Explainable AI for Cost Prediction of Medical Insurance

Abstract

The Health Insurance market is expanding at a staggering rate worldwide due to increased awareness of healthcare needs. Smoking remains a major health hazard, contributing to life-threatening conditions such as lung cancer and heart disease. Medical Insurance provides financial relief to individuals and their families during any health emergency. However, with such growing demand, the system somehow becomes inefficient in analyzing the data and ensuring a fairer price for all. While predictive modeling has been employed for cost estimation, this alone is not sufficient, owing to the 'black-box' nature of ML models. So in this research, SHAP and LIME, two of the most used Explainable Artificial Intelligence(XAI), are being used to understand and interpret the results of our ML Model. For this, Ensemble Techniques such as Random Forest, XgBoost, and Adaboost are used. XgBoost performed the best among all with R-Squared =0.8672, RMSE=0.35, MAE=0.19, and MPSE=0.02. This underscores its effectiveness compared to traditional statistical and machine learning approaches. Furthermore, SHAP and LIME were used to identify the key features influencing medical insurance charges and to interpret feature interactions. The research not only uncovers the major players in assessing the insurance cost but also emphasizes the role of Explainable AI in promoting fairness and transparency in healthcare insurance pricing.

1 Introduction

Medical insurance plays a pivotal role in protecting individuals and families from financial shocks associated with unexpected health events. However, determining premiums remains a complex task, particularly for populations engaging in high-risk behaviors such as smoking. Smoking is a well-documented driver of chronic illnesses, including cancers, cardiovascular disease, and respiratory disorders, substantially increasing both the frequency and cost of medical claims[1]. Smoking is a well-known cause of long-term health challenges such as cancer, heart disease, and respiratory problems. This makes medical claims happen more frequently and cost more. Because of this, insurance companies often charge smokers premiums that are 30 percent to 50 percent higher than nonsmokers, as they are more likely to get sick and eventually become a financial liability[2]. Accurate risk stratification and fair premium pricing, especially for smokers, present ongoing challenges.

Standard actuarial methods work well for standard risk factors, but they cannot fully capture the complex, nonlinear relationships between demographic, clinical, and behavioral factors that affect insurance costs. With the rise of machine learning, predictive accuracy

has been improved through advanced algorithms capable of learning complex relationships from large, heterogeneous health datasets.

However, the deployment of these machine learning models introduces concerns related to interpretability, transparency, and fairness[3]. Health insurance decisions have significant personal and societal impacts, and "black box" predictions can make people, regulators, and even insurers less trusting of these models; hence, there is a growing shift towards XAI[3]. XAI methodologies, such as SHAP [4], LIME[5], and counterfactual analysis, enhance the understanding of AI models, allowing the users to have a complete understanding of their assumptions, strengths, and shortcomings[3].

2 Research Significance

The present research addresses these issues by:

- Quantitatively assessing the cost impact of smoking on health insurance premiums using advanced ML techniques.
- Applying leading XAI tools to uncover and visualize the key drivers behind insurance cost predictions for smokers;
- Evaluating fairness and transparency in premium-setting to support ethical, user-centered insurance practices.

By integrating XAI with machine learning in the context of medical insurance, this study aims to provide a transparent, evidence-based framework for fairer pricing mechanisms, ultimately contributing to greater trust and equity in healthcare financing.

3 Literature Review

3.1 Overview of Health Insurance

A health insurance policy offers financial support during medical emergencies. It is an agreement between the policyholder and an insurance provider that helps cover expenses related to sickness or any injury. Under this agreement, the policyholder pays a regular premium and , in return, the insurer covers a wide range of medical services such as hospitalization, surgery, medication, and preventive care. Like other types of insurance, the risk is shared between a group of individuals.An insurer can create a regular finance structure, like a monthly premium or payroll tax, to supply the funds to pay for the health care benefits outlined in the insurance agreement by estimating the total risk of health risk and health system expenses over the risk pool. In case of private insurance, the contract may be lifelong or renewable on an annual or monthly basis. Additionally, it may be required for all citizens in case of national plans.

3.2 Smoking and Health Risks

The tobacco epidemic remains one of the biggest public health crises the world has ever faced. It is responsible for more than 7 million deaths each year and contributes significantly to long-term suffering from tobacco-related diseases[1]. Smokers are 2-4 times more vulnerable to heart attacks and account for about 80% of cases of chronic obstructive pulmonary disease (COPD)[6]. Due to these reasons, health insurance is crucial for smokers.

Key health risks for smokers include:

- Lung cancer and Blood cancer
- Heart diseases
- Chronic obstructive pulmonary disease (COPD)
- Reduced life expectancy (approximately 13+ years shorter than non-smokers)
- Higher surgical complication rates
- Reduced medication effectiveness[2]

3.3 Importance of Risk Differentiation

In most health insurance systems, people are grouped into a risk pool. Health plans can be broadly classified into two types:

1. Individual Health Plans
2. Group-based Plan

In individual health plans, smokers are usually charged 20-50% more than non-smokers[2]. This is known as smoking loading. So, here the non-smokers are somewhat protected.

In group-based plans, both smokers and non-smokers are placed in the same pool. So, since smokers face higher health-related risks, they incur higher health bills, so the overall cost of claims for the insurer rises. Eventually, the insurance company raises premiums for everyone in the group, leading to non-smokers paying more because of the shared pool.

ML-based predictions would help the insurers to differentiate smokers from non-smokers more precisely and charge non-smokers according to their own risk. But there are certain challenges that need to be addressed:

- Bias: It must be ensured that the ML models do not inherit the biases present in the training data or other external factors[7].
- Opacity: ML models operate as black-box models. The 'black box' refers to the inability to interpret how the model produced the particular prediction or made its decision.[8].

These challenges can be mitigated with the use of XAI in ML models. Interpretability provided by XAI is crucial when complex machine learning models are deployed for high-stakes decision-making situations[9].This becomes more important in sectors like healthcare where XAI helps clinicians , patients and stakeholders by providing interpretable justifications creating trust among them.[10].

4 Related Work

The intersection of machine learning and healthcare insurance has evolved dramatically over the past few years, and researchers increasingly recognize that predictive accuracy alone is not sufficient for real-world deployment. This evolution has been driven by three key factors: regulatory demands for algorithmic transparency, the need for fair risk assessment across diverse populations, and the growing sophistication of explainable AI techniques.

4.1 Evolution of Machine Learning in Insurance Cost Prediction

Early attempts at automating insurance cost prediction relied heavily on traditional statistical methods. However, the landscape began to change around 2020 when researchers began exploring ensemble methods. Orji and Ukwandu [11] demonstrated that combining XGBoost, Gradient Boosting, and Random Forest could achieve impressive predictive performance, with their XGBoost model reaching an R^2 of 0.88 on the widely-used Kaggle medical insurance dataset. What made their work particularly noteworthy was not just the accuracy, but their systematic application of SHAP values to understand feature contributions, marking one of the first comprehensive XAI analyses in this domain.

The momentum continued with Srinivasagopalan's innovative approach using artificial neural networks, achieving a remarkable prediction accuracy of 92. 72% [12]. His work challenged the prevailing wisdom that ensemble methods were always superior, showing that deep learning could outperform traditional approaches when properly implemented. However, the black-box nature of neural networks underscored the growing need for interpretability solutions.

More recently, researchers have begun to address the practical challenges of deployment. The 2024 study published in the Journal of Engineering Sciences focused on mobile deployment of RF and XGBoost predictors, achieving an R^2 of 0.871 while demonstrating the feasibility of real-time prediction systems [13]. This practical orientation reflects the field's maturation from academic exercises to production-ready solutions.

4.2 Comparative Analysis of Existing Approaches

Table 1 presents a systematic comparison of key contributions in explainable medical insurance cost prediction, highlighting the evolution from simple accuracy-focused studies to comprehensive interpretability frameworks. The comparative analysis presented in Table 1 reveals several important gaps in the existing literature. Most of them have focused either on achieving high predictive accuracy or on using a single interpretability framework. This proposed study aims to bridge this gap by combining two complementary XAI approaches (SHAP and LIME) and applied it across multiple ensemble models (RF, GBM, and AdaBoost) to deliver both global and instance-level insights into feature contributions. Our dual-framework approach allows a richer, more transparent interpretation of how factors like age, BMI, and smoking status influence premiums.

...

Table 1: Comparative Analysis of Related Work in Explainable AI for Medical Insurance Cost Prediction

Ref.	Dataset & Size	ML Models	XAI Method(s)	Best Result	Notable Contribution
[6]	Kaggle Medical-Cost (986 records)	XGB, GBM, RF	SHAP, ICE	$R^2 = 0.88$	First comprehensive XAI analysis for premium pricing
[12]	Enhanced dataset (synthetic enriched)	ANN	Permutation FI	Accuracy 92.72%	Demonstrates ANN superiority over ensemble methods
[13]	Kaggle (1,338 records)	LR, RF, GB, XGB	-	$R^2 = 0.871$	Mobile deployment framework
[11]	Multi-factor dataset (various sizes)	RF, GB, LR, SVM	—	GB: 87.78%	Gradient Boosting optimization study
[14]	Underwriting context	Various	LIME, SHAP, counterfactuals	—	Regulatory compliance framework for XAI
[15]	Claims prediction dataset	SVM, DT, RF, LR, XGB, KNN	Feature Analysis	XGB: $R^2 = 79\%$, RF: $R^2 = 77\%$	Focus on fairness and bias detection
[10]	Healthcare dataset	Logistic Regression	LIME, SHAP	Comparative study	Systematic comparison of LIME and SHAP methodologies

5 Explainable AI in Healthcare

5.1 A Brief Introduction

Explainable artificial intelligence refers to set of methodologies which help the users to understand the logic behind the decision taken by the model. Rather than treating model result as opaque predictions, XAI techniques aim to clarify why a model made a particular decision, thereby enhancing interpretability and enabling users to assess the model's reasoning. This not only builds trust, but it also makes deployment more responsible and useful, especially for users who don't know much about technology. [3].

5.2 Importance in Healthcare

One of the biggest problems with using AI in important areas like healthcare is that people don't trust the model output because they don't understand how the model works. Getting a better understanding of how the algorithm was made can greatly improve this trust. This will make people more confident in the results produced by AI applications, rather than just seeing them as prototypes.[3].

- **Transparency:** Transparency involves giving clear insights into how AI models work. This principle emphasizes on clear and easy understanding of model's processes, data usage, and decision-making pathways. Transparent systems allow users to assess how the inputs are transformed into outputs and check the accuracy of the AI's predictions with transparent systems.[14].

- **Interpretability:** It is the primary principle of XAI.This means that users should be able to understand how the AI model works. Interpretability makes sure that everyone involved, including doctors and patients, can understand how an AI system comes

to its conclusions[16]. This is achieved through visualizations that illustrates model behavior, rule-based explanations, and feature importance scores which shows which input feature influenced the most in model,s decision.

- **Trustworthiness:** One of XAI's main goals is to make people trust AI systems. XAI helps users trust the AI's decisions by giving them clear and easy-to-understand explanations. This makes AI more widely accepted and used in decision-making[3].

5.3 Key XAI Methods

5.3.1 SHAP : SHapley Additive exPlanations

SHAP is a feature-oriented method which uses game theory methods to explain the outcomes of a machine learning model[4]. It gives a detailed understanding of how each input feature affects the prediction. This not only ensures fairness but also makes it easier for everyone to understand. It is model agnostic, which means it can be used with any machine learning model regardless of architecture, training method and type[4]. For each sample $x = [x_1, x_2, \ldots, x_n]$, the contribution of each feature x_j to the model prediction $f(x)$ is computed using Shapley values, treating the features $x_1, x_2, \ldots, x_n$ as players in a cooperative game defined by a characteristic function v and player set $N = \{x_1, \ldots, x_n\}$. The Shapley value $\phi_j(v)$ for feature x_j is defined as:

$$\phi_j(v) = \sum_{S \subseteq N \setminus \{j\}} W_{S,N} \left(v(S \cup \{j\}) - v(S) \right), \tag{1}$$

where S is subset of features that doesnt include the features x_j, and $W_{S,N}$ is the weight factor given by:

$$W_{S,N} = \frac{|S|!\,(|N| - |S| - 1)!}{|N|!}. \tag{2}$$

Here:

- $|N|!$ is total number of ways of forming a coalition
- $|S|$ is size of subset S,
- $|S|!$ is number of ways coalition S can be formed.

As described by Sadeghi et al.[6], the SHAP explanation of the model can be expressed as:

$$g(z') = \phi_0 + \sum_{j=1}^{n} \phi_j z'_j, \tag{3}$$

where:

- $z' = [z'_1, \ldots, z'_n] \in \{0, 1\}^n$ is a binary vector representing whether a feature is included in a coalition,
- $z'_j = 0$ tells that the j-th feature is excluded from the coalition,

- $z'_j = 1$ tells that the j-th feature is included,
- $\phi_0 = E[f(x)]$ is the average model prediction over all feature coalitions.

SHAP provides a mathematically rigorous and interpretable framework to figure out how much each feature affects a model's prediction.[4].

5.3.2 LIME: Locally Interpreted Model-agnostic Explanations

It is a technique that helps us understand how complex machine learning models make decisions[5]. Instead of trying to explain the entire model, it focuses on one prediction at a time. It creates slightly altered versions of the input, for example by graying out parts of an image, and then checks how these changes affect the model's output. Using this information, LIME builds a simple, interpretable model that mimics the behavior of the original one just around that input. This allows us to see which part of the input has significant influence on the prediction.

6 Materials and Methodology

The cost of medical insurance may be influenced by a variety of factors such as age, BMI, smoking status, etc.However not all of these factors have the same influence; some contribute more to the cost while some contribute less. Our goal is to pinpoint the important factors influencing the insurance cost in the medical sector and how we can make premiums fairer and transparent. In the proposed work, we follow a structured approach to analyze the impact of features involved and predict the insurance costs.

6.1 Dataset

The medical insurance cost dataset is sourced from Kaggle repository [17]. It consists of various factors influencing medical expenses, such as age, sex, BMI, smoking status, number of children, and region. This dataset comprises of 2700 rows and 7 columns.

Table 2 shows the statistical analysis of the features taken

Table 2: Descriptive Statistics for Medical Insurance Dataset

Statistic	**Age**	**BMI**	**Children**	**Charges**
Count	1338	1338	1338	1338
Mean	39.207	30.663	1.095	13270.42
Std Dev	14.050	6.098	1.205	12110.01
Min	18.000	15.960	0	1121.87
25%	27.000	26.296	0	4740.29
50% (Median)	39.000	30.400	1	9382.03
75%	51.000	34.694	2	16639.91
Max	64.000	53.130	5	63770.43

6.2 Exploratory Data Analysis(EDA)

To understand feature distributions and potential relationships with insurance charges, we conducted exploratory data analysis:

1. **Numerical Features:**
   The distributions of numerical variables such as *age*, *BMI*, *children*, and *charges* were examined using histograms. The analysis revealed the following patterns:

   - *Age* exhibited a broadly uniform distribution across the dataset.
   - *BMI* values were concentrated within the "overweight" and "obese" categories, with the presence of a few outliers.
   - *Charges* displayed a pronounced right-skewed distribution, with several high-cost outliers.

2. **Categorical Features:**
   Count plots for *sex*, *smoker*, and *region* highlighted the following observations:

   - The dataset demonstrated a near-balanced representation of *sex*.
   - A clear imbalance between *smokers* and *non-smokers* was observed.
   - The *region* variable showed a relatively uniform distribution across categories. All the results are shown in Figure 5

6.3 Feature Engineering

To improve model performance and interpretability, the raw dataset underwent several feature engineering steps.

- **Binning and Categorization:** Age was discretized into five risk-based intervals (18–29, 30–39, 40–49, 50–59, and 60–64), enabling the model to capture age-related risk tiers rather than treating age as a continuous variable. Similarly, BMI was categorized according to standard medical cutoffs: underweight (< 18.5), normal (18.5–25), overweight (25–30), and obese (> 30).

- **Statistical Methods** Statistical approaches like correlation matrix and t-test are used to extract important features from the dataset. Correlation matrix is used to determine the correlation or strength between two variables while the t-test measures the difference between the mean of two groups relative to variation in each group. We got Smoker vs Non-smoker charges t-test as t=46.66 with p=0 indicating large difference between the two groups compared to the variation within each group.p value of 0 tells that the results are highly statistical significant.

6.4 Data Preprocessing

Numerical features were standardized using z-score normalization using StandardScaler, while the categorical features were transformed using one-hot encoding via OneHotEncoder. Thereafter, all the features were selected and the data was split into a training and testing dataset, with 80% being used for training and 20% for testing.

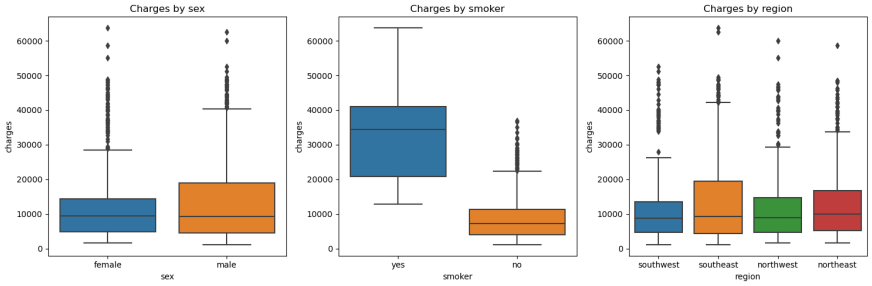


Figure 1: Charges Against Different Features

7 Predictive Modeling Approach and Evaluation Metrics

In this study, three machine learning techniques—Random Forest, Gradient Boosting Machines (GBM), and AdaBoost—were selected to predict medical insurance charges. Several studies have demonstrated that tree-based ensemble algorithms outperform traditional models when dealing with high-dimensional, heterogeneous health datasets [15]. Moreover, these algorithms not only provide high predictive accuracy but also facilitate interpretability through feature importance measures, making them well-suited for integration with Explainable AI (XAI) techniques such as SHAP and LIME [4, 5]. Their strong performance and interpretability potential have made them the algorithms of choice in numerous insurance cost prediction studies [18, 19].

The following sections provide a brief overview of each model, including its principles and key advantages.

7.1 Random Forest

The Random Forest is an ensemble learning method that uses decision trees as its base learners [20]. It is an application of the bagging method, which operates by constructing multiple decision trees each trained on a random set of data. In case of classification, the results are issued by a process called majority voting classifier and for regression we take the average of all the trees. This leads to increased accuracy and less errors.

7.1.1 Advantages of Random Forest

- **Robustness and Stability:** Random Forest aggregates predictions from multiple decision trees, reducing overfitting and improving generalization on unseen data.

- **Handles High-Dimensional Data:** It can manage large numbers of features and automatically model nonlinear relationships and complex feature interactions.
- **Resistant to Outliers & Noise:** The ensemble nature of the Random Forest buffers individual trees from being overly influenced by anomalous data.
- **Feature Importance:** It provides estimates of feature importance, which helps in interpretability for tabular datasets [20].

7.2 Gradient Boosting Machine

Gradient Boosting is a type of machine learning technique that uses boosting in a functional space, using pseudo-residuals as the target instead of residuals as in traditional boosting which enhances the predictive results. [21]. The fundamental principle of this technique is to build the new base-learners such that they are highly correlated with the negative gradient of the ensemble's loss function. Using the standard squared-error loss as an example, the learning process would lead to consecutive error-fitting, although the loss functions used are completely up to interpretation. [22].

7.2.1 Advantages of Gradient Boosting Machine

- **High Accuracy:** Gradient Boosting produces highly accurate predictions on tabular data.
- **Flexibility:** It is used for solving complex datasets and is compatible with both classification and regression problems,
- **Feature Selection:** It assigns a weight to each feature depending on how much it contributes for reducing the loss.
- **Handling Missing Data:** They are capable of addressing the missing data on their own.

7.3 AdaBoost

AdaBoost is a boosting method that emphasizes on misclassified data points in subsequent models by iteratively adjusting the weights which were initially assigned to all the training samples. This makes this effective in reducing bias and variance which makes it best for classification problems but can be sensitive to outliers.

Although AdaBoost is traditionally used with weak base learners, such as decision stumps, it has also been proven to be remarkably effective with stronger models such as deeper decision trees, frequently resulting in even greater overall accuracy [23].

7.3.1 Advantages of AdaBoost

- **Focus on Hard-to-Classify Cases:** AdaBoost adaptively increases the weight of misclassified observations so that the subsequent trees could focus on correcting those difficult instances.

- **Simple and Effective:** It achieves good performance with fewer trees compared to bagging methods.
- **Reduces Overfitting:** By combining weak learners (typically shallow trees), it produces a strong ensemble with reduced variance and improved stability.
- **Works Well on Small Data:** AdaBoost is particularly effective for small datasets.

7.4 Model Evaluation

Evaluating machine learning models is a critical step to assess their predictive performance, generalization capability, and overall reliability. The choice of evaluation metrics depends on the type of problem—whether it is regression, classification, or another task.

In this study, the target variable (*charges*) is continuous, making this a regression problem. Therefore, we employ the following metrics to compare model performance:

7.4.1 R-Squared (R^2)

R^2 is a statistical metrics that indicates the extent to which the independent variables can predict the change in the dependent variable. It generally ranges from 0 to 1.

- $R^2 = 0$ indicates that the model has novariability in the target to explain.
- $R^2 = 1$ indicates perfect predictions where the predicted values match the actual values.

The formula is:

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

where y_i are the true values, $\hat{y}_i$ are the predicted values, and $\bar{y}$ is the mean of actual values.

7.4.2 Root Mean Squared Error (RMSE)

RMSE measures the square root of the average squared differences between the predicted and actual values, bringing the error to the same scale as the target variable:

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

7.4.3 Mean Absolute Error (MAE)

MAE measures the average magnitude of prediction errors without considering their direction. It provides a straightforward measure of prediction accuracy:

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

7.4.4 Mean Absolute Percentage Error (MAPE)

MAPE expresses the error as a percentage of the actual values, making it easy to interpret:

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

This metric helps identify how far predictions deviate from the actual values on a relative scale.

7.4.5 Residual Analysis

Residual plots were used to examine prediction errors across target values. For RF and GBM, residuals were tightly clustered around zero , as seen in Fig 2 and Fig 3 respectively , suggesting strong generalization and minimal bias. AdaBoost(Fig 4) displayed slightly higher variance in residuals, indicating reduced stability for certain high-cost cases.

By analyzing the residual distribution, it was evident that errors increased marginally for very high premium values, particularly among smokers with high BMI, which aligns with the dominant SHAP findings for these features.

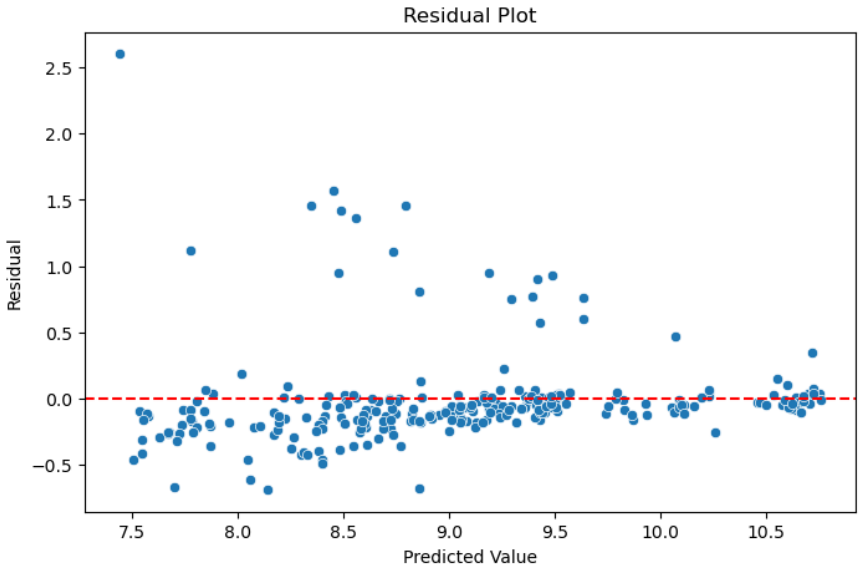


Figure 2: Random Forest Residual Plot

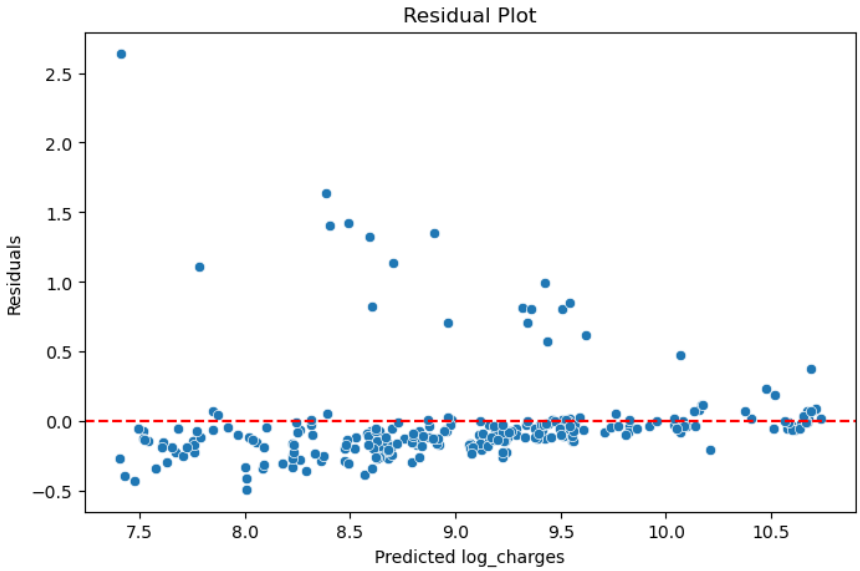


Figure 3: GBM Residual Plot

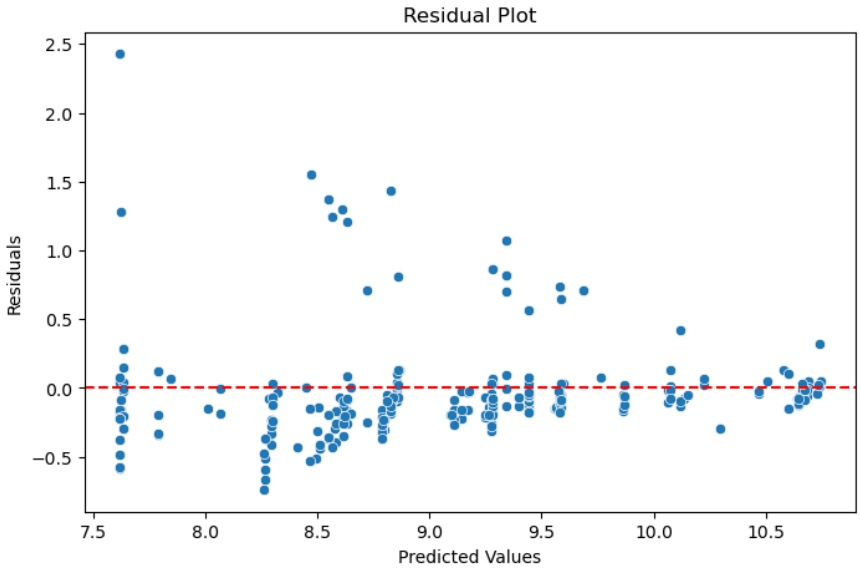


Figure 4: ADA Residual Plot

8 Experimental Result Analysis

This section presents the performance evaluation of the ensemble models Random Forest, Gradient Boosting Machine, and AdaBoost, followed by interpretability analysis using SHAP and LIME, and residual analysis to assess prediction errors.

Table 3: Performance Metrics for Ensemble Models

Model	R^2	**RMSE**	**MAE**	**MAPE**
GBM	0.8672	0.35	0.1900	0.0200
RF	0.8562	0.36	0.1945	0.0218
AdaBoost	0.8544	0.36	0.2174	0.0244

8.1 Model Performance Evaluation

The performance of the models was assessed using four evaluation metrics: R^2, RMSE, MAE, and MAPE as described . Table 3 summarizes the results of the models used for predictions.

Among the models, GBM exhibited the highest R^2 (0.8672) and the lowest RMSE (0.35), indicating superior predictive accuracy and its ability to capture complex, non-linear patterns. RF achieved a slightly lower R^2 (0.8562) but a marginally better MAE (0.1945), suggesting strong stability and robustness. AdaBoost, while competitive, recorded higher errors (MAE = 0.2174, MAPE = 0.0244), reflecting its sensitivity to noisy instances and less stable predictions compared to GBM and RF.

The superior performance of GBM can be attributed to its sequential boosting framework, which reduces residual errors iteratively, whereas RF leverages bagging and random feature selection to control overfitting. AdaBoost's tendency to overemphasize misclassified points likely contributed to its relatively weaker performance.

8.2 Model Interpretability with SHAP

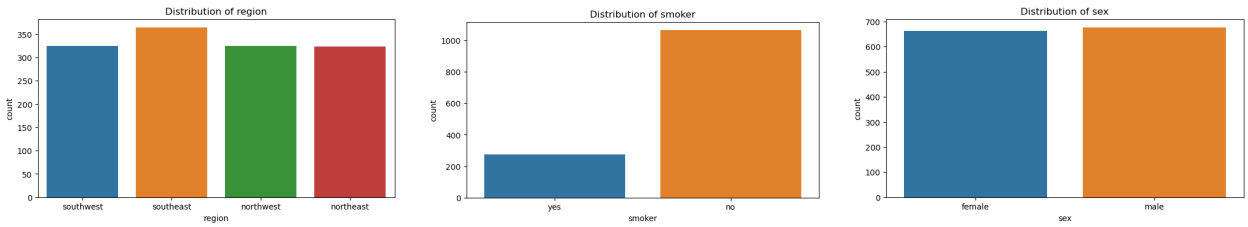
To gain deeper insights into model behavior, SHAP (Shapley Additive exPlanations) was employed to analyze feature contributions across all three models. SHAP bar plots, beeswarm plots, and dependence plots were generated for interpretation.

The SHAP bar plot for RF(Figure 6) revealed that *age* was the dominant driver of predicted charges, followed by the *smoker-BMI* interaction, which highlights the amplified health risk for smokers with high BMI (obese category). Smoking-related features (*smoker-age*, *smoker-yes*) consistently ranked among the top contributors. GBM produced similar feature importance trends, with age and smoker-BMI interactions dominating the predictions(Figure 19). In contrast, AdaBoost's SHAP bar plot(Figure 13) emphasized *smoker-BMI* as the unequivocally dominant factor, followed by BMI and age, due to its sharper partitioning behavior. Features like *children*, *sex*, and *region* had negligible influence across all models. In all the models, children , sex and demographic features have minor effect as compared to other factors discussed above.

The Beeswarm plot also revealed similar interpretations. Y-axis represents the features ranked by their average absolute SHAP values and X-axis represents SHAP values. Positive values for a given feature push the model's prediction closer to the feature being examined . In contrast, negative values push towards the opposite . Color Gradation helps in showing lower to higher values for each feature when compared to other observations with features having higher values(red) pushes the value upwards i.e increase the costs . On the other hand , lower feature values (blue) cause downward movement in predictions, acting as risk-reducing factors. In Random Forest (Figure 7) and GBM(Figure 20) , Age, smoker-BMI,

and smoker-yes routinely pushed predictions upward in high-value cases (red) and in cases of lower age or absent smoker status (blue), strong downward SHAP contributions are seen. In Adaboost (Figure 14), smoker status is the most prominent feature having strongest impact on model's predictions. So for smokers, the predicted charges increases a lot while for non-smokers it decreases.Furthermore ,for smoker-BMI interactions the costs get amplified even more with model predicting very high charges.

Dependence plots further illustrated these relationships. They show the relationship between a specific feature and the predicted outcome for each instance within the data. Dependence graph for age (Coloured by smoker) in Random Forest(Figure 8) showed nearly linear increase in SHAP values for increasing age. Smokers (highlighted by the colour scale) exhibits even higher SHAP values with age , showing that old smokers are at highest risk .In GBM , again Nearly linear trend in SHAP values with increasing age is observed, Big jump is seen for old smokers and this is revealed by Interaction colour-mapping in the upper-most right-hand quadrant of the plot, skyrocketing the prices(Figure 9) .In Adaboost,for BMI below a certain value, the predicted medical charges are much lower but once BMI crosses that threshold, costs jump sharply upward, no matter what other factors say.(Figure 10)



(a) Distribution of Region (b) Distribution of Smoker (c) Distribution of Sex

Figure 5: Categorical distributions in the dataset.

8.3 Instance-level Explanations with LIME

LIME explains the model's predictions by taking individual instances helping us in understanding how model approached every instances and made their predictions. This method gives a local interpretation of the model.

We took three instances 10,50,100 for every model. Red Bars are the features which decrease the costs while Green bars are the features which increase. For Instance 10, who is a smoker, the lime interpretablity for RF(Figure‘23) and GBM model(Figure 12) , it says that Age and Smoker are key factors for pushing the premium cost while BMI and other smoker factors makes it even worse. In AdaBoost(Figure 17), Smoker-BMI almost single-handedly determine the prediction. Instance 100, is a non-smoker as smoker-BMI ¡ 0. Being a non-smoker reduces the cost prediction. However age and BMI are still in play as they affect the predictions negatively but not much stating that the person is young and hence these are important factors for non-smoker. Instance 50 is again a non-smoker but age here is a positive factor pushing the costs up which implies the person is older. In Table 4 we present a comparison of actual and predicted values for selected individual instances using different ensemble models.

Table 4: LIME Explanations and Model Predictions for Selected Instances

Instance	Actual	RF Pred	GBM Pred	ADA Pred
10	9.81	9.76	9.86	9.82
50	9.54	9.56	9.60	9.54
100	8.42	8.53	8.58	8.60

To get an overview for all the instances we drew a LIME heatmap to determine the major contributing feature . Deep red colour indicates strong positive contribution while blue indicates strong negative contribution. Red is consistently seen in smoker related areas indicating a strong contributor to prices as shown in Figure 11,Figure 22,Figure 16.

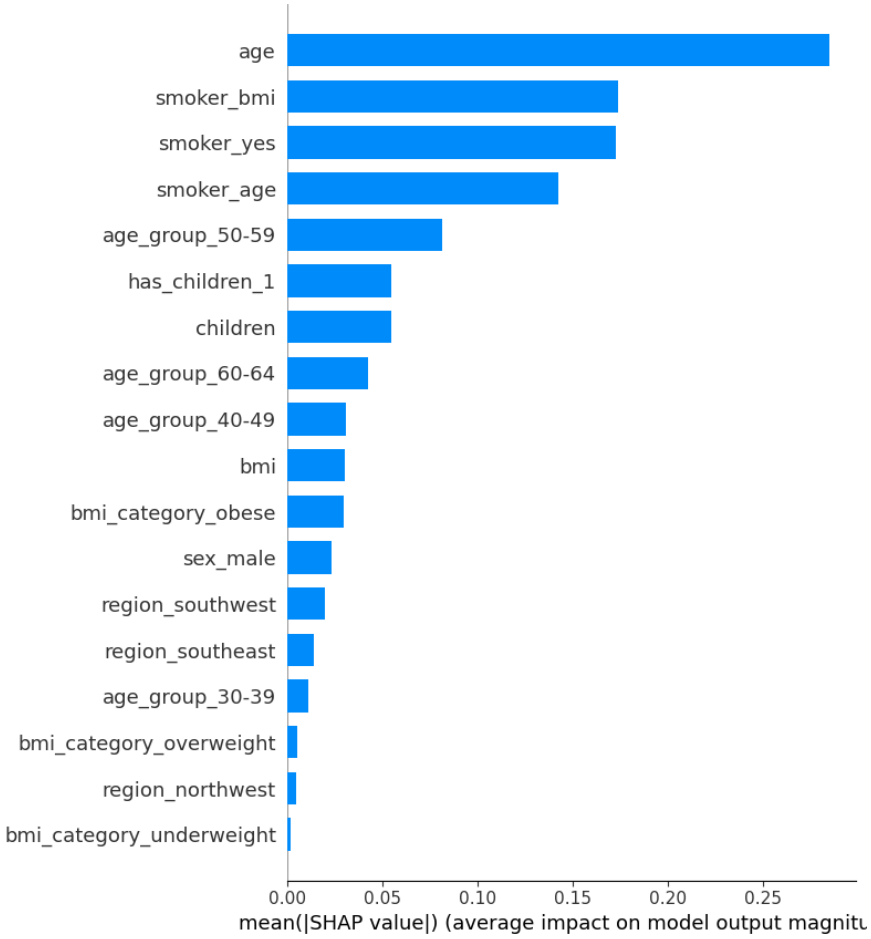


Figure 6: Random Forest Average SHAP Value

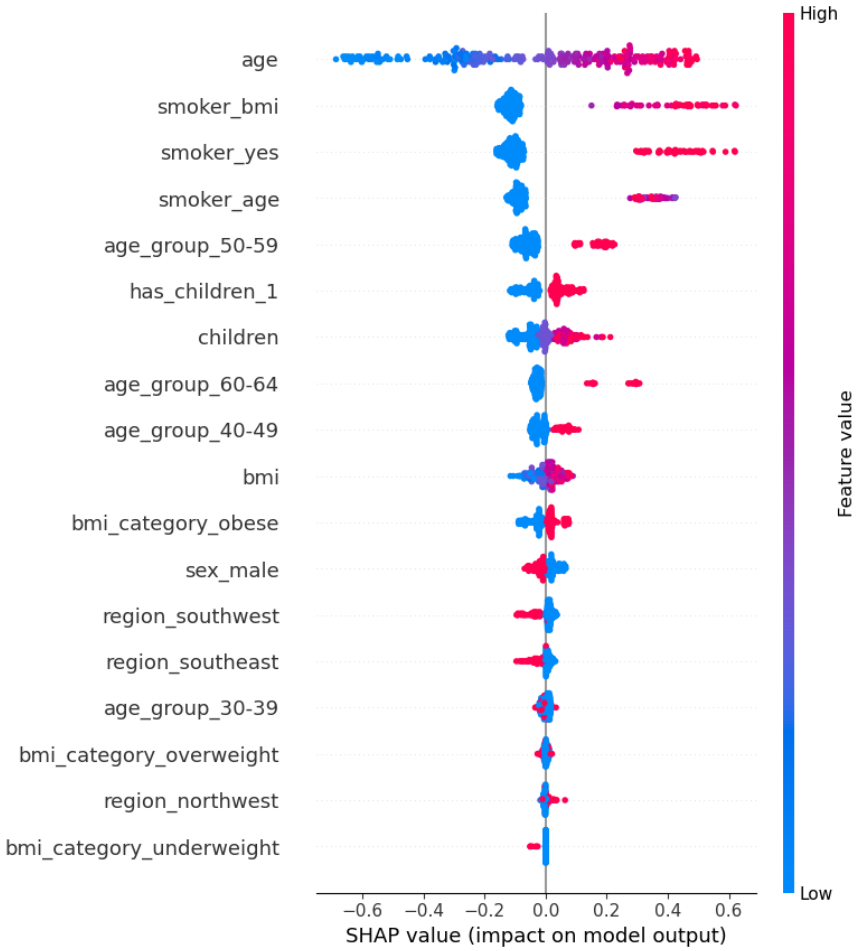


Figure 7: Random Forest SHAP Value Plot

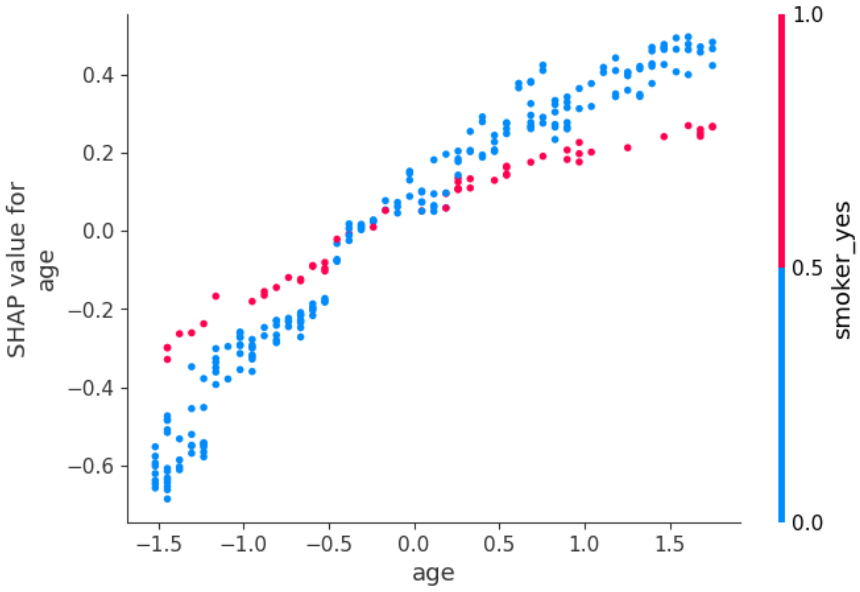


Figure 8: Random Forest SHAP Dependence Plot

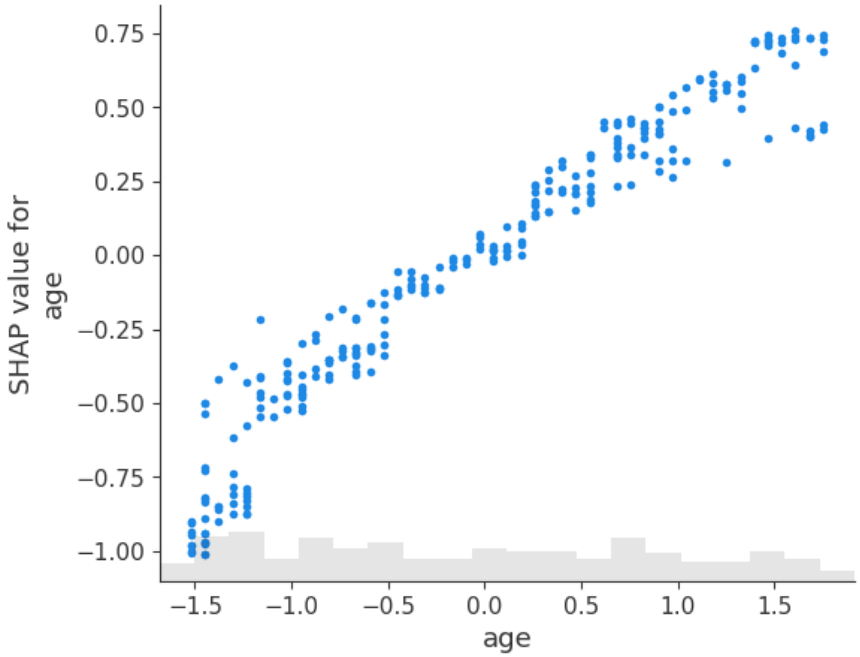


Figure 9: GBM Dependence Plot

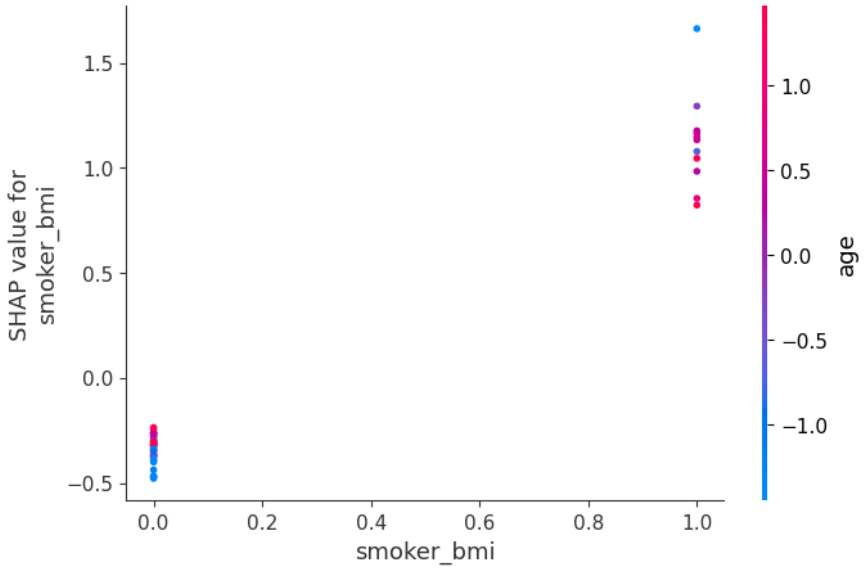


Figure 10: Adaboost Dependence Plot

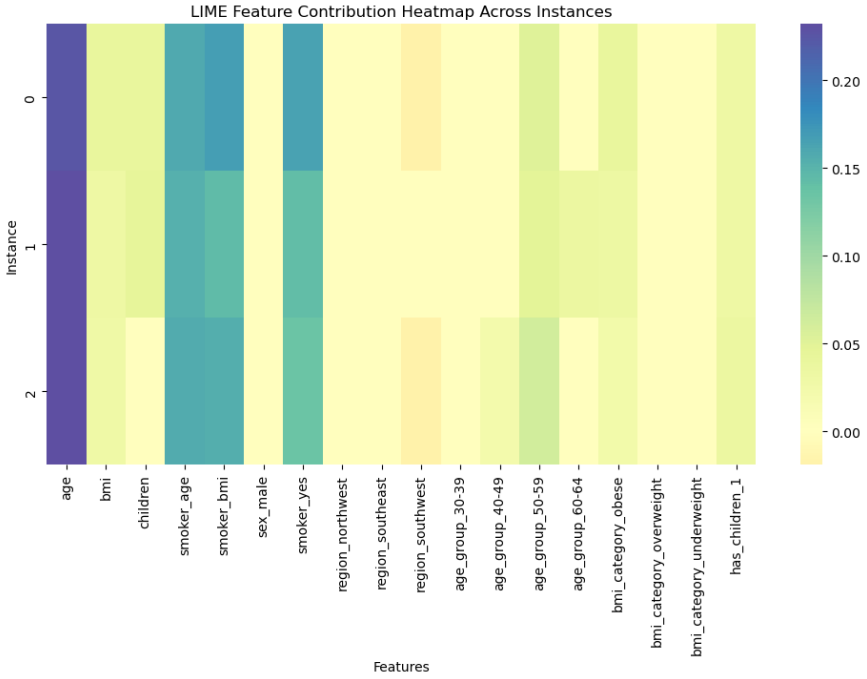


Figure 11: Random Forest LIME Heatmap Visualization

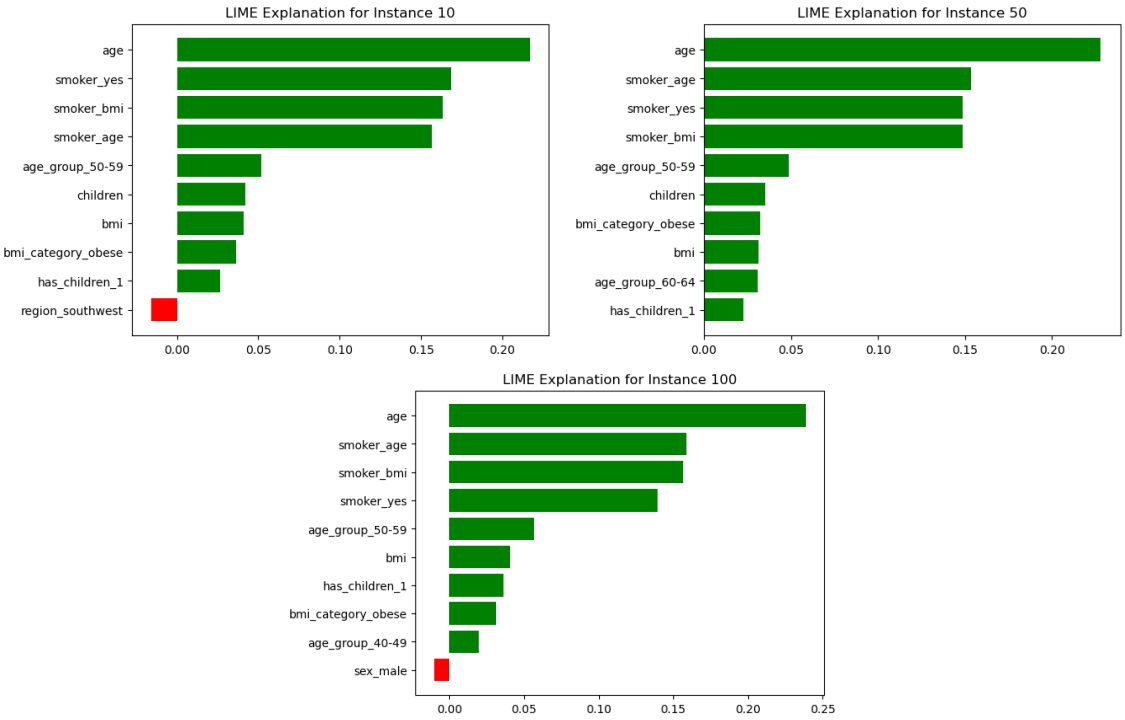


Figure 12: Random Forest LIME Instance-Based Explanations (10, 50, 100)

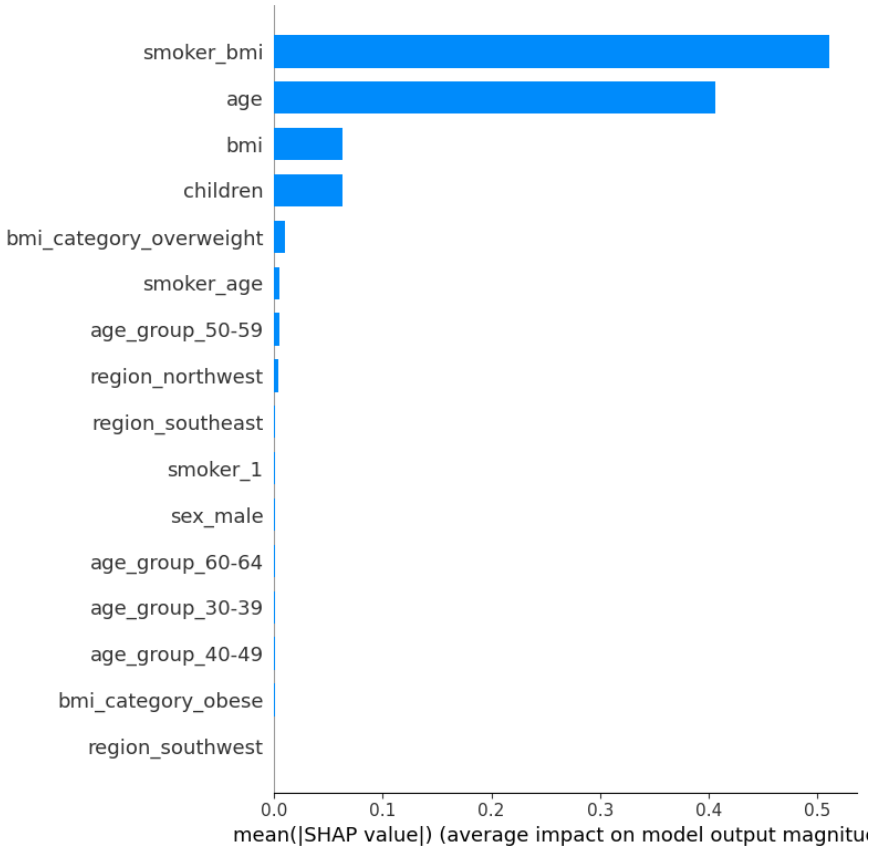


Figure 13: AdaBoost Mean SHAP Values

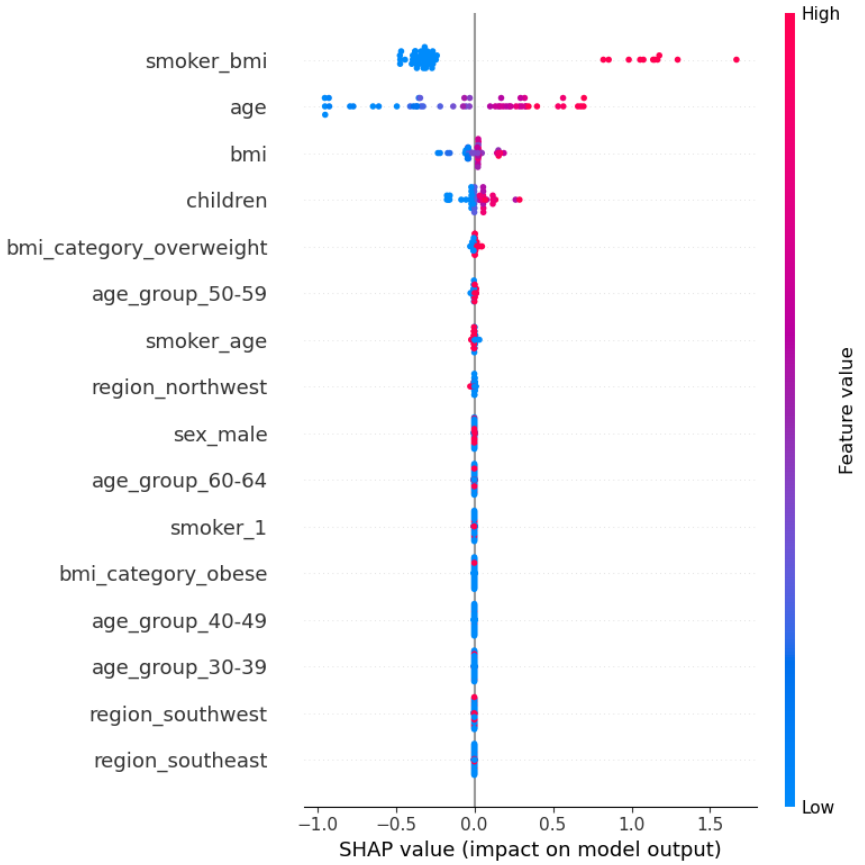


Figure 14: AdaBoost SHAP Values

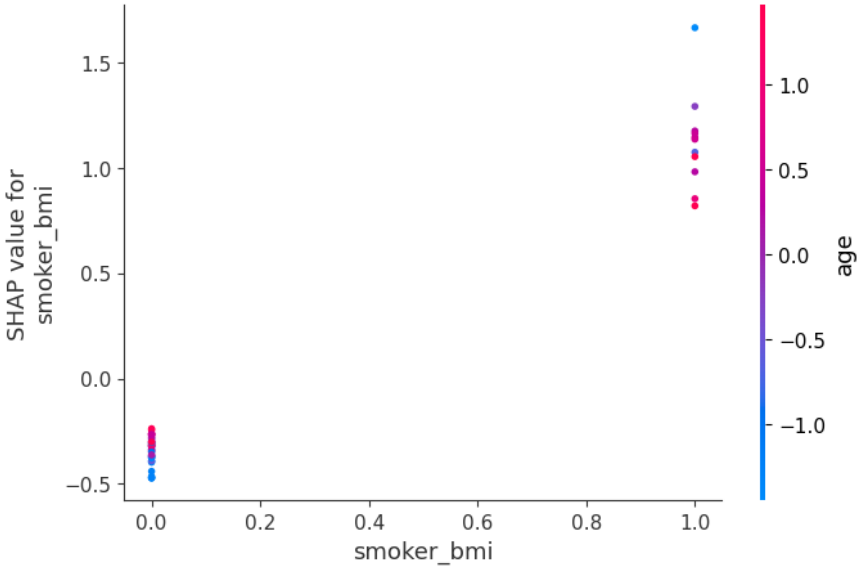


Figure 15: AdaBoost Smoker vs Age SHAP Comparison

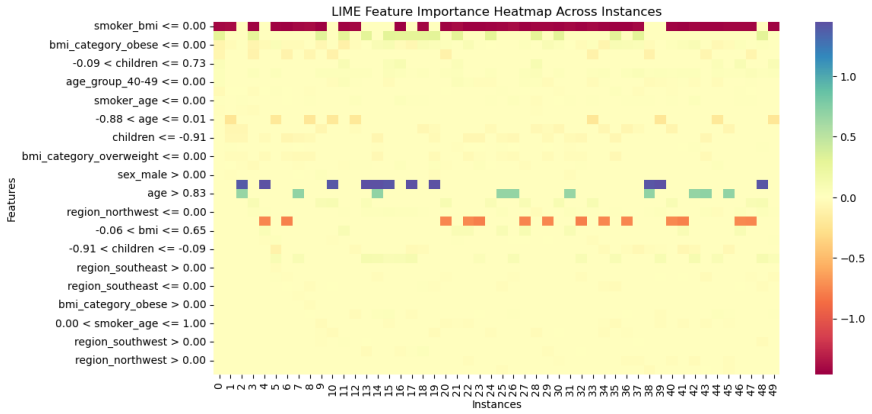


Figure 16: AdaBoost LIME Heatmap

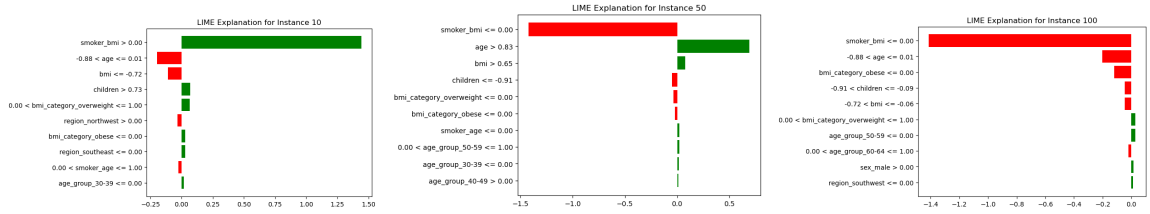


Figure 17: AdaBoost LIME Instances (10–100)

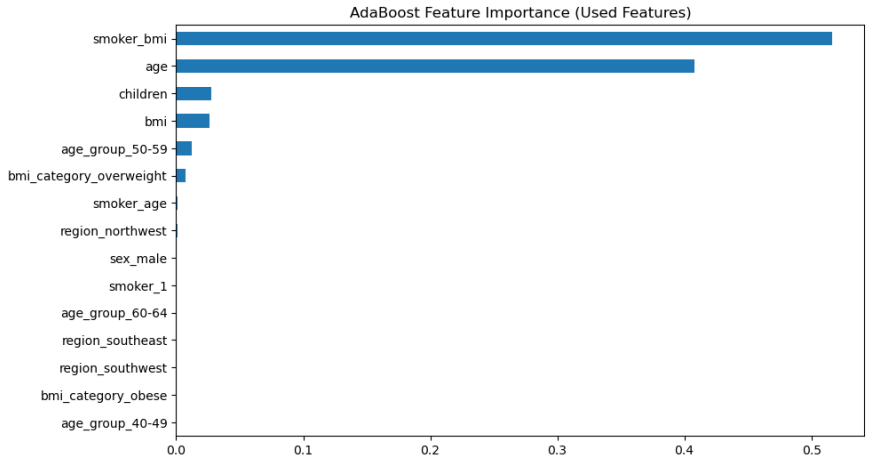


Figure 18: AdaBoost Feature Importance

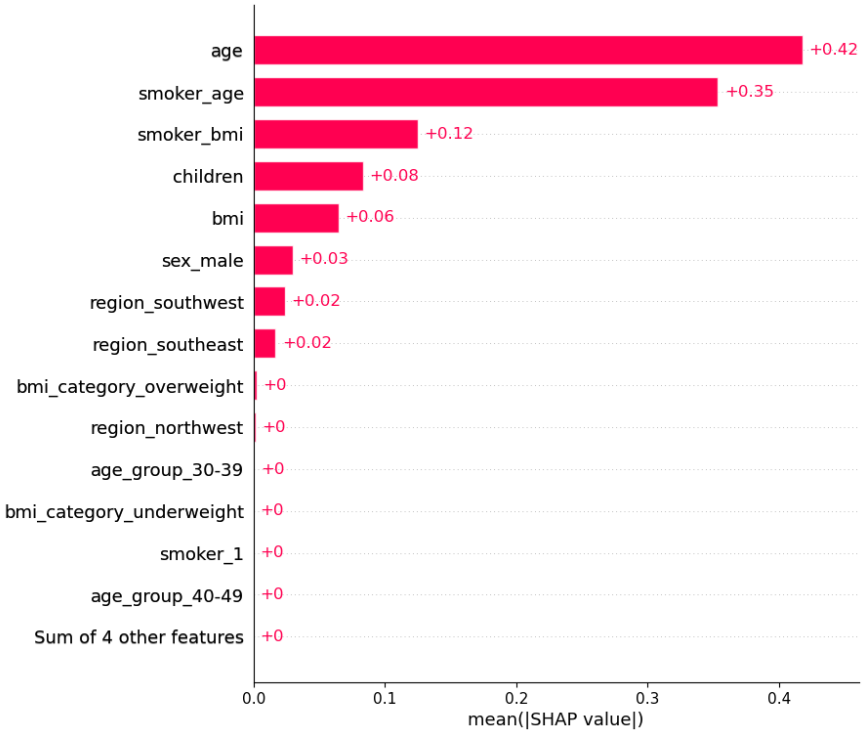


Figure 19: GBM Mean SHAP Values

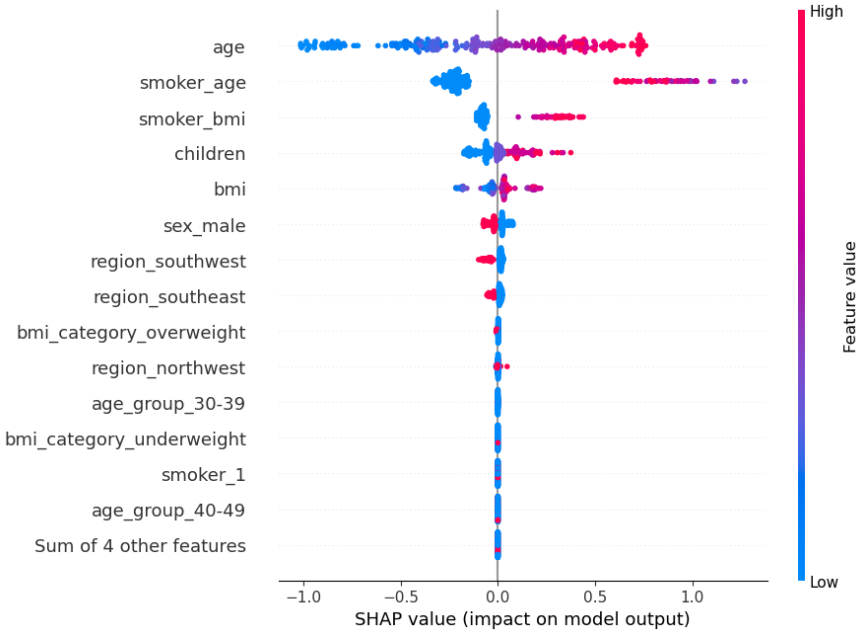


Figure 20: GBM SHAP Values

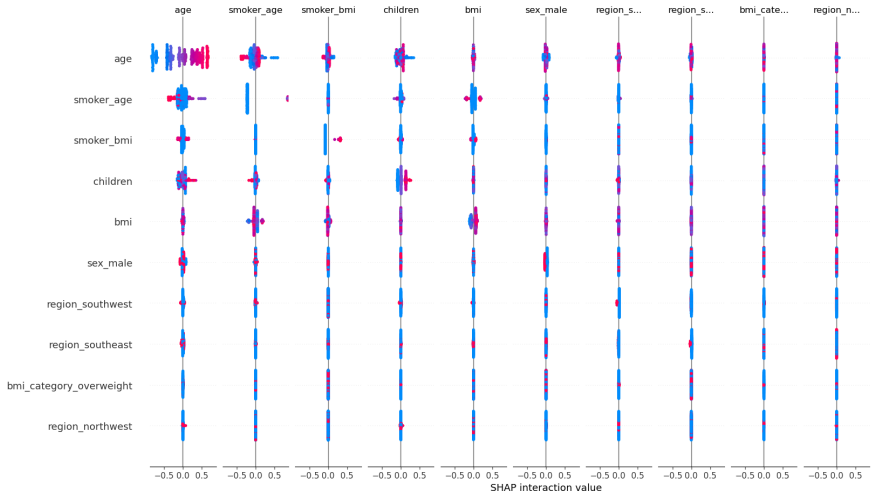


Figure 21: GBM SHAP Interaction Values

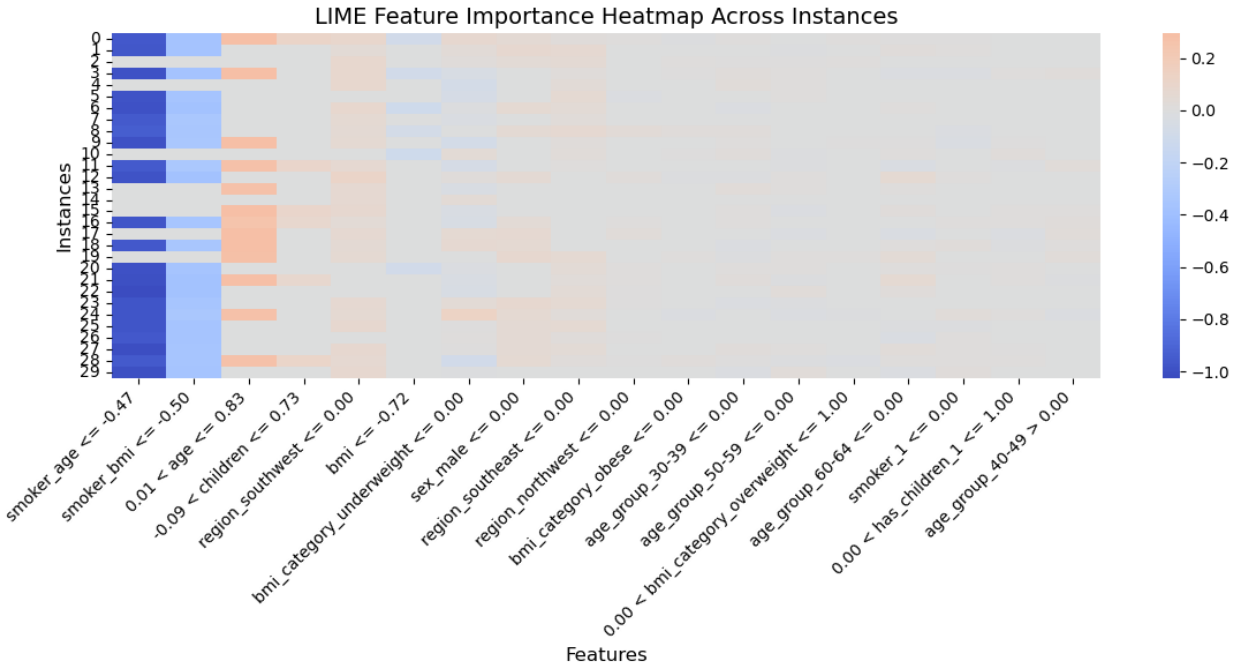


Figure 22: GBM LIME Heatmap

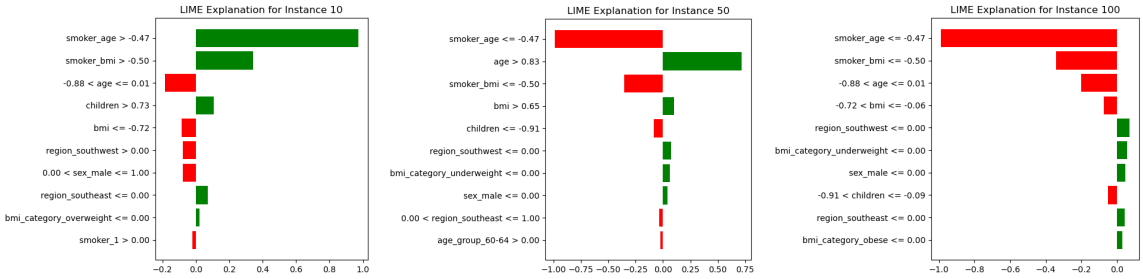


Figure 23: GBM LIME Instance Explanations

9 Conclusion and Future work

This study underscores the primary role of XAI in insurance cost predictions, particularly for smokers and non-smokers. By employing various ensemble learning models, we systematically investigated the importance of features present and how they influence decision making. The results revealed that age and smoker status are the two most important contributing features to increasing premium; their interaction further worsens the effect [4, 5, 15]. These insights can help insurers design pricing systems that ensure fairness and transparency, especially when evaluating individuals based on their smoking habits. Moreover, such models can guide insurance buyers toward selecting policies best suited to their personal health profiles [24]. This research also opens doors for insurers and policymakers to incorporate AI into their decision-making frameworks.

But there are certain barriers which need to be addressed:

- Limitation of large and high-quality insurance datasets, which are necessary to capture the complexities of real-world issues.
- Lack of data on individuals' past medical history, presence of chronic illnesses, frequency of hospital visits, etc., which are crucial.
- The performance and interpretability of models might not generalize well to insurance data from different regions, populations, or insurance policies with different pricing models.

Addressing these limitations will require future collaborations between data scientists and policymakers, which will not only help in improving insurance policies but also build products that are equitable, explainable, and scalable for real-world deployment in the insurance domain.

References

[1] W. H. Organization, "Who report on the global tobacco epidemic, 2019: Offer help to quit tobacco use," World Health Organization, 2019.

[2] M. McLennan, "How life insurance companies view smokers," 2023, available at: https://www.marshmclennan.com/ [Accessed: 2023-09-01].

[3] D. Gunning, "Explainable artificial intelligence (xai)," Defense Advanced Research Projects Agency (DARPA), Tech. Rep., 2017.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[6] V. Sadeghi, S. Pooyan, A. Jalili, R. Moradzadeh, M. Zare, and N. Yazdi, "Explainable artificial intelligence-driven insights into smoking prediction using machine learning and clinical parameters," *Scientific Reports*, vol. 12, pp. 1–12, 2022.

[7] E. O'Reilly, "Bias and fairness in machine learning: a review," *Journal of Artificial Intelligence Research*, 2019, preprint.

[8] L. Cavique, M. Areias, and A. M. Madureira, "The black box of machine learning: Interpretability and transparency," *ACM Computing Surveys*, 2022.

[9] A. Weller, "Challenges for transparency," *arXiv preprint arXiv:1708.01870*, 2017.

[10] M. Mia and M. Hossain, "Explainable artificial intelligence in healthcare: A comprehensive survey," *IEEE Access*, vol. 10, pp. 174 207–174 234, 2022.

[11] I. Orji and D. Ukwandu, "Predictability of medical insurance cost using various machine learning models and explainable ai," *International Journal of Computer Applications*, vol. 176, no. 36, pp. 25–31, 2020.

[12] S. Srinivasagopalan, "Medical insurance cost prediction using artificial neural network," *International Journal of Science and Research*, vol. 9, no. 6, pp. 123–126, 2020.

[13] N. Sharma and A. Das, "Mobile deployment of random forest and xgboost for health insurance prediction," *Journal of Engineering Science*, vol. 15, no. 2, pp. 78–85, 2024.

[14] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[15] M. Sharifi, S. Sarkhosh, M. Gharacheh, and H. Yazdi, "Predictive modelling of healthcare costs using machine learning algorithms," *Health Policy and Technology*, vol. 8, no. 3, pp. 254–262, 2019.

[16] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[17] Kaggle, "Medical cost personal datasets," available at: https://www.kaggle.com/datasets/mirichoi0218/insurance [Accessed: 2023-09-01].

[18] Z.-H. Zhou, "Ensemble methods in insurance risk classification," in *Ensemble Methods*. Springer, 2012, pp. 201–217.

[19] J. Wu, P. Li, and Y. Wang, "A comparative study of ensemble learning methods for insurance fraud detection," *IEEE Access*, vol. 8, pp. 134 254–134 262, 2020.

[20] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[21] W. contributors, "Gradient boosting — wikipedia, the free encyclopedia," 2023, available at: https://en.wikipedia.org/wiki/Gradient_boosting [Accessed: 2023-09-01].

[22] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[23] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[24] A. Kumar, S. Singh, and S. Gupta, "Comparative study of machine learning algorithms for insurance cost prediction," *International Journal of Computer Applications*, vol. 975, p. 8887, 2021.