IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 6, June 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2023.1206013 |

Health Insurance Cost Prediction Using Regression Models

Chinmaye B P, Harshitha G, Pratibha Pujari

UG student, Department of CSE, Nagarjuna College of Engineering & Technology, Karnataka, India

UG student, Department of CSE, Nagarjuna College of Engineering & Technology, Karnataka, India

Assistant Professor, Department of CSE, Nagarjuna College of Engineering & Technology, Karnataka, India

ABSTRACT: Individuals opt for insurance policies as a safeguard against potential financial setbacks, providing a safety net in unforeseen circumstances affecting them or their belongings, given the unpredictability of the future. Insurance coverage extends to various aspects of life, including homes, businesses, vehicles, health, and education. Among these, health insurance holds particular importance due to the prevalence of numerous serious illnesses and accidents in today's society. The rising costs of medical treatment pose significant challenges, both in terms of comprehension and sustainability. This study utilizes machine learning techniques such as Random Forest, Linear Regression, and Cross Validation, alongside datasets sourced from KAGGLE's repository, to predict health insurance expenditures. By evaluating the performance metrics of these algorithms, one can discern the most effective method for cost prediction and deepen their understanding of the topic.

KEYWORDS: health insurance, dataset, regression, random forest, repository

I. INTRODUCTION

Health insurance is a comprehensive term encompassing financial coverage for a multitude of health-related issues, spanning from minor injuries to major accidents and beyond. Various machine learning algorithms have been employed to construct models capable of estimating an individual's health insurance expenses using diverse datasets. These models typically utilize classification methods such as Random Forest and regression techniques like Linear Regression, in addition to Cross Validation. The proposed model relies on six primary factors for cost estimation: age, gender, geographical location, body mass index (BMI), smoking habits, and number of dependents. It's noteworthy that health insurance costs may vary significantly among individuals and across insurance providers. Many individuals residing in rural areas may lack awareness regarding government-sponsored health insurance schemes for those below the poverty line, resulting in low uptake of insurance. Early prediction of health insurance costs can facilitate improved financial planning for insurance purchases. Engineers have collaboratively devised various machine learning algorithms for this purpose, leveraging datasets containing diverse parameters. Combining different classification methods can enhance the accuracy of health insurance cost predictions. These predictions can be made accessible and cost-effective, utilizing real-life data from both healthy and sick individuals. The system forecasts potential outcomes based on various factors, offering estimates of health insurance costs in US dollars. Moreover, an application could be developed to provide convenient and reliable predictions based on basic factors such as age, gender, and location.

Section 2 of this paper present a review of related work, then Section 3 gives explanation of the proposed work, the features used along with the block diagram of the proposed method. Moreover, the result of the model is presented in Section 4 and finally the conclusion is outlined in Section 5.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2023.1206013 |

II. RELATED WORK

Health Insurance Cost Prediction can be performed with the help of many Machine learning algorithms. It can be performed by using different Regression Models, single technique or by combination of multiple techniques. Few papers also show the comparison of multiple techniques used and choose the technique with best accuracy to detect heart disease.

The model proposed uses methods like Generalized Linear Model where a mean function and a variance function are specified and the parameters are estimated using these structural assumptions, to build the health insurance prediction model, Log-linear, GLM and a statistical model is used. The most complex Statistical method used to solve this is Markov Chain model, The most essential attribute of the algorithm is that it combines the models by allowing optimization of arbitrary function [1].

The work in [2] presents the model where the feature selection of the health insurance cost prediction is done using weighted evidential regression. In this work the author has the fact that although prediction of the cost of the health is useful for the budget management, but most of people don't know the reason behind the prediction. So this model along with weighted regression model it uses K-NN.

The model built in [3] is flexible and easy to use. It used many regression models namely Multi Linear Regression, lasso, Ridge, and DNN Regression these models are tested and compared, In this the author has focused on Regression analysis between the outcomes and associated variables.

The model proposed in [4] used a large scale dataset which they claim to be about 242075 individuals over past three years, but the algorithm used is decision tree. Datasets has been processed in python using machine learning algorithms and Random Forest algorithms. Selected data from database under training set is trained with different algorithm like KNN, Adaptive Boost, Decision Tree and K-mean.

In this they used computational approach for predicting medical insurance cost, in which the domains that they have choosen is applied mathematics, soft computing, and fuzzy logic, can be done using the hybrid machine learning techniques as in [5]. It uses support vector regression, ridge regression, stochastic gradient, Hybrid Random Forest with Linear model. The proposed model uses all features without any restrictions of feature selection.

The proposed model classifies the patient based on majority vote of several machine learning models to provide more accurate solutions than having only one model. Hard Voting Ensemble Model is a technique where multiple machine learning models are combined, and the prediction of result is based on R-squared & mean squared error to predict the most effective model.

The work in [6] uses Decision Tree, k-nearest neighbours, SVM, to process the datasets. The datasets used are classified in terms of medical parameters. The study also examined the impact of the preprocessing steps like normalization and imputation on model performance.

An application is built which can predict vulnerability of health insurance cost prediction. Main aim behind developing the application is to make it user-friendly so that regular monitoring of the price is made possible. To fulfill these requirements, paper [8] states that the factors used are Deep learning techniques like CNN's and RNN's is used to build the application.

Large amount of data is produced and collected by the healthcare organization on the daily basis. Weighted Association Rule is a type of data mining technique used for extracting the data directly from the electronic records. The work proposed in [9] is health insurance cost prediction used Big data and Machine learning to improve prediction accuracy and model robustness.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2023.1206013 |

The proposed model in [10] is hybrid combination of two different methods namely Decision Tree and Logistic Regression based on ensemble learning. This combination is compared with seven other models. HGBDTLR is hybrid gradient boosting decision tree with logistic regression.

III. PROPOSED METHODOLOGY

The proposed work uses the Random Forest Regression, Linear Regression, and Cross Validation Techniques. Health insurance data collected is already pre-processed. The dataset which is collected from the KAGGLE's repository contains a total of 1300 instances and 6 features. The algorithm requires input from the user (such as age, sex, smoker, bmi, no. of children and region).and Cross Validation technique is used to test the model, and the model with acceptable accuracy for the prediction is then obtained.

TABLE I: DATA DESCRIPTION

NAME	DESCRIPTION
Age	Person's age
Sex	Male/Female
Smoker	Whether the person smoker or not
BMI	Body Mass Index of the person
No. of children's	No. of kids of the person
Region	Where the customer lives: Southwest, Southeast, Northwest, Northeast

There are 6 features used for detecting the cost of the Health Insurance, Features used are described & mentioned. These are the required and most important features for any health insurance cost prediction model.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2023.1206013 |

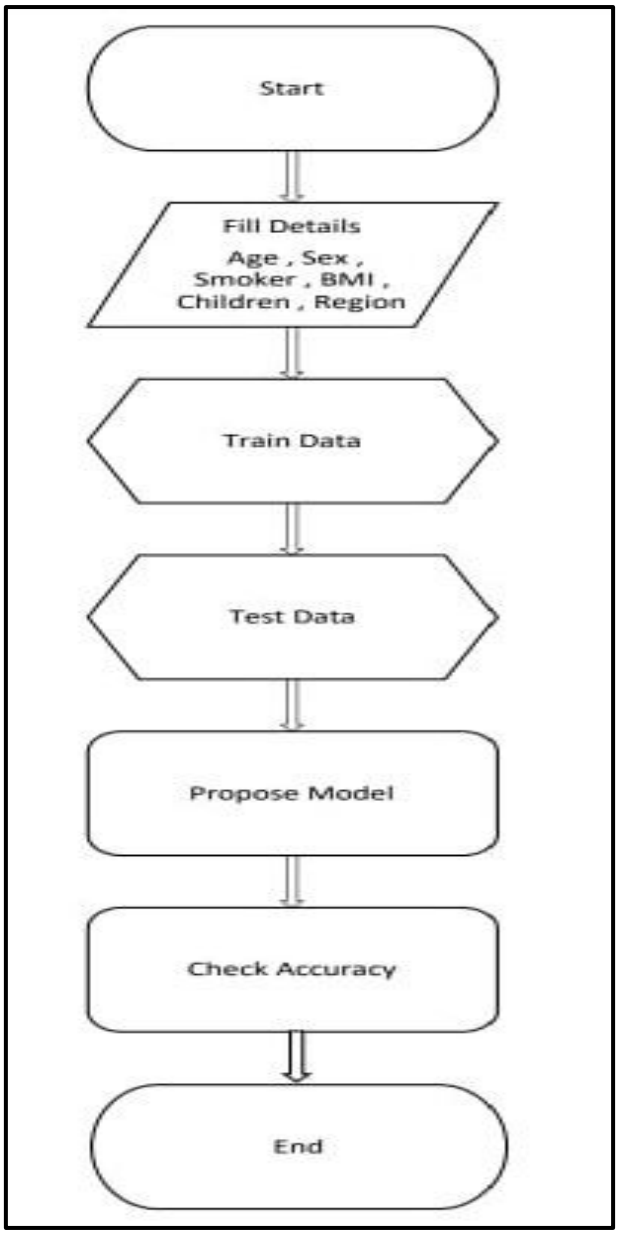

Figure 1: Data Flow diagram of Health Insurance Cost Prediction

Patient data is collected for all 6 features that proposed model uses for health insurance cost prediction. As mentioned earlier in this section, dataset has 1300 instances of 6 features of patient data and is stored and accessed from CSV files. Preprocessing is where dataset is checked for missing values or corrupted values, if found any those are dropped. So user initially is drived to the user interface of the proposed model, where the user usually fills his/her data, then the data enter the data pre-processing stage, where the missing values are checked and then it is sent to the models used which predicts the cost from the processed data using regression models, at last the system sends the predicted cost of the health insurance to the user.

IV. RESULTS AND DISCUSSION

The proposed model used 6 attributes and 1300 instances of patient data. At every different set of the dataset instances accuracy is calculated. The best hyperparameters are chosen for every different dataset selected. The output terminal shows predicted amount in US dollars.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2023.1206013 |

Figure 2: Output of the predicted amount

V. CONCLUSION AND FUTURE ENHANCEMENTS

The proposed work the amount of the Health Insurance. Here, the output obtained is in the form of US dollars. The detection is performed by using a model built using different regression models & etc. The data collected was preprocessed and all the 6 features are used in the model has obtained the result. In Future metaheuristic algorithms can be used to modify the parameters of machine learning and deep learning approaches on multiple medical health related databases

REFERENCES

[1] Predicting Health Care Costs Using Evidence Regression by Belisario Panay 1, Nelson Baloian José A. Pino 1,Sergio Peñafiel 1 Horacio Sanson 2 and Nicolas Bersano 2.Available online: https://www.mdpi.com/2504-3900/31/1/74

[2] Feature Selection for Health Care Costs Prediction Using Weighted Evidential Regression by Belisario Panay, Nelson Baloian, José A. Pino, 1 Sergio Peñafiel :https://www.mdpi.com/1424-8220/20/16/4392

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2023.1206013 |

[3] Hanafy M., Mahmoud O.M.A. Predict Health Insurance Cost by Using Machine Learning and DNN Regression Models. Int. J. Innov. Technol. Explor. Eng. 2021;10:137– 143. doi: 10.35940/ijitee.C8364.0110321. [CrossRef] [Google Scholar]
[4] Health Insurance Amount Prediction by Nidhi Bhardwaj, Rishab Anand Delhi, India.(Publisher: International Journal of Engineering Research & Technology (IJERT)) :https://www.ijert.org/research/health-insurance-amount-prediction-IJERTV9IS050700.pdf
[5] A Computational Intelligence Approach for Predicting Medical Insurance Cost by Ch Anwar Ul Hassan, Jawaid Iqbal, Saddam Hussain : https://www.hindawi.com/journals/mpe/2021/1162553/
[6] Health Insurance Premium Prediction with Machine Learning. [(accessed on 9 May 2022)]. Available online: https://thecleverprogrammer.com/2021/10/26/healthinsurance- premium-prediction-with-machine-learning/
[7] ul Hassan C.A., Iqbal J., Hussain S., AlSalman H., Mosleh M.A.A., Sajid Ullah S. A Computational Intelligence Approach for Predicting Medical Insurance Cost. Math. Probl. Eng. 2021; 2021:1162553. doi: 10.1155/2021/1162553. [CrossRef] [Google Scholar]
[8] Cevolini A., Esposito E. From Pool to Profile: Social Consequences of Algorithmic Prediction in Insurance. Big Data Soc. 2020;7 doi: 10.1177/2053951720939228. [CrossRef] [Google Scholar]
[9] Van den Broek-Altenburg E.M., Atherly A.J. Using Social Media to Identify Consumers' Sentiments towards Attributes of Health Insurance during Enrollment Season. Appl. Sci. 2019;9:2035. doi: 10.3390/app9102035. [CrossRef] [Google Scholar]
[10] Health Insurance Premium Prediction with SVM & DNN models. [(accessed on 9 May 2018)]. Available online: https://thecleverprogrammer.com/2021/10/26/healthinsurance- premium-prediction-with-svm and dnn models /

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details