Journal of Computer and Electrical Sciences (JCES)- Volume-1 Issue-4, April 2026
ISSN: 3049-2602

PERFORMANCE ANALYSIS OF MACHINE LEARNING ALGORITHMS TO PREDICT HEALTH INSURANCE PREMIUM

MUBEENA SHAIK[1], DR. K. SIREESHA[2]

[1]Student, Department of Computer Science and Engineering,
Andhra Loyola Institute of Engineering and Technology, Vijayawada,
Andhra Pradesh, India
[2]Associate Professor, Department of Computer Science and Engineering,
Andhra Loyola Institute of Engineering and Technology, Vijayawada,
Andhra Pradesh, India.
Email id: mubeenashaik134@gmail.com

Abstract: Health insurance premium prediction is a critical task in the modern insurance industry. Accurate premium estimation helps companies design fair policies while enabling customers to understand financial commitments. Traditional approaches often fail to capture complex relationships among multiple influencing factors such as age, BMI, lifestyle, and region. This paper presents a comprehensive performance analysis of machine learning algorithms including Linear Regression, Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting. The study evaluates models using MAE, MSE, RMSE, and R^2 score. Results indicate that ensemble models outperform traditional techniques.

Keywords: Machine Learning, Health Insurance, Prediction, Random Forest, Gradient Boosting, Data Analysis

1.INTRODUCTION

Health insurance is essential for safeguarding individuals and families against sudden and costly medical expenses by offering financial support during emergencies. However, calculating premiums is challenging because it involves several factors like age, BMI, lifestyle habits such as smoking, number of dependents, and location. Traditional methods often rely on manual calculations and basic statistics, which may not fully capture these complex relationships. With the rise of data and modern technology, machine learning provides a more efficient approach by analyzing large datasets, identifying patterns, and improving accuracy over time. This study focuses on applying and comparing different machine learning models to predict insurance premiums more effectively, helping companies make smarter and more informed decisions, while also enhancing transparency and fairness in premium estimation.

2. LITERATURE SURVEY

Researchers have widely explored machine learning techniques in the insurance sector to estimate policy costs and assess risk more accurately. Linear Regression is commonly used due to its simplicity and ease of interpretation, but it struggles with complex, non-linear data patterns found

https://doi.org/10.5281/zenodo.19598216

in real-world scenarios. To address this, models like Decision Trees are applied as they can better capture interactions and non-linear relationships between variables. More advanced approaches such as Random Forest and Gradient Boosting further enhance prediction accuracy by combining multiple models, reducing errors and overfitting. Additionally, Artificial Neural Networks are used for handling large and complex datasets, and overall, studies show that ensemble methods tend to outperform traditional single-model techniques in predicting insurance premiums.

3. PROPOSED SYSTEM

The proposed system focuses on predicting health insurance premiums using different machine learning algorithms. It considers key factors such as age, Body Mass Index (BMI), smoking habits, and the number of children, as these significantly influence insurance costs. By analyzing these inputs, the system identifies patterns and relationships within the data to generate accurate and realistic predictions.

The process involves data preparation, training multiple machine learning models, and evaluating their performance to select the most suitable one. This approach improves prediction accuracy by choosing the best-performing model. It also reduces the time and manual effort required for decision-making. Additionally, it ensures more consistent and reliable results compared to traditional methods. Overall, this makes the system more efficient and practical for use.

Furthermore, the system can be extended into a web-based application, allowing users to easily access real-time predictions through a simple and interactive interface. This makes the system more user-friendly and accessible to a wider audience. It can also support continuous updates and improvements based on new data, ensuring better performance over time. In addition, integrating it with databases helps in efficient data storage and management. The system can be designed to handle multiple users simultaneously without affecting performance. Overall, this enhances practicality and makes it suitable for real-world deployment.

https://doi.org/10.5281/zenodo.19598216

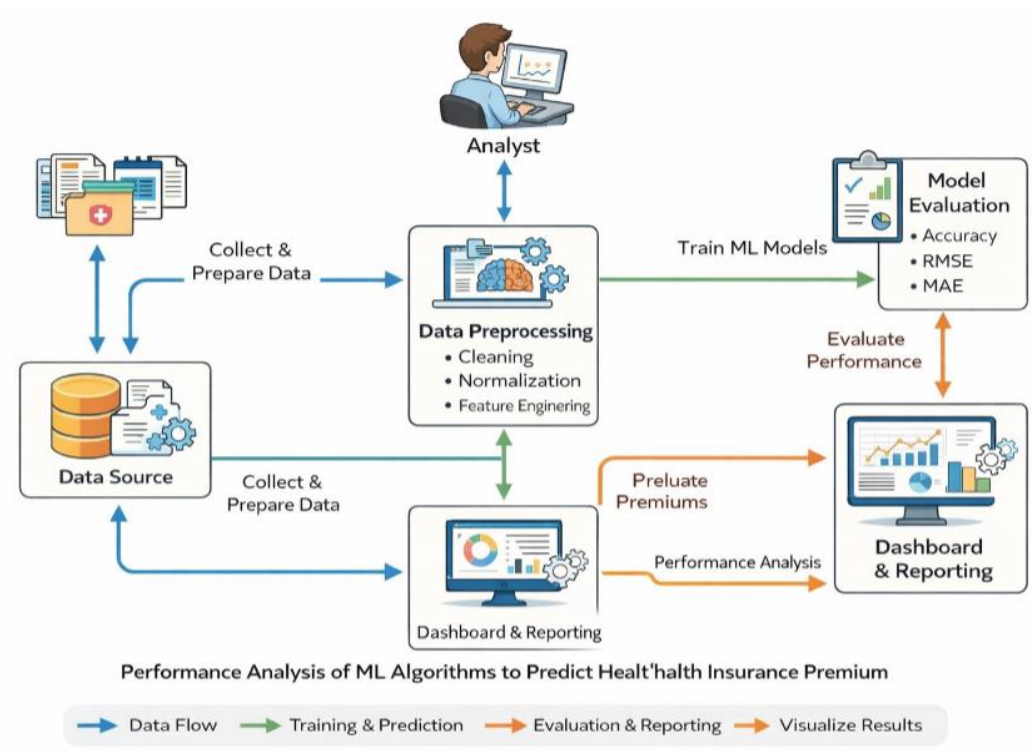

Fig1: Proposed System

4. METHODOLOGY

1.Data Preprocessing: Data preprocessing involves cleaning and preparing the dataset by handling missing values, converting categorical data into numerical form, and scaling numerical features. These steps improve data quality and help the models learn more effectively.

2.Data Splitting: The dataset is divided into training and testing sets to ensure proper evaluation of the models. This helps in checking how well the model performs on new, unseen data.

3.Model Training: Different machine learning algorithms are trained using the prepared dataset. This allows the models to learn patterns and relationships between input features and the target variable.

4.Model Evaluation: The performance of the models is evaluated using metrics such as MAE, MSE, RMSE, and R^2 score. These metrics help in comparing models and selecting the most accurate one.

5.Feature Engineering: Feature engineering focuses on selecting the most important variables that influence predictions. This improves both the accuracy and efficiency of the models.

6.Overall Process: All these steps are combined in a structured way to ensure reliable and accurate predictions. This approach enhances the overall performance and effectiveness of the system.

https://doi.org/10.5281/zenodo.19598216

5. RESULTS AND DISCUSSION

The results of the experiments show that all the machine learning models used in this study are able to predict health insurance premiums with a good level of accuracy. Linear Regression acts as a basic model, providing simple and easy-to-understand results, but it is limited when it comes to capturing complex patterns in the data. On the other hand, Decision Tree models perform better by handling non-linear relationships and interactions between different features more effectively.

More advanced techniques like Random Forest and Gradient Boosting deliver the best performance because they use an ensemble approach, combining multiple models to improve accuracy. Support Vector Machine (SVM) produces acceptable results but requires proper tuning of parameters to work effectively. Overall, the results clearly show that ensemble methods are more reliable and accurate compared to individual models, making them a better choice for predicting health insurance premiums.

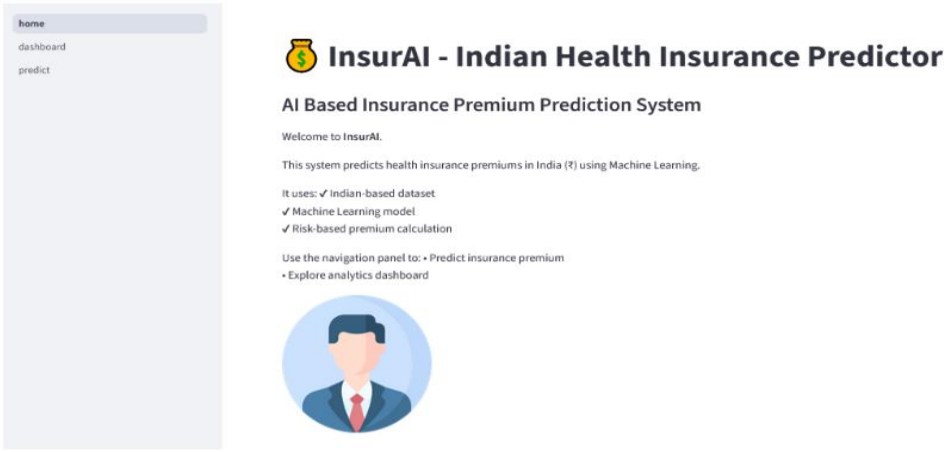

Fig 2: Home Page

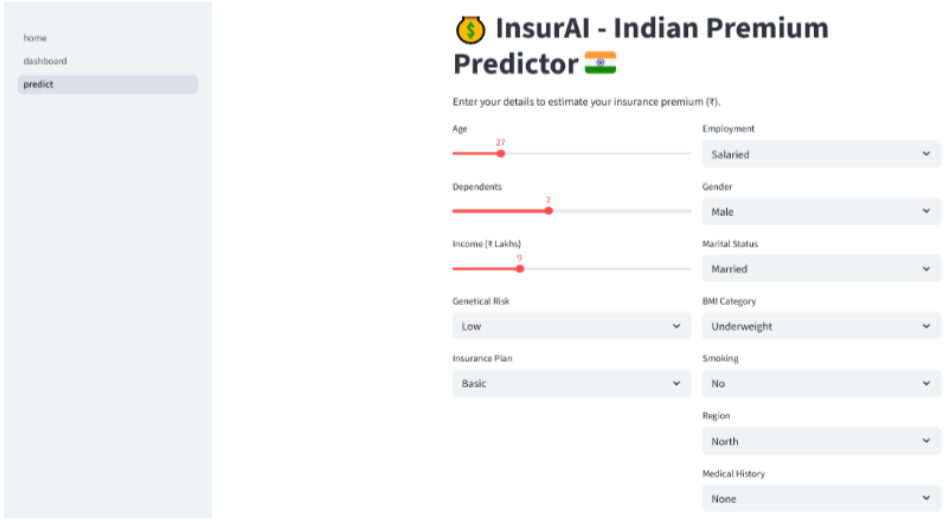

Fig 3: Insurance Premium Prediction Page

https://doi.org/10.5281/zenodo.19598216

Fig 4: Prediction Results

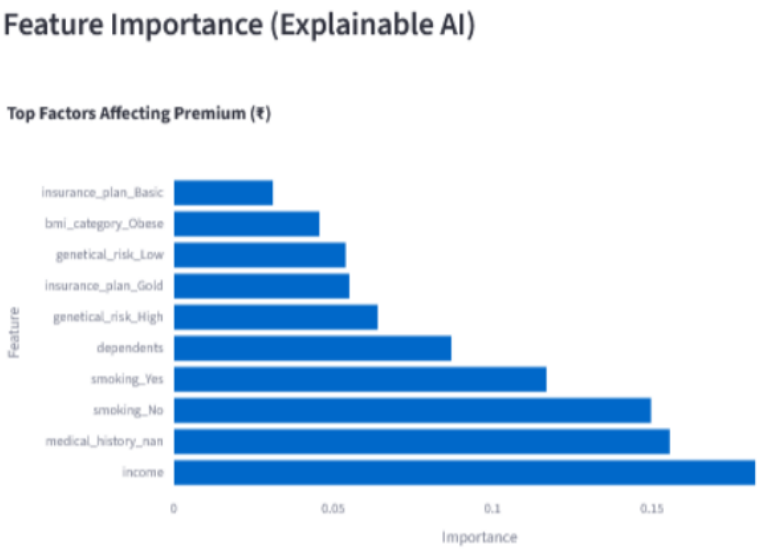

Fig 6: Future Importance

6. CONCLUSION

Machine learning proves to be highly effective in predicting health insurance premiums by improving both accuracy and reliability. Compared to traditional methods, these models use data-driven approaches to analyze large amounts of information efficiently. They are capable of identifying hidden patterns and relationships that are not easily captured through manual calculations. This leads to more precise and consistent prediction results. Overall, the use of machine learning enhances the quality of premium estimation.

https://doi.org/10.5281/zenodo.19598216

Among the different techniques, ensemble methods like Random Forest and Gradient Boosting show the best performance. By combining multiple models, they handle complex data patterns more effectively and reduce prediction errors, resulting in better accuracy.

The proposed system can be highly beneficial for insurance companies in automating the premium calculation process. It reduces manual effort and speeds up decision-making with reliable predictions. At the same time, it helps customers understand how factors like age, BMI, and lifestyle influence their insurance costs. This improves transparency and builds trust between companies and users. Additionally, the system provides a user-friendly experience through clear and accessible results. Overall, it makes the entire insurance process more efficient and practical for real-world use

7. FUTURE SCOPE

Future improvements can focus on increasing the accuracy of the system by incorporating advanced deep learning techniques. These models are capable of identifying more complex patterns in data, which can lead to better and more precise predictions compared to traditional machine learning methods.

The system can also be developed into a real-time web or mobile application, making it more accessible and convenient for users. This would allow individuals to obtain instant premium predictions anytime and from anywhere, improving the overall user experience.

Furthermore, integrating the system with healthcare databases and IoT devices can provide more accurate and up-to-date information. The use of advanced methods such as XGBoost and other AI-based analytics can further enhance the performance, reliability, and efficiency of the prediction system.

REFERENCES

[1] Leo Breiman, “Random Forests,” *Machine Learning Journal*, vol. 45, no. 1, pp. 5–32, 2001.
[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.
[3] Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O’Reilly Media, 2019.
[4] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
[5] Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.
[6] Scikit-learn Developers, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, Deep Learning, MIT Press, 2016.
[8] Andreas C. Müller and Sarah Guido, Introduction to Machine Learning with Python, O’Reilly Media, 2016.
[9] Jake VanderPlas, *Python Data Science Handbook*, O’Reilly Media, 2016.
[10] UCI Machine Learning Repository, *“Machine Learning Datasets,”* Available: https://archive.ics.uci.edu

https://doi.org/10.5281/zenodo.19598216