International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5, June 2019

# Naïve Classification Approach for Insurance Fraud Prediction

**Bhavna Batra, Sheetal Kundra**

***Abstract :*** ***An approach which can be used for the prediction of future potentials on the basis of present information is known as prediction analysis. This study is relied on the fraudulent discovery in the insurance business. A number of approaches have been projected up to now for the fraudulent discovery in insurance sector. These approaches mainly rely on machine learning algorithms. The insurance fraud detection is the major issue of prediction analysis. The insurance fraud detection has three phases which are pre-processing, feature extraction and classification. The naïve bayes classification approach is proposed in this work for the insurance fraud prediction. The proposed algorithm is implemented in python and results are analyzed in terms of accuracy, execution time.***

***Keywords: Insurance Fraud detection, voting method, naïve bayes.***

## I. INTRODUCTION

The huge quantity of information is generated inside dissimilar applications in daily routine. The proper management of this data or information is a complex task. Generally, large size databases or folders are utilized for the storage of this huge quantity of information. This information is retrieved by the customers according to their obligation. The large sized storage areas and folders are formed for the storage of this big sized data [1]. The suitable extraction of significant information from such enormous database is the major area of concern which rises inside these schemes. According to the prerequisite of consumer, the significant data is to be withdrawal from the folder which is the main reason behind the development of different mechanisms [2]. The procedure which is used for the efficient storage and extraction of information inside some applications according to the need is specified as data mining process. Information detection is the other name given to the data mining. The information amassed inside enormous database is retrieved with the help of this procedure. The extracted or retrieved information can be utilized as significant data for some other purpose further [3]. The data mining process is an imperative branch of KDD although both data mining and KDD are very much alike. The unprocessed information is retrieved through database. Then KDD procedure transformed this data into valuable information which is utilized in different applications. The occurrence of fraud directly affects any kind of insurance cover.

**Manuscript published on 30 June 2019.**
***** Correspondence Author (s)
**Bhavna Batra,** Student, ME-CSE (Cloud Computing) Chandigarh University , Ajitgarh, Punjab, India
**Sheetal Kundra,** Associate Professor, AIT, Chandigarh University, Ajitgarh, Punjab, India



The data inside the claims of insurance or the solutions of queries rising from an insurance application structure is not ingenuous or absolute [4]. A assert is proposed on the basis of mislead or false situations involving the overstatement of an authentic allege. With a purpose of increasing profit inside an insurance agreement, ambiguous or making false contract with an insuring agent is checked. The insurance plan holder or third parties assert can entrust an insurance deception beside an insurance plan. The asserts for apparition travelers, untrue damage in highway catastrophe, cosmopolitan allege or prearranged offense spheres are only mere of these kind of circumstances in which deception is present in huge manner [5]. The degree of insurance deception is different for dissimilar nations. There are also some definite factors connected to the performance of marketplace or the restricted pervasiveness of any exacting kind of insurance, which influence the degree of fraudulent [6]. Though, the main goals in this study are the creation of an extent for pertaining that the present counter-fraudulent proposals are unbeaten and checking for the requirement of additional proposals [7]. The different marketplaces gathers accurate data associated with the occurrence of fraudulent. The insurance company and the client or policy holder give his permission on some points through an insurance policy. Several kinds of jeopardy are enclosed in these points. In a case when data given by both the sides is similar and both sides are trustworthy, the compensation outstanding through the plan creator and also the payment paid by the client will efficiently demonstrate probability of the loss occurring and the imposed expected losses. The insurance plan holder is not all the time truthful in the genuine scenario [8]. Different kinds of fraudulent activities can be developed if the data irregularity leans towards the consumer. In other words it can be said that when insurance plan owner comprises more information about the insurance plan than insurance manager or corporation, then fraudulent occurs. Cover fraudulent happens in a case when insurance plan owner or client is not exposed to the insurance firm or insurer [9]. In this situation, any kind of data which can influence the likelihood of thrash occurrence turn out to be the cause of cover fraud incident. The fraudulent is described as a type of purposeful activity in order to cheat or mislead some human being or organization for attaining financial earnings [10]. On the basis of investigations, this determined and unlawful fraudulent activity can be established or classified in terms of different means. The exterior fraud or customer fraud is a type of fraud in which the deception is carried out exterior to the organization [11]. The inside fraud or administration fraud is a kind of fraud which is executed through the higher-level management. The significant dissimilarities amid consumer activities scrutiny and fraudulent scrutiny methods are to be highlighted importantly [12].

A low optimistic pace can be utilized for the detection of recognized fraudulent behaviors of the deception scrutiny technique. The name and form of fraudulent actions that are given inside the prophesy data sample are retrieved through the scheme [13]. The kinds of frauds being faced by the scheme can be found out easily [14]. If the experiment information does not comprise any fraudulent signatures, no alarm rings. Thus, there can be tremendous diminishing of the false positive rate. The new deceptions cannot be identified as the partial and precise fraudulent proceedings are utilized for the learning of fraudulent scrutiny scheme [15]. Therefore, the false negatives rate turn out to be very elevated on the basis of the information that how clever the impostors are. A suite of interconnected nodules that are intended for emulating the functioning of person intellect is identified as artificial neural network (ANN) [16]. A prejudiced link is allocated to every nodule present in the neighboring layer of every nodule. The key acknowledged from the linked nodules is collected and the masses are utilized in association with an easy purpose for computing the yield standards by the solitary nodules. The Artificial Immune System (AIS) is a recently evolved sub-domain which is developed on the basis of organic metaphor of the immune structure [17]. The demarcation can be made amid the self and non-self-cells or amid the injurious and non-injurious cells by the immune structure. The natural development was the inspiration behind the evolution of genetic algorithms. The chromosomes are identified as a populace of applicant resolutions which occur in the structure of dual threads [18]. The chromosomes are considered as best possible resolution and provided through genetic algorithm. The idea that the probability of continued existence and imitation is higher in the suite of robust associates of definite populace is applicable in this scenario. The hidden Markov model is a twofold entrenched speculative procedure which is used for the generation of extremely complex speculative procedures [19]. Inside the fundamental scheme, a Markov procedure that comprises unnoticed state is implicated to be accessible.

The remaining portion of this paper contains the “Literature Review” section that has the table of comparison. The part Research Methodology explores the evaluation of data accuracy and classification. The next part contains result and discussion. The concluding portion is examined in the section “Conclusion”.

## II. LITERATURE REVIEW

Lamberti, et.al (2018) proposed a analysis in which smart contract sensors information was used to understand a scheme for on-demand insurance. A mobile application and an electronic tool were fitted on the vehicles of customers for generating a prototype. For scheduling the automatic modifications and collecting pictures of vehicles, the interaction with smart contract was carried out for modify the policy coverage physically [20]. The environmental circumstances were observed and the alterations were launched with the help of electronic equipment. The projected resolution was capable for lowering the policy modification costs.

Subudhi, et.al (2018) projected a novel fraudulent detection approach in vehicle insurance area. A method identified as adaptive oversampling technique was utilized for the implementation of the projected methodology [21]. Within the alternative class spaces, an adaptive oversampling method was utilized for the information inequity accessible in the formerly maintained dataset. Many tests on the basis of automobile insurance dataset were executed in order to evaluate the projected method. Different tests performed on practical dataset clearly demonstrated the competence of the projected scheme.

kareem, et.al (2017) projected a new scheme for the detection of fraudulent in health insurance alleges. Some definite qualities on the claim credentials were scrutinized for recognizing the association or connection amid them for the identification of deceptions [22]. The necessities of health insurance business were not fulfilled. Therefore, the recognition of deceptions in health insurance enlarged the level of investigators attention in data mining. Therefore, the thriving fortitude of connected qualities could deal with the inconsistency of information in the fraudulent alleges which could provide assistance in reducing the possibilities of fraudulent in the health insurance.

Kenyon, et.al (2017) presented an investigation of big data and data science applications utilized for the prediction of fraudulent insurance alleges. Large data, data science and predictive analytics were applied along with a use-case in an interim insurance business [23]. A privacy protection approach was projected for predicting the insurance alleges fraud. The proposed approach utilized enormous data samples for the generation of policies beside with the interest of cross-broker and cross-insurer exploitations. The tested outcomes depicted that the regulations generated in this technique were more precious than the existing approaches.

Anbarasi, et.al (2017) proposed a research relevant to the fraudulent discovery in the health insurance information. The proposed approach integrated the proactive and retrospective analysis [24]. The disjointed character of anomalous actions was managed by incorporating the reactive and proactive schemes for generating an enhanced method. The doubtful activity of health concerned trace was recognized with the help of outlier relied predictors for health insurance fraudulent discovery. In future, the character of health care fraudulent existence can be identified by realizing future enhanced online fraudulent discovery technologies.

Wang, et.al (2017) projected a new scheme for the identification of automobile insurance fraudulent based on deep learning [25]. The proposed approach utilized LDA relied text analytics. Latent Dirichlet Allocation technique was used for the extraction of the text characteristics incasing in the text images of accidents. Further deep neural networks were trained on the presented information. The tested outcomes demonstrated that for the detection of automobile insurance fraudulent, the combination of Latent Dirichlet Allocation and deep neural network approach was extremely important and competent.


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

**Table 2.1: Table of Compassion**

| Ref. No. | Year | Technique Used | Pros and Cons |
|---|---|---|---|
| 20. | 2018 | Proposed a analysis in which smart contract sensors information was used to understand a scheme for on-demand insurance | The projected resolution was capable for lowering the policy modification costs. The damages, and smart contracts leveraged to automatically trigger reimbursements were not detected here. |
| 21. | 2018 | Projected a novel fraudulent detection approach in vehicle insurance area. A method identified as adaptive oversampling technique was utilized for the implementation of the projected methodology | Different tests performed on practical dataset clearly demonstrated the competence of the projected scheme. |
| 22. | 2017 | Projected a new scheme for the detection of fraudulent in health insurance alleges | Therefore, the thriving fortitude of connected qualities could deal with the inconsistency of information in the fraudulent alleges which could provide assistance in reducing the possibilities of fraudulent in the health insurance. |
| 23. | 2017 | Presented an investigation of big data and data science applications utilized for the prediction of fraudulent insurance alleges | The tested outcomes depicted that the regulations generated in this technique were more precious than the existing approaches. |
| 24. | 2017 | Proposed a research relevant to the fraudulent discovery in the health insurance information | The doubtful activity of health concerned trace was recognized with the help of outlier relied predictors for health insurance fraudulent discovery. However, the nature of the health care fraud occurrences was not identified. |
| 25. | 2017 | Projected a new scheme for the identification of automobile insurance fraudulent based on deep learning | The tested outcomes demonstrated that for the detection of automobile insurance fraudulent, the combination of Latent Dirichlet Allocation and deep neural network approach was extremely important and competent. |

## III. RESEARCH METHODOLOGY

The prognostic modeling can be executed with the help of any of the huge-scale data samples accessible nowadays both from social media platforms, trade modeling and so on. The prognostic modeling of data samples is advancing recognition amid the educational shareholders because of the improvement in messaging and information level. The customers are sighted on the basis of new approaches through predictive modeling analysis which provide assistance in the modernization of educational institutions. Enormous quantity of information is produced inside every application which is amassed in huge size folders. Extremely competent technologies are mandatory for the management of these huge folders and the information accessible inside them. The investigators have projected different methods for facilitating the numerous customers and their functions. In order to handle such huge quantity of information, data mining and knowledge discovery approaches are utilized. The procedure which is used for the efficient storage and extraction of information inside some applications according to the need is specified as data mining process. The information already present inside the folders is utilized for the generation of evocative, comprehensible and prognostic forms with the help of data mining. The presented investigative study utilizes Naïve Bayes Classifier. This is sort of probabilistic classifier which is used for the recognition of any doubts so that the true values remain unaffected. The normal circulation inside this scheme is utilized for showcasing the mathematical properties. The intended discretization is utilized for dealing with numeric properties. The plan using which the description demonstration is produced like the group names can be selected for questioning of occurrences, the vectors of emphasize values are produce by pointer is identified as Naïve bayes. Here, through several restricted suite, the group characters are ventured.

The computation performed on the basis of standard necessities is composed by this classifier as well. The constituent is computed with the help of this bayes classifier for the provided group symbols. The Naïve Byaes classifier uses small assess of organized data in order to appraise the constraints that are utilized for performing categorization.



Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

The past information of last 10 years connected to credit cards is utilized for performing the key information attainment. The information such as through which data is to be retrieved , at what occasion and from which place describe the transactions made which are all incorporated inside the past information.

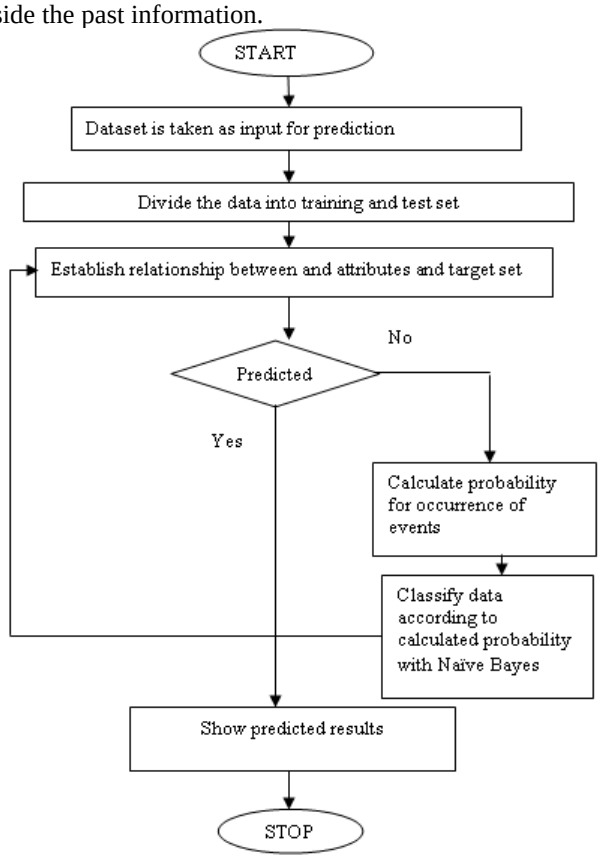


**Fig 1: Proposed Methodology**

## IV. RESULT AND DISCUSSION

The proposed model is implemented for the insurance policy fraud detection. The proposed model has the three phases which are pre-processing, feature extraction and classification. The naïve bayes classification model is implemented in this work for the insurance fraud policy detection. The performance of proposed model is testing in terms of accuracy and execution time

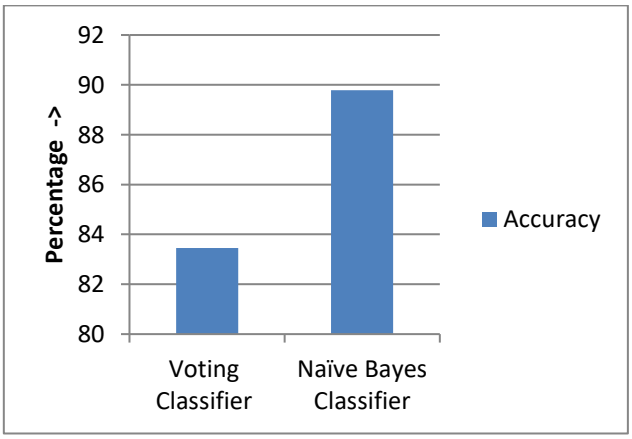


**Fig 2: Accuracy Comparison**

As shown in figure 2, the accuracy of the voting based classification model is compared with naïve bayes classification. It is analyzed that naïve bayes classifier has high accuracy as compared to voting method

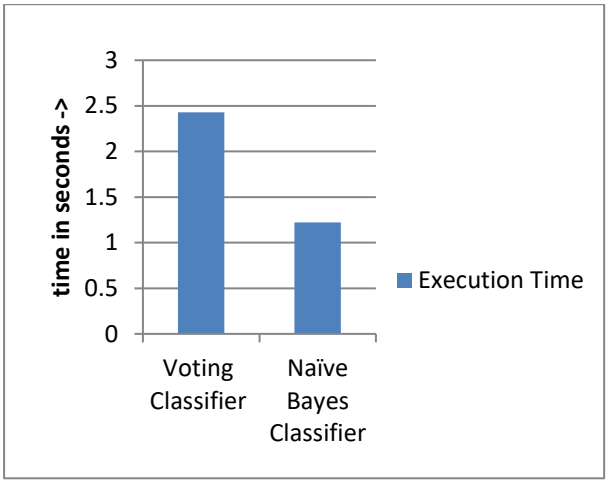


**Fig 3: Execution Time Comparison**

As shown in figure 2, the execution time of voting based classifier is compared with naïve bayes classifier. It is analyzed that naïve bayes has low execution time due to its low complexity as compared to voting method.

## V. CONCLUSION

In this paper, it is concluded insurance fraud detection is the major challenge of prediction detection. The insurance fraud detection has the three major steps which are pre-processing, feature extraction and classification. The naïve bayes classification approach is used in this work for the insurance fraud detection. The proposed model is implemented in python and results are analyzed in terms of accuracy, execution time. The proposed model give approx 5 percent high results as compared to voting classifier.

## REFERENCES

1. Chenfei Sun ; Qingzhong Li ; Hui Li ; Yuliang Shi ; Shidong Zhang ; Wei Guo, "Patient Cluster Divergence Based Healthcare Insurance Fraudster Detection" , IEEE Access Year: 2019 , Volume: 7 Page s: 14162 – 14170
2. S. Viaene, R.A. Derrig ; G. Dedene, "case study of applying boosting naive Bayes to claim fraud diagnosis", IEEE Transactions on Knowledge and Data Engineering, Year: 2004 , Volume: 16 , Issue: 5 Page s: 612 – 620
3. M. Sternberg; R.G. Reynolds, "Using cultural algorithms to support re-engineering of rule-based expert systems in dynamic performance environments: a case study in fraud detection", IEEE Transactions on Evolutionary Computation Year: 1997 , Volume: 1 , Issue: 4 Page s: 225 - 243
4. J.R. Dorronsoro; F. Ginel ; C. Sgnchez ; C.S. Cruz, "Neural fraud detection in credit card operations", IEEE Transactions on Neural Networks Year: 1997 , Volume: 8 , Issue: 4 Page s: 827 – 834
5. Aastha Bhardwaj, Rajan Gupta, "Financial Frauds: Data Mining based Detection – A Comprehensive Survey", 2016, International Journal of Computer Applications (0975 – 8887)

6. Yongchang Gao ; Chenfei Sun ; Ruican Li ; Qingzhong Li ; Lizhen Cui ; Bin Gong, "An Efficient Fraud Identification Method Combining Manifold Learning and Outliers Detection in Mobile Healthcare Services", IEEE Access Year: 2018 , Volume: 6 Page s: 60059 – 60068
7. Yufeng Kou, "Survey of fraud detection techniques", 2004, Networking, Sensing and Control,IEEE International Conference
8. Shunzhi Zhu, Yan Wang, Yun Wu, "Health Care Fraud Detection Using Nonnegative Matrix Factorization", The 6th International Conference on Computer Science & Education (ICCSE 2011) August 3-5, 2011. SuperStar Virgo, Singapore
9. Zhongyuan Zhang, Tao Li, Chris Ding, Xiangsun Zhang, "Binary Matrix Factorization with Applications", Proceeding ICDM '07 Proceedings of the 2007 Seventh IEEE International Conference on Data Mining Pages 391-400.
10. Mohammad Sajjad Ghaemi. Class Lecture, Topic: "Clustering and Nonnegative Matrix Factorization". DAMAS LAB, Computer Science and Software Engineering Department, Laval University. Apr.12, 2013.
11. Haesun Park. Class Lecture, Topic: "Nonnegative Matrix Factorization for Clustering". School of Computational Science and Engineering Georgia Institute of Technology Atlanta, GA, USA, July 2012.
12. Fashoto Stephen G., Owolabi Olumide, Sadiku J., Gbadeyan Jacob A, "Application of Data Mining Technique for Fraud Detection in Health Insurance Scheme Using Knee-Point K-Means Algorithm", Australian Journal of Basic and Applied Sciences, 7(8): 140-144, 2013 ISSN 1991- 8178.
13. Leonard Wafula Wakoli. "APPLICATION OF THE K-MEANS CLUSTERING ALGORITHM IN MEDICAL CLAIMS FRAUD/ABUSE DETECTION." MSc Thesis, Jomo Kenyatta University Of Agriculture And Technology, 2012.
14. Yaqi Li, Chun Yan, Wei Liu, Maozhen Li, "Research and Application of Random Forest Model in Mining Automobile Insurance Fraud", 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)
15. Richard A. Bauder, Taghi M. Khoshgoftaar, Aaron Richter, Matthew Herland, "Predicting Medical Provider Specialties to Detect Anomalous Insurance Claims", 2016 IEEE 28th International Conference on Tools with Artificial Intelligence
16. S. N. John, Okokpujie Kennedy O. C. Anele, F. Olajide, Chinyere Grace Kennedy, "Real time Fraud Detection in the Banking Sector Using Data Mining Techniques/Algorithm", 2016, IEEE
17. Riya Roy, Thomas George K, "DETECTING INSURANCE CLAIMS FRAUD USING MACHINE LEARNING TECHNIQUES", 2017 International Conference on circuits Power and Computing Technologies [ICCPCT]
18. K. Supraja, S.J. Saritha, "Robust Fuzzy Rule based Technique to Detect Frauds in Vehicle Insurance", International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017)
19. Aayushi Verma, Anu Taneja, Anuja Arora, "Fraud Detection and Frequent Pattern Matching in Insurance claims using Data Mining Techniques", Proceedings of 2017 Tenth International Conference on Contemporary Computing ( IC3)
20. Fabrizio Lamberti, Valentina Gatteschi, Claudio Demartini, Matteo Pelissier, Alfonso Gómez, and Victor Santamaria, "Blockchains Can Work for Car Insurance: Using Smart Contracts and Sensors to Provide On-Demand Coverage", 2018, IEEE Consumer Electronics Magazine, Volume: 7, Issue: 4
21. Sharmila Subudhi, Suvasini Panigrahi, "Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud", 2018, 2nd International Conference on Data Science and Business
22. Saba kareem, Dr. Rohiza Binti Ahmad, Dr. Aliza Binit Sarlan, "Framework for the Identification of Fraudulent Health Insurance Claims using Association Rule Mining", 2017 IEEE Conference on Big Data and Analytics (ICBDA)
23. David Kenyon, J.H.P Eloff, "Big Data Science for Predicting Insurance Claims Fraud", 2017, IEEE
24. Dr.M.S. Anbarasi, S. Dhivya, "FRAUD DETECTION USING OUTLIER PREDICTOR IN HEALTH INSURANCE DATA", INTERNATIONAL CONFERENCE ON INFORMATION, COMMUNICATION & EMBEDDED SYSTEMS (ICICES 2017)
25. Yibo Wang, Wei Xu, "Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud", 2017, DECSUP 12895

## AUTHORS PROFILE

**Bhavna Batra** is pursuing Master of Engineering in Computer Science and Applications in Cloud Computing specialization from Chandigarh University, Gharuan, India. She has done Bachelors of Engineering from DCRUST Murthal University, India (2016). Her research interest includes machine learning and prediction analysis.

**Sheetal Kundra** received her PhD degree from Department of Computer Science and Engineering, IKGPTU ,Jalandhar. She has done his Master in Technology (Computer Science and Engineering) from Rayat Institute Of Engineering And IT, Railmajra, India (2010). Currently, She is working as an associate professor at Chandigarh University, Gharuan, Mohali, Punjab, India. She has published more than 20 research papers in well-known reputed journals and international conferences. Her research interest includes Exploring Ground WaterPossibility in Case based recommendation system using Nature Inspired Techniques.


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)