פרויקט – חלק ב'

Table of Contents

Contents

Decision Trees

1. בניית עץ החלטה ותהליך Hyperparameter Tuning

```
decision_trees_param_grid = {
    'max_depth': np.arange(1, 51, 1),
    'criterion': ['entropy', 'gini'],
    'min_samples_split': np.arange(2, 101, 2)
}
```

1.1. בחרנו לכוונן את הפרמטרים הבאים:

- Max_Depth – פרמטר זה קובע את העומק המקסימלי של העץ. ככל שהעומק גדול יותר, כך העץ יכול לתפוס קשרים מורכבים יותר בדאטה אולם ישנה סכנה ל- overfitting ולפגיעה ביכולת ההכללה של המודל. מנגד, עץ בעל עומק נמוך אמנם יהיה פשוט יותר, אך לא בהכרח יתפוס את המורכבויות שבדאטה. טווח הערכים שבדקנו הוא בין 1 ל-50.

- Criterion – פרמטר זה קובע על פי איזה מדד האלגוריתם יבחר את הפיצ'ר הכי טוב לפיצול העץ. בחנו חישוב אי הודאות באמצעות שני המדדים הבאים: ['entropy', 'gini']. מדד gini מחשב את ההסתברות לסווג בצורה לא נכונה סאמפל שנבחר אקראית אם הוא סווג רנדומלית בהתאם להסתברות לקבל כל לייבל. לעומת זאת, מדד האנטרופיה בודק את כמות הידע שנרוויח מבחירת כל פיצ'ר.

- Min_Sample_Split – קובע את מספר הסאמפלים המינימלי הדרוש כדי לפצל את העץ בהתאם לפיצ'ר מסוים. ערכים גבוהים בפרמטר זה אמנם יובילו לעצים פשוטים יותר עם סיכוי נמוך יותר להתאמת יתר, אך תיתכן פגיעה בתפיסת מורכבויות בנתונים ואילו ערכים נמוכים יובילו לאפקט ההפוך. טווח ערכים בין 2 ל-100 בקפיצות של 2.

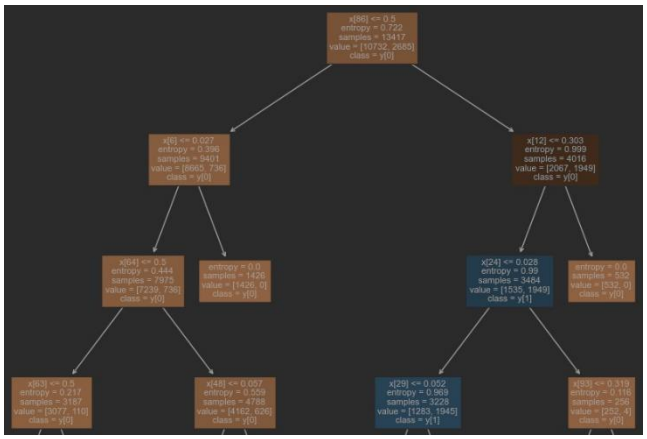
1.2. אחוזי דיוק שהתקבלו עבור המודל הטוב ביותר - ניתן להסיק כי המודל למד בצורה טובה את סט האימון, ויש לו יכולת הכללה טובה בגלל שהוא נותן ביצועים טובים גם עבור סט הוולידציה. מעבר לכך, ניתן לומר שהפיצ'רים שנבחרו מכילים הרבה מידע שעוזר למתן תחזית מדויקת.

```
Training accuracy:  0.991
Validation accuracy:  0.971
```

2. יכולת ההסברה של המודל

מודל עץ ההחלטה מסוגל לתת הסבר ברור ומובן לתחזיות שהוא מספק בניגוד להרבה מודלים אחרים, מה שצריך לעשות הוא לעקוב מהשורש עד לעלה המבוקש כדי להבין למה המודל סיווג אותו בלייבל מסוים. יכולת זו מקנה את האפשרות להבין אילו פיצ'רים הם החשובים והמשפיעים ביותר על הסיווג הסופי של המודל, מה שייתן לנו תובנות בנוגע למודעות ומה משפיע יותר על הסיווג שלהן. במידת הצורך נוכל לעדכן את הדאטה סט לאחר מכן לאור התוצאות שנקבל.

3. גרף עץ ההחלטה (העץ המלא – נספח 1)



3.1. מסקנות מהגרף:

- הפיצול הראשון שתורם הכי הרבה ליכולת לסווג את המודעות הוא על הפיצ'ר qualification_entrylevel, כאשר מודעה תקבל את הערך 1 אם במאפיינים required education / experience מופיעות מילים כמו: high school או entry level. מודעות עם ערך 1 ישלחו לצד ימין של העץ עם סבירות גבוהה יחסית להיות מסווגות כמזויפות, ואילו מודעות עם ערך 0 יסווגו לצד שמאל בו רוב הסיכויים שיסווגו כאמיתיות. ניתן להסיק כי הפיצ'רים המתייחסים לדרישות הקדם של המשרה מעידים ברמה טובה על סיווג המודעה, כאשר דרישות קדם נמוכות יובילו לסיווג מזויף בסבירות גבוהה יותר.

- לאחר הפיצול הראשון, ניתן לראות בצד ימין אנטרופיה מאוד גבוהה (0.999) ממנה ניתן להסיק כי כמות הסאמפלים שסווגו כשאמיתיים וכמזויפים די זהה. כלומר, התפלגות שווה בין שתי הקטגוריות. לעומת זאת, הצומת בצד שמאל עם אנטרופיה נמוכה יחסית של 0.396 אשר מעידה על הפיצול הראשון שהוא איכותי ביכולת הפללה שלו.

- הפיצ'רים לפיצול בשכבה השנייה שניהם מתייחסים למאפיין company ולמילים מסוימות שהוא מכיל. ניתן להסיק מכך כי הכיתוב במודעות בנוגע לחברה משמעותי ליכולת לסווג את המודעה. בנוסף, ניתן לראות כי שני הצמתים בהן הסאמפלים מסווגים כמודעות מזויפות (הכחולים) מפוצלות גם הן על ידי פיצ'רים שנובעים מהכיתוב ב company, אם המילים 'include' או 'get' מוכלות במאפיין זה. נתון זה מדגיש את החשיבות של מאפיין זה לסיווג המודעות.

3.2. פונקציית חשיבות:

```
Features importance scores:
qualification_Entry level: 0.21
company_drive: 0.06
description_product: 0.04
description_client: 0.04
company_includ: 0.04
company_get: 0.03
description_sale: 0.03
company_base: 0.03
country_US: 0.02
company_platform: 0.02
```

- נציג כאן את 10 הפיצ'רים עם החשיבות הגבוהה ביותר. באופן כללי ניתן לראות את ההתאמה גבוהה בין אופן פיצול העץ לבין חשיבות הפיצ'רים.

- הפיצ'ר עם החשיבות הגדולה ביותר באופן משמעותי לעומת יתר הפיצ'רים ממוקם בשורש העץ ומתייחס לרקע הלימודי והתעסוקתי של המשרה במודעה המפורסמת.

- ניתן לראות כי הפיצ'רים הקשורים לטקסט שמופיע ב- company הם משמעותיים ביותר. אם נסכום את חשיבות פיצ'רים אלה מבין 10 הפיצ'רים המשמעותיים ביותר נגיע ל- 0.18 (הפיצ'ר המשוקלל יהיה מדורג שני בחשיבותו). חשוב לציין כי ישנם פיצ'רים נוספים ששייכים לאותו מאפיין מקורי שאינם כלולים בעשרה הראשונים. ניתן להסיק מכך כי טקסט זה קריטי ביותר לסיווג של מודעות כמזויפות או אמיתיות.

Artificial Neural Networks

1. הרצת רשת הנוירונים בערכי ברירת מחדל

1.1. משמעות הקונפיגורציה - מספר הנוירונים בשכבת הכניסה יהיה כמספר הפיצ'רים בדאטה סט. במקרה שלנו מדובר ב-100 פיצ'רים והמשמעות היא שהרשת תוכל ללמוד באמצעות כל אחד מהפיצ'רים. ברשת ברירת המחדל ישנה רק שכבה חבויה אחת עם 100 נוירונים, כאשר כל נוירון ברשת יקבל קלט מכל אחד מהנוירונים בשכבת הכניסה. בשכבת הפלט יהיה נוירון אחד בסיווג בינארי שמייצג את החיזוי של המודל אם מדובר במודעה אמיתית או מזויפת. ככלל מדובר בקונפיגורציה די פשוטה שתרוץ יחסית מהר, אך לא בהכרח תצליח לזהות דפוסים וקשרים מורכבים.

1.2. אחוזי דיוק שהתקבלו - ניתן להסיק כי המודל למד בצורה טובה את סט האימון. בנוסף, ניכר כי יש לו יכולת הכללה טובה לאור הביצועים הטובים עבור סט הוולידציה. מאחוזי הדיוק שהתקבלו נראה שאין התאמת יתר.

```
Training accuracy:  0.988
Validation accuracy:  0.969
```

2. תהליך Hyperparameters Tuning

2.1. בחרנו לכוונן את הפרמטרים הבאים:

- מספר שכבות חבויות – כיוונון של פרמטר זה משפיע על היכולת לתפוס דפוסים מורכבים בדאטה, ובהתאם אם יצרנו התאמת יתר או היעדר התאמה של המודל לנתוני האימון. הוספת שכבות תגדיל את היכולת של המודל לזהות דפוסים מורכבים מצד אחד, ומצד שני עלולה להוביל להתאמת יתר ולזמן ריצה ארוך יותר. בחרנו בטווח ערכים בין שכבה אחת לשלוש שכבות כדי לבצע זיהוי דפוסים מורכבים יותר בדאטה אך יחד עם זאת לא לגרום לזמני ריצה ארוכים מדי.

- מספר נוירונים בכל שכבה – כיוונון פרמטר זה משפיע גם הוא על יכולת המודל לזהות דפוסים מורכבים. גם כאן הגדלה של מספר הנוירונים תאפשר יכולת למידה של דפוסים מורכבים מצד אחד, אך מצד שני עלולה להוביל להתאמת יתר וזמני ריצה ארוכים. בחרנו לבחון כמות נוירונים בקפיצות קבועות – 50, 100, 150, 200, 250 ובמספר זהה אם היו 2 או 3 שכבות חבויות. בחרנו בערכים אלה לאחר מחקר באינטרנט בהקשר לכמות נוירונים סבירה עבור הדאטה סט שלנו וכמו כן לא רצינו להעמיס על זמן הריצה של האלגוריתם ולעבור דרך כל ערך שלם של נוירונים בטווח מסוים.

- פונקציית אקטיבציה – פונקציות אקטיבציה מאפשרות לרשת ללמוד דפוסים לא לינאריים בדאטה באופן שונה, ובכך הן משפיעות על היכולת של הרשת ללמוד את הנתונים ולבצע הכללה בעתיד. פונקציית Relu מתאימה יותר לרשתות עמוקות כי היא פשוטה ומאפשרת זמן ריצה מהיר והיא טובה בזיהוי דפוסים מורכבים. ואילו פונקציית Logistic (סיגמואיד)

מתאימה לרשתות פחות עמוקות ולמשימות של סיווג בינארי (כמו במקרה שלנו) והיא פחות טובה בזיהוי דפוסים מורכבים. בחנו את שתי הפונקציות הנ"ל.

- Learning_rate_init – כיוונון פרמטר זה מאפשר להשפיע על מהירות תהליך הלמידה כאשר בחירת ערך מתאים תוביל להתכנסות מהירה יותר של הרשת לערכה האופטימלי. ככל שהערך גבוה יותר זה יכול להוביל להתכנסות מהירה יותר של הרשת לערכה האופטימלי כי כל צעד במהלך Epoch יהיה גדול יותר. מצד שני, זה עלול להוביל לפספוס של המשקולות על הקשתות עם הערכים האופטימליים. כאשר הערך של פרמטר זה הוא נמוך יותר אמנם הסיכוי לפספס את הערך האופטימלי הוא נמוך יותר, אולם ההתכנסות של הרשת תהיה איטית יותר יחד עם סכנה להתכנסות לערכים לא אופטימליים.

2.2. טבלה:

בטבלה זו ניתן לראות את ההיפר-פרמטרים השונים כפונקציה של אחוז הדיוק על סט הוולידציה. כפי שניתן לראות, מוצגות 20 הקומבינציות המובילות מתוך 48 סה"כ. ניתן לראות כי קצב למידה של 0.01 הועדף על פני כל האופציות למעט שתיים ונראה שהוא מתאים יותר לדאטה שלנו. כמו כן, בקומבינציות הטובות ביותר כמות השכבות החבויה הייתה נמוכה, אחת או שתיים, וניתן להסיק מכך כי הדאטה שלנו לא בהכרח מורכב מאוד ולכן אינו מצריך שימוש ביותר שכבות חבויות. באופן כללי ניתן לומר כי גם אחוזי הדיוק על סט האימון וגם על סט הוולידציה הם גבוהים, אין פערים גדולים ביניהם וניתן להניח שלרשת יש יכולת הכללה טובה.

	num_layers	hidden_layer_sizes	activation	learning_rate_init	train_score	val_score 1
14	1	(200,)	logistic	0.01	0.999369	0.995131
6	1	(100,)	logistic	0.01	0.999319	0.994992
7	1	(100,)	logistic	0.10	0.998102	0.994974
10	1	(150,)	logistic	0.01	0.999340	0.994903
2	1	(50,)	logistic	0.01	0.999226	0.994872
0	1	(50,)	relu	0.01	0.999013	0.994852
24	2	(150, 150)	relu	0.01	0.999203	0.994604
8	1	(150,)	relu	0.01	0.999242	0.994229
30	2	(200, 200)	logistic	0.01	0.999213	0.993930
42	3	(150, 150, 150)	logistic	0.01	0.999181	0.993836
16	2	(50, 50)	relu	0.01	0.999011	0.993768

2.3. אחוזי דיוק ברשת הטובה ביותר – ברשת הטובה ביותר המופיעה בראש הטבלה, ניתן לראות כי אחוזי הדיוק בסט האימון ובסט הוולידציה הם גבוהים מאוד ובלי הפרש משמעותי ביניהם. נסיק מכך כי המודל בעל יכולת הכללה ואינו במצב של התאמת יתר.

2.4. הבדל בין רשת ברירת מחדל לרשת הטובה ביותר – אחוז הדיוק על סט האימון ברשת הטובה ביותר הוא גבוה ב-0.002 מרשת ברירת המחדל, ואילו אחוז הדיוק על סט הוולידציה גבוה ב-0.008. על ידי כיוונון היפר-פרמטרים ניתן למצוא קונפיגורציה שתואמת יותר את אופי הדאטה ובהתאם לתת תוצאות דיוק טובות יותר לעומת הקונפיגורציה של ברירת המחדל.

3. <u>מטריצת מבוכה</u> – מתוך 13,417 סאמפלים סה"כ, 13,305 סווגו נכונה (בין אם מזויף או אמיתי) כלומר 99.1% דיוק בסיווג. ההסתברות של המודל לסווג מודעה כאמיתית

```
[[10659    73]
 [   39  2646]]
```

במידה והיא אכן אמיתית הוא 99.6% לעומת ההסתברות של המודל לסווג מודעה כמזויפת במידה והיא אכן מזויפת שהוא 97.3%.

SVM

1. אימון מודל וביצוע Hyperparameter tuning

1.1. טבלת סיכום היפר-פרמטרים:

היפרפרמטר	טווח ערכים שנבדק	ערך נבחר במודל הטוב ביותר
C	0.1, 1, 10, 100	1
Penalty	L2	L2
Loss	Hinge, Squared_Hinge	Squared_Hinge

1.2. אחוזי דיוק המודל:

```
Training accuracy: 0.928
Validation accuracy: 0.923
```

1.3. משוואת הישר המפריד:

- מקדמי הפיצ'רים במשוואת הישר המפריד משקפים את המשקולות שהוקצו לכל תכונה בתהליך הסיווג של מודעה כמזויפת או לא. כאשר ערך מוחלט גבוה יותר מצביע על השפעה חזקה יותר של אותו פיצ'ר על גבול ההחלטה. הסימן החיובי או השלילי של המקדמים מציין את כיוון ההשפעה. מקדמים חיוביים מצביעים על כך שהגדלת הערכים של הפיצ'ר המתאים תורמים לסיווג הנקודות למחלקה החיובית (המודעה אכן מזויפת), בעוד שמקדמים שליליים מצביעים על ההפך (מודעה אמיתית). כלומר, גודל המקדמים יכול לספק תובנות לגבי החשיבות היחסית של פיצ'ר לצורך הסיווג.

- בנספח 2 ניתן לראות את משוואת הישר עבור הדאטה סט שלנו המורכב מ-100 פיצ'רים עליו הפעלנו את מודל ה – SVM. מרבית הפיצ'רים עם המקדמים המשמעותיים שייכים לפיצ'ר המקורי company. נתון זה מתיישב עם המסקנות שקיבלנו בשלבים קודמים שמדובר ברכיב עיקרי ומהותי בסיווג. אולם, להפתעתנו הפיצ'רים שנמצאו קודם לכן כפיצ'רים בעלי החשיבות הגבוהה ביותר כמו qualification_entrylevel או company_get / include אינם מקבלים את המקדמים הגבוהים ביותר במודל זה. כלומר, על פי מודל זה הם אינם הפיצ'רים המשמעותיים ביותר לסיווג וזה בשונה מהמסקנות שהתקבלו קודם לכן.

Clustering

1. הסבר על סט הנתונים והשוני בין המודלים - אלגוריתם האשכול אמנם ירוץ על אותו סט נתונים כמו זה שהשתמשנו במודלים הקודמים, אולם בניגוד אליהם נסיר ממנו את הלייבלים של הסאמפלים. נסיר את עמודת משתנה המטרה fraudulent אשר מייצגת האם מודעה מסוימת הינה מזויפת או לא. הסיבה לכך היא שאישכול הינה טכניקת למידה בלתי מונחית, מה שאומר

שאין לנו שום לייבלים להנחות את האלגוריתם. נרצה להריץ את האלגוריתם על סט נתונים לא מתויג. בנוסף, מטרת משימת האשכול היא לקבץ מודעות דרושים דומות על סמך תכונותיהן, ולזהות דפוסים או חריגות כלשהן בנתונים – בשונה מהמודלים הקודמים שהם היו מסוג למידה מונחית, בהם מטרת משימת הלימוד הייתה לתת תחזית עבור מודעה אם היא מזויפת או לא.

2. <u>הרצת אלגוריתם K-Medoids</u>

2.1. סט הנתונים שלנו כרגע כולל 100 פיצ'רים שכולם נומריים לאחר העיבוד המקדים והנרמול שביצענו עליהם. בשלב של עצי ההחלטה ביצענו בדיקה של feature importance וראינו כי ישנם לא קצת פיצ'רים עם חשיבות מזערית. אי לכך ובהתאם לזאת, עבור שלב זה נבחר רק את הפיצ'רים שקיבלו חשיבות של 0.02 ומעלה ובכך כמות הפיצ'רים תצטמצם ל-15 בלבד. על הפיצ'רים האלו נבצע PCA כך שלבסוף יתקבלו 2 פיצ'רים החשובים ביותר.

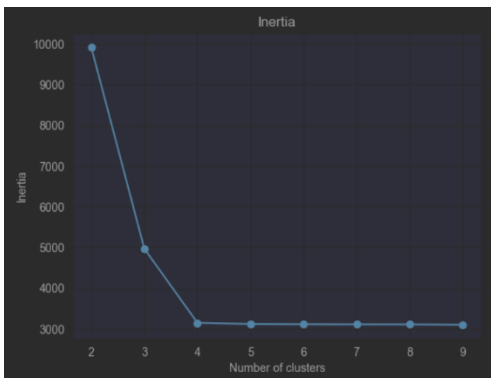
2.2. לאור העובדה כי ביצענו עיבוד מקדים על כלל הפיצ'רים המקוריים והם כבר לא בסוג הדאטה המקורי שלהם אלא כולם נומריים בסקאלה אחידה, נצטרך מטריקת מדידת מרחק שמתאימה לפיצ'רים כאלה. מטריקת ברירת המחדל של האלגוריתם הינה מרחק אוקלידי שזו מטריקה יחסית בסיסית שרגישה לחריגים. לאור הסיבות שצוינו לעיל, בחרנו להשתמש במטריקת מנהטן למדידת מרחק שמסוגלת להתמודד עימם בצורה טובה יותר ושאינה מושפעת מהיחידות והנרמול של כל פיצ'ר בנפרד.

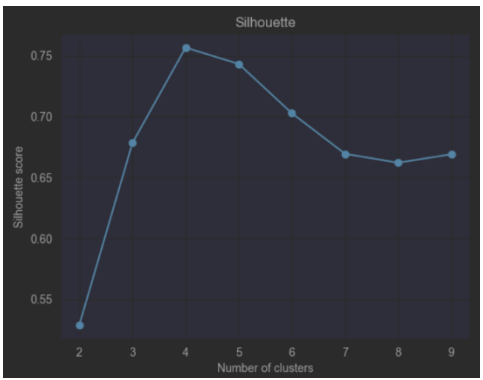
2.3. נצפה לראות שני קלאסטרים באופן דומה לחלוקה הטבעית של הדאטה המתויג, אחד עבור מודעות מזויפות והשני עבור מודעות אמיתיות.

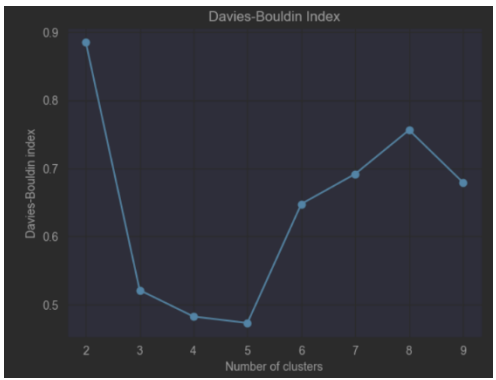
3. <u>השוואת תוצאות האשכול</u>

3.1. קריטריונים להשוואה:

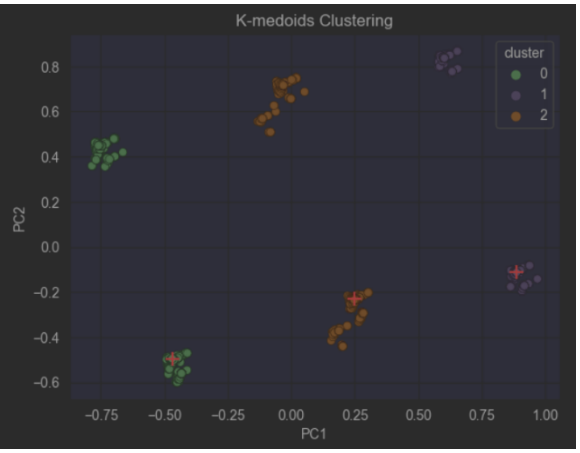
- אינרציה – מודד את המרחקים הריבועיים בין כל נק' בקלאסטר לבין נק' המרכז במטרה לצמצם מרחק זה. נחפש את נק' ה"מרפק" בגרף ובהתאם נבחר את מספר האשכולות.
- Silhouette – מודד את איכות ההפרדה בין קלאסטרים, בודק באיזה רמה סאמפל מתאים לקלאסטר שלו בהשוואה לקלאסטרים אחרים. על פי מדד זה נחפש בגרף את העלייה הגדולה ביותר ונבחר את כמות הקלאסטרים בהתאם.
- Davies Bouldin – מחשב את ההפרדה בין קלאסטרים ואת הקומפקטיות שלהם. על פי מדד זה נחפש בגרף את הירידה הגדולה ביותר ונבחר את כמות הקלאסטרים בהתאם.







3.2. מספר המחלקות שנבחר – מתרשים האינרציה ניכר כי "המרפק" נוצר כאשר מספר הקלאסטרים הינו שלושה בדיוק. מתרשים Silhouette ניתן לראות כי העלייה הגדולה ביותר מתרחשת בין 2 ל-3 קלאסטרים ומתרשים Davies Bouldin ניתן לראות כי הירידה הגדולה ביותר מתרחשת בין 2 ל-3 קלאסטרים. לפיכך, שלושת הגרפים מצביעים על כך שכמות הקלאסטרים שנרצה לבחור היא שלושה.



3.3. ייחוס משמעות לכל קלאסטר – באופן כללי ציפינו לקבל שני קלאסטרים, אולם מהתרשימים עולה כי כדאי לבחור ב-3 במקום. מתרשים הפיזור הנ"ל ניתן לראות כי ישנם 6 קלאסטרים ולא 3, כלומר יש איזושהי מורכבות פנימית בתוך כל קלאסטר שהאלגוריתם לא הצליח לזהות. לאור העובדה כי קיימים יותר משני קלאסטרים שמייצגים מודעות מזויפות או אמיתיות קשה לייחס משמעות לכל קלאסטר שהתקבל.

Evaluation

1. <u>השוואה בין המודלים</u>

מודל	דיוק אימון	דיוק ולידציה
DT	0.991	0.971
MLP	0.990	0.995
SVM	0.928	0.923

- נבחר במודל MLP שמקבל את אחוזי הדיוק הגבוהים ביותר על סט הוולידציה מבין שלושת המודלים שבחנו. מעבר לדיוק הגבוה יותר, מודל זה מאפשר לזהות דפוסים מורכבים בנתונים וזמן הריצה שלו היה דומה לזמן הריצה של אלגוריתם ה- DT.

Improvements

1. <u>מסקנות שעלו לגבי הנתונים או המודל:</u>

- כמות הפיצ'רים בעלי משמעות זניחה היא גבוהה, ראינו זאת כאשר ביצענו בדיקת feature importance בעצי ההחלטה וכאשר נעזרנו בתוצאות בדיקה זו לפני הזנת הדאטה סט למודל ה K-Medoids (התקבלו 15 פיצ'רים). ייתכן כי תהליך העיבוד שעשינו בחלק א' לא היה מספיק טוב. לכן, כדי לבחון שיפור בנתונים נוכל לבצע סינון של הפיצ'רים שקיימים בדאטה סט על פי הניקוד ב feature importance כמו שעשינו בשאלת הקלאסטרינג, או להשתמש באלגוריתם RFE שנעזרנו בו בחלק א' בשלב ה- Feature Selection על מנת לתת לו לבחור כמות מצומצמת של פיצ'רים לעומת ה-100 שיש כרגע בדאטה סט. שיפור זה יכול להאיץ את יכולת הריצה של המודל, ובהתאם נוכל לבצע כיוונון היפר-פרמטרים במודלים שונים בצורה מהירה יותר על ידי הורדת מימד זו בנתונים.

- מרבית הפיצ'רים בדאטה סט שלנו הם תוצר של עיבוד פיצ'רים טקסטואליים, ומיעוט הפיצ'רים הם נומריים, בינאריים או קטגוריאליים במקור. ייתכן כי פספסנו כאן מידע רב ערך לטובת משימת הלמידה. לכן, כדי לבחון שיפור בנתונים נוכל לבצע את תהליך העיבוד על הפיצ'רים מחדש עם מתן דגש נמוך יותר לפיצ'רים טקסטואליים, למשל על ידי חילוץ פחות פיצ'רים כאלו מכל פיצ'ר טקסט מקורי. אפשרות נוספת היא העמקה חוזרת בפיצ'רים שאינם טקסט, לנסות לזהות דפוסים שלא זיהינו בעבר ובהתאם לחלץ מהם פיצ'רים רלוונטיים יותר למשימת הלימוד. ייתכן שעיבוד חוזר כזה יכול להניב פיצ'רים אינפורמטיביים יותר מאלו הקיימים כרגע.

- בנוסף, על אף שקיבלנו אחוזי דיוק גבוהים במודלים שבדקנו, ייתכן כי ישנם אלגוריתמי למידה נוספים שיוכלו לספק לנו אחוזי דיוק אפילו גבוהים יותר על סט הוולידציה ובעתיד על סט המבחן. לכן, כדי לבחון שיפור במודל נרצה לבחון אלגוריתמים כאלו שלא למדנו עליהם וטרם נחשפנו להשפעה שלהם על אחוזי הדיוק.

2. <u>רעיונות לשיפור הביצועים:</u>

- שיפור בנתונים – רצינו להתמודד עם בעיית הפיצ'רים בעלי תרומה זניחה למודל שתוארה קודם לכן. השתמשנו בשיטת RFE שנעזרת באלגוריתם של רגרסיה לוגיסטית על מנת לבחור את 25 הפיצ'רים המשמעותיים ביותר בדאטה סט. בחרנו במספר 25 לאור בדיקת חשיבות הפיצ'רים שנעשתה בשלב הקלאסטרינג בה קיבלנו 15 פיצ'רים מעל רף מסוים.

- שיפור במודל – ניסינו למצוא מודל שיתאים יותר לדאטה סט שלנו וייתן תחזית מדויקת יותר. בחרנו לבחון מודל משולב המשתמש במכלול של מסווגים במקום מסווג בודד. החלטנו לשלב את כלל המסווגים שקיבלנו לאחר ביצוע כיוונון ההיפר-פרמטרים עבור כל אחד מהמודלים DT, MLP , SVM. כלומר, כדי לבנות את ה- ensemble בחרנו את המודלים הטובים ביותר שמצאנו מכל סוג ונשלב אותם. בחרנו בסכמה של הצבעה קלה מכיוון שבפרמטר זה נלקח בחשבון רמות האמון של התחזיות עבור כל אחד מהמסווגים.

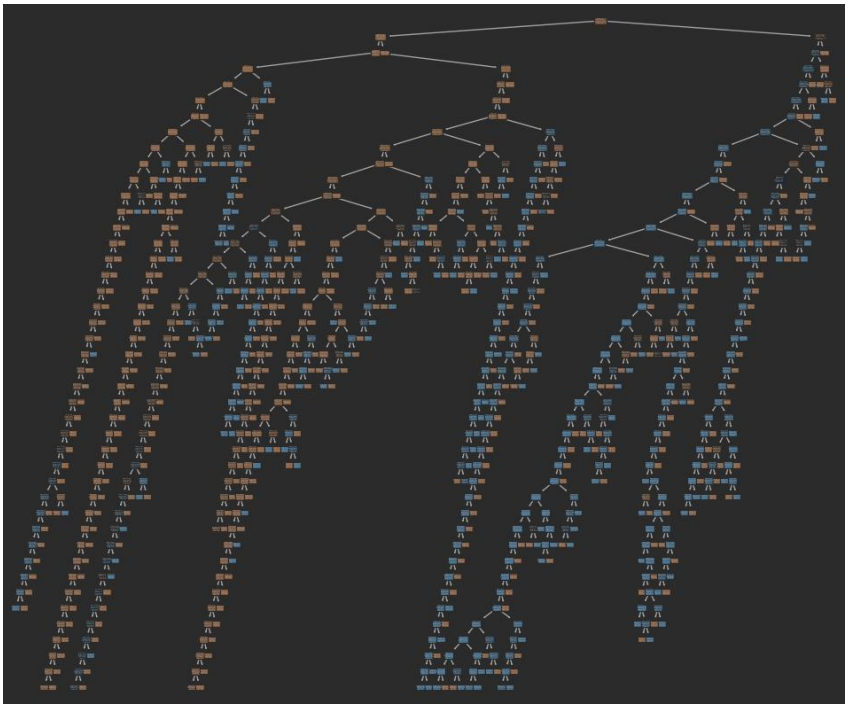
- המודל החדש לא שיפר את אחוזי הדיוק על סט הולידציה אלא פגע בהם, גם כאשר בדקנו על סט הנתונים המצומצם שבוצע עליו RFE והוא מכיל 25 פיצ'רים וגם על סט הנתונים המקורי שמכיל 100 פיצ'רים. קודם כל להערכתנו סט הנתונים המצומצם היה מצומצם מדי ולא כלל פיצ'רים אינפורמטיביים למשימת הלמידה. בנוסף ייתכן שהמודל החדש נתן תוצאות פחות טובות משום שלא ביצענו עליו כיוונון היפרפרמטרים, בנוסף אולי אחד מהמודלים הוא דומיננטי במיוחד לעומת האחרים בשקלול של המודל החדש מה שמוביל לשגיאות גבוהות יותר.

```
Ensemble Classifier ROC Accuracy Score - After RFE (25 Features): 0.876
```

```
Ensemble Classifier ROC Accuracy Score - Without RFE (100 Features): 0.958
```

נספחים

1. אלגוריתם עצי החלטה – העץ המלא:



2. SVM – משוואת "ישר" ההפרדה עבור דאטה סט עם 100 פיצ'רים:

$$
\begin{aligned}
y = {} & -0.06 * benefits_{grow} - 0.03 * benefits_{project} - 0.08 * company_{across} - 0.09 \\
& * company_{agenc} - 0.00 * company_{also} - 0.44 * company_{back} - 0.25 \\
& * company_{base} - 0.14 * company_{build} - 0.27 * company_{care} + 0.05 \\
& * company_{day} - 0.11 * company_{dedic} - 0.90 * company_{digit} + 0.29 \\
& * company_{drive} - 0.07 * company_{employ} - 0.51 * company_{empow} + 0.34 \\
& * company_{enabl} - 0.69 * company_{engag} + 0.02 * company_{enhanc} - 0.51 \\
& * company_{ensur} - 0.03 * company_{enterpris} - 0.57 * company_{environ} - 0.63 \\
& * company_{fast} - 0.32 * company_{find} - 0.20 * company_{found} - 0.33 \\
& * company_{get} - 0.95 * company_{great} - 1.25 * company_{high} - 0.81 \\
& * company_{highli} - 0.22 * company_{home} - 0.66 * company_{includ} + 0.42 \\
& * company_{increas} - 0.06 * company_{integr} - 0.31 * company_{interact} - 0.43 \\
& * company_{interest} - 0.74 * company_{larg} - 0.35 * company_{learn} - 0.19 \\
& * company_{life} - 1.23 * company_{like} - 0.34 * company_{long} - 0.26 \\
& * company_{manufactur} - 0.03 * company_{mobil} - 0.20 * company_{nation} \\
& - 0.32 * company_{number} - 0.30 * company_{open} - 0.10 * company_{oper} \\
& - 0.24 * company_{optim} + 0.04 * company_{passion} + 0.35 * company_{platform} \\
& - 0.13 * company_{posit} - 0.29 * company_{safe} - 0.15 * company_{satisfact} \\
& - 0.33 * company_{social} - 0.44 * company_{startup} - 0.09 * company_{take} \\
& - 0.04 * company_{talent} - 0.05 * company_{think} - 0.21 * company_{top} \\
& - 0.26 * company_{trust} - 0.41 * company_{understand} - 0.01 * company_{want} \\
& - 0.85 * company_{web} + 0.10 * country_{AU} - 0.20 * country_{GR} + 0.06 \\
& * country_{MY} + 0.24 * country_{US} - 0.13 * description_{app} - 0.02 \\
& * description_{care} - 0.07 * description_{client} - 0.06 * description_{creat} \\
& - 0.16 * description_{dd} - 0.08 * description_{exist} - 0.09 \\
& * description_{expand} - 0.12 * description_{first} - 0.18 * description_{fun} \\
& - 0.11 * description_{grow} - 0.03 * description_{growth} - 0.10 \\
& * description_{locat} - 0.05 * description_{love} - 0.09 * description_{passion} \\
& - 0.08 * description_{play} + 0.02 * description_{product} - 0.03 \\
& * description_{sale} + 0.17 * description_{url} - 0.10 * description_{user} - 0.06 \\
& * description_{web} + 0.15 * industry_{manufacturing} + 0.28 \\
& * qualification_{Entry\ level} - 0.22 * required_{experience_{Associate}} - 0.35 \\
& * required_{experience_{Mid\ Senior\ Level}} - 0.04 * requirements_{background} - 0.04 \\
& * requirements_{candi} - 0.16 * requirements_{face} - 0.13 * requirements_{lift} \\
& - 0.09 * requirements_{market} - 0.09 * requirements_{photoshop} - 0.32 \\
& * requirements_{php} - 0.07 * requirements_{sell} - 0.11 * requirements_{social} \\
& - 0.09 * requirements_{us} - 0.10 * title\_develop - 4.55
\end{aligned}
$$