Experiment – 3 Simple Linear Regression

Pinni Venkata Abhiram

20BAI1132

Introduction

Building a Simple Linear Regression model to predict the prices of Medical Insurance when we enter all the necessary details.

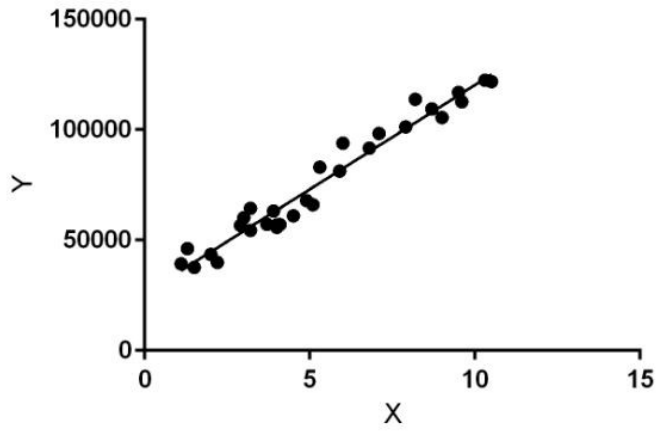
The fields in the data are - Age, Sex , BMI , Children, Smoker, Region, Charges

The model takes in the values of all the fields above except region because it has no importance in this matter and charges because it's our target and makes a regression model based on it.

Methodology

The methodology used in here is the concept of Linear regression

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 being used.



Dataset

The dataset used is a .csv file which contains the following rows

Age, Sex , BMI , Children, Smoker, Region, Charges

Age , BMI , Children , Charges are the rows which have a integer value. The smoker row is having yes / no as it's values . The Region has Northeast , Northwest , Southeast and Southwest.

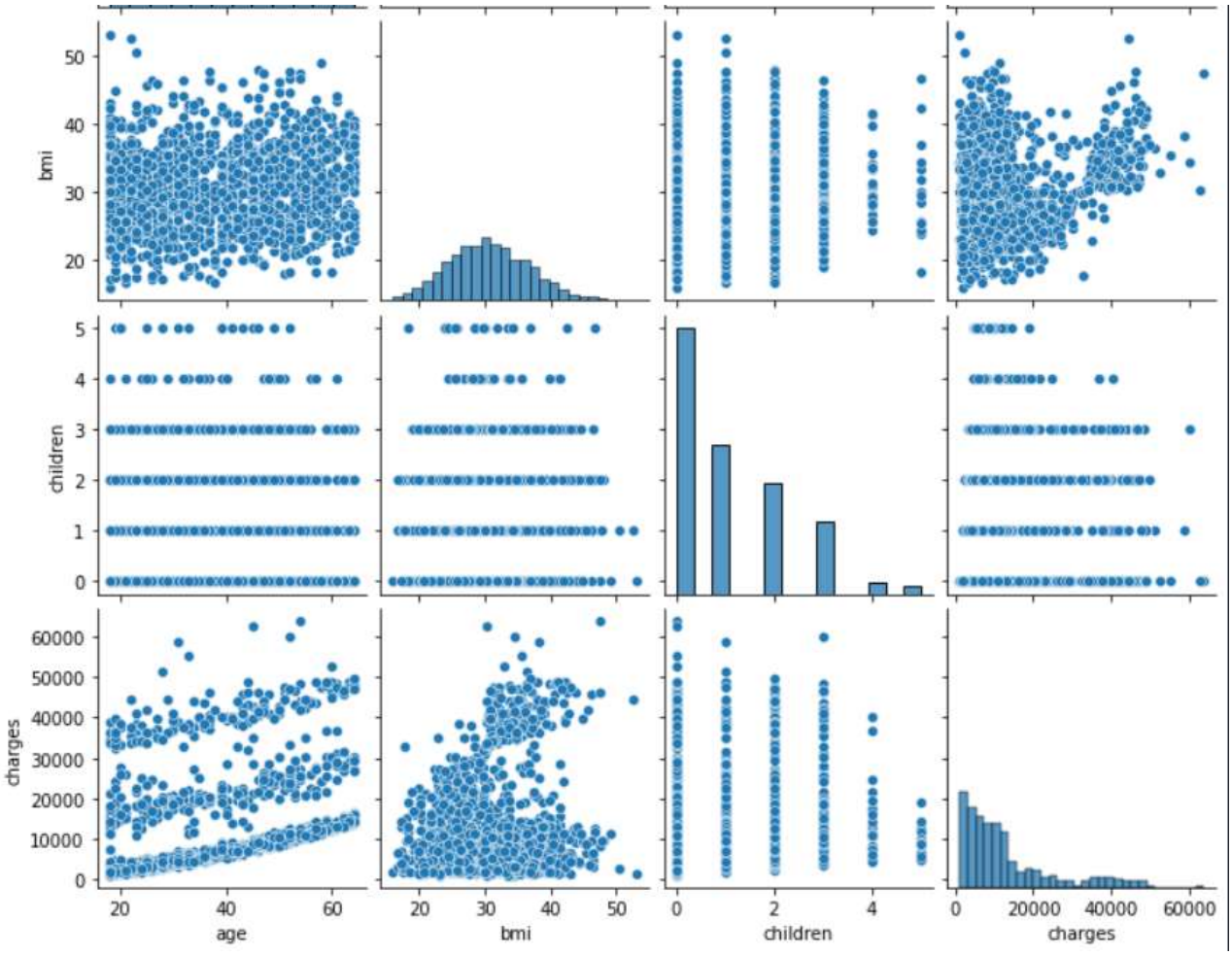
We use Age, Sex, BMI , Children, Smoker for the development of the model and we predict Charges i.e. the cost of insurance.

	A	B	C	D	E	F	G	H
1	Index	age	sex	bmi	children	smoker	region	charges
2	1	19	female	27.9	0	yes	southwest	16884.92
3	2	18	male	33.77	1	no	southeast	1725.552
4	3	28	male	33	3	no	southeast	4449.462
5	4	33	male	22.705	0	no	northwest	21984.47
6	5	32	male	28.88	0	no	northwest	3866.855
7	6	31	female	25.74	0	no	southeast	3756.622

Experiments and Results

The experiment required numpy , pandas , seaborn , sklearn libraries for analysis and model building.

The data is imported using numpy then we make plots based on the data using seaborn for our better understanding



We use correlation value we get from pandas and we use the sklearn library to divide the data into training set and test set.

We import Linear Regression function from sklearn and get the linear regression model with the train set.

Score of the Linear Regression Model is

```
score = linear_regression.score(x_test, y_test)
print("Model Score with test dataset is : ",score.round(3))
```

✓ 0.4s

```
Model Score with test dataset is :  0.731
```

We test the model by predicting with certain set of user inputs.

```
row10 = x_train[8:9]
row10
```

✓ 0.9s

	age	bmi	children	sex_female	sex_male	smoker_no	smoker_yes
737	26	23.7	2	0	1	1	0

```
linear_regression.predict(row10)
```

✓ 0.9s

```
array([3095.80350136])
```

We can see the predicted value is 3095.

```
ans_row = y_train[8:9]
ans_row
```

✓ 0.1s

```
737    3484.331
Name: charges, dtype: float64
```

The actual value is 3484 , it’s pretty close , so we can say that this model is successfully implemented.

Conclusion

The linear regression model of Machine Learning is successfully implemented and can predict the Medical Insurance costs using Python Libraries and Jupyter Notebook interface.

The Medical Insurance costs increase drastically if the person is a smoker. There is a slight increase in costs of female customers compared to male customers for Insurance.

References

https://www.kaggle.com/mirichoi0218/insurance

https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html

https://www.python-graph-gallery.com/92-control-color-in-seaborn-heatmaps