פרויקט בינה עסקית- הגשה 3

עילי סולומון	208934299
גל נוברט	208287953
יובל אמסלם	208998963

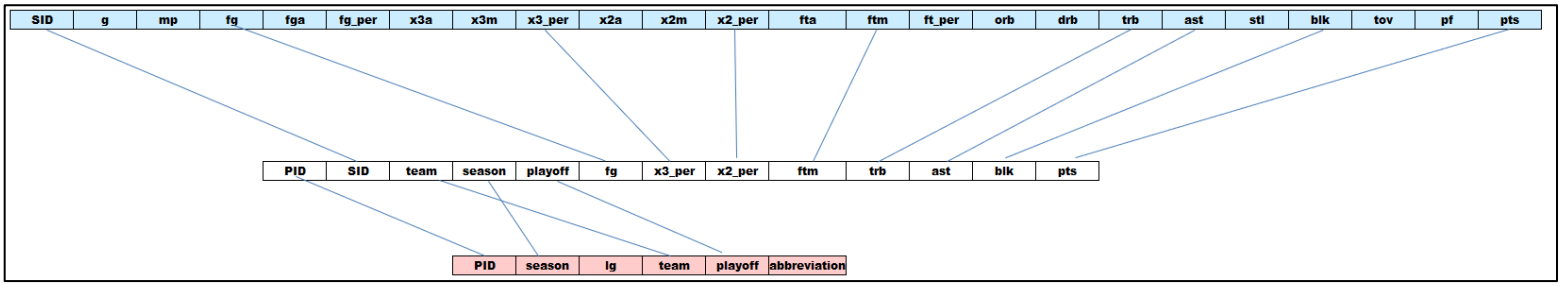
מאגרי הנתונים

- קובץ 1 – 'teamstats' - מכיל נתונים על סטטיסטיקות מתקדמות עבור כל קבוצת כדורסל בליגת ה- NBA לכל עונה.
- קובץ 2 –'teamsplayoff' - מכיל נתונים על עלייה לשלב ה-Playoff עבור כל קבוצה לכל עונה.

חלק 1- STTM

ביצענו STTM בקובץ Excel לפי הנחיות ה-ETL.

להלן סכמה ויזואלית של ה-STTM לטובת ההמחשה:



לינק לקובץ Excel של ה-STTM והתרשים - STTM

חלק 2- data mining techniques

עבור ה-DW שבנינו לאחר ה-ETL

1. **תהליכי ה-KDD**
   - בלימוד מונחה: תהליך ה- KDD אותו אנו מבצעים הוא ניבוי אם קבוצה עלתה או לא עלתה לפלייאוף הNBA. על כן, נבחר בשיטת Predictive בעזרת רגרסיה לוגיסטית.
   - בלימוד לא מונחה: תהליך ה-KDD מסוג Descriptive, בשיטת Clustering בעזרת K-means.

2. **הטכניקות שנממש על אוסף הנתונים**

a. סוגי הנתונים- רוב הנתונים שלנו (8 עמודות) הן עמודות נומריות, 2 עמודות הם עמודות שמיות ועוד עמודת מטרה שהיא עמודה בינארית- עלה/לא עלה ל-Playoff.
על כן, נבחר בטכניקת רגרסיה לוגיסטית מאחר והיא מאפשרת זיהוי קשר בין משתנים- הקשר בין המשתנים המסבירים למשתנה המוסבר, נבצע רגרסיה לוגיסטית על בסיס העמודות הנומריות הרציפות שלנו על מנת לנבא האם הקבוצה עלתה/ לא עלתה למשחקי הפלייאוף.

b. תרחישי usecase עם דוגמה מספרית:
1. מנהל קבוצה מעוניין לבדוק האם יש קשר בין המשתנים למשתנה המוסבר כלומר האם עדיף לקבוצה להשקיע במדד אחוז השלשות ובמדד הריבאונד התקפה כדי למקסם תוצאות של מדד אחר הקשור אליו על מנת להשיג את תוצאות הטובות ביותר עבור הקבוצה ולהצליח להגיע לפלייאוף. בתהליך זה נרצה להבין האם הרגרסיה הלוגיסטית שנבצע על המשתנים תספק לנו הסבר על הקשרים בין המשתנים למשתנה המוסבר.
לדוגמא קבוצת בוסטון סלטיקס בעונת 2020 עמדה על אחוז שלשות של 37% ו-14 ריבאונדים התקפיים למשחק דבר שגרם לאסיסטים רבים (16 למשחק) ועזר לקבוצה להעפיל לפלייאוף.
2. אתר הימורים בארץ רוצה להבין האם קיים קשר בין הנתונים הסטטיסטיים של הקבוצות לבין ההצלחה שלהם להגיע למשחקי הפלייאוף. אתר ההימורים מעוניין לבצע ניתוח של הנתונים על מנת להציע יחסי הימורים שתואמים את החיזוי. אתר ההימורים פנה אלינו על מנת לדעת מה הם סיכויי העפלה למשחקי הפלייאוף על פי מדד הזריקות ל-3. השאלה היא האם ככל שאחוז הזריקות ל-3 עולה מעל 40%, היחס של הקבוצה צריך להיות נמוך יותר? כלומר, הסיכוי של הקבוצה לנצח, קטן.

3. **מדד דמיון עבור ה-DW**

בלימוד מונחה: מדד דמיון לרגרסיה לוגיסטית הוא פשוט ערך התוצאה (הסיגמויד) מכיוון שהנתונים שלנו ושיטת החיזוי מתבססים על עמודות מסוג נומרי רציף. פונקציית סיגמויד (בצורת s) כזו שמוציאה ערכים בין 0 ל-1 כך שבעזרתה רואים את הדמיון בין הנתונים.
בלימוד לא מונחה: המרחק האוקלידי אשר מביע את ההפרשים בין הנקודות וכך ניתן לדעת את הדמיון ביניהם.

4. **ההשערות לשאלות העסקיות**

	Supervised	Unsupervised
KPI	עמודת המטרה, עמודת ה-playoff	האם יש הבדל בין מאפייני המשחקים לפי שנים?
H0	הקבוצה לא עולה ל-Playoff	אחוז הקליעות לשלוש נקודות לא משפיע
H1	הקבוצה עולה ל-Playoff	אחוז הקליעות לשלוש נקודות כן משפיע
	כלל ההכרעה יהיה לפי מדד ה-KPI, עמודת ה-target, במות הפעמים שהעפילו.	כלל הכרעה האם קיים הבדל ברמת מובהקות 0.05

חלק 3- שאילתות

שאילתות SQL ו- Window Functions :

```
SELECT team, pts, playoff, SUM(playoff) OVER(PARTITION By team) as Total Playoff

FROM Fact

Where season > 2000


SELECT team, season, pts LEAD(pts,1) OVER (PARTITION BY team ORDER BY season) AS Next Season pts

FROM Fact

Where season >2000


SELECT team, , playoff, MAX(ast) OVER(PARTITION By team, year) as ASTlead

FROM Fact


SELECT Team, playoff,

DENSE_RANK() OVER (ORDER BY x3) AS Rankx3

FROM fact


SELECT team, playoff,

NTILE(3) OVER ( ORDER BY x3) AS RankX3

FROM Fact


SELECT team, playoff,

NTILE(3) OVER ( ORDER BY x3) AS RankX3

where "Season"=2020
```

חלק 4- ניהול גרסאות

קישור לפרויקט שלנו ב- Github