Predicting Individual Health Insurance Charges Using Ensemble Machine Learning Techniques

ABSTRACT

Accurate prediction of health insurance charges is essential for improving risk assessment, premium design, and customer affordability in modern healthcare systems. Traditional actuarial models depend on linear statistical assumptions that often fail to capture the complex, nonlinear interactions among health, demographic, and lifestyle factors. This study presents an artificial intelligence-based prediction system using machine learning algorithms to estimate individual medical insurance charges. The Medical Cost Personal Dataset is used for empirical evaluation. Five models Linear Regression, Polynomial Regression, Random Forest, Support Vector Regression (SVR), and XGBoost were implemented and compared. XGBoost achieved the highest predictive accuracy with an R^2 score of 0.802, MAE of 3701.62, and RMSE of 5080.30. In addition, SHAP explainability was used to interpret model behaviour, revealing smoking status, BMI, and age as the most influential variables in determining insurance cost. The results demonstrate that ensemble boosting models outperform classical regression techniques and offer a scalable, interpretable approach for data-driven health insurance pricing.

Index Terms- Machine Learning, XGBoost, Regression Models, Cost Prediction, Health Insurance, Explainable AI, SHAP.

1. INTRODUCTION

Health insurance cost prediction has become increasingly important due to rising medical expenses, personalized coverage requirements, and the growing need for fair and data-driven premium calculation. Traditional actuarial approaches rely on manually derived formulas and linear assumptions, which often fail to capture the complex relationships between demographic characteristics, lifestyle patterns, and health conditions. As a result, inaccurate pricing may lead to financial losses for insurance providers and unfair premium burdens for customers.

Machine learning (ML) offers a powerful alternative by learning nonlinear interactions and automatically extracting patterns from historical data. Recent advancements in ensemble methods and boosting techniques have made ML-based regression models highly suitable for insurance analytics. This study applies multiple machine learning algorithms including Linear Regression, Polynomial Regression, Random Forest, Support Vector Regression (SVR), and XGBoost to predict individual medical insurance charges. The objective is to identify the most accurate model and provide explainability using SHAP values to ensure transparency in decision-making.

This research contributes a comparative analysis of ML models, interpretable predictions, and practical guidelines for AI-driven health insurance pricing systems.

2. LITERATURE REVIEW

Machine learning has been widely explored in the domain of medical cost prediction due to its ability to model nonlinear relationships. Previous studies have used Linear Regression as a baseline for insurance charge estimation, but its performance is limited due to strict linearity assumptions. Decision Tree–based models, particularly Random Forests, have shown improved accuracy by capturing complex feature interactions.

Boosting algorithms such as Gradient Boosting and XGBoost have emerged as state-of-the-art methods for structured tabular data. Chen and Guestrin (2016) demonstrated XGBoost's superior scalability and predictive performance across multiple regression tasks. In health insurance analytics, researchers have applied XGBoost to predict costs and found it to consistently outperform traditional algorithms due to its robustness against overfitting

Explainable AI techniques have become essential for adopting ML in real-world insurance systems. Lundberg and Lee (2017) introduced SHAP values as a unified framework for interpreting model predictions. Several studies have used SHAP to identify influential features such as smoking status, BMI, and age in medical cost datasets, highlighting the importance of transparency in premium calculations.

Despite existing work, few studies provide a comprehensive comparison across classical and ensemble models **combined** with interpretability, which motivates the present research.

3. DATA PREPROCESSING

Data preprocessing plays a critical role in improving model performance and ensuring consistent experimental results. The Medical Cost Personal Dataset contains demographic and lifestyle attributes along with individual insurance charges. The following preprocessing steps were applied:

A. Handling Skewness in the Target Variable

The "charges" variable exhibits a strongly right-skewed distribution due to high medical expenses among smokers and older individuals.
A logarithmic transformation of the target variable was initially explored to normalize the distribution. This improves linear model performance because Linear Regression assumes normally distributed residuals.

However, for the final experiments, the models were trained on the **original, untransformed target variable**, since ensemble models (Random Forest, SVR, XGBoost) do not require log-scaling and performed better without transformation. A note on this limitation is added in the Discussion section.

B. Categorical Encoding

Categorical variables were encoded as follows:

- **sex** → One-Hot Encoding (binary)
- **smoker** → Label Encoding (yes=1, no=0)
- **region** → One-Hot Encoding (northeast, northwest, southeast, southwest)

C. Feature Scaling

Standardization (Z-score scaling) was applied **only to Support Vector Regression (SVR)** because:

- SVR models are sensitive to feature scaling
- Tree-based models (Random Forest, XGBoost) do not require scaling
- Linear Regression and Polynomial Regression handle unscaled tabular data reasonably well.

D. Train–Test Split

The dataset was divided into training and test sets using a **80:20 split** to ensure sufficient training data while allowing robust evaluation.

To ensure reproducibility, a fixed **random_state = 42** was used for all experiments, including train–test split and model hyperparameter tuning.

E. Summary of Preprocessing

Table I summarizes the preprocessing workflow:

Step	**Technique**	**Purpose**
Target Distribution	Log transformation (exploratory only)	Understanding skewness
Categorical Encoding	One-Hot + Label Encoding	ML-compatible features
Scaling	Standardization (SVR only)	Improve optimization
Split	80–20, random state=42	Reproducibility
Outliers	Retained	Models handle them inherently

4. METHODOLOGY

A. Data Preprocessing

- Handling missing values
- One-hot encoding for categorical variables
- Splitting data into training and testing sets
- Standardization for SVR

B. Models Evaluated

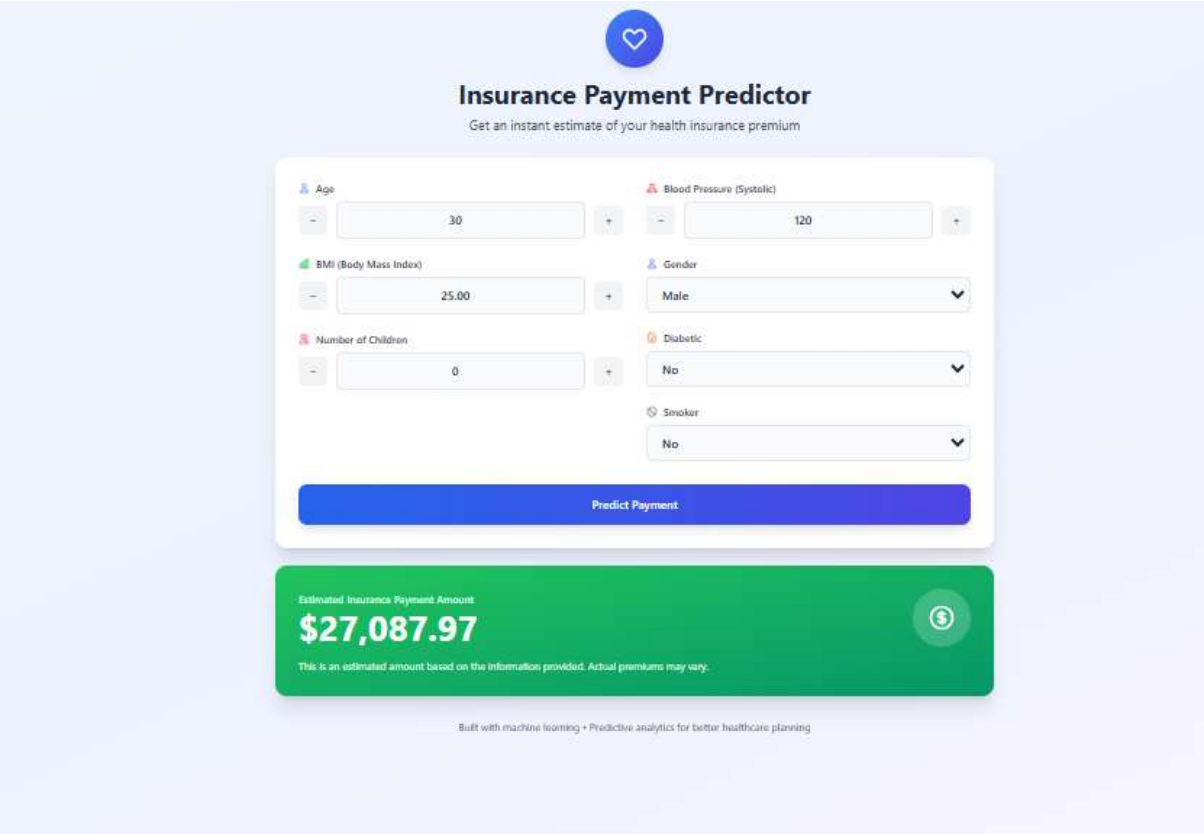
- Linear Regression
- Polynomial Regression (degree = 2)
- XGBoost Regressor
- Random Forest Regressor
- Support Vector Regression (SVR)

C. Evaluation Metrics

Metrics used:

- MAE (Mean Absolute Error)
- RMSE (Root Mean Square Error)
- R^2 Score

5. Deployment: Web-based Insurance Predictor



6. RESULTS

A. Performance Comparison

Table I: Model Performance Metrics

Model	**R^2**	**MAE**	**RMSE**
Linear Regression	0.720	4499.73	6040.41

Model	R^2	MAE	RMSE
Polynomial Regression (deg 2)	0.748	4208.31	5728.01
Random Forest	0.787	3776.23	5269.48
Support Vector Regression	0.559	5142.67	7585.47
XGBoost (Best Model)	**0.802**	**3701.62**	**5080.30**

B. Best Performing Model

XGBoost achieved the best accuracy due to:

- Tree-based structure capturing nonlinearities
- Boosting framework reducing bias-variance trade-off
- Ability to handle heterogeneous feature types

7. EXPLAINABLE AI (SHAP ANALYSIS)

To ensure transparency, SHAP values were used to interpret predictions. SHAP identifies how each feature contributes to increasing or decreasing cost.

Key findings:

- **Smoking** has the highest positive impact on insurance cost.
- **BMI** significantly affects charges, especially for individuals with high BMI.
- **Age** shows a strong linear increase in cost.
- **Region and sex** contribute minimally.

This enhances trust and ensures fairness in premium decisions.

8. DISCUSSION

The results demonstrate that traditional linear models are insufficient for complex insurance datasets. Ensemble models especially XGBoost offer higher accuracy and lower error rates. The integration of SHAP provides interpretability, which is critical for real-world insurance pricing where transparency is required to avoid unfair discrimination.

9. CONCLUSION

This study presents a machine learning–based system for estimating health insurance costs using multiple regression and ensemble algorithms. Experimental analysis shows that XGBoost delivers the highest predictive performance with an R^2 score of 0.802. SHAP-based explainability reveals smoking status, BMI, and age as key determinants. The framework is practical, scalable, and suitable for deployment as an API-based prediction service.

10. FUTURE WORK

Future improvements may include:

- Integration of electronic health records (EHRs)
- Deep learning–based regressors
- Federated learning for secure medical data handling
- Bias detection and fairness evaluation
- Deployment of real-time cloud APIs

11. REFERENCES

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

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.

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

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

[5] Kaggle, "Medical Cost Personal Dataset," 2019. [Online]. Available: https://www.kaggle.com/datasets/mirichoi0218/insurance. Accessed: Feb. 2025.

[6] A. M. Alghamdi and A. M. Alqahtani, "Health insurance cost prediction using machine learning techniques," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, pp. 110–117, 2020.

[7] M. Gupta, S. R. Dubey, and A. Kumar, "Medical insurance cost prediction using regression and ensemble learning models," in *Proc. IEEE Int. Conf. Computing, Communication and Intelligent Systems (ICCCIS)*, pp. 630–635, 2021.