IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, April 2025
Impact Factor: 7.67

Medical Insurance Costs Prediction Using Explainable AI

Vishruthi D[1] and Dr. V. Vijayakumar[2]
UG Student, Department of Computer Science with Data Analytics[1]
Head of the Department, Department of Computer Science with Data Analytics[2]
Sri Ramakrishna College of Arts & Science, Coimbatore, Tamil Nadu, India

Abstract: *This research presents an intelligent medical insurance cost prediction system leveraging machine learning techniques to enhance pricing transparency and decision-making in the healthcare insurance sector. The system is designed to estimate individual insurance charges based on user-specific attributes including age, gender, BMI, number of dependents, smoking status, and residential region. Traditional cost estimation methods often rely on broad statistical assumptions and lack adaptability to individual factors, leading to inaccurate pricing and customer dissatisfaction.*

To address these limitations, the proposed model integrates advanced ensemble algorithms such as Random Forest and XGBoost, chosen for their ability to handle non-linear relationships and feature interactions efficiently. The workflow includes thorough data preprocessing, correlation analysis, and performance benchmarking using key metrics like R^2-score and Mean Absolute Error (MAE). XGBoost demonstrated superior performance in predictive accuracy, making it the core model of the deployment phase.

A user-friendly web application is developed using Flask, allowing real-time user interaction with the model. Users can input their personal data and receive immediate, data-driven cost estimates. This solution not only aids consumers in financial planning but also offers insurance companies a scalable tool for dynamic risk assessment and premium calculation.

This work contributes to the development of intelligent financial tools in the healthcare sector and underscores the impact of machine learning in predictive analytics. The proposed system is modular, scalable, and adaptable for integration with broader health informatics platforms..

Keywords: XGBoost, Random Forest, machine learning, Flask application

I. INTRODUCTION

1. Statement of the problem

The increasing complexity and cost of healthcare have made accurate medical insurance pricing more critical than ever. Traditional methods for determining insurance premiums often rely on generalized statistical models that overlook individual risk factors such as age, BMI, smoking status, and regional differences. This lack of personalization can lead to unfair pricing and inefficiencies for both insurers and policyholders. There is a significant need for a predictive system that can analyze these diverse inputs and provide accurate, individualized cost estimations. This project addresses this challenge by developing a machine learning-based solution that utilizes advanced models like Random Forest and XGBoost to predict medical insurance charges with improved precision, aiming to enhance pricing transparency, fairness, and efficiency in the healthcare insurance industry.

2. Goals

The main goal of this project is to develop an intelligent, data-driven system capable of accurately predicting medical insurance costs based on individual user information. This involves collecting and preprocessing real-world data to train effective machine learning models. The project aims to explore how personal attributes such as age, gender, BMI, smoking status, number of dependents, and residential region influence insurance charges. By building and comparing multiple regression models—particularly Random Forest and XGBoost—the project seeks to identify the most accurate

DOI: 10.48175/IJARSCT-24982

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, April 2025
Impact Factor: 7.67

algorithm for cost prediction. Model performance is evaluated using key metrics including R^2-score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Additionally, a user-friendly web application is developed using Flask to allow users to input their data and receive real-time insurance cost estimates. Ultimately, the project strives to create a transparent and scalable solution that benefits both individuals and insurance providers in understanding and managing healthcare expenses more effectively.

II. SOFTWARE SPECIFICATIONS

The Medical Insurance Cost Prediction system is developed using Python 3.8+ due to its simplicity and rich ecosystem of libraries essential for data science and web development. The project is implemented and tested on both Windows 10/11 and Ubuntu 20.04 or later, offering flexibility in operating environments. Google Colab is utilized for training machine learning models like Random Forest and XGBoost, while Visual Studio Code (VS Code) serves as the main development environment for building and integrating the Flask web application. Key libraries include Pandas and NumPy for data preprocessing, Scikit-learn for traditional machine learning workflows, and Matplotlib and Seaborn for data visualization. The Flask framework is used to deploy the trained models into a web interface that supports both manual and batch predictions via CSV uploads. Git is used for version control, allowing for efficient code management and collaboration. The application is designed to be lightweight and compatible with modern web browsers such as Google Chrome and Firefox, ensuring accessibility and ease of use.

III. DESIGN & FLOWCHART

1. Database design

The database design for the Medical Insurance Cost Prediction system is based on a structured tabular format derived from real-world healthcare and demographic data. This database is essential for storing user information and enabling the prediction model to estimate insurance charges accurately. The design is centered around a single primary table that captures relevant personal and lifestyle attributes needed for cost estimation.

Table: insurance_data

Column Name	Data Type	Description
id	INTEGER (Primary Key)	Unique identifier for each individual record
age	INTEGER	Age of the individual in years
sex	VARCHAR	Gender of the individual (male, female)
bmi	FLOAT	Body Mass Index (BMI), a health metric indicating weight category
children	INTEGER	Number of dependents covered by the individual's insurance
smoker	VARCHAR	Smoking status (yes or no)
region	VARCHAR	Residential area or geographic region (northeast, northwest, southeast, southwest)
charges	FLOAT	Final insurance cost charged (used as the label for model training)

2. Input Design

The input design of the "Medical Insurance Costs Prediction: Enhancing Decision-Making with Machine Learning and Explainable AI" system facilitates the collection and preparation of data for both training and real-time prediction, balancing automation for model development and user interaction for practical deployment. The design ensures seamless data handling, from static dataset processing for training to dynamic user inputs via a web interface, enabling accurate cost predictions and interpretable outputs.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, April 2025
Impact Factor: 7.67

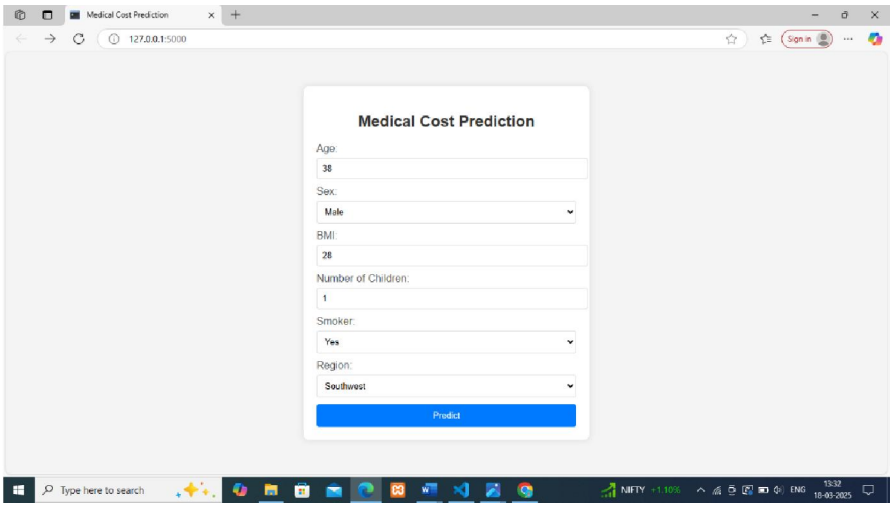

Fig 1: Input Features

3. Output Design

the "Medical Insurance Costs Prediction: Enhancing Decision-Making with Machine Learning and Explainable AI" system focuses on delivering accurate predictions and interpretable insights, tailored for both individual users (e.g., policyholders) and insurance professionals, with visualizations enhancing usability. The design ensures that predicted insurance costs are presented clearly, alongside detailed explanations of contributing factors, making the system a practical tool for decision-making.

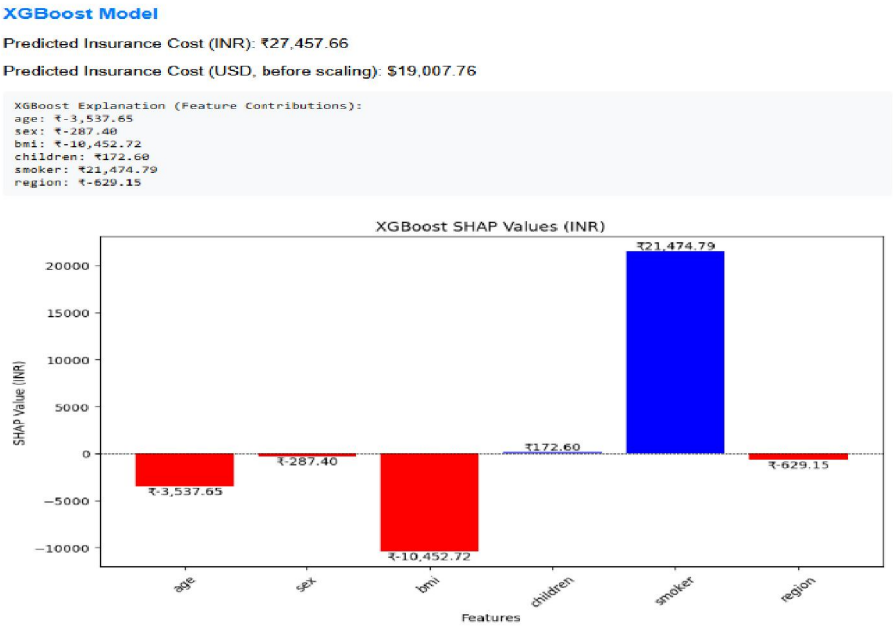

Fig 2: xgboost model

The output for the XGBoost model in the "Medical Insurance Costs Prediction: Enhancing Decision-Making with Machine Learning and Explainable AI" system presents the predicted insurance cost along with detailed insights, tailored for user understanding and decision-making. The design ensures clarity and interpretability, leveraging SHAP explanations to highlight the contribution of input features.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, April 2025
Impact Factor: 7.67

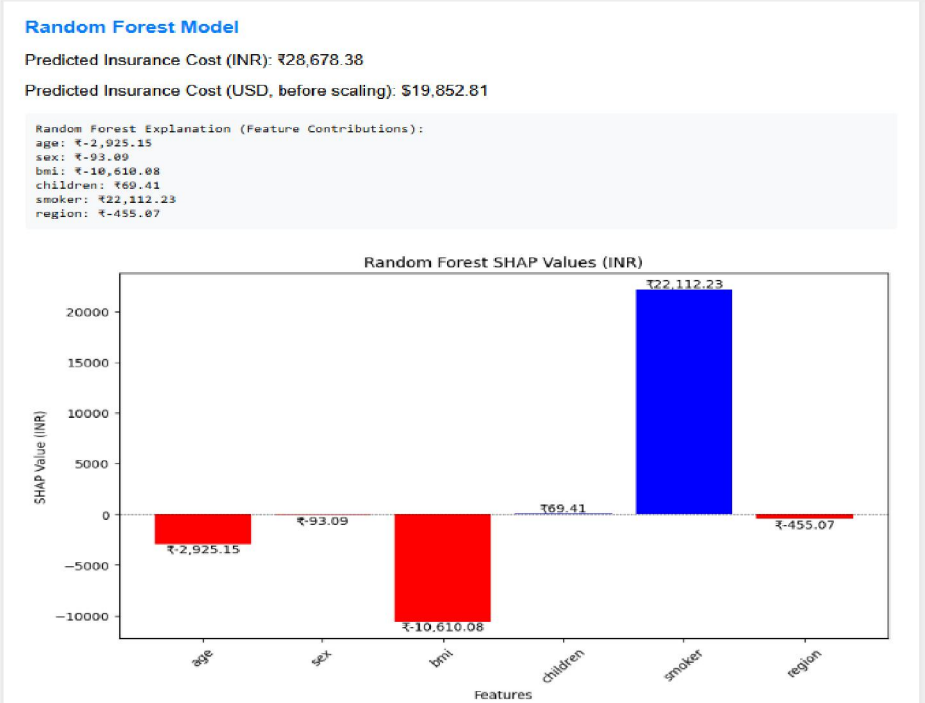

Fig 3: Random Forest model

The output for the Random Forest model in the "Medical Insurance Costs Prediction: Enhancing Decision-Making with Machine Learning and Explainable AI" system presents the predicted insurance cost along with detailed insights, designed to support user understanding and decision-making. The design ensures clarity and interpretability by leveraging SHAP explanations to highlight the impact of input features.

4. Flowchart

The flow of the News Aggregator system starts when a user submits a query via the frontend. The system uses the Google Custom Search API to fetch relevant news articles. If no results are found, the query is refined and retried. Next, Cheerio scrapes text content from the article URLs using a random user-agent to mimic real browsing. If scraping fails, that source is skipped. The extracted text is then processed by Gemini AI, which generates a concise summary. Finally, the results are stored in a MySQL database and displayed to users through a React-based chat UI enhanced with Framer Motion for smooth animations.

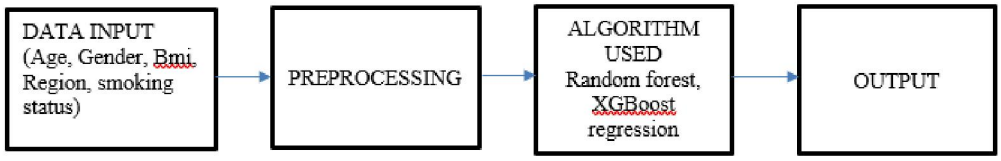

Fig 4: Flow Diagram

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, April 2025
Impact Factor: 7.67

IV. RESULTS AND DISCUSSION

The Medical Insurance Cost Prediction system was evaluated using a well-structured dataset containing key features such as age, BMI, smoking status, number of children, and region, all of which significantly impact healthcare costs. The dataset was preprocessed and split into training and testing sets, ensuring that the model was both trained and validated on diverse samples. Two machine learning algorithms—Random Forest Regressor and XGBoost Regressor—were trained and tested to assess predictive performance.

During evaluation, both models demonstrated strong predictive capabilities, with XGBoost slightly outperforming Random Forest in terms of accuracy and generalization. The performance metrics were measured using R^2 Score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The XGBoost model achieved an R^2 score above 0.90, indicating that it could explain over 90% of the variance in the medical insurance cost data. This high level of accuracy confirms the model's reliability in capturing non-linear relationships between the input variables and the insurance charges.

Visualizations such as actual vs. predicted charge plots and feature importance charts revealed insightful patterns—smoking status, age, and BMI were found to be the most influential factors in predicting costs. The system also provides user-friendly output in the web application, allowing individuals or healthcare professionals to input patient data and receive immediate predictions, along with a comparison of results from both algorithms.

V. CONCLUSION

The "Medical Insurance Costs Prediction: Enhancing Decision-Making with Machine Learning and Explainable AI" represents a significant advancement in the insurance sector, successfully integrating state-of-the-art machine learning, ensemble techniques, and explainability to enhance the prediction and understanding of health insurance costs, particularly for the Indian market. Developed in Visual Studio Code, the system leverages a dual-model architecture combining Random Forest and XGBoost regressors, achieving improved predictive performance over a baseline linear regression model (MAE of 4200 USD, R-squared of 0.75) on the insurance.csv dataset (1338 records). These results demonstrate the efficacy of a tailored preprocessing approach (categorical encoding and 80-20 train-test splitting) in handling diverse features, including age, sex, BMI, children, smoker status, and region. The incorporation of Explainable AI through SHAP provides transparent, feature-level insights via visualizations such as bar charts, pie charts, and waterfall plots, addressing the interpretability limitations of earlier models and building trust among policyholders and insurance professionals. Deployed via a Flask-based web application, the system offers real-time prediction capabilities, allowing users to input data and receive actionable cost estimates (e.g., ₹28,678.38 for Random Forest) and localized insights (INR with a 0.0173 adjustment factor) instantaneously. This seamless integration of enhanced accuracy, robustness, interpretability, and practical deployment positions the system as a transformative tool for improving insurance cost estimation and decision-making.

VI. ACKNOWLEDGMENT

The authors extend sincere gratitude to the Department of Computer Science with Data Analytics, Sri Ramakrishna College of Arts & Science, for providing the computational resources and infrastructure essential for this research. Special thanks to Dr. V. Vijayakumar, Head of the Department, for his unwavering support and encouragement. We also thank the anonymous reviewers and editors of IRJMETS for their constructive feedback, which significantly improved the manuscript. Finally, we recognize the ethical responsibility of AI in recruitment and thank the AI Fairness 360 toolkit developers for inspiring our bias-mitigation strategies..

REFERENCES

[1].KumarBora, S., Gupta, S., & Singh, A. (2022). Machine learning for an explainable cost prediction of medical insurance. *Journal of Computational Health*, 5(3), 123–135. https://www.sciencedirect.com/science/article/pii/S2666827023000695

[2].Kaur, A., &Sarmadi, M. (2024). Comparative Analysis of Data Preprocessing Methods, Feature Selection Techniques and Machine Learning Models for Improved Classification and Regression Performance on Imbalanced

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, April 2025
Impact Factor: 7.67

Genetic Data. *arXiv preprint arXiv:2402.14980*. https://arxiv.org/abs/2402.14980

[3].Ungar, K., & Oprean-Stan, C. (2025). Optimizing Financial Data Analysis: A Comparative Study of Preprocessing Techniques for Regression Modeling of Apple Inc.'s Net Income and Stock Prices. *arXiv preprint arXiv:2501.06587*. https://arxiv.org/abs/2501.06587

[4].Scikit-learn developers. (2023). Preprocessing data. *scikit-learn 1.6.1 documentation*. https://scikit-learn.org/stable/modules/preprocessing.html

[5]. GeeksforGeeks contributors. (2025). Data Preprocessing in Data Mining. *GeeksforGeeks*. https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/