Diabetes Prediction for Pima Indian Diabetes Dataset

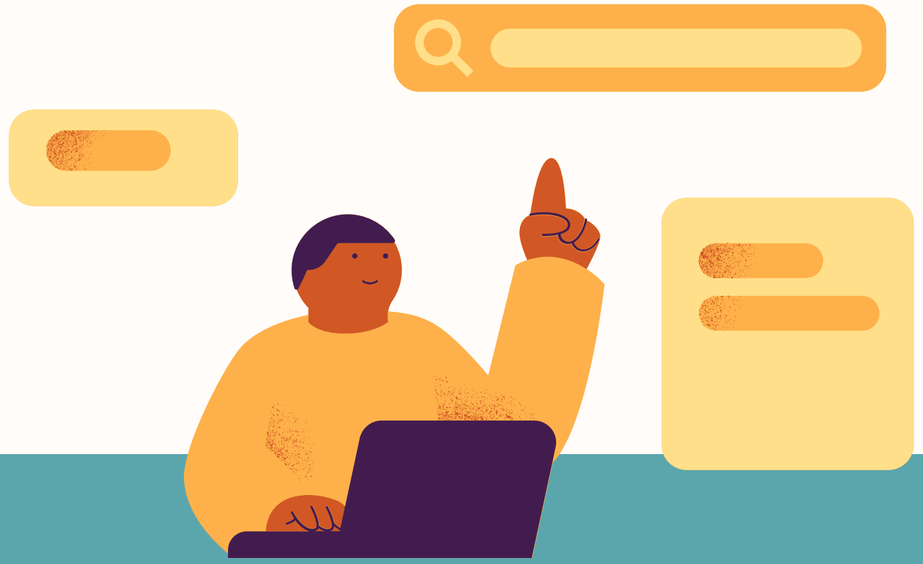
Presented by,

Priyanka Pai

Analytics for Healthcare and Pharma

Content

- Background
- Objective
- Data Description
- Data Preprocessing
- EDA
- Model Comparison



What is Diabetes?

Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high.

The most common types of diabetes are:

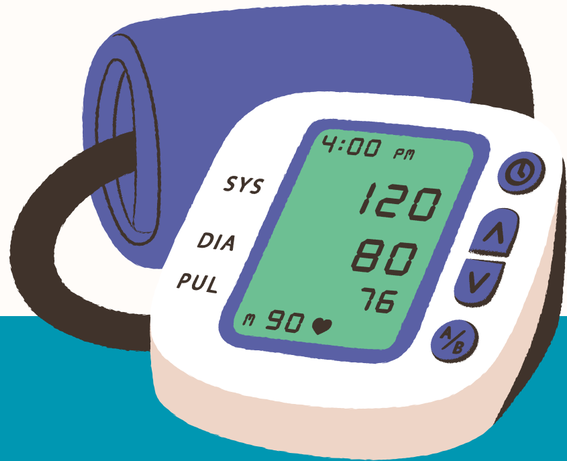
- **Type 1 diabetes** is a chronic condition in which the pancreas produces little or no insulin. It is usually diagnosed in children and young adults, although it can appear at any age.
- **Type 2 diabetes** is a chronic condition that affects how the body processes blood sugar (glucose). This type of diabetes occurs most often in middle-aged and older people.
- **Gestational diabetes** is a form of high blood sugar affecting pregnant women.

Objective

The objective of the dataset is **to diagnostically predict whether or not a patient has diabetes**, based on certain diagnostic measurements included in the dataset.

Using the following modeling techniques:

- Logistic Regression
- Random Forest Classifier
- AdaBoost Classifier (Adaptive boosting)



About the data

The data was collected and made available by “The National Institute of Diabetes and Digestive and Kidney Diseases”. All patients in this belong to the Pima Indian heritage (subgroup of Native Americans) and are females of ages 21 and above.

Variables	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration over 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)2)
DiabetesPedigreeFunction	Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
Age	Age (years)
Outcome	Class variable (0 if non-diabetic, 1 if diabetic)

Data Preprocessing

Here, we can see that the columns Glucose, Blood Pressue, Skin Thickness, Insulin and BMI have value 0, which is impossible.

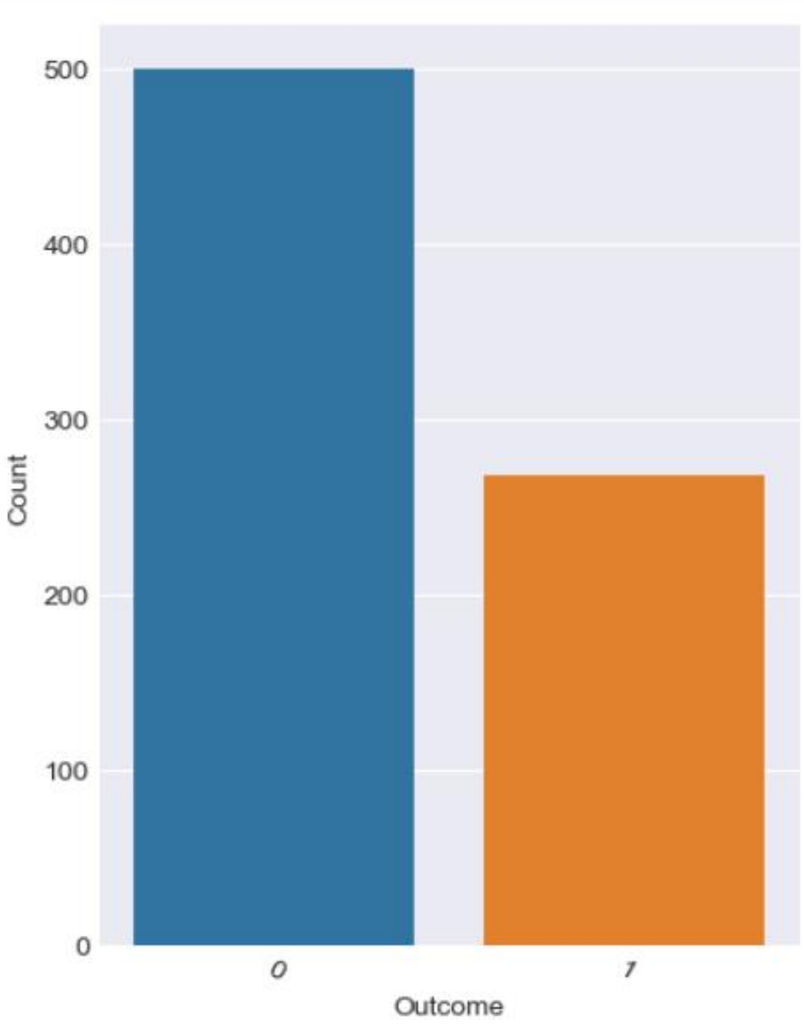
	Pregnancies	Glucose	BP	SkinThickness	Insulin	BMI	DPF	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

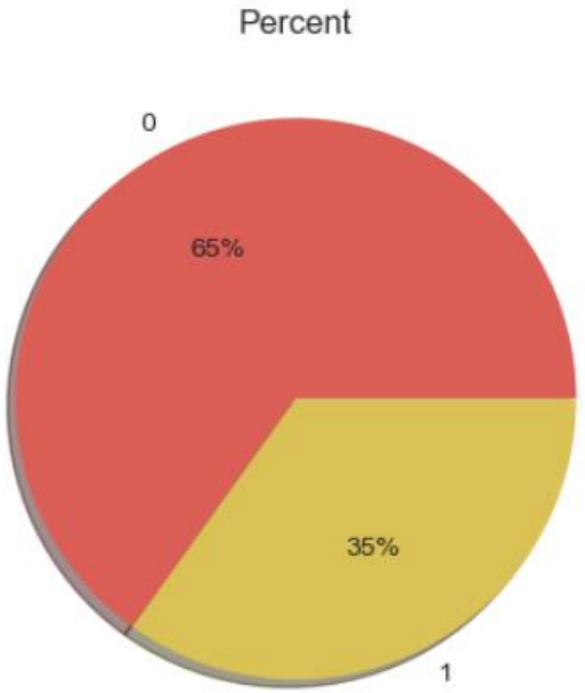
Replaced these zero values with NaN, and later replaced it with either mean or median of the variable.

Exploratory Data Analysis

Outcome

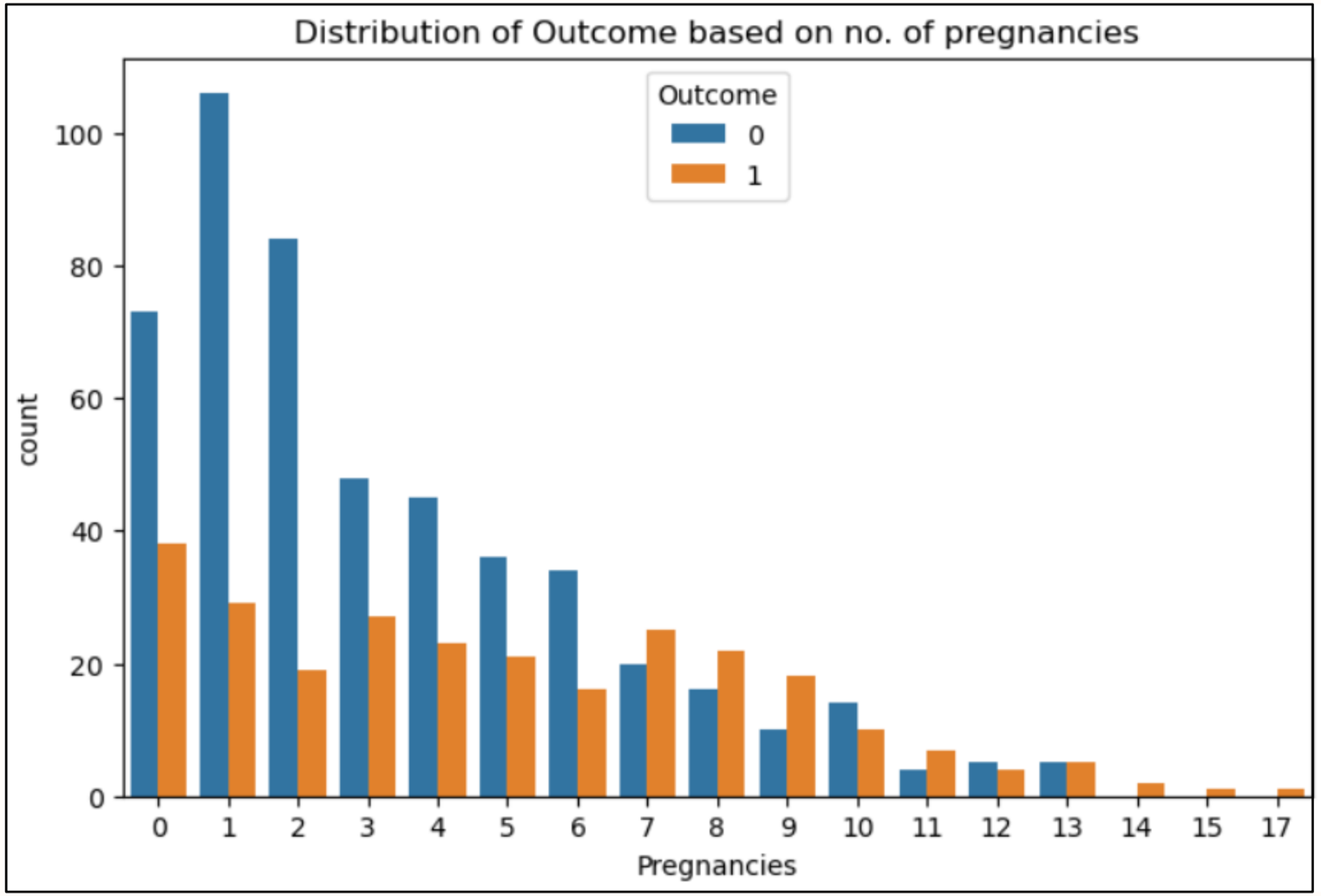
Most of the patients in the dataset are not diabetic, i.e., 65%





Pregnancies

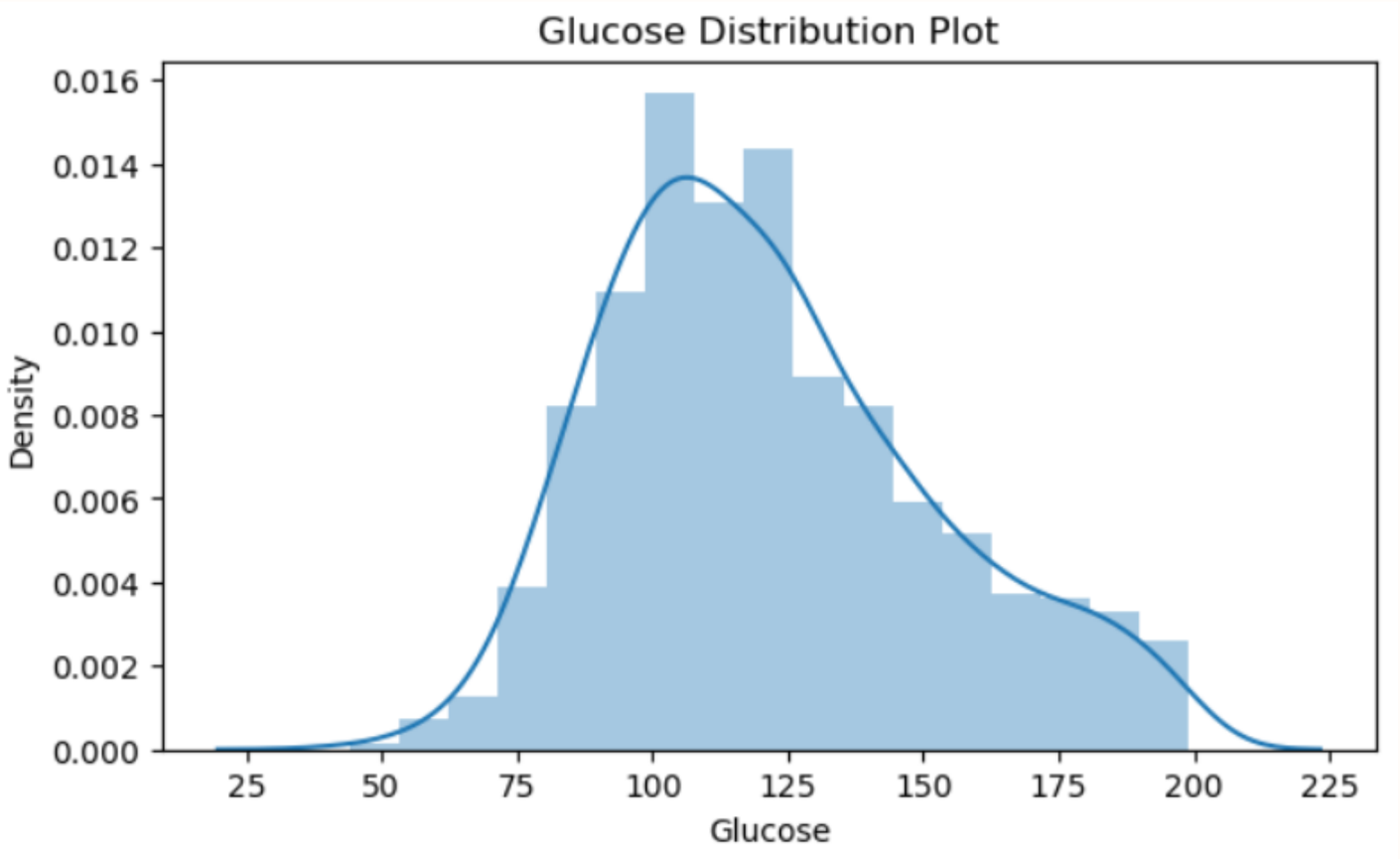
After pregnancy, people are more likely to have diabetes.



Glucose

There were 5 data points with 0 value which was replaced with NaN.

From the distribution, we can observe that there is not much skewness present in the data. So, replaced the missing values with the mean of the data.

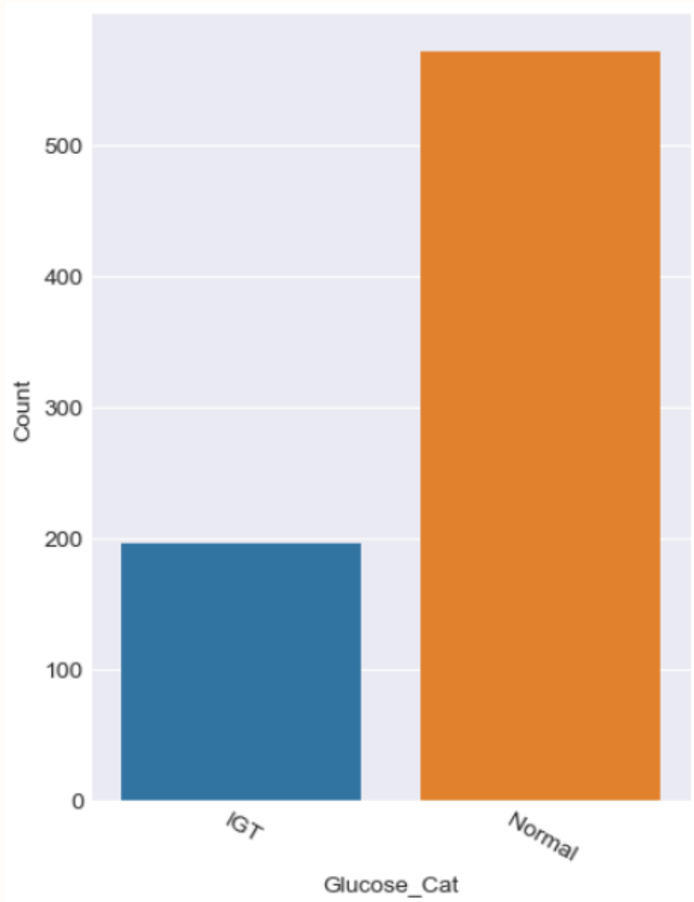


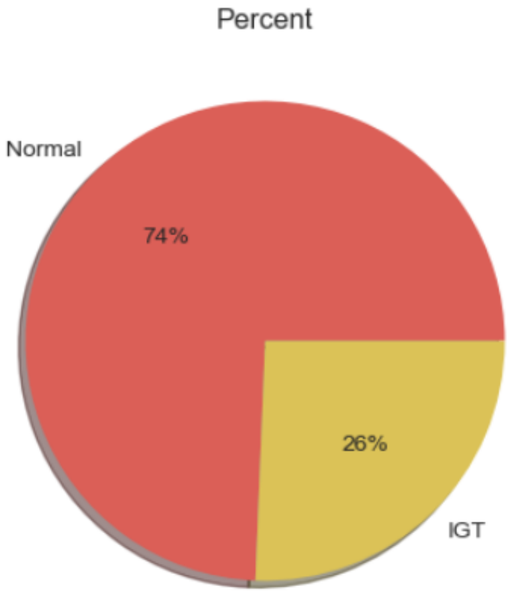
Glucose

Categorized the Glucose levels into:

- **Normal** - below 140 mg/dL
- **IGT** (Impaired Glucose tolerance) - between 140 and 200 mg/dL
- **DM** (Diabetes Mellitus) - above 200 mg/dL

None of the women have a glucose level above 200mg/dL and 74% of them have normal blood glucose level.

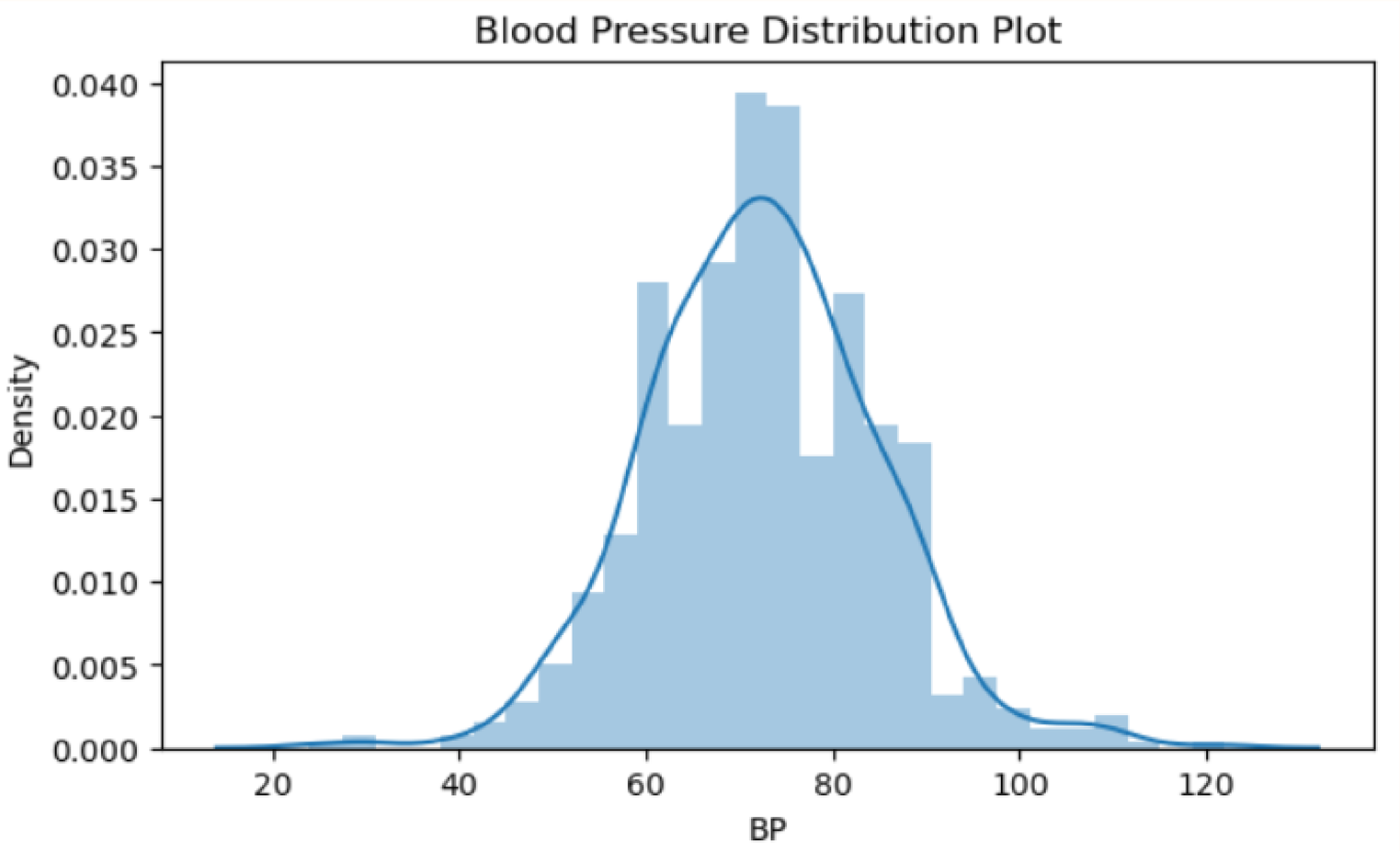




Blood Pressure

There were 35 data points with 0 value which was replaced with NaN.

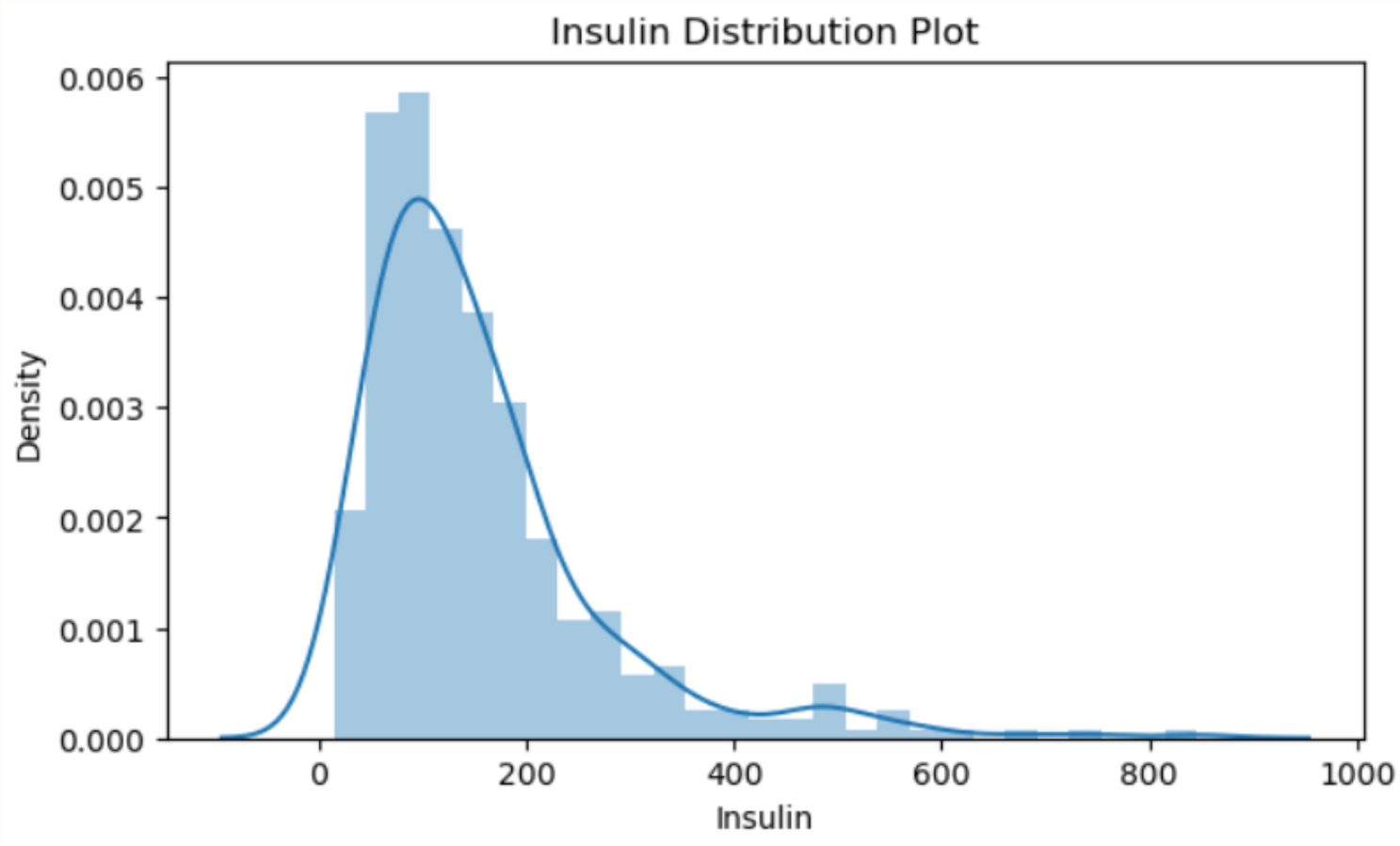
From the above distribution, we can observe that there is not much skewness present in the data. So, replaced the missing values with the mean of the data.



Insulin

There were 374 data points (48% of the data) with 0 value which was replaced with NaN.

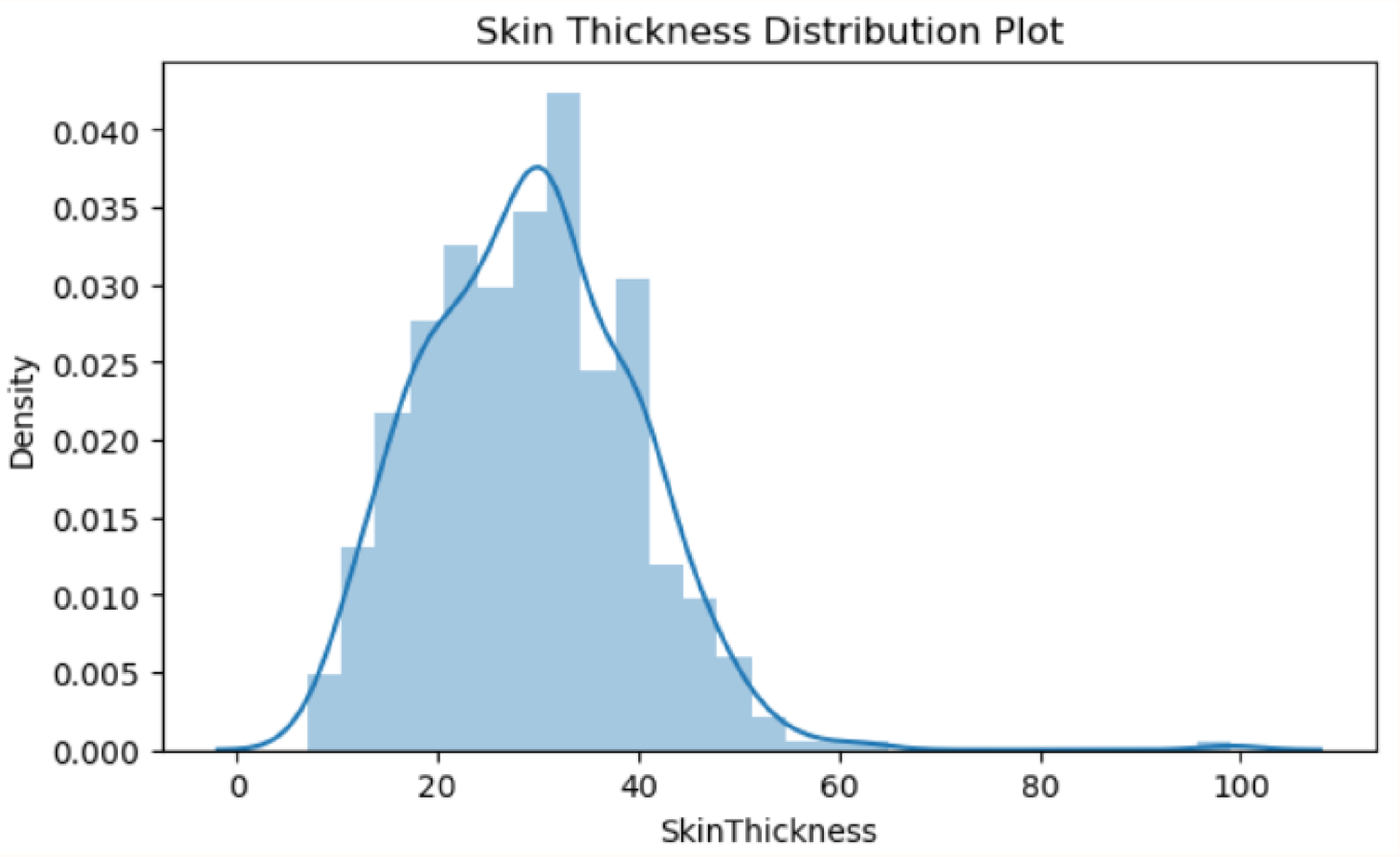
From the distribution plot we can observe that the Insulin data right skewed. So, replaced the missing values with median of the data.



Skin Thickness

There were 227 data points with 0 value which was replaced with NaN.

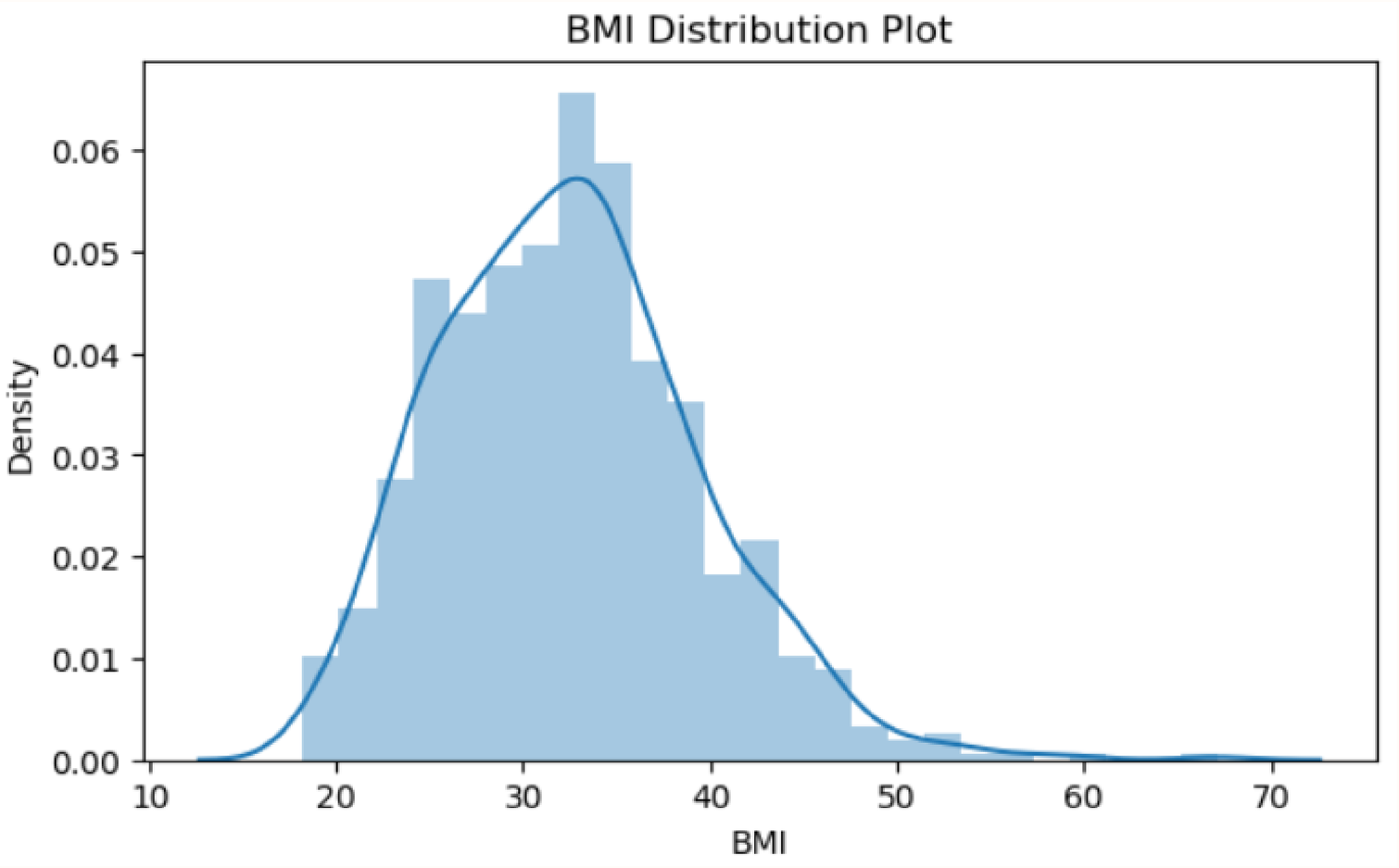
From the distribution plot we can observe that the SkinThickness data right skewed. So, replaced the missing values with median of the data.

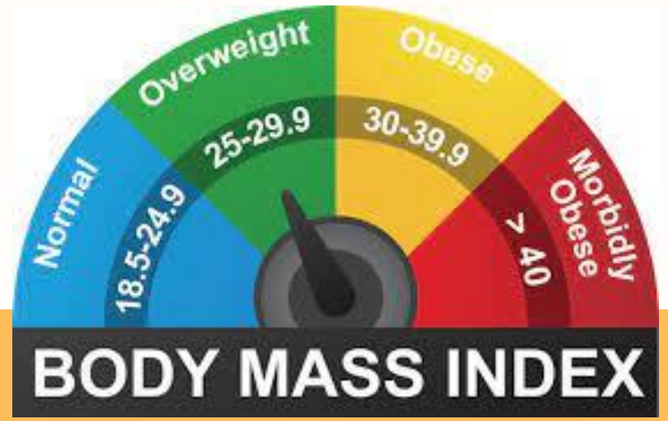


BMI

There were 11 data points with 0 value which was replaced with NaN.

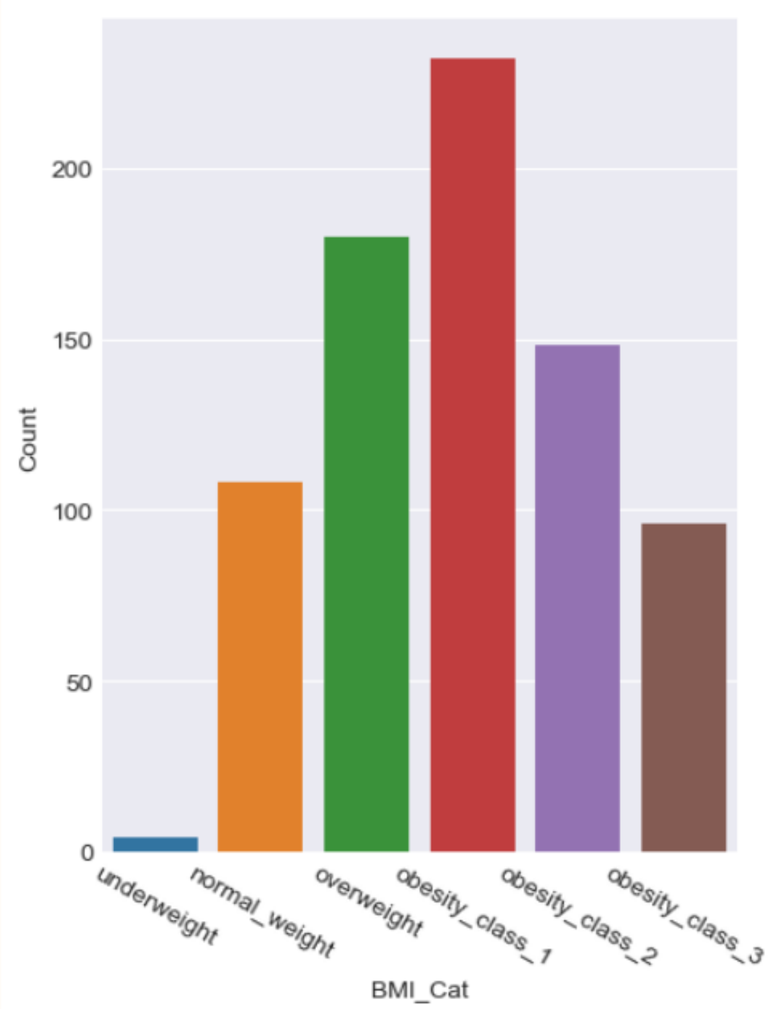
From the distribution plot we can observe that the Insulin data right skewed. So, replaced the missing values with median of the data.

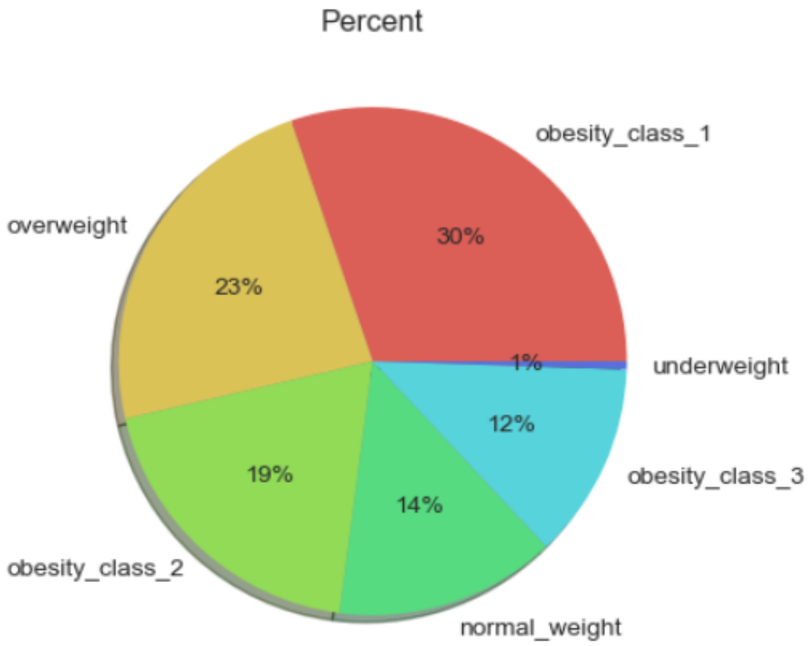




BMI

Only 1% of the women were underweight. While the rest have a BMI of more than 25 kg/m^2.



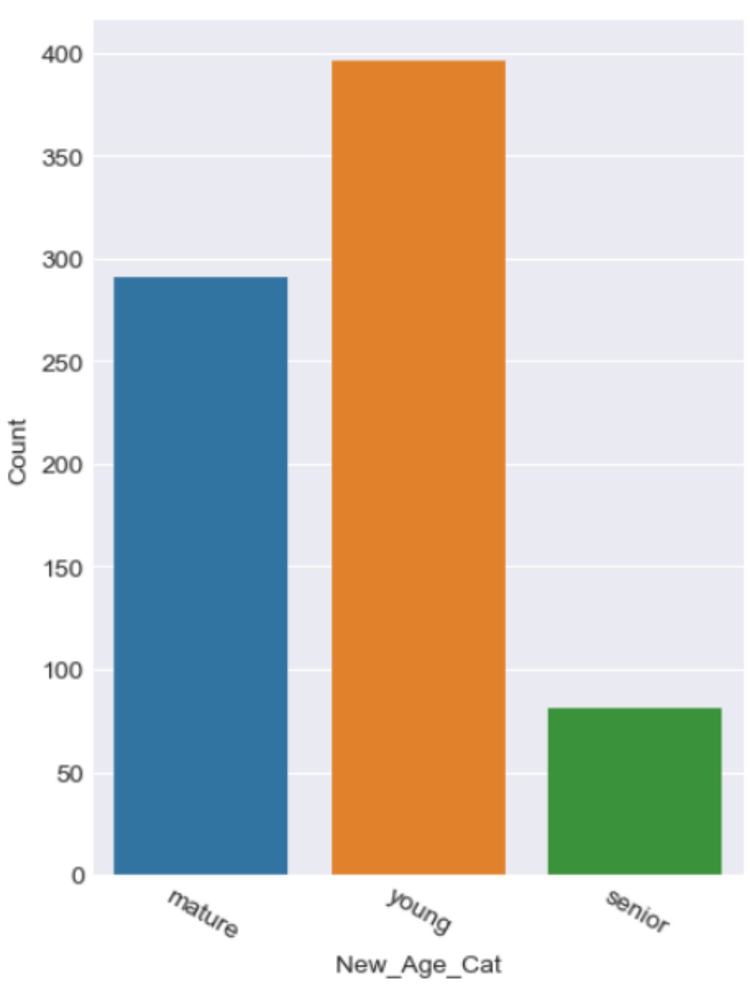


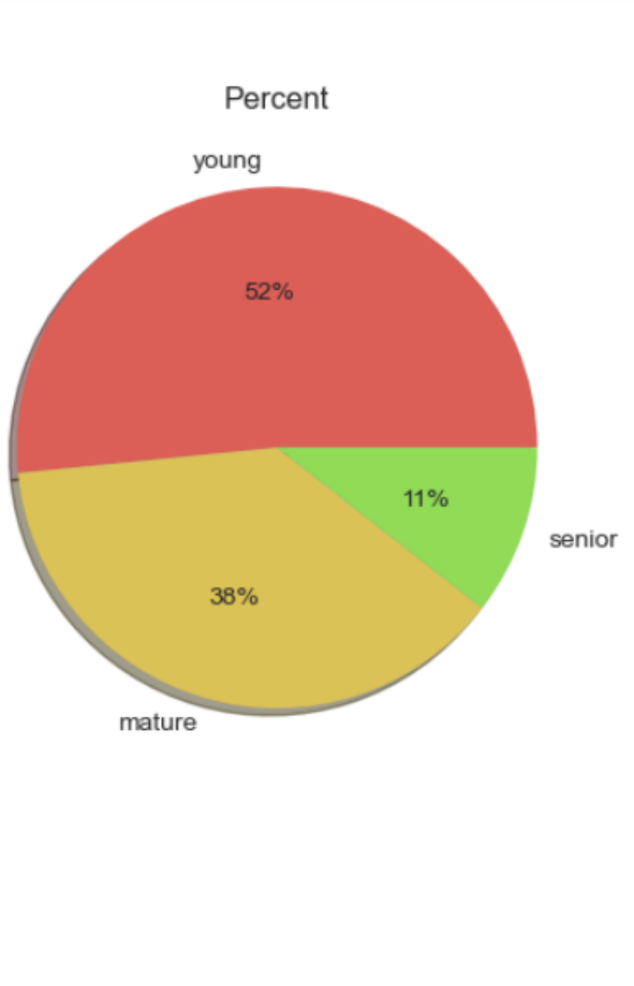
Age

Categorized the Age column into:

- **young** - below 30 years
- **mature** - between 30 and 50 years
- **senior** - above 50 years

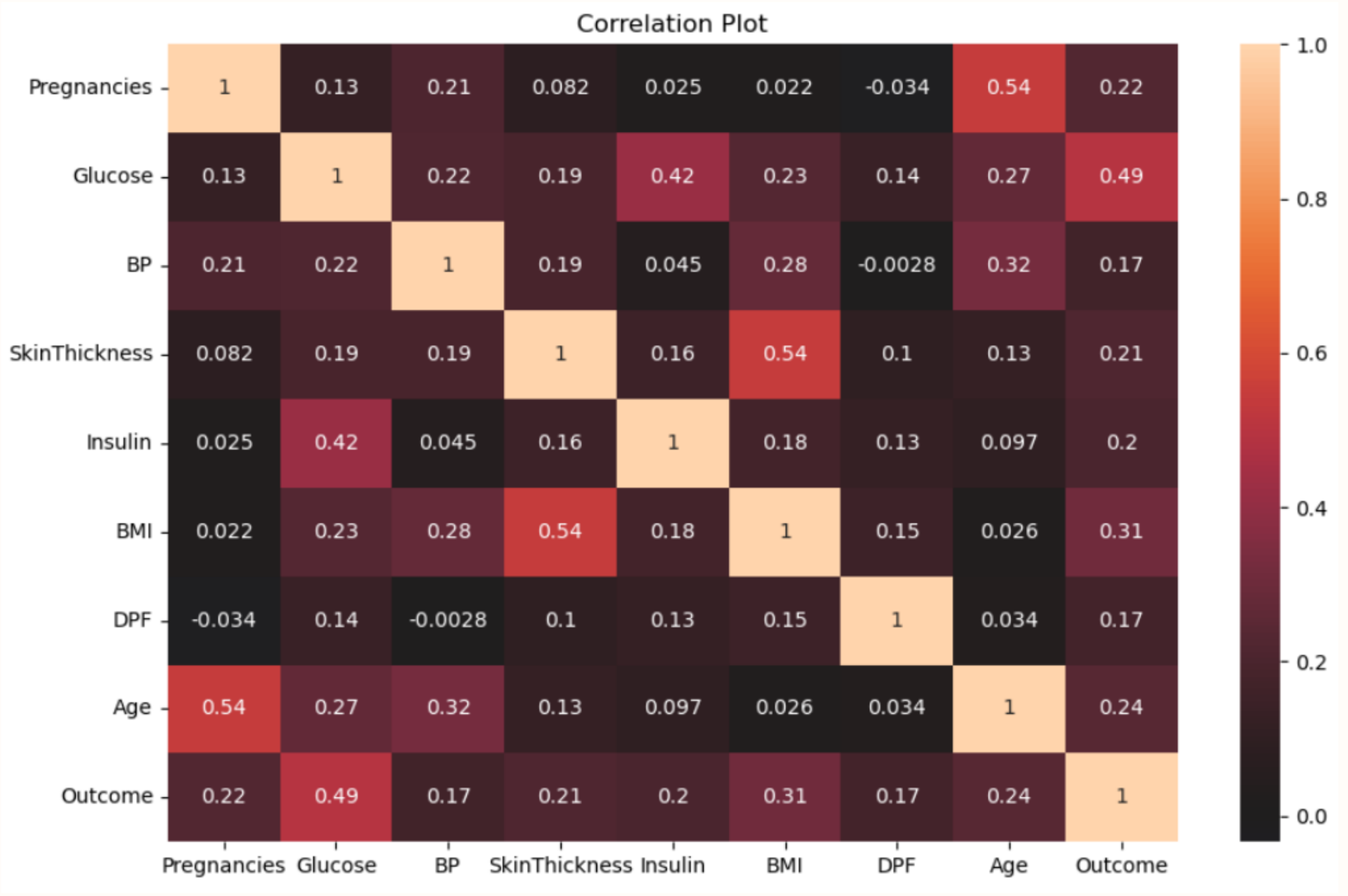
Here, more than **50%** of the women are **young** (aged 21 to 30 years).





Correlation Plot

There is no multicollinearity problem in this dataset.

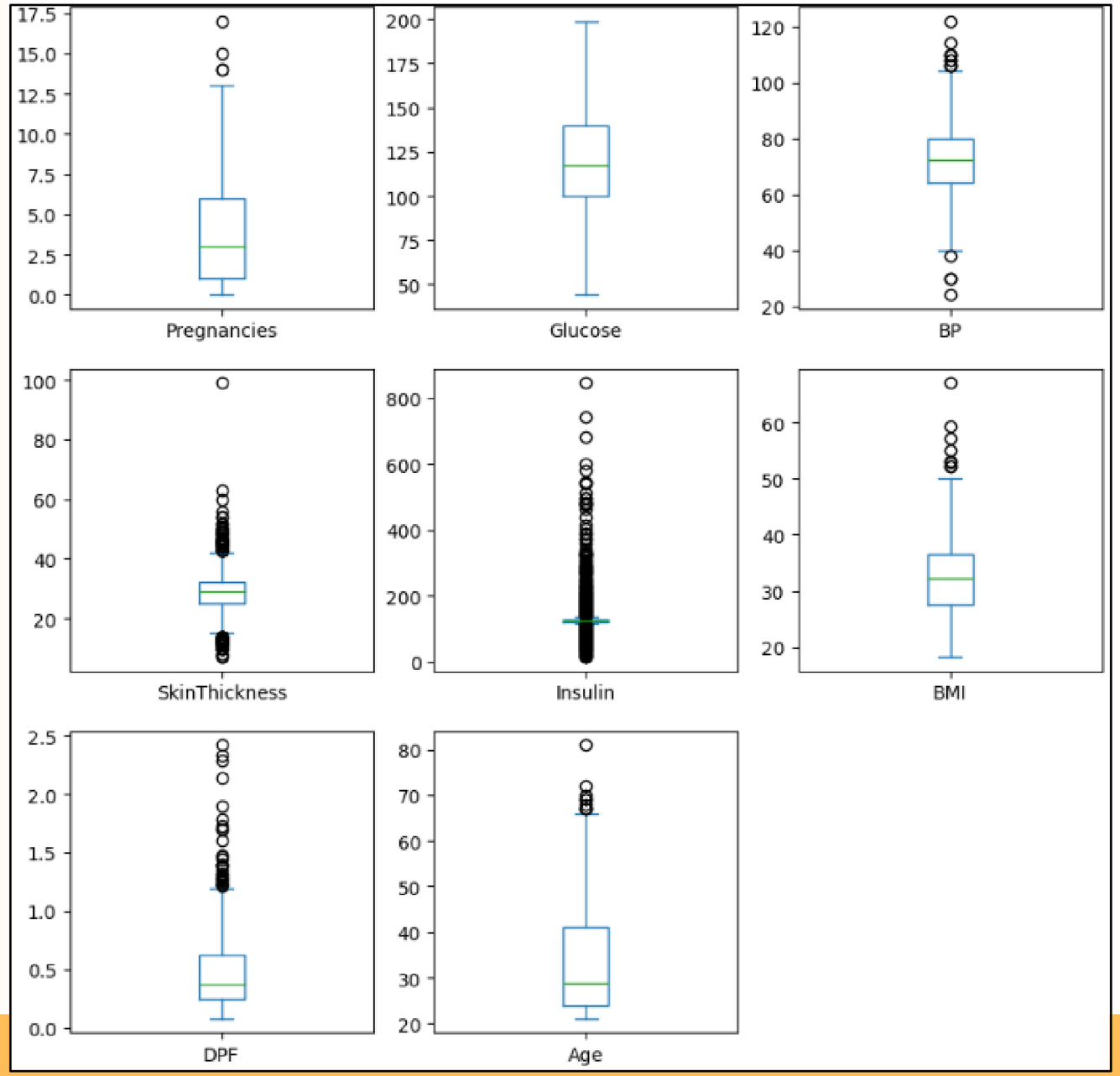


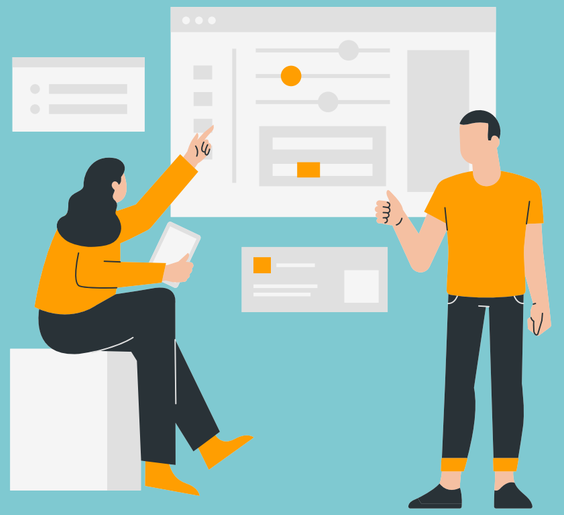
Box Plot

All the variables in the dataset have outliers.

Hence, we standardize the columns.

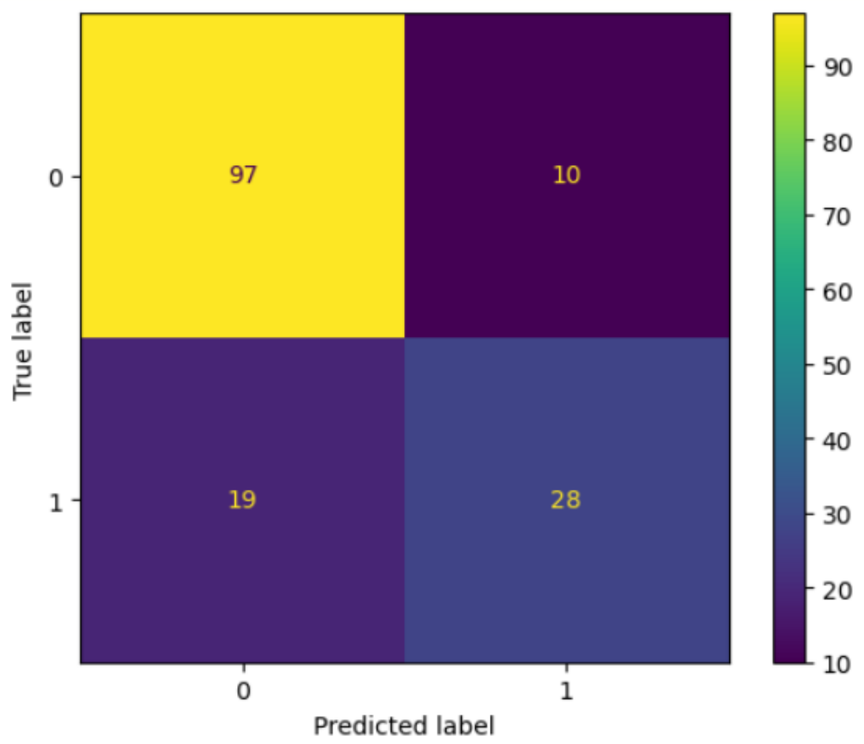
Here, the values of all the columns are scaled in such a way that they all have a mean equal to 0 and standard deviation equal to 1. This scaling technique works well with outliers.





Model Building

Logistic Regression

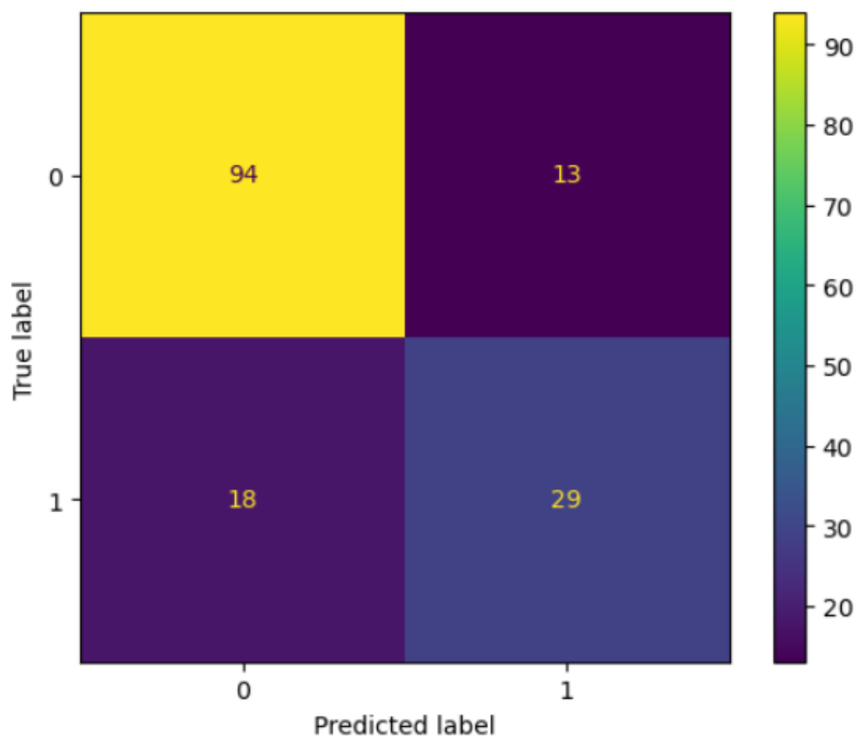



```
log_reg =LogisticRegression()
log_reg.fit(X_train, y_train)

# test data
pred_log = log_reg.predict(X_test)
accuracy = accuracy_score(y_test, pred_log)
print(f"Accuracy on Test Data: {accuracy*100}%")
```

```
Accuracy on Test Data: 81.16883116883116%
```

Random Forest Classifier

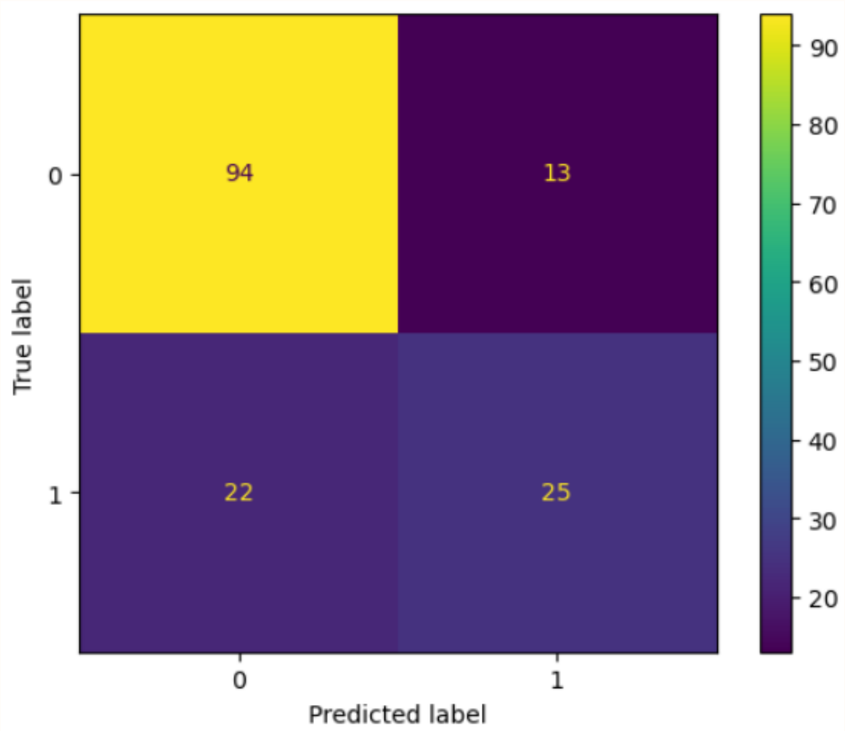



```
rf =RandomForestClassifier()
rf.fit(X_train, y_train)

# Accuray On Test Data
pred_rf = rf.predict(X_test)
accuracy = accuracy_score(y_test, pred_rf)
```

```
Accuracy on Test Data: 79.87012987012987%
```

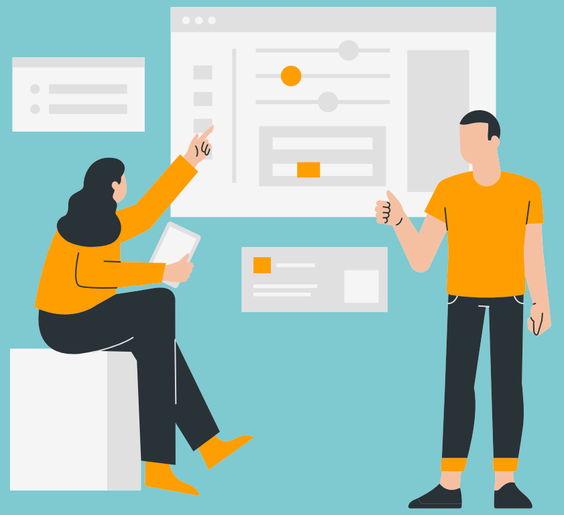
Adaptive Boosting Classifier




```python
abc =  AdaBoostClassifier(n_estimators=50, learning_rate=0.1, random_state=0)
abc.fit(X_train, y_train)

# Accuray On Test Data
pred_ada = abc.predict(X_test)
accuracy = accuracy_score(y_test, pred_ada)
print(f"Accuracy on Test Data: {accuracy*100}%")
```

```
Accuracy on Test Data: 77.27272727272727%
```



Model Comparison

Model	Accuracy
Logistic Regression	81.17
Random Forest Classifier	79.87
AdaBoost Classifier	77.27

The **logistic regression model** outperformed the other two models as the accuracy of the model is **81%** which indicates that it is a good fit.

Not only based ont the accuracy but also based on the count of correctly predicted classes vs the errors, logistic regression is considered to be a better classification model for this dataset.

THANK YOU!