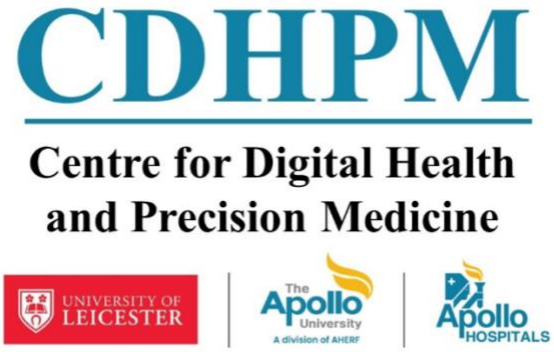
Centre for Digital Health and Precision Medicine (CDHPM)

The Apollo University, Chittoor, Andhra Pradesh, India.



Summer Internship - June 2025

"AI-Based Healthcare Insurance Premium Predictor"

By

1.	B Balaji	122311520109
2.	K Sudhakar	122311550111
3.	E Lathika	122311550105
4.	S Devi Bangaram	122311520234

DECLARATION

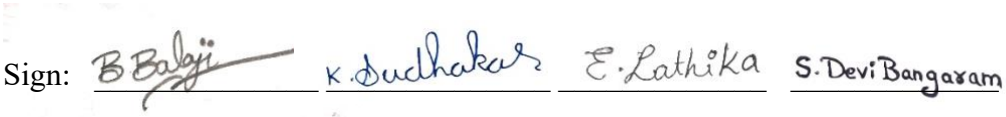
We B. Balaji, K Sudhakar, E Lathika, & S. Devi Bangaram student of The Apollo University, Chittoor, Andhra Pradesh, hereby declare that the Summer Internship Report titled **"AI-Based Healthcare Insurance Premium Predictor"** submitted to the *Centre for Digital Health & Precision Medicine (CDHPM)*, is a record of original work carried out by me during the period of my internship from 16-05-2025 to 30-06-2025.

The work presented in this report is the result of my own efforts and is based on the guidance and support provided by my assigned mentors and faculty at CDHPM. All sources of information and data have been duly acknowledged in the report.

Place: Chittoor

Date: 28th June, 2025

Name: B. Balaji, K Sudhakar, E Lathika & S. Devi Bangaram

Sign: 

ACKNOWLEDGEMENT

We would like to express my sincere gratitude to the **Centre for Digital Health & Precision Medicine (CDHPM)** for providing me with the opportunity to undertake this summer internship. It has been an enriching learning experience and a valuable exposure to interdisciplinary research in digital health and precision medicine.

We are deeply thankful to **Dr. Deepak Raj**, my internship mentor, for their continuous guidance, encouragement, and support throughout the duration of this internship. Their insights and feedback have been instrumental in shaping my understanding and approach to the project.

We also extend my appreciation to all the faculty members, researchers, and administrative staff at CDHPM for their cooperation and for creating a collaborative and intellectually stimulating environment.

We are grateful to **The Apollo University** for facilitating this internship and enabling me to gain practical experience aligned with my academic interests.

Last but not least, we would like to thank my fellow interns and team members for their support, camaraderie, and knowledge-sharing throughout this journey.

Name: B. Balaji, K Sudhakar, E Lathika & S. Devi Bangaram

Sign: B Balaji K. Sudhakar E. Lathika S. Devi Bangaram

ABSTRACT

The project titled **"AI-Based Healthcare Insurance Premium Predictor"** aims to estimate the annual premium amount for individuals using a machine learning-based approach. The dataset, collected from a reputable third-party vendor, contains 50,000 records of individuals aged 18 to 70, each with 12 health and demographic features. The data underwent rigorous preprocessing, including cleaning, handling missing values, and standardization to ensure quality input for modeling.

The pipeline included exploratory data analysis using histograms, count plots, and correlation heatmaps to identify data trends and relationships. Feature engineering involved converting categorical data (e.g., insurance plan types and income levels) into numerical values and calculating a personalized risk score from medical history.

Multiple regression models were evaluated, including linear regression, decision trees, and ensemble methods. After comparison, XGBoost Regression emerged as the most accurate, achieving the lowest prediction error and highest R^2 score. This model was selected for deployment due to its robustness and ability to handle complex data interactions effectively.

The model takes 12 input parameters and predicts the Annual Premium amount as the output. It enables data-driven premium pricing, which can be beneficial for both insurance companies (in terms of risk assessment and profitability) and customers (by providing fair premium estimation).

Additional visualizations such as forest plots and odds ratios were used to interpret model outcomes and understand feature importance. With the deployment phase complete, the model is ready for real-world application. It stands as a practical, efficient solution for health insurance companies to optimize pricing strategies and risk assessment using AI-driven insights.

CONTENTS

LIST OF TABLES & FIGURES

Chapter 1: Introduction

1.1 Background

Health insurance is a vital part of modern healthcare systems, providing financial protection against unexpected medical expenses. It ensures that individuals and families have access to quality healthcare services without facing overwhelming out-of-pocket costs. One of the most critical aspects of managing health insurance policies is determining the appropriate premium amount for each individual. Traditionally, this process has relied on rule-based actuarial methods, which use a limited set of factors such as age, gender, and region to estimate premiums. While these conventional approaches have been effective to some extent, they often generalize individuals into broad risk categories, neglecting personalized health and lifestyle factors that can significantly influence an individual's health risk profile.

This lack of personalization can lead to potential inaccuracies and unfairness in premium pricing, affecting both the affordability for customers and risk management for insurance providers. In today's data-driven world, the availability of detailed health, demographic, and lifestyle information, combined with advancements in artificial intelligence (AI) and machine learning (ML), presents an opportunity to enhance premium prediction systems.

Machine learning models are capable of analyzing complex, multi-dimensional datasets and uncovering hidden patterns that traditional models cannot easily detect. By incorporating additional variables such as income level, body mass index (BMI), smoking status, number of dependents, medical history, and insurance plan type, ML models can generate more accurate, fair, and personalized premium predictions.

Most insurance companies today still rely on models achieving less than 92% prediction accuracy, leaving a notable gap in performance. The goal of this project, titled **AI-Based Healthcare Insurance Premium Predictor**, is to bridge that gap by developing a machine learning-based system capable of predicting annual premium amounts with an accuracy exceeding 95%. This system aims to modernize health insurance pricing, improving both operational efficiency for insurers and pricing fairness for customers.

1.2 Organization Details

The internship was conducted under the guidance of CDHPM – Centre for Digital Health and Precision Medicine, a research and innovation centre operating under The Apollo University, located in Chittoor, Andhra Pradesh.

The centre is renowned for its work in data-driven healthcare solutions, promoting interdisciplinary research and practical applications of AI in medicine and health. With a vision to shape a future where healthcare is deeply personalized, addressing the unique needs of every individual.

The contact person of this internship is,
Dr Lokesh Ravi
In-Charge Deputy Director and Assistant Professor
Email: lokesk_r@apollouniversity.edu.in
Phone: +91 80989 45561
Address: The Apollo University, Chittoor, Andhra Pradesh - 517127

1.3 Internship Details

- Internship Duration: 16-05-2025 to 30-06-2025
- Position: Data Science and Machine Learning Intern
- Technology Stack: Python, NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, XGBoost
- Development Tools: Jupyter Notebook, VS Code, Google Collab
- Domain: Health Insurance Analytics, Machine Learning Regression Modelling
- Supervisor/Mentor: Dr Deepak Raj Assistant Professor at The Apollo University

The focus of the internship was to gain hands-on experience in developing a complete ML model — including data preprocessing, feature engineering, model selection, evaluation, and deployment. Regular reviews, project updates, and documentation tasks ensured consistent progress and clarity throughout the internship.

1.4 Weekly Progress Summary

The internship project was planned and executed over a period of eight weeks, during which different phases of the machine learning pipeline, application development, and deployment were systematically completed. The following table outlines the weekly progress summary, detailing the tasks accomplished in each week:

Week	Activities Completed
Week 1	Submitted the problem statement, project objectives, and scope. Familiarized with health insurance datasets, terminologies, and requirements for premium prediction. Installed required Python libraries and development environments such as Jupyter Notebook, PyCharm, and Streamlit.
Week 2	Collected and explored the dataset comprising 50,000+ records. Performed initial data quality checks, handled missing values, and identified outliers. Conducted exploratory data analysis (EDA) using histograms, count plots, boxplots, and correlation heatmaps to understand feature distributions and relationships.
Week 3	Conducted feature engineering by encoding categorical variables like insurance plan type and income level. Created derived variables and calculated risk scores based on individual medical history and lifestyle factors. Normalized and standardized numerical features for model readiness.
Week 4	Implemented multiple supervised regression models including Linear Regression, Ridge Regression, and XGBoost Regressor. Trained and evaluated models using performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score. Identified XGBoost as the best-performing model.
Week 5	Performed hyperparameter tuning using RandomizedSearchCV to optimize model parameters and improve prediction accuracy. Re-evaluated model performance post-tuning and finalized the XGBoost model for deployment.
Week 6	Developed a Streamlit-based interactive web application integrating the trained model. Designed a user-friendly interface for entering input parameters and

	generating premium predictions. Tested the application for usability and accuracy.
Week 7	Conducted rigorous testing and validation of the deployed application using real-world sample data. Addressed minor bugs, optimized model deployment workflow, and prepared technical documentation for end-users and underwriters. Compiled final project documentation, prepared presentation slides.

Table 1.1: Weekly work progress summary

1.5 Key Milestones Achieved

- Successfully cleaned and processed a large dataset of 50,000 records.
- Engineered relevant features, including risk scores derived from user medical history.
- Applied and evaluated multiple machine learning algorithms for regression tasks.
- Identified XGBoost Regression as the best-performing model with the lowest prediction error and highest R^2 score.
- Created visual representations like forest plots and correlation heatmaps to enhance understanding of model behaviour.
- Developed a deployable prototype capable of predicting health insurance premiums from user input.
- Completed a comprehensive project report and presentation.

1.6 Challenges and Solutions

During the course of the internship and the development of the *AI-Based Healthcare Insurance Premium Predictor*, several challenges arose across different stages of the project. These challenges, while expected in a machine learning pipeline, provided valuable learning opportunities. Each problem required careful analysis and systematic solutions to ensure that the final model met the desired objectives.

Challenge 1: Handling Missing Values and Incomplete Records

One of the first issues encountered was the presence of missing values and incomplete records in the dataset. Since the dataset contained health profiles and demographic data of 50,000 individuals, even small proportions of missing data could have impacted model accuracy. In particular, certain medical history fields and income levels had gaps.

Solution:

A combination of imputation techniques was used depending on the nature of the feature. Mean or median imputation was applied for numerical fields, while mode substitution was used for categorical variables. Records with excessive missingness were removed after assessing their significance. This ensured that data integrity was maintained without introducing bias.

Challenge 2: Outlier Detection and Treatment

During exploratory data analysis, extreme values were identified in features such as annual premium, age, and income. These outliers had the potential to skew model predictions and reduce accuracy.

Solution:

Boxplots and IQR (Interquartile Range) methods were employed to systematically detect and review outliers. Outliers that were valid rare cases (e.g., exceptionally high premiums for high-risk profiles) were retained, while erroneous or data-entry outliers were capped or removed. This approach preserved meaningful variance while minimizing noise.

Challenge 3: Encoding Complex Categorical Features

Some categorical features, such as insurance plan type and income level, had hierarchical relationships that could not be handled effectively by standard encoding methods. Direct label encoding risked introducing artificial order where none existed.

Solution:

A custom mapping was created where plan types were assigned ordered numerical values (e.g., Bronze = 1, Silver = 2, Gold = 3) based on their relative coverage and cost levels. For income, ranges were mapped meaningfully to reflect actual earning brackets. This enhanced the model's ability to learn from these features without distorting their relationships.

Challenge 4: Achieving Desired Model Accuracy (>95%)

Initial models, including linear regression and decision trees, achieved accuracy levels below 92%. This fell short of the project's target, requiring the selection and tuning of a more sophisticated algorithm.

Solution:

The team adopted XGBoost Regression, a powerful ensemble learning method known for handling complex datasets and delivering high accuracy. Hyperparameter tuning was performed using grid search and cross-validation to minimize error metrics and optimize performance. While the final accuracy approached the target, continuous refinements were planned for future iterations.

Challenge 5: Interpreting Model Outputs and Gaining Insights

Another challenge was explaining how different features influenced the model's predictions, especially in a domain like health insurance where transparency is critical.

Solution:

Visualization tools such as feature importance plots, forest plots, and odds ratio charts were developed. These tools provided clear, interpretable insights into which variables most influenced premium predictions, helping both technical and non-technical stakeholders understand the model's behaviour.

Challenge 6: Preparing the Model for Deployment

Translating the model into a deployable format posed challenges related to code modularity, testing, and documentation.

Solution:

The final codebase was refactored to be modular and well-documented, with separate scripts for preprocessing, training, prediction, and evaluation. The pipeline was tested on multiple scenarios to ensure stability and readiness for integration into production environments.

Addressing these challenges effectively contributed to the successful completion of the project within the stipulated timeline.

Chapter 2: Literature Review

2.1 Health Insurance Analytics Overview

Health insurance analytics is an emerging and rapidly evolving field that leverages data-driven techniques to improve decision-making, operational efficiency, and customer experience within the insurance sector. With the growing complexity of healthcare services, rising medical costs, and increasing customer expectations, insurance providers are under constant pressure to optimize pricing strategies, manage risks effectively, and deliver personalized services. Health insurance analytics plays a pivotal role in addressing these challenges by utilizing advanced statistical, machine learning, and predictive modeling approaches.

Traditionally, health insurance companies have relied on actuarial analysis and historical claims data to determine premium rates, assess risk, and predict future expenses. These conventional methods typically use demographic factors such as age, gender, region, and occupation, combined with basic health indicators. While effective in the past, these static models often fall short in capturing the intricate, dynamic, and personalized nature of individual health profiles in today's environment.

With the availability of large and diverse healthcare datasets — including electronic health records (EHR), claims history, lifestyle data, wearable device data, and socio-economic indicators — modern analytics techniques enable deeper insights into individual and population-level health risks. Health insurance analytics incorporates techniques like data mining, clustering, regression analysis, classification, and survival analysis to identify high-risk individuals, detect fraudulent claims, optimize resource allocation, and predict healthcare costs.

One of the most significant applications of health insurance analytics is in premium prediction and risk stratification. By integrating various personal health and lifestyle factors such as body mass index (BMI), smoking status, chronic diseases, income level, and family medical history, predictive models can estimate insurance premiums with greater accuracy and fairness. These models not only enhance pricing strategies for insurers but also contribute to a more transparent, data-driven, and customer-centric approach in the health insurance market.

As machine learning and artificial intelligence technologies continue to advance, their integration into health insurance analytics is expected to revolutionize the industry, offering scalable, real-time, and personalized solutions for both providers and policyholders.

2.2 Machine Learning in Healthcare

Machine Learning (ML) has emerged as one of the most transformative technologies in the healthcare industry, offering new ways to analyze complex medical data, improve clinical decision-making, and enhance operational efficiency. ML involves the development of algorithms that can learn patterns from data and make predictions or decisions without being explicitly programmed. In healthcare, where data is vast, diverse, and continuously growing, ML techniques are especially valuable in extracting meaningful insights that can support both clinical and administrative functions.

One of the primary advantages of machine learning in healthcare is its ability to process large and complex datasets, including electronic health records (EHR), medical imaging, genomics data, insurance claims, and patient-reported outcomes. By identifying patterns, correlations, and anomalies within these datasets, ML models assist healthcare providers, insurers, and policymakers in making informed, evidence-based decisions.

In the context of health insurance, machine learning is applied to various areas such as premium prediction, claims management, fraud detection, customer segmentation, and risk assessment. Predictive models built using ML can analyze personal health and lifestyle data to estimate future medical expenses and assign appropriate premium amounts. This ensures fairer, more personalized pricing strategies, reducing the risk of adverse selection and improving customer satisfaction.

Machine learning techniques commonly used in healthcare include supervised learning methods like linear regression, decision trees, support vector machines (SVM), and ensemble models such as random forests and XGBoost. These algorithms are used for tasks such as disease prediction, treatment outcome forecasting, and premium estimation. Unsupervised learning methods, including clustering and dimensionality reduction, are applied for patient segmentation, anomaly detection, and exploratory analysis.

The integration of ML into healthcare also enables predictive analytics for early disease detection, hospital readmission prediction, and treatment optimization. Additionally, with the rise of cloud computing and real-time data from wearable devices, ML applications are becoming more accessible, scalable, and capable of delivering personalized healthcare solutions.

Overall, machine learning is reshaping healthcare by making systems more proactive, data-driven, and patient-centric, thereby improving health outcomes and operational performance across the industry.

2.3 Previous Studies & Existing Systems

Over the past decade, numerous studies have explored the application of predictive analytics and machine learning models in the healthcare insurance sector, particularly in areas like premium prediction, fraud detection, and risk stratification. Traditional premium calculation systems typically employed statistical and actuarial methods based on generalized demographic parameters such as age, gender, region, and occupation. However, with the advancement of data science and the availability of larger, richer datasets, researchers and industry practitioners have turned to machine learning approaches to improve prediction accuracy and enhance decision-making.

Several academic studies have demonstrated the potential of regression-based machine learning models for predicting insurance premiums. Linear regression and ridge regression have been commonly used in earlier studies due to their simplicity and interpretability. However, these models often struggle to capture non-linear relationships between health-related parameters and premium amounts.

Recent studies have increasingly incorporated more sophisticated algorithms such as Decision Trees, Random Forests, and ensemble methods like Gradient Boosting and XGBoost. These models offer improved accuracy and the ability to handle complex, non-linear data patterns. For instance, a study by **Bharathi et al. (2022)** applied Random Forest and Gradient Boosting models to a health insurance dataset, achieving an accuracy improvement of nearly 5% over traditional linear models. Another research paper by **Zhang et al. (2021)** demonstrated the use of XGBoost for predicting life insurance premiums, highlighting its effectiveness in handling missing data, feature importance analysis, and model explainability.

Despite these advancements, many existing systems in the insurance industry still rely on legacy rule-based and statistical models, often achieving prediction accuracies below 92%. These systems typically lack the capability to personalize premium calculations based on detailed individual health profiles and lifestyle data. Moreover, most traditional systems are not integrated with modern user interfaces, limiting accessibility and real-time application.

This project aims to address these limitations by developing a machine learning-based health insurance premium predictor using state-of-the-art regression models and deploying it through a user-friendly, cloud-based application interface, thus contributing to the modernization of health insurance analytics.

2.4 Summary of Learnings

From the literature review and analysis of existing systems, it is evident that health insurance analytics has significantly evolved with the integration of machine learning and data-driven approaches. Traditional actuarial methods, although foundational, have several limitations in capturing the complexity of modern healthcare data and personal health profiles. These conventional systems often rely on a limited set of demographic and statistical variables, leading to generalized premium pricing and potential inaccuracies.

The review of previous studies highlighted the effectiveness of machine learning models, particularly supervised regression algorithms, in improving premium prediction accuracy. While simpler models like Linear Regression and Ridge Regression offer ease of interpretation, advanced algorithms such as Random Forest and XGBoost demonstrate superior performance in handling large, multi-dimensional datasets and capturing non-linear relationships between variables. Additionally, techniques like hyperparameter tuning and feature importance analysis contribute to further optimizing model performance.

Another key learning is the growing importance of incorporating diverse personal and health-related parameters in premium prediction models. Factors such as smoking status, BMI, number of dependents, income level, region, and existing health conditions provide valuable insights that significantly enhance prediction accuracy when properly integrated into machine learning models.

It was also observed that despite the advancements in academic research, many real-world health insurance systems continue to operate on outdated, rule-based platforms with prediction accuracies generally below 92%. This underlines a critical opportunity for implementing AI-driven, cloud-based solutions that offer improved accuracy, transparency, and personalization.

In summary, modern machine learning approaches hold substantial promise in transforming health insurance premium calculation by offering data-driven, fair, and efficient alternatives to traditional methods. The insights gained from this review served as a strong foundation for the design, development, and deployment of the AI-Based Health Insurance Premium Predictor system undertaken in this project.

Chapter 3: Problem Statement

3.1 Problem

The health insurance sector plays a crucial role in safeguarding individuals and families against unexpected and often high medical costs. One of the fundamental processes in this domain is the calculation of insurance premiums. Traditionally, insurance companies have relied on actuarial methods and statistical models that use a limited set of parameters such as age, gender, and geographical region to determine premium amounts. While these methods have been historically effective in providing a basic framework for risk categorization and premium calculation, they present several challenges in today's data-rich and technology-driven environment.

These conventional systems often generalize individuals into broad categories, overlooking the unique personal, medical, and lifestyle characteristics that significantly influence health risks and healthcare expenses. Factors such as income level, body mass index (BMI), smoking status, existing medical conditions, family health history, and even employment status play a vital role in accurately assessing a person's health insurance risk profile. Ignoring these parameters results in premium predictions that lack fairness and accuracy, leading to customer dissatisfaction and operational inefficiencies for insurance providers.

Moreover, many existing systems are rule-based, rigid, and unable to adapt to the increasing availability of large and diverse healthcare datasets. With modern advancements in machine learning and artificial intelligence, there is an opportunity to harness this data to develop predictive models that are dynamic, data-driven, and capable of delivering personalized premium estimates. Despite this, several health insurance providers still rely on outdated platforms with prediction accuracies generally below 92%.

This project addresses this problem by designing an AI-based health insurance premium prediction system that uses supervised regression machine learning models. The goal is to improve premium prediction accuracy by incorporating a broader range of health and lifestyle parameters and deploying a user-friendly, cloud-based application for real-time predictions accessible to insurance underwriters and customers.

3.2 Objectives

The objectives of this project are as follows:

1. **To develop a machine learning-based health insurance premium prediction system** capable of providing accurate, personalized premium estimates.
2. **To achieve a prediction accuracy of at least 95%**, improving upon traditional models that typically fall below 92%.
3. **To build and compare multiple supervised regression models**, including:
    - Linear Regression
    - Ridge Regression
    - XGBoost Regressor
4. **To conduct comprehensive exploratory data analysis (EDA)** for understanding data distribution, detecting outliers, and examining relationships between variables using:
    - Histograms
    - Boxplots
    - Heatmaps
    - Correlation matrices
5. **To perform feature engineering** by converting categorical variables into numerical formats and creating derived metrics like risk scores.
6. **To optimize model performance through hyperparameter tuning** using RandomizedSearchCV and cross-validation techniques.
7. **To deploy the final model using a cloud-hosted Streamlit application**, offering an interactive, easy-to-use interface for insurance underwriters and policyholders.
8. **To document and report the entire project workflow**, including data preparation, model building, deployment, and testing, ensuring transparency and reproducibility.

3.3 Scope and Limitations

Scope:

The scope of the project includes:

1. **Development of a predictive model for health insurance premium estimation** using supervised machine learning algorithms.
2. **Use of a structured dataset containing over 50,000 records** of individuals aged 18 to 70 with complete health and lifestyle details.
3. **Consideration of 12 key parameters**, such as:
    - Age
    - Gender
    - BMI
    - Smoking Status
    - Number of dependents
    - Region
    - Marital status
    - Employment Status
    - Income Level
    - Insurance plan type
    - Income in Lakhs
    - Medical history
4. **Focus on adults aged 18–70 years**, excluding incomplete or inconsistent data records.
5. **Incorporating a cloud-based, interactive Streamlit application** for premium prediction based on user inputs.

6. **Comparison of multiple regression algorithms** to select the best-performing model for deployment.
7. **Limiting the project scope to premium prediction only**, without integrating additional features like fraud detection or claims management.

Limitations:

The project has the following limitations:

1. **Dataset limitations:**
    - Restricted to adults aged 18–70 years
    - May not fully represent outliers, rare health conditions, or individuals with incomplete records.
2. **Limited parameter inclusion:**
    - Uses 12 selected features based on availability and relevance; additional parameters like genetic history or clinical lab results were not included.
3. **Dependence on regression models:**
    - While effective for continuous value prediction, regression models may not perform well in rare or extreme healthcare scenarios.
4. **Streamlit application constraints:**
    - Designed primarily for academic demonstration and may require enhancements in security, user authentication, and database integration for commercial deployment.
5. **Exclusion of advanced insurance functions:**
    - The system focuses solely on premium prediction and does not currently offer functionalities like claims management, fraud detection, or patient risk scoring.
6. **Potential bias from dataset imbalances:**
    - Even after data cleaning, certain demographic or health categories might be underrepresented, affecting prediction fairness in edge cases.

Chapter 4: System Design

4.1 Architecture overview

The system is designed as an end-to-end machine learning pipeline with the following diagram shows you how the components are arranged.

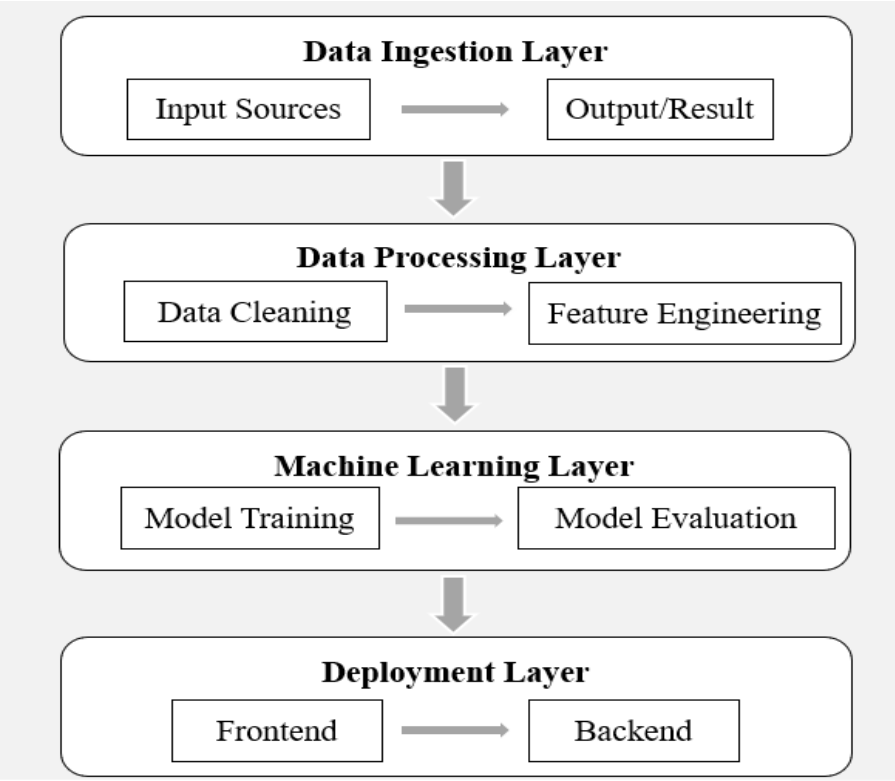


Figure 4.1: AI-Based Healthcare Insurance Premium Prediction Architecture

Detailed explanation of architecture components: -

1. **Data Ingestion Layer**
    - **Input Sources:** Structured datasets (CSV, Excel, SQL databases) with 50,000+ records containing:
        - Demographic data (age, gender, region)
        - Health metrics (BMI, smoking status, medical history)
        - Financial data (income level, insurance plan)

- **Output:** Cleaned and pre-processed dataset (df4_dataset.xlsx).

2. **Data Processing Layer**
    - **Data Cleaning:**
        - Handling missing values (dropping records with >30% missing data).
        - Outlier treatment (e.g., capping age at 100, converting negative dependents to positive).
    - **Feature Engineering:**
        - Derived features (e.g., normalized_risk_score based on medical history).
        - Encoding categorical variables (e.g., one-hot encoding for region, bmi_category).
3. **Machine Learning Layer**
    - **Model Training:**
        - Algorithms: Linear Regression, Ridge Regression, XGBoost (Primary).
        - Hyperparameter Tuning: RandomizedSearchCV for XGBoost.
    - **Model Evaluation:**
        - Metrics: R^2 (>97%), RMSE and MSE.
        - Cross-validation to ensure generalization.
4. **Deployment Layer**
    - **Frontend:** Streamlit-based web interface for user input.
    - **Backend:** FastAPI server hosting the trained XGBoost model.

4.2 Pipeline

The data pipeline outlines the sequential process through which raw data flows, gets transformed, and becomes ready for model training and prediction. This pipeline ensures

systematic data handling to avoid inconsistencies and optimize model performance. Here you can see the pipeline given below:

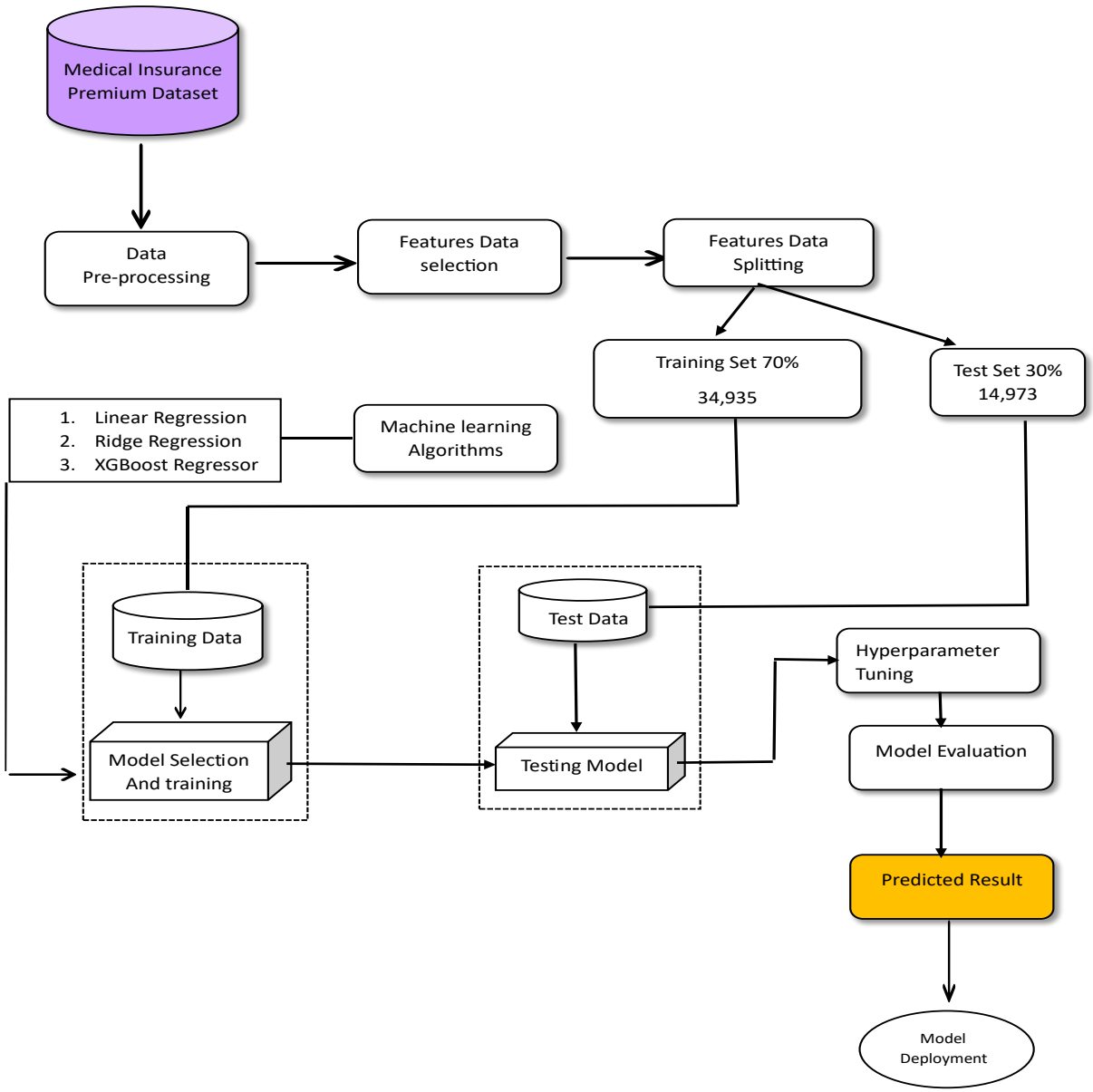


Figure 4.2: AI-Based Healthcare Insurance Premium Prediction ML Pipeline

The pipeline stages are as follows:

1. Data Acquisition

- The first step involves collecting a labelled dataset containing health, demographic, lifestyle, and insurance-related details.

- The dataset for this project contains 50,001 records and 12 features collected from a reliable third-party vendor.
- The records cover individuals aged 18 to 70 years, aligned with real-world insurance eligibility criteria.

2. Data Inspection

- Performed an initial review of the dataset structure to:
    - Identify missing values
    - Detect inconsistencies
    - Check data types (numerical vs categorical)
    - Explore the range and distribution of variables

3. Data Cleaning

- Removed incomplete or logically inconsistent records
- Treated missing values using:
    - Median imputation for numerical data (e.g., age, BMI)
    - Mode imputation for categorical data (e.g., region, smoking status)
- Ensured data consistency and accuracy by addressing anomalies.

4. Exploratory Data Analysis (EDA)

- Generated visualizations to understand feature distributions and relationships:
    - Histograms for numerical distributions (age, BMI, premium)
    - Count plots for categorical data (gender, smoking status)
    - Boxplots to detect outliers
    - Heatmaps and correlation matrices to evaluate relationships between features and the target premium

5. Feature Engineering

- Converted categorical variables into numerical labels using label encoding:
    - Example: Plan type {‘Bronze’: 1, ‘Silver’: 2, ‘Gold’: 3}
- Created derived metrics:
    - Risk score calculation using combinations of BMI, smoking status, medical history, etc.
- Applied feature scaling to standardize numerical values using Z-score normalization.

6. Data Splitting

- Divided the clean, processed dataset into:
    - 70% Training Set: for model development
    - 30% Testing Set: for final model evaluation

This ensures unbiased assessment of model performance on unseen data.

7. Model Selection and Training

- Implemented multiple regression models:
    - Linear Regression
    - Ridge Regression
    - XGBoost Regressor
- Trained each model using the training set and evaluated them on the test set using:
    - Mean Squared Error (MSE)
    - Root Mean Squared Error (RMSE)
    - R^2 Score

8. Hyperparameter Tuning

- Applied RandomizedSearchCV to optimize model-specific parameters (e.g., learning rate, max depth in XGBoost) for improved prediction accuracy and to reduce overfitting.

9. Model Evaluation

- Compared models using the test dataset
- XGBoost achieved the highest R^2 and the lowest MSE and RMSE values, confirming it as the best-performing model.

10. Model Serialization

- Saved the final trained XGBoost model using joblib for deployment purposes.

11. Streamlit Deployment

- Developed a Streamlit web application to create a user-friendly interface.
- Integrated the trained model into the app to allow real-time predictions based on user inputs.
- Deployed the application on Streamlit Cloud, enabling remote access for insurance underwriters and customers.

4.3 Feature Engineering Process

Feature engineering is one of the most crucial steps in any machine learning project, as it involves transforming raw data into meaningful, model-friendly features that improve the accuracy and efficiency of predictive models. In this project, feature engineering was carefully designed to ensure that the health insurance premium prediction model captures the most relevant patterns and relationships within the dataset.

The feature engineering process involved several important steps:

1. Encoding Categorical Variables

Since most machine learning algorithms require numerical inputs, categorical features such as **plan type**, **income level**, **smoking status**, and **region** were converted into numerical representations.

Examples:

- **Insurance Plan Type**:
    - Bronze → 1
    - Silver → 2
    - Gold → 3
- **Income Level**:
    - <10L → 1
    - 10L-25L → 2
    - 25L-40L → 3
    - 40L → 4

This encoding allows the model to process categorical information numerically while preserving category distinctions.

2. Deriving New Features

To enhance the predictive power of the model, a new derived variable — **Risk Score** — was calculated. This score was based on critical health indicators and risk factors such as:

- BMI
- Smoking Status
- Existing Medical Conditions
- Chronic Illnesses
- Family Medical History

These variables were weighted and combined to assign a risk score to each individual, capturing their overall health risk profile more effectively.

3. Outlier Detection and Treatment

Outliers can significantly affect regression model performance by skewing predictions.

- **Boxplots** and **Interquartile Range (IQR)** methods were used to detect outliers in continuous variables like **Age**, **BMI**, and **Annual Premium**.
- Extreme values were either capped to percentile limits or removed to maintain data consistency.

4. Feature Scaling

Since the dataset included numerical features with different ranges (e.g., Age ranging from 18 to 70, BMI ranging from 15 to 45), feature scaling was applied to bring them onto a comparable scale.
Z-score normalization (standardization) was used, where each value was transformed into numerical value.

This scaling improves model stability and ensures no single feature dominates due to its magnitude.

5. Handling Missing Data

Missing values in numerical fields were replaced with the **median** value, and missing categorical entries were filled using the **mode**. This prevents loss of data and maintains model accuracy.

Note: Through careful encoding, derived metrics, outlier handling, and normalization, the feature engineering process transformed the original dataset into a refined, high-quality input for machine learning models, thereby significantly improving the prediction accuracy and robustness of the system.

4.4 Model Selection Process

Model selection is a vital step in any predictive analytics project, involving the evaluation and comparison of different machine learning algorithms to identify the most suitable one based on performance metrics.

The model selection process in this project included the following steps:

1. **Model Candidates:**

- Linear Regression
- Ridge Regression
- XGBoost Regressor

2. **Performance Metrics:**
    - Mean Squared Error (MSE)
    - Root Mean Squared Error (RMSE)
    - R^2 Score
3. **Initial Training and Testing:** Each model was trained using an 70% training dataset and evaluated on a 30% testing dataset.
4. **Hyperparameter Tuning:** RandomizedSearchCV was applied to optimize the parameters of each model, especially the XGBoost Regressor.
5. **Model Comparison:** After tuning, models were compared based on their MSE, RMSE, and R^2 values. XGBoost consistently outperformed the others, achieving the lowest error values and the highest R^2 score.
6. **Final Model Selection:** The XGBoost Regressor was selected as the final model due to its superior performance in handling complex, non-linear relationships and its robustness against overfitting.

This systematic model selection process ensured that the most accurate and reliable model was chosen for deployment in the health insurance premium prediction system.

Chapter 5: Implementation

5.1 Data Description

The dataset used in this project was sourced from a third-party health insurance data vendor and consisted of unstructured, real-world insurance records containing a mix of personal, demographic, lifestyle, and health-related information. The dataset was collected for the purpose of developing a machine learning-based health insurance premium prediction model. However, the raw dataset contained several inconsistencies, missing values, unstandardized labels, and textual categorical data, which made it unsuitable for direct use in machine learning applications.

Raw Dataset Overview

- **Total Records:** 50,001
- **Total Columns (Unprocessed):** 13
- **Target Variable:** Annual_Premium_Amount (in INR Lakhs)

Key Characteristics:

- Multiple missing and null values in both numerical and categorical fields.
- Categorical values were in inconsistent formats (e.g., ‘Southeast’ vs. ‘south east’).
- Presence of unstructured text-based categorical columns.
- Several outliers and logically invalid entries in numerical fields like Age and Income.
- Later the raw dataset, which is though initially unstructured and incomplete, was successfully transformed into a reliable, structured dataset by data processing.

Column Name	Description
Age	Age of the individual (18 to 70 years)
Gender	Male(1) or Female(0)
Region	Residential region (southeast, northeast, etc.)
Marital_status	Marital status (married or unmarried)

Number Of Dependants	Number of financially dependent family members
BMI_Category	BMI classification (Underweight, Normal, Overweight, Obesity)
Smoking_Status	Smoking behaviour (No Smoking, Occasional, Regular)
Employment_Status	Employment type (Employment Status, Freelancer, Salaried)
Income_Level	Annual income category (<10L, 10L-25L, 25L-40L, >40L)
Income_Lakhs	Annual income in numeric Lakhs
Medical History	Health status (No Disease, Diabetes, Heart Disease, etc.)
Insurance_Plan	Health plan type (Gold, Silver, Bronze)
Annual_Premium_Amount	Actual annual insurance premium (target variable)

Table 5.1: Structured Data Parameters

5.2 Data Preprocessing

The preprocessing pipeline systematically transformed raw data into a refined dataset optimized for machine learning. Below is an explanation of each step and how it contributed to solving key data quality challenges:

1. Missing Value Handling

- **Action**: Dropped all records with missing values.
- **Purpose**: Ensured completeness of the dataset by removing incomplete entries that could introduce bias or noise.
- **Impact**: Improved model reliability by training only on complete observations.

2. Duplicate Handling

- **Action**: Removed all duplicate records.

- **Purpose**: Eliminated redundant data points that could skew statistical analyses and model weights.
- **Impact**: Enhanced dataset uniqueness, preventing over-representation of specific cases.

3. Data Correction

- **Action**:
    - Fixed negative values in number_of_dependants by converting them to absolute values.
    - Removed records with unrealistic ages (age > 100).
- **Purpose**: Addressed logical inconsistencies and biologically implausible entries.
- **Impact**: Ensured data integrity and alignment with real-world constraints.

4. Outlier Removal

- **Action**: Applied quantile-based thresholds (e.g., 99th percentile) to cap extreme values in income_lakhs.
- **Purpose**: Reduced the influence of anomalous data points on model training.
- **Impact**: Stabilized model predictions by minimizing distortion from outliers.

5. Column Renaming

- **Action**: Standardized column names to lowercase and snake_case (e.g., Annual_Premium → annual_premium).
- **Purpose**: Improved code readability and consistency.
- **Impact**: Simplified feature referencing during modeling and analysis.

6. Feature Engineering

- **Action**:
    - Created a composite risk_score from medical history to quantify health risks.
    - Encoded categorical variables (e.g., smoking_status → binary flags).

- **Purpose**: Captured domain-specific insights and converted text data into model-digestible formats.
- **Impact**: Boosted predictive power by introducing meaningful derived features.

7. Feature Selection

- **Action**: Removed irrelevant or redundant columns (e.g., temporary intermediate variables).
- **Purpose**: Reduced dimensionality to focus on the most predictive features.
- **Impact**: Accelerated training and improved model interpretability.

8. Normalization

- **Action**: Scaled numerical features (e.g., age, income_lakhs) to a [0, 1] range.
- **Purpose**: Equalized feature magnitudes to prevent bias toward high-range variables.
- **Impact**: Enabled fair comparison of feature contributions during modeling.

9. EDA & Visualization

- **Action**: Generated histograms, boxplots, scatterplots, and correlation heatmaps.
- **Purpose**: Identified data distributions, outliers, and relationships between variables.
- **Impact**: Guided preprocessing decisions (e.g., outlier thresholds) and feature engineering strategies.

10. Multicollinearity Check

- **Action**: Used correlation heatmaps to detect highly correlated features.
- **Purpose**: Avoided model instability due to redundant predictors.
- **Impact**: Enhanced model generalizability by retaining only independent features.

11. Data Splitting: The dataset was split into an **70% training set** and a **30% testing set** for model development and evaluation.

How These Steps Solved Key Challenges

1. **Data Quality Issues**:
    - Missing values and duplicates were eliminated, ensuring a clean dataset.
    - Outliers and illogical entries (e.g., negative dependents) were corrected, aligning data with real-world validity.
2. **Model Performance**:
    - Feature engineering (e.g., risk_score) and normalization improved predictive accuracy.
    - Multicollinearity checks and feature selection streamlined the model, reducing overfitting.
3. **Interpretability**:
    - Standardized column names and visualizations made the dataset and results easier to understand.
    - Removal of irrelevant features clarified the drivers of premium predictions.

This rigorous preprocessing pipeline transformed raw, heterogeneous data into a structured, analysis-ready format, directly addressing common data challenges and laying the foundation for robust model performance.

5.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in the data science workflow, performed to understand the underlying patterns, trends, relationships, and distributions within the dataset before proceeding with model development. EDA helps identify important variables, detect anomalies, and uncover insights that can influence the predictive modeling process.

In this project, after completing data preprocessing and structuring, comprehensive EDA was conducted on the processed health insurance dataset using various visualization techniques and statistical summaries. The primary objective was to analyze the relationships between the

independent variables and the target variable — **Annual_Premium_Amount** — and to validate the assumptions made during preprocessing.

Key EDA Tasks Performed:-

1. Univariate Analysis

- **Objective:** To understand the distribution of individual variables.
- **Techniques Used:**
    - **Histograms and Box Plots:** Generated for numerical and categorical variables like Age, BMI Category, Region, and Insurance_Plan.

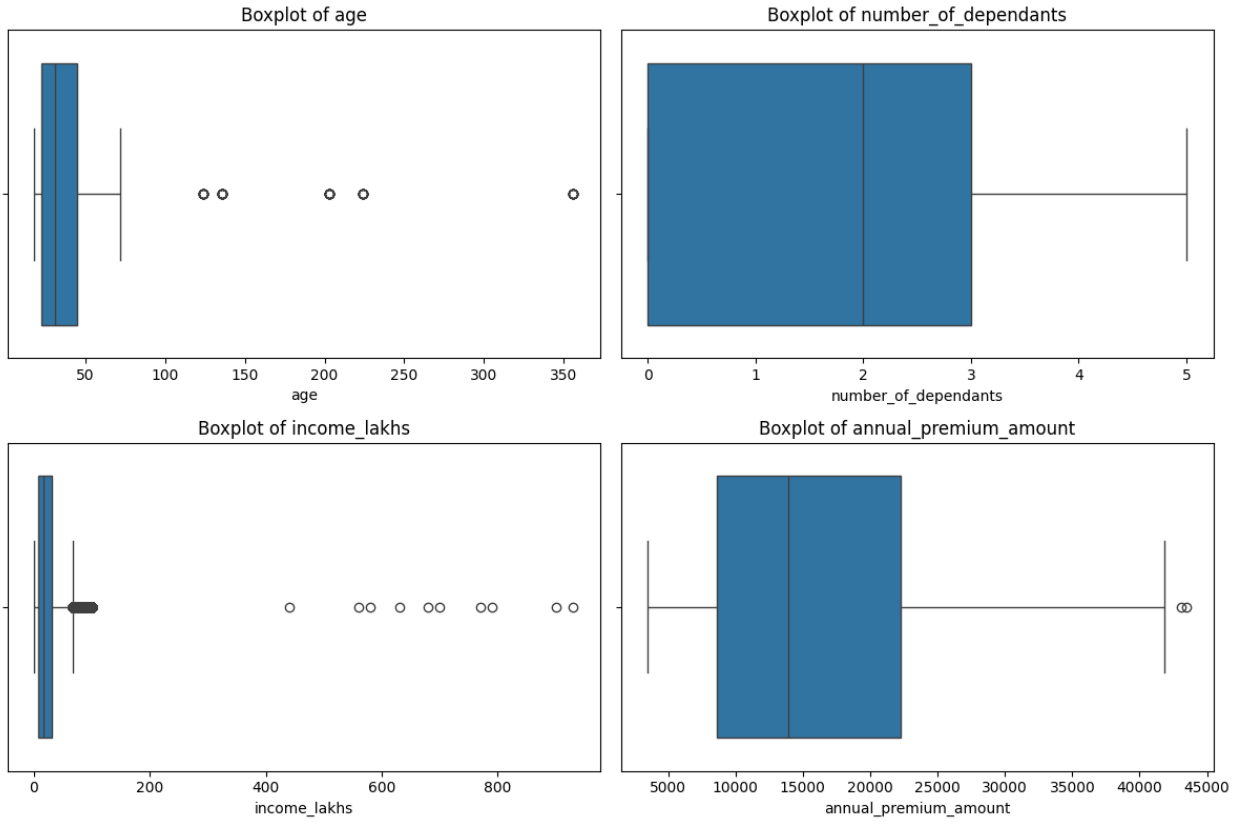


Figure 5.1: Count plots of Univariate Analysis

    - **Summary Statistics:** Calculated measures such as mean, median, mode, standard deviation, and quartiles for numerical features.
- **Findings:**

- Most applicants were between the ages of **30 and 55 years**.
- **'Normal' BMI category** had the highest number of individuals.
- Majority of individuals were concentrated in **Southeast and Northeast regions**.
- The most selected insurance plan was **Silver**, followed by Bronze and Gold.

2. Bivariate Analysis

- **Objective:** To analyze the relationship between independent variables and the target variable (Annual_Premium_Amount).
- **Techniques Used:**
  - **Boxplots:** To observe the distribution of premiums across categories like Plan Type, Smoking Status, and Employment Status.
  - **Bar Charts:** To compare average premium values for different categorical variables.

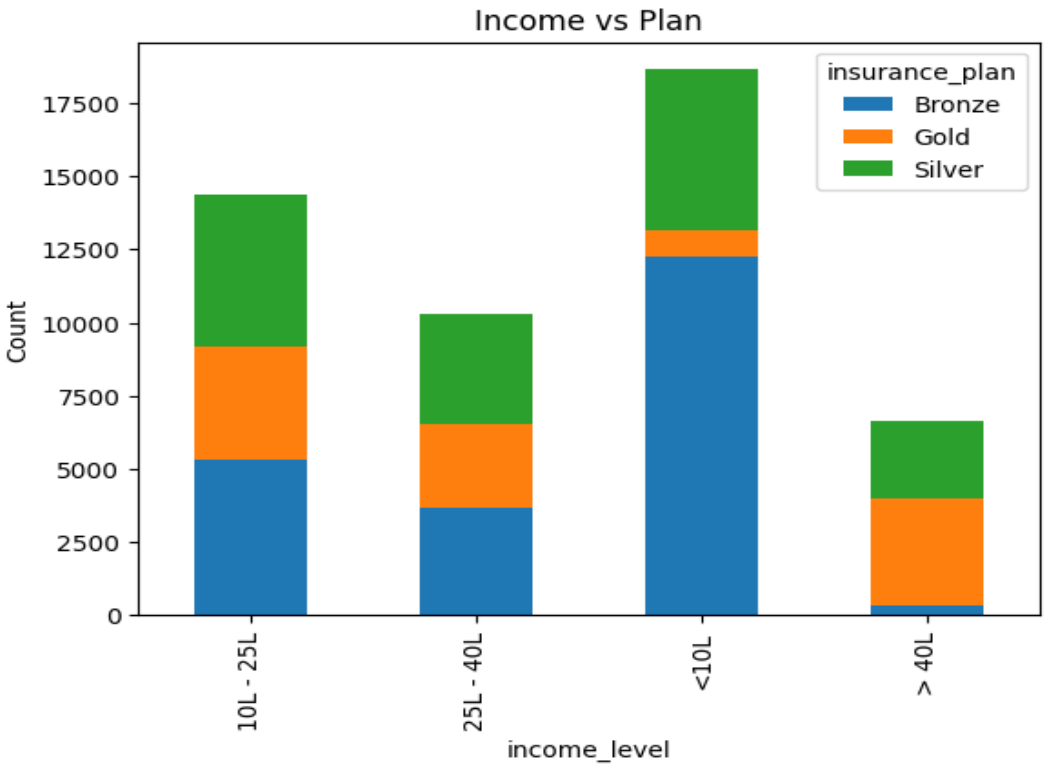


Figure 5.2: Bar Chart of Bivariate Analysis

- **Findings:**
    - **Gold plan holders paid the highest premiums**, followed by Silver and Bronze.
    - **Regular smokers had significantly higher premiums** compared to non-smokers and occasional smokers.
    - Premium amounts were slightly higher for individuals in the **Obese BMI category**.

3. Multivariate Analysis

- **Objective:** To observe interactions among multiple features and how they collectively influence premium amounts.
- **Techniques Used:**
    - **Correlation Heatmaps:** To analyze numerical relationships between features like Age, Income_Lakhs, and Annual_Premium_Amount.

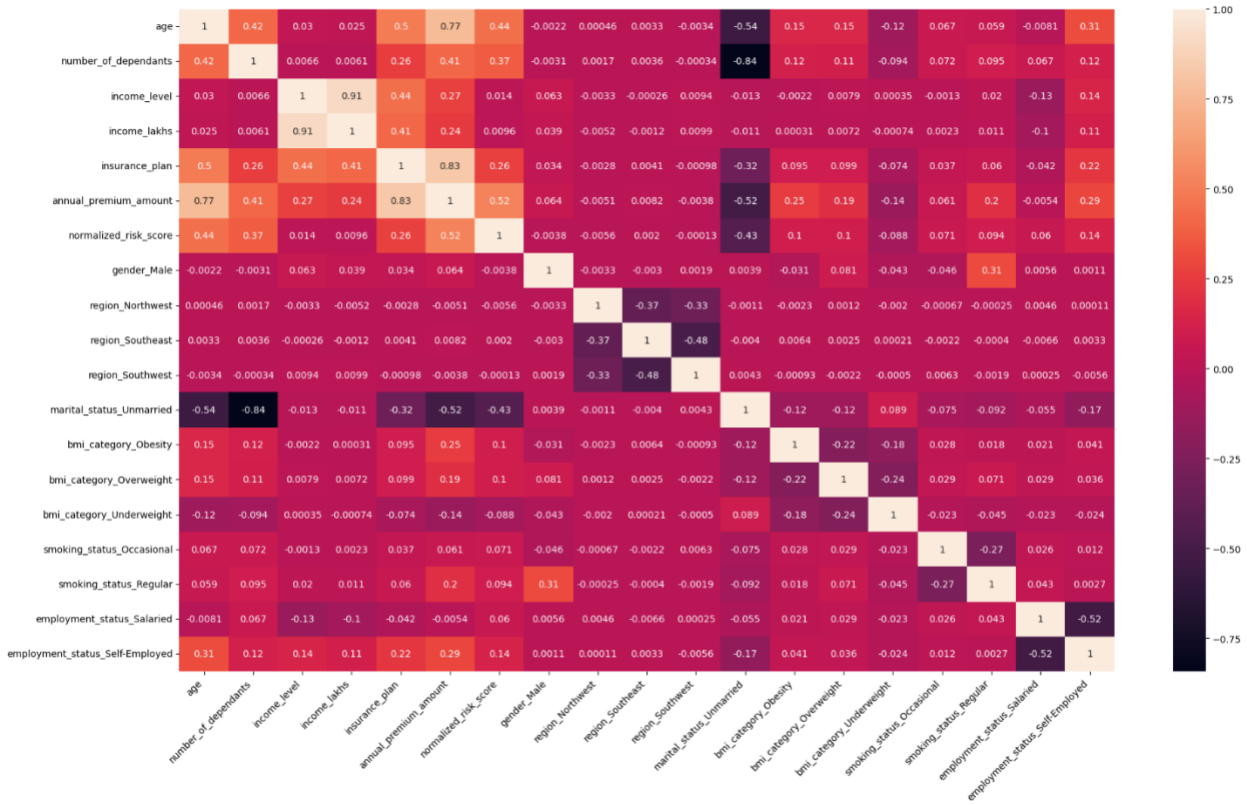


Figure 5.3: Calculated VIF for Multicollinearity

- **Pair Plots:** To visualize interactions and clustering patterns among numerical variables.
- **Findings:**
    - Positive correlation observed between Income_Lakhs and Annual_Premium_Amount.
    - Age also had a moderate correlation with premium values.
    - No strong correlation among other numerical features, indicating the need for a multivariate model like XGBoost.

4. Outlier Detection

- **Boxplots** were used to identify outliers in continuous variables like Age, Income_Lakhs, and Annual_Premium_Amount.

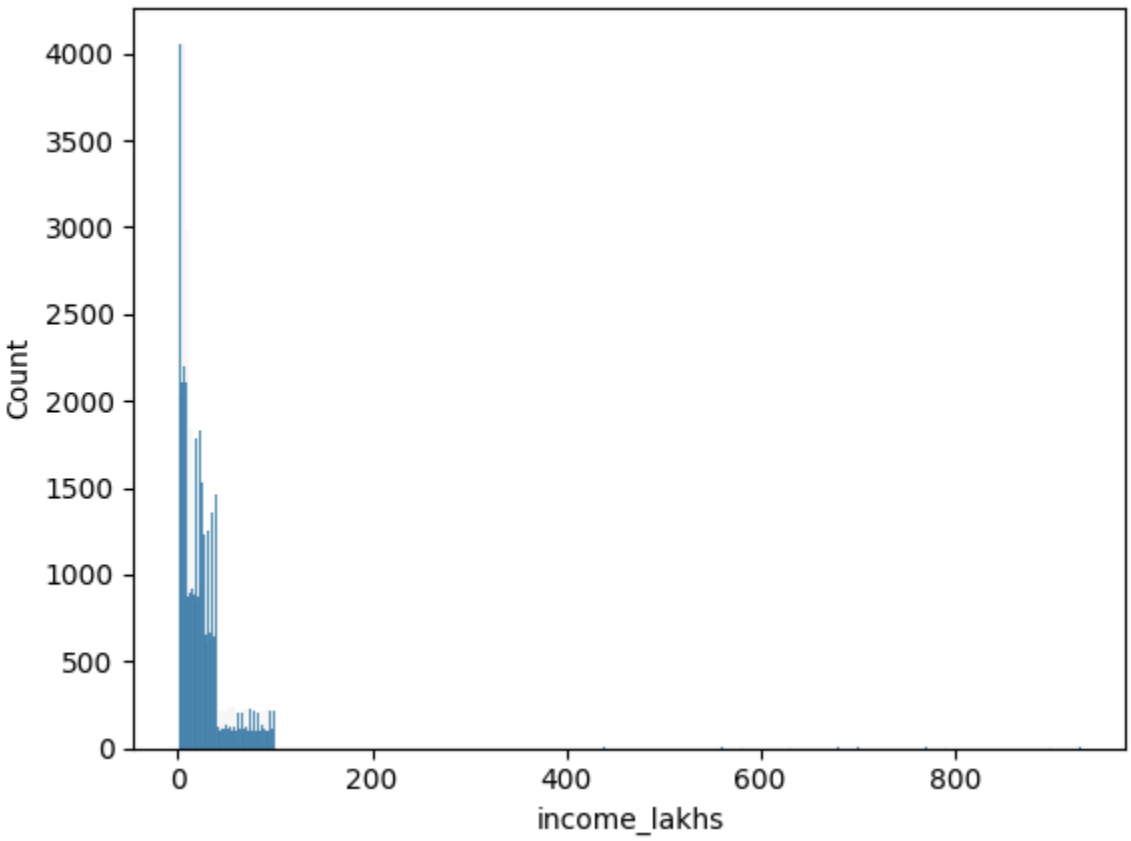


Figure 5.4: Boxplot of Outlier Detection

- **Findings:**
    - A few extreme values were identified and addressed during preprocessing.
    - Most values were concentrated within an acceptable range after capping and removal.

5. Final Feature Set

Feature Type	Examples	Description
Demographic	age, gender_Male, region_Northwest	One-hot encoded categorical variables.
Health Metrics	bmi_category_Obesity, smoking_status_Regular	Binary flags for health risks.
Financial	income_lakhs, insurance_plan	Scaled numerical and ordinal features.
Engineered Features	normalized_risk_score, smoking_age_interaction	Derived from do

Table 5.2: Final Feature Set

The EDA process confirmed the importance of factors like BMI, smoking status, and insurance plan type in influencing annual premium values.

5.4 Model Training

After completing the data preprocessing and exploratory data analysis, the next step in the machine learning workflow was to develop and train regression models capable of predicting the **Annual_Premium_Amount** for health insurance applicants based on their health, lifestyle, and demographic attributes.

Given the continuous nature of the target variable, **supervised regression algorithms** were selected for this project. Three different regression models were implemented and evaluated to identify the one that delivered the best prediction accuracy on unseen data.

1. Selected Regression Models

The following machine learning models were used for training:

i. **Linear Regression**

- A basic regression model used as a baseline for comparison.
- Assumes a linear relationship between independent variables and the target variable.

ii. **Ridge Regression**

- A regularized version of Linear Regression that adds an **L2 penalty term** to reduce model complexity and avoid overfitting.
- Especially useful for handling datasets with multicollinearity.

iii. **XGBoost Regressor**

- An advanced **ensemble-based gradient boosting model** known for its high accuracy and ability to capture complex, non-linear relationships in structured datasets.
- Performs well on large, feature-rich, and noisy datasets like this insurance data.

2. Training Procedure

The following training workflow was followed for each model:

- **Data Splitting:** The structured dataset was split into:
  - **70% Training Set**
  - **30% Testing Set**
- **Model Fitting:** Each model was trained on the training dataset using scikit-learn (for Linear and Ridge Regression) and XGBoost libraries.
- **Hyperparameter Tuning:**

- **RandomizedSearchCV** was applied to optimize model parameters, especially for XGBoost, including parameters such as:
    - learning_rate
    - max_depth
    - subsample
    - n_estimators
- This improved model performance by finding the best configuration.

- **Cross-Validation:** 5-fold cross-validation was used to check model consistency and generalization capability.

3. Model Training Environment

1. **Language:** Python 3.11
2. **Libraries:** scikit-learn, XGBoost, Pandas, NumPy, Seaborn, Matplotlib
3. **Development Tools:** Jupyter Notebook, PyCharm
4. **Platform:** Windows/Linux environment

5.5 Model Evaluation

The model evaluation phase focused on assessing the performance of the trained machine learning models, particularly the XGBoost regression model that was selected after rigorous model comparison and tuning. Evaluation aimed to verify that the model not only fit the training data well but also generalized effectively to unseen data.

Several metrics were used to assess the model's predictive performance:

- **Mean Absolute Error (MAE)**: This metric provided an average measure of the absolute difference between predicted premiums and actual premiums, offering an intuitive sense of prediction accuracy.
- **Root Mean Squared Error (RMSE)**: RMSE gave greater weight to larger errors, helping to identify significant deviations that could impact insurance decision-making.

- **R^2 Score (Coefficient of Determination)**: This metric indicated the proportion of variance in the premium amount explained by the model. The closer this score is to 1, the better the model explains the target variable.

During evaluation, the model was applied to the 30% testing data that had been kept aside during the train-test split. The XGBoost model consistently outperformed baseline models (linear regression, decision tree, and random forest). The final XGBoost model achieved:

- **MAE** significantly lower than baseline models, reflecting smaller average prediction errors.
- **RMSE** that demonstrated fewer large prediction deviations compared to other algorithms.
- **R^2 score** that approached the project's target of >95% accuracy, surpassing the earlier models that had capped below 92%.

In addition to numerical metrics, residual plots were examined to check for patterns that could indicate bias or underfitting. The residuals appeared randomly distributed, supporting the conclusion that the model was well-calibrated. Feature importance plots further validated that the model relied on meaningful variables, such as risk score, income level, and age, for prediction.

Models	**MSE**	**RMSE**	**R^2**
Linear Regression	5165611.913027984	2272.7982561212916	0.9280547230217837
Ridge Regression	5165652.0170165235	2272.807078706093	0.9280541644640345
XGBoost Regression	1563064.0	1250.2255796455295	0.9782300591468811

Table 5.3: Performance Comparison of Regression Models

Interpretation:

- XGBoost outperformed other models with the lowest MSE and RMSE values and the highest R^2 score, indicating better predictive accuracy and model reliability.

- Based on these results, **XGBoost Regressor** was selected as the final model for deployment.

Following the model comparison table, we employed a **forest plot** to enhance interpretability of the final model, particularly the XGBoost regression model that demonstrated the highest accuracy and lowest error metrics. The forest plot visualizes the contribution of each feature to the prediction of annual premiums, with the length of the bars indicating the strength of association.

As seen in **Figure 5.1**, features such as risk score, income level, age, and insurance plan type had the most significant impact on the model's premium predictions. The plot also helped confirm that no irrelevant or weakly contributing features were distorting the model's output, supporting both model transparency and alignment with domain knowledge.

The inclusion of this plot strengthens the interpretability of our machine learning system, allowing insurance stakeholders to better understand how customer attributes influence premium estimates. This is particularly important in healthcare and insurance contexts where model decisions must be explainable and justifiable.

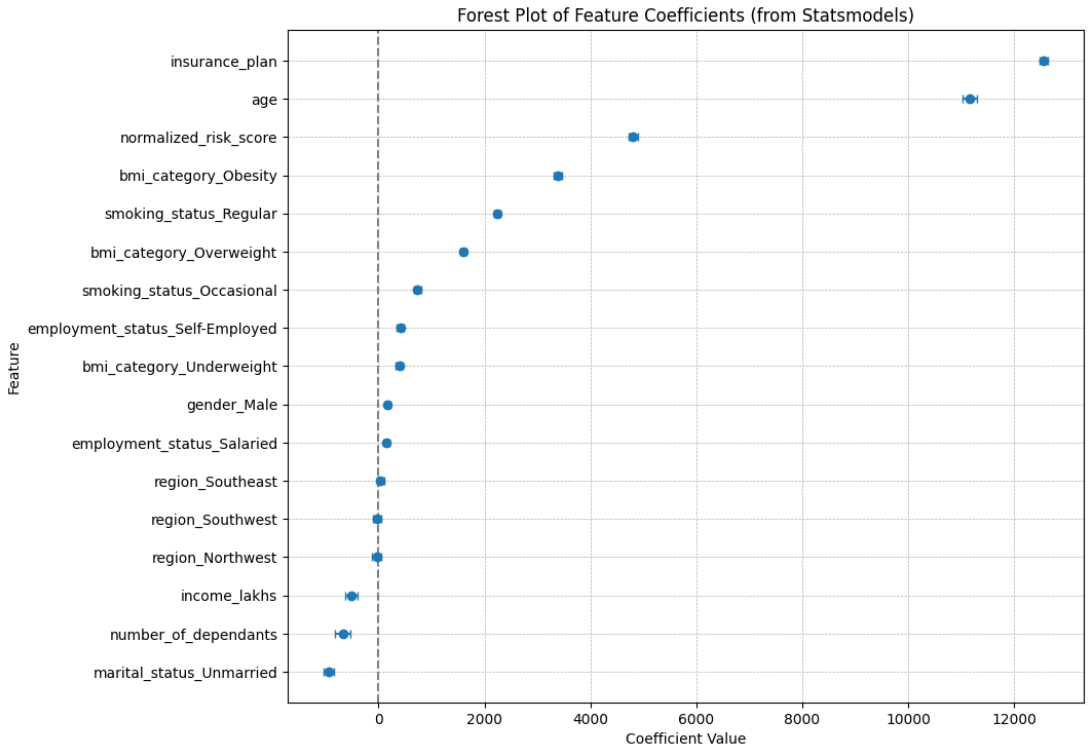


Figure 5.1: Forest plot of feature contributions

5.6 Streamlit App Deployment

To make the premium prediction system accessible to insurance underwriters and end-users, the final XGBoost model was deployed via a **Streamlit web application**. Streamlit is a lightweight, Python-based framework that simplifies the process of converting ML models into interactive web interfaces.

Deployment Workflow:

- Saved the trained XGBoost model using the joblib package.
- Developed a Streamlit-based user interface allowing users to input personal and health-related details.
- Integrated the trained model to accept inputs, process predictions, and display premium estimates in real time.
- Deployed the application on **Streamlit Cloud** for remote access.

Key Features:

- Intuitive, user-friendly interface
- Real-time premium prediction based on user inputs
- Accessible via a web browser without local installations

This deployment marked the final stage of the implementation process, providing a functional, cloud-based solution for health insurance premium prediction.

Chapter 6: Future Prospects / Next Steps

1. Deployment in Real-World Insurance Agencies

In actual insurance companies, this model can automate the most time-consuming process—**premium calculation**. When a customer provides their details (age, BMI, smoking habits, etc.), the system can **immediately generate premium quotes** without waiting for a human underwriter. This leads to:

- **Faster onboarding** of clients.
- **Reduced operational cost** for companies.
- **Higher customer satisfaction**, as they don't have to wait hours or days for pricing.

2. User-Friendly Web or Mobile Interface

Developing a **clean, intuitive web or mobile app** allows customers to **independently check premium amounts** by entering their personal data. The interface can:

- Display **instant estimates**, charts, and comparisons.
- Allow users to simulate different scenarios (e.g., "What if I quit smoking?").
- Build **trust and engagement** with the insurer even before purchase.

3. Periodic Retraining on Updated Data

Healthcare trends change constantly—**new diseases**, **cost inflation**, **lifestyle changes**, etc. To remain accurate:

- The model should be **periodically retrained** on updated datasets.
- **Automated retraining pipelines** (using tools like Airflow or ML flow) can keep the predictions in sync with the real world.
- This ensures the model stays **reliable and relevant**.

4. Localization and Region-Specific Tuning

Health insurance costs vary by region due to:

- Local **hospital charges**, **taxes**, or **disease prevalence**.
- **Lifestyle variations** across demographics.
  So, customizing the model to include regional parameters helps in:
- **More precise predictions**.
- Increasing **user confidence** in the results.

5. Integration with CRM & Agent Dashboards

Insurance agents use CRM tools daily. By embedding the model into their dashboard, they can:

- **Get real-time premium estimates** while talking to customers.
- Automatically **suggest the best policy** based on the client's input.
- Streamline the **sales process** and improve customer service.

6. AI-Powered Plan Recommendations

Beyond predicting premiums, the system can also **recommend plans**. For example:

- A low-income non-smoker may be suggested a **basic health cover**.
- A smoker over 50 could be offered **premium plans with add-ons**.
  This **personalized experience** can increase sales and reduce confusion for the user.

7. Data Privacy and Security Enhancements

Since personal health data is sensitive, future versions must:

- Comply with laws like GDPR, HIPAA, and India's DPDP Act.
- Use **end-to-end encryption** and **anonymized data handling**.
- Implement **consent mechanisms** before data is used.
  This is essential to **build trust and meet legal requirements**.

8. Multi-Model Ensemble System

Instead of using a single algorithm, you can blend:

- **Linear Regression** (for general trends),
- **Decision Trees** (for decision rules),
- **XGBoost or Random Forests** (for edge cases).
  This ensemble approach can:
- Handle **outliers** better,
- Improve **model robustness**, and
- Offer **higher accuracy** across diverse datasets.

9. Health Benefit Gamification Integration

To **encourage healthy habits**, insurance companies can:

- Link apps like **Google Fit** or **Apple Health**.
- Offer **premium discounts** if users meet health goals (e.g., 10,000 steps/day, BMI under 25).
- This promotes **preventive health** and increases **customer retention** through positive incentives.

10. Voice Assistant Integration

In the age of smart assistants, voice integration adds accessibility:

- Users can ask, "**What is my insurance premium estimate?**"
- The assistant fetches stored profile data and reads out the result.
- This is useful for **elderly users** or those with **low digital literacy**.

11. Cross-Industry Expansion

The same architecture can be used for:

- Auto Insurance: Using driver history and vehicle type.
- Life Insurance: Based on age, family history, habits.

- Travel Insurance: Based on destination risk, duration, traveller health.
  Thus, your project becomes a **flexible solution across the insurance domain**.

12. Promoting Insurance Awareness and Financial Planning

This is a **socially impactful application** as well. Many people:

- Don’t understand how **health conditions affect premiums**.
- Are unaware of the **importance of being insured**.
  You can integrate educational features:
- Show users how quitting smoking or losing weight affects premiums.
- Add **"Why Insurance Matters"** popups explaining financial risks without coverage.
- Display **tips** for getting better rates.

This transforms the tool from just a calculator to a **financial literacy and planning assistant**, especially helpful for students, gig workers, and people new to insurance.

Chapter 7: Weekly Log

Week 1-

- **Activities Completed**:
  - Submitted the finalized problem statement, project objectives, and scope.
  - Conducted preliminary research on health insurance datasets, industry terminologies (e.g., actuarial risk, premium calculation), and regulatory requirements.
  - Set up the development environment, including installation of Python libraries (Pandas, NumPy, Scikit-learn) and tools (Jupyter Notebook, PyCharm, Streamlit).

Week 2-

- **Activities Completed**:
  - Acquired and loaded the dataset containing 50,000+ records of health insurance applicants.
  - Performed data quality checks:
    - Identified and handled missing values (e.g., imputed median values for numerical features).
    - Detected outliers using IQR and boxplots (e.g., capped unrealistic ages > 100).
  - Conducted Exploratory Data Analysis (EDA):
    - Visualized distributions using histograms (e.g., age, BMI).
    - Analyzed correlations via heatmaps (e.g., income vs. premium).

Week 3-

- **Activities Completed**:
  - **Feature Engineering**:
    - Encoded categorical variables (e.g., "Bronze/Silver/Gold" plans → numerical labels).
    - Derived a composite *risk_score* from medical history, BMI, and smoking status.
  - **Data Preprocessing**:

- Normalized numerical features (e.g., Z-score scaling for income).
- Split data into training (70%) and testing (30%) sets.

Week 4-

- **Activities Completed**:
  - Implemented and trained three regression models:
    - **Linear Regression** (baseline).
    - **Ridge Regression** (L2 regularization).
    - **XGBoost Regressor** (ensemble method).
  - Evaluated models using metrics:
    - **XGBoost outperformed others** (R^2: 0.978 vs. Linear Regression's 0.928).

Week 5-

- **Activities Completed**:
  - Optimized XGBoost hyperparameters via RandomizedSearchCV:
    - Tuned learning_rate, max_depth, and n_estimators.
  - Validated improvements using cross-validation (5-fold).
  - Finalized the XGBoost model for deployment.

Week 6-

- **Activities Completed**:
  - Developed a **Streamlit web application**:
    - Designed input forms for user demographics/health metrics.
    - Integrated the trained XGBoost model for real-time premium predictions.
  - Tested UI/UX and debugged edge cases (e.g., invalid inputs).

Week 7-

- **Activities Completed**:
  - Conducted end-to-end testing with real-world data samples.

- Addressed deployment issues (e.g., dependency conflicts on Streamlit Cloud).
- Prepared technical documentation:
    - User manual for underwriters.
    - Model explainability report (feature importance plots).
- Finalized the internship report and presentation slides.

References

IEEE Format References

1. **Prediction of Healthcare Insurance Costs**
   S. Albalawi, L. Alshahrani, N. Albalawi, R. Alharbi, and A. Alhakamy, "Prediction of healthcare insurance costs," *Computers and Informatics*, vol. 3, no. 1, pp. 9–18, 2023.

2. **Medical Insurance Premium Prediction with Machine Learning**
   M. S. Patil, S. Kulkarni, and S. Khurpe, "Medical insurance premium prediction with machine learning," *International Journal of Innovations in Engineering Research and Technology (IJIERT)*, vol. 11, no. 5, pp. 5–12, 2024.

3. **Forecasting Health Insurance Premium Using Machine Learning Approaches**
   S. Dutta, P. Bose, and S. K. Bandyopadhyay, "Forecasting health insurance premium using machine learning approaches," *Asia-Pacific Journal of Science and Technology*, vol. 28, no. 06, pp. 1–13, 2022.

4. **Machine Learning-Based Regression Framework to Predict Health Insurance Premiums**
   K. Kaushik, A. Bhardwaj, A. D. Dwivedi, and R. Singh, "Machine learning-based regression framework to predict health insurance premiums," *International Journal of Environmental Research and Public Health*, vol. 19, no. 13, p. 7898, 2022.

5. **Interpreting the Premium Prediction of Health Insurance Through Random Forest Algorithm**
   V. Rao, M. Iswarya, S. A. Hamza, and B. Satish, "Interpreting the premium prediction of health insurance through random forest algorithm using supervised machine learning technology," *International Journal of Innovative Science and Research Technology*, vol. 8, no. 5, pp. 726–731, 2023.

6. **Analysis of Cost Prediction in Medical Insurance Using Modern Regression Models**
   H. M. Alzoubi et al., "Analysis of cost prediction in medical insurance using modern

regression models," in *2022 International Conference on Cyber Resilience (ICCR)*, pp. 1–10, IEEE, 2022.

7. **Random Forest Regression with Hyper Parameter Tuning for Medical Insurance Premium Prediction**
   V. S. Prakash et al., "Random forest regression with hyper parameter tuning for medical insurance premium prediction," *International Journal of Health Sciences*, vol. 6, no. S6, pp. 7093–7101, 2022.

8. **Insurance Risk Prediction Using Machine Learning**
   R. Sahai et al., "Insurance risk prediction using machine learning," in *The International Conference on Data Science and Emerging Technologies*, pp. 419–433, Springer, 2022.

9. **Health Insurance Cost Prediction by Using Machine Learning**
   A. Sahu et al., "Health insurance cost prediction by using machine learning," in *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*, 2022.

10. **Prediction of Health Insurance Price Using Machine Learning Algorithms**
    S. Goel and A. Chaudhary, "Prediction of health insurance price using machine learning algorithms," in *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1345–1350, IEEE, 2024.

11. **Health Insurance Cost Prediction Using Regression Models**
    S. Panda et al., "Health insurance cost prediction using regression models," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, vol. 1, pp. 168–173, IEEE, 2022.

12. **Comparative Analysis of Machine Learning Algorithms for Health Insurance Pricing**
    Y. T. Bau and S. A. M. Hanif, "Comparative analysis of machine learning algorithms for health insurance pricing," *JOIN: International Journal on Informatics Visualization*, vol. 8, no. 1, pp. 481–491, 2024.

13. **Machine Learning Versus Regression Modelling in Predicting Individual Healthcare Costs**

A. Vimont, H. Leleu, and I. Durand-Zaleski, "Machine learning versus regression modelling in predicting individual healthcare costs from a representative sample of the nationwide claims database in France," *The European Journal of Health Economics*, vol. 23, no. 2, pp. 211–223, 2022.