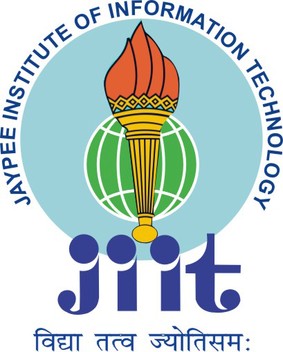
# OPEN SOURCE SOFTWARE

# PROJECT REPORT



**TITLE : Medical Insurance Cost Prediction Using Machine Learning.**

**Submitted to: Mrs. Sonal**

**Team Members**

|  |  |  |
| --- | --- | --- |
| **NAME** | **ENROLL** | **BATCH** |
| **Pishishta Singh** | **21103215** | **B8** |
| **Tejas Trived** | **21103250** | **B8** |
| **Mohd Yasir Khan** | **21103251** | **B8** |

# Chapter-1:

**Introduction and Motivation**

People seek out various means of protecting themselves as well as their families when life is most at risk. Working hard and earning money satisfies merely their basic comforts of day-to-day existence. Risk cannot often be eliminated. By leveraging the money they have invested to safeguard them through financial support, financial companies have created a variety of solutions to shield both individuals and businesses from a variety of hazards. One cannot afford to instantly spend a large sum of money in times of emergency or unforeseen health concerns. Therefore, it makes sense to use insurance to save money for the future. The loss caused by various hazards is reduced or completely eliminated through insurance policies. Today, health insurance is especially important. Every person is either a part of the public or private insurance system.

The amount covered by each policy that the customer must pay for must be precisely measured by the insurance provider. Every insurance provider has a different policy with unique terms and conditions. They evaluate a variety of factors when determining the policy amount. The overarching policy itself changes if any of the contributing components change. Therefore, it is essential to calculate the policy's value accurately. Due to the sensitive nature of the information, skilled individuals are employed for this purpose. These people compute the insurance premium using a variety of tools.

To generalise the approach to calculating the policies, machine learning methods are applied. Using historical insurance data, these ML algorithms train the computer. As input parameters for the model, the variables that are used to measure consumer payments are described. The customer's cost for health insurance is predicted by the machine learning algorithm or model using past experience. A highly accurate model has been developed. Effort and time requirements are decreased. The term "continuous" refers to the fact that insurance is based on numerous variables or conditions. The most effective method for computing the insurance premium when dealing with these continuous variables is regression modelling.

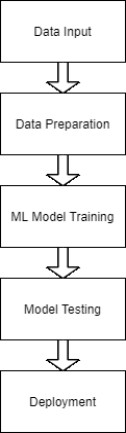
# Problem Statement

To build a model for Medical Insurance Cost Prediction Using Machine Learning. In this project machine learning algorithms such as multiple linear regression and random forest regressor have been used.

# Literature review

Author Muhammad Arief et al. [1], used XGBoost machine learning model to find the insurance claim frequency and severity which is considered to be most important one to find claims faster with accurate result. Sahar F. Sabbeh [1-3], compared all the machine learning models with their performance result to yield customer retention. Among the other models, the derived result showed that Random Forest and Ada boost algorithms perform better and attained 96% accuracy. Cunningham et al. [4-5], presented a paper to find the classification using K nearest neighbour. Kayri et al. [6-8], compared the results of three algorithms such as Random Forest classifier, Multiple Linear regression, and Artificial Neural Network for atmospheric data. Leo Breiman [9] gave a detailed study for prediction using Random Forest algorithm. E. Wang et al., and Singh et al., suggested the techniques to calculate the claims for health insurance and car insurance respectively.

Machine Learning is defined as the collection of algorithms and statistical models to train the machine with previous experience and patterns without being programmed explicitly. The major goal of Machine Learning is to make computers learn by themselves. It has been proved to be one of the most familiar and useful technological advancements of the past decades. It allows companies to adapt digital transformation and grow with the new age of technology. Fig. 2. shows the steps in Machine Learning.



### Supervised Learning

Fig. 2. Steps in Machine Learning

Supervised learning algorithms create a mathematical model according to a set of data that contains both the inputs and the desired outputs. Usually a random part of data is selected from the complete dataset known as training data, or in other words a set of training examples. Training data has one or more inputs and a desired output, called as a supervisory signal. What’s happening in the mathematical model is each training dataset is represented by an array or vector, known as a feature vector. A matrix is used for the representation of training data. Supervised learning algorithms learn from a model containing a function that can be used to predict the output from the new inputs through iterative optimization of an objective function. The algorithm correctly determines the output for inputs that were not a part of the training data with the help of an optimal function.

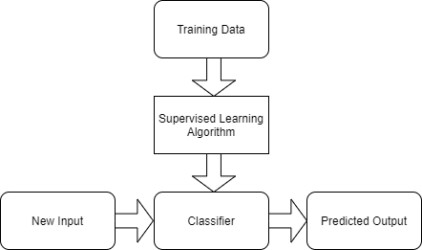


Fig. 3. Operation of Supervised Learning

### Unsupervised Learning

In this learning, algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. Test data that has not been labeled, classified or categorized helps the algorithm to learn from it. What actually happens is unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data. The main application of unsupervised learning is density estimation in statistics. Though unsupervised learning encompasses other domains involving summarizing and explaining data features also.

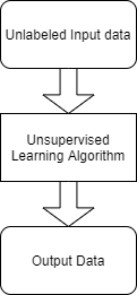


Fig. 4. Unsupervised Learning process

### Reinforcement Learning

In this type of Machine learning algorithm, agents play a major role. Agent learns from the experience instead of training the dataset as in Supervised Learning and decides what action to be taken. It is employed to observe the best possible path or behaviour in a specific situation. The input is the initial state or start of the model. Training is given with the input of the model. The model represents the state through which the user has to decide based on the output whether it is punishment or reward. The best solution is identified based on the maximum reward.

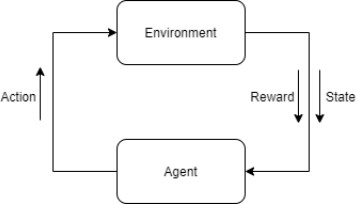


Fig. 3. Reinforcement learning

**REGRESSION**

Regression analysis allows us to quantify the relationship between outcome and associated variables. Many techniques for performing statistical predictions have been developed, but, in this project, three models - Multiple Linear Regression (MLR), Decision tree regression and Gradient Boosting Regression were tested and compared.

1. *Multiple Linear Regression*

Multiple linear regression can be defined as extended simple linear regression. It comes under usage when we want to predict a single output depending upon multiple input or we can say that the predicted value of a variable is based upon the value of two or more different variables. The predicted variable or the variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable) and the variables being used in predict of the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or regressor variables).

1. *Decision tree regression*

Regression or classification models in decision tree regression builds in the form of a tree structure. The dataset is divided or segmented into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision tree with decision nodes and leaf nodes is obtained as a final result. These decision nodes have two or more branches, each representing values for the attribute tested. Decision on the numerical target is

represented by leaf nodes. The topmost decision node corresponds to the best predictor in the tree called root node. Numerical data along with categorical data can be handled by decision trees.

1. *Gradient Boosting Regression*

This algorithm for Boosting Trees came from the application of boosting methods to regression trees. The basic idea behind this is to compute a sequence of simple trees, where each successive tree is built for the prediction residuals of the preceding tree. For predictive models, gradient boosting is considered as one of the most powerful techniques. Gradient boosting involves three elements:

* 1. An optimized loss function.
  2. An additive model to add weak learners to minimize the loss function.
  3. A weak learner to make predictions

### Influence of Machine Learning in the Field of Insurance

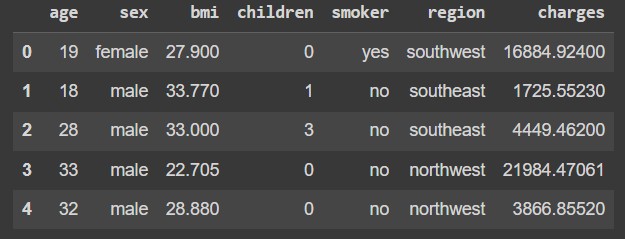
Insurance companies use Machine learning algorithms to improve their working efficiency, to gain customer service, and also to detect fraudulent occurrences. The following five ways define how Machine Learning has changed the insurance industry.

* + - Process Automation: The insurance industry comprises many legal requirements and regulations. They tend to process thousands of claims, and respond to customer queries at all times. Machine learning automatically processes the claims and reduces the human effort. Risk
    - Classification: Insurance companies can take greater risk and their only aim is achieving good profit. While evaluating the risk associated with properties, classification and calculation are done based on historical data. Machine learning offers a number of methods and tools for classifying risks and provides accurate and best pricing models which reduce loss ratios in the insurance company.
    - Insurance Underwriting: Machine learning provides underwriting services in the area of health insurance. It provides insurance coverage for diseases, disability and death. Machine learning tools help to analyse the vast data such as insurance claims, customer membership, medical records and so on. It helps to process the data efficiently which provides cost reduction, and detection of frauds leads to business improvement of healthcare insurance.
    - CLV prediction: Customer Lifetime Value (CLV) represents the value of a customer to the company. It is derived from the difference value between

the revenue and expenses of the company. Insurance companies have to predict the customer behaviour data and also access the customer’s profitability for the Insurer by using Machine Learning tools.

# Chapter-2: Exploratory Data Analysis Dataset Description

Data for this research came primarily from D'marco, a Kaggle user. The collection consists of 1338 records with 7 attributes. The following attributes are listed in Fig. 1: age, gender, bmi, children, smoker, and charges. The data

was organised and kept in a csv file.

The dataset is unsuitable for direct regression.In order to use the data with different regression techniques, dataset cleaning becomes crucial.Not every attribute in a dataset has an effect on the prediction.

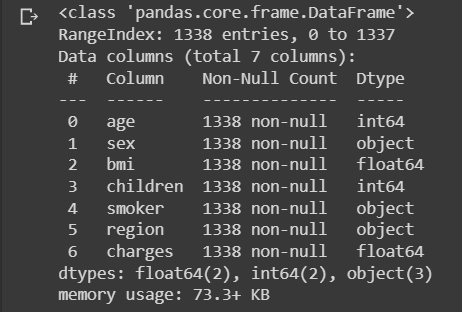
Some attributes even reduce accuracy, therefore it becomes necessary to delete them from the features of the code. Getting rid of these characteristics not only helps with accuracy but also with performance in general and speed.

In health insurance many factors such as pre-existing body condition, family medical history, Body Mass Index (BMI), marital status, location, past insurances etc affects the amount. According to our dataset, age and smoking status has the maximum impact on the amount prediction with smoker being the one attribute with maximum effect. Children attribute had almost no effect on the prediction, therefore this attribute was removed from the input to the regression model to support better computation in less time.

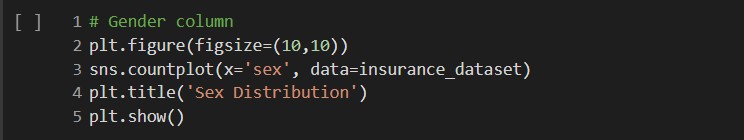
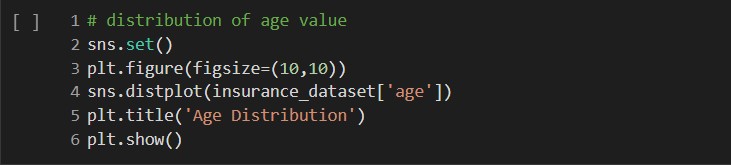
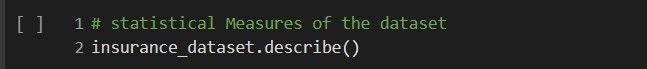
# Exploratory Data Analysis

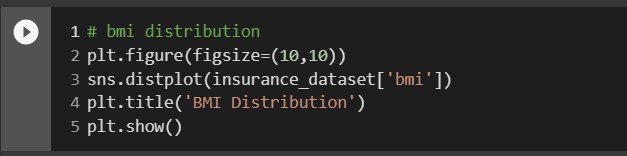


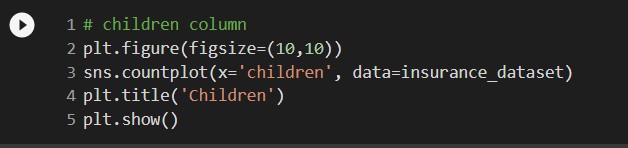
## Result:

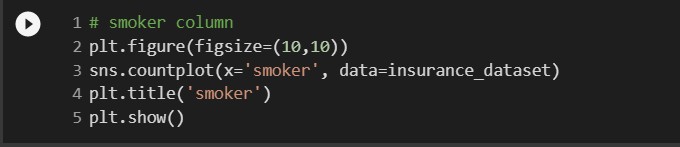


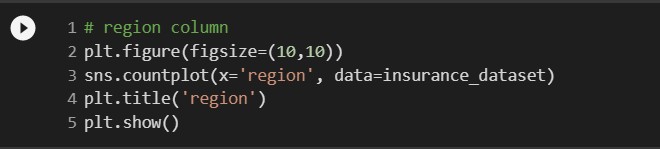
**Observation**: As we can see, there are a total 1338 records with 7 attributes each consisting of non-null values.

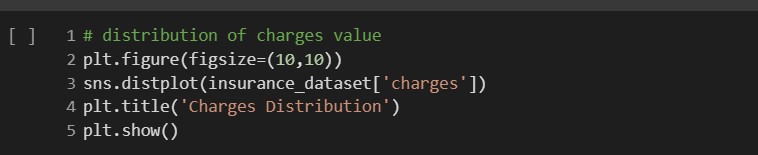




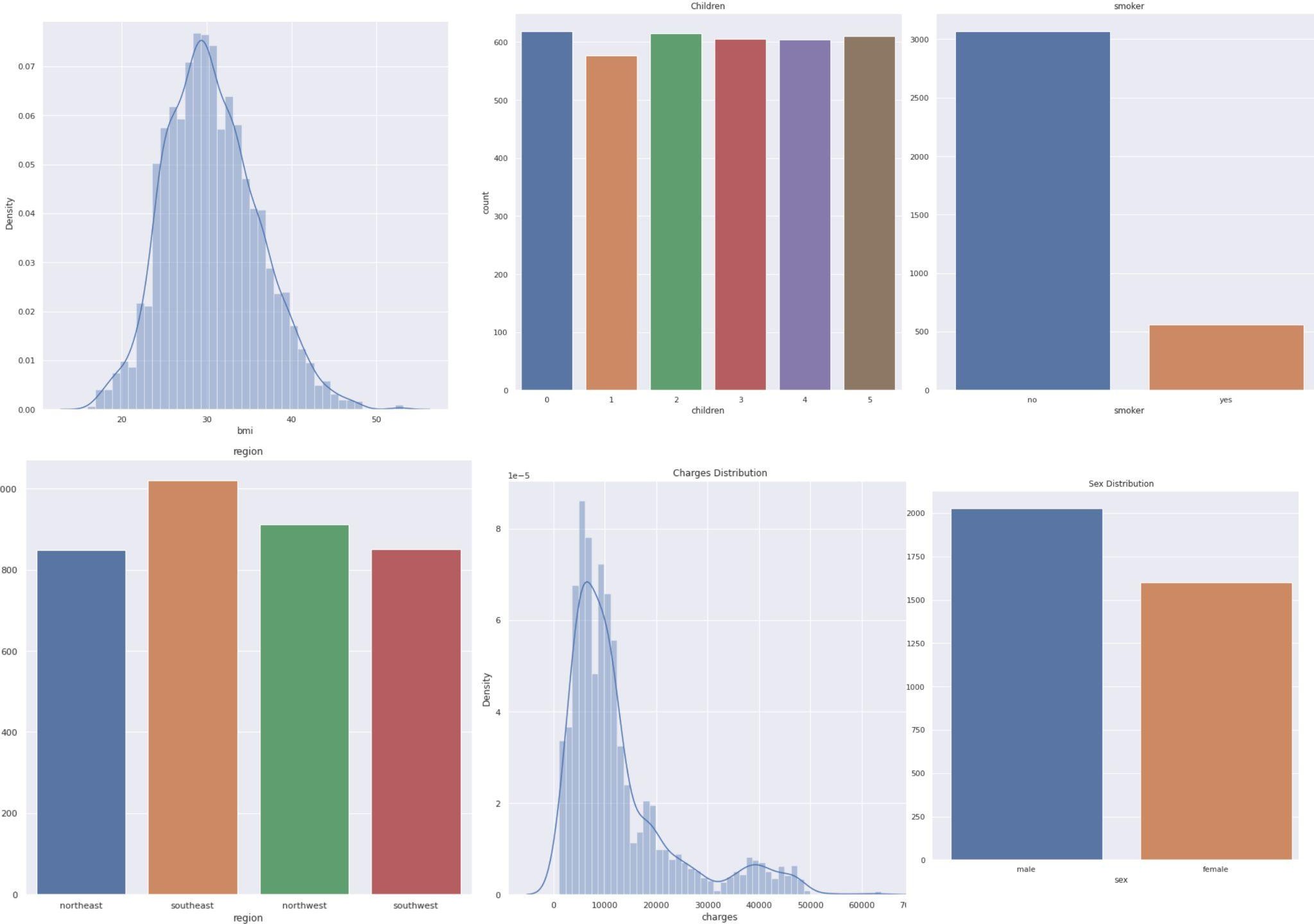




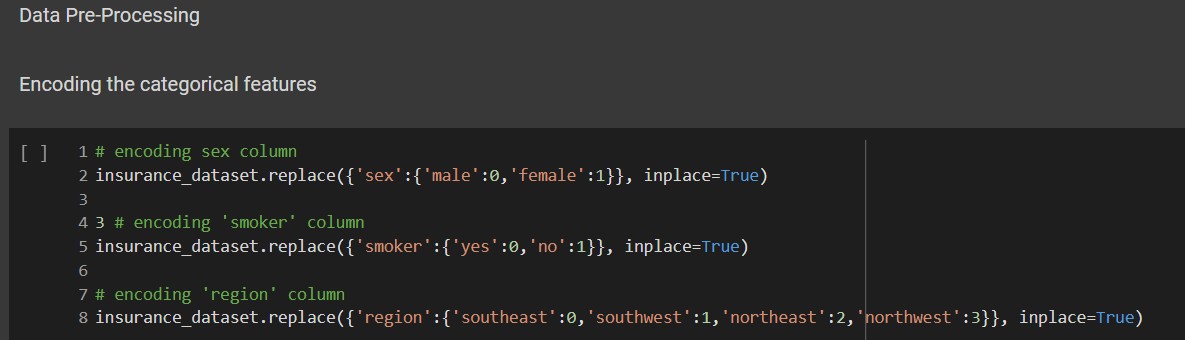


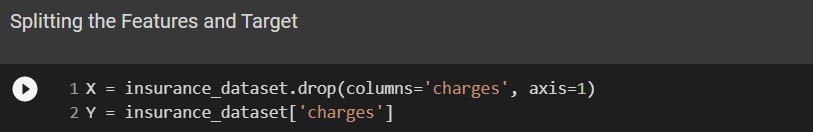


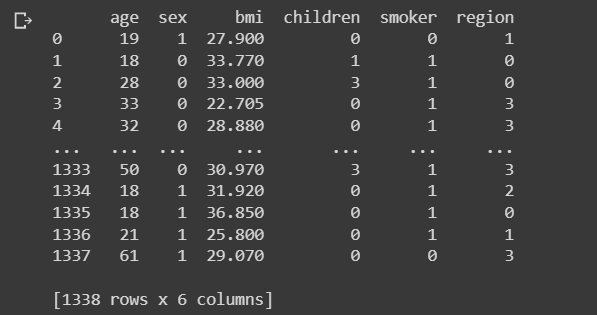
## Result:

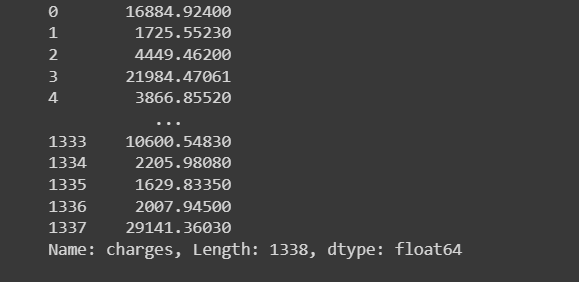


**Chapter-3: Preprocessing**





**Result:-**



**Observations:-** In the first code snippet, we are using the replace method to replace the character values with numeric values. While in the second snippet, drop method has been used to split the features and target.

# Chapter-4: Implementation Results and Analysis

## Implementation Set-up (software used)

In this section of the report, we are mentioning different tools and libraries which are used in our working model.

Google Colab

Colab is a free Jupyter notebook environment that runs entirely in the cloud. Most importantly, it does not require a setup and the notebooks that you create can be simultaneously edited by your team members

Python version

Python 3.7 is used for this project.

Python is a very useful programming language.It is object oriented and interpreted.It is a high level language.There are lots of in-built libraries in Python for machine learning purposes which we can use easily.

Windows Version

Colab and python 3 can be used in all the operating systems including Windows, iOS and Linux. It is most useful in Linux but can be used in windows

as well.It can be run on windows xp,vista,7,8 and the latest version windows 10 as well.

Tools & Libraries

* + - Numpy:- Used to support Panda frameworks.
    - Panda:- To Create Dataframe of the Image Pixel Values.
    - Sklearn:-Python Library used for machine learning and statistical modelling including classification.
    - Matplotlib:-Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. Create publication quality plots.

# Machine Learning Algorithms Used:-

Here we have used multiple machine learning algorithms to test and predict the medical insurance cost. R squared value is also calculated for each model to determine which of the used algorithms has the highest accuracy.

## Multi-Linear Regression

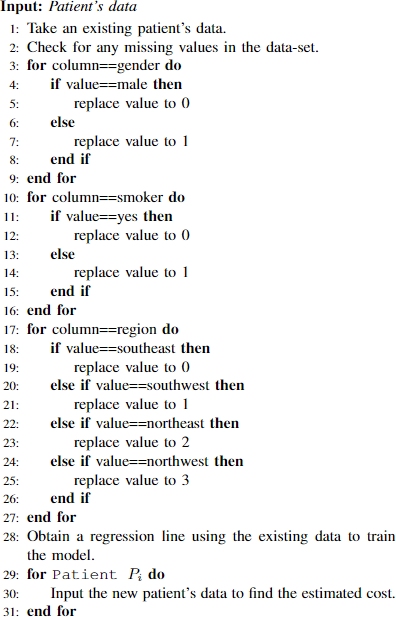
This is a more complex version of the simpler linear regression. For example, we use it when we want to predict the output of a single variable from a large number of inputs or when the expected value of one variable is based on the values of two or more other variables. When attempting to predict a variable's future value, we employ independent variables (also called predictors, explanatory variables, or regression coefficients) to help us make that prediction. Given the large number of independent variables, multiple linear regression has been applied in this case. The gender, smoker, and region columns in this table are encoded for calculating purposes. In column gender, 0 for male and 1 for female and in column smoker, 0 for yes and 1 for no. Similarly in the region, southeast:0, southwest:1, northeast:2, northwest:3.

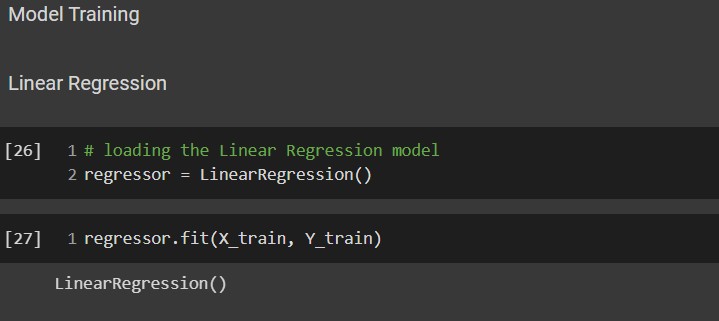
The mathematical equation for cost estimation,

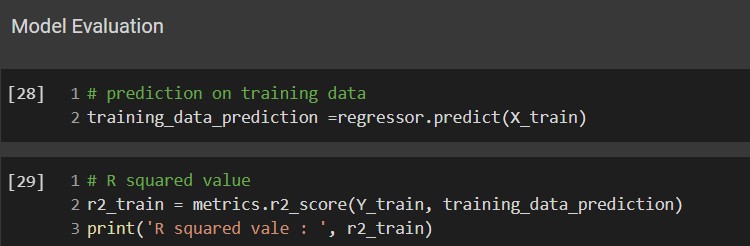
Y = αo+α1×x1+α2×x2+α3×x3+α4×x4+α5×x5+α6×x6 Y = Dependent variable and

x1, x2, x3, …… xn = multiple independent variables

The result Y will be the final cost of the insurance which can be further divided into premium subscriptions according to the beneficiary’s needs.1 age of the beneficiaries.

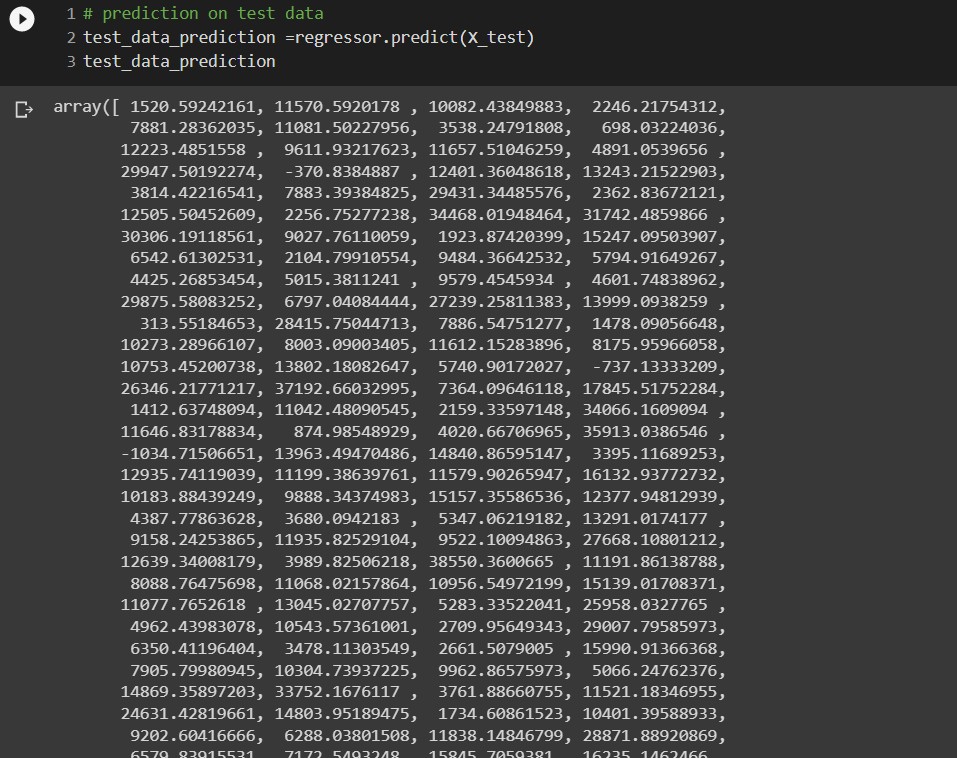




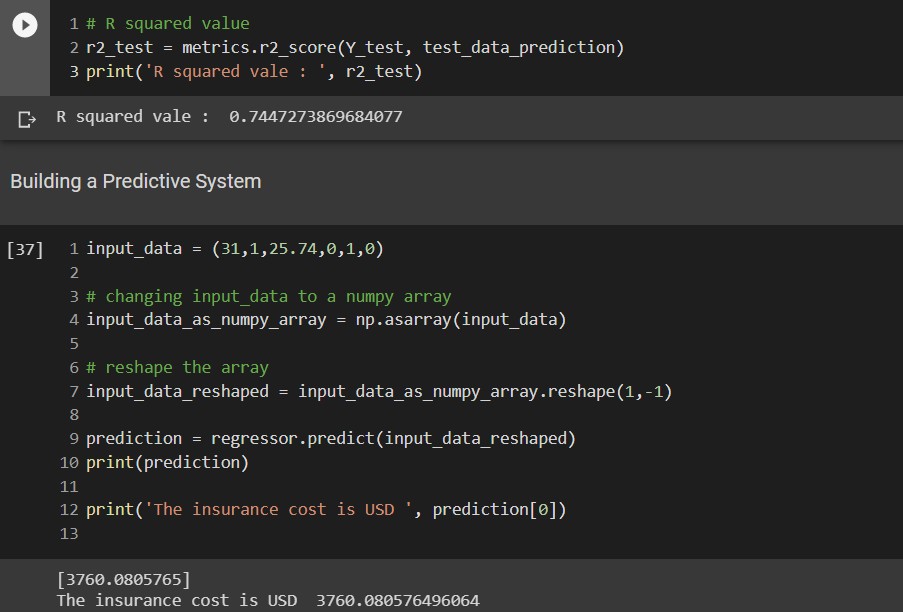


Result for training data and error computation:-





Result for testing data and error computation:



## Ridge Regression

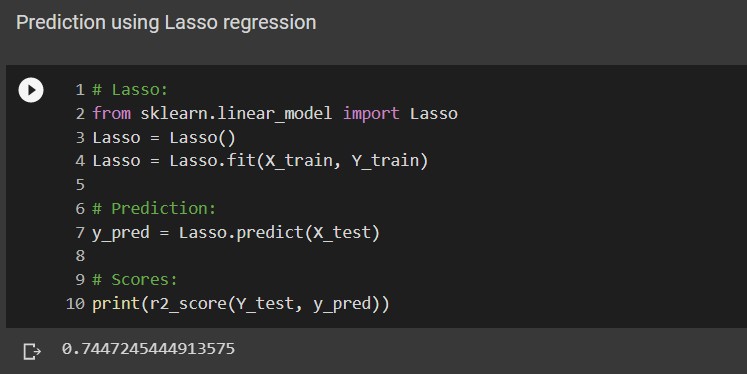
Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.



## Lasso Regression

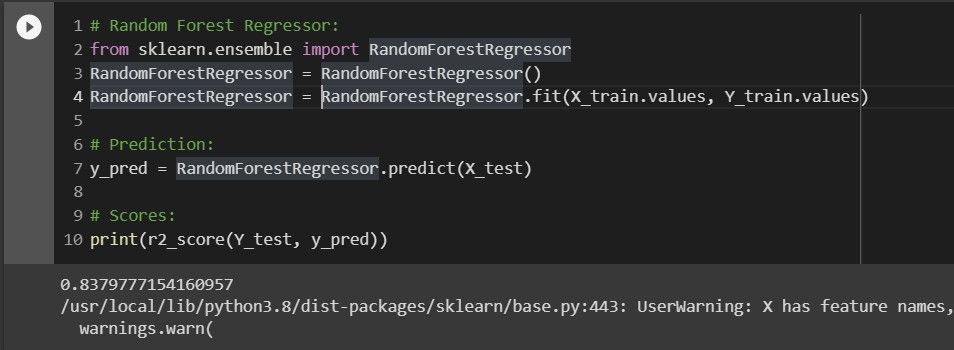
Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

Lasso Regression uses L1 regularization technique (will be discussed later in this article). It is used when we have more features because it automatically performs feature selection.

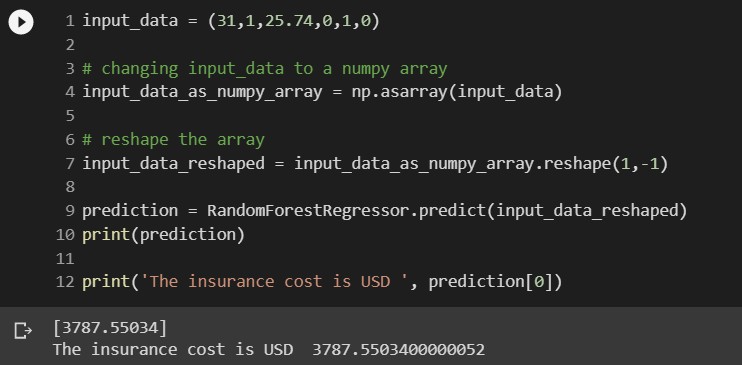


## Random Forest Regressor

The Random Forest Classifier algorithm is suitable for both classification and regression. The basic concept behind this algorithm is ensemble learning which means that combining multiple classifiers to solve a particular complex problem leads to the improvement in performance. Instead of creating a single decision tree, it creates multiple decision trees based on the dataset and the average is taken to predict the output.



Result:



# Chapter-5: Conclusion

An investigation into individual health insurance data is conducted using two regression models.Out of the four models used above, Random forest regressor has better accuracy. Any unneeded attribute was removed from each of the features. Premiums are determined by a person's health rather than the terms and conditions of another insurance provider. Some other algorithms can be employed to predict premiums based on data and improve accuracy. People and insurance companies can work together to deliver better and more health-focused coverage as a result of this.

# References

1. Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression.

Risks, 7(2), 70

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. C. C. a. A. Semanskee, "Analysis of UnitedHealth Group’s Premiums and Participation in ACA Marketplaces," 2016.
3. Prasad, K.S., Reddy, N.C.S. & Puneeth, B.N. A Framework for Diagnosing Kidney Disease in Diabetes Patients Using Classification Algorithms. SN COMPUT. SCI. 1, 101 (2020)
4. Varun K L Srivastava, N. Chandra Sekhar Reddy, Dr. Anubha Shrivastava, "An Effective Code Metrics for Evaluation of Protected Parameters in Database Applications", International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.1.3, 2019.
5. Singh, Anshy, Shashi Shekhar, and Anand Singh Jalal. "Semantic based image retrieval using a multi-agent model by searching and filtering replicated web images." 2012 World Congress on Information and Communication Technologies. IEEE, 2012.
6. R. Nafeena Abdul Munaf, K. Karthikeyan, S. Joe Patrick Gnanaraj, T.A. Sivakumar, N. Muthukumaran, "An ML Approach for Household Power Consumption," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 2022, pp. 784-791.
7. N. Muthukumaran, N. R. G. Prasath and R. Kabilan, "Driver Sleepiness Detection Using Deep Learning Convolution Neural Network Classifier," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), Palladam, India, 2019, pp. 386-390.
8. Sturm, Roland, Ruopeng An, Josiase Maroba, and Deepak Patel. The S.P.Panimalar, R.T. Subhalakshmi (2021). “ Privacy and Security aspects of COVID 19 Image in Big Data Era”, in International Journal of Innovative Research in Science , Engineering and Technology, vol. 10, Issue 1pp 529-537.