CS 412 Introduction to Machine Learning, Spring 2018
University of Illinois at Chicago
Homework 5: Mini-Project (Write-Up)

Dataset used: Young People Survey, task 2: understanding how likely a person is to spend money on healthy, quality and good food.

(a) Approach to task:

- Understand the dataset to better facilitate selection and application of models by plotting the various features given in the data
- Clean and pre-preprocess the data:
    - Split the data into X and Y
    - Train data: 70%, Validation data: 10%, Test data: 20%
    - Found 2 missing values in the target label column, discarded the two rows without Y values
    - Look for NaN in the columns and replace with the mean value and mode value for numerical and categorical data types respectively
    - On plotting the distribution of data based on the class label, I found the class label with score 1 to have a low representation, hence I resampled the data and added additional rows with class label 1
    - Encoded the categorical data types using LabelEncoder module, representing all the string labels as numerical encoded values
- Logistic Regression model was applied on the dataset with no hyperparameter tuning or feature selection and this model has been used as the baseline for the mini-project
- Performed feature correlation (each of the available 149 X fields was correlated with the target Y value). The top 30 correlated features were selected for subsequent models (SVM, Logistic Regression, Random Forest)
- SVM seems to work best for this dataset, with a penalty value of 2 and kernel as RBF.
- I also used Recursive Feature Elimination (RFE) method to select and use the most suitable features. Applied these features to the Logistic Regression model. However, the performance of this model was not as good as SVM with correlated features.
- Credits to Sklearn, Statistics, various ML libraries

(b) I choose the SVM with RBF kernel to model this dataset. SVM works well with numerical data and its performance is not impacted by the number of data points present in the dataset. With the kernel trick, the data is transformed, allowing SVM to find a more optimal boundary. SVM also gave me stable results over multiple runs.

(c) Accuracy is my primary source of evaluation. The F1 score calculated from precision and recall is also used to understand how each class us represented.

(d) This homework uses Python, Jupyter Notebook and the vast Python library for machine learning.

(e) Below table provides an overview of performances of the few models tried.

Dataset	Baseline Model Logistic Regression	Proposed Model SVM with RBF (Correlated Features)	Random Forest with Correlated Features	Logistic Regression with Correlated Features	Logistic Regression with RFE
Validation Data Accuracy	32.18%	45.97%	44.82%	27.58%	33.33%
Test Data Accuracy	31.48%	46.29%	42.05%	38.42%	36.11%

(f) Tuning of hyperparameters is being done on the validation data split. In case of points being misclassified, I would want to re-look at the hyperparameters to see which one produces a better result. For example, row index 49 of the dataset led to the prediction with score 4 while the true value is 2. Similarly, index 752 predicts with the given features a score of 3 while the true value is 2.