Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Δ.Π.Μ.Σ. Επιστήμη Δεδομένων & Μηχανική Μάθηση

Εξαμηνιαία Εργασία στην Εξόρυξη Γνώσης από Δεδομένα

Γιώργος Βερνίκος Α.Μ.: 03400005

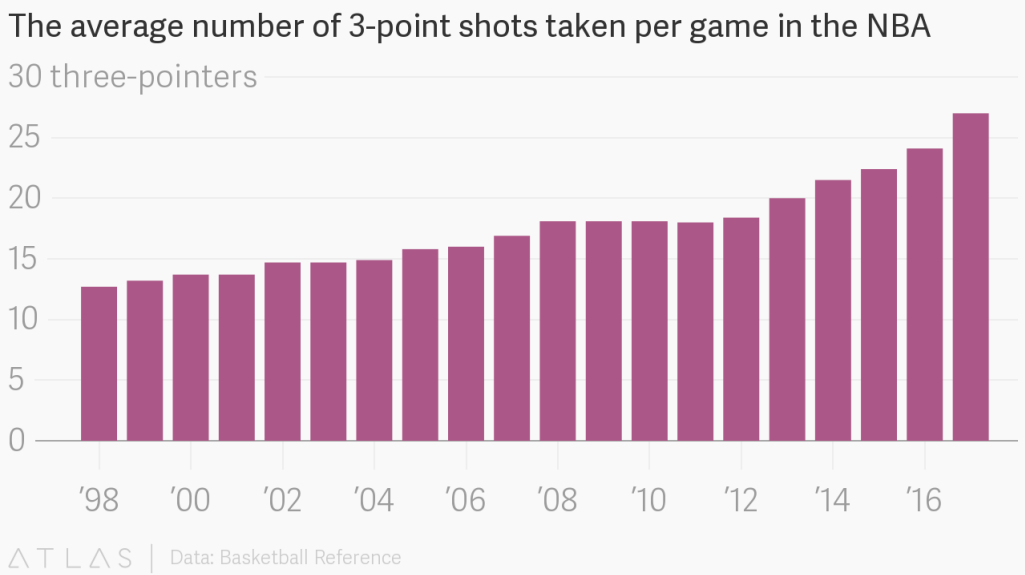
Νικηφόρος Μανδηλαράς Α.Μ.: 03400022

Χρήστος Σπυρόπουλος Α.Μ.: 03400035

Παραδοτέα 3/3/19

1 ΕΙΣΑΓΩΓΗ

Στο πέρασμα των χρόνων η εξέλιξη του αθλητισμού έμελλε να συνδεθεί στενά με την εξέλιξη της τεχνολογίας, συνεισφέροντας έτσι στην αναβάθμιση των ίδιων των αθλημάτων, αλλά και στην εμβάθυνση στο τρόπο που τα προσεγγίζουν, τόσο οι επαγγελματίες αθλητές όσο και το φιλοθεάμον κοινό. Αυτή η εξέλιξη είχε ως αποτέλεσμα την εισαγωγή της στατιστικής και της ανάλυσης δεδομένων σταδιακά σε όλο και περισσότερα αθλήματα, είτε ατομικά είτε ομαδικά. Σε κάθε άθλημα η επίδραση της ανάλυσης δεδομένων διαφέρει, καθώς είναι πολλοί οι εμπλεκόμενοι παράγοντες. Μπορούμε όμως να διακρίνουμε κάποιους τομείς όπου χρησιμοποιείται κατά κόρον, όπως είναι η διαμόρφωση στρατηγικών, η βελτιστοποίηση της τεχνικής, η δημιουργία του ρόστερ μιας ομάδας, η επιλογή του είδους των προπονήσεων κ.λ.π. Μάλιστα σε μερικές περιπτώσεις έρχονται στατιστικά αποτελέσματα να αλλάξουν τον τρόπο που παίζεται το παιχνίδι, όπως για παράδειγμα στο χώρο του μπάσκετ, όπου μελέτες έχουν καταλήξει πως το σουτ μέσης απόστασης έχει χαμηλότερο αναμενόμενο κέρδος έναντι άλλων επιλογών στην επίθεση, όπως για παράδειγμα το τρίποντο. Αυτή η παρατήρηση είχε ως αποτέλεσμα την παρακάτω εικόνα.



Εικόνα 1: *Ο μέσος αριθμός τριπόντων που εκτελούνται κατά τη διάρκεια ενός αγώνα NBA.*

Ο ρόλος της στατιστικής ανάλυσης στα σπορ μοιάζει να έχει εδραιωθεί πλήρως σήμερα, ώστε κάθε ομάδα που ασκεί πρωταθλητισμό σε υψηλό επίπεδο να συμπεριλαμβάνει στα επιτελεία της αναλυτές . Στο μπάσκετ αυτή η εξέλιξη συμβαίνει την τελευταία δεκαετία όπου όπως παρατηρούμε στο

παραπάνω διάγραμμα είναι πολύ πιο έντονη η χρήση του τριπόντου, σε βαθμό τέτοιο που μπορεί να επιφέρει και αλλαγές στους κανονισμούς του αθλήματος.

2 Σύνολο Δεδομένων

Το σετ δεδομένων που επιλέξαμε να μελετήσουμε προέρχεται από την Euroleague και την SAP και αφορά αγώνες για τις season 2016-17 και 2017-2018. Πιο συγκεκριμένα περιέχει δεδομένα για κάθε φάση όλων των παιχνιδιών της Euroleague τα τελευταία δύο χρόνια. Το csv αρχείο που έχουμε διαθέσιμο αποτελείται από 121 χιλιάδες εγγραφές η κάθε μία από αυτές έχει 51 στήλες οι οποίες περιέχουν πληροφορία, όπως επιτιθέμενος παίκτης, παίκτες που συνεισφέρουν στη φάση, επιτιθέμενη/αμυνόμενη ομάδα, ενέργειες που πραγματοποιούνται, έκβαση καθώς και πολλές ακόμα. Όμως, όπως ήταν αναμενόμενο χρειάστηκε να επέμβουμε αρκετά στο σύνολο δεδομένων μας όπως θα αναφέρουμε και στην συνέχεια. Σκοπός μας λοιπόν σε αυτή την εργασία είναι η εξόρυξη γνώσης στο χώρο των sports analytics και ειδικότερα στο χώρο της καλαθοσφαίρισης.

3 Σχετικές μέθοδοι

Μια κλασσική μέθοδος που χρησιμοποιείται ευρέως στο ποδόσφαιρο μελετά τις ενέργειες των παικτών όταν δεν έχουν οι ίδιοι την μπάλα στην κατοχή τους. Δεδομένου ότι αναφερόμαστε σ' ένα άθλημα όπου οι αγώνες διαρκούν 90 λεπτά και στο οποίο συμμετέχουν 22 παίκτες καταλαβαίνουμε πως ο κάθε ποδοσφαιριστής είναι κάτοχος της μπάλας το πολύ για 3 λεπτά κατά μέσο όρο. Συνεπώς η απορία που γεννάται, είναι, πως ενεργεί τον υπόλοιπο χρόνο και πώς μπορούν να βελτιστοποιηθούν οι κινήσεις προς όφελος της ομάδας. Τα μοντέλα λοιπόν που δημιουργούνται σε αυτή την περιοχή προσμετρούν τη θέση των παικτών στο γήπεδο, την ταχύτητα τους και την απόσταση που βρίσκονται από την μπάλα. Έτσι το μοντέλο εκτιμά την αξία της κάθε περιοχής του γηπέδου με στόχο να δώσει κατεύθυνση στους παίκτες σε ποιες περιοχές πρέπει να κινούνται, που και πότε πρέπει να ανοίξουν χώρους αλλά και ποιους χώρους πρέπει να διαφυλάξουν.

Μια άλλη κλασσική μέθοδος στα sports analytics είναι η προσπάθεια πρόβλεψης της πιθανότητας να τραυματιστεί ένας παίκτης στο επερχόμενο ματς. Τα μοντέλα σε αυτή την περίπτωση χρησιμοποιούν δεδομένα από αγώνες και όπως φαίνεται τα χαρακτηριστικά που επηρεάζουν περισσότερο ένα τραυματισμό είναι και τα αναμενόμενα, όπως η μέση ταχύτητα που τρέχει ένας παίκτης κατά τη διάρκεια ενός αγώνα, η μέση απόσταση που καλύπτει, τα παιχνίδια που έχει παίξει και κατά μέσο όρο τα λεπτά που έχει αγωνιστεί στο εκάστοτε ματς. Βέβαια τα μοντέλα σε κάθε άθλημα θα διαφέρουν αλλά η κεντρική ιδέα είναι ίδια ανεξαρτήτως του αθλήματος.

4 Προεπεξεργασία Δεδομένων

Η προεπεξεργασία των δεδομένων είναι ένα πολύ σημαντικό κομμάτι πριν προχωρούμε με την εξαγωγή συμπερασμάτων, έτσι και στην δική μας περίπτωση ήταν αρκετά αυτά που έπρεπε να κάνουμε. Αρχικά απομακρύναμε στήλες που περιείχαν πληροφορία είτε επικαλυπτόμενη είτε που μπορούσε να εξαχθεί από άλλες στήλες, όπως οι συνθέσεις των ομάδων. Ακόμα απομακρύναμε στήλες που δεν παρουσίαζαν κάποιο ενδιαφέρον για την μελέτη μας όπως η ημερομηνία του παιχνιδιού καθώς και στήλες με πολλές κενές εγγραφές ή στήλες με σχεδόν μηδενική διακύμανση. Φροντίσαμε επίσης να απομακρύνουμε εγγραφές που είχαν μη ρεαλιστικές τιμές όπως για παράδειγμα σκορ αγώνα μικρότερο των 50 πόντων. Απ' την άλλη δημιουργήσαμε νέες μεταβλητές από τα δεδομένα που είχαμε τις οποίες θα χρειαστούμε στη συνέχεια για την εκπαίδευση του μοντέλο που θα χρησιμοποιήσουμε. Αυτές οι μεταβλητές είναι για παράδειγμα οι συνολικές ασίστ σε ένα παιχνίδι, οι συνολικοί αιφνιδιασμοί που εκδήλωσε μια ομάδα, ο αριθμός των τριπόντων που σούταρε, τα φάουλ που διέπραξε κ.λ.π. Προχωρήσαμε επίσης και σε διακριτοποιήση μεταβλητών όπως ο χρόνος στον οποίο εκδηλώνει επίθεση μια ομάδα.

5 Διερευνητική Ανάλυση

Σε αυτή την ενότητα θα παρουσιάσουμε κάποια ευρήματα που προέκυψαν έπειτα από μελέτη των δεδομένων.

Το πρώτο κομμάτι του σετ δεδομένων μας που έχει μεγάλο ενδιαφέρον είναι ο αριθμός των ασίστ του κάθε παίκτη αλλά και της κάθε ομάδας ξεχωριστά. Θα ξεκινήσουμε παρουσιάζοντας ένα πίνακα που δείχνει τους παίκτες με τις περισσότερες ασίστ.

	assist.name	team	count
1	Nick Calathes	Panathinaikos	418
2	Vassilis Spanoulis	Olympiacos	323
3	Konstantinos Sloukas	Fenerbahce	314
4	Kevin Pangos	Zalgiris	300
5	Luka Doncic	Real Madrid	286
6	Nando de Colo	CSKA	216
7	Sergio Llull	Real Madrid	209
8	Nikolaos Zisis	Brose	205
9	Thomas Heurtel	Barcelona	191

Εικόνα 2 : Οι εννέα παίκτες με τις περισσότερες ασίστ.

Από τον παραπάνω πίνακα μπορούν να βγουν ενδιαφέροντα συμπεράσματα. Αρχικά βλέπουμε ότι οι παίκτες με τις περισσότερες ασίστ είναι και οι εννέα ευρωπαίοι, και ακόμα πιο συγκεκριμένα στις πρώτες θέσεις βρίσκονται τρεις Έλληνες παίκτες, αποτέλεσμα που μπορεί να πηγάζει από πολλούς παράγοντες όπως, την ικανότητα του ίδιου του παίκτη, τον τρόπο με τον οποίο έχει μάθει να βλέπει το άθλημα και σίγουρα παίζει ρόλο και το προφίλ της ομάδας στην οποία αγωνίζεται.

```
team
Darussafaka          417
Unics                459
Valencia             513
Khimki               534
Unicaja              537
Galatasaray          566
Olympiacos           913
Panathinaikos        936
Crvena Zvezda        944
Milano               972
Efes                 995
Brose               1013
Maccabi TelAviv     1021
Fenerbahce          1071
Baskonia            1109
Barcelona           1113
CSKA                1117
Zalgiris            1125
Real Madrid         1195
```

Εικόνα 3 : Ο αριθμός των ασίστ που έχει βγάλει μια ομάδα.

Εδώ μετράμε τις ασίστ που έχει σημειώσει μία ομάδα ανεξάρτητα με το αν έχει παίξει μία ή δύο χρονιές στη Euroleague. Το πρώτο συμπέρασμα που μπορούμε να βγάλουμε και κυρίως για τις ομάδες που έχουν παίξει δύο χρονιές στη Euroleague είναι ότι οι ομάδες που είχαν τους παίκτες που ατομικά είχαν τις περισσότερες ασσίστ δεν βρίσκονται στην κορυφή της λίστας. Όπως για παράδειγμα ο Παναθηναικός και ο Ολυμπιακός. Στον αντίποδα η Ρεάλ Μαδρίτης και η Ζαλγκίρις Κάουνας βρίσκονται και στις πρώτες θέσεις. Αυτό μπορεί να εξηγηθεί από το γεγονός ότι ο Παναθηναϊκός και ο Ολυμπιακός στηρίζονται κυρίως στις προσπάθειες ενός παίκτη, ενώ ομάδες όπως η Ρεάλ και η Ζαλγκίρις μπορεί να έχουν και παίκτες ατομικά που να ξεχωρίζουν αλλά συγχρόνως να λειτουργούν αποτελεσματικά και ως ομάδα. Το επόμενο αποτέλεσμα που θα παραθέσουμε είναι ο λόγος ασίστ/λαθών. Η συγκεκριμένη μετρική χρησιμοποιείται κατά κόρον στο σύγχρονο μπάσκετ αφού ουσιαστικά εκφράζει την αναλογία μεταξύ των επιτυχημένων οργανωμένων επιθέσεων σε σχέση με τις επιθέσεις που τελείωσαν από λάθος της ομάδας. Καλύτερα αποτελέσματα σε αυτή τη μετρική συνεπάγονται καλύτερη λειτουργία της ομάδας κυρίως στο επιθετικό κομμάτι άρα και μεγαλύτερες πιθανότητες για συνολικά επιτυχημένη πορεία.

```
team
Unics              1.254098
Olympiacos         1.287729
Darussafaka        1.299065
Milano             1.357542
Maccabi TelAviv    1.398630
Panathinaikos      1.448916
Zalgiris           1.451613
Crvena Zvezda      1.452308
Barcelona          1.453003
Brose              1.453372
Efes               1.509863
CSKA               1.521798
Baskonia           1.529655
Valencia           1.549849
Fenerbahce         1.556686
Khimki             1.589286
Unicaja            1.598214
Galatasaray        1.650146
Real Madrid        1.788922
dtype: float64
```

Εικόνα 4 : Το ratio ασσίστ / λαθών για όλες τις ομάδες που αγωνίστηκαν στην Euroleague.

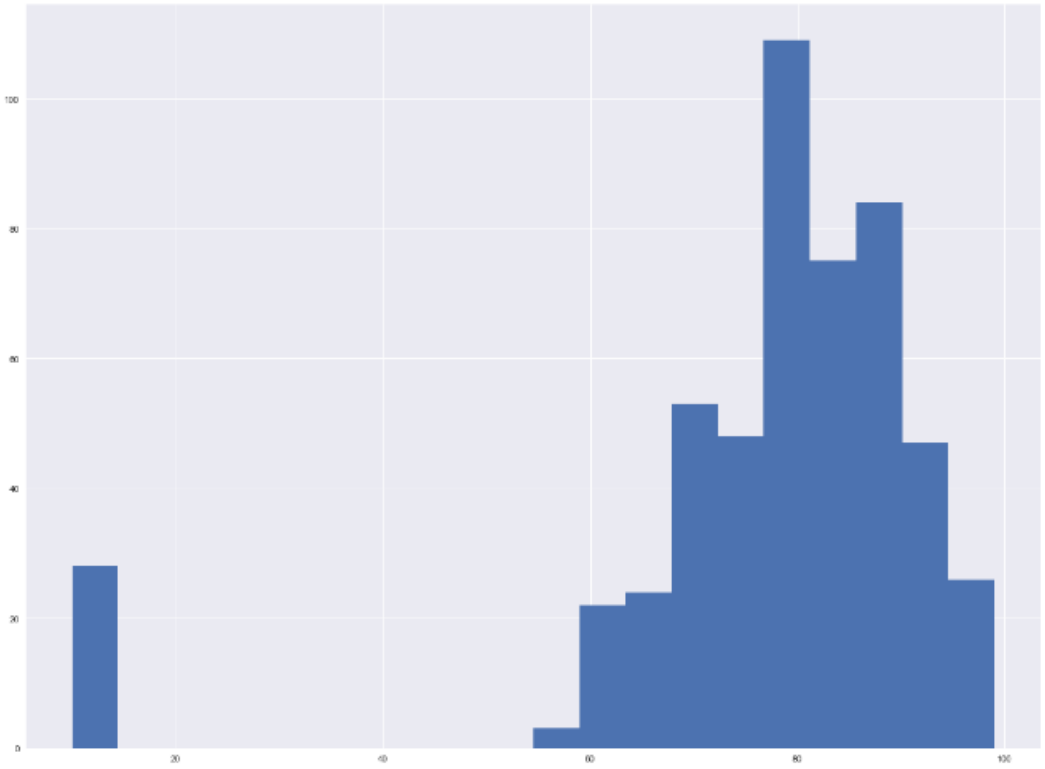
Εδώ τα αποτελέσματα μας δίνουν για ακόμα μια φορά την Ρεάλ Μαδρίτης στην κορυφή της λίστας. Από την άλλη βλέπουμε την Ζαλγκίρις να έχει πολύ χαμηλό ratio ενώ ήταν δεύτερη σε ασίστ γεγονός που μας δείχνει ότι μάλλον θα έχει κάνει πολλά λάθη. Ενώ όμως αυτή η μετρική θα μας έδινε ουσιαστικά μια εποπτεία και για τα λάθη που κάνει μια ομάδα η επόμενη εικόνα θα μας δείξει ότι κάτι τέτοιο δεν ισχύει.

```
team
Darussafaka        321
Valencia           331
Khimki             336
Unicaja            336
Galatasaray        343
Unics              366
Panathinaikos      646
Crvena Zvezda      650
Efes               659
Real Madrid        668
Fenerbahce         688
Brose              697
Olympiacos         709
Milano             716
Baskonia           725
Maccabi TelAviv    730
CSKA               734
Barcelona          766
Zalgiris           775
dtype: int64
```

Εικόνα 5 : Ο αριθμός των λαθών της κάθε ομάδας.

Η παραπάνω εικόνα μας δείχνει ποιες ομάδες κάνουν τα λιγότερα λάθη. Βλέπουμε λοιπόν ότι ο Παναθηναϊκός είναι η ομάδα με τα λιγότερα λάθη από αυτές που έχουν αγωνιστεί δύο σεζόν. Στο μοντέλο μας χρησιμοποιήσουμε τον αριθμό των λαθών και τον αριθμό των ασίστ σαν δύο διαφορετικά χαρακτηριστικά.

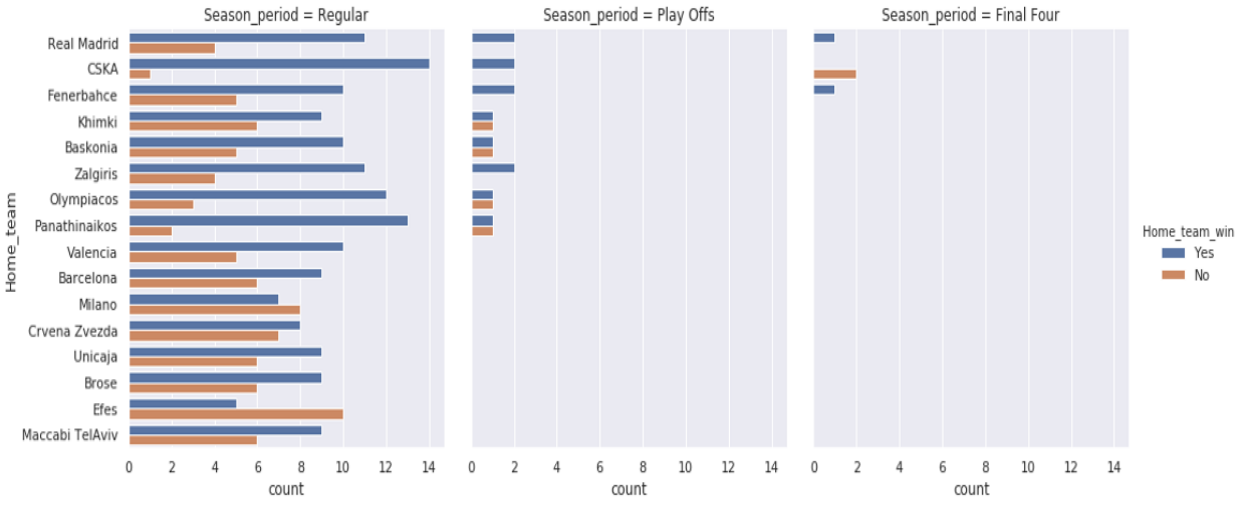
Τώρα ήρθε η ώρα να μιλήσουμε για τα αποτελέσματα μιας ομάδας είτε αυτή αγωνίζεται εντός είτε εκτός έδρας. Θα ξεκινήσουμε με ένα γράφημα που μας δείχνει τους πόντους που πετυχαίνει μία ομάδα όταν αγωνίζεται εντός έδρας.



Εικόνα 6 : Οι πόντοι που έβαλε η γηπεδούχος ομάδα βρίσκονται στον οριζόντιο άξονα.

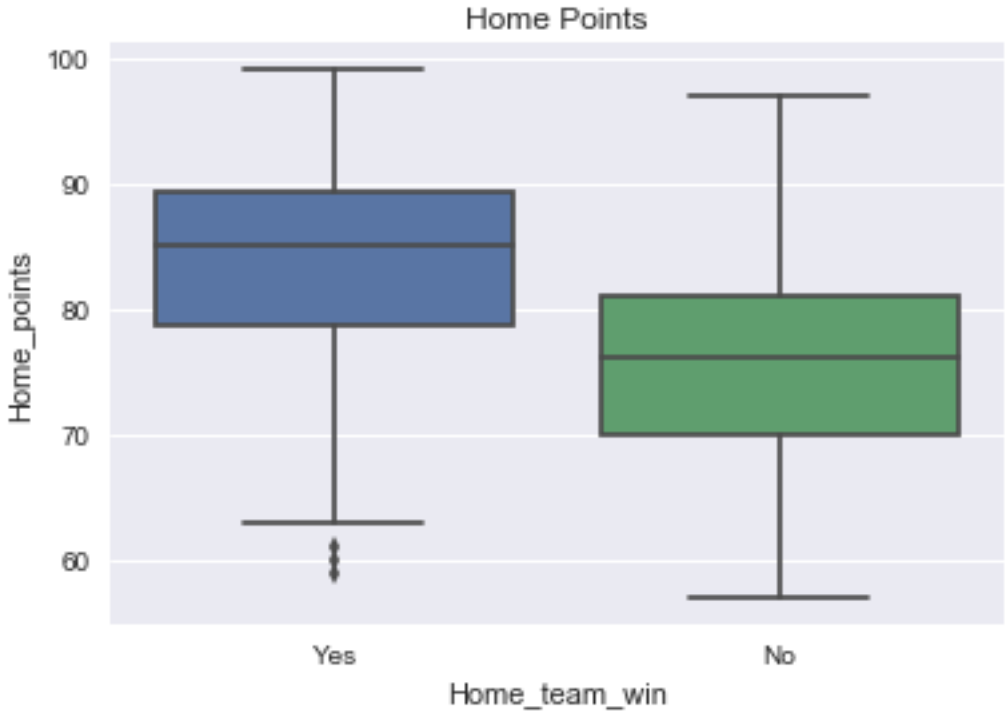
Από αυτή την εικόνα μπορεί να καταλάβει κανείς εύκολα ότι στο σετ δεδομένων υπάρχουν μερικές τιμές που πρέπει να αφαιρεθούν, αφού μια ομάδα είναι στατιστικά αδύνατον να έχει σκοράρει κάτω από 30 πόντους σε παιχνίδι που ολοκληρώθηκε ομαλά (Όλοι οι αγώνες της Euroleague είναι τέτοιοι).

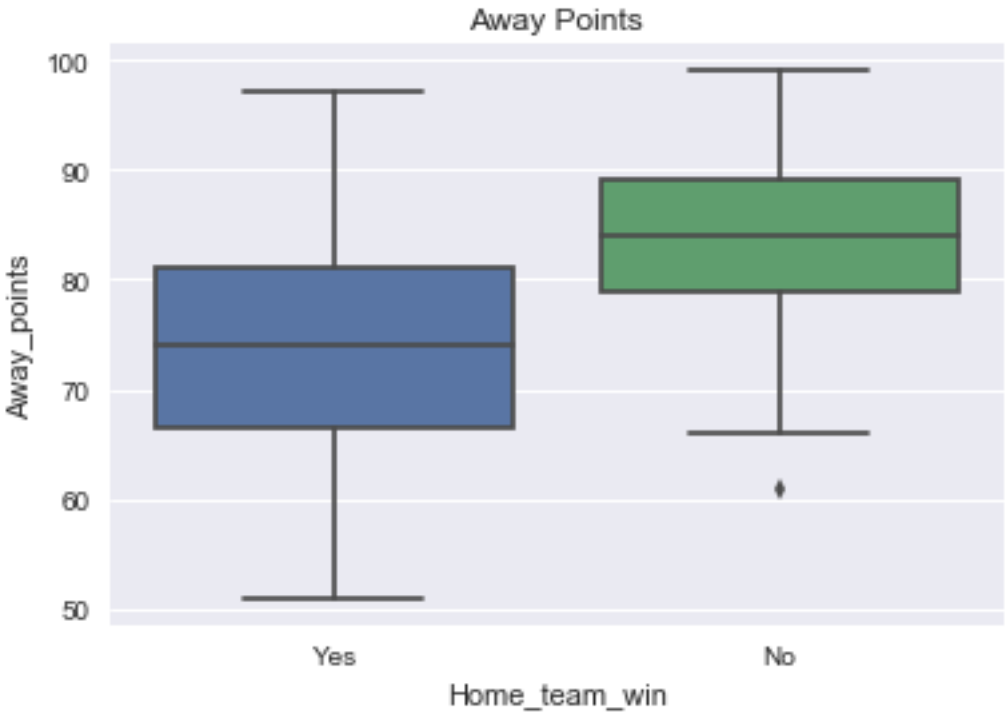
Ακόμα κάναμε ένα catplot έτσι ώστε να δούμε πόσες νίκες είχαν στην έδρα τους οι ομάδες στην κανονική περίοδο στα play-off και στο Final Four.



Εικόνα 8 : Αποτελέσματα εντός έδρας για όλες τις ομάδες.

Και τώρα θα περάσουμε σε ένα συμπέρασμα που βγάλαμε από την ανάλυση των δεδομένων, το οποίο φαντάζει λογικό αλλά δίνει πολύ μεγάλη πιθανότητα νίκης στην ομάδα που θα ακολουθήσει αυτήν τη στρατηγική για να νικήσει ένα παιχνίδι.

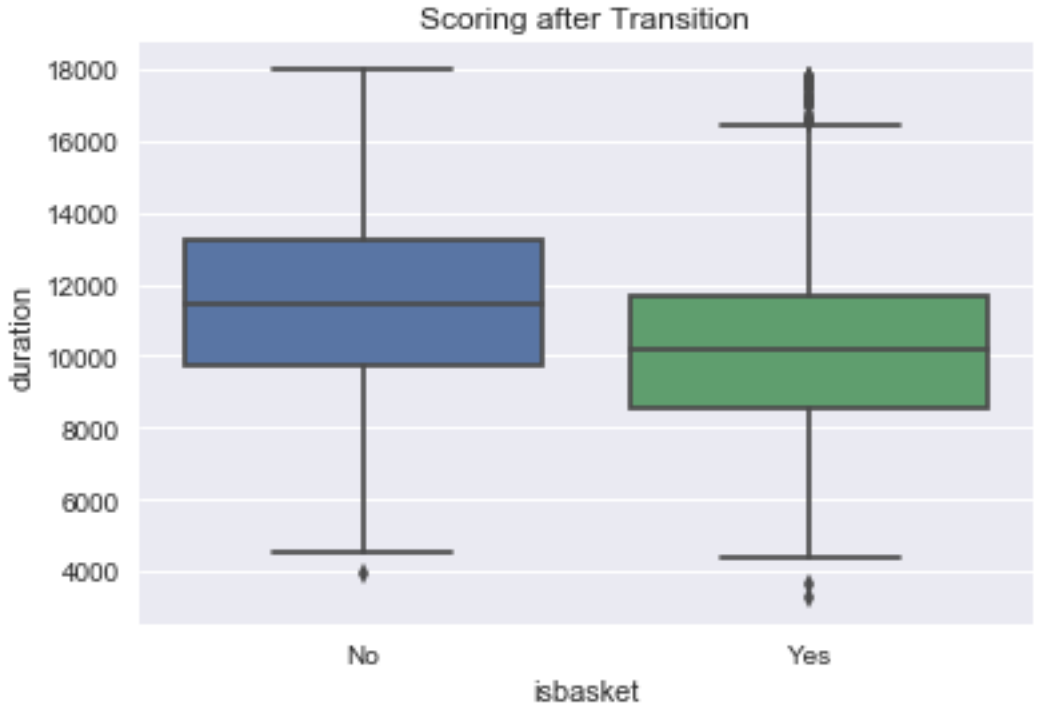




Εικόνα 9: Οι πόντοι που σκοράρει η γηπεδούχος ομάδα και φιλοξενούμενη ομάδα στις περιπτώσεις που νικάει ο γηπεδούχος.

Από τα δύο παραπάνω boxplot έχουμε ένα πολύ ενδιαφέρον συμπέρασμα, που έχει να κάνει με το πόσους πόντους πρέπει να επιτρέψεις στον αντίπαλο σου να σκοράρει. Από τα δύο διαγράμματα φαίνεται ότι στην περίπτωση που νικάει η γηπεδούχος ομάδα ο φιλοξενούμενος έχει σκοράρει στο 75% των περιπτώσεων λιγότερους από 80 πόντους. Αλλά και όταν χάνει η γηπεδούχος ομάδα πάλι στο 75% των περιπτώσεων έχει βάλει λιγότερους από 80 πόντους. Συνεπώς λοιπόν για να μπορέσει μία ομάδα να διεκδικήσει με αξιώσεις θα πρέπει να στοχεύσει πρωταρχικά στο να «κρατήσει» την αντίπαλό της χαμηλά στο σκορ, αντί να επιδιώξει υψηλότερο σκοράρισμα.

Στη συνέχεια αναπαριστούμε την επιτυχία/αποτυχία μιας επίθεσης αιφνιδιασμού ως προς τα δευτερόλεπτα που απαιτήθηκαν για την εκδήλωση της.



Εικόνα 10 : Σκορ μετά από αιφνιδιασμό.

Σε αυτή την περίπτωση μετρήσαμε τις προσπάθειες που στο σετ δεδομένων μας υπήρχε η αναφορά σε αιφνιδιασμό. Οι περιπτώσεις στις οποίες είχαμε αιφνιδιασμό στο σετ δεδομένων μας είναι 8553. Από αυτές σημειώθηκε καλάθι στις 4668. Συνεπώς Το αποτέλεσμα όπως φαίνεται μας δείχνει ότι αν σε περίπτωση αιφνιδιασμού σουτάρεις πριν τα 10 δευτερόλεπτα έχεις πιθανότητα να σκοράρεις κοντά στο 70% (αφού εκεί βρίσκεται το 50% των περιπτώσεων που μπήκε καλάθι και το 25% εκείνων που δεν μπήκε καλάθι). Τώρα αν σουτάρεις μετά τα 12 δευτερόλεπτα δηλαδή δεν εκδηλώσεις τον

αιφνιδιασμό και πας σε μια οργανωμένη επίθεση η πιθανότητα να σκοράρεις είναι 37%. Τέλος αν σουτάρεις ανάμεσα στα 10-12 δευτερόλεπτα η πιθανότητα να σκοράρεις είναι 55%. Είναι ένα αποτέλεσμα που εκ πρώτης όψεως φαίνεται λογικό, αλλά σίγουρα δεν θα μπορούσαμε να περιμένουμε τόσο μεγάλη διαφορά στην πιθανότητα σκοραρίσματος μεταξύ των τριών αυτών χρονικών διαστημάτων. Γιατί διαισθητικά, είναι λογικό όταν χάνεται ένας αιφνιδιασμός και περνάμε σε κανονική επίθεση να μειώνεται το ποσοστό θετικής κατάληξης της. Αλλά μια πτώση από 70% σε 37%, ίσως οδηγεί στο συμπέρασμα ότι ακόμα και τις φορές που οι ομάδες δεν έχουν και τόσο καλές επιλογές στον αιφνιδιασμό πρέπει να ρισκάρουν γιατί τους περιμένει ένα δύο φορές πιο δύσκολο έργο αν δεν το πράξουν.

6 Χαρακτηριστικά που δημιουργήσαμε για τα μοντέλα μας

Αρχικά ο σκοπός της εργασίας μας είναι να εξάγουμε γνώση από τα δεδομένα που διαθέταμε. Όπως δείξαμε και παραπάνω αναλύσαμε τα δεδομένα που μας δόθηκαν και βγάλαμε χρήσιμα συμπεράσματα για να μας βοηθήσουν στην κατασκευή του μοντέλου μας. Συνεπώς χρησιμοποιώντας τα δεδομένα που μας δόθηκαν αλλά και δημιουργώντας εμείς καινούργια χαρακτηριστικά δημιουργήσουμε μοντέλα που θα εκπαιδεύονται στους αγώνες τις κανονικής περιόδου και θα προσπαθούν να προβλέψουν τους νικητές στα Play-off και στο Final Four. Με την μορφή που είχε λοιπόν το σετ δεδομένων μας και αφορούσε συγκεκριμένες φάσεις από όλα τα παιχνίδια της Euroleague αλλά για τον κάθε παίκτη ξεχωριστά έπρεπε να τα προσαρμόσουμε πρώτα ως προς τις ομάδες. Οπότε ομαδοποιήσαμε τα δεδομένα κάθε παίκτη ανάλογα με την ομάδα στην οποία αγωνιζότανε και προσπαθήσαμε να βγάλουμε γενικά χαρακτηριστικά για την ομάδα και όχι για τους παίκτες ξεχωριστά.

Τα χαρακτηριστικά που εισάγαμε στο μοντέλο μας αφορούσαν τους μέσους όρους τις κάθε ομάδας σε έντεκα κατηγορίες

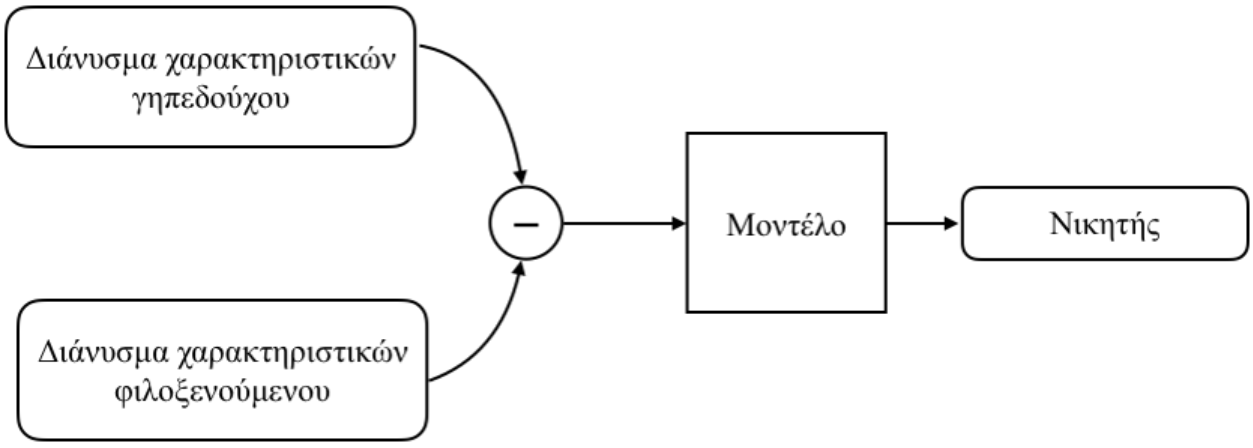
1. Προσπάθεις δύο πόντων
2. Εύστοχες προσπάθειες δύο πόντων
3. Προσπάθεις τριών πόντων
4. Εύστοχες προσπάθειες τριών πόντων
5. Βολές
6. Εύστοχες Βολές
7. Ασσίστ
8. Λάθη
9. Ριμπάουντ
10. Φάουλ
11. Pick and Roll

	pts2_att	pts2_made	pts3_att	pts3_made	ft_att	ft_made	assists	turnovers	fouls	rebounds	pnrs
team											
Barcelona	34.000000	18.766667	23.766667	9.166667	15.200000	11.266667	16.866667	13.966667	16.200000	8.200000	25.766667
Baskonia	44.666667	27.966667	25.333333	8.966667	19.866667	15.200000	17.300000	12.566667	15.000000	9.333333	28.033333
Brose	36.200000	23.066667	23.133333	9.400000	15.033333	11.733333	17.866667	12.066667	15.033333	6.133333	32.300000
CSKA	44.900000	29.500000	25.133333	10.200000	26.700000	21.933333	19.400000	13.600000	17.866667	7.833333	31.766667
Crvena Zvezda	41.966667	23.966667	20.266667	6.800000	17.200000	12.866667	15.933333	11.566667	14.800000	7.733333	30.566667
Darussafaka	43.966667	25.500000	27.300000	10.366667	22.133333	16.266667	13.900000	10.700000	15.800000	9.966667	31.933333
Efes	52.200000	30.133333	25.966667	8.933333	22.800000	16.733333	18.000000	11.233333	15.366667	10.766667	31.333333
Fenerbahce	41.566667	24.966667	25.700000	10.400000	20.133333	15.333333	16.466667	11.933333	16.366667	7.400000	28.500000
Galatasaray	37.966667	21.666667	22.600000	9.133333	13.500000	9.800000	18.866667	11.433333	14.233333	7.766667	32.300000
Maccabi TelAviv	39.900000	23.066667	20.700000	8.033333	16.400000	11.966667	16.866667	13.600000	14.366667	7.000000	31.266667
Milano	43.666667	26.900000	20.300000	7.100000	19.733333	14.733333	16.666667	13.766667	16.100000	9.233333	27.733333
Olympiacos	46.566667	29.100000	29.200000	10.366667	23.366667	17.366667	14.566667	11.966667	17.366667	9.766667	29.300000
Panathinaikos	38.100000	23.500000	28.566667	10.200000	18.433333	13.233333	14.933333	10.766667	15.566667	8.266667	28.966667
Real Madrid	44.400000	29.800000	31.033333	11.700000	21.800000	16.766667	20.366667	11.466667	16.033333	9.866667	25.633333
Unics	40.600000	21.766667	19.500000	7.133333	18.033333	14.366667	15.300000	12.200000	16.600000	9.333333	33.566667
Zalgiris	37.766667	22.666667	21.400000	8.300000	17.266667	13.766667	18.066667	13.066667	16.666667	9.233333	30.866667

Εικόνα 11 : Χαρακτηριστικά που χρησιμοποιήσαμε στο μοντέλο μας.

7 Πώς λειτουργεί το μοντέλο μας

Το μοντέλο μας παίρνει σαν είσοδο το διάνυσμα των χαρακτηριστικών της γηπεδούχου ομάδας και αφαιρεί από αυτό τα χαρακτηριστικά της φιλοξενούμενης ομάδας χωρίς να μας ενδιαφέρει το ποια είναι η κάθε ομάδα. Όπως είπαμε και παραπάνω η έξοδος του είναι το αποτέλεσμα του αγώνα και αξιολογούμε τα μοντέλα μας βάση του accuracy score.



Εικόνα 12 : Ο τρόπος λειτουργίας του μοντέλου μας.

Χρησιμοποιήσαμε τη βιβλιοθήκη sklearn κανονικοποιήσαμε τα χαρακτηριστικά μας έτσι ώστε να ακολουθούν Κανονική Κατανομή με μέση τιμή **0** και διασπορά **1**. Παρακάτω σας παραθέτουμε τα

μοντέλα της sklearn που χρησιμοποιήσαμε και τα αποτελέσματα που μας έδωσαν. Όπως αναφέραμε και νωρίτερα κατασκευάσαμε δυο διαφορετικά μοντέλα για την κάθε season έχοντας κάθε φορά ως training set τους αγώνες της κανονικής περιόδου.

Model	Accuracy
Logistic Regression	70.00%
3 Nearest Neighbour	60.00%
5 Nearest Neighbour	66.67%
Naïve Bayes	66.67%
SVM (rbf kernel)	60.00%
SVM (linear kernel)	70.00%
Decision Trees	60.00%

Εικόνα 13 : Τα αποτελέσματα των μοντέλων μας για την season 2016-2017

Όπως βλέπουμε από τα αποτελέσματα των μοντέλων μας η πρόβλεψη ενός αποτελέσματος σε ένα ομαδικό άθλημα είναι μία πολύ δύσκολη διαδικασία. Τα μοντέλα μας δίνουν ένα προβάδισμα στις ομάδες που παίζουν εντός έδρας. Γεγονός που είναι απολύτως λογικό γιατί όπως είδαμε και σε προηγούμενο γράφημα, οι ομάδες με δυνατή έδρα έχουν περάσει στην επόμενη φάση. Τα καλύτερα ποσοστά τα πετυχαίνουμε με τον SVM (linear kernel) και με το Logistic Reggresion.

8 Επεκτάσεις

- Θα μπορούσαμε να χρησιμοποιήσουμε κάποιο σετ δεδομένων που αφορά τις προπονήσεις της κάθε ομάδας με στόχο την συλλογή πληροφορίας που ίσως δεν εμφανίζεται στο σετ δεδομένων μας.
- Θα ήταν χρήσιμο να εμπλουτίσουμε τα δεδομένα μας χρησιμοποιώντας μεταβλητές όπως τα αποτελέσματα της κάθε ομάδας στους πιο πρόσφατους αγώνες ή τους τραυματισμούς που μπορεί να έχει σε μια χρονική περίοδο, τροφοδοτώντας έτσι το μοντέλο με μια εικόνα της αγωνιστικής της φόρμας.
- Ο βασικός λόγος που δεν χρησιμοποιήσαμε νευρωνικά δίκτυα είναι το περιορισμένο σύνολο δεδομένων που είχαμε στη διάθεση μας. Αν συμπεριλαμβάναμε εγγραφές και από άλλες

αναμετρήσεις (ακόμα και διαφορετικών πρωταθλημάτων, καθώς αυτό που μας ενδιαφέρει είναι να μπορέσουμε να αποτυπώσουμε στο μοντέλο μας τους κανόνες που διέπουν το παιχνίδι, ανεξάρτητα από το ποιες ομάδες αναμετρώνται κάθε φορά) θα μπορούσαμε να συνθέσουμε πολύ πιο σύνθετα μοντέλα.