פרויקט קורס כריית מידע בנושא: Netflix DB



בהגשת: שיר אמסלם 313585903

אלמאזה פארס 315734285

סול קרנכל 208950568

איתי לוי 205641426

בהנחיית: עומר כספי

# תקציר:

המהפכה הדיגיטלית בשידורי הטלוויזיה הביאה עמה נגישות ותפוצה אדירה בעולם. צריכת תכני המדיה נמצאים בעלייה מתמדת, והתעשייה פונה לשימוש בשירותי סטרימינג. בפרויקט זה אנו נתמקד בנטפליקס אשר נחשבת כיום לחברת הטלוויזיה בתשלום הגדולה בעולם, ולשחקן הראשי בזירה העולמית של תעשיית הבידור.

עקב מגוון רחב של תכנים, עלה הצורך לבסס וליעד תכנים למגוון רחב של צרכנים; בעזרת IMDb מאגר מידע המאפשר לכל צופה לכתוב ביקורת צפייה, ניתן להתאים תכנים חדשים לצופה הנלהב.

בעבודה זו, נבחן אילו מאפיינים מניבים דירוג גבוה יותר לתכנים השונים – מהי השפעת הבמאי על דירוג התוכן? מהי השפעת השחקן הראשי? האם מדינות הפקה שונות זוכות לדירוג גבוה באופן עקבי או שמא אין לכך משמעות? מאפיין נוסף הוא אורך התוכן – האם ישנה השפעה של אורך התוכן על הדירוג?

באמצעות ממצאים אלו ניתן יהיה לדעת אילו גורמים משפיעים על דירוג גבוה יותר. ככל שתוכן זוכה לדירוג גבוה יותר, כך תוכן עתידי בעל מאפיינים דומים, צפוי להצלחה.

# מבוא:

## מידע כללי:

**NETFLIX**

נטפליקס היא חברה אמריקאית אשר הוקמה בשנת 1997 בלוס גאטוס קליפורניה. היא החלה את דרכה כחברה להשכרת תקליטורי DVD, ובשנת 2007 התחילה לתמוך בהפקות תכנים.

החברה התרחבה לפעילות בין-לאומית בשנת 2010 לקנדה, אמריקה הלטינית והאיים הקריביים. בשנת 2011 תפסה נטפליקס את המקום הראשון בדירוג התעבורה באינטרנט בצפון אמריקה, ובכך עברה את ביטורנט. הדבר נתפס כמשמעותי במאזן בין תכנים חוקיים בתשלום תוך שמירת זכויות יוצרים לבין שימושים חינמיים/פיראטיים. ובשנת 2013, נטפליקס החלה בהפקת תוכן מקורי משלה, כאשר סדרת המקור הראשונה שלה הייתה "בית הקלפים".

בינואר 2016, נטפליקס הרחיבה את פעילותה ל-190 מדינות, בין היתר לישראל, ובאותה שנה החלה להציע כתוביות בעברית לחלק מהתכנים בשירות. לאחר מכן, ביולי 2017, החלה נטפליקס להציע דיבוב לעברית לחלק מתכני הילדים, והציגה ממשק מלא בעברית כולל פוסטרים בעברית לתכנים בשירות.

נכון לשנת 2019, נטפליקס מעסיקה כ-8600 עובדים, ודיווחה על מחזור הכנסות של 20.1 מיליארד דולרים, כאשר ההכנסות נטו הן של 1.9 מיליארד דולרים. נטפליקס נחשבת לרוכשת תכני וידאו הגדולה ביותר בעולם. יש להם 183 מיליון מנויים ברחבי העולם, כאשר 69 מיליון מתוכם בארצות הברית.

כיום, עיקר פעילות החברה הוא במתן שירותי צפייה ישירה בשיטת Over the Top של ספריית סרטים ותוכניות טלוויזיה, שחלק מהם תוכן מקורי, תמורת מינוי חודשי.

לצורך התאמת תכנים, נטפליקס פועלת על פי תאוריית הזנב הארוך של אנדרסון. כלומר בזמן שרוב חנויות הסרטים הן בעלות מקום מדף מוגבל, לנטפליקס (הפועלת על טכנולוגיית ענן) יש אפשרות לאחסן אינספור כותרים, הניתנים להתאמה עבור כל משתמש.   
ובזכות שירות ההמלצות של נטפליקס, המפעילה מנגנון חיזוי העדפות צפייה של המשתמשים, הצופים נהנים ממגוון רחב של סרטים. ובנוסף נחשפים לתכנים חדשים התואמים להם; זאת בשל האפשרות שלהם לדרג את התכנים השונים שצפו בהם – דירוג חיובי או שלילי (לעומת דירוג מספרי של מסד הנתונים IMDb).

**מילה על IMDb**

Internet Movie Database - IMDb (בסיס נתוני סרטים אינטרנטי) אתר אינטרנט בבעלות אמזון. המכיל בסיס נתונים מקוון של מידע על סרטים וסדרות טלוויזיה, החל מפרטים בסיסיים כמו שמות השחקנים, הבמאי, צוות הפקה, תקציר עלילה וביקורות, ועד פרטי טריוויה, שגיאות המשכיות בסרט, פסקול וקשרים לסרטים אחרים. בנוסף לנתונים על סרטים ומשחקי וידאו, יש באתר חדשות על סרטים וסדרות טלוויזיה, וסקירת אירועים כדוגמת טקס האוסקר.

IMDb הוא אתר חינמי, וכן אין צורך להירשם אליו על מנת לקבל גישה לנתונים שבאתר. משתמשים רשומים באתר יכולים לדרג את הסרטים, הסדרות והפרקים לפי טעמם האישי בסולם של 1 עד 10. כל משתמש רשאי להצביע לכל סרט פעם אחת בלבד, כאשר הצבעה נוספת תבטל את ההצבעה הקודמת. ההצבעות מסוננות ומשוקללות בדרכים שונות על מנת ליצור ממוצע משוקלל.

## ה-Data Base הנבחר, והשימוש בו:

מאגר ה-DB שליקטנו מכיל מידע על תכני סטרימינג שונים שנטפליקס מציעה בספרייה שלה למנויים, ואת הדירוג שהתכנים קיבלו במסד הנתונים המקוון IMDb.

עבור כל תוכן נשמר המידע הבא:

* מספר מזהה של התוכן נומינאלי (int)
* סוג התוכן נומינאלי (string)
* שם התוכן נומינאלי (string)
* במאים /במאיות נומינאלי (string)
* צוות השחקנים (קאסט) נומינאלי (string)
* מדינות הפקה נומינאלי (string)
* תאריך הוספת התוכן לשירות נטפליקס נומינאלי אורדינלי (date)
* שנת הוצאה נומינאלי אורדינלי (date)
* דירוג IMDb נומינאלי אורדינלי (int)
* משך זמן התוכן נומינאלי (int)

כפי שניתן להבחין טרם השימוש במאגר מידע זה, קיימים בו חוסרים:

תאים ריקים או בעלי תוכן "NONE"/"BAN", הנחשבים ריקים גם הם.

ישנן עמודות הכוללות מידע רלוונטי, שאותן נרצה לפצל:  
צוות השחקנים- הפיצול הרצוי הוא השחקן הראשי (השחקן הראשון בכל שורה) מול שאר השחקנים, אותם נשליך.  
עמודות נוספות אותן נרצה לפצל הן במאים, מדינות וז'אנרים.  
לא ניראה בסקירה ראשונית שיש חשיבות לתאריך ושנת הוצאת התוכן. אולם נפצל גם את תאריך ההוצאה בשל כפילות בשנה- כאשר נסיר בפיצול זה את השנה. נשמור את השנה בעיקר על מנת לבדוק שאין קורלציה בין יותר זמן קיים בשירות לדירוג יותר גבוה.

בניגוד לכך ישנו טור העלול לגרום להתאמת יתר של התוצאות שלנו במודל, כותרת התוכן. כלומר בטור זה לא יבוצע השימוש העיקרי, כך שנשמיט אותו ואת טור תיאור התוכן. בנוסף הטור אינדקס אינו רלוונטי לשימושינו ואינו מספק מידע כלל- גם אותו נשמיט. ונשקול במהלך העבודה להסיר את התאריך המדויק כי לא ניראה כי קיים בו צורך.

כמו כן, במאגר קיימים תאים שתוכנם אינו ברור (ג'יבריש). תאים אלו מכילים מלל בשפה זרה של מדינות ההפקה. למשל בהפקת מקור מיפן, תאים מסוימים עלולים להיות כתובים בשפת המקור שהיא יפנית, ולכן מופיעים ב- DB כג'יבריש. מצב זה קורה מאחר ו- DB זה אינו תומך Unicode.

## הזנחות:

במאגר המידע, בעמודה של צוות השחקנים (קאסט) לא מצוין בבירור מידע אודות השחקנים הראשיים בתוכנית/סרט. לכן, עקב חוסר מידע זה, אנו נבצע הנחה כי השחקן הראשון שמצוין בכל שורה בעמודה זו, הינו השחקן הראשי בתוכנית/סרט ונבצע פיצול לפי הנחה זו.

ישנם תאים אשר מופיעים כג'יבריש, מדובר בתאים שתוכנם הוא בשפה זרה. את חלקם נמחק בשל חוסרים בעמודות אחרות, כאשר את השורות המלאות נשאיר למרות אי היכולת של פיענוח UNICODE.

חשיבות צוות הבמאים, המדינות ועוד הינו לפי סדר הופעתם ברשימה. כך שבעט הצורך נתעלם מהם.

## מטרות הפרויקט:

* סיווג מאפיינים שונים של תוכן לפי סיכוייהם לקבל דירוג גבוה.
* מציאת שילוב המאפיינים אשר מניבים את דירוג הצופים הגבוה ביותר.
* בדיקת חוקיות בין אורך התוכן לבין הדירוג הניתן על ידי הצופים.

עבור כל אחת מן המטרות האלה נבצע כריית המידע, במטרה להגיע למסקנות בהתאם ולייעול מאגר המידע.

## הנחות:

* ככל שערך במאפיין מסוים מופיע ב-DB פעמים רבות יותר, ככה הדירוג שלו יהיה גבוהה יותר.
* כאשר משך סרט עולה על 90 דקות, הדירוג הנצפה יהיה נמוך.
* תכנים ישנים יותר יכולים לקבל דירוג גבוהה יותר בשל כך שהיו זמינים זמן רב יותר למשתמשים.

## קישור למאגר: <https://www.kaggle.com/sarahjeeeze/imdbfile> (29.4.2020)

## מטודות:

* Python 3.7
* CMD windows version
* Excel
* Weka

# Pre-Processing:

## Data Reduction and Data Transformation:

### מחיקת טורים ללא משמעות עבור הניתוח:

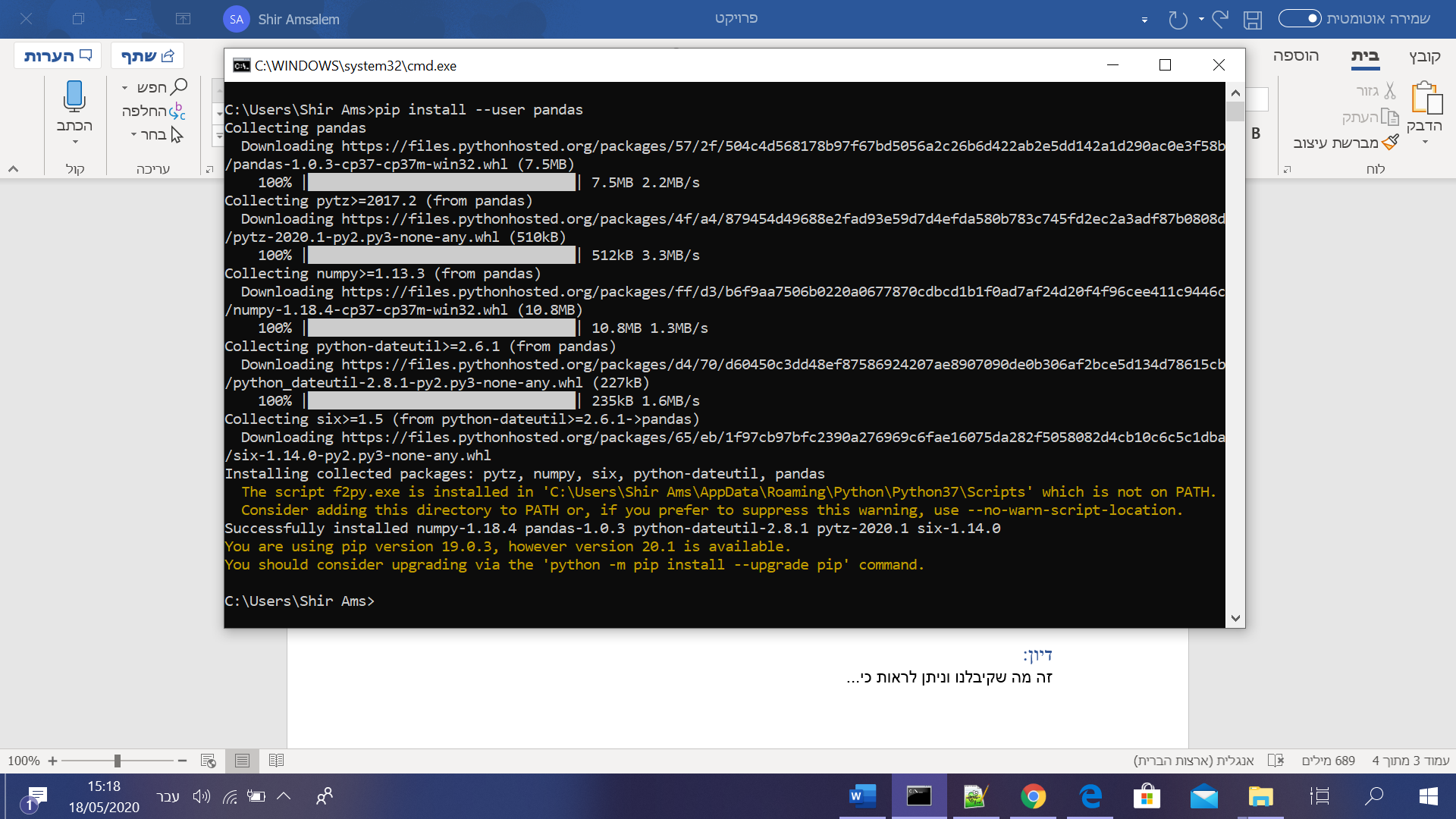
בהתאם לכתוב במבוא, נבצע צמצום של טורים שאינם רלוונטיים לאנליזה שלנו.  
בעזרת Excel ניתן למחוק את הטור הלא רלוונטי של "אינדקס".  
כך שלמעשה ה-DB שלנו ניראה ככה:

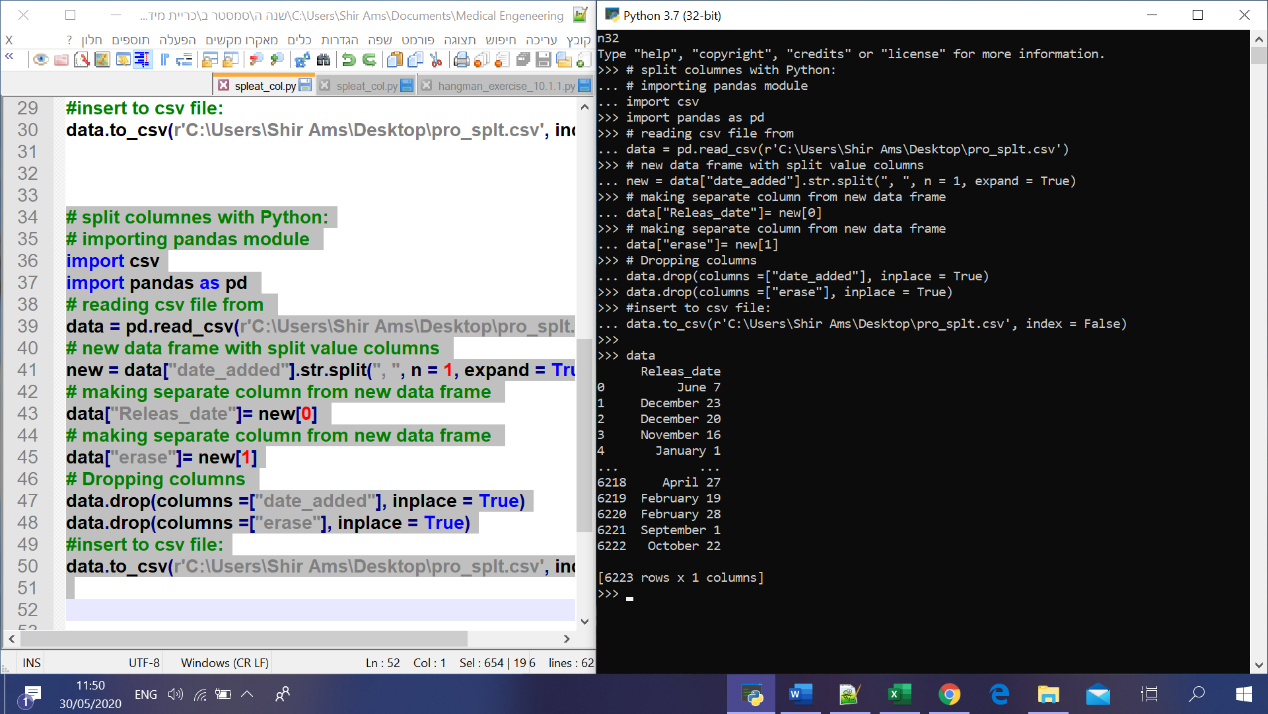


לאחר מכן נשמיט את הטורים תיאור וכותרת גם כן.

עכשיו נרצה לזרוק את החלק בעמודה של "תאריך הוספה" אשר מכיל את השנה, המצוינת בטור אחר (קורלציה). לצורך כך יש לעקוף את גישת ה-Administer ל-CMD של Windows ובכך להוריד את הספרייה: pandas לשימוש בסביבת Python:

בתמונה הבאה ניתן לראות כי הורדת הספרייה נעשתה באופן תקין ונשמרה בספריות של גרסה 3.7 של פייטון, בהקלדת השורה: pip install --user pandas.



לאחר מכן נכניס את הקוד הבא (אותו ניתן למצוא בקובץ: "split-col.PY") לצורך פיצול עמודה מסויימת למספר עמודות, שמירה בקובץ, וצפיה ב-DATA לולידציה: 

קוד זה בוצע עבור תאריך הוספה, לצורך שימור חודש ויום; עבור השחקנים, אותה פיצלנו לשחקן ראשי ומחקנו את השאר; ועבור הטורים "במאים", "מדינות" ו"ז'אנרים"- פיצלנו ושמרנו את שלושת הערכים הראשונים, בהנחה שהם הרלוונטיים ביותר. את הקובץ מסוג CSV עליו עבדנו, את הקוד עבור כל אחד מהפיצולים, וקובץ וורד המתאר לעומק על כל פיצול והקוד שלו- זמינים בקבצי הפרויקט בתיקיית "spleat\_Cols".

ניתוח הקוד בקצרה:

* תחילה העתקנו את הטור המיועד לפיצול לקובץ בשם "pro-splt", שמרנו וסגרנו.
* יבאנו את הספריות הרלוונטיות:  
  CSV- לצורך גישה ועריכת קובץ מסוג זה עם כל המטודות הקימות לו.  
  Pandas- לצורך פיצול הטורים (אותו הורדנו במיוחד ע"י CMD).
* יצרנו משתנה מטיפוס data frame הקורא את המידע בקובץ ה-CSV של ה-DATA שלנו.
* הגדרנו משתנה הכולל פיצול של המידע לשני טורים (n=1, expand=True) הפיצול מתבצע לאחר הגעה לאינדקס של התו\ מחרוזת המצוינת בתוך המירכאות.
* ואז לכל טור במשתנה data בחרנו איזה חלק של הטור ילך לכל צד.
* כאשר היינו צריכים יותר מפיצול אחד, ביצענו את 3-4 פעם נוספת.
* לאחר מכן השלכנו את הטורים הלא רלוונטיים.
* לפני סגירת סביבת העבודה, רשמנו את שם הארגומנט וצפינו בתוצאה הסופית של הפירוק.

להלן התוצאה: 

נציין שוב כי את פירוט הפיצול המלא, ניתן לראות בתיקיה: "split\_Cols".

### מחיקת טורים בעלי מידע מועט מידי לניתוח:

לצורך מציאת טורים מסוג זה, השתמשנו בפונקציות מובנות של Excel.  
עבור כל טור ביצענו את החישובים הבאים:

חישוב תאים ריקים בטור:

עבור טורים בעלי תאים בהם רשום bam/none:

חישוב אחוז התאים המלאים מתוך הטור:

כאשר:

* **X**- אות התור בקובץ ה-Excel
* **Y**- מיקום תוצאת החישוב תאים ריקים עבור כל טור.

יש לציין שעבור טורים מלאים לחלוטין, לא היה צורך בשימוש החישוב השני.

הטבלה הבאה כוללת את התאים החסרים עבור כל טור:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **title** | **release\_year** | **rating** | **duration** |  |
| 0 | 0 | 0 | 0 | **COUNTBLANK** |
| 0 | 0 | 0 | 0 | **BLANK [%]** |
| **Country2** | **Country3** | **Releas\_date** | **Main\_Acter** |  |
| 5314 | 5921 | 11 | 570 | **COUNTBLANK** |
| 85.22855 | 94.96391 | 0.176423 | 9.141941 | **BLANK [%]** |
| **Janner1** | **Janner2** | **Janner3** | **Country1** |  |
| 0 | 1450 | 3582 | 476 | **COUNTBLANK** |
| 0 | 23.25581 | 57.44988 | 7.634322 | **BLANK [%]** |
| **Director1** | **Director2** | **Director3** | **type** |  |
| 1969 | 5800 | 6184 | 0 | **COUNTBLANK** |
| 31.57979 | 93.02326 | 99.18204 | 0 | **BLANK [%]** |

ניתן לראות כי ברוב טוריי המידע הראשיים: דירוג, משך זמן, שנת הוצאה, כותרת, ז'אנר ראשי וסוג תוכן, הינם מלאים לחלוטין.  
טורים כדוגמת תאריך הוצאה, שחקן ראשי, מדינה ראשונה, וז'אנר שני, הינם טורים בהם החוסרים מינוריים, ועדיין ניתן לבדוק את ההשפעה של החומרים המצוינים בהם למיצויי המטרות.

קיים מצב גבולי עבור טור הבמאי הראשי. יש לנו בו חוסרים של 31.5%, מספר גבוהה יחסית. בגלל שמצב זה אינו עולה על 50% חוסרים, וניתן עדיין לכרות מידע למציאת מטרתנו, נשתמש בכלל ה-4265 שורות המלאות למציאת השפעת הבמאים הראשיים, בגפם ובצירופים שכיחים עם מאפיינים נוספים. במידה ולא נמצא מספיק מידע בנושא, נשליך את מטרה זו ונתמקד במאפיינים מלאים יותר כדוגמת שחקנ\ית ראשי\ת.

בנוסף, קיימים חוסרים רבים עבור הטורים: במאי משני ראשון ומשני שני, מדינה משנית ראשונה ושנייה, וז'אנר שלישי. אולם לא נרצה לזרוק את שורות אלה. הסיבה להשארתם בפעילותינו היא השימוש ככל מדינה (לדוגמא) כאינדיבידואל. כלומר בהמשך אנו נבדוק עבור כל מדינה מה ממוצע הדירוג שקיבלה. כך שלא משנה כמה מדינות מפיקות את אותו התוכן, עבור כל אחת מהן הדירוג ישכלל בממוצע המדינה.

לאחר שלב הדיסקרטיזציה הוחלט להשליך את הטור המכיל את יום העלאת התוכן לנטפליקס. מהסיבה המצוינת במבוא, ובשל העובדה שישנם יותר מידי תאריכים העלולים לגרום להתאמת יתר במודל שלנו. כלומר שילוב של חוסר רלוונטיות, ומודל לא מתאים לניבוי ערכים חדשים.

### השלמת תאים חסרים:

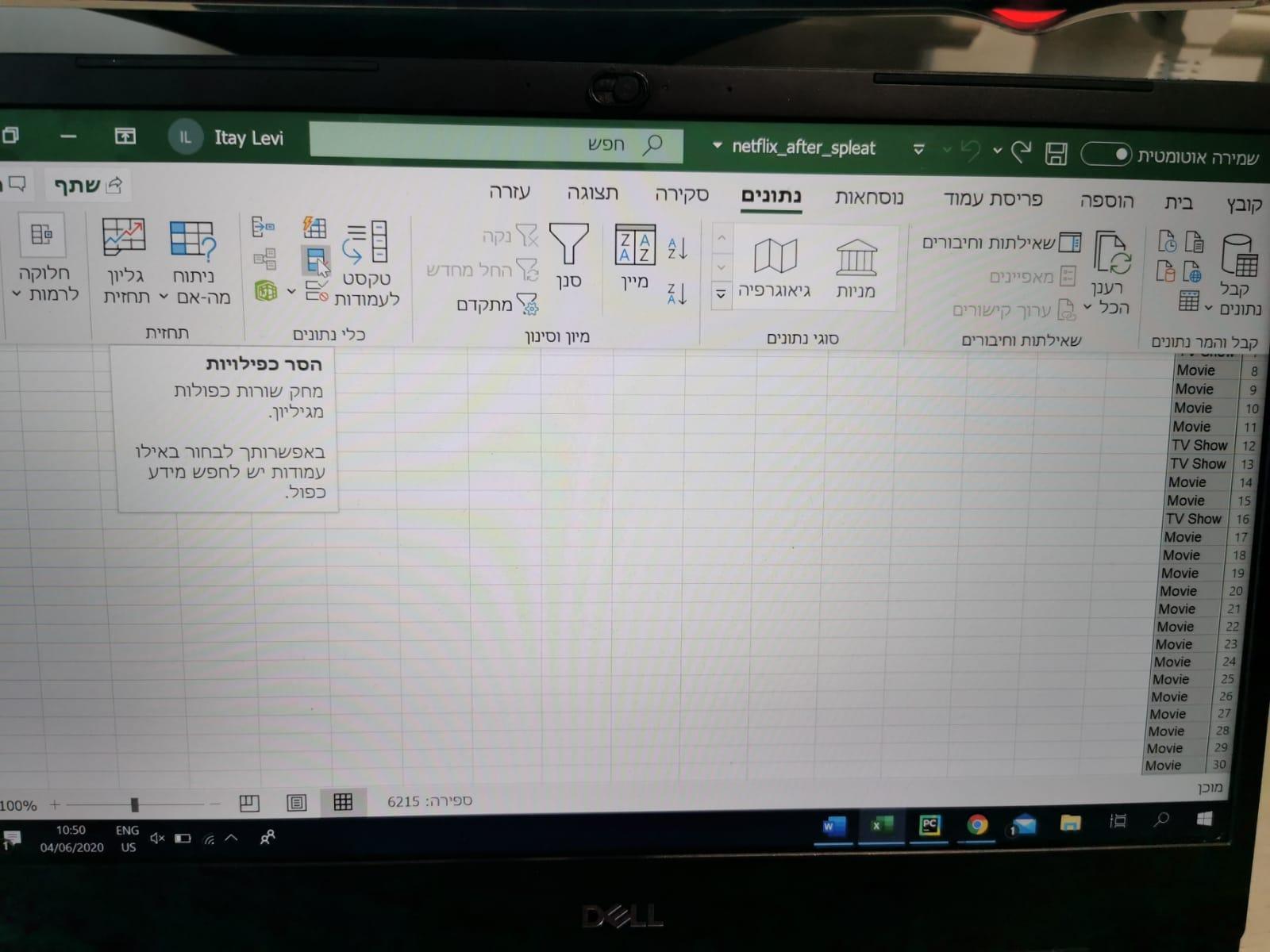
במידע מסוג זה לא ניתן להשלים חוסרים. הסיבה לכך היא שמדובר בערכים נומינליים, אינם ניתנים לחישוב. לכן כל תוכן חסר אינו ניתן להשלמה בעזרת השורות האחרות כי אין שום תלות וחוקיות נראית לעיין ביניהן.

### מחיקת שורות:

נעשתה מחיקת שורות בהן רוב המלל הוא בכתב ג'יבריש, חסרים בהם רוב המאפיינים הרלוונטיים לניתוח. היו סך הכל כ-20 שורות העונות על תנאים אלה אותן הסרנו בצורה ידנית ע"י מיון ב-EXCEL.

### דיסקרטיזציה:

עבור מידע המוגש בצורה נומינאלית, יש לבצע דיסקרטיזציה לצורך הפשטת האנליזה של ה-DB. לצורך כך השתמשנו ב-EXCEL וביצענו את הפעולות הבאות:

את העמודה הרצויה העתקנו לגיליון חדש (קובץ סיומת ".xlsx"), והסרנו כפילויות.

התוצר הינו טבלה המכילה עמודה עם ערכים ייחודיים, ועמודה עם מספור של כל אחד מהם.

לאחר מכן בגיליון הראשי, בוצע השימוש בפונקציה הבאה:

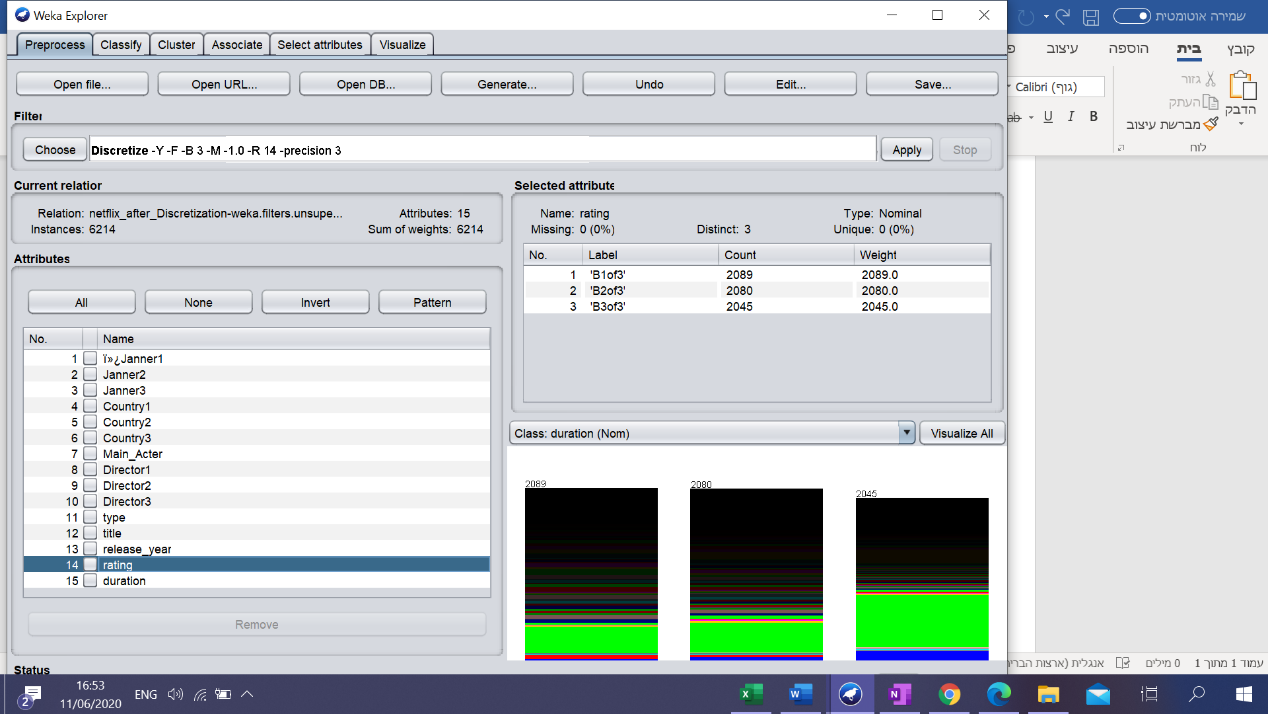
כאשר:

* **X**- זו העמודה של הערכים אותם אנו רוצים לעדכן.
* **z**- זה השורה האחרונה של הטבלה.
* **2**- מיקום עמודת הערכים האפשריים בטור הסופי
* **0**- **FALSE** כלומר יש לבצע שכלל הערכים יועברו בצורה מדויקת על פי הטבלה, ואין להשתמש בערכים לא מדויקים.

בתוך תיקיית "discretization" ניתן לראות את התוצאה של פעולות אלה בקובץ "netflix\_after\_Discretization", ופירוט נוסף הכולל תמונות מסך בקובץ וורד בשם "דיסקרטיזציה".

לצורך בחירת קבוצות עם ערכי **דירוג** חדשים, היה אילוץ לבצע בנוסף דיסקרטיזציה ע"י Weka.  
מטרת הדיסקרטיזציה במקרה זה היא להימנע מדירוגים בעלי משקל קטן מאוד, מה שגורר לערכי TN ו-FN גבוהים. במילים אחרות למנוע מצב שבו אחוזי התאמת המודל העתידי למציאות אינו תקין.

התהליך בוצע על תור זה כאשר הוא מוגדר כערך נומרי. בעזרת החלון שנפתח ע"י:  
Choose 🡪 Unsupervised 🡪 Attributes 🡪 Discretize  
ביצענו דיסקרטיזציה המחלק את הטור לשלוש קבוצות בעלות משקל יחסית זהה.

התוצאה הייתה טור נומינאלי בעל הערכים הבאים:

לאחר קבלת ערכים אלו, שינינו ב-EXCEL את הערכים למספרים 1-3.  
תיאור מלא של הפעולה ניתן למצוא בקובץ "דירוג" בתיקיה "discretization".

# תוצאות:

להלן התוצאות:

## אלגוריתמים בשימוש:

יש לציין כי עבור כלל האלגוריתמים יש לנו דגימות רבות, לכן אין לנו צורך ב-CROSSVALIDATION. כמו כן, עבדנו לרוב עם ML - SUPERVISED מאחר וכל המשתנים שלנו הם נומינליים וסיווגם ידוע מראש.

הסיבה שאין שימוש ב-SUPERVISED בהפיכת המשתנים נובעת מהעבודה שבתחילה כלל המשתנים היו מסוג נומרי. באמצעות ה-UNSUPERVISED ניתן לבצע תהליך של הפיכת המשתנים לנומינליים. רק לאחר מכן השימוש ב-SUPERVISED יהיה רלוונטי לביצוע.

להלן פירוט האלגוריתמים באנליזה שבוצעה:

### אלגוריתם לחישוב קורלציה:

לצורך חישוב קורלציה היה שימוש בקובץ EXCEL בו עבור כל אחת מהעמודות בוצעה בדיקה עם כל אחד מהטורים האחרים.

כאשר האלגוריתם לקורלציה עבור כל בדיקה בין התורים הוא:

* **ערכי X,Y –** הינם הטורים הרלוונטיים עליהם מבוצע האלגוריתם.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ז'אנר1** | **ז'אנר2** | **ז'אנר3** | **מדינה1** | **מדינה2** | **מדינה3** | **שחקן** | **במאי1** | **במאי2** | **במאי3** | **סוג** | **שנה** | **דירוג** | **אורך** |  |
| 1 |  |  |  |  |  |  |  |  |  |  |  |  |  | **ז'אנר1** |
| 14.0% | 1 |  |  |  |  |  |  |  |  |  |  |  |  | **ז'אנר2** |
| 1.7% | 4.5% | 1 |  |  |  |  |  |  |  |  |  |  |  | **ז'אנר3** |
| 14.2% | 9.2% | 7.4% | 1 |  |  |  |  |  |  |  |  |  |  | **מדינה1** |
| 7.7% | 7.6% | 6.3% | 2.2% | 1 |  |  |  |  |  |  |  |  |  | **מדינה2** |
| 6.2% | 5.8% | 16.6% | 0.6% | 11.9% | 1 |  |  |  |  |  |  |  |  | **מדינה3** |
| 0 | 6.2% | 5.1% | 5.1% | 3.6% | 3.5% | 1 |  |  |  |  |  |  |  | **שחקן** |
| 0 | 0.5% | 1.4% | 3% | 1.2% | 1.6% | 1.4% | 1 |  |  |  |  |  |  | **במאי1** |
| 3.5% | 0.8% | 12.5% | 0.4% | 5.9% | **37.7%** | 3% | 1.5% | 1 |  |  |  |  |  | **במאי2** |
| 4.8% | 12.1% | 4.2% | 8% | 12.5% | **100%** | 3.8% | **24.9%** | 10.3% | 1 |  |  |  |  | **במאי3** |
| 19.9% | 19% | 21.9% | 14.5% | 6.1% | 4.3% | 6.6% | 0 | 6.2% | 12.8% | 1 |  |  |  | **סוג** |
| 7.3% | 6.5% | 7.9% | 8.3% | 1.1% | 4.8% | 0.8% | 0.4% | 3.7% | 2.7% | 17.1% | 1 |  |  | **שנה** |
| 8.3% | 4.3% | 1.6% | 8.1% | 1% | 7.6% | 0.2% | 0.1% | 2.2% | 5% | **28.8%** | 5.2% | 1 |  | **דירוג** |
| **28.3%** | 21.4% | 20.6% | 12.8% | 6.4% | 6.6% | 5.9% | 0.4% | 5.2% | 19.2% | **88.9%** | 22.7% | 19% | 1 | **אורך** |

הטבלה המלאה המתקבלת:

ניתן לצפות בהסבר מלא ובקובץ החישוב בתיקיה: "קורלציה".

בתוך הקבצים- יש לציין כי בשל טעות בהזנת שנת הוצאה, התוצאות הרלוונטיות עבור טור זה תוקנו בקורלציה עבורה ולא בקורלציה עבור שאר הטורים לפני שעלינו על הטעות הזנה. אך הערכים המצוינים בטבלה תקינים לחלוטין במלואם.

**בעניין התוצאות:**יש לשים לב שרוב הערכים קטנים מאוד מ-1, כלומר קיימת שונות נרחבת.  
אולם ניתן להבחין בקורלציה עבור מספר טורים:

* **אורך- סוג** : כאשר אנו מדברים על סדרות טלוויזיה, אורך התוכן הוא בעונות, מספרים קטנים יחסית. אולם כאשר מדובר בסרטים אז מדובר בערכים המדודים בדקות, כלומר לרוב ערכים גבוהים. כך שפרט למספר מועט מאוד של תכנים בעלי זמן קיצוני יחסית, היכולים להיכלל כערכים בסיווג של התכונה השנייה, ישנה קורלציה ברורה בין טורים אלה.
* **דירוג- סוג** : אולם אין קורלציה בין דירוג לאורך, אך היא קיימת עבור סוג. עוצמת הקורלציה היא כ-29 אחוז, כך שניתן עדיין לשקול להשאיר את שני הטורים. קשר זה מראה כי קיימת קורלציה מסוימת בין דירוג- אורך כפי שהנחנו בתחילת העבודה.
* **במאי1 – במאי 3** : קיימת קורלציה של כ-25 אחוזים. הסיבה לכך היא שהערכים הקיימים בטור של הבמאי השלישי, קיימים גם בסדר שונה בבמאי הראשון. ניתן להבחין בקשרים דומים עבור: ז'אנר1-ז'אנר2, מדינה2-מדינה3, ובמאי2-במאי3. נחשב בגבול הסביר, לכן לא יוסרו.
* **מדינה3 – במאי3** : מצב קיצוני של 100 אחוז התאמה. זה נובע בשל העובדה שרוב מוחלט של התאים בשני הטורים ריקים. נשקול להסיר את טור הבמאי השלישי בגלל שכמות התאים הריקים עבורו גדולה יותר. עבור הקשר: **מדינה3 - במאי2** מדובר ב37% דומות. כך שיש בכל זאת לשקול להסיר את הטורים השלישיים המתארים את המדינות הפעילות והבמאים, ולהוסיף הנחה חדשה.
* **ז'אנר1 – אורך**: ניתן להסביר את הקשר ביניהם ע"י העובדה שמדובר בשמות ז'אנרים מסוימים התואמים רק לאחד מסוגי התכנים. למשל " TV Dramas" הייחודי עבור סדרות, או "Horror Movies" התואם לסרטים.

לקראת שימוש באלגוריתמים, ובהתאם לממצאים של הקורלציה, נסיר את העמודות: סוג, במאי שלישי ומדינה שלישית. ובהתאם לכך נוסיף הזנחות חדשות כי לקחנו את שתי המדינות העיקריות והבמאים העיקריים הקשורים לתוכן. וניתן לאשר כי ההנחה כי קיים קשר בין אורך התוכן לבין הדירוג נכונה.

### אלגוריתם רגרסיה:

באמצעות שימוש באלגוריתם רגרסיה ניתן למצוא קשר בין משתנים שונים. אם נמצא קשר כזה נוכל לנתח את הנתונים בעזרתו.

למרות שאין צורך ב-cross-validation, עדיין השתמשנו בו באלגוריתם זה לצורך הנוחות כבתמונה הבאה:

באלגוריתם זה, עמודות שמשו ליצירת מודל חיזוי שבנתה התוכנה למציאת ניבוי הדירוג:

ביצענו יישום של המודל על הנתונים, אשר החזירו דיוק של 41.2%. בשונה מרנדומליות 1/3 שרמת הדיוק היא 33%. ניתן לומר שעבור 33% לא מדובר בתוצאה טובה יותר. אם היה דירוג של שתי אפשרויות היחס בין המאפיינים היה הפוך- כלומר יותר יעיל מה-41.2% שקיבלנו על ידי מציאת המקדמים (Coefficients). אולם בגלל שיש לנו שלושה ערכים סופיים אפשריים מדובר בתוצאה הנחשבת להימור. כדומה למטבע היפותטי בעל שלוש פאות.

ניתן לראות כי דירוג התוכן יהיה יותר גבוה כאשר אורך זמן הסרט/סדרה ארוכים יותר, בניגוד להשארתנו. הסיבה לכך היא שאורך התוכן של סרטים הוא ארוך יותר משל סדרות- כלומר סרטים מדורגים גבוהה יותר. בנוסף או בנפרד למאפיין תכנים חדשים- שמניבים תוצאות גבוהות גם כן, וסותרות בכך את השארתינו הנוספת. כמו כן, ניתן להסביר זאת שעם השנים הצופים נטו לדרג יותר לטובת הסרטים והסדרות באתר IMDb בזכות הפקה בתקציב גבוה יותר, בשילוב עם יכולות טכנולוגיות מתקדמות להפקת סרטים חדשים.

סיבה נוספת היא שסרטים שנמשכים זמן רב יותר זוכים לדירוג גבוה, בגלל העלילה עמוקה ומרתקת יותר, הגורמת לצופה להתעמק בתוכן למרות אורכו. וגם היכולת לעצור את הסרט, ולהמשיכו במועד מאוחר יותר לבחירתם.

נשים לב כי השחקן הראשי על פי מודל זה אינו מהווה כלל כפקטור להצלחת התוכן.

תיאור מלא של האלגוריתם ניתן למצוא בקובץ "רגרסיה" בתיקיה "algorithems".

### אלגוריתם Bayes:

Naive Bayes היא טכניקה פשוטה לבניית מסווגים: דגמים המקצים תוויות כיתות למופעי בעיה, המיוצגים כוקטורים של ערכי תכונה, שבהם תוויות הכיתה נמשכות מתוך קבוצה סופית כלשהי. אין אלגוריתם אחד להכשרת מסווגים כאלה, אלא משפחה של אלגוריתמים המבוססים על עיקרון משותף: כל מסווגי Bayes התמימים מניחים שהערך של תכונה מסוימת אינו תלוי בערך של כל תכונה אחרת בהתחשב במשתנה הכיתה.

עבור סוגים מסוימים של דגמי הסתברות, ניתן לאמן מסווגי Bayes נאיביים ביעילות רבה במסגרת למידה מפוקחת. ביישומים פרקטיים רבים, אומדן הפרמטרים עבור דגמי Bayes התמימים משתמש בשיטת הסבירות המרבית; במילים אחרות, אפשר לעבוד עם מודל Bayes הנאיבי מבלי לקבל את ההסתברות הבייסיאנית או להשתמש בשיטות בייסיות כלשהן.

יתרון של Bayes התמים הוא שהוא דורש רק מספר קטן של נתוני אימונים בכדי להעריך את הפרמטרים הדרושים לסיווג. ה-ROC שקיבלנו בעל ערך דיוק מרשים של: 82.44%.

כאשר התאמה למציאות מבוטאת בתמונה הבאה:תמונה שמכילה צילום מסך

התיאור נוצר באופן אוטומטי

תיאור מלא של האלגוריתם ותוצאותיו ניתן למצוא בקובץ "מודל בייז" בתיקיה "algorithems".

### אלגוריתם RandomForest:

אלגוריתם זה, השייך לתחום ה-Machine Learning, מורכב מעצי החלטה ותפקידו לשפר את תהליך החיזוי. באלגוריתם זה יש שימוש בבחירת קבוצות TEST אקראיות, בעזרתן יוצרים את עצי ההחלטה. כל עץ החלטה בנפרד לא נותן תוצאה מספקת, אולם כאשר משתמשים בהם לאלגוריתם נקבל רמת דיוק גבוהה יותר כפי שניתן לראות בהמשך. כל זה לצורך סיווג איכותי ככל הניתן.

עבור כל אחד מהמאפיינים נחשב את החשיבות שלו לדירוג, זאת על מנת שנוכל לצפות ברווח הממוצע המשוקלל מהמידע (אשר נמדד על בסיס מדד אנטרופי או על פי מדד ג'יני) עם משקולות שפרופורציונאליות למספר התצפיות שנלקח בצומת מסוים.

מכיוון שמדובר באלגוריתם מסובך ומורכב, ולאחר ניסיונות חוזרים ונשנים לשימוש בו על כלל ה-DATA ללא הצלחה, הוחלט כי נבצע את האלגוריתם מספר פעמים עבור קבוצות של כמות טורים מוגבלת.

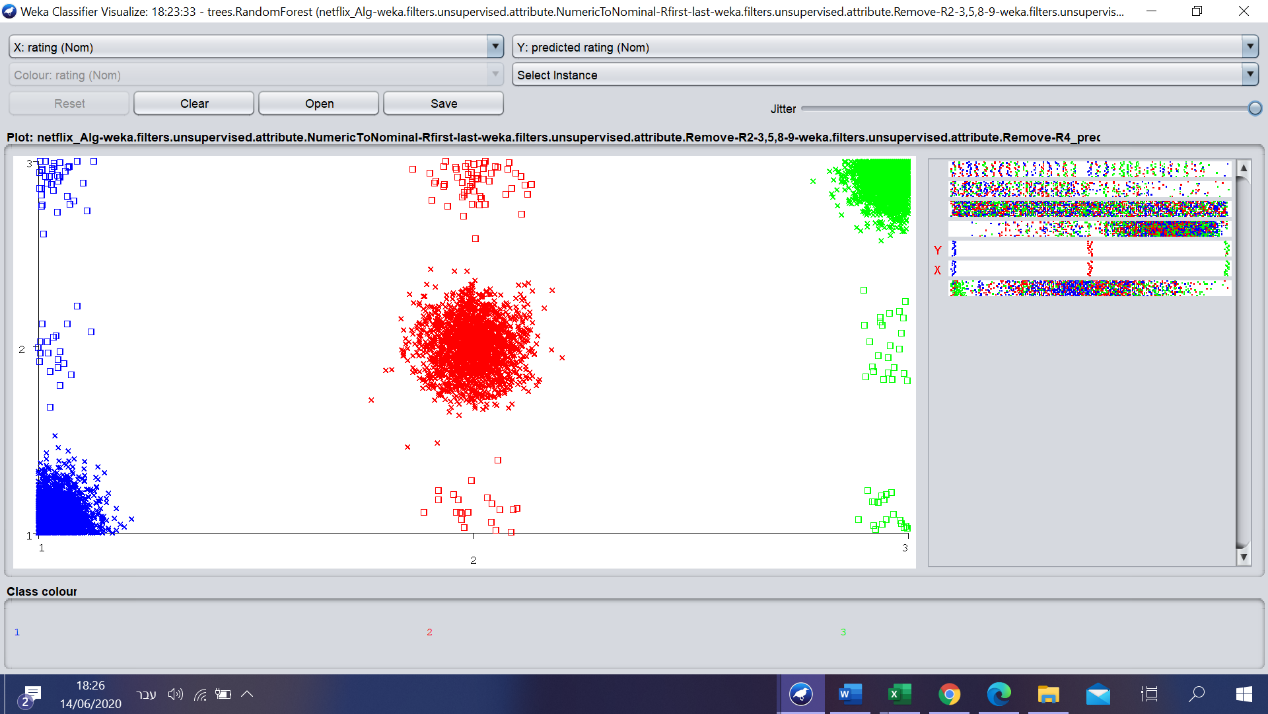
כלל הקבוצות כוללות את המאפיין "דירוג", כאשר שאר הטורים כל פעם מתחלפים. אולם לא נבצע פעולה זאת על כלל הטורים, כלומר הטורים של הבמאים אינם נלקחים בחשבון באלגוריתם זה כלל, בשל הערכים המגוונים מדי. כאשר אין שימוש בטור שחקן ראשי ניתן להשתמש ביותר טורים באנליזה, ולקבל דיוק גבוה גם כן.

כמו כן, עבור טור השחקן הראשי, יש לחשוד בהתאמת יתר של המודל בגלל ריבוי הערכים. לכן אם במסמך זה כל שימוש בשחקן ראשי יניב תוצאות טובות מידי, יהיה עלינו כנראה לוותר עליהן.

הרצת מודל האלגוריתם עבור 5 טורים:



עבור הרצה זו קיבלנו בתוצאה הכללית:



בגרף זה אנו רואים את כלל הדגימות, כאשר ציר ה-X מתאר את ערכי הדירוג האמיתיים של קבוצת הטסט, וציר Y את הניבוי של הדגימות בקבוצות אלו. כל המסומנים ב-X הינם ערכים התואמים את הציפיות למציאות, אולם כל המסומנים ב-  מציינים דגימות שתוצאותיהן שונות מהמצופה.

בבדיקת ROC, כלומר TP ו-FP, כידוע מטרתנו היא להימנע משטח קטן מתחת לערכים. כלומר יש להימנע ממצב בו הגרף שלנו ליניארי וקרוב לאלכסון. כפי שניתן לראות ה-ROC המתקבל הינו ריבועי, כלומר השטח תחתיו גבוהה מאוד מה שתורם לאמינות גבוהה של המודל.

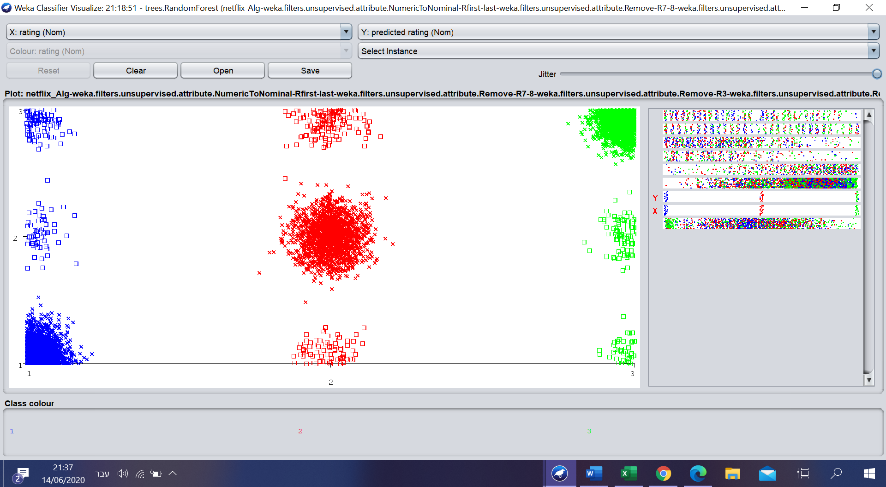
ROC עבור דירוג 1 – הדירוג הנמוך ביותר:

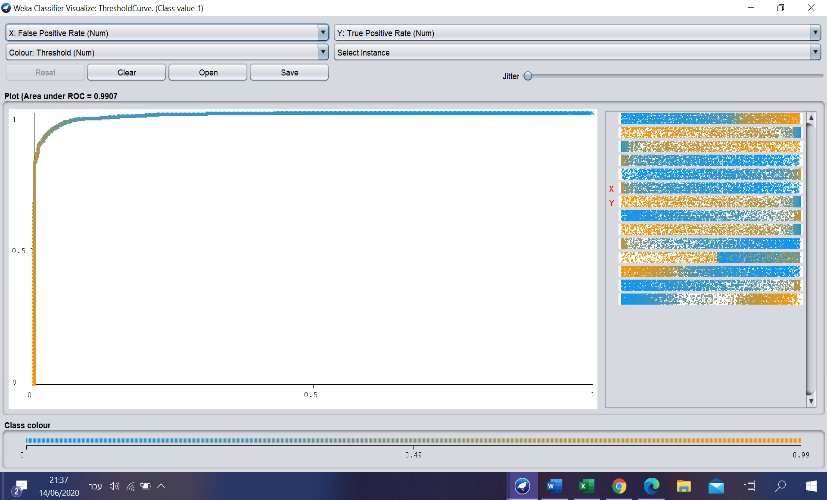
ROC עבור דירוג 2 – הדירוג הבינוני:

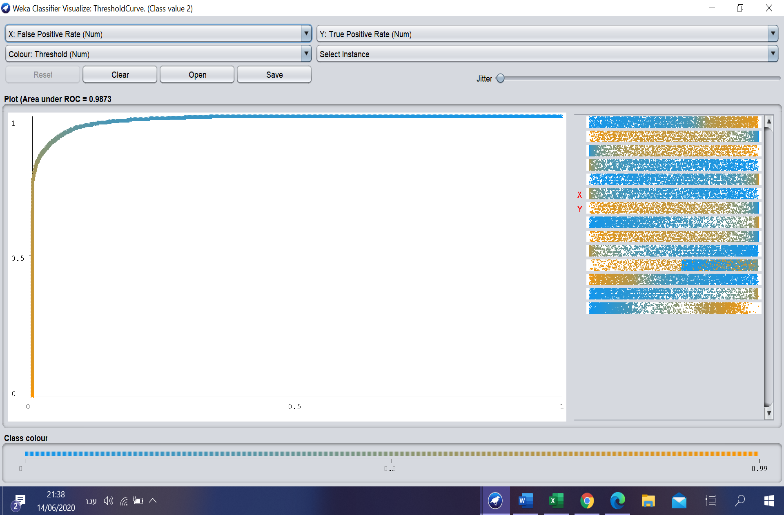
ROC עבור דירוג 3 – הדירוג הגבוה:

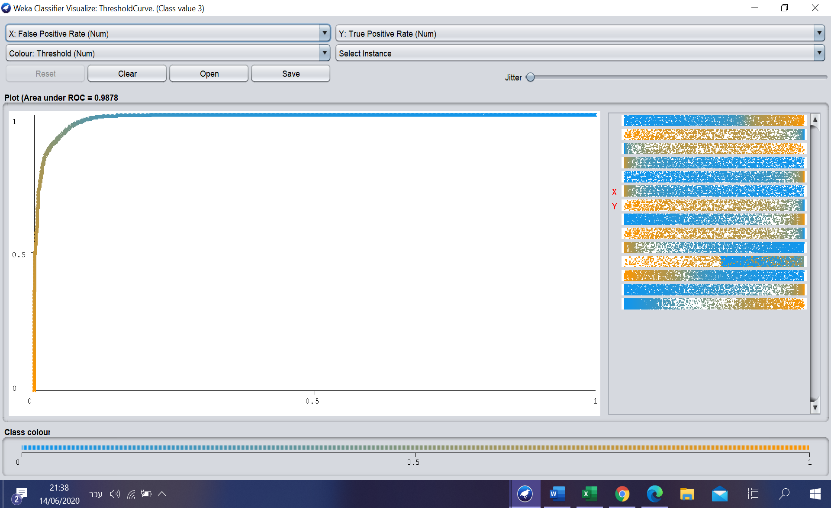
הרצה נוספת עבור 7 טורים (ללא שחקן ראשי):



לאחר הרצה זו קיבלנו בתוצאה הכללית:

ROC עבור דירוג 1 – הדירוג הנמוך ביותר:

ROC עבור דירוג 2 – הדירוג הבינוני:

ROC עבור דירוג 3 – הדירוג הגבוה:

בכל הרצה, עבור קבוצות הדירוג (X נבחר, ו-Y עם Z) התקבלו שתי תוצאות confusion matrix. האחת עבור מקסימום רווח \ מחיר, ושניה עבור מינימום. כלומר בסימון COST: כאשר מדובר במקסימום- ערכו של המחיר הוא מקסימאלי, וכאשר מינימום - ערכו של המחיר מינימאלי.

תיאור מלא של האלגוריתם ניתן למצוא בקובץ "RandomForest" בתיקיה "algorithems".

ניתן לראות שעבור כל המדדים וההרצות האלגוריתם הניב אחוזי דיוק גבוהים מאוד (מעל 95%).

# סיכום ומסקנות:

בפרויקט זה, בחרנו את האלגוריתמים המתאימים ביותר למענה המטרות שהצבנו, ובדקנו את נכונות ההנחות והשפעת ההזנחות גם כן. ביצוע פרויקט זה הניב קשים לניבוי בשל העובדה שרוב המידע נומינאלי, והרבה ערכים הם שונים מאוד מערכים אחרים של אותו טור, כך שנאלצנו להתמודד עם הסכנה של התאמת יתר לערכים מסוימים למציאת המטרות שלנו. אולם לצורך מציאת דפוס ויעילות ה-DB, בוצעו ע"י אמצעים מתקדמים חישובים אשר הפכו מידע זה ליעיל עבור יוצרי תוכן בחברת נטפליקס.

## דיון:

עבור כל אחת ממטרותינו נענה לפי המודלים ששומשו בתוצאות:

* **סיווג מאפיינים שונים של תוכן לפי סיכוייהם לקבל דירוג גבוה.**  
  נימצא כי במודל רגרסיה המאפיינים המובילים הינם- שנת הוצאה, מדינה וז'אנר (מסודר בהתאם להשפעתם).ככל שהתוכן חדש יותר, הוא נחשב לאיכותי יותר בעיניי המשתמשים.לגבי המדינה בה התוכן הופק, מדינות בעלות השפעה יחסית גבוהה לפי מודל זה. ויש ז'אנרים הנצפים יותר ע"י מגוון רחב יותר של צופים מז'אנרים אחרים. אולם מולםהשפעת השחקן הראשי הוא 0, כלומר אין כלל השפעה לשיוך שחקן מפורסם לדמות מסוימת. ישנה סתירה בין תוצאות אלה לבין ההנחות הראשוניות, זה הוא נושא מעניין ומפתיע.  
  עבור אלגוריתם Bayes ניתן לראות את השפעת כל אחד מהמאפיינים בנפרד בצורה איכותית עם דיוק של 82.44%. נמצא שם כי יש השפעה לז'אנר השני ועוד, אך ההשפעה הגדולה ביותר היא של טור השחקנים- בניגוד לרגרסיה. קימת השפעה של אורך התוכן על הדירוג, אך היא נמוכה יחסית. וגם השפעת המדינות היא הנמוכה ביותר.
* **מציאת שילוב המאפיינים אשר מניבים את דירוג הצופים הגבוה ביותר.**  
  במודל רגרסיה ניתן לראות את ההשפעה של כל אחד מהמאפיינים על הדרוג בדיוק של 41.2 אחוזים. כלומר בנוסף לנאמר מלעיל, נמצא דפוס השפעת השילוב בין מאפיינים.עבור RandomForest ניתן לראות את השפעת שילוב התכנים בשתי קבוצות התוצאות המוצגות. בגלל שעם הטור "שחקן ראשי" ובלעדיו הדיוק נשאר ברמה גבוהה- ניתן להסיק עבור חישובים נוספים באלגוריתמים אחרים שאין overfitting עבור טור זה, וניתן להסיק מסקנות להתאמתו.
* **בדיקת חוקיות בין אורך התוכן לבין הדירוג הניתן על ידי הצופים.**  
  השפעת אורך התוכן על הדירוג נמצאה כחיובית, למרות השערתנו בעניין. ניתן לראות זאת בתוצאות הרגרסיה.

## מסקנות:

על פי מודל הרגרסיה- ניתן לקחת שחקן פחות מפורסם וזה לא יגרום להורדה בדירוג התוכן**,** כלומרחיסכון בעלויות ליהוק. אולם במודל Bayes המדויק יותר, אין להשליך את חשיבות בחירת השחקן הראשי, כלומר יש להשקיע בבחירתו של שחקן המעביר את התוכן בצורה יעילה ומרגשת.

תכנים חדשים מניבים יעילות גבוהה יותר בכל הנוגע לדרוג. יש להזרים כמה שיותר תכנים בהתחשבות בז'אנר ומדינת ההפקה.

אורך התוכן הוא בעל קשר ישיר להצלחת התוכן. כלומר לרוב משתמשים יתנו ציון גבוה יותר לסרטים מאשר סדרות. יש להתחשב ברעיון כי ה-IMDB מוכר יותר כמדרג סרטים ופחות מוכר עבור דירוג סדרות ומשחקים וכו'...

כאשר מדובר בשילוב של מאפיינים יש לשים לב להשפעת שילוב הז'אנר והמדינה. למשל שיוך טוב יותר של "לוקיישנים" לצילום עם הערך מוסף לתוכן. כלומר מומלץ למרות העלויות לעיתים לפזר את שטחי הצילום לתוצאה איכותית יותר וריאלית יותר עבור הצופים.