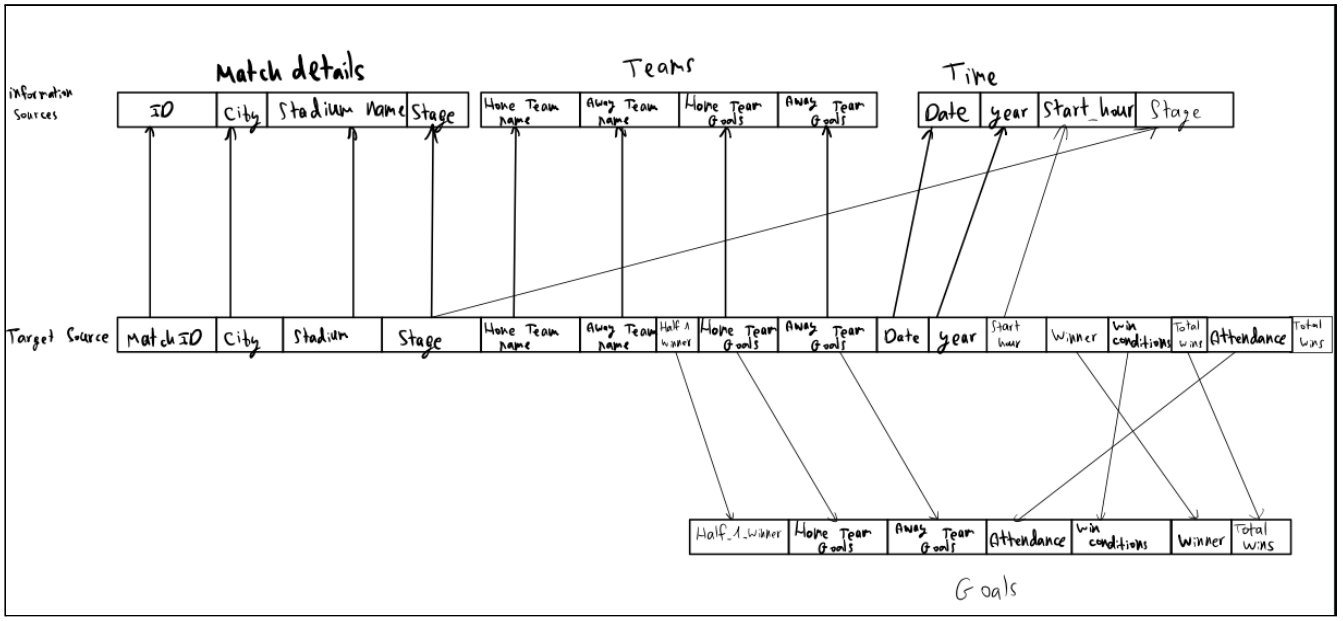
בינה עסקית - מטלה 3

מגישות: עמית ניצן, ניצן זיפלינגר ונועה בורג

חלק 1: STTM

קישור ל- STTM WorldCUP.xlsx : STTM

סכימה:



חלק 2:

א. תהליך הKDD שלנו כולל:

- Data Cleaning - Finding the missing data, removing noise, removing low - quality data.
- Data Selection - Seperate the data into relevant sets.
- Data Transaformation- the data will be put together. Transform the data into suitable form.
- Data mining - Extract relevant and useful patterns from the transformed data. We will use statistical methods and classification methods.
- Pattern Evaluation - Visualiztion of the patterns that were found throgh graphs and time plots.

נבצע תהליך Predictive כיוון שנרצה לחזות מי המדינה אשר תנצח את המונדיאל הבא.

עבור שאלת הSUPERVISED שלנו נשתמש באלגוריתם מתחום ניתוח האשכולות – KNN.

אלגוריתם השכן הקרוב – אלגוריתם זה מהיר יחסית וקל לשימוש.

הרעיון באלגוריתם זה, הוא שכשאר מגיעה תצפית חדשה אז בוחנים לאילו תצפיות קיימות היא "קרובה" ומניחים פיישהערך שלה יהיה הסיווג הנפוץ בקרב התצפיות הקרובות.

עבור שאלת ה UNSUPERVISED נשתמש באלגוריתם K-means, המטרה שלנו היא קיבוץ נתונים לפי מאפייני דמיון שונים ובכך נחזה האם הקבוצה תנצח.

אופן הפעולות : נאפיין את הצרכים הנדרשים לפרויקט זה. לאחר מכן, נבנה מסד נתונים. על מסד הנתונים נבצע ניקוי וסינון כפילויות ולבסוף נבצע טרנספורמציה למידע.

ב. 1. ניישם את אלגוריתם k-means על מנת לבצע פעולת clustering לדאטה. זהו האלגוריתם הנפוץ והאמין ביותר מסוג אלגוריתמי מבוססי מרחק, הוא מאפשר התמודדות עם נתונים ללא התפלגות מוגדרת, כמו הנתונים בפרויקט שלנו. לאחר הגדרת האשכולות ניתן לעבוד בעזרת האלגוריתם ללא נתוני המקור, דבר שייקל על העבודה שלנו מפני שיש לנו סוגי דאטה רבים ושונים. הclusters מתחלקים לפי מרכזי הכובד (דמיון), דבר שיאפשר לנו לראות באופן ברור את ההבדלים ואת הפערים בין קבוצות הכדורגל השונות והשלבים (stages &rounds), אשר מהווים חלק מרכזי מהדאטה שלנו.
2. *לצורך קבלת ההחלטה אילו שחקנים ישחקו באיזה משחק ומתי (מחצית ראשונה או שניה) יש לנתח את היסטוריית המשחקים ולחזות את הסיכויים לפי משחקי העבר. נשתמש במאפיינים הבאים: תוצאת המשחק, כמות גולים במחצית הראשונה, כמות גולים במחצית השניה, האם משחק בית\חותץ, קבוצה יריבה.

לדוגמא: במשחק שנערך ב14/06/2014 קולומביה מול יוון, כחלק משלב "Group C" ניתן לראות שקולומביה ניצחה במשחק. והיא גם הובילה במחצית הראשונה.

*לצורך מיקסום הנצחונות של קבוצה מסוימת, המאמן רוצה לדעת האם יש בקבוצה קושי בזמן המחצית השניה. לצורך כך נאסוף ונסווג את המשחקים שבהם הקבוצה שלו (של המאמן) הפסידה, לאחר מכן נסתכל על נתוני המשחקים ונראה האם במשחקים האלו הקבוצה הובילה במחצית הראשונה ולאחר מכן התהפכו התוצאות והקבוצה בסופו של דבר הפסידה. דבר זה מעיד (ככל הנראה) על חולשה בצוות השחקנית שמשחק החל מהמחצית הראשונה ויכול להוביל לכך שהפתרון לבעיה יהיה בחילוף מאסיבי של שחקנית בזמן המחצית השניה.

2. המדד הנבחר הוא מדד אוקלידי.

בעזרת מדד זה נחלק את הנתונים לאשכולות שונים. המדד הזה הוא המתאים ביותר מיוון שהנתונים החשובים שלנו ברובם נומרים ולכן העמודות יהיו מוגדרות על פי מרחקים.

3. השערת המחקר הראשונה – האם ניתן לחלק את הדאטה לאשכולות שונים לפי דמיון של נצחונות וחוזק קבוצות, ובכך לדעת האם האשכול החזק הוא בעל הסיכויים הטובים ביותר לנצחון.

H0 – לא ניתן לחלק את הנתונים לאשכולות דומים

H1 – ניתן לחלק את הנתונים לאשכולות דומים

השערת מחקר שניה – האם ניתן לחזות כי קבוצה מסוימת תנצח בהפרש הגדול מ1 מהיריבה שלה.

H0 – אין קשר סטטיסטי בין הנתונים לבין כמות הגולים הצפויה

H1 – קיים קשר סטטיסטי בין הנתונים לבין כמות הגולים הצפויה

*לצורך כך נשתמש במבחן פישר.

מבחן פישר: מבחן סטטיסטי לבדיקת השערת אי תלות בין 2 משתנים איכותיים. נבחר אלפא=0.05 לשם הערכת הקשר.

חלק 3: שאילתות SQL:

א.

השאילתא הבאה מאפשרת לנו לחשב מי הקבוצה המנצחת בכל משחק, האם היא קבוצת הבית או קבוצת החוץ. השאילתא מציגה גם את שם הקבוצה והאם היא הקבוצה המארחת (host) או הקבוצה האורחת (guest).

```
SELECT

   'host'     AS side,
   home_team_name  AS team_id,
   home_team_goals AS goals,
   CASE WHEN home_team_goals > away_team_goals THEN 1 WHEN home_team_goals = away_team_goals THEN 0 ELSE -1 END AS victory
FROM
   Teams

UNION ALL

SELECT

   'guest'     AS side,
   away_team_name  AS team_id,
   away_team_goals AS goals,
   CASE WHEN away_team_goals > home_team_goals THEN 1 WHEN away_team_goals = home_team_goals THEN 0 ELSE -1 END AS victory
FROM
   Teams
```

ב.השאילתא הבאה מציגה עבור כל קבוצה, אם היא הייתה מארחת או מתארחת במשחק מסוים ואת מספר הגולים הכולל שהבקיעה. שאילתא זו יכולה לענות על כמה שאלות עסקיות, למשל:

- האם קבוצה מסוימת מבקיעה יותר כאשר היא מתארחת או כאשר היא מארחת? האם קבוצות מרגישות יותר "בנוח" במגרש שלהן?
  למשל, אפשר לראות שברזיל הבקיעה סה"כ 180 גולים כאשר היא הייתה הקבוצה המארחת, אבל לעומת זאת במשחקי חוץ ,היא הבקיעה 45 גולים בלבד.

```
SELECT home_team_name, SUM(home_team_goals)
OVER(PARTITION BY home_team_name) AS home_team_victories,
away_team_name, SUM(away_team_goals)
OVER(PARTITION BY away_team_name) AS away_team_victories
FROM WorldCupMatchess
```

ג. השאילתא הבאה מציגה את מספר הגולים הממוצע של קבוצת הבית בכל שעת משחק. השאילתא עונה על השאלה העסקית העוסקת בשעות המשחקים, ובעזרת נתונים אלה יוכלו להסיק בעלי התפקידים (מאמנים, קפטן, מנהל מקצועי, בעלים) באילו שעות השחקנים מתפקדים הכי טוב.
למשל, בעזרת השאילתא הזו אפשר לראות שבין השעות 14:00-16:00 השחקנים הצליחו להבקיע את הכמות הכי גדולה של גולים.

```sql
SELECT matchid,start_hour, AVG(home_team_goals)
        OVER(PARTITION BY start_hour) AS average_goals_per_hour
FROM WorldCupMatchess
ORDER BY average_goals_per_hour DESC
```

ד. השאילתא הבאה משתמשת בפונקציית Ntile על מנת לחלק את מספר הגולים שהובקעו במשחק ל-4 חלקים. בעזרת השאילתא ניתן לדעת עבור כל מס' גולים לאיזה אחוזון (רבעון) הוא שייך, ועל ידי איזו קבוצה הוא הובקע, האם היא הייתה הקבוצה המארחת או קבוצת הביץ ובאיזה אצטדיון היה המשחק.

```sql
SELECT winner,
  CASE WHEN winner = home_team_name THEN 'home team'
  WHEN winner = away_team_name THEN 'Away Team' ELSE 'TIE' END AS 'home/away', stadium,home_team_goals,away_team_goals,
  NTILE (4) OVER(ORDER BY home_team_goals,away_team_goals) AS RankNo
FROM WorldCupMatchess
```

ה. השאילתא הבאה מציגה לנו את "3 המנצחות הגדולות". על ידי שימוש בפונקציית COUNT, ספרנו כמה ניצחונות יש לכל קבוצה ועל ידי שימוש בLIMIT וDESC הגבלנו את התוצאה ל3 הראשונות.

```sql
SELECT winner, COUNT(winner)
FROM Goals
WHERE winner <> 'Tie'
GROUP BY winner
ORDER BY COUNT(winner)  DESC
LIMIT 3
```

Winner	COUNT(winner)
Brazil	71
Italy	45
Argentina	44

ו. בעקבות הפלט שקיבלנו עבור השאילתא הקודמת, בו ראינו שברזיל מובילה בפער גם על המדינות החזקות, רצינו לדעת כמה גולים סה"כ הובקעו בכל משחק שבו ברזיל שיחקה.

```sql
SELECT matchid,SUM(home_team_goals+away_team_goals) OVER(PARTITION BY matchid) AS 'average_goals_in_game'
FROM WorldCupMatchess
WHERE winner = 'Brazil' AND (home_team_name = "Brazil" OR away_team_name = "Brazil");
```

חלק 4

קישור לגיטהאב: https://github.com/NoaBurg/BI-Project