Asia-Pacific Journal of Science and Technology: Volume: 28. Issue: 06. Article ID.: APST-28-06-10. Research Article

APST

Asia-Pacific Journal of Science and Technology

https://www.tci-thaijo.org/index.php/APST/index

Published by the Research and Graduate Studies Division,
Khon Kaen University, Thailand

Forecasting health insurance premium using machine learning approaches

Shawni Dutta[1], Payal Bose[1] and Samir K. Bandyopadhyay[1,*]

[1]Department of Computer Science and Multimedia, Lincoln University College, Selangor, Malaysia
[*]Corresponding author: drsamir@lincoln.edu.my

Received 7 January 2022
Revised 23 June 2022
Accepted 16 November 2022

Abstract

A medical emergency can impact anybody at any moment and have a significant psychological and economic consequence. Health insurance encompasses different costs including hospital charges, medicine costs, physician consultation fees, etc. The importance of health insurance cannot be underestimated, given the exponential rise in healthcare expenses. This research has attempted to forecast the cost of health insurance premiums. To address this problem, an automated system can be built that analyzes an individual's health complications and forecasts associated costs. Machine learning-based techniques can be employed to design the automated system. This research work will use ensemble-based machine learning approaches to evaluate an individual's risk and anticipate premium costs. This will allow the insurance firms to set a minimal fee with a higher profit in order to attract more policyholders. Random Forest, CatBoost, Extra Trees, AdaBoost, Extreme Gradient Boost, and Gradient Boost models are popular ensemble-based algorithms that are applied to an individual's health information and used to estimate premium price. The automated model can be developed by applying the Random Forest model with a Mean Square Error (MSE) of 0.01 and Mean Absolute Error (MAE) of 0.0394, according to the comparative study of these employed models. The graphical representation of the comparative analysis is depicted in Figure 1A and 1B based on MAE and MSE respectively. The research finds relative ranking among the interfering factors for determining medical cost for an individual. According to the Random Forest model's findings, a person's age has the greatest impact on determining the related premium.

Keywords: Ensemble approach, Feature importance, Health insurance price, Machine learning, Random forest, Regression analysis

1. Introduction

Uncertain illness or sickness cannot be predicted beforehand, but people can prepare themselves from a monetary perspective to get rid of unwanted situations. This economic imbalance condition can be overcome by subscribing to health insurance. Health insurance facilitates people to take much-needed financial assistance when a medical emergency arises. In any insurance business, it is imperative to adjust the premium amount for the policyholders considering the fact that there will be loss of premium amount due to customers opting out of the policy. Hence, the premium amount is based on this anticipated loss of revenue. The insurer should charge a reasonable premium amount for their clients with a sensible profitability. Scientific and technical advancements are allowing health insurance coverage to be optimized to individual health risk profiles. Exponential expenses in medical costs can be reduced by lowering the out-of-pocket expenditures. These expenditures should be well-planned and saved using a medical policy to conceal the future costs. The influence of health insurance on individual medical expenses is of interest to policymakers.

There are certain factors which can lead to insurance premium fluctuation. A study [1] has revealed that the people of the age group 21-30 years are quite pleased with the medical insurance schemes including the costs. However, people above 60 years are less satisfied with the premium cost incurred. Again, satisfaction in customer support schemes offered by health insurance companies for the people in the 31-40 years age bracket is relatively higher than the people above 60 years. The relationship among diabetic patients and premium price rate explained by [2] the authors. Another study [3] elaborated the different chronic diseases that may have an impact on the

insurance price level. This research work contributes towards forecasting/projection of health insurance cost by examining several health-related aspects that were put forward by other research works discussed in this article. In fact, the entire world is facing coronavirus disease starting in 2019 (COVID-19) pandemic situation which is addressed by certain research articles [4,5]. Prior estimation of medical cost for an individual at an advanced time will assist in combating pandemic situations.

Rapid improvements in machine learning and artificial intelligence, along with easily available processing capacity to run these algorithms on even the tiniest of devices, have resulted in significant breakthroughs in scientific discovery and commercial applications across a wide range of fields and industries. As a result, modern enterprises and technical advancements must be accompanied with suitable legislation that will regulate the related risks and allow the industry to thrive [6]. Any industry recognizes the importance of building a smart automated system for making informed decisions. This requirement prompts the development of smart tools using Machine Learning-based methodologies. The study encourages the development of a smart healthcare system that uses Machine Learning-based technologies to provide insight into health-related expenditure estimation. A step-by-step procedure to construct such automated decision-making model is outlined as follows. This study has meant to forecast the medical insurance premium amount depending on some interfering factors of the clients such as age, height, weight, diabetes tendency, problem of blood pressure, whether the patient has undergone through any transplants or not, if the patient has any chronic disease or not, cancer history in the family, number of surgeries faced by the patient, and allergies. In this study, an automated technique has been proposed for discovering the hidden relationship between these considered factors and recommending the premium amount for that patient. Machine learning (ML) based methods can be applied to develop this automation. Machine learning techniques are advantageous in this situation because they allow for the discovery of spontaneous groupings among the patients' background details. The investigation of this patient information can lead to the forecast of health insurance premiums. Ensemble learning is a dominant machine learning technique which can increase the predictive model's efficiency by assembling multiple learners. Ensemble methods have been used in this research to predict premium amount for an individual policy holder. Different ensemble models namely, Extra Trees, Random Forest, CatBoost, Adaboost, Gradient Boosting, Extreme Gradient Boosting methods are applied to an individual's health information for obtaining the premium rate at an early stage. The presented work ensures the performance of these models will be optimized by fine-tuning their related parameters. An exhaustive comparison among the models is carried out so that the most promising performing models can be identified and can be utilized to implement the automatic premium rate recommender tool. Thegraphical representation of the comparative analysis is depicted in Figure 1A and 1B based on Mean Absolute Error (MAE) and Mean square Error (MSE) respectively.

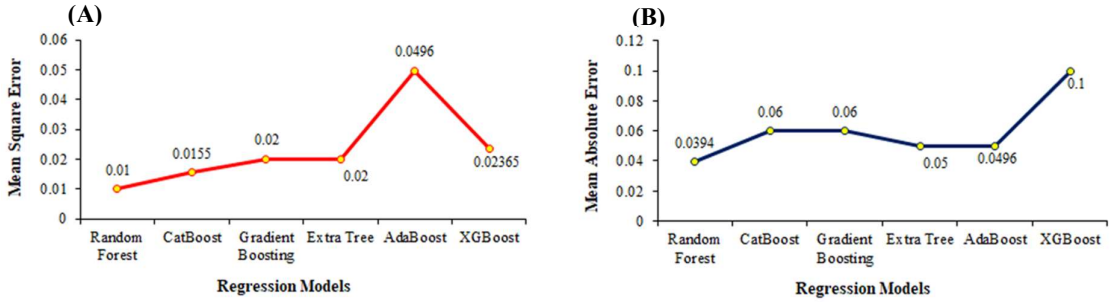

Figure 1 Comparative Analysis Among the Ensemble based Regression Models based on (A) MAE, (B) MSE.

2. Materials and methods

2.1 Health insurance cost prediction

This section discusses research efforts related to price prediction involving machine learning techniques. The first sub-section of this article discusses published research works in the health insurance price prediction domain. The following subsection summarizes other price prediction algorithms that have been successfully applied to other research domain like time series analysis problem in Stock market data analysis. Several papers have been written on the topic of claim prediction on premium price of health insurance. A detailed summary is described in Table 1.

Table 1 A Brief summary.

Problem Statement	Method Used	Result
Predict Health Insurance Cost [7]	Machine Learning and DNN Regression Models	Mean Absolute Error (MAE) = 0.17448, Root Mean Square Error (RMSE) = 0.38018, R-squared = 85.8295
The Prediction Model of Medical Expenditure [8]	Best models for prediction SVR methods with XGBoost	MAE = 0.1302, Mean Absolute Square Error (MASE) = 0.2067, and Mean Absolute Percentage Error (MAPE) = 0.0094
Prediction of Health Insurance Cost [9]	Multiple Linear Regression Technique	Mean Absolute Percentage Error (MAPE) = 3% (around), R2=0.7615896
Health Insurance Amount Prediction [10]	Multiple Linear Regression	Accuracy of gradient boosting decision tree regression = 99.5%
Predicting Health Care Costs [11]	Evidence Regression	R-Square = 0.44
Forecasting high-cost for the high-need patient to spend in health care [12]	Machine learning approaches	Overall R-squared > 0.7

2.2 Similar price prediction system

Several other studies on price prediction systems are being conducted by various researchers. Yan and Yang [13] proposed an encoder-decoder model of attention mechanism for predicting the stock market price based on two features and time. They used Long Short-Term Memory (LSTM) neural network encoding and decoding. The experimental results show that by introducing the attention strategies, forecast errors can be minimized than the other methods. Ye et al. [14] conducted China's empirical research for medical insurance companies' equity performance in the epidemic period. The authors developed a predictive model using the integrated autoregressive model and two neural network models, back propagation, and long short-term memory. Finally, the LSTM model outperforms the other two models based on MAE and MSE. Another research. [15] discussed various prediction techniques of stock market price prediction. They examined the most modern stock market prediction algorithms in this article. They used Artificial Neural Networks, Neuro-Fuzzy Systems, Time Series Linear Models, and Recurrent Neural Network models, as well as their advantages and disadvantages. The authors [16] proposed a method to forecast the stock market trend. They created two methods, the first model for daily prediction and another model for monthly prediction. The models are built using supervised machine learning algorithms. They observed on a daily prediction model, supervised machine learning algorithms achieved up to 70% accuracy and the monthly prediction model correlated with the next month's prediction. A supervised machine learning technique has been proposed by Shastri and Rengarajan [17] for forecasting car prices in India. To conduct the predictive analytics, the data has been retrieved from the Quikr website using some existing models, namely, Multiple linear regression analysis, Random Forest regressor using Randomized search Cross validation technique. Next, the forecasts are evaluated and contrasted to check the ones that produce the enhanced predictive results. Asghar, and Mehmood et al. [18], proposed a method that assists both the buyer and seller in purchasing and selling their old used vehicle. The best choice for their vehicle has been predicted by this study to support an informed decision-making process for both personal and business scenarios.

Due to the ongoing pandemic situation, the entire world is facing health as well as financial challenges as discussed by [19,20]. Accurate medical cost prediction will aid a person in managing their financial goals in order to overcome the financial load imposed by the pandemic condition. Hence, the presented study will be appropriate to fight against such a catastrophic crisis.

2.3 Methods

Regression Analysis is a statistical procedure. A data instance is mapped to a real-valued prediction variable via regression. This explains how the predictors respond to changes in criteria values. Regression presumes that the target data fits into a recognized sort of function which represents the dataset [21]. The regression problem is modeled by using machine learning based methods in this present study. Ensemble learning-based regression is a specialized domain of machine learning which can integrate the predictions of more than one method into a single prediction. The application of ensemble-based models is quite evident and observed in numerous studies [22,23]. These models are more robust than a single model and can provide higher forecasting abilities. As a result, the ensemble machine learning based approaches are preferred in this study. There are three types of ensemble

methods, namely, Bagging, Boosting and Stacking ensemble. However, the authors have used Bagging and Boosting ensemble approaches in their research work.

2.3.1 Bagging techniques

The bagging ensemble method, popularly regarded as Bootstrap Aggregation, follows a parallel ensemble approach. This method can provide a low-variance forecast by collecting extra data throughout the learning stage and replacing the original dataset with a random selection. The training procedure is conducted multiple times to teach and construct several models. As a result, it creates an assemblage of distinct models. The assemblage regression outcome is constructed using the average of each obtained prediction from distinct models. The robustness of single model alone is lower than these ensemble models.

2.3.1.1 Random forest algorithm

This ML method is utilized for regression as well as classification problems. This algorithm is founded on the concept of ensemble technique. This technique assembles several decision trees with the Boosting and Aggregation or Bagging technique [24]. The total number of trees and the model's accuracy are directly related. Random Forest, rather than depending on an individual decision tree, gathers the predictive result from each tree and forecasts the final outcome depending on the highest vote count of predictions. Random forest is quite efficient in avoiding the over-fitting problem by generating random selections of features and using those subsets to create smaller trees.

2.3.1.2 Extra trees algorithm

Extra Trees is another ensemble-based bagging technique. It has promising properties such as scalability and high efficiency. This method works by combining the results of many de-correlated decision trees, each of which has been provided with training on a set of data. A "forest" is formed by the gathering of base learners. In the case of a regression problem, the predictions of these learners are derived by computing the average prediction value. [25]. The Extra Trees algorithm slightly differs from the bagging and random forest method. Each decision tree in this approach is applied to the whole training dataset, while Random Forest makes decisions using a bootstrap sample of a training dataset. Random Forest and Extra Trees methods sample feature instances in a random pattern at every decision tree split point.

2.3.2 Boosting technique

The boosting method follows a sequential ensemble approach which incorporates many prediction forms. It creates a powerful prediction model by lowering the bias error. This method, like the Bagging method, uses random sampling to generate several training datasets. This strategy takes numerous weak learners and turns them into strong learners by merging them. This approach iteratively modifies the weights to each generated model during the training session. When comparing learners with good prediction results to those with wrong predictions, larger weights were allocated to the former.

2.3.2.1 Gradient boosting

Gradient Boost (GBR) [26] is a well-known ensemble-based learning technique. This model is used to perform regression and classification tasks. There are three components to this boosting method: a hyper tuned loss function, weak learners for predictions, and an additive model for adding the weak models which helps in minimizing the loss function. This learning approach creates highly correlated new base learners with negative gradients of the loss function. The purpose of the loss function is to improve intuition; hence, the loss function should be selected carefully by the researcher. If the predicted error function is the usual squared-error loss, the learning strategy anticipates sequential error-fitting.

2.3.2.2 AdaBoost

Freund and Schapire [27] proposed the first boosting ensemble technique, Adaptive Boosting or AdaBoost, in 1996. A meta-estimator is another name for this technique. The AdaBoost approach is modeled by utilizing weak learners, such as decision trees, and adding weights to incorrect values. This method can address classification and regression problems, but it was primarily designed to solve classification problems with only two classes and enhance the productivity of decision-making trees. This approach is resistant to overfitting for less noisy datasets.

The self-averaging property is also present in this method. This method automatically adjusts the parameters on the data. When this Regressor is used, the generalization error is low [28].

2.3.2.3 Extreme gradient boosting

This approach implements the Gradient Boosting algorithm in a more efficient and effective manner. The base learner in extreme gradient boosting (XGBoost) [29] is always awful at the remainder, so when all the predictions are accumulated, the less-promising ones are wiped out. Next, the remaining promising predictions are assembled to infer the final forecast. The XGBoost method is robust enough to prevent overfitting as the number of trees grows. The capacity to predict new test data may be impaired by a high value for a specific learning rate. This causes the learning rate to be reduced and the number of trees is increased. This results in the calculation becomes more costly that requires a longer time to complete.

2.3.2.4 CatBoost

Yandex's CatBoost is an open-source machine learning technique. It is simple to integrate with deep learning frameworks such as TensorFlow and ML. It exhibits cutting-edge prediction without the considerable training as needed by other ML methods. This method exhibits excellent support for more descriptive data formats. CatBoost produces cutting-edge results and can compete with any leading machine learning algorithm in terms of performance. CatBoost reforms categorical instances to numeric values by employing certain statistics on categorical feature combinations and numerical feature combinations. It is simple to use and user-friendly [30].

Figure 2 explains the subdivisions of ensemble approaches along with the employed models by the study.

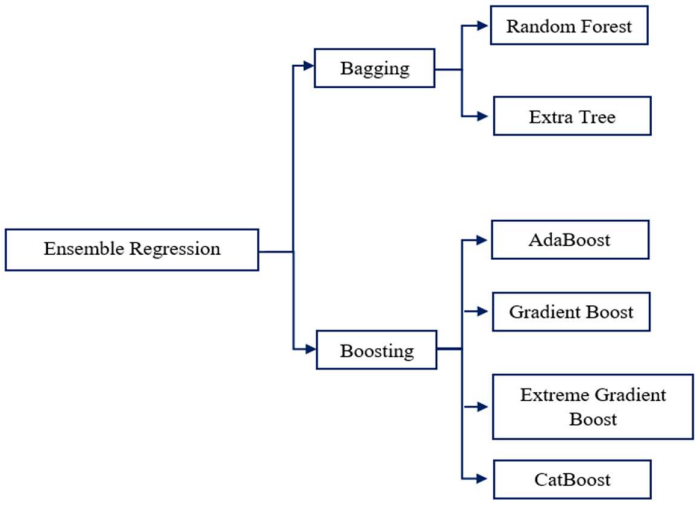

Figure 2 Types of Ensemble Regression.

2.3.2.5 Regression results evaluator

The efficiency of any regression method is justified against the deviation measurement between the forecast outcome and original data. The efficiency of prediction outcome is measured using two well-known metrics: MAE [31] and MSE [31]. The MSE, as given in the Equation (1) is defined as the difference in squared values between the actual and forecasted data (1). The model's efficiency increases as MSE decreases.

The MAE function is another loss function utilized for regression performance evaluation. The absolute difference present within actual and forecasted instances are obtained and summed up to constitute the MAE term. equation (2) has shown the MAE function. When the MAE value is 0, a perfect forecast is obtained.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_{\mathrm{Actual}} - Y_{\mathrm{Predicted}}) \quad (1)$$

Here n = No. of training set instances, $Y_{\mathrm{Actual}}Y_{\mathrm{Actual}}$= Actual Values, $Y_{\mathrm{Predicted}}Y_{\mathrm{Predicted}}$= Predicted Values

2.4 Materials

2.4.1 Dataset description

For pursuing the health insurance prediction, an existing dataset has been collected from Kaggle data repository [32]. The dataset has a total of 986 patient instances along with the necessary features (or attributes). The attributes present in the dataset include numerous features such as Age, Diabetes, Blood Pressure Problems, Any Transplants, Any Chronic Diseases, Height, Weight, Known Allergies, History of Cancer in Family. Number Of Major Surgeries, and Premium-Price. The detailed description of the dataset in terms of feature types and value range for each attribute is shown in Table 2. The variable Premium-Price is the target variable of the dataset which needs to be predicted. The target or dependent variable is continuous in nature hence the problem is implemented as a Regression problem. Figure 3 provides the insight of the histogram representation of the dataset. Another correlation matrix is plotted in Figure 4 to understand the relationship among the variables present in the dataset. The correlation matrix implies that the age and premium price attributes are highly correlated with each other.

Table 2 The detailed description of the dataset.

Attribute Name	Description of the attribute	Values present under the attribute
Age	Patient's Age	18-66
Diabetes	Whether the patient is diabetic or not.	0- non-diabetic 1-Diabetic
Blood Pressure Problems	Whether the patient is having any Blood Pressure related problem or not	0-No Blood Pressure related problem 1- Blood Pressure related problem exists
Any Transplants	Whether the patient has undergone or not	0-No transplant 1- Transplantation happened
Any Chronic Diseases	Whether the patient is having any chronic disease or not	0-No chronic disease 1- Chronic disease exist
Height (in centimeter)	Patient's height	145-188
Weight (in kilogram)	Patient's weight	51-132
Known Allergies	Whether person has any kind of allergy or not	0- No allergies 1-Allergy exists
History Of Cancer in Family	Whether the person has a cancer history in the family	0- No history 1-Cancer history
Number Of Major Surgeries	How many numbers of surgeries the patient has experienced	0-3
Premium Price	Premium price allotted to the individual.	15000-40000

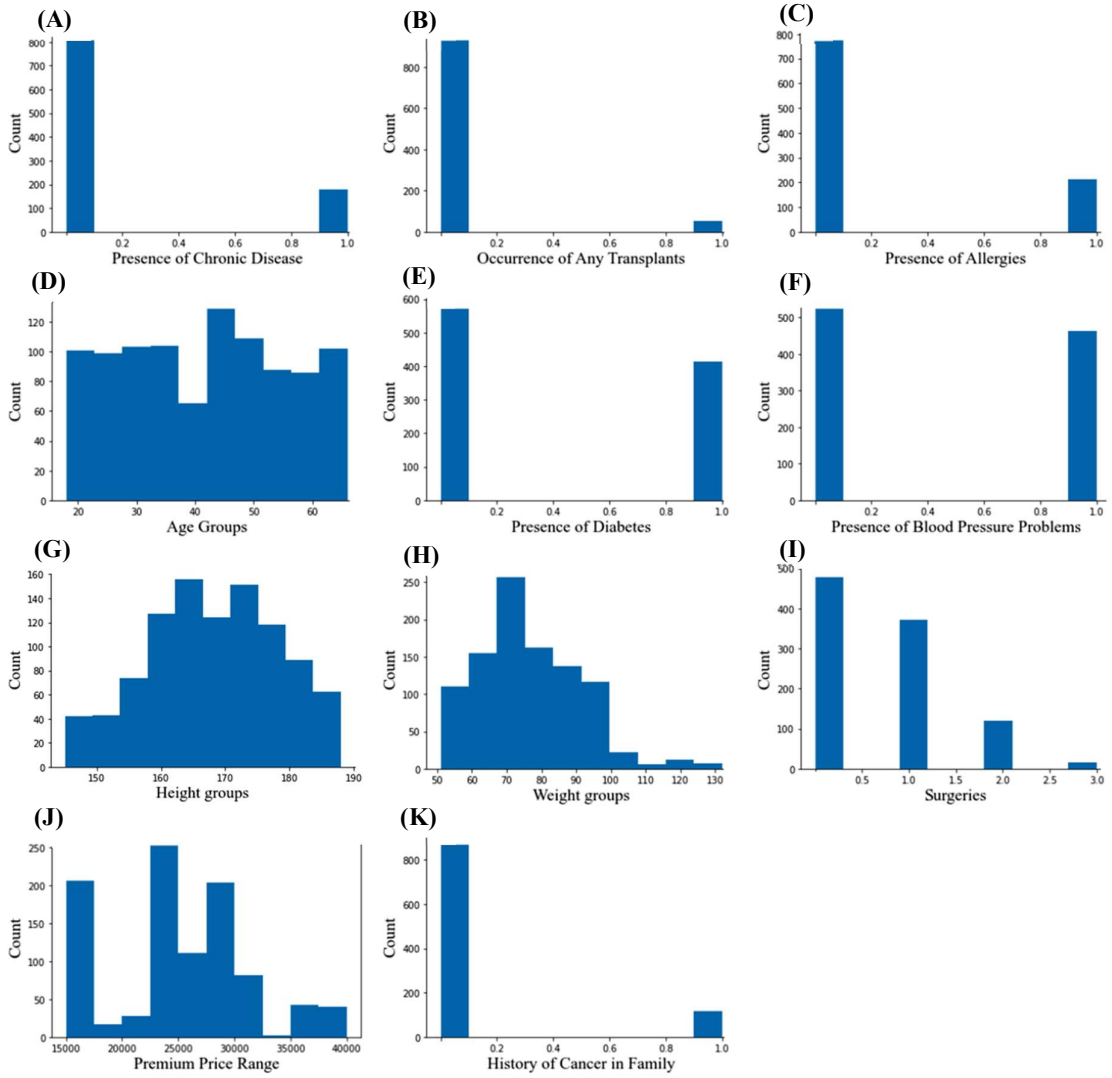

Figure 3 Histogram Interpretation of Dataset Attributes (A) Age (B) Any Chronic Diseases (C) Any Transplants (D) Blood Pressure Problems (E) Diabetes (F) Height (G) History of Cancer in Family (H) Known Allergies (I) Number of Major Surgeries (J) Premium Price (K) Wight.

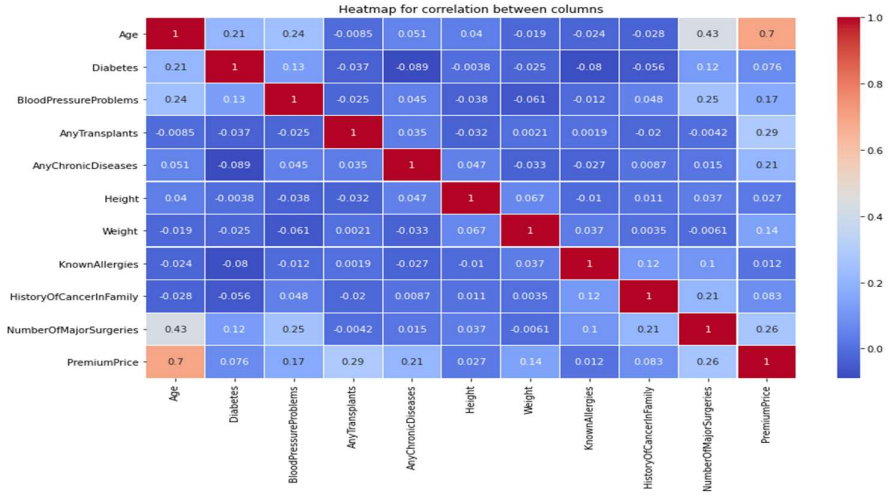

Figure 4 Correlation Matrix Plot Among the Attributes present in the Dataset.

2.4.2 Methodology

The entire medical insurance premium price will be identified as a regression problem in this investigation. This problem falls under the topic of multiple regression because the dataset contains numerous independent variables that are used to predict the desired outcome (insurance cost). Machine learning-based methods are used to analyze this regression problem. Machine learning models can be implemented for prediction. Since the data labels of this considered domain are essentially numeric in nature, the application of machine learning based regression techniques are most appropriate. The aforementioned task is accomplished by this study with the help of a multi-step procedure.

As the dataset collection process is over, the dataset requires to be preprocessed for obtaining a cleaned dataset. The first step to be applied on the dataset is feature scaling which is the process of normalizing the features of the dataset. The dataset has certain attributes such as age, height, weight, and premium price which have a larger range of values than the other attributes present in the data. The larger valued features may dominate the other small-valued features of the data. In the case of the feature variables that are highly unequal, the ML algorithms iterate inefficiently to obtain the optimum feature value. Hence, it is essential to apply feature scaling operations where the aforementioned attributes namely age, height, weight and premium price are scaled within a range between 0 and 1.

After the feature scaling step is implemented, the resultant data is divided in the ratio of 7:3. The larger portion of the data will be regarded as the training dataset and the other part will be the testing data. The training and testing data is distinguished by the existence of the dependent or target attribute (medical premium). The training set along with the dependent variable is fed into the regressor model. The model will analyze the hidden relationship among the data and the output value and gather domain knowledge. Later, that knowledge will be evaluated in the testing phase by using testing data. The purpose of the testing procedure is to infer prediction from the unknown instances.

This research work has applied several ensemble-based regression models such as Random Forest, CatBoost, Gradient Boost, XGBoost, Extra trees and AdaBoost for predicting the medical insurance premium price. All these models have encountered a variety of hyper-parameter optimization to ensure better efficiency. The algorithm of the entire process is shown in 2.4.2.1.

2.4.2.1 Algorithm

Step-1: Collect the medical premium dataset.
Step-2: Scaling the features of the dataset is performed as a preprocessing step for prediction.
Step-3: Partition the cleaned data with a ratio of 7:3. Use the major data partition for training and minor for testing.
Step-4: Build the ensemble regressor models namely, Random Forest, CatBoost, Extra Tree, GradientBoost, AdaBoost, and XGBoost.
Step-5: Fine-tune hyper-parameter values for each model.
Step-6: Fit the training data to each of the ensemble regressor.
Step-7: Each model acquires the knowledge and applies it to retrieve predictions for the test data.
Step-8: Calculate the MSE and MAE for each regressor and analyze them to select the promising predictive model.
The block diagram of the above algorithm is described in Figure 5.

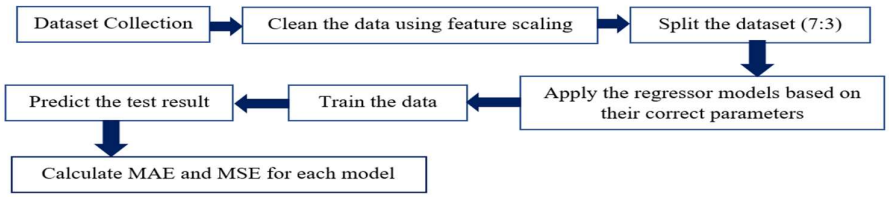

Figure 5 Graphical Representation of Proposed Workflow to Develop Insurance Price Prediction System.

3. Results and discussion

3.1 Hyperparameter tuning and results

Incorporation of multiple weak learners along with their prediction makes the ensemble learning a promising field to explore. This strategy of combining weak learners can produce powerful and improved learners which can provide less error-prone predictive analysis. The elementary parameters of any ML model should be chosen wisely to yield appreciable prediction outcomes. To carry out the hyper-parameter tuning, the estimator count (i.e., the number of base models) for each employed regressor model has been varied within a specified range between 50

to 500 with a step size of 50. It is to be mentioned that the number of estimators is essentially the number of base learners considered for each model. In this section, the performance of each ensemble model is shown by varying the estimator count. In the case of the Random Forest regressor, it is explained in Figure 6(A) and (B) that the estimator size of 100 exhibits the minimized error rate in terms of both MSE and MAE. Hence, the exhaustive comparison among the different number of estimators reveals that this model can attain the numeric prediction with MSE of 0.01 and MAE of 0.0394. The required time to execute the optimized random forest model is 0.173 sec. The CatBoost regression model, the estimator count of 150 has reached the minimized MSE of 0.0155 and MAE of 0.0609. The performance variation among the specified number of base learners is shown in Figure 7 (A) and (B). The required time to execute this model is 0.171541 sec. The Gradient boosting regressor has attained the optimized MSE of 0.0168 and MAE of 0.0607. The estimator count of 300 has shown the best value for MSE whereas the best value for MAE is produced by 250 estimators. Here, a confusion arises regarding optimized estimator count determination. This is why time to build the regressor is taken into consideration to clear the confusion. Experiments have also shown the time required to build the model using 250 estimators requires 0.145611 sec whereas the time required for building the tree having 300 estimators will be 0.176525 sec. This reveals that the gradient boost regressor should be developed by taking 250 estimators. The performance evaluation of this model has been depicted in Figure 8 (A) and (B) with respect to MSE and MAE. As shown in Figure 9 (A) and (B), the extra trees regressor model attains the reduced MSE of 0.0162 and MAE of 0.05 for the estimator count of 50 and 400 respectively. Again, to obtain the best predictive model, the time of execution is taken as the tiebreaker similar to the methods applied to obtain Gradient boosting regressor. The time required to execute this ensemble model with 50 estimators is 0.071773 sec and that of 400 estimators will be 0.558519 sec. Hence, the extra trees regressor model with 50 estimators is recommended. The XGBoost model has reached the minimum error rate of 0.02344 and 0.09722 in terms of MSE and MAE respectively. The lowest MSE and MAE were observed in Figure 10 (A), (B) for the estimator size of 50 and 500 respectively. Since there is a trade-off between choosing the best estimator size, the time needed to implement these models required to be evaluated. The model with 50 estimators takes 0.09722s whereas the model having 400 estimators needs 0.118 sec. Hence, the XGBoost model with estimator size 50 is regarded as the best model. During execution of the adaboost model the same minimized value of MSE and MAE is obtained. The estimator size of 100 has reached the least error value of 0.0496 with a time of 0.2014 sec. The resultant performance is depicted in Figure 11(A) and (B).

The aforementioned implementations are conducted using the scikit-learn library provided by the Python programming language.

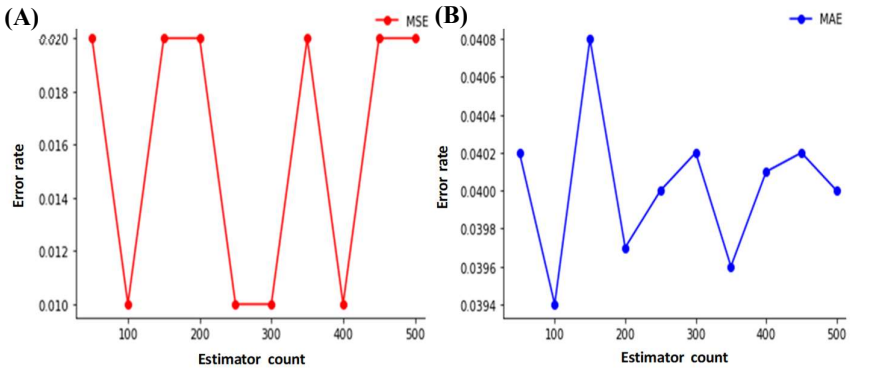

Figure 6 Performance analysis of random forest using (A) MSE, (B) MAE.

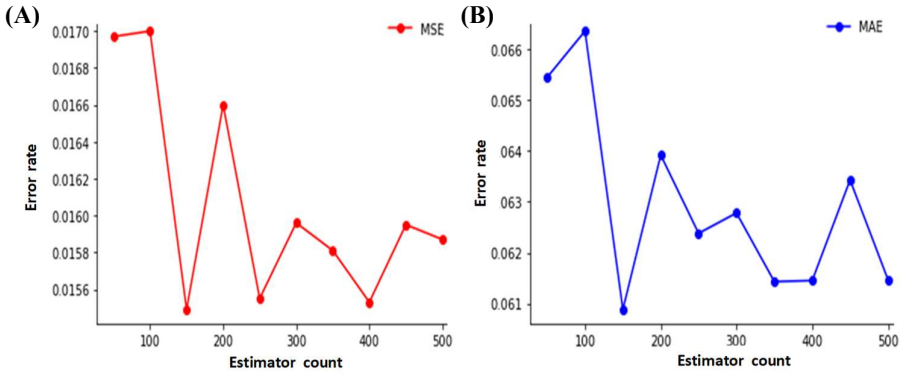

Figure 7 Performance analysis of CatBoost regression using(A) MSE, (B) MAE.

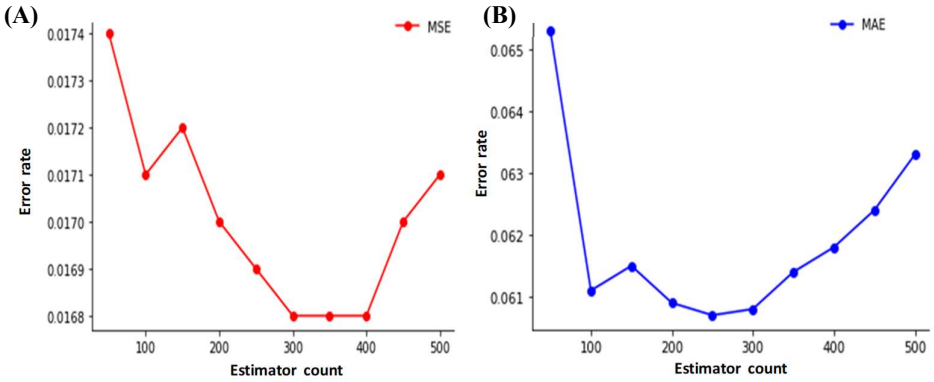

Figure 8 Performance analysis of gradient boost regression using (A) MSE, (B) MAE.

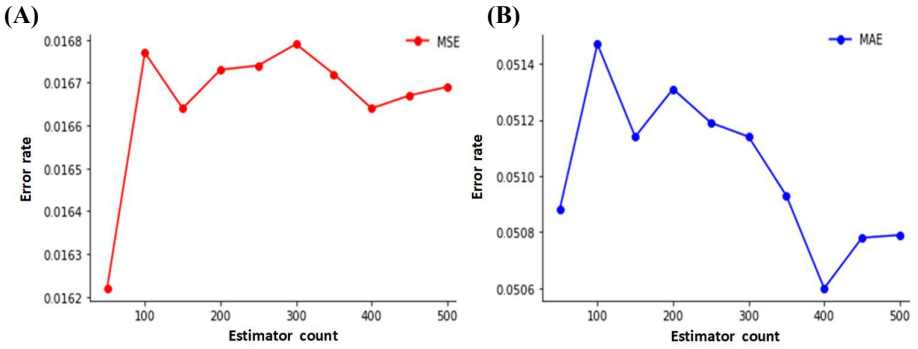

Figure 9 Performance analysis of extra tree regression using (A) MSE, (B) MAE.

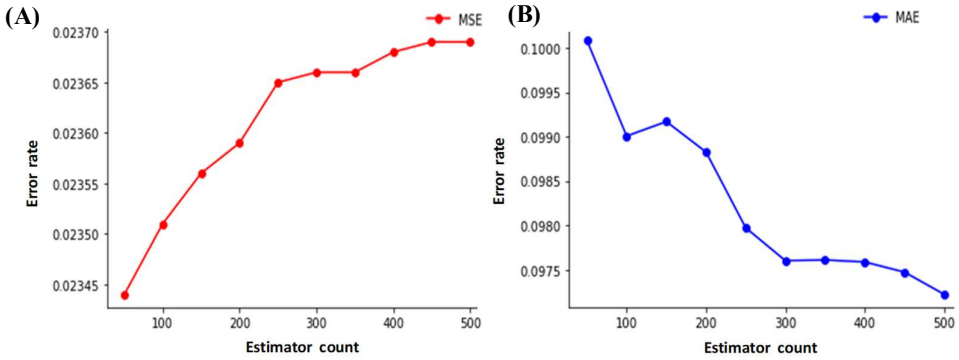

Figure 10 Performance analysis of XGBoost regression model using (A) MSE (B) MAE.

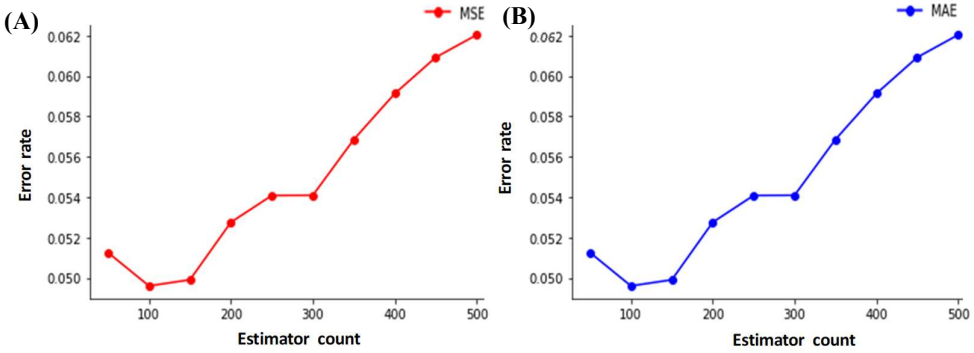

Figure 11 Performance analysis of adaboost regression model using (A) MSE, (B) MAE.

During the hyper-parameter tuning step, the lowest error rate is determined for each ensemble-based model. After completing the hyper-parameter fine-tuning process for each employed model, another exhaustive analysis has been accomplished among all the models in order to explore the best predictive model. The study of Table 3 identifies that the random forest model has exhibited the lowest error rate as compared to other employed models. The minimum error rate of 0.01 and 0.0394 with respect to MSE and MAE has been observed for predicting the medical premium price. Hence, this ensemble model can be employed to construct an automated tool that can provide insights into the health insurance subscription cost.

The features of the dataset have been provided as input to the regressor model in order to assess the premium cost in advance. Each feature should make some contribution while assessing the premium price. Now, the features should be ranked with respect to their importance in the predictive analysis. Since the random forest model needed to construct the automated model, the feature importance matrix is retrieved and shown in Figure 10. The experimental results have confirmed that the age attribute is the one that contributes the most to making the prediction. However, the weight and frequency of a person's transplants are also important to this predictive model. The contribution of other attributes is significantly lower than that of the attributes of age, weight, and frequency of transplants. Table 3 describes the comparative analysis of the described regression models.

Table 3 Comparative analysis of regression models.

Regressor Model Name	MSE	MAE
Random Forest	0.01	0.0394
CatBoost	0.0155	0.06
Gradient Boosting	0.02	0.06
Extra Tree	0.02	0.05
AdaBoost	0.0496	0.0496
XGBoost	0.02365	0.10

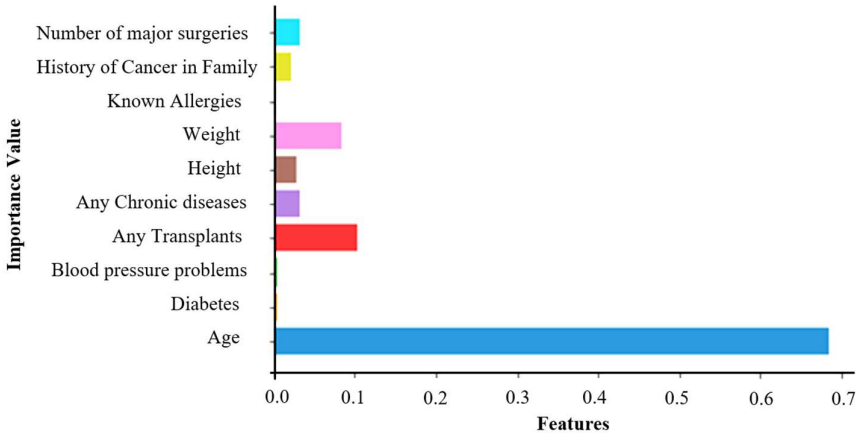

Figure 12 Graphical plot of random forest feature importance.

3.2 Discussion

Public health as well as anything that can prolong life is of utmost importance. As a result, one of the most crucial responsibilities is to build a health-care system. For any country a better healthcare system ensures the better life of every citizen. The most significant factor is the expense of health insurance. As a result, assessing and forecasting the cost of health insurance based on different age groups has proven to be a difficult process. This fact has been verified by the application of ML based techniques. As mentioned in Figure 12, age turns out to be the most important factor for predicting health insurance cost. This result can be regarded as the best predictive outcome as it exhibits the optimized MSE value of 0.01.

The research work carried out in [7] and [8], the MSE values were found to be 0.17448 and 0.1302 respectively. As compared to research studies [7] and [8], the current study has provided more improved predictive performance for medical subscription cost.

Based on the predictive outcome, this automated decision-making system can be applicable to the following real-life scenarios. This system can be beneficial to the companies as they can get an idea regarding the subscription cost of an individual patient based on their health histories. As a consequence, they can set a suitable premium with a profitable margin. This automated model can also be useful for the patients for the advance

premium estimation. Due to the advance estimation, patients optimize their additional expenditures and manage necessary resources.

In recent times, the world is facing a massive life-threatening challenge due to COVID-19 [33,34]. Hence, due to the currently ongoing pandemic created by COVID-19, this tool may be beneficial as it can address the health-related expenses caused during the COVID-19 disease. However, post COVID-19 disease, the medication and treatment expenditures can be estimated by the predictive model by incorporating necessary features in the model.

3.3. Implications

Any disease prediction system in health-care can be supported by the implemented model. This will enable us to identify the disease along with the associated costs required for the treatment of that disease. The automated model requires the relevant factors of a particular disease to reveal the associated cost at an early stage. This can be beneficial for patients in terms of managing treatment resources and expenses.

3.4. Limitations

The implemented predictive system has exhibited quite promising forecasts for medical cost estimation. The model can be further improved in the following areas. The model will perform better when more patient details (data-points) are used to train the model. Inclusion of medical histories of the patients in the model will enable it for better prediction outcomes.

4. Conclusion

An automated procedure has been proposed in this study that would examine the many parameters considered and recommend a premium amount cost for patients. Due to the vast amount of data, ensemble regression models were utilized instead of typical regression models. Finally, the comprehensive analysis among the implemented models demonstrates that the random forest regression method may be used to create the intelligent data-driven model. The experimental results show that the cost of health insurance can be estimated using the best forecasting results, which have MSE values of 0.01 and MAE values of 0.0394.

5. References

[1] Meenakshisundaram KS, Krishnekumaar ST. Age Factor-A Basic Parameter for Health Insurance-A Study with Special Reference to Chennai City among Standalone Health Insurers. Int J Manag Human. 2020;4(5):78-88.

[2] Lee SH, Brown SL, Bennett AA. The relationship between insurance and health outcomes of diabetes mellitus patients in Maryland: a retrospective archival record study. BMC Health Serv Res. 2021;21:495.

[3] Institute of Medicine (US). Committee on the Consequences of Uninsurance. Washington, D.C.: National Academy Press; 2002.

[4] Su Z, McDonnell D, Cheshmehzangi A, Abbas J, Li X, & Cai Y. The promise and perils of Unit 731 data to advance COVID-19 research. BMJ Global Health. 2021;6(5):e004772.

[5] Maqsood A, Abbas J, Rehman G, Mubeen R. The paradigm shift for educational system continuance in the advent of COVID-19 pandemic: Mental health challenges and reflections. Current Res Behav Sci. 2021;2:100011

[6] Perc M, Ozer M, Hojnik. Social and juristic challenges of artificial intelligence. Palgrave Commun.2019;5:61.

[7] Wernly B, Mamandipoor B, Baldia P, Jung C, Osmani V. Machine learning predicts mortality in septic patients using only routinely available ABG variables: a multi-centre evaluation. Int J Med Inform. 2021;145:104312.

[8] Huang YC, Li SJ, Chen M, Lee TS. The Prediction Model of Medical Expenditure Appling Machine Learning Algorithm in CABG Patients. Healthcare (Basel). 2021;9:710.

[9] Sharma DK, Sharma A. Prediction of health insurance emergency using multiple linear regression technique. Eur J Mol Clin Med. 2020;7:98-105.

[10] Nidhi Bhardwaj, Rishabh Anand. Health Insurance Amount Prediction. Int J Eng Res. 2020;V9(05):1008-1011.

[11] Panay B, Baloian N, Pino JA, Peñafiel S, Sanson H, Bersano N. Predicting health care costs using evidence regression. In: Bravo J, González I, editors. The 13th International Conference on Ubiquitous Computing and Ambient Intelligence UCAmI 2019; 2019 Dec 2-5; Toledo, Spain. Basel; MDPI; 2019. p.74.

[12] Yang C, Delcher C, Shenkman E, Ranka S. Machine learning approaches for predicting high cost high need patient expenditures in health care. Biomed Eng. 2018;17:1-20.

[13] Yan Y, Yang D. A stock trend forecast algorithm based on deep neural networks. Sci Program. 2021;2:1-7.
[14] Ye R, An N, Xie Y, Luo K, Lin Y. An Empirical Study on the Equity Performance of China's Health Insurance Companies During the COVID-19 Pandemic-Based on Cases of Dominant Listed Companies. Front Pub Health. 2021;9:663189.
[15] Rao PS, Srinivas K, Mohan AK. A survey on stock market prediction using machine learning techniques. In: Kumar A, Paprzycki M, Gunjan VK, editors. ICDSMLA 2019. Singapore: Springer; 2019. p. 601.
[16] Nayak A, Pai MM, Pai RM. Prediction models for Indian stock market. Proc Comput Sci. 2016;89:441-449.
[17] Shastri R, Rengarajan A. Prediction of Car Price using Linear Regression. Int J Trend Sci Res Dev. 2021;5:866-869.
[18] Asghar M, Mehmood K, Yasin S, Khan ZM. Used Cars Price Prediction using Machine Learning with Optimal Features. Pakistan J Eng Technol. 2021;4:113-119.
[19] Yoosefi LJ, Abbas J, Moradi F, Salahshoor MR, Chaboksavar F, Irandoost SF, et al. How the COVID-19 pandemic effected economic, social, political, and cultural factors: A lesson from Iran. Int J Soc Psychiatry. 2021;67:298-300.
[20] Abbas J, Wang D, Su Z, Ziapour A. The Role of Social Media in the Advent of COVID-19 Pandemic: Crisis Management, Mental Health Challenges and Implications. Risk Manag Health Policy. 2021;14:1917-1932.
[21] Boelaert J, Ollion É. The great regression. Rev Fr Sociol. 2018;59:475-506.
[22] Onan, A. An ensemble scheme based on language function analysis and feature engineering for text genre classification. J Inf Sci. 2018;44:28-47.
[23] Onan A. Classifier and feature set ensembles for web page classification. J Inf Sci. 2016;42(2):150-165.
[24] Livingston F. Implementation of Breiman's random forest machine learning algorithm. ECE591Q Mach Learn J Paper. 2005:1-13.
[25] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63:3-42.
[26] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013;7:21.
[27] Schapire RE. Explaining adaboost. Στο: Empirical inference. Berlin: Springer; 2013:37-52.
[28] Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Aug 13-17; San Francisco California: USA. New York: Association for Computing Machinery; 2016.
[29] Hong J. An Application of XGBoost, LightGBM, CatBoost Algorithms on House Price Appraisal System. Hous Finance Res. 2020;4:33-64.
[30] Rong S, Bao-wen Z. The research of regression model in machine learning field. InMATEC Web of Conferences 2018. EDP Sci. 2018;176:01033.
[31] Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res. 2005;30:79-82.
[32] TejasBard. Medical Insurance Premium Prediction Predict Yearly Medical Cover Cost (₹) Daily, Kaggle Data v2 [Internet]. 2021 [cited 2021 August 20]. Available from: https://www.kaggle.com/tejashvi14 /medical-insurance-premium-prediction.
[33] Aqeel M, Abbas J, Shuja K.H, Rehna T, Ziapour A, Yousaf I, et al. The influence of illness perception, anxiety and depression disorders on students mental health during COVID-19 outbreak in Pakistan: a Web-based cross-sectional survey. Int J Human Rights Health. 2021;14:1-14.
[34] NeJhaddadgar N, Ziapour A, Zakkipour G, Abbas J, Abolfathi M, Shabani M. Effectiveness of telephone-based screening and triage during COVID-19 outbreak in the promoted primary healthcare system: a case study in Ardabil province, Iran. Z Gesundh Wiss. 2020;29:1-6.