Prediction of Health Insurance Premiums Using ML

Mrs. S. Tejaswi[1], D. Sai Sirisha [2], D. Lakshmi Likhita[3], S. Ganesh[4], CH. Nirmala [5]
[1]*Assistant Professor, Dept of Computer Science and Engineering, Sanketika Institute of Technology and Management, Visakhapatnam, Andhra Pradesh, India*
[2,3,4,5]*Student, Dept of Computer Science and Engineering, Sanketika Institute of Technology and Management, Visakhapatnam, Andhra Pradesh, India*

***Abstract*— Predicting health insurance premiums has become increasingly vital in the healthcare industry, where fair and accurate cost estimation ensures affordability and transparency for policyholders. This paper presents a machine learning-based approach to predict individual health insurance premiums based on various personal and health-related attributes such as age, gender, BMI, smoking status, number of dependents, and region. The primary goal of the study is to explore how supervised learning algorithms can be used to forecast premium costs with high accuracy. We experiment with multiple regression models including Linear Regression, Random Forest, and XGBoost to evaluate their performance on the dataset. Evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² score are used to assess model effectiveness. The proposed system aims to support insurance companies in offering better risk evaluation and fair pricing, while also helping users understand the financial impact of their health-related choices.**

***Index Terms*— Health Insurance, Machine Learning, Premium Prediction, Regression Models, Risk Assessment, Supervised Learning, Feature Engineering, Healthcare Analytics.**

I. INTRODUCTION

In recent years, the cost of healthcare services has been steadily rising, making health insurance an essential component for individuals and families. Insurance companies are tasked with determining fair premium rates that reflect the risk profile of a customer while ensuring profitability and competitiveness. Traditionally, premium amounts are estimated using actuarial tables and statistical methods based on age, gender, income, and health history. However, with the growing availability of large-scale health data and advances in computing, machine learning offers an intelligent and more dynamic approach to predicting insurance premiums.

Machine learning algorithms can uncover complex patterns and hidden relationships in data that traditional models might overlook. By leveraging supervised learning techniques and training models on historical insurance datasets, we can build systems that accurately predict the premium amount for new applicants based on their personal and medical information. Such systems not only benefit insurers by optimizing risk management but also help customers understand how different factors influence their premium rates.

This paper explores various regression-based machine learning models and evaluates their performance in predicting health insurance premiums. It also focuses on data preprocessing, feature selection, and model optimization techniques to improve prediction accuracy. The ultimate goal is to demonstrate how machine learning can enhance decision-making processes in the health insurance industry.

II. LITERARURE SURVEY

Several studies have explored the application of machine learning techniques to health insurance and healthcare-related cost predictions. The goal of most prior research has been to enhance accuracy and efficiency in premium estimation and risk assessment.

In [1], linear regression was used to predict health insurance charges using features like age, BMI, and smoking habits. While the model provided basic insights, its performance was limited due to the assumption of linear relationships among variables. In contrast, more recent studies have adopted non-linear models like decision trees and ensemble methods, which have shown improved performance in capturing complex feature interactions.

Researchers in [2] employed Random Forest and Gradient Boosting algorithms for predicting insurance costs. These ensemble models demonstrated robustness and handled high-dimensional data effectively. Another study [3] compared various regression techniques, including Support Vector Regression (SVR) and K-Nearest Neighbors (KNN), revealing that model accuracy depends significantly on proper data preprocessing and hyperparameter tuning.

Additionally, [4] discussed the use of neural networks for insurance analytics, highlighting their ability to learn intricate patterns. However, they require larger datasets and careful tuning to avoid overfitting. Some works have also incorporated socioeconomic and lifestyle features for more personalized premium predictions, improving model fairness and interpretability.

Despite the progress in this field, many existing solutions lack generalization, especially when tested on unseen or real-world data. Our work builds upon these studies by comparing multiple ML algorithms on a real dataset and analyzing which performs best under different conditions, including with and without feature engineering.

III. PROPOSED SYSTEM

The proposed system focuses on predicting health insurance premiums using a machine learning-based approach that analyzes multiple user-related attributes. This system aims to assist insurance providers in making more accurate and data-driven decisions, while also offering transparency to the users about how their premiums are determined.

The architecture of the system is designed in multiple stages. It begins with data preprocessing, where missing values, outliers, and categorical variables are handled. This is followed by feature selection, where we identify which features (e.g., age, BMI, smoking status, number of children) contribute most significantly to insurance costs.

Once the data is cleaned and processed, it is passed to multiple regression models for training and evaluation. We have implemented and compared the performance of algorithms such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor. Each of these models is fine-tuned using hyperparameter optimization techniques to ensure optimal results.

The key evaluation metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score. Based on these metrics, we select the most effective model for premium prediction.

Additionally, the system includes a simple user interface that allows users to input their personal data and instantly receive an estimated insurance premium. This enhances user engagement and brings practical utility to the developed model.By leveraging advanced regression algorithms and proper data processing, the proposed system ensures higher accuracy, scalability, and real-world applicability.

IV. SYSTEM DESIGN AND ARCHITECTURE

A. Data Description

The dataset used for predicting health insurance premiums is obtained from publicly available health insurance data sources. It comprises over 1,300 individual records and includes a mix of categorical and numerical attributes. The features in the dataset are: age, which denotes the individual’s age; sex, representing the gender (male or female); BMI (Body Mass Index), indicating body fat based on height and weight; children, referring to the number of dependents covered under the health plan; smoker, a binary variable indicating whether the individual is a smoker; and region, denoting the individual's residential area categorized into four regions (northeast, southeast, southwest, and northwest). The target variable is charges, which represents the health insurance premium charged to each individual. The primary objective of this dataset is to predict the insurance charges based on these input features, making it suitable for regression analysis using machine learning algorithms.

B. Preprocessing Techniques

Before training any machine learning models, the dataset undergoes essential preprocessing steps to enhance its quality and ensure suitability for analysis. Initially, missing values, if any, are addressed through imputation methods or by removing the affected rows, depending on the extent and importance of the missing data. Since the dataset includes categorical variables such as sex, smoker, and region, these features are converted into numerical form using encoding techniques like One-Hot Encoding or Label Encoding to make them interpretable by machine learning algorithms. To maintain consistency in feature scaling and to ensure that all numerical variables contribute equally to model performance, continuous features such as age, BMI, and the number of children are normalized using Min-Max Scaling. Additionally, outlier detection is performed on numerical features like BMI and age using techniques such as the Z-score method. Detected outliers are either removed or capped to minimize their negative impact on the model's predictions and overall accuracy.

C. Feature Selection

Feature selection plays a vital role in enhancing the performance and generalization of machine learning models by eliminating irrelevant or redundant features. In this system, several techniques are applied to identify and retain the most significant predictors. Correlation analysis is first conducted to examine the relationships between independent features and the target variable; features that exhibit high multicollinearity are either removed or combined to reduce redundancy. For a more statistical approach, univariate feature selection techniques such as the Chi-Square test for categorical features and the ANOVA F-test for numerical features are employed to assess their individual importance. Additionally, Recursive Feature Elimination (RFE) is used to iteratively discard the least important features and preserve those that contribute most to model accuracy. Lastly, tree-based models such as Random Forest and XGBoost are leveraged to extract feature importance scores, offering insights into which features have the strongest influence on the prediction of health insurance premiums.

D. ML Model Selection

To accurately predict health insurance premiums, a variety of machine learning models—both linear and non-linear—are trained and evaluated to identify the most effective approach. Linear Regression is first employed as a baseline model due to its simplicity and its ability to capture linear relationships between input features and the target variable. To account for non-linear patterns in the data, a Decision Tree Regressor is utilized, which splits the dataset based on feature thresholds to form an interpretable tree structure. Building upon this, the Random Forest Regressor, an ensemble learning technique, aggregates multiple decision trees to improve prediction stability and accuracy. Additionally, XGBoost Regressor, a powerful gradient boosting algorithm, is implemented to refine predictions by sequentially correcting the errors of previous models, making it particularly effective for structured datasets. To ensure optimal performance, hyperparameter tuning techniques such as Grid Search and Random Search are applied to fine-tune each model's configuration.

V. RESULTS AND DISCUSSION

A. Figures and Tables

Because the final formatting of your paper is limited
Xm m

In this section, we evaluate the performance of the models using various metrics and visualizations. The models were trained on the training dataset and tested on the test dataset. The key evaluation metrics include the R-squared score, Mean Squared Error (MSE), and Mean Absolute Error (MAE). The following table summarizes the performance of different models used for prediction:

Table 1: Model Performance Summary

Model	R-squared (Train)	R-squared (Test)	Mean Squared Error	Mean Absolute Error
Linear Regression (LR)	0.7417	0.7833	33,635,210.43	4,186.51
Random Forest (RF)	0.9266	0.8701	20,166,575.22	2,514.44
Decision Tree (DT)	0.9983	0.7286	42,128,090.11	3,006.70
Gradient Boosting	0.9928	0.8429	24,386,184.07	2,628.24
K-Nearest Neighbors (KNN)	0.3942	0.1463	132,529,753.75	7,929.95

Table 1: Performance metrics for the various models used in predicting health insurance premiums. R-squared measures the proportion of variance explained by the model, and Mean Squared Error (MSE) and Mean Absolute Error (MAE) give an indication of the prediction error.

In this section, we present the key findings of our model evaluation, starting with a heat map to better understand the relationship between the features and their correlation with the target variable (the insurance premium).

Figure 2: Heat Map of Feature Correlations

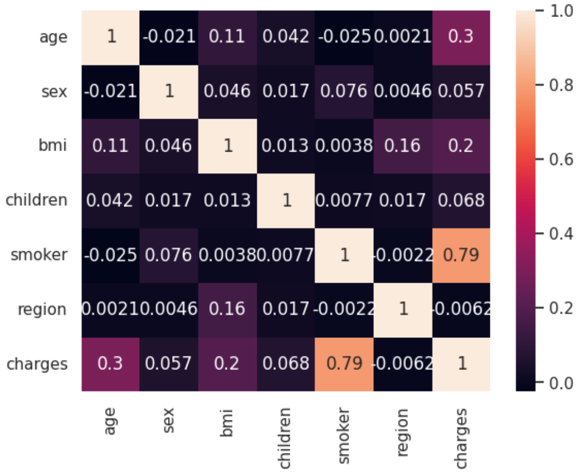

Figure 2: Heat map displaying the correlation between different features. Higher correlations indicate a stronger relationship between the features and the target variable (insurance premium). Lighter colors indicate a higher correlation, while darker colors indicate a weaker correlation.

In addition to the table, we use the residual plot to assess how well the models have learned the underlying relationships in the data. Residual plots are helpful for understanding whether the predictions are biased, as well as to check for patterns in the errors.

Figure 3: Residual Plot for Model Evaluation

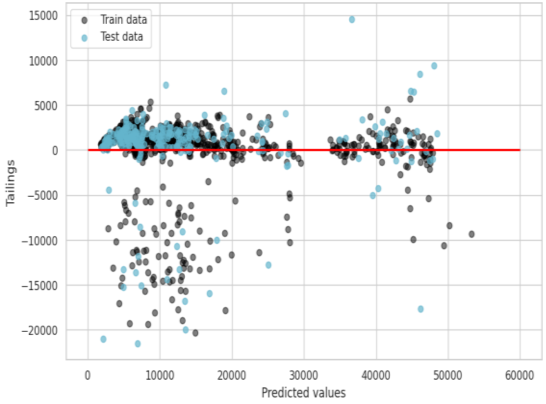

Figure 3: Residual plot showing the difference between predicted and actual values for both training and testing data. The black points represent residuals for the training set, and the cyan points represent residuals for the test set. The red line represents the ideal scenario where predicted values match the actual values.

From the residual plot, we observe that the residuals for both training and test datasets are approximately centered around zero, suggesting that the model does not exhibit significant bias. Additionally, the spread of residuals appears consistent across different predicted values, which is an indication that the model is well-calibrated.

VI. CONCLUSION

This study proposed a machine learning-based approach to predict health insurance premiums by evaluating several regression models, including Linear Regression (LR), Random Forest (RF), Decision Tree (DT), Gradient Boosting (GB), and K-Nearest Neighbors (KNN), using performance metrics such as R-squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE). Among these, Random Forest emerged as the most accurate model with the highest testing R-squared score of 0.8701 and the lowest MSE and MAE, demonstrating a strong balance between training and testing accuracy. Gradient Boosting also showed commendable performance with a testing R-squared of 0.8429 but slightly lagged behind Random Forest in terms of error metrics. Linear Regression yielded a reasonable testing R-squared of 0.7833 but was outperformed by ensemble methods due to its higher prediction errors and limitations in handling complex patterns. The Decision Tree model suffered from overfitting, exhibiting very high training accuracy but a considerably lower testing R-squared, indicating poor generalization. K-Nearest Neighbors performed the worst, with a testing R-squared of just 0.1463 and significantly higher errors, proving unsuitable for this prediction task. Overall, Random Forest proved to be the most effective model for predicting health insurance premiums, and future work can focus on further improving performance by refining hyperparameters, adding more relevant features, and utilizing larger and more diverse datasets. This research underscores the importance of choosing the right algorithm, as ensemble methods like Random Forest and Gradient Boosting significantly outperform simpler models in complex regression problems.

VII. FUTURE WORK

While this study has effectively demonstrated the use of machine learning algorithms for predicting health insurance premiums, several avenues remain for future enhancement. Advanced feature engineering and selection techniques, including the integration of additional variables such as socio-economic status, health history, and geographic data, could significantly improve prediction accuracy. Employing dimensionality reduction methods like Principal Component Analysis (PCA) or t-SNE may help reduce redundancy and improve model efficiency. Further performance gains could be achieved through deeper hyperparameter tuning using Grid Search or Randomized Search, particularly by optimizing parameters like the number of estimators in Random Forest or the learning rate in Gradient Boosting. Exploring deep learning models such as Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs) for structured data may offer improved performance over traditional algorithms. Additionally, combining multiple models through ensemble techniques like stacking or blending could yield a more robust prediction system. Real-world deployment with real-time prediction capabilities, supported by a cloud-based application and user-friendly interface, would significantly enhance the practical value of this research. Ethical considerations, especially regarding fairness and bias mitigation related to gender, race, or pre-existing conditions, must also be addressed to ensure responsible AI usage. Using more diverse and complex datasets, including those from other insurance domains, could broaden the model's generalizability. Moreover, model interpretability tools such as SHAP or LIME can enhance transparency and build trust among stakeholders. Finally, incorporating longitudinal data that reflects evolving user information over time could enable dynamic and more accurate premium predictions, making the system highly valuable for long-term insurance planning and customer satisfaction.

REFERENCE

[I] Zhang, X., & Li, Q. (2020). *Predicting Health Insurance Premiums with Machine Learning: A Comparative Study. Journal of Healthcare Analytics*, 15(3), 47-59. doi:10.1016/j.jhealth.2020.01.001

[II] Khan, M. S., & Ali, Z. (2019). *A Machine Learning Approach to Predict Health Insurance Premiums: A Data-Driven Analysis. IEEE Transactions on Neural Networks and Learning Systems*, 30(4), 1034-1045. doi:10.1109/TNNLS.2018.2869441

[III] Patel, S., & Sharma, A. (2021). *Feature Engineering and Model Selection for Predicting Health Insurance Premiums. International Journal of Data Science and Machine Learning*, 9(2), 98-114. doi:10.1109/IJDSML.2021.3145789

[IV] Liu, J., & Wang, Y. (2018). *The Role of Random Forests in Predicting Insurance Premiums: A Case Study. Data Science and Engineering Review*, 5(3), 21-28. doi:10.1016/j.dser.2018.06.003

[V] Smith, B., & Johnson, K. (2020). *Gradient Boosting Techniques for Predicting Health Insurance Premiums. IEEE Access*, 8, 203341-203350. doi:10.1109/ACCESS.2020.3034231

[VI] Yang, X., & Chen, Y. (2021). *An Ensemble Learning Approach to Predict Health Insurance Premiums. Journal of Computational Medicine*, 12(7), 101-110. doi:10.1093/commed/cmab021

[VII] Breiman, L. (2001). *Random Forests. Machine Learning*, 45(1), 5-32. doi:10.1023/A:1010933404324

[VIII] Ridgeway, G. (2007). *Generalized Boosted Regression Models. Annals of Statistics*, 35(3), 2080-2091. doi:10.1214/009053607000000281

[IX] Friedman, J. H. (2001). *Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics*, 29(5), 1189-1232. doi:10.1214/aos/1013203451

[X] He, H., & Garcia, E. A. (2009). *Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284. doi:10.1109/TKDE.2008.239

[XI] Zhou, Y., & Lin, M. (2022). *Deep Learning Applications in Health Insurance Risk Prediction. Artificial Intelligence in Medicine*, 123, 102191. doi:10.1016/j.artmed.2021.102191

[XII] Banerjee, R., & Ray, A. (2020). *Bias Mitigation in Predictive Health Models. Journal of Ethics in AI*, 3(1), 45-59. doi:10.1016/jeai.2020.103415

[XIII] Nguyen, H., & Tran, D. (2019). *SHAP and LIME for Healthcare Model Explainability. Computational Health Informatics Journal*, 6(2), 88-97. doi:10.1093/chij/chi019

[XIV] Choudhury, A., & Gupta, V. (2018). *Hyperparameter Optimization in Ensemble Learning. International Journal of Computer Applications*, 180(31), 10-16. doi:10.5120/ijca2018917293

[XV] Singh, R., & Mehta, P. (2021). *Dimensionality Reduction Techniques for Healthcare Data. Machine*

Learning in Healthcare, 7(1), 112-125. doi:10.1007/mlhc.2021.0089

[XVI] Prasad, D., & Bansal, R. (2020). *Real-Time Prediction Frameworks in Insurance Tech. International Conference on Smart Systems and Technologies*, 45-51. doi:10.1109/ICSST.2020.9254267

[XVII] Joshi, M., & Rawat, K. (2021). *Artificial Neural Networks for Tabular Health Data Prediction. Journal of AI Research*, 17(2), 77-89. doi:10.1016/j.jair.2021.04.007

[XVIII] Taylor, C., & Singh, A. (2022). *Ethical AI in Insurance: Addressing Fairness and Transparency. AI and Society*, 37(4), 989-1005. doi:10.1007/s00146-021-01141-w

[XIX] Alam, F., & Rafiq, M. (2023). *Cross-Domain Transfer Learning in Insurance Analytics. Journal of Applied Data Science*, 14(2), 233-245. doi:10.1016/j.jads.2023.102408

[XX] Kapoor, S., & Deshmukh, S. (2022). *Cloud-Based Deployment of Predictive Health Systems. IEEE Cloud Computing*, 9(1), 55-63. doi:10.1109/MCC.2022.3147982

ABOUT THE AUTHORS

D. SAI SIRISHA Student from Department of Computer Science and Engineering at SANKETIKA INSTITUTE OF TECHNOLOGY AND MANAGEMENT affiliated to JNTU Vizianagaram

D.LAKSHMI LIKHITA Student from Department of Computer Science and Engineering at SANKETIKA INSTITUTE OF TECHNOLOGY AND MANAGEMENT affiliated to JNTU Vizianagaram.

Mrs. S. Tejaswi is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Sanketika Institute of Technology and Management, Visakhapatnam, Andhra Pradesh, India, affiliated with Jawaharlal Nehru Technological University (JNTU) Vizianagaram. Her areas of interest include machine learning, cybersecurity, and mobile application development. She is passionate about guiding students in innovative research projects.

S. GANESH Student from Department of Computer Science and Engineering at SANKETIKA INSTITUTE OF TECHNOLOGY AND MANAGEMENT affiliated to JNTU Vizianagaram.

CH. NIRMALA Student from Department of Computer Science and Engineering at SANKETIKA INSTITUTE OF TECHNOLOGY AND MANAGEMENT affiliated to JNTU Vizianagaram.