**Machine Learning Regression**

**Assignment-Regression Algorithm**

**Insurance charges prediction**

**Multiple linear Regression R2-Score = 0.7894790349867009**

1.Find out the 3 -Stage of Problem Identification

Stage1- Machine Learning

Stage2- supervised Learning

Stage3- Regression

2.Name the project

Insurance charges prediction

3.Tell basic info about the dataset (Total number of rows, columns)

1. Total number of rows

1338 rows

2. Total number of columns – 6 Columns

1. age 2. Sex 3. Bmi 4. Children 5. Smoker 6. Charges

|  | **age** | **sex** | **bmi** | **children** | **smoker** | **charges** |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | 19 | female | 27.900 | 0 | yes | 16884.92400 |
| 1 | 18 | male | 33.770 | 1 | no | 1725.55230 |
| 2 | 28 | male | 33.000 | 3 | no | 4449.46200 |
| 3 | 33 | male | 22.705 | 0 | no | 21984.47061 |
| 4 | 32 | male | 28.880 | 0 | no | 3866.85520 |
| ... | ... | ... | ... | ... | ... | ... |
| 1333 | 50 | male | 30.970 | 3 | no | 10600.54830 |
| 1334 | 18 | female | 31.920 | 0 | no | 2205.98080 |
| 1335 | 18 | female | 36.850 | 0 | no | 1629.83350 |
| 1336 | 21 | female | 25.800 | 0 | no | 2007.94500 |
| 1337 | 61 | female | 29.070 | 0 | yes | 29141.36030 |

1338 rows × 6 columns

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |

**4.use to categorical to numerical coverted**

dataset=pd.get\_dummies(dataset)

dataset

| **age** | **bmi** | **children** | **charges** | **sex\_female** | **sex\_male** | **smoker\_no** | **smoker\_yes** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 19 | 27.900 | 0 | 16884.92400 | 1 | 0 | 0 | 1 |
| 1 | 18 | 33.770 | 1 | 1725.55230 | 0 | 1 | 1 | 0 |
| 2 | 28 | 33.000 | 3 | 4449.46200 | 0 | 1 | 1 | 0 |
| 3 | 33 | 22.705 | 0 | 21984.47061 | 0 | 1 | 1 | 0 |
| 4 | 32 | 28.880 | 0 | 3866.85520 | 0 | 1 | 1 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 1333 | 50 | 30.970 | 3 | 10600.54830 | 0 | 1 | 1 | 0 |
| 1334 | 18 | 31.920 | 0 | 2205.98080 | 1 | 0 | 1 | 0 |
| 1335 | 18 | 36.850 | 0 | 1629.83350 | 1 | 0 | 1 | 0 |
|  |  |  |  |  |  |  |  |  |
| 1336 | 21 | 25.800 | 0 | 2007.94500 | 1 | 0 | 1 | 0 |
| 1337 | 61 | 29.070 | 0 | 29141.36030 | 1 | 0 | 0 | 1 |

1338 rows × 8 columns

**5.Input/output split the dataset**

**Input split**

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

1338 rows × 7 columns

**Output Split**

| **charges** |
| --- |
| 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 |

1338 rows × 1 columns

**6.Train-set and test-set split**

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

x\_train,x\_test,y\_train,y\_test=train\_test\_split(independent,dependent,test\_size=0.30,random\_state=0)

**7.model creation**

Using multiple linear regression algoritham

And using fit methed to created model

**from sklearn.linear\_model import LinearRegression**

**regressor=LinearRegression()**

**regressor.fit(x\_train,y\_train)**

**# view the wieght/slope value**

weight=regressor.coef\_

weight

array([[ 257.8006705 , 321.06004271, 469.58113407,

20.87412859, -20.87412859, -11709.3335956 ,

11709.3335956 ]])

**# view the bias/intercept value**

bias=regressor.intercept\_

bias

array([-368.78537899])

**# Evaluvate metrics**

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

**8.** **Develop a good model with r2\_score**

**#R2\_square with using metrics called a test to predicted y test**

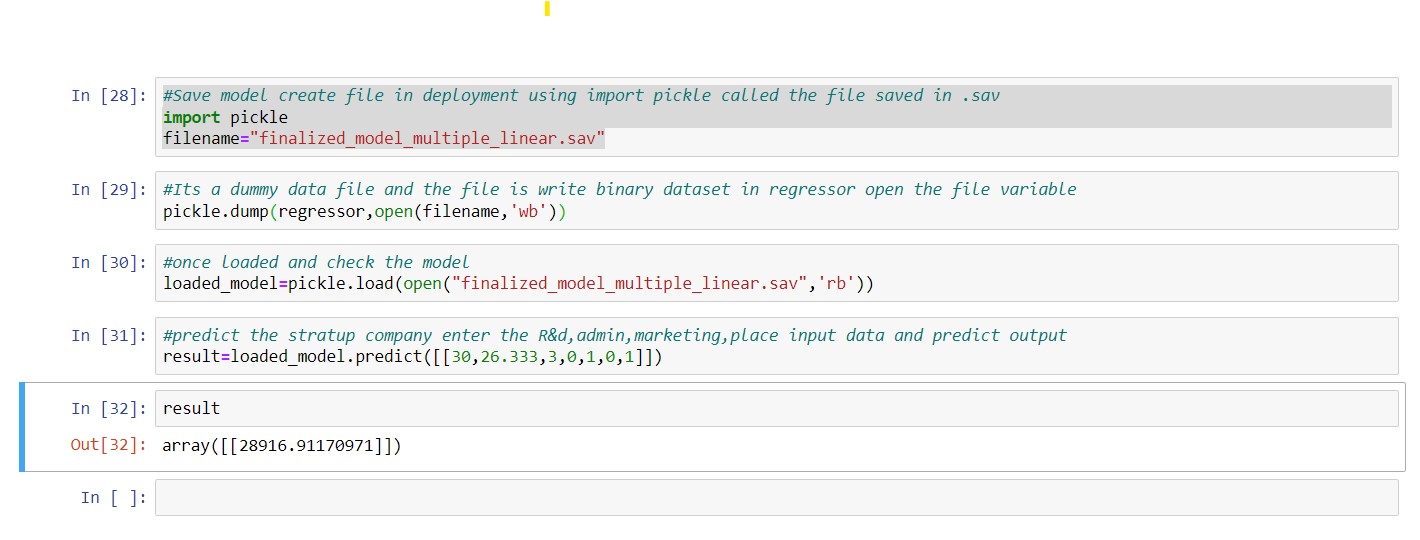
from sklearn.metrics import r2\_score

r\_score=r2\_score(y\_test,y\_pred)

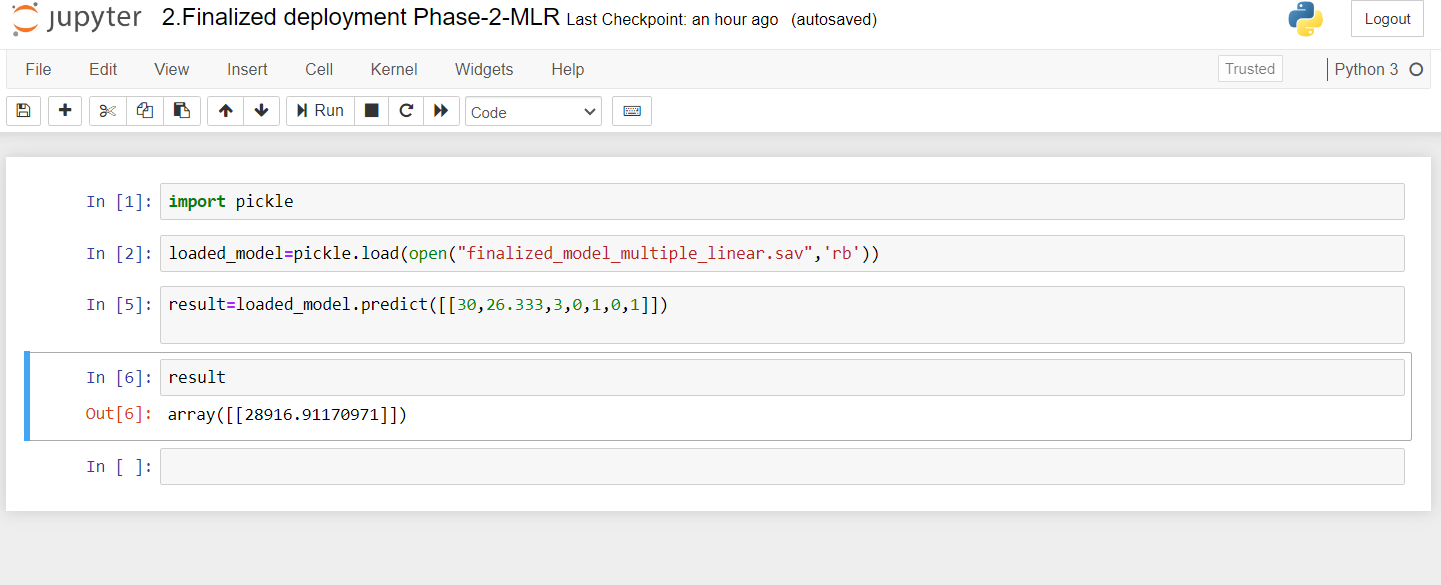
**r\_score**

**0.7894790349867009**

**9.Save mode and Finalized model create new file**



**10.deployment creation**

****