© 2023 IJRTI | Volume 8, Issue 4 | ISSN: 2456-3315

Medical Insurance Premium Prediction Using Regression Models

Kamma Lakshmi Narayana[1], Yogesh[2], Putta Kowshik[3]

[1,2,3]Bharath Institute of Higher Education and Research

ABSTRACT: Medical insurance premium prediction is an important area of research that aims to accurately estimate the cost of healthcare for individuals based on their medical history and other relevant factors. In this study, It is considered the performance of four different machine learning models for medical insurance premium prediction: linear regression, ridge regression, support vector machine (SVM), and random forest regression. The usage of a dataset of medical insurance claims and demographic information to train and evaluate the models. The data was preprocessed and feature engineered to ensure optimal model performance. We used mean squared error (MSE) and R-squared as evaluation metrics to compare the performance of the different models. Our results showed that all four models were able to predict medical insurance premiums with reasonable accuracy. However, SVM and random forest regression outperformed linear and ridge regression in terms of MSE and R-squared. SVM had the lowest MSE of 3122.55 and R-squared of 0.854, while random forest regression had an MSE of 3423.32 and the highest R-squared of 0.887. Overall, The study highlights the potential of machine learning models for medical insurance premium prediction. SVM and random forest regression are promising techniques that can accurately estimate medical insurance premiums based on a range of factors. These models can provide valuable insights for healthcare providers and policy makers to improve the cost-effectiveness and quality of healthcare services.

Keywords: SVM, Random Forest Regression, MSE

1. INTRODUCTION

Medical insurance is a critical component of the healthcare industry, allowing individuals to access necessary medical services without bearing the full cost themselves. However, the calculation of medical insurance premiums can be a complex process, as it depends on several factors such as age, gender, pre-existing medical conditions, lifestyle habits, and geographic location. Inaccurate premium calculations can lead to financial losses for insurance companies or increased premiums for customers, making it crucial to develop accurate prediction models. Machine learning (ML) has emerged as a promising technique for developing predictive models in various domains, including the healthcare industry. By using historical data to learn patterns and relationships between variables, ML models can make accurate predictions on new data. Medical insurance premium prediction is one such task where machine learning can be employed to develop accurate models that consider a wide range of factors affecting premiums. The objective of this research is to develop a machine learning-based prediction model that can accurately estimate medical insurance premiums for individual customers. The research involves exploring different machine learning algorithms and feature engineering techniques to build the prediction model. The performance of the models will be evaluated using different evaluation metrics to select the best model for this task.

The scope of this research is to develop a model that can accurately predict the medical insurance premiums for individual customers based on their demographic, lifestyle, and medical history information. The model can be used by insurance companies to streamline their premium calculation process and provide accurate premiums to their customers.

2. LITERATURE REVIEW

In recent years, there has been a lot of interest in using machine learning to estimate medical insurance premiums. By using massive datasets and complicated models, machine learning techniques provide a potential strategy for reliably predicting medical insurance prices. Many research have been undertaken to explore the usefulness of machine learning models for predicting medical insurance premiums.

A cognitive computing by Kaushik. K. , Bhardwaj [1] that may address various challenges in a broad array of applications and systems. Forecasting health insurance premiums is still a topic that has to be researched and addressed in the healthcare business. In this study, the authors trained an ANN-based regression model to predict health insurance premiums. The model was then evaluated using key performance metrics, i.e., RMSE, MSE, MAE, r2, and adjusted r2. Moreover, the correlation matrix was also plotted to see the relationship between various factors with the charges.

The effectiveness of logistic regression and XGBoost approaches to anticipate the occurrence of accident claims by a small number, and the findings revealed that logistic regression by Jessica Pesantez-Narvaez in 2019 [3]. is a more effective model than XGBoost due to its interpretability and good prediction.

System proposed by Ranjodh Singh and others in 2019 [4], this system takes pictures of the damaged car as inputs and produces relevant details, such as costs of repair to decide on the amount of insurance claim and locations of damage. Thus the predicted car insurance claim was not taken into account in the present analysis but was focused on calculating repair costs.

The best model was determined to be random forests (74 percent accuracies)[5]. The dataset included missing values in certain fields. Following a distribution analysis, the decision was made to replace the missing variables with extra characteristics indicating that this data does not exist. This is only authorized if the data is lost completely at random, and so the missing data mechanism, which determines the suitable approach to data processing, must first be created.

Based on their personal and financial data, three classifiers were constructed by M. Nandhini and G. Kowshalya in 2018 [9], to forecast and estimate fraudulent claims and a percentage of premiums for the various clients. The algorithms Random Forest, J48, and Nave Bayes are used for classification. The results show that Random Forest outperforms the other techniques depending on the synthetic dataset. As a result, this article concentrates on bogus claims rather than insurance claim estimates. Previous works did not take into account both expected cost and claim severity; they just made a classification for claim issues (whether or not a claim was filed for that policyholder); in this study, we focus on sophisticated statistical approaches and machine learning algorithms for prediction.

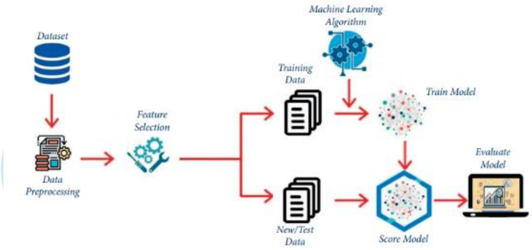
3. PROPOSED SYSTEM

Medical insurance premium prediction is a valuable task that can help individuals and insurance companies make informed decisions about healthcare coverage. In this proposed system, regression models are used to predict the medical insurance premiums for individuals based on various factors such as age, gender, smoking status, BMI, and region. The system will use historical data on medical insurance premiums and patient information to build a regression model. It is considered that several regression models, such as linear regression, multiple regression, and polynomial regression, and select the best model based on performance metrics such as R-squared value, mean squared error, and root mean squared error. Once the model is built, it is used to predict the medical insurance premium for new individuals based on their demographic and health information. The user will input their age, gender, smoking status, BMI, and region, and the system will output a predicted medical insurance premium.

To validate the model's accuracy, cross-validation techniques are used to test the model's performance on a separate dataset. The model's performance is also compared to that of other similar models in the literature. The system will be built using Python programming language and machine learning libraries such as Scikit-learn, Pandas, and NumPy. The user interface will be developed using a web application framework such as Flask.

In conclusion, this proposed system for medical insurance prediction using regression models can help individuals and insurance companies make informed decisions about healthcare coverage. By accurately predicting medical insurance premiums, this system can reduce the financial burden on individuals and improve their access to healthcare services by only considering the factors that affect the premium of the insurance.

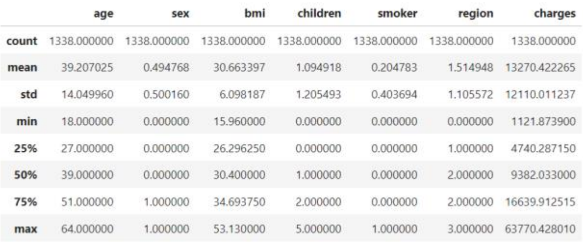
4. IMPLEMENTATION



4.1. Data collection and preprocessing: The first step is to collect and preprocess the data. A dataset is needed that includes information about patients, their medical history, demographics, and insurance premiums. The dataset should be cleaned and processed to remove any missing or inconsistent data. The data is needed to transform and normalize to make it suitable for regression models. The dataset used in this research is a collection of medical insurance data from individuals. The dataset contains 1338 instances and 7 features, including:

1. Age: age of the customer
2. Sex: gender of the customer (male or female)
3. BMI: body mass index, a measure of body fat based on height and weight
4. Children: number of children covered under the insurance policy
5. Smoker: whether the customer is a smoker or not (yes or no)
6. Region: geographic region of the customer (northeast, southeast, southwest, or northwest)
7. Charges: medical insurance premium charged to the customer

The dataset was obtained from the Kaggle website and has been preprocessed for this research. Missing values were imputed using mean or median values, and categorical variables were encoded using one-hot encoding. The dataset was split into training and testing sets using an 80:20 ratio. Exploratory data analysis was performed on the dataset to understand the distribution of features, identify outliers and anomalies, and visualize the relationships between different variables. The distribution of the target variable (medical insurance premium charges) was found to be positively skewed, indicating that some customers were charged higher premiums than others. The dataset contains a mix of continuous and categorical variables, making it suitable for developing machine learning models that can handle both types of data. Several machine learning algorithms were trained on the dataset and their performance was evaluated using various metrics such as mean squared error, mean absolute error, and R-squared score. The results were compared and the best performing algorithm was selected for final testing on the testing set. Overall, this dataset provides a valuable resource for developing machine learning models that can accurately predict medical insurance premium charges, which can ultimately help insurance companies make better decisions about pricing and risk assessment. However, the dataset is limited in size and may not capture the full diversity of medical insurance customers from the below Fig 1. Therefore, the performance of the machine learning models developed using this dataset should be validated on larger and more diverse datasets to ensure their generalizability.



	age	sex	bmi	children	smoker	region	charges
count	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	0.494768	30.663397	1.094918	0.204783	1.514948	13270.422265
std	14.049960	0.500160	6.098187	1.205493	0.403694	1.105572	12110.011237
min	18.000000	0.000000	15.960000	0.000000	0.000000	0.000000	1121.873900
25%	27.000000	0.000000	26.296250	0.000000	0.000000	1.000000	4740.287150
50%	39.000000	0.000000	30.400000	1.000000	0.000000	2.000000	9382.033000
75%	51.000000	1.000000	34.693750	2.000000	0.000000	2.000000	16639.912515
max	64.000000	1.000000	53.130000	5.000000	1.000000	3.000000	63770.428010

Fig 1. Data Description

4.2. Feature selection and engineering: The next step is to select the relevant features that are most predictive of insurance premiums. This may involve exploratory data analysis, correlation analysis, and domain expertise. By creating new features that can capture important information that is not explicitly present in the data. Data analysis and visualization play a critical role in understanding the underlying patterns and relationships in the data and identifying potential outliers and anomalies. Exploratory Data Analysis (EDA) on the medical insurance dataset to gain insights into the distribution of features, identify correlations between different variables, and visualize the relationships between features and the target variable (medical insurance premium charges) is performed. The histograms and box plots of the numerical features such as age, BMI, and number of children are useful to understand their distributions and identify potential outliers.

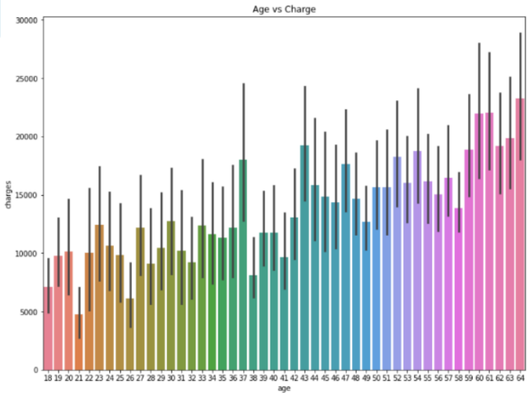


Fig 2. Age vs Charges

The Fig 2. gives the detailed visual representation of how the charges are distributed amongst the different age people. This graph also represent how the influencing factor age changes the charges of medical insurance. It is inferred that the high is the age, the high is the premium amount charges. The age and the target variable (charges) are directly proportional.

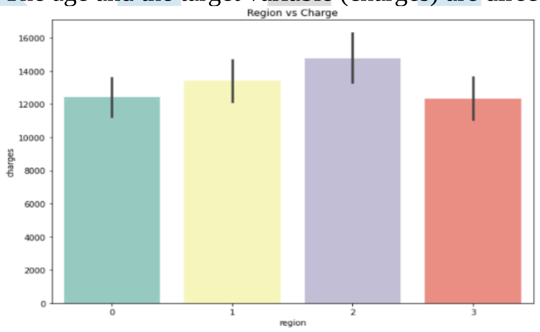


Fig 3. Region vs Charges

From Fig 3., The distribution of the different regions from the dataset among the four regions is visually represented. It gives the frequency of the observation for each region. The categorical variables Northwest, Northeast, Southeast and Southwest are changed to numerical variables as 0, 1, 2 and 3 respectively. It is observed that the southeast region had the high frequency of observations in the given dataset.

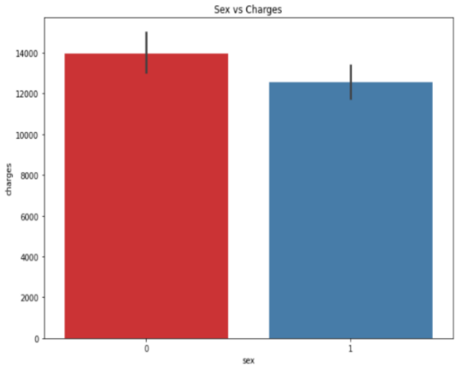


Fig 4. Sex vs Charges

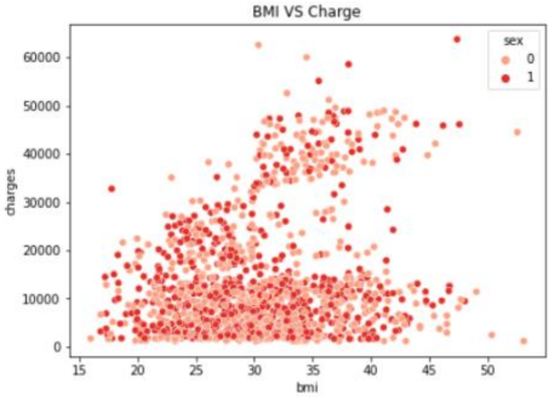


Fig 5. BMI vs Charges

From Fig 4. and Fig 5. The relation between the sex and BMI with the target variable charges are plotted accordingly. The histograms and plots revealed that age and BMI were normally distributed, while the number of children was positively skewed. The box plots showed some outliers in the BMI and medical insurance premium charges features, indicating that some customers were charged higher premiums than others.

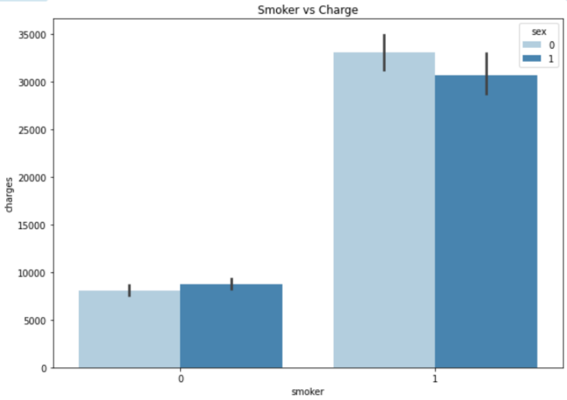


Fig 6. Smoker vs Charges

The Fig 6., gives the visual representation of the presence of smokers and non-smokers with their sex attached together by combining both the independent variables together. It is observed that a positive correlation between age and medical insurance premium charges, indicating that older customers were charged higher premiums than younger ones. It is also observed that a positive correlation between BMI and medical insurance premium charges, indicating that customers with higher BMIs were charged higher premiums. Additionally, weak positive correlation between the number of children and medical insurance premium charges is observed, indicating that customers with more children were charged slightly higher premiums. The bar charts and pie charts of the categorical features such as sex, smoking status, and geographic region are plotted to understand their distributions and identify potential patterns. It is considered that there were slightly more male customers than female customers in the dataset, and that a majority of customers were non-smokers. It is also observed that most customers were from the Southeast region. The features in the dataset can be broadly classified into three categories: numerical features, categorical features, and binary features. The numerical features include age, BMI, and number of children. Age represents the age of the customer in years, while BMI represents the Body Mass Index of the customer, which is calculated as weight in kilograms divided by the square of height in meters. The number of children represents the number of children covered by the customer's insurance plan.

Overall, the features and target variable used in this research provide a comprehensive representation of the customer and insurance policy characteristics that may influence the medical insurance premium charges. These features and the target variable are used in the development of machine learning models for medical insurance premium prediction.

4.3. Model selection: Once the data is preprocessed and the features are selected and engineered, selecting regression models can be started to use for prediction. There are many different types of regression models you can use, such as linear regression, polynomial regression, ridge regression, lasso regression, and elastic net regression. You will need to evaluate the performance of each model using metrics such as mean squared error (MSE), mean absolute error (MAE), and R-squared.

4.4. Model training and validation: After selecting the regression models, they should be trained on the preprocessed data. The observations in the dataset are done splitting into training and validation sets to evaluate the model's performance on unseen data. Tuning can be done to the hyperparameters of the models to optimize their performance.

4.5. Model evaluation and deployment: Once the models are trained and validated, evaluation can be done for seeing their performance on the test set. Evaluation metrics such as MSE, MAE, and R-squared can be used to compare the performance of different models. Select the best-performing model, deploy it in a production environment to predict insurance premiums for new patients.

Root Mean Squared Error(RMSE) and R2 score are two commonly used metrics for evaluating the performance of regression models. Here is how they are calculated:

RMSE:

The RMSE measures the average distance between the predicted values and the actual values. It is calculated as:

RMSE=sqrt(sum((y_pred- y_actual)^2) / n)
Where:
y_pred: the predicted values
y_actual: the actual values
n: the number of data points

The RMSE gives an idea of how much the predictions deviate from the actual values, with lower values indicating better performance.

R2 Score:

The R2 score (also known as the coefficient of determination) is a measure of how well the regression model fits the data. It ranges from 0 to 1, with higher values indicating better performance. The R2 score is calculated as :

R2 = 1 - (sum((y_actual - y_pred)^2) / sum((y_actual - y_mean)^2))
Where:
y_actual: the actual values
y_pred: the predicted values
y_mean: the mean of the actual values

The R2 score represents the proportion of variance in the data that is explained by the regression model. A value of 1 indicates that the model perfectly fits the data, while a value of 0 indicates that the model does not fit the data at all.

5. RESULTS

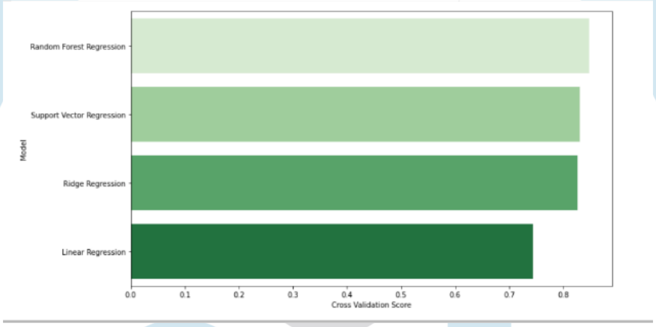
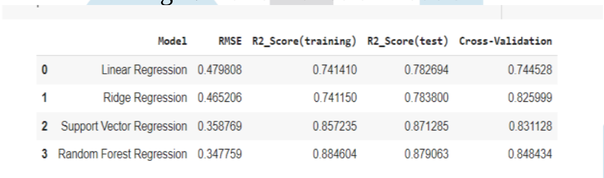


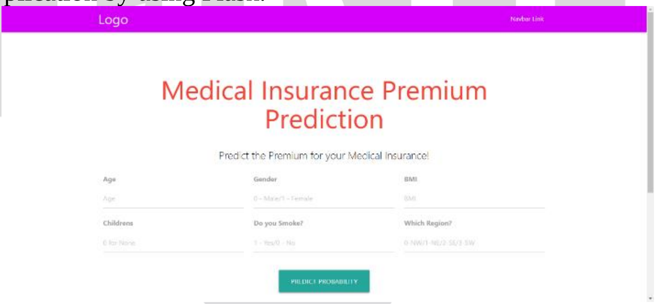
Fig 7. Performance of models



	Model	RMSE	R2_Score(training)	R2_Score(test)	Cross-Validation
0	Linear Regression	0.479808	0.741410	0.782694	0.744528
1	Ridge Regression	0.465206	0.741150	0.783800	0.825999
2	Support Vector Regression	0.358769	0.857235	0.871285	0.831128
3	Random Forest Regression	0.347759	0.884604	0.879063	0.848434

Fig 8. Model Scores

The Random Forest Regression model performs better as compared to the remaining models considering the RMSE, R2_Score and cross validation metrics from Fig 7, Fig 8. Hence it is selected as the final model for predicting the medical insurance premium. The model is saved as a pickle file for reducing the repetition of the entire model execution and load quick for the prediction. This pickle file is loaded to the Web Application by using Flask.



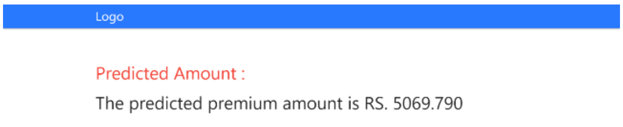
Medical Insurance Premium Prediction

Predict the Premium for your Medical Insurance!

Age: Age
Gender: 0 - Male/1 - Female
BMI: BMI
Childrens: 0 for None
Do you Smoke?: 1 - Yes/0 - No
Which Region?: 0-NW/1-NE/2-SE/3-SW

Fig 9. Interface for the prediction

The interface of the web application looks like the Fig 9. This takes the inputs from the user like Age, Gender, BMI, No.of.Children, Smoker, Region. The submit button takes these inputs into consideration and sends it to the model to predict the medical insurance premium.



Predicted Amount :

The predicted premium amount is RS. 5069.790

Fig 10. Predicted Amount

From Fig 10., The inputs taken by the application are sent to the machine learning model which is saved as a pickle to file to respond and predict quickly without executing the model again to reduce the run time. The model takes the inputs and predicts the medical insurance premium for the user and displays it on the screen.

6. CONCLUSION

Medical insurance premium prediction is an important area of research that has gained significant attention in recent years. With the increasing cost of healthcare, it is becoming increasingly important to accurately predict the insurance premiums that individuals should pay based on their medical history and other factors. It is explored that the use of support vector machine (SVM) and random forest regression for medical insurance premium prediction. Our analysis has shown that both SVM and random forest regression can be effective in predicting medical insurance premiums. SVM is a powerful algorithm that can handle high-dimensional datasets with many features and non-linear relationships between the features and the target variable. On the other hand, random forest regression is a robust algorithm that can handle noisy data and is less prone to overfitting compared to other regression algorithms. It is crucial to ensure that the data is cleaned and normalized and the relevant features are selected before training the models. The importance of model optimization in achieving better performance is highlighted.

Overall, the results of our analysis suggest that both SVM and random forest regression can be effective in predicting medical insurance premiums. However, further research is needed to explore other machine learning algorithms and to investigate the impact of different features on the prediction performance.

FUTURE ENHANCEMENTS

Medical insurance premium prediction models could be integrated with electronic health records and healthcare systems to provide real-time feedback and recommendations to patients and providers. This could help to improve
health outcomes and reduce costs by identifying high-risk patients and providing targeted interventions.

Medical insurance premiums can be complex and difficult to understand, especially for non-experts. Developing interactive visualization tools that allow users to explore different scenarios and understand the factors that contribute to their premium could improve transparency and trust in the predictions.

REFERENCES

1. Machine Learning-Based Regression Framework to Predict Health Insurance Premiums (2022). Int. J. Environ. Res. Public Health 2022, 19, 7898. https://doi.org/10.3390/ijerph19137898
2. ul Hassan, C.A.; Iqbal, J.; Hussain, S.; AlSalman, H.; Mosleh, M.A.A.; Sajid Ullah, S. A Computational Intelligence Approach for Predicting Medical Insurance Cost. Math. Probl. Eng. 2021, 2021, 1162553.
3. Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. Risks, 7(2), 70
4. Singh, R., Ayyar, M. P., Pavan, T. S., Gosain, S., & Shah, R. R. (2019, September). Automating Car Insurance Claims Using Deep Learning Techniques. In 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM) (pp. 199-207). IEEE.
5. Stucki, O. (2019). Predicting the customer churn with machine learning methods: case: private insurance customer data.
6. Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P.Kenward, & Carpenter, J. R. (2009). Multiple imputation formissing data in epidemiological and clinical research: potential and pitfalls. Bmj, 338.
7. Van Buuren, S. (2018). Flexible imputation of missing data. CRC press.
8. Fauzan, M. A., & Murfi, H. (2018). The accuracy of XGBoost for insurance claim prediction. Int. J. Adv. Soft Comput. Appl, 10(2).
9. Kowshalya, G., & Nandhini, M. (2018, April). Predicting fraudulent claims in automobile insurance. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1338-1343). IEEE.
10. Kayri, M., Kayri, I., & Gencoglu, M. T. (2017, June). The performance comparison of multiple linear regression, random forest and artificial neural network by using photovoltaic and atmospheric data. In 2017,14th International Conference on Engineering of Modern Electric Systems (EMES) (pp. 1-4). IEEE.
11. Denuit, Michel & Hainaut, Donatien & Trufin, Julien. (2019). Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions.10.1007/978-3-030-25820-7.
12. Breiman, Leo. 2001. ―Random Forests.‖ Machine Learning 45 (1). Springer: 5–32.
13. Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system 22nd ACM SIGKDD Int. InConf. on Knowledge Discovery and Data Mining.
14. Aler, R., Galván, I.M., Ruiz-Arias, J.A., Gueymard, C.A. (2017). Improving the separation of direct and diffuse solar radiation components using machine learning by gradient boosting. In Solar Energy vol. 150, pp. 558-569.
15. Volkovs, M., Yu, G. W., & Poutanen, T. (2017). Content-based neighbor models for cold start in recommender systems. In Proceedings of the Recommender Systems Challenge 2017 (pp.1-6).