Αναγνώριση πρότυπων Project 2022

Ονοματεπώνυμο : Κωνσταντίνος -Ηλίας Χονδροριζος

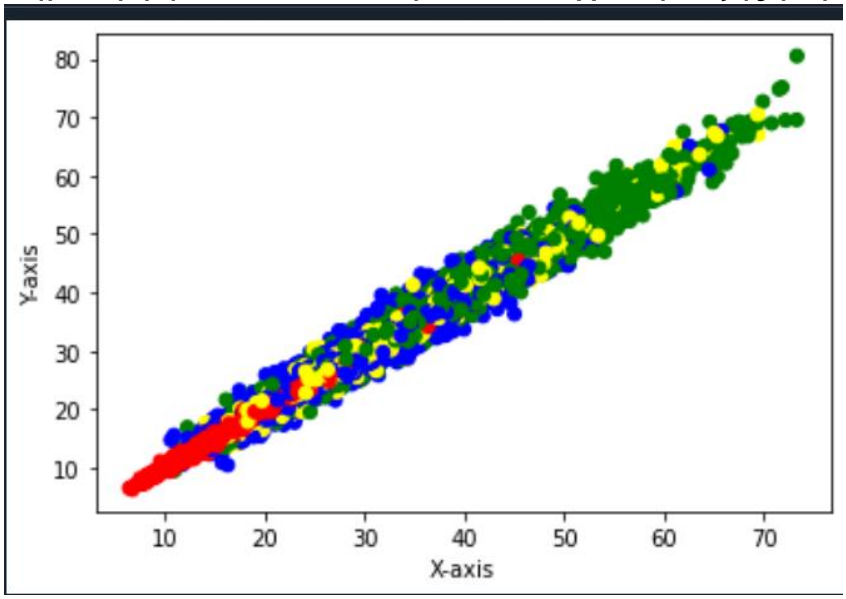
A.E.M. : 3812

Ερώτημα 1ο)

Αρχικά ,κατέβασα και διάβασα τα περιεχόμενα του αρχείου σύμφωνα με τις προϋποθέσεις της εκφώνησης στην συνάρτηση def LoadData(imageFile,labelFile,Filesize) με πληροφορίες που άντλησα από εδώ: https://stackoverflow.com/questions/40427435/extract-images-from-idx3-ubyte-file-or-gzip-via-python

Ερώτημα 2ο)

Στη συνέχεια ,στον πίνακα M2(=$\widehat{M}$) αποθήκευσα αντικείμενα της κλάσης Vector που αναπαριστά ένα διάνυσμα με όλα τα χαρακτηριστικά που απαιτούνται στην εργασία .Έπειτα δημιούργησα ένα scatterplot που έχει την εξής μορφή:

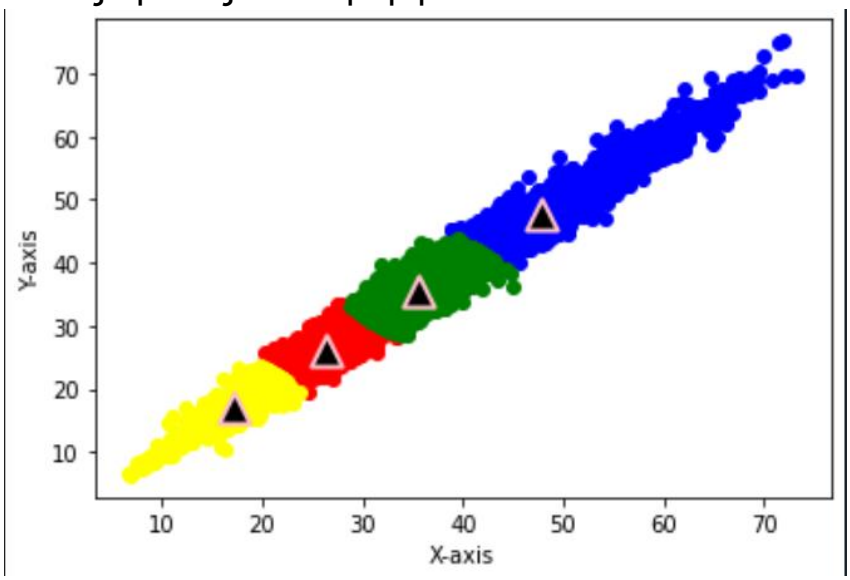


Ερώτημα 3ο)

Όπως μπορούμε να καταλάβουμε από το πάνω Plot ,τα δεδομένα δεν είναι κατανεμημένα με τέτοιο τρόπο ώστε να διευκολύνετε η ομαδοποίηση με την χρήση του K-means αλγορίθμου .Αυτό επιβεβαιώνετε αν ρίξουμε μια ματιά στο average του purity των τεσσάρων κλάσεων:

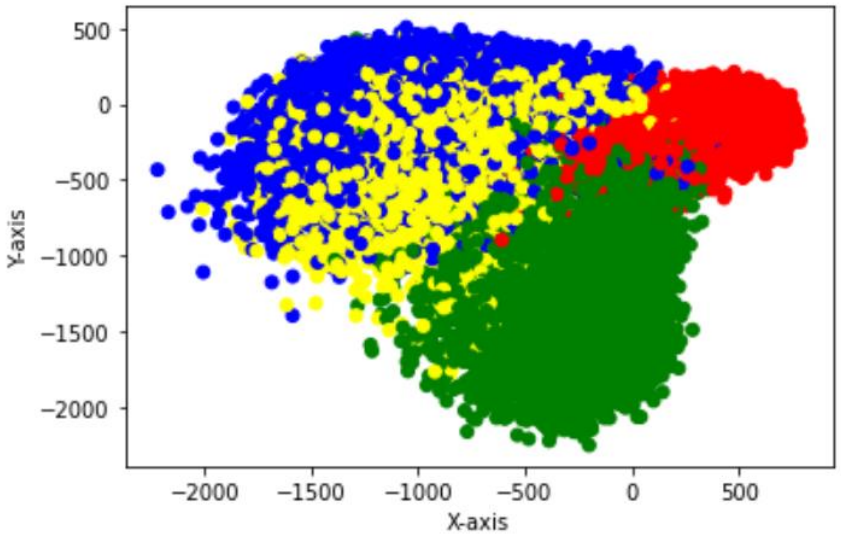
```
Exercise 3 purity : 0.498958
```

και τις ομάδες να διαμορφώνονται έτσι:



<u>Ερώτημα 4ο)</u>

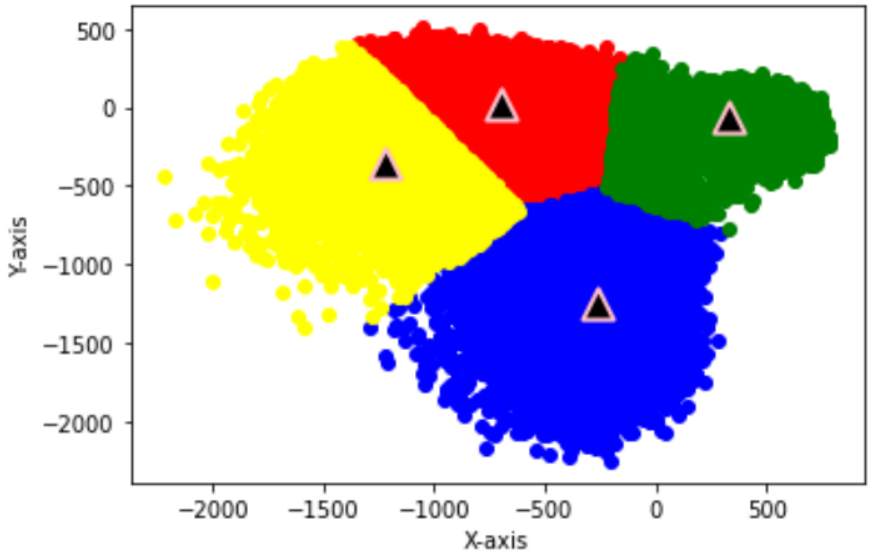
Με την χρήση του PCA αλγορίθμου και την δραστική μείωση των διαστάσεων των δεδομένων μας ,παρατηρούμε ότι τα πράγματα διορθώνονται και περνούν μορφή (δυο διαστάσεις) πιο φιλική ως προς τον clustering αλγόριθμο μας:



Πράγματι ,τα αποτελέσματα του purity των κλάσεων μας βελτιώνονται σημαντικά και φτάνουν τα παρακάτω ποσοστά για αριθμό διαστάσεων 2,25,50 και 100 αντίστοιχα :

```
Exercise 4 (V=2) purity :0.717348
Exercise 4 (V=25) purity :0.717390
Exercise 4 (V=50) purity :0.717389
Exercise 4 (V=100) purity :0.717390
Max Purity (V=100) : 0.717390
```

Με τις ομάδες τελικά να περνούν μορφή:



<u>Ερώτημα 5ο)</u>

-