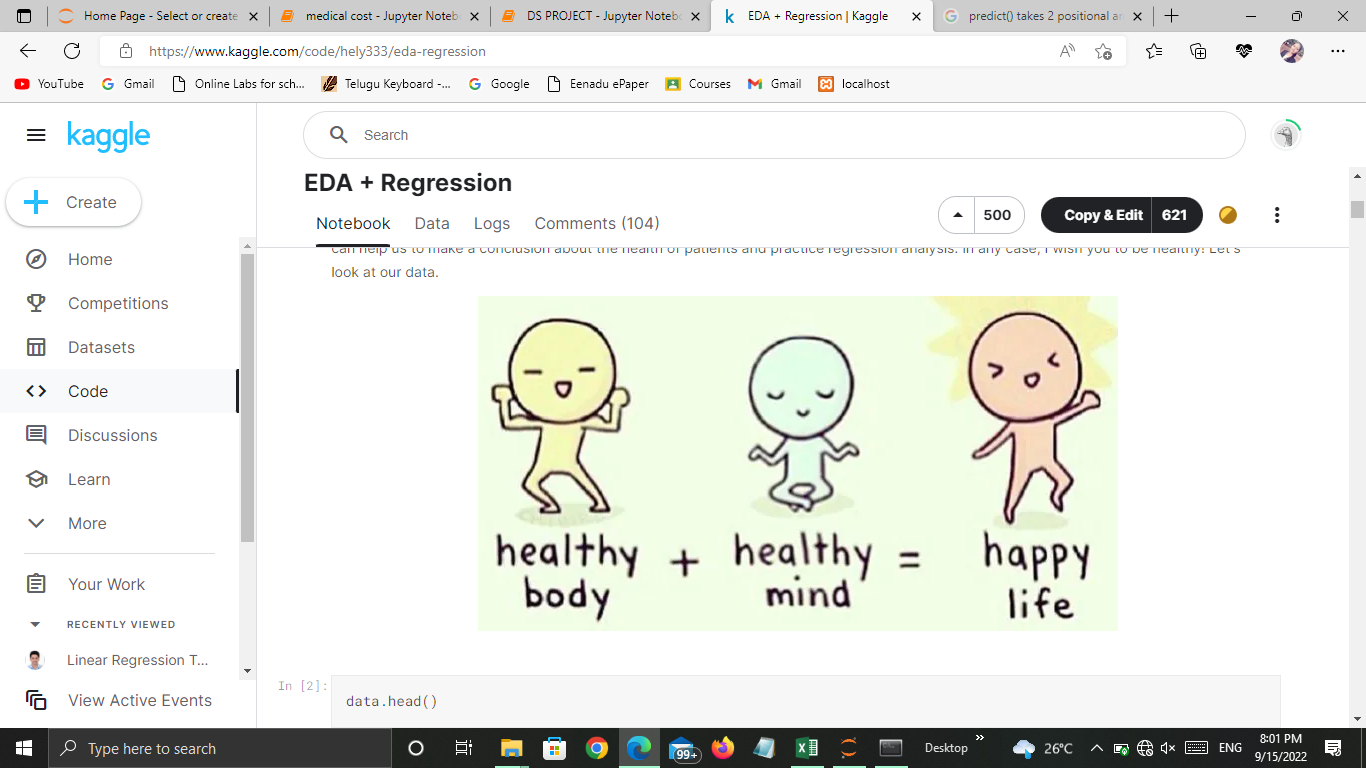
MEDICAL COST PREDICTION



**Problem Statement**

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.

**Objective**

· Predict the future medical expenses of subjects based on certain features building a robust machine learning model.

· Identifying the factors affecting the medical expenses of the subjects based on the model output.

**Prerequisite libraries**

import pandas as pd

import numpy as np

import matplotlib.pyplot as pl

import seaborn as sns

from sklearn.preprocessing import LabelEncoder

from sklearn.linear\_model import LinearRegression

from sklearn.model\_selection import train\_test\_split

## Dataset

## data=pd.read\_csv("D:\medical cost prediction\medical cost.csv")

print(data)

| **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 |
| **...** | ... | ... |  | ... | ... | ... | ... | ... |
| **1333** | 50 | male |  | 30.970 | 3 | no | northwest | 10600.54830 |
| **1334** | 18 | female |  | 31.920 | 0 | no | northeast | 2205.98080 |
| **1335** | 18 | female |  | 36.850 | 0 | no | southeast | 1629.83350 |
| **1336** | 21 | female |  | 25.800 | 0 | no | southwest | 2007.94500 |
| **1337** | 61 | female |  | 29.070 | 0 | yes | northwest | 29141.36030 |

1338 rows × 7 columns

For this project, the data has been imported from the machine learning repository. The dataset contains 1338 rows and 7 columns. The columns present in the dataset are ‘age’,’ sex’,’bmi’, ’children’, smoker’, ’region’, and ‘charges’. The charges column is the target column and the rest others are independent columns. Independent columns are those which will predict the outcome.

The first column is Age. Age is an important factor for predicting medical expenses because young people are generally more healthy than old ones and the medical expenses for Young People will be quite less as compared to old people.

The Next column is sex, which has two Categories in this column: Male and Female. The sex of the person can also play a vital role in predicting the medical expenses of a subject.

After that, you have the ‘bmi’ column, then**BMI is Body Mass Index*.***

For most adults, an ideal BMI is in the 18.5 to 24.9 range.

For children and young people aged 2 to 18, the BMI calculation takes into account age and gender as well as height and weight. If your BMI is less than 18.5, you are considered underweight. People with very low or very high ‘bmi’ are more likely to require medical assistance, resulting in higher costs.

The fourth column is the ‘children’ column, which contains information on how many children your patients have. Persons who have children are under more pressure because of their children’s education, and other needs than people who do not have children.

The fifth is the ‘smoker’ column. The Smoking factor is also considered to be one of the Most Important factors as the people who smoke are always at risk when their age reaches 50 to 60.

Next is the ‘region’ column. Some Regions are Hygienic, Clean, Neat, and Prosperous, But some Regions are not, and this information affects health which is related to medical expenses.

**DatasetDownloadLink:**<https://drive.google.com/uc?id=1WJxHVqSZNkTRLWIdlJR0VIW0vcaCGRPz>

**Data Pre-processing**

data.isnull().sum()

age 0

sex 0

bmi 0

children 0

smoker 0

region 0

charges 0

dtype: int64

**Encoding of categorical columns**

In our dataset, we have three categorical columns: Sex, Smoker, and Region

Sex consists of males and females and it has been observed from the data analysis that males have higher medical expenses than females. As a result, we have encoded male as 1 and female as 0.

#sex

le = LabelEncoder()

le.fit(data.sex.drop\_duplicates())

data.sex = le.transform(data.sex)

print(data.sex)

0 0

1 1

2 1

3 1

4 1

..

1333 1

1334 0

1335 0

1336 0

1337 0

Name: sex, Length: 1338, dtype: int32

The Smoker column consists of smokers and nonsmokers and it has been observed the smoker has more medical expenses than the nonsmoker ones so we encoded smokers as 1 and nonsmokers as 0.

# smoker or not

le.fit(data.smoker.drop\_duplicates())

data.smoker = le.transform(data.smoker)

print(data.smoker)

0 1

1 0

2 0

3 0

4 0

..

1333 0

1334 0

1335 0

1336 0

1337 1

Name: smoker, Length: 1338, dtype: int32

The Region column comprises four segments: Southeast, Southwest, Northeast, Northwest.

It has been noticed from the data analysis the Southeast region has the highest expenses followed by Northeast, Northwest, and Southwest, so we have encoded those regions as 3,2,1,0 respectively.

#region

le.fit(data.region.drop\_duplicates())

data.region = le.transform(data.region)

print(data.region)

0 3

1 2

2 2

3 1

4 1

..

1333 1

1334 0

1335 2

1336 3

1337 1

Name: region, Length: 1338, dtype: int32

**Correlation**

Finding the realtion of the columns with the target column charges.

data.corr()['charges'].sort\_values()

region -0.006208

sex 0.057292

children 0.067998

bmi 0.198341

age 0.299008

smoker 0.787251

charges 1.000000

Name: charges, dtype: float64

## Data Visualization

Here we have used a pie chart to plot the Smoker Column, as the Smoker column has only two values: **Yes and No**. We have found 20.48% of the subjects are smokers and 79.52 % are non-smoker.

Using a Count plot we have shown the subjects having children ranging from 0 to 5 and it has been computed and observed from the count plot also that those who are having no children are highest in number.

* Number of Subjects having no children- 574
* Number of Subjects having one child- 324
* Number of Subjects having two children- 240
* Number of Subjects having three children- 157
* Number of Subjects having four children- 25
* Number of Subjects having five children-18
* We have again used a pie chart to plot the number of inhabitants in the region column which consists of four segments: Northeast, Northwest, Southeast, Southwest

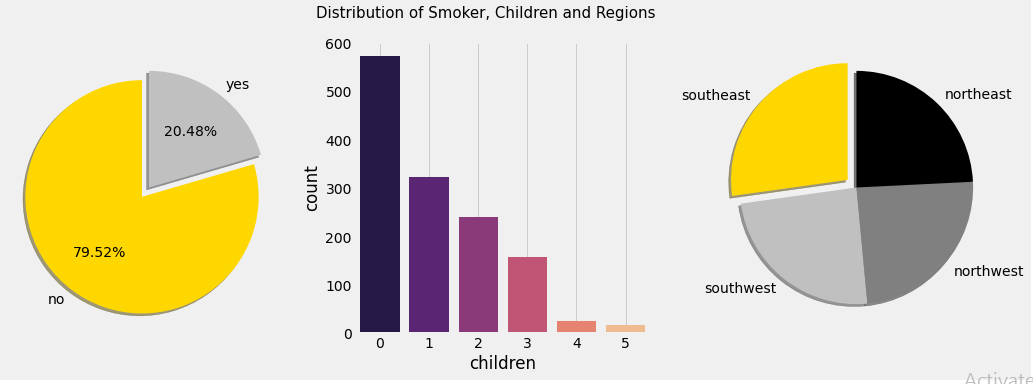
The number of Southwest and Northwest are the same and the value is 324, but the number of inhabitants in Northeast and Southeast are respectively 324 and 364.

The plots are shown below.

sns.pieplot(x=”smoker”,kind=”count”,palette=”ch:,25”,data=data)

sns.catplot(x="children", kind="count", palette="ch:.25", data=data,height= 6)

sns.pieplot(x=”region”,kind=”count”,palette=”ch:,25”,data=data)



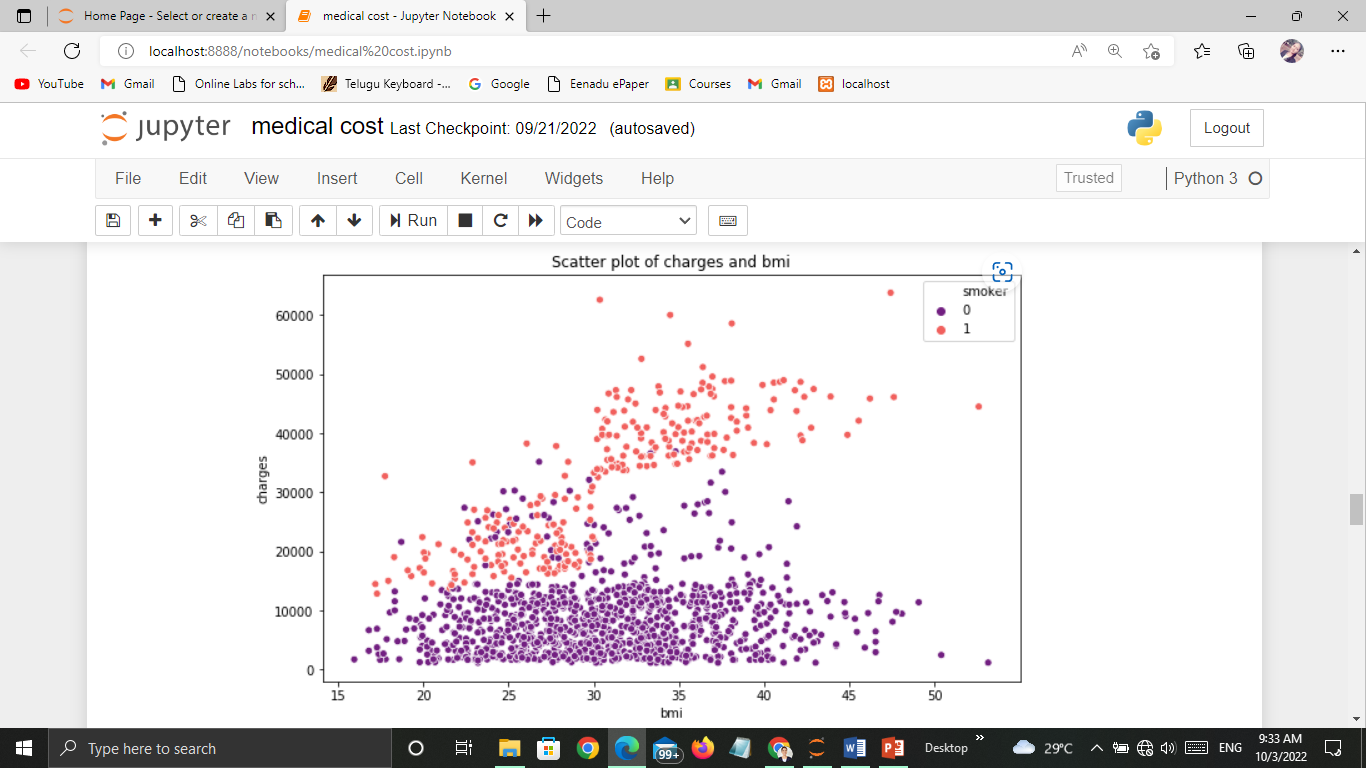
Distribution of Smoker, Children, and Regions

pl.figure(figsize=(10,6))

ax = sns.scatterplot(x='bmi',y='charges',data=data,palette='magma',hue='smoker')

ax.set\_title('Scatter plot of charges and bmi')

sns.lmplot(x="bmi", y="charges", hue="smoker", data=data, palette = 'magma',height= 8)



**Data Splitting**

We have split the data in the entire dataset in Train and Test with a ratio of 80:20. We have trained the model with trained data and then we applied the model to the test dataset to check the performance of the model.

x= data.drop(['charges','region'], axis = 1)

print(x)

|  | **age** | **sex** | **bmi** | **children** | **smoker** |
| --- | --- | --- | --- | --- | --- |
| **0** | 19 | 0 | 27.900 | 0 | 1 |
| **1** | 18 | 1 | 33.770 | 1 | 0 |
| **2** | 28 | 1 | 33.000 | 3 | 0 |
| **3** | 33 | 1 | 22.705 | 0 | 0 |
| **4** | 32 | 1 | 28.880 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... |
| **1333** | 50 | 1 | 30.970 | 3 | 0 |
| **1334** | 18 | 0 | 31.920 | 0 | 0 |
| **1335** | 18 | 0 | 36.850 | 0 | 0 |
| **1336** | 21 | 0 | 25.800 | 0 | 0 |
| **1337** | 61 | 0 | 29.070 | 0 | 1 |

1338 rows × 5 columns

y= data.charges

print(y)

0 16884.92400

1 1725.55230

2 4449.46200

3 21984.47061

4 3866.85520

...

1333 10600.54830

1334 2205.98080

1335 1629.83350

1336 2007.94500

1337 29141.36030

Name: charges, Length: 1338, dtype: float64

x\_train,x\_test,y\_train,y\_test = train\_test\_split(x,y,test\_size=0.21,random\_state=0)

**Model Building**

Linear Regression was applied to predicted Future Medical Expenses for your Patients based on certain features such as *Age, Gender, Region, Smoking Behavior, and Number of children.*

lr = LinearRegression().fit(x\_train,y\_train)

y\_train\_pred = lr.predict(x\_train)

y\_test\_pred = lr.predict(x\_test)

print(lr.score(x\_test,y\_test))

0.8000126868215097

**Prediction**

a= lr.predict([[21,0,25.800,0,0]])

a[0]

1759.1069456992336

b=lr.predict([[32,1,30.97,3,0]])

b[0]

7435.977717289981

## Conclusion:

We came to know that the Most Important Factor to Predict the Medical Expenses of a subject is Smoking Behavior and Age, that means, smoking is Bad for Health, as already know that and which inevitably increases medical expenses as due to smoking one is likely to fall ill more than the nonsmokers.  
We also found that with increasing of age, one needs to take some more care and precautions for your health as with the increase of age health becomes fragile so they go for frequent medical check-up, likely to fall ill quickly as with the increase of age immunity falls so they adopt measures to stay healthy by taking medicines and engaging in some physical activities like jogging, walking, Yoga which causes an increase of medical expenses.