Journal of Science and Technology
ISSN: 2456-5660 Volume 8, Issue 07 (July -2023)
www.jst.org.in *DOI:https://doi.org/10.46243/jst.2023.v8.i07.pp45 - 57*

IDENTIFYING HEALTH INSURANCE CLAIM FRAUDS USING MACHINE LEARNING CONCEPT

Mrs. K. Aarati[1], A. Shirisha[2], K. Bindhu[3], Ch. Vandhana[4], Ch. Hasika[5]
[1]Assistant Professor, Department of CSE, Malla Reddy Engineering College for Women, Hyderabad, Telangana, India
[2,3,4,5]UG Scholar, Department of CSE, CSE,Malla Reddy Engineering College for Women, Hyderabad,Telangana, India
mrecwaarati27sep@gmail.com
shirishaankam07@gmail.com, binduketham@gmail.com, vandhanach28282@gmail.com, hasikareddycholleti@gmail.com

To Cite this Article

Mrs. K. Aarati, A. Shirisha, K. Bindhu, Ch. Vandhana, Ch. Hasika , IDENTIFYING HEALTH INSURANCE CLAIM FRAUDS USING MACHINE LEARNING CONCEPT" *Journal of Science and Technology, Vol. 08, Issue 07,-July 2023, pp45-57*

Article Info
Received: 27-06-2023 Revised: 28-06-2023 Accepted: 11-07-2023 Published: 18-07-2023

***Abstract*— Patients depend on health insurance provided by the governmentsystems, 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. In this paper, we perform a comparative analysis on various classification algorithms, namely Support Vector Machine (SVM), Decision-Tree (DT), K-Nearest Neighbor (KNN), Logistic Regression (LR), to detect the health insurance fraud. The effectiveness of the algorithms are observed on the basis of performance metrics: Precision, Recall and F1-Score**

Keywords—Insurance, Healthcare, Fraud Detection, SVM, Decision Tree, Logistic Regression, KNN

I. INRODUCTION

The major issue faced by insurance companies is a 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 toclaimants. Insurance agents and company employees may also commit insurance fraud. Common frauds include "padding" (inflating claims), misrepresenting facts onan insurance application, submitting claimsfor injuries or damage that never occurred,and staging accidents.

In a fraud detection scenario in a supervised learning method we can find outfraud and legal cases from training data butin 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. We present a solution to the fraudulent claim detection in medical domain using Machine learning concepts like SVM, Logistic regression, Decisiontree and KNN algorithms.

II. LITERATURE SURVEY

Sun et al f presented a novel approach for detecting frauds, called Patient Cluster Divergence-based Healthcare Insurance Fraudster Detection (PCDHIFD) in presence of camouflage responses. For theexperimental purpose, the health caredataset was chosen and the dataset comprised of around 40M admission records of 15000 patients of the previous five years.

The proposed technique worked in 3steps for three basic records: Life history ofpatients, diagnosis record, and medical practitioners attended. Steps we use in this sequence: first of all, a patient graph was constructed based on most similar info for the patient level hospital admission. Then aclustering-based graph algorithm was used for finding the peak and real meaning for individual clusters. Lastly, the difference inthe patient cluster was found and the probability of fraud for each patient was calculated. The comparison was made with other state of the art algorithms i.e., Decision Trees, Support Vector Machines, GridLOF, BP Growth, MLP and LSTM. Itwas claimed that the proposed approach produced the highest accuracy.

III. METHODOLOGY

Looking into protection claims misrepresentation space needs a reasonable unmistakable output on what extortion is asa consequence of its normally lumped related to mishandle and squander. Notwithstanding, extortion and misuse visit a situation where help administration is acquired yet not gave or remuneration of assets is made to outsider insurance agencies. Misrepresentation and misuse unit of estimation further clarified as help

providers accepting kickbacks, patients looking for medications that unit of estimation no doubt unsafe to them, (for example, looking for medications to fulfil addictions), and therefore the remedy of administrations understood to be extra. Protection extortion is Associate in nursing deliberate demonstration of beguiling, covering, or distorting information that winds up in help edges being paid to a non- open or bunch. Record examining and analyst examination is a zone of protection misrepresentation location. Cautiousrecord inspecting can uncover suspicious providers and policyholders. It is the best gratitude to review all cases individually. Finally evaluating the findings of each model and algorithms with a set of data thathaven't been used in the training phase of the model to check how accurate is this model and find out if there is any overfitting or underfitting before the selection of the model or modifying it if it was possible to come up with a better anda more accurate model for the end-user.

A. SUPPORT VECTOR MACHINE(SVM)

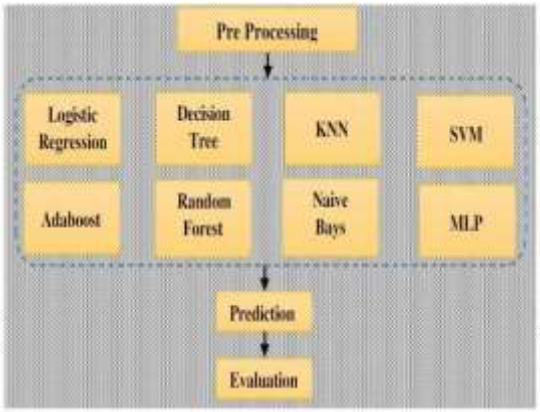


Figure 1: Fraud Detection System

A. DATA FLOW DIAGRAM

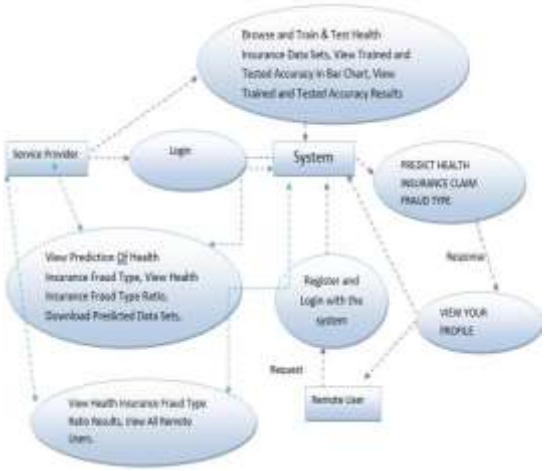


Figure 2: Data flow diagram

IV. TYPES OF CLASSIFICATIONALGORITHMS

Many machine learning algorithms are being used in various fields of research to help in solving the real-world problems. Mostly used machine learning classification algorithms are discussed below:

In classification tasks a discriminant machine learning technique aims at finding, based on an *independent and identically distributed* (*iid*) training dataset, a discriminant function that can correctly predict labels for newly acquiredinstances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point *x* and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involvesoutlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surfacethat best separates the different classes in the feature space.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to *genetic algorithms* (*GAs*) or *perceptron's*, both of which are widely used for classification in machine learning.For perceptron's, solutions are highly

dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptron's is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

B. LOGISTIC REGRESSION

Logistic regression analysis studies the association between a categoricaldependent variable

and a set of independent (explanatory) variables. The name *logistic regression* is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name *multinomial logistic regression* is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different fromthat of multiple regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is

discriminant analysis. This is becauselogistic regression does not assume that the independent variables are normallydistributed, as discriminant analysis does. This program computes binary logisticregression and multinomial logisticregression on both numeric and categoricalindependent variables. It reports on theregression equation as well as the goodnessof fit, odds ratios, confidence limits,likelihood, and deviance. It performs acomprehensive residual analysis including diagnostic residual reports and plots. It canperform an independent variable subset selection search, looking for the bestregression model with the fewestindependent variables. It providesconfidence intervals on predicted valuesand provides ROC curves to helpdetermine the best cutoff point forclassification. It allows you to validateyour results by automatically classifyingrows that are not used during the analysis.

C. DECISION TREE CLASSIFIER

Decision tree classifiers are usedsuccessfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision-making knowledge from the supplied data. Decision tree can be generated fromtraining sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with thisclass

Step 2. Otherwise, let T be some test with possible outcomes O1, O2…, On. Eachobject in

S has one outcome for T so the test partitions S into subsets S1, S2… Sn where each object in Si has outcome Oi for

T. T becomes the root of the decision tree and for each outcome Oi, we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

D. KNN CLASSIFIER

Simple, but a very powerful classificationalgorithm

- Classifies based on a similarity measure
- Non-parametric
- Lazy learning
  - Does not "learn" until the test exampleis given

Whenever we have a new data to classify,we find its K-nearest neighbors from the training data

Example

- Training dataset consists of k-closestexamples in feature space
- Feature space means, space with categorization variables (non-metric variables)

- Learning based on instances, and thus also works lazily because instance close to the input vector for test or predictionmay take time to occur in the training dataset

V. EXPERIMENTAL RESULTS

For performance evaluation, we havecomputed five metrics: accuracy,Precision, Recall, F1-Score, and confusionmatrix.

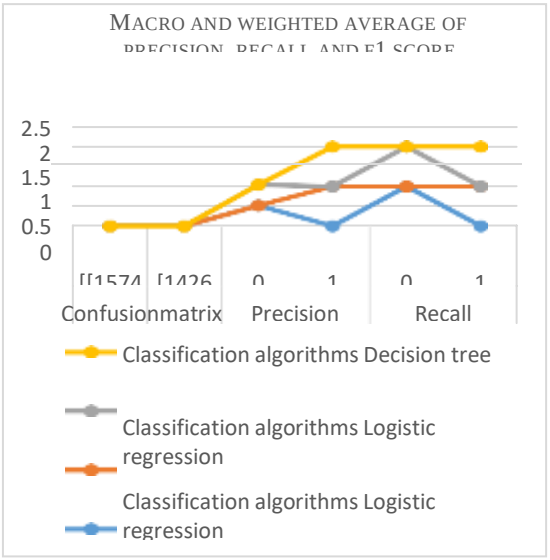
Where Precision is the portion of relative cases among the retrieved occasions, whileRecall is the division of the aggregate sumof relative cases that are retrieved. F1- Score is the average of Precision and Recall, while the Confusion Matrix is the measure of performance of an ML algorithm as

as explained in Table 1 and Table 2. In this paper, we consider an auto insurance fraud detection dataset and execute a sample that contains 110 customers with corresponding attributes. Table I shows that the eight

classification models have been validated using evaluation metrics such as precision, recall, and F1-score with correspondingMacro and weighted average as in Table 2.The results of the experiment have shown that Decision-Tree outperforms in all aspects such as execution time,

no sensitive to outliers, and the reduction of noise. The results obtained using the Classificationalgorithm outshines using real sample obtained from the reliable repository. For all the experiments in this section, the

performance shown is based on the test dataset. The Precision, Recall, F1-Score are computed using the equation. (1), (2) and (3)



Precision=True Positive/ (True Positive +False Positive) ... (1)

Recall = True Positive/ (True Positive + False Negative) … (2)

F1 Score=(2*Precision*Recall)/ (Precision +Recall) …. (3)

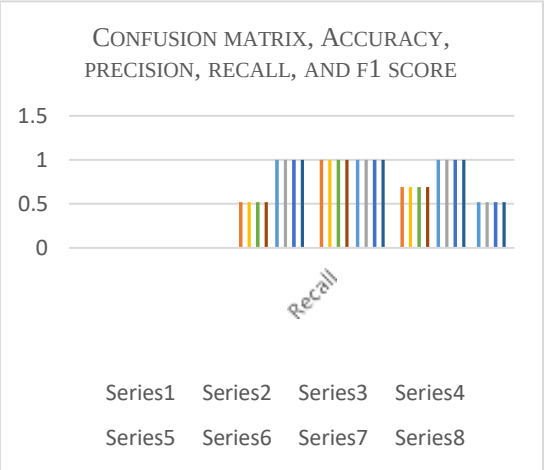
TABLE I. MACRO AND WEIGHTEDAVERAGE OF PRECISION, RECALL AND F1 SCORE

Figure 3: Macro and weighted averageof precision, recall and f1 score.

TABLE II. CONFUSION MATRIX, ACCURACY, PRECISION, RECALL, AND F1SCORE

Metrics	Average	Classification algorithms			
		SVM	Logistic Regression	Decision tree	KNN Classifier
Precision	Macro	0.26	0.26	0.26	0.26
	weighted	0.28	0.28	0.28	0.28
Recall	Macro	0.50	0.50	0.50	0.50
	weighted	0.52	0.52	0.52	0.52
F1 Score	Macro	0.34	0.34	0.34	0.34
	Weighted	0.36	0.36	0.36	0.36

Metrics	Classification algorithms							
	SVM		Logistic regression		Decision tree		KNN classifier	
Confusion matrix	[[1574 0] [1426 0]]		[[1574 0] [1426 0]]		[[1574 0] [1426 0]]		[[1574 0] [1426 0]]	
Precision	0	0.52	0	0.52	0	0.52	0	0.52
	1	0.00	1	0.00	1	0.00	1	0.00
Recall	0	1.00	0	1.00	0	1.00	0	1.00
	1	0.00	1	0.00	1	0.00	1	0.00
F1 Score	0	0.69	0	0.69	0	0.69	0	0.69
	1	0.00	1	0.00	1	0.00	1	0.00



user name, email, address and admin authorize the users.

Remote User

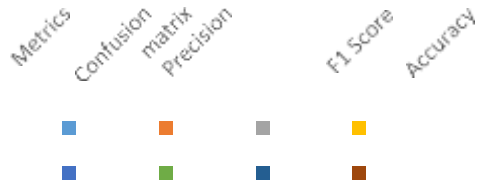


Figure 4: Confusion Matrix, Accuracy,Precision, Recall, and F1 score.

IMPLEMENTATION

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can dosome operations such as

Login, Browse and Train & Test Health Insurance Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction of Health Insurance FraudType, View Health Insurance Fraud Type Ratio, Download Predicted Data Sets, View Health Insurance Fraud Type Ratio Results, View All Remote Users

View and Authorize Users

In this module, the admin can view the listof users who all registered. In this, theadmin can view the user's details such as,

In this module, there are n numbers of usersare present. User should register beforedoing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT HEALTH INSURANCE CLAIM FRAUD TYPE, VIEW YOUR PROFILE.

Figure 5: Home Page

Figure 6: Login Page

Figure 7: User details

Figure 8: Login Details

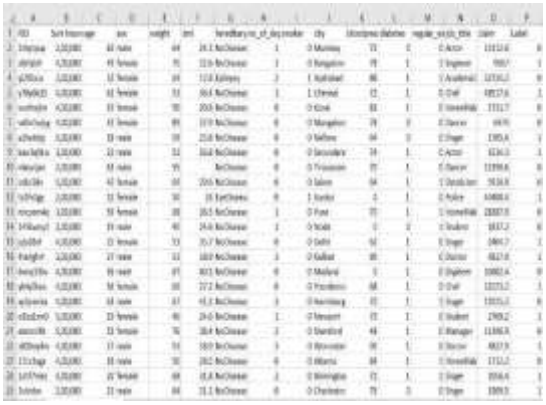


Figure 9: Dataset details

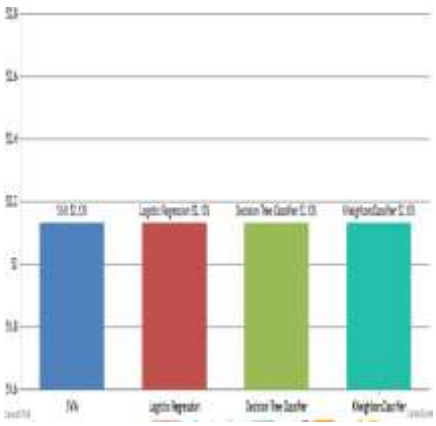


Figure 10: Output Results

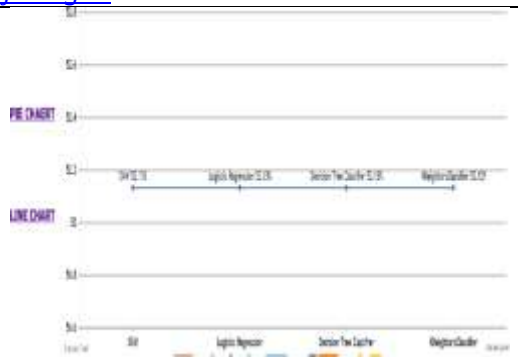


Figure 11: Output Results

VI. CONCLUSION AND FUTUREWORK

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 andefficiency 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 minimalinformation.

In the future, the fraud detection methodcan be extended to the Adaptive Neuro- Fuzzy Inference System (ANFIS) which is the combination of both Neuro-Fuzzy and Neural Networks. Hence, the prediction can be made more accurate and Hidden Markov Model (HMM) to predict fraud using internal factors.

VII. REFERENCE

[1] W. Zhang and X. He, “An anomaly detection method for Medicare fraud detection,” in Big Knowledge (ICBK), 2017 IEEE International Conference on. IEEE, 2017, pp. 309–314

[2] A. Urunkar, A. Khot, R. Bhat and N. Mudegol, "Fraud Detection and Analysis for Insurance Claim using Machine Learning," *2022 IEEE InternationalConference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, THIRUVANANTHAPURAM, India,2022,pp. 406-411, doi: 10.1109/SPICES52834.2022.9774071.

[3] R. A. Bauder and T. M. Khoshgoftaar, "Aprobabilistic programming approach for outlier detection in healthcare claims," in Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on. IEEE, 2016, pp. 347–354.