**Homework 3+4 – Business Intelligence**

**קבוצה:**  Q

**חברי הקבוצה:** גלית אביוב 204862726

קורל נמימי 318330693

דניאל רביב 206812851

ענבר נדלר 208640110

**חלק 1 – STTM**

 **מימוש תהליך ה-ETL Pipeline**

**סכמה ויזואלית אשר מציגה את ה- STTM**

Information Source

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **RK**  מס' סידורי | Age  גיל | **GID**  מס' סידורי משחק | MP  מס דקות משחק לכל משחק | FG%  אחוז יעדי שדה | PTS  נקודות למשחק |

Target Structure

|  |  |
| --- | --- |
| **RK**  מס' סידורי | Age  גיל |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **RK**  מס' סידורי לשחקן | **GID**  מס' סידורי משחק | Age  גיל | MP  מס דקות משחק לכל משחק | PTS  נקודות למשחק | FG%  אחוז יעדי שדה |

|  |  |
| --- | --- |
| **GID**  מס' סידורי משחק | MP  מס דקות משחק לכל משחק |

**חלק 2 – Data mining techniques**

1. **מהו תהליך ה- KDD אשר אתם מבצעים? יש לפרט אם בחרתם בשיטת Predictive או Descriptive.**

בשלב הראשון יש להבין מהו הידע הנדרש על מנת להגיע למטרות השונות. לאחר מכן, נייצר Target Data Set ע"י בחירת מערך נתונים, נבצע ניקוי ועיבוד מקדים של המידע ולבסוף נבצע Transformation למידע (אם יש צורך). בשלב הבא נבין מהו סוג התהליך שעלינו לבצע, במקרה שלנו נשתמש בשיטת Descriptive ונבחר באלגוריתם Clustering - נחלק לקבוצות ע"י טווח גילאים כאשר בכל קבוצה נבדוק את דרגת החשיבות ומכאן נסיק תובנות אודות הדפוסים השונים.

1. **מה הטכניקות אשר תממשו על אוסף הנתונים?**
   1. **יש להתייחס לסוגי הנתונים**
   2. **יש להראות שני תרחישי Use case עם דוגמה מספרית (ניתן להמציא ערכים אקראיים לצורך ההמחשה בלבד)**

האלגוריתם בו נשתמש על מנת לזהות האם ניתן לחלק את השחקנים לקבוצות גיל מסוימות אשר יסבירו את דרגת חשיבותם בקבוצה בה הם נמצאים הוא אלגוריתם KNN K-Near Neighbors.

תרחישי שימוש:

1. נניח כי אנחנו אנליסטים בקבוצת כדורסל בישראל ומטרתנו לבחון איזה שחקן עלינו להביא – האם להביא שחקן צעיר בעלות נמוכה יחסית כאשר אנחנו יודעים שנצטרך להשקיע זמן ומאמצים, או האם להביא שחקן בוגר בעלות גבוה כאשר אנחנו יודעת שנקבל את התפוקה ברמה המיידית. השחקן הראשון בן 30 קולע בממוצע 25 נק' למשחק ומשחק 35 דק' בממוצע במשחק. עלות השחקן – 2 מיליון ₪. השחקן השני בן 21, קולע בממוצע 12 נק' למשחק ומשחק 17 דק' בממוצע. עלותו 400 אלף ₪. כעת עלינו לבחור מדגם של שחקנים מקבוצות שונות בליגה, להפעיל את האלגוריתם המבוקש ולהחליט כמה הגיל משפיע על החשיבות השחקן במונחי כסף וכך נקבל החלטה.
2. נניח כי אנחנו סוכני שחקנים אשר מעוניינים להגדיל את כמות השחקנים שברשותנו. במסגרת קיבולת השחקנים שביכולתנו להוסיף, עלינו לדעת באיזה טווחי גילאים נרצה להתמקד – האם נרצה להתמקד בשחקנים צעירים שכעת הרווח עליהם יהיה נמוך אך הם בעלי פוטנציאל גבוה להמשך, או שנרצה להתמקד ב"כוכבים" שכבר הוכיחו את עצמם, והרווח עליהם יהיה מידי אך לא לאורך זמן. השחקנים המבוגרים בממוצע יכולים להכניס לחברה כ-70 אלף ₪ אך בשל גילם הגבוה סכום זה יכול להיות הסכום היחיד אותו נקבל חלק מהעסקה. השחקנים הצעירים בממוצע יכולים להכניס לחברה בממוצע כ- 10 אלף ₪ אך במידה והפוטנציאל שלהם יתממש, סכום זה יכול לעלות בצורה אקספוננציאלית. כעת עלינו לבחון לכמה קבוצות אלגוריתם KNN יחלק מדגם של שחקנים בהתאם לגילם וליכולות המשחק שלהם וכך לקבל החלטה המבוססת ע"י הנתונים.
3. **הגדירו מדד דמיון עבור ה DW שבניתם. ניתן להשתמש במדדי דמיון אשר ראינו בכיתה לחילופין להחליט על מדד דמיון שרלוונטי עבור העבודה שלכם. יש להסביר בקצרה מדוע מדד זה אכן מהווה דמיון בין הנתונים.**

הנתונים הינם נתונים נומריים ולכן נשתמש במדד הדמיון Euclidean Distance עבור ה DW בנינו. הקרבה מוגדרת על ידי ההפרש (המרחק) לאורך הסקאלה של כל משתנה, המומר למדד דמיון. מרחק זה מוגדר כ- מרחק אוקלידי . בעזרת מדד זה יעזור לנו לבחון את האשכולות וכך לקבץ נתונים דומים לאשכולות אלו.

1. **עבור כל אחת מהשאלות העסקיות, יש לנסח 2 השערות: H1, H0 ולהציע דרך כיצד להחליט איזו השערה לקבל ומדוע.**

**שאלה עסקית 1-**

האם אפשר לנבא את כמות הדק' ששחקן ישחק לפי מס' נתונים סטטיסטיים: כמות נק' ששחקן קלע לסל (PTS) , כמות הנק' ששחקן קלע בתוך קשת ה-3 (FG) ?

הגדרת השערות:

h0: אין קשר בין נתונים סטטיסטים (המפורטים מטה) לבין כמות הדק' שישחק

h1: יש קשר בין נתונים סטטיסטים (המפורטים מטה) לבין כמות הדק' שישחק

נתונים סטטיסטים - כמות נק' ששחקן קלע לסל (PTS) , כמות הנק' ששחקן קלע בתוך קשת ה-3 (FG)

* נשתמש במבחן F על מנת להחליט איזו השערה לקבל, וזאת מכיוון שמבחן זה בוחן השערות התלויות במספר פרמטרים.

**שאלה עסקית 2-**

האם אפשר לחלק את בסיס הנתונים לקבוצות גיל שונות כאשר לכל קבוצת גיל קיימת דרגת חשיבות שונה בקבוצה בה הם נמצאים?

\*דרגת חשיבות מחושבת ע"פ שילוב של כמות הדק' ששחקן קיבל וכמות הנק' שקלע במשחק.

הגדרת השערות:

h0: אין הבדל בהצלחה בין הגילאים המבוגרים לצעירים

h1: שחקנים בגילאים בוגרים מצליחים יותר משחקנים בגילאים צעירים

* המדגם שלנו הוא גדול יותר מ-30 ולכן נשתמש במבחן Z על מנת להחליט איזו השערה לקבל.

**חלק 3 – שאילתות**

select

case when PTS >15 and MP >22 then 'Important' else 'Less important' end as player\_value,

avg(Age) as avg\_age,

count(\*) as num\_of\_players

from dbase.fact

group by 1

order by 2;

select case when Age>25 then 'Adults' else 'Young' end as Age\_group,

avg(MP) as avg\_minutes\_per\_player,

avg(PTS) as avg\_points\_per\_player

from dbase.fact

group by Age\_group

order by avg\_points\_per\_player desc;

select Age,

MP,

PTS,

avg(FG%) over(partition by Age)

from dbase.fact

order by PTS desc;

select Age,

PTS,

count(\*) over(partition by PTS) as num\_of\_players,

avg(MP) over (partition by PTS)as avg\_minutes\_per\_player

from dbase.fact

group by 1,2;

select MP,

PTS,

Age,

FG%,

ratio\_to\_report(PTS) over(partition by Age) as PTS%

from dbase.fact

order by Age;

select

MP,

avg(age),

avg(PTS) as avg\_points\_per\_minutes

from dbase.fact

group by MP

order by 3;

**חלק 4 – ניהול גרסאות**

קישור לפרויקט GitHub: