Final work Data Science II class

background:

This work summarizes the material that was learnt during Data Science II class. The purpose of this work is to show our understanding of the material and gaining some experience using pandas, and advanced ML algorithms.

Workflow:

My workflow was to first understand the data by playing with it and plotting graphs. This gave me the opportunity to gain insights about the data and see which approach I should take, like choosing the n_components parameter for the PCA. After that I ran ML algorithms. the algorithms I used were:

- knn
- random forest
- logistic regression
- gaussian NB
- adaboost
- xgboost

From here I used ensemble techniques like voting ,stacking , and bagging to try and improve the base estimators.

I was only able to do tuning on the third part, due to the large time it took to run, where even though I ran it on the azure server it took a few days for a run to complete.

First part:

Due to only entering the D.S program this semester, and because this part depends on the previous course, I was given permission not to do this part.

Second part -Fashion mnist:

In this part we were asked to identify between 28x28 pictures of clothes. The options were:

0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot

This is a multi-class problem.

Each row represents a 28x28 picture.

In the playing with the data stage I noticed that we can represent 85% of the data with only 40 features, Which was quite amazing at first, then I ran all the Ml models. I saw that only 3 of them passed the 85% accuracy mark, so I tried stacking and voting on them. also I tired to improve the best model which was xgboost.

The best model was using boosting on xgboost with the score of 88% on the validation and 87% on the test.

Third part - Dogs Vs Cats:

This data set contains 25000 pictures of dogs and cats. our goal is to make an ML model without using neural networks, and try to achieve the best score possible.
In the playing with the data stage I resized the photos to 32x32 RGB and noticed that we can represent 90% of the data with only 400 features (which is roughly 10% of the original features count). I also tried to split the histogram of the picture and see if that had any effect, and sadly it didn't. Because the test had no label I had to split the data into 3 parts by myself. The models gave me an accuracy of 58%-67%.
In previous runs i took the 3 models that has passed the 60% accuracy mark, but then I saw that I had better result only with the adaboost and xgboost, I also tried bagging with the xgboost as the base estimator. which was better the the voting classifier by only 0.5%.

I got 68% on the test data

Forth part - hands:

This data contains recorded data from hands motion in various states.
In the playing with the data stage I plotted one of the subject and tried to gain insights on the data. From what I saw there was a big correlation between the difference of both hands and the state that they are in.
So instead of using the original features I made my own that points to differences in various features.
In the preprocessing stage I needed to work on the data before passing it to the ML models. First I combined the RightHand data to all of the Alone states data, then I cleared the first 7 seconds of the data.
Instead of using the original clips I took the maximum values of each 2 second period. I chose the maximum because we are dealing with differences and it made more sense to me. then I concat all of the dataframes into one. because we are talking about time series I cant use train_test split, so I just had to take the last 10% of the rows and use it as the data.
on this data set the best scoring model was random forest with 75%.
on the ensemble stage i saw that there was not a lot of gain, but because there was so little data compared to the other dataset, I could

run a grid search, which boosted out accuracy from 75% to 83%. and on the test we got 96% accuracy!