תיעוד פרויקט במדעי הנתונים עידן אלאשוילי ושקד כהן:

<u>פרק 1 - מבוא:</u>

כדוריד הוא ענף ספורט קבוצתי בו הרכב פותח מונה 7 שחקנים, המחולקים לעמדות: 2 פינות, 2 מקשרים, רכז, פיבוט ושוער.

מטרת המשחק היא להבקיע יותר שערים מהקבוצה היריבה.

השחקנים משתמשים בידיהם על מנת לתפוס, למסור, ולזרוק את הכדור, שוערים יכולים גם לקיים מגע בכדור עם רגליהם כל עוד הם בתוך רחבת השוער.

אורך משחק הוא 60 דקות למשחקי הליגה, משחקי גביע ושלבי הנוקאאוט בטורנירים ימשכו להארכה של 10 דקות מחולקות למחציות של 5 דקות, במקרה של תיקו נוסף תהיה הארכה נוספת, ואם לא המשחק לא יוכרע אז פנדלים.

ה- **Bundesliga** היא הליגה הבכירה ביותר בכדוריד בגרמניה ונחשבת לאחת הליגות המובילות והחזקות בעולם בתחום זה. הליגה נוסדה בשנת 1966, והיא כוללת כיום 18 קבוצות שמתחרות ביניהן במהלך העונה המתחילה בסתיו ומסתיימת באביב.

בליגה משתתפים שחקנים ברמה הגבוהה ביותר, כולל שחקנים בינלאומיים מהנבחרות המובילות בעולם. הקבוצות המתחרות בליגה ידועות בשל המשחק המהיר, הטקטיקות המתקדמות, והאווירה התחרותית הגבוהה.

פרק 2 -הבעיה/הפער שננסה לפתור:

אנחנו מנסים לחזות כיצד נתוני שחקנים יכולים להשפיע על מיקום הקבוצה בליגה.

כדי לענות על שאלה זו אנו ניעזר במודל חיזוי לדירוג הטבלה בליגה הגרמנית בכדוריד באמצעות נתונים אישיים של כל שחקן.

לבסוף נרצה להשוות בין הטבלה המקורית של כל עונה לבין הטבלה שהמודל ייצר ונראה האם יהיו שינויים משמעותיים בין הטבלאות, או שאולי הטבלאות ייצאו זהות לחלוטין.
אנו מנבאים כי הטבלאות יהיו דומות אך לא זהות, כך שכל קבוצה תהיה במרחק של לכל היותר 5 מקומות בטבלת החיזוי ביחס לטבלה המקורית לכל עונה.

במהלך הפרויקט נשתמש ב Random Forest Regressor (פירוט בפרק 3) על מנת ליצור את החיזוי, תוך אימון המודל על גבי 3 עונות של מידע המחולק למידע נתוני שחקן אישיים ונתוני שוער אישיים עליהם יפורט בפרק 4.

פרק 3 - השיטה

1)Scraping

בתחילת הדרך השתמשנו בדאטאסט יחיד המונה את כל הערכים על עונת 2022/23 בליגה הגרמנית.
לאחר התקדמות הבחנו כי מאגר זה אינו מספק את הצרכים של הפרויקט שלנו ולכן התחלנו בחיפוש דאטאסטים רבים עם נתונים נרחבים ביחס לקודם.
מצאנו באתר הליגה הגרמנית הרשמי מאגר מאוד רחב עם יותר מ-10 שנים של עונות בהם יש לנו מידע על כל שחקן ועל כל קבוצה ששיחקו באותה העונה, בחרנו להשתמש לצורך הפרויקט ב-4 העונות האחרונות, 2020/21-2023/24.

נתקלנו במספר בעיות כדי להשיג את כל המידע, ראשית היה עלינו לעשות Data Scraping לאתר על מנת לייבא את הנתונים מקובץ HTML לקובץ CSV .
בעת ביצוע Scraping נתקלנו בבעיה שהאתר מכיל בתוך עמודת שמות השחקנים את שם הקבוצה שלהם ואת שם העמדה המרכזית שלהם, ובנינו קוד פייתון שיוכל לזהות את המצב ולתקן אותו שיוסיף לנו עמודות חדשות של שם הקבוצה והעמדה המרכזית, וישנה את העמדה של שם השחקן שתכיל רק את השם המלא של השחקן.

2)Merging:

תחילה, נטענו ונוקו מערכי נתונים שונים המכילים סטטיסטיקות של שחקנים, שוערים, לעונות 2020/2021-2023/2024, הנתונים נאספו ונרשמו לפי קבוצות, כך שכל קבוצה קיבלה ממוצע של הסטטיסטיקות שלה.

התאמת התכונות:

לאחר איסוף הנתונים, התאימו את העמודות של מערך הנתונים החדש לעמודות ששימשו באימון המודל, כדי לוודא שהמודל יכול לעבד את הנתונים החדשים כראוי.

3)Random Forest Regressor:

Random Forest הוא אלגוריתם למידת מכונה שמשתמש במספר רב של עצי החלטה (Decision Trees) כדי לחזות ערכים רציפים (במקרה של Regressor). בכל עץ החלטה מתקבלת תחזית, ו-Random Forest משלב את כל התחזיות האלו כדי להגיע לתחזית סופית (במקרה של Regressor, זה יהיה ממוצע התחזיות).

כיצד האלגוריתם עובד?

יצירת עצים שונים: האלגוריתם יוצר מספר רב של עצי החלטה. כל עץ מאומן על דגימה שונה של הנתונים. זה עוזר להקטין את ההטיה (bias) ולהגדיל את הדיוק.

שימוש במשתנים שונים: בכל צומת של עץ ההחלטה נבחרים משתנים שונים באקראי כדי להחליט על הפיצול, מה שעוזר להקטין את הקורלציה בין העצים.

תחזית סופית: במקרה של Regressor, כל עץ מנבא ערך רציף, והתחזית הסופית מתקבלת על ידי ממוצע התחזיות של כל העצים.

כיצד זה מיושם בקוד:

פיצול הנתונים: הנתונים מתחלקים למערך אימון (80%) ומערך בדיקה (20%). מערך האימון משמש לבניית המודל, ומערך הבדיקה משמש להערכת הביצועים של המודל.

MAE: מציג כמה בממוצע התחזיות שונות מהערכים האמיתיים. (אצלנו 6.32)

R^2 Score: מציג את האיכות הכללית של המודל, כאשר ערך קרוב ל-1 מראה על התאמה טובה בין התחזיות לנתונים האמיתיים. (אצלנו 0.7)

שימוש ב-Random Forest Regressor בקוד מאפשר לבצע תחזיות מדויקות יחסית לגבי מספר הנקודות שקבוצה תקבל בעונה, בהתבסס על נתונים היסטוריים של השחקנים והקבוצות.

פרק 4 - תיאור הדאטאסטים:

בתחילת הדרך השתמשנו בדאטאסט יחיד המונה את כל הערכים על עונת 2022/23 בליגה הגרמנית.
לאחר התקדמות הבחנו כי מאגר זה אינו מספק את הצרכים של הפרויקט שלנו ולכן התחלנו בחיפוש דאטאסטים רבים עם נתונים נרחבים ביחס לקודם.
מצאנו באתר הליגה הגרמנית הרשמי מאגר מאוד רחב עם יותר מ-10 שנים של עונות בהם יש לנו מידע על כל שחקן ועל כל קבוצה ששיחקו באותה העונה, בחרנו להשתמש לצורך הפרויקט ב-4 העונות האחרונות, 2020/21-2023/24.

נתקלנו במספר בעיות כדי להשיג את כל המידע, ראשית היה עלינו לעשות Data Scraping לאתר על מנת לייבא את הנתונים מקובץ HTML לקובץ CSV .
בעת ביצוע Scraping נתקלנו בבעיה שהאתר מכיל בתוך עמודת שמות השחקנים את שם הקבוצה שלהם ואת שם העמדה המרכזית שלהם, ובנינו קוד פייתון שיוכל לזהות את המצב ולתקן אותו שיוסיף לנו עמודות חדשות של שם הקבוצה והעמדה המרכזית, וישנה את העמדה של שם השחקן שתכיל רק את השם המלא של השחקן.

יש לנו את הנתונים האישיים לכל שחקן על גבי אותן 4 עונות:

- כמות הבקעות של פנדלים
- אחוזי הבקעת פנדלים
- כמות הבקעות כאשר השחקן שיחק בעמדה מסוימת(מקשר,מרכז,פינה,פיווט)
- אחוזי הבקעות לכל עמדה
- כמות הבקעות בהתקפות מתפרצות
- אחוזי הבקעה בהתקפות מתפרצות

ונתונים של שוערים:

- כמות הצלות כוללת
- הצלות פנדלים
- הצלות לכל עמדה(מקשר,מרכז,פינה,פיווט)
- הצלות של התקפות מתפרצות
- הבקעות (בכדוריד כאשר שחקן מקבל הרחקה של 2 דקות קבוצות בדרך כלל מוציאות את השוער כדי להכניס שחקן להתקפה, בהסתמך על חילוף מהיר כאשר ההתקפה מסתיימת כך שנוצר מצבים שבהם שוער מבקיע בזריקה משער לשער).

אל כל אלו מתווספים גם הנתונים של באיזה קבוצה השחקן משחק, וכמות המשחקים ששחקן שיחק, על מנת לאמוד את הנתונים שלו להשפעה על הקבוצה.
לדוגמא: שחקן ששיחק 5 משחקים והבקיע בהם 20 שערים יגרום ליותר השפעה מאשר שחקן ששיחק 20 משחקים והבקיע 20 שערים.
כמות המשחקים המקסימלית הינה 38 משחקים.

פרק 5 - הסבר תהליך המחקר

איסוף נתונים -

מכיוון שלא היה קיים dataset היינו צריכים ללכת לאתר הרשמי של הליגה הגרמנית בכדוריד ולחלץ משם את הנתונים לקובץ html .לאחר מכן באמצעות קוד שרשמנו המרנו את ה-html לטבלה בקובץ csv. מכיוון שבאתר הרשמי הנתון של שם קבוצה ומיקום השחקן רשומים באותה שורה היינו צריכים לכתוב עוד קוד על מנת לפצל את המידע לפי תבנית שזיהינו.

עיבוד הנתונים -

עיבדנו את הנתונים והתאמנו אותם על מנת שיוכלו להתאים לאימון של מודל החיזוי.

לדוגמה הורדנו את התו "%" והחלפנו בערך מספרי של אחוזים (0.7 =70%)

ו "-" הוחלף ב NaN

שימוש ב - RandomForestRegressor

בחרנו בציר ה - X נתונים שלא מסגירים את כמות הנקודות שקבוצה עשתה.

בחרנו בציר ה - Y את נקודות הקבוצה .

אימנו את המודל לחזות ב test size 0.2 שזה אומר ש20% מהנתונים ישמשו לסט הבדיקה. ו80% לסט אימון.

כלומר כל פעם נתנו לו 80% ורצינו לראות איך הוא מנבא את ה20% הנותרים.

הצגת תוצאות -

בנינו הצגה ויזואלית בעזרת matplot בה השווינו את הטבלה האמיתית לעומת הטבלה שהמודל שלנו חזה.

בטבלה ניתן לראות את ההבדלים בין החיזוי לבין הטבלה האמיתית.

פרק 6 - תוצאות

Predicted Handball League Standings (2023-2024)

Rank	Team
1	SCM (→ 0)
2	SGF (↑ 1)
3	MTM (↑ 2)
4	THW (→ 0)
5	RNL (↑ 7)
6	GUM (→ 0)
7	BER (↓ 5)
8	TBV (↑ 2)
9	WET (↑ 4)
10	HSV (↓ 1)
11	HBW (↑ 7)
12	TVB (↓ 1)
13	LEI (↓ 5)
14	BHC (↑ 3)
15	FAG (→ 0)
16	EIS (↓ 2)
17	HCE (↓ 1)
18	HAN (↓ 11)

המודל חזה במדויק את מיקומה של הקבוצה האלופה והראה שתפס למעשה את הביצועים של הקבוצות החזקות ביותר. זה מעיד על אמינותו בקצה הגבוה יותר של הדירוג.
יש תזוזות ניכרות בדרגות עבור מספר קבוצות, כמו RNL שעולה 7 מקומות ו- LEI יורדת 5 מקומות. זה מצביע על כך שאמנם המודל טוב יחסית, אך עשויים להיות לו אתגרים בניבוי אמצע הטבלה שבהם השונות גבוהה יותר.

הטבלה מדגישה שונות מסוימת בתחתית הטבלה, כאשר קבוצות קיבלו שינוי של מספר מקומות, אך לא באופן משמעותי ביחס לאמצע הטבלה.

האנובר (HAN) התמקמה במקום האחרון בטבלת המודל, כנראה בגלל שהם היו חריגים לטובה במהלך העונה, יתכן גם שהיו גורמים חיוביים בלתי צפויים במהלך העונה, שלא בהכרח קשורים לנתוני השחקנים.

לעומת זאת, ייתכן והמודל לא הצליח להעריך אותם כראוי בשל הערכה נמוכה של נתונים המדגישים את קבוצה זו ספציפית.

לסיום, המודל הצליח לחזות כמעט באופן מושלם את הטבלה ביחס לשאלת החקר שלנו, עם אחוזי פגיעה של 83.33%! (מרחק של לכל היותר 5 מקומות)

פרק 7 - מסקנות

המודל כמעט והצליח באופן מושלם לממש את הציפייה שלנו בתחילת המחקר, כאשר קיימות רק 3 קבוצות אשר חרגו.

הנתון המוערך ביותר הינו מספר המשחקים ששחקן שיחק, אשר הגיוני לחלוטין, המודל העריך בצורה נכונה שחקן ששיחק את מירב המשחקים בעונה לעומת שחקן ששיחק מספר מועט.

הקבוצה במקום הראשון זהה בשתי הטבלאות, מה שמראה על דיוק המודל בלקיחת הנתונים על הקבוצות החזקות.

אמצע הטבלה היה קשה לחיזוי, כמעט ואין קבוצות במיקומים זהים בשתי הטבלאות, אך רוב הקבוצות אכן נמצאות בטווח זה בשתי הטבלאות.
האנובר חוו ירידה משמעותית בטבלה החזויה, כנראה בגלל הערכת נתונים לא מדויקת של המודל או הפתעות לטובה במהלך העונה.

לאחר המחקר, ניתן להסיק כי לנתוני השחקנים יש חלק חשוב בהשפעה על מיקומה של קבוצה בליגה, ניתן לראות זאת בכך שהאלופה האמיתית היא האלופה שנחזתה,
הקבוצות שהיו בחלק העליון של הטבלה נשארו כך, ומירב הקבוצות בתחתית הטבלה נשארו כך גם כן.