Homework 5 - Written Part

Problem Statement:
Using the Young People Survey dataset predict a person's "empathy" on a scale from 1 to 5

This is a multiclass classification problem with 5 classes. The Young People survey data has 1010 samples and 150 feature one being "Empathy" itself. Most of the columns have values for abstract quantities like how much someone like a movie or activity and some columns being values like height and weight.

I used SVC model to solve this problem. I had help from the cheat sheet available at http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html . This was a classification model which can handle the amount of data we had and works well with PCA.

Preprocessing:
I used Pandas to read the input CSV file. Initial step was to convert text values to numeric. For this i took the set of values for each text column and arranged them in a logical order. For example in column 'smoking', never smoking gets a value of 0 or 5 while 'always smoking' get 5 or 0 respectively. Also binary values like 'gender' get 0/1 values. Once that is done, the data is split 70/30 for training and testing, Validation data is split at the time of training using k_fold strategy. then i used sklearn imputer to fill the missing values. I used most_frequent strategy for this since it would be better for columns like 'right_handed'. Then the values are scaled using sklearn standard scaler which scales the data using standard deviation and centres the data by subtraction mean. Once all that is done, the data is exported and saved in a variable called PeopleData.

Solution/Experimental Setup:
Initially i thought that the no of features may be too high and thought how features like height influence empathy and might even be bad for the classification. But that was not for me to decide which features are important. I had tried both feature selection and feature extraction. For feature selection, i am using sklearn recursive feature elimination which eliminates some features. I found the link useful while trying to implement that. Then i did PCA on top of that. This article helped me in implementation. After that i did sklearn GridSearchCV over a set of parameters and found that rbf kernel with gamma value of 0.01 and C = 1 worked best for the data with the best accuracy score. The original data was initially split into 3 with 60:20:20 split. I used the validation data to determine the constants varance_threshold for PCA and no of features to select for recursive feature elimination. Once the values have been determined, i changed the split back to 70:30 since the validation data was not used since the GridSearchCV used k-fold validation with the training data to determine the best modal.

Final Score Evaluation
The model has 39.8% accuracy on test data. The baseline accuracy with dummy modal was 30%.