



(RESEARCH ARTICLE)



Vital care insurance prediction using machine learning

Ashok Kumar Pasi, Lasya Palarapu, Akshitha Mailaram, Laxmi Prasanna Kanithi * and Deekshith Bommana

Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India.

World Journal of Advanced Research and Reviews, 2025, 25(02), 456-464

Publication history: Received on 25 December 2024; revised on 31 January 2025; accepted on 02 February 2025

Article DOI: <https://doi.org/10.30574/wjarr.2025.25.2.0368>

Abstract

The Vital Care Insurance Prediction System leverages machine learning, particularly linear regression, to estimate insurance costs based on user-specific attributes. It evaluates key factors such as age, gender, BMI, dependents, geographic region, medical risk, lifestyle, and occupation to enhance prediction accuracy. Unlike conventional actuarial models, this system provides dynamic forecasts and includes confidence metrics, ensuring greater transparency in cost estimation. The integration of machine learning enables a more adaptive and precise approach to risk assessment, improving the efficiency of insurance planning. A user-friendly Streamlit interface ensures accessibility, offering real-time results to both individuals and insurance professionals. The interactive "pop-up" feature enhances user engagement by presenting insights in a structured manner. This system bridges the gap between healthcare and finance, optimizing insurance decision-making processes. By increasing prediction accuracy and simplifying access to information, the system empowers users with data-driven insights, aiding them in making well-informed choices. This innovation ultimately enhances affordability and efficiency in the insurance sector, benefiting both providers and policyholders.

Keywords: Machine Learning; Personalized Insurance Recommendations; Real-Time Insurance Cost Prediction; User-Friendly Stream-lit Interface

1. Introduction

Vital Care Insurance plays a crucial role in safeguarding individuals against unexpected medical expenses by providing financial support for healthcare needs. Vital Care Insurance offers financial protection by covering medical expenses, ensuring individuals can access quality healthcare without excessive costs. With rising medical expenses, an efficient and accurate insurance system is essential for managing unexpected healthcare costs. Traditional insurance models use broad risk classifications, often leading to inconsistent pricing. Some individuals may pay higher premiums than necessary, while others might not receive adequate coverage. To overcome these challenges, machine learning is incorporated into insurance prediction models to improve accuracy and fairness.

The Vital Care Insurance Prediction System applies linear regression to assess various factors, including age, gender, BMI, dependents, lifestyle, occupation, and medical history. By analyzing this data, the system generates personalized cost estimates, moving away from one-size-fits-all pricing. Unlike traditional actuarial methods, this system provides real-time predictions and includes confidence metrics to enhance transparency. This ensures fair pricing based on actual health risks rather than generalized categories.

A user-friendly Stream-lit interface enhances accessibility, allowing users to input data and receive instant cost predictions. The inclusion of interactive pop-up insights helps individuals understand how different factors impact their insurance costs, making the system more engaging and informative. This system benefits both insurance providers and policyholders. Insurers can optimize pricing strategies while reducing financial risks, and customers receive accurate,

* Corresponding author: K Laxmi Prasanna

transparent, and affordable insurance options tailored to their needs. Additionally, healthcare institutions can use these insights for better financial planning and patient care management.

By integrating machine learning with risk assessment, the Vital Care Insurance Prediction System enhances efficiency, accuracy, and accessibility in the insurance sector. This approach leads to fairer pricing and better decision-making, making healthcare coverage more reliable and cost-effective for all.

2. Related Work

Research on health insurance cost prediction has evolved significantly with the integration of machine learning techniques. Traditional methods relied on actuarial models and statistical analysis to estimate insurance costs based on generalized risk factors. These approaches often lacked personalization and failed to accurately reflect an individual's specific health risks.

Recent studies have explored the use of machine learning algorithms such as linear regression, decision trees, and neural networks to enhance the accuracy of insurance cost prediction. These models analyze multiple factors, including age, gender, BMI, medical history, lifestyle habits, and occupation, to generate personalized cost estimates. By leveraging predictive analytics, these approaches help insurance providers create fair and competitive pricing strategies while improving policyholder satisfaction.

Several researchers have also worked on real-time insurance cost estimation using web-based interfaces. Platforms like Streamlit and Flask have been used to build interactive applications where users can input their personal details and receive instant predictions. The incorporation of visual analytics further enhances the user experience by displaying insights into how different factors impact premium calculations. Another area of related research focuses on transparency and explainability in insurance pricing. Studies emphasize the importance of confidence metrics and interpretability tools, ensuring that policyholders understand how their insurance costs are determined.

3. Existing System

Traditional medical insurance prediction systems rely on basic risk estimation models that assess a narrow range of factors, such as age, BMI, and smoking status. These models use predefined actuarial tables and fixed risk categories, which results in generalized cost estimations rather than personalized predictions. The scope of these models is limited, as they do not account for critical lifestyle details, medical history, or other personalized health information. As a result, these systems often fail to accurately represent an individual's health profile, leading to inefficiencies in pricing and risk assessment.

The existing system has several significant limitations. Firstly, it only analyzes a small set of variables, excluding more important information about a person's lifestyle and medical background. Secondly, the lack of advanced analytical techniques in these models leads to less precise predictions, meaning users may not receive accurate insurance cost estimates. Furthermore, the system does not adapt to real-time data, making it ineffective in addressing the dynamic nature of healthcare needs.

4. Proposed Model

The proposed Vital Care Insurance Prediction System, as outlined in the final review, addresses the limitations of traditional medical insurance models by incorporating advanced machine learning techniques. Unlike conventional systems that only consider a limited set of factors, this model analyzes a broader range of user attributes, including medical history, lifestyle choices, and other health risks. By leveraging predictive analytics, such as linear regression, the system can provide more accurate and personalized insurance cost estimations.

Additionally, the model features interactive user interfaces that enable real-time data integration, allowing for adaptive learning based on evolving healthcare needs. This dynamic approach ensures that insurance predictions are not only more precise but also responsive to changes in the user's health profile. Overall, the Vital Care Insurance Prediction System enhances transparency and fairness, offering users deeper insights into the factors affecting their insurance costs, thereby improving decision-making for both policyholders and insurers.

5. Methodology

The development of this project, "Prediction of Vital Care Insurance Using Machine Learning Techniques," follows a structured approach to ensure accuracy, efficiency, and user-friendliness. The methodology consists of multiple stages, including data collection, preprocessing, visualization, model building, evaluation, and prediction.

5.1. Data Collection

A dataset containing key attributes such as age, sex, BMI, number of children, smoker status, region, medical history, lifestyle, occupation, and insurance charges was sourced from Kaggle. This dataset serves as the foundation for training and testing the machine learning model.

5.2. Data Preprocessing

To enhance the quality and reliability of the dataset, several preprocessing steps were applied:

Handling missing values to ensure data completeness.

Normalizing numerical values to standardize input features.

Encoding categorical variables, such as smoker status and region, into numerical representations.

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn import metrics
import matplotlib.pyplot as plt
import seaborn as sns
```

Figure 1 Libraries imported

5.3. Data Visualization

Visualization techniques were used to explore and analyze the dataset effectively. Charts and graphs were employed to identify patterns, correlations, and trends among different features. This step helps in understanding the impact of various attributes on insurance costs.

5.4. Model Development

A machine learning model, specifically Linear Regression, was implemented to predict insurance costs based on input attributes. The dataset was divided into training and testing sets in an 80:20 ratio to evaluate the model's performance accurately.

5.5. Model Training and Evaluation

The model was trained using the training dataset, and its performance was assessed using the testing dataset. Various evaluation metrics, such as Mean Squared Error (MSE) and R-squared score, were used to measure the accuracy and reliability of the predictions.

5.6. Prediction

Once trained, the model was used to predict insurance costs for new user inputs. Users provide details such as age, BMI, and smoking status, and the model generates an estimated insurance cost based on learned patterns.

5.7. Future Enhancements

To improve accuracy and functionality, future work will focus on:

- Incorporating additional predictive factors such as genetic history and environmental conditions.
- Experimenting with advanced machine learning algorithms like Random Forest and Neural Networks.
- Enabling real-time predictions through an interactive API and dashboard.

6. System Architecture

The architecture of the Vital Care Insurance Prediction System is designed to handle user inputs, process data efficiently, and provide accurate insurance cost predictions. It consists of three key layers:

- **Data Layer:** This layer is responsible for storing and managing the dataset. It includes:
 - *Data Collection:* The system retrieves insurance-related data, including age, BMI, smoker status, medical history, and lifestyle factors.
- **b. Processing Layer:** This is the core computational component where machine learning techniques are applied. It consists of:
 - *Data Preprocessing Module:* Handles missing values, normalizes numerical data, and encodes categorical features.
 - *Machine Learning Model:* Implements Linear Regression for predicting insurance costs.
 - *Evaluation & Optimization:* Ensures the model performs well using metrics such as Mean Squared Error and R-squared.
- **c. Application Layer:** This layer enables interaction between the system and users:
 - *User Interface:* Allows users to input personal details such as age, BMI, and lifestyle habits.
 - *Prediction Output:* Displays estimated insurance costs and insights through an intuitive interface.
 - *Visualization Tools:* Provides graphical representation of data trends to help users understand cost predictions.

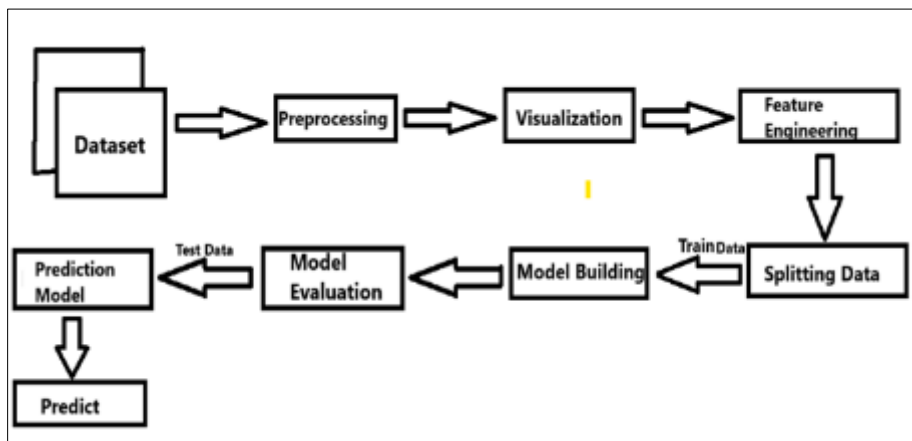


Figure 2 System Architecture

It illustrates how input data flows through various processing layers, resulting in meaningful outputs.

6.1. Algorithm Used

The algorithm is responsible for learning and predicting insurance.

6.1.1. Linear Regression

Linea Linear regression is a fundamental statistical technique used to understand the relationship between a dependent variable and one or more independent variables. It is commonly applied in predictive modeling, trend analysis, and forecasting. The primary goal of linear regression is to find a line that best fits the data by minimizing the differences between actual and predicted values. This is achieved using the least squares method, which reduces the sum of squared errors.

In simple linear regression, the relationship is expressed using the equation $Y = mX + b$, where Y represents the dependent variable, X is the independent variable, m is the slope, and b is the intercept. Multiple linear regression extends this concept by incorporating multiple independent variables to improve predictive accuracy.

Several assumptions are necessary for linear regression to provide reliable results. These include linearity, where the relationship between variables should be straight-line; independence, meaning residuals should not be correlated; homoscedasticity, where variance remains constant across all values of the independent variable; and normality of residuals. If these assumptions are violated, the model's accuracy may decrease. Linear regression remains a powerful and widely used technique due to its simplicity and interpretability, making it an essential tool in data analysis and machine learning.

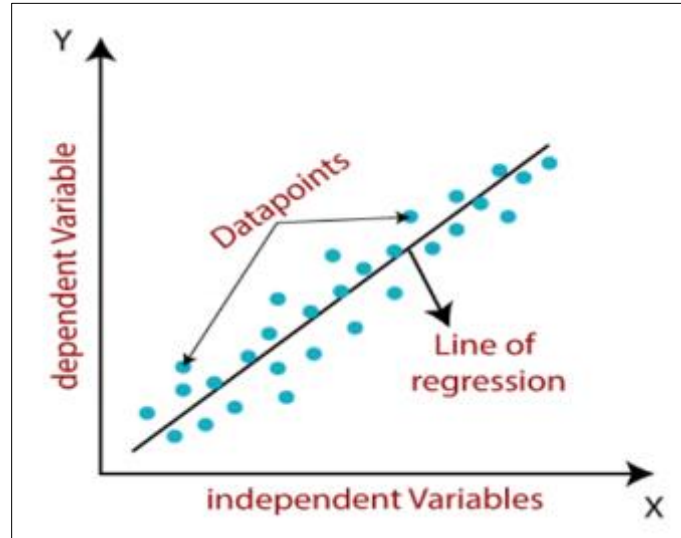


Figure 3 Linear Regression

6.1.2. R^2 metrics:

The R^2 (R-squared) metric is a statistical measure that evaluates how well a regression model fits the data. It represents the proportion of the variance in the dependent variable that is explained by the independent variables. The value of R^2 ranges from 0 to 1, where 1 indicates a perfect fit and 0 means the model explains no variance.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Figure 4 R^2 metric formula

Where:

SS_{tot} (Total Sum of Squares) measures the total variance in the data.

SS_{res} (Residual Sum of Squares) represents the unexplained variance.

A higher R^2 value suggests a better-fitting model, but it does not indicate causation. In multiple regression, an adjusted R^2 is used to account for the number of predictors, preventing overfitting.

6.1.3. Real-Time Display

The project utilizes Streamlit to create an interactive and user-friendly interface for real-time insurance cost predictions. Users enter details like age, BMI, smoking status, and medical history, and Streamlit processes the input through a trained Linear Regression model. The predicted insurance cost or risk category is then displayed instantly.

7. Results

Vital Care Insurance Prediction

Please enter your details below to predict the insurance price.

Enter your age
18 100

Select your sex
Male v

Enter your BMI value
0.00 - +

Enter number of children
0 18

Select your region
1 4

Select medical risk
Low v

Select your lifestyle
Sedentary v

Select your occupation
Desk/Home v

Predict Price

Figure 5 User Interface

Vital Care Insurance Prediction

Please enter your details below to predict the insurance price.

Enter your age
 32

Select your sex

Enter your BMI value

Enter number of children
 9

Select your region
 2

Select medical risk

Select your lifestyle

Select your occupation

Predict Price

Predicted Insurance Price: ₹4830.0

Prediction Successful! 🎉

Figure 6 Medical Insurance Prediction

8. Conclusion

The Vital Care Insurance Prediction System enhances traditional insurance models by integrating machine learning for more accurate and personalized cost estimations. Unlike conventional methods, it considers a wider range of health and lifestyle factors, improving prediction accuracy and adaptability. The system's interactive interface allows real-time data updates, ensuring dynamic risk assessment. By increasing transparency and fairness, it empowers users with insights into their insurance costs and contributing factors. This data-driven approach benefits both policyholders and insurers by optimizing pricing strategies and reducing financial uncertainties, ultimately making healthcare insurance more efficient, accessible, and aligned with individual health needs.

Compliance with ethical standards



Disclosure of conflict of interest



No conflict of interest to be disclosed.

References

- [1] Bharti, Ayushi, and Malik, Lokesh. "Regression Analysis and Prediction of Medical Insurance Cost." International Journal of Creative Research Thoughts, vol. 10, no. 3, March 2022.
- [2] Genita, Jonelle Angelo S., Asuncion, Paul Richie F., and Victoriano, Jayson M. "Performance Evaluation of Regression Models in Predicting the Cost of Medical Insurance." arXiv preprint arXiv:2304.12605, April 2023.
- [3] Reddy, G. Akshara, and Madhuri, N. Latha. "Medical Health Insurance Price Prediction." International Journal of Novel Research and Development, vol. 9, no. 4, April 2024. Kumar, P. Ashok. "Event Based Time Series Sentiment Trend Analysis."
- [4] Orji, Ugochukwu, and Ukwandu, Elochukwu. "Machine Learning For An Explainable Cost Prediction of Medical Insurance." arXiv preprint arXiv:2311.14139, November 2023.
- [5] Morid, Mohammad Amin, et al. "Healthcare Cost Prediction: Leveraging Fine-grain Temporal Patterns." arXiv preprint arXiv:2009.06780, September 2020.
- [6] PANDI, CHIRANJEEVI, THATIKONDA SUPRAJA, P. ASHOK KUMAR, and RALLA SURESH. "A SURVEY: RECOMMENDER SYSTEM FOR TRUSTWORTHY."
- [7] Li, Zhengxiao, Huang, Yifan, and Cao, Yang. "Analyzing Covariate Clustering Effects in Healthcare Cost Subgroups: Insights and Applications for Prediction." arXiv preprint arXiv:2303.05793, March 2023.
- [8] Reddy, G. Akshara, and Madhuri, N. Latha. "Medical Health Insurance Price Prediction." International Journal of Novel Research and Development, vol. 9, no. 4, April 2024.
- [9] Kumar, P. Ashok, B. Vishnu Vardhan, and Pandi Chiranjeevi. "Correction to: Investigating Context-Aware Sentiment Classification Using Machine Learning Algorithms." In *XVIII International Conference on Data Science and Intelligent Analysis of Information*, pp. C1-C1. Cham: Springer Nature Switzerland, 2023.
- [10] Genita, Jonelle Angelo S., Asuncion, Paul Richie F., and Victoriano, Jayson M. "Performance Evaluation of Regression Models in Predicting the Cost of Medical Insurance." arXiv preprint arXiv:2304.12605, April 2023
- [11] Bharti, Ayushi, and Malik, Lokesh. "Regression Analysis and Prediction of Medical Insurance Cost." International Journal of Creative Research Thoughts, vol. 10, no. 3, March 2022.
- [12] Kumar, E. Puneeth, and Krishna, P. Pavan. "Medical Expense Prediction Using Machine Learning." Hindustan University, 2022.
- [13] Chakraborty, Manomita, et al. "Health Insurance Cost Prediction Using Regression Models." 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), IEEE, 2022.

Author's short biography

<p>Mr Ashok Kumar Pasi: Mr. Ashok Kumar Pasi is working as an Assistant Professor in the Department of Computer Science and Engineering (Data Science) at ACE Engineering College, Hyderabad, Telangana. He has 15+ years of teaching experience and one year in the software industry. Holding a B.Tech, M.Tech, and Ph.D., his research focuses on Machine Learning and Deep Learning. His aim is to inspire students and contribute to advancements in technology through his work.</p>	
<p>Lasya Palarapu: Lasya Palarapu is currently pursuing a final-year B.Tech degree at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). She has a growing interest in data science and programming and enjoys exploring new technologies. Lasya strives to enhance her skills and apply them effectively in her field of study.</p>	

<p>Akshitha Mailaram: Akshitha Mailaram is currently pursuing a final-year B.Tech degree at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). She is passionate about technology and problem-solving, with a strong desire to expand her knowledge in software development and data science. Akshitha looks forward to applying her skills to real-world challenges.</p>	
<p>Laxmi Prasanna Kanithi: I am Laxmi Prasanna Kanithi, a final-year B.Tech student at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). I have a keen interest in Machine Learning and Data Science and enjoy exploring predictive analytics and intelligent systems. I am always eager to enhance my skills and apply innovative techniques to solve real-world challenges.</p>	
<p>Deekshith Bommana: Deekshith Bommana is currently pursuing a final-year B.Tech degree at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). He is enthusiastic about learning and improving his technical skills, particularly in software development and data science. Deekshith aims to gain practical experience and contribute to innovative projects.</p>	