International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Medical Insurance Premium Prediction using Machine Learning

***D. Rohan[1], P. Rohan[2], B. Rohith[3], A. Rohith[4], T. Rohith[5], K. Rupa Sri[6], K. Manoj Sagar[7]***

[1]2111CS020406, [2]2111CS020407, [3]2111CS020408, [4]2111CS020409, [5]2111CS020410, [6]2111CS020411, B. Tech [CSE] MRUH
[7]Assistant Professor, MRUH
[1]2111CS020406@mallareddyuniversity.ac.in, [2]2111CS020407@mallareddyuniversity.ac.in, [3]2111CS020408@mallareddyuniversity.ac.in,
[4]2111CS020409@mallareddyuniversity.ac.in, [5]2111CS020410@mallareddyuniversity.ac.in, [6]2111CS020411@mallareddyuniversity.ac.in,

**ABSTRACT:**

Most of the work goes to medical bills. Health expenditure accounts for approximately 30% of GDP. Health expenditures are highest in developing countries, both in absolute terms and as a percentage of the economy. The state covers a significant part of the medical expenses of the elderly through health insurance. Rising health care costs, combined with impending retirement and the baby boomer generation's subsequent lack of access to Medicare, pose a significant financial burden. Therefore, every available tool should be used to limit health care costs. In this study, we will develop a method to estimate the cost of treatment using machine learning algorithms, which will help inform patients about the cost. Technology can also help regulators identify providers that are often more expensive and impose fines if necessary.

## 1. Machine Learning:

The random forest regression algorithm will be used in machine learning to predict medical costs. We also aim to test experiments and compare results using different learning models (e.g., gradient boosted trees and linear regression) on the same data. It helps in estimating health insurance costs early. Additionally, people can easily mistakenly believe that they are paying for expensive health insurance that they do not need. Our research does not give the exact cost of a particular doctor, but it does give a general picture of the cost a person will be charged for health insurance.

## 2. Introduction:

We live in a world full of threats and uncertainties. Such individuals, families, fixed assets and related products are subject to different risks and the level of risk may differ. These risks include the risk of disease, death if not protected, and loss of property or assets [1]. But most risks cannot be avoided, so the financial community has developed many products to protect people and organizations from these risks using capital financing. Therefore, insurance is one of the policies that reduce or eliminate the cost of damage caused by various risks. Cost of individual life insurance. It is therefore very important that the insurance company is sufficiently clear when assessing the cost of services for certain policies and the insurance premiums that must be paid for the same. Many parameters or factors play an important role in estimating the insurance cost and each of them is important. If something is ignored or changed when calculating the price, all official prices will change. Therefore, it is important that this job is done with the right people. Therefore, the possibility of human error is high and therefore insurance agents also use different tools to calculate insurance premiums. So machine learning is useful here. Machine learning can expand the functions or processes of policy making. These machine learning models can learn on their own. This model has been examined on previous insurance data. The model can estimate the true cost of the policy by using appropriate criteria to evaluate the payment according to its components. This reduces the number of employees and resources and increases the profitability of the company. Therefore, accuracy can be improved by machine learning. Our goal is to estimate insurance costs. The amount of the insurance premium depends on the difference. Therefore, the cost of insurance is fixed. Regression is the best choice for our needs. We use multiple linear regression in this analysis because there are many independent variables used to calculate the dependent (target) variable. This study uses health insurance cost data. [2] First, preprocess the dataset. We then show the regression model on the training data and finally evaluate the model on the testing data. In this paper, we used various regression models such as multiple linear regression, pruning tree regression, and gradient boost regression. Gradient boosting was found to give the most accuracy with an r-squared value of 86.7853. Collaborating with new methods for insurance cost estimation is the main goal of this project.

## 3. Literature Review

### *3.1 Literature Review:*

This section provides a review of knowledge discovery and machine learning studies. There is a lot of literature discussing the difficulty of requesting a quote. “I fix a car insurance claim with remote information,” says Jessica. This study compares the performance of expected regression and XGBoost techniques in predicting accident probability on a small scale, and the results show that regression is better than XGBoost due to better definition and stronger prediction of the model. [4] Invented by Ranjodh Singh in 2019, this technology uses images of damaged cars as input and produces key points (such as repair costs) to make decisions based on the number of claims and the location of the damage. Therefore, the need for vehicle insurance will not be taken into account in the tender analysis, instead the focus will be on the plan for repair costs. Oskar Sucki 2019, The purpose of this analysis is to analyze customer churn forecast. Random Forest is considered the simplest model (75% correct). There are missing values in some fields in the data set. After performing a high-level analysis of the classification, we chose to replace the missing variables with additional features indicating missing data. This is usually allowed, given that data is optionally completely lost, so first create the lost data adjusting the processing method accordingly. In 2018, Muhammad rFauzan used XGBoost's accuracy to predict sentences in this article. Compare the output with the performance of XGBoost, methods such as AdaBoost, Random Forest, Neural Networks. XGBoost provides more accurate Gini models. Lighting for public access to the Kaggle dataset in the city of Seguro. This data contains many NaN values, but this article uses median and median transformation to handle missing values. However, these simple and naive ideas turned out to be wrong. Therefore, their goal is to find a cubic centimeter method like XGboost that will be good for many missing problems. G. Kosalia, M. Nandini. In 2018, this research developed a classification system to predict the best results based on users' personal and financial information on different products to predict and predict unfair claims. Random Forest, J48 and Naive Bayes algorithms were chosen for classification. The results show that Random Forest outperforms other prediction methods on fake data. Therefore, this article does not address predictive claims, but focuses on false claims. While previous studies did not consider the weight of each agreement or request, only creating a classification for the problem request (even if it was not requested for that record holder), in this study we tend to focus on advanced mathematics. algorithms and deep neural networks to predict health insurance premiums.

### *3.2 Regression :*

Multivariate analysis can be a prognostic technique that investigates the relationship between dependent variables (target variables) and independent variables (predictor variables) [5]. This technology is used to estimate and predict the time structure of the project to obtain the relationship between variables. For example, during this analysis I needed to evaluate the relationship between insurance rates (objective variable) and 6 independent variables (age, BMI, number of children, personal area of residence or gender, whether the client smokes or not). About regression strategy. As mentioned earlier, many analyzes estimate the shares of two or more variables. I predict insurance premiums using a fully variable regression model of six free variables, and using this regression we can predict future health insurance premiums, supported by current and historical data. There are several advantages of using regression analysis as follows:- 1) It shows the relationship between variables and experimental variables. 2) It shows that for many free variables the strength of the results goes beyond the number of variables. Regression analysis helps compare results by measuring the difference between different indicators compared to independent and related variables [6]. These results enable business researchers, information analysts, and data scientists to develop and apply best methods for predicting multiple variables at their own standards.

## 4. Problem Statement:

Health insurance costs continue to rise at an alarming rate. In 2019 alone, average house prices increased by 8.1%. This trend is expected to continue and increase the burden on employers and individuals. Forecasting plays an important role in controlling insurance costs and increasing employee satisfaction. Legal cost estimates can help employers create more competitive packages and help people plan their budgets accordingly. Modern systems often fail to capture the nuances of human health, leading to inaccuracies. The main challenge of the project is to use machine learning techniques to develop powerful predictive models that can predict health insurance premiums based on a variety of demographic and social factors.

## 5. Methodology:

### *5.1 Machine Learning:*

Machine learning is a branch of computer science and artificial intelligence that uses data and algorithms to replicate the way humans learn. These algorithms are designed to use statistical methods for classification or prediction to reveal important insights during the data mining process. When used correctly, the results of these insights can significantly contribute to business and application growth. (S. Ramakrishnan, 2016) Data shows that age and smoking have the biggest impact on insurance coverage, while smoking has the biggest impact. However, factors such as family medical history, body weight, marital status, and region of residence also play a role. Data shows that age and smoking have the biggest impact on insurance premiums, while smoking has the biggest impact. However, factors such as family history, body measurements, marital status, and region of residence also play a role.

### *5.2 Linear Regression Algorithm:*

Linear regression is a machine learning algorithm based on the concept of "supervised learning". It is used to predict the value of variable (y) based on the value of variable (x). Essentially, this means that linear regression is used to determine how well the variables are related to the independent variables, and then the prediction is based on that relationship. (H. Goldstein, 2012) Again, prediction of future outcomes.

### *5.3 Support Vector Machine Algorithm:*

SVM or Support Vector Machine is a widely used supervised learning algorithm used to solve classification and regression problems, focusing on classification in machine learning (T. Han, 2020). border. The site can be divided into different classes, making it useful for new information for the future. This well-defined boundary is called the hyperplane and is created using points called extreme vectors and support vectors. SVM is a popular choice due to its ability to classify data well and handle high density. This is a preliminary prediction. This well-defined boundary is called a hyperplane.

### *5.4 Random Forest Regression:*

The bootstrapped random forest approach uses multiple decision trees constructed from data and combined through the learning process. This method generally provides accurate prediction and classification by averaging the results of selected trees. (X.Zhu, C.Ying, J.Wang, 2021)

### *5.5 Training:*

Once the necessary data is generated and prepared, the model can begin the training and testing phase. The main goal of the training phase is to select the appropriate model for the task at hand. This may include deciding on the best model or determining the best value for a model (V. Roth, 2014). In some cases, this process is called sample selection because there are many samples that can be tested and selected, and eventually the best one is chosen; This is created using points called cloud vectors and support vectors.

### *5.6 Prediction:*

The formula used to estimate health insurance premiums is based on the relationship between certain characteristics and labels. The accuracy of this estimate depends on how accurate the expected value is. Multiple features, methods, and training-test split sizes are used in the model to increase accuracy. Research has shown that the amount of data used for training correlates with accuracy, and larger training sessions lead to better results. The model also uses various algorithms to predict premium and show how each feature interacts (Kaggle, resource recovery)

## 6. Experimental Results:

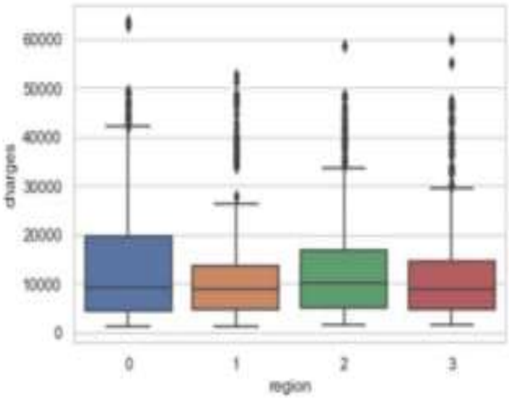


Fig.1.Boxplot of Medical charges per Region

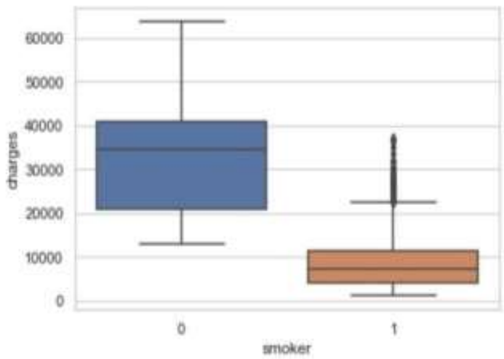


Fig.2. Boxplot of Medical charges per Smoking Status

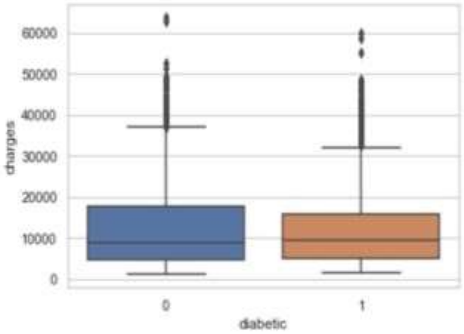


Fig.3. Boxplot of Medical charges per Diabetic Status

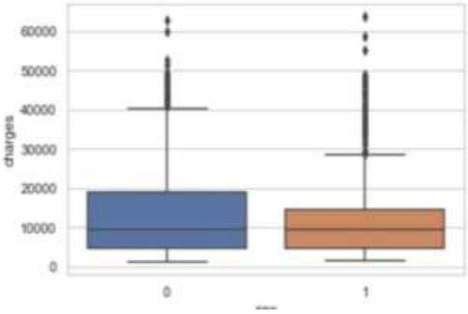


Fig.4.Boxplot of Medical charges per Gender

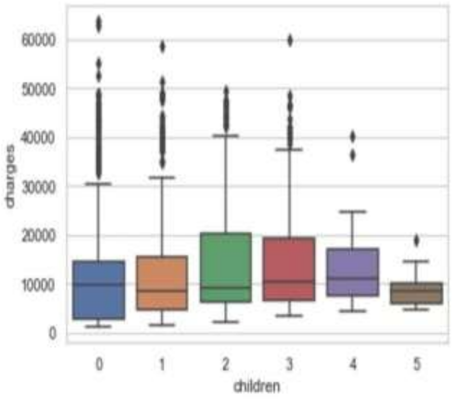
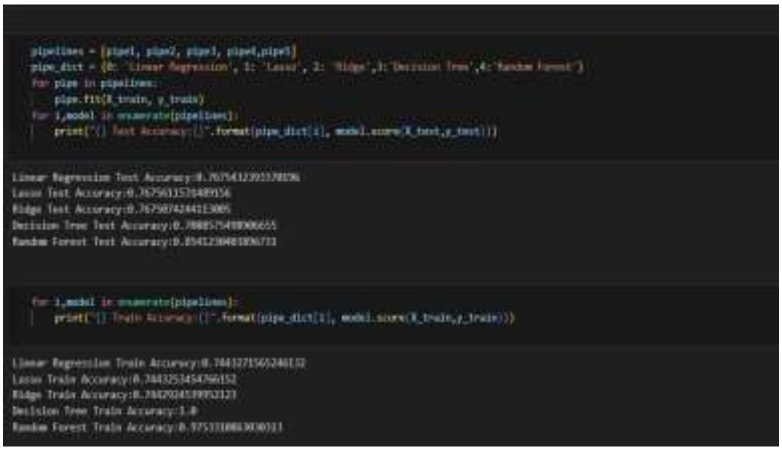


Fig.5.Boxplot of Medical charges per Children



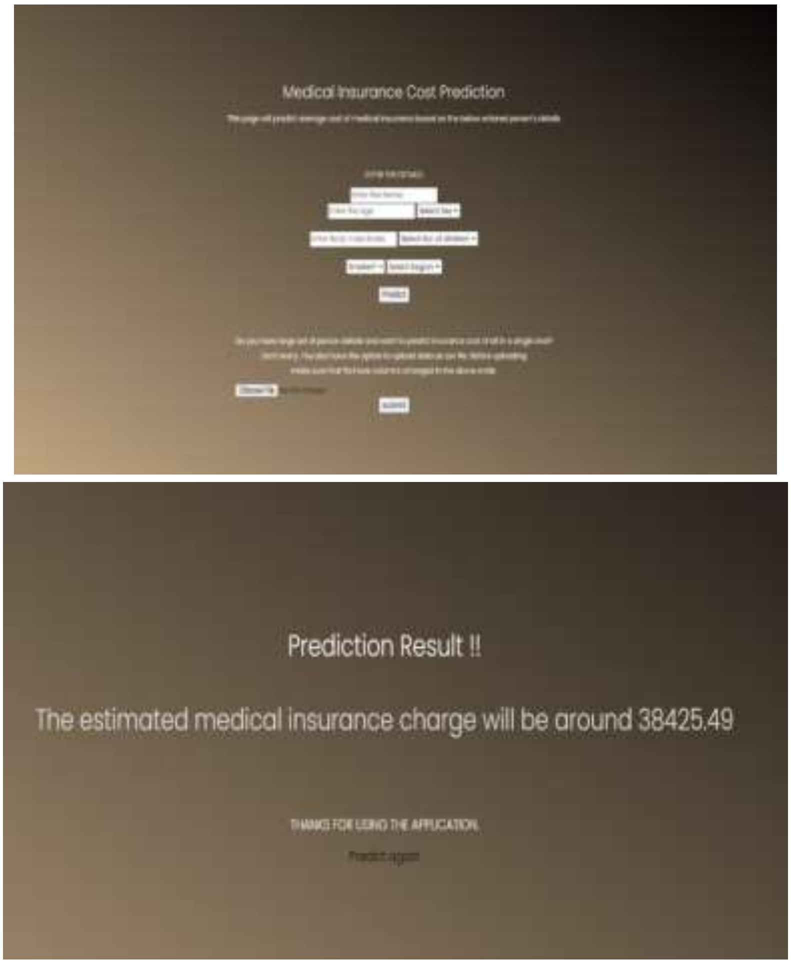
```
pipelines = [pipe1, pipe2, pipe3, pipe4,pipe5]
pipe_dict = {0: 'Linear Regression', 1: 'Lasso', 2: 'Ridge',3:'Decision Tree',4:'Random Forest'}
for pipe in pipelines:
    pipe.fit(X_train, y_train)
for i,model in enumerate(pipelines):
    print("{} Test Accuracy:{}".format(pipe_dict[i], model.score(X_test,y_test)))
```

```
Linear Regression Test Accuracy:0.7675432393370196
Lasso Test Accuracy:0.7675611571489156
Ridge Test Accuracy:0.7675874244113005
Decision Tree Test Accuracy:0.7088575498906655
Random Forest Test Accuracy:0.8541238481806733
```

```
for i,model in enumerate(pipelines):
    print("{} Train Accuracy:{}".format(pipe_dict[i], model.score(X_train,y_train)))
```

```
Linear Regression Train Accuracy:0.7443271565246132
Lasso Train Accuracy:0.7443253454766152
Ridge Train Accuracy:0.7442924539952123
Decision Tree Train Accuracy:1.0
Random Forest Train Accuracy:0.9753310863030313
```

Fig.6. Performance of different algorithms



Medical Insurance Cost Prediction

Prediction Result !!

The estimated medical insurance charge will be around 38425.49

THANKS FOR USING THE APPLICATION.

Fig.7. Output

## 7. Conclusion:

The accuracy of random forest regression is 87.006% (Stucki, Finland, 2019). Research has shown that the amount of data used for training affects accuracy, with a larger training size producing better results (Kenward, J.A., 2019). The accuracy of the linear regression algorithm is 76.75%. The decision tree algorithm is 70.88%. It can solve both classification and regression problems.

## 8. Future Work:

It would be useful to test at least one million records in the future to evaluate the scalability of the system. Distributed systems such as Spark and Hadoop can be used to process big data and improve system performance. Currently, the algorithm is being trained and tested using thousands of data sets (Donald W. Marquardt, 2012).

### 9. References:

[1] Gupta, S., & Tripathi, P. (2016, February). A leading trend of data analytics with health insurance in India. In 2016 International Conference on Innovation and Challenges in CS (ICICCS-INBUSH) (pp. 64-69). IEEE.

[2] Yerpude, P., Gudur, V.: Prediction modeling of crime

dataset using data mining. Int. J. Data Min. Knowl. Manage. Process (IJDKP) 7(4) (2017)

[3] Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Prediction vehicles insurance claims using telematics data—

XGBoost versus logistic regression. Risks, 7(2), 704. Singh, R., Ayyar, M. P., Pavan, T. S., Gosain, S., & Shah, R. R. (2019, September). Automating Vehicle Car Insurance Claims Using Deep Learning Techniques. In 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM) (pp. 199-207). IEEE.

[4] Stucki, O. (2019). Predicting the customer churn with machine learning methods: case: private insurance customer data.

[5] Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for not present data in epidemiological and clinical research: potential and pitfalls. Bmj, 338.