Model training outcome for predicting obesity levels

**Abstract**

For training a model with higher accuracy in predicting obesity level of a person with given life habits, we applied different methods on the UCI obesity dataset. The results are promising. Both the Random Forest and Neural Network models scored the highest accuracy, 0.95. The other model scores are LDA:0.889, QDA:0.018, Logistic: 0.806, Decision Tree: 0.871, SVM:0.936 and K-Means Clustering: 0.142.

**Introduction**

Through the years, obesity has become one of the biggest and the most dangerous health problem among the society. Obesity can induce different diseases such as diabete,hypertension and various cerebro-cardiovascular diseases, which can cause serious results or even further, deaths.

To help with preventing the worst situations happening, our group decided to build a model can successfully predict the obesity level of a person based on the information of his or her eating habits and physical condition.

The dataset we are going to use to train the model is downloaded from UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+)>

This dataset contains 17 variables and 2111 observations. The variable related to obesity level is NObesity and it is categorized into 7 categories: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III.

The variables related to eating habits are: The attributes related with eating habits are: Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC).

The variables related to physical conditions are: Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS).

**Methods**

For training the model, we decided to use the following methods: Logistic Classification, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Support Vector Machine, Decision Tree, Random Forest, Neural Network and a K-Means Clustering method as a comparison to the previous supervised learning methods.

Both Logistic and SVM are designed originally for binary classification. In order to complete the job for multiclass classification, we need to use one-versus-rest method, which splits the multiclass classification process into multiple binary classification processes.

Both QDA and LDA are approximating the Bayes classifier for the boundary to separate different features. LDA provides a linear decision boundary while QDA provides a quadratic decision boundary.

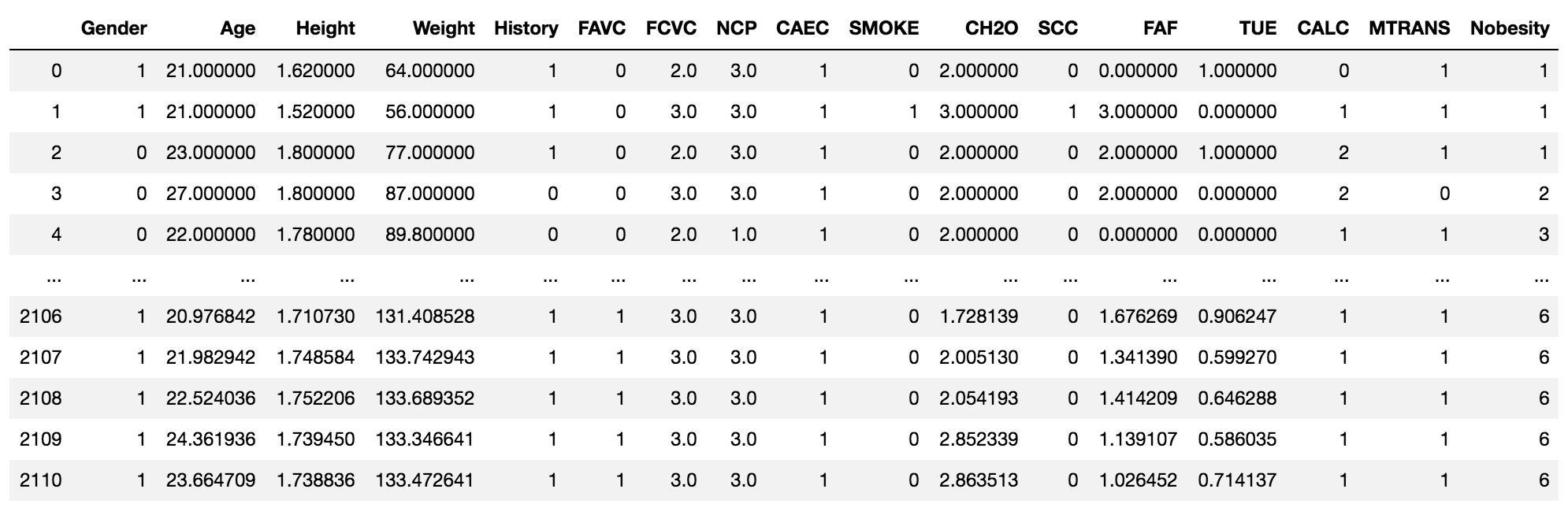
Decision Tree method uses a flowchart-like structure and makes separation divided predictor space to find out the result falling into this space with certain splitting algorithm. However, Decision Tree method is very easy to be overfitted. Therefore, Random Forest method is used. The randomness mechanisms in Random Forest method will help solve the overfitting problem.

Neural Network is a method that imitates the learning processes of human brains with hidden layers and neuron units.

K-Means Clustering is overall different from the previous method. It is an unsupervised machine learning process that separate observations into k clusters with the closest means.

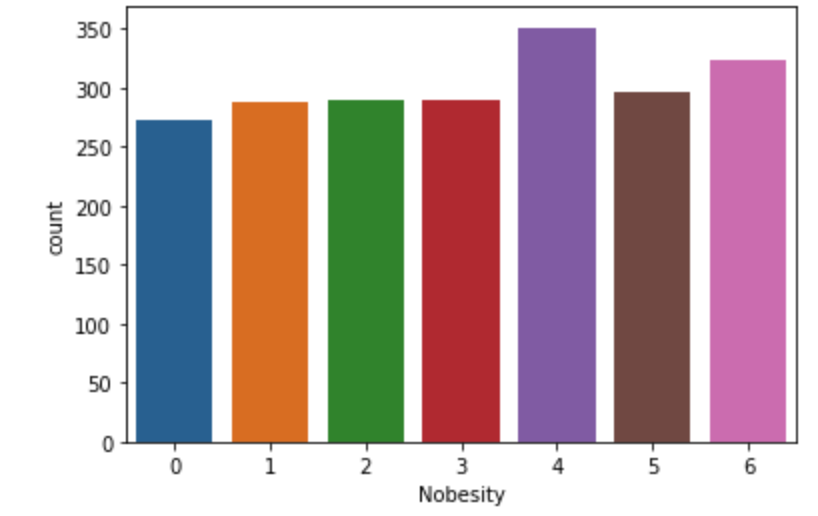
**Exploratory Data Analysis**

Most of the categorical variable in the dataset is recorded in words. In order to make it suitable for the machine learning process, we re-coded those into ordinal variables with numbers stand for each level. (Details will be provided in supplementary document)



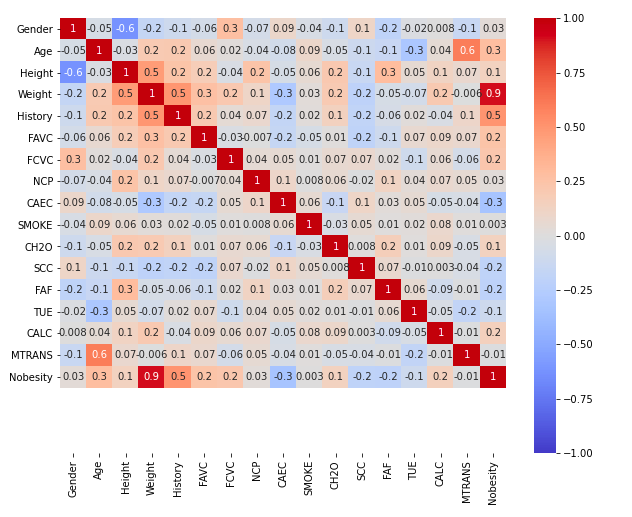
**Figure 1. Recoded Dataset**

After that, we checked the frequency of independent variable, which is the Nobesity in the dataset.



**Figure 2. Frequency of Independent Variable**

From the histogram we can tell that the data are overall balanced and none of them have extremely high or low frequency.



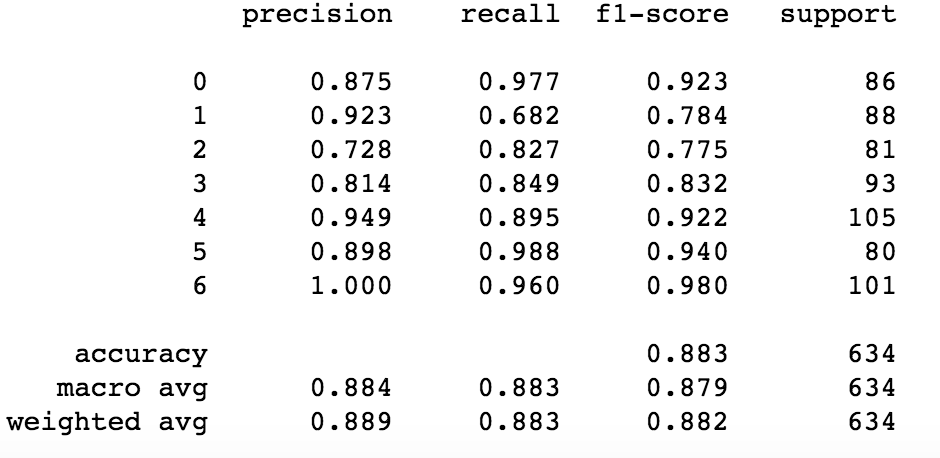
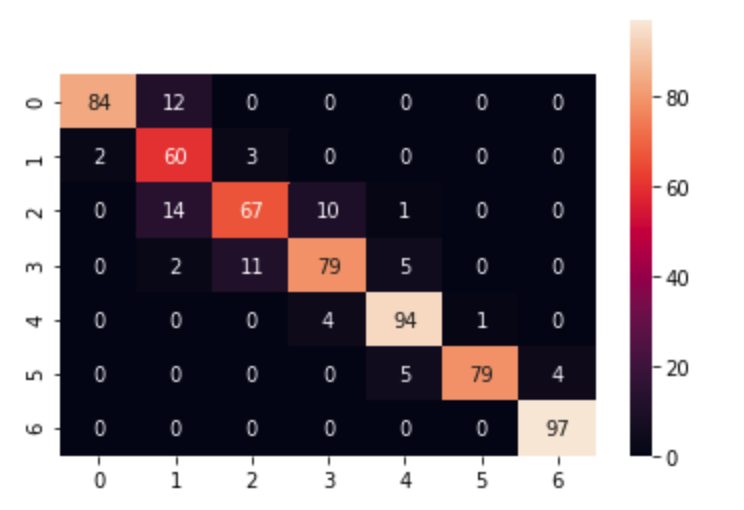
**Figure 3. Correlation Check**

We can find out that variable weight and Age are most correlated to the level of the obesity from the correlation heat map.

**Modeling**

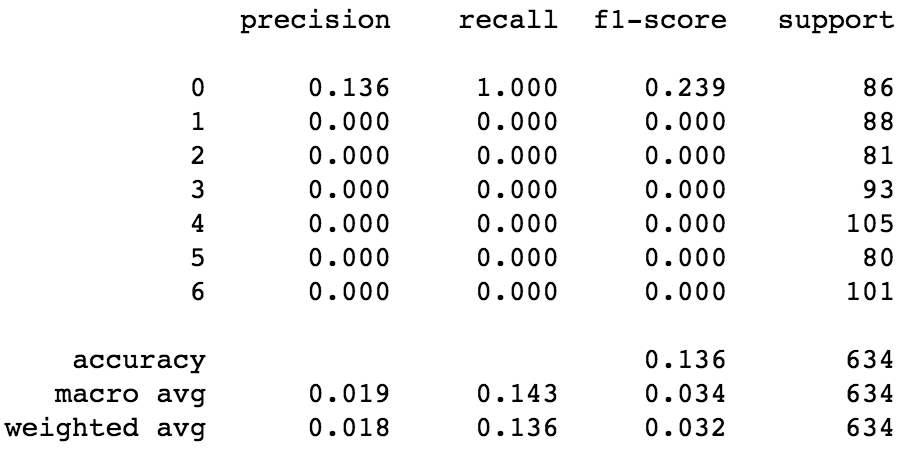
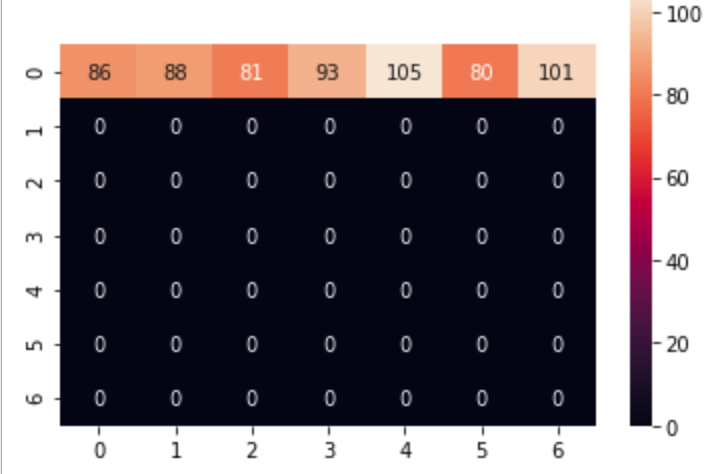
We start the modeling process with splitting the dataset into training and testing sets with a ratio of 7:3. Then we used several different methods with training set and test the model with testing set to find out a model with the highest accuracy.

**LDA**

****

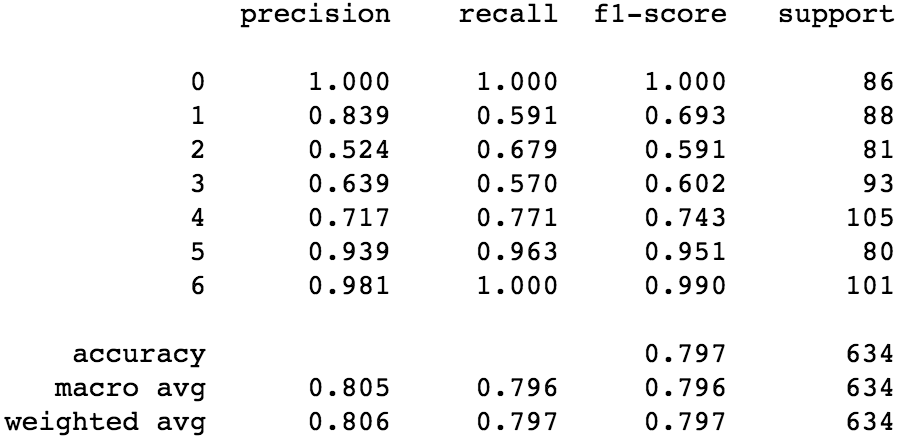
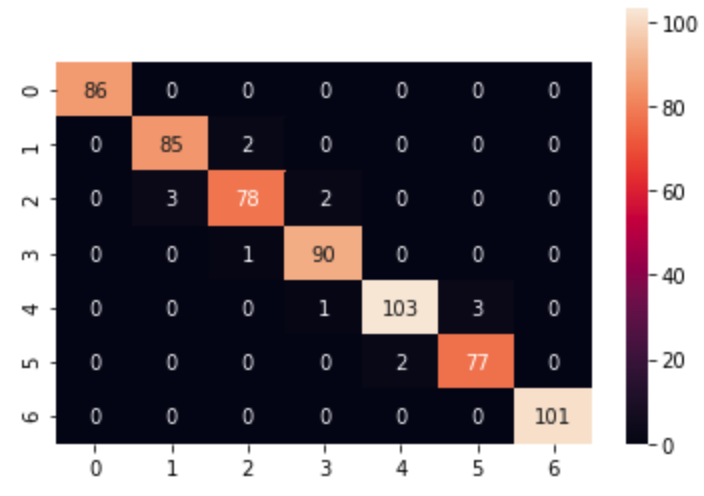
**Figure 4. Confusion Matrix and Accuracy Score for LDA model**

**QDA**

****

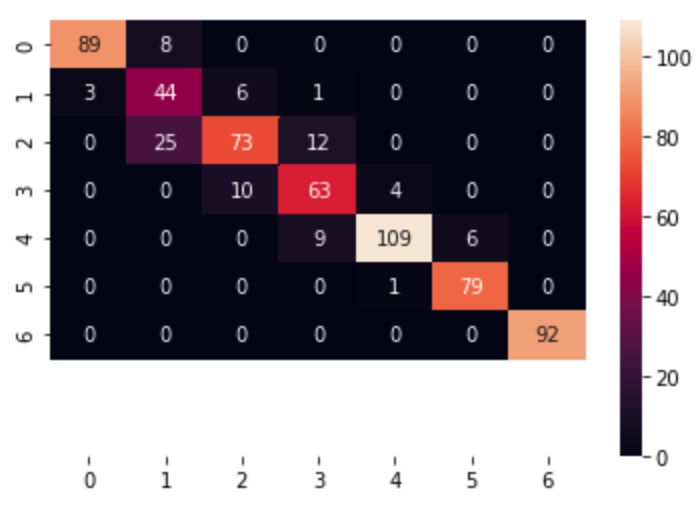
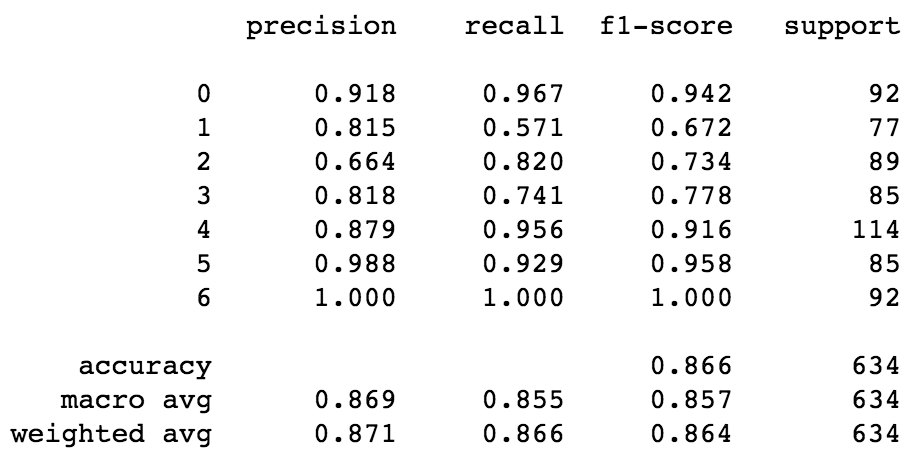
**Figure 5. Confusion Matrix and Accuracy Score for QDA model**

**Logistic**

****

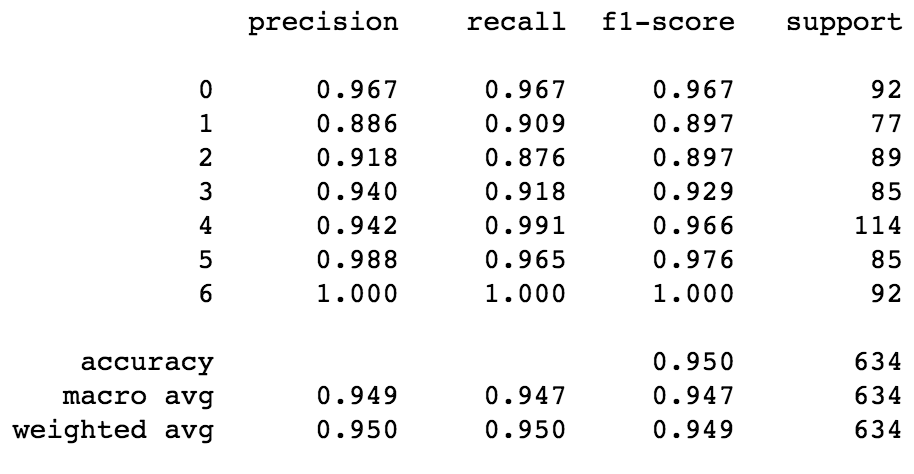
**Figure 6. Confusion Matrix and Accuracy Score for Logistic model**

**Decision Tree**

**** ****

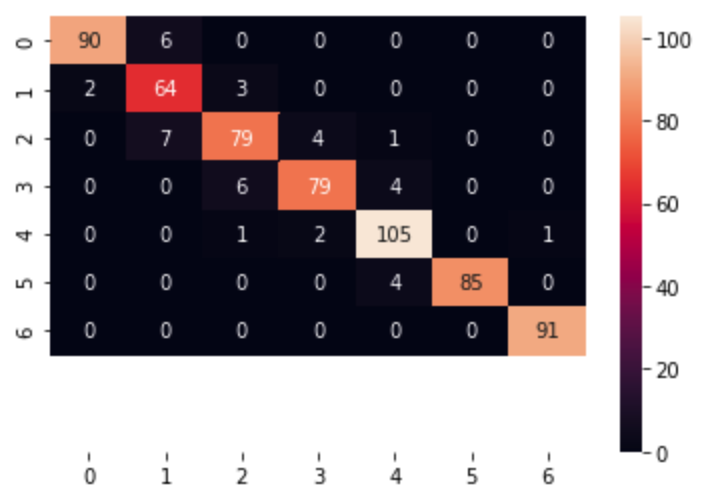
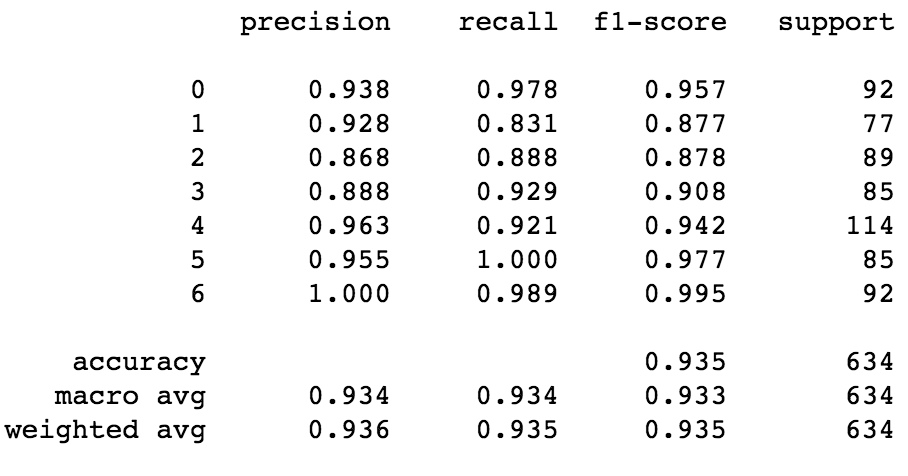
**Figure 7. Confusion Matrix and Accuracy Score for Decision Tree**

**Random Forest**

**** ****

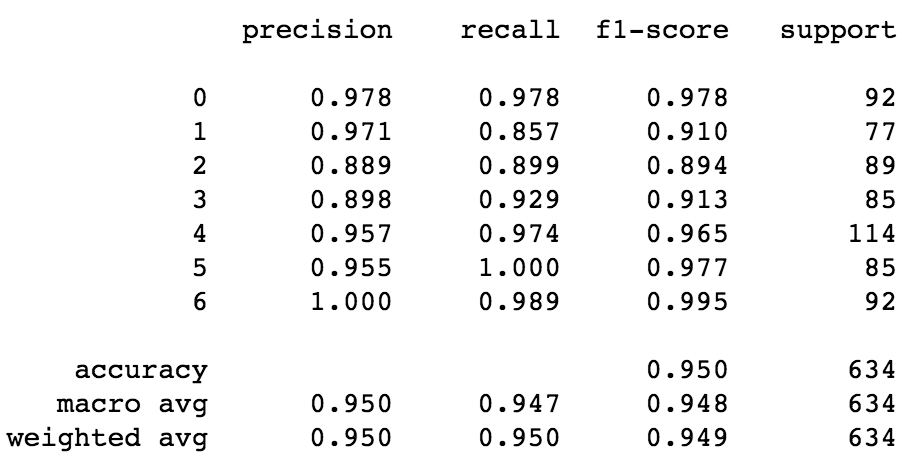
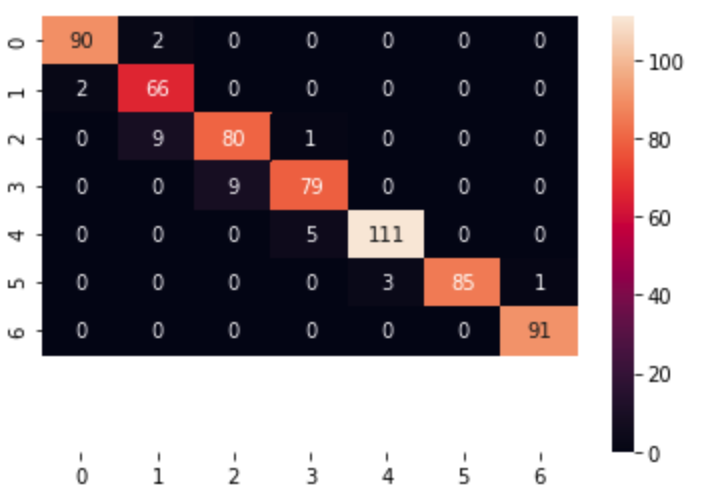
**Figure 7. Confusion Matrix and Accuracy Score for Random Forest Model**

**Support Vector Machine**

**** ****

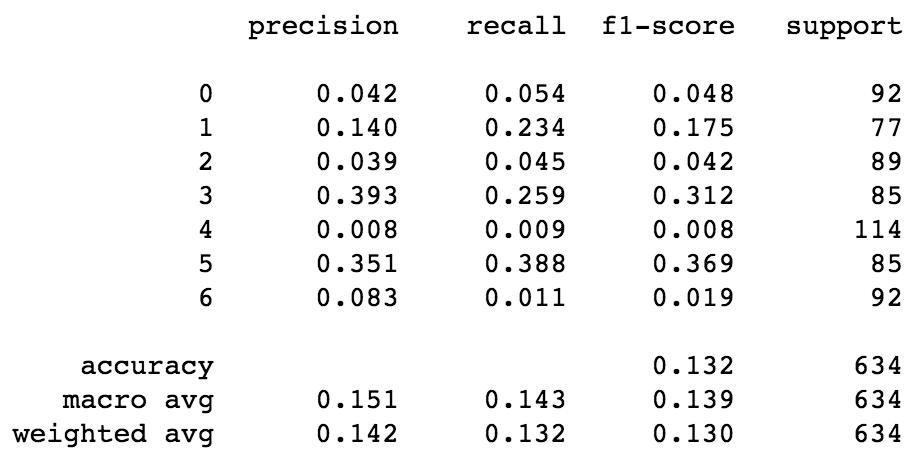
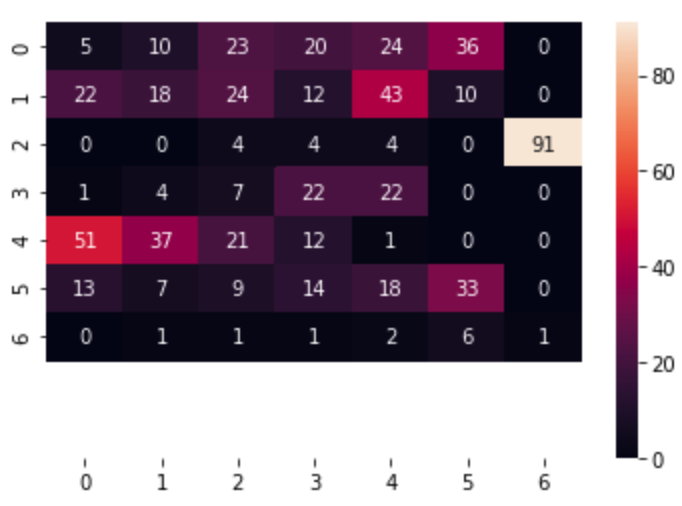
**Figure 7. Confusion Matrix and Accuracy Score for SVM**

**Neural Network**

****

**Figure 7. Confusion Matrix and Accuracy Score for Neural Network Model**

**Clustering (K-Means)**

****

**Figure 8. Confusion Matrix and Accuracy Score for K-Means Clustering**

**Discussion**

Among all the methods, it is not surprised to see that K-Means Clustering is not performing well since clustering algorithms are not designed for doing classification. Besides K-Means, QDA has also not performed ideally in classifying this dataset. One possible guess is that the projection boundary is not even close to the shape of quadratic.

For the SVM, the linear kernel performs much better than radial kernel, which also partially verified the hypothesis above about the boundary problem.

For Neural Network, 5 layers with 4 neuron units performs the best. However, the current knowledge of mine could not give a proper explanation about the reason for choosing this combination.