Μηχανική Μάθηση και Μεγάλα Δεδομένα στον Επιχειρηματικό Αθλητισμό: Η περίπτωση της Αγγλικής Premier League.

Togias Panagiotis
TEI of Western Greece, Department of Business Administration
ptogias@outlook.com

Theologos Konstantinos
TEI of Western Greece, Department of Business Administration
kos.theologos@gmail.com

Περίληψη

Στη παρούσα έρευνα τίθεται υπό δοκιμή η εφαρμογή μεθόδων μηχανικής μάθησης και εξόρυξης δεδομένων a priori και Διαβαθμιζόμενη Ενδυνάμωση (με τη χρήση Generalized Boosted μοντέλου) για τον εντοπισμό κρυφών σχέσεων ανάμεσα σε δεδομένα όσο και για τη δημιουργία υποθέσεων στις οποίες αναζητούμε τους παράγοντες αυτούς οι οποίοι έχουν μεγαλύτερη σημαντικότητα για τη διαμόρφωση του τελικού αποτελέσματος ενός αγώνα ποδοσφαίρου της αγγλικής Premier League τη σεζόν 2016, καθώς ακόμα στοχεύουμε στο να ενταχθεί μια σχετικά νέα έννοια στον χώρο της επιστήμης των δεδομένων στην Ελλάδα, αυτή της Διαβαθμιζόμενης Ενδυνάμωσης (Gradient Boosting), εξηγώντας με συνοπτικό και σχετικά απλό τρόπο τη λειτουργία της, ελπίζοντας στην περαιτέρω διάδοσή της στον ελληνικό επιστημονικό τομέα. Παρατηρήθηκε ότι η περίπτωση με την μεγαλύτερη επιτυχία ήταν αυτή της πρόβλεψης της νίκης της γηπεδούχου ομάδας (55,4% θετική απόδοση μοντέλου) έναντι της νίκης της φιλοξενούμενης (30,1% θετική απόδοση μοντέλου) ενώ στη περίπτωση της ισοπαλίας υπήρξε αδυναμία στην εξαγωγή ενός ικανοποιητικού ποσοστού επιτυχίας καθώς κατάφερε να αποδώσει μέγιστο ποσοστό σωστής πιθανότητας πρόβλεψης 40,9% (με βάση 70%) και χρίζει επιπλέον έρευνας στον τομέα του συνόλου δοκιμής και εκπαίδευσης. Η σχετική σημαντικότητα των ανεξάρτητων μεταβλητών κατέδειξε ότι στις πρώτες θέσεις, για την πρόβλεψη του τελικού αποτελέσματος, έρχεται το αποτέλεσμα του ημιχρόνου, οι ομάδες αυτές καθ' αυτές, τα συνολικά σουτ και ο διαιτητής κάθε αγώνα. Τα αποτελέσματα του αλγορίθμου a priori εξήγαγαν μία κατάσταση στην οποία παρατηρήθηκαν κανόνες συσχέτισης, έχοντας ως "συμπέρασμα" τον δείκτη των γκολ της γηπεδούχου στο ημίχρονο να είναι μηδέν (0) και τα γκολ της γηπεδούχου στο τελικό αποτέλεσμα του αγώνα να είναι ένα (1), μέρος των οποίων επιβεβαιώνουν τα αποτελέσματα της διαβαθμιζόμενης ενδυνάμωσης.

Generalized Boosted Model, big data, machine learning, a priori, gradient boosting, premier league

1. Εισαγωγή

Η χρήση των δεδομένων στο ποδόσφαιρο-το πιο δημοφιλές άθλημα στον κόσμο-έχει δει μεγάλη ανάπτυξη τα τελευταία χρόνια. Το άθλημα ξεκίνησε να αντιλαμβάνεται την χρησιμότητα της ανάλυσης δεδομένων και του πραγματικού πλεονεκτήματος που αυτή του προσφέρει. Στην περίπτωση της Premier League και πιο συγκεκριμένα του Αγγλικού πρωταθλήματος θα εστιάσουμε την ερευνά μας. Είναι το πιο ακριβό και πολυδιαφημισμένο πρωτάθλημα στον κόσμο, με εκατομμύρια τηλεθεατές να παρακολουθούν την εξέλιξή του καθημερινά. Είναι ένα καινοτόμο πρωτάθλημα που προσπαθεί να εξελιχθεί μέσα από διάφορες τεχνολογικές προόδους. Μια από αυτές είναι και η ανάλυση των δεδομένων.

Σήμερα και τα 20 γήπεδα της Premier League στην Αγγλία είναι εξοπλισμένα με ένα σετ από 8-10 ψηφιακές κάμερες που παρακολουθούν και εντοπίζουν κάθε παίκτη μέσα στον αγωνιστικό χώρο. 10 σημεία δεδομένων συλλέγονται κάθε δευτερόλεπτο για κάθε έναν από τους 22 παίκτες που βρίσκονται στο γήπεδο, παράγουν δηλαδή 1,4 εκατομμύρια σημεία δεδομένων ανά παιχνίδι. Οι αναλυτές έπειτα θα αποκωδικοποιήσουν τα δεδομένα για κάθε τάκλιν, σουτ ή πάσα και θα τα προωθήσουν στους προπονητές και τους αναλυτές απόδοσης, προκειμένου να αποκτήσουν γνώση για το τι ακριβώς συνέβη στην διάρκεια του αγώνα με ή χωρίς την μπάλα.

Στο σύγχρονο παιχνίδι η χρήση των δεδομένων έχει να κάνει με την εύρεση της λεπτομέρειας. Αυτού του 1% που μπορεί να εκμεταλλευτούμε για να ανακαλύψουμε τις αδυναμίες του αντιπάλου και κάνει την διαφορά μεταξύ νίκης και ήττας. Είναι κάτι περισσότερο από απλή εφαρμογή δεδομένων σε κομμάτια τακτικής. Η αντικειμενική πληροφόρηση χρησιμοποιείται από όλα τα σωματεία (football clubs) την ενίσχυση της αποτελεσματικότητας και την ανάπτυξη των διαδικασιών που επιτρέπουν στον οργανισμό να είναι όσο το δυνατόν προετοιμασμένος από το γήπεδο μέχρι την αίθουσα συνεδριάσεων.

Το αντικείμενο της παρούσας έρευνας είναι η δοκιμή και η εφαρμογή μεθόδων μηχανικής μάθησης και εξόρυξης δεδομένων (Apriori και Διαβαθμιζόμενη Ενδυνάμωση) για τον εντοπισμό κρυφών σχέσεων ανάμεσα σε δεδομένα όσο και για τη δημιουργία υποθέσεων στις οποίες αναζητούμε τους παράγοντες αυτούς που έχουν μεγαλύτερη σημαντικότητα για τη διαμόρφωση του τελικού αποτελέσματος ενός αγώνα ποδοσφαίρου της Αγγλικής Premier League τη σεζόν 2016. Ακόμα, η έρευνα, στοχεύει στο να εντάξει μια σχετικά νέα έννοια στον χώρο της επιστήμης των δεδομένων στην Ελλάδα, αυτή της Διαβαθμιζόμενης Ενδυνάμωσης (Gradient Boosting), εξηγώντας με συνοπτικό και σχετικά απλό τρόπο τη λειτουργία της, ελπίζοντας στην περαιτέρω διάδοσή της στον ελληνικό επιστημονικό τομέα, καθώς και στη μελλοντική διάθεση των αποτελεσμάτων στον γενικότερο τομέα του επιχειρηματικού αθλητισμού.

Οι δείκτες οι οποίοι χρησιμοποιήθηκαν είναι οι εξής:

Key to results data

Div	Κατηγορία Πρωταθλήματος
Date	Ημερομηνία διεξαγωγής
HomeTeam	Γηπεδούχος
AwayTeam	Φιλοξενούμενος
FTHG	Γκολ γηπεδούχου σε ολόκληρο αγώνα
FTAG	Γκολ φιλοξενούμενου σε ολόκληρο αγώνα
HTHG	Γκολ γηπεδούχου σε ένα ημίχρονο
HTAG	Γκολ φιλοξενούμενου σε ένα ημίχρονο
HTR	Αποτέλεσμα ημιχρόνου

Match Statistics

Attendance	Θεατές
Referee	Διαιτητής
HS	Σουτ γηπεδούχου
AS	Σουτ φιλοξενούμενου
HST	Σουτ γηπεδούχου στον στόχο
AST	Σουτ φιλοξενούμενου στον στόχο
HHW	Δοκάρια γηπεδούχου
AHW	Δοκάρια φιλοξενούμενου
HC	Κερδισμένα κόρνερ γηπεδούχου
AC	Κερδισμένα κόρνερ φιλοξενούμενου
HF	Φάουλ γηπεδούχου (που έχει διαπράξει)
AF	Φάουλ φιλοξενούμενου (που έχει διαπράξει)
HO	Οφσάιντ γηπεδούχου
AO	Οφσάιντ φιλοξενούμενου
HY	Κίτρινες κάρτες γηπεδούχου
AY	Κίτρινες κάρτες φιλοξενούμενου
HR	Κόκκινες κάρτες γηπεδούχου
AR	Κόκκινες κάρτες φιλοξενούμενου
HBP	Πόντοι καρτών γηπεδούχου

2. *Μέθοδοι*

2.1. *Εξαγωγή κανόνων συσχέτισης A priori*

Η εξαξωγή κανόνων συχέτισης είναι μια κοινή τεχνική που χρησιμοποιείται για να βρεθούν συσχετίσεις μεταξύ πολλών μεταβλητών. Ο αλγόριθμος A priori είναι ένας κλασικός αλγόριθμος για την εκμάθηση κανόνων συσχέτισης και έχει σχεδιαστεί για να λειτουργεί με βάσεις δεδομένων που περιέχουν κρυφές και μη συσχετίσεις. Όπως είναι σύνηθες στην εξόρυξη κανόνων συσχέτισης, δεδομένου ενός συνόλου από στοιχειοσύνολα (αρχικά δεδομένα), ο αλγόριθμος προσπαθεί να βρει υποσύνολα που είναι κοινά σε τουλάχιστον έναν ελάχιστο αριθμό C από τα στοιχειοσύνολα.

Ο A priori χρησιμοποιεί προσέγγιση «από κάτω προς τα πάνω», όπου τα συχνά υποσύνολα επεκτείνουν ένα στοιχείο ανα χρονική περίοδο και οι ομάδες των υποψηφίων υποσυνόλων ελέγχονται έναντι των δεδομένων. Ο αλγόριθμος τερματίζει όταν δεν βρεθούν περαιτέρω επιτυχείς επεκτάσεις. Χρησιμοποιεί τεχνική αναζήτησης τύπου “brith first” και μια δομή δέντρου για να μετρήσει τα υποψήφια υποσύνολα, αποτελεσματικά και παράγει σύνολα μήκους k από στοιχειοσύνολα μήκους k - 1. Από εκεί και πέρα κλαδεύει τα στοιχειοσύνολα με όχι και τόσο συχνό μοτίβο.

Ο σκοπός των κανόνων συσχέτισης είναι να αποκαλύπτουν ενδιαφέρουσες σχέσεις μεταξύ δεδομένων. Γι' αυτό το λόγο χρησιμοποιούνται ορισμένα μέτρα τα οποία αξιολογούν το επίπεδο σημαντικότητας του κάθε κανόνα συσχέτισης. Αυτά είναι [6]:

- **Confidence (Strength, Εμπιστοσύνη):** Η εμπιστοσύνη ενός κανόνα συσχέτισης είναι το ποσοστό των περιπτώσεων που καλύπτονται από το LHS του κανόνα και οι οποίες καλύπτονται επίσης από το RHS. Μια τιμή της εμπιστοσύνης κοντά στο 1 είναι ένδειξη ενός σημαντικού κανόνα συσχέτισης.
- **Support (Υποστήριξη):** Η υποστήριξη ενός κανόνα συσχέτισης είναι το ποσοστό όλων των περιπτώσεων στο σύνολο δεδομένων που ικανοποιούν έναν κανόνα, δηλαδή ικανοποιούν το LHS και το RHS του κανόνα. Η υποστήριξη μπορεί να θεωρηθεί ως ένδειξη του πόσο συχνά ένας κανόνας εμφανίζεται σε ένα σύνολο στοιχείων και κατά συνέπεια πόσο σημαντικός είναι ο κανόνας.
- **Lift:** Ορίζεται ως η εμπιστοσύνη διαιρούμενη με το ποσοστό όλων των περιπτώσεων που καλύπτονται από το RHS. Είναι ένα μέτρο της σπουδαιότητας της συσχέτισης και είναι ανεξάρτητο από την κάλυψη.

2.2. *Διαβαθμιζόμενη Ενδυνάμωση (Gradient Boosting)*

Για τις ανάγκες ορισμού μιας ελληνικής ορολογίας του όρου “Gradient Boosting”, προτείνουμε τον “Διαβαθμιζόμενη Ενδυνάμωση”, καθώς περιγράφει με ακρίβεια την λειτουργία του μετα-μαθησιακού σχήματος. Ως διαβάθμιση ορίζεται μια πρόοδος ή μετάβαση από ένα αντικείμενο στο άλλο. Για παράδειγμα από ένα χρώμα στο άλλο ή από μια απόχρωση ενός χρώματος σε μια άλλη.

Η ενδυνάμωση στο σύνολό της είναι μία τεχνική συνόλου (ensemble technique), της οποίας το προγνωστικό μοντέλο προκύπτει από ένα σύνολο πιο απλών εκτιμητών πρόβλεψης χρησιμοποιώντας πολλαπλά δέντρα κατηγοριοποίησης. Όντας μια από τις σημαντικότερες και πιο πρόσφατες καινοτομίες στο πεδίο της ταξινόμησης δεδομένων, λειτουργεί με τη διαδοχική εφαρμογή ενός αλγορίθμου ταξινόμησης σε επαναστάθμισμένες εκδοχές του πεδίου εκπαίδευσης και στη συνέχεια λαμβάνοντας μια σταθμισμένη πλειοψηφία της διαδοχικότητας παράγονται ταξινομητές με μικρότερα ποσοστά σφάλματος [2][3].

Η διαβαθμιζόμενη ενδυνάμωση είναι μια τεχνική μηχανικής μάθησης για ταξινόμηση, η οποία παράγει ένα μοντέλο πρόβλεψης με τη μορφή ενός συνόλου από άλλα, στατιστικώς αδύναμα μοντέλα πρόβλεψης δένδρων αποφάσεων. Δημιουργεί το τελικό μοντέλο με διαδικασία step-wise όπως και άλλες μέθοδοι ενδυνάμωσης και τα γενικεύει, επιτρέποντας τη βελτιστοποίηση μιας τυχαίας διαφορίσιμης συνάρτησης απώλειας [1].

Αντικειμενικά, ο τρόπος λειτουργίας ενός μοντέλου διαβαθμισμένης ενδυνάμωσης είναι με την δημιουργία μιας απλής πρόβλεψης και μιας σειράς δέντρων αποφάσεων, με κάθε δέντρο να προσπαθεί να διορθώσει το σφάλμα πρόβλεψης του δέντρου που βρισκόταν πριν από αυτό. Η διαβαθμισμένη ενδυνάμωση έχει την ικανότητα να αναγνωρίζει τα "ελαττώματα" των αδύναμων τελεστών πρόβλεψης με διαβαθμίσεις στην συνάρτηση απώλειας αντί στις μεγάλες τιμές από τα βάρη των αντίστοιχων δεδομένων στην γενική ενδυνάμωση.

Πιο αναλυτικά, τα σύνολα από τα οποία δημιουργούνται τα μοντέλα διαβαθμιζόμενης ενδυνάμωσης, δημιουργούνται ως εξής [4]:

Οι προβλέψεις των δέντρων αποφάσεων που έχουν δημιουργηθεί προστίθενται

$$D(x) = d_{tree1}(x) + d_{tree2}(x) + \cdots + d_{tree\ldots}(x)$$

i. Το επόμενο δέντρο κατηγοριοποίησης (δέντρο 3) προσπαθεί να "καλύψει" την απόκλιση μεταξύ της εξαρτημένης μεταβλητής $f(x)$ και της παρούσας πρόβλεψης από το μέχρι στιγμής υπάρχον σύνολο αναδιαμορφώνοντας τα κατάλοιπα

Άρα, μέχρι στιγμής η πρόβλεψη είναι:

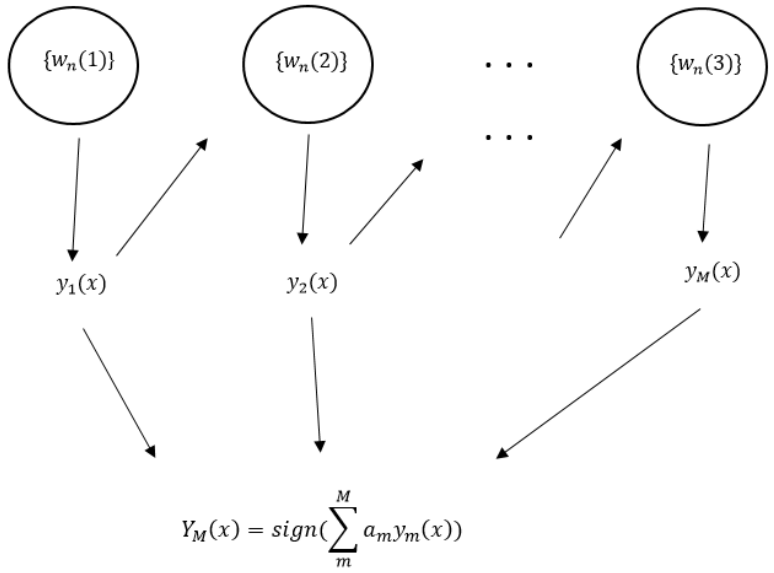
$$D(x) = d_{tree1}(x) + d_{tree2}(x) + d_{tree3}(x)$$

ii. Με την ίδια διαδικασία, το δέντρο 4 θα συμπληρώσει σωστά τα προηγούμενα και θα μειώσει το σφάλμα εκπαίδευσης του συνολικού δέντρου. Η ιδανική περίπτωση θα είναι το δέντρο 4 μαζί με την υπάρχουσα συνολική πρόβλεψη να εξάγει ακριβώς την εξαρτημένη μεταβλητή $f(x)$:

$$f(x) = d_{tree4}(x) + D(x)$$

Στις περισσότερες περιπτώσεις η δομή των καταλοίπων δείχνει τα δεδομένα τα οποία δεν χρησιμοποιήθηκαν για την δημιουργία ενός μοντέλου πράγμα που τα καθιστά έμπιστο οδηγό για τον έλεγχο αποδοτικότητας ενός μοντέλου και εν συνεχεία, της αναβάθμισής του. Η συνάρτηση με την οποία εκφράζονται είναι της μορφής [7]:

$$e_i = y_i - \hat{y}_i \qquad (1)$$



Εικόνα 1 Σχηματική απεικόνιση της λειτουργίας ενδυνάμωσης

Στην εικόνα 1 περιγράφουμε ξανά τον τρόπο λειτουργίας της περίπτωσής μας, με τον κάθε βασικό ταξινομητή $y_m(x)$ να είναι εκπαιδευμένος με αντίστοιχο συναπτικό βάρος (ρυθμός εκπαίδευσης) από τα δεδομένα εκπαίδευσης, στα οποία κάθε συναπτικό βάρος $w_n(m)$ εξαρτάται από την απόδοση του προηγούμενου βασικού ταξινομητή $y_{m-1}(x)$. Μόλις όλοι οι βασικοί ταξινομητές έχουν εκπαιδευτεί, συνδυάζονται από τον αλγόριθμο ενδυνάμωσης για να δώσουν τον τελικό ταξινομητή $Y_M(x)$.

Ο αλγόριθμος εκμάθησης που χρησιμοποιήσαμε, ο GBM (Generalized Boosted Model), όπως αναφέρεται στη βιβλιογραφία, είναι ένας συνδυασμός της εκθετικής συνάρτησης απώλειας του αλγορίθμου AdaBoost προσαρμοστικής ενδυνάμωσης και του αλγορίθμου διαβαθμιζόμενης καθόδου (gradient descent) του Friedman.

Ρυθμός μάθησης

Πριν την ώρα που θα ενταχθεί ένα δέντρο κατηγοριοποίησης στο συνολικό δέντρο, οι προβλέψεις αυτού πολλαπλασιάζονται από κάποιο συναπτόμενο βάρος, το οποίο καλείται ρυθμός μάθησης ή "η" σε μαθηματικές παραστάσεις και είναι μια από τις σημαντικότερες παραμέτρους της διαβαθμιζόμενης ενδυνάμωσης. Για παράδειγμα, σε περίπτωση διαβαθμιζόμενης ενδυνάμωσης με 10 δέντρα κατηγοριοποίησης και μεταβαλλόμενο μέγεθος παιδιών (κλαδιών) αυτών, παρατηρούμε ότι όσο το μέγεθος αυτό μεγαλώνει, οι τιμές των καταλοίπων μικραίνουν αλλά με περισσότερο θόρυβο με συνέπεια να δημιουργείτε μη συνέχεια. Έτσι, όσο μεγαλύτερος είναι ο ρυθμός μάθησης, τόσο μεγαλύτερη είναι η διαφορά από δέντρο σε δέντρο και η μη συνέχεια [4].

Για την εξαγωγή των καλύτερων δυνατών αποτελεσμάτων προτείνουμε μικρές τιμές ρυθμού μάθησης ($0{,}001 < \eta < 0{,}01$) και μεγάλο αριθμό δέντρων κατηγοριοποίησης.

3. Αποτελέσματα

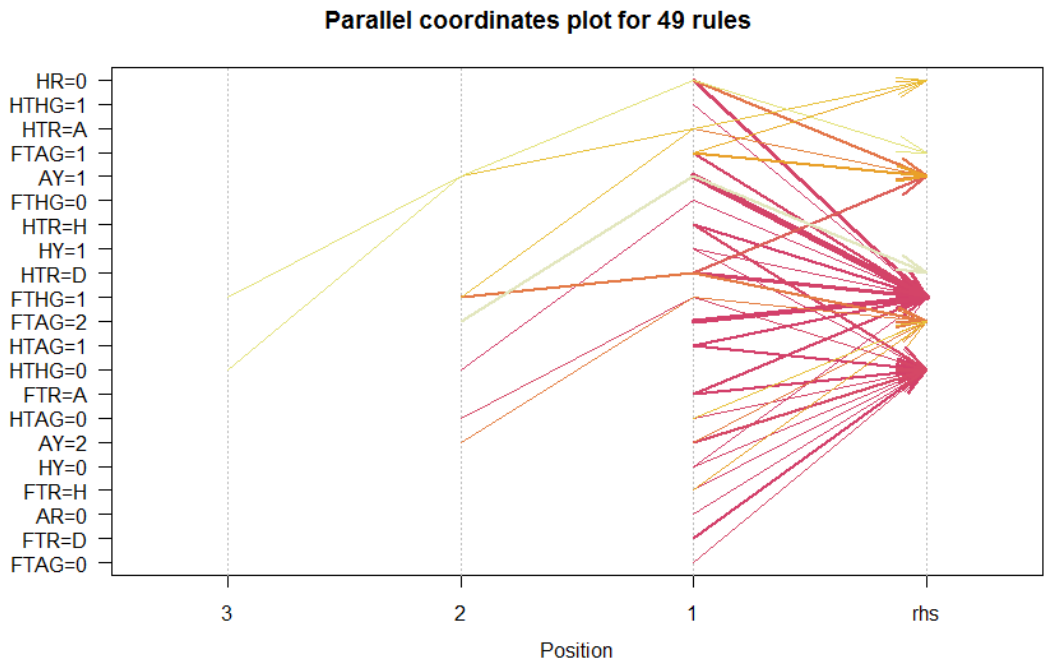
3.1. Κανόνες A priori

Για αρχή και για να βρεθούν οι κανόνες οι οποίοι εμφανίζονται πιο συχνά, ορίσαμε τις τιμές των μέτρων ισχύος των κανόνων support και confidence σε 20% και 70%, αντίστοιχα. Μετά από αφαίρεση διπλότυπων προέκυψαν 42 κανόνες. Μερικοί [1]από αυτούς είναι:

Rule		Support	Confidence	Lift
RHS	**LHS**			
1. {HTHG=0,HTAG=0}	{HTR=D}	0.3157895	1	2.261905
2. {HTR=H,HR=0}	{FTR=H}	0.2394737	0.7913043	1.915259
3. {FTR=H,HTAG=0,HR=0}	{FTAG=0}	0.2315789	0.7154472	2.210325
4. {HTAG=0,HTR=H}	{FTR=H}	0.2210526	0.7924528	1.918039
5. {HTAG=0,HTR=H,HR=0}	{FTR=H}	0.2105263	0.8	1.936306
6. {FTR=H,HTAG=0,AR=0}	{FTAG=0}	0.2052632	0.7155963	2.210785

Πίνακας 1 Κάνόνες apriori με υψηλή συχνότητα

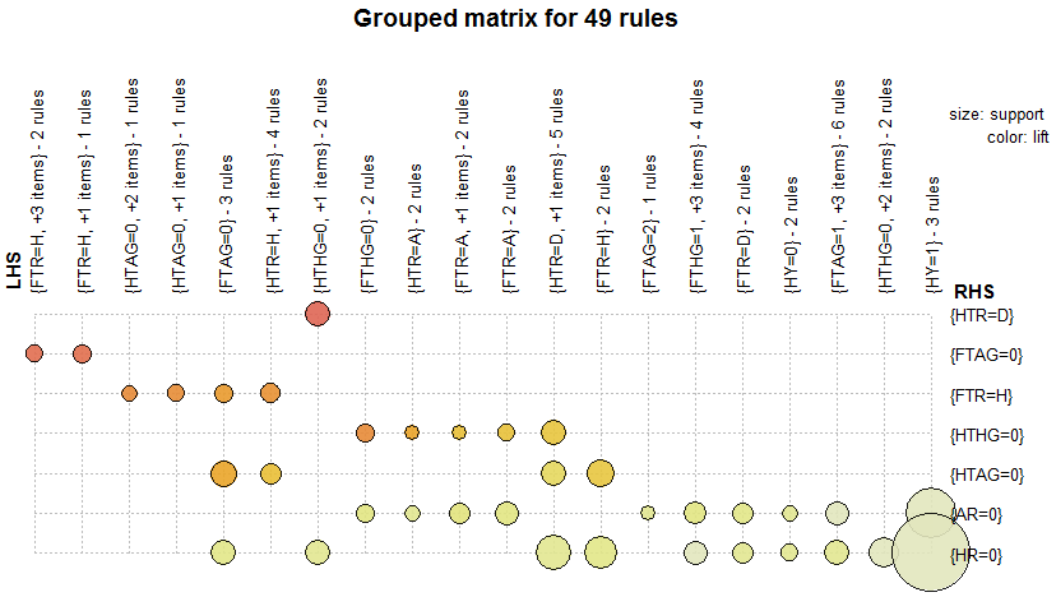
Από τους παραπάνω, ενδιαφέρον παρουσιάζει ο κανόνας με αριθμό 4 ο οποίος αναφέρει ότι στο 79% των παιχνιδιών όπου ο δείκτης των γκολ της φιλοξενούμενης στο ημίχρονο (HTAG) είναι μηδέν (0) και το αποτέλεσμα ημιχρόνου (HTR) είναι νίκη της γηπεδούχου, τότε το τελικό αποτέλεσμα του αγώνα (FTR) ισούται με νίκη της γηπεδούχου ομάδας.
Στο σύνολό τους, οι 42 κανόνες μπορούν να αναπαρασταθούν γραφικά ως εξής:



Εικόνα 2 Διάγραμμα παράλληλων συντεταγμένων

Στο διάγραμμα παράλληλων συντεταγμένων, η πορεία του κάθε βέλους αντιπροσωπεύσει το σύνολο (RHS) και το τέλος του βέλους, τον κανόνα (LHS). Παρατηρούμε ότι η πλειοψηφία των κανόνων, έχουν ως "συμπέρασμα" τον δείκτη των γκολ της γηπεδούχου στο ημίχρονο να είναι μηδέν (0) και τα γκολ της γηπεδούχου στο τελικό αποτέλεσμα του αγώνα να είναι ένα (1).

[1] Όλοι οι εξαγόμενοι κανόνες, μαζί με αλλαγές στις τιμές των μέτρων αξιολόγησης, μπορούν να βρεθούν στο σύνδεσμο: https://1drv.ms/w/s!AqbCaGLnF5dXnGMiez0L43DJuIEc



Εικόνα 3 Απεικόνιση βάση ομαδοποίησης

Στην απεικόνιση βάση ομαδοποίησης των κανόνων, η ομαδοποίηση έχει γίνει αυτόματα με clustering και αναπαριστάτε με το σχήμα ενός κύκλου για κάθε περίσταση. Το μέγεθος του κύκλου δηλώνει την τιμή της υποστήριξης (Support) του εκάστοτε κανόνα, ενώ το χρώμα τη τιμή του μέτρου Lift (η τάση προς το σκούρο δηλώνει και μεγαλύτερη τιμή).

Εκτός από τους κανόνες με μεγάλη συχνότητα, εξετάζουμε την περίπτωση για τους κανόνες με μικρή συχνότητα. Για την περίπτωση αυτή, επιλέξαμε να εξετάσουμε κανόνες οι οποίοι αφορούν τους διαιτητές των παιχνιδιών, καθώς είναι και ο παράγοντας με το περισσότερο ενδιαφέρον ανάμεσα στους φιλάθλους.

Rules		Support	Confidence	Lift
RHS	**LHS**			
{AS=12,HF=7}	{Referee=A Marriner}	0.01053	1	15.83333
{AF=13,HC=6}	{Referee=M Clattenburg}	0.01053	0.8	10.13333
{FTHG=1,HTAG=0,HF=4}	{Referee=C Pawson}	0.01053	0.8	10.85714
{FTHG=2,HTAG=0,HF=14}	{Referee=J Moss}	0.01053	0.8	12.16000
{FTAG=1,HTR=A,AF=12}	{Referee=M Oliver}	0.01053	0.8	11.69231
{FTR=D,HST=4,AST=5}	{Referee=C Pawson}	0.01053	0.8	10.85714
{HTR=D,HST=4,AST=5}	{Referee=C Pawson}	0.01053	0.8	10.85714
{HTR=H,AF=10,AY=2}	{Referee=M Clattenburg}	0.01053	0.8	10.13333
{FTAG=0,HF=11,AY=0}	{Referee=M Dean}	0.01053	0.8	9.21212
{FTAG=1,HTHG=0,HTAG=1,AF=12}	{Referee=M Oliver}	0.01053	0.8	11.69231
{FTHG=1,HTR=D,HST=6,HR=0}	{Referee=M Oliver}	0.01053	0.8	11.69231
{FTR=D,HTR=D,HST=4,AC=6}	{Referee=C Pawson}	0.01053	0.8	10.85714
{FTR=D,HTHG=0,HTR=D,AC=6}	{Referee=C Pawson}	0.01053	0.8	10.85714
{FTR=D,HTAG=0,HTR=D,AC=6}	{Referee=C Pawson}	0.01053	0.8	10.85714
{FTR=D,HTHG=0,HTAG=0,AC=6}	{Referee=C Pawson}	0.01053	0.8	10.85714
{FTR=H,HTAG=0,HF=11,AY=0}	{Referee=M Dean}	0.01053	0.8	9.21212

Πίνακας 2 Κάνόνες apriori με χαμηλή συχνότητα

Με την ίδια διαδικασία, απορρίφθηκαν οι πλεονάζοντες κανόνες και απέμειναν 17 κανόνες οι οποίοι περιέχουν 6 από τους 19 συνολικά διαιτητές της Premier League της σεζόν 2016. Ενώ θα μπορούσαμε να πάρουμε τους κανόνες έναν προς έναν και να τους αναλύσουμε, θεωρούμε καλύτερη τη διαγραμματική προσέγγιση συναρτήσει των κανόνων για την συγκεκριμένη περίπτωση, καθώς παρουσιάζεται μια καλύτερη συνολική εικόνα.

Εικόνα 4 Γράφημα κανόνων με χαμηλή συχνότητα

Παρατηρείτε ότι υπάρχουν δύο διαιτητές (M. Clattenburg, A. Marinner) οι οποίοι συνδέονται σε παιχνίδια για τα οποία ο αλγόριθμος εξήγαγε κανόνες με τον ένα από αυτούς να περιέχει αποτέλεσμα ημιχρόνου υπέρ της γηπεδούχου και αρκετά μεγάλο αριθμό φάουλ και των δύο αντιπάλων.

Τα στατιστικά παιχνιδιών των διαιτητών M. Oliver, M. Dean, J. Moss και C. Pawson δείχνουν να συνδέονται έχοντας σαν "κέντρο" έναν αριθμό παιχνιδιών στα οποία διαιτητής ήταν ο C. Pawson.

3.2. Διαβαθμιζόμενη Ενδυνάμωση

Αρχίζοντας την εκπαίδευσης του gradient μοντέλου, επιλέξαμε τα τρία σενάρια για τα οποία θα παρουσιαστούν αποτελέσματα.

- Σημαντικότητα [2]δεικτών που οδηγούν σε νίκη της γηπεδούχου ομάδας
- Σημαντικότητα δεικτών που οδηγούν σε νίκη της φιλοξενούμενης ομάδας
- Σημαντικότητα δεικτών που οδηγούν σε ισοπαλία

[2] Η σημαντικότητα ενός δείκτη ορίζεται ως το ποσοστό στο οποίο συμβάλει η συγκεκριμένη ανεξάρτητη μεταβλητή για να προβλεφθεί η εξαρτημένη μεταβλητή

Η απόδοση του κάθε μοντέλου για κάθε υπόθεση εκλέχθηκε με τη δημιουργία τριών κατηγοριών και βάση εξαγόμενης ποσοστιαίας πιθανότητας. Αναλυτικά:

- Για προβλέψεις με πιθανότητα μεγαλύτερη ή ίση από 0.7 θεωρήσαμε ότι το μοντέλο έχει προβλέψει, σε ανεκτά όρια, το τελικό αποτέλεσμα του αγώνα. (Success)

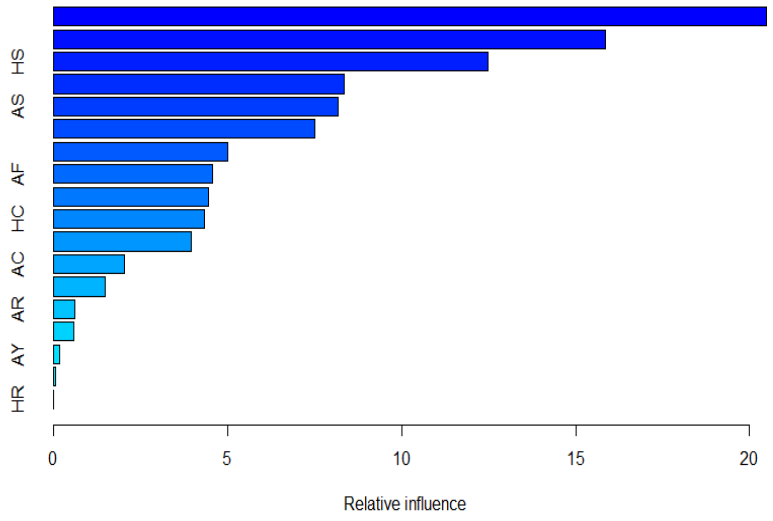
- Για προβλέψεις, με πιθανότητα μικρότερη από 0.7, θεωρήσαμε ότι το μοντέλο απέτυχε να προβλέψει σωστά το τελικό αποτέλεσμα του αγώνα. (Low Prediction)

Εκτός από τις δύο κατηγορίες υπάρχει και αυτή που στην οποία το μοντέλο απέτυχε στο να αναγνωρίσει την νίκη είτε με μεγάλο είτε με μικρό ποσοστό. (Other Result)

Να σημειώσουμε ότι για την επιλογή ενός βέλτιστου αριθμού δέντρων κατηγοριοποίησης, επιλέξαμε να λάβει μέρος έλεγχος σταυρωτής επικύρωσης (cross validation) στα δεδομένα σε συνέχεια των παραμετροποιήσεων [3]του μοντέλου.

3.2.1. Υπόθεση νίκης γηπεδούχου ομάδας

Variable	Importance (%)
HTR	20.48
AwayTeam	15.86
HS	12.49
HomeTeam	8.35
AS	8.17
AST	7.49
HST	5.00
AF	4.55
Referee	4.44
HC	4.33
HTAG	3.96
AC	2.03
HF	1.47
AR	0.59
HTHG	0.57
AY	0.17
HY	0.04
HR	0.00



Εικόνα 5 Σχετική συμαντικότητα ανεξάρτητων μεταβλητών - Υπόθεση νίκης γηπεδούχου

Παρατηρούμε ότι η μεταβλητή με τη μεγαλύτερη σημαντικότητα στην πρόβλεψη της τελικής νίκης της γηπεδούχου είναι το αποτέλεσμα του ημιχρόνου (HTR) με 20,5 % με την φιλοξενούμενη ομάδα να ακολουθεί με 15,8% και τα συνολικά σουτ της γηπεδούχου να ακολουθούν. Οι κόκκινες κάρτες της φιλοξενούμενης και γηπεδούχου (AR, HR), τα γκολ ημιχρόνου γηπεδούχου (HTHG), οι κίτρινες κάρτες φιλοξενούμενης και γηπεδούχου (AY, HY) παρουσίασαν σημαντικότητα κάτω από 1%.

Βάση των παραπάνω, δίνουμε τα παρακάτω περιγραφικά στατιστικά για την πρόβλεψη

[3] Cross validation folds = 10, ρυθμός μάθησης = 0,001

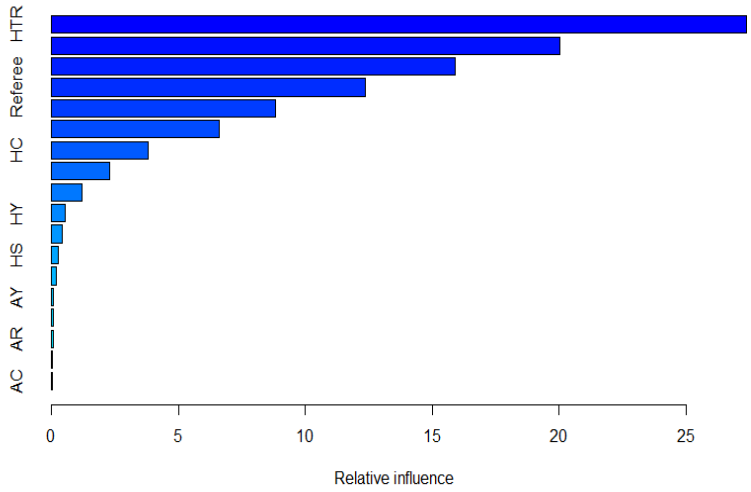
Probability	Result	Prediction Success Rate
Min. : 0.03988 1st Qu. : 0.17203 Median : 0.35569 Mean : 0.41302 3rd Qu. : 0.64837 Max. : 0.95466	Home: 157 Away: 116 Draw: 107	Success: 87 Low Prediction: 70 Other Result: 190

Πίνακας 3 Περιγραφικά στατιστικά πιθανοτήτων πρόβλεψης και επιτυχίες ανα κατηγορία - Υπόθεση νίκης γηπεδούχου

και εκφράζουμε το μοντέλο με απόδοση στην πρόβλεψη 55,4%.

3.2.2. Υπόθεση νίκης φιλοξενούμενης ομάδας

Variable	Importance (%)
HTR	27.36
HomeTeam	20.03
AwayTeam	15.89
Referee	12.36
HST	8.80
AST	6.61
HC	3.81
HTAG	2.28
HR	1.20
HY	0.52
HTHG	0.42
HS	0.28
AS	0.19
AY	0.09
HF	0.06
AR	0.05
AF	0.04
AC	0.01



Εικόνα 6 Σχετική σημαντικότητα ανεξάρτητων μεταβλητών – Υπόθεση νίκης φιλοξενούμενης

Πρώτη σε σημαντικότητα μεταβλητή αποδείχθηκε ο δείκτης του αποτέλεσμα του ημιχρόνου (HTR) με 27,36 % με τις ομάδες καθ' εαυτών (20,03% για HomeTeam και 15,89% για AwayTeam), ενώ τέταρτη σε σημαντικότητα μεταβλητή είναι ο διαιτητής (Referee) του παιχνιδιού με ποσοστό 12,36%. Σημαντικό είναι ότι από τις 18 ανεξάρτητες μεταβλητές, οι 9 (50%) παρουσίασαν σημαντικότητα κάτω του 1% με αυτές να περιλαμβάνουν τους δείκτες μεταξύ άλλων των κόκκινων καρτών της φιλοξενούμενης και γηπεδούχου (AR), των γκολ ημιχρόνου γηπεδούχου (HTHG) και των κίτρινων καρτών της φιλοξενούμενης (AY).

Probability	Result	Prediction Success Rate
Min. :0.0550 1st Qu.:0.1340 Median :0.2207	Home: 157 Away: 116 Draw: 107	Success: 35 Low Prediction: 81 Other Result: 264

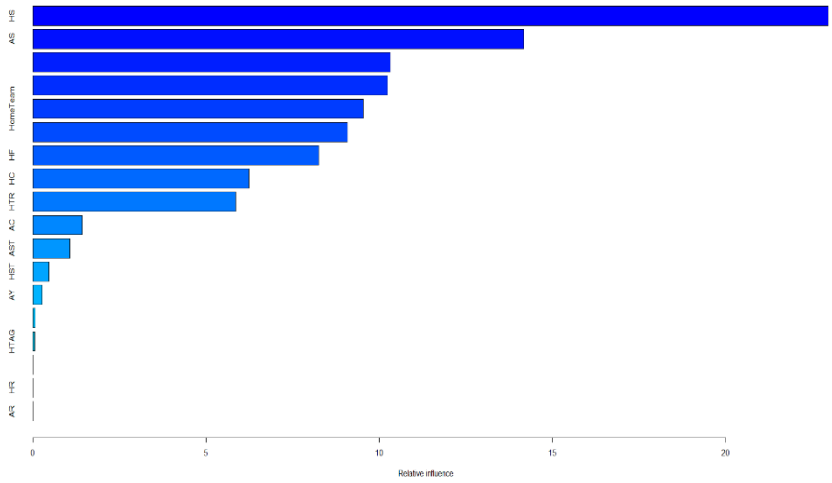
Mean :0.3046 3rd Qu.:0.4230 Max. :0.8713		

Πίνακας 4 Περιγραφικά στατιστικά πιθανοτήτων πρόβλεψης και επιτυχίες ανα κατηγορία - Υπόθεση νίκης φιλοξενούμενης

εκφράζουμε το μοντέλο με απόδοση στην πρόβλεψη 30,1%.

3.2.3. Υπόθεση ισοπαλίας ομάδων

Variable	Importance (%)
HS	22.95
AS	14.17
AwayTeam	10.31
AF	10.24
HomeTeam	9.54
Referee	9.08
HF	8.24
HC	6.24
HTR	5.86
AC	1.43
AST	1.06
HST	0.46
AY	0.27
HY	0.08
HTAG	0.05
HTHG	0.00
HR	0.00
AR	0.00



Εικόνα 7 Σχετική σημαντικότητα ανεξάρτητων μεταβλητών – Υπόθεση ισοπαλίας

Σε αυτό το σενάριο παρατηρούμε ότι η μεταβλητή με τη μεγαλύτερη σημαντικότητα στην πρόβλεψη της ισοπαλίας είναι τα σουτ των γηπεδούχων (HS) με 23% ενώ ακολουθούν τα σουτ της φιλοξενούμενης (AS - 14%), η φιλοξενούμενη ομάδα (AwayTeam 10,3%), και τα φάουλ της φιλοξενούμενης (AF 10.2%). Οι κόκκινες κάρτες της φιλοξενούμενης και γηπεδούχου (AR, HR), τα γκολ ημιχρόνου γηπεδούχου (HTHG), οι κίτρινες κάρτες φιλοξενούμενης και γηπεδούχου (AY, HY) παρουσίασαν σημαντικότητα κάτω από 1%, μεταξύ άλλων.

Probability	Result	Prediction Success Rate
Min. :0.1965 1st Qu.:0.2456 Median :0.2772 Mean :0.2806 3rd Qu.:0.3118 Max. :0.4097	Home: 157 Away: 116 Draw: 107	Success: 0 Low Prediction: 107 Other Result: 273

Πίνακας 5 Περιγραφικά στατιστικά πιθανοτήτων πρόβλεψης και επιτυχίες ανα κατηγορία - Υπόθεση ισοπαλίας

Στην περίπτωση της πρόβλεψης της ισοπαλίας, ο αλγόριθμος μαζί με το εξαγόμενο μοντέλο απέτυχε στο να παράγει κάποιο ποσοστό επιτυχίας μη έχοντας κάποια πιθανότητα σωστής πρόβλεψης πάνω από 70%. Παρατηρούμε ότι η μεγαλύτερη εξαγόμενη είναι ποσοστό της τάξης του 40%. Πρώιμες αιτιολογίες για την αποτυχία είναι το μικρό διαθέσιμο δείγμα πάνω στο οποίο εκπαιδεύτηκε το μοντέλο και η γενικότερη πολυπλοκότητα των παραγόντων από τους οποίους προκύπτει η ισοπαλία σε κάποιον αγώνα ποδοσφαίρου.

4. Συμπεράσματα

Έχοντας μία συνολική εικόνα, όσον αφορά τη δυνατότητα της διαβαθμιζόμενης ενδυνάμωσης να προβλέψει την έκβαση ενός αγώνα, παρατηρήθηκε ότι η περίπτωση με την μεγαλύτερη επιτυχία ήταν αυτή της πρόβλεψης της νίκης της γηπεδούχου ομάδας (55,4% απόδοση μοντέλου) έναντι της νίκης της φιλοξενούμενης (30,1% απόδοση μοντέλου) ενώ στη τελευταία περίπτωση υπήρξε αδυναμία στην εξαγωγή ενός ικανοποιητικού ποσοστού επιτυχίας καθώς κατάφερε να αποδώσει μέγιστο ποσοστό σωστής πιθανότητας πρόβλεψης 40,9% (στο όριο του 70%) και χρίζει επιπλέον έρευνας στον τομέα του συνόλου δοκιμής και εκπαίδευσης. Γενικότερα, το αποτέλεσμα ημιχρόνου παίζει καταλυτικό ρόλο (20,5%) όπως επίσης και τα συνολικά σουτ του γηπεδούχου (15,8%) που είναι δυνατότητα απειλής για τον αντίπαλο. Η σχετική σημαντικότητα των ανεξάρτητων μεταβλητών κατέδειξε ότι στις πρώτες θέσεις, για την πρόβλεψη του τελικού αποτελέσματος, έρχεται το αποτέλεσμα του ημιχρόνου, οι ομάδες αυτές καθ' αυτές, τα συνολικά σουτ και ο διαιτητής κάθε αγώνα.

Τα αποτελέσματα του αλγορίθμου a priori εξήγαγαν μία κατάσταση στην οποία παρατηρήθηκαν κανόνες συσχέτισης, έχοντας ως "συμπέρασμα" τον δείκτη των γκολ της γηπεδούχου στο ημίχρονο να είναι μηδέν (0) και τα γκολ της γηπεδούχου στο τελικό αποτέλεσμα του αγώνα να είναι ένα (1), μέρος των οποίων επιβεβαιώνουν τα αποτελέσματα της διαβαθμιζόμενης ενδυνάμωσης. Με την μελέτη που κάνουμε μπορούμε πιο συγκεκριμένα να πούμε πως εμφανίζεται μεγάλη δυνατότητα παρακολούθησης από τις ομάδες της Premier League καθώς η εξισορρόπηση του παθητικού των γκολ σε μηδέν από τον γηπεδούχο του δίνει εμφανές προβάδισμα απέναντι στην αντίπαλη ομάδα για την νίκη στον αγώνα. Άρα είναι ευδιάκριτο ότι οι ομάδες πρέπει να προσαρμόσουν και βελτιώσουν την άμυνά τους στο πρώτο τέταρτο του γηπέδου και να εξαλείψουν τους κινδύνους για την εστία τους. Ακόμα, στη περίπτωση όπου κανόνες οι οποίοι αφορούν το σύνολο των διαιτητών, συχνότερη εμφάνιση είχαν οι: M. Oliver, M. Dean, J. Moss και C. Pawson. Ονόματα τα οποία με μια σύντομη αναζήτηση στη Google, εμφανίζουν αρκετά δημοσιεύματα με χαρακτηριστικό περιεχόμενο.

Βιβλιογραφικές Αναφορές

[1]. Carnevale, R., (2015). Data Stuff – Gradient Boosting Part 1.

[2]. Friedman, J. H., (2001). Greedy Function Approximation: A Gradient Boosting Machine.

[3]. FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R., (2000). Additive Logistic Regression: a Statistical View of Boosting.

[4]. Rogozhnikov, A., (2016). Brilliantly Wrong - Gradient Boosting.

[5]. Βαρυτιμίδης, Χ. Ι., (2008). Ανίχνευση αντικειμένων και ημιαυτόματος χαρακτηρισμός εικόνων. Αθήνα

[6]. Μαργαρίτης, Σ., Θεολόγος, Κ. & Σπύρου, Ν., (2016). Εφαρμογές Υπολογιστικής Νοημοσύνης στις Γνωστικές Επιστήμες (Μελέτη περίπτωσης: Συσχέτιση Χαρακτηριστικών Προσωπικότητας και Συναισθηματικής Νοημοσύνης). Πάτρα

[7]. Τόγιας, Π. & Μαργαρίτης, Σ., (2016). Data Analytics και Ευφυή Συστήματα Πρόβλεψης Δεδομένων σε Χρονοσειρά. Αθήνα

Ευχαριστίες

Θα θέλαμε να ευχαριστήσουμε τη κα. Μότση Δ. Κατερίνα για την βοήθειά της στην προεπεξεργασία των δεδομένων, στο κομμάτι της εξόρυξης γνώσης, όσο και για τα σχόλιά της τα οποία βοήθησαν στην βελτιστοποίηση του κειμένου της έρευνας.