Homework 5

Problem:

You are working for a non-profit that is recruiting student volunteers to help with Alzheimer's patients. You have been tasked with predicting how suitable a person is for this task by predicting how empathetic he or she is. Using the Young People Survey dataset (https://www.kaggle.com/miroslavsabo/young-people-survey/), predict a person's "empathy" on a scale from 1 to 5. You can use any of the other attributes in the dataset to make this prediction.
I have chosen the sklearn python library to solve this problem as it provides all the necessary functions for machine learning.

Preprocessing:

Since sklearn library requires all features to be numerical, I have converted the values of categorical features to numerical values. I have also replaced the missing values in the data with the most frequent value appearing in the corresponding column to make better predictions. Having irrelevant features can decrease the accuracy of the model and increase training time. So, I have chosen the best 100 features from the data using univariate selection.

Environment Setup:

I have separated the column 'Empathy' as the value to be predicted i.e. Y and used the remaining columns as the input X. I have split the data into 80% training data and 20% testing data. Then I split the training data into 80% training and 20% development data and used 10-fold cross validation for the training data on some basic models to select the best model. I have used 'accuracy' as the measure of success.

Evaluation:

I have selected Logistic Regression as my model as it gave the highest accuracy. Then I have tuned the hyperparameters for Logistic Regression on the development data using GridSearch and got an accuracy of 95%.
I have used these hyperparameters to predict the value of 'Empathy' on the test data and got an accuracy of approximately 91%.

Credits:

http://scikit-learn.org/stable/modules/classes.html#module-sklearn.base
http://scikit-hhlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html#sklearn.lineli_model.LogisticRegression
http://scikit-learn.org/stable/modules/grid_search.html
https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/
https://towardsdatascience.com/working-with-missing-data-in-machine-learning-9c0a430df4ce