פרויקט מסכם בקורס "מחט בערימת דאטה" (67978)

1. **שם הפרויקט**: חיזוי בחירת שחקני קולג' ב- NBA Draft לשנת 2017.

2. **חברי הקבוצה**: אייל וילנסקי (eyal.vil , eyal.vilensky@mail.huji.ac.il) , אריק גופמן (guffi, eric.gofman@mail.huji.ac.il) וגבריאל יהב (qt05w, gabriel.yahav@mail.huji.ac.il)

3. **תיאור הבעיה :** חיזוי 11 הבחירות הראשונות ל NBA Draft הקרוב מבין שחקני המכללות בארה"ב בעונת 2016/2017 .

4. **נתונים** : ה Data בפרויקט שלנו נלקח מהאתר http://www.sports-reference.com והוא נחלק ל 3 קבוצות - training data, test data והנתונים על-גביהם מבוצע החיזוי .

4.1. training data - קובץ xlsx בשם 2011-2016_All_Players (צורף ל ZIP ההגשה), המכיל נתונים סטטיסטיים (שם שחקן, קטגוריית גיל, המחוז בו שיחק,עמדה, אחוזים מהשדה, אחוזים מ 3, אחוזים מעונשין, ממוצע נקודות , ממוצע עבירות , ממוצע אסיסטים, ממוצע חטיפות , ממוצע איבודים, ממוצע נקודות, ממוצע ריבאונדים התקפה והגנה) של שחקני מכללות ששיחקו בין העונות 2011-2012 ל 2015-2016 והעמידו שורה סטטיסטית כמפורט : שיחקו לפחות 20 משחקים בעונה וגם בממוצע קלעו לפחות 13 נק' למשחק או מסרו לפחות 2 אסיסטים בממוצע או קטפו לפחות 5 ריבאונדים בממוצע או חסמו לפחות 1.5 פעמים בממוצע.
כל entry ב training data תויג בעזרת label שחושב באופן הבא :
$120 - 2 * (position - 1)$, כאשר position מציין את המיקום בדראפט שנבחר השחקן (אם לא נבחר אז 0). סה"כ כ- 3000 רשומות. בנוסף לקובץ אשר מכיל את הנתונים על כלל השחקנים, הוספנו שלושה קבצים נוספים אשר בהם חילקנו את השחקנים ע"פ העמדה שלהם – גארדים, פורוורדים וסנטרים – זאת על מנת לבחון האם ה-coefficients משתנים בין עמדה לעמדה, וע"י כך להשיג פרדיקציה טובה יותר בהתאם לעמדה בה משחק השחקן.
יש לציין כי את ארבעת הקבצים הללו המרנו לקבצי טקסט כדי לאפשר קריאה נוחה יותר שלהם.

4.2. test data- קובץ csv בשם 2010-2011(צורף ל zip ההגשה), המכיל נתונים סטטיסטיים (אותם המדדים בדומה ל training data) של שחקני מכללות ששיחקו בעונת 2010-2011 והעמידו שורה סטטיסטית כמפורט: שיחקו לפחות 20 משחקים בעונה וגם שיחקו בממוצע לפחות 15 דקות במשחק וגם קלעו לפחות 5 נקודות בממוצע. סה"כ כ 2000 רשומות.

4.3. הנתונים על-גביהם מבוצע החיזוי - קובץ csv בשם 2016-2017 (צורף ל zip ההגשה) , המכיל נתונים סטטיסטיים(אותם מדדים כמו מקודם) של שחקני מכללות שמשחקים בעונה הנוכחית, קרי 2016-2017, והעמידו עד כה שורה סטטיסטית כמפורט : שיחקו לפחות 13 משחקים וגם משחקים בממוצע 15 דקות למשחק וגם קלעו לפחות 5 נקודות בממוצע. סה"כ כ-2000 רשומות

5. <u>הפתרון המוצע</u> : על-מנת לפתור את הבעיה הנתונה השתמשנו במודל חיזוי מסוג רגרסיה ליניארית מרובה:

ראשית, ייצגנו את השורה הסטטיסטית של שחקן כלשהו כ feature vector שבו כל feature מייצג אחד מבין המדדים הבאים : קבוצת גיל, המחוז בו שיחק השחקן, ממוצע נקודות, ממוצע אסיסטים, ממוצע ריבאונדים בהתקפה ובהגנה, ממוצע חסימות, ממוצע איבודים, ממוצע חטיפות, אחוזים מהשדה, אחוזים מהעונשין, אחוזים מ 3. לאחר מכן, בנוסף לקובץ שבו שמרנו את הנתונים על כלל השחקנים, הוספנו שלושה קבצים מחולקים ע"פ העמדה בה השחקן משחק – גארדים, פורוורדים וסנטרים.

השלב הבא בתהליך הוא שלב הלמידה . כתבנו סקריפט בשפת R בשם **DraftPrediction** שמקבל כקלט 4 קבצים המכילים training data , ולומד את ה coefficient של כל feature עבור כל קובץ: הקובץ עם כלל נתוני השחקנים, וקובץ עבור כל עמדה של שחקן. הסקריפט כותב את הפלט שלו ל-4 קבצי טקסט, אחד עבור כל קובץ קלט אשר נמצאים בתיקייה **Vectors**.

עבור שלב הפרדיקציה כתבנו סקריפט פייתון בשם DraftPredictor, אשר מקבל כקלט קובץ עם נתוני שחקנים על-גביהם מעוניינים לבצע חיזוי. הסקריפט מחשב את ציון השחקן באופן הבא:

- מבצע מכפלה פנימית בין נתוני השחקן לפלט של קובץ ה-R עבור כלל נתוני השחקנים, כלומר מחשב את ציון השחקן ע"פ ה-Coefficients הכלליים.
- מבצע מכפלה פנימית בין נתוני השחקן לפלט של קובץ ה-R עבור העמדה הספציפית של השחקן, כלומר מחשב את ציון השחקן ע"פ ה-Coefficients של העמדה שלו.
- סוכם ביניהם

לבסוף, מדפיס הסקריפט את 11 השחקנים עם הציון הגבוה ביותר – כלומר 11 השחקנים שע"פ נתוניהם האישיים ייבחרו באחד מ 11 המקומות הראשונים בדראפט הקרוב.

6. <u>הניסויים</u>:

- קריטריון ההערכה: החלטנו להעריך את התוצאות שלנו אל מול אתרי הספורט המובילים ESPN ו-DraftExpress, אשר מספקים חיזוי משלהם לבחירות הדראפט הקרוב. קריטריון ההערכה הראשוני שחשבנו עליו היה חישוב סכום ההפרשים בין התוצאות שלנו לבין החיזוי של אחד מהאתרים – כלומר אם שחקן, ע"פ הסקריפט שלנו, צפוי להיבחר ראשון אך באחד האתרים הוא רק רביעי – נוסיף לסכום 3 - כך נעבור על כלל השחקנים אשר התקבלו ב-11 הבחירות הראשונות, וככל שהתוצאה נמוכה יותר אז החיזוי שלנו מוצלח יותר. בהמשך, הגענו למסקנה שכפי הנראה המספר שיתקבל יהיה חסר משמעות (לדוגמא, אם שחקן חזוי להיבחר ראשון ע"י ESPN ואצלנו כלל לא מופיע אז התוצאה שלנו לא תיפגע) החלטנו כי אנו מעוניינים רק לבחון את התחזית שלנו ביחס לשחקנים הבולטים. פירושו של דבר : **קריטריון ההערכה יהיה אחוז השחקנים שצפויים להיבחר ב-5 המקומות הראשונים באתרי החיזוי, אשר מופיעים אצלנו ב-11 הראשונים.** לאחר בירור קצר מצאנו כי האתר DraftExpress נחשב לאמין, ולכן החלטנו לבדוק את התוצאות שלנו ביחס לתוצאותיו שלו. בתוך כך, קבענו כי הצלחה תיחשב ל-80% ומעלה.

על-מנת להעריך את ביצועיי האלגוריתם שלנו ואת איכות תוצאותיו ביצענו 2 ניסויים:

❖ ניסוי 1 : בניסוי זה מדדנו עד כמה תוצאות החיזוי של האלגוריתם לדראפט הקרוב (2017) זהה להערכות שמתפרסמות באתרי ספורט מובילים. לשם כך, ליקטנו מהאתרים draftexpress ו ESPN, הנחשבים מובילים בתחום חיזוי בחירות דראפט, את הערכותיהם אילו שחקנים ייבחרו ב X המקומות הראשונים בדראפט הקרוב. אחר כך, כפי שתואר בסעיף 4 - "הפתרון המוצע" , ביצענו פרדיקציה בעזרת סקריפט ה Python בשם DraftPredictor, שקיבל כקלט את קובץ ה csv של נתוני השחקנים מעונת 2016-2017 ואת קבצי המשקלים של ה features (שאותם מצאנו כאמור ע"י סקריפט ה R) והפלט היה 11 השחקנים הראשונים שאנו צופים שייבחרו בדראפט הקרוב. את התוצאות שקיבלנו השוונו לאלו שחזו EPSN ו draftexpress .
ניתן לראות את התוצאות שלנו בטבלה הבאה:

מיקום בתחזית שלנו	שם	קולג'	מיקום בתחזית של DraftExpress	מיקום בתחזית של ESPN
1	Markelle Fultz	Washington	1	1
2	Jonathan Isaac	Florida State	6	6
3	Dennis Smith Jr.	North Carolina State	4	5
4	Lonzo Ball	UCLA	2	2
5	Jayson Tatum	Duke	5	7
6	TJ Leaf	UCLA	23	15
7	Robert Williams	Texas A&M	12	16
8	Miles Bridges	Michigan State	11	12
9	Justin Patton	Creighton	13	13
10	Bruce Brown	Miami	NA	NA
11	Josh Jackson	Kansas	3	3

ניתן לראות כי כל השחקנים אשר חזויים להיבחר ב-5 הראשונה ע"פ DraftExpress גם מופיעים אצלנו ב-11 הבחירות הראשונות. כלומר 100% מהשחקנים.

❖ ניסוי 2 : בניסוי זה מדדנו עד כמה תוצאות החיזוי של האלגוריתם לדראפט שנערך בשנת 2011 זהה לבחירות האמיתיות בדראפט של אותה השנה. לשם כך, תחילה ליקטנו את בחירות הדראפט משנת 2011 בעזרת ויקיפדיה. אחר כך, כפי שתואר בסעיף 4 - "הפתרון המוצע" , ביצענו פרדיקציה בעזרת סקריפט ה Python בשם DraftPredictor, שקיבל כקלט את קובץ ה csv של נתוני השחקנים מעונת 2010-2011 ואת קבצי המשקלים של ה features (מיוצר

כאמור ע"י סקריפט ה R) והפלט היה 11 השחקנים הראשונים שהסקריפט צפה שנבחרו ב 2011. את התוצאות שקיבלנו השוונו לתוצאות האמת מאותה השנה.

התוצאות נראו כך:

שם שחקן	חיזוי שלנו	בחירת דראפט אמיתית
Kyrie Irving	1	1
Tobias Harris	2	19
Jordan Williams	3	36
Tristan Thompson	4	4
JaJuan Johnson	5	27
Marshon Brooks	6	25
Jordan Hamilton	7	26
Reggie Jackson	8	24
Brandon Knight	9	8
Iman Shumpert	10	17
Trey Thompkins	11	37

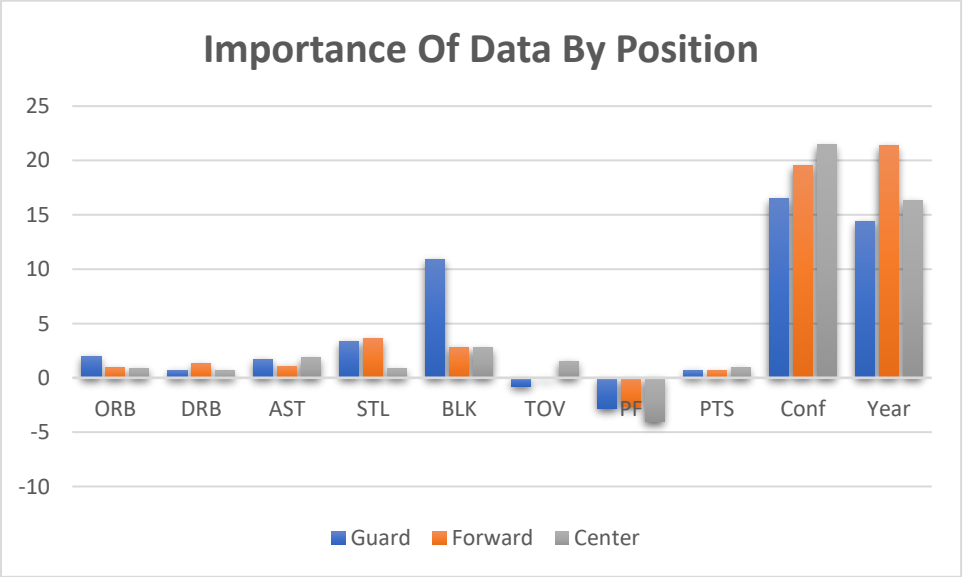
כאן ניתן לראות כי רק 2 שחקנים שנבחרו בחמשת הבחירות הראשונות מופיעים אצלנו ב-11. כלומר – 40% התאמה בלבד. אך אם נוריד את השחקנים אשר לא שיחקו במכללות ונבחרו בדראפט זה, נקבל כי ברנדון נייט, אשר חזינו כי ייבחר 9, נבחר 5 מבין שחקני המכללות, ועל כן ניתן לומר כי ישנם 60% התאמה.

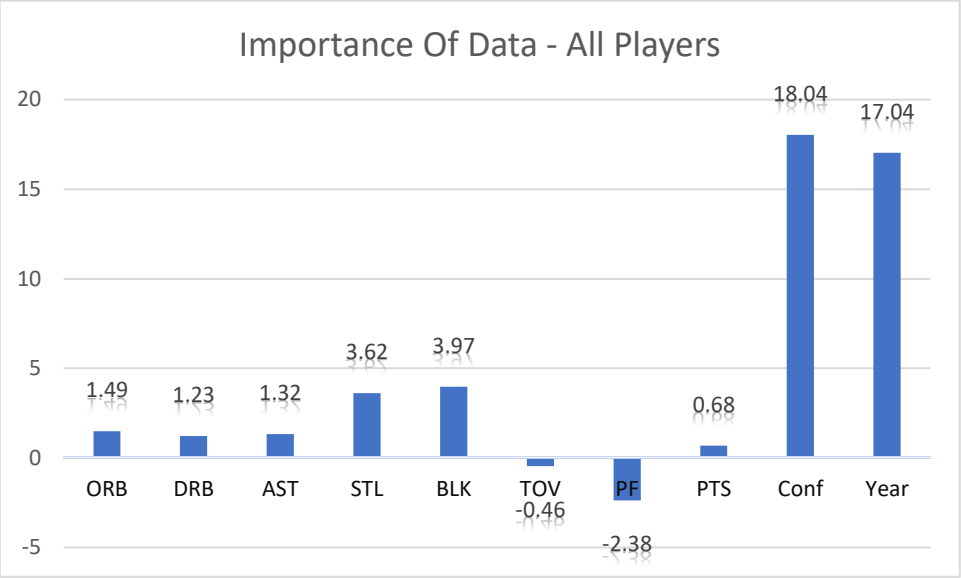
<u>חשיבות נתון סטטיסטי על מיקום בדראפט:</u>

כאמור, חילקנו את חשיבות הנתונים ל-2:

- נתונים כלליים, כלומר לכל השחקנים.
- נתונים ע"פ עמדה – גארד, פורוורד וסנטר.

בגרף, נדגים את חשיבותו של נתון מסויים לציון הסופי של השחקן – ככל שציונו של שחקן גבוה יותר כך ע"פ המודל שלנו הוא ייבחר גבוה יותר בדראפט.





הסבר קצר – בגרף אנו מראים במספרים עד כמה נתון משפיע על הציון הסופי של שחקן קולג'ים- לדוגמא, נניח כי לשחקן יש בממוצע 6 ORB (כלומר ריבאונדי התקפה) למשחק, אז זה יעלה לו את הציון ב-9 נקודות (כלומר 6*1.5). דבר זה נכון לכל הנתונים חוץ מ-Year(מס' השנים שהשחקן שיחק בקולג') ו-Conf(המחוז בו הקבוצה של השחקן משחקת) שאלו הם נתונים קטגוריים, כלומר לא ניתן למדוד אותם כמותית - בגרף זה אנו מציגים מהו המקסימום שנתון זה יכול להשפיע על הציון הסופי.

<u>קשיים בהם נתקלנו</u>

- איסוף הנתונים היה קשה מחשבנו ודרש מאמץ רב. כמות השחקנים שמשחקים בליגת הקולג'ים היא גדולה ועל כן החלטנו להתמקד רק בשחקנים בעלי בעלי נתוני סף מסוימים.
- בתקופה זו בשנה אין מידע בנוגע לאילו שחקנים בכלל מתכוונים להירשם לדראפט הקרוב, כלומר יכול להיות שאנו חוזים ששחקנים מסויימים ייבחרו בדראפט הקרוב על אף שהם כלל אינם מתכוונים לגשת. לדוגמא - אנו חוזים כי ברוס בראון מקבוצת מיאמי ייבחר במקום ה- 10 אולם עפ"י כתבות באינטרנט הוא בכלל מתכוון לגשת לדראפט 2018.
- מספר קבוצות רב: אין ספק כי הקולג' בו שחקן משחק משפיע על בחירתו בדראפט, אך מכיוון שישנו מספר רב של קבוצות לא היה באפשרותינו להפוך נתון זה למשתנה קטגורי. לפיכך החלטנו לבסוף להתפשר ולהשתמש במחוז בו משחקת הקבוצה לנתון קטגורי.
התפשרות זו פגעה בתוצאות שלנו בכך שקולג'ים חזקים אך משתתפים במחוז יחסית חלש נפגעו בציון, כך לדוגמא שחקנים ממכללת אריזונה, מכללה חזקה אך במחוז חלש לא הופיעו אצלנו בתחזיות.
- בסופו של דבר הרגרסיה הליניארית על הנתונים איננה מושלמת – ניתן לראות שהפלט של סקריפט ה-R הביא תוצאות עם Adjusted-R סביב ה 0.3 - נתון חלש יחסית, כלומר אין באמת התאמה מושלמת בין הנתונים לבין הבחירה בדראפט. יתכן ואם היינו משתמשים בסטטיסטיקות מעט יותר

מתקדמות (כמו מדד פלוס מינוס, או PER למשל) היינו מקבלים התאמה יותר גבוהה.

7. <u>תוספות אפשרויות לעתיד</u> :(שאינן חלק מהפרויקט במתכונתו הנוכחית)

- ❑ הוספת sentiment feature ל feature vector של כל שחקן - בעזרת Twitter API ו crawling על אתרי ספורט מובילים שמסקרים את ליגת המכללות נאתר ציוצים ומאמרים שבהם מתויגים ומאוזכרים שחקני מכללות פעילים מהעונה הנוכחית . לכל text נבצע sentiment analysis , כך שה sentiment הכולל של שחקן יהיה שקלול כלל ה sentiment-ים המקושרים אליו בטקסטים השונים. ה sentiment feature יאפשר לקבל ממד נוסף של חווי באשר לשחקן כדוגמת מאפייני אישיות והתנהגות שלא ניתן לדלות מדפי הסטטיסטיקה בלבד.
- ❑ הרחבת מאגר השחקנים שלגביהם מבוצע חיזוי האם ייבחרו בדראפט הקרוב, כך שייכלול גם שחקנים פעילים שאינם שחקני מכללות. הדבר ייעשה באמצעות שימוש ברשתות חברתיות לצורך איתור שחקנים שעתידים להגיש מועמדותם לדראפט ואיסוף סטטיסטיקות רלוונטיות עבור שחקנים בקטגורייה זו(לדוגמא אליפות עולם עד גיל 21, אליפויות יבשתיות עד גיל 21 וכו)

8. <u>סיכום קצר</u>:

חזינו את בחירות הדראפט הקרובות ע"י איסוף נתונים של שחקני קולג' משנים קודמות וקישורם לבחירות הדראפט אשר בהן השחקנים נבחרו. מצאנו כי המדדים אשר משפיעים באופן המשמעותי ביותר על בחירת שחקנים בדראפט הם:

- המחוז בו משחקת קבוצתו של שחקן – ככל ששחקן משתתף בליגה איכותית ו\או משחק בשירותיה של קבוצה מובילה(באותה הליגה בה משחקת) סיכויו להיבחר במקום טוב בדראפט גבוהים יותר.
- נתונים המתקשרים ליכולותיו האתלטיות של השחקן – ניתן לראות כי המדדים הקלאסיים המשפיעים ביותר על בחירתו של שחקן הינם חסימות וחטיפות – מדדים שבד"כ גבוהים במיוחד אצל שחקנים אתלטיים.

על אף שראינו כי התאמת נתוני השחקן למיקומו בדראפט ע"י רגרסיה ליניארית אינו מושלם כלל, עדיין שיטה זו סיפקה תוצאות מרשימות. שחקנים שאתרי הספורט המובילים מנבאים כי צפויים להיבחר בבחירות הראשונות בדראפט הופיעו גם בצמרת החיזוי שלנו, כלומר רגרסיה מרובה על חיזוי התוצאות.