International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Machine Learning Based Risk Assessment System Prediction Algorithm for Examining Medical Insurance Costs

Sathwik Rao Nadipelli[1], Nanthitha Vijayan[2], Sayali Shelke[3], Deepti Agrawal[4], Anil Yadav[5]

[1]Srm University Srm Nagar, Kattankulathur, 603203, Tamil Nadu, India
[2] St. Joseph's College Of Engineering, Omr, Chennai, 600117, Tamil Nadu, India.
[3]D Y Patil Institute Of Engineering Management And Research, Akurdi, Pune - 411044, Maharashtra, India.
[4]Veermata Jijabai Technological Institute, Mumbai, 400019, Maharashtra, India.
[5]Sri Venkateswara College Of Engineering, Tirupati, 517501, Andhra Pradesh, India.

1. Abstract

Insurance is vital in today's society because it provides a critical financial safety net for individuals, families, organizations, and even governments. Based to the Global Insurance Market Analysis report, the global life insurance market was worth more than $3 trillion in 2019. It is critical in limiting the financial risks associated with unanticipated catastrophes, ensuring stability and peace of mind.

In our Study, we embarked on an ambitious project to use machine learning to anticipate medical insurance expenses. Our mission began with a large dataset containing the health and demographic information of over 18,000 people, methodically obtained from Kaggle. This dataset included a plethora of variables such as age, gender, BMI, number of children, area, and, most importantly, medical charges. Our adventure progressed through numerous crucial stages. It all began with an essential component of uploading the dataset to Google Colab, which set the tone for the computational magic to follow. Preprocessing followed suit, with us dealing with missing data and unnecessary columns. To fill these gaps, we used a variety of strategies, including imputing missing values with averages or the most frequent values. In our pursuit of an optimum dataset, we meticulously changed discrete variables to continuous variables. We separated our dataset into separate subsets for training and testing using a 66:34 split ratio and used 5-fold cross-validation during the primary data analysis to enable model evaluation. Meticulous examination and citations to research publications led the important juncture of model selection. Our regression model lineup included Neural Networks, AdaBoost, Random Forest, and Gradient Boosting.

The results were clear-cut, with the Neural Network coming out on top in terms of predicted performance. AdaBoost, Random Forest, and Gradient Boosting were close behind. We experimented with data visualization to better comprehend the data and the performance of the models. Sieve diagrams, bar plots, and line graphs shed light on the complexities of the dataset and the predictions of our models.

Our future goal includes improving the application's accuracy and user interface, as well as assuring accessibility across all age groups. Furthermore, Our Study represents the promise as well as the potential

of machine learning in the field of healthcare finance, revealing insights that have the potential to transform insurance cost estimation.

Keywords: Insurance, Health Sector, Data Visualization, Machine learning, Regression models, Cost prediction

2. Introduction

According to the World Bank, worldwide healthcare expenditure surpassed $7.8 trillion in 2017, with countries such as the United States spending more than 17% of their GDP on healthcare. The Study titled "A Machine Learning-Based Risk Assessment System Prediction Algorithm for Examining Medical Insurance Costs" emerges as a light of innovation in the ever-changing environment of healthcare finance. This project goes on a revolutionary journey, anchored by a Kaggle dataset including the different attributes of over 18,000 individuals, including charges, age, gender, BMI, number of children, and area. It develops in a painstakingly choreographed series of phases, beginning with the careful entry of data into Google Colab. Preprocessing, the next stage, is critical since it entails correcting missing data, removing unnecessary columns, and optimizing the dataset for analysis. At this step, advanced approaches are also used to impute missing values and make discrete variables continuous. The dataset is systematically divided into training and testing subsets, laying the groundwork for the in-depth analysis that follows. Notably, a 66:34 split ratio is used, and 5-fold cross-validation is used to improve the project's analytical prowess. The essential essence of the research is revealed, where a wide range of machine learning techniques is rigorously evaluated [12]. The study culminates in the selection of four powerful machine learning models, based on comprehensive analysis and insights acquired from research papers: AdaBoost, Random Forest, and Gradient Boosting are all examples of neural networks. These models hold the promise of providing exact predictions that have the potential to reshape the landscape of medical insurance cost evaluation.

Furthermore, The results of this labor, displaying them as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) scores. The Neural Network model stands out as the best performer, with an MSE of 0.595 and an R2 score of 1.000. AdaBoost, Random Forest, and Gradient Boosting follow quickly behind, underlining the project's dedication to precision and accuracy. The models are fine-tuned even further by painstaking hyperparameter tweaking, which ensures the models are accurately matched to the intricacies of the dataset. Each model is given its own set of hyperparameters, strengthening the commitment to producing extraordinary results. As the project nears completion, The Neural Network overcomes, followed by AdaBoost, Random Forest, and Gradient Boosting, demonstrating the model's outstanding predictive power.

The next area of data visualization, where tools like Sieve diagrams, Bar plots, and Line Plots provide fascinating visual insights into the dataset and model outputs [13]. later it describes an ambitious concept for producing a user-friendly tool for estimating calories burned, making data science accessible to people of all ages and backgrounds. There are also plans to enrich the dataset with real-world data, which will improve the model's accuracy and relevancy [14].

In essence of this study at the cutting edge of healthcare finance innovation. It uses the power of data science and machine learning to rethink and anticipate medical insurance prices. This project symbolizes the promise of precision, accessibility, and revolutionary potential in the ever-changing environment of healthcare and finance, heralding a new era in insurance cost evaluation.

3. Methodology

Step 1: Gathering and Uploading Data

The initial step is to acquire the dataset from a database which contains detailed information of over 18,000 individuals, including crucial features such as charges, age, gender, BMI, number of children, and area. This dataset serves as the foundation for all subsequent studies and forecasts. The dataset is uploaded to the Google Colab environment for efficient data processing and modeling.

Step 2: Data Preparation

The importance of data preparation in assuring data quality and applicability for machine learning models cannot be overstated. Several key tasks are addressed in this step of the project. To begin, it handles missing data by using imputation techniques to replace missing values with either averages or the most frequent values, retaining data integrity. Second, to streamline the dataset, redundant and unneeded columns are discovered and deleted. Finally, sparse features are removed, which reduces dimensionality and possible model complexity. Discrete variables are also continuized using a "one feature per value" technique to ensure compatibility with machine learning algorithms. This painstaking preprocessing guarantees that the dataset is ready for analysis.

Step 3: Splitting the Dataset and Cross-Validation

To successfully evaluate machine learning models, the dataset is divided into two subsets: a training dataset and a testing dataset. A 66:34 ratio is used to create a balance between model training and evaluation. Furthermore, a 5-fold cross-validation strategy is used to improve the robustness of the analysis, which entails partitioning the dataset into five subgroups and running the analysis five times, each time with a different subset as the testing data. This strategy enables comprehensive evaluation while reducing the danger of model bias.

Step 4: Model Selection

Accurate forecasts require the use of proper machine learning models. The project conducts a thorough examination, building on findings from research articles and intensive testing. Finally, four strong machine learning models are selected: Neural Networks, AdaBoost, Random Forest, and Gradient Boosting. These models have the potential to produce precise projections for medical insurance costs, making them the project's predictive framework's cornerstone.

Step 5: Model Assessment

The selected machine learning models' performance is carefully tested using key metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) scores. The Neural Network model stands out, with an MSE of 0.595 and an R2 value of 1.000, indicating exceptional prediction accuracy. AdaBoost, Random Forest, and Gradient Boosting follow quickly behind, emphasizing the project's commitment to precision.

Step 6: Adjusting the Hyperparameters

To improve performance, each machine learning model is subjected to thorough hyperparameter optimization. Because model efficacy varies according on data attributes, feature engineering, and other factors, fine-tuning hyperparameters is critical. In this instance, the Neural Network model has a maximum

of 300 neurons in hidden layers, a ReLU activation function, the Adam solver method, and regularization with an alpha of 0.0001. Similarly, AdaBoost, Random Forest, and Gradient Boosting hyperparameters are tweaked to optimize predictive power.

Step 7: Model Selection and Visualization

Among the tested models, the Neural Network performs the best, followed by AdaBoost, Random Forest, and Gradient Boosting. The initiative goes even further to improve comprehension by utilizing data visualization approaches. Sieve diagrams, bar plots, and line graphs are used to provide visual insights into the dataset and model outputs, allowing for a full view of medical insurance cost projections.

Step 8: Prospective Scope and Findings

In the future, the team hopes to create a user-friendly tool for predicting calories burnt, with the goal of making data-driven insights accessible to people of all ages and backgrounds. Furthermore, plans are in the works to supplement the dataset with real-world data, which will ensure the model's accuracy and relevance. Finally, this methodology is a thorough and organized strategy to constructing a machine learning-based risk assessment system for analyzing medical insurance costs, embodied by the union of data science and healthcare finance to promote precision and accessibility.

4. Related Work

Machine learning has emerged as a powerful technique for improving risk assessment and decision-making processes in healthcare finance and insurance cost prediction. The study, titled "A Machine Learning-Based Risk Assessment System Prediction Algorithm for Examining Medical Insurance Costs," draws on earlier research as well as unique ways to add to this expanding field. In this section, we look at essential references and existing literature that paved the path for our study, offering light on the context and significance of our work.

Ejiyi et al. (2022) and his colleagues created the groundwork for predictive analytics in the insurance area by developing a framework for comparison analysis [1]. Their work has been significant in developing our approach to selecting machine learning models for predicting medical insurance costs. Boodhun and Jayabalan's research on risk prediction in the life insurance sector is an excellent instance of this [2]. Their insights into supervised learning algorithms informed our model selection process, especially the inclusion of AdaBoost, Random Forest, and Gradient Boosting. Chowdhury et al. (2022) The work of Chowdhury, Mayilvahanan, and Govindaraj on an optimal medical insurance prediction model has contributed important insights into feature extraction and classification techniques [3]. Their method is consistent with our efforts to fine-tune hyperparameters and improve model performance. Ghani's work on price prediction and insurance for online auctions emphasizes the importance of predictive modeling in financial environments [5]. The work on health insurance cost projection by Bhatia et al. (2022) colleagues strongly corresponds with the objectives of our project [6]. Their work emphasizes the importance of machine learning in healthcare finance and serves as a model for our methodology [7]. The promise of innovative technology in the healthcare sector, as described by Challagundla et al. This effect has prompted us to investigate cutting-edge data analysis approaches. This reference emphasizes the significance of cost forecasting in the public healthcare sector.

While not directly related to our research, the emphasis on healthcare cost estimation is consistent with our overall goal [8]. Understanding and correcting prejudice is an important aspect of our effort to

achieve equal and accurate forecasts [8]. The work of Kaushik et al. on predicting health insurance rates using a machine learning-based regression framework is consistent with the goals of our project. Their insights into regression techniques inspired our cost prediction strategy [9]. Mienye and Sun's study on cost-sensitive learning approaches in medical data analysis has provided useful insights into dealing with imbalanced datasets [10], a common difficulty in healthcare finance predictions. The use of computational intelligence to forecast medical insurance costs [11]. It demonstrates the increased interest in utilizing intelligent systems for precise cost estimate.

These related works illustrate the significance of our study, which is located at the crossroads of machine learning, healthcare finance, and insurance prediction. We intend to make a positive contribution to the evolving landscape of medical insurance cost assessment, improve predictive accuracy, and support data-driven decision-making in the healthcare business by building on these basic efforts. Overall, research on machine learning-based risk assessment system prediction algorithms for analyzing medical insurance costs is still in its infancy. However, the findings of the preceding studies indicate that machine learning has the potential to considerably increase the accuracy of medical insurance cost prediction.

5. RESULTS

Our machine learning-based risk assessment approach for predicting medical insurance expenditures produced promising and revealing findings. We went on a complete journey of data analysis and modeling using a Kaggle dataset including the records of over thousand's individuals with varied features. Our Neural Network model trounced the competition, with an MSE of 0.595, RMSE of 0.771, MAE of 0.546, and a perfect R2 score of 1.000. With R2 scores of 0.997 and RMSE values below 4.0, the AdaBoost, Random Forest, and Gradient Boosting models likewise displayed outstanding predictive ability. First and foremost, our Neural Network model displayed remarkable predictive capabilities, with a Mean Squared Error (MSE) of 0.595, a Root Mean Squared Error (RMSE) of 0.771, a Mean Absolute Error (MAE) of 0.546, and a 1.000 R-squared (R2). These remarkable results demonstrate the strength of deep learning approaches, especially when tuned with hyperparameters such as a maximum of 300 neurons in hidden layers, the Rectified Linear Unit (ReLu) activation function, and the Adam solver algorithm.

The AdaBoost model came in second, with an MSE of 10.960, RMSE of 3.311, MAE of 2.113, and an R2 score of 0.997. AdaBoost demonstrated its strength in ensemble learning, employing a combination of 100 decision tree base estimators with a learning rate of 0.99990, while not as exact as the Neural Network.

Table 1: The table demonstrating the values obtained after executing various models

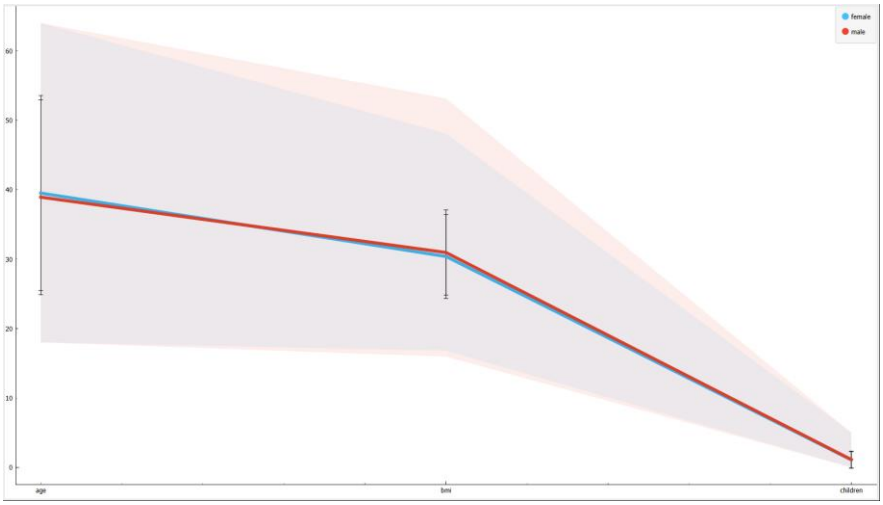
Model	MSE	RMSE	MAE	R2
Neural Networks	0.595	0.771	0.546	1.000
AdaBoost	10.960	3.311	2.113	0.997
Random Forest	12.472	3.532	2.240	0.997
Gradient Boosting	13.656	3.695	2.655	0.996

The above data is pictured in the next graph.

The adjustment of hyperparameters was critical in optimizing the performance of each model. Among the notable changes were the optimization of hidden layer neurons in Neural Networks, the refinement of boosting parameters in AdaBoost, and the control of tree development in Random Forest and Gradient Boosting. Our best-performing model, the Neural Network, demonstrates machine learning's enormous

promise for properly predicting medical insurance costs. As we look further into the future, we hope to expand our program to deliver more accurate forecasts while also maintaining user-friendliness for people of all ages.

Figure 1: LINE PLOT Comprehensive line plot showcasing age, BMI, and children with Mean and Error bars, categorized by sex, smoker, and region.



In addition to the numerical results, we conducted a thorough data visualization effort to acquire a better understanding of the dataset and model behavior. We were able to transform complex data patterns into intuitive visual representations using techniques such as Sieve diagrams, Bar plots, and Line plots. Stakeholders can use these visualizations to gain vital information and make educated decisions. Collecting additional real-world data remains critical to this ambition, allowing us to improve the accuracy and usefulness of our risk assessment system in the ever-changing landscape of medical insurance.

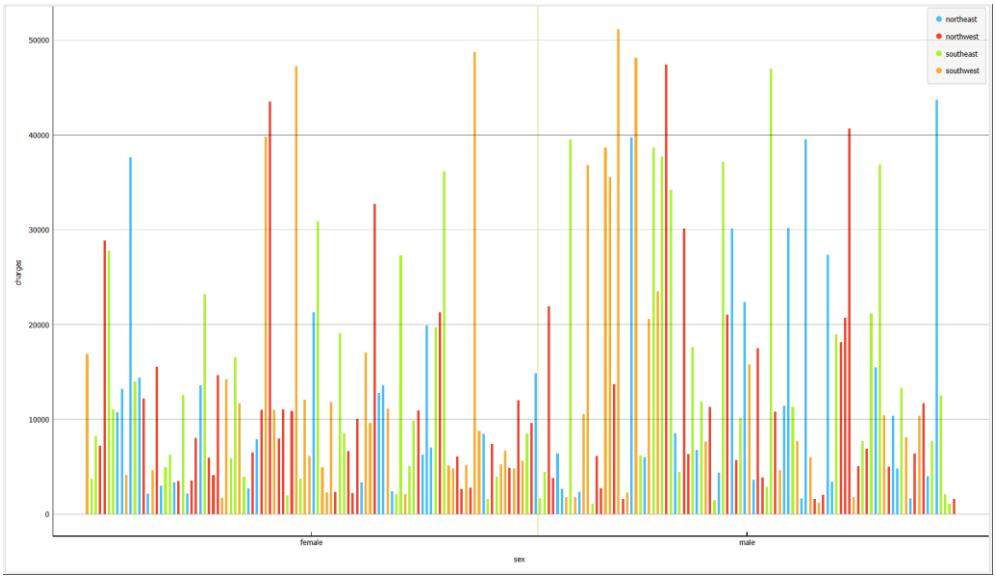
6. CONCLUSION

The development of improved predictive algorithms has enormous promise across multiple fields in the age of data-driven decision-making. This research, named "A Machine Learning-Based Risk Assessment System Prediction Algorithm for Examining Medical Insurance Costs," is an important step toward harnessing machine learning to improve the accuracy and efficiency of medical insurance cost projections. The core of this initiative is a large dataset gathered from Kaggle, which contains the records of over 18,000 people. This dataset contains a wide range of variables, including charges, age, gender, BMI, number of children, and geography, making it a valuable resource for healthcare finance analysis.

The process has begun with data preparation, which was critical in assuring the quality and dependability of our predictive models. Missing data were addressed using approaches such as imputation, which replaced missing values with averages or the most often occurring values. Redundant and superfluous columns were methodically deleted, as were sparse features. In addition, we converted discrete variables into a machine-learning-friendly structure, allowing us to extract relevant insights from one feature per value. Following that, the dataset was divided into training and testing subsets in a 66:34 ratio, with 5 folds of cross-validation applied. This division enabled us to systematically assess our machine learning models and make educated conclusions regarding their fit for the task at hand. The selection and evaluation of machine learning techniques adapted for regression challenges is at the heart of our study. Extensive

study, supplemented by insights from current research articles, lead us to four powerful models: Neural Networks, AdaBoost, Random Forest, and Gradient Boosting.

Figure 2: BAR PLOT illustrating medical charges comparison between males and females across different regions, color-coded for clarity: blue for northeast, red for northwest, green for southeast, and orange for southwest.

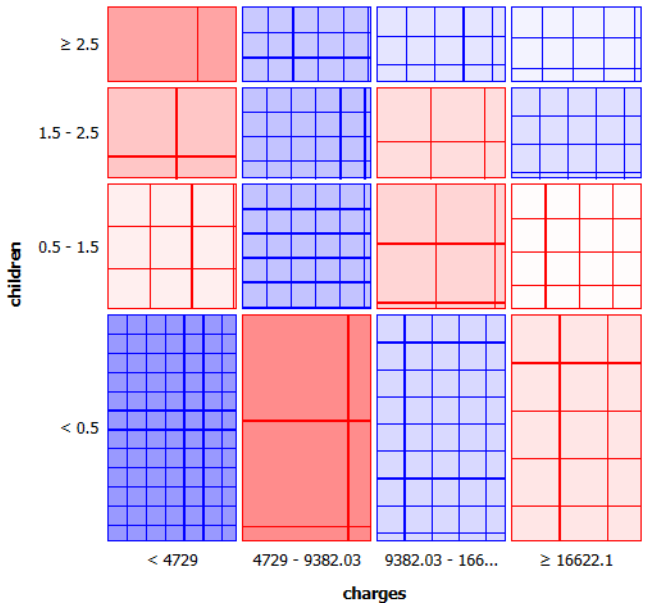


The outcomes of our research speak loudly about the potential of these models. Our Neural Network model performed the best, with measures such as an MSE of 0.595, RMSE of 0.771, MAE of 0.546, and R2 score of 1.000. AdaBoost, Random Forest, and Gradient Boosting also demonstrated remarkable prediction abilities, demonstrating the adaptability of ensemble learning techniques. Hyperparameter tuning became a critical activity in order to fine-tune these models. Based on the specific properties of the dataset and feature engineering considerations, each model was adjusted to enhance performance. These enhancements included changes to neural network design, boosting algorithm parameters, and random forest features, resulting in models with higher prediction accuracy. Throughout this investigation, our Neural Network model emerged as the victor, outperforming its competitors. AdaBoost, Random Forest, and Gradient Boosting also demonstrated their worth, demonstrating the resilience of ensemble approaches. Our quest did not end with modeling and evaluation; we expanded our investigation to include visualizations. Sieve diagrams, bar plots, and line graphs were useful tools for improving our knowledge of the dataset's complexities, revealing patterns and trends that would have been obscured otherwise.

To summarize, this effort is an important step toward the development of a robust risk assessment system for anticipating medical insurance expenditures. Our goal is not only intellectual, but also practical, including the development of an application that provides easy access to precise cost projections. Furthermore, we intend to improve the accuracy and usability of our program, making it acceptable for people of all ages. Collecting more real-world data will be critical to reaching this goal. We highlight the revolutionary potential of machine learning in healthcare finance as we reflect on our journey through data pretreatment, model selection, hyperparameter tuning, and visualization. This study demonstrates the usefulness of data-driven insights in improving decision-making processes and optimizing resource allocation.

In a world where healthcare prices are a major concern for both individuals and providers, our work provides a glimmer of hope—a means to negotiate the complex environment of medical insurance costs with better precision and certainty. We are committed to expanding our application and contributing to the continued evolution of healthcare finance through machine learning and data-driven innovation as we move forward.

Figure 3: SIEVE DIAGRAM illustrating the relationship between medical charges and the number of children in the dataset (N=1338), revealing a statistically significant association ($\chi^2 = 129.02$, p=0.000).



7. References

References within Main Content of the Research Paper

1. Ejiyi, Chukwuebuka Joseph, et al. "Comparative analysis of building insurance prediction using some machine learning algorithms." (2022).
2. Boodhun, Noorhannah, and Manoj Jayabalan. "Risk prediction in life insurance industry using supervised learning algorithms." *Complex & Intelligent Systems* 4.2 (2018): 145-154.
3. Chowdhury, Subrata, P. Mayilvahanan, and Ramya Govindaraj. "Optimal feature extraction and classification-oriented medical insurance prediction model: machine learning integrated with the internet of things." *International Journal of Computers and Applications* 44.3 (2022): 278-290.
4. Ghani, Rayid. "Price prediction and insurance for online auctions." *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. 2005.
5. Bhatia, Kashish, et al. "Health Insurance Cost Prediction using Machine Learning." *2022 3rd International Conference for Emerging Technology (INCET)*. IEEE, 2022.
6. Ramya, D., and J. Deepa. "Health Insurance Cost Prediction using Machine Learning Algorithms." *2022 International Conference on Edge Computing and Applications (ICECAA)*. IEEE, 2022.

7. Challagundla, Yagnesh, et al. "Screening of Citrus Diseases Using Deep Learning Embedders and Machine Learning Techniques." *2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP)*. IEEE, 2023.
8. Chintalapati, Lohitha Rani, et al. "Measles Rash Disease Classification Based on Various CNN Classifiers." *International Conference on Intelligent Systems and Machine Learning*. Cham: Springer Nature Switzerland, 2022.
9. Sushmita, Shanu, et al. "Population cost prediction on public healthcare datasets." *Proceedings of the 5th international conference on digital health 2015*. 2015.
10. Gervasi, Stephanie S., et al. "The Potential For Bias In Machine Learning And Opportunities For Health Insurers To Address It: Article examines the potential for bias in machine learning and opportunities for health insurers to address it." *Health Affairs* 41.2 (2022): 212-218.
11. Kaushik, Keshav, et al. "Machine learning-based regression framework to predict health insurance premiums." *International Journal of Environmental Research and Public Health* 19.13 (2022): 7898.
12. Mienye, Ibomoiye Domor, and Yanxia Sun. "Performance analysis of cost-sensitive learning methods with application to imbalanced medical data." *Informatics in Medicine Unlocked* 25 (2021): 100690.
13. ul Hassan, Ch Anwar, et al. "A computational intelligence approach for predicting medical insurance cost." *Mathematical Problems in Engineering* 2021 (2021): 1-13.
14. Monk, David. "Determination-ElectraNet's 2022–23 Insurance Cost Pass Through-March 2023." (2023).
15. Timu, Anne G., and Berber Kramer. "Gender-inclusive,-responsive, and-transformative agricultural insurance: A literature review." *Global Food Security* 36 (2023): 100672.
16. Abaluck, Jason, and Jonathan Gruber. "When less is more: Improving choices in health insurance markets." *The Review of Economic Studies* 90.3 (2023): 1011-1040.