הצעת פרויקט בינה עסקית

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.

*** קישורים למקורות המידע וניתוח הנתונים הרלוונטיים:**

Colab-Data Exploration

Colab-Modeling

מקור מידע- אתרKaggle

שאלות המחקר

♦ **Supervised –** האם הקבוצה העפילה למשחקי הפלייאוף?

KPI- עמודת המטרה, עמודת ה-playoff				
T	R	A	M	S
היעד תחום בזמן לכל עונה ונקבע לאחר תום המשחקים הסדירים וצבירת הנקודות של כל קבוצה.	היעד רלוונטי מכיוון שהמוניטין של כל קבוצה ומועדון נקבע ע"י ההישגים של הקבוצה. עליה לפלייאוף היא אחת מהישגים של קבוצה ומועדון	היעד בר השגה מכיוון שכל קבוצה יכולה להגיעה למשחקי הפלייאוף בתום העונה הסדירה	היעד מדיד בסולם בינארי כאשר המדד מקבל 1 כהצלחה ו -0 ככישלון	שאלת המחקר היא ספציפית וברורה. השאלת מתמקדת בהעפלת הקבוצה למשחקי הפלייאוף

♦ **Unsupervised -** איזה מאפיינים משפיעים על דמיון בין משחקי כדורסל?

KPI – האם יש הבדל בין מאפייני המשחקים לפי שנים?				
T	R	A	M	S
היעד תחום בזמן מכיוון שיש לנו סט אחורה של 70 שנה בלבד.	היעד רלוונטי מכיוון שצבירת נקודות הינה גורם דומיננטי להצלחת קבוצות.	היעד בר השגה מאחר ויש עמודת נתונים של הנקודות שאותן שהקבוצות יכולות להשיג.	היעד מדיד מאחר ויש עמודת נתונים של הנקודות.	שאלת המחקר היא ספציפית וברורה, כי הוא מדד להצלחה במשחק.

*עמודת point כמדד להצלחה (לאחר clustering, נבדוק בכל קלאסטר האם הקבוצות בו בעלות מספר נק' דומה).

תהליך הבנת הנתונים

♦ **מדדי פיזור:**

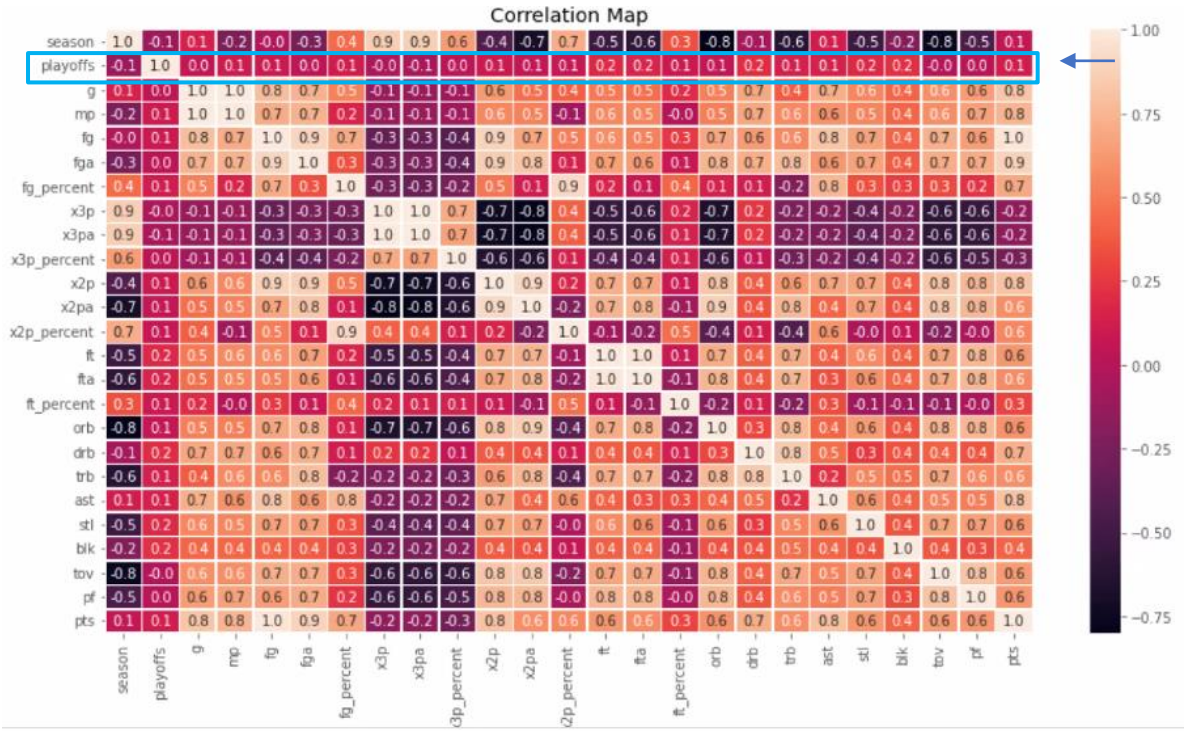
- **count:** number of entries
- **mean:** average of entries
- **std:** standard deviation
- **min:** minimum entry
- **25%:** first quantile
- **50%:** median or second quantile
- **75%:** third quantile
- **max:** maximum entry

	season	g	mp	fg	fga	fg_percent	x3p	x3pa	x3p_percent	x2p	...	ft_percent	orb	drb	trb	ast	stl	blk	tov	pf
count	1783.000000	1782.000000	1593.000000	1782.000000	1782.000000	1782.000000	1340.000000	1340.000000	1340.000000	1782.000000	...	1782.000000	1453.000000	1453.000000	1730.000000	1782.000000	1394.000000	1394.000000	1465.000000	1782.000000
mean	1991.182838	79.044332	19373.216573	3125.957912	6904.686308	0.451116	407.278358	1162.461194	0.330249	2819.699214	...	0.750409	1028.900895	2473.063317	3681.429480	1811.546577	652.376614	403.015782	1288.151536	1826.461279
std	19.965183	7.358682	1536.881375	512.229244	880.680686	0.037219	273.925581	735.314412	0.047610	629.944748	...	0.031794	225.661400	267.553038	630.445945	322.298289	110.381168	84.711705	229.125986	272.587284
min	1947.000000	11.000000	2640.000000	432.000000	1020.000000	0.246000	6.000000	24.000000	0.104000	426.000000	...	0.590000	202.000000	348.000000	550.000000	215.000000	72.000000	39.000000	196.000000	253.000000
25%	1975.000000	82.000000	19755.000000	2943.000000	6532.000000	0.441000	161.750000	501.500000	0.314000	2408.500000	...	0.733000	876.000000	2366.000000	3386.250000	1675.250000	582.000000	347.250000	1147.000000	1682.250000
50%	1993.000000	82.000000	19805.000000	3179.500000	6933.000000	0.457000	394.000000	1122.500000	0.344000	2672.500000	...	0.752000	1023.000000	2475.000000	3556.000000	1839.500000	649.000000	397.000000	1261.000000	1856.000000
75%	2008.000000	82.000000	19855.000000	3487.500000	7423.750000	0.473000	589.000000	1621.750000	0.362000	3424.000000	...	0.771000	1177.000000	2619.000000	3805.000000	2016.000000	720.000000	453.000000	1429.000000	2022.000000
max	2022.000000	84.000000	20460.000000	4059.000000	9295.000000	0.545000	1323.000000	3721.000000	0.428000	4018.000000	...	0.839000	1845.000000	3316.000000	6131.000000	2575.000000	1059.000000	716.000000	2011.000000	2470.000000

8 rows × 24 columns

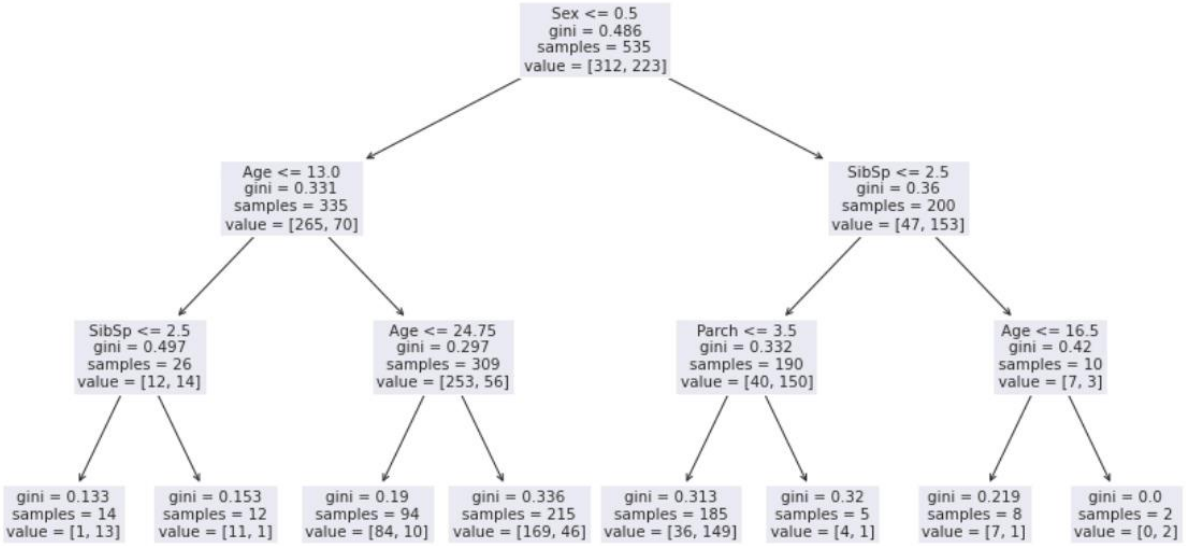
♦ **תלויות וקשרים (קורלציה):**

ניתן לראות בתרשים הבא כי כלל ה-features אדישים לעמודת המטרה.



- **עבור שאלת ה- supervised, מציאת האנטרופיה של כל אחת מהתכונות:**

נבצע חישוב אנטרופיה על כל העמודות שיש לנו למעט עמודת ה-Playoff, עמודת המטרה.



- **חישוב מדדי ה-gini-index וה-information-gain-** עבור 2 התכונות בעלות האנטרופיה הנמוכה ביותר ('fg_precent', 'ft_precent'):

```
[94] gini_index(data, "fg_percent")
```

```
0.9879648209240428
```

```
gini_index(data, "ft_percent")
```

```
0.9898101856575482
```

```
[96] !pip install info_gain
     from info_gain import info_gain
     info_gain.info_gain(data["ft_percent"], data["playoffs"])
```

```
Collecting info_gain
  Downloading info_gain-1.0.1-py3-none-any.whl (3.3 kB)
Installing collected packages: info-gain
Successfully installed info-gain-1.0.1
0.06293094447883973
```

```
!pip install info_gain
from info_gain import info_gain
info_gain.info_gain(data["fg_percent"], data["playoffs"])
```

```
Requirement already satisfied: info_gain in /usr/local/lib/python3.7/dist-packages (1.0.1)
0.10929974756964711
```

<u>מסקנות מתחקור הנתונים</u>

מאחר וכלל ה-features אדישים לעמודת המטרה, אין מולטי קולינריות, כלומר כל התכונות הן בלתי תלויות במטרה. על כן, הניבוי יהיה יותר טוב והמסקנות שנוציא מכאן יהיו אמינות.

בנוסף, העמודות- 'ft_precent' ,'fg_precent', הן בעלות ה-information-gain הנמוך ביותר, כלומר נותנות לנו הכי מעט אינפורמציה ביחס לעמודות האחרות ולכן נשקול האם להשאירם בתהליך ניקוי הנתונים.