2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) | 979-8-3503-1920-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSES60034.2023.10465334

An Efficient Health Insurance Prediction System using Machine learning

A.Vinora[1], V. Surya[1], Dr.E.Lloyds[2], B.Kathir Pandian[1], R.Nancy Deborah[1] and A. Gobinath[1]

[1]Department of Information Technology Velammal College of Engineering and Technology Madurai, India
[2]Department of Psychiatry Governemnet Sivagangai Medical College Sivagangai, India

E-mail : vnr@vcet.ac.in suryaranji7542@gmail.com, doclloyds@gmail.com, kathirpadma24@gmail.com, rnancydeborah@gmail.com, agn@vcet.ac.in

Abstract- Health insurance policies provide financial assistance to cover medical expenses and mitigate the financial impact of illnesses. Various factors contributes the cost of healthcare and health insurance. Predicting health insurance costs early can assist in determining the appropriate coverage amount and identifying potential benefits. ML can improve the efficiency of insurance policy in the insurance industry. In healthcare, ML algorithms excel at forecasting high-cost medical expenses. Machine learning can enhance the effectiveness of insurance program language within the insurance sector. In the Insurance business Machine Learning (ML) algorithms help to anticipate and make the process effortless between the policyholder and the insurer as the distance has been zeroed with the help of technology in the form of digital insurance. In contrast to traditionally available health insurance Digital insurance with the help of AI, ML help to predict the value required to be insured of a specific individual based on factors such as age, gender, BMI, Smoking routine, geographical location, and the number of children. Hence Model using XGBRF Regressor is trained to foretell the minimum insurance premium for an individual.

Keywords: Machine learning (ML), Health insurance, Prediction, XGBRFRegressor.

I. INTRODUCTION

In a world filled with Uncertainty, all forms of life are subjected to a variety of risk factors. The primary concern is the threat caused due to illness or prolonged illness leading to death. Hence, the Effects of illness worsen with the financial inability of a person to accommodate Medicare. This issue can be foreseen and effectively handled with the help of investing in health insurance. A Policy that helps to cover all the expenses related to Medicare is called health insurance after paying the premium for a specific tenure of time. It helps to reduce or eliminate the expenses incurred by various risk factors. The major difficulty with handling health insurance is the inability to predict a specific premium for yearly installments. Hence by using Machine learning algorithms a predefined amount can be derived to give the user an estimate to be invested for healthcare to their future benefit. Medical insurance helps in the deduction of expenses incurred due to various health-related risk factors and forecasting a healthcare cost helps to improve accountability in the medical sector. The Premium of an individual varies and can be estimated on a collection of factors such as age, gender, locality, etc. Example: The premium amount of individuals older than 50 years is more when compared to young adults since the probability of health complications is high in them.

Medical expenses are a significant and frequently recurring cost in human existence. An individual's lifestyle and various physical factors determine the illnesses or conditions they may experience, which in turn impact their healthcare costs.

The increasing occurrence of unpredictable diseases and accidents has led to a substantial rise in the demand for health insurance. Health insurance plans serve as a means to alleviate the financial burden during medical emergencies. Individuals, at all stages of life, must contend with uncertainties and risks. Numerous factors contribute to the determination of healthcare insurance costs. Machine Learning algorithms have been extensively employed across various domains, leveraging extensive data for predictive purposes with notably high accuracy rates. One significant application is the estimation of hospital expenses within the realm of healthcare.This document covers related research, the methodology, the results, and the conclusion.

II.RELATED WORK

Ahmed I. Taloba et.al worked on a multiview learning architecture that helped to predict one of the best methods used for the prediction of medical care expenses incurred due to hospitalization. They studied the various key factors to be considered for medical care and hospitalization with one of them being BMI (Body Mass Index) based on which obesity was determined that serves one of the root causes of major health problems. Out of themachine learning algorithms employed, namely random forest, linear regression analysis, and naive Bayes classifier, linear regression demonstrated the highest accuracy in predicting healthcare costs as a whole. [1].

Keshav Kaushik et.al recommended the integration of ML andAIinto the prediction of healthcare insurance premiums. The model was the key factor for the emergence of digital health insurance that narrows down the possible distance between the consumer and the policy issuer and helps to ease the process of documentation where the services are even at a faster rate but their trained model deployed the use of ANN based regression network that obtained higher levels of accuracy with an r2_score of 0.75 considering the key performance metrics such as age, BMI, Smoker, number of children and geolocation [2].

Ch. Anwar ul Hassan et.al suggested the utilization of machine learning techniques for predicting health insurance premiums. Their study involved a comparison of various algorithms, including Linear Regression, Support Vector Regression, XGBoost, Random Forest Regressor, and KNN. The model predicted higher accuracy rates using SGB on the dataset that was used for training and testing from Kaggle [3].

M. A. Aefa and colleagues present two techniques for parameter estimation in the mixture model: the maximum likelihood method and the Bayesian method. The maximum likelihood method centers on identifying parameter values that optimize the likelihood function, representing the probability of observing the data given those parameters. Conversely, the Bayesian method focuses on determining the posterior distribution of the parameters, which reflects the probability of the parameters in light of the data and additional prior information [4].

The prevalence of obesity and weight gain among the population leads to an increase in health risks which is a major concern and also an essential parameter to be considered to evaluate the health premium [5]

R. A. Ganaie et.al proposed a new distribution called the weighted power Shanker distribution that can model various lifetime data. The paper explores its properties and applications and shows that it outperforms some existing distributions [6].

J. Liu, D. Capurro et.al introduced a new way for parameter calculation of the generalized exponential distribution using record values. The paper derives the likelihood function and compares the method with existing ones using simulation and real data [7].

H. N. Alhazmi et.al introduced a predictive model aimed at estimating healthcare expenses for organizations resembling orphanages in Saudi Arabia. This predictive model has the added benefit of assisting orphanages in forecasting healthcare costs. It aids in determining the approximate government funding required for the welfare of orphans and abandoned children through the utilization of Machine Learning, Business Analytics, Naive Bayes, and Simple Linear Regression. Among these approaches, the combination of Random Forest and business analytics yielded the highest level of accuracy. [8].

M. H. Abu-Moussa et.al applied the weighted power Shanker distribution to three real life data sets related to the lifetime of electronic components, the time to failure of mechanical systems, and the survival time of patients with lung cancer [9].

S. Sana et.alstudied and solved the problem of estimating and predicting the parameters of the generalized exponential distribution using lower record values. The authors compare the maximum likelihood and Bayesian methods using simulation and real data [10].

M. Ravaut, H. Sadeghi, K. K. Leung et al., proposed the model that could be used to identify individuals at high risk of developing diabetes complications. These individuals could then be targeted with preventive interventions, such as lifestyle changes or early medical treatment. The model could also be used to identify individuals who have already developed diabetes complications. These individuals could then be monitored more closely and provided with more intensive care [11].

Jeremy A. Irvin et.al incorporated machine learning along with Social Determinants of Health (SDH) to enhance risk adjustments for effective payment details of a health plan. It often targets budgetary needs and provides better care for high-risk individuals using the linear regression method. Risk modeling is done to provide better healthcare services [12].

Akashdeep Bhardwaj et.al used Artificial Neural Networks (ANN) for Health insurance claim prediction. It was aimed to foretell the anticipated medical claims that would occur in a year for a specific medical health insurance company that helped to include prefix amounts in their annual budgets. The recurrent neural network aided various business organizations in decision making that helps to increase the profit margin of the company To overcome the dilemma of accurate investment aggregate, we can integrate(ML), and Artificial Neural Network (ANN) models to predict these costs and simplify investment in the health insurance field since it is a mandatory investment to be made by any individual [13].

Reliability in prediction also plays a vital role where the growth of a model is unrecognized until validated for a particular data set [14].

L.Hu et.al uses machine learning to identify and rank the determinants of high healthcare costs for breast cancer patients. They compared different models and finds that deep neural network is the best [15].

M. Shyamala Devi et.al proposed a medical insurance prediction model using linear and ensemble regression techniques to provide accuracy. Finding a random target on multiple lines with multiple searchers is the problem that studies and solves. It deployed the use of the UCI machine repository to analyze the data, compare the mechanisms such as polynomial regression, and random forest regression, and obtain results. Random forest mechanism provided higher accuracy results for the particular repository with R2Score evaluated as a parameter before and after scaling [16].

Hong J. Kan et.al used machine learning in risk adjustments towards predicting health insurance costs for adults. The model implemented Lasso regression which provided superior prediction results with high accuracy that serves as a boon for policy providers to help the risk population for the benefit [17].

Based on the related works inputs were taken to incorporate an efficient machine learning algorithm to provide accurate prediction results for estimating the health insurance of an individual.

III. METHODOLOGY

Dataset Description:
The dataset employed in this study was obtained from Kaggle and exhibits no instances of missing or undefined values. It pertains to Insurance Premium Charges within the United States and consists of 1338 rows of insured individuals' data, all of which are valid, with Insurance charges associated with specific attributes of the insured. inTable-I:

TABLE 1 FACTORS TAKEN INTO CONSIDERATION FOR MEDICAL INSURANCE COST PREDICTION

S.no.	Feature Name	Description	Value
1.	Age	Age of the particular person applying for health insurance	Integer value Range: 18 to 64
2.	Sex	Gender	Female, Male
3.	Body mass index (BMI)	Comprehending the human physique: weights significantly divergent from the norm in relation to height	A standardized body weight index (in kg/m^2) derived from the height-to-weight proportion, ideally falling within the range of 18.5 to 25
4.	Children	Children to the applicant(if applicable)	Integer value
5.	Smoker	SmokeHabit.	Smoker,Non-smoker
6.	Region	Location.	Northeast, Northwest, Southeast, Southwest
7.	Charges	Medical costs .	Integer Value

Data Pre-processing:
Data Pre-processing is a necessary process to build an ideal model. It makes the input data suitable for our machine learning model. The Dataset Contains both Categorical (Categorized into some groups) and Numerical values which is difficult for a model to predict. Hence, data are normalized (organizes data for more efficient access), which converts the values in a range between -1.0 to 1.0 or 0.0 to 1.0.

Data Normalization is done by type cast the data into a specified type such as float, double, etc. Type casting is the process of converting a value into another format.

The Categorical values are Smoker, Sex, and Region and the Numerical values are Age, BMI, Children, and charges. All these values were normalized into type float.

Feature Engineering:
It is the process of finding the best features that have a positive influence on the prediction of the target value. We performed Feature Selection with the help of the Correlation Matrix.

Correlation Matrix:
Fig 1.A correlation matrix is essentially a tabular representation illustrating the relationships between different features and target variables. In this context, we calculated the correlation using Pearson's correlation coefficient, determined through the provided formula.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

In this context, "n" denotes the amount of information, "Σx" stands for the cumulative total of the first variable's values, "Σy" represents the cumulative total of the second variable's values, "Σxy" indicates the summation of the product of the first and second values, "Σx2" refers to the cumulative total of the squares of the first value, and "Σy2" signifies the cumulative total of the squares of the second value.

Correlation Matrix for features and target:
Here n=1338 and X values are age, BMI, region, smoker, children, sex and Y value is Charges. With the help of the correlation matrix, we found that the feature ‘region’ is negatively correlated with the charges. So, we removed that feature, since it’s not contributing to the target variable. Some other features like sex, and children contribute less to the target variable. We decided to include those features since they are not making such a huge difference as the feature 'region'.

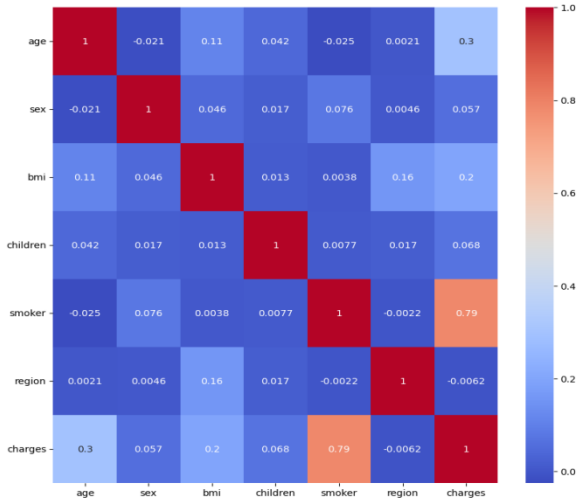

Fig. 1 Correlation matrix

Splitting the input data:
The input data was divided into training and testing sets following the Pareto Principle, allocating 80% of the data to the training part and 20% to the test part. The random nature of this split was employed to ensure the model's impartiality and prevent bias towards specific data. The model was then trained on the training set and assessed using XGBRFRegressor, which combines the Random Forest algorithm with XGBOOST.

A Random Forest serves as a meta-estimator by fitting multiple decision trees for classification on subsets of the dataset, utilizing averaging to address overfitting and improve predictive accuracy. When combined with XGBOOST, it aids in prioritizing the features used to forecast health insurance premiums. It measures the extent to which the output variable can be predicted from the input independent variables and assesses how closely the model replicates observed results based on the degree of variance explained by the model. It's important to note that our model doesn't provide an exact aggregate value for any health insurance company but offers a general estimate of the investment required for an individual's health insurance

.

Fig 2 Depicts the flow of overall working process.

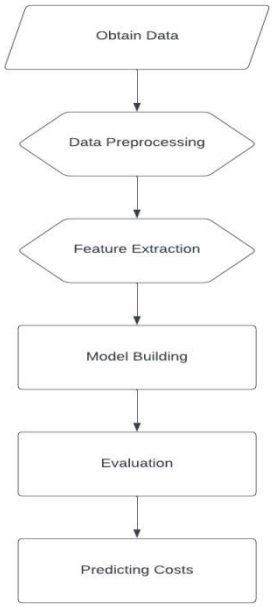

Fig. 2 Working methodology

IV. RESULTS

XGBRFRegressor has been applied to predict the premium incurred for an individual. Based on validation and testing, the model has achieved R2_score= 87.34% in Fig 3, Mean Absolute Error (MAE) = 2511.03 in Fig 4, Mean Squared Error (MSE) = 19429024.51 in Fig 5. This model was implemented on a 12 GB RAM machine with 107 GB storage. The model developed is inexpensive and accessible to all. We have also hosted a website(https://health-insurance-predictor-k19.streamlit.app/) for individuals to get an idea about the premium amount based on input values such as age, gender, BMI, Number of children , smoker and region where the categorical and non-categorical data are converted to float values on which XGBRFRegressor is applied and their health insurance aggregate amount will be predicted (Fig 6). We have run different algorithms on the data set that we have obtained from Kaggle and have got better R2_score can be viewed in the tabular column listed below for XGBRFRegressor in Table II.

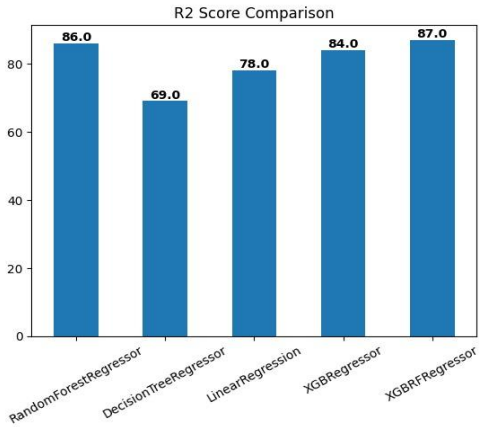

Fig. 3 XGBRFRegressor performance Vs other ML models considering their R2 Score

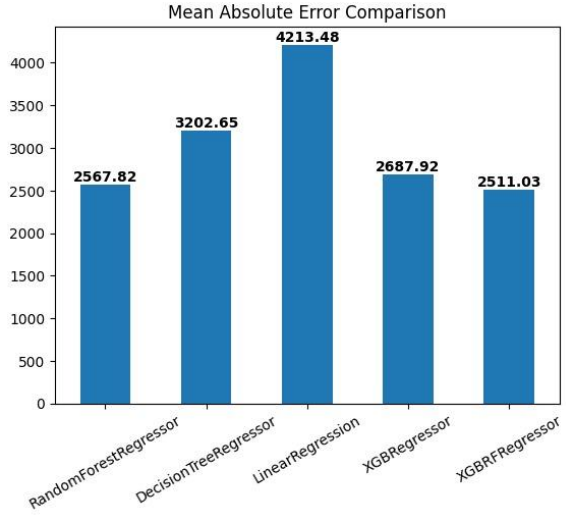

Fig. 4.XGBRFRegressor performance Vs other ML models considering their MAE values

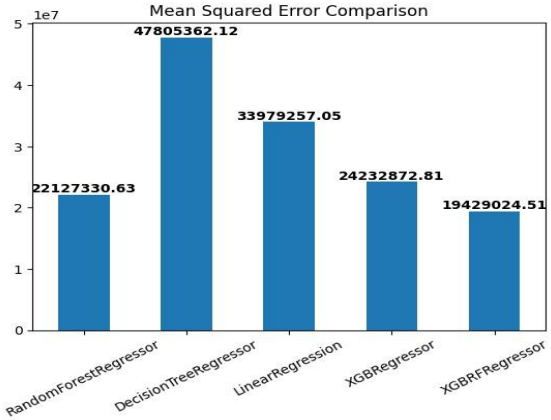

Fig. 5.XGBRFRegressor performance Vs other ML models considering their MSE score.

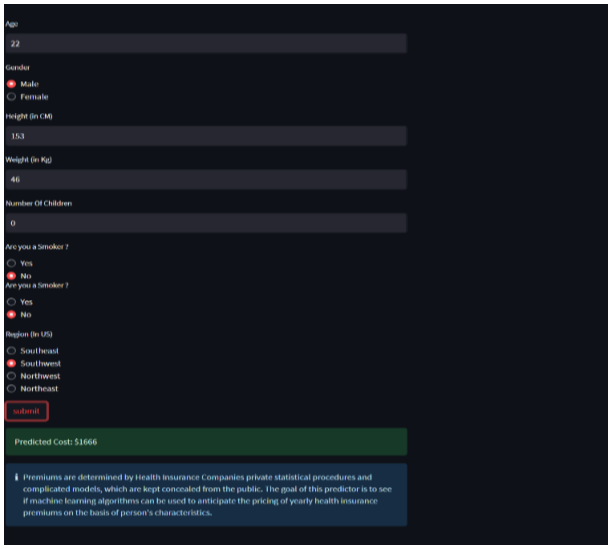

Fig.6. Website Screenshot

V. Conclusion

The Research focuses on machine learning algorithms that can be used for effective price prediction along with XGBRFRegressor which helps to prioritize the parameters used for prediction. The findings that are listed in the table II shows that XGBRFRegressor provides better accuracy than other machine learning models with a R2_score= 87.34%. XGBRFRegressor can therefore be used to predict and estimate the insurance cost when compared to other machine learning models. ML reduces the amount of effort incorporated in policy making it helps to estimate the premium on an individual in a sort span of time. It helps to work with enormous amount of data and also assist the users efficiently.

TABLE 2 VALUES OBTAINED FROM VARIOUS ML MODELS.

Algorithm	R2_Score	MAE	MSE
Random Forest Regressor	86	2567.82	22127330.63
DecisionTree Regressor	69	3202.65	47805362.12
Linear Regression	78	4213.48	33979257.05
XGB Regressor	84	2687.92	24232872.81
XGBRF Regressor	87	2511.03	19429024.51

References

[1] Ahmed I. Taloba , Rasha M. Abd El-Aziz , Huda M. Alshanbari , and Abdal-Aziz H. El-Bagoury "Estimation and Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning" Hindawi , Journal of Healthcare Engineering Volume 2022, Article ID 7969220, 10 pages,2022.

[2] Keshav Kaushik, Akashdeep Bhardwaj , Ashutosh Dhar Dwivedi and Rajani Singh "Machine Learning-Based Regression Framework to Predict Health Insurance Premiums" international Journal of Environmental Research and Public

Health ,15 Pages ,2022.
[3] Ch. Anwar ul Hassan, Jawaid Iqbal, Saddam Hussain , Hussain AlSalman , "A Computational Intelligence Approach for Predicting Medical Insurance Cost" Hindawi, Mathematical Problems in Engineering, Volume 2021, Article ID 1162553, 13 pages
[4] M. A. Aefa, M. Mahmoud, and M. M. Nassar, "Parameter estimation for a mixture of inverse chen and inverse compound Rayleigh distribution based on type-I hybrid censoring scheme," Journal of Statistics Applications & Probability, vol. 10, no. 3, pp. 647–663, 2021.
[5] W. A. Afifi and A. H. El-Bagoury, "Optimal multiplicative generalized linear search plan for a discrete randomly located target," Information Sciences Letters, vol. 10, no. 1, pp. 153–158, 2021.
[6] R. A. Ganaie, V. Rajagopalan, and S. Aldulaimi, "The weighted power shanker distribution with characterizations and applications of real life time data," Journal of Statistics Applications & Probability, vol. 10, no. 1, pp. 245–265, 2021.
[7] J. Liu, D. Capurro, A. Nguyen, and K. Verspoor, "Early prediction of diagnostic-related groups and estimation of hospital cost by processing clinical notes," NPJ Digital Medicine, vol. 4, no. 1, pp. 1–8, 2021.
[8] H. N. Alhazmi, A. Alghamdi, F. Alajlani, S. Abuayied, and F. M. Aldosari, "Care cost prediction model for orphanage organizations in Saudi Arabia," IJCSNS, vol. 21, no. 4, p. 84, 2021.
[9] M. H. Abu-Moussa, A. M. Abd-Elfattah, and E. H. Hafez, "Estimation of stress-strength parameter for Rayleigh distribution based on progressive type-II censoring," Information Sciences Letters, vol. 10, no. 1, pp. 101–110, 2021.
[10] S. Sana and M. Faizan, "Bayesian estimation using lindley's approximation and prediction of generalized exponential distribution based on lower record values," Journal of Statistics Applications & Probability, vol. 10, no. 1, pp. 61–75, 2021.
[11] M. Ravaut, H. Sadeghi, K. K. Leung et al., "Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data," NPJ digital medicine, vol. 4, no. 1, pp. 1–12, 2021.
[12] J. A. Irvin, A. A Kondrich, M Ko et al., "Incorporating machine learning and social determinants of health indicators into prospective risk adjustment for health plan payments," BMC Public Health, vol. 20, no. 1, pp. 608–610, 2020
[13] Akashdeep Bhardwaj, "Health Insurance Claim Prediction Using Artificial Neural Networks" International Journal of System Dynamics Applications Volume 9 • Issue 3 • 19 pages, 2020.
[14] L. Hu, L. Li, J. Ji, and M. Sanderson, "Identifying and understanding determinants of high healthcare costs for breast cancer: a quantile regression machine learning approach," BMC Health Services Research, vol. 20, no. 1, pp. 1066–1110, 2020.
[15] N. I. Jha, I. Ghergulescu, and A.-N. Moldovan, "OULAD MOOC dropout and result prediction using ensemble, deep learning and regression techniques," in Proceedings of the 11th International Conference on Computer Supported Education CSEDU, no. 2, pp. 154–164, Heraklion, Crete, Greece, MAY 2019.
[16] M. P. Shyamala Devi, M. Swathi, V. Purushotham Reddy et al., "Linear and ensembling regression based health cost insurance prediction using machine learning," in In: Smart Computing Techniques and Applications. Smart Innovation, Systems and Technologies, S. C. Satapathy, V. Bhateja, M. N. Favorskaya, and T. Adilakshmi, Eds., vol. 224, Singapore, Springer, 2019.