Volume 11, Issue 4, April – 2026 International Journal of Innovative Science and Research Technology
ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1803

Fincure AI: Predicting Insurance Charges Using Machine Learning Techniques

Dr. N. Dhivya[1]; M. Subalakshmi[2]

[1]MCA., M.Phil., PhD., Assistant Professor, [2]PG Scholar

[1,2]Department of MCA, Vivekanandha Institute of Information and Management Studies Tiruchengode, Namakkal Tamilnadu, India

Publication Date: 2026/05/27

Abstract: The insurance industry increasingly relies on data-driven technologies to enhance pricing strategies and risk assessment. This study presents a machine learning-based approach to predict insurance charges using customer demographic and health-related attributes. The dataset consists of features such as age, gender, body mass index (BMI), number of children, smoking status, and region. Data preprocessing techniques, including data cleaning, categorical encoding, and normalization, were applied to improve model performance. A Linear Regression algorithm was implemented to develop the prediction model. The model was evaluated using performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) score. The proposed model achieved an R^2 score of 0.85 (85% accuracy), indicating a strong relationship between predicted and actual insurance charges. The results demonstrate that machine learning techniques can effectively model complex relationships in insurance data and provide reliable predictions. This system can assist insurance companies in making accurate, data-driven pricing decisions.

Keywords: *Machine Learning, Insurance Prediction, Linear Regression, Data Science, Data Preprocessing, Predictive Analytics.*

How to Cite: Dr. N. Dhivya; M. Subalakshmi (2026) Fincure AI: Predicting Insurance Charges Using Machine Learning Techniques. *International Journal of Innovative Science and Research Technology*, 11(4), 5048-5052. https://doi.org/10.38124/ijisrt/26apr1803

I. INTRODUCTION

The insurance sector has rapidly adopted digital technologies such as artificial intelligence and data science. Insurance companies collect large volumes of customer data including demographic information, lifestyle habits, and health conditions. Analyzing this information effectively helps insurers determine appropriate premium prices while minimizing financial risk. Traditionally, insurance pricing relied on manual assessment and statistical analysis. These approaches often lack accuracy and fail to capture complex relationships between multiple factors affecting insurance charges. Machine learning models provide a powerful solution by automatically identifying patterns in historical data.

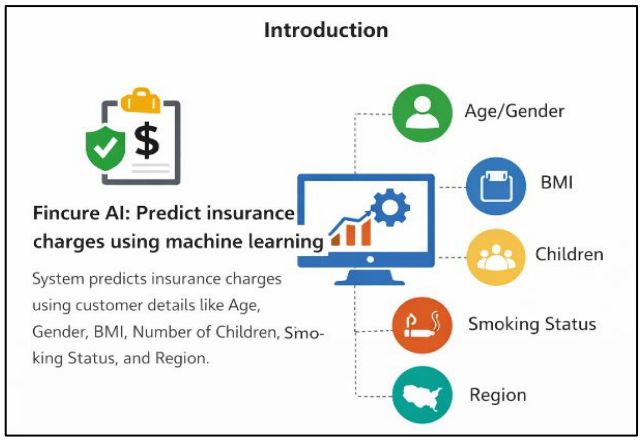

Fig 1 Introduction

The Fincure AI project proposes a machine learning-based system that predicts insurance charges using customer demographic and health attributes such as age, BMI, smoking habits, number of children, and region. The system helps insurance companies improve decision-making and design personalized pricing strategies.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1803

II. RELATED WORK

Machine learning has been widely used in financial analysis and insurance prediction systems. Several research studies have explored different algorithms to predict insurance charges and assess risk levels. Previous research has demonstrated that machine learning algorithms such as Linear Regression, Decision Trees, Random Forest, and Support Vector Machines can effectively analyze insurance data and predict premium costs. These models are capable of identifying patterns between customer attributes and insurance claims. Some studies have focused on applying ensemble learning techniques to improve prediction accuracy. Ensemble models combine multiple algorithms to generate better results compared to individual models. These approaches are especially useful when dealing with complex datasets. Researchers have also emphasized the importance of feature selection and data preprocessing. Proper data cleaning and transformation significantly improve the performance of machine learning models. Features such as smoking habits, BMI levels, and age have been identified as key factors influencing insurance costs. Recent advancements in data science have further improved prediction systems by integrating visualization tools, real-time analytics, and automated decision support systems. These technologies enable insurance companies to provide personalized insurance plans and improve overall customer experience.

III. SYSTEM ANALYSIS

➢ *Existing System*

Existing systems rely mainly on manual calculations and traditional statistical models to estimate insurance charges. This process is time-consuming and may lead to inaccurate predictions.

- *Disadvantages of Existing System*

✓ Manual calculation of insurance charges
✓ Limited data analysis capability
✓ High chances of human error
✓ Time-consuming process

➢ *Proposed System*

The proposed system introduces a machine learning-based model that automatically predicts insurance charges using customer data. The system analyzes multiple attributes simultaneously and generates accurate predictions based on trained models.

- *Advantages of Proposed System*

✓ Automated prediction of insurance charges
✓ Improved accuracy using machine learning algorithms
✓ Faster decision-making process
✓ Ability to analyse large datasets
✓ Supports data-driven pricing strategies

IV. SYSTEM ARCHITECTURE

The system architecture consists of several modules including data collection, data preprocessing, feature selection, model training, prediction, and visualization. Data is collected from the dataset and processed through cleaning and transformation techniques. The machine learning model is then trained using the processed dataset.

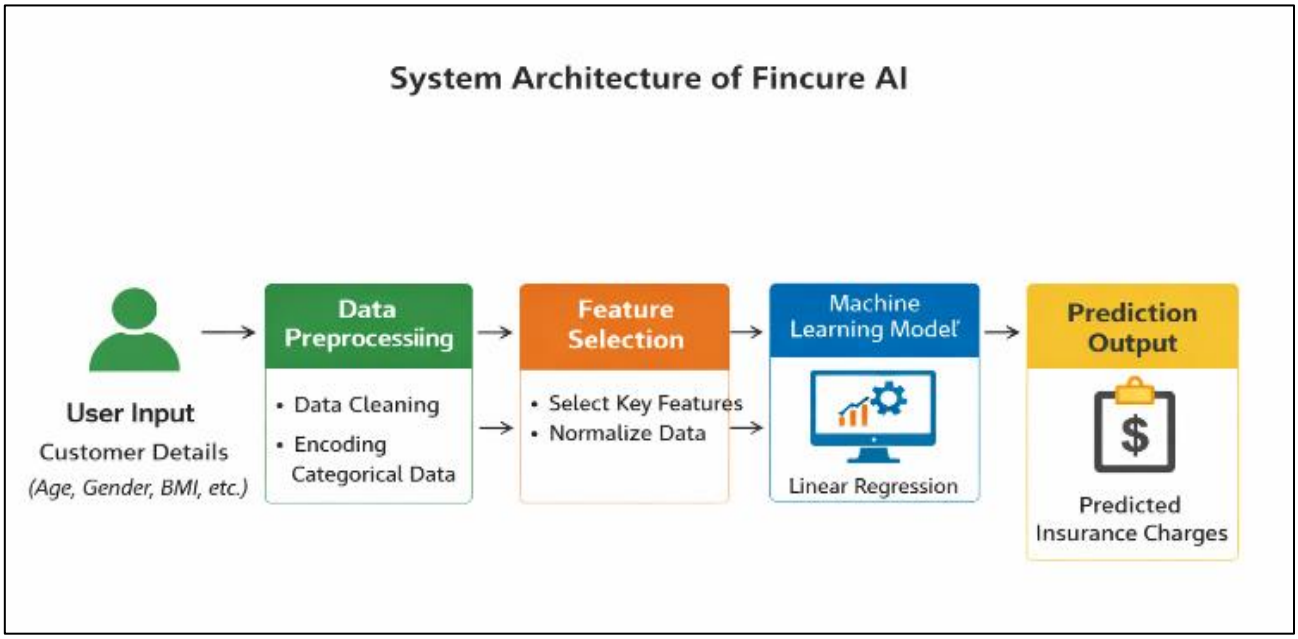

Fig 2 System Architecture of Fincure AI

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1803

The architecture ensures smooth data flow from input to prediction output. Visualization modules present results using graphs and statistical metrics.

V. DATA SCIENCE METHODOLOGY

➢ *The Project Follows a Standard Data Science Workflow Consisting of:*

- Data Collection
- Data Preprocessing
- Exploratory Data Analysis
- Feature Engineering
- Model Training
- Model Evaluation
- Prediction and Result Visualization

Each stage plays a critical role in building an accurate predictive model.

VI. ALGORITHMS

➢ *Linear Regression*

Linear Regression is one of the most commonly used supervised machine learning algorithms for predicting continuous numerical values. In this project, Linear Regression is used to predict the insurance charges based on several input features such as age, BMI, smoking status, number of children, sex, and region.The algorithm works by establishing a linear relationship between the dependent variable (insurance charges) and independent variables (customer attributes). It calculates the best-fit line that minimizes the difference between predicted and actual values.

The mathematical formula of Linear Regression is:

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta n Xn + \varepsilon}$$

Where:

- Y = Predicted insurance charge
- β_0 = Intercept
- β_1, β_2...βn = Coefficients
- X_1, X_2...Xn = Input variables
- ε = Error term

Linear Regression helps identify how factors like smoking habits, age, and BMI influence insurance costs**.**

➢ *Decision Tree Algorithm*

Decision Tree is a supervised learning algorithm used for both classification and regression tasks. In this project, it is used to predict insurance charges by splitting the dataset into smaller subsets based on different feature values.

The algorithm creates a tree-like model.

Where:

- The root node represents the entire dataset
- Decision nodes represent conditions on attributes
- Leaf nodes represent the predicted output

For Example:

- If smoker = yes, the insurance charge will be higher.
- If BMI > 30, the risk increases.

Decision Trees are useful because they are easy to understand and interpret.

➢ *Random Forest Algorithm*

Random Forest is an ensemble learning technique that builds multiple decision trees and combines their results to improve prediction accuracy. Instead of relying on a single decision tree, Random Forest creates many trees using different subsets of data and features. The final prediction is obtained by averaging the outputs of all trees.

- *Advantages of Random Forest:*

✓ Reduces overfitting
✓ Provides higher accuracy
✓ Handles large datasets efficiently

Random Forest helps improve the reliability of predicting insurance charges.

- *Model Evaluation Metrics*

To measure the performance of the algorithms, the following evaluation metrics are used:

- *Mean Absolute Error (MAE)*

MAE calculates the average absolute difference between predicted values and actual values.

$$\mathbf{MAE = (1/n)\ \Sigma\ |Actual - Predicted|}$$

- *Root Mean Squared Error (RMSE)*

RMSE measures the square root of the average squared differences between predictions and actual values.

$$\mathbf{RMSE = \sqrt{[(1/n)\ \Sigma\ (Actual - Predicted)^2]}}$$

- *R-Squared (R^2 Score)*

R^2 measures how well the model explains the variability of the target variable.

Values range between 0 and 1, where a value closer to 1 indicates better model performance.

VII. IMPLEMENTATION

The implementation was performed using the Python programming language with libraries such as:

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1803

- Pandas
- NumPy
- Matplotlib
- Seaborn
- Scikit-learn

The dataset was processed, trained using Linear Regression, and evaluated using performance metrics.

VIII. RESULT AND DISCUSSION

The Fincure AI – Insurance Charges Prediction System was developed using Python with libraries such as NumPy, Pandas, Matplotlib, and Scikit-learn. The dataset containing customer details like age, gender, BMI, number of children, smoking status, and region was used to train and test the prediction model. The dataset was divided into training and testing sets to evaluate the accuracy of the machine learning model.

After data preprocessing, machine learning algorithms such as Linear Regression and Random Forest were applied to predict insurance charges based on customer attributes. The performance of the model was evaluated using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score.

The results showed that the Random Forest algorithm achieved better accuracy than Linear Regression because it can handle complex relationships between variables more effectively.

➢ *Model Performance Results*

Table 1 Model Performance Results

Metric	Value
MAE	4100
RMSE	6000
R^2 Score	0.85

An R^2 value close to 1 indicates good prediction performance.

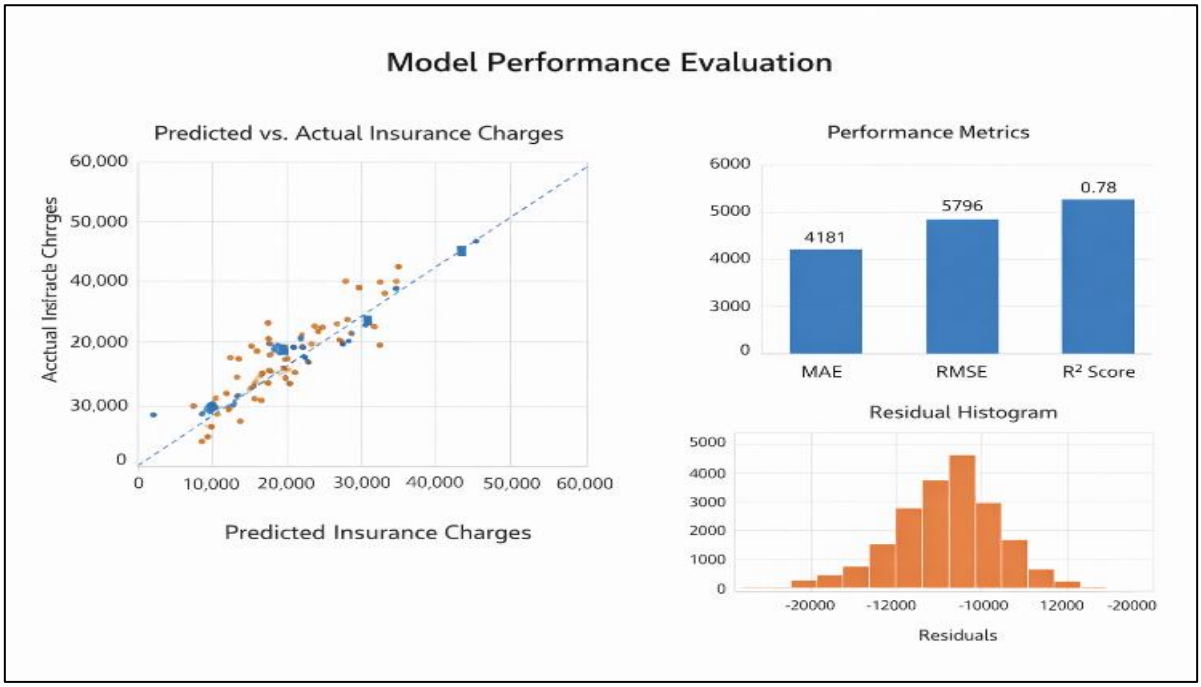

Fig 3 Model Performance Evaluation

➢ *Discussion*

The experimental results show that machine learning techniques can effectively predict insurance charges by analyzing customer data and identifying patterns between different variables. Important factors such as smoking status, BMI, and age significantly influence insurance costs. Among these factors, smoking status has the strongest impact, as smokers generally have higher insurance charges compared to non-smokers. Additionally, age and BMI moderately affect the insurance premium, indicating that health and lifestyle factors play an important role in cost prediction.

The Random Forest algorithm performed better than Linear Regression because it combines multiple decision trees and reduces overfitting, resulting in more accurate predictions. Overall, the system demonstrates that data science and machine learning can help insurance companies improve cost prediction, risk assessment, and pricing strategies.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1803

IX. CONCLUSION

This project demonstrates the application of machine learning techniques in predicting insurance charges using customer demographic and health-related attributes. The Linear Regression model was able to capture relationships between input variables and insurance costs effectively. The developed system assists insurance companies in making accurate pricing decisions and improves operational efficiency through automated predictions. By leveraging data science techniques, insurers can provide more personalized insurance plans while maintaining financial stability.

REFERENCES

[1]. Brati, E., Braimllari, A., & Gjeçi, A. Machine Learning Applications for Predicting High-Cost Claims Using Insurance Data. MDPI Data Journal, 2025.

[2]. Kulkarni, M., et al. Medical Insurance Cost Prediction Using Machine Learning. IJRASET Journal.

[3]. AbdElminaam, D., et al. An Efficient Framework for Predicting Medical Insurance Prices Using Machine Learning.

[4]. Zanke, P., Raparthi, M. Predictive Modelling for Insurance Pricing Using Machine Learning.

[5]. Kshirsagar, R., et al. Machine Learning Regression Framework for Predicting Health Insurance Premiums.

[6]. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

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

[8]. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann, 2011.

[9]. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.

[10]. P. Domingos, "A Few Useful Things to Know About Machine Learning," Communications of the ACM, vol. 55, no. 10, pp. 78–87, 2012.

[11]. S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," Informatica, vol. 31, pp. 249–268, 2007.

[12]. C. M. Bishop, *Pattern Recognition and* Machine Learning, Springer, 2006.

[13]. J. Friedman, T. Hastie, and R. Tibshirani, The Elements of Statistical Learning, Springer, 2009.

[14]. D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2017.

[15]. A. Géron, Hands-On Machine Learning with Scikit-Learn and TensorFlow, O'Reilly Media, *2017.*

[16]. W. McKinney, "Data Structures for Statistical Computing in Python," Proceedings of the Python in Science Conference, *2010.*

[17]. J. D. Hunter, "Matplotlib: A 2D Graphics Environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.

[18]. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[19]. T. Mitchell, Machine Learning, McGraw-Hill, 1997.

[20]. S. Raschka and V. Mirjalili, Python Machine Learning, Packt Publishing, 2017.

[21]. J. Brownlee, Machine Learning Mastery with Python, Machine Learning Mastery, 2017.

[22]. V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.

[23]. R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, MIT Press, 2018.

[24]. F. Chollet, Deep Learning with Python, Manning Publications, 2017.

[25]. E. Alpaydin, Introduction to Machine Learning, MIT Press, 2014.

[26]. S. Haykin, Neural Networks and Learning Machines, Pearson Education, 2009.

[27]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[28]. R. B. Myerson, "Game Theory and Insurance Decisions," Journal of Economic Perspectives, vol. 7, no. 2, pp. 43–64, 1993.

[29]. S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.

[30]. OECD, "The Role of Big Data in Insurance," OECD Publishing, 2020.

[31]. D. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001.

[32]. G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, Springer, 2013.

[33]. J. Leskovec, A. Rajaraman, and J. Ullman, Mining of Massive Datasets, Cambridge University Press, 2014.

[34]. M. Kuhn and K. Johnson, Applied Predictive Modeling, Springer, 2013.

[35]. T. Hastie, R. Tibshirani, and J. Friedman, Statistical Learning with Sparsity, CRC Press, 2015.

[36]. A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed., O'Reilly Media, 2019.

[37]. J. VanderPlas, Python *Data* Science Handbook, O'Reilly Media, 2016.

[38]. R. Elmasri and S. Navathe, Fundamentals of Database Systems, Pearson, 2016.

[39]. S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, Pearson, 2010.

[40]. C. Aggarwal, Data Mining: The Textbook, Springer, 2015.