**חיזוי תוצאות משחק כדורגל**

במדינת ישראל מדי שנה מתרחשים מאות משחקי כדורגל, וברחבי העולם מדי שנה מתרחשים אלפי משחקי כדורגל. כך שנראה כי מובן מאליו מדוע חיזוי תוצאות משחקי כדורגל הוא נושא כה חשוב, אך בכל זאת ננסה להסביר זאת בכמה מילים.

קיימות ראיות למשחקי כדורגל כבר משנת 300 לפני הספירה ומאז המשחק רק הלך והפך לחלק בלתי נפרד מהתרבות האנושית, זהו משחק **עתיק**. **הימורים** על כדורגל הם דבר **שבשגרה ואהדה לקבוצות** מחברת בין אנשים. כלל הפריטים הללו יחד מחזקים את המוטיבציה הרבה בחיזוי תוצאות משחקים.

בתהליך המחקר ניסינו להבין אילו **פיצ'רים** משפיעים על תוצאות המשחק, אילו **מודלים** מספקים אחוזי דיוק גבוהים יותר ואילו **מדדים** מספקים השוואה מתאימה בין מודלים.

ה-data set -  לקוח מהאתר https://www.kaggle.com/hugomathien/soccer

נושא המחקר - בהינתן נתונים על משחק, מה הסיכוי שקבוצת הביתה תנצח (יסווג כ-1), קבוצת הבית תפסיד

(יסווג כ- 1-) או שיהיה מצב של תיקו (יסווג כ-0).

בדו"ח זה נציג **שני מאמרים** אשר את האלגוריתמים שלהם מימשנו על הנתונים מה- dataset ככל שיכולנו להתאים את הפיצ'רים, וכן **מודל נוסף** אשר יצרנו אנחנו לפי המסקנות שהסקנו.

1. **סקירה מדעית**

על מנת לבחור את מודלי חיזוי תוצאות משחקי הכדורגל המתאימים ביותר עבורנו, עברנו על **עשרות מאמרים** הבוחנים את תחום החיזוי בעולם הספורט והכדורגל בפרט מזוויות שונות. המיקוד היה בעיקר על מאמרים שמשתמשים באלגוריתמי חיזוי שונים על datasets של משחקי כדורגל מהעבר, במטרה לחזות האם קבוצה תנצח משחק עתידי או לא. ה-datasets שנתקלנו בהם היו מגוונים מאוד, וכללו תוצאות משחקים מהעבר, אירועים ממשחקים, יכולות של שחקנים וקבוצות וכן הימורים.  
מתוך כל המאמרים, **בחרנו את העשרה הרלוונטיים ביותר**, וסקרנו אותם ע"י מספר פרמטרים:

1. כמות הפעמים שהמאמר צוטט
2. כמות ה-features במודל שהמאמר סוקר
3. כמה מה-features נמצאים ב-dataset שקיבלנו
4. האחוז שלהם מסך ה-features שהמאמר סוקר
5. האלגוריתם שמומש במודל
6. אחוז ההצלחה של האלגוריתם

עבור כל פרמטר סומנו שלושת המאמרים הרלוונטיים ביותר, כאשר העדפנו לבחור מאמרים שיש להם מספר גבוה ככל הניתן של features מה-dataset שקיבלנו, שהיחס שלהם מכמות ה-features של המודל יהיה הגבוה ביותר, ושאחוז ההצלחה של המודל שלהם יהיה גבוה ככל הניתן. בטבלאות 1, 2 ניתן לראות את תוצאות הממצאים.

לפיכך, בעבודה זו נסקור את המודלים של שני המאמרים הבאים:

* An ensemble based Genetic Programming system to predict English football premier league games [1]
* A novel way to Soccer Match Prediction [2]



**טבלה 1:** ממצאי סקירה ספרותית



**טבלה 2:** המשך ממצאי סקירה ספרותית

1. **סקירה – מאמר ראשון**

המאמר An Ensemble Based Genetic Programming System to Predict English Football Premier League Games [1]מנסה לחזות את תוצאות משחקי הכדורגל של הליגה האנגלית. תוצאה מוגדרת באופן הבא: ניצחון קבוצת הבית, תיקו או הפסד קבוצת הבית. באופן כללי, עולם החיזוי של תוצאות משחקי כדורגל, הינו מורכב ומכיל בתוכו גורמי סיכון ואי וודאות רבים ומשתנים. בנוסף לכך, הגורמים המשפיעים מגיעים ממגוון רחב של תחומים: אנושיים, חברתיים, עונתיים, סביבתיים ועוד. כל אלו יכולים להסביר את הקושי והמורכבות בחיזוי משחקי כדורגל בדיוק גבוה.

במאמר בוצעה סקירה על חיזוי תוצאות משחקי כדורגל באמצעות אלגוריתמי למידת מכונה שונים. מהסקירה עולה כי החיזוי באמצעות השיטות: רשת נוירונים, תכנות גנטי ורשת בייסיאנית הינו הגבוה והטוב ביותר מבחינת אחוזי דיוק ביחס לשיטות אחרות. חשוב להדגיש, כי עבור השימוש בתכנות גנטי, יצא אחוז דיוק הגבוה ביותר, בסך של 76% ועל כן הוחלט לפתור את הבעיה הנ״ל באמצעות שיטה זו.

* 1. אופן החיזוי המוצע במאמר

בניית מערכת מבוססת **תכנות גנטי** (על ידי אלגוריתמים גנטיים) לטובת חיזוי משחקי הכדורגל בליגה האנגלית. היתרון בשיטה זו, מתבסס על הצלחת האבולוציה בפתרון בעיות. בתהליך זה ניקח אוכלוסייה של פתרונות ונבחר מתוכם רק את המתאימים ביותר לפתרון בעיה. הפתרונות מעורבבים ולהם מוסיפים רעש, נקבל דור חדש של פתרונות הקרוב צעד נוסף לפתרון הבעיה הנתונה. נחזור על התהליך מספר רב של פעמים (דורות) ולבסוף נגיע לפתרון הקרוב ביותר. כתוצאה מכך, נגיע לאחוז דיוק בניבוי תוצאות המשחקים הגבוה ביותר.

* + 1. האלגוריתם

נגדיר את פונקציית החיזוי r:

**f (x1, . . ., x25) = r**

נגדיר ערכי גבולות עבור r:

1: if (r > 1 or r == 1)

2: then it is a home win game.

3: else if (r < 1 and r > -1)

4: then it is a draw game.

5: else // r < -1 or r == -1

6: it is an away win game.

נגדיר את החלקים מהם בנויה המערכת:

* **Terminal Set:** x1, . . ., x25 and 7 random constants with the range [−5, 5]
* **Function Set:** +, −, x, % (protected division, the denominator cannot be 0).
* **Fitness measure:** the number of the wrongly predicted games.
* **GP parameters:** population = 500, the maximum length of the program = 1000, probability of crossover operation = 0.8 and probability of mutation operation = 0.1.
* **Termination criterion:** the fitness measure is 0 or up to 100 generations

שלבי האלגוריתם:

נריץ n פעמים את המערכת וניקח את התוכנית המוצלחת ביותר שהתקבלה בסיומה של כל הרצה. ניעזר בתוכניות שהתקבלו על מנת לייצר הצבעת רוב. ניתן לכל אחת מהן להצביע במשחק ספציפי וניקח את הרוב. בצורה פורמלית:

let be a set of GP-generated functions. , compute . where, are the input features.  
 The output of each function is corresponding to the three possible outcomes. The final output is the one which is supported by most of the functions.   
Assumptions:

1. The functions are all independent from each other and each one can make the right choice with a probability p.
2. The probabilities of the two wrong choices left are equal, i.e. each.
3. The number of functions is odd, here we consider the majority number as more than one half. That is, the system will only return the outcome with more than one half vote. If no outcome gets more than one half vote, the system will return “undecided”. Therefore, if there is a majority vote, it gives an overall correction accuracy which can be computed by using the following binomial formula:

\*where n is the number of the functions

בנספח ג' מפורטת רשימת הפיצ׳רים מהמאמר אשר ניתן היה לחלץ מה dataset הנתון.

בחלק 5 בדו"ח זה מפורטות תוצאות המודל וכן השוואה בין המודלים.

1. **סקירה – מאמר שני**

במאמר A novel way to Soccer Match Prediction [2] הוצגה השאלה האם ניתן לחזות תוצאות של משחקי כדורגל, בדגש על חיזוי תוצאות קבוצות חלשות, ע"י מחקר המבוסס על נתונים ממשחקי מחשב הממחישים את הנתונים הרלוונטיים בצורה וירטואלי. החידוש בבחינה זו הוא שטכנולוגיית משחקי הוידאו כיום מדויקת להפליא והינה תוצאה של מחקרים מעמיקים שהושקעו בהם זמן ומאמצים רבים על ידי חברות המשחקים על מנת להמחיש את המשחק בצורה מציאותית ככל שניתן. לצורך חיזוי תוצאות משחקי האמת נאספו מתוך משחקי הוידאו נתונים על שחקני הקבוצות המקבילים לשחקנים בקבוצות האמיתיות, ונאספו למעלה מ-33 פיצ'רים המאפיינים כל אחד מהשחקנים. הפיצ'רים כוללים תכונות פיזיות כגון מהירות, כוח, ותאוצה לצד יכולות טכניות כגון מסירות, בעיטות, ודיוק.  
עבור כל אחד מהשחקנים מיוצגת היכולות האישיות שלו על סולם הערכים 1-100, תכונות אלו משתנות עם הזמן ומתעדכנות לצד שינויים ביכולותיו של השחקן. הוגדרה פונקציית אגריגציה אשר משלבת את יכולות השחקנים לייצוג ערכי הקבוצה עבור כל אחת מהקבוצות במשחק הרצוי לחיזוי על ידי מערכת למידת מכונה מבוקרת.

* 1. אופן החיזוי המוצע במאמר

לצורך חיזוי תוצאות אופטימלי נבדקו חמישה מודלים שונים החוזים את תוצאות משחקי האמת ע"פ הנתונים הנ"ל, מתוכם**Linear SVM** (Support Vector Machine) הביא את התוצאות הטובות ביותר העומדות על כ-0.8 אחוזי דיוק לעומת המודלים הנוספים שנבדקו במאמר והם -RBF (Radial Basis Function) SVM, Logistic Regression,   
SGD (Stochastic Gradient Descent) and Multivariate NB (Negative Binomal) model.  
בשל אילוץ יצירת פלט בינרי לרוב המסווגים שנבדקו התוצאות הוגדרו כ-0 לניצחון קבוצת הבית או 1 אחרת.  
בנוסף, מוצגת במאמר דרך בחינה נוספת המתבססת על בחינת אסטרטגיית המשחק של כל אחת מהקבוצות.  
לצורך בחינה זו בוצע נרמול של ערכי היכולות של הקבוצה ובעזרתו סווגה אסטרטגיית המשחק שלה. חשוב לציין שבשתי הבחינות החיזוי בוצע על בסיס הנתונים של שלושת המשחקים האחרונים של אותה קבוצה באותו הרכב שחקנים לצורך יצירת רלוונטיות מקסימלית של הערכת יכולות השחקן במשחק הנוכחי. שתי השיטות, בשל שימוש בנתונים המדויקים ביותר על כל שחקן, מאפשרות ייצוג נכון גם לקבוצות חלשות יותר, ואופציות חיזוי ניצחונן מול קבוצות חזקות על בסיס בחירת אסטרטגיות המשחק שלהן.  
במאמר מוצגת יכולת של העולם הוירטואלי לייצג נתונים על העולם האמיתי בצורה מדויקת במיוחד, ניתן היה לאסוף נתונים אלו גם מתוך מידע על יכולות השחקנים בפועל אך לצורך הרצת מודלי למידת מכונה וחיזוי קל יותר להשתמש בנתונים קיימים וזהים הקיימים ומתועדים מבעוד מועד בעולם משחקי המחשב.

בהתאם לנתונים במאמר, התכונות של כל אחת משתי הקבוצות במשחק נתון, מיוצגת ע"י ערכם של 11 השחקנים הפעילים באותו משחק. מימשנו במודל פיצ'רים המוצגים במאמר לצורך ייצוג תכונותיו של כל שחקן - 22 הפיצ'רים המוגדרים בנספח ג' (הנתונים יוצאו מתוך טבלת נתוני השחקנים), הערך הסופי לייצוג תכונות של כל שחקן מוגדר להיות סכום ערכי הערכים הנ"ל.

כחלק מניסיונות האופטימיזציה במהלך ההרצות, התגלה כי כאשר הפרמטר creation עומד על 20, נוצר מצב של **overfitting** במודל, ו-creation שערכו 10 היה בעל הערכים האופטימליים מתוך הניסיונות שבוצעו (ערכים של 5,10,20).

בחלק 5 בדו"ח זה מפורטות תוצאות המודל וכן השוואה בין המודלים.

1. **המודל שלנו - Random forest**

המודל בו השתמשנו הוא מודל random forest שבו הסיווג מתבצע על פי עץ החלטה, כאשר בכל פעם נדגם חלק אחר מהנתונים עליו מתבצע האימון, ולבסוף הסיווג מתבצע על פי מיצוע של החיזוי.

המודל נבנה בעזרת ספריית sklearn.ensemble תוך שימוש בפונקציית RandomForestClassifier עם הפרמטרים criterion=gini, max\_depth=500 עבור המודל עם התוצאות הטובות ביותר (ראה חלק 4.2 וגם חלק 5).

בהשוואה אנחנו נציג את המודל עם אחוזי הדיוק הגבוהים ביותר מבין כל המודלים שבדקנו.

הפיצ'רים בהם השתמשנו למודל זה הינם הפיצ'רים משני המאמרים וכן מספר פיצ'רים נוספים. את רשימה הפיצ'רים המלאה ניתן לראות בנספח ג'.

* 1. תיאור הנתונים, פיצ'רים

חוזקה של קבוצת כדורגל נמדדת על ידי יכולותיה, הן ההתקפיות והן ההגנתיות. לכן יצרנו 2 פיצ׳רים חדשים מתוך טבלת Team\_Attributes המייצגים יכולות אלו. הפיצ׳ר הראשון מייצג את יכולת הקבוצה מבחינה התקפית והשני – את היכולת ההגנתית. התכונות שלקחנו לפן ההתקפי:

buildUpPlaySpeed, buildUpPlayDribbling, buildUpPlayPassing, chanceCreationPassing, chanceCreationCrossing, chanceCreationShooting

התכונות שלקחנו לפן ההגנתי:

defencePressure, defenceAggression, defenceTeamWidth

סכמנו את כל התכונות השייכות לפן ההתקפי וכך יצרנו את הפיצ׳ר – sum\_attack.  
סכמנו את כל התכונות השייכות לפן ההגנתי וכך יצרנו את הפיצ׳ר – sum\_defense.  
עשינו זאת עבור כל קבוצה במשחק (קבוצת בית וחוץ) ולכן קיבלנו סה״כ 4 פיצ׳רים למשחק.

בנוסף לפיצ'רים אלו, השתמשנו בפיצ'רים משני המאמרים עבור יכולות של שחקן וקבוצה.

מאחר והנתונים לא היו איכותיים עבור כל משחק בנפרד, ביצענו ממוצע על פני השנים עבור היכולות של שחקן ושל קבוצה על מנת ליצור את הפיצ'רים הללו.

* 1. הרצת ניסוי

על מנת לבחון את המודל הטוב ביותר מבין האפשרויות השונות בנינו 14 מודלים שונים וביצענו השוואה ביניהם. **על מנת להימנע מ-overfitting** הגדרנו את המדד לעומק העץ בצורה שונה כך שיצרנו מודלים שונים עם הגבלה של עומק העץ לעומק של 1, 2, 5, 10, 50, 250, 500. בנוסף, השתמשנו במדדי פיצול שונים ליצירת העץ, במדד Gini וכן לפי entropy. כך יצרנו 14 מודלים שונים (לפי שבעה המדדים לעומק ושני המדדי לפיצול). ביצענו השוואה בין מודלים אלה כפי שמפורט בחלק הבא.

1. **השוואה בין המודלים**

השתמשנו בשלושה מדדים להשוואה בין המודלים, שלושתם בודקים את דיוקי המודלים לפי מטריקות שונות. הבחירה שלנו במדדים אלו נעשתה לאחר שראינו כי אלה מדדים נפוצים לפיהם מעריכים מודלים רבים במאמרים שקראנו, וכן מדדים אלה ניתנים לשימוש על המודלים שבדקנו.

מדד ראשון בו השתמשנו הוא **accuracy score.** מדד זה בודק כמה מתוך התוצאות שהחזיר המודל הן בדיוק התוצאות שהיו אמורות לחזור לפי המידע האמיתי.

מדד שני הינו ה F1 score (נקרא גם **F-measure**). מדד זה הוא ממוצע הרמוני של ה precision וה-recall עבור סיווג מסוים, שיקרא X לצורך ההסבר. ה precision הוא מספר התוצאות שסווגו כ X על ידי המודל שהן אכן נכונות, כלומר אכן מסיווג X, חלקי מספר התוצאות הכולל שסיווג המודל כ X. ה-recall הוא מספר התוצאות שסווגו כ X על ידי המודל, חלקי מספר התוצאות שהן X לפי המידע האמיתי. כלומר מדד זה מוסיף מידע על ה accuracy כי הוא לוקח בחשבון גם את ערכי ה false positive ולא רק את ה true positive עבור הסיווג.

מדד שלישי הינו **k fold cross validation.** במדד זה מתבצעת חלוקה של ה data ל k חלקים שווים. המודל יוצר את פונקציית החיזוי על k-1 חלקים ומבצע את החיזוי על החלק הנותר. דבר זה נעשה k פעמים, כלומר, פעולת החיזוי מתבצעת על כל חלק מ k החלקים. עבור כל סבב מתבצעת בדיקת דיוק החיזוי כך שלבסוף מתקבלות k תוצאות דיוק של המודל. כאשר מבצעים בדיקה זו עבור כלל המודלים ניתן להשוות את אחוזי הדיוק בסבבים השונים וכן לוודא כי אחוזי הדיוק עבור מודל מסוים נותרים באותו סדר גודל עבור כל הסטים של ה data.

* 1. השוואת המודלים על פי המדדים

האימון נעשה החל מעונה 2008/2009 עד עונה 2014/2015.

הבדיקה נעשתה על עונה 2015/2016.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **F-measure** | **accuracy score** | **k fold cross validation (average)** |
| **SVM linear** | Win: 0.42736009 Tie: 0 Loss: 0.61142857 | 0.47978793903247186 | 52.054 |
| **Random forest** | Win: 0.39256424 Tie: 0.23893805 Loss: 0.54296738 | 0.42909211398277 | 43.24008273 |
| **Genetic prediction** | Win: 0.60235085 Tie: 0.13085765 Loss: 0 | 0.4305472038484666 | 39.45 |

טבלה 3: נתוני המדדים עבור שלושת המודלים

**בנספח א'** מופיעים הגרפים של מדד ה- k-fold לאורך 10 המודלים השונים שנבנו.

**בנספח ב'** מופיעים boxplot עבור אלגוריתם random forest שמכיל 14 מודלים שונים.

על פי ההשוואה ניתן לראות כי המודל random forest שיצרנו עם מדד Gini לפיצול העץ ועומק של 500 הוא המודל הטוב ביותר, אמנם אחוז הדיוק הממוצע שלו מעט נמוך יותר מאשר המודל SVM linear, אך הביצועים שלו טובים יותר כאשר מתבוננים על מדד ה f measure עבור כל סיווג בנפרד.

1. **שיטת ברירת פיצ'רים**

על מנת לבחור את הפיצ'רים הטובים ביותר עבור כל מודל, השתמשנו בשיטת **Recursive Feature Elimination**.

בהינתן משקלים לפיצ'רים, שיטה זו בוחרת את הפיצ'רים המתאימים ביותר על ידי בדיקה באופן רקורסיבי של סטים קטנים יותר ויותר של פיצ'רים, כך שלבסוף נבחרים הפיצ'רים הטובים ביותר לפי המשקלים שהתקבלו.

החישוב של משקלי הפיצ'רים נעשה על ידי מודל Logistic Regression.

בחנו שיטות נוספות לביצוע feature selection כגון שימוש ב Principal Component Analysis. אולם, לאחר קריאת מאמרים בנושא ובפרט את המאמר Predicting Game Day Outcomes in NFL Games [3] אשר השווה בין השיטות החלטנו לבחור בשיטה הרקורסיבית עבור המודל שלנו. בנספח ג' ניתן לראות עבור כל מודל את הפיצ'רים שנבחרו מתוך הפיצ'רים שהוכנסו.

1. **סיכום**

בדו"ח זה ראינו שלוש דרכים שונות מבין עשרות שבדקנו לחיזוי תוצאות משחקי כדורגל. להלן נקודות עיקריות לסיכום:

* חיזוי תוצאות משחקי כדורגל הוא דבר מורכב אשר דורש מידע רב מהמשחקים על מנת להגיע לדיוק גבוה. קיימות גישות שונות לשימוש בנתונים אלה וראינו שלא כל גישה מתאימה לכל סט של נתונים.
* הוספת פיצ'רים נוספים ורבים מהנתונים לא בהכרח תביא את המודל לשיפור בדיוק. ייתכן ויש פיצ'רים שתלויים אחד בשני או שמורידים את הסיכוי לחיזוי נכון, לכן חשוב לבצע **בחירת פיצ'רים** על פי מודל מאומן **לפני הרצת המודל המסווג**. ביצוע בחירת פיצ'רים בצורה זו הגדילה את אחוזי הדיוק במודלים שהרצנו.
* **שימוש בשלושה מדדים** להשוואה בין המודלים אפשר לנו לראות את התוצאות מכמה נקודות מבט ולבחור לבסוף את המודל האיכותי ביותר – הסתכלנו על הדיוק בכל סיווג בנפרד וכן על אחוזי דיוק כלליים באמצעות המדדים השונים.
* **במודל ה random forest** קיימים מאמרים רבים אשר בודקים את תחום ה **overfitting** בהתאם לעומק העץ. היינו שמחים להמשיך ולחקור בנושא זה על מנת להבין את ההשלכות של בניית המודל בצורות שונות לעומק.

למדנו רבות מעבודה זו על המודלים השונים, אך אנו מבינים כי מודל חיזוי איכותי אינו יכול להיבנות בזמן כה קצר ויש להשקיע בכך עבודת מחקר מעמיקה. זהו קצה קצהו של הקרחון, ואנו מקווים להמשיך וללמוד את התחום.

**ביבליוגרפיה**

[1] T. Cui, J. Li, J. Woodward and A. Parkes, "An Ensemble Based Genetic Programming System to Predict English Football Premier League Games", 2013.

[2] J. Shin and R. Gasparyan, "A novel way to Soccer Match Prediction", 2014.

[3] J. Klein, A. Frowein and C. Irwin, "Predicting Game Day Outcomes in National Football League Games", 2018.

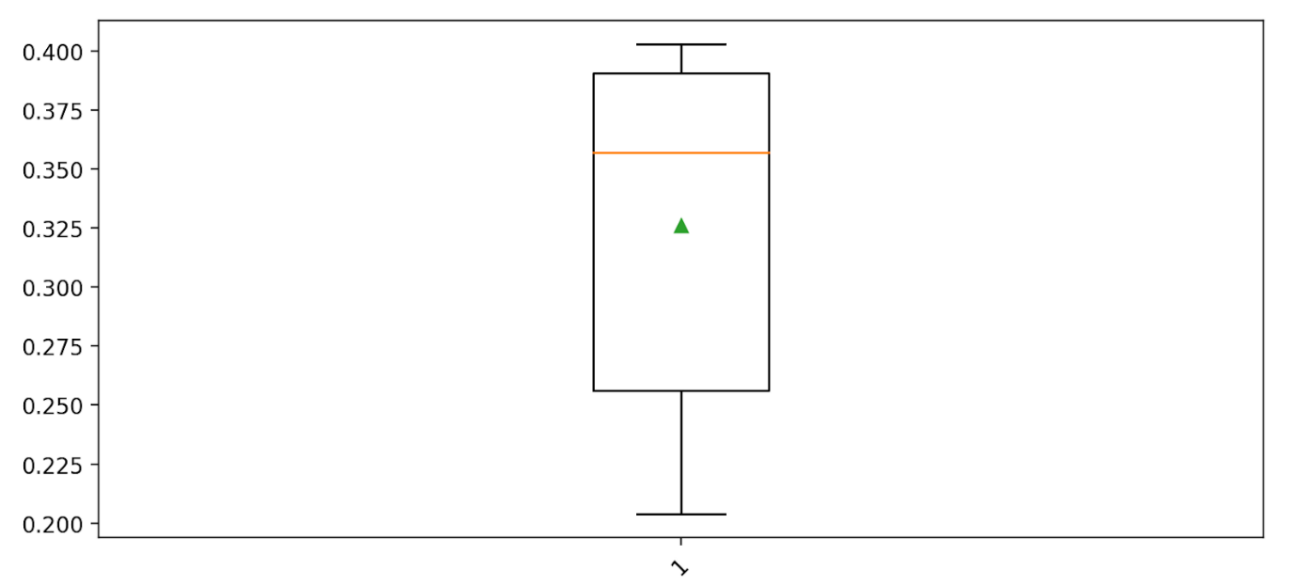
**נספח א'**

השוואה של מדד ה- k-fold עבור המודלים השונים:

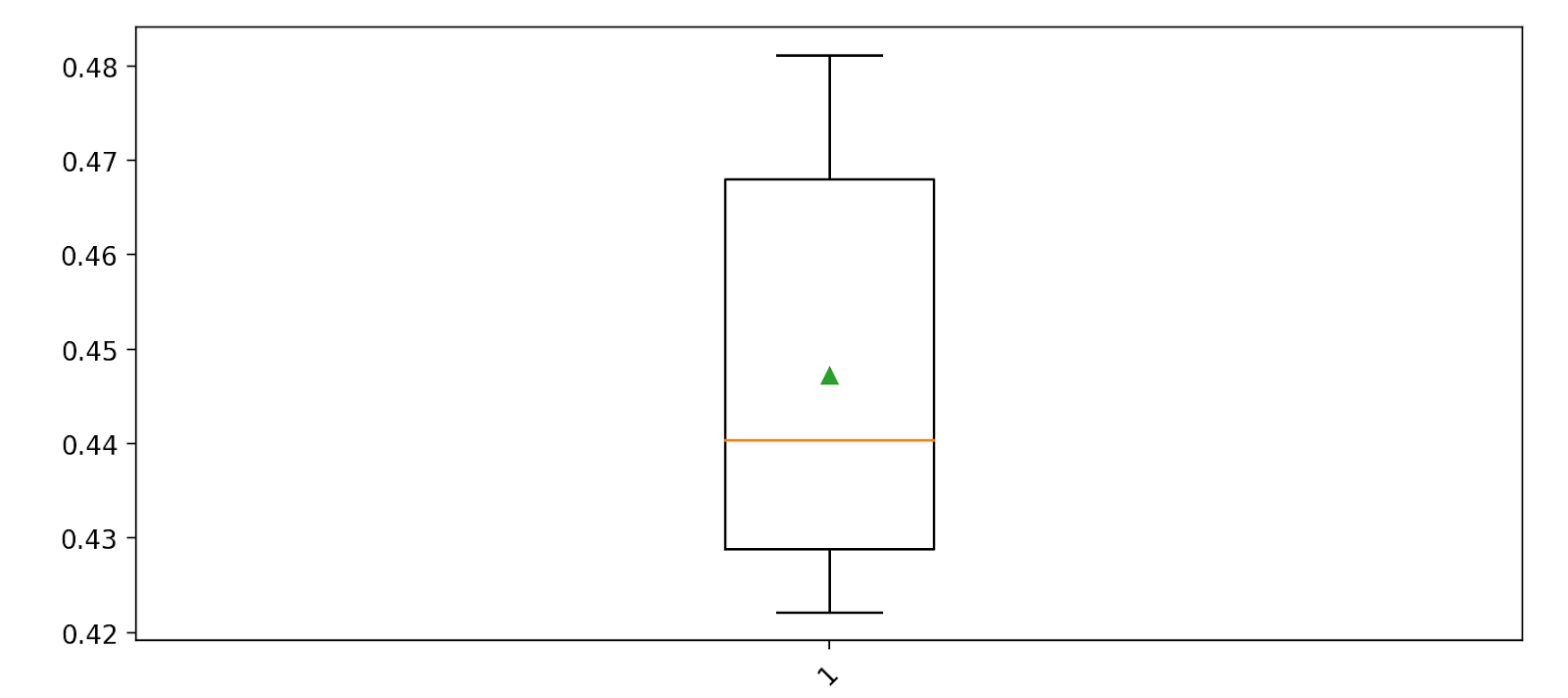
**נספח ב'**

boxplot עבור אלגוריתם random forest שמכיל 14 מודלים שונים.

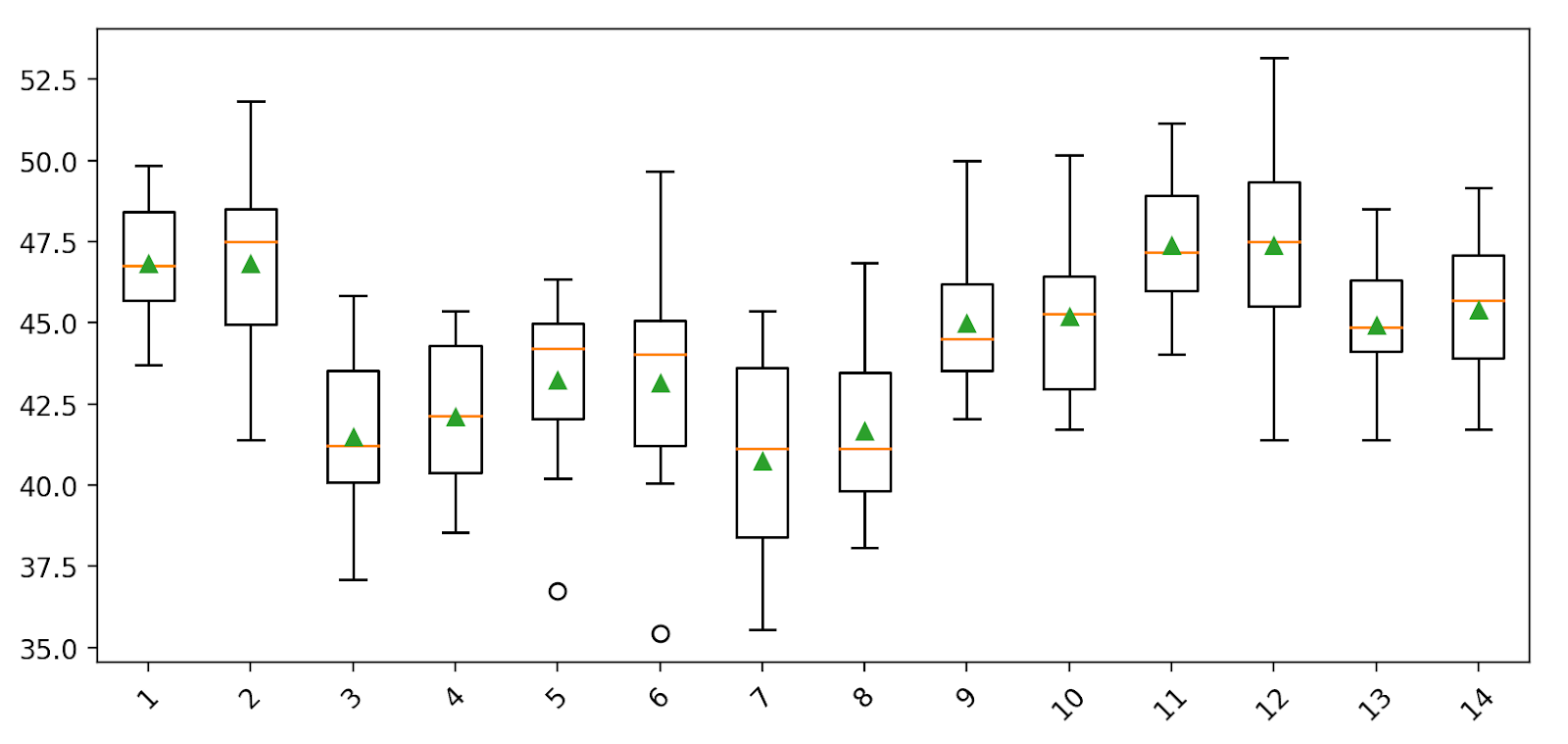
עבור f-measure

****

עבור  accuracy score

****

עבור  k fold cross validation



**נספח ג'**

|  |  |
| --- | --- |
| **פיצ׳ר** | **סימון** |
| **מס׳ ניצחונות בחמשת הסיבובים האחרונים של קבוצת הבית** | **x3** |
| **מס׳ ניצחונות בחמשת הסיבובים האחרונים של קבוצת החוץ** | **x4** |
| מס׳ תיקו בחמשת הסיבובים האחרונים של קבוצת הבית | x5 |
| מס׳ תיקו בחמשת הסיבובים האחרונים של קבוצת החוץ | x6 |
| מס׳ הפסדים בחמשת הסיבובים האחרונים של קבוצת הבית | x7 |
| מס׳ הפסדים בחמשת הסיבובים האחרונים של קבוצת החוץ | x8 |
| מס׳ ניצחונות בחמשת משחקי הבית האחרונים של קבוצת הבית | x11 |
| **מס׳ תיקו בחמשת משחקי הבית האחרונים של קבוצת הבית** | **x12** |
| **מס׳ הפסדים בחמשת משחקי הבית האחרונים של קבוצת הבית** | **x13** |
| מס׳ ניצחונות בחמשת משחקי החוץ האחרונים של קבוצת החוץ | x14 |
| **מס׳ תיקו בחמשת משחקי החוץ האחרונים של קבוצת החוץ** | **x15** |
| מס׳ הפסדים בחמשת משחקי החוץ האחרונים של קבוצת החוץ | x16 |

פיצ'רים בהם השתמשנו עבור כל מודל

טבלה 1: פירוט פיצ'רים עבור מודל GP אשר התאימו לנתונים מה dataset הנתון – מודגשים הפיצ'רים שנבחרו על ידי מודל ה feature selection

|  |  |
| --- | --- |
| **הסבר** | **מדד** |
| יכולות שוער – מיקום ורפלקסים | gk\_reflexes + gk\_positioning |
| יכולות שוער – בעיטה | gk\_kicking |
| יכולות שוער – יכולת צלילה ושליטה | gk\_handling+gk\_diving |
| יכולות שחקן - תיקול בגלישה (גליץ׳) | sliding\_tackle |
| יכולות שחקן - תיקול בעמידה ושמירה אישית | standing\_tackle'+ marking |
| יכולות שחקן - בעיטות עונשין, ראיית משחק ומיקום | Penalties+vision+positioning |
| יכולת שחקן – יירוטים | Interceptions |
| יכולת שחקן – אגרסיביות ובעיטות ארוכות | aggression+long\_shots |
| יכולת שחקן – חוזק | Strength |
| יכולת שחקן – תיקול בעמידה וניתור | standing\_tackle + jumping |
| יכולת שחקן – עוצמת בעיטה וחוזק | shot\_power+strength |
| יכולת שחקן – יציבות ויכולת תגובה | balance + reactions |
| יכולת שחקן – גמישות וזריזות | Agility |
| יכולת שחקן – מהירות ספרינט והאצה | sprint\_speed + acceleration |
| יכולת שחקן – שליטה בכדור ומסירות ארוכות | ball\_control +long\_passing |
| יכולת שחקן – דיוק בעיטות חופשיות | free\_kick\_accuracy |
| יכולת שחקן – סיבוב הכדור בבעיטות חופשיות | Curve |
| יכולת שחקן – כדרור | Dribbling |
| יכולת שחקן – הרמה ובעיטת וולה (מהאוויר) | Volley+ crossing |
| יכולת שחקן – סיום (הבקעת גול) ודיוק בנגיחה | heading\_accuracy+ finishing |
| יכולת שחקן – הרמה ופוטנציאל | crossing+potential |
| יכולת שחקן – דירוג כללי | overall\_rating |

טבלה 2: פירוט פיצ'רים עבור מודל SVM אשר התאימו לנתונים מה dataset הנתון

**במודל SVM** השתמשנו בכל הפיצ'רים מעלה **עבור כל שחקן בקבוצה**. מודל ה feature selection בחר שחקנים עבור כל קבוצה. עבור קבוצת הבית את השחקנים: 2, 6, 7, 8, 10, 11. עבור הקבוצה האורחת את השחקנים: 2, 6, 7, 8, 11.

במודל **random forest** השתמשנו בכל הפיצ'רים המצוינים מעלה משני המאמרים. בנוסף אליהם השתמשנו בפיצ'רים שתוארו בחלק 4.1. מודל ה feature selection בחר את הפיצ'רים הבאים:

**מטבלה 1:**

x3, x4, x5, x6, x7, x11, x12, x14, x15

**שחקנים עם פיצ'רים מטבלה 2:**

עבור קבוצת הבית, מספרי השחקנים: 1, 5, 7, 10.

עבור הקבוצה האורחת: 6, 7, 8, 11

**מהפיצ'רים החדשים** שיצרנו נבחר הפיצ'ר המתאר את ההגנה של קבוצת הבית.