HEALTH INSURANCE COST PREDICTION:

1. INTRODUCTION:

Problem Statement:

Building a machine learning model which helps the health
insurance companies in predicting premiums for their customers based on certain important factors.

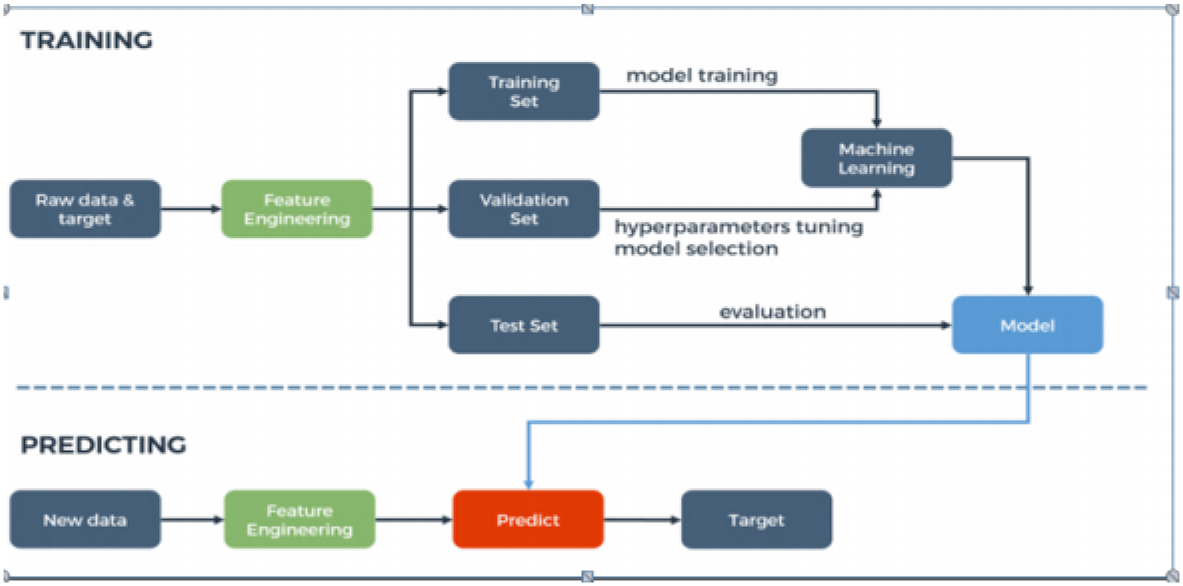
- **Overview**:
  The aim of the project is to build a machine learning model which helps the health insurance companies in predicting the premium health insurance offers for their customers by following the health laws in the country.
- **Purpose:**
  To build a machine learning model that helps the health insurance companies to provide premium offers to the customers based on certain factors and statistics.

2. LITERATURE SURVEY:

- **Existing Problem:** Health insurance companies face a problem in determining the premium insurances for their customers. They need a trained model which predicts what kind of premiums are suitable to a customer. The companies should also follow certain rules set by the health care law in that country. In order to determine premium offers the companies have to consider certain factors and also go statistically to give importance to a customer.
- **Proposed Solution:**
  The main motto of the project is to provide the health insurance companies with a machine learning model using the machine learning algorithms and python programming that can predict the customer's eligibility for different premium insurance services based on certain factors like age, gender, BMI, weight, previous health issues, the region he/she belongs to, if he is a smoker or not etc. The model is tested based on the accuracy and performance of the model.

3. THEORETICAL ANALYSIS:

- **Block Diagram:**

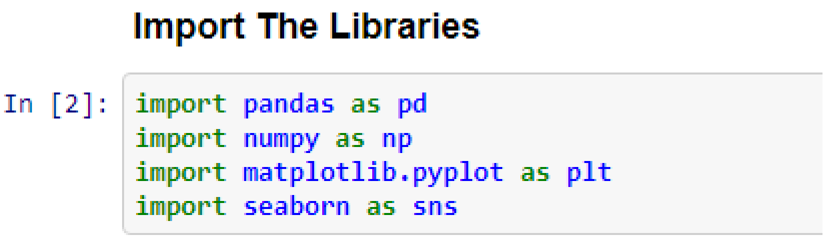


- **Hardware Software Designing:**
  Python based computer vision and knowledge about various machine learning algorithms and how to implement them. Knowledge of how to use Jupiter notebook for the deployment of the project.

4. EXPERIMENTAL INVESTIGATIONS:

EXPERIMENTAL INVESTIGATIONS:

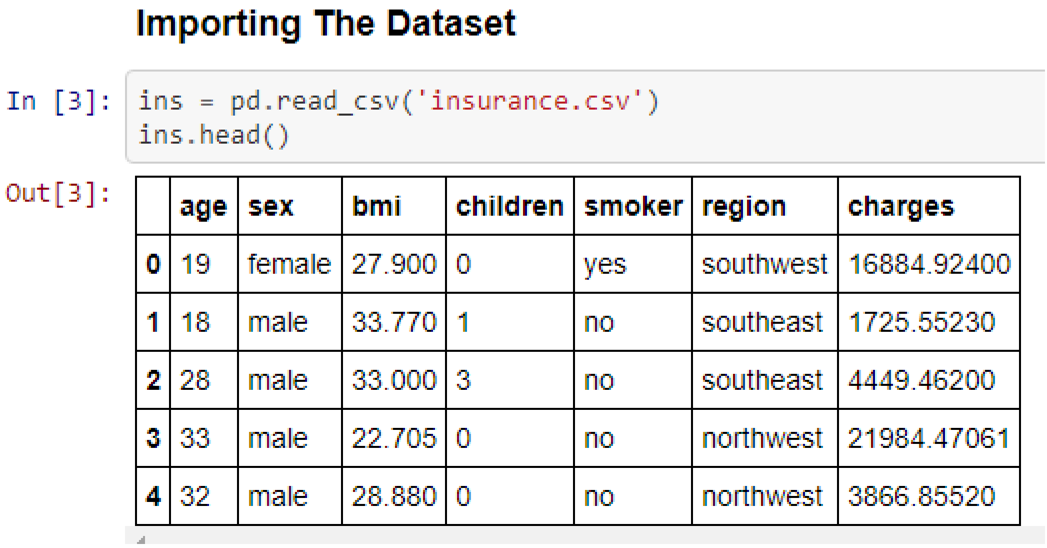
1. Collect the data set. For this I downloaded the dataset from Kaggle.
2. Before building the machine learning model the data collected should be preprocessed. In the data preprocessing the following steps are to be followed:
   - Importing the required libraries.



Import The Libraries

```
In [2]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
```

   - Importing the dataset.



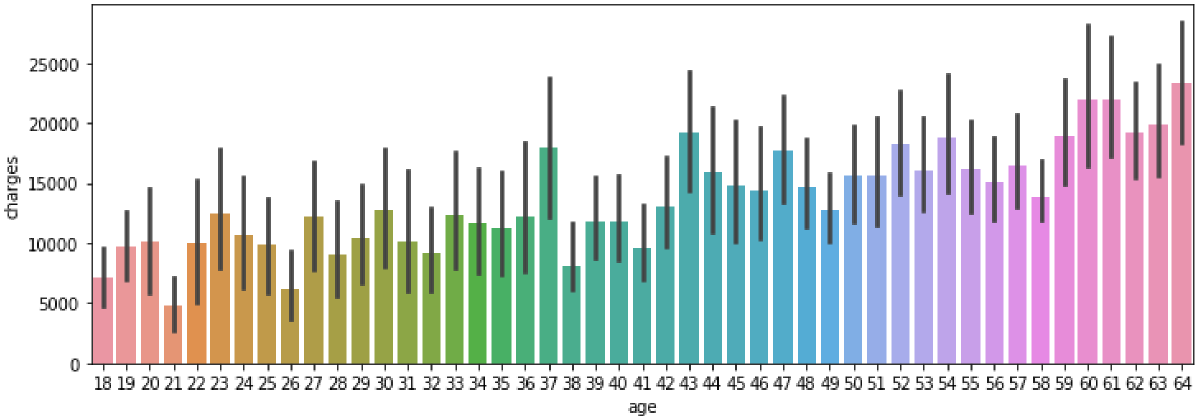
Importing The Dataset

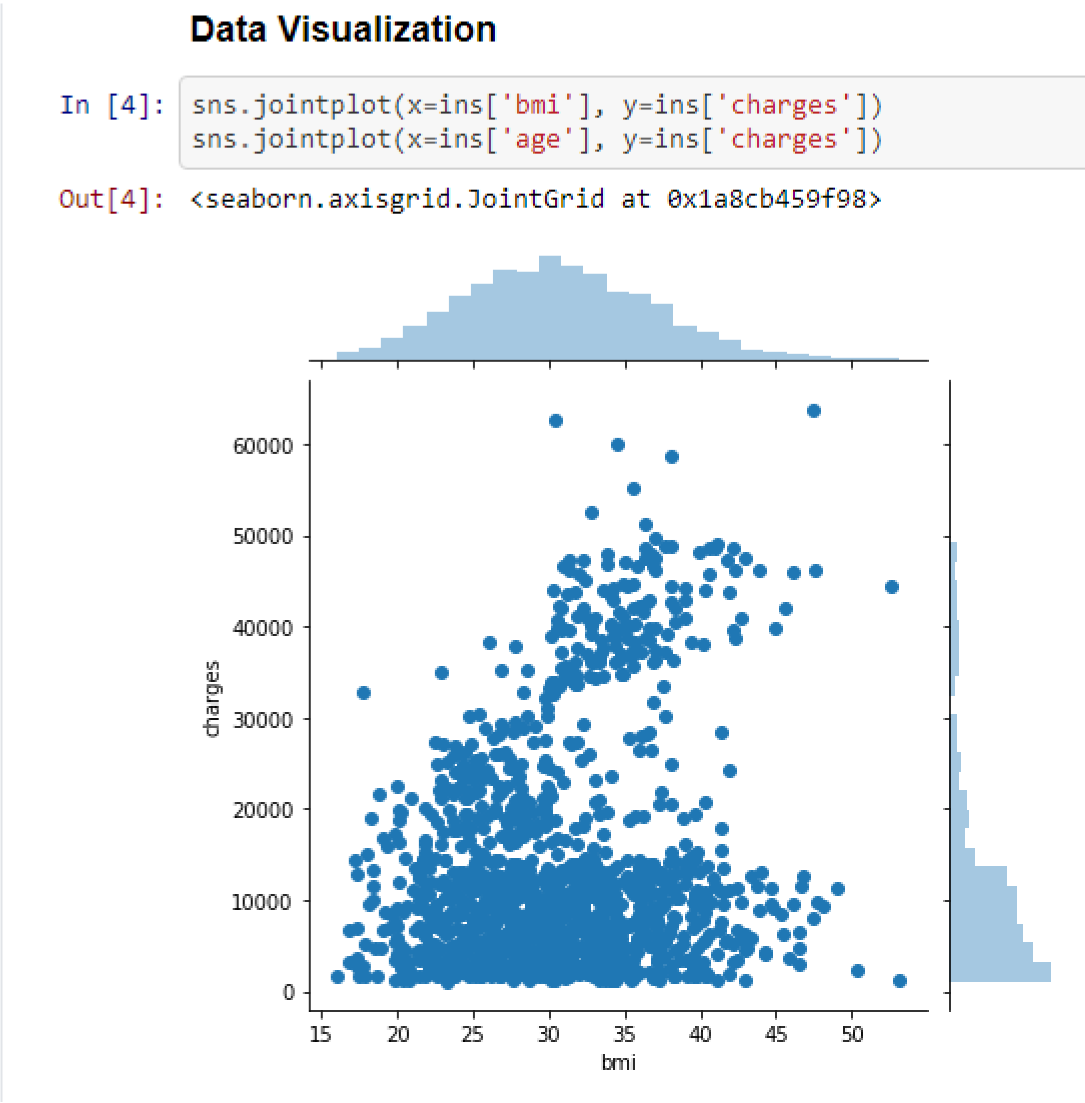
```
In [3]: ins = pd.read_csv('insurance.csv')
        ins.head()
```

Out[3]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

- Visualizing the data. By this we can understand the columns in the dataset and also know about the missing data and null values.



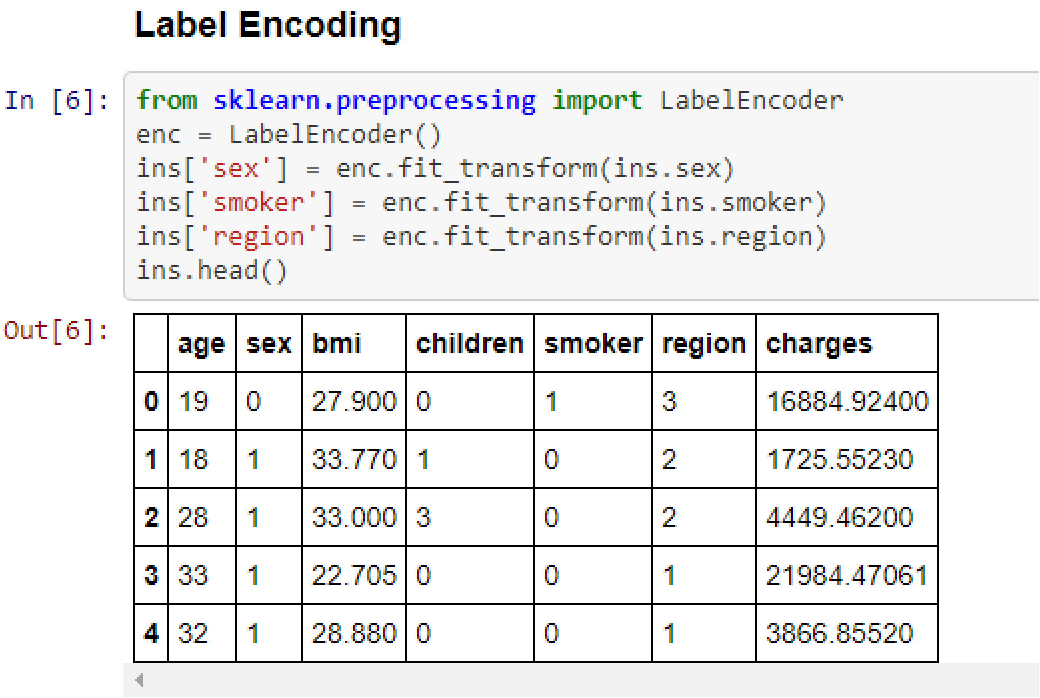


Data Visualization

```
In [4]: sns.jointplot(x=ins['bmi'], y=ins['charges'])
        sns.jointplot(x=ins['age'], y=ins['charges'])
```

```
Out[4]: <seaborn.axisgrid.JointGrid at 0x1a8cb459f98>
```

- After this we should take care about the missing data.
- Then we should do label encoding followed by one hot encoding.

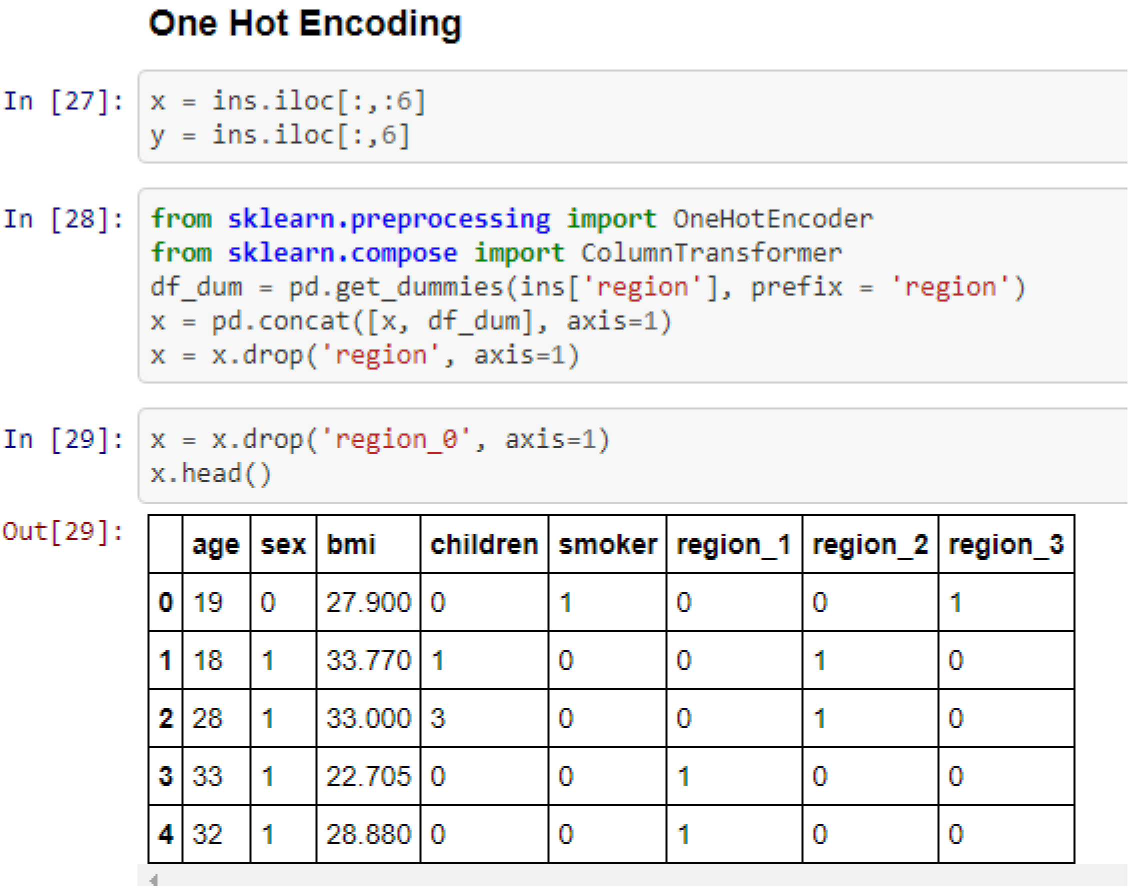


Label Encoding

In [6]:
```
from sklearn.preprocessing import LabelEncoder
enc = LabelEncoder()
ins['sex'] = enc.fit_transform(ins.sex)
ins['smoker'] = enc.fit_transform(ins.smoker)
ins['region'] = enc.fit_transform(ins.region)
ins.head()
```

Out[6]:

	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	1	3	16884.92400
1	18	1	33.770	1	0	2	1725.55230
2	28	1	33.000	3	0	2	4449.46200
3	33	1	22.705	0	0	1	21984.47061
4	32	1	28.880	0	0	1	3866.85520



One Hot Encoding

In [27]:
```
x = ins.iloc[:,:6]
y = ins.iloc[:,6]
```

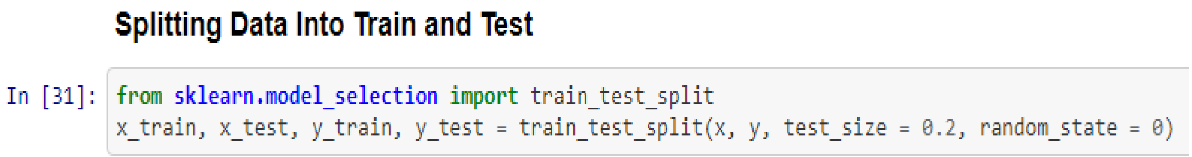
In [28]:
```
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
df_dum = pd.get_dummies(ins['region'], prefix = 'region')
x = pd.concat([x, df_dum], axis=1)
x = x.drop('region', axis=1)
```

In [29]:
```
x = x.drop('region_0', axis=1)
x.head()
```

Out[29]:

	age	sex	bmi	children	smoker	region_1	region_2	region_3
0	19	0	27.900	0	1	0	0	1
1	18	1	33.770	1	0	0	1	0
2	28	1	33.000	3	0	0	1	0
3	33	1	22.705	0	0	1	0	0
4	32	1	28.880	0	0	1	0	0

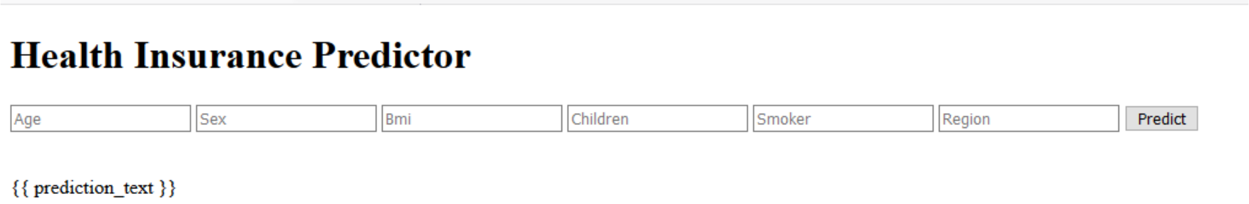
- Then feature scaling of the data should be done.
- Then we split the data into training data and test data



Splitting Data Into Train and Test

```
In [31]: from sklearn.model_selection import train_test_split
         x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 0)
```

3. After data preprocessing, we need to choose the best machine learning algorithm that can be used for building the model.
4. After building the prediction model we need to build the application. This involves the following steps:
    - Creating an HTML file.



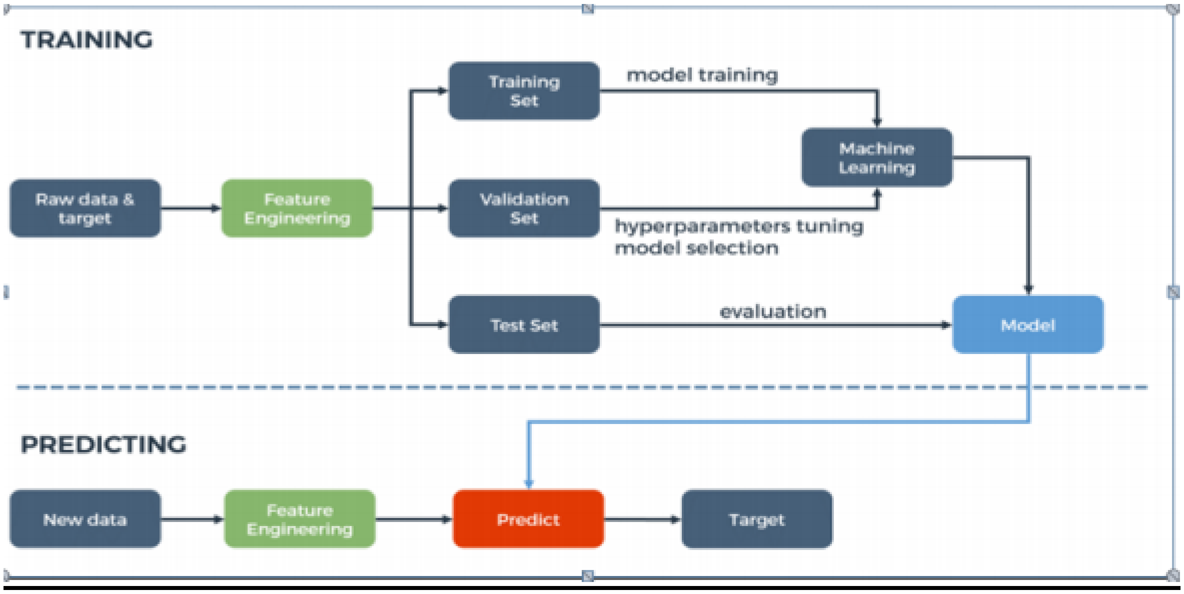
Health Insurance Predictor

Age | Sex | Bmi | Children | Smoker | Region | Predict

{{ prediction_text }}

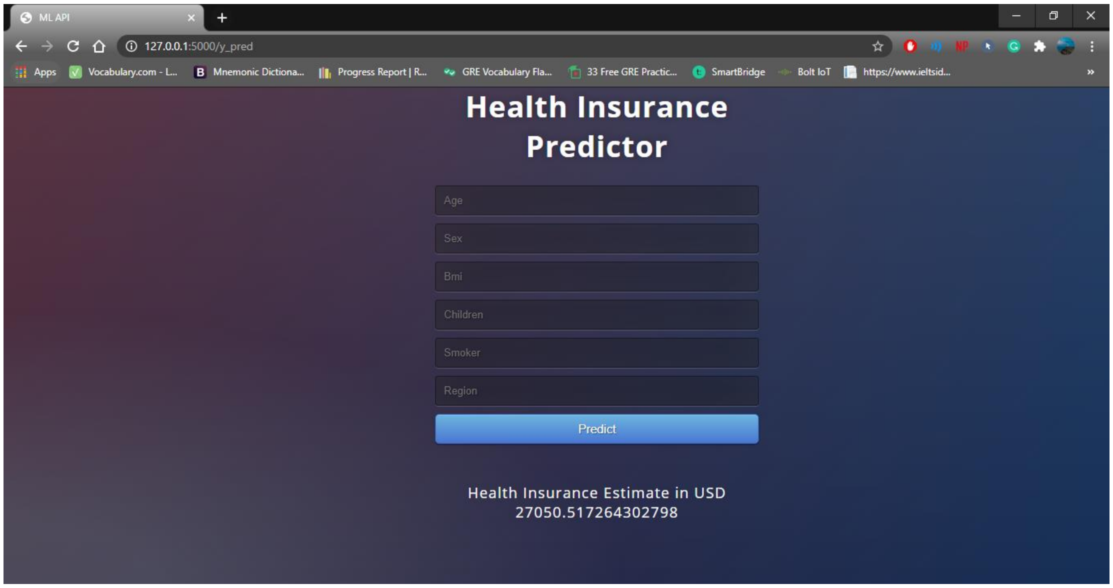
    - Building the required python code.
5. After this the model is to be deployed.
6. Then we test the model against the test dataset and check the accuracy of the prediction model.
7. Next, we need to make the node-RED flow.
8. Once the model is deployed it can be seen on the user interface.

5. FLOWCHART:



6. RESULT:

The machine learning model that works on the regression algorithm predicts the cost of the health insurance package that can be given to a customer based on the details of the customers depending upon various factors.



Health Insurance Estimate in USD 27050.517264302798

7. ADVANTAGES AND DISADVANTAGES:

- **Advantages:**
  This model helps the health insurance companies to predict the premium offers in insurance that can be provided to a particular customer based on factors like age, gender, weight, previous health issues etc.
- **Disadvantages:**
  Sometimes the cost requirement may not be dependent on gender, region they belong to.

8. APPLICATIONS:

This model predicts the various health insurance premiums applicable to a given customer and can be used by the health insurance companies in order to improvise their services and can also bring new changes to their policies based on the statistics.

9. CONCLUSION:

This was a great experience with Smart bridge learning new and interesting things and also applying them in real time. Coming to my project I can say that the machine learning model that is created to predict the cost of health insurance has a wide range of applications and makes the work of health insurance companies simpler. This gives all the predictions just by giving basic details of the customer. I learnt a lot from this project and also thank all the mentors and the bootcamp that was very supportive and helpful at every point of work.

10. FUTURE SCOPE:

This can be implemented more efficiently by adding some more factors depending on which the model predicts the health insurance cost that can be provided.

11. BIBLIOGRAPHY:

- Kaggle for downloading the dataset.
- Reference links given for the project individual tasks.