International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

# HEALTH INSURANCE PRICE FORECASTING

Subha Indu .S [1], Professor

Vidhyalakshmi.M[2], Student

Department of Software Systems,

Sri Krishna Arts and Science College,

## Abstract:

This initiative entails the development of a health insurance prediction system utilizing machine learning. Given the heightened significance of health insurance post the Covid-19 pandemic, numerous endeavors have been made to address this issue. Analyzing the factors influencing health insurance costs poses a considerable challenge, and employing a regression model proves effective in comprehending intricate patterns for predicting medical insurance expenses. The dataset utilized for this study was sourced from Kaggle.

This project employs machine learning algorithms to establish a connection between medical costs. The objective is to devise a methodology for predicting healthcare expenses, leveraging machine learning algorithms to guide individuals toward more affordable options. Furthermore, this technology can assist policymakers in identifying providers with higher costs, enabling potential corrective measures. The Random Forest algorithm is employed to forecast insurance costs in this study. Various machine learning models, such as KNN and Linear Regression, will be experimented with on the same dataset to compare results. Early estimation of health insurance expenses is beneficial, preventing individuals from potentially investing in unnecessary and costly coverage. While our research does not provide specific amounts tailored to individual insurance providers, it does offer a general understanding of the potential costs individuals may encounter in securing their health insurance.

## I INTRODUCTION

The aim of this study is to aid individuals in gauging the financial requirements for health insurance based on their individual health status. By doing so, individuals can prioritize relevant health aspects of insurance over unnecessary ones. In today's world, possessing health insurance is crucial, and the majority of people maintain affiliations with either public or private health insurance

providers. The factors influencing insurance costs can vary among different companies. Furthermore, individuals in rural areas may not be aware that the Indian government offers free health insurance to those below the poverty line. Despite this, the process can be intricate, leading some rural residents to opt for private health insurance or forego insurance altogether. Additionally, individuals may be at risk of being misguided into investing in costly health insurance plans that may not be necessary. While our research doesn't provide specific amounts tied to any particular health insurance provider, it does offer a general understanding of potential costs individuals may encounter in securing their health insurance. It's important to note that this is a preliminary estimate and should not be the sole consideration when selecting health insurance, as it doesn't align with any specific company. Early estimation of health insurance costs is beneficial in guiding individuals to carefully consider the required amount.

## II BACKGROUND STUDY

Insurance serves as a protective measure against various risks, either eliminating or reducing the financial losses incurred. The cost of insurance is influenced by multiple factors, contributing to the formulation of insurance policies. Utilizing machine learning (ML) in the insurance sector enhances the efficiency of policy wording. This research explores the application of different regression models to predict insurance costs, including Multiple Linear Regression, Generalized Additive Model, Support Vector Machine, Random Forest Regressor, CART, XGBoost, k-Nearest Neighbors, Stochastic Gradient Boosting, and Deep Neural Network. The study concludes that the Stochastic Gradient Boosting model outperforms others, yielding an MAE value of 0.17448, RMSE value of 0.38018, and an R-squared value of 85.8295 [3].

In this thesis, Nidhi Bhardwaj and Rishabh Anand analyze personal health data to predict insurance amounts for individuals. The study employs three regression models: Multiple Linear Regression, Decision Tree Regression, and Gradient Boosting Decision Tree Regression, comparing and contrasting their performance. The models are trained using a dataset, and the training process leads to predictions. Subsequently, the predicted amounts are compared with the actual data to validate the models. The study reveals that both Multiple Linear Regression and Gradient Boosting algorithms outperform Linear Regression and Decision Tree. Particularly, Gradient Boosting stands out

as the most suitable choice due to its efficient computational time while achieving comparable performance to Multiple Regression [5].

In the exploration conducted by Gaurav Kumar and Nidhi Prajapati, the anticipation of health insurance costs emerges as a pivotal aspect for insurance providers, guiding premium determinations, resource allocations, and strategic decision-making. This paper delves into a comprehensive study on health insurance cost prediction through the application of machine learning techniques [6]. The research involved the collection and preprocessing of data related to demographics, medical history, and lifestyle factors within a representative sample population. Various machine learning algorithms, including support vector machine, deep learning, linear regression, decision tree, and random forest, were employed to construct multiple models. Performance evaluations, based on metrics such as mean squared error and R-squared, demonstrated the high accuracy of the developed models in predicting health insurance costs. The outcomes underscore the efficacy of machine learning algorithms in this context, offering valuable insights for insurance providers to enhance decision-making processes. The study also advocates for continued research to refine models and gain a deeper understanding of the factors influencing health insurance costs [6].

## III PROBLEM DEFINITION

Health insurance price forecasting using machine learning involves leveraging algorithms and statistical models to predict future health insurance premiums based on historical data and relevant features. The goal of this project is to develop a machine learning model that can accurately forecast health insurance premiums for individuals based on relevant factors. The model must adhere to industry regulations and compliance standards governing health insurance pricing. Ensuring fairness, transparency, and ethical considerations in the predictions is paramount.

## IV PROPOSED SYSTEM

Employing prediction is instrumental in achieving accurate results for health insurance prediction. The identification of relationships between independent and dependent variables is facilitated through prediction methods. Extracting concealed insights related to health insurance price prediction is accomplished by utilizing a dedicated health insurance price prediction database. The enhancement of the proposed model involves the application of

a composite of three classifiers: Linear Regression, Random Forest, and KNN algorithm. Implementation of the suggested model is carried out using Python, and the results substantiate an accuracy rate of 86%.

**Flow Diagram**

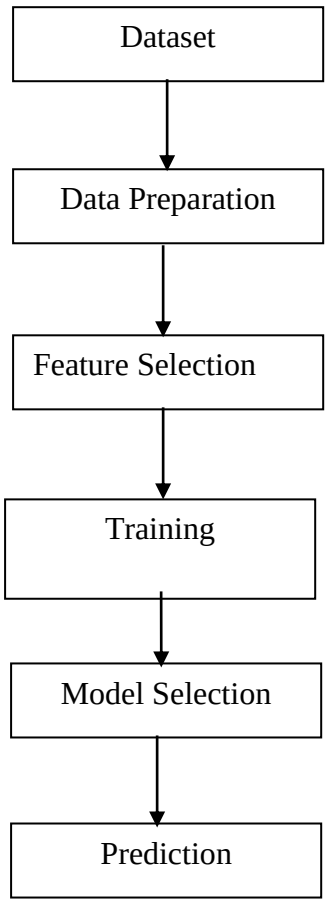


Fig.1. Flow Diagram

**V RESULT AND DISCUSSION**

| ML Algorithm | Accuracy |
|---|---|
| KNN | 77% |
| RF | 86% |
| Linear Regression | 76% |

Table1. Performance Comparison

The suggested model incorporates an automated classification system for price forecasting, designed to categorize the input textual dataset. Utilizing features extracted from LR, RF, and KNN algorithms, the classification process is executed. Notably, the RF model demonstrates a notable level of accuracy, contributing to the overall precision of the proposed model, which achieves an accuracy rate of 86.6%.

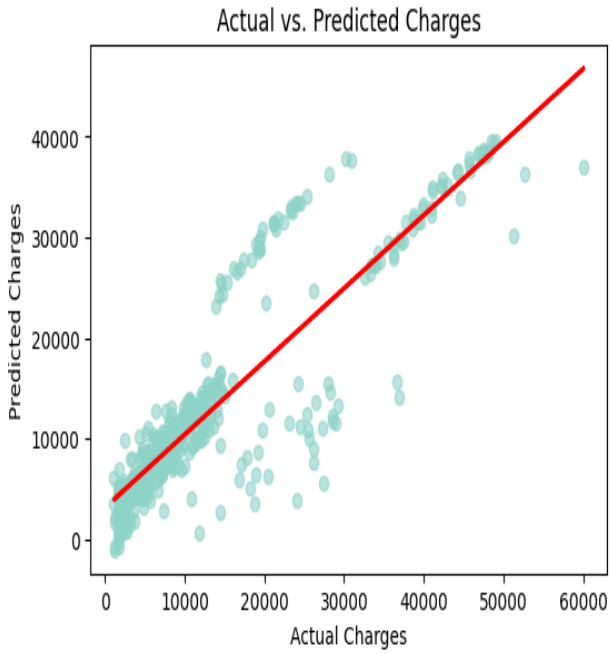


Fig.2. Predicted Charges

Fig.2, Predicting health insurance prices involves utilizing various factors and methodologies to estimate future costs.

International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

While I can't generate real-time data or specific graphs

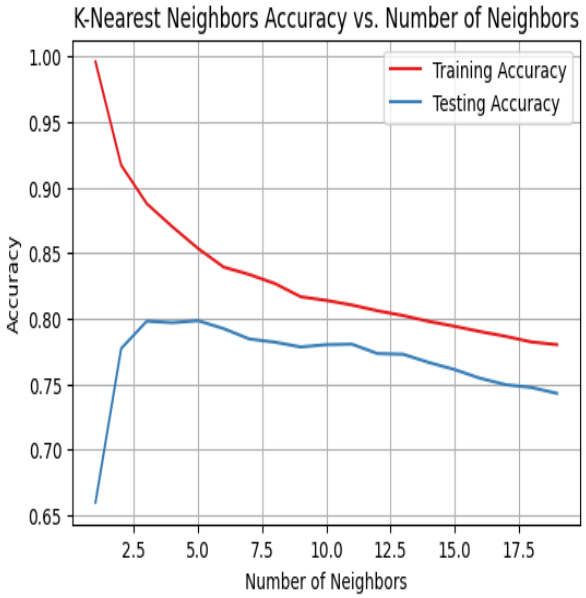


Fig.3. KNN Training and Testing Accuracy

Fig3, K-Nearest Neighbors (KNN) is a supervised machine learning algorithm used for classification and regression tasks. It's a simple algorithm that works by finding the k-nearest data points in the training set to a given test data point and making predictions based on the majority class (for classification) or average (for regression) of those neighbors. Training accuracy measures how well the model performs on the data it was trained on. Testing accuracy, also known as validation accuracy, measures how well the model generalizes to new, unseen data.

## VI CONCLUSION

In conclusion, the utilization of machine learning and data analysis techniques in health insurance price forecasting brings considerable advantages to both insurance providers and consumers. Accurate prediction of future insurance premiums enables providers to optimize pricing strategies, assess risks more efficiently, and improve overall business performance. For consumers, this means the ability to make informed decisions regarding insurance coverage, leading to better financial planning and healthcare choices. The health insurance price forecasting process involves meticulous steps, including data collection, preprocessing, feature selection, model selection, training, evaluation, and continuous monitoring. Employing Random Forest algorithms and other machine learning techniques allows insurers to analyze historical data and various influencing factors, creating robust forecasting models.

In summary, the application of advanced machine learning techniques in health insurance price forecasting plays a pivotal role in shaping the insurance landscape's future. Embracing challenges, staying abreast of the latest technologies, and adhering to ethical standards empower insurers to leverage data-driven insights for informed decisions, improved customer

experiences, and contributions to a more efficient and equitable healthcare system.

**Reference**

[1]Kafuria, A. D. (2022). Predictive Model for Computing Health Insurance Premium Rates Using Machine Learning Algorithms. International Journal of Computer (IJC), 44(1), 21-38.

[2] Sahare, A. N. (2023). Forecasting Medical Insurance Claim Cost with Data Mining Techniques (Doctoral dissertation, Dublin, National College of Ireland).

[3] Hanafy, M., & Mahmoud, O. M. A. (2021). Predict Health Insurance Cost by Using Machine Learning and DNN Regression Models. Int. J. Innov. Technol. Explor. Eng, 10(3), 137-143.

[4] Kathrin, D., Günther, I., & Harttgen, K. (2021). Using machine learning to predict health insurance enrolment and take-up of health services.

[5] Nidhi Bhardwaj , Rishabh Anand "Health Insurance Amount Prediction" International Journal of Engineering Research & Technology (IJERT) http://www.ijert.org ISSN: 2278-0181 Vol. 9 Issue 05, May-2020.

[6] Gaurav Kumar, Nidhi Prajapati "Health Insurance Cost Prediction App" International Journal of Research Publication and Reviews Vol 4, no 6, pp 1284-1287 June 2023.

[7] Chinthala Shreekar , Maloth Kiran, Dubbudu Sumanth, Preethi Jeevan "Cost Prediction of Health Insurance" International Research Journal of Engineering and Technology (IRJET) Volume: 10 Issue: 01 | Jan 2023.

[8] Sam Goundar, Suneet Prakash, Pranil Sadal, Akashdeep Bhardwaj "Health Insurance Claim Prediction Using Artificial Neural Networks" International Journal of System Dynamics ApplicationsVolume 9 • Issue 3 • July-September 2020.

[9] Gupta, S., & Tripathi, P. (2016, February). An emerging trend of big data analytics with health insurance in India. In 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH) (pp. 64-69). IEEE.

[10] Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. Risks, 7(2), 70

[11] Singh, R., Ayyar, M. P., Pavan, T. S., Gosain, S., & Shah, R. R. (2019, September). Automating Car Insurance Claims Using Deep Learning Techniques. In 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM) (pp. 199-207). IEEE.

[12] Stucki, O. (2019). Predicting the customer churn with machine learning

methods: case: private insurance customer data.

[13] Fauzan, M. A., & Murfi, H. (2018). The accuracy of XGBoost for insurance claim prediction. Int. J. Adv. Soft Comput. Appl, 10(2).

[14] Kowshalya, G., & Nandhini, M. (2018, April). Predicting fraudulent claims in automobile insurance. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1338-1343). IEEE.

[15] Kayri, M., Kayri, I., & Gencoglu, M. T. (2017, June). The performance comparison of multiple linear regression, random forest and artificial neural network by using photovoltaic and atmospheric data. In 2017 14th International Conference on Engineering of Modern Electric Systems (EMES) (pp. 1-4). IEEE.