SCIENCE SET
OPEN ACCESS PUBLISHERS

Research Article

Global Journal of Translational Science and Integrated Technologies

Predicting Medical Insurance Claims Using Machine Learning: A Data-Driven Approach for Improved Risk Assessment

Bu Kai Yean[1], Angelina Toh[1], Ong Zheng He[1], Jason[1], Soobia Saeed[1*]

Department of Computer Science, Taylors University, Malaysia

[*]Corresponding author: Soobia Saeed, Department of Computer Science, Taylors University, Malaysia.

Submitted: 30 December 2025 **Accepted:** 06 January 2026 **Published:** 13 January 2026

Citation: *Bu, K. Y., Toh, A., Ong, Z. H., Jason, & Saeed, S. (2025). Predicting Medical Insurance Claims Using Machine Learning: A Data-Driven Approach for Improved Risk Assessment. Glo J Tran Sci & Int Tec, 2(1), 01-07.*

Abstract

The present study thoroughly examines the prediction of health insurance claims by demographic and health-related features to the end of supporting more precise and uniform risk evaluation in the insurance sector. The dataset, which was downloaded from Kaggle, contains the following variables: age, BMI, blood pressure, diabetes status, smoking status, gender, number of children, and area. Comprehensive preprocessing was performed to completely eliminate missing values, duplicates, and categorical inconsistencies while also performing equal feature scaling through utilizing median/mode imputation, categorical standardization, label and one-hot encoding, and robust scaling. Four distinct types of regression models were created and assessed utilizing both train-test split and K-Fold Cross Validation: Linear Regression, Decision Tree Regression, Random Forest Regression, and K-Nearest Neighbors. The performance of these models was evaluated using R-Squared (R2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The Random Forest model showed the highest predictive accuracy and consistency across all metrics and therefore outperformed the other models in both evaluation settings. The results point out the success of ensemble methods in the frame of revealing complex relationships in healthcare insurance datasets and also indicate the necessity of clean, well-processed data for the improvement of predictive performance. Suggested approaches for enhancing the quality of the dataset include an oversampling or stratified sampling method to eliminate the problem of data imbalance, as well as applying more sophisticated imputation methods like KNN Imputer and Iterative Imputer for better data quality. This work has shown how machine learning can be a strong ally for insurance companies; they would be able to charge fair premiums and make informed policy decisions with the help of this data-driven approach.

Keywords: Linear Regression, Decision Tree Regression, Random Forest Regression, and K-Nearest Neighbors.

Introduction

Background of Study

The insurance industry is undergoing a transformation due to the variation of patients and their health conditions, which are mainly the causes of risk assessment and cost prediction difficulties. Besides, the insurance claim amount can vary, depending on the viewpoint, and be inconsistent and subjective [1]. Hence, a fair amount of claim is a must to conduct a win-win situation for both insurer and policymakers. With the ever-increasing medical cost, the insurance sector needs to implement data-driven techniques to not only make the right decisions but also to revise their policies accordingly [2, 3].

The project under discussion conducts an analysis of a medical insurance claim dataset incorporating patient demographics and health-related features like age, gender, BMI, blood pressure, diabetic status, number of kids, smoking status, and area of living. Our goal is to unravel the patterns and dependencies that can potentially impact the insurance claim by scrutinizing and comprehending these features [4, 5]. The main goal of this project is to create a predictive model that forecasts insurance claims based on risk evaluation, which is an important factor when setting premiums for insurers. In addition, the application of various data mining techniques enables the creation of a just and reasonable decision-making process relating to insurance for both the insurers and the insurance companies. Furthermore,

this research work connects the dots between the requirements of the real world and data science, thus, facilitating more precise insurance claims for both insurers and insurers [6, 7].

Dataset

Data Source

The paper makes use of a dataset that is available publicly on Kaggle under the name “Insurance Claim Analysis: Demographic and Health” created by The Devastator. The data set comes with 1,340 rows and 11 columns which gives a brief overview of people's demographics and health-related characteristics that are important for the prediction of insurance claims. The main features include the person's age, sex, weight (BMI), blood pressure, diabetes (yes/no), smoking (yes/no), number of children, and the area of residence. The very first analysis using df.shape, df.head(), and df.tail() has shown the dataset's structure and pointed out the issue of missing data, mostly affecting the age and region variables. Descriptive statistics derived from df.describe() have made it clearer regarding the dataset's characteristics, an average BMI of 30.67 for the group which indicates that the majority of them are overweight, and a wide variability of claims which is shown by a high standard deviation. The blood pressure readings also varied a lot, since the upper quartile of the records had readings higher than 86, inferring a slight trend of the represented ones having high blood pressure. More analyses, like bar charts and scatter plots, gave a clearer picture of the distributions and relationships of the features. For instance, the diabetes variable was found to have nearly the same number of “diabetic” and “non-diabetic” individuals, whereas box plot analysis indicated that smokers typically have higher claim amounts than the non-smokers. The scatter plots showed that one of the factors affecting the other is high blood pressure, which in turn leads to higher claim amounts. Moreover, with a higher BMI, the patient’s blood pressure is likely to be high. A gender-wise study of smoking habits showed a small difference, with the number of male smokers being larger than that of female smokers. The analysis of the dataset revealed the existence of significant patterns that could be used as a solid basis for predictive modeling, nevertheless, the dataset still requires a lot of preprocessing due to the presence of missing values, inconsistent formatting of categorical variables, and different scaling of numerical variables. With the help of df.info(), one can find out the features that have missing values along with their respective data types. The chart indicates that the variables region and age have missing values and this will be checked again in the preprocessing phase in Section 3.2.2.

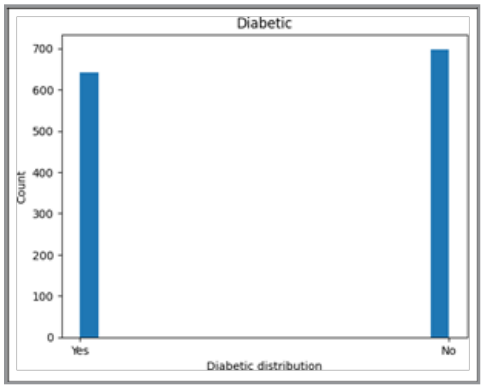

Figure 1: Bar Chart of Statistic

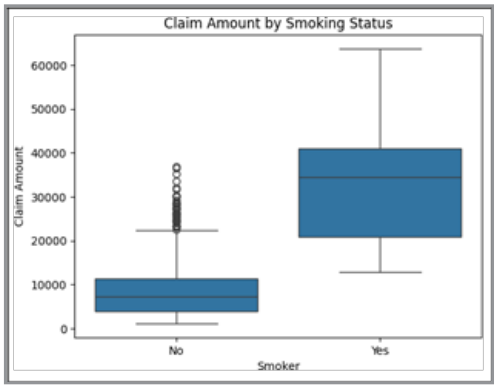

Figure 2: Box Plot of Diabetic Claim Amount by Smoking Status

A bar chart was selected to facilitate the analysis of the distribution of diabetic features, as illustrated in Figure 1. The distribution of patients classified as 'diabetic' and 'non-diabetic' is equal, leading to the conclusion that this dataset exhibits a balanced distribution. A box plot, presented in Figure 2, was employed to examine the relationship between smoking status and claim amount, indicating that non-smokers typically have a lower claim amount, whereas smokers tend to have a relatively higher claim amount. Although there may be potential outliers, they are still considered valid data since they remain within the acceptable range for claim amounts, which could be influenced by other factors.

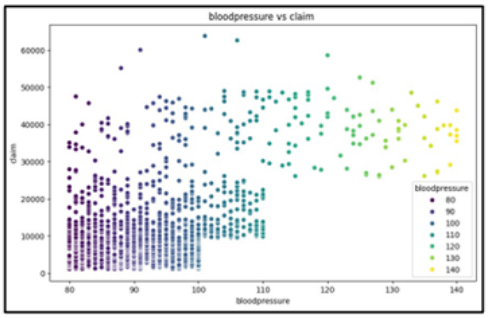

Figure 3: Scatter Plot of pressure vs claim

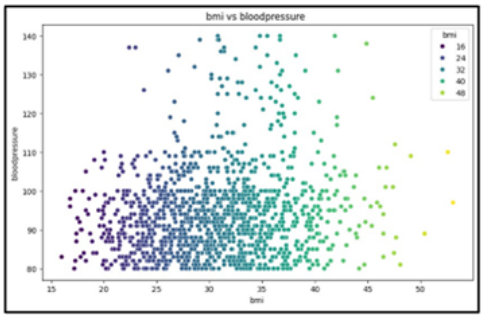

Figure 4: Scatter Plot of blood BMI vs blood pressure

The scatter plot presented in Figure 4 is employed to elucidate the connection between blood pressure and claim amount. The plot indicates that there is a grouping of low blood pressure readings in the data set that correlates with low claim amounts. It is also noted that there are no data points showing the situation of high blood pressure and low claim amount, so one can see that blood pressure is a very influencing factor when it comes to claim amount. Besides, the scatter plot allows visualization of the correlation between the two variables, namely BMI and blood pressure, hence the latter has been found to be one of the factors that develop the condition. According to Figure 4, it is observed that the obese patients are usually sweating their blood pressure but the thinner ones are at a normal range as depicted in the graph. This implies that great body weight people tend to be treated for hypertension more often than their normal-weight counterparts.

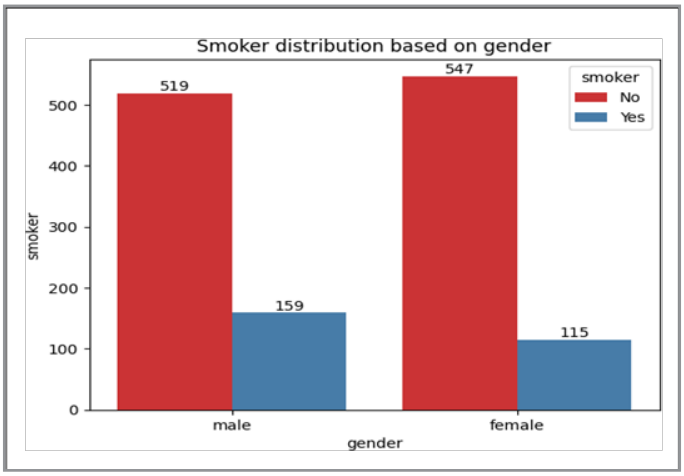

Figure 5: Bar Chart of Smoker Distribution Based on Gender

The bar chart depicted in Figure 5 indicates the presence of 159 male smokers and 115 female smokers while non-smokers are recorded as 519 and 549 for males and females respectively. Hence, a minor difference between the sexes can be noticed concerning smoking habits.

Data-Related Issues and Preprocessing

The investigation of the Insurance Claim Analysis dataset uncovered numerous data-quality issues that could widely influence the very machine learning accuracy and reliability. This section describes the main issues that were found during the exploratory phase, and furthermore, outlines the preprocessing strategies that were applied in order to get a clean, consistent, and fully analyzable dataset for model development [8].

Key Data Issues

The examination of the initial dataset first brought to light the very problems that are typical in health and demographic datasets in real-world scenarios. Missing entries in both the numerical and categorical datasets were considered to be the main cause of the eventual biased decisions or even the total failure of the analyzers or the models, if such problems were left unnoticed. Even different categories had inconsistencies within them—for instance, there were cases where mixed-up lettering and unnecessary spaces at the start or at the end of the words caused the same thing to be counted several times and made it hard for the computer to give one correct interpretation. On top of that, there were various text dissimilarities and unnecessary variations in the strings which were taking toll on the categorical processing [9, 10]. The checking for duplicates in the dataset was another procedure since duplicate records lead to misrepresentation of certain patterns, thus impacting the learning process of the model. Besides that, there were also outliers that were a source of concern and mainly in the case of numerical attributes, as their presence could be the reason statistics would be interpreted incorrectly, or the model training would be affected. Moreover, a number of categorical variables had not been transformed into numerical ones yet, so they were not suited for many machine learning algorithms [11-13]. So, it was indispensable to deal with such problems via systematic preprocessing in order to have a final dataset that was clean, standardized, and suitable for high-quality predictive modeling.

Data Preprocessing

Dropping Uninformative Data

The preprocessing procedure kicked off by getting rid of non-informative columns like the index and Patient_ID. To the model, these identifiers bear no predictive worth and do not have a significant correlation with the amount of the claim. By getting rid of such features, the model would become more efficient, the noise would be reduced, and interpretability would be made easier.

Handling Missing Values

The management of missing values was so to speak a very crucial preprocessing operation that the completeness and reliability of the dataset depended on it. The missing entries were initially spotted using the .isnull().sum() method which made it possible to see at a glance the magnitude and the places where the missing data was. The rows which had several missing values were deleted, while numerical columns that were left with gaps were filled by the median value of each feature respectively. The mode was used for categorical variables—this means that the missing entries were replaced by the category that occurs most frequently. A dataset comparison before and after the imputation confirmed that all the missing values had, indeed, been treated. The procedure also ensured that the final dataset was free from null entries, thus completely removing the chance of getting errors during the training of the model.

Handling Duplicates

In order to maintain data integrity, the dataset was assessed for duplicates with the help of the .duplicated() function. The investigation verified that there were not any duplicate rows, which meant that every record was a different case. The prevention of duplicating data ensures that the learning process is not distorted, and the modeling results are reliable.

Categorical Standardization

The standardization of categorical variables was done in order to get rid of inconsistencies in the text format. The differences in letters and spacing were corrected by changing all categorical strings to lowercase and removing excess whitespace. This action made the categories such as gender, smoker status, and diabetic status uniform so that they could not be misinterpreted by the machine learning models which consider the text variations as separate categories.

Feature Encoding

Categorical features were transformed into a numerical format via label encoding and one-hot encoding that were applicable to machine learning algorithms simultaneously. Label encoding was assigned to the binary categorical variables such as gender, diabetic status, and smoking status that were converted to 0–1 numerical representation. For the region variable, which consists of several categories, one-hot encoding was applied to build binary columns for every region. This method guarantees that the model does not confound the differences in regions for non-existent ordinal implication. After encoding, the original region column was removed while the new binary features were retained for training.

Feature Scaling

Scaling using the Robust method was performed on numerical features such as age, BMI, blood pressure, and claim amount. This technique that employs the median and interquartile range was selected because of its property of being immune to outliers that could otherwise be a great source of model distortion if not adjusted. Also, there were huge differences in ranges of numerical variables especially the claim amount variable, which was very high, thus scaling placed all features on a similar range. This transformation in turn made sure that the model did not give undue emphasis to the features with larger numerical ranges due to their numerical range and provided the model with all attributes for effective learning.

Final Dataset Preview

The last look at the cleaned dataset assured the proper implementation of the preprocessing steps. The dataset that was obtained contained no missing values, duplicates, textual inconsistencies, or unscaled features. Properly organized, the encoded categorical variables and the scaled numerical features indicated that the data was completely ready for model training and testing.

Methodology

In this section, the paper will further discuss the four primary data mining techniques utilized in this research: Linear Regression, Decision Tree Regression, Random Forest, and K-Nearest Neighbors. These models were selected because of their reliability and accuracy in generating predictions.

Linear Regression

Model Initialization

The Linear Regression model was initialized with the help of the Linear Regression() function which is part of the sklearn.linear_model library. This is the first step to setting up the model which eventually will be able to catch linear relationships between the independent variables and the amount of insurance claim. The model is able to handle the training process using the prepared dataset after it is initialized.

Train-Test Process

In order to test how well the model works, the dataset was divided into a training set and a test set in an 80:20 ratio. The feature set (X) was obtained by eliminating the target variable, Claim, while the rest of the variables were used for training the model. The training of the Linear Regression model was done on the training data with the help of the linear_model.fit(X_train, y_train) method which means the model has absorbed the patterns and the correlations that will be later used for making predictions on the testing set.

Decision Tree Regression

Model Initialization

The Decision Tree Regression model was initialized with the help of the Decision Tree Regressor(random_state=42) function from the sklearn.tree library. The data is divided into branches using the tree structure approach based on the thresholds set for each feature. A tree-like structure that is interpretable and can also capture non-linearity in the data is thus created.

Train-Test Process

Just like Linear Regression, the Decision Tree model was trained via the tree_model.fit(X_train, y_train) function. The model strives to minimize prediction error through the development of optimal splits in the training data. After training, the model predicts claims for the test data that was not previously seen.

Random Forest Regression

Model Initialization

The Random Forest model was set up via the RandomForestRegressor(n_estimators=100, random_state=42) class. By means of ensemble methods, the decision trees are constructed, and their predictions are thereafter combined to produce the final RF output that is more precise and stable. The downside of overfitting is completely eliminated, and the upside of prediction robustness is greatly amplified.

Train-Test Process

The Random Forest model underwent training using the identical train-test split structure. The forest_model.fit(X_train, y_train) function was used, and the model learned from different parts of the training data to effectively generalize patterns. The ensemble averaging not only helps in delivering improved predictive accuracy but also allows the model to cope with complex feature interactions.

K-Nearest Neighbors Regression

Model Initialization

The K-Nearest Neighbors (KNN) Regression model was established through the KNeighborsRegressor(n_neighbors=5) class. The model assumes that the amount of claim will be the average

of the values of the five closest data points (neighbors) in the training set, which means it will be highly sensitive to the distribution and scaling of the data.

Train-Test Process

The KNN model was trained with knn_model.fit(X_train, y_train) right after the dataset underwent scaling. Because of its distance-based approach, the model takes advantage of the closeness of identical data points to make predictions. The model trained later was applied to determine prediction values for claims in the test set that had not been seen.

Model Validation

Metrics Employed

Four performance metrics have been chosen to evaluate the effectiveness of regression models (Chugh, 2020).

R-Squared (R2)

R2 represents the proportion of variance in the dependent variable that can be explained by the model using the independent variables. A higher R2 value indicates that the model is better at explaining the variance in the outcome.

Mean Squared Error (MSE)

MSE is defined as the average of the squared differences between the actual and predicted values. The squaring of errors makes this metric particularly sensitive to outliers, as it heavily penalizes larger errors (Djellouli, 2018). Models that demonstrate superior performance will exhibit lower MSE values.

Root Mean Squared Error (RMSE)

RMSE is derived from the square root of MSE, which facilitates a more intuitive understanding of the error compared to MSE. Models that perform better will have lower RMSE values.

Mean Absolute Error (MAE)

MAE is calculated as the average of the absolute differences between the actual and predicted values. Compared to MSE, MAE is more robust to outliers since it considers the absolute difference rather than the squared difference. Models with better performance will show lower MAE values.

K-Fold Cross Validation

The K-fold cross-validation technique is utilized to validate the results of regression models. This method partitions the dataset into k equal segments, known as folds, and fits the model to k-1 folds, using the remaining fold as the test set for the trained model (Chugani, 2024). This process is repeated until each fold has been utilized as a test set once. The average of the metrics obtained from each iteration is calculated, providing a more stable and reliable assessment of the model's performance compared to a single train-test split, which may be biased based on the specific data division employed.

Interpretation of Results

The results of the Train-Test Split and K-Fold Cross Validation are detailed in Table 5.3. In the Train-Test Split, Random Forest emerged as the top performer with an R2 of 82.30%, followed by Linear Regression at 73.98%, Decision Tree at 63.08%, and K-Nearest Neighbors at 62.98%. The R2 performance of Decision Tree and K-Nearest Neighbors is nearly the same, with Decision Tree having a slight advantage over K-Nearest Neighbors. However, when comparing the MAE values of both models, K-Nearest Neighbors (0.4531) outperforms Decision Tree (0.4737), suggesting that K-Nearest Neighbors is more effective in predicting claims that are closer to the actual claim amount. Additionally, K-Nearest Neighbors exhibits marginally higher MSE and RMSE scores compared to Decision Tree, indicating that it experiences larger prediction errors, which accounts for its lower R2 value. Random Forest demonstrates the lowest MSE, RMSE, and MAE values, signifying that its predictions for claims are the most dependable among all the models. The metrics derived from K-Fold Cross Validation provide a basis for comparison against the metrics obtained from the Train-Test Split. All models experienced a slight decline in performance; however, Random Forest remains the top-performing model among the four. The only notable performance difference is that K-Nearest Neighbors has slightly surpassed Decision Tree, with R2 values of 55.83% and 55.57%, respectively. This suggests that the specific training and testing set utilized by the Train-Test Split method may favor the Decision Tree to a certain degree. Nevertheless, an analysis of the standard deviations of both models across the folds indicates that K-Nearest Neighbors demonstrates significantly more variability in performance, as evidenced by its standard deviation of 7.64%, which exceeds that of the Decision Tree at 6.42%. The remaining metrics, including MSE, RMSE, and MAE, did not reveal any substantial differences compared to the Train-Test Split metrics, indicating that the prediction errors of the models remain consistent across different training and testing sets, despite slight fluctuations in the R2 value. Consequently, it can be concluded that Random Forest is the superior model due to its strong performance in both training and testing sets, as well as its reliability across various folds in K-Fold Cross Validation, making it a trustworthy choice for the insurance claim regression task.

Table 1: Interpretation of Results for Train-Test Split and K-Fold Cross Validation

Model	Train-Test Split				K-Fold Cross Validation			
	R2 (%)	MSE	RMSE	MAE	R2 (%)	MSE	RMSE	MAE
Linear Regression	73.98	0.3027	0.5502	0.4116	68.80 ± 2.74	0.3135 ± 0.0522	0.5579 ± 0.0471	0.4201 ± 0.0263
Decision Tree	63.08	0.4294	0.6553	0.4737	55.57 ± 6.42	0.4382 ± 0.0208	0.6618 ± 0.0157	0.4695 ± 0.0150
Random Forest	82.30	0.2059	0.4538	0.3352	78.80 ± 1.86	0.2129 ± 0.0338	0.4598 ± 0.0374	0.3442 ± 0.0205
K-Nearest Neighbors	62.98	0.4306	0.6562	0.4531	55.83 ± 7.64	0.4419 ± 0.0962	0.6612 ± 0.0680	0.4597 ± 0.0412

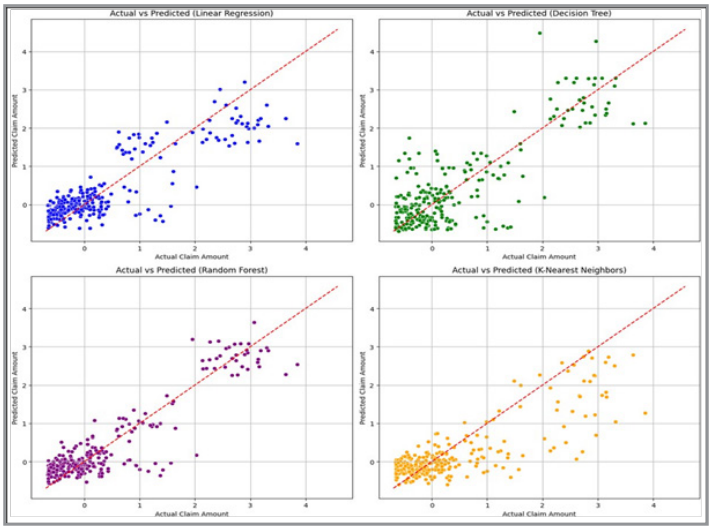

Figure 6: Visualization of the Results

The scatterplots illustrating the actual values in comparison to the predicted values for each regression model, utilizing the Train-Test Split method, are depicted in Figure 6. The predictions made by the Decision Tree and K-Nearest Neighbors exhibit greater variability than those of Linear Regression and Random Forest. Specifically, the Decision Tree tends to forecast larger claims on average for higher actual values, whereas K-Nearest Neighbors typically predict smaller claims on average. These scatterplots effectively illustrate the Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) for both the Decision Tree and K-Nearest Neighbors. Notably, K-Nearest Neighbors demonstrates a lower average dispersion for smaller actual claims, while the Decision Tree shows a higher dispersion, resulting in a greater MAE for the Decision Tree. Conversely, K-Nearest Neighbors presents slightly larger outliers compared to the Decision Tree, which negatively impacts its R-squared value, as well as leading to elevated MSE and RMSE values. In the comparison between Linear Regression and Random Forest, it is evident that Random Forest significantly outperforms Linear Regression, as the majority of the predicted claims are accurately clustered near the ideal prediction line, whereas the predictions from Linear Regression are widely dispersed. This disparity is reflected in their respective performance metrics, with Random Forest achieving more favorable results than Linear Regression across all evaluated parameters.

Conclusion and Future Work

Conclusion

The project outlined in this paper successfully developed and assessed several machine learning regression models aimed at predicting insurance claim amounts based on demographic and health characteristics. Following comprehensive data preprocessing, which included addressing missing data, standardizing categories, encoding, and applying robust scaling methods, four models were trained and evaluated: Linear Regression, Decision Tree Regression, Random Forest Regression, and K-Nearest Neighbors Regression. Ultimately, the Random Forest model emerged as the most effective among all tested models for prediction, demonstrating the highest R-square values and the lowest errors in the MSE, RMSE, and MAE metrics. The Random Forest model demonstrates its superiority over the other models by successfully uncovering intricate relationships in the data and generalizing them well across different validation sets. The findings suggest that even though simple methods still play an important role in prediction, more complex ones, such as Random Forest, can be used together to provide more accurate and trustworthy insurance claim predictions. The gap between data science techniques and insurance practice is thus narrowed down, leading to the generation of insights that might be very useful for the insurers in their decision-making process concerning risk evaluation.

Future Work

The presence of potential imbalances in a dataset may impact the accuracy of the model and its credibility as well. A possible moment of imbalances due to the Insurance Claim Analysis dataset could be the lack of uniform distribution of claim amounts. This means that if a greater part of the data is comprised solely of one type of range of claim values, for instance, low claim values as opposed to having an even distribution, this could result in poor model performance which might lead to a disruption in the prediction of insurance and risks. Thereby, a suggestion to overcome these data imbalances such as using techniques like oversampling, under sampling or stratified sampling to make the training set more balanced can be made. The recommendation to deal with potential data imbalances plays a crucial role in ensuring that the data is more dependable for prediction purposes. Not only that, but the use of more sophisticated imputation methods is another improvement that is advised. Missing data is normally managed by substituting it with median or simply excluded from the dataset. However, more advanced methods can be used, for instance, K-Nearest Neighbors (KNN) Imputer which selects the best similar data and fills in the missing values, and Iterative Imputer which predicts missing values based on other features. The use of these two methods not only makes the dataset neater but also more accurate for predictions.

References

1. Khela, F. (2024). Predicting health insurance claim costs: A data-driven approach using machine learning. Rochester Institute of Technology.
2. Nadipelli, S. R., Vijayan, N., Shelke, S., Agrawal, D., & Yadav, A. (2024).A Machine Learning Based Risk Assess-

ment System Prediction Algorithm for Examining Medical Insurance Costs.
3. Yu, S. H., Su, E. C. Y., & Chen, Y. T. (2018). Data-driven approach to improving the risk assessment process of medical failures. International journal of environmental research and public health, 15(10), 2069.
4. Moon, S., Jagadeesh, D., Sultana, S. R., & Prasanna, M. L. (2025). AI-Driven Machine Learning Model for Health Insurance Claim Prediction. International Journal of Communication Networks and Information Security, 17(3), 900-906.
5. Matloob, I., Khan, S., Bashir, B., Rukaiya, R., Khan, J. A., & Alfraihi, H. (2025). Data driven healthcare insurance system using machine learning and blockchain technologies. PeerJ Computer Science, 11, e2980.
6. Dritsas, E., & Trigka, M. (2022). Data-driven machine-learning methods for diabetes risk prediction. Sensors, 22(14), 5304.
7. Chiu, Y. L., Jhou, M. J., Lee, T. S., Lu, C. J., & Chen, M. S. (2021). Health data-driven machine learning algorithms applied to risk indicators assessment for chronic kidney disease. Risk management and healthcare policy, 4401-4412.
8. Dey, S., Basak, D., & Mondal, K. C. (2025). Advancement in Insurance Risk Prediction: A Review of Data-Driven Approaches. Journal of Convergence in Technology and Management: Global Nexus, 1(2), 71-76.
9. Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC medical informatics and decision making, 19(1), 1-15.
10. Akter, J., Roy, A., Ara, J., & Ghodke, S. (2025). Using machine learning to detect and predict insurance gaps in US healthcare systems. Journal of Computer Science and Technology Studies, 7(7), 449-458.
11. Pandiri, L., & Chitta, S. (2024). Machine Learning-Powered Actuarial Science: Revolutionizing Underwriting and Policy Pricing for Enhanced Predictive Analytics in Life and Health Insurance.
12. khadijat Aremu, B., & Peggy, O. O. (2025). Advanced machine learning-driven business analytics for real-time health risk stratification and cost prediction models.
13. Liza, I. A., Hossain, S. F., Saima, A. M., Akter, S., Akter, R., Al Amin, M., ... & Marzan, A. (2025). Heart Disease Risk Prediction Using Machine Learning: A Data-Driven Approach for Early Diagnosis and Prevention. British Journal of Nursing Studies, 5(1), 38-54.