Cureus Journal of Computer Science

Open Access Original Article

Predictive Precision: Unraveling Health Insurance Claim Patterns With Logistic Regression and Decision Trees

Received 01/23/2025
Review began 02/12/2025
Review ended 02/27/2025
Published 03/05/2025

DOI:
https://doi.org/10.7759/s44389-025-03010-y

Vaishnavi P. Thakre [1] , Rohit D. Poul [1] , Ankush D. Sawarkar [1]

1. Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded, IND

Corresponding author: Vaishnavi P. Thakre, vpthakre17@gmail.com

Abstract

Healthcare insurance is a type of insurance policy that covers the risk of medical expenses during any unfortunate crisis. Understanding the pattern in the insurance field is essential for improving policy design, managing risks, and detecting fraud. This study provides a comprehensive comparison of logistic regression with other machine learning models, such as decision tree and support vector machine, to evaluate their effectiveness in predicting health insurance claims. Its performance, evaluated using metrics like precision, recall, and F1 score, shows balanced results, though less robust than the decision tree model. Logistic regression demonstrated balanced performance with an accuracy of 82% and superior interpretability, making it valuable for actionable insights. While the decision tree model achieved the highest accuracy (96%), its complexity, particularly in terms of interpretability and potential overfitting, may limit its real-world applicability. By exploring the trade-offs between accuracy and interpretability, this study provides insurers with practical guidance for optimizing operations, enhancing customer service, and mitigating financial losses.

Categories: Data Analysis, Data Engineering, Machine Learning (ML)
Keywords: insurance claim, logistic regression, health insurance, classification, decision tree, svm

Introduction

Insurance involves a contract between an individual or entity (the insured) and an insurance company (the insurer). The insurer agrees to offer financial protection or reimbursement for specified losses or damages outlined in the insurance policy under certain circumstances [1]. Insurance is a legal agreement between an individual or entity, referred to as the insured, and an insurance company, known as the insurer. The primary objective of insurance is to reduce the financial impact of unforeseen circumstances, including accidents, illnesses, natural disasters, or death [2]. There are different kinds of insurance, such as health, life, auto, and home insurance. Each type of insurance is tailored to provide protection against specific risks [3].

Health insurance plays a crucial role in healthcare by offering financial security and enhancing the availability of medical care [4]. Health coverage is a vital part of healthcare, offering financial security and increasing the availability of medical services [2]. These insurance plans help ease the financial strain of medical costs by paying for all or a portion of healthcare bills. This makes crucial treatments, procedures, and medications more affordable for individuals and families. Additionally, health insurance provides access to preventive care services like immunizations, screenings, and routine check-ups, which are crucial for detecting and managing health issues early on [5]. Health insurance utilizes risk pooling to distribute the financial risk among a larger group of people, which ultimately helps to reduce healthcare costs for everyone [6]. Adhering to regulatory requirements frequently requires having health insurance, which helps individuals support the healthcare system as a whole. This also helps decrease the burden on healthcare providers and taxpayers caused by uncompensated care costs [4]. Affordable and inclusive health insurance coverage is crucial for individuals who want to safeguard themselves from the potentially steep expenses of medical care [5]. Health insurance policies generally require individuals to meet specific eligibility criteria to qualify for coverage. These criteria can vary depending on the type of insurance plan, the insurance company, and local regulations [3]. The qualifications for health insurance coverage typically differ based on the type of insurance plan, the insurance provider, and local laws [2]. Insurance eligibility criteria may be governed by regulations and laws at the national, state, or local level [4].

Logistic regression is a powerful statistical technique used for binary classification. It models the relationship between input variables (features) and the probability of an event occurring. The output variable in logistic regression is binary, making it suitable for scenarios like insurance claims [7]. Logistic regression can be a useful statistical technique for identifying whether a person is likely to be eligible to claim insurance or not. In the context of insurance eligibility, the dependent variable could represent whether an individual successfully claims insurance (1) or not (0) [7].

How to cite this article
Thakre V P, Poul R D, Sawarkar A D (March 05, 2025) Predictive Precision: Unraveling Health Insurance Claim Patterns With Logistic Regression and Decision Trees. Cureus J Comput Sci 2 : es44389-025-03010-y. DOI https://doi.org/10.7759/s44389-025-03010-y

The objective of this research paper is to investigate the utility of logistic regression in predicting insurance claims and assessing eligibility for health insurance coverage. Developing a predictive model helps insurers optimize their operations, improve customer service, and mitigate financial losses by identifying high-risk policies or customers early on. The performance of logistic regression models in predicting insurance claims based on available predictors will also be evaluated. The study will investigate the impact of different variables on the likelihood of making an insurance claim and assess the significance of each predictor.

Authors	Methodology	Key Findings	Ref.
Sun et al. (2019)	Demonstrated methods to identify the fraudsters behaviors or camouflage behaviors by computing the similarity between the patient-level hospital admissions and constructing the similarity graph and then clustering them to extract the semantic meaning of each cluster through a graph-based density peak clustering algorithm.	They noted that some fraudsters pretend to be patients to get the insurance money. In recent literature, such incidents are termed as camouflage. Additionally, the authors noted that the healthcare insurance data are heterogenous and longitudinal in nature. The fraudsters can take advantage of this and can hide their activities in the vast data. Moreover, the fraudsters' behavior keeps evolving, making it difficult to predict and find patterns.	[8]
DeVoe et al. (2009)	The studies highlighted the importance of understanding how family coverage patterns affect children's access to healthcare services. With significant shifts in the patterns of coverage in the past decade, it is difficult to identify pattern between the family coverage and the children’s insurance care.	This study revealed that children of age group 2-27 years living with at least one parent were 73.6% insured, whereas children living with both their parents were 8.0% uninsured. The remaining 18.4% experienced discordant patterns of coverage, where children and parents had different insurance statuses. In comparison, during the same period, approximately 82.9% of the overall U.S. population was insured, while 17.1% remained uninsured.	[9]
Smith et al. (2000)	The study illustrated how data mining and machine learning models are capable of analyzing complex patterns in customer retention and insurance claims, where traditional methods fall short.	Handling and processing large volumes of insurance claim data necessitate the use of advanced computational tools. Machine learning techniques have emerged as critical in processing such data and extracting essential insights for decision-making. The authors illustrated how data mining and machine learning models are capable of analyzing complex patterns in customer retention and insurance claims, where traditional methods fall short.	[10]
Konrad et al. (2019)	The paper highlights the key challenges researchers face when using health insurance claims data for advanced data analyses, such as coding irregularities, temporal changes in coding practices, and missing information. The paper provides practical solutions and recommendations to address these challenges, such as grouping similar codes, analyzing temporal and provider effects, and working closely with data partners to understand the intricacies of the data. Authors emphasizes the importance of standardizing the fusion of aggregated claims data to improve the quality and replicability of research findings.	Authors made the use of data mining approaches to analyze the insurance claims data. They used a data-driven clustering approach to identify homogeneous subgroups of services for a given condition. They used data analytics tools to extract insights about comorbidities, treatment quality, and disease progression within the identified clusters.	[11]
Seo et al. (2012)	Authors developed an algorithm to estimate cancer incidence rates using national health insurance claims data from 2005 to 2008. They identified incident cancer cases in 2008 as those admitted to the hospital that year without a prior admission for the same cancer diagnosis from 2005 to 2007. Compared the incidence rates calculated from the claims data to the incidence rates from the National Cancer Registry of Korea.	The incidence rate of all cancers found using insurance claims data was very similar to the rate found in the national cancer registry. The age-, gender-, and disease-specific incidence rates were also similar between the two data sources. Insurance claims data can be a useful and cost-effective resource for health services research if appropriate algorithms are applied.	[12]
Saripalli et al. (2017)	Used machine learning classification methods to identify claims prone to rejection or denial. Investigated the reasons for claims denial - Engineered features using Claim Adjustment Reason Codes (CARCs) with high information gain. Developed a novel,	The study developed machine learning methods to accurately identify healthcare claims that are prone to denial or rejection. The study investigated the reasons for claims denial and recommended methods to engineer relevant features using CARCs to improve the accuracy of the machine	[13]

	state-of-the-art machine learning engine for automating and containing claims denial risks	learning models. The machine learning approach presented in the study is a novel, significant advancement in the state of practice for automating and containing healthcare claims denial risks.	
Arunkumar et al. (2021)	Used a publicly available Medicare dataset to classify providers as fraudulent or non-fraudulent. Employed the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance in the dataset. Utilized a hybrid approach combining clustering and classification techniques. Tested multiple machine learning algorithms to determine the most efficient one for the task	The researchers were able to detect fraudulent healthcare insurance claims using machine learning algorithms. The researchers used SMOTE to address the class imbalance problem in the data. The researchers used a hybrid approach based on clustering and classification to detect fraud.	[14]
Roy and George (2017)	Used machine learning techniques to detect auto/vehicle insurance fraud. Evaluated the performance of the machine learning model using a confusion matrix to calculate metrics like accuracy, precision, and recall	Insurance fraud is a significant problem, costing the industry over $40 billion annually. The study focused on using machine learning techniques to detect auto/vehicle insurance fraud. The performance of the machine learning models was evaluated using metrics like accuracy, precision, and recall.	[15]
Nabrawi and Alanazi (2023)	The study used a retrospective cohort design, analyzing a dataset of healthcare claims to build machine learning models for fraud detection. The dataset was highly imbalanced, with most cases labeled as fraud. The authors used the SMOTE technique to balance the dataset. The authors used various evaluation metrics to assess the reliability of the machine learning models, including accuracy, precision, recall, ROC, and AUC. The three machine learning models used were random forest, logistic regression, and artificial neural networks.	The random forest model was the best performing model, with an accuracy of 98.21%, and identifying policy type, education, and age as the most significant features contributing to healthcare fraud. The logistic regression and artificial neural network models also performed well, with accuracies of 80.36% and 94.64%, respectively. All three models demonstrated strong evaluation metrics, indicating their effectiveness in detecting healthcare fraud.	[16]
Ramani et al. (2024)	The study explored various machine learning models, including Random Forest Regression, Multiple Linear Regression, XGBoost Regression, Gradient Boosting Regression, and Decision Tree Regression, to predict health insurance claim amounts.	The experimental results demonstrated the efficacy of these machine learning models in making accurate predictions. The Random Forest model achieved a remarkable 96.7% accuracy, outperforming the other models tested.	[17]
Saraswat et al. (2023)	Used machine learning classification algorithms to predict health insurance claims. Developed a tool to help organizations decide whether to provide insurance to employees. Detected insurance fraud within the organization	The study aimed to develop a machine learning tool to help companies decide which employees should receive insurance coverage, in order to save time and resources. The study applied machine learning techniques to the problem of predicting health insurance claims, which has not been extensively explored before. The study's findings include the ability to save time and resources for organizations, as well as the potential to detect insurance fraud.	[18]

TABLE 1: Summary of Work Done on Insurance Claims and Healthcare Insurance Claims

AUC: Area Under the ROC Curve, ROC: Receiver Operator Characteristic

Machine learning and advanced data analysis methods have proven essential in addressing key challenges within the insurance industry, such as fraud detection, claim prediction, and customer behavior analysis. Researchers have employed diverse approaches, including clustering techniques, to identify unusual patterns, predictive models for analyzing claims, and classifiers like random forests and decision trees for enhanced accuracy. Strategies like dataset balancing and feature selection further optimize outcomes. These advancements not only improve decision-making processes but also offer scalable and efficient solutions to automate operations and reduce risks across various insurance applications (Table *1*).

This study addresses gaps in understanding health insurance claim prediction and eligibility assessment by focusing on the utility of logistic regression. While previous research has explored various machine learning methods for fraud detection or claim prediction, this study is unique in its emphasis on logistic regression's interpretability and practical applicability. By analyzing key predictors and their impact on claim likelihood, the study provides actionable insights often overlooked in broader analyses. Unlike past studies, which primarily prioritize model accuracy, this research investigates the statistical significance of predictors and their real-world implications for operational efficiency, customer service, and risk mitigation. Through its targeted approach, the study contributes to a deeper understanding of health insurance claims, offering insurers reliable and interpretable tools to optimize decision-making and

enhance financial stability.

The study aims to develop a predictive model that enables insurers to optimize operations, improve customer service, and mitigate financial losses by identifying high-risk policies or customers early. The performance of logistic regression will be assessed in terms of its accuracy, interpretability, and practical application in insurance claim prediction. Additionally, the research investigates the statistical significance and practical impact of key predictors on claim likelihood, providing actionable insights to enhance decision-making in the insurance sector.

Materials And Methods

The initial step involves defining the issue and creating a hypothesis suggesting that demographic, health, and lifestyle elements can be used to predict insurance claims. Subsequently, the dataset is collected and preprocessed. This involves addressing missing data, encoding categorical variables, normalizing features, and managing class imbalance through techniques such as Synthetic Minority Over-sampling Technique. The dataset is then divided into five folds for cross-validation to ensure each fold is used as a test set once and as part of the training set four times. Decision tree, logistic regression, and support vector machine (SVM) models are trained on the training data, and their performance is evaluated on the test data using metrics such as accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (ROC-AUC) (Figure *1*).

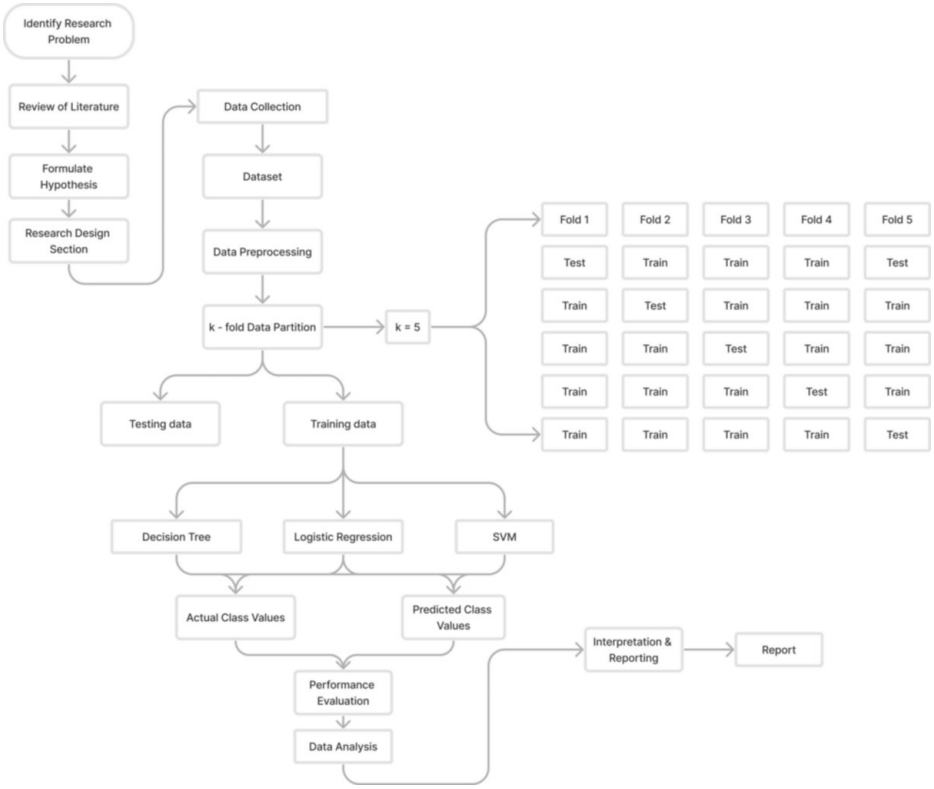

FIGURE 1: Workflow Description of Model Implementation

SVM: Support Vector Machine

The analysis focuses on comparing the performance of the models, identifying the most predictive features, and addressing challenges related to small sample size and class imbalance. The final report provides a summary of the methodology, preprocessing steps, model performance, and conclusions. It aims to offer insights into predicting insurance claims and assist the insurance industry in identifying high-risk individuals and enhancing health outcomes.

In this study, we applied a systematic methodology to analyze the dataset, identify key predictors, and develop a logistic regression model for predicting health insurance claims based on eight parameters, including age, BMI, smoking status, and medical costs. Our approach involved starting with the preprocessing of the data, which included cleaning and transforming the data to remove any imbalance in the dataset. We then delved deeper into the data by analyzing and then applying the model. Through this process, we were able to identify key patterns and trends in the dataset regarding fraud and non-fraud

activities that were not initially apparent.

The dataset used in this study, the "Sample Insurance Claim Prediction Dataset" from Kaggle, is a synthetic dataset designed for research and educational purposes. It consists of 1,338 samples, and we utilized all available data points for training and evaluation to ensure comprehensive model assessment [19]. The dataset consists of nine columns, each providing unique insights into the characteristics and behaviors of policyholders. It provides information on policyholders' characteristics, including age, gender, BMI, average daily steps, number of children, smoking status, residential region, and medical costs billed by health insurance. Derived from the "Medical Cost Personal Datasets," this dataset has been updated with sample values for analysis, making it a comprehensive and reliable source. This dataset offers an opportunity to explore the effectiveness of various techniques in overcoming these obstacles and developing a reliable classification model to anticipate insurance claims accurately.

To ensure data accuracy, all columns have been appropriately identified with their respective data types: integers for categorical variables (sex, children, smoker, region, insurance claim) and floats for continuous variables (age, BMI, charges). Furthermore, no missing values are present in any of the columns, as indicated by the output of df. isnull().sum() and df. notnull().sum(). Therefore, the dataset is complete and does not require imputation or removal of missing values. The distribution of data and skewness of each attribute have been examined. Age and sex exhibit minimal skewness, indicating relatively balanced distributions. However, BMI, children, smoker, region, and insurance claim status show varying degrees of skewness, which may require normalization or transformation during model training. The insurance claim status (target variable) appears slightly imbalanced, with approximately 58.5% of policyholders claiming insurance. Addressing class imbalance may be necessary during model training to prevent bias and improve prediction accuracy.

The correlation matrix revealed values ranging from -1 to 1, indicating the strength and direction of relationships between features. A weak positive correlation (0.3) between age and medical charges suggests that medical expenses slightly increase with age, likely due to age-related health conditions. A moderate positive correlation (0.79) between smoking status and medical charges highlights significantly higher expenses for smokers, consistent with established health risks associated with smoking. A weak negative correlation (-0.41) between the number of children and medical charges suggests that individuals with more children tend to incur slightly lower medical expenses, possibly influenced by demographic factors. A weak positive correlation (0.11) between age and insurance claim status indicates older individuals are slightly more likely to file insurance claims. Negligible correlations (~0.05) were observed for sex and region, suggesting minimal impact of these features on medical charges (Figure *2*).

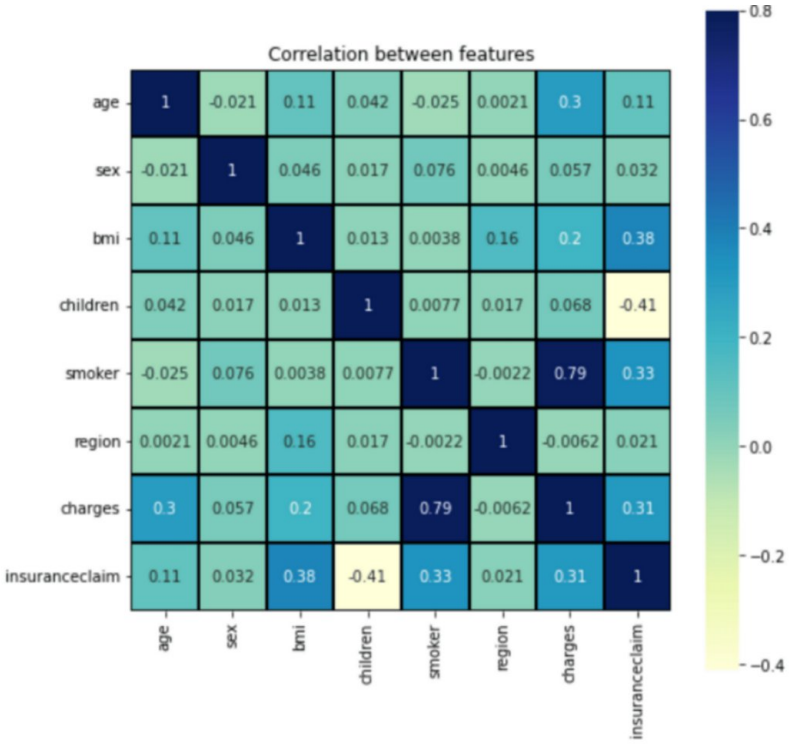

FIGURE 2: Correlation Matrix

Logistic regression was chosen as a model because it aligns closely with the study's objective of predicting binary outcomes, whether an insurance claim will be made or not. Its simplicity and computational efficiency make it a practical tool for real-world applications, particularly in the insurance domain. Additionally, logistic regression is widely valued for its interpretability, as it provides clear coefficients that indicate the influence of each predictor on the likelihood of an outcome. This makes it particularly useful for deriving actionable insights, such as identifying key risk factors like age, BMI, or smoking status, which are critical for insurers aiming to optimize operations and mitigate financial losses.

The decision tree model was included for its ability to capture non-linear relationships and interactions among variables, which are often present in complex datasets like healthcare insurance. Decision trees inherently perform feature selection during training, which simplifies the process of identifying significant predictors. Moreover, the decision tree achieved the highest performance among the models tested, with an accuracy of 95.52%, making it a strong candidate for predicting insurance claims. Its interpretability and ease of use also make it a valuable addition for comparative purposes, providing a clear understanding of how features influence the model's decisions.

SVM was selected as a comparative model due to its robustness in handling high-dimensional data and its ability to model non-linear decision boundaries using kernel functions. While SVM is generally effective in many classification tasks, its performance in this study was limited, achieving an accuracy of only 66%. This lower performance highlights the challenges SVM faces with imbalanced datasets, such as those seen in insurance claims. Including SVM provided a useful benchmark against simpler models like logistic regression and more adaptable models like decision trees, showcasing the strengths and weaknesses of each approach.

The study employed several metrics to evaluate model performance comprehensively. Accuracy was used to measure the proportion of correctly classified instances, providing an overall effectiveness score. To address class imbalance, precision and recall were utilized to evaluate the reliability of positive predictions and the model's ability to identify actual claims, respectively. The F1 score, a harmonic mean of precision and recall, offered a balanced metric for imbalanced datasets. ROC-AUC assessed the model's ability to distinguish between claim and non-claim classes, while the confusion matrix provided detailed insights into true positives, false positives, true negatives, and false negatives.

Mathematically,

Precision: The ratio of true positives (correctly predicted instances) to the sum of true positives and false positives (incorrectly predicted instances). It indicates the accuracy of positive predictions.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{1}$$

Recall: The ratio of true positives to the sum of true positives and false negatives (missed instances). It indicates the ability of the model to find all the positive instances.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2}$$

F1 score: The harmonic mean of precision and recall, providing a balanced measure of both metrics. It is a single metric that summarizes the model's performance.

$$F_1\,\text{Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Support: The number of instances in each class, indicating how many instances belong to each class in the dataset.

$$\text{Support}_{\text{positive}} = \text{True Positive} + \text{False Negative} \tag{4}$$
$$\text{Support}_{\text{negative}} = \text{True Negative} + \text{False Positive} \tag{5}$$

Accuracy: The proportion of correctly classified instances out of all instances. It measures the overall performance of the model in correctly predicting class labels.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}} \tag{6}$$

GridSearchCV was employed in the study to optimize the hyperparameters of the machine learning models, including logistic regression, decision trees, and SVM. Hyperparameters, such as the regularization parameter for logistic regression, maximum depth for decision trees, or kernel type for SVM, significantly influence model performance. Instead of relying on default values, GridSearchCV systematically searched through a predefined grid of hyperparameter combinations to identify the configuration that maximized model performance.

Using cross-validation, GridSearchCV evaluated each combination of hyperparameters by splitting the training data into multiple folds. This ensured that the selected hyperparameters generalized well to unseen data, reducing the risk of overfitting or underfitting. By automating the tuning process, GridSearchCV enabled the study to achieve optimal results, improving the predictive accuracy, precision, recall, and overall robustness of the models.

Results

The model was implemented in this research using the scientific Python development environment on the Anaconda platform, running on the Windows 10 operating system. To carry out the study, several libraries, including NumPy, pandas, matplotlib, and sklearn, were utilized to develop the predictive model.

The evaluation of our models revealed clear differences in their predictive capabilities for health insurance claims, with the decision tree model emerging as the most effective. It achieved the highest performance across all metrics, with an accuracy of 96%, precision of 94%, recall of 99%, and an F1 score of 96%, showcasing its ability to handle the complexities of the dataset. The decision tree's strength lies in its capacity to model non-linear relationships, capture feature interactions, and inherently perform feature selection during training, making it well suited for tasks involving varied predictor relationships. In contrast, logistic regression demonstrated balanced and reliable performance, with an accuracy of 82%, precision of 81%, recall of 88%, and an F1 score of 85%, serving as a dependable baseline model. Its interpretability allowed for actionable insights, such as identifying smoking status, BMI, and age as significant predictors of claims, which is particularly valuable for insurers seeking to understand and mitigate risks. However, logistic regression's linear nature limits its ability to capture complex interactions among features, which the decision tree handles adeptly. SVM, on the other hand, performed the poorest, with an accuracy of 66%, precision of 76%, recall of 59%, and an F1 score of 66%. Despite SVM's ability to model non-linear decision boundaries, it struggled with the dataset's class imbalance and feature complexity, underscoring its limitations in this specific context without further optimization (Table *2*).

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.82	0.81	0.88	0.85
Decision Tree	0.96	0.94	0.99	0.96
SVM	0.66	0.76	0.59	0.66

TABLE 2: Model Results Comparison

SVM: Support Vector Machine

A comparative analysis of the models highlights the decision tree's dominance, particularly in datasets requiring nuanced feature interactions and robust handling of imbalances. Similar trends have been observed in other studies, where decision trees outperform other classifiers in handling complex relationships within data. Meanwhile, logistic regression remains a practical choice due to its simplicity, interpretability, and reasonable accuracy, making it suitable for real-world applications. SVM's lower performance emphasizes the need for careful consideration of dataset characteristics, such as imbalance and feature scaling, when selecting models. These results were achieved through a systematic methodology that included data preprocessing, such as addressing missing values, encoding categorical variables, and normalizing features. Hyperparameter tuning via GridSearchCV and five-fold cross-validation further ensured that the models were rigorously optimized and evaluated. In conclusion, the decision tree model proved to be the most effective for predicting health insurance claims in this study, while logistic regression served as a robust and interpretable alternative, and SVM highlighted the challenges of applying certain algorithms to imbalanced datasets. This comparative study underscores the importance of selecting models based on dataset characteristics and project goals to ensure optimal outcomes.

Discussion

The decision tree model emerged as the most effective in this study, outperforming logistic regression and SVM in predicting health insurance claims. With an accuracy of 96%, a recall of 99%, and an F1 score of 96%, it demonstrated its ability to handle the complexity of the dataset. The decision tree's inherent feature selection and capacity to model non-linear relationships allowed it to capture significant interactions among predictors, such as smoking status, BMI, and age, which are critical for understanding claim likelihood. Additionally, its interpretability, simplicity in visualizing decision-making, and high predictive power make it a reliable choice for health insurance claim prediction tasks.

In comparison, logistic regression offered balanced and interpretable results, with an accuracy of 82% and an F1 score of 85%. While its linear nature limits its ability to model complex feature interactions, its coefficients provided clear insights into the influence of predictors, such as the strong impact of smoking on medical expenses. This interpretability is particularly valuable in operational scenarios where understanding risk factors is as important as prediction accuracy. SVM, on the other hand, struggled with the dataset's characteristics, achieving only 66% accuracy and showing limitations in handling class imbalance and feature relationships, making it unsuitable for this specific application without significant tuning.

Compared to existing research on health insurance claims, this study introduces key improvements and addresses critical gaps. Previous studies often prioritized model accuracy, leveraging advanced machine learning techniques such as random forests or ensemble methods, to achieve high predictive performance. While effective, these models are often criticized for their lack of interpretability and the computational cost associated with their implementation. By contrast, this research emphasizes both accuracy and interpretability, combining the high predictive capability of the decision tree with the practical insights provided by logistic regression. Unlike other studies that primarily focus on identifying fraud or claim probability at a superficial level, this research delves deeper into understanding the statistical significance and real-world implications of predictors, offering insurers actionable insights into optimizing their operations.

Additionally, the methodology adopted in this study enhances the robustness of the results. Existing research frequently employs simple data preprocessing or lacks comprehensive hyperparameter tuning, which can lead to biased or suboptimal outcomes. This study overcame these challenges through rigorous preprocessing steps, including handling categorical variables, balancing classes, and tuning hyperparameters using GridSearchCV. These methodological improvements ensure that the models were not only accurate but also generalizable to unseen data, a factor often overlooked in prior work.

In summary, the decision tree model's superior performance, combined with the structured and robust approach of this study, positions it as the best choice for health insurance claim prediction. While existing research has contributed valuable insights into the broader application of machine learning in insurance, this study provides a more targeted and practical perspective by balancing predictive power with interpretability. This dual focus on accuracy and actionable insights equips insurers with tools that are not only effective but also aligned with real-world decision-making needs, ultimately improving financial stability and customer satisfaction in the insurance industry.

Conclusions

In conclusion, this research emphasizes the decision tree model as the most effective approach for predicting health insurance claims, showcasing exceptional performance with an accuracy of 96%, a recall of 99%, and an F1 score of 96%. Its ability to handle complex interactions between key predictors, such as smoking status, BMI, and age, while maintaining interpretability, makes it an ideal choice for insurers looking to optimize their decision-making processes. While the decision tree excels in capturing non-linear relationships, logistic regression remains valuable for its simplicity, efficiency, and interpretability, offering clear insights into risk factors for transparent decision-making. This study not only highlights the importance of model accuracy but also emphasizes the need for transparency and practical application in real-world scenarios. By comparing multiple models, it offers a well-balanced trade-off between predictive power and interpretability where decision trees provide high accuracy (96%) and feature importance insights, while logistic regression offers a simpler, more interpretable approach suitable for operational decision-making in the insurance industry. Moreover, the rigorous methodology adopted in this study, including thorough preprocessing and hyperparameter tuning, ensures that the models are not only accurate but also generalizable, paving the way for future advancements in health insurance claim prediction. Ultimately, this work contributes valuable insights that can aid insurers in making informed decisions, improving financial stability, and enhancing customer satisfaction.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of

the work.

Concept and design: Vaishnavi P. Thakre

Acquisition, analysis, or interpretation of data: Vaishnavi P. Thakre, Rohit D. Poul, Ankush D. Sawarkar

Drafting of the manuscript: Vaishnavi P. Thakre, Rohit D. Poul, Ankush D. Sawarkar

Critical review of the manuscript for important intellectual content: Vaishnavi P. Thakre, Ankush D. Sawarkar

Supervision: Ankush D. Sawarkar

Disclosures

Human subjects: Consent was obtained or waived by all participants in this study. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Xie Y, Schreier G, Chang DCW, Neubauer S, Liu Y, Redmond SJ: Predicting days in hospital using health insurance claims. IEEE Journal of Biomedical and Health Informatics. 2015, 19:1224-1233. 10.1109/jbhi.2015.2402692
2. Rayan N: Framework for analysis and detection of fraud in health insurance. 2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS), Singapore. 2019, 47-56. 10.1109/CCIS48116.2019.9073700
3. Rukhsar L, Bangyal WH, Nisar K, Nisar S: Prediction of insurance fraud detection using machine learning algorithms. Mehran University Research Journal of Engineering & Technology. 2022, 41:33-40.
4. Kumar R, Duggirala A: Health insurance as a healthcare financing mechanism in India: key strategic insights and a business model perspective. Vikalpa. 2021, 46:112-128. 10.1177/02560909211027089
5. Faulkner LA, Schauffler HH: The effect of health insurance coverage on the appropriate use of recommended clinical preventive services. American Journal of Preventive Medicine. 1997, 13:453-458. 10.1016/s0749-3797(18)30141-7
6. Withagen-Koster AA, van Kleef RC, Eijkenaar F: High-risk pooling for mitigating risk selection incentives in health insurance markets with sophisticated risk equalization: an application based on health survey information. BMC Health Services Research. 2024, 24:273. 10.1186/s12913-024-10774-x
7. Antwi S, Zhao X: A logistic regression model for Ghana National Health Insurance claims. International Journal of Business and Social Research. 2012, 139-47.
8. Sun C, Li Q, Li H, Shi Y, Zhang S, Guo W: Patient cluster divergence-based healthcare insurance fraudster detection . IEEE Access. 2019, 7:14162-14170. 10.1109/access.2018.2886680
9. DeVoe JE, Tillotson CJ, Wallace LS: Children's receipt of health care services and family health insurance patterns. The Annals of Family Medicine. 2009, 7:406-413. 10.1370/afm.1040
10. Smith KA, Willis RJ, Brooks M: An analysis of customer retention and insurance claim patterns using data mining: a case study. Journal of the Operational Research Society. 2000, 51:532-541. 10.1057/palgrave.jors.2600941
11. Konrad R, Zhang W, Bjarndóttir M, Proaño R: Key considerations when using health insurance claims data in advanced data analyses: an experience report. Health Systems. 2019, 9:317-325. 10.1080/20476965.2019.1581433
12. Seo HJ, Oh IH, Yoon SJ: A comparison of the cancer incidence rates between the National Cancer Registry and insurance claims data in Korea. Asian Pacific Journal of Cancer Prevention. 2012, 13:6163-6168. 10.7314/apjcp.2012.13.12.6163
13. Saripalli P, Tirumala V, Chimmad A: Assessment of healthcare claims rejection risk using machine learning. 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), Dalian, China. 2017, 1-6. 10.1109/HealthCom.2017.8210758
14. Arunkumar C, Kalyan S, Ravishankar H: Fraudulent detection in healthcare insurance. Advances in Electrical and Computer Technologies. Sengodan T, Murugappan M, Misra S (ed): Springer, Singapore; 2021. 711:1-9. 10.1007/978-981-15-9019-1_1
15. Roy R, George KT: Detecting insurance claims fraud using machine learning techniques. 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Kollam, India. 2017, 1-6. 10.1109/ICCPCT.2017.8074258
16. Nabrawi E, Alanazi A: Fraud detection in healthcare insurance claims using machine learning. Risks. 2023, 11:160. 10.3390/risks11090160
17. Ramani K, Kumar ST, Datta PP, Jamuna P, Nithin KS: Predicting health insurance claim amount through machine learning algorithms. 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), Bangalore, India. 2024, 1-6.

10.1109/ICITEICS61368.2024.10625132
18. Saraswat BK, Singhal A, Agarwal S, Singh A: Insurance claim analysis using traditional machine learning algorithms. 2023 International Conference on Disruptive Technologies (ICDT), Greater Noida. 2023, 623-628. 10.1109/ICDT57929.2023.10150491
19. https://www.kaggle.com/datasets/easonlai/sample-insurance-claim-prediction-dataset.