Antardrishti IUD Journal
of Interdisciplinary Research
Journal home page: www.iudehradun.edu.in

COMPARISONS OF MACHINE LEARNING ALGORITHMS FOR FRAUDULENT ANALYSIS IN FINANCIAL SECTOR

Nishant Mathur[1*] and Dr. Mukul Jain[2]

[1]Assistant Professor and Research Scholar, [2]Assistant Professor, ICFAI Tech School
ICFAI University, Rajawala Road, Selaqui, Dehradun- 248011, Uttarakhand, India

Corresponding author's Email: nishant.m@iudehradun.edu.in

ABSTRACT

With the approach of new advances accessible to deal with the economy around the world, it's anything but a reality obscure that there is significant dependence on the utilization of credit only techniques for exchanges. This likewise results from the need to manage the disadvantages associated with utilization of hard cash. Here, a significant job is played by the utilization of charge cards which work with significant inclusion of the web. With this, additionally emerges the negligence of monetary misrepresentation which is scaling up rapidly, bringing about gigantic misfortunes to the economy. Machine Learning classification is techniques used to order information and predicting accuracy. Point of this inspection is correlation of AI calculations on various datasets. The center of AI is that the program figures out how to produce the data that we furnish and uses to order information with the learning experience store in it. Here, we give a framework to compare the algorithms with different data type. The framework is based on category theory on credit card data for comparing various machine learning algorithms. A relative study of popular in field of supervised learning techniques executed using Decision Tree, Random forest, Logistic regression, K-Nearest-Neighbor, Neural-Network and Support Vector Machine.

Key words: **Decision Tree, Random forest, Logistic regression, K-Nearest-Neighbor, Neural-Network and Support Vector Machine**

1. INTRODUCTION

Monetary misrepresentation is said to happen when somebody unlawfully denies one of their cash or damage their monetary abundance through tricky practices. With the assistance of numerous measurable strategies and assessment examination standards to a great extent checked on to reviews the primary elements for the development of scoring model for credit scoring **[1].** As scoring idea gives better plan to comprehend client practices and their philosophies utilizing various highlights, this thereafter help to as chiefs, especially in area like banking, to anticipate conduct of their customer various sorts cheats in monetary appeared in Table I.

The good learner in machine is the one that can differentiate the two classes impeccably if no same points have different labels and there are no different features have the same label for better classification in machine learning enhancement. Some machine technique also explores how big data generate relation between context and 'things' refer in 'Internet of Things', in 'Internet of Everything' and in 'Internet of People with Things,' and how all these are used to generate 'signs' of semiotic **[2]** to make learning of machine.

Our key goal is to predict decisions based on a specific issue. Many machine learning methods have been used in alike regression and classification to accomplish this goal **[3].** When the prediction goal is a discrete value or a class label Classification is used to design model and Regression is the appropriate method to use in case when prediction goal is continuous.

Table I Different type of Financial Fraud

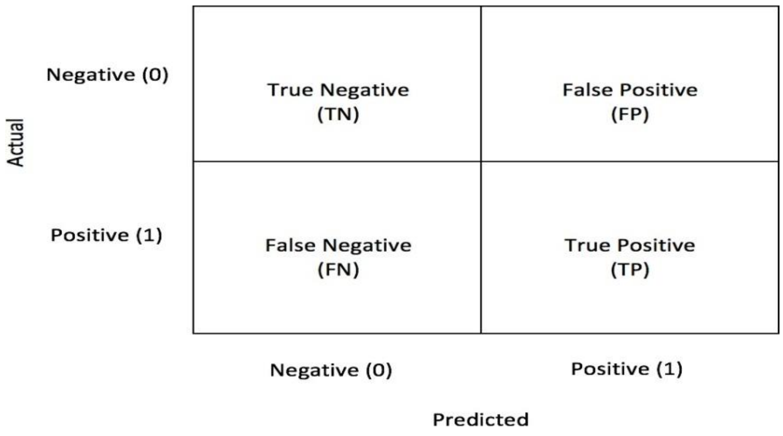
S. No	Types	Description
1	Ponzi Schemes.	Investment schemes which offer a high dividend yield. Any additional capital is now used to pay off previous investors
2	Pyramid Schemes	Schemes that offer big returns to shareholders based solely on the recruitment of everyone else, not on the benefit of actual investment
3	Identity Fraud	Someone personalises you to rob capital or uses your private details.
4	Phishing	Internet banking customers are contacted by e-mail and asked to provide one 's account login, password, and personal information to a webpage that appears to be affiliated with their rightful bank. This information is then used to rob money from the system.

5	Card Fraud	This occurs when a credit card is stolen. The card is still available, and the robber performs illegal transactions with it before the bank is alerted.
6	Skimming	This means detailed outline from a credit card when it is being used in a legal transaction. The scam artist whacks the card into an automated skimming system that stores all details on the fingerprint scanner which could be used to buy or to replicate the card electronically.
7	Counterfeit Card	The fraudster robbed cards to produce false cards or to market them. The perpetrator seldom realizes that he already has the true card
8	Advanced Fee Scam	Such scams are normally made by mail, by email or by telephone with a substantial amount of money if you can send massive sums of money from your country. The fraudster demands for details and administrative fees on the bank account.
9	Fraud Transfer Scam	You are requested by email to deposit into their bank account and to give it in exchange for a profit.
10	Fake Prizes	The suspect believes that you must have won a reward that does not exist and requests your credit card information in order to pay for postage and storage costs.
11	Inheritance Scam	You get an email concerning an unredeemed wealth and are persuaded to pay a premium in exchange for guidance on how to claim it.
12	Wills and legacies	The scammer would deliver an email pretending to be the suspect's legal counsel and requesting a payment in advance.

2. CLASSIFICATION ALGORITHMS AND EXPERIMENTAL PROCEDURE

Models Evaluations

For evaluations of output and accuracy prediction percentage we compute cross matrix, It's a Square matrix containing the output **[5]** of a model on a dataset, where N denotes the number of features used as class labels in the classifier (Fig 1). Finally, we may calculate recall and precision values using a combination of statistics methods such as True Negative (TN), True Positive (TP), False Positive (FP), and False Negative (FN), as seen in Table II, which allows us to obtain real and expected values



Actual		Predicted	
		Negative (0)	Positive (1)
Actual	Negative (0)	True Negative (TN)	False Positive (FP)
	Positive (1)	False Negative (FN)	True Positive (TP)

Fig 1. Confusion cross matrix layout for actual and predicted computation.

Table II Different statistics methods of accuracy prediction.

S. No	Methods	Description
1	True Positive (TP)	In which the real value (e.g., fraud) remained positive and thus the calculated return also became positive.
2	False Positive (FP)	When the real value (e.g., normal) is negative but even the calculated return is positive.
3	True Negative (TN)	When the real value (e.g., normal) appeared negative and even the average payout also became negative.
4	False Negative (FN)	Where the real worth was positive but even the expected value remained negative (e.g., fraud).
5	Recall	The proportion of correctly identified to real positive cases essentially indicates how often these true positives became discovered (recalled) out of some of the positive class cases.
6	Precision	The percentage of correctly identified between number of true positives positives clearly indicates why so many true positives have discovered.

Decision tree

Decision Tree is basically a technique in data science which helps us to make a decision. It is a learning of symmetrical and geographical represents to all the solution of decision problem. In Decision Tree all the internal nodes represented the test condition. All the root nodes divide the data, and all the leaf represents as a class, edge is representing as an attribute of a test node. And it is focusing a discussion when a group is making for a decision.

This algorithm is used for the conditional control statement in operation research, mainly to analysis the decision and to indentify the strategy to reach the goal which makes it popular in machine language. It is simple to interpret and understand to construct the decision tree **[6].** Decision tree bisect the space into a small region and has linear single decision boundary. Using confusion matrix shown in Fig 2 predicting accuracy percentage of 82.0 using decision tree algorithm.

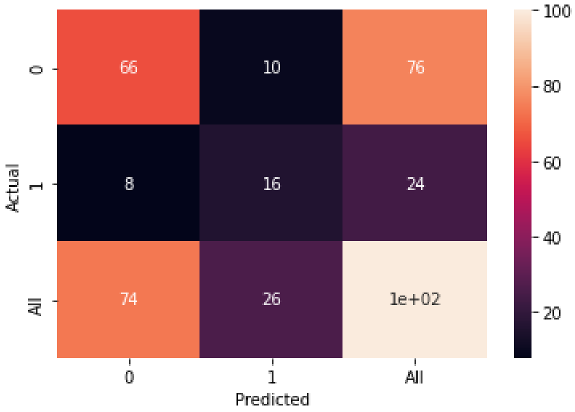


Fig 2 Decision Tree Confusion matrix using Heat Map.

Logistic Regression

This machine learning algorithms perform a binary expression and used in classification problem, which works on probability to predict and analyzed algorithm on continuous variable is Logistic Regression. It also describes the output as a discrete variable 0 or 1. It also identifies relationship between binary variable, dependent on more ordinal or independent variable.

This algorithm is invented by statistician Sir David Roxbee Cox in 1958, it predates in the field of machine language **[7].** This algorithm is the direct method of probability model to compute the probability of an event occurrence and to detect problem. It is also used to predicting binary classes, target outcome and binary variable nature. It is a great effectual algorithm working on low variance and can easily update to new data for stochastic and gradient descent. Logistic regression performs better than random forest when noise number variable is less than explanatory variables or equal to its number. Using confusion matrix shown in Fig 3 predicting accuracy percentage of 76.0 using logistic regression algorithm.

Random Forest

Random Forest algorithm is the supervised technique that construct on multiple sets of decision tree. The tree also grows up very deep to learn irregular patterns and their outfit of a training. It manages the goal of variance of reducing, averaging deep decision of a tree on a set of training. The idea is to combine the multiple decision trees and get the final output rather than to get an individual relying of a decision of a tree. The random tree is suffered from the low bias and high variants, where low bias means model can be fit in the training data set and the high variants means the performance is good at a training of a data set. Random Forest can be classified as a kind of a bagging of a tree for multiple learning to make a better predictive performance model. This algorithm provides the high accuracy **[7].**

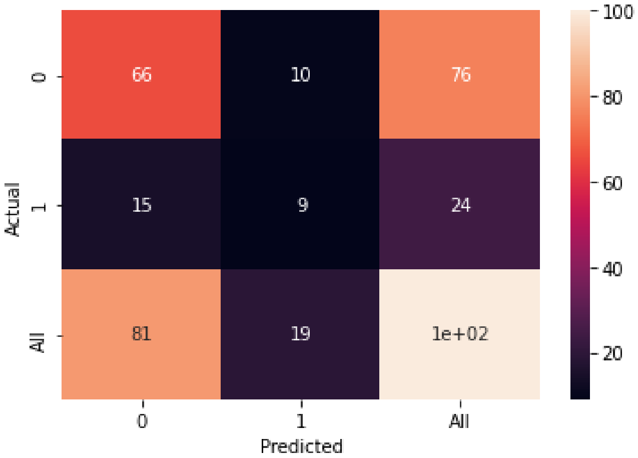


Fig 3 Logistic Regression Confusion matrix using Heat Map

It has the power to handle the large set of a data with their higher dimensionality while performing on large database it works efficiently. It also provides good accuracy with default setting and make easy parallel relatively that will handle missing values and continue to maintain accuracy in large set of data.

In this algorithm; there is case of when there are more trees, which won't allow over-fitting of tree as a model. This method also has a balance and unbalance error of data. It also manages large sets of input variables without its deletion which offers the method for detecting variable interactions. Using confusion matrix shown in Fig 4 predicting accuracy percentage of 81.0 using random forest algorithm.

KNN (K-NEAREST NEIGHBOURS)

KNN refers to number of nearest neighbors that the classifier was uses to make its predictions. KNN is a supervised classification algorithm in which we have some data points or data vectors which is separated into several categories and tries classification prediction in a new sample from that population set. It uses to classify data and provide instance-based learning type in only proximate function are delayed locally while computation.

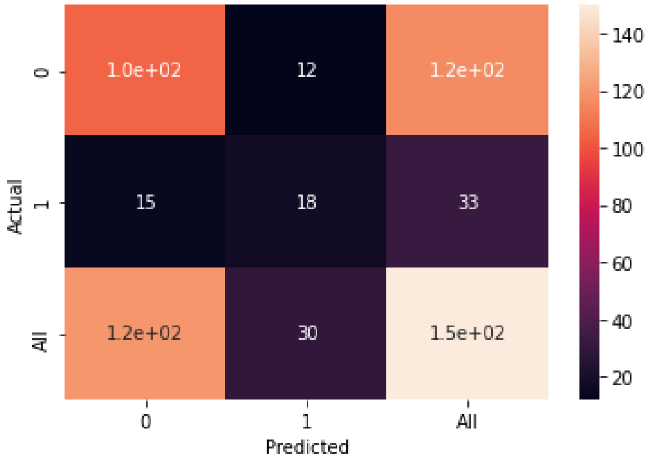


Fig 4 Random Forest Confusion matrix using Heat Map.

KNN works on the concepts of finding the distances between the point on which data query with all the remaining points in the data by selecting the specified k number examples that are close to the query occurred and then vote for the most frequent label identified. It computes Euclidean distance which is a proximity distance and tries to identify who are its neighbors, provides simple and easy implementation which gives Robust to noisy training data and Learns complex models easily. It partitions the data space into n regions number consisting of the collection of points closest to a particular point known as Voronoi Partition Space **[8].** Using confusion matrix shown in Fig 5 predicting accuracy percentage of 75.0 using KNN algorithm.

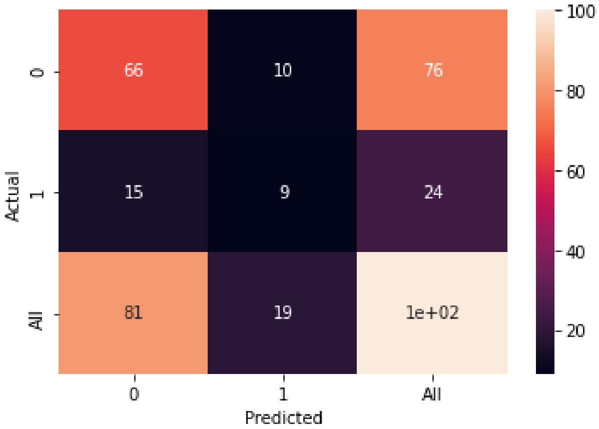


Fig 5 KNN Confusion matrix using Heat Map.

NEURAL NETWORK

Neural Network is learning algorithms within machine learning. NN forms a base of deep learning, a subset of ML. It is stirred by human brain structure. NN assimilate data and train them to distinguish data into different patterns and structures**.** The patterns which they recognize are vector form numerical, into real world data like images, sound, text or time series. This helps to transfer into cluster and classify to formed pixels of each leaf such that it will be broken down depending on its dimensions. Then these pixels are represented into vector matrices and later provided to input layer of neural network. NN have perceptions of accepting into inputs layer and then process it by passing them from the input layer to hidden layer and finally to the output layer. When data input transferred to hidden layer, an initial random weight is assigned to each input node in layer. Further each preceptor is passed through designed activation function which determines whether the preceptor is in activation state or not.

An activated preceptor is then transmitted with data to a next presented layer **[9].** Like this, the data is spread into forward layers through existed neural network interconnection until the preceptor reach to its final expected output layer. On this output layer, probability is computed to decide that whether the data belongs into class A or class B. In case where the predicted output went wrong, we train NN by using Back Propagation method. Initially, while designing the NN at input state with random values. The quality to self-learn from training model examples makes them strong and compatible learner by this neural network learns itself and does not have to programmed every time again. This makes it efficacious in solving numbers of classifications, clustering and regression problems areas using confusion matrix shown in Fig 6 predicting accuracy percentage of 80.66 using Neural Network algorithm.

SVM (SUPPORT VECTOR MACHINE)

SVM is an also comes under supervised ML algorithm used in regression and mostly in classification and. Its implementation process moved from low dimension to higher dimensions data implementation like support vector classifier **[4]** splits the higher dimensions data into multiple groups, if data is in 1-D, SVC work as single point on a 1-D number line. If data is 2-D, SVC results in a line. If data is 3-D, SVC forms a plane. If data is 4-D or more dimensions, SVC is a hyper plane.

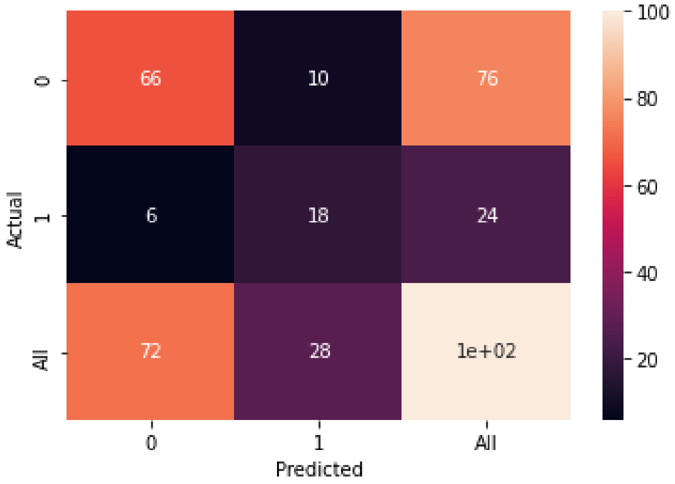


Fig 6 Neural Network Confusion matrix using Heat Map.

To transform the data in a higher dimension (for e.g. 3-D), the boundary is identified and classification is performed. However, computation becomes more expensive when large numbers of dimensions occurred in that space. This is when kernel trick comes into role plays when without computing the coordinates it gives access to work into the original dimensions space of features. To transform data into higher dimensions SVM proposed less expensive and more efficient methods to computed. For example: In polynomial kernel it has degree of polynomial (d) parameter. When d is equal to one the relationship between each pair is computed by kernel for the observation in one Dimension with the relationships defines between them to opt better SVC. When in two-dimension space, polynomial kernel computes 2-D relationships. When in three-dimension polynomial kernel computes 3-D relationships. When d = 4 or more, then we get even more dimensions to identifies best SVC.

System increases dimensions by adjusting value of d in polynomial kernel and relation between each pair of observation are used to find SVC. Radial kernel is the commonly used to find SVC in infinite dimensions. SVM process by shifting the data into high dimension space and identifying relatively high dimension SVC which effectively classify the observations noted. It really works well with a clear separation marked margin and which predict effectively in high dimensional spaces **[10].** It works effectively where the number of samples is less then number of dimensions. Here prediction accuracy become high and provides better generalization performance, for making memory efficient decision function also called support vectors uses subset of training points. Using confusion matrix shown in Fig 7 predicting accuracy percentage of 77.00 using SVM algorithm

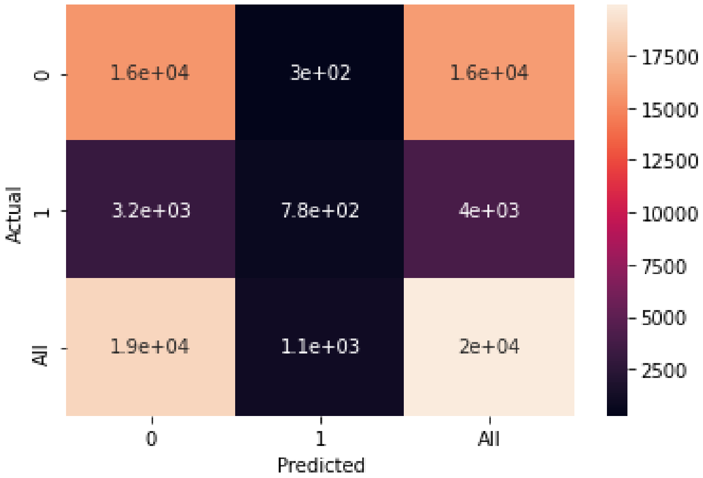


Fig 7. SVM Confusion matrix using Heat Map

GRADIENT BOOSTING CLASSIFIER

Ensembles are also used successfully in another machine learning method such as Boosting. Many trees emerge sequentially when boosted. Each tree attempts to construct a model that effectively forecasts what previous trees were unable to forecast. Finally, the procedure aggregates subsequent models and uses an average or a plurality vote to determine the final prediction. Machine learning prediction models that do not perform well are usually the poor predictors.

In Random Forest, parallelize boosting is not possible for quick execution. When working with big data, though, we can fetch the data piece by piece due to its sequential existence, so that the algorithm still holds the previous estimators in the series. In comparison to the Gradient Boosting classifier, which uses only the deviance loss, which is similar to the cost function of a logistics regression, multiple loss functions are used in Gradient Boosting regression. When there are so many consecutive regression coefficients in a GBM model, it is susceptible to over fitting, and indeed the model begins to match the dimensionality of data. It is critical to check the productivity of the synchronized values of the amounts of explanatory variables and the activation functions. If highest levels are transferred in the largest value parameter, convergence may be delayed significant computational pressure imposed by the GBM protocols. Holding processing rate constant and maximizing multivariate regression and scope with preference to the teration assessment tool, the optimization can be achieved.

The difference between the learning functions of the series occurs because the models depend on both the features and the examples weighted by the values of the vector. So, the principle is not how boosting works but rather the optimization process for getting the weight and the power of the summed functions, which weak learners can't determine. Using confusion matrix shown in Fig 8 predicting accuracy percentage of 80.0 using Gradient Boosting Classifier.

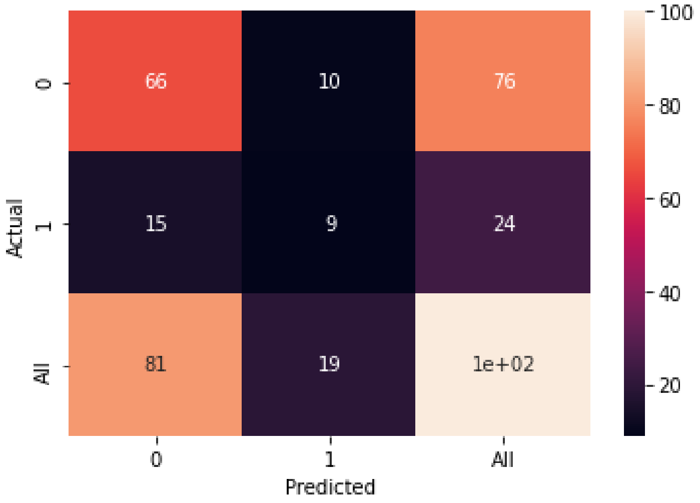


Fig 8 Gradient Boosting Classifier Confusion matrix using Heat Map

3. RESULTS AND DISCUSSION

Machine learning efficiency of various algorithms was tested on different data sets. Different results have been obtained by comparison. From different algorithms computed different accuracy as shown in Table III when the data set was on lower side 50 KNN gives better accuracy comparing with other algorithms, when considering data set of 100 tuple logistic become little better among all, at 500 data tuple random and decision start producing better accuracy.

Table III Average Computation on Different Data Sets

Algorithms	Training Data Size								Average
	50	100	500	1000	5000	10,000	20,000	30,000	
SVM	68	74	79	82	83	79	75	74	76.75
Random Forest	72	74	97	97	97	97	97	97	91.00
Neural Network	68	79	77	69	75	79	80	80	75.88
Decision Tree	56	66	97	97	97	97	97	97	88.00
k-Nearest Neighbor	84	76	80	83	83	79	78	78	80.13
Logistic Regression	76	80	80	80	80	80	80	80	79.50
Gradient Boosting	72	76	80	80	80	80	80	80	78.50

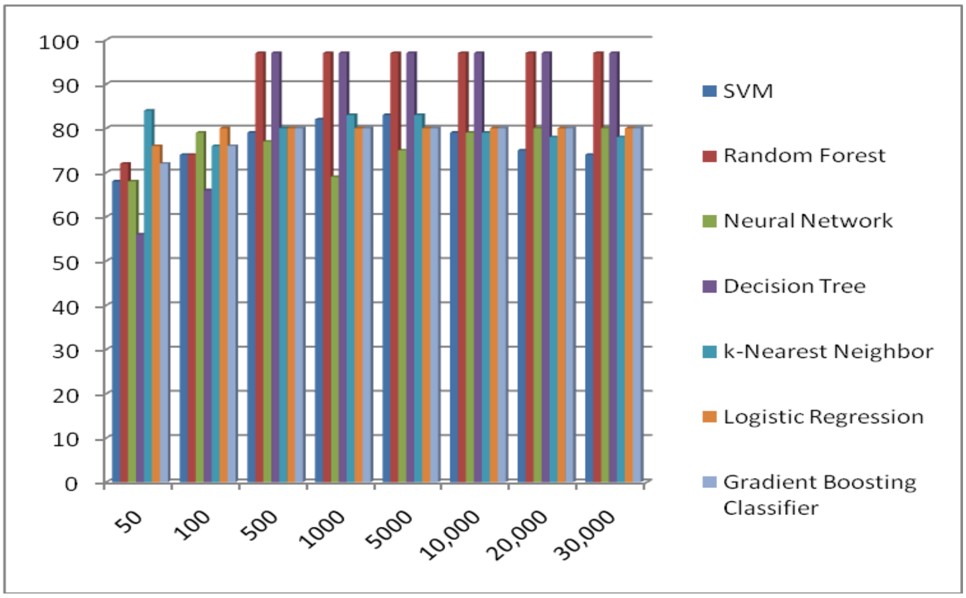


Fig 9. Algorithms comparison on different dataset.

As when compared with remaining algorithms, in all SVM predict less average accuracy of 76.75 %. It has also been observed Decision Tree and ensemble Random Forest predict similar accuracy of 97 % for every different combination of tuples taken from 1000,5000,10000, 20000 to 30000, for same dataset Logistic Regression and Gradient Boosting compute not highest but consistent accuracy of 80 %. Whereas highest accuracy was predicted by Random Forest Algorithm on an average of 91 % accuracy even dataset has been raised from 500 to 30,000 tuples compared with others algorithm.

CONCLUSION

Different algorithm like Support Vector Machine, Random Forest, Neural Networks, Decision Tree, k-Nearest Neighbor, Logistic Regression and Gradient Boosting Classifier used for the study for UCI Credit card value datasets Repository shown in Fig 9. Machine learning assumes significant job in expectation and investigation in various application regions from money to medication, from space science to science utilizing quantities of calculations and methods. Our experimentation has come about similar and its investigation to investigate notable grouping methods utilized for information digging using AI. Arrangement calculation plays significant in getting sorted out the information and helps in information marking.

As KNN requires no training time and training of Neural Networks is time consuming in that feature KNN is better than NN. KNN requires tuning only one hyper parameter (i.e. value of k), while NN involves many parameters involves like size monitoring and controlling of network structure to optimize the procedure. NN is a classifier that consider parameters that manages tuning of hyper-parameters during the training phase of model while SVM is a classifier without classifier which identify a linear vector for class separation for labelling it. Generally, NN outstrip SVM in case of large number of instances in training. In case of multi-classes problems NN compute probabilities for each class to manage different classes while SVM manages by producing output as single binary for independent classifiers of type one-versus-all in same multi class problems.

The investigation has indicated that every procedure assessed distinctive rate precision, as it is hard to recognize one classifier that can characterize all the informational indexes with a similar exactness. Where on the off chance that Logistic Regression performs better when amount of clamor factors is not exactly illustrative factors numbers, then again Random Forest processed higher genuine positive and bogus positive rate as dataset increments in numbers. As Decision Tree separate the space into a little district and Logistic Regression fits single lines to partition a space precisely into two and has straight single choice limit.

REFERENCES

[1]. Abdou, Hussein A. and Pointon, John. Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of The Literature in Intelligent Systems in Accounting, Finance and Management, Vol.18, pages 59-88, 2011

[2]. O'Leary, D.E., ‘Big Data’, The ‘Internet of Things’ And The ‘Internet of Signs’. Intell. Sys. Acc. Fin. Mgmt., 20: 53-65, 2013.

[3]. Sahin, Y., Duman, E. Detecting Credit Card Fraud by Decision Trees and Support Vector Machines, in Proceedings of the International Multi Conference of Engineers and Computer Scientists Hong Kong, 2011.

[4]. Huang, S. Y. Fraud Detection Model by Using Support Vector Machine Techniques, Chiayi, Taiwan: International Journal of Digital Content Technology & its Applications, 2013.

[5]. Ehramikar, S. The Enhancemeat of Credit Card Fraud Detectioa Systems using Machine Learning Methodology, Master of Applied Science Thesis, University of Toronto, 2000.

[6]. John, O. A., Adebayo, O. A., Samuel, A. O. Credit card fraud detection using machine learning techniques: A comparative analysis, ICCNI 2017: International Conference on Computing, Networking and Informatics, Lagos, Nigeria2017.

[7]. Rajamani, R., Rathika, M. Credit Card Fraud Detection using Hidden Morkov Model and Neural Networks, in Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications, 2015.

[8]. Anohhin, I. Data mining and machine learning for fraud detection, Master thesis, Tallinn, 2017.

[9]. Padvekar SA, Kangane PM, Jadhav KV Credit card fraud detection system. Int J Eng Comput Sci 5(4):16183–16186, 2016.

[10]. Khare N, Sait SY Credit card fraud detection using machine learning models and collating machine learning models. Int J Pure Appl Math 118(20):825–838, 2018.

[11]. Banerjee R, Bourla G, Chen S, Kashyap M, Purohit S, Battipaglia J Comparative analysis of machine learning algorithms through credit card fraud detection. New Jersey’s Governor’s School of Engineering and Technology, Piscataway, pp 1–10, 2018.

[12]. Mishra A, Ghorpade C (2018) Credit card fraud detection on then skewed data using various classification and ensemble techniques. In: 2018 IEEE International students’ conference on electrical, electronics and computer science, SCEECS, 2018.

[13]. Xuan S, Liu G, Li Z, Zheng L, Wang S, Jiang C Random forest for credit card fraud detection. In: ICNSC 2018-15th IEEE International conference on networking, sensing and control, pp 1–6, 2018.

[14]. Hordri NF, Yuhaniz SS, Azmi NFM, Shamsuddin SM Handling class imbalance in credit card fraud using resampling methods. Int J Adv Comput Sci Appl 9(11):390–396, 2018.

[15]. Marques, Bernardo P. and Alves, Carlos F. Using clustering ensemble to identify banking business models, in Intelligent Systems in Accounting, Finance and Management}, Vol. 27, number 2, pages 66-94, 2020.