# NETFLIX

IMDb rating analysis

עומרי יולזרי

## מבוא

"נטפליקס" היא חברה אמריקאית שמושבה בלוס גאטוס, קליפורניה. היא הוקמה בשנת 1997 על ידי ריד הייסטינגס ומארק רנדולף בסקוטס ולי. "נטפליקס" היא תוכנת סטרימינג שמשדרת תוכניות טלויזיה וסרטים לטלוויזיה, למובייל ולמחשב. ל"נטפליקס" יש 139 מיליון לקוחות שמשלמים מידי חודש עבור התכנים המוצעים להם.

**הצורך העסקי של "נטפליקס" הינו להגדיל את מאגר הלקוחות ולשמר את הקיימים.** בשוק יותר מבעבר, קיימות יותר ויותר חברות שמתחרות ב"נטפליקס". לכן, קיים הצורך לבדוק כיצד לגרום לתיעדופה על פני המתחרים.

"נטפליקס" כל הזמן בוחנת איזה תכנים חדשים יכנסו למאגר התוכניות והסרטים שלה. הdata set אותו בחרנו מאתר "Kaggle" מכיל בתוכו מעל 6000 רשומות של סרטים וסדרות שמופעים ב"נטפליקס". לרשומות אלו נאספו 12 תכונות שביניהן דירוג התכנים המוצעים ב"נטפליקס" באתר "IMDB".

**הבעיה העסקית אותה אנו רוצים לפתור היא איך להגדיל את מאגר הלקוחות ולשמר את הקיימים.** על מנת לפתור בעיה זו, "נטפליקס" תכניס למאגר התכנים שלה סדרות שיקבלו רייטינג גבוה. רייטינג גבוה בדרך כלל מתבסס על דירוג גבוה באופן יחסי שנמצא באתרי דירוג מובילים כגון "IMDB".

מהסתכלות על הdata set התגלה לפנינו שישנם לא מעט סדרות/סרטים שמקבלים דירוג נמוך ביחס לתכנים אחרים. באמצעות הניתוח אותו למדנו בקורס "בינה עסקית", נוכל להגיע למסקנות ותובנות שבהן נוכל להבין את סיבת הדירוג של כל תוכן.

לאחר סקירת הנושא, שאלת המחקר והבעיה העסקית אותה אנו רוצים לפתור הן כלדלהלן:

<u>שאלת המחקר –</u> איך אפשר למנוע מ"נטפליקס" רכישה של סרטים ותוכניות טלוויזיה בעלות דירוג נמוך ביחס לתכנים אחרים.

<u>הבעיה העסקית אותה אנו רוצים לפתור –</u> מניעה מ"נטפליקס" לרכוש תכנים שיקבלו רייטינג נמוך.

<u>מרכיבי מאגר הנתונים:</u>

עמודות : קוד תוכן (נומינלי), סוג התוכן (נומינלי), שם התוכן (נומינלי), במאי (נומינלי), שחקנים (נומינלי), מדינה שהפיקה את הסרט (נומינלי), תאריך הוספה ל"נטפליקס" (נומינלי), תאריך הוצאה לאור (נומינלי), דירוג לפי אתר "IMDB" (נומינלי - אורדינלי), משך התוכן (נומרי), קטגוריית תכנים (נומינלי), תיאור התוכן (נומינלי).

מספר רשומות : 6234

מספר תכונות: 12

## Pre-processing

<u>לאחר סקירת הטבלה, החלטנו להוריד את עמודת הבאות:</u>

עמודות Description, Show_id, Cast ו Title מכילות מלל רב אשר עלול לגרום למצב של over fitting שעלול להטות את אמינות המסקנות שנבצע בעזרת האלגוריתמים השונים על הdata set.

לאחר סקירת כלל העמודות ובדיקת כמות התאים הריקים בכל עמודה הגענו למסקנה שנוריד עמודת director. בעמודה זו חסרים מעל 30% מהערכים בתאים, לעומת שאר העמודות שקיימות בדאטה שלא עולות על 8%. לכן, מדובר באחוז גבוה באופן יחסי אשר מעיד על חוסר אמינות בנתונים, עמודה זו היא נומינלית ולכן לא ניתן לבצע עליה פעולות חישוביות על מנת למלא את הנתונים החסרים.

בנוסף, עמודת listed in מכילה בתוכה מספר לא מבוטל של ז'אנרים הנכללים בכל סרט או סדרה ב"נטלפיקס". הערכים הנתונים לא מאפשרים חקירה יסודית ומספיק איכותית מכיוון שלא ניתן לנתח אותם ולהסיק עליהם מסקנות בצורה טובה.

העמודה האחרונה אותה ניפנו היא עמודת country. עמודה זו אמנם מכילה את המדינות בהם יוצרו התכנים שהתווספו ל"נטפליקס", אך ברוב התאים היו רשימות של יותר ממדינה אחת, מה שהקשה מאוד על ניתוח העמודה ועל היעילות שהיא יכולה לתרום לנו בניתוח מרכיבי הדירוג של כל תוכן.

<u>מילוי ערכים חסרים בעמודות-</u>

תחילה, בדקנו כמה תאים חסרים בכל עמודה באקסל באמצעות פונקציית COUNTIF ובדקנו מה היחס בין כמות הערכים החסרים לכמות הרשומות.

Rating- קיימים 331 תאים חסרים. כדי למלא אותם חישבנו את הממוצע עבור התאים המלאים ולאחר מכן את כל התאים החסרים מילאנו בממוצע שחושב.

Date added- קיימים 10 תאים חסרים. בעמודה זו מצאנו בעזרת תוכנת "פייתון" את התאריך השכיח ומילאנו את הנתונים החסרים בעזרת תאריך זה.

<u>פיצול שדות –</u>

בכדי לקבל ראייה מעמיקה יותר על הנתונים, פיצלנו את עמודת תאריך הוספת התוכן ל"נטפליקס" לשנת הוספה וחודש הוספה.

<u>דיסקרטיזציה –</u>

כדי שנוכל לבצע את האלגוריתמים, עשינו דיסקרטיזציה לדירוג של כל התכנים כדי להמירם מערכים רציפים WEKA:לאחידים, הדירוגים חולקו כך על פי תוכנת

```
Class
'(-inf-5.15]'
'(5.15-5.75]'
'(5.75-6.15]'
'(6.15-6.55]'
'(6.55-6.785]'
'(6.785-7.05]'
'(7.05-7.35]'
'(7.35-7.65]'
'(7.65-8.05]'
'(8.05-inf)'
```

המרה לערכים בינאריים-

מעמודה Type המרנו את הערכים מסוג נומינלי לסוג בינארי א-סימטרי.

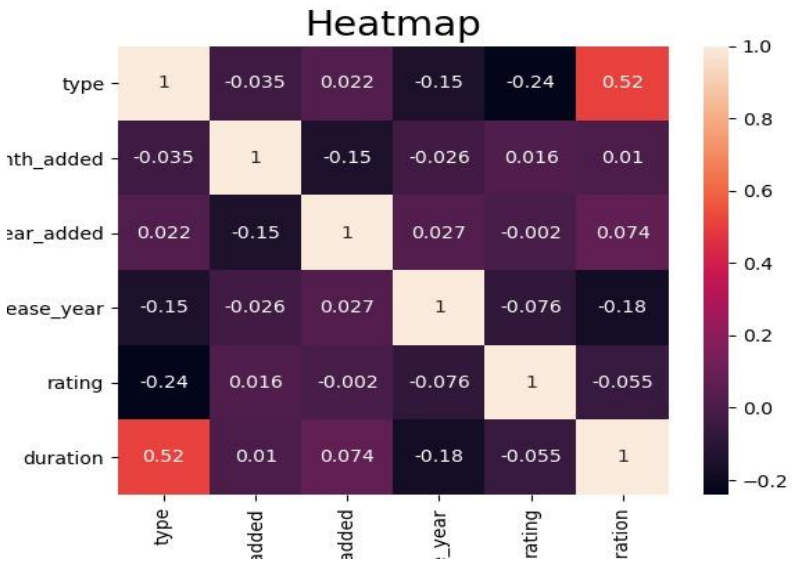
1 - Movies

0 - TV Shows

קורלציה-

בחלק זה, בחנו האם קיימות קורלציות גבוהות בין תכונות מסוימות.את הקורלציה בדקנו בעזרת תוכנת פיתון על ידי הרצת פונקציית Heatmap. במידה ותתקבל התאמה של 0.9 ומעלה או 0.9- ומטה בין שתי תכונות, עלינו לרדד את מימדי הdata בעמודה נוספת. כאשר קיימת התאמה גבוהה מאוד בין שתי תכונות אז תכונה אחת "מספרת" על השניה ולכן מספיקה עמודה אחת לניתוח.

פלט מפת חום בתוכנת Python-




ניתן לראות כי ההתאמה הגבוהה ביותר היא בין עמודת type לעמודת duration והיא 0.52. התאמה זו הגיונית ביוון שבדרך כלל משך ארוך של תוכן יסווג אותו כסרט (ערכו 1 אחרי הבינאריזציה). ההתאמה אינה גבוהה די כדי להוריד תכונה ולכן לא נוריד יותר עמודות. בנוסף, רוב הקורלציות סובבות סביב ה-0 ולכן רוב העמודות עצמאיות ובלתי תלויות אחת בשניה.

## אלגוריתמים

לאחר תהליך ה-pre-processing העלנו את קובץ הנתונים המעודכן לתוכנת Weka. בשלב הראשון, ביצענו דסקטיזציה לנתונים. לאחר מכן הרצנו על הנתונים את אלגוריתמי הקלסיפיקציה הבאים: Random Forest, Decision Tree ו Naive Bayes. את האלגוריתמים הרצנו במצב cross validation כאשר Folds = 10 לשם טיוב הנתונים.

לניתוח הנתונים הגדרנו כך את ה-Confusion Matrix:

True Positive – חזינו קבוצת דירוג לתוכן מסוים ואכן התוכן השתייך לקבוצה זו.

True Negative – חזינו שתוכן מסוים לא ישתייך לקבוצת דירוג מסוימת, התוכן באמת לא השתייך לקבוצת הדירוג.

False Positive – שגיאה מסוג 1. חזינו שתוכן ישתייך לקבוצת דירוג מסוימת אך הוא שייך לאחרת.

False Negative -שגיאה מסוג 2. חזינו שתוכן לא ישתייך לקבוצת דירוג מסויימת אך התבדנו שהוא כן שייך אליה.

המטרה שלנו היא שהאלגוריתמים יפיקו לנו ניבויים כמה שיותר מדויקים. נרצה שמדד ה-True Positive יהיה כמה שיותר גבוה בעוד שנרצה לצמצם את מדד ה- False Positive. הצלחה זו תעזור ל"נטפליקס" לעלות את צריכת התכנים ובמקביל למזער את הפסדי החברה ככל האפשר.

מדד נוסף לדיוק האלגוריתמים הינו גרף ROC. מטרתינו היא להגיע לשטח מקסימלי מתחת לגרף, לכן נחפש את מדד ה-ROC הנותן את התוצאה המקסימלית.

## אלגוריתמי הקלסיפיקציה

Random Forest

בשיטה זו האלגוריתם מבצע מספר רב של חישוב עצי החלטה ומהם מחשב את ההחלטה המתאימה ביותר.

העצים נוצרים לרוב על ידי דגימה מתוך המאפיינים או מתוך התצפיות. כלומר, כל אחד מהעצים נותן תוצאה לא אופטימאלית אך באופן כללי, על פי רוב החיזוי בדרך זו משתפר.

ניתן לחשב את חשיבותו של כל אחד מהמאפיינים השונים ביער אקראי באמצעות הרווח הממוצע המשוקלל מהמידע עם משקולות שפרופורציונאליות למספר התצפיות שנלקח בצומת מסוים.

אלגוריתם זה צפוי להניב את שטח ההתאמה הטוב ביותר.

פלט ההרצה בתוכנת Weka

-Accuracy

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| | 0.170 | 0.089 | 0.171 | 0.170 | 0.171 | 0.082 | 0.599 | 0.168 | '(-inf-5.15]' |
| | 0.129 | 0.090 | 0.128 | 0.129 | 0.129 | 0.039 | 0.569 | 0.114 | '(5.15-5.75]' |
| | 0.123 | 0.087 | 0.122 | 0.123 | 0.123 | 0.036 | 0.526 | 0.103 | '(5.75-6.15]' |
| | 0.122 | 0.118 | 0.123 | 0.122 | 0.123 | 0.005 | 0.510 | 0.121 | '(6.15-6.55]' |
| | 0.204 | 0.105 | 0.207 | 0.204 | 0.206 | 0.100 | 0.563 | 0.187 | '(6.55-6.785]' |
| | 0.109 | 0.093 | 0.107 | 0.109 | 0.108 | 0.015 | 0.509 | 0.099 | '(6.785-7.05]' |
| | 0.129 | 0.097 | 0.129 | 0.129 | 0.129 | 0.032 | 0.514 | 0.119 | '(7.05-7.35]' |
| | 0.077 | 0.094 | 0.081 | 0.077 | 0.079 | -0.017 | 0.517 | 0.097 | '(7.35-7.65]' |
| | 0.108 | 0.095 | 0.104 | 0.108 | 0.106 | 0.013 | 0.543 | 0.102 | '(7.65-8.05]' |
| | 0.187 | 0.091 | 0.186 | 0.187 | 0.186 | 0.096 | 0.628 | 0.160 | '(8.05-inf)' |
| Weighted Avg. | 0.138 | 0.097 | 0.138 | 0.138 | 0.138 | 0.041 | 0.548 | 0.129 | |

-Confusion Matrix

=== Confusion Matrix ===

| a | b | c | d | e | f | g | h | i | j | <-- classified as |
|---|---|---|---|---|---|---|---|---|---|---|
| 91 | 54 | 54 | 62 | 52 | 45 | 47 | 49 | 46 | 34 | a = '(-inf-5.15]' |
| 57 | 66 | 58 | 60 | 52 | 51 | 38 | 46 | 48 | 35 | b = '(5.15-5.75]' |
| 57 | 58 | 61 | 66 | 55 | 53 | 41 | 45 | 34 | 25 | c = '(5.75-6.15]' |
| 67 | 67 | 71 | 80 | 69 | 61 | 83 | 54 | 54 | 48 | d = '(6.15-6.55]' |
| 49 | 52 | 51 | 80 | 133 | 51 | 60 | 55 | 52 | 70 | e = '(6.55-6.785]' |
| 47 | 47 | 52 | 57 | 51 | 56 | 48 | 67 | 47 | 43 | f = '(6.785-7.05]' |
| 45 | 47 | 40 | 79 | 57 | 43 | 71 | 50 | 55 | 63 | g = '(7.05-7.35]' |
| 43 | 43 | 46 | 62 | 60 | 65 | 51 | 41 | 57 | 63 | h = '(7.35-7.65]' |
| 45 | 45 | 33 | 54 | 54 | 54 | 55 | 43 | 55 | 71 | i = '(7.65-8.05]' |
| 32 | 36 | 32 | 51 | 58 | 43 | 57 | 59 | 79 | 103 | j = '(8.05-inf)' |

ההתאמה הכללית במודל הינה 0.548. ה- roc הכי גבוה הוא של קבוצת הדירוגים האחרונה (מ8.05 עד 10) והיא 0.628.

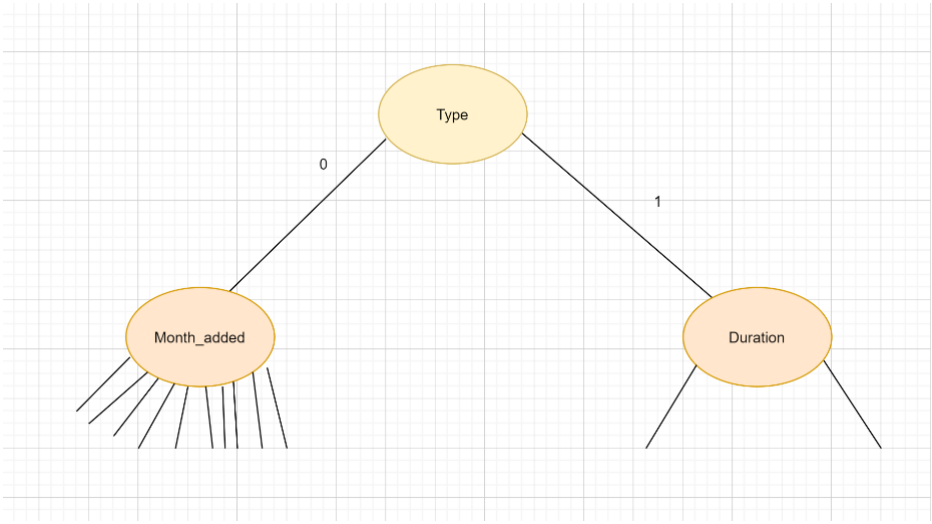
בconfusion matrix האלכסון מראה לנו כמה פעמים הערך שהיה צריך להופיע אכן הופיע (true positive). לדוגמא, קבוצה j שדיוקה יחסית גבוה לשאר הקבוצות קיבלה ב-true positive את הערך 103. כלומר מכלל החיזויים של האלגוריתם עץ אקראי שהתוכן הוא בעל דירוג של 8-10 הוא צדק 103 פעמים.

Decision Tree

המודל בוחן את הנתונים לפי עץ החלטה אותו הוא בונה לפי מובהקות הקשר בין המשתנים המסבירים למשתנה המוסבר. העץ מבצע כל פעם חלוקה בין המשתנים, החלוקה נעשית לפי האנטרופיה הגדולה ביותר. העץ מסתיים כאשר כל הדגימות נמצאות ב"עלה" מאותו הסוג או כאשר נגמרו הדגימות ולא התאפשר ליצור עץ מספיק טוב שיוכל לספק את הסיווג המתאים לכל דגימה.

כעת, בעקבות כך שהעץ החלטה שנוצר מאוד גדול ואינו מספיק ברור ויזואלי על פי תוכנת הWeka, נמחיש בעזרת תיאור ויזואלי של התחלת העץ את המשתנה המסביר שמקטין באופן מקסימלי את האי וודאות.

להלן תיאור ויזואלי של העץ-




מהסתכלות בעץ, ניתן לראות כי המשתנה המסביר העיקרי שמקטין באופן מקסימלי את האי וודאות הינו סוג התוכן, לאחר מכן המשתנים המסבירים המשניים הינם אורך התוכן אשר יוצאים ממנו 2 הסתעפויות וחודש צירוף התוכן ל"נטפליקס" ממנו יצאו כ10 הסתעפויות.

פלט ההרצה בתוכנת Weka-

Accuracy-

=== Detailed Accuracy By Class ===

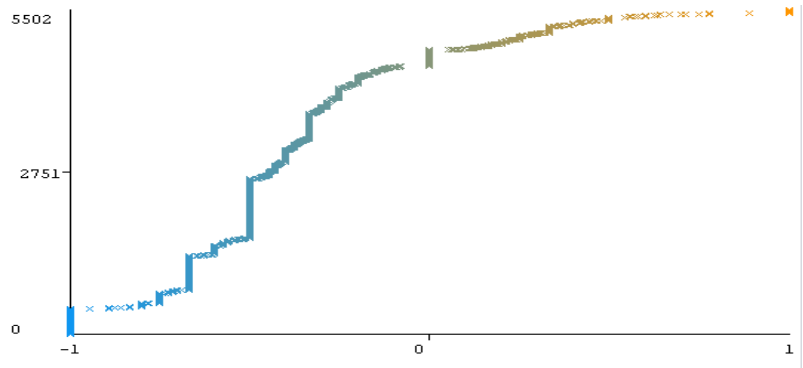
| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| | 0.215 | 0.126 | 0.155 | 0.215 | 0.181 | 0.078 | 0.584 | 0.124 | '(-inf-5.15]' |
| | 0.149 | 0.105 | 0.127 | 0.149 | 0.137 | 0.041 | 0.530 | 0.106 | '(5.15-5.75]' |
| | 0.113 | 0.090 | 0.110 | 0.113 | 0.112 | 0.022 | 0.510 | 0.099 | '(5.75-6.15]' |
| | 0.240 | 0.181 | 0.152 | 0.240 | 0.186 | 0.049 | 0.528 | 0.126 | '(6.15-6.55]' |
| | 0.225 | 0.112 | 0.213 | 0.225 | 0.219 | 0.110 | 0.569 | 0.169 | '(6.55-6.785]' |
| | 0.087 | 0.069 | 0.116 | 0.087 | 0.100 | 0.021 | 0.530 | 0.103 | '(6.785-7.05]' |
| | 0.056 | 0.070 | 0.082 | 0.056 | 0.067 | -0.017 | 0.514 | 0.106 | '(7.05-7.35]' |
| | 0.075 | 0.064 | 0.112 | 0.075 | 0.090 | 0.014 | 0.513 | 0.100 | '(7.35-7.65]' |
| | 0.065 | 0.050 | 0.117 | 0.065 | 0.083 | 0.019 | 0.560 | 0.108 | '(7.65-8.05]' |
| | 0.196 | 0.083 | 0.208 | 0.196 | 0.202 | 0.116 | 0.573 | 0.132 | '(8.05-inf)' |
| Weighted Avg. | 0.147 | 0.098 | 0.141 | 0.147 | 0.141 | 0.047 | 0.542 | 0.119 | |

במודל זה נתמקד בשני מדדים:

– ROC

ההתאמה הכללית של המודל הינה 0.542. זוהי התאמה בינונית של המודל מכיוון שנשאף שהוא יתקרב ככל היותר ל1.

עקומת ROC-




PRECISION–על פי הנתונים ניתן לראות שמדד הPRECISION הוא 0.141. ניתן להסיק מכך שזהו נתון נמוך אשר יקשה עלינו לזהות את דפוסי הדירוגים השונים של התכנים.

Naïve Byes

אלגוריתם זה מתבסס על תורת ההסתברות ומניח כי אין תלות בין תכונות האובייקטים המסווגים.

תהליך השימוש במסווג בייסאני מתחלק לשני שלבים:

1. המסווג מקבל training set – אוסף של דוגמאות והסיווג (classification) שלהן.
כל דוגמה מיוצגת על ידי וקטור של ערכים, כאשר כל ערך מציין את הערך שמקבלת הדוגמא בפיצ'ר (מדד מסוים).
2. בהינתן דוגמה חדשה שהסיווג שלה אינו ידוע, המסווג צריך לחזות את הסיווג שלה.

המסווג נקרא "נאיבי" מכיוון שהוא מניח שכל פיצ'ר הוא בלתי תלוי בפיצ'רים אחרים.

מטרתו של האלגוריתם היא לסווג אובייקט לאחד מכל הקטגוריות כאשר האובייקט מאופיין על ידי וקטור התכונות. האלגוריתם מתבסס על תורת ההסתברויות הניתנות להערכה מתוך מדגם מייצג. בנוסף מניחים את ההנחה ה"נאיבית" שאם ידועה הקטגוריה אזי התכונות אינן תלויות זו בזו. את התוצאות המתקבלות משווים לכל הקטגוריות שבחרנו ונבחר את הקטגוריה שנותנת את התוצאה המקסימלית.

פלט ההרצה בתוכנת Weka-

- Accuracy

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| | 0.288 | 0.130 | 0.192 | 0.288 | 0.231 | 0.133 | 0.625 | 0.154 | '(-inf-5.15]' |
| | 0.178 | 0.108 | 0.145 | 0.178 | 0.160 | 0.063 | 0.605 | 0.126 | '(5.15-5.75]' |
| | 0.058 | 0.059 | 0.088 | 0.058 | 0.070 | -0.001 | 0.555 | 0.100 | '(5.75-6.15]' |
| | 0.208 | 0.172 | 0.140 | 0.208 | 0.167 | 0.030 | 0.555 | 0.134 | '(6.15-6.55]' |
| | 0.148 | 0.105 | 0.159 | 0.148 | 0.154 | 0.045 | 0.558 | 0.146 | '(6.55-6.785]' |
| | 0.054 | 0.068 | 0.075 | 0.054 | 0.063 | -0.017 | 0.530 | 0.098 | '(6.785-7.05]' |
| | 0.067 | 0.060 | 0.109 | 0.067 | 0.083 | 0.008 | 0.508 | 0.107 | '(7.05-7.35]' |
| | 0.050 | 0.041 | 0.114 | 0.050 | 0.069 | 0.012 | 0.516 | 0.106 | '(7.35-7.65]' |
| | 0.029 | 0.053 | 0.053 | 0.029 | 0.037 | -0.032 | 0.549 | 0.104 | '(7.65-8.05]' |
| | 0.294 | 0.160 | 0.169 | 0.294 | 0.214 | 0.106 | 0.651 | 0.150 | '(8.05-inf)' |
| Weighted Avg. | 0.140 | 0.098 | 0.126 | 0.140 | 0.127 | 0.035 | 0.565 | 0.124 | |

שטח ההתאמה הכללי במודל זה עומד על 0.565. ה ROC הכי גבוה הוא של קבוצת הדירוגים האחרונה (8.05 עד 10) והוא 0.651. ניתן לראות כי שרמת הדיוק אמנם גבוה מהשאר אך אינה גבוה במיוחד.

-Confusion Matrix

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   <-- classified as
 156  75  33 102  51  29  19  13  13  50 |   a = '(-inf-5.15]'
 108  93  39 104  37  34  31  10  21  46 |   b = '(5.15-5.75]'
  89  90  29 113  41  32  21  21  17  50 |   c = '(5.75-6.15]'
  95 107  53 138  62  43  33  20  32  82 |   d = '(6.15-6.55]'
  77  60  29 112  98  28  46  32  32 147 |   e = '(6.55-6.785]'
  91  51  33  98  44  28  43  26  32  75 |   f = '(6.785-7.05]'
  49  55  34  85  70  55  37  17  40 113 |   g = '(7.05-7.35]'
  58  46  28  95  64  47  32  27  39 105 |   h = '(7.35-7.65]'
  50  37  29  74  61  36  41  39  15 136 |   i = '(7.65-8.05]'
  38  28  21  65  87  40  38  31  44 163 |   j = '(8.05-inf)'
```

במטריצת הבלבול, קבוצה J שדיוקה יחסית גבוה לשאר הקבוצות קיבלה ב true positive את הערך 163. כלומר, מתוך כלל הפעמיים שחזתה את אחוז הרייטינג צדקה ב163 מהמקרים.

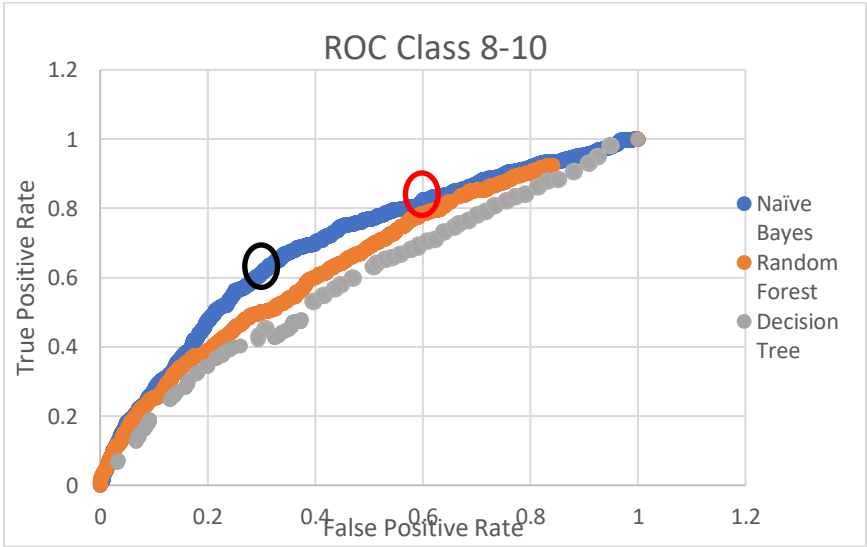
## סיכום ומסקנות

על מנת לבדוק איזה מבין האלגוריתמים שבחנו קודם לכן בעל רמת הדיוק הגבוהה ביותר נשתמש במדד ROC.

ROC הינו מדד המציג בצורה ויזואלית את הביצועים של האלגוריתמים. בעזרת מדד זה נבחן אילו מבין האולגוריתמים מביא אותנו לרמת הדיוק הגבוהה ביותר ונבחר באלגוריתם עם השטח הגדול ביותר הקרוב לחלקו העליון של ציר ה-Y. ציר ה- Y (true positive rate) הינו מדד ה- Sensitivity והשאיפה שיהיה כמה שיותר גבוה. ציר ה-x (false positive rate) הינו מדד ה-Specificity והוא מחווה לנו על מדד השגיאה של האלגוריתם בכל נקודה בגרף.

בחרנו להתמקד בניתוח שתי קבוצות הדירוגים בעלות מדד ה-ROC הגבוה ביותר בכל האלגוריתמים. קבוצות דירוג 0-5 ו-8-10.

להלן גרף השוואת ה-ROC של קבוצת הדירוגים 8-10 לפי כל האלגוריתמים שנבחנו:



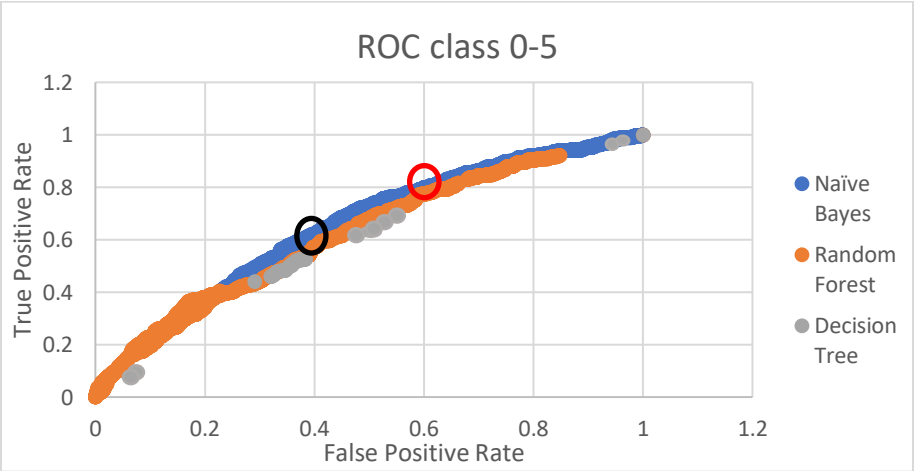
האלגוריתם בעל השטח הגדול ביותר הינו Naïve Bayes. ניתן לראות שלאורך כל התרשים הוא נמצא מעל שני האלגוריתמים האחרים.

נתייחס לשתי נקודות עבודה בתרשים ROC:

1. נקודת העבודה העליונה מסומנת בתרשים באדום. בנקודה זו נדע לחזות ש 80% מהתכנים שניבאנו להם רייטינג גבוה (טווח הניקוד בין 8 ל-10) אכן נמצאו בקבוצה זו. עם זאת, קיימת שגיאה של 0.6. כלומר, אומנם ב80% התוכן שנרכש היה עם אחוז רייטינג גבוה אך ב-60% מהמקרים קנינו לשווא תכנים שאחוז הרייטינג שלהם לא היה גבוה כפי שצפינו שיהיה.
2. נקודת העבודה השנייה היא הנקודה התחתונה המוקפת בעיגול שחור. בנקודה זו אחוז החיזוי של התכנים לסיווג נכון לקבוצה 8-10 הינו רק 60% אך השגיאה יורדת ל30%. אמנם הסיכוי לנבא סיווג נכון לתוכן ירד משמעותית אך גם מדד השגיאה לסיווג לקבוצה זו ירד.

מבין שתי הנקודות הנ"ל נעדיף את נקודת העבודה השנייה (שחורה). נמליץ ל"נטפליקס" על נקודת עבודה זו כיוון שעדיין קיים אחוז גבוה של תכנים שאכן יקבלו דירוג גבוה. כתוצאה מכך, "נטפליקס" תגדיל את כמות המנויים. בנוסף, השגיאה קטנה יחסית ולא תגרום לנזק משמעותי מבחינה כלכלית.

<u>להלן גרף השוואת ה-ROC של קבוצת הדירוגים 0-5 לפי כל האלגוריתמים שנבחנו:</u>




מהגרף ניתן לראות, שבין הטווחים 0-0.25 האלגוריתם המוביל הינו Random Forest אך ברוב התרשים (0.25-1) האלגוריתם המוביל הוא Naïve Bayes. לכן, אלגוריתם זה הוא בעל השטח הגדול ביותר ונתמקד בו.

נתייחס לשתי נקודות עבודה בתרשים ROC:

1. נקודת העבודה העליונה המסומנת בתרשים באדום. בנקודה זו נדע לחזות ש 80% מהתכנים שניבאנו להם רייטינג נמוך (טווח הניקוד בין 0 ל-5) אכן נמצאו בקבוצה זו. עם זאת, קיימת שגיאה של 0.6. כלומר, אומנם ב80% התוכן שנבדק היה עם אחוז רייטינג נמוך אך ב-60% מהמקרים נמנע מקניית תכנים בעלי רייטינג גבוה יותר.
2. נקודת העבודה השנייה היא הנקודה התחתונה המוקפת בעיגול שחור. בנקודה זו אחוז החיזוי של התכנים לסיווג נכון הינו רק 60% אך השגיאה יורדת ל40%. אמנם הסיכוי לנבא סיווג נכון לתוכן ירד אך גם מדד השגיאה לסיווג לקבוצה זו ירד במספר אחוזים זהה.

מבין שתי הנקודות הנ"ל נעדיף את נקודת העבודה הראשונה. בנקודה זו נוכל לחזות אחוז גבוה של תכנים בעלי רייטינג נמוך העלולים לגרום להגירה שלילית של מנויי "נטפליקס". מצד שני, אחוז השגיאה הינו גבוה ויכול לגרום ל'פספוס' של תכנים איכותיים יותר. סיכון זה אינו מהווה איום מבחינה כלכלית כיוון שבכל מקרה זהו תוכן שלא נרכש ו"נטפליקס" לא הוציאו עליו כסף.

## לסיכום

אלגוריתם Naïve Bayes יעזור לחברת "נטפליקס" לנבא איזה תכנים יקבלו דירוג גבוה ובכך להחליט בעתיד על התכנים שכדאי לה לרכוש.

מבין שתי נקודות העבודה שנבחרו מבין שני **הטווחים אותם בחנו** נבחר את נקודת העבודה העדיפה ביותר עליה נמליץ ל"נטפליקס".

**נקודת העבודה עליה נמליץ היא נקודת העבודה שנבחרה בטווח הדירוגים 8-10 (הנקודה השחורה).** בנקודה זו קיים סיכוי של 60% לבחירת תוכן בעל דירוג גבוה ומנגד 30% לשגיאה. המטרה שלנו היא להמליץ ל"נטפליקס" על תכנים שיביאו לה קהל נוסף. בבחירת נקודת עבודה זו קיימת שגיאה נמוכה יחסית לאחוז ההצלחה. צריך לקחת בחשבון שאחוזי השגיאה מכילים גם תכנים של רייטינג גבוה יחסית (6.5-8). כלומר, גם מתוך 30 אחוזי השגיאה, קיימים תכנים שנרצה שנטפליקס ירכשו.

נקודת העבודה העליונה ב-ROC של טווח הדירוגים 0-5 היא בעלת אחוזי הצלחה גבוהים יותר מנקודת העבודה שנבחרה בטווח הקודם. לעומת זאת, היחס שלה בין אחוזי ה-TP ל-FP הוא נמוך מהיחס של הנקודה הנבחרת בטווח 8-10 (1.34 לעומת 2).

בנוסף, המטרה העיקרית שלנו היא לחזות בשביל "נטפליקס" תכנים בעלי דירוג גבוה ולא למנוע מהם קנייה של תכנים בעלי דירוג נמוך. לכן, בהיבט העסקי, נמליץ ל"נטפליקס" לרכוש תכנים בעלי שילובי התכונות שדומים לתכנים שסווגו לקבוצת הדירוגים 8-10. כאשר, לפי המסקנות שהגענו להם יש את הסיכוי להיות הכי מדויקים. בכך, "נטלפיקס" יצליחו לרכוש סדרות בעלות רייטניג גבוה וימזערו קניית תכנים עתידיים בעלי רייטניג נמוך.