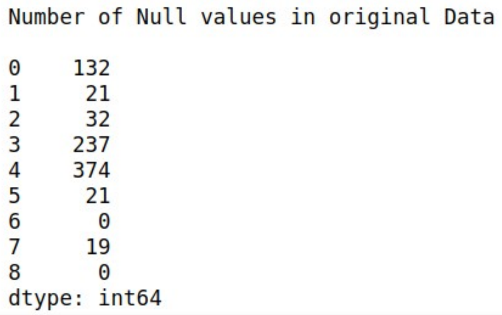
ASSIGNMENT - PIMA INDIAN DIABETES

1) Problem Statement

The type of dataset and problem is a classic supervised binary classification. Given a number of elements all with certain characteristics (features), we want to build a machine learning model to identify people affected by type 2 diabetes.
To solve the problem we will have to analyse the data, do any required transformation and normalisation, apply a machine learning algorithm, train a model, check the performance of the trained model and iterate with other algorithms until we find the most performant for our type of dataset.

2) Missing Data Handling

Because of the small size of the dataset, entire row or column with missing values cannot be deleted as it will further reduce the dataset size resulting in loss of information.
Therefore, to handle missing data, we are finding the median of each column and replacing NaN with it.



```
Number of Null values in original Data

0    132
1     21
2     32
3    237
4    374
5     21
6      0
7     19
8      0
dtype: int64
```

Fig. - No. of missing values

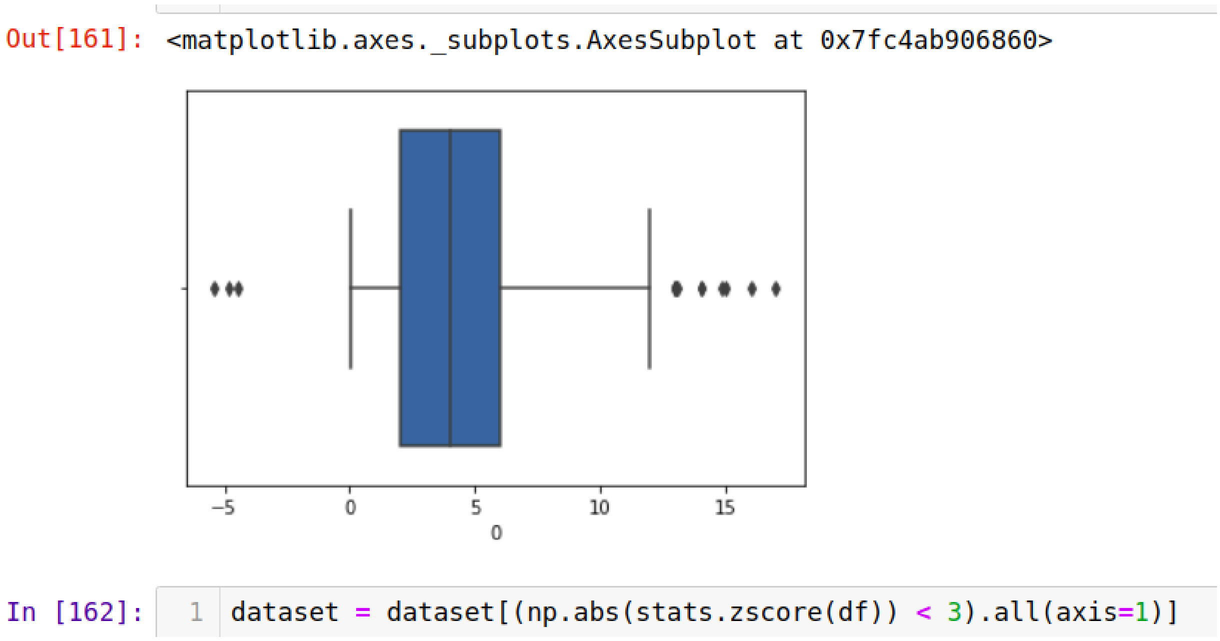
3) Data Preprocessing

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set.

Detecting and Removing Outliers

Outliers are very important because they affect the mean and median which in turn affects the error (absolute and mean) in any data set. When you plot the error you might get big deviations if outliers are in the data set.

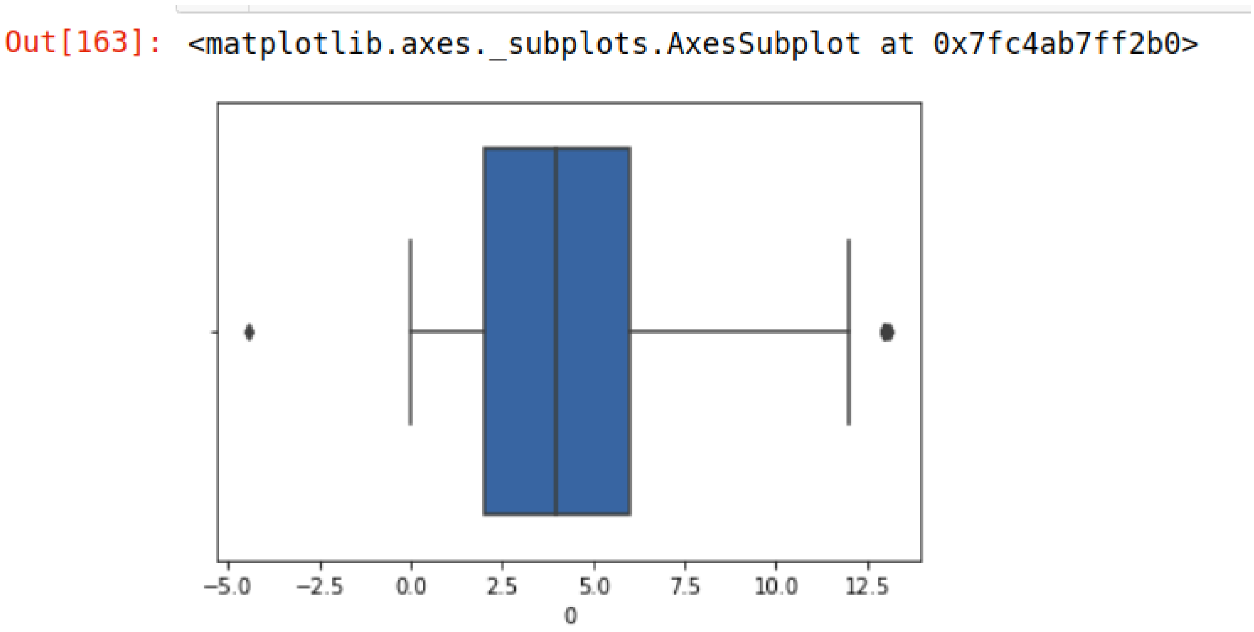
Boxplot is probably one of the most common type of graphic. It gives a nice summary of one or several numeric variables. The line that divides the box into 2 parts represents the median of the data. The end of the box shows the upper and lower quartiles. The extreme lines shows the highest and lowest value excluding outliers.



```
Out[161]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc4ab906860>
```

```
In [162]: dataset = dataset[(np.abs(stats.zscore(df)) < 3).all(axis=1)]
```

Fig. - With Outliers



```
Out[163]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc4ab7ff2b0>
```

Fig. - Without Outliers

The **Z-score** is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured.

The data points which are way too far from zero will be treated as the outliers. In most of the cases a threshold of 3 or -3 is used i.e if the Z-score value is greater than or less than 3 or -3 respectively, that data point will be identified as outliers.

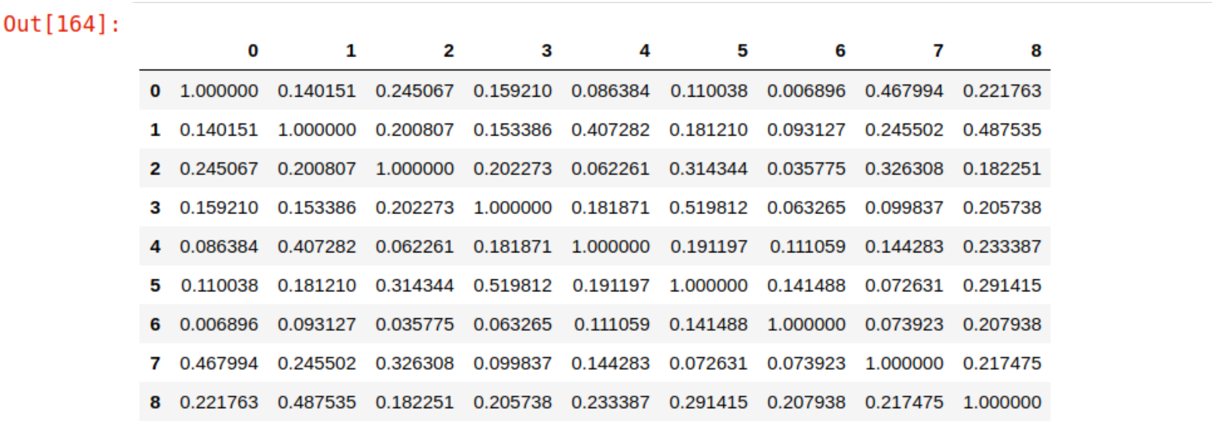
4) Feature Extraction

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process.

Finding Correlation

The correlation matrix is an important tool to understand the correlation between the different characteristics. The values range from -1 to 1 and the closer a value is to 1 the bettere correlation there is between two characteristics. Therefore, we conclude that there is -

- No obvious high correlation between independent variables.
- No obvious relationship between diastolic blood pressure and diabetes.
- No obvious relationship between age and diabetes.



Out[164]:

	0	1	2	3	4	5	6	7	8
0	1.000000	0.140151	0.245067	0.159210	0.086384	0.110038	0.006896	0.467994	0.221763
1	0.140151	1.000000	0.200807	0.153386	0.407282	0.181210	0.093127	0.245502	0.487535
2	0.245067	0.200807	1.000000	0.202273	0.062261	0.314344	0.035775	0.326308	0.182251
3	0.159210	0.153386	0.202273	1.000000	0.181871	0.519812	0.063265	0.099837	0.205738
4	0.086384	0.407282	0.062261	0.181871	1.000000	0.191197	0.111059	0.144283	0.233387
5	0.110038	0.181210	0.314344	0.519812	0.191197	1.000000	0.141488	0.072631	0.291415
6	0.006896	0.093127	0.035775	0.063265	0.111059	0.141488	1.000000	0.073923	0.207938
7	0.467994	0.245502	0.326308	0.099837	0.144283	0.072631	0.073923	1.000000	0.217475
8	0.221763	0.487535	0.182251	0.205738	0.233387	0.291415	0.207938	0.217475	1.000000

Fig. - Correlation Matrix

Feature Scaling

It is a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalise the data within a particular range.

MinMaxScaler rescales the data set such that all feature values are in the range [0, 1] as shown in the right panel below.

5) Model Building

Splitting the Data

Now that we have transformed the data we need to split the dataset in two parts: a training dataset and a test dataset. Splitting the dataset is a very important step for supervised machine learning models. Training set : Test set = 70:30.

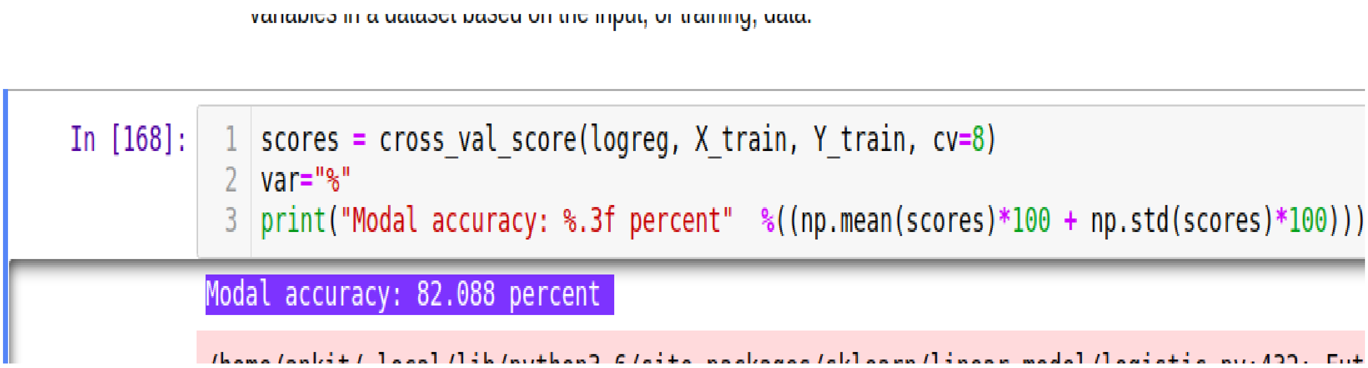
Logistic Regression

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Accuracy of the Model

Machine learning model accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training, data.

So, using the Logistic Regression Model, we were able to attain an overall model accuracy of 82.088 percent.



variables in a dataset based on the input, or training, data.

```
In [168]: scores = cross_val_score(logreg, X_train, Y_train, cv=8)
          var="%"
          print("Modal accuracy: %.3f percent"  %((np.mean(scores)*100 + np.std(scores)*100)))
```

```
Modal accuracy: 82.088 percent
```

```
/home/ankit/.local/lib/python3.6/site-packages/sklearn/linear_model/logistic.py:432: Fut
```

Fig. - Accuracy of the model