**עבודת גמר – מבוא למדעי הנתונים סמסטר א׳:**

דינה מיזלס-319092201

בחרתי לעשות את העבודה על חיזוי רייטינג סרטים.

כאשר בתחילת העבודה בחרתי כי הרייטינג יהיה מחולק ל6 קטגוריות רייטינג – 0,1,2,3,4,5

**השארת אפס:** הדירוג יהיה הממוצע בין 0-5 כלומר 3.5 כאשר נבצע עליו - (round(0

-כלומר מתעגל ובמקרה שלנו יתעגל ל 3 .

זאת השארת האפס כיון שלקחתי בחשבון כי חלק ידרגו גבוה וחלק נמוך ולכן זה יהיה הממוצע בין השנים.

**ניקו וסידור הדאטה:**

עבדתי עם שלושה דאטות , כאשר דאטה הבסיס הכיל 24 פיצרים, דאטה נוספת שבה היה את הפיצר רייטינג אותו רציתי לחזות ובה היה 4 פיצרים , ודאטה נוספת שתפקידה היה למזג על ידי הפיצרים את שתי הדאטות האחרות אחד לשני .

דאטה הבסיס אשר הכילה 24 פיצרים –

שלב ראשון בדקתי את כמות הערכים החסרים בכל פיצר

'tagline','homepage','belongs\_to\_collection'

וכתוצאה מכך בחרתי למחוק שלושה פיצרים -

כיון שבהם היה יותר ערכים חסרים מקיימם, ועמודה שכזאת עלולה לגרם לחוסר דיוק המודל .

החלטתי ל'מלא' חלק מערכים החסרים בערך המקסימלי או בממוצע או 'אין מידע' או עמודה אחרת שהערכים בהם דומים.

בחלק מהפיצרים אשר לדעתי הייתה משמעות מיוחדת לכל ערך , או שלא מצאתי דרך מדויקת ביותר למילוי הערכים החסרים החלטי שכיון שיש לי בדאטה מעל 45000 שורות יהיה נכון להוריד את אותן שורות של ערכים חסרים מאשר למלא ערכים אילו במידע לא מדויק שעלול להשפיע על המודל.

לאחר ניקוי הדאטה עברתי למיזוג 3 הדאטות וזאת במטרה להכניס לדאטה הבסיס את עמודת הרייטינג שאותה נרצה לחזות.

חיברתי את שני דאטות המשנה על ידי עמודה משותפת ויצרתי דאטה חדשה.

tmdbId

Id

גיליתי שעמודת בעלת ערכים משותפים לעמודת אשר בדאטה הבסיס , לכן החלפתי את שם עמודת המשנה לשם זהה לזה של עמודת הבסיס ומזגתי ביניהם.

החלפת כל המשתנים להיות מסוג נומרי –

חלק המרתי למספרים, בחלק הייתי רק צריכה לשנות את הסוג לנומרי , ובחלק העבודה הייתה קצת יותר מסובכת כמו הפיצר גאנר – בה היו בין גאנר אחת לשבע גאנרים בכל שורה ולכן בניתי פונקציה שתפריד בין הקטגוריות , לקחתי את עשרת הקטגוריות העיקריות המרתי אותם למספרים, וכך גם את השורות.

עוד דבר שנתקלתי בו הוא שם הסרט או תקציר הסרט , והחלטתי לבדוק את השפעת מספר האותיות שלהם על המודל ולכן יצרתי פונקציה שתפקידה לספור את האותיות בשם הסרט וכו...

רציתי לחזות את הרייטינג לכן חילקתי אותו ל6 קטגוריות .(על ידי פקודה שמעגלת את המספר )

חלקתי את הדאטה 30% אימון ו70% בדיקה .

לאחר הרצת פקודה למציאת המודל קיבלתי מטריצת שגיאות בגודל 6\*6 , במטריצה היו שלושה עמודות בעלות ערכים של אפס עמודה אחת בעלת ערכים מספרים שונים ועמודה נוספת בעלת ערכים של אפס עם מספרים נוספים.

הסתכלתי שוב על מספר הערכים שהיו לי בכל קטגוריה של הרייטינג חשבתי ונעזרתי באינטרנט והבנתי כי הערכים של האפס הם כתוצאה מכך שלא היה למודל מספיק מידע בכדי לסווג.

בעקבות כך החלטתי לחזור על התהליך אך הפעם במקום לחלק את הרייטינג ל 6 אני אחלק ל 3 ואכן המודל סיוג בצורה טובה יותר אך עדיין לא מספיק ועדיין היו ערכים של אפס שהראו כי אין מספיק מידע למודל.

לכן החלטתי לנסות שוב ולחלק את החיזוי רייטינג לשני עמודות .

לדעתי המודל לא היה טוב ולא היה בו מספיק מידע שאפשר לו לסווג בצורה טובה(כיון שרוב הסרטים היו מדורגים בטווח שבין 3-4)

KNN

בנוסף השתמשתי באלגוריתם בוא בדקתי את השכנים הקרובים מ1-30 , כאשר הדיוק הגבוה ביותר היה כאשר לפי 30 השכנים הקרובים.

Naive bayes

השתמשתי באלגוריתם נוסף לחיזוי- בוא השתמשתי בחלוקה של הרייטינג לשני סיוגים- ובכך ניסיתי לחזות את ההסתברות לרייטינג.

המודל לא היה טוב מהשארת האפס.

נהניתי מאד לעשות את העבודה , למדתי הרבה על מהלך ניקוי הדאטה ועל דרכים ופקודות שונות .