Homework 3 – Business Intelligence

קבוצה: B

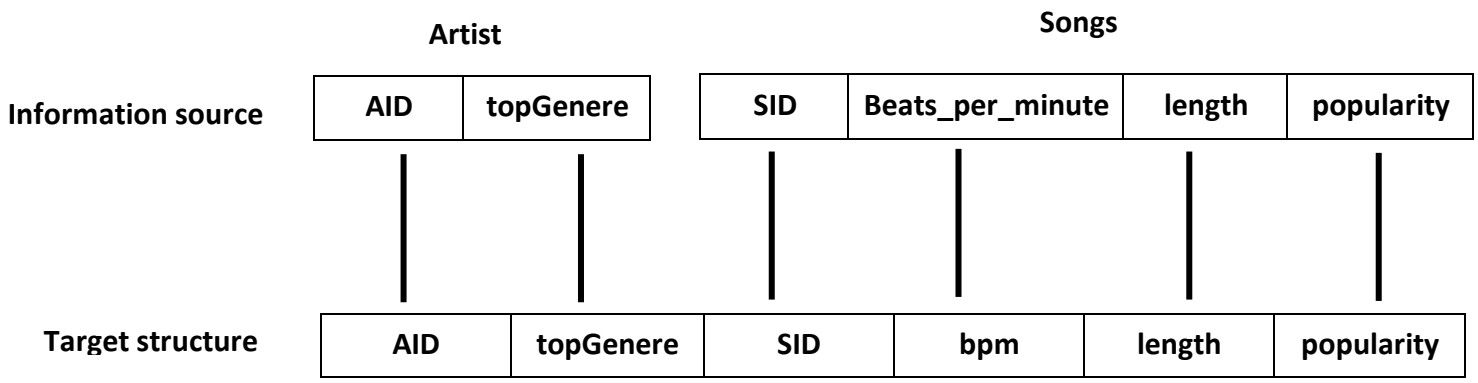
מגישות:
מיתר ירון, ת.ז.- 204263818
הדר פרץ, ת.ז.- 315970020
עדי קדם, ת.ז.- 312234818

חלק 1 – STTM

1. תהליך STTM

Version	Source Table	Source Column	Source Type	Transformation	Target Table	Target Column	Target Type	Target Length	Default Value	Note
1	Artist	AID	int	NOT NULL	dsArtist	AID	int			
1	Artist	AID	int	NOT NULL	songsFact	AID	int			
1	Artist	topGenere	VARCHAR		dsArtist	topGenere	VARCHAR	50	NoGenere	
1	Artist	topGenere	VARCHAR		songsFact	topGenere	VARCHAR	50	NoGenere	
1	Songs	SID	int	NOT NULL	dsSongs	SID	int			
1	Songs	SID	int	NOT NULL	songsFact	SID	int			
1	Songs	beats_per_minute	int		dsSongs	bpm	int		-1	
1	Songs	beats_per_minute	int		songsFact	bpm	int		-1	
1	Songs	length	int		dsSongs	length	int		-1	
1	Songs	length	int		songsFact	length	int		-1	
1	Songs	popularity	int		dsSongs	popularity	int		-1	
1	Songs	popularity	int		songsFact	popularity	int		-1	

2. סבימה ויזואלית של STTM



חלק 2 - data mining techniques

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

תהליך הKDD (גילוי ידע ממסד נתונים) הוא תהליך הכרחי למענה על שאלות עסקיות.
תהליך זה כולל בתוכו 5 שלבים:
Selection – בחירת בסיס הנתונים המתאים. אנו בחרנו בבסיס נתונים של spotifiy המכיל בתוכו את כל הפרמטרים שאנו זקוקים על מנת לענות על השאלה העסקית שלנו.
Preparation – שלב הכנת וניקוי הנתונים ע"י בחירת מערך נתונים הכולל בתוכו את העמודות הרלוונטיות בלבד וטיפול בערכים חסרים.
Transformation – ביצוע שינויים רלוונטיים על הנתונים הקיים. לדוגמה שינוי עמודה קטגוריאלית לעמודה נומרית.
Data Mining – בחירת שיטה Predictive או Descriptive בהתאם לשאלה העסקית ובחירת אלגוריתם מתאים.
בפרויקט שלנו בחרנו בשיטת Predictive מכיוון שנרצה לענות על השאלה האם שיר פופולרי או לא? - בהתאם לעמודת מטרה.

בפרויקט שלנו נשתמש בשיטת Predictive במטרה לחזות האם שיר עם מאפיינים מסוימים הוא פופולרי או לא.

2. נשתמש בטכניקת Nearest Neighbors, על סוגי נתונים נומריים וקטגוריאליים.
דוגמאות לתרחישים בשימוש בטכניקה זו:
א. חברת תקליטים רוצה לאתר את השירים הכי פופולריים של האומנים החתומים בה על מנת להכניסם למצעד הלהיטים. מחסן הנתונים הזמין לנו:

BPM	Genre	Year	Song	Artist
171	canadian contemporary r&b	2020	Blinding Lights	The Weekend

אלגוריתם ה- KNN יחפש את התצפיות הקרובות ביותר לנתונים שלנו ויסווג את מידת הפופולריות בהתאם.

ב. מפיק מוזיקלי רוצה לסווג שירי פופ לפי פופולריות כדי לדעת באיזה קצב כדאי לו להשתמש בהפקת שירים עבור זמרת פופ חדשה וכדי להחליט מה יהיה אורכם האידיאלי. מחסן הנתונים הנתון:

Length	BPM	Song
200	95	Watermelon Sugar

אלגוריתם ה- KNN יחפש את התצפיות הקרובות ביותר לנתונים שלנו ויסווג את מידת הפופולריות בהתאם.

3. מדד הדמיון שנגדיר הוא Sorensen-Dice, משום שהוא רלוונטי ל-DW המכיל עמודות נומריות וקטגוריאליות.

4. **שאלה עסקית ראשונה**:

השערת H0:

שירים בסגנונות פופ בעלי BPM מעל 110, שאורכם פחות מ – 180 שניות פופולריים באותה מידה של שירים בסגנונות ראפ בעלי מדדים הפוכים.

השערת H1:

שירים בסגנונות פופ בעלי BPM מעל 110, שאורכם פחות מ – 180 שניות יותר פופולריים מאשר שירים בסגנונות ראפ בעלי מדדים הפוכים.

שאלה עסקית שניה:

השערת H0:

לשירים פופולריים לא קיימים מאפיינים משותפים.

השערת H1:

לשירים פופולריים קיימים מאפיינים משותפים.

על מנת להחליט איזו השערה נקבל בכל אחת מהשאלות העסקיות, נשתמש במבחן F משום ששתי ההשערות תלויות במספר פרמטרים (סגנון, BPM ואורך).

חלק 3 – שאילתות SQL

```
Select SID, topGenere, bpm, AVG(bpm)

OVER(PARTITION BY topGenere) as AvgBpm

From fact


Select SID, topGenere, bpm, COUNT(SID)

OVER(PARTITION BY topGenere) as CountSongs

From fact


Select SID, popularity, AVG(length)

OVER(PARTITION BY topGenere ORDER BY popularity) as AvgLength

From fact
```

```
Select SID, AID, popularity, Count(SID)
OVER(PARTITION BY AID ORDER BY popularity) as CountSongs
From fact

select case when bpm>110 then 'fast' else 'slow' end as Song_speed,
avg(popularity) as avg_popularity,
length
from songsFact
group by Song_speed
order by length desc;

select case when bpm >110 and length < 180 then '1' else '0' end as suspect_as_Hit,
popularity,
from songsFact
group by suspect_as_Hit
order by popularity;
```