דוח מדעי– קבוצה 16

הקדמה

משחק הכדורגל הוא ענף הספורט הקבוצתי הפופולרי והנפוץ ביותר בעולם.
הפופולריות הרבה בקרב משחק הכדורגל יצרה עניין רב במשחק ובעיקר בתוצאותיו. בד בבד עם התפתחות הכדורגל התפתח גם ענף ההימורים על תוצאות המשחקים בכדורגל. משחר ההיסטוריה אהבו בני אדם להמר על תוצאות של משחקי ספורט שונים דוגמת מרוצי סוסים באנגליה. עם התפתחות ההימורים סוגיות של מודל חיזוי כדורגל הפכו גם הם לפופולריות בשנים האחרונות וגישות רבות ומגוונות הוצעו במטרה להעריך את התכונות המובילות קבוצת כדורגל לנצח או להפסיד במשחק. ישנם שלושה סוגים של גישות שנחשבות לחיזוי תוצאות משחקי הכדורגל: גישות סטטיסטיות, למידת מכונת וגישה בייסניות. הפרויקט שלנו התמקד בחיזוי בעזרת אלגוריתם KNN שהינו גישת חיזוי של למידת מכונה, לחיזוי תוצאות משחק הכדורגל (ניצחון, הפסד או תיקו) ב-11 ליגות אירופאיות בשנים 2008-2016.

סקירת ספרות

אירועי ספורט זהו ענף שמאז ומתמיד עניין אוכלוסיות רבות. אחד מענפי הספורט הפופולריים ביותר הוא משחקי **כדורגל**. במשחק כדורגל משתתפות שתי קבוצות – קבוצת בית וקבוצת חוץ. מטרת כל אחת מן הקבוצות היא הכנסת כדור המשחק לתוך שערה של השנייה. כל קבוצה מונה אחד עשר שחקנים. נקודות המשחק מצטברות עבור כל קבוצה על ידי הבקעת הכדור במלוא היקפו לתחום השער של הקבוצה היריבה – אירוע זה נקרא הבקעת גול. עבור כל אחד מהשערים, ניצב שחקן שתפקידו הוא שוער הקבוצה והוא למעשה מגן על השער מפני הבקעת גולים של הקבוצה היריבה. הקבוצה המנצחת בתום המשחק היא הקבוצה שזכתה למרב הנקודות בתום הזמן החוקי של המשחק הנמשך לרוב כ90 דקות. ישנם 3 תוצאות משחק אפשריות : ניצחון קבוצת הבית , תיקו וניצחון קבוצת חוץ.

לאור הפופולריות הגדולה של הכדורגל יחד עם הנתון כי מספר תוצאות משחקי הכדורגל הוא נמוך - **חיזוי** תוצאות משחקי כדורגל הפך לאתגר מאוד מעניין. כיום יש קידמה לאיסוף נתונים בכל תחום ובפרט בענף הכדורגל כך שעל בסיס נתונים אלו פותחו טכניקות וכלים באמצעותם בני אדם יכולים לבצע ניתוח לנתונים שנאספו והפקת מידע רלוונטי. ישנן טכניקות רבות באמצעות ניתן לבצע כריית מידע, העיקריות שבהן –רשת נוירונים(ANN - Artificial Neural Networks), עצי החלטה(Decision Trees), רשתות בייסניות(Bayesian Networks), תמיכת מכונות ווקטוריות(SVM - Support Vector Machines), KNN-K nearest neighbor, Data Driven Bayesian, Naïve Bayesian .

יחד עם זאת, חיזוי התוצאות אינו דבר קל משום שהוא תלוי בגורמים רבים שיכולים להשפיע על תוצאות משחקי הכדורגל, כמו עבודת צוות, מיומנות, מזג-האוויר, יתרון ביתי ועוד גורמים רבים ואחרים. בעקבות הגורמים השונים והמרובים המשפיעים על תוצאות משחק הכדורגל, הקריטריונים בהם בחרנו להתמקד בעת חיפוש המאמרים הינם – טכניקות לכריית מידע בעלות תוצאות חיזוי של מעל 50% . קריטריון נוסף, מאמרים המציגים ניסויים בעלי גורמים שונים ומציגים את השפעתם(ניתן לראות בטבלה 1 את סיכום הקריטריונים הללו).

בחרנו להתמקד ב4 מאמרים עיקריים([1],[2],[3],[4]) בהם הושגו תוצאות החיזוי הטובות ביותר בניסויים שערכו(כפי שמוצג בטבלה 1).
במאמר [1] ו-[2] **הבעיה** שתוארה הינה הקושי לחזות את תוצאות הכדורגל המדויקות בשל השפעה של כל כך הרבה גורמים כמו מורל הצוות, כישורים, יתרון בייתי ורבים אחרים. **הפתרון** המוצע במאמרים אלו הינו שימוש בשיטת רשתות בייסניות שהוכיחה את עצמה כבר בעבר כבעלת כוח חיזוי למזג האוויר וספורט.
במאמר [1] בדקו את הפתרון המוצע מעלה ע"י **ניסויי** שערכו - במהלך ניסויי זה בוחנים מספר גורמים שונים שהם הגורמים העיקריים המשפיעים על חיזוי תוצאות המשחקים. הגורמים העיקריים שנבחנו מחולקים לשני סוגים – גורמים פסיכולוגיים וגורמים לא פסיכולוגים. למשל הגורמים הפסיכולוגיים כוללים בתוכם את מזג האוויר, את תוצאות חמשת המשחקים האחרונים וכו' והגורמים הלא פסיכולוגיים כוללים בתוכם למשל את כמות פציעות השחקן הראשי, הגיל הממוצע של השחקנים וכו'. הניסויים מבוצעים באמצעות תוכנת NETICA המספקת הטמעת אלגוריתמי כריית נתונים. הניסוי התבצע כך – כל משחק של ברצלונה בליגה הספרדית חושב בנפרד מכיוון שלכל משחק כפי שראינו יש גורמים משפיעים שונים בהתאם למשחק .לכן הניסוי חזר 38 פעמים עבור כל המשחקים שהתבצעו בעונה

2008-2009. **תוצאות** הניסויי מראות כי באמצעות רשתות בייסניות ניתן להגיע לתוצאות חיזוי מאוד גבוהות – בהשוואה לתוצאות האמת של עונה 2008-2009 ביחס לתוצאות החיזוי יש התאמה של 92% .
ניתן לראות כי הקריטריונים לבחירת המאמר(אחוזי חיזויי) מתבטאים פה באחוז הצלחה מאוד גבוה.

במאמר [2] בדקו את הפתרון המוצע מעלה ע"י **ניסויי** שערכו – במהלך ניסויי זה בוחנים מספר גורמים שונים שהם הגורמים העיקריים המשפיעים על חיזוי תוצאות המשחקים. הניסויים מבוצעים באמצעות תוכנת WEKA המספקת הטמעת אלגוריתמי כריית נתונים. הניסוי התבצע כך – כל משחק חושב בנפרד מכיוון שלכל משחק כפי שראינו יש גורמים משפיעים שונים בהתאם למשחק . לכן הניסוי חזר 380 פעמים עבור כל המשחקים השונים בשלושת העונות בנפרד. המחקר התבסס על ליגת העל האנגלית בעונות 2010-2011,2011-2012,2012-2013 . הליגה כוללת בתוכנה 20 קבוצות וכל קבוצה משחקת פעמיים – פעם אחת במשחק בית ופעם אחת במשחק חוץ. **תוצאות** הניסויי מראות כי באמצעות רשתות בייסניות ניתן להגיע לתוצאות חיזוי מאוד גבוהות – בניסוי זה ממוצע תוצאות החיזוי עבור 3 העונות הינו 75.09% . ניתן לראות כי הקריטריונים לבחירת המאמר(אחוזי חיזויי) מתבטאים פה באחוז הצלחה מאוד גבוה.

ממאמרים [1] ו-[2] אנו מסיקים כי רשתות בייסניות היא שיטה טובה לחיזוי תוצאות משחקי כדורגל, אך רצינו לאמת זאת באמצעות ניסוי נוסף שהוצג במאמר [3] ובוחן מספר מודלי חיזוי. **הבעיה** שתוארה במאמר [3] היא שהמודל כלל בתוכו מידע שהיה רלוונטי עבור שתי עונות בלבד, היות והמידע הכיל בתוכו ערכים על שחקני מפתח ספציפיים שלאחר שתי העונות כבר לא היו חלק ממועדון הקבוצה. בשל חוסר הרלוונטיות של ערכים אלו היה חשש שתוצאות החיזוי יהיו נמוכות ביחס למודלים אחרים. לכן **הפתרון** שהוצע במאמר הינו לבצע **ניסוי** הבוחן מודלי חיזוי נוספים ביחס למודל הBN .
מודלי החיזוי הנוספים אותם הניסוי בחן הם – MC4 Decision trees, Naïve Bayesian , Data Driven Bayesian , KNN-K nearest neighbor . הניסוי בחן סט נתונים מסוים המיוחס למשחקים ששחקן שכינויו הינו 'Spurs' וכך הכריע האם BN היא שיטה מדויקת וטובה. **תוצאות הניסוי** מראות כי כאשר משתמשים באותם סטי נתונים עבור עונות שלמות שיטת הBN הוכחה כמדויקת וטובה יותר ביחס למודלי החיזוי האחרים שהוצעו(כפי שמתואר בטבלה הבאה המופיעה במאמר [3]).

Train period–Test period	Number of correct predictions by learner					
	Most common	MC4	Naive BN	Hugin BN	Expert BN	KNN
Overall average percentage	40.05%	41.72%	47.86%	39.69%	59.21%	50.58%

המסקנות העולות מהמאמר הן שלמרות שהמודל היה לא רלוונטי ביחס לנתונים של שחקני המפתח שהשתנו במהלך שתי עונות, תוצאות הניסוי מאשרות את הפוטנציאל המצוין של מודל זה. למודל זה יש יכולות לחזות באופן גבוה מבלי להשתמש בסט נתונים כל כך גדול. יתרה מכך, שיטת חיזוי BN היא פשוטה יותר באופן יחסי לשאר מודלי החיזוי וניתן לבצע שימוש חוזר במבנה שלה עבור בעיות חיזוי אחרות. ניתן לראות כי הקריטריונים לבחירת המאמר(אחוזי חיזויי) מתבטאים פה באחוז הצלחה גבוה(59.21%)ביחס לרף שהצבנו בקריטריונים של בחירת המאמר.

במאמר [4] **הבעיה** שתוארה במאמר זה, היא מהו המודל חיזוי הטוב ביותר, באמצעותו ניתן יהיה לחזות את תוצאות המשחקים. **הפתרון** המוצע במאמר הוא ביצוע **ניסוי** הבוחן מספר מודלי חיזוי, בינהם : decision tree, SVN,KNN , ו- LR – Linear Regression . סט הנתונים הנבחן בניסוי מוצג בטבלה 1. תוצאות הניסוי מראות כי בעזרת מודל הKNN הגיעו לחיזוי תוצאות נכונות של 7 משחקים מתוך 9 שלמעשה זה 77.18% אחוזי דיוק.
ניתן לראות כי הקריטריונים לבחירת המאמר(אחוזי חיזויי) מתבטאים פה באחוז הצלחה גבוה(77.18%)ביחס לרף שהצבנו בקריטריונים של בחירת המאמר.

ניתן לראות כי במאמר [5] שיטת החיזוי שהוצעה היא באמצעות מודל – "Hidden Markov Process " ותוצאות החיזוי שהוצגו במאמר הינן בעלות דיוק של 55.64% . במאמר [6] שיטת החיזוי שהוצעה היא Naive Bayes ותוצאות הניסוי שם הן 54.702%. במאמרים [7] ו- [8] שיטת החיזוי שהוצעה היא באמצעות SVM – Support Vector Machines. תוצאות החיזוי בניסויים שהוצגו שם הינן 50.8% ו- 53.3% בהתאמה. בחרנו לא להתמקד במאמרים אלו משום שביחס למאמרים האחרים שסקרנו, הבחנו כי תוצאות החיזוי היו גבוהות מעל 59% ולכן הם יותר עומדים בקריטריונים של בחירת המאמרים שהצבנו.

סיכום הסקירה:
[טבלה 1]

מאמר	שיטת חיזויי	סט הנתונים הנבדק		תוצאות חיזוי
[1]	Bayesian Networks	גורמים פסיכולוגיים	גורמים לא פסיכולוגיים	92%
		weather History_of_5last_games Result_against_for_teams Home_game ability_front_team Psychological_state	Average_of_players_age Injuried_main_players ave_match_in_week Performane_of_main_players performance_of_all_players ave_goal_in_all_home ave_goal_for_Home	
[2]	Bayesian Networks	Home Team Away Team Home Team Shots Away Team Shots Home Team Shots on Target Away Team Shots on Target Home Team Corners Away Team Corners Home Team Fouls Committed Away Team Fouls Committed Home Team Yellow Cards Away Team Yellow Cards Home Team Red Cards Away Team Red Cards Half Time Home Team Goals Half Time Away Team Goals Full Time Home Team Goals Full Time Away Team Goals		75.09%
[3]	Bayesian Networks	Attack	מייצג את איכות המתקפה של השחקן 'Spurs' (low, medium , high).	59.21%
		Spurs_quality	היכולת הכללית של הקבוצה בה 'Spurs' משחק(low, medium , high).	
		Performance	כמה טובים היו ביצועי הקבוצה בהינתן ליכולות שלהם ושל הקבוצה היריבה(low, medium , high).	
[4]	KNN	Team id Avg goals scored per match this season Avg goals conceded per match this season Result of previous match Result of two matches ago Result of three matches ago Result of four matches ago Result of five matches ago Team was in a lower league previous year Number of matches coached by current coach Team hired new coach during previous month Top-scorer suspended or injured Top-assist suspended or injured Avg goals scored by top-scorer		77.18%

		Avg assists given by top-assist Days since previous match Percentage of wins this season Percentage of lose this season Percentage of draw this season	

[יש לציין כי המאמרים ברובם מבצעים השוואה בין מספר שיטות חיזוי אך שיטות החיזוי המצוינות בטבלה הנ"ל הן השיטות שהניבו את תוצאות החיזוי הגבוהות ביותר בכל מאמר ומאמר].

תיאור הנתונים:

סט הנתונים עליו בחרנו לבצע את הניסוי הוא הוא סט של 11 ליגות אירופאיות בשנים 2008-2016 .
מתוך סט הנתונים, במהלך עיבוד הנתונים בחרנו את התכונות הבאות על מנת לאמן את המודל:
<u>הערה:</u> הנתונים המוצגים בטבלה הם כלל הנתונים אותם עיבדנו בשלב ה- pre processing .
הנתונים בצהוב הם הנתונים אותם בחרנו בסופו של דבר למודל שלנו לאחר הרצת מודלים רבים ובחירת התכונות שנתנו לנו את מודל החיזוי המדויק ביותר.

שם נתון	תיאור	פירוט חישוב
match_api_id	מזהה משחק	
home_team_api_id	מזהה קבוצת בית במשחק	
away_team_api_id	מזהה קבוצת חוץ במשחק	
5Last_Gamesaway_team_api_id	ממוצע חמשת תוצאות של משחקים אחרונים עבור קבוצת הבית	חולקו נקודות לכל תוצאה: ניצחון -3 הפסד- 0 תיקו- 1 חושב סכום התוצאות עבור 5 משחקים אחרונים של קבוצת הבית וחולק ב15
5Last_Gameshome_team_api_id	ממוצע חמשת תוצאות של משחקים אחרונים עבור קבוצת החוץ.	חולקו נקודות לכל תוצאה: ניצחון -3 הפסד- 0 תיקו- 1 חושב סכום התוצאות עבור 5 משחקים אחרונים של קבוצת החוץ וחולק ב15
five_last_meetings_for away_team_api_id	ממוצע חמשת תוצאות של משחקים אחרונים **בין הקבוצות במשחק** עבור קבוצת החוץ.	חולקו נקודות לכל תוצאה: ניצחון -3 הפסד- 0 תיקו- 1 חושב סכום התוצאות עבור 5 משחקים אחרונים של קבוצת החוץ וחולק ב15
five_last_meetings_for_home_team_api_id	ממוצע חמשת תוצאות של משחקים אחרונים **בין הקבוצות במשחק** עבור קבוצת הבית	חולקו נקודות לכל תוצאה: ניצחון -3 הפסד- 0 תיקו- 1 חושב סכום התוצאות עבור 5 משחקים אחרונים של קבוצת הבית וחולק ב15
avg_performance_of_main_home_players	ממוצע הרייטינג של ה5 הטובים בקבוצת הבית	חושב ממוצע הרייטינג של 5 השחקנים בעלי הרייטינג הכי גבוה בקבוצת הבית
avg_performance_of_all_home_players	ממוצע הרייטינג של כל השחקנים בקבוצת הבית	חושב ממוצע הרייטינג של כל השחקנים בקבוצת הבית
avg_performance_of_main_away_players	ממוצע הרייטינג של ה5 הטובים בקבוצת החוץ	חושב ממוצע הרייטינג של 5 השחקנים בעלי הרייטינג הכי גבוה בקבוצת החוץ
avg_performance_of_all_away_players	ממוצע הרייטינג של כל השחקנים בקבוצת החוץ	חושב ממוצע הרייטינג של כל השחקנים בקבוצת החוץ

home_team_avg_game_week	ממוצע משחקים בשבוע לקבוצת הבית	חישוב ממוצע משחקים לשבוע של 10 משחקים אחרונים על לתאריך המשחק הנוכחי עבור קבוצת בית
away_team_avg_game_week	ממוצע משחקים בשבוע לקבוצת החוץ	חישוב ממוצע משחקים לשבוע של 10 משחקים אחרונים על לתאריך המשחק הנוכחי עבור קבוצת חוץ
home_team_avg_age	ממוצע הגילאים של שחקני קבוצת הבית	חישוב ממוצע גילאי השחקנים שהשתתפו ב5 משחקים אחרונים עד לתאריך המשחק לקבוצת הבית
away_team_avg_age	ממוצע הגילאים של שחקני קבוצת החוץ	חישוב ממוצע גילאי השחקנים שהשתתפו ב5 משחקים אחרונים עד לתאריך המשחק לקבוצת החוץ
ave_goal_for_home_team	ממוצע גולים של קבוצת הבית	חישוב ממוצע גולים ל10 המשחקים אחרונים עד לתאריך המשחק הנוכחי לקבוצת הבית.
ave_goal_for_away_team	ממוצע גולים של קבוצת החוץ	חישוב ממוצע גולים ל10 המשחקים אחרונים עד לתאריך המשחק הנוכחי לקבוצת החוץ.
class	משתנה המטרה- תוצאות המשחק: ניצחון,הפסד,תיקו	

כל השדות נורמלו לפי שיטת standardization :

$$\frac{value - mean}{std}$$

Value- ערך המקור בשדה
Mean- ממוצע הערכים לתכונה
Std- סטיית התקן לתכונה

תיאור האלגוריתם:

לאור הממצאים שעלו בסקירת הספרות, לאחר שסקרנו מספר אלגוריתמים שונים, שהמובילים ביניהם היו רשתות ביסנייות וKNN, לבסוף בחרנו לעבוד עם שיטת - k-Nearest Neighbors algorithm KNN. אלגוריתם זה מבוסס מופעים בו הפונקציה מקורבת באופן מקומי בלבד וכל החישובים נדחים עד סיווגה. אלגוריתם זה הוא מבין האלגוריתמים הפשוטים בתחום למידת מכונה.

האלגוריתם עובד בצורה הבאה: עבור וקטור תכונות חדש שנכנס לשם חיזוי, האלגוריתם מסווג אותו לפי K השכנים הקרובים שלו. סיווג הווקטור יעשה לפי רוב מבין K השכנים של הווקטור.

הבחירה הטובה ביותר של K תלויה בנתונים, בדרך כלל ערכים גבוהים של K יכולים לצמצם השפעה של רעשים אך יגרמו לגבולות בין התכונות להיות פחות מובהקים. הדיוק של האלגוריתם יכול להיפגע ע"י נוכחות של תכונות לא רלוונטיות וגורמי רעש. מאמצי מחקר רבים הושקעו עבור בחירת תכונות שיתרמו לסיווג.

הרצת ניסוי/הערכה:

תחילה בחרנו את התכונות איתן נרצה לאמן את המודל שלנו בעזרת המאמרים אותם סקרנו בסקירת הספרות ובעזרתן בנינו Data frame בחלוקה למשחקים (פירוט החישובים מופיע בטבלה שמוצגת תחת תיאור הנתונים). על העמודות בוצעו נרמול נתונים בשיטת standardization . בשלב השני חילקנו את מאגר הנתונים שלנו לסט אימון וסט מבחן כפי שנתבקשו כאשר סט המבחן נבחר עבור השנים 2015-2016 והשנים 2008-2014 נבחר להיות סט האימון עבור המודל.

לאחר עיבוד והכנת התכונות איתן אנו רוצים לחזות את המודל. הורצו מספר ניסויים במטרה למקסם את אחוז הדיוק במודל. את הניסויים הבאים הרצנו על 5 המודלים הבאים:

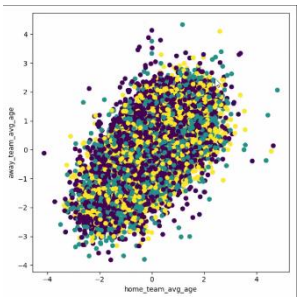
1. KNN (k-Nearest Neighbors)
2. RFC (Random Forest Classifier)
3. NB (Gaussian NB)
4. LR (Logistic Regression)
5. Ada Boost Classifier

תחילה הרצנו את שלושת המודלים על כל התכונות המוצגים בטבלה הנ"ל. תוצאות אלו לא הניבו לנו תוצאות מדויקות (כ-40 אחוז דיוק עבור המודלים הנתונים) אך נראה כי המודל KNN דייק יותר ביחס לשאר המודלים שנבחנו ולכן בחרנו לבצע פעולות נוספות על מנת לשפר אותו.

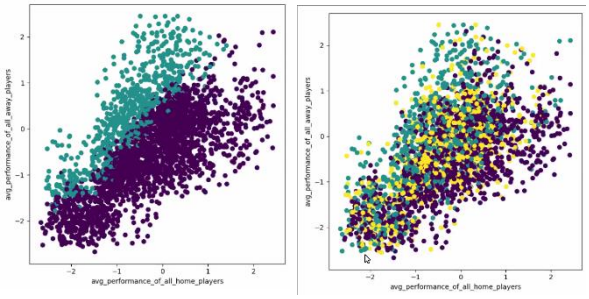
בניסיון לשפר את התוצאות הראשוניות בוצעו הפעולות הבאות:

1. חלוקת הפיצ'רים הבאים ל- binning בגדלים שונים:
   - home_team_avg_age
   - away_team_avg_age
   - avg_performance_of_all_home_players
   - avg_performance_of_main_home_players
   - avg_performance_of_all_away_players

2. ניסוי וטעיה למציאת מקסימום גלובלי למספר השכנים באלגוריתם KNN – כלומר משחק עם כמות השכנים באלגוריתם על מנת למצוא את הכמות הטובה יותר.
3. הוספה והורדה של תכונות לאלגוריתם על מנת לזהות ראשית את התכונות ה"חזקות" ביותר המשפיעות על דיוק החיזוי ובנוסף לזהות תכונות אשר גורעות מהדיוק ולהוריד אותן .
   על מנת להחליט אילו תכונות משפיעות יצרנו גרפים של התפלגות הערכים (סגול- ניצחון, טורקיז-הפסד, צהוב- תיקו) בתכונות כגון:



בגרף זה רואים את התפלגות הערכים של הגיל הממוצע של השחקים בקבוצה הבית והחוץ (away_team_avg_age, home_team_avg_age) . ניתן לראות שתכונה זו אינה מוסיפה ערך מוסף לסיווג תוצאות המשחק מאחר וכל תוצאות המשחק (הנקודות) מקובצות באותו אזור ואינן מתפזרות, לכן למודל קשה לסווג לפי תכונה זאת והוחלט להוציא אותה.

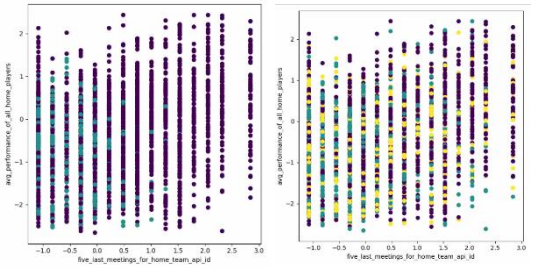


הגרפים מציגים את תוצאות המשחק ביחס לתכונות של ביצועי השחקים בקבוצת הבית והחוץ (avg_performance_of_all_home_players, avg_performance_of_all_away_players). הגרף הימני מייצג את התוצאות האמיתיות של המשחק והגרף השמאלי מייצג את התוצאות של מודל החיזוי. ניתן לראות כי לפי תוצאות מודל החיזוי ככול שלקבוצת הבית ביצועים טובים יותר בניגוד לקבוצת החוץ, המודל חוזה את תוצאת המשחק כניצחון לקבוצת הבית והפסד לקבוצת החוץ.

ניתן לראות בגרף הימני של תוצאות האמת כי:

- ככל שביצועי השחקנים של קבוצות הבית גבוהים יותר ובנוסף ביצועי השחקנים של קבוצות החוץ נמוכים ביחס לקבוצות הבית, יש סיכוי גבוה יותר לנצחון של קבוצת הבית ולהפך.
- ניתן לראות שכאשר ביצועי השחקנים של שני הקבוצות נמוך יותר קשה לחזות את תוצאת המשחק.
- ניתן לראות לפי מודל החיזוי שיש יתרון לקבוצת הבית (ישנם יותר ניצחונות לקבוצות הבית) ובנוסף ניתן לראות גם שכשביצועי השחקנים נמוכים של שני הקבוצות יש יותר סיכוי לקבוצת הבית לנצח.
- ניתן לראות כי תוצאות התיקו מפוזרים על פני כל הגרף ולכן ניתן להסיק שביצועי השחקנים אינם משפעים על באופן מובהק על הסיכוי לתוצאת התיקו.

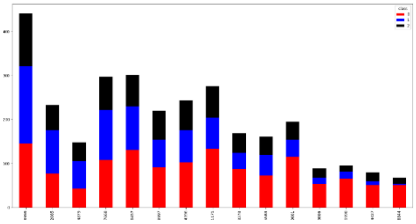
ניתן לראות כי ישנה צפיפות של נקודות סגולות בגרפים הבאים כאשר ביצועי קבוצת הבית גבוהים יותר וצפיפות של נקודות טורקיז עבור קבוצת החוץ.



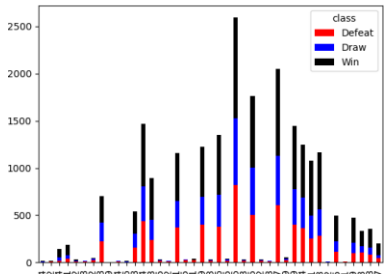
הגרפים מציגים את תוצאות המשחק ביחס לתכונות של ביצועי השחקנים בקבוצת הבית, אל מול מספר הנקודות שצברה קבוצת הבית בחמשת המשחקים האחרונים מול קבוצת החוץ במשחק. (avg_performance_of_all_home_players, five_last_meetings_for_home_team_api_id).

הגרף הימני מייצג את התוצאות האמיתיות של המשחק והגרף השמאלי מייצג את התוצאות של מודל החיזוי. ניתן לראות שתכונות אלה תורמות לחיזוי. לפי הגרף הימני (נתוני אמת) ככל שקבוצת הבית השיגה יותר נקודות בחמשת המפגשים האחרונים מספר הנקודות הסגולות (ניצחון) רב יותר. כמו כן, ניתן לראות שככל שביצועי השחקנים גבוהים יותר, מספר הנקודות הסגולות גדל. וכן, מודל החיזוי (גרף שמאל) מציג מגמה דומה לנתוני האמת.

הגרפים הבאים מציגים גם הם הסבר לבחירת התכונות שנכנסו לאימון המודל (אדום – ניצחון לקבוצת הבית, כחול- הפסד לקבוצת הבית, שחור – תיקו):



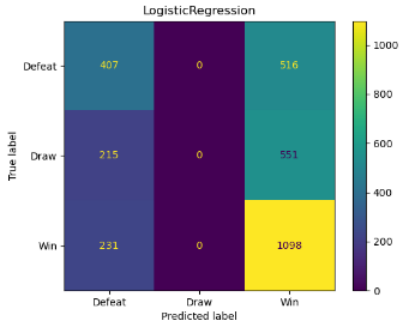
בגרף הבא רואים כמות הנקודות שהשיגה קבוצת הבית בחמשת המפגשים האחרונים (five_last_meetings_for_home_team_api_id), כאשר בצד הימני נמצאות הקבוצות שהשיגו הכי הרבה נקודות. ניתן לראות שכאשר קבוצת הבית השיגה מספר נמוך של נקודות יש הסתברות של כשליש לניצחון\הפסד\תיקו וכאשר קבוצת הבית השיגה הרבה נקודות במפגשים האחרונים יש לה סיכוי גבוה לנצח גם את המפגש הנוכחי.

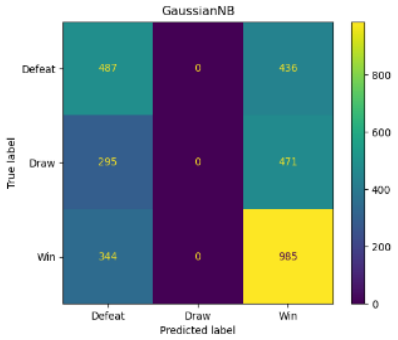


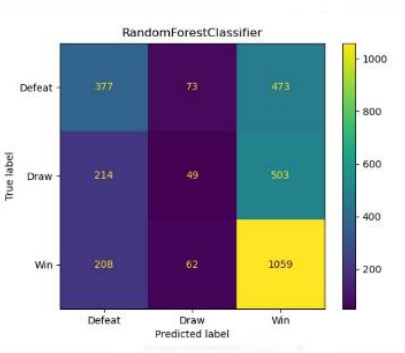
בגרף הבא אנו רואים את כמות הגולים הממוצעת של קבוצת הבית בחמשת המשחקים האחרונים (ave_goal_for_home_team). ניתן לראות שעבור כל כמות ממוצעת של שערים תמיד יש עדיפות גבוהה לקבוצת הבית. לכן, משתנה זה אינו עוזר לנו בחיזוי ובחרנו שלא להשתמש בו.

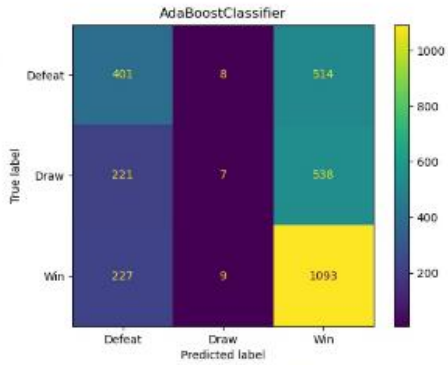
תוצאות ומסקנות:

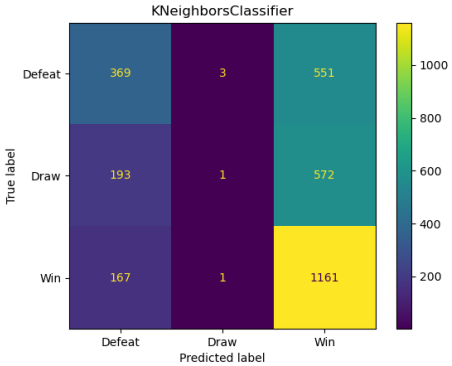
במהלך מודל החיזוי הרצנו כמה מודלי חיזוי: KNN, LR, AdaBoostClassifier, RFC, NB .
להלן תוצאות החיזוי שהתקבלו עבור מודלי החיזוי השונים:

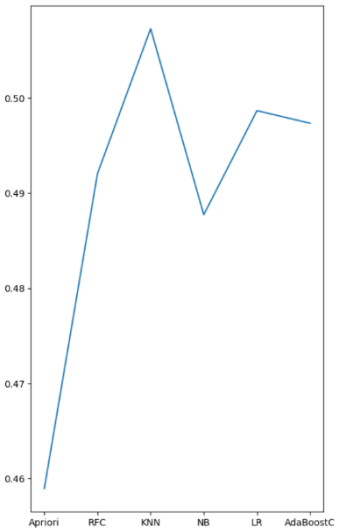












כפי שעלה מסקירת הספרות ומהרצת ניסויי החיזוי, המודל KNN חזה את תוצאות המשחק בצורה הטובה ביותר.

מתוצאות הניסוי עלה כי התכונות :

1. 'five_last_meetings_for_away_team_api_id'
2. 'five_last_meetings_for_home_team_api_id'
3. 'avg_performance_of_main_home_players'
4. 'avg_performance_of_all_home_players'
5. 'avg_performance_of_main_away_players'
6. 'avg_performance_of_all_away_players'

הן התכונות אשר השפיעו בצורה החזקה ביותר על תוצאות החיזוי. ולכן, הוחלט לאמן את המודל בעזרת תכונות אלו. יתר התכונות שהוצגו בטבלת תיאור הנתונים (אך לא הוכנסו לאימון המודל) היוו "רעשי רקע" (הוצג בהרצת ניסוי/הערכה) והורידו את אחוזיי הדיוק.

בנוסף, מצאנו כי חלוקה ל- binning שביצענו בניסוי, אינה תרמה למודל ולכן הוחלט שלא להשתמש בה. כמו כן, התגלה כי מספר השכנים הטוב ביותר עבור מודל הKNN בהתחשב בתכונות שנבחרו, הוא 330.

```
class
Defeat    6166
Draw      5398
Win       9810
Name: match_api_id, dtype: int64
0.4589473684210526
---------------------RandomForestClassifier---------------------------
0.49204771371769385
              precision    recall  f1-score   support

      Defeat       0.47      0.41      0.44       923
        Draw       0.27      0.06      0.10       766
         Win       0.52      0.80      0.63      1329

    accuracy                           0.49      3018
   macro avg       0.42      0.42      0.39      3018
weighted avg       0.44      0.49      0.44      3018

---------------------KNeighborsClassifier---------------------------
0.5072895957587806
              precision    recall  f1-score   support

      Defeat       0.40      0.51      0.45       729
        Draw       0.00      0.20      0.00         5
         Win       0.87      0.51      0.64      2284

    accuracy                           0.51      3018
   macro avg       0.42      0.40      0.36      3018
weighted avg       0.76      0.51      0.59      3018
```

```
---------------------GaussianNB---------------------------
0.487740225314778
              precision    recall  f1-score   support

      Defeat       0.43      0.53      0.48       923
        Draw       0.00      0.00      0.00       766
         Win       0.52      0.74      0.61      1329

    accuracy                           0.49      3018
   macro avg       0.32      0.42      0.36      3018
weighted avg       0.36      0.49      0.41      3018

---------------------LogisticRegression---------------------------
0.49867461895294896
              precision    recall  f1-score   support

      Defeat       0.48      0.44      0.46       923
        Draw       0.00      0.00      0.00       766
         Win       0.51      0.83      0.63      1329

    accuracy                           0.50      3018
   macro avg       0.33      0.42      0.36      3018
weighted avg       0.37      0.50      0.42      3018

---------------------AdaBoostClassifier---------------------------
0.49734923790589797
              precision    recall  f1-score   support

      Defeat       0.47      0.43      0.45       923
        Draw       0.29      0.01      0.02       766
         Win       0.51      0.82      0.63      1329

    accuracy                           0.50      3018
   macro avg       0.42      0.42      0.37      3018
weighted avg       0.44      0.50      0.42      3018
```

להלן סיכום תוצאות המודלים:

מודל	אחוז חיזוי
RFC	49.2%
KNN	50.7%
NB	48.7%
LR	49.8%
AdaBoostClassifier	49.7%

REFERENCES

[1] Farzin Owramipur, Parinaz Eskandarian, and Faezeh Sadat Mozneb , "Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team"

[2] Nazim Razali1, Aida Mustapha1, Faiz Ahmad Yatim2 and Ruhaya Ab Aziz, "Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)"

[3] A. Joseph, N.E. Fenton, M. Neil, "Predicting football results using Bayesian nets and other machine learning techniques "

[4] Adão Baptista Pereira Lopes, "Application of Machine Learning Algorithms for Automatic Classification of Problems Football*"

[5] Zheyuan Fan, Yuming Kuang, Xiaolin Lin, "Chess Game Result Prediction System"

[6] Niek Tax and Yme Joustra, "Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach"

[7] João Gomes, Filipe Portela, Manuel Filipe Santos, " Decision Support System for predicting Football Game result"

[8] Chinwe Peace Igiri, "Support Vector Machine–Based Prediction System for a Football Match Result "