DS 203 : Programming for Data Science
Tutorial and Assignment Sheet – 7
Basic Machine Learning Practice

3. Predict the rating of a clothing items based on other variables for the data at URL https://www.kaggle.com/jmmvutu/summer-products-and-sales-in-ecommerce-wish, using the following general guidelines:
   a. Decide whether the variable to be predicted is discrete or continuous.
   b. Decide if this is a supervised or an unsupervised problem. For the former, find the target variable.
   c. Decide on a measure of performance, e.g. accuracy, area under ROC curve, F1-score, sensitivity, specificity, variance explained, Davies-Bouldin criteria, reconstruction error, root mean square error (RMSE), mean absolute error, RMSE normalized by standard deviation of the target variable, etc.
   d. Decide which variables might be relevant for predicting the target variable.
   e. Decide which variables are usable.
   f. Convert categorical variables into one-hot bit dummy variables, and standardize (or normalize) the continuous variables.
   g. If there are too many variables, then consider dimension reduction techniques.
   h. Consider ML frameworks that work with the type of problem and the input variables, and try to order them based on whether they are likely to succeed based on the number of variables and the number of samples. For example, some ML frameworks fare better with fewer samples and higher dimensions (e.g. LASSO regression), while others are scalable with more samples (e.g. neural networks, RF, and kernelized SVM). Pick two-three ML frameworks.
   i. Divide the data into training, validation, and testing subsets, roughly in 70:15:15 ratio.
   j. List hyper-parameters for the ML frameworks selected, and form hyper-parameter grids.
   k. Train the ML frameworks, and test their performance on validation subsets.
   l. Select the ML framework and hyperparameter combinations with the best validation performance, and test them on the test data.
   m. Comment whether the test results indicate if the model is usable

4. Repeat the exercise for predicting gestures based on muscle activity for the following dataset: https://www.kaggle.com/kyr7plus/emg-4.

5. Compress the 64 input dimensions in the same dataset https://www.kaggle.com/kyr7plus/emg-4 to an appropriate number of dimensions using PCA such that the RMSE reconstruction error is less 1% of the standard deviation of the L2 norm of the 64 variable input. Plot a graph of dimensions retained versus normalized RMSE.