iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** III **Month of publication:** March 2022

DOI: https://doi.org/10.22214/ijraset.2022.40702

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

Implementation of Credit Card Fraud Detection Using Random Forest Algorithm

K. Deepika[1], M. Pavan Sai Nagenddra[2], M Vamshi Ganesh[3], N. Naresh[4]
[1, 2, 3, 4]*Electronics and Communication Engineering Dept, TKR College of Engineering and Technology, Hyderabad, Telangana, India*

Abstract: ***Credit card fraud processing is presently the most frequently arising problem in the present world. This is due to the rise in both online transaction and ecommerce platforms. To detect these fraudulent activities the credit card fraud detection system was introduced, this project main aim is to focus on the machine learning algorithms. The voting based classification algorithm approach is applied for credit card fraud detection. We use different types of classification algorithms such as SVM, Naïve bayes and Random forest. We consider their results based on confusion matrix for the above classification algorithms. We analyze their performance based on accuracy, precision, recall and f1-score. We compare random forest algorithm with other algorithm. We considered random forest algorithm has greatest accuracy, precision, recall and F1-score, considered as the best algorithm that is used to detect the fraud.***
Keywords: Fraud detection, Naive Bayes, SVM, and Random Forest.

I. INTRODUCTION

Card payments are very popular payment method in nowadays. Card payments are quite easy to perform on merchant side by presenting credit card or on internet by announcing credit card details: number, expiring date and security code. As result of low level of security card payments are influence of fraudulent abuse. Also, another important reason is increase in mobile devices use for initialization of payments. In this study Naïve Bayes, SVM, random forest machine learning classifiers are performed on the dataset that contains transactions made by credit cards in September 2013 by European cardholders. This dataset contains 284.807 transactions where 492 are fraud. In dataset there are 31 features where 31st is a binary variable with 0 regular transactions and 1 as fraud transaction. The performances of algorithms are evaluated through following performance matrices: precision, recall and precision-recall, f1-sore and accuracy.

II. PROBLEM DEFINITION

Credit card fraud is a sober and major growing problem in banking industries. With the advent of the rise of many web services provided by banks, banking frauds are also on the increase. Banking systems always have a strapping security system in order to detect and prevent fraudulent activities of any category of transactions. Totally eliminating banking fraud is almost unfeasible, but we can however minimize the frauds and prevent them from happening by machine learning techniques like Data Mining. It represents how to utilize these data and find useful information from data has become an urgent need for detection of fraud. Therefore, data mining technology has become an effective method for detection of fraud. Thus we are developing a fraud detection system for credit cards using decision tree induction algorithm for security using Data Mining Technique

III. METHODOLOGY

In this project we investigate through Random forest, SVM, naïve bayes models and test its performance from accuracy, recall, precision and f1 score.

A. NAÏVE BAYES

This algorithm learns the probability of an object with certain feature belonging to a particular class. This algorithm assumes that the probabilities of individual feature are independent of each other which are quite hard to happen in this real world it is the reason to be called naïve. It can be described as probability of an event will occur based on another which has already occurred. It can be written as

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

$$P(A/B) = \frac{P\left(\frac{B}{A}\right).P(A)}{P(B)}$$

1) P (A/B): Conditional probability- Probability of occurrence of event A given the event B is true.

2) P (B/A): Likelihood probability- Probability of the occurrence of event B given the event A is true.

3) P (A), P (B) are occurrence of event A and B respectively.

B. Support Vector Machine (SVM)

It is supervised learning algorithms, which is used for both classifications as well as regression problems. It is mainly used for classification problems in Machine learning. SVM classifies the two classes using hyper plane. This hyper plane should have the largest margin in a high dimensional space to separate given data into classes. The margin between the 2 classes represents the longest distance between closest data points of those classes.

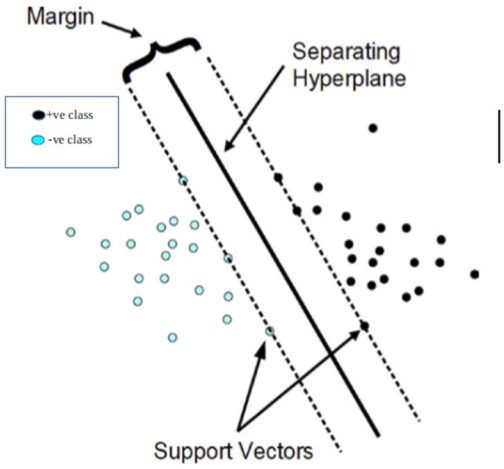


Figure: Support Vector Machine

C. Random Forest

It is a machine learning technique that constructs multiple decision trees. The final decision is made based on the outcome of the majority of the decision trees. It can be used for both regression and classification purposes. But this algorithm is mainly used for classification purpose. This algorithm creates decision trees on the sample data and gets the prediction from each sample data. It is called as ensemble method. This algorithm is better than the single decision tree because it reduces the over-fitting by averaging the results.

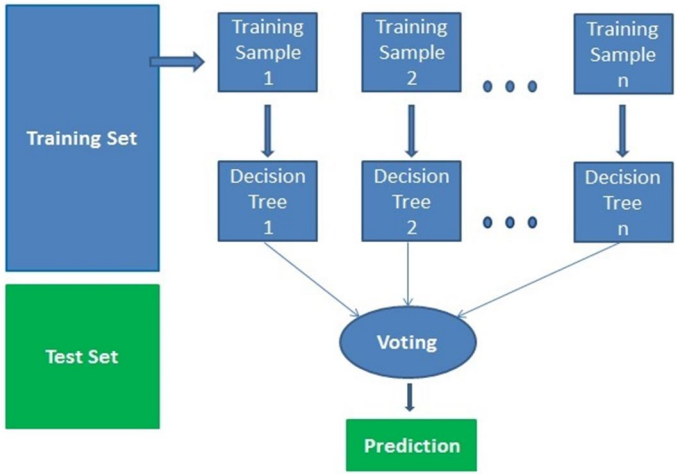


Figure: Random Forest

D. Confusion Matrix

It is a specific table that is used to measure the performance of the algorithm. It is used to summarize the performance of the classification algorithm. It shows the error in performance of algorithm in the form of matrix hence it's called as error matrix. The matrix is based on actual and predicted parameters.

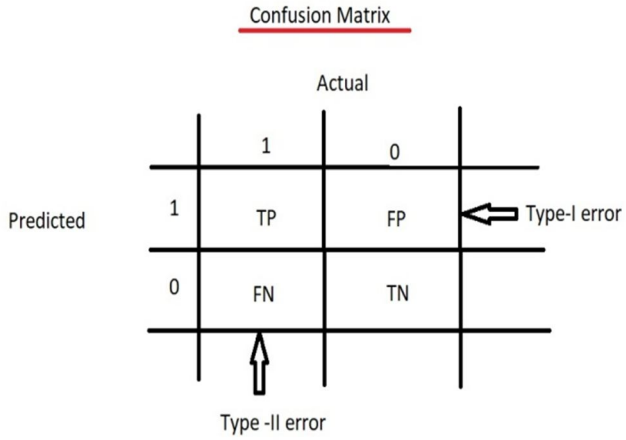


Figure: Confusion matrix

1) *True Positive (TP):* The model has predicted YES, and the actual value was also true (YES).
2) *False Positive (FP):* The model has predicted YES, but the actual value was NO. It is also called as Type-I error.
3) *False Negative (FN):* The model has given prediction NO, and the actual value was YES, it's also called as Type-II error.
4) *True Negative (TN):* The model has given prediction NO, and the actual value was also NO.

We can perform various calculations using this matrix such as accuracy, precision etc,

E. Accuracy

It is the proportion of correct classifications (true positives and negatives) from overall number of cases. It defines hoe often the model predicts the correct output. The formula for accuracy is given below

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

F. Precision

It is the proportion of correct positive classification (true positives) from cases that are predicted as positive. It can be calculated using the below formula.

$$\textbf{Precision} = \frac{TP}{TP+FP}$$

G. Recall

It is the proportion of correct positive classifications (true positives) from cases that are actually positive. It can be calculated using this formula.

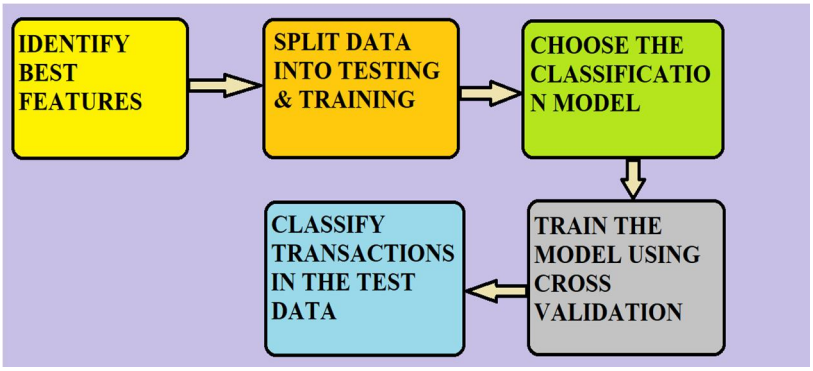
$$\textbf{Recall} = \frac{TP}{TP+FN}$$

H. F1-Score

It is defined as the harmonic mean between precision and recall. F1 –score is maximum if the recall and precision values are equal. It can be calculated using this formula.

$$\textbf{F1-score} = \frac{2(Precision*Recall)}{Precison+Recall}$$

IV. FUNCTIONAL DIAGRAM



1) *Identify the Best Feature:* Selecting the best feature among the large number of feature. We cannot build the best model when there is large number of features. We should build the simplest model possible.
2) *Splitting the Data Into Training and Testing Data:* When we train the model we should ensure that testing data should not disturb the training process.
3) *Choose the Classification Model:* To classify the transaction as genuine and fraud we have selected naïve bayes classification model.
4) *Train the Models using the Cross Validate:* We train the model using the dataset with process called cross validation. Cross validation is process is used to improve the performance of model over fixed train and test split of dataset.
5) *Classify Transaction in the Testing Data:* We use testing data to classify transaction into genuine and fraud.

V. RESULTS

The dataset when applied for different algorithms gives us different outputs. The random forest algorithm is applied for the dataset and the results are obtained as below:

A. *Testing Data Ouput (Random Forest)*

The accuracy of testing data is 100%, the recall value is 1.00 for genuine and 0.78 for fraud transactions (100% for genuine and 78% for fraud). The precision value of genuine transaction is 100% and for fraudulent transaction is 91%. F1-score for genuine transaction is 100% and for fraud transaction is 84%.

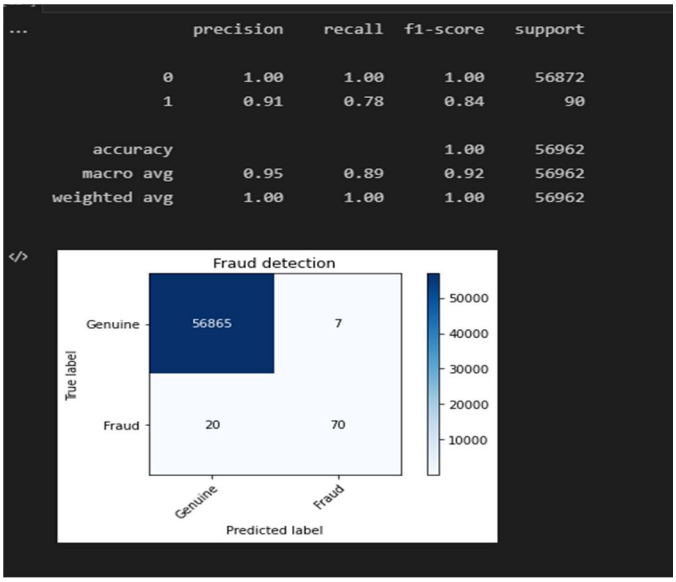


Figure: Confusion matrix for Random forest

B. Training Data Ouput (Random Forest)

The accuracy of training data is 100%, the recall value is 1.00 for genuine and 0.99 for fraud transactions (100% for genuine and 99% for fraud). The precision value of genuine transaction is 100% and for fraudulent transaction is 99%. F1-score for genuine transaction is 100% and for fraud transaction is 99%.

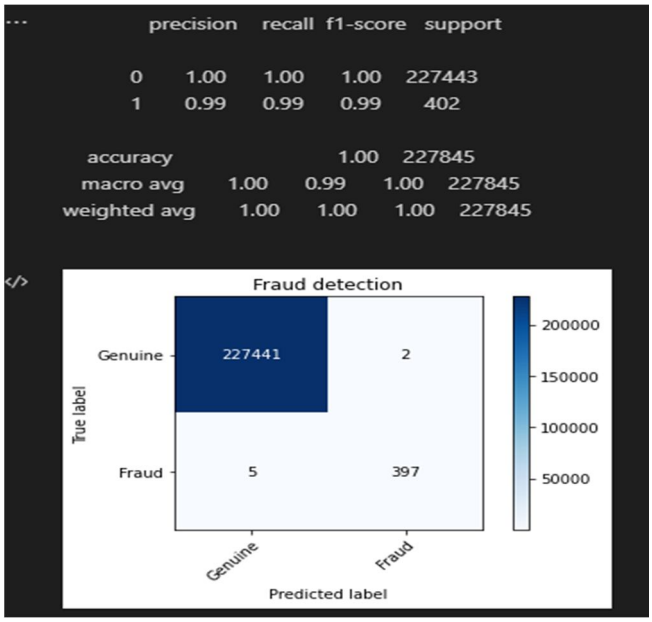


Figure: Confusion matrix for Random forest

The SVM algorithm is applied for the dataset and the results are obtained as below:

C. Testing Data Ouput (SVM)

The accuracy of testing data is 100%, the recall value is 1.00 for genuine and 0.66 for fraud transactions (100% for genuine and 66% for fraud). The precision value of genuine transaction is 100% and for fraudulent transaction is 96%. F1-score for genuine transaction is 100% and for fraud transaction is 79%.

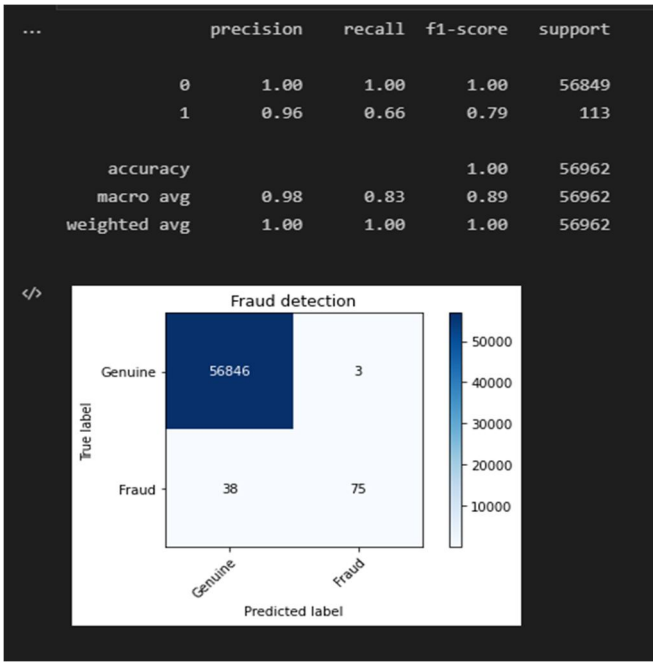


Figure: Confusion matrix for SVM

D. Training Data Output (SVM)

The accuracy of training data is 100%, the recall value is 1.00 for genuine and 0.81 for fraud transactions (100% for genuine and 81% for fraud). The precision value of genuine transaction is 100% and for fraudulent transaction is 99%. F1-score for genuine transaction is 100% and for fraud transaction is 89%.

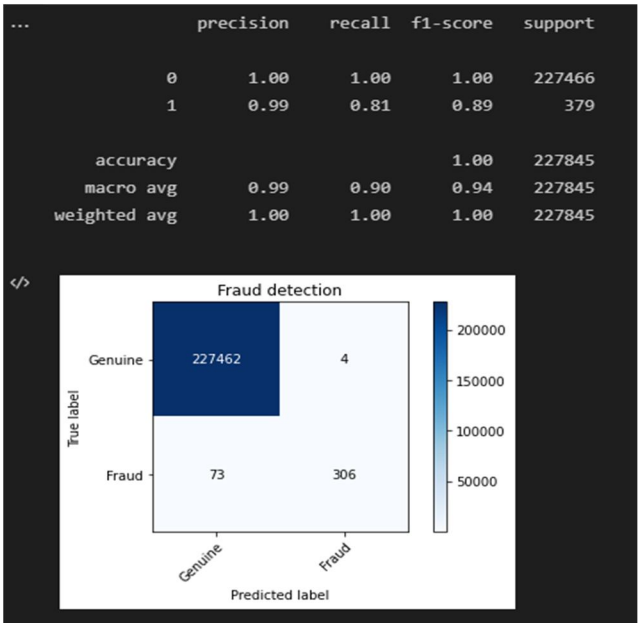


Figure: Confusion matrix for SVM

The naïve bayes algorithm is applied for the dataset and the results are obtained as below:

E. Testing Data Output (NAÏVE BAYES)

The accuracy of testing data is 99%, the recall value is 0.99 for genuine and 0.88 for fraud transactions (99% for genuine and 88% for fraud). The precision value of genuine transaction is 100% and for fraudulent transaction is 14%. F1-score for genuine transaction is 99% and for fraud transaction is 24%.

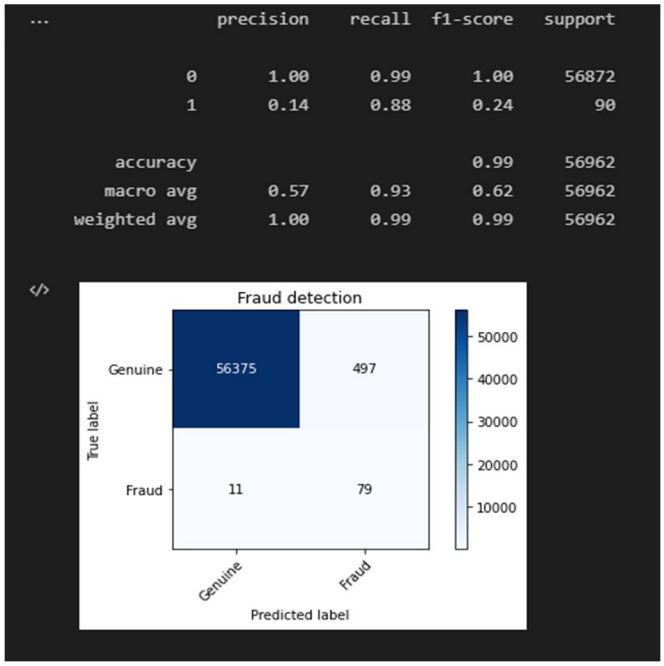


Figure: Confusion matrix for Naïve bayes

F. Training Data Output (NAÏVE BAYES)

The accuracy of training data is 99%, the recall value is 0.99 for genuine and 0.84 for fraud transactions (99% for genuine and 84% for fraud). The precision value of genuine transaction is 100% and for fraudulent transaction is 15%. F1-score for genuine transaction is 99% and for fraud transaction is 25%.

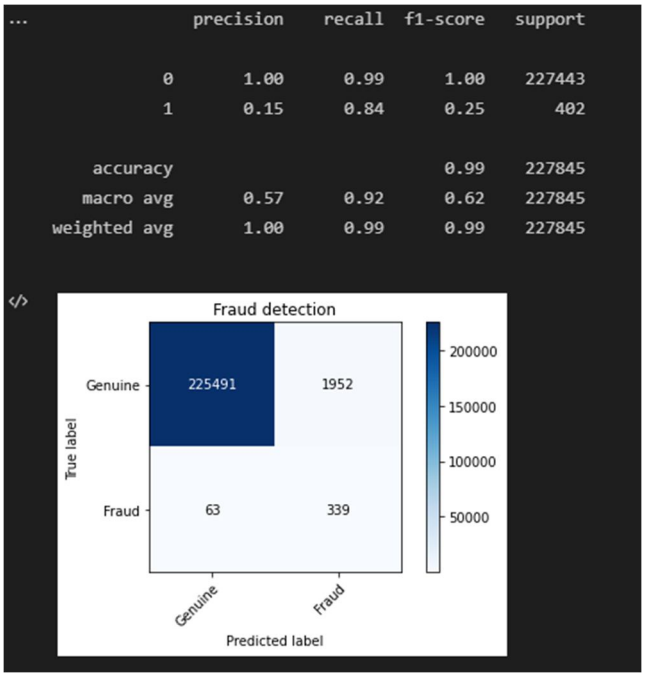


Figure: Confusion matrix for Naïve bayes

VI. CONCLUSION

From our analysis we can conclude that accuracy of random forest and SVM are similar. The accuracy of naïve bayes algorithm is a bit low than the other two algorithms. On average the precision, recall and f1-score of Random forest algorithm has the highest value than the other two algorithms. Hence we conclude that the Random forest algorithm works best than SVM and naïve bayes to detect the credit card fraud.

Algorithms	Accuracy	Recall	Precision	F1-score
Random forest	100%	78%	91%	84%
Support Vector Machine	100%	66%	96%	79%
Naïve bayes	99%	88%	14%	24%

REFERENCES

[1] World line and the Machine Learning Group, 'Credit Card Fraud Detection at Kaggle', Credit Card Fraud Detection Dataset, 2013.

[2] Admel Husejinović, (January 2020), Credit card fraud detection using naïve Bayesian and C45 decision tree classifiers.

[3] P. R. Shimpi, 'Survey on Credit Card Fraud Detection Techniques', Int. J. Eng. Computer. Sci., 2016.

[4] S. N. John, C. Anele, O. O. Kennedy, F. Olajide, and C. G. Kennedy, 'Realtime Fraud Detection in the Banking Sector Using Data Mining Techniques/Algorithm', in 2016 International Conference on Computational Science and Computational Intelligence (CSCI), 2016, pp. 1186–1191.

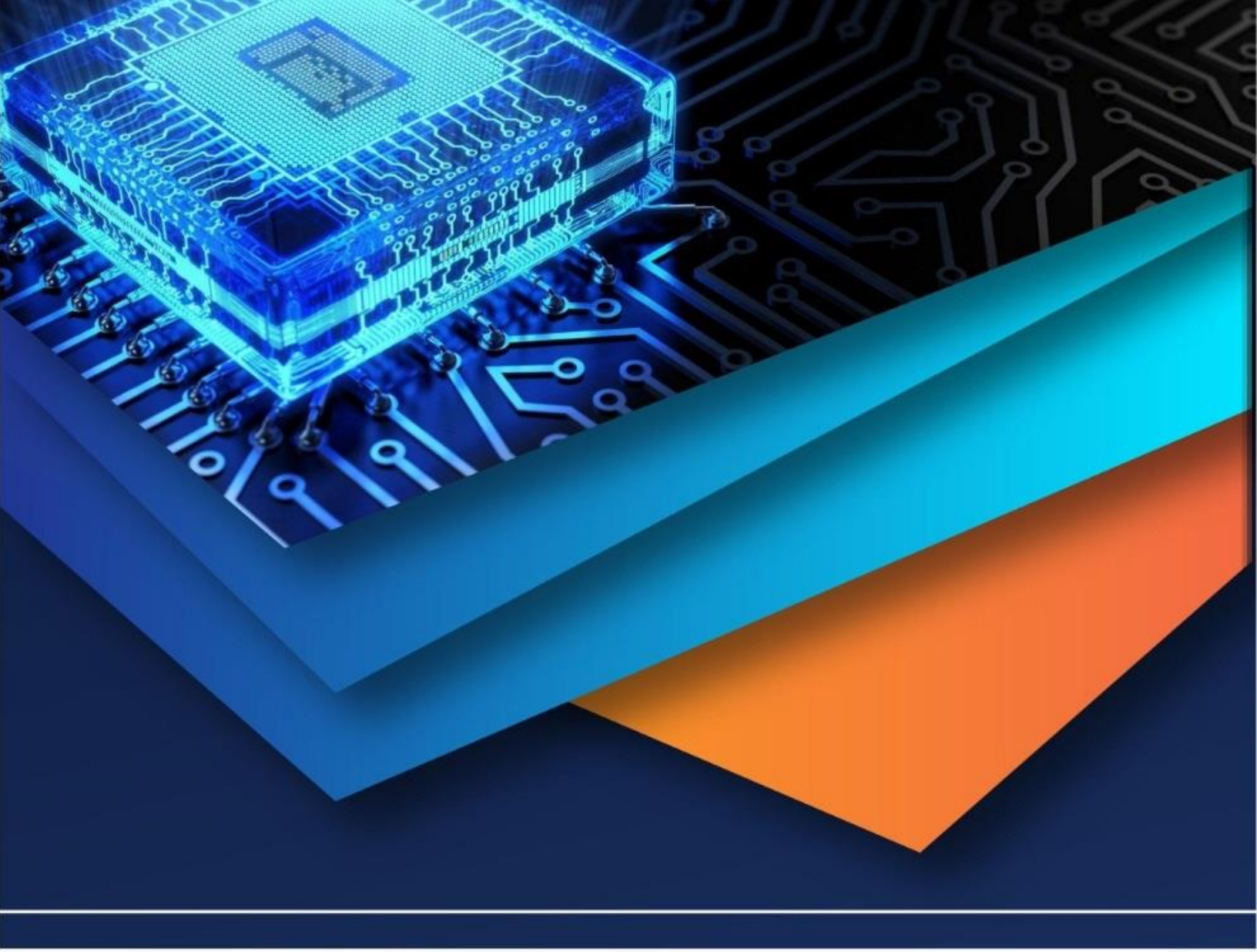
AUTHORS BIODATA

N.NARESH received his B. Tech degree in Electronics and Communication Engineering from ANRK KODAD in 2011 and he completed his M. Tech from VBIT HYDRABAD in 2014. Presently he is working as Assistant Professor in TKR college of Engineering and Technology, HYD.

K.DEEPIKA, pursuing B. Tech final year in Electronics and Communication Engineering from TKR college of Engineering and Technology in 2021-2022, HYD

M PAVAN SAI NAGENDRA, pursuing B. Tech final year in Electronics and Communication Engineering from TKR college of Engineering and Technology in 2020-2021, HYD.

M VAMSHI GANESH, pursuing B. Tech final year in Electronics and Communication Engineering from TKR college of Engineering and Technology in 2021-2022, HYD.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY