iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71763

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Research Work on Predictive Premium Model of Medical Cost

Anurag Shrivastava[1], Kripa Shankar Pathak[2], Ayush Narayan[3], Muskan Chaudhary[4], Gaurav Ghildiyal[5]
Department Of Computer Science & Engineering, Babu Banarasi Das Northern India Institute Of Technology Lucknow, Uttar Pradesh, India

Abstract: ***Insurance is a protection policy which minimizes or eliminates the impacts of expenses loss caused by different risks. A number of factors determine risk costs. These factors dictate how insurance packages are designed. The effectiveness of certain clauses within an insurance policy can be improved through the use of machine learning (ML). In this study, we utilize individual local health data for forecasting insurance amounts tailored to specific person groups. For the purpose of evaluating the performance of these algorithms, nine regression models were used: Linear Regression, XG Boost Regression, Lasso Regression, Random Forest Regression, Ridge Regression, Decision Tree Regression, KNN Model, Support Vector Regression, and Gradient Boosting Regression. The model was trained on the provided dataset, which included a portion of the data as training data. After training the model, it was tested against real data. The validation of the model was done by comparing the predicted data which was assumed abundant. After that the comparison was carried out between the accuracy of these models.***
We aim to provide some valuable insights for researchers, practitioners, and policymakers for effective decision-making in healthcare contexts by exploiting machine learning methodologies.
Keywords: Healthcare, Insurance, Underwriting, Premium Prediction, Machine Learning, Predictive Modelling, And Data-Driven Decision-Making.

I. INTRODUCTION

An example of a rapidly expanding industry around the world is digital health. Over the past five years, the number of digital health companies worldwide has increased by 100% [1]. In developed countries, health insurance has two main problems: increasing expenditures for health care services and a growing population of uninsured people. There is a strong grassroots political initiative that seeks to solve these problems. In this region, governments have already committed several hundred million dollars to develop the digital health sector. Private health insurance is particularly important within the health care system for patients with uncommon diseases [2] because, in the case of these patients, medical and preventive insurance can substantially reduce the expense of treatments. The reality is that we dwell in a frightening and uncertain world. houses, companies, buildings. The most startling prospect is the acquisition or loss of property and possessions. Happiness and health are the main concern of people and the very existence of everything. Medical insurance forms an important part of an industry. However, projecting medical expenditures is complex due to most finances being received from patients with uncommon ailments and conditions. For prediction of data, numerous ml algorithms and deep learning methods are employed. Two of the major aspects considered are the training time and accuracy; the latter being critical. Most modern machine learning algorithms train on data within an acceptable timeframe. Unfortunately, the prediction outcomes using these methods are not very reliable. Deep learning models are capable of uncovering concealed structure s, but applying them in real-time scenarios is limited to the amount of time spent in training [3].

II. LITERATURE SURVEY

This subsection covers the ongoing work concerning information retrieval and machine learning techniques. The prediction of claims has been approached in a number of articles. “Using telematics data for predicting automobile insurance claims,” Jessica Pesantez-Narvaez claimed the authorship on Year 2019. This study, although limited in scope, attempted to test the predictive power of logistic regression against XG Boost in the prediction of accident states. The results and vibrations indicated that logistic regression is a superior model to XG Boost for the reasons of its interpretability and predictability [4]. The above mentioned studies do not seek claim issues assuming the cost and extent of the claims are ignored. The problem, however, is solving costs of healthcare services where very sophisticated statistical methods, ML, and deep neural networks are employed.

A key challenge in enhancing clinical decision support systems is to discover solutions that enable efficient processing of individual cases, thorough data analysis without gaps, personalized approaches, and full integration of the latest advancements in medical science through cloud services. Many current clinical decision support systems fail to offer complete informational assistance for the diagnostic and treatment process. As a result, there is an issue with insufficient data or a high volume of missing data.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

III. METHODOLOGY

A. Dataset Description

We obtained the data set from the Kaggle website [5] in order to calculate the cost of this model prediction. The data set is split into two categories: training data and test data, and it has seven attributes. The majority of the data used is for testing, with just around 20% being used for training. The training data set is used to create a model that forecast medical insurance cost by year, and test data set is used to assess the regression model.

B. Data Analysis

There were 1338 rows and 7 columns in our data set. The charges variable, which has a float value, is our aim. Maximum number of individuals in our dataset range in age from 18 to 22.5, and the majority of them are male. Few have more than three children, and the majority of them have a BMI between 29.26 and 31.16. In this dataset, four main regions are taken into account: northeast, northwest, southeast, and southwest. The largest concentration of smokers is in the southeast, where 1064 out of 1338 people smoke.

C. Model Specification

The goal of the study is to forecast insurance cost based on variety of factors including age, sex, children's number, location, BMI and whether or not person smokes. All of these characteristics aid in our ability to calculate the price of health insurance. Several regression models are used in this study to calculate the cost of insurance. Data is used for training 80% of time and testing 20%.

IV. TECHNOLOGIES, TOOLS AND TECHNIQUES

A. Technologies

- Electronic Health Records (EHRs): EHRs are a crucial source of data for building predictive models. They contain comprehensive patient information, including diagnoses, treatments, and medical history, which are valuable for predicting future healthcare costs.
- Big Data Technologies: ML algorithms often require large datasets to learn complex relationships. Big data technologies like Hadoop and Spark are used to manage and process these vast amounts of data.
- Cloud Computing: Cloud computing platforms (e.g., AWS, Google Cloud, Azure) provide scalable infrastructure for storing, processing, and deploying ML models.

B. Tools

- Programming Languages: Python and R are popular programming languages used in ML for data analysis, model building, and implementation.
- ML Frameworks: Libraries and frameworks like TensorFlow, PyTorch, and Scikit-learn provide pre-built algorithms and tools for building and deploying ML models.

C. Techniques

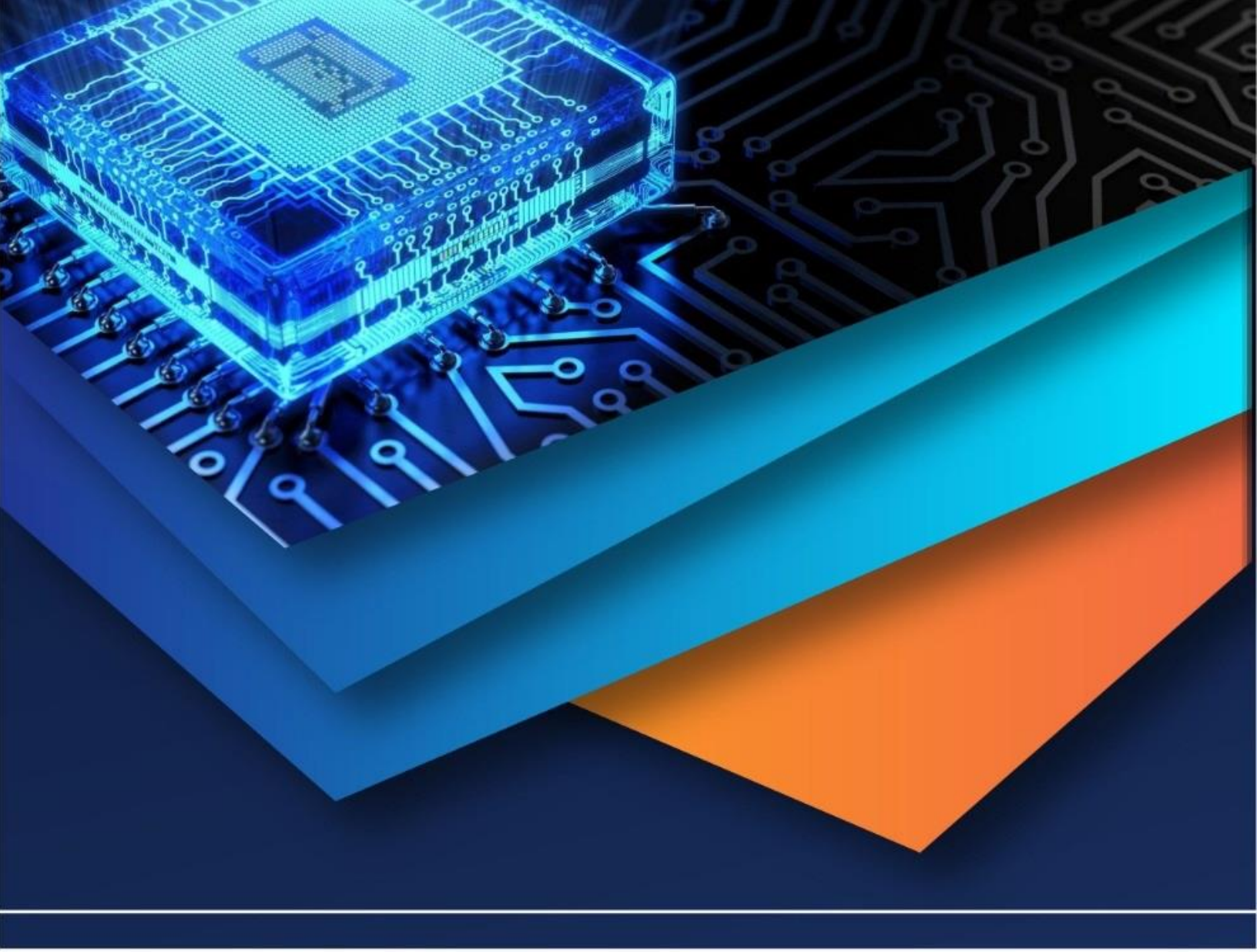
- Regression Analysis: Regression models are used to predict continuous variables like medical costs based on input features.
- Decision Trees and Ensemble Methods: Decision trees and ensemble methods like random forests and gradient boosting algorithms are used for classification and regression tasks.
- Data Preprocessing: Cleaning and transforming data to ensure it is suitable for use in ML models.
- Model Evaluation: Evaluating the performance of ML models using metrics like accuracy, precision, recall, and F1score.
- Model Selection: Choosing the best model based on its performance and interpretability.
- Model Deployment: Deploying ML models into production systems to make predictions on new data.

V. CONCLUSION

In order to forecast health insurance prices based on provided factors in a Kaggle site medical cost individual data set, the study combines ML regression models. Table IV is a list of the outcomes. By predicting insurance rates based on a variety of factors, insurance policy firms may attract consumers and save time. Machine learning may significantly reduce these individual efforts in price analysis since ML models can compute costs quickly while doing so would take a person a long time. Large volumes of data can also be handled via machine learning techniques. The work might be improved in the future by building a web application based on the XGBoost or Gradient Boosting algorithm and using a larger dataset than that used in this study.

REFERENCES

[1] "Digital Health 150: The Digital Health Startups Transforming the Future of Healthcare | CB Insights Research", CB Insights Research, 2022. [Online]. Available: https://www.cbinsights.com/research/report/digital-health-startups-redefininghealthcare.
[2] J. H. Lee, “Pricing and reimbursement pathways of new ophan drugs in South Korea: A longitudinal comparison. in healthcare,” Multidisciplinary Digital Publishing Institute, vol. 9, no. 3, pp. 296, 2021.
[3] 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
[4] N. Shakhovska, S. Fedushko, I. Shvorob and Y. Syerov, “Development of mobile system for medical recommendations,” Procedia Computer Science, vol. 155, pp. 43–50, 2019
[5] MedicalCost Personal Datasets: https://www.kaggle.com/datasets/mirichoi0218/insurance
[6] J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz, "Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression, " Risks, vol. 7, no. 2, p. 70, Jun. 2019, doi: 10.3390/risks7020070.
[7] M. hanafy and O. Mahmoud, "Predict Health Insurance Cost by using Machine Learning and DNN Regression Models", International Journal of Innovative Technology and Exploring Engineering, vol. 10, no. 3, pp. 137-143, 2021. Doi: 10.35940/ijitee.c8364.0110321



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY