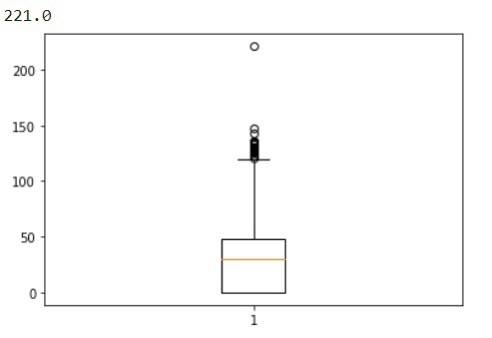
**ΕΡΓΑΣΙΑ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ**

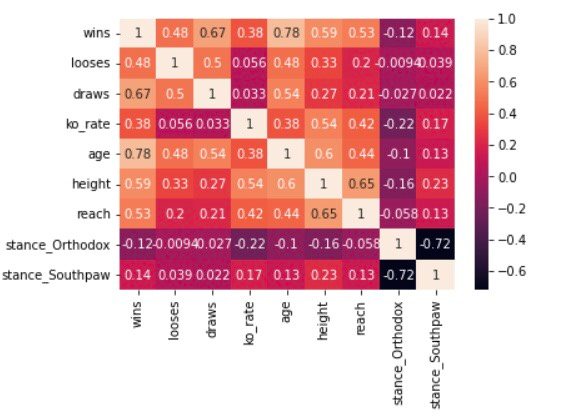
ΚΑΛΟΓΕΡΟΠΟΥΛΟΣ ΔΗΜΗΤΡΗΣ MTN2202

H παρούσα εργασία έχει ως στόχο την κατασκευή μοντέλου το οποίο δέχεται ως είσοδο διάφορα στοιχεία επαγγελματιών μποξέρ pre game και με βάση αυτά προβλέπει το τελικό αποτέλεσμα -δηλαδή το ποιος επικράτησε έναντι του άλλου-.Στην πορεία της, δυο διαφορετικά dataset που περιέχει το ένα στοιχεία των αθλητών και το δεύτερο σημαντικούς αγώνες και στοιχεία αυτών όσον αφορά την εξέλιξη τους, συνενώνονται σε ένα πιο εκτεταμένο το οποίο τελικώς χρησιμοποιείται για τη διαδικασία της πρόβλεψης του αποτελέσματος των αγώνων.

Σε ένα πρώτο επίπεδο, πραγματοποιείται το καθάρισμα των δεδομένων στο dataset “fighters”.Διαγράφεται η στήλη που αφορά τις εθνικότητες των αθλητών και στη συνέχεια γίνεται χειρισμός των τιμών που λείπουν συνολικά στο dataset Τα Unknowns γίνονται 0.0( από string σε float) κάτι που θα χρειαστεί και στη συνέχεια ,και αφαιρείται το ‘%’ από την στήλη KO.Επίσης το preprocessing των δεδομένων συνεχίζεται στις στήλες height και reach ώστε να μείνουν στο τέλος μόνο float τιμές (σε μετρά και εκατοστά αντίστοιχα).Επιπροσθέτως , γίνεται encoding στην στήλη stance σε 0 και 1 ανάλογα με το εάν ο αθλητής είναι αριστερόχειρας ή δεξιόχειρας. Ενδιαφέρον παρουσιάζουν οι τιμές στην στήλη age όπου υπάρχουν (πέραν από missing values) ακραίες τιμές μη ρεαλιστικές όπως η τιμή 221 (χρόνων).

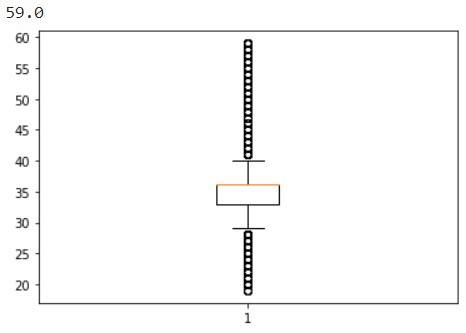


Μάλιστα χωρίς να έχει γίνει ακόμα χειρισμός των τιμών της στήλης ,παρατηρούμε από το correlation matrix ότι είναι high correlated οι τιμές age με τις νίκες.



Έτσι η επιλογή ήταν να αντικατασταθούν οι τιμές που λείπουν (δηλαδή τα 0) καθώς και οι ακραίες τιμές (άνω των 60 ) που ρεαλιστικά δεν εκφράζουν ενεργούς αθλητές , με το μέσο όρο των υπόλοιπων τιμών.

Έτσι τελικά το boxplot γίνεται :



Αντίστοιχα έγινε ο χειρισμός στις τιμές που λείπουν από το στήλες “height” και “reach”, δηλαδή έγινε αλλαγή των μηδενικών τιμών με το μέσο όρο των υπολοίπων. Μετά από όλες τις δοκιμές που διενεργήθηκαν, τη μεγαλύτερη ακρίβεια την έδινε η επιλογή χειρισμού των ακραίων τιμών στις 3 προαναφερθείσες στήλες. Δοκιμάστηκε να παραμείνει ως έχει το dataset καθώς και να γίνουν είτε μόνο σε μια στήλη είτε ανά δύο οι αντικαταστάσεις των μηδενικών τιμών. Επιπρόσθετα ,διαχωρίζεται το επώνυμο από τα ονόματα των αθλητών ( η τελευταία λέξη δηλαδή ελέγχοντας να μην είναι JR) ώστε στη συνέχεια να γίνει η συνένωση με το άλλο dataset.

Όσον αφορά το δεύτερο dataset με την ονομασία “popular matches”,αρχικά διαγράφονται οι κολώνες “date” και “place”.Στη συνέχεια πραγματοποιείται handling των missing values συνολικά στο dataset στις στήλες : 'opponent\_1\_estimated\_punch\_power',

        'opponent\_2\_estimated\_punch\_power',

        'opponent\_1\_estimated\_punch\_resistance',

        'opponent\_2\_estimated\_punch\_resistance',

        'opponent\_1\_estimated\_ability\_to\_take\_punch',

        'opponent\_2\_estimated\_ability\_to\_take\_punch', 'opponent\_1\_rounds\_boxed',

        'opponent\_2\_rounds\_boxed', 'opponent\_1\_round\_ko\_percentage',

        'opponent\_2\_round\_ko\_percentage', 'opponent\_1\_has\_been\_ko\_percentage',

        'opponent\_2\_has\_been\_ko\_percentage', 'opponent\_1\_avg\_weight',

        'opponent\_2\_avg\_weight'

με τον μέσο όρο τις κάθε μιας. Όπως και στο προηγούμενο dataset έτσι και σε αυτό ,για να μπορέσει να γίνει η συνένωση ,από τις στήλες opponent 1 και opponent 2 ,δημιουργούνται ,ακριβώς με τον ίδιο τρόπο δυο νέες στήλες που συγκρατούν το επώνυμο των αθλητών.Eκ νέου μετατρέπουμε σε float τα στοιχεία του dataset ώστε να μπορέσουν να εφαρμοστούν τα διάφορα μοντέλα. Τελευταίο βήμα στο preprocessing των δεδομένων αποτελεί η εξαγωγή από τη στήλη verdict των νικητών ( ή της ισοπαλίας ) που το όνομα του νικητή είναι η πρώτη λέξη της στήλης. Έτσι δημιουργείται μια νέα στήλη (winners) με τους νικητές.

Με βάση τα επώνυμα που προέκυψαν στο πρώτο data sets γίνεται η σύνδεση με το “popular matches” χρησιμοποιώντας ,αντίστοιχα τα επώνυμα των δυο αντίπαλων. Δηλαδή για κάθε έναν αντίπαλο προστίθενται τα στοιχεία που υπάρχουν στο dataset fighters. Έτσι προκύπτει ένα καινούργιο, το οποίο περιέχει όλες τις τιμές για κάθε αθλητή ,το αποτέλεσμα στους μεταξύ τους αγώνες και τα ονόματα τους.

Στο καινούργιο dataset δημιουργείται μια νέα στήλη με τιμές το 1 εάν έχει κερδίσει ο πρώτος αντίπαλος , το 2 εάν κέρδισε ο άλλος και 0 εάν ήρθε ισοπαλία ή εάν είναι άγνωστο το αποτέλεσμα.

Εν συνεχεία σβήνουμε στο νέο dataset τις στήλες με τους χαρακτήρες δηλαδή τα πλήρη ονόματα των αθλητών και τα επώνυμα τους καθώς και τα ονόματα των νικητών ( έχει κρατηθεί στη μεταβλητή winners\_2 όμως το ποιος κέρδισε) ώστε να υπάρχουν μόνο float χαρακτήρες.

Είναι σημαντικό ότι με τη χρήση διάφορων ιστογραμμάτων φάνηκε ότι κανένα feature δεν είχε υψηλή συσχέτιση με τις 3 τιμές της στήλης winners .Επίσης με τη δοκιμή διαγραφής διάφορων features κάθε φορά πάντα τα διάφορα scores ήταν χειρότερα από αυτά που θα παρουσιαστούν παρακάτω.

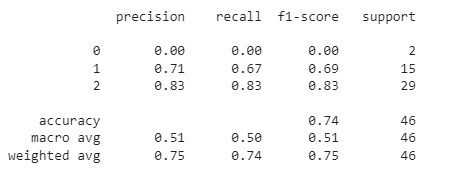
Στο split έγινε stratification ώστε σε train και test να τηρείται η αναλογία των κλάσεων.

Επίσης πραγματοποιήθηκε standardization (προφανώς μόνο στα train δεδομένα).

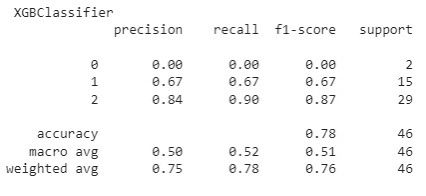
Για κάθε μία από τις μεθόδους διενεργήθηκε stratified K-fold και grid search με διάφορες παραμέτρους κάθε φορά.

Με βάση τις υπερπαραμέτρους που προέκυψαν κάθε φορά έγινα train τα παρακάτω μοντέλα και με τα αντίστοιχα αποτελέσματά τους.

Ο SVM με παραμέτρους : C=0.01 kernel='linear' gamma=0.1

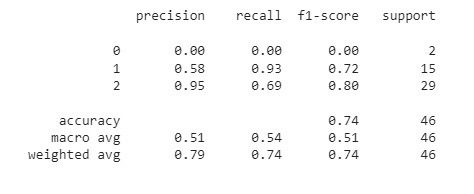


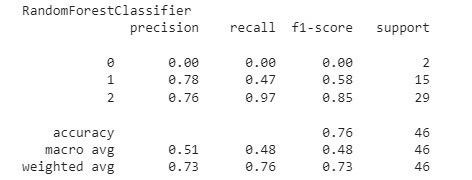
O ΧGBOOST Classifier δίνει τα παρακάτω αποτελέσματα με παραμέτρους : max\_depth=4, learning\_rate=0.5, n\_estimators=200



O Decision\_tree δίνει τα παρακάτω αποτελέσματα με τις παραμέτρους

max\_depth=3,min\_samples\_leaf=1,min\_samples\_split=5



Ο RandomForestClassifier δίνει με τις παραμέτρους    n\_estimators = 100, max\_depth = 4

Mε μικρές αποκλίσεις ,λοιπόν με βάση την weightedF1 o XGboost ήταν ο καλύτερο ταξινομητής ενώ με βάση την macroAvg ήταν όλοι ίδιοι με εξαίρεση τον RandomForest όπου ελάχιστα χαμηλότερη απόδοση