**Aim :** Predicting early onset of Diabetes with certain input Data with BI and Neural Network Algorithms

**Data Description:**

Pima Indian Diabetes Dataset from Kaggle

The datasets consist of several medical predictor (independent) variables and one target (dependent) variable, Outcome. Independent variables include

The number of pregnancies

BMI

Insulin level

Age

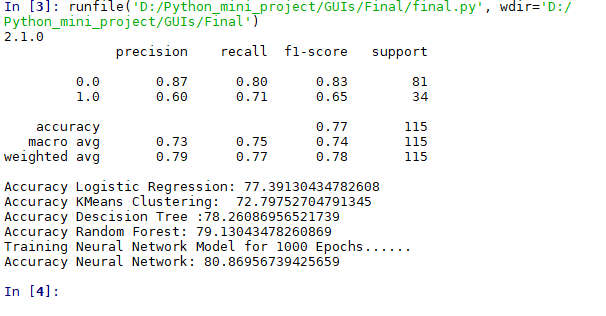
Glucose levels,

Blood Pressure

Diabetes Pedigree function ( a mathematical value calculated with historical diabetes data)

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

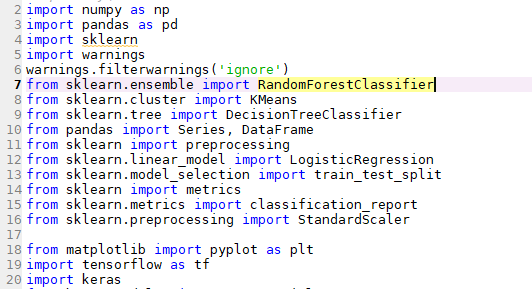
**Results :**

****

**Approach:**

1. **Dependencies**

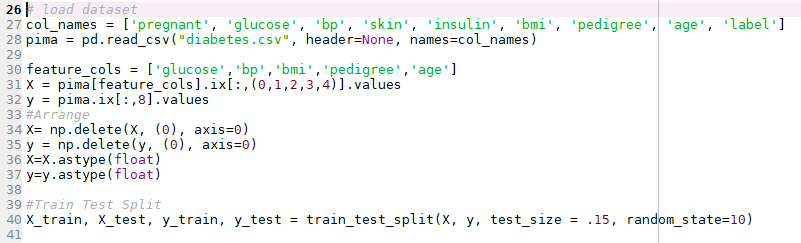
The Following dependencies need to be imported in python before starting

****

1. **Data Preprocessing and cleaning.**

A lot of the Missing Data values have been imputed and certain tables that are unnecessary towards the prediction have been truncated in the dataset.

Also the data has been split into training and test sets for validation

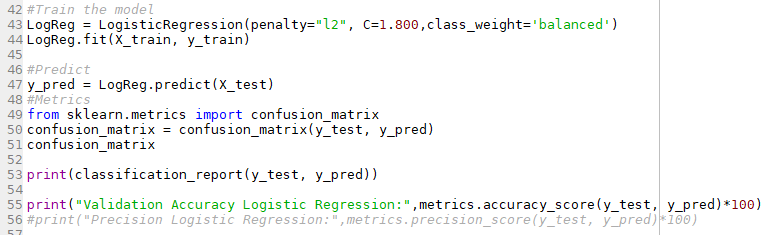


Our Model uses the following features for prediction

1. Glucose content in the blood
2. Blood Pressure
3. Body Mass Index
4. Diabetes pedigree function
5. Age

**3) Logistic Regression Model**

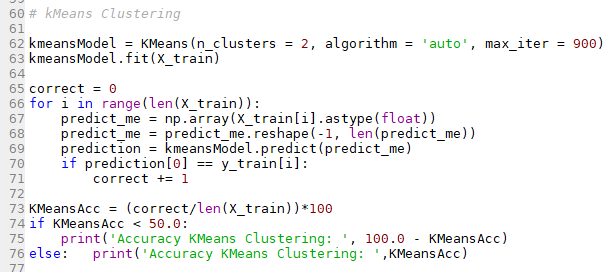
A logistic regression classifier was modeled with ‘l2’ penalty and balanced weights for prediction on the data



Validation Accuracy of **77.391%** was achieved on this model

**4) KMeans Clustering**

The KMeans Clustering Algorithm ran for 900 iterations with the number of clusters = 2 for the Boolean prediction values.



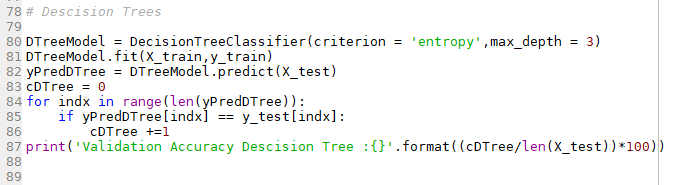
Accuracy of **72.797%** was achieved on this model.

Since KMeans in unsupervised the predictions were run directly on training data.

**5) Decision Trees**

A Decision Tree classifier with max depth 3 was modeled in python.

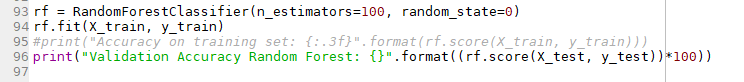
Validation Accuracy of **78.26%**  was achieved on this model.



**6) Random Forest Classifier**

The Random Forest Classifier was modeled with 100 estimators.

The second best score was achieved on this model with **79.13%**  validation accuracy.

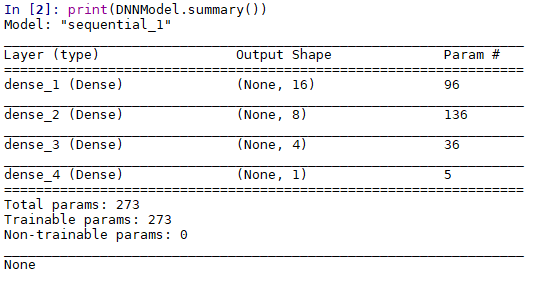


**7 ) Dense Neural Network.**

Custom DNN was modeled with Keras and Tensorflow 2.1 as backend.

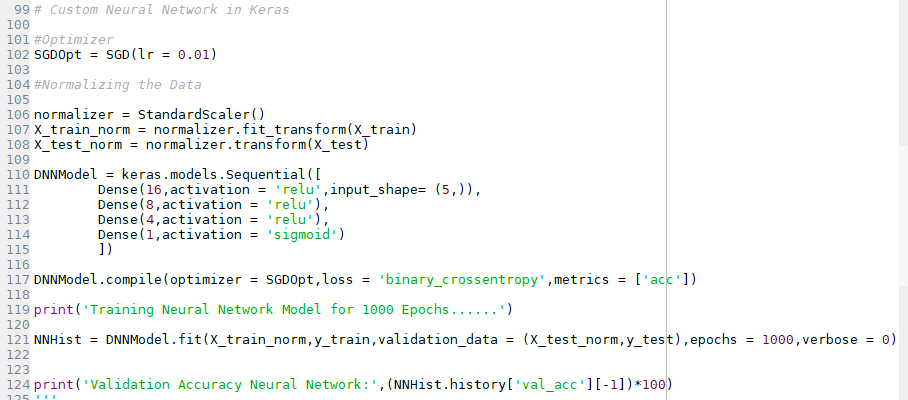
After testing a multitude of architectures, the following one was considered to show the highest validation Accuracy of **80.869%.**

1. For the Network Optimizer Stochastic Gradient Descent with a learning rate of 0.02 was chosen.
2. Data was normalized before feeding through the DNN.
3. Following is the model summary



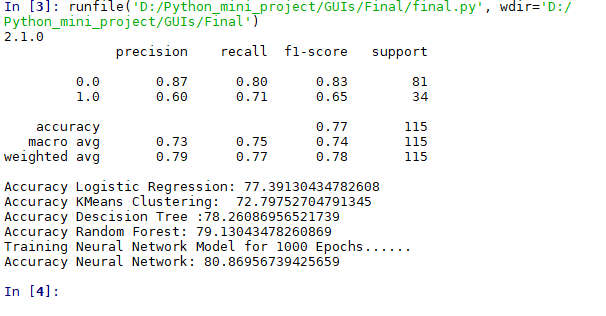
1. The Network was trained over 1000 Epochs.

The Code:



**8) Interpretations.**

With all the Classifiers Tested for Validation Accuracy. Here is a final result comparison



-The Highest Validation Accuracy was predicted on the custom DNN which was modeled after trial and error for different architectures.

-The Second Highest Validation Accuracy was Measured on the Random Forest classifier with accuracy very close to the DNN.

-KMeans Clustering achieved the Lowest accuracy as opposed to the other models.

**Conclusion:**

Hence we have successfully modeled various DMBI algorithms to predict data on the Pima Indian Diabetes dataset.

All of the tasks were handled in python with the Scikit Learn and TensorFlow Libraries.

The DNN and Random Forest Classifier models show promising results on the dataset.