International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 09 | Sep 2023 www.irjet.net p-ISSN: 2395-0072

Minimum Health Insurance Premium prediction using health parameters

P Sudheer Benarji[1], Kollipara Praveen Kumar[2], Manishetty Sushanth[3], Meghana Gogi[4], Donuru Neeraja[5]

[1]Professor, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad
[2345]Under Graduate Student, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad

---

Abstract - *Health insurance is important nowadays, and almost every individual is having a link with the government or private health insurance companies. Factors determining the amount of insurance vary from company to company. People in rural areas are not aware of the fact that the government of India provide free health insurance to people below poverty line. It is very complex procedure and some rural people either buy some private health insurance or will not invest money in health insurance at all. People can be fooled easily about the minimum amount of the insurance and they may land into expensive health insurances. Idea is to analyze the personal health data to predict insurance amount for individuals. The project uses health related parameters to predict the minimum insurance premium amount.*

Key Words: (Health Parameters, Prediction, Premium, Pipelines.

1. INTRODUCTION

In providing health insurance premiums to patients most of the premium-providing companies are asking the patients about the minimum amount they require. This is a critical task for patients to choose, because it may lead to selecting an amount more than they require or may select a lesser amount. Most of the patients don't have any idea about how much it cost to cure the immediate disease that may occur due to their current health condition. The aim of this project is to allow a person to get an idea about the necessary amount required according to their own health status. Later they can approach any health insurance company and their schemes & benefits keeping in mind the minimum amount predicted from our project. The project gives enough idea about the amount associated with an individual for his/her own health insurance. Our project does not give the exact amount required for any health insurance company but gives enough idea about the amount associated with an individual for his/her own health insurance. The healthcare sector produces a very large amount of data related to patients, diseases, and diagnosis, but since it has not been analyzed properly, it does not provide the significance which it holds along with the patient healthcare cost

2. Literature Review

Mohamed Hanafy (2021) Insurance is a policy that eliminates or decreases loss costs occurred by various risks. Machine learning (ML) for the insurance industry sector can make the wording of insurance policies more efficient. And we will compare the results of models, for example, Multiple Linear Regression, Support Vector Machine, Random Forest Regressor, etc. All the health parameters are not considered. This paper offers the best approach to the Stochastic Gradient Boosting model with an MAE value of 0.17448, RMSE value of 0.38018and R-squared value of 85.8295.

Ch Anwar Ul Hassan, Jawaid Iqbal [2021] A medical insurance cost dataset is acquired from the KAGGLE repository for this purpose, and machine learning methods are used to show how different regression models can forecast insurance costs and to compare the models' accuracy. Considered many Machine Learning algorithms for better accuracy. The results shows that the Stochastic Gradient Boosting (SGB) model outperforms the others with a cross-validation value of 0.0.858 and RMSE value of 0.340 and gives 86% accuracy.

Ayushi Bharti, Lokesh Malik [2022] In this study they have used the Kaggle dataset of Medical Insurance cost and trained our model on attributes like age, sex, BMI, children. smoker, region which contains 1070 training instances and 267 testing instances. The data was used in the model after pre-processing it. The best rating was received by using Random forest. In the future, we can practice more strategies to get even more accurate scores for medical insurance costs.

Preet Jayendra kumar Modi, Vraj Jatin Naik (2021) It is a web application which is developed for tracking the details of the insurance policy, customer details and company details. Our System gives us a predicted value of premium by looking at your data and our system also has other functionalities like policy comparison, premium payment, etc.

Shyamala Devi , Swathi Pillai (2021) The raw dataset and the feature scaled dataset is applied to all the Ensembling Regression models and the performance is analyzed through intercept, MAE, MSE, R2Score, and EVS. Anova Test Reults shows that the variable 'region' does not influence the target as the F-statistic value is 0.14. All health parameters are not

considered for prediction of the premium price. Experimental results show that polynomial regression is achieving 88% of R2Score before and after feature scaling.be used. Other font types may be used if needed for special purposes.

Ghosh Madhumita (2022) Blockchain technology is based on a sequence of blocks, where each block carries a certain amount of information. Medical records can be cryptographically secured in the health insurance ecosystem with blockchain technology. Here, blockchain technology model is used to create a user interface for storing data block wise. Multiple Linear Regression algorithm gives the better result.

Nidhi Bhardwaj, Rishabh Anand (2020) It analyse the personal health data to predict insurance amount for individuals. The predicted amount was compared with the actual data to test and verify the model. Gradient boosting is best suited in this case because it takes much less computational time to achieve the same performance metric, though its performance is comparable to multiple regression. The accuracy of the model is 14 more but didn't considered all health parameters for prediction.

Chaparala Jyothsna, K. Srinivas, Bandi Bhargavi (2022) The goal is to anticipate a person's insurance costs and to identify patients with health insurance policies and medical information, regardless of whether or not they have any health problems. it was determined that Gradient Boosting was the most accurate of all the methods, with an accuracy of 87 percent. Finally, using the best model, the Telegram-integrated chatbot is trained with instructions to communicate with the user and estimates the insurance premium.

Keshav Kaushik , Akashdeep Bhardwaj (2022) There is a direct link between the insurer and the policyholder when the distance between an insurance business and the consumer is reduced to zero with the use of technology, especially digital health insurance. This research trained and evaluated an artificial intelligence network-based regression based model to predict health insurance premiums. The experimental results displayed an accuracy of 92.72%.

Kashish Bhatia, Shabeg Singh Gill, Navneet Kamboj, (2022) Features in the dataset that are used for the prediction of insurance cost include: Age, Gender, BMI, Smoking Habit, number of children etc. We used linear regression and also determined the relation between price and these features. Various health parameters are not considered for prediction. We trained the system using a 70-30 split and achieved an accuracy of 81.3%.

3. Existing system

We are on a planet full of threats and uncertainty. The Threats may cause a risk of death or health to the people. Insurance is a policy that eliminates or decreases the cost occurred by various risks. The count of people taking insurance policies are increasing drastically but people are unaware of the amount they should get from the policy. The existing systems are predicting the insurance premium amount. The model is using BMI(Body Mass Index), number of children, age, gender and many more for prediction.

4. Proposed system

Policy companies will ask the clients, how much amount they require and clients makes mistake in choosing the amount. There is a chance that client asks for less amount than required. The main aim of the project is to use wide range of features in order to get accurate results. The features include age, diabetics, chronic disease, surgeries and many other health parameters. The model will be trained with the above features using machine learning algorithms. The model will predict the minimum insurance amount the client required.

5. System Architecture

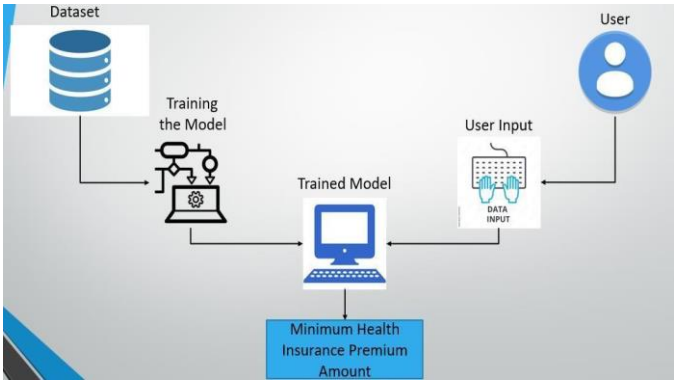

Fig 1: System Architecture

It is a conceptual model that describes the structure, viewpoints, and behavior of a system. A formal description and representation of a system that is organized in a way that facilitates reasoning about the system's structures and behaviours is known as an architecture description. A system architecture consists of system components and created sub-systems that work together to implement the overall system.

There are four main tiers in our project. The first and important tier is ML Datasets which were brought from various resources and are used to train the model so that accurate predictions are made.

The second tier is the logic tier. This tier is used to train the model using various algorithms which are imported using the sklearn module. The Random Forest algorithm is used to train and fit the model.

The third tier is the user tier. This tier allows users to give input to the trained model. Users enter various health parameters like age, diabetes, any transplants, any chronic

diseases, height, weight, any allergies, number of major surgeries, etc.

The fourth tier is the trained tier. This tier takes input from the user and predicts the minimum health insurance premium amount.

6. Modules

A. Pipeline:

Various machine learning algorithms are considered and stored in pipeline. All the machine learning algorithms that are available in the pipeline are used to train the model with the dataset.

B. Model Selection:

RMSE and R2 score values for the above models are calculated. Then the model with best figures of RMSE and R2 score is considered as final model for the prediction.

C. Prediction:

Various health inputs are taken from the user and the above selected model is being used to predict the minimum health insurance premium amount.

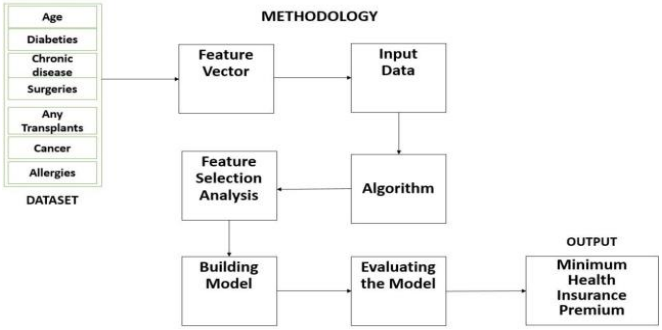

Fig -2: Methodology

7. UML Diagrams

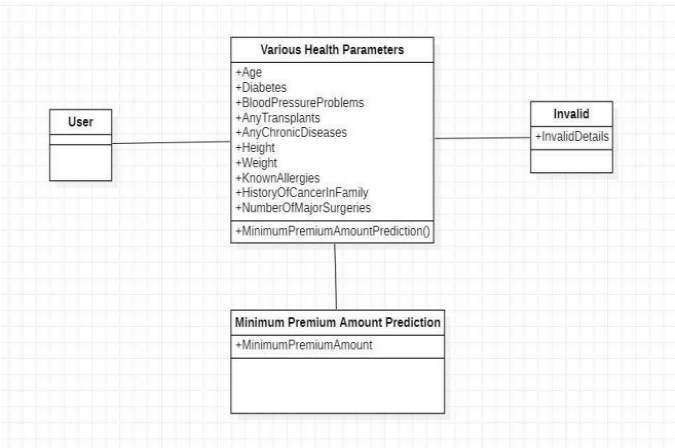

Fig 3: Class Diagram

In the class diagram, the main class is various health parameters which are connected to user and minimum premium amount prediction. It contains the attributes like Age, Diabetes, Blood Pressure Problems, any Transplants, any chronic diseases, Height, Weight, Known Allergies, History of cancer in family, Number of Major surgeries. The function present in the model is used to analyse the parameters and predict the insurance amount. The user class is used to provide input and display.

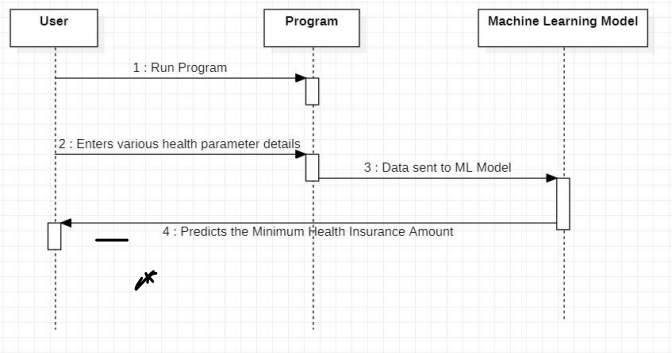

Fig 4: Sequence diagram

The lifelines are:

- User
- Program
- Machine Learning model

Firstly, the datasets are fetched into the Jupyter Notebook. Then the datasets are trained and tested using various algorithms. When the user gives health parameters as input, the model analyses parameters and predicts the minimum health insurance amount.

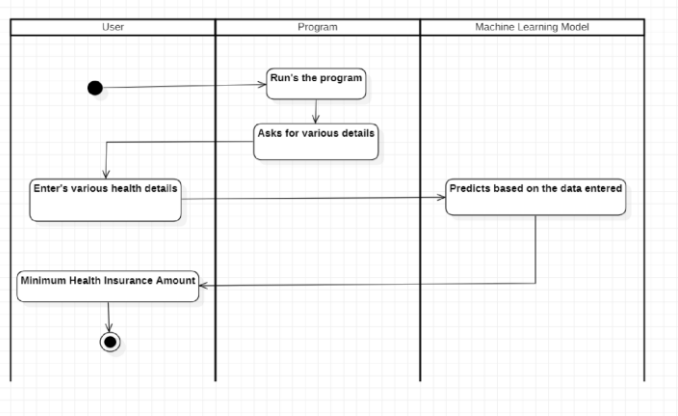

Fig5: Activity Diagram

This activity diagram represents the activity of the whole system. The activity starts with the user, the user enters the health parameters. Then the model analyses parameters and predicts the minimum health insurance amount.

8. RESULTS

Data analysis performed on the collected data set. By checking on basic parametric requirements. To do that so first checked whether the collected data set is perfect.

```
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Age                         986 non-null    int64
 1   Diabetes                    986 non-null    int64
 2   BloodPressureProblems       986 non-null    int64
 3   AnyTransplants              986 non-null    int64
 4   AnyChronicDiseases          986 non-null    int64
 5   Height                      986 non-null    int64
 6   Weight                      986 non-null    int64
 7   KnownAllergies              986 non-null    int64
 8   HistoryOfCancerInFamily     986 non-null    int64
 9   NumberOfMajorSurgeries      986 non-null    int64
 10  PremiumPrice                986 non-null    int64
dtypes: int64(11)
memory usage: 84.9 KB
```

Fig - 6: Data set info

To recheck the collected data set, plotted the graph with age and premium. As age increases premium amount should also increase.By below picture it is verified accordingly.

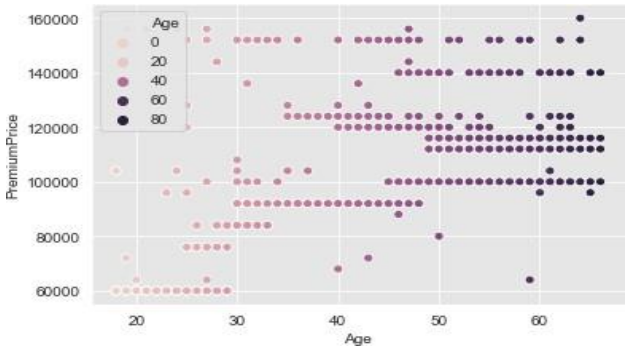

Fig 7(a): Age vs Premium Price

After training the model it's again verified with the predicted data with age as the parameter.

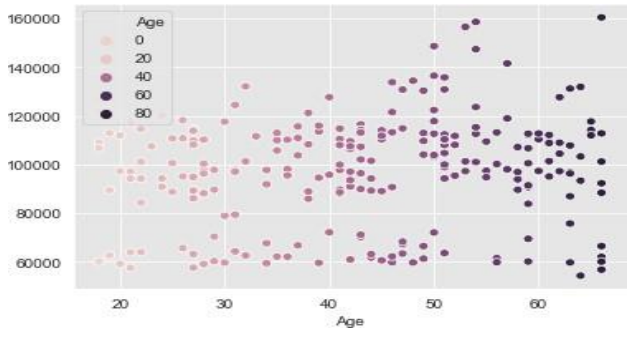

Fig. 7(b): Age vs Premium Price (Predicted data)

So by above 2 figures it's been verified that the data is collected and trained perfectly. To make it more user friendly, implemented the code by asking the user inputs and predicted the minimum health premium accordingly.

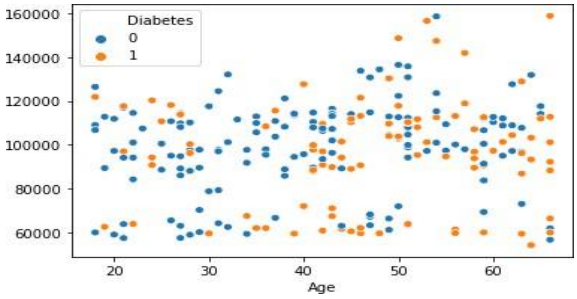

Fig. 8(a): Age vs Premium Price with diabetes

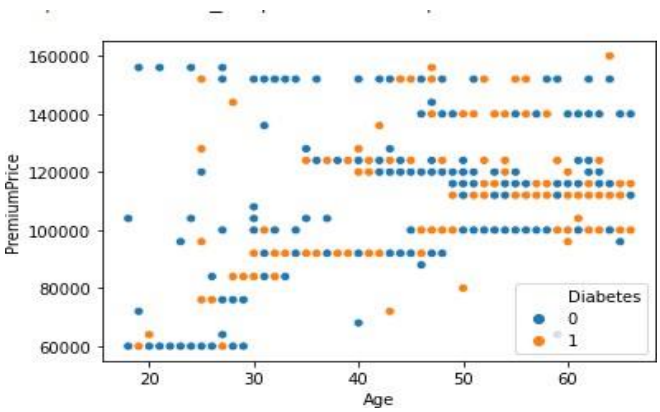

Fig. 8(b): Age vs Premium Price with diabetes (Predicted Data)

```
Enter your Age 45
Do you have Diabetes ?- Enter 1 if YES otherwise 0 -7
Diabetes value should be 0 or 1
Do you have Diabetes ?- Enter 1 if YES otherwise 0 0
Do you have any Blood Pressure Problems ?- Enter 1 if YES otherwise 0 0
Do you went through any Transplants ?- Enter 1 if YES otherwise 0 0
Do you have any Chronic Diseases ?- Enter 1 if YES otherwise 0 0
Enter your Height in cm 155
Enter your Weight in kgs 57
Do you have any Allergies ?- Enter 1 if YES otherwise 0 0
Any of your ancestors suffered from cancer ?- Enter 1 if YES otherwise 0 0
Enter number of major surgeries happened to you 0
[99920.]
```

Fig. 9: Final Result

9. CONCLUSIONS

Use of machine learning provided many advantages in predicting the minimum health insurance amount. The model saved time by preventing all the complex calculations

and giving the minimum health insurance amount in less computational time. The model also saved the money of the policy holders and the insurers.

We used 7 essential attributes and used seven regression techniques particularly Linear Regression, K Neighbours Regressor, Decision Tree Regressor, Random forest, Gradient Boosting Regressor, LGBM Regressor and XGB Regressor. The data was used in the model after pre-processing it. The best rating was received by using Random forest. We got even more accurate scores for medical insurance premium amount.

REFERENCES

[1] Mohamed Hanafy, Assiut University" Predict Health Insurance Cost by using Machine Learning and DNN Regression", Research gate, 348559741,2021.

[2] Shyamala Devi, Swathi Pillai , Vel Tech "Linear and Ensembling Regression Based Health Cost Insurance Prediction Using Machine Learning ",Research gate, 353231212,2021.

[3] Ch Anwar Ul Hassan, Jawaid Iqbal"A Computational Intelligence Approach for Predicting Medical Insurance Cost Hindawi journal,1162553,2021.

[4] Kashish Bhatia, Shabeg Singh Gill, Navneet Kamboj, Manish Kumar" Health Insurance Cost Prediction using Machine Learning" IEEExplore,984201,2022.

[5] Chaparala Jyothsna, K. Srinivas, Bandi Bhargavi" Health Insurance Premium Prediction using XGboost Regressor" IEEExplore, 9793258,2022.

[6] Preet Jayendrakumar Modi, Vraj Jatin Naik "Insurance Management with Premium Prediction " IJRASET,2022

[7] Ghosh Madhumita "Health Insurance Premium Prediction using Blockchain Technology and Random Forest Regression Algorithm" IJOEST article,346,2022.

[8] Keshav Kaushik , Akashdeep Bhardwaj "Machine Learning-Based Regression Framework to Predict Health Insurance Premiums" NCBI,35805557,2022.