Neural Networks Project Lab Report
05/2021

Κουτσουρελάκης Χαρίλαος
ΤΠ4591

Για το project του εργαστηρίου επέλεξα το Glass Identification dataset.

Το dataset αποτελείται απο 214 Instances (γραμμές) και 10 Attributes (στήλες). Υπάρχει και μία επιπλέον στήλη η οποία περιέχει τα χαρακτηριστικά των κλάσεων.

Οι στήλες με τη σειρά μας δείχνουν:

1. ID number
2. RI: Δείκτης διάθλασης
3. Na: Νάτριο
4. Mg: Μαγνήσιο
5. Al: Αλουμίνιο
6. Si: Πυρίτιο
7. K: Κάλιο
8. Ca: Ασβέστιο
9. Ba: Βάριο
10. Fe: Σίδηρος
11. Τύπος γυαλιού (7 class attributes)

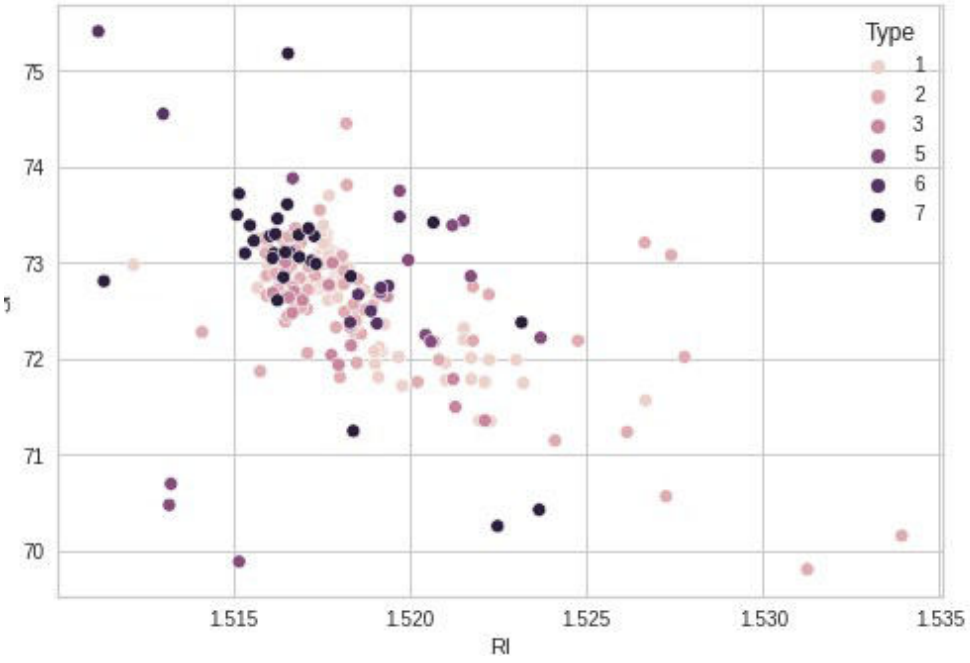
Οι κατηγορίες που βρίσκονται στην 11η στήλη είναι και τα Targets που χρησημοποιώ αργότερα στον κωδικα και αποτελούνται από:

- (163) Window glass
    - (87) float processed
        - (70) building windows
        - (17) vehicle windows
    - (76) non-float processed
        - (76) building windows
        - (0) vehicle windows
- (51) Non-window glass
    - (13) containers
    - (09) tableware
    - (29) headlamps

Αναφορές για τον κώδικα

1) **Στο πρώτο βήμα** διαβάζω το dataset με το pandas και έπειτα το κανω εμφανίζω με την εντολή df.describe(). Θα μπορούσα να χρησημοποιήσω το df.head() αλλά δεν εμφάνιζε τα δεδομένα το ίδιο αναλυτικά.

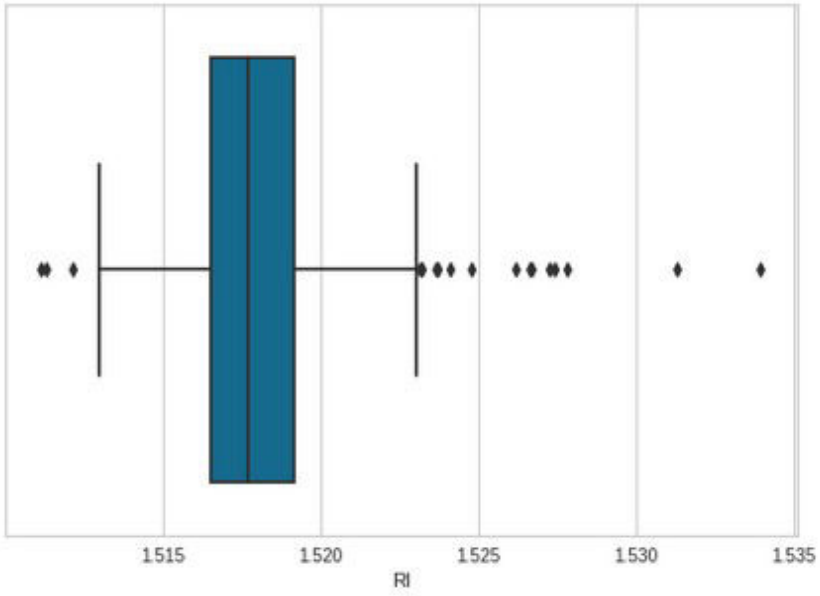
Έπειτα κανω visualize τα δεδομένα μου χρησημοποιώντας το scatter plot του πακέτου seaborn. Τα δεδομένα του x, y τα επέλεξα τυχαία και έβαλα extra τους τύπους των δεδομένων στην παράμετρο hue για να δώσω μια χρωματική απεικόνιση στα δεδομένα μου για καλύτερα αποτελέσματα.



2) **Στο δεύτερο βήμα** προσπάθησα να απεικονίσω τα outliers μέσω της εντολής boxplot του πακέτου seaborn. Δοκίμασα τυχαία μερικές απο τις κατηγορίες του dataset και σχολίασα σχετικά με μερικά απο τα αποτελέσματα από τα οποία βρηκα

Έπειτα, χρησημοποίησα την εντολή Z-score (συμβουλεύτηκα τα documents της python στο Internet) και εμφάνισα (πιστεύω) κάποια γενικά outliers του dataset.

Το συγκεκριμένο boxplot αφορά την κατηγορία RI του dataset το οποίο δείχνει 14 πιθανά outliers.



3) **<u>Στο τρίτο βήμα</u>** μετασχημάτισα τις κατηγορίες των αποτελεσμάτων μου σε δυαδικό συστημα (bits-format) ούτως ώστε να εκπαιδεύσω το δίκτυο μου. Έκανα pre-process τα δεδομένα μου με τη χρήση της MinMaxScaler και δημιούργησα ενα training set και test set με την αρχική τιμή split-rate του 70%-30%.

4) **<u>Στο τέταρτο βήμα</u>** δημιούργησα ενα νευρωνικό δίκτυο με 30 Νευρώνες στο κρυφό layer και 7 Νευρώνες εξόδου (Output Neurons).

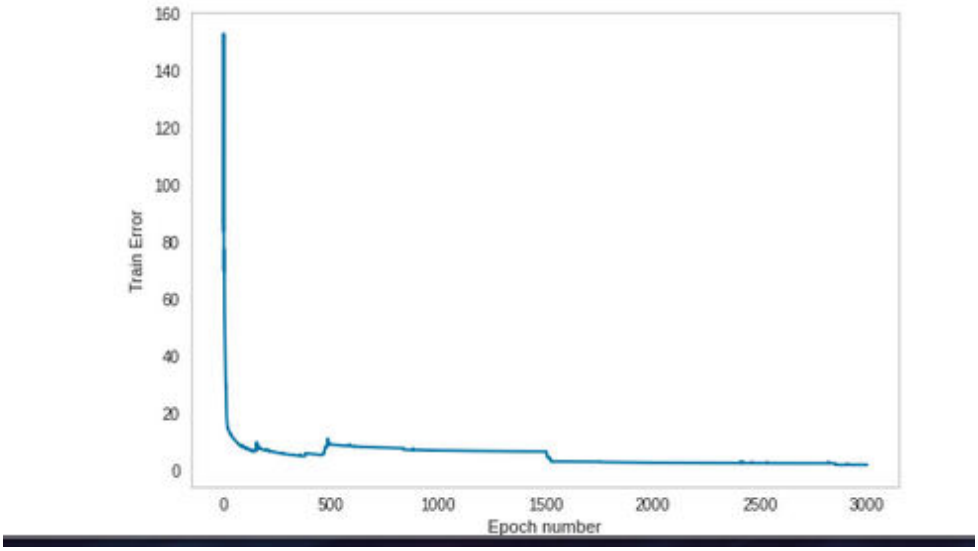
Χρησημοποίησα τις παραμέτρους που δώθηκαν

- Learning rate: 0.3
- Maximum number of epochs to train: 3000

- Performance goal: 1e-5 (0.00001)
- Epochs between displays: 100

Τα αποτελέσματα μου ήταν τα εξής:

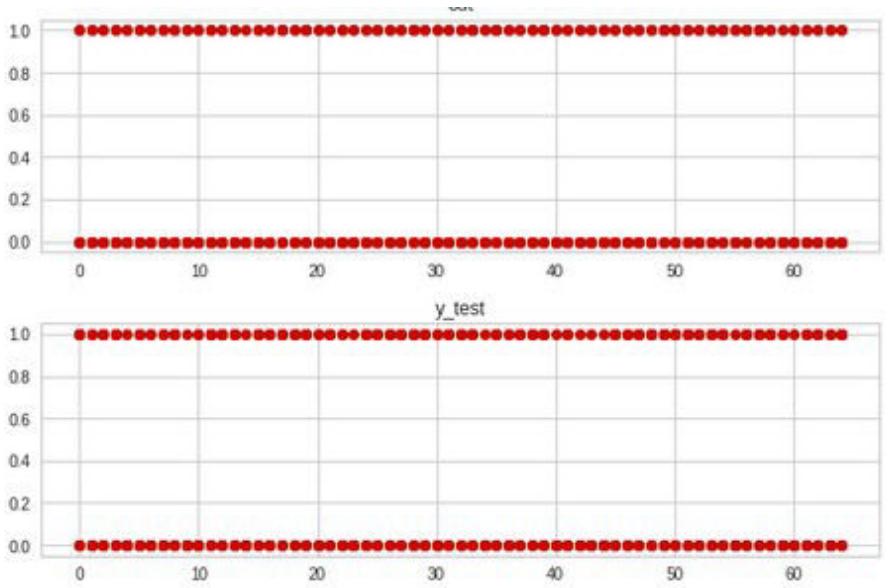
```
Epoch: 100; Error: 7.885324448708324;
Epoch: 200; Error: 7.067623893341116;
Epoch: 300; Error: 5.655717351630038;
Epoch: 400; Error: 5.871163477669583;
Epoch: 500; Error: 9.134853582623414;
Epoch: 600; Error: 8.557605229165121;
Epoch: 700; Error: 8.148412956853385;
Epoch: 800; Error: 7.893329242661972;
Epoch: 900; Error: 7.221939420827654;
Epoch: 1000; Error: 7.035756166574992;
Epoch: 1100; Error: 6.942393085298566;
Epoch: 1200; Error: 6.852240437256544;
Epoch: 1300; Error: 6.76794526690401;
Epoch: 1400; Error: 6.723607932425594;
Epoch: 1500; Error: 6.678915281705645;
Epoch: 1600; Error: 3.1464539264041114;
Epoch: 1700; Error: 3.109297023718715;
Epoch: 1800; Error: 2.944844112411492;
Epoch: 1900; Error: 2.840341530312142;
Epoch: 2000; Error: 2.7643017909478758;
Epoch: 2100; Error: 2.711498320606321;
Epoch: 2200; Error: 2.6609869881708903;
Epoch: 2300; Error: 2.631534677175393;
Epoch: 2400; Error: 2.601225344617522;
Epoch: 2500; Error: 2.571968002628214;
Epoch: 2600; Error: 2.549454875096454;
Epoch: 2700; Error: 2.5353396100369885;
Epoch: 2800; Error: 2.5242340020209606;
Epoch: 2900; Error: 2.016467443663887;
Epoch: 3000; Error: 2.01224905448891;
The maximum number of train epochs is reached
```



Το τελικό accuracy ήταν 12.307692307692308. Δεν είναι πολύ υψηλό, αλλά μετά

από διάφορους πειραματισμούς με τους νευρώνες στο hidden layer κατέληξα σε αυτό το συνδυασμό.

5) **<u>Στο πέμπτο βήμα</u>** παρουσιάζω και συγκρίνω το test set με την matplotlib.



6) **<u>Στο έκτο βήμα</u>** καταγράφω τα αποτελέσματα των πειραμάτων μου με διαφορετικούς συνδυασμούς σε νευρώνες, layers και εποχές και τα παρουσιάζω σε table mode μέσω του πακέτου pandas.

Δοκίμασα 6 διαφορετικούς συνδυασμούς όπως αναγράφεται στην παρακάτω εικόνα. Δόκιμασα από πολυ μικρό μέχρι αρκετά μεγάλο αριθμο νευρώνων (hidden layer) και το μέγιστο εύρος των εποχών μου ήταν 3000.

Τα καλύτερα αποτελέσματα τα πήρα από τον συνδυασμό 3 (επειδή είχε το μεγαλύτερο accuracy.

	Epochs	Neurons_of_hidden_layer	Output_Neurons	Error_from_last_epoch	Accuracy
0	3000	3	7	61.107383	3.076923
1	1100	10	7	29.015639	16.923077
2	2500	21	7	36.169077	16.923077
3	500	30	7	50.641819	21.686084
4	1400	50	7	13.148390	9.230769
5	3000	100	7	4.899278	3.076923

Best results were those for 30 Neurons and 500 Epochs

7) **Στο βήμα εφτά**, δοκιμάζω τα παρακάτω ποσοστά με τον καλύτερο συνδυασμό που μου έδωσε το προηγούμενο βήμα (συνδυασμός νούμερο 3).

	Split Rate	Epochs	Neurons_of_hidden_layer	Output_Neurons	Error_from_last_epoch	Accuracy
0	50%-50%	500	30	7	26.137946	24.299065
1	60%-40%	500	30	7	20.135638	22.093023
2	80%-20%	500	30	7	74.029240	9.302326
3	90%-10%	500	30	7	79.807571	31.818182

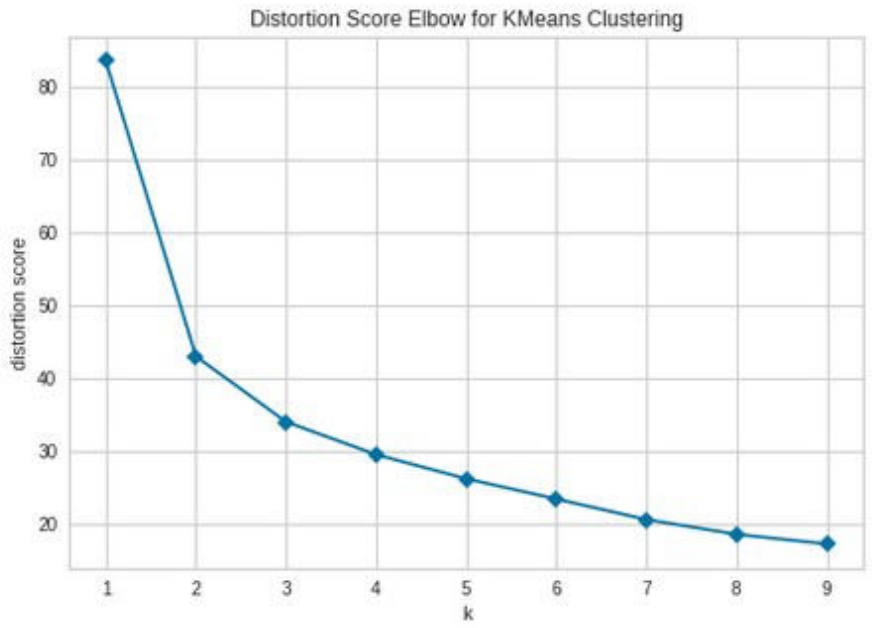
####### What do you conclude?

8) **Στο βήμα 8** δημιουργώ ένα δίκτυο Kohonen και του δίνω το dataset μου. Τα αποτελέσματα του δικτύου για 1000 εποχές είναι:

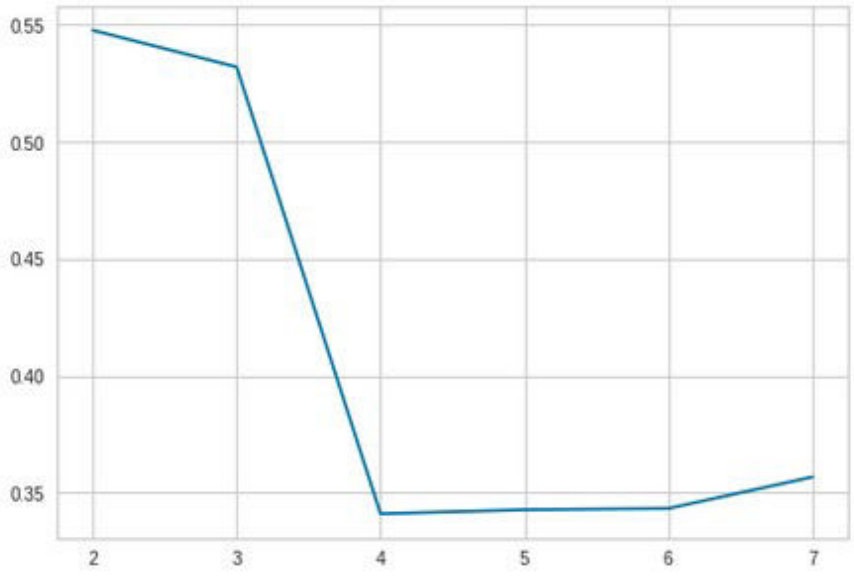
```
Epoch: 200; Error: 141.6398110771792;
Epoch: 400; Error: 141.46862894193893;
Epoch: 600; Error: 141.4704089096092;
Epoch: 800; Error: 141.4704382245136;
Epoch: 1000; Error: 141.47043868908543;
The maximum number of train epochs is reached
```

Έπειτα χρησημοποιώ την μέθοδο elbow και silhouette score για να διευκρινίσω τον αριθμό των clusters.

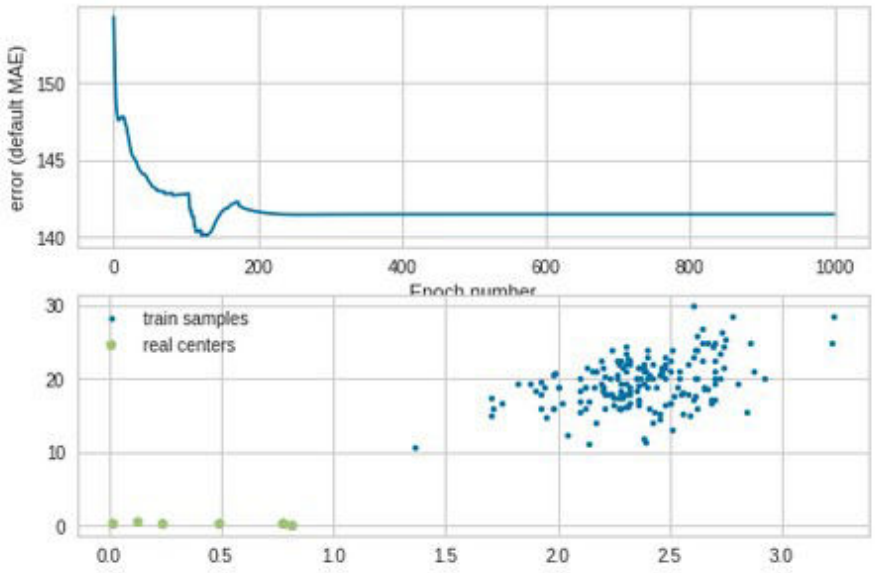
Elbow score



Silhouette score



Τέλος, εκπαιδεύω το δίκτυο και εμφανίζω τα αποτελέσματα



9) Κατά την άποψη μου, ένα νευρωνικό δίκτυο, για να είναι αποτελεσματικό χρειάζεται μια έμπιστη βάση δεδομένων, τόσο από πλευράς δεδομένων όσο και ακρίβειας αυτών. Αν υπάρχει διαχωρισμός των κατηγοριών των δεδομένων των οποίων θέλουμε να δώσουμε στο δίκτυο μας, θα υπάρξει σαφώς καλύτερη ανταπόκριση απο δίκτυο και επομένως καλύτερα αποτελέσματα.

Όσο καλύτερα είναι τα δεδομένα μας, τόσο καλύτερα θα μπορέσει το νευρωνικό δίκτυο να υπολογίσει τα patterns τα οποία αναζητάμε.