Data visualization course final project

Itamar Kraitman, I.D 208925578

I'm glad to introduce the final project of data visualization course of semester B. In this project I worked on four notebooks, each notebook with different target.

Notebook number one

The first notebook is continuation of the classification notebook from previous semester, the mission was to improve the accuracy of this notebook using methods and algorithms we learned in this semester. In this notebook I dealt with imbalanced data and used PCA. In this notebook I trained several models such as: RandomForest, Bagging, AdaBoost, XGBoost, knn and Stacking.

Notebook number 2

The second notebook mission was to classify FMNIST. Gray scaling wasn't needed because the images are already in black and white. After visualizing I applied PCA. In this Notebook I used several models such as: LogisticRegression, RandomForest, AdaBoost, XGBoost and Voting. Finally, I tested the best model with the test data set.

Notebook number 3

The mission of this notebook was to classify dogs vs cats data set. In this notebook I decided not to convert the images into grayscale for the simple reason that after gray-scaling the results were worse than before. In this notebook I applied PCA with 0.95 components and trained some models. In this notebook I used several models such as: logisticRegression, XGBoost, Stacking and Voting.

Notebook number 4

This notebook was the trickiest stage in this project, for the reason it required a lot of self-learning and diving deep into pandas, but I feel I cope with it in pretty good way and got good results. First step I took, was to upload both training data and validation data in order to work on them simultaneously, I created two list, one for each set because I believed it will be easer to me to work in this way. Secondly, using pandas I perform some methods in order to process the data sets to meet the requirements Roi wrote in the instruction for this project. The nest step I tool was to applying PCA with four different number of components- 0.8,0.85,0.9,0.95 (meanwhile I noticed that 0.8 and 0.85 had almost the same results with only minor differences so I drop 0.8) and trained several models (same models on each precent) in order to catch the precent that will give me the best scores along with as little as possible number of features. The appropriate precent turned out to be 0.95 with 12 features out of 20 (three features were dropped before). Lastly, I tested the best model with the validation set. The models I used in this notebook are: knn, RandomForest, XGBoost, AdaBoost and Stacking.