How to Cite:
Prakash, V. S., Bushra, S. N., Subramanian, N., Indumathy, D., Mary, S. A. L., & Thiagarajan, R. (2022). Random forest regression with hyper parameter tuning for medical insurance premium prediction. *International Journal of Health Sciences*, *6*(S6), 7093–7101. https://doi.org/10.53730/ijhs.v6nS6.11762

Random forest regression with hyper parameter tuning for medical insurance premium prediction

Dr. V. S. Prakash
Assistant Professor, Department of Computer Science, Kristujayanti College, Bengaluru
Email: vsprakash@kristujayanti.com

Dr. S. Nikkath Bushra
Associate Professor St. Joseph's Institute of Technology, Chennai
Email: ferozbushra@gmail.com

Nalini Subramanian
Associate Professor/IT, Rajalakshmi Engineering College, Thandalam
Email: mrgn.nalini@gmail.com

Dr. D. Indumathy
Associate Professor, Department of ECE/Rajalakshmi Engineering College
Email: indumathy.d@rajalakshmi.edu.in

S. Angel Latha Mary
Professor and Head Department of Computer Science and Business Systems, Sri Eshwar College of Engineering, Coimbatore
Email: xavierangellatha@gmail.com

Dr. R. Thiagarajan
Associate Professor/IT,Prathyusha Engineering College,Chennai
Email: rthiyagarajantpt@gmail.com

***Abstract*---**The proposed effort has the purpose of predicting an individual's insurance expenses also identifying people having medical insurance plans and clinical data, irrespective of their health concerns. A patient will require many types of health insurance. Regardless of the type of insurance coverage a person has, it is feasible to estimate their health insurance expenditures depends on the degree of critical care they get. The random forest Regression is one of the regressors used in this investigation. When the accuracies were compared, hyper parameter tuning was the most effective of all the approaches, with a 98 percent accuracy. Finally, the prediction fit

Manuscript submitted: 9 April 2022, Manuscript revised: 18 June 2022, Accepted for publication: 27 July 2022

will calculate the insurance expense of the user and calculate the insurance costs.

***Keywords*---**random forest regression, hyper parameter, machine learning, prediction.

Introduction

Humans deal with a lot with hazards and uncertainty. People, houses, companies, institutions, and assets are all subject to numerous sorts of danger, which might vary. Death, disease, and the property loss or goods are among the risks. People's lives are centred on their happiness and health. Yet, since risks can be prevented, the financial industry has developed a range of products to shield people and institutions from them by compensating them with financial resources. As a consequence, insurance is a form that lowers the costs associated with specific risks. Health insurance is a coverage that protects against medical expenditures. After paying a premium in health insurance, a person acquired a medical insurance plan obtains coverage. A number of factors influence the insurance cost. Because many factors determine the cost of a healthcare plan, the insurance premium value differs from person to person. Consider age: a youth is significantly less likely to suffer from major health problems than an older one. As a reason, treating the old is more costlier than nursing the young. As a result, an elderly adult is required to pay a larger charge than a younger man. The amount due differs from individual because [1] several parameters impact the insurance costs of a medical insurance plan.

In health coverage, machine learning is competent of executing numerous tasks at a significantly faster pace in order to properly anticipate identify the illness and offer the best drug treatments to the patient. AI may collect data, analyse it, and then present the right output to the user. This shortens the diagnosis-treatment-recovery cycle by reducing the time needed to discover illnesses and errors. Chatbots, for example, are used by health care providers or organisations to acquire basic information ahead to a consultation with a physician if you want to choose an online appointment. This helps the doctor understand the problem before starting the consultation process. As a consequence, both the physician as well as the patient benefit from time savings. Even though the healthcare industry rapidly digitises, massive volumes of data will eventually be generated and collected. Because more raw data equals more effort, this will merely increase the burden for healthcare practitioners. Machine learning can evaluate the data and provide insights to patients and healthcare professionals. It is a more efficient method of diagnosing illnesses. On data evaluation, medical care in the country is extremely intricate and tough to comprehend, and people frequently pay the costs. Machine learning in medical care has the potential to improve the treatment efficiency. It can also help healthcare professionals spend more time providing proper therapy, reducing burnout among medical specialists. Here are some instances of how artificial intelligence affects healthcare:

- Healthcare scope is restricted in developing or underdeveloped countries.
- Health records were about time consuming.

- The threat of resistance to antibiotics is being decreased.
- Insurances are processed more quickly.
- Affordable health insurance plans

The following are the aspects of this work: The subject of predicting insurance has not been thoroughly found out after much investigation. Patients, clinics, doctors, and insurers may profit from the suggested artificial intelligence model and complete their jobs more quickly and efficiently.

- To estimate health insurance rates, the authors developed a Radom forest-based regression model.
- The model was assessed using key performance indicators.
- The proposed model's total accuracy was 98 percent.
- To increase accuracy, hyperparameter adjustment was used.

The following is how this paper is organised: Chapter 1 started with an overview to the topic and idea, followed by a discussion of the most recent related research in this domain. Chapter 3 covers the implementation's working technique. Finally, Chapter 4 offers the paper's conclusion and displays the findings and comments.

Literature Review

The study presented by van den Broek-Altenburg [2] sought to discover the customers thought on health care insurance by examining their ideas said on Twitter. The idea was to use sentiment analysis to learn how individuals feel about medical insurance and doctors. The authors used a Software Development kit to gather tweets on Twitter with the terms "healthcare insurance" or "health plan" throughout the 2016-2017 registration period in the United States. Insurance is a policy that reduces or eliminates the expenses of loses caused by various perils. Several [3] factors influence the insurance costs. These factors influence the creation of insurance programmes. These factors influence the creation of medical insurance. Artificial intelligence can assist the insurance sector in improving policy wording effectiveness.

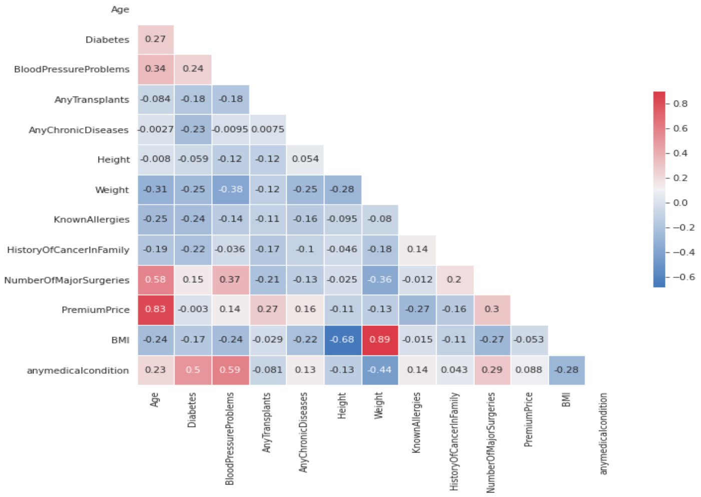
Risk assessment is crucial in the life insurance sector for categorising applicants. Companies use screening process to make application judgments and set insurance product price. Because of the proliferation of data and developments in business intelligence, the review process may be automated to speed up submissions or programmes. The research in [4] sought to identify methods to apply prediction insights to assess risk analysis for health insurance businesses. The study made use of a data set including over a hundred anonymized features. Dimensionality reduction was used to choose salient characteristics that might improve the predictive power of the models.

Fauzan and Murfi [5] employed Gradient boosting to solve the insurance prediction problem and assess its accuracy. They have compared XGBoost's effectiveness to that of Stochastic gradient descent, Random Forest, Random GB, and Neural Network ensemble learning methods, as well as online learning methods. The simulation results show that XGBoost beats other approaches in terms of standardised Gini. Individuals are slowly purchasing such insurance, allowing scam artists to take advantage of them. Customer insurance fraud includes, among other things, exaggerated demands and post-dated policies. Insurance fraud happens on the vendor side when non-existent businesses implement requirements and neglect to pay premiums, among many other things. In the work [6] they have compared various types of categorisation methods.

In the work, the K-means technique [7] and the Elbow approach were employed to correctly organise users into an adequate group based on similarities. Using a provided criteria, the medical insurance premium quote was anticipated for each group of individuals based on this study. Estimating the cost of people's health insurance is an important step toward improving clinical accountability. Sailaja et al. used multiple regression models to analyse personal health information in order to anticipate insurance rates for persons in this study [8]. A variety of factors influence insurance premiums. The adoption of a Stacked Regression model to forecast insurance premiums for individuals may benefit users. Dutta et al. [9] calculated the healthcare costs using prediction analysis. Actuaries in the insurance industry use a number of numerical approaches to anticipate yearly medical claims expense. This amount must be included in the financial statement budgets. In most cases, improper estimation has a negative influence on a success. Goundar et al. [10] shown how to construct a convolutional neural networks (ANN) capable of predicting the medical claims. The goal was to reduce the average absolute error by adjusting configuration parameters such as epoch, transfer function, and neurons in different layers once the machine learning algorithms were built. To anticipate the yearly claim amounts, feeding forward as well as RNN were used.

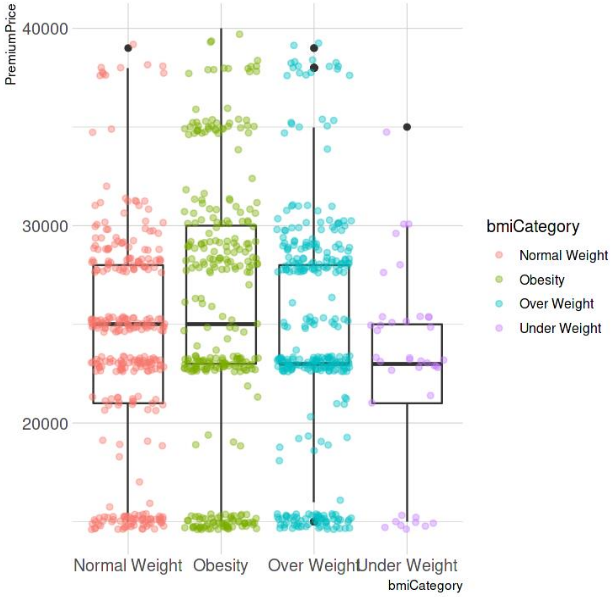
Proposed Methodology

The authors of this study employed the Random forest algorithm for implementation and built a machine learning model to identify medical insurance costs. Initially, the dataset as well as the required libraries and modules were downloaded. The dataset had over 1000 columns of data containing information such as age, diabetes, blood pressure, transplants, chronic illnesses, height, weight, allergies, disease history, and significant operations. This information was used to forecast the cost of medical insurance. Following that, an exploratory statistical analysis was carried out. The dataset was tested for null values in this stage. The statistical summary of the dataset was examined for the missing values in the data set. The dataset was examined in this stage to determine the relationships among the various columns. The dataset was divided into age groups, and the association among various parameters was examined.



Matrix showing relevance of the data with every parameter

We trained the prediction model in this stage, but first they cleaned the dataset. The data was scaled and just the quantitative values were obtained. The data was scaled using a conventional scaler. Before providing the data to the model, it is necessary to scale it. The logit model was trained when the dataset had been properly tried to scale.



BMI Category and Premium Price

A Random Forest is an ensemble approach that can do both classification and regression pitfalls by aggregating the methods as Bootstrap Aggregation, or bagging. Bagging in the Random Forest approach entails learning each prediction model on a distinct data sample, using replacement sampling. The basic knowledge is to use numerous decision trees to determine the exact output rather than depending on independent decision trees. Collection of decision trees are called as random forest. This is to state that a Random Forest is made up of several trees that are built in a "random" manner.

- Each tree is built from a separate set of rows, and a new set of characteristics is chosen for splitting at each node.
- Each of the trees provides a unique forecast.
- These forecasts are then combined to get a single outcome. Averaging enhances the accuracy of a Random Forest over a single Decision Tree and decreases overfitting.
- A Random Forest Regressor prediction is an average of the forecasts made by the trees in the forest.

Hyperparameter tuning

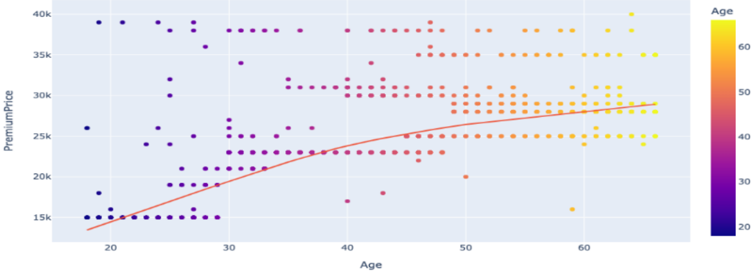
Because tuning is based on results and result based findings rather than theory, the easiest way to identify the ideal settings is to test several possible combinations and assess the performance of each model. However, assessing each classifier solely on the training data set can lead to overfitting, among the most fundamental difficulties in machine learning. If we optimise the model for training examples, our model will perform admirably on the training set but will be unable to generalise to updated data, such as in a test set. Overfitting occurs when a model performs well on the training phase but badly on the test set. This results in a model that understands the training dataset quite well but cannot be transferred to the test set. To tune the hyperparameters, we run multiple iterations of the full K-Fold CV technique, each time with a different model setup. Then we analyze all of the models, choose the top one, training it on the entire training set, and assess it on the testing set. The following hyper parameters were tuned and the performance of the regression model was improved to achieve the high performance.

Estimator - maximum number of nodes in the foreset
Maximum Features – Max feature on node division
Maximum depth - the maximum number of tiers in each tree
Sample split = the minimum number of data before split
_depth = maximum number of tiers in each tree
bootstrap = sampling procedure for data points

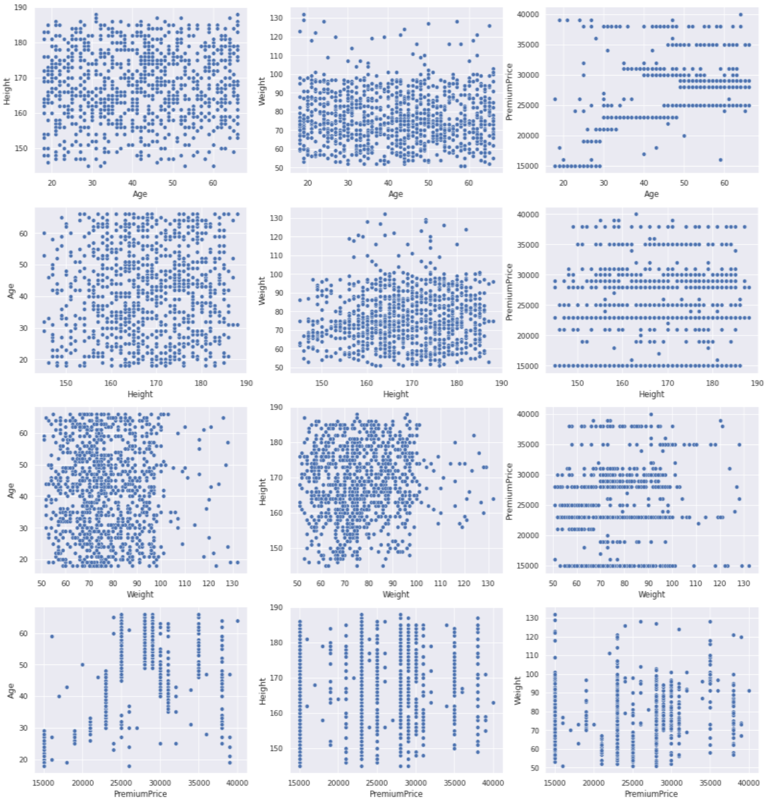
OOB estimate of error rate: 10.5%

Experimental Results

The pairplot diagram depicted the correlation between the dataset's distinct columns. A pairplot is a matrix that displays all of the scatter plots in all of the various configurations in our data. The plot was plotted after the pairplot diagram, as illustrated in Figure. We can can see it as age climbed, so did the cost of living. As a result, the costs and age have a linear connection.

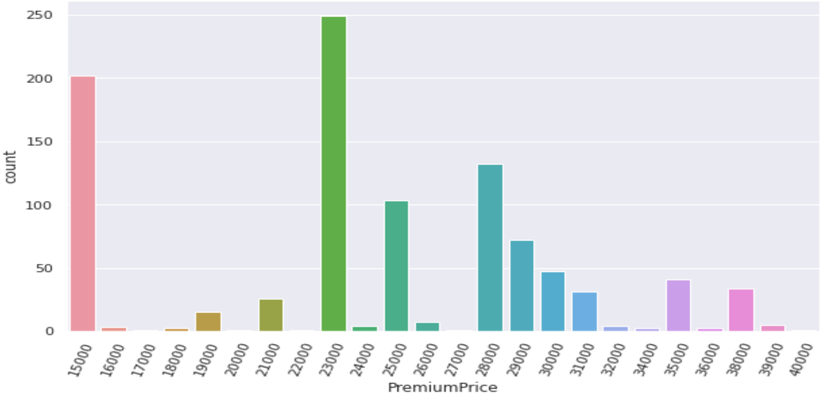


Age and Premium Relevance



Premium Price with various factors

The pairplot was then plotted, as shown in Figure. They are used to determine which characteristics explain the relationship between two variables or create the most distinct groupings. In our dataset, creating basic lines or generating a linear distinction contributed in the creation of certain rudimentary classification models.



Premium Price

The random forest classifier along with the hyper parameters were tuned and the results were displayed below with the efficiency.

```
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.
[Parallel(n_jobs=-1)]: Done  18 out of  18 | elapsed:    4.8s finished
[Parallel(n_jobs=-1)]: Using backend ThreadingBackend with 4 concurrent workers.
[Parallel(n_jobs=-1)]: Done  33 tasks      | elapsed:    0.1s
[Parallel(n_jobs=-1)]: Done  60 out of  60 | elapsed:    0.2s finished
```

ACCURACY : 0.984944

Conclusion

Machine learning is so well to tasks which are frequently done by individuals at a slower pace in the area of medical insurance. AI and machine learning can analyse and evaluate large amounts of data to expedite and improve health insurance procedures. The influence of computer vision on medical insurance will save both patients and insurers time and money. AI will undertake routine tasks, freeing up insurance professionals to focus on procedures that will enhance the users experience. People, institutions, clinicians, and health insurers will profit from ML's capacity to execute tasks that are now done by humans but are considerably quicker and less costly when done by machine learning. Machine learning is one aspect of historical data exploitation. The proposed method employs the Random forest classifier with hyperparameter tuning and accuracy of health care premium prediction. The model achieves the accuracy of nearly 98.4 percentage using the regression model. This area of insurance forecasting has not been well investigated and requires further investigation.

References

1. Boodhun, N.; Jayabalan, M. Risk Prediction in Life Insurance Industry Using Supervised Learning Algorithms. *Complex Intell. Syst.* 2018, *4*, 145–154.
2. Dutta, K.; Chandra, S.; Gourisaria, M.K.; GM, H. A Data Mining Based Target Regression-Oriented Approach to Modelling of Health Insurance Claims. In

Proceedings of the 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 8–10 April 2021; pp. 1168–1175.

3. Fauzan, M.A.; Murfi, H. The Accuracy of XGBoost for Insurance Claim Prediction. *Int. J. Adv. Soft Comput. Appl.* 2018, *10*, 159–171. Available online: https://www.claimsjournal.com/news/national/2013/11/21/240353.htm (accessed on 9 May 2022).
4. Goundar, S.; Prakash, S.; Sadal, P.; Bhardwaj, A. Health Insurance Claim Prediction Using Artificial Neural Networks. *Int. J. Syst. Dyn. Appl.* 2020, *9*, 40–57.
5. 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
6. Health Insurance Premium Prediction with Machine Learning. Available online: https://thecleverprogrammer.com/2021/10/26/health-insurance-premium-prediction-with-machine-learning/ (accessed on 9 May 2022).
7. Mukhtar, A. U. S., Budu, B., Sanusi B, Y., Mappawere, N. A., & Azniah, A. (2022). The effect of reproductive health education with multimedia video learning on the improvement of fluor albus prevention behavior young woman pathologist. International Journal of Health & Medical Sciences, 5(1), 75-79. https://doi.org/10.21744/ijhms.v5n1.1841
8. Preethi, P., & Asokan, R. (2021). Modelling LSUTE: PKE Schemes for Safeguarding Electronic Healthcare Records Over Cloud Communication Environment. Wireless Personal Communications, 117(4), 2695-2711.
9. Preethi, P., Asokan, R., Thillaiarasu, N., & Saravanan, T. (2021). An effective digit recognition model using enhanced convolutional neural network based chaotic grey wolf optimization. Journal of Intelligent & Fuzzy Systems, (Preprint), 1-11.
10. R.Thiagarajan ,R.Jothikumar,T.Rubeshkumar,P.Jayalakshmi,M.Baskar“Enhanced Resemblance Measures forIntegration in Image-Rich Information Networks”, Journal of Critical Reviews,ISSN- 2394-5125 Vol 7, Issue 16, July 2020.
11. R.Thiagarajan,N.R .Rajalakshmi , M. Baskar ,P. Jayalakshmi “A Novel Solution for Economizing Water by a Mix of Technologies with a Low Cost Approach”,International Journal of Advanced Science and Technology Vol. 29, No. 7, April 2020,
12. Sailaja, N.V.; Karakavalasa, M.; Katkam, M.; Devipriya, M.; Sreeja, M.; Vasundhara, D.N. Hybrid Regression Model for Medical Insurance Cost Prediction and Recommendation. In Proceedings of the 2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT), Visakhapatnam, India, 13–14 November 2021; pp. 93–98.
13. Suryasa, I. W., Rodríguez-Gámez, M., & Koldoris, T. (2021). The COVID-19 pandemic. *International Journal of Health Sciences*, *5*(2), vi-ix. https://doi.org/10.53730/ijhs.v5n2.2937
14. 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.