European Journal of Technique

journal homepage: https://dergipark.org.tr/en/pub/ejt

Vol.14, No.1, 2024

Research Article

Medical Insurance Cost Prediction MedCost: Machine Learning Ensemble Approaches

Murat EMEÇ[1*]

[1*]Istanbul University, Department of Information Technology, 34116, Fatih, Istanbul, Turkey. (e-mail: murat.emec@istanbul.edu.tr).

ARTICLE INFO

Received: Oct., 16. 2023
Revised: Dec., 31. 2024
Accepted: Jan, 14. 2024

Keywords:
Medical cost
Healthcare insurance
Machine Learning
Ensemble methods
Regression

Corresponding author: *Murat EMEÇ*

ISSN: *2536-5010* | e-ISSN: *2536-5134*

DOI: https://doi.org/10.36222/ejt.1375677

ABSTRACT

Healthcare insurance costs are a significant concern for individuals and providers. Accurately predicting these costs can assist in financial planning and risk assessment. This study explores machine learning ensemble methods to predict healthcare insurance costs based on various factors, including age, sex, body mass index (BMI), number of children, smoking status, and region. Additionally, new features were introduced by incorporating the mean and standard deviation of BMI and smoking habits, which are known to affect insurance costs substantially.

The study began with a comprehensive statistical analysis of the dataset, followed by feature engineering to enhance its predictive power. Categorical variables such as sex, smoking status, and region were appropriately encoded. Two datasets were constructed: one containing all the original features and the other having the engineered features. Ensemble learning methods, including Bagging, Stacking, and the proposed MedCost-AdaBoost model, were employed to predict the insurance costs for both datasets. The results revealed that the MedCost-AdaBoost model outperformed the other methods in terms of lower Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values, along with higher R-squared (R^2) scores.

These findings underscore the effectiveness of ensemble learning techniques in predicting healthcare insurance costs, with feature engineering playing a crucial role in improving prediction accuracy. Despite certain limitations, such as the dataset size, this study provides valuable insights for researchers and professionals in the healthcare insurance industry. Future research could explore additional factors and larger datasets to enhance the predictive models in this domain further.

1. INTRODUCTION

Health insurance is a vital financial tool that facilitates access to healthcare services and offsets unexpected medical expenses. However, health insurance companies' risk assessment and premium determination processes are complex and involve several factors. These factors include the insured's age, sex, medical history, region of residence, insurance plan coverage, etc. Machine learning and data analytics play essential roles in health insurance cost estimation.

Health insurance companies use big data and machine-learning techniques to offer fairer premiums to customers and better manage their financial risks. Therefore, health insurance cost estimations are essential for insurance companies and PHs. This study focuses on health insurance cost estimation and examines methods for improving these estimates using machine learning ensemble approaches. Additionally, the importance of health insurance cost estimation and the advantages of accurate estimation are essential areas of discussion.

Health insurance cost estimation is essential for several reasons.

- Fair Premium Setting: Health insurance companies must offer fair and competitive premiums to their customers. Accurate cost estimates help set insurance premiums fairly. This implies more favorable and competitive insurance prices for customers.
- Risk Management: Health insurance companies use accurate cost estimates to forecast payments and manage financial risks. Inaccurate estimates can cause financial difficulties for insurance companies.
- Insured Making Informed Decisions: Individuals need accurate information to understand their health insurance options and costs. Precise cost estimates help insureds make informed decisions.

Access to Health Services: Accurate cost estimates can help insurance companies provide individuals access to affordable

healthcare services. This contributes to making healthcare services accessible to the broader population.

Therefore, accurate and precise health insurance cost estimations are essential. Machine-learning ensemble approaches offer a way to improve these estimates and make better decisions. This study examines how health insurance cost estimation can be improved and how ensemble methods can be used.

The primary motivation and contributions of this study are as follows:

- Accurately estimating health insurance costs is challenging with current diagnostic methods. In particular, individuals' health profiles are complex and include age, sex, smoking habits, BMI, and other essential factors.
- Traditional methods are often inadequate for estimating health insurance costs. For example, some models do not adequately consider certain factors, while others cannot accurately identify effects and relationships.
- MedCost-AdaBoost, the method presented in this study, can make more precise estimates, especially by considering certain health profile factors. In particular, with new features, including determinants such as BMI and smoking habit, it has a predictive capability beyond traditional methods.
- This study aims to contribute to related studies to estimate health insurance costs better. The highlights of MedCost-AdaBoost contribute to more accurate financial planning and risk assessment in industry and academia.

The remainder of this paper works as Section 2. In Section 3, the materials and methods used are described. Section 4 presents the results and evaluation. A discussion and comparison are presented in Section 5. Finally, conclusions are presented in Section 6.

2. RELATED WORKS

Emerging technologies and growing data sources have become important research topics in health insurance cost estimation, machine learning, and data analytics. This section reviews the literature on health insurance cost estimations and machine-learning ensemble methods. Health insurance cost estimation is essential for healthcare providers, insurance companies, and individuals, and accurate estimates play a critical role in financial planning and resource management. In this context, they understand the contributions of previous studies on health insurance cost estimation and how machine learning ensemble methods can be used to form the basis of this research.

Research using various machine learning algorithms and data mining techniques has enriched the literature on health insurance and cost estimations. Studies such as those in [1,2] have highlighted how traditional machine learning algorithms can predict insurance claims and medical insurance prices. To improve the interpretability of the models, explainable artificial intelligence (XAI) methods were investigated, and health insurance datasets were considered [3]. XGBoost and machine learning-based regression methods have been investigated for health insurance premium prediction [4,5] and focused on current methods of estimating health insurance costs [6-8]. Finally, studies have been conducted to assess health insurance claim costs using data mining techniques [9-10]. Related studies reflect various health insurance cost estimation approaches and shed light on future research. In this context, to update the literature, we have also included critical recent works [11-17]. These references reflect current applications of ensemble methods for estimating health insurance costs and new developments in the research field. Recently, many studies have focused on regression techniques based on machine learning techniques. [18-21].

A review of these studies showed that machine learning and data mining techniques have been widely used in health insurance cost estimations, and significant results have been achieved. However, each of these approaches has advantages and limitations. Therefore, this study adopted a strategy called ensemble modeling. Ensemble modeling allows us to reach more robust and stable predictions by combining different machine learning algorithms. This study uses this approach to evaluate stacking, bagging, and boosting ensemble methods and examines their impact on health insurance cost prediction. The results based on the ensemble model can significantly improve health insurance cost estimations by providing more precise and reliable estimates.

3. MATERIALS AND METHODS

This section presents detailed information on the dataset, methods, and analyses used for health insurance cost estimation. The main objective of this study was to estimate health insurance costs using different ensemble methods and evaluate the performance of these estimates. First, the dataset's sources, components, and importance are explained. The data preprocessing steps, feature engineering techniques, and ensemble methods are described in detail. This study's health insurance cost estimation methodology was carefully designed and implemented to obtain accurate and reliable estimates.

3.1. Materials

3.1.1. Data Set and Features

The dataset used to understand and estimate the key factors affecting health insurance costs and the characteristics of this dataset are critical. This section discusses the data sources that form the basis of our health insurance cost estimation analysis and the essential features of these sources. To better understand the various factors determining health insurance costs, we examined patient characteristics such as age, sex, body mass index (BMI), number of children, smoking habits, region of residence, and individual health costs. These characteristics provide the primary data for estimating and analyzing health insurance costs. Therefore, by exploring this dataset and its characteristics more closely, we aim to better understand these variables' impact on health insurance cost estimation. The features of the dataset are as follows.

- Age is an essential factor in health insurance costs. Usually, older individuals pay higher insurance fees because of the increased risk of health problems.
- Sex: Sex can affect health insurance fees. Women, especially those of childbearing age, may tend to use more health services, but the effect may vary depending on other factors.
- Body mass index (BMI): BMI assesses whether a person has a healthy body weight. A high BMI can indicate an increased risk of chronic disease, leading to higher healthcare costs.

- Child: The number of dependent children can affect insurance costs. More dependents usually pay higher insurance costs.
- Smokers: Smoking is a risk factor for many health problems, and smokers often have higher health insurance rates.
- Region: The region where you live can affect health insurance costs. The costs of living, healthcare costs, and health habits in different areas can vary.
- Charges: Individual medical costs incurred by health insurance. All other factors influence them, and the main objective of this analysis is to understand the relationship between independent variables (age, sex, BMI, children, smokers, and region) and costs.

Assessing the impact of these essential features and datasets on health insurance cost estimation is one of the central aims of this study. Using these characteristics, we attempt to estimate health insurance costs by evaluating different ensemble methods. A better understanding of how age, sex, BMI, number of children, smoking habits, and region of residence affect health insurance costs is needed to provide an essential roadmap for future health insurance planning and decisions.

3.1.2. Statistical analysis of the dataset

This section focuses on the statistical analysis of the dataset used to estimate health insurance costs. The statistical properties of the dataset, its distribution, and basic summary statistics formed the basis of this analysis. We will also use correlation analysis to examine the relationships between the features and visual studies to understand the overall structure of the dataset better. These statistical analyses will help us to understand the factors influencing health insurance costs and how these factors are interrelated. This approach helps us identify the variables of interest when building future forecasting models.

TABLE I
THE STATISTICAL DISTRIBUTION OF THE MEDICAL COST DATASET

Feature(s)	Count	Mean	STD	MIN	MAX
AGE	1338	39.207	14.049	18	64
SEX	1338	0.5	0.5	0.0	1
BMI	1338	30.664	6.097	1.596	5.313
CHILDREN	1338	1.094	1.205	0	5
SMOKER	1338	0.2	0.4	0	1
REGION	1338	1.5	1	0	3

The statistics in Table 1 describe the distribution and key features of each characteristic of the dataset. For example, age generally comprises young and middle-aged individuals, sex is evenly distributed, BMI is usually between 26 and 34, and the proportion of smokers is low. It also shows that wages are spread over a wide range and have large standard deviations. These statistics help us better understand the characteristics of the dataset and allow us to prepare for the analysis. In particular, the wide distribution of wages is essential for studying their impact on cost estimations, as this can help us understand how health insurance costs are associated with different factors. More advanced statistical methods, such as correlation and visual analyses, will help us examine the relationships between these characteristics in more detail.

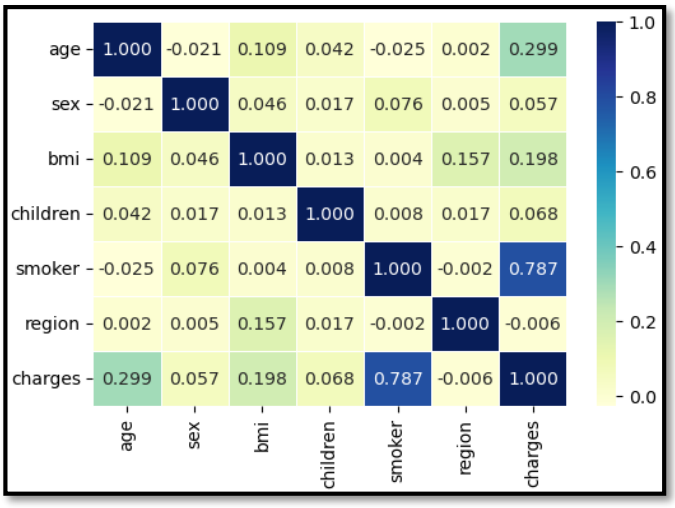

Figure 1. The correlation matrix of the medical cost dataset

The correlation matrix shown in Figure 1 measures the relationships between the variables in the dataset. This correlation determines the strength and direction of the relationship between two variables. The following is an interpretation of the correlations in the matrix:

- Age–charge correlation (0.299): A positive correlation of 0.299 was shown between age and charge. This finding indicates that insurance charges generally increase with age.
- Sex-charge correlation (0.057): A weak positive correlation exists between sex and charges. However, this correlation indicates that sex is not among the most critical factors in determining charges.
- Body mass index (BMI) – was correlated (0.198): A positive correlation of 0.198 existed between BMI and medical status. This finding indicates that higher BMI is generally associated with higher health insurance charges.
- Child-charge correlation (0.068): A weak positive correlation exists between the number of children and charges. However, this correlation was also lower than that of the other factors.
- Smokers-charge correlation (0.787): A strong positive correlation exists between smokers and charges. Smokers generally pay higher health insurance costs.
- Region charge correlation (-0.006): A very low negative correlation exists between the region and charges. This indicates that the region's influence on determining the charges is almost negligible.

These correlations show the relationships between factors that affect health insurance costs. Smoking and BMI appear to have a more significant impact on wages. Predictions. The grid search method was preferred for hyperparameter tuning. The MedCost-AdaBoost method combines weak learners with weights to develop a robust forecasting model. This model positively impacts the ability of the model to predict health insurance costs more accurately. The results and evaluation section presents the results and performance of the proposed method in more detail.

3.1.3. Feature engineering

Feature engineering, an essential step in the data analysis and modeling process, is being applied to estimate health insurance costs more accurately and better understand the factors behind these costs. In this phase, we plan to add new features and transform the existing features to enrich and better understand our current dataset. This approach allows us to understand better the critical factors that influence health insurance costs, thus enabling us to construct more powerful and precise predictive models. We will also code categorical variables appropriately, making them understandable in our model. We are fortunate in handling incomplete data because our dataset is complete, which allows us to obtain more reliable results. At this stage, we aim to step forward in health insurance cost estimation by making our dataset more robust and meaningful. The feature engineering stages are listed below:

- Adding New Features: To examine the effects of BMI and smoking on wages, new features were created from these two features. This is logical because these two features can significantly affect health insurance costs.

- Adding the mean and standard deviation: Adding the mean and standard deviation for BMI and smoking status is essential for measuring the distribution and variability of these characteristics in the dataset. This makes it easier to understand how common these characteristics are and how much they can vary.

Feature engineering is an essential step toward making the analysis and modeling of the dataset more robust [22]. In particular, adding new features and a better understanding existing features can improve the model's ability to predict health insurance costs. In addition, coding categorical variables can help the model better use the tables.

3.1.4. Data preprocessing

Data preprocessing, one of the critical steps in this study, forms the basis of our model for estimating health insurance costs. At this stage, we analyzed, cleaned, and prepared our dataset. First, we code our categorical variables appropriately and put them into a format machine-learning algorithm. In terms of missing data processing, there were no missing data, which increased the reliability of the dataset. New features are added using feature engineering methods so that our model can be fed with more information. Finally, we split our dataset appropriately for model training and evaluation. Data preprocessing is a fundamental step in more accurately predicting health insurance costs and is essential in helping us increase the model's power.

The steps listed below describe the key stages of the data preprocessing process and indicate that you are preparing to migrate to the model.

The following are descriptions of these steps:

- Encoding Categorical Variables: Convert categorical variables such as "sex," "smoker," and "region" into numeric values via the encoder method and introduce these variables into a format that machine learning algorithms can understand.
- Missing value: No missing data in the dataset means incomplete data processing methods are unnecessary. This means the dataset is clean and complete and helps the machine learning model produce reliable results.
- Creating datasets: Creating a dataset with new features added using feature engineering methods helps the model obtain more information. By creating two datasets, one has two separate datasets containing the original and new features, allowing us to examine how the model performs in different scenarios.
- Splitting the dataset: Separating the training and test datasets was essential for training and evaluating the model. Thus, we can determine the generalization ability of the model.

The data are divided as follows:

  - Train set (80%): 1.070
  - Test set (20%): 268

These preprocessing steps form the basic infrastructure of the health insurance cost estimation model. Using these datasets, one can train the machine learning algorithms and proceed to the evaluation phase.

3.2. Methods

3.2.1. Ensemble methods

This study uses various ensemble methods to estimate health insurance costs. These methods allow for more robust forecasting by combining multiple underlying models. The following is a brief introduction to these methods:

- Bagging (Bootstrap Aggregating): Bagging is an ensemble method in which several base models are trained independently on the same dataset, and their predictions are subsequently aggregated. Each base model was trained on a random dataset sample; thus, different models were generated with different samples. This method can increase the stability of the model while reducing the variance.
- Stacking: Stacking is an ensemble method in which the predictions of multiple base models are combined using a higher-level model (meta-model). Base models can be built on different features or algorithms, and metamodels use the predictions of these base models to generate the final prediction. This method is effective when different models complement each other and achieve better forecasting performance.
- Boosting: Boosting is an ensemble method that combines weak forecasting models to construct a robust model. The base models were trained sequentially, and each model was introduced by focusing on the errors of the previous model. In this way, the number of mistakes is reduced, and more robust predictions are obtained. Popular boosting methods include AdaBoost, gradient boosting, and XGBoost.

This study performed experiments to predict health insurance costs using ensemble methods. The performance of each method was evaluated separately, and the results are presented comparatively.

3.2.2. Proposed MedCost-AdaBoost Ensemble Methods

This section details the proposed AdaBoost ensemble method for forecasting health insurance costs. AdaBoost is an efficient ensemble method that aims to construct a robust prediction model by combining weak learners. The proposed architecture utilizes vulnerable learners, which are decision trees. Decision trees are suitable options for capturing the complexity of a dataset and understanding the effects of critical features.

The proposed model architecture includes a customized structure, as shown in Figure 2, to predict health insurance costs. This model architecture consists of three primary layers: an input layer, a MedCost-AdaBoost ensemble layer, and an output layer.

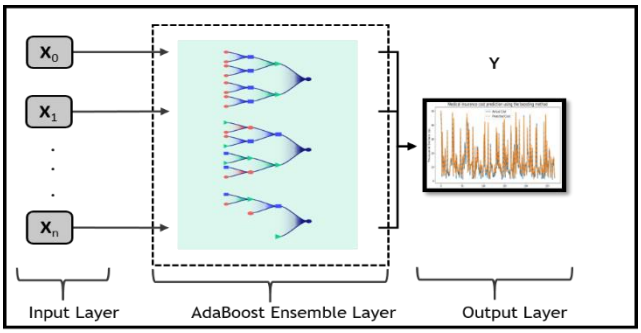

Figure 2. Proposed MedCost-AdaBoost Ensemble Architecture

- Input Layer: The first layer of the model, the input layer, contains the dataset's characteristics. These features include factors that affect health insurance costs, such as age, sex, body mass index (BMI), number of children, smoking status, and region (X0,…, Xn). The input layer uses these features and starts processing them.
- MedCost-AdaBoost Ensemble Layer: The layer that forms the core of the model contains the AdaBoost ensemble method. AdaBoost is an ensemble method that combines weak learners with a robust prediction model. This layer connects vulnerable learners like decision trees to obtain more robust predictions.
- OutPut Layer: The final layer of the model, the output layer, is where the predictions are generated and the results are obtained. This layer makes predictions from the AdaBoost Ensemble Layer and predicts medical insurance costs (Y).

The model architecture in Figure 2 was customized and optimized to predict health insurance costs by applying the AdaBoost ensemble method using the features in the dataset. Setting specific parameters during the training of the AdaBoost model is important. These parameters affected the performance and stability of the model. In particular, parameters such as the depth of the trees, learning rate, and number of estimators (n_estimators) must be chosen carefully [23]. These parameters help the model avoid overfitting and produce accurate predictions. The grid search method was preferred for hyperparameter tuning. The optimal parameters after tuning are listed below:

- Max_depth=5
- Max_features='auto'
- Max_leaf_nodes=70
- Min_samples_leaf=10
- Min_weight_fraction_leaf=0.1
- Splitter='best'
- N_estimators=10
- Learning_rate= 0.01

The MedCost-AdaBoost method combines weak learners with weights to develop a robust forecasting model. This model positively impacts the ability of the model to predict health insurance costs more accurately. The results and evaluation section presents the results and performance of the proposed method in more detail.

4. RESULTS AND EVALUATION

4.1. Evaluation Metrics

In this study, which used the AdaBoost ensemble method to predict health insurance costs, different model parameters and feature engineering approaches were examined. We are now exploring the results of this study to assess the effectiveness of this method and its impact on the ability to predict health insurance costs. This study highlights the importance of using various evaluation metrics to measure a model's performance and understand how it can be integrated into real-world applications [24].

1. Mean absolute error (MAE): MAE is a metric that measures how far the predictions are from the actual values. A lower MAE indicates a better prediction model.

$$MAE = \frac{\Sigma\,(p - a)^2}{n} \quad (1)$$

In Formula (1):

- p: predicted value
- a: actual value
- n: represents the number of observations.

2. Root mean squared error (RMSE): The RMSE is a metric that measures how much error a prediction makes relative to the actual values. A lower RMSE indicates a better prediction model.

$$RMSE = \sqrt{\left(\frac{\Sigma\,(p - a)^2}{n}\right)} \quad (2)$$

The values of p, a, and ne in formula (2) were used as defined in formula (1).

3. R^2 (R-squared): R^2 is a metric that measures how much of the variance in the dataset is explained by the prediction model. It has a value between 0 and 1, with higher R^2 values indicating better model performance:

The formula R^2:

$$R^2 = 1 - \frac{SSR}{SST} \quad (3)$$

- The sum of squares of residuals (SSR) represents the sum of squares of the actual value and forecast differences.
- Total Sum of Squares (SST): represents the sum of squares of the deviations from the mean of the actual values.

These are the standard metrics used to evaluate model performance and measure the accuracy of predictions. The performance of the proposed model was assessed in detail using these metrics.

4.2. Evaluation of model performance

The performance and model results of different ensemble methods for estimating health insurance costs were evaluated. The work is based on the original dataset containing all features and the MedCost dataset, which results from feature engineering. Bagging, stacking, and the proposed MedCost-AdaBoost model were applied to both datasets. For each ensemble method, mean absolute error (MAE), root mean squared error (RMSE), and R2 were used to evaluate the model's performance. These results will help us better understand which ensemble method is more effective for both datasets and how dataset characteristics affect the prediction performance. The results of the ensemble methods for both datasets are presented in Tables 2 and 3, respectively.

TABLE II
ALL-FEATURES DATASET

Ensemble Method(s)	MAE	RMSE	R^2
Bagging	2.4850	4.6641	84.46
Stacking	2.7359	4.5078	85.48
MedCost-AdaBoost	2.3052	4.2500	87.10

The evaluation metrics in Table 2 show the different results for each ensemble method. In this study, we evaluated these results. MedCost-AdaBoost has lower MAE and RMSE values than the other two methods. This indicated that the model predictions were closer to the actual values. In addition, the R2 value of MedCost-AdaBoost was greater than that of the others. This suggests that the model can explain a more significant proportion of the variance of the dataset, signaling a better prediction performance. Stacking performed moderately in terms of MAE and RMSE values, with a slightly higher R2 value. This offered an advantage over the other two models. Bagging was higher than the others in terms of MAE and RMSE values and had a lower R2 value, indicating that the predictions were less accurate. In conclusion, the MedCost-AdaBoost model shows better forecasting performance than the results obtained using all features. This model appears to be the preferable ensemble method for estimating health insurance costs.

TABLE III
MOST-ADABOOST FEATURE LEARNING DATASET

Ensemble Method(s)	MAE	RMSE	R^2
Bagging	3.9538	4.9418	89.75
Stacking	2.4656	4.7428	90.56
MedCost-AdaBoost	0.6749	0.9144	95.75

The evaluation metrics in Table 3 show the different results for each ensemble method. These results were evaluated as follows:

- MedCost-AdaBoost: This model performed exceptionally well on the MedCost dataset using all features. Both the MAE and RMSE values were shallow, and the R^2 value was high. This indicates that the model predictions are very close to the actual values and can explain a large part of the variance of the dataset.
- Stacking: The Stacking model showed moderate performance regarding MAE and RMSE, but the R^2 value was relatively high. This model offers more precise predictions than other methods.
- Bagging: The Bagging model is higher than MedCost-AdaBoost and Stacking regarding MAE and RMSE values and has a lower R^2 value. This indicates that the model's predictions are less precise and can explain less variance than those of the other two models.

In conclusion, the MedCost-AdaBoost model performed best on the MedCost dataset. In particular, the MAE and RMSE values were shallow, and the R2 value was extremely high. This model appears to be an effective ensemble method for estimating health insurance costs. In Tables 2 and 3, the MedCost-AdaBoost model after feature engineering on the MedCost dataset performs best when all features are used and after feature engineering. This model can better predict health insurance costs by improving the accuracy and precision of the predictions. Although the Bagging and Stacking models performed well, MedCost-AdaBoost was more effective.

5. DISCUSSION AND COMPARISON

This study aims to evaluate different ensemble methods for estimating health insurance costs using two other datasets. First, the results were obtained using Bagging, Stacking, and the proposed MedCost-AdaBoost model on the original dataset containing all the features. The same methods were applied to a new dataset created after feature engineering. We can evaluate which ensemble method and dataset performs better by comparing the results obtained for the two datasets.

Although the first dataset (All Features) performed well for the Bagging and Stacking models, the prediction performance of these models was lower than that of MedCost-AdaBoost. The MedCost-AdaBoost model obtained this dataset's lowest MAE and RMSE values and stood out as the model with the highest R2 value. These results show that the proposed ensemble method, MedCost-AdaBoost, can predict health insurance costs better when all features are used.

The new dataset created after feature engineering (Without-Feature Engineering) performed better, with higher R2 values for the Bagging and Stacking models. However, the MedCost-AdaBoost model still achieved the lowest MAE and RMSE values for this dataset, and the R2 value was high. This shows that the MedCost-AdaBoost model exhibited the best prediction performance on the without-feature engineering dataset.

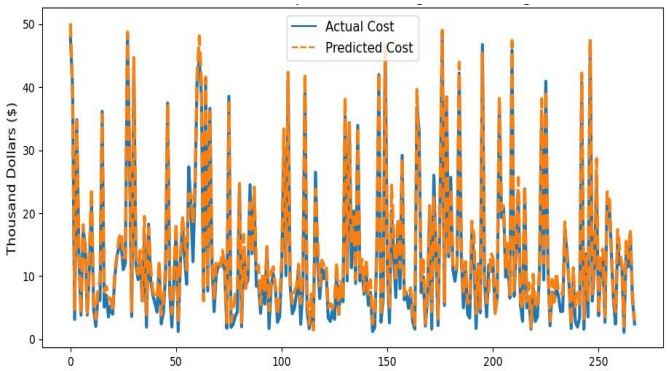

Figure 3. Forecast graph of Proposed MedCost-AdaBoost

Figure 3 shows the relationship between estimated and actual health insurance costs. The x-axis of the graph represents the "Actual Cost" values, and the y-axis represents the "Estimated Cost" values. An overview of the graph shows that, in general, the costs are in reasonably good agreement with the estimates. Instead of forming a line between the actual and estimated expenses, the dots appear densely packed. This shows that the forecasts are generally close to the actual costs and that the model performs well. However, there was a more significant deviation between the actual and estimated costs at some points. In particular, some estimates are lower than the actual costs of high-attraction insurance policies. These points may need to be considered, and the model may need to be improved to predict such cases better. Overall, this graph shows that the model successfully predicts health insurance costs and generally provides good agreement with actual expenses. However, some aspects can be further improved.

In conclusion, this study emphasizes that the MedCost-AdaBoost model is an effective ensemble method for predicting health insurance costs and performs well on a without-feature engineering dataset. This model can be a valuable tool for the cost estimation of health insurance companies and healthcare providers. Although the Bagging and Stacking models also performed well, the MedCost-AdaBoost model was more efficient and could make more precise predictions.

This study has some limitations. First, the dataset size is limited, which may limit its generalizability and require validation on a larger dataset. Moreover, the choice of characteristics used in the study may have excluded some other essential factors. For example, not including genetic factors, chronic diseases, or socioeconomic status may affect the model's predictive power. Therefore, future studies with more extensive and diverse datasets and a more comprehensive selection of characteristics may help make these models more robust and generally valid.

6. CONCLUSION

This study examined the usability and performance of machine learning methods in predicting health insurance costs. The dataset included factors such as age, sex, body mass index (BMI), number of children, smoking habit, and region and aimed to understand and predict the impact of these factors on health insurance costs. The statistical analysis of the dataset was performed at the beginning of the study. This analysis allowed us to examine the basic statistical properties and the distribution of the dataset. Subsequently, new features are added to the dataset through a feature engineering step. In particular, we have added the mean and standard deviation values of BMI and smoking status. Categorical variables were coded appropriately.

The main focus of this study was to evaluate the performance of ensemble learning methods. Health insurance costs were estimated on both datasets using Bagging, Stacking, and the proposed MedCost-AdaBoost model. The results show that the MedCost-AdaBoost model outperformed the other two methods. The MedCost-AdaBoost model achieved lower MAE and RMSE values and higher R2 values on the dataset with all features and the dataset generated after feature engineering. These results emphasize that ensemble learning methods are practical for predicting health insurance costs. This also shows that feature engineering can improve the model's performance and is vital in obtaining accurate predictions. However, this study has some limitations. The limited size of the dataset and exploration of potential possibilities for adding more features provide an opportunity for future work. Moreover, considering other factors affecting insurance costs and using a larger dataset may improve forecasting performance.

In conclusion, this study shows that ensemble learning methods are practical tools for predicting health insurance costs, and feature engineering plays a vital role in obtaining accurate predictions. This study can be a valuable reference for researchers and professionals seeking to improve cost estimations in the health insurance industry. Future studies can advance research by considering more factors and working on larger datasets.

ACKNOWLEDGEMENT

I want to thank my dear family, wife, and kids for their patience with my academic studies.

REFERENCES

[1] Saraswat, B. K., Singhal, A., Agarwal, S., & Singh, A. (2023, May). Insurance Claim Analysis Using Traditional Machine Learning Algorithms. In 2023 International Conference on Disruptive Technologies (ICDT) (pp. 623-628). IEEE.

[2] Vijayalakshmi, V., Selvakumar, A., & Panimalar, K. (2023, January). Implementation of Medical Insurance Price Prediction System using Regression Algorithms. In 2023, the 5th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 1529-1534). IEEE.

[3] Bora, A., Sah, R., Singh, A., Sharma, D., & Ranjan, R. K. (2022, October). Interpretation of machine learning models using xai-a study on health insurance dataset. In 2022, the 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) (pp. 1-6). IEEE.

[4] Jyothsna, C., Srinivas, K., Bhargavi, B., Sravanth, A. E., Kumar, A. T., & Kumar, J. S. (2022, May). Health Insurance Premium Prediction using XGboost Regressor. In 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 1645-1652). IEEE.

[5] Kaushik, K., Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Machine learning-based regression framework to predict health insurance premiums. International Journal of Environmental Research and Public Health, 19(13), 7898.

[6] Chittilappilly, R. M., Suresh, S., & Shanmugam, S. (2023, May). A Comparative Analysis of Optimizing Medical Insurance Prediction Using Genetic Algorithm and Other Machine Learning Algorithms. In 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) (pp. 1-6). IEEE.

[7] Albalawi, S., Alshahrani, L., Albalawi, N., & Alharbi, R. (2023). Prediction of healthcare insurance costs. Computers and Informatics, 3(1), 9-18.

[8] Praveen, M., Manikanta, G. S., Gayathri, G., & Mehrotra, S. (2023, February). Comparative Analysis of Machine Learning Algorithms for Medical Insurance Cost Prediction. In International Conference On Innovative Computing and Communication (pp. 885-892). Singapore: Springer Nature Singapore.

[9] Sahare, A. N. (2023). Forecasting Medical Insurance Claim Cost with Data Mining Techniques (Doctoral dissertation, Dublin, National College of Ireland).

[10] Hassan, C. A., Iqbal, J., Hussain, S., AlSalman, H., Mosleh, M. A., & Sajid Ullah, S. (2021). A computational intelligence approach for predicting medical insurance cost. Mathematical Problems in Engineering, 2021, 1-13.

[11] Demirci, F., Emec, M., Gursoy Doruk, O., Ormen, M., Akan, P., & Hilal Ozcanhan, M. (2023). Prediction of LDL in hypertriglyceridemic subjects using an innovative ensemble machine learning technique. *Turkish Journal of Biochemistry*, (0).

[12] Kaya, Y., Yiner, Z., Kaya, M., & Kuncan, F. (2022). A new approach to COVID-19 detection from X-ray images using angle transformation with GoogleNet and LSTM. *Measurement Science and Technology*, *33*(12), 124011.

[13] Hemdan, E. E. D., El-Shafai, W., & Sayed, A. (2023). CR19: A framework for preliminary detection of COVID-19 in cough audio signals using machine learning algorithms for automated medical diagnosis applications. *Journal of Ambient Intelligence and Humanized Computing*, *14*(9), 11715-11727.

[14] AKDAĞ, S., Kuncan, F., & Kaya, Y. (2022). A new approach for classification of congestive heart failure and arrhythmia by downsampling local binary patterns with LSTM. *Turkish Journal of Electrical Engineering and Computer Sciences*, *30*(6), 2145-2164.

[15] Kaya, Y., & Kuncan, F. (2022). A hybrid model for classification of medical data set based on factor analysis and extreme learning machine: FA+ ELM. *Biomedical Signal Processing and Control*, *78*, 104023.

[16] Wu, X., Tang, H., Zhu, Z., Liu, L., Chen, G., & Yang, M. S. (2023). Nonlinear strict distance and similarity measures for intuitionistic fuzzy sets with applications to pattern classification and medical diagnosis. *Scientific reports*, *13*(1), 13918.

[17] Ayvaz, E., Kaplan, K., Kuncan, F., Ayvaz, E., & Türkoğlu, H. (2022). Reducing Operation Costs of Thyroid Nodules Using Machine Learning Algorithms with Thyroid Nodules Scoring Systems. *Applied Sciences*, *12*(22), 11559.

[18] Yurtsever, M., & Emeç, M. (2023). Potable Water Quality Prediction Using Artificial Intelligence and Machine Learning Algorithms for Better Sustainability. *Ege Academic Review*, *23*(2), 265-278.

[19] Orenc, S., Acar, E., & Özerdem, M. S. (2022, October). The Electricity Price Prediction of Victoria City Based on Various Regression Algorithms. In *2022 Global Energy Conference (GEC)* (pp. 164-167). IEEE.

[20] Gönenç, A., Acar, E., Demir, İ., & Yılmaz, M. (2022, October). Artificial Intelligence Based Regression Models for Prediction of Smart Grid Stability. In *2022 Global Energy Conference (GEC)* (pp. 374-378). IEEE.

[21] Ruzgar, S., & Acar, E. (2022). The statistical neural network-based regression approach for prediction of the optical band gap of CuO. *Indian Journal of Physics*, *96*(12), 3547-3557.

[22] Emeç, M., & Özcanhan, M. H. (2023). Veri Ön İşleme ve Öznitelik Mühendisliğinin Yapay Zekâ Yöntemlerine Uygulanması. MÜHENDİSLİKTE ÖNCÜ VE ÇAĞDAŞ ÇALIŞMALAR, 33-54.

[23] Emeç, M., & Özcanhan, M. H. (2023). Makine Öğrenmesi Algoritmalarında Hiper Parametre Belirleme. MÜHENDİSLİKTE ÖNCÜ VE ÇAĞDAŞ ÇALIŞMALAR, 71-98.

[24] Alzoubi, H. M., Sahawneh, N., AlHamad, A. Q., Malik, U., Majid, A., & Atta, A. (2022, October). Analysis Of Cost Prediction In Medical Insurance Using Modern Regression Models. In 2022 International Conference on Cyber Resilience (ICCR) (pp. 1-10). IEEE.

BIOGRAPHIES

Murat Emeç completed his computer engineering undergraduate degree at Ege University in 2015. In 2017, he received his master's in management information systems from Dokuz Eylül University. In 2022, he received his PhD in computer engineering from Dokuz Eylül University. Between 2010 and 2020, he worked as a senior software specialist in the Dokuz Eylül information technology (IT) department. Between 2021 and 2022, he worked at Marmara University as a lecturer and then as a lecturer doctor. He has been working at Istanbul University since September 2022. His research interests are the Internet of Things, Data Science, Artificial Intelligence, and Machine Learning.