iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70775

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Predictive Modeling for Healthcare Insurance Costs Using Machine Learning

Lokesh Khedekar[1], Bhawana Rakshit[2], Prutha Sawale[3], Siddhant Wasnik[4], Yashraj Shinde[5]
Department of Computer Science and Engineering (Artificial Intelligence), Vishwakarma Institute of Technology, Pune, India

Abstract: ***Insurance policies help to reduce financial losses by covering various risks, including medical expenses. The task of grading academic assignments is typically cumbersome, unequal, and also charged with some human bias due to personal judgment especially when it is subjective, for instance, essays and short hand answers. The paper presents an AI grading system that automates both objective and subjective assignment grading based on state-of-the-art technology. The system includes Optical Character Recognition (OCR) for processing handwriting, NLP models for evaluating essays and textual responses, and machine learning algorithms for objective questions in multiple-choice and fill-in-the-blank formats. This system also delivers detailed feedback to improve learning outcomes. After significantly reducing grading time while remaining fair and accurate, this system presents a scalable and efficient solution for modernization in educational evaluation processes.***
Keywords: Medical insurance, machine learning, Random Forest, regression, prediction, feature scaling, cross-validation.

I. INTRODUCTION

Out-of-pocket payments have largely been dominant in a healthcare system of most developing countries, giving rise to barriers to universal health coverage because of inefficiencies, inequities, and high costs. Health insurance serves as one critical mechanism for people to manage the financial risks associated with health care through providing coverage for treatment and related services. This means that high premium rates charged by insurance companies leave most people uninsured while causing delay in accessing healthcare and eventually leading to increased mortality.

Accurate predictions of individual healthcare expenses are fundamentally important to stakeholders and health departments because a lot of factors influence the costs of both insurance and healthcare. Such estimates enable health insurers and healthcare delivery organizations to plan adequately and allocate limited resources more efficiently. An understanding of possible future costs also gives the patients the better opportunity to select the more appropriate insurance plan for themself, in conformity with their financial means, including suitable deductibles and premiums.

Health insurance rate calculation is also a very complex issue, because each insurer has to find prices that balance profitability and affordability for the beneficiary. Such rates are usually founded on the probability of specific medical events occurring within a population pool. However, it would be highly challenging to make an exact estimate about medical expenses since those expensive conditions are rare and not easy to predict. The incidence of some diseases is also very different among various people and population segments, so there is a need for a fair and just model of premium calculation that considers these individual factors.

This paper utilizes demographic and behavioral data to construct a model for predicting healthcare costs. Most existing studies have relied on conventional statistical methodology, but the paper uses advanced machine-learning algorithms to increase accuracy of prediction. We compare four models: Linear Regression, K-Nearest Neighbors (KNN), Support Vector Regression, and Random Forest Regressor, to determine which one is effective enough. This concludes that the Random Forest model fares better in achieving greater predictive accuracy and generalization capabilities than the rest. Indeed, machine learning proves particularly valuable to the health care sector as data-driven methods in this sector increasingly become essential; it forecasts the expenditures on high-cost, high-need patients who will inevitably support resource reallocation and risk management in insurance industries.

II. LITERATURE REVIEW

[1]Sazzad Hossen "Medical Insurance Cost Prediction Using Machine Learning".This research paper introduces how several kinds of ML regression models may be utilized to predict medical insurance costs. It further emphasizes the requirement for insurance in order to protect individuals financially in the event that they are suffering from some rare diseases. The research uses a Kaggle dataset that includes features such as age, gender, BMI, no. of children, smoking status, and region to predict insurance charges.

©IJRASET: All Rights are Reserved

For this analysis, nine regression models were utilized, and these included the following: Linear Regression; XGBoost Regression; Lasso Regression; Random Forest Regression; Ridge Regression; Decision Tree Regression; KNN Model; Support Vector Regression; and Gradient Boosting Regression. Among them, the XGBoost Regression model had the best result with an R-squared value of 0.8681, with a Mean Absolute Error of 2381.567 and a Root Mean Squared Error of 4450.4433. Gradient Boosting and Random Forest models performed well too with R-squared values of 0.8679 and 0.8382 respectively. It brings out the facts that forecasting medical expenditure is not an easy task because of different uniqueness of data from various patients from the dataset. It means though more machine learning algorithms provide fast prediction results, accuracy varies markedly. Future work can include creating a web app using the best models up to this point and even more substantial datasets in order to realize more predictive power. The work, thus, shows how much more feasible it is to apply machine learning to optimizing the cost estimation process for medical insurance, both as an insurance provider and consumer, by accelerating and improving the accuracy with which one can analyze pricing estimates.

[2] Dr. S. M. Iqbal, Sayali D. Ghatol, Prerana V. Jadhav, Nikita D. Raspalle, "Health Insurance Cost Prediction Using Machine Learning" This paper intends to predict the cost of health insurance by applying machine learning algorithms and deals with a very essential challenge which is present in health care. According to the authors, multiple factors affect the cost of insurance and prediction of this amount is highly critical for stakeholders like the insurer and patient. The research places more importance on the better prediction models, mainly in developing countries in which out-of-pocket payment occurs and prevents healthcare access.

It utilizes a KAGGLE dataset with almost 1,000 demographics and health attributes for multiple individuals. Many machine learning algorithms are used in the paper, such as Extreme Gradient Boosting and Random Forest Regression, to come up with predictive models that can be used in different estimations of health insurance costs. Comparing the accuracy of different regression models helps identify the best approach for estimating the costs of health insurance for an individual.

The studied literature manifested an increasing concern in the application of machine learning for healthcare costs predictions; more studies demonstrated an exploration of various algorithms and techniques. The research highlighted the influence that machine learning might exert on efficiency when creating an insurance policy and the prediction for those who appear as high-cost patients.

The methodology section describes the process of data analysis, including preprocessing of data, feature engineering, and the splitting of the dataset into training/testing splits. A structured model training and validation approach is also drawn up, with a focus on the need for an iterative refinement of machine learning models toward optimal performance.

This study, therefore, contributes to how machine learning may be capitalized in predicting health insurance costs-that is, where insights into the complexities of insurance pricing and possibly improved accuracy in cost estimation can be shown.

[3]Kashish Bhatia, Shabeg Singh Gill, Navneet Kamboj, Manish Kumar,Rajesh Kumar Bhatia, "Health Insurance Cost Prediction using Machine Learning".Health insurance cost predictability through machine learning is found to have drawn much attention with recent times and post the COVID-19 pandemic due to critical importance of health coverage. Various studies that aimed at propounding more accurate estimates of cost prediction in medical insurance improved the efficiency of insurance systems through different machine learning techniques. Common attributes for such models include demographic and lifestyle factors such as age, gender, BMI, smoking habits, and the number of dependents or children, all being good indicators of health care needs and risks. Researchers have applied models like linear regression, decision trees, and even more complex methods like random forests and neural networks. These models give insight into the nature of the relationship these features have with the insurance premiums. Generally, linear regression has proven quite useful because it's quite simple and can display linear relationships between variables. The study provided a dataset sourced from Kaggle, which included personal medical data that allowed the training of machine learning models for the estimation of costs based on suitable attributes. Training was conducted with a 70-30 train-test split ensuring that strength in performance evaluation will be eminent, and it was achieved with an accuracy of 81.3%, showing that the model is strong to predict. On the other hand, the literature does indicate that though linear regression is good, the presence of models with the complexity other than being linear has the potential for further enhancing the prediction accuracy while considering nonlinear interaction between features. Moreover, present research work on this line of study emphasizes how the selection of the feature and preprocessing of data are important tasks to get better performance and generalizability of these models.

[4] Md Mohtaseem Billa, Dr. Tapsi Nagpal, "Medical Insurance Price Prediction Using Machine Learning". It supports the machine learning approach in predicting the price of medical insurance with which health care services also face climbing prices and complexities of health services. It further establishes an efficient model of prediction that will provide proper insights to stakeholders about pricing strategies and risk management. It employs a rich dataset in providing demographic details, history of illness, lifestyle exposure, and insurance coverage for the patient. Several machine learning algorithms, namely regression, decision trees, and random forests, were run and compared for better model performance as well as their interpretability.

This paper reviewed various machine learning approaches, ranging from regression-based methods, time series forecasting methods, ensemble methods, deep learning strategies, hybrid models, and many other issues on their strengths and weaknesses. Issues for future research in implementing machine learning in medical insurance price prediction include data accessibility, feature selection, model interpretability, scalability, and generalization. Opportunities are then given to future research to make such models more accurate and useful.

In the introduction, the medical insurance price has been painted to be much more complex and dependent on such factors as demographical characteristics and economic trends, and for such a complex situation, traditional actuarial methods often fail to capture relationships in complex situations because the pertinent relationships among the factors could often be complex. Further with this intricacy, "dynamically changing market conditions" can be unraveled through machine learning methodologies.

In the methodology section, it describes the data set used for which is downloadable from Kaggle, split into a training data set and a test data set, with attributes in the form of age, gender, BMI, number of children, smoking status, and region. Further detail obtained from the underlying data elucidates that it consists of 2,773 entries and further explains how these apply to insurance charges. Discussion of data preprocessing: In this aspect, there is no missing value of the dataset available which is required to perform the apt machine learning analysis.

Overall, this study gives a view to the researcher, practitioner, and policymaker about the use of machine learning methodologies that are informative to health care and decision-making.

[5]Ugochukwu Orji, Elochukwu Ukwandu, "Machine learning for an explainable cost prediction of medical insurance". In applying ML models in cost prediction of medical insurance, the paper was interested in the explainability aspect related to the proposed models. In fact, three regression-based ensemble ML models namely Extreme Gradient Boosting (XGBoost), Gradient-boosting Machine (GBM), and Random Forest (RF) are applied in analyzing 986 records in the KAGGLE repository. This research will mirror the growing interest in predictive modeling in healthcare as a result of the increased demand on insurance companies to be productive and more efficient.

The results show that all the models performed very well. With XGBoost at the top in terms of overall performance though computationally more demanding while the RF model had lower prediction errors, together with fewer resources. In this analysis, two XAi methods have been compared viz: SHapley Additive exPlanations (SHAP) and Individual Conditional Expectation (ICE) plots; to understand key factors affecting the medical insurance premium. The case of ICE plots demonstrates clear interactions in variables, in which one can only make general statements based on SHAP. Still, both approach the same outcome.

This, consequently, brings much emphasis on the correct actuarial modeling involved in the determination of insurance premium, especially considering such significant variations and complexities surrounding medical insurance cost determinations. The authors strive to make some discoveries that could be helpful to policymakers, insurers, and would-be buyers in making informed decisions regarding their choice of the right insurance policies. The research contributes to the current discussion on the role ML plays within healthcare, calling for more transparency and accountability in predictive analytics to help build a higher level of trust with stakeholders.

III. RELATED WORK

Generally, machine learning has dominated the approach to regression tasks of this type. Indeed, although some earlier work in this domain was primarily based on linear models, such as Linear Regression purely based on the assumption of linearity between input features and the target variable, there is still merit in making use of such types of models, though they might not reflect the complexity often encountered in the data. Contrarily, non-linear models, such as Random Forest and K-Nearest Neighbors (KNN), performed better in explaining complex variable-variable interactions, where that translates into more accurate predictions in a rather wide range of applications.

The applications include forecasting critical outcomes, such as disease progression, readmissions into hospitals, and fraudulent insurance claims. Thus, these applications demonstrate the potential of applying machine learning techniques in the management of healthcare data. However, our work is fundamentally different from those applied to this area because it will comparatively analyze a variety of machine learning models which have been specifically designed for predicting the medical insurance cost. This paper compares the performance of each model regarding how well each model can manage real-world insurance data based on some metrics, for instance, predictive accuracy and robustness. The algorithms tested and compared in a systematic fashion include Linear Regression, KNN, Support Vector Regression (SVR), and Random Forest among others. In this regard, a major approach toward predicting healthcare costs is used. Such a study contributes both to existing literature and as a source of useful input for healthcare providers and insurers as they continue to strive to optimize their pricing strategies and risk management models.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

IV. DATA DESCRIPTION

The dataset for this study is mostly demographic in nature, dealing with health, comprising:
Age: the age of the proposer.
Sex: Sex of the insured.
Whether one is overweight or not, that's BMI: Body Mass Index.
Children: Number of children / dependents provided with insurance.
Smoker Status: Whether the insured person is a smoker.
Region: The region in which the policyholder lives.
Charges: Insurance premium, which is the amount to be forecasted.
The dataset had no missing values and was therefore suitable for model training. The categorical feature variables, including sex, smoker status, and region, had to be encoded into a numerical format for model compatibility.

V. METHODS

A. Data Preprocessing:

Very few preprocessing steps were required before the models could actually start training.
Feature Encoding: It makes all categorical variables such as sex, region, etc. smoker numerical using OneHotEncoding.
Scaling: Algorithms like KNN, SVR are sensitive to scale. Therefore, we have normalized the numerical features of the datasets using StandardScaler so that all of them will come on an equal scale.

B. Model Selection:

We have trained four different machine learning models for this work:
A basic regression model assuming a linear relation of features with the target. Linear Regression (LR).
KNN: It is a nonlinear model based upon the value predicted by the closest neighbors in the dataset.
Support Vector Regression (SVR): While for regression it fittingly reaches the best line within thresholds to minimize the prediction errors.
RF: Random Forest Regressor. Ensemble algorithm with multiple decision trees that average out predictions to increase accuracy and reduce overfitting.

C. Cross-Validation:

To increase the strength of predictions even further, we applied 5-fold cross-validation on these models. This means that our training data were split into five folds: we trained on different folds and tested them on the remaining data. The average Mean Squared Error across all folds is reported as a cross-validation score.

D. Performance Metrics:

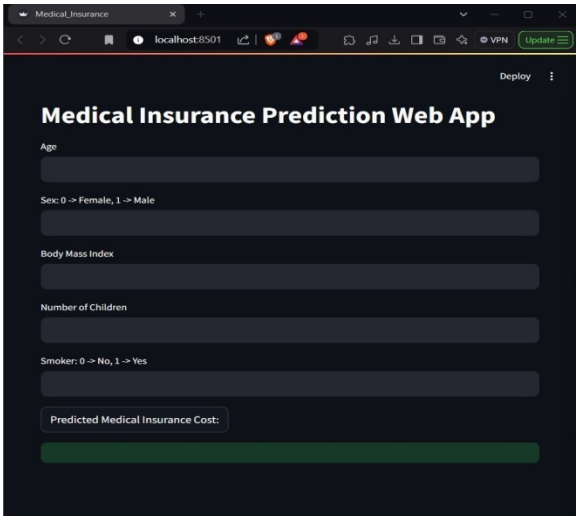
We measure the performance of each model on the test set in terms of the above.
R^2 Score: This is the extent to which your model explains variance in the target variable.
MSE: Mean Squared Error Average square difference of prediction and actual values.
MAE: Measures the mean absolute deviation of the predicted and actual values.

VI. DISCUSSION

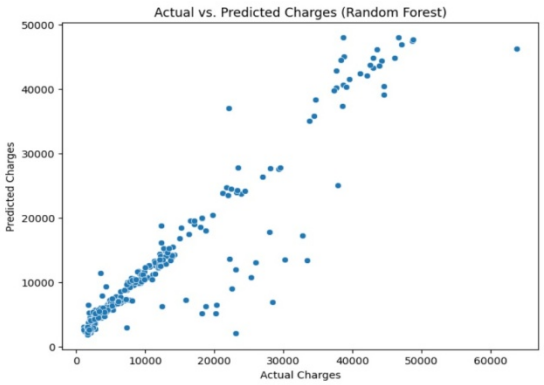
The best performance for the charge prediction for medical insurance is found to be in the Random Forest model: it seems possible to deal with complex interactions between features and is very accurate. The difference of performance between RF and the models above indicates that ensemble methods are more suitable for insurance charge prediction, where relationships between a feature-for example, age, smoking, BMI-and target variables are probably nonlinear.
Optimal performance was achieved by both KNN and SVR models after scaling the features. The other one, Random Forest, is rather insensitive to feature scaling: a characteristic of its invariable potential capabilities.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

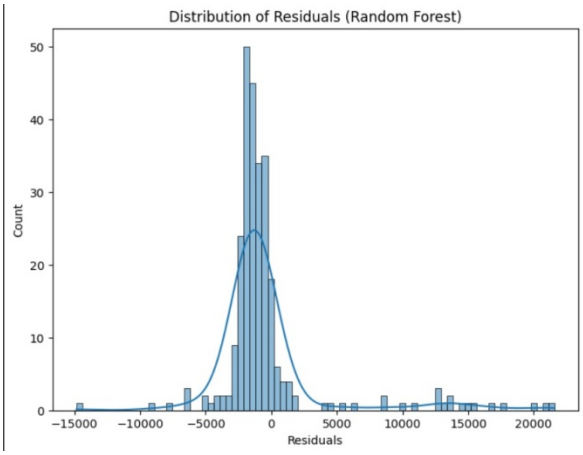
VII. RESULTS



Thus, the web application so designed can estimate the medical insurance charges based on the region and other input parameters such as age and BMI. In this regard, the developed interface of the web application is using an interactive Random Forest regression model for easy estimation of the charges.



The graph of actual medical insurance charges versus the values actually obtained by the Random Forest regression model are plotted. Most of the line closely follows along the diagonal, so the model can be said to be working well enough as to predict charges with minimal deviation in most cases, albeit there are some outliers for which the model does not quite capture the variance in the dataset.



The residual plot does evidence the bias of the predicted over the actual charges. If residual follows a more or less normal pattern around zero, then it does indicate that the errors in the model are random, with no bias in predictions; thus, this would demonstrate that the model performed fairly well on all kinds of cases.

```
[ ]
 ------------------------------------------------
  Training Model LR
Cross-validation MSE: 38064899.879597545
Testing R-squared: 0.7835929767120722
Mean Squared Error: 33596915.851361476
Mean Absolute Error: 4181.194473753652

 ------------------------------------------------
  Training Model RF
Cross-validation MSE: 22845347.097363733
Testing R-squared: 0.8722924273131
Mean Squared Error: 19826438.66153469
Mean Absolute Error: 2451.5552593281895

 ------------------------------------------------
  Training Model KNN
Cross-validation MSE: 33511315.3525976
Testing R-squared: 0.7981792099067552
Mean Squared Error: 31332421.65064374
Mean Absolute Error: 3560.316902205224

 ------------------------------------------------
  Training Model SVR
Cross-validation MSE: 158218026.5440263
Testing R-squared: -0.06892210909501029
Mean Squared Error: 165948801.50051147
Mean Absolute Error: 8611.328379758412
```

The cross validated and tested various regression models by using metrics of the test set. It has the cross-validation MSE of the Linear Regression model, which was 38,064,899.88, the test set R-squared of 0.78, MSE, 33,596,915.85; MAE is 4,181.19. RF was superior to LR. Although its cross-validation MSE is 22, 845, 347.10, on the test set, it surpasses with an R-squared of 0.87, with a cross-validation MSE of 19, 826, 438. 66 and an MAE of 2,451.56. KNN performed moderately since its cross-validation MSE is 33, 511, 315.35, on the test set, MSE of 31, 332, 421.65, and MAE of 3,560.32 and R-squared of 0.80. In contrast, the Support Vector Regression model performed very poorly: cross-validation MSE was achieved as 158,218,026.54 with a negative test R-squared of -0.07. The MSE remained high at 165,948,801.50 while the MAE was 8,611.33. The R-squared values and error in the Random Forest model indicate the best generalization performances.

```
] print('MSE train data: %.2f, MSE test data: %.2f' % (
      mean_squared_error(y_train, forest_train_pred),
      mean_squared_error(y_test, forest_test_pred)))

  MSE train data: 13554236.27, MSE test data: 19617514.85

] print('R2 train data: %.2f, R2 test data: %.2f' % (
      r2_score(y_train, forest_train_pred),
      r2_score(y_test, forest_test_pred)))

  R2 train data: 0.91, R2 test data: 0.87
```

The MSE was computed on both the training and test datasets. So, on the training set, it was 13,554,236.27 and on the test set, it was 19,617,514.85, meaning that there is a small amount of overfitting happening as error is a bit higher when the model is making predictions on unseen data. Also, the R-squared value was **0.91** for training data and **0.87** for test data. Here also, it is showing that the model is picking up quite a significant share of variance in both datasets. Though test set performance is marginally degraded, good high R-squared values indicate good predictive powers and generalization of the model as a whole.

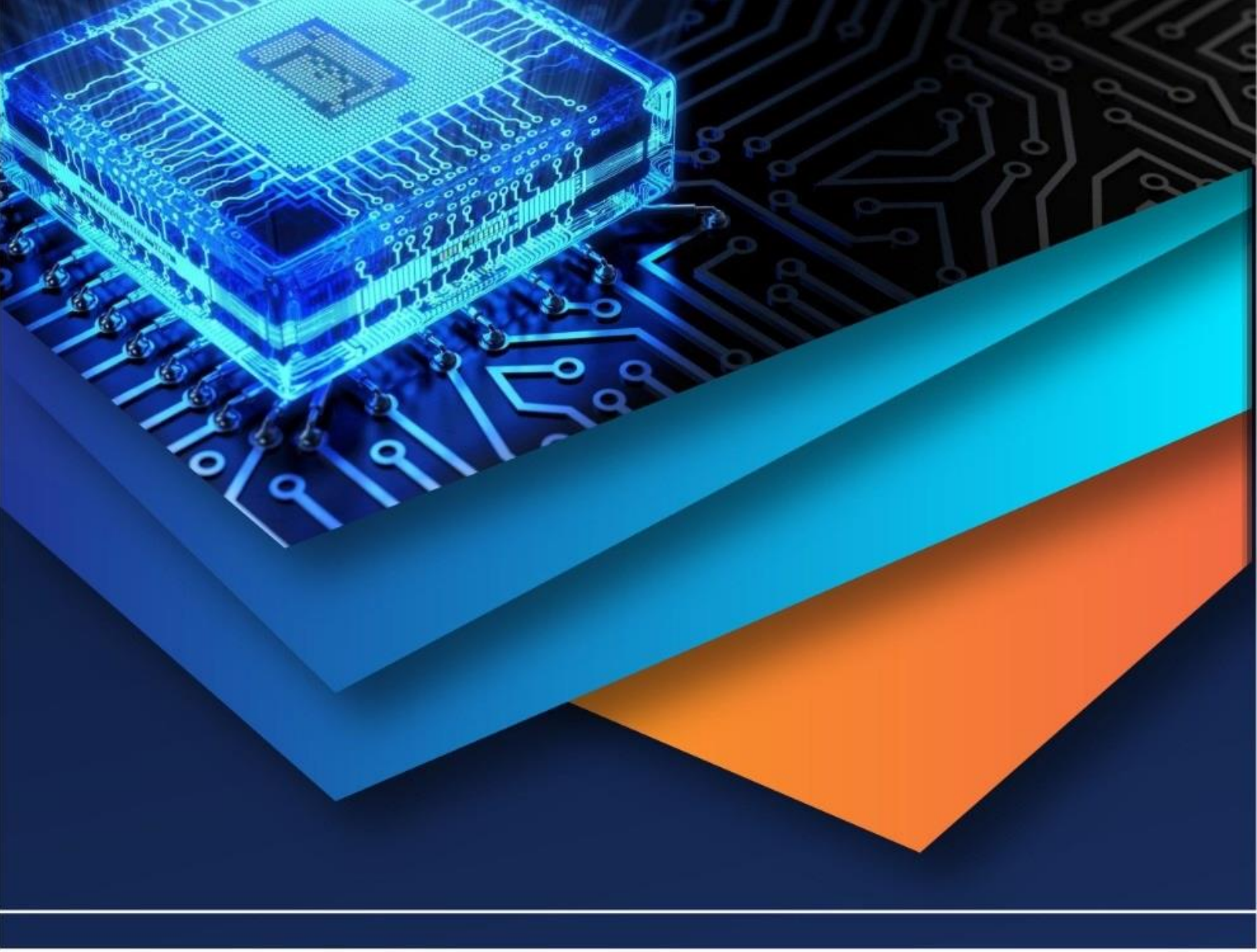
VIII. CONCLUSION

This study compared the performance of four machine learning models for predicting medical insurance charges. The Random Forest Regressor proved to be the most effective model, providing the highest R^2 score and the lowest Mean Squared Error. These results suggest that Random Forest is well-suited for cost prediction tasks involving non-linear relationships between features.

Future research could focus on further enhancing model performance by incorporating more advanced techniques like XGBoost or Gradient Boosting, and feature engineering to derive new features from the existing dataset. Additionally, hyperparameter tuning for models like SVR and Random Forest could improve their predictive accuracy.

REFERENCES

[1] Sazzad Hossen "Medical Insurance Cost Prediction Using Machine Learning". October 2023 DOI:10.13140/RG.2.2.31456.25604 Thesis for: Medical Insurance Cost Prediction
[2] Dr. S. M. Iqbal, Sayali D. Ghatol, Prerana V. Jadhav, Nikita D. Raspalle, "Health Insurance Cost Prediction Using Machine Learning" .
[3] Kashish Bhatia, Shabeg Singh Gill, Navneet Kamboj, Manish Kumar,Rajesh Kumar Bhatia, "Health Insurance Cost Prediction using Machine Learning".
[4] Md Mohtaseem Billa, Dr. Tapsi Nagpal, "Medical Insurance Price Prediction Using Machine Learning".
[5] Ugochukwu Orji, Elochukwu Ukwandu, "Machine learning for an explainable cost prediction of medical insurance".
[6] S. Panda, B. Purkayastha, D. Das, M. Chakraborty and S. K. Biswas, "Health Insurance Cost Prediction Using Regression Models".
[7] T. T, S. H. T, V. K. V and K. R, "Medical Insurance Cost Analysis and Prediction using Machine Learning," 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 113-117, doi: 10.1109/ICIDCA56705.2023.10100057
[8] R. D, M. S. K and D. J, "Health Insurance Cost Prediction using Machine Learning Algorithms," 2022 International Conference on Edge Computing and Applications (ICECAA), Tamilnadu, India, 2022, pp. 1381-1384, doi: 10.1109/ICECAA55415.2022.9936153.
[9] A. Vinora, V. Surya, E. Lloyds, B. Kathir Pandian, R. N. Deborah and A. Gobinath, "An Efficient Health Insurance Prediction System using Machine learning," 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2023, pp. 1-5, doi: 10.1109/ICSES60034.2023.10465334.
[10] Kaushik K, Bhardwaj A, Dwivedi AD, Singh R. Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. Int J Environ Res Public Health. 2022 Jun 28;19(13):7898. doi: 10.3390/ijerph19137898. PMID: 35805557; PMCID: PMC9265373.
[11] Sahu, Ajay and Sharma, Gopal and Kaushik, Janvi and Agarwal, Kajal and Singh, Devendra, Health Insurance Cost Prediction by Using Machine Learning (February 22, 2023). Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2022.
[12] Mukund Kulkarni, Dhammadeep D. Meshram, Bhagyesh Patil, Rahul More, Mridul Sharma, Pravin Patange "Medical Insurance Cost Prediction using Machine Learning"
[13] Uber Rides Prediction using Machine Learning Lokesh S Khedekar, Ajay S Chhajed, Ravishankar C Bhaganagre, Naina S Kokate, Swaraj Patil-2025 International Conference on Electronics and Renewable Systems (ICEARS) DOI:10.1109/ICEARS64219.2025.10940984
[14] Creative sustainability: Transforming household waste in India–A public survey on awareness and participation, L Khedekar, A Pandit, R Apte, B Adke, P Ambade, A Aher,Challenges in Information, Communication and Computing Technology DOI:10.1201/9781003559085-7
[15] Team portal website: Development & constructing bridges between teams,L. Khedekar, R. Dane, N. Dgama, S. Dangat, R. Dagade, V. Dahatonde, Challenges in Information, Communication and Computing Technology DOI:10.1201/9781003559085-30
[16] Innovating Healthcare: Developing a Comprehensive Patient Record Tracker System for Enhanced Medical Data Management and Patient Care,Lokesh Khedekar, Atharva Dhananjay Mohite, Arnav Meghan Kamat, Arpit Anil Topugol, Purva Dipak Atale, Pranay Suresh Asniyekar,SSRN 5086771
[17] AgriTech: Technology Driven E-Commerce Platform for Sustainable Agricultural Development ,Lokesh Khedekar, Radhika Dagade, Vaibhav Dahatonde, Rohit Dane, Sanskar Dangat, Prem Deore, Nevan Dgama, 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)