אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev

הפקולטה למדעי ההנדסה
המחלקה להנדסת מערכות תוכנה ומידע

קורס: הכנה לפרויקט

נושא:
מערכות חיזוי לתוצאות משחקי כדורגל

מרצה: ד"ר ארנון שטורם
מנחה: אנטולי שוסטרמן
מגישים:
רותם מיארה 301886776
מיטל יניב 307938969
שי ארץ קדושה 203276258
גל בוזגלו 203886528
יסמין אברהם 208063453

1. מבוא

הכדורגל הוא ענף הספורט מהפופולאריים ביותר בעולם. בכל משחק מתחרות שתי קבוצות, כאשר מטרתה של כל קבוצה היא להכניס את הכדור לתוך השער של הקבוצה היריבה- תחריש המכונה 'גול' . הקבוצה המנצחת היא זו שהכניסה את מספר הגולים הגדול ביותר. כאשר שתי הקבוצות הכניסו מספר גולים זהה המשחק נגמר בתיקו.

על מנת שנוכל לחזות תוצאה של משחק כדורגל, לא נוכל להשתמש בתוכנה פשוטה או בגורם אנושי בגלל המאפיינים הרבים הקיימים למשחק- תוכנה פשוטה לא תוכל להריץ כמות גדולה של נתונים והגורם האנושי יוכל לנחש על סמך ידע קודם שלו אך זה לא יהיה על כמות נתונים גדולה והדיוק לא יהיה מרבי. כדי שנוכל להשיג דיוק רב בחיזוי התוצאות נעשה שימוש בלמידת מכונה.

למידת מכונה (Machine Learning) הוא תת-תחום במדעי המחשב המשלב בתוכו את תחומי הבינה מלאכותית, הסטטיסטיקה והאופטימיזציה. זוהי טכניקה לניתוח נתונים ע"י למידה מדוגמאות, ניסיון ומינימום התערבות של גורמים אנושיים. המטרה המרכזית של התחום הוא לבצע טיפול ממחושב בנתונים רבים מהעולם האמיתי עבור בעיה מסוימת, כשלא ניתן לפתור זאת באמצעות תוכנת מחשב רגילה[1]. השימוש בלמידת מכונה בא לידי ביטוי בעיקר בחיזוי תרחישים מסוימים וזיהוי תבניות מתוך מאגר נתונים קיים.

במחקר שלנו נשתמש בלמידת מכונה על מנת לסוג תוצאות של משחקי כדורגל .

2. סקירה ספרותית

חיזוי תוצאות משחקי כדורגל מעסיקה רבים החל ממעריצים ועד אנשי עסקים גדולים. בנוסף בעיה זו זוכה לעניין גם בקרב חוקרי למידת מכונה בגלל היותה בעיה קשה, מכיוון שתוצאות המשחקים תלויות בהרבה מאוד פרמטרים כגון: מורל הקבוצה, כישורי השחקנים, מיקום המשחק ועוד[2].

בחלק זה נסקר בעיות דומות בתחום חיזוי תוצאות הכדורגל איתן ניסו חוקרים להתמודד בעזרת למידת מכונה. נציג את סוג הנתונים עליהם עבדו, מודלים שונים לחיזוי, הפיצ'רים בהם השתמשו, שיטות שונות לאימון המודל ומסקנותיהם.

2.1. הבעיות השונות איתן ניסו להתמודד ממערכות אחרות

במהלך מחקרים רבים התגלו בעיות או כשלים במערכות שנחקרו או חוסר נתונים מסוים העלולים להביא לדיוק לא מיטבי של המערכת ולכן חוקרים רבים התבססו על מחקרי עבר והמשיכו לחקור על בסיס מחקרים אחרים לשפרם ולהרחיבם.
בחלק זה נסקור את הבעיות ממערכות אחרות איתם ניסו להתמודד חוקרים רבים במהלך השנים. ראשית, ניסו לפתח מודל שיעזור לחזות את התוצאות של משחקי כדורגל (ניצחון, הפסד ותיקו) תוך הבנה מהם המשתנים שמשפיעים על תוצאות המשחקים[6]. ניתן לראות כי החוקרים נתקלו בבעיה המתבססת על הסתמכות על מעט מדיי נתונים(התחלת עונה עד מחזור 17) מה שיוצר טעות גדולה כשרוצים לחזות את המשחקים הראשונים בעונה, מתן חיזוי עבור קבוצה אחת בלבד וצורך ב-input נוסף עבור כל משחק [3].

שנית, שימוש בקומבינציות של if-then המבוססות על סטטיסטיקה נטו כמו למשל ההסתברות שקבוצת בית תנצח שווה להפרש הנקודות של הקבוצה בליגה בערך המוחלט כפול 0.53 ועוד 0.448. המערכות שנבדקו מתבססות על שיטות סטטיסטיות כאשר קבוצה A משחקת נגד קבוצה B כאשר קבוצה A היא המארחת [4].
ובנוסף, ניסו לחזות את תוצאות משחקי כדורגל לצורך שימוש אנושי ולא תעשייתי (כמו בתעשיית ההימורים או לשיפור היכולת של קבוצות) בעקבות ביצועיו של פול התמנון שחזה את התוצאות של המשחקים במונדיאל 2010 [5].

2.2. Datasets

על מנת לחזות את תוצאות המשחקים היה על החוקרים למצוא נתונים מתאימים שעליהם הם יערכו את הניסויים. בפרק זה נסקור את ה Data sets שהשתמשו במחקרים רבים שעליהם התבסס חיזוי תוצאות משחקי כדורגל.
מרבית הנתונים למחקרים נלקחו לרוב מן הליגה האנגלית, הליגה הספרדית, האיטלקית והטורקית.
הליגה האנגלית: סט האימון כולל תוצאות של 10 עונות (מעונה 2003-2002 עד עונה 2012-2011) וסט הבדיקה כולל 2 עונות (2014-2013). בכל עונה השתתפו 20 קבוצות, ובכל עונה כל קבוצה שיחקה פעמים מול קבוצה אחרת (בית וחוץ). סה"כ בקובץ האימון היו 3800 משחקים ובקובץ הבדיקה 760 משחקים.
עבור כל משחק המידע שנלקח הוא: הקבוצות ששיחקו, מי הקבוצה האורחת ומי המארחת, תוצאת המשחק, זהות המנצח, מספר הגולים שכל קבוצה הבקיעה [5].

הליגה הספרדית: 5 שנים אחורה נתוני משחק הכוללים מס' שערים, מס' בעיטות במשחק, מס' בעיטות למסגרת במשחק כדי לנסות לנבא תוצאה סופית. נתונים של כל קבוצה נגד כל קבוצה בליגה – מאזנים של

ניצחונות תיקו והפסדים בין הקבוצות כדי לנסות לנבא כיצד המשחק ייגמר. נתונים על שחקנים מהמשחק פיפ"א 18 כדי לנסות לנבא מי יהיו השחקנים שיבקיעו במשחק. נתונים על הרכבי הקבוצות והטקטיקות שלהם ע"מ לנסות לנבא את הרכב הקבוצה למשחק הקרוב [3].

הליגה האיטלקית: שימוש בתוצאות המשחקים של הסיבוב הראשון בליגה האיטלקית הבכירה כאשר הסיבוב השני ישמש כסט אימון. חוסר שימוש בתוצאות 6 המחזורים הראשונים מפאת גורמים שמשפיעים מאוד בתחילת העונה ויכולות לעוות את אימון המודל. והכנת הדאטה [4].

הליגה הטורקית: הנתונים כללו מידע 3060 משחקים עבור 33 קבוצות וכל המידע הזה נפרש על 10 עונות בין השנים 2007-2017. לאחר מחשבה החליטו שמכיוון ובסוף העונה ישנה תחרות רק בין הקבוצות במקומות הראשונים שיש להם רצון לנצח כמו גם אלו שבתחתית ואלו שבאמצע הטבלה אין להם סיבה תחרותית לשחק ולכן הוחלט להתעלם מ-4 המשחקים האחרונים בכל עונה. בנוסף בעקבות כך שב- 5 השבועות הראשונים לליגה עדיין אפשר לבצע העברות של שחקנים בין קבוצות ולשנו את הרכב הקבוצה הוחלט להתעלם מ-5 השבועות הראשונים. כלומר המשחקים שהתחשבו בהם היו משבוע 6 עד שבוע 30 כלומר, במקור היו 34 שבועות בסך הכל [6].

2.3. שיטות לחיזוי

במשך השנים פותחו מגוון רחב של אלגוריתמים לחיזוי וזיהוי תבניות, הנפוצות שבהן: ANN , SVM, Naïve Bayes ועוד. בחלק זה נסקור מספר אלגוריתמים הממפים את הקלט לאחת מקבוצות הידועות מראש.
חוקרים רבים משתמשים באלגוריתמים אלו ומסווגים את המשחק הנתון כקלט לאחת משלושת הקבוצות הבאות: ניצחון, הפסד או תיקו.

2.3.1. Bayesian Networks

טכניקה לסיווג פשוטה ויעילה העובדת טוב על נתונים רועשים. זהו מודל הסתברותי החוזה את הקלט עפ"י נתונים קודמים. במאמר [2] שנכתב עבור חיזוי תוצאות משחקי כדורגל ומבסס את המודל שבנה על רשתות בייסאניות נרשם דיוק של כ-92% בניבוי התוצאות. דיוק זה הוא בין הגבוהים שמצאנו בין המאמרים, אך חיסרון אחד בולט הוא כי נתוני המודל איתם עבדו החוקרים משתייכים לקבוצה אחת בלבד.

2.3.2. Decision Trees

משפחת אלגוריתמים לסיווג רשומות לקבוצות ידועות מראש. טכניקות הקיימות באלגוריתמים אלו הם: Random Forestו- Gradient Boosting Trees. בסקירה נפרט על טכניקת Random forest שבה מצאנו תוצאות טובות במספר ניסויים שהתקיימו בעבר.

2.3.2.1 Random Forest

אלגוריתם ממשפחת אלגוריתמי עצי החלטה המשמש לסיווג, רגרסיה ועוד. שיטה זו בונה מספר עצי החלטה תוך כדי אימון המודל ומסווגת וחוזה באמצעות עצים אלו. נוכחנו לראות כי שיטה זו נפוצה מאוד בקרב חוקרים המנסים לחזות תוצאות למשחקי כדורגל ואף מגיעה לאחוזי דיוק גבוהים. כך למשל במחקר [7] הבודק שישה שיטות שונות לחיזוי משחקי כדורגל, נמצא כי שיטת Random Forest היא השיטה השנייה המביאה את הדיוק הגבוה ביותר (אחרי ANN). מחקר חיזוי נוסף מהשנה האחרונה [6] המשווה בין שמונה מודלים שונים מצא כי אלגוריתם זה הביא לדיוק המרבי של כ- 74.6% על סט הנתונים שלו.

2.3.3. SVM

טכניקה נפוצה המשמשת לסיווג ורגרסיה, ובה מייצגים את רשומות האימון כווקטורים במרחב הלינארי. במחקר [5] המשווה בין מודלים שונים עבור משחקים בליגה האנגלית ובנוסף משווה בין 3 שיטות שונות של SVM מצא כי המודלים של שיטת הSVM מביאים את הדיוק הגבוה ביותר עבור נתונים אלו.

2.3.4. KNN

בדומה לאלגוריתם SVM גם באלגוריתם זה התצפיות מיוצגות במרחב הווקטורי, אך הסיווג עבור התצפית החדשה נקבע עפ"י K השכנים הקרובים. אלגוריתם זה הופיע במחקרים עבור ניסוי לחיזוי תוצאות כדורגל [6] [7] נמצא כי מודל זה לא מביא תוצאות נמוכות אך הוא אינו בין המודלים עם הדיוק הגבוה עבור datasetsשונים.

2.3.5. Logistic Regression

מודל סטטיסטי המתבסס על רגרסיה- כלומר הקשר בין משתנים מסבירים למשתנה המוסבר. שיטה זו לא הופיעה בהרבה מהניסויים שבוצעו, אך המחקר [3] שיטה זו הביאה את הדיוק הגבוה ביותר מבין כל המודלים שנבדקו.

2.4. אימון המודל

אימון המודל נעשה בשתי שלבים עיקריים: הראשון הוא בחירת סט האימון והשני הוא הרצת סט האימון במודל. לשלב בחירת סט האימון והרצתו במודל קיימות מספר גישות. במחקר [4] ביצעו החוקרים את האימון באיטרציות וחלוקת הנתונים למספר קבוצות. באיטרציה הראשונה נעשה על המודל אימון בלבד, והחל מהשלב השני המערכת מנסה לנבא תוצאות ומקבלת תיקון כאשר הסיבוב מסתיים. גישה נפוצה יותר לאימון המודל וחלוקת הנתונים היא K-Fold cross validation עליה נפרט בהמשך. שפת התכנות בה השתמשו מספר חוקרים [3][5] היא Python המכילה מספר ספריות המשמשות ללמידת מכונה כגון: Sci-Kit Learn, Pandas ועוד.

2.4.1. K-Fold Cross Validation

בשיטה זו קובעים את פרמטר K ומחלקים את הנתונים ל K קבוצות באופן רנדומאלי. כל מודל חיזוי אומן ונבדק K פעמים, כאשר בכל פעם סט האימון והלמידה מוגדר על K-1 קבוצות, והקבוצה הנותרת משמשת לבדיקת המודל . במספר מחקרים שבדקנו [6] [3] נבחר לבסוף K=10 לאחר שזהו המספר שהביא תוצאות אופטימליות.

2.5. בחירת פיצ'רים למודל

לאחר איסוף הנתונים, דיוק ומהירות התחזיות יהיה תלוי בבחירה נכונה של הפיצ'רים המתאימים ביותר. במערכות קיימות ניתחו גישות שונות בהן נעשה שימוש בחיזוי תוצאות משחקי הכדורגל, סטטיסטיקה ולמידת מכונה.

כך למשל, במאמר שבדק על הליגה הספרדית [3] השתמשו בשיטה שהוצעה על ידי צוות מומחים ממומבאי הודו, כדי לבחור פיצ'רים סופיים לחיזוי. הפיצ'רים שנבחרו היו - מזהה קבוצת בית, מס' בעיטות לקבוצת בית, מס' בעיטות למסגרת לקב' בית, מס' קרנות לקב' בית, מס' כרטיסים צהובים לקב' בית, מס' כרטיסים אדומים לקב' בית – ובנוסף כל התכונות הללו גם לקב' חוץ. כמו כן, אחוז הניצחונות של קב' הבית מול קב' החוץ.

לעומת זאת, במאמר שבדק על הליגה האנגלית [5] הפיצ'רים שנבחרו למודל הם- האם הקבוצה היא מארחת או אורחת ופיצ'ר נוסף אם הקבוצה היא ברצף ניצחונות.

עבור רצף ניצחונות בחרו להתעלם מ-7 המשחקים הראשונים של העונה, כלומר יתחשבו בנתונים שלהם אבל במהלכם לא יוגדר רצף עבור קבוצה אלא מהמשחק השמיני והלאה. ובנוסף קבעו שרצף הוא עד 7 משחקים בכל המודלים (לאחר בדיקת אופטימליות שנעשתה).

נוכל להרחיב את התכונות לדיוק מירבי של המודל. בין הפיצרים ניתן למצוא:

פיצ'רים לזיהוי כמו- איזו עונה, איזה שבוע ומי הקבוצה. פיצ'רים על הקבוצה עצמה כמו: מספר הזרים בקבוצה, הערך כלכלי של הקבוצה, גיל המאמן, כמה זמן הקבוצה בליגה וכו'. פיצ'רים על כל משחק כמו: שעת משחק, משחק ביץ / חוץ, מס' מסירות, מס' בעיטות לשער, מס' כרטיסים צהובים, מס' כרטיסים אדומים, ממוצע גיל השחקים שעל המגרש, מס' קרנות, מס' נבדלים, מס' פאולים, איבודי כדור, מס' גולים שהכניסו, מס' גולים שספגו ותוצאה (ניצחון ,תיקו ,הפסד). במאמר שבחן את הליגה הטורקית [6] הוחלט שמכל הפיצ'רים הללו התוצאה תשמש כמסווג לחיזוי והפרש השערים לתוצאות מודל רגרסיה.

2.6. תוצאות החיזוי

בליגה הספרדית, אלו הן תוצאות החיזוי לפי הניסויים שבוצעו בשיטות השונות:

Logistic Regression – דיוק של 71.63%
Random Forest – דיוק של 69.9%
Linear SVM – דיוק של 66.95%
Naïve Bayes – דיוק של 63.57%

במאמר אחר שבדק חיזוי תוצאות טורניר בליגת האלופות [7] בעזרת אותן שיטות וכן שיטות נוספות תוצאות החיזוי היו:

Logistic Regression – דיוק של 48%-68.8%
Random Forest – דיוק של 50%-65.6%
Linear SVM – דיוק של 66.95%
Naïve Bayes – דיוק של 47%-56%
KNN – דיוק של 53%-62.5%

בליגה האנגלית:

1. תוצאות stochastic gradient descent algorithm - עבור שני סיווגים (ניצחון והפסד) הוא: תוצאת אחוז השגיאה עבור סט הבדיקה- 0.38 ועבור סט האימון תוצאה אחוז השגיאה של 0.34. תוצאות של

one vs all stochastic gradient descent עבור שלושה סיווגים (ניצחון תיקו והפסד הם: עבור סט הבדיקה תוצאת שגיאה של 0.6 ועבור סט האימון אחוז תוצאת שגיאה של 0.39.

2. תוצאות של Naïve bayes עבור שלושה סיווגים (ניצחון, תיקו והפסד) הם: עבור מודל gaussian של naïve bayes התקבל עבור סט האימון אחוז שגיאה של 0.52 ועבור סט הבדיקה אחוז שגיאה של 0.56. עבור המודל multinomial של naïve bayes התקבל עבור סט האימון אחוז שגיאה של 0.51 ועבור סט הבדיקה אחוז שגיאה של 0.54.

3. תוצאות של hidden markov model עבור שלושה סיווגים (ניצחון, תיקו והפסד) הם: עבור סט האימון התקבל אחוז שגיאה של 0.52 ועבור סט הבדיקה אחוז שגיאה של 0.56.

בליגה הטורקית סיכום תוצאות החיזוי מוצגות ב-confusion matrix עבור כל מודל, כלומר תוצאת סיווג והסיווג בפועל (ניצחון הפסד, תיקו) עבור כל צירוף קיים TP,TN,FP,FN.

נסכם כן את אחוז הדיוק של כל מודל:

1. Naive Bayes- אחוז דיוק של 51.90%.
2. Decision Trees (CART)- אחוז דיוק של 59.70%.
3. Neural networks- אחוז דיוק של 55.91%.
4. SVM- אחוז דיוק של 58.80%.
5. KNN (K=5)- אחוז דיוק של 58.80%.
6. Gradient Boosted Tress- אחוז דיוק של אחוז דיוק של 58.80%.
6. Gradient Boosted Tress- אחוז דיוק של 74.50%.
7. Random Forest- אחוז דיוק של 74.60%.

נעשתה בדיקה והורידו את כמות הסיווגים לשניים במקום ניצחון, תיקו והפסד ל- יש נקודות/ אין נקודות.

התוצאות עבור שני המודלים המובילים היו:

1. Random Forest- 86.3% אחוז דיוק.
2. Gradient Boosted Tree - אחוז דיוק של 86.4%.

הטבלה הבאה מסכמת את המאמרים העיקריים בהם השתמשנו בסקירת הספרות.

מאמר	סט נתונים - dataset	מודלים שהשתמשו בהם	פיצ'רים נבחרים	הדיוק הכי גבוה- באיזה מודל זה התקבל
Predicting Soccer Match Results in the English Premier League	הליגה האנגלית. 10 עונות . בכל עונה 20 קבוצות. סה"כ 3800 משחקים הליגה האנגלית. 2 עונות. בכל עונה 20 קבוצות. סה"כ 760 משחקים	1. Baseline 2. Naïve Bayes 3. Hidden Markov Model 4. SVM 5. Random Forest 6. One vs All stochastic gradient decent	האם הקבוצה היא מארחת או אורחת ופיצ'ר נוסף אם הקבוצה היא ברצף ניצחונות. עבור רצף ניצחונות בחרו להתעלם מ-7 המשחקים הראשונים של העונה, כלומר יתחשבו בנתונים שלהם אבל במהלכם לא יוגדר רצף עבור קבוצה אלא מהמשחק השמיני והלאה. ובנוסף קבעו שרצף הוא עד 7 משחקים בכל המודלים	המודלים שהראו תוצאות הכי טובות הם: SVM עם gaussian kernel (RBF), SVM עם gram kernel (linear kernel), random forest ו- one vs all SGD model
Predicting the Outcome of a Football Game: A Comparative Analysis of Single and Ensemble Analytics Methods	הנתונים למחקר נאספו על הליגה הטורקית. הנתונים כללו מידע 3060 משחקים עבור 33 קבוצות וכל המידע הזה נפרש על 10 עונות בין השנים 2007-2017. לאחר מחשבה החליטו שמכיוון בסוף העונה ישנה תחרות רק בין הקבוצות במקומות הראשונים שיש להם רצון לנצח כנ"ל אלו שבתחתית ואלו שבאמצע הטבלה אין להם סיבה תחרותית לשחק הוחלט להתעלם מ-4 המשחקים האחרונים בכל עונה. בנוסף בעקבות כך שב- 5 השבועות הראשונים לליגה עדיין אפשר לבצע העברות של שחקנים בין קבוצות ולשנו את הרכב הקבוצה הוחלט להתעלם מ-5 השבועות הראשונים. כלומר המשחקים שהתחשבו בהם היו משבוע 6 עד שבוע 30 (במקור היו 34 שבועות סה"כ).	על ה-data המקורי עם ערכים חסרים השתמשו במודלים שיכולים להתמודד עם ערכים חסרים: 1. Naïve bayes 2. Decision Trees Ensembled Methods הכוללים שתי טכניקות: 3. Gradient Boosting Trees 4. Random Forest. 5. Neural Networks hidden 2 layers, 10 nodes per layer) 6. SVM (with RBF) 7. KNN כאשר K=5.	פיצ'רים לזיהוי כמו: עונה, שבוע והקבוצה. פיצ'רים על הקבוצה עצמה כמו: מספר הזרים בקבוצה, הערך כלכלי של הקבוצה, גיל המאמן, כמה זמן הקבוצה בליגה וכו'. ופיצ'רים על כל משחק כמו: שעת משחק, משחק ביץ / חוץ, מס' מסירות, מס' בעיטות לשער, מס' כרטיסים צהובים, מס' כרטיסים אדומים, ממוצע גיל השחקים שעל המגרש, מס' קרנות, מס' נבדלים, מס' פאולים, איבודי כדור, מס' גולים שהכניסו, מס' גולים שספגו ותוצאה (ניצחון ,תיקו ,הפסד).	אחוז הדיוק הכי גבוה שהושג עבור שלושה סיווגים (ניצחון, תיקו והפסד) היה 74% ו- 86% עבור שני סיווגים (יש/ אין נקודות). Random Forest- 86.3% אחוז דיוק. Gradient Boosted Tree- אחוז דיוק של 86.4%.

Prediction of Football Match Score and Decision Making Process	נתונים על הליגה הספרדית הבכירה 5 שנים אחורה כמו למשל נתוני משחק(מס' שערים, מס' בעיטות במשחק, מס' בעיטות למסגרת במשחק וכו') כדי לנסות לנבא תוצאה סופית. נתונים של כל קבוצה נגד כל קבוצה בליגה – מאזנים של ניצחונות תיקו והפסדים בין הקבוצות כדי לנסות לנבא כיצד המשחק ייגמר. נתונים על שחקנים מהמשחק פיפ"א 18 כדי לנסות לנבא מי יהיו השחקנים שיבקיעו במשחק. נתונים על הרכבי הקבוצות והטקטיקות שלהם ע"מ לנסות לנבא את הרכב הקבוצה למשחק הקרוב.	השתמשו במס' שיטות והשוו ביניהם לאחר מכן בתוצאות, השיטות: Logistic Regression Random Forest ANN Linear SVM Naïve Bayes	מזהה קב' בית, מס' בעיטות לקב' בית, מס' בעיטות למסגרת לקב' בית, מס' קרנות לקב' בית, מס' כרטיסים צהובים לקב' בית, מס' כרטיסים אדומים לקב' בית – בנוסף כל התכונות הללו גם לקב' חוץ. אחוז הניצחונות של קב' הבית מול קב' החוץ.	Logistic Regression – דיוק של 71.63%
A Comparative Study on Neural Network based Soccer Result Prediction	שימוש בתוצאות המשחקים של הסיבוב הראשון בליגה האיטלקית הבכירה(כאשר הסיבוב השני ישמש כסט אימון). חוסר שימוש בתוצאות 6 המחזורים הראשונים מפאת גורמים שמשפיעים מאוד בתחילת העונה ויכולות לעוות את אימון המודל. הכנת הדאטה עבור שתי שיטות שישמשו אותנו בהמשך: Learning vector quantization A Learning vector quantization B	שיטות מבוססות רשתות נוירונים. השיטות יודעות לחזות האם יהיה ניצחון לקבוצה המארחת, תיקו או ניצחון לקבוצה האורחת.	Learning vector quantization A – רשת נוירונים אשר כקלט מקבלת 4 תכונות: רייטינג בית של קבוצה מארחת רייטינג חוץ של קבוצה מארחת רייטינג בית של קבוצה אורחת. רייטינג חוץ של קבוצה אורחת. Learning vector quantization B – רשת נוירונים אשר כקלט מקבלת 2 תכונות: רייטינג בית של קבוצה מארחת רייטינג חוץ של קבוצה אורחת.	Learning vector quantization B – דיוק של 53.25% LVQ B בעלת דיוק גבוה יותר, אך זה נובע בגלל העובדה שאחוזי הניבוי שלה עבור ניצחון לקבוצת הבית גבוהים, אחוז זיהוי התיקו/ ניצחון לקב' החוץ נמוך יותר משיטה LVQ A.
Predicting football scores using machine learning techniques	שימוש ב2 סוגי DATASET שונים והשוואה ביניהם. הראשון מבוסס מידע מקדים על הקבוצה לפי הפיצ'רים הנבחרים על בסיס משחקי שלב הבתים. השני מבוסס מידע מקדים ודעת מומחי כדורגל.	השיטות שנבחרו: Naïve Bayes, Bayesian Networks, LogitBoost, KNN, Random Forest, ANN	1. תוצאות 6 המשחקים האחרונים. 2. תוצאת המפגש האחרון בין הקבוצות. 3. מיקום בטבלה. 4. מס' שחקני סגל ראשון פצועים. 5. ממוצע כיבוש וספיגת שערים פר משחק.	שיטת הANN הביאה לתוצאות החיזוי הטובות ביותר: ANN – 54%-68.8% (אחוזי דיוק)

לסיכום, במאמר Prediction of Football Match Score and Decision Making Process, הפיצ'ר אחוז הניצחונות של קבוצת הבית מול קבוצת החוץ העלה את הדיוק ב-8%. ביחס לעבודות אחרות בנושא עדיין יש מקום רב לשיפור גם בניסיון לחזות מי תנצח וגם בניסיון לחזות מי יבקיע.

במאמר A Comparative Study on Neural Network based Soccer Result Prediction, Learning vector quantization B בעלת דיוק גבוה יותר, אך זה נובע בגלל העובדה שאחוזי הניבוי שלה עבור ניצחון לקבוצת הבית גבוהים, אחוז זיהוי התיקו/ ניצחון לקבוצת החוץ נמוך יותר משיטה Learning vector quantization A.

הוכח כי תוצאות השיטות הפועלות על בסיס ANN טובות יותר מאלו המתבססות על סטטיסטיקות בלבד. זאת ועוד נמצא במחקר כי המידע אמור לעבור טרנספורמציה נכונה כדי שרשת הנוירונים תוכל להשתמש בו בצורה הטובה ביותר ולספק תוצאות טובות, אין צורך לספק הרבה מאוד מידע כקלט לרשת כי זהו דבר שיכול להוביל לאי התכנסות ולתוצאה לא חד משמעית. בנוסף, עדכון המערכת כאשר מחזור נגמר משפיע מהותית על תוצאות המערכת.

במאמר Predicting Soccer Match Results in the English Premier League, המודלים שהראו תוצאות הכי טובות הם: SVM עם gaussian kernel (RBF), SVM עם gram kernel (linear kernel), random forest ו-one vs all SGD model (הראו את זה באמצעות confusion matrix, ועקומות ROC עבור תוצאות חיזוי ותוצאות ממשיות של ניצחון תיקו והפסד עבור כל מודל). אומנם תוצאת השגיאה הכי נמוכה הייתה של one vs all SGD model אבל הוא היה תלוי מאוד בגודל סט האימון ועבור כל שינוי בגדול התוצאות השתנו משמעותית. דבר נוסף one vs all SGD model וגם linear SVM שניהם חזו בצורה נמוכה מאוד תוצאות של תיקו. ולכן השימוש במודלים אלה לניבוי בעולם האמיתי עבור הימורים הוא מאוד בעייתי. יש לציין שגם במודלים האחרים החיזוי של תיקו היה נמוך אבל לא נמוך מאוד כמו שני המודלים הללו. לסיכום המסקנה העיקרית שהתוצאות שהתקבלו הם טובות עבור ניתוח אנושי אבל לא מספיק מדויקות עבור שימוש בתעשייה עבור הימורים.

במאמר Predicting the Outcome of a Football Game: A Comparative Analysis of Single and Ensemble Analytics Methods, אחוז הדיוק הכי גבוה שהושג עבור שלושה סיווגים (ניצחון, תיקו והפסד) היה 74% ו-86% עבור שני סיווגים (יש/ אין נקודות). ייתכן שהתוצאות הנמוכות עבור חלק מהמודלים נבעו מכך שהיו לא מעט ערכים חסרים ולכן אמינות המודל נפגעה. ייתכן ואם היו משלימים את הנתונים בנתונים נכונים היו מתקבלות תוצאות הרבה יותר טובות בחלק מהמודלים. היכולת לסווג בצורה נכונה את הסיווג "תיקו" בכל המודלים הייתה יחסית קשה. בנוסף בכל המודלים שמבוססים על עצים עשו

בדיקה מי הפיצ'ר המשמעותי יותר שמביא תוצאות מדויקות יותר על סמך הפיצול הראשון בעץ וניתן היה לראות שהפיצ'ר של כמות הנקודות שהושגו היה המשמעותי ביותר, אחריו מיקום הקבוצה בטבלה ולאחר מכן האם הקבוצה הייתה 10 שנים ברציפות בליגה.
המסקנות שהסקנו הן שישנו מחסור משמעותי של מחקרים בתחום חיזוי תוצאות משחקי כדורגל ולכן יש לעשות מחקר רב נוסף על ליגות שונות, קלט שונה, מודלים ושיטות שבנויות אחרת כדי להשיג מערכות חיזוי עם תוצאות טובות, תוצאות רבות שהתקבלו היו נמוכות עבור חלק מהמודלים כנראה עקב כך שהיו לא מעט ערכים חסרים ולכן אמינות המודל נפגעה וייתכן שאם היו משלימים את הנתונים בנתונים נכונים היו מתקבלות תוצאות הרבה יותר טובות בחלק מהמודלים.

3. תיאור המערכת והאלגוריתם

3.1. התייחסות לדאטה

לצורך בניית המודל שעבורו נתבקשנו לייצר פיצ'רים מסוימים ע"מ לנסות ולחזות תוצאות משחקים עתידיות השתמשנו בבסיס נתונים שניתן לנו מסוג sql lite שמכיל מידע על משחקים ממס' ליגות שונות במשך 8 שנים, פרטים על יכולות אישיות של שחקנים ששיחקו באותם משחקים ויחסי הימורים של אתרים שונים שניתנו לכל משחק. בהתאם לסקירה שביצענו עבור ניתוח הפיצ'רים המועילים ביותר לחיזוי, בחרנו להשתמש בשתי טבלאות עיקריות בבסיס הנתונים שהן טבלת המשחקים וטבלת השחקנים (המכילה יכולות אישיות של כל שחקן).
בטבלת המשחקים בחרנו לקחת את העמודות העיקריות שלדעתנו ניתן היה לעבד אותן לפיצ'רים מרכזיים שישפיעו על תוצאות המודל למשל, כמות שערים של כל קבוצה במשחק, מי השחקנים ששיחקו בהרכב בכל קבוצה במשחק, יחסי הימורים של 4 סוכנויות הימורים מרכזיות עבור תוצאת משחק ועוד.
בטבלת השחקנים ביצענו שאילתה שמציגה את היכולות האישיות של כל שחקן במספר מדדים, בחרנו לקחת את העמודה של הדירוג הכולל המשוקלל עבור כל שחקן ולאחר מכן ביצענו מיזוג עם טבלת המשחקים ושייכנו לכל משחק את היכולות המשוקללות של השחקנים ששיחקו באותו משחק בהרכב.
בנוסף, כפי שקראנו במאמרים בסקירה הספרותית, בעזרת העמודות שציינו לעיל בחרנו ליצור פיצ'רים כמו למשל, הפרשי שערים ב-10 משחקים אחרונים של קבוצת הבית וקבוצת החוץ, מס' הניצחונות של קבוצת הבית והחוץ ב-10 משחקים האחרונים, מס' הניצחונות של קבוצת הבית מול קבוצת החוץ ב-4 המפגשים האחרונים ביניהן ומס' הניצחונות של קבוצת החוץ מול קבוצת הבית ב-4 המפגשים האחרונים ביניהן (ההתייחסות היא למפגשים באצטדיון הביתי של כל קבוצה).

3.2. שיטות ותיאור הניסוי

המודל שלנו מתבסס על בסיס נתונים של sql-lite , לצורך השימוש בבסיס הנתונים עשינו ייבוא לספריית sql-lite ע"מ שנוכל ליצור חיבור ולבסיס הנתונים ולמשוך את הדאטה.
גרסת הפייתון שבה השתמשנו היא 2.7 .
הספריות:

- Pandas - כדי לעבוד עם אובייקטים של data-frame במשיכת הנתונים מבסיס הנתונים.
- Numpy – בשביל לבצע פעולות שונות על אובייקטים של מערכים.
- SKLearn – ספרייה עיקרית שמכילה את כל המודלים שהרצנו על סט הנתונים שיצרנו לאחר עיבוד, בנוסף השתמשנו בכלים שלה לביצוע הערכה על המודל.
- Matplotlib – נעזרנו בה כדי להציג באופן ויזואלי את תוצאות המודלים (למשל גרפים).
- Os.path – ספריה שבעזרתה בדקנו נתיבים של קבצים (האם קיים או לא).
- Time – עזרה לנו למדוד את זמני הריצה של ייצור הפיצ'רים ובניית המודלים השונים.
- Warning – אפשרה לנו להתעלם מאזהרות שבאו בעקבות שחזור קוד בחלקים מסוימים.

הקוד שלנו מתחלק ל-2 חלקים עיקריים: החלק שבו אנו קוראים את הדאטה, מחלקים אותו לסט אימון וסט מבחן, מייצרים את הפיצ'רים שבהן נשתמש בהמשך. חלק נוסף שבו אנו מריצים את המודלים שבחרנו על סט האימון וסט המבחן עם הפיצ'רים, מקבלים תוצאות חיזוי של כל מודל ומבצעים הערכה והשוואה בין המודלים.

חלק ראשון – על מנת לייעל זמני ריצה, החלטנו לשמור קבצי CSV המכילים את הדאטה המעובד ועליהם להריץ את המודלים בהמשך. במידה וקבצי CSV אלו עדיין לא נוצרו אנו יוצרים אותם, אם קיימים נעבור לחלק הבא של הרצת המודלים.

יצירת קבצי ה-CSV – במידה וזו ריצה ראשונה ניצור קובץ CSV של סט אימון וקובץ CSV של סט מבחן. סט המבחן שלנו נבחר להיות עונת 15/16 וסט האימון נבחר להיות כל שאר העונות 08/09-14/15 .

עיבוד מידע מטבלת משחקים לטבלת פיצ'רים ראשונה - מאחר וקיימות עמודות עם מידע חסר מאוד כמו כרטיסים אדומים למשל, החלטנו להסיר את העמודות הללו ולקחת עמודות ספציפיות עם מידע יותר שלם. מתוך העמודות שבחרנו הוצאנו רשומות שבהן היה ערך NULL באחת העמודות (בחרנו לפעול כך משום שניסינו גם לנרמל את הנתונים ע"י ממוצעים ולהשתמש בהם בצורה מנורמלת אך אלו הפיקו תוצאות חיזוי פחות טובות).

במעבר על הטבלה עם כלל הערכים יצרנו עבור קבוצת הבית וקבוצת החוץ בכל משחק את הפיצ'רים הבאים:

- הפרש שערים של קבוצת הבית (שערי זכות פחות שערי חובה) ב-10 המשחקים האחרונים שלה
- הפרש שערים של קבוצת החוץ ב-10 המשחקים האחרונים שלה.
- כמות הניצחונות של קבוצת הבית ב-10 המשחקים האחרונים שלה.
- כמות הניצחונות של קבוצת החוץ ב-10 המשחקים האחרונים שלה.
- כמות הניצחונות של קבוצת הבית על קבוצת החוץ ב-4 המשחקים האחרונים ביניהן.
- כמות הניצחונות של קבוצת החוץ על קבוצת הבית ב-4 המשחקים האחרונים ביניהן.

מאחר ובנתונים שקיבלנו אין תוצאות מוגדרות לפי ניצחון, תיקו והפסד אלא רק כמה שערים הבקיעה כל קבוצה במשחק, יצרנו פונקציה הלוקחת את כמות השערים של כל קבוצה במשחק ותייגנו את התוצאה לפי קבוצת הבית, כלומר במידה וקבוצת הבית ניצחה במשחק מסוים, המשחק יקבל תיוג של ניצחון. במידה וקבוצת הבית הפסידה, המשחק תויג כהפסד (המטרה הסופית היא לחזות ניצחון/תיקו/הפסד).

עיבוד טבלת נתוני שחקנים לטבלת פיצ'רים שנייה – עבור כל משחק לקחנו את הדירוג הכולל של שחקני שתי הקבוצות.

עיבוד טבלת נתוני יחסי הימורים לטבלת פיצ'רים שלישית - עבור כל משחק ביצענו נרמול של נתוני יחסי ההימורים לסקלה שבין 1-0 כלומר, לקחנו את יחסי ההימורים של כל סוכנות הימורים והמרנו אותה ליחס הסתברותי של הסיכויים לקבל במשחק מסוים ניצחון, תיקו או הפסד.

לאחר סיום העיבוד ביצענו JOIN לפי מזהה משחק בין כל טבלאות הפיצ'רים לטבלה אחת מאוחדת של פיצ'רים.

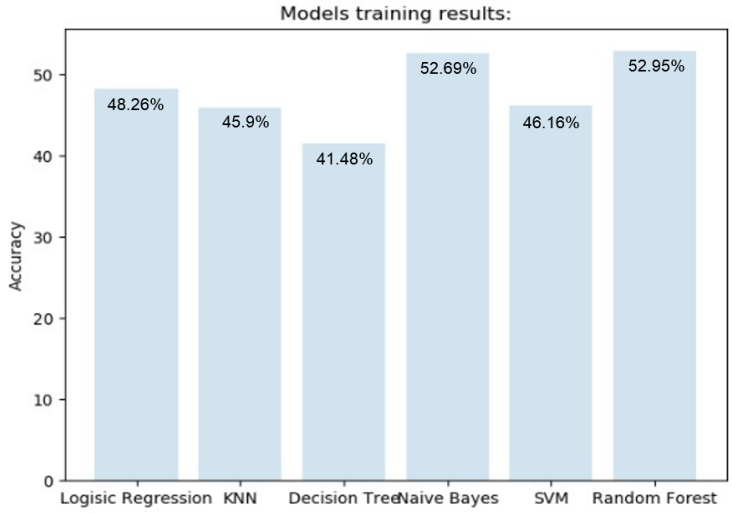
את כלל הפעולות הללו ביצענו גם על סט האימון וגם על סט המבחן ושמרנו אותם בקבצי CSV עבור כל סט ע"מ שבפעם הבאה, במידה ונרצה להריץ שוב מתוך כוונה לשנות משהו רק במודלים עצמם, לא נצטרך להכין מחדש את כלל הדאטה.

חלק שני - טענו את קבצי ה-CSV והחסרנו מהם את עמודת התיוג שיצרנו ע"מ לאמן את המודל לפי קובץ האימון. ביצענו K-FOLD שהתחלק ל-10 חלקים במטרה למנוע over-fitting של הנתונים על כל מודל שבחרנו במהלך האימון. בחרנו לבחון את סט האימון בעזרת 6 מודלים: Logistic Regression, KNN, Decision Tree, Naïve Bayes, SVM, Random Forest – מהם נבחר המודל בעל תוצאות האימון הטובות ביותר ועליו הרצנו את סט המבחן(קובץ CSV נוסף שממנו הסרנו את עמודת התיוג).

כדי לבחור את ה-hyper parameters הטובים ביותר עבור כל מודל שהרצנו על סט האימון, השתמשנו בפונקציה שנקראת GridSearchCV אליה הכנסנו כל מודל עם פרמטרים שונים. אותם הפרמטרים שחזרו הם אלו שהכנסנו למודלים שהרצנו על מנת לבחון את סט האימון.

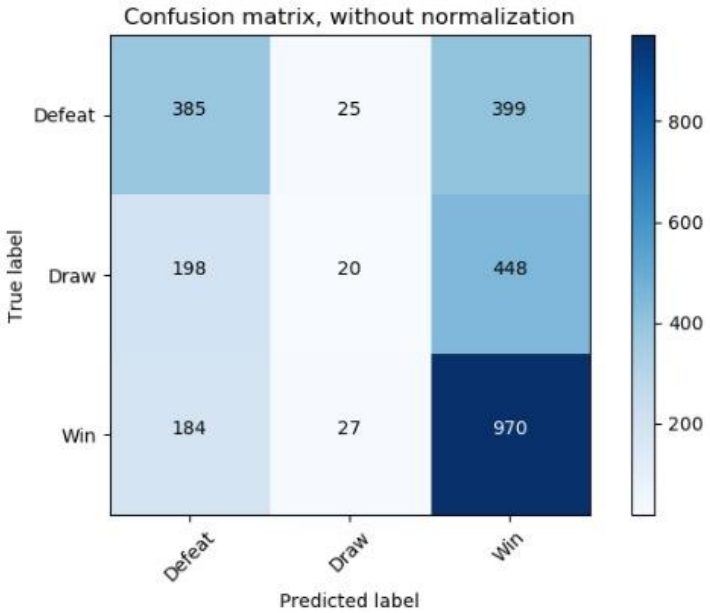
3.3. תוצאות והערכות

מבין כלל המודלים שהרצנו על סט האימון בעזרת שימוש ב-cross validation המודל שהחזיר את ממוצע התיוגים המדויקים הגבוה ביותר היה מודל Random Forest עם hyper parameters הבאים - n_estimators=25, min_samples_split=25, max_depth=10 שהביאו לממוצע של : 52.95% .



על סמך תוצאה זו, בחרנו להריץ את סט המבחן בעזרת מודל זה עם הפרמטרים שקיבלנו ולבחון את התוצאות שהתקבלו. בחרנו לבצע הערכה על התוצאות של המודל על סט המבחן בעזרת מדדים נוספים כמו דיוק המודל, precision, recall, f1 score.

הדיוק של המודל על סט המבחן היה 51.77% .



יצרנו Confusion Matrix של תוצאות ההרצה של המודל על סט המבחן ובעזרתה ניתן לראות ולבחון בצורה ויזואלית את טיב התוצאות.
ניתן לראות כי תוצאת תיקו הינה הכי קשה לחיזוי עם 20 חיזויים נכונים מתוך 666 משחקים שנגמרו בתיקו.
חיזוי של תוצאת ניצחון עבור קבוצת הבית היא בהסתברות גבוהה מאוד כאשר ישנם 970 חיזויים נכונים מתוך 1181 משחקים שנגמרו בניצחון קבוצת הבית.

```
Classification Report:
              precision    recall  f1-score   support

      Defeat       0.50      0.48      0.49       809
        Draw       0.28      0.03      0.05       666
         Win       0.53      0.82      0.65      1181

   micro avg       0.52      0.52      0.52      2656
   macro avg       0.44      0.44      0.40      2656
weighted avg       0.46      0.52      0.45      2656
```

כדי לקבל את תוצאות שאר המדדים השתמשנו בפונקציה Classification report שמציגה עבור כל תיוג(ניצחון/תיקו/הפסד) את מדדי ה- precision, recall, f1 score.
המדד המשמעותי שבחרנו להסתכל עליו הוא f1-score שמשקלל את הממוצע ההרמוני של ה-precision וה-recall ולמעשה נותן לנו הערכה ברורה יותר לטיב תוצאות המודל.
ניתן לראות זאת במדד ה-micro avg אשר מבצע את החישוב שלו על סמך כל התוצאות של כל הסיווגים (לעומת המאקרו שמבצע חישוב פר תיוג ועושה ממוצע ביניהם)
קיבלנו תוצאה של 0.52 וניתן לומר שהמודל שלנו פוגע בחיזוי שלו במעט יותר מחצי מהמשחקים.

4. סיכום

לאור סקירת הספרות עשינו שימוש במספר מודלים מאומנים לקבלת חוות דעת רחבה יותר בעניין הסיווג ובדקנו את השיטות שנתנו בניסויים הקודמים דיוק גבוה: Naïve Bayes, Random Forest, KNN, SVM, Decision Tree ו- Logistic Regression ו- בנוסף השתמשנו K-Fold Validation כאשר k=10 לאימון המודל ובדיקתו, שיטה שהופיעה בהרבה ניסויים שנעשו בעבר. הנתונים והפיצ'רים בניסוי שלנו לא דומים במלואם לנתונים בסקירה אך מבוססים על הפיצרים שהוצעו והגדילו בהם את הדיוק בחיזוי כגון: הפרשי שערים, מספר ניצחונות של קבוצת הבית וקבוצת החוץ במשחקים האחרונים.

התוצאות שהתקבלו מראות כי חיזוי משחקי כדורגל אינו פשוט. הצלחנו להשיג דיוק מעל ל 50% אך הדיוקים שקיבלנו בבדיקת המודלים היו באחוזי דיוק פחות גבוהים מהסקירה. אנו מניחים כי אם היו ברשותנו נתונים סטטיסטים נוספים ולא חסרים כמו בעיטות לשער, קרנות, עברות וכו' תוצאות הדיוק שלנו היו משתפרות. בעת בניית המודל בדקנו האם ניתן בכל זאת לשלב את העמודות עם הערכים חסרים ע"י השלמתם לממוצעי העמודה, אך הדיוק שהתקבל היה נמוך יותר ולכן העדפנו להתעלם מעמודות אלו ולהגדיל את הדיוק.

יש עוד הרבה מה לעשות בתחום, בעת ביצוע ההשוואות התגלו מספר חולשות למערכת שלנו כגון: סיווג הנתונים לשלושה סיווגים- ניצחון הפסד או תיקו הקשו על החיזוי ובייחוד סיווג של תיקו לא היה מוצלח . בנוסף, היינו מציעים להשלים את הנתונים החסרים בנתונים נכונים ואמיתיים כך שייתכן והיו מתקבלות תוצאות הרבה יותר טובות בחלק מהמודלים. כמו כן, נכחנו לראות כי לאור סקירת הספרות הרחבה שביצענו ישנו מחסור גדול של מחקרים בתחום חיזוי תוצאות משחק כדורגל. ולכן יש לעשות מחקרים נוספים ורחבים יותר, תוך חקירת קלטים שונים ושימוש במודלים שונים של למידת מכונה עם מספר פונקציות שונות.

ביבליוגרפיה

[1] Introduction to Machine Learning. Ethem Alpaydin, MIT press, 2004.

[2] Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team. Farzin Qwramipur, Parinaz Eskandarian, Faezeh Mozneb, 2013.

[3] Prediction of Football Match Score and Decision Making Process. IT_SAKEC, 2018.

[4] A Comparative Study on Neural Network based Soccer Result Prediction. Burak Galip Aslan, Mustafa Murat Inceoglu,2007.

[5] Predicting Soccer Match Results in the English Premier League. Ben Ulmer, Matthew Fernandez.

[6] Predicting the Outcome of a Football Game: A Comparative Analysis of Single and Ensemble Analytics Methods. Enes Eryarsoy, Dursum Delen, 2019.

[7] Predicting football scores using machine learning techniques. Josip Hucaljuk, Alen Rakipovic, 2011.

[8] An improved Prediction System for Football a Match Result. University of Port Harcourt, Nigeria,2014.