

Health Insurance Recommendation System using Optimized Grid Search and Regression Models

Kaladevi R^{1*}, Uma Rani V², Senthamizh Selvi S³, Dilli Babu M⁴, Revathi A⁵

^{1*}Professor, AI &DS Department, Saveetha Engineering College, Chennai, Tamil Nadu, India.

^{1*}Email: kalaramar26@gmail.com

²Associate Professor, Computer Science and Engineering Department, Saveetha Engineering College, Chennai, Tamil Nadu, India.

³Associate Professor, Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Chennai, Tamil Nadu, India.

⁴Associate Professor, IT Department, Panimalar Engineering College, Chennai, Tamil Nadu, India.

⁵Associate Professor, Department of Computational Intelligence, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.

ARTICLE INFO

ABSTRACT

Received: 30 Sept 2024

Revised: 29 Nov 2024

Accepted: 10 Dec 2024

Introduction: Health insurance schemes help cover medical expenses by distributing financial risk among many individuals. With various insurance options available, choosing the right provider and predicting costs can be challenging. Predictive modeling and machine learning techniques play a important role in analyzing past data, identifying patterns in customer behavior, and supporting informed decision-making for new insurance plans.

Objectives: The main aim of this research is to assist individuals in selecting appropriate medical insurance providers and estimating associated costs using predictive models. By leveraging historical data, the study seeks to improve cost prediction accuracy and enhance decision-making in the health insurance sector.

Methods: This study utilizes medical provider datasets along with cost prediction data to develop predictive models. A total of 12 regression classifiers are applied to analyze the data. To optimize performance, Grid Search Cross-Validation is used for fine-tuning the models. This ensures better accuracy and reliability in predicting insurance costs.

Results: From analysis, X-Gradient Boost, Random Forest, and Extra Trees models demonstrated the highest accuracy. These models achieved R^2 scores greater than 98%, indicating their effectiveness in capturing the relationship between input features and insurance costs. The Extra Trees model perform well with an R^2 score of 0.99 during training and 0.88 during testing. Additionally, these models provide low Root Mean Squared Error (RMSE) values, confirming their reliability in making precise predictions.

Conclusions: The findings suggest that machine learning models, especially tree-based regressors like Extra Trees, X-Gradient Boost, and Random Forest, can effectively predict medical insurance costs with high accuracy. By leveraging predictive modeling, both insurance providers and customers can make informed decisions regarding cost estimation and plan selection.

Keywords: Health insurance, Grid search, hyper parameter, Regression, Classifiers.

INTRODUCTION

Increasing population and pollution have resulted in greater exposure to hazards. Modern lifestyles, longer working hours, and extended life expectancy contribute to both chronic and unexpected medical expenses [1]. To manage these risks, insurance becomes essential, helping people cover costs related to accidents, diseases, and other unexpected events [2]. Insurance is broadly classified into two categories: general insurance and health insurance. General insurance covers various risks, such as vehicle accidents, property damage, natural disasters (like floods and earthquakes), and theft [3, 4]. Health insurance, on the other hand, focuses on covering medical expenses, providing individuals and their families with financial protection in case of illness or injury [5].

As healthcare costs continue to rise and non-communicable diseases become more prevalent, health insurance is increasingly seen as a necessity [6]. The COVID-19 pandemic underscored the unpredictability of life and the critical need for comprehensive health coverage [7]. Health insurance helps reduce the financial burden of medical treatments, especially as life expectancy increases. Global public health spending reached \$8.5 trillion in 2019, accounting for nearly 10% of global GDP [8]. In India, the insurance industry consists of 57 companies, 24 of which provide life insurance. Notable players include public sector giants like LIC and New India, alongside private companies like ICICI, HDFC, and SBI [9]. These firms compete in providing life and non-life insurance products to the population.

Studies indicate that Medicare enrollees are generally more satisfied with their insurance coverage than those with commercial insurance plans. Those newly covered by Medicare report higher satisfaction with care. However, comparisons between Medicaid and commercial insurance have yielded mixed results, with research often focusing on specific populations and services [10]. Predicting health insurance costs is vital for both individuals and the healthcare system. Accurate predictions help insurance companies set premiums and manage healthcare expenses more effectively. Machine learning techniques such as Lasso, Ridge, Random Forest and Decision Tree are widely used for cost prediction. Techniques like Grid Search Cross Validation optimize model performance, ensuring accurate and efficient predictions for better decision-making.

OBJECTIVES

The primary goal of this research is to design a health insurance recommendation system that helps individuals choose the most suitable insurance policy by providing provider details and cost estimates. The system aims to simplify decision-making by offering data-driven insights into different health insurance plans.

To achieve this, the study explores various regression models to predict health insurance costs accurately. By analyzing past data, these models can forecast potential expenses, allowing users to make informed choices based on estimated costs. The research further evaluates the performance of multiple regression algorithms to determine the most effective model for integration into the recommendation system.

To enhance the accuracy and efficiency of the predictive models, the study employs GridSearchCV, a hyperparameter tuning technique that optimizes model parameters. This fine-tuning process ensures that the selected model delivers the best possible predictions, improving the reliability of the recommendation system. Ultimately, this approach helps both insurance providers and consumers make better, data-driven decisions regarding health insurance policies.

BACKGROUND

Universal health coverage -UHC assures that public can use quality medical services without financial crisis. This includes health-related services such as prevention, promotion, treatment, palliative care, and rehabilitation [11]. Authors are chosen from 68 potential research papers by 118 researchers to analyze the impact of utilizing medical insurance policies across various countries. According to the supply-demand equation, these insurance policies reduce the cost of healthcare services, resulting in increased demand [12]. This chapter reviews the gap on the relationship between medical insurances, the use of health care services, and health benefits for particular conditions and practices, along with health condition and mortality [13]. It is known that a consistent, Significant correlation among medical insurance coverage and outcomes from health-related services from numerous studies using diverse data portals and systematic approaches. The evidence also suggests that medical insurance is related with proper implementation of healthcare services which leads to good health outcomes for adults. It focuses on primary prevention, cancer care, chronic disease management, hospital-based care, general medicine, and mortality.

Madan Mohan analysed the contribution of medical insurance to the development of general insurance in India [14]. He used predictive analysis to find a relation between dependent terms such as profit or loss and the independent terms, i.e., medical insurance premiums. Research results show a correlation between earned premium and underwriting loss. Li and Dou explored the impact of health insurance for urban and remote residents and basic health insurance for city employees on the practices of essential public health services, also reviewing the mediating effect of social integration [15]. They collected data from 169,989 Chinese individuals in 2017. The authors implemented the Bootstrap method from structural equation modelling in order to evaluate the intermediate role of social integration.

A health insurance awareness review was conducted by the authors in India [16]. For the review, resources were taken from Scopus, MEDLINE, Web of Science (WoS), Social Science Research portal, and the gie development portal, covering the duration from Jan (2010) to July (2020). Official web portals and related references were also used. The aim of this study is to identify obstacles experienced in India to elevate awareness of medical insurance and to provide evidence for the usefulness of mitigating such obstacles on the awareness and promotion of medical insurance for the native Indian people. Kaushik et al. proposed a ML-based health insurance premium calculation method [17]. They were predicted the cost associated with health insurance incurred by patients based on the parameters age, male or female, BMI (body mass index), number of children, geolocation and smoking habits. An artificial neural network (ANN) model was used and analysed, with evaluation results showing with an accuracy of 92.72%. The authors evaluated their system's performance with key performance measures.

Sahu et al. proposed a machine learning-based approach to predict medical costs [18]. It helps policymakers identify policy providers with premiums. The Random Forest Regression algorithm is applied to predict medical expenses. They also tested the issue with other ML models such as linear regression and Gradient Boosted Trees. Their outcomes suggest that the general public can estimate treatment costs.

Kulkarni et al. implemented a computational intelligence approach using ML algorithms to predict medical insurance costs [19]. They used a dataset from Kaggle and regression algorithms such as linear regression, decision tree and Gradient Boosting Regression, with Streamlit as a framework. They obtained an R^2 score for linear regression of 75%, decision tree regression of 69%, and Gradient Boosting Regression of 86.9%. Bhatia et al. used the dataset of USA's medical cost personal from Kaggle, with 1,338 tuples [20]. Notable features present in the dataset are gender, age, BMI, No.of.Children, smoking habits for predicting insurance costs. The authors applied linear regression model and identified the relation between price and the given features. For training and testing, their system used a 70-30 split, achieving an accuracy of 81.3%.

The proposed health recommendation system applies Support Vector Regression (SVR), Linear Regression, Decision tree, Multiple Linear Regression, Stochastic Gradient Boosting, Ridge Regressor, XGBoost, Random Forest Regressor and k-Nearest Neighbors [21]. A health insurance cost dataset is obtained from the Kaggle data repository, and ML models are used to project how various regression methods can forecast insurance costs and analyse the models' performance. The outcomes proven that the Stochastic Gradient Boosting (SGB) method outperforms the others regressors. SGB has a cross-validation score of 0.858, an RMSE score of 0.340, and the accuracy is 86%. This research demonstrates the performance of different regression models to predict insurance costs [22]. The results of methods like Generalized Additive Model, Multiple Linear Regression, Random Forest Regressor, k-Nearest Neighbors, Support Vector Machine, Random Forest Regressor, CART, XGBoost, Deep Neural Network and Stochastic Gradient Boosting are compared. The optimum result is obtained from Stochastic Gradient Boosting method with an MAE value of 0.17448, RMSE score of 0.38018, and R-squared score of 85.8295.

The aforementioned researchers analysed the importance and impact of health insurance and predicted insurance costs using various datasets and machine learning algorithms. However, a research gap remains in identifying, normalizing the proper dataset, and applying more regression algorithms to predict accurate premiums for medical insurance.

METHODS

This research work attempts to calculate the premium for health insurance schemes with the help of regression algorithms. Regression algorithms are a subset of ML algorithms which are mainly used for predictive purposes. These models are used to find the correlation between goal (dependent variable) to the independent variables.

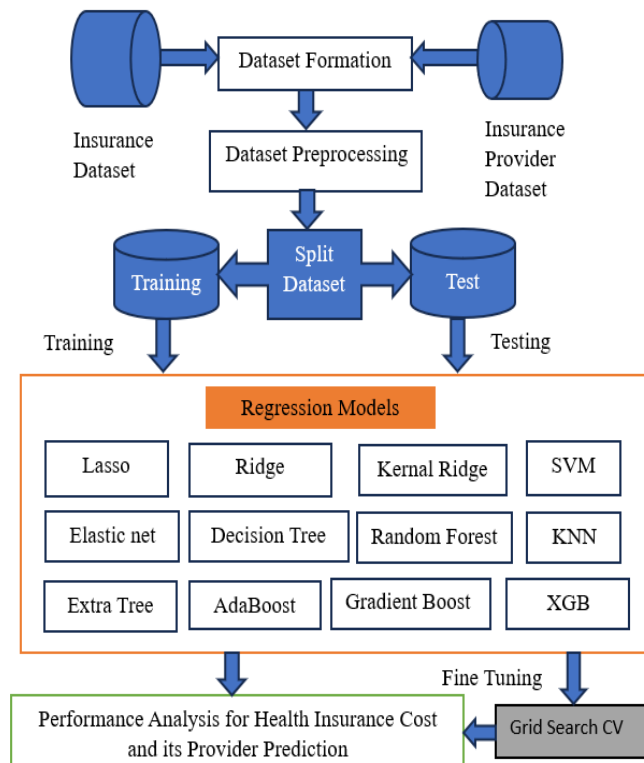


Figure 1. Prediction flow of Health Insurance Recommendation System

The proposed approach considers various regression algorithms such as Lasso, Ridge, Kernel Ridge, SVM, Elastic Net, Decision Tree, Random Forest, KNN, Extra tree, AdaBoost, Gradient Boost and XGB for identifying effective prediction strategies. The entire process flow is shown in Figure.1.

DATA SET FORMATION

In the initial phase data are collected from UCI repositories with 8 healthcare attributes and 1338 tuples in the collection. The first five rows of health care attributes Age, gender, BMI, number of children, smokers, and costs in medical insurance are shown in Figure.2.

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Figure 2. Five rows of Health Insurance Cost Data set

Another data set which comprises information like insurance provider, policy name, police type, estimated amount per year and claim amount which is depicted in Fig.3. The histogram in Fig.4 shows the distribution of attributes in the health insurance cost data set.

	Provider	Payment	Claim
0	Care supreme	7111	500000
1	Care supreme	7343	700000
2	Care supreme	8107	1000000
3	Niva Re-assure	6802	500000
4	Star Health Assuure	7675	500000

Figure 3. Five rows of Health Insurance Provider Dataset

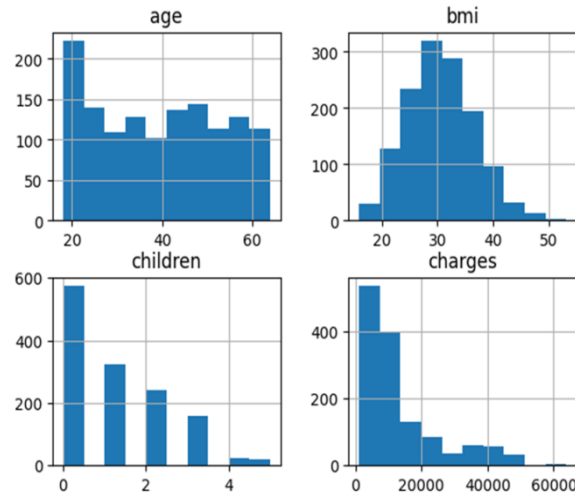


Figure 4. Histogram of attributes in health insurance cost data set

DATA PREPROCESSING

Data preprocessing is the necessary step to clean and structuring data before training. It is a way of formatting a huge volume of data after removal of null attributes, null values and unwanted characters. In this insurance cost prediction dataset, the null values and null attributes are removed by eliminating the corresponding rows and the parameters BMI, age and charges are normalized by standard scalar mechanism. Standard scalar normalization helps to remove mean and scaling features to unit variance. The standard value of sample x is measured by equation(1).

$$Z = \frac{X - \mu(X)}{\sigma(X)} \quad (1)$$

REGRESSION MODELS

Regression models are useful for predicting continuous data. In this research, 12 popular regression models are used for medical insurance cost prediction. Each model has unique characteristics that make it suitable for various predictive tasks. These models are discussed below [23, 24], along with their fundamental regression formulas.

Lasso Regression

Lasso Regression-Least Absolute Shrinkage and Selection Operator, relies on shrinkage, which pulls coefficients toward the mean. This regularization technique is highly effective in models with fewer parameters, improving prediction accuracy. The Lasso regression formula is defined by equation (2).

$$\min_{\beta_0, \beta} \left\{ \frac{1}{2n} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (2)$$

Ridge Regression

Ridge Regression, or Tikhonov regularization, is useful for models with multicollinearity, where independent variables are highly correlated. It addresses this issue by penalizing the size of coefficients. The Ridge regression is shown by equation (3).

$$\min_{\beta_0, \beta} \left\{ \frac{1}{2n} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\} \quad (3)$$

Support Vector Regression and Elastic Net

SVR, a variation of SVM, models complex relationships between input and target variables using a kernel function. Elastic Net combines L1 (Lasso) and L2 (Ridge) penalties for regularization, making it applicable for datasets with related features. Its formula is defined by equation (4).

$$\min_{\beta_0, \beta} \left\{ \frac{1}{2n} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \left(\alpha \sum_{j=1}^p |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^p \beta_j^2 \right) \right\} \quad (4)$$

Random Forest and Decision Trees

Random Forest (RF) is an ensemble learning method which creates many decision trees and averages their decision values to improve accuracy. The formula for RF is defined by equation.(5).

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B T_b(x) \quad (5)$$

Decision Trees divide the data set based on rules, making them useful for both classification and regression tasks. The expected value for decision trees is calculated by equation (6).

$$EV = (\text{First outcome} \times \text{likelihood}) + (\text{First outcome} \times \text{likelihood}) - \text{Cost} \quad (6)$$

KNN and Extra Trees

KNearest Neighbors (KNN) regression approximates the relationship between independent variables and outcomes by averaging the values of the k-nearest data points. The distance between points are computed by using eq.(7).

$$\text{dist}(x, z) = \left(\sum_{r=1}^d |x_r - z_r|^p \right)^{\frac{1}{p}} \quad (7)$$

Extra Trees also predict results by aggregating decisions from multiple decision trees, but they differ from Random Forest by splitting nodes randomly. The formula is given in eq.(8).

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N h_i(x) \quad (8)$$

Ada Boost and Gradient Boost

AdaBoost combines weak classifiers into a stronger one by focusing on misclassified samples, assigning greater weight to difficult cases. The formula is given in eq(9).

$$\hat{f}(x) = \sum_{t=1}^T \alpha_t h_t(x) \quad (9)$$

Gradient Boosting model develops a strong model by orderly combining weak models, with the consequent model learning from the errors of the previous one. Its formula is:

$$\hat{f}(x) = \sum_{t=1}^T \gamma h_t(x) \quad (10)$$

XGBoost

XGBoost is an optimized gradient boosting algorithm that combines the predictions of multiple base learners, typically decision trees, into a final prediction. Regularization helps prevent overfitting. The formula is given by eq(11).

$$\hat{y}_i = \sum_{k=1}^K f_k(x(i)) \quad (11)$$

Cross Validation

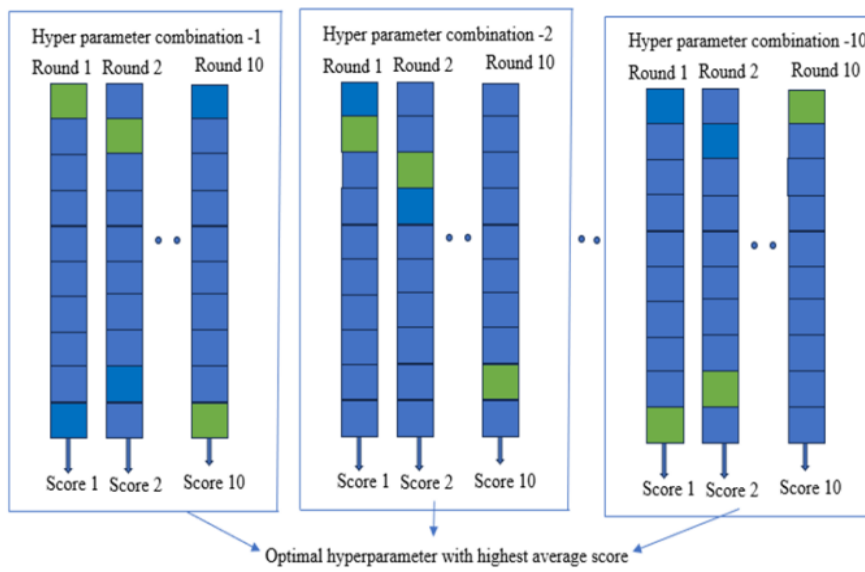


Figure 5. Hyper-parameter selection in Grid Search CV

Cross validation is one of the procedures to check the model is fitted for unseen data. Generally, it can be carried out by holdout or k-fold cross validation. K-fold cross-validation is a technique for evaluating predictive models and mitigate overfitting. The given dataset is splitter into k subsets or folds. The ML model is trained and tested k times, each time using a different subset for the validation purpose. Performance metrics is calculated as the average value derived from all folds and it is used to estimate the system's overall performance. This method helps in model selection, assessment and hyper parameter tuning, providing a more reliable measure of a model's effectiveness. Grid Search CV is a kind of k-fold cross validation which searches through multiple hyperparameters of the model and reports the optimal one. It divides the dataset into k folds and for every kth fold find the performance of the classifier. For each fold, it calculates the cross-validation score value for the fitted function. Finally, the hyperparameter with highest average score is selected to improve the accuracy of regression models. This is depicted in Figure 5. The cross validation score for fitted function is calculated by eq (12).

$$cvs(f) = \frac{1}{n} \sum_{i=1}^n \mathcal{T}(y_i, f^{k(i)}(x_i)) \quad (12)$$

where n is the size of the insurance dataset, k is the number of sub-folds, f is the fitted function and T is the loss function.

RESULTS AND DISCUSSION

For experimental results, the medical insurance cost data set is analyzed by using visual studio code along with Jupiter notebook. It uses sklearn packages for regression model and statistics analysis. It uses plotly and seaborn packages for graph generation. This section first describes the evaluation metrics taken for comparison. Then the results obtained before cross validation and after cross validation are discussed. Finally, it selects the best model for deployment and provide suggestions to choose their medical insurance providers and their medical insurance cost associated with them.

Evaluation Metrics

The Regression methods performance is measured by R² score, Mean Absolute Error (MAE) and Mean Square Error (MSE). In regression R² is a statistical measure for fitness of regression line and actual data. It is a coefficient of determination measured by using the eq(13).

$$R^2 = 1 - \frac{RSS}{TSS} \quad (13)$$

where RSS is a sum of squares of residuals from original value Y_i and predicted value which is defined by eq (14).

$$RSS = \sum_i (y_i - \hat{y}_i)^2 \quad (14)$$

TSS is a sum of squares between original value Y_i and average value .

$$TSS = \sum_i (y_i - \bar{y})^2 \quad (15)$$

RMSE.MAE is calculated using the formula (16) and (17).

$$RMSE = \sqrt{\frac{1}{N} \times \sum_i (y_i - \hat{y}_i)^2} \quad (16)$$

$$MAE = \frac{1}{N} \sum_i (y_i - \hat{y}_i)^2 \quad (17)$$

Results of Regression model before cross validation

The twelve regression models are trained and evaluated by splitting data set into 80% for training and 20% for testing. The r² score, RMSE, MAE of regression model is evaluated using training and test data set and the results are shown in Table 1. From this Table 1, the SVR have lower performance which has less R² score -0.08, high RMSE score 11168 and high MAE 6882 during training and provide low performance during testing. The EN and AB provide better R² score 0.46, 0.59 for training and 0.46, 0.59 for testing compared to SVR but lower performance with other models.

Table 1. Performance of regression models before CV

Model	Training data set			Test data set		
	R ²	RMSE	MAE	R ²	RMSE	MAE
Lasso	0.74	5642	3793	0.72	5801.0	3818.4
Ridge	0.74	5643	3795	0.72	5800.5	3819.3
KR	0.85	4239	2470	0.83	4311.2	2515.4
EN	0.74	5642	3793	0.71	5701.0	3818.6
DT	0.99	301.90	15.20	0.88	4820.1	1513.6
SVR	0.13	9981.1	4785	0.12	3627.9	5012.6
KNN	0.76	4293.1	2702	0.67	6702.8	3727.9
RF	0.98	1276.8	537	0.90	3550.0	1370.4
ET	0.99	301.9	15.2	0.91	3639.3	1163.1
AB	0.60	6808.7	6395	0.61	6461.1	6261.1
GB	0.90	3385.5	1039	0.88	3931.7	1361.3
XGB	0.98	1141.4	591	0.87	3856.7	1739.7

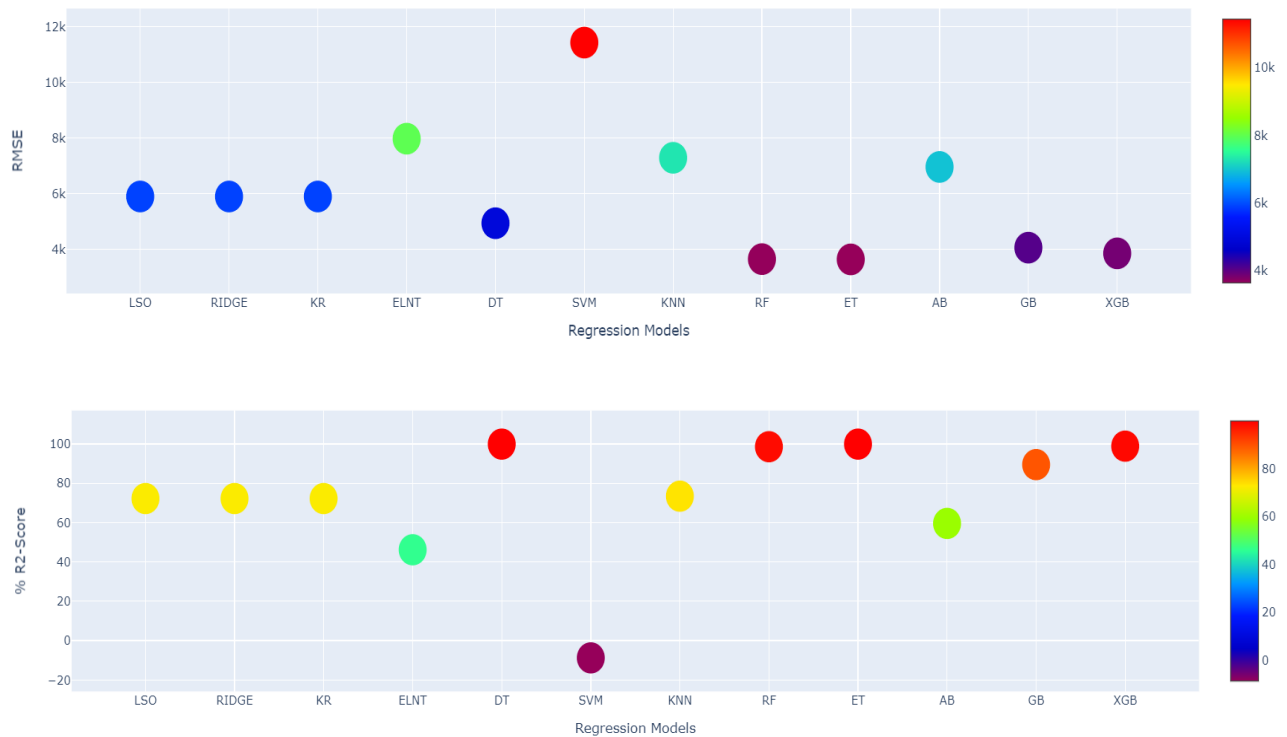


Figure 6. RMSE score for regression models before cross validation

Figure 7. R² score for regression models before cross validation

The Lasso, ridge, kernel ridge provide better R² score (0.72) but lesser than RF, DT, ETC and XGB. The Extra tree classifier provide good R² score (0.9), less MAE (15.2), less RMSE (train RMSE is 341, test RMSE is 3639) score compared with RF, DT, XGB. The RMSE score for regression models are shown in Fig. 6. From this figure, SVM is poorly performed model which provide high RMSE score compared with others. The RF, ET and XGB provide lower RMSE value compared with other models. The R² score for regression models are shown in Fig. 7. From this DT, RF, ET, GB and XGB provide good R² score (above 0.8) compared with other regression models.

Results of Regression model after cross validation

From this DT, RF, ET, GB and XGB provide good R² score (above 0.8) compared with other regression models. The grid search CV is employed over the models to reduce the over fitting and fine tune the performance of regression models. The performance of regression models is illustrated in Table 2. From this Table 2, the SVR have lower performance which has less R² score 0.131 which is higher than previous score without cross validation (score -0.088 in Table 1), high RMSE score 9981.147 and high MAE 4785.512 during training and provide low performance during training but cross validation helps to reduce this value compared from table 1.

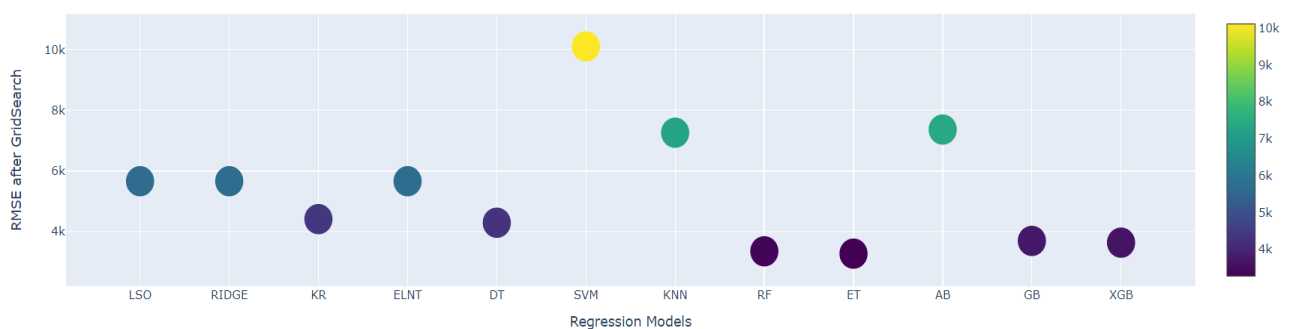


Figure 8. RMSE score for regression models after cross validation



Figure 9. R2 score for regression models after cross validation

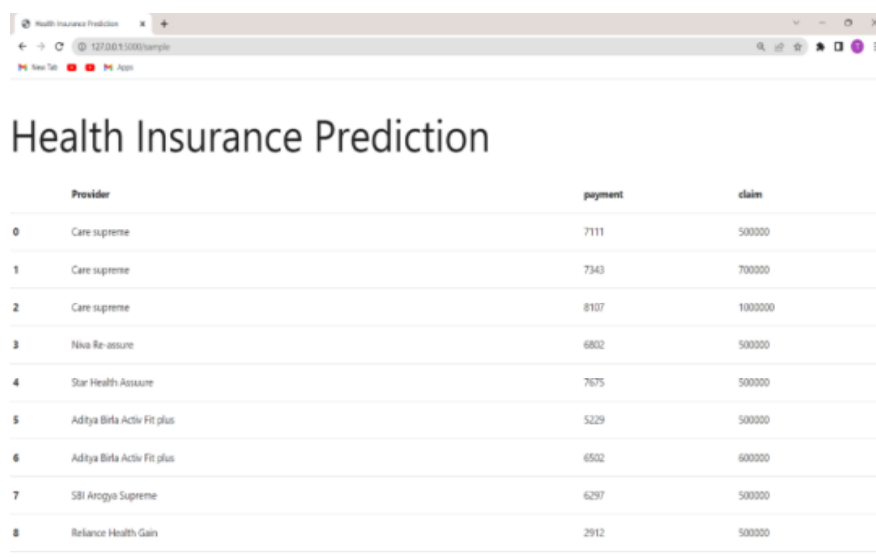
Table 2. Performance of regression models before CV

Model	Training dataset			Test dataset		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Lasso	0.72	5642	3793	0.71	5901	3918.4
Ridge	0.72	5643.	3795	0.71	5900	3919.8
KR	0.72	5643.	3795	0.71	5900	3919.9
EN	0.46	7854	5715	0.46	7978	5809.5
DT	0.99	341.9	15	0.79	4940	1613.6
SVR	-0.08	11168	6882	-0.10	11432	7202.2
KNN	0.73	5519	2702	0.55	7291	3827.9
RF	0.98	1276	537	0.88	3650	1480.4
ET	0.99	341.9	15	0.88	3639	1273.1
AB	0.59	6808.7	6395	0.59	6962	6461.1
GB	0.89	3474.1	1840	0.86	4061	2079.8
XGB	0.94	1386.3	591	0.86	3856	2087.0

Similarly for all regression models RMSE score is reduced after cross validation. The resultant RMSE score and R2 score is shown in Fig.8 and Fig.9. From this analysis, the extra tree classifier is selected for final prediction and model is deployed on web application. The prediction of health insurance cost is shown in Fig.10.

Figure 10. Health Insurance Cost Prediction

The developed health insurance recommendation system provides the list of medical insurance providers and the cost of medical insurance along with claim and payment details. This is shown in Figure 11.



	Provider	payment	claim
0	Care supreme	7111	500000
1	Care supreme	7343	700000
2	Care supreme	8107	1000000
3	Niva Re-assure	6802	500000
4	Star Health Assure	7675	500000
5	Aditya Birla Activ Fit plus	5229	500000
6	Aditya Birla Activ Fit plus	6502	600000
7	SBI Arogya Supreme	6297	500000
8	Reliance Health Gain	2912	500000

Figure 11. Health Insurance Scheme Prediction

CONCLUSION

Health insurance cost prediction is necessary to support individuals in selecting insurance plans, learning about appropriate deductibles and insurance rates based on their health. This research uses various regression learning models from linear regression to XGB for medical insurance prediction. The machine learning regression model performance is improved by grid search cross validation mechanism. Among the selected twelve classifiers, extra tree classifier gives good response for Medclaim cost prediction and provider selection. From experimental results, extra tree regression models give R^2 score is 99.9 during training and 91.9 R^2 score during testing. It gives low RMSE score compared with other regression models. This recommendation system helps the user to select the best providers based on cost estimation. For this research uses 75 medical insurance providers and their Medclaim cost prediction is carried out by machine learning algorithm. In future, we incorporate the several features of medical insurance providers and give suggestions according to it.

Conflicts Of Interest

The authors declare no conflict of interest.

REFERENCES

- [1] C. Edelman and E. C. Kudzma, Health promotion throughout the life span-e-book. Elsevier Health Sciences, 2021. Accessed: Sep. 25, 2024.
- [2] C.-H. Tu, D. E. Andersson, O. F. Shyr, and P.-H. Lin, "Earthquake Risk, Flooding Risk and Housing Prices: Evidence from Taichung, Taiwan," *Appl. Spatial Analysis*, vol. 16, no. 2, pp. 923–938, Jun. 2023, doi: 10.1007/s12061-023-09513-2.
- [3] S. Bansal and Y. Jin, "Heterogeneous Effects of Obesity on Life Expectancy: A Global Perspective," *Annu. Rev. Resour. Econ.*, vol. 15, no. 1, pp. 433–554, Oct. 2023, doi: 10.1146/annurev-resource-022823-033521.
- [4] K. Baicker, A. Chandra, and M. Shepard, "Achieving Universal Health Insurance Coverage in the United States: Addressing Market Failures or Providing a Social Floor?," *Journal of Economic Perspectives*, vol. 37, no. 2, pp. 99–122, 2023.
- [5] Z. Mustafa Busu et al., "Significance and Empowerment Through Self-hygiene According to Modern Medical and the Relationship with Health in Curbing the COVID-19 Epidemic," in *From Industry 4.0 to Industry 5.0*, vol. 470,
- [6] A. Hamdan, A. Harraf, A. Bualay, P. Arora, and H. Alsabatin, Eds., in *Studies in Systems, Decision and Control*, vol. 470. , Cham: Springer Nature Switzerland, 2023, pp. 1037–1049. doi: 10.1007/978-3-031-28314-7_87.
- [7] S. Albahri et al., "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion," *Information Fusion*, vol. 96, pp. 156–191, 2023.

- [8] S. Dubey, S. Deshpande, L. Krishna, and S. Zadey, "Evolution of Government-funded health insurance for universal health coverage in India," *The Lancet Regional Health-Southeast Asia*, vol. 13, 2023, Accessed: Sep. 25, 2024.
- [9] U. Phalswal, V. K. Neeraja, P. Dixit, and A. K. Bishnoi, "Government health insurance schemes and their benefits to the Indian population: An overview," *National Journal of Community Medicine*, vol. 14, no. 05, pp. 340–345, 2023.
- [10] H. Mahmoud, E. Abbas, and I. Fathy, "Data mining and ontology-based techniques in healthcare management," *IJIEI*, vol. 6, no. 6, p. 509, 2018, doi: 10.1504/IJIEI.2018.096549.
- [11] H. Azeez, "A Novel Binary Drawer Algorithm for Feature Selection in AI Application.," *International Journal of Intelligent Engineering & Systems*, vol. 17, no. 4, 2024, Accessed: Sep. 25, 2024.
- [12] G. S. M. Kalyani, "A Novel Ranking Approach to Improved Health Insurance Cost Prediction by Comparing Linear Regression to Random Forest," *Journal of Survey in Fisheries Sciences*, vol. 10, no. 1S, pp. 2030–2039, 2023.
- [13] D. Erlangga, M. Suhrcke, S. Ali, and K. Bloor, "The impact of public health insurance on health care utilisation, financial protection and health status in low-and middle-income countries: a systematic review," *PloS one*, vol. 14, no. 8, p. e0219731, 2019.
- [14] M. M. Dutta, "Health insurance sector in India: an analysis of its performance," *Vilakshan-XIMB Journal of Management*, vol. 17, no. 1/2, pp. 97–109, 2020.
- [15] Y. Li and D. Dou, "The influence of medical insurance on the use of basic public health services for the floating population: the mediating effect of social integration," *Int J Equity Health*, vol. 21, no. 1, p. 15, Dec. 2022, doi: 10.1186/s12939-022-01623-6.
- [16] B. Reshmi, B. Unnikrishnan, S. S. Parsekar, E. Rajwar, R. Vijayamma, and B. T. Venkatesh, "Health insurance awareness and its uptake in India: a systematic review protocol," *BMJ open*, vol. 11, no. 4, p. e043122, 2021.
- [17] 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.
- [18] Sahu, G. Sharma, J. Kaushik, K. Agarwal, and D. Singh, "Health Insurance Cost Prediction by Using Machine Learning," in *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*, 2022.
- [19] Mukund Kulkarni, Dhammadeep D. Meshram, Bhagyesh Patil, Rahul More, Mridul Sharma, Pravin Patange, Medical Insurance Cost Prediction using Machine Learning, *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, Volume 10 Issue XII Dec 2022.
- [20] K. Bhatia, S. S. Gill, N. Kamboj, M. Kumar, and R. K. Bhatia, "Health insurance cost prediction using machine learning," in *2022 3rd International Conference for Emerging Technology (INCET)*, IEEE, 2022, pp. 1–5. Accessed: Sep. 25, 2024.
- [21] Ch. A. Ul Hassan, J. Iqbal, S. Hussain, H. AlSalman, M. A. A. Mosleh, and S. Sajid Ullah, "A Computational Intelligence Approach for Predicting Medical Insurance Cost," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–13, Dec. 2021, doi: 10.1155/2021/1162553.
- [22] M. Hanafy and O. M. A. Mahmoud, "Predict health insurance cost by using machine learning and DNN regression models," *Int. J. Innov. Technol. Explor. Eng.*, vol. 10, no. 3, pp. 137–143, 2021.
- [23] G. Nagappan and V. U. Rani, *Disruptive Technologies for Sustainable Development*. CRC Press, 2024. Accessed: Sep. 25, 2024.
- [24] S. Vuddanti, V. R. K. Jamili, S. R. Bommareddy, V. Rotta, and V. Pagadala, "Machine Learning Insights into Personalized Insurance Pricing," in *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, IEEE, 2024, pp. 923–927. Accessed: Sep. 27, 2024.