# **Homework 3 Group D – Business Intelligence**

מגישים:

דור דלמדיגו – 205973175

עדן גבריאל – 318868700

חנן חן – 301647814

## **חלק 1 – STTM**

סכמת STTM ממקורות המידע אל מחסן הנתונים:

Diagram

Description automatically generated

## **חלק 2 – data mining techniques**

1. בשביל שאלת המחקר המונחית של הפרויקט, נשתמש בתהליך ה-KDD מסוג **Predictive.** מכיוון שאנו רוצים לעשות פעולות שינבאו מה יש לנו בדטה.

בשביל שאלת המחקר הלא מונחית,נשתמש בתהליך ה-KDD מסוג **Descriptive.** מכיוון שעל פי שאלת המחקר שלנו, אנו רוצים לעשות פעולות שיסבירו מה יש לנו בדטה.

1. טכניקות למימוש:
   1. בשביל תהליך הניבוי נשתמש בשיטת רגרסיה לוגיסטית מכיוון ששאלת המחקר המונחית היא בעיה בינארית והשיטה מתאימה לערכים הקטגוריאליים כמו סוג הסרט. בנוסף, נשתמש גם באלגוריתמים כמו עצי החלטה ו-naïve base שמתבססים על חוסר ודאות.

בשביל תהליך ההסברה, נשתמש בשיטת Clustering ונמדוד אותה בעזרת מדד silhouette שבודק אם יש קשר בין האיברים בקלאסטר. או בעזרת Association Rules.

* 1. Use case ראשון יהיה לנסות לנבות, על פי עץ החלטה, כאשר האיבר הראשון יהיה הוא התכונה בעלת הinformation gain- הכי גבוה, או ג'יני/אנט' הכי נמוכים, האם סרט מסוג דרמה מעל אורך 120 דקות שהתפרסם בשנים 2019- 2016 יצליח יותר בקופות מאשר סרט פשע מתחת ל 120 דקות באותן שנים?

Use case שני יהיה לנסות להסביר, בלי מידע מקדים על הנתונים, נרצה לחלק לקבוצות כך שהאיברים בכל קבוצה דומים זה לזה. ולבדוק מה מאפיין סרטים שהדירוג שלהם מעל 8.5. למשל, האם מה שמאפיין אותם זה אורך סרט של 100 דקות או שילוב של במאי ושחקן מסוים.

1. אין לנו דטה רציף. אלא ערכים קטגוריאליים או בינאריים ולכן נשתמש במדד דמיון **Jaccard**. ניקח 2 עמודות או יותר שנרצה לבדוק, והמדד יבדוק כמה פעמים האיברים דומים זה לזה חלקי האיחוד של כולם. (חיתוך חלקי איחוד). נותן לנו יחס בין X ל-Y.
2. שאלת מחקר מונחית:

H0 – סרטי דרמה מעל 120 דקות שהתפרסמו בשנים 2016-2019 לא מרוויחים יותר בקופות מאשר סרטי פשע מתחת ל120 דקות או שמרוויחים אותו דבר.

H1- סרט מסוג דרמה מעל אורך 120 דקות שהתפרסם בשנים 2019- 2016 יצליח יותר בקופות מאשר סרט פשע מתחת ל 120 דקות באותן שנים.

נעשה זאת על ידי חלוקה של סרטי קומדיה וסרטי פשע, נעשה ממוצע רווחים ונבדוק. בגלל שהדטה שאני לומד עליו יש לו כבר מראש את התשובות.

שאלת מחקר לא מונחית:

H0 – אין מאפיינים משותפים לסרטים שהדירוג שלהם מעל 8.5.

H1- קיימים פרמטרים משותפים המאפיינים סרטים שהדירוג שלהם מעל 8.5.

נעשה זאת על ידי מבחן כי בריבוע כי הוא משמש אותנו לערכים קטגוריאליים על פי התדירות שהם מופיעים.

## **חלק 3 – שאילתות**

Application

Description automatically generated with medium confidenceשאילתות לשאלת המחקר המונחית:

1)

A picture containing table

Description automatically generated

2)

שאילתות לשאלת המחקר הלא המונחית:

Table

Description automatically generated3) ) הסרטים בעלי דירוג 8.5 ומעלה והמאפיינים שלהם.

Table

Description automatically generated4) מאפייני הסרטים מחולקים למחיצות של במאי + שחקן.

Text

Description automatically generated5) KPI שהוגדר:

6) הסרט הכי מרוויח בכל שנה בתקופת הזמן המוגדרת: (ב1992 לא היה סרט שעמד בתנאים).

Table

Description automatically generated

## **חלק 4 – GitHub**

קישור לפרויקט: https://github.com/Dalma1995/MoviesProject.git