Predictive Analytics for Health Insurance Pricing Using ML and Explainable AI

ABSTRACT

Accurate prediction of health insurance costs is essential for improving risk assessment, premium calculation, and financial planning within the insurance sector. Traditional actuarial models rely on manually crafted statistical assumptions, limiting their ability to capture complex nonlinear interactions among demographic and lifestyle features. This study presents a machine learning–based framework for predicting insurance charges using multiple regression and ensemble algorithms. Five models Linear Regression, Polynomial Regression, Random Forest, Support Vector Regression (SVR), and XGBoost were trained and evaluated on the Medical Cost Personal Dataset.

Experimental results demonstrate that XGBoost outperforms all baseline models, achieving an R^2 score of **0.802**, MAE of **3701.62**, and RMSE of **5080.30**. Furthermore, the study employs SHAP-based explainable AI techniques to identify key contributing factors affecting prediction outcomes, with smoking status, BMI, and age emerging as the most influential variables. The findings highlight the effectiveness of tree-based ensemble models in complex cost-prediction tasks and provide actionable insights for data-driven insurance pricing.

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

1. INTRODUCTION

Health insurance pricing plays a crucial role in determining affordability, risk management, and long-term sustainability of insurance providers. However, estimating medical costs is challenging due to the influence of multiple nonlinear factors such as age, BMI, lifestyle habits, regional variations, and smoking behavior. Classical actuarial methods—including generalized linear models—often fail to model these nonlinear relationships effectively.

Machine learning (ML) techniques provide a powerful alternative by automatically identifying patterns in historical data and learning complex relationships without manually engineered assumptions. This paper develops and evaluates an AI-driven insurance cost prediction system built using Python and tested on a real-world dataset. The objective is to compare multiple ML algorithms and identify the most reliable model for high-accuracy cost estimation.

2. LITERATURE REVIEW

Previous studies have explored ML techniques for insurance analytics. Linear Regression has been widely used as a baseline model due to its simplicity but struggles with nonlinear patterns. Polynomial Regression improves curve fitting but is prone to overfitting on multi-dimensional data.

Ensemble learning models, specifically Random Forest and Gradient Boosting, have demonstrated superior performance in cost prediction tasks owing to their robustness and ability to capture complex feature interactions. Recent research has highlighted XGBoost as a state-of-the-art algorithm for regression problems due to its optimized boosting strategy.

Explainable AI methods such as SHAP have been proposed to overcome the "black-box" concern of ML models by attributing feature importance and enhancing transparency. However, limited research integrates SHAP-based explainability into health insurance cost prediction models—a gap this study aims to address.

3. DATASET DESCRIPTION

The Medical Cost Personal Dataset contains 1,338 records with the following attributes:

a. Age
b. BMI
c. Sex
d. Smoking Status
e. Number of Children
f. Region
g. Charges (Target Variable)

Categorical features were encoded using one-hot or label encoding techniques. Numerical values were normalized where required.

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

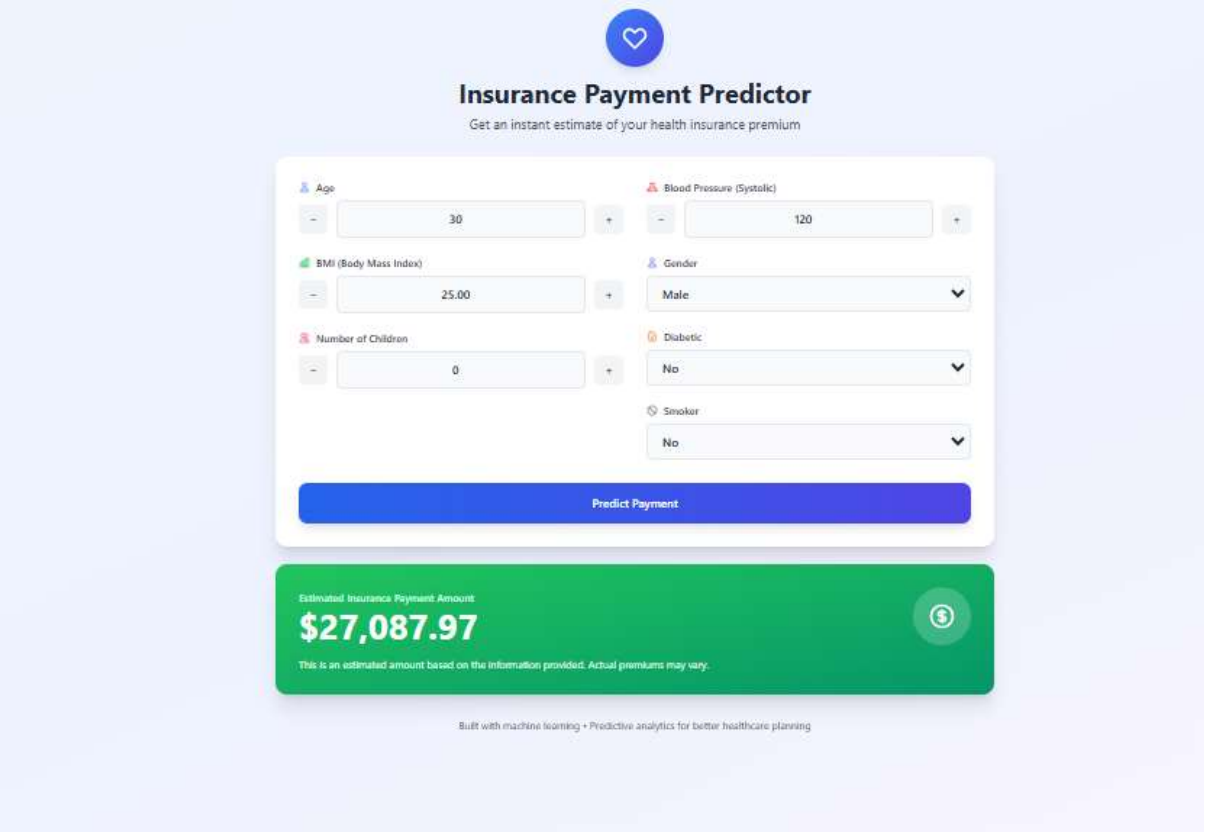
- Linear Regression
- Polynomial Regression (degree = 2)
- XGBoost Regressor
- Random Forest Regressor
- Support Vector Regression (SVR)

C. Evaluation Metrics

Metrics used:

- MAE (Mean Absolute Error)
- RMSE (Root Mean Square Error)
- R² Score

5. Deployment: Web-based Insurance Predictor



Insurance Payment Predictor

Get an instant estimate of your health insurance premium

Age: 30

Blood Pressure (Systolic): 120

BMI (Body Mass Index): 25.00

Gender: Male

Number of Children: 0

Diabetic: No

Smoker: No

Estimated Insurance Payment Amount

$27,087.97

This is an estimated amount based on the information provided. Actual premiums may vary.

Built with machine learning • Predictive analytics for better healthcare planning

6. RESULTS

A. Performance Comparison

Table I: Model Performance Metrics

Model	**R²**	**MAE**	**RMSE**
Linear Regression	0.720	4499.73	6040.41
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] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. 22nd ACM SIGKDD*, 2016.

[2] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, 2017.

[3] Medical Cost Personal Dataset, Kaggle.

[4] J. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, 2001.

[5] L. Breiman, "Random forests," *Machine Learning*, 2001.