iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VII **Month of publication:** July 2020

DOI: https://doi.org/10.22214/ijraset.2020.30734

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: 7.429
Volume 8 Issue VII July 2020- Available at www.ijraset.com

Analysis of Vehicle Insurance Data to Detect Fraud using Machine Learning

P Sai Pranavi [1], Sheethal H D[2], Sharanya S Kumar[3], Sonika Kariappa[4], Swathi B H[5]
[1]*Computer Science Department, Vidyavardhaka College Of Engineering, Mysuru, India*

Abstract: ***There are more than thousands of insurance companies in the world which handles large amount of data. Insurance fraud has become the most common among the organizations across the industries. Insurance industry is one among the most growing industries hence fraud detection becomes very important in this current world. Fraud detection can be implemented in various fields like banking, insurance, financial sectors and information security systems. There are many approaches by which fraud can be detected. In this paper, we use various techniques of machine learning for the detection of vehicle insurance fraud. We make use of Random forest and KNN algorithm for the accuracy detection the insurance fraud. The performance calculation is done by calculation of confusion matrix.***

Index Terms: Fraud, Fraud detection, Machine Learning, Insurance, KNN, Random forest, Confusion Matrix.

I. INTRODUCTION

Fraud is a criminal misrepresentation which causes hindrance for both the individual and the organization. [1] The insurance industries are now implementing various techniques for the effective management of fraud. There are two major types of fraud, hard insurance fraud and the soft insurance fraud. When few people intentionally fake an accident then that type of fraud is known as the hard insurance fraud. And When the person has insurance claim that is valid but falsifies the part of the claim is known as soft insurance fraud. So the organisations should implement various techniques that are helpful for the fraud detection to increase the customer satisfaction. [4]

When the number of undetected fraud cases increases then the premium amount also increases to compensate the losses, which in turn affects the insured parties. [12] With the increase in the number of fraud cases, we can detect the frauds by implementing various techniques with the help of the data obtained from many other similar cases. Searches should be possible by using data innovation as an answer for discover an example and afterward recognize misrepresentation that happens dependent on the information of vehicle protection throughout the years. [5]

Insurance fraud detection is a difficult issue, given the assortment of misrepresentation designs and moderately small proportion of known frauds in regular examples. While building identification models, the investment funds from misfortune anticipation should be offset with cost of false alarms. Various machine learning methods consider improving prescient precision, empowering misfortune control units to accomplish higher inclusion with low false positive rates. [8] In this paper, numerous machine learning strategies for misrepresentation location are introduced and their exhibition on different data collection sets are analysed. The feature engineering impact, parameter tweaking and feature selection are investigated with the objective of obtaining prevalent prescient execution. [11]

Besides, the extortion specialists may confront numerous unfavourable circumstances while detecting the vehicle protection misrepresentation cases for the most part happen because of two reasons. Firstly, any absent or wrong case data makes the extortion identification testing challenging. Furthermore, it is additionally discovered that the quantity of noxious cases is significantly less than the absolute cases submitted. This uneven dispersion (information imbalance) prompts progressively troublesome extortion recognition. Moreover, the majority of the supervised classifiers create inefficient classification models with unequal information, since they tend to order all the information focuses as certifiable class (significant class tests) and overlook the deceitful focuses (minority class tests) [3].

Insurance frauds spread the scope of inappropriate exercises which an individual may submit so as to accomplish a good result from the insurance company. This could run from organizing the occurrence, distorting the circumstance including the applicable on-screen characters and the reason for incident and lastly the degree of damage caused. In this paper we use various techniques of machine learning from the detection of fraudulent cases. The feature selection algorithm is used for the selecting the most target variables. KNN and Random Forest are the other algorithms used for the fraud detection, hence they provide the accurate results for the given dataset.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VII July 2020- Available at www.ijraset.com

II. LITERATURE SURVEY

A literature survey takes the current and past theories behind the subjects, which is being examined. It is a representation of these theories and information aimed to give the researchers an idea of desertation before they present their findings. We try to present a few words on those technologies and tools which helped us to develop our project.

Sharmila Subudhi, et al. proposed fraud detection methodology in the vehicle insurance area with the help of an adaptive oversampling method (ADASYN).

By employing the ADASYN the data imbalances can be removed from the original claim dataset. Using different classifiers, the anomalous records are classified from the normal ones. The outcomes of the proposed algorithm justify the efficiency of the balanced data set over an unbalanced one [3].

Tessy Badriyah, et al. proposed detection algorithm to detect the fraud. It develops prediction modelling in the field of detection to detect the fraud using Nearest Neighbour based Method (distance and density based) and Statistics Methods (interquartile range). The results obtained from them are compared with that of the other results performed by others using the same dataset. Hence, then they determine the experiment results obtained are superior in some cases [5].

Aisha Abdallah, et al. proposed the collaboration of the FDS with the FPS for the control and reduction the E-commerce systems. There are many issues and challenges that effect the performance of the FDS like concept drift, supports real time detection, and many. Hence this paper aims to provide a systematic overview of these issues and challenges that block the FDS performance. Further state of the art FDS is selected in the E-commerce systems. [12]

StijnViaene, et al. proposed various methods for the detection of fraud claims. The insurers use automatic detection systems that helps the users decide whether they have to conduct an investigation on fraudulent claim detection. The model incorporates screening, examination, arrangement of case stages. It is actualised in safety net providers case for taking care of procedures. Claim handling is a procedure that starts from claim occurrence and closes with the payment for the damages caused [13].

Ali Ghorbani, et al. [4] utilized a portion of the data mining procedures for extraction of information and examples on huge data to lead the insurance industry. K-means clustering procedure is applied on informational collection with Euclidean separation.

Rekha Bhowmik [11] analyses fraud detection techniques to predict fraud patterns from the data. Data mining techniques are associated with supervised learning and unsupervised learning. Algorithms are utilized to isolate the information and for descriptive classification rule that can be utilized for new instance.

Dongxu, et al. [6] CoDetect fraud detection framework is used which can detect the fraud as well as the fraud activities patterns. Give and set up a way to deal with weighted chart in the financial network and afterward joins the properties of links and hubs. Shows the various situations in the financial fraud.

Xinxin Jiang, et al. [7] The insurance datasets are more likely imbalanced and heterogeneous. And the algorithms can be applied to those which are balanced and homogenous. Therefore, a parallel neural network has been proposed for such datasets which are imbalanced and heterogeneous.

Ke Nian, et al. [9] unsupervised spectral ranking for anomaly of interdependence relation to predict the fraud. Data Mining and Machine learning are used to detect fraud cases and can reduce the economic losses and Predictive strategy expand the detection rate, limit the false positive rate and can quickly distinguish and develop the fraud plans.

Sebastián M. Palacio, et al. [2] semi supervised procedures and another metric that uses the cluster score which can be utilized for fraud detection which can manage the pragmatic difficulties. The principle strategy incorporates transposing unaided models into managed models utilizing the cluster score metric.

Our survey is mainly based on the Automobile Fraud Detection. There are many data mining and machine learning techniques. After going through list of paper we feel safe to say that there are many technologies used for Automobile Insurance Fraud Detection with its own advantage and drawbacks.

III. IMPLEMENTATION

In this proposed system we deal with the vehicle insurance fraud detection using machine learning algorithms. The auto insurance fraud is the most eminent insurance fraud among the various other types of frauds. The fraud users claim insurance using fake accident reports. Hence fraud detection becomes an important task for the organizations to avoid losses. In this system, we focus on detection of auto insurance fraud by using, various machine learning technique.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VII July 2020- Available at www.ijraset.com

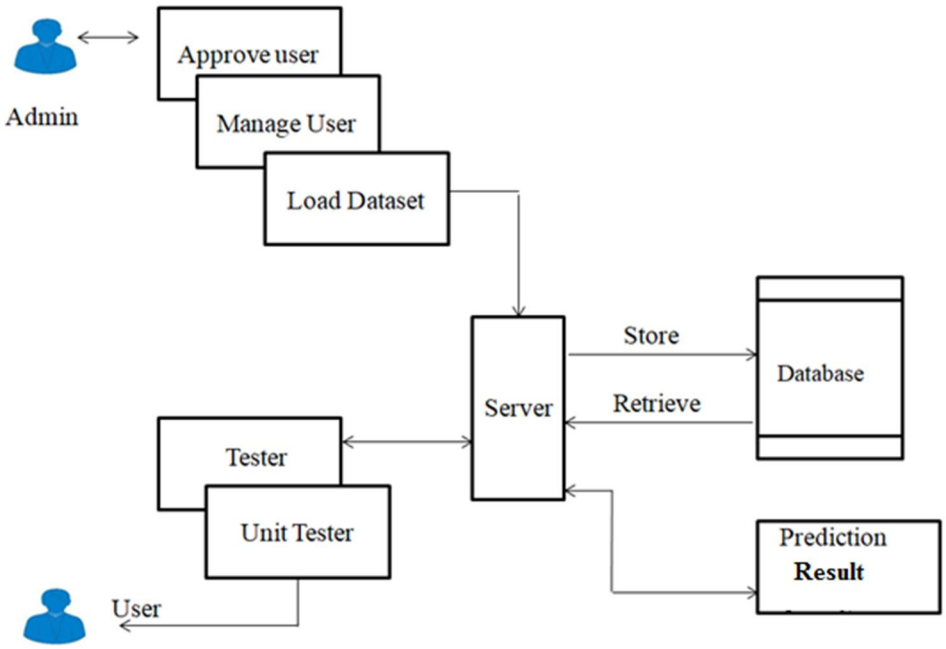


Figure 1: Architecture of the proposed system

Figure 1 represents the architecture of the proposed system for the auto insurance fraud detection which shows how the user and admin works. To build up a more profound comprehension, it merits experiencing the overall work process of the AI procedure, the procedure comprises of 5 phases:

1) *Information Intake:* Right away, the dataset is stacked from the record and is spared in memory.
2) *Information Change:* At this point, the information that was stacked at stage 1 is changed, cleared, and standardized to be reasonable for the calculation. Information is changed over so it lies in a similar range, has a similar organization, and so forth. Now highlight extraction and choice, which are examined further, are proceeded too. Notwithstanding that, the information is isolated into sets – 'preparing set' and 'test set'. Information from the preparation set is utilized to construct the model, which is later assessed utilizing the test set.
3) *Model Training:* At this stage, a model is assembled utilizing the chose calculation.
4) *Model Testing:* The model that was constructed or prepared during stage 3 is tried utilizing the test informational collection, and the delivered outcome is utilized for building another model, that would think about past Models, for example "learn" from them.
5) *Model Deployment:* At this stage, the best model is chosen (either after the characterized number of cycle or when the required outcome is accomplished). The Figure 2 speaks to the overall work process procedure of the 5 phases: Data consumption, Data change, Model preparing, Model testing, Model deployment.

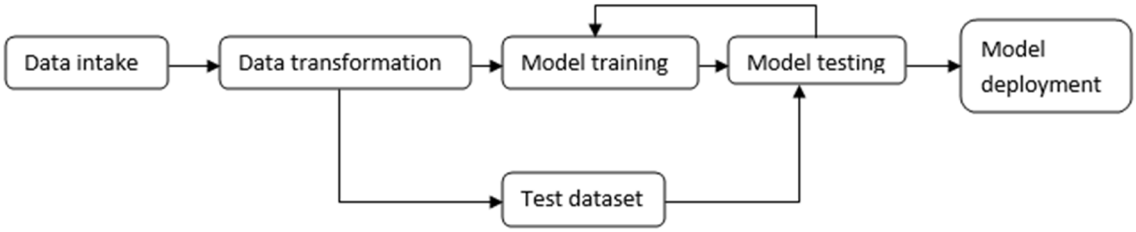


Figure 2 : General workflow process

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VII July 2020- Available at www.ijraset.com

IV. ALGORITHMS USED

A. *Select K Best*

Feature determination is a technique where we pick those highlights in our information that contribute most to the objective variable. As it were, we pick the best indicators for the objective variable. The classes in the sklearn. feature_selection module can be utilized for include choice/dimensionality decrease on test sets, either to improve estimators' exactness scores or to help their exhibition on exceptionally high-dimensional datasets. The benefits of this is it lessens overfitting, improve exactness and decreases preparing time.

Feature selection strategy: SelectKBest

Score work:

For regression: f_regression, mutual_info_regression

For classification: chi2, f_classif, mutual_info_classif

SelectKBest has a default conduct actualized, so you can compose select = SelectKBest () and afterward call select.fit_transform(X, y) For this situation SelectKBest utilizes the f_classif score work.

The SelectKBest class just scores the highlights utilizing a capacity (for this situation f_classif yet could be others) and afterward "evacuates everything except the k most noteworthy scoring highlights".

B. *K-Nearest Neighbours*

K-Nearest Neighbours (KNN) is one of the most least difficult, precise AI calculations. KNN is a non-parametric computation, suggesting that it doesn't make any assumptions about the data structure. In genuine issues, data now and again consents to the overall speculative notions, making non-parametric counts a conventional response for such issues. KNN model depiction is as direct as the dataset – there is no learning required, the entire getting ready set is taken care of. KNN can be used for both classification and regression problems. In the two issues, the estimate relies upon the k getting ready events that are closest to the information model. In the KNN classification issue, the yield would be a class, to which the data model has a spot, foreseen by the predominant part vote of the k closest neighbours. In the regression problem, the yield would be the property estimation, which is regularly a mean estimation of the k nearest neighbours. The nearest neighbour division can be resolved using Euclidean Distance:

Euclidean Distance = $\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$; p and q are the points in n-space (1)

The estimation of k assumes a vital job in the expectation precision of the calculation. Smaller estimations values of k will probably bring about lower precision, particularly in the datasets with much noise, since each case of the preparation set presently has a higher weight during the decision process. Larger the estimation values of k bring down the exhibition of the calculation. In addition, if the estimation value is excessively high, the model can over fit, making the class limits less unmistakable and bringing about lower exactness once more or resulting in the lower accuracy. As an overall methodology, it is encouraged to choose k using the formula:

$$k = \sqrt{n} \quad (2)$$

The algorithm for k nearest neighbour is

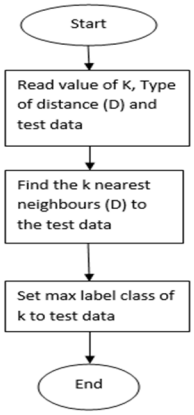


Figure 3: Flowchart of KNN Algorithm

C. Random Forest

Random Forest is one of the most known machine learning algorithm. It requires basically no information planning and demonstrating or modelling yet for the most part brings about precise outcomes. Random Forests depend on the decision trees portrayed in the past segment. All the more explicitly, Random Forests are the assortments of decision trees, creating a superior forecast exactness. That is the reason it is known as a 'forest' – it is essentially a lot of decision trees. The fundamental thought is to develop various decision trees dependent on the autonomous subsets of the dataset. At every node, n factors out of the list of capabilities are chosen arbitrarily, and the best split on these factors is found. In this venture, among KNN and random forest calculations, whichever calculation gives an increasingly precise outcome will be utilized to foresee the extortion in the protection information. The calculation for Random forest is

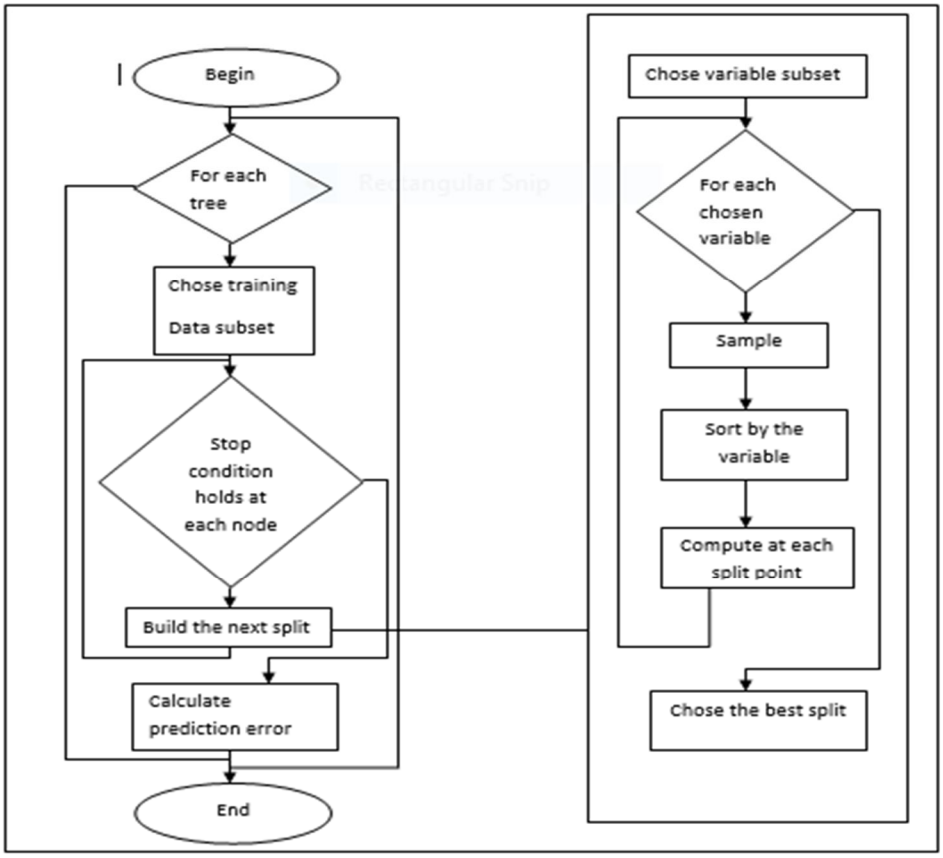


Figure 4: Flowchart of Random Forest Algorithm

V. RESULTS

As fraud poses a serious problem in the current society, it has to be resolved. In order to resolve these problems, we can build systems which predict fraud in the data given. These systems are built using various machine learning techniques like naïve Bayes, KNN, random forest, neural networks. In this paper we have discussed about various ML techniques and how it is implemented in the systems and how accurate it is in predicting the fraud. Later these techniques are compared using five criteria from different perspectives.

In Random Forest, training data is chosen randomly. Each trained tree gives its own classification result so this analyzes the missing data and calculates the errors. KNN algorithm stores the data for further classification instead of making calculations of the data. Among KNN and random forest algorithms, whichever algorithm gives a more accurate result will be used to predict the fraud in the insurance data. The system identifies whether the claim is fraud or not by considering the information given by users. Using Random Forest and KNN algorithms insurance fraud is predicted accurately. The information obtained from the user is evaluated with the dataset and tells the users whether the claim is accepted or rejected. For classification problems while using KNN algorithm with an even number of classes, it is advised to choose an odd *k* since this will eliminate the possibility of a tie during the majority vote. The Figure 3 shows the comparison graph for the algorithms KNN and Random Forest:

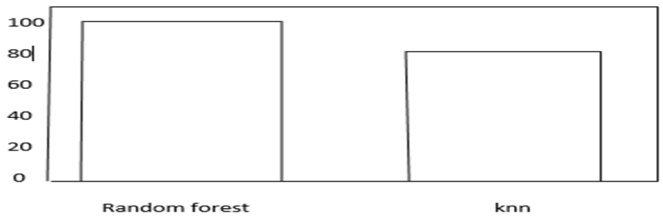
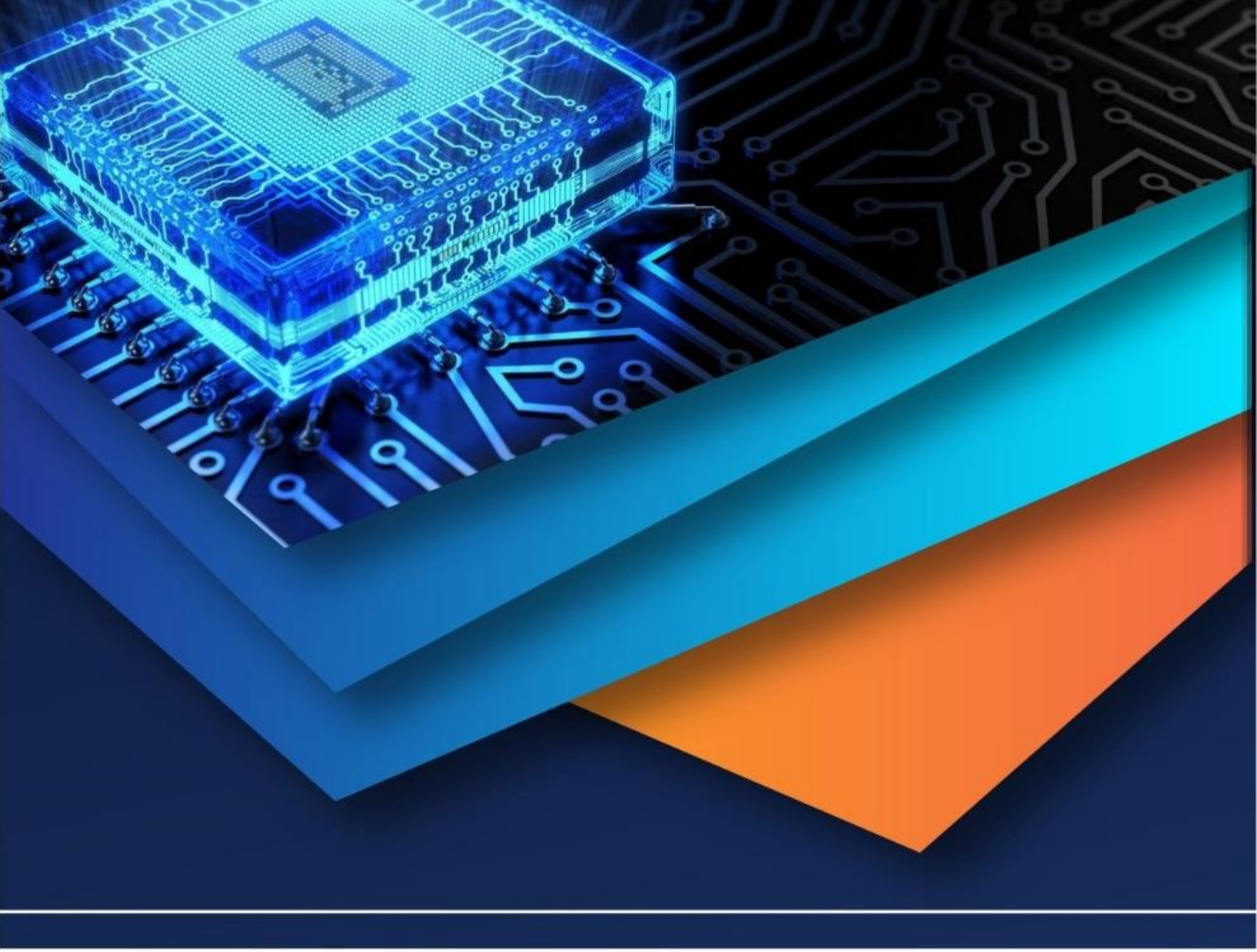


Figure 3: Comparison results of the algorithms KNN and Random Forest on the dataset

The drawback of the KNN algorithm is the bad performance on the unevenly distributed datasets. Thus, if one class vastly dominates the other ones, it is more likely to have more neighbours of that class due to their large number, and, therefore, make incorrect predictions.

REFERENCES

[1] A Predictive Modelling for Detecting Fraudulent Automobile Insurance Claims, Hojin Moon, Yuan Pu, Cesarina Ceglia,Journal of Theoretical Economics Letters, 2019, 9, 1886-1900
[2] Abnormal Pattern Prediction: Detecting Fraudulent Insurance Property Claims with Semi-Supervised Machine-Learning, Sebastián M. Palacio, Data science journal, 2019
[3] Detection of Automobile Insurance Fraud Using Feature Selection and Data Mining Techniques, Sharmila Subudhi, International Journal of Rough Sets and Data Analysis · July 2018
[4] Fraud Detection in Automobile Insurance using a Data Mining Based Approach, Ali Ghorbani and Sara Farzai, International journal of Mechatronics, Electrical and Computer technology, Vol. 8(27), Jan. 2018, PP. 3764-3771
[5] Nearest Neighbour and Statistics Method based for Detecting Fraud in Auto Insurance, TessyBadriyah, LailulRahmaniah, IwanSyarif, 2018 IEEE, 978-1-5386-8066-7/18
[6] CoDetect: Financial Fraud Detection with Anomaly Feature Detection, Dongxu,DejunMu, LibinYang and Xiaoyan Cai, 2169-3536 2018 IEEE.
[7] Cost-sensitive Parallel Learning Framework for Insurance Intelligence Operation, Xinxin Jiang, Shirui Pan, Member, IEEE, Guodong Long, Fei Xiong, Jing Jiang, and Chengqi Zhang, IEEE transactions on industrial electronics, 2018
[8] A State-of-the-Art Review of Machine Learning Techniques for Fraud Detection Research, Sinayobye Janvier Omar, Kiwanuka Fred, Kaawaase Kyanda, 2018 ACM/IEEE Symposium on Software Engineering in Africa
[9] Auto insurance fraud detection using unsupervised spectral ranking for anomaly, KeNian, Haofan Zhang, Aditya Tayal, Thomas Coleman, Yuying Li, The Journal of Finance and Data Science 2 (2016) 58e75
[10] Fraud detection system, Aisha Abdallah, MohdAizainiMaarof, Anazida Zainal, Journal ofNetworkandComputerApplications68(2016)90–113
[11] Detecting Auto Insurance Fraud by Data Mining Techniques, Rekha Bhowmik, Journal of Emerging Trends in Computing and Information Sciences, Volume 2 No.4, April 2011
[12] Journal of computer and network Applications, Fraud detection system: A Survey, Aisha Abdallah, Mohd Aizaini Maarof, Anazida Zainal, 68 (2016) 90-113.
[13] Strategies for detecting fraudulent claims in the automobile insurance industry, Stijn Viaene, Mercedes sAyuso, Montserrat Guillen, Dirk Van Gheel, Guido Dedene, European Journal of Operational Research 176 (2007) 565–583ss
[14] Claims Auditing in automobile insurance: fraud detection and deterrence objectives, Sharon Tennyson, Pau Salsas-Forn, The Journal of Risk and Insurance, 2002, Vol. 69, No. 3, 289-3



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY