International Journal of Advances in Engineering and Computer Science

ISSN: 3121-6382

Volume 1, Issue 2, December, 2025
(An Academians Publishers of Research & Academic Resources)

Research Article

Received: **10-11-2025** Accepted: **29-11-2025** Published: **28-12-2025**

Health Insurance Cost Prediction Using Machine Learning Based Regression

Ramesh Prasad Bhatta
Central Department of CSIT
Far Western University, Mahendranagar
PhD Scholar MJPRU, Bareilly, India

Abstract: Accurate prediction of health insurance costs is essential for effective financial planning and policy formulation in the healthcare sector. With the growing availability of healthcare-related data, machine learning methods have become increasingly useful for estimating medical expenses based on individual characteristics. This study observes the performance of four regression-based machine learning models KNN, Ridge Regression, Lasso Regression, and Extreme Gradient Boosting (XGBoost) using the medical most personal on dataset obtained from Kaggle. The model performance was evaluated using the R^2-score and Root Mean Square Error (RMSE). The results show that XGBoost achieved the highest prediction accuracy with an R^2-score of 0.88 and an RMSE of 3217.53. In comparison, KNN, Ridge, and Lasso achieved R^2-scores of 0.76, 0.74, and 0.72, respectively, with higher RMSE values. The results specify XGBoost is more effective in catching complex relationships within insurance cost data, foremost to improved prediction accuracy. These findings highlight XGBoost as the most effective model for accurate prediction of health insurance costs. The implication was to develop the best model for accurately predicting medical costs which can considerably benefit insurance firms in risk assessment and premium computation.

Keywords: *Machine Learning, Regression, RMSE, XGBoost, KNN.*

Corresponding Author: Ramesh Prasad Bhatta

1. Introduction

Healthcare costs are currently the world's most urgent problem. Healthcare cost predictions are emerging as a crucial instrument for enhancing responsibility in the healthcare sector. The huge amounts of patient, illness, and diagnosis data produced by the healthcare sector are useless because of incorrect analysis, though the high cost of treating real patients. (Sommers, 2020; Kumar et al., 2024). The expense of losses brought on by a range of risks may be covered or reduced by a health insurance policy. The cost of healthcare and insurance is influenced by many factors (Milovic, 2012). Many stakeholders and health authorities rely on prediction models for accurate cost estimates of each therapy (Mathur & Gupta, 2023; Morid et al., 2017).

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 (Basile et al., 2023; Ramya et al., 2022) 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 (Thejeshwar et al., Healthcare delivery organizations and health insurers can more effectively allocate limited resources for care management and create long-term plans with the aid of accurate cost estimates (Basile et al., 2023; Ramya et al., 2022). Additionally, patients can select insurance plans with appropriate rates and deductibles if they are informed of their anticipated future costs beforehand. These factors impact the development of insurance plans (Thejeshwar et al., 2023).

Stakeholders and health authorities must utilize prediction models to effectively forecast individual healthcare expenses because so many factors affect insurance or healthcare prices (Duijvestijn et al., 2023). Health insurers and healthcare delivery organizations require precise cost estimates for long-term planning and allocating scarce resources for care management (Gabriel, 2024).

The cost of health care is rising every day. As the number of new diseases that infect people rises, forecasting health costs becomes essential. People also recognize the significance of spending on health care. Every aspect of life is impacted by the field of machine learning. Machine learning models are also used in many health-related applications in the healthcare industry. We conducted a predicate analysis on medical health insurance costs in this study. In this study, I created a model to forecast a person's health insurance costs according to their gender. The Kaggle dataset consists of 1338 rows of data with the following attributes: age, gender, smoker, BMI, children, region, and insurance charges. Medical records and costs reported by health insurance companies are included in the data. The study was implemented using the Python programming language.

The contribution of this study is to:

1. Use a publicly available dataset to assess the efficacy of the most widely used machine learning techniques for healthcare cost prediction.
2. Determine significant variables influencing health insurance costs and provide details about significant predictors.
3. Assess model accuracy using metrics like R2 and RMSE to provide a quantitative basis for model selection.

1. Literature review

Machine learning is capable of managing numerous variables and identifying connections that were previously difficult to see. This aids insurers in better understanding the factors that influence medical expenses, allowing them to more precisely set rates and effectively manage risks. This improves everyone's access to healthcare while also assisting insurance companies in maintaining their financial stability.

Using photos of the damaged vehicle as input, this system generates pertinent information, including repair prices to determine the insurance claim amount and damage locations. As a result, the current analysis concentrated on estimating repair costs rather than the anticipated auto insurance claim (Singh, Ayyar, Pavan, Gosain, & Shah, 2019).

The literature assessment several studies that apply machine learning algorithms to predict health insurance premiums and healthcare expenditures. Vijayalakshmi, Selvakumar, and Panimalar (2023) utilized a dataset comprising 24 relevant features to estimate insurance costs. Their study implemented multiple regression techniques using the R programming language, including Linear Regression (LR), Decision Tree (DT), Lasso, Ridge, Random Forest (RF), Elastic Net, Support Vector Regression (SVR), K-Nearest Neighbors (KNN), and Neural Networks. Among these methods, the Random Forest Regression (RFR) model demonstrated superior performance, achieving an R-squared value of 0.9533.

Marinova and Todorova (2023) evaluated model performance using metrics such as R-squared, Root Mean Square Error (RMSE), and training time. Experimental results indicated that the Bagging-based model incorporating the BMI feature achieved an accuracy of 0.94 on the test dataset. Additionally, models developed using the Bagging approach recorded a Mean Squared Error (MSE) of 0.06 for blood pressure and smoking-related features. However, models based on Support Vector Machines (SVM) required comparatively longer training times.

Thejeshwar et al. (2023) aimed to enhance public awareness of insurance pricing by enabling fair and accurate premium estimation. Their study demonstrated that training machine learning models on relevant datasets significantly improved prediction accuracy. The proposed models were analyzed and validated by comparing predicted values with actual data, confirming the effectiveness of the approach.

Random Forest Regression (RFR) was selected because, when compared with competing methods, it achieved the highest accuracy rate of 87% while requiring significantly less processing time (Thejeshwar et al., 2023).

Dutta et al. (2021) emphasized the importance of estimating the patient’s share of healthcare expenditures. Their study applied multiple regression-based techniques, including Decision Tree (DT), Random Forest (RF), Polynomial Regression, and Linear Regression, to improve prediction accuracy. Among these methods, RFR demonstrated the best performance in predicting health insurance premiums, attaining an R^2 score of 0.862533.

Baro, Oliveira, and De Souza Britto Junior (2022) analyzed three datasets to extract key features related to medical specialty, the International Classification of Diseases (ICD), and event records. The study utilized a large dataset consisting of 34,930 patient records and 38,524 medical events. Two ensemble learning methods, Random Forest (RF) and Gradient Boosting (GB), were employed to evaluate and benchmark the dataset. The highest predictive performance (AUC = 0.82) was achieved when Gradient Boosting was used to combine models derived from all three feature sets.

Luo et al. (2021) developed several prediction models, including Linear Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Classification and Regression Tree (CART), and Backpropagation Neural Network (BPNN), using real-world data collected from asthma patients in a major Chinese city between 2012 and 2014. Risk analysis revealed that circulatory system disorders (23.83%; 95% CI: 15.95–35.22) and respiratory diseases (adjusted odds ratio: 36.38%; 95% CI: 27.61–47.86) were the most significant comorbidities influencing treatment costs.

Billa and Nagpal (2024) conducted a comparative analysis of different machine learning algorithms for medical insurance price prediction. Their study highlighted the effectiveness of ensemble methods over traditional regression models, suggesting that ensemble approaches provide better predictive performance.

The research gap shows that while most studies rely on conventional algorithms, there is a lack of research into more sophisticated machine learning models, such as ensemble approaches or deep learning. Many of these studies only look at small subgroups of persons or characteristics, therefore their conclusions are not applicable to the real world. Computational efficiency is still an issue, especially for sophisticated models like SVM, and there are yet no real-time prediction systems that can adjust to changing data. Further research on the ethical aspects and wider societal ramifications of predictive models is needed, particularly in the area of insurance cost prediction, in order to better understand how to reduce prejudice and ensure equity in these systems (Patra et al., 2024).

2. Methodology

The various steps and phases that comprise the research technique are depicted in the data flow diagram in Figure 1. The first step in the process is to collect the medical cost personal dataset from a KAGGLE repository, which consists of 1388 entries and seven features. Cleaning and preparing the dataset for analysis is the aim of data preparation. This entails eliminating duplicate entries and verifying missing values, which may result from insufficient data entry or device malfunctions, in order to preserve data integrity. Through feature extraction, important factors such as age, BMI, and smoking status are found to have a considerable impact on medical expenses. Min-max scaling ensures consistency between features by normalizing the data to a range of 0 to 1.

The dataset is usually split into training and testing sets after preprocessing, with a typical ratio of 30% for testing and 70% for training. Multiple regression models, such as Ridge, Lasso, XGBoost, and KNN, are used to forecast medical insurance costs. The accuracy and reliability of each model in predicting insurance costs are assessed using metrics such as the R2 score and RMSE.

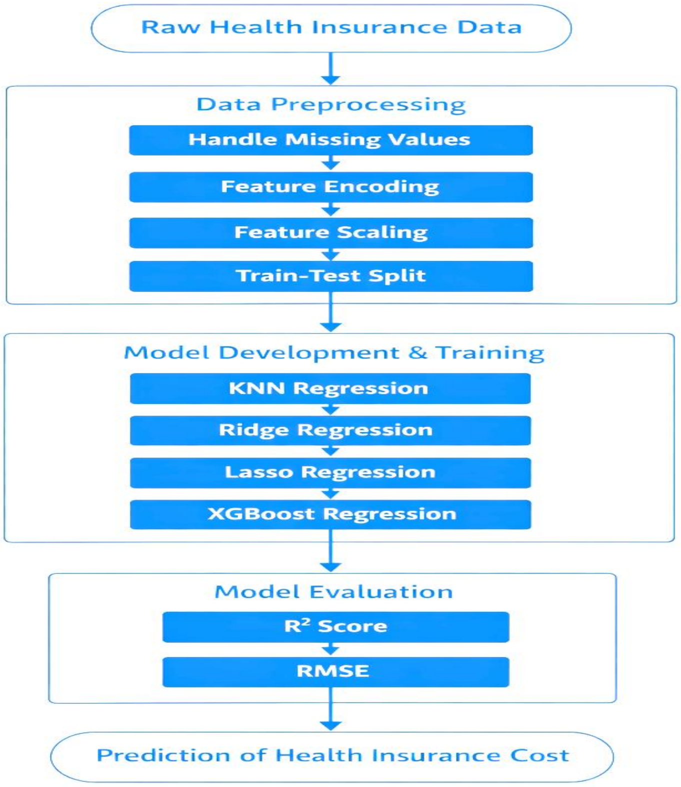

Figure 1. Health insurance cost data flow diagram

3.1. Data Collection

The KAGGLE repository is the source of individual medical cost datasets. The collection has seven characteristics with non-null attributes, totaling 1388 items per column. The dataset consists of following attributes.

age: age of the insurances primary beneficiary
sex: insurances' primary beneficiary gender (female or male)
bmi: body mass index,
children: number of children covered by health insurance
smoker: whether the primary beneficiary is a current smoker or not
region: the primary beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
charges: individual medical costs billed by health insurance

3.2. Data Preprocessing

Data pre-processing is the process of preparing unstructured data for use in a more organized dataset. To put it another way, preprocessing refers to any modifications performed to the dataset before it is fed into the algorithm, even though data is gathered from a variety of sources.

- Handling the missing value: Data loss can be caused by a variety of circumstances, including missing files, inadequate data entry, and equipment problems.
- Remove duplicate value: To guarantee that duplicate data is permanently removed, use the

3.3. Feature Extraction

By employing domain knowledge to identify appropriate features from raw data, feature engineering in machine learning seeks to improve the efficiency of ML algorithms. The most significant variables in the medical insurance cost dataset are age, BMI, and smoking status.

3.4. Feature Scaling with Min-Max Scaler

A particular feature scaling technique called "min-max scaling" rescales all of a feature's values to fall between a predetermined minimum and maximum range, usually 0 and 1. This procedure aids in preserving uniformity and comparability between various aspects. The subsequent The dataset is scaled by equation.

$$\tilde{x} = (x - \min(x))/(\max(x) - \min(x))\ldots\ldots\ldots 1$$

3.5 Data Splitting

Splitting datasets is a crucial step in machine learning modeling that helps with training and model evaluation. The dataset is divided into two distinct subsets: a training set and a testing set. Testing uses thirty percent of the data and during the experimental phase, 70% of the data is used for training.

3.6. Machine Learning Models

To estimate health insurance costs, select from a variety of regression models, including the KNN, Ridge, Lasso, and XGB models that are explained below:

3.6.1 K-Nearest Neighbor

After identifying the K data items or training patterns that are most similar to an input pattern, the KNN technique chooses the model class with the most models. The number of nearest neighbors that will be taken into account for class labels in test data are predicted using the K value. K's neighboring class cast their votes to select K. Use the Euclidean Distance formula (2) to determine the distances between neighbors:

$$d_x y = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \ \ldots\ldots\ldots 2$$

The prediction ˆi y for a new point x is given by formula

$$\bar{y}_i = 1/k \sum_{i=1}^{k} y_i \ \ldots\ldots 3$$

3.6.2 Ridge regression is used when dealing with multi-collinearity data.

It is a method for enhancing the analysis of multi-collinear data. In this case, we can estimate the regression model's coefficient if there is a substantial correlation between the independent variables. To avoid overfitting, a *LL* 2 regularization term is added in ridge regression, a kind of linear regression. Ridge regression's cost function is defined by

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

where:

y_i = is the true value.y_i = is the predicted value.

λ = is the regularization, parameter

j $_\theta$= are the coefficients of the model.

3.6.3 Lasso regression

Ridge regression and Lasso, commonly referred to as the Selection Operator and Least Absolute Shrinkage, are quite comparable. In machine learning, Lasso regression is used to identify a significant subset of variables. In general, Lasso regression prediction Lasso regression, like Ridge regression, uses *LL* 1 regularization to encourage sparsity in the model coefficients. are more accurate than those generated by alternative models, as indicated by equation.

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

where θ_j stands for the coefficients' absolute values, urging some to be precisely zero.

.

3.6.4 XGBoost Regressor

The XGBoost model combines gradient boosting with tree-based boosting. The fundamental structure consists of several CARTs, each of which adds the necessary value to produce the final prediction outcome. the values that the decision tree had previously projected. After each decision tree has finished training, an agreement is obtained. Pruning decision trees created during XGBoost model training is necessary to avoid overfitting, which occurs when each new tree is learned using the previously trained tree. In an attempt to reduce error, the XGBoost model uses each tree's error as an input to train subsequent trees. By gradually reducing prediction error, this procedure promotes the model's expected outcome to be closer to the actual value.XGB prediction model may thus be written as

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

where $x \in Rm$, $y \in R$, x is the eigenvector, y is the sample label, and kth decision

tree is represented by $k_i f\,(x_i)$.

3. Results and Discussion

This section should include a detailed analysis of datasets that were obtained from publicly available sources. This section also includes the R2-score and presents the machine learning model expertness and results in terms of RMSE.

4.1. Data Analysis

The data analysis that examines for wide patterns in the collected data is called exploratory data analysis (EDA). Outliers and prominent data features are examples of these patterns. EDA is a crucial initial step in any analysis of data. The visualization graphs are shown below. The association between BMI, area, number of children, smoking status, sex, age, and medical expenses is depicted in Figure 2's Dependencies of Medical Charges heat map. Green indicates strong negative correlations, purple indicates strong positive links, and beige indicates weak or nonexistent associations. Dark purple indicates that medical costs are higher for smokers.

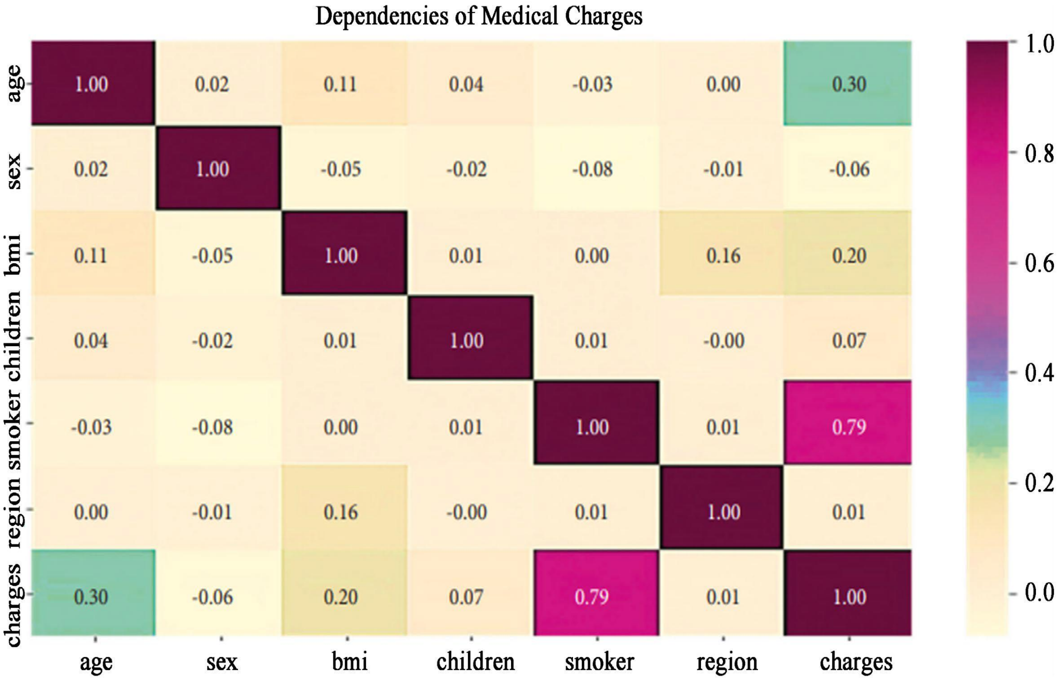

Figure 2: Dependencies of Medical charges on heat map.

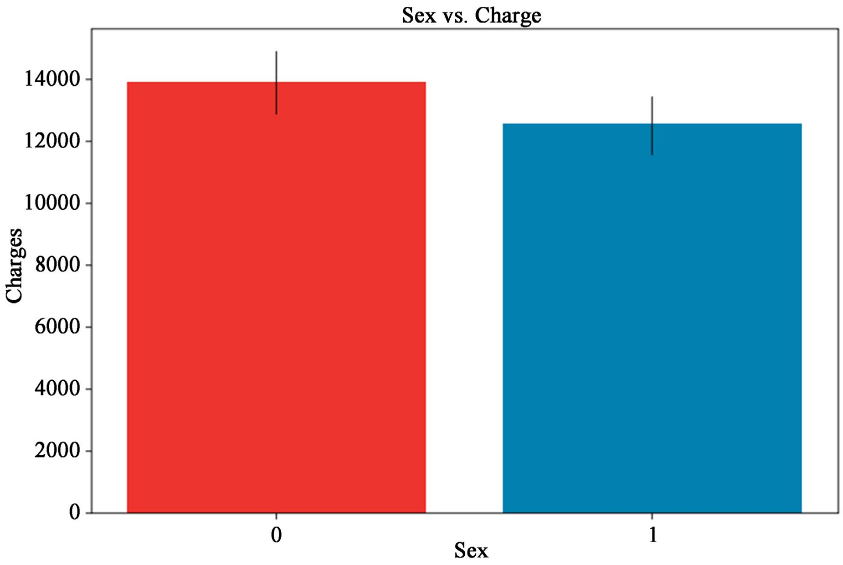

Figure 3: Sex and Insurance Cost features.

In Figure 3, the blue bar for Sex Category "1" is just below 12,000 and the red bar for Sex Category "0" is little over 12,000; the error lines show minor variations between the two categories. Figure 4 displays the age plot, with the vertical axis representing the count, which ranges from 0 to 200, and the horizontal axis representing age, with intervals of 10 years between 20 and 60. The graph displays a variable distribution among age groups, with subsequent counts decreasing below 200 and the greatest count occurring around age 20. It makes demographic research and pattern recognition easier.

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

top. The x and y axes have ranges of 15 to 50 and 0 to 140, respectively. The tallest bar encircling a BMI of 25 indicates that this value has the highest frequency in the sample. Understanding how BMI levels differ across the population is made simpler by this image.

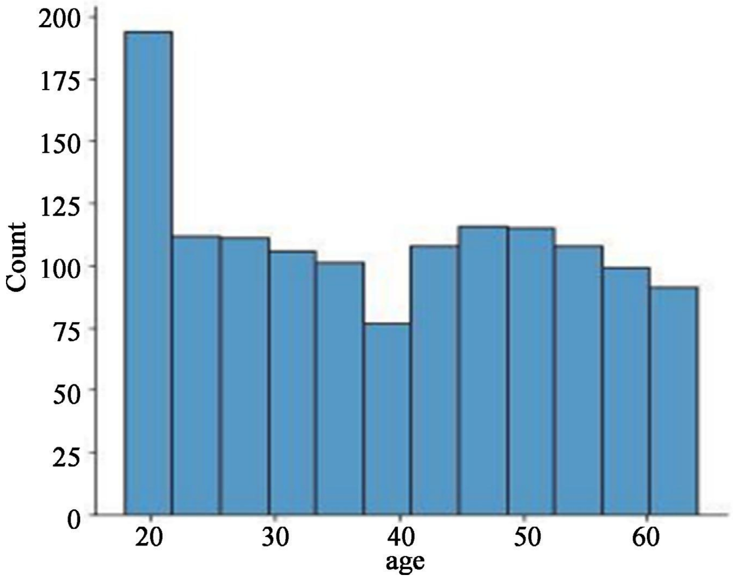

Figure 4: Plot for age.

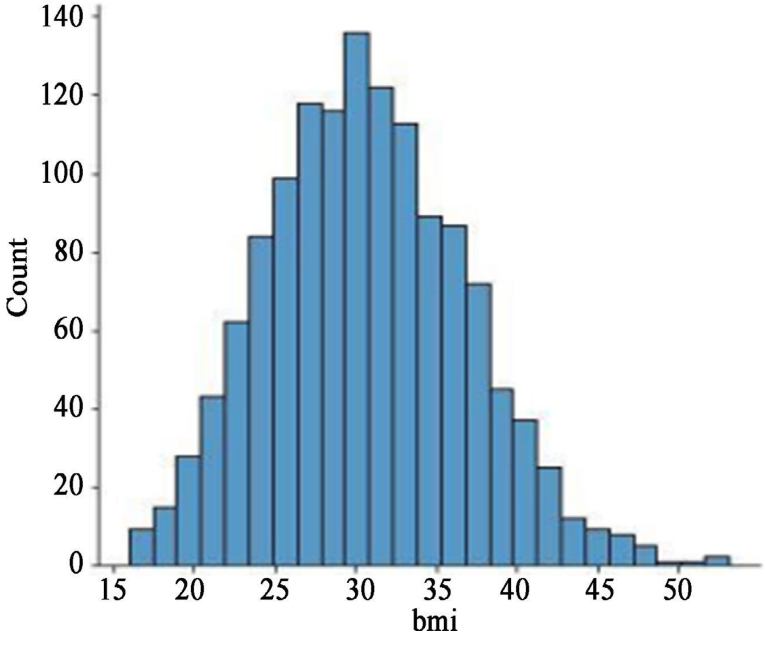

Figure 5 : Histogram for BMI

Performance Matrix

Use the evaluation error matrix to assess the quality of machine learning models. To compare various algorithms, we must assess metrics like R2-square and Root Mean Squared Error.

1.RMSE

The RMSE is computed by calculating the MSE's square root. The RMSE formula is

$$RMSE = \sqrt{1/n \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$$

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

R-squared (R^2) is another name for the T constant of purpose. It is a measure of statistics. It establishes how near the data are to the fitted regression line. A formula is used to determine R^2.

$$R^2 = 1 - \frac{Unexplained\, Variation}{Total\, Variation}$$

where, The SSR (sum of squared residuals) is the sum of the squared variances between the expected and actual values. The total squared difference between the calculated and real SST (Total Sum of Squares). The mean and values of the target variable. The percentage of the dependent variable's volatility that can be predicted using the independent variables in a regression model is measured by the R^2 metric, also known as the coefficient of determination. A model that accurately predicts the dependent variable has an R^2 of 1, while a model that fails to account for any variability in the data has an R^2 of 0.

4.2 Results

The results of the machine learning model experiment are shown in this section. XGB has an RMSE of 3217.53and an R2-score of 87.94.

The scatter plot of XGBoost Regression in Figure 6 illustrates the relationship between actual and projected expenses. The red dots show that real expenses often rise. Given projected prices, indicating that the XGBoost regression model correctly forecasts costs using actual values.

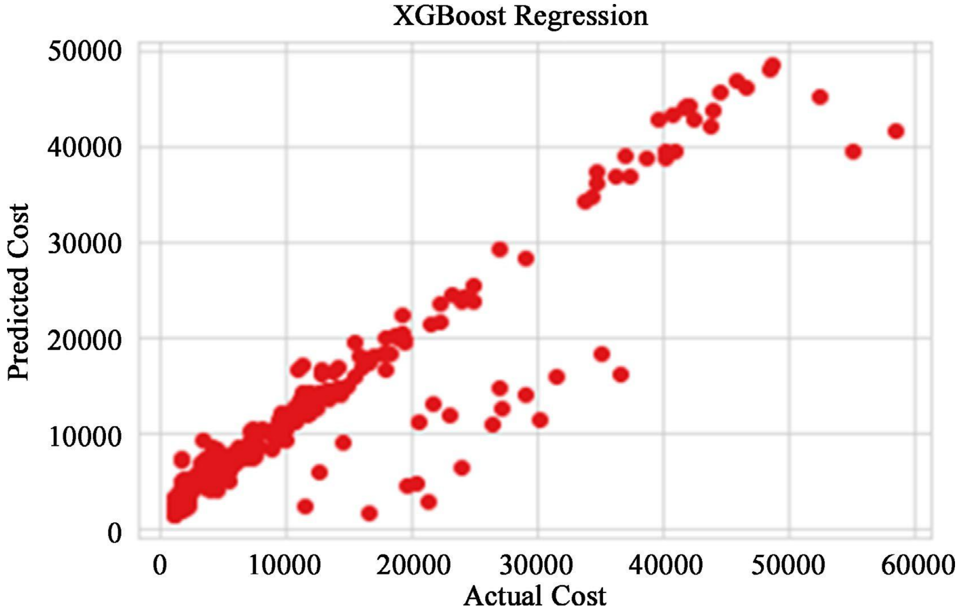

Figure 6: RMSE comparison between models.

Comparative Evaluation

This section shows the outcomes of a machine learning method on the dataset, such as KNN, Ridge, Lasso, and XGB. At this point, evaluate the predictive power of the model. Bar graphs, Table 1, and figures are used to show the results. Figure 7 shows the outcomes of comparing the R2-scores of the different models.

Table 1: Comparative analysis for medical health insurance costs.

.Models	R^2-score	RMSE
KNN	76.43	5321.84
Ridge	74.29	5586.21
Lasso	72.16	5894.33
XGB	87.94	3217.53

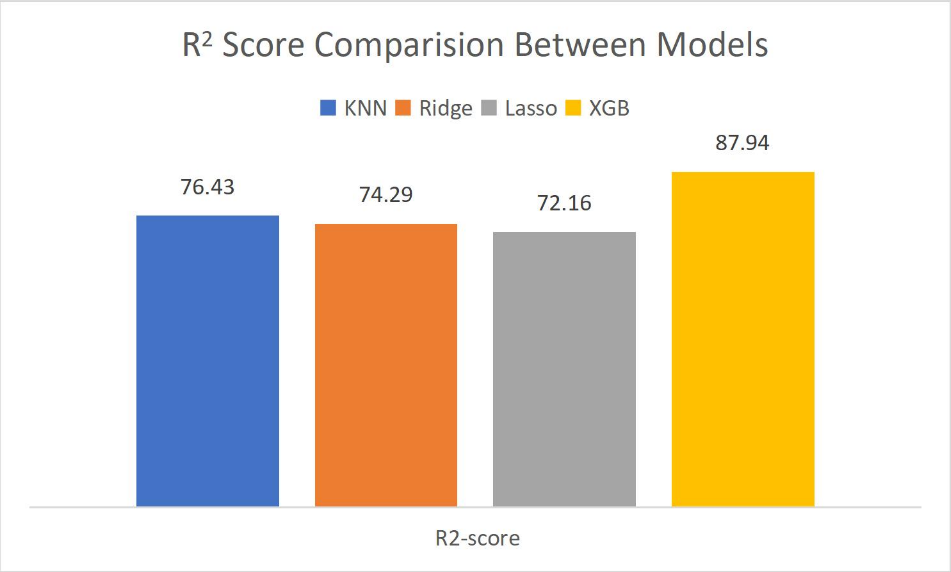

Figure 7: R^2-score comparison between models.

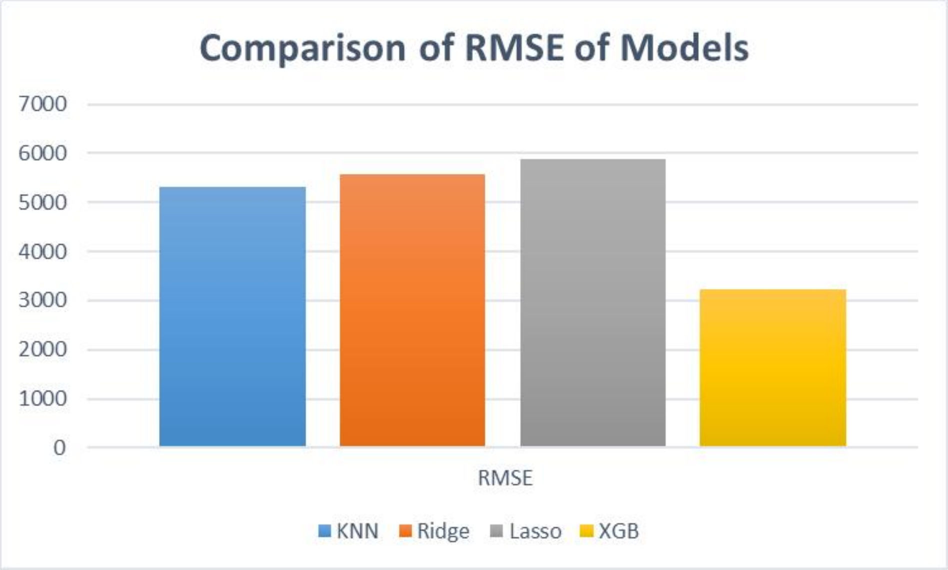

Figure 8: RMSE comparison between models.

4. Discussion

The comparative results show clear performance differences among the evaluated models. XGBoost (XGB) reveals the best predictive capability, achieving the highest R² score (87.94) and the lowest RMSE (3217.53), indicating superior accuracy and minimal prediction error. This suggests that XGB is more effective in capturing complex, non-linear relationships in the dataset.

Among the traditional regression approaches, KNN performs better than Ridge and Lasso, with a higher R² score (76.43) and lower RMSE (5321.84), reflecting a moderate ability to model data variability. Ridge regression slightly outperforms Lasso, as evidenced by its higher R² (74.29 vs. 72.16) and lower RMSE (5586.21 vs. 5894.33), indicating that L2 regularization is more suitable than L1 regularization for this dataset.

5. Conclusion

The forecasting health insurance premiums is essential for both customers and insurance companies. This study aspect at using regression methods to forecast health insurance amounts. A personal dataset of insurance premiums for medical expenses and related variables that have the most effects on insurance prices is used in a study. According to the study, geography and gender had comparatively little impact on insurance costs, whereas age and BMI were the primary determinants. To create prediction models, the study used several regression methods, including KNN, XGBoost Regression, Lasso, and Ridge regression.

The XGBoost performed the best among the models in terms of accuracy and predictive power. with an R2-score of 87.94 and an RMSE of 3217.53. Because of the small sample size, inferring the findings to wider populations may be more challenging. To improve prediction accuracy, future research could concentrate on growing the dataset to include more entries and other variables, such as lifestyle or medical history. The results showed age, BMI, and smoking status as some of the most significant determinants of insurance costs, which is consistent with medical knowledge of health risk factors. In the future study multimodal dataset will be used and Ml with DL model will be utilized for better prediction.

References

Milovic, B. (2012). Prediction and decision making in health care using data mining. *International Journal of Public Health Science,* *1*(2), 126–136. https://doi.org/10.11591/ijphs.v1i2.1380

Morid, M. A., Kawamoto, K., Ault, T., Dorius, J., & Abdelrahman, S. (2017). Supervised learning methods for predicting healthcare costs: Systematic literature review and empirical evaluation. *AMIA Annual Symposium Proceedings, 2017*, 1312–1321.

Yang, C., Delcher, C., Shenkman, E., & Ranka, S. (2018). Machine learning approaches for predicting high cost high need patient expenditures in health care. *BioMedical Engineering OnLine, 17*, 131. https://doi.org/10.1186/s12938-018-0568-3

Anumandla, S. K. R., Yarlagadda, V. K., Vennapusa, S. C. R., & Kothapalli, K. R. V. (2020). Unveiling the influence of artificial intelligence on resource management and sustainable development: A comprehensive investigation. *Technology & Management Review, 5*(1), 45–65.

Sommers, B. D. (2020). Health insurance coverage: What comes after the ACA? *Health Affairs,* *39*(3), 502–508. https://doi.org/10.1377/hlthaff.2019.01416

Dutta, K., Chandra, S., Gourisaria, M. K., & GM, H. (2021). A data mining based target regression-oriented approach to modelling of health insurance claims. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1168–1175). IEEE. https://doi.org/10.1109/iccmc51019.2021.9418038

Iorliam, I. B., Ikyo, B. A., Iorliam, A., Okube, E. O., Kwaghtyo, K. D., & Shehu, Y. I.

(2021). Application of machine learning techniques for okra shelf life prediction. *Journal of Data Analysis and Information Processing, 9*(3), 136–150. https://doi.org/10.4236/jdaip.2021.93009

Luo, L., Yu, X., Yong, Z., Li, C., & Gu, Y. (2021). Design comorbidity portfolios to improve treatment cost prediction of asthma using machine learning. *IEEE Journal of Biomedical and Health Informatics, 25*(6), 2237–2247. https://doi.org/10.1109/jbhi.2020.3034092

Baro, E. F., Oliveira, L. S., & de Souza Britto Junior, A. (2022). Predicting hospitalization from health insurance data. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2790–2795). IEEE. https://doi.org/10.1109/smc53654.2022.9945601

Ramya, D., Manigandan, S. K., & Deepa, J. (2022). Health insurance cost prediction using machine learning algorithms. In *2022 International Conference on Edge Computing and Applications (ICECAA)* (pp. 1381–1384). IEEE. https://doi.org/10.1109/icecaa55415.2022.9936153

Basile, L. J., Carbonara, N., Pellegrino, R., & Panniello, U. (2023). Business intelligence in the healthcare industry: The utilization of a data-driven approach to support clinical decision making. *Technovation, 120*, 102482. https://doi.org/10.1016/j.technovation.2022.102482

Duijvestijn, M., de Wit, G. A., van Gils, P. F., & Wendel-Vos, G. C. W. (2023). Impact of physical activity on healthcare costs: A systematic review. *BMC Health Services Research, 23*, 572. https://doi.org/10.1186/s12913-023-09556-8

Marinova, G., & Todorova, M. (2023). Regression analysis for predicting health insurance. In *2023 4th International Conference on Communications, Information, Electronic and Energy Systems (CIEES)* (pp. 1–4). IEEE. https://doi.org/10.1109/ciees58940.2023.10378755

Mathur, S., & Gupta, S. (2023). Classification and detection of automated facial mask to COVID-19 based on deep CNN model. In *2023 IEEE 7th Conference on Information and Communication Technology (CICT)* (pp. 1–6). IEEE. https://doi.org/10.1109/cict59886.2023.10455699

Thejeshwar, T., Sai Harsha, T., Vamsi Krishna, V., & Kaladevi, R. (2023). Medical insurance cost analysis and prediction using machine learning. In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)* (pp. 113–117). IEEE. https://doi.org/10.1109/icidca56705.2023.10100057

Vijayalakshmi, V., Selvakumar, A., & Panimalar, K. (2023). Implementation of medical insurance price prediction system using regression algorithms. In *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1529–1534). IEEE. https://doi.org/10.1109/icssit55814.2023.10060926

Gabriel, J. (2024). A machine learning-based web application for heart disease prediction. *Intelligent Control and Automation, 15*(1), 9–27. https://doi.org/10.4236/ica.2024.151002

Kumar, V. V., Sahoo, A., Balasubramanian, S. K., & 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*. Advance online publication. https://doi.org/10.1080/00207543.2024.2316884

Nori, N. (2024). Machine learning based virtual screening for biodegradable polyesters. *Journal of Materials Science and Chemical Engineering, 12*(8), 1–11. https://doi.org/10.4236/msce.2024.128001

Patra, G. K., Kuraku, C., Konkimalla, S., Boddapati, V. N., Sarisa, M., & 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*(4), 581–596

Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. (2022). *International Journal of Environmental Research and Public Health, 19*(7898). https://doi.org/10.3390/ijerph19137898

ul Hassan, C. A., Iqbal, J., Hussain, S., AlSalman, H., Mosleh, M. A. A., & Sajid Ullah, S. (2021). A computational intelligence approach for predicting medical insurance cost. *Mathematical Problems in Engineering, 2021*, 1162553. https://doi.org/10.1155/2021/1162553

Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks, 7*(2), 70. https://doi.org/10.3390/risks7020070

Singh, R., Ayyar, M. P., Pavan, T. S., Gosain, S., & Shah, R. R. (2019, September). Automating car insurance claims using deep learning techniques. In *2019 IEEE Fifth International Conference on Multimedia Big Data(BigMM)* (pp. 199–207). IEEE. https://doi.org/10.1109/BigMM.2019.00039