MEDICAL INSURANCE COST PREDICTION

Team Members:
20BDS0146 - VENNELA G
20BDS0172 - AMOGH A M
20BDS0270 - A KIREETI
20BDS0352 - A HARI PRIYA

Report submitted for the

Final Project Review of

Course Code: CSE3045

Predictive Analysis

Slot: E2 Slot

Professor: Dr. GUNAVATHI C

INTRODUCTION:

- we are on a planet full of threats and uncertainty. people, households, companies, properties, and property are exposed to different risk forms.
- And the risk levels can vary. these dangers contain the risk of death, health, and property loss or assets.
- life and wellbeing are the greatest parts of people's lives. but, risks cannot usually be avoided, so the world of finance has developed numerous products to shield individuals and organizations from these risks by using financial capital to reimburse them.
- Insurance is, therefore, a policy that decreases or removes loss costs incurred by various risks.
- Concerning the value of insurance in the lives of individuals, it becomes important for the companies of insurance to be sufficiently precise to measure or quantify the amount covered by this policy and the insurance charges which must be paid for it.
- Various variables estimates these charges. each factor of these is important. if any factor is omitted when the amounts are computed, the policy changes overall.
- It is therefore critical that these tasks are performed with high accuracy. as human mistakes are could occur, insurers use people with experience in this area.
- They also use different tools to calculate the insurance premium. ml is beneficial here. ml may generalize the effort or method to formulate the policy.
- These ml models can be learned by themselves. the model is trained on insurance data from the past.
- The requisite factors to measure the payments can then be defined as the model inputs, then the model can correctly anticipate insurance policy costs.
- This decreases human effort and resources and improves the company's profitability. thus the accuracies can be improved with ml.

LITERATURE REVIEW SUMMARY TABLE:

Literature review Summary table:

Authors and Year (Reference)	***Title (Study)***	***Concept / Theoretical model/ Framework***	***Methodology used/ Implementation***	***Dataset details/ Analysis***	***Relevant Finding***
MOHAMMED HANFEY, Omar M. A. Mahmoud (2021)	Predict Health Insurance Cost by using Machine Learning and DNN Regression Models	The research uses various machine learning regression models and deep neural networks to forecast charges of health insurance based on specific attributes, on medical cost personal data set from Kaggle	The findings are summarized in Table IV. shows that Stochastic Gradient Boosting offers the best efficiency, with an RMSE value of 0.380189, an MAE value of 0.17448, and an accuracy of 85.8	the data set is separated into two-part the first part called training data, and the second called test data; training data makes up about 80 percent of the total data used, and the rest for test data The training data set is applied to build a model as a predictor of medical insurance cost year and the test set will use to evaluate the regression model	Machine learning (ML) for the insurance industry sector can make the wording of insurance policies more efficient. This study demonstrates how different models of regression can forecast insurance costs

Dr. Akhilesh Das Gupta (2020)	Health Insurance Amount Prediction	The goal of this project is to allows a person to get an idea about the necessary amount required according to their own health statu	Three regression models naming Multiple Linear Regression, Decision tree Regression and Gradient Boosting Decision tree Regression have been used to compare and contrast the performance of these algorithms	The primary source of data for this project was from Kaggle user Dmarco. The dataset is comprised of 1338 records with 6 attributes. Attributes are as follow 'age', 'gender', 'bmi', 'children', 'smoker' and 'charges'	We see that the accuracy of predicted amount was seen best i.e. 99.5% in gradient boosting decision tree regression. Other two regression models also gave good accuracies about 80%.
Saddam Hussain, Mogeeb A. A. Mosleh	A Computational Intelligence Approach for Predicting Medical Insurance Cost	In this study, we used supervised ML models to demonstrate and compare the accuracy of various regression models, including Linear Regression (LR), Stochastic Gradient Boosting (SGB), XGBoost (XGB)	The proposed research approach uses Linear Regression, Support Vector Regression, Ridge Regressor, Stochastic Gradient Boosting, XGBoost, Decision Tree, Random Forest Regressor, Multiple Linear Regression, and k-Nearest Neighbors	The medical cost personal datasets are obtained from the KAGGLE repository. This dataset contains seven attributes	Data mining (DM) and machine learning (ML) techniques are widely used for insurance cost prediction and medical fraud detection. Using the Extreme Gradient Boosting algorithm, we improved the accuracy of a decision tree classifier for predicting healthcare insurance fraud.

Nataliya Shakhovska1, Valentyna Chopiyak2 and Michal Gregus ml3	An Ensemble Methods for Medical Insurance Costs Prediction Task	The paper reports three new ensembles of supervised learning predictors for managing medical insurance costs. The open dataset is used for data analysis methods development.	bagging shows its weakness in generalizing the prediction. The stacking is developed using K Nearest Neighbors (KNN), Support Vector Machine (SVM), Regression Tree, Linear Regression, Stochastic Gradient Boosting	The medical insurance payments dataset [29] was selected. It consists of 7 attributes and 1338 vectors. The task is to predict individual payments for health insurance.	two feature selection technics for the comparison of the prediction accuracy of the different machine learning algorithms were applied. The weak components for the design an ensemble models were found

OBJECTIVE OF THE PROJECT:

Let's say that you are a machine learning expert or you are a data scientist and there is a medical insurance company and this company wants to create automatic system that can predict what is the medical insurance cost of a person will be.

- How we can build a machine learning System that can predict what is the medical insurance cost of a person.
- Insurance is a policy that eliminates or decreases loss costs occurred by various risks. Various factors influence the cost of insurance.
- These considerations contribute to the insurance policy formulation.

- Machine learning (ML) for the insurance industry sector can make the wording of insurance policies more efficient.
- This study demonstrates how different models of regression can forecast insurance costs.
- And we will compare the results of models, for example, Multiple Linear Regression, Generalized Additive Model, Support Vector Machine, Random Forest Regressor, CART, XGBoost, k-Nearest Neighbors, Stochastic Gradient Boosting, and Deep Neural Network.

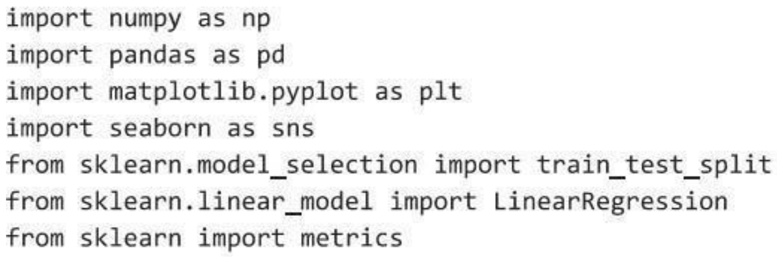
INNOVATION COMPONENT OF THE PROJECT:

- Model drug development collaborations that maximize IP and drug discovery.
- Simulate PRO (Patient Reported Outcomes) for care quality improvement and outcomes.
- Accelerate time to market for new therapies with strategic portfolio modeling.
- Predict market access and optimize resource allocation for new therapies.
- Predict high risk patients for ACO (accountable care organization) and hospitals.
- Leverage advanced analytics to reduce hospital readmissions.
- Simulate connected health consumer and recommend technology interventions that drive healthy behavior change.

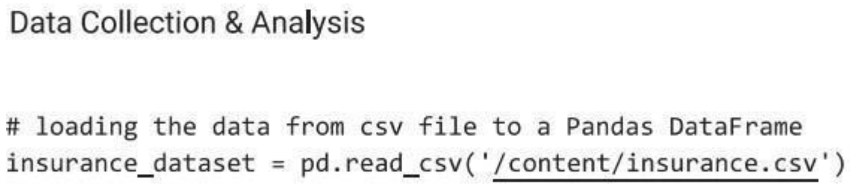
WORK DONE AND IMPLEMENTATION:

- Linear Regression will be implemented with automatic feature selection using backward elimination.
- Starting from using all features, the backward elimination process will iteratively discard some and evaluate the model until it finds one with the lowest Akaike Information Criterion (AIC).
- Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models based on information loss. Lower AIC means better model.

Importing Data Set:

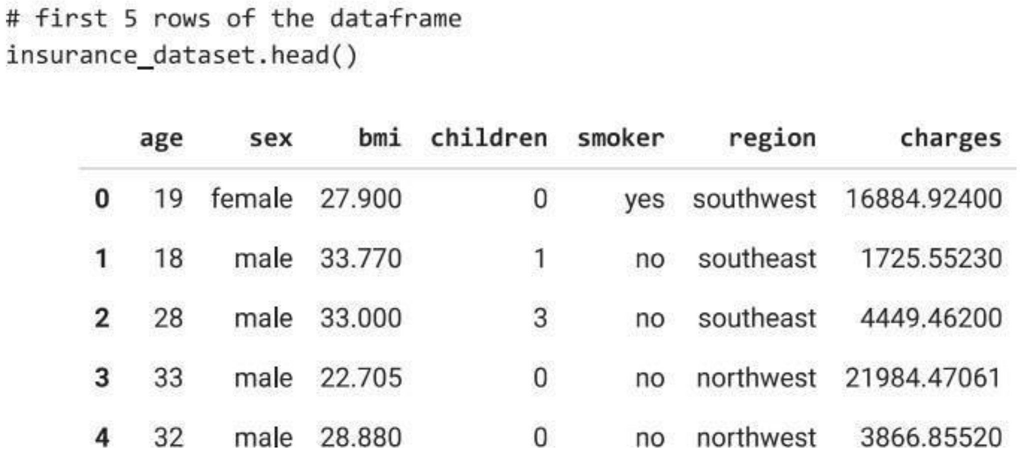


```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
```



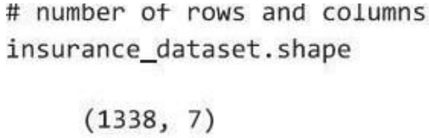
Data Collection & Analysis

```
# loading the data from csv file to a Pandas DataFrame
insurance_dataset = pd.read_csv('/content/insurance.csv')
```



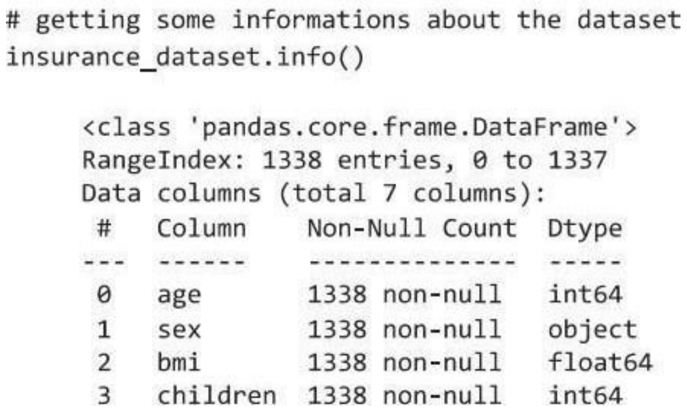
```
# first 5 rows of the dataframe
insurance_dataset.head()
```

	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



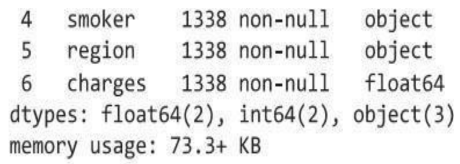
```
# number of rows and columns
insurance_dataset.shape

     (1338, 7)
```

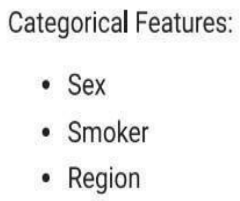


```
# getting some informations about the dataset
insurance_dataset.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 1338 entries, 0 to 1337
     Data columns (total 7 columns):
      #   Column    Non-Null Count  Dtype
     ---  ------    --------------  -----
      0   age       1338 non-null   int64
      1   sex       1338 non-null   object
      2   bmi       1338 non-null   float64
      3   children  1338 non-null   int64
```

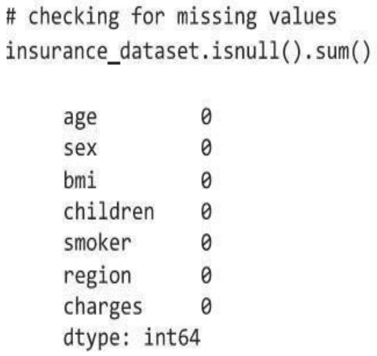


```
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```



Categorical Features:

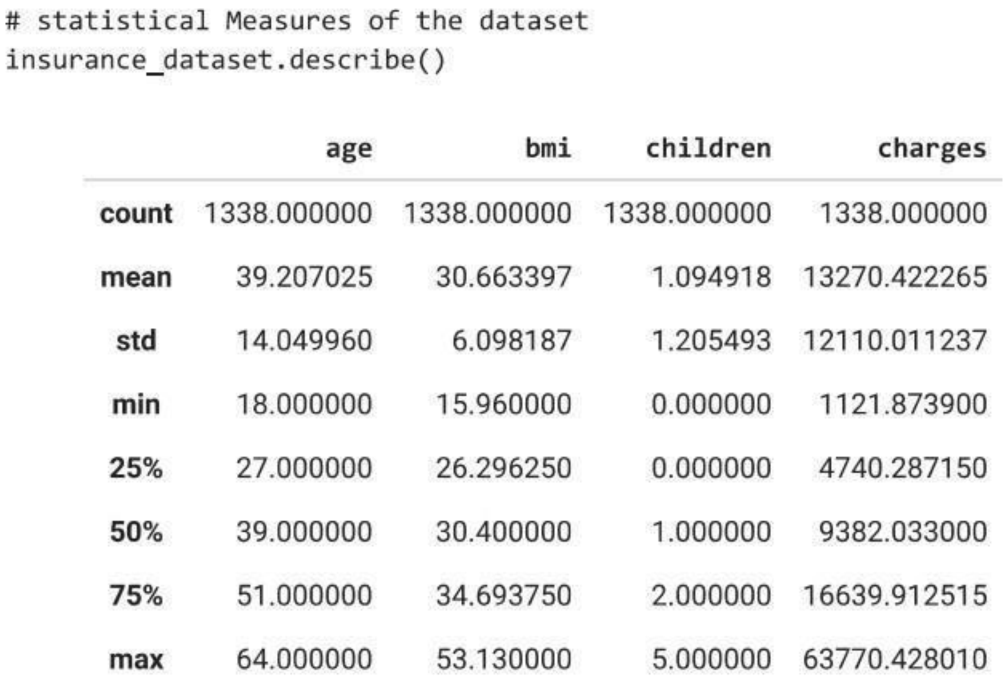
- Sex
- Smoker
- Region



```
# checking for missing values
insurance_dataset.isnull().sum()
```

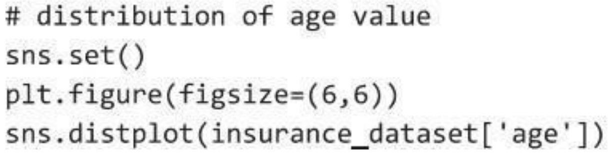
```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

Data analysis:

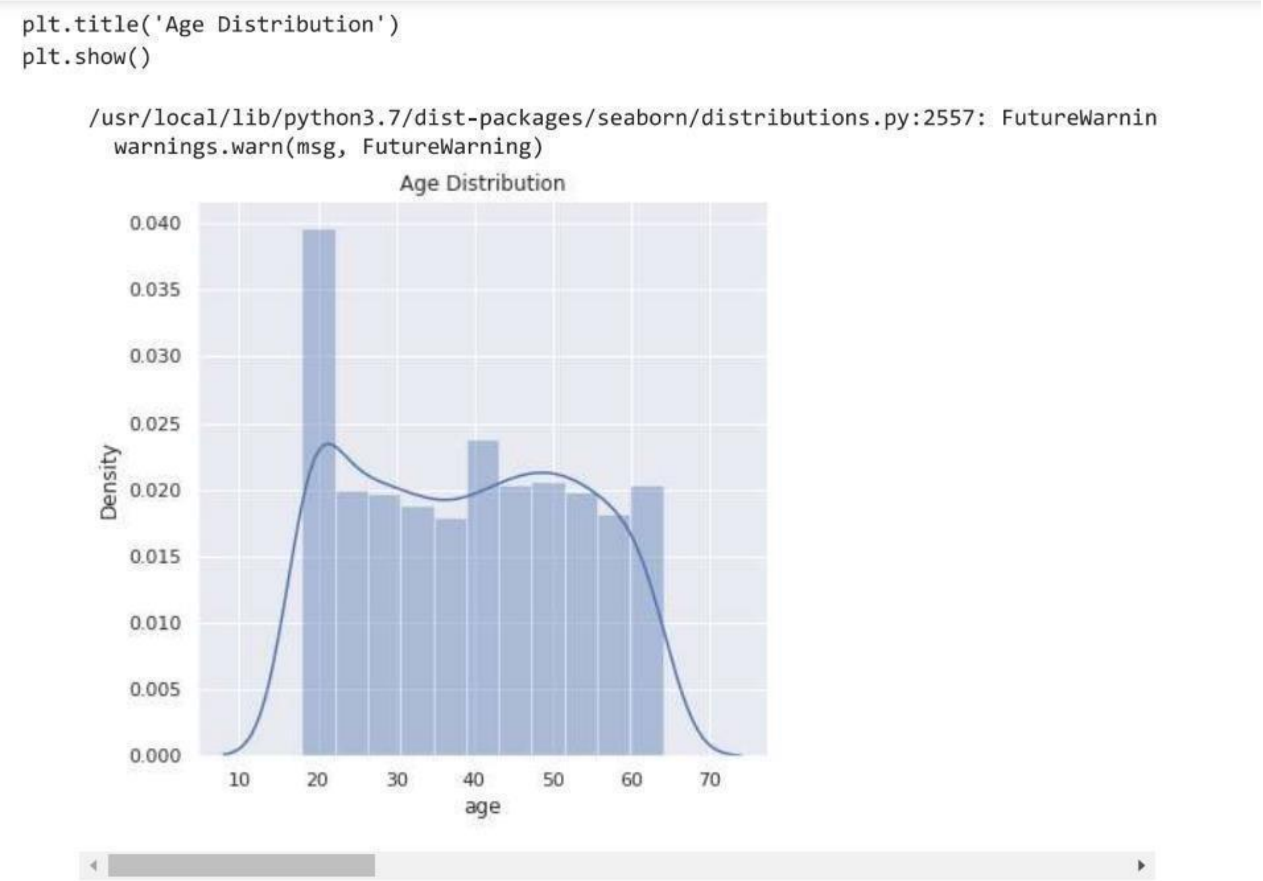


```
# statistical Measures of the dataset
insurance_dataset.describe()
```

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

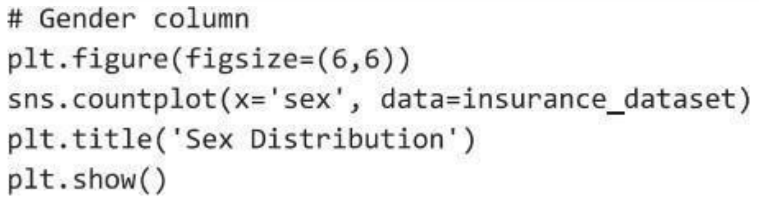


```
# distribution of age value
sns.set()
plt.figure(figsize=(6,6))
sns.distplot(insurance_dataset['age'])
```

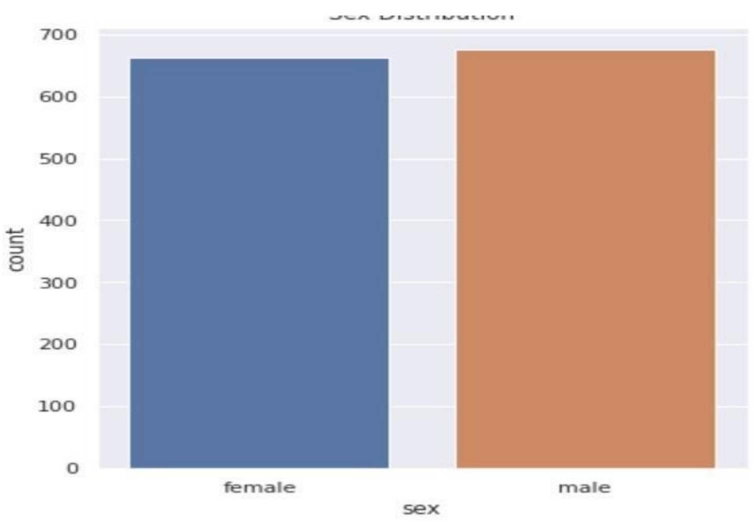


```
plt.title('Age Distribution')
plt.show()

    /usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2557: FutureWarnin
      warnings.warn(msg, FutureWarning)
```



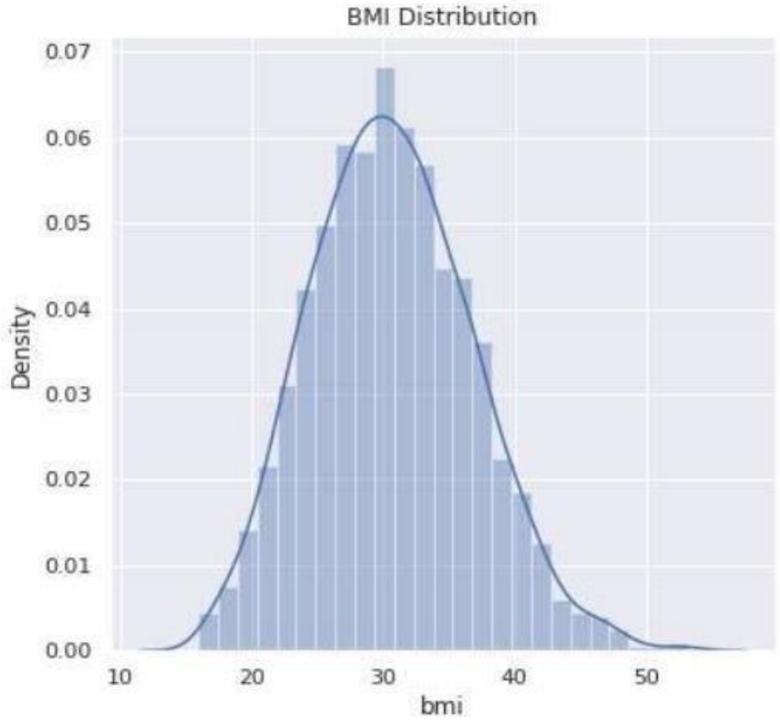
```
# Gender column
plt.figure(figsize=(6,6))
sns.countplot(x='sex', data=insurance_dataset)
plt.title('Sex Distribution')
plt.show()
```



```
[ ]  insurance_dataset['sex'].value_counts()
```

```
male      676
female    662
Name: sex, dtype: int64
```

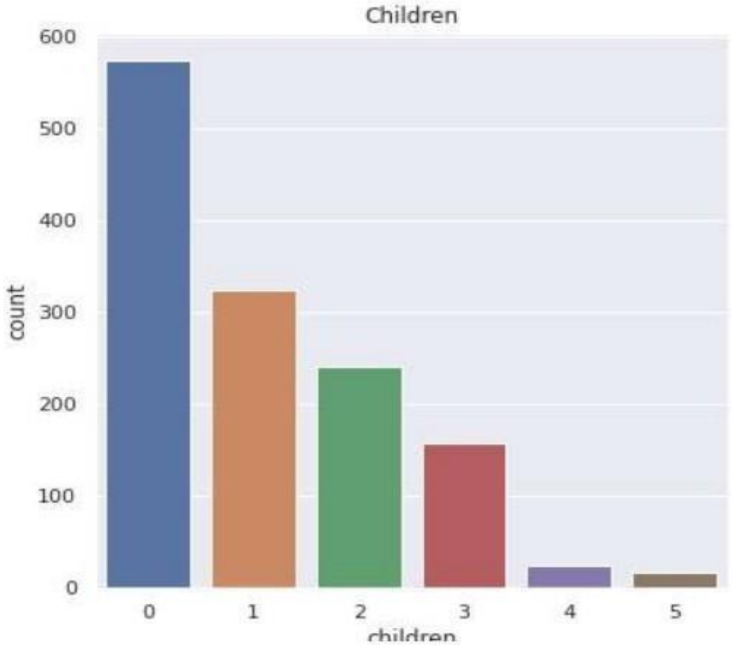
```
# bmi distribution
plt.figure(figsize=(6,6))
sns.distplot(insurance_dataset['bmi'])
plt.title('BMI Distribution')
plt.show()
```



Normal BMI Range --> 18.5 to 24.9

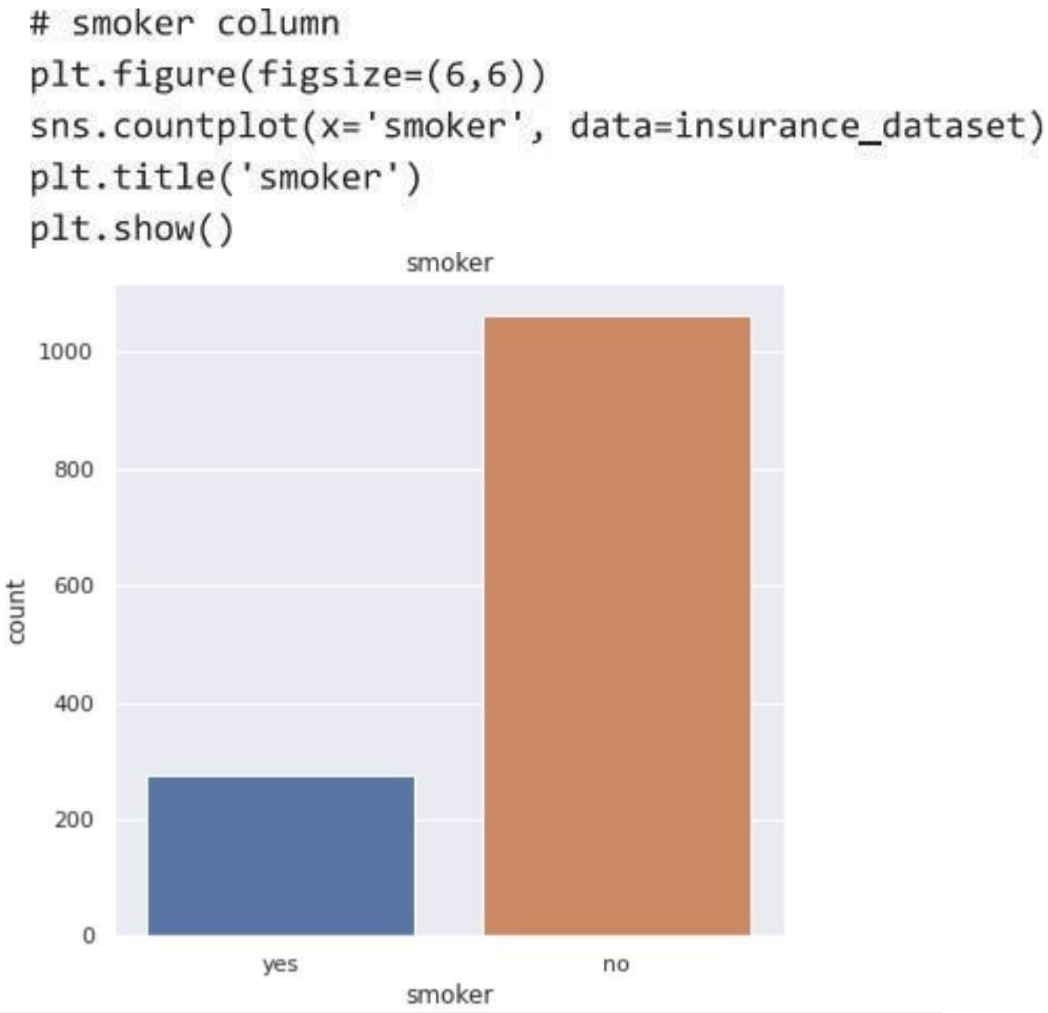
```
# children column
plt.figure(figsize=(6,6))
sns.countplot(x='children', data=insurance_dataset)
```

```
plt.title('Children')
plt.show()
```

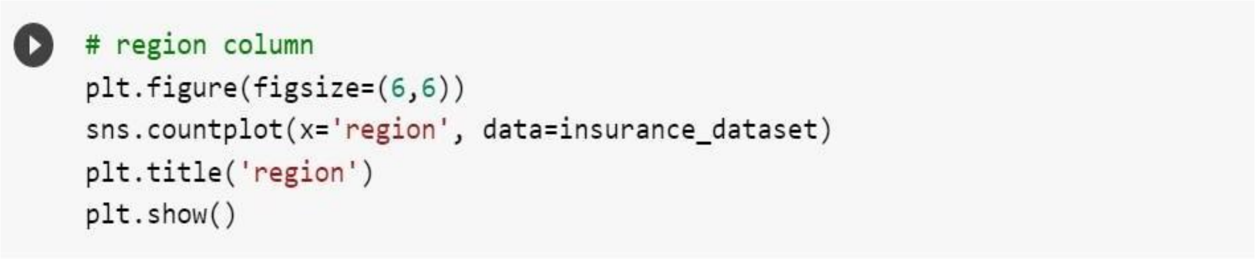


```
insurance_dataset['children'].value_counts()
```

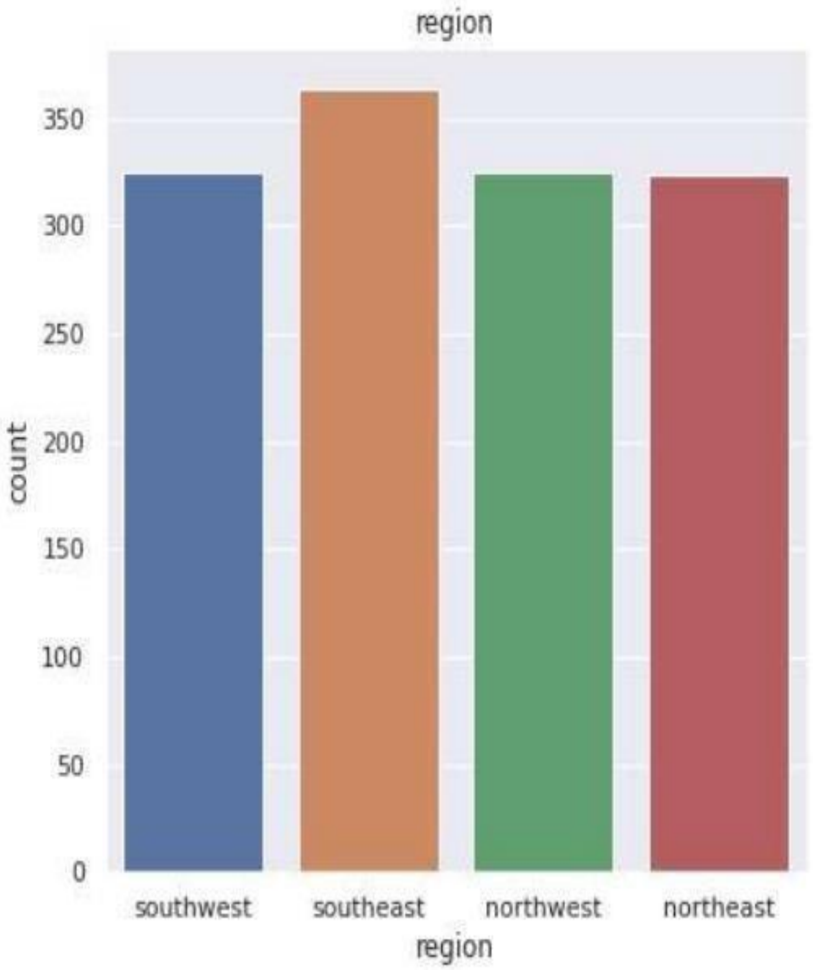
```
0    574
1    324
2    240
3    157
4     25
5     18
Name: children, dtype: int64
```



```
# smoker column
plt.figure(figsize=(6,6))
sns.countplot(x='smoker', data=insurance_dataset)
plt.title('smoker')
plt.show()
```

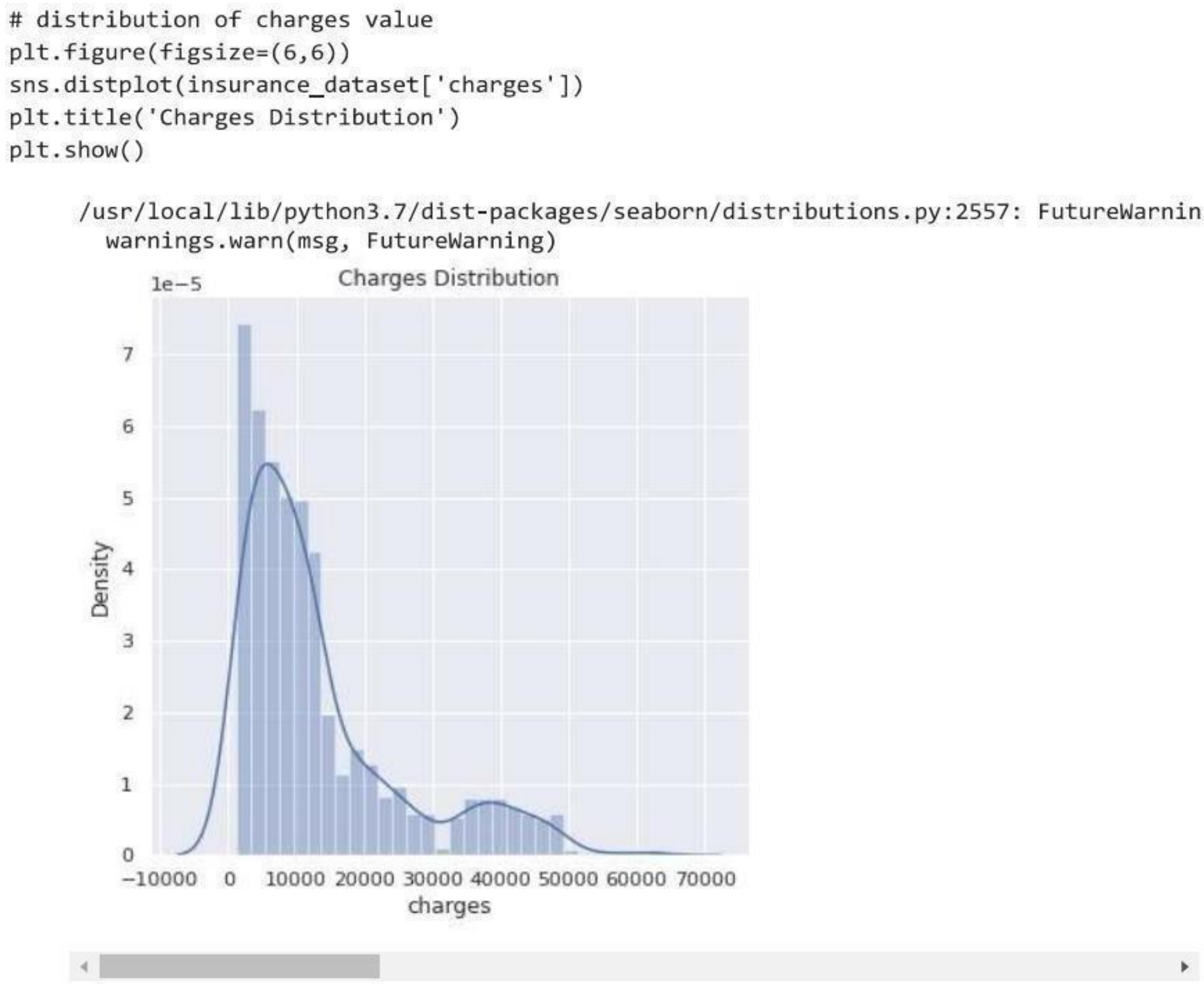


```
# region column
plt.figure(figsize=(6,6))
sns.countplot(x='region', data=insurance_dataset)
plt.title('region')
plt.show()
```



```
insurance_dataset['region'].value_counts()
```

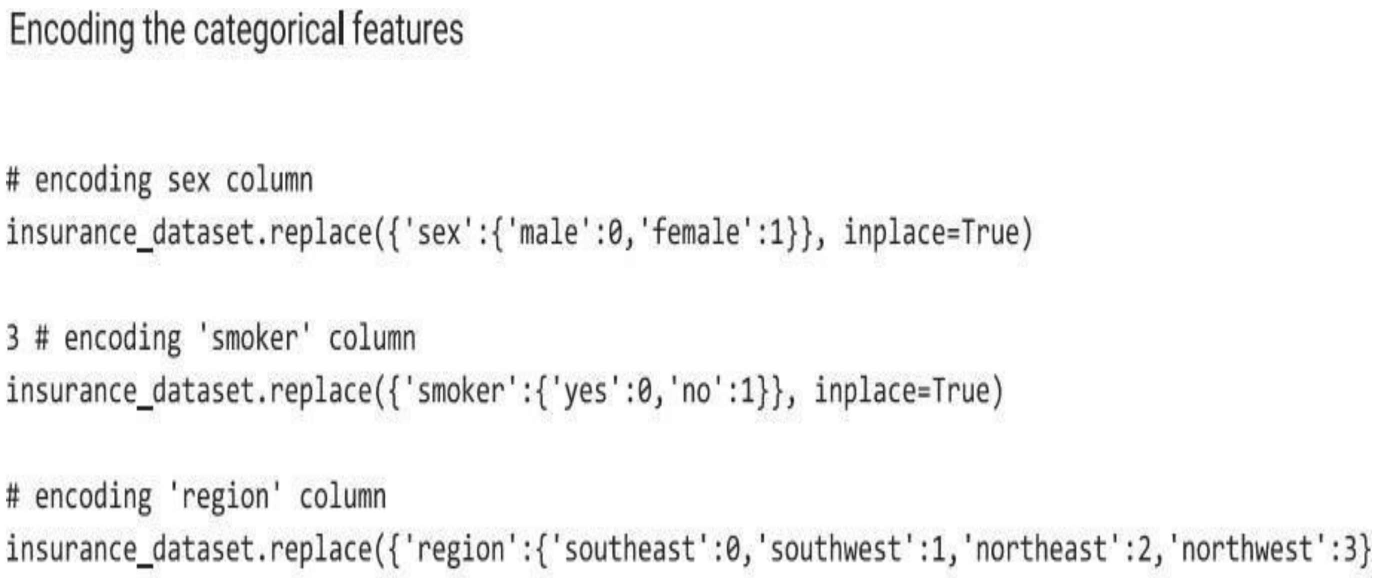
```
    southeast    364
    northwest    325
    southwest    325
    northeast    324
    Name: region, dtype: int64
```



```
# distribution of charges value
plt.figure(figsize=(6,6))
sns.distplot(insurance_dataset['charges'])
plt.title('Charges Distribution')
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2557: FutureWarnin
  warnings.warn(msg, FutureWarning)
```

Data Preprocessing:



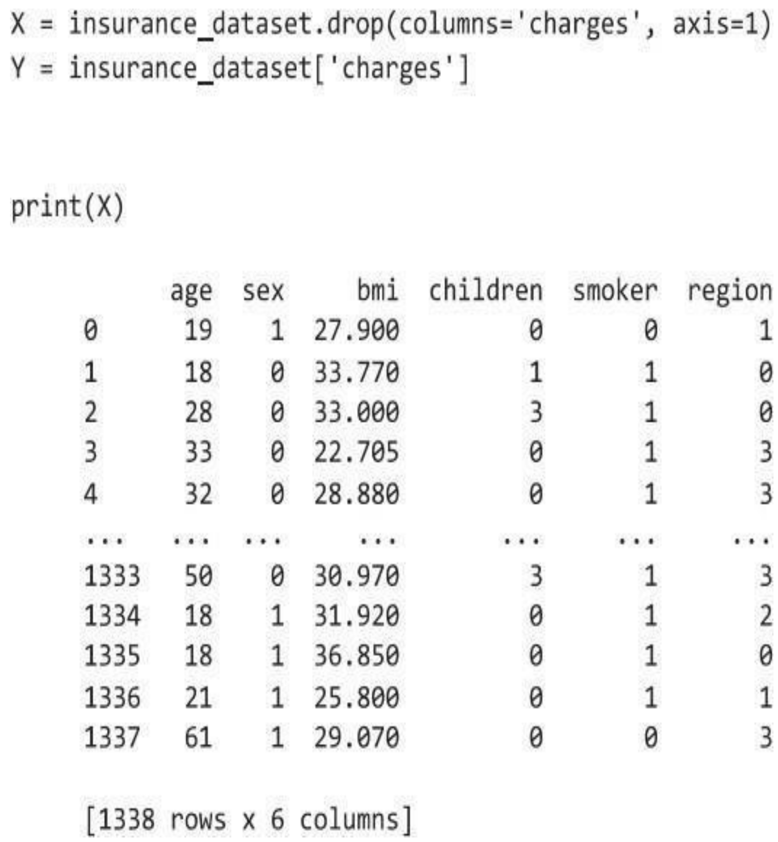
Encoding the categorical features

```
# encoding sex column
insurance_dataset.replace({'sex':{'male':0,'female':1}}, inplace=True)

3 # encoding 'smoker' column
insurance_dataset.replace({'smoker':{'yes':0,'no':1}}, inplace=True)

# encoding 'region' column
insurance_dataset.replace({'region':{'southeast':0,'southwest':1,'northeast':2,'northwest':3}
```

Splitting the Features and Target

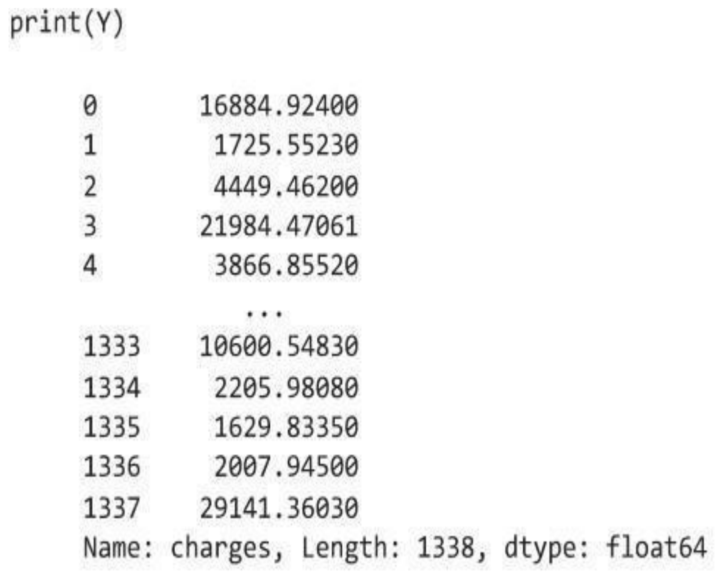


```
X = insurance_dataset.drop(columns='charges', axis=1)
Y = insurance_dataset['charges']
```

```
print(X)
```

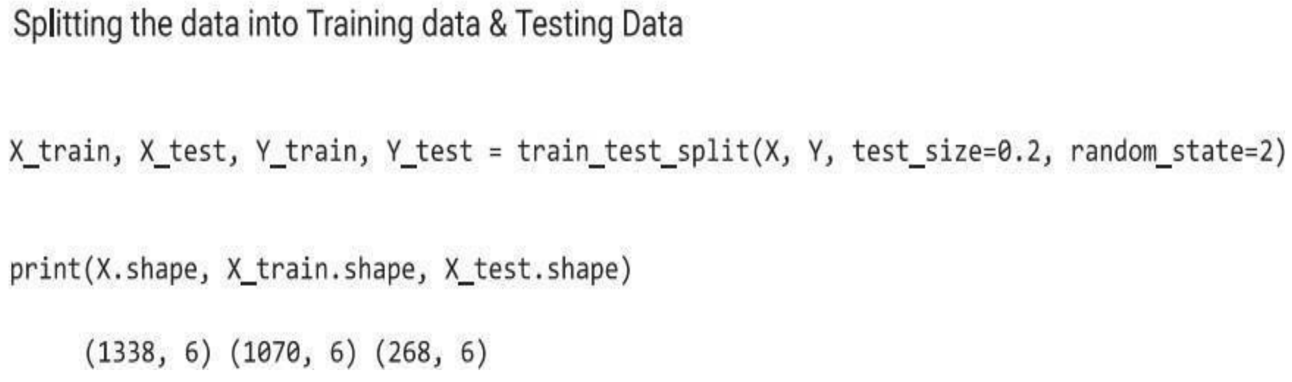
```
      age  sex     bmi  children  smoker  region
0      19    1  27.900         0       0       1
1      18    0  33.770         1       1       0
2      28    0  33.000         3       1       0
3      33    0  22.705         0       1       3
4      32    0  28.880         0       1       3
...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

[1338 rows x 6 columns]
```



```
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
```



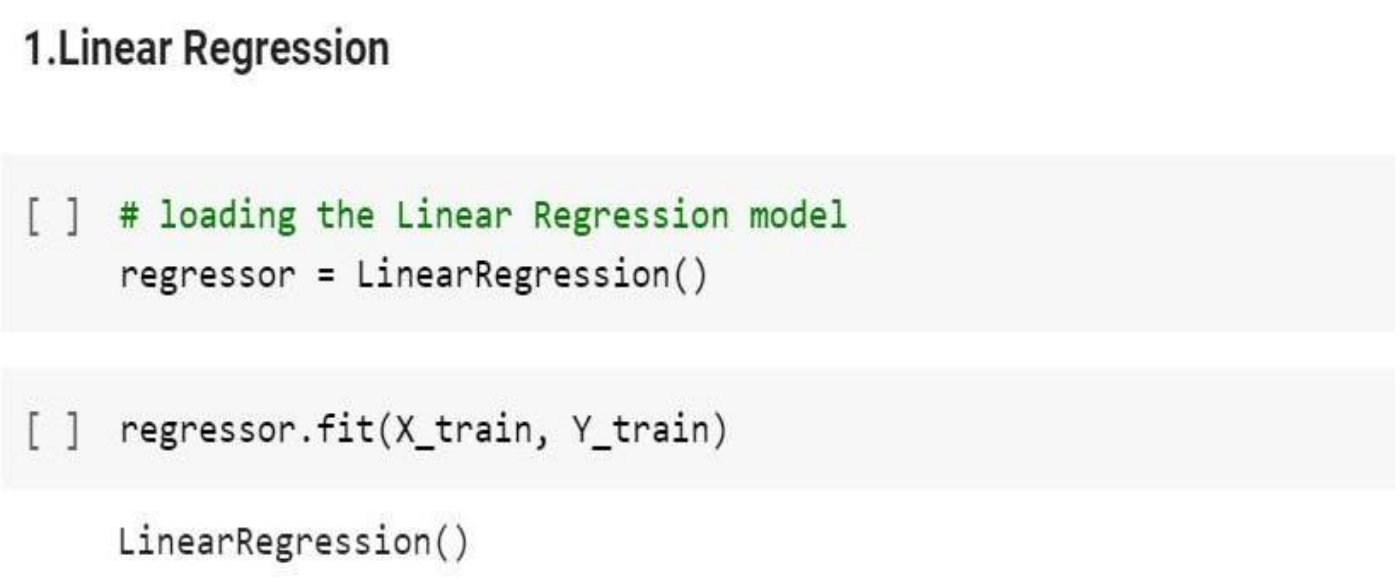
Splitting the data into Training data & Testing Data

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)
```

```
print(X.shape, X_train.shape, X_test.shape)

     (1338, 6) (1070, 6) (268, 6)
```

Model training:

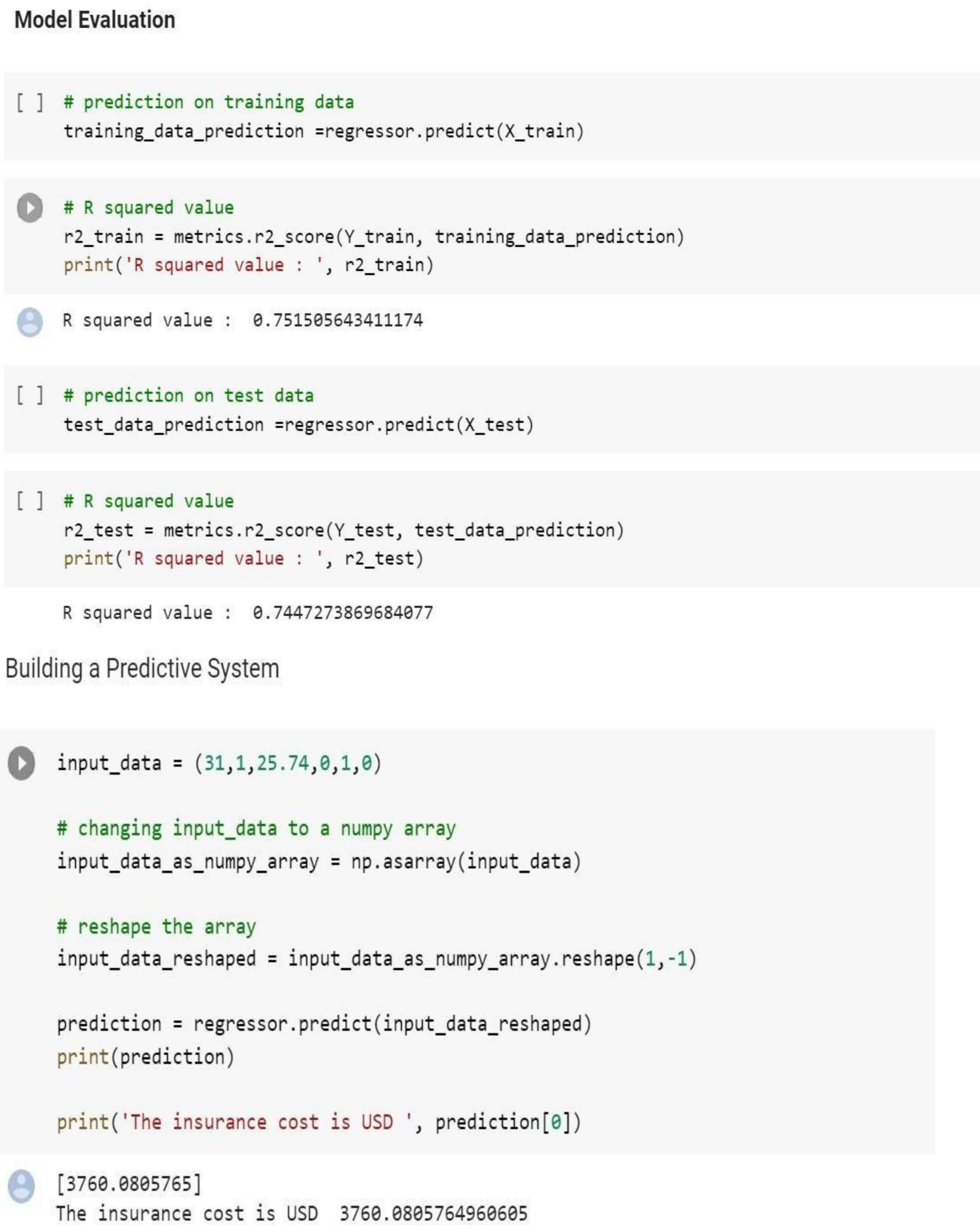


1.Linear Regression

```
# loading the Linear Regression model
regressor = LinearRegression()
```

```
regressor.fit(X_train, Y_train)
```

```
LinearRegression()
```



Model Evaluation

```
# prediction on training data
training_data_prediction =regressor.predict(X_train)
```

```
# R squared value
r2_train = metrics.r2_score(Y_train, training_data_prediction)
print('R squared value : ', r2_train)
```

```
R squared value :  0.751505643411174
```

```
# prediction on test data
test_data_prediction =regressor.predict(X_test)
```

```
# R squared value
r2_test = metrics.r2_score(Y_test, test_data_prediction)
print('R squared value : ', r2_test)
```

```
R squared value :  0.7447273869684077
```

Building a Predictive System

```
input_data = (31,1,25.74,0,1,0)

# changing input_data to a numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the array
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = regressor.predict(input_data_reshaped)
print(prediction)

print('The insurance cost is USD ', prediction[0])
```

```
[3760.0805765]
The insurance cost is USD  3760.0805764960605
```

DATA SET USED:

Columns,

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

TOOLS USED:

The python programming which was done in google collaboratory.

This data set is taken from Kaggle

MODEL’S USED:

- So linear regression model is most of a statistical model rather than a machinelearning model but you know it is the base for other models.

- Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task.

- Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

- Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

OUTCOME:

- Finally, predictive system that given all these parameters it can find the medical insurance cost so as we have achieved what we wanted to do in thisparticular project.

- It gives us the insurance cost is 3760 which is very close to the real value. This particular data point the medical insurance cost is 3756 dollars and the value predicted by model is 3760 which is very very close.

- So it tells us the model is performing kind of very good.

SUMMARY:

- Built a machine learning System that can predict what is the medical insurance cost of a person.
- These independent variables are assumed to have statistical significance in determining insurance premium cost of the customer.
- The first step in machine learning is to collect the data. Here we need insurance cost data. Once we have this insurance cost data we need to do some data analysis so we need to analyse this data to understand whether it can give us some meaning.
- So, the next part is data pre-processing so once we have the data we cannot feed it directly to our machine learning algorithm so we need to do some processing on it this is where data pre-processing comes into play.
- Then the next step will be to split our data into training data and testing data.
- So once we split the data into training data and testing data, we feed this training data to our machine learning model linear regression model.
- Once we train this linear regression model we will evaluate this model to check how well it is performing. So, once we feed new data, this model can predict what the insurance cost will be.