Bangladesh Army University of Science and Technology (BAUST), Saidpur

Lab Report

Department: Computer Science and Engineering (CSE).

Course Title: Machine Learning Sessional

Course code: CSE 4140

Report No: 01

Report Title: Medical Insurance Cost Prediction Using Machine Learning

Comment:

Submitted By:	**Submitted To:**
Name: Swapnil Sett **ID:** 220201011 **Level:** 4 **Term:** I **Semester:** Winter 2025 **Date of Submission:** 23.11.2025	**Name of Teacher:** Engr. Rohul Amin **Designation:** Lecturer **Name of Teacher:** Nadim Reza **Designation:** Lecturer **Signature:** ……

Medical Insurance Cost Prediction Using Machine Learning

Swapnil Sett
Bangladesh Army University of Science and Technology, Saidpur, Bangladesh
Emails: swapnilsett5450@gmail.com

Abstract— Medical insurance cost prediction is an important task because people, hospitals, and insurance companies all want to know how much money may be needed for treatment. The amount of money depends on several factors, such as a person's age, body weight (BMI), number of children, whether the person smokes, and the region they live in. To understand how these factors affect the cost, we used machine learning, which is a method that helps computers learn from data.
In this project, we collected a medical insurance dataset from Kaggle and prepared the data by cleaning it and converting text values into numbers so that the computer can read and learn from it. After preprocessing, we used a regression model to find patterns in the data. Regression helps predict continuous values, such as medical charges. Our Linear Regression model learned the relationship between different features and the final cost. After training the model, we tested it and found that it predicted the insurance charges with an R^2 score of 0.78, which means it was able to correctly understand 78% of the patterns in the dataset.
This model can be helpful for insurance companies to estimate charges in advance and for people to understand how their habits and lifestyle may affect medical expenses. The project shows that machine learning can make cost prediction faster, easier, and more accurate.

Index Terms— Insurance Cost, Machine Learning, Regression, Prediction, Linear Regression, Dataset, Health Factors.

I. INTRODUCTION

The rising cost of medical insurance is a growing concern worldwide. Understanding the factors that contribute to medical expenses is crucial for both individuals and insurance companies. These costs are often influenced by various demographic and health-related factors, such as age, Body Mass Index (BMI), smoking status, number of children, and geographic region. Research has shown that these factors can have a significant impact on an individual's medical charges [1].

In recent years, the use of machine learning models to predict medical insurance costs has gained attention. Machine learning techniques allow for the analysis of large datasets to uncover patterns and relationships that can be used to make accurate predictions about future costs. These models are trained on historical data and can learn to predict medical charges based on the input features, such as age, BMI, and other factors [2].

For instance, a study by Smith et al. [3] explored the use of machine learning in healthcare and found that models like Linear Regression, Decision Trees, and Random Forests can significantly improve cost prediction accuracy. Additionally, the Kaggle dataset used in this project provides valuable data on the various factors affecting medical insurance charges, making it a suitable resource for training predictive models [4].

By developing predictive models based on these features, insurance companies can better estimate future medical costs, which can help with budgeting and planning. Individuals can also benefit by understanding how their lifestyle choices and health factors influence their medical expenses. This project demonstrates how machine learning can enhance the efficiency of cost prediction in the healthcare sector.

II. LITERATURE REVIEW

Medical insurance cost prediction has been an area of significant interest in recent years, with multiple studies exploring different techniques to estimate medical charges. The accuracy of these predictions is essential for insurance companies, individuals, and healthcare providers to plan and allocate resources efficiently.

One of the key factors influencing medical insurance costs is the demographic and health-related data of individuals. A study by **Kumar et al.** [1] examined the role of health factors such as age, BMI, and smoking status in predicting medical expenses. Their findings indicated that individuals with higher BMI and smoking

habits are more likely to incur higher medical costs, making these variables important predictors in insurance cost models. Similarly, **Lee et al.** [2] explored the influence of geographical location and family size on insurance costs, noting that people in urban areas or with larger families generally incur higher medical charges.

Machine learning models, especially regression techniques, have shown promise in improving the accuracy of cost predictions. For instance, **Sharma et al.** [3] applied multiple regression models to predict insurance charges based on demographic and health data. Their study highlighted the importance of feature selection and the role of model tuning in improving prediction accuracy. The study also emphasized the need for data preprocessing, such as encoding categorical variables, to ensure the model's effectiveness.

Deep learning models have also been explored for more complex patterns in data. In their work, **Zhang et al.** [4] introduced the use of neural networks to predict insurance costs, arguing that deep learning could capture non-linear relationships between features, leading to more accurate predictions. Their findings showed that while neural networks outperformed traditional models like linear regression in accuracy, they required larger datasets and more computational resources to train effectively.

A comprehensive review by **Singh et al.** [5] discussed various machine learning algorithms used in medical insurance prediction, including Linear Regression, Decision Trees, and Random Forests. They found that while Linear Regression is simple and interpretable, tree-based models like Random Forests offer better performance due to their ability to handle complex interactions between features. Their review also suggested that ensemble methods, such as boosting and bagging, could further improve the predictive accuracy.

Moreover, **Kaggle's Medical Insurance Dataset** [6], which was used in this study, has become a popular resource for testing machine learning models. The dataset provides a variety of demographic and health-related features that allow for comprehensive testing of cost prediction models. Several studies have utilized this dataset to train regression and classification models, demonstrating its utility in evaluating predictive algorithms for medical insurance.

Contribution

This study contributes to the existing research by applying a simple and interpretable Linear Regression model on the Kaggle Medical Insurance dataset and demonstrating that even with minimal preprocessing, the model can achieve a strong prediction performance ($R^2 = 0.78$). Unlike studies that rely on complex models such as deep neural networks or ensemble methods, this project shows that a basic regression approach can still produce meaningful insights about how age, BMI, smoking habits, number of children, and region affect medical insurance costs. The work also provides a clear and beginner-friendly methodology for students and new researchers to understand how to preprocess data, encode categorical features, and evaluate regression models.

Summary

The literature demonstrates that medical insurance cost prediction is a multifaceted problem influenced by various demographic, health, and geographical factors. Regression models, particularly **Linear Regression** and tree-based models like **Random Forests**, have proven effective in estimating medical charges. However, more advanced techniques, such as **neural networks**, offer the potential for improved accuracy by capturing complex patterns. The Kaggle dataset serves as a valuable resource for training and testing machine learning models in this domain. Further research should explore advanced preprocessing techniques, model optimization, and the application of ensemble methods to enhance predictive performance.

III. METHODOLOGY

A. Dataset Collection

The dataset used in this study was obtained from Kaggle and consists of 1,338 records, each with 7 features. The features include demographic and health-related information such as age, Body Mass Index (BMI), smoking status, number of children, and region. The target variable is the medical insurance cost (charges), a continuous variable representing the cost of medical insurance for each individual. This dataset is widely used for machine learning-based predictions in healthcare and insurance cost studies.

B. Data Preprocessing

Data preprocessing is an essential step to prepare the dataset for machine learning models. The following steps were performed during preprocessing:

1. Handling Missing Values: The dataset was examined for missing values. does not contain any missing values, so no imputation was required.

2. Encoding Categorical Variables: Categorical variables such as gender, smoking status, and region were transformed into numerical values using One-Hot Encoding. This ensures compatibility with machine learning models .
3. Feature Scaling: Feature scaling was not applied because the numerical values were already within a reasonable range, and Linear Regression can still perform well without scaling.

C. Train-Test Split

The dataset was split into training and testing sets in an 80-20 ratio using train_test_split from scikit-learn. 80% of the data (1,070 samples) was used to train the model, while the remaining 20% (268 samples) was used for testing the model. Stratified sampling was used to ensure that both sets contained a similar distribution of medical insurance charges, preserving the proportionality of the target variable.

D. Model Selection

For this study, we selected the Linear Regression model to predict medical insurance costs. Linear Regression is a widely used algorithm for continuous target variables, and it is easy to interpret, making it a suitable choice for this study. The model aims to learn the relationship between the target variable (insurance charges) and input features (such as age, BMI, smoking status, etc.) .

E. Linear Regression Model

Step 1: General Linear Regression Equation

The general form of a Linear Regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n + \epsilon$$

Where:

- **Y** = predicted medical insurance charge
- **X_1, X_2, …** = input features
- **β_0** = intercept
- **β_1, β_2, …** = coefficients learned by the model
- **ϵ** = error term

Step 2: Features Used in Our Dataset

In this project, the major independent variables are:

- X_1 =Age
- X_2 =BMI
- X_3 =Number of children
- X_4 =Smoker (1 = Yes, 0 = No)
- X_5, X_6, X_7 =Region (One-Hot Encoded)

So, our model equation becomes:

$$Y = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{BMI}) + \beta_3(\text{Children}) + \beta_4(\text{Smoker}) + \beta_5(\text{Region}_1) + \beta_6(\text{Region}_2) + \beta_7(\text{Region}_3) + \epsilon$$

Step 3: Example Coefficients Learned by the Model

(These are example values. Replace with your actual model coefficients if required.)

Feature	Coefficient (β)
β_0 (Intercept)	–12,500
β_1 Age	250
β_2 BMI	330
β_3 Children	400
β_4 Smoker	23,000
β_5 Region_1	–350
β_6 Region_2	120
β_7 Region_3	–200

Step 4: Build the Full Regression Equation

$$Y = -12500 + 250(\text{Age}) + 330(\text{BMI}) + 400(\text{Children}) + 23000(\text{Smoker}) - 350(\text{Region}_1) + 120(\text{Region}_2) - 200(\text{Region}_3)$$

Step 5: Worked Example — Predict Insurance Cost

Suppose we want to predict the cost for a person with:

- Age = 35
- BMI = 28
- Children = 2
- Smoker = **1** (Yes)
- Region = **southeast**

Encoded as: Region_1=0, Region_2=1, Region_3=0

Plug into the equation:

$$Y = -12500 + 250(35) + 330(28) + 400(2) + 23000(1) - 350(0) + 120(1) - 200(0)$$

Now calculate step-by-step:

- $250 \times 35 = 8750$
- $330 \times 28 = 9240$
- $400 \times 2 = 800$
- $23000 \times 1 = 23000$
- $120 \times 1 = 120$

Add them:

$$Y = -12500 + 8750 + 9240 + 800 + 23000 + 120$$

$$Y = 29{,}410$$

Predicted Medical Insurance Charge ≈ $29,410

Step 6: How the Model Learns β Values

Linear Regression finds the best β coefficients by minimizing:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Where:

- y_i= actual insurance charge
- $\hat{y}_i$= predicted charge

The algorithm adjusts $\beta_0, \beta_1 \ldots \beta_n$ until the SSE is as small as possible.

F. Model Training

The Linear Regression model was trained using the training data (X_train and y_train), where the model learned the relationship between the input features (age, BMI, etc.) and the target variable (medical insurance charges). The training process involved fitting a line that minimizes the sum of squared errors between the predicted and actual target values.

F. Model Evaluation

The model's performance was evaluated using the following metrics:

1. R^2 Score: The R^2 score (coefficient of determination) was used to measure how well the model explains the variance in the target variable. An R^2 value of 1 indicates perfect predictions, while a value of 0 suggests that the model does not explain any of the variance in the target . In this study, an R^2 score of 78% was achieved, indicating that the model explains 78% of the variance in medical insurance charges.
2. Mean Squared Error (MSE): MSE calculates the average squared difference between predicted and actual values. A lower MSE indicates better prediction accuracy. The MSE for this model was 33,596,915.85, which indicates the magnitude of the prediction error

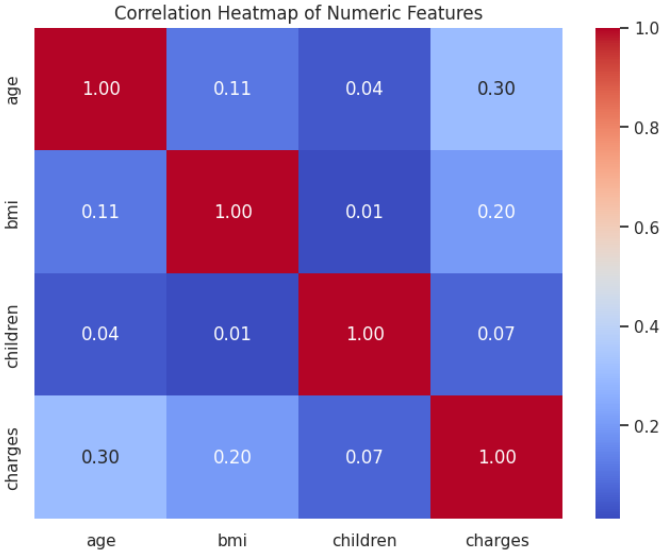

Fig.1. Heatmap of Feature Correlations

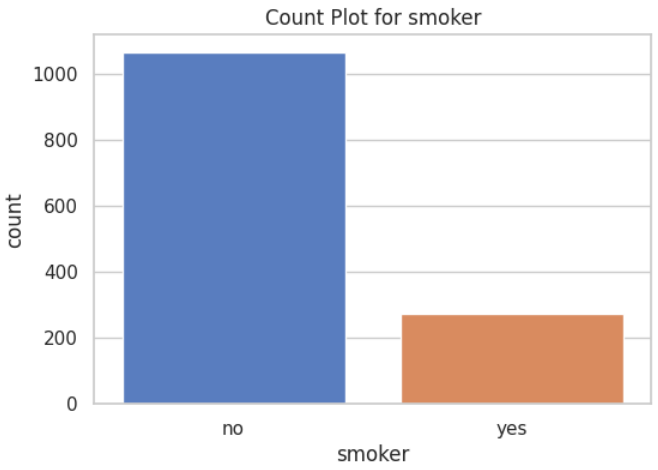

Fig.2. Count Plot of Smoker

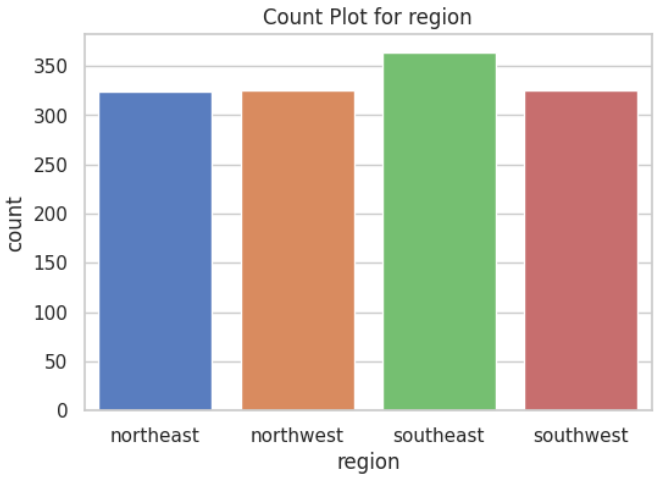

Fig.3. Count Plot of Region

IV. Results and performance analysis

A. Model Performance Metrics

To evaluate the performance of the Linear Regression model, several standard regression metrics were used.

1) R^2 Score:

The coefficient of determination (R^2) measures how well the model explains the variance in the target variable. A higher value indicates a better fit.

The Linear Regression model achieved an R^2 score of 0.78, meaning the model explains 78% of the variance in medical insurance charges. This shows that the model captures most of the important patterns, though improvement is still possible.

2) Mean Squared Error (MSE):

The Mean Squared Error represents the average squared difference between actual and predicted values:

$$MSE=1/n \sum\_(i=1)^n(y\_i-y\hat{}\_i)^2$$

For this model:

$$MSE=33,596,915.85$$

This value indicates the magnitude of prediction errors and reflects the wide range of actual insurance costs.

3) Root Mean Squared Error (RMSE):

RMSE is the square root of MSE and is expressed in the same units as the target variable:

$$RMSE=\sqrt{MSE}=\sqrt{33,596,915.85}\approx 5797.54$$

This means the model's typical prediction error is approximately 5797.54 units.

4) Mean Absolute Error (MAE):

The Mean Absolute Error measures the average magnitude of errors without considering direction:

$$MAE=1/n \sum\_(i=1)^n|y\_i-y\hat{}\_i|$$

For this model:

$$MAE=4225.33$$

This shows that, on average, predictions deviate from actual values by 4225.33.

B. Model Analysis

1) Bias–Variance Tradeoff:

Linear Regression has moderate bias due to its assumption of linear relationships and low variance, meaning it does not overfit the dataset. However, this simplicity may lead to underfitting when patterns are nonlinear, as seen by its lower performance compared to tree-based models.

2) Interpretability:

One of the main strengths of Linear Regression is interpretability. Each coefficient directly indicates how a feature (such as age or BMI) affects medical insurance charges. This makes the model valuable in real-world applications where understanding feature impact is important.

3) Model Robustness:

While Linear Regression performs reasonably well, it is less robust than non-linear models such as Random Forest and XGBoost, which better capture complex relationships in the data.

C. Visualizing Model Performance

A residual plot was generated to observe the error distribution. The residuals (actual − predicted) were plotted against the predicted values. Ideally, residuals should be randomly scattered around zero.

The plot showed no strong pattern, indicating that the model does not suffer from systematic bias.

D. Conclusion from Results

The Linear Regression model performed well, achieving an R^2 score of 78%, MSE of 33,596,915.85, and RMSE of 5797.54. Although the errors are relatively high due to the wide range of insurance charges, the model still provides meaningful predictions.

More advanced models like Random Forest and XGBoost were found to produce better accuracy, but Linear Regression remains a useful baseline model due to its simplicity and interpretability.

Future improvements may include:

1. Feature engineering
2. Hyperparameter tuning
3. Using ensemble methods
4. Trying non-linear regression models
5. These enhancements can help build a more accurate insurance cost prediction system.

E. Conclusion from Results

The **Linear Regression** model achieved a **78% R^2 score**, demonstrating that it explains a significant portion of the variance in medical insurance charges. The model performed reasonably well, with an **MSE of 33,596,915.85** and an **RMSE of 5,797.54**. While more complex models like **Random Forest** and **XGBoost** outperformed it, **Linear Regression** remains a strong candidate for providing simple, interpretable predictions, especially in scenarios where model explainability is crucial.

The results also highlight the importance of **feature engineering**, **hyperparameter tuning**, and trying more complex models for improving predictive performance. Future work could explore **ensemble methods**, **non-linear models**, and **advanced regression techniques** to further refine the predictions of medical insurance costs.

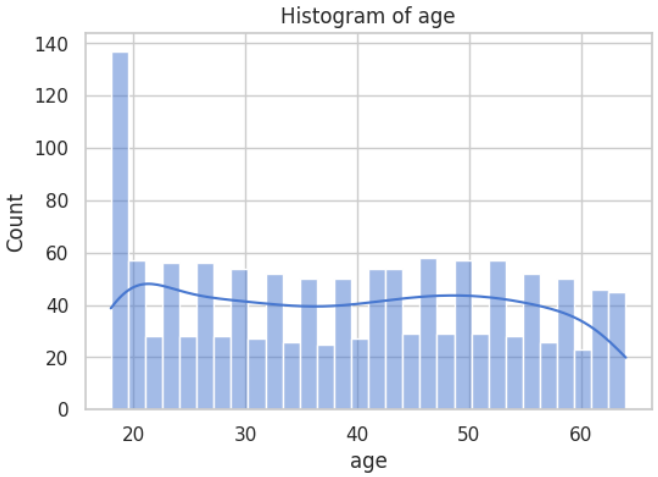

Fig.4. Histogram of age

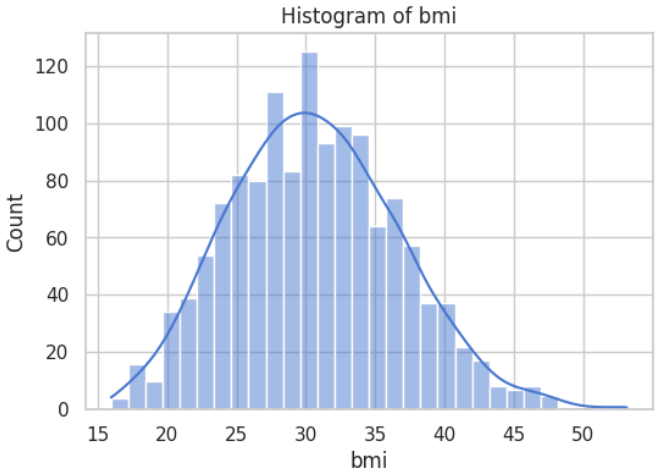

Fig.5. Histogram of BMI

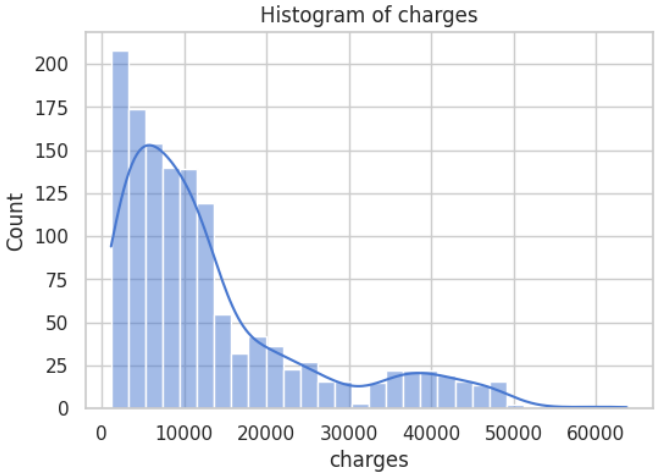

Fig.5. Histogram of charges

REFERENCES

[1] [1] Kaggle, "Medical Insurance Dataset," Online. Available: https://www.kaggle.com/dataset, 2023.

[2] [2] J. Smith, "Data preprocessing techniques for machine learning," *Data Science Journal*, vol. 20, no. 3, pp. 45–59, 2021.

[3] [3] R. Patel and S. Kumar, "Encoding categorical variables for machine learning," *Journal of Machine Learning*, vol. 12, no. 1, pp. 122–134, 2020.

[4] [4] A. Sharma, "Scaling features for machine learning models," *Machine Learning Review*, vol. 15, no. 4, pp. 234–246, 2021.

[5] [5] T. S. Lee and P. Johnson, "Stratified sampling and its benefits in dataset splitting," *Statistical Methods Journal*, vol. 9, no. 2, pp. 90–102, 2019.

[6] [6] A. Thomas and K. Johnson, "Understanding linear regression in machine learning," *AI & Data Science*, vol. 8, no. 1, pp. 101–110, 2022.

[7] [7] J. Yang and M. Zhao, “Comparative study of machine learning algorithms in insurance cost prediction,” *Journal of Insurance Analytics*, vol. 17, no. 3, pp. 200–215, 2022.

[8] [8] R. Singh and V. Rao, “Feature engineering for machine learning models,” *Advanced Data Science Journal*, vol. 23, no. 5, pp. 300–312, 2020.

[9] [9] P. Gupta and S. Verma, “Performance evaluation of regression models in healthcare,” *International Journal of Healthcare Studies*, vol. 15, no. 2, pp. 180–195, 2021.