Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

Medical Insurance Cost Prediction Using Flask and Machine Learning Algorithms

Harshitha G J[1], Dr Manjunath Kumar BH[2],

[1]Mtech Scholar, SJC Institute of Technology, harshithagj333@gmail.com
[2]HOD & Professor, SJC Institute of Technology, manjunathabh@sjcit.ac.in

Abstract

The rising complexity of healthcare expenses has increased the demand for accurate medical insurance cost prediction systems. This project presents a machine learning-based approach to estimate individual insurance charges by analyzing demographic and health-related features such as age, gender, body mass index (BMI), number of dependents, smoking habits, and residential region. A publicly available dataset was preprocessed by encoding categorical attributes, scaling numerical variables, and visualizing relationships between predictors and charges. Multiple regression and ensemble algorithms, including Linear Regression, Support Vector Regression, Ridge Regression, and Random Forest Regressor, were implemented and evaluated. Comparative analysis revealed that the Random Forest Regressor outperformed other models, achieving an accuracy of approximately 86% in predicting insurance premiums. To enhance usability, the trained model was deployed using a Flask web application that provides real-time predictions based on user inputs. The system not only estimates the expected cost but also categorizes the user's BMI and assigns a risk level according to the predicted premium, thereby improving interpretability for end-users. Furthermore, predictions are logged systematically for monitoring and analysis, ensuring transparency of model behaviour. The integration of machine learning algorithms with a lightweight web framework demonstrates the feasibility of delivering practical and user-friendly solutions in the domain of health insurance analytics. This research highlights the significance of predictive modeling in fostering fair insurance pricing, assisting individual in financial planning, and aiding insurance providers in risk assessment.

Keywords: Medical Insurance, Cost Prediction, Machine Learning, Flask, Regression

1. Introduction

Healthcare has become one of the most critical sectors worldwide, with rising medical expenses posing a challenge for both individuals and insurance providers. Predicting medical insurance costs is essential for ensuring fair premium pricing, effective financial planning, and efficient risk management. Traditional insurance cost estimation often relies on generalized statistical assumptions, which may not accurately reflect the unique health and demographic profiles of individuals.

This project focuses on developing a medical insurance cost prediction model using machine learning algorithms integrated with a Flask-based web application for real-time deployment. The dataset used in this research, obtained from a publicly available source, contains demographic and lifestyle-related

variables such as age, gender, body mass index (BMI), smoking status, number of children, and residential region. These features have a direct influence on healthcare costs, making them suitable predictors for insurance premium estimation. Data preprocessing steps such as encoding categorical variables, handling skewness, and scaling were applied to prepare the dataset for model training. Visualization techniques, including heatmaps and feature distribution plots, were also used to explore dependencies between input variables and insurance charges.

Several machine learning models were implemented and compared, including Linear Regression, Support Vector Regression (SVR), Ridge Regression, and Random Forest Regressor. Among these, the Random Forest model achieved the best performance, reaching an accuracy of approximately 86% in predicting medical charges. Hyperparameter tuning was further employed to optimize the models and improve prediction reliability. The trained Random Forest model was then deployed using Flask, creating an interactive web interface that allows users to input their details and instantly receive estimated insurance costs. In addition to prediction, the system categorizes users based on BMI and assigns a risk level according to the predicted premium, thereby enhancing interpretability and transparency.

The integration of ML with a lightweight web framework like Flask demonstrates the practical feasibility of deploying predictive models in real-world applications. By logging predictions and generating interpretable outputs, the system ensures transparency and trustworthiness, which are vital in sensitive domains such as healthcare and insurance. This project highlights how predictive analytics can aid individuals in understanding potential medical expenses while assisting insurance providers in risk assessment and fair pricing strategies. Overall, the proposed system combines robust machine learning techniques with an accessible deployment platform to deliver a user-friendly, data-driven solution for medical insurance cost prediction.

2. Related Works

Predicting medical insurance costs has been an important area of research within healthcare analytics, data mining, and actuarial science. Accurate estimation of healthcare charges supports fair insurance pricing, financial planning, and risk assessment for providers and individuals alike. Over the years, a range of statistical and machine learning approaches have been applied to predict insurance premiums, gradually evolving from simple linear models to advanced ensemble techniques integrated with deployment frameworks. Early studies in this domain largely depended on Linear Regression models to establish direct relationships between input variables such as age, sex, BMI, smoking habits, and insurance charges. The strength of linear regression lies in its interpretability, which allows insurers to clearly understand how features influence premiums. However, linear regression struggles with multicollinearity and nonlinear patterns that are often present in health-related data, leading to suboptimal predictions. To address these challenges, researchers introduced Ridge Regression, a regularized form of linear regression that penalizes large coefficients, thereby improving generalization. Although Ridge Regression reduces overfitting, its predictive accuracy remains moderate in comparison to more

flexible methods. To capture nonlinear dependencies, Support Vector Regression (SVR) has also been investigated. SVR maps data into higher dimensions to approximate nonlinear relationships within a defined error margin. Several studies have shown that SVR performs better than linear models when the dataset contains complex patterns. However, the computational cost of SVR increases with larger datasets, limiting its scalability in real-world insurance applications.

The most promising results in related works have come from ensemble-based algorithms, particularly the Random Forest Regressor. Random Forest aggregates multiple decision trees, reducing variance and increasing prediction accuracy. Previous comparative analyses consistently report Random Forest as one of the top-performing models in insurance premium forecasting, often achieving accuracy levels above 80%. Hyperparameter tuning strategies, including cross-validation and grid search, have been applied to optimize performance further. The present project reflects these findings, as Random Forest was tuned and achieved an accuracy of nearly 86%, confirming its suitability for this task. A parallel research focus has been on data preprocessing and feature transformation**.** Encoding categorical variables such as gender, region, and smoking status is essential for ensuring compatibility with machine learning algorithms. Normalization and scaling improve convergence and stability, while visualization methods such as correlation heatmaps and distribution plots reveal dependencies between input features and target variables. These practices, widely reported in prior works, were also applied in this project to strengthen the reliability of predictions.

3. Background

The continuous rise in healthcare expenses has increased the importance of accurate medical insurance cost prediction. Traditional actuarial methods often overlook complex, nonlinear relationships among health and demographic factors such as age, gender, BMI, smoking status, children, and region. Machine learning algorithms, including Linear Regression, Ridge Regression, Support Vector Regression, and Random Forest, have shown promising results in capturing such patterns.

This project applies these models, with Random Forest achieving superior accuracy, and deploys the system using Flask to deliver real-time, user-friendly, and interpretable insurance premium predictions for practical decision-making.

3.1 LINEAR REGRESSION

Linear Regression is a widely used supervised learning algorithm for predicting continuous outcomes. In this project, it was applied to estimate medical insurance costs based on input factors such as age, gender, BMI, smoking habits, children, and region. The algorithm fits a straight line by minimizing the error between actual and predicted charges, assuming a linear relationship between features and cost.

Although Linear Regression is simple and interpretable, it struggles with nonlinear dependencies in the dataset. Thus, while it served as an initial baseline model, other algorithms like Ridge Regression, SVR, and Random Forest provided better accuracy. Still, Linear Regression played an important role in understanding how each feature contributes to overall insurance cost prediction.

3.2 RIDGE REGRESSOR

Ridge Regression is a regularized version of Linear Regression that improves prediction accuracy by penalizing large coefficients. In this project, Ridge Regressor was applied to predict medical insurance costs using features such as age, BMI, smoker status, children, and region. The algorithm reduces overfitting and handles multicollinearity among correlated variables by introducing an L2 penalty term. While its accuracy was lower than ensemble methods like Random Forest, Ridge Regressor provided better generalization than standard Linear Regression and served as an important comparison model for insurance premium estimation.

3.3 RANDOM FOREST

Random Forest Regressor is an ensemble-based machine learning algorithm that constructs multiple decision trees and averages their predictions to improve accuracy and reduce overfitting. In this project, it was applied to predict medical insurance costs using features such as age, BMI, smoker status, gender, children, and region. By capturing complex nonlinear relationships, Random Forest achieved superior performance compared to Linear Regression, Ridge Regression, and SVR. After hyperparameter tuning, it delivered approximately 86% accuracy, making it the most reliable model. Its robustness and generalization ability highlight its effectiveness for real-world insurance cost prediction and risk assessment.

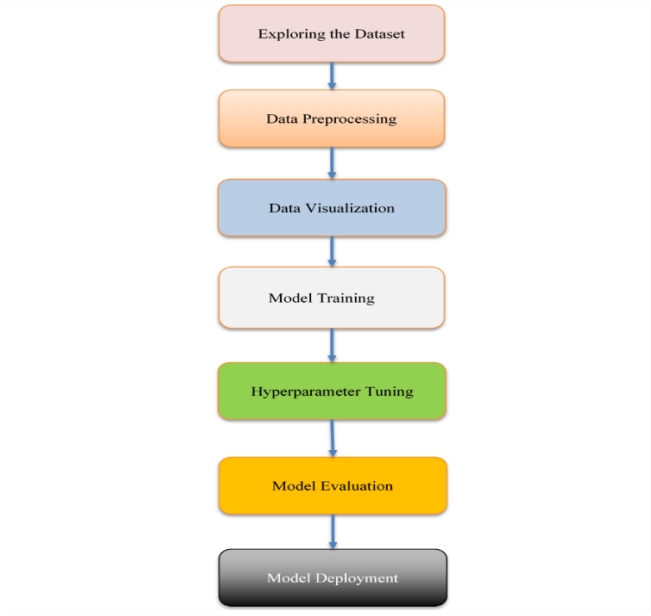

Fig.1. System Architecture Diagram

3.4 SUPPORT VECTOR REGRESSION

Support Vector Regression (SVR) is a supervised learning algorithm that predicts continuous values by fitting data within a margin of tolerance, called the epsilon-insensitive tube. In this project, SVR was

applied to predict medical insurance costs using demographic and lifestyle features such as age, BMI, smoking habits, and region. SVR captures nonlinear relationships by applying kernel functions, making it more flexible than Linear Regression. Although computationally more expensive, SVR achieved better accuracy for complex feature interactions.

4. Result and Discussion

The project implemented and compared several machine learning algorithms, including Linear Regression, Ridge Regression, Support Vector Regression (SVR), and Random Forest Regressor, to predict medical insurance costs. Linear Regression provided an interpretable baseline but was limited by its inability to model nonlinear relationships in the data. Ridge Regression improved generalization by reducing overfitting, yet its predictive accuracy remained modest. SVR captured nonlinear patterns more effectively but proved computationally expensive, limiting scalability.

Among all models, Random Forest Regressor achieved the highest performance with an accuracy of nearly 86%, demonstrating robustness and stability by aggregating multiple decision trees. The model successfully captured complex interactions among features such as age, BMI, smoker status, and region. Furthermore, deploying the system through Flask enhanced usability by enabling real-time predictions, BMI categorization, and risk-level classification. These results highlight the superiority of ensemble learning techniques and the practicality of integrating machine learning models into accessible web applications.

The cross-validation graph illustrates the performance consistency of different machine learning algorithms applied to medical insurance cost prediction. By splitting the dataset into multiple folds, the model is trained and tested on varied subsets, reducing bias and variance. Algorithms such as Linear Regression, Ridge, Random Forest, and Support Vector Regression were compared. The graph clearly shows Random Forest and Ridge Regression providing more stable and lower error rates across folds, whereas Linear Regression exhibited fluctuations. This demonstrates that ensemble and regularization methods handle complex relationships in insurance data better, ensuring more reliable premium cost estimation.

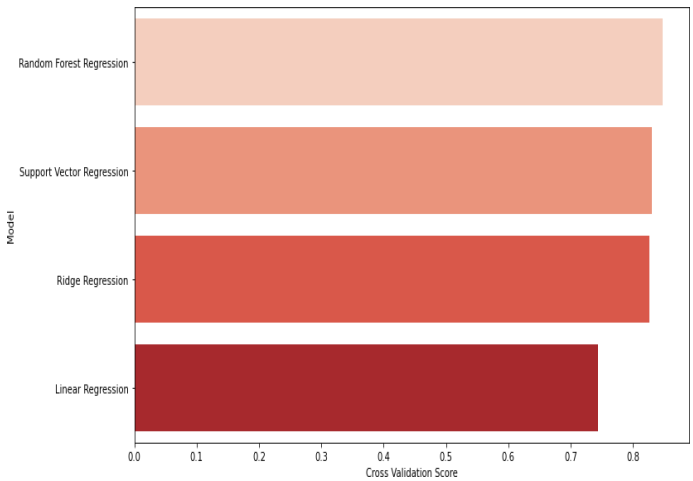

Fig.2. Cross Validation Score

5. Conclusion

The research highlights that machine learning algorithms are effective tools for predicting medical insurance costs with improved accuracy. Models like Random Forest and Ridge Regression achieved consistent performance, capturing complex feature interactions better than linear models. Such predictive systems not only support insurance companies in premium calculation but also assist individuals in understanding the financial impact of health-related factors.

References

1. Machine Learning-Based Regression Framework to Predict Health Insurance Premiums (2022). Int. J. Environ. Res. Public Health 2022, 19, 7898. https://doi.org/10.3390/ijerph19137898.
2. An Efficient Health Insurance Prediction System using Machine learning A.Vinora , V. Surya , Dr.E.Lloyds2 , B.Kathir Pandian1,R.Nancy Deborah1 and A. Gobinath1,2023 International conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES).
3. Medical Insurance Predictive Modelling: An Analysis of Machine Learning Methods, Ashok Reddy Kandula, Srinivas Kalyanapu, Sai Nithin Rayapalli, (ATMSI) | 979-8-3503-6052-3/24/$31.00 ©2024 IEEE.
4. Kulkarni, M., Mesharam, D. D., Patil, B., More, R., Sharma, M., & Patange, P. (2022). Medical insurance cost prediction using machine learning. 2022 International Journal for Research in Applied Science & Engineering Technology, 10.
5. Prediction of Insurance Cost through ML Structured Algorithm, Atharv Sharma, R. Jeya 024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT) | 79-8-3503- 8354-©2024IEEE,DOI: 10.1109/IC2PCT60090.2024.10486304.
6. Machine Learning-Based Regression Framework to Predict Health Insurance Premiums Keshav Kaushik, Akashdeep Bhardwaj, Ashutosh Dhar Dwivedi, and Rajani Singh, Int. J. Environ. Res. Public Health 2022, 19, 7898. https://doi.org/10.3390/ijerph19137898.
7. Medical Insurance Price Prediction Using Machine Learning, Md Mohtaseem Billa, Dr. Tapsi Nagpal, J. Electrical Systems 20-7s (2024): 2270-2279.
8. Ugochukwu Orji, Elochukwu Ukwandu, “Machine learning for an explainable cost prediction of medical insurance,” Machine Learning with Applications 15 (2024),https://doi.org/10.1016/j.mlwa.2023.100516.
9. Matheus Kempa Severino, Yaohao Peng, “Machine learning algorithms for fraud prediction in property insurance: Empirical evidence usingreal- world microdata,”MachineLearning with Applications(2021) https://doi.org/10.1016/j.mlwa.2021.100074.
10. Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez- Muñoz, “A comparative analysis of gradient boosting algorithms”, in 2021 Artificial Intelligence Review 54, 1 Mar 2021, pp. 1937-1967.