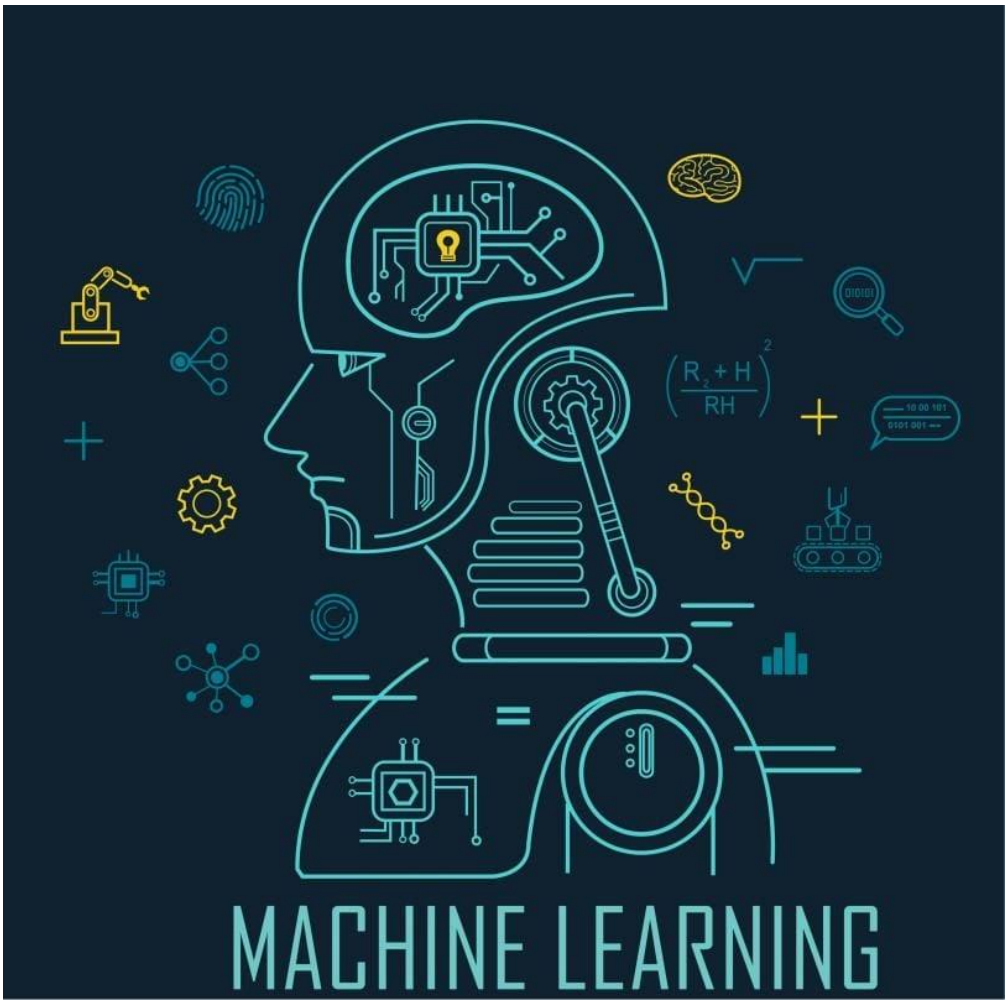
פרויקט לימוד מכונה חלק ב'



מגיש:

יאיר גזית- 314720517

תוכן עניינים

Model Training

- בחרתי להשתמש בCV על ידי holdout כאשר חילקתי את סט האימון 80% וסט הואלדציה 20%. נבדוק את המדד AUC-ROC על הסט האימון והואלידציה.

Decision Trees

1) Hyperparameter Tuning: בחרתי להשתמש בGrid search עם cv=10.

<u>ההיפר פרמטרים אותם בחרתי לכוון:</u>

- Criterion: באיזה מדד העץ בוחר את המאפיין על מנת לפצל אותו, בחרתי באופציה Entropy אשר נלמדה בכיתה.
- Max_depth: עד איזה עומק העץ יגדל. טווח הערכים שהגדרתי הינו 3-5 מכיוון שחששתי מoverfitting. האופציה שנבחרה היא: 4
- Min_samples_leaf: המספר המינימאלי של דגימות הנדרש להיות עלה בעץ, טווח הערכים שלקחתי הינו 2,3,4,5,6,8,10,12 הערך שנבחר הינו 2
- Min_samples_split: המספר המינימאלי של דגימות הנדרש לפיצול צומת פנימית בעץ. טווח הערכים שלקחתי הינו 8,10,12, הערך שנבחר הינו 8
- ccp_alpha: הסף שמעליו ענפים של עץ ההחלטה יגזרו, מסייע למניעת over fitting על ידי גזירת חלקים מהעץ שאינם מספקים שיפור משמעותי. טווח הערכים שלקחתי הינו 0.01,0.015,0.02,0.025,0.03 האופציה שנבחרה היא 0.01

<u>אחוזי הדיוק שקיבלתי:</u>

אחוזי הדיוק המתקבלים על ידי AUC-ROC עבור המודל הטוב ביותר:

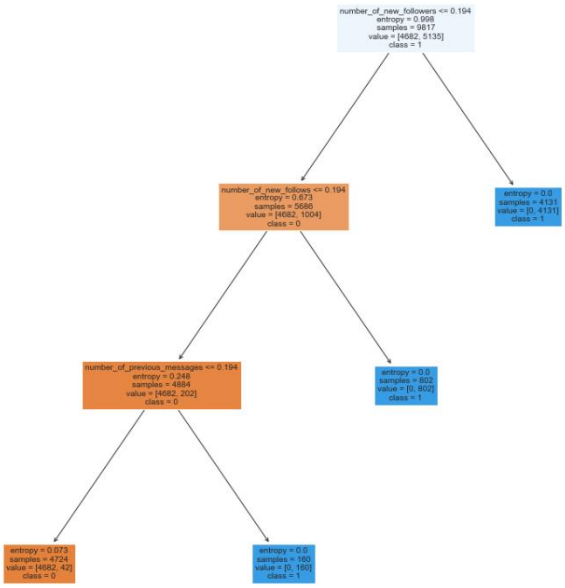
סט אימון: 0.996 סט ואלידציה: 0.995 המשמעות היא שיש כאן חשש לoverfitting, או שסט הנתונים שקיבלנו אינו מורכב מספיק וכי מודל הלמידה לומד באופן כמעט מושלם את הסיווגים למחלקות שלילי וחיובי. התוצאות קרובות בין סט האימון לסט הואלידציה אך קרובות מאוד ל1. תוצאות אלו נובעות מכך שבחרתי למודל היפר פרמטרים מתאימים ושאופן לימוד הdataset יחסית טוב ומתאים לו, אך ניתן לראות כי ישנם 3 מאפיינים בלבד אשר העץ מתבסס עליהם. <u>נספח 1</u>

2) Interpretability:

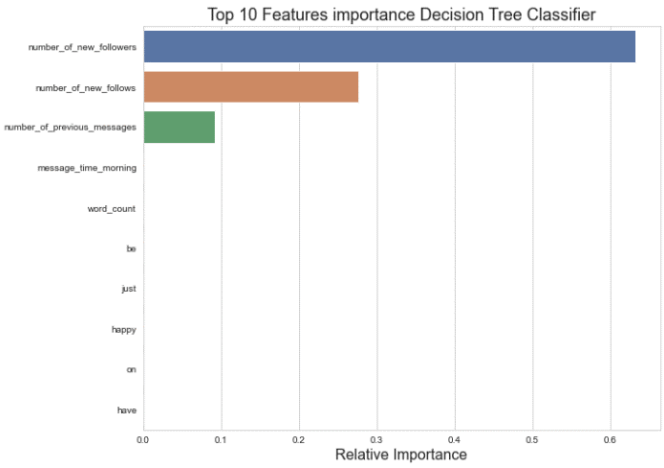
עץ ההחלטה הוא אלגוריתם המייצר תנאים של אם...אז.., לכן יכולת ההסברה של עץ ההחלטה מתייחסת ליכולת להבין ולפרש כיצד המודל עושה סיווגים ואת התחזיות שלו, זאת אומרת כיצד הוא קיבל את ההחלטה שלו.

יכולת זו יכולה לסייע לנו במשימת הלמידה בכמה דרכים: ניתן להשתמש בעצי החלטה כדי לקבוע את החשיבות היחסית של כל מאפיין למשתנה המטרה וכך להבין את הגורמים שמשפיעים על התוצאה. בנוסף, ניתן להשתמש בעצי ההחלטה כדי לאבחן שגיאות שנעשו על ידי המודל, על ידי בחינת נתיב ההחלטה המוביל לתחזית שאינה נכונה. כמו כן, ניתן להשתמש בעצי ההחלטה על מנת לזהות אזורים בהם ניתן לשפר את המודל, על ידי בחינת כללי ההחלטה והתוצאות שהתקבלו בעקבותיהם.

3) גרף העץ:



- הפיצולים שנעשו בכל צומת מייצגים את המאפיינים אשר הפרידו בין המחלקות השונות. Number_of_followers בראש העץ, זה מרמז על כך שתכונה זו היא האינפורמטיבית ביותר בהבחנה בין שתי המחלקות ותורמת לסדר הגדול ביותר בנתונים. מורכבות העץ הינה קטנה, כאשר עומק העץ קטן ביחס לכמות הפיצ'רים ולכן ניתן לשער שמורכבות הבעיה אינה גדולה.
- Feature importance:



ניתן להסיק שלמשימת הלימוד שלנו, רק 3 מאפיינים חשובים למודל על מנת שיוכל לסווג את הסנטימנט של ההודעה לחיובי או שלילי. אך בגלל שמודל בחר רק 3 מאפיינים מראה לנו כי אולי יש לבחון אותו בשנית, ולבצע בו שינויים אשר יוכלו לתת לנו תוצאות טובות יותר, המודל לא לקח כלל את מאפייני הטקסט בחשבון, מפאת קוצר הזמן לצערי לא אוכל להתעכב ולבדוק זאת לעומק.

Artificial Neural Networks

1) הרצת המודל הדיפולטיבי:

- מספר הנוירונים בשכבת הכניסה: קלט המודל, מס' הנוירונים ככמות המאפיינים שיש במודל, כמות המאפיינים הינה: 55
- מספר שכבות חבויות: היפר פרמטר בעל כיוונון, ישתנה בהתאם למימד או מורכבות הנתונים. כאשר יש מימד גבוה נרצה פחות שכבות חבויות, כאשר הדאטה עם הרבה מאפיינים ביחס לנתונים נרצה יותר שכבות חבויות. במודל הדיפולטיבי כמו השכבות החבויות היא 1.
- מספר נוירונים בכל שכבה: היפר פרמטר בעל כיוונון, ישתנה בהתאם לממד או מורכבות הנתונים, במודל הדיפולטיבי יש שכבה אחת עם 100 נוירונים
- מספר נוירונים בשכבת היציאה: במשימת סיווג כמו בפרויקט שלנו תהיה ככמות המחלקות שיש למשתנה המטרה כלומר יהיה 0 או 1 בהתאם לסיווג למחלקה הרלוונטית
- אחוזי הדיוק המתקבלים על ידי AUC-ROC עבור המודל הדיפולטיבי: סט אימון: סט ואלידציה: המשמעות היא שגם כאן הייתי צופה לover fitting מכיוון שהתוצאות קרובות ל1, אין הפרש גדול בין שתי התוצאות.

נספח 2

כיוונון היפר פרמטרים:

- Hidden_layer_size: קובע כמה שכבות חבויות יהיו במודל ועבור כל שכבה חבויה, כמה נוירונים יהיו. ככל שתגדל כמות השכבות החבויות או כמות הנוירונים, נפתח מודל מורכב יותר עם זמן ריצה גדול מאוד. במידה ואין צורך בכך בעקבות שלדאטה יש מימד גבוה, יכול להוביל לעוד יותר התאמת יתר ולפגוע בתוצאות המודל על סט המבחן. הקטנת ההיפר פרמטר תתאים למודל שאינו מורכב בעל מימד גבוה וכך נמנע מהתאמת יתר.

  מצד שני, לא יספק תוצאות טובות כאשר המודל הינו מודל מורכב. הטווח שבחרתי הינו בין 50-100 בסופו של דבר הערך הנבחר הינו 50.

- Activation: קובע את פונקציית האקטיבציה שתפקידה לתרגם את הקלט שנכנס לנוירון לפלט שייצא ממנו. בחרתי בין שני פונקציות אקטיבציה: logistic וrelu הסיבה הינה שאני סבור כי פונקציה קבועה לא תתאים לבעיה. לבסוף נבחר בפונקציה logistic.
- Learning_rate_init: קובע את קצב ההתקדמות בפונקציית הLOSS בכל איטרציה, שמעדכנת את המשקלים. המוטיבציה לכוונן אותו היא שיש טרייד-אוף בין זמן ריצה לבין הדיוק במציאת מינימום בפונקציה. המשמעות של הגדלת פרמטר זה היא שזמן הריצה יהיה יותר מהיר אך הסיכוי לפספס מינימום יגדל ולהפך. טווח הערכים שבחרתי הינו 0.001 ,0.0001, לפי התוצאות נראה את ההשפעה של משתנה זה על תוצאות המודל. הקפיצות יחסית גדולות על מנת לחסוך בזמן הריצה. לאחר כיוונון נבחר 0.001
- Alpha: היפר פרמטר זה שולט בעוצמת הרגוליזציה, שהיא טכניקה למניעת התאמת יתר ברשתות נוירונים, על ידי הענשת פונקציית הפסד הגדלת האלפא תגדיל את הרגוליזציה תפחית את התאמת היתר ולהפך. זהו היפר- פרמטר חשוב לכיוונון, טווח הערכים שבחרתי הוא 0.01, 0.001, 0.0001. הערך הדיפולטיבי הוא 0.0001. רציתי לבחון את ערכים גדולים יותר כדי להפחית את התאמת היתר. לאחר כיוונון הערך הנבחר הוא 0.01

אחוזי הדיוק במודל המדויק ביותר הם: סט אימון: 0.999 סט ווליידציה: 0.999 נספח 2.1

- ניתן להסיק מתוצאות אלו כי ישנה חשיבות רבה לכוונון ההיפר-פרמטרים, מכיוון שהתוצאות השתפרו מהמודל הדיפולטיבי ולכן סט הנתונים שנבחר הינו מתאים.
- ההיפר-פרמטרים במודל הדיפולטיבי אינם מותאמים לכל סט נתונים ולכל משימת לימוד ולכן יש הבדל באחוזי הדיוק בין המודלים

היפר פרמטר	ערך
Hidden_layer_size	50
Activation	logistic
Learning_rate_init	0.001
alpha	0.01

2) מטריצת מבוכה

ניתן להסיק כי המודל מסווג בצורה טובה סנטימנט חיובי או שלילי של ההודעה, סך הכל יש לו מרווח טעויות קטן, עדיין נחשוש מפני over fitting.

```
Confusion matrix:
[[1157    7]
 [  21 1270]]
```

השוואה בין המודלים:

אלגוריתם	סט אימון	סט ואלידציה
עץ החלטה	0.996	0.995
ANN	0.999	0.999

1) למרות שANN בעל ערך טוב יותר במדד הROC-AUC של סט האימון וסט הואלידציה, אני אבחר במודל עץ ההחלטה מכיוון שהציון שלו גם גבוה, אך נראה שבהשוואה מבין שני האלגוריתמים, יש לו פחות פוטנציאל לover-fitting
2) אני אבחר במודל עץ ההחלטה, הדיוק שהתקבל על הסט הואלידציה הינו: 0.995

שיפור המודל הנבחר:

בעיות שזיהיתי במודל במהלך העבודה:

- בעיה במודל- עומק העץ הנבחר – היפר פרמטר זה עומד על 4 בלבד, מה שאומר שהאלגוריתם לא רואה במורכבות הבעיה, והינו נאמן למאפיינים בודדים בלבד מתוך כל המאפיינים הפוטנציאלים.
- לפי תרשים גרף העץ שלעיל, ניסיתי להוריד את הפיצ'רים שעושים סדר באנטרופיה בצורה קיצונית, שאלו הם number_of_new_followers, number_of_new_follows, number_of_previous messages. מכיוון שהם לוקחים בחשבון רק את מדדי כמות ואינם מתייחסים כלל לטקסט ההודעה. בנוסף, הם מסווגים באופן כמעט מוחלט את העץ לשני המחלקות. מכיוון שאני רוצה שהמודל יבחן יותר פיצ'רים החלטתי לוותר עליהם- כאשר ויתרתי על הפיצ'רים הללו קיבלתי מודל פחות טוב שעדיין לא מתייחס לחלק נכבד מן הפיצ'רים:

```
Average AUC-ROC score from CV: 0.921
AUC-ROC training score: 1.000
Fitting 10 folds for each of 1680 candidates, totalling 16800 fits
Best parameters found:  {'ccp_alpha': 0.01, 'criterion': 'entropy', 'max_depth': 5, 'min_samples_leaf': 2,
'min_samples_split': 2}
Best AUC-ROC score from CV:  0.9681778641607341
AUC-ROC training score (optimized): 0.970
AUC-ROC validation score: 0.971
                          feature  importances
0   number_of_previous_messages     0.827558
1      Earlier_date_of_follower     0.077911
2        Earlier_date_of_follow     0.073362
3               email_verified     0.021170
4          message_time_morning     0.000000
5              avg_word_length     0.000000
6                         with     0.000000
7                           be     0.000000
8                         just     0.000000
9                        happy     0.000000
10                          on     0.000000
```

ניסיתי לבצע אלגוריתם random forest אשר יודע להתמודד עם בעיות יחסית מורכבות, אך גם אלגוריתם זה נראה שלא מתאים מכיוון שהמודל לא מסווג את הבעיה כמורכבת ואלו התוצאות שקיבלתי:

```
Train AUC-ROC score (without tuning): 1.000
Test AUC-ROC score (without tuning): 1.000
Train AUC-ROC score (with tuning): 1.000
Test AUC-ROC score (with tuning): 1.000
Best hyperparameters: {'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_depth': 10, 'bootstrap':
False}
```

תוצאות אלו מעידות על over-fitting ולכן גם מודל זה אינו מתאים לבעיה שלנו.

- בעקבות כך שסט הנתונים הראשוני שקיבלתי היה זקוק לתיקונים רבים, נאלצתי לבצע פעולות רבות על בסיס מסקנות אישיות. כנראה שפעולות אלו גרמו לעיוות כלשהו בנתונים האמיתיים ולכן גם לעיוות מסוים בתוצאות.

נספחים

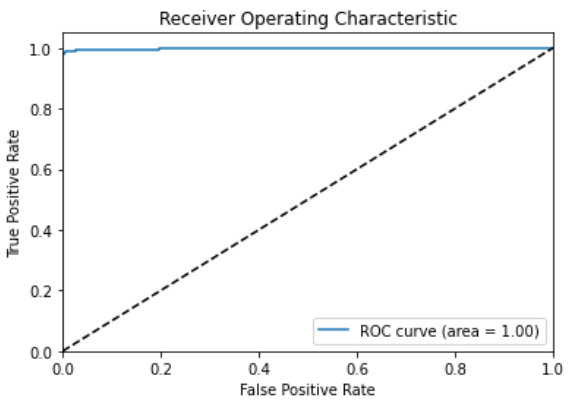
נספח 1:

```
Fitting 10 folds for each of 360 candidates, totalling 3600 fits
Best parameters found:  {'ccp_alpha': 0.01, 'criterion': 'entropy', 'max_depth': 4, 'min_samples_leaf': 2,
'min_samples_split': 8}
Best AUC-ROC score from CV:  0.9959104148178488
AUC-ROC training score (optimized): 0.996
AUC-ROC validation score: 0.995
                          feature  importances
0        number_of_new_followers     0.632223
1          number_of_new_follows     0.276068
2    number_of_previous_messages     0.091708
3           message_time_morning     0.000000
4                     word_count     0.000000
5                             be     0.000000
6                           just     0.000000
7                          happy     0.000000
8                             on     0.000000
9                           have     0.000000
10                          good     0.000000
11                            me     0.000000
12                           day     0.000000
13               avg_word_length     0.000000
14                sentence_count     0.000000
15                  message_year     0.000000
16                    char_count     0.000000
17                           not     0.000000
18                 email_company     0.000000
19         Latest_date_of_follow     0.000000
20      Latest_date_of_follower     0.000000
21       Earlier_date_of_follow     0.000000
22     Earlier_date_of_follower     0.000000
23                     seniority     0.000000
24                    blue_tick     0.000000
25                email_verified     0.000000
26                          with     0.000000
27                           was     0.000000
28         message_time_evening     0.000000
29                            no     0.000000
30       message_time_afternoon     0.000000
31                    platform_x     0.000000
32            platform_whatsapp     0.000000
33              platform_tiktok     0.000000
34             platform_telegram     0.000000
35           platform_instagram     0.000000
```

נספח 2: גרף ותוצאות ANN

```
Train AUC-ROC score: 1.000
Train Accuracy score: 1.000
Test AUC-ROC score: 0.997
Test Accuracy score: 0.982
[[1125   13]
 [  31 1286]]
```

:Roc curve



נספח 2.1: גרף ותוצאות לANN הנבחר

```
Best parameters found: {'learning_rate_init': 0.001, 'hidden_layer_sizes': (50, 50), 'alpha': 0.01, 'activation':
'logistic'}
Train AUC-ROC score: 0.999
Test AUC-ROC score: 0.999
Confusion matrix:
[[1157    7]
 [  21 1270]]
```

:ROC Curve