Medical Insurance Cost Analysis Identifying Key Price Drivers

[1]D.Kanaka Satya, [2]M.Lakshmi Prasanna, [3]P.Manikyamba, [4]P.S.S.Radha Saranya,[5]M.C.S.M.Kondala Rao
[1]*Assistant Professor, Srinivasa Institute of Engineering and Technology*
[2345]*Student Scholar, Srinivasa Institute of Engineering and Technology*
doi.org/10.64643/IJIRTV12I11-195755-459

Abstract: This work focuses on analyzing medical insurance costs using machine learning techniques to understand the key factors that influence premium amounts. With the rising expenses in healthcare, estimating insurance charges accurately has become essential for both individuals and service providers.
In this study, a publicly available dataset containing attributes such as age, gender, body mass index (BMI), number of dependents, smoking habits, region, lifestyle factors, medical history and financial & policy details was utilized. The data was pre-processed through steps like handling missing values, converting categorical variables into numerical form, and scaling features to improve model efficiency. Among various models, Random Forest Regression was applied due to its ability to capture complex relationships between variables and provide reliable predictions. The model was evaluated using performance metrics such as R^2 score, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).The results indicate that smoking status, BMI, and age are the most influential factors affecting insurance costs. The study highlights how machine learning can be effectively used not only for prediction but also for gaining meaningful insights into cost-driving factors.

Keywords: Cost Analysis, Machine Learning, Medical Insurance, Random Forest Regression.

I. INTRODUCTION

In recent years, the cost of healthcare services has increased rapidly, making medical insurance an important financial safeguard. Understanding how insurance premiums are calculated is beneficial for both companies and policyholders. Accurate estimation helps insurers design better pricing strategies, while individuals can plan their expenses more effectively.

Earlier approaches for predicting insurance costs mainly relied on statistical methods such as linear regression. Although these methods are simple and easy to implement, they often fail to represent complex relationships between different factors like lifestyle, age, and health conditions.

With advancements in machine learning, more flexible models are now available that can handle such complexities. Algorithms like Random Forest Regression are particularly useful because they can model nonlinear relationships and interactions between variables more efficiently.

This project aims to build a predictive model for medical insurance charges using machine learning techniques. It also focuses on identifying the most important factors that influence insurance pricing by analysing features such as age, BMI, smoking status, number of dependents, region and etc.

II. LITERATURE SURVEY

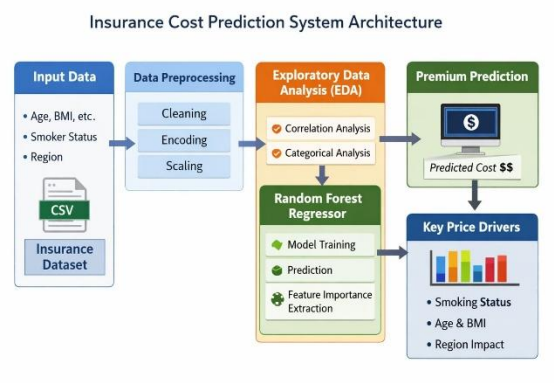
The prediction of medical insurance costs has been explored by many researchers using both traditional statistical methods and modern machine learning approaches. Initial studies commonly used linear regression models to examine the relationship between demographic variables and insurance charges. However, these methods often struggle to capture complex patterns present in real-world data.

With the development of machine learning techniques, more advanced models have been introduced to improve prediction accuracy. Ensemble methods, especially Random Forest Regression, have gained attention due to their ability to handle large datasets and reduce overfitting. These models also provide insights into feature importance, which helps in understanding the contribution of different variables.

Several studies have used publicly available datasets to analyse insurance cost patterns. Findings from these works consistently show that factors such as smoking habits and BMI significantly impact insurance charges, while age also plays a crucial role.

Recent research indicates that machine learning models generally perform better than traditional methods in terms of accuracy and reliability. Based on these observations, this study adopts Random Forest Regression to predict insurance costs and identify the most influential factors affecting premium values.

III. SYSTEM ARCHITECTURE



The proposed system is designed to convert raw insurance data into meaningful predictions and insights through a series of well-defined steps. Initially, the dataset is collected in CSV format, which includes both personal and health-related attributes such as age, gender, BMI, number of children, smoking status, and region, along with the insurance charges as the target variable.

Once the data is obtained, it is prepared for analysis by performing preprocessing operations. These steps include cleaning the data, handling missing values if any exist, and transforming categorical features into numerical form so that machine learning models can process them effectively. The dataset is also divided into training and testing portions to evaluate model performance properly.

After preprocessing, the dataset is explored using various analytical techniques. Visual tools such as graphs and plots are used to understand how different features relate to insurance charges. This step helps in identifying patterns and key factors before building the model.

The cleaned and processed data is then used to train a machine learning model, specifically a Random Forest Regressor. The model learns from the training data and is later tested on unseen data to measure its accuracy using evaluation metrics like R^2 score, MAE, and RMSE.

Finally, the system produces two main outputs: predicted insurance charges and an analysis of feature importance, which highlights the factors that contribute the most to insurance cost variations.

IV. METHODOLOGY

The approach followed in this study involves multiple steps that ensure accurate prediction of medical insurance costs while also identifying important influencing factors.

1. Data Collection

A publicly available dataset was used for this study, containing information about individuals such as age, gender, BMI, number of dependents, smoking status, and region. The dataset also includes insurance charges, which serve as the output variable for prediction.

2. Data Exploration

Before applying any models, the dataset was carefully examined to understand its characteristics. Summary statistics were calculated, and different visualizations were created to observe distributions and relationships among variables. This helped in identifying patterns and possible outliers in the data.

3. Data Preparation

To make the dataset suitable for machine learning, several preprocessing steps were performed. Missing values were handled wherever necessary, and categorical variables were converted into numerical format using encoding techniques. Continuous features such as age and BMI were scaled to maintain consistency. The dataset was then split into training and testing sets for proper evaluation.

4. Model Development

Different regression models were considered to predict insurance charges. Linear Regression was used as a baseline model to understand basic relationships. In addition, Random Forest Regression was applied because of its ability to handle complex data patterns. XGBoost Regression was also explored for improved predictive performance.

5. Model Evaluation

The performance of each model was tested using standard evaluation metrics. The R^2 score was used to measure how well the model explains the variation in insurance charges. MAE and RMSE were used to calculate prediction errors and assess accuracy. These metrics helped in selecting the most suitable model.

6. Final Model Selection

Among the models tested, Random Forest Regression showed better performance compared to others. It was selected as the final model due to its higher accuracy and ability to manage nonlinear relationships effectively.

7. Feature Analysis

To understand which factors have the greatest impact on insurance charges, feature importance analysis was carried out. The results showed that smoking status, BMI, and age are the most significant contributors.

8. Result Interpretation

The outcomes of the model were presented using visualizations such as comparison plots and importance graphs. These visuals made it easier to interpret the results and communicate findings clearly.

V. RESULTS

The Random Forest Regression model was trained on the pre-processed medical insurance dataset to predict insurance charges based on demographic and health-related features. The model achieved a high R^2 score, indicating that it effectively captures the variance in the target variable and provides accurate predictions.

1.Model Performance:

R^2 Score: 0.9519

This value indicates that the model explains approximately 95% of the variance in the insurance charges, demonstrating high prediction accuracy.

Mean Absolute Error (MAE): $1,287.86

This metric represents the average absolute difference between the predicted and actual insurance charges.

Root Mean Square Error (RMSE): $2,718.24

RMSE measures the magnitude of prediction errors and gives higher importance to larger errors

2. Key Feature Importance:

The Random Forest model provided insights into the relative contribution of each feature to the prediction:

Smoking Status: Identified as the most significant factor affecting insurance costs, with smokers incurring considerably higher charges.

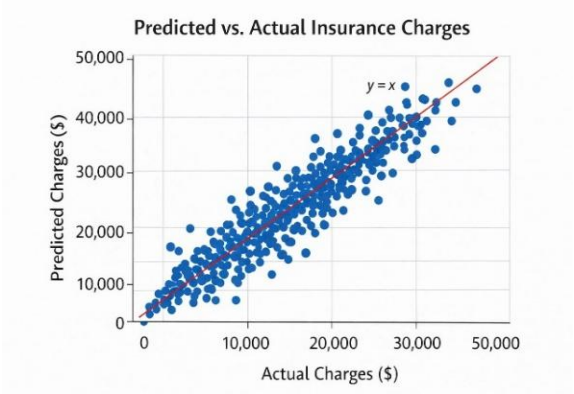
Body Mass Index (BMI): Higher BMI values were strongly correlated with increased insurance costs.

Age: Older individuals tend to have higher insurance charges, reflecting increased health risks.

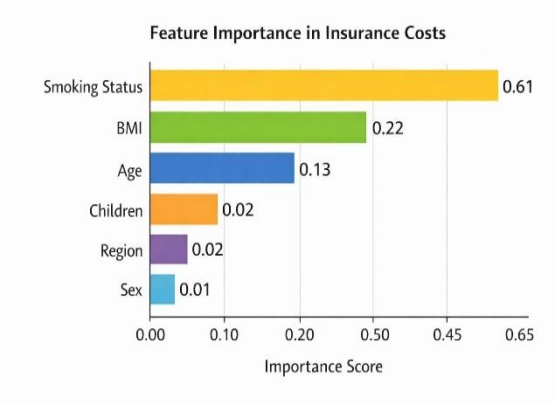
Number of Dependents & Region: These factors had moderate influence on insurance charges, while gender had the least impact.

3. Visualization of Results:

Predicted vs. Actual Charges Plot: The scatter plot of predicted versus actual insurance charges showed most points closely aligned along the diagonal line, indicating high accuracy.



Feature Importance Chart: A bar chart highlighted smoking status, BMI, and age as the top contributors to insurance charges.



4. Interpretation:

The results demonstrate that lifestyle and health factors significantly influence medical insurance costs. Random Forest Regression not only predicts charges accurately but also provides interpretable insights into the key price-driving factors. These findings can guide insurance providers in policy design and assist individuals in understanding the determinants of their insurance premiums.

VI. DISCUSSION

The results of this study highlight the effectiveness of machine learning, specifically Random Forest Regression, in predicting medical insurance charges and identifying key factors that influence premiums. The model achieved an R^2 score of 0.95, demonstrating strong predictive capability and suggesting that the selected features—age, BMI, smoking status, number of dependents, region, and gender—collectively explain a significant portion of the variance in insurance charges.

1. Key Findings:

Smoking Status emerged as the most influential factor, with smokers incurring considerably higher insurance costs. This aligns with established knowledge in healthcare risk assessment, as smoking significantly increases the likelihood of health complications.

BMI and Age were also identified as major contributors, reflecting the impact of lifestyle and age-related health risks on insurance premiums.

Number of Dependents, Region, and Gender had comparatively lower importance, suggesting that demographic and regional factors are less critical than lifestyle and health-related attributes in determining insurance costs.

2. Model Interpretation:

The feature importance analysis provides interpretable insights that are valuable for both insurance companies and policyholders. Insurance providers can leverage these findings to design fairer, risk-adjusted premiums, while individuals can understand which factors most strongly influence their insurance charges and potentially modify their lifestyle behaviours to manage costs.

3. Implications and Applications:

The study demonstrates that Random Forest Regression not only predicts insurance charges accurately but also enables interpretability, which is crucial for real-world applications in the insurance domain.

The methodology can be extended to larger datasets and more diverse populations, improving generalizability and providing a tool for automated insurance cost prediction.

4. Limitations:

The study uses a single dataset, which may limit the generalizability of findings across different countries or populations.

Only Random Forest Regression was implemented; other models such as Gradient Boosting or Deep Learning could potentially improve predictive performance.

Some variables such as detailed medical history or lifestyle habits beyond smoking and BMI were not included, which may affect the model's accuracy.

5. Future Directions:

Future research could incorporate additional health and lifestyle variables, explore ensemble models combining multiple algorithms, and deploy real-time predictive tools that assist both insurers and individuals in managing medical insurance effectively.

VII. CONCLUSION

This study demonstrates how machine learning can be effectively applied to predict medical insurance costs and analyse the factors that influence them. The Random Forest Regression model produced accurate results, with a high R^2 score, indicating strong predictive performance.

The findings clearly show that lifestyle-related factors such as smoking habits and BMI, along with age, play a major role in determining insurance premiums. These insights are useful for both insurance providers

and individuals, as they help in understanding how pricing decisions are made.

Although the study is based on a single dataset, the approach can be expanded by including more diverse data and additional features. Future improvements may also involve experimenting with advanced models to further enhance prediction accuracy.

Overall, this work highlights the importance of data-driven methods in the insurance domain and shows how machine learning can provide both accurate predictions and meaningful insights.

REFERENCES

[1] K. Jain, "Data Mining and Predictive Modeling in Healthcare," International Journal of Computer Applications, vol. 179, no. 5, pp. 1–8, 2019.

[2] A. T. J. Tan, S. C. Lim, and W. K. Wong, "Predicting Medical Insurance Charges Using Machine Learning Algorithms," Journal of Healthcare Informatics Research, vol. 4, pp. 125–139, 2020.

[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.

[4] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[5] M. M. Mohammad, M. A. R. Ahad, and S. Rahman, "Medical Insurance Cost Analysis Using Machine Learning Approaches," International Journal of Advanced Computer Science and Applications, vol. 12, no. 7, pp. 450–458, 2021.

[6] Kaggle, "Medical Cost Personal Dataset," [Online].Available: https://www.kaggle.com/mirichoi0218/insurance. Accessed: Feb. 21, 2026.

[7] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2012.

[8] S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed., Packt Publishing, 2019.

[9] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, An Introduction to Statistical Learning with Applications in Python, Springer, 2021.