CMSC6950 - Comp Based Tools and Applications

Fatema Yeasmin Chowdhury
Student Number: 202193356
MUN email: fychowdhury@mun.ca

Project Title: Bangladeshi Paddy Trend Analysis in Coastal Areas

The main purpose of this project is to predict the best suitable crop type given the weather condition and find out the best model by comparing two classifiers based on data scaling techniques and feature selection techniques.

The dataset I am using on this project is collected from Nasa power view and Bangladesh Agricultural Book. According to the task, I have created a separate python script called CMSC.py, where I have built a total of four classes for plotting, data scaling, outlier removing and running models. These four classes have their own functionalities to perform different tasks. In addition I have also defined three functions namely recursive_feature_elimintation, feautre_importance, and PSO_feature_slection_with_result.

The dataset I have used had six features named Temperature, Humidity, Precipitation, Wind speed, Moisture, and Crops. Among them, the first five are independent features and the feature named Crops is the dependent one.
We will try to build a model which can predict the types of crops to firm given the values of five independent features.

In our dataset, there are three types of Crops named Aus, Aman, and Boro.

At first, I checked if there were any null values in the dataset. Luckily, I found none. Then I have boxplotted the independent features to see if there were outliers. I have found some outliers and removed them using the Interquartile Range Method. I have decided to use two models named Support Vector Machine and Random Forest. I have trained and tested these two models with the raw data and recorded the accuracy as baseline accuracy. Then I have performed two separate data scaling techniques on the independent features named Min Max and Standard Scaling. Then I again trained and tested the models with these two types of scaled data and recorded the accuracy to compare with the baseline accuracy. Following that, I wanted to check if all the independent features were needed to achieve high accuracy. I have performed three types of features selection. First one was Recursive Feature Elimination. I performed this feature selection technique on both types of scaled data. Then I trained and tested the models using the features selected by Recursive Feature Elimination and recorded the accuracies. Then I performed feature selection using the feature importance attribute which can be found from Random Forest Model and Support Vector Machine when they are trained using some data. The technique is that I need to train the Random Forest and Support Vector Machine using the whole dataset. The models assign some importance to the independent features by observing their contribution to model training. We can see the importance and select some

features based on the importance given by the models. In this way, I have selected some features for both types of scaled data. Using those selected features, I trained and tested the models and recorded the accuracy. Finally, I have performed feature selection using a different technique called particle swarm optimization. It performs a heuristic search to select the optimal feature subset. I performed particle swarm optimization for feature selection for both types of scaled data and recorded the accuracies. It should be noted that I have performed 5-fold repeated cross-validation for every train-test scenario I mentioned. The accuracies I recorded can be seen in the following table. It can be seen that the baseline Random Forest model performed better than other variants which mean the dataset is good enough to work with no scaling and feature selection. However, I received a similar performance from the random forest when the data was scaled using a standard scaler. In addition, I also saw a similar performance using the features selected by PSO for standard scaled data when the Random Forest Model was used. So to conclude, I found Random Forest to perform better than SVM in my exploration of the dataset I used. Moreover, my exploration says that the raw data is good enough to go for the task but if we want to reduce computational complexity by removing 1 or 2 features we can use the Particle Swarm Optimization technique for feature selection.

In the following table, we can see the model and the accuracy

	Model_variant	Accuracy
0	Baseline_SVC	85.148988
1	Baseline_RF	93.377049
2	MinMax_Scaled_SVC	90.457512
3	MinMax_Scaled_rf	93.046761
4	Standard_Scaled_SVC	91.399057
5	Standard_Scaled_rf	93.046761
6	RFE_Features_MinMax_Scaled_SVC	91.143334
7	RFE_Features_MinMax_Scaled_RF	91.634650
8	RFE_Features_Standard_Scaled_SVC	91.143334
9	RFE_Features_Standard_Scaled_RF	91.656498
10	FeatureImportance_Features_Standard_Scaled_SVC	91.057177
11	FeatureImportance_Features_S	92.062687

	tandard_Scaled_RF	
12	FeatureImportance_Features_MinMax_Scaled_SVC	90.051873
13	FeatureImportance_Features_MinMax_Scaled_RF	92.190961
14	PSO_Features_MinMax_Scaled_RF	92.982590
15	PSO_Features_MinMax_Scaled_SVC	91.635337
16	PSO_Features_Std_Scaled_RF	93.260917
17	PSO_Features_Std_Scaled_SVC	91.827782