

## ● Project on Medical Cost Prediction using Machine Learning Mode ●

- **Aim** - Everyone's life revolves around their health. Good health is essential to all aspects of our lives. Health refers to a person's ability to cope up with the environment on a physical, emotional, mental, and social level. Because of the quick speed of our lives, we are adopting many habits that are harming our health. One spends a lot of money to be healthy by participating in physical activities or having frequent health check-ups to avoid being unfit and get rid of health disorders. When we become ill we tend to spend a lot of money, resulting in a lot of medical expenses. So, an application can be made which can make people understand the factors which are making them unfit, and creating a lot of medical expenses, and it could identify and estimate medical expense if someone has such factors.
- **Steps to be taken in the Project is sub-divided into the following sections. These are:**
  - Loading necessary libraries such as numpy , pandas , sklearn etc.
  - Loading the dataset as CSV file and showing first five rows.
  - Drop the unnecessary columns from dataset.
  - Data preprocessing or (Data cleaning) performed by the one hot encoding in this process we change categorical data into numerical data and the technique is called feature Engineering.
  - Visualization of Medical expenses with features using Tableau.
  - Splitting the cleaned data into dependent and independent variables.
  - Splitting the data into train and test sets with train\_test\_split using sklearn library.
  - Import different kind of Regression Models and Train that model with the help of .fit().
  - Predicting the trained models and then checking their accuracy of the model using accuracy score.
  - Then recall the train\_test\_split and split the data into training and testing set with different models.
  - Then predicting the trained models and checking the accuracy of model and check the accuracy difference.

Shahjan  
Khan

- And finally predict whether the medical expense value generated or not by checking the medical expenses for the new customer.

## **Step-1** – Loading Necessary Libraries used in machine learning.

Loading Necessary Libraries

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

## **Step-2** – Loading and Showing the first five rows of the dataset with .head().

Load and Show Dataset

```
data=pd.read_csv("/content/medical_cost.csv")
data.head()
```

	Id	age	sex	bmi	children	smoker	region	charges
0	1	19	female	27.900	0	yes	southwest	16884.92400
1	2	18	male	33.770	1	no	southeast	1725.55230
2	3	28	male	33.000	3	no	southeast	4449.46200
3	4	33	male	22.705	0	no	northwest	21984.47061
4	5	32	male	28.880	0	no	northwest	3866.85520

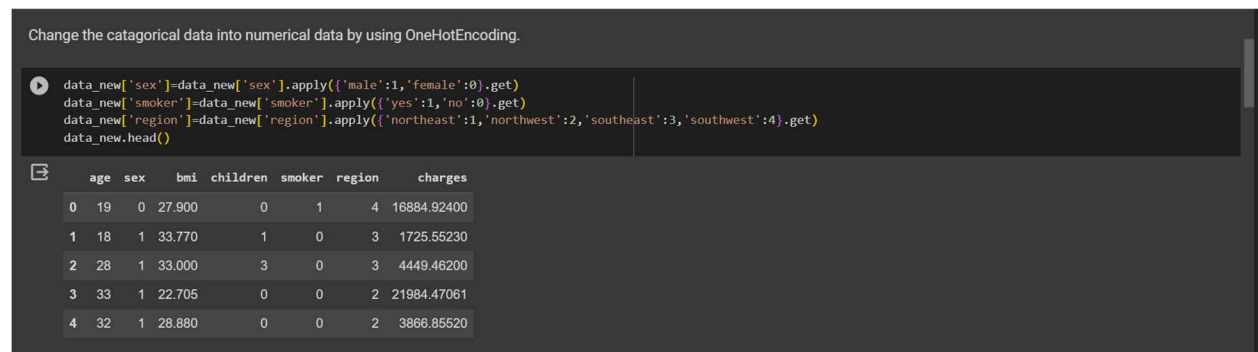
## **Step-3** – Drop the unnecessary columns from dataset.

Dropping Unnecessary Column

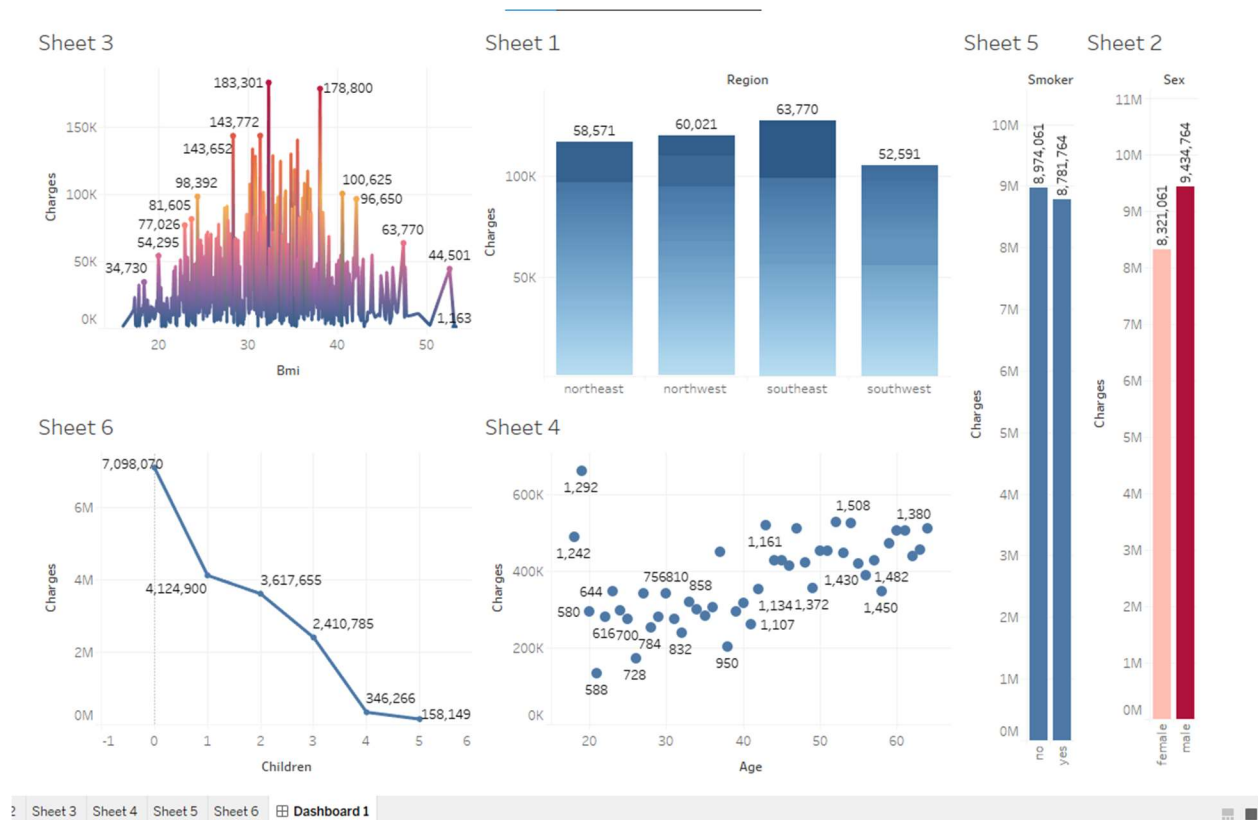
```
[ ] data_new=data.drop(['Id'],axis=1)
data_new.head()
```

	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

## **Step-4** – Data preprocessing or (data cleaning) by one hot encoding with feature engineering technique.



## **Step-5** – Visualization of Medical Expenses with Features using Tableau.



## **Step-6** – Splitting Cleaned Data into Dependent and Independent Variables.

Divide the Cleaned data into dependent(y) and Independent(x) variables.

```
[ ] x=data_new.drop(['charges'],axis=1)
    y=data_new['charges']
```

## **Step-7** – Splitting the Cleaned data into Training and Testing set.

Splitting the Cleaned data into Training and Testing set

```
[ ] from sklearn.model_selection import train_test_split
    x_train,x_test,y_train,y_test=train_test_split(x,y,train_size=0.8)
```

## **Step-8** – Import first ML Model Linear\_Regression().

Creating Machine Learning Model Using Linear Regression Algorithm.

```
[ ] from sklearn.linear_model import LinearRegression
    regression=LinearRegression()
```

## **Step-9** – Train the model using .fit() function.

Train the Model with .fit() function

```
regression.fit(x_train,y_train)
```

```
LinearRegression
LinearRegression()
```

## **Step-10** – Predict the trained model using .predict() function.

Predict the trained model using .predict() function

```
predictions=regression.predict(x_test)
```

Shahjan  
Khan

**Step-11** – Check the accuracy with the help of accuracy score.

```
Check the Accuracy with accuracy score.

regression.score(x,y)

0.7506873604195824
```

**Step-12** – Repeat the process using other machine learning model which is Decision tree regressor.

```
Creating a Machine Learning Model Using Decision Tree Regression Model

[ ] from sklearn.tree import DecisionTreeRegressor
    tree=DecisionTreeRegressor()

Train the Model

[ ] tree.fit(x_train,y_train)

DecisionTreeRegressor
DecisionTreeRegressor()

Make predictions on model

[ ] predictions=tree.predict(x_test)
```

**Step-13** – Checking the accuracy of Decision tree regressor model.

```
Check the accuracy score of model

[ ] tree.score(x,y)

0.9413618580172949
```

**Step-14** – Repeat the process using Random Forest Regressor.

Shahjan  
Khan

```
Creating a Machine Learning Model Using Random Forest Model

[ ] from sklearn.ensemble import RandomForestRegressor
    forest=RandomForestRegressor()

[ ] forest.fit(x_train,y_train)

+ RandomForestRegressor
RandomForestRegressor()

[ ] predictions=forest.predict(x_test)
```

**Step-15** – Check the accuracy of Random forest regressor model.

```
forest.score(x,y)

0.9495154058417709

+ Code + Text
```

**Step-16** – Check the accuracies.

1. Using Linear Regression – 75% accuracy.
2. Using Decision Tree Regressor – 94% accuracy.
3. Using Random Forest Regressor – 95% accuracy.

**Step-17** – Checking the medical expenses for the new customer.

1. First Create the information data for the new customer and convert it in a data-frame.

```
Creating the Medical Cost for a new Patient

[ ] new_data={'age':34,'sex':1,'bmi':29.800,'children':2,'smoker':0,'region':2}

[ ] my_data=pd.DataFrame(new_data,index=[1])

my_data

age sex bmi children smoker region
1 34 1 29.8 2 0 2
```

Shahjan

Khan

2. Predict the Medical cost for the new customer.

```
Predicting the Medical Cost for New Customer

[ ] new_predictions=forest.predict(my_data)

[ ] print("The Medical Cost Price for the New Customer is",new_predictions)

The Medical Cost Price for the New Customer is [8124.5667859]
```

**The Medical Cost for the new customer is 8124.57 Rs.**

**Conclusion** – This Project can easily demonstrate the reasons for producing a certain forecast regarding potential healthcare expenses, which is a useful capacity in the healthcare area. This Project is used to estimate the healthcare costs of the patients such as obesity (BMI) using certain devices such as smartphones and smart devices. For estimation, by the use of linear regression, Decision tree regressor and Random forest regressor (supervised learning) performs more accurately. By providing comprehensive evidence, regression methodology can be effectively used for prognosis in conjunction with the dataset. The domain and time accuracy will determine the prediction model and the estimation of healthcare expenses. The proposed method reduces the risk of overfitting, and also, training time is less. This method is effective in estimating the healthcare costs of patients with an accuracy rate of 95%. The extensive tests on a real-time world database have confirmed the efficiency of our method.

