**Objective**

The objective of the project is to “use the machine learning workflow to process and transform pima Indian data to create a prediction model. This model predicts which people are likely to develop diabetes with 70% or greater accuracy”.

**Pre-processing/Data Cleansing**: Before building a model, the data is explored. Every variable is to be scrutinized and is checked to handle the outliers. Missing values are investigated properly to check their significance. Missing values can also be imputed by the mean or the median of a variable

**Python Libraries Used:** I am using pandas (to read the data, identifying correlated features, cleaning the data, molding the data, checking true/false ratio), matplotlib (for plotting).

**Model Building:** After performing the initial exploration, the data is divided into two groups of training and testing sets. If the model is built on full data, there can be a risk of over fitting and the model might fail in the real world. To reduce this, the data is divided into two sets of 70% training and 30% testing.

By building a model on 70% and testing on rest of 30% untouched data, the effectiveness of a model could be assessed. In addition, the relationship between the variables are be tested. The relation between each independent variable and the response variable is checked such that the independent variables behaved in a similar way in a model as they behaved individually when run against the response variable. Additionally, highly correlated independent variables are removed otherwise there could be a possibility of signs of coefficients of variables being reversed.

**Checking Performance Measures**: This section discusses the different measures used to assess the performance of a model. As discussed above, data split is done through 70% of data in training and 30% of data in testing. Different predictive learning techniques learn the data from the training set and build a model on it. It then implements the model on the testing set to check the performance of a model. There are different measures to check the performance of the model.

The following links provided are some of the references:

<https://www.kaggle.com/uciml/pima-indians-diabetes-database/data> to data

<https://jakevdp.github.io/PythonDataScienceHandbook/05.05-naive-bayes.html> to textbook reference for machine learning with python

<https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>

----- for logistic regression with scikit learn

<https://machinelearningmastery.com/get-your-hands-dirty-with-scikit-learn-now/> view of code for different algorithms in machine learning with python

<https://machinelearningmastery.com/machine-learning-in-python-step-by-step/> step by step explanation of dealing with data using python.