International Journal Of Innovative Trends In Multi Disciplinary Scientific Research
ISSN (Online): 3139 – 2687
(www.ijtmsr.co.in) Volume 9, Issue 6, 2026

Medical Insurance Cost Prediction Using Machine learning

Thallam Venkata Keerthi[1], V. Rani[2], Sagi Manogna Lakshmi[3], Vengala Mahesh Babu[4], Rampilla Bhanu Teja[5]

Department of Computer Science and Engineering (AI/ML), NRI Institute of Technology, Agiripalli, Andhra Pradesh
E-mail: rani.v@nriit.edu.in

Abstract
This project focuses on building a practical system that estimates medical insurance premiums using user infor- mation and machine learning techniques. A Random Forest regression model is applied to study how factors such as age, body mass index, smoking behavior, gender, number of dependents, and region influence the final insurance cost. Instead of only generating a single price, the system also compares premiums across different companies, highlights the variables that affect the cost the most, and allows users to explore "what-if" situations by modifying their inputs. The model is trained and tested on an open dataset and its accuracy is measured using Mean Absolute Error, Root Mean Squared Error, and the R^2 score. The results show that the proposed system produces reliable predictions while also helping users understand why their premium changes, making it useful for real-life insurance planning and decision making.

Keywords—Medical Insurance Cost Prediction, Machine Learning, Random Forest Regressor, Feature Importance, What- If Analysis, Healthcare Analytics

INTRODUCTION

Rising healthcare expenditures have become a serious financial challenge for people and governing bodies across different countries [1], [2]. As healthcare technology develops and average life expectancy rises, the associated financial demands continue to expand. In this environment, health insurance plays a key role by supporting access to medical services while minimizing the impact of unpredictable expenses on individuals [3], [4]. Reliable prediction of insurance costs is now a key requirement, helping providers design balanced premium rates and enabling consumers to better organize their financial commitments [5]. In particular, these models also provide valuable support to governments and policymakers by revealing trends that assist in planning and allocating healthcare resources [6]. At the same time, Machine Learning (ML) has gained importance as a tool for analyzing large-scale healthcare data and predicting medical charges [7]. Ensemble learning approaches, for example, Random Forest, Gradient Boosting Regressor (GBR), and XGBoost, have shown better

results than standard linear techniques [8], [9]. These models are capable of learning nonlinear trends and complex relationships between variables that are common in healthcare data. Existing studies also highlight the effectiveness of machine learning in predicting insurance costs by using factors such as age, BMI, smoking habits, and number of dependents [10].In this work, a Random Forest–based prediction system is developed to estimate medical insurance costs while also helping users compare multiple insurance companies to choose the most affordable or suitable plan. Rather than providing only a predicted value, the framework focuses on transparency and practical decision support. The main contributions of this work include:

- Construction of an ensemble Random Forest model for insurance cost estimation.
- Evaluation against standard regression techniques for performance validation
- Analysis of feature importance to explain model behavior.
- What-if analysis for personalized evaluation.
- Company-wise premium comparison for informed insurer choice.

II. PROBLEM FORMULATION

Medical insurance cost prediction can be formulated as a supervised regression problem, where the objective is to learn a mapping between a set of input attributes and a continuous-valued target variable representing insurance charges. Let $X = \{x1, x2, \ldots, xn\}$ denote the input feature vector for an individual policyholder, where features include age, BMI, smoking status, number of dependents, gender, and region. The target variable y represents the corresponding medical insurance cost.

The goal of the predictive model is to estimate a function $f(\cdot)$ such that:

$$\hat{y} = f(X) \qquad (1)$$

where $\hat{y}$ is the predicted insurance premium. The function $f(\cdot)$ should minimize the prediction error while maintaining generalization to unseen data. Given the nonlinear nature of healthcare cost determinants, ensemble-based regression techniques are well-suited for approximating this function.

III. RELATED WORK

Several studies have investigated the application of machine learning techniques for predicting healthcare and medical insurance costs. Taloba et al. developed a regression-based machine learning model for estimating hospitalization costs using features such as patient age and length of hospital stay, reporting a high prediction accuracy of 97.91% using linear regression techniques [11]. Panay et al. proposed a Dempster–Shafer theory–based regression framework for healthcare cost

International Journal Of Innovative Trends In Multi Disciplinary Scientific Research
ISSN (Online): 3139 – 2687
(www.ijtmsr.co.in) Volume 9, Issue 6, 2026

estimation in Japanese hospitals, emphasizing interpretability, although the reported performance was moderate with an R2 value of 0.44 [12].

Emec developed an ensemble-based technique called MedCost-AdaBoost, which achieved better results than bag- ging and stacking methods for predicting medical insurance expenses on a publicly available dataset [13]. Sabarinath and Mathew compared several regression approaches—Ridge Re- gression, Support Vector Regression (SVR), and Lasso—and identified SVR as the most accurate, reaching 82.6% accuracy [14]. Hanafy and Mahmoud evaluated eight machine learning algorithms, including Random Forest, XGBoost, and Deep Neural Networks, and demonstrated that Gradient Boosting achieved the highest performance with an R2 score of 85.83% [15].

Shakhovska et al. explored ensemble regression models combined with feature selection techniques such as Boruta to improve prediction accuracy in noisy medical insurance datasets [16]. Orji and Ukwandu focused on enhancing model explainability using SHAP and Individual Conditional Expec- tation (ICE) techniques, reporting that XGBoost achieved the highest predictive accuracy while Random Forest offered faster computation and improved interpretability [17].

Pfutzenreuter and de Lima compared regularized linear regression with tree-based ensemble methods. While linear regression achieved an R2 score of 0.73, Random Forest and XGBoost models demonstrated superior performance after hyperparameter tuning [18]. Farouka et al. evaluated six machine learning models and found that Decision Trees achieved the lowest Mean Absolute Percentage Error (MAPE), highlighting their suitability for lightweight and resource-constrained deployment [19]. Zhang et al. employed Gradient Boosting Regression to capture nonlinear feature interactions and concluded that ensemble-based models provided the best overall predictive performance [20].

Overall, these studies demonstrate that ensemble learning methods such as Random Forest, XGBoost, and Gradient Boosting Regressor consistently outperform traditional regres- sion models in terms of accuracy and robustness. However, challenges including limited interpretability, susceptibility to overfitting, and dependence on relatively small datasets re- main. These limitations motivate the development of ma- chine learning frameworks that combine strong predictive performance with transparency, scalability, and user-oriented decision-support capabilities.

BLOCK DIAGRAM

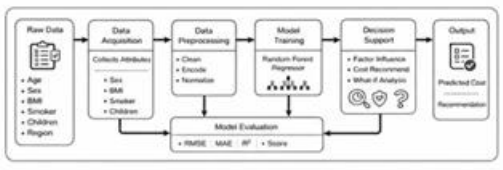


Fig. 1. Block diagram of the proposed medical insurance cost prediction framework.

The overall architecture of the proposed system is illustrated in Fig. 1. The framework follows a modular pipeline consisting of data acquisition, data preprocessing, model training, model evaluation, and decision support. Each module performs a distinct function, enabling a structured and efficient trans- formation of raw input data into meaningful predictions and recommendations.

Initially, the data acquisition module collects relevant user attributes such as age, sex, BMI, smoking status, number of children, and region. These inputs are processed in the preprocessing stage, where data cleaning, encoding of cate- gorical variables, and normalization of numerical features are performed to ensure compatibility with the learning algorithm. The processed data is then supplied to the model training module, which employs a Random Forest Regressor to learn relationships between input features and insurance charges.

Subsequently, the model evaluation module assesses predic- tive performance using standard regression metrics to verify accuracy and generalization. Finally, the decision support module interprets the predicted insurance cost by identifying influential factors, enabling what-if analysis, and recommend- ing the most cost-effective insurance option. This modular architecture enhances scalability, simplifies maintenance, and supports seamless integration with real-world insurance ana- lytics platforms.

V. METHODOLOGY

A. Data Preprocessing

Before building the model, the dataset was analyzed and processed to reduce noise and ensure consistency. The data preparation involved several steps:

- To lessen the impact of extreme values on model learning, unusually high and low observations were managed by applying winsorization at the 1st and 99th percentile levels.
- When there were missing entries, numerical attributes were filled with their mean values. Categorical attributes were filled with the most common category in the dataset.
- Since learning algorithms cannot work directly with cat- egorical variables, these features were turned into binary indicators using one-hot encoding. This process avoided creating unintended ordinal relationships.
- After completing all preprocessing tasks, the dataset was divided into two parts: 80% was used for training the model, and the remaining 20% was set aside for evaluating predictive performance.

B. Random Forest Regressor

Random Forest is an ensemble learning algorithm that

constructs multiple decision trees using bootstrap sampling and random feature selection. For regression tasks, the final prediction is obtained by averaging the predictions of individ- ual trees:

C. Mathematical Interpretation of Random Forest

Random Forest regression operates by constructing an en- semble of T decision trees, each trained on a bootstrap sample of the training dataset. For a given input vector x, each decision tree produces an independent prediction ft(x). The final prediction is computed as the arithmetic mean of individual tree outputs. This averaging process reduces variance and improves prediction stability. Random feature selection at each split ensures decorrelation among trees, enhancing generalization performance. The ensemble nature of Random Forest enables the model to approximate complex nonlinear functions without explicit functional assumptions.

D. Hyperparameter Configuration

The Random Forest model is configured with the following hyperparameters:

- n estimators = 200
- max depth = 10
- min samples split = 5
- min samples leaf = 2
- random state = 42
- n jobs = -1

These settings balance model complexity and generalization performance.

E. Design Rationale and Modeling Strategy

This was developed to address several issues previously brought forth by traditional pricing techniques for insurance coverage. There are many issues involved in medical insurance cost estimations. There are various factors such as age factors, lifestyle factors, and health-related factors. Most of these factors tend to interact with each other too. Due to this reason, traditional techniques such as Linear Regression or Ridge Regression are not necessarily capable enough to address them properly.

For this issue, various machine learning algorithms have been applied in the current study, such as Gradient Boost- ing, XGBoost, and Random Forest. These models have been applied and evaluated for the task with the same set of data. Based on the analysis, the Random Forest Regressor algorithm has been chosen for this task due to its properties of having stable predictions and being able to deal with the correlations within the features. In turn, this approach decreases the vari- ability in predictions and involves growing a forest of decision trees on bootstrap samples. Also, by determining the factors that affect the output and testing the model for different input situations, it will be easier to interpret the system and build user confidence.

VI. ALGORITHM

Algorithm 1 Medical Insurance Cost Prediction Using Ran- dom Forest:
Step-1: Import all the required libraries such as NumPy, Pandas, Flask
2: Step-2: Load the medical insurance dataset.
3: Step-3: Preprocess the dataset which includes handling missing values, encoding categorical features, and scaling numerical attributes.
4: Step-4: Split the dataset into two parts: 80:20
5: Step-5: Initialize the Random Forest Regressor with tuned hyperparameters.
6: Step-6: Train the model using the training dataset.
7: Step-7: Evaluate the model performance using MAE, RMSE, and R^2 metrics on the testing dataset.
8: Step-8: Perform feature importance analysis to identify the most influential factors affecting insurance cost.
9: Step-9: Generate the predicted insurance premium for a given user profile.
10: Step-10: Provide decision-support recommendations based on the predicted cost and influencing factors.

VII. EXPERIMENTAL SETUP

All experiments were conducted using Python 3.13 and a system equipped with an Intel i5 processor and 8 GB of RAM. The proposed Random Forest–based model was evaluated in comparison with several baseline regression models, including Linear Regression, Ridge Regression, Gradient Boosting Regressor, and XGBoost Regressor, to ensure a comprehensive and fair performance assessment.

A. Training and Testing Pipeline

The dataset was partitioned into training and testing subsets using an 80:20 split ratio to enable unbiased evaluation. The training set was utilized for learning model parameters, while the testing set was reserved exclusively for performance assessment on unseen data. Random Forest inherently employs bootstrap sampling, which provides an internal mechanism for validation through the aggregation of multiple decision trees. Predictions generated on the test set were compared with actual insurance charges to compute evaluation metrics. This pipeline ensures reproducibility, fairness, and robustness in model assessment.

B. Overfitting Prevention and Generalization Control

Overfitting was mitigated through both algorithmic design and controlled hyperparameter selection. The Random Forest algorithm inherently reduces variance through bootstrap aggre- gation of multiple trees. In addition, constraints on tree depth and minimum samples per node were applied to prevent ex- cessive model complexity. These measures limit memorization of training data and promote stable and consistent predictive performance on unseen samples.

International Journal Of Innovative Trends In Multi Disciplinary Scientific Research
ISSN (Online): 3139 – 2687
(www.ijtmsr.co.in) Volume 9, Issue 6, 2026

C. Model Evaluation Metrics

Model evaluation is a critical component of the experimental framework, as it establishes objective criteria for comparing predictive performance across different models. In this study, three widely adopted regression metrics were employed: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R2).

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \qquad (2)$$

Root Mean Squared Error (RMSE):

RMSE measures the square root of the mean of the squared differences between actual and predicted values. It penalizes larger errors more heavily, thereby emphasizing significant deviations. A lower RMSE indicates superior predictive accuracy. It is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \qquad (4)$$

where $\bar{y}$ denotes the mean of the actual values.

These evaluation metrics collectively provide a comprehensive assessment of prediction accuracy, error magnitude, and explanatory power, thereby enabling reliable comparison among the regression models employed in this study.

Coefficient of Determination (R2):

The R2 score rep-resents the proportion of variance in the dependent variable that is explained by the model, indicating how well predicted values approximate the actual data. An R2 value closer to 1 signifies stronger explanatory capability. It is expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (5)$$

where $\bar{y}$ denotes the mean of the actual values.

These evaluation metrics collectively provide a comprehensive assessment of prediction accuracy, error magnitude, and explanatory power, thereby enabling reliable comparison among the regression models employed in this study.

VIII. RESULTS AND DISCUSSION

A. Predicted vs Actual Analysis

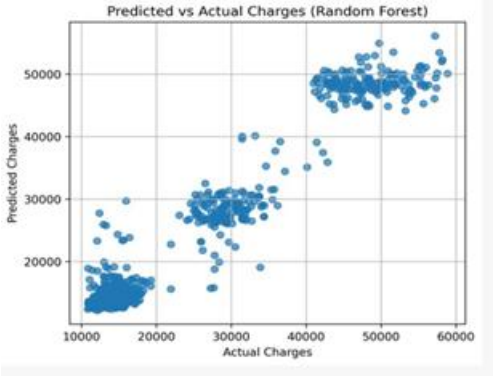


Fig. 2. Predicted versus actual medical insurance charges using the Random Forest Regressor.

The **predicted versus actual** plot is used to visually evaluate the predictive performance of the Random Forest model by comparing the estimated insurance charges with the corresponding observed values. As illustrated in **Fig. 2**, the majority of the data points are closely aligned with the diagonal reference line, indicating a strong agreement between the predicted and actual insurance charges.

This close correspondence demonstrates that the proposed Random Forest model effectively captures the complex relationships between the input features and insurance charges across a wide range of values. Although minor deviations from the diagonal line are observed, they represent normal prediction errors rather than systematic inaccuracies. Overall, the tight clustering of data points around the reference line confirms the high accuracy, reliability, and generalization capability of the proposed model for medical insurance cost prediction.

B. Residual Analysis

Residual analysis was performed to evaluate the error behavior of the Random Forest model by examining the differences between the actual and predicted insurance charges. As illustrated in **Fig. 3**, the residuals are symmetrically distributed around zero without any noticeable systematic pattern, indicating that the model does not exhibit significant prediction bias. Furthermore, the dispersion of residuals remains relatively consistent across different charge ranges, suggesting that the prediction errors are stable for both low- and high-premium values. Moreover, the absence of structured trends or heteroscedasticity in the residual plot confirms that the Random Forest model effectively captures the underlying nonlinear relationships within the dataset.

These observations validate the robustness and reliability of the proposed framework, demonstrating that the prediction errors are predominantly random rather than resulting from model misspecification. Overall, the residual analysis indicates that

International Journal Of Innovative Trends In Multi Disciplinary Scientific Research
ISSN (Online): 3139 – 2687
(www.ijtmsr.co.in) Volume 9, Issue 6, 2026

the model generalizes well and provides accurate predictions across the entire range of insurance charges.

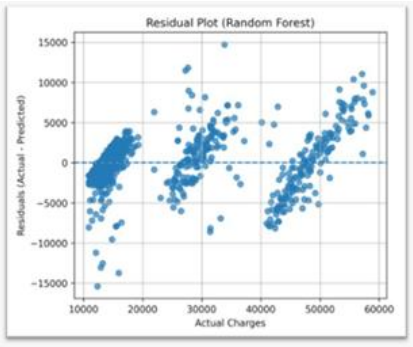


Fig. 3. Residual distribution of Random Forest predictions

C. Comparative Performance

A comparative evaluation of multiple regression models was conducted to assess their effectiveness in predicting medical insurance costs. The performance of each model was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R2). These metrics collectively provide insights into prediction accuracy, error magnitude, and explanatory power.

TABLE I. PERFORMANCE COMPARISON OF REGRESSION MODELS

Model	MAE	RMSE	R^2 Score
Random Forest Regressor	1152.34	2058.36	0.9756
XGBoost Regressor	1305.86	2234.34	0.9713
Gradient Boosting Regressor	1743.36	2918.89	0.9510
Linear Regression	4188.18	5937.97	0.7972
Ridge Regression	4191.16	5938.23	0.7972

D. Feature Importance Analysis

Feature importance analysis is used to quantify the contribution of each input variable toward the model's prediction of medical insurance costs. The Random Forest Regressor provides an inherent mechanism for estimating feature importance based on the reduction in prediction error achieved by each feature across decision trees. This enables interpretability by identifying which attributes most strongly influence the estimated premium.

The feature importance analysis reveals that smoking status is the most dominant factor influencing medical insurance premiums, accounting for the majority of the model's predictive power. The number of children and regional location exhibit a moderate effect on the predicted cost, while age and sex contribute comparatively less. These findings indicate that lifestyle-related factors have a stronger impact on insurance pricing than basic demographic attributes. The interpretability of the Random Forest model enables transparent understanding of how individual characteristics affect premium estimation, thereby supporting personalized recommendations and informed decision making within the proposed framework.

Table II. Feature importance of key factors

Feature	Importance Score
Smoking Status	0.7468
Children	0.0353
Age	0.0084
Region	0.0068
Sex	0.0044

E. What-If Analysis

The proposed framework enables scenario-based evaluation by allowing users to modify input attributes such as smoking status or BMI and observe corresponding changes in predicted premiums. This improves transparency and supports personalized decision making.

F. Recommendation Module

Based on the predicted premium and the identified influ- ential features, the proposed system generates personalized recommendations for users. By leveraging feature importance scores and what-if analysis, the framework suggests actionable steps that can potentially reduce insurance costs. For instance, if smoking status is identified as a dominant contributor to higher premiums, the system advises smoking cessation as a long-term cost-reduction strategy. Similarly, elevated BMI values trigger recommendations related to maintaining a healthier lifestyle, such as regular physical activity and dietary adjustments. In addition, the system provides comparative recommenda- tions across insurance providers by highlighting the company offering the lowest predicted premium for a given user profile. This enables users to make informed choices by selecting the most cost-effective plan while understanding the factors that drive pricing variations. Through these targeted suggestions, the proposed model extends beyond prediction to function as a decision-support tool, promoting transparency, user awareness, and financially responsible healthcare planning.

X. CONCLUSION

This paper presented a Random Forest–based framework for medical insurance cost prediction that integrates accurate forecasting, comparative model evaluation, interpretability, and

user-oriented decision support. The proposed approach not only demonstrates strong predictive capability across multiple machine learning models but also provides transparency through feature importance analysis and scenario-based what-if evaluation, enabling users to understand how different attributes influence insurance premiums. Beyond prediction, the framework extends into practical applicability by offering personalized recommendations and comparative insights across insurance providers, allowing users to identify the most cost-effective option for their specific profile. By combining prediction, explanation, and actionable guidance within a unified system, the proposed model func- tions as a comprehensive decision-support tool for healthcare insurance analytics. This makes it well-suited for real-world adoption, where both accuracy and interpretability are essential for informed financial and healthcare planning.

REFERENCES

1. A. I. Taloba, M. I. ElAdl, M. A. Abd ElAal, and M. A. El-Hoseny, "Estimation and Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning," Journal of Healthcare Engineering, vol. 2022, Article ID 7969220, 2022.
2. B. Panay, M. Petit-Renaud, and T. Denœux, "Predicting Health Care Costs Using Evidence Regression," in UCAmI 2019: Ubiquitous Computing and Ambient Intelligence, Springer, pp. 31–42, 2019.
3. M. Emec, "Medical Insurance Cost Prediction MedCost: Machine Learning Ensemble Approaches," European Journal of Technique (EJT), vol. 14, no. 1, pp. 51–60, 2024.
4. U. S. Sabarinath and A. Mathew, "Medical Insurance Cost Prediction," Indian Journal of Data Communication and Networking, vol. 4, no. 4, pp. 45–50, 2024.
5. M. Hanafy and O. M. A. Mahmoud, "Predict Health Insurance Cost by Using Machine Learning and DNN Regression Models," International Journal of Innovative Technology and Exploring Engineering, vol. 10, no. 3, pp. 137–143, 2021.
6. N. Shakhovska, I. Pasichnyk, M. Komar, and V. Stryzhak, "An Ensemble Methods for Medical Insurance Costs Prediction Task," Computers, Materials & Continua, vol. 71, no. 1, pp. 1575–1590, 2022.
7. U. Orji and E. Ukwandu, "Machine Learning for an Explainable Cost Prediction of Medical Insurance," Machine Learning with Applications, vol. 15, Article 100516, 2023.
8. T. C. Pfutzenreuter and E. P. de Lima, "Machine Learning in Healthcare Management for Medical Insurance Cost Prediction," in Proceedings of the XLI Encontro Nacional de Engenharia de Produc¸a˜o, 2022.
9. M. Farouka, R. Zahran, A. El-Demerdash, and D. Abdou, "An Efficient Framework for Predicting Medical Insurance Costs Using Machine Learning," Journal of Computing and Communication, vol. 3, no. 2, pp. 55–64, 2024.
10. B. Zhang, H. Huang, and D. Wang, "Medical Insurance Cost Prediction Using Gradient Boosting Regression: A Machine Learning Approach," Informatica, vol. 49, no. 1, pp. 9–22, 2025.
11. A. S. Baazeem and A. M. Khedr, "Machine Learning Techniques for Health Insurance Cost Prediction: A Comparative Study," International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, pp. 667–673, 2021.
12. J. M. Woolhandler and D. U. Himmelstein, "The Rising Costs of Healthcare and the Role of Insurance," New England Journal of Medicine, vol. 384, no. 8, pp. 740–749, 2021.
13. A. F. Mafi, T. F. Moller, and G. J. Linden, "Economic Burden of Chronic Illness and Rising Insurance Costs," Healthcare Finance Journal, vol. 9, no. 2, pp. 115–122, 2020.
14. D. Mishra and N. Rao, "Insurance in the Era of Big Data: Machine Learning-Based Underwriting," Journal of Financial Services Technology, vol. 6, no. 1, pp. 33–42, 2022.
15. R. Gupta, "Predictive Analytics for Health Insurance Premium Estimation Using Decision Trees," Data Science in Healthcare, vol. 3, no. 4, pp. 89–96, 2021.
16. Y. Liu and S. Zhang, "Feature Engineering and Supervised Learning in Health Cost Prediction," in Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, pp. 232–239, 2020.
17. M. Kumar and P. S. Reddy, "Multi-Factor Health Premium Estimation Using SVM and Logistic Regression," Journal of Applied Computing and Informatics, vol. 17, no. 1, pp. 1–10, 2021.
18. T. Y. Wang and R. Ali, "An ANN-Based Model for Insurance Pricing," International Journal of Artificial Intelligence and Applications, vol. 10, no. 3, pp. 57–65, 2020.
19. [R. Rao and M. Singh, "Medical Insurance Charge Forecasting Using Ensemble Models," Machine Learning and Healthcare Analytics, vol. 4, no. 2, pp. 78–84, 2021.
20. K. Patil, S. Roy, and M. Sharma, "Comparative Analysis of Random Forest and XGBoost for Insurance Cost Prediction," in International Conference on Intelligent Systems Design and Applications, pp. 122–130, 2022.