Journal of Actuarial, Finance and Risk Management (JAFRM)
Vol. 2(2), December 2023.

Comparison Between Machine Learning Regression Modelling to Predict Individual Premium Price

Srava Chrisdes Antoro[1], Elisabeth Gloria Manurung[2*], Jessykapna Randalline [3], Maria Yus Trinity Irsan[4]

[1]Universitas Gunadarma, Depok,16424, Indonesia
[2 ,3,4]Universitas Presiden,,Bekasi Regency, 17550, Indonesia
[*]Corresponding author: elisabeth.manurung@student.president.ac.id

Abstract—Machine Learning (ML) applications in healthcare have emerged as powerful tools for predicting and diagnosing diseases, surpassing traditional medical expert capabilities. This study explores the integration of machine learning, specifically regression methods such as K-Nearest Neighbours (KNN), Random Forest, and Boosting, in the context of digital health insurance. The aim is to bridge the gap between insurance providers and policyholders, establishing a direct and efficient connection that transforms the process of creating health insurance policies.In this research, our primary objective is to predict health insurance premium prices accurately. We trained and evaluated regression models on a dataset incorporating factors such as age, diabetes, blood pressure issues, height, and weight. Our experimentation revealed that the KNN method outperformed other regression techniques, achieving an impressive accuracy of 87.73%, followed by Boosting at 87.19%, and Random Forest at 87%. The authors thoroughly analysed the models using key metrics to assess their effectiveness.

Keywords—*Regression, KNN, Random Forest, Boosting, Premium*

I. INTRODUCTION

Health, being paramount in our lives, is influenced by various factors that drive the need for reliable health assurance. Health insurance serves as a crucial mechanism to safeguard individuals against unpredictable health-related expenses. Whether obtained individually or through group plans offered by employers, the significance of health insurance cannot be overstated. However, as the complexities of health insurance costs increase, so does the challenge of accurately predicting and managing these expenses.

Understanding the dynamics of health insurance costs is pivotal for effective resource allocation, policy planning, and risk assessment. Traditional manual calculations become cumbersome and time-consuming as datasets grow in complexity and volume. In this context, the integration of machine learning models emerges as a promising solution, offering efficiency and precision to insurance companies.

Machine learning methods, with their adaptability and predictive capabilities, have become invaluable for modeling and forecasting health insurance costs. Despite the transformative potential of these technologies, a comprehensive comparison of various machine learning models, especially within the context of health insurance cost prediction, remains underexplored.

Several studies on medical price estimation have been published in the health sector in different contexts [1],[2],[3]. Machine learning makes a variety of assumptions, but its effectiveness depends on selecting algorithms that closely match a particular problem domain and following appropriate procedures to build, train, and deploy models.

This paper addresses this gap by exploring and comparing different machine learning models in predicting health insurance costs. It aims to evaluate the effectiveness of these models and provide insights into their application within the healthcare domain. By establishing a clear link between the importance of health insurance and the potential of machine learning, this research seeks to contribute to the advancement of predictive modeling in the healthcare sector.

II. LITERATURE REVIEW

A. Premium

The premium is a sum of money paid by the insured party and received by the insurer as compensation for any damage, loss, or when the policyholder experiences a loss. The amount of the premium can be determined through risk selection carried out by the under writer, or if the company has already selected the risks at the request of potential policyholders, then the prospective policyholders pay insurance premiums based on their respective risk levels. The amount of the premium for participation in the insurance that must be paid has been set by the insurance company, considering the circumstances of the policyholders [4].

B. Deductibles

Deductibles (or also known as claim deductibles) are the amount of money that policyholders must pay before the insurance company starts covering the insurance claim expenses. When policyholders file an insurance claim, they have to pay the deductible first from their own pockets before the insurance company pays the remaining amount of the claim that exceeds the deductible. Deductibles are designed to prevent frequent small claims, making policyholders responsible for the initial costs before insurance becomes active. Generally, the higher the deductible amount, the lower the premiums that policyholders have to pay [5].

C. Machine Learning

Machine learning (ML) technology, developed to learn autonomously without direct user intervention, leverages disciplines such as statistics, mathematics, and data mining. ML has the ability to analyze and learn from data without explicit programming, making it adaptable to various tasks. Supervised and unsupervised learning are two fundamental ML techniques, with this paper utilizing methods such as K-Nearest Neighbor (KNN), Random Forest, and Boosting.

D. Modelling

Following data preparation and exploratory data analysis, the modeling phase involves predicting health insurance costs based on dataset factors. Machine learning models employed include:

K-Nearest Neighbor (KNN) is a supervised learning algorithm for classification and regression. KNN predicts by considering the k-nearest neighbors of a data point, taking the average or mode of their labels for classification or target values for regression. Key parameters include the choice of k values, distance metrics, and weighting techniques [7].

Random forestis an ensemble learning algorithm for classification and regression. Random Forest combines multiple decision trees generated from random data samples. The final prediction is based on the majority of predictions from all trees. Key parameters include the number of trees, tree depth, and randomly selected features [8].

Boosting Algorithm is an ensemble learning algorithm, such as AdaBoost, where models are created sequentially, learning from the mistakes of previous models to improve performance gradually. Key parameters include the number of models, learning rate, and tree depth in Gradient Boosting and XGBoost [8].

The selection of suitable machine learning models is followed by parameter tuning to find the optimal combination for optimal model performance.

III. METHODOLOGY

A. Data Set

The dataset, sourced from Kaggle, comprises 986 entries with six variables. A preview of the dataset is presented in Table 1.

TABLE 1
PREVIEW DATA OF MEDICAL PREMIUM PERSONAL DATASET

Age	Diabetes	Blood Pressure	Height	Weight	Premium
45	NO	NO	155	57	25000
60	YES	NO	180	73	29000
36	YES	YES	158	59	23000
52	YES	YES	183	93	28000
38	NO	NO	166	88	23000
30	NO	NO	160	69	23000

B. Data Processing,

Before applying machine learning analysis, the dataset undergoes preprocessing, including handling missing values, outlier checks, and variable transformations. Exploration Data Analysis (EDA) reveals a mix of numeric ('Age', 'Weight', 'Height', ‘Premium Price’) and categorical ('Diabetes', 'Blood Pressure Problems’) variables. To enable machine learning algorithms to process categorical data, we employ the following methods:

1. Label encoding. Label encoding refers to the process of converting word labels into numerical form, enabling algorithms to effectively process and analyze the data.

2. One hot encoding. A One hot encoding is a technique used to represent categorical variables as binary vectors, enabling a more expressive representation of categorical data. The process involves mapping categorical values to corresponding integer values through label encoding. Subsequently, each integer value is transformed into a binary vector where all elements are zero, except for the index corresponding to the integer value, which is indicated by a 1.
3. Dummy Variable Trap. The Dummy variable trap occurs when there is multicollinearity among the independent variables, meaning that two or more variables are highly correlated, to the extent that one variable can be accurately predicted from the others.

Train-Test-Split

The dataset is divided into training (80%) and testing (20%) sets, totaling 788 and 198 samples, respectively, allowing the model's performance evaluation on unseen data.

Standardization

Numerical data, specifically 'Age,' undergoes standardization, transforming it to a scale between 0 and 1.

Model Building

1. K-Nearest Neighbor (KNN): A non-parametric regression method with an efficient training phase and the capability to grasp intricate patterns in the target variable without making assumptions about data distribution [9].
2. Random Forest Regression: A meta-estimator constructing multiple decision trees, enhancing predictive accuracy through averaging and preventing overfitting [10].
3. Boosting Regression: An ensemble learning method combining weak learners to create a strong predictive model [11].

Model Evaluation

This research uses Mean Squared Error (MSE) and Mean Absolute Error (MAPE) to evaluate the model. The following equations are the formula for MSE and MAPE respectively [12][13].

$$MSE = \frac{\sum_{i=1}^{n}(A_t - F_t)^2}{n} \tag{1}$$

$$MAPE = \frac{\sum_{i=1}^{n}\left(\frac{A_t - F_t}{A_t}\right)}{n} \times 100\% \tag{2}$$

where A_t is the actual value, F_t is the predicted value, and n is the amount of data. Table 2 outlines the interpretation of MAPE values. [14]

TABLE 2
INTERPRETATION OF TYPICAL MAPE VALUES

MAPE	Intrepretation
<10	Highly accurate forecasting
10 – 20	Good forecasting
20 – 50	Reasonable forecasting
>50	Inaccurate forecasting

Model Validation

Comparison and analysis of statistical metrics, particularly Mean Absolute Percentage Error, provide insights into the accuracy and robustness of machine learning models. A smaller MAPE value indicates a more accurate prediction of premium prices [15][16].

IV. RESULT AND ANALYSIS

This section delves into a detailed analysis of the regression models' training and evaluation, aiming to provide a comprehensive understanding of the predictive performance.

The dataset is initially divided into 20% test data and 80% training data. During the training phase, an optimizer is utilized to fine-tune the model parameters, optimizing its performance.

Table 3 below summarizes the key performance metrics, including Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE), for each regression model:

TABLE 3
MEAN SQUARE ERROR (MSE) AND MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)

Method	**MSE**	**MAPE**
KNN	28173.13	12.26
RF	30325.82	13.00
Boosting	22508.41	12.81

- K-Nearest Neighbor (KNN): MSE: 28173.13 MAPE: 12.26. The KNN model demonstrates a moderate MSE and MAPE, indicating a reasonable level of accuracy in predicting premium prices. The model's performance suggests that it effectively captures patterns in the data, but there is room for improvement.
- Random Forest (RF): MSE: 30325.82 MAPE: 13.00. The Random Forest model exhibits slightly higher MSE and MAPE compared to KNN. While the model performs reasonably well, the elevated metrics suggest a need for further refinement to enhance predictive accuracy.
- Boosting: MSE: 22508.41 MAPE: 12.81. The Boosting model stands out with a lower MSE and competitive MAPE. This indicates that the Boosting algorithm is more effective in capturing the underlying patterns in the dataset, leading to a more accurate prediction of premium prices.

From the detailed analysis, it is evident that all three methods fall within the category of "Good Forecaster," as per the interpretation in Table 2. However, the KNN model stands out with the lowest MSE, suggesting superior predictive accuracy. These results provide valuable insights into the strengths and weaknesses of each regression method, guiding potential adjustments and improvements for future analyses.

In conclusion, while all three methods demonstrate a reasonable level of accuracy, the KNN algorithm emerges as the most effective model in predicting health insurance premium prices based on the given dataset.

V. CONCLUSION

In the ever-evolving landscape of health insurance, the integration of machine learning stands as a transformative force. Particularly adept at handling tasks traditionally executed at a slower pace by humans, artificial intelligence and machine learning technologies have demonstrated their prowess in analyzing vast datasets, thereby streamlining and simplifying health insurance operations. This not only promises time and cost savings for both policyholders and insurers but also positions AI to automate routine tasks, allowing insurance experts to focus on processes that enhance the overall policyholder experience.

The efficiency of machine learning in undertaking tasks currently carried out by humans holds significant advantages for patients, hospitals, physicians, and insurance providers. The speed and cost-effectiveness with which machine learning processes data present opportunities for improved service delivery and resource utilization. As a critical component of cognitive computing, machine learning addresses diverse challenges across various applications and systems, especially when leveraging historical data.

This study specifically investigates the application of three regression strategies—K-Nearest Neighbor (KNN), Random Forest, and Boosting—for predicting health insurance premium prices. The evaluation metric, Mean Absolute Percentage Error (MAPE), provides a nuanced understanding of the predictive accuracy of each method. The results are as follows: K-Nearest Neighbor (KNN): MAPE: 12.26%, Accuracy: 87.73%, Classification: Good forecasting level. Random Forest: MAPE: 13.00%, Accuracy: 87.00%, Classification: Good forecasting level, Boosting: MAPE: 12.81%, Accuracy: 87.19%, Classification: Good forecasting level.

The detailed comparison reveals that K-Nearest Neighbor emerges as the most accurate model, achieving an impressive accuracy rate of 87.73%. The other two methods, Random Forest and Boosting, also demonstrate commendable accuracy but slightly trail behind KNN.

The study's findings contribute valuable insights to the field of health insurance premium prediction. The demonstrated success of machine learning models, particularly K-Nearest Neighbor, opens avenues for further research and application in refining predictive modeling. Future studies may explore additional variables, fine-tune model parameters, and delve into more complex machine learning architectures to push the boundaries of forecasting accuracy.

In conclusion, the application of machine learning in health insurance premium prediction holds promise for improving the precision and efficiency of forecasting. As technology continues to evolve, ongoing research and development in this area will be crucial for advancing the capabilities and impact of machine learning in the healthcare industry.

REFERENCES

[1] B. Purkayastha, S. Panda, D. Das, & M. Chakraborty, "Health Insurance Cost Prediction Using Regression Models," 2022.

[2] A. Bharti & L. Malik, "Regression Analysis and Prediction of Medical Insurance Cost," *International Journal of Creative Research Thoughts*, vol. 10 no. 3, March 2022.

[3] Tejashvi, 2021,*"Medical Insurance Premium Prediction,"* [Online]. Available: https://www.kaggle.com/datasets/tejashvi14/medical-insurance-premium-prediction [Accessed October 7th, 2023]

[4] M. Johnny, B. Purwoko, & E.E. Merawaty, "Pengaruh Premi Bruto, Cadangan Klaim, Cadangan Premi, dan Pembayaran Klaim terhadap ROA (Suatu Survey pada Perusahaan Asuransi Umum Tercatat di BEI)," *Jurnal Ekbang,* vol. 3, no. 1, 2020.

[5] Allianz Indonesia, 4 Juni 2018, "Memaksimalkan Manfaat Lewat Deductibel," [Online]. Available: https://www.allianz.co.id/explore/memaksimalkan-manfaat-lewat-deductible.html [Accessed October 9th, 2023]

[6] M. Attaran, & P. Deb, "Machine Learning: The New 'Big Thing' for Competitive," *International Journal of Knowledge Engineering and Data Mining*, vol. 5, no. 4, pp. 277-305, January 2018.

[7] S. Uddin, I. Haque, H. Lu, M.A. Moni, & E. Gide, "Comparative Performance Analysis of K-Nearest Neighbor (KNN) Algorithm and Its Different Variants for Disease Prediction," *Scientific Reports,* 2022.

[8] S.K. Kiangala, & Z. Wang, "An Effective Adaptive Customization Framework for Small Manufacturing Plants Using Extreme Gradient Boosting-XGBoost and Random Forest Ensemble Learning Algorithms in an Industry 4.0 Environment," *Machine Learning with Applications*, vol. 4, June 2021.

[9] R. Goyal, P. Chandra, & Y. Singh, "Suitability of KNN Regression in the Development of Interaction based Software Fault Prediction Models," *International Conference on Future Software Engineering and Multimedia*, December 2014.

[10] "Scikit-Learn," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html [Accessed October 12th, 2023]

[11] I. D. Mienye & Y. Sun,, "A Survey of Ensemble Learning: Concepts, Algorithms, Application, and Prospects," in *IEEE Access*, vol. 10, pp. 99129-99149, September 2022.

[12] Khoiri, 16 Desember 2020, "Pengertian dan Cara Menghitung Mean Absolute Percentage Error (MAPE)," [Online]. Available: https://www.khoiri.com/2020/12/pengertian-dan-cara-menghitung-mean-absolute-percentage-error-mape.html [Accessed October 11th, 2023]

[13] F. Soufitri & E. Purwawijaya, "Analisis Kualitas Rancangan Point of Sale Menerapkan Metode Mean Squared Error," *Jurnal Media Informatika Budidarma*, vol. 6(4), October 2022.

[14] J.J.M. Moreno, A.P. Pol, A.S. Abad, & B.C. Blasco, "Using the R-MAPE Index as a Resistant Measure of Forecast Accuracy," *Psicothema*, 2013.

[15] F.H. Hamdanah & D. Fitrianah, "Analisis Performansi Algoritma Linear Regression dengan Generalized Linear Model untuk Prediksi Penjualan pada Usaha Mikro, Kecil, dan Menengah," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI),* vol. 10, no. 1, pp. 23-32, March 2021.

[16] A.N. Hanif, "Prediksi Biaya Asuransi Kesehatan," 2023, [Online]. Available: https://www.kaggle.com/code/alwannabilhanif/prediksi-biaya-asuransi-kesehatan [Accessed October 14th, 2023]