**דוח** **פרויקט**

**מגישים: 208721373 +209494756**

**חלק ראשון : מודל**

תחילה הקמנו תיקייה לפרויקט ופתחנו את קובץ הTRAIN ו TEST.

לאחר שהטענו את כל החבילות שרלוונטיות לקוד עברנו על העמודות ובחנו את המידע , סימנו את העמודות שהן לרוב עם ערכים כדי לדעת אילו עמודות פחות משפיעות על הניבוי .

בלמידת מכונה קיימים שתי דאטות בסיסיות שמחולקות לTRAIN ו TEST לשם כך נבנה את המודל על הTRAIN ואז נריץ על הTEST על מנת לתת חיזוי, על מנת לבדוק את המידע בצורה הטובה ביותר החלטנו לעשות cross validation המפצל את המידע למספר מקטעים וכל פעם בודק קטע אחד בהשוואה לקטעים האחרים.

בחרנו לחלק ל שישה מקטעים כי מצאנו שחלוקה ליותר מקטעים או פחות מקטעים מביאה תוצאות ניבוי פחות טובות .

נציין שלכל מודל ביצענו cv מארבע מקטעים ועד עשרה מקטעים.

הסרנו את העמודות הבאות מתוך ההנחה שהן לא עמודות המשפיעות על הניבוי -

question , exclamation, punctuation, questions\_ratio, excs\_ratio, fBombN, f\_ratio, anger , anticipation, disgust, fear, joy, negative, positive, sadness , surprise, trust

כעת נפרט על סוגי המודלים שבחרנו כחלק מהתהליך:

**random forest**

תחילה הרצנו את random forest שמתבסס על עץ החלטות ששואל שאלות של כן/לא שמחלק את המידע בצורה המדויקת ביותר שיש.

כל ענף בעץ מחזיר תשובה ולבסוף התשובה עם מספר הקולות הרבים ביותר היא התשובה הנבחרת ,מודל זה הפיק (59009) (59504) אחוז.

**BAGET CARET**

אחרי מודל זה עברנו למודל לינארי שמשתמש בlikelihood וצורת קנסות לשגיאות , מודל שמסוגל לעבוד עם צורות רבות של מידע .

נעשה שימוש באותה חלוקה של ה test ואותו מספר של cv אך המודל לא היה מדויק יותר(55841)

**Logistic model tree**

מודל לינארי זה קרוי גם כמודל בינומי המתבסס על הסתברות , טוווח הערכים נע בין מינוס אינסוף לאינסוף , מודל זה לא הניב תוצאות טובות יותר (57524)

**Knn**

בנק זו החלטנו לחזור אחורה ולנסות להשתמש ב מודל KNN (KNEAREST NEIGHBOR) אשר מתבסס על סביבה הקרובה של מידע הדומה זה לזה או במצב שלנו על תצפיות דומות על מנת לתת מענה בצורת ניבוי לשאלה מסוימת, מודל זה הרגיש מראש פחות ונכחנו לראות שאכן התוצאה הנמוכה היותר באה ממנו (52376)

**GLMNET**

המודל הבא שניסינו היה glmnet שעובד על אותו סגנון כמו baget caret והוא הפיק (57821)

**Conditional Inference Tree**

מודל שמתבסס על כך שאין צורך להגדיר פרמטר כדי להשתמש בעץ ההחלטות בצורה רקורסיבית.

מודל זה הפיק (56633)(57524)

**ROC-Based Classifier**

מודל שמאפשר לנו להעריך ולהצדיק את ביצועי המודל במערך הנתונים, עקומת **ROC** מייצגת את עקומת ההסתברות של הערך

ומטרתה למצוא את הקשר בין שיעור חיובי אמיתי של מודל ושיעור חיובי כוזב

באמצעות המודל קיבלנו (54563)(56633).

לבסוף, המודל האחרון שבדקנו הוא:

**Conditional Inference Random Forest(CFOREST)**

מתבסס על הדרך הסטנדרטית של חלוקה לעצים עם שינוי קל בו שבמודל זה אין יחס מועדף למשתנה אחד יותר מלאחר

מודל זה הביא מספר תוצאות אשר אחת מהן היא הנוכחית (59306)(59207)(59900)(59801)(58910)(59108)

**חלק שני : מחקר נתונים וניתוח**

**א . בחירת המשתנים ותמצית**

בחלק זה התחלנו לחשב ממוצעים שונות וסטיות תקן לעמודות לפי החלוקה של עמודת label( ערכים 0 ו 1 ).

המטרה של החישובים הייתה לבדוק באילו עמודות קיימים ההבדלים המשמעותיים יותר בין ציוץ סרקסטי לבין ציוץ לא סרקסטי

לאחר חישוב ידני באמצעות אקסל לכלל העמודות ובדיקה באמצעות R מצאנו כי קיימים מספר עמודות שבהן השוני בין ציוץ סרקסטי לציוץ לא סרקסטי הוא מובהק יותר .

מצאנו כי הפערים המשמעותיים מבחינת ממוצע הם בעמודות הבאים:

wordN:

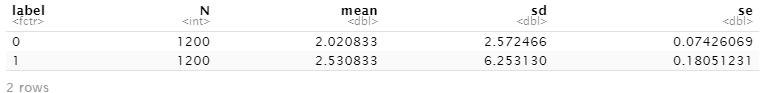
1. 11.15500
2. 10.81667

Table

Description automatically generated with low confidence

CapsN:

1. 2.020833
2. 2.530833



בנוסף בעמודת subreddit בדקנו לפי קטגוריה איפה יש פער בין ציוץ סרקסטי לציוץ לא סרקסטי

ומצאנו כי בעמודות הבאות הפער הוא המשמעותי יותר

Worldnews:

1. 175/274 = 0.64 %
2. 99/274 = 0.36 %

Politics:

1. 166/434 = 0.38 %
2. 268/434 = 0.62 %

Pcmasterrace:

1. 77/203 = 0.37 %
2. 126/203 = 0.63 %

Nfl:

1. 56/145 = 0.38 %
2. 89/145 = 0.62 %

News:

1. 65/189 = 0.34 %
2. 124/189 = 0.66 %

Askreddit:

1. 365/629 = 0.65 %
2. 264/329 = 0.41 %

**ב. גרפים:**

יצרנו מספר גרפים כדי להראות את הנתונים שניתחנו .

Chart, box and whisker chart

Description automatically generatedגרף ראשון – אורך מילה ממוצעת לפי label כאשר הגרף הוא מסוג boxplot המתאר את הממוצע בקו האמצעי והקווים הרציפים האופקיים של הגרף מסמנים את האחוזים 75 ו 25 (עליון ותחתון) מבחינת הטווח . הנק השחורות מסמנות את המילים שבהן הערכים הקיצוניים שמראים אם הם גבוהים מאוד או נמוכים מאוד מהממוצע

Chart, histogram

Description automatically generatedגרף שני – שני גרפים שמתארים את הספירה של האותיות לפי label , הכחול מסמל ציוץ סרקסטי והאדום מסמל ציוץ לא סרקסטי

גרף מספר שלוש – גרף שמתאר את הכמות של הציוצים סרקסטים (label 1) והציוצים הלא סרקסטים (label 0) בכל נושא שקיים תחת עמודת subreddit

A screenshot of a computer

Description automatically generated with medium confidence

**ג.מבחנים סטטיסטים**

ערכנו 2 מבחנים בין על מנת לראות האם הטענות שלנו לגבי סימני קריאה ו הנושא שצוייץ לגביו קשורים לניבוי

המבחן הראשון שביצענו הוא מבחן t

אנחנו מאמינים שסימני קריאה יכולים לעזור לחזות בין תגובות שיש בהן מילים סרקסטיות לתגובות שאין בהן.

השערת האפס - אין הבדל בכמות סימני קריאה בין ציוצים סרקסטים לבין לציוצים לא סרקסטים

השערת האחד - ישנו הבדל בכמות סימני קריאה בין ציוצים סרקסטים לבין לציוצים לא סרקסטים

Graphical user interface, text

Description automatically generated

ניתן לפסול את השערת האפס לפי p value ולכן ניתן להגיד שסימני קריאה עשויים לעזור בניבוי בין ציוצים סרקסטים לבין לציוצים לא סרקסטים

המבחן השני הוא מבחן חי בריבוע אי תלות

אנחנו מאמינים שישנה תלות בין label ו subreddit

השערת האפס – אין תלות בין label ו subreddit

Graphical user interface, text

Description automatically generated

ניתן לראות שלפי p value ניתן לדחות את השערה ה אפס וישנה תלות בין label ו subreddit