How to Cite:
Arun, G. K., & Rajesh, P. (2022). Analysis and prediction for credit card fraud detection dataset using data mining approaches. *International Journal of Health Sciences*, *6*(S5), 4155–4173.
https://doi.org/10.53730/ijhs.v6nS5.9534

Analysis and prediction for credit card fraud detection dataset using data mining approaches

G. K. Arun
Department of Computer and Information Science, Annamalai University, Annamalai Nagar – 608 002, Tamil Nadu, India.
Email: arunnura2370@gmail.com

P. Rajesh
PG Department of Computer Science, Government Arts College, Chidambaram – 608 102 (Deputed from Department of Computer and Information Science, Annamalai University, Annamalai Nagar), Tamil Nadu, India
Corresponding Author: rajeshdatamining@gmail.com

***Abstract*---**Analysis of fraud detection is one of the recent research areas in data analytics. Credit card fraud is rapidly increasing in different ways significantly with modern technology. The Credit card consumers and processing financial company have losing large amount annually. The fraudsters rapidly and continuously try to find and follow novel technologies to commit the illegal actions against consumers. In COVID-19 pandemic situations online shopping trends growing exponentially and parallelly increase Credit card fraud rapidly. In this paper we trained different data mining techniques with statistical approaches for credit card fraud detection. Numerical illustrations to prove the proposed results.

***Keywords*---**Credit Card, Data Mining, Regression Model, Normalization, and Correlation Coefficient.

I. Introduction

Data mining is the discovery of information or models for pre-existing data. The corresponding models made up of one of several things. In data analytics research, the statisticians were the first to use the term “Data Mining”. The term data mining was a critical term mentioning to challenges to retrieve hidden information that was not stayed by the data. The statistical based researchers view data mining concepts as the creation of a statistical model, that is, a fundamental probability distribution from which the visible data is drawn [1]. The credit card customers and industry has considered computing statistical models for automation. Recently, these types of statistical models have been used to

various areas likes academic research, especially with respect to share markets and e-commerce. The credit card fraud-detection domain is s growing field in Covid periods and presents many challenging and issues for using data mining techniques [2].

The global e-commerce will predict to reach almost 7k billion US dollars in 2023 as shown in Fig. 1. E-commerce frauds mainly focused on credit cards as a means of payment through credit card adoption. The credit card fraud ratio is the same or perhaps slightly growing indicted worldwide as shown in Fig. 2 [3]. The annual value of consumers losses on card-not-present (CNP) fraud for debit and credit cards issued in the United Kingdom (UK) alone for 2019 amounted to 470.2 million GBP [4]. Over the years, technology has significantly changed, and so have the fraud patterns. Today, CNP is a dominant type of fraud, as visible in Fig. 2, and it is reported that [5]: CNP fraud accounted for 1.43 billion in fraud losses in 2018 that means increase of 17.7% compared with 2017 [6].

II. Related Works

Credit card fraud detection is one of the most essential research area discovered domains of fraud detection and its depend on the computerization recorded transactions to detect fraudulent behavior. Every time a credit card is used, transaction data, composed of several attributes (e.g., credit card identifier, transaction date, recipient, amount of the transaction), are stored in the databases of the service provider. However, a single transaction information is typically not sufficient to detect a fraud occurrence and the analysis must consider aggregate measures like total spent per day, transaction number per week or average amount of a transaction [7].

Data mining technique used in solving unsolvable problems in the field of credit fraud detection problem. Credit card fraud detection mainly used to identify those transactions that are fraudulent into two main classes of valid (genuine) and fraudulent transactions. Credit card fraud finding is constructed on analysis of a card's spending behavior. Many techniques have been applied to credit card fraud detection, artificial neural network, genetic algorithm, support vector machine, frequent itemset mining, decision tree, migrating bird's optimization algorithm and naïve bayes [8].

Now a days most of the research compare their proposed related review of literature and its algorithms with a benchmark algorithm. The comparison using different standard binary for classification measurement called as misclassification error, receiver operating characteristic (ROC), Kolmogorov–Smirnov (KS), F1Score [9] or AUC statistics [10]. The above measures may not be the most appropriate evaluation methods for evaluating the credit card fraud detection models and assume that different misclassification errors carry the same cost and similarly with the correct classified transactions [11]. Now a days most of the researchers are using statistical modeling tools used to solve the task, observe the true singularities in the real world only through a proxy given as a restricted set of point-wise observations. In credit-card fraud detection, the true phenomenon of interest is the genuine purchase behavior of card holders or, likewise, the malicious behavior of fraudsters [12]. In our experiment, we use

random forest as a classifier [13]. The popularity of decision tree models in data mining is owed to their simplification in algorithm and flexibility in handling different data attribute types [14]. However, single-tree model is possibly sensitive to specific training data and easy to overfit [15] and [16].

Seven classification algorithms such as J48, Random Tree (RT), Decision Stump (DS), Logistic Model Tree (LMT), Hoeffding Tree (HT), Reduce Error Pruning (REP) and Random Forest (RF) are used to measure the accuracy. Data mining and machine learning techniques build with WEKA (Waikato Environment for Knowledge Analysis) tools which is used to finding experimental results of weather related dataset. Out of seven classification algorithms, Random tree algorithm outperforms other algorithms by yielding an accuracy of 85.714% [17] and the authors discuss another paper clearly data mining with decision tree approaches using medical related data [18].

The real-world research mainly focuses on unsupervised learning techniques can help the fraud detection systems to find various anomalies relating to the credit card issues. In this research paper authors present a hybrid supervised and unsupervised techniques to improve the fraud detection accuracy. Unsupervised outlier scores, figured at different levels of granularity, are compared and tested on a real, annotated, credit card fraud detection dataset [19].

III. Credit Card Fraud Detection

The pandemic situations, the credit card companies can diagnose fraudulent credit card transactions so that customers are not exciting for items that they did not obtaining. The following Fig. 1 and Fig. 2 indicate credit card fraud detection increase year by year and explain evaluation of the total value of credit card fraud within single euro payments area. Finally, it is also vital for businesses not to detect transactions which are honest as fraudulent, if not, the companies would keep blocking the credit card, and which may lead to client disappointment. So at this time are two important expects of this analysis:

- What would happen when the company will not be able to detect the fraudulent transaction and would not confirm from a customer about this recent transaction whether it was made by him/her.
- In contract, what would happen when the company will detect a genuine transaction as fraudulent and keep calling customer for confirmation or might block the card.

The datasets [20] contain transactions that have 492 frauds out of 284,807 transactions. So, the dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. When we try to build the prediction model with this kind of unbalanced dataset, then the model will be more inclined towards to detect new unseen transaction as genuine as our dataset contains about 99% genuine data. As our credit card dataset is highly imbalanced, use accuracy score as a metric because it will be usually high and misleading, we should focus on f1-score, precision/recall score or confusion matrix.

Due to confidentiality issue, original features v1, v2, ... v_{28} have been transformed with PCA, however, we may guess that these features might be originally credit

card number, expiry date, CVV, cardholder name, transaction location, transaction date-time, etc. Two features have not been changed with PCA are 'Time' and 'Amount'. In this research feature 'Time' contains the seconds forgotten between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount for the corresponding card customer, this feature can be used for cost-sensitive learning and the feature 'Class' denoted as response variable, and it takes value 1 in case of fraud and 0 otherwise.

IV. Experimental Setup and Methods

A. Multiple Linear Regression (MLR)

In statistical analysis, linear regression is a basic and most of the researchers are commonly used prediction problems. The overall interpretation of regression analysis is to clearly be explained two types of things, set of predictor variables do amazing jobs for predicting the futures based on dependent variables and variables are important predictors of the outcome variable. This regression which is used to explain the relationship between one dependent variable and one or more independent variables.

Multiple linear regression model known as multiple regression. It is used to finding the prediction and a statistical technique which is include several explanatory variables is to be used to predict the outcome of the response or predicted variable. The advantages of MLE model are to model the linearity between the explanatory or independent variable and response or dependent variable and the MLE equations as defined in the following.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon \qquad (1)$$

Where, y_i denoted as response or dependent variable, x_i means explanatory or independent variables, β_0 means y-intercept that means terms of constant, β_p coefficients of the given slope for the given independent variables and ε denoted as error terms in the given model also known as residuals [21].

B. Symbolic Representation

The main objective for symbolic representation is to be used to understand the nature of the parameter. The symbolic representation is one the methods for solving the statistical related problems. In this research mainly focused three major parameters namely time, confidentiality, amount, and class label. The parameters 'time' assigned as 'θ', sources of 'confidentiality' level assigned as 'μ'. Similarly, other fields like 'amount' of transactions named as 'δ' and the 'class' label assigned as 'η' respectively.

C. Correlation Coefficient

The linear correlation coefficient which is used to the given data that measures the strength of the given data and the linear relationship between two variables namely x and y. The linear correlation coefficient indicates the sign and direction of the linear relationship between x and y. When r nothing but correlation

coefficient and the range of near 1 to −1. The correlation coefficient equation returns near 1 which means linear relationship is strong and the correlation coefficient return near 0, the linear relationship is weak. The linear correlation coefficient equation is presented in equation (2).

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (2)$$

D. Decision Tree

The decision tree approaches used to involve the probability-based decision making which is represented the outcomes in a tree-shaped diagram to chart out a course of action in a statistical probability analysis. In this approach used to solve complex problems or branches and each possible branch indicate the decision tree could be a possible outcome. In this research using four different decision tree approaches namely M5, random forest, random tree, and REP tree. M5 model decision tree approaches is used to learner for regression task which is used to predict values of numerical response variable Y. The M5 decision tree approaches having linear regression functions are used to find the leaf nodes. M5 decision tree approaches very useful for finding the forecasting in various research area which can predict continuous numerical attributes. Random forests are an ensemble learning method for finding classification, regression, and other related tasks of decision trees at training time. It is used to solve the classification tasks; the output of the random forest is the class selected by most trees.

Random tree is easy to intellectualize but will naturally suffer from extreme variance, which makes them not good in terms of precision. One way to affect this restriction is to make several alternatives of a single decision tree by choosing every time a separate subset of the same training set in the context of randomization-based ensemble methods. Random Trees (RT) belong to a class of machine learning algorithms which is used to makes predictions by averaging over the predictions of several independent base models. REP decision tree is a method for finding to generate a tree from a given dataset. It is careful to be an extension of C45 decision tree approach by enhancing the pruning stage by using Reduced Error Pruning (REP).

E. Mean Absolute Error (MAE)

In statistical analysis the mean absolute error is used to measure the model error between paired observations expressing the similar singularity. For statistical examples of 'y' and 'x' include the comparisons of predicted versus observed values and subsequent time versus initial time. The mean absolute error which is used to an alternative technique of measurement.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad (3)$$

Where, y_i denoted as prediction and x_i means true values and 'n' denoted as total number of data points or number of observations.

F. Root Mean Squared Error (RMSE)

Root mean square error is the standard deviation of the residuals called prediction errors. Residuals is used to measure of how far from the regression line data points presented. RMSE is a measure of how spread out these residuals and used to find it tells you how concentrated the data is around the line of best fit. The RMSE mainly used for forecasting or predicting the future of commonly used in climatology, forecasting, and regression analysis to verify experimental results.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \widehat{x_i})^2}{N}} \qquad (4)$$

Where *'i'* means individual variable in the corresponding columns, *'N'* means number of non-missing data points, x_i means actual observations of time series data and $\hat{x}_i$ means estimated times series.

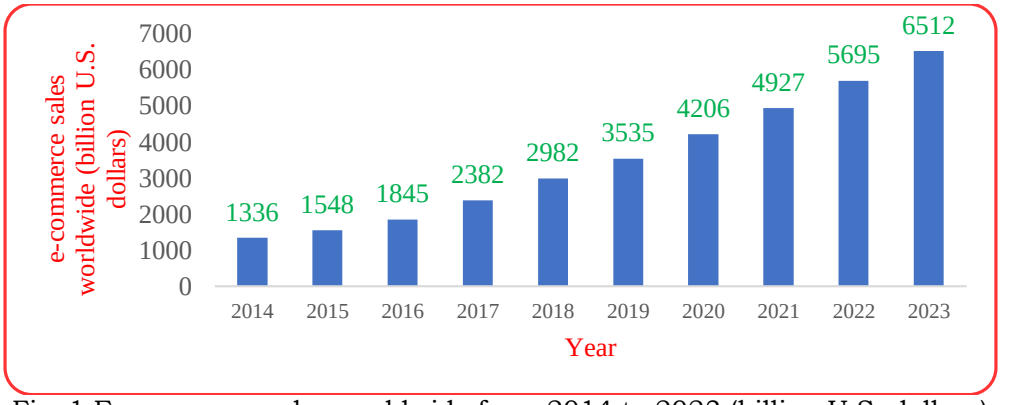

Fig. 1 E-commerce sales worldwide from 2014 to 2023 (billion U.S. dollars)

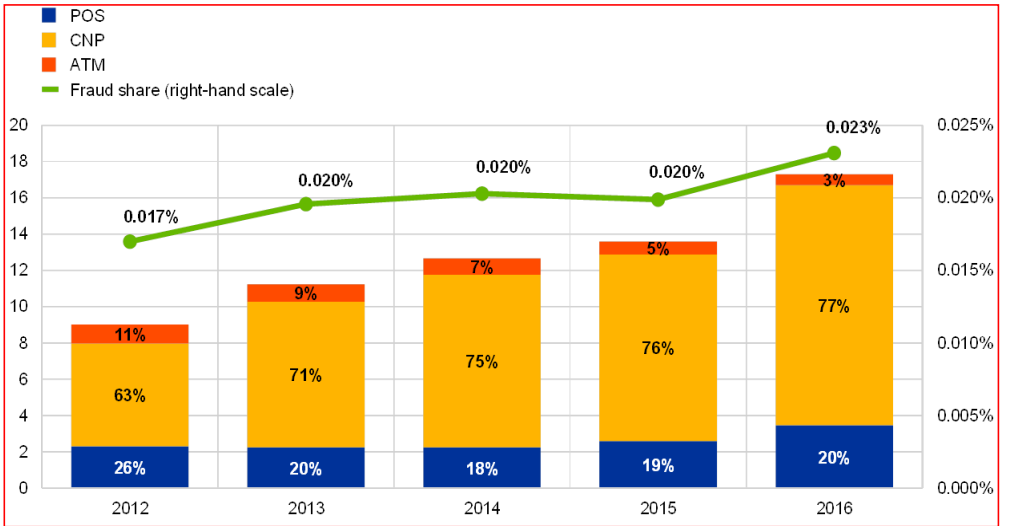

Fig. 2 Evolution of the total value of credit card fraud within Single Euro Payments Area

Table 1. Features in German Credit Dataset [25].

Features	Description
C1	Credit card transaction time in seconds
C2-C29	Confidentiality features which are indicated v1 to v28
C30	Transferred amount
C31	Class which is indicated whether the transaction called fraud or not means (1 and 0)

As our credit card dataset is highly imbalanced, use accuracy score as a metric because it will be usually high and misleading, we should focus on f1-score, precision/recall score or confusion matrix.

Table 2. Credit card fraud detection dataset with time, confidentiality features and transferred amount

C1	C2	C3	C4	C5	C6	C7	C8	C9		C29	C30	C31
Time	v1	v2	v3	v4	v5	v6	v7	v8	...	v28	Amount	Class
0	-1.3598	-0.0728	2.5363	1.3782	-0.3383	0.4624	0.2396	0.0987	...	-0.0211	149.62	0
0	1.1919	0.2662	0.1665	0.4482	0.0600	-0.0824	-0.0788	0.0851	...	0.0147	2.69	0
1	-1.3584	-1.3402	1.7732	0.3798	-0.5032	1.8005	0.7915	0.2477	...	-0.0598	378.66	0
1	-0.9663	-0.1852	1.7930	-0.8633	-0.0103	1.2472	0.2376	0.3774	⋯	0.0615	123.5	0
2	-1.1582	0.8777	1.5487	0.4030	-0.4072	0.0959	0.5929	-0.2705	⋯	0.2152	69.99	0
2	-0.4260	0.9605	1.1411	-0.1683	0.4210	-0.0297	0.4762	0.2603	⋯	0.0811	3.67	0
4	1.2297	0.1410	0.0454	1.2026	0.1919	0.2727	-0.0052	0.0812	⋯	0.0052	4.99	0
7	-0.6443	1.4180	1.0744	-0.4922	0.9489	0.4281	1.1206	-3.8079	⋯	-1.0853	40.8	0
7	-0.8943	0.2862	-0.1132	-0.2715	2.6696	3.7218	0.3701	0.8511	⋯	0.1424	93.2	0
9	-0.3383	1.1196	1.0444	-0.2222	0.4994	-0.2468	0.6516	0.0695	⋯	0.0831	3.68	0

G. Relative Absolute Error (RAE)

Relative Absolute Error (RAE) is used to find the performance of a predictive model. The Relative Absolute Error is expressed as a ratio, comparing a mean error or residual to errors produced by a trivial or naive model. The RAE approximation error in a data value is the discrepancy between an exact value and some approximation to it.

$$\delta = \left|\frac{u_A - u_E}{u_E}\right| . 100\% \qquad (5)$$

Where δ means percentage of the error occurs in the corresponding model, u_A actual value observed and u_E expected values.

H. Root Relative Squared Error (RRSE)

The root relative squared error is used to find the relative to a simple predictor had been used. The simple predictor and find the average of the actual values. Based on statistical analysis, the root relative squared error Ei of each individual model and i is evaluated by the RRSE equation.

$$E_i = \sqrt{\frac{\sum_{j=1}^{n}\left(P_{(ij)}-T_j\right)^2}{\sum_{j=1}^{n}\left(T_j-\bar{T}\right)^2}} \quad (6)$$

Where P(ij) means predicted values by the corresponding individual model i means record of j (out of n records), Tj stands for target value for record of j and T ¯ presented using the given formula:

$$\bar{T} = \frac{1}{n}\sum_{j=1}^{n} T_j$$

For RRSE called a perfect fit, the numerator is equal to 0 and E_i = 0. The E_i index ranges between 0 to infinity, with 0 corresponding to the ideal for the corresponding model.

I. Data Normalization

Highlight scaling is a strategy used to institutionalize the scope of independent variables or feature of data. In information handling, it is otherwise called data normalization and is generally performed the information preprocessing step is accustomed to carry all values into the range [0, 1]. This is additionally called unity-based standardization. This can be summed up to limit the scope of values in the dataset between any self-assertive point 'a' and 'b' and assign (0.1, 0.9) respectively, the similar methods discussed using various dataset [22, 23, 24].

$$X' = a + \frac{(X - X_{\min})(b-a)}{X_{\max} - X_{\min}} \quad (7)$$

Where 'a' and 'b' namely define the maximum and minimum range, 'X' means corresponding column value, X_{min} denoted as minimum value of the corresponding column and X_{max} denoted as maximum value of the column.

V. Experimental

The current situations to increase most of the people for using debit/credit card in day-to-day activities. In this manner for using cards trust & safety is a major problem for using all platform. Fraud detection and prevention at every step of the credit/debit card user journey. The above dataset could be inferred that the 28 anonymized features and 2 non anonymized features such as amount and Class which means whether the transaction was a fraud or not. The statistical description and summary show in Table 1 having some preliminary statistical measurements.

Table 3. Credit card fraud detection dataset with descriptive statistics

Descriptive Statistics	Time (θ)	Confidentiality (μ)	Amount (δ)	Class (η)
Minimum	0	-7.084	0	0
Maximum	172792	0.802	25691.160	1
Mean	94813.860	0	88.350	0.002
Standard Deviation	47488.150	0.198	250.120	0.042

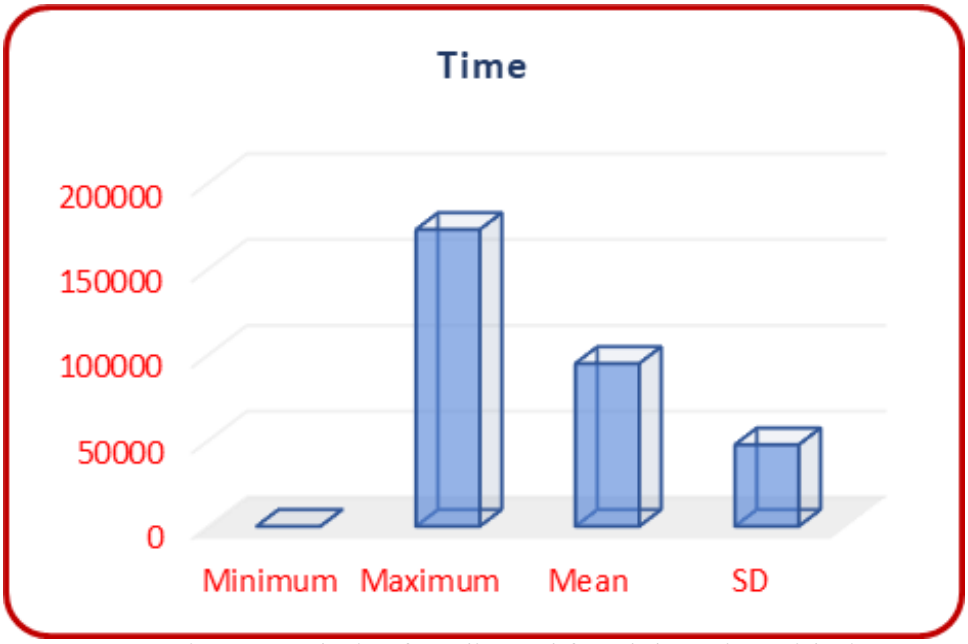

Fig. 3a Analysis of credit card fraud detection with time

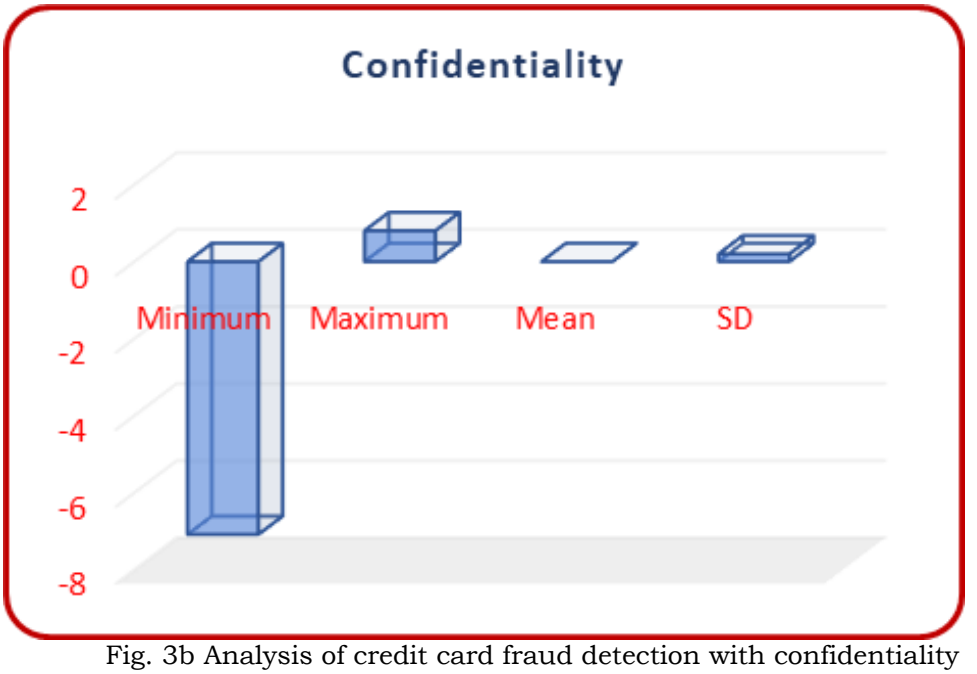

Fig. 3b Analysis of credit card fraud detection with confidentiality

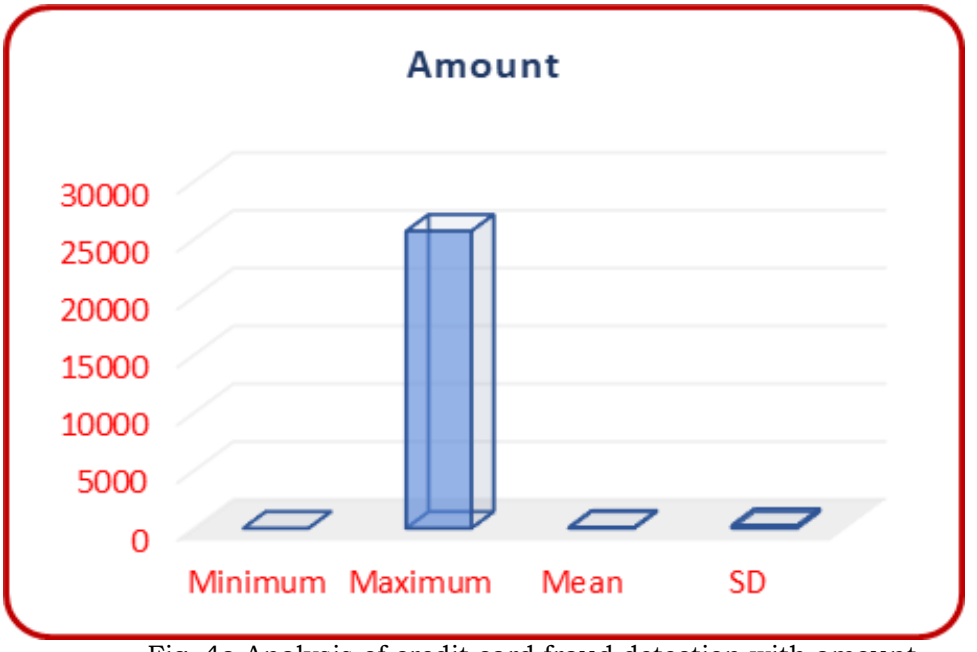

Fig. 4a Analysis of credit card fraud detection with amount

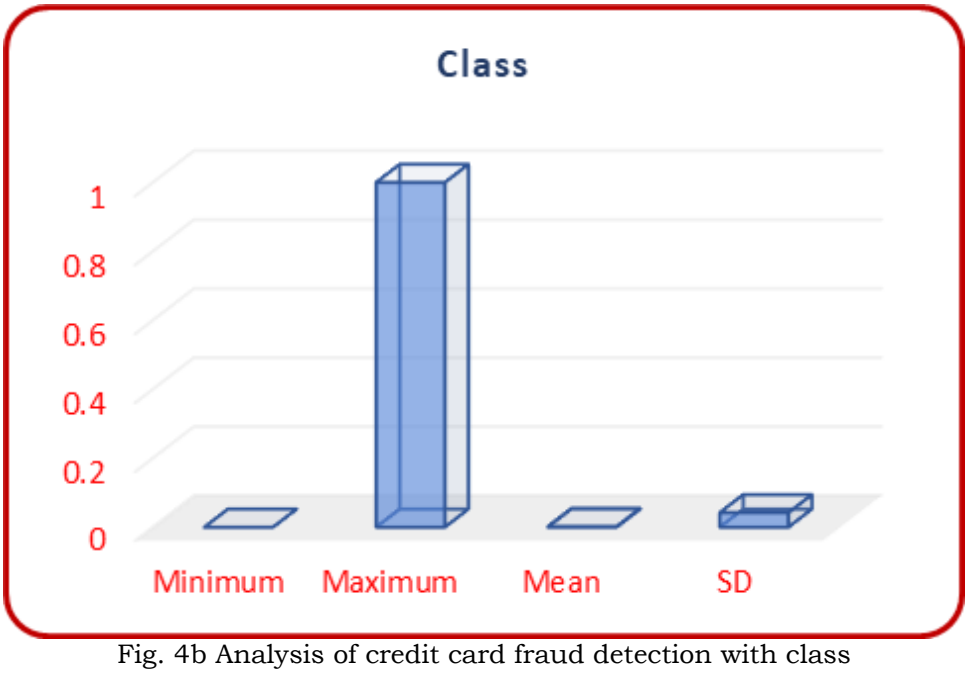

Fig. 4b Analysis of credit card fraud detection with class

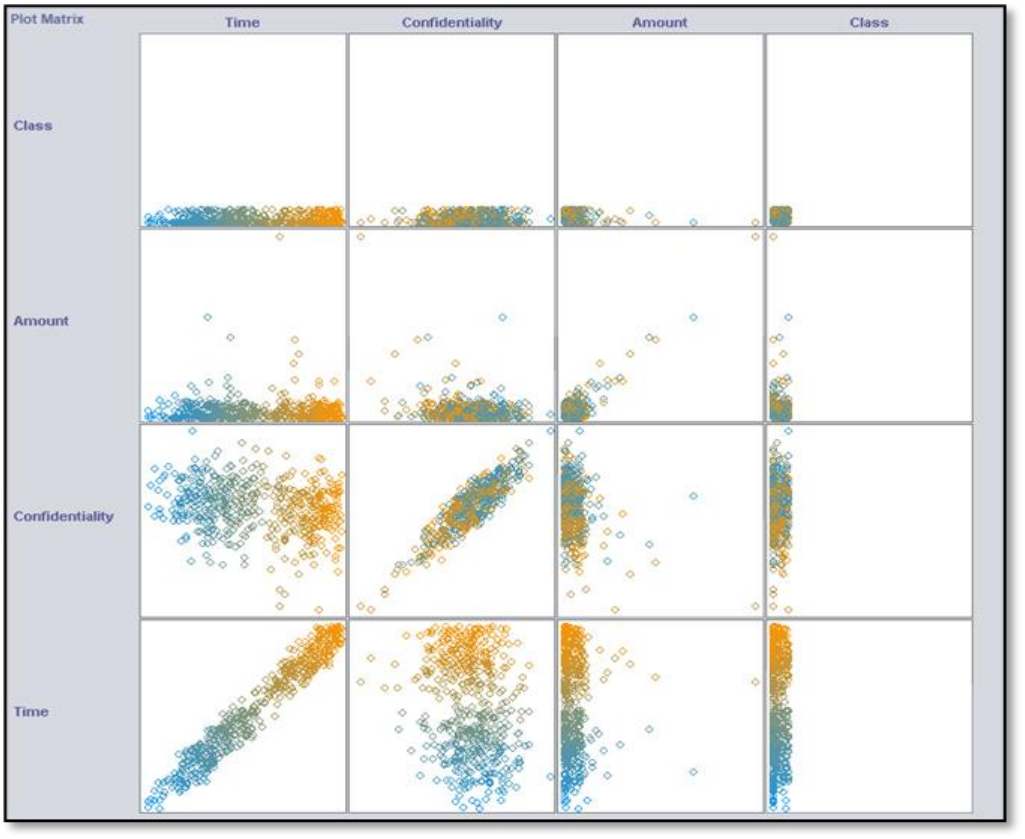

Fig. 5 Clustering analysis for credit card fraud detection

Table 4a. Multiple linear regression approaches

Name of the Parameters	Regression Equations
Time (θ)	-41198.1685 * μ + -9.6678 * δ + -75896.4961 * η + 95799.1183
Confidentiality (μ)	-0 * θ + -0.0002 * δ + -1.5103 * η + 0.0765
Amount (δ)	-0.0003 * θ + -339.9048 * μ + -482.3367 * η + 113.6039
Class (η)	-0 * θ + -0.0719 * μ + -0 * δ + 0.0079

Table 4b. Multiple linear regression approaches and its accuracy parameters

Name of the Parameters	Time taken (Seconds)	Correlation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
Time (θ)	0.36	0.16	41912.04	46887.85	97.93	98.74
Confidentiality (μ)	0.22	0.42	0.12	0.1797	93.87	90.77
Amount (δ)	0.21	0.25	102.25	242.07	98.76	96.78
Class (η)	0.21	0.33	0.01	0.04	291.54	94.46

Table 5. Finding accuracy for credit card fraud detection data using normalization

Name of the Parameters	Time Taken (Seconds)	Correlation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
Time (θ)	0.3600	0.1600	0.9000	0.9000	0.1164	0.9000
Confidentiality (μ)	0.2200	0.4200	0.1000	0.1000	0.1000	0.1000
Amount (δ)	0.2100	0.2500	0.1020	0.1041	0.1198	0.7033
Class (η)	0.2100	0.3300	0.1000	0.1000	0.9000	0.4704

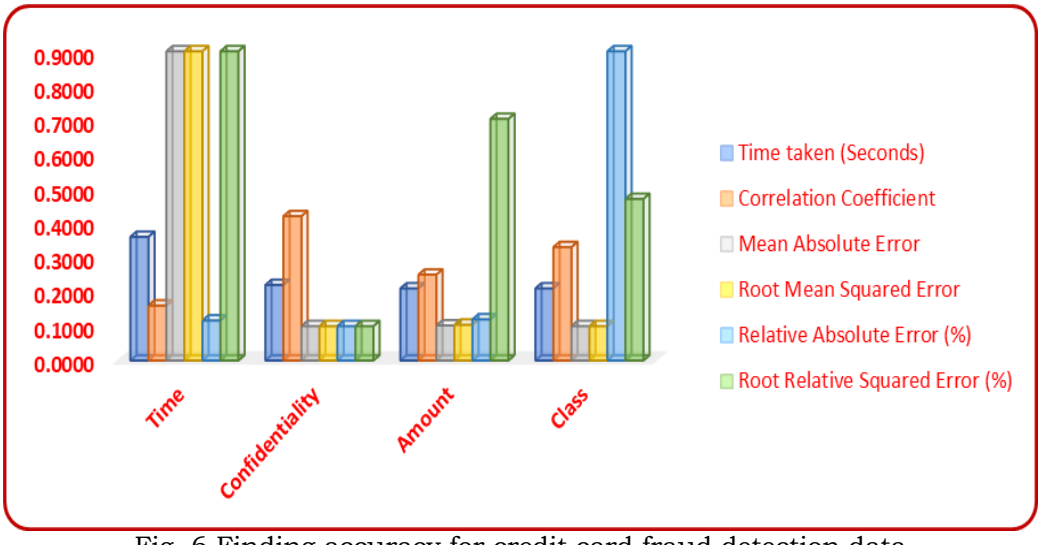

Fig. 6 Finding accuracy for credit card fraud detection data

Table 6. Decision tree approaches using cross-validation

Decision Tree Classifier	Size of the Tree	Time taken to build model (Seconds)	Correlation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
M5	240	12.05	0.5744	0.0023	0.0340	66.5190	81.8706
Random Forest	100	171.03	0.6246	0.0019	0.0325	56.5373	78.3061
Random Tree	1137	3.23	0.4438	0.0019	0.0434	54.6672	104.5627
REP Tree	77	3.19	0.5585	0.0022	0.0346	64.1355	83.3015

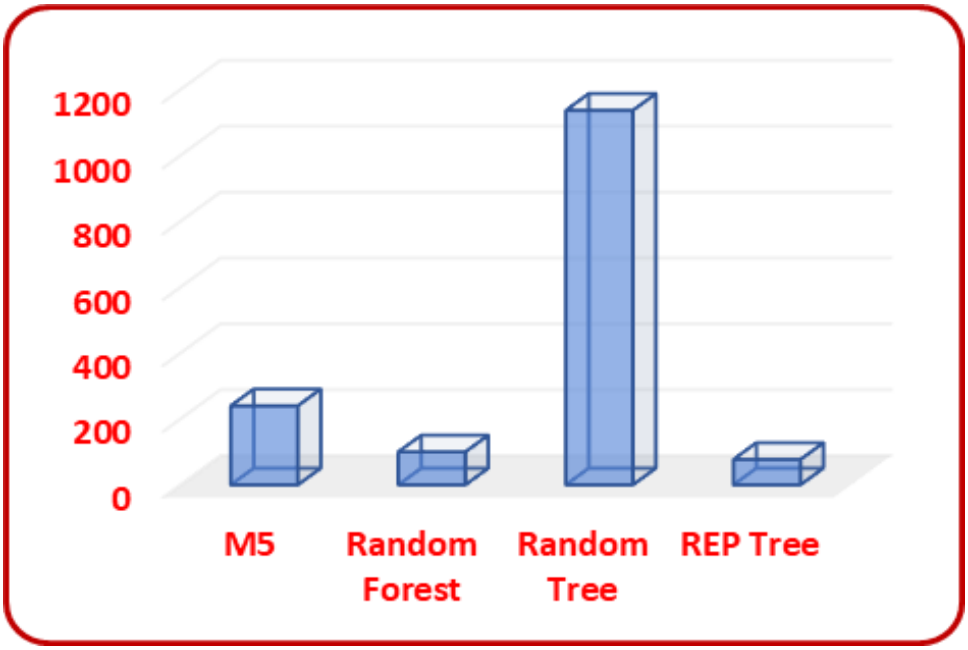

Fig. 7 Decision tree comparison using tree size to build the tree

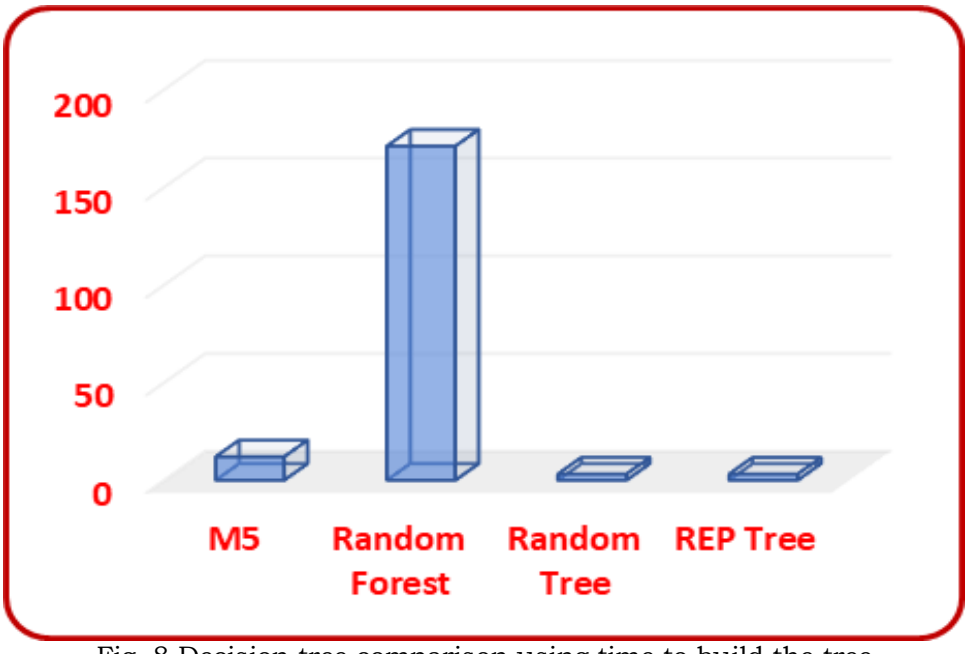

Fig. 8 Decision tree comparison using time to build the tree

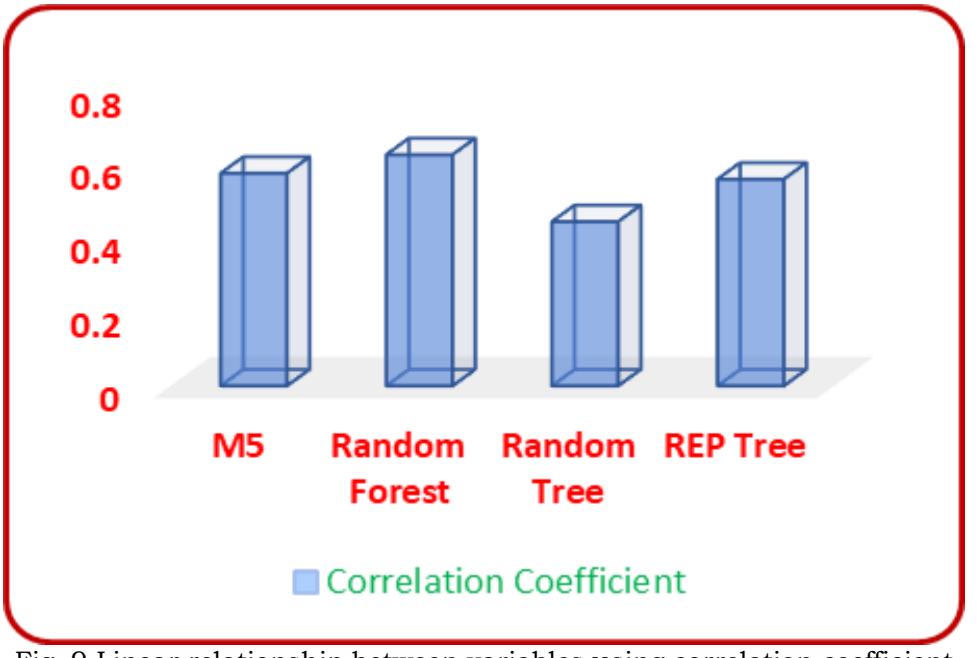

Fig. 9 Linear relationship between variables using correlation coefficient

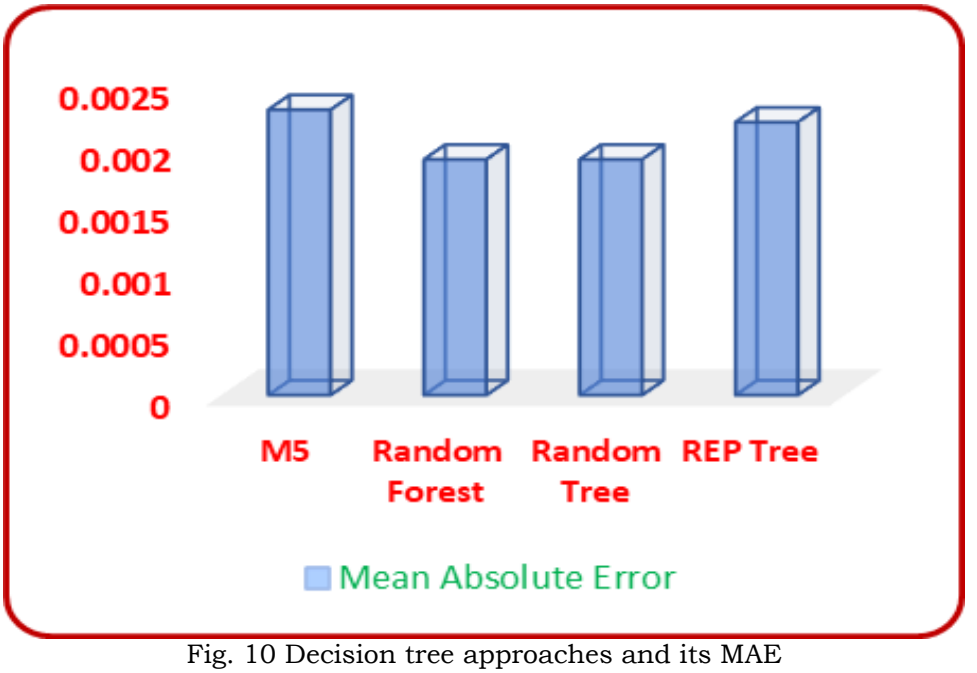

Fig. 10 Decision tree approaches and its MAE

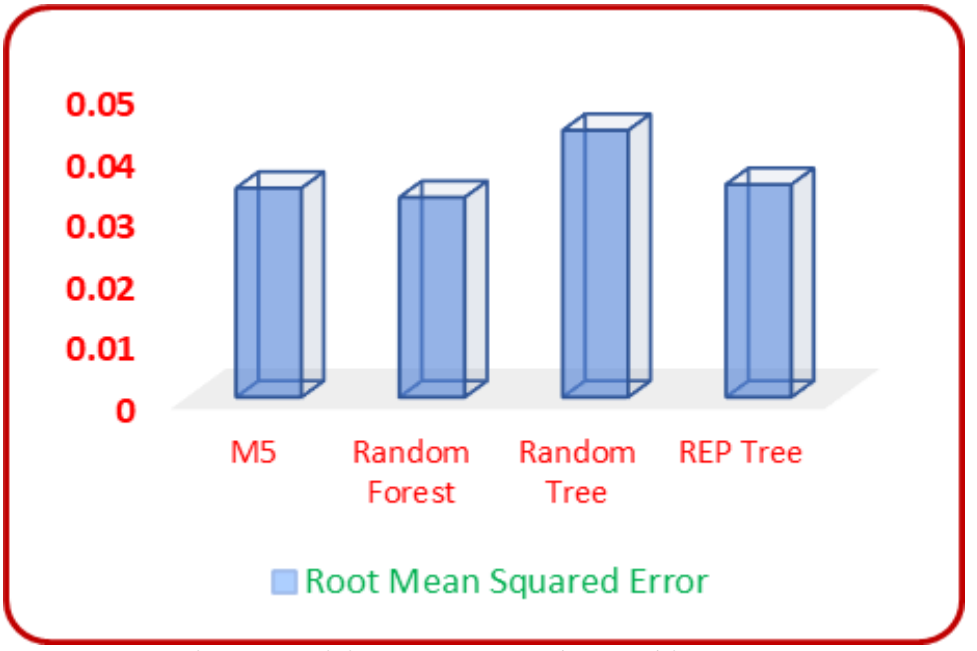

Fig. 11 Decision tree approaches and its RMSE

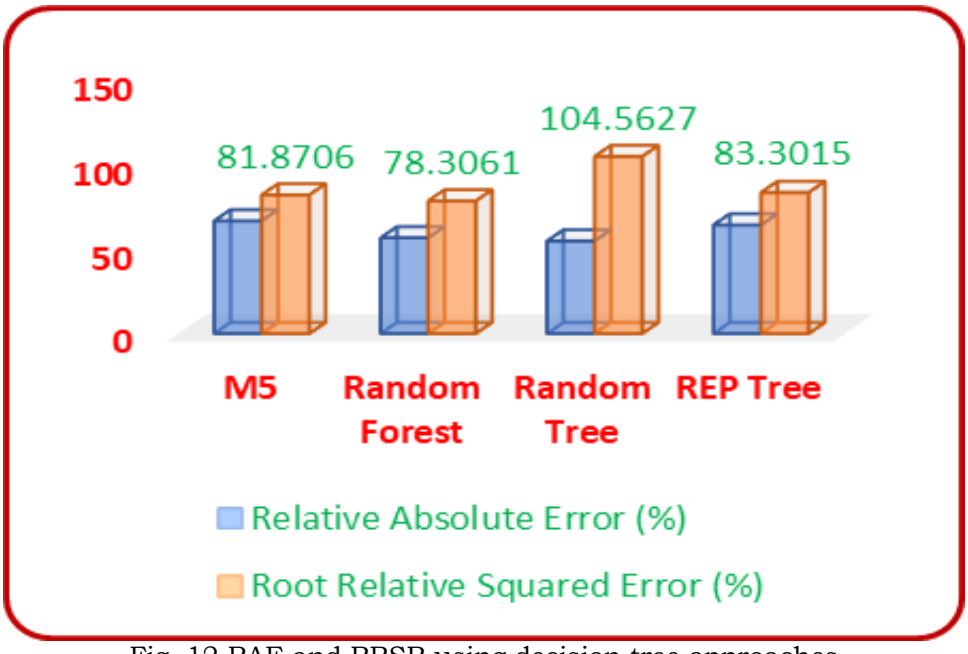

Fig. 12 RAE and RRSR using decision tree approaches

VI. Summary and Conclusions

For using Table 2 and the datasets [20] contain transactions that have 492 frauds out of 284,807 transactions. In this case, the fraud transactions occur 0.1954% and the genuine transaction transactions occurs 99.8040%. Based on the descriptive statistics, mean is the average samples of credit card transactions level observed data.

The entire dataset having the average amount of transactions Rs. 88.35 and the standard deviation which is used to find the amount of variation or dispersion of amount Rs. 250.12. The maximum amount of credit card transition Rs. 25691.16 and the minimum Rs. 0, the numerical illustrations show in Table 3, Fig. 3 and Fig. 4 based on time and amount of transactions. Clustering analysis for credit card fraud detection using four parameters namely time, confidentiality, amount, and class, the corresponding clustering analysis shows in Fig. 4.

Multiple linear regression which is used to predict the future with symbolic notations using equation (1). In this case, the dataset having four types of parameters namely time (θ), confidentiality (μ), amount (δ) and class (η). MLE used to finding forecasting time, confidentiality, amount and class equations, the corresponding equations shows in Table 4. Fig. 5, indicate the plot matrix which is used to understanding the clustering pattern in various combinations of the parameters namely time (θ), confidentiality (μ), amount (δ) and class (η), the corresponding diagrammatical representation shows in Fig. 5.

Multiple linear regression approaches followed by finding four response or dependent variable y_i using equation (1) relating to the independent variable namely time (θ), confidentiality (μ), amount (δ) and class (η) and the table having

different scale of data namely second, percentage and error terms shows in Table 4. In this case different scales of data not suitable for numerical analysis, in this case various scales of data converted into similar scale using the familiar normalization equation (7) and show the corresponding normalized valued present in Table 5 and Fig. 6. The normalization techniques are very essential for researchers in data mining and data analytics, and it is used to very helpful to write the necessary interpretation.

The decision tree approaches used to involve the probability-based decision making which is represented the outcomes in a tree-shaped diagram. In Table 6, indicates four decision tree approaches and its accuracy parameters namely correlation coefficient, mean absolute error, root mean squared error, relative absolute error, and root relative squared error shows in Table 6. The Fig. 7 bar chart indicates, comparison between four decision tree approaches, REP decision tree burning size are less and random tree size is high. The time taken to build the model having the random forest spend more time and REP tree take less time compared to other approaches result shows in Fig. 8. The linear correlation coefficient which is used to measures the strength of the given data and the linear relationship between two variables namely x and y using equation (2). In this case random forest consuming high correlation and random tree return low correlation, the results show in Fig. 9. In statistical analysis the Mean Absolute Error (MAE) is used to measure the model error between paired observations. In this approach used to measure MAE, the M5 decision tree return high, and the random forest and random tree approaches return low and the related result shows in Fig. 10.

The Root Mean Squared Error (RMSE) is the standard deviation of the residuals called prediction errors. Residuals is used to measure of how far from the regression line data points presented. In this approach random tree return high value and M5 decision tree return low value, corresponding result shows in Fig. 11. Relative Absolute Error (RAE) is used to find the performance of a predictive model and the root relative squared error is used to find the relative to a simple predictor had been used. The related results show in Fig. 12.

References

[1] J. Leskovec, A. Rajaraman and J. D. Ullman, Mining of massive data sets, Cambridge university press. (2020).

[2] P. K. Chan, W. Fan, A. L. Prodromidis and S. J. Stolfo, Distributed data mining in credit card fraud detection, IEEE Intel. Sys. and their App. 14(6) (1999) 67-74.

[3] I. Mekterovic, M. Karan, D. Pintar and L. Brkic, Credit card fraud detection in card-not-present transactions: where to invest, Appli. Sci. 11(15) (2021) 6766

[4] (2022) Statista retail e-commerce sales worldwide from 2014 to 2023 (in Billion U.S. Dollars). [online]. Available: https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales

[5] (2022) Statista value of annual losses on “card-not present” fraud on uk-issued debit and credit cards in the United Kingdom (UK) from 2002 to

2019. [online]. Available: https://www.statista.com/statistics/286245/united-kingdom-uk-card-not-present-fraudlosses/).
[6] E. C. Bank, Sixth report on card fraud, European central bank: Frankfurt, Germany. (2020).
[7] A. Dal Pozzolo, C. Andrea, L. Olivier, W. Yann-Ael, B. Serge and Gianluca, learned lessons in credit card fraud detection from a practitioner perspective, Exp. Syst. with Appli. 41(10) (2014) 4915-4928.
[8] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, Credit card fraud detection using machine learning techniques: A comparative analysis, IEEE Int. Con. on Comp. Net. and Info. (2014) 1-9.
[9] D. J. Hand, C. Whitrow, N. M. Adams, P. Juszczak and D. J. Weston, performance criteria for plastic card fraud detection tools, J. of the Oper. Res. Soc. 59(7) (2007) 956–962.
[10] (2022) Binary classifier evaluation matrices: error rate, KS statistics, AUROC, (online), Available: http://rstudio-pubs-static.s3.amazonaws.com/303414_fb0a43efb0d7433983fdc9adcf87317f.html.
[11] A. Correa Bahnsen, D. Aouada, A. Stojanovic, B. Ottersten, Feature engineering strategies for credit card fraud detection, Exp. Sys. with Appli. 51 (2016) 134–142.
[12] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P. E. Portier, L. He-Guelton and O. Caelen, Sequence classification for credit-card fraud detection, Exp. Sys. with Appli. 100 (2018) 234-245.
[13] L. Breiman, Random forests. Mach. Lear. 45(1) (2001) 5-32.
[14] J. R. Quinlan, Induction on decision tree, Mach. Lear. 1(1), (1986) 81-106.
[15] S. Bhattacharyya, S. Jha, K. Tharakunnel and J. C. Westland, Data mining for credit card fraud: a comparative study, Deci. Supp. Sys. 50(3) (2011) 602-613.
[16] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang and C. Jiang, Random Forest for credit card fraud detection, IEEE 15th Int. Con. on Net., Sen. and Con. (2018) 1-6.
[17] P. Rajesh and M. Karthikeyan, A comparative study of data mining algorithms for decision tree approaches using WEKA tool, Adv. in Nat. and Appli. Sci. 11(9) (2011) 230-243.
[18] P. Rajesh, M. Karthikeyan, B. Santhosh Kumar and M. Y. Mohamed Parveesh, Comparative Study of Decision Tree Approaches in Data Mining Using Chronic Disease Indicators (CDI) Data, J. of Compu. and Theo. Nano. 16(4) (2019) 1472–1477.
[19] Rinartha, K., & Suryasa, W. (2017). Comparative study for better result on query suggestion of article searching with MySQL pattern matching and Jaccard similarity. In 2017 5th International Conference on Cyber and IT Service Management (CITSM) (pp. 1-4). IEEE.
[20] F. Carcillo, Y. A. Le Borgne, O. Caelen, Y. Kessaci, F. Oble and G. Bontempi, Combining unsupervised and supervised learning in credit card fraud detection, Infor. Sci. 557, (2021) 317-331.
[21] (2022) Credit card fraud detection dataset (online), Available: https://www.kaggle.com/dktalaicha/credit-card-fraud-detection-using-smote-adasyn/data.
[22] (2022) Multiple linear regression, (Online), Available: https://www.investopedia.com/terms/m/mlr.asp

[23] P. Rajesh and M. Karthikeyan, Data mining approaches to predict the factors that affect the agriculture growth using stochastic model, Int. J. of Comp. Sci. and Engg. 7(4) (2019) 18-23.
[24] P. Rajesh and M. Karthikeyan, and R. Arulpavai, Data mining approaches to predict the factors that affect the groundwater level using stochastic model, AIP Con. Proc. 2177(1), (2019) 020079.
[25] P. Rajesh, M. Karthikeyan and R. Arulpavai, Predication of labour demand in agriculture based on comparative study of different data using data mining and stochastic approach, Int. J. of Engg. Sci. Inv. 2 (2018) 86-97.
[26] Ernawati, E., Baso, Y. S., Hidayanty, H., Syarif, S., Aminuddin, A., & Bahar, B. (2022). The effects of anemia education using web-based she smart to improve knowledge, attitudes, and practice in adolescent girls. International Journal of Health & Medical Sciences, 5(1), 44-49. https://doi.org/10.21744/ijhms.v5n1.1831
[27] G. K. Arun and P. Rajesh, Design of metaheuristic feature selection with deep learning based credit card fraud detection model, Second International Conference on Artificial Intelligence and Smart Energy (ICAIS). (2022) 191-197.