

A Predictive Analytics Approach for Health Insurance Cost Modeling in Pharmaceutical Healthcare

Ms. Anamika Soni

Assistant Professor, Department of Computer Sciences and Applications

Mandsaur University, Mandsaur

anamika.soni@meu.edu.in

Corresponding Author: * Anamika Soni (anamika.soni@meu.edu.in)

Received: 5 September 2026 | Revised: 10 September 2026 | Accepted: 28 September 2026

Abstract—Health insurance policies provide financial support to individuals by covering medical expenses and reducing the financial burden associated with diseases, injuries, and healthcare treatments. Several demographic and health-related factors influence the cost of health insurance, making accurate cost prediction an important task for insurance providers and policyholders. Early prediction of health insurance expenses can assist in determining suitable coverage plans and improving financial decision-making. In this study, a supervised machine learning-based approach is proposed for Health Insurance Cost Prediction using the Medical Insurance Cost Dataset. Data preprocessing techniques, including missing value handling, duplicate removal, outlier detection, one-hot encoding, and feature standardization, were applied to enhance data quality and model performance. Four machine learning algorithms, namely Random Forest (RF), AdaBoost, LightGBM, and Extreme Gradient Boosting (XGBoost), were implemented and compared. Experimental results reveal that the proposed XGBoost model achieved superior performance with an R-Square (R^2) value of 99% and a Mean Absolute Error (MAE) of 721.05, outperforming the other models. The findings demonstrate the effectiveness of XGBoost in capturing complex relationships among insurance-related factors and generating highly accurate cost predictions. Therefore, the proposed model can serve as a reliable decision-support tool for health insurance pricing, risk assessment, and healthcare financial planning.

Keywords—Health Insurance Premiums, Predictive Modelling, Machine Learning Algorithms, Statistical Methods, Healthcare Cost Prediction, Medicare Costs, Cost Control Mechanisms.

I. INTRODUCTION

Healthcare plays a vital role in improving the quality of life and ensuring the well-being of individuals. However, the rising cost of medical treatments, pharmaceutical services, and healthcare facilities has increased the financial burden on patients and insurance providers [1][2][3]. The insurance sector therefore serves as an essential mechanism for reducing financial risks associated with medical expenses by providing financial protection and access to quality medical services [4][5]. Accurate estimation of insurance premiums is crucial

for both insurers and policyholders, as it supports fair pricing, effective risk assessment, and informed financial planning. However, determining medical insurance costs is a challenging task because they are influenced by numerous factors, including demographic characteristics, lifestyle habits, medical conditions, and coverage-related attributes [6].

Modern medical systems involve complex interactions among patients, healthcare providers, pharmaceutical organizations, and insurance companies [7][8][9]. Rising medical expenditures and increasing demand for quality care have highlighted the need for reliable cost estimation methods [10]. Accurate prediction of insurance expenses can help organizations optimize premium structures, allocate resources efficiently, and improve long-term financial sustainability [11]. At the same time, individuals can benefit from better awareness of the factors influencing their insurance costs and future expenditures [12].

Traditional actuarial approaches often rely on predefined statistical assumptions and limited variables, which may not adequately capture the complex relationships among multiple risk factors [13][14]. As a result, prediction accuracy may be reduced when dealing with large and diverse datasets. Predictive analytics offers a data-driven solution by identifying hidden patterns and relationships within historical records, enabling more precise estimation of insurance costs and supporting evidence-based decision-making [15][16].

Recent advances in Machine Learning (ML) and Deep Learning (DL) have significantly transformed predictive analytics by enabling intelligent data-driven decision-making. These techniques can automatically learn complex patterns from large datasets, capture nonlinear relationships among variables, and generate highly accurate predictions [17][18]. ML and DL models have been widely adopted in insurance analytics for risk assessment, premium estimation, fraud detection, and customer profiling [19]. Furthermore, their ability to continuously improve with new data makes them highly suitable for dynamic and evolving insurance environments. Consequently, these technologies provide a reliable foundation for developing advanced insurance cost

prediction systems with improved accuracy and efficiency [20][21].

A. Motivation and Contribution

The increasing cost of healthcare services, medical treatments, and pharmaceutical expenses has created a significant financial burden on individuals and insurance providers. Accurate prediction of health insurance costs is essential for effective premium estimation, risk assessment, and financial planning. However, the complex relationships among demographic, lifestyle, and health-related factors make cost estimation a challenging task. This motivates the use of advanced machine learning techniques to develop reliable prediction models that can improve estimation accuracy and support better decision-making in the healthcare insurance sector. This research offers several key contributions as listed below:

- Developed a machine learning-based framework for Health Insurance Cost Modeling in Pharmaceutical Healthcare using the Medical Insurance Cost Dataset.
- Conducted exploratory data analysis and feature importance analysis to identify the key factors influencing health insurance costs, such as age, BMI, and smoking status.
- Implemented and compared multiple machine learning models, including Random Forest (RF), AdaBoost, LightGBM, and XGBoost, for insurance cost prediction.
- Proposed an XGBoost-based prediction model that achieved superior performance and provided highly accurate health insurance cost predictions compared to other machine learning models.
- Demonstrated the effectiveness of machine learning techniques in accurately estimating healthcare insurance expenses and supporting pharmaceutical healthcare financial planning.
- Provided a reliable decision-support approach that can assist insurance providers in premium estimation, risk assessment, and cost management.

Accurate health insurance cost prediction is essential for effective healthcare financial planning, premium estimation, and risk assessment in the pharmaceutical healthcare sector. This study is justified by the need to address the limitations of traditional estimation methods in capturing complex relationships among demographic and health-related factors. The novelty of the proposed work is the creation of an XGBoost-based prediction framework that includes all necessary steps: data preprocessing, feature importance analysis and comparison of various machine learning models. By identifying the most influential factors affecting insurance costs and providing highly accurate predictions, the proposed approach offers a reliable and interpretable decision-support tool for health insurance cost modeling.

B. Organization of the Paper

The rest of the paper is structured as follows: The concept of medical insurance and related works are discussed in Section II. Section III describes the methodology and approaches taken. Section IV contains the experimental outcomes and the results of the determinant analysis. Finally, Section V summarizes the findings and provides recommendations for future work.

II. LITERATURE REVIEW

A comprehensive review of recent studies on Health Insurance Cost Prediction was conducted to support the development of this research. Table I summarizes the proposed models, datasets used, key findings, and limitations identified in previous studies, providing valuable insights for the proposed approach.

Mishra et al. (2026) The study uses five machine learning models: Linear Regression, Decision Tree, K-Nearest Neighbors (KNN), Random Forest, and XGBoost, on data from 1,338 individuals including age, BMI, smoking habits, location, and air and water quality. Linear Regression achieved an R^2 score of 0.7296, while Random Forest performed best with an R^2 score of 0.9654. The results show that smoking status and BMI are major factors affecting insurance costs, demonstrating that ML can improve the accuracy and fairness of insurance premium prediction.[22].

Mohan et al. (2025) estimate health insurance claim costs using machine learning based on health risk factors such as age, sex, BMI, smoking status, and city. A Kaggle health insurance dataset is used, involving data preprocessing, feature engineering, and model selection, with performance evaluated using R^2 and Mean Absolute Error. Among all models, Random Forest achieved the highest accuracy of 96.7%, highlighting the potential of machine learning to improve insurance cost forecasting, pricing strategies, and financial management [23].

R and Arunraj (2025) proposes a metaheuristic optimization algorithm for feature selection and hyperparameter tuning, and compares its performance with Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Bayesian Optimization. Models were evaluated using MAE, MSE, RMSE, and R^2 score. The proposed algorithm, when applied to Random Forest and Gradient Boosting, achieved the best performance with an R^2 score of 92.5% and an RMSE of 2510.23, improving accuracy by 4.3%. The results show that ensemble-based models outperform regression models in insurance cost prediction, with smoking status and BMI identified as key predictors. The study highlights the effectiveness of metaheuristic optimization and suggests combining it with deep learning for further improvements [24].

Sharma and Jeya (2024) examines a variety of factors such as age, sex, BMI, smoking and number of children plus to predict health insurance costs. What makes it unique is the use of linear regression to unpack the strong relationship between these factors and insurance costs. carefully partitioned data set, with 70% for training, and achieved an impressive 81.3% accuracy. This ethically driven study enhances understanding of the complexities of post-COVID-19 health insurance spending. The use of linear regression stands out as a powerful rational forecasting tool, and represents a remarkable advancement in healthcare cost forecasting. The importance of the research is to provide insight into the changing landscape of post-pandemic health insurance costs, increase knowledge, and contribute to decision-making in health care policy [25].

Vuddanti et al. (2024) introduce the experienced numerous health problems during the COVID-19 pandemic. The creation and assessment of predictive models for the purpose of calculating medical costs are the subjects of investigation. This study has utilized the 4968-row USA medical insurance

dataset from Kaggle. The dataset's features allow a summary of an individual's personal and medical circumstances. Regression analysis and gradient boosting are the two machine learning methods utilized to build predictive models. The analysis reflected the link between features and costs. After training the model, a 94% accuracy rate has been obtained [26].

T et al. (2023) public awareness about health insurance and promote fair pricing by considering factors such as people's attitudes, decision-making processes, and demand- and supply-side perceptions. The effectiveness of three regression models—Linear Regression, Support Vector Regression, and Random Forest Regressor—was evaluated using a training dataset. The predicted and actual values were compared to validate the models. Among all methods,

Random Forest Regressor achieved the highest accuracy of 87% with lower computation time, making it the most suitable approach [27].

Panda et al. (2022) explores predictive modeling to enhance insurance business operations and services using machine learning (ML). A real-time insurance cost prediction system, ML Health Insurance Prediction System (MLHIPS), is developed to help insurance companies rapidly estimate premiums and reduce health expenditure. The system employs Ridge Regression, Lasso Regression, Simple Linear Regression, Multiple Linear Regression, and Polynomial Regression models. Among these, Polynomial Regression achieved the best performance with an RMSE of 5100.53 and an R^2 value of 0.80 [28].

TABLE I. RECENT STUDIES ON HEALTH INSURANCE COST PREDICTION USING MACHINE LEARNING TECHNIQUES

Author	Proposed Work	Dataset	Results	Limitations & Future Work
Mishra et al. (2026)	Compared Linear Regression, Decision Tree, KNN, Random Forest, and XGBoost for health insurance cost prediction.	Medical Insurance Dataset (1,338 records) containing age, BMI, smoking status, location, and environmental factors.	Random Forest achieved the highest $R^2 = 0.9654$, outperforming other models. Smoking status and BMI were identified as key predictors.	Limited dataset size may affect generalization. Future work can explore larger datasets and advanced ensemble techniques.
Mohan et al. (2025)	Developed a machine learning framework for predicting health insurance claim costs using demographic and health-related attributes.	Kaggle Health Insurance Dataset.	Random Forest delivered the best performance with 96.7% accuracy.	Study focused on traditional ML models; future research may incorporate deep learning and additional health indicators.
R and Arunraj (2025)	Proposed a metaheuristic optimization algorithm for feature selection and hyperparameter tuning in insurance cost prediction.	Health Insurance Cost Dataset.	Optimized Random Forest and Gradient Boosting achieved $R^2 = 92.5\%$ and RMSE = 2510.23.	Computational complexity of optimization methods is high. Future work can integrate deep learning with metaheuristic optimization.
Sharma and Jeya (2024)	Applied Linear Regression to analyze the relationship between demographic factors and insurance costs.	Health Insurance Dataset with age, sex, BMI, smoking status, and number of children.	Achieved 81.3% prediction accuracy using a 70:30 train-test split.	Linear Regression may not capture complex nonlinear relationships. Future studies can employ advanced ensemble models.
Vuddanti et al. (2024)	Developed predictive models using Regression Analysis and Gradient Boosting for medical cost estimation.	USA Medical Insurance Dataset (4,968 records) from Kaggle.	Obtained 94% prediction accuracy.	Limited evaluation of advanced optimization methods. Future work can investigate hybrid machine learning approaches.
T et al. (2023)	Evaluated Linear Regression, Support Vector Regression, and Random Forest Regressor for insurance price prediction.	Health Insurance Dataset.	Random Forest achieved the highest accuracy of 87% with lower computational time.	Limited feature diversity and model exploration. Future work may include ensemble boosting algorithms.
Panda et al. (2022)	Proposed the ML Health Insurance Prediction System (MLHIPS) using multiple regression techniques.	Health Insurance Cost Dataset.	Polynomial Regression achieved the best performance with RMSE = 5100.53 and $R^2 = 0.80$.	Prediction accuracy remains moderate. Future work should investigate advanced ensemble and boosting methods.

Research Gaps: The existing studies on health insurance cost prediction have primarily focused on traditional machine learning models such as Linear Regression, Random Forest, KNN, and Support Vector Regression, achieving moderate to high prediction accuracy. Although some studies incorporated optimization techniques and ensemble methods, limitations remain in terms of prediction accuracy, model generalization, feature diversity, and handling complex nonlinear relationships among insurance-related factors. Moreover, research was often based on relatively small-scale data sets and did not include much advanced analysis on advanced boosting algorithms. Therefore, there is a need for a more robust and accurate predictive framework that leverages advanced ensemble learning techniques, such as XGBoost, to improve health insurance cost estimation and support better decision-making in pharmaceutical healthcare applications.

III. RESEARCH METHODOLOGY

The proposed methodology uses the Medical Insurance Cost Dataset to predict health insurance charges. Data preprocessing techniques such as missing value handling, duplicate removal, outlier detection, one-hot encoding, and standardization were applied to improve data quality. The dataset was split into 80% training and 20% testing data, and the XGBoost regression model was trained for prediction. Finally, model performance was evaluated using R^2 and Mean Absolute Error (MAE) metrics. Fig. 1. illustrates the proposed flowchart for Health Insurance Cost Prediction using machine learning.

The following section describes the individual stages of the proposed approach in detail:

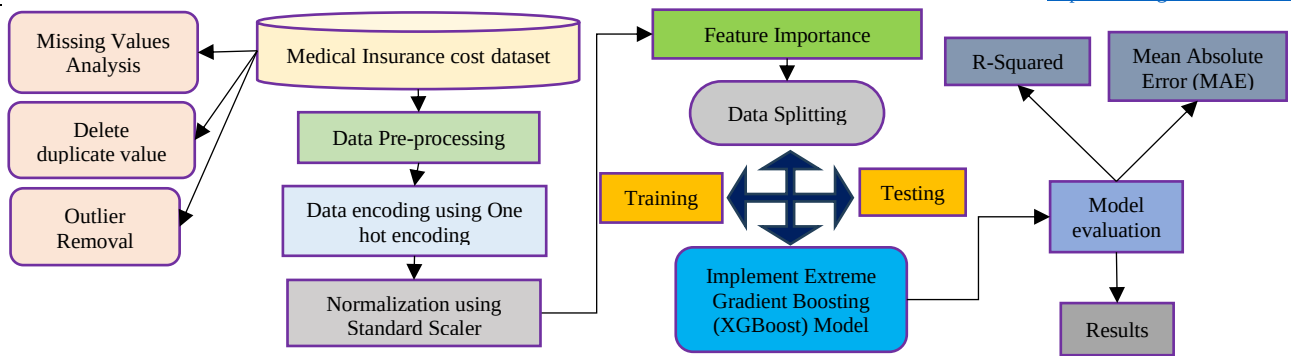


Fig. 1. Proposed Flowchart for Health Insurance Cost Prediction Using Machine Learning

A. Data Gathering and Analysis

This study utilizes the Medical Insurance cost dataset. The dataset of medical insurance expenses used to create the ML model for cost prediction was obtained from Kaggle and comprises 1,338 records with seven attributes. Data visualizations such as bar plots and heatmaps were used to examine distribution, feature correlations etc., are given below:

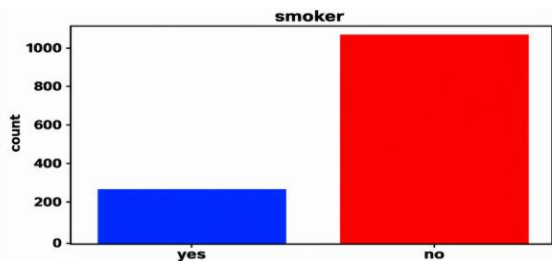


Fig. 2. Bar Graph of Class Distribution

Fig. 2. shows the distribution of the smoking status classes in the data. The graph displays a clear class imbalance with around 1060 records for the non-smoker ("no") class and around 275 records for the smoker ("yes") class. This means that most people in the dataset are non-smokers. Understanding this class distribution is important because such an imbalance can influence the performance of machine learning models, potentially causing them to favour the majority class during prediction.

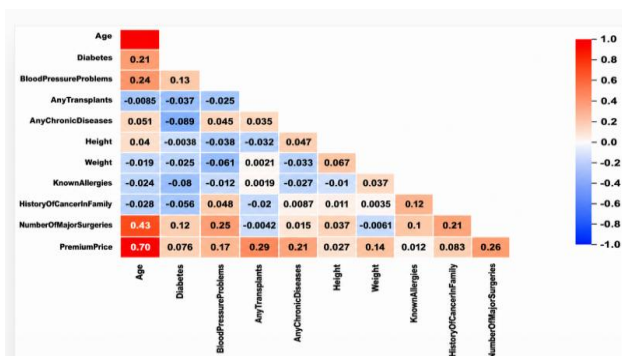


Fig. 3. Correlation Heatmap

Fig. 3. shows the correlation among the dataset features and the target variable, Premium Price. Age exhibits the strongest positive correlation with Premium Price (0.70), indicating that premium costs generally increase with age. Other factors, such as the number of Major Surgeries, Any Transplants, and Any Chronic Diseases, show moderate

positive correlations, while most remaining features have weak relationships with the target variable.

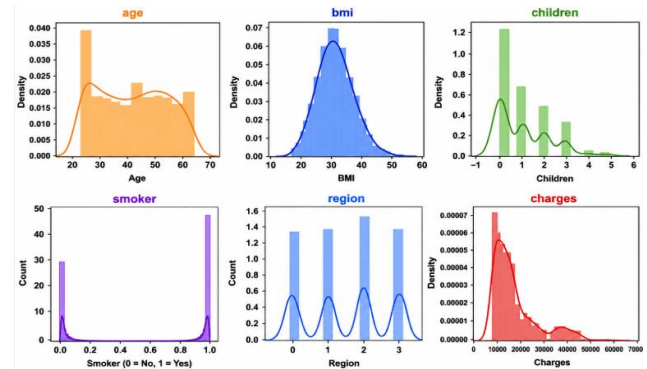


Fig. 4. Histogram Plot for Distribution of Features

Fig. 4. presents the distribution of the dataset features using histogram plots. The variables age, BMI, and children show varying distributions, with BMI following an approximately normal distribution, while age and children exhibit moderate skewness. The charges feature is highly right-skewed, indicating that most individuals incur lower medical expenses, whereas a smaller number have significantly higher charges, and the categorical features smoker and region display the frequency distribution of their respective classes.

B. Data Pre-Processing

The Medical Insurance Cost Dataset was utilized for data preparation and preprocessing to ensure data quality and model readiness. The preprocessing procedure included Missing Value Analysis, duplicate value removal, outlier removal, label encoding of categorical variables, and data normalization. These steps improved the consistency, reliability, and compatibility of the dataset, making it suitable for training and evaluating machine learning models. The key stages of the preprocessing workflow are described below:

- **Missing Values Analysis:** Dealing with missing values is the first stage in data cleansing. Absent data were detected using the `is null ()`, `sum ()` function. Records containing null values were removed using the `dropna ()` method to maintain data integrity.
- **Delete duplicate values:** Make use of the Remove Duplicates method to guarantee that duplicate data is removed from the dataset forever.
- **Outlier Removal:** Extreme data points were identified and removed to improve data quality and enhance model accuracy.

C. Data Encoding Using One Hot Encoding

Data encoding is the technique used to convert categorical data to numerical representation and prepare it for efficient use by machine learning algorithms. In this study, One-hot encoding is applied to the region column, with different columns for each unique region category, represented in the form of 1's and 0's. This ensures that the categorical features for all of the data, since sex, smoker, and region are categorical in nature, get encoded and support the machine learning model.

D. Normalization using Standard Scaler

Given the different scales of each descriptor, the dataset was standardized using the StandardAero () method to transform the data so that the mean of the resulting distribution is zero and the standard deviation is one. This transformation is achieved by subtracting the mean value of each observation and dividing by the standard deviation, as shown in Equation (1):

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

where z is the transformed value of the feature, x is the original value of each descriptor, μ is the mean, and σ is the standard deviation of the feature in the dataset.

E. Feature Importance

The feature importance analysis was conducted to determine the most important features in the prediction of medical insurance costs. The results revealed that features such as age, BMI, and smoking status contributed significantly to the prediction of insurance charges. This analysis helps improve model interpretability and provides insights into the key factors driving healthcare expenses [29].

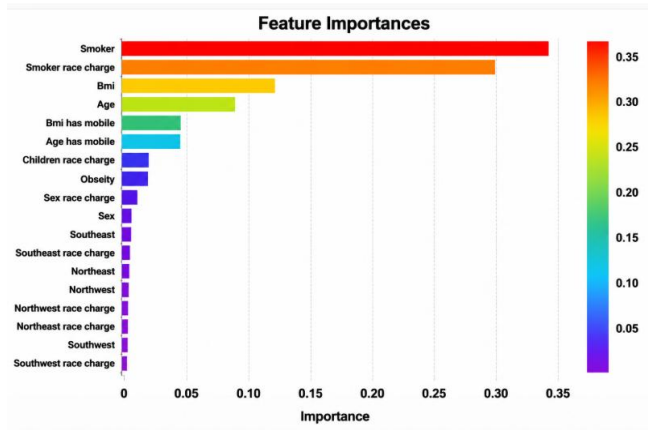


Fig. 5. Feature Importance Plot

Figure 5 illustrates the relative importance of features used in the insurance cost prediction model. The results revealed that the most influential factors are smoker status, smoker-related charges, BMI, and age, and that the regional and demographic properties are relatively less influential on the prediction outcome.

F. Data Splitting

Dataset splitting is a crucial process in ML modelling that supports the model at every stage, from training to evaluation.

The dataset is divided into two distinct subsets: the training set and the testing set. During the experimental phase, 80% of the data is utilized for training and 20% for testing.

G. Proposed Extreme Gradient Boosting (XGBoost) Model

In this work, a supervised machine learning-based boosting model, namely Extreme Gradient Boosting (XGBoost), is proposed for Health Insurance Cost Prediction. XGBoost is a powerful ensemble learning algorithm that constructs multiple decision trees sequentially, where each subsequent tree is trained to correct the errors of the previous trees. For regression tasks, XGBoost minimizes an objective function consisting of a loss function and regularization terms, thereby improving prediction accuracy while reducing overfitting. The mathematical formulation of the XGBoost regression model is expressed in Equation (2):

$$y = f(x) \quad (2)$$

where y is the predicted property price, x is the vector of input features (such as square footage, number of bedrooms, etc.), and $f(x)$ is the XGBoost model that predicts y based on x . To compute $f(x)$, XGBoost builds an ensemble of decision trees that are trained to minimize the mean squared error (MSE) loss function. The model combines the predictions from multiple decision trees to arrive at a final prediction. The general form of the XGBoost regression model can be expressed as Equation (3):

$$y = \sum_{k=1}^K f_k(x) \quad (3)$$

where $f_k(x)$ is the prediction of the k -th decision tree, and K is the total number of decision trees in the ensemble. The prediction of each tree is a weighted sum of the leaf values of the tree, which are learned during training. The prediction of the XGBoost model for a given input x is obtained by summing the predictions of all the decision trees in the ensemble.

The proposed XGBoost model was configured with the following hyperparameter values: number of estimators ($n_estimators$) = 500, learning rate = 0.05, maximum tree depth (max_depth) = 6, minimum child weight (min_child_weight) = 3, subsample ratio = 0.8, column sample by tree ($colsample_bytree$) = 0.8, gamma = 0.1, and random state = 42. These parameters were selected to balance model complexity, minimize overfitting, and improve prediction accuracy for health insurance cost estimation.

H. Evaluation Metrics

Evaluating models is important during ML project development since it lets you judge their performance and present their results more clearly. But it is not easy to predict the exact value of a regression model which is why it try to explain how close the predictions and actual observations are.

It evaluated the models by applying 2 metrics (R^2), and Mean Absolute Error (MAE):

R-Squared (R^2): R^2 is the squared form of the Correlation Coefficient (R) and is used to assess how well a model fit. It also denotes the proportion of the predicted price that is explained by the features (see Equation (4)).:

$$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

Mean Absolute Error (MAE): MAE is a commonly used metric for evaluating the accuracy of a predictive model. It measures the average magnitude of errors in a set of predictions, without considering their direction. A lower MAE value suggests better performance. The formula to calculate MAE is (see Equation (5)).

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - y_i^p| \quad (5)$$

Where:

y is the actual value,

y is the predicted value, and n is the number of observations.

IV. RESULTS AND DISCUSSION

This section presents findings from ML experiments conducted to predict costs in health insurance claims using the Medical Insurance Cost Dataset. ML models were trained on a system with a 3.3 GHz Intel dual-core i6 processor, 1 TB RAM, and Windows 10 operating system. Evaluation of the model's performance was conducted using key regression metrics, namely the R-Square (R^2) and Mean Absolute Error (MAE), as presented in Table II. The proposed Extreme Gradient Boosting (XGBoost) model achieved an R^2 value of 99% and an MAE of 721.05, demonstrating excellent predictive accuracy and a very low prediction error for health insurance cost estimation.

TABLE II. RESULTS OF PROPOSED MODELS FOR HEALTH INSURANCE COST PREDICTION

Matrix	Extreme Gradient Boosting (XGBoost)
MAE	721.05
R-Square	99

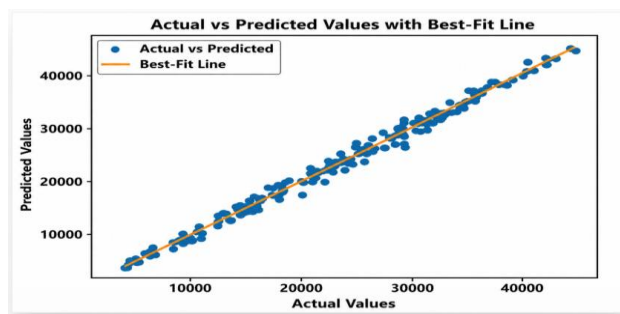


Fig. 6. Actual vs Predicted Health Insurance Costs with Best-Fit Line

Fig. 6 illustrates the relationship between the actual and predicted health insurance costs generated by the XGBoost regression model. The blue scatter points represent individual observations, while the orange best-fit line indicates the overall trend between actual and predicted values. The close alignment of the data points with the best-fit line demonstrates the model's high predictive accuracy and strong correlation between observed and estimated costs. The minimal deviation from the line suggests that the model effectively captures the underlying patterns in the health insurance dataset.

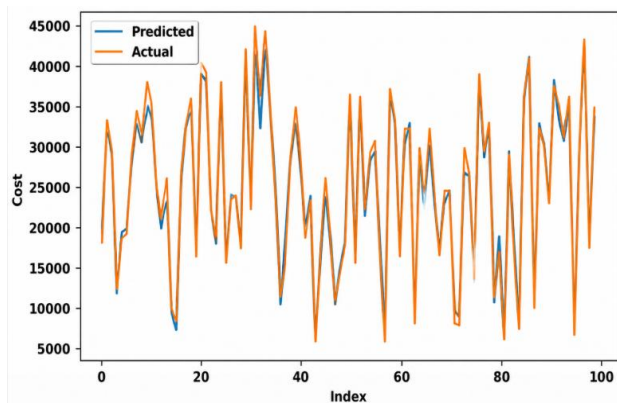


Fig. 7. Line Plot of Actual vs Predicted Medical Insurance Costs

Fig. 7. shows the comparison of real health insurance costs and health insurance cost predicted by the XGBoost model for each observation. The blue line represents the predicted values, while the orange line denotes the actual values. The close overlap between the two curves indicates that the model accurately captures variations in insurance costs and produces reliable predictions. The minimal differences between the actual and predicted values demonstrate the effectiveness and robustness of the proposed model in health insurance cost estimation.

A. Comparative Analysis

The effectiveness of the proposed model is tested with the comparative evaluation of the accuracy presented in Table III, which are compared with other existing models. The performance of LightGBM, Random Forest (RF), AdaBoost, and Extreme Gradient Boosting (XGBoost) was evaluated using Mean Absolute Error (MAE) and R-Square (R^2) metrics. Among all models, XGBoost achieved the best performance with an MAE of 721.05 and an R^2 value of 99%, indicating superior prediction accuracy and minimal error. In comparison, LightGBM, RF, and AdaBoost attained R^2 values of 86.8%, 84%, and 87.1%, respectively, with higher prediction errors. These results demonstrate that the proposed XGBoost model significantly outperforms the other machine learning models for health insurance cost prediction.

TABLE III. COMPARISON OF DIFFERENT MACHINE LEARNING MODELS FOR HEALTH INSURANCE COST PREDICTION

Model	MAE	R2
LightGBM[30]	2381.5	86.8
RF[31]	1379.9	84
AdaBoost[32]	2.3052	87.1
XGBoost	721.05	99

The proposed Extreme Gradient Boosting (XGBoost) model offers several advantages for health insurance cost prediction. It achieved a high R^2 value of 99%, indicating excellent predictive capability, while maintaining a low MAE of 721.05, which reflects minimal prediction error. It is capable of representing complicated relationships between features, dealing with non-linear data patterns and mitigating overfitting via regularization methods. These characteristics make it a robust and reliable model for accurate health insurance cost estimation.

V. CONCLUSION

Insurance is a financial protection mechanism that helps individuals reduce the risk of monetary loss arising from unexpected events. It can cover various aspects of life, including health, automobiles, education, business, and property. Among these, health insurance is particularly important due to the increasing prevalence of diseases, medical emergencies, and rising healthcare expenses. In this study, health insurance costs were predicted using machine learning algorithms, namely Random Forest (RF), AdaBoost, LightGBM, and XGBoost, on the Medical Insurance Cost Dataset. The performance of these models was evaluated using the R^2 metric, where XGBoost achieved the highest prediction accuracy of 99%, outperforming AdaBoost (87.1%), LightGBM (86.8%), and RF (84%). The findings indicate that XGBoost effectively captures complex relationships among insurance-related factors and provides highly accurate cost predictions, making it a suitable tool for health insurance cost estimation and financial decision-making.

A. Limitation and Future Work

A limitation of this study is that the model was developed and evaluated using a single health insurance dataset, which may limit its generalizability to other populations and healthcare environments. Future work can focus on incorporating larger, more diverse datasets from multiple healthcare and insurance sources to improve model robustness and predictive reliability. Additionally, advanced deep learning and hybrid ensemble approaches can be explored to further enhance prediction performance. The inclusion of additional medical, pharmaceutical, and behavioral factors may also provide more comprehensive insurance cost estimation.

- **Funding:** This research received no external funding.
- **Institutional Review Board Statement:** Not applicable.
- **Informed Consent Statement:** Not applicable.
- **Data Availability Statement:** The data supporting the findings of this study are available from the corresponding author upon reasonable request.
- **Conflicts of Interest:** The author declares no conflict of interest.

How to cite this article:

A. Soni, "A Predictive Analytics Approach for Health Insurance Cost Modeling in Pharmaceutical Healthcare," *J. Glob. Res. Electron. Commun.*, vol. 2, no. 9, pp. 1–8, Sep. 2026. Doi.XXX

REFERENCES

- [1] S. Shivam, V. Nitalapati, T. P. Patel, A. K. Padhy, M. Kumari, and R. Purushothaman, "Transformer Based Framework for Imbalanced Transaction Fraud Detection in FinTech Systems," in *2026 14th International Symposium on Digital Forensics and Security (ISDFS)*, Boston, MA, USA: IEEE, Mar. 2026, pp. 1–6. doi: 10.1109/ISDFS69419.2026.11459102.
- [2] R. Snehamrutha, "Patient Engagement Strategies in Community Pharmacies and their Effect on Vaccination Uptake and Medication Synchronizations," *ESP J. Eng. Technol. Adv.*, vol. 3, no. 3, pp. 163–173, 2023, doi: 10.56472/25832646/JETA-V3I7P120.
- [3] P. Mishra, "Design Of Intelligent Healthcare It Infrastructure Using Graph Theory, Network Analysis, And Artificial Intelligence," *Int. J. Appl. Math.*, vol. 38, no. 12s, pp. 2267–2280, Dec. 2025, doi: 10.12732/ijam.v38i12s.1547.
- [4] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.
- [5] S. M. Burwell, "Setting Value-Based Payment Goals — HHS Efforts to Improve U.S. Health Care," *N. Engl. J. Med.*, vol. 372, no. 10, pp. 897–899, Mar. 2015, doi: 10.1056/NEJMp1500445.
- [6] V. M. Montori, N. L. Wilczynski, D. Morgan, and R. B. Haynes, "Optimal search strategies for retrieving systematic reviews from Medline: analytical survey," *BMJ*, vol. 330, no. 7482, p. 68, Jan. 2005, doi: 10.1136/bmj.38336.804167.47.
- [7] K. Bhatia, S. S. Gill, N. Kamboj, M. Kumar, and R. K. Bhatia, "Health Insurance Cost Prediction using Machine Learning," in *2022 3rd International Conference for Emerging Technology (INCET)*, IEEE, May 2022, pp. 1–5. doi: 10.1109/INCET54531.2022.9824201.
- [8] V. Chaturvedi, "Disease Diagnostic Systems based on AI-Applications in Healthcare: Models, Challenges, and Future Directions," *Int. J. Emerg. Res. Eng. Technol.*, vol. 6, no. 4, pp. 207–217, Dec. 2025, doi: 10.63282/3050-922X.IJERET-V6I4P125.
- [9] S. Mahmud, "AI And Data Analytics For Enhancing Home Healthcare: Optimizing Patient Outcomes And Resource Allocation," *Front. Appl. Eng. Technol.*, vol. 2, no. 01, pp. 78–100, February, Jan. 2025, doi: 10.70937/faet.v2i01.61.
- [10] J. A. Cenita, P. R. Asuncion, and J. Victoriano, "Performance Evaluation of Regression Models in Predicting the Cost of Medical Insurance," *Int. J. Comput. Sci. Res.*, vol. 7, pp. 2052–2065, Jan. 2023, doi: 10.25147/ijcsr.2017.001.1.146.
- [11] A. Joon, S. A. Pahune, S. S. S. Neeli, and B. Boddu, "A Robust Artificial Intelligence-Based Financial Forecasting System to Enhance Accuracy with Machine Learning Algorithms," in *Proceedings of the Second International Conference on Advanced Computing and Systems*, W. Ding, A. Chakrabarti, M. Chakraborty, and S. Chakraborty, Eds., Cham: Springer Nature Switzerland, 2026, pp. 522–533. doi: 10.1007/978-3-032-20253-6_40.
- [12] N. Shakhovska, N. Melnykova, V. Chopiyak, and M. Gregus ml, "An Ensemble Methods for Medical Insurance Costs Prediction Task," *Comput. Mater. Contin.*, vol. 70, no. 2, pp. 3969–3984, 2022, doi: 10.32604/cmc.2022.019882.
- [13] N. Shakhovska, N. Melnykova, V. Chopiyak, and M. Gregus ml, "An Ensemble Methods for Medical Insurance Costs Prediction Task," *Comput. Mater. Contin.*, vol. 70, no. 2, pp. 3969–3984, 2022, doi: 10.32604/cmc.2022.019882.
- [14] M. hanafy and O. M. A. Mahmoud, "Predict Health Insurance Cost by using Machine Learning and DNN Regression Models," *Int. J. Innov. Technol. Explor. Eng.*, vol. 10, no. 3, pp. 137–143, Jan. 2021, doi: 10.35940/ijitee.C8364.0110321.
- [15] K. Kaushik, A. Bhardwaj, A. D. Dwivedi, and R. Singh, "Machine Learning-Based Regression Framework to Predict Health Insurance Premiums," *Int. J. Environ. Res. Public Health*, vol. 19, no. 13, pp. 1–15, Jun. 2022, doi: 10.3390/ijerph19137898.
- [16] V. S. Prakash, S. N. Bushra, N. Subramanian, D. Indumathy, S. A. L. Mary, and R. Thiagarajan, "Random forest regression with hyper parameter tuning for medical insurance premium prediction," *Int. J. Health Sci. (Qassim)*, pp. 7093–7101, Aug. 2022, doi: 10.53730/ijhs.v6nS6.11762.
- [17] B. Langenberger, T. Schulte, and O. Groene, "The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data," *PLoS One*, vol. 18, no. 1, p. e0279540, Jan. 2023, doi: 10.1371/journal.pone.0279540.
- [18] H.-H. König et al., "Effects of multiple chronic conditions on health care costs: an analysis based on an advanced tree-based regression model," *BMC Health Serv. Res.*, vol. 13, no. 1, p. 219, Dec. 2013, doi: 10.1186/1472-6963-13-219.
- [19] S. Kukar, "Explainable AI Models for Transparent and Ethical Customer Relationship Management Decision," in *2026 IEEE 2nd International Conference on Secure IoT, Assured and Trusted Computing (SATC)*, Houston, TX, USA: IEEE, Mar. 2026, pp. 1–7. doi: 10.1109/SATC69565.2026.11542554.
- [20] C. A. Powers, C. M. Meyer, M. C. Roebuck, and B. Vaziri, "Predictive Modeling of Total Healthcare Costs Using Pharmacy Claims Data," *Med. Care*, vol. 43, no. 11, pp. 1065–1072, Nov. 2005, doi: 10.1097/01.mlr.0000182408.54390.00.
- [21] X. Guo, W. Gandy, C. Coberley, J. Pope, E. Rula, and A. Wells, "Predicting Health Care Cost Transitions Using a Multidimensional Adaptive Prediction Process," *Popul. Health Manag.*, vol. 18, no. 4, pp. 290–299, Aug. 2015, doi: 10.1089/pop.2014.0087.
- [22] K. Mishra, K. Mishra, L. Kumar, K. Kumar, and A. Kumar, "Enhancing the Forecasting of Medical Insurance Premiums Through the Use of Multiple Methods," in *2026 International Conference on Intelligent Systems in Engineering, Secured Systems and Cybersecurity (ICISESSC)*, IEEE, Apr. 2026, pp. 1022–1026. doi: 10.1109/ICISESSC68634.2026.11542626.
- [23] S. Mohan, S. Sharma, S. Agrawal, and S. Kamatchi, "Optimization of Insurance Claim Cost Prediction Through Health Data and Machine Learning," in *2025 Fifth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, IEEE, Jan. 2025, pp. 1–7. doi: 10.1109/ICAECT63952.2025.10958928.
- [24] M. R and M. Arunraj, "Enhancing Healthcare Insurance Cost Prediction with a Metaheuristic Optimization Approach: Comparative Analysis of Machine Learning Models," in *2025 3rd International Conference on Inventive Computing and Informatics*

- (ICICI), IEEE, Jun. 2025, pp. 1225–1232. doi: 10.1109/ICICI65870.2025.11069726.
- [25] A. Sharma and R. Jeya, "Prediction of Insurance Cost through ML Structured Algorithm," in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, 2024, pp. 495–500. doi: 10.1109/IC2PCT60090.2024.10486304.
- [26] S. Vuddanti, V. G. S. K. R. K. Jamili, S. R. Bommareddy, V. Rotta, and V. Pagadala, "Machine Learning Insights into Personalized Insurance Pricing," in *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, IEEE, Jul. 2024, pp. 923–927. doi: 10.1109/ICSCSS60660.2024.10625562.
- [27] T. T, S. H. T, V. K. V, and K. R, "Medical Insurance Cost Analysis and Prediction using Machine Learning," in *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, IEEE, Mar. 2023, pp. 113–117. doi: 10.1109/ICIDCA56705.2023.10100057.
- [28] S. Panda, B. Purkayastha, D. Das, M. Chakraborty, and S. K. Biswas, "Health Insurance Cost Prediction Using Regression Models," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, IEEE, May 2022, pp. 168–173. doi: 10.1109/COM-IT-CON54601.2022.9850653.
- [29] R. Snehamrutha, "Role of In-Process Quality Control (IPQC) in Reducing Deviations during ANDA-Oriented Pharmaceutical Manufacturing," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 11, no. 6, pp. 554–568, Nov. 2025, doi: 10.32628/CSEIT251117150.
- [30] D. Yadav, "Machine Learning-Based Cost Prediction of Health Insurance Claims Using Medical Dataset," *J. Glob. Res. Multidiscip. Stud. Mach.*, vol. 1, no. 6, pp. 6–12, 2025.
- [31] U. Orji and E. Ukwandu, "Machine learning for an explainable cost prediction of medical insurance," *Mach. Learn. with Appl.*, vol. 15, no. November 2023, p. 100516, 2024, doi: 10.1016/j.mlwa.2023.100516.
- [32] M. Emeç, "Medical Insurance Cost Prediction MedCost: Machine Learning Ensemble Approaches," *Eur. J. Tech.*, vol. 14, no. 1, pp. 88–95, 2024, doi: 10.36222/ejt.1375677.