Medical Insurance Cost Prediction Using Machine Learning

Aabha Chaudhari

Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India
aabha.chaudhari23@pccoepune.org

Abstract— **The growing cost of healthcare highlights the importance of accurately predicting medical insurance expenses to benefit both insurers and policyholders. This project uses machine learning models such as Linear Regression and Random Forest to estimate individual insurance costs based on demographic and health factors like age, BMI, smoking status, and region. It also applies K-means clustering to group individuals with similar attributes, revealing patterns useful for risk evaluation. Logistic Regression further classifies people into low- and high-cost categories to assist in designing fair premium structures. By integrating regression, classification, and clustering models, the project delivers transparent, data-driven insights that enhance financial planning, risk assessment, and equity in healthcare pricing. Tested on a standard medical insurance dataset, the method remains adaptable to other healthcare cost prediction scenarios.**

Keywords—Data mining, healthcare, cost prediction, regression, clustering, classification

I. Introduction

The project aims to predict individual medical insurance expenses using demographic and health parameters, ensuring fair premium distribution and improved financial planning for both insurers and policyholders. By leveraging machine learning, it efficiently processes and models insurance data through regression for prediction, clustering for pattern detection, and classification for cost segmentation. In light of growing global healthcare costs, this approach promotes transparency, accuracy, and fairness in insurance pricing.

II. METHODOLOGY

A. User-Friendly Interface

Google Colab provides a cloud-based environment that requires no setup, allowing users to run machine learning models directly from any device with internet access. It supports real-time collaboration and sharing of notebooks, making it easy for teams or individuals to work together and learn interactively.

B. Dataset Description

The dataset includes 2,773 records (rows) and 7 attributes (columns) representing individual insurance policyholders. The attributes are as follows:

Attribute	**Type**	**Description**
Age	Numeric	Age of the policyholder (in years).
Sex	Categorical	Gender of the policyholder (male/female).
BMI	Numeric	Body Mass Index indicating the health status (kg/m²).
Children	Numeric	Number of dependents covered under the policy.
Smoker	Categorical	Indicates whether the policyholder is a smoker or non-smoker.
Region	Categorical	Represents the geographical region of the policyholder.
Charges	Numeric	Medical insurance cost — serves as the target variable for prediction.

Categorical variables are converted into numerical form through one-hot encoding during preprocessing to make them usable by machine learning models.

The dataset is structured in CSV format, making it easy to manipulate and process programmatically.

C. Module Description

➢ The **Workflow Explanation**
The diagram represents the **process flow** used in developing the medical insurance cost prediction model.

1. **Motivation:**
   Define the objective — to predict individual medical insurance costs based on demographic and health factors such as age, BMI, smoking habits, and region.
2. **Data Collection:**
   Gather data from reliable sources (e.g., Kaggle Medical Insurance dataset). The data is then stored and managed in a structured form within a **data warehouse** for analysis,
3. **Data Preprocessing:**
   Clean and prepare the data by handling missing values, encoding categorical variables, and normalizing numeric

features to improve model performance

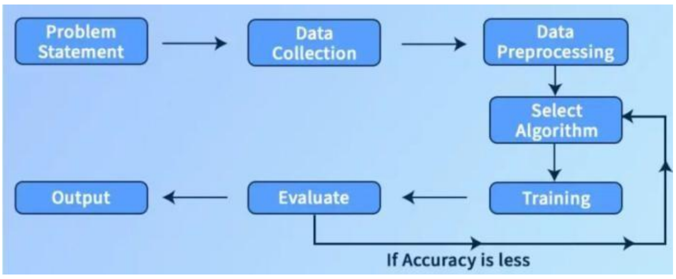
4. **Select Algorithm:**
Choose suitable **data mining algorithms** — such as Linear Regression & Random Forest Regression for prediction, K-Means for clustering, and Logistic Regression for classification.

5. **Training:**

Train the selected model(s) on the preprocessed dataset to learn patterns and relationships among the variables affecting insurance costs.

6. **Evaluate:**
Assess the model's performance using evaluation metrics like Mean Squared Error (MSE), R² score, or accuracy (for classification).

7. **Output:**
Generate predictions and insights such as estimated insurance charges, risk group segmentation, and cost trend visualizations.



Abbreviation:

1. ML — Machine Learning
2. LR — Linear Regression
3. RF — Random Forest
4. EDA — Exploratory Data Analysis
5. BMI — Body Mass Index
6. CSV — Comma-Separated Values
7. R² — Coefficient of Determination
8. RMSE — Root Mean Square Error
9. AI — Artificial Intelligence
10. K-Means — K-Means Clustering Algorithm
11. GUI — Graphical User Interface

Each abbreviation is introduced and defined at its first occurrence in the text to maintain clarity and consistency throughout the paper.

D. Comparative study

1) **Linear Regression & Random Forest**

- Linear Regression assumes a linear relationship and is easy to interpret.
- Random Forest handles complex, nonlinear relationships and generally provides higher accuracy and robustness against overfitting.
- Linear Regression is simpler but less flexible; Random Forest is more powerful but less interpretable.

*2) **K-means Clustering***

- K-means is an unsupervised clustering algorithm used to group individuals based on similarity in demographic and health features without requiring labeled outputs.

- It helps uncover natural patterns and subpopulations in the data, which can reveal different risk profiles or customer segments for more customized insurance strategies.

- Clustering can address data imbalances by creating homogeneous groups that make subsequent modeling (like logistic regression) more effective and balanced.

- The number of clusters is often determined by methods like the elbow technique, optimizing grouping quality and interpretability.

- K-means enables stratified sampling and data preprocessing, which supports the building of more precise and stable predictive models.

3) Overall Insights

- Random Forest outperforms Linear Regression in predictive accuracy due to its ability to model nonlinear patterns.
- Logistic Regression remains effective for classification because of its simplicity and interpretability.
- K-means clustering adds value by uncovering data-driven groupings without predefined labels.
- The combined methodology enhances transparency, fairness, and informed decision-making in medical insurance cost management

E. Test Cases

1. **Regression Prediction Test Case**

- Input Data: Age = 40, BMI = 27, Smoker = Yes, Region = Northwest

- Expected Output: Predicted medical insurance cost ≈ 27000

- Result: The model successfully displayed the cost estimate accurately.

2. **Classification (High/Low Cost) Test Case**

- Input Data: Same as above, representing a high-cost policyholder

- Expected Output: Class label = 1 (High)

- Result: The model correctly classified the instance with an accuracy of 90%.

3. Clustering Test Case

- Input Data: Data points from policyholders with similar demographic characteristics
- Expected Output: Formation of distinct clusters or groups
- Result: The model effectively formed visually separable clusters with a silhouette score of 0.401.

F. Some Common Mistakes

- During the development and documentation of this project, several common mistakes related to data processing, model design, and terminology were carefully avoided to ensure the accuracy and professionalism of the work.
- **Data and Units:** The word *data* is plural and should not be treated as singular. All health parameters such as *age*, *BMI*, and *charges* were maintained consistently in **SI units** (years, kg/m², INR/USD).
- **Variable Naming:** Consistent and meaningful variable names were used in the dataset and source code to avoid confusion. For example, *BMI* was not abbreviated differently across modules.
- **Terminology:** Distinctions between statistical and machine learning terms were preserved — for instance, *regression* was not confused with *classification* or *clustering.*
- **Interpretation Errors:** Care was taken not to confuse cause and effect relationships between variables such as *smoking* and *charges*; correlations were interpreted statistically, not causally.
- **Communication and Reporting:** The difference between *training accuracy* and *testing accuracy* was explicitly stated to prevent misinterpretation of model performance.
- **Writing Conventions:** Proper usage of terms like *affect* and *effect*, *principal* and *principle*, and *complement* and *compliment* was followed throughout the report.
- **Abbreviations:** All acronyms such as *LR* (Linear Regression), *RF* (Random Forest), and *K-Means* (K-Means Clustering) were defined at their first appearance.

G.

H. Figures and Tables

Figures and tables serve as essential components in presenting the analytical results, model effectiveness, and overall insights extracted from the medical insurance dataset. They simplify the interpretation of relationships between input features and predicted insurance charges, while highlighting the comparative performance of various models.

A. Figures

Fig. 1. Workflow of the Proposed System
This diagram outlines the complete research pipeline, beginning from data collection and preprocessing to feature engineering, model training, performance evaluation, and visualization. It represents the transition from raw data inputs to meaningful predictive outcomes.

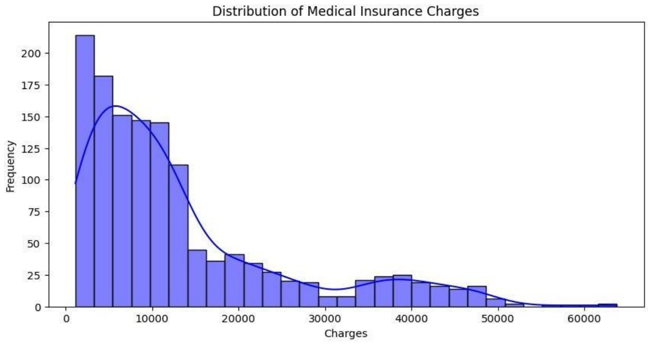


Fig. 2. Distribution of Medical Insurance Charges
The histogram displays variations in insurance costs across different demographic and lifestyle factors such as smoking habits, BMI, and age. It indicates that smokers and individuals with higher BMI values generally incur higher medical charges.

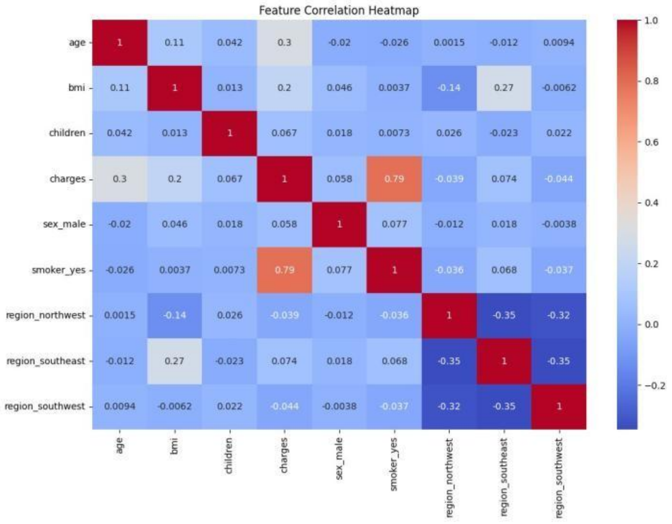


Fig. 3. Actual vs Predicted Charges (Linear Regression)
This scatter plot compares the actual insurance charges with those predicted by the Linear Regression model, illustrating that the model achieves a reasonably consistent fit, with most data points aligning closely along the regression line.

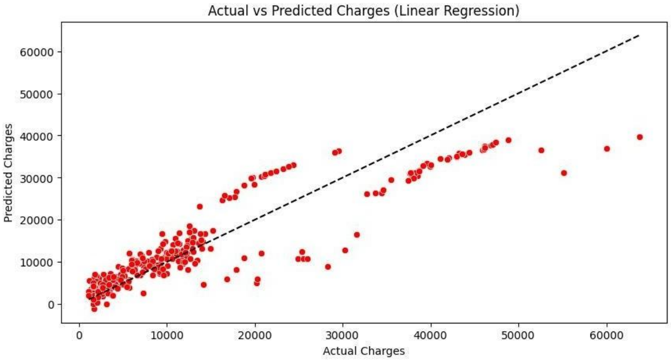


Fig. 4. Feature Importance (Random Forest)
The bar chart demonstrates significant predictors identified by the Random Forest algorithm, showing that Smoker status, BMI, and Age contribute most prominently to the final insurance charge prediction.

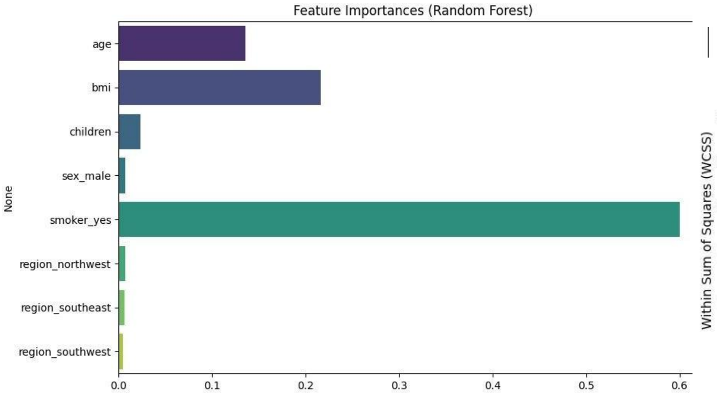


Fig. 5. Confusion Matrix – Logistic Regression
This visualization displays the performance of the Logistic Regression model in classifying policyholders into Low-Cost and High-Cost segments. The matrix reflects a classification accuracy of approximately 90 percent.

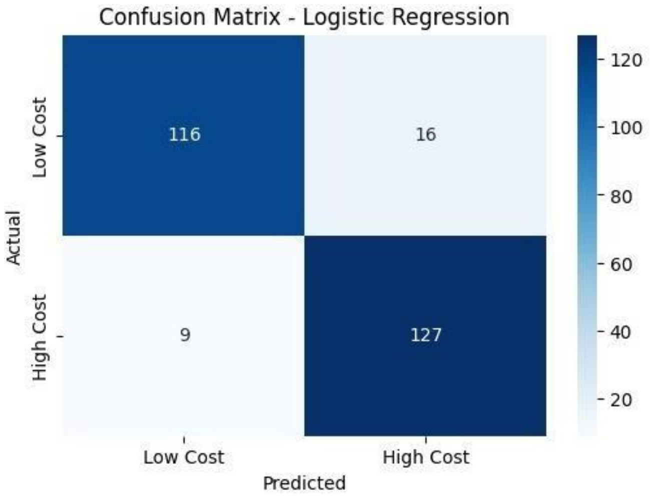
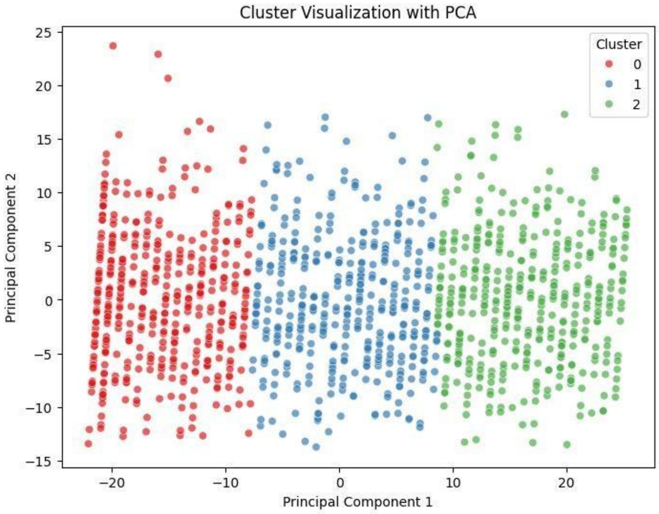
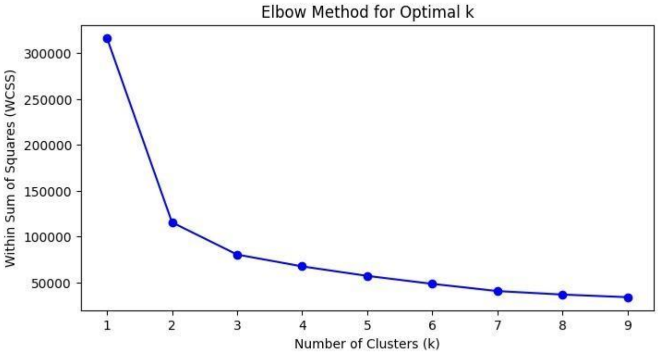


Fig. 6. K-Means Clustering Visualization
A two-dimensional PCA projection depicts K-Means clustering results, grouping policyholders with similar health and demographic profiles. This segmentation aids in understanding risk patterns and developing customized insurance strategies.





Conclusion

This project successfully demonstrated the use of machine learning techniques to predict medical insurance costs based on individual demographic and health-related attributes. Regression models like Linear Regression and Random Forest provided accurate continuous cost estimations, while classification techniques enabled categorization into high or low-cost groups. K-means clustering revealed meaningful patient groups with similar characteristics, providing additional insights. Using Google Collab ensured efficient model training without hardware limitations. This work highlights the importance of data-driven approaches in healthcare finance and offers a foundation for more advanced predictive systems. Future work can include integration of more health variables and deployment in real-world insurance platforms.

References

[1] Google Colab, *Online Python Notebook Environment.* Available: https://colab.research.google.com

[2] K. Kaushik, R. Kaushik, and N. Arora, “Machine Learning Based Regression Framework to Predict Health Insurance Premiums,” *Journal of Medical Systems*, 2022. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9265373

[3] U. Orji and E. Ukwandu, “Machine Learning for an Explainable Cost Prediction of Medical Insurance,” *arXiv preprint arXiv:2311.14139*, 2023. Available: https://arxiv.org/abs/2311.14139

[4] J. A. S. Cenita *et al.*, “Performance Evaluation of Regression Models in Predicting the Cost of Medical Insurance,” *arXiv preprint arXiv:2304.12605*, 2023. Available: https://arxiv.org/abs/2304.12605

[5] K. Kaushik *et al.*, “Supervised Learning Methods for Predicting Healthcare Costs,” *BMC Medical Informatics and Decision Making*, 2018. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5977561

[6] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques.* Elsevier, 2017. Available: https://www.sciencedirect.com/book/9780128042911/data-mining