A Comparative Analysis of Regression Techniques in Predicting Medical Insurance Charges

Emil John Llanes
College of Computing and Information Technologies
National University
Manila, Philippines
Llanes@students.national-u.edu.ph

Danielle Joseph Octaviano
College of Computing and Information Technologies
National University
Manila, Philippines
octavianodp@students.national-u.edu.ph

Luis Ryan Sanisit
College of Computing and Information Technologies
National University
Manila, Philippines
sanisitld@students.national-u.edu.ph

***Abstract*—A growing concern for individuals and policymakers is the rising cost of medical insurance which signifies a growing need for accurate predictive models to predict medical insurance costs. In this study, the performance of four regression models were evaluated, namely Multiple Linear Regression (baseline), K-Nearest Neighbors, Support Vector Regression, and Random Forest Regression, to determine the most effective method for predicting medical insurance costs. An exhaustive feature selection and hyperparameter tuning with RandomizedCVSearch were applied to all models. The models are evaludated using 3-fold cross-validatoin based on Root Mean Squared Error (RMSE). The study reveals that Random Forest Regression achieved the best overall accuracy and consistency with a mean RMSE of 1758.64. KNN showed the highest accuracy but reduced reliability, while SVR performed the weakest among the models. Through the study, it was found out that ensemble based approach like Random Forest proved to be the best for capturing complex and nonlinear patterns such as medical insurance costs.**

***Index Terms*—medical insurance cost prediction, regression models, Random Forest, Support Vector Regression, K-Nearest Neigbors, RMSE**

I. Introduction

The rapid increase in healthcare insurance premiums across the world places a heavy financial strain on both individuals and families. Worldwide, health-related expenses account for about 10% of household income, and this percentage keeps rising as more families are compelled to pay for their own medical treatment, indicating a growing concern with cost control [1]. The ability to accurately forecast medical insurance costs is consequently essential for reducing financial risk and facilitating well-informed choices for both policyholders and insurers.

Although precise cost estimation is increasingly important, the complexity of health data continues to be a challenge for many predictive algorithms. This experiment tackles the problem of finding which regressor algorithm has the least Root-Mean-Square Error. Several regression-based predictive models, including Linear Regression as baseline of the models, Support Vector Regression, K-Nearest Neighbors, and Random Forest Regression, are applied and compared in the study to determine which model is the most accurate and reliable for anticipating insurance prices. This study emphasizes the comparative evaluation of different regression techniques to determine which performs best in predicting insurance price, in contrast to many previous studies that concentrate on a single model for convenience or efficiency. When dealing with high-dimensional nonlinear healthcare data, the precision is limited by the fact that existing models often only use one forecast strategy. This disparity affects policy makers, individuals, and insurers who rely on precise risk assessments and pricing. The suggested comparative framework can be used to improve sustainability and equity in healthcare budgeting systems, actuarial analysis, and insurance premium pricing.

The significance of this problem lies in its wide-ranging impact on economic stability and healthcare accessibility. Inaccurate cost predictions can lead to mispriced insurance premiums, unexpected out-of-pocket expenses, and financial stress for policyholders. For insurers, poor estimation models can result in inefficient risk management and unsustainable pricing strategies. By using AI and machine learning to improve predictive accuracy, this study contributes to better financial planning, more equitable premium structures, and enhanced policy design. The findings can be valuable to insurance companies, policymakers, and data scientists seeking to optimize cost prediction systems and strengthen the financial resilience of healthcare markets [2].

II. LITERATURE REVIEW

Regression analysis is a statistical method to find out the relationships between variables. It is used to investigate the causal effect of one variable upon another. In Machine

Learning (ML), regression is one of the core methods for predicting continuous values based on a given data pattern.

As discussed by [3], predictive methods of Machine Learning have become the most used tool for industrial applications like banking and real-estate fields, where data driven decision making is critical. Similarly in healthcare, there is a growing interest to use ML approaches like regression models to predict insurance costs and improve policies. Predicting medical insurance cost using Machine Learning techniques like Regression helps solve issues like accountability and transparency in healthcare expenditure. According to [4], one of the reasons for high medical costs is low accountability resulting from unnecessary medical procedures or use of medications. By taking advantage of regression based ML models, healthcare providers can better understand cost determinants, detect anomalies, and promote more data-informed financial and clinical decisions.

Several regression techniques are used for medical insurance cost prediction. One fundamental approach is Linear Regression. However, more advanced techniques are required when relationships between data become non-linear [5]. More complex approaches like Decision Tree, Random Forest, Support Vector, and Gradient Boosting Regression have been shown to provide better predictive performance in such cases [6].

Recent studies demonstrated significant progress in predicting health insurance costs using both statistical and machine learning approaches. The three identified key factors that contributes to the increase of health insurance through various analytical methods such as correlation analysis, k-means clustering, principal component analysis, multiple regression, lasso regression, support vector machine, and random forest are BMI, blood pressure, and smoking [7]. Machine learning models that include polynomial, decision tree, random forest, and gradient boosting has shown better performance compared to traditional methods of calculations evaluating using MSE, RMSE, and R-Squared metrics [8]. On another hand eXtreme Gradient Boosting (XGBoost) have been used and represents as being highly effective tree boosting system used by data scientists to achieve state-of-the-art results for these types of machine learning challenges [9]. These machine learning advances indicate potential to transform premium pricing strategies, resulting in a more customized and financially sustainable outcomes within healthcare insurance [2], [8].

Prior studies, tackled to predict health insurance costs with the use of machine learning. A journal made by Kaushik et. al. published from Environmental Research and Public Health, worked on an artificial neural network-based regression model on health insurance premiums achieved an accuracy of 97.72%; with age, gender, body mass index, number of children, smoking habits, and geolocation used as parameters [2]. Additionally, Researchers' Devi et. al. from the Institute of Technology Chennai experimented on both linear and ensemble regression models. Results found that polynomial regression achieved 88% R2Score while Random Forest Regression achieved 86% R2Score. On top of that, the researchers formed the idea that the variable region has little influence on healthcare cost, suggesting that the geographic location is not a useful predictor within their model [6]. Both Kaushik et. al. and Devi et. al. approach on using age, gender, body mass index, smoking habits, number of children, and geolocation shows consistency suggesting that these features are recognized as being an important factor for predicting healthcare costs [2], [6]. Whilst the idea of using a single model to find which factors better affect healthcare costs, Researchers from National Institute of Technology Silchar developed a comprehensive real-time healthcare insurance cost prediction system called ML Health Insurance Prediction System (MLHIPS) integrating the idea of machine learning to a user friendly application. The researchers incorporated multiple regression models, resulting in a R2Score of 0.80% on their best model polynomial regression [5]. Although a number of papers use varying evaluation metrics, there are only a handful of studies in which comparison and analysis is possible. Some report R-squared values, others' focal point is on accuracy percentages. Resulting on there not being a standardized benchmark for model performance. Our study seeks to improve this challenge by compiling regression models and choosing the best possible parameters for each model then comparing them to each other to find the most optimal model with the least Mean Squared Error (MSE).

III. Methodology

The development and implementation of the project is divided into 4 phases: data collection, data pre-processing, model training, and model evaluation. This structure follows the standard Machine Learning modeling pipeline ensuring a systematic workflow. Each phase is further discussed in the next section.

A. Data Collection

The dataset used is collected from Kaggle, a popular data science community. It comprises 2772 rows and seven (7) columns: age, sex, body mass index (BMI), children, smoking status, region, and annual insurance prices. Although the specific data source was not explicitly mentioned in the Kaggle page, it was understood that the dataset is designed primarily for education and research purposes.

B. Data Pre-Processing

To determine consistency throughout the training process the dataset is cleaned to achieve the goal of the experiment. The following features; smoker and age has blank values indicating (**?**) labeling them as null variables, classifying them by;

- Within the smoker feature was filled in by getting the mode of the whole smoker parameter. The mode of the whole parameter was used due to assumptions that the amount of smokers was more than the amount of non-smokers.

- As per the null variables within age feature, it was filled out by getting the mean of the whole parameter. Making the assumption of the range of the null variables would be the average of the age parameter.

C. Experimental Setup

The setup used for the experiment is Google Colab for the software running the training process for the regression models. For the training process to work as intended a cell was first initialized containing the primary libraries; pandas 2.2.2, numpy 2.0.2, seaborn 0.13.2, matplotlib 3.10.0, and the built in module itertools within python 3.12.12.

Scatter plots and correlation heatmaps were used to further identify which features has a high relationship to the change of insurance prices. Upon the completion of the pre-processing step, the data was standardized and prepared for model training. The dataset was divided into training and testing sets using an 80–20 split, where 2,218 samples were used to train the models and 554 samples were reserved for testing. Each subset contained the same seven features—age, gender, bmi, children, smoker, region, and insurance price-ensuring consistent feature representation across both sets for accurate model evaluation.

In order to determine the most optimal model, exhaustive feature tuning was conducted, a process that systematically evaluates all possible combinations of features for each regression model. During this process, every configuration was tested and assesed using Root Mean Squared Error (RMSE) to identify the combination that produced the best predictive performance. This was to ensure that the final selected model achieved the most accurate result possible within the dataset.

D. Algorithm

Support Vector Regression (SVR)

Support Vector Regression (SVR) is a supervised learning method derived from Support Vector machines (SVM), that is capable of predicting continuous values. SVR was chosen because it models linear and non-linear relationships effectively while remaining robust to noise and variations in the data through the use of a $\varepsilon - insensitive$ loss function [10], [11]. In SVR, the regression function is represented as a hyperplane in an n-dimensional feature space [6]:

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b \tag{1}$$

Where $\mathbf{w}$ is the weight vector, $\mathbf{x}$ is the input feature vector, and $\mathbf{b}$ is the bias term. Around this hyperplane is the $\varepsilon - insensitive$ tube or loss function which can be represented as follows:

$$|y_i - f(\mathbf{x}_i)| \leq \varepsilon \tag{2}$$

Errors within the tube are ignored, and only variations exceeding the tube are penalized [11].

K-Nearest Neighbors (KNN)

A particular kind of non-parametric method is K-Nearest Neighbors. utilized to forecast a fresh data point's ongoing result by looking at its closest neighbors' outcomes in the feature room. This method works especially well with complex data. structures because it doesn't assume anything about In contrast to Linear Regression, the underlying data relationships as well as Support Vector Regression. But this approach is prone to overfitting, which is a significant drawback [12]. The Euclidean distance is usually the most generally utilized metric to evaluate how close data points are. The expected result is calculated by taking the average of the outcomes of the k-nearest neighbors. This can be expressed mathematically as [12]:

$$\hat{y} = \frac{1}{k} \sum_{i \in N_k} y_i \tag{3}$$

Where:

- $\hat{y}$ denotes the predicted value,
- k is the number of nearest neighbours,
- $N_k(x)$ represents the set of k closest training samples to x, and
- y_i is the actual target value of the i-th neighbour.

Random Forest

Random forest is a supervised learning technique by combining multiple decision trees to model the complex interactions and nonlinear relationships in the data. It was described as the most successful general-purpose method in modern times [13]. It makes a prediction by averaging the result of its ensemble of decision trees to make a more reliable and accurate model. Mathematically, this is represented as [14]:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{4}$$

Where:

- $\hat{y}$ is the final predicted value,
- B is the total number of decision trees in the forest, and
- $T_b(x)$ is the prediction from the b-th individual tree.

E. Training Procedure

Each regression model was trained and optimized using a systematic search strategy by combining exhausting search of feature subset and hyperparameter tuning. The exhaustive feature search works by finding the most performing combination of features over all possible combinations.The dataset was divided into training and testing subsets. For each feature subset in the training data, RandomizedSearchCV was used to optimize model parameters with 3 fold cross validation to efficiently explore and find the best configuration.

F. Evaluation Metrics

The metrics used to assess and evaluate the model is Root Mean Squared Error (RMSE). RMSE assesses the average magnitude of error in a regression or statistical model in which a low predicted value is more precise than a higher predicted value [15], [16]. Penalizing larger errors more heavily, effective for identifying models that perform well consistently and avoiding large deviations of values. Evaluating models with the use of RMSE has been widely regarded as a standard within statistical and machine learning models [17], especially in fields like healthcare insurance prediction, in which large cost deviations have significant real-world consequences.

The results were measured and compared solely using RMSE to evaluate the predictive accuracy and performance for each regression model. The baseline model for all the model evaluations is dictated by the best possible score using Multiple Linear Regression.

G. Baselines and Comparative Models

The selected baseline model was Multiple Linear Regression (MLR) because of its simplicity and ease of use. It is a suitable benchmark for assessing more complex regression models as it assumes a linear relationship between the data. To assess improvements over the baseline model, three more models mainly, Support Vector Regression (SVR), Random Forest Regression, and K-Nearest Neighbours (KNN) were trained using the same dataset and preprocessing steps. Each model underwent 3 fold cross validation, with Root Mean Squared Error (RMSE) as the primary performance metrics. The comparison of the models were further discussed and analyzed in the Results and Discussion section.

IV. Results and Discussion

This section presents the experimental results and analysis of the different regression models used in predicting healthcare insurance costs. The models were evaluated using Root Mean Squared Error (RMSE) to determine each of the models' predictive performance and accuracy. The findings discussed is in relation to the study's objective, highlighting the comparison of the different types of regression models; KNN, SVM, Random Forest, and Multiple Linear Regression as the baseline comparator.

A. One-shot Performance

Table 1. Performance Summary of Regression Models

Model	Feature Combination	Hyperparameters	RMSE
Multiple Linear Regression (Baseline)	age, bmi, children, smoker, region	positive=False, fit_intercept=True	6322.1611
KNN	gender, bmi, children, smoker	weights=distance, p=1, n_neighbors=5	4272.0351
SVR	smoker	kernel=rbf, gamma=scale, epsilon=0.1, C=10	10483.4739
Random Forest	age, bmi, children, smoker, region	n_estimators=300, min_samples_split=2, min_samples_leaf=1, max_features=log2, max_depth=20	2694.7542

Table 1 consists of the best possible feature set and hyperparameters for each of the regression models used during the experiment. The table consists of the feature combination, hyperparameters and the lowest possible RMSE from each model. Observing the data from within the table the model with the lowest RMSE score is Random Forest with 2694, being followed by KNN with 4272, which in contrast SVR coming last with 10483 above our baseline (Multiple Linear Regression) with 6322.

B. Cross-Validation Performance

Table 2. Cross-Validation Results of Regression Models

Model	Mean RMSE	Std RMSE
KNN	411.04	307.18
Multiple Linear Regression	6077.12	130.89
SVR	12802.75	368.67
Random Forest	1758.64	141.71

Table 2 indicates cross-validation performance of each regression model. Upon observing the result, Random Forest achieves the lowest mean RMSE (1758.64) with a low standard deviation (141.71), demonstrating high and stable performance. Meanwhile, KNN shows a lower mean RMSE (411.04) with higher standard deviation (307.18) compared to Linear Regression's mean RMSE (6077.12) with lower standard deviation (130.89). With SVR's performance being the worst of the bunch with a mean RMSE of 12802.75 and standard deviation of 368.67.

C. Model Comparison

Table 3. Cross-Validation Performance of Regression Models vs Linear Regression

Model	Mean RMSE	Std RMSE	% Improvement vs Baseline
KNN	411.04	307.18	93%
Linear Regression (Baseline)	6077.12	130.89	–
SVR	12802.75	368.67	-111%
Random Forest	1758.64	141.71	71%

Table 3 presents the cross-validation performance of each regression model compared to the Linear Regression baseline. Random Forest achieves a mean RMSE of 1758.64 with a low standard deviation of 141.71, showing both high accuracy and stable performance across folds. KNeighborsRegressor has a lower mean RMSE of 411.04 but a higher standard deviation of 307.18, indicating more variability between folds despite its strong average performance, resulting in the highest percentage improvement over baseline at 93%. Linear Regression serves as the baseline with a mean RMSE of 6077.12 and a low standard deviation of 130.89. SVR performs worst, with a mean RMSE of 12802.75, standard deviation of 368.67, and a negative percentage improvement of –111%, reflecting consistently poor performance relative to the baseline.

D. Statistical Significance

Table 4. Pairwise Statistical Significance of Models

Model 1	Model 2	p-value	Significant ($p \leq 0.05$)?
KNN	LinearRegression	0.0000	Yes
KNN	SVR	0.0000	Yes
KNN	RandomForestRegressor	0.0001	Yes
LinearRegression	SVR	0.0000	Yes
LinearRegression	RandomForestRegressor	0.0000	Yes
SVR	RandomForestRegressor	0.0000	Yes

Table 4 shows the pairwise statistical significance of the models based on cross-validation RMSEs. A key concern in model comparison studies is that researchers present estimates of model performance with little evidence on whether they reflect true differences in model performance [18]. The use of paired t-tests with cross validation addresses statistical evidence to determine the model performance differences are genuine rather than random variation. After determining the p-values of each model against each other we can determine that the predictive performance of each one are significant.

In summary, Random Forest is the most reliable and effective model, providing both high accuracy and consistent performance across folds. KNN can achieve very low errors under certain conditions, but its high variability reduces overall reliability. Linear Regression delivers stable predictions, though its accuracy is limited by the assumption of linearity. SVR consistently underperforms and fails to generalize effectively. These findings underscore the importance of evaluating models not solely on single-test results, but also using cross-validation metrics to account for both predictive accuracy and stability.

E. Experimental Constraints

The experimental constraints encountered throughout the model training include; limited size of the dataset, which may not fully represent the diversity of healthcare insurance data across different regions or demographic groups. This limitation does not generalize the models to broader populations. Additionally, model sensitivity posed a challenge, although the study focused on identifying the best possible feature combination for each model, limitations within computational resources and methodological scope restricted the extent of hyperparameter optimization. Furthermore the study did not incorporate deeper statistical validation methods, such as extended significance testing or residual analysis, which could have strengthened the reliability of the findings.

F. Conclusion

A comparative evaluation of multiple regression algorithms namely Multiple Linear Regression, K-Nearest Neighbors, Support Vector Regression, and Random Forest Regression, for predicting medical insurance premium or cost is conducted in this study. By using cross-validation and RMSE as evaluation metrics, the results revealed that Random Forest has the most balanced and consistent performance with a mean RMSE of 1758.64 with a low standard deviation of 141.71. While K-Nearest Neighbors achieved the lowest RMSE on some folds, it has a higher standard deviation indicating reduced reliability. Conversely, Linear Regression, while revealing a stable performance, it lacked the flexibility to accurately capture complex datasets. Finally, Support Vector machine showed the weakest performance overall.

Throughout the study, it was shown that ensemble-based methods like Random Forest can effectively improve accuracy in medical insurance cost prediction. This study aims to contribute to a more accurate insurance pricing, better financial risk management, and data-driven decision-making for insurers and policymakers.

Future research may extend this work by incorporating feature importance analysis, additional ensemble methods, and a more diverse datasets to further improve the prediction performance and generalization.

References

[1] G. I. Alobo, “Perspective chapter: Health insurance across worldwide health systems – why it matters now,” in *Health Insurance Across Worldwide Health Systems*, A. I. Tavares, Ed. London: IntechOpen, 2024, ch. 8. [Online]. Available: https://doi.org/10.5772/intechopen.1003031

[2] 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, 2022.

[3] U. E. Orji, C. H. Ugwuishiwu, J. C. N. Nguemaleu, and P. N. Ugwuanyi, “Machine learning models for predicting bank loan eligibility,” in *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*, 2022, pp. 1–5.

[4] B. Panay, N. Baloian, J. A. Pino, S. Peñafiel, H. Sanson, and N. Bersano, “Predicting health care costs using evidence regression,” in *Proceedings*, vol. 31, no. 1. MDPI, 2019, p. 74.

[5] S. Panda, B. Purkayastha, D. Das, M. Chakraborty, and S. K. Biswas, “Health insurance cost prediction using regression models,” in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, vol. 1, 2022, pp. 168–173.

[6] M. Shyamala Devi, P. Swathi, M. Purushotham Reddy, V. Deepak Varma, A. Praveen Kumar Reddy, S. Vivekanandan, and P. Moorthy, “Linear and ensembling regression based health cost insurance prediction using machine learning,” in *Smart Computing Techniques and Applications: Proceedings of the Fourth International Conference on Smart Computing and Informatics, Volume 2*. Springer, 2021, pp. 495–503.

[7] S. Wen, “Health insurance claim amount prediction using statistical methods,” in *Proceedings of the 2nd International Conference on Software Engineering and Machine Learning*, vol. 73, 2024.

[8] R. Ejjami, “Machine learning approaches for insurance pricing: a case study of public liability coverage in morocco,” *International Journal for Multidisciplinary Research (IJFMR)*, pp. 1–23, 2024.

[9] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[10] M. Awad and R. Khanna, “Support vector regression,” in *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*. Springer, 2015, pp. 67–80.

[11] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, “Machine learning for internet of things data analysis: A survey,” *Digital Communications and Networks*, vol. 4, no. 3, pp. 161–175, 2018.

[12] P. Srisuradetchai and K. Suksrikran, “Random kernel k-nearest neighbors regression,” *Frontiers in big Data*, vol. 7, p. 1402384, 2024.

[13] D. Borup, B. J. Christensen, N. S. Mühlbach, and M. S. Nielsen, “Targeting predictors in random forest regression,” *International Journal of Forecasting*, vol. 39, no. 2, pp. 841–868, 2023.

[14] Z. Mustaffa and M. H. Sulaiman, “Random forest based wind power prediction method for sustainable energy system,” *Cleaner Energy Systems*, p. 100210, 2025.

[15] J. Frost. (n.d.) Root mean square error (rmse). Accessed: 2025-10-15. [Online]. Available: https://statisticsbyjim.com/regression/root-mean-square-error-rmse/

[16] A. Jierula, S. Wang, T.-M. Oh, and P. Wang, “Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data,” *applied sciences*, vol. 11, no. 5, p. 2314, 2021.

[17] T. O. Hodson, “Root mean square error (rmse) or mean absolute error (mae): When to use them or not,” *Geoscientific Model Development Discussions*, vol. 2022, pp. 1–10, 2022.

[18] J. B. Nasejje, A. Whata, and C. Chimedza, “Statistical approaches to identifying significant differences in predictive performance between machine learning and classical statistical models for survival data,” *PLOS ONE*, vol. 17, no. 12, p. e0279435, 2022. [Online]. Available: https://doi.org/10.1371/journal.pone.0279435