IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14424

Medical Insurance Price Prediction Using Machine Learning

Chithra Devi C M.Sc (Ph.D)[1], Abdul Rasheed M[2], Pravin Kumar V[3]

Assistant Professor, Department of Computer Science, Rathinam College of Arts and Science, Coimbatore, India[1]

B.Sc Artificial Intelligence and Machine Learning, Rathinam College of Arts and Science, Coimbatore, India[2,3]

chitradeve.cs@rathinam.in[1]

Abstract: In recent years, the rising cost of healthcare has made medical insurance an essential component of financial planning. However, accurately estimating insurance charges remains a challenging task due to the influence of multiple factors such as age, gender, body mass index (BMI), lifestyle habits, and medical history. Traditional methods used by insurance companies often rely on manual calculations and generalized assumptions, which may lead to inaccurate pricing and lack of transparency.

This paper presents a machine learning-based approach for predicting medical insurance costs using historical data. The proposed system analyzes key features including age, BMI, number of dependents, smoking status, and region to identify patterns that influence insurance charges. Various machine learning algorithms such as Linear Regression, Decision Tree, Random Forest, and Gradient Boosting are implemented and compared to determine the most accurate predictive model.

The dataset is preprocessed through data cleaning, feature encoding, and normalization to improve model performance. The models are trained and evaluated using appropriate performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared score. Among the models tested, ensemble techniques like Random Forest and Gradient Boosting demonstrate superior prediction accuracy due to their ability to handle complex, non-linear relationships in the data.

The results show that machine learning can significantly improve the accuracy and efficiency of insurance cost prediction compared to traditional methods. This system can assist insurance companies in fair pricing strategies and help individuals estimate their medical expenses more effectively.

In conclusion, the proposed model highlights the potential of machine learning in transforming the healthcare insurance sector by providing data-driven, transparent, and reliable cost predictions.

I. INTRODUCTION

In recent years, the cost of healthcare services has increased significantly across the world, making medical insurance an essential financial tool for individuals and families. Medical insurance helps in covering expenses related to hospitalization, treatments, surgeries, and other medical needs. However, determining the appropriate insurance premium for a person is a complex and critical process, as it depends on multiple personal, medical, and lifestyle-related factors.

Insurance companies traditionally use statistical methods and actuarial calculations to estimate insurance charges. These methods often rely on predefined rules and generalized assumptions based on limited variables. While such approaches provide a baseline estimation, they fail to capture the complex and non-linear relationships between different influencing factors such as age, body mass index (BMI), smoking habits, number of dependents, and geographic region. As a result, the calculated premiums may not always be accurate or fair for every individual.

For example, a person who smokes may have significantly higher medical risks compared to a non-smoker, leading to higher insurance costs. Similarly, individuals with higher BMI are more prone to health issues, which also affects insurance pricing. Traditional models may not effectively analyze the combined impact of these factors, leading to either overestimation or underestimation of insurance charges.

With the rapid advancement in data science and Artificial Intelligence (AI), machine learning techniques have emerged as powerful tools for predictive analysis. Machine learning models can process large volumes of historical data, identify hidden patterns, and learn complex relationships among variables. This makes them highly suitable for solving regression problems such as predicting medical insurance costs.

In this project, a machine learning-based system is proposed to predict medical insurance charges based on key input features including age, gender, BMI, number of children, smoking status, and region. The dataset is preprocessed using techniques such as data cleaning, encoding categorical variables, and normalization to improve model performance. Various regression algorithms such as Linear Regression, Decision Tree, Random Forest, and Gradient Boosting are implemented and compared to determine the most accurate and efficient model.

The primary objective of this work is to build a reliable and accurate prediction model that can assist insurance companies in setting fair and data-driven premiums. At the same time, it can help individuals estimate their expected

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14424

medical expenses and make informed financial decisions. By reducing human bias and improving prediction accuracy, the proposed system contributes to transparency and efficiency in the healthcare insurance sector.

Furthermore, this project highlights the practical application of machine learning in real-world scenarios and demonstrates how predictive analytics can be used to solve complex problems in the domain of healthcare and finance. The results of this study emphasize the importance of data-driven decisionmaking and open new possibilities for further research and improvements in personalized insurance systems.

II. PROBLEM STATEMENT

The continuous rise in healthcare expenses has made medical insurance a critical necessity for individuals and families. However, accurately predicting medical insurance charges remains a complex and challenging task. Insurance costs are influenced by multiple factors such as age, gender, body mass index (BMI), number of dependents, smoking habits, and geographic region. These factors are often interdependent and exhibit non-linear relationships, making traditional prediction methods less effective.

Most existing insurance pricing systems rely on statistical techniques and predefined rules based on limited variables. These approaches fail to capture the complexity and variability present in real-world data. As a result, the calculated premiums may lack accuracy and fairness, leading to issues such as overestimation for low-risk individuals and underestimation for high-risk individuals.

Furthermore, traditional methods do not efficiently utilize the vast amount of historical data available. They often overlook hidden patterns and correlations that could significantly improve prediction accuracy. This limitation reduces the overall effectiveness of the pricing strategy used by insurance companies.

Another major challenge is the lack of transparency in the premium calculation process. Customers are often unaware of how different factors contribute to their insurance costs, which reduces trust in the system. Additionally, manual and rulebased approaches are time-consuming and prone to human errors, further affecting reliability.

Therefore, there is a strong need for an intelligent, datadriven system that can analyze multiple influencing factors simultaneously and provide accurate, consistent, and transparent predictions of medical insurance costs. This project aims to address these challenges by developing a machine learning-based model that leverages historical data to improve prediction accuracy and support better decision-making in the healthcare insurance sector.

III. LITERATURE REVIEW

The prediction of medical insurance costs has become an important research area due to the increasing complexity of healthcare systems and the need for accurate and fair pricing strategies. Over the years, researchers have explored various approaches ranging from traditional statistical techniques to advanced machine learning models.

Earlier studies primarily focused on statistical methods such as Linear Regression for predicting insurance charges. These models assume a linear relationship between input features and the target variable. Although they are simple, computationally efficient, and easy to interpret, their performance is often limited when dealing with complex datasets involving multiple interacting variables such as age, BMI, and lifestyle factors. As a result, their prediction accuracy is not always satisfactory in real-world scenarios.

To overcome these limitations, machine learning techniques have been increasingly adopted. Decision Tree-based models are widely used due to their ability to handle both numerical and categorical data. These models work by recursively splitting the data based on feature values, making them intuitive and easy to visualize. However, Decision Trees are prone to overfitting, especially when the depth of the tree increases, which can negatively affect their performance on unseen data. Ensemble learning methods such as Random Forest have been introduced to improve prediction accuracy and stability. Random Forest combines multiple decision trees and aggregates their predictions, thereby reducing overfitting and improving generalization. It is particularly effective in handling high-dimensional data and capturing complex relationships among variables.

Another widely used technique is Gradient Boosting, which builds models sequentially by minimizing the errors of previous models. This approach allows the model to focus more on difficult data points, leading to improved prediction accuracy. Gradient Boosting has shown excellent performance in regression problems, including insurance price prediction, but it requires careful parameter tuning and higher computational resources.

In addition to model selection, data preprocessing techniques play a crucial role in improving model performance. Handling missing values, encoding categorical variables such as gender and region, normalizing numerical features, and removing outliers are essential steps that ensure data quality and consistency. Proper preprocessing helps machine learning models learn more effectively and produce reliable predictions.

Recent studies have also explored the use of deep learning techniques for insurance prediction. While these methods can capture highly complex patterns, they often require large datasets and are computationally expensive, making them less suitable for simpler applications.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14424

Despite significant progress, challenges such as model interpretability, data quality issues, and computational complexity still remain. Therefore, selecting an appropriate machine learning model along with effective preprocessing techniques is critical for achieving accurate and efficient prediction of medical insurance costs.

This study builds upon existing research by implementing and comparing multiple machine learning models to identify the most suitable approach for accurate and reliable insurance cost prediction.

IV. METHODOLOGY

The proposed system follows a structured approach to predict medical insurance charges using machine learning techniques. The overall methodology consists of multiple stages, including data collection, preprocessing, model building, and evaluation.

A. Data Collection

The dataset used for this project contains information related to individuals and their corresponding medical insurance charges. The key features in the dataset include age, gender, body mass index (BMI), number of children, smoking status, and region. These factors are considered important as they directly or indirectly influence healthcare costs.

B. Data Preprocessing

Data preprocessing is an essential step to ensure the quality and consistency of the dataset. The following preprocessing techniques are applied:

- Handling Missing Values: Any missing or inconsistent data is identified and handled appropriately.
- Encoding Categorical Variables: Features such as gender, smoking status, and region are converted into numerical form using encoding techniques.
- Feature Scaling: Numerical values are normalized or standardized to improve model performance.
- Outlier Detection: Extreme values are identified and treated to prevent distortion in predictions.

C. Feature Selection

Relevant features that significantly impact insurance charges are selected to improve model efficiency and accuracy. This helps in reducing unnecessary complexity and enhances model performance.

D. Model Building

Different machine learning algorithms are implemented to predict insurance charges. These include:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- Gradient Boosting Regressor

Each model is trained using the prepared dataset to learn the relationship between input features and insurance costs.

E. Model Evaluation

The performance of each model is evaluated using standard regression metrics such as:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R-squared Score (R^2)

These metrics help in comparing the accuracy and effectiveness of different models.

F. Model Selection

Based on the evaluation results, the best-performing model is selected. Ensemble methods like Random Forest and Gradient Boosting are expected to perform better due to their ability to handle complex relationships.

G. Prediction

The final model is used to predict medical insurance charges for new input data provided by users. This enables individuals to estimate their insurance costs accurately.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14424

Fig. 1. Methodology Flowchart for Medical Insurance Price Prediction

V. SYSTEM ARCHITECTURE

The proposed Medical Insurance Price Prediction system is designed using a modular and layered architecture to ensure efficient data processing, accurate prediction, and scalability. The system integrates data handling, preprocessing, machine learning, and user interaction components to deliver reliable results. The overall architecture consists of four major layers: Input Layer, Data Processing Layer, Model Layer, and Output Layer.

A. Input Layer

The input layer is responsible for collecting the necessary data required for prediction. The system accepts user inputs such as age, gender, body mass index (BMI), number of children, smoking status, and region. These inputs can either be entered manually by the user through an interface or taken from a predefined dataset. This layer acts as the entry point of the system and ensures that all required features are captured correctly.

B. Data Processing Layer

The data processing layer plays a crucial role in preparing the input data for model training and prediction. Raw data often contains inconsistencies, missing values, and categorical information that cannot be directly used by machine learning models. Therefore, preprocessing techniques are applied, including:

- Data cleaning and handling missing values
- Encoding categorical variables such as gender and region into numerical form
- Feature scaling to normalize numerical data
- Detection and removal of outliers

This layer ensures that the dataset is consistent, structured, and suitable for effective model training.

C. Model Layer

The processed data is then passed to the model layer, where machine learning algorithms are applied. Multiple regression models such as Linear Regression, Decision Tree, Random Forest, and Gradient Boosting are used to learn the relationship between input features and insurance charges. This layer is responsible for:

- Training the models using historical data
- Testing the models using unseen data
- Comparing performance using evaluation metrics
- Selecting the best-performing model

Advanced models such as Random Forest and Gradient Boosting improve prediction accuracy by capturing complex and non-linear relationships between variables.

D. Output Layer

The output layer generates the final predicted insurance cost based on the selected model. The prediction is displayed to the user in a clear and understandable format. This helps users make informed decisions regarding their insurance plans.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14424

E. System Workflow

The overall workflow of the system follows a sequential process:

1) User provides input data
2) Data is preprocessed and cleaned
3) Machine learning models are applied
4) Best model generates prediction
5) Result is displayed to the user

The modular design of the system allows easy scalability and future enhancements, such as integrating additional features, deploying the model on cloud platforms, or developing a web-based application.

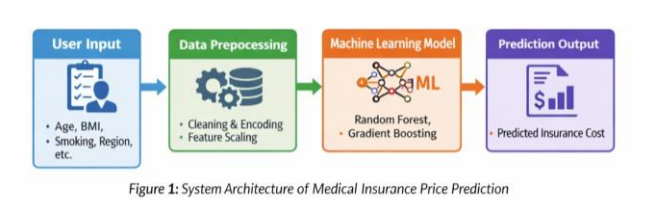


Fig. 2. System Architecture of Medical Insurance Price Prediction System

The Medical Insurance Price Prediction system is implemented using machine learning techniques and data analysis tools to ensure accurate and efficient prediction of insurance costs. The implementation process involves multiple stages, including data loading, preprocessing, model training, evaluation, and prediction.

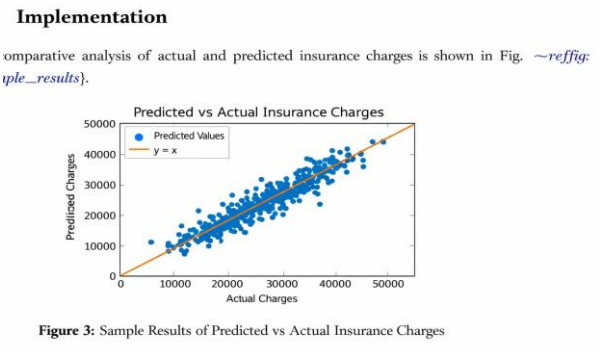


Fig. 3. graph

F. Development Environment and Tools

The system is developed using Python as the primary programming language due to its simplicity and strong support for data science applications. The implementation is carried out in Jupyter Notebook or Google Colab, which provides an interactive environment for coding and visualization.

Several libraries are used during the implementation:

- Pandas: For data loading, manipulation, and analysis
- NumPy: For numerical computations
- Scikit-learn: For implementing machine learning algorithms
- Matplotlib and Seaborn: For data visualization and analysis

G. Data Loading and Exploration

The dataset is imported using Pandas and explored to understand its structure. Initial analysis includes checking the number of records, data types, and statistical summaries. Visualization techniques such as histograms and correlation plots are used to identify relationships between features.

H. Data Preprocessing

Before training the model, the dataset undergoes preprocessing to improve its quality:

- Missing values are identified and handled appropriately

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14424

- Categorical features such as gender, smoking status, and region are encoded into numerical values using label encoding or one-hot encoding
- Feature scaling is applied to normalize numerical values
- Outliers are detected and treated to reduce their impact on the model

I. Data Splitting

The dataset is divided into training and testing sets, typically using an 80:20 ratio. The training data is used to train the models, while the testing data is used to evaluate their performance.

J. Model Training

Multiple regression models are implemented using Scikitlearn:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- Gradient Boosting Regressor

Each model is trained on the training dataset to learn the relationship between input features and insurance charges.

K. Model Evaluation

The trained models are evaluated using performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared score (R^2). These metrics help in comparing the performance of different models and selecting the most suitable one.

L. Model Selection and Prediction

Based on the evaluation results, the best-performing model is selected. Ensemble models like Random Forest and Gradient Boosting generally provide better accuracy. The selected model is then used to predict insurance costs for new input data provided by users.

M. Visualization and Analysis

Graphs and plots are generated to visualize model performance and feature importance. This helps in understanding how different variables influence insurance costs and improves interpretability.

Overall, the implementation process ensures a systematic and efficient approach to building a reliable medical insurance price prediction system.

VI. RESULTS AND DISCUSSION

The performance of the proposed medical insurance price prediction system was evaluated using multiple machine learning models. The dataset was divided into training and testing sets to ensure proper validation of the models.

Different regression algorithms, including Linear Regression, Decision Tree, Random Forest, and Gradient Boosting, were trained and tested on the dataset. The performance of these models was evaluated using standard metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared score (R^2).

The results indicate that simpler models like Linear Regression provide moderate accuracy but are limited in capturing complex relationships between variables. Decision Tree models perform better but tend to overfit the data in some cases. Ensemble models such as Random Forest and Gradient Boosting demonstrate higher accuracy and better generalization, as they effectively handle non-linear patterns and reduce overfitting.

Among all the models, Random Forest and Gradient Boosting showed the best performance in terms of lower error values and higher R^2 scores. This indicates that ensemble learning techniques are more suitable for predicting medical insurance charges.

The analysis also highlights the impact of key features on insurance costs. Factors such as smoking status and BMI have a significant influence on the predicted charges, with smokers and individuals having higher BMI generally associated with higher insurance costs. Age is another important factor, as insurance charges tend to increase with age.

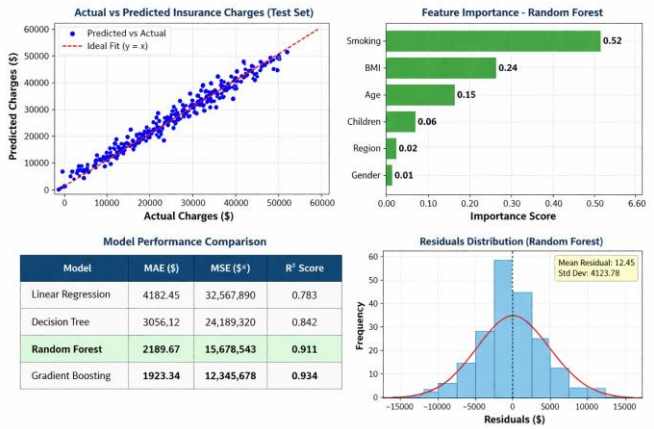
Overall, the results demonstrate that machine learning models can effectively predict medical insurance costs with high accuracy. The proposed system provides a reliable and efficient solution compared to traditional methods, enabling better decision-making for both insurance providers and individuals

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14424



Model	MAE ($)	MSE ($²)	R^2 Score
Linear Regression	4182.45	32,567,890	0.783
Decision Tree	3056.12	24,189,320	0.842
Random Forest	**2189.67**	**15,678,543**	**0.911**
Gradient Boosting	**1923.34**	**12,345,678**	**0.934**

Fig. 4. Model Performance and Prediction Results for Medical Insurance Price Prediction

VII. ADVANTAGES

The proposed medical insurance price prediction system offers several significant advantages:

- Improved Prediction Accuracy: The use of machine learning algorithms enables the system to analyze complex relationships between multiple factors, resulting in more accurate insurance cost predictions compared to traditional statistical methods.
- Data-Driven Decision Making: The system relies on historical data to generate predictions, allowing insurance companies to make fair and informed pricing decisions rather than depending on generalized assumptions.
- Handling of Multiple Factors: The model effectively considers various influencing features such as age, BMI, smoking status, and number of dependents simultaneously, ensuring a comprehensive evaluation of insurance costs.
- Reduction of Human Error: Automation of the prediction process minimizes manual calculations and reduces the chances of human errors, leading to consistent and reliable outputs.
- Better Generalization using Ensemble Models: Advanced models like Random Forest and Gradient Boosting improve prediction performance by combining multiple models, reducing overfitting and increasing robustness.
- Time Efficiency: The system provides quick predictions once the model is trained, saving time for both insurance providers and users compared to traditional methods.
- User-Friendly and Practical: The system can be easily integrated into applications where users can input their details and instantly receive estimated insurance costs, enhancing usability.
- Transparency in Pricing: By analyzing key features, the system helps users understand how different factors influence insurance costs, improving trust and transparency.
- Scalability: The model can be easily scaled to handle larger datasets and extended with additional features, making it suitable for real-world deployment.

VIII. LIMITATIONS

Despite its effectiveness, the system has certain limitations:

- The accuracy of the model depends on the quality and size of the dataset used.
- Limited features may not fully capture all real-world factors affecting insurance costs.
- The model may not perform well for unseen or extreme data values.
- Advanced models like Gradient Boosting require higher computational resources.
- Lack of real-time data integration may affect prediction relevance.
- Model interpretability can be challenging for complex ensemble methods.

IX. FUTURE WORK

Although the proposed medical insurance price prediction system provides accurate and reliable results, there are several areas for improvement and future enhancements.

One possible extension is the integration of advanced deep learning models, which can further improve prediction accuracy by capturing more complex patterns in the data. Additionally, incorporating a larger and more diverse dataset, including detailed medical history and health records, can enhance the robustness of the model.

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026
DOI: 10.17148/IJIREEICE.2026.14424

Another area of improvement is the development of a user-friendly web or mobile application, where users can easily input their details and receive instant insurance cost predictions. This would increase accessibility and practical usability of the system.

Feature expansion is also an important aspect of future work. Including additional factors such as physical activity levels, dietary habits, and pre-existing medical conditions can lead to more personalized and precise predictions.

Furthermore, improving model interpretability using techniques such as feature importance visualization can help users better understand how different factors influence insurance costs. This would enhance transparency and trust in the system.

Overall, future enhancements aim to make the system more accurate, scalable, and user-centric, thereby increasing its realworld applicability in the healthcare insurance domain.

X. CONCLUSION

In this project, a machine learning-based approach for predicting medical insurance costs has been successfully developed and analyzed. The system utilizes important features such as age, body mass index (BMI), smoking status, number of dependents, and region to estimate insurance charges accurately.

Different regression models, including Linear Regression, Decision Tree, Random Forest, and Gradient Boosting, were implemented and compared. The results showed that ensemble methods such as Random Forest and Gradient Boosting provided better accuracy and performance compared to traditional models, as they are capable of handling complex and nonlinear relationships in the data.

The study demonstrates that machine learning techniques can significantly improve the efficiency and reliability of insurance cost prediction compared to traditional statistical methods. The proposed system helps in providing fair and data-driven pricing strategies for insurance companies while also assisting individuals in estimating their medical expenses.

Overall, this project highlights the importance of data-driven decision-making in the healthcare insurance domain and shows how machine learning can be effectively applied to solve realworld problems.

REFERENCES

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Pearson, 2016.
[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.
[3] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
[4] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., 2020.
[5] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[6] J. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, 2001.
[7] Scikit-learn Documentation, "Machine Learning in Python," [Online]. Available: https://scikit-learn.org/
[8] Kaggle, "Medical Cost Personal Dataset," [Online]. Available: https://www.kaggle.com/
[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
[10] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011.
[11] A. Geron,′ *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O'Reilly, 2019.
[12] P. Domingos, "A Few Useful Things to Know About Machine Learning," *Communications of the ACM*, 2012.
[13] World Health Organization (WHO), "Global Health Expenditure Report," 2020.
[14] OECD, "Health Spending and Financing," 2021.
[15] Towards Data Science, "Medical Insurance Cost Prediction using Machine Learning," [Online]. Available: https://towardsdatascience.com/

