

שם הפרויקט: דירוגי סדרות וסרטים

מגיש:

אביב בסלו 322294315

משפט המתאר את הפרויקט: סדרות\סרטים ודירוגם ע"י מספר אטרי ביקורת (IMDb,Rotten Tomatoes).

שלב 3:

קובץ df_1.pkl מכיל מידע כללי על סרטים וסדרות כגון: שנת הפקה, משך זמן הסרט\סדרה ודיוגים מכל הפלטפורמות כגון דיסני, פריים וידאו וכו'.

ההאנפורמציה של הנתונים ניתן לראות שיש 5380 אינדקסים, 14 עמודות, מכילות שם עמודה, כמו ערכיהם לא חסרים יש בכל עמודה ואיזה סוג עמודה כל אחת.

למשל, אחת מהעמודות היא country שהיא בעצם הסוג שקיים, בה ניתן לראות שישנם בסה"כ 4004 ערכים לא חסרים והיא מסוג object. עמודה נוספת release year שמצויה בקובץ, בה ניתן לראות שישנם בסה"כ 5380 ערכים לא חסרים והיא מסוג int64.

בסטטיסטיקה ההיסטורית של הנתונים ניתן לראות את המינים השונים לכל עמודה כמו חיזון, ממוצע, מינימום ומקסימום. לדוגמה, בעמודה של release year הממוצע הוא 2012.642193 .

שלב 5:

טבלה 1:

pv1_df	Movie	TV Show
type		
country		
Argentina	7.500000	6.928571
Australia	NaN	7.210811
Austria	NaN	6.500000
Belarus	NaN	6.800000
Belgium	NaN	7.262500
Brazil	4.700000	6.666001
Canada	7.200000	6.940000

בטבלה הראשונה הגדרתי 'country' כ-index , עמודות כ- 'type' (Movie, TV Show) וערכי IMDb כדירוג והחלתי על טבלת הציר פונקציית צבירה של ממוצע.

טבלה 2:

pv2_df	13+	16+	18+	7+	all
Age	0	84	90	82	67
type	79	96	100	93	83
Movie					

בטבלה השנייה הגדרתי 'type' כ-index , עמודות כ- 'Age' וערכי Rotten Tomatoes כדירוג החלתי על טבלת הציר פונקציית צבירה של מקסימום ובכל ערכי ה-Nan הוצגו כ-0 .

בטבלה זו ניתן לראות לכל קבוצה גיל את המקסימום דירוג של Rotten Tomatoes לפי סרט וסדרה. בקטgorיה של 13+ בסרטים ניתן לראות שאין דירוג כלל.

טבלה 3:

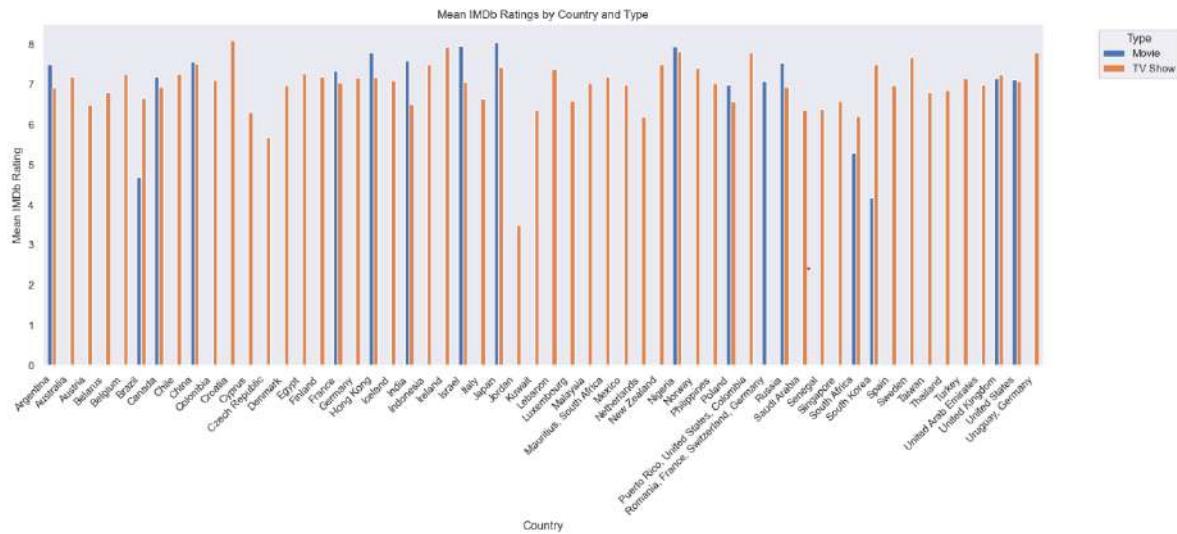
type	Movie	TV Show
release year		
1904	0	1
1931	0	1
1932	0	1
1934	0	1
1943	0	1
...
2017	18	587
2018	12	550

בטבלה השלישי הגדրתי 'release year' כ-index, עמודות כ-'type' וערבי בטבלה השלים את הדרישה על טבלת הציר פונקציית צבירה של ספירה וככל ערבי ה-NaN הוצגו כ-0.

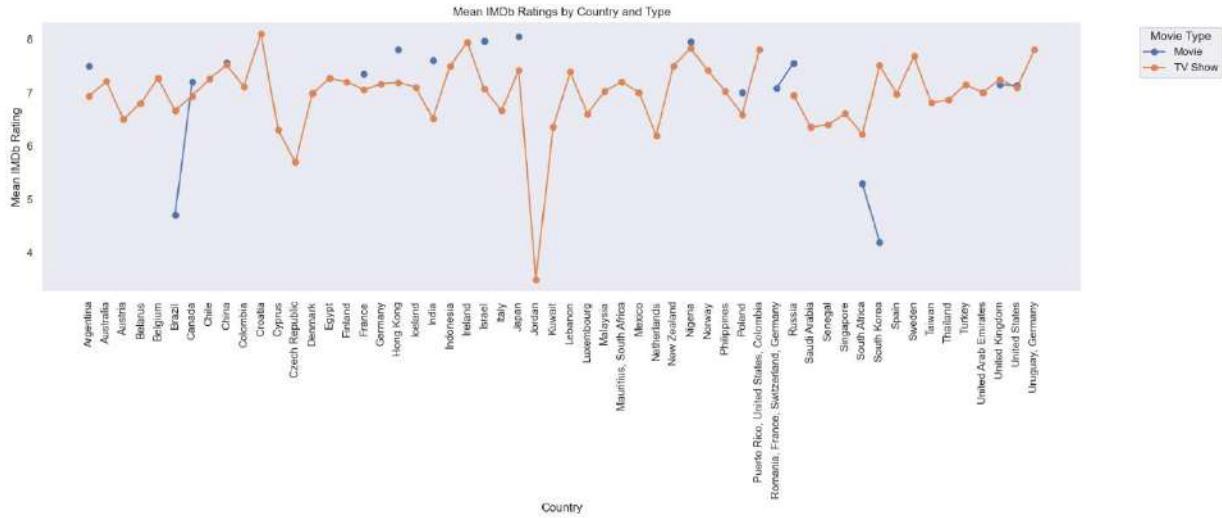
בטבלה זאת ניתן לראות ממות של סרטים וסדרות שקיימת לכל שנה.

יש שנים שיש להם רק סרטים או רק סדרות ויש שנים שיש גם ו גם.

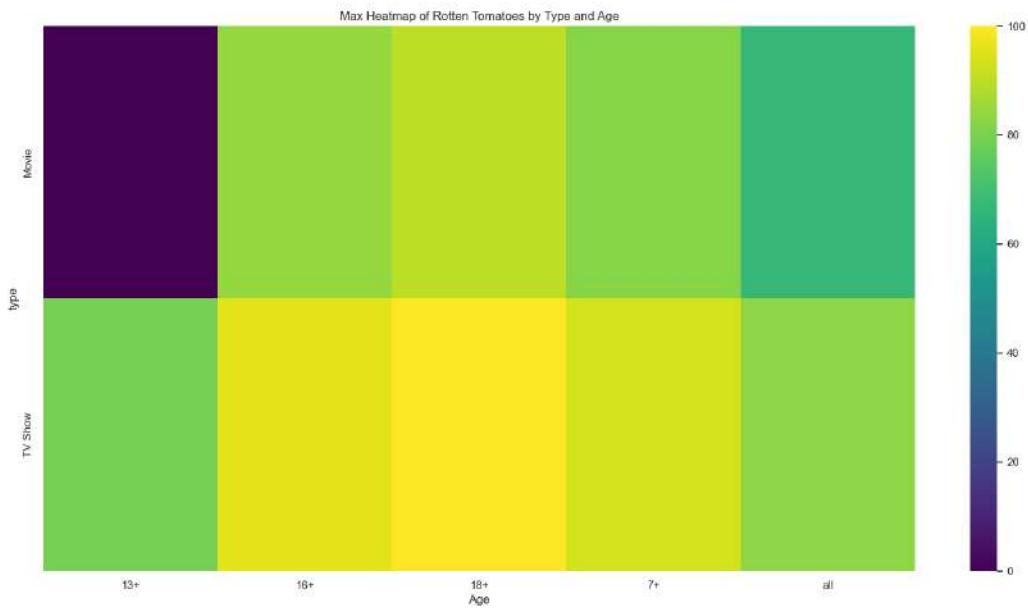
שלב 6:



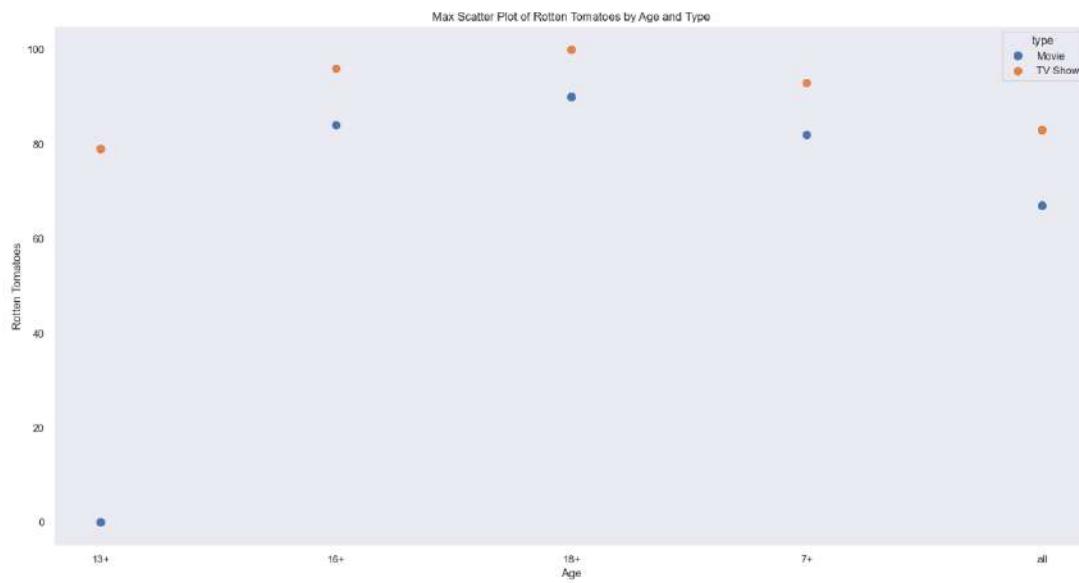
graf 1 : מייצג את טבלת הציר הראשונה, מייצג את ממוצע הדירוגים לכל מדינה לפי סרט(בצבע כחול) וסדרה(בצבע כתום). בגרף זהה ניתן לראות שממוצע הדירוג הכי נמוך (כ-3.5) נמצא בירדן בקטגורית סדרות. בשאר המדינות ניתן לראות שטוווח ממוצע הדירוגים הוא באזורי ה-5-7.5.



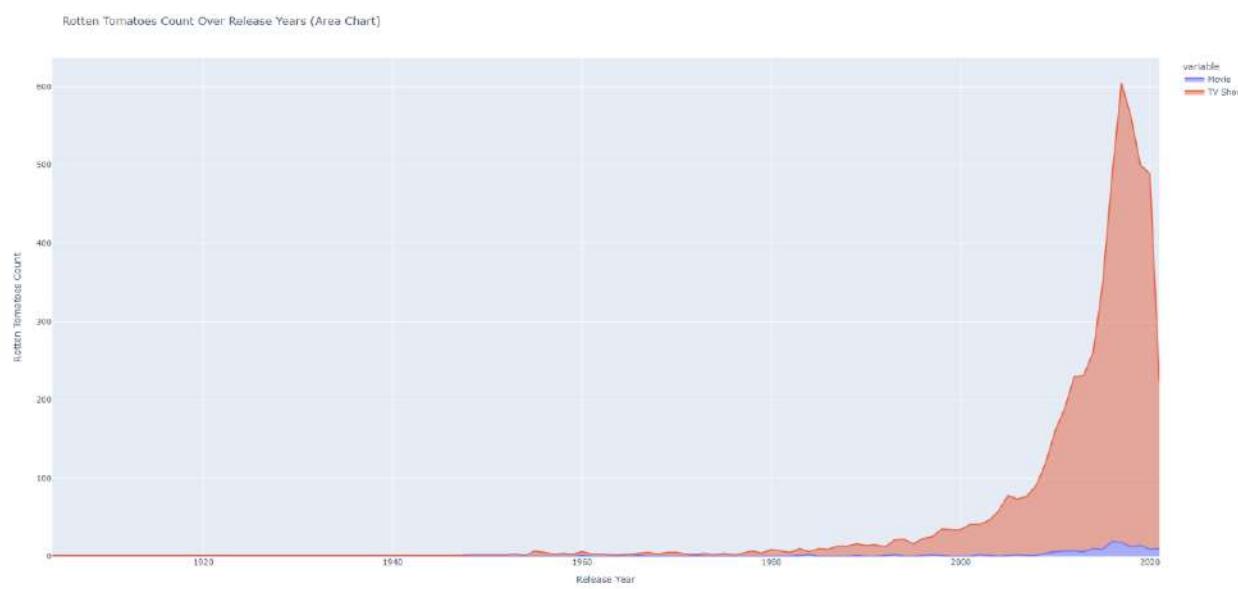
graf 2: מיצג את טבלת הציר הראשונה, מיצג את ממוצע הדירוגים לכל מדינה לפי סרט(בצבע חום) וסדרה(בצבע כחום). ממוצע הדירוג בקטגוריות הסדרות נמצא לחוב באותו הטווח בין 6 ל-8 לעומת קטגוריות הסרטים יש יותר פיזור כאשר רוב ממוצע דירוג זה נמצא בחלק העליון ובודדים בחלק התחתון.



graf 3: מיצג את טבלת הציר השניה, מיצג את מקסימום הדירוג לכל קבוצת גיל לפי סרט וסדרה. ניתן לראות בגרף שהציון המקסימלי הינו נמצא בקבוצת הגיל 13+ בסרטים, הציון המקסימלי הינו גבוה נמצא בקבוצת הגיל 18+ בסדרות. בסדרות ניתן לראות שרובם עם דירוג גבוה מ-75 ובסרטים זה יחסית מפוזר יש דירוגים גבוהים ויש גם נמוכים.

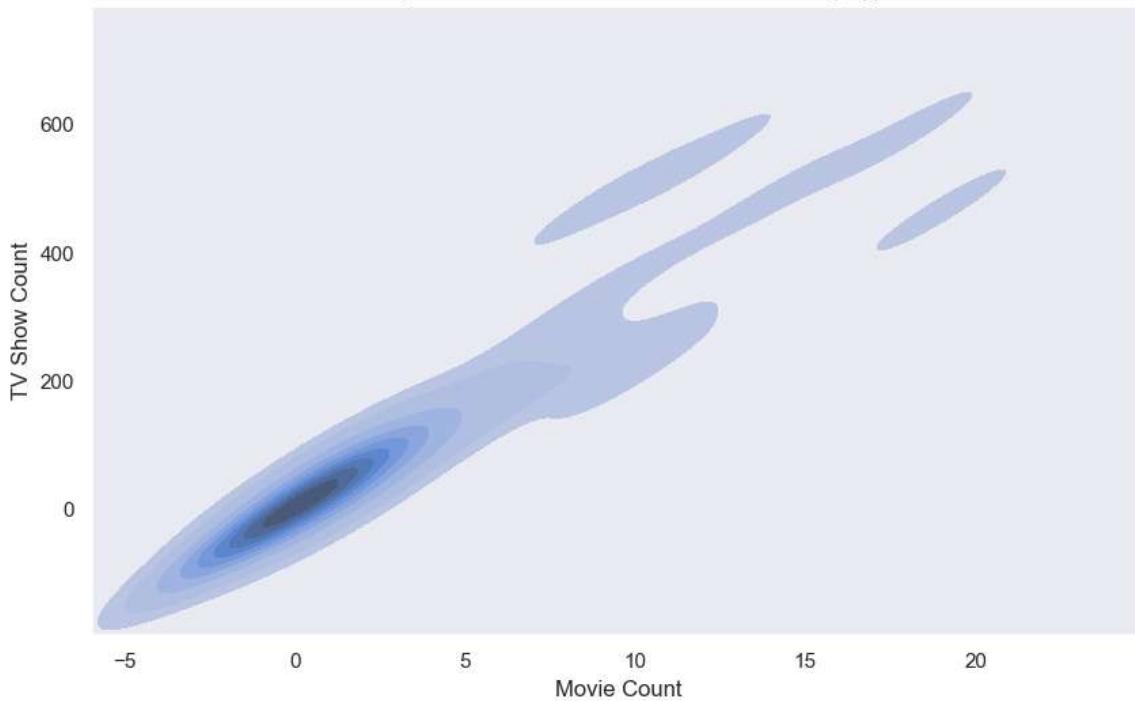


גרף 4: מייצג את טבלת הציר השנייה, מייצג את מקסימום הדירוג לכל קבוצת גיל לפי סרט וסדרה. ניתן לראות בגרף שהציגו המקסימלי hei נמור נמצא בקבוצת הגיל 13+ בסרטים. ניתן לראות שבין סדרות לסרטים באותום קבוצות גיל הדירוגים בסדרות יותר גבוהים מהדירוגים הסרטים.



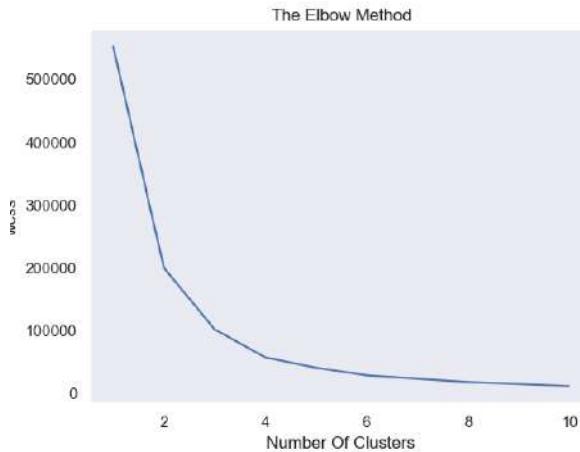
גרף 5: מייצג את טבלת הציר השלישי, מייצג אתכמות הדירוג הקיימת לכל שנת יציאה לאור לפי סרט (צבע כחול) וסדרה (צבע אדום). ניתן לראות בגרף עלייה בכמות דירוגים בסדרות לאורך השנים בערך משנת 2000 כשכמות הדירוג הייתה 34 וב-2017 הייתה 584 ולאחר מכן ישבה יצירה. לעומת זאת סרטים רואים שיש עליה של כמות הדירוג באזרע שנות 2010 אבל בכללי אין עליה דרסטית לעומת כמות דירוג הסדרות.

Density Contour Plot of Rotten Tomatoes Count by Type



graf 6: מייצג את טבלת הציר השלישי, מייצג את במות הדירוג הק"מאת לפ"י סרט וסדרה. בגרף זה ניתן לראות שיש קורולציה בכמות דירוג בין הסדרות לסרטים באשר במעטות הדירוג קטנות. לאחר מכן ניתן לראות שהקשר בכמות דירוג בין הסדרות לסרטים פוחת באשר במעטות הדירוג של הסדרות גדלה בצורה הרבה יותר גדולה במהלך השנים לעומת במות דירוג הסרטים ובכך ניתן לראות שבחילוק מהgraf קשור זה נעלם.

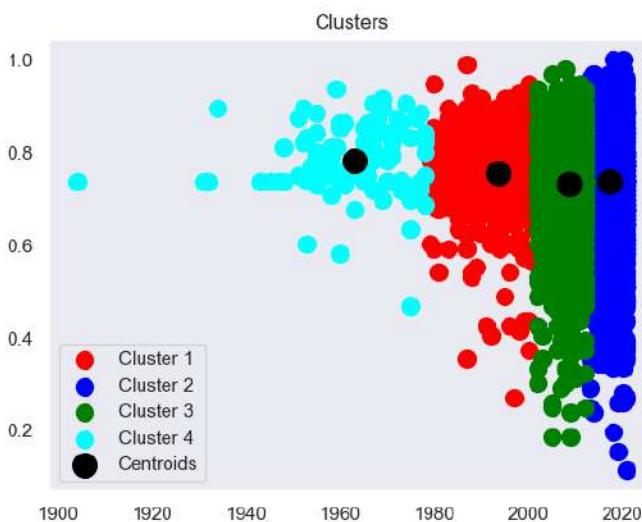
שלב 7:



בשיטת המפרק אני מחפש את המספר האופטימלי של האשכבות.

WCSS הוא סכום המרחק בריבוע בין כל נקודה לבין מרכז הצביר (האשכול).

ניתן לראות מהגרף שהערך האופטימלי הוא $K=4$.

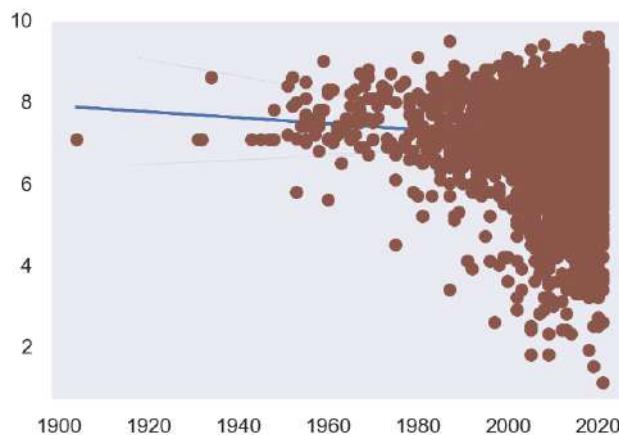


ניתן לראות שמדובר בויזואלייזציה של האשכבות.

המרכזים של כל קלטסר הוא בצעע שחוו.

ניתן לראות שלכל קלטסר יש צבע שונה, בклטסר 2 הכהול ניתן לראות שיש טווח גדול יותר של דירוג. בכל הקלטסרים למעט קלטסר 4 תבלת, רואים שיש צפיפות גדולה. כל המרכזים של הקלטסרים נמצאים מ-1960-.

שלב 8:



לקן רגסיה יש שיפור שלילי שמצוין על ירידה בדירוג IMDb לאורך השנים.

האזור המוצל סביב קו הרגסיה מצין את השונות הפוטנציאלית בדירוג IMDb.

נקודות הנתונים מוקבצות על קו הרגסיה, זה מצביע על קשר ליניארי חזק מכיוון שיש צפיפות וקורלציה גבוהה יותר מאשר שנות 2000.

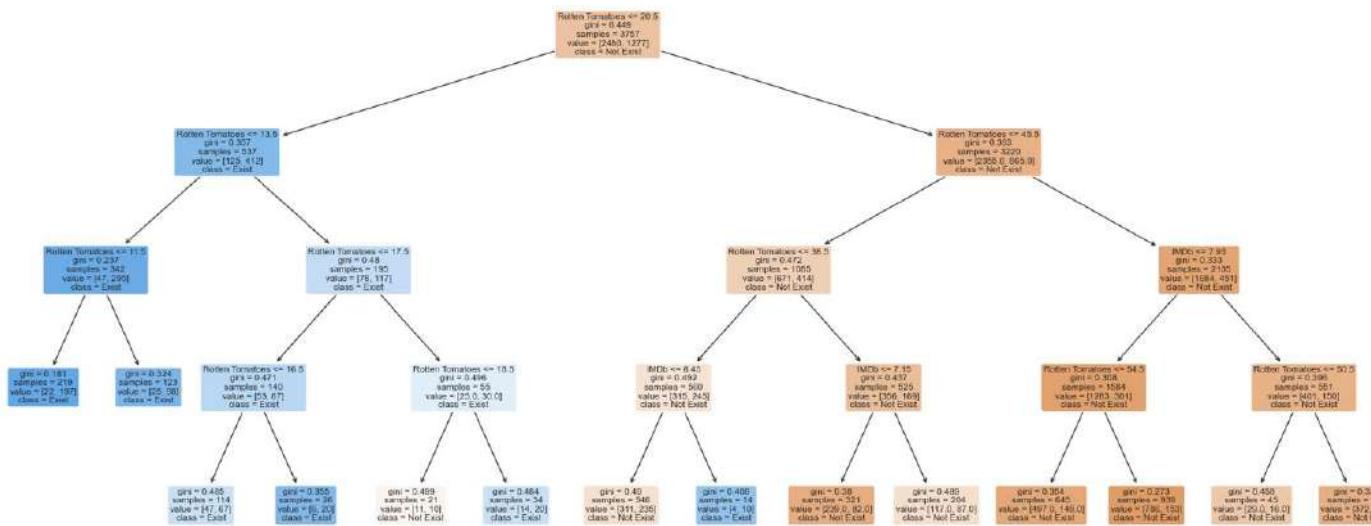
Significance: Yes
R-squared: 0.005327300810787139
Adjusted R-squared: 0.0051419350450047085

ניתן לראות את מדרדי הרגסיה והרגסיה יצאא 0.0053 ולכן רגסיה זאת מובהקת, ביון שהוא קטן יותר מ-0.05.

מדד הרגסיה Adjusted R-squared יצא 0.00514.

שלב 9:

Gini



בעץ החלטה על בסיס מודל החלטה של גINI לוקחים 70 אחוז מהנתונים לאימון - 30 אחוז לבדיקה, מדדיי ב- 3757 ערכים, על פי דירוגי IMDb ו-Rotten Tomatoes ו- Not Exist classes שמחולקים ל- Exist . Prime Video. בחלק העליון של העץ נתוני הגי'י הוא 0.449 שהוא ממד גבוה. במתוך הרמות שיש בעץ היא 4 ובעומק זה הגעתי לממד הדיק הגבוה ביותר במדוד זה. העץ החלטה מתחלק לנواتים שהם מתחת ל- 20.5 ומעלה ל- 20.5 ב-Rotten Tomatoes . כמשמעותם על צד שמאל של העץ, ממד הגי'י משתפר ונهاיה 0.357 על בסיס 537 ערכים. ברמה 2 הגעתי לממד גי'י טוב יותר שהוא 0.237 על בסיס 342 ערכים. ברמה 3 הגעתי לממד גי'י הכי טוב שהוא 0.181 על בסיס 219 ערכים ושהוא נמצא ב-Class Exist . במשמעותם על צד ימין של העץ שהמדדים פחות טובים כיון שמדד הגי'י הכי טוב שהגעתי אליו בצד זה הוא 0.273 שנמצא ב- Class – Not Exist .

רמת הדיק הצעיר במודול זה היא 71.5% שמצוין על רמת דיק גבוהה יחסית.

F1 score – במקרה הנ"ל הציון הוא 0.403

Recall score – במקרה הנ"ל הציון הוא 0.279

Precision – במקרה הנ"ל יצא 0.72 . מצינו את חלקן של התוצאות החיבויות האמיתיות מתוך כל התוצאות החיבויות.

Confusion Matrix – מטריצה זאת נותנת פירוט מפורט של סיווגים נכונים ושגויים, מראה 155 חיוביות אמיתיות, 997 שליליות אמיתיות, 60 חיוביות שגויות ו- 399 שליליות שגויות.

ROC AUC – במקרה הנ"ל הציון הוא 0.702 וזה מצביע על ביצועים יחסית טובים של המודול.

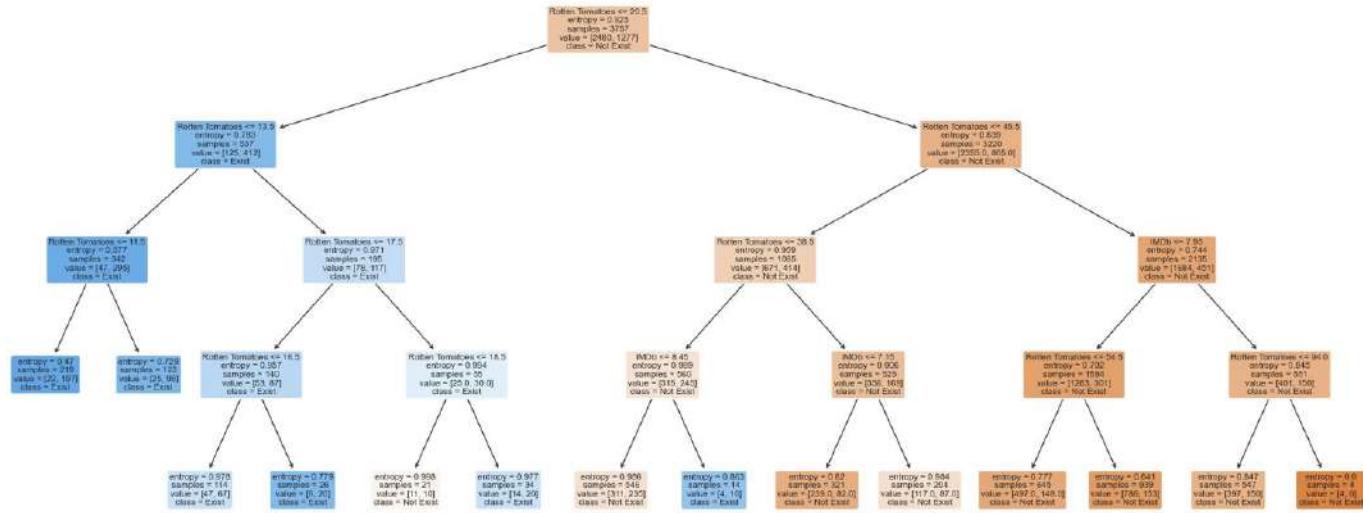
```

gini - Model Performance:
Accuracy = 0.7150837988826816
F1_score = 0.40312093628088425
Recall_score = 0.27978339350180503
Precision = 0.7209302325581395

Confusion Matrix:
[[997 60]
 [399 155]]
ROC AUC 0.7021378876938684

```

Entropy



בעז החלטה על בסיס מודל החלטה של entropy לוחמים 70 אחוז מהנתונים לאימון ו-30 אחוז לבדיקה, אני מודד ב-3757 ערבים, על פי דירוג IMDb ו-Rotten Tomatoes ו-IMDb ו-Prime Video. בחלק העליון של העז נתוני entropy הוא 0.925 שהוא ממד שמצביע על איזה גובה. כמות הרמות שיש בעז היא 4 ובוונק זה הגיעו לרמת הדיקוגרף הגבוהה ביותר במדד זה. העז החלטה מתחילה לננתונים שהם מתחת ל-20.5 ב-Rotten Tomatoes. במשמעותם על צד שמאל של העז, ממד entropy משתפר ונעה 0.783 על בסיס 537 ערבים. בrama 3 הגיעו למדד entropy הכי נמוך בצד שמאל של העז שהוא 0.47 על בסיס 219 ערבים שהוא נמצא ב-Class Exist. כאשר ממד entropy יותר נמוך זה מצביע על פחות איזה סדר בנתונים. במשמעותם על צד ימין של העז ניתן לראות בצד זה שמדד entropy הכי נמוך הוא 0.0 על בסיס 4 ערבים שנמצא ב-Class Not Exist – .

```
entropy - Model Performance:
Accuracy = 0.7150837988826816
F1_score = 0.40312093628088425
Recall_score = 0.27978339350180503
Precision = 0.7209302325581395

Confusion Matrix:
[[997 60]
 [399 155]]
ROC AUC 0.7014701713520659
```

רמת הדיקוגרף שהגעתי במודל זה היא 71.5% שמצביע על רמת דיקוגרף גבוהה יחסית.

F1 score – במקרה הנ"ל הציון הוא 0.403 .

Recall score – במקרה הנ"ל הציון הוא 0.279 .

Precision – במקרה הנ"ל יצא 0.72 . מצין את חלקן של התוצאות החיוביות

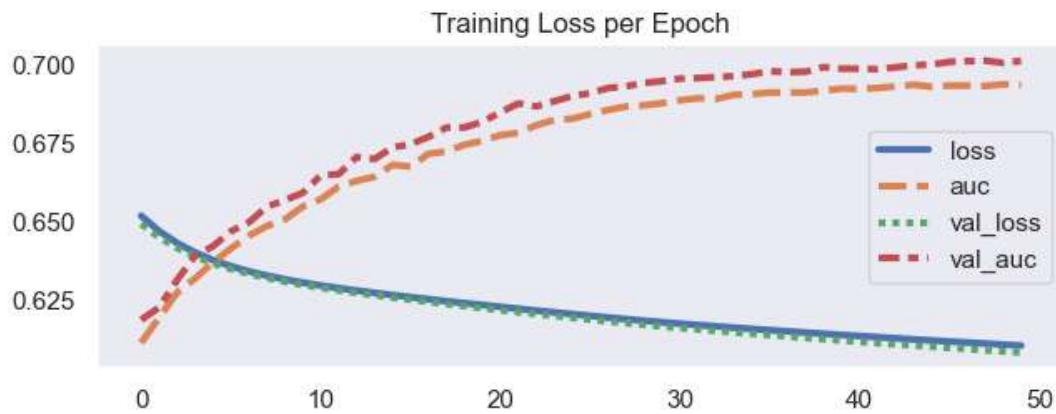
ה眞יות מתחום כל התוצאות החיוביות.

Confusion Matrix – מטריצה זאת נותנת פירוט מפורט של סיווגים נכוןים ושגויים, מראה 155 חיוביות אמיתיות, 997 שליליות אמיתיות, 60 חיוביות שגויות ו-399 שליליות שגויות.

ROC AUC – במקרה הנ"ל הציון הוא 0.701 וזה מצביע על ביצועים יחסית טובים של המודול.

לבסוף אחרי שביצעת את שני עצי החלטה אלה, ממד היגי'י מראה רמת דיקוגרף גבוהה יותר במעט ממדד entropy . לעומת זאת בשאר הממדדים קיבלו ממדים זהים בשני המודלים.

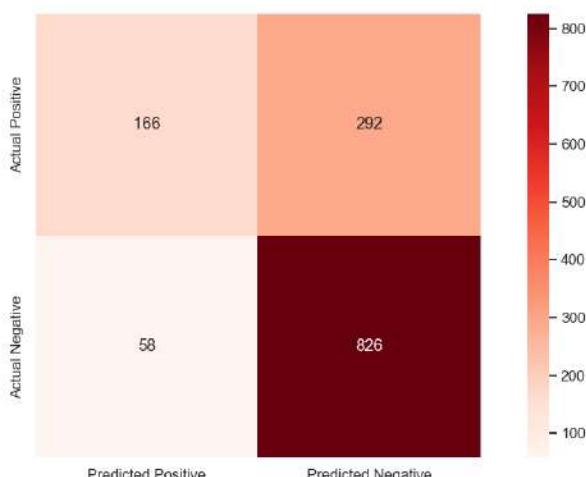
שלב 10:



בגרף Training Loss per Epoch אני מודד את ביצועי המודל בכל epoch , במודל לוקחים 75 אחוז מהנתונים לאימון ו-25 אחוז לבדיקה וניתן לראות את נתוני ה-AUC וה-VAL_AUC נמצאים בטוחה של בין 0.6 לבין 0.7 על בסיס 50 epochs . ה-AUC מייצג את רמת הדיק על גבי נתוני האימון. ה-VAL_AUC מייצג את רמת הדיק על גבי נתוני האמת. שניהם עולים בהתאם לכמות epoch עולה, כאשר נתונים VAL_AUC גבוהים כמעט מה-AUC .

שניהם מגעים לרמת דיק יחסית גבוהה באזור ה-0.7 , אם זאת בשלבים מסוימים בקורס יוכלו לראות שיפורים (לקראת ה-epochs האחרונים).

ניתן לראות שה-loss (מייצג את ההפסד של המודל) וה-val_loss (מייצג את ההפסד של המודל בהתאם להערכתו) יורד לבינן 0.65-0.66. ניתן לראות שגם loss ה-50 עדין נמצא בערך 0.65. ניתן לראות שגם לערך epoch ה-50 נמצא בשיפור יחסית נמוך שלילי. בתוצאה לכך ניתן לראות שככל שכמות epoch גדולה, כמות ההפסד במודל יורדת.



ניתן לראות שהרשות העצבית הפשויה הזאת מספקת ביצועי מודל עבור 0.73 AUC . $(166+826)/(1342) = 0.73$

ניתן לראות שחייזי שלילי עם תוצאה שלילית אמיתית קיבל תוצאה של 826 לעומת חייזי חיובי עם תוצאה נכון נבונה אמיתית שקיביל 292. חייזי נכון עם תוצאה נכון נבונה אמיתית קיבל 166 לעומת חייזי נכון עם תוצאה שלילית אמיתית שקיביל 58 . ניתן לראות מההתוצאות שקיבלנו שמודל זה מנביא יחסית נכון ברוב המקרים.