ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ
Προαιρετική Άσκηση 1
2017-2018
Τοσουνίδης Βασίλειος
ΑΜ: 109 / 15

1. Εισαγωγή:

Για τον σκοπό της εργασίας επιλέχθηκε ένα dataset από το Kaggle το οποίο περιλαμβάνει πληροφορίες για τα βιομηχανικά χαρακτηριστικά ορθοπαιδικών ασθενών καθώς και μία κατάταξη ανάλογα με τις τιμές τους. Τα δεδομένα είναι κατηγοριοποιημένα σε 2 αρχεία, τα οποία σχετίζονται. Στο πρώτο οι ασθενείς είναι διαχωρισμένοι σε δύο κατηγορίες “Normal” και “Abnormal” ανάλογα με τις μετρήσεις τους, ενώ στο δεύτερο σε “Normal”, “Hernia” και “Spondylolisthesis”. Κάθε ασθενής παρουσιάζεται από έξι βιομηχανικά χαρακτηριστικά τα οποία που προέρχονται από το σχήμα και τον προσανατολισμό της λεκάνης και της οσφυϊκής σπονδυλικής στήλης.

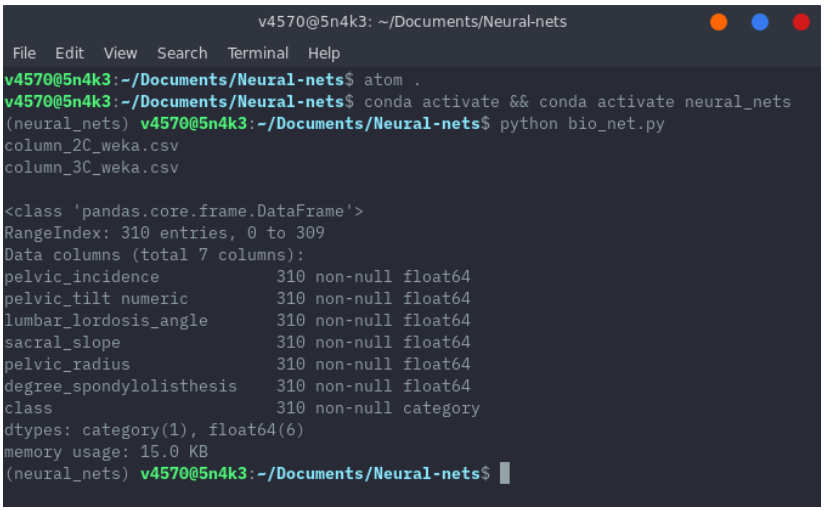
2. Εργασιακό περιβάλλον

Για την ολοκλήρωση της εργασίας έγινε χρήση της γλώσσας προγραμματισμού Python (έκδοση 3) και μερικών βιβλιοθηκών κατάλληλων για διαχείρηση και οπτικοποίηση δεδομένων αλλά και εκπαίδευση του μοντέλου. Συγκεκριμένα αυτές είναι:

- **Tensorflow:** Είναι μία βιβλιοθήκη ανοιχτού κώδικα για υψηλής απόδοσης αριθμητική επεξαργασία, η οποία δημιουργήθηκε από την Google.
- **Scikit-learn:** Είναι μια βιβλιοθήκη ανοιχτού κώδικα η οποία περιέχει πληθώρα αλγορίθμων για μηχανική μάθηση όπως random forests, k-means κλπ.
- **Matplotlib:** Είναι μία βιβλιοθήκη ανοιχτού κώδικα για οπτικοποίηση δεδομένων σε plots.
- **Pandas:** Είναι μία βιβλιοθήκη ανοιχτού κώδικα η οποία παρέχει εύκολες στη χρήση δομές δεδομένων για υψηλού επιπέδου αριθμητικούς υπολογισμούς.

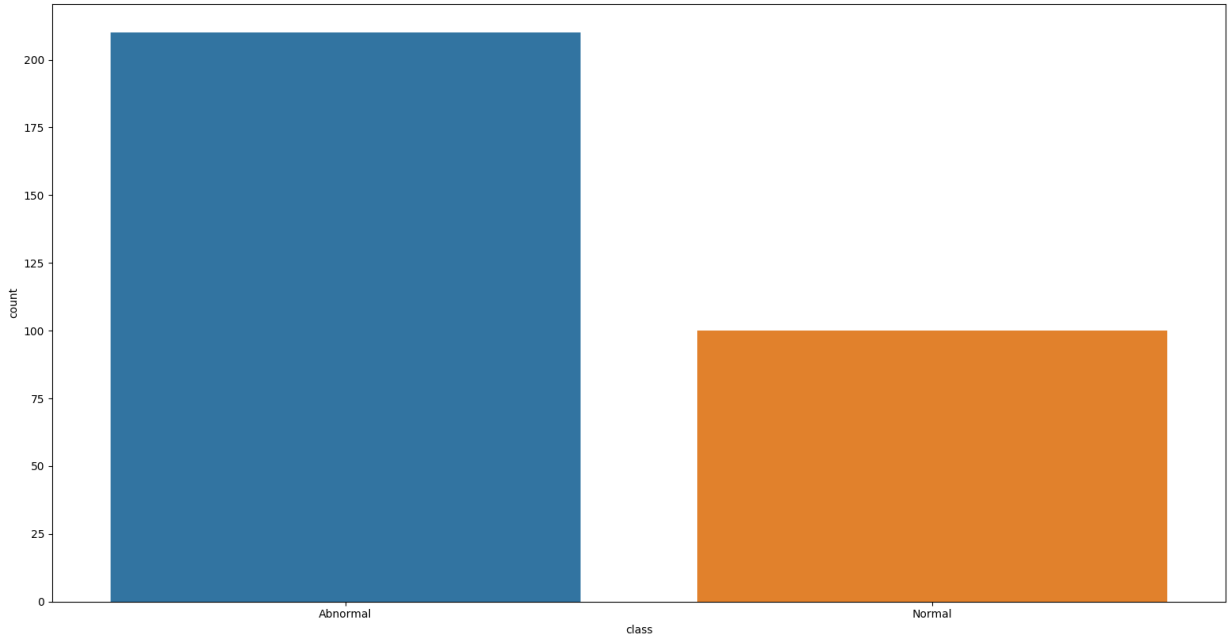
3. Ανάλυση δεδομένων

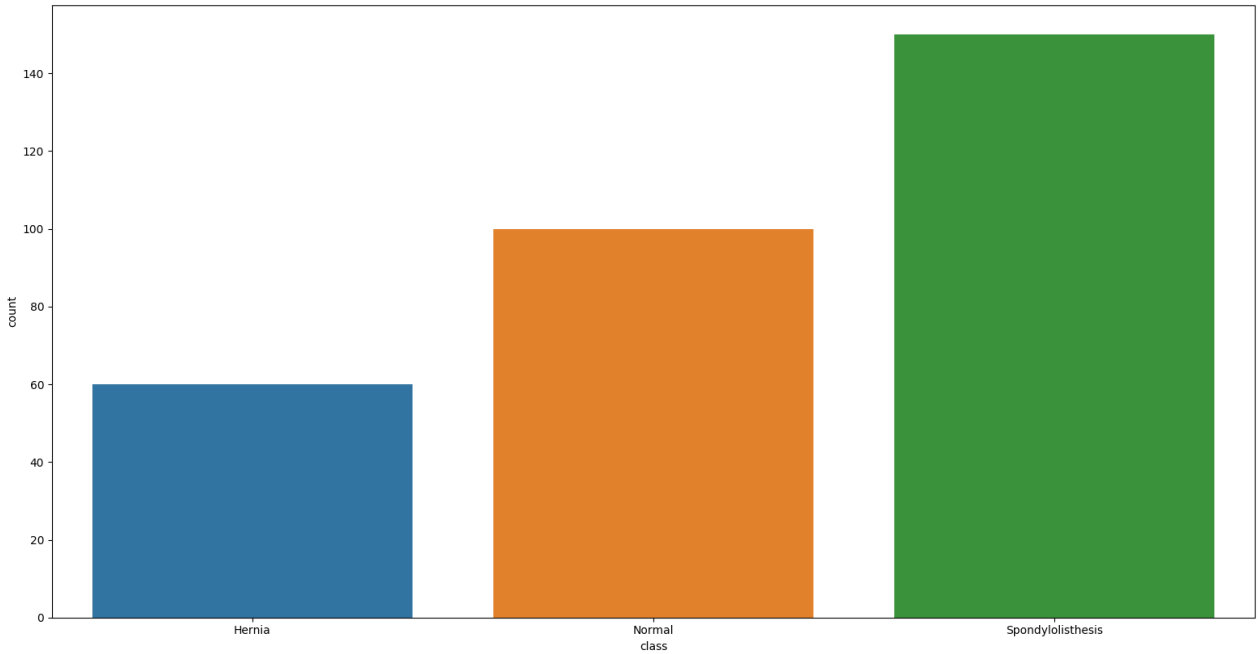
Αρχικά πρέπει να εξετάσουμε τα δεδομένα για την συνοχή τους, το πλήθος τους, τον τύπο τους, αλλά και για την ύπαρξη ή μη NaN στοιχείων. Αρχικά με χρήση της βιβλιοθήκης Pandas έγινε ανάγνωση των δεδομένων σε ένα dataframe αντικείμενο, το οποίο είναι ουσιαστικά ένα πολυδιάστατο array. Με χρήση εντολής της βιβλιοθήκης έγινε εκτύπωση των πληροφοριών του dataset στο terminal.



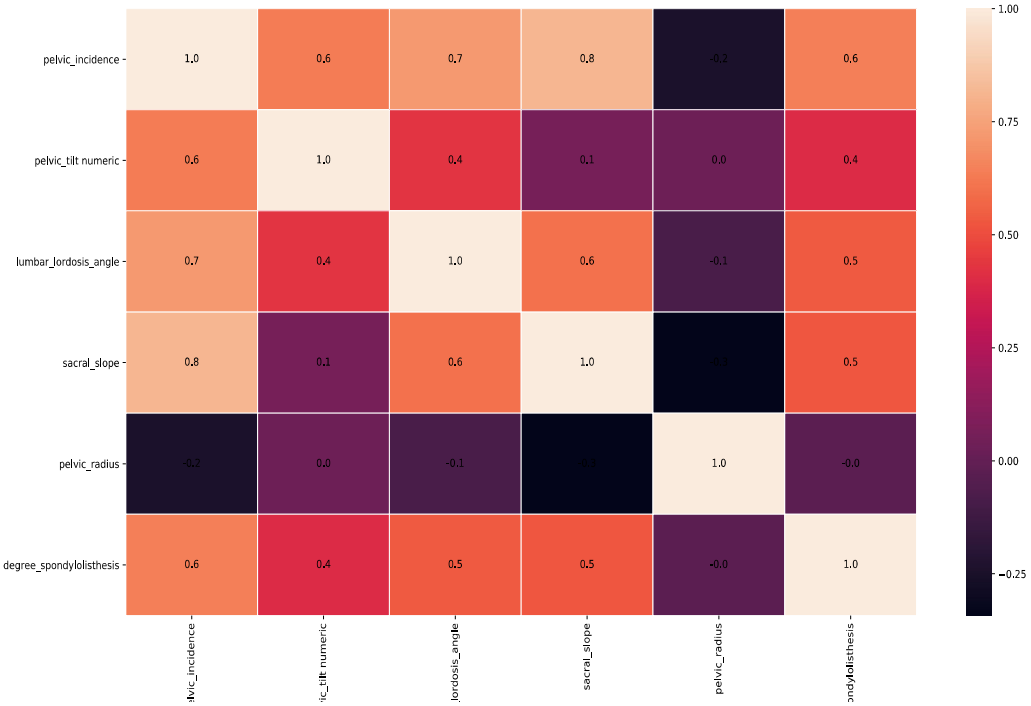
Από την παραπάνω ανάλυση συμπεραίνουμε ότι έχουμε 310 εισαγωγές, 7 στήλες, 6 για τις μετρήσεις και 1 για την κλάση του κάθε ασθενή. Οι τιμές είναι της μορφής float και δεν έχουμε null στοιχεία οπότε δεν χρειάζεται να κάνουμε κάποια αλλαγή. Η μόνη αλλαγή που έγινε ήταν η μετατροπή της στήλης με τις κλάσεις σε τύπο “category” της βιβλιοθήκης για να μπορέσουμε να κάνουμε πιο εύκολη την εκπαίδευση.

Σε αυτά τα plot φαίνεται ο διαχωρισμός των ασθενών για τα δύο αρχεία:





Και εδώ μπορούμε να δούμε το πως συσχετίζονται οι τιμές μεταξύ τους ανά δύο:



Αυτό το διάγραμμα θα μας φαινόταν πολύ χρήσιμο σε περίπτωση που θέλαμε να χρησιμοποιήσουμε κάποιες από τις τιμές. Έτσι με βάση αυτό θα διαλέγαμε αυτές με τη μεγαλύτερη συσχέτιση.

4. Μοντέλο

Αφού μαζεύτηκαν όλες οι απαραίτητες πληροφορίες μπορούμε πλέον να δώσουμε τα δεδομένα στο μοντέλο για να εκπαιδευτεί. Αρχικά πρέπει να καθορίσουμε τους δύο άξονές μας (x και y), έτσι για τον άξονα x διαλέξαμε όλες τις στήλες εκτός της τελευταίας που περιέχει τις κλάσεις και για τον άξονα y τις κλάσεις. Το επόμενο βήμα που έγινε ήταν να χωρίσουμε τα δεδομένα έτσι ώστε κάποια να χρησιμοποιηθούν για την εκπαίδευση και κάποια για την επικύρωση των αποτελεσμάτων της. Τα ποσοστά αυτά αναλογούν σε 70% για την εκπαίδευση και 30% για την επικύρωση. Για τον διαχωρισμό αυτό χρησιμοποιήθηκε μια συνάρτηση της βιβλιοθήκης scikit-learn (train_test_split).

Ο αλγόριθμος που χρησιμοποιήθηκε για την εκπαίδευση είναι ο Random_Forest, οποίος μετά από πειράματα παρουσίασε τα καλύτερα αποτελέσματα. Για την λειτουργία αυτού του αλγόριθμου χρειάζονται κάποιες παράμετροι, όπως ο αριθμός των συνολικών δέντρων, βημάτων, κόμβων κλπ, οι οποίες καθορίστηκαν πειραματικά σε αρκετά καλές τιμές. Χρησιμοποιήθηκε η υλοποίηση του αλγορίθμου της βιβλιοθήκης Tensorflow ο οποίος είναι αρκετά αποδοτικός. Λόγω του μικρού πλήθους των δεδομένων δε φαίνεται η υπεροχή αυτής της υλοποίησης του αλγορίθμου, αλλά σε περισσότερα δεδομένα η διαφορά θα ήταν αισθητή. Αφού το μοντέλο εκπαιδευτεί δοκιμάζουμε με τα δεδομένα επαλίθευσης και τα αποτελέσματα για τα δύο αρχεία είναι τα εξής:

- **column_2c:** 0.8172043
- **column_3c:** 0.8494624