**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, ואפילו הימורים איזוטריים כמו מה סכום מספר החולצות של השחקנים שיכבשו גול או האם איזו קבוצה תבעט את כדור הקרן האחרון. 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.

מצד שני יש לציין שהימורי הספורט לא נותנים יחס זכיה שווה עבור שתי הקבוצות במשחק, אלא מתעדפים את אחת הקבוצות ונותנים לה יחס נמוך יותר-הקבוצה שסוכנות ההימורים חושבת שהיא הפייבוריטית. כלומר-בקביעת היחסים על ידי סוכנות ההימורים קיים אלמנט של הבנה בכדורגל וייתכן מאוד שגם משולבים היבטי בינה מלאכותית. גם מהמרי ספורט רבים משוכנעים שידע עוזר להם להרוויח, ויש אנשים שטוענים לרווח קבוע ומתמשך, וקיימים אתרים שנותנים עצות למהמרים שפועלים בהצלחה מסוימת כבר שנים רבות.

נרצה לבחון טענה זו של דנון ושות' אל מול הטענות הרבות לכך שהם טועים . 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. קיימים מחקרים שמטרתם לחזות בהסתברות גבוהה מי הקבוצה שתנצח-זה נושא מעניין אבל לא הנושא שבחרנו להתמקד בו. לא קשה להגיע לאחוזי דיוק טובים למדי לגבי איזו קבוצה תנצח-קל יחסית להמר תמיד על הקבוצה הפייבוריטית. העניין הוא שהימור בשיטה כזאת יוביל כנראה להפסד בטווח הארוך כפי שיוסבר בהמשך. החוכמה אינה להמר על הקבוצה שהכי סביר שתנצח אלא על הקבוצה שיחס הזכיה עבורה משתלם ביחס להערכה שלנו את סיכויי הניצחון שלה.

הערה - חשוב לציין שכוונתנו היא אקדמית בלבד ואין בה בכדי לעודד אנשים להמר, או להמליץ להמר בשיטה מסוימת על חשבון אחרות או לתת טיפים למהמר, ואנחנו לא לוקחים אחריות על כל שימוש במסקנות פרויקט זה למטרות שאינן אקדמיות נטו.

הערה נוספת: בפרסומות ל"ווינר" הישראלי מופיעה הערה שהזכיה תלויה במזל בלבד-רמיזה לכך שאין לצפות לאלמנט משמעותי של יכולת.

עוד יש לציין שלצורך הפרויקט לא בוצע שום הימור, אמנם השתמשו בנתונים של סוכנות הימורים בריטית שאינה זמינה למשחק בישראל(לא חוקי בישראל להמר באתרים מחו"ל) אבל לא נעשה שום הימור אמיתי - הכל בסימולציה שמבוצעת אצלנו במחשב.

בתמונה-"פול התמנון"-שזכה לפרסום במונדיאל 2010 על ידי אחוזי הצלחה גבוהים מאוד ב"חיזוי" איזו קבוצה תנצח, על ידי הליכה לקופסת האוכל המתאימה. למרות שכמובן שתוצאות אלה מקריות לחלוטין (בהנחה שלא שמו אוכל שהוא מעדיף באחת הקופסאות, ולטענת הבעלים שלו-שמו אוכל זהה בשתי הקופסאות) עדיין ההצלחה של התמנון הפכה אותו לאגדה בקרב חובבי כדורגל ומהמרי ספורט. בתמונה זו הוא "מנבא"(נכונה) שספרד תנצח את הולנד בגמר מונדיאל 2010. לצערנו התמנון כבר מת ולא יוכל להוות אצלנו תכונה עבור משחק כלשהו...



**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.

העדפנו להתמקד בהימור הבסיסי יחסית ולא דברים יותר איזוטריים יותר כמו שצוינו למעלה, משום שהימורים אלה נפוצים בכל סוכנויות ההימורים, כמעט לכל משחק שקיים, ולרוב התוחלת שלהם טובה יותר, בנוסף אלה נתונים שלהערכתנו יש יותר אלמנט של היסק לוגי מאשר מזל נטו מול נתון כמו מספרי החלוצות של הכובשים.

**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).

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. גם קשה להעריך השפעה של שחקן בודד בפני עצמו אלא יש חשיבות לשיטת המשחק, שיתוף פעולה עם שחקנים אחרים, כושר השחקן(אם הוא חזר מפציעה) ועוד. 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. **Elo rating** או בעברית **מד כושר**–שיטת דירוג שהתפתחה במקור עבור שחמט והורחבה לתחום הכדורגל ומשחקים תחרותיים רבים נוספים-וידוע כמדד חזק עבור חיזוי מי מנצח. עיקרון השיטה הוא שניצחון על קבוצה טובה זה יותר "נחשב" מאשר ניצחון על קבוצה נחותה, והפסד לקבוצה נחותה נחשב יותר גרוע מאשר הפסד לקבוצה טובה.

באתר שממנו אנחנו לוקחים את דירוגי הelo, שמתוחזקים קבוע מתוצאות של שנים רבות(על ידי סקריפט משלהם) מכל הליגות הגדולות, נלקחים בחשבון הפרמטרים הבאים בנוסף לדירוג הקלאסי:

* הפרש השערים במשחק-ככל שניצחון הוא בהפרש גדול יותר ככה הוא משפיע יותר
* משחקי בית וחוץ-ניצחון בית נחשב פחות מאשר ניצחון חוץ, שכן יותר קל לנצח בבית.

פירוט מלא של השיטה מופיע באתר הבא:

<http://clubelo.com/System>

\*יש לציין שהאתר לוקח בחשבון גם משחקים במפעלים שאנחנו לא מסתכלים עליהם-כמו ליגת האלופות וגביעים.

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

למרות שיש שיטה להמיר תוצאות אלה לחיזוי הסתברויות החלטנו שלא לעשות זאת אלא להתחשב בדירוג הelo כתכונה נוספת עבור כל משחק.

הדירוג המקסימלי האפשרי לקבוצה הוא קצת יותר מ2000(ולרוב 2000 זה חסם עליון), הנתון הנמוך ביותר שיש לנו נתונים עליו הוא קצת מעל 1000 ולכן החלטנו שכל קבוצה שאין עליה משום מה נתונים עבור משחק נתון אז ניתן לה ערך של 1000-בגלל שאנחנו מניחים שקבוצות שאין עליהן נתונים יהיו כנראה קבוצות שעלו מליגה נמוכה ואינן קבוצות חזקות כל כך.

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.

מחשב המחשב FIFA הוא משחק מחשב שמהווה סוג של סימולטור כדורגל, אהוב מאוד על חובבי כדורגל ומשחקי מחשב בעולם. במשחק מיוצגות הקבוצות האמיתיות, עם שחקנים שאמורים לשקף את השחקנים מהעולם האמיתי. הייצוג אינו רק במראה והשם של השחקנים והקבוצות אלא גם בנתונים שקובעים את יכולת השחקן במשחק-למשל מהירות, גובה, יכולת הבקעה, יכולת הגנה וכו'.

נתונים אלה נקבעים על ידי מומחי כדורגל מרחבי העולם, וזוכים לא אחת לתגובות מצד השחקנים האמיתיים שמשתמשים בזה כסוג של הערכת מומחים עבורם. לכן אנחנו משערים שניתן להיעזר בציונים אלה כדי להסיק איזו קבוצה טובה יותר.

החלטנו שלא להיכנס לרזולוציה של שחקנים בודדים אלא להסתכל על ממוצע הציון של הקבוצה(של ההרכב הפותח), מסיבות דומות לאלו שהוזכרו קודם לגבי למה אנחנו לא מסתכלים על ביצועי שחקנים בודדים.

1. **"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.

ההיגיון אומר שמה שחוזה הכי טוב משחק בין שתי קבוצות הוא משחק שכבר היה בין הקבוצות-וככל שהתנאים יותר דומים אז הוא יחזה טוב יותר. יש לציין שבליגות הגדולות יש שני סיבובים בלבד בכל עונה-משחק בית ומשחק חוץ, ככה שלא ייתכן שבעונה עונה יהיה משחק נוסף שמתקיים עם אותם תנאים בדיוק(במסגרת ליגה-התעלמנו מגביעים וטורנירים כלל אירופיים) לצערנו...

הערה: ניתן היה להתחשב ביחסים שהתקבלו מחברת ההימורים בתור תכונה של המשחק, או ביחסים שמספקות חברות אחרות-שכן זה מידע שגלוי בפני מהמר, ומהמרים רבים מתחשבים בו. עם זאת די ברור שזה פרמטר מאוד מוצלח מבחינת חיזוי ההסתברויות שכן חברות ההימורים צריכות לחזות טוב את הסיכויים... לכן כנראה שתכונות אלה היו משתלטות על כל התכונות האחרות וקובעות כמעט לבדן את התוצאה.

לפיכך ראינו בזה נתון לא מעניין שיפגע ברמת העניין של הפרויקט ולא ישפר אותה.

**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 סוכנויות הימורים מובילות 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.

\*יש לציין שיחסי ההימורים משתנים מספר פעמים לפני המשחק –אנחנו מניחים שמדובר ביחס העדכני ביותר שפורסם זמן קצר מאוד לפני תחילת המשחק-אבל זה לא בהכרח המצב. בכל מקרה-זה כן יחס שבשלב מסוים היה זמין למהמרים, וסביר להניח שהנתונים נאספו בצורה אחידה עבור כל המשחקים והליגות.

את נתוני ה-**elo**  השגנו מהאתר[3]: <http://clubelo.com>. מיינו אותם לפי תאריכים וקבוצות ושייכנו אותם למשחקים הרלוונטיים.

המידע שנאסף הוא עבור חמש ליגות כדורגל מהבכירות בעולם: הליגה האנגלית, הצרפתית, הגרמנית, הספרדית והאיטלקית, הליגות הראשונה והשנייה. בליגות אלה מהימנות תוצאות המשחקים גבוהה ביותר ותופעות פסולות של שוחד פחות נפוצות בהן, וגם איסוף המידע עליהן יותר אמין ומסודר מאשר ליגות אחרות, וציוני שחקנים של קבוצות כאלה קיימים במשחק FIFA. 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, בתוצאות משחקים ב90 הדקות החוקיות(כולל זמן פציעות) 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". למשל תוצאת תיקו בגביע אינה אומרת תיקו אלא הארכה(ואולי פנדלים) ואין סיבה לקבוצה להעדיף ניצחון על פני תיקו כזה. לעומת זאת בליגה אין לזה משמעות. יש גם משחקים עם תנאי עליה יותר מורכבים כמו בתים, תוצאות גומלין וכדומה שיכולות לגרום לכך שקבוצה לא אכפת לה להפסיד-ורצינו להימנע מהם.
* 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 ונראה בזה "רעש" קטן.
* 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, או שהמידע שיש כרגע עוד לא מובהק סטטיסטית. אבל מצד שני יש לה דירוג גבוה בדירוגים הכלליים שהוספנו.

עם זאת יש לציין שגם סוכנות ההימורים סובלת מאותו חיסרון בדיוק וגם עבור קובעי היחסים שלהם קיימת אותה בעיה, לכן אפשר להניח שהיתרון של סוכנות ההימורים עלינו אינו גדול יותר בשלב מוקדם של העונה.

* קבוצת הבית נוטה לנצח יחסית לקבוצות החוץ-עובדה זו נבדקה סטטיסטית על קבוצת האימון וגם על קבוצת המבחן. זו עובדה ידועה לכל חובב כדורגל. זה מתבטא בין היתר בכך שקיימים פורמטים כמו ליגת האלופות(שלא כללנו בפרויקט) בהם שערי חוץ שווים יותר מאשר שערי בית-כלומר גם התאחדות הכדורגל האירופית מודעת להטיה זו. היחס הוא: \****להשלים***\*

הסיכוי לתוצאת תיקו ולניצחון חוץ דומה.

\*יש לציין שבעונת 2020 בגלל מגיפת הקורונה נוצר מצב שמשחקים אינם ללא קהל ולכן החשיבות של משחקי בית וחוץ מעט יורדת. עם זאת, מצב זה יפגע באופן דומה גם בסוכנות ההימורים וזו רק עונה אחת מבין מספר עונות שבדקנו. עדיין ראינו לנכון לציין זאת.

* במהלך כיוונון הפרמטרים השתמשנו ביחסי ההימורים של . בשלב המבחן בדקנו את ביצועי המסווג גם עבור יחסים אחרים.
* **הערה חשובה:** ישנו הבדל 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. מצב זה כפי שיובהר בהמשך גורם לא אחת לתוצאות קשות יותר לפענוח ולקשרים מורכבים יחסית בין פעולת אלגוריתם הלמידה ודיוקו לבין מה יהיה הרווח שלנו בפועל.

**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.

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 והוא לא יודע בהכרח להבין שיש תכונות שהקשר ביניהן הכרחי. למשל הוא עלול להסתכל על elo של קבוצת הבית, לראות שהוא גבוה ולהעריך שהיא תנצח-בהתעלמות מכך שייתכן שלקבוצה השנייה יש elo דומה או אפילו עדיף אבל היא קבוצת החוץ וזו תכונה שונה שנבדקת במקום אחר בעץ. עם זאת ייתכן שבמקרים מאוד מסוימים זה אולי עדיף-למשל אם האלגוריתם יזהה שקבוצה שמנצחת תמיד בבית היא קבוצה חזקה מאוד-גם אם הקבוצה השנייה מנצחת בחוץ בדרך כלל(אבל לא תמיד) וההפרש ביניהן כביכול קטן.

כאשר נדרש פיצול בצומת בעץ, הוא משתמש באחת התכונות מבלי להתחשב בתכונה האחרת המשלימה לה.

1. הצגת התכונות כהפרש בין שתי הקבוצות.

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 מקבל השוואת התכונות של קבוצת הבית לתכונות של קבוצת החוץ ויכול להעריך בקלות יחסית את יחסי הכוחות ביניהן.

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 ), ולמשל הוא יוכל בצורה כזאת למצוא אולי קשרים מורכבים יותר כמו הפרש בין ניצחונות של קבוצת הבית להפסדים של קבוצת החוץ שנעלמים בשיטת ההצגה השנייה.

**אסטרטגיית הימורים**

קודם כל נסביר כיצד עובדים הימורי ספורט מהסוג שמכונה money line:

לכל אפשרות בהימור סוכנות ההימורים מגדירה יחס זכיה. אם אדם מהמר על אופציה כלשהי ואופציה זו אכן התרחשה אז הוא מקבל את כספו חזרה כפול יחס הזכיה. למשל יחס זכיה 3 אומר שמי ששם 10 שקלים יקבל 30 שקל חזרה(וסה"כ 20 שקל רווח).

לרוב יחסי הזכיה אמורים לשקף במידה רבה איזו קבוצה טובה יותר וכנראה נקבעים לפי שיטות שונות שחברות ההימורים מעריכות את הקבוצה הטובה יותר, שאנחנו מניחים שמשלבות גם בינה מלאכותית אבל לא רק.

נזכיר כי יחסי ההימורים פועלים לרעת המהמר-סוכנות ההימורים "מקצצת" אחוזים מהיחס שהם חושבים שבאמת מוצדק כי אחרת לא ירוויחו-ואנחנו בדקנו על ידי הימור רנדומלי אלף פעמים ועל ידי חישובים הסתברותיים פשוטים את גודל הקיצוץ. יש לציין שסוכנות ההימורים שונות לוקחות אחוזים שונים ולכן אין לצפות לאותן תוצאות בין סוכנויות שונות.

\*בהערה נוסיף שסוכנויות ההימורים לא בהכרח מקצצות אחוזים באופן זהה מקבוצות הבית והחוץ-ואין דרך טובה לדעת זאת, אלא רק להעריך כמה הקיצוץ הממוצע עבור שלישיית אפשרויות למשחק. לראיה נראה בהמשך שבמקרה של המשחקים וסוכנות ההימורים שלנו-נלקחים אחוזים שונים עבור קבוצות הבית והחוץ. ייתכן גם שישנם שיקולים נוספים בקביעת יחסים מעבר להסתברות טהורה-למשל פסיכולוגיים, לא נוכל להתחשב בכך אבל צריך להיות מודעים לאפשרות הזאת.

\*\*נהמר תמיד רק על משחק בודד, ולא קומבינציות הימורים ושיטות מורכבות יותר שגורמות לתלויות בין משחקים ולסיבוך של הבעיה. העדפנו להתמקד במשחק בודד בלבד-בצורה שתשקף טוב את האלגוריתם הלומד שלנו שחוזה הסתברויות עבור משחק בודד.

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. באופן דומה הימור על תיקו או על החוץ בתקווה שהיחסים עליהם גבוהים יותר.*
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 האנדרדוג-בהנחה שאמנם הוא בדרך כלל מפסיד אבל היחס עליו גבוה יותר.*
5. *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. \****אני לא בטוח שאני רוצה להיכנס לזה, זה גם לא לגמרי נאיבי\***

*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.* \***לא הבנתי מה אתה אומר פה\***

*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. יש לציין שאם האלגוריתם מעריך שאין שום הימור משתלם אז הוא לא יהמר.*

***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.*

*אנו מניחים כי אחוזי פגיעה טובים אינם ערובה לרווח לאורך זמן-כדוגמה ניקח מקרה של קבוצה שמנצחת 90% מהמשחקים בעונה, והיחס עליהם הוא בממוצע 1.1. אם היינו מהמרים תמיד שקבוצה זו תנצח אז נסיים עם 99% מהכסף שהימרנו עליו(לאורך העונה)-כלומר הפסדנו! לעומת זאת אם אנחנו פוגעים בהסתברות של בלבד אבל יחס הזכיה שלנו הוא* 3.5  *הרי שנסיים את העונה עם רווח של 1.5%-עדיף מאשר אם היינו מהמרים יותר "שמרני".*

***השיטה שבחרנו, פרמטרי ההצלחה ולמה זה מהווה אתגר***

*בדרך כלל בעיות למידה מונחית מחולקות בגדול לשתי קבוצות: בעיות סיווג ובעיות רגרסיה.*

*הבעיה שלנו היא לא באף אחת משתי קבוצות אלה.*

*כדי לעשות רגרסיה אנחנו צריכים שיהיה לנו לכל משחק את ה*result *שמנבא בדיוק מה היה הסיכוי האמיתי לכל תוצאה-ניצחון בית, חוץ ותיקו. הבעיה שאי אפשר לעשות כזה דבר! הנתון הזה לא קיים, שכן התוצאה של משחק היא טרינארית-סיווג, ולא ווקטור הסתברויות של מספרים ממשיים.*

*עם זאת-אנחנו לא רוצים שהאלגוריתם שלנו יחזיר רק סיווג כי אז נקבל תשובה טרינארית שמשקפת "מה התוצאה הכי סבירה" לדעת האלגוריתם שלנו, שזה לא מה שיאפשר לנו בהכרח רווח כפי שאמרנו.*

*בנוסף יש בעיה נוספת: מהי פונקציית ה*score *שלנו לפיה נדע איזה אלגוריתם ואילו פרמטרים ותכונות הצליחו? האופציה היחידה היא להמר על משחק ולבדוק אם הרווחנו. אבל משחק בודד אינו מספק מבחינה הסתברותית. כדי לשקף בצורה טובה יותר ביצועי הימורים החלטנו להמר על אלפי משחקים ולהחזיר את הרווח/הפסד המצטבר עבור האלגוריתם הלומד(והתכונות), ולהשתמש בזה כפרמטר להצלחה.*

*לסיכום: האלגוריתם הלומד שלנו מתאמן בתור בעיית סיווג, מחזיר ווקטור של 3 הסתברויות(למשחק בודד) בצורה דומה לרגרסיה-אבל הפרמטר להצלחה הוא מספר ממשי שמהווה הרווח(היחסי) עבור הימור על אלפי משחקים. זה מהווה אתגר-כי האלגוריתם לא לומד ישירות לפי פונקציית המטרה שלו ו"לא יודע" שהשימוש בו אינו ישיר. ייתכן שפרמטרים שמעלים את דיוק המסווג יגרמו דווקא לפגיעה ברווח הסופי בגלל נטיית יתר להמר תמיד על הקבוצה הטובה יותר.*

***אתגר נוסף-קשר מורכב בין איכות אלגוריתם הלמידה לבין הרווח***

*ניקח כדוגמה פשוטה למדי: נגיד שהייתה לנו דרך לדעת שהסיכוי ה"אמיתי" לתוצאה במשחק נתון הוא באמת . נניח שהיחס המתקבל מחברת ההימורים הוא .*

*אם ככה ברור שההימור המשתלם הוא תוצאת הבית(אינדקס 0).*

*אבל אז נגיד שהאלגוריתם שלנו לא מדויק במאה אחוז, ולכן קיבלנו . במקרה הזה ההימור המשתלם הוא תוצאת החוץ ולכן האלגוריתם שלנו יבחר להמר על תוצאת החוץ.*

*מצד שני, התוצאה האם הרווחנו או לא הרווחנו מהימור כלשהו היא לא מספר ממשי רציף אלא דבר מאוד בדיד ובינארי: או שכן(ואז הרווחנו לפי היחס) או שלא(ואז לא הרווחנו).*

*כלומר תוצאות רציפות של האלגוריתם הלומד עוברות דיסקטריזציה לתוצאה בינארית של רווח או הפסד.*

*אחרי אלפי משחקים בתקווה ההסתברות אמורה להתאזן ככה שאלגוריתם טוב אכן יוביל לבחירת הימור טובה בכל משחק, ואף לרווח. עם זאת בגלל העניין הזה ייתכנו הבדלים גדולים מאוד בתוצאה הסופית של רווח או הפסד וגודלם אפילו אם ההבדל בדיוק המסווג/רגרסיה הוא קטן.*

*מסיבה זו גם לא נוכל להשתמש בפונקציית ה"כמה היינו מרוויחים" עבור ערך* result *ברגרסיה-שכן הפונקציה של הרווח תלויה ב****משחקים רבים*** *ולא באחד בודד, ולא ניתן להגדיר עבור משחק בודד כמה האלגוריתם הלומד הצליח או נכשל.*

***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.*

**אלגוריתמי למידה**

בדרך כלל נעשה שימוש בספריית sklearn של פייתון, והאלגוריתמים שקיימים בה, למעט כמה יוצאים מהכלל.

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-פונקציה שקיימת כבר ונקראת predict\_proba. אם נאפשר לעצים להיפרש עד לעלים בודדים יותר מצב שלא נקבל ווקטור הסתברווית אלא ווקטור שאחת התוצות היא 1.0 ושתי האחרות הן 0.0 ו0.0, לא מצב רצוי מבחינתנו עבור חיזוי תוצאות הימורים. ולכן אנחנו מגבילים את מינימום האפשרויות בעלה.

לא ביצענו גיזום מאוחר משום שגם ככה העצים גזומים למדי, וזמני הריצה די גדולים.

תמיד אפשרנו שימוש בכל התכונות כדי שלא להוביל לעץ בודד ומאוד לא יציב שתלוי מאוד בבחירה האקראית, עם בחירת תכונות best בכל צעד ולא אקראית. כמו כן השתמשנו במשקול-מכיוון שנצחונות בית נפוצים בהרבה מאשר שאר התוצאות הרי שיכולה להיות הטיה של התוצאות לטובת בית. משקלנו לפי היחסים בקבוצת האימון בין ניצחון בית, תיקו וניצחון חוץ.

לרוב עצים נוטים לoverfit ושונות גדולה, אבל בכל זאת ננסה להשתמש גם בעצים בודדים, וגם להשתמש בעצים הטובים יותר כבסיס לgradient boost, ונעדיף להשתמש ביערות:

יער אקראי:

במקרה זה ניתן לתת לכל אחד מהעצים להתפרש במלואם, אם כי לא חובה לאפשר זאת כי זה יכול לגרום לovefit (בדקנו פרמטרים על עומק מקסימלי לעץ ומספר התכונות בעלה), ואחוזי העלים של כל העצים ביחד הם ההסתברות המבוקשת. השתמשנו באלגוריתם כפי שהוא מופיע בויקיפדיה, כלומר בכל שלב באלגוריתם הפיצול הוא לפי התכונה הטובה יותר מבין תת קבוצה אקראית (max\_features) בגודל שורש של מספר התכונות או בגודל log2 של מספר התכונות. בנוסף נעשה שימוש בbootstrap עם max samples(כלומר כל עץ מאומן על מספר משחקים זהה למספר הכולל, אבל עם חזרות על חלק מהמשחקים באקראי כמה פעמים).

גם פה נעשה שימוש במשקול.

התקווה היא שכל עץ יפצה על הטעויות של עצים אחרים ובאמצעות מספר רב של עצים נוכל להגיע לאלגוריתם יציב, עם שונות קטנה יותר אבל עדין הטיה נמוכה.

Extra random forest:

אלגוריתם דומה מאוד לקודם, עם שני הבדלים:

* 1. לא עושים boostrap אלא לוקחים את כל המשחקים עבור כל עץ. אפשרי בsklearn שכן לקחת bootstrap אבל זה לא האופציה הדפולטית ולא האלגוריתם המקובל שמוזכר בויקיפדיה.
  2. נקודת הפיצול עבור כל תכונה היא אקראית ולא נקודת הפיצול האופטימלית. עם זאת עדיין נבחרת תמיד התכונה הטובה ביותר מבחינת gini/entropy מבין התכונות הנבדקות עבור הפיצול שנבחר.

שינויים אלה אמורים לפי הדוקומנטציה של sklearn להקטין מעט את השונות ולהגדיל מעט את ההטיה ביחס לעצים אקראיים רגילים. האם זה יעזור גם אצלנו? עוד נדע.

Gradient boosting:

נעשה שימוש באלגוריתם זה מסוג זה, על 3 מסווגי בסיס שונים:

1. כלום-מסווג האפס.
2. מסווג טיפש-האופציה הדפולטית. מסווג נאיבי שלא משתמש בשום אלגוריתם.
3. עץ שלנו-עבור כל מספר וסוג תכונות נבחר העץ הבודד שנתן את התוצאה הטובה ביותר עבור תכונות אלה.

יש לציין שעצי הרגרסיה של האלגוריתם Gradient bossting משתפרים בכיוון הגרדיאנט, בהנחה של פונקציית loss של \*מסווג\*, שזו אינה פונקציית ההפסד/רווח שלנו. כלומר, ייתכן שהעצים שיצאו יהיו מאוד מדויקים מבחינת בחירת המנצח אבל יתנו לו סיכוי טוב מידי, וסיכוי נמוך מידי לקבוצה הפחות טובה לדעתם. לכן אנחנו מניחים שאלגוריתם זה יבחר באחוזים גבוהים יותר בקבוצה הטובה ובהתאם לכך יזכה באחוזים גבוהים יותר, אם כי לא בהכרח ביותר כסף-הימור מאוד "שמרני".

בסופו של דבר האלגוריתם מחזיר את ההסתברות המשוקללת לפי יער עצי הרגרסיה שנוצרו, עם משקל שונה לכל אחד בהתאם לlearning rate.

אנחנו מקווים שעם זאת אלגוריתם זה יצליח להשיג תוצאות טובות לא פחות מאשר אלגוריתמי העצים האחרים שהשתמשנו בהם.

לא בדקנו את כל הפרמטרים הרבים של אלגוריתם זה, אבל כן בדקנו פרמטרים שעוסקים ביער העצים שנוצרו(מספר עלים, עומק וכו'). ניסינו מספר ערכים של learning rate שקטנים או שווים מ0.1 לפי המלצה של הדוקומנטציה, וגודל שונה של דגימות-subsample-מה שמכונה גם stoachastic gradient boost

מטעמי זמני ריצה גדולים בחרנו max features = ‘log2’ ביצירת כל פיצול בעץ, ולא ערכים גדולים יותר, שכן יוצרי sklearn ממליצים על מספר תכונות קטן יחסית לזירוז זמני הריצה הגדולים.

**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.

\****תרחיב יותר כיצד מחושבות ההסתברויות אולי? נשקול את זה***

ובכל אחת מהשיטות בדקנו מספר פותרים שונים עם פרמטרים שונים. כמו כן בדקנו עם התוצאות עם ובלי משקול לתוצאות-שכן בניגוד לעצים זמני הריצה פה לא גדולים כל כך ויכולנו להרשות לעצמנו לבדוק יותר אפשרויות פרמטרים.

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.

ברגרסיה לינארית נבדקו כמה solvers שונים, וכמה שיטות שונות לרגורליזציה ועונש. יש לציין שאמנם המימוש של הרגרסיה הלינארית עצמה הוא של sklearn אבל ההמרה להסתברויות היא אלגוריתם שלנו.

***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*