Optimizing Data-Driven Healthcare Cost Prediction Using XGBoost ML Algorithm

Maddu Naga Supraja
Dept. of CSE
Bapatla Engineering College, Bapatla
Andhra Pradesh – 521356, India
maddunagasupraja@gmail.com

Manikala Sai Vardan
Dept. of CSE
Bapatla Engineering College, Bapatla
Andhra Pradesh -521356,India
saivardhanmanikala@gmail.com

Kommasani Durga Manohar Reddy
Dept. of CSE
Bapatla Engineering College, Bapatla
Andhra Pradesh – 521356, India
Kommasanimanoharreddy05@gmail.

Karri Sekhar Venkata Prasad
Dept. of CSE
Bapatla Engineering College,Bapatla,
Andhra Pradesh – 521356, India
sekharkarri2001@gmail.com

ABSTRACT

An insurance policy lowers or completely removes the costs related to declining returns caused by different risks. A variety of things affect the cost of insurance. These elements have an impact on how insurance plans are made. In the insurance industry, machine learning (ML) has promise for increasing the effectiveness of insurance policy terms. Actual modelling of insurance claims has emerged as a major field of study in the health insurance industry in recent years, primarily for the purpose of determining appropriate rates. This is essential for drawing in new insured's, keeping the ones you already have, and managing current plan participants well. However, it can be difficult to create an accurate forecast model for medical insurance prices because of the multitude of factors that influence them and their inherent complexities. The expected costs of health insurance could be greatly impacted by a number of factors, such as provider characteristics, lifestyle decisions, health status, accessibility in a given area, and demographic information. Actuarial research into predictive modeling in healthcare is still going strong, as more insurance companies look to leverage ML technologies to increase productivity and efficiency. Regression-based ensemble machine learning models that incorporate different Extreme Gradient Boosting (XGBoost) techniques are used in this study to forecast medical insurance expenses.

KEYWORDS: Medical Insurance, Machine Learning (ML), XGBoost, Predictive Modeling, Insurance Cost Prediction, Ensemble Regression Models, Healthcare Analytics.

1. INTRODUCTION

Health is one of the most vital aspects of human life, influencing not only physical well-being but also emotional, mental, and social stability. In today's fast-paced world, individuals strive to secure their future through financial planning, investments, and savings. However, unexpected medical emergencies can derail even the most carefully laid plans. Such unforeseen expenses often lead to financial instability, affecting long-term goals such as education, retirement, or property ownership. This highlights the growing necessity of having comprehensive health insurance coverage.As a result, nearly everyone is insured by some form of health insurance, be it government or corporate.

Health insurance serves as a crucial financial buffer against the high costs of medical care. By paying regular premiums, individuals ensure access to healthcare services through reimbursement policies or cashless treatments. This system prevents sudden out-of-pocket expenditures that could disrupt an individual's financial equilibrium. Insurance not only provides financial security but also peace of mind, encouraging individuals to seek timely medical assistance without hesitation.

Determining the appropriate premium for each policyholder is a complex task. Factors such as age, gender, smoking habits, pre-existing health conditions, region, and family medical history all influence insurance costs. Traditionally, these premiums were calculated manually by underwriters using generalized risk models. However, such manual processes are prone to human error, inconsistencies, and oversimplified assumptions, often resulting in unfair or inaccurate pricing.

To overcome these challenges, the insurance industry is increasingly turning to machine learning (ML) technologies. ML models can process large volumes of historical insurance data, identify complex patterns, and make highly accurate predictions. This data-driven approach reduces reliance on manual estimation and enhances the reliability and fairness of insurance pricing. Moreover, ML systems continuously improve over time as more data becomes available, further refining their predictions.

This project leverages the power of ML by implementing the Extreme Gradient Boosting (XGBoost) algorithm, an advanced and efficient ensemble learning technique. XGBoost is particularly suited for structured data and has proven successful in numerous prediction and classification tasks. In this context, it is trained on historical healthcare insurance data to predict future insurance charges with high precision. The model captures non-linear relationships and interdependencies among features, which traditional models might overlook.

2. LITERATURE REVIEW

Linear Regression is commonly used to predict continuous outcomes like insurance costs by modeling relationships between variables such as age and BMI. Logistic Regression handles classification tasks like claim approval. While both are simple and interpretable, they often fail to capture non-linear feature interactions. Frees and Valdez (2008) noted that these models perform poorly on complex healthcare datasets [1].

Decision Trees split data based on feature conditions, offering clear, rule-based predictions that are easy to interpret. They are useful in insurance pricing but often suffer from overfitting and instability. Weiss and Indurkhya (1995) found that while effective for risk modeling, decision trees performed inconsistently on complex, large-scale datasets [2].

Support Vector Machines (SVM) are effective for classification tasks like identifying high-risk individuals in insurance. They perform well with high-dimensional data but are computationally intensive and sensitive to kernel choice. Wu et al. (2007) found SVM reliable for healthcare cost prediction, with moderate results in regression scenarios (SVR) [3].

Random Forests, as an ensemble of Decision Trees, enhance accuracy and reduce overfitting by aggregating multiple models. Bertsimas et al. (2018) found them superior to individual trees and logistic regression for predicting insurance costs [4]. They are robust and useful for feature analysis but may lack transparency for regulatory needs.

Bagging, an ensemble method, boosts model stability and generalization by training multiple models on varied data subsets. It's effective in premium prediction, especially with noisy or incomplete data. Though simpler than boosting, it still improves performance over single-model approaches [5].

Clustering algorithms like K-Means group individuals by health profiles to aid in segmentation and pricing strategies. Su et al. (2010) showed their usefulness in identifying customer groups and detecting fraud [6]. While not used for direct premium prediction, clustering supports preprocessing and strategic decision-making.

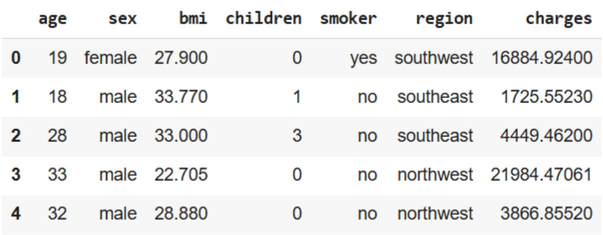
3. DATASET DESCRIPTION

Kaggle was the key data source for this project. There are 2773 records in the collection, each with six properties. 'Age,' 'gender,' 'BMI,' 'children, 'smoker,' and 'charges' are the attributes [12]. The information was organized and saved as a CSV file. The data collection does not lend itself to direct regression analysis. As a result, before using the dataset in multiple regression algorithms, it must be cleansed. The prediction is not affected by every attribute in a dataset. Some qualities impair accuracy to the point that they must be removed from the code's features. Removing these characteristics improves accuracy as well as overall performance and speed.

Table 1: Dataset Overview

Name	Description
Age	age of the client
BMI	body mass index
the number of children	number of children the client has
Gender	male/female
Smoker	Whether a client is a smoker or not
Region	whether a client lives southwest, southeast, northeast, or northwest
charges (target variable)	medical cost the client pays

smoking status was the most dominant predictor of insurance costs. This is logical and expected, as smoking is a well-known risk factor that significantly increases the likelihood of health issues, resulting in higher premiums. Age was the next most important feature, reflecting the increased risk of medical expenses as individuals grow older. Body Mass Index (BMI) also showed high importance, indicating that individuals with higher BMI values may be at risk of obesity-related conditions, which influence insurance pricing. Other features like the number of children, sex, and region had relatively lower importance but still contributed to refining the predictions.



	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Figure 1: Sample of Health Insurance Dataset

4. PROPOSED SYSTEM

The proposed system is designed to predict healthcare insurance charges using a machine learning-based approach that ensures higher accuracy and automation compared to traditional manual methods. It leverages the XGBoost regression algorithm, known for its performance and ability to handle complex, non-linear relationships between features. The goal is to create a data-driven solution that can estimate insurance premiums based on personal and demographic inputs provided by the user.

The system utilizes a historical dataset containing various attributes such as age, gender, body mass index (BMI), smoking status, number of children, and region. These features are critical in determining the cost of medical insurance and serve as inputs to the model. The XGBoost algorithm is trained on this data to learn patterns and relationships that influence premium amounts. To further enhance the model's performance, hyperparameter tuning is performed using RandomizedSearchCV, which helps optimize model parameters for better accuracy and generalization.

To make the model accessible and user-friendly, it is deployed through a web interface built using the Flask framework. This interface allows users to input their details via a form and receive instant predictions of their estimated insurance charges. The model runs in the background, processing the input data and generating a result in real time.

Overall, the proposed system offers a scalable, efficient, and practical approach for predicting healthcare costs. It reduces human effort, improves accuracy, and provides transparency in premium calculation, making it valuable for both insurance providers and policyholders.

4.1 Working process of XGBoost Algorithm

XGBoost (Extreme Gradient Boosting) is an efficient and scalable machine learning algorithm that builds an ensemble of decision trees in a sequential manner, where each new tree corrects the errors made by the previous ones. It starts with an initial prediction and iteratively minimizes the residual errors using gradient descent optimization. XGBoost incorporates regularization techniques to avoid overfitting and improve model generalization. It also supports missing value handling, column and row sampling, and parallel processing, making it faster and more accurate than traditional boosting methods. This combination of speed, accuracy, and robustness makes XGBoost ideal for complex prediction tasks like healthcare cost estimation.

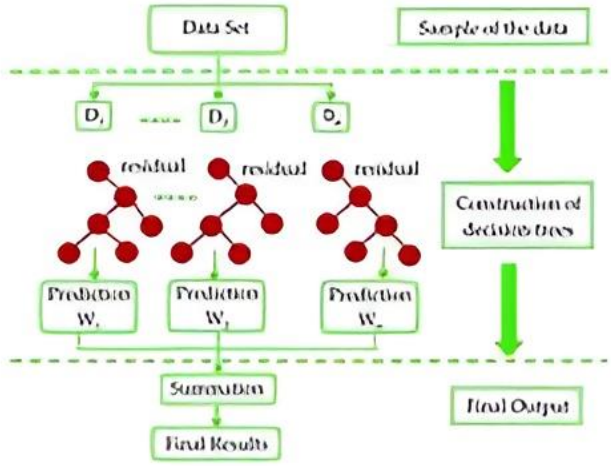


Figure 2: Working process of XGBoost Algorithm

4.2 Benefits of XGBoost Algorithm

The XGBoost algorithm offers several key benefits that make it highly effective for predictive modeling tasks. It is known for its exceptional speed and performance, achieved through optimized parallel processing and efficient handling of sparse data. XGBoost includes built-in regularization techniques, which help reduce overfitting and improve model generalization. It can capture complex, non-linear relationships between features and supports both classification and regression problems. Additionally, it handles missing values internally and provides insights into feature importance, making it both powerful and interpretable for real-world applications like healthcare cost prediction.

4.3 Implementation

4.3.1 Importing Libraries

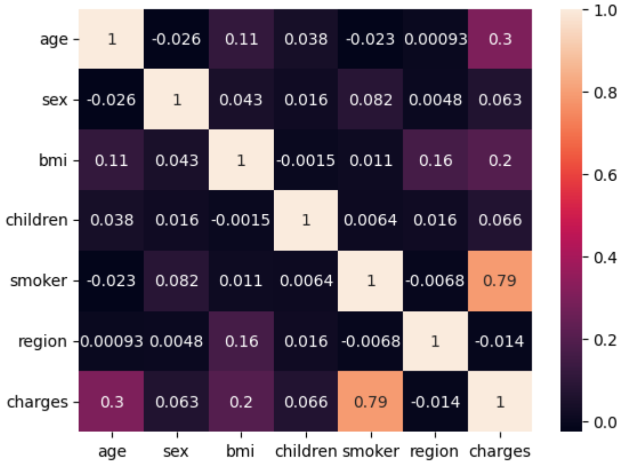
The implementation starts by importing essential libraries for data processing, modeling, and deployment. **Pandas** and **NumPy** handle data operations, while **Matplotlib** and **Seaborn** assist in visualization. **Scikit-learn** supports preprocessing and evaluation, and **XGBoost** is used for building the regression model. **Flask** enables the creation of a user-friendly web interface for real-time predictions.

4.3.2 Loading and Exploring the dataset

The dataset is loaded using the pandas.read_csv() function, followed by initial exploration using methods such as head() to preview the data, info() to examine data types and structure, and describe() to view summary statistics. Exploratory Data Analysis (EDA) is then conducted using visualizations and correlation matrices to identify patterns, outliers, and potential imbalances. This step provides a comprehensive understanding of the dataset and guides the necessary preprocessing steps for model development.

4.3.3 Data Preprocessing

The raw data is processed to make it suitable for machine learning by performing several key transformations. Although the dataset generally lacks missing values, null checks are conducted as a standard precaution. Categorical features such as sex, smoker, and region are converted into numeric format using Label Encoding. Inputs are also checked for consistency, and irrelevant columns, if any, are removed. While XGBoost can handle unscaled data efficiently, data scaling is considered only if significant outliers are present. These steps ensure the dataset is clean, well-structured, and ready for effective model training.



	age	sex	bmi	children	smoker	region	charges
age	1	-0.026	0.11	0.038	-0.023	0.00093	0.3
sex	-0.026	1	0.043	0.016	0.082	0.0048	0.063
bmi	0.11	0.043	1	-0.0015	0.011	0.16	0.2
children	0.038	0.016	-0.0015	1	0.0064	0.016	0.066
smoker	-0.023	0.082	0.011	0.0064	1	-0.0068	0.79
region	-0.00093	0.0048	0.16	0.016	-0.0068	1	-0.014
charges	0.3	0.063	0.2	0.066	0.79	-0.014	1

Figure 3: Correlation between Features

4.3.4 Splitting the dataset

After preprocessing, the dataset is divided into input features (X) and the target variable (y), which represents the insurance charges. The data is then split into a training set, typically comprising 80% of the data for model learning, and a testing set, usually 20%, for model validation. This split is performed using the train_test_split() function from sklearn.model_selection, ensuring that the model is trained and evaluated on separate, non-overlapping data, which helps in assessing its generalization performance.

4.3.5 Model Building using XGBoost Algorithm

An instance of XGBRegressor is created and initially trained with default parameters to establish a baseline. XGBoost builds decision trees sequentially, where each new tree corrects errors from the previous ones, effectively minimizing prediction errors. This boosting technique captures both linear and non-linear relationships, making it ideal for complex tasks like insurance price prediction. The result is a highly accurate and efficient predictive model.

4.3.6 Hyperparameter Tuning

To improve the model's performance, RandomizedSearchCV is used for hyperparameter tuning. A predefined grid of parameters such as n_estimators, max_depth, learning_rate, gamma, subsample, and colsample_bytree is explored to identify the optimal configuration. Cross-validation is applied to assess each parameter combination, and the best-performing set is selected based on the highest R² score. This optimization step enhances the model's accuracy, reduces prediction errors, and improves generalization by minimizing the risk of overfitting.

4.3.7 Model Evaluation

Once the model is trained, its performance is assessed using several key evaluation metrics. The R² Score measures how well the model explains the variance in the target variable, indicating overall accuracy. The Mean Absolute Error (MAE) provides the average magnitude of prediction errors, offering a straightforward interpretation of model reliability. Meanwhile, the Root Mean Squared Error (RMSE) gives greater weight to larger errors, helping identify significant deviations. Together, these metrics offer a comprehensive view of the model's predictive accuracy and effectiveness in estimating healthcare insurance costs.

4.3.8 Saving the Model

After the model is trained and evaluated, it is crucial to save it for future predictions without the need for retraining. This ensures the model is retained in its best-performing state and can be smoothly integrated into a deployment setup. In this project, the trained XGBoost model is serialized using the pickle library and stored as xgb_model.pkl. This approach significantly reduces computational overhead during deployment, allowing the model to be quickly loaded and used for predicting insurance charges based on user inputs. The saved model is later utilized in the Flask application to deliver real-time predictions through a user-friendly web interface.

4.3.9 Create Front-End using Flask

To make the model interactive and user-accessible, a web application is developed using Flask. A straightforward HTML form captures input features such as age, sex, BMI, number of children, smoking status, and region. Upon submission, the data is sent to the Flask backend, where it is preprocessed and fed into the trained XGBoost model for prediction. The model then generates an estimate of the insurance charges, which is displayed to the user through the same web interface. This real-time interaction makes the application practical, user-friendly, and easily accessible through a standard web browser.

4.4 Comparing with other Algorithms

Several machine learning algorithms were explored for predicting healthcare insurance costs, each with its own strengths and limitations. Multi-Linear Regression is simple and interpretable but struggles with non-linear relationships and multicollinearity. Decision Tree Regression offers easy visualization and rule-based decisions but is prone to overfitting and instability with complex data. Random Forest Regression, an ensemble of decision trees, improves accuracy and reduces overfitting, but its predictions lack transparency and are computationally heavier. Support Vector Regression (SVR) performs well in high-dimensional spaces but is sensitive to kernel selection and computationally intensive for large datasets. XGBoost, a gradient boosting algorithm, addresses many of these issues by offering high accuracy, fast computation, and regularization to prevent overfitting, making it the most effective choice for this project despite its slightly higher complexity.

Table 2:Comparision of Algorithms

Algorithm	**Mean Absolute Error (MAE)**	**Mean Squared Error (MSE)**	**Root Mean Squared Error (RMSE)**	**R-Squared (R²)**
Linear Regression	4260.99	34,515,553.66	5874.99	77.76%
Decision Tree Regression	3285.99	44,580,876.15	6676.89	71.28%
Random Forest Regression	2688.10	23,723,428.01	4870.67	84.71%
XGBoost Regression	1263.52	7,405,386.66	2721.28	95.17%

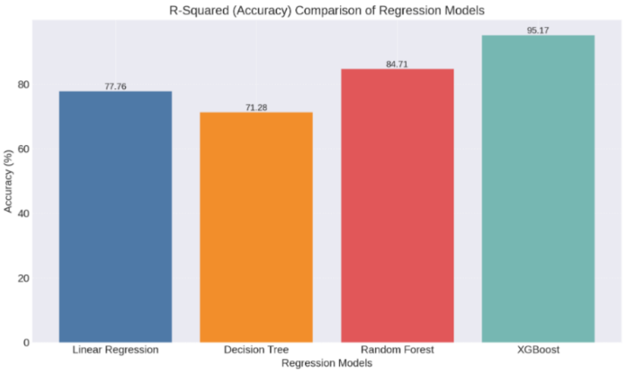


Figure 4:R2 Score of Algorithms

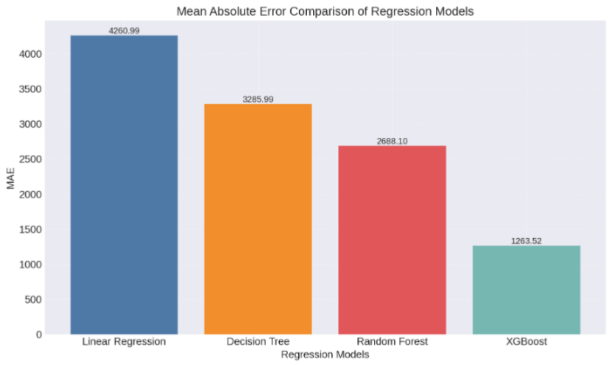


Figure 5:MAE of Algorithms

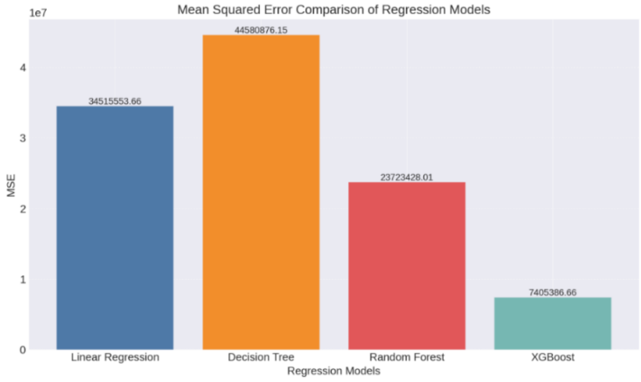


Figure 6:MSE of Algorithms

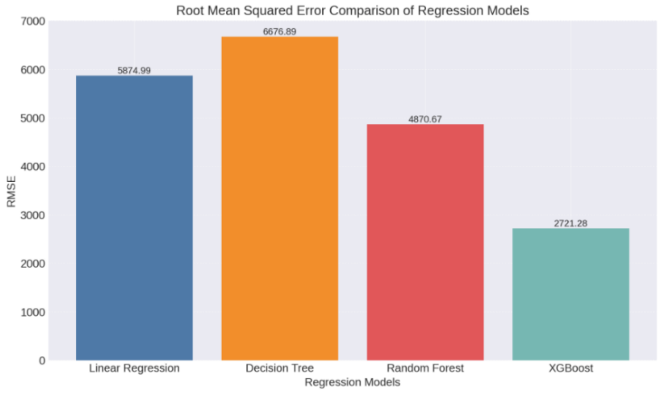


Figure 7: RMSE of Algorithms

5. SYSTEM DESIGN

The system design outlines the structure and interaction of various components involved in predicting healthcare insurance costs. It follows a modular approach to ensure smooth data flow from user input to prediction output. Each module, including the frontend interface, backend processing, and the prediction engine, is designed for a specific task to enhance efficiency and maintainability. This design enables real-time predictions while ensuring the system remains user-friendly, accurate, and scalable.

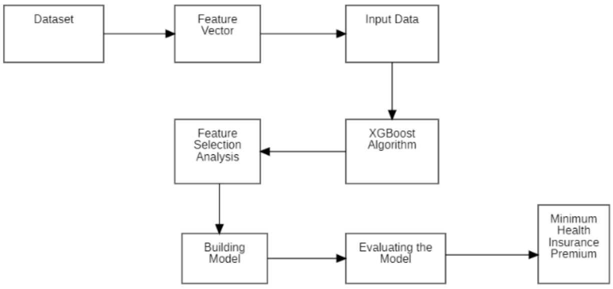


Figure 8: Flow Diagram

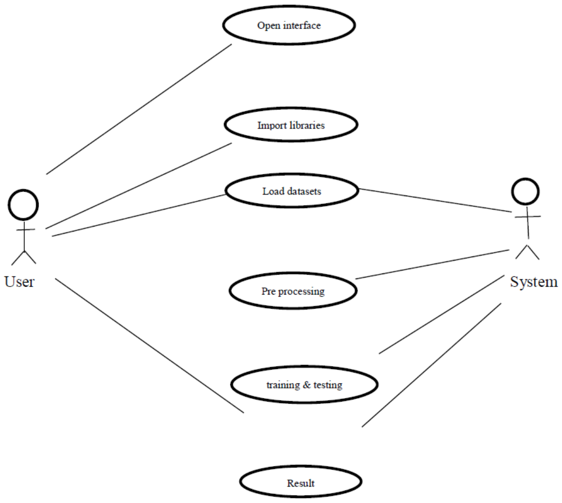


Figure 9: Use case Diagram

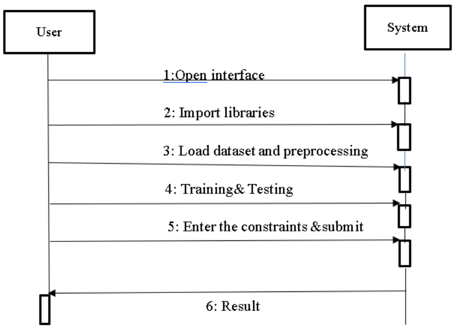
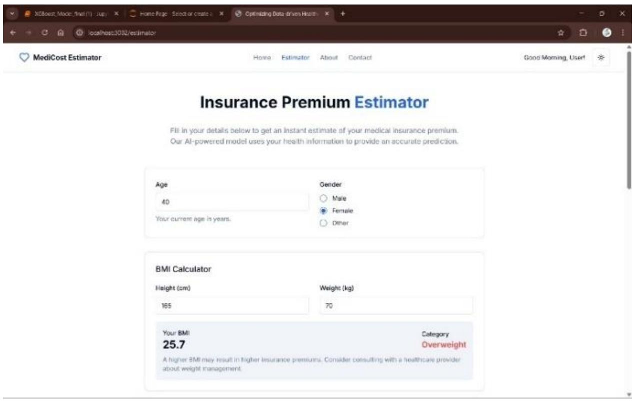


Figure 10: Sequence Diagram

6. RESULTS AND DISCUSSION

At 95 percent, the accuracy of the expected amount predicted using XGBoost model was the best. All of the other regression models had similar accuracy, with a range of 70% to 80% accuracy. A table showing the accuracy percentages of various attributes for all four models is provided in the following section for each model separately and merged.

The below figures showcase the user interface of the web application built using Flask for predicting medical insurance premiums. The application, titled Healthcare Cost Prediction using XGBoost provides a clean and interactive form where users can input essential health and demographic information. It includes fields for Age, Gender, Height, Weight, Smoking Status, Region, and Number of Children. Upon entering the required details, the user can click the Calculate Premium button to receive an instant prediction of their insurance cost. This interface ensures ease of use and improves accessibility for users seeking accurate insurance estimates.


XGBoost_Model_final (1) - Jup
Home Page - Select or create a
Optimizing Data-driven Health
localhost:3002/estimator
MediCost Estimator
Home
Estimator
About
Contact
Good Morning, User!

Insurance Premium Estimator

Fill in your details below to get an instant estimate of your medical insurance premium. Our AI-powered model uses your health information to provide an accurate prediction.

Age
40
Your current age in years.

Gender
○ Male
● Female
○ Other

BMI Calculator

Height (cm)
165

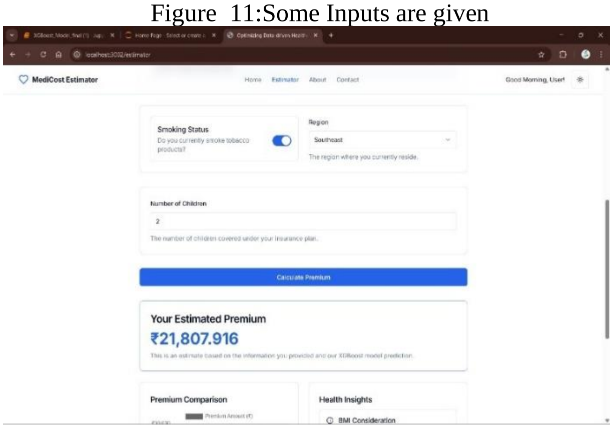
Weight (kg)
70

Your BMI
25.7

Category
Overweight

A higher BMI may result in higher insurance premiums. Consider consulting with a healthcare provider about weight management.

Figure 11:Some Inputs are given



Smoking Status
Do you currently smoke tobacco products?

Region
Southeast
The region where you currently reside.

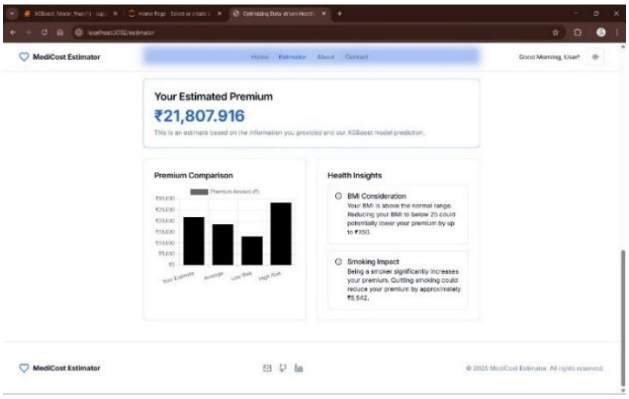
Number of Children
2
The number of children covered under your insurance plan.

Your Estimated Premium
₹21,807.916
This is an estimate based on the information you provided and our XGBoost model prediction.

Premium Comparison
Premium Amount (₹)

Health Insights
BMI Consideration

Figure 12: Estimator predicting Insurance Premium



Your Estimated Premium

₹21,807.916

This is an estimate based on the information you provided and our XGBoost model prediction.

Premium Comparison

Health Insights

BMI Consideration
Your BMI is above the normal range. Reducing your BMI to below 25 could potentially lower your premium by up to ₹350.

Smoking Impact
Being a smoker significantly increases your premium. Quitting smoking could reduce your premium by approximately ₹6,542.

Figure 13: Showing Health Insights

7. CONCLUSION AND FUTURE SCOPE

In this project, we developed a predictive model for health insurance pricing using machine learning algorithms, focusing on the XGBoost algorithm. The dataset included features like age, BMI, gender, smoking status, and region to estimate insurance charges. Among the various regression models tested—Linear Regression, Support Vector Machine, Random Forest, Gradient Boosting, and Decision Tree—XGBoost outperformed others due to its ability to handle complex, non-linear relationships. Hyperparameter tuning with RandomizedSearchCV further improved the model's accuracy, and evaluation metrics such as R², MSE, RMSE, and MAE validated its strong performance.

Additionally, feature importance analysis highlighted key predictors, aiding in the design of fair and personalized policies. Overall, XGBoost proved to be a reliable tool for real-world insurance cost prediction, supporting better risk assessment and operational efficiency.The effect of several variables on the calculated amount was examined. A person's age and smoking history were found to have a significant impact on every algorithm. An unneeded attribute was removed from each of the features. Premiums are determined by a person's health rather than the terms and conditions of another insurance provider. Some algorithms can be employed in the next years to predict premiums based on data. People and insurance companies can work together to deliver better and more health-focused coverage as a result of this.

Exploring advanced ensemble techniques and deep learning models could further enhance predictive accuracy beyond what XGBoost offers, especially for more complex datasets. Techniques like model stacking, neural networks, or time-series forecasting can provide dynamic pricing solutions, while AutoML can streamline hyperparameter tuning and reduce development time. Integrating real-time data from hospital management systems or wearable health devices can enable dynamic premium adjustments based on ongoing health trends. Additionally, implementing sub-models for fraud detection can safeguard against anomalies and ensure fair pricing. Incorporating feedback loops—such as insurer decisions or user appeals—will allow the model to retrain periodically, adapt to evolving patterns in healthcare and insurance, and maintain high accuracy and relevance over time.

8. REFERENCES

[1] E. W. Frees and E. A. Valdez, "Understanding relationships using copulas," North American Actuarial Journal, vol. 12, no. 1, pp. 1–25, 2008.

[2] S. M. Weiss and N. Indurkhya, "Rule-based machine learning methods for functional prediction," Journal of Artificial Intelligence in Medicine, vol. 6, no. 4, pp. 367–384, 1995.

[3] J. Wu, J. Xie, and P. Liu, "Support vector machine modeling for health care cost prediction," Expert Systems with Applications, vol. 33, no. 1, pp. 1–5, 2007.

[4] D. Bertsimas, J. Dunn, and A. R. McKinney, "Predicting healthcare costs and utilization using machine learning," Health Services Research, vol. 53, no. 4, pp. 2366–2379, 2018.

[5] L. Breiman, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, 1996.

[6] M. Su, C. Wang, and P. Chen, “A clustering-based approach to customer segmentation in the insurance industry,” Expert Systems with Applications, vol. 37, no. 9, pp. 6322–6328, 2010.

[7] Health Insurance of India’s missing middle”, Niti Ayog India, Oct 2021, [Online]. Available: https://www.niti.gov.in

[8] Lahiri B, Agarwal N. “Predicting healthcare expenditure increase for an individual from Medicare data”.

[9] Douglas C Montgomery, Elizabeth A Peck and G Geoffrey Vining, “Introduction to linear regression analysis”, John Wiley & Sons, vol. 821, 2012.

[10] Izmie, A. A., et al. "Healthcare Management and Medical Insurance with Predictive Analytics Using Machine Learning." International Research Journal of Innovations in Engineering and Technology 7.10 (2023): 49.