Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2024, 12, 581-596
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

An Analysis and Prediction of Health Insurance Costs Using Machine Learning-Based Regressor Techniques

Gagan Kumar Patra[1], Chandrababu Kuraku[2], Siddharth Konkimalla[3], Venkata Nagesh Boddapati[4], Manikanth Sarisa[5], Mohit Surender Reddy[4]

[1]Tata Consultancy Services, Charlotte, NC, USA
[2]Mitaja Corportaion, Woodlawn, MD, USA
[3]Adobe, Seattle, WA, USA
[4]Microsoft, Charlotte, NC, USA
[5]Ally Financial Inc, Charlotte, NC, USA
Email: gagankpatra@outlook.com, chandrababu.kuraku@gmail.com, Siddharth.konkimalla@gmail.com, VenkataNagesh.boddapati@student.ctuonline.edu, Mk2703@outlook.com, Mohitreddy17@gmail.com

How to cite this paper: Patra, G.K., Kuraku, C., Konkimalla, S., Boddapati, V.N., Sarisa, M. and Reddy, M.S. (2024) An Analysis and Prediction of Health Insurance Costs Using Machine Learning-Based Regressor Techniques. *Journal of Data Analysis and Information Processing*, **12**, 581-596.
https://doi.org/10.4236/jdaip.2024.124031

Received: September 30, 2024
Accepted: November 9, 2024
Published: November 12, 2024

Abstract

One of the most significant annual expenses that a person has is their health insurance coverage. Health insurance accounts for one-third of GDP, and everyone needs medical treatment to varying degrees. Changes in medicine, pharmaceutical trends, and political factors are only a few of the many factors that cause annual fluctuations in healthcare costs. This paper describes how a system may analyse a person's medical history to display their insurance plans and make predictions about their health insurance premiums. The performance of four ML models—XGBoost, Lasso, KNN, and Ridge—is evaluated using R2-score and RMSE. The analysis of medical health insurance cost prediction using Lasso regression, Ridge regression, and K-Nearest Neighbours (KNN), and XGBoost (XGB) highlights notable differences in performance. KNN has the lowest R^2-score of 55.21 and an RMSE of 4431.1, indicating limited predictive ability. Ridge Regression improves on this by an R^2-score of 78.38 but has a higher RMSE of 4652.06. Lasso Regression slightly edges out Ridge with an R^2-score of 79.78, yet it suffers from an advanced RMSE of 5671.6. In contrast, XGBoost excels with the highest R^2-score of 86.81 and the lowermost RMSE of 4450.4, demonstrating superior predictive accuracy and making it the most effective model for this task. The best method for accurately predicting health insurance premiums was XGBoost Regression. The findings beneficial for policymakers, insurers, and healthcare providers as they can use this information to allocate resources more efficiently and enhance cost-effectiveness in

the healthcare industry.

Keywords

Medical Cost, Health Insurance, Cost Prediction, Medical Cost Personal Datasets, Machine Learning

1. Introduction

Healthcare costs are now the world's most critical problem. A significant instrument for enhancing accountability in healthcare is now people's healthcare expense predictions. Due to a lack of proper analysis, the vast quantities of patient, illness, and diagnosis data produced by the healthcare business are meaningless, despite the high cost of treating real people [1] [2]. The cost of losses brought on by a range of risks may be covered or reduced by a health insurance policy. Healthcare and insurance costs are influenced by a number of variables [3]. Numerous stakeholders and health authorities rely on prediction models for accurate cost estimates of individual treatment [4] [5]. Accurate cost estimations are useful in helping healthcare delivery organizations and health insurers make long-term plans and allocate scarce resources for care management in a more efficient manner [6] [7]. Additionally, by being aware of their expected future costs in advance, patients may choose insurance plans with suitable rates and deductibles. The creation of insurance policies is influenced by these factors [8].

Stakeholders and health authorities must properly estimate individual healthcare expenditures using prediction models due to the large number of factors that affect insurance or healthcare costs [9]. Having reliable cost projections is crucial for healthcare delivery organisations and health insurers when it comes to long-term planning and allocating scarce resources for care management [10]. The insurance industry may benefit from ML by using it to improve the efficiency of policy language [11]. Forecasting expensive, high-need patient spending is one area where ML algorithms excel in healthcare [12]-[14]. In this work, we compared and contrasted the performance of several regression models for forecasting health insurance premiums using supervised ML models.

1.1. Contribution and Aim of Paper

The contribution of this study lies in developing a robust approach to forecasting medical insurance costs employing advanced ML techniques. This research has made the following contributions:

- Predicting healthcare insurance costs employing a computational intelligence technique based on ML is an area that needs further investigation.
- Using a publicly available dataset, we compare the efficiency of the most extensively used ML methods for healthcare cost forecasting.
- Identifies crucial attributes influencing medical insurance costs, providing

insights into significant predictors.

- Uses evaluation metrics like R^2 and RMSE to assess model accuracy, offering a quantitative basis for model selection.
- Contributes to healthcare and insurance industries, potentially improving cost estimation and pricing strategies.

1.2. Structure of Paper

The remainder of the paper adheres to this format. Section II provides an overview of medical health insurance. Section III provides a detailed description of the method. In Section IV, the results, analysis, and conclusions are contrasted. The study's findings and possible directions for the future are outlined in depth in Section V.

2. Literature Review

In the literature review, various studies are discussed regarding the prediction of health insurance premiums and healthcare costs using machine learning algorithms that are provided in this section.

In Vijayalakshmi, Selvakumar and Panimalar (2023), For insurance cost prediction, the dataset with 24 characteristics was used, which included all relevant attributes. It is implemented utilising R programming's regression methods, including LR, DT, Lasso, Ridge, RF, Elastic Net, SVR, KNN, and Neural Network. Using an RS squared value of 0.9533, RFR demonstrated superior performance [15].

In Marinova and Todorova, (2023) evaluate the model's performance, we used R-Squared, RMSE, and training time as measurement measures. There is a 0.94 accuracy rate for the BMI feature model using the Bagged method, according to the testing data. Models built using the Bagged technique have a MSE of 0.06 for the attributes Smoker and Blood Pressure. It takes more time to train models that were constructed using the SVM compared to other methods [16].

In Thejeshwar *et al.* (2023), seeks to educate the general people about insurance in order to facilitate their acquisition at an accurate and reasonable cost. The models' predictions were enhanced by training on a dataset. By contrasting the estimated quantity with the actual data, the model was examined and verified. They evaluated the models' levels of accuracy. This approach works well in RFR because, compared to competing methods, it finds the performance measure with the greatest accuracy rate (87%) with much less processing time [8].

In Dutta *et al.* (2021), emphasizes the need of calculating the patient's portion of the healthcare expenditure. The best prediction analysis is achieved by using a variety of data employing regression techniques, including DT, RF, polynomial regression, and linear regression. The best method for accurately predicting health insurance premiums was RFR, which achieved a r2 score of 0.862533 when utilised as intended [17].

In Baro, Oliveira and De Souza Britto Junior (2022), investigate three data sets to extract characteristics, namely, medical specialty, the International Classification of

Diseases and an account of the event. Also, a dataset with 34,930 patient records totalling 38,524 medical occurrences was provided. In order to evaluate and establish a standard for this fresh dataset, we have used two popular ensemble techniques: RF and GB. When the models from the three feature sets that were examined were combined using GB, the best outcomes (AUC = 0.82) were obtained [18].

In Luo *et al.* (2021), information gathered between 2012 and 2014 from actual asthmatic patients in a large Chinese city was used to train prediction models, such as LR, RF, SVM, classification regression tree, and BPNN. According to the risk analysis of comorbidity on cost, the two main risks for asthmatic patients that impact treatment costs are respiratory diseases (36.38% in the adjusted odds ratio (95% Confidence Interval: 27.61%, 47.86%) and disorders of the circulatory system (23.83%; 95%CI: 15.95%, 35.22%) [19].

Table 1 below summarizes prior research on predicting health insurance premiums using ML and DL methodologies.

Table 1. Comparative study on health insurance cost prediction using machine and deep learning methods.

Author	Dataset	Methods	Performance	Limitation/Contribution
Vijayalakshmi, Selvakumar, and Panimalar [15]	24 features related to insurance cost	Linear Regression, DT, Lasso, Ridge, Random Forest, Elastic Net, SVR, KNN, Neural Network (R)	Best: Random Forest (R^2 = 0.9533); Metrics: MSE, RMSE, MAE, MAPE, R^2, Adj. R^2, EVS	• Accurate insurance cost prediction with minimal manual work. • May not generalize to other datasets or different insurance types.
Marinova and Todorova [16]	BMI, smoker, blood pressure	Bagged Algorithm, SVM	Best: Bagged Algorithm (Accuracy = 0.94, MSE = 0.06 R^2, RMSE, Training Time	• Improved model performance for specific health features. • SVM has high training time. Results specific to BMI, Smoker, and Blood Pressure features.
Thejeshwar *et al.* [8]	Variables related to public awareness and insurance demand	Linear Regression, SVM, Random Forest	Best: Random Forest (Accuracy = 87%); Metrics: Accuracy	• Helps in pricing insurance premiums accurately and efficiently. • Study focused more on awareness rather than broader insurance cost determinants.
Dutta *et al.* [17]	Health insurance cost prediction	DT, RF, Polynomial Regression, LR	Best: Random Forest (R^2 = 0.862533); Metrics: R^2, RMSE, MSE	• Random Forest excels in predicting health insurance costs. • Limited comparison with more advanced models like Neural Networks or Gradient Boosting.
Baro, Oliveira *et al.* [18]	38,524 records from 34,930 patients	RF, GB	Best: Gradient Boosting (AUC = 0.82); Metrics: AUC	• New dataset for hospitalization prediction. • Results limited to the provided dataset; further validation needed for other use cases.
Luo *et al.* [19]	Real-world data of asthmatic patients	Logistic Regression, Random Forest, SVM, CART, Neural Network	Best: Random Forest, Comorbidity Portfolio; Metrics: AUC, Sensitivity (46.89% improvement in AUC)	• Contribution: Advanced cost prediction and comorbidity management for asthmatic patients. • Focused on asthma patients, limiting generalizability to other conditions.

Research Gaps

Most studies depend on conventional algorithms; however, the research gap reveals a lack of investigation into advanced models of machine learning, such as ensemble approaches or deep learning. Many of these studies only look at small subsets of features or populations, which means their results are not relevant to the real world. Computational efficiency is still an issue, especially for complicated models like SVM, and there is an absence of real-time prediction systems that are responsive to changing data. There is a need for more study on the ethical implications and wider social effects of predictive models [20]-[22], particularly in the area of insurance cost estimation, so that we may better understand how to reduce bias and ensure equality in these systems.

3. Research Design

The research approach entails various steps and phases, as shown in the data flow diagram in **Figure 1**. The methodology begins with collecting a medical cost personal dataset by a KAGGLE repository, which comprises 1388 entries and seven features. Cleansing and preparing the dataset for analysis is the purpose of data preprocessing. This entails verifying missing values, which may result from incomplete data entries or equipment malfunctions, and removing duplicate entries to ensure data integrity. Feature extraction finds that important features like age, BMI, and smoking status are major factors affecting medical costs. To ensure consistency among features, min-max scaling is implemented, which normalizes the values to a range of 0 to 1. After the dataset has been preprocessed, it is often partitioned and separated into testing and training sets, often with a 70% training and 30% testing ratio. Multiple regression models, such as Ridge, Lasso, XGBoost, and KNN, predict medical insurance costs. The accuracy and reliability of each model in predicting insurance charges are assessed using metrics such as the R^2 score and RMSE. These evaluations facilitate the comparison and selection of the most effective model.

The following lists every step and stage of the data flow diagram in **Figure 1** are briefly explained below:

3.1. Data Collection

Personal datasets on medical costs are sourced from the KAGGLE repository. There is a grand total of 1388 items per column in the dataset, including seven characteristics with nonnull attributes.

3.2. Data Preprocessing

Data pre-processing is the process of preparing unprocessed data for use in a more organised dataset. Stated differently, even while data is gathered from a variety of sources, it is not acquired in a format that has been processed and is ready for analysis. Preprocessing is any alteration done to the dataset before supplying it to the algorithm. The preprocessing techniques listed below are explained in:

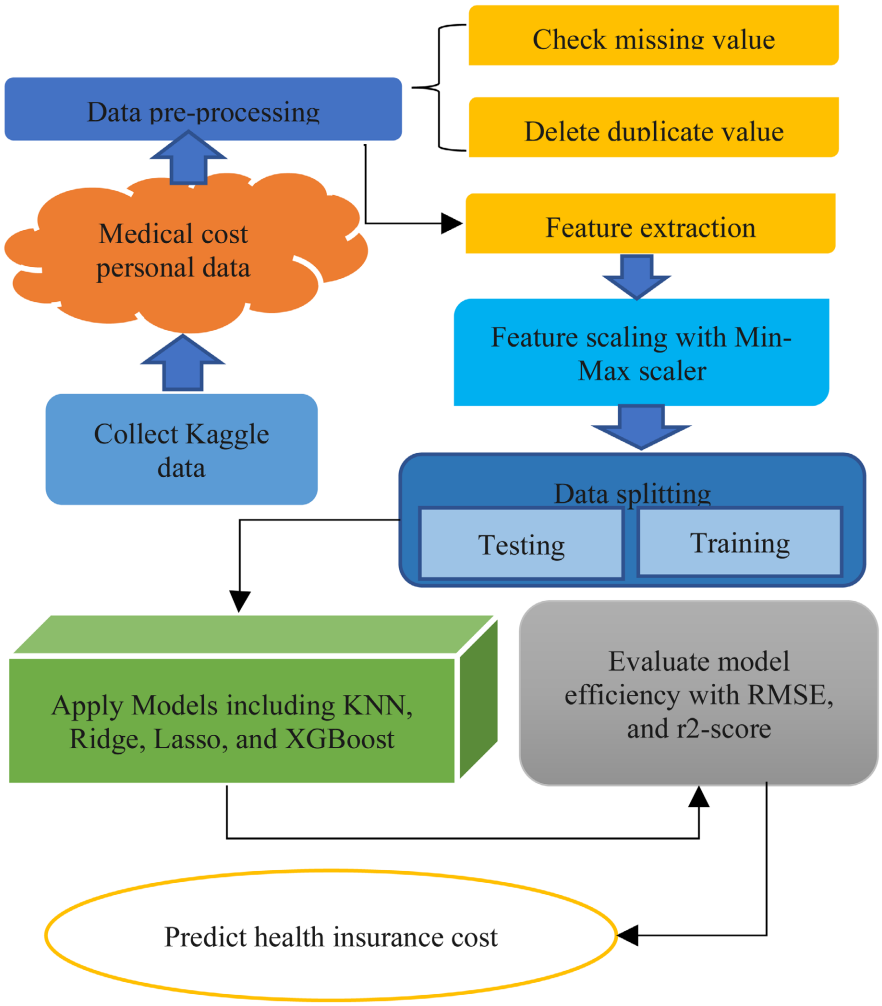

Figure 1. Health insurance cost data flow diagram.

- **Check missing value:** Equipment malfunctions, incomplete data entry, lost files, and numerous other factors can result in data loss.
- **Delete duplicate value:** Utilize the Remove Duplicates function to ensure that duplicate data is permanently eliminated.

3.3. Feature Extraction

Feature engineering in ML aims to increase the effectiveness of ML algorithms by using domain expertise to extract relevant features from unprocessed data. Factors like age, BMI, and smoking status are the most influential in the medical insurance cost dataset [23].

3.4. Feature Scaling with Min-Max Scaler

Min-max scaling is a specific method within feature scaling that rescales all values of a particular feature to fit within a predefined minimum and maximum range, typically 0 and 1. This process helps maintain consistency and comparability among different features. The following Equation (1) scale the dataset.

$$\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

3.5. Data Splitting

In machine learning modeling, dataset splitting is an essential step that aids in various stages, from training to evaluating the model. There are two separate subsets of the dataset: a testing set and a training set. 30% of the data is used for testing and 70% of the data is used for training throughout the experimental phase.

3.6. Applying Machine Learning Models

Choose various regression models, such as the KNN, Ridge, Lasso, and XGB models that are described below, to estimate health insurance costs:

1) K-nearest neighbor

The KNN approach finds the K data items or training patterns that are nearest to an input pattern, and then it chooses the model class with the highest number of models. The number of nearest neighbours that will be taken into account for predicting class labels in test data is known as the K value. K was selected by neighbouring K's class vote. Utilise the Euclidean Distance formula (2) to determine distances between neighbours [24]:

$$dxy = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

The prediction $\hat{y}_i$ for a new point x is given by formula (3):

$$\hat{y}_i = \frac{1}{k} \sum_{i=1}^{k} y_i \tag{3}$$

where y_i are the target values of the k nearest neighbors.

2) Ridge regression

Ridge regression is used when dealing with data that exhibit multi-collinearity. It is a method for fine-tuning the analysis of multi-collinear data. If the independent variables have a strong correlation, we may approximate the regression model's coefficient in this scenario. In order to avoid overfitting, ridge regression is a kind of linear regression that incorporates a L 2 regularisation term. The formula defines the cost function for Ridge regression (4):

$$j(\theta) = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \tag{4}$$

where:

y_i = is the true value.

$\hat{y}_i$ = is the predicted value.

λ = is the regularization parameter.

θ_j = are the coefficients of the model.

3) Lasso regression

Ridge regression is extremely similar to Lasso, also referred to as the Selection Operator and Least Absolute Shrinkage. In ML, Lasso regression is used to pick a

significant subset of variables. In general, Lasso regression prediction Lasso regression is similar to Ridge regression but employs L 1 regularization, promoting sparsity in the model coefficients. s are more accurate than those made by other models formulate as Equation (5).

$$j(\theta) = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{n} |\theta_j| \tag{5}$$

where $|\theta_j|$ represents the absolute value of the coefficients, encouraging some coefficients to be exactly zero.

4) XGBoost regressor

The XGBoost model [25] is an integrated model based on gradient boosting and tree-based. The fundamental structure is composed of numerous CARTs, which calculate the final prediction result by adding the desired value as well as each decision tree's predicted values from the past. Once each decision tree has finished training, a consensus is obtained. Pruning decision trees created during XGBoost model training is necessary to avoid overfitting, which occurs when each new tree is learnt using the previously trained tree. For the purpose of minimizing error, the XGBoost model uses the error that each tree produces as an input to train subsequent trees. By gradually reducing prediction error, this procedure forces the model's predicted outcome to be closer to the actual value. Assume the sample used for training is (x_i, y_i). XGB prediction model may thus be written as (63) [26]:

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i), f_k \in F \tag{6}$$

where $x \in R^m$, $y \in R$, x is the eigenvector, y is the sample label, and kth decision tree is represented by $f_k(x_i)$.

4. Results and Discussion

Datasets derived from freely accessible sources should have their analytical results detailed in this section. The machine learning model experents and their outcomes are presented in terms of RMSE, and R^2-score is also provided in this section.

4.1. Data Analysis

A kind of data analysis known as exploratory data analysis (EDA) seeks to discover broad patterns from the collected information. Notable data features and outliers are all part of these patterns. EDA is a critical initial stage in any data analysis. Histograms and boxplots are graphical methods for analysing the data's distribution. Some visualization graphs are given in below:

The heat map Dependencies of Medical Charges in **Figure 2** demonstrates sex, age, correlation between BMI, area, number of children, smoking status, and medical costs. Purple suggests strong positive connections, green suggests strong negative correlations, and beige suggests weak or no relationships. Smokers incur greater medical bills; therefore, dark purple indicates that.

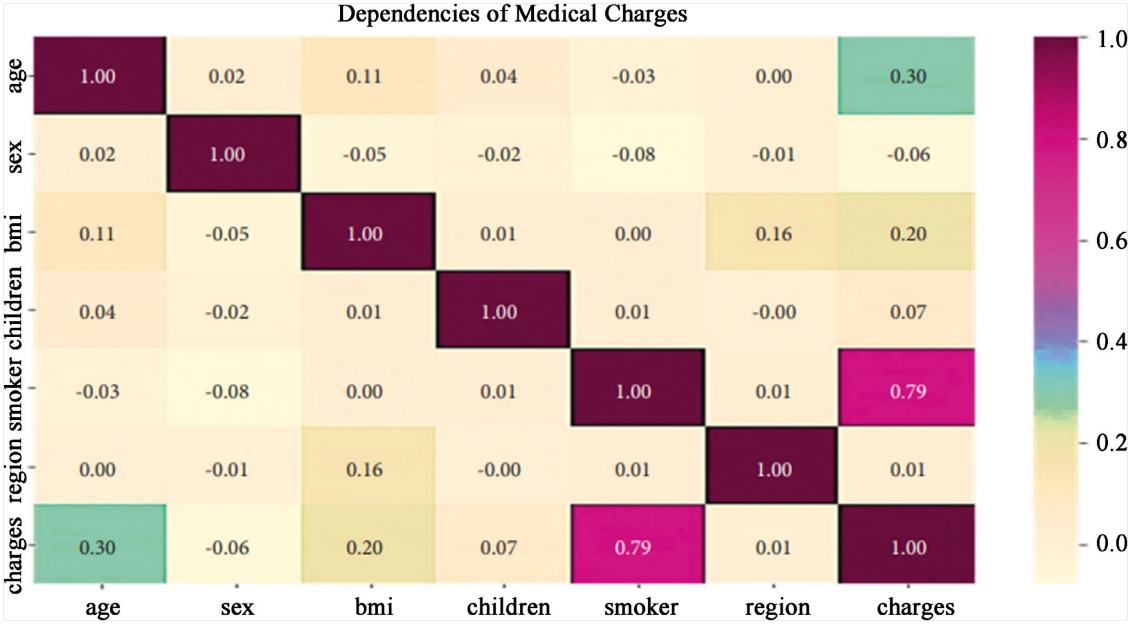

Figure 2. Elation matrix with a heat map.

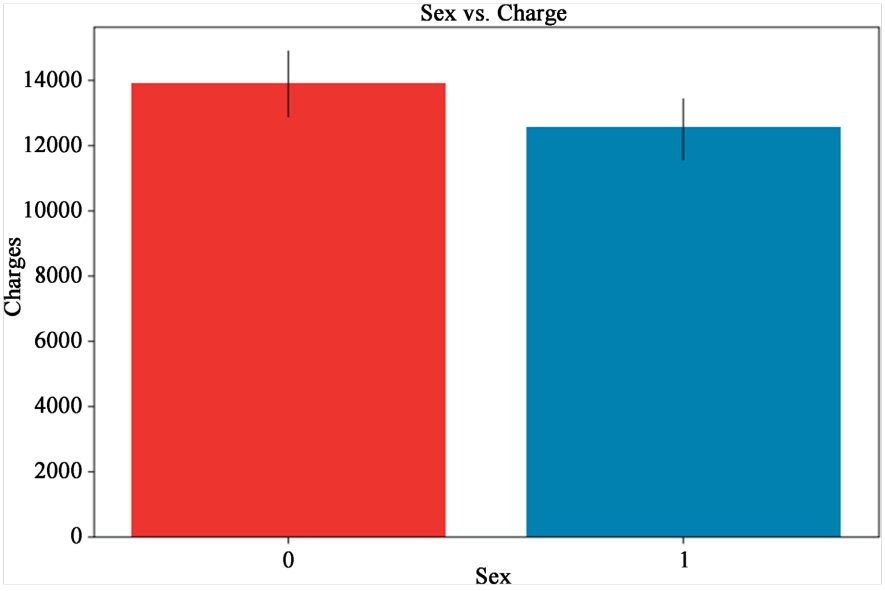

Figure 3. Sex vs. insurance charge features.

The red bar in Figure 3 indicates that typical insurance charges for Sex Category "0" are just over 12,000, while the blue bar for Sex Category "1" are just below 12,000; the error lines indicate small differences between the two groups.

Figure 4 shows the plot for age, where the horizontal axis represents age, with intervals of 10 years from 20 to 60 vertical axis represents the count, ranging from 0 to 200. The graph displays variable distribution across age groups, with the highest count around age 20 and subsequent counts below 200. It aids in demographic analysis and pattern understanding.

Figure 5 shows the distribution of BMI values as a histogram superimposed on

top of a line graph. The x-axis ranges from 15 to 50, and the y-axis ranges from 0 to 140. The tallest bar around a BMI value of 25 shows that this value has the highest frequency in the dataset. This visualization aids in understanding the spread of BMI values across the population.

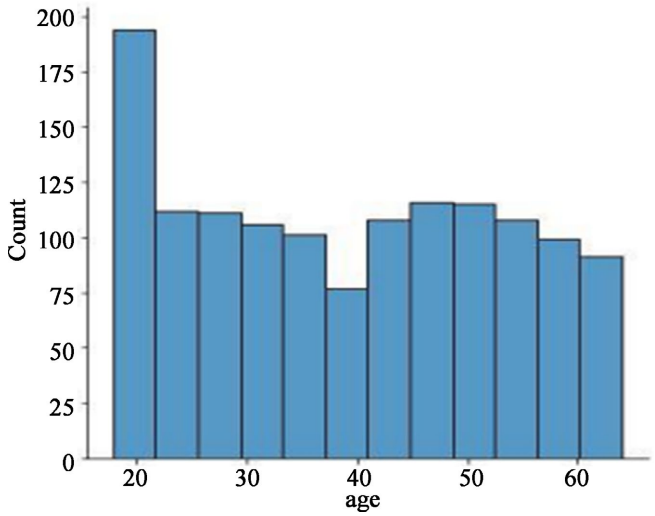

Figure 4. Plot for age.

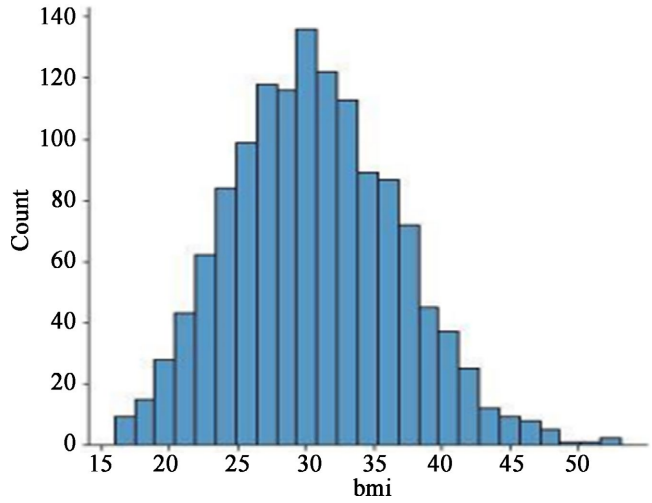

Figure 5. Plot for body mass index.

4.2. Performance Matrix

To evaluate the quality of ML models, use the evaluation error matrix. Metrics like as R^2-square, and Root Mean Squared Error need to be measured for us to be able to compare different algorithms.

1) RMSE

The RMSE is computed by calculating the MSE's square root. The RMSE formula is (7):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t' - y_t\right)^2}\ . \tag{7}$$

where

y_t' is a forecasted value,

y_t is an original value, and,

n is the sum of all the test set values.

2) R^2-score

T constant of purpose is also referred to as R-squared (R^2). It is a statistical metric. It determines the degree to which the data are in close proximity to the regression line that has been fitted. R^2 is calculated using formula (8).

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}} \tag{8}$$

where,

- The total of the squared deviations between the expected and real values is referred to as the SSR (sum of Squared Residuals).
- The sum of the squared differences between the computed and actual SST (Total Sum of Squares).
- Target variable values and the mean.

The R^2 metric, also referred to as the coefficient of determination, measures the percentage of the dependent variable's volatility that can be anticipated based on the independent variables in a regression model. It has a range of 0 to 1, where a model that predicts the dependent variable perfectly has an R^2 of 1 and a model that does not explain any variability in the data has an R^2 of 0.

4.3. Experiment Result

This section displays the machine learning model experiment outcomes. The following Table 2 shows the XGB model achieves 86.81% R-score.

Table 2. Results of XGB model on medical cost personal dataset

Measures	XGB
R^2-score	86.81
RMSE	4450.4

Figure 6 XGBoost Regression's scatter plot shows actual and forecasted costs related to one another. The red dots demonstrate that actual costs tend to increase with expected prices, suggesting that the XGBoost regression model accurately uses actual values to make cost predictions.

4.4. Comparative Analysis

In this section, present the outcomes of a machine learning method, including KNN [7], Ridge [27], Lasso [28], and XGB on the dataset. In this step, evaluate how well the model could predict. It presents the result in the procedure of bar graphs, Table 3, and figures.

Figure 7 shows the results of comparing the R^2-scores of the different models.

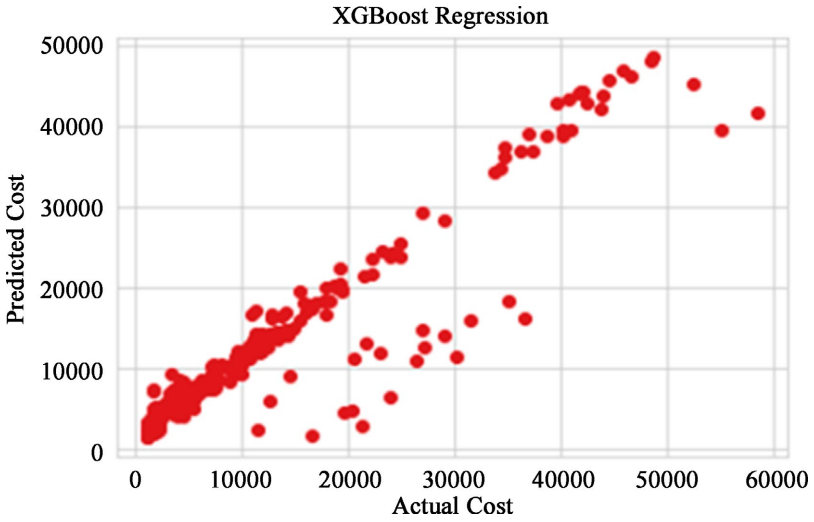

Figure 6. Predicted cost using XGBoost Regression.

Table 3. Comparative analysis for medical health insurance cost prediction.

Models	R^2-score	RMSE
KNN	55.21	4431.1
Ridge	78.38	4652.06
Lasso	79.78	5671.6
XGB	**86.81**	**4450.4**

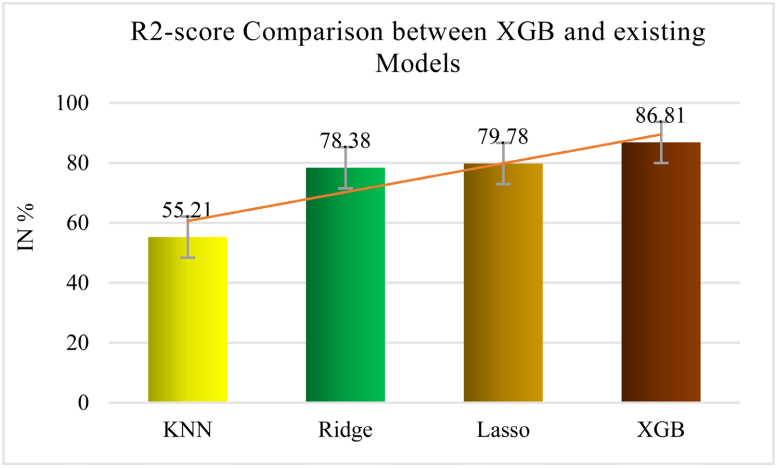

Figure 7. R^2-score comparison between models.

XGB fits the data the best, with a R^2-score of 86.81. In a regression model, the R^2-score, also known as the coefficient of determination, quantifies the percentage of the dependent variable's variation that is accounted for by the independent variables. Lasso follows closely with an R^2-score of 79.78, slightly outperforming Ridge at 78.38. KNN has the lowest R^2-score at 55.21, reflecting the lowest accuracy among the models in predicting the target variable. The overall XGB model outperforms other models.

Figure 8 RMSE comparison reveals that KNN has an RMSE of 4431.1, slightly better than XGB with an RMSE of 4450.4, indicating better accuracy in predictions. Ridge Regression follows with an RMSE of 4652.06, and Lasso has the highest RMSE at 5671.6, suggesting it has the greatest prediction error among the models.

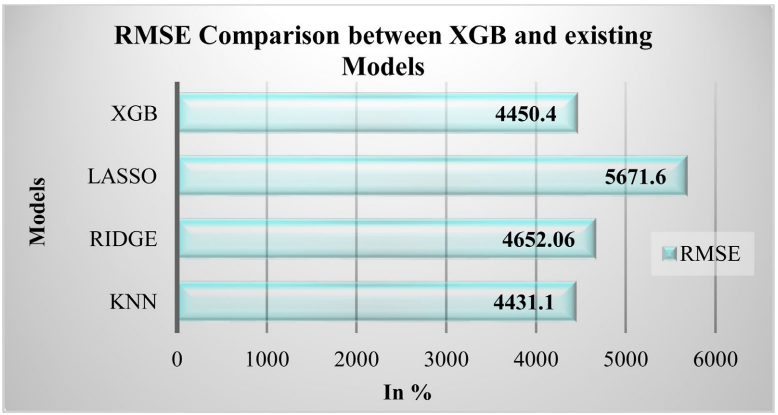

Figure 8. RMSE comparison between models.

5. Conclusion and Future Study

It's critical for insurance firms and customers to predict health insurance premiums. This paper explores the use of regression approaches to forecast health insurance premiums. A study uses a medical cost personal dataset of insurance premiums and related variables that have the biggest effects on insurance rates. According to the study, geography and gender had very little impact on insurance costs, with age and BMI being the primary determinants. To create predictive models, the study used various regression approaches, including KNN, XGBBoost Regression, Lasso, and Ridge regression. Among the models, XGBoost showed the best performance with an R^2-score86.81 and an RMSE4450.4, outperforming the other models according to accuracy and predictive power. The comparative analysis demonstrates the superior fit and prediction accuracy of XGBoost, making it the most suitable model for this dataset. The XGBoost model performed well, although the research had several drawbacks. Because of the limited sample size, it may be harder to extrapolate the findings to bigger populations. To improve prediction accuracy, future research might concentrate on growing the dataset to include more entries and more characteristics, such lifestyle or medical history. In further study, we will adjust the settings of ML and DL techniques on various datasets linked to medical health using metaheuristic and nature-inspired algorithms.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Kumar, V.V., Sahoo, A., Balasubramanian, S.K. and Gholston, S. (2024) Mitigating

Healthcare Supply Chain Challenges under Disaster Conditions: A Holistic AI-Based Analysis of Social Media Data. *International Journal of Production Research*. https://doi.org/10.1080/00207543.2024.2316884

[2] Sommers, B.D. (2020) Health Insurance Coverage: What Comes after the Aca? *Health Affairs*, **39**, 502-508. https://doi.org/10.1377/hlthaff.2019.01416

[3] Milovic, B. (2012) Prediction and Decision Making in Health Care Using Data Mining. *International Journal of Public Health Science* (*IJPHS*), **1**, 126-136. https://doi.org/10.11591/ijphs.v1i2.1380

[4] Mathur, S. and Gupta, S. (2023) Classification and Detection of Automated Facial Mask to COVID-19 Based on Deep CNN Model. 2023 *IEEE 7th Conference on Information and Communication Technology* (*CICT*), Jabalpur, 15-17 December 2023, 1-6. https://doi.org/10.1109/cict59886.2023.10455699

[5] Morid, M.A., Kawamoto, K., Ault, T., Dorius, J. and Abdelrahman, S. (2017) Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. *AMIA Annual Symposium Proceedings*, **2017**, 1312-1321.

[6] Basile, L.J., Carbonara, N., Pellegrino, R. and Panniello, U. (2023) Business Intelligence in the Healthcare Industry: The Utilization of a Data-Driven Approach to Support Clinical Decision Making. *Technovation*, **120**, Article ID: 102482. https://doi.org/10.1016/j.technovation.2022.102482

[7] Ramya, D., Manigandan, S.K. and Deepa, J. (2022) Health Insurance Cost Prediction Using Machine Learning Algorithms. 2022 *International Conference on Edge Computing and Applications* (*ICECAA*), Tamilnadu, 13-15 October 2022, 1381-1384. https://doi.org/10.1109/icecaa55415.2022.9936153

[8] Thejeshwar, T., Sai Harsha, T., Vamsi Krishna, V. and Kaladevi, R. (2023) Medical Insurance Cost Analysis and Prediction Using Machine Learning. 2023 *International Conference on Innovative Data Communication Technologies and Application* (*ICIDCA*), Uttarakhand, 14-16 March 2023, 113-117. https://doi.org/10.1109/icidca56705.2023.10100057

[9] Duijvestijn, M., de Wit, G.A., van Gils, P.F. and Wendel-Vos, G.C.W. (2023) Impact of Physical Activity on Healthcare Costs: A Systematic Review. *BMC Health Services Research*, **23**, Article No. 572. https://doi.org/10.1186/s12913-023-09556-8

[10] Gabriel, J. (2024) A Machine Learning-Based Web Application for Heart Disease Prediction. *Intelligent Control and Automation*, **15**, 9-27. https://doi.org/10.4236/ica.2024.151002

[11] Nori, N. (2024) Machine Learning Based Virtual Screening for Biodegradable Polyesters. *Journal of Materials Science and Chemical Engineering*, **12**, 1-11. https://doi.org/10.4236/msce.2024.128001

[12] Anumandla, S.K.R., Yarlagadda, V.K., Vennapusa, S.C.R. and Kothapalli, K.R.V. (2020) Unveiling the Influence of Artificial Intelligence on Resource Management and Sustainable Development: A Comprehensive Investigation. *Technology & Management Review*, **5**, 45-65.

[13] Yang, C., Delcher, C., Shenkman, E. and Ranka, S. (2018) Machine Learning Approaches for Predicting High Cost High Need Patient Expenditures in Health Care. *BioMedical Engineering OnLine*, **17**, Article No. 131. https://doi.org/10.1186/s12938-018-0568-3

[14] Iorliam, I.B., Ikyo, B.A., Iorliam, A., Okube, E.O., Kwaghtyo, K.D. and Shehu, Y.I. (2021) Application of Machine Learning Techniques for Okra Shelf Life Prediction. *Journal of Data Analysis and Information Processing*, **9**, 136-150. https://doi.org/10.4236/jdaip.2021.93009

[15] Vijayalakshmi, V., Selvakumar, A. and Panimalar, K. (2023) Implementation of Medical Insurance Price Prediction System Using Regression Algorithms. 2023 5*th International Conference on Smart Systems and Inventive Technology* (*ICSSIT*), Tirunelveli, 23-25 January 2023, 1529-1534. https://doi.org/10.1109/icssit55814.2023.10060926

[16] Marinova, G. and Todorova, M. (2023) Regression Analysis for Predicting Health Insurance. 2023 4*th International Conference on Communications*, *Information*, *Electronic and Energy Systems* (*CIEES*), Plovdiv, 23-25 November 2023, 1-4. https://doi.org/10.1109/ciees58940.2023.10378755

[17] Dutta, K., Chandra, S., Gourisaria, M.K. and GM, H. (2021) A Data Mining Based Target Regression-Oriented Approach to Modelling of Health Insurance Claims. 2021 5*th International Conference on Computing Methodologies and Communication* (*ICCMC*), Erode, 8-10 April 2021, 1168-1175. https://doi.org/10.1109/iccmc51019.2021.9418038

[18] Baro, E.F., Oliveira, L.S. and de Souza Britto Junior, A. (2022) Predicting Hospitalization from Health Insurance Data. 2022 *IEEE International Conference on Systems*, *Man*, *and Cybernetics* (*SMC*), Prague, 9-12 October 2022, 2790-2795. https://doi.org/10.1109/smc53654.2022.9945601

[19] Luo, L., Yu, X., Yong, Z., Li, C. and Gu, Y. (2021) Design Comorbidity Portfolios to Improve Treatment Cost Prediction of Asthma Using Machine Learning. *IEEE Journal of Biomedical and Health Informatics*, **25**, 2237-2247. https://doi.org/10.1109/jbhi.2020.3034092

[20] Farhud, D.D. and Zokaei, S. (2021) Ethical Issues of Artificial Intelligence in Medicine and Healthcare. *Iranian Journal of Public Health*, **50**, i-v. https://doi.org/10.18502/ijph.v50i11.7600

[21] Mondal, H. and Mondal, S. (2024) Ethical and Social Issues Related to AI in Healthcare. *Methods in Microbiology*, **55**, 247-281. https://doi.org/10.1016/bs.mim.2024.05.009

[22] Nill, A., Laczniak, G. and Thistle, P. (2017) The Use of Genetic Testing Information in the Insurance Industry: An Ethical and Societal Analysis of Public Policy Options. *Journal of Business Ethics*, **156**, 105-121. https://doi.org/10.1007/s10551-017-3554-y

[23] Rohilla, V., Chakraborty, S. and Kumar, R. (2022) Deep Learning Based Feature Extraction and a Bidirectional Hybrid Optimized Model for Location Based Advertising. *Multimedia Tools and Applications*, **81**, 16067-16095. https://doi.org/10.1007/s11042-022-12457-3

[24] Okfalisa, Gazalba, I., Mustakim, and Reza, N.G.I. (2017) Comparative Analysis of K-Nearest Neighbor and Modified K-Nearest Neighbor Algorithm for Data Classification. 2017 2*nd International conferences on Information Technology*, *Information Systems and Electrical Engineering* (*ICITISEE*), Yogyakarta, 1-2 November 2017, 294-298. https://doi.org/10.1109/icitisee.2017.8285514

[25] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. P*roceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[26] Anmol, Aggarwal, S. and Jahan Badhon, A. (2022) Medical Insurance Cost Prediction Using Machine Learning Algorithms. In: Maurya, S., Peddoju, S.K., Ahmad, B. and Chihi, I., Eds., *Cyber Technologies and Emerging Sciences*, Springer, 271-281. https://doi.org/10.1007/978-981-19-2538-2_27

[27] Ul Hassan, C.A., Iqbal, J., Hussain, S., AlSalman, H., Mosleh, M.A.A. and Sajid Ullah,

S. (2021) A Computational Intelligence Approach for Predicting Medical Insurance Cost. *Mathematical Problems in Engineering*, **2021**, Article ID: 1162553. https://doi.org/10.1155/2021/1162553

[28] Christobel, Y.A. and Subramanian, S. (2022) An Empirical Study of Machine Learning Regression Models to Predict Health Insurance Cost. *Webology*, **19**, 1677-1685.