ISSN No:-2456-2165

Interpreting the Premium Prediction of Health Insurance Through Random Forest Algorithm Using Supervised Machine Learning Technology

V.Srinivasa Rao [1]
[1]Asst. Professor, NRI Institute of Technology, A.P, India-521212

M. Iswarya[2]
[2]UG Scholar, Dept. of IT, NRI Institute of Technology, A.P, India-521212

SK. Ameer Hamza[3]
[3]UG Scholar, Dept. of IT, NRI Institute of Technology, A.P, India-521212

B. Satish[4]
[4]UG Scholar, Dept. of IT, NRI Institute of Technology, A.P, India-521212

Abstract:- In this study, we examine individual insurance amounts using health data. The performance of these algorithms has been compared using the three regression models employed in this study: multiple linear regression, decision tree regression, and decision tree regression. The dataset is used to train the models, and the training then assists in producing more predictions. Later, the model will be tested and verified by comparing the anticipated quantity with the actual data. These models' accuracy levels will then be compared. The decision tree and linear regression are outperformed by the random forest regression algorithm, according to the analysis. It enables a person to understand the required amount based on their health situation. They might examine any health insurance company, their plans, and the benefits while keeping in mind the anticipated amount from the project. Later, the predicted amount will be compared with the real amount. This can also be quite beneficial to someone who wants to concentrate more on the useful aspects of insurance than the health-related ones. In addition, most people are susceptible to being duped regarding the cost of insurance and may unnecessarily purchase expensive medical coverage. This project does not provide the precise sum needed by any health insurance provider, but it does provide a general sense of the sum needed by an individual for their personal health insurance. Prediction is inaccurate and does not apply to any organization; therefore, it should not be the only factor considered when choosing a health insurance plan. First, estimating the cost of health insurance is extremely beneficial and helps in better examining the amount required so that a person can be confident that the amount he or she is going to justify It can also provide you with a wonderful idea for maximizing your health insurance profits.

***Keywords*:-** *Health Insurance Premium Prediction, Linear Regression, Decision Tree Regression, Multiple Regression Algorithm, Machine Learning, Python, Deep Learning, Insurance Amount Prediction, Random Forest Regression Algorithm.*

I. INTRODUCTION

The goal of this exercise is to examine various features to see how they relate to one another and to plot a multiple linear regression based on characteristics of people like age, physical or family condition, and location against their current medical expenses in order to predict future medical expenses of people and assist medical insurance companies in determining how much to charge for coverage. The project's primary objective is to provide people with a general sense of the amount needed based on their personal health status. They can then adhere to any health insurance company's policies and benefits while taking the anticipated funding from our project into consideration. This can assist someone in concentrating more on the health-related aspects of insurance as opposed to the pointless ones. These days, health insurance is virtually always required, and almost all of them are connected to a public or commercial health insurance organisation. The variables affecting the cost of insurance vary from business to business. Additionally, relatively few individuals in rural areas are aware that the Indian government offers free health insurance to those who fall below the poverty level.

It is a highly complicated process, and some rural residents opt to forgo health insurance altogether or only purchase a small amount of private coverage. In addition, consumers can be duped into purchasing expensive health insurance unnecessarily if they are given false information about the cost of the policy. This project does not provide the precise sum needed by any health insurance provider, but it does provide a good indication of the cost associated with a person's own health insurance.

Prediction is unreliable and ungoverned by any corporation, so it should not be the only factor considered while choosing a health insurance policy. Early health insurance cost estimation might help with the required amount. Where a person can ensure that the amount, they choose is appropriate. Additionally, it can give guidance on how to get more advantages from health insurance. This study aims to provide a person with an understanding of the

ISSN No:-2456-2165

required amount based on their personal health situation. Later, customers can look at any health insurance provider and their plans and advantages while taking the anticipated funding from our initiative into consideration.

Large-scale data acceleration has been proposed using a variety of methods. For the greatest performance in a variety of applications, some clustering techniques have been developed over the past few decades. There are numerous distinct rule discovery algorithms, thanks to earlier research. The many clustering algorithms can be divided into four categories: hybrid k-means, Parkinson's disease, situation understanding for intelligent online learning platforms, and artificial neural networks. Based on how closely they relate to one another, the initial centroids of the K-means clustering algorithm are generated at random from k points. Directional information is contained in the circular k-means (CK-means) cluster vectors. A new competitive k-means algorithm was discovered through research to address the inconsistent outcomes of traditional k-means, which scale poorly for very large data sets. In order to determine the ideal number of clusters for the k-means algorithm, Kumar conducted research to present a taxonomy of clustering techniques. He also investigated combining machine learning and intelligent systems with k-means. [Fig.1]

II. TECHNOLOGIES USED

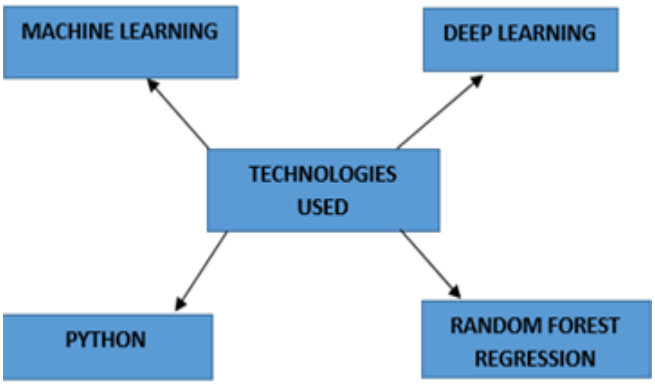

Fig 1 Technologies Used

➢ *Machine Learning:*

Internet search engines, spam-sniffing email filters, websites that offer personalised advice, banking software that can spot unusual transactions, and numerous mobile apps that use voice recognition are all examples of applications that use machine learning. A subfield of artificial intelligence and computer science called "machine learning" uses data and algorithms to mimic how people learn while slightly increasing the accuracy of the results. These days, the technology has a wide range of potential applications, some of which have higher stakes. Future developments will significantly affect this society and could support the UK economy.

Machine learning, for instance, can give us readily available "personal assistants" to help us manage our lives; it might significantly improve the transportation system by utilising autonomous vehicles; and it could also significantly improve the healthcare system by enhancing disease diagnosis or tailoring therapy. In security applications, such as examining and analysing email or internet usage, machine learning [Fig. 2] can also be used. These and other technological applications need to be investigated right away, and steps will be taken to ensure that they are used for the benefit of society. While machine learning is distinct from robotics, there are several areas where they intersect.

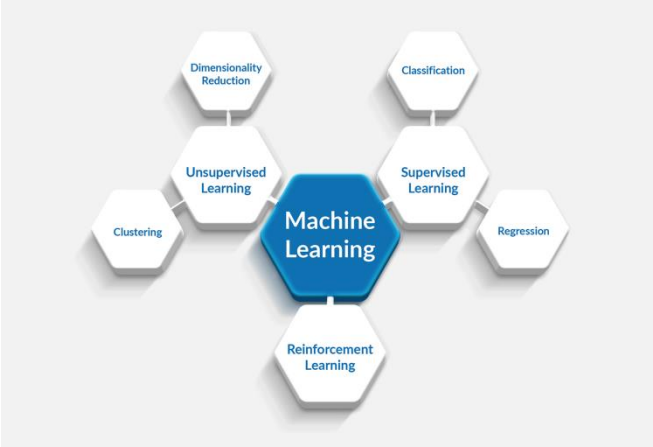

Fig 2 Machine Learning

➢ *Deep Learning:*

Machine learning, which is essentially a neural network with three or more layers, is divided into deep learning and machine learning. These neural networks aim to mimic how the human brain operates by making it capable of quickly learning from massive volumes of data. The accuracy of predictions made by a neural network with a single layer can still be improved and refined with the inclusion of hidden layers. Deep learning powers a variety of artificial intelligence programmes and tools that advance automation by carrying out mental and physical tasks without the need for human intervention. Technology based on deep learning extends beyond common goods and services. [Fig.3]

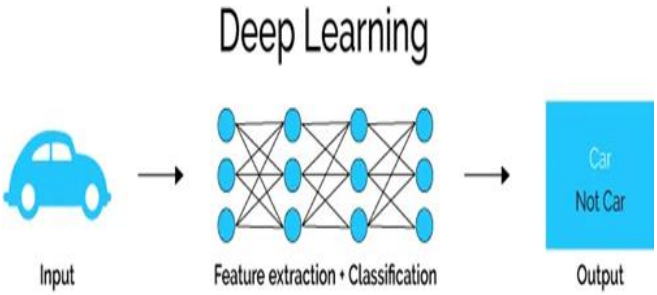

Fig 3 Deep Learning

➢ *Python:*

Guido van Rossum created Python, a high-level, interpreted programming language. It was primarily created to offer a language with a very simple and easy grammar that is easy to read and understand. Numerous programmers started to gradually cling to Python for coding because of the language's shorter codes and ease of writing. Additionally, it contains built-in features and can work as procedural, functional, or object-oriented programming.

ISSN No:-2456-2165

Additionally, it is a platform-neutral programming language. As a result, it is free and open source, has a large library of support, can be used for a wide range of tasks, and many programmers find it to be far simpler to learn and use than many other languages. Additionally, it features built-in memory management strategies and exception handling. It is short, dense, and dynamically typed; thus, there are no declarations. Indentation is the most important aspect of Python since it controls how statements flow. Artificial intelligence is a feature of Python [Fig. 4] that makes it useful in a variety of industries. Additionally, it serves as the foundational language for the Raspberry Pi and is used in game development and information security. The best programming language is Python, which is great. Although it is very easy to read, it is also incredibly forceful.

Fig 4 Python

➢ *Random Forest Regression:*

Among the supervised learning techniques, Random Forest is a well-known machine learning technique. In machine learning, it is applied to problems involving regression as well as classification. It is based on the idea of ensemble learning, which is a method of combining different classifiers to solve a difficult problem and enhance the performance of the model.

As suggested by its name, Random Forest is a classifier that uses several decision trees on different subsets of the input dataset and averages the results to increase the dataset's predicted accuracy. Instead of using a single decision tree for planning, the random forest uses forecasts from each tree and predicts the outcome based on which predictions received the most votes.

Therefore, for forecasting the cost of health insurance, random forest regression [Fig. 5] outperforms linear, multiple, and decision tree regression algorithms.

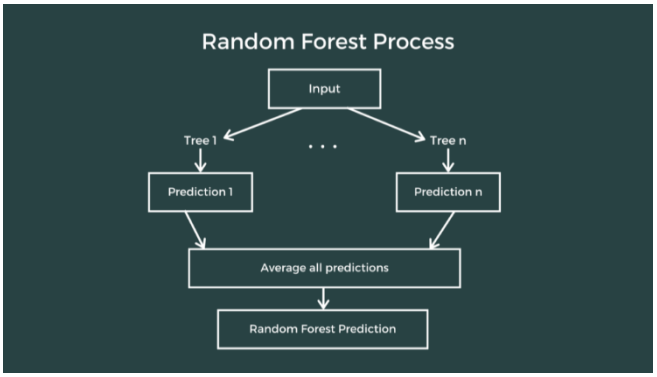

Fig 5 Random Forest

III. SOFTWARE REQUIREMENTS SPECIFICATION

SRS is a comprehensive description of how the system should function. It is typically approved at the conclusion of the requirements engineering phase. It outlines how software systems will communicate with all internal hardware and modules, as well as with other programmes and human users, in a variety of situations that are similar to real-world ones.

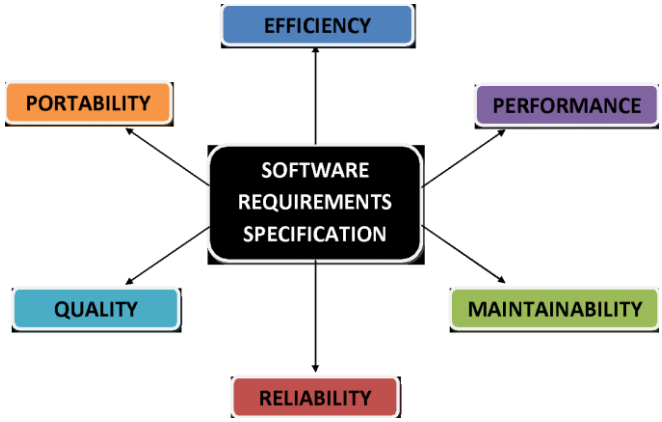

Fig 6 SRS

- *Reliability*:

It is more reliable. It can perform both regression and classification tasks easily. A random forest brings out good predictions that can be understood easily. It can handle huge datasets efficiently. This algorithm provides a higher level of accuracy in predicting the outcomes over the decision tree algorithm. [Fig.6]

- *Quality*: The quality of this project is good and it is very efficient.

- *Maintainability*:

Maintenance of software will be clean and done by the administrator keeps the information safe without any failure or error.

- *Efficiency*: It would be more efficient for users to use it. It provides a good prediction for health insurance.

- *Portability*: It should be portable on any system.

- *Performance*: Performance is good and efficient because it would have done a good work to the users.

IV. EXISTING SYSTEM

A linear regression, multiple regression, and decision tree regression methodology is used in the existing system to predict health insurance premiums. Linear regression is a very sensitive tool to use in the current system. It could have an impact on those using the system. In addition, a protracted and difficult analysis and calculation process is involved. In the current system of predicting health insurance premiums, it also does not fit complex datasets properly.

ISSN No:-2456-2165

➢ *Linear Regression:*

The primary purpose of linear regression analysis is to forecast the value of a variable based primarily on the value of another variable. The dependent variable is the one that needs to be predicted. The variable for which you are attempting to predict the value of the other variable is referred to as the independent variable. The mathematical formula used in linear-regression models is simple to understand and can be used to make predictions. The business world and academic research both greatly benefit from the use of linear regression.

➢ *Multiple Regression:*

A single dependent variable can be analyzed in relation to a few independent variables using the statistical technique known as multiple regression. The main goal of multiple regression analysis is to use known independent variables to predict the value of a single known dependent variable. The use of multiple regression analysis enables researchers to evaluate the significance of each predictor to the relationship as well as the strength of the relationship between an outcome and various predictor variables, frequently with the effect of other predictors being statistically eliminated.

➢ *Disadvantages of Exixting Sysytem:*

- *The disadvantage in this existing system is it does not give the exact amount of prediction.*
- *Its efficiency is very less.*
- *The prediction process is slow.*
- *Linear regression is noise and overfitting.*
- *In this existing system using linear regression is more sensitive.*
- *Another algorithm multiple regression is also not good at predicting the accurate result.*

V. PROPOSED SYSTEM

In this project, we suggest a random forest algorithm to boost system performance. The supervised learning method includes the very well-known machine learning algorithm Random Forest. In machine learning, it is used for both classification and regression issues. Random Forest is a classifier that uses several decision trees on numerous subsets of the input dataset and averages the results to increase the dataset's predictive accuracy. So the algorithm with the best performance for this task is the random forest algorithm.

➢ *Random Forest Regression Alogorithm:*

An algorithm for machine learning called Random Forest primarily uses the supervised learning approach. It is primarily used for classification and regression issues in machine learning. It is largely based on the idea of ensemble learning, which is the process of combining various classifiers to solve a very complex problem and enhance the performance of the model. As the name indicates, Random Forest is a classifier that consists of a few decision trees on many different subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. On behalf of relying on one decision tree, the random forest takes the predictions from each tree and, based on that, predicts most of the predictions, and it also predicts the final output. Therefore, random forest regression outperforms linear, multiple, and decision tree regression in terms of accuracy in predicting the cost of health insurance.

➢ *Advantages of Proposed System:*

- *The proposed system employs the Random Forest algorithm, which is capable of both classification and regression tasks.*
- *It makes accurate and reliable predictions here.*
- *It is simple to comprehend.*
- *In this case, it can effectively handle big and large datasets.*
- *The random forest algorithm in the proposed system offers a high level of accuracy.*
- *It offers a useful way to handle the missing data and is flexible and simple to use.*
- *The health care industry primarily uses the random forest algorithm.*

VI. SYSTEM ARCHITECTURE

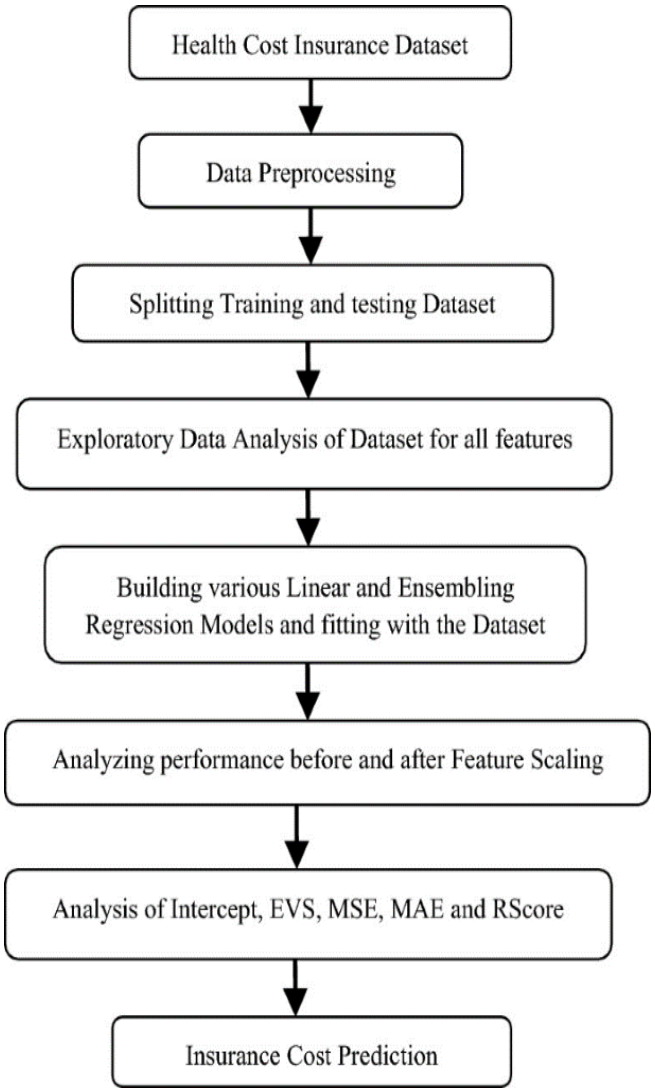

Fig 7 System Architecture

VII. FUTURE SCOPE

The health insurance premium prediction for the future provides the precise predicted amount, allowing people to easily understand their status. The prediction of premium amounts focuses more on an individual's health than on the policies of other insurance companies. These models can be used to predict the precise amount of data that will be collected in the upcoming years. Additionally, it may aid insurance companies in operating. [Fig.7]

VIII. CONCLUSION

To forecast changes in insurance based on attributes, this project employs a variety of machine learning regression models. A health insurance company billed each user, and this model is used to forecast the insurance claim each user will submit. Businesses' efficiency will increase as a result of this. The model can handle a huge amount of data. I hope you enjoyed reading this article on predicting health insurance premiums using machine learning and the random forest regression algorithm. To make health insurance operations simpler, artificial intelligence and machine learning are very capable of analysing and estimating large volumes of data. For insurers, the predicted health insurance premiums will result in time and cost savings. Our model had a 92.72% accuracy rate.

REFERENCES

[1]. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., "Scikitlearn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, Oct 2011.
[2]. E. Wang; and G. Gee, "Larger Issuers, Larger Premium Increases: Health insurance issuer competition post-ACA," 2015
[3]. C. C. a. A. Semanskee, "Analysis of UnitedHealth Group's Premiums and Participation in ACA Marketplaces," 2016.
[4]. Ng, "Machine learning for Housing Price Prediction Mobile Application," Masters, Department of Computing, Imperial College London, 2015
[5]. Sturm, Roland. The effects of obesity, smoking, and drinking on medical problems and costs. Health affairs 21, no. 2 (2002): 245-253.
[6]. Sturm, Roland, Ruopeng an, Josiase Maroba, and Deepak Patel. The effects of obesity, smoking, and excessive alcohol intake on healthcare expenditure in a comprehensive medical scheme. South African Medical Journal 103, no. 11 (2013): 840-844
[7]. Kim, David D., and Anirban Basu. Estimating the medical care costs of obesity in the United States: systematic review, meta-analysis, and empirical analysis. Value in Health 19, no. 5 (2016): 602- 613.\
[8]. Han, Kimyoung, Minho Cho, and Kihong Chun. Determinants of health care expenditures and the contribution of associated factors: 16 cities and provinces in Korea, 2003-2010 Journal of Preventive Medicine and Public Health 46, no. 6 (2013): 300.
[9]. Han, Kimyoung, Minho Cho, and Kihong Chun. Determinants of health care expenditures and the contribution of associated factors: 16 cities and provinces in Korea, 2003-2010 Journal of Preventive Medicine and Public Health 46, no. 6 (2013): 300.
[10]. Sharma, Ashish, Ashish Sharma, and Anand Singh Jalal. "Distance-based facility location problem for fuzzy demand with simultaneous opening of two facilities." International Journal of Computing Science and Mathematics 9.6 (2018): 590-601.
[11]. Singh, Anshy, Shashi Shekhar, and Anand Singh Jalal. "Semantic based image retrieval using multi-agent model by searching and filtering replicated web images." 2012 World Congress on Information and Communication Technologies. IEEE, 2012.
[12]. Shekhar, Shashi, et al. "A WEBIR crawling framework for retrieving highly relevant web documents: evaluation based on rank aggregation and result merging algorithms." 2011 International Conference on Computational Intelligence and Communication Networks. IEEE, 2011.
[13]. Varun K L Srivastava, N. Chandra Sekhar Reddy, Dr. Anubha Shrivastava, "An Effective Code Metrics for Evaluation of Protected Parameters in Database Applications", International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.1.3, 2019. doi.org/10.30534/ijatcse/2019/1681.32019
[14]. Prasad, K.S., Reddy, N.C.S. & Puneeth, B.N. A Framework for Diagnosing Kidney Disease in Diabetes Patients Using Classification Algorithms. SN COMPUT. SCI. 1, 101 (2020)

BIOGRAPHIES

V. Srinivasa Rao is currently working as a Asst. Professor in the Department of Information technology at NRI Institute of technology, Pothavarappadu, Agiripalli, Krishna(dist), India.

M. Iswarya is currently studying B. Tech with specification of Information Technology in NRI Institute of Technology. She done a mini project Health insurance premium prediction.

ISSN No:-2456-2165

SK. Ameer Hamza is currently studying B. Tech with specification of Information Technology in NRI Institute of Technology. He had done a mini project Health insurance premium prediction.

B. Satish is currently studying B. Tech with specification of Information Technology in NRI Institute of Technology. He had done a mini project Health insurance premium prediction.