A Study of Machine Learning Techniques for Predictive Analysis of Health Insurance

Nuzhat Noor Islam Prova
Seidenberg School of CSIS
Pace University
New York, USA
np23474n@pace.edu

Abstract— The use of machine learning algorithms for health insurance is a key area that aims to improve the accuracy and efficacy of these processes. This field focuses on anomaly detection and predictive modeling using a range of techniques, such as XGB regressors, random forest regressors, decision tree regressors, and k-neighbors regressors. Applications of decision tree regression are well recognized for their interpretability and simplicity of usage. An individual may successfully pinpoint the insurance system's shortcomings with a performance score of 0.85. In addition, the random forest regressor works wonderfully, this regressor is well-known for its capacity to manage complex datasets and reduce overfitting. The use of comparable machine learning algorithms in health insurance ensures accurate diagnosis of medical conditions, whilst error detection in insurance systems protects against errors and inconsistencies. Decision Tree (0.85), KNN (0.67), Random Forest (0.83), and XGB (0.70). XGB and KNN regressors are outperformed by Decision Tree and Random Forest models in terms of accuracy in predicting. In terms of overall R2 score, Decision Tree Regressor performs best, demonstrating its ability to accurately capture the variance of the target variable.

Keywords— Health insurance, Medical sector, Machine learning, Decision tree regressor, Random forest regressor

I. INTRODUCTION

Modern technologies must be used to increase accuracy and efficiency in the constantly evolving insurance and healthcare sectors. Health insurance error detection is one well-known industry spearheading this shift. Error detection, health problem prediction, and decision-making process acceleration are all made possible by machine learning approaches. Numerous approaches used in this discipline, such as decision tree regressors, hybrid models, random forest regressors, XGB regressors, and k-neighbor regressors, serve the general goals of anomaly identification and predictive modeling. The decision tree regressors exhibit remarkable interpretability and user-friendliness, yielding a robust performance score of 0.8464. Their versatility makes them useful tools for accurate forecasting and mistake detection, underscoring the potential benefits of a comprehensive approach to improving insurance systems. Medical insurance conversion is aided by the efficacy of random forest regressors, which are renowned for their capacity to manage complex datasets well and minimize overfitting. The relevance of the XGB Regressor (0.8254) and KNeighbors Regressor (0.7033) in this paradigm shift is highlighted by their performance rating, although with varying degrees of accuracy. The use of these machine learning algorithms has enormous promise for the insurance sector, in addition to protecting insurance systems against errors and inconsistencies. By ensuring precise identification of medical conditions, these models pave the way for a revolutionary impact on risk assessment, resource allocation, and decision-making in the insurance and healthcare industries. As these models advance, it is becoming more evident that they have the ability to fundamentally change the insurance and healthcare sectors. This study examines the transformative potential of state-of-the-art machine learning techniques, offering a glimpse of a future in which precision, efficiency, and innovation will unite to fundamentally transform these vital sectors. The intricate rationale for using machine learning to look into health insurance prediction is based on the need and potential in the insurance and healthcare industries. The primary motivator is the possibility that advances in machine learning may significantly improve the accuracy of health insurance. The goal is to provide more personalized, effective, and accurate treatment plans by using algorithms such as hybrid models and decision tree regressors. The project intends to decrease costs for policyholders and insurance firms by enhancing the capacity of insurance systems to identify mistakes in claims, preventing fraud, and streamlining operations. The paper recognizes that machine learning may help close the access gap to high-quality insurance and healthcare. Innovative applications, such as telemedicine, may break down geographical and socioeconomic barriers, enabling more equal access to essential services for everyone.

II. LITERATURE REVIEW

In [1] the dataset that was utilized contained 986 records and can be accessed by the general public through the KAGGLE repository. When compared to the SHAP analysis, which appeared to exhibit a higher level of abstraction, we discovered that the ICE plots displayed the interactions between each variable in greater depth. The authors of this study hope that the contributions of this study will assist regulators, insurers, and potential buyers of medical insurance in the process of making decisions regarding the selection of appropriate policies that are tailored to fit their particular requirements. In [2] The purpose of this study is to investigate the increasing usefulness of health insurance estimates in the wake of the COVID-19 epidemic. Because we are currently in a situation in which several efforts are being made to address this significant problem, our research makes use of a dataset obtained from Kaggle, which contains 1338 items that impressively represent the intricacies of medical spending in the United States. It is remarkable that we were able to attain an accuracy of 81.3% after carefully partitioning our data set, with 70% being used for training. Our comprehension of the intricacies of health insurance spending after COVID-19 is improved as a result of this study, which is driven by ethical considerations. In [3] determine the parameters that are connected with the utilization of Kampo formulations, we created a machine learning model and applied it to a database of health insurance claims. For the purpose of developing the training and testing sets, respectively, a sample of ten percent of the persons who enrolled in the JMDC Claims Database in 2018 and 2019 was utilized. The testing set was subjected to the application of models in order to compute the C-statistics. Furthermore, the performance of simplified scores that utilized either ten or five variables was considered and evaluated. During the training set, there were 338,924 participants, whereas during the testing set, there were 399,174 participants. The models were able to determine not only the common factors that are related with various Kampo formulations, but also the unique qualities that are connected with individual formulations. In [4] only ten percent of the counties in the United States were included in the cluster that had the highest life expectancy, which is the most efficient cluster classification. There are three unique clusters in the United States. Those machine learning clusters that are the most effective do not identify the clusters that have the most significant health care inequities. When applied to data at the county level, machine learning clustering reveals that the key factors that determine the composition of clusters are access to health care and infrastructure. In the Social Determinants of Health (SDOH) database, which was compiled by the Agency for Healthcare Research and Quality, we employed more than 650 variables that were compiled from 24 distinct databases. We clustered the data at the county level using k-means, which is a non-hierarchical machine learning clustering algorithm. In [5] the purpose of this study is to investigate this field with the objective of utilizing the capabilities of machine learning and deep learning methodologies in order to find opportunities for health insurance adoption among the existing customer base, regardless of the insurance holdings that they currently possess. This research highlights TabPFN's remarkable superiority over other algorithms by demonstrating a significant improvement in accuracy, which went from 52.62% to 63.10%. The incredible effectiveness of TabPFN, in particular with regard to the management of tabular data, is highlighted by this significant development. In [6] Through the application of Machine Learning Regression techniques, the purpose of this study is to forecast appropriate medical insurance expenses as a result of the patient's biological and demographic characteristics. On a dataset based in the United States, four models are implemented. Lasso, Elastic Net Regression, Gradient Boosting Regressor, and AdaBoost Regressor are all types of regression. I utilized a number of different loss functions in order to determine which model was the most accurate for each parameter. The approach that is being proposed will assist organizations in developing medical insurance policies that are more focused on the public interest, which will not only be beneficial to the users but will also increase the organization's revenue. In [7] the research presents three novel ensembles of supervised learning predictors for the purpose of regulating the expenses of medical insurance. For the purpose of developing methodologies for data analysis, the open dataset is utilized. This work intends to construct three novel ensembles for the purpose of predicting individual insurance costs in order to achieve a high level of accuracy in the prediction process. The quality of the ensemble that was developed through the use of the RootMean Squared Error metric is 1.47 times higher than that of the best weak predictor (SVR).

III. Proposed methodology

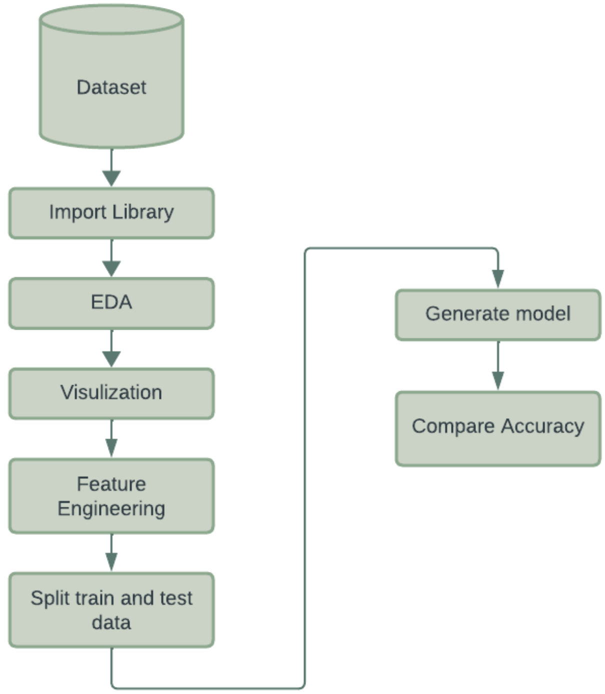

Fig. 1. The proposed methodology of the work

Our row dataset is preprocessed before we work with it. This is why we are modifying it in many ways. The machine is unable to handle row data.Following preprocessing, we get our goal value using an algorithm model. We make a series of advancements toward this continuous aim before arriving at the destination.

A. Dataset

The dataset of 79,210 rows and 8 columns that we obtained online was modified somewhat to meet our needs. Preprocessing is necessary to make sure that our various modeling techniques work with one other. The dataset was divided, with 70% designated for training and 30% for testing, to allow for a thorough assessment of the model's performance. The data will be preprocessed according to the distinct behavior of each approach, resulting in an optimized fit into the corresponding models for precise analysis and prediction. The data will be stored as a CSV file.

B. Data Discription

There are 8 columns and 79,210 items in the dataset. The monetary amounts of claim payments are shown in the 'Amount' column. "Severity" is a rating that indicates how serious the harm is to the patients; it goes from 1 for emotional distress to 9 for death. "Age" denotes the claimant's age expressed in years. "Private Attorney" indicates whether a private attorney was hired to represent the claimant. "Marital Status" keeps track of claimants' marital status. "Specialty" refers to the particular medical specialization of the plaintiff in the case. The term "insurance" designates the kind of health insurance that individuals own. In conclusion, the patient's gender is classified by 'Gender'. Each column's descriptive statistics will provide light on the data's variability, central tendency, and distribution, which will help with the preprocessing and modeling activities that follow.

C. Data Visualization

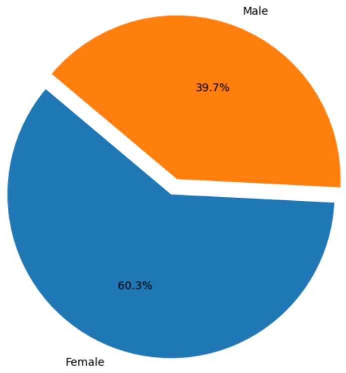

Fig. 2: Gender distribution Pie chart

In figure 2, are showing 60.3% female and 39.7% male in this dataset**.**

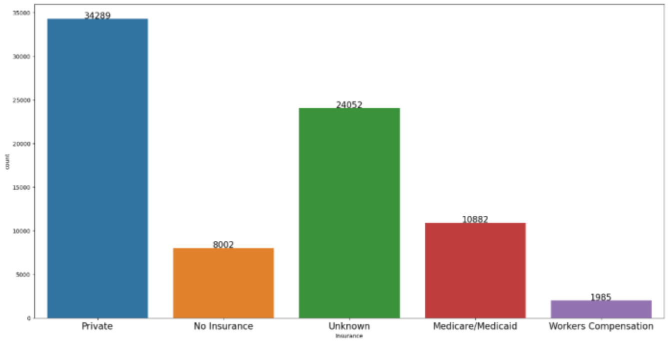

Fig. 3: Insurance distribution bar chart

This set of data has 5 different kinds of insurance. Private, no insurance, unknown, medi-care, and workers' compensation are some of them. Figure 3 shows how the insurance is split: 34289 people have private insurance, 8002 people have no insurance, 24052 people unknown insurance they have, 10882 people have Medicare, and 1985 people have workers' compensation.

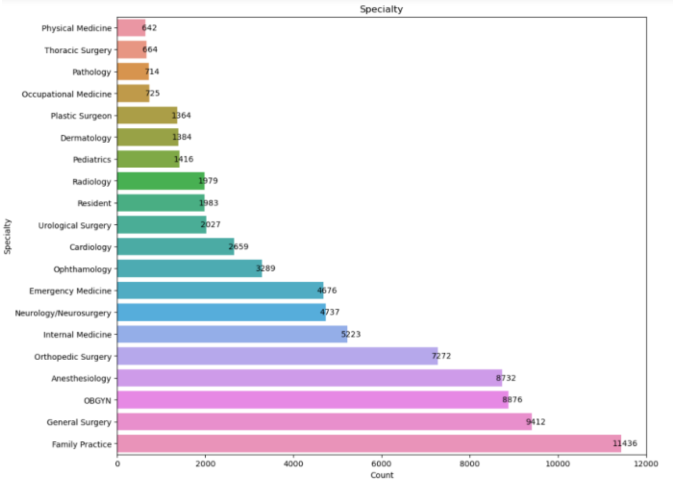

Fig.4: Specialty distribution bar chart

The number of practitioners across different medical specializations is seen in figure 4. With 11,436 practitioners, family practice has the largest number of practitioners. General surgery and obstetrics and gynecology have 9,412 and 8,876 practitioners, respectively. Both orthopedic surgery and anesthesiology, with 8,732 and 7,272 practitioners, respectively, are well represented. With 5,223 practitioners, Internal Medicine comes next, followed by Neurology/Neurosurgery with 4,737 practitioners and Emergency Medicine with 4,676 practitioners. There are fewer practitioners in cardiology and ophthalmology, with 2,659 and 3,289, respectively. Radiology, pediatrics, dermatology, plastic surgery, urological surgery, and residents are the specialties with the lowest representation.

D. Label Encoding

Label encoding is a method that is used in the field of machine learning for the purpose of converting descriptive variables into numerical representations. As a result, categorical data is converted into a format that can be submitted to machine learning algorithms. This is accomplished by assigning a distinct number to each category that is included inside a feature.

E. Model Generate

Decision Tree Regressor:

DecisionTreeRegressor applies supervised machine learning to regression challenges. A decision tree-based learning component, that predicts numerical outcomes continuously. It produces a binary tree structure by continually separating the dataset into smaller groups by characteristics and values. Data is divided by the best characteristics and values at each tree node using the algorithm. Reduce target variable variance in the partitioned subset. This approach continues until a target stopping point is achieved, such as the maximum tree depth, the minimum leaf node samples, or other criteria. After tree building, each terminal node (leaf) has a prediction value, usually the average of the samples' target values. It can handle category and numerical data and discover non-linear relationships between goal variables and descriptors. If not regularized or modified, it tends to overfit the training set and perform badly on unknown data.

Random Forest Regressor:

Random forest regression solves regression issues via ensemble learning. In this Random Forest method modification, multiple decision trees boost prediction accuracy. To estimate regression tree averages, algorithms generate numerous decision trees during training. A random forest builds several decision trees using replacement-based random samples of the training data set. Some dataset properties and data components are used in each tree. The plants are constructed concurrently and separately. Regression trees provide numbers. Random forest regression's final predictor is each tree's output mean. The model predicts new data using the average tree prediction. Random forest regression handles huge, high-dimensional datasets well. It performs effectively with few hyperparameter modifications even when overfitting.

XGB Regressor:

XGBRegressor is Extreme Gradient Boosting. Regressor is a powerful supervised machine learning technique for regression problems. The XGBoost (Extreme Gradient Boosting) library is notable for its performance in machine learning challenges and real-world applications. Gradient Boosting underpins it. Stronger prediction models like decision trees incorporate weak learners progressively. New models are trained to additively fix past model addition deficiencies. Each model aims to reduce residual errors from previous iterations. XGBoost has learning rate, maximum tree depth, subsampling, and column subsampling regularization hyperparameters. They reduce overfitting and improve model generalization to fresh data. XGBOST values efficiency and performance. It calculates quicker than gradient boosting implementations using parallel processing, tree pruning, and cache awareness.The complex tree building methods of XGBoost enable rapid tree pruning and split search, enhancing generalization and minimizing overfitting.

KNN Regressor:
The supervised machine learning approach KNeighborsRegressor solves regression issues. It is an instance-based learning method that modifies the k-nearest neighbor (KNN) regression methodology. tutorial. Memory is used to store the training dataset. KNeighborsRegressor predicts new data points by finding the k-nearest neighbors to the given data point using a distance metric like Euclidean distance. is KNeighborsRegressor's representation of 'k'. how many neighbors to consider while predicting. Users must set this hyperparameter A larger 'k' number may provide smoother predictions but ignore local patterns, whereas a lower value may produce a more flexible model but be more noisy. The KNeighborsRegressor algorithm is easy to understand and utilize. However, the distance measure and number of neighbors ('k') may affect performance. It may not work well in high-dimensional environments or datasets with irrelevant or noisy properties.

IV. Results and Discussion

Table 1. Performances of Different Classifiers

Algorithm Name	**R2 Score**
Decision Tree Regressor	0.85
Random Forest Regressor	0.83
XGB Regressor	0.70
KNN Regressor	0.67

The R2 score measure is used to display the performance assessment results of several regression methods in the table. With an R2 value of 0.85, the Decision Tree Regressor demonstrated its efficacy in capturing the variation of the dependent variable. The Random Forest Regressor, which has an R2 value of 0.83 and good predicting ability, comes in second. With a score of 0.70, the XGB Regressor performed somewhat worse in terms of prediction accuracy than the random forest and decision tree models. Finally, out of all the algorithms that were assessed, the KNN Regressor had the lowest R2 score of 0.67, suggesting somewhat worse predictive ability. Based on the R2 score measure, the Decision Tree and Random Forest Regressors fared better overall than the other models.

V. Conclusion

Machine learning methods for predictive analysis in the field of health insurance are worth looking into because they could help people make better decisions and get better health results. This research looks at how different machine learning methods can be used to look at old data and guess what will happen in the future with health insurance use, costs, and risk rating. Using advanced methods like regression, classification, and grouping, this study aims to find trends in large, complicated healthcare records. This will help lawmakers, insurance, and healthcare workers by giving them useful information. This study looks at these methods in detail in order to help improve resource sharing, risk management, and the general efficiency of health insurance systems. This will lead to a more sustainable and flexible healthcare environment in the long run .In the table 1, the R2 score is used to show how well different regression methods worked in a performance review. The Decision Tree Regressor did a good job of catching the change of the dependent variable with an R2 value of 0.85. The Random Forest Regressor comes in second. It has an R2 number of 0.83 and can make good predictions. With a score of 0.70, the XGB Regressor wasn't quite as good at making predictions as the random forest and decision tree models. Lastly, the KNN Regressor had the lowest R2 score of all the algorithms that were looked at, which means it was not as good at making predictions. The R2 score showed that the Decision Tree and Random Forest Regressors did better than the other models in general.

References

[1] U. Orji and E. Ukwandu, “Machine learning for an explainable cost prediction of medical insurance,” Machine Learning with Applications, vol. 15, p. 100516, Mar. 2024. doi:10.1016/j.mlwa.2023.100516

[2] A. Sharma and R. Jeya, “Prediction of insurance cost through ML structured algorithm,” 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), Feb. 2024. doi:10.1109/ic2pct60090.2024.10486304

[3] H. Yamana et al., “Machine learning-based models for outpatient prescription of Kampo formulations: An analysis of a health insurance claims database,” Journal of Epidemiology, vol. 34, no. 1, pp. 8–15, Jan. 2024. doi:10.2188/jea.je20220089

[4] D. M. Bowser, K. Maurico, B. A. Ruscitti, and W. H. Crown, “American clusters: Using machine learning to understand health and health care disparities in the United States,” Health Affairs Scholar, vol. 2, no. 3, Feb. 2024. doi:10.1093/haschl/qxae017

[5] J. Z. Chu, J. C. Than, and H. S. Jo, “Deep learning for cross-selling health insurance classification,” 2024 International Conference on Green Energy, Computing and Sustainable Technology (GECOST), Jan. 2024. doi:10.1109/gecost60902.2024.10475046

[6] H. M. Alzoubi et al., “Analysis of cost prediction in medical insurance using modern regression models,” 2022 International Conference on Cyber Resilience (ICCR), Oct. 2022. doi:10.1109/iccr56254.2022.9995926

[7] N. Shakhovska, N. Melnykova, V. Chopiyak, and M. Gregus ml, “An ensemble methods for Medical Insurance Costs Prediction Task,” Computers, Materials &amp; Continua, vol. 70, no. 2, pp. 3969–3984, 2022. doi:10.32604/cmc.2022.019882

[8] J. J.-C. Ying, P.-Y. Huang, C.-K. Chang, and D.-L. Yang, “A preliminary study on Deep Learning for predicting social insurance payment behavior,” 2017 IEEE International Conference on Big Data (Big Data), Dec. 2017. doi:10.1109/bigdata.2017.8258131

[9] K. Dutta, S. Chandra, M. K. Gourisaria, and H. GM, “A data mining based target regression-oriented approach to modelling of health insurance claims,” 2021 5th International Conference on Computing Methodologies andCommunication(ICCMC),Apr.2021.doi:10.1109/iccmc51019.2021.9418038

[10] H. M. Alzoubi et al., “Analysis of cost prediction in medical insurance using modern regression models,” 2022 International Conference on CyberResilience(ICCR),Oct.2022. doi:10.1109/iccr56254.2022.9995926

[11] U. Orji and E. Ukwandu, “Machine learning for an explainable cost prediction of medical insurance,” Machine Learning with Applications, vol. 15, p. 100516, Mar. 2024. doi:10.1016/j.mlwa.2023.100516

[12] K. Kaushik, A. Bhardwaj, A. D. Dwivedi, and R. Singh, “Machine learning-based regression framework to predict health insurance premiums,” International Journal of Environmental Research and Public Health, vol. 19, no. 13, p. 7898, Jun. 2022. doi:10.3390/ijerph19137898

[13] S. Chowdhury, P. Mayilvahanan, and R. 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, vol. 44, no. 3, pp. 278–290, Feb. 2020. doi:10.1080/1206212x.2020.1733307

[14] N. J. Ravindran and P. Gopalakrishnan, “Predictive analysis for healthcare sector using Big Data Technology,” 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT), Aug. 2018. doi:10.1109/icgciot.2018.8753090

[15] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from Healthcare Communities,” IEEE Access, vol. 5, pp. 8869–8879, 2017. doi:10.1109/access.2017.2694446