**Report**

**HEALTH INSURANCE COST PREDICTION**

Mentor: Akash Maurya Sir

Group Members:

1. Piyush More

2. Aditya Patil

3. Fensta Praiselin

4. Neelima

**Introduction**

Why do we need a health insurance you might ask, especially if you're generally healthy. But there are many reasons for having one.

First, your health matters the most, and health insurance is a way of helping protect your health.

Here are few ways health insurances can protect you: -

* Health insurances are designed to help prepare for unexpected moments in your life, like accidents or illness.
* Without health insurance, you could be faced with paying medical bills yourself, and health care can be expensive.
* Health insurance can keep you on track with preventive care — which may be covered by insurance at 100%, in many cases. Regular doctor visits may help you stay healthier over time.

Anyone can get a health insurance irrespective of their age, sex, or their place of living. And this project helps them to get an estimate of the price of their health insurance based on the factors listed below

**Data Description**

The goal of this project is to allow a person to get an idea about the necessary amount required according to their own health status. Later they can comply with any health insurance company and their schemes & benefits keeping in mind the predicted amount from our project. This can help a person in focusing more on the health aspect of an insurance rather than the futile part.

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.

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 predicted, explanatory or regressor variables).

**2. 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 node. 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 tress.

# 3. Random Forest Regressor

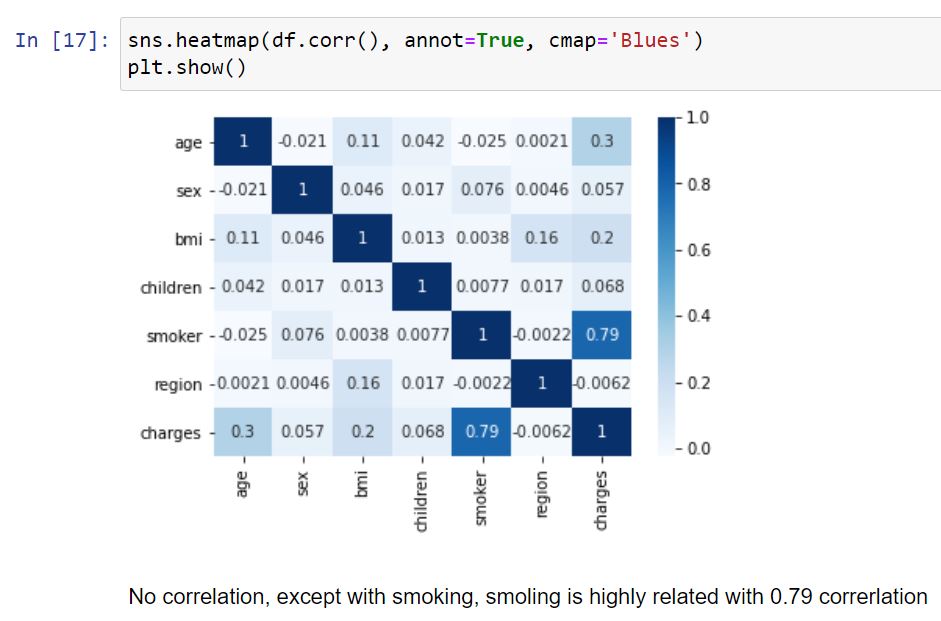
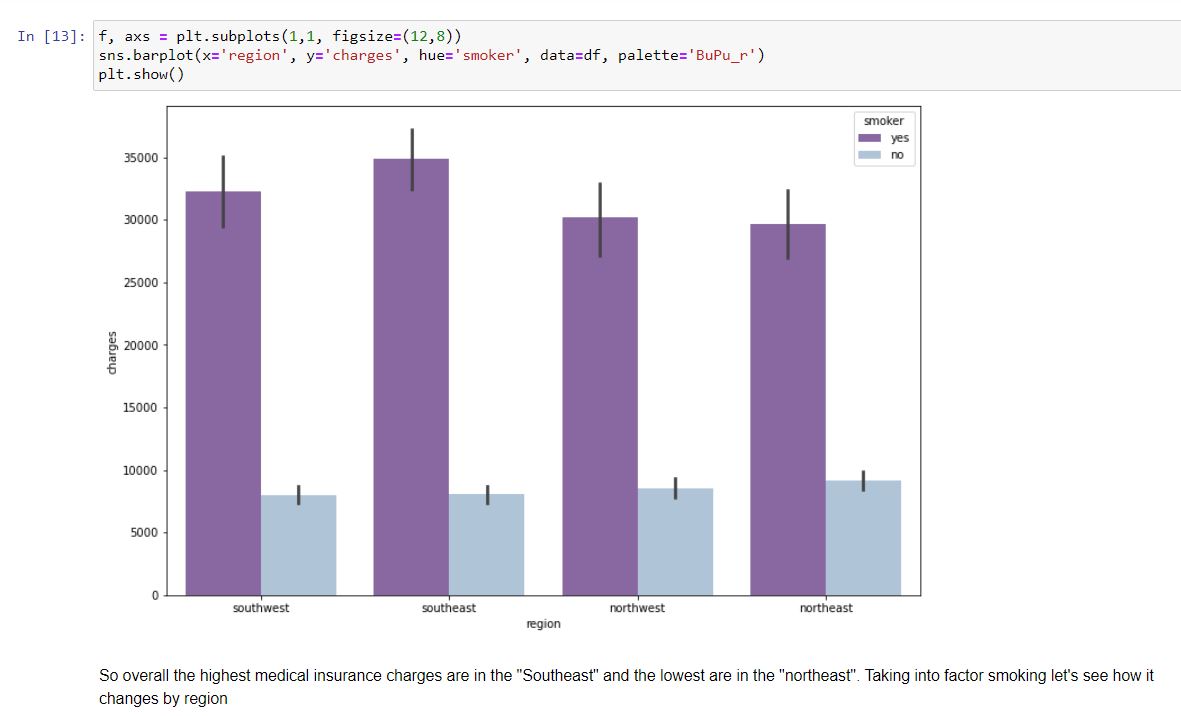
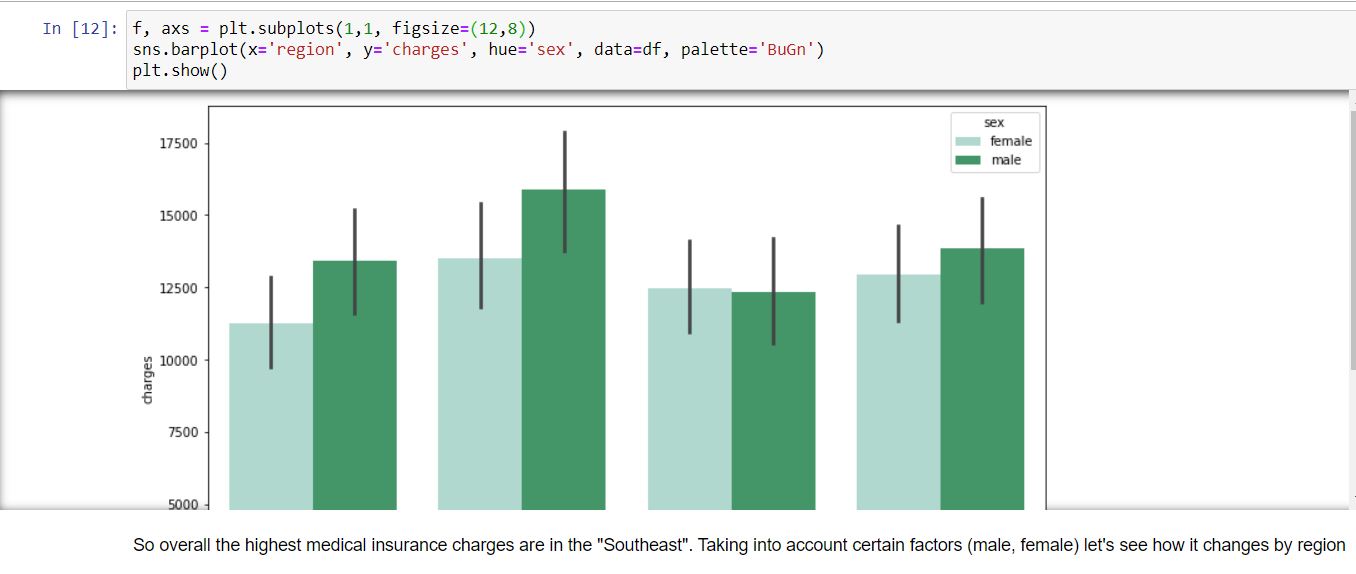
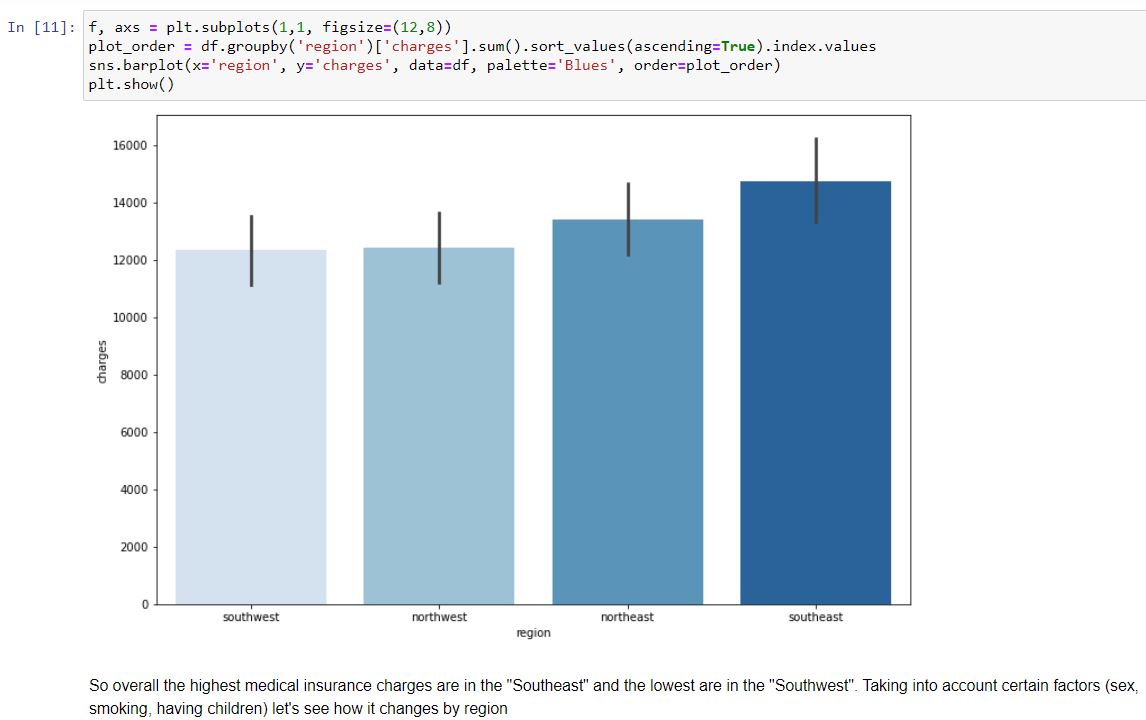
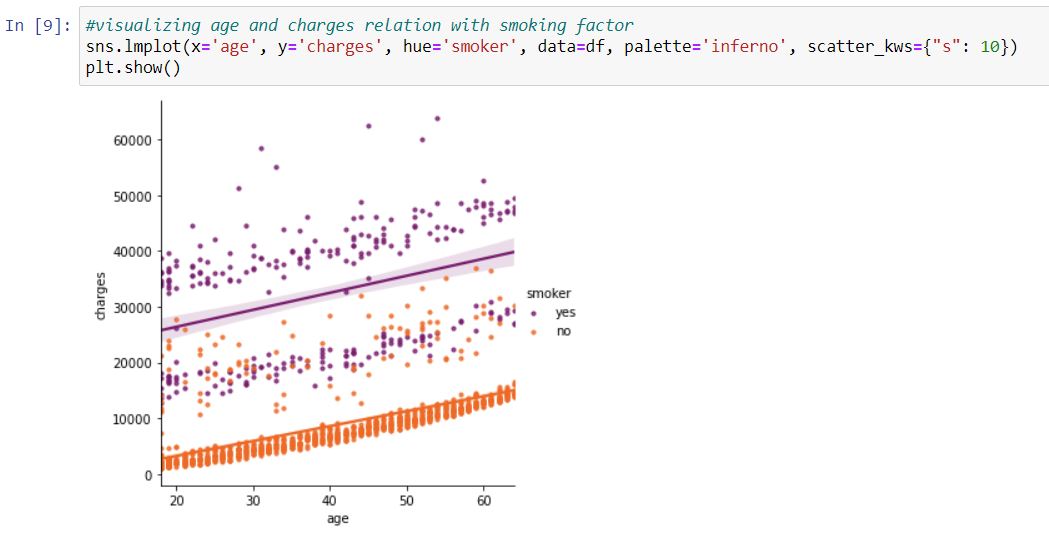
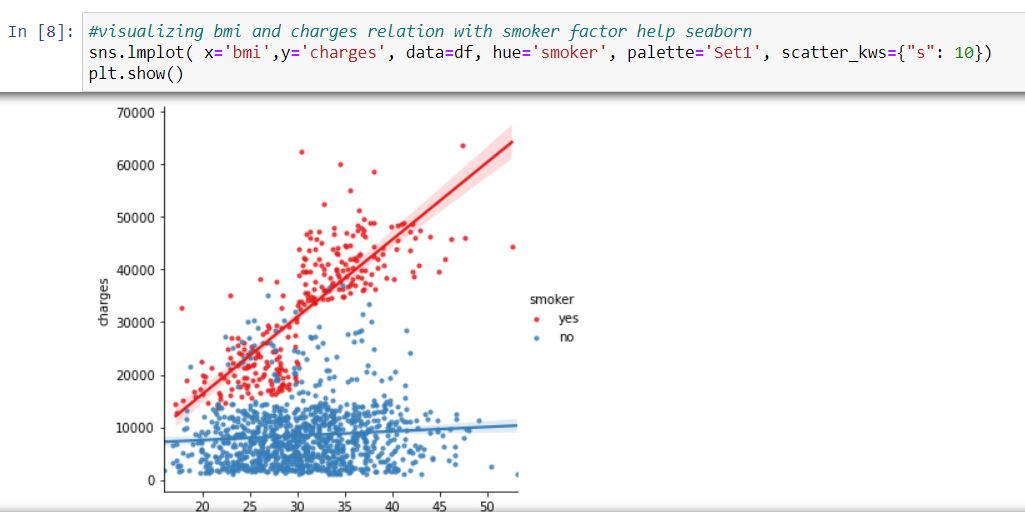
A random forest is a Meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

**Approach**

We started proceeding with reading with data and we found that various factors that affected the health insurance cost were age, sex, bmi, children, smoker and region .So our target variable was charges .Then we started with processing the data in which we preformed data cleaning, data description and found that the data was having no null values and the minimum cost of insurance was 1122 and around 75% of charges were 17000 The maximum cost was 63770 .We started visualizing 3 factors .That is smoker ,region charges and found that highest medical charges are in south east region . We converted categories into numerical form using label encoder.

We plotted heat map using seaborn for the data core relation and found that smoker and charges were highly co-related with 0.79 co-relation. Then we splitted the data into training and testing set and stated training models Multiple Linear regression , Random Forest regression, Decision Tree.

**Visualization**



**Algorithms**

As we mentioned above we have used various regression model on these data set and they resulted in different accuracy we have used Multiple Linear regression model which as accuracy of 0.79. Next mode was Random Forest regression with accuracy of 0.86 and the last model was Decision Tree with accuracy of 0.72.

**Comparison**

From the above three model Random Forest regression is most accurate and best suited for data set.

**Result and Discussion**

From the above Models we found Random Forest Model most applicable to our problem of health insurance cost prediction with accuracy of 0.79

**Conclusion**

* Prediction Health insurance cost depends on person to person with various factors.
* From given data set we found that smoker from southeast region were factors that are mainly responsible for affecting cost of insurance.
* We analyzed the data through visualization, found that smoker and charges was highly correlated.
* So we train various models and found that most accurate model is Random forest model with accuracy of 86%.

**Future work**

Using this model we can develop an application for people to predict the cost .

**Difficulties faced**

Difficulty faced during choosing a model, analyzing various factors.

**References**

Websites:

* <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html>
* <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>
* <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>
* <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>