International Journal of Engineering Research and Science & Technology

IJERST

ISSN : 2319-5991

www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issuse.1, March 2024

IDENTIFYING HEALTH INSURANCE CLAIM FRAUDS USING MACHINE LEARNING CONCEPTS

Chandra Sunilram[1],Danda Bhavana[2],Karnati Abhisriraj[3],Dr M Gayatri[4]

[1,2,3]B.Tech Student, Department of CSE (Internet of Things), Malla Reddy College of Engineering and Technology, Hyderabad, India.

[4]Associate Professor, Department of CSE (Internet of Things), Malla Reddy College of Engineering and Technology, Hyderabad, India.

ABSTRACT

Patients depend on health insurance provided by the government systems, private systems, or both to utilize the high-priced healthcare expenses. This dependency on health insurance draws some healthcare service providers to commit insurance frauds. Healthcare is one of the largest financial sectors. It has the massive amount of data containing health records, clinical data, insurance claim, provider and patient information. Although the number of such service providers is small, it is reported that the insurance providers lose billions of dollars every year due to frauds. In this paper, we formulate the fraud detection problem over a minimal, definitive claim data consisting of medical diagnosis and procedure codes. In this paper, we perform a comparative analysis on various classification algorithms, namely Support Vector Machine (SVM), Decision-Tree (DT), K- Nearest Neighbor (KNN), and Logistic Regression (LR), to detect the health insurance fraud. The effectiveness of the algorithms is observed on the basis of performance metrics: Precision, Recall and F1-Score. Our experimental results demonstrate promising outcomes in identifying fraudulent records.

Keywords: Machine Learning, Support Vectot Machine, Decision Tree, K- Nearest Neighbor,Logistic Regression.

I.INTRODUCTION

With the rapid growth of healthcare services, health insurance fraud detection has become important measure to ensure efficient use of public funds. Traditional fraud detection methods have tended to focus on the attributes of a single visit and have ignored the behavioural relationships of multiple visits by patients. Machine Learning (ML) is a sub-area of Artificial Intelligence with the main aim to mimic human intelligence abilities. ML focuses on constructing models with high prediction capabilities. The most basic feature is “Learning” which is done by looking at the given data. The two basic learning techniques are Supervised and Unsupervised. In supervised learning, we are provided with fully labelled data that means in

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issuse.1, March 2024

the training data against each input we have the desired result as well. It is highly useful for solving problems of classification and regression. In classification, the aim is to predict a discrete value whereas regression deals with continuous data. On contrary, in an unsupervised learning paradigm, we are provided with unlabelled data where results are not known.

The major issue faced by insurance companies is fraud that causes immense loss to insurance companies sometimes beyond repair. Fraud may be committed at different points by applicants, policyholders, third-party claimants, or professionals who provide services to claimants. Insurance agents and company employees may also commit insurance fraud. Common frauds include "padding" (inflating claims), misrepresenting facts on an insurance application, submitting claims for injuries or damage that never occurred, and staging accidents. In a fraud detection scenario in a supervised learning method, we can find out fraud and legal cases from training data but in unsupervised learning, we cannot infer which one is a fraud case and which one is legal. We formulate the fraud detection problem over a minimal, definitive claim data consisting of medical diagnosis and procedure codes.

II.LITERATURE REVIEW

Fraud Detection and Analysis for Insurance Claim using Machine Learning,Abhijeet Urunkar,Amruta Khot,Rashmi Bhat,Nandinee Mudegol,Insurance Company working as commercial enterprise from last few years have been experiencing fraud cases for all type of claims. Amount claimed by fraudulent is significantly huge that may causes serious problems, hence along with government, different organization also working to detect and reduce such activities. Such frauds occurred in all areas of insurance claim with high severity such as insurance claimed towards auto sector is fraud that widely claimed and prominent type, which can be done by fake accident claim. So,we aim to develop a project that work on insurance claim data set to detect fraud and fake claims amount. The project implement machine learning algorithms to build model to label and classify claim. Also, to study comparative study of all machine learning algorithms used for classification using confusion matrix in term soft accuracy, precision, recall etc. For fraudulent transaction validation, machine learning model is built using PySpark Python Library.

III.EXISTING SYSTEM

The existing system for health insurance fraud detection primarily relies on manual in- vestigations and rule-based algorithms. Insur- ance companies employ teams of experts to manually review claims and identify potential fraud based on suspicious patterns or behaviors. This approach is time-consuming and can be limited in its ability to handle large volumes of data and complex fraud schemes. While these algorithms provide some level of automation, they may not be able to adapt to evolving fraud patterns or capture subtle fraudulent activities, leading to potential

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issuse.1, March 2024

false positives and negatives. As a result, the current system may struggle to effectively combat sophisticated health insurance fraud.

IV.PROPOSED SYSTEM

The main purpose of the project is to detect health insurance claim frauds. Generally, many people depend on the insurance for their treatments in hospital by taking these as advantage some people trying to do fraud.so we came up with this project to identify frauds and help people who are in need of insurance. To detect health insurance frauds, we used some algorithms like support vector method (SVM), Logistic regression, K-Nearest Neighbor (KNN), Decision tree. These algorithms classify given data into categorize like fraud and not fraud.

V.METHODOLGY

Data Collection and Preprocessing

Collect historical health insurance claim data, including both legitimate and potentially fraudulent claims. Clean and preprocess the data by handling missing values, outliers, and data quality issues.Feature engineering: Create relevant features from the raw data that can help in fraud detection. These features may include claim amounts, diagnosis codes, provider information, patient demographics, etc.

Data Splitting

Divide the dataset into training, validation, and test sets. The training set is used to train the model, the validation set for tuning hyperparameters, and the test set for final evaluation.

Exploratory Data Analysis (EDA)

Conduct EDA to gain insights into the data, identify patterns, and understand the distribution of features in legitimate and fraudulent claims.

Feature Selection

Select the most relevant features that contribute to fraud detection. Feature selection techniques like mutual information, recursive feature elimination, or L1 regularization can be used.

Model Selection

Choose machine learning models suitable for fraud detection. Commonly used models include:

- Logistic Regression
- Decision Trees
- Random Forest
- Gradient Boosting (e.g., XGBoost, LightGBM)
- Neural Networks (Deep Learning)

Model Training

Train the selected models on the training dataset. Use appropriate evaluation metrics (e.g., F1-score, precision, recall) to assess model performance during training.

Hyperparameter Tuning

Optimize model hyperparameters using techniques like grid search or randomized search

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issuse.1, March 2024

on the validation set to improve model performance.

Ensemble Methods

Combine multiple models using ensemble techniques like stacking or bagging to enhance fraud detection accuracy.

Model Evaluation:

Evaluate the trained models on the test dataset using appropriate metrics.Analyze the confusion matrix to understand false positives and false negatives.

Threshold Optimization:

Identify an optimal threshold for the classification of claims into fraudulent or legitimate categories. Fine-tune this threshold to strike a balance between precision and recall, aligning with specific business requirements.

Model Deployment:

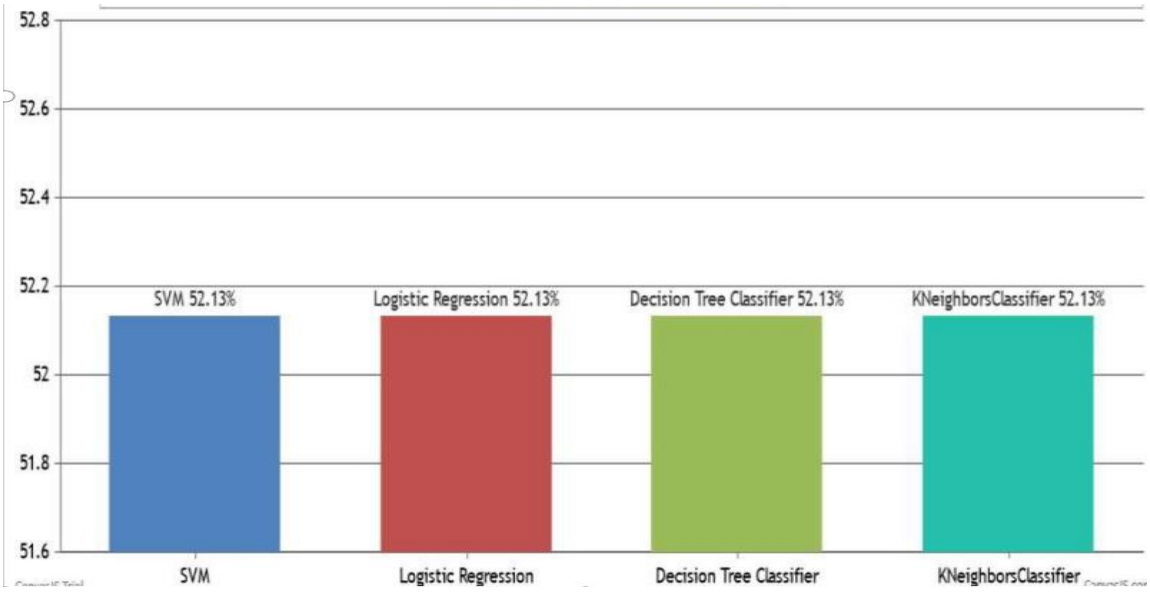
Initiate the deployment of the trained model into a live production environment, enabling the automatic real-time detection of fraudulent claims.

Continuous Monitoring and Feedback Loop:

Establish a continuous monitoring mechanism to assess the model's performance within the production environment. Collect feedback on predictions and implement a feedback loop to iteratively retrain and enhance the model over time, ensuring its sustained effectiveness.

Regulatory Compliance:

Verify that the fraud detection system adheres to healthcare regulations and privacy laws, including but not limited to HIPAA in the United States. Prioritize compliance to ensure the system's ethical use and safeguarding of sensitive information.

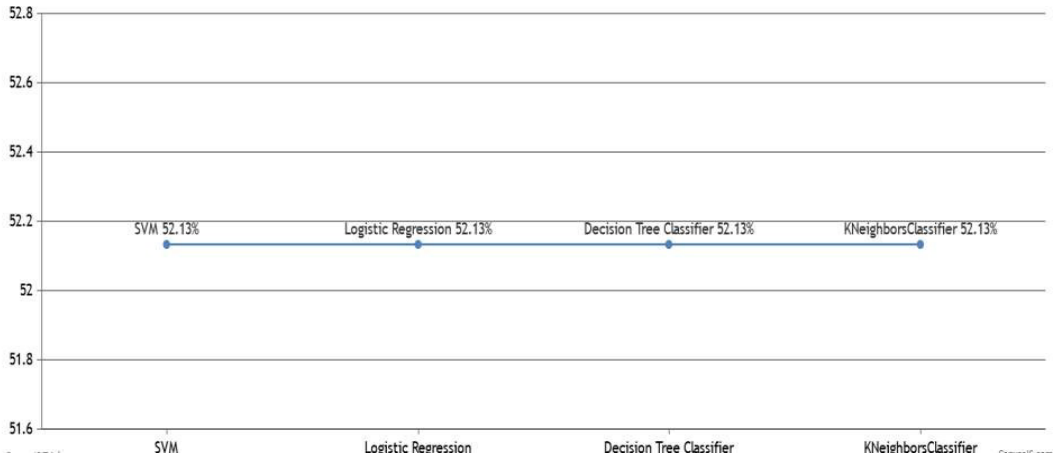


International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issuse.1, March 2024

PIE CHAERT

LINE CHART



VIEW ALL REMOTE USERS !!!

USER NAME	EMAIL	Gender	Address	Mob No	Country	State	City
Ashok	Ashok123@gmail.com	Male	#8928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
Manjunath	tmksmanju13@gmail.com	Male	#892,4th Cross,malleshwaram	9535866270	India	Karnataka	Bangalore
user	user@gmail.com	Female	Karimnagar	9993487243	India	Telangana	Karimnagar

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issuse.1, March 2024

VI.CONCLUSION

We pose the problem of fraudulent insurance claim identification as a feature generation and classification process by utilizing these concepts, healthcare organizations can improve the accuracy and efficiency of fraud detection. By using Machine learning algorithms like SVM, Logistic regression, Decision tree and KNN. Our results demonstrate an improvement scope to find fraudulent healthcare claims with minimal information. In the future, the fraud detection method can be extended to the Adaptive NeuroFuzzy Inference System (ANFIS) which is the combination of both Neuro-Fuzzy and Neural Networks. Hence, the prediction can be Model (HMM) to predict fraud using internal factors

.

VII. REFERENCES

[1] K. Ulaga Priya and S. Pushpa, “A Survey on Fraud Analytics Using Predictive Model in Insurance Claims,” Int. J. Pure Appl. Math., vol. 114, no. 7, pp.755–767, 2017.

[2] E. B. Belhadji, G. Dionne, and F. Tarkhani, “A Model for the Detection of Insurance Fraud,” Geneva Pap. Risk Insur. Issues Pract., vol. 25, no. 4, pp. 517–538, 2000, doi: 10.1111/1468-0440.00080.

[3] “Predictive Analysis for Fraud Detection.” https://www.wipro.com/analytics/comparative- analysis-of-machine-learning-techniques-for-%0Adetectin/.

[4] F. C. Li, P. K. Wang, and G. E. Wang, “Comparison of the primitive classifiers with extreme learning machine in credit scoring,” IEEM 2009 - IEEE Int. Conf. Ind. Eng. Eng. Manag., vol. 2, no. 4, pp. 685–688, 2009, doi: 10.1109/IEEM.2009.5373241.

[5] V. Khadse, P. N. Mahalle, and S. V. Biraris, “An Empirical Comparison of Supervised Machine Learning Algorithms for Internet of Things Data,” Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018, pp. 1–6, 2018, doi: 10.1109/ICCUBEA.2018.8697476.

[6] S. Ray, “A Quick Review of Machine Learning Algorithms,” Proc. Int. Conf. Mach. Learn. Big Data,Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019, pp. 35–39, 2019, doi: 10.1109/COMITCon.2019.8862451.