My task is to predict if the individual will pay more money for healthy food or not.The dataset I have to used is Kaggle Young people survey dataset.

I have chosen several ML techniques like Data preprocessing, Exploratory data analysis, Data Classification and finally prediction and accuracy as part of evaluation of my classification.
I chose to evaluate success using the score of a classification model.Then finally used evaluate method where appropriate so as to get RMSE of my prediction.I have used Scikit learn it is free and easy to use and provides many libraries to perform operations and also the parameters can be tuned with ease using functions in this library.Moreover,Scikit learn offers a elaborate documentation which makes it very easy for us to follow and implement.I have used several classifiers like ExtraTreesClassifier and Logistic Regression after selecting most valuable features based on F Score.Compared accuracy of each the best was SVM using pipeline.The parameters were tuned on validation dataset.

Algorithm

Here I will describe the major functions used to perform the analysis,details can be found in codes and the writeup/comments in notebook.After the data preprocessing and Exploratory data analysis I have done data modelling where I performed classification using RFE function of scikit learn library.Then after that I have used SelectKBest to select 50 features and then make another dataFrame on that I have used extraTreesClassifier so as to improve dataset.I have the respective score of estimator to judge accuracy.

Results

Using the above methods of Scikit learn and tuning parameters as per dataset.I have performed analysis on development data.I have achieved accuracy of about 68%.I have computed Area under ROC Curve and Log Loss .Basis of the fact that strongly correlated fact with the target set will predict values better.I had also carried out several visualization analysis to support my assumption of these strongly related values and I chose them and got accuracy and calculated RMSE.The decision boundary for both classifier is smooth and it well separates the data.More graphs and results analysis can be seen in the notebook.

GIthub Link:https://github.com/harsh543/MachineLearning_homework5