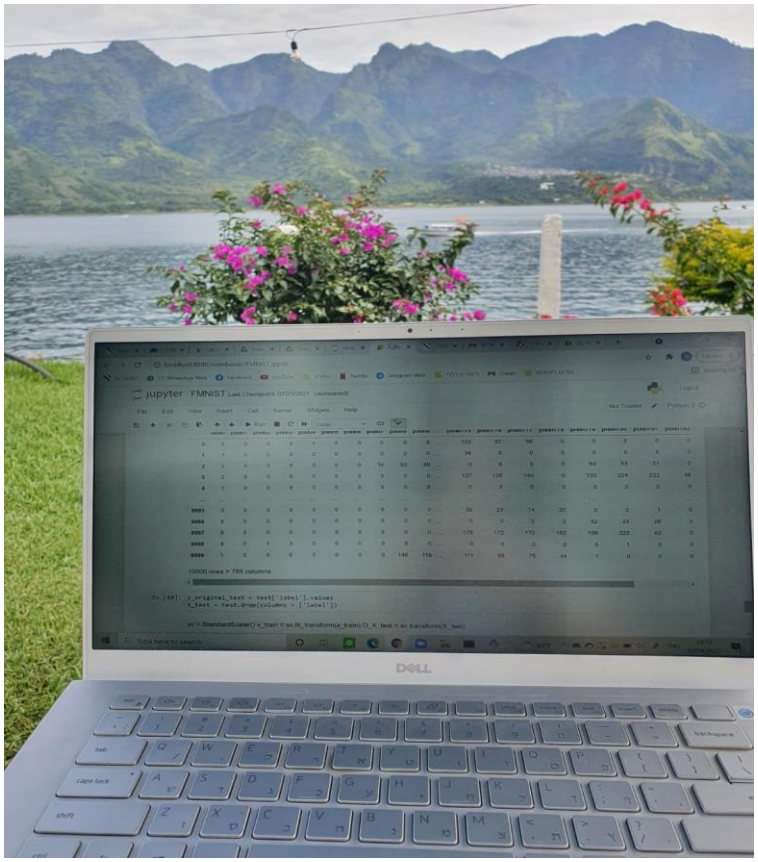
Methods and explanation of Data Science final project

Submitter- Shauli Taragin

Date- 30.08.21



Note: This project was done while on vacation in Guatemala and Costa Rica☺

Explanation of my agenda throughout this project:

1. My main agenda while working on the process of this project was to use the minimal number of components Whilst still achieving high scores. A sort of a tradeoff, and I tweaked my models plenty of times in order to succeed in this goal. Ultimately, I am very pleased with the results I achieved. That is, I reached high scores and used a small number of dimensions.
2. I used randomized search instead of grid search because in our datasets running a simple grid search took extremely long. Therefore, we used randomized search instead.
3. For that same reason I only implemented the searches on models which I saw had the potential to perform well (the initial accuracy score of those said models were high).
4. I used pipelines for preprocessing. These pipelines consisted of a standard scalar and PCA. Learning from the book I comprehended that this is the best way to use these methods. I also used pipelines for the K-Means models.
5. In the F-mnist data set I dealt with a balanced data set and therefore did not plot a confusion matrix/classification report because the accuracy score was enough.
6. The Hands sync notebook was a little tricky since it took a while to create and clean the data set. Additional explanation on how this was done can be found in the Hands sync notebook.

Models and Methods I have used during this project:

Models:

- K Neighbors Classifier
- Logistic Regression
- K-mean(using mini batches)
- Random Forest Classifier
- XGB Classifier
- Ada Boost Classifier
- Gaussian Classifier
- Voting Classifier (hard and soft)
- Bagging Classifier(pasting as well)
- Stacking Classifier

Methods:

- Classification report
- Data Scaling
- Dimensionality reduction using PCA
- Grid/Randomized Search
- Pipeline
- Ensemble Methods