A Project Report on

Optimizing Data-Driven Healthcare Cost Prediction Using XGBoost ML Algorithm

Submitted in partial fulfillment for award of

Bachelor of Technology

Degree

in

Computer Science and Engineering

By

M.Naga Supraja (Y21ACS499) **M.Sai Vardan (Y21ACS503)**

K.D.Manohar Reddy (Y21ACS482) **K.S.V.Prasad (Y20ACS470)**

Under the guidance of
Mr.N.Srikanth, M.Tech.,(Ph.D.)
Assistant Professor

Department of Computer Science and Engineering
Bapatla Engineering College
(Autonomous)
(Affiliated to Acharya Nagarjuna University)
BAPATLA – 522 102, Andhra Pradesh, INDIA
2024-2025

Department of
Computer Science and Engineering

CERTIFICATE

This is to certify that the project report entitled **Optimizing Data-Driven Healthcare Cost Prediction Using XGBoost ML Algorithm** that is being submitted by M.Naga Supraja (Y21ACS499), M.Sai Vardan (Y21ACS503), K.Durga Manohar Reddy (Y21ACS482), K.Sekhar Venkata Prasad (Y20ACS470) in partial fulfillment for the award of the Degree of Bachelor of Technology in Computer Science & Engineering to the Acharya Nagarjuna University is a record of bonafide work carried out by them under our guidance and supervision.

Date:

Signature of the Guide	**Signature of the HOD**
Mr.N.Srikanth	**Dr. M. Rajesh Babu**
Assistant Professor	**Assoc. Prof. & Head**

DECLARATION

We declare that this project work is composed by ourselves, that the work contained herein is our own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

M.Naga Supraja (Y21ACS499)

M.Sai Vardan (Y21ACS503)

K.Durga Manohar Reddy (Y21ACS482)

K.Sekhar Venkata Prasad (Y20ACS470)

Acknowledgement

We sincerely thank the following distinguished personalities who have given their advice and support for successful completion of the work.

We are deeply indebted to our most respected guide **Mr.N.Srikanth,** M.Tech.,(Ph.D), Asst.Professor, Department of CSE, for his valuable and inspiring guidance, comments, suggestions and encouragement.

We extend our sincere thanks to **Dr. M. Rajesh Babu**, Assoc. Prof. & Head of the Dept. for extending his cooperation and providing the required resources.

We would like to thank our beloved Principal **Dr.N.Rama Devi** for providing the online resources and other facilities to carry out this work.

We would like to express our sincere thanks to our project coordinator

Dr.P.Pardhasaradhi, Prof. Dept. of CSE for his helpful suggestions in presenting this document.

We extend our sincere thanks to all other teaching faculty and non-teaching staff of the department, who helped directly or indirectly for their cooperation and encouragement.

M.Naga Supraja (Y21ACS499)

M.Sai Vardan (Y21ACS503)

K.Durga Manohar Reddy (Y21ACS482)

K.Sekhar Venkata Prasad (Y20ACS470)

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.

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

Table of Contents

List of Figures

List of Tables

1 Introduction

Health is a crucial aspect of every individual's life, influencing physical, emotional, mental, and social well-being. While people plan for the future by investing in assets, unforeseen medical emergencies can cause sudden financial strain, disrupting long-term goals like education, savings, or retirement plans. Health insurance acts as a financial shield, covering medical expenses and preventing economic instability. By paying a premium, policyholders gain access to medical treatments through reimbursements or cashless hospital services.

Determining insurance premiums accurately is complex due to factors like age, medical history, and lifestyle. Traditional manual calculations often result in errors and pricing inconsistencies. Machine learning (ML) addresses these issues by analyzing historical data to uncover hidden patterns. This improves the accuracy of cost estimation and minimizes human errors. As a result, ML enhances efficiency in insurance processes. It also ensures fair and data-driven pricing for policyholders.

This project utilizes machine learning, specifically the Extreme Gradient Boosting (XGBoost) algorithm, to predict medical insurance costs more accurately. By training on past insurance data, the model automates premium calculations, reducing errors and improving efficiency. The integration of ML ensures fair pricing, enhances decision-making for insurers, and benefits policyholders by offering more reliable insurance cost predictions.

1.1 Background

Medical insurance serves as a crucial financial safety net, shielding individuals and families from the burden of unexpected medical expenses. The rising costs of healthcare services make it essential for people to have a reliable insurance policy that can cover hospitalization, medication, and other medical treatments. However, determining the appropriate premium for each individual remains a complex challenge due to the wide range of factors influencing health risks. Lifestyle choices, pre-existing medical conditions, age, and geographic location all play significant roles in determining the cost of insurance coverage. Insurers must carefully assess these variables to offer fair and accurate premium rates while ensuring profitability. Traditional actuarial models have long been used for this purpose, but they often fail to capture intricate patterns in large datasets. This calls for more advanced techniques such as machine learning to enhance predictive accuracy..

Machine learning, particularly advanced algorithms like Extreme Gradient Boosting (XGBoost), has revolutionized insurance premium prediction by analyzing vast amounts of data with high efficiency. The ability to incorporate multiple influencing factors, such as BMI, smoking habits, and medical history, ensures a more precise estimation of risk. Furthermore, predictive models assist in reducing fraudulent claims and minimizing financial losses for insurance providers. A well-optimized premium prediction system benefits both insurers and policyholders, as it ensures fairness and affordability. With the continuous advancements in AI and data analytics, the future of insurance pricing is expected to become more transparent and data-driven. As a result, integrating sophisticated algorithms into insurance pricing frameworks is becoming a necessity rather than an option..

1.2 Importance of Healthcare Cost Prediction

Accurate insurance price prediction is essential for maintaining a fair and sustainable insurance system. For insurance companies, precise premium estimation helps in managing financial risks and ensuring profitability. If premiums are set too low, insurers may face substantial losses due to high claim payouts. Conversely, if premiums are too high, potential customers may find insurance unaffordable, leading to decreased policy subscriptions.

A well-calculated premium ensures that the insurer can cover claims while maintaining competitive pricing. Additionally, accurate predictions help insurance companies comply with regulatory requirements, as authorities often mandate fair pricing practices to protect consumers. By leveraging advanced predictive models, insurers can minimize uncertainties and set premiums that accurately reflect the risk profile of each policyholder.

For policyholders, precise premium estimation ensures fairness and affordability. Lower-risk individuals avoid overpaying, while high-risk customers are charged appropriately. Accurate pricing encourages more people to opt for health insurance, improving overall healthcare access. Machine learning techniques, such as XGBoost, refine pricing models by considering factors like age, medical history, and lifestyle. This data-driven approach creates personalized policies and enhances transparency in insurance pricing. Ultimately, fair premium prediction benefits both insurers and customers by balancing costs and risks effectively.

1.3 Problem Statement

Accurately predicting medical health insurance premiums is a critical task for insurance providers. Premium estimation depends on several factors such as age, gender, medical history, BMI, and lifestyle habits. Traditional statistical models often fail to capture the complex, nonlinear relationships among these variables. This can lead to inaccuracies in pricing, affecting both insurers and policyholders. Underpricing can result in financial losses, while overpricing may discourage customers. Therefore, a smarter and more adaptable approach is required.

The problem becomes more challenging due to the variability in individual health conditions and regional healthcare costs. People living in urban areas may face higher treatment charges compared to those in rural regions. Lifestyle choices like smoking and diet also influence health risks, making it harder to create a standard model. Each person's health profile is unique and dynamic, requiring models that can adjust accordingly. This diversity makes fair and accurate premium prediction a difficult task. Hence, more advanced techniques are needed to tackle this issue.

Additionally, the availability of large-scale healthcare data creates opportunities and challenges. These datasets often contain missing values, imbalanced samples, or confidential information that must be handled carefully. Standard models may struggle with such data and produce biased or unreliable results. Ensuring both the accuracy and ethical use of data is critical in model development. Data security and privacy concerns also limit access to valuable information. This further complicates the prediction process for insurers.

To solve these problems, the project utilizes the XGBoost algorithm, a powerful machine learning technique. XGBoost is known for its efficiency, scalability, and ability to handle complex datasets. It helps in modeling nonlinear patterns and interactions among variables with high accuracy. The goal is to build a robust model that can deliver fair and personalized premium predictions. This approach benefits insurers by reducing risk and customers by promoting transparent pricing. It represents a step forward in the application of AI in healthcare finance.

1.4 Motivation

The rising complexity of healthcare systems and individual risk factors has made insurance pricing more challenging. Traditional methods fall short in delivering precise premium estimations. This project is motivated by the need to leverage modern technology for fair, accurate, and personalized insurance pricing.

1.4.1 Increasing Demand for Personalized Insurance:

As people become more health-conscious, they expect insurance plans tailored to their specific needs. A one-size-fits-all approach no longer appeals to modern consumers. Personalized premiums based on individual health profiles ensure fairness and transparency. Predictive models help assess risk more accurately, enabling insurers to offer customized plans. This growing demand motivates the use of data-driven technologies in the insurance sector.

1.4.2 Complexity of Health Risk Factors:

Health insurance pricing depends on numerous factors like age, BMI, smoking habits, and pre-existing conditions. These variables interact in non-linear ways, making prediction difficult with traditional methods. Understanding the influence of these complex interactions is essential for setting fair premiums. Machine learning models, especially XGBoost, excel at identifying hidden patterns. This motivates the application of such models for better prediction accuracy

1.4.3 Limitations of Traditional Approaches:

Conventional actuarial models often fall short when dealing with large and diverse datasets. They typically assume linearity and lack the adaptability needed for modern data structures. Additionally, they are less effective in handling missing values or outliers. Machine learning overcomes these issues with flexible, automated learning techniques. The limitations of older methods highlight the need for more advanced tools like XGBoost.

1.4.4 Need for Fair and Accurate Premium Estimation:

Health insurance premiums must reflect an individual's actual risk to ensure fairness. Traditional pricing models often overgeneralize, leading to overcharging low-risk individuals and undercharging high-risk ones. This imbalance affects customer trust and company profitability. By using machine learning, we aim to bring fairness and precision to premium prediction. Personalized pricing encourages more people to adopt health insurance. It also helps companies remain competitive and financially stable.

1.4.5 Benefits for Insurance Providers :

Accurate premium prediction helps insurance companies manage risk more effectively. It reduces chances of underestimating or overestimating premiums, which can lead to losses or customer dissatisfaction. With better forecasting, companies can set optimal prices and remain competitive in the market. Machine learning enhances business decision-making and resource allocation. This drives motivation to incorporate AI-based models into pricing strategies.

1.5 Objective

To achieve accurate and fair health insurance premium predictions, this project outlines several clear goals. These goals include:

i. **To Develop a Predictive Model Using XGBoost:** XGBoost is selected for its speed, accuracy, and ability to handle complex data. The objective is to train this model on real-world health insurance data. It should effectively estimate premium values based on individual inputs.

ii. **To Improve Fairness in Premium Estimation:** The project strives to reduce biases in pricing caused by traditional models. Machine learning enables fairer treatment by considering a broader range of variables. This improves trust and transparency for both insurers and customers.

iii. **To Identify Key Factors Influencing Premiums:** The project aims to determine which personal and medical features (e.g., age, BMI, smoking habits) most affect insurance costs. Understanding these variables is essential for building accurate predictive models. It also aids in designing fair and transparent pricing strategies.

iv. **To Enhance Prediction Accuracy Using Real-World Data:** By training on actual insurance datasets, the model aims to reflect real-life trends and conditions. This ensures the predictions are not just theoretical but practically applicable. Accurate predictions are crucial for risk management and pricing strategies.

v. **To Address Challenges in Regional and Lifestyle Variations:** Health expenses vary by region, and lifestyle habits influence risk profiles. The model considers these factors to generate more personalized premium estimates. It ensures location and habits are fairly integrated into pricing.

By achieving these objectives, we can enhance the efficiency, fairness, and accuracy of premium prediction models. By leveraging machine learning, particularly XGBoost, the project offers a data-driven approach to modern insurance challenges. The outcomes can support both insurers and policyholders in making informed decisions.

1.6 Existing System

In traditional health insurance systems, premium prediction is primarily based on actuarial methods and linear statistical models. These approaches rely heavily on historical data and predefined risk factors such as age, gender, and pre-existing conditions. While effective in the past, these models often fail to capture complex, non-linear relationships between multiple health indicators and lifestyle choices.

Moreover, manual intervention and rule-based systems dominate the premium calculation process. Underwriters assess risk based on medical reports, questionnaires, and sometimes even in-person interviews. This makes the process time-consuming and prone to human bias. It also lacks personalization, often resulting in either overestimated or underestimated premium amounts for individuals.

The existing systems also struggle to incorporate real-time or dynamic data such as changes in lifestyle, habits, or recent medical diagnoses. This static nature of conventional models makes them less adaptive to changes in an individual's health profile. As a result, insurance pricing remains rigid and may not reflect the actual risk involved, limiting its fairness and efficiency.

Another major drawback of the existing systems is their limited use of advanced data analytics and automation. Most traditional models do not leverage large datasets or modern computational tools to improve prediction quality. This restricts the scalability of the system, especially when dealing with a growing customer base and diverse medical histories. The absence of adaptive learning also means that the system cannot improve over time, leading to stagnation in pricing strategies. These limitations highlight the need for integrating intelligent, data-driven models like XGBoost in modern insurance systems.

Furthermore, scalability poses a major challenge in existing premium prediction systems. As insurance companies handle increasingly large datasets, traditional statistical methods become inefficient and slow. These models often lack the computational power and flexibility to process complex, high-volume data in real-time. This is especially problematic for large insurers managing millions of policyholders.

Without automation and intelligent algorithms, adapting to evolving healthcare patterns and personalized data becomes difficult. As a result, the system's responsiveness and predictive accuracy significantly decline.

1.7 Scope of the Project

The scope of this project is centered on developing a predictive model for estimating medical health insurance premiums using the Extreme Gradient Boosting (XGBoost) algorithm. The model aims to consider various personal and health-related attributes such as age, BMI, smoking status, number of dependents, and geographical region. It focuses on improving the accuracy and fairness of premium pricing compared to traditional actuarial and linear models.

This project is limited to structured datasets and does not incorporate real-time or unstructured data sources like electronic health records or wearable device outputs. However, the approach is scalable and adaptable for future integration with such data. The scope also includes performance evaluation, comparison with other models, and interpretation of key features influencing premium costs.

2 Literature Review

Linear Regression has been a foundational approach in predicting continuous outcomes like insurance costs. It models the linear relationship between input variables such as age, BMI, and smoking status and the insurance premium. Logistic Regression, on the other hand, has been used for classification tasks like claim approval or assessing risk categories. Despite their interpretability and simplicity, these models often fail to capture non-linear interactions between features. Frees and Valdez (2008) found that although these techniques provide a good baseline, they perform poorly on heterogeneous healthcare datasets [1].

Decision Trees are rule-based models that split the dataset into branches based on feature conditions, leading to predictions at the leaves. They are highly interpretable and easy to implement, making them useful in decision-making environments like insurance pricing. However, overfitting and instability in tree structures are common drawbacks. A study by Weiss and Indurkhya (1995) demonstrated that decision trees could model risk and pricing but performed inconsistently on larger and more complex datasets [2].

Support Vector Machines (SVM) are well-suited for classification problems in insurance, such as identifying high-risk individuals. SVMs use hyperplanes to separate data points in high-dimensional space and have been effective when the feature count is large. Wu et al. (2007) applied SVM for predicting healthcare costs and found it reliable, though computationally expensive and sensitive to kernel selection. Its performance in regression scenarios (SVR) also showed moderate promise [3].

Random Forests, an ensemble of Decision Trees, improve predictive accuracy and reduce overfitting by combining results from multiple models. Bertsimas et al. (2018) showed that Random Forests outperformed individual decision trees and logistic regression in estimating insurance claims and healthcare costs [4]. Their strength lies in robustness and feature importance evaluation, though they may lack transparency for regulatory purposes.

Other ensemble methods like Bagging (Bootstrap Aggregating) have been applied to improve the stability and performance of base learners such as Decision Trees. Bagging helps reduce variance and increases generalization by training multiple models on different subsets of the data. These methods have demonstrated effectiveness in premium prediction tasks, particularly when data is noisy or has missing values. While not as advanced as boosting methods, they still offer meaningful improvements over single-model approaches [5].

Clustering algorithms like K-Means and Hierarchical Clustering have been used to group individuals based on health profiles and risk factors. These techniques help insurers identify customer segments with similar characteristics, such as age groups, health conditions, or spending behavior. For instance, Su et al. (2010) demonstrated the use of clustering in segmenting policyholders for targeted policy pricing and fraud detection. However, clustering is unsupervised and does not directly predict premium amounts, limiting its utility in exact cost estimation. It is often used as a preprocessing step to improve the performance of supervised learning models. Despite these limitations, clustering provides valuable insights for market segmentation and strategic decision-making [6].

3 Proposed System

The implementation of the proposed system follows a structured and sequential workflow to ensure model efficiency and accuracy. Each phase—from initial planning to final deployment—is aligned to contribute effectively to the prediction model's development. Proper segmentation of tasks made it easier to identify milestones and address challenges at each step. The methodology ensures both technical robustness and real-world applicability. With careful attention to each task, the project maintains consistency, usability, and precision throughout. The following subsections provide a breakdown of these essential development stages.

3.1 Task

The tasks involved in the project span multiple stages of development, starting from planning and ending with the final documentation. Each task was essential in building a reliable and efficient prediction system. Below is a breakdown of each stage and its contribution to the overall project.A task-oriented approach helped us maintain focus throughout the project lifecycle. By breaking the problem into well-defined phases, we could effectively manage our time and resources while monitoring progress. Each sub-task was interlinked and built upon the outcomes of previous stages.

3.1.1 Project Planning and Setup:

The initial phase involved defining the project scope, identifying the problem statement, and selecting the appropriate tools and technologies. We decided on using Python and Jupyter Notebook for model development, along with libraries such as Pandas, Scikit-learn, and XGBoost. The project goals were clearly established to ensure clarity and direction throughout development.

3.1.2 Data Collection and Preparation:

The dataset used in this project was sourced from publicly available medical insurance records and included key demographic and health-related attributes. Before training the model, the data underwent cleaning to handle missing values and remove outliers that could affect prediction quality. Essential features like age, sex, BMI, smoking status, number of children, and region were retained after ensuring their relevance.

An initial exploratory analysis was conducted to understand data distribution, spot trends, and uncover potential correlations. This process helped inform feature engineering steps that followed. The final cleaned dataset served as a strong foundation for building an accurate and reliable prediction model.

3.1.3 Feature Engineering:

Feature engineering focused on transforming data into a suitable format for the XGBoost model. Categorical variables such as 'region', 'sex', and 'smoker' were converted using one-hot encoding. Numerical features like BMI and age were normalized to ensure consistent scaling.

Additional features were derived from existing variables to improve model performance. Interactions between variables were explored to capture complex patterns. Multicollinearity was checked to avoid redundant data that could affect accuracy.

Irrelevant or less impactful features were excluded based on correlation analysis. These preprocessing steps ensured that the model was trained on clean, informative, and optimized inputs.

3.1.4 Model Selection and Development:

XGBoost was chosen as the core algorithm for this project due to its ability to efficiently process structured data, support regularization techniques, and deliver high predictive performance. Its gradient boosting framework allows it to build strong models from multiple weak learners, effectively handling complex patterns in the dataset. The model was initially trained on health and demographic variables such as age, BMI, smoking status, and region to estimate insurance charges. Parameter tuning began with default values guided by domain knowledge, followed by optimization through cross-validation techniques. XGBoost's ability to reduce overfitting using L1 and L2 regularization made it ideal for this regression task. Its scalability and fast execution further supported the need for real-time premium prediction.

3.1.5 Training and Evaluation:

The dataset was split into training and testing sets. The model was trained using the training set and evaluated using metrics like MAE, MSE, and R^2 on the testing set. Hyperparameter tuning was performed using Grid Search. Cross-validation was also employed to ensure that the model generalized well across unseen data.

3.1.6 Integration and Deployment:

Once the XGBoost model was trained and evaluated, it was integrated into a simple yet functional web interface using **Flask**, a lightweight Python web framework. This allowed users to enter inputs like age, BMI, smoking status, and receive predicted insurance charges. The interface was designed to be user-friendly and accessible even to non-technical users. This step made the model practical for real-world usage scenarios.

The deployment process ensured that the model could run efficiently on a server and respond to user queries in real-time. The backend handled preprocessing and prediction tasks seamlessly, ensuring consistency between training and prediction. This integration phase bridged the gap between development and user interaction. It marked the final step in making the model usable beyond theoretical evaluation.

3.1.7 Testing and Validation:

The deployed system was rigorously tested using a wide range of input combinations to ensure both accuracy and robustness. Various real-world and hypothetical scenarios were fed into the model, including normal cases and outliers, to evaluate its adaptability. Extreme values such as high BMI or advanced age were introduced to assess the model's reliability under unusual conditions. In addition to model evaluation, unit testing was conducted on each functional module to verify independent performance. User testing helped in refining the web interface, ensuring a smooth and intuitive user experience. The combined approach ensured that the system met both technical and usability standards.

3.1.8 Documentation and Reporting:

Each stage of the project was carefully documented to ensure clarity and traceability. Reports included methodology, code structure, and model evaluation results. Clear visuals and graphs were used to present findings effectively. Proper formatting and structure were maintained throughout the documentation.

In addition to technical content, the report also included references, acknowledgements, and appendices. All sources were cited appropriately to ensure academic honesty. The final report was reviewed for completeness, accuracy, and presentation quality. This comprehensive documentation supports future modifications and scalability.

3.2 Dataset

The dataset used in this project is sourced from a publicly available health insurance database consisting of 1,338 records and 7 key features. These features are selected based on their direct or indirect influence on insurance pricing. The dataset includes both numerical and categorical attributes, providing a rich combination of information for building a predictive model.The attributes include:

i. **Age**: Age of the policyholder.

ii. **Sex**: Gender of the individual.

iii. **BMI**: Body Mass Index.

iv. **Children**: Number of children covered by insurance.

v. **Smoker**: Smoking status (yes/no).

vi. **Region**: Residential area (northeast, northwest, southeast, southwest).

vii. **Charges**: The insurance charges (target variable).

The data was preprocessed and checked for missing values, which were not present. Feature encoding and normalization were applied where necessary. Visualization was performed to understand trends and relationships between variables such as BMI and smoking status with charges.

3.3 Input

The input to the model includes the following user-provided information:

i. **Age**: Represents the age of the individual applying for insurance.

ii. **Sex**: Indicates the gender of the individual.

iii. **BMI**: Indicates the body mass index, a key metric for assessing obesity.

iv. **Smoking Status**: Identifies whether the individual is a smoker or non-smoker.

v. **Number of Children**: Refers to the number of dependents covered under the policy.

vi. **Region**: Denotes the geographical area where the policyholder resides.

These features were selected based on their significant influence on the premium values. They were normalized and encoded appropriately to match the format

3.4 Output

The output of the system is a predicted health insurance premium, expressed as a numerical value. This prediction is generated after processing the user inputs through the trained XGBoost model. The result reflects an estimate of the insurance charges that an individual might incur based on factors like age, BMI, smoking status, and other demographic attributes.

This output serves as a decision-support tool for insurance providers, enabling them to quote premiums more accurately and fairly. It also benefits users by offering a clearer understanding of how their personal health and lifestyle choices influence the cost of their insurance. The prediction encourages awareness and transparency in premium calculation.

To enhance usability, the output is displayed in a clean and interpretable format through the deployed user interface. Users receive immediate feedback upon entering their details. Additionally, the model can be retrained with updated data to keep predictions in line with evolving health trends and pricing policies.

4 Algorithm

Extreme Gradient Boosting (XGBoost) is an optimized gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework and has become one of the most popular algorithms for structured/tabular data due to its speed and performance.

XGBoost is based on the principle of boosting, where multiple weak learners (typically decision trees) are combined to form a strong learner. The core idea is to fit a model on the residual errors of the previous models to gradually reduce the overall prediction error. This method uses gradient descent optimization to minimize a loss function, which, in our case, is the difference between predicted and actual insurance charges.

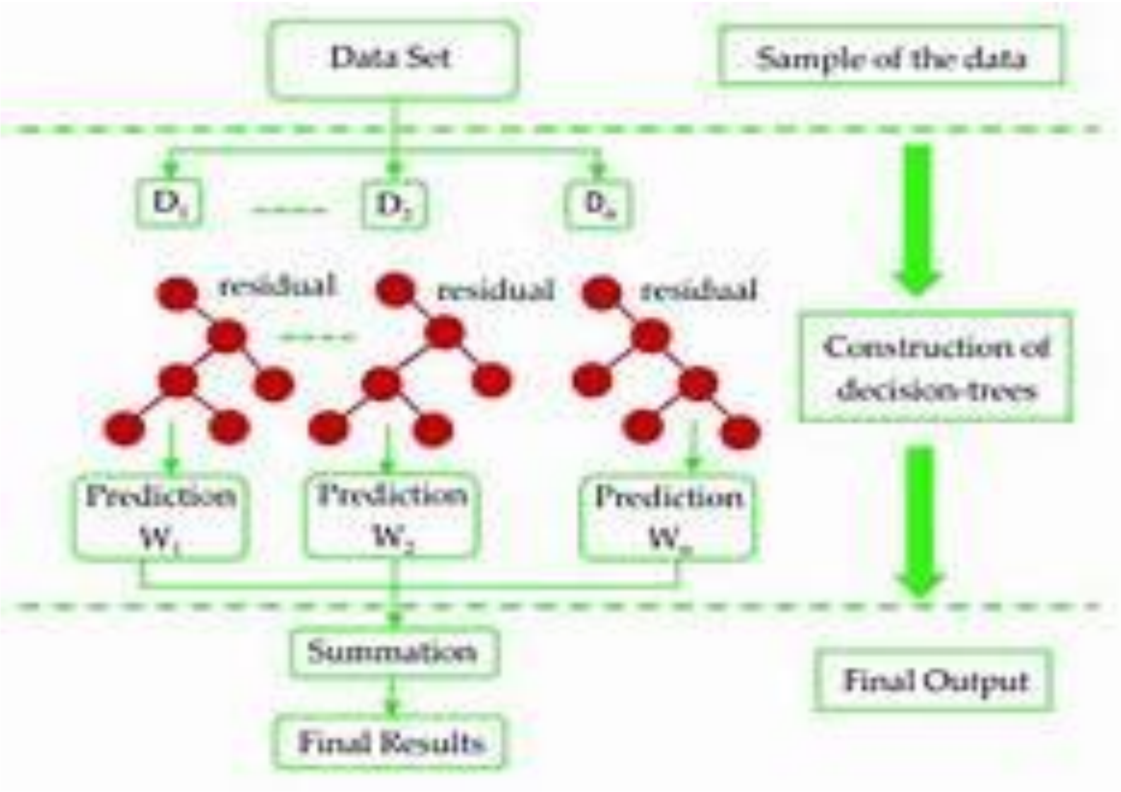


Figure 4.1 Working Process of XGBoost

Unlike traditional gradient boosting, XGBoost introduces several enhancements:

i. **Regularization (L1 and L2)**: Adds penalties to the complexity of the model (number of leaf nodes), reducing overfitting and improving model generalization.

ii. **Handling Missing Values**: Automatically learns the best way to handle missing data during training without needing imputation.

iii. **Tree Pruning**: Employs a max-depth parameter and a pruning approach that removes unnecessary branches, improving accuracy and efficiency.

iv. **Parallelization**: Unlike traditional boosting which is sequential, XGBoost constructs trees in parallel at each iteration, making it extremely fast.

v. **Weighted Quantile Sketch**: Allows the algorithm to handle large-scale datasets with high performance and low memory usage.

In our project, XGBoost is used as a regression model to predict the medical insurance premium costs. The features provided to the model include:

a. Age

b. Sex

c. BMI

d. Number of children

e. Smoking status

f. Region

These variables are preprocessed and passed into the model, which outputs the expected insurance charges based on learned patterns from historical data.

4.1 Mathematical Foundation

XGBoost minimizes the following regularized objective function:

$$\mathcal{L}(t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k)$$

Where:

- l is the loss function (e.g., Mean Squared Error for regression)
- $\hat{y}_i^{(t)}$ is the prediction of the i-th instance at the t-th iteration
- $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda||w||^2$ is the regularization term
- T is the number of leaves in the tree, and w is the vector of scores for each leaf

This objective ensures that the model is not only fitting the training data well but is also penalized for being overly complex.

4.2 Hyperparameters Tuned in Our Project

- **Learning rate (eta)**: Controls the contribution of each tree.
- **Max depth**: Limits the depth of each tree to prevent overfitting.
- **Number of estimators**: Total number of trees built.
- **Subsample**: Fraction of data used per iteration.
- **Colsample_bytree**: Fraction of features used for each tree.

4.3 Performance Metrics Used

- **Mean Absolute Error (MAE)**:

  MAE is the average of the absolute differences between predicted and actual values. It gives an idea of how far off predictions are from actual results, on average. Since it doesn't square the errors, it treats all errors equally, making it easier to interpret. A lower MAE indicates better model performance.

Formula:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- **Mean Squared Error (MSE):** MSE is the average of the squared differences between the predicted and actual values. Squaring gives more weight to larger errors, which helps identify models that make occasional large mistakes. MSE is more sensitive to outliers than MAE. A lower MSE reflects a better-performing model**.**

  Formula:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- **R-squared (R² Score):** R² measures how well the predicted values approximate the actual data. It represents the proportion of variance in the dependent variable that is predictable from the independent variables. An R² score close to 1 indicates a strong model, while a value near 0 suggests poor predictive power.

  Formula:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

These metrics were chosen to evaluate how close the model predictions are to the actual insurance charges. XGBoost demonstrated superior performance compared to baseline models like linear regression or decision trees.

4.4 Advantages of XGBoost Algorithm

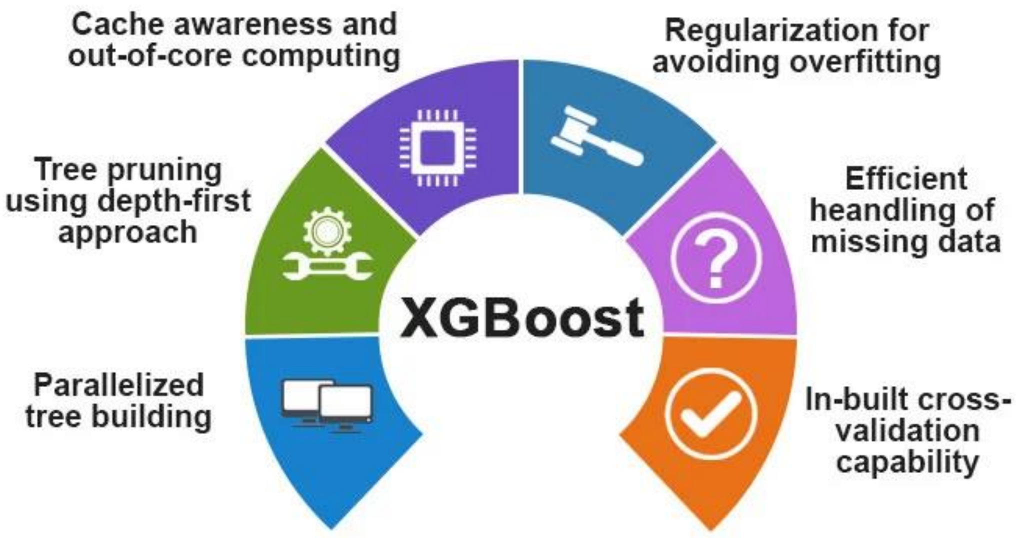

Cache awareness and out-of-core computing
Regularization for avoiding overfitting
Tree pruning using depth-first approach
Efficient heandling of missing data
XGBoost
Parallelized tree building
In-built cross-validation capability

Figure 4.2 Features of XGBoost

i. **Excellent Handling of Both Linear and Non-linear Relationships**

XGBoost is based on decision trees, which naturally handle both linear and complex non-linear patterns in data. This allows the model to capture intricate interactions between features that traditional linear models may miss. As a result, it performs well in real-world scenarios with diverse data types and distributions.

ii. **Built-in Regularization Improves Model Robustness**

XGBoost includes L1 (Lasso) and L2 (Ridge) regularization terms in its objective function. These techniques help prevent overfitting by penalizing overly complex models, ensuring that the model generalizes well to unseen data. This makes XGBoost particularly effective for datasets with noise or irrelevant features.

iii. **Fast Training Even on Large Datasets**

XGBoost is optimized for speed and performance. It uses advanced techniques such as parallel processing, tree pruning, and cache-aware access patterns, which significantly reduce training time. Even with large datasets, it maintains high efficiency without compromising accuracy.

iv. **Well-Supported with Detailed Documentation and Community Resources** XGBoost is widely used and maintained, offering rich documentation, tutorials, and examples. It has a strong community of developers and users, making it easier to troubleshoot issues, learn new techniques, and stay up-to-date with best practices.

4.5 Feature Importance Analysis

In this project, feature importance analysis was performed to understand which input features most significantly impact the model's prediction of health insurance charges. XGBoost, being a tree-based ensemble algorithm, inherently provides feature importance scores based on how frequently and effectively each feature is used in splitting the data during training. These importance scores help in identifying the key driving factors behind the model’s decisions, thereby enhancing transparency and interpretability. Understanding these contributions is crucial for validating the model and ensuring that its behavior aligns with domain knowledge in healthcare and insurance.

The results from the feature importance analysis revealed that 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.

To further illustrate these findings, a feature importance plot was generated using XGBoost’s built-in visualization tools. The bar graph clearly depicted the comparative importance of all input features, with smoker, age, and bmi standing out at the top. This analysis not only confirmed that the model was learning meaningful patterns from the data but also provided actionable insights. For example, insurance companies can use this information to better assess risk factors and tailor their pricing strategies accordingly. Overall, feature importance analysis adds depth to model evaluation by combining performance with interpretability.

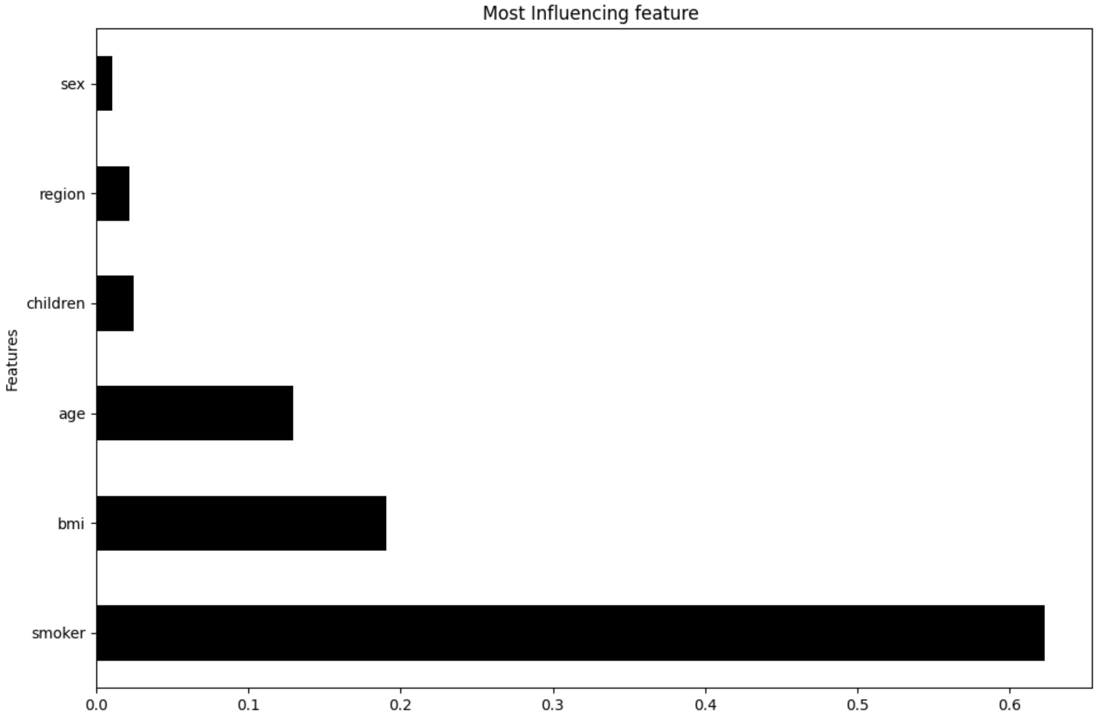


Figure 4.3 Feature Importance

4.6 Comparision with Other Algorithms

To evaluate the effectiveness of the XGBoost model, we compared its performance with other commonly used machine learning algorithms including Linear Regression, Decision Tree Regression, and Random Forest Regression. All models were trained on the same dataset with identical preprocessing steps to ensure a fair comparison. The evaluation was based on several performance metrics such as R² Score, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) on both the training and testing datasets.

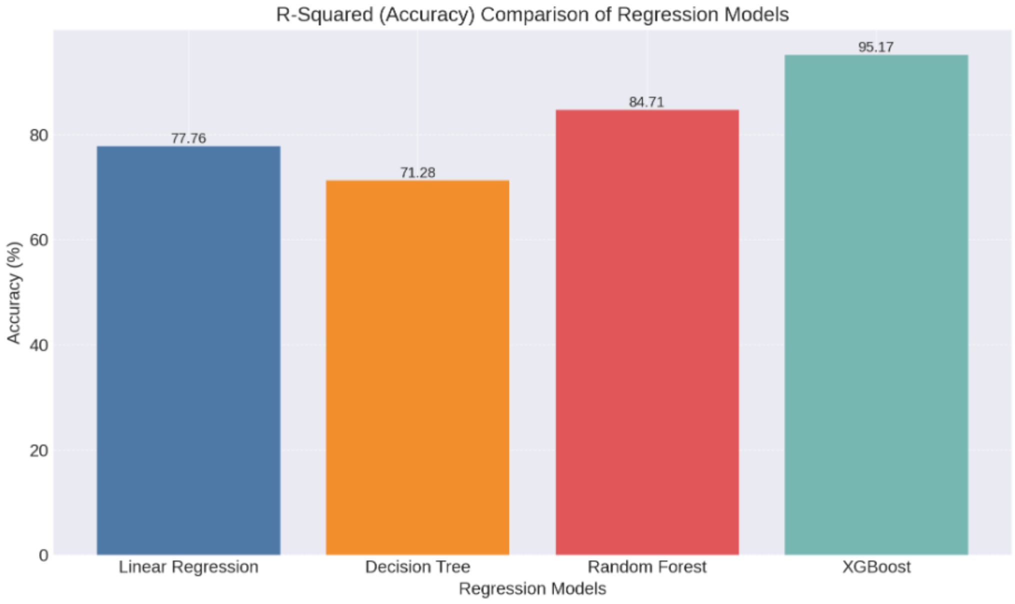


Figure 4.4 Comparision of Accuracy of Algorithms

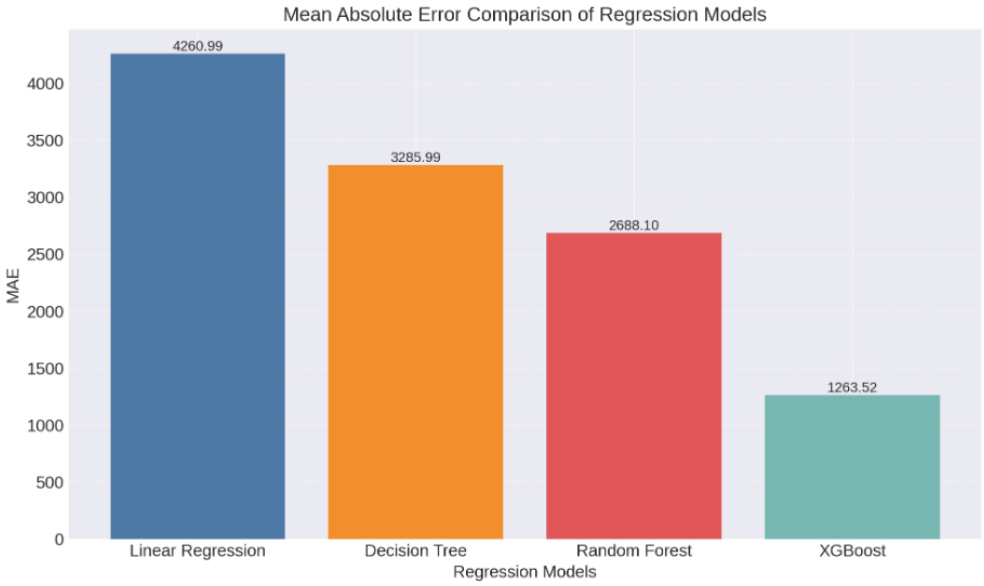


Figure 4.5 Mean Absolute Error of Algorithms

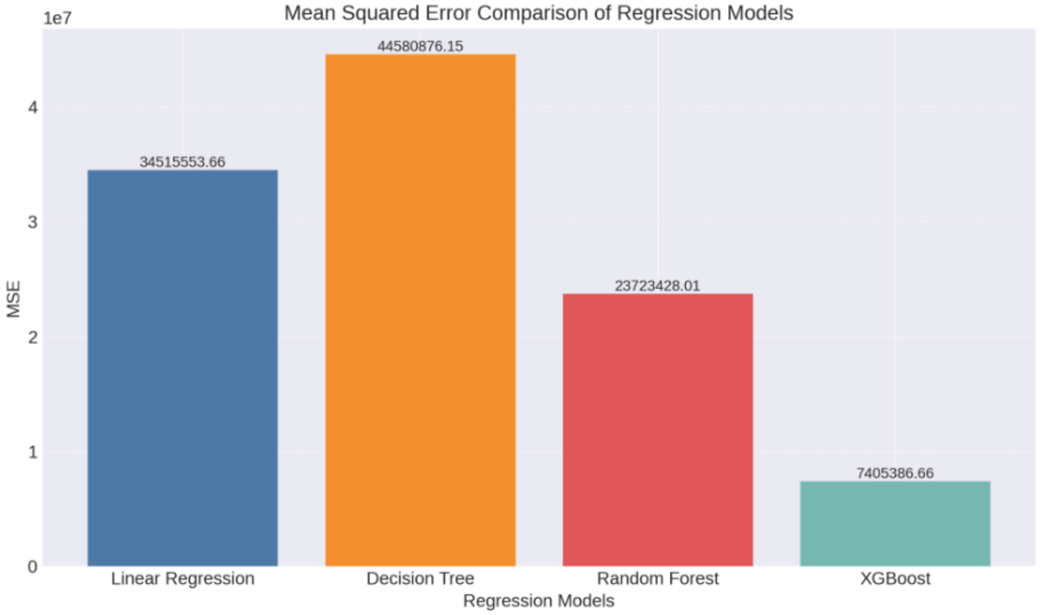


Figure 4.6 Mean Squared Error of Algorithms

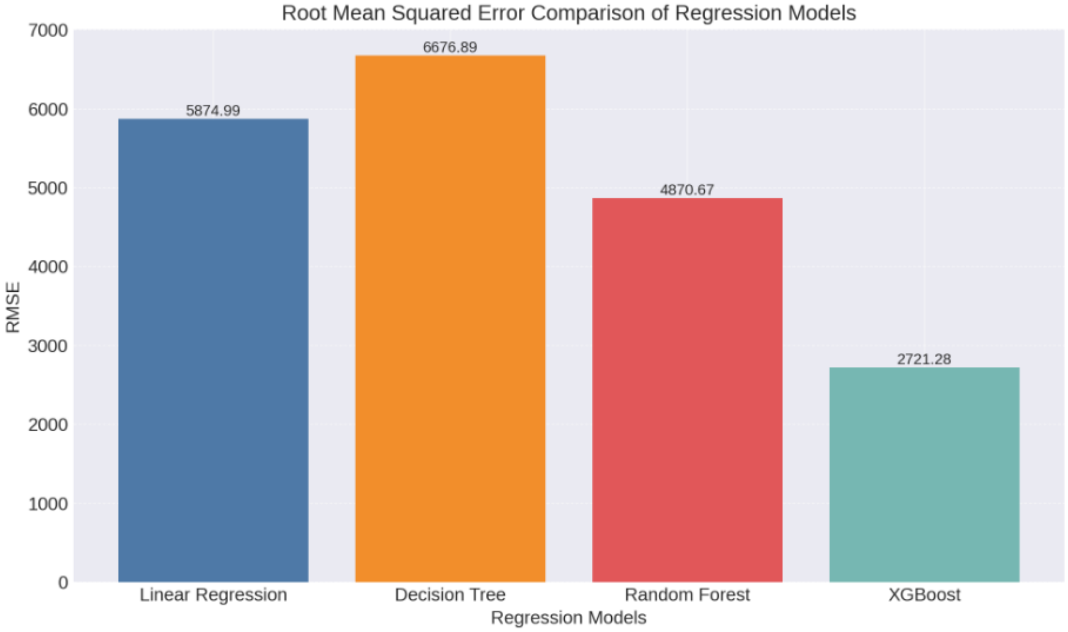


Figure 4.7 RMSE of Algorithms

From the results, it was evident that XGBoost outperformed all other models. While Linear Regression showed the lowest performance due to its assumption of linearity in data, Decision Tree and Random Forest performed better but still fell short in terms of generalization and accuracy. After hyperparameter tuning, XGBoost achieved an impressive R² Score of 99.71% on training data and 95.18% on testing data, indicating strong predictive power with minimal overfitting. In comparison, Random Forest followed next with good performance, while Decision Tree showed signs of overfitting, and Linear Regression had the weakest results due to its inability to capture complex relationships in the data.

This comparison highlighted the superior capability of XGBoost in handling non-linear patterns, managing overfitting through regularization, and delivering high accuracy. These advantages make XGBoost the most suitable choice for our insurance price prediction problem, where capturing complex interactions among features is crucial for accurate estimation.

Table 4.1 Comparing R2 Score of algorithms

Algorithm	**R2 Score**
Linear Regression	72.90%
Decision Tree	85.69%
Random Forest	87.28%
Support Vector Regression(SVR)	83.21%
XGBoost	95.18%

5 System Design

The System Design chapter provides a structural overview of the proposed prediction system, detailing how various components interact to deliver the final output. A well-planned system architecture ensures efficiency, modularity, scalability, and accuracy. This chapter outlines the architecture, user interfaces, data flow, and system modules of the application.

5.1 System Architecture

The system follows a modular architecture, where each component has a specific role:

i. **Frontend Interface**: A web-based input form allows users to enter information such as age, gender, BMI, number of children, smoking status, and region.

ii. **Backend Processing**: Inputs are processed and passed through preprocessing functions to ensure they match the model's expected format.

iii. **Prediction Engine**: The XGBoost model is loaded and used to generate predictions.

iv. **Output Interface**: The predicted insurance charge is displayed to the user in an intuitive format.

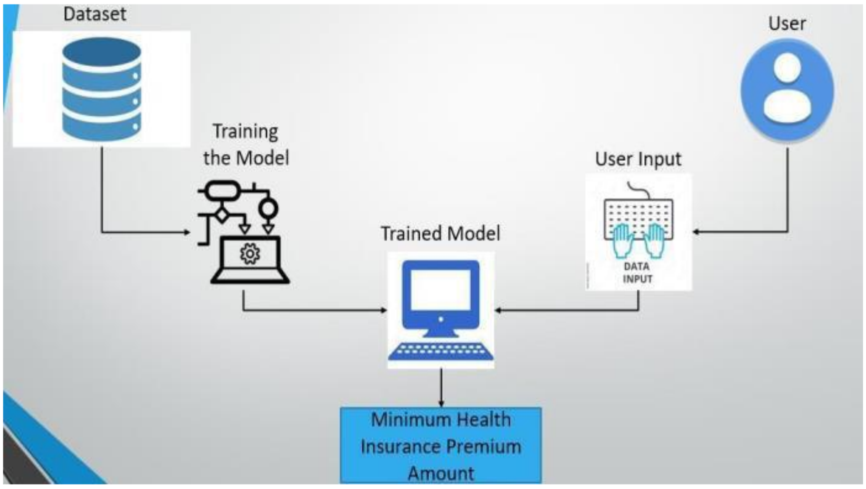


Figure 5.1 Architecture Diagram

5.2 Flow Diagram

The flow diagram represents the step-by-step process of predicting the minimum health insurance premium using the XGBoost algorithm. It starts with the dataset, from which relevant features are extracted to form a feature vector. These features are used as input data for the model. The XGBoost algorithm is then applied, and feature selection analysis is conducted to identify the most impactful features. Using these insights, the model is built and evaluated for accuracy. Finally, the trained model predicts the most suitable insurance premium for individuals based on their personal attributes. This systematic approach ensures accurate and reliable predictions.

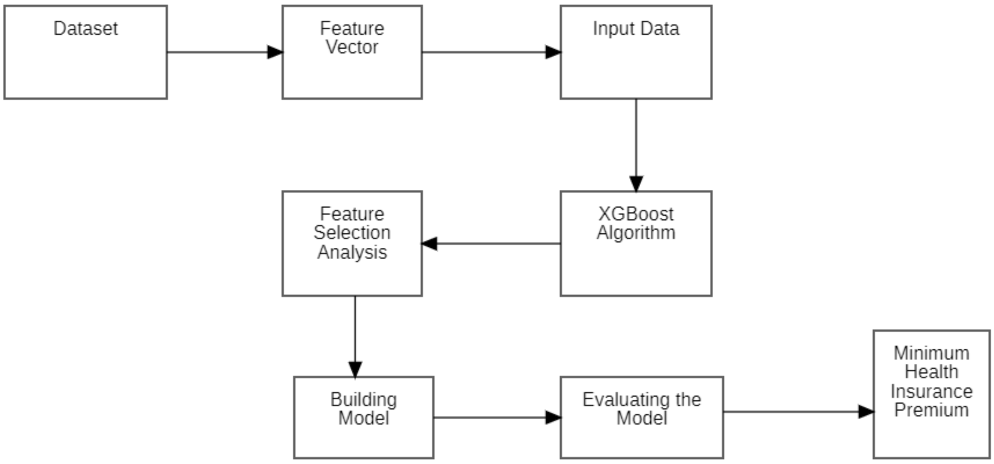


Figure 5.2 Flow Diagram

5.3 Sequence Diagram

The sequence diagram illustrates the interaction between the user and the system in a health insurance price prediction process. Initially, the user opens the interface, prompting the system to import the required libraries. The system then loads the dataset and performs necessary preprocessing steps. After that, training and testing of the model are carried out. Finally, the user enters specific constraints and submits them, upon which the system processes the input and returns the predicted result.

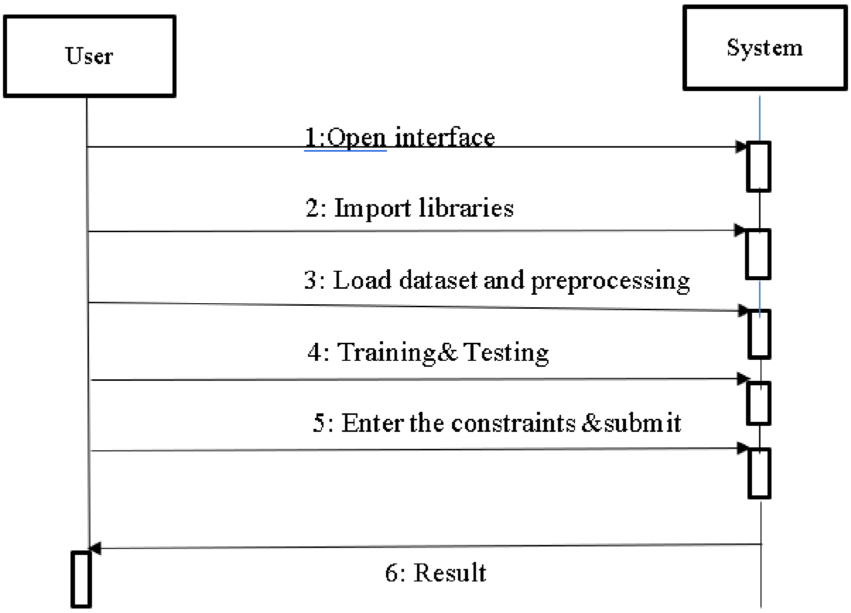


Figure 5.3 Sequence Diagram

5.4 Activity Diagram

The activity diagram represents the step-by-step process of a machine learning workflow. It begins with importing the dataset, followed by checking for any missing values. If missing data is found, it is filled appropriately before moving on. Once the data is complete, the dataset is analyzed and then split into training and testing sets. The next steps involve building the model and finally checking its accuracy score to evaluate its performance.

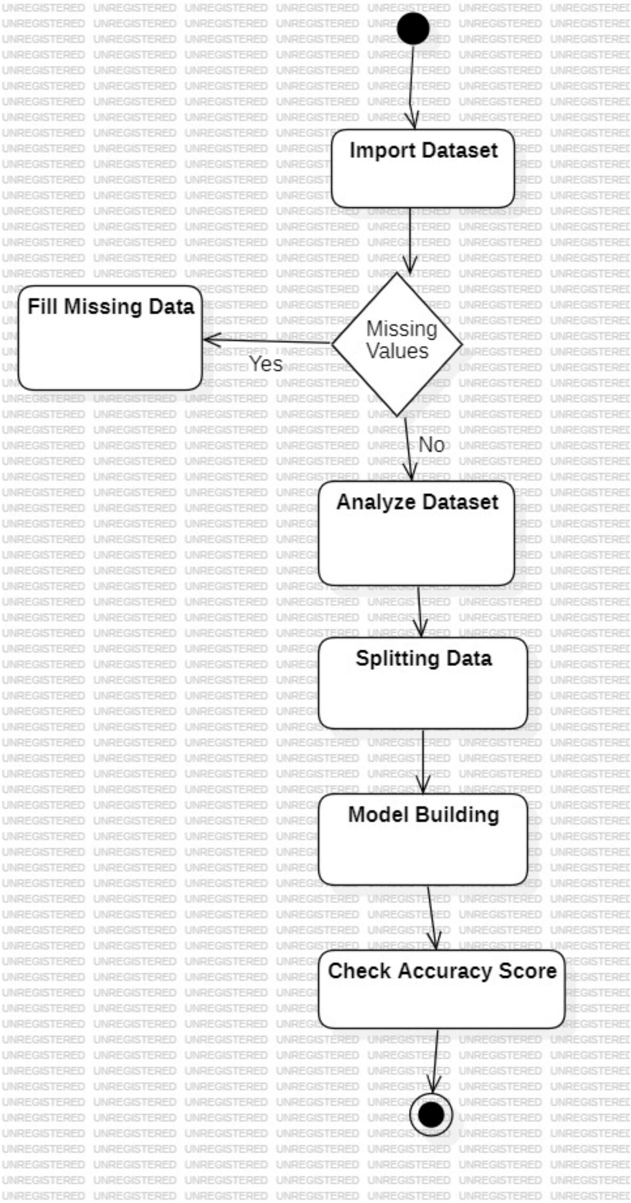


Figure 5.4 Activity Diagram

5.5 Use Case Diagram

A use case diagram represents the interaction between the user and the system in a simplified visual form. It highlights the various functionalities the system provides, such as importing data, preprocessing, training the model, and generating results. The user is shown as an actor who initiates these actions. Each action is represented as a use case within the system boundary. This diagram helps in understanding the system’s functionality from a user's perspective.

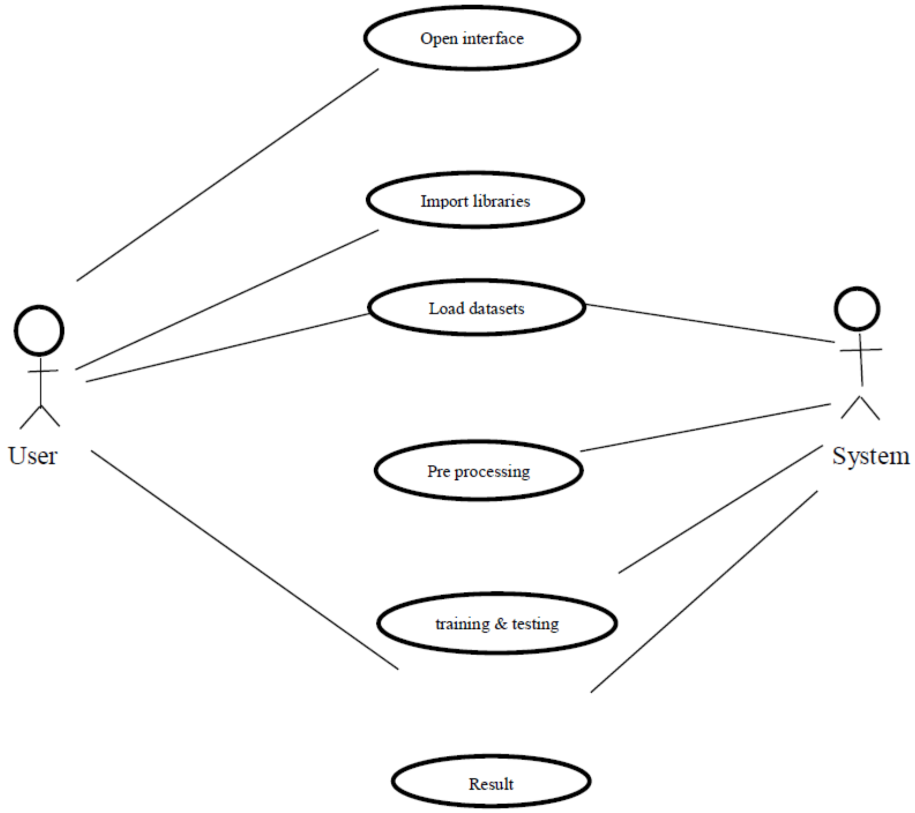


Figure 5.5 Use Case Diagram

5.6 Class Diagram

A class diagram represents the static structure of a system by showing its classes, attributes, methods, and the relationships between them. It defines how different entities in the system interact with each other. Each class is shown as a box divided into sections for the class name, attributes, and operations. Relationships like inheritance, association, and dependency are illustrated using connecting lines. This diagram helps in understanding the overall architecture and design of the system.

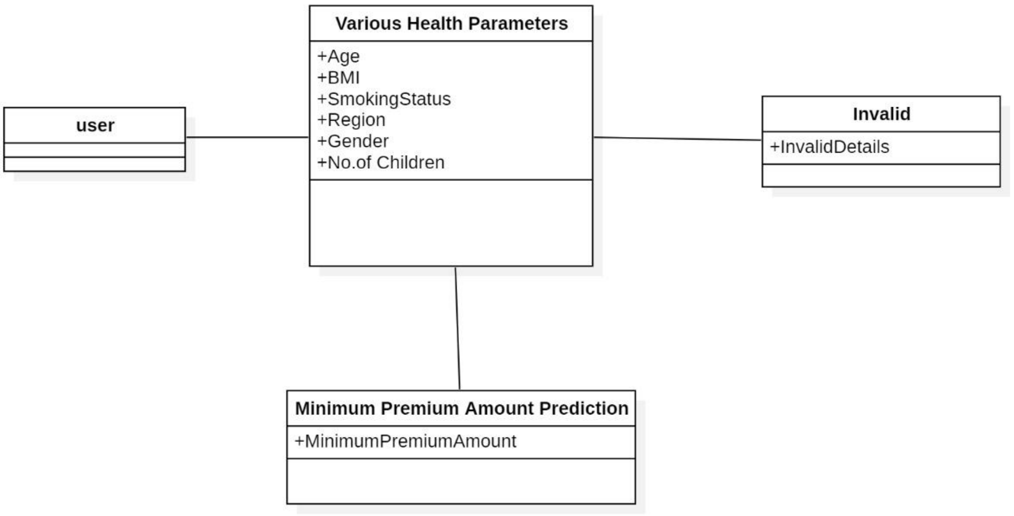


Figure 5.6 Class Diagram

6 Implementation

We developed a robust system for predicting health insurance prices using the XGBoost algorithm. By analyzing various customer attributes such as age, BMI, smoking status, and region, the model accurately estimates insurance premiums. The implementation process involved data preprocessing, model training, evaluation using performance metrics, and hyperparameter tuning to enhance accuracy. This approach enables insurance providers to set fair and data-driven premium rates. The system ensures efficient and reliable premium prediction, contributing to better decision-making in the insurance domain.

6.1 Requirements

Before building and deploying the Health Insurance Price Prediction model using XGBoost, it is essential to ensure that all necessary requirements are in place. These requirements include both hardware and software components that support data preprocessing, model training, evaluation, and visualization. Having the right setup helps ensure smooth development, better model performance, and efficient handling of data. This section outlines the key hardware specifications and software tools needed to successfully implement the project.

6.1.1 Hardware Requirements

The hardware requirements define the essential physical components needed to develop, train, and test the Health Insurance Price Prediction system effectively. Since the model involves data preprocessing, machine learning algorithm implementation, and evaluation, a reliable computing environment with sufficient processing power and memory is necessary. Adequate hardware ensures smooth execution of tasks such as training the XGBoost model, handling large datasets, and generating accurate predictions without performance issues.

i. RAM (min 16GB)
ii. iHard Disk (min 128GB)
iii. CPU (i5 Processor)
iv. X64 based Processor.
v. 64-bit operating system

6.1.2 Software Requirements

The software requirements specify the tools, platforms, and programming environments essential for developing and deploying the Health Insurance Price Prediction system. These include programming languages, libraries, frameworks, and other supporting software that facilitate data handling, model training, and evaluation. The right combination of software ensures efficient development, smooth integration, and accurate predictions using the XGBoost algorithm. Additionally, software tools assist in visualizing results and improving the overall functionality of the system.

i. **Operating System**: Windows 10 / Linux / macOS
ii. **Programming Language**: Python 3.8 or above
iii. **Development Environment**: Jupyter Notebook / VS Code / PyCharm
iv. **Web Browser**: Chrome / Firefox (for any web-based interfaces or visualizations)
v. **Anaconda** for simplified package management and environment setup

6.1.3 Libraries

i. **NumPy**

NumPy (Numerical Python) is a core library for numerical computations. It provides support for large multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on them efficiently.

ii. **Pandas**

Pandas is a powerful data manipulation and analysis library. It provides data structures like DataFrames that make it easy to handle structured data, perform cleaning, transformation, and aggregation tasks.

iii. **Matplotlib**

Matplotlib is a plotting library used to create static, animated, and interactive visualizations in Python. It was used in the project for visualizing data distributions and evaluation results.

iv. **Seaborn**

Built on top of Matplotlib, Seaborn is used for making statistical graphics. It simplifies the process of creating attractive and informative visualizations, such as heatmaps and pair plots.

v. **Scikit-learn (sklearn)**

Scikit-learn is a comprehensive machine learning library. It was used for data preprocessing, splitting datasets, training models, evaluating performance, and performing hyperparameter tuning through tools like RandomizedSearchCV.

vi. **XGBoost**

XGBoost is an optimized and scalable implementation of gradient boosting. It was the primary algorithm used in this project for building the insurance price prediction model due to its superior speed and performance.

vii. **Math**

The math library provides access to mathematical functions like square root, used in the calculation of evaluation metrics such as RMSE (Root Mean Squared Error).

6.2 Code

This chapter elaborates on the step-by-step coding process involved in building and evaluating the Optimizing Data-Driven Healthcare Cost Prediction model using the XGBoost algorithm.

6.2.1 GitHub Link

https://github.com/saivardhanmanikala/Medical_Insurance.git

The above repository contains the complete source code, dataset, and related files for this project through the above GitHub repository link. The repository is well-structured and includes separate files for data preprocessing, model training, hyperparameter tuning, and evaluation.

6.2.2 Importing Required Libraries

The implementation begins with importing essential libraries that support data handling, visualization, machine learning, and web development.

i. **Pandas** and **NumPy**: For handling structured data and numerical operations.

ii. **Matplotlib** and **Seaborn**: For generating informative visualizations.

iii. **Scikit-learn**: For data preprocessing, model training, evaluation, and hyperparameter tuning.

iv. **XGBoost**: For building the regression model due to its efficiency and predictive power.

v. **Flask**: A lightweight Python framework used to build the user-facing web interface.

These libraries enable efficient data flow, model construction, and deployment.

6.2.3 Loading and Exploring the Dataset

The dataset is loaded using the pandas.read_csv() function. Initial exploration is carried out using:

i. **Head()** to preview data,

ii. **info()** to understand the structure and data types,

iii. **describe()** to view summary statistics.

Exploratory Data Analysis (EDA) is performed using plots and correlation matrices to detect patterns, outliers, or imbalances in the data. This step ensures a clear understanding of the dataset and informs preprocessing strategies.

6.2.4 Data Preprocessing

The raw data is processed to prepare it for modeling:

i. **Missing Value Handling**: Although this dataset typically contains no missing values, null checks are performed as a standard step.

ii. **Categorical Encoding**: Features like sex, smoker, and region are encoded using Label Encoding to convert them into numeric format.

iii. **Feature Formatting**: All inputs are checked for consistency, and unnecessary columns, if any, are removed.

iv. **Data Scaling (Optional)**: Since XGBoost handles unscaled data effectively, this step is usually skipped unless outliers dominate.

These transformations ensure the data is clean, structured, and suitable for machine learning.

6.2.5 Splitting the Dataset

After preprocessing, the dataset is split into features (X) and the target variable (y, i.e., insurance charges). The data is then divided into:

i. **Training Set** (typically 80%): Used for model learning.

ii. **Testing Set** (typically 20%): Used for model validation.

The train_test_split() function from sklearn.model_selection is used for this purpose, ensuring the model is trained and tested on separate data.

6.2.6 Model Building Using XGBoost

An instance of XGBRegressor is created to perform the regression task. Initially, the model is trained using default parameters to establish a baseline performance and understand its initial accuracy. XGBoost, the core algorithm used, constructs an ensemble of decision trees in a sequential manner, where each new tree attempts to correct the errors made by the previous ones. This boosting technique helps in minimizing prediction errors over iterations.

The model is highly efficient and is capable of capturing both linear and non-linear relationships in the data, making it suitable for complex regression tasks like insurance price prediction.This step establishes a predictive model with high accuracy and efficiency.

6.2.7 Hyperparameter Tuning

To enhance model performance, RandomizedSearchCV is employed for hyperparameter optimization:

i. A parameter grid is defined for attributes like n_estimators, max_depth, learning_rate, gamma, subsample, and colsample_bytree.

ii. Cross-validation is used to evaluate each combination of parameters.

iii. The best configuration is selected based on the R² score.

This step helps in minimizing error and increasing model generalization by avoiding overfitting.

6.2.8 Model Evaluation

Once trained, the model is evaluated using several metrics:

i. **R² Score**: Indicates the proportion of variance explained by the model.

ii. **Mean Absolute Error (MAE)**: Represents the average error magnitude.

iii. **Root Mean Squared Error (RMSE)**: Highlights larger errors more significantly.

These metrics collectively give a robust view of the model's predictive accuracy and performance.

6.2.9 Saving the Model

After the model has been trained and evaluated, it is essential to save it so that it can be reused for making predictions without retraining. This step ensures that the model is preserved in its best-performing state and can be easily integrated into a deployment environment. In this project, the pickle library is used to serialize and save the trained XGBoost model into a file named xgb_model.pkl. Saving the model significantly reduces computation time during deployment, as the model can be quickly loaded and used to predict insurance charges based on user inputs. This saved file is later imported in the Flask application to serve predictions through the web interface.

6.2.10 Creating the Front-End Using Flask

To make the model interactive and accessible to users, a web application is developed using Flask. A simple HTML form is designed to capture user input for features such as age, sex, BMI, number of children, smoker status, and region. Once the form is submitted, the data is sent to the Flask backend, where it is preprocessed and passed to the trained XGBoost model for prediction.

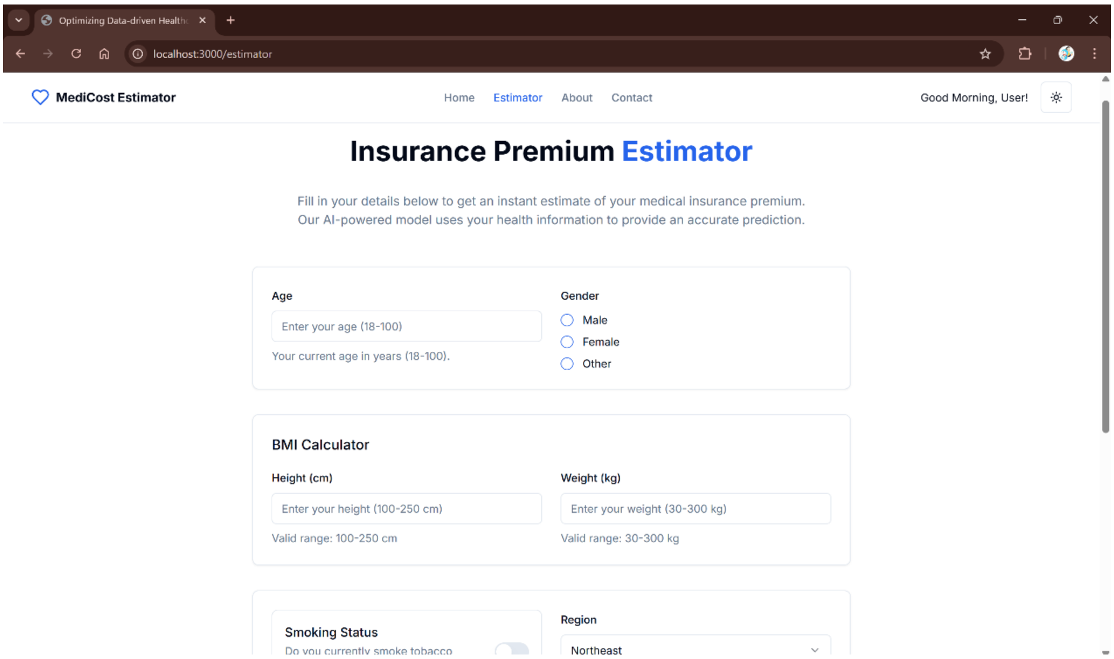
The model processes the input and returns the estimated insurance charges, which are then displayed to the user through the web interface, enabling real-time and user-friendly interaction.This front-end interface enables users to receive real-time predictions through a browser, making the application user-friendly and practical.

7 Results

This chapter presents the results obtained from the XGBoost regression model used for predicting health insurance charges. The model's performance was evaluated using metrics like R² Score, MAE, and RMSE. After training and tuning, the model demonstrated reliable prediction accuracy. A web interface was also created using Flask to display real-time outputs. The following sections highlight the evaluation and prediction outcomes in detail.

7.1 User Interface

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.



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

Enter your age (18-100)

Your current age in years (18-100).

Gender

○ Male

○ Female

○ Other

BMI Calculator

Height (cm)

Enter your height (100-250 cm)

Valid range: 100-250 cm

Weight (kg)

Enter your weight (30-300 kg)

Valid range: 30-300 kg

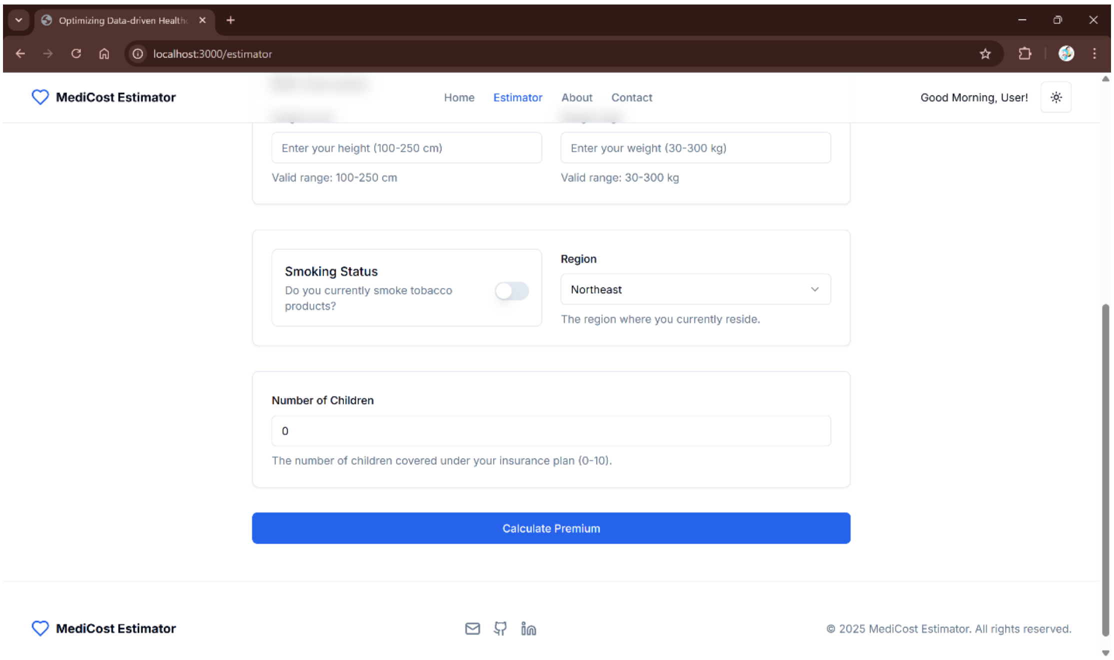
Smoking Status

Do you currently smoke tobacco

Region

Northeast

Figure 7.1 User Interface(1)



Enter your height (100-250 cm)

Valid range: 100-250 cm

Enter your weight (30-300 kg)

Valid range: 30-300 kg

Smoking Status

Do you currently smoke tobacco products?

Region

Northeast

The region where you currently reside.

Number of Children

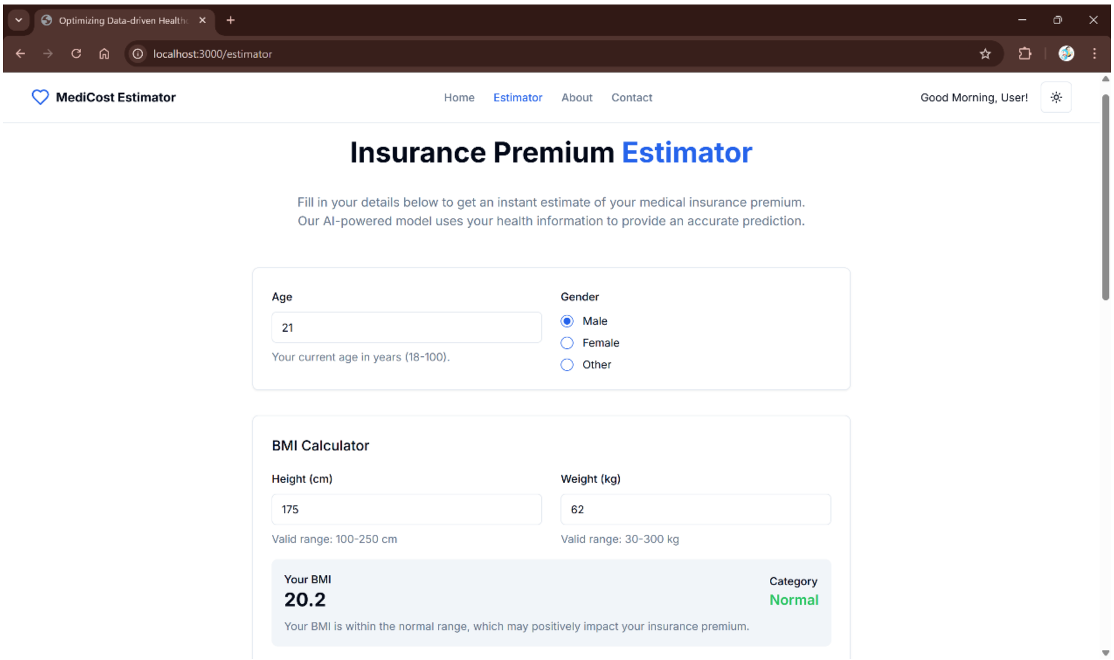
0

The number of children covered under your insurance plan (0-10).

© 2025 MediCost Estimator. All rights reserved.

Figure 7.2 User Interface(2)

7.2 Ouput when inputs are given



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: 21

Your current age in years (18-100).

Gender

● Male
○ Female
○ Other

BMI Calculator

Height (cm): 175

Valid range: 100-250 cm

Weight (kg): 62

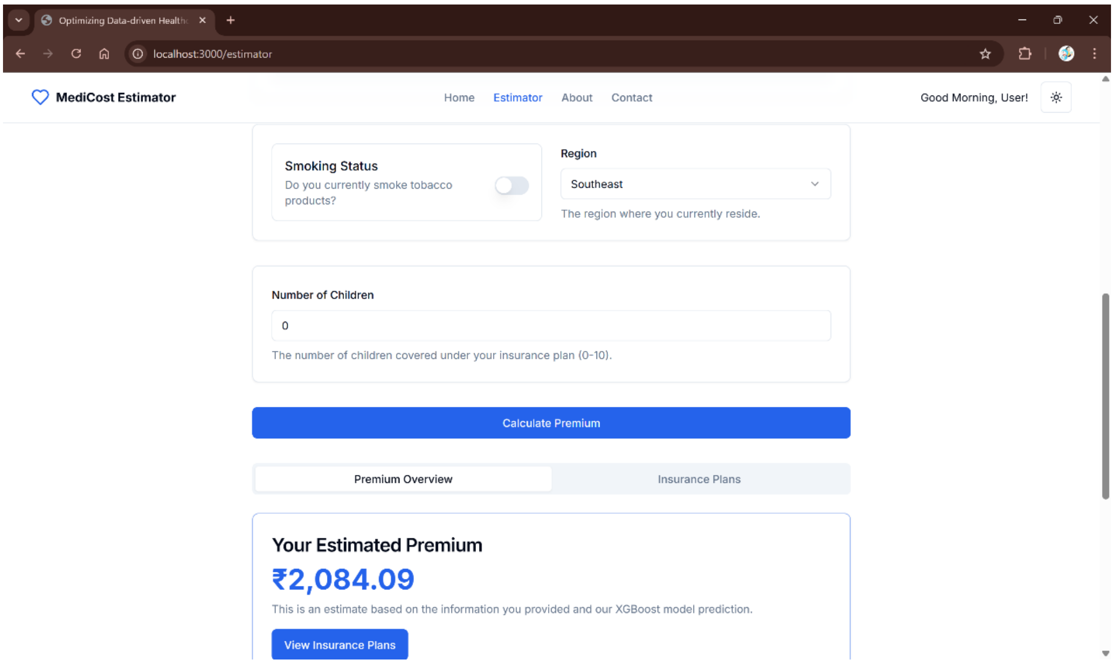
Valid range: 30-300 kg

Your BMI 20.2

Category Normal

Your BMI is within the normal range, which may positively impact your insurance premium.

Figure 7.3 Some inputs are given



Smoking Status

Do you currently smoke tobacco products?

Region: Southeast

The region where you currently reside.

Number of Children: 0

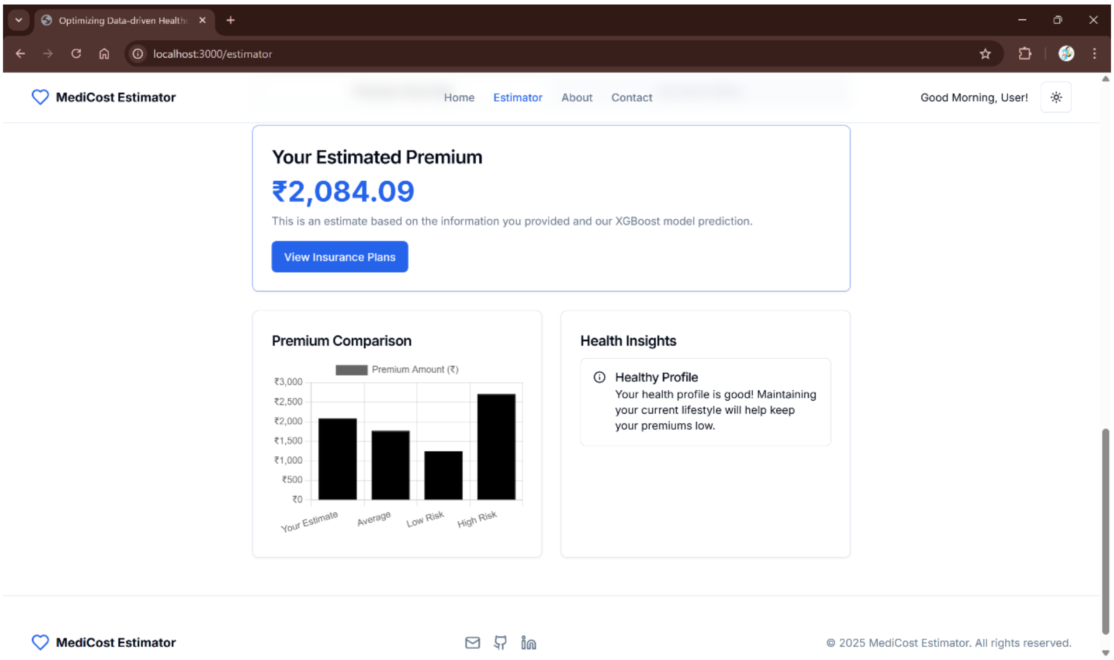
The number of children covered under your insurance plan (0-10).

Your Estimated Premium

₹2,084.09

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

Figure 7.4 Remaining inputs are given and calculate premium



Your Estimated Premium

₹2,084.09

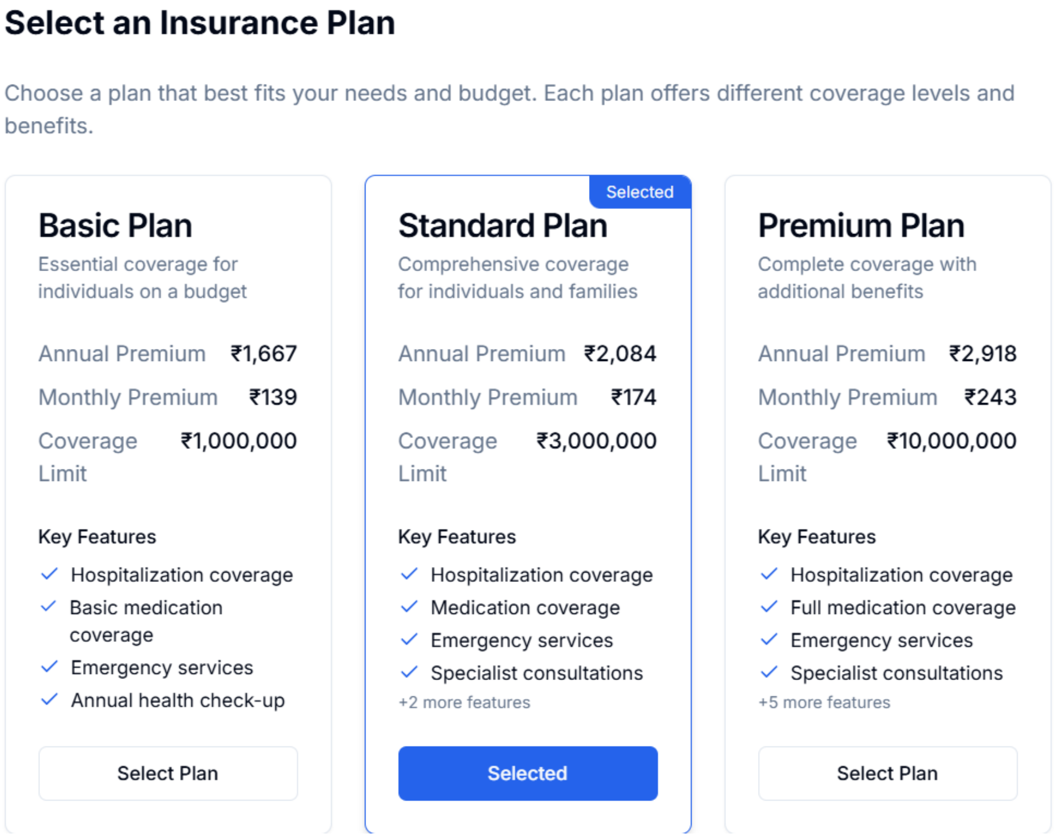
This is an estimate based on the information you provided and our XGBoost model prediction.

Health Insights

Healthy Profile
Your health profile is good! Maintaining your current lifestyle will help keep your premiums low.

© 2025 MediCost Estimator. All rights reserved.

Figure 7.5 Displaying premium



Select an Insurance Plan

Choose a plan that best fits your needs and budget. Each plan offers different coverage levels and benefits.

Basic Plan
Essential coverage for individuals on a budget

Annual Premium	₹1,667
Monthly Premium	₹139
Coverage Limit	₹1,000,000

Key Features
- ✓ Hospitalization coverage
- ✓ Basic medication coverage
- ✓ Emergency services
- ✓ Annual health check-up

Standard Plan
Comprehensive coverage for individuals and families

Annual Premium	₹2,084
Monthly Premium	₹174
Coverage Limit	₹3,000,000

Key Features
- ✓ Hospitalization coverage
- ✓ Medication coverage
- ✓ Emergency services
- ✓ Specialist consultations

+2 more features

Premium Plan
Complete coverage with additional benefits

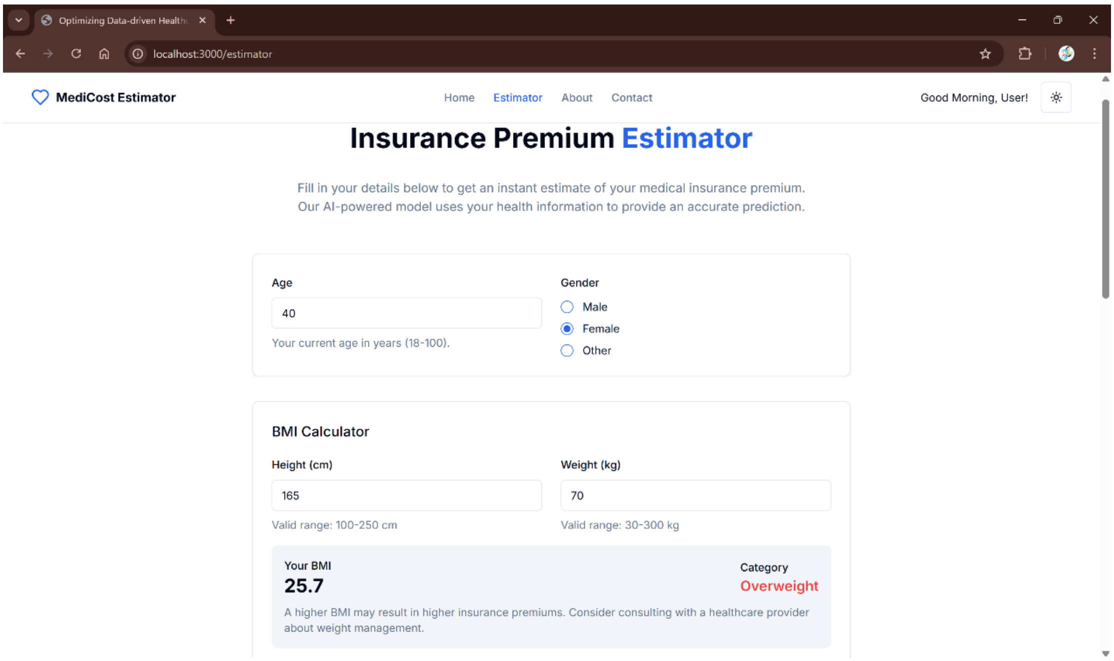
Annual Premium	₹2,918
Monthly Premium	₹243
Coverage Limit	₹10,000,000

Key Features
- ✓ Hospitalization coverage
- ✓ Full medication coverage
- ✓ Emergency services
- ✓ Specialist consultations

+5 more features

Figure 7.6 Suggesting plans

7.3 Output When another inputs are given



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 (18-100).

Gender

○ Male
● Female
○ Other

BMI Calculator

Height (cm)

165

Valid range: 100-250 cm

Weight (kg)

70

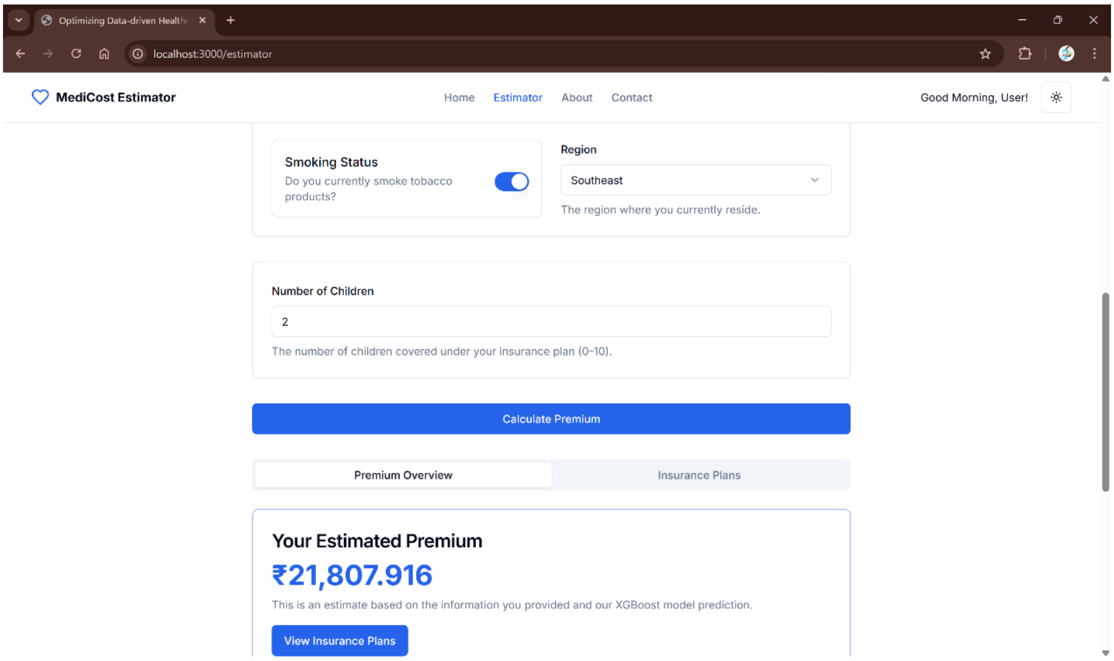
Valid range: 30-300 kg

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 7.7 Another inputs(1)



Smoking Status

Do you currently smoke tobacco products?

Region

Southeast

The region where you currently reside.

Number of Children

2

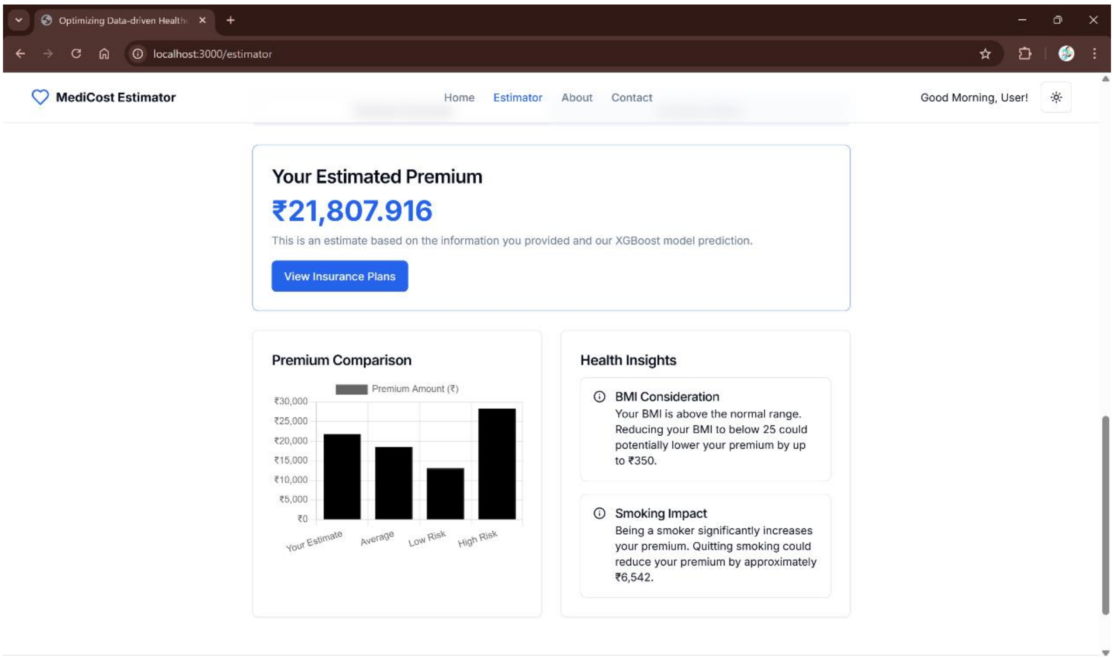
The number of children covered under your insurance plan (0-10).

Your Estimated Premium

₹21,807.916

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

Figure 7.8 Remaining inputs are given and calculate premium(2)



Your Estimated Premium

₹21,807.916

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

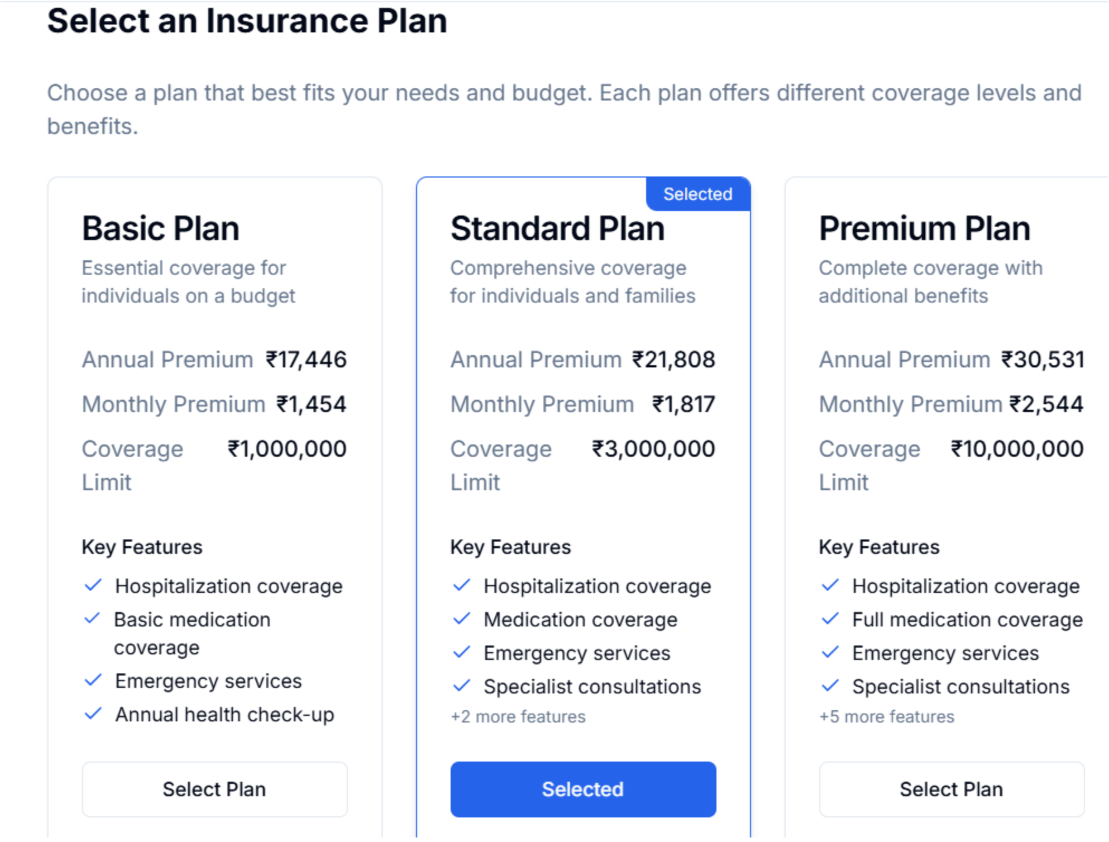
Premium Comparison

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 7.9 Showing health Insights



Select an Insurance Plan

Choose a plan that best fits your needs and budget. Each plan offers different coverage levels and benefits.

Basic Plan

Essential coverage for individuals on a budget

Annual Premium ₹17,446

Monthly Premium ₹1,454

Coverage Limit ₹1,000,000

Key Features

- ✓ Hospitalization coverage
- ✓ Basic medication coverage
- ✓ Emergency services
- ✓ Annual health check-up

Standard Plan

Comprehensive coverage for individuals and families

Annual Premium ₹21,808

Monthly Premium ₹1,817

Coverage Limit ₹3,000,000

Key Features

- ✓ Hospitalization coverage
- ✓ Medication coverage
- ✓ Emergency services
- ✓ Specialist consultations

+2 more features

Premium Plan

Complete coverage with additional benefits

Annual Premium ₹30,531

Monthly Premium ₹2,544

Coverage Limit ₹10,000,000

Key Features

- ✓ Hospitalization coverage
- ✓ Full medication coverage
- ✓ Emergency services
- ✓ Specialist consultations

+5 more features

Figure 7.10 Suggesting Plans

8 Conclusion

In this project, we developed a predictive model for health insurance prices using machine learning algorithms, with a strong focus on the XGBoost algorithm. The dataset used included various personal and demographic attributes such as age, BMI, gender, smoking status, and region. Our primary objective was to accurately estimate insurance charges based on these features. After testing multiple regression algorithms including Linear Regression, Support Vector Machine, Random Forest, Gradient Boosting, and Decision Tree, we identified XGBoost as the most effective model due to its strong performance and ability to handle complex, non-linear relationships in the data.

XGBoost provided a high R² score on both training and testing data, indicating its excellent generalization capability. To further enhance the performance, we applied hyperparameter tuning using techniques like RandomizedSearchCV, which improved the accuracy of our predictions significantly. We also evaluated the model using standard metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), all of which confirmed XGBoost's superiority over the other models. Additionally, we explored feature importance to better understand the influence of each attribute, which can help insurance companies in designing fair and personalized insurance policies.

In conclusion, the XGBoost model proved to be the best fit for our health insurance price prediction task. Its ability to provide accurate predictions, combined with effective parameter tuning and robust evaluation, makes it highly suitable for real-world applications in the insurance domain. This model can be used to assist insurers in risk assessment and premium calculation, thereby improving operational efficiency and customer satisfaction.

9 Future Enhancement

Lastly, exploring advanced ensemble techniques and deep learning models could be considered to potentially outperform XGBoost in complex datasets. Techniques such as stacking multiple algorithms, using neural networks, or incorporating time-series forecasting (for dynamic pricing models) could be valuable future directions. Also, applying automated machine learning (AutoML) for hyperparameter tuning could reduce development time and further optimize model performance. With continuous data updates and feedback loops, the model can also be trained periodically to adapt to evolving patterns in health and insurance industries.

The model can be connected with hospital management systems or health-tracking devices (e.g., Fitbit, Apple Health) to automatically retrieve user health metrics. This real-time data integration will enable dynamic premium prediction based on ongoing health trends.Another valuable addition would be a sub-model to detect potential fraud or anomalies in submitted insurance claims or input data. This can protect companies from losses and ensure fairness in pricing.

By incorporating feedback from actual outcomes—such as the final premium approved by the insurer or customer appeal responses—the model can evolve beyond static predictions. This feedback loop enables the system to retrain itself periodically, adapting to changes in policy trends and customer behavior. As a result, the model maintains higher predictive accuracy and relevance over time. It also helps in minimizing errors and increasing fairness in premium predictions. Ultimately, this makes the system more intelligent, user-focused, and aligned with real-world decision-making.

10 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.