Journal of Computing and Communication Vol.3, No.2, PP. 55-64, 2024

An Efficient Framework for Predict Medical Insurance Costs Using Machine Learning

Maged **Farouk**[a], Nashwa S **Ragab**[a], Diaa **Salama***[a,b], Omnia **Elrashidy**[a], Loaa Allah **Youssef**[a], Abdelrahman **Ehab**[a], Eman **Mohamed**[a], Hana **Mohamed**[a], Chahd **Islam**[a,] Marina **Rafat**[a], Reda **Elazab**[a]

[a] *Department of Business Information Systems, Faculty of Business Administrative, Alamiem University, Alamein, Egypt*

[b] *Faculty of Computers Science, Misr International University, Cairo, Egypt*

[c], *Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt*

*Corresponding Author: Diaa Salama [***diaa.salama@miuegypt.edu.eg***]

ARTICLE DATA

Article history:
Received 07 Jan 2024
Revised 24 July 2024
Accepted 24 July 2024
Available online

Keywords:
Medical Insurance prediction
Prices of insurances
Machine learning
Tree
Linear regression

ABSTRACT

One of the applications of machine learning in predicting medical insurance prices considering health and economic factors is because this branch analyzes how healthcare resources are allocated and how healthcare outcomes are determined. The production of medical insurance prices encounters challenges rooted in data accuracy and ethical consideration of machine learning models. In this paper, we proposed an efficient framework for predicting medical insurance prices and a delicate balance between accuracy and fairness to ensure the efficacy and ethical soundness of the pricing process using five machine learning algorithms MAPE , R2. On four different datasets, Cross-validation number of folder:5 and the best result on MAPE is a tree with the smallest number of errors was 3.9%, Cross-validation number of folder:10 and the best result on MAPE is a tree with the smallest number of errors was 3.5%, Random sampling training set size 80% and testing 20% the best result on MAPE is a tree with the smallest number of errors was 4.1%,%, Random sampling training set size 90% and testing 10% the best result on MAPE is Tree with the smallest number of errors was 4%. The best result of all datasets on MAPE is Tree.

1. Introduction

Health economics is a growing exploration field. The number of handbooks in this field has increased as well. Health economics is a field within the more general field of applied microeconomics. But unlike other microeconomic operations, like public economics, labor economics, and artificial association, health economics is extensively tutored outside economics departments, similar to public health schools (ag. in hospital administration programs). Public policy, business, nursing, drugstore, and medical insurance prices. The demand for health economics stems much further from an essential interest in health care than a primary interest in farming [1].

Prophetic modeling in healthcare continues to be an active actuarial exploration as further insurance companies aim to maximize the eventuality of Machine learning (ML) approaches to increase their productivity and effectiveness [2].

In recent times, actuarial modeling of insurance claims has become a crucial exploration area in the health insurance sector and is substantially applied in setting effective premiums (Duncan et al., 2016). This is vital for attracting and retaining insureds and for effective management of being plan members. Still, due to the variety of factors that drive medical insurance costs and the complications therein, there's a bit of a challenge in accurately a predictive model for it. Factors like demographic information, health status, geographic access, life choices, provider characteristics, etc., can dramatically impact the anticipated medical insurance costs. Other vital factors like the scope of coverage, type of plan, deductible, and the age of a client when they subscribe also play a major part in determining the implicit cost of medical insurance.

The significance of an effective and transparent medical insurance system cannot be overemphasized, considering the need for universal healthcare content and the challenges of the COVID-19 epidemic (Orji et al., 2022b). Orji et al. (2022a) described how the prophetic analytics points of ML have become the most employed point for artificial operations and industrial applications. The ongoing regulatory and market changes in the health industry motivate actuarial exploration into prophetic modeling in health care. ML algorithms have proven to yield accurate results in predicting high-cost, high-need case expenditures.

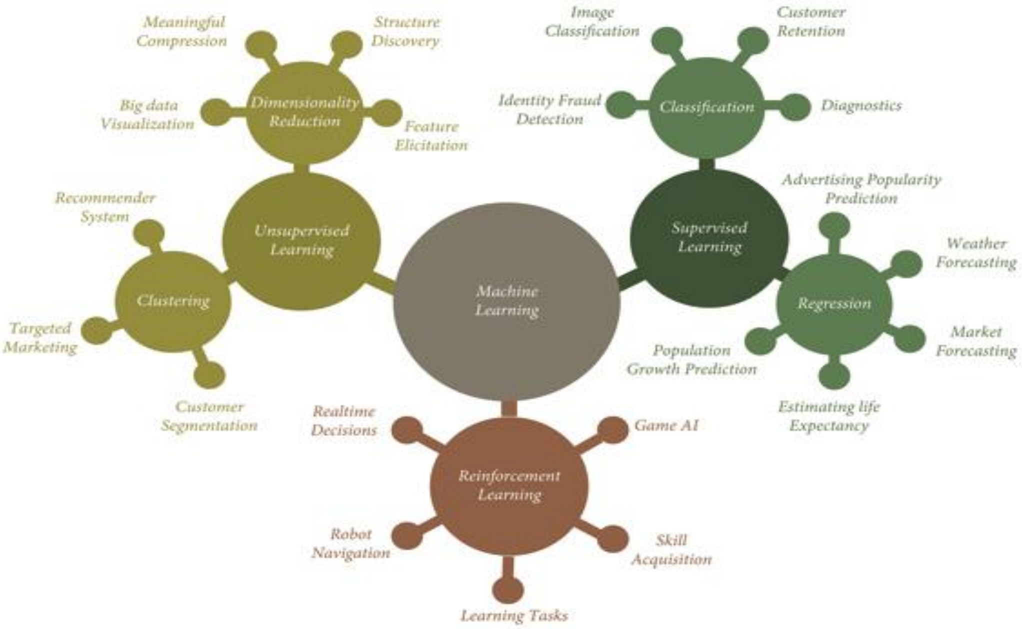

FIGURE 1. Types of Machine Learning

Therefore, insurance companies are increasingly turning to ML to improve their policies and premium settings (Yang et al., 2018). Still, the high performance of ML algorithms in healthcare, as noted by (Panay et al., 2019), is somewhat offset by their black-box nature. If the prophetic analytics involving cases of patients' personal and clinical information are not well understood or explained, they could be vulnerable to bias. Fortunately, the emergence of Explainable Artificial Intelligence (XAi) methods allows those involved – cases of patients, healthcare managers, and insurers - to gain further insight into the logic behind predictions, consequently promoting transparency and acceptability. As a result, predicting medical insurance costs with high certainty would solve problems such as responsibility and transparency, enabling control over all parties involved in patient care (Panay et al., 2019) [3].

The main contribution of this paper can be summarized as follows: We used four different dataset tables of Cross-Validation, Number of folds:5, table of Cross-Validation, Number of folds:10, table of Random Sample, training set size: 80% and testing size: 20%, and table of Random Sample, training set size: 90%, testing size: 10%.and we used six algorithms random Forest, Nouran network, decision tree, KNN, SVM, Linear regression and the best result was decision tree and Random forest.

The rest of the paper can be organized as follows: We will discuss the other five main points Related work, Methodology, Results, and discussion and conclusion.

3. Related Work

Many attempts to automate medical insurance have recently been made. Various algorithms and machine translation approaches are used in these efforts. Several authors used neural machine translation of medical insurance, similar to the method proposed in this paper.

In [4], the authors presented three ensemble ML models, Boost, GBM, and RF were deployed for medical insurance cost prediction using the medical insurance cost dataset from KAGGLE's repository. The comparative result from our prediction showed that all the models achieved impressive outcomes. In contrast, the Boost model achieved the best R2 score of 86.470 % and RMSE of 2231.524, and it also took a more substantial computing resource to complete, as shown in Table 7. The RF model achieved a better MAE and MAPE outcome of 1379.960 and 5.831 %, respectively, and was the fastest model to build and consumed less memory than the Boost model. The result showed that the GBM model had more large-scale prediction errors than the Boost and RF models.

In [5], the authors reported a similar study that employed the sequence-to-sequence approach. The authors propose the Advanced Aspects of Computational Intelligence and Applications of Fuzzy Logic and Soft Computing by using a set of ML algorithms; a computational intelligence approach is applied to predict healthcare insurance costs. The medical insurance dataset was obtained from the KAGGLE repository. It was utilized for training and testing the Linear Regression, Ridge Regressor, Support Vector Regression, Boost, Stochastic Gradient Boosting, Decision Tree, Random Forest Regressor, k-nearest Neighbors, and Multiple Linear Regression ML algorithms. The regression of this dataset followed the steps of preprocessing, feature engineering, data splitting, regression, and evaluation. The resultant outcome revealed that Stochastic Gradient Boosting (SGB) achieved a high accuracy of 86% with an RMSE of 0.340.

In [6], the authors presented a Machine Learning-Based Regression Framework to Predict Health Insurance Premiums, in this study, the authors trained an ANN-based regression model to predict health insurance premiums. The model was then evaluated using key performance metrics, i.e., RMSE, MSE, MAE, r2, and adjusted r2. The accuracy of our model was 92.72%. Moreover, the correlation matrix was also plotted to see the relationship between various factors and charges. This domain of insurance prediction has not been fully explored and requires thorough research.

In [7] An article used individuals' health data to forecast their insurance premiums. To assess and evaluate the performance of various algorithms, regression was utilized. The dataset was used to train the models, and the training results were utilized to make predictions. The model was then tested and verified by comparing the anticipated quantity to the actual data. The accuracy of these models was later compared. According to the findings, multiple linear regression and gradient boosting algorithms outperformed linear regression and decision trees. Gradient boosting was suitable in this scenario since it required far less computing time to attain the same performance measure, although its performance was equivalent to multiple regression. Finally, in this study, we deployed two algorithm methods, MAPE and R2, for a deeper analysis of how individual features in the dataset affect the overall outcome of our models. We further compared the performance of each algorithm method deployed, and the best was Tree.

4. Methodology

4.1 Dataset Description

The first dataset consists of 6 features; the dataset was split into two positions: 66% for training and 34% for testing. A detailed description of the features can be found below. Age Category represents which age range the person belongs to. Sex represents gender; the human is male or female. BMI is an abbreviation for Body Mass Index, which is computed by multiplying a person's weight in kg by the square of their height. A smoker represents a smoking status smoker or non-smoker.

TABLE I

Features of Medical Insurance Prices

Feature	Type	Values
Age	Numerical	From 18 to 64
Sex	Classification	Male or Female
BMI	Numerical	From 18 to 53.1
Children	Numerical	From 0 to 5
Smoker	Classification	Yes or no
Region	Classification	Southwest or Northwest

The first dataset consists of 6 features; the dataset was split into two positions: 80% for training and 20% for testing. A detailed description of the features can be found below. The age category represents which age range the person belongs to. Diabetes whether the Person Has Abnormal Blood Sugar Levels or not. Blood Pressure Problems Whether the Person Has Abnormal Blood Pressure Levels. Any Transplant represents any Major Organ Transplants. Any Chronic Diseases represent Whether the Customer Suffers from Chronic Ailments Like Asthma. Height represents the height of the customer. Weight represents the weight of the customer. Known Allergies represent Whether the Customer Has Any Known Allergies. History of cancer in the family represents Whether Any Blood Relative of the Customer Has Had Any Form of Cancer. Number of major surgeries represents The Number of Major Surgeries The Person Has Had.

TABLE II

Features Medical Insurance Premium Prices

Feature	Type	values
Age	Numerical	From 18 to 66
Diabetes	Numerical	0 or 1
Blood Pressure Problems	Numerical	0 or 1
Any Transplants	Numerical	0 or 1
Any Chronic Diseases	Numerical	0 or 1
Height	Numerical	From 145 to 188
Weight	Numerical	From 51 to 132
Known Allergies	Numerical	0 or 1
History of cancer in the family	Numerical	0 or 1
Number of major surgeries	Numerical	From 0 or 3

The mentioned datasets were passed into different Machine Learning algorithms: neural network, SVM, KNN, linear regression, Decision tree, and Random Forest. For each of the algorithms there, statistics were generated, these statistics were MSE, RMSE, MAE, MAPE, R2. The results were charted and compared. The results, charts, and the discussion can be found later in the paper.

4.2 Used Algorithm

1) Neural Network: A Neural Network is a machine learning algorithm inspired by the structure and functioning of the human brain. It consists of interconnected nodes, or artificial neurons, organized into layers. It is supervised and unsupervised because neural networks are highly flexible and can be adapted to various learning paradigms, making them powerful tools in machine learning [8]. The focus was to drop and decrease the mean absolute chance and percentage error by conforming parameters like time, learning rate, and neurons in different layers. Feed-forward and intermittent neural networks were enforced and implemented to forecast the yearly claims amount data [9].

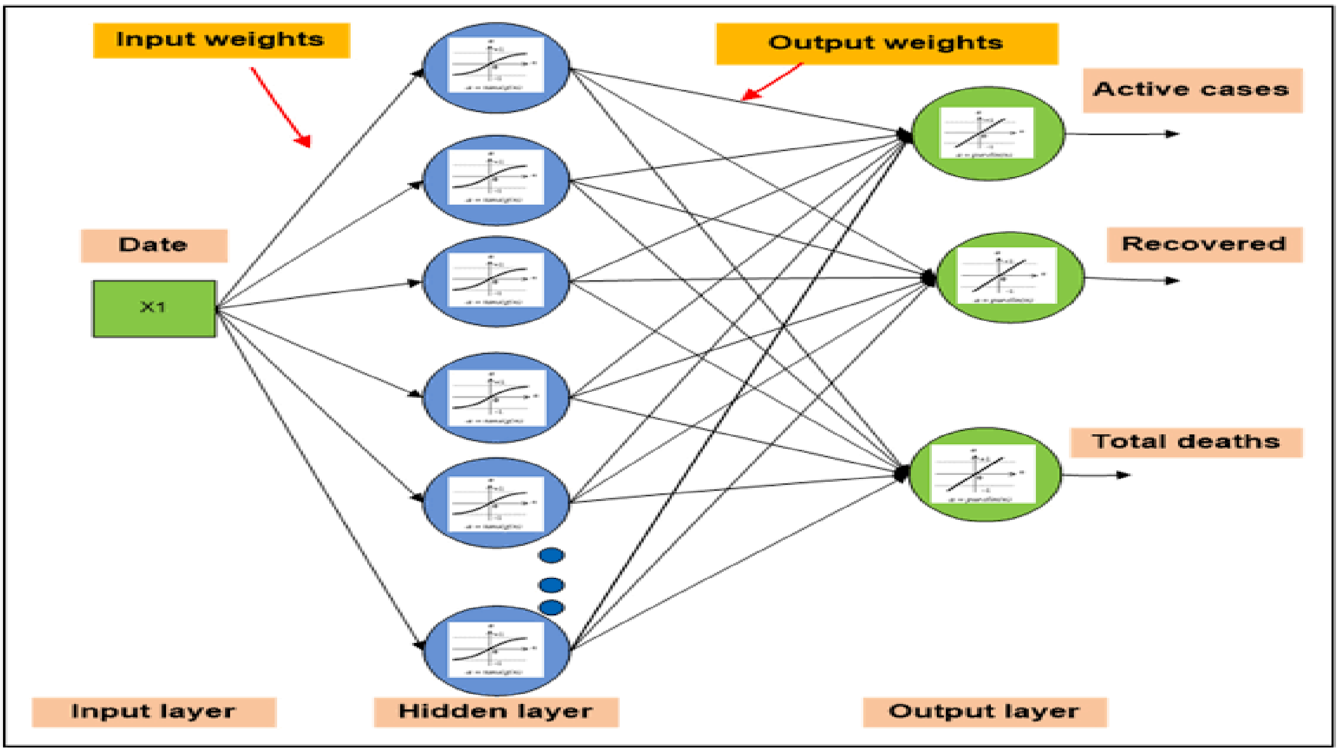

FIGURE 2. Neural Network

2) Random forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset [10]. A random forest is a supervised machine learning algorithm used for classification and regression problems in machine learning. Classification: Predict the input. Regression: Predict the output. The random forest reduces the overfitting and increases accuracy, robustness, features important, and scalability [11].

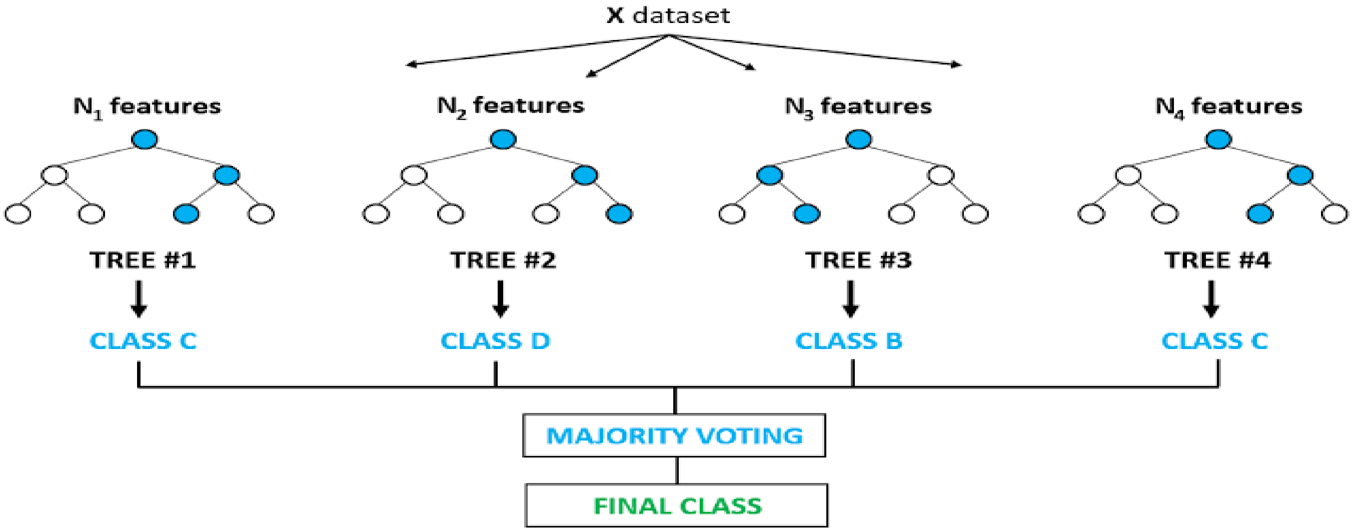

FIGURE 3. Random Forest

3) A support vector machine (SVM): is defined as a machine learning algorithm that uses supervised learning models to solve complex classification [12], regression, and outlier detection problems by performing optimal data transformations that determine boundaries between data points based on predefined classes, labels, or outputs It can handle complex datasets and generalize well to new, unseen data [13].

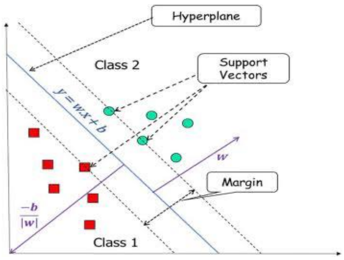

FIGURE 4. SVM

4) K-neatest neighbors (KNN): a non-parametric, supervised learning classifier that uses proximity to make classifications or predictions about the grouping of an individual data point [14].

It is used for regression tasks; the prediction will be based on the mean and median of the K closet observations. Is used for classification purposes; the mode of the closet observations will serve for prediction. KNN is a simple yet effective algorithm for small to moderately-sized datasets when the decision boundary is not highly complex [15].

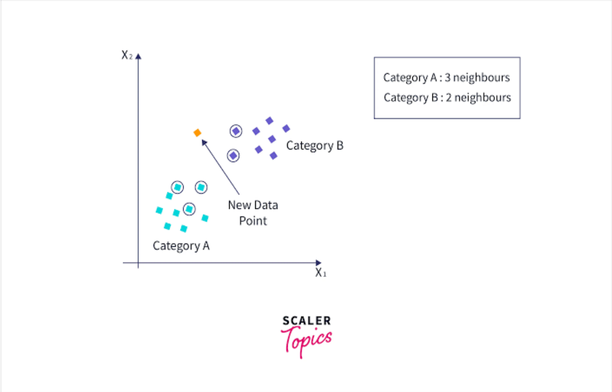

FIGURE 5. KNN

5) Linear regression: a supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features The algorithm estimates the coefficients of the linear equation that best fits the data [16,17].

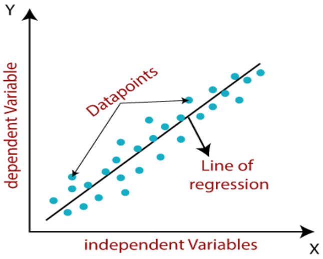

FIGURE 6. Linear regression

6) Decision tree: is a Supervised learning technique that can be used for classification and Regression problems, but it is mostly preferred for solving Classification problems. A Decision tree has two nodes: the Decision Node and the Leaf Node [18]. Decision nodes are used to make decisions with multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the tests are performed based on features of the given dataset [19]. In a decision tree, the output is mostly "yes" or "no" [20].

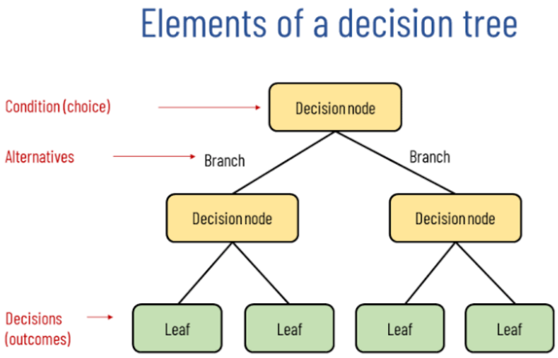

FIGURE 7. Decision Tree

5. Results and Discussion

This Table of Cross Validation, Number of folds:10 and we will focus on the result on MAPE and R2. Results of Medical Insurance Prices:

TABLE III
Statistics of algorithms with 10 K-folds

Model	MAPE	R2
Random forest	0.046	0.784
Tree	0.035	0.78
Linear Regression	0.117	0.63
KNN	0.108	0.535
SVM	0.218	-0.045
Neural Network	0.979	-14.504

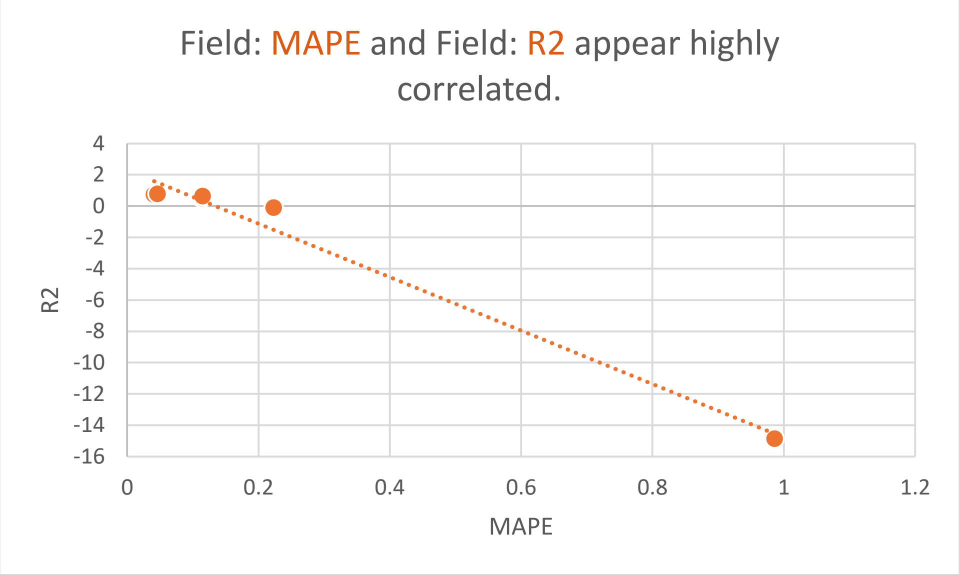

FIGURE 8.First dataset performance chart with ten k-fold

We depend on MAPE Results: The first and best result with the lowest number of errors is the Decision Tree 0.035. The second-best result was Random Forest with 0.046 number of errors. The third best result was KNN with 0.108 number of errors. The fourth best result was Linear regression with 0.117 number of errors. The fifth best result was SVM with 0.218 number of errors. The worst result was Neural Network with 0.979 number of errors.

This Table of Random Sample, Training set size: 80%, Testing size:20% and we will focus on the result on MAPE and R2.

TABLE V:
Statistics of algorithms with 80/20 data split

Model	MAPE	R2
Neural network	0.986	-14.855
SVM	0.223	-0.092
KNN	0.114	0.505
Linear regression	0.115	0.639
Tree	0.041	0.738
Random forest	0.046	0.790

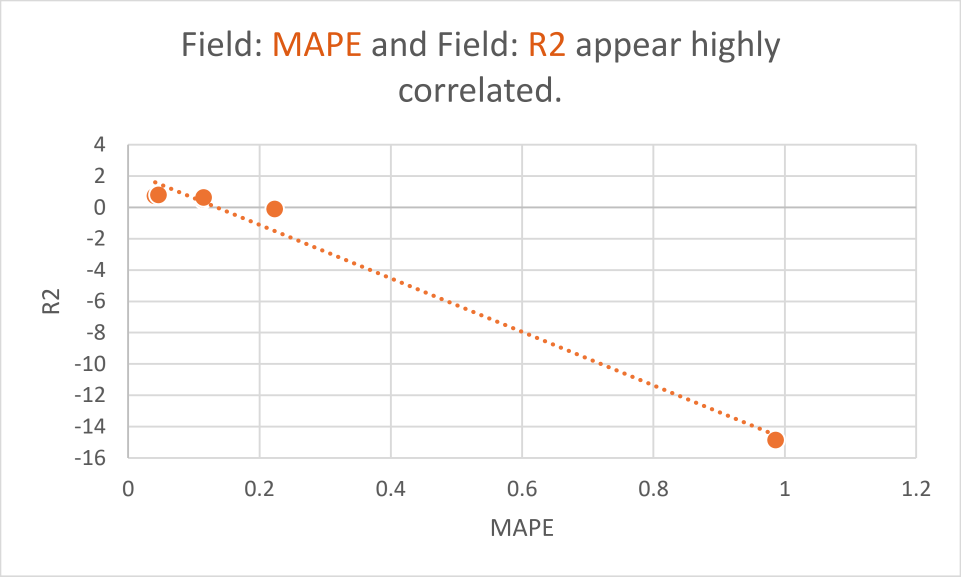

FIGURE 9. First dataset performance With data split 80/20.

We depend on MAPE Results: The first and best result with the lowest number of errors is the Tree 0.041. The second-best result was Random Forest with 0.046 number of errors. The third best result was KNN with 0.114 number of errors. The fourth best result was Linear regression with 0.115 number of errors. The fifth best result was SVM with 0.223 number of errors. The worst result was a Neural network with 0.986 number of errors.

Result of Medical Insurance Premium Prices:
This Table of Cross Validation, Number of folds:10 and we will focus on the result on MAPE and R2.

TABLE VI
Statistics of algorithms with 10 k-folds

Model	MAPE	R2
Random Forest	0.046	0.784
Descion Tree	0.035	0.78
Linear Regression	0.117	0.63
kNN	0.108	0.535
SVM	0.218	-0.045
Neural Network	0.979	-14.504

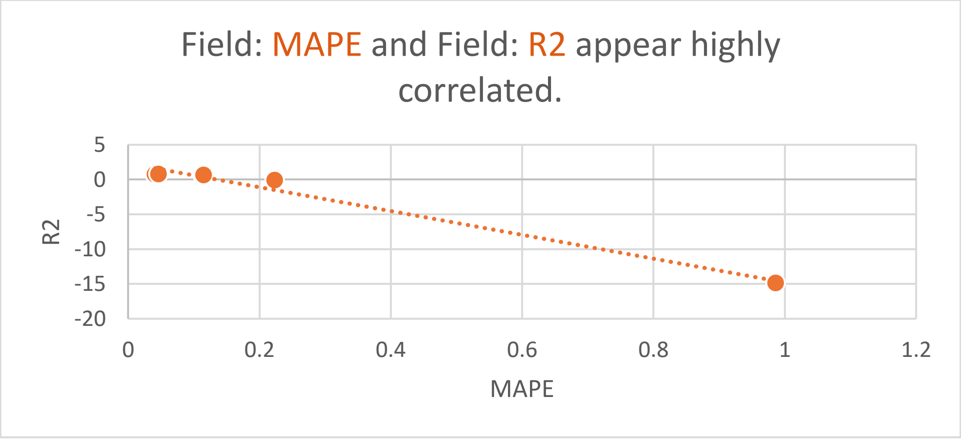

FIGURE 10.Second dataset performance chart with 10 k-fold

We depend on MAPE Results: The first and best result with the lowest number of errors is the Tree 0.035. The second-best result was Random Forest with 0.046 errors. The third best result was KNN with 0.108 number of errors. The fourth best result was Linear regression with 0.117 number of errors. The fifth best result was SVM with 0.218 number of errors. The worst result was a Neural network with 0.979 number of errors.

This Table of Random Sample, Training set size: 80%, Testing size:20% and we will focus on the result on MAPE and R2.

TABLE VII

Statistics of algorithms with 80/20 data split

Model	MAPE	R2
Neural network	0.986	-14.855
SVM	0.223	-0.092
KNN	0.114	0.505
Linear regression	0.115	0.639
Decision tree	0.041	0.738
Random forest	0.046	0.790

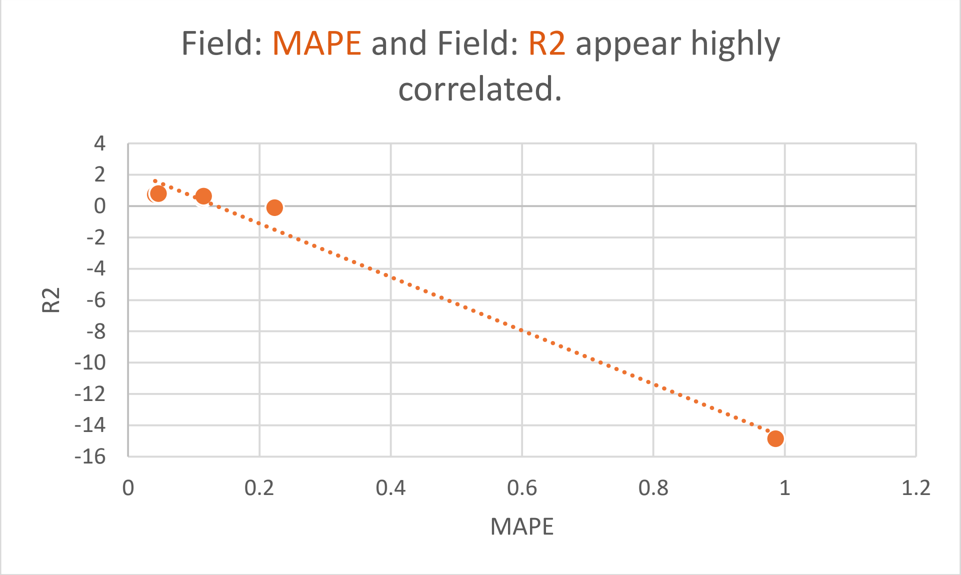

FIGURE 11. First dataset performance With data split 80/20.

We depend on MAPE Results: The first and best result with the lowest number of errors is the Tree 0.041. The second-best result was Random Forest with 0.046 number of errors. The third best result was KNN with 0.114 number of errors. The fourth best result was Linear regression with 0.115 number of errors. The fifth best result was SVM with 0.223 number of errors. The worst result was a Neural network with 0.986 number of errors.

6. Conclusion

We used four different datasets: table of Cross-Validation, Number of folds:5, table of Cross-Validation, Number of folds:10, table of Random Sample, the training set size: 80% and testing size: 20%, and table of Random Sample, the training set size: 90%, testing size: 10%. We Focused our dataset on MAPE and R2.

The best result on MAPE of Cross Validation, Number of folds:10 and in the algorithm of Decision Tree with 3.5% number of errors. The best result on R2 of Random Sample, the training set size: 90%, testing size:10% in the algorithm of Neural Network with -15.206 number of errors.

References

[1] Sloan, F. A., & Hsieh, C. R. (2017). *Health economics*. MIT Press.

[2] ul 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.

[3] Orji, U., & Ukwandu, E. (2024). Machine learning for an explainable cost prediction of medical insurance. Machine Learning with Applications, 15, 100516.

[4] Orji, U., & Ukwandu, E. (2024). Machine learning for an explainable cost prediction of medical insurance. Machine Learning with Applications, 15, 100516.

[5] ul 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.

[6] 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.

[7] Boodhun, N.; Jayabalan, M. Risk Prediction in Life Insurance Industry Using Supervised Learning Algorithms. Complex Intell. Syst. 2018, 4, 145–154. [Google Scholar] [CrossRef] [Green Version]

[8] Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11).

[9] Dongare, A. D., Kharde, R. R., & Kachare, A. D. (2012). Introduction to artificial neural network. International Journal of Engineering and Innovative Technology (IJEIT), 2(1), 189-194.

[10] Rigatti, S. J. (2017). Random forest. Journal of Insurance Medicine, 47(1), 31-39.

[11] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

[12] Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. Cancer genomics & proteomics, 15(1), 41-51.

[13] Noble, W. S. (2006). What is a support vector machine?. Nature biotechnology, 24(12), 1565-1567.

[14] Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.

[15] Kramer, O., & Kramer, O. (2013). K-nearest neighbors. Dimensionality reduction with unsupervised nearest neighbors, 13-23.

[16] Weisberg, S. (2005). Applied linear regression (Vol. 528). John Wiley & Sons

[17] Groß, J. (2003). Linear regression (Vol. 175). Springer Science & Business Media.

[18] Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

[19] Suthaharan, S., & Suthaharan, S. (2016). Decision tree learning. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, 237-269.

[20] Su, J., & Zhang, H. (2006, July). A fast decision tree learning algorithm. In Aaai (Vol. 6, pp. 500-505).