סקירה מדעית – קבוצה 2

איתי כץ, ינון בן זכרי, עידו עפרוני, אלון גולומבק, תומר זיינפלד

חיזוי בכדורגל

אנליזה והימורים בתחום הספורט הפכו בשנים האחרונות לתחומים הנמצאים בעלייה לאור הנגשת האינטרנט ותחומים שונים של למידת מכונה, ולתוך כל זאת נכנס עולם הספורט ופרט הכדורגל. מכיוון שמדובר בענף מורכב ודינמי בהשוואה לענפים אחרים, וכן תוצאות המשחק מושפעת ממשתנים רבים כמו מורל הקבוצה, יכולות השחקנים, השתלשלות אירועי המשחק ועוד, כדורגל מהווה מוקד עניין עבור מחקרים רבים.

משחק הכדורגל הוא הענף הנפוץ ביותר בעולם. כיום אנשים מתעניינים במשחק גם בצד העסקי, בייחוד בהשקעה בקבוצות כדורגל, הימורים וכלי עזר לצוותים המקצועיים. אם תוצאות הקבוצה לא יעלו בקנה אחד עם שאיפות המשקיעים, הם עלולים להפסיק את מימון הקבוצה. כמו כן, האימונים כיום מבוססים על למידה מנתוני עבר וניתוחים סטטיסטיים שונים על פעולותיהם של השחקנים. על כן, כיום ישנו מחקר רב בתחום של חיזוי תוצאות המשחקים מראש על מנת לשפר את הישגי המועדונים והכנתם לקראת המשחקים. בתחומי ההימורים, חלק מהשיטות המצליחות בקרב קהל המהמרים בספורט בכלל ובכדורגל בפרט היא הסתמכות רחבה על הימורי עבר. מהמרים לוקחים בחשבון את תוצאות ההימורים ומתחשבים בהצלחתם או בכישלונם.

קיימים מאמרים רבים ושיטות רבות לטובת חיזוי תוצאות משחקי כדורגל: שימוש בשיטות למידה סטטיסטיות, שימוש במידע שנאסף ממשחקי הוידאו של FIFA, חיזוי תוצאות על ידי שימוש במספר שיטות של למידת מכונה (Naive Bayes, Bayesian networks, KNN, random forest).

במסמך זה נסקור כעשרה מאמרים שעסקו בתחום זה, נפרט על שלושה מהם וכן נפרט על דרך המימוש שלהם בשפת Python.

סקירת ספרות

מאמר [1]

הכותבים בוחנים את החיזוי תוצאות המשחקים ששוחקו על ידי הקבוצה Tottenham בשנים 1995-96 באמצעות המודל Bayesian networks. במאמר זה ביצועי ה-BN נמדדים בהשוואה לארבע שיטות נוספות של machine learning: naive BN, BN learnt from statistical relationships in data, KNN ועץ החלטה. לשם כך הכותבים משתמשים במספר פיצ'רים בינהם: נוכחות שלושה שחקנים ספציפיים: Sherringham, Anderson, Armstrong, תפקידו במגרש של השחקן Wilson – האם הוא midfield או לא, מיקום גאוגרפי של המשחק – האם משוחק במגרש הבית של קבוצת Spurs או חוץ ואיכות הקבוצה היריבה – פיצ'ר זה נמדד על פי מדד של שלוש נקודות: high, medium ו-low.

ניתן לראות בטבלה הבאה (המבוססת על 78 משחקים ו-30 פיצ'רים: 28 שחקנים, מגרש בית/חוץ ואיכות הקבוצה היריבה) את דיוק המודל בהשוואה לארבעת השיטות הנוספות שצוינו במאמר זה:

Comparison of learner accuracy with expert model data

Train period–Test period	Number of correct predictions by learner					
	Most common	MC4	Naive BN	Hugin BN	Expert BN	KNN
95/96–95/96 season	16 (42.11%)	25 (65.79%)	22 (57.89%)	23 (60.53%)	20 (52.63%)	27 (71.05%)
96/97–96/97 season	18 (47.37%)	26 (68.42%)	25 (65.79%)	26 (68.42%)	25 (65.79%)	32 (84.21%)
Average for full seasons	17 (44.74%)	25.5 (67.11%)	23.5 (61.83%)	24.5 (64.47%)	22.5 (59.21%)	29.5 (77.63%)
Period 1–period 234 95/96	12 (42.86%)	8 (28.57%)	7 (25.00%)	8 (28.57%)	14 (50.00%)	9 (32.14%)
Period 12–period 34 95/96	7 (38.89%)	5 (27.78%)	9 (50.00%)	0 (0.00%)	10 (55.56%)	8 (44.44%)
Period 123–period 4 95/96	2 (25.00%)	4 (50.00%)	3 (37.50%)	2 (25.00%)	3 (37.50%)	4 (50.00%)
Sum for 1995/1996 periods	21 (38.89%)	17 (31.48%)	19 (35.19%)	10 (18.52%)	27 (50.00%)	21 (38.89%)
Period 1–period 234 96/97	11.5 (41.07%)	11 (39.26%)	12 (42.86%)	13 (46.43%)	19 (67.86%)	7 (25.00%)
Period 12–period 34 96/97	7.5 (41.67%)	6 (33.33%)	8 (44.44%)	6 (33.33%)	10 (55.56%)	8 (44.44%)
Period 123–period 4 96/97	5 (62.50%)	4 (50.00%)	2 (25.00%)	2 (25.00%)	3 (37.50%)	3 (37.50%)
Sum for 1996/1997 periods	24 (44.44%)	21 (38.89%)	22 (40.74%)	21 (38.89%)	32 (59.26%)	18 (33.33%)
Period 23 95/96–period 4/1 95/97	6 (33.33%)	7 (38.89%)	7 (30.89%)	7 (30.89%)	9 (50.00%)	8 (44.44%)
Period 234 95/96–period 1 96/97	4 (40.00%)	7 (70.00%)	3 (30.00%)	6 (60.00%)	6 (60.00%)	5 (50.00%)
Period 34 95/96–period 12 96/97	8 (40.00%)	14 (70.00%)	9 (45.00%)	11 (55.00%)	15 (75.00%)	11 (55.00%)
Period 4 95/96–period 123 96/97	6 (20.00%)	6 (20.00%)	8 (26.67%)	4 (13.33%)	22 (73.33%)	7 (23.33%)
Period 4/1 95/97–period 23 96/97	6.67 (33.33%)	6 (30.00%)	8 (40.00%)	6 (30.00%)	16 (80.00%)	8 (40.00%)
Season 95/96–season 96/97	13 (34.21%)	22 (57.89%)	13 (34.21%)	21 (55.26%)	25 (65.79%)	14 (36.84%)
Sum for cross season periods	43.67 (32.11%)	62 (45.59%)	48 (35.29%)	55 (40.44%)	93 (68.38%)	53 (38.97%)
Overall average percentage	40.05%	45.77%	42.26%	40.58%	59.21%	47.21%
Overall disjoint training/data sets	38.48%	38.65%	35.74%	32.62%	59.21%	37.06%

טבלה 1

כאשר מסתכלים על סט אימון ובדיקה נפרדים עבור מספר עונות, ניתן לראות כי expert BN מספק את התוצאות הטובות ביותר, זאת מכיוון שכל שינוי בין עונות שלא קשור לפיצ'רים ראשיים אינו משפיע על חיזוי המודל. יש לציין כי בממוצע אחוזי הדיוק של מודל זה נמוכים ממודלים אותם בחרנו, והפיצ'רים שעליהם מתבסס אינו עולה בקנה אחד עם ה-Dataset שניתן לנו.

מאמר [2]

הכותבים סוקרים מודל לחיזוי תוצאות משחקי ספורט המבוסס על רשתות נוירונים. הכותבים מתבססים על מספר פיצ'רים הקשורים למשחק עצמו, פיצ'רים שנגזרו באמצעות אלגוריתמים למציאת פיצ'רים חדשים וכן פיצ'רים שניתנו על ידי מומחים. מדובר בסקירה, ובהצעה למודל עבור חיזוי תוצאות ספורט, על כן הכותבים לא ציינו במפורש את אחוזי דיוק המודל. לא בחרנו לממש את מודל זה, מכיוון שלא פורטו אחוזי הדיוק של המודל.

מאמר [3]

הכותבים מנסים ליצור מודל לסיווג תוצאת המשחק לניצחון בית, ניצחון חוץ או תיקו במשחקי כדורגל שהתקיימו בליגה האנגלית הראשונה באמצעות שימוש במודלים יערות אקראיים מבוססים עצי החלטה ולמידה עמוקה. המודל משתמש ב-18 פיצ'רים, ביניהם: קבוצת הבית וקבוצת החוץ, שם השופט, ועבור כל קבוצה: מספר הניצחונות, מספר הניצחונות כקבוצת בית וכקבוצת חוץ, מספר הפעמים שהושג תיקו ותוצאות חמשת המשחקים האחרונים. הכותבים מציגים את תוצאות המודל על ידי confusion matrix:

Class	Accuracy	Precision	Recall
Home Win	79.09%	70.58%	88.42%
Draw	81.81%	82.75%	40.67%
Away Win	79.09%	63.88%	69.69%
Average	80.00%	72.40%	66.26%

טבלה 2

ניתן להסיק מהטבלה כי ממוצע אחוזי הדיוק הינו 80%. על אף הדיוק הגבוה, החלטנו שלא לבחור במודל זה כיוון ישנה התחשבות מהותית בפיצ'ר 'שופט המשחק' אשר אינו נמצא ב-Dataset שלנו.

מאמר [4]

הכותבים מנסים ליצור מודל לסיווג תוצאת המשחק לניצחון בית, ניצחון חוץ או תיקו במשחקי כדורגל במדינה אחת באמצעות דגימת משחקים ממדינות נוספות. הכותבים משתמשים במודל של מערכת דירוג דינמית בשילוב עם רשת בייסיאנית היברידית. המודל הוא חלק מתחרות לסיווג של משחקי כדורגל באמצעות למידת מכונה כאשר מערכת הדירוג מתבססת על ניקוד שמתייחס ליכולותיה של קבוצה בהשוואה לקבוצות אחרות בליגה מסוימת, ולאחר מכן תוצאות אלו משמשות כקלט ל-BN לטובת חיזויים עתידיים. ה-dataset עליו מבוסס המאמר מורכב מכ-216,743 משחקים מליגות שונות ברחבי העולם המייצגים את סט האימון וכ-206 משחקים שהתקיימו בשנת 2017 המייצגים את סט הבדיקה. לכל רשומה קיים מידע אודות שמות קבוצות הבית והחוץ, הליגה אליה משויך המשחק, תאריך המשחק והתוצאה הסופית של המשחק במונחי סכם של גולים. דיוק המודל Dolores נמדד בהשוואה למודלים נוספים כחלק מתחרות שהתקיימה בשנת 2017. להלן טבלה המייצגת את תוצאות התחרות בהתאם שנקבעה בעזרת שימוש בפונקציית RPS [11]:

Position	Participant	RPS	Relative performance (%)
1	Team OH	0.206307	100
2	Team ACC	0.208256	99.06
3	Team FK	0.208651	98.88
4	Team HEM	0.217665	94.78
5	Team EB	0.225827	91.36
6	Team LJ[a]	0.231297	89.2
7	Team AT	0.398058	51.83
8	Team LHE	0.451456	45.7
9	Team EDS	0.451456	45.7

טבלה 3

יש לשים לב כי הקבוצה ACC מייצגת את מודל Dolores ומדורגת במקום השני בתחרות עם שגיאת חיזוי הגבוהה בכ-0.94 אחוזים מהמקום הראשון. הסיבה לאי בחירת המודל לטובת המודלים שלנו מכיוון שמדובר במודל מורכב ביחס למודלים האחרים שסקרנו, הכולל בתוכו חישובים מתמטיים רבים ולמידה חישובית ועל כן העדפנו בבחירת מודלים המכיל שיטות חיזוי שיותר מוכרות לנו.

מאמר [5]

החוקרים מנסים לסווג את משתנה המטרה לניצחון בית, ניצחון חוץ או תיקו, בהתבסס על נתונים מ-3 עונות רצופות (2010,2011,2012) בליגה האנגלית לכדורגל. לשם כך, החוקרים בנו רשת בייסיאני, שיטה שלטענתם הוכיחה את עצמה בעבר בחיזויי מזג אוויר ומשקי כדורגל. החוקרים משתמשים בתוכנה בשם WEKA על מנת להריץ את המודל שלהם וב-dataset של 20 קבוצות שמשחקות אחת נגד השניה פעמיים בעונה (משחק בית וחוץ), כלומר 380 משחקים בעונה. הפיצ'רים עליהם מסתמכים החוקרים הינם קבוצת הבית, קבוצת החוץ, ועל הפיצ'רים הבאים עבור כל אחת מהקבוצות (בית וחוץ): בעיטות לשער, בעיטות למסגרת, קרנות, עבירות שבוצעו ,כרטיסים צהובים, כרטיסים אדומים, גולים עד המחצית, גולים במשחק מלא.

החוקרים ווידאו את דיוקו של המודל באמצעות fold cross validation-10 (כלומר בכל איטרציה 90% מתוך 380 המשחקים בעונה יאמן את המודל וישמש כ-training set והשאר כ-test set לבחינת דיוק המודל) וממוצע הדיוק של כל איטרציה עבור כל עונה מוצג להלן:

עונה	דיוק חיזוי באחוזים
2010-2011	75.26
2011-2012	79.47
2012-2013	70.53

ממוצע של 75.09 אחוזי דיוק לעונה. חשוב לציין כי על אף אחוזי הדיוק הגבוהים במודל זה, לא בחרנו בו מכיוון שרק חלק מה-dataset שלנו מכיל את הפיצ'רים שהשתמשו בהם במודל, ולכן יש פחות חומר לאימון המודל.

מאמר [6]

החוקרים מנסים לבנות מערכת לתמיכה בקבלת החלטות ושיפור הסיכוי לניצחון לצוותים מקצועיים וכן סיווג תוצאות לניצחון בית, ניצחון חוץ או תיקו בהתבסס על נתונים מ-5 עונות רצופות (2016,2015,2014,2013,2012), בליגה הספרדית לכדורגל מ-29 קבוצות. החוקרים בנו מודלים במספר שיטות וביניהן: Logistic Regression, Random Forest, Artificial Neural Network, Linear SVM and Naïve Bayes.

החוקרים משתמשים במספר databaseים שונים עם פיצ'רים שונים כמפורט להלן:

- **Match history** - TeamID(מדורג לפי איכות הקבוצה), Shots, Shots on Target, Corners, Yellow Cards, Red Cards עבור כל קבוצה במשחק.
- **Team vs Team** - Home Team's win percentage against the given Away Team.
- **Goal history** - goals, shots, shots on target לחיזוי התוצאה המדויקת של המשחק
- **Player stats** - 46 Player attributes from fifa 18 לשם קביעת ההרכב הטוב ביותר
- **Team stats** - Team stats from fifa 18.

לאחר ניקוי הנתונים וחלוקתם לפיצ'רים ומשתני מטרה, החוקרים מאמנים מודל רגרסיה לוגיסטית על ה-DATA הנ"ל. לשם אבחון הדיוק, משתמשים ב- fold cross validation-10.

בעת הגעת אינפוט חדש מהיוזר, משתמשים במודל המאומן הנ"ל, מריצים רגרסיה לוגיסטית בכדי לקבוע את תוצאת המשחק והכובשים, משתמשים בנתוני השחקנים השונים בכדי לקבוע הרכב טוב ביותר, ולבסוף, משתמשים בנתוני הקבוצות על מנת לקבוע חולשות וחוזקות של כל קבוצה על מנת לקבוע כיצד ניתן לשפר את סיכוי הניצחון לכל קבוצה.

תוצאות המאמר:

	LR	**RF**	**ANN**	**Linear SVM**	**NB**

Match history DB	63.94%	61.53%	63.1%	58.25%	58.63%
Match history + Team vs Team DB	71.63%	69.9%	69.2%	66.95%	63.57%

לא בחרנו במודל זה מכיוון שחסר בו מידע על תוכן הפיצ'רים בהם השתמשו, וכן חסרים לא מעט פיצ'רים ב-dataset שלנו שמהותיים למודל זה. כמו כן, לא למדנו רגרסיה לוגיסטית ולכן העדפנו להשתמש במודלים אחרים עם אחוזי דיוק גבוהים יותר שאנו כן מכירים.

מאמר [7]

החוקרים מנסים לחזות את מספר הגולים שכל קבוצה תבקיע במשחקים של השבוע האחרון בליגה, בהתבסס על נתונים מ-6 עונות קודמות בליגה האירנית ו-29 שבועות בליגה הנוכחית (2014). לשם כך, החוקרים בנו רשת נוירונים, שבשכבת הקלט מכילה 10 קלטים שונים (מפורטים בהמשך), 2 שכבות אמצע עם 20 נוירונים ושכבת הפלט שמחזירה 2 פלטים כאשר כל אחד מייצג את מספר הגולים עבור קבוצה. כל נוירון בשכבת הקלט מכיל ייצוג של קבוצת הבית, קבוצת החוץ, מספר עונה, מספר השבוע ועבור כל אחת מהקבוצות (בית וחוץ) - ממוצע נקודות בארבע המשחקים האחרונים, ממוצע נקודות שהושגו בליגה וממוצע הנקודות של קבוצות שהתמודדו מול הקבוצה ב-4 המשחקים האחרונים שלה.

החוקרים בונים את המודל באמצעות ANN לפי המבנה המפורט לעיל, כאשר קצב הלמידה הינו 0.05, פונקציית הלמידה היא scaled conjugate gradient, פונקציית האקטיבציה בשכבה הראשונה הינה log-sigmoid ובשכבה השניה הינן Poslin, Purelin, Satlin, Satlins, Tansigmoid, Logsigmoid.

המודל בחן 8 משחקים, והורץ 30 פעמים על מנת להגיע לתוצאה יציבה. לתוצאות עשו טבלת וגרף moving range (X-MR), והסירו את הנקודות שחוצות את גבולות ה- Control Chart Limits.

באשר לדיוק חיזוי תוצאות המשחק הדיוק של המודל היה נמוך. אך באשר לדיוק המודל בחיזוי ניצחון בית/חוץ/תיקו – החוקרים הצליחו לחזות 5 מתוך 8 תוצאות נכון (כלומר דיוק של 62.5%).

Table 17 Final MatchesPredicted and Actual Results for Last Week

Results	The last weeks matches							
	Match 1	Match 2	Match 3	Match 4	Match 5	Match 6	Match 7	Match 8
Prediction Results	2 0	2 0	0 1	2 1	0 2	0 2	2 0	2 0
Actual Results	1 1	2 1	0 1	1 0	0 1	0 1	1 2	0 0

טבלה 4

חשוב לציין כי לא בחרנו במודל זה ממספר סיבות: דיוק המודל אינו הגבוה ביותר שנתקלנו בו, הליגה האיראנית מאוד שונה באופייה וברמתה מהליגה האירופית שעליה ה-dataset מסתמך, חלק מהפיצ'רים חסרים ב- dataset או שנדרשים חישובים מאוד מורכבים על מנת להגיע אליהם וכן הפונקציות שמשתמשים בהם בשכבות הביניים אינן מוכרות לנו לעומק.

מאמר [8]

החוקרים מנסים לחזות את המשחק שיגמר בתיקו עם מספר הגולים הגבוה ביותר ככלל ואת תוצאת המשחק בפרט. זאת, בהתבסס על נתונים מעונת 2002/03 בליגה האנגלית לכדורגל עבור 20 קבוצות. במאמר החוקר פורט מספר שיטות הסתברותיות וביניהן התפלגות פאוסונית ובינומית שלילית, לחיזוי כללי בליגה – ולאחר מכן יורד לרזולוציה של חיזוי תוצאת משחק אינדיבידואלי באמצעות התפלגות פאוסונית – שעליה הוא ממליץ בסופו של דבר. הפיצ'רים עליהם מסתמך המאמר:

λ1, λ2 = the average goals rate per match at home and away respectively.

החוקר יוצר מראש טבלה ובה $\lambda 1, \lambda 2$ לכל קבוצה בליגה ע"פ ה-dataset. לאחר מכן החוקר משתמש בנוסחה:

$$\Pr(\text{Home goals} = x, \text{Away goals} = y) = \left(e^{-\lambda_1} * \frac{\lambda_1 x}{x!}\right) \times \left(\frac{e^{-\lambda_2}\lambda_2 y}{y!}\right) = \lambda_1^x \lambda_2^y \frac{e^{-\lambda_1+\lambda_2}}{x!y!}$$

$$\lambda_1 - average\ goals\ rate\ of\ home\ team\ in\ home\ game, \lambda_2 - average\ goals\ rate\ of\ away\ team\ in\ away\ match$$

כאשר עבור כל תוצאה אפשרית של המשחק (במאמר מוצגות כל האפשרויות עד 3 שערים, ואופציה נוספת ל-4 שערים ומעלה), מציבים את x ו-y ומקבלים את ההסתברות לתוצאה זו. החוקר מציין לאחר מכן דרכים לחשב יחסי הימורים – ועל מנת להכריע אם יהיה ניצחון בית/חוץ או תיקו הוא סוכם את כל ההסתברויות העונים על סיווג זה.

Birmingham versus West Ham	
Birmingham win	0.409
West Ham win	0.303
Draw	0.288
Total	1.000

West Ham. / *Birmingham*	*0*	*1*	*2*	*3*	*4 or more*	*Total*
0	0.110	0.110	0.055	0.018	0.006	0.299
1	0.133	0.133	0.066	0.022	0.077	0.361
2	0.081	0.081	0.040	0.013	0.004	0.219
3	0.032	0.032	0.016	0.005	0.002	0.087
4 or more	0.013	0.013	0.006	0.002	0.000	0.034
Total	0.369	0.369	0.183	0.060	0.019	1.000

החוקר מציין כי עדיף לבחור בהתפלגות פאוסונית על התפלגות בינומית שלילית כיוון שהאחרונה אינה פרקטית ודורשת איסוף נתונים רב וחישובים מסובכים. לא בחרנו במודל זה מכיוון שאחוזי דיוק המודל אינם מפורטים.

המודלים שמומשו במסגרת הסקירה

מאמר [9]

החוקרים מנסים לסווג את משתנה המטרה לניצחון בית, ניצחון חוץ או תיקו, בהתבסס על ההימורים שניתנו עבור כל אחת מהתוצאות. עבור כך, החוקרים משתמשים במודל K-Nearest-Neighbors על מנת לנבא את תוצאות המשחקים באמצעות השוואה עם משחקים שתוצאותיהם ידועות ובעלי מאפיינים דומים למשחק המנובא.

KNN משמש לחיזוי רשומות חדשות במערך נתונים על סמך סיווגים ידועים ברשומות במערך האימון. הדמיון מחושב במרחב בעל n מימדים, והוא מוגדר באמצעות המרחק הוקטורי בין נקודה המייצגת רשומה קיימת לבין הנקודה של הרשומה החדשה שברצוננו לחזות את סיווגה במרחב זה.

x^u - מיקום נקודה של רשומה חדשה בעלת סיווג שאינו ידוע במרחב

x_i^j - מיקום נקודה של רשומה במערך הנתונים בעלת סיווג ידוע במרחב

$$dist_i = |x^u - x_i^j|$$

אלגוריתם זה מתבצע בצורה הבאה:

1. בחירת ערך K.
2. חישוב כל ה-$dist_i$ בין הרשומה החדשה לבין הרשומות בעלות הסיווג הידוע.
3. דירוג ערכי ה-$dist_i$ מהקטן ביותר לגדול ביותר.
4. בחירת ה-k מקומות הראשונים מסעיף 3.
5. בחירת הסיווג של הרשומה החדשה לפי סיווג רוב השכנים בקבוצה מסעיף 4.

מהאלגוריתם המתואר לעיל, במודל ה-KNN רשומה חדשה תסווג לפי רוב של k הנקודות הקרובות ביותר לרשומה החדשה במרחב. כאשר רשומה חדשה מסווגת במודל ה-KNN באופן שונה ממה שהיא אמורה להיות מסווגת במציאות, מתרחשת שגיאה. השאיפה שלנו במודל זה היא לבחור ערך ל-k כך שהשגיאה תהיה מינימלית.

במהלך הניסוי המתואר במאמר, נבחרו ערכי k בין הערכים 1 ל-20, תוך ניסיון לשגיאה מינימלית בתוצאות הדיוק של המודל. הנתונים הכילו תוצאות של 153 משחקים שבהן שיחקו 18 קבוצות בליגה הטורקית, כאשר 17 שבועות הראשונים של הליגה בשנת 2015/2016 נלקחו בתור סט האימון, והאחרונים נלקחו בתור סט המבחן. על מנת לשמור על דיוק גבוה, נבחרו בכל בדיקה פיצ'רים בנפרד כדי לא להעמיס על המודל. התוצאות תוארו בטבלה הבאה:

		Full time result			Double chance		
Features	Number of events predicted	k	Number of correct predictions	Percentage of correct predictions	k	Number of correct predictions	Percentage of correct predictions
Name	153	9	69	45.10	1	119	77.78
Cost	153	19	73	47.71	11	119	77.78
Bet	153	**14**	**83**	**54.25**	9	120	78.43
Frequency	153	14	75	49.02	20	123	80.39
Intersection	153	11	79	51.63	8	123	80.39
Rating	153	4	76	49.67	7	118	77.12
Mark	153	2	81	52.94	**1**	**127**	**83.01**
Consistency	153	2	69	45.10	2	117	76.47

טבלה 7

מהטבלה ניתן למפות את המדדים המרכזיים על פיהם הכותבים ניסו לחזות את נתוני המשחק, וניכר כי אחוזי הדיוק הגבוהים ביותר היו עבור אחוזי ההימורים עבור כל משחק. על כן, על פי הכותבים הימורים הינו מדד טוב עבור חיזוי תוצאות המשחק וסיווג לניצחון, הפסד או תיקו.

בסקירה שלנו בחרנו לממש את מודל זה מכיוון שבסט הנתונים שקיבלנו, יש מידע כולל על כלל אחוזי ההימורים עבור כל משחק, ממספר אתרי הימורים אשר מהוויים תשתית טובה לבדיקה של המודל. על מנת לממש את המודל שלנו השתמשנו בפיצ'רים *LBH, LBD, LBA* שהם אחוזי ההימורים שניתנו עבור כל תוצאה (ניצחון קבוצת הבית, תיקו וניצחון קבוצת החוץ) באתר ההימורים *London Betting Shop*. קבוצת האימון הייתה כל העונות מלבד עונה 2016\2015 בסט הנתונים וקבוצת הבדיקה הייתה עונה 2016\2015. ה-*K* שנבחר היה 350.

תהליך בניית אלגוריתם החיזוי בוצע באופן הבא:

1) הכנת הנתונים:
   - *a*) על פי המאפיינים *home_team_goal, away_team_goal* חישבנו האם קבוצת הבית ניצחה, האם היה תיקו או האם קבוצת החוץ ניצחה ובהתאמה הוספנו את הסימונים 1,0,-1 לעמודה בשם *predict*.
   - *b*) מתן סימון עבור כל עונה מ-0 עד 7 כאשר עונה 2016\2015 קיבלה את הסימון 7.
   - *c*) חלוקת הנתונים לסט אימון וסט בדיקה.
2) הרצת המודל:
   - *a*) הרצת המודל *KNN* עם 350 שכנים
   - *b*) ביצוע בדיקות טיוב המודל אשר כללו שינוי ערך ה-*K* וכן בדיקה של אחוזי הימורים מהאתרים השונים.

<u>הממצאים:</u>

	precision	recall	f1-score	support
-1	0.52	0.41	0.46	1012
0	0.50	0.02	0.03	855
1	0.50	0.86	0.64	1459
accuracy			0.51	3326
macro avg	0.51	0.43	0.38	3326
weighted avg	0.51	0.51	0.43	3326

אחוזי הדיוק יצאו 51%. ייתכן והדיוק נמוך יחסית מכמה סיבות: הליגות שעליהן ביצענו את הניסוי הינן תחרותיות יותר מהליגה הטורקית ושוני בפיצ'רים שנלקחו בחשבון בין אתרי ההימורים.

מאמר [10]

הכותבים מנסים לסווג את משתנה המטרה לניצחון בית, ניצחון חוץ או תיקו בהתבסס על נתוני המשחק FIFA.
הכותבים מציינים את האתגר באיסוף מידע עבור פרויקטים מבוססי למידת מכונה מתחום החיזוי ומציעים אלטרנטיבה אחרת של מקור איסוף מידע לגבי יכולות שחקנים ומאפייני קבוצות. אלטרנטיבה זו הינה איסוף נתונים ממשחק המחשב "FIFA", שלטענת המחבר, מדויקים - משום שמטרתם הינה להקנות למשתמש/לשחקן סימולציה "אמיתית" לנתונים מהעולם האמיתי, דבר שנובע מאיסוף מידע נרחב בכדי להפוך את משחק זה לכמה שיותר מציאותי. מידע זה יכול לשמש את תהליך למידת המכונה בפרויקטים שמטרתם לחזות תוצאות משחקים עם נתונים מתאימים.
לפי המאמר רמת דיוק במודל נקבע על פי וקטור של 66 פיצ'רים שעליהם נסביר בהרחבה בהמשך,
במאמר נעשה שימוש במודלי חיזוי שונים (ניתן לראות בטבלה מטה) כאשר המודל שהניב את רמת הדיוק מקסימלית הינו מודל *Linear SVM*. המודל משמש לחיזוי רשומות, כאשר דוגמאות האימון הינם ווקטור של פיצ'רים במרחב הליניארי, כך שלבסוף החיזוי יישיך את החיזוי כמקדם בינארי.

Model	Home Win	Draw	Away Win
Linear SVM	78%	80%	78%
RBF SVM	69%	81%	80%
Logistic Reg	70%	75%	76%
SGD	64%	70%	67%
Multinomial NB	78%	70%	75%

טבלה 8

הכותבים מציינים כי המודל נבחר משום שאנו רוצים להסתמך על נתונים שנאספו על ידי ציידי כישרונות ומומחים בתחום הכדורגל ומסתמכים על כישרונותיהם ויכולותיהם של השחקנים במציאות. בדרך זו, ניתן גם לאמוד את ביצועי הקבוצה שאליה השחקן שייך. הכותבים טוענים כי בקבוצות שלהן יש שחקנים אשר עם יכולות התקפה גבוהות יותר, ינצחו קבוצה שמעסיקה שחקנים עם יכולות הגנה טובות יותר. על כן, יש צורך לסווג את השחקנים בקבוצה ואת

הקבוצה עצמה על ידי SVM ולדעת אם קבוצה נתונה תשויך לניצחון על פני קבוצה אחרת. הפיצ'רים נלקחו מהאתר sofifa.com.להלן הרשימה המלאה:

Attacking	Skill	Movement	Power	Mentality	Defending	Goalkeeping
Crossing	Dribbling	Acceleration	Shot Power	Aggression	Marking	GK Diving
Finishing	Curve	Sprint Speed	Jumping	Interceptions	Standing Tackle	GK Handling
Heading Accuracy	Free Kick Accuracy	Agility	Stamina	Positioning	Sliding Tackle	GK Kicking
Short Passing	Long Passing	Reactions	Strength	Vision		GK Positioning
Volleys	Ball Control	Balance	Long Shots	Penalties		GK Reflexes

טבלה 9

עבור כל תכונה מרכזית ישנם פיצ'רים אשר קשורים אליו לדוגמא:
Attacking כולל את הפיצ'רים הבאים: Crossing, Finishing, Heading Accuracy, Short Passing, Volleys.
לכל פיצ'ר ישנו ערך מספרי בין 0-100 אשר מגדיר כמה השחקן שולט ביכולת המוזכרת. בשל העובדה כי הנתונים מתוארים לנו עבור כל שחקן, ואנו רוצים לתאר את יכולות הקבוצה (אשר מכילה 11 שחקנים) אנו נגדיר פונקציה אגרגציה שתתאר לנו קבוצה כאוסף של מאפיינים של שחקנים המרכיבים את הקבוצה.
בכדי למנוע החלקת יתר הכותבים בוחרים רק את המאפיינים הדומיננטיים של כל קבוצה באופן הבא:
[attacking,skill,movement,power,mentality,defending,goalkeeping]

$$\text{Team Virtual Features(team)} = \begin{cases} \sum \text{top 4 Attacking} \in [0, 2000] \\ \sum \text{top 5 Skill, Movement, P Power, Mentality} \in [0, 2500] \\ \sum \text{top 4 Defending} \in [0, 1200] \\ \sum \text{top 1 Goalkeeping,} \in [0, 500] \end{cases}$$

תרשים חישוב מאפיינים מרכזיים עבור שחקן

פונקציית האגרגציה מחשבת את הסכימה של כל שחקני המגרש האחרים לפי המאפיינים העיקריים שציינו לעיל, ולאחר מכן בוחרת את השחקנים עם היכולות המקסימליות באותו התחום. ישנם 3 תרחישים במשחק: ניצחון של קבוצת הבית, ניצחון של קבוצת היריב, ותיקו. לשם הניסוי הכותבים מגדירים את התרחישים בתור נתונים בינאריים, לדוגמא אם קבוצת הבית תנצח ניתן 0 אחרת ניתן לה 1, זאת אומרת בכדי לגלות את התרחישים האחרים תיקו, והפסד אנו נבצע את אותו התהליך אלה שהמשתנים האלה יהפכו למשתנה אשר אותו נרצה לחזות על פי מודל הsvm.

לדוגמא:

$$Y = \begin{cases} 0 \text{ if home team won} \\ 1 \text{ otherwise} \end{cases}$$

להלן אחוזי הדיוק בניסוי במאמר:

TABLE II: Prediction results comparison

Real Data			
Model	Home Win	Draw	Away Win
Linear SVM	73%	75%	71%
Logistic Reg	73%	72%	74%
Virtual Data			
Model	Home Win	Draw	Away Win
Linear SVM	78%	80%	78%
RBF SVM	69%	81%	80%
Logistic Reg	70%	75%	76%
SGD	64%	70%	67%
Multinomial NB	78%	70%	75%

טבלה 10

ניתן לראות מטבלה 10 כי עבור המודל Linear SVM אחוזי הדיוק היו הגבוהים ביותר. בשל אחוזי הדיוק הגבוהים ביחס לשאר המאמרים, ולאור העובדה כי ה-Dataset שקיבלנו מכיל את כל הנתונים על יכולותיהם של השחקנים מהמשחק FIFA, בחרנו ומימשנו את המודל באופן הבא:

1. על פי המאפיינים *home_team_goal, away_team_goal* חשבו האם קבוצת הבית ניצחה, האם היה תיקו או האם קבוצת החוץ ניצחה ובהתאמה הוסיפו את הסימונים 1,0,-1 לעמודה בשם *predict*.
2. בצעו GroupBy ב-Dataset של יכולות השחקנים על פי מספר מזהה השחקן.
3. עבור רשומות ריקות: חשב ממוצע עבור כל פיצ'ר ב-Dataset ומלא אותן.

4. עבור כל שחקן ועבור כל מאפיין עיקרי מטבלה 9 - חשב את סכום המאפיינים עפ"י תרשים החישוב המוצג מעלה וצור ב-Dataset פיצ'ר מתאים למאפיין.
5. הסר רשומות מה-Dataset שהכיל משחקים ללא מידע שלם על השחקנים שהשתתפו.
6. חלק לעונות לפי labels שונים 0-7 (שנים 2008-2016)
7. חלק לTraining ו Testing כך שעונות המסווגות כlabels 0-6 יסווגו כ-Training והשאר Testing.
8. בחר את הפיצ'רים הגבוהים ביותר (כפי שתואר לעיל) עבור השחקנים בכל קבוצה.
9. סכום את כל הפיצ'רים המרכזיים כפי שמתואר בתרשים החישוב עבור כל קבוצה וצור וקטור.
10. הרץ את מודל ה-SVM באמצעות הכנסת הוקטור שנוצר בשלב 9, ורשומת ה-Predict שיצרת בשלב 1.

התוצאות שקיבלנו לאחר הרצת האלגוריתם עבור ניצחון קבוצת בית, תיקו וניצחון קבוצת חוץ בהתאמה מימין לשמאל:

	precision	recall	f1-score	support
0	0.60	0.20	0.29	923
1	0.73	0.94	0.82	2095
accuracy			0.71	3018
macro avg	0.66	0.57	0.56	3018
weighted avg	0.69	0.71	0.66	3018

	precision	recall	f1-score	support
0	0.00	0.00	0.00	637
1	0.75	1.00	0.85	1868
accuracy			0.75	2505
macro avg	0.37	0.50	0.43	2505
weighted avg	0.56	0.75	0.64	2505

	precision	recall	f1-score	support
0	0.55	0.48	0.51	432
1	0.64	0.70	0.67	571
accuracy			0.61	1003
macro avg	0.59	0.59	0.59	1003
weighted avg	0.60	0.61	0.60	1003

כלומר, 70% דיוק בממוצע, כאשר דיוק תיקו וניצחון חוץ בעלי הדיוק הגבוה ביותר.

המודל שאנחנו בנינו

מתמקד בתכונות הקשורות להימורי המשחק. מדדי היחס של ההימורים משקפים תכונות רבות במשחק, ובמהלך קביעת נתונים אלו נלקחים בחשבון נתונים המעידים על רמת הקבוצה מול הקבוצה היריבה. כך למשל, אם במשחק של קבוצה א' מול קבוצה ב', נקבע יחס הימור נמוך לקבוצה א', נסיק מכך כי קבוצה א' נחשבת קבוצה חזקה יותר מקבוצה ב', וכל זה בהתחשב בנתונים כמו תוצאות משחקים קודמים, אסטרטגיות משחק קודמות, ביצועי השחקנים ועוד. במודל זה, אנחנו בודקים את הקשר בין יחסי ההימורים לסיווג התוצאה.

לשם כך, ועל מנת להתנהל עם משתנים קטגוריאלים כיאה למודל ID3 - בחרנו להמיר את יחס ההימורים המספרי לצורה קטגוריאלית (High,Medium,Low). לטעמנו, האלגוריתם קל להבנה ולמימוש, בעל זמן ריצה נמוך וגמיש (בחירת פרמטרים לבניית העץ כגון עומק וסיבוכיות).

ID3 הינו אלגוריתם לסיווג בעזרת עצי החלטה. אלגוריתם זה הוא מסוג supervised learning algorithms, כלומר, הוא פועל לפי למידה מונחית - טכניקה המאפשרת לבנות "מכונות" שלומדות להכליל פתרונות על בסיס מאגר גדול של דוגמאות פתורות. כל קודקוד בעץ מייצג תכונה, ענף מייצג החלטה ועלה מייצג תוצאה. העץ נבנה תוך חישוב האנטרופיה של כל הרשומות וה-Information Gain בהתאם בעזרת הנוסחאות:

$$Entropy(s) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_V|}{|S|} Entropy(S_v)$$

כאשר c מציין את מספר אפשרויות ההחלטה, ו-p_i מייצג את ההסתברות להחלטה של החלטה $v.\ i$ מציין את מספר האפשרויות שתכונה יכולה לקבל. את פיצול הקודקודים ביצענו ע"י חישוב ה-Gain המקסימלי עבור כל אחד מהפיצ'רים. בהגעת רשומה חדשה ניתן לרדת במורד העץ בהתאם לתכונות של הרשומה עד אשר נגיע לעלה שיציין את ההחלטה לסיווגה.

שלבי מימוש המודל בוצעו בצורה הבאה:

1. טען את קובץ הנתונים.
2. בחר את הפיצ'רים הבאים: תוצאות המשחק, העונה, מספר שערי קבוצת החוץ, מספר שערי קבוצת הבית, ויחסי ההימורים.
3. חלק את המודל ל-training set ו-testing set כאשר עונת 2015/2016 הינה קבוצת הבדיקה וכל שאר העונות קבוצת האימון.

4. השלם נתונים חסרים בפיצ'ר של יחסי ההימור בממוצע של כל עמודה.
5. נרמל את הפיצ'ר של יחסי ההימור לפי הערך המקסימלי והמינימלי בעזרת הנוסחה:

$$data(i,j) = \frac{data(i,j) - data(j)_{min}}{data(j)_{max} - data(j)_{min}}$$

6. חלק את הפיצ'רים של היחסי ההימור למשתנים קטגוריאליים ל-High, Medium, Low.
7. הרץ את המודל ובדוק דיוק.

בעת הרצת המודל ובדיקתו נעשו שינויים בפרמטרים השונים: מספר ה-bins בכל עמודה בעת חלוקה למשתנים קטגוריאליים, עומק העץ, מספר רשומות מינימלי בכל קדקוד בעץ וסיבוכיות העץ ומספר הפיצ'רים שבהם התחשבנו. הדיוק הגבוה ביותר התרחש כאשר מספר ה-bins היה 5, ערכו של עומק העץ היה 10, בכל קודקוד היו מינימום 5 רשומות, סיבוכיות העץ הייתה 100 וכאשר התחשבנו רק בפיצ'רים BWH, BWD ו-BWA:

	precision	recall	f1-score	support
-1	0.71	0.14	0.24	1012
0	0.00	0.00	0.00	855
1	0.46	0.99	0.63	1459
accuracy			0.48	3326
macro avg	0.39	0.38	0.29	3326
weighted avg	0.42	0.48	0.35	3326

ניתן לראות כי דיוק המודל הוא 48%. הסיבה לאחוזים נמוכים אלה יכול לנבוע חוסר השקיפות בנוגע לצורת החישוב של יחסי ההימור. סיבה נוספת יכולה להיות חלוקה לא מספקת בין קבוצת האימון וקבוצת המבחן.

סיכום

מספר מאמר	1	2	3	4	5	6	7	8	9	10	שלנו
מודל	BN	ANN	יערות אקראיים מבוססי עצי החלטה	רשת בייסינית היברידית	BN	LR,RF,ANN, Linear SVM, NB	ANN	הסתברות פואסונית ובינומית שלילית	KNN	Linear SVM	ID3
נתונים	משחקים שוחקו ע"י טוטנהאם בשנים 1995-6	לא מצוין במפורש	משחקי הליגה הראשונה באנגליה	משחקים מליגות שונות ברחבי העולם	עונות 2010-2012 בליגה האנגלית	משחקי הליגה הספרדית בשנים 2012-2016	משחקי הליגה האיראנית בשנת 2014	משחקים משנת 2002-3 בליגה האנגלית	משחקי הליגה הטורקית בשנת 2015-16	נתוני המשחק פיפא	משחקי ליגות שונות משנת 2008-2016
דיוק מקסימלי	60%	לא מצוין במפורש	80%	לא מצוין במפורש	75%	71.63%	62.5%	לא מצוין במפורש	83.01%	80%	47%

מסקנות

חיזוי תוצאות בכדורגל הינו תחום מרתק להרבה בעלי עניין – מאוהדים וצוותים מקצועיים ועד בעלי קבוצות, מעסקים כדוגמת הימורים ואנליזה ועד חוקרים בתחום הבינה המלכותית שמוצאים תחום זה כמרתק לאור המשתנים הרבים הכלולים בו והקושי בחיזויו.

במאמר זה סקרנו מספר מאמרים בתחום זה שכוללים בתוכם מודלים שונים עם הצלחות שונות, כאשר בסופו של דבר מימשנו שני מודלים שמוצעים במאמרים אלו והרחבנו על מודל נוסף שמימשנו בעצמנו. במודלים אלו השתמשנו בשיטות חיזוי מסוג K-nearest neighbors, SVM ו- ID3 לחיזוי התוצאות כאשר אחוזי הדיוק אליהם הגענו היו 70%, 51% ו-48% בהתאמה. כל מודל אומן על נתונים ופיצ'רים שונים המופיעים ב-dataset כאשר סט האימון הינו כל הנתונים עד 2015, והשאר הוקדש לבדיקת המודל.

ניתן להסיק כי בניגוד לשיטות אחרות, SVM - המתבסס על יכולות שחקנים, הניב לנו את התוצאות הטובות ביותר.

על מנת לשפר את מודל זה, אנחנו מציעים דרכים נוספות לחישוב ערכי הוקטורים המרכזיים עבור כל קבוצה. עבור ליגות שונות, יש טקטיקות מקובלות שונות המאפיינות את הקבוצות הזוכות. לדוגמה, בליגה הגרמנית הקבוצות לרב משתמשות בהרכב התקפי, ועל כן נרצה לתת משקל כבד יותר ליכולות ההתקפיות של הקבוצה. כדי לעשות זאת, בחישוב הוקטור הראשי של ההתקפה ניקח שחקן נוסף, ונוריד שחקן על חשבון ערך הוקטור ההגנתי. לעומת זאת, בליגה האנגלית ההגנה חזקה הינה מרכיב חשוב בניצחון משחקי ליגה, ועל כן נוריד מוקטור ההתקפה שחקן, ונעביר אותו לוקטור ההגנה. באופן זה, אנו נגבש מודלים שיוכלו לחזות בצורה מדויקת יותר עבור ליגה מסוימת.

ביבליוגרפיה

[1] Arabzad, S. Mohammad et al. “Football Match Results Prediction Using Artificial Neural Networks; The Case of Iran Pro League.” (2014).

[2] Bunker, Rory P. and Fadi A. Thabtah. “A machine learning framework for sport result prediction.” *Applied Computing and Informatics* 15 (2019): 27-33.

[3] Constantinou, Anthony C.. “Dolores: a model that predicts football match outcomes from all over the world.” *Machine Learning* 108 (2018): 49-75.

[4] Croucher, John S.. “Using statistics to predict scores in English Premier League soccer.” (2004).

[5] Esme, Engin and Mustafa Servet Kiran. “Prediction of Football Match Outcomes Based On Bookmaker Odds by Using k-Nearest Neighbor Algorithm.” *International Journal of Machine Learning and Computing* 8 (2018): 26-32.

[6]Joseph, Anito et al. “Predicting football results using Bayesian nets and other machine learning techniques.” *Knowl. Based Syst.* 19 (2006): 544-553.

[7] Pugsee, Pakawan and Pattarachai Pattawong. “Football Match Result Prediction Using the Random Forest Classifier.” *ICBDT2019* (2019).

[8] Razali, Nazim et al. “Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL).” (2017).

[9] Shin, JongHo and Robert Gasparyan. “A novel way to Soccer Match Prediction.” (2014).

[10] Zaveri, Nilay and Pramila P. Shinde. “Prediction of Football Match Score and Decision Making Process.” (2018).

[11] Wheatcroft, Ed. (2019). Evaluating probabilistic forecasts of football matches: The case against the Ranked Probability Score.