e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

( Peer-Reviewed, Open Access, Fully Refereed International Journal )

Volume:06/Issue:02/February-2024 Impact Factor- 7.868 www.irjmets.com

# HEALTH INSURANCE COST PREDICTION USING MACHINE LEARNING

**Sakshi Mishra[*1], Ritwik Kapoor[*2], Yukti[*3], Mahesh G[*4]**

[*1,2,3,4]Department Of Computer Science And Engineering, JSS Academy Of Technical Education, Noida, India.

DOI : https://www.doi.org/10.56726/IRJMETS49200

## ABSTRACT

Amidst the backdrop of escalating healthcare costs, a substantial share of the GDP is allocated to health-related expenditures. This study employs machine learning algorithms, including Random Forest Regression, Gradient Boosted Trees, Linear Regression, and Support Vector Machine, to forecast health insurance costs. The primary objective is to empower individuals in making informed decisions about health coverage based on their unique health attributes. Additionally, the research seeks to aid policymakers in identifying providers with higher costs and implementing targeted cost-containment measures. By evaluating algorithm performance on a health insurance dataset, the study underscores the significance of early cost estimation to guide individuals in selecting suitable coverage. In addressing the pressing need for effective management of healthcare expenses, the findings of this research contribute not only to individual decision-making but also provide valuable insights for policymakers striving to strike a balance between quality healthcare provision and fiscal responsibility. The utilization of machine learning in predicting health insurance costs is pivotal for creating a more transparent and efficient healthcare ecosystem. This research endeavors to foster a nuanced understanding of cost dynamics, empowering both individuals and policymakers in navigating the complexities of the contemporary healthcare landscape.

## I. INTRODUCTION

The relentless rise in healthcare costs, constituting nearly 30% of the GDP, underscores the urgent need for effective strategies in managing health-related expenditures. As the baby boomer generation approaches retirement, their eligibility for Medicare intensifies the demand for innovative tools to control healthcare costs. This research addresses this challenge by leveraging machine learning algorithms to predict medical costs, providing a potential solution to guide patients towards affordable healthcare options. Utilizing a dataset encompassing critical factors such as age, smoking status, family medical history, BMI, marital status, and geography, our study seeks to offer individuals a general sense of the costs they may incur for health insurance.

The medical cost trend, projected to be 7.0% year on year in 2024, surpassing trends in previous years, further emphasizes the urgency for proactive cost management. Additionally, the surge in health insurance premium collections by 40% in 2020, driven by public demand to safeguard against the Covid-19 pandemic, highlights the dynamic nature of the healthcare landscape. Anticipating a substantial rise in medical plan costs in 2023, albeit trailing overall inflation, which stands at approximately 8.5% year over year, the study aims to provide a timely and predictive framework for individuals and policymakers alike to navigate the evolving healthcare cost landscape.

## II. LITERATURE SURVEY

The paper underscores the importance of early estimation of health insurance costs to prevent individuals from being misled into paying for unnecessary or expensive health insurance plans. While the research does not provide an exact amount required by any specific health insurance provider, it aims to offer a general sense of the costs individuals may incur for their health insurance. The findings presented in the literature review suggest that Gradient Boosting Decision Tree Regression demonstrated the highest accuracy rate (87.776%) in predicting medical costs. In comparison, Linear Regression and Random Forest achieved correct predictions about 80% of the time. However, the Support Vector Machine did not perform well and was not considered a reliable predictor in this context. [1]

This paper introduces the challenges of forecasting healthcare costs, highlights the potential of mobile data, and proposes a novel multitask learning-based framework for interpretable medical cost interval prediction, addressing the limitations of traditional methods. The paper uses multitasking framework employs multitask learning to predict sub cost intervals using multidimensional data collected from mobile devices. It follows the multitask learning paradigm, where sub cost intervals are predicted first, and then the total cost interval is predicted based on these subcost predictions. The paper mentions the use of logistic regression methods for data preprocessing to improve the speed of model training and convergence. Additionally, a ResNet structure is employed to maintain network identity mapping. [2]

This paper acknowledges that machine learning offers a variety of techniques for predicting medicine expenditures using historical data and other healthcare variables. Traditional models like multilayer perceptron (MLP), long short-term memory (LSTM), and convolutional neural network (CNN) have been used for this purpose. The review notes that generative approaches, such as generative adversarial networks (GANs), have not been extensively explored for time-series prediction of medicine-related expenditures. GANs are introduced as a potential solution to address the challenges posed by individual diversity in health status and the complex factors influencing costs. The V-GAN model, using an LSTM generator and a CNN discriminator, outperforms other GAN-based prediction models, as well as LR, GBR, MLP, and LSTM models in predicting medical expenditures of patients. The evaluation includes root mean square error (RMSE) on training and test data and the percentage of real-like samples reported by the discriminator model. [3]

This paper discusses the importance of accurately measuring the value of health insurance in people's lives, emphasizing the challenges faced by insurance companies in calculating health insurance charges through traditional processes. The intervention of humans in this process is noted to sometimes lead to faulty or inaccurate results, and as data increases, manual calculations become lethargic and time-consuming. The review suggests that machine learning (ML) models can be highly beneficial in automating and improving the accuracy of the insurance charge calculation process. The primary objective of the proposed model, named ML Health Insurance Prediction System (MLHIPS), is to provide rapid estimation and prediction of insurance charges incurred by a patient at a hospital. The model is designed to aid insurance companies in determining premiums quickly, thereby reducing health expenditure. The proposed model incorporates and demonstrates various regression models, including Multiple Linear Regression, Ridge Regression, Simple Linear Regression, Lasso Regression, and Polynomial Regression. [4]

This paper discusses the pervasive threats and uncertainties individuals face globally, ranging from health risks to financial losses, and highlights the importance of insurance as a means to mitigate these risks. Given the complexity of factors influencing insurance costs, the financial industry seeks accurate measurement of the sum covered and associated insurance fees. Human errors in this process prompt the use of various technologies, including machine learning (ML), to determine insurance premiums. The goal of the study is to predict health insurance costs using ML regression models, specifically multiple linear regression. The dataset from Kaggle.com is utilized for this purpose. Among the regression models tested, Gradient Boosting emerges as the most efficient, achieving an accuracy of 86.86%. The conclusion is that Gradient Boosting outperforms other regression models in estimating insurance costs. The potential benefits of using ML in insurance pricing include attracting clients, saving time in program creation, and reducing individual efforts in policymaking. [5]

In this paper, the authors employ three regression-based ensemble machine learning models—Extreme Gradient Boosting (XG Boost), Gradient-boosting Machine (GBM), and Random Forest (RF)—to predict medical insurance costs. They utilize Explainable Artificial Intelligence (XAI) methods, specifically Shapley Additive explanations (SHAP) and Individual Conditional Expectation (ICE) plots, to identify and explain the key determinant factors influencing medical insurance premium prices in a dataset comprising 986 records, publicly available on the Kaggle repository. The evaluation of the models involves four performance metrics: R-squared ($R^2$), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The results indicate impressive outcomes for all models, with XG Boost achieving the best overall performance but requiring more computational resources. The study aims to assist policymakers, insurers, and potential buyers in making informed decisions when selecting medical insurance policies. [6]

This paper explores the challenges posed by various risks individuals and organizations face, emphasizing the importance of insurance as a financial tool to mitigate or eliminate the costs associated with these risks. The focus is on the precision required by insurance companies to measure the coverage amount and associated charges accurately. The paper highlights the critical role of various parameters in estimating insurance charges, emphasizing the need for high accuracy in these calculations due to the potential for human errors. The goal of the study is to predict insurance costs, and regression is identified as the best approach. Multiple linear regression is specifically chosen due to the presence of multiple independent variables used to calculate the dependent (target) variable. The dataset used in the study pertains to the cost of health insurance, and preprocessing is conducted before training regression models on the training data. Several regression models, including multiple linear regression, Decision Tree Regression, and Gradient Boosting Regression, are employed and evaluated based on testing data. The article concludes that Gradient Boosting Regression provides the highest accuracy, with an R-squared value of 86.7853. [7]

This paper discusses the role of insurance policies in mitigating expenses associated with various risks and highlights the influence of multiple factors on insurance prices. It proposes the use of machine learning (ML) to enhance the efficiency of insurance policy terms in the industry. The study focuses on forecasting insurance amounts for different categories of people using individual and local health data. Nine regression models, including Linear Regression, XG Boost Regression, Lasso Regression, Random Forest Regression, Ridge Regression, Decision Tree Regression, KNN Model, Support Vector Regression, and Gradient Boosting Regression, are employed and trained using a dataset. Predictions are made using the training data, and the models' accuracy is evaluated by comparing the predictions with actual data. The optimal model is identified as XGBoost, providing Mean Absolute Error (MAE) of 2381.567, Mean Squared Error (MSE) of 19806356.6067, Root Mean Squared Error (RMSE) of 4450.4433, and R-squared value of 0.8681. Gradient Boosting and Random Forest are also recognized as top-performing models with R-squared values of 0.8679 and 0.8382, respectively. [8]

## III. METHODOLOGY

**Machine Learning Overview:** Machine Learning, a branch of both computer science and AI, entails leveraging data and algorithms to emulate human learning processes. These specialized algorithms aim to make classifications or predictions through statistical techniques, revealing crucial insights during data mining.

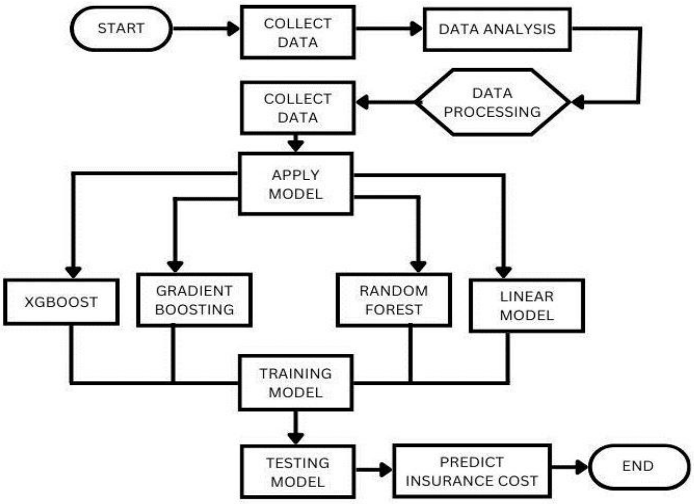


**Figure 1:** Procedure

Examining the data, it becomes evident that age and smoking status exert the most significant impact on insurance amounts, with smoking exhibiting the most pronounced effect. Nevertheless, additional factors such as family medical history, BMI, marital status, and geography also contribute to the overall assessment. This underscores the multifaceted nature of the variables influencing insurance costs, with age and smoking status emerging as primary determinants. Tabular comparisons of algorithms on various parameters.

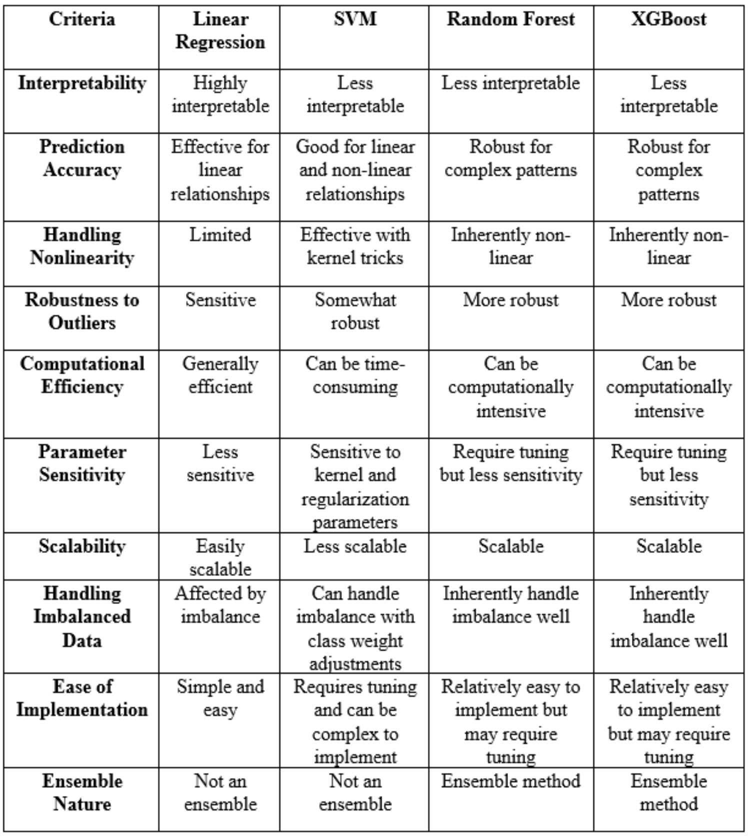
**Linear Regression Algorithm:** Linear regression, a supervised learning algorithm, operates on the principle of predicting a dependent variable (y) based on an independent variable's (x) value. In essence, it quantifies the relationship between these variables, enabling accurate predictions (H. Goldstein, 2012). This tool proves invaluable for data analysts, unravelling intricate patterns and relationships in data to enhance predictions about future outcomes.

**Support Vector Machine Algorithm:** Support Vector Machine (SVM), a widely embraced supervised learning algorithm, excels in solving classification and regression problems, with a primary focus on classification in machine learning. SVM aims to identify an optimal line or decision boundary, known as a hyperplane, effectively segregating multi-dimensional spaces into distinct classes. Leveraging extreme vectors and support vectors, SVM is adept at classifying data and managing high-dimensional spaces.

**Random Forest Regression:** Random Forest Regression adopts a bootstrapping approach, employing multiple decision trees derived from data and amalgamating them through ensemble learning techniques. By averaging the results of randomly selected trees, this method often yields precise predictions and classifications.

**XG Boost Algorithm:** XG Boost, an extension of gradient boosting, is a powerful supervised learning algorithm widely employed for classification and regression tasks. This algorithm enhances predictive accuracy by combining the strengths of multiple weak learners, typically decision trees, in a sequential manner. Similar to Random Forest, XG Boost employs ensemble learning, achieving robust performance in predictive modelling scenarios.

**Table 1.** Algorithm Comparison



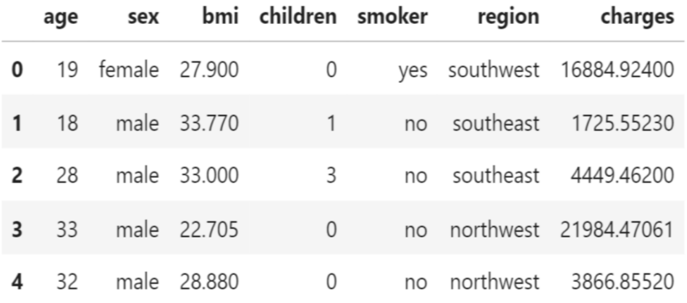
| Criteria | Linear Regression | SVM | Random Forest | XGBoost |
|---|---|---|---|---|
| **Interpretability** | Highly interpretable | Less interpretable | Less interpretable | Less interpretable |
| **Prediction Accuracy** | Effective for linear relationships | Good for linear and non-linear relationships | Robust for complex patterns | Robust for complex patterns |
| **Handling Nonlinearity** | Limited | Effective with kernel tricks | Inherently non-linear | Inherently non-linear |
| **Robustness to Outliers** | Sensitive | Somewhat robust | More robust | More robust |
| **Computational Efficiency** | Generally efficient | Can be time-consuming | Can be computationally intensive | Can be computationally intensive |
| **Parameter Sensitivity** | Less sensitive | Sensitive to kernel and regularization parameters | Require tuning but less sensitivity | Require tuning but less sensitivity |
| **Scalability** | Easily scalable | Less scalable | Scalable | Scalable |
| **Handling Imbalanced Data** | Affected by imbalance | Can handle imbalance with class weight adjustments | Inherently handle imbalance well | Inherently handle imbalance well |
| **Ease of Implementation** | Simple and easy | Requires tuning and can be complex to implement | Relatively easy to implement but may require tuning | Relatively easy to implement but may require tuning |
| **Ensemble Nature** | Not an ensemble | Not an ensemble | Ensemble method | Ensemble method |

## IV. DATA SOURCE

This article explores a dataset available on the Kaggle website, intended for training and testing purposes. The dataset, stored in a well-organized CSV file, can be accessed through the provided link. It comprises 7 columns and a total of 1338 rows, making it a valuable resource with a total number of 9366 entries. To accurately predict health insurance costs, it is essential to preprocess the dataset before applying regression algorithms. The analysis reveals that age and smoking status exert the most significant impact on insurance costs, with smoking demonstrating the most substantial effect. Other factors, including family medical history, BMI, marital status, and geography, also contribute to the prediction. Notably, children's property has minimal impact and was consequently excluded from the input for the regression model to enhance efficiency and accuracy.

It is noteworthy that the observed impacts of age and smoking status on insurance costs are preliminary estimates and do not align with any specific company. The algorithms employed are designed to perform classifications or predictions using statistical techniques, uncovering key insights in data mining processes.

These insights, can serve as crucial growth indicators for businesses and applications when utilized correctly. Recognizing the influential roles of age and smoking status in insurance costs empowers individuals to make more informed decisions, providing valuable suggestions.



| | age | sex | bmi | children | smoker | region | charges |
|---|---|---|---|---|---|---|---|
| 0 | 19 | female | 27.900 | 0 | yes | southwest | 16884.92400 |
| 1 | 18 | male | 33.770 | 1 | no | southeast | 1725.55230 |
| 2 | 28 | male | 33.000 | 3 | no | southeast | 4449.46200 |
| 3 | 33 | male | 22.705 | 0 | no | northwest | 21984.47061 |
| 4 | 32 | male | 28.880 | 0 | no | northwest | 3866.85520 |

**Figure 2:** Data Fields

## V. CONCLUSION

In conclusion, our preliminary analysis suggests that Gradient Boosting Decision Tree Regression shows promising potential, exhibiting the highest accuracy rate of 87.776% in predicting the amount. While linear regression and random forest also demonstrate decent predictive capabilities, Support Vector Machine lags behind as an unreliable predictor in this context. Notably, XG Boosting Regression emerges as the superior model, consistently delivering high accuracy across the evaluated attributes.

However, our exploration is far from complete. We foresee exciting avenues for future implementation. Introducing unpredictability through the feature selection process, particularly with the Random Forest algorithm, holds promise for further enhancing prediction accuracy. Furthermore, assessing the scalability of our system remains a critical area for improvement. Experimentation with larger datasets, ideally with millions of records, will provide valuable insights into the system's performance under increased load.

## VI. REFERENCE

[1] Ajay Kumar Sahu, Gopal Sharma, Janhvi Kaushik, Kajal Agarwal, Devender Singh (2023). Health Insurance Cost Prediction by using Machine Learning. SSRN id-4366801

[2] Yongjie Yan,Guang Yu,and Xiangbin Yan(2020). Online Doctor Recommendation with Convolutional Neural Network and Sparse Inputs.

[3] Mehmood Ali Mohammed (2023). Use of Machine Learning in Optimizing Medical Appointment Schedules. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 14 (1).

[4] Lin Sun,Tingqi Wang, Bei Hui, Yun Li,and Ling Tian(2022). Explainable and Personalized Medical Cost Prediction Based on Multitask Learning over Mobile Devices. Research Article, Hindawi. Article ID 8966266.

[5] Shruti Kaushik, Abhinav Choudhury, Sayee Natarajan, Larry A. Pickett, Varun Dutt (2020). Medicine Expenditure Prediction via a Variance- Based Generative Adversarial Network. IEEE Xplore, ISSN-2169-3536.

[6] Sudhir Panda, Biswajit Purkayastha, Dolly Das, Manomita Chakraborty, Saroj Kumar Biswas (2022). Health Insurance Cost Prediction Using Regression Models. International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON).

[7] Asst. Prof. Ms. Madhuri Thorat, Mohasin Patel, Yog Kute, Muskan Sharma, Shweta Bhosale (2022). Medical Insurance Cost Prediction Using Machine Learning.

[8] Ugochukwu Orji, Elochukwu Ukwandu(2023). Machine learning for an explainable cost prediction of medical insurance.

[9] Mukund Kulkarni, Dhammadeep D. Meshram, Bhagyesh Patil, Rahul More, Mridul Sharma, Pravin Patange. Medical Insurance Cost Prediction using Machine Learning. International Journal for Research in Applied Science & Engineering Technology (IJRASET**)** ISSN: 2321-9653.