



ISSN: 2321-2152

IJMECE

*International Journal of modern
electronics and communication engineering*

E-Mail

editor.ijmece@gmail.com

editor@ijmece.com

www.ijmece.com

MEDICAL INSURANCE COST PREDICTION USING MACHINE LEARNING

¹ MK Mehveen, ² Kavya Manchanpally, ³ Manasa SomeSETTY, ⁴ Harinitha G

¹ Associate professor in Department of Information Technology, Bhoj Reddy Engineering College for Women

^{2,3,4} UG Scholars in Department of Information Technology, Bhoj Reddy Engineering College for Women

Abstract

Medical expenses are among the most common recurring costs in an individual's life. A person's lifestyle and various physiological factors play a crucial role in determining the likelihood of developing certain diseases or disorders, which in turn influence their medical expenditures. Several studies have identified significant contributors to higher medical costs, including smoking habits, age, and Body Mass Index (BMI). The objective of this study is to analyze and establish the relationship between personal medical expenses and various influencing factors. Using significant attributes as predictors, we develop a **Linear Regression model** and compare its performance with a **Random Forest model** to forecast individual medical expenditures. The results of our research indicate that smoking, age, and higher BMI are strongly correlated with increased medical costs, making them key determinants of healthcare spending. The regression model achieved a prediction accuracy exceeding **75%**, demonstrating its effectiveness. According to the **World Health Organization (WHO)**, global personal healthcare spending is rising at a faster rate than overall economic growth. This surge is attributed primarily to factors such as smoking, ageing, and obesity. By utilizing health insurance datasets containing variables such as smoking status, age, number of dependents, residential region, and BMI, this study aims to uncover meaningful correlations between these factors and medical expenses, thereby contributing valuable insights into healthcare cost prediction.

Keywords: Machine Learning (ML), Random Forest, Health Insurance Prediction, Medical Factors, Linear Regression, Prediction Accuracy, Data Analysis, Healthcare Spending

I. INTRODUCTION

The cost of healthcare is increasing rapidly across the world. With the emergence of new diseases and viral infections, forecasting medical expenses has become an essential task. Accurate prediction of healthcare costs enables

governments and policymakers to make informed health-related decisions and assists individuals in understanding the significance of managing medical expenditures effectively. Machine Learning (ML) has become a transformative technology that influences almost every aspect of modern life,

including the healthcare sector. ML-based models are increasingly being applied to medical systems for diverse purposes such as disease diagnosis, treatment recommendations, and cost prediction. In this study, we focus on predictive analysis of **medical insurance expenses** using supervised learning techniques.

The primary objective of this work is to develop a model capable of forecasting an individual's medical insurance cost based on demographic and lifestyle factors such as **age, gender, BMI, smoking status, number of dependents, and residential region**. The dataset used for this study is obtained from **Kaggle**, containing **1,338 records** with the aforementioned attributes, along with the corresponding **insurance charges**. The data includes medical and financial information typically recorded by health insurance providers.

To achieve reliable prediction, several regression algorithms were implemented and compared to determine their performance in estimating insurance costs. The project was developed using the **Python programming language**, leveraging its extensive data analysis and machine learning libraries.

Medical insurance cost prediction plays a vital role in both the healthcare and insurance industries, contributing to **risk assessment, policy pricing, and financial planning**. Accurate cost estimation ensures fair premium

calculations for insurers and helps policyholders plan their finances more effectively. Traditionally, such estimations relied on **actuarial and statistical models**, which depend on predefined mathematical assumptions. However, with the rise of machine learning, more adaptive and data-driven models can now deliver higher accuracy and deeper insights into cost prediction.

II LITERATURE SURVEY

Prediction in medicine is not a new concept; however, with the advent of data-driven technologies, clinical predictions have become increasingly common and precise. Predictive models are now integral to several areas of healthcare, ranging from **risk categorization in critical care units** to determining the need for specific treatments such as **anticoagulant therapy (CHADS₂)** and **cholesterol medication (ASCVD)**. Systems such as **APACHE (Acute Physiology and Chronic Health Evaluation)** are widely used for assessing mortality risk in intensive care units.

The rapid advancement of **Machine Learning (ML)** and **Artificial Intelligence (AI)** has enabled the development of predictive models for a wide variety of clinical applications. Modern ML algorithms can process vast amounts of medical data to answer hundreds of clinical questions — from **early sepsis detection** to **diagnostic imaging** with near-human or even superhuman accuracy. These

intelligent systems enhance the ability of clinicians to make informed decisions and improve patient outcomes through data-driven insights.

However, the use of real-world medical data presents several challenges. Unlike traditional clinical datasets that are carefully curated to minimize bias, modern medical data sources often originate from **unstructured or heterogeneous systems**, such as electronic health records, billing data, and patient self-reports. These datasets may contain **incomplete information, inconsistent labeling, or biases** introduced through **patient self-selection, misinterpretation of medical indications, or disparities in data recording**. Consequently, such limitations can lead to **unintentional biases** or even **discriminatory outcomes** in predictive healthcare models.

Therefore, while the potential of data analysis and machine learning in clinical prediction is immense, it also necessitates **ethical and methodological vigilance** to ensure fairness, accuracy, and reliability.

In the broader context, identifying and managing **high-risk patients** remains a key research focus in predictive healthcare analytics. Governments, hospitals, and health insurance organizations employ distinct metrics to assess risk.

Hospitals typically focus on parameters such as **readmission rates, length of stay, and**

cumulative mortality risk, as accurate prediction of these metrics can help in **optimal resource allocation** and **enhanced operational efficiency**.

Health insurers, on the other hand, prioritize **insurance risk prediction**, as they must balance healthcare reimbursements with fixed premium incomes. Inaccurate or biased risk assessment models can lead to **financial losses** and **inequitable policy pricing**.

Overall, the literature indicates that predictive modeling in healthcare—powered by machine learning—plays a critical role in improving patient management, resource utilization, and financial sustainability. The ongoing challenge lies in developing **robust, unbiased, and interpretable models** that can be effectively deployed in real-world healthcare environments. Feature extraction applied to an image database is a **backend process** that operates independently of user interaction. In this process, distinctive characteristics or attributes are extracted from each image to represent its content in a more compact and computationally efficient form. These extracted features are significantly smaller in size compared to the original images and are stored in a **feature database** as numerical matrices or feature vectors, which are later used for **similarity measurement**.

The collection of all feature vectors constitutes the **feature database** for the image repository.

When a user submits a **query image** to retrieve visually similar images, the system enters the **frontend phase**. In this phase, features of the query image are extracted using the same feature extraction method employed for the database images, resulting in a corresponding **query feature vector**.

Subsequently, the **similarity between the query vector and the feature vectors in the database** is computed using predefined **distance metrics** (such as Euclidean, Manhattan, or cosine similarity). Based on the computed distances, the system retrieves and ranks the images with the **least distance values**, thereby providing a set of images most similar to the query image as the final output.

III. EXISTING SYSTEM

In today's world, medical expenses play a vital role in determining an individual's financial stability. The cost of healthcare treatments is often high, requiring people to rely on medical insurance to manage unforeseen medical expenditures. To address this, machine learning techniques have been increasingly utilized to predict medical insurance costs and treatment expenses based on various influencing parameters such as age, gender, smoking habits, BMI, and medical history.

In the existing system, **machine learning-based health insurance cost prediction** involves collecting historical data of policyholders, followed by **data**

preprocessing, feature extraction, and model training using algorithms such as **Linear Regression** and **Decision Trees**. These models analyze the relationship between patient attributes and their corresponding insurance claims to predict future costs or expected medical expenses. The system takes user inputs (e.g., personal and lifestyle details) and estimates both the **treatment cost** for specific diseases and the **insurance claim amount** that may be approved by the insurance provider.

However, the existing systems have several **limitations** that affect their performance, reliability, and practical applicability.

Problems in the Existing System

1. Data Privacy and Security Risks:

Health insurance datasets contain highly sensitive and confidential personal information. Improper handling or insufficient security measures can lead to data breaches or privacy violations.

2. Lack of Transparency and Interoperability:

Many machine learning models—especially complex ones like deep neural networks—function as “black boxes,” making it difficult to interpret how specific predictions are generated. This lack of transparency hinders trust and accountability in the prediction results.

3. Data Quality Issues:

Medical datasets often contain missing, inconsistent, or noisy data due to human error or incomplete records, leading to reduced model accuracy and biased predictions.

4. Complexity in Modeling:

Designing and tuning machine learning models for medical cost prediction requires domain expertise and careful parameter optimization, which can be complex and time-consuming.

5. High Computational Cost:

Training large-scale predictive models, especially on extensive insurance datasets, demands high computational power and processing resources.

6. Over-Reliance on Historical Data:

Most existing models depend heavily on historical data, which may not reflect recent trends, policy changes, or emerging diseases, thereby limiting prediction accuracy and adaptability.

IV. PROPOSED SYSTEM

The proposed system aims to develop an efficient and accurate **medical insurance cost prediction model** using machine learning techniques. The dataset used in this study consists of **1,338 records** containing both personal and medical information obtained

from a publicly available health insurance dataset. The attributes include **age, gender, BMI, illnesses, smoking status, and insurance charges**. Among these features, **insurance charges** serve as the **dependent variable**, while the remaining attributes act as **independent variables** that influence the prediction outcome.

In regression-based predictive modeling, the primary objective is to estimate the value of the dependent variable (insurance cost) based on the relationships identified with the independent variables. The system follows a structured methodology consisting of **data collection, preprocessing, feature encoding, model training, and evaluation**.

Data Preprocessing

Data preprocessing is a critical step to ensure data quality and model accuracy. The raw dataset may contain **missing values, inconsistent records, or categorical variables** that must be transformed into numerical form for machine learning algorithms to process effectively.

1. Handling Missing Values:

Missing data can negatively affect model performance and may lead to inaccurate predictions. Therefore, missing values are identified and removed or imputed to maintain dataset integrity.

2. Encoding Categorical Variables:

Since machine learning models work primarily with numerical data, categorical features such as **gender**, **smoker status**, and **region** are transformed using **Label Encoding** and **One-Hot Encoding** techniques. This conversion ensures that the categorical information is represented numerically without introducing bias or order dependency.

After preprocessing, the clean and encoded dataset is used to train regression models, which learn the correlation between various independent features and the target insurance charge. The trained model can then predict medical insurance costs for new user inputs with improved accuracy and efficiency.

V. METHODOLOGY

The foundation of the medical insurance cost prediction model lies in the quality and preparation of the dataset. The dataset used in this project was sourced from **Kaggle**, which contains **1,338 records** representing individual health and demographic information. The dataset includes the following key attributes:

- **Age:** Age of the individual (numeric)
- **Gender:** Male or Female (categorical)
- **BMI:** Body Mass Index, a measure of body fat (numeric)

- **Children:** Number of dependents (numeric)
- **Smoker:** Smoking status – Yes or No (categorical)
- **Region:** Residential region (categorical)
- **Charges:** Medical insurance cost (numeric, dependent variable)

5.1 Data Preprocessing

To ensure the dataset is clean and suitable for machine learning algorithms, several preprocessing steps were applied:

1. Handling Missing Values:

The dataset was examined for missing or null values. Any incomplete records were either removed or imputed to maintain data consistency and prevent training errors.

2. Encoding Categorical Variables:

Label Encoding was applied to binary categorical features such as *gender* and *smoker* to convert text labels into numeric form.

One-Hot Encoding was applied to multi-category features like *region* to represent each category as a binary vector without introducing ordinal relationships.

3. Feature Scaling:

Numerical attributes such as *age* and *BMI* were normalized or standardized to ensure all features contribute equally to the model and to improve convergence during training.

4. Feature Correlation Analysis:

A correlation heatmap was used to identify the relationship between features and the target variable (*charges*). Features such as **age**, **BMI**, and **smoker** showed a strong correlation with medical expenses and were prioritized for model training.

VI IMPLEMENTATION

The implementation phase involves dividing the system into several functional modules, each responsible for handling specific operations. A module represents an independent unit of a program, designed to perform a particular task efficiently. Each module communicates with others to ensure smooth data flow and system integration. The detailed implementation of the proposed system is organized into the following modules:

1. User Setup Module:

This module is responsible for collecting essential user information such as age, gender, Body Mass Index (BMI), number of children, smoking status, and region. The system validates the entered data to ensure accuracy and completeness,

detecting and correcting any missing or incorrectly formatted inputs. This ensures the integrity and consistency of the information used for further processing.

2. Dataset Collection Module:

This module focuses on acquiring a reliable and high-quality dataset containing relevant medical and demographic attributes. The dataset includes key features such as age, sex, region, BMI, and smoking status—critical for predicting medical insurance costs. During this stage, irrelevant or redundant data is removed to improve analysis efficiency and ensure that only meaningful information is retained. The refined dataset is then passed to the preprocessing module for further preparation.

3. Data Preprocessing and Feature Extraction Module:

Data preprocessing plays a vital role in ensuring the dataset is clean, standardized, and ready for model training. Feature selection techniques are employed to identify the most significant attributes, reducing dimensionality and improving the computational efficiency of the system. Standardization is applied to bring all feature values to a common scale, which enhances the performance and accuracy of the machine learning model used in prediction.

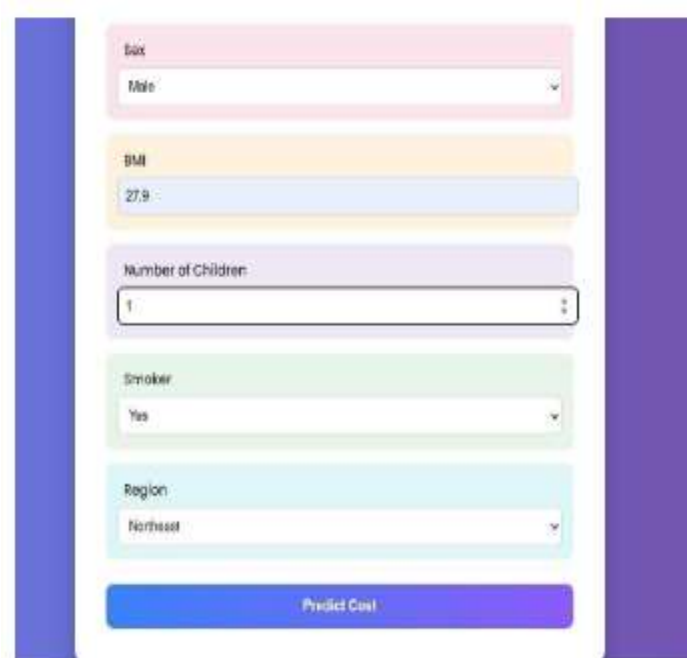
4. Prediction Output Display Module:

In this module, the system utilizes a pre-trained machine learning model to generate insurance cost predictions based on user-provided inputs such as age, gender, BMI, smoking status, number of dependents, and region. The model processes the input data and produces an estimated medical insurance cost. The output is displayed in a clear, user-friendly interface, allowing users to easily interpret the results and understand how their personal information influences the predicted cost.

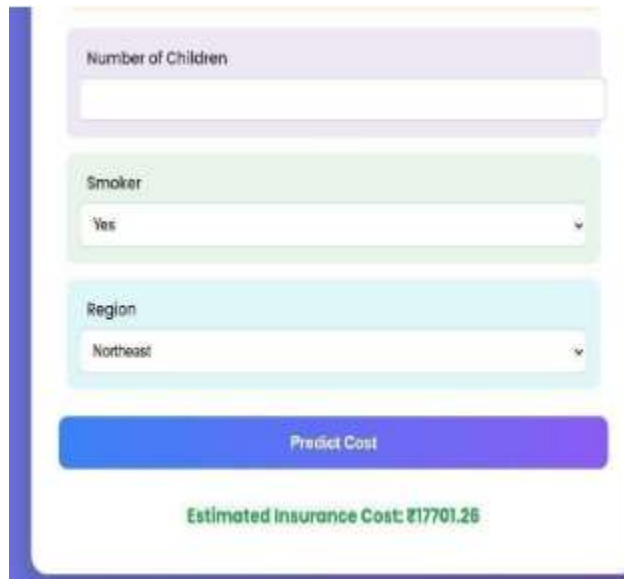


The image shows a web form titled "Insurance Cost Predictor" with a logo at the top. The form includes input fields for "Age", "Sex" (with a dropdown menu showing "Male"), and "BMI".

VII RESULTS

The image shows the same web form as before, but with additional input fields: "Number of Children" (with a dropdown menu showing "1"), "Smoker" (with a dropdown menu showing "Yes"), and "Region" (with a dropdown menu showing "Northeast"). A large blue "Predict Cost" button is at the bottom.



VIII CONCLUSION

Machine learning demonstrates remarkable potential in predicting health insurance costs by analyzing a diverse range of factors such as age, gender, medical history, lifestyle habits, and socio-economic conditions. Through advanced data-driven models, insurers can gain deeper insights into risk assessment, optimize premium pricing, and design personalized insurance plans tailored to individual needs. These predictive capabilities not only enhance decision-making for insurers but also empower policyholders with more transparent and fair pricing structures.

Moreover, machine learning plays a crucial role in detecting fraudulent claims by identifying unusual or inconsistent patterns within large datasets, helping insurers save significant financial resources. The integration of such models also facilitates the prediction of

healthcare trends, allowing organizations to make proactive and informed policy decisions.

Despite its advantages, certain challenges persist, including concerns related to data privacy, algorithmic bias, and the need for continuous model updates to maintain accuracy. However, with the adoption of explainable AI techniques and the inclusion of diverse and representative data sources, the trustworthiness and reliability of predictive models can be further strengthened. As technology continues to evolve, machine learning will remain a cornerstone in transforming the health insurance industry into a more efficient, transparent, and equitable system.

REFERENCES

- [1] Angela D. Kafuria, "Predictive Model for Computing Health Insurance Premium Rates Using Machine Learning Algorithms," *International Journal of Computer (IJC)*, vol. 44, no. 1, pp. 21–38, 2022.
- [2] Ch. Anwar ul Hassan, Jawaid Iqbal, Saddam Hussain, Hussain AlSalman, Mogeab A. A. Mosleh, and Syed Sajid Ullah, "A Computational Intelligence Approach for Predicting Medical Insurance Cost," *Mathematical Problems in Engineering*, vol. 2021, Article ID 1162553, 13 pages, 2021.
- [3] Natalya Shakhovska, Nataliia Melnykova, Valentyna Chopiyak, and Michal, "An

Ensemble Method for Medical Insurance Costs Prediction Task,” *Computers, Materials & Continua*, vol. 70, pp. 3969–3984, 2022.

[4] Mohamed Hanafy, “Predict Health Insurance Cost by Using Machine Learning and DNN Regression Models,” *International Journal of Innovative Technology and Exploring Engineering*, vol. 10, p. 137, 2021.

[5] G. Satya Mounika Kalyani and Rama Parvathy L., “A Novel Ranking Approach to Improved Health Insurance Cost Prediction by Comparing Linear Regression to Random Forest,” *Journal of Survey in Fisheries Sciences*, vol. 10, no. 1S, pp. 2030–2039, 2023.

[6] Kaggle Dataset – “Medical Cost Personal Dataset,” Available at:
<https://www.kaggle.com/datasets/mirichoi0218/insurance>.

[7] GitHub Repository – Available at:
<https://github.com/>.

[8] TutorialsPoint – Web Development Tutorials. Available at:
https://www.tutorialspoint.com/web_development_tutorials.html.

[9] W3Schools – Web Development Resources. Available at: <https://www.w3schools.com>.