<u>מטלת גמר קורס למידת מכונה</u>

מגישים:
יהונתן אסקוכידו - 206326548
שמואל הרוש - 210037495

תיאור של המאגרים :

א. Ligue 1 - ליגה צרפתית -

כל התוצאות הסופיות של העונות 1999 עד 2019:

	year	Home Team	Away Team	Home Team Goals	Away Team Goals	Winner
0	1999	Bordeaux	Bastia	3	2	0.0
1	1999	Monaco	St Etienne	2	2	2.0
2	1999	Auxerre	Nancy	2	1	0.0
3	1999	Lyon	Montpellier	1	2	1.0
4	1999	Marseille	Sedan	3	0	0.0

הוספנו את עמודת WINNER כאשר:
0 - ניצחון בית
1- ניצחון חוץ
2- ללא הכרעה

ב. Premier League - ליגה אנגלית -

כל התוצאות הסופיות של הליגה מ 1993 עד 2021:

	year	Home Team	Away Team	Home Team Goals	Away Team Goals	Winner
0	1993	Arsenal	Coventry	0	3	1.0
1	1993	Aston Villa	QPR	4	1	0.0
2	1993	Chelsea	Blackburn	1	2	1.0
3	1993	Liverpool	Sheffield Weds	2	0	0.0
4	1993	Man City	Leeds	1	1	2.0

הוספנו את עמודת WINNER כאשר:
0 - ניצחון בית
1- ניצחון חוץ
2- ללא הכרעה

ג. Laliga - ליגה ספרדית

כל התוצאות הסופיות של עונות 2014 עד 2021

	year	Home Team	Away Team	Home Team Goals	Away Team Goals	Winner	Home Team Points	Away Team Points
0	2014	MALAGA	ATHLETIC	1	0	0.0	3.0	0.0
1	2014	SEVILLA FC	VALENCIA	1	1	2.0	1.0	1.0
2	2014	GRANADA	DEPORTIVO	2	1	0.0	3.0	0.0
3	2014	ALMERIA	ESPANYOL	1	1	2.0	1.0	1.0
4	2014	EIBAR	REAL SOCIEDAD	1	0	0.0	3.0	0.0

הוספנו את עמודת:

WINNER כאשר:
0 - ניצחון בית
1- ניצחון חוץ
2- ללא הכרעה

Home/Away Team points כאשר
3 - נקודות נצברות עבור ניצחון
1 - נקודה נצברת עבור תיקו
0 - ללא נקודות עבור הפסד

שאלה 1:

בהינתן 30 מחזורים ראשונים של עונה עם תוצאות המשחק נרצה לדעת אם בהינתן התוצאות הסופיות של שמונת המחזורים האחרונים המסווגים שלנו יהיו מספיק "פקחים" כדי לדעת לחזות את זהות המנצחת.
לצורך כך השתמשנו במאגר של הליגה האנגלית (בשנת 2020-2021).
כלומר אימנו את המודלים השונים על 30 המשחקים הראשונים ובחנו אותם על שמונת מחזורי הסיום.

- אנו מצפים מכל המודלים לחזות ב 100 אחוזי הצלחה.

מודל ראשון : Logistic Regression

בחרנו במודל זה כי הוא יכול לעשות פעולה מתמטית פשוטה על הפיצ'רים ("Home Team Goals" - "Away Team Goals") ואז להשתמש בפונקצית סיגמויד ולחזות בצורה נכונה.

התוצאות :

```
Logistic Regression Accuracy:  1.0

Logistic Regression Matrix:
[[32  0  0]
 [ 0 35  0]
 [ 0  0 13]]

Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        32
         1.0       1.00      1.00      1.00        35
         2.0       1.00      1.00      1.00        13

    accuracy                           1.00        80
   macro avg       1.00      1.00      1.00        80
weighted avg       1.00      1.00      1.00        80
```

כצפוי המודל ידע להכריע מהי זהות המנצחת.

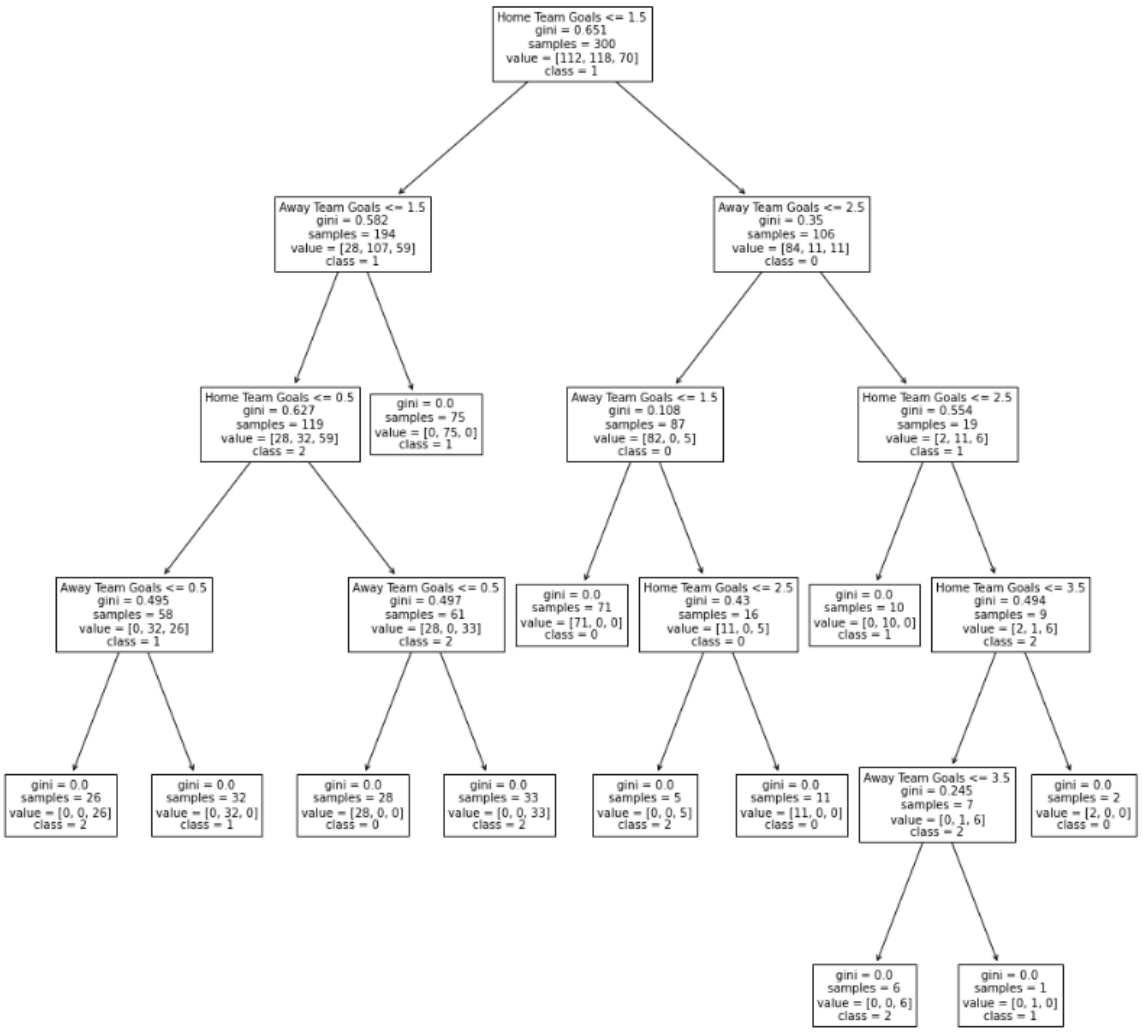
מודל שני : Decision Trees

רצינו לבדוק אם ה Decision Trees יכול להשוות בין 2 פיצ'רים היות והוא משתמש בשיטה לגילוי פיצ'רים חשובים (הוא יכול להשוות בין שערי בית לשערי חוץ).

תוצאות :

```
Decision Tree Accuracy :  1.0

Decision Tree Matrix:
[[32  0  0]
 [ 0 35  0]
 [ 0  0 13]]

Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        32
         1.0       1.00      1.00      1.00        35
         2.0       1.00      1.00      1.00        13

    accuracy                           1.00        80
   macro avg       1.00      1.00      1.00        80
weighted avg       1.00      1.00      1.00        80
```

התוצאות אכן טובות , נרצה לדעת מה היו קודקודי ההחלטה ועומק העץ :



כפי שניתן לראות אין כאן השוואה בין הפיצ'רים עצמם, אלא עומדן של כל פיצ'ר (באיזה טווח ערכים הוא נמצא).
על בסיס עומדן זה המודל ידע לסווג נכונה.

כל מסלול בעץ סוגר טווחים על נקודות הקבוצות. ועל בסיס אותו טווח ניתן לדעת בדיוק את התוצאה הסופית בעלה ואת זהות המנצחת.

לדוגמא : העלה הרביעי משמאל נותן לנו 33 דוגמאות מסווגות לתוצאה 1-1 :

```
Entrée [150]: X_train[(X_train['Home Team Goals'] == 1 ) & (X_train['Away Team Goals'] == 1 )].count()

   Out[150]: Home Team          33
             Away Team          33
             Home Team Goals    33
             Away Team Goals    33
             dtype: int64
```

מודל שלישי : Adaboost :

נסיון 1:

טכניקה נוספת לסיווג שרצינו לבחון הינה ADABOOST. בנוסף רצינו לבחון האם מסווגים חלשים שלכאורה יכולים לעשות את העבודה יהיו מספיק טובים.

- ציפינו שהמודל יתן לנו גם 100 אחוזי הצלחה אך להפתעתנו התוצאות :

```
AdaBoost Accuracy :  0.7375

AdaBoost Matrix:
[[17  0 15]
 [ 0 29  6]
 [ 0  0 13]]

Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      0.53      0.69        32
         1.0       1.00      0.83      0.91        35
         2.0       0.38      1.00      0.55        13

    accuracy                           0.74        80
   macro avg       0.79      0.79      0.72        80
weighted avg       0.90      0.74      0.76        80
```

לאחר חקירה במודל ADABOOST של SKLEARN המסווג הדיפולטיבי הוא עץ החלטה בעומק 1 וכפי שראינו עצי ההחלטה משתמשים בעומדנים ולא בהשוואת פיצ'רים ולכן זה כנראה לא הספיק (המסווגים היו חלשים מדי).

נסיון 2:

כעת הגדרנו את עץ ההחלטה בו משתמש המודל להיות בעומק 2 ומכאן התוצאות :

```
AdaBoost Accuracy :  1.0

AdaBoost Matrix:
[[32  0  0]
 [ 0 35  0]
 [ 0  0 13]]

Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        32
         1.0       1.00      1.00      1.00        35
         2.0       1.00      1.00      1.00        13

    accuracy                           1.00        80
   macro avg       1.00      1.00      1.00        80
weighted avg       1.00      1.00      1.00        80
```

After "boosting" the model we got perfect results.

מודל רביעי : Naive Bayes

רצינו לבדוק מודל הסתברותי עם אי תלות בין הפיצ'רים השונים. התוצאות :

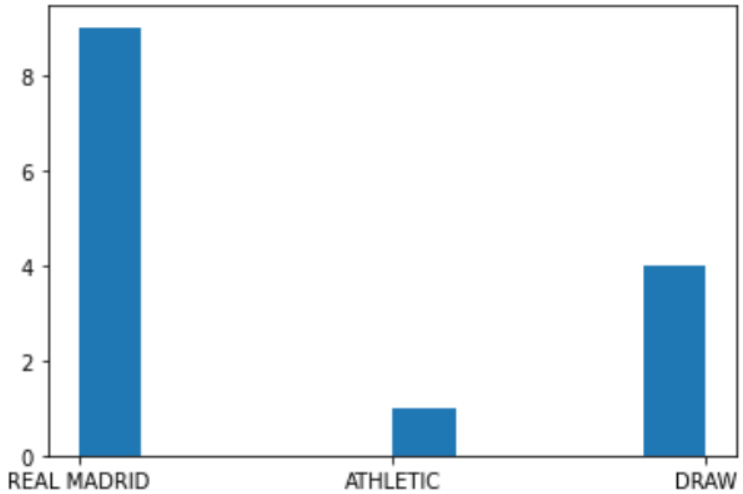
```
Naive Bayes Accuracy :  0.925

Naive Bayes Matrix:
[[29  0  3]
 [ 0 32  3]
 [ 0  0 13]]

Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      0.91      0.95        32
         1.0       1.00      0.91      0.96        35
         2.0       0.68      1.00      0.81        13

    accuracy                           0.93        80
   macro avg       0.89      0.94      0.91        80
weighted avg       0.95      0.93      0.93        80
```

כפי שניתן לראות , המודל לא עבד בצורה משולמת מכיוון שישנה תלות מסוימת בין פיצ'רים שונים (מנצח תלוי בשערי חוץ ובית יחד) ולכן המודל לא נתן מאה אחוז הצלחה.

שאלה 2:

בהינתן זהות המנצחת בין השנים 2014-2019 במשחק המסקרן שיש בליגה הספרדית בין ריאל מדריד לבין אתלטיק בילבאו (משחק שיש בו הרבה פוליטיקה בספרד - הבאסקים נגד הקסטילינים)
רצינו לחזות את זהות המנצחת ב2 המפגשים של עונת 2021 ושל 2 המפגשים בעונת 2022 (הכנסנו ידנית את תוצאות המשחק של השנה הנוכחית שלא הייתה קיימת בדאטהסט)

ניתוח מספר הנצחונות:



מכיוון שהדאטה לא מאוזן , נצפה מהמודלים השונים לחזות שריאל מדריד תנצח.

מודל Logistic Regression :

השתמשנו במודל זה כי הוא יודע להתמודד טוב עם דאטה שאינו מאוזן תוצאות:

עבור שנת 2021:

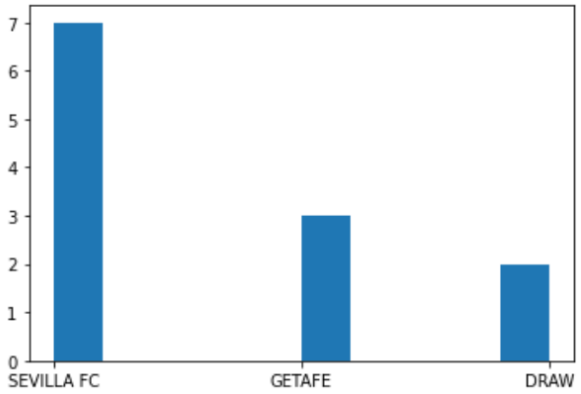
```
Logistic Regression Accuracy:  1.0

Logistic Regression Matrix:
[[1 0]
 [0 1]]

Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00         1
         1.0       1.00      1.00      1.00         1

    accuracy                           1.00         2
   macro avg       1.00      1.00      1.00         2
weighted avg       1.00      1.00      1.00         2
```

עבור שנת 2022:

```
True:
    year    Home Team    Away Team  Winner
0  2021  REAL MADRID     ATHLETIC       0
1  2021     ATHLETIC  REAL MADRID       1

Prediction:
    year    Home Team    Away Team  Winner
0  2021  REAL MADRID     ATHLETIC     0.0
1  2021     ATHLETIC  REAL MADRID     1.0
```

כפי שניתן לראות המודל עבד בצורה טובה עם הדאטה הלא מאוזנת וענה נכון (הטה לטובת ריאל מדריד).

כעת ננסה לענות על אותה שאלה אך במשחק שהדאטה בו הוא קצת יותר מאוזן. לצורך כך נשתמש במודל Naive Bayes בדאטה של המשחק סביליה נגד חטאפה (משחק בו 2 הקבוצות יותר שוות כוחות). השתמשנו במודל זה משום שהוא משתמש בהסתברויות ומחזיר את המחלקה עם ההסתברות הגבוה ביותר לכן יתאים יותר לדאטה יותר מאוזנת . הדאטה :



כפי שניתן לראות סך הנצחונות של חטאפה ותוצאות התיקו **מעט** קטן יותר מסך הנצחונות של סביליה (בניגוד לדאטה הקודם בו הכף היית מוטת משמעותית לטובת ריאל מדריד).

תוצאות :

```
Naive Bayes Accuracy :  1.0

Naive Bayes Matrix:
[[1 0]
 [0 1]]

Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00         1
         1.0       1.00      1.00      1.00         1

    accuracy                           1.00         2
   macro avg       1.00      1.00      1.00         2
weighted avg       1.00      1.00      1.00         2
```

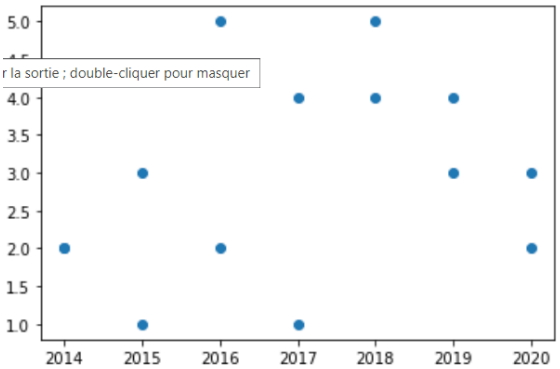
כפי שניתן לראות המודל מחזיר תוצאה טובה על אף שהדאטה יחסית מאוזנת , ההבדל הקטן הוא זה שהכריע לטובת זהות המנצחת.

שאלה 3 :

אנו נרצה לחזות כמה שערים יובקעו במשחק המסקרן בין ריאל מדריד לויאריאל בשנת 2021 ,לצורך כך הוספנו עמודה שבה מופיעים סך השערים של כל משחק בין 2 הקבוצות בין השנים 2014-2020 :

	year	Home Team	Away Team	Total goals
51	2014	VILLARREAL	REAL MADRID	2
248	2014	REAL MADRID	VILLARREAL	2
529	2015	VILLARREAL	REAL MADRID	1
717	2015	REAL MADRID	VILLARREAL	3
802	2016	REAL MADRID	VILLARREAL	2
999	2016	VILLARREAL	REAL MADRID	5
1322	2017	REAL MADRID	VILLARREAL	1
1515	2017	VILLARREAL	REAL MADRID	4
1689	2018	VILLARREAL	REAL MADRID	4
1877	2018	REAL MADRID	VILLARREAL	5
1929	2019	VILLARREAL	REAL MADRID	4
2262	2019	REAL MADRID	VILLARREAL	3
2372	2020	VILLARREAL	REAL MADRID	2
2656	2020	REAL MADRID	VILLARREAL	3

גרף של סך השערים לאורך השנים :



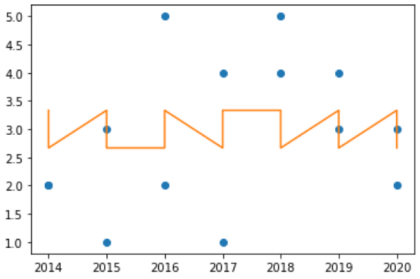
נרצה לבחון אם לשנים עצמם ישנה השפעה לכן חילקנו ל2 נסיונות :

נסיון 1: ללא שימוש בשנים בשימוש בLinear Regression:

בחרנו במודל הנ"ל מכיוון שהוא טוב לחיזוי רגרסיה (לא סיווג כלשהו) . המשחקים שנרצה לחזות :

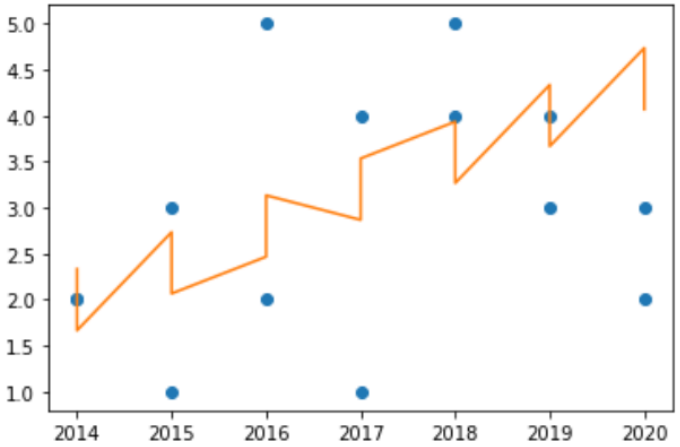
2372	2020	VILLARREAL	REAL MADRID	2
2656	2020	REAL MADRID	VILLARREAL	3

תוצאות :
בשני המשחקים המודל חזה שיובקעו 3 שערים. מכיוון שלא כללנו את השנים המודל ראה רק את הקבוצות ואת סך השערים ולכן עשה מעין ממוצע של סך השערים והחזיר 3 . הקו הלינארי:



כעת נרצה לבחון אם המודל משתנה כשנכלול גם את השנים.

תוצאות :
עבור המשחק הראשון המודל חזה 5 שערים ועבור המשחק השני הוא חזה 4 שערים.(בפועל היו 2,3).כפי שניתן לראות בגרף של סך השערים לאורך השנים , ישנה עליה בכמות השערים לאורך השנים , לכן המודל ציפה שהעלייה תימשך גם עבור שנת 2021 (מה שלא קרה בפועל) ולכן שגה. הקו הלינארי :



מסקנה - חיזוי לפי ממוצע לעתים מספק תוצאות לא רעות.

שאלה 4.א:

בעת סיום עונת המשחקים הסדירה (לאחר 38 משחקים) נקבעים זהות האלופה (מי שמסיימת ראשונה), זהות היורדות לליגת המשנה(שלושת המקומות האחרונים) וזהות העולות לליגת האלופות (בליגה הספרדית ארבעת המקומות הראשונים). אנו נרצה לחקור עם המודלים השונים בהינתן טבלת ניקוד סופית האם המודלים ידעו לסווג אלו מהקבוצות יורדות ליגה , אלו עולות לליגת האלופות ואלו מסיימות עונה ללא השפעה כלשהי. לצורך כך יצרנו טבלת ניקוד סופית לכל עונת משחקים בליגה הספרדית בין השנים 2014-2020 :
כמו כן הוספנו עמודה שמציינת את ההשפעה של כל קבוצה לעונה הבאה (כאשר 0 מסמל על ירידת ליגה , 1 סטטוס קוו ו2 עלייה לליגת האלופות)

	year	Home Team	points	pred
0	2014	CORDOBA	20.0	0.0
1	2014	ALMERIA	32.0	0.0
2	2014	DEPORTIVO	35.0	0.0
3	2014	EIBAR	35.0	1.0
4	2014	GRANADA	35.0	1.0
...	...	...	...	...
15	2020	REAL SOCIEDAD	62.0	1.0
16	2020	SEVILLA FC	77.0	2.0
17	2020	BARCELONA	79.0	2.0
18	2020	REAL MADRID	84.0	2.0
19	2020	ATLETICO MADRID	86.0	2.0

- לא ערבבנו את הדאטה לכן נצפה שגם מודלים מסווגים חלשים ידעו להתמודד עם הנ"ל.

נסיון ראשון עם מודל ADABOOST:

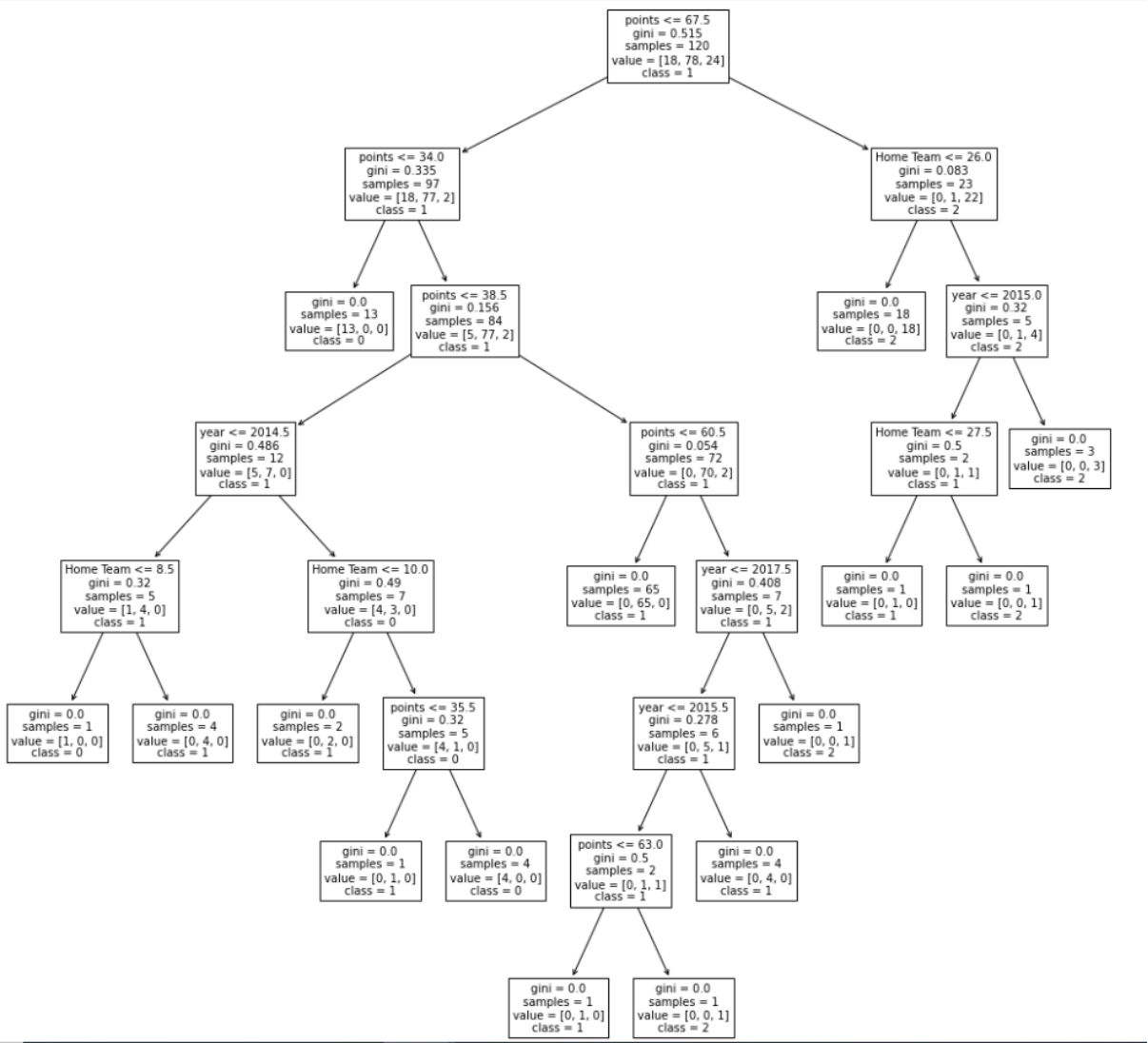
בנסיון זה השתמשנו בברירת מחדל (עומק עץ ההחלטה הינו 1) תוצאות :

```
AdaBoost Accuracy :  0.6

AdaBoost Matrix:
[[ 2  1  0]
 [ 3 10  0]
 [ 0  4  0]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.40      0.67      0.50         3
         1.0       0.67      0.77      0.71        13
         2.0       0.00      0.00      0.00         4

    accuracy                           0.60        20
   macro avg       0.36      0.48      0.40        20
weighted avg       0.49      0.60      0.54        20
```

- בעומק 1 העומדן הוא ככל הנראה או על מספר הנקודות שהצטברו על מס' הקבוצה , מה שלא מספק כמובן.

נסיון שני עם מודל Decision Tree :

עץ ההחלטה :



התוצאות :

```
Decision Tree Accuracy :  0.85

Decision Tree Matrix:
[[ 3  0  0]
 [ 1 10  2]
 [ 0  0  4]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.75      1.00      0.86         3
         1.0       1.00      0.77      0.87        13
         2.0       0.67      1.00      0.80         4

    accuracy                           0.85        20
   macro avg       0.81      0.92      0.84        20
weighted avg       0.90      0.85      0.85        20
```

כפי שניתן לראות התוצאות טובות יותר ממודל ADABOOST (מכיוון שכאן אנו לוקחים בחשבון גם את כל הפרמטרים , קודקודי ההחלטה) אך עדיין התוצאות לא הכי מספקות.

נסיון שלישי : מודל ADABOOST עם עץ החלטה בעומק 3 :

תוצאות :

```
AdaBoost Accuracy :  0.9

AdaBoost Matrix:
[[ 3  0  0]
 [ 2 11  0]
 [ 0  0  4]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.60      1.00      0.75         3
         1.0       1.00      0.85      0.92        13
         2.0       1.00      1.00      1.00         4

    accuracy                           0.90        20
   macro avg       0.87      0.95      0.89        20
weighted avg       0.94      0.90      0.91        20
```

ניתן להגיע לתוצאות יותר טובות עם מספר רב של מסווגים חלשים משמעותית. כפי שנאמר " טובים השניים מן האחד " .

שאלה 4.ב

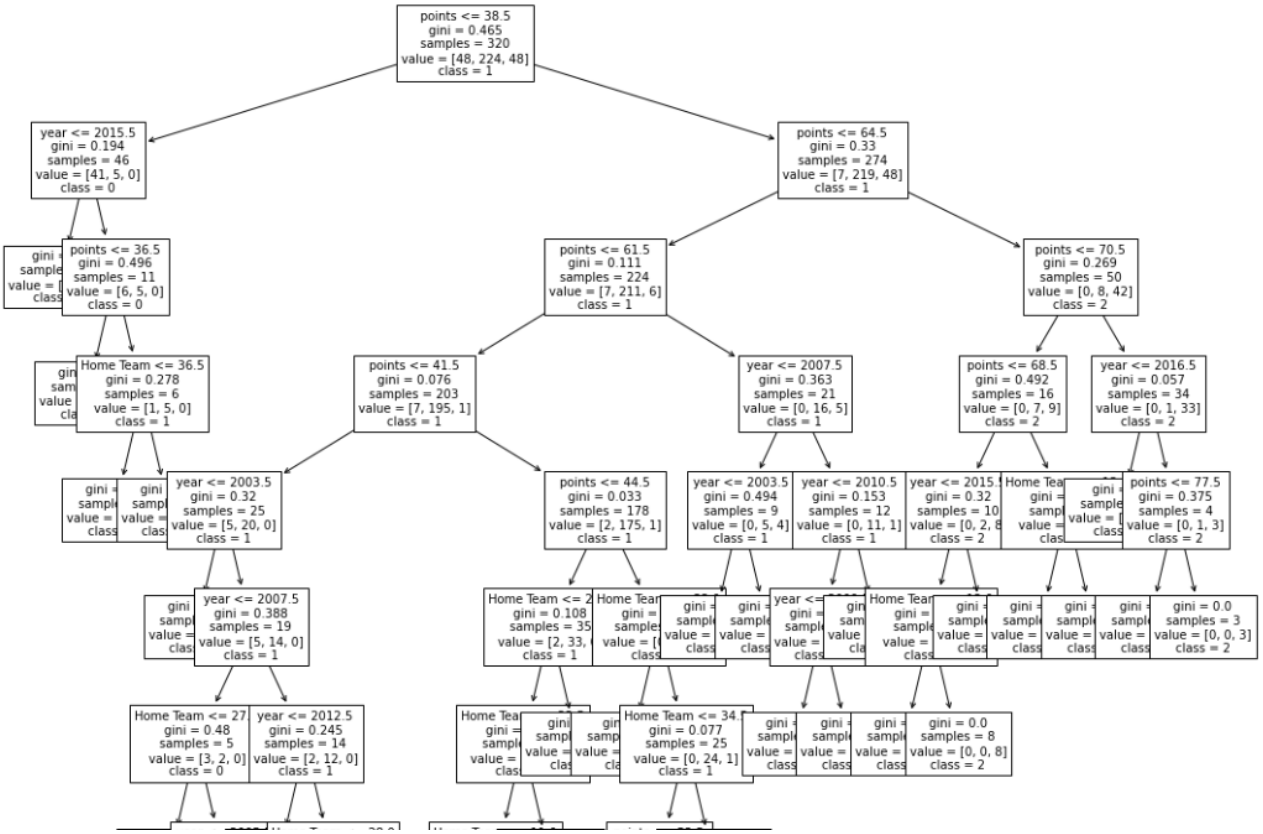
כעת נרצה לענות על אותה שאלה כמו בסעיף הקודם רק עם ערבוב של הדאטה . כמו כן הדאטה שלנו תהיה הליגה הצרפתית בין השנים 2002-2019 (בצרפת ישנם רק 3 עולות לליגת האלופות). הדאטה :

	year	Home Team	points	pred
0	2002	Troyes	31.0	0.0
1	2002	Sedan	36.0	0.0
2	2002	Le Havre	38.0	0.0
3	2002	Ajaccio	39.0	1.0
4	2002	Montpellier	40.0	1.0
...	...	...	...	...
335	2018	Marseille	61.0	1.0
336	2018	St Etienne	66.0	1.0
337	2018	Lyon	72.0	2.0
338	2018	Lille	75.0	2.0
339	2018	Paris SG	91.0	2.0

340 rows × 4 columns

נסיון ראשון - עץ החלטה

עץ ההחלטה:



התוצאות :

```
Decision Tree Accuracy :  0.8

Decision Tree Matrix:
[[ 3  0  0]
 [ 2 12  0]
 [ 0  2  1]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.60      1.00      0.75         3
         1.0       0.86      0.86      0.86        14
         2.0       1.00      0.33      0.50         3

    accuracy                           0.80        20
   macro avg       0.82      0.73      0.70        20
weighted avg       0.84      0.80      0.79        20
```

כפי שניתן לראות , אמנם עץ ההחלטה הוא מורכב יותר (נובע מכך שבדאטה הנוכחית ישנם יותר שנים מאשר בסעיף הקודם) אך הערבוב לא השפיע על קודקודי ההחלטה. אנו מניחים שהתוצאות פחות טובות כי הליגה הצרפתית פחות יציבה מהליגה הספרדית.

נסיון שני : Naive Bayes

התוצאות :

```
Naive Bayes Accuracy :  0.9

Naive Bayes Matrix:
[[ 3  0  0]
 [ 2 12  0]
 [ 0  0  3]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.60      1.00      0.75         3
         1.0       1.00      0.86      0.92        14
         2.0       1.00      1.00      1.00         3

    accuracy                           0.90        20
   macro avg       0.87      0.95      0.89        20
weighted avg       0.94      0.90      0.91        20
```

להפתעתנו המודל הבייסיאני השיג תוצאות טובות יותר מהמודל של עץ ההחלטה , הדבר יכול לנבוע מהעובדה שמודל בייסיאני מניח שאין תלות בין הפיצ'רים השונים מה שבמקרה הזה עוזר לנו לשפר את התוצאות שכן אין תלות בקבוצה ביחס לשאר הפיצ'רים (לעומת עץ ההחלטה שבה הקבוצה כן נלקחת בחשבון).

נסיון נוסף שימוש ב Logistic Regression :

תוצאות :

```
Logistic Regression Accuracy:  0.85

Logistic Regression Matrix:
[[ 3  0  0]
 [ 3 11  0]
 [ 0  0  3]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.50      1.00      0.67         3
         1.0       1.00      0.79      0.88        14
         2.0       1.00      1.00      1.00         3

    accuracy                           0.85        20
   macro avg       0.83      0.93      0.85        20
weighted avg       0.93      0.85      0.87        20
```

רצינו לנסות מודל נוסף כדי לראות אם ישנו שיפור כלשהו , במודל זה לא ראינו שיפור ניתן להסביר זאת בעבודה שהמודל נותן משקולת (אמנם קטנה כנראה אבל עדיין קיימת) לפיצ'ר הקבוצה.

עוד נסיון שימוש ב ADABOOST עם עץ החלטה בעומק 1:

בהמשך לנסיונות הקודמים במודל זה ובהמשך לנאמר לעיל התוצאות לא היו שונות :

```
AdaBoost Accuracy :  0.8

AdaBoost Matrix:
[[ 3  0  0]
 [ 3 10  1]
 [ 0  0  3]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.50      1.00      0.67         3
         1.0       1.00      0.71      0.83        14
         2.0       0.75      1.00      0.86         3

    accuracy                           0.80        20
   macro avg       0.75      0.90      0.79        20
weighted avg       0.89      0.80      0.81        20
```

כעת ננסה עם ADABOOST עם עץ החלטה בעומק 3 :

תוצאות :

```
AdaBoost Accuracy :  0.95

AdaBoost Matrix:
[[ 3  0  0]
 [ 1 13  0]
 [ 0  0  3]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.75      1.00      0.86         3
         1.0       1.00      0.93      0.96        14
         2.0       1.00      1.00      1.00         3

    accuracy                           0.95        20
   macro avg       0.92      0.98      0.94        20
weighted avg       0.96      0.95      0.95        20
```

ואכן כפי שנאמר " טובים השניים מן האחד"

שאלה מס' 5:

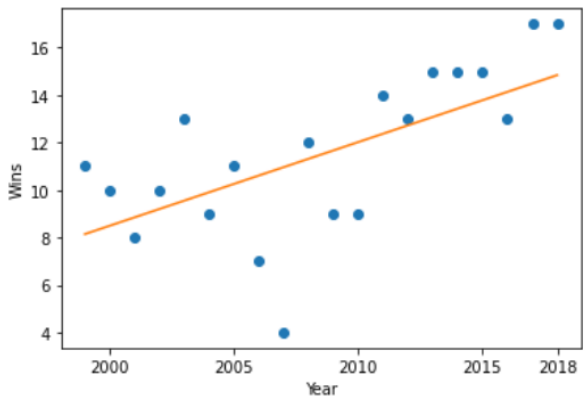
בשאלה זו נרצה לחזות את העוצמה של מגרש הבית של הקבוצה הדומיננטית ביותר בליגה הצרפתית , הלא היא פאריס סן ז'רמן. נרצה לדעת כמה נצחונות בית תשיג פאריס בעונת 2018-2019 בהינתן כל התוצאות של משחקי הבית שלה בשנים האחרונות .

	year	Winner
0	1999	11
1	2000	10
2	2001	8
3	2002	10
4	2003	13
5	2004	9
6	2005	11
7	2006	7
8	2007	4
9	2008	12
10	2009	9

טבלה זו מראה לנו כמה נצחונות בית השיגה פאריס בעונה מסוימת .

לצורך החיזוי נשתמש במודל Linear Regression

כפי שראינו מודל זה טוב לחיזוי רגרסיה.
התוצאות :
המודל חיזה לנו שפאריס תנצח בעונת 2018-2019 בסך הכול 15 משחקים . (בפועל פאריס ניצחה 17 משחקים).
הקו הלינארי :



כפי שניתן לראות עד שנת 2012 (השנה בה הגיעו המשקיעים הקטארים והשקיעו כסף במועדון) פריז הייתה קבוצת ליגה ממוצעת , לאחר מכן פריז השתפרה משמעותית בסך נצחונות הבית אך השנים הללו השפיעו על המודל ולכן הוא לא חזה בצורה מדויקת.