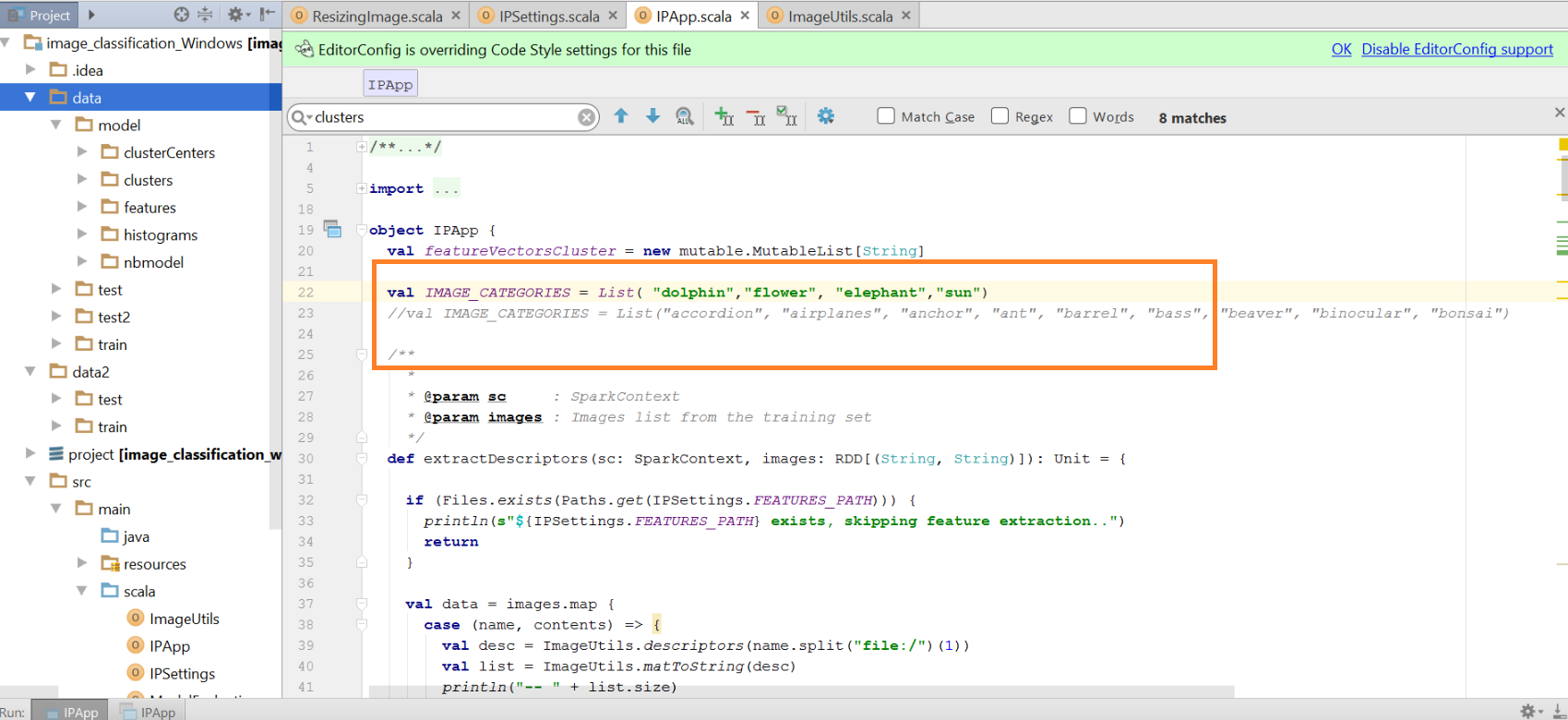
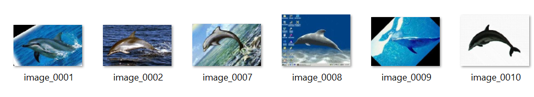
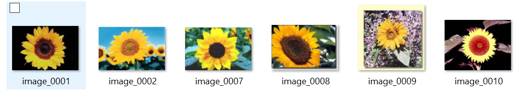
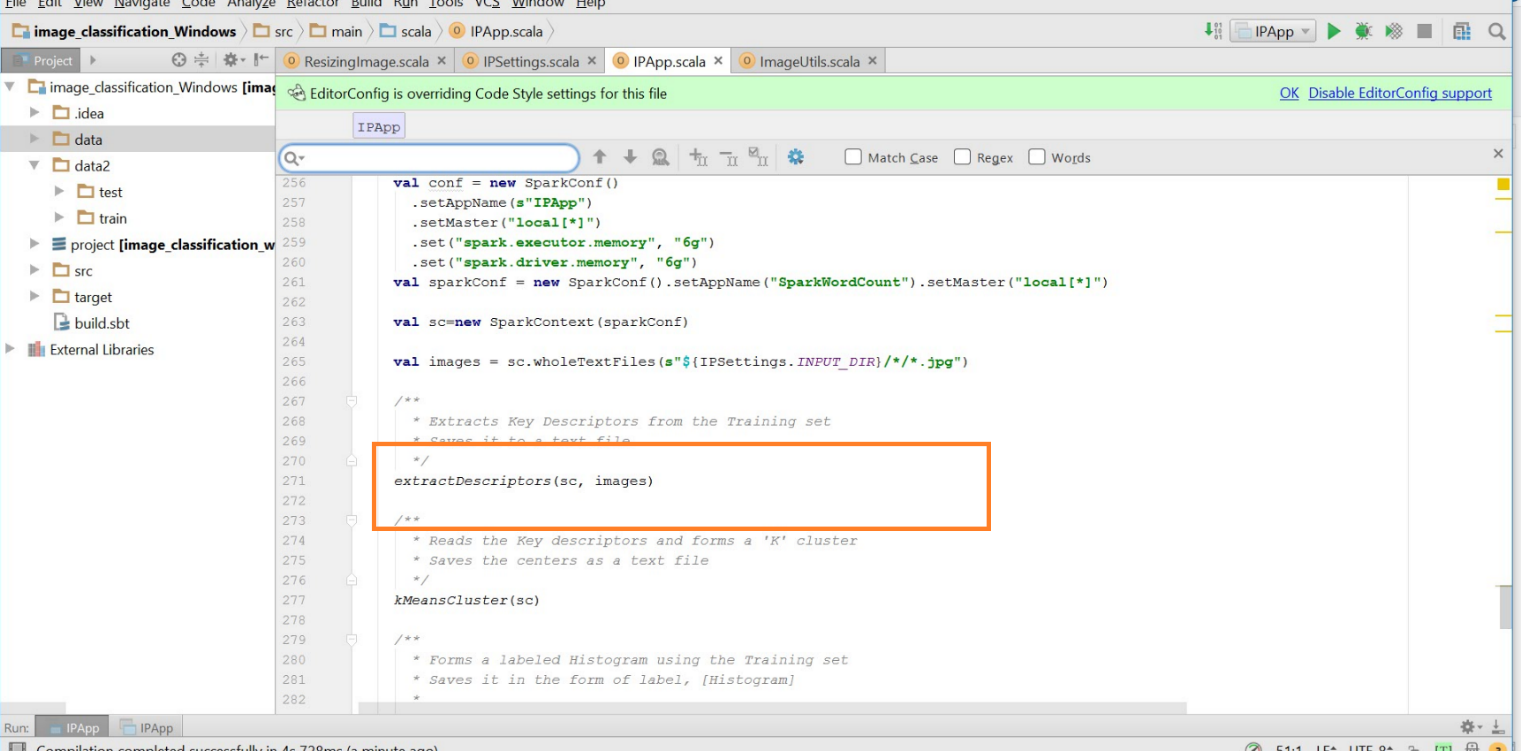
**CS5542 – Big Data Analytics and Apps**

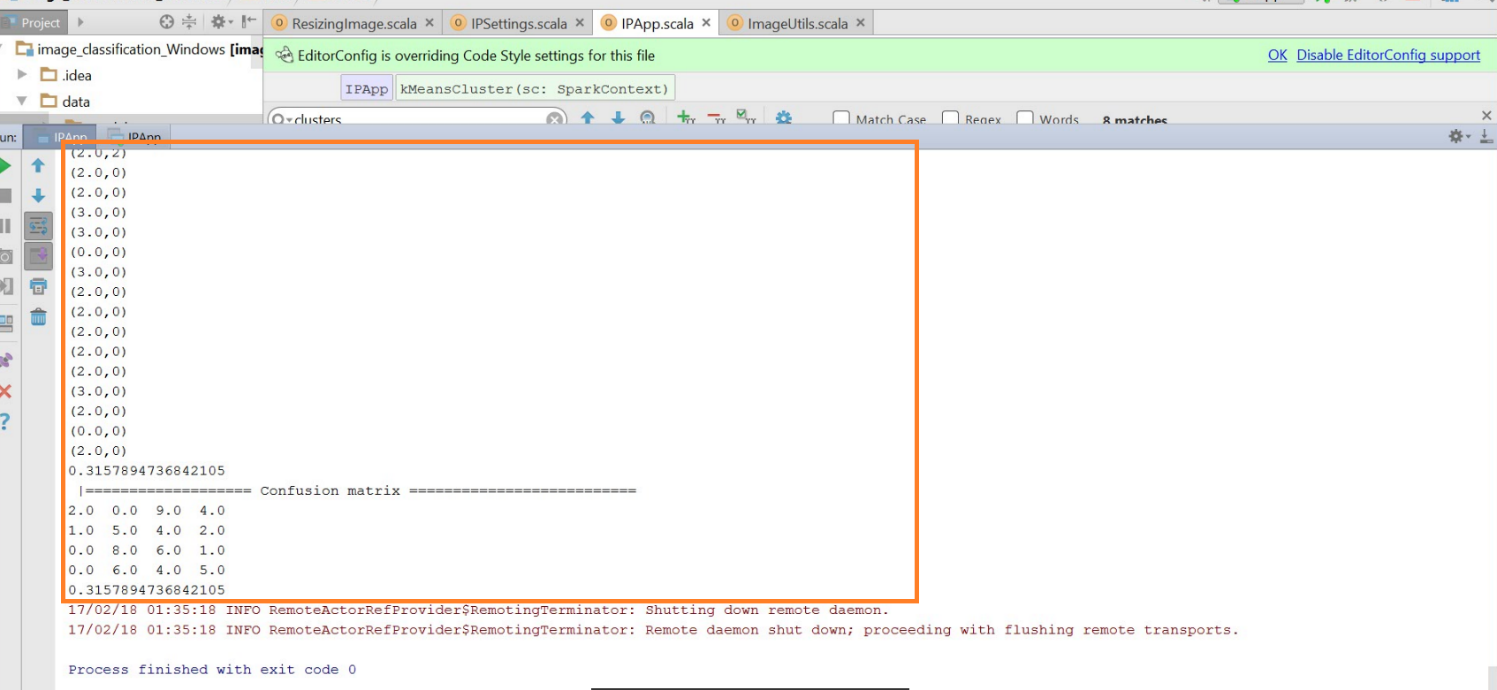
**Lab 4 – Assignment Submission**

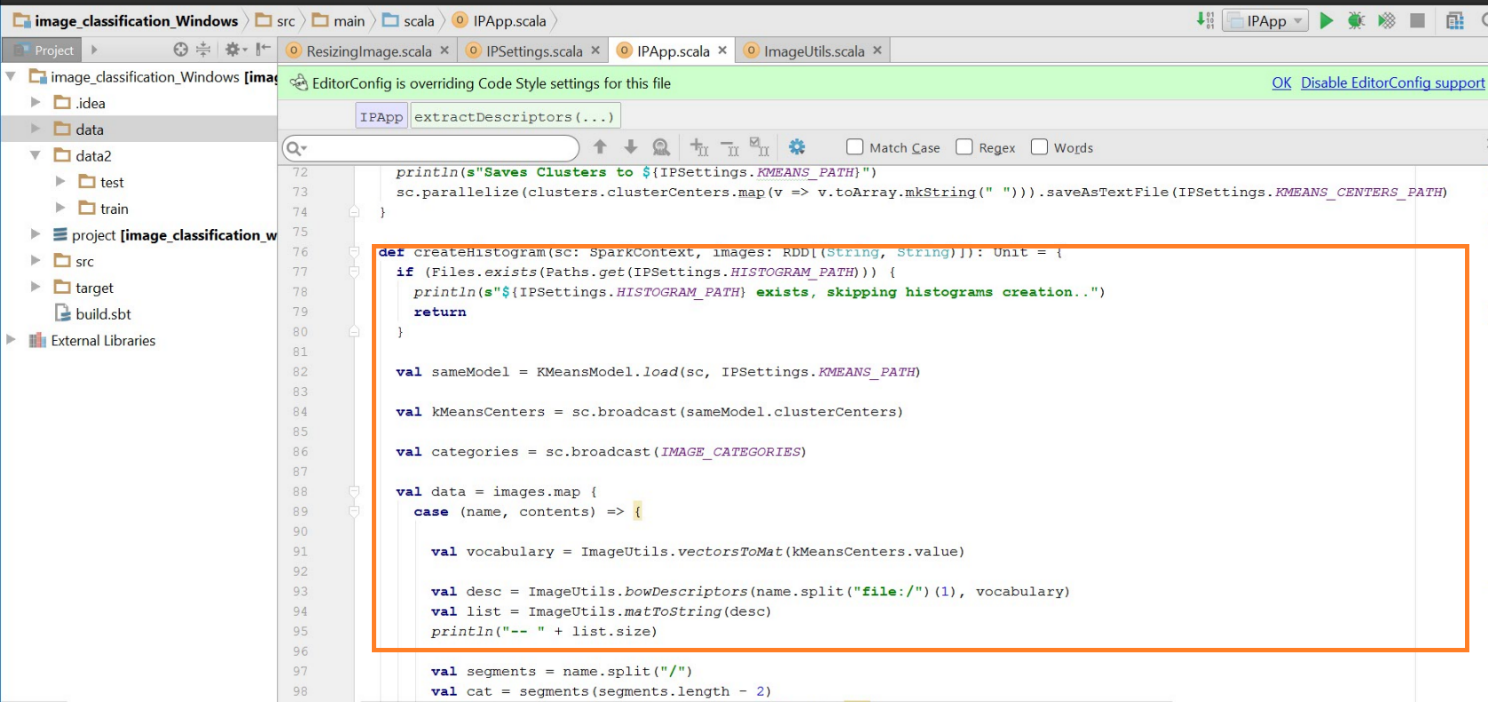
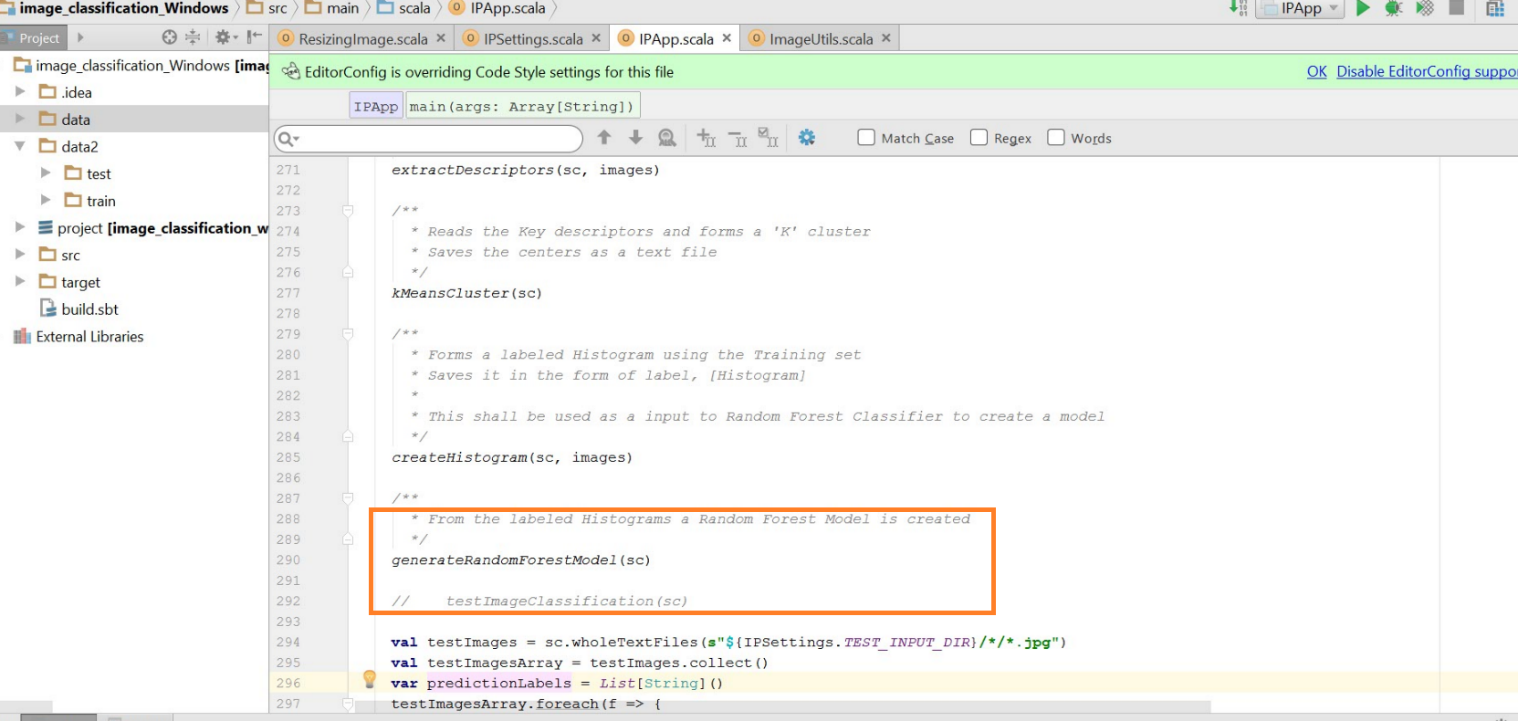
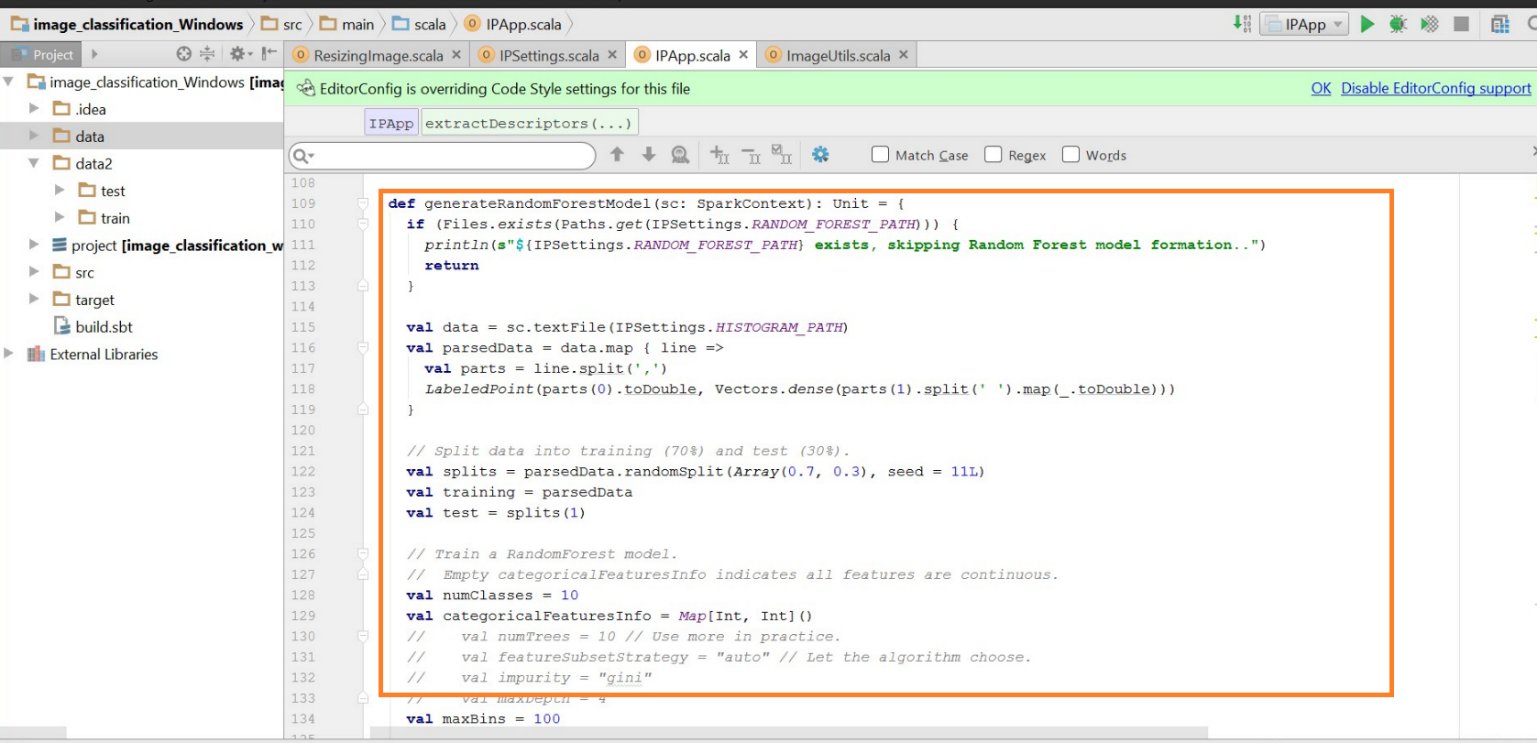
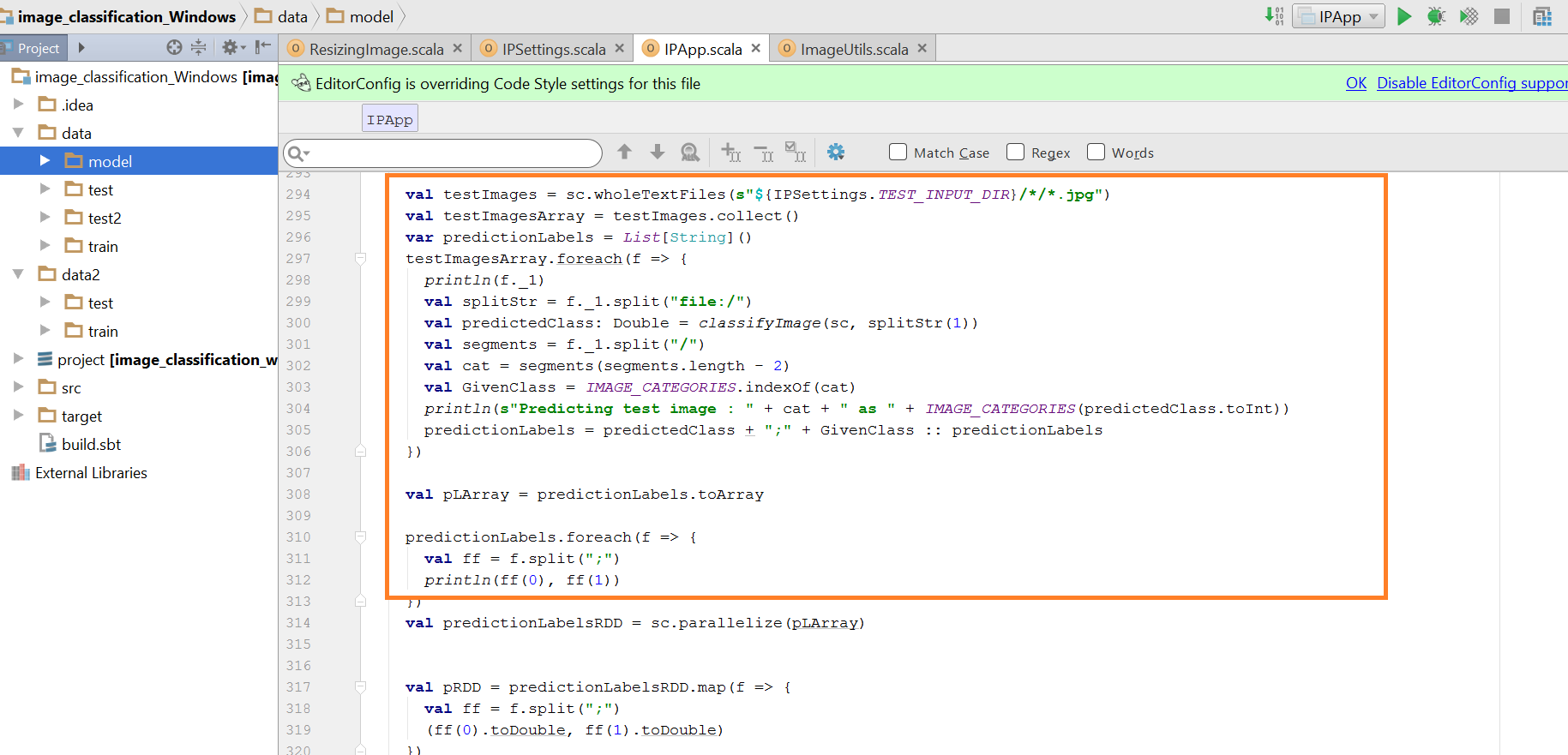
Name: Lava Kumar S

Class ID : 38

**Question**  
Create your own dataset for Image Classification Problem. Use the workflow as discussed in the Tutorial 4 Session using Decision Tree Algorithm. Report the accuracy and confusion matrix obtained. In the Wiki Page, include a brief description of your dataset and purpose behind image classification problem.  
I have created a list with four classes **dolphin, elephant, flower, sun.**  
  
  
The training images corresponding to the list are.

  
  
  
  
  
  
  
  
From the main functions all the functions are called in sequence.  
**SIFT.**  
Scale-invariant feature transform (SIFT) is an algorithm in computer vision to detect and describe local features in images. or any object in an image, interesting points on the object can be extracted to provide a "feature description" of the object. This description, extracted from a training image, can then be used to identify the object when attempting to locate the object in a test image containing many other objects. To perform reliable recognition, it is important that the features extracted from the training image be detectable even under changes in image scale, noise and illumination. Such points usually lie on high-contrast regions of the image, such as object edges. Another important characteristic of these features is that the relative positions between them in the original scene shouldn't change from one image to another. For example, if only the four corners of a door were used as features, they would work regardless of the door's position; but if points in the frame were also used, the recognition would fail if the door is opened or closed. Similarly, features located in articulated or flexible objects would typically not work if any change in their internal geometry happens between two images in the set being processed. However, in practice SIFT detects and uses a much larger number of features from the images, which reduces the contribution of the errors caused by these local variations in the average error of all feature matching errors.  
Firstly the main program calls the descriptors and the the keypoints are calculated.  
  
  
  
  
**KeyMeansClustering**  
k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.  
Kmeans is calculated for different cluster values. Based on the cluster size the output has varied.  
  
**kmeans for 100 clusters**  
  
**kmeans for 900 clusters:**



**Bag Of Words**  
These values are passed to the histogram functions which generate the bag of words. The function calculates the histograms which is passed to the random forest function  
  
  
**Decison Trees & Random Forests**  
Decision tree learning uses a decision tree as a predictive model which maps observationsT about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a finite set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.  
The histogram values are passed to the decision tree with corresponding depth and the confusion matrix are created.  
  
  
Then the corresponding function is called and the values are generated.  
  
  
  
**Training**  
Then the training data set is passed to the functions to calculate the accuracy. These are passed to test function to classify and calculate confusion matrix.  
  


**ConfusionMatrix**  
A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.  
Confusion matrix is calculated for different cluster sizr.  
**Confusion Matrix for 100 clusters**  
