Health Insurance Amount Prediction using Machine Learning



A Minor Project Report

in partial fulfillment of the degree

**Bachelor of Technology**

in

**Computer Science & Engineering**

**By**

19K41A0573 K. Sudeeptha

19K41A05C0 B. Jamini

19K41A0571 K. Avinash

20K45A0509 Shaik Afsar

**Under the Guidance of**

**Dr. D. Kothandaraman**

**Submitted to**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

# S.R.ENGINEERING COLLEGE(A),ANANTHASAGAR, WARANGAL

**(Affiliated to JNTUH, Accredited by NBA)**

**May, 2021.**



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**CERTIFICATE**

This is to certify that the Minor Project Report entitled “Health Insurance Amount Prediction using Machine Learning” is a record of bonafide work carried out by the student(s) K. Sudeeptha, B. Jamini, K. Avinash, Shaik Afsar bearing Roll No(s) 19K41A0573, 1941AO5C0, 19K41A0571, 20K45A0509 during the academic year 2021-22 in partial fulfillment of the award of the degree of ***Bachelor of Technology*** in **Computer Science & Engineering** by the Jawaharlal Nehru Technological University, Hyderabad

**Supervisor Head of the Department**

**External Examiner**

**ACKNOWLEDGEMENT**

* I would like to express my gratitude to our guide, D. Kothandaraman sir, who guided me throughout this project. I could say that this project would not have achieved this accuracy without his assistance. Sir has been our great support in helping gathering the dataset and sorting out complex errors of our code and enhancing our understanding of the project.
* I would also like to thank V. Venkatraman sir, who introduced us to the magnificent python language and made us familiar with machine learning of Neural Networks.
* I would like to thank our HOD, Shashikala mam, who has ensured a required importance in slotting time for minor project and enabled us to have interaction with our guides and helped with any queries arising in between.
* I also would like to express gratitude to my Team – B11 comprising K.Sudeeptha, B.Jamini, K.Avinash, Shaik Afsar who have worked as a fruitful and healthy team with all productive thoughts and effort with good co-ordination.
* I would also like to express gratitude to all the authors of the websites, research papers, publications, technicians I have referred to in order to gain knowledge for this project.
* Last but not the least, I would also like to express thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

**ABSTRACT**

 Insurance is a policy that eliminates or decreases loss costs occurred by various risks. Various factors influence the cost of insurance. These considerations contribute to the insurance policy formulation. Machine learning (ML) for the insurance industry sector can make the wording of insurance policies more efficient. This study demonstrates how different models of regression can forecast insurance costs. And we will compare the results of models, for example, Multiple Linear Regression, Random Forest Regression, Decision tree.

**CONTENTS**

1. INTRODUCTION 6
   1. EXISTING SYSTEM 7
   2. PROPOSED SYSTEM 7
2. LITERATURE SURVEY 8
   1. RELATED STUDY
   2. SYSTEM STUDY
3. DESIGN 9
   1. REQUIREMENTS SPECIFICATIONS
   2. UML DIAGRAMS
4. IMPLEMENTATION
   1. MODULES 10
   2. OVERVIEW OF TECHNOLOGY 14
5. TESTING
   1. TEST CASES 16
   2. TEST RESULTS 17
6. RESULTS 19
7. CONCLUSION 20
8. FUTURE SCOPE 20

BIBLIOGRAPHY 21

1. **INTRODUCTION**

We are on a planet full of threats and uncertainty. People, households, companies, properties, and property are exposed to different risk forms. And the risk levels can vary. These dangers contain the risk of death, health, and property loss or assets. Life and wellbeing are the greatest parts of people's lives. But, risks cannot usually be avoided, so the world of finance has developed numerous products to shield individuals and organizations from these risks by using financial capital to reimburse them. Insurance is, therefore, a policy that decreases or removes loss costs incurred by various risks. Concerning the value of insurance in the lives of individuals, it becomes Important for the companies of insurance to be sufficiently precise to measure or quantify the amount covered by this policy and the insurance charges which must be paid for it. Various variables estimates these charges

Each factor of these is important. If any factor is omitted when the amounts are computed, the policy changes overall. It is therefore critical that these tasks are performed with high accuracy. As human mistakes are could occur, insurers use people with experience in this area. They also use different tools to calculate the insurance premium. Ml is beneficial here. Ml may generalize the effort or method to formulate the policy. these ml models can be learned by themselves. The model is trained on insurance data from the past. The requisite factors to measure the payments can then be defined as the model inputs, then the model can correctly anticipate insurance policy costs.

This decreases human effort and resources and improves the company's profitability. Thus the accuracies can be improved with ml. our objective is to forecast insurance charges in this article. The value of insurance fees is based on different variables. As a result, insurance fees are continuous values. The regression is the best choice available to full fill our needs. We use multiple linear regression in this analysis since there are many independent variables used to calculate the dependent(target) variable. For this study, the dataset for cost of health insurance is used.

Pre-processing of the dataset done first. then we trained regression models with training data and finally evaluated these models based on testing data. In this article, we used several models of regression, for example, multiple linear regression, decision tree (cart), and Random Forest Regression. It is found that the Random Forest Regression provides the highest accuracy with an r-squared value of 87.07829. the key reason for this study is to include a new way of estimating insurance costs.

* 1. **EXISTING SYSTEM**

Health insurance is a necessity nowadays, and almost every individual is linked with a government or private health insurance company. Factors determining the amount of insurance vary from company to company. Also people in rural areas are unaware of the fact that the government of India provide free health insurance to those below poverty line. It is very complex method and some rural people either buy some private health insurance or do not invest money in health insurance at all. Apart from this people can be fooled easily about the amount of the insurance and may unnecessarily buy some expensive health insurance.

Our project does not give the exact amount required for any health insurance company but gives enough idea about the amount associated with an individual for his/her own health insurance.

* 1. **PROPOSED MODEL**

Prediction is premature and does not comply with any particular company so it must not be only criteria in selection of a health insurance. Early health insurance amount prediction can help in better contemplation of the amount

needed. Where a person can ensure that the amount he/she is going to opt is justified. Also it can provide an idea about gaining extra benefits from the health insurance.

1. **LITREATURE SURVEY**
   1. **RELATED STUDY**

Machine learning is helpful for a variety of situations. The prediction of dependent variable values from independent variables is one of the uses of this methodology. The objective of the study is to predictive the insurance cost based on age, BMI, child number, the region of the person living, Gender, and whether a client is smoking or not. These features contribute to our target variable prediction of insurance costs.

For the measurement of the cost of insurance, several regression models are applied in this study. The dataset is split into two sections. One part for model training and the other part for model evaluation or testing. In this study, the data set is separated into two-part the first part is called training data and the second called test data, training data makes up about 80 percent of the total data used, and the rest for test data. Every one of these models is trained with the training data part and then evaluated with the test data.  And we used  Mean absolute error (MAE), root mean squared error (RMSE) and R-squared As a standard for evaluating these models.

* 1. **SYSTEM STUDY**

For the measurement of the cost of insurance, several regression models are applied in this study. The dataset is split into two sections. One part for model training and the other part for model evaluation or testing. In this study, the data set is separated into two-part the first part is called training data and the second called test data, training data makes up about 80 percent of the total data used, and the rest for test data. Every one of these models is trained with the training data part and then evaluated with the test data.  And we used  Mean absolute error (MAE), root mean squared error (RMSE) and R-squared As a standard for evaluating these models.

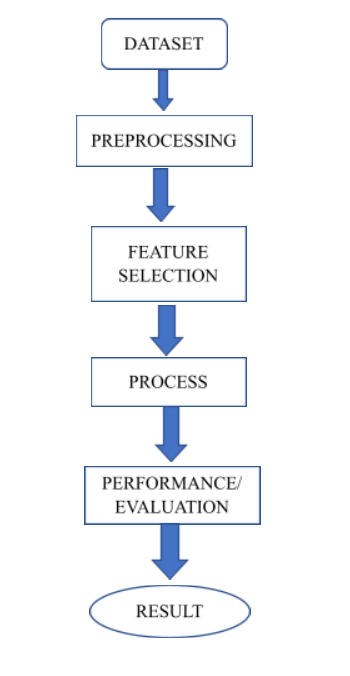
1. **DESIGN**
   1. **REQUIREMENTS SPECIFICATION**

**Hardware Requirements**

* **System** : Pentium 4, Intel Core i3, i5, i7 and 2GHz Minimum
* **RAM** : 4GB or above
* **Hard Disk** : 10GB or above
* **Input**  : Keyboard and Mouse
* **Output** : Monitor or PC

**Software Requirements**

* **OS** : Windows 8 or Higher Versions
* **Platform** : Visual Studio Code
* **Program Language** : Python 3.2
  1. **UML DIAGRAMS**



1. **IMPLIMENTATION**
   1. **MODULES**

Segregating the project into different modules can be useful to achieve clarity while implementing each module. By this process we can have a clear idea and when we have changes in a particular module it stays in the respective module and the process can be easy.

**Data Collection**

We have collected data from Kaggle website. We have 6 attributes which will help us to prepare a model for predicting the insurance amount. It has a total of 1338 data information. Given below are the input and output features of the data.

**Input Features**

* Age: Age in years
* Gender: Gender (1 = male; 0 = female)
* BMI: Body Mass Index, objective index of body weight (kg/ m^2) using the ratio of height to weight
* Children: Number of children dependent.
  + Value between 0-5
* Smoker: Smoking or not(1=yes,0=no)
* Region: Area they belong to.
  + Value 1: southwest
  + Value 2: southeast
  + Value 3: northwest
  + Value 4: northwest

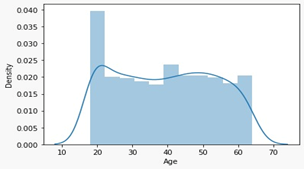
**Output Features**

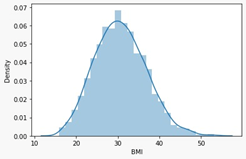
* The output will be the prediction value.

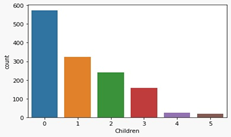
The data set is separated into two-part the first part called training data, and the second called test data; training data makes up about 80 percent of the total data used, and the rest for test data The training data set is applied to build a model as a predictor of medical insurance cost year and the test set will use to evaluate the regression model.

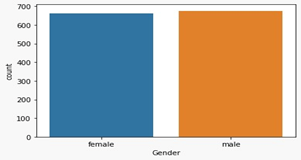
**Data Visualization using Graphs**

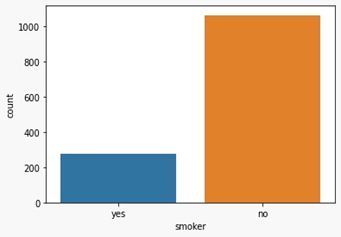
For the better and deep understanding of the data we use graphs. The below are the graphs which we have obtained from the data. These help us to understand the data efficiently.

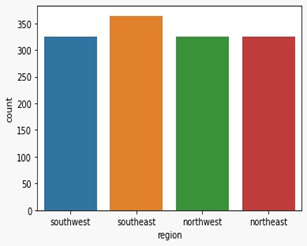












**Data Preprocessing**

Data pre-processing is a technique that is used to convert raw data into a clean dataset. The data is gathered from different sources is in raw format which is not feasible for the analysis. Preprocessing for this approach takes 4 simple yet effective steps. They are:

1.Attribute selection

2. Cleaning missing values

3. Training and Test data

4. Feature Scaling

5. Evaluating the data

|  |  |
| --- | --- |
| Age | Numeric [18 to 64; mean=39.207025] |
| BMI | Numeric [15.96 to 53.13; mean=30.66] |
| Gender | Male or Female |
| Children | Numeric [0 to 5; mean=1.0949] |
| Smoker | Yes or no |
| Region | southwest, southeast, northwest, or northeast |
| Charges (target variable) | Numeric [1121.873 to 63770.428; mean=13270.422] |

The dataset includes seven variables, as shown in table 1. every one of these attributes has some contribution to estimate the cost of the insurance, which is our dependent variable. In this stage, the data is scrutinized and updated properly to efficiently apply the data to the ML algorithms. First of all, all unknown values are cleaned. The unknown numerical values are replaced with the mean. The target variable (charges) would then be examined.

**Correlation Matrix**

A correlation matrix is simply a table that displays the correlation. The measure is best used in variables that demonstrate a linear relationship between each other. The fit of the data can be visually represented in a scatterplot. coefficients for different variables.

***How it is calculated?***

A correlation matrix is a table showing correlation coefficients between sets of variables. Each random variable (Xi) in the table is correlated with each of the other values in the table (Xj)... The diagonal of the table is always a set of ones because the correlation between a variable and itself is always 1. Let's perform the Correlation matrix to understand the relation between the dependent variable and the independent variable and within the independent variable.



**Fig-3**: Correlation Matrix

Enough EDA performs on the data to evaluate the dataset and gather knowledge about the data. Let's perform some Machine Learning model and Experimentation to create a model that helps us to achieve our goal we state in the problem definition.

* 1. **OVERVIEW OF TECHNOLOGY**

We have used different machine learning models to solve our problem.

* Multiple Regression
* Random Forest
* Decision Tree

**Multiple Regression**

Multiple regression is a statistical technique that can be used to analyse the relationship between a single dependent variable and several independent variables. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the values of the single dependent value.

In practice, we often have more than one predictor. For example, with the data set used in this study, we may wish to understand if independent variables (6 independent variables), (linearly) related to the dependent variable (charges). this is referred to as the multiple linear regression (MLR) model. An MLR model with independent features and Y results can be calculated as in the following equation

  X1 X2 X3   Xi

Y=a0X0+a1X1+a2X2+…….+atXt+u

In the above equation, u is the residual regression while a is the weight of each independent          variable or parameter assigned.

**Random Forest**

Random forest is used for both regression and classification-based applications. This algorithm is flexible and easy to use. Most of the times this algorithm gives accurate results even without hyper tuning the parameters. It builds many decision trees which on merging forms as a forest. While building the decision trees, adds more randomness to the model. This algorithm searches for the best feature in the random subset of features, which results in the formation of a better model.

With the help of the sklearn library, we can measure the relative importance of each feature in prediction. By finding the feature importance, we can drop the features that have less importance in the prediction process. The main limitation of random forest is that many trees can make the algorithm too slow and ineffective for real-time predictions. In general, these algorithms are fast to train but quite slow to create predictions once they are trained. A more accurate prediction requires more trees, which results in a slower model. In most real-world applications, the random forest algorithm is fast enough but there can certainly be situations where run-time performance is important and other approaches would be preferred.

  With the aid of the following step we can understand how the Random Forest algorithm works.

Step 1 − First, select random samples from a particular dataset.

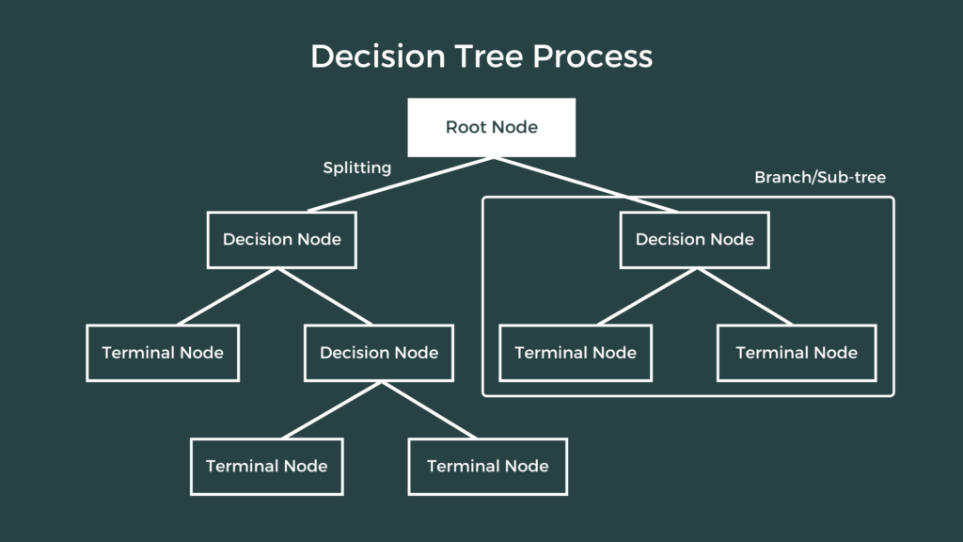
Step 2 - Next, for each sample this algorithm will build a decision tree. Then every decision tree will predict the result.

Step 3 − Every predicted result will be voted in this step.

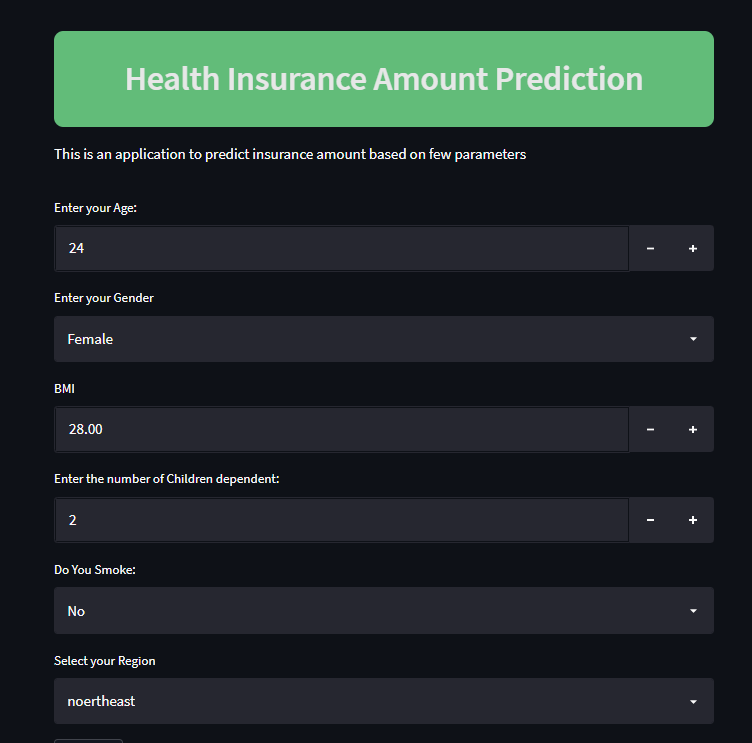
Step 4 − Finally, the final prediction result will be selected as the most voted prediction result.

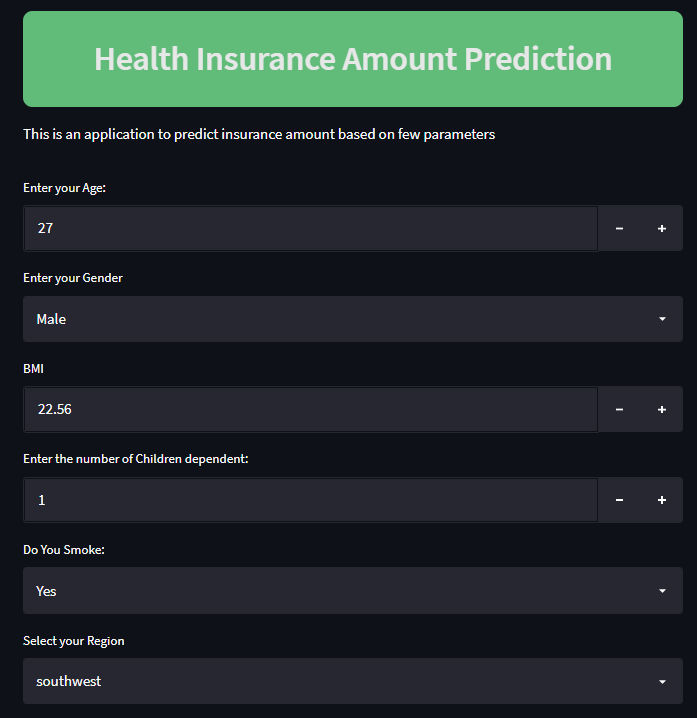
**Decision Tree**

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called **root node**. Decision trees can handle both categorical and numerical data.

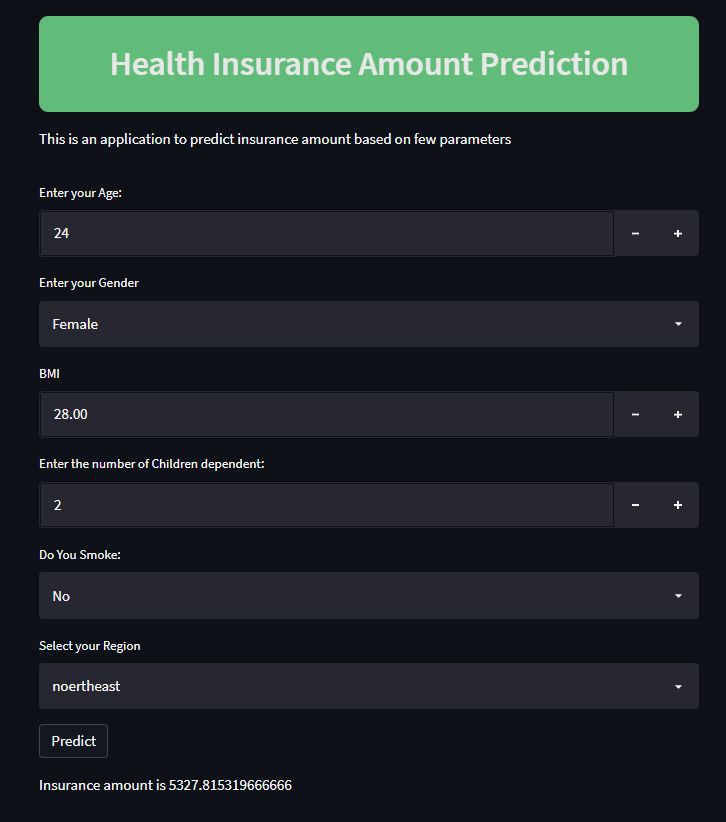


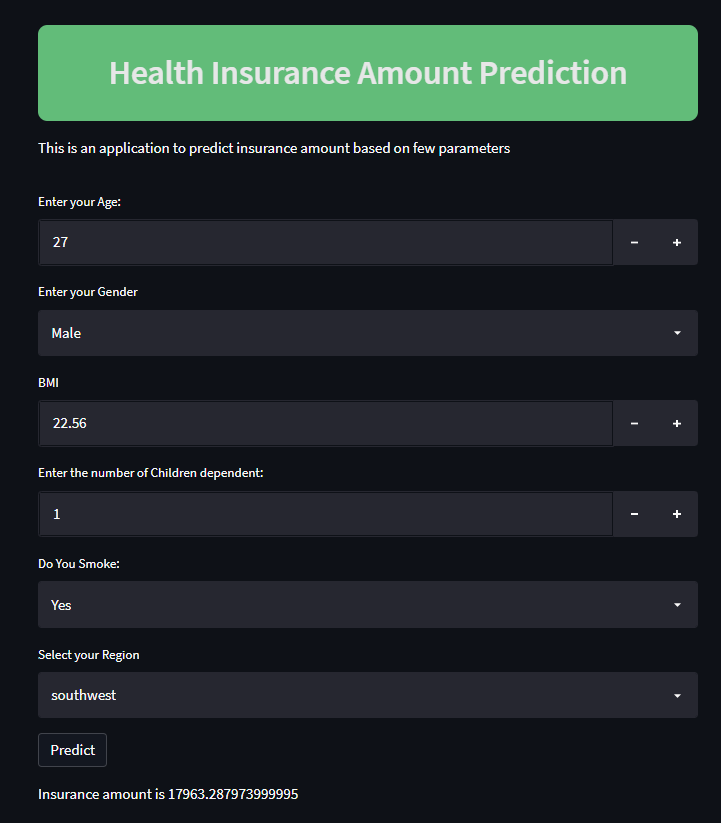
1. **TESTING**
   1. **TEST CASES**

****

****

* 1. **TEST RESULTS**

****

****

1. **RESULTS**

We have used 3 different machine learning algorithms such as Multiple Regression, Random Forest and Decision Tree. Random Forest model gave the highest accuracy of 84% where Multiple Regression giving 77.1% and Decision Tree giving 77.4%.

|  |  |  |
| --- | --- | --- |
| **Model** | **Training Accuracy** | **Testing Accuracy** |
| **Multiple Regression** | 74.36639 | 77.14801 |
| **Random Forest** | 93.36814 | 87.07829 |
| **Decision Tree** | 97.00801 | 77.44010 |

We have to consider the testing accuracy. As we have got the highest accuracy for the random forest model we will use that algorithm for preparing our model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Error/Model** | **Multiple Regression** | **Random Forest** | **Decision Tree** |
| **Root mean square error** | 0.09678 | 0.07279 | 0.0699 |
| **Mean square error** | 0.0093 | 0.0052 | 0.0048 |
| **Mean absolute error** | 0.0686 | 0.0366 | 0.03649 |

1. **CONCLUSION**

 The research uses various machine learning regression models and deep neural networks to forecast charges of health insurance based on specific attributes, on medical cost personal data set from Kaggle.com. The Random Forest Regression offers the best efficiency, with an RMSE value of 0.07291, an MAE value of 0.036628, and an accuracy of 87.078291. Random Forest Regression can therefore be used in the estimation of insurance costs with better performance than other regression models. Forecasting insurance costs based on certain factors help insurance policy providers to attract consumers and save time in formulating plans for every individual. Machine learning can significantly minimize these individual efforts in policymaking, as ML models can do cost calculation in a short time, while a human being would be taking a long time to perform the same task. This will help businesses improve their profitability. The ML models can also manage enormous amounts of data.

1. **FUTURE SCOPE**

In the future we can easily extend the model by using different machine learning models. Not only the machine learning models we can use other prediction models also. In the present model we have usen only a few attributes. We can also increase change the attributes and extend the project.

**BIBLIOGRAPHY**

1. Gupta, S., & Tripathi, P. (2016, February). An emerging trend of big data analytics with health insurance in India. In 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH) (pp. 64-69). IEEE.
2. Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus
3. logistic regression. Risks, 7(2), 70
4. Ms.sonal patil, Mr.Mayur Agrawal, Ms.Vijaya R. Baviskar “Efficient Processing of Decision Tree using ID3 & improved C4.5 Algorithm”, International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1956-1961.
5. Stucki, O. (2019). Predicting the customer churn with machine learning methods: case: private insurance customer data.
6. Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj, 338.
7. Van Buuren, S. (2018). Flexible imputation of missing data. CRC press.
8. Kowshalya, G., & Nandhini, M. (2018, April). Predicting fraudulent claims in automobile insurance. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1338-1343). IEEE.
9. Kayri, M., Kayri, I., & Gencoglu, M. T. (2017, June). The performance comparison of multiple linear regression, random forest by using photovoltaic and atmospheric data. In 2017 14th International Conference on Engineering of Modern Electric Systems (EMES) (pp. 1-4). IEEE.
10. Denuit, Michel & Hainaut, Donatien & Trufin, Julien. (2019). Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions. 10.1007/978-3-030-25820-7.
11. Breiman, Leo. 2001. ―Random Forests. ‖ Machine Learning 45 (1). Springer: 5–32.
12. Aler, R., Galván, I.M., Ruiz-Arias, J.A., Gueymard, C.A. (2017). Improving the separation of direct and diffuse solar radiation components using machine learning by gradient boosting. In Solar Energy vol. 150, pp. 558-569.
13. Volkovs, M., Yu, G. W., & Poutanen, T. (2017). Content-based neighbor models for cold start in recommender systems. In Proceedings of the Recommender Systems Challenge 2017 (pp. 1-6).