**SMART BRIDGE INTERNSHIP**

**TEAM NAME :** Instinct Seekers

**Team Leader :**

Saikiran Seethamraju

**Team Members:**

Rubeena Fathima

Tahmeena Kousar

Mididoddi Vinisha

**HEALTH INSURANCE PREDICTION**

**INTRODUCTION:**

Health is one of the most important assets of our life which directly reflects in any form of progress or development.Here we allow users to get instant guidance on their Health Insurance through Health Insurance Prediction System which uses Machine Learning. Machine Learning is an application of Artificial Intelligence(AI) that provides system an ability to automatically learn and improve. Health Insurance is a service that takes away the financial burden that you might have to bear in case of medical emergency at any stage of life. When we have a huge data set on which we would like to perform predictive analysis or pattern recognition, machine learning is the way to go. Machine Learning (ML) is the fastest rising arena in computer science, and health informatics is of extreme challenge. The aim of Machine Learning is to develop algorithms which can learn and progress over time and can be used for predictions. Machine Learning practices are widely used in various fields and primarily health care industry has been benefitted a lot through machine learning prediction techniques. It offers a variety of alerting and risk management decision support tools, targeted at improving patients' safety and healthcare quality. With the need to reduce healthcare costs and the movement towards personalized healthcare, the healthcare industry faces challenges in the essential areas like, electronic record management, data integration, and computer aided diagnoses and disease predictions. Machine Learning offers a wide range of tools, techniques, and frameworks to address these challenges.

This system is fed to predict Health Insurance charge of each individual by considering various parameters.In this system we use regression algorithms.

**Objectives of Research:**

The Main objectives are:

* To find awareness levels of Health Insurance Systems,
* To estimate the charges of Health Insurance.
* To determine the satisfaction of customers regarding unpredictable life occurrences related to things like accidents, illness or even death.
* To ensure equitable distribution of health care costs among different income groups.

**Problem Statement:**

The Health Insurance Prediction System is designed by using supervised regression algorithms to predict the accurate charges based on unpredictable life occurrences related to things like accidents, illness or even death.

**Review of Literature:**

The health insurance sector is widely considered as one of the most important in making predictions using Machine Learning . More and more, Machine Learning has been considered as a practice that facilitates health insurance performance through using data and information efficiently within the healthcare sectors. Therefore, it is said that in order to understand the relation between Machine Learning and healthcare, we first need to understand what are the technologies used in healthcare.

Machine Learning algorithms have developed over the last few years not only as a automated learning system, but also as a strategic provider that develops and integrates health insurance sector’ infrastructures to facilitate and ensure quality of service.

**Data Collection:**

The implementation is done on the health insurance dataset which has 1340 records with 6 attributes.

**STEP 1:** Data Gathering:

Collecting data allows to capture a record of past events so that we can use data analysis to find recurring patterns. From those patterns, we build predictive [models](https://www.datarobot.com/wiki/model) using machine learning [algorithms](https://www.datarobot.com/wiki/algorithm) that look for trends and predict future changes .Predictive models are only as good as the data from which they are built, so good data collection is crucial to develop high-performing models.

**STEP 2:** Data Pre-processing:

Beforeimplementing any modeling techniques, the dataset needs to be pre-processed. The data need to be error-free and contain relevant information. For this we handle the missing data. Generally, there are many solutions such as:

null value replacement

mode/median/average value replacement

deleting the whole record

Model based imputation — regression, k-nearest neighbours etc

Handlingmissingdatain the wrong way can cause disasters.

**STEP3:**Feature Extraction:

In this step we extract required information from whole dataset.Including more relevant features helps to improve prediction power. Here we decide which key factors are important and effect the output and which doesn’t.The feature which is completely unrelated to the output can be removed.

**STEP 4:** Model Training:

A common strategy is to take all available label data, and split it into training and evaluation subsets.

The rule of splitting the data is 80–20 percent training & testing sets respectively.

The ML system uses the training data to train models to see patterns, and uses the evaluation data to evaluate the predictive quality of the trained model. The ML system evaluates predictive performance by comparing predictions on the evaluation data set with true value using a variety of metrics.

Here the metrics used is r2\_score.

STEP 5: Prediction:

**If prediction is <75% compared to true value then it turns back to model training and if >75% then the system is taken for deployment.**

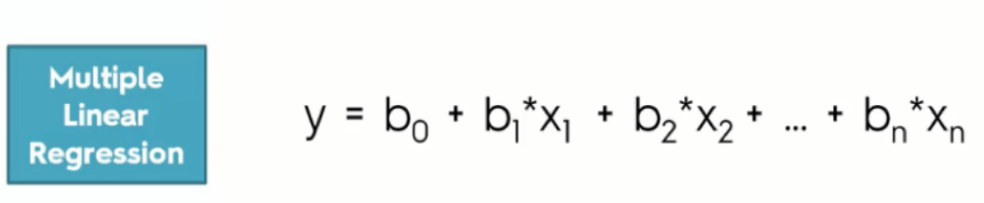
Machine Learning is using data to answer questions.So prediction is the step where we get to answer some questions.This is the point of all this work,where the value of machine learning is realized.We can finally use our model to predict.

**Methodology**

Regression is an approach to model the relationship between a dependent response variable and one or more explanatory variables. It considers data as a series, rather than considering each time point separately. A regression model is therefore useful in predicting what is likely to happen in the next time period, or even in the far future. An additional advantage of the regression method is that it can account for multiple factors, which may affect the trend rate.

**1.Multi Linear Regression**:

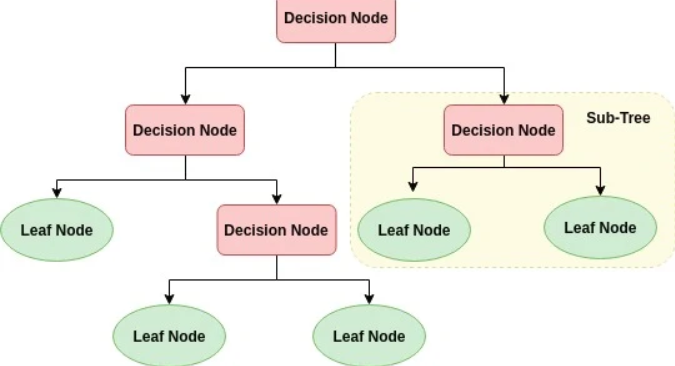
Multi Linear Regression is the most common form of linear regression analysis.  As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. Multilinear regression is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable). Multiple linear regression depicts the relationship between one dependent variable and two or more independent variables. Each value of the dependent variable x is associated with the value of independent variable y. The regression line p for independent variables x1, x2,…,xp is deﬁned to be:



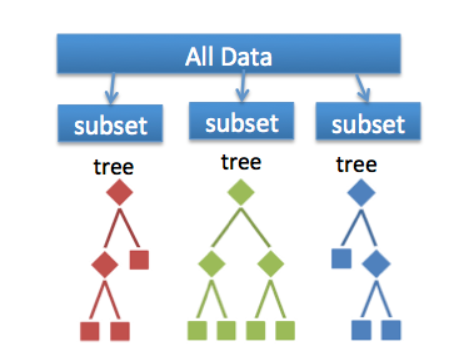
**2.Decision Tree** :Basically, a decision tree is a graphical representation. That makes use of branching method to exemplify all possible outcomes of a decision. Basically, in a decision tree, the internal node represents a test on the attribute. As each branch of the tree represents the outcome of the test. And also the leaf node represents a particular class label. i.e. the decision made after computing all the attributes. Further, we have to represent classification through the path from a root to the leaf node. Decision Trees are supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Decision tree is simple to understand and to interpret and it requires little data preparation.

Classification of Decision Tree:



**3.Random Forest** : In machine learning, the random forest algorithm is also known as the random forest classifier. It is a very popular classification algorithm. One of the most interesting thing about this algorithm is that it can be used as both [classification and regression algorithm.](https://dataaspirant.com/2014/09/27/classification-and-prediction) The random forest algorithm is an algorithm for machine learning, which is a forest. We know the forest consists of trees. The trees being mentioned here are [decision trees](https://en.wikipedia.org/wiki/Decision_tree).  Therefore, the random forest algorithm comprises a random collection of a forest tree. It is an addition to the decision tree algorithm. So basically, what a random forest algorithm does is that it creates multiple decision trees and merges them together to obtain a more stable and accurate prediction. In general, the more trees in the forest, the more robust would be the prediction and thus higher accuracy.



**Exploratory Data Analysis:**

**Exploratory data analysis** (**EDA**) is an approach to [analyzing](https://en.wikipedia.org/wiki/Data_analysis) [data sets](https://en.wikipedia.org/wiki/Data_set) to summarize their main characteristics.

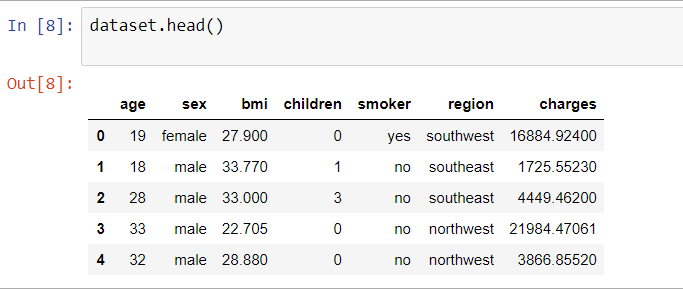
We are performing data analysis on Health Insurance dataset.

**STEP1**:Importing libraries

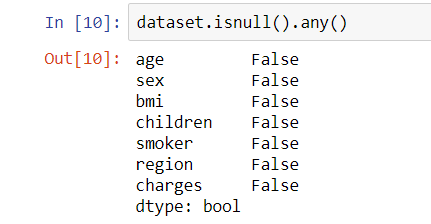
**STEP2**: Importing dataset

**STEP3**: Using head() we can view first 5 records of our dataset i.e shown below.

****



**STEP 4**: Handling missing data.



Since we don’t have any null values in dataset.Hence there is no need to remove any values.

**STEP 5**: Separating independent and dependent variables in dataset.

Independent variables in our dataset are:

Age,Sex,BMI,Children,Smoker,Region.

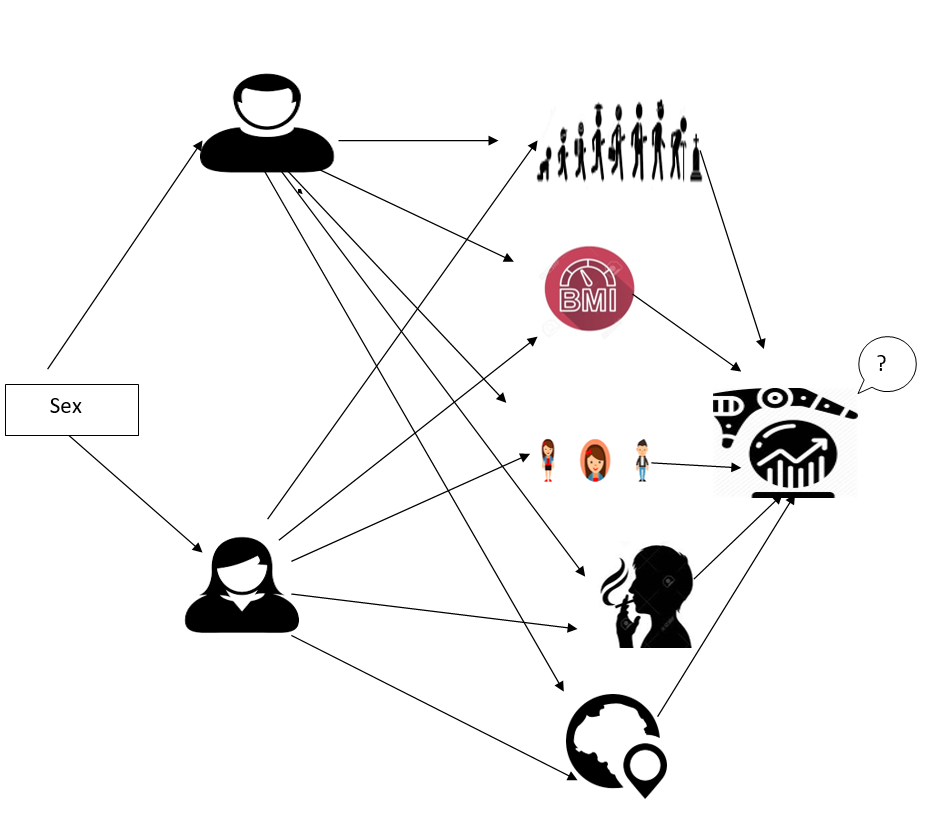
Dependent variable in our dataset is:

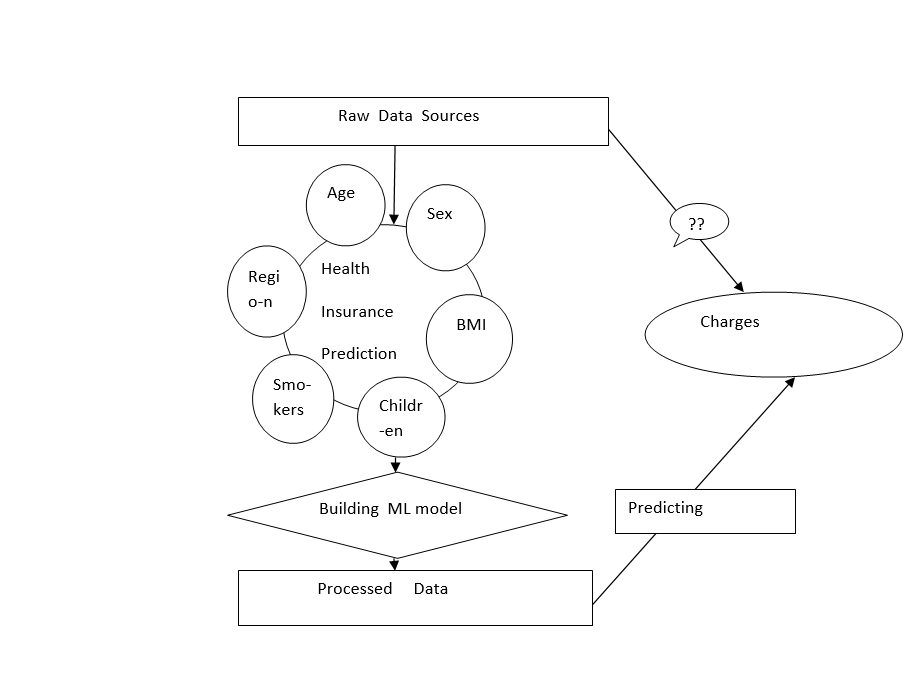
Charges.

**STEP 6** :Performing Label Encoding to columns(containing strings) Sex,Smoker,Region.

**STEP 7** : Performing Training and testing on data.

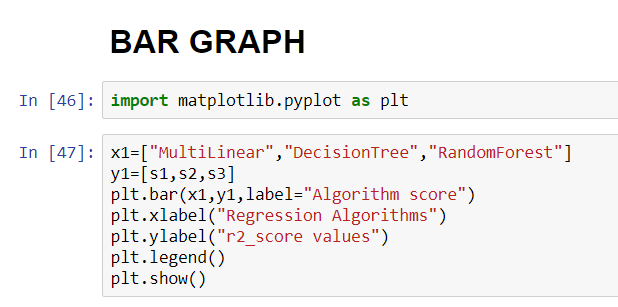
**Figures and Tables:**



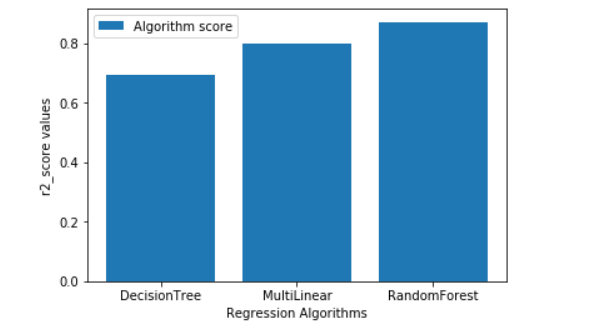


**Statistical techniques and**

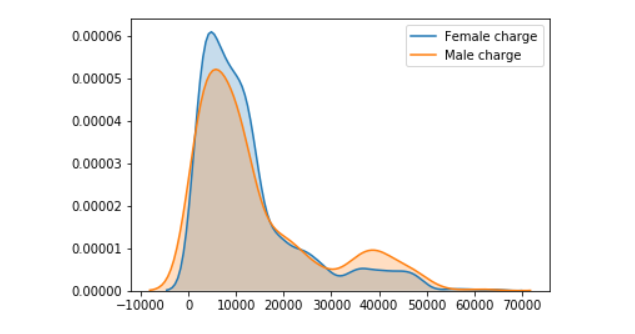
**data visualization:**



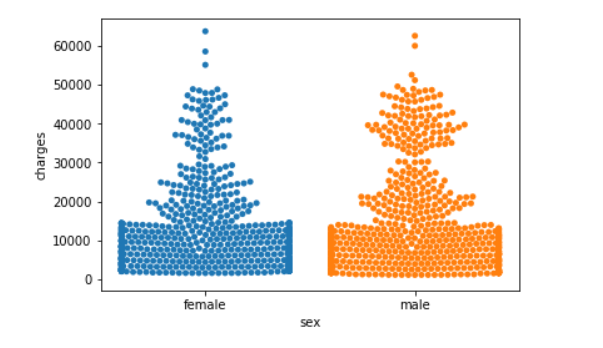
The output of the bargraph is:



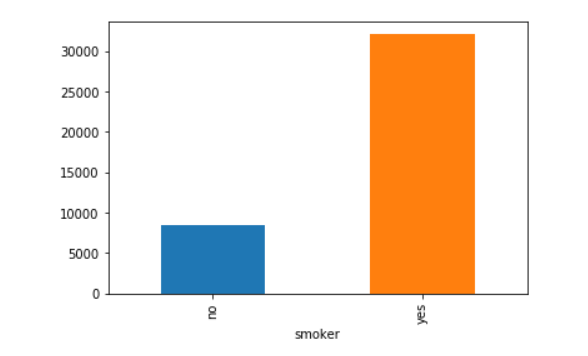
Comparing the charges of male and female by using kdeplot :



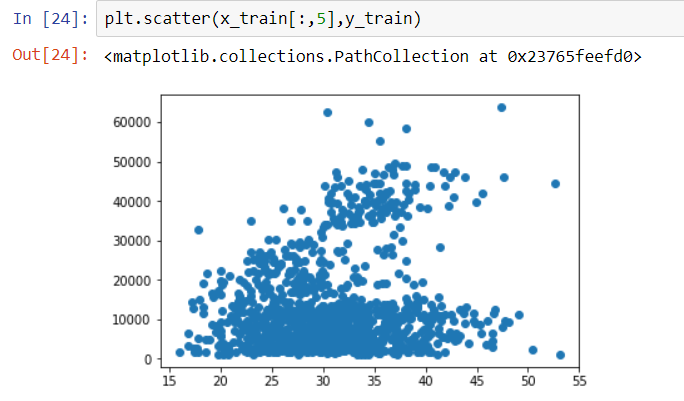
Comparing the charges of male and female by using swarmplot :



The impact of smoke on charges :



Scattering the points by using scatter():



**Data modeling using Supervised**

**ML techniques:**

Supervised ML techniques used in our model are:

Model built using Multilinear Regression:

**STEP 1**: Importing Linear Regression package.

**STEP 2**: Predicting values and fitting model by using x\_train and y\_train parameters.

**STEP 3** : Finding accuracy by using r2\_score.

Model built using Decision Tree:

**STEP 1**: Importing DecisinTreeRegressor package.

**STEP 2**: Predicting values and fitting model by using x\_train and y\_train parameters.

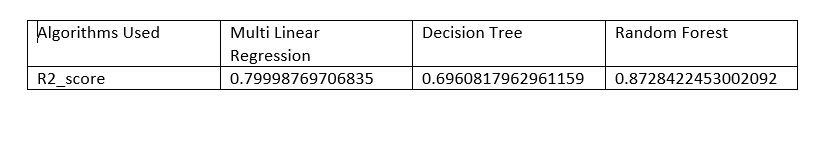
**STEP 3** : Finding accuracy by using r2\_score.

Model built using Random Forest:

**STEP 1**: Importing RandomForestRegressor package.

**STEP 2**: Predicting values and fitting model by using x\_train and y\_train parameters.

**STEP 3** : Finding accuracy by using r2\_score.

Comparative Analysis of Regression Algorithms: 

Thus, by comparing the r2\_score of Regression algorithms,Random Forest Algorithm is slightly better than the multilinear regression and decision tree. Hence, Random Forest can be used in scoring any new claim data.

**Findings and Suggestions:**

**Findings:**

From the above Regression Algorithms, we have calculated the r2\_score by using Multi linear Regression, Decision Tree, Random Forest.Random Forest Algorithm is slightly better than the multilinear regression and decision tree.

**Suggestions:**

Health Insurance System should suggest the policy according to the individuals requirements.

**Conclusion:**

In this project , the concepts of machine learning have been implemented for building a prediction model which can help in detecting health insurance charges claims made by thecustomers by training the model using supervised learning technique.Thismodel is validated using Multilinear regression, Decision tree, Random forest.

By comparing different statistics derived from these three models, Random forest was found to be slightly better than multilinear regression and decision treefor the purpose of making predictions and for scoring any new claim data.

**Result:**

