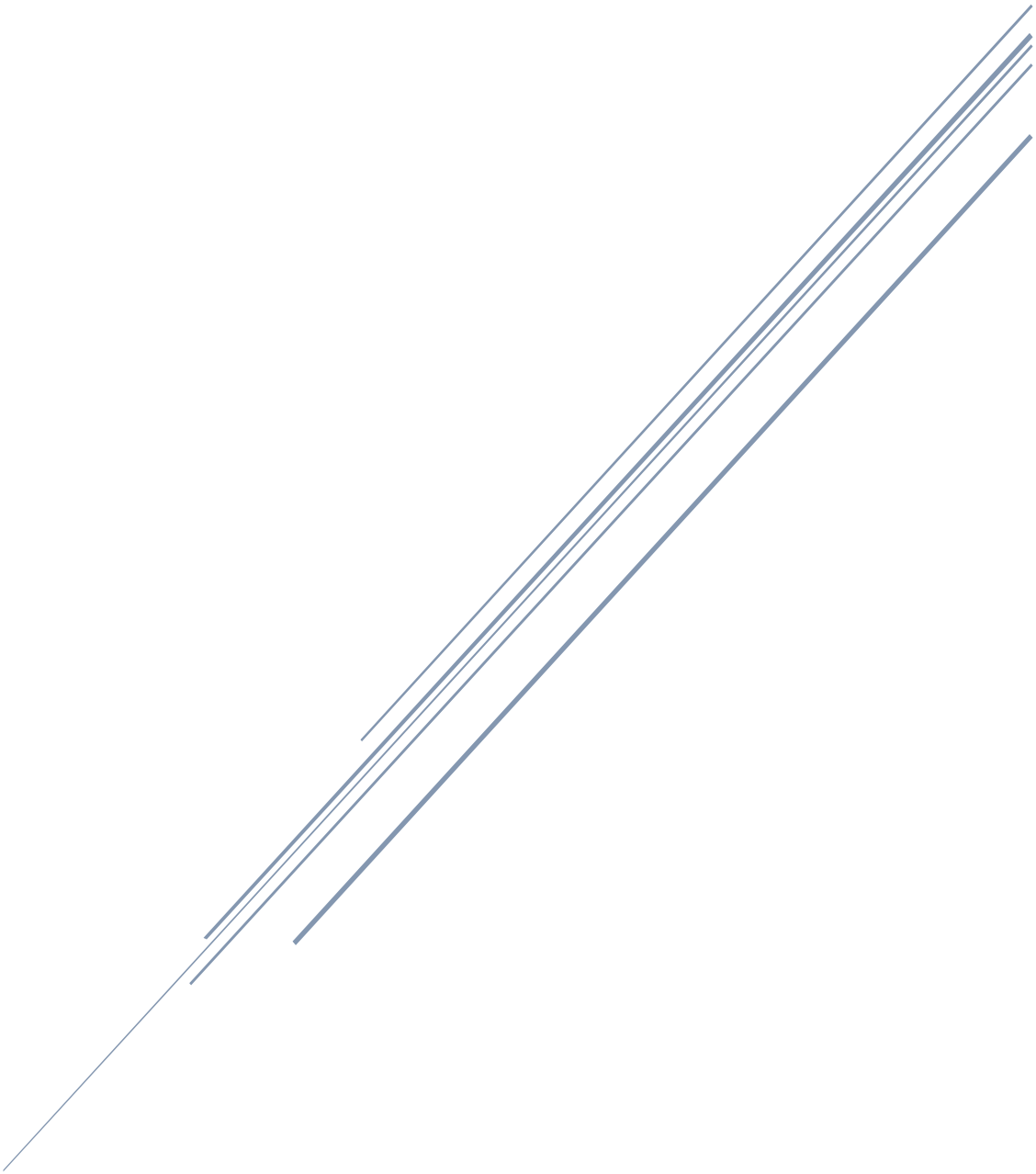
CSE1015 – MACHINE LEARNING ESSENTIALS

Lab – 5



Pinni Venkata Abhiram
20BAI1132

Experiment – 5 Logistic Regression

Pinni Venkata Abhiram

20BAI1132

Introduction

Building a Logistic Regression model to predict if a certain individual is suffering with breast cancer or not.

The fields in the data are - mean_radius, mean_texture, mean_perimeter, mean_area, mean_smoothness, diagnosis

The model is built using sklearn logistic regression and various plots for correlation provided by seaborn module and statsmodels for statistical summary.

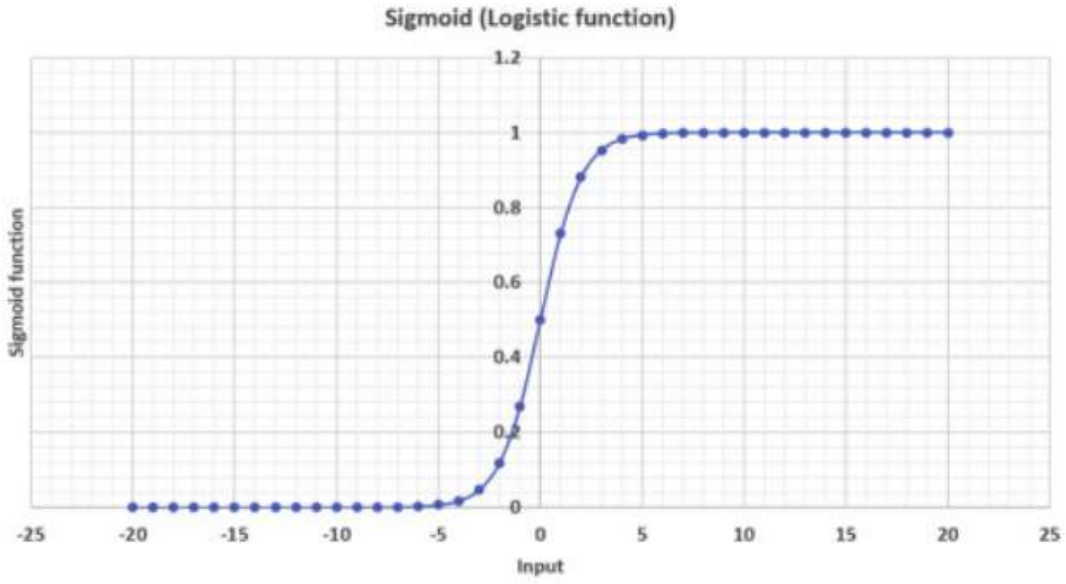
Methodology

The methodology used in here is the concept of Logistic regression

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category.

The best way to think about logistic regression is that it is a linear regression but for classification problems. Logistic regression essentially uses a logistic function defined below to model a binary output variable. The primary difference between linear regression and logistic regression is that logistic regression's range is bounded between 0 and 1. In addition, as opposed to linear regression, logistic regression does not require a linear relationship between inputs and output variables. This is due to applying a nonlinear log transformation to the odds ratio.

$$Logistic\ function = \frac{1}{1+e^{-x}}$$



Sigmoid Graph

Dataset

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

mean_radius, mean_texture, mean_perimeter, mean_area, mean_smoothness, diagnosis

All the rows are having a integer value.

We use mean_radius, mean_texture, mean_perimeter, mean_area, mean_smoothness for the development of the model and we predict the diagnosis i.e. if the patient has cancer or not.

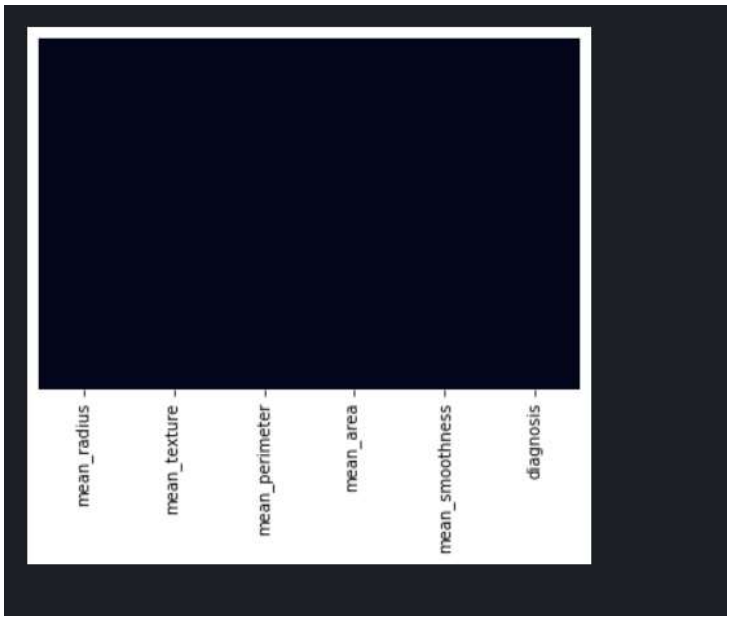
The dataset has 569 rows and 6 columns of data and we split 30% of the data for testing and 70% of data for training purpose.

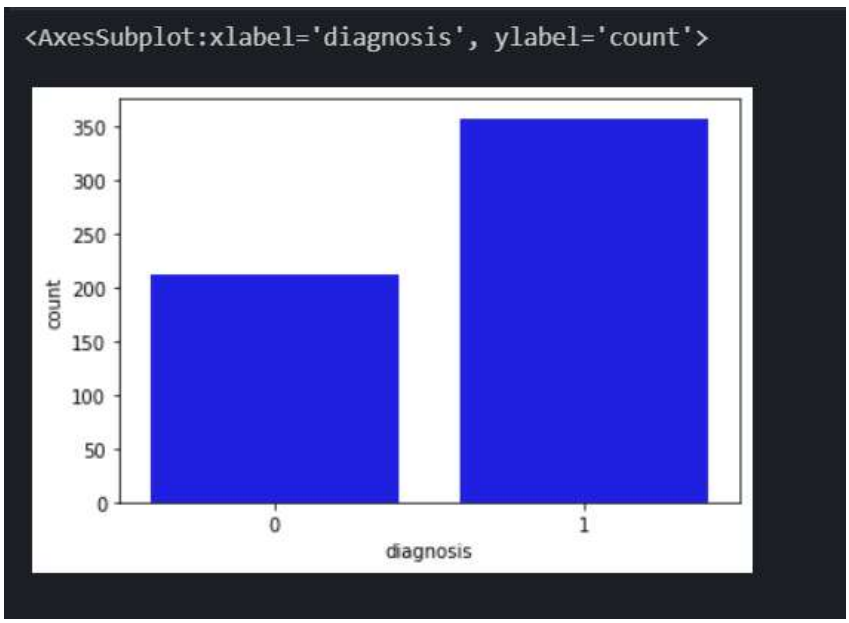
	A	B	C	D	E	F	G	H
1	Index	age	sex	bmi	children	smoker	region	charges
2	1	19	female	27.9	0	yes	southwest	16884.92
3	2	18	male	33.77	1	no	southeast	1725.552
4	3	28	male	33	3	no	southeast	4449.462
5	4	33	male	22.705	0	no	northwest	21984.47
6	5	32	male	28.88	0	no	northwest	3866.855
7	6	31	female	25.74	0	no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

Experiments and Results

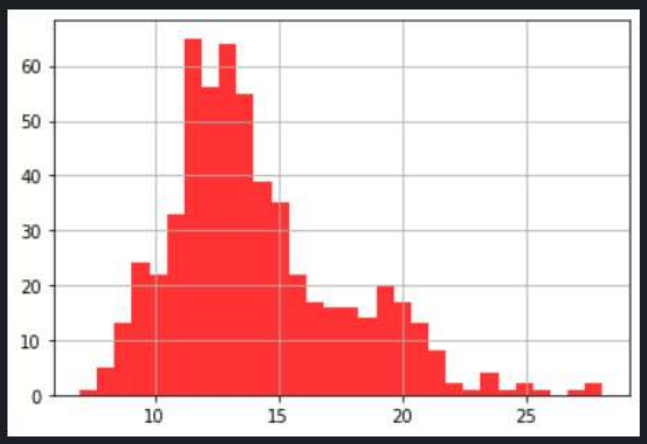
The experiment required numpy , pandas , seaborn , sklearn , statsmodels , matplotlib libraries for analysis and model building.

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





```
<AxesSubplot:>
```



And many more plots.

We check if the data has any NaN values i.e. missing values and we try to remove them , this is the data cleaning process.

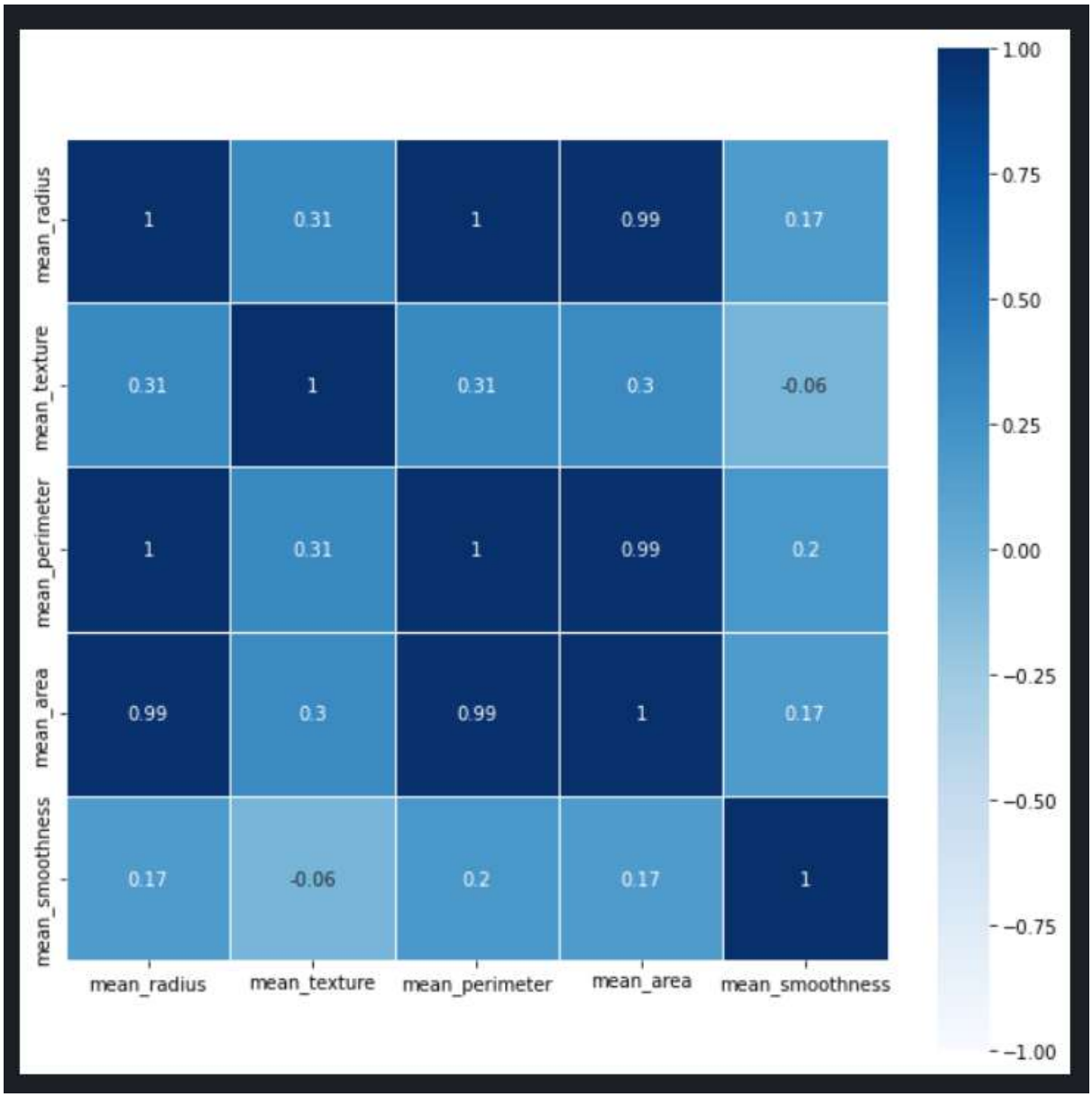
We use the sklearn library to divide the data into training set and test set.

```
y = df.diagnosis.values
x = df.drop("diagnosis",axis=1)
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.3)
```

```
scaler = MinMaxScaler()
cols = x_train.columns
temp = scaler.fit_transform(x_train)
x_train = pd.DataFrame(temp,columns=cols)

cols = x_test.columns
temp = scaler.fit_transform(x_test)
x_test = pd.DataFrame(temp,columns=cols)
```

We use correlation value we get from sklearn library and plot a correlation heatmap using seaborn library.



Correlation heatmap

Now we check the Statistical summary to check the P – values and if the P – value of a column is greater than 0.7 we need to remove it. We observe that P – Values are less so we don't remove any column of our data.

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.9052	0.093	20.574	0.000	1.723	2.087
mean_radius	1.5044	1.725	0.872	0.384	-1.886	4.895
mean_texture	-0.7221	0.112	-6.442	0.000	-0.943	-0.502
mean_perimeter	-5.2698	1.760	-2.995	0.003	-8.730	-1.810
mean_area	2.2229	0.639	3.476	0.001	0.966	3.480
mean_smoothness	-0.6117	0.128	-4.791	0.000	-0.863	-0.361

Now we check the Varience Inflation Factor or VIF values and see if they are large. If there are large values then we remove one column and look at the VIF values again and we remove the columns till the VIF values are in double or single digit values.

VIF values at the start

	variables	VIF
0	const	37.988653
1	mean_radius	386.899609
2	mean_texture	1.154779
3	mean_perimeter	388.252924
4	mean_area	39.647102
5	mean_smoothness	1.544868

VIF values after removing mean_perimeter because it's huge value

	variables	VIF
0	const	31.087124
1	mean_radius	38.947812
2	mean_texture	1.119790
3	mean_area	38.927010
4	mean_smoothness	1.046060

Now that we have removed mean_perimeter the values came down a lot and now we can build our model using sklearn's logistic regression function and then we can view the result using classification report.

	precision	recall	f1-score	support
0	0.92	0.76	0.83	59
1	0.89	0.96	0.92	112
accuracy			0.89	171
macro avg	0.90	0.86	0.88	171
weighted avg	0.90	0.89	0.89	171

Classification Report

To get the accuracy and number of correct outcomes when the model is tested we can use the confusion matrix where we can directly get the number of true positives , true negetives , false positives and false negetives.

```
True Negative: 45
False Positive: 14
False Negative: 4
True Positive: 108
Correct Predictions 89.5 %
```

Finally we can say that the model is 89.5% accurate when we use the test dataset to predict when the model is built with 70% of the data.

Conclusion

The Logistic regression model of Machine Learning is successfully implemented and can predict the if the patient has breast cancer using Python Libraries and Jupyter Notebook interface.

References

https://www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset

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

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

https://seaborn.pydata.org/generated/seaborn.heatmap.html

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

https://www.statisticshowto.com/variance-inflation-factor/