NAME: YAMINI AALLA **STUDENT ID:014331018**

Building a Classification model for detecting the Image, which one of 11 classes it belongs.

Dataset Description:
The dataset contains features extracted from image beside the images. Images are represented as 887 feature vector which are composed by concatenating majority of HOG features, followed by 256 Hist, LBP, RGB and DF features.
There are input files given train.dat, test.dat and train.labels where train.data file has 21186 records , test.dat has 5296 records, train.labels has all 11 labels for each record in train file.

Feature Selection:
For extracting the important features from dataset I have chosen SelectKBest from Sklearn with score function as f_classif and number of top features as 48. Function takes train data and labels as input and filters data and given first k features with greater scores. Before applying featureselection there are 887 features with the help of SelectKBest it narrowed down to 48 features.

Handling Imbalanced Data:
For handling imbalanced chose SMOTE technique from imbearn library. The goal is to balance dataset by increasing the minority classes by replicating them. The minority class data point is chosen at random and find k nearest neighbors for minority class by calculating Euclidean distance and new instance is generated in between these two points. The new point generated will be convex combination of those two neighboring points.

Splitting Data for Validation:
Using train_test_split method from sklearn split the data to 70:30 ratio and performed all the classification models.

Classification Models:

Random Forest Classifier:
Random Forest Classifier is an ensemble method with base estimator as decision tree.
It uses multiple decision trees and aggregates the results which help in improving the accuracy.
Parameters Chosen:
(n_estimators=120, random_state = 0)

KNN Classifier:
KNN classify the data by calculating the k- nearest neighbors points and picking the majority class.
For this solution chose K = 3.
Parameters Chosen:
(n_neighbors=3)

Decision Tree Classifier:
Decision tree is a supervised machine learning algorithm which splits data based on test conditions and stops when all are records are on same class or have same values.

Parameters Chosen:
(random_state = 0)

Extra Tree Classifier:
This is an ensemble method. It selects samples without replacement. It builds multiple decision trees and uses subsets of features at each split and aggregates the results for improving accuracy. The no of estimators chosen are 600.
Parameters Chosen:
(n_estimators=650)

AdaBoost Classifier:
This is a boosting algorithm which tried to combine weak classifiers to strong. Which has a default base estimator as decision tree, and it is helpful for solving difficult cases.
I Chose base estimator as Extra tree classifier as it is giving best F1 score compared to other classifiers.
Parameters Chosen:
(base_estimator=alg_ext, random_state=0)

Evaluation metrics:
For evaluation of models chose F1 score as a metric.

Approach:
1.For feature selection used SelectKBest with score function as f classify and k=48.
2.For handling imbalanced data used SMOTE technique which over samples the data belonging to minority class.
3.Split the data to train and validation in 70:30 ratio.
4.Applied multiple classification models on the obtained balanced data.
5.Chose the best working classifier by comparing there F1 scores for validation data
6. Trained the best classifier on entire data and predicted the labels for test data.

Results and Analysis:

Classifier	**F1 Score**
Random Forest Classifier	0.955
K-Nearest Neighbors Classifier	0.942
Decision Tree Classifier	0.903
Extra Tree Classifier	0.960
Ada Boost Classifier	0.962

AdaBoost Classifier with base estimator as Extra Tree classifier outperformed other models by giving F1 score of 0.962.

At the time of submission F1 Score is 0.8524 and Rank in Leaderboard is 7.