Detection Healthcare Frauds in Insurance Industry by Healthcare Service Providers

MSc Research Project
Data Analytics

Vinay Reddy Pannala
Student ID: X20138261

School of Computing
National College of Ireland

Supervisor: Bharathi Chakravarthi

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Vinay Reddy Pannala
Student ID:	X20138261
Programme:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Bharathi Chakravarthi
Submission Due Date:	30/01/2022
Project Title:	Detection Healthcare Frauds in Insurance Industry by Healthcare Service Providers
Word Count:	6359
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	30th JAN 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	Q
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	Q
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	Q

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Detection Healthcare Frauds in Insurance Industry by Healthcare Service Providers

Vinay Reddy Pannala
X20138261

Abstract

The frauds in insurance industry have exponentially increased over the past few years, this frauds claims leads to increase in spending's on insurance claims for insurance companies. According to the leading insurance companies, the loss in revenues for the organization is about 15%, which might be a potential threat to the insurance companies, so the organizations have to work on this fraud claims to protect the organization from potential risks as well as keeping the trust among the insurance policy holder. Hence the main goal of this project is to identify the frauds in insurance industry which is carried out by the fraud service providers mainly involves the physicians, as these physicians knows the codes for costliest procedures to claim which is impacting the insurance companies thus this leads the insurance companies to increase the premiums. The models implemented were into machine learning and neural network models in which logistic regression outperformed all the machine learning models and auto encoders from neural network models performed well with two hidden layers than the original auto encoder implementation.

1 Introduction

1.1 Background and Importance

The frauds in insurance industry have become multifold increase, which is concern area for most of the leading insurance industries all over the world. Most of the fraud identification done using manual verification which takes a lot of human effort and can cause huge loss in revenue and time for the organization, which can lead to increase in premium amount ultimately becomes costly for the customers to pay for the premiums and may result in losing the customer base. Thus, insurance companies need to focus to retain the customer by providing the fair premiums which only possible by eliminating the potential risk in the industry such as automating the identification of frauds claims.

In General, the health care frauds have been committed by the customers which are intentional, the lot of researches have been done in identifying the fraud claims done by the patients in health care industry and in the field of auto insurance claims, but there are few studies conducted on the healthcare service providers which mainly involves the hospital groups, physicians and peers, who comes together to involve in this kind of fraud. Thus, this project will try to identify the potential fraud claims which help the organization in terms of loss of revenue using machine learning.

1.2 Research Question

RQ: What extent can machine learning algorithms and neural network algorithms helpful in predicting the insurance fraud claims by health service providers using classification and auto encoder model?

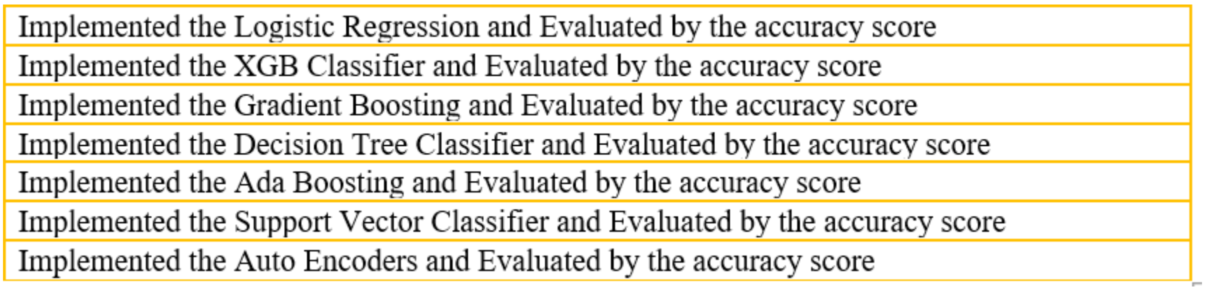
1.3 Research Objectives

For this research to conduct, the objectives begin with the literature review in the field of, fraud claims in insurance industry, the second objective was to collect the dataset, which was collected from the Kaggle which is publicly available. The next objective begins with the cleaning of dataset, such as handling missing values, nan values and identifying the important factors from the dataset to make it ready for the model implementation. The fourth objective was to implementation of various classification models with the cleaned dataset and evaluation. Final objective was conclusion and future work. The below are objectives discussed.

- Objective 1: Literature review in the field of insurance claims with respect to frauds.
- Objective 2: Identification of the data set and its ethical concerns.
- Objective 3: The pre processing of the dataset prior to the model implementation
- Objective 4: Feature extraction for the model implementation
- Objective 5: Implementation of the various classification and neural network models
- Objective 6: Final objective is to conclude the research and future work.

1.4 Research Project Contributions

The tasks were carried out in the two phases overall the project, such as the executing the machine learning models together after a data modification the neural network models were implemented on the different inpatient and outpatient health insurance data after configuring the required python libraries. The classification models were evaluated by accuracy and confusion matrix.



Implemented the Logistic Regression and Evaluated by the accuracy score
Implemented the XGB Classifier and Evaluated by the accuracy score
Implemented the Gradient Boosting and Evaluated by the accuracy score
Implemented the Decision Tree Classifier and Evaluated by the accuracy score
Implemented the Ada Boosting and Evaluated by the accuracy score
Implemented the Support Vector Classifier and Evaluated by the accuracy score
Implemented the Auto Encoders and Evaluated by the accuracy score

Figure 1: Objectives

The next phase of the research begins with the literature review for health care frauds claims in insurance industry. For the research approach the CRISP DM methodology was used the final parts of the research consists of the model implementations, evaluations and its results along with the conclusion and future work.

2 Literature Review of Frauds in Insurance Industry

2.1 Introduction

(Kowshalya and Nandhini (2018)) conducted research on insurance frauds be referred as, the healthcare insurance industries have taken into consideration as it is most popularly fastest growing sectors across the world and data concerned in this discipline is increasing exponentially, which has massive vulnerable to frauds, accordingly their research was performed on the fraud claims which were done by individual entities, however their studies were not at the organizational data but by means of generating the artificial record. Hence, for this research three different types of algorithms were used such as J48, Random Forest, Na¨ıve Bayes to calculate the premium for the customers based at the financial and private facts. Random forest model has performed better than other machine learning models in predicating the frauds claims.

During research, (Dhieb et al. (2020)), they have built advanced new system SISBAR for identification of fraud detection to help insurance organization by means of the usage of the block chain along with machine learning algorithms. The important goal of the research to identify the healthcare frauds and classify them. This research was performed by gathering the real information by using the usage of XGBoost, VFDT algorithms for detecting and classification of fraudulent claims in auto insurance and analysing the customers risk level. The Acritical intelligence, block chain and machine learning used to carry out the checks and simulation within the region of vehicle insurance, SISBAR system efficaciously helped the firms to reduce their claim refunds and enhance their overall performance. Their finding consists of client future behaviour and future claims. XGB outperformed with the higher accuracy as compared with other algorithms.

The comparative study of machine learning algorithms by using (Bauder and Khoshgoftaar (2017)), for predicting the frauds carriers has implemented using the supervised, unsupervised and hybrid models. 2015 healthcare statistics is gathered and the imbalances inside the class are reduced by way of the oversampling and 80-20 beneath sampling methods. During their sturdy, it is achieved that predicting the frauds in insurance industry is viable via the usage of the supervised and unsupervised machine learning models and hybrid models, finally the supervised models outperformed the alternative models by means of the use of particular strategies elegance imbalances sampling and the results are varies based on the issuer kind.

The fraud auditing and detection manual by (Busch (2012)) is extensively followed, and its analysis of existing data statistics in opposition to the predefined set of regulations was a big success in auditing the frauds, which worked efficiently, however its standard performance is not performed as expected, which is inaccurate at times. Addition to those fraudsters all approaches find out a way to overcome this built-in fraud detection guidelines. As there can be necessity to develop a possible solution that can be done by using of machine learning methodologies with enterprise information can come across the outliers.

2.2 Identification of Frauds using specific patterns followed

New trends in healthcare fraud detection can help in reduction of the losses by way of detecting the frauds at in advance stage and may help in increasing the equity in insurance industry. This study is conducted on the patient's behaviour who frequently claims for carrier that is truly fraud, in widespread legacy researches have executed to become aware of the fraudsters who frequently commits the fraud, with the aid of figuring out the normal behaviour of patients. During their study, the research is on locating the odd behaviour of the patients who behaves likes the regular patients and proposed affected person cluster divergence the usage of the healthcare insurance fraud detection that is PCDHIFD. The experimental effects have outperformed the legacy evaluation processes by using 15%, this study conducted with the aid of (Sun et al. (2019)).

During their observe (Liu et al. (2016)), on detecting the fraud in healthcare claims with the aid of analysing the data, the primary purpose in their examine is to discover the community-based totally frauds, wherein the entities such as starting from patients, issuer, pharmacy and son on are involved. As this network-primarily based fraud detection is first of its type. During their research, data records together with the information of thousands and thousands of patients and healthcare service providers and claimed offerings, medications concerning the multiple entity relationships and also focused on the man or woman entity, person attributes, pairwise relationships, the very last purpose of their look at was using graph techniques to find the clusters of frauds, which include doctors regarding their mutual coordinator and man or woman associated with pharma medicines, graph structure evaluation has performed main role in study of figuring out theses clusters.

The research performed by (Bauder et al. (2018)),at the detection of the healthcare frauds, that is ensuing in the higher rates for the beneficiaries and which is sometimes cannot be affordable for insurance beneficiaries , as a result this study is critical to are expecting the frauds and reduce the fraud claims by means of using the unsupervised algorithms, including KNN, Unsupervised Random Forest, in which unsupervised Random forest performed better , AUC is used as primary metric for estimating the accuracy, and KNN turned into carried out the use of the 1 neighbour and 5 neighbours, in which 5 neighbours KNN accomplished nicely, their very last findings algorithms was no longer up to the mark and executed poorly and has high false fantastic fee and suggested this can be applied as destiny paintings with right vendors statistics and automobile tuning of the functions used in device gaining knowledge of algorithms.

2.3 Existing Implemented Algorithms in Insurance domain

The studies conducted by (Rayan (2019)) the use of the supervised machine studying can is expecting the frauds primarily based on the historic data, but those algorithms can't be performed properly when the claims do not comply with the conventional regular claims which can be frauds. Thus, an experimental model desires to be developed for this study for having an alternative model which can be used as for extraordinary behaviour of claims, this research used a hybrid model of patients facts by way of combining the supervised and unsupervised with the guideline engine with anomaly detection. The integration of existing statistics of all of the entities worried on this fraud claims with this hybrid model alongside with the brand-new rule engine anomaly. The hybrid model was carried out as, model begin with KNN set of rules after which the result information passes as input to the decision tree and successful in predicting the beneficiaries claims most effective.

The outlier detection studies (Thomas and Judith (2020)) by, for detecting the frauds in insurance coverage is on most cancers statistics and cardiology collectively to discover the frauds claimed via beneficiaries by the use of the hybrid version with the svm, artificial neural network and evaluated by means of the use of f-1 score. The problem of their version is it cannot be finished big data and most effective finished with sure type of sicknesses and identifies the outliers regardless of the ordinary fraud behaviour.

The analysis by (Kirlidog and Asuk (2012)), at the turkey insurance claim information by using the svm, and gaussian navie bayes for predicting the frauds, the system advanced to discover and differentiate among legitimate and fraudulent claims, via that is conventional manner back in 2012, there have been no techniques for finding the limits. As that is completely primarily based at the historic facts and might have its own limitations for locating contemporary frauds claims.

Multiple research on fraud claims the use of device studying had been implemented but this study carried out through (Saldamli et al. (2020)), makes use of the block chain for finding the healthcare frauds by way of developing higher internet application through using reactjs and relaxation API service and automatic software program application which integrates with the blockchain which stores the health information which is shipped and saved within the more than one servers and authorized by centralised method. The method not concerned any gadget gaining knowledge of techniques as an alternative whole set of net utility gear have been used.

The auto insurance fraud claims have implemented through (Muranda et al. (2020a)), for detecting the fraud sample in car insurance enterprise by way of the usage of SVM and imbalances inside the information were eliminated by the adaptive synthetic sampling approach (ADASYN) technique, that is once more the legacy version and used simplest to classify the clusters of frauds, which may be underperformed with the modern-day fraud techniques adopted by fraudsters.

The statistical evaluation of locating the outlier inside the US health care is implemented through (Jing Li (2008)), to categorise the fraud behaviours and identifying the behavioural styles of the entities involved. This study has conducted on insurance fraud detection by categorising the different frauds and implementing the machine learning algorithms for better results and evaluated accordingly.

The survey of medical frauds claims by (Xie et al. (2016)), to come across the insurance frauds for China insurance organizations with large information is analysed by using Local outlier detection algorithm LOF. The outlier analysis will help the models in detecting the insurance frauds by eliminating the unnecessary things from the data by which models can only focus on the available data, then learn and implement it with frauds claims data(Muranda et al. (2020b)). Which introduces a way in which unwanted records to be omitted for improving the accuracy of model, and used Hadoop dispensed structures for performing fashions on huge information.

One of the literatures (Biwalkar et al. (2021)), makes use of statistical strategies and gadget learning algorithms on widespread health care records and discover the suitable appropriate model for the better accuracy while working with the healthcare facts. Though the selection tree and svn, knn are commonly used other than those they have got used chi-square and ANOVA, t-exams statistical methods to discover the health insurance on low value and frauds in healthcare and so on. SPSS used as statistical analysis and SAS and RStudio used for powerful predictive analytics.

Another Approach of fraud detection in healthcare is accomplished within the field of upcoding thinking about there are limited researches on this vicinity. Where (Richard Bauder (2017)), have implemented the device getting to know strategies which include supervised and unsupervised fashions are explored, although there is limitation on the auditing the fitness care information. Analysis centred mainly on upcoding where fraud claims have accomplished by way of converting the operation expenses using upcoding techniques by special entities worried in health care industry.

One of the literature evaluations by using (Hancock and Khoshgoftaar (2020)), primarily based at the overall performance of algorithms applied on large data, for frauds claims detection have used XGBoost, CatBoost algorithms with the under-sampling approach especially that specialize in improving the performance of fashions as there are lot studies have performed previously, with the aid of converting the sampling size from 100 to 250 till 1000 the special fashions were experimented and good overall performance achieved than the previous research by way of under sampling techniques.

The insurance organizations combined all together more than thousand companies across all countries which collects trillions of dollars premiums in each year. Any individual or peers or physicians make false insurance claims is a good way to achieve repayment or benefits to which they may be not entitled is insurance fraud. The total amount of these frauds is envisioned to be more than 40 billion. Hence by detecting these frauds in earlier stage will help the organizations to be profitable and other genuine customers can be benefited. The traditional approaches such as heuristic fraud detection and frauds identification is in auto sector has become most prominent frauds. In their research they have analysed the machine learning algorithm and evaluated by using the precision and recall (Roy and George (2017)).

The frauds in insurance industry have costed a lot in terms of cutting their revenue from operations every year in the past ten years, emerging technology has helped the fraudsters more than legacy techniques, hence detection of such incident earlier stage can be beneficial for the organization which can result in good profits growth for the organizations and its customer also get befitted from the proper premium and claims(Rawte and Anuradha (2015)). Hence the research done by , using the supervised and unsupervised techniques for combining the advantages of both techniques together to form a novel hybrid approach to detect frauds in insurance industry.

In identifying the frauds, the classification algorithms have played a huge role, when it comes to training the imbalanced data sets to the datasets with the improper data having outliers, missing values and many more. With the existing data with less information, it is very difficult for the algorithms to predict the proper output, hence the research is proposing to use adaptive synthesis sampling method (ADASYN) to eliminate any such data which can not be part of productive output. Hence used SVM to classify the frauds claims.

3 Research Methodology

3.1 Introduction

The research was done in the field of health care domain with identification of frauds in insurance claims which were majorly done by the health care service providers. As it is related industry business model, CRISP-DM methodology was used to carry out the research. The model's implementations took place and properly managed by following the CRISP-DM methodology.

3.2 Approach of the research methodology for fraud identification

For every research to be carried out, executing the steps from the CRISP-DM methodology is essential, as it is majorly for research which are related to the businesses. The research starts with the business understanding of the health care insurance industry and data understanding, data preparation, data cleaning, data modelling, evaluation and deployment were discussed below figure.

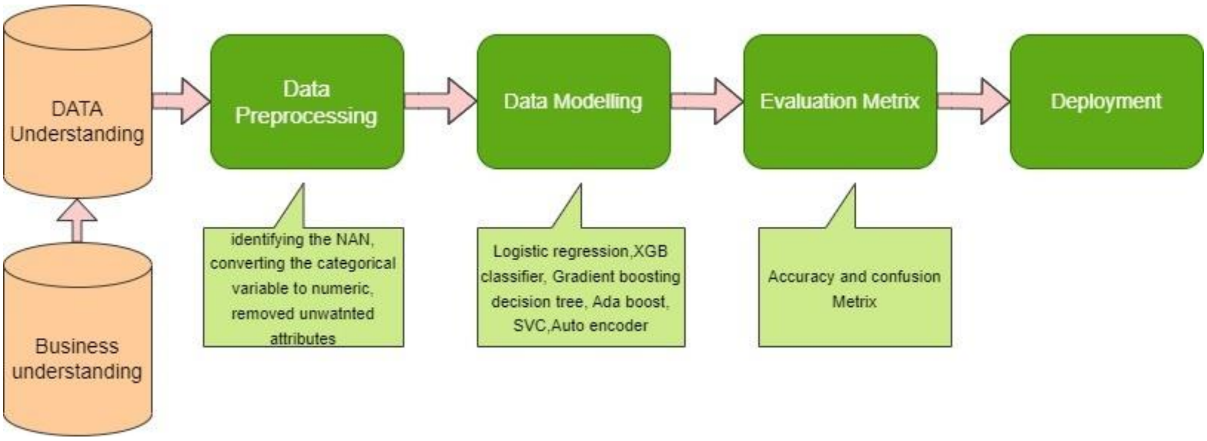

DATA Understanding
Business understanding
Data Preprocessing
Data Modelling
Evaluation Metrix
Deployment
identifying the NAN, converting the categorical variable to numeric, removed unwatnted attributes
Logistic regression,XGB classifier, Gradient boosting decision tree, Ada boost, SVC,Auto encoder
Accuracy and confusion Metrix

Figure 2: Methodology Approach

Business Understanding: This is the important step from the methodology to any insurance organization in terms of revenue, this research will take care of identifying the frauds claims done healthcare service providers, but in general the frauds claim from the patients, physicians, health care service providers as well. This research mainly focuses on the frauds done by the healthcare service providers, which helps the companies to identify the fruads in earlier stage by using the machine learning techniques which will result in growth of revenue and retain the customers by providing the right insurance premium.

Data Understanding: In this step the data set is collected, which consists of various different data sets of patients types such as inpatients data, outpatient data, beneficiary data, train data, and the same will repeat for the test data as well. The train data consists of the 5410 rows with 25 attributes, the test data consists of the 1353 rows, 1 attribute that is fraud claim or not with which we can validate the model. The beneficiary train data consists of 138556 rows and 25 attributes. The beneficiary test data consists of 63968 rows, 25 attributes. The inpatient train data consists of 40474, 30 attributes. The inpatient test data consists of the 9551 rows, 30 attributes. The outpatient train data consists of 517737 rows, 27 attributes. The outpatient test data consist of the 125841 rows and 27 attributes, which is collected from the Kaggle, available publicly as the open source.

Data Preparation: The step consists of the data cleaning such as the finding the missing values, nan values and make the data set ready before applying any machine learning techniques. The data set is prepared in such a way that it should be executable by the machine learning models without interruption, such as while implementing machine learning models some models does not allow categorical variables to be present in the data set. Hence converting such type of variables to numeric was done in this pre-processing step.

Data Modelling: As the target variable is to identify the frauds claims or not fraud claim which comes under binary classification, hence the classification machine learning models were implemented such as Logistic regression, XGB Classifiers, Gradient boosting, Decision tree classifier, ADA boost, Auto Encoders from neural network machine models for the health insurance data set to find the frauds health service providers.

Evaluation: The models implemented in the previous step were evaluated by evaluation matrix such as confusion matrix which will give insights of the model performance to identify the desired output obtained or not. With the help of accuracy, the best model was identified by comparing the with other implemented models. Some other evaluation metrics were also considered such as precision, recall, f1.

Deployment: The final stage is to form textual presentation of its actual findings, though the model implementation was not only the ultimate step but it has to be properly documented its findings. The information documented has to be proper such that the final user has to understand the get benefitted from the addition of the model implementation. It can be an addition to the insurance industry in identifying the fraud claims.

3.3 Data collection Pre-processing

The data set was collected from the Kaggle [1] which was publicly available with no ethical restrictions, the data set consists of the two parts train and test data in which the data is grouped together that the outpatient data, inpatient data, and the beneficiary data

[1] https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis

the data set with the fraud claims done or not. The data set was then processed before implementation of the model implementation for further analysis. The data set contains the information in csv file formats, later the required csv libraries were imported in python libraries.

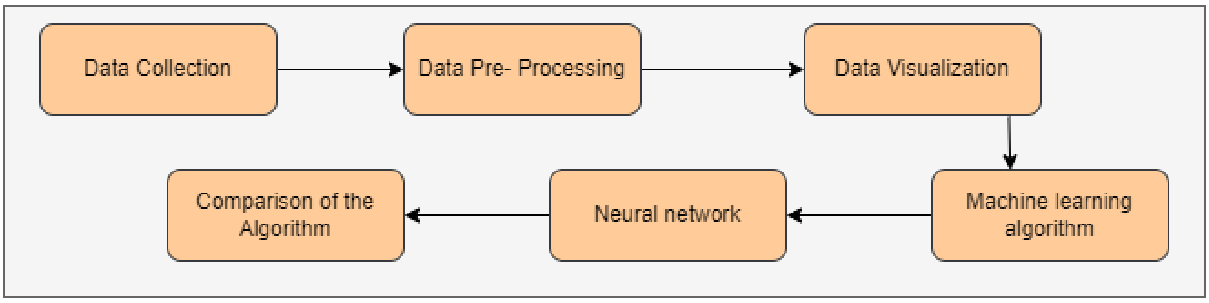


Figure 3: Data Collection and Preprocessing steps

The data set was verified with the missing values and nan values, they were handled properly with appropriate solution such as mean replacement. Some of the models does not allow the categorical variables while running the models hence the categorical variables into numeric variables which helped the models to run smoothly. The data set was spitted into train and test data then classification models were implemented and evaluated.

3.4 Architectural and Technical Design

The research was implemented in a three phase as shown in the figure 4, the first phase the data set was visualised using the python plot library using the important correlation between the features were visually represented. In the second phase the classification models were implemented in different phased manner, the machine learning models such as logistic regression, xgb boost, svc, adaboost, apart from this from the neural network models the auto encoders was implemented. The final phase data persistence phase, the python language was used in this phase for implementing the models.

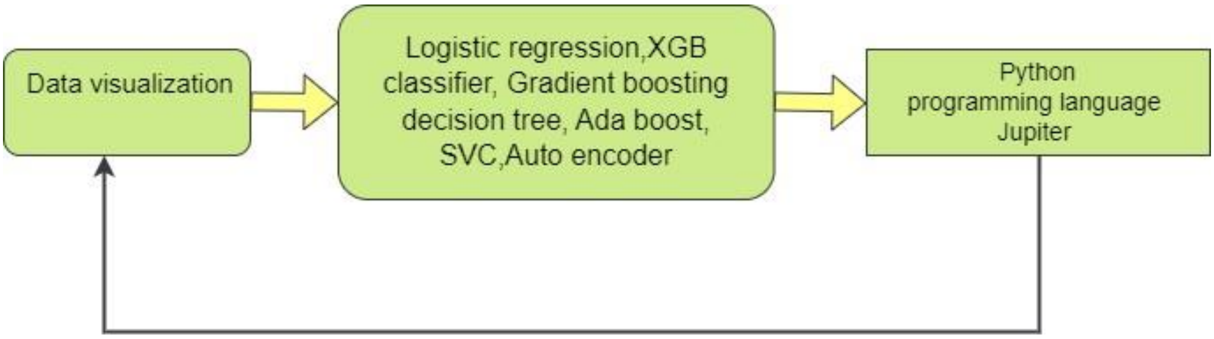

Data visualization
Logistic regression,XGB classifier, Gradient boosting decision tree, Ada boost, SVC,Auto encoder
Python programming language Jupiter

Figure 4: Architectural Design

3.5 Methodology Conclusion

The steps mentioned above were specified about methodology, the research implementation done in the three-phase manner which are visualization, where the in-depth analysis of the data features was done and correlation of the variables were identified. The second phase in the data persistence layer where the data manipulation and the language of implementation and the required libraries were collected. At the end where the business model where the machine learning models were implemented.

4 Implementation

In this step first the data set was collected and then the machine learning models and the neural network models were implemented with the help of the accuracy and confusion matrix the results of the model were analysed further to analyse and identify the research question which was mainly to identify the frauds claims filed by the health care service providers.

4.1 Introduction

The data set was originally collected from the website Kaggle which is publicly available and which has no ethical concerns related to the data privacy and ethical concerns. The model implementation and evaluation was done with the several machine models and one neural network models with multiple epochs to achieve better results which were evaluated by accuracy and confusion matrix.

4.2 Data Analysis

The data analysis phases the relation between the features were evaluated to find out the best suitable features from all the data sets which helped the models to perform the better and some of the graphs were discussed below. The analysis also includes the categorical variables conversion from categorical to numeric so that smooth implementation of the models takes place.

From the figure 5, after the both the train and test data analysed together to find the potential fraud distribution in aggregated claim transitional data was described such as fraud claims were almost more than half of the non-frauds claims. The count of the non-fraud claims are around the 350000 then fraud claims were about 200000.

From the figure 6, it is observed that the maximum number of claims are from the state number 5, which were converted to numeric while implementing the machine learning models. This analysis can further help the research model, by identifying the states in which most umber of claims are being filed and can focus further more on them can identify the fraud claims. Some of the machine learning models can not allow the variables names in string format, hence the state names were converted to the numeric values such that the models can execute without interruption.

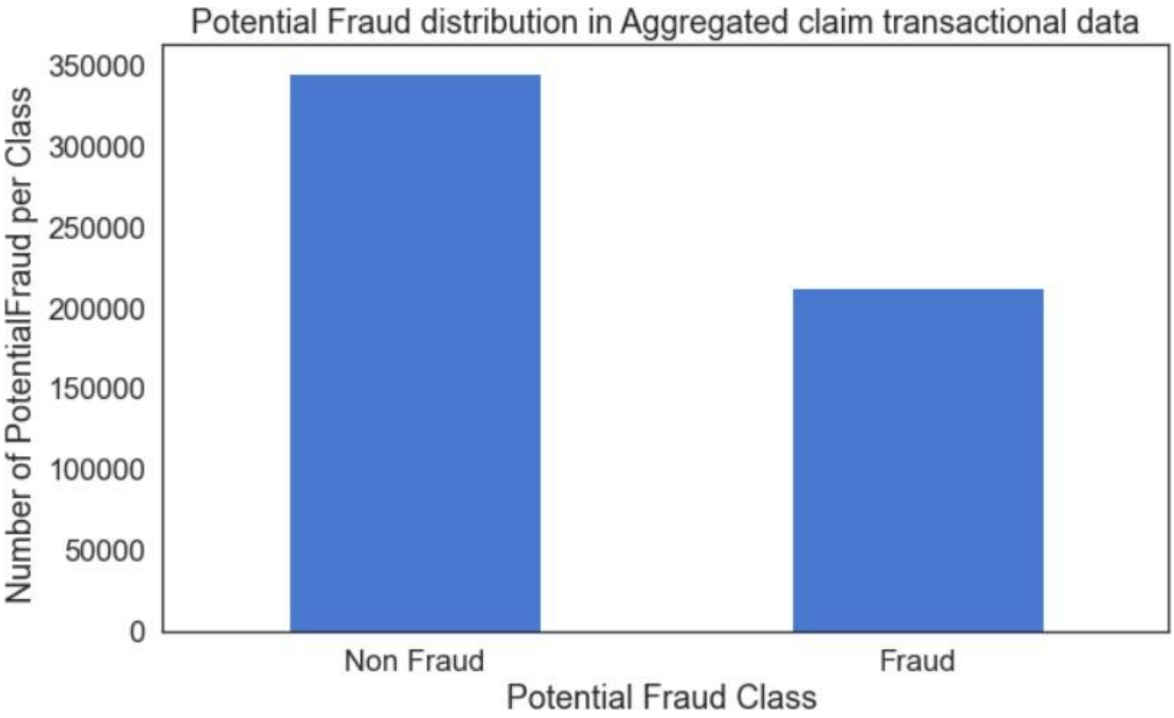


Figure 5: Potential fraud claims

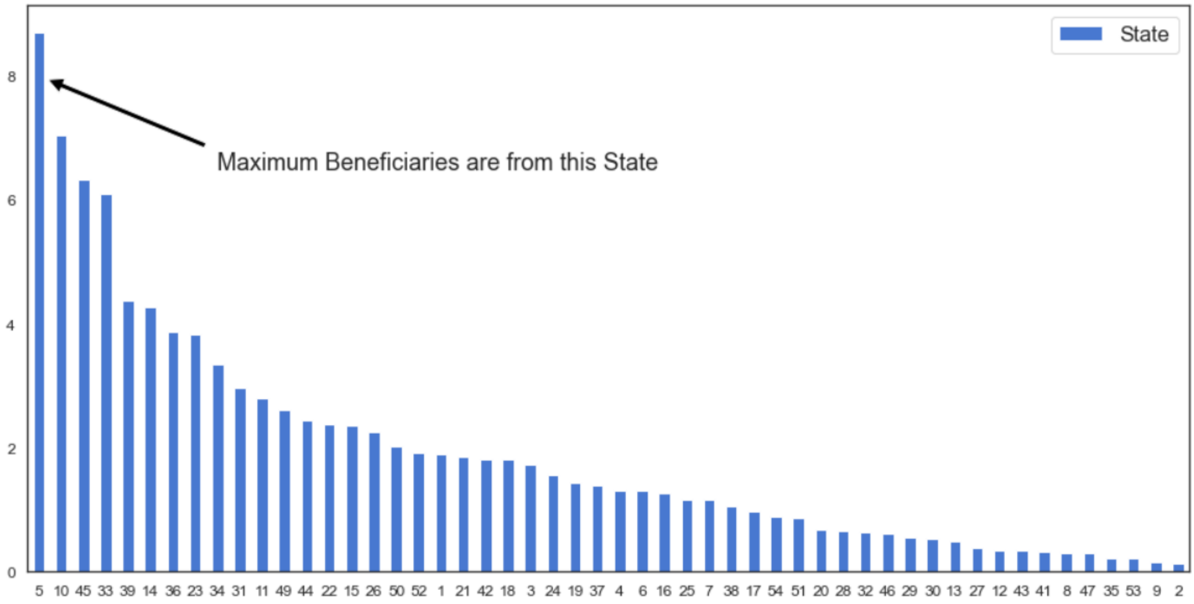


Figure 6: Beneficiaries from each state

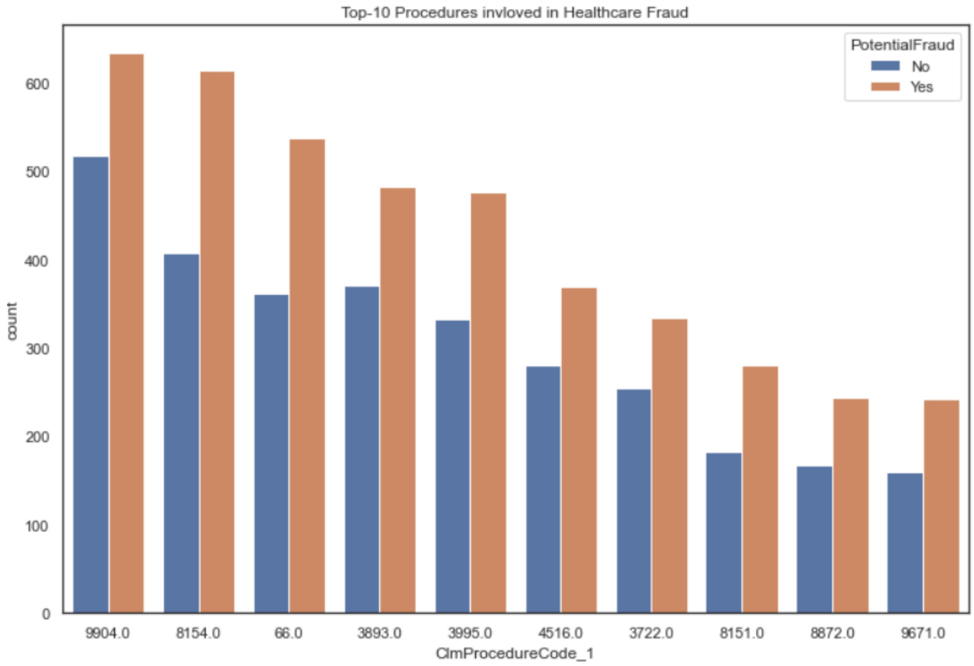


Figure 7: Number of Claim Procedures

From the figure 7, it is observed that the top ten procedures provided in healthcare fraud, the potential frauds number is high for the particular procedure and less for the particular procedure , these can help significantly in finding the frauds when claiming frequently for the same procedure and which has been costlier procedure, which can cost the organization and can cause huge loss , which can be minimised by the adoption these machine learning algorithms to help mitigate the risk and help grow the profits for the organization.

From the above figure 8 it is observed that the claim dignosied code has filed for fraud or not, and the count of the frauds have been identified , all the dignosied codes are converted from string format to numeric format such that the models were executes without interruption.

From the figure 9 it is observed that the number of frauds done by the existing physicians which will help the train the model and can predict the frauds done by the physicians in future by analysing the which procedure are claiming mostly and what are the frequent procedures are costlier and claiming in regular basis. These are the things the model gets trained with to help predict future frauds in healthcare fraud claims identification.

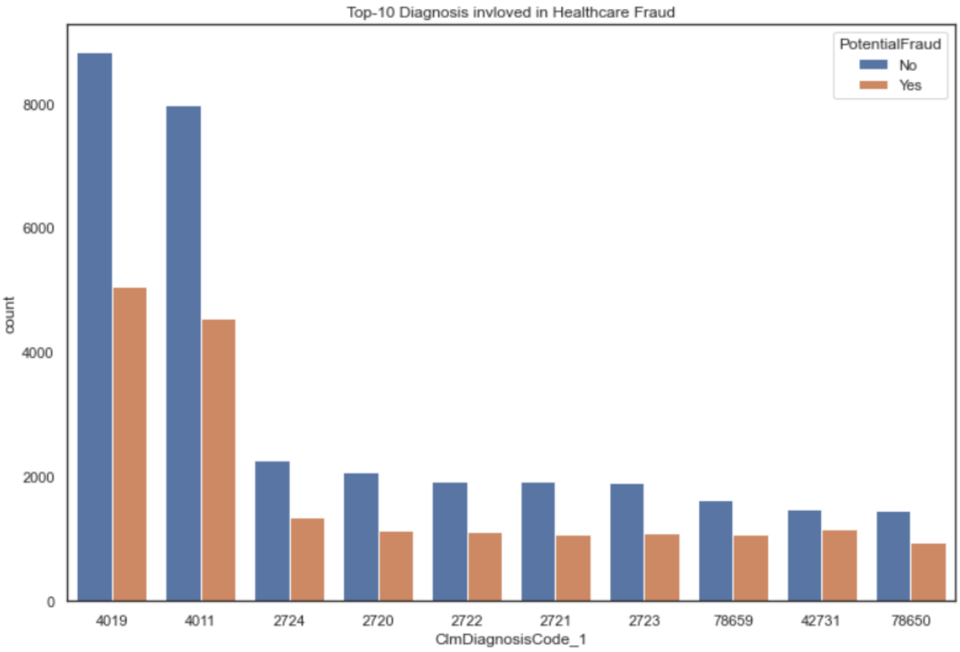


Figure 8: Number of Claim Diagnosis

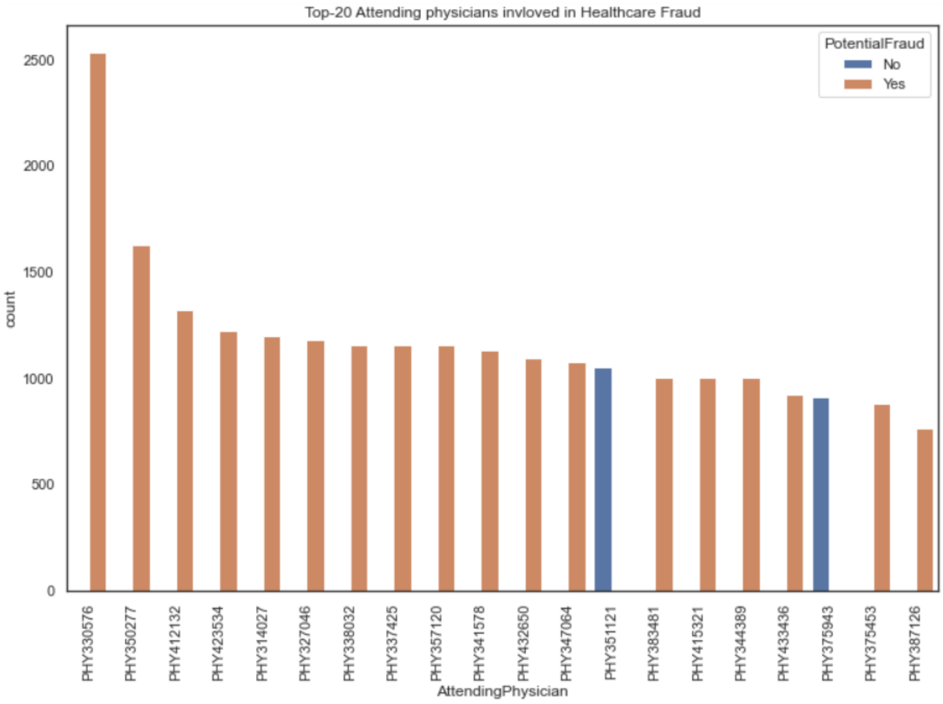
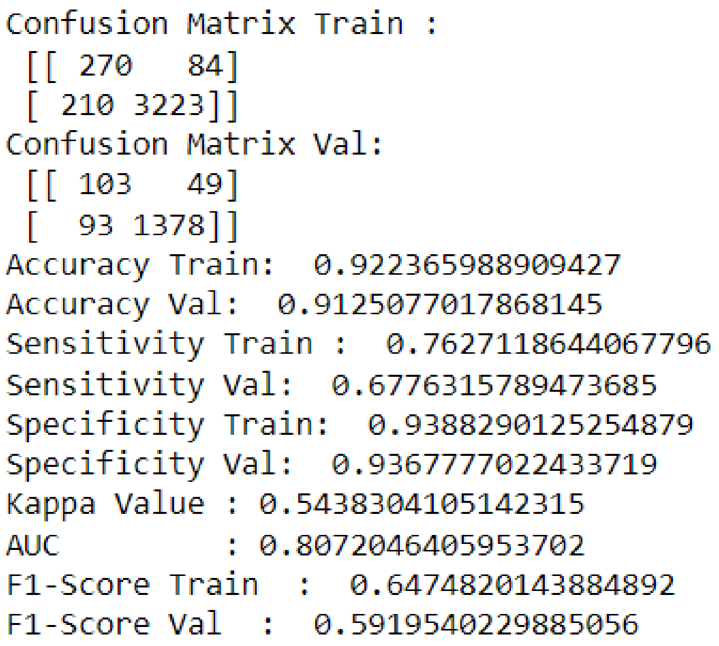


Figure 9: Attending Physcians

4.3 Model Implementation

As the research topic was on classification problem, the logistic regression model was one of the efficient model in identifying the classification problems, The data set was divided into the train and test data set and then the model was implemented with the proper data modelling and important features were finalised before implementing the model. The result of the model were discussed, the train and test data was evaluated by using the confusion matrix and the model was successful in identifying the frauds claims about 91% in both train and test data.

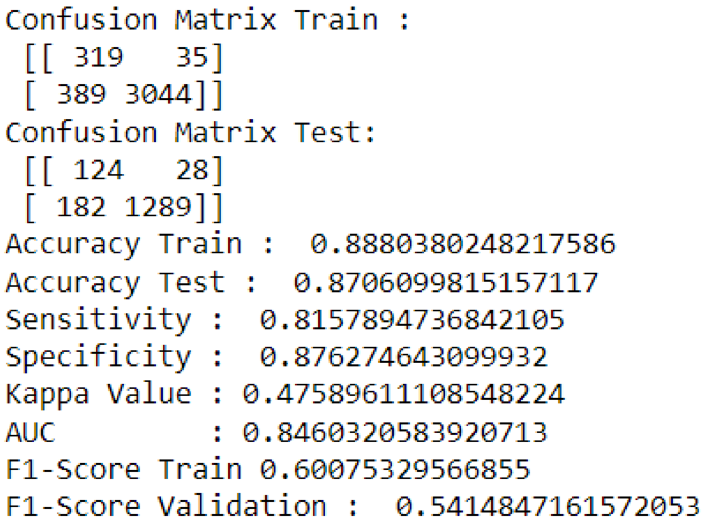


```
Confusion Matrix Train :
 [[ 270   84]
 [ 210 3223]]
Confusion Matrix Val:
 [[ 103   49]
 [  93 1378]]
Accuracy Train:  0.922365988909427
Accuracy Val:  0.9125077017868145
Sensitivity Train :  0.7627118644067796
Sensitivity Val:  0.6776315789473685
Specificity Train:  0.9388290125254879
Specificity Val:  0.9367777022433719
Kappa Value : 0.5438304105142315
AUC         : 0.8072046405953702
F1-Score Train  :  0.6474820143884892
F1-Score Val  :  0.5919540229885056
```

Figure 10: Logistic Regression

The research problem was to identify the frauds in healthcare sector, as it is the classification problem several classification models were implemented, the most used machine learning model random forest for both classification problems was applied and the result was evaluated by confusion matrix and accuracy. The accuracy of the model was around 88% for both the train and test data set, which was less than the logistic model.

At the end the models were implemented together to find out the best performing model for both the data sets and the logistic regression model was outperformed all the models. The results were discussed below. The ADA boost algorithm was significant in predicting the classification binary problem such as fraud or not, and it was the first real time boosting machine models which is helpful in all the fields. For ADA boosting algorithm the f1 score was around 53%. The decision tree classifier algorithm was efficient in identifying the classification model, where it contains binary tree and non-binary tree node structures, the decision starts from the root node which helped to achieve the goal, for the decision tree classifier the f1 score was around 43%.

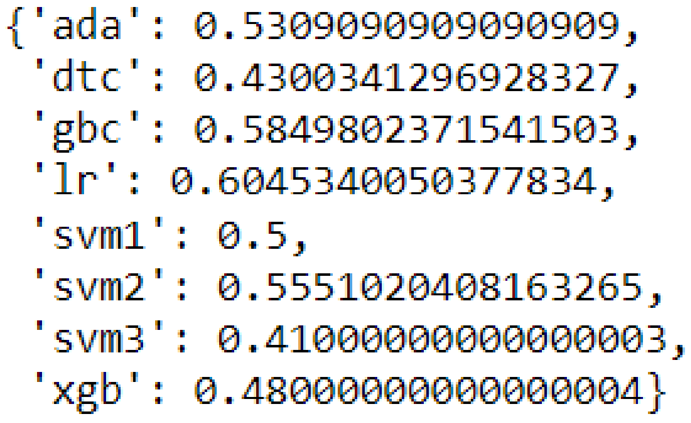


```
Confusion Matrix Train :
 [[ 319   35]
 [ 389 3044]]
Confusion Matrix Test:
 [[ 124   28]
 [ 182 1289]]
Accuracy Train :  0.8880380248217586
Accuracy Test :  0.8706099815157117
Sensitivity :  0.8157894736842105
Specificity :  0.876274643099932
Kappa Value : 0.47589611108548224
AUC         : 0.8460320583920713
F1-Score Train 0.60075329566855
F1-Score Validation :  0.5414847161572053
```

Figure 11: Random Forest

The gradient boosting algorithm was one of the most significant algorithms in boosting algorithms to identify the classification problems, the gradient boosting classifier, logistic regression model was 60%. The support vector machine algorithm is mostly used in both the regression and classification problem, generally it performs better than the other classification algorithms. For support vector machine f1 score was around 55%. XGB model f1 score was 48%, but among all the machine learning algorithms, f1 score of the logistic regression f1 score outperformed all the other models in terms of identifying the fraud claims in health care data set.

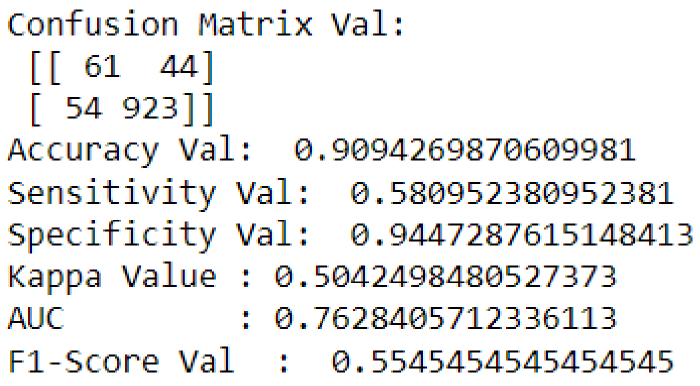
From the neural network models the auto encoder model was implemented with the different ephocs in multiple implementations, the result was improved with the variation of ephocs count, the auto encoders model was implemented to verify the performance of neural network models compared to machine learning models, and model outperformed over all the machine learning models with the good accuracy in predicting the frauds



```
{'ada': 0.5309090909090909,
 'dtc': 0.4300341296928327,
 'gbc': 0.5849802371541503,
 'lr': 0.6045340050377834,
 'svm1': 0.5,
 'svm2': 0.5551020408163265,
 'svm3': 0.41000000000000003,
 'xgb': 0.48000000000000004}
```

Figure 12: Models Results Evaluation

claims. With 2 different hidden layers the model outperformed. The results were discussed below.



```
Confusion Matrix Val:
 [[ 61  44]
 [ 54 923]]
Accuracy Val:  0.9094269870609981
Sensitivity Val:  0.580952380952381
Specificity Val:  0.9447287615148413
Kappa Value : 0.5042498480527373
AUC         : 0.7628405712336113
F1-Score Val  :  0.5545454545454545
```

Figure 13: Auto Encoders

5 Conclusion and Future Work

As the insurance organization's main goal is to have good revenue, which will come from having more number customer base and retention rate of the customers, every health insurance organization is have some risks to mitigate one such risk is the frauds claims in insurance. This research was mainly focused on the identifying the potential frauds which can cause the insurance companies a lot in terms of revenue.

Various machine learning models were implemented, After the research work done in the same field of work the data was collected and the data set was analysed and modified as required per the model implementation, the model implementation starts with Ada boosting model which has accuracy of 53%, then decision tree classifier was implemented with the accuracy of 43%, gradient boosting model was implemented with the model accuracy of 58%, the logistic regression model was which was outperformed all the models with 60% accuracy, the model support vector machine was implemented with the 55% accuracy, then the at the end the xgb model was implemented, then with neural network model the auto encoders have been implemented with the different ephocs and different hidden layers to get the best accuracy which was achieved around 90%, which outperformed the machine learning models. The auto encoders model helped the insurance industries in identifying the fraud claims with good accuracy by analysing the outpatient, inpatient and beneficiaries data in right manner, this machine learning models which can help the insurance organization reduce the fraud claim risks by pre identifying the fraud, this research was identified the major threats the insurance organizations were facing.

The major things were addressed in this research was feature importance by using the correlation of the attributes , the attending physicians can claim for the health insurance in most costliest diagnosis, the frequent claims done by the same physicians and the patients information being forged similar kind of attributes from the inpatient, outpatient , beneficries data set were used by the models to identify the frauds, these can be improved further with the advancements in the technologies in machine learning and as well as neural network models with different epochs and different hidden layers of various neural network models.

6 Acknowledgement

I would especially thank my supervisor Bharathi Chakravarthi for supporting me through out the research.

References

Bauder, R. A. and Khoshgoftaar, T. M. (2017). Medicare fraud detection using machine learning methods, *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 858–865.

Bauder, R., da Rosa, R. and Khoshgoftaar, T. (2018). Identifying medicare provider fraud with unsupervised machine learning, *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 285–292.

Biwalkar, A., Gupta, R. and Dharadhar, S. (2021). An empirical study of data mining techniques in the healthcare sector, *2021 2nd International Conference for Emerging Technology (INCET)*, pp. 1–8.

Busch, R. S. (2012). Healthcare fraud: Auditing and detection guide. new york: Wiley.

Dhieb, N., Ghazzai, H., Besbes, H. and Massoud, Y. (2020). A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement, *IEEE Access* **8**: 58546–58558.

Hancock, J. and Khoshgoftaar, T. M. (2020). Performance of catboost and xgboost in medicare fraud detection, *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 572–579.

Jing Li, Kuei-Ying Huang, J. J. . J. S. (2008). A survey on statistical methods for health care fraud detection., *A survey on statistical methods for health care fraud detection.*, pp. 1–5.

Kirlidog, M. and Asuk, C. (2012). A fraud detection approach with data mining in health insurance, *Procedia - Social and Behavioral Sciences* **62**: 989–994.

Kowshalya, G. and Nandhini, M. (2018). Predicting fraudulent claims in automobile insurance, *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 1338–1343.

Liu, J., Bier, E., Wilson, A., Guerra-Gomez, J. A., Honda, T., Sricharan, K., Gilpin, L. and Davies, D. (2016). Graph analysis for detecting fraud, waste, and abuse in healthcare data, *AI Magazine* **37**(2): 33–46.
URL: *https://ojs.aaai.org/index.php/aimagazine/article/view/2630*

Muranda, C., Ali, A. and Shongwe, T. (2020a). Detecting fraudulent motor insurance claims using support vector machines with adaptive synthetic sampling method, *2020 61st International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS)*, pp. 1–5.

Muranda, C., Ali, A. and Shongwe, T. (2020b). Detecting fraudulent motor insurance claims using support vector machines with adaptive synthetic sampling method, *2020 61st International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS)*, pp. 1–5.

Rawte, V. and Anuradha, G. (2015). Fraud detection in health insurance using data mining techniques, *2015 International Conference on Communication, Information Computing Technology (ICCICT)*, pp. 1–5.

Rayan, N. (2019). Framework for analysis and detection of fraud in health insurance, *2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pp. 47–56.

Richard Bauder, T. M. K. . N. S. (2017). A survey on the state of healthcare upcoding fraud analysis and detection. health serv outcomes res method, *A survey on the state of healthcare upcoding fraud analysis and detection. Health Serv Outcomes Res Method*, pp. 1–8.

Roy, R. and George, K. T. (2017). Detecting insurance claims fraud using machine learning techniques, *2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT)*, pp. 1–6.

Saldamli, G., Reddy, V., Bojja, K. S., Gururaja, M. K., Doddaveerappa, Y. and Tawalbeh, L. (2020). Health care insurance fraud detection using blockchain, *2020 Seventh International Conference on Software Defined Systems (SDS)*, pp. 145–152.

Sun, C., Li, Q., Li, H., Shi, Y., Zhang, S. and Guo, W. (2019). Patient cluster divergence based healthcare insurance fraudster detection, *IEEE Access* **7**: 14162–14170.

Thomas, R. and Judith, J. (2020). Hybrid outlier detection in healthcare datasets using dnn and one class-svm, *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1293–1298.

Xie, Z., Li, X., Wu, W. and Zhang, X. (2016). An improved outlier detection algorithm to medical insurance, *in* H. Yin, Y. Gao, B. Li, D. Zhang, M. Yang, Y. Li, F. Klawonn and A. J. Tallón-Ballesteros (eds), *Intelligent Data Engineering and Automated Learning – IDEAL 2016*.