**מבוא**

כמו לרבים ממשחקי הספורט, גם סביב משחק הכדורגל התפתחה תרבות הימורים ענפה. מהימורים פשוטים על זהות המנצחת ועד להימורים יותר מסובכים כמו מי יבקיע ראשון או כמה אחוזי החזקת כדור יהיו לקבוצה מסוימת. המשך קיומם של סוכנויות ההימורים הם העדות החזקה ביותר להיותם הצד המרוויח מכל הסיפור. הצלחות קטנות של אנשים מתגמדות למול סכומי הכסף שמגלגלים אותן סוכנויות. הרבה אנשים סבורים שחוכמתם עומדת להם בכדי לנצח את משחק ההימורים ולהרוויח. האם הם צודקים? האם ידע בכדורגל משפר את סיכויי הזכייה?

במאמר של דנון ושות'[1] נטען כי אין לידע ובקיאות במשחק השפעה משמעותית על אחוזי הפגיעה של מהמרים. לצורך הניסוי נלקחו שלוש קבוצות של מהמרים כל אחת עם רמת הבנה שונה בכדורגל: מהמרים, חובבים וחסרי ניסיון. ביצועי שלוש הקבוצות הייתה זהה.

נרצה לבחון טענה זו. אנו רוצים להוכיח כי ידע מוקדם ויכולת ניתוח יכולים לסייע בזכייה (אם כי במקרה זה אנו עומדים להשתמש ביכולות החישוב של מחשב). כמו כן, נראה כי אחוזי הפגיעה בתוצאה הנכונה אינם הפרמטר היחיד בו יש להתחשב.

הערה - חשוב לציין שכוונתנו היא אקדמית בלבד ואין בה בכדי לעודד אנשים להמר.

**תיאור הפתרון המוצע לבעיה**

הרעיון הוא מימוש אלגוריתם למידה שיעריך, על סמך נתוני עבר, את ההסתברות לכל אחת משלוש התוצאות האפשריות למשחק: ניצחון לקבוצת הבית, לקבוצת החוץ או תיקו. על בסיס ההסתברות שנקבל נממש אסטרטגיית הימור בתקווה שתמקסם את הרווח.

**הגדרות**

האובייקט אותו אנו רוצים לסווג הוא **משחק** בין קבוצת בית לקבוצת חוץ. תכונות המשחק הן תכונות הקבוצות המשתתפות בו בהצגה כזו או אחרת (יבואר בהמשך). המושג אותו אנו לומדים אחוזי ההסתברות לכל אחת מהתוצאות.

באופן פורמלי:

פונקציה כלשהי.

כאשר מתקיים:

**תכונות**

התכונות של משחק הן בעיקרן תכונות של הקבוצות המשתתפות במשחק.

תכונות גלובליות כמו מזג אוויר או שופט, הן בעיתיות משתי סיבות: האחת קשה להשיג נתונים על תכונות אלו אחורה בזמן. סיבה נוספת היא שהאובייקט הוא משחק. האלגוריתם לא יודע להבדיל בין ריאל מדריד לברצלונה ולא יודע להעריך איך שחקניה מתמודדים למשל במזג אוויר קשה. אנו מניחים שהשפעתם של גורמים גלובליים אלה על הביצועים של שתי הקבוצות קרובה להיות זהה.

כמו כן, בבחירת התכונות התייחסנו לרמת הקבוצה ולא ירדנו לרזולוציית השחקנים. בחירה זו נובעת משלוש סיבות עיקריות: ראשית, לא תמיד קל להשיג מידע על כל השחקנים בכל קבוצה. שחקנים רבים הם אלמונים ואין עליהם הרבה מידע. שנית, איכות השחקנים בדרך כלל מגולמת באיכות הקבוצה. הסיבה השלישית היא שגם אם שחקן משחק בקבוצה מסוימת זה לא אומר שהוא ישתתף במשחק. בכל קבוצה משחקים 11 שחקנים בלבד מתוך 18. אף שהרכב הפתיחה בדרך כלל קבוע, קשה להעריך אלו שחקנים ישתתפו במשחק בעקבות אירועים לא צפויים כמו הרחקות, פציעות וחילופים. עם זאת, קיימת עבודה מעניינת בתחום, אשר מסתמכת אך ורק על נתוני השחקנים שמגיעה לרווח של למשחק בממוצע[2].

את התכונות שנבחרו ניתן לחלק לשלושה סוגים:

1. **נתונים סטטיסטיים**:

עבור כל קבוצה אספנו נתונים סטטיסטיים במשך העונה הנוכחית וקומבינציות שלהן: שערים למשחק, נקודות בטבלה, הפרש שערים, כרטיסים, בעיטות למסגרת, מספרים משחקים ללא שערים\ללא ספיגת שערים נתונים דומים מחמשת המשחקים האחרונים ועוד...

הרשימה המלאה מצורפת בנספחים.

סך כל התכונות בקטגוריה זו עבור כל אחת מהקבוצות הוא 100.

1. דירוגים כלליים:

השתמשנו בשני דירוגים חיצוניים:

1. **World Football Elo rating** או בעברית **מד כושר לנבחרות כדורגל** –שיטת דירוג שהתפתחה במקור עבור שחמט והורחבה לתחום הכדורגל. הפרמטרים שנלקחים בחשבון חישוב הדירוג:

* דירוג הקבוצה לפני המשחק
* חשיבות המסגרת בה נערך המשחק (במקרה שלנו משחקי ליגה בלבד)
* הפרש השערים במשחק
* תוצאת המשחק
* התוצאה הצפויה במשחק

דירוג זה הוא דינמי ומשתנה לאחר כל משחק. לכל משחק יש להתאים את הדירוג המתאים של כל קבוצה לפי תאריך.

1. דירוג קבוצות של משחקי וידאו (fifa) – זהו מדד חצי שנתי הניתן לכל קבוצה ע"י יצרני משחקי הוידאו. אנו מניחים שמטרת היצרנים היא לתת חווית משחק קרובה למציאות ולכן הדירוג שלהם מקצועי ומבוסס.
2. **"ראש בראש"** : היסטוריית תוצאות המשחקים בין הקבוצות הספציפיות במשחק. קטגוריה זו כוללת שלוש תכונות: הפרש ניצחונות בין הקבוצות, הפרש שערים בין הקבוצות והפרש ניצחונות בין הקבוצות כאשר ניצחונות רחוקים יותר בזמן קיבלו משקל נמוך יותר.

**איסוף מידע**

את המידע הסטטיסטי חילצנו מהמידע הקיים באתר football-date[2] . אתר זה מכיל רשימות של משחקים עם הנתונים הבאים: מספר שערים, בעיטות לכיוון השער, בעיטות למסגרת, כרטיסים, ויחסי ההימורים של מספר מועדני הימורים מובילים בעולם. עבור כל משחק, חישבנו לכל אחת מהקבוצות את ביצועיה לאורך העונה הנוכחית עד לאותו משחק, תוך שאנו מתחשבים בעובדה שאחת משחקת כעת בבית ואחת בחוץ.

את נתוני ה-**elo**  השגנו מהאתר[3]: \_\_\_\_\_\_\_\_\_

מיינו אותם לפי תאריכים וקבוצות ושייכנו אותם למשחקים הרלוונטיים.

את נתוני משחקי הוידאו אספנו מהאתר[4] \_\_\_\_. גם כאן שייכנו את הדירוג המתאים לפי תאריך למשקים הרלוונטיים.

המידע שנאסף הוא עבור חמש ליגות כדורגל מהבכירות בעולם: הליגה האנגלית, הצרפתית, הגרמנית, הספרדית והאיטלקית. בליגות אלה מהימנות תוצאות המשחקים גבוהה ביותר ותופעות פסולות של שוחד פחות נפוצות בהן. נאסף מידע על 14 עונות: מעונה 2006/2007 ועד לעונה האחרונה 2019/2020. בקבוצת האימון נבחרו כ-16980 משחקים. בקבוצת הוולידציה אשר שימשה לבחירת תכונות וכיוונן פרמטרים היו \_\_\_ משחקים. בקבוצת המבחן היו .

**מספר הערות לגבי המידע:**

* המשחקים בקבוצת האימון קודמים בתאריך לקבוצת הוולידציה וקבוצת המבחן.
* השתמשנו אך ורק במשחקי ליגה מכיוון שבמשחקים מסוג זה בדרך כלל מטרת כל קבוצה היא למקסם את מספר הניצחונות ומספר השערים שלה, בעוד שבמפעלים אחרים כמו משחקי גביע לפעמים קיימת "הסתפקות במועט".
* על מנת למקסם את הדמיון בין קבוצת הוולידציה לקבוצת המבחן, המשחקים לקבוצות אלה נבחרו לסירוגין כך שהמרחק בזמן של משחקי קבוצת הוולידציה וקבוצת המבחן ממשחקי קבוצת האימון פחות או יותר זהים.
* משחקי המחזור הראשון של כל עונה הושמטו מכיוון שכמעט ולא קיים מידע רלוונטי לגביהם.
* בדיעבד התברר כי בקבוצת האימון קיימת קבוצה של 302 משחקים כפולים שהוכנסו לשם בטעות. זהו מספר קטן ביחס לכלל המשחקים ולכן הנחנו שהשפעתם זניחה.
* התכונות עברו נורמליזציה לפי ערכי מקסימום ומינימום של קבוצת האימון.
* אנו צופים כי משחקים במחזורים הראשונים עלולים להוות דוגמאות אימון בעייתיות מכיוון שהקבוצות עדיין לא הפגינו את היכולת שלהן בצורה מלאה. לדוגמא קבוצה חזקה שבמשחקיה הראשונים עדיין לא התבססה וביצועיה לא טובים, אבל מצד שני יש לה דירוג גבוה בדירוגים הכלליים שהוספנו.
* בדרך כלל קבוצת הבית מנצחת.
* במהלך כיוונון הפרמטרים השתמשנו ביחסי ההימורים של . בשלב המבחן בדקנו את ביצועי המסווג גם עבור יחסים אחרים.
* **הערה חשובה:** ישנו בדל מהותי בין הסיווג של דוגמאות האימון לבין הפלט שאנו רוצים להשיג מהמסווג שלנו. בעוד שהראשון הוא סיווג לשלשה כלומר: ניצחון בית, תיקו או ניצחון חוץ, לעומתו, הפלט המצופה הוא הסתברות לכל אחת מהתוצאות. מדוע אנו רוצים דווקא את ההסתברות ולא להכריע מה תהיה תוצאות המשחק ואיך אנו משיגים הסתברות יבואר בהמשך.

**אלגוריתמי למידה**

על מנת לחשב את אחוזי ההסתברות לכל אחת משלוש התוצאות האפשריות השתמשנו בשתי משפחות עיקריות של אלגוריתמי למידה:

**עצים:**

עצים בדרך כלל משמשים על מנת לסווג את הדוגמאות לסוג מסוים. על מנת לקבל הסתברות לכל אחד מהסוגים (תוצאות) השתמשנו בשיטות הבאות:

עץ עם מספר דוגמאות בעלים: במקרה כזה, כאשר מגיעים לעלה, בודקים כמה דוגמאות מכל סוג קיימות בעלה. מחלקים במספר הכולל ומקבלים הסתברות לכל אחת מהתוצאות.

יער אקראי: במקרה זה ניתן לתת לכל אחד מהעצים להתפרש עד עלה יחיד, ואחוזי ההצבעה של העצים הם ההסתברות המבוקשת.

\_\_\_\_\_ extra וגראדיאנט

**רגרסיה:**

שיטה זו משתמשת בדרך כלל לניבויים וכמו כן להערכת הסתברויות. השתמשנו בשיטות הבאות:

רגרסיה לוגיסטית: שיטה זו מניחה קשר לינארי בין התכונות לבין הסיווג. על בסיס דוגמאות האימון היא מחפשת מקדמים מתאימים. שיטה זו מיועדת בעיקרה לסיווג בינארי. על מנת להשיג סיווג טרינארי השתמשנו בשיטות הבאות:

1. רגרסיה מולטי – הרחבה של שיטת הרגרסיה הלוגיסטית ליותר משני סוגים.
2. One vs all: בשיטה זו בונים שני מסווגים: האחד מחשב את סיכויי ההצלחה של קבוצת הבית, השני מחשב את סיכויי ההצלחה של קבוצת החוץ ומה שנשאר הוא הסיכוי לתוצאת תיקו.

רגרסיה לינארית: באופן דומה לרגרסיה הלוגיסטית, גם כאן מניחים קשר לינארי בין התכונות לסיווג. בשונה מהלוגיסטית, תוצאת המסווג היא מספר. על מנת לחשב הסתברות, לקחנו את התוצאה וחישבנו את המרחק היחסי של המספר לכל אחד מהסיווגים של דוגמאות האימון . מרחק זה חלקי סכום המרחקים נותן לנו הסתברות.

**הצגת התכונות במודלים השונים**

ניסינו להציג את התכונות בשני אופנים:

1. פריסת כל התכונות של שתי הקבוצות כתכונות נפרדות של המשחק. בשיטה זו לא הוכנסו תכונות ה-"ראש בראש". לדוגמא מספר השערים בממוצע לעונה של קבוצת הבית היא תכונה אחרת ממספר השערים לעונה של קבוצת החוץ.
2. הצגת התכונות כהפרש בין שתי הקבוצות.

אנו צופים כי הצגה זו היא בעייתית עבור עצים. זאת כיוון שהאלגוריתם לא יודע לשייך את התכונות הקשורות אחת לשנייה. מבחינתו כל תכונה עומדת בפני עצמה. כאשר נדרש פיצול בצומת בעץ, הוא משתמש באחת התכונות מבלי להתחשב בתכונה האחרת המשלימה לה.

ההצגה השנייה היא הפרש בין תכונות קבוצת הבית לקבוצת החוץ והוספת תכונות ה-"ראש בראש" (שהיא גם סוג של הפרש). בהצגה זו אנו צופים שיפור עבור עצים מכיוון שכעת המסווג יודע לקשר בין התכונות של קבוצת הבית לתכונות של קבוצת החוץ.

לגבי הרגרסיה אנו מניחים כי השינוי לא יהיה משמעותי. את הקשר בין התכונות האלגוריתם עצמו קובע ע"י בחירת מקדמים. יתכן שהוא מוצא אפילו מקדמים מתאימים יותר מאשר פעולת חיסור פשוטה (מקדם )

**אסטרטגיית הימורים**

ראשית נזכיר כי יחסי ההימורים פועלים לרעת המהמר.

קיימות מספר אסטרטגיות הימורים:

הימורים "עיוורים":

1. הימור רנדומלי. מכיוון שינן 3 תוצאות אפשריות אחוזי הפגיעה בהימור הנכון הוא .
2. *הימור על קבוצת הבית. ברוב המשחקים קבוצת הבית מנצחת.*
3. *הימור על הפייבוריטית. יחסי ההימורים נותנים אחוזי זכייה נמוכים לקבוצה הפייבוריטית. זה נותן לנו אינדיקציה מה הייתה הערכה של סוכני ההימורים.*
4. *הימור על קבוצה שבד"כ מנצחת. לדוגמא בעונת 2019/2020 ריאל מדריד ניצחה ב-26 מתוך 38 משחקים. כלומר אם היינו מהמרים תמיד על ריאל מדריד היינו פוגעים ב- מהפעמים.*

*הערכה מושכלת של תוצאת המשחק:*

1. *הערכה על בסיס ידע מוקדם של המהמר. הכרת משחק הכדורגל ותנאיו. הכרת הקבוצות ויכולותיהן.*
2. *הימור על פי המלצות של מומחי כדורגל\הימורים.*

*בשתי האחרונות ההימור הולך על התוצאה הסבירה ביותר על סמך השיקולים שנלקחו בחשבון.*

*אסטרטגיית ההימור שלנו היא חישוב ההסתברות לכל אחת מהתוצאות. הכפלה ביחסי ההימורים. הימור על התוצאה עם תוחלת הזכייה הגבוהה ביותר בתנאי שהיא גדולה מ-1.*

***דוגמא:***

*בהינתן המשחק בין Napoli קבוצת בית לבין קבוצת החוץ שהתקיים ב-02.05.2016*

*ניתנו יחסי ההימורים הבאים:*

|  |  |  |
| --- | --- | --- |
| *בית* | *תיקו* | *חוץ* |
| 1.17 | 7.5 | 17 |

קל לראות כי הערכת סוכנות ההימורים היא לטובת קבוצת הבית. התוחלת היא הפסד ב-

.

נניח כי המסווג החזיר לנו את אחוזי ההסתברות הבאים:

|  |  |  |
| --- | --- | --- |
| *בית* | *תיקו* | *חוץ* |
| 0.6 | 0.3 | 0.1 |

*במקרה זה אנו נהמר על תוצאת התיקו אשר בתוחלת מקבלת יחס זכייה של . ספציפית כאן התוצאה הייתה ניצחון לקבוצת הבית ואז היינו מפסידים במקרה זה. הרעיון הוא שגם אם היינו צודקים הרווח לא היה משמעותי. בהימור לאורך זמן הרווח יתקבץ סביב התוחלת ונרוויח.*

*קבוצת הבית צריכה לקבל הסתברות של לפחות 0.85 כדי להיות רלוונטית.*

*בשלב זה אנו מניחים כי אחוזי פגיעה טובים אינם ערובה לרווח לאורך זמן.*

***בחירת תכונות***

*Rfe וכו'\_\_\_\_\_\_\_*

***כיוונון פרמטרים***

*מספר הפרמטרים לכל מסווג הוא גדול. מספר האפשרויות לבחירת פרמטרים הוא אקספוננציאלי במספר הפרמטרים האפשריים. על מנת לברור פרמטרים מתאימים השתמשנו בשיטה הבאה:*

*הגרלת מספר גדול של קומבינציות של פרמטרים. מיון הקומבינציות לפי רווח. עבור התוצאות הטובות ביותר עם פרמטרים נתונים, להגדיר תווך חיפוש סביב אותה נקודה מתוך הנחה שלאיכות המסווג כפונקציה של הפרמטרים אין שיפועים חדים ולכן באזור של תוצאה טובה סביר למצוא תוצאות טובות אחרות, בתקווה יותר טובות.*

*כיוונון זה נעשה על קבוצת וולידציה בגודל כמעט זהה לקבוצת המבחן.*

***ביבליוגרפיה***

# *[1]* Football gambling three arm-controlled study: gamblers, amateurs and laypersons, [Ronen Huberfeld](https://pubmed.ncbi.nlm.nih.gov/?term=Huberfeld+R&cauthor_id=22890307), [Roman Gersner](https://pubmed.ncbi.nlm.nih.gov/?term=Gersner+R&cauthor_id=22890307), [Oded Rosenberg](https://pubmed.ncbi.nlm.nih.gov/?term=Rosenberg+O&cauthor_id=22890307), [Moshe Kotler](https://pubmed.ncbi.nlm.nih.gov/?term=Kotler+M&cauthor_id=22890307), [Pinhas N Dannon](https://pubmed.ncbi.nlm.nih.gov/?term=Dannon+PN&cauthor_id=22890307), Psychopathology, August 12

# [2] Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics Johannes Stübinger 1,\* , Benedikt Mangold 2 and Julian Knoll 3

[3]<https://www.football-data.co.uk/>.

[4] אתר של ה-elo

[5] האתר של ה fifa

***נספחים***

*רשימת תכונות:*

*הערה – במ"ל – בממוצע למשחק.*

*במל"ב – בממוצע למשחק בית.*

*דירוג elo לקבוצת הבית*

*דירוג fifa לקבוצת הבית*

*לכל התכונות הבאות יש תכונה דומה נוספת המתייחסת רק למשחקי בית ששוחקו.*

*מספר ניצחונות לקבוצת הבית במ"ל*

*מספר תוצאות תיקו לקבוצת הבית במ"ל*

*מספר הפסדים לקבוצת הבית במ"ל*

*מספר שערים לקבוצת הבית במ"ל*

*מספר שערים שספגה קבוצת הבית במ"ל*

*הפרש שערים לקבוצת הבית*

*מספר נקודות למשחק לקבוצת הבית במ"ל*

*מספר משחקים עם רשת נקיה לקבוצת הבית*

*מספר משחקים בלי הבקעות לקבוצת הבית*

*מספר בעיטות במ"ל*

*מספר בעיטות למסגרת במ"ל*

*מספר בעיטות של היריבה במ"ל*

*מספר בעיטות למסגרת של היריבה במ"ל*