

Application Flow

Logistic Regression is one of the most fundamental algorithms for classification in the Machine Learning world.

But before proceeding with the algorithm, let's first discuss the lifecycle of any machine learning model. This diagram explains the creation of a Machine Learning model from scratch and then taking the same model further with hyperparameter tuning to increase its accuracy, deciding the deployment strategies for that model and once deployed setting up the logging and monitoring frameworks to generate reports and dashboards based on the client requirements. A typical lifecycle diagram for a machine learning model looks like:



Introduction

In linear regression, the type of data we deal with is quantitative, whereas we use classification models to deal with qualitative data or categorical data. The algorithms used for solving a classification problem first predict the probability of each of the categories of the qualitative variables, as the basis for making the classification. And, as the probabilities are continuous numbers, classification using probabilities also behave like regression methods. Logistic regression is one such type of classification model which is used to classify the dependent variable into two or more classes or categories.

Why don't we use Linear regression for classification problems?

Let's suppose you took a survey and noted the response of each person as satisfied, neutral or Not satisfied. Let's map each category:

Satisfied – 2

Neutral – 1

Not Satisfied – 0

But this doesn't mean that the gap between Not satisfied and Neutral is same as Neutral and satisfied. There is no mathematical significance of these mapping. We can also map the categories like:

Satisfied – 0

Neutral – 1

Not Satisfied – 2

It's completely fine to choose the above mapping. If we apply linear regression to both the type of mappings, we will get different sets of predictions. Also, we can get prediction values like 1.2, 0.8, 2.3 etc. which makes no sense for categorical values. So, there is no normal method to convert qualitative data into quantitative data for use in linear regression. Although, for binary classification, i.e. when there only two categorical values, using the least square method can give decent results. Suppose we have two categories Black and White and we map them as follows:

Black – 0

White - 1

We can assign predicted values for both the categories such as $Y > 0.5$ goes to class white and vice versa. Although, there will be some predictions for which the value can be greater than 1 or less than 0 making them hard to classify in any class. Nevertheless, linear regression can work decently for binary classification but not that well for multi-class classification. Hence, we use classification methods for dealing with such problems.

Logistic Regression

Logistic regression is one such regression algorithm which can be used for performing classification problems. It calculates the probability that a given value belongs to a specific class. If the probability is more than 50%, it assigns the value in that particular class else if the probability is less than 50%, the value is assigned to the other class. Therefore, we can say that logistic regression acts as a binary classifier.

Working of a Logistic Model

For linear regression, the model is defined by: $y = \beta_0 + \beta_1 x$ - (i)

and for logistic regression, we calculate probability, i.e. y is the probability of a given variable x belonging to a certain class. Thus, it is obvious that the value of y should lie between 0 and 1.

But, when we use equation(i) to calculate probability, we would get values less than 0 as well as greater than 1. That doesn't make any sense. So, we need to use such an equation which always gives values between 0 and 1, as we desire while calculating the probability.

Sigmoid function

We use the sigmoid function as the underlying function in Logistic regression. Mathematically and graphically, it is shown as:



Why do we use the Sigmoid Function?

1) The sigmoid function's range is bounded between 0 and 1. Thus it's useful in calculating the probability for the Logistic function. 2) It's derivative is easy to calculate than other functions which is useful during gradient descent calculation. 3) It is a simple way of introducing non-linearity to the model.

Although there are other functions as well, which can be used, but sigmoid is the most common function used for logistic regression. We will talk about the rest of the functions in the neural network section.

The logistic function is given as:



Let's see some manipulation with the logistic function:



We can see that the logit function is linear in terms with x .

Prediction



Cost Function



The cost function for the whole training set is given as :



The values of parameters (θ) for which the cost function is minimum is calculated using the gradient descent (as discussed in the Linear Regression section) algorithm. The partial derivative for cost function is given as :



Multiple Logistic Function

We can generalise the simple logistic function for multiple features as: 

And the logit function can be written as:



The coefficients are calculated the same we did for simple logistic function, by passing the above equation in the cost function.

Just like we did in multilinear regression, we will check for correlation between different features for Multi logistic as well.

We will see how we implement all the above concept through a practical example.

Multinomial Logistics Regression(Number of Labels >2)

Many times, there are classification problems where the number of classes is greater than 2. We can extend Logistic regression for multi-class classification. The logic is simple; we train our logistic model for each class and calculate the probability($h\theta x$) that a specific feature belongs to that class. Once we have trained the model for all the classes, we predict a new value's class by choosing that class for which the probability($h\theta x$) is maximum. Although we have libraries that we can use to perform multinomial logistic regression, we rarely use logistic regression for classification problems where the number of classes is more than 2. There are many other classification models for such scenarios. We will see more of that in the coming lectures.

Learning Algorithm

The learning algorithm is how we search the set of possible hypotheses (hypothesis space \mathcal{H}) for the best parameterization (in this case the weight vector \mathbf{w}). This search is an optimization problem looking for the hypothesis that optimizes an error measure.

There is no sophisticated, closed-form solution like least-squares linear, so we will use gradient descent instead. Specifically we will use batch gradient descent which calculates the gradient from all data points in the data set.

Luckily, our "cross-entropy" error measure is convex so there is only one minimum. Thus the minimum we arrive at is the global minimum.

To learn we're going to minimize the following error measure using batch gradient descent.

$$e(h(\mathbf{x}_n), y_n) = \ln(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n})$$
$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N e(h(\mathbf{x}_n), y_n) = \frac{1}{N} \sum_{n=1}^N \ln(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n})$$

We'll need the derivative of the point loss function and possibly some abuse of notation.

$$\frac{d}{d\mathbf{w}} e(h(\mathbf{x}_n), y_n) = \frac{-y_n \mathbf{x}_n e^{-y_n \mathbf{w}^T \mathbf{x}_n}}{1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}} = -\frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}}$$

With the point loss derivative we can determine the gradient of the in-sample error:

$$\nabla E_{\text{in}}(\mathbf{w}) = \frac{d}{d\mathbf{w}} \left[\frac{1}{N} \sum_{n=1}^N e(h(\mathbf{x}_n), y_n) \right] \quad (1)$$

$$= \frac{1}{N} \sum_{n=1}^N \frac{d}{d\mathbf{w}} e(h(\mathbf{x}_n), y_n) \quad (2)$$

$$= \frac{1}{N} \sum_{n=1}^N \left(-\frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}} \right) \quad (3)$$

$$= -\frac{1}{N} \sum_{n=1}^N \frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}} \quad (4)$$

Our weight update rule per batch gradient descent becomes

$$\mathbf{w}_{i+1} = \mathbf{w}_i - \eta \nabla E_{\text{in}}(\mathbf{w}_i) \quad (5)$$

$$= \mathbf{w}_i - \eta \left(-\frac{1}{N} \sum_{n=1}^N \frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}_i^T \mathbf{x}_n}} \right) \quad (6)$$

$$= \mathbf{w}_i + \eta \left(\frac{1}{N} \sum_{n=1}^N \frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}_i^T \mathbf{x}_n}} \right) \quad (7)$$

where η is the learning rate.

Evaluation of a Classification Model

In machine learning, once we have a result of the classification problem, how do we measure how accurate our classification is? For a regression problem, we have different metrics like R Squared score, Mean Squared Error etc. what are the metrics to measure the credibility of a classification model?

Metrics In a regression problem, the accuracy is generally measured in terms of the difference in the actual values and the predicted values. In a classification problem, the credibility of the model is measured using the confusion matrix generated, i.e., how accurately the true positives and true negatives were predicted. The different metrics used for this purpose are:

- Accuracy
- Recall
- Precision
- F1 Score
- Specifity
- AUC(Area Under the Curve)

- RUC(Receiver Operator Characteristic)

Confusion Matrix

A typical confusion matrix looks like the figure shown.



Where the terms have the meaning:

- ▮ **True Positive(TP):** A result that was predicted as positive by the classification model and also is positive
- ▮ **True Negative(TN):** A result that was predicted as negative by the classification model and also is negative
- ▮ **False Positive(FP):** A result that was predicted as positive by the classification model but actually is negative
- ▮ **False Negative(FN):** A result that was predicted as negative by the classification model but actually is positive.

The Credibility of the model is based on how many correct predictions did the model do.

Accuracy

The mathematical formula is :

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

Or, it can be said that it's defined as the total number of correct classifications divided by the total number of classifications.

Recall or Sensitivity

The mathematical formula is:

$$\text{Recall} = \frac{TP}{(TP+FN)}$$

Or, as the name suggests, it is a measure of: from the total number of positive results how many positives were correctly predicted by the model.

It shows how relevant the model is, in terms of positive results only.

Let's suppose in the previous model, the model gave 50 correct predictions(TP) but failed to identify 200 cancer patients(FN). Recall in that case will be:

Recall= $\frac{50}{(50+200)} = 0.2$ (The model was able to recall only 20% of the cancer patients)

Precision

Precision is a measure of amongst all the positive predictions, how many of them were actually positive. Mathematically,

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

Let's suppose in the previous example, the model identified 50 people as cancer patients(TP) but also raised a false alarm for 100 patients(FP). Hence,

$$\text{Precision} = \frac{50}{(50+100)} = 0.33 \text{ (The model only has a precision of 33\%)}$$

But we have a problem!!

As evident from the previous example, the model had a very high Accuracy but performed poorly in terms of Precision and Recall. So, necessarily *Accuracy* is not the metric to use for evaluating the model in this case.

Imagine a scenario, where the requirement was that the model recalled all the defaulters who did not pay back the loan. Suppose there were 10 such defaulters and to recall those 10 defaulters, and the model gave you 20 results out of which only the 10 are the actual defaulters. Now, the recall of the model is 100%, but the precision goes down to 50%.

A Trade-off?



As observed from the graph, with an increase in the Recall, there is a drop in Precision of the model.

So the question is - what to go for? Precision or Recall?

Well, the answer is: it depends on the business requirement.

For example, if you are predicting cancer, you need a 100 % recall. But suppose you are predicting whether a person is innocent or not, you need 100% precision.

Can we maximise both at the same time? No

So, there is a need for a better metric then?

Yes. And it's called an *F1 Score*

F1 Score

From the previous examples, it is clear that we need a metric that considers both Precision and Recall for evaluating a model. One such metric is the F1 score.

F1 score is defined as the harmonic mean of Precision and Recall.

The mathematical formula is:
$$F1\ score = \frac{2 * ((Precision * Recall))}{(Precision + Recall)}$$

Specificity or True Negative Rate

This represents how specific is the model while predicting the True Negatives. Mathematically,

Specificity = $\frac{TN}{(TN + FP)}$ Or, it can be said that it quantifies the total number of negatives predicted by the model with respect to the total number of actual negative or non favorable outcomes.

Similarly, False Positive rate can be defined as: (1- specificity) Or, $\frac{FP}{(TN + FP)}$


ROC(Receiver Operator Characteristic)

We know that the classification algorithms work on the concept of probability of occurrence of the possible outcomes. A probability value lies between 0 and 1. Zero means that there is no probability of occurrence and one means that the occurrence is certain.

But while working with real-time data, it has been observed that we seldom get a perfect 0 or 1 value. Instead of that, we get different decimal values lying between 0 and 1. Now the question is if we are not getting binary probability values how are we actually determining the class in our classification problem?

There comes the concept of Threshold. A threshold is set, any probability value below the threshold is a negative outcome, and anything more than the threshold is a favourable or the positive outcome. For Example, if the threshold is 0.5, any probability value below 0.5 means a negative or an unfavourable outcome and any value above 0.5 indicates a positive or favourable outcome.

Now, the question is, what should be an ideal threshold?

The following diagram shows a typical logistic regression curve. 

- The horizontal lines represent the various values of thresholds ranging from 0 to 1.

- Let's suppose our classification problem was to identify the obese people from the given data.
- The green markers represent obese people and the red markers represent the non-obese people.
- Our confusion matrix will depend on the value of the threshold chosen by us.
- For Example, if 0.25 is the threshold then
 - TP(actually obese)=3
 - TN(Not obese)=2
 - FP(Not obese but predicted obese)=2(the two red squares above the 0.25 line)
 - FN(Obese but predicted as not obese)=1(Green circle below 0.25line)

A typical ROC curve looks like the following figure. 

- Mathematically, it represents the various confusion matrices for various thresholds. Each black dot is one confusion matrix.
- The green dotted line represents the scenario when the true positive rate equals the false positive rate.
- As evident from the curve, as we move from the rightmost dot towards left, after a certain threshold, the false positive rate decreases.
- After some time, the false positive rate becomes zero.
- The point encircled in green is the best point as it predicts all the values correctly and keeps the False positive as a minimum.
- But that is not a rule of thumb. Based on the requirement, we need to select the point of a threshold.
- The ROC curve answers our question of which threshold to choose.

But we have a confusion!!

Let's suppose that we used different classification algorithms, and different ROCs for the corresponding algorithms have been plotted. The question is: which algorithm to choose now? The answer is to calculate the area under each ROC curve.

AUC(Area Under Curve)



- It helps us to choose the best model amongst the models for which we have plotted the ROC curves
- The best model is the one which encompasses the maximum area under it.
- In the adjacent diagram, amongst the two curves, the model that resulted in the red one should be chosen as it clearly covers more area than the blue one

Python Implementation

```
In [2]: #Let's start with importing necessary libraries
```

```

import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import Ridge, Lasso, RidgeCV, LassoCV, ElasticNet, ElasticNetCV, LogisticRegression
from sklearn.model_selection import train_test_split
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.metrics import accuracy_score, confusion_matrix, roc_curve, roc_auc_score
import matplotlib.pyplot as plt
import seaborn as sns
# import scikitplot as skl
sns.set()

```

In []:

In [3]: `data = pd.read_csv("diabetes.csv") # Reading the Data`
`data.head()`

Out[3]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

In [5]: `data.columns`

Out[5]: Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
dtype='object')

In [6]: `data.shape`

Out[6]: (768, 9)

In [7]: `data.describe().T`

Out[7]:

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00

SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

In [8]: `data.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Pregnancies            768 non-null    int64
1   Glucose                768 non-null    int64
2   BloodPressure          768 non-null    int64
3   SkinThickness          768 non-null    int64
4   Insulin                768 non-null    int64
5   BMI                   768 non-null    float64
6   DiabetesPedigreeFunction 768 non-null    float64
7   Age                   768 non-null    int64
8   Outcome                768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Observation:- All the values are either int or float there is no object or string in the dataset.

In [11]: `data.isnull().sum()`

```
Out[11]: Pregnancies            0
Glucose                0
BloodPressure          0
SkinThickness          0
Insulin                0
BMI                   0
DiabetesPedigreeFunction 0
Age                   0
Outcome                0
dtype: int64
```

Observation:- There is no null value in the data

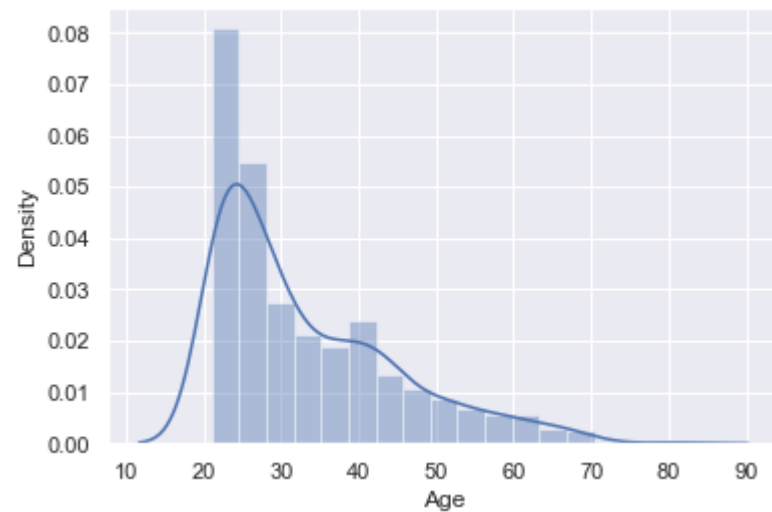
It seems that there are no missing values in our data. Great, let's see the distribution of data:

```
In [16]: sns.distplot(data['Age']) # distribution plot for Age column
```

```
/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
```

```
warnings.warn(msg, FutureWarning)
```

```
Out[16]: <AxesSubplot:xlabel='Age', ylabel='Density'>
```



```
In [19]: # let's see how data is distributed for every column
plt.figure(figsize=(20,25), facecolor='white')
plotnumber = 1

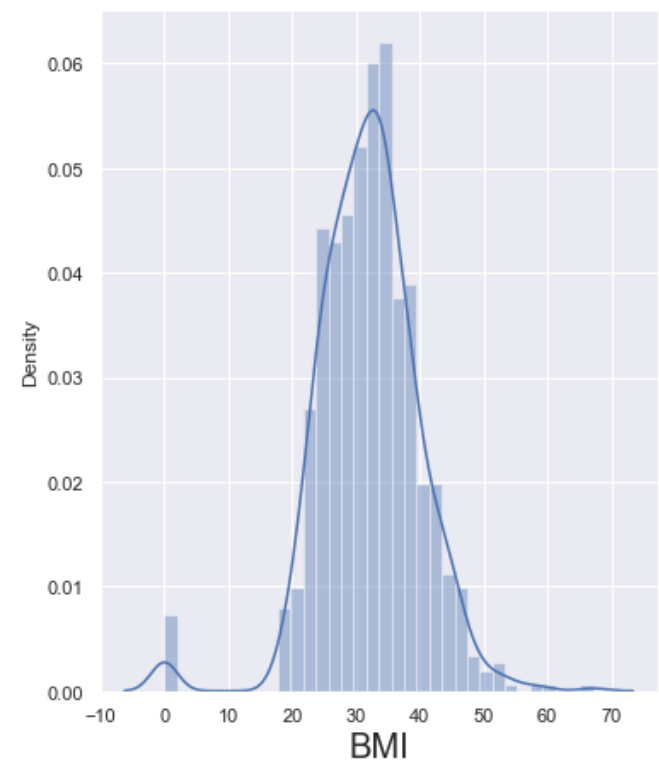
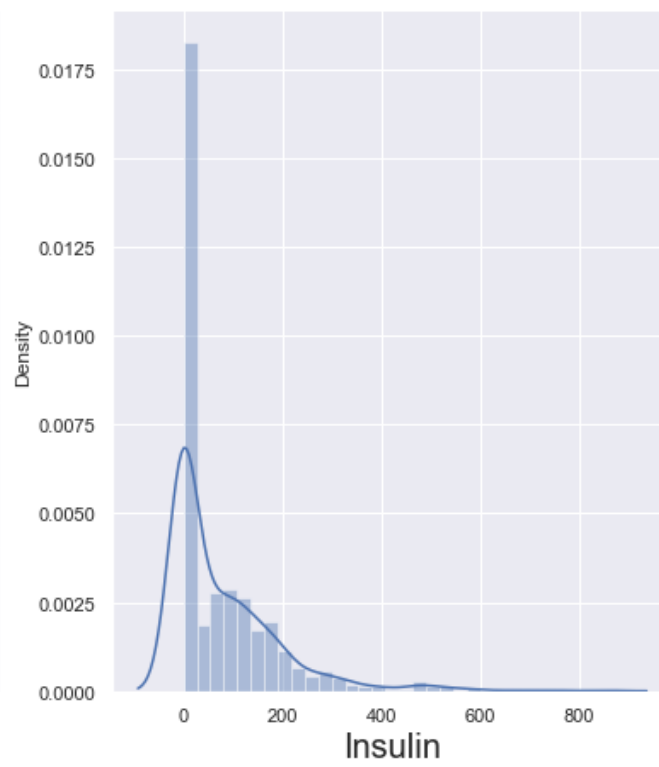
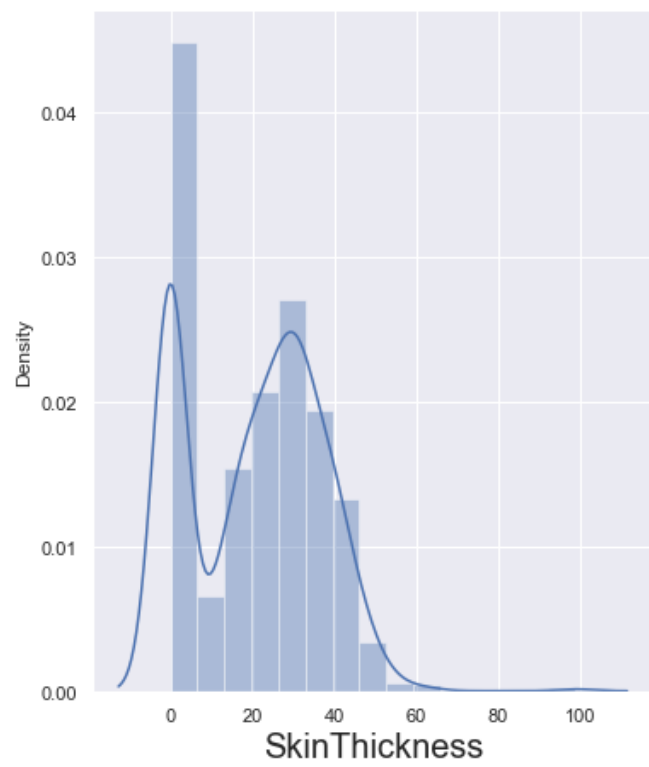
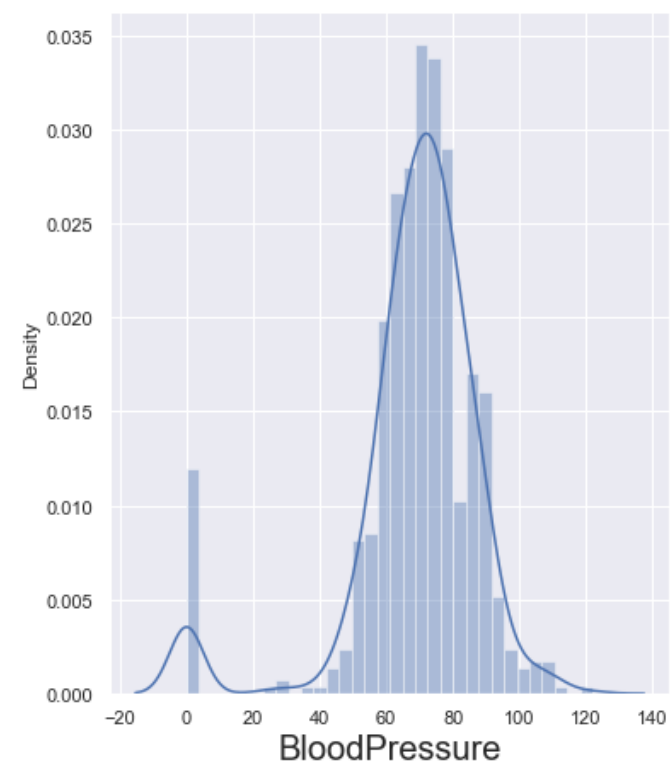
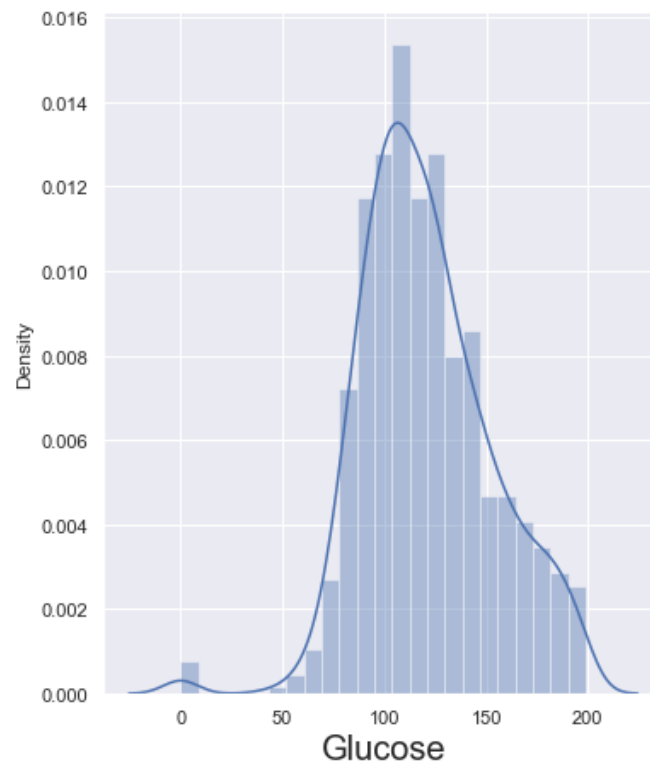
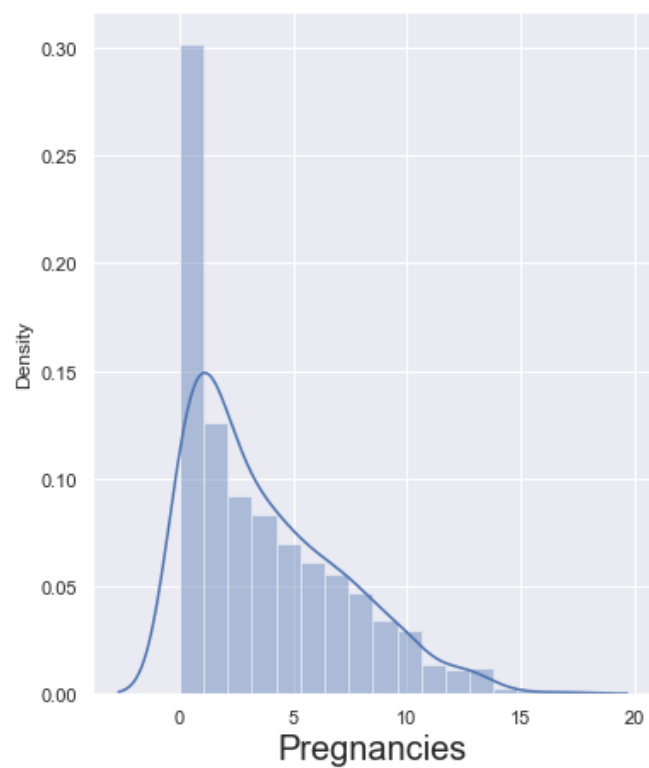
for column in data:
    if plotnumber<=9 :      # as there are 9 columns in the data
        ax = plt.subplot(3,3,plotnumber)
        sns.distplot(data[column]) # plot the distribution plot.
        plt.xlabel(column,fontsize=20)
        #plt.ylabel('Salary',fontsize=20)
        plotnumber+=1
plt.show()
```

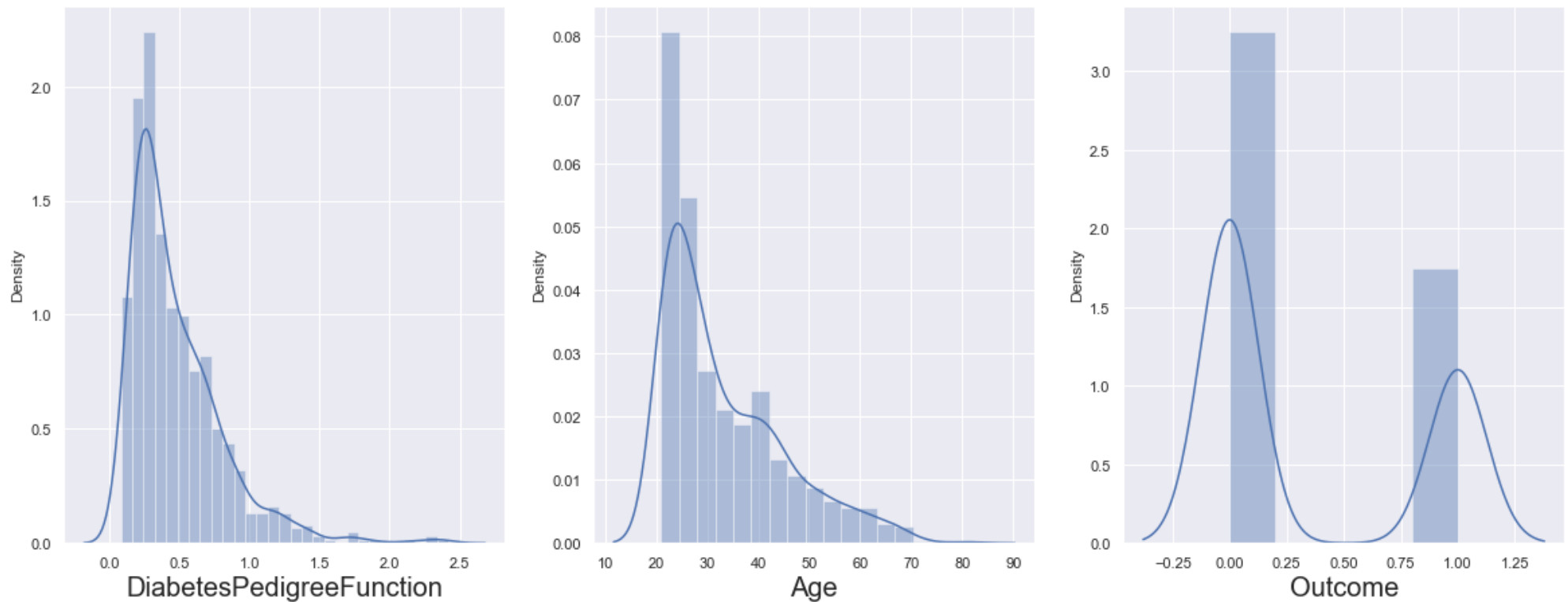
```
/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
```

```
warnings.warn(msg, FutureWarning)
```

```
/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
```

```
warnings.warn(msg, FutureWarning)
/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated
function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).
    warnings.warn(msg, FutureWarning)
/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated
function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).
    warnings.warn(msg, FutureWarning)
/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated
function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).
    warnings.warn(msg, FutureWarning)
/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated
function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).
    warnings.warn(msg, FutureWarning)
/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated
function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).
    warnings.warn(msg, FutureWarning)
/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated
function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).
    warnings.warn(msg, FutureWarning)
/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated
function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).
    warnings.warn(msg, FutureWarning)
```





We can see there is some skewness in the data, let's deal with data.

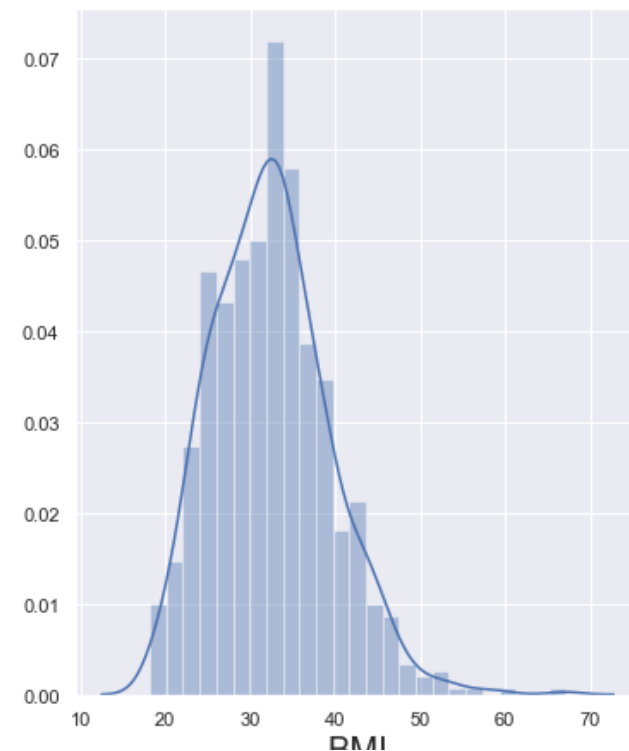
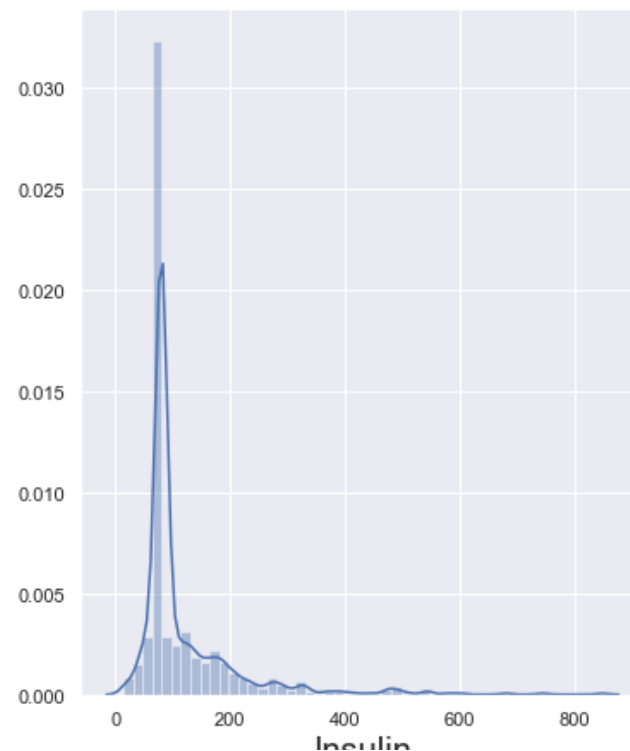
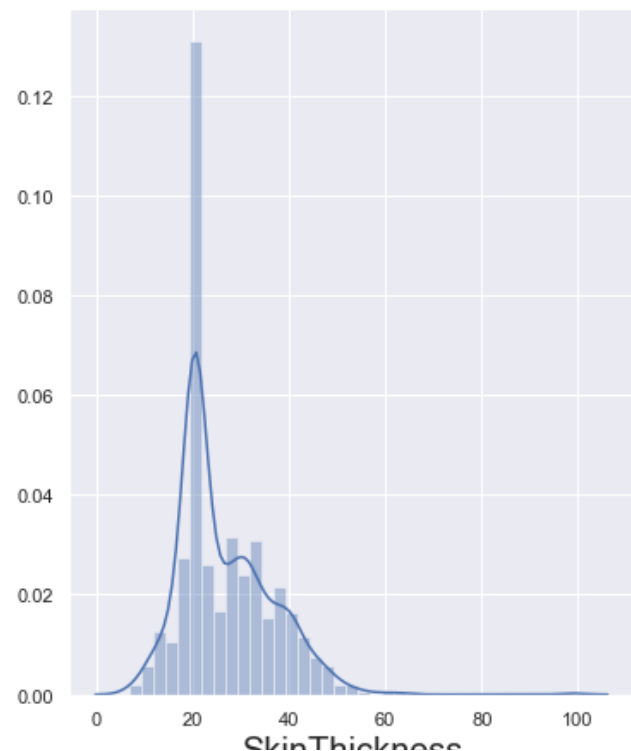
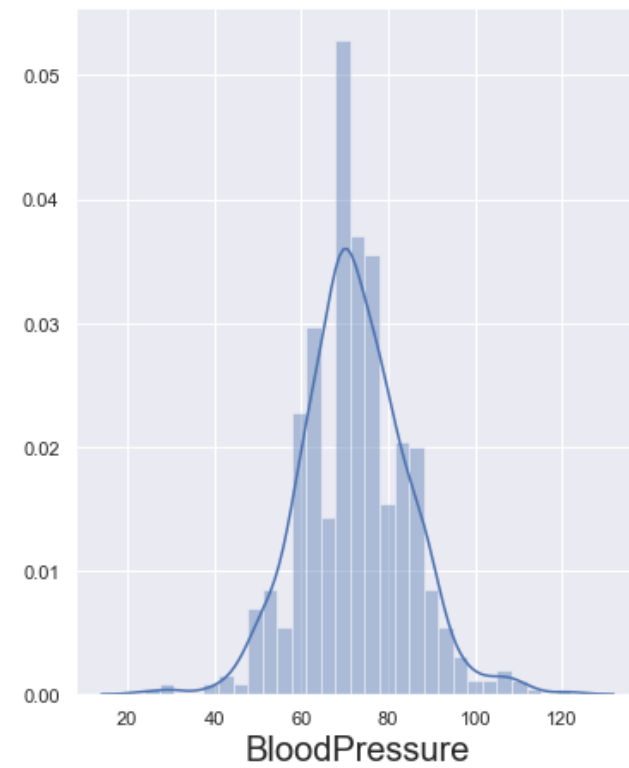
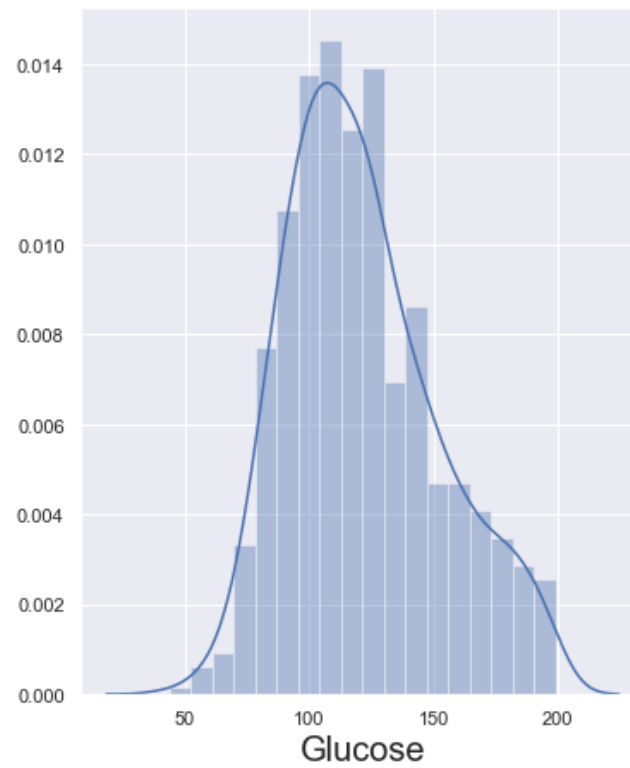
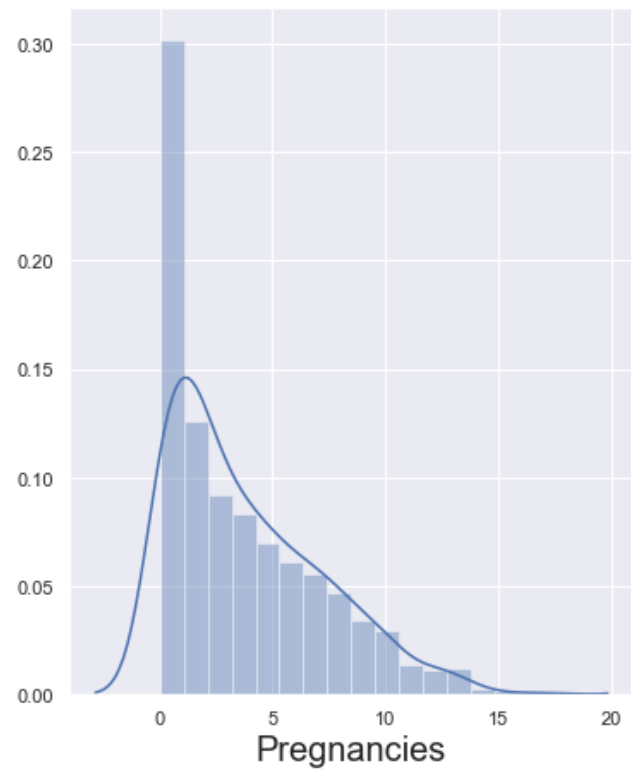
Also, we can see there few data for columns Glucose, Insulin, skin thickness, BMI and Blood Pressure which have value as 0. That's not possible. You can do a quick search to see that one cannot have 0 values for these. Let's deal with that. we can either remove such data or simply replace it with their respective mean values. Let's do the latter.

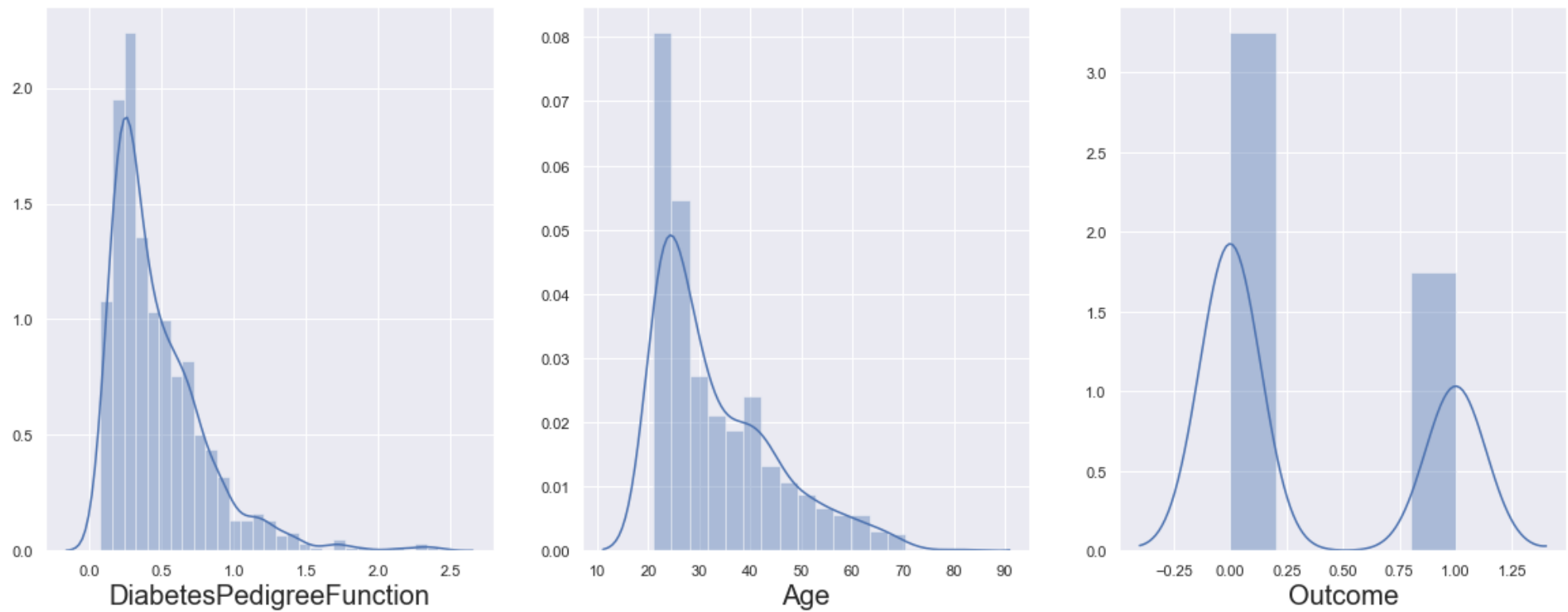
```
In [6]: # replacing zero values with the mean of the column
data['BMI'] = data['BMI'].replace(0,data['BMI'].mean())
data['BloodPressure'] = data['BloodPressure'].replace(0,data['BloodPressure'].mean())
data['Glucose'] = data['Glucose'].replace(0,data['Glucose'].mean())
data['Insulin'] = data['Insulin'].replace(0,data['Insulin'].mean())
data['SkinThickness'] = data['SkinThickness'].replace(0,data['SkinThickness'].mean())
```

```
In [7]: # let's see how data is distributed for every column
plt.figure(figsize=(20,25), facecolor='white')
plotnumber = 1

for column in data:
    if plotnumber<=9 :
```

```
ax = plt.subplot(3,3,plotnumber)
sns.distplot(data[column])
plt.xlabel(column,fontsize=20)
#plt.ylabel('Salary',fontsize=20)
plotnumber+=1
plt.show()
```

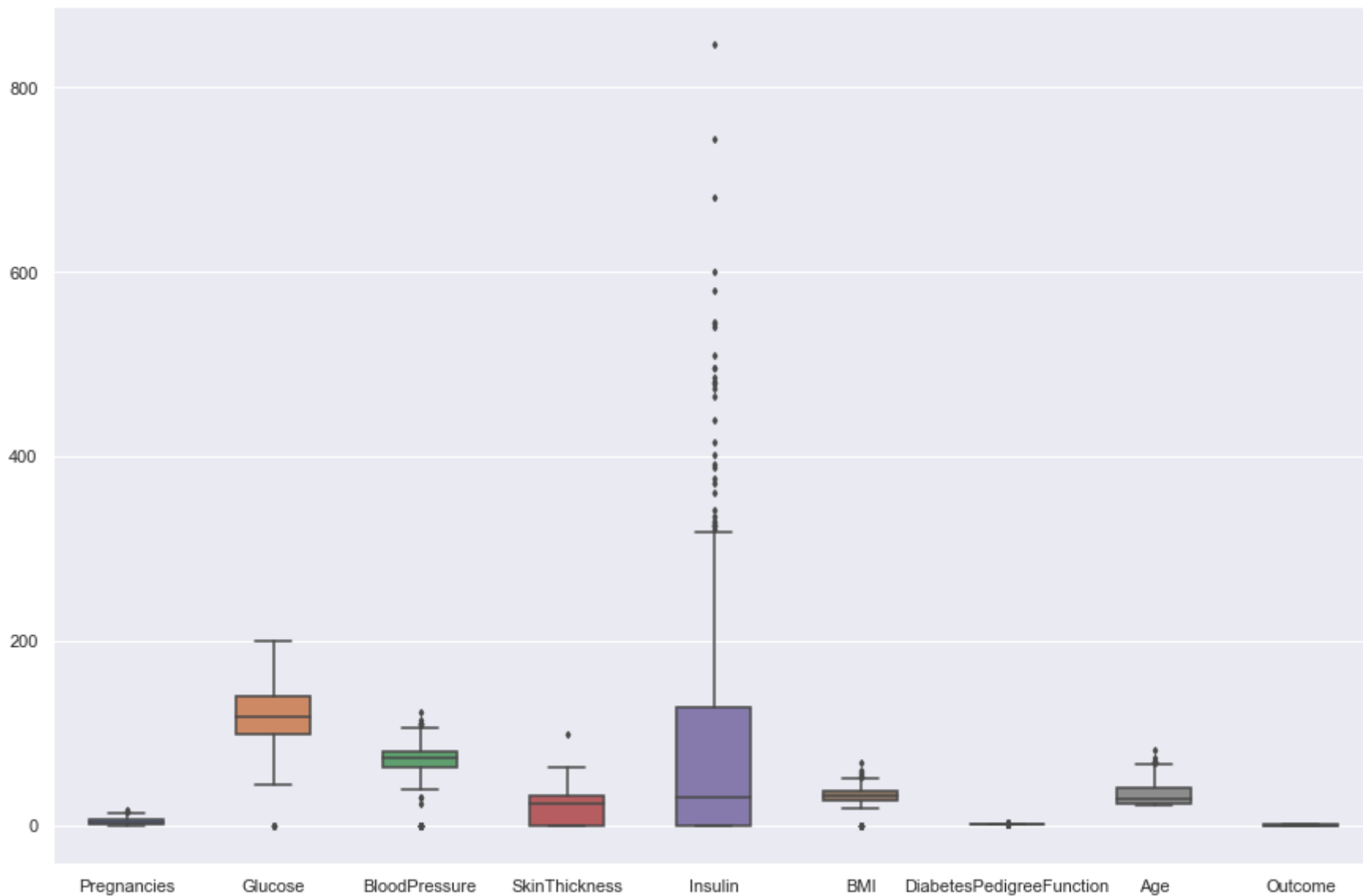





Now we have dealt with the 0 values and data looks better. But, there still are outliers present in some columns. Let's deal with them.

```
In [20]: # Check the outliers using box plot.
fig, ax = plt.subplots(figsize=(15,10))
sns.boxplot(data=data, width= 0.5,ax=ax, fliersize=3)
```

```
Out[20]: <AxesSubplot:>
```



From observation we are assuming and taking some percentage of data in cosideration and trying to remove the putlier.

```
In [27]: # we are removing the top 2% data from the Pregnancies column
q = data['Pregnancies'].quantile(0.98)
data_cleaned = data[data['Pregnancies']<q]
```

```
In [28]: # we are removing the top 1% data from the BMI column
q = data_cleaned['BMI'].quantile(0.99)
data_cleaned = data_cleaned[data_cleaned['BMI']<q]
```

```
In [29]: # we are removing the top 1% data from the SkinThickness column
q = data_cleaned['SkinThickness'].quantile(0.99)
```

```
data_cleaned = data_cleaned[data_cleaned['SkinThickness']<q]
```

```
In [30]: # we are removing the top 5% data from the Insulin column
q = data_cleaned['Insulin'].quantile(0.95)
data_cleaned = data_cleaned[data_cleaned['Insulin']<q]
```

```
In [31]: # we are removing the top 1% data from the DiabetesPedigreeFunction column
q = data_cleaned['DiabetesPedigreeFunction'].quantile(0.99)
data_cleaned = data_cleaned[data_cleaned['DiabetesPedigreeFunction']<q]
```

```
In [32]: # we are removing the top 1% data from the Age column
q = data_cleaned['Age'].quantile(0.99)
data_cleaned = data_cleaned[data_cleaned['Age']<q]
```

```
In [26]: # Before cleaning the shape of data is
data.shape
```

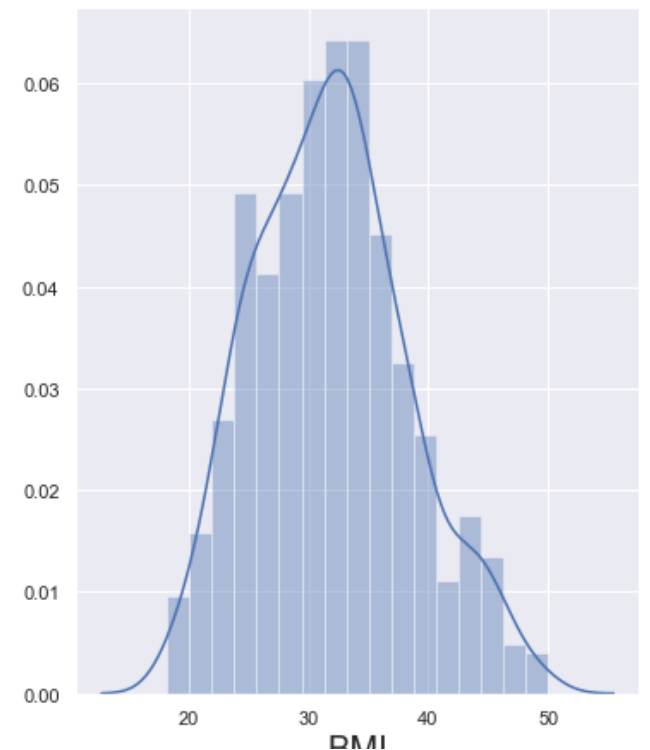
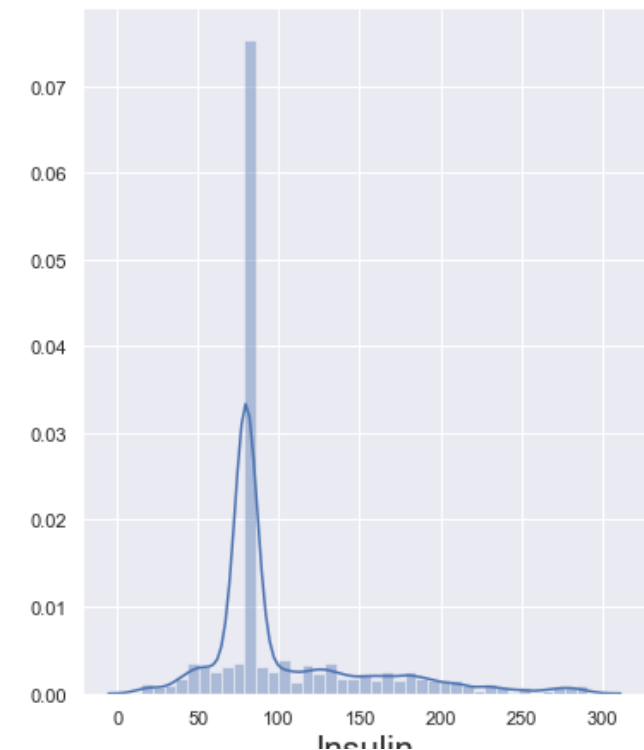
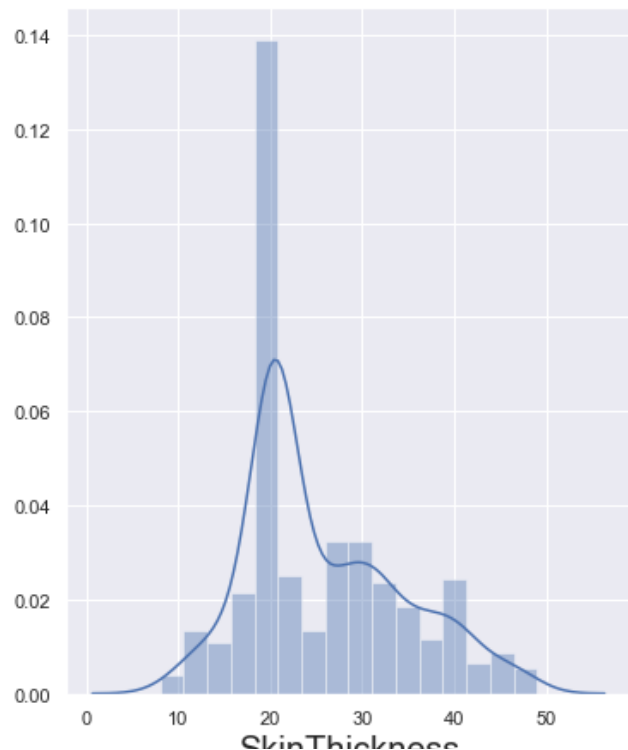
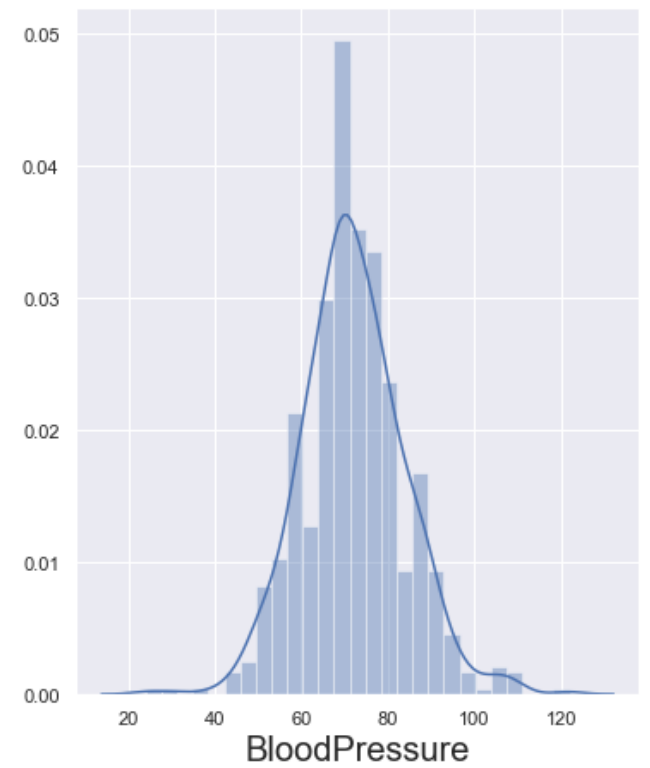
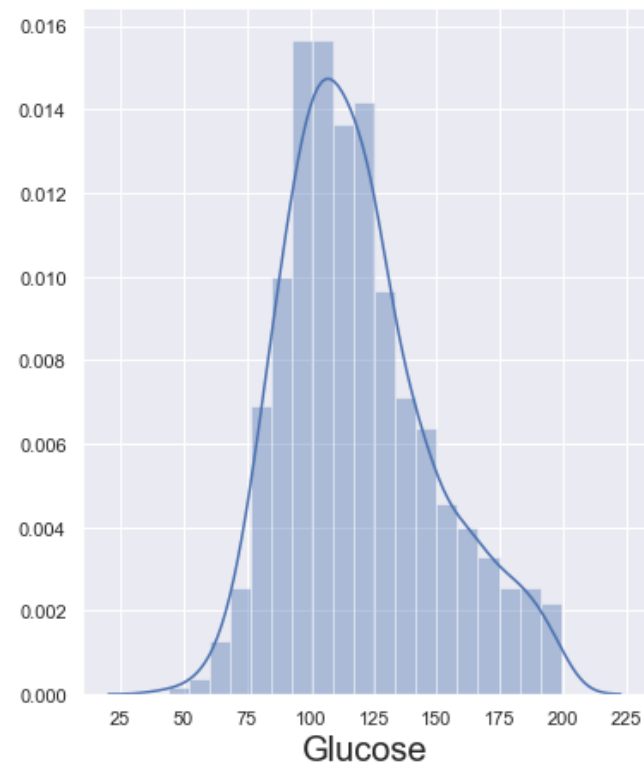
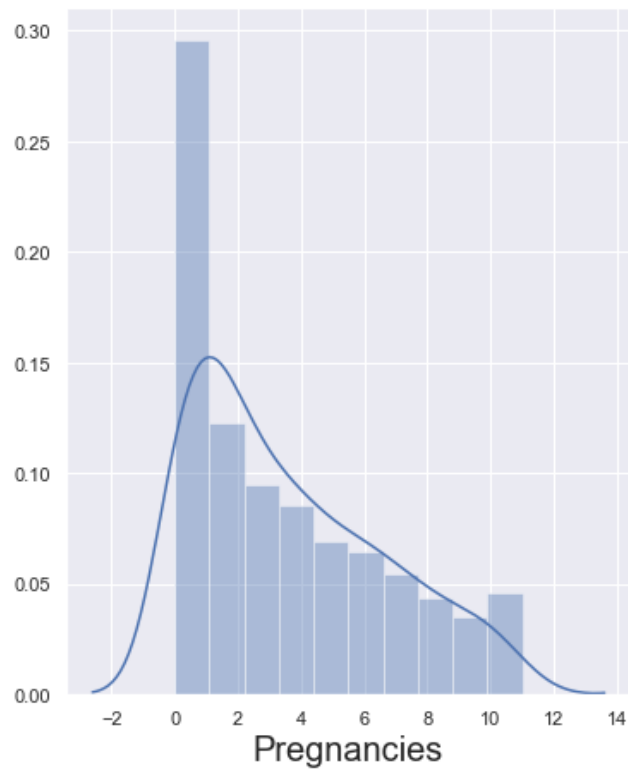
```
Out[26]: (768, 9)
```

```
In [24]: # After cleaning check the shape of sata
data_cleaned.shape
```

```
Out[24]: (674, 9)
```

```
In [10]: # let's see how data is distributed for every column
plt.figure(figsize=(20,25), facecolor='white')
plotnumber = 1

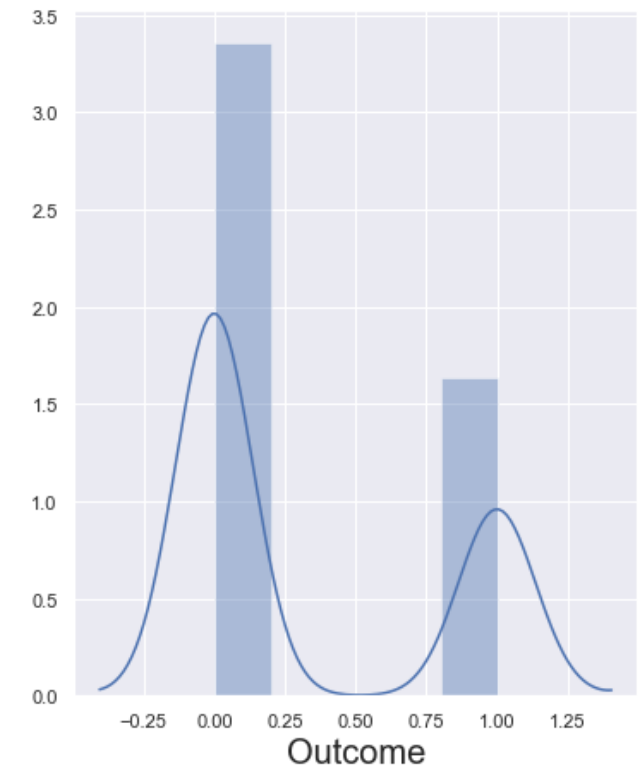
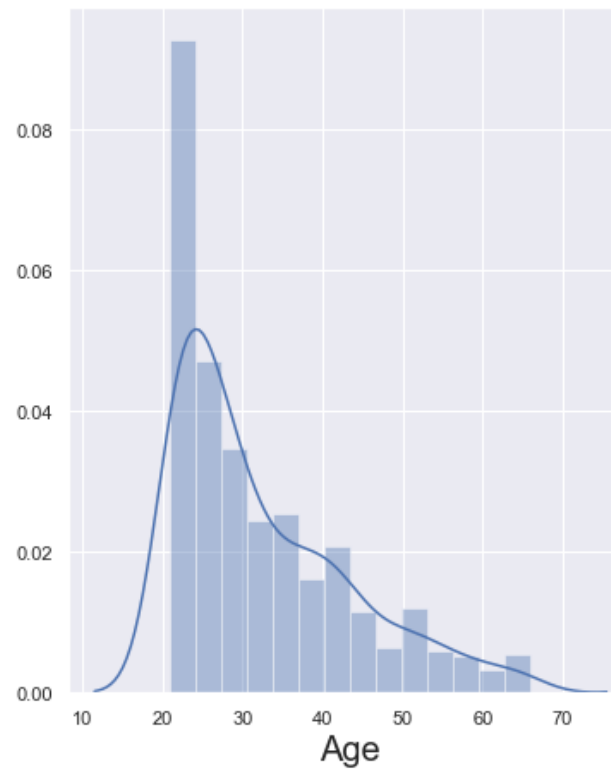
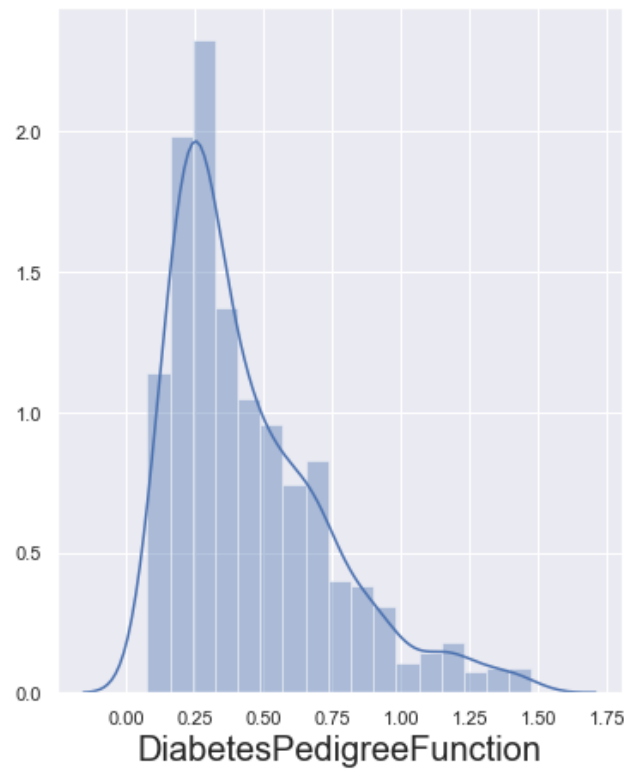
for column in data_cleaned:
    if plotnumber<=9 :
        ax = plt.subplot(3,3,plotnumber)
        sns.distplot(data_cleaned[column])
        plt.xlabel(column,fontsize=20)
        #plt.ylabel('Salary',fontsize=20)
        plotnumber+=1
plt.show()
```



SKIN thickness

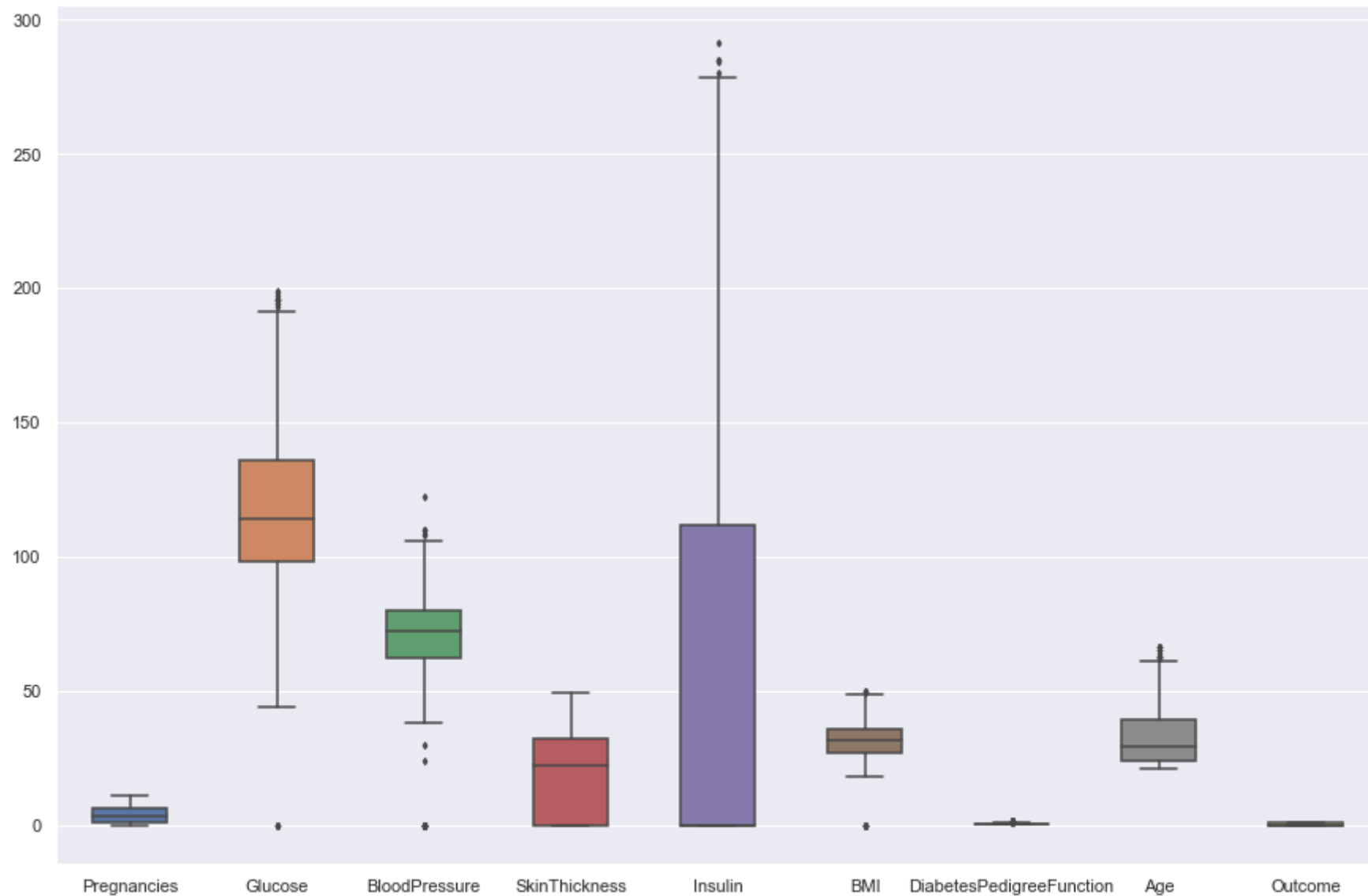
Insulin

BMI



```
In [33]: # Again Check the outliers using box plot.
fig, ax = plt.subplots(figsize=(15,10))
sns.boxplot(data=data_cleaned, width=0.5, ax=ax, fliersize=3)
```

Out[33]: <AxesSubplot:>



The data looks much better now than before. We will start our analysis with this data now as we don't want to lose important information. If our model doesn't work with accuracy, we will come back for more preprocessing. To remove the outlier further we can use log transformation on box-cox transformation

```
In [41]: # Keep the independent data/ i/p data inside X-variable
X = data_cleaned.drop(columns = ['Outcome'])
X.head()
```

```
Out[41]:
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	6	148	72	35	0	33.6	0.627	50

1	1	85	66	29	0	26.6	0.351	31
2	8	183	64	0	0	23.3	0.672	32
3	1	89	66	23	94	28.1	0.167	21
5	5	116	74	0	0	25.6	0.201	30

```
In [43]: # Put the o/p data i.e dependent variable inside y-variable
y = data_cleaned['Outcome']
y.head()
```

```
Out[43]: 0    1
1    0
2    1
3    0
5    0
Name: Outcome, dtype: int64
```

Before we fit our data to a model, let's visualize the relationship between our independent variables and the categories.

```
In [45]: # let's see how data is distributed for every column
plt.figure(figsize=(20,25), facecolor='white')
plotnumber = 1

for column in X:
    if plotnumber<=8 :
        ax = plt.subplot(3,3,plotnumber)
        sns.stripplot(y,X[column]) # strip plot is user to check the distribution of data.
        plotnumber+=1
plt.tight_layout()
```

/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.

warnings.warn(

/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.

warnings.warn(

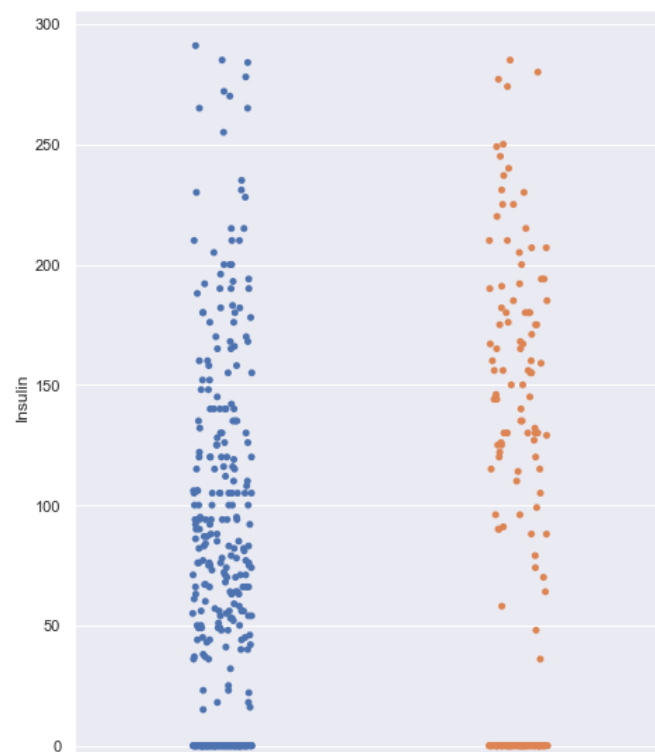
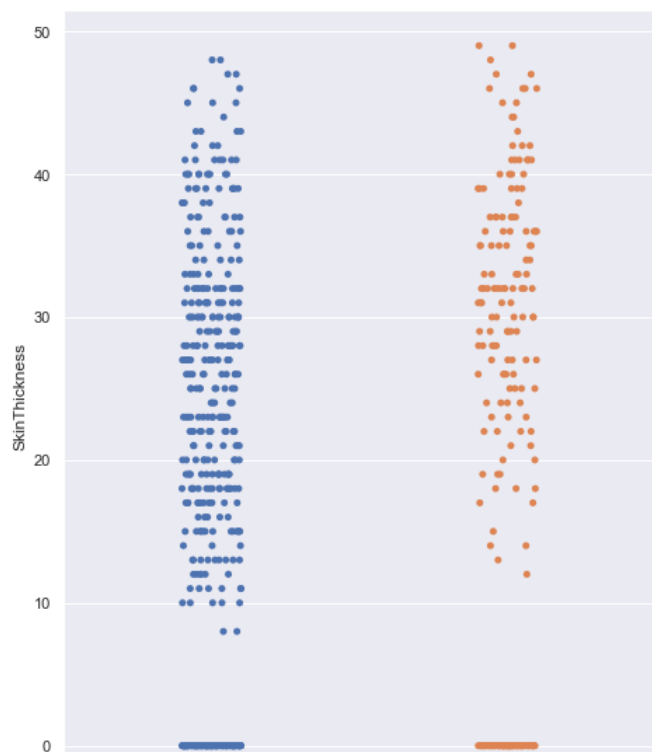
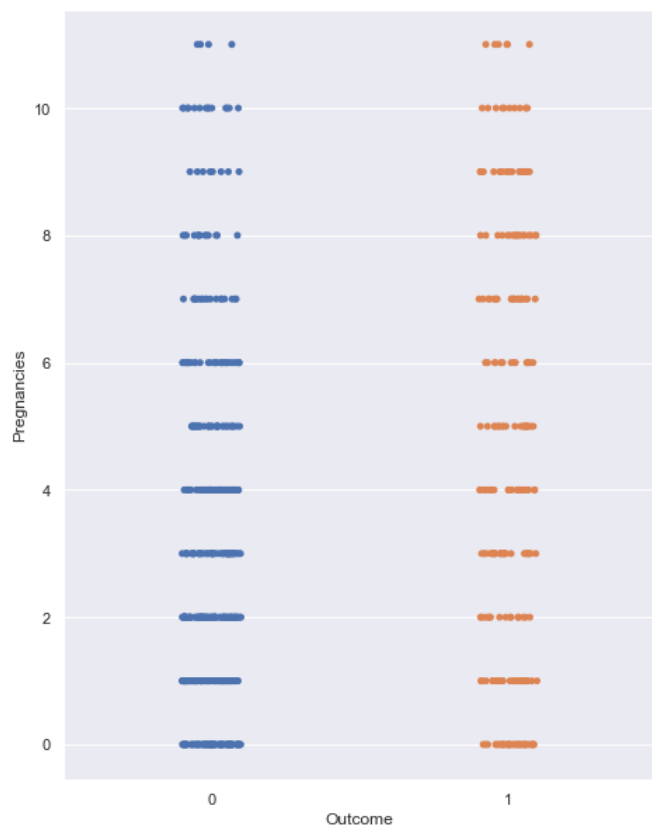
/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.

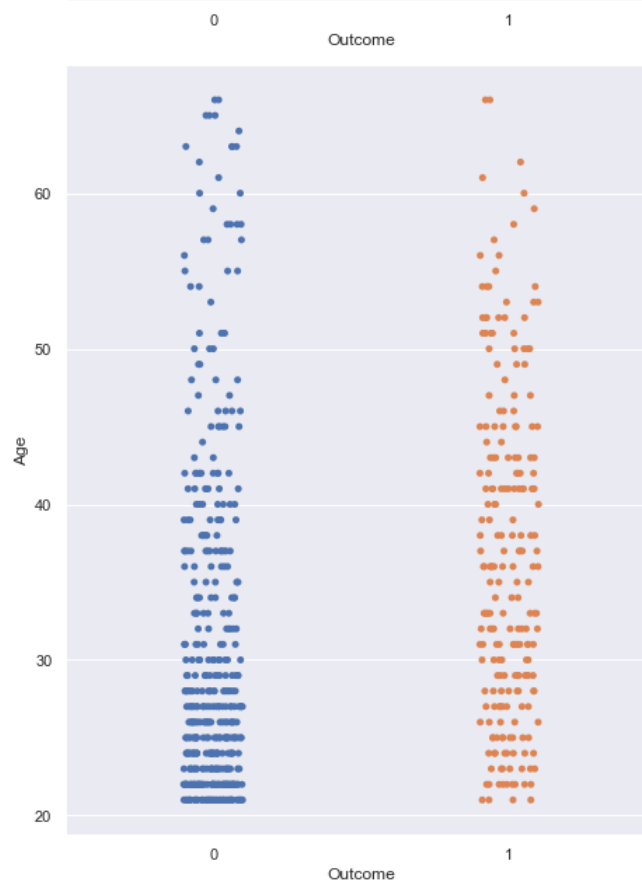
warnings.warn(

/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.

warnings.warn(


```
/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
/Users/madhu/opt/anaconda3/lib/python3.9/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```





Great!! Let's proceed by checking multicollinearity in the dependent variables. Before that, we should scale our data. Let's use the standard scaler for that.

```
In [46]: scalar = StandardScaler()
X_scaled = scalar.fit_transform(X)
```

This is how our data looks now after scaling. Great, now we will check for multicollinearity using VIF(Variance Inflation factor)

```
In [47]: X_scaled
```

```
Out[47]: array([[ 0.79675391,  0.95778209,  0.16945124, ...,  0.2886168 ,
        0.63048454,  1.60141519],
       [-0.86479354, -1.07527845, -0.13992159, ..., -0.66301789,
        -0.33807867, -0.13270648],
       [ 1.46137289,  2.08726017, -0.24304586, ..., -1.11164568,
        0.78840246, -0.04143692],
       ...,
       [ 0.46444442,  0.08647043,  0.16945124, ..., -0.71739702,
```

```
-0.71006309, -0.22397605],
[-0.86479354,  0.24782444, -0.44929441, ..., -0.18720054,
-0.34509724,  1.3276065 ],
[-0.86479354, -0.81711203,  0.06632696, ..., -0.1464162 ,
-0.464413 , -0.86286298]])
```

```
In [15]: # Perform multicollinearity
vif = pd.DataFrame()
vif["vif"] = [variance_inflation_factor(X_scaled,i) for i in range(X_scaled.shape[1])]
vif["Features"] = X.columns

#let's check the values
vif
```

```
Out[15]:
```

	vif	Features
0	1.431075	Pregnancies
1	1.347308	Glucose
2	1.247914	BloodPressure
3	1.450510	SkinThickness
4	1.262111	Insulin
5	1.550227	BMI
6	1.058104	DiabetesPedigreeFunction
7	1.605441	Age

All the VIF values are less than 5 and are very low. That means no multicollinearity. Now, we can go ahead with fitting our data to the model. Before that, let's split our data in test and training set.

```
In [49]: # Segregate the data into train and test data. 25% data will be used as test data.
x_train,x_test,y_train,y_test = train_test_split(X_scaled,y, test_size= 0.25, random_state = 355)
```

```
In [50]: # train the data
log_reg = LogisticRegression()
log_reg.fit(x_train,y_train)
```

```
Out[50]: LogisticRegression()
```

```
In [82]: # Standard Scaler Object
with open('sandardScaler.sav', 'wb') as f:
    pickle.dump(scaler,f)
```

```
In [84]: # Model Saving or pickling our model
import pickle
# Writing different model files to file
with open('logreg.pkl', 'wb') as f:
    pickle.dump(log_reg, f)
```

```
In [87]: # Load the logreg file
with open('logreg.pkl', 'rb') as f:
    pickle.load(f)
```

Let's see how well our model performs on the test data set.

```
In [69]: # predict the o/p
y_pred = log_reg.predict(x_test)
```

```
In [70]: y_pred
```

```
Out[70]: array([0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0,
        1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1,
        0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
        0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
        0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1,
        0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
        0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0,
        1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0])
```

```
In [71]: # 82% Accuracy
accuracy = accuracy_score(y_test, y_pred)
accuracy
```

```
Out[71]: 0.8284023668639053
```

```
In [72]: # Confusion Matrix
conf_mat = confusion_matrix(y_test, y_pred)
conf_mat
```

```
Out[72]: array([[109,  8],
        [ 21, 31]])
```

```
In [73]: true_positive = conf_mat[0][0]
true_positive
```

```
Out[73]: 109
```

```
In [74]: false_positive = conf_mat[0][1]
false_negative = conf_mat[1][0]
```

```
true_negative = conf_mat[1][1]
```

```
In [75]: # Breaking down the formula for Accuracy
Accuracy = (true_positive + true_negative) / (true_positive + false_positive + false_negative + true_negative)
Accuracy
```

```
Out[75]: 0.8284023668639053
```

```
In [76]: # Calculate Precision
Precision = true_positive / (true_positive + false_positive)
Precision
```

```
Out[76]: 0.9316239316239316
```

```
In [77]: # Recall
Recall = true_positive / (true_positive + false_negative)
Recall
```

```
Out[77]: 0.8384615384615385
```

```
In [78]: # F1 Score
F1_Score = 2 * (Recall * Precision) / (Recall + Precision)
F1_Score
```

```
Out[78]: 0.882591093117409
```

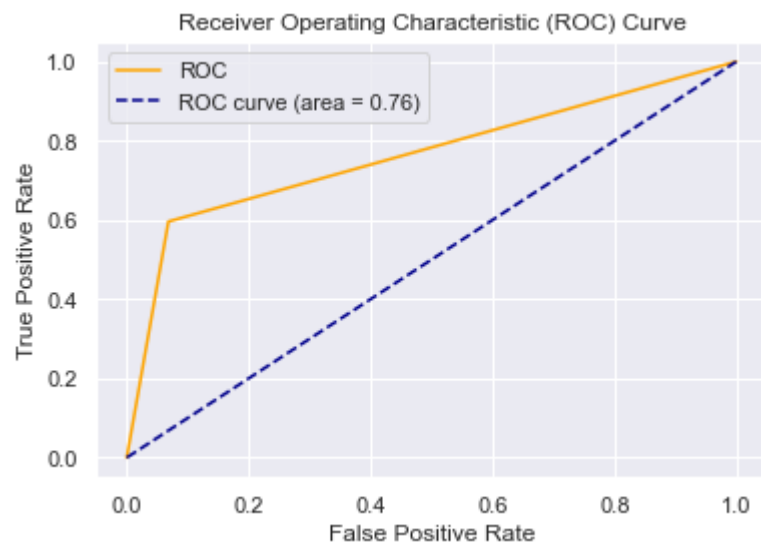
```
In [79]: # Area Under Curve
auc = roc_auc_score(y_test, y_pred)
auc
```

```
Out[79]: 0.7638888888888889
```

ROC

```
In [80]: fpr, tpr, thresholds = roc_curve(y_test, y_pred)
```

```
In [81]: plt.plot(fpr, tpr, color='orange', label='ROC')
plt.plot([0, 1], [0, 1], color='darkblue', linestyle='--', label='ROC curve (area = %0.2f)' % auc)
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend()
plt.show()
```



What is the significance of Roc curve and AUC?

In real life, we create various models using different algorithms that we can use for classification purpose. We use AUC to determine which model is the best one to use for a given dataset. Suppose we have created Logistic regression, SVM as well as a clustering model for classification purpose. We will calculate AUC for all the models separately. The model with highest AUC value will be the best model to use.

Advantages of Logistic Regression

- It is very simple and easy to implement.
- The output is more informative than other classification algorithms
- It expresses the relationship between independent and dependent variables
- Very effective with linearly separable data

Disadvantages of Logistic Regression

- Not effective with data which are not linearly separable
- Not as powerful as other classification models
- Multiclass classifications are much easier to do with other algorithms than logistic regression
- It can only predict categorical outcomes

Cloud Deployment (Heroku)

Once the training is completed, we need to expose the trained model as an API for the user to consume it. For prediction, the saved model is loaded first and then the predictions are made using it. If the web app works fine, the same app is deployed to the cloud platform. The application flow for cloud deployment looks like:



Pre-requisites for cloud deployment:

- Basic knowledge of flask framework.
- Any Python IDE installed(we are using PyCharm).
- A Heroku account.
- Basic understanding of HTML.

Steps before cloud deployment:

We need to change our code a bit so that it works unhindered on the cloud, as well.

- Add a file called 'gitignore' inside the project folder. This folder contains the list of the files which we don't want to include in the git repository. My gitignore file looks like:

.idea

As I am using PyCharm as an IDE, and it's provided by the IntelliJ Idea community, it automatically adds the .idea folder containing some metadata. We need not include them in our cloud app.

- Add a file called 'Procfile' inside the 'reviewScrapper' folder. This folder contains the command to run the flask application once deployed on the server:

web: gunicorn app:app

Here, the keyword 'web' specifies that the application is a web application. And the part 'app:app' instructs the program to look for a flask application called 'app' inside the 'app.py' file. Gunicorn is a Web Server Gateway Interface (WSGI) HTTP server for Python.

- Open a command prompt window and navigate to your 'reviewScrapper' folder. Enter the command 'pip freeze > requirements.txt'. This command generates the 'requirements.txt' file

The requirements.txt helps the Heroku cloud app to install all the dependencies before starting the webserver.

After performing all the above steps the project structure will look like:



Deployment to Heroku:

- After installing the Heroku CLI, Open a command prompt window and navigate to your project folder.
- Type the command **heroku login** to login to your heroku account.
- After logging in to Heroku, enter the command **heroku create** to create a heroku app. It will give you the URL of your Heroku app after successful creation. Or alternatively, you can go to the heroku website and create an app directly.
- Before deploying the code to the Heroku cloud, we need to commit the changes to the git repository.
- Type the command **git init** to initialize a local git repository.
- Enter the command **git status** to see the uncommitted changes.
- Enter the command **git add .** to add the uncommitted changes to the local repository.
- Enter the command **git commit -am "make it better"** to commit the changes to the local repository.
- Enter the command **git push heroku master** to push the code to the heroku cloud.
- After deployment, heroku gives you the URL to hit the web API.
- Once your application is deployed successfully, enter the command **heroku logs --tail** to see the logs.

In []: