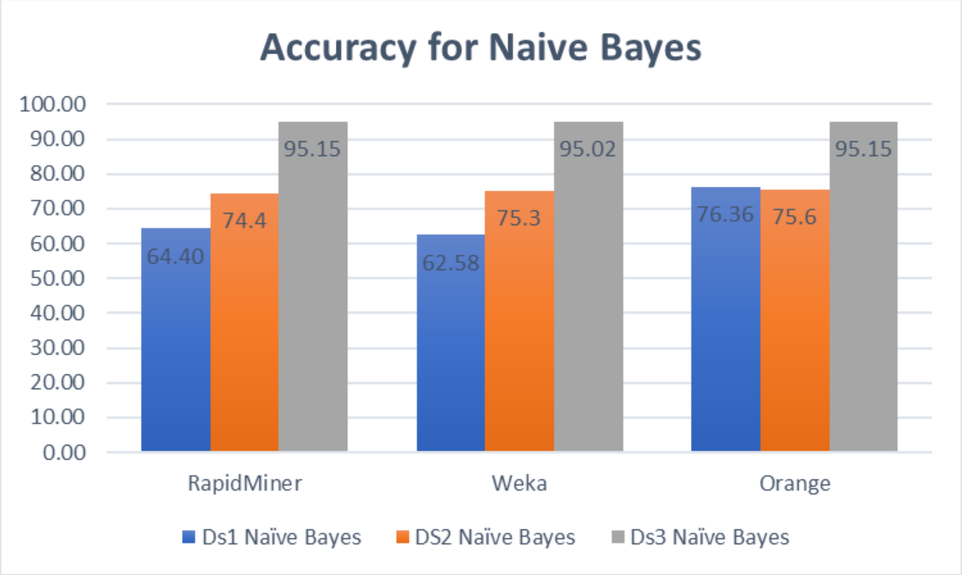

Fig.3 Accuracy for Naïve Bayes using da1, ds2and ds3

5.2.2 Comparison of accuracy using Decision Tree for DS1, DS2and Ds3

In fig.4 we evaluate the accuracy of different datasets i.e. Ds1, Ds2 and Ds3 on Rapid Miner, weka and orange tool for decision Tree. In the case of Ds1, the Rapid Miner tool gives more accuracy when we used the Naïve Bayes algorithm as compared to the orange tool and the weka tool. In the case of Ds2, the weka tool gives more accuracy when we used the Naïve Bayes algorithm as compared to the Rapid Miner tool and the weka orange Tool. In the case of Ds3, the weka tool gives more accuracy when we used the Naïve Bayes algorithm as compared to the Rapid Miner and weka tool. The overall result for the Decision Tree algorithm shows that the weka tool gives maximum Accuracy for Ds2 and ds3 datasets. This is 80.73% for Ds1, 73.3% for DS2 and 98.11 for Ds3.

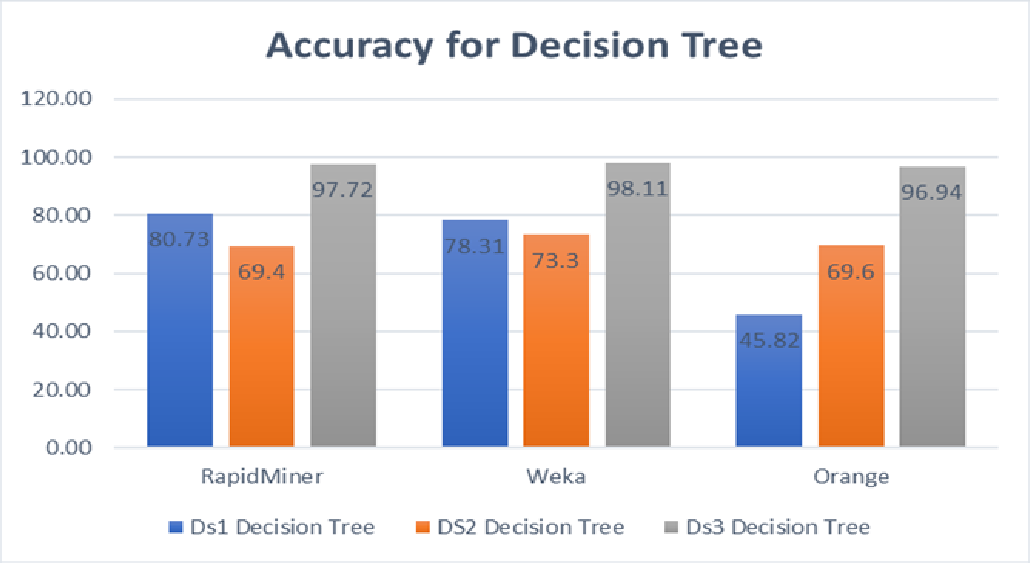

Fig.4 Accuracy for Decision Tree using da1, ds2and ds3

5.3 Error Rate

5.3.1 Comparison of Error Rate using Naïve Bayes for DS1, DS2and DS3

In fig.5 we evaluate the Error Rate of different datasets i.e. Ds1, Ds2, and Ds3 on Rapid Miner, weka and orange tool for Naïve Bayes. In the case of Ds1, the orange tool gives the lowest error rate when we used the Naïve Bayes algorithm as compared to the Rapid Miner tool and weka tool. In the case of Ds2, the orange tool gives less error rate when we used a Naïve Bayes algorithm as compared to the Rapid Miner tool and weka tool. In the case of Ds3, the orange tool and Rapid Miner tool gives less error rate when we used a Naïve Bayes algorithm as compared to the weka tool. The overall result for the naïve Bayes algorithm using dataset ds1, ds2and ds3 by using error rate parameter shows that the orange tool gives the lowest error Rate for all three datasets. Which is 23.64% for Ds1, 24.4% for DS2 and 4.85for Ds3.

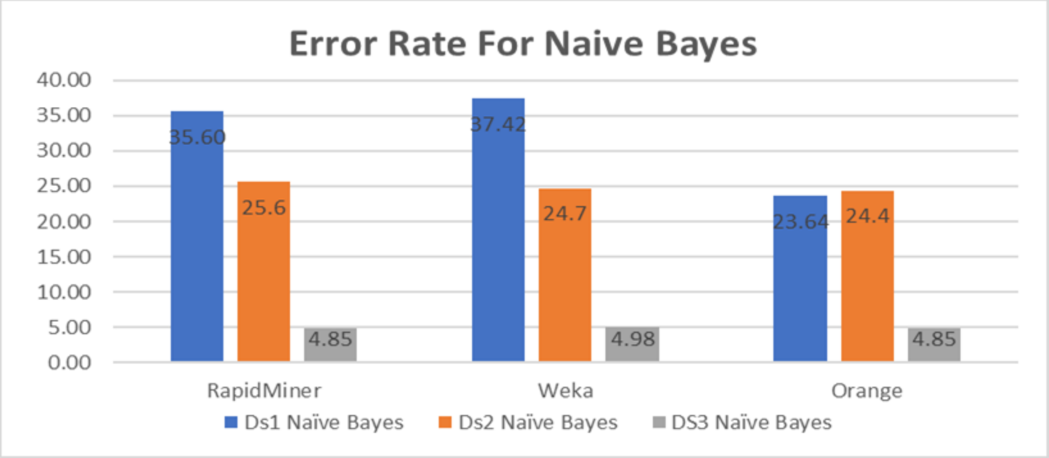

Fig.5 Error Rate for Naïve Bayes using da1, ds2and ds3

5.3.2 Comparison of Error Rate Using Decision Tree for DS1, DS2and Ds3

In fig.6 we evaluate the Error Rate of different datasets i.e. Ds1, Ds2 and Ds3 on Rapid Miner, weka and orange tool for Decision Tree. In the case of Ds1, the Rapid Miner tool gives the lowest error rate when we used the Decision Tree algorithm as compared to the orange and weka tool. In the case of Ds2, the weka tool gives less error rate when we used the Decision Tree algorithm as compared to the Rapid Miner tool and orange tool. In the case of Ds3, the Weka tool gives less error rate when we used the Decision Tree algorithm as compared to the Rapid Miner and orange tool. The overall result for the Decision Tree algorithm by using dataset ds1, ds2and ds3 on the basis of the error rate parameter shows that the Weka tool gives the lowest error Rate for two datasets. Which is 26.7% for Ds2, 1.89% for DS3 and for ds1 Rapid Miner tool gives error rate 19.27%.

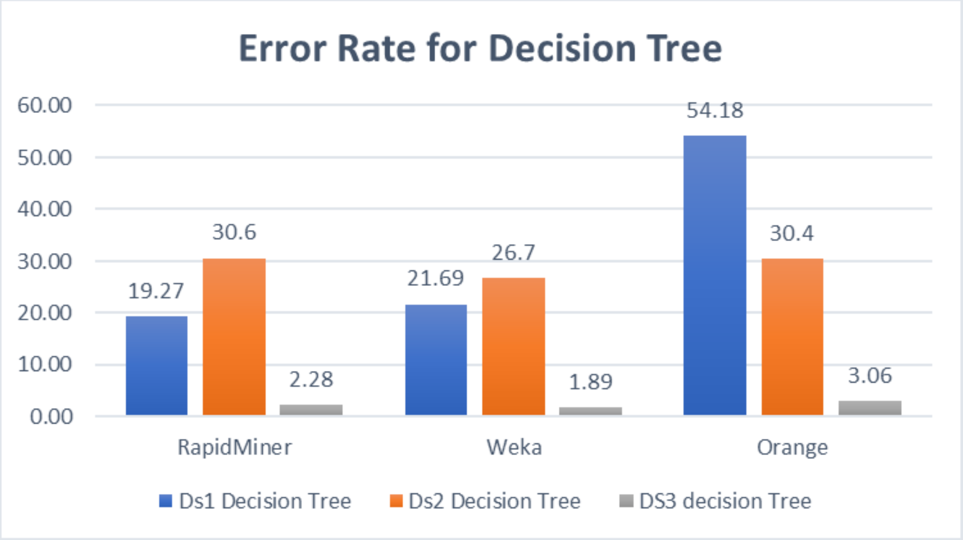

Fig.6 Error Rate for Naïve Bayes using da1, ds2and ds3

VI. CONCLUSION AND FUTURE SCOPE

From the above result it has been concluded that for the naïve Bayes algorithm using the accuracy parameter for ds1, ds2 and ds3 orange tool gives the highest accuracy with 76.36%, 75.6%, and 95.15%. For the Decision Tree algorithm using accuracy, the parameter shows that the weka tool gives maximum Accuracy for Ds2 and ds3 datasets, which is 73.3% for DS2 and 98.11 for Ds3. In the case of Error Rate parameter for naïve Bayes algorithm using dataset ds1, ds2and ds3 show that the orange tool gives the lowest error Rate for all three datasets. This is 23.64% for Ds1, 24.4% for DS2 and 4.85 for Ds3. Error Rate for Decision Tree algorithm by using dataset ds1, ds2and ds3 shows that the Weka tool gives the lowest error Rate for two datasets. Which is 26.7% for Ds2, 1.89% for DS3.From the above overall result, it shows that the orange tool with the naïve Bayes algorithm shows the highest accuracy and lowest error rate for all three datasets. Form the above overall result it is concluded that for our research orange tool with naïve Bayes algorithm works better for credit card fraud detection.

For future scope, the same algorithms can be run on different datasets and other data mining tools can be used instead of Rapid Miner, orange and weak tools to analyze the accuracy of algorithms on different datasets.

VII. REFERENCES

[1] Kumari, Priyanka, and Smita Prava Mishra. "Analysis of Credit Card Fraud Detection Using Fusion Classifiers." Computational Intelligence in Data Mining. Springer, Singapore, 2019. 111-122.

[2] Dhankhad, Sahil, Emad Mohammed, and Behrouz Far. "Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study." 2018 IEEE International Conference on Information Reuse and Integration (IRI). IEEE, 2018.

[3] Xuan, Shiyang, et al. "Random forest for credit card fraud detection." 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). IEEE, 2018.

[4] Zheng, Lutao, and et al. "A new credit card fraud detecting method based on behavior certificate." 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). IEEE, 2018.

Weblink

[5] Jog, Anita, and Anjali A. Chandavale. "Implementation of Credit Card Fraud Detection System with Concept Drifts Adaptation." Intelligent Computing and Information and Communication. Springer, Singapore, 2018. 467-477.

[6] Gahlaut, Archana, and Prince Kumar Singh. "Prediction analysis of risky credit using Data mining classification models." 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2017.

[7] Lakshmi Haritha Medida."A Comparative Analysis of Datasets Classification Using Machine Learning Techniques "International Journals of Innovations & advancements in computer science, 2018, 518-525

[8] https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwiswZ-jj-

[9] http://www.cs.waikato.ac.nl

[10] https://my.rapidminer.com/nexus/account/index.html#downloads

[11] https://orange.biolab/si/download/.orange

[12] https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

[13] https://www.openml.org/d/31

[14] https://www.kaggle.com/shubhamjoshi2130of/abstract-data-set-for-credit-card-fraud-detection