International Journal of Information Management Data Insights 1 (2021) 100012

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

journal homepage: www.elsevier.com/locate/jjimei

Application of machine learning and data visualization techniques for decision support in the insurance sector

Seema Rawat[a], Aakankshu Rawat[a], Deepak Kumar[b,*], A. Sai Sabitha[a]

[a] Department of Information Technology, Amity School of Engineering and Technology (ASET), Amity University Uttar Pradesh, Sector 125, Noida-201313, Gautam Buddha Nagar, Uttar Pradesh, India.

[b] Amity Institute of Geoinformatics & Remote Sensing (AIGIRS), Amity University Uttar Pradesh, Sector 125, Noida-201313, Gautam Buddha Nagar, Uttar Pradesh, India.

ARTICLE INFO

Keywords:
Insurance
Claim analysis
Classification
Data mining
Exploratory data analysis
Feature selection
InsurTech
Machine learning

ABSTRACT

The insurance industry has a giant role in the sustainable economic growth of any country. With an increase in the number of insurance buyers, it has become an absolute necessity for an insurance company to have a detailed claim analysis system in place. Claim Analysis is performed by insurance companies to distinguish between fraudulent and genuine claims. Apart from that, Claim Analysis can also be used to understand the client strata in a much better way and implement the results further during the underwriting and acceptance/denial stage of policy enrollment. The main objective of this research work is to identify meaningful and decisive factors for claim filing and acceptance in a learning context through exploratory data analysis (EDA) and feature selection techniques. Also, machine learning algorithms are applied to the datasets and are evaluated using performance metrics.

1. Introduction

Insurance is financial protection against any type of risk. This protection is defined by a contract between two parties: the insurer and the insured (beneficiary). The insurer is the company that sells the policies and the insured is the person or the party which buys the insurance policy for its benefits. The insurer agrees to take risk of an insured entity against future events in return for monetary compensation known as Premium (Aswani, Ghrera, Chandra, & Kar, 2020). In case of an unforeseen event, the insurer must pay a claim to the insured i.e. the benefit amount to be paid to the beneficiary according to the policy document. Depending on the types of risk involved, insurance policies are of different types. Life Insurance, Travel Insurance, Health Insurance, Auto Insurance and Property Insurance are some of the different Line of Business' (LOB) in the insurance industry (Bacry et al., 2020). Another type of insurance is Re-insurance in which an insurance company purchases insurance from another insurance company to protect itself from financial risk due to the event of an enormous claim Barry & Charpentier (2020). The whole industry runs on the concept of risk or financial loss reduction (Batra, Jain, Tikkiwal, & Chakraborty, 2021). In such a deal, where the insurer has to ensure the client against any type of financial loss against any unforeseen event, there must be a way for the insurer or the insurance company to manage their transactions to pay claims as well as generate sufficient profit to survive in the industry Blackstone (2013).

"InsurTech" is the term used in association with bringing new technologies and innovations in the field of insurance to impact the regulatory practices of insurance markets. Here comes the concept of "Big Data". Big Data has transformed the way insurance companies deals with the enormous amount of data that they receive. Due to the advancement of technologies, the volume of data to be dealt with has enormously increased and this data can be structured, semi-structured as well as unstructured (Chakraborty & Kar, 2017; Kar, 2016). Initially, and even now, many insurance companies use actuarial formulas for underwriting and mortality table for life expectancy assessment. Big Data technologies help to analyze data and extract important information from it that can lead to better decision making and strategic business development as compared to the above-mentioned traditional data-processing techniques (Chowdhury, Mayilvahanan, & Govindaraj, 2020; Karhade et al., 2019). Data Mining plays an important role in various aspects of the insurance industry such as risk assessment, fraud detection, underwriting analysis, claim analysis, marketing analytics, product development, customer profiling etc (D. Das, Chakraborty, & Banerjee, 2020; S. Das, Datta, Zubaidi, & Obaid, 2021). Along with Data Mining, the industry is shifting towards machine learning algorithms to predict using the analysis done on big datasets for better fraud detection,

* Correpsonding author.
E-mail addresses: srawat1@amity.edu (S. Rawat), aakankshurawat@gmail.com (A. Rawat), deepakdeo2003@gmail.com, dkumar12@amity.edu (D. Kumar), assabitha@amity.edu (A.S. Sabitha).

https://doi.org/10.1016/j.jjimei.2021.100012
Received 4 February 2021; Received in revised form 17 March 2021; Accepted 18 March 2021

KYC verification, behavioural policy evaluation and custom claim settlements. The use of Machine Learning (ML) is increasing tremendously in the insurance industry despite initial resistance by the industry because of its explicitly recursive approach in predictive modelling which helps to improve the model at each repetition. In this analysis, we will deal with the claim analysis aspect of the insurance industry (Dave, Patwa, & Pandit, 2021; Doupe, Faghmous, & Basu, 2019). ML is used in claim analysis & processing for triaging claims, identifying outlier claims & even fraud, and automating where possible i.e. reducing the human intervention in claim processing and making the whole process hassle-free (Gupta, Kar, Baabdullah, & Al-Khowaiter, 2018; Kakhki, Freeman, & Mosher, 2020). The use of ML algorithms in this process helps the company to understand the beneficiary's claim filing pattern as well their claim acceptance pattern which can be used to optimize the whole process flow for policy enrollment Kar (2016).

The purpose of this research is to understand how machine learning algorithms can help in insurance companies to deduce patterns in various segments/branches of InsurTech (claim analysis is done in this research paper). In this analysis, two datasets are used to perform claim analysis using different classification algorithms. Three feature selection algorithms have been used to reduce the dimensionality of the data and to improve the results of the analysis. The algorithms are finally evaluated and compared based on four widely approved and trusted metrics: accuracy, precision, recall and f1-score. The following sections of the research work are literature review, proposed computational methodology, experimentation, results, discussion and conclusion.

2. Literature review

2.1. All about "InsurTech"

InsurTech is simply bringing technology to the world of insurance. InsurTech companies have been able to come up with several business models to improve the contracting process as well the claims management process, helping customers to clearly understand where their premium goes as well as how their premiums are structured. InsurTech technologies and innovations have helped to reach out to all the sections of the economy, especially the lower-income bracket and the less developed markets. Most of the InsurTech companies are startups and due to the huge scale of customers they attract, many reinsurance companies invest in them. But while bringing new technologies to the market, they must be aware of the regulations of the insurance industry and meet them. Many countries have a well-established regulatory sandbox approach to allow InsurTech companies to enter the market so that such startups can grow their business models as well as keep up with the regulatory requirements. India has a regulation policy in place for micro insurance regulations since 2015 (Kasy, 2018; Kaur, Sharma, & Mittal, 2018). Apart from the benefits, InsurTech needs to balance the need for innovation as well as the need for the protection of data. There is a need for close inspection of the internal controls that ensure that the client's data use is free of bias and the law is adhered to.

2.2. Need for artificial intelligence (AI) & machine learning (ML) in insurance

The insurance industry is one of the oldest industries in the world with marine insurance as the first form of insurance, used for self-insurance. With the increase in client data, there is a huge scope for improvement in methodologies used in the sector for policy enrollment and claims settlement (Khan, Bashir, & Qamar, 2014; Knighton et al., 2020). Artificial Intelligence is the buzzword of the decade and it simply means the enhancement of machine to simulate human intelligence (Kose, Gokturk, & Kilic, 2015; Kraus, Feuerriegel, & Oztekin, 2020). AI encompasses ML and predictive analysis which is taking the insurance industry by storm. AI helps in almost all the business needs of an insurance company. AI speeds up the process of underwriting, as well as improves risk selection and pricing strategies (Larson & Sinclair, 2021; Maehashi & Shintani, 2020). Not only that, but different insights can also be drawn from the client's data as well as personalization of policies has become easier due to faster underwriting and analysis (McGlade & Scott-Hayward, 2019; Mita et al., 2021). Claim processing has become automated using mobile applications (Nian, Zhang, Tayal, Coleman, & Li, 2016). A thorough analysis can help with fraud detection. Insurance fraud is a 40 billion dollar industry per year, hence, the use of ML techniques can help to notify brokers in case of fraud to further investigate. The analysis tools used to gather insights on customer behaviour can also be used to gather insights on employees to retain valuable talent. This can be done by understanding the behaviour, interests, learning styles of the brokers as well as their job satisfaction and the potential to look out for a job somewhere else (Ozbayoglu, Gudelek, & Sezer, 2020; Sengupta et al., 2020). Lastly, AI & ML can be used for the effective marketing of insurance policies. Table 1 summarizes the major work done in the insurance sector.

2.3. Claim analysis in insurance sector

Statistics has been in use in the insurance industry from the onset of the industry. There is a whole discipline for the use of statistics in the insurance industry known as Actuarial Science. With the massive increase in data processed by the industry, Predictive Analytics is coming into the limelight (Pal, Mandana, Pal, Sarkar, & Chakraborty, 2012; Yang et al., 2021). It encompasses data mining, predictive modelling, and machine learning techniques like classification, regression, clustering and outlier detection to make accurate and fast predictions about unforeseen events in the future using the current data (Palanisamy & Thirunavukarasu, 2019).

Claim Analysis is an important aspect of Predictive Analytics in the insurance industry as approximately 80% of the premium revenue generated is spent on claims by the insurance companies Pappas & Woodside (2021). Hence, it is essential to do a thorough analysis of claims to improve cash flow. By analyzing the insurance data, relations between various factors (variables) is observed and a function is derived to model predictions (Petropoulos, Siakoulis, Stavroulakis, & Vlachogiannakis, 2020). These predictions can be used for making decisions. Apart from the structured data used by the companies, there is a huge scope of unstructured data that provides vital information (Pourhabibi, Ong, Kam, & Boo, 2020; Waring, Lindvall, & Umeton, 2020). It can be used to cluster the different categories of beneficiaries and calculate the expected claim payout and processing for these categories (Pramanik et al., 2020). There are certain Key Performance Indicators (KPI) of insurance claims such as claims cycle time, customer satisfaction, fraud detection, claims recovery and claim handling costs (Richter & Khoshgoftaar, 2018). It is observed that out of the enormous amount of data that an insurance company has access to, it utilizes only 10-15% of that data (Ringshausen et al., 2021; Saggi & Jain, 2018). ML can help to increase the utilization of data keeping in mind the KPIs of claim, to automate many routine processes to reduce claims cycle time, increase customer satisfaction, combat fraud, optimize claims recovery and reduce claim handling costs .

3. Proposed computational methodology

In this analysis, two different datasets are used to perform claim analysis using ML classification algorithms. Classification algorithms are a type of Supervised ML algorithms. Supervised Learning is used when the data is divided into a set of input variables or features and there is a corresponding output or target variable. Supervised algorithms are further divided into classification and regression algorithms. Classification algorithms are used when the target variable is categorical in nature and regression algorithms are used when the target variable is continuous. For example, if an insurance company has to predict the premium

Table 1
Research work in the Insurance Industry.

S. No.	Authors	Year	Application	Techniques used
1.	Bartl et al.	(2020)	Claim Prediction in Export Credit Finance	Decision Tree, Random Forest, Neural Networks (NN) & Probabilistic Neural Networks (PNN) for prediction and Accuracy, Cohen's K & R-square for assessment
2.	Baudry et al.	(2019)	Claim Reserving	Non-parametric ML model used for prediction, chain ladder method used for assessment w.r.t. bias and variance of estimates.
3.			Insurance Claim Analysis	Naïve Bayes, Naïve Bayes Updatable, Multi-Layer Perceptron, J48, Random Tree, LMT, Random Forest used for prediction with Recall, Precision, F-Measure, MCC, PRC & ROC Area for assessment
4.			Risk Prediction in Life Insurance	Univariate & Bivariate analysis for data visualization, PCA & Correlation-based feature selection for dimensional reduction and multiple linear regression, multilayer perceptron, REPTree & Random Tree for prediction
5.	Dehghanpour et al.	(2018)	Portfolio Insurance Strategy	Adaptive Neuro-Fuzzy Inference Systems (ANFIS) for prediction with the Markowitz portfolio optimization model for determining optimal portfolio weights
6.	Kang et al.	(2018)	Aggregate Auto-Insurance Data Analysis	Feature Selection techniques to classify the dataset into homogenous risk groups
7.	Panigrahi et al.	(2018)	Auto-Insurance Fraud Detection	Univariate, L1 based & Tree-based feature selection and Decision Tree, Naïve Bayes, KNN & Random Forest classification algorithms
8.	Patil et al.	(2018)	Survey on ML techniques used for Fraud Detection	Supervised, Unsupervised and Hybrid- Bagging, Boosting, Stacking & Ensemble Learners
9.	Quan et al.	(2018)	Predictive Analytics of Claims	Multivariate Decision Trees compared using R2, Gini, ME, MPE, MSE, MAE and MAPE
10.	Wang et al.	(2018)	Auto-Insurance Fraud Detection	Deep Learning model with Latent Dirichlet Allocation (LDA) based text analytics
11.			Role of Data Mining in the Insurance Industry	Classification, Clustering, Regression, Association, Summarization
12.	Rao et al.	(2013)	Factors influencing Claims in General Insurance, India	Regression analysis
13.	Guelman	(2012)	Insurance Lost Cost Modeling	Gradient Boosting (GB) compared to the Linear Model approach
14.	Salcedo-Sanz et al.	(2004)	Insolvency Prediction	Simulated Annealing (SA) and Walsh Analysis for feature selection using SVM as underlying classifier
15.	Viaene et al.	(2002)	Insurance Claim Fraud Detection	Logistic Regression, C4.5 Decision Tree, K-Nearest Neighbor, Bayesian Multilayer Neural Network, Naïve Bayes and SVM for classification and PCC, AUROC and ROC curves for assessment

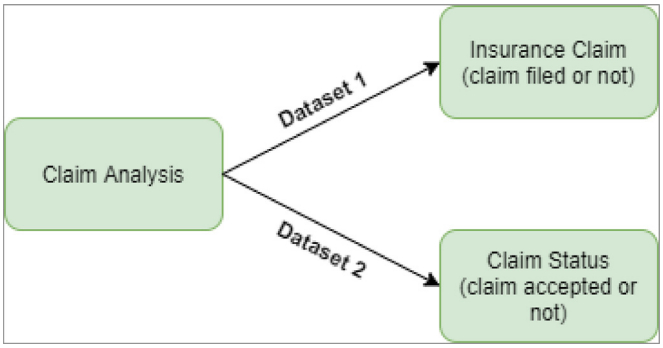

Fig. 1. Outcomes of claim analysis done in this analysis.

or claim amount, then regression algorithms can be used. In this analysis, the outcomes of the analyses using the two datasets are different, depending on the type of the datasets and the variables (features) involved.

Apart from Supervised Learning, Unsupervised Learning is also implemented in the insurance industry. In the case of unsupervised learning, there is no target variable. Unsupervised algorithms are further classified into Clustering and Association algorithms. Some use cases of unsupervised learning are to find customers having similarities in various attributes, analyze customer attrition using clustering algorithms and discovering associations that improve or promotes the business in the case of association algorithms.

As shown in Fig. 1, the target variable in both datasets is categorical. Hence, classification algorithms are used to do the analyses. Whenever ML algorithms are applied to a dataset, a framework is used to outline the process. This framework simplifies the process as well as makes it easier to understand. The framework used in this analysis is designed based on Yufeng Guo's 7 Steps of Machine Learning. Fig. 2 shows the framework used to perform claim analysis.

3.1. Data collection

Data Collection initiates the onset of the ML process. Data can be collected through various sources and methods. Searching and sharing is the most effective and common way to collect data. Data can be searched on the web or obtained from a 'data lake' and be shared through the web. Another method in use is Data Augmentation. In this method instead of collecting new data, existing data is augmented with external data to increase the diversity of the dataset. This approach is mostly used in deep learning to develop and train large artificial neural networks (ANN). The last way to collect data is through crowdsourcing and generating synthetic datasets.

The datasets used in the analysis are collected from Kaggle.com and Github.com. Further description is given in the "Case Studies" section.

3.2. Data preparation

Data Preparation is the process of transforming data such that it is suitable for an ML algorithm. It can have a great impact on the performance of the model. It consists of data cleaning, exploratory data analysis (EDA), normalization and dimensionality reduction processes.

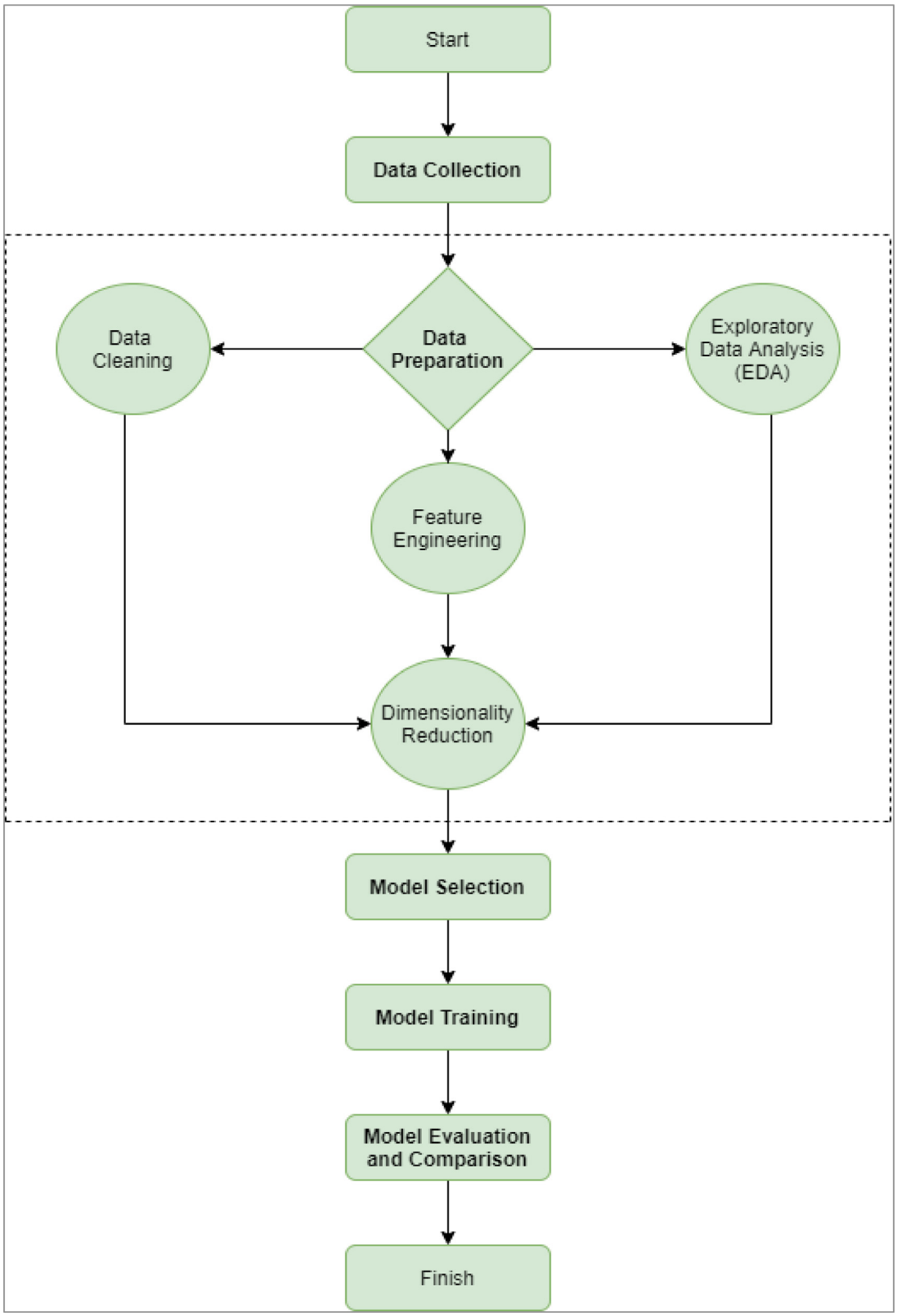

Fig. 2. ML Framework used for Claim Analysis.

3.2.1. Data cleaning

Data Cleaning is the first step after the data is retrieved. It consists of detecting and removing inaccurate, false, incomplete, corrupt, or irrelevant records from the dataset. Data cleaning is also called "Data Pre-Processing". One of the most common approaches for data cleaning is variable-by-variable cleaning. In this approach, illegal or misspelt features values are removed from the dataset based on certain factors such as the minimum and maximum value should not be outside the permissible range, the variance and standard deviation should not be more than the threshold value and there should not be any misspelt values in

the dataset. Data values are either removed or manipulated depending on the coarseness of the data value. In case there are missing feature values, the feature is either eliminated if there are many missing values or the missing values are replaced by a dummy value (treating the missing value itself as a new value). Mean substitution is the most common approach for treating missing values.

3.2.2. Exploratory data analysis (EDA)

EDA is an aid to understanding the data before applying any machine learning model to it. It is done by visualizing data with the help of different graphs to understand the different characteristics of the data like hidden relationships among various features, which is not possible by simply looking at the dataset Table 3.

3.2.3. Feature engineering

Feature engineering is one of the most fundamental and important parts of data preparation before moving further to model training and evaluation. In this stage, new features are created based on EDA performed on the dataset and existing knowledge of the domain of the dataset to improve the performance of the model. It is one of the most difficult and time-consuming parts of data preparation. Surveys show that data scientists spend almost 80% of their time on data preparation. The new features are created based on different calculations between the existing features. The new features might be a ratio, or a mathematical transformation or any statistical or scientific formula to generate a more significant feature. Feature engineering can be done both manually by statisticians and by using feature encoding techniques in the case of categorical variables. There is a general misconception that feature engineering can only be beneficial for linear regression or text classification problems. Feature engineering has proved to be greatly beneficial for support vector machines, random forests, neural networks, and gradient boosting machines. Encoding is important as machine learning itself is based on mathematical models and algorithms, so most of the algorithms cannot classify between categorical and continuous values. Encoding follows two methodologies: nominal and ordinal. Nominal Encoding is performed where the order of the data is not of much importance and vice-versa.

Apart from encoding, there are other techniques for feature engineering such as normalization. Normalization is used for scaling all the values in a dataset in a fixed range between 0 and 1. The formula used for normalization is.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Although it does improve the numerical scalability of the model, it should not be used always it can harm the performance of a model.

3.2.4. Dimensionality reduction

Dimensionality Reduction is simply reducing the dimensionality of your features. There are two approaches to the process: Feature Selection and Feature Extraction. In this analysis, only feature selection is utilized to reduce the dimensionality of the feature set as feature extraction is more suitable for data used for pattern recognition or image processing where meaningful inferences cannot be obtained just by looking at the data.

3.2.4.1. Feature selection. Having all features considered for modelling can reduce the predictability of the model. It is always preferred to select features that contribute more to the target variable. It can be done using manual methods like univariate selection where each feature is evaluated to decipher its importance. Statistical methods like variance and Pearson correlation are used for univariate analysis. But univariate analysis is more reliable when the data is linear, also it is exceedingly difficult to perform univariate analysis on a large dataset. In such a case, the multivariate analysis can be performed. There are three methods of performing multivariate analysis: filter, wrapper and embedded.

Following are the feature selection methods used in the analysis:

3.2.4.1.1. Chi-Square test. It is a type of statistical filter method that is used to evaluate the correlation between different features using their frequency distribution. In this method, feature selection is based on the intrinsic properties of the features and is independent of any ML algorithm.

3.2.4.1.2. Recursive feature elimination (RFE). It is a type of wrapper method used for feature selection. The term “wrapper” is used because this method wraps up a classifier in a feature selection algorithm. In RFE, features are recursively removed from the dataset based on an external estimator used which is the classifier. The classifier assigns weights to a feature based on its performance. It is a greedy algorithm that seeks to generate the best performing subset.

3.2.4.1.3. Tree-based feature selection. It is a type of embedded method in which there is an inbuilt method for feature importance which generates a set of features along with their importance. Embedded methods can be used with the help of algorithms that have inbuilt feature selection methods.

3.3. Model selection

After the data preparation is done and the data is divided into training and testing tests, suitable models need to be selected to do the training. Since it is a classification problem, appropriate classifiers need to be selected to conduct the classification. Both the datasets used in the analysis fall under the category of binary classification. The classification algorithms used in this analysis are Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, Gaussian Naïve Bayes, Bernoulli Naïve Byes, Mixed Naïve Bayes and K-Nearest Neighbors.

3.4. Model training

After model selection, the training data is trained using the models selected with all features initially. Then, classification algorithms are applied only to the features that are selected through feature selection techniques.

3.5. Model evaluation and comparison

Finally, the models are compared with each other along with the feature selection techniques to come up with the best model and feature selection technique for the two datasets. Four metrics are used to evaluate the models in the analysis: Precision, Recall, F1 Score and Accuracy. All four metrics are given equal importance rather than relying on one performance metric.

4. Experimentation results

The analysis consists of two case studies: one of the health insurance sector and the other one of the travel insurance sector. The output in both cases is different as shown in Fig. 1. The first one is from the perspective of the beneficiary and the second one is from the perspective of the insurance company.

4.1. Case Study 1: health insurance

For the case study, the dataset is obtained from Kaggle.com. It consists of 1338 rows and 9 columns: 8 features and 1 target variable. Table 2 describes the features and target variable of the dataset for better understanding.

Once the data is collected, the next step is data preparation. First, the data is checked for any missing values. There are no missing values in the dataset. Next, different statistics of the features are observed.

Before proceeding further for EDA, one-hot encoding is performed for the feature ‘region’ to better understand the distribution of policyholders in different regions along with other features. The feature itself is removed from the dataset and four new features ‘NorthEast USA’,

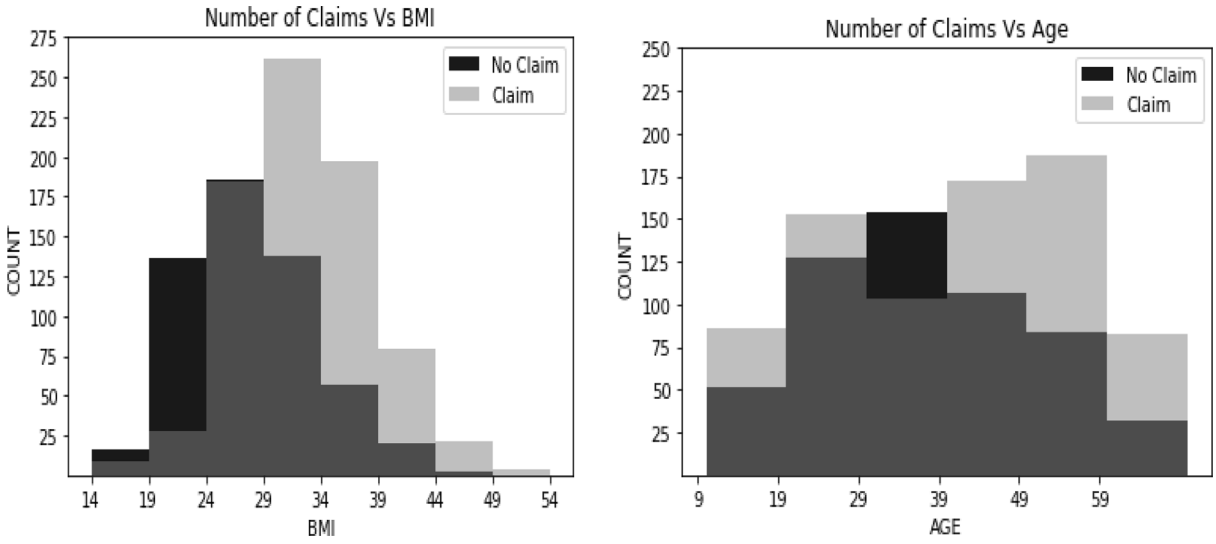

Fig. 3. Graphical Representation of relationship b/w Number of Claims and the BMI of the beneficiary & Number of Claims and the Age of the beneficiary.

Table 2
Description of Health Insurance Dataset.

S.No.	Column Heading	Description
1.	age	Age of the beneficiary
2.	sex	Gender of the beneficiary (female = 0, male = 1)
3.	bmi	Body Mass Index of the beneficiary, i.e. the ratio of weight to height (kg / m^2), ideally 18.5 to 25
4.	steps	Average walking steps per day of the beneficiary
5.	children	Number of children or dependents of the beneficiary
6.	smoker	Smoking status of the beneficiary (non-smoker = 0, smoker = 1)
7.	region	The place of residence of the beneficiary in the US (northeast = 0, northwest = 1, southeast = 2, southwest = 3)
8.	charges	Individual medical costs billed by health insurance
9.	insurance claim	Whether the beneficiary files a claim or not (Yes = 1, No = 0)

Table 3
Statistics of features of the Health Insurance Dataset.

Statistics	Age	BMI	Steps	Children	Charges
Min	18	15.96	3000	0	1121.87
Max	64	53.13	10010	5	63770.42
Mean	39.2	30.66	5328.62	1	13270.42

'NorthWest USA', 'SouthEast USA',' SouthWest USA' are added to the dataset.

By observing Fig. 3, it can be inferred that the policyholders having a BMI between 14 and 24 claim the least, policyholders having a BMI between 24–29 have a neutral response and the policyholders with a BMI of more than 29 are most likely to file a claim. Also, it can be concluded that the age bracket of 29–39 has the maximum number of policyholders that have not filed a claim.

By observing Fig. 4, it can be inferred that while most of the smokers file a claim, non-smokers are neutral towards filing a claim and the sex of the policyholder does not affect the same. Whether a male or female, if the policyholder is a smoker, he or she will most probably file a claim.

From Fig. 5, it can be inferred that policyholders with charges less than 9999 are least likely to file a claim and policyholders with charges more than 39,999 are going to file a claim. It can also be inferred that policyholders with an average of 3–6k steps day claim the most and the policyholders with more than 6k steps per day claim the least and the policyholders with no children claim the most. Apart from these findings, it is also observed that the policyholders of NorthWest USA are neutral to filing a claim while policyholders of SouthEast USA are more likely to file a claim as compared to other regions.

The final set of features selected before proceeding to feature selection are 'age', 'sex', 'BMI', 'steps', 'children', 'smoker', 'NorthEast USA', 'NorthWest USA', 'SouthEast USA', 'SouthWest USA', 'charges' and the target variable 'insurance claim'.

From Table 6 and Fig. 8, it can be inferred that Logistic Regression, Random Forest and Decision Tree Classifiers show the best performance. The Decision Tree Classifier has the best performance among these three as three out of the four performance metrics have a higher value in Decision Tree classifier as compared to the rest of the two classifiers. One thing to note is that no single performance evaluation metric is given more importance.

Now, once again the dataset is trained using the same eight classifiers but with feature selection techniques to observe the changes in the result as well as evaluate the best feature selection technique for all the classifiers.

From Table 7 and Fig. 8, it can be inferred that the Logistic Regression, Random Forest and Decision Tree Classifiers show the best performance once again. Random Forest is the best among these three classifiers as all four of its performance metrics have a higher value than the rest of the classifiers. Using Chi-Squared Test, four features are eliminated leaving the dataset with seven features: 'age', 'BMI', 'steps', 'children', 'smoker', 'NorthWest USA' and 'Southeast USA'. The performance of the SVM Classifier has decreased, there is an improvement in the performance of Gaussian and Mixed NB Classifiers & the performance of the KNN Classifier has increased tremendously.

From Table 8 and Fig. 8, it can be inferred that Logistic Regression, Random Forest, Decision Tree, Gaussian NB and Mixed NB show the best performance. Random Forest is once again the best among all the classifiers. Logistic Regression, Gaussian NB and Mixed NB show equal performance. The performance of SVM, Gaussian NB, Bernoulli NB and Mixed NB classifiers increased, the rest of the classifiers' performance has decreased as compared to the results obtained from the Chi-Squared Test. As Chi-Squared Test is a statistical test and has no involvement of any ML model, the number of features considered important by this test i.e. seven is considered as the number of comparing all the feature selection techniques. The seven most important features selected using

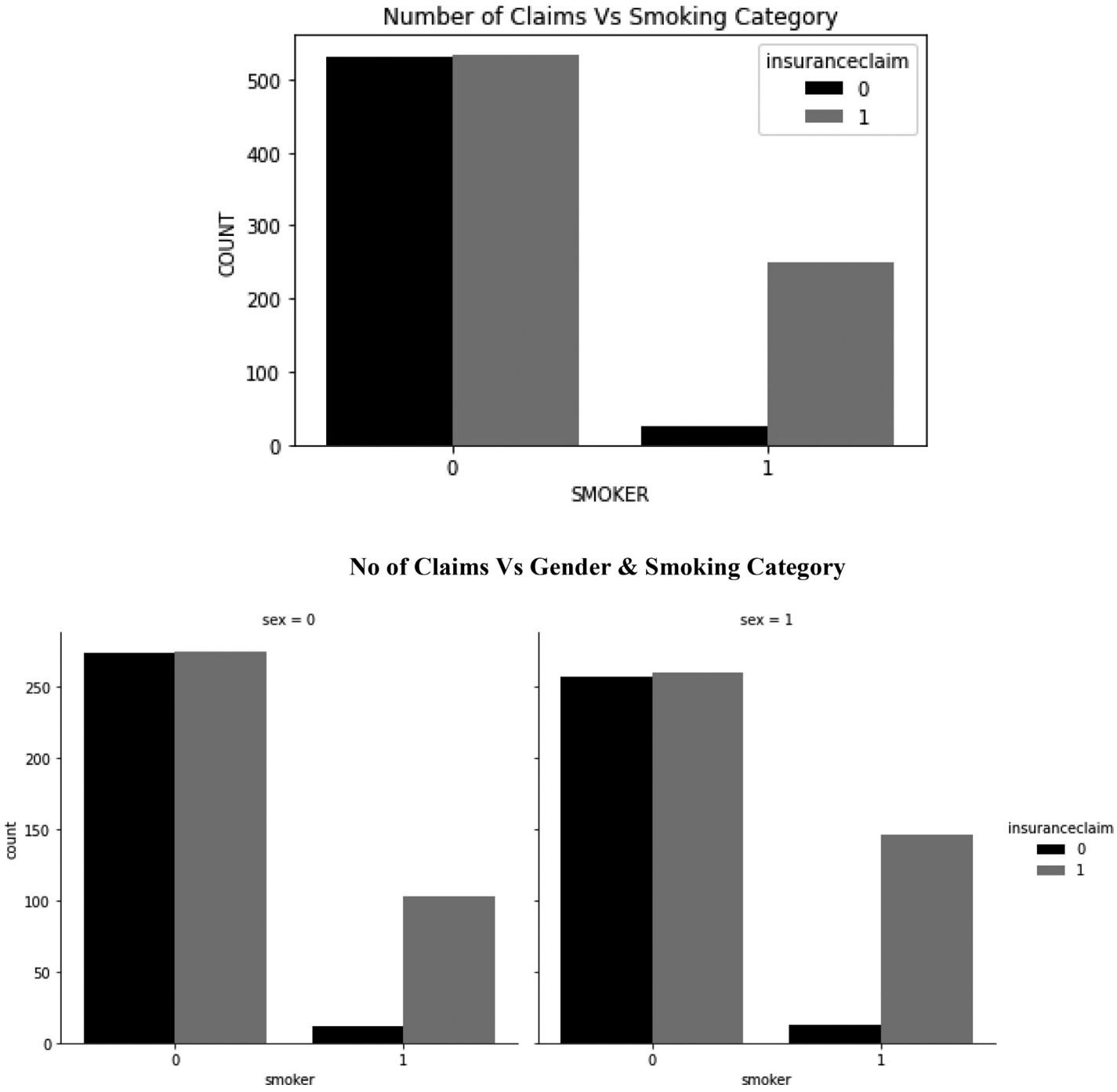

Fig. 4. Graphical Representation of the relationship between Number of Claims and the smoking category of the policyholders, Number of Claims and the smoking category when the policyholders are female & Number of Claims and the smoking category when the policyholders are male.

RFE are: 'NorthEast USA', 'age', 'BMI', 'charges', 'children', 'smoker' and 'steps'.

From Table 9 and Fig. 8, it can be inferred that Logistic Regression, Random Forest and Decision Tree classifiers show the best performance. Decision Tree is the best among all the classifiers. Except for the Decision Tree, all the classifiers' performance has decreased as compared to RFE. The seven most important features selected using the Tree-Based Feature Importance Method are: 'children', 'steps', 'BMI', 'charges', 'age', 'smoker' and 'sex'.

By observing the performance of all the eight classifiers with and without feature selection it can be concluded that Decision Tree is the best classifier without feature selection and Random Forest is the best classifier with feature selection. RFE is the best feature selection method as most of the models performed their best after RFE. KNN is the only classifier that performed best with the features selected by Chi-Squared Test. Decision Tree gives the best result with the Tree-Based method. Logistic Regression gives the best result without feature selection. Gaussian NB and Mixed NB produce the same results in all the cases; hence it can be concluded that continuous variables are more important than categorical variables for the dataset. The best set of features, therefore, is: 'NorthEast USA', 'age', 'BMI', 'charges', 'children', 'smoker' and 'steps'.

4.2. *Case study 2:* travel insurance

For the case study, the dataset is obtained from Kaggle.com. It consists of 62,288 rows and 12 columns: 11 features and 1 target variable.

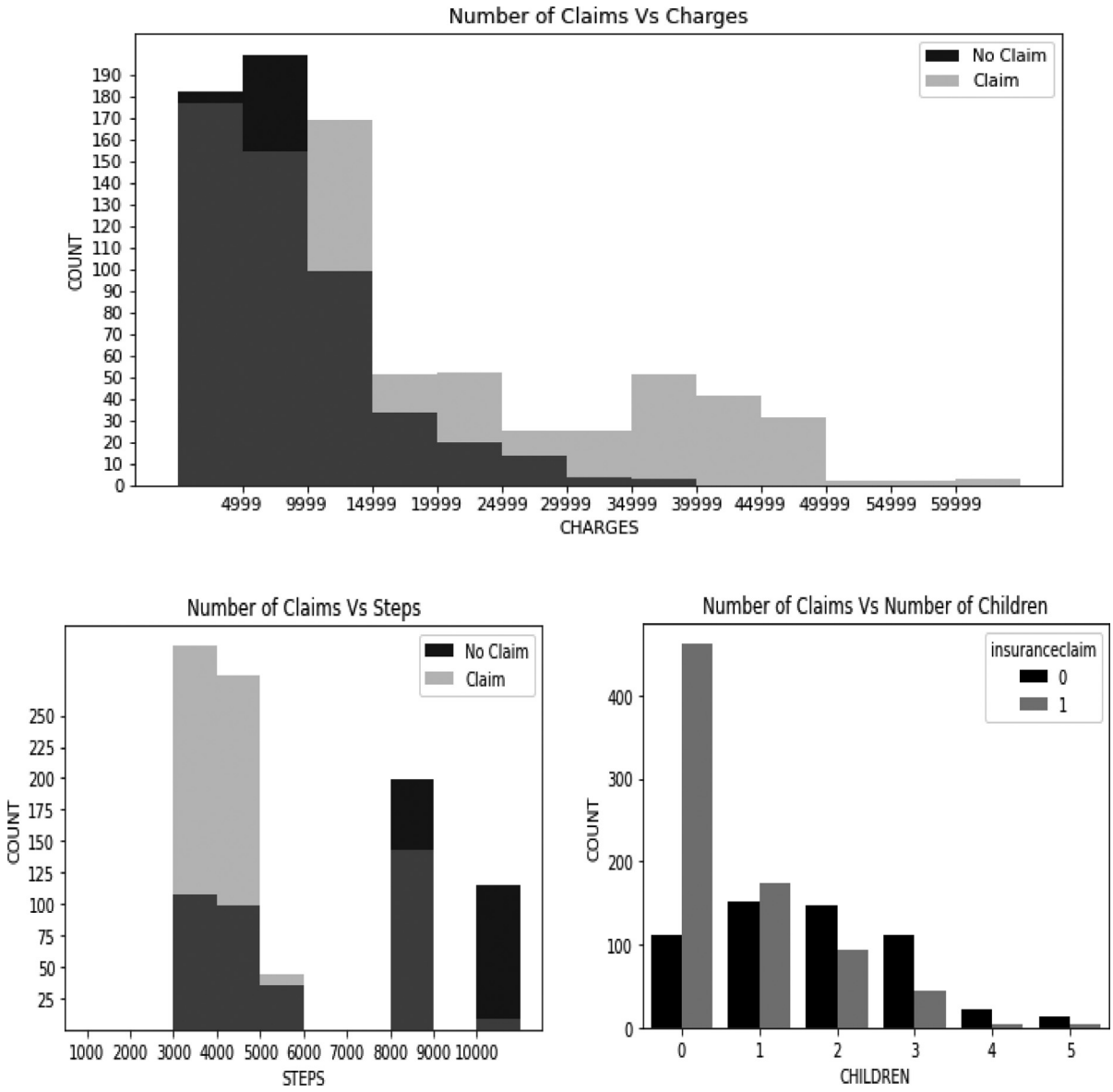

Fig 5. Graphical representation of the relationship between Number of Claims by the policyholder and the Charges billed by the health insurance, Number of Claims and the average walking steps of the policyholder per day & Number of Claims and the number of children of the policyholder.

Table 4 describes the features and target variable for a better understanding of the dataset.

Once the data is collected, the next step is data preparation. First, the data is checked for any missing values. There are 39,575 missing values for Gender i.e. 63.54 %. Hence, the column is dropped from the dataset before proceeding for further evaluation. Also, the ID column is dropped as it has no significance in predicting claim acceptance. Next, different statistics of the features are observed.

According to the statistics observed as in Table 5, the maximum age is 118 which is not a suitable age to travel for anybody, also most of the insurance companies do not provide insurance to people above 85 years of age, hence considering 100–118 as an outlier category which comprises of 1.44% of the total policyholders, it is replaced by 99 hence keeping 99 as the maximum age of a policyholder. The minimum negative value of Net Sales is justified as net sales are calculated as the difference between the value for which the insurance was sold and the expenses incurred, or the claim amount paid by the insurance company to the policyholder/beneficiary. So the net sale may be negative if the claim amount paid or even if the claim amount is not paid it can be negative when the claim is rejected and the expenses incurred for doing the investigation is more than the actual policy amount paid by the policyholder. The minimum value of duration is -2 is which not possible under any circumstances and the maximum value is 4881. Even if the unit for the duration is considered to be in days then also this value is not possible as travel insurance policies can be applied for a maximum duration of 1–2 years in the case of an Annual Plan. Considering a maximum duration of 731 days i.e. one year and one leap year, the values above 731 and below 1 are imputed as they constitute only 0.026% of the whole dataset. Values above 731 are imputed as 731 and values below 1 are imputed by the median value of duration.

By observing Fig. 6, it can be inferred that Net Sales and Acceptance % are directly proportional to each other and only 3 agencies have a good amount of sales. Also, Agency 'C2B' despite having high net sales and acceptance % provide a low commission to the mediator or maybe most of their sales are direct and there is no mediator in between.

From Fig. 7, it can be inferred that products having high Commission Value have high Net Sales as well as high Acceptance %. Apart from these findings, it is also observed that most of the Travel Agency

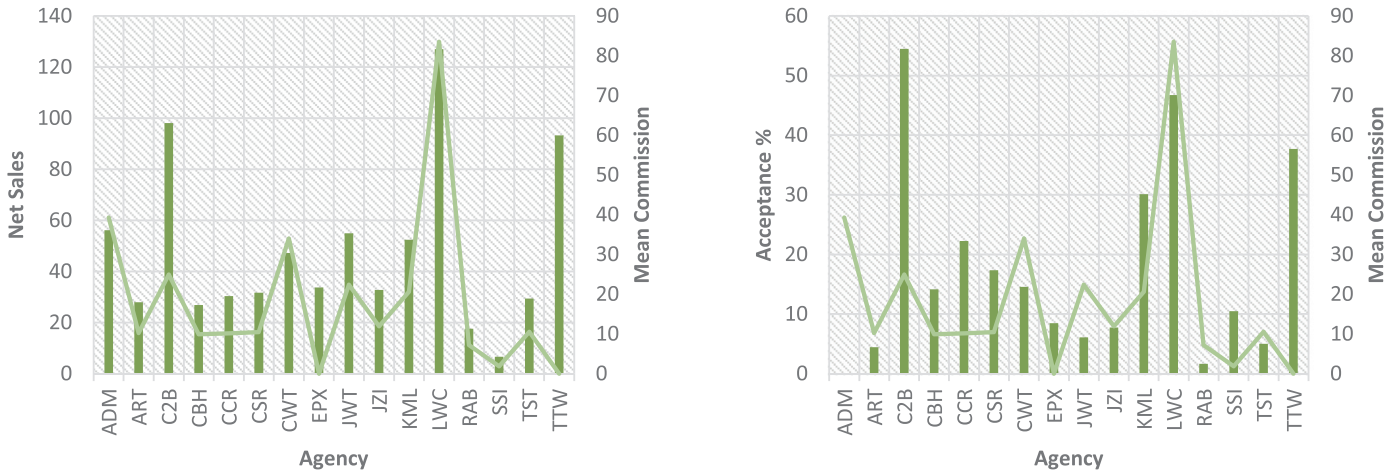

Fig. 6. Graphical representation of the relationship between Agency and Net Sales of policies & Agency and Acceptance % of claims along with the mean commission.

Table 4
Description of Travel Insurance Dataset.

S.No.	Column Heading	Description
1.	Agency	Name of the insurance agency
2.	Agency Type	Type of agency: travel or airlines
3.	Distribution Channel	Distribution channel of the insurance agency: online or offline
4.	Product Name	Name of the insurance policies (products)
5.	Duration	Duration of travel of the policyholder
6.	Destination	Destination of travel
7.	Net Sales	The total amount of sales of the insurance policies
8.	Commission (in value)	The commission received to the mediator (agent)
9.	Gender	Gender of the policyholder
10.	Age	Age of the policyholder
11.	ID	ID of the policyholder
12.	Claim	Claim status of the insurance policy: accepted or denied

Table 5
Statistics of the features of the Travel Insurance Dataset.

Statistics	Age	Commission	Net Sales	Duration
Min	0	0	-389	-2
Max	118	262.76	682	4881
Mean	39.67	12.83	50.71	60.96

claims are denied and despite having a low count of claims under Airlines Agency, around 40% of the claims are accepted. Before moving onto the next comparison, the age is divided into three groups: Child (less than 21 years old), Adult (21–50 years old) and Senior (above 50 years old).

The next step after EDA is Feature Engineering. In this stage, label, dummy, and frequency encoding are performed to deal with the categorical columns. Frequency Encoding is done for Destination, Agency, and Product Name columns. Dummy Encoding is done for Agency Type and Distribution Channel columns & Label Encoding is performed for Destination Category. Destination Category is decided based on Destination Risk. Destination Risk is calculated based on the count of claims. If the value is more than 0.3 i.e. more than 30% of the policyholders (travellers) have claimed the destination is marked as 'High Risk'. Similarly, if the value of risk is between 0.2 and 0.3, then the destination is marked as 'Moderate Risk' and if the value of risk is between 0 and 0.2, then the destination is marked as 'Low Risk'.

The final set of features selected before proceeding to feature selection are 'Age', 'Commission (in value)', 'Duration', 'Net Sales', 'Dest_freq_encoding', 'Agency_freq_encoding', Product_Name_freq_encoding', 'Destination Category (labels)', 'Agency Type_Travel Agency', 'Distribution Channel_Online' and the target variable 'Claim'.

After feature engineering, the next step is to do dimensionality reduction. In both the datasets, as previously discussed there is no need for feature extraction. In this section, first of all, both the datasets are trained and evaluated using the aforementioned eight classifiers without performing feature selection. Then feature selection is performed using three different methods, namely filter, wrapper and embedded methods & then the models are evaluated using four performance metrics. The feature selection techniques used under these methods in this analysis are the Chi-Squared Test, Recursive Feature Elimination using Logistic Regression Classifier & Tree-Based Feature Importance using ExtraTrees Classifier.

From Table 10 and Fig. 9, it can be inferred that Random Forest, Decision Tree and KNN show the best performance. Random Forest is best among all the classifiers as all four of its performance metrics have a higher value than the rest of the classifiers.

Now, once again the dataset is trained using the same eight classifiers but with feature selection techniques to observe the changes in the

Table 6
Classification of Health Insurance dataset without Feature Selection.

Metrics Used	Classification Algorithms Used							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.8137255	0.96078	0.95098	0.52941	0.80392	0.73529	0.80392	0.5490196
Recall	0.8556701	0.95146	0.9898	0.7013	0.73874	0.68182	0.73874	0.6021505
F1 Score	0.8341709	0.9561	0.97	0.60335	0.76995	0.70755	0.76995	0.574359
Accuracy	0.8768657	0.96642	0.97761	0.73507	0.81716	0.76866	0.81716	0.6902985

Table 7
Classification of Health Insurance dataset with Chi-Squared Test.

Metrics Used	Classification Algorithms Used							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.773913	0.99107	0.97321	0.5625	0.875	0.76786	0.875	0.723214
Recall	0.872549	0.97368	0.9646	0.62376	0.75385	0.66667	0.75385	0.771429
F1 Score	0.820276	0.9823	0.96889	0.59155	0.80992	0.71369	0.80992	0.746544
Accuracy	0.854478	0.98507	0.97388	0.67537	0.82836	0.74254	0.82836	0.794776

Table 8
Classification of Health Insurance dataset with Recursive Feature Elimination.

Metrics Used	**Classification Algorithms Used**							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.7714286	0.99048	0.96639	0.54622	0.80672	0.79832	0.80672	0.5714286
Recall	0.8350515	0.97196	0.95833	0.80247	0.83478	0.73077	0.83478	0.7234043
F1 Score	0.8019802	0.98113	0.96234	0.65	0.82051	0.76305	0.82051	0.6384977
Accuracy	0.8507463	0.98507	0.96642	0.73881	0.84328	0.77985	0.84328	0.7126866

Table 9
Classification of Health Insurance dataset with Tree Based Feature Importance.

Metrics Used	**Classification Algorithms Used**							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.754237	0.9322	0.97458	0.55932	0.80508	0.69492	0.80508	0.542373
Recall	0.89899	0.99099	1	0.74157	0.79832	0.76636	0.79832	0.752941
F1 Score	0.820276	0.9607	0.98712	0.63768	0.80169	0.72889	0.80169	0.630542
Accuracy	0.854478	0.96642	0.98881	0.72015	0.82463	0.77239	0.82463	0.720149

Table 10
Classification of Travel Insurance dataset without Feature Selection.

Metrics Used	**Classification Algorithms Used**							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.91919	0.98727	0.97604	0.97323	0.83818	0.86154	0.83818	0.939743
Recall	0.799163	1	1	0.83515	0.90457	0.8981	0.90457	0.999147
F1 Score	0.854985	0.99359	0.98787	0.89892	0.87011	0.87944	0.87011	0.968535
Accuracy	0.750361	0.98981	0.98082	0.82477	0.79965	0.81088	0.79965	0.951116

Table 11
Classification of Travel Insurance dataset with Chi-Squared Test.

Metrics Used	**Classification Algorithms Used**							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.965773	0.98706	0.97662	0.97461	0.83895	0.86033	0.83895	0.942103
Recall	0.837788	1	1	0.83266	0.90204	0.89452	0.90204	1
F1 Score	0.89724	0.99349	0.98817	0.89806	0.86935	0.87709	0.86935	0.970189
Accuracy	0.823086	0.98965	0.9813	0.82301	0.79828	0.80711	0.79828	0.953684

result as well as evaluate the best feature selection technique for all the classifiers.

From Table 11 and Fig. 9, it can be inferred that Random Forest, Decision Tree and KNN classifiers show the best performance. Random Forest is once again the best among all the classifiers. Using Chi-Squared Test, only one feature is discarded leaving the dataset with nine features: 'Age', 'Commission (in value)', 'Duration', 'Net Sales', 'Dest_freq_encoding', 'Agency_freq_encoding', 'Product_Name_freq_encoding', 'Destination Category (labels)' and 'Agency Type_Travel Agency'. The performance of Logistic Regression, Decision Tree and KNN classifiers has increased and the performance of the remaining classifiers has decreased as compared to modelling without feature selection.

From Table 12 and Fig. 9, it can be inferred that Random Forest, Decision Tree and KNN classifiers show the best performance. Random Forest is the best among all the classifiers used for the dataset. Except for Decision Tree and KNN, the rest all the classifiers have increased performance as compared to the results of the Chi-Squared Test. Also, the performance of the Bernoulli NB classifier has decreased as compared to its performance without any feature selection technique. As Chi-Squared Test is a statistical test and has no involvement of any ML model, the number of features considered important by this test

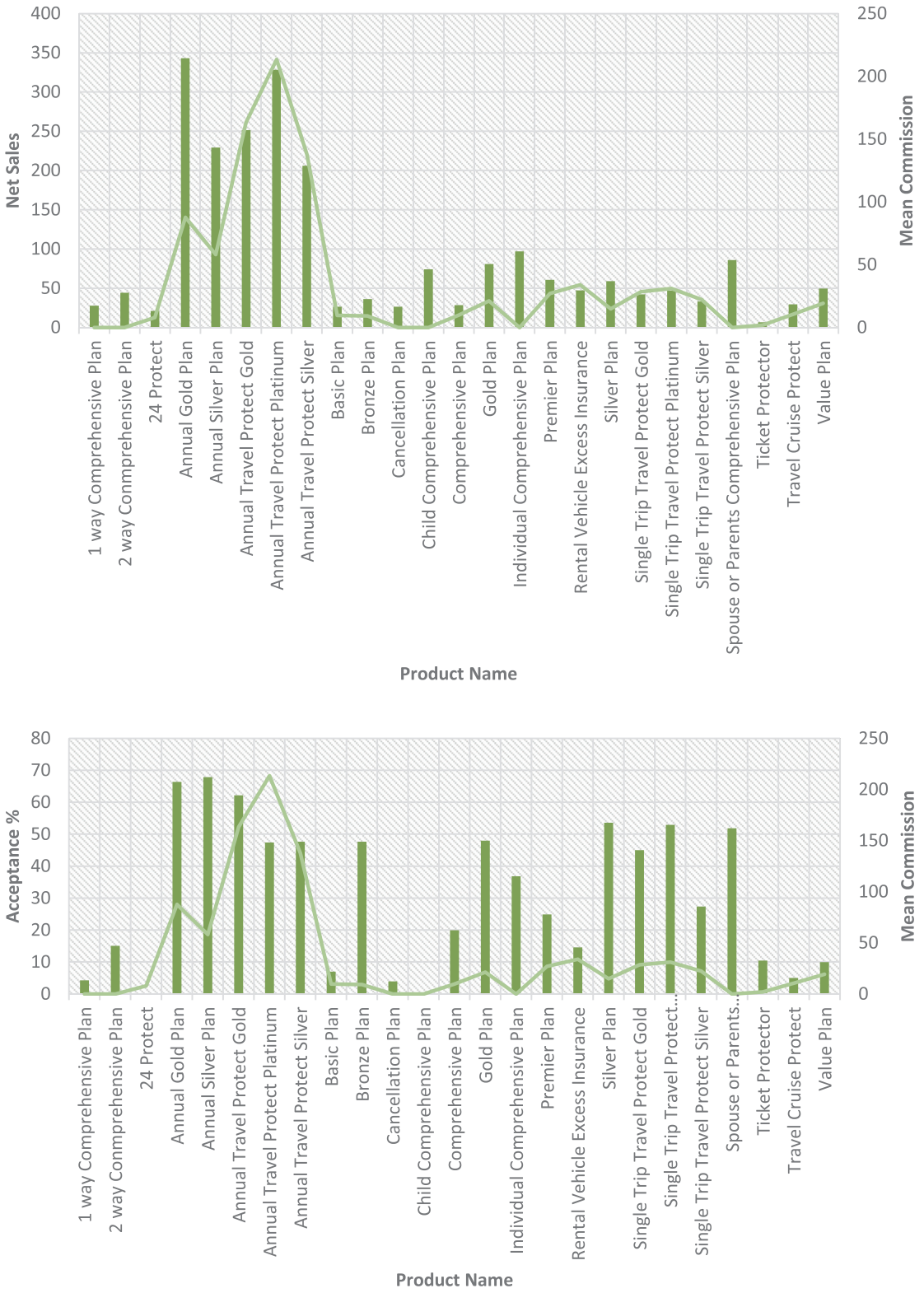

Fig. 7. Graphical Representation of the relationship between Product Name and Net Sales along with Mean Commission & between Product Name and Acceptance % along with mean commission.

i.e. nine is considered as the number of comparing all the feature selection techniques. The nine most important features identified by RFE are: 'Age', 'Agency Type_Travel Agency', 'Commission (in value)', 'Dest_freq_encoding', 'Destination Category (labels)', 'Distribution Channel_Online', 'Duration', 'Net Sales' and 'Product_Name_freq_encoding'.

From Table 13 and Fig. 9, it can be inferred that Random Forest, Decision Tree and KNN classifiers show the best performance. Random is once again the best among all the classifiers. The performance of Logistic Regression, Random Forest, Decision Tree and Bernoulli NB classifiers has increased, rest four classifiers' performance has decreased compared

Fig. 8. Graphical Representation of the performance of the aforementioned classifiers without feature selection, with Chi-Squared Test, with RFE (using LogisticRegression Classifier) and with Tree-Based Feature Importance (using ExtraTreesClassifier).

Table 12
Classification of Travel Insurance dataset with Recursive Feature Elimination.

Metrics Used	Classification Algorithms Used							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.966322	0.98751	0.97601	0.97401	0.85927	0.85767	0.85927	0.938331
Recall	0.839476	1	1	0.83533	0.88913	0.89825	0.88913	0.999574
F1 Score	0.898444	0.99371	0.98786	0.89936	0.87395	0.87749	0.87395	0.967985
Accuracy	0.825574	0.98997	0.98074	0.82493	0.80093	0.80767	0.80093	0.950153

Table 13
Classification of Travel Insurance dataset with Tree Based Feature Importance.

Metrics Used	Classification Algorithms Used							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.965455	0.98794	0.97659	0.97408	0.83886	0.86277	0.83886	0.936095
Recall	0.844541	1	1	0.83306	0.90348	0.89626	0.90348	0.999569
F1 Score	0.90096	0.99394	0.98816	0.89807	0.86997	0.8792	0.86997	0.966792
Accuracy	0.829347	0.99037	0.9813	0.82333	0.79965	0.81056	0.79965	0.948788

to RFE. The nine best features selected by the Tree-Based Feature Importance method are: 'Duration', 'Age', 'Net Sales', 'Dest_freq_encoding', 'Commission (in value)', 'Agency_freq_encoding', 'Destination Category (labels)', 'Product_Name_freq_encoding' and 'Agency Type_Travel Agency'. These are the same features as selected by Chi-Squared Test. The performance of the models using both feature selection methods is also almost the same.

By observing the performance of all the eight classifiers with and without feature selection it can be concluded that Random Forest is the best classifier both with and without feature selection. Chi-Squared and Tree-Based Feature Importance methods are the best feature selection techniques for the dataset as four models perform their best with features selected from these two. The slight difference between the results of these two techniques is neglected as both have selected the same set of features and any ML model trains itself continuously & both the techniques are not applied parallelly. Gaussian NB and Mixed NB produce the same results in all the cases; hence it can be concluded that continuous variables are more important than categorical

Fig. 9. Graphical Representation of the performance of the aforementioned classifiers without Feature Selection, with Chi-Squared Test, with RFE (using LogisticRegression Classifier), with Tree-Based Feature Importance (using ExtraTreesClassifier).

variables for the dataset. Bernoulli NB classifier performs best without any feature selection. Therefore, the best set of features for the dataset is: 'Duration', 'Age', 'Net Sales', 'Dest_freq_encoding', 'Commission (in value)', 'Agency_freq_encoding', 'Destination Category (labels)', 'Product_Name_freq_encoding' and 'Agency Type_Travel Agency'.

5. Discussion

With the help of the experimentation conducted, various characteristics of the client strata have been deduced in correlation with the insurance claim and claim status. Further, machine learning models have been used to train the model to predict the claim status/insurance claim with accuracy (Testoni & Boeri, 2015) . Feature selection techniques are used to reduce the dimensionality of the datasets as well as increase the accuracy of the trained models Guo & Yu (2016). The results have been further discussed in the following section. Claim Analysis of both the datasets is performed successfully with the help of eight classification algorithms. For both datasets, Random Forest is the best classifier with suitable feature selection methods. For the first and second dataset, Logistic Regression and Bernoulli NB classifiers have performed better with all the features. Also, the KNN classifier has performed better with Chi-Squared Test in the case of both the datasets despite the Chi-Squared Test being a filter method that selects features solely based on their correlation with each other based on their frequency distribution. Although wrapper methods like RFE improve the model's performance by a better margin than the filter and embedded methods, it is not always the same, as observed in the analysis. The performance of the feature selection method also depends on the dataset, the models used for training and evaluation of the dataset & the models used for feature selection. In the analysis, LogisticRegression and ExtraTrees classifiers are used with RFE and Tree-Based Feature Importance embedded method, respectively. The choice of model used to fit the dataset for RFE is not much of a concern, it will not bring a huge lot of difference in the performance of the model. ExtraTrees classifier is used for the Tree-Based Feature Importance method as it is an extremely randomized classifier and is computationally less expensive than other tree-based algorithms. Unlike the results of the healthcare insurance dataset, all the eight classifiers used in the travel insurance dataset perform extremely well as wells as the Chi-Squared Test identifies nine out of the ten features selected for model training as important. This is because the features of the second dataset are highly engineered already to increase the model's performance. It is also observed that the data is high imbalanced based on the target variable in the second dataset, 80% of the claims are not accepted. This may result in biased predictions even though the model is properly trained.

5.1. Theoretical contributions and implications

It can be deduced that introducing technologies such as Machine Learning into the field of insurance can be very helpful. InsurTech as a whole can help identify and understand the customer in a much better way than the narrow definition defined by the insurance industry regarding their needs and investing patterns (Doupe et al., 2019; McGlade & Scott-Hayward, 2019). Using claim analysis, customer claiming patterns and their demography can be understood which can further help to improve policies and decide more viable premiums for the customers (Doupe et al., 2019; Karhade et al., 2019). Also, by understanding the insurance company's acceptance patterns, the policies can be modified to monitor their profit/loss ratio.

5.2. Implications for practice

It is observed that using feature selection techniques is indeed very beneficial before classifying data using classification algorithms (Dave et al., 2021). Not all attributes are equally significant and using feature selection techniques help to choose the best subset of attributes for optimal results (Nian et al., 2016). It reduces the overfitting of data, increases the accuracy of the algorithm used and also reduces the computation time (Larson & Sinclair, 2021). For example, in the Travel Insurance case study, using logistic regression with all attributes results in a model which has an accuracy of 0.750361, while using logistic regression along with tree-based feature selection results in an accuracy of 0.829347.

6. Conclusions

Despite the innumerable advantages of InsurTech only 28% of the big companies go for partnership with InsurTech companies and even less than 14% actively participate in incubator programs or ventures (water house Coopers 2016). Due to sparse partnerships, it is difficult for new InsurTech companies to survive for long. This brings into question the sustainability of this model for scaling. Also, most of the funds of startups go into the distribution of policies which makes it difficult for them to invest in other units of insurance i.e. underwriting, claims to service and maintain regulatory compliance.

The future scope of this research work can be to address this problem of high data imbalance using resampling of the dataset, clustering the abundant class or by simply applying Synthetic Minority Oversampling Technique (SMOTE), XGBoost or Adaptive Boosting (AdaBoost) algorithms .

Author contributions

Dr Seema Rawat conceived and designed the study, and Mr Aakankshu Rawat performed the research, and Dr A. Sai Sabitha with Dr Deepak Kumar analyzed the data and contributed to editorial input.

Declaration of Competing Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

Aswani, R., Ghrera, S. P., Chandra, S., & Kar, A. K. (2020). A hybrid evolutionary approach for identifying spam websites for search engine marketing. *Evolutionary Intelligence* (0123456789). 10.1007/s12065-020-00461-1.

Bacry, E., Gaïffas, S., Leroy, F., Morel, M., Nguyen, D. P., Sebiat, Y., & Sun, D. (2020). SCALPEL3: A scalable open-source library for healthcare claims databases. *International Journal of Medical Informatics, 141*(May). 10.1016/j.ijmedinf.2020.104203.

Barry, L., & Charpentier, A. (2020). Personalization as a promise: Can big data change the practice of insurance? *Big Data and Society, 7*(1). 10.1177/2053951720935143.

Batra, J., Jain, R., Tikkiwal, V. A., & Chakraborty, A. (2021). A comprehensive study of spam detection in e-mails using bio-inspired optimization techniques. *International Journal of Information Management Data Insights, 1*(1), Article 100006. 10.1016/j.jjimei.2020.100006.

Blackstone, E. H. (2013). Generating new knowledge in cardiac interventions. *Anesthesiology Clinics, 31*(2), 217–248. 10.1016/j.anclin.2012.12.006.

Chakraborty, A., & Kar, A. K. (2017). Swarm intelligence: A review of algorithms. *Modeling and Optimization in Science and Technologies*, 10, 475–494. https://doi.org/10.1007/978-3-319-50920-4_19

Chowdhury, S., Mayilvahanan, P., & Govindaraj, R. (2020). Optimal feature extraction and classification-oriented medical insurance prediction model: machine learning integrated with the internet of things. *International Journal of Computers and Applications, 0*(0), 1–13. 10.1080/1206212X.2020.1733307.

Das, D., Chakraborty, C., & Banerjee, S. (2020). A Framework development on big data analytics for terahertz healthcare. terahertz biomedical and healthcare technologies. Elsevier Inc. 10.1016/b978-0-12-818556-8.00007-0.

Das, S., Datta, S., Zubaidi, H. A., & Obaid, I. A. (2021). Applying interpretable machine learning to classify tree and utility pole related crash injury types. *IATSS Research*. 10.1016/j.iatssr.2021.01.001.

Dave, H. S., Patwa, J. R., & Pandit, N. B. (2021). Facilitators and barriers to participation of the private sector health facilities in health insurance & government-led schemes in India. *Clinical Epidemiology and Global Health, 10*(January), Article 100699. 10.1016/j.cegh.2021.100699.

Doupe, P., Faghmous, J., & Basu, S. (2019). Machine learning for health services researchers. *Value in Health, 22*(7), 808–815. 10.1016/j.jval.2019.02.012.

Guo, Y., & Yu, S. (2016). A new histogram based shape descriptor in image retrieval. *International Journal of Signal Processing, Image Processing and Pattern Recognition, 9*(4), 233–246. 10.14257/ijsip.2016.9.4.22.

Gupta, S., Kar, A. K., Baabdullah, A., & Al-Khowaiter, W. A. A (2018). Big data with cognitive computing: A review for the future. *International Journal of Information Management, 42*, 78–89 (April). 10.1016/j.ijinfomgt.2018.06.005.

Kakhki, F. D., Freeman, S. A., & Mosher, G. A. (2020). Applied machine learning in agromanufacturing occupational incidents. *Procedia Manufacturing, 48*, 24–30 (2019). 10.1016/j.promfg.2020.05.016.

Kar, A. K. (2016). Bio inspired computing - a review of algorithms and scope of applications. *Expert Systems with Applications, 59*, 20–32. 10.1016/j.eswa.2016.04.018.

Karhade, A. V., Ogink, P. T., Thio, Q. C. B. S., Broekman, M. L. D., Cha, T. D., Hershman, S. H., …, & Schwab, J. H. (2019). Machine learning for prediction of sustained opioid prescription after anterior cervical discectomy and fusion. *Spine Journal, 19*(6), 976–983. 10.1016/j.spinee.2019.01.009.

Kasy, M. (2018). Optimal taxation and insurance using machine learning — Sufficient statistics and beyond. *Journal of Public Economics, 167*, 205–219. 10.1016/j.jpubeco.2018.09.002.

Kaur, P., Sharma, M., & Mittal, M. (2018). Big data and machine learning based secure healthcare framework. *Procedia Computer Science, 132*, 1049–1059. 10.1016/j.procs.2018.05.020.

Khan, F. H., Bashir, S., & Qamar, U. (2014). Author ' s personal copy TOM : Twitter opinion mining framework using hybrid classi fi cation scheme. *Decision Supp. Syst., 57*(January), 245–257.

Knighton, J., Buchanan, B., Guzman, C., Elliott, R., White, E., & Rahm, B. (2020). Predicting flood insurance claims with hydrologic and socioeconomic demographics via machine learning: Exploring the roles of topography, minority populations, and political dissimilarity. *Journal of Environmental Management, 272*, Article 111051. 10.1016/j.jenvman.2020.111051.

Kose, I., Gokturk, M., & Kilic, K. (2015). An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Applied Soft Computing Journal, 36*, 283–299. 10.1016/j.asoc.2015.07.018.

Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research, 281*(3), 628–641. 10.1016/j.ejor.2019.09.018.

Larson, W. D., & Sinclair, T. M. (2021). Nowcasting unemployment insurance claims in the time of COVID-19. *International Journal of Forecasting* xxxx. 10.1016/j.ijforecast.2021.01.001.

Maehashi, K., & Shintani, M. (2020). Macroeconomic forecasting using factor models and machine learning: an application to Japan. *Journal of the Japanese and International Economies, 58*(March), Article 101104. 10.1016/j.jjie.2020.101104.

McGlade, D., & Scott-Hayward, S. (2019). ML-based cyber incident detection for electronic medical record (EMR) systems. *Smart Health, 12*, 3–23. 10.1016/j.smhl.2018.05.001.

Mita, Y., Inose, R., Goto, R., Kusama, Y., Koizumi, R., Yamasaki, D., …, & Muraki, Y. (2021). An alternative index for evaluating AMU and anti-methicillin-resistant Staphylococcus aureus agent use: A study based on the national database of health insurance claims and specific health checkups data of Japan. *Journal of Infection and Chemotherapy*, (xxxx). 10.1016/j.jiac.2021.02.009.

Nian, K., Zhang, H., Tayal, A., Coleman, T., & Li, Y. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. *The Journal of Finance and Data Science, 2*(1), 58–75. 10.1016/j.jfds.2016.03.001.

Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications: A survey. *Applied Soft Computing Journal, 93*, Article 106384. 10.1016/j.asoc.2020.106384.

Pal, D., Mandana, K. M., Pal, S., Sarkar, D., & Chakraborty, C. (2012). Fuzzy expert system approach for coronary artery disease screening using clinical parameters. *Knowledge-Based Systems, 36*, 162–174. 10.1016/j.knosys.2012.06.013.

Palanisamy, V., & Thirunavukarasu, R. (2019). Implications of big data analytics in developing healthcare frameworks – a review. *Journal of King Saud University - Computer and Information Sciences, 31*(4), 415–425. 10.1016/j.jksuci.2017.12.007.

Pappas, I. O., & Woodside, A. G. (2021). Fuzzy-set qualitative comparative analysis (fsQCA): Guidelines for research practice in Information Systems and marketing. *International Journal of Information Management, 58*, Article 102310 (September 2020). 10.1016/j.ijinfomgt.2021.102310.

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Vlachogiannakis, N. E. (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting, 36*(3), 1092–1113. 10.1016/j.ijforecast.2019.11.005.

Pourhabibi, T., Ong, K. L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems, 133*(April), Article 113303. 10.1016/j.dss.2020.113303.

Pramanik, M. I., Lau, R. Y. K., Azad, M. A. K., Hossain, M. S., Chowdhury, M. K. H., & Karmaker, B. K (2020). Healthcare informatics and analytics in big data. *Expert Systems with Applications, 152*, Article 113388. 10.1016/j.eswa.2020.113388.

Richter, A. N., & Khoshgoftaar, T. M. (2018). A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artificial Intelligence in Medicine, 90*, 1–14 (September 2017). 10.1016/j.artmed.2018.06.002.

Ringshausen, F. C., Ewen, R., Multmeier, J., Monga, B., Obradovic, M., van der Laan, R., & Diel, R. (2021). Predictive modeling of nontuberculous mycobacterial pulmonary disease epidemiology using German health claims data. *International Journal of Infectious Diseases, 104*, 398–406. 10.1016/j.ijid.2021.01.003.

Saggi, M. K., & Jain, S. (2018). A survey towards an integration of big data analytics to big insights for value-creation. *Information Processing and Management, 54*(5), 758–790. 10.1016/j.ipm.2018.01.010.

Sengupta, S., Basak, S., Saikia, P., Paul, S., Tsalavoutis, V., Atiah, F., …, & Peters, A. (2020). A review of deep learning with special emphasis on architectures, applications and recent trends. *Knowledge-Based Systems, 194*, Article 105596. 10.1016/j.knosys.2020.105596.

Testoni, C., & Boeri, A. (2015). Smart governance: urban regeneration and integration policies in Europe. Turin and Malmö case studies. *City and Community, 26*(27), 28.

Waring, J., Lindvall, C., & Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine, 104*, Article 101822 (October 2019). 10.1016/j.artmed.2020.101822.

Yang, C., Yang, Z., Wang, J., Wang, H.-Y., Su, Z., Chen, R., & Zhou, Z. (2021). Estimation of prevalence of kidney disease treated with dialysis in China: A study of insurance claims data. *American Journal of Kidney Diseases*. 10.1053/j.ajkd.2020.11.021.