Assignment 2

Vathana Him

November 1, 2021

1 Abstract

The purpose of this assignment is to use the data of the image created from assignment 1 to build classification models in order to classify the choosen images. This assignment utilized images from UCI respository as sample datasets that will be used to train a machine learning model. Images that were processed represented three fruits spanish pear, fuji apple, watermelon. These images were labeled as Image0, Image1, and Image2 respectively. These labels was then encoded to take in the values of 0, 1, and 2. Two machine learning was used in order to classify these images. The SGDClassifier with an Elastic Net penalty was used to classify a two-class dataset and Random Forest classifier was used to classify a multi-class dataset. Prior to training the machine learning model, additional methods were taken into account in feature selection and data scaling in order to reduce the size of the data and train the model with a sufficient outcome. Finally, the results between the Elastic Net and Random Forest model was examined.

2 Task 1

In order to begin working on this project, there were python libraries that needed to be installed. Thus, the libraries that were used were pandas, sklearn, matplolib, and seaborn. The data for the images that were choosen was first divided into a test and train split. This test and train split divided the dataset for each respective images into a 80:20 ratio of train-test split. Then a histogram for each training and test sets were generated for the nonoverlapping and non overlapping images.

Seaborn was primarly used for plotting histogram and scatterplot. The plot was created with a helper function created in the code that requires the input of a dataframe and two columns name in order to generate a plot.

The two features that were choosen for the histogram plots were features 54 and 56. In the non-overlapping images01, it is evident that features 54 and features 56 follows the same left skewed distribution based on 2.1 Figure 1. Based on 2.1 Figure 2, the test set of the non-overlapping images01, the distribution follow a similar left skewed distribution pattern. The mean for both the test and train set for non-overlapping images01 of features 54 and 56 were both 184 and 183 respectively. Thus, there was not a large deviation of mean from feature 54 and 56 from the train and test set of non-overlapping images01. Similary features 54 and 56 were also choosen to plot histograms for the non-overlapping dataset of images012. Based on 2.1 Figure 3 and Figure 4, it can be concluded that the test and train set followed the same left skewed distribution pattern. The mean and variances of the respective train and test set did no vary to a large degree as their mean was 173 for all and their variances ranging from 1970 to 1980. This suggests that the testing and splitting method for the non-overlapping images provided a good randomeses among the batches of dataset, which would help against any biases when training a machine learning model.

Subsequently, the two features that were choosen for the overlapping images01 dataset were also features 54 and 56. Based on 2.1 Figure 5 and Figure 6. The train and test set followed a left skewed distribution pattern. Their respective mean was both 184 and their variance was around 1500 and 1450 respectively. This suggests that both dataset had similar characteristics, and the train-test split didn't alter the date in a way that would cause the train-test split to largely deviate from each other. Finally, 2.1 Figure 7 and 8 showed the test and train set of feature 54 and 56 in the images012 dataset. Similary, these histograms follow the same left skewed distribution patter. Their mean did not deviation from each other largely as both feature had a mean of 172 for the train set and a mean of 174 for the test set. Their variance also differed in a small quanity with their respective variance being in the 1900s and 2000s. This again suggests that the train and test split preserved the integrity of the dataset, while also providing a degree of randomness suitable to training a machine learning model.

2.1 Figures Histogram

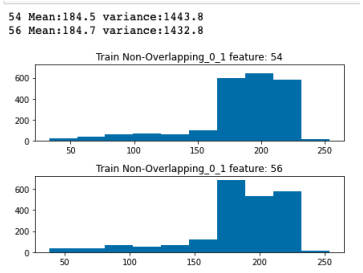


Figure 1: Image01 Non-Overlapping Train

Figure 2: Image01 Non-Overlapping Test

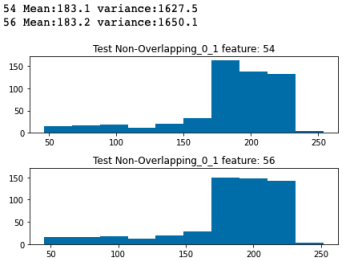


Figure 3: Image012 Non-Overlapping Train

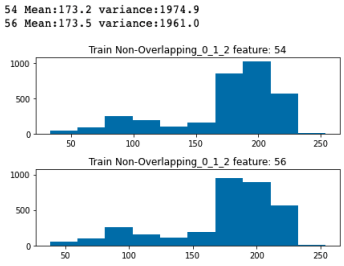


Figure 4: Image012 Non-Overlapping Test

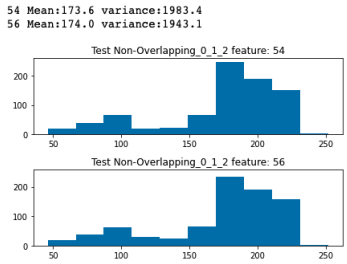


Figure 5: Image01 Overlapping Train

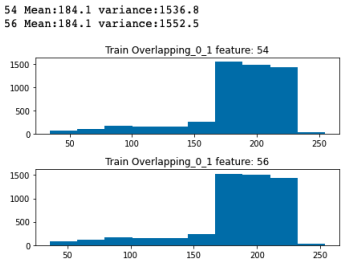


Figure 6: Image01 Overlapping Test

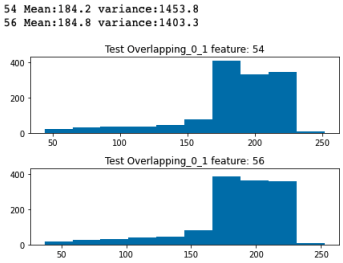


Figure 7: Image012 Overlapping Train

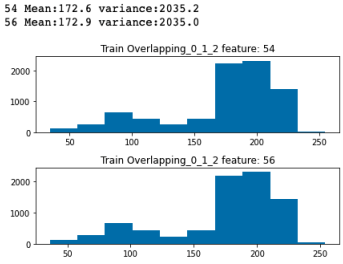
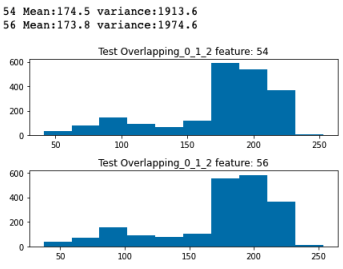


Figure 8: Image012 Overlapping Test



The scatterplot in 2.2 Figure 9 and Figure 10, represents the relationship of the black and white RGB color between features 54 and 56 of the train and test split dataset of the non-overlapping image01 dataset. Both features in the train and test split follow a similar positively correlated pattern. The data was also randomly distributed across all range of the black and white color spectrum. This suggests a great degree of randomness among the dataset. The non-overlapping image012 dataset in 2.2 Figure 11 and Figure 12 follow a positively linear pattern but deviated slightly from their test and train set. In 2.2 Figure 11, image 2 had more of a cluster around the darker RGB values and image 0 and 1. This suggests that image 2 may contain a darker texture than that of image 0 and 1. This same pattern was also found in the test set.

In the overlapping dataset, 2.2 Figure 13 and Figure 14 of image01 showed that the relationship between feature 54 and 56 follow a positively linear pattern. The same pattern can be found in both the train and test set. Thus, the method of splitting the data into train and test did not destroy the original characteristics of the dataset and its integrity. Similarly, the overlapping plot of image012 in 2.2 Figure 15 and Figure 16, followed the same positively linear relationship between features 54 and 56. It also displayed similar characteristics to the non-overlapping dataset of image012 as image 2 had a large cluster on the darker RGB values than that of image 0 and 1 due to its darker texture. This behavior was consistent through the non-overlapping and overlapping split, therefore, this suggests that the dataset was not altered to a large degree during its split.

2.2 Figures Scatterplot

Figure 9: Image01 Train

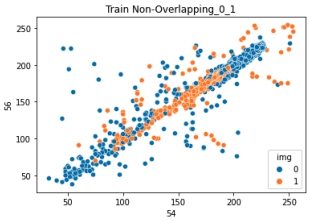


Figure 10: Image01 Test

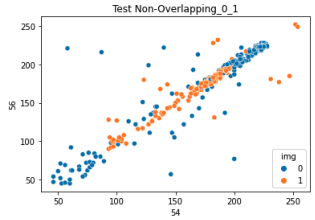


Figure 11: Image012 Train

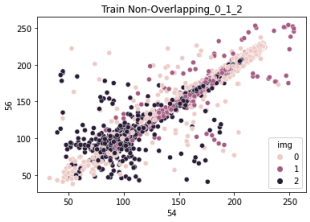


Figure 12: Image012 Test

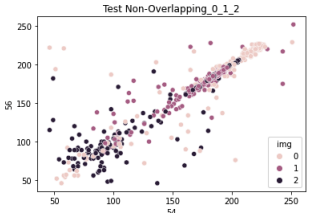


Figure 13: Image01 Train

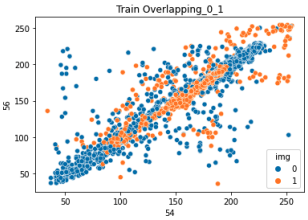


Figure 14: Image01 Test

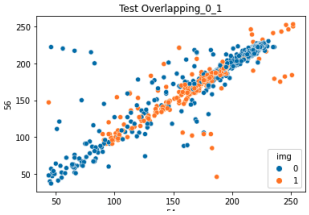


Figure 15: Image012 Train

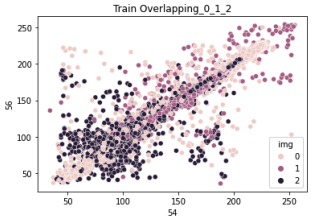


Figure 16: Image012 Test