International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 04 | Apr 2024 www.irjet.net p-ISSN: 2395-0072

Health Insurance Cost Prediction Using Machine Learning

Dr. S. M. Iqbal[1], Sayali D. Ghatol[2], Prerana V. Jadhav[3], Nikita D. Raspalle[4]

[1]Professor, Dept. of CSE Engineering, PRMIT&R college, Maharashtra, India
[2]Student, Dept. of CSE Engineering, PRMIT&R college, Maharashtra, India
[3] Student, Dept. of CSE Engineering, PRMIT&R college, Maharashtra, India
[4]Student, Dept. of CSE Engineering, PRMIT&R college, Maharashtra, India

---

Abstract - *Insurance is a policy that helps to cover up all losses or decrease losses in terms of expenses incurred by various risks. Several factors influence the insurance cost. Various factors contribute to the determination of insurance policy costs. Predicting medical insurance costs remains a challenging issue in the healthcare industry. Predicting medical insurance costs is still a problem in the healthcare industry and thus it requires more investigation and improvement. Machine learning is one of the computational intelligence aspects that may address diverse difficulties in a wide range of applications and systems when it comes to the exploitation of historical data. Using a series of machine learning algorithms, a computational intelligence approach will be proposed for predicting healthcare insurance costs. The system will predict the approximate cost of health insurance for a person by using the dataset from KAGGLE. A medical insurance cost dataset was acquired from the KAGGLE repository for this purpose, and machine learning algorithms were used to show how different regression models can predict insurance costs and to compare the models' accuracy*.

Keywords: **Machine learning (ML), Artificial-intelligence(AI), Health insurance(HI), Premium cost, Regression algorithm(RA).**

1. Introduction

Healthcare systems in developing countries depend heavily on out-of-pocket payments, the mechanism that is a barrier to universal health coverage, as it contributes to inefficiency, inequity, and cost. Health insurance serves as a means for individuals in different countries to manage the financial risk associated with medical expenses. It provides coverage against the costs incurred from medical treatment and related services. However, due to the high rates that are charged by insurance companies, many people are without health insurance and so fail to access timely health services which results in high death rates. A health insurance policy is a policy that covers or minimizes the expenses of losses caused by a variety of hazards. Accurately predicting individual healthcare expenses using prediction models is crucial for various stakeholders and health departments, as numerous factors influence the cost of insurance or healthcare. Accurate cost estimates can help health insurers and, increasingly, healthcare delivery organizations to plan for the future and prioritize the allocation of limited care management resources. Furthermore, knowing ahead of time what their probable expenses are for the future can assist patients in choosing insurance plans with appropriate deductibles and premiums [1]. These factors contribute to the formulation and development of insurance policies. However, health insurance rates calculations are often complex as they need to determine the rates that are acceptable to both insurance companies and beneficiaries; Insurance companies need to make money by collecting more money than they spend on the medical expenses of their beneficiaries, hence making a profit and continue to stay in businesses. These companies price the premiums based on the probability of certain events occurring among a pool of people [2]. However, the medical expenses and other associated costs are difficult to estimate because the costliest conditions are rare and seemingly random. Another complex part of estimating medical expenses is that the occurrence of certain diseases differs from one person to another and from one segment of the population to another. Therefore, there is a need for a fair premium calculation model that suits the unique population factors. In this regard, this study used demographic and behavioral data from the patients to develop a predictive model. While previous studies used conventional statistical methods, this study used machine learning logarithms to develop a predictive model. It compares the performance of several models to find the most suitable. In the insurance sector, machine learning can help enhance the efficiency of policy wording. In healthcare, machine learning algorithms are particularly good at predicting high-cost, high-need patient expenditures. machine learning can be categorized into three different types. There are three main types of machine learning: supervised, unsupervised, and reinforcement learning. Supervised machine learning involves tasks such as classification and regression, where the data is labeled, and the algorithm learns from provided input-output pairs. Unsupervised machine learning, on the other hand, is used for tasks like clustering, where the data is unlabeled, and the algorithm discovers patterns or structures within the data. Reinforcement learning is a type of learning where the algorithm learns through trial and error by interacting with an environment to achieve a goal [5].

2. Literature Review

In the literature review, various studies are discussed regarding the prediction of health insurance premiums and healthcare costs using machine learning algorithms.

One study recommends the use of Extreme Gradient Boosting (XGBoost) and Random Forest Regression (RFR) to develop more accurate models for predicting premiums [1]. Another study employs gradient-boosting models for predicting medical insurance costs [2]. A computational intelligence approach using regression-based machine learning algorithms is proposed for predicting healthcare insurance costs [3].

Additionally, new ensembles are developed for individual insurance cost prediction to improve prediction accuracy [4]. Various regression models are explored to forecast insurance costs, with comparisons made among them [5]. Furthermore, novel ranking techniques with machine learning algorithms are applied to classify cost prediction in health insurance [6].

Another approach involves training and evaluating an artificial intelligence network-based regression model to predict health insurance premiums based on individual features [8]. The aim is to predict future high-cost patients using machine learning algorithms such as Random Forest, Gradient Boosting Machine, Artificial Neural Network, and Logistic Regression [10].

Moreover, the temporal consistency of healthcare expenditures in a state Medicaid program is studied using predictive machine learning models, particularly for high-cost, high-need patients [12]. Another method utilizing machine learning algorithms is proposed for predicting medical costs, aiming to guide patients to affordable care and assist policymakers in identifying costly providers [13].

Finally, an artificial neural network model is developed for predicting annual medical claims [14]. Overall, these studies highlight the importance of machine learning techniques in predicting health insurance premiums and healthcare costs, with various algorithms and approaches being explored for improved accuracy and efficiency.

3. Methodology

he medical cost personal datasets were sourced from the KAGGLE repository [15]. This dataset comprises eleven attributes, as outlined in Table 1. Provided by nearly 1000 customers voluntarily, the data includes various health-related parameters. The premium prices, denoted in INR (₹), represent annual costs for the customers.

Attribute	Data Description
Age	The age of the person
Diabetes	Whether The Person Has Abnormal Blood Sugar Levels
Blood Pressure Prob	Whether The Person Has Abnormal Blood Pressure Levels
Any Transplants	Any Major Organ Transplants
Any Chronic Disease	Whether Customer Suffers From Chronic Ailments Like Asthma, Etc.
Height	Height Of Customer
Weight	Weight Of Customer
Known Allergies	Whether The Customer Has Any Known Allergies
History Of Cancer In Family	Whether Any Blood Relative Of The Customer Has Had Any Form Of Cancer
Number Of Major Surgeries	The Number Of Major Surgeries That The Person Has Had
Premiums	Premium Prices for A Whole Year

Table1. Overview of the dataset

3.1 Data Analysis

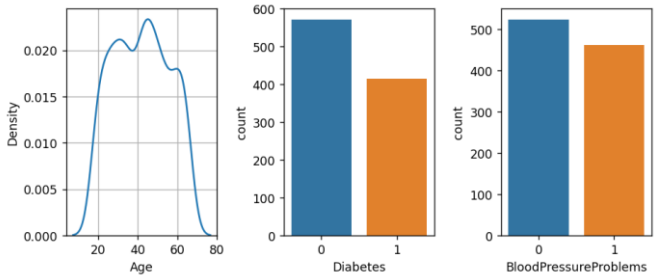

Fig 1. Distribution of Age, Diabetes, and Blood pressure problems

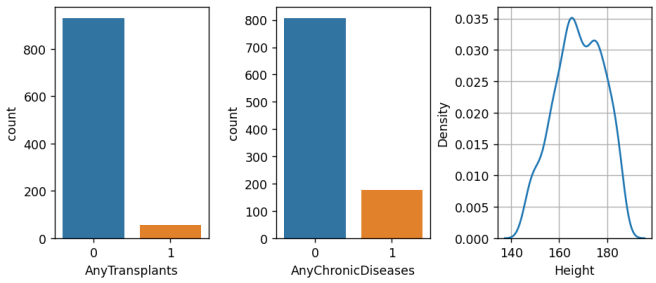

Fig 2. Distribution of Any Transplants, Any Chronic Diseases, Height

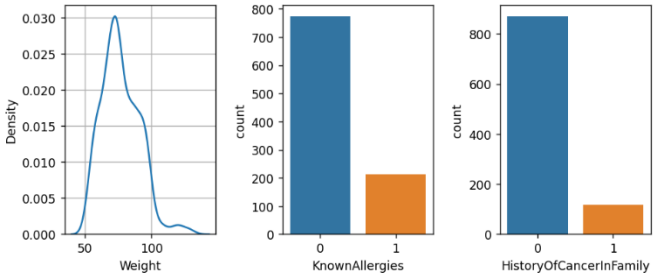

Fig 3. Distribution of Weight, Known Allergies, History of Cancer in Family

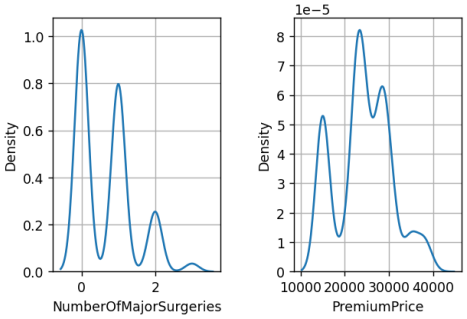

Fig 4. Distribution of Number of Major Surgeries, Premium Price

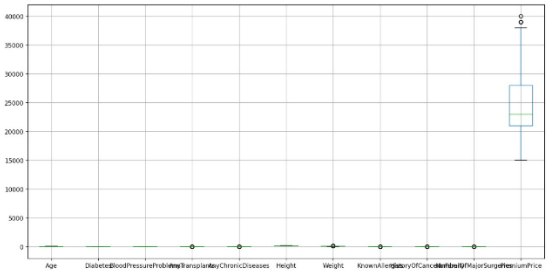

Fig 5. Candlestick pattern

3.2 Block Diagram

We conducted machine learning techniques on medical insurance data. The dataset for medical insurance costs was obtained from KAGGLE's repository [15], and we conducted data preprocessing. Subsequently, we performed feature engineering to select the relevant features. The dataset was then split into two parts: a training set and a test set. Approximately 70% of the total data was allocated for training purposes, with the remaining portion reserved for testing. The training set was utilized to develop a model for predicting medical insurance costs for the year, while the test set was employed to assess the performance of the regression models. To facilitate regression analysis on the dataset, categorical values were converted into numerical values. The steps of our methodology are illustrated in Figure 6.

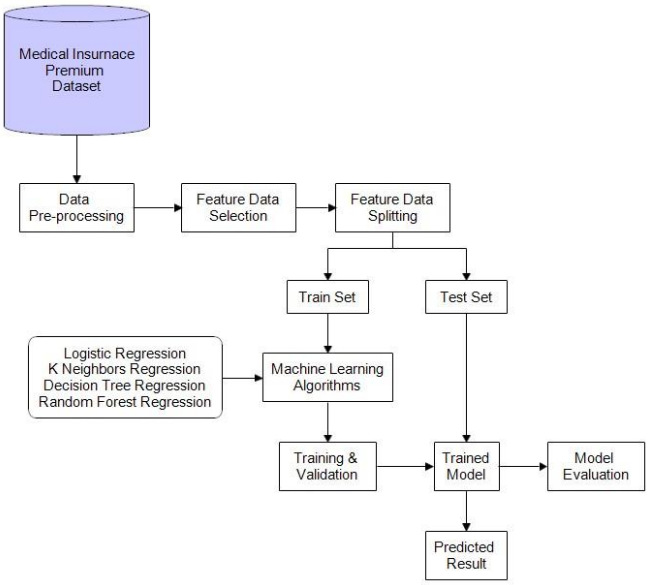

Fig 6. Working methodology

3.3 Training Model Flowchart

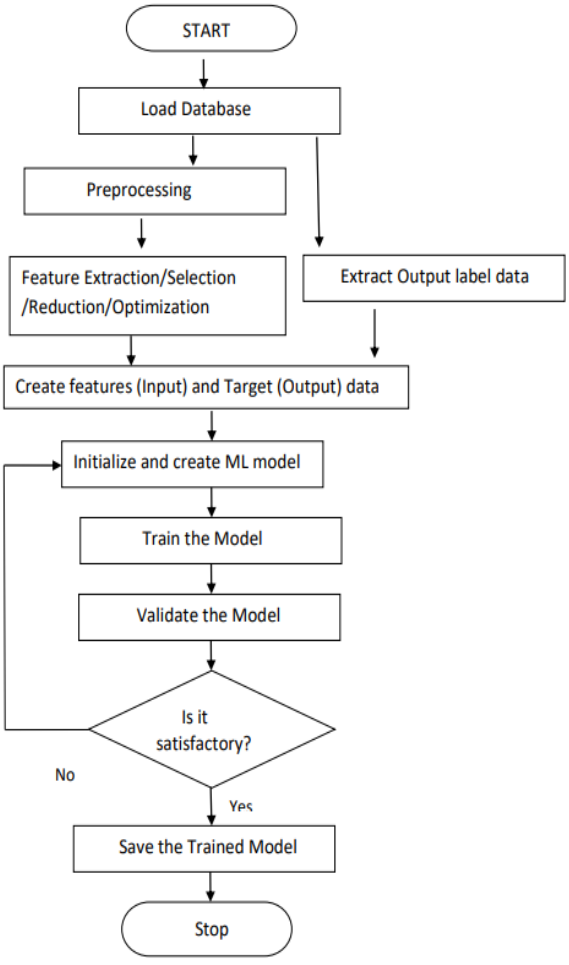

Fig 7. Training Model Flowchart

3.4 Testing Model Flowchart

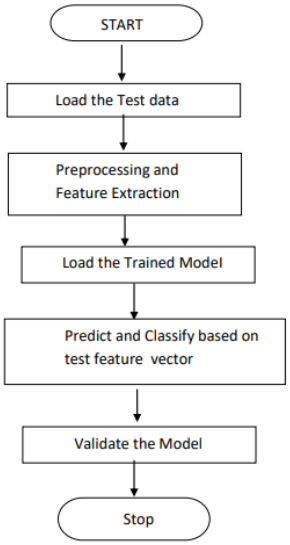

Fig 8. Testing Model Flowchart

Training algorithm explanation

As we know, in supervised approach has a dataset that includes input data with several input attributes and its equivalent output label data for training and testing. So in general, data need to be divided in two ways based on validation strategies, which is further explained, as 70-30% or 60-20- 20%for train, test or train, validate, and test. So input attributes that depend on the application scenario will be given to the preprocessing operation which includes several operations that again depend on input data, after features are extracted from preprocess data, if needed have feature reduction/selection depending on the methodology. Then extracted feature will be input START Load the Test data Load the Trained ModeI Predict and Classify based on test feature vector Validate the Model Stop Preprocessing and Feature Extraction features for machine learning implementation and its equivalent label will be taken directly from the original dataset as output/ target vector. Then need to initialize the parameters regarding selected /shortlisted ML algorithms and build the architecture for them. After that, we have to train the architecture that we built using a specific toolbox or toolkit depending on the implementation language and its platform. After successful learning or training, one needs to cross-check the trained model whether it is effectively trained or not. So for verification, there is a procedure called Validation. In validation there are two ways to do that, first directly use train data for validation, for that train data need to be simulated on the trained model to get 100% test results then only we can say, the model is validated. The second option is to use a validation dataset for it to get 90-100% results to achieve maximum efficiency. If validation is not satisfactory as per the desired test value result as mentioned for both options, then need to do several trial-and-error things to achieve its desired result.

1. Retrain the model by changing the network parameter value and again check for improvement in validation.

2. If the first option doesn't work with the possible parameters value option, make a change in the features step, extract more distinct features from attribute data or select such features that are more unique according to its mentioned output category, or reduce the features length to avoid feature data complexity. After again train as per procedure with step 1 to get desired validation.

3. Again if the second step doesn't work, then need to work on pre-processing operations like noise removal or many others and again check for validation with steps 1 and 2.

4. Again if the third step also didn't work by considering particular ML algorithms with step1,2,3. Then change ML algorithms to do the same with repeat trial and error from step 1 to step 4.

So after getting successful validation, we need to store the trained model for further testing results.

Testing algorithm explanation

In testing, whatever final preprocessing and feature extraction method is selected while training, needs to apply the same for the testing procedure. Then the trained model which is locally stored or from cloud storage, needs to apply it on the test input feature vector to get the predicted or classified result which can be verified from the test database. Further, the efficiency of the test model is evaluated using a confusion matrix and ROC curve from which we get statistical parameter values that can show the overall project performance.

Now, while doing all this we assumed that dataset splitting is done as discussed above (70-30% or 60-20-20% for train, test or train, validate and test respectively) manually in one random manner. But there are no possible ways to split a dataset as dataset size increases number of splitting possibilities increases. It is your good luck if, in one attempt of splitting, you get maximum accuracy of the model by doing the above procedure of training and testing, then it is fine. If still, it is still not, then there is a method to improve the accuracy of an overall model called

- K- fold cross validation method

In K-fold, 'k' stands for the number of groups that a given data sample is to be split into. So we have to check the value of 'K' to achieve the maximum value of your desired project model

3.5 Feature Engineering

Feature engineering in machine learning involves extracting meaningful features from raw data, often leveraging domain knowledge to enhance the performance of ML algorithms. Since the effectiveness of machine learning models heavily depends on the quality of input data during training, feature engineering plays a critical role as a preprocessing step. It entails selecting the most pertinent attributes from the raw training data, tailored to both the predictive task and the type of model being utilized.

In the medical insurance cost dataset, the attributes listed are deemed crucial factors influencing the premium amount. Feature scaling is a common standardization technique used to normalize features and mitigate the impact of large-scale differences on the models. Unlike feature transformation, which converts data from one type to another, feature scaling adjusts data in terms of range and distribution while preserving its original data type

Min-max scaling is a specific method within feature scaling that rescales all values of a particular feature to fit within a predefined minimum and maximum range, typically 0 and 1. Each value of a data point for the selected feature is recalculated relative to the specified minimum and maximum feature values, resulting in a new feature value for that data point. This process helps maintain consistency and comparability among different features.

$$\tilde{x} = \frac{x - min(x)}{max(x) - min(x)}$$

3.6 Dataset Splitting

In machine learning modeling, dataset splitting is an essential step that aids in various stages from training to evaluating the model. It serves as a crucial sub-step, enabling a realistic assessment of model performance and facilitating generalization to unseen data. Ensuring an adequate quantity of training data relative to testing and validation sets is imperative for effective learning without bias towards any specific class or category**.**

Following the identification of the best-performing model, the test set is employed to assess its performance quantitatively. The dataset is divided into two distinct subsets: a training set and a testing set. In the experimentation process, a common splitting ratio of 70% for training data and 30% for testing data is often utilized to strike a balance between training sufficiency and model evaluation accuracy.

3.7 Machine Learning Algorithms, Training and Validation

A machine learning model is a computational tool that utilizes algorithms to make predictions based on input data. Unlike traditional methods that rely on predefined equations, these models learn directly from the provided dataset. By analyzing a known set of input data and corresponding responses (outputs), the model is trained to generate predictions for new, unseen data

Supervised learning, the simplest form of machine learning, operates on the principle of input-output pairs. In this approach, the input data, also known as training data, is paired with known labels or results as outputs. A function is then created using the training dataset, which is subsequently applied to unseen data to make predictions. Supervised learning is task-oriented and evaluated using labeled datasets. It can be further categorized into two main types: classification and regression-based approaches

Classification models are utilized to conclude from categorical observations. These algorithms predict outcomes by categorizing data into different groups based on observed features. In contrast, regression models are employed when the output variable is continuous

In the current experimentation, the regression models used are as follows: [mention specific models used]. These models are applied to predict continuous variables based on the provided dataset.

3.7.1 Logistic Regression (LR)

Logistic regression (LR) is a supervised machine learning algorithm specifically crafted for binary classification tasks. It predicts the probability of an outcome, event, or observation. The resulting outcome is typically binary, represented as yes/no, 0/1, or true/false. By analyzing the relationship between one or more independent variables, logistic regression categorizes data into discrete classes. This method is extensively used in predictive modeling to estimate the mathematical probability of whether an instance belongs to a particular category or not.

3.7.2 K Neighbors Regression (KNR)

K-nearest neighbors (KNN) regression is a non-parametric method that approximates the relationship between independent variables and a continuous outcome. It's a supervised machine learning algorithm that's used for classification and regression problems. KNN regression stores all available cases and predicts the numerical target based on a similarity measure. It does this by averaging observations in the same neighborhood. The size of the neighborhood can be set by the analyst or chosen using cross-validation.

3.7.3 Decision Tree Regression (DTR)

The decision tree is a method used to build regression or classification models, presented in a tree-like structure. It partitions a dataset into smaller subsets while constructing a corresponding decision tree. The resulting tree consists of decision nodes and leaf nodes.

In decision tree regression, the features of an object are examined, and a model is constructed in the form of a tree to predict future data, generating continuous output. Continuous output implies that the result is not discrete; rather, it represents a range of values, providing meaningful predictions for continuous variables.

3.7.4 Random Forest Regression (RFR)

Random forest regression is a supervised learning algorithm employing an ensemble learning approach for regression tasks. It is a versatile technique used for predicting numerical values. By aggregating the predictions of multiple decision trees, random forest regression aims to mitigate overfitting and enhance accuracy.

In random forest regression, several decision trees serve as base learning models. The algorithm involves randomly selecting rows and features from the dataset to create sample datasets for each model. During training, random forest constructs numerous decision trees and aggregates their predictions to produce the final output. For regression tasks, the output is typically the mean prediction of the individual trees, aiming to provide a more robust prediction.

Model Evaluation

In model evaluation, we assess the performance of the model using various metrics such as R Square, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

The function calculates four performance metrics:

1. RMSE (Root Mean Squared Error): This metric represents the square root of MSE. RMSE quantifies how far the model's predictions deviate from the actual values. Lower RMSE values indicate better model performance.

2. MSE (Mean Squared Error): MSE measures the average squared difference between the actual values and the predicted values. A lower MSE suggests that the model's predictions are closer to the actual values, indicating better performance.

3. R2 (R-squared or Determination Coefficient): R2 evaluates how well the independent variables explain the variability of the dependent variable. It ranges between 0 and 1, with higher values indicating a better fit of the model to the data. A value closer to 1 signifies that the model fits the data well

These metrics collectively provide insights into the accuracy and effectiveness of the model in predicting the target variable.

4. Experimental Results and Discussion

The dataset contains health-related parameters of the customers. By using the health parameters attributes, an artificial intelligence model is built by performing the training and validation on the train feature dataset. A trained model is created that predicts the yearly medical cover cost of a customer based on their health parameters.

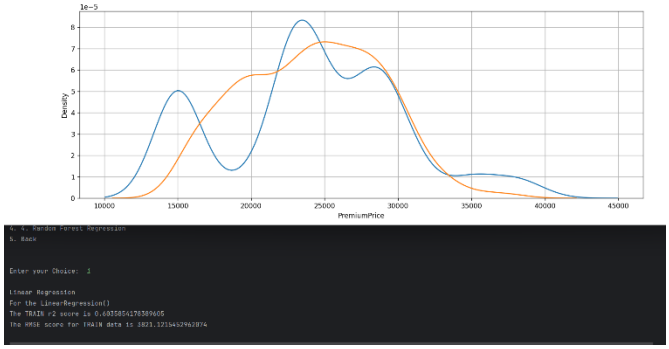

Fig 9. LRNet_Trained_model

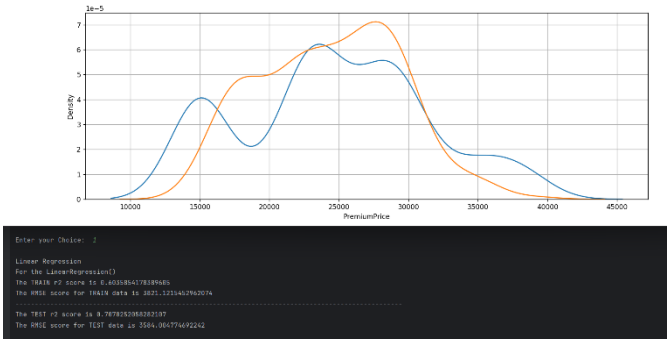

Fig 10. LRNet_Testing_model

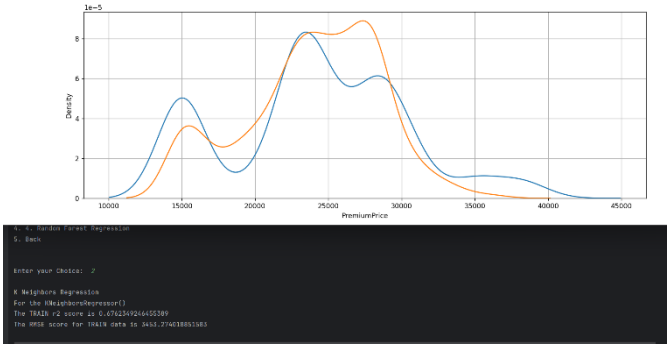

Fig 11. KNRNet_Trained_model

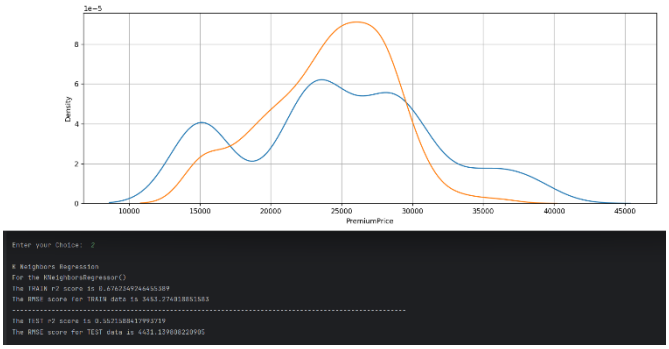

Fig 12. KNRNet_Testing_model

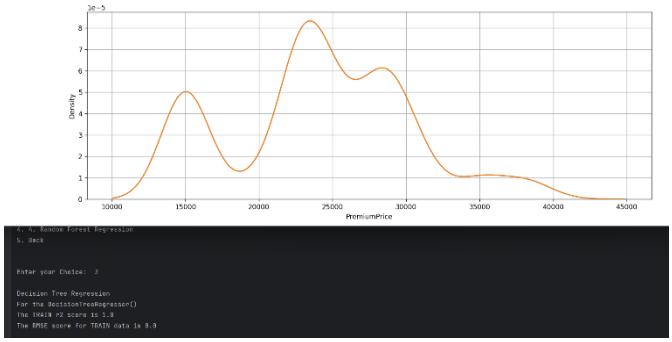

Fig 13. TRNet_Trained_model

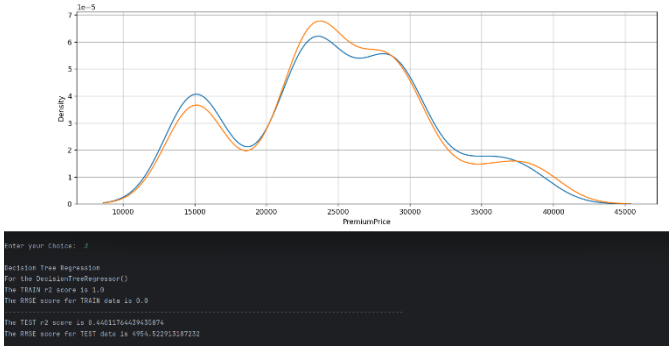

Fig 14. TRNet_Testing_model

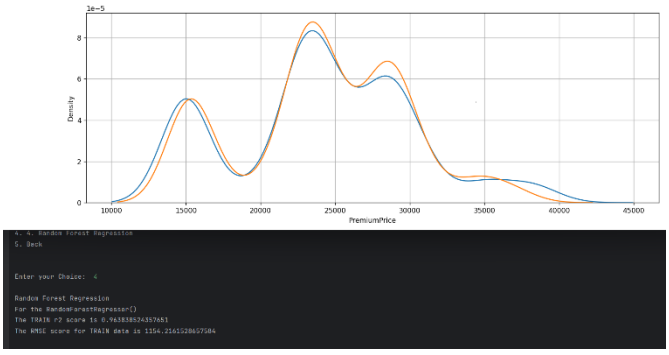

Fig 15. RFRNet_Trained_model

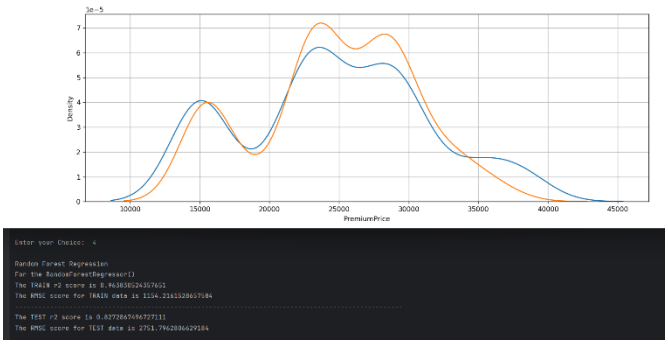

Fig 16. RFRNet_Testing_model

Here the training and testing of regression models are done and each model RMSE and R2 score is calculated for LR, KNR, DTR, and RFR. Also, the elapsed time for training and testing is calculated for each regression model.

Algorithms	Training		Testing	
	RMSE	R2_Score	RMSE	R2_Score
LR	3821.1	0.6035	3584.0	0.7070
KNR	3453.2	0.6762	4431.1	0.5521
DTR	0.0	1.0	3720.3	0.6843
RFR	1179.1	0.9622	2630.0	0.8422

Table2. Model performance

Fig 17. Health insurance cost prediction system demo

In Figure 11: we can see how the medical insurance cost prediction system predicts the cost for LR, KNR, DTR, and RFR.

5. Conclusion

In the realm of healthcare, machine learning presents itself as a powerful tool capable of performing tasks at a faster pace compared to human counterparts. The integration of machine learning into health insurance processes stands to revolutionize the industry, resulting in significant time and cost savings for both policyholders and insurers alike. By automating repetitive tasks, artificial intelligence enables insurance experts to allocate their time toward enhancing the overall experience for

policyholders. This shift not only streamlines administrative processes but also improves the efficiency of patient care, benefiting stakeholders across the healthcare ecosystem, including patients, hospitals, physicians, and insurance providers.

This paper has explored various machine learning regression models tailored for predicting health insurance charges based on specific attributes. By leveraging these models, insurance providers can expedite the formulation of tailored plans for individuals, thereby saving considerable time and effort in policymaking. Overall, the integration of machine learning holds the potential to revolutionize the health insurance landscape, making processes faster, more affordable, and ultimately more responsive to the needs of both policyholders and insurers.

References

1. Angela D. Kafuria, "Predictive Model for Computing Health Insurance Premium Rates Using Machine Learning Algorithms", International Journal of Computer (IJC), 2022, 44(1), 21–38.

2. Cenita, Jonelle Angelo & Asuncion, Paul Richie & Victoriano, Jayson. "Performance Evaluation of Regression Models in Predicting the Cost of Medical Insurance". International Journal of Computing Sciences Research, 2023, (ISSN print: 2546-0552; ISSN online: 2546-115X) Vol. 7, pp. 2052-2065 doi: 10.25147/ijcsr.2017.001.1.146.

3. Ch. Anwar ul Hassan, Jawaid Iqbal, Saddam Hussain, Hussain AlSalman, Mogeeb A. A. Mosleh, Syed Sajid Ullah, "A Computational Intelligence Approach for Predicting Medical Insurance Cost", Mathematical Problems in Engineering, vol. 2021, Article ID 1162553, 13 pages, 2021. https://doi.org/10.1155/2021/1162553

4. Shakhovska, Natalya & Melnykova, Nataliia & Chopiyak, Valentyna & machine learning, Michal., "An Ensemble Methods for Medical Insurance Costs Prediction Task", Computers, Materials and Continua., 2022, 70. 3969-3984. 10.32604/cmc.2022.019882.

5. Hanafy, Mohamed. "Predict Health Insurance Cost by using MACHINE LEARNING and DNN Regression Models", International Journal of Innovative Technology and Exploring Engineering, 2021, Volume-10. 137. 10.35940/ijitee.C8364.0110321.

6. G. Satya Mounika Kalyani, Rama Parvathy L, "A Novel Ranking Approach to Improved Health Insurance Cost Prediction by Comparing Linear Regression to Random Forest", Journal of Survey in Fisheries Sciences, 2023, 10(1S) 2030-2039.

7. Mukund Kulkarni, Dhammadeep D. Meshram, Bhagyesh Patil, Rahul More, Mridul Sharma, Pravin Patange, "Medical Insurance Cos3t Prediction using MACHINE LEARNING", International Journal for Research in Applied Science & Engineering Technology (IJRASET), 2022, ISSN: 2321-9653; Volume 10 Issue XII Dec 2022.

8. 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

9. Prakash, V. S., Bushra, S. N., Subramanian, N., Indumathy, D., Mary, S. A. L., & Thiagarajan, R. "Random Forest regression with hyperparameter tuning for medical insurance premium prediction", International Journal of Health Sciences, 2022, 6(S6), Page 9 7093–7101. https://doi.org/10.53730/ijhs.v6nS6.11762

10. Langenberger B, Schulte T, Groene O, "The application of MACHINE LEARNING to predict high-cost patients: A performance-comparison of different models using healthcare claims data". PLoS ONE 18(1): e0279540. 2023 https://doi.org/10.1371/journal.pone.0279540

11. K. Dutta, S. Chandra, M. K. Gourisaria and H. GM, "A Data Mining based Target Regression-Oriented Approach to Modelling of Health Insurance Claims," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1168-1175, doi: 10.1109/ICCMC51019.2021.9418038.

12. Yang, C., Delcher, C., Shenkman, E. et al., "MACHINE LEARNING approaches for predicting high-cost high need patient expenditures in health care". BioMed Eng OnLine 17 (Suppl 1), 131 (2018). https://doi.org/10.1186/s12938-018-0568-3

13. 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.

14. Goundar, S.; Prakash, S.; Sadal, P.; Bhardwaj, A. "Health Insurance Claim Prediction Using Artificial Neural Networks". Int. J. Syst. Dyn. Appl. 2020, 9, 40–57.

15. Dataset link: https://www.kaggle.com/datasets/tejashvi14/medical-insurance-premium-prediction/data