Data Science Final Project

Overview

- This project is the final project of Data Science course and written in python(jupyter notebook).
- It is direct continuation of the task from last semester, you can find the last project Here.
- In this project, we walked through machine learning from basic models to more advanced one.
- In the project I tried to move forward along with the book (Hands-on Machine Learning) and slowly improve the models.
- This project was built from 4 parts:

Notebook 1:

- Improving the project from last semester with the new knowledge we gained in the semester.
- In this part I started directly from the Ensemble Learning because in its previous part we tested a lot of regular models.
- The main models were: Ada boost(with DecisionTree), Xgboost and RandomForest.
- In this case the Ada boost give us the best result **(73.27%)** the Ada boost improved the result by 1.3% (from last semester). Because these are health tests we want to reduce the features as much as possible.
- A reduction with the help of feature_importances function is more effective then PCA in this case because in this way we can reduce the cost of the tests and make it easier for the subject (Coincidentally, the result is also better).
- After reducing **6 out of 13 features** with the help of feature_importances function, we reached **72.77** percent and downloaded almost half of the features!

Model	features	mean accuracy
AdaBoost	13	73.27
AdaBoost	6	72.77

Notebook 2:

- Prediction of Fashion-MNIST Dataset.
- In this part, first i presented basic information and then started testing models (no pre-processing was needed besides dividing by 255).
- I have used some basic models and also in Ensemble models the results are as follows:
  result befor PCA

Model	mean accuracy
KNeighbors	85.0
LogisticRegression	85.1
DecisionTree	79.4
xgboost	90.3
GradientBoosting	83.4

After getting the results on all the Data I tried to reduce dimensions with PCA, After printing the cumsum of "pca.explained_variance_ratio_" I chose 200 n_components because it represents the vast majority of the Data.
result after PCA

Model	mean accuracy
LogisticRegression	85.1
xgboost	88.7

My final result is that xgboost with PCA use only 25% of the data with 88.79% mean accurancy (vs 100% of the data with 90.3%) so we will prefer to use the model after PCA!

Notebook 3:

- Prediction of Dogs vs. Cats dataset
- At first I resized all the images using an Bicubic interpolation, converted each image to a row in a large table (50000×12289),then i label the images.
- I have used some basic models and also in Ensemble models for after getting the result i try to reduce dimensions in 2 ways, PCA and convert the images from RGB to grayscale the results are as follows:

Model	mean accuracy
LogisticRegression	60.0
KNeighbors	55.1
DecisionTree	55.8
RandomForest	66.1
xgboost	66.8
RandomForest pca	65.6
xgboost pca	65.6
RandomForest gray	64.7
xgboost gray	64.6

features vs accuracy:

Model	features	mean accuracy
xgboost	12288	66.8
xgboost pca	1454	65.6
xgboost gray	4097	64.6

- Because the accuracy percentages are low I would choose the most accurate model but if we lack processing power we will selecte the model after the PCA!

Notebook 4

- classify between three situations in the way people communicate with each other, Spontan, Sync and Alone.
- At first I read all the Data from the csv files and built one big Data frame.
- Drop erorrs from the DF (right heands in Alone, null valus etc.) and tack every 10th row.
- After the pre prossing i add some visualizations
- Modeling with diffrent models:

Model	mean accuracy
LogisticRegression	89.2
RandomForest	81.5
Naive Bayes	86.6
AdaBoost	67.6
xgboost	98.4
voting	94.7
Stacking	97.9

After PCA

Model	mean accuracy
LogisticRegression	87.2
RandomForest	80.0
Naive Bayes	86.5
AdaBoost	80.6
xgboost	95.7
voting	94.3
Stacking	95.8

- I decided to stick with the xgboost before the PCA.
- The result of xgboost on the validation Data was **88.44%** when the main error is when the model predict spontan but in fact it should be synchronized.

About:

This project is part of Data Science course of Ariel university and made for study purposes.
This project was made by Or Trabelsi, for more information please contact me, email - ortrsa@gmail.com.