Homework 3
COMP 379
Brian Nguyen
10/24/2021

This program is meant to implement a Linear Regression (LR), K-Nearest Neighbors (KNN), and Scikit-learn's (SKL) Dummy Classifier (DC) models to solve a classification problem in the form of non-linearly separable dataset, comparing their performances (by accuracies and macro F1 scores) in analyzing training datasets and predicting upon development and testing datasets.

The chosen dataset, *wine.data*, was preprocessed by randomly splitting it into three pairs of subsets, a training (70%), development (15%), and testing (15%) subset. These subsets were then standardized (after splitting to prevent data leakage) to ensure all features contribute equally to the models' fitting and learning functions. The chosen metrics to evaluate the prediction performance of the models included their accuracy rates and macro F1 scores to accommodate for potential imbalanced class distributions affecting the learning functions of the models.

To implement and test SKL's default LR model, it was fitted using the standardized training subset and evaluated on the standardized development subset. With a default C-value of 0.1, the model yielded a prediction accuracy rate of 96.30% and macro F1 score of 95.85%. After tweaking the LR model's C-value across a range from 0.00001 to 10000 and comparing their classification reports and confusion matrices, it was found that C-values of 0.1 and 0.01 yielded prediction accuracy rates and macro F1 scores of 100% on the standardized development subset.

To implement and test the KNN model I cooked up in my own kitchen, it was fitted using the standardized training subset and evaluated on the standardized development subset. After tweaking the KNN model's k-value across a range from 1 to 125, it was found that several k-values yielded prediction accuracy rates and macro F1 scores of 100% on the standardized development subset before accuracy rates began to decrease drastically after a k-value of 81 (see *Fig. 1*). This KNN model was adjusted to utilize a k-value of 23, the first k-value occurrence with perfect metrics.

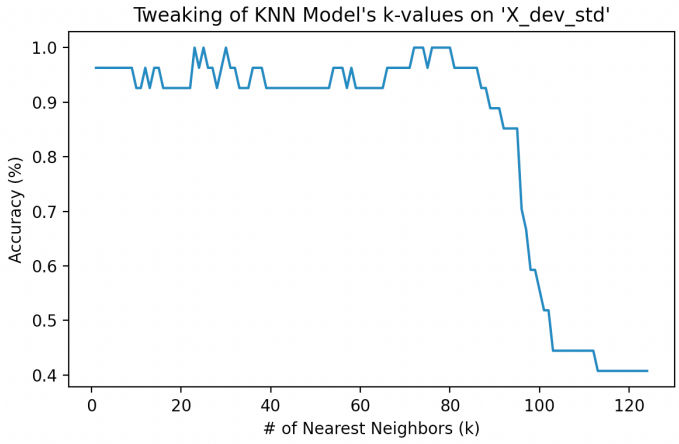

Fig. 1: Tweaking of KNN Model's k-values on Standardized Development Subset

To implement and test SKL's DC model with stratified, most frequent, prior, and uniform strategies, it was fitted using the standardized training subset and evaluated on the standardized development subset. The stratified strategy yielded an accuracy rate of 37.04% and macro F1 score of 33.33%. Both the most frequent and prior strategies yielded accuracy rates of 40.74% (highest of all DC models) and macro F1 scores of 19.30% (lowest of all DC models). The uniform strategy yielded an accuracy rate of 33.33% and macro F1 score of 32.75%.

To compare the prediction performances of the three models, they were all fitted using the standardized training subset and evaluated on the standardized testing subset. It was found that both the LR model with a C-value of 0.1 and KNN model with a k-value of 23 yielded accuracy rates and macro F1 scores of 100%. The DC model with a stratified strategy yielded an accuracy rate of 40.74% (highest of all DC models) and macro F1 score of 37.84%. Both the most frequent and prior strategies yielded accuracy rates of 37.04% and macro F1 scores of 18.02% (lowest of all DC models). Finally, the uniform strategy yielded an accuracy rate of 22.22% and macro F1 score of 22.07%.

Classifier Prediction Performances on Standardized Testing Subset		
Classifier	**Accuracy Rate**	**Macro F1 Score**
LR w/ C = 0.1	100	100
KNN w/ K = 23	100	100
DC w/ Strategy…		
Stratified	40.74	37.84
Most Frequent	37.04	18.02
Prior	37.04	18.02
Uniform	22.22	22.07

Fig. 2: Classifier Prediction Performances on Standardized Testing Subset