סקירת ספרות

משחק הכדורגל הוא ענף הספורט הקבוצתי הפופולרי ביותר בעולם. המשחק מושך מיליוני אוהדים וצופים מכל רחבי הגלובוס, וכיום משחק הכדורגל אינו ענף ספורט בלבד אלא מאורע המהווה תוכן בידורי, סממן תרבותי וחברתי ואף אפיק השקעה המניע תעשייה ענפה. בעקבות ההתעניינות הרבה בענף הכדורגל, נאספו במרוצת השנים כמויות מידע עצומות אודות קבוצות ומשחקי כדורגל, המפורסמות לרוב בצורה פומבית לקהל הרחב. כחלק ממגמה אשר הולכת וגוברת בשנים האחרונות, ישנם ניסיונות רבים מטעם בעלי עניין שונים, להפיק ידע ותובנות אסטרטגיות ממידע זה לצורך ניתוח ביצועים, קבלת החלטות, הערכת סיכונים וכן חיזוי תוצאות משחקים.

מערכות חיזוי, הן מודלים לקביעת הערכה של משתנה כלשהו בנקודת זמן עתידית. הערכה זו מבוצעת לרוב על ידי שיטות שונות לניתוח סטטיסטי של המידע הקיים, תוך התייחסות למשתנים נוספים המשפיעים על ערכי המשתנה החזוי. תהליך בניית מודל חיזוי, מורכב מפעולות איסוף ועיבוד המידע הגולמי, ניתוח ועיבוד הנתונים וההערכת תוצאות החיזוי המתקבלות. על פי [1] Yezus et.al ניתן לחלק תהליך זה בתחום חיזוי משחקי ספורט ל-5 שלבים מרכזיים : בחירת מאגר משחקים אותו ננתח, קביעת משתנים מסבירים למודל, חילוץ ועיבוד המידע, ניסוי מגוון אלגוריתמי בינה מלאכותית ושיפור האלגוריתם הממומש.

בשל פופולריות המשחק, ארגונים רבים השקיעו מאמצים על מנת לחקור ולשפר את היכולת לחיזוי תוצאות המשחקים. חיזוי תוצאות משחקים, מביא תועלת רבה למועדוני הכדורגל, ומסייע לדרג המקצועי לקבל החלטות מושכלות לצורך שיפור הביצועים של הקבוצה [2].

השימוש בשיטות של כריית מידע ובינה מלאכותית, הביא בשנים האחרונות לשיפור משמעותי ברמת הדיוק עבור החיזוי בתחומים רבים, ולכן יושמו גם בתחום חיזוי המשחקים בענף הכדורגל. חוקרים בתחום פרסמו מודלים רבים המשתמשים בשיטות שונות ביניהם- רשתות נוירונים מלאכותיות, עצי החלטה, מודלים בייסיאנים, רגרסיה לוגיסטית, SVM ועוד.

בסקירה זו נבחן שימוש באלגוריתמים שונים לחיזוי תוצאות משחקי כדורגל, בניסיון לקבוע איזו שיטה מביאה לדיוק הגבוהה ביותר. המאפיינים על פיהם ניתחנו את האלגוריתמים השונים מתייחסים לתהליך עיבוד המידע, מורכבות המודל ומגבלותיו, ערכי הסיווג הנתמכים, דיוק תוצאות המודל ואימותו.

רשת בייסיאנית

במאמרו של [3] Razali, N , הוצעה גישה לחיזוי משחקי כדורגל על ידי שימוש ברשת בייסיאנית.
רשת בייסיאנית היא מודל הסתברותי גרפי, המייצג את פונקציית ההסתברות המשותפת של סט משתנים אקראיים, ומציג קשרים סיבתיים כגרף מכוון חסר מעגלים. ברשת זו, הקודקודים מייצגים משתנים (בדידים או רציפים) כאשר קיומה של קשת בין שני קודקודים, מייצג תלות בין שני המשתנים. מבנה הרשת מביא לידי ביטוי את חוזק הקשר בין המשתנים, ומאפשר עדכון של קשרים אלה במידה ומוזן לו מידע נוסף.
במודל אותו הציעו החוקרים, נותחו 18 משתנים שונים מנתונים של 380 משחקי הליגה האנגלית לכדורגל, במשך 3 עונות בין השנים 2010-2013. בתוצאות המחקר הציגו רמת דיוק ממוצעת (בין שלושת העונות אשר נבחנו) העומדת על 75.09% עבור חיזוי תוצאות משחקי כדורגל. מדידת רמת הדיוק התבצעה באמצעות שיטת k-fold cross validation לצורך אימות ביצועי המודל. התכונות אשר נבחנו מתייחסות לאירועים שהתרחשו במהלך משחק בודד (מספר קרנות, מספר בעיטות, מספר כרטיסים צהובים ואדומים וכדומה..) ולכן אינן משקפות תלות בין משחקים שונים. נראה כי השימוש ברשת בייסיאנית הוביל לרמה

דיוק גבוהה באופן יחסי לאלגוריתמים חיזוי שונים אשר נעשה בהם שימוש לחיזוי תוצאות משחקי כדורגל.

עצי החלטה

במאמרו של [4], נבחנו שלושה אלגוריתמים שונים, מבוססי עץ החלטה, על מנת לבדוק את התאמתם לבעיית חיזוי תוצאות משחקי כדורגל כמודל סיווג (ניצחון, הפסד, תיקו).
במחקרם, בוצע שימוש במאגר מידע אודות 3800 משחקי כדורגל מהליגה האנגלית לאורך 10 עונות בין השנים 2007-2016. המידע הגולמי כלל נתונים סטטיסטים אודות המשחקים, וכן נתוני התחזיות שפורסמו על ידי סוכנויות הימורים. המודלים שנבנו במחקר זה התבססו על הנתונים הסטטיסטים בלבד.
מתוך הנתונים הסטטיסטים, נבחנו 15 תכונות ראשוניות, ובאמצעות שימוש במודל backward wrapper נבחרו מתוכן 10 תכונות מובהקות אשר ישמשו כקלט לבניית מודל החיזוי.
החוקרים בנו 3 מודלים להשוואה – C5.0, Random Forest Algorithm ו- XGBoost, וקיבלו אחוזי דיוק של 64.87%, 68.55%, ו-67.89% בהתאמה. כדי לאמוד את דיוק המודלים, נעשה שימוש בשיטת K-fold cross validation. על פי מסקנות החוקרים, מודלים אלו הראו תוצאות טובות משל מודלים מבוססי SVM ו-logistic regression, אך ביצועיהם נופלים מאלו המשתמשים ברשתות בייסאניות וברשתות נוירונים. החוקרים משערים כי ניתן לשפר את ביצועי המודלים, על ידי שילוב נתונים ממאגרי מידע בעלי תכונות נוספות, וכן שיפור תהליך בחירת התכונות.

ANN

במאמר [5], בוצעה השוואה בין מודל Artificial Neural Network ומודל Logistic Regression בפתרון בעיית חיזוי תוצאות משחקי כדורגל. במחקר, בוצע שימוש במאגר מידע אודות 110 משחקי כדורגל מהליגה האנגלית בעונה אשר התקיימה בין השנים 2014-2015. החוקרים השתמשו בתכונות אשר הופקו משלוש קטגוריות: נתוני המשחק, היסטוריית המשחקים של הקבוצות ואינדקס ביצועים אבסטרקטי אשר נבנה באמצעות תהליך כריית נתונים והפקת ידע Knowledge discovery in database (KDD).
בתהליך ניקוי והכנת הנתונים, השתמשו החוקרים בטכניקת ANN להשלמת הנתונים החסרים עבור כל תכונה, המאפשרת הערכה מושכלת של הערכים החסרים.
על מנת לייעל ולשפר את תהליך הלמידה של המודלים, בוצע תהליך אופטימיזציה על מבנה הרשת, המשקולות ההתחלתיות ופרמטרים נוספים באמצעות אלגוריתם גנטי המדמה תהליך אבולוציוני טבעי.

רמת הדיוק אשר התקבלה עבור מודל ה-ANN הינה 75.04% ללא תהליך האופטימיזציה, ו-85% לאחר השימוש במודל הגנטי. מודל ה-LR השיג רמת הדיוק של 93%, אך סוג מודל זה מוגבל בערכי סיווג בינאריים, ולכן מאפשר לחזות עבור כל קבוצה ערך של ניצחון/הפסד בלבד, זאת בניגוד למודל הבעיה הדורש סיווג מורכב יותר של הערכים (ניצחון/הפסד/תיקו).
כמו כן, החוקרים מצאו כי אינדקס ביצועי השחקנים שהופק בתהליך כריית המידע, הינו התכונה המשמעותית והמובהקת ביותר עבור רשת הנוירונים (זו שקיבלה את המשקל הגדול ביותר).

חוקרים רבים ניסו ליישם שיטות נוספות לחיזוי תוצאות משחקי כדורגל אך לא הגיעו לרמת דיוק מספק.
במאמר[6] בוחנים את רמת הדיוק אשר מתקבלת על ידי שימוש באלגוריתם SVM לחיזוי תוצאות משחקים. ממסקנות המאמר עולה השימוש באלגוריתם הנ"ל אינו מספק עבור חיזוי תוצאות משחקי כדורגל כיוון שתוצאת הדיוק שהתקבלה עבור החיזוי עומדת על כ- 53.3% בלבד. בנוסף נבחנו אלגוריתמים שונים שהביאו לתוצאות שאינן מספקות ולכן לא נכללו בסקירה זו.

טבלת השוואת תכונות:

	Bayesian Networks	Trees	ANN+ LR	feature available in Kaggle dataset
Home Team	X			X
Away Team	X			X
Home Team Shots	X	X	X	X
Away Team Shots	X	X	X	X
Home Team Shots on Target	X	X		X
Away Team Shots on Target	X	X		X
Home Team Corners	X	X	X	X
Away Team Corners	X		X	X
Home Team Fouls Committed	X	X		X
Away Team Fouls Committed	X	X		X
Home Team Yellow Cards	X			X
Away Team Yellow Cards	X	X		X
Home Team Red Cards	X			X
Away Team Red Cards	X			X
Half Time Home Team Goals	X	X		X
Half Time Away Team Goals	X	X		X
Full Time Home Team Goals	X		X	X
Full Time Away Team Goals	X		X	X
Home Team odds			X	X
Away Team odds			X	X
Home Team Attack Strength			X	
Away Team Attack Strength			X	
Home Players' performance			X	
Away Players' performance			X	
Home managers' win			X	
Away managers' win			X	
Home streak			X	X
Away streak			X	X

טבלה להשוואה בין האלגוריתמים שנבחנו:

Authors	Research Title	Data Set Used	Techniques Applied	Prediction Accuracy	Limitation
[3]	Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)	3 seasons of football data match history (2010-2012) in the English Premier League	Bayesian Networks (BNs)	75.09%	learn the conditional probability links between nodes required large amount of data.
[4]	Football Match Prediction with Tree Based Model Classification	10 seasons of football data match history (2007-2016) in the English Premier League	Random Forest, C5.0, XGBoost	68.55%, 64.87%, 67.89%	
[5]	An Improved Prediction System for Football a Match Result	match history records, performance index record and football spreadsheet was extracted from different data sources in the internet	ANN, LR	85%, 93%	Complicated data cleaning and feature extraction. in LR - results only obtain 2 values (don't support Draw values)

דו"ח מדעי

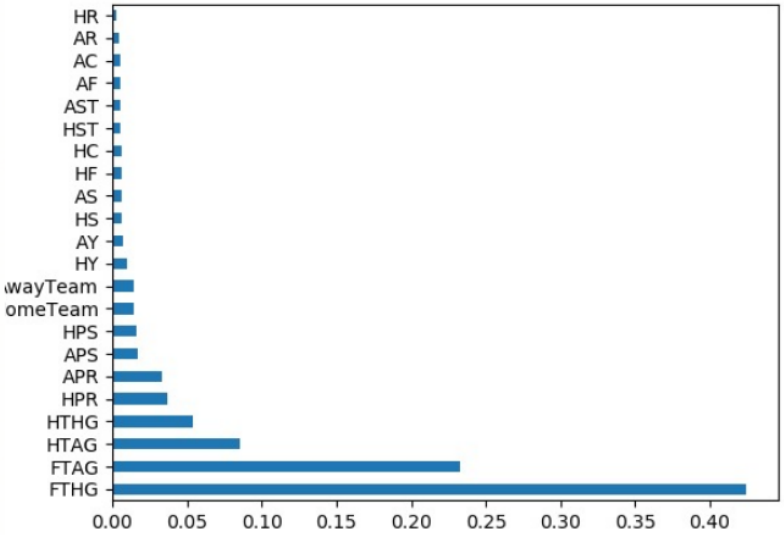
מודל חיזוי

בעקבות הטבלה לעיל, בה פורטו הפיצ'רים הרלוונטיים עבור כל מודל חיזוי שנסקר במאמרים, ביצענו מיפוי עבור הפיצ'רים הרלוונטיים וכיצד הם ממופים ב-DB הנתון. עבור כל פיצ'ר בדקנו באיזה טבלה הוא נמצא וכיצד הוא שמור מבחינת דאטה ב-DB. בעקבות המיפוי יכולנו לדעת יותר טוב באילו טבלאות נצטרך להשתמש ועבור איזה פיצ'רים נצטרך לבצע data extraction.

	Bayesian Networks	Trees	ANN+ LR	Table Name	Type	Relevant Tags
Home Team	X			Match	Integer	
Away Team	X			Match	Integer	
Home Team Shots	X	X	X	Match	XML	team
Away Team Shots	X	X	X	Match	XML	team
Home Team Shots on Target	X	X		Match	XML	team
Away Team Shots on Target	X	X		Match	XML	team
Home Team Corners	X	X	X	Match	XML	team
Away Team Corners	X		X	Match	XML	team
Home Team Fouls Committed	X	X		Match	XML	team
Away Team Fouls Committed	X	X		Match	XML	team
Home Team Yellow Cards	X			Match	XML	team and card_type='Y'
Away Team Yellow Cards	X	X		Match	XML	team and card_type='Y'
Home Team Red Cards	X			Match	XML	team and card_type='R'
Away Team Red Cards	X			Match	XML	team and card_type='R'
Half Time Home Team Goals	X	X		Match	XML	Goal and elapsed<=45
Half Time Away Team Goals	X	X		Match	XML	Goal and elapsed<=45
Full Time Home Team Goals	X		X	Match	Integer	
Full Time Away Team Goals	X		X	Match	Integer	

בחירת הפיצ'רים

בבניית המודל ביצענו תחילה בחינה וניתוח של הנתונים הקיימים בבסיס הנתונים, ע"מ להבין אילו פיצ'רים יכולים להיות רלוונטים. ביצענו הצלבה בין הפיצ'רים שראינו שהם חשובים על פי המאמרים שקראנו, לבין הטבלאות והפי'צרים שתואמים להם בדאטה סט. לפיצ'רים הללו הוספנו פיצ'רים נוספים, אשר ראינו שיש קורלציה בינהם לבין האם הקבוצה ניצחה או לא, כמו הדירוג הכללי של השחקנים. ע"מ לבחור רק את הפיצ'רים שיכולים לנבא בצורה הטובה ביותר האם קבוצת הבית תנצח או לא, בחנו את חשיבותו ותרומתו של כל פיצ'ר ע"י שימוש בפונקציה **feature_importances_**, ולהלן התוצאות:



כפי שניתן לראות, באופן לא מפתיע, שני הפיצ׳רים שמנבאים את התוצאות בצורה הטובה ביותר הינם מס׳ הגולים שהבקיעה כל קבוצה (FTHG ו-FTAG). לא ראינו לנכון לקחת בחשבון את הפיצ׳רים הללו, מהסיבה הפשוטה שברגע שידוע מס׳ הגולים שהביעה כל קבוצה, אין צורך במודל שינבא מי הקבוצה המנצחת.

לעומת זאת, את מס׳ הגולים שהובקעו עד המחצית (HTAG ו-HTHG), כן ראינו לנכון לקחת עבור המודל שכן דבר זה אינו מעיד על הניצחון באופן חד משמעי, וגם באתרי הימורים ניתן להמר על הקבוצה המנצחת עד המחצית. פיצ׳רים נוספים שלא הכנסו למודל שלנו הם ה-AwayTeam ו-HomeTeam, וזאת מכיוון שאחרי מספר ניסיונות שהרצנו מודלים שונים עם הפיצ׳רים הללו, הגענו לאחוזי דיוק גבוהים יותר כאשר הרצנו את אותם מודלים ללא שני הפיצ׳רים הללו. לבסוף, בהתבסס על המידע שקיבלנו, ולאחר ששללנו את הפיצ׳רים שהזכרנו, בחרנו 14 פיצ׳רים אשר משמעותיים לניבוי הקבוצה המנצחת.

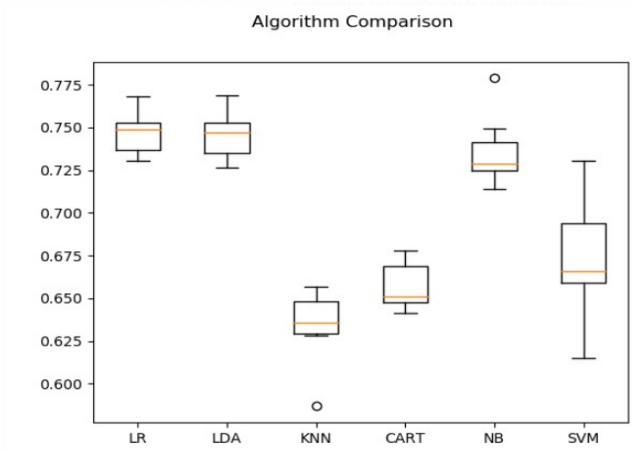
<u>בחירת מודל</u>

לאחר קריאת המאמרים וחומר רב בנושא, ומתוך הבנה שכל מודל ייתן תוצאות שונות בהתאם לפיצ׳רים, החלטנו לבדוק איזה מודל יתאים בצורה מיטבית לפי׳צרים שלנו ע״י ביצוע השוואה של מודלים שונים. המודלים אותם בחרנו להשוואה הינם:

- Logistic regression – מסווג תצפיות כאשר המשתנה התלוי הינו בינארי, אשר מתאים לסיווג נתונים.
- Linear discriminant analysis – מסווג עם גבול החלטה ליניארי, המתבסס על חוק בייס.
- k-Nearest Neighbors Classification – טכניקה של למידה בלתי מונחית (Unsupervised Learning) חסר פרמטרים לסיווג ולרגרסיה מקומית.
- Decision Tree Classification – טכניקה של למידה מונחית (supervised learning), חסרת פרמטרים, המשתמשת לניתוח נתונים לסיווג ורגרסיה.
- Gaussian Naive Bayes – שיטה בסיווג בייסיאני נאיבי, אשר הוא מסווג הוא מסווג הסתברותי שמסתמך על הפעלת משפט בייס.
- Support Vector Machines – טכניקה של למידה מונחית (supervised learning) המשמשת לניתוח נתונים לסיווג ולרגרסיה.

כאמור בחרנו במודלים אלו מלכתחילה, גם בהתבסס על המאמרים שקראנו וגם ע״י הוספה של מודלים נוספים אשר מתאימים לסיווג בינארי כפי שעשינו בעבודה זו.

בכדי לבחון איזה מהמודלים יניב לנו תוצאות טובה ביותר, הרצנו את כולם, כאשר כל מודל קיבל את אותם פיצ׳רים ואת אותו סט של מידע, וביצענו השוואה לגבי אחוזי הדיוק שכל מודל הניב לנו. אימנו את כל אחד מהמודלים הללו באמצעות **K-cross validation**, כאשר K הוא 10, וביצענו ממוצע בין כל התוצאות עבור כל ההרצות עבור כל מודל. השימשו ב-K-cross validation עזר לנו גם להימנע מ-**Overfitting**. להלן התוצאות:




```
LR: 0.746769 (0.011588)
LDA: 0.745018 (0.013427)
KNN: 0.635045 (0.018685)
CART: 0.656741 (0.012273)
NB: 0.734506 (0.018196)
SVM: 0.673013 (0.033984)
```

כפי שניתן לראות בהינתן הפיצ'רים שנבחרו, מודל ה-Logistic regression הניב את אחוזי הדיוק הגבוהים ביותר, וגם בעל שונות נמוכה, ועל כן בחרנו בו. כמו כן, עבור כל אחד מהאלגוריתם בחנו את הפרמטרים **precision**, **recall** ו-**f1-score**, וגם במדדים אלו ה-LR הניב את התוצאות הגבוהות ביותר.

סיכום

לאחר שבחרנו את המודל על בסיס נתוני ה-training, כלומר כל המידע על אודות כלל הליגות חוץ מהשנים 2015-2016, ביצענו **Hyperparameter Tuning**, וזאת בכדי לבחון עם אילו פרמטרים כדי להריץ את המודל בכדי לקבל את התוצאות המיטביות. זאת עשינו ע"י שימוש ב-**GridSearch**, כאשר הפרמטרים רצינו לבצע להם כיוונון, הנם: **penalty, solver, C**, ואלו ערכי הפרמטרים שיצאנו הכי טובים עבור המודל:

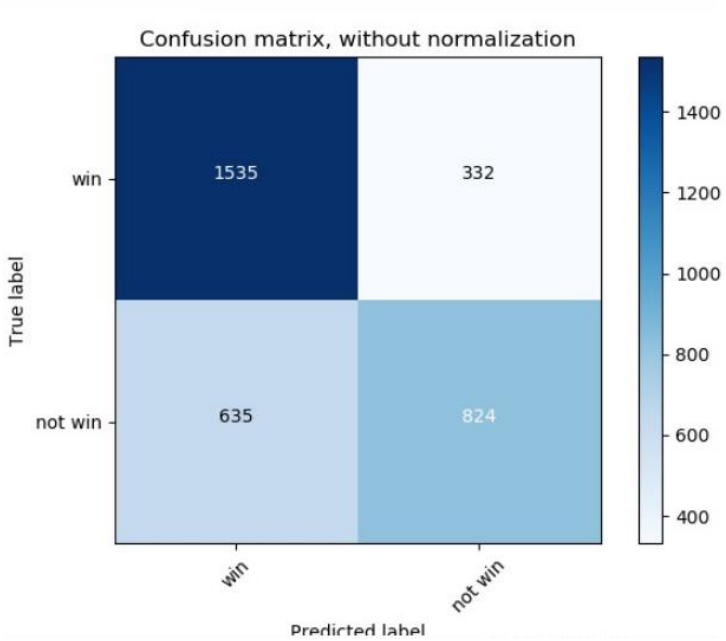
```
('Best Penalty:', 'l2')
('Best C:', 1291.5496650148827)
('Best Solver:', 'saga')
```

לאחר מכן, הרצנו את המודל על השנים הללו, בכדי לבחון את התוצאות. להלן הפלט של המודל לאחר ההרצה ותוצאות החיזוי:

```
LR: 0.709260
              precision    recall  f1-score   support

         0.0       0.71      0.82      0.76      1867
         1.0       0.71      0.56      0.63      1459

   micro avg       0.71      0.71      0.71      3326
   macro avg       0.71      0.69      0.70      3326
weighted avg       0.71      0.71      0.70      3326
```



כפי שניתן לראות המודל שלנו צדק ב-2359 מתוך 3326, כלומר הוא מצליח לנבא **בדיוק של 0.709**

שחזור יצירת המודל

לאחר מיפוי הדאטה, והבנה מהם ההכנות שנצטרך לעשות עבור כל פיצ'ר, התחלנו בחילוץ הדאטה מה-DB.

בשלב הראשון, ביצענו שאילתת SQL על מנת לחלץ את העמודות הרלוונטיות מהטבלאות המבוקשות.

```
SELECT Match.id, League.name AS league_name, season, stage, date,
 shoton, shotoff, goal, corner, foulcommit, card, HT.team_long_name AS home_team,
 AT.team_long_name AS away_team, HT.team_api_id AS home_team_api_id, AT.team_api_id AS away_team_api_id,
 home_team_goal, away_team_goal, home_player_1, home_player_2 , home_player_3 ,home_player_4, home_player_5,
 home_player_6, home_player_7, home_player_8, home_player_9, home_player_10, home_player_11, away_player_1,
 away_player_2, away_player_3, away_player_4, away_player_5, away_player_6, away_player_7, away_player_8,
 away_player_9, away_player_10, away_player_11
FROM Match
JOIN League on League.id = Match.league_id
LEFT JOIN Team AS HT on HT.team_api_id = Match.home_team_api_id
LEFT JOIN Team AS AT on AT.team_api_id = Match.away_team_api_id
WHERE season not like '2015/2016' and goal is not null
ORDER by date
LIMIT 100
```

בשלב השני, ביצענו DATA EXTRACTION עבור כל פיצ'ר.

<u>עבור פיצ'רים שהיו בתצורת XML ביצענו את התהליך הבא:</u>

1. המרה של ה XML לאובייקט מסוג JSON.
2. מעבר על האובייקט ועל כל ה- attribute שלו.
3. חילוץ המידע מהתגית\ תגיות הרלוונטיות והכנסה לווקטור חדש.

נראה מס' דוגמאות חילוץ מ-XML, בכדי לתאר את התהליך שביצענו.

<u>עבור פיצ'ר Card מה-DB</u>

ראשית, יצרנו ארבעה וקטורים חדשים (וריקים) לתיאור סוגי הכרטיסים עבור הקבוצות המשתתפות במשחק: Home ו-Away . לאחר מכן עברנו שורה שורה בדאטה סט ועבור כל תגית XML ביצענו פירוק והמרה של ה-XML לאובייקט מסוג JSON. עבור כל ערך במערך ה-values של האובייקט , בדקנו באיזה קבוצה מדובר וספרנו לכל קבוצה כמה כרטיסים היו לה באותו המשחק. בנקודה זו ביצענו הפרדה על פי סוג הכרטיס- מונה עבור כרטיס צהוב ומונה עבור כרטיס אדום. בסוף המעבר על תגית הXML, לקחנו את ארבעת המונים שלנו (שניים עבור כל קבוצה ששיחקה במשחק) והוספנו אותם לווקטורים שיצרנו בהתחלה. הווקטורים החדשים שנוצרו בתחילה מולאו ולאחר


```
{
    "card": {
        "value": [
            {
                "comment": "y",
                "stats": {
                    "ycards": "1"
                },
                "event_incident_typefk": "70",
                "elapsed": "19",
                "card_type": "y",
                "player1": "30749",
                "sortorder": "0",
                "team": "9790",
                "n": "26",
                "type": "card",
                "id": "375310"
            },
            {...},
            {...},
            {...},
            {...}
        ]
    }
}
```


המעבר על כל הדאטה סט הם נראים כך:

Home Team Yellow Cards	Away Team Yellow Cards	Home Team Red Cards	Away Team Red Cards
1	2	0	0
4	1	0	1

כל שורה מייצגת משחק מסוים וסך כל הכרטיסים המופיעים בשורה מסוימת הם סך הכרטיסים שניתנו במהלך המשחק.

```
{
    "goal": {
        "value": [
            {
                "comment": "n",
                "stats": {
                    "goals": "1",
                    "shoton": "1"
                },
                "event_incident_typefk": "71",
                "elapsed": "12",
                "player1": "30872",
                "sortorder": "0",
                "team": "9823",
                "id": "375301",
                "n": "21",
                "type": "goal",
                "goal_type": "n"
            },
            {...},
            {...},
            {...}
        ]
    }
}
```


<u>עבור פיצ'ר Goals מה-DB</u>

ראשית, יצרנו שני וקטורים חדשים (וריקים) לתיאור הקבוצות המשתתפות במשחק: Home ו-Away . לאחר מכן עברנו שורה שורה בדאטה סט ועבור כל תגית XML ביצענו פירוק והמרה של ה-XML לאובייקט מסוג JSON. עבור כל ערך במערך ה-values של האובייקט , בדקנו באיזה קבוצה מדובר וספרנו לכל קבוצה את כמות השערים שהיו לה עד המחצית באותו משחק. ביצענו זאת על ידי בדיקה של פיצ'ר elapsed, המתאר לנו את הדקה בה הובקע השער. בסוף המעבר על תגית ה-XML, לקחנו את שני ה-count שלנו (אחד עבור כל קבוצה ששיחקה במשחק) והוספנו אותם לווקטורים שיצרנו בהתחלה.

הווקטורים החדשים שנוצרו בתחילה מולאו ולאחר המעבר על כל הדאטה סט הם נראים כך:

Half Time Home Team Goals	Half Time Away Team Goals
1	0
0	0
0	1

כל שורה מייצגת משחק מסוים וסך כל השערים המופיעים בשורה מסוימת הם סך השערים שהובקעו עד המחצית באותו המחשק שניתנו במהלך המשחק.

פי'צרים אחרים שנמצאים ב-DB באופן מפורש, הוכנסו ל-dataFrame בצורתם הגולמית.

בשלב השלישי, יצרנו פיצ'רים ייחודיים עבורנו נתונים וטבלאות נוספות. ראשית יצרנו ארבעה וקטורים ריקים. לאחר מכן, עבור כל קבוצה, עברנו על כלל השחקנים שלה וסכמנו עבור כל קבוצה את סה"כ ה-overall_rating ואת סה"כ ה-strength. מכיוון שנתונים אלו היו חסרים (לא בכל קבוצה היו נתונים על כל 11 השחקנים המשחקים המשחק), ביצענו ממוצע בהתאם למס' השחקנים שעבורם היה המידע. בסוף המעבר על כלל נתוני הקבוצות, הוספנו את הנתונים הנ"ל לווקטורים שיצרנו בהתחלה.

הווקטורים החדשים שנוצרו בתחילה ומולאו לאחר התהליך נראים כך:

home_players_overall	away_players_overall	home_players_strength	away_players_strength
75.8	49.6	89.2	74.5
56.3	81.2	76.2	48.7

בשלב הרביעי, ביצענו Data pre-processing בו עבור כל פיצ'ר מילאנו ערכים חסרים בערכי הממוצע שלו.

בשלב החמישי איחדנו את כלל הווקטורים מהשלבים הקודמים לכדי דאטה סט אחיד.

להלן עמודות הדאטה סט:

Name in Dataset	Description
HomeTeam	data.home_team_api_id
AwayTeam	data.away_team_api_id
FTHG	data.home_team_goal
FTAG	data.away_team_goal
AY	away_team_yellows
HY	home_team_yellows
AR	away_team_reds
HR	home_team_reds
HS	home_team_shots
AS	away_team_shots
HST	home_team_shots_target
AST	away_team_shots_target
HTHG	home_team_half_time_goals
HTAG	away_team_half_time_goals
HC	home_corner
AC	away_corner
HF	home_foul
AF	away_foul
HPR	home_players_overall
APR	away_players_overall
HPS	home_players_strength
APS	away_players_strength

בנוסף לעמודות אלו, צירפנו אל ה-DataFrame את וקטור החיזוי – Winner, שמסמל האם קבוצת הבית ניצחה במשחק או לא.

בשלב השישי, ביצענו בדיקה על הפיצ'רים שבחרנו והרצנו את השיטה feature_importances על מנת שנבחר את 14 הפיצ'רים המשמעותיים ביותר. בנוסף, ביצענו ניסויים חוזרים ונשנים בכדי לבחור את הפיצ'רים הטובים ביותר. הפיצ'רים שנבחרו הם:

Name in Dataset	Description
HS	home_team_shots
AS	away_team_shots
HST	home_team_shots_target
AST	away_team_shots_target
HTHG	home_team_half_time_goals
HTAG	away_team_half_time_goals
HC	home_corner
HF	home_foul
AF	away_foul
HPR	home_players_overall
APR	away_players_overall
HPS	home_players_strength
APS	away_players_strength
AY	away_team_yellows

בשלב השביעי, הרצנו את האלגוריתם LogisticRegression על סט האימון באופן הבא:

```
LR = LogisticRegression()
LR.fit(X, Y)
```

לבסוף, כעת המודל שלנו מאומן על בסיס סט האימון וכל שנותר הוא לחזות באמצעותו האם קבוצת הבית מנצחת עבור כל משחק שנכניס אל המודל את הפיצ'רים שנבחרו.

ביבליוגרפיה

[1] M. Faculty, A. Yezus, and A. Igoshkin, "Predicting outcome of soccer matches using machine learning"
Saint-Petersbg. Univ, 2014

[2] Darwin Prasetio, Dra. Halili, "Predicting football match results with logistic regression"
2016 International Conference on Advanced Informatics: Concepts, Theory, And Application .(ICAICTA), 2016

[3] Razali, N., Mustapha, A., Yatim, F.A., Aziz, R.A. "Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)". In IOP Conference Series: Materials Science and Engineering, vol. 226, no. 1, (2017).

[4] Yoel F. Alfredo, Sani M. Isa, "Football Match Prediction with Tree Based Model Classification"
International Journal of Intelligent Systems and Applications (IJISA), Vol.11, No.7, pp.20-28, 2019.
ijisa.2019.07.03/10.5815 : DOI

[5] C. P. Igiri and E. O. Nwachukwu, "An Improved Prediction System for Football a Match Result"
. IOSR Journal of Engineering, vol. 04, no. 12, pp. 12-20, 2014

[6] C. P. Igiri, "Support Vector Machine—Based Prediction System for a Football Match Result", IOSR Journal of Computer Engineering (IOSR-JCE), vol. 17, pp. 21-26, 2015.