Επίθετο: Μορφιαδάκης Όνομα: Εμμανουήλ AM:3150112

Εξόρυξη γνώσης από δεδομένα ομάδων μπάσκετ χρονολογιών 2015-2019

Περιγραφή πτυχιακής εργασίας

Το θέμα της πτυχιακής εργασίας ήταν να πάρω ένα αρχείο από το Kaggle το οποίο περιέχει δεδομένα για ομάδες μπάσκετ τα χρονικά διαστήματα 2015 με 2019.Σε αυτά τα δεδομένα έπρεπε να προβλέψω την πιθανότητα που έχει η κάθε ομάδα να κερδίσει μια μέση ομάδα καθώς και το ποσοστό των δίποντων και τρίποντων.

Για να τα προβλέψω αυτά αποφάσισα να εκπαιδεύσω τους αλγορίθμους Decision Tree,Random Forest και Linear Regression για τα δεδομένα 2015 και 2018 και ύστερα να τεστάρω πόσο καλά προβλέπει τα παραπάνω για την χρονολογία 2019.

Περιγραφή αλγορίθμων που χρησιμοποίησα

Linear Regression:Ο Linear Regression προσπαθεί να φτιάξει μια γραμμική συνάρτηση (y=Σwixi+b)που να προβλέπει το output όσο πιο καλά γίνεται για τα δεδομένα. Για να το πετύχει αυτό εκπαιδεύεται με δεδομένα που γνωρίζει το output και για να αποφύγει το overfitting δεν αφήνει τα wi να πάρουν πολύ μεγάλες τιμές.

Decision Tree Regressor:Ο Decision Tree Regressor προσπαθεί να βρει μια πατέντα που ακολουθούν τα δεδομένα φτιάχνοντας ένα δέντρο. Το δέντρο αυτό έχει για κόμβους ανισότητες των attributes(πχ G>40) και για φύλα τιμές.Αν ισχύει η συνθήκη του κόμβου το δεδομένο εισάγεται στο κάτω δεξιό υποδέντρο αλλιώς στο αριστερό. Ανάλογα το φύλο που θα εισαχθεί το output παίρνει μια τιμή.

Random Forest Regressor:Ο Random Forest φτιάχνει ένα σύνολο από Decision Trees.Αφού γίνει αυτό εισάγει κάθε δεδομένο που θα πάει για εκπαίδευση στα Decision Trees και η τιμή που του αποδίδεται είναι ο μέσος όρος των τιμών στα φύλλα που κατατάχθηκε το δεδομένο.

Οι ενέργειες που έκανα και τα συμπεράσματά μου

Ξεκινώντας πήρα το dataset από αυτόν τον σύνδεσμο. Κατέβασα το αρχείο cbb.csv το οποίο περιέχει τα παρακάτω columns:

Ορολογία	Περιγραφή ορολογίας.
TEAM	Ομάδα.
CONF	Η αθλητική σύσκεψη που συμμετέχει η ομάδα.
G	Ο αριθμός των παιχνιδιών που έπαιξε η ομάδα μια χρονιά.
W	Ο αριθμός των παιχνιδιών που νίκησε η ομάδα.
ADJOE	Μια εκτίμηση των πόντων που θα σκοράρει η ομάδα(ανά 100 κατοχές) εναντίων μιας μέτριας ομάδας.
ADJDE	Μια εκτίμηση των πόντων που θα σκοράρει η ομάδα(ανά 100 κατοχές) εναντίων μιας μέτριας ομάδας.

BARTHAG	Η πιθανότητα η ομάδα να νικήσει μια μέση ομάδα.
EFG_O	Ποσοστό επιτυχημένων καλαθιών που σκόραρε η ομάδα.
EFG_D	Ποσοστό επιτυχημένων καλαθιών που δέχθηκε η ομάδα.
TOR	Το ποσοστό των turnover που δέχθηκε η ομάδα.
TORD	Το ποσοστό των turnover που έκανε η ομάδα.
ORB	Το ποσοστό των επιθετικών rebound που έκανε η ομάδα.
DRB	Το ποσοστό των αμυντικών rebound που έκανε η ομάδα.
FTR	Η συχνότητα των ελεύθερων βολών που εκτέλεσε η ομάδα.
FTRD	Η συχνότητα των ελεύθερων βολών που εκτέλεσαν η αντίπαλες ομάδες.
2P_O	Το ποσοστό των δίποντων που εκτέλεσε η ομάδα.
2P_D	Το ποσοστό των δίποντων που δέχθηκε η ομάδα.
3P_O	Το ποσοστό των τριπόντων που εκτέλεσε η ομάδα.
3P_D	Το ποσοστό των τριπόντων που δέχθηκε η ομάδα.
ADJ_T	Η εκτίμηση των κατοχών ανά 40 λεπτά που θα έχει η ομάδα εναντίων μιας μέσης ομάδας.
WAB	wins above bubble.
POSTSEASON	Το που είναι η ομάδα π.χ (Champions,First Four κλπ).
SEED	
YEAR	Χρόνος.

Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε για το project ήταν η python,τα εργαλεία που χρησιμοποιήθηκαν ήταν το Anaconda Navigator και το Jupiter Notebook.Το αρχείο διαβάζεται από την βιβλιοθήκη pandas μέσα από την μέθοδο read_csv.Οι αλγόριθμοι εκπαίδευσης έγιναν import από την βιβλιοθήκη sklearn.Επίσης από αυτήν την βιβλιοθήκη υπολογίστηκε και το score για τον κάθε αλγόριθμο. Χρησιμοποιώ ως train τα δεδομένα με χρονολογία 2015-2018 και για test τα δεδομένα με χρονολογία 2019.

Train_size:1404 Test_size:353.

Αρχικά κράτησα όλα τα attributes για να εκπαιδευτούν οι αλγόριθμοι εκτός των POSTSEASON και SEED και TEAMS επειδή τα attributes POSTSEASON και SEED για κάποια στοιχεία δεν είχαν τιμή και το attribute TEAMS δεν το κράτησα επειδή αυτό δεν έχει σημασία. Έπειτα χρησιμοποίησα τον Label Encoder για να μετατραπούν σε αριθμητικές τιμές τα ορίσματα στο CONF με σκοπό να εκπαιδευτούν οι αλγόριθμοι. Από κάτω είναι τα αποτελέσματα του πειράματος:

Score: 0.9775911605290136
Prediction of BARTHAG using Decision Tree
TABLE A1

TEAM	G	W	BARTHAG	Predicted_BARTHAG	ERROR
Virginia	38	35	0.9736	0.9702	0.0034
Houston	37	33	0.9439	0.9507	0.0068
Gonzaga	37	33	0.9744	0.9576	0.0168
Duke	38	32	0.9646	0.9517	0.0129
Buffalo	35	32	0.8819	0.9014	0.0195
Michigan St.	39	32	0.9597	0.9308	0.0289
Texas Tech	38	31	0.9696	0.9507	0.0189
UC Irvine	36	31	0.7458	0.7458	0.0000
Tennessee	36	31	0.9488	0.9648	0.0160
Wofford	32	30	0.8892	0.9049	0.0157

Score: 0.9884696857070043
Prediction of BARTHAG using Random Forest
TABLE A2

TEAM	G	W	BARTHAG	Predicted_BARTHAG	ERROR
Virginia	38	35	0.9736	0.9643	0.0093
Houston	37	33	0.9439	0.9517	0.0078
Gonzaga	37	33	0.9744	0.9515	0.0229
Duke	38	32	0.9646	0.9597	0.0049
Buffalo	35	32	0.8819	0.9338	0.0519
Michigan St.	39	32	0.9597	0.9561	0.0036
Texas Tech	38	31	0.9696	0.9452	0.0244
UC Irvine	36	31	0.7458	0.7715	0.0257
Tennessee	36	31	0.9488	0.9509	0.0021
Wofford	32	30	0.8892	0.9002	0.0110

Score: 0.9817055957384296
Prediction of BARTHAG using Linear Regression
TABLE A3

TEAM	G	W	BARTHAG	Predicted_BARTHAG	ERROR
Virginia	38	35	0.9736	1.0000	0.0264
Houston	37	33	0.9439	1.0000	0.0561
Gonzaga	37	33	0.9744	1.0000	0.0256
Duke	38	32	0.9646	1.0000	0.0354
Buffalo	35	32	0.8819	0.8794	0.0025
Michigan St.	39	32	0.9597	1.0000	0.0403
Texas Tech	38	31	0.9696	1.0000	0.0304
UC Irvine	36	31	0.7458	0.6953	0.0505
Tennessee	36	31	0.9488	1.0000	0.0512
Wofford	32	30	0.8892	0.8848	0.0044

Score: 0.9219407647954928
Prediction of 2P_O using Decision Tree
TABLE A4

TEAM	G	W	2P_O	Predicted_2P_O	ERROR
Virginia	38	35	52.5	51.8	0.7
Houston	37	33	51.3	51.8	0.5
Gonzaga	37	33	61.4	58.2	3.2
Duke	38	32	58.0	54.0	4.0
Buffalo	35	32	55.7	53.0	2.7
Michigan St.	39	32	54.3	54.3	0.0
Texas Tech	38	31	52.8	53.3	0.5
UC Irvine	36	31	50.1	50.2	0.1
Tennessee	36	31	55.4	55.5	0.1
Wofford	32	30	53.9	52.5	1.4

Score: 0.9609100894766307
Prediction of 2P_O using Random Forest
TABLE A5

TEAM	G	W	2P_O	Predicted_2P_O	ERROR
Virginia	38	35	52.5	52.8	0.3
Houston	37	33	51.3	51.5	0.2
Gonzaga	37	33	61.4	59.3	2.1
Duke	38	32	58.0	54.5	3.5
Buffalo	35	32	55.7	54.0	1.7
Michigan St.	39	32	54.3	54.5	0.2
Texas Tech	38	31	52.8	53.0	0.2
UC Irvine	36	31	50.1	49.8	0.3
Tennessee	36	31	55.4	55.3	0.1
Wofford	32	30	53.9	54.9	1.0

Score: 0.9832159580844068
Prediction of 2P_O using Linear Regression
TABLE A6

TEAM	G	W	2P_O	Predicted_2P_O	ERROR
Virginia	38	35	52.5	52.7	0.2
Houston	37	33	51.3	51.3	0.0
Gonzaga	37	33	61.4	61.4	0.0
Duke	38	32	58.0	57.7	0.3
Buffalo	35	32	55.7	54.9	0.8
Michigan St.	39	32	54.3	54.3	0.0
Texas Tech	38	31	52.8	52.6	0.2
UC Irvine	36	31	50.1	49.8	0.3
Tennessee	36	31	55.4	55.2	0.2
Wofford	32	30	53.9	54.8	0.9

Score: 0.8715275509036727
Prediction of 3P_O using Decision Tree
TABLE A7

TEAM	G	W	3P_O	Predicted_3P_O	ERROR
Virginia	38	35	39.5	40.6	1.1
Houston	37	33	35.5	35.8	0.3
Gonzaga	37	33	36.3	36.9	0.6
Duke	38	32	30.8	34.0	3.2
Buffalo	35	32	33.7	34.0	0.3
Michigan St.	39	32	37.8	37.3	0.5
Texas Tech	38	31	36.5	36.6	0.1
UC Irvine	36	31	35.9	36.5	0.6
Tennessee	36	31	36.7	36.9	0.2
Wofford	32	30	41.4	40.1	1.3

Score: 0.9285234356001574
Prediction of 3P_O using Random Forest
TABBLE A8

TEAM	G	W	3P_O	Predicted_3P_O	ERROR
Virginia	38	35	39.5	40.5	1.0
Houston	37	33	35.5	36.0	0.5
Gonzaga	37	33	36.3	38.6	2.3
Duke	38	32	30.8	34.3	3.5
Buffalo	35	32	33.7	34.2	0.5
Michigan St.	39	32	37.8	38.2	0.4
Texas Tech	38	31	36.5	36.6	0.1
UC Irvine	36	31	35.9	36.1	0.2
Tennessee	36	31	36.7	37.8	1.1
Wofford	32	30	41.4	40.9	0.5

Score: 0.9651819316733338
Prediction of 3P_O using Linear Regression
TABLE A9

TEAM	G	W	3P_O	Predicted_3P_O	ERROR
Virginia	38	35	39.5	39.6	0.1
Houston	37	33	35.5	35.6	0.1
Gonzaga	37	33	36.3	36.6	0.3
Duke	38	32	30.8	30.9	0.1
Buffalo	35	32	33.7	33.0	0.7
Michigan St.	39	32	37.8	37.8	0.0
Texas Tech	38	31	36.5	36.3	0.2
UC Irvine	36	31	35.9	35.4	0.5
Tennessee	36	31	36.7	36.6	0.1
Wofford	32	30	41.4	42.3	0.9

Από ότι φαίνεται στα TABLES(A1 με A3) το μεγαλύτερο score για την πρόβλεψη του BARTHAG ήταν του Random Forest.Αλλά για την πρόβλεψη του 2P_O και του 3P_O από ότι φαίνεται στα TABLES (A4 με A9)το μεγαλύτερο score ήταν στον Linear Regression.Αυτό γιατί υπάρχει γραμμική συσχέτιση μεταξύ των attributes για την πρόβλεψη του 2P_O και του 3P_O.

Το score στον Linear Regression(TABLE A3) για την πρόβλεψη του BARTHAG είναι μεγαλύτερο από το Decision Tree(TABLE A1) είτε επειδή έτυχε με το συγκεκριμένο δείγμα για Test είτε επειδή υπάρχει γραμμική σχέση μεταξύ κάποιον χαρακτηριστικών που περνάν ως ορίσματα στον αλγόριθμο εκπαίδευσης.Από ότι φαίνεται στα TABLES(A4 με A9) το score στον Linear Regression για την πρόβλεψη των 2P_O και 3P_O είναι το μεγαλύτερο επειδή υπάρχει γραμμική σχέση μεταξύ των χαρακτηριστικών που περνάν ως ορίσματα στον αλγόριθμο εκπαίδευσης καθώς για την πρόβλεψη των 2P_O και 3P_O σε όλα τα πειράματα είχε το υψηλότερο score. Επίσης πιστεύω πως για τον υπολογισμό του BARTHAG υπάρχει συνάρτηση που παίρνει ως ορίσματα υποσύνολο των attributes.

Έπειτα χρησιμοποιώ για εκπαίδευση τα attributes(G,W).Κρατώντας αυτά τα attributes παρατηρώ αυτά τα αποτελέσματα:

```
Score: 0.6313960880300524
Prediction of BARTHAG using Decision Tree
TABLE B1
              TEAM   G   W  BARTHAG  Predicted_BARTHAG   ERROR
6          Virginia  38  35   0.9736             0.9345  0.0391
1726        Houston  37  33   0.9439             0.9345  0.0094
29          Gonzaga  37  33   0.9744             0.9345  0.0399
16             Duke  38  32   0.9646             0.9345  0.0301
1513        Buffalo  35  32   0.8819             0.9345  0.0526
32    Michigan St.  39  32   0.9597             0.9345  0.0252
3       Texas Tech  38  31   0.9696             0.9345  0.0351
1506     UC Irvine  36  31   0.7458             0.9345  0.1887
1754     Tennessee  36  31   0.9488             0.9345  0.0143
1526       Wofford  32  30   0.8892             0.7081  0.1811


Score: 0.6453758271522153
Prediction of BARTHAG using Random Forest
TABLE B2
              TEAM   G   W  BARTHAG  Predicted_BARTHAG   ERROR
6          Virginia  38  35   0.9736             0.9350  0.0386
1726        Houston  37  33   0.9439             0.9350  0.0089
29          Gonzaga  37  33   0.9744             0.9350  0.0394
16             Duke  38  32   0.9646             0.9350  0.0296
1513        Buffalo  35  32   0.8819             0.9350  0.0531
32    Michigan St.  39  32   0.9597             0.9350  0.0247
3       Texas Tech  38  31   0.9696             0.9350  0.0346
1506     UC Irvine  36  31   0.7458             0.9350  0.1892
1754     Tennessee  36  31   0.9488             0.9350  0.0138
1526       Wofford  32  30   0.8892             0.7044  0.1848
```

Score: 0.6656978886505613
Prediction of BARTHAG using Linear Regression
TABLE B3

	TEAM	G	W	BARTHAG	Predicted_BARTHAG	ERROR
6	Virginia	38	35	0.9736	1.0000	0.0264
1726	Houston	37	33	0.9439	1.0000	0.0561
29	Gonzaga	37	33	0.9744	1.0000	0.0256
16	Duke	38	32	0.9646	1.0000	0.0354
1513	Buffalo	35	32	0.8819	0.9749	0.0930
32	Michigan St.	39	32	0.9597	1.0000	0.0403
3	Texas Tech	38	31	0.9696	1.0000	0.0304
1506	UC Irvine	36	31	0.7458	0.9676	0.2218
1754	Tennessee	36	31	0.9488	0.9676	0.0188
1526	Wofford	32	30	0.8892	0.8645	0.0247

Score: 0.26184130339517464
Prediction of 2P_O using Decision Tree
TABLE B4

	TEAM	G	W	2P_O	Predicted_2P_O	ERROR
6	Virginia	38	35	52.5	54.1	1.6
1726	Houston	37	33	51.3	54.1	2.8
29	Gonzaga	37	33	61.4	54.1	7.3
16	Duke	38	32	58.0	54.1	3.9
1513	Buffalo	35	32	55.7	54.1	1.6
32	Michigan St.	39	32	54.3	54.1	0.2
3	Texas Tech	38	31	52.8	54.1	1.3
1506	UC Irvine	36	31	50.1	54.1	4.0
1754	Tennessee	36	31	55.4	54.1	1.3
1526	Wofford	32	30	53.9	54.1	0.2

Score: 0.2675289356204259
Prediction of 2P_O using Random Forest
TABLE B5

	TEAM	G	W	2P_O	Predicted_2P_O	ERROR
6	Virginia	38	35	52.5	54.4	1.9
1726	Houston	37	33	51.3	54.4	3.1
29	Gonzaga	37	33	61.4	54.4	7.0
16	Duke	38	32	58.0	54.4	3.6
1513	Buffalo	35	32	55.7	54.4	1.3
32	Michigan St.	39	32	54.3	54.4	0.1
3	Texas Tech	38	31	52.8	53.4	0.6
1506	UC Irvine	36	31	50.1	53.4	3.3
1754	Tennessee	36	31	55.4	53.4	2.0
1526	Wofford	32	30	53.9	53.6	0.3

Score: 0.24938076636458906
Prediction of 2P_O using Linear Regression
TABLE B6

	TEAM	G	W	2P_O	Predicted_2P_O	ERROR
6	Virginia	38	35	52.5	54.3	1.8
1726	Houston	37	33	51.3	53.8	2.5
29	Gonzaga	37	33	61.4	53.8	7.6
16	Duke	38	32	58.0	53.2	4.8
1513	Buffalo	35	32	55.7	53.9	1.8
32	Michigan St.	39	32	54.3	53.0	1.3
3	Texas Tech	38	31	52.8	52.8	0.0
1506	UC Irvine	36	31	50.1	53.3	3.2
1754	Tennessee	36	31	55.4	53.3	2.1
1526	Wofford	32	30	53.9	53.8	0.1

Score: -0.005579708921825688
Prediction of 3P_O using Decision Tree
TABLE B7

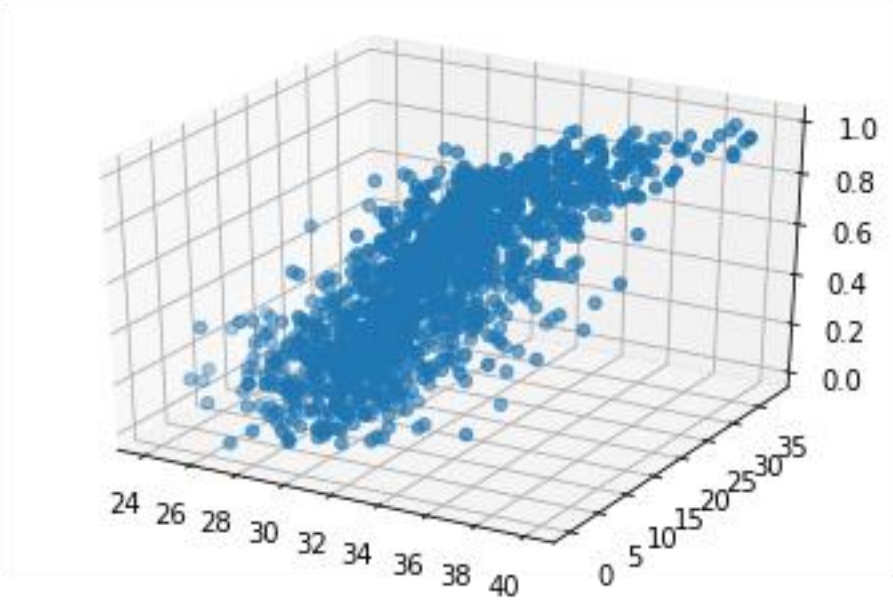
	TEAM	G	W	3P_O	Predicted_3P_O	ERROR
6	Virginia	38	35	39.5	38.1	1.4
1726	Houston	37	33	35.5	38.1	2.6
29	Gonzaga	37	33	36.3	38.1	1.8
16	Duke	38	32	30.8	38.1	7.3
1513	Buffalo	35	32	33.7	38.1	4.4
32	Michigan St.	39	32	37.8	38.1	0.3
3	Texas Tech	38	31	36.5	38.1	1.6
1506	UC Irvine	36	31	35.9	38.1	2.2
1754	Tennessee	36	31	36.7	38.1	1.4
1526	Wofford	32	30	41.4	38.1	3.3

Score: 0.028161466430863258
Prediction of 3P_O using Random Forest
TABLE B8

	TEAM	G	W	3P_O	Predicted_3P_O	ERROR
6	Virginia	38	35	39.5	37.8	1.7
1726	Houston	37	33	35.5	37.8	2.3
29	Gonzaga	37	33	36.3	37.8	1.5
16	Duke	38	32	30.8	37.8	7.0
1513	Buffalo	35	32	33.7	37.8	4.1
32	Michigan St.	39	32	37.8	37.8	0.0
3	Texas Tech	38	31	36.5	37.8	1.3
1506	UC Irvine	36	31	35.9	37.8	1.9
1754	Tennessee	36	31	36.7	37.8	1.1
1526	Wofford	32	30	41.4	37.8	3.6

```
Score: 0.08519265909370799
Prediction of 3P_O using Linear Regression
TABLE B9
              TEAM   G   W  3P_O  Predicted_3P_O  ERROR
6         Virginia  38  35  39.5            38.0    1.5
1726       Houston  37  33  35.5            37.7    2.2
29         Gonzaga  37  33  36.3            37.7    1.4
16            Duke  38  32  30.8            37.3    6.5
1513       Buffalo  35  32  33.7            37.8    4.1
32    Michigan St.  39  32  37.8            37.1    0.7
3       Texas Tech  38  31  36.5            37.0    0.5
1506     UC Irvine  36  31  35.9            37.4    1.5
1754     Tennessee  36  31  36.7            37.4    0.7
1526       Wofford  32  30  41.4            37.9    3.5
```

Από ότι φαίνεται από τα TABLES(B1 με B3)υπάρχει μια γραμμική συσχέτιση μεταξύ των attributes(G,W,BARTHAG) επειδή το υψηλότερο score το έχει ο Linear Regressor έχοντας ως output το BARTHAG.Για να κοιτάξω αυτήν την συσχέτιση απεικόνισα σε τρισδιάστατο πίνακα τα δεδομένα με τα παραπάνω attributes(G,W,BARTHAG).



Όπως φαίνεται η γραμμική συσχέτιση είναι σε ένα βαθμό ισχυρή.

Επόμενο βήμα ήταν να δω την συμπεριφορά του αλγόριθμου αλλάζοντας τα δεδομένα εκπαίδευσης. Οι αλλαγές φαίνονται στον παρακάτω πίνακα:

YEAR	BARTHAG	2P_O	3P_O
2015	0.5	19.8	22.8
2016	0.6	20.8	50
2017	0.55	30.6	43.1

2018	0.65	60.9	58.2

Το score με αυτές τις αλλαγές ήταν πολύ χαμηλό. Από αυτό παρατηρείται πως υπάρχει κάποια λογική που υπολογίζονται το BARTHAG το 2P_O και το 3P_O.

Αφού στο Decision Tree και στον Random Forest άλλαξα τις παραμέτρους min_samples_split,max_leaf_nodes και max_depth με σκοπό να σταματήσει το overfitting, το score αυξήθηκε λίγο αλλά παρόλες τις αλλαγές μεγαλύτερο score είχε ο Linear Regression.

Έπειτα για εκπαίδευση κράτησα τα attributes(ADJOE,ADJDE,TOR,TORD,ORB,DRB,ADJ_T) επειδή υπέθεσα πως κάποια attributes προδίδουν το BARTHAG(όπως τα WAB,G,W),το 2P_O και το 3P_O.Όμως το score για το BARTHAG ήταν υψηλό και πάλι. Παρακάτω είναι τα αποτελέσματα:

Score: 0.9755045534229869
Prediction of BARTHAG using Decision Tree
TABLE C1

	TEAM	G	W	BARTHAG	Predicted_BARTHAG	ERROR
6	Virginia	38	35	0.9736	0.9558	0.0178
1726	Houston	37	33	0.9439	0.9254	0.0185
29	Gonzaga	37	33	0.9744	0.9558	0.0186
16	Duke	38	32	0.9646	0.9254	0.0392
1513	Buffalo	35	32	0.8819	0.9254	0.0435
32	Michigan St.	39	32	0.9597	0.9558	0.0039
3	Texas Tech	38	31	0.9696	0.9254	0.0442
1506	UC Irvine	36	31	0.7458	0.6915	0.0543
1754	Tennessee	36	31	0.9488	0.9558	0.0070
1526	Wofford	32	30	0.8892	0.8642	0.0250

Score: 0.9850320899895498
Prediction of BARTHAG using Random Forest
TABLE C2

	TEAM	G	W	BARTHAG	Predicted_BARTHAG	ERROR
6	Virginia	38	35	0.9736	0.9334	0.0402
1726	Houston	37	33	0.9439	0.9273	0.0166
29	Gonzaga	37	33	0.9744	0.9334	0.0410
16	Duke	38	32	0.9646	0.9334	0.0312
1513	Buffalo	35	32	0.8819	0.8889	0.0070
32	Michigan St.	39	32	0.9597	0.9334	0.0263
3	Texas Tech	38	31	0.9696	0.9273	0.0423
1506	UC Irvine	36	31	0.7458	0.7314	0.0144
1754	Tennessee	36	31	0.9488	0.9314	0.0174
1526	Wofford	32	30	0.8892	0.9052	0.0160

Score: 0.9803530613389121
Prediction of BARTHAG using Linear Regression
TABLE C3

	TEAM	G	W	BARTHAG	Predicted_BARTHAG	ERROR
6	Virginia	38	35	0.9736	1.0000	0.0264
1726	Houston	37	33	0.9439	1.0000	0.0561
29	Gonzaga	37	33	0.9744	1.0000	0.0256
16	Duke	38	32	0.9646	1.0000	0.0354
1513	Buffalo	35	32	0.8819	0.8749	0.0070
32	Michigan St.	39	32	0.9597	1.0000	0.0403
3	Texas Tech	38	31	0.9696	1.0000	0.0304
1506	UC Irvine	36	31	0.7458	0.6970	0.0488
1754	Tennessee	36	31	0.9488	1.0000	0.0512
1526	Wofford	32	30	0.8892	0.8886	0.0006

Score: 0.3496679918204313
Prediction of 2P_O using Decision Tree
TABLE C4

	TEAM	G	W	2P_O	Predicted_2P_O	ERROR
6	Virginia	38	35	52.5	55.0	2.5
1726	Houston	37	33	51.3	53.1	1.8
29	Gonzaga	37	33	61.4	53.1	8.3
16	Duke	38	32	58.0	53.1	4.9
1513	Buffalo	35	32	55.7	50.5	5.2
32	Michigan St.	39	32	54.3	53.1	1.2
3	Texas Tech	38	31	52.8	50.5	2.3
1506	UC Irvine	36	31	50.1	49.0	1.1
1754	Tennessee	36	31	55.4	53.1	2.3
1526	Wofford	32	30	53.9	53.1	0.8

Score: 0.43324654151207476
Prediction of 2P_O using Random Forest
TABLE C5

	TEAM	G	W	2P_O	Predicted_2P_O	ERROR
6	Virginia	38	35	52.5	52.7	0.2
1726	Houston	37	33	51.3	52.3	1.0
29	Gonzaga	37	33	61.4	53.1	8.3
16	Duke	38	32	58.0	52.8	5.2
1513	Buffalo	35	32	55.7	51.7	4.0
32	Michigan St.	39	32	54.3	52.8	1.5
3	Texas Tech	38	31	52.8	53.4	0.6
1506	UC Irvine	36	31	50.1	48.6	1.5
1754	Tennessee	36	31	55.4	52.7	2.7
1526	Wofford	32	30	53.9	52.5	1.4

Score: 0.5242612181413259
Prediction of 2P_O using Linear Regression
TABLE C6

	TEAM	G	W	2P_O	Predicted_2P_O	ERROR
6	Virginia	38	35	52.5	54.3	1.8
1726	Houston	37	33	51.3	51.3	0.0
29	Gonzaga	37	33	61.4	55.0	6.4
16	Duke	38	32	58.0	52.8	5.2
1513	Buffalo	35	32	55.7	52.0	3.7
32	Michigan St.	39	32	54.3	53.8	0.5
3	Texas Tech	38	31	52.8	53.7	0.9
1506	UC Irvine	36	31	50.1	48.2	1.9
1754	Tennessee	36	31	55.4	54.6	0.8
1526	Wofford	32	30	53.9	53.1	0.8

Score: 0.1394654168846846
Prediction of 3P_O using Decision Tree
TABLE C7

	TEAM	G	W	3P_O	Predicted_3P_O	ERROR
6	Virginia	38	35	39.5	39.3	0.2
1726	Houston	37	33	35.5	35.8	0.3
29	Gonzaga	37	33	36.3	36.7	0.4
16	Duke	38	32	30.8	36.7	5.9
1513	Buffalo	35	32	33.7	37.0	3.3
32	Michigan St.	39	32	37.8	39.3	1.5
3	Texas Tech	38	31	36.5	37.0	0.5
1506	UC Irvine	36	31	35.9	35.7	0.2
1754	Tennessee	36	31	36.7	39.3	2.6
1526	Wofford	32	30	41.4	37.0	4.4

Score: 0.2954030503132269
Prediction of 3P_O using Random Forest
TABLE C8

	TEAM	G	W	3P_O	Predicted_3P_O	ERROR
6	Virginia	38	35	39.5	38.9	0.6
1726	Houston	37	33	35.5	36.8	1.3
29	Gonzaga	37	33	36.3	38.9	2.6
16	Duke	38	32	30.8	38.7	7.9
1513	Buffalo	35	32	33.7	36.4	2.7
32	Michigan St.	39	32	37.8	38.9	1.1
3	Texas Tech	38	31	36.5	36.8	0.3
1506	UC Irvine	36	31	35.9	35.3	0.6
1754	Tennessee	36	31	36.7	38.7	2.0
1526	Wofford	32	30	41.4	37.5	3.9

```
Score: 0.36525986175630876
Prediction of 3P_O using Linear Regression
TABLE C9
              TEAM   G   W  3P_O  Predicted_3P_O  ERROR
6         Virginia  38  35  39.5            39.7    0.2
1726       Houston  37  33  35.5            36.0    0.5
29         Gonzaga  37  33  36.3            38.4    2.1
16            Duke  38  32  30.8            36.0    5.2
1513       Buffalo  35  32  33.7            35.7    2.0
32    Michigan St.  39  32  37.8            37.9    0.1
3       Texas Tech  38  31  36.5            37.1    0.6
1506     UC Irvine  36  31  35.9            34.1    1.8
1754     Tennessee  36  31  36.7            39.0    2.3
1526       Wofford  32  30  41.4            37.9    3.5
```

Από ότι κατάλαβα για να προβλέψουμε το 2P_O και το 3P_O σχετικά καλά πρέπει να κρατήσουμε σχεδόν όλα τα attributes και να χρησιμοποιήσουμε τον Linear Regressor καθώς υπάρχουν γραμμικές συσχετίσεις. Αυτό φαίνεται αν συγκρίνουμε το score στα TABLES(A4 με A9)που είναι υψηλό με τα TABLES(B4 με B9 και C4 με C9) που είναι χαμηλό.

Έπειτα πρόσθεσα στα attributes το 2P_D και το 3P_D.Τα αποτελέσματα ήταν περίπου ίδια με πριν.

Τελικά συμπεράσματα

Συνοψίζοντας παρατηρώ πως μπορούμε να προβλέψουμε το BARTHAG πολύ καλά ακόμα και με σχετικά λίγα δεδομένα τα οποία δεν το προδίδουν. Αυτό φαίνεται στα TABLES(A1 με A3 και C1 με C3). Επίσης πιστεύω πως ο Linear Regressor επιτυγχάνει συνήθως μεγαλύτερο score από τον Decision Tree για output το BARTHAG επειδή τα δεδομένα έχουν γραμμική συσχέτιση σε σχέση με το BARTHAG. Αυτό παρατηρείται στα TABLES(A1,A2,C1,C2).Παρ όλα αυτά καλή δουλειά κάνουν και ο Decision Tree και ο Random Forest. Αυτό φαίνεται στα TABLES(A2,A3,C2,C3).Επίσης τα 2P_O και τα 3P_O μπορούμε να τα προβλέψουμε καλύτερα με το Linear Regression κρατώντας τα περισσότερα attributes επειδή υπάρχει ισχυρή γραμμική συσχέτιση. Αυτό φαίνεται στα TABLES(A6,A9)