**ML Project**

**Part B**

****

|  |  |  |
| --- | --- | --- |
| **Test** | **Train** |  |
| 0.2 | 0.199955 | **Fraudulent** |
| 0.8 | 0.800045 | **Real** |
| 3,340 | 13,393 | **Number of records** |

חלוקת הנתונים ל-Train ול-Test **-** בחרנו לפצל את סט הנתונים שלנו כך ש -20% מוקצים לטובת סט הבחינה ו – 80% לסט האימון החלוקה התבצעה בהתאם לנוהל הנהוג בתחום למידת המכונה. ניתן לראות שהנתונים בסט האימון ובסט הבחינה מאוזנים עבור ערך משתנה המטרה :

שיטת הולדיציה שבחרנו היא K-fold cross validation.

# Decision Trees

בנינו עץ החלטה מלא על פי הגדרות ברירת המחדל של חבילת DecisionTreeClassifier, קיבלנו עץ בעל 29 רמות ותוצאות הדיוק שהתקבלו עבור עץ זה: Train ROC\_AUC score = 1.0 , Test ROC\_AUC score = 0.8256

התוצאות מעידות על יכולת סיווג טוב של המודל ואינן מעידות על התאמת יתר.

בשלב הבא ניגשנו לכיוונון הפרמטרים – נסביר את משמעותו ואת טווח ערכיו של כל פרמטר בטבלה הבאה:

|  |  |  |
| --- | --- | --- |
| **היפרפרמטר** | **משמעות** | **טווח ערכים** |
| Max depth | היפרפרמטר זה מהווה חסם עליון לעומק העץ.  הגדרת עומק מקסימלי עוזרת לשלוט במורכבות העץ ולמנוע overfitting. | 29 - 1  כאשר 29 הוא עומק העץ המירבי שקיבלנו בהרצה דיפולטיבית . |
| Criterion | היפרפרמטר זה בוחר את הפונקציה שמשמשת את המודל למדידת איכות הפיצול בכל צומת בעץ. שני הקריטריונים המקובלים הם "gini" ו"entropy".  מדד ג'יני - בוחן את ההסתברות לסיווג שגוי של אלמנט שנבחר באקראי מצומת נתון.  מדד אנטרופיה – בוחן את מידת אי הוודאות שיש לנו לגבי התיוג בקבוצה. ערך המדד נע בין 0 ל-1 כך שצומת עם אנטרופיה של 0 הוא טהור (כל הנתונים בצומת מסווגים לאותה מחלקה).  שני הקריטריונים שואפים למצוא את הפיצול הבא הטוב ביותר שימקסם את ההומוגניות או יפחית את ה"טומאה" של משתנה המטרה בתוך כל ענף שנוצר של העץ. | טווח ערכים קטגוריאלי המכיל את שתי הפונקציות :  Gini , Entropy |
| Max features | היפרפרמטר זה קובע את המספר המרבי של תכונות (פיצ'רים) שיש לקחת בחשבון כאשר המודל בוחר את הפיצול הטוב ביותר בכל צומת. כיוונון של מספר הפיצ'רים עוזר לשלוט במורכבות המודל ויכול להפחית את הסיכון להתאמת יתר. | 3 אופציות אפשריות למספר הפיצ'רים בהם נתחשב בכל איטרציה :  None – נתחשב בכלל הפיצ'רים.  Log2 – מספר הפיצ'רים בו נתחשב הוא log2 של מספר הפיצ'רים הכולל.  Sqrt – מספר הפיצ'רים בו נתחשב הוא השורש הריבועי של מספר הפיצ'רים הכולל. |

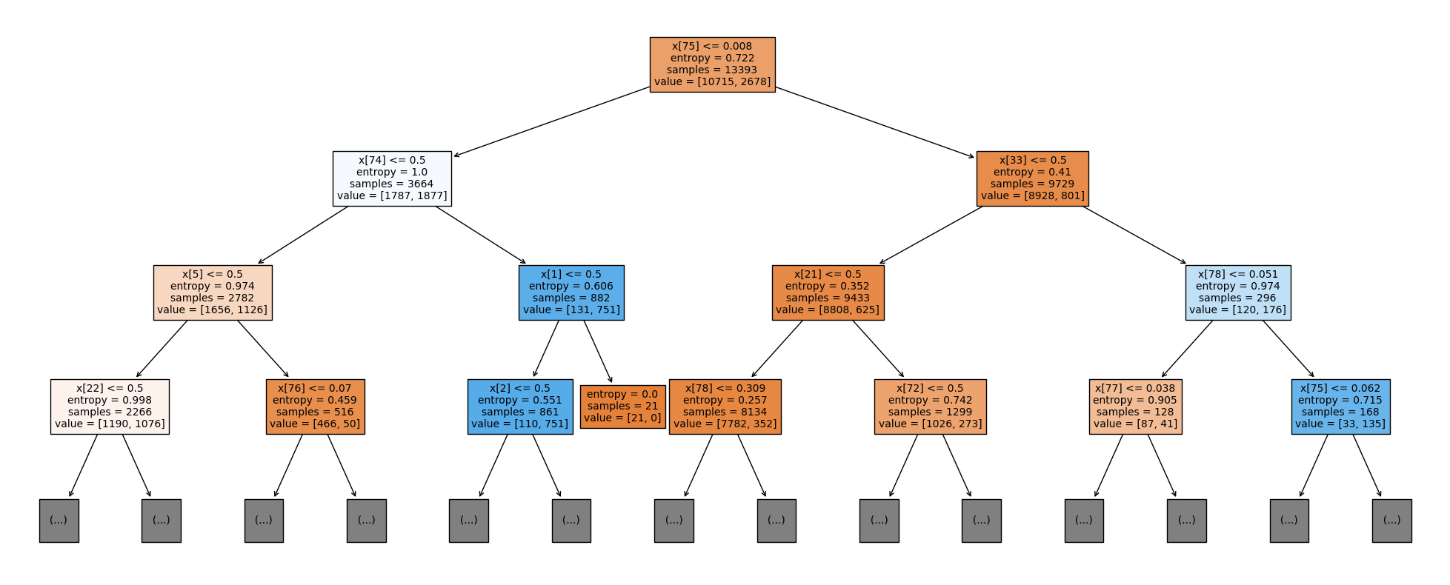
הקונפיגרציה המיטבית שנבחרה לאחר הרצת grid search :

Max depth = 20 , Criterion = entropy , Max features = None

**אחוזי הדיוק שהתקבלו עבור המודל הטוב ביותר :** עבור סט האימון : 0.9995 ,עבור סט הולדיציה : 0.8622

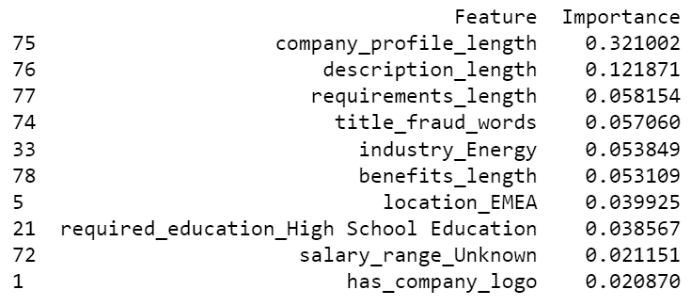
התוצאה הגבוהה על סט האימון מעידה על כך שמודל העץ מתאים ברמה טובה לנתונים שלנו באופן כזה כך שהמודל יכול לסווג ביעילות בין שתי המחלקות, כמובן שהתוצאה על סט האימון לא מבטיחה הצלחה עבור נתונים שהמודל לא ראה קודם. התוצאה שהתקבלה עבור סט הולדיציה מעידה על כך שהמודל מתמודד בצורה טובה יחסית עם נתונים שלא ראה קודם לכן נסיק שהוא בעל יכולת הכללה טובה ומצליח לסווג נכונה מודעות חדשות ברוב המקרים.

**יכולת ההסברה של מודל עץ החלטה ותרומתה למשימת הלמידה -** יכולת ההסברה של המודל מתייחסת לקלות שבה בני אדם יכולים להבין ולפרש את תהליך קבלת ההחלטות של המודל. עצי החלטה מייצגים באופן ברור את כללי ההחלטה וניתן בקלות להתחקות אחר תהליך ההחלטה שביצע המודל. יכולת ההסברה של המודל מאפשרת לנו להבין את כללי ההחלטה ולזהות את הפיצ'רים החשובים ביותר שמבחינים בין מודעות אמיתיות למזויפות. ההבחנה הזו תאפשר לנו לשפר את יכולות הסיווג של המודל ואף לזהות תכונות נוספות שנרצה לקחת בחשבון.

**גרף העץ שהתקבל :**

**התבוננות במבנה העץ יכולה לספק תובנות על הבעיה ועל חשיבות המאפיינים השונים :**

* בחינת עומק העץ - התבוננות בעומק העץ יכולה ללמד אותנו על מורכבות הבעיה, ככל שהעץ יותר עמוק כך נדרשים יותר כללי החלטה כדי לסווג מודעה כמזויפת או אמיתית. במקרה שלנו בעץ האופטימלי יש 20 רמות, העץ יחסית עמוק וניתן לומר שבהינתן הפיצ'רים הקיימים משימת הסיווג מורכבת.
* צמתי ההחלטה - גרף העץ מציג בכל צומת פנימי קריטריון פיצול, על ידי בחינת הפיצולים הללו אפשר לזהות את המאפיינים החשובים ביותר בהבחנה בין מודעות אמיתיות ומזויפות. הפיצול הראשון במקרה שלנו הוא על פי אורך פרופיל החברה מה שמצביע על כך שזהו גורם מכריע בקביעת האותנטיות של מודעה.

**חשיבות המשתנים -** נציג את עשרת המשתנים המשמעותיים ביותר בטבלת חשיבות המשתנים שהתקבלה משימוש בתכונת feature\_importances של המודל :

**מסקנות מטבלת חשיבות המשתנים בנוגע למשימת הלימוד -** בחינת ציוני החשיבות של הפיצ'רים מאפשרת לזהות את התכונות בעלות ההשפעה הגדולה ביותר על משימת הסיווג – במקרה שלנו אנחנו מזהים כי אורך פרופיל החברה, אורך התיאור של המשרה ואורך רשימת הדרישות של המשרה הם שלושת הפיצ'רים המשמעותיים ביותר שמבחינים באופן המשמעותי ביותר בין מודעה מזויפת לאמיתית.ניתן לראות שמלבד שני הפיצ'רים הראשונים בטבלה חשיבות הפיצ'רים יחסית נמוכה מה שמרמז על שני מצבים אפשריים - הפיצ'רים שחילצנו לא משפיעים באופן משמעותי על משימת הסיווג ועלינו לשים לב ליכולות החיזוי של המודל או לחילופין קיימים קשרים בין הפיצ'רים שמייתרים את חלקם.

# Artificial Neural Networks

הרצנו ואימנו את מודל ה - MLPClassifier בערכי ברירת המחדל שלו, נסביר את משמעות הקונפיגורציה שלו:

* **מספר הנוירונים בשכבת הכניסה** - מספר הפיצ'רים הכולל במודל, במקרה שלנו 78 פיצ'רים.
* **מספר השכבות החבויות** – במודל MLP המידע זורם משכבת הקלט לשכבת הפלט דרך שכבות חבויות ועלינו להחליט מה מספר השכבות האידיאלי בהן יעבור ויעובד המידע. הערך הדיפולטיבי הוא 1, כלומר שכבה אחת בלבד בין שכבת הקלט לפלט.
* **מספר נוירונים חבויים בכל שכבה** – כל שכבה חבויה מכילה סט נוירונים שמבצעים חישובים על הדאטה שהם מקבלים מהשכבה הקודמת ומעבירים לשכבה הבאה. הערך הדיפולטיבי הוא 100 נוירונים (קודקודים) בשכבה.
* **מספר הנוירונים בשכבת הפלט** - מספר המחלקות במודל, במקרה שלנו 2 מחלקות אליהן המודל מסווג (1 ו – 0 סיווג למודעה מזויפת או אמיתית).

**אחוזי הדיוק שהתקבלו עבור המודל הדיפולטיבי - עבור סט האימון:** 0.9992 , **עבור סט הולידציה:** 0.9350

התוצאה הגבוהה על סט האימון מעידה על כך שהמודל מתאים ברמה טובה לנתונים ויכול לסווג ביעילות בין שתי המחלקות. התוצאה שהתקבלה עבור סט הולדיציה מעידה על כך שהמודל מתמודד בצורה טובה מאוד עם נתונים חדשים לכן נוכל לומר שהמודל הדיפולטיבי בעל יכולת הכללה טובה ומצליח לסווג נכונה מודעות חדשות ברוב המקרים.

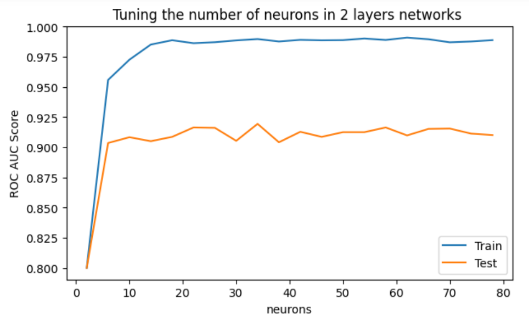
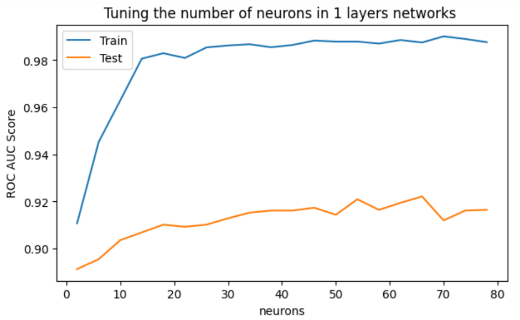
**MLP Hyperparameter Tuning**

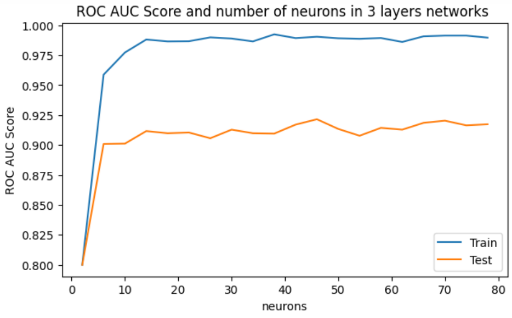
|  |  |  |  |
| --- | --- | --- | --- |
| **היפרפרמטר** | **מוטבציה לכוונון** | **משמעות של הקטנה/הגדלה** | **טווח ערכים ומדוע** |
| Number of hidden layers | מספר השכבות החבויות משפיע על יכולת המודל ללמוד ייצוגים מורכבים של הנתונים. מטרת הכוונון של מספר השכבות החבויות היא למצוא את המספר האידיאלי של שכבות בו קיים איזון בין מורכבות המודל והיכולת שלו לתפוס דפוסים מורכבים לבין שמירה על ביצועי ההכללה טובים. | הגדלת מספר השכבות החבויות מגדילה את היכולת של המודל ללמוד תכונות מורכבות יותר ולזהות יחסים מורכבים בין פיצ'רים. עם זאת, כמות גדולה מידי של שכבות נסתרות עלולה להביא למצב של התאמת יתר. הקטנה של מספר השכבות החבויות הופכת את המודל לפשוט יותר אך עלולה להגביל את יכולות הלמידה שלו. | טווח הערכים שבחרנו לכיוונון הוא : 1-3 שכבות.  מקובל לבחון מספר שכבות שנע בין 1 ל- 5 כדי להמנע  מ – overfitting  התחלנו מהמספר הקטן ביותר ועלינו תוך מעקב אחרי הביצועים של המודל. |
| Hidden layer sizes | גודל כל שכבה חבויה משפיע על יכולות הלמידה של המודל. המטרה בכיוונון היפרפרמטר זה היא למצוא למצוא את רמת הגרעיניות הנחוצה בשכבות החבויות כך שחילוץ התכונות והדפוסים יהיה אופטימלי בהתחשב במורכבות הבעיה ובנתונים הזמינים. | הגדלת מספר הנוירונים בשכבות החבויות מאפשרת למודל ללמוד דפוסים מורכבים יותר וללכוד פרטים עדינים יותר בנתונים, יחד עם זאת נזכור כי הגדלה של מספר הנוירונים עלולה להוביל להתאמת יתר. הקטנת מספר הנוירונים בשכבות החבויות תפשט את המודל אך עלולה להגביל את יכולות הלמידה שלו. | טווח הערכים שבחרנו לכיוונון הוא : 2-78 נוירונים בשכבה.  בחרנו את טווח הערכים על פי כלל אצבע מקובל שמצאנו באינטרנט שאומר שמספר הנוירונים בשכבות החבויות צריך לנוע בין מספר המחלקות למספר הפיצ'רים במודל. |
| Learning rate | קצב הלמידה שולט בגודל הצעד על פיו המודל משנה את משקלי הקשתות במהלך אימון המודל.  כוונון קצב הלמידה נועד למציאת האיזון האופטימלי בין מהירות ההתכנסות לבין יציבות המודל במהלך תהליך האימון. | קצב למידה גבוה יותר מאפשר למודל להתכנס מהר יותר במהלך האימון, אך הוא גם מגביר את הסיכון לחריגה מהפתרון האופטימלי. לעומת זאת, קצב למידה נמוך יותר יכול לשפר את יציבות ודיוק האימון אך הוא גורם להתכנסות איטית ועלול לגרום לכך שהמודל יתקע בפתרונות לא אופטימליים. | טווח הערכים שבחרנו לכיוונון הוא מערך קצבי הלמידה הבאים:  [0.1,0.01,0.001,0.0001]  נתחיל מקצב הלמידה הגבוה ביותר ונרד בסולם הערכים.  בחרנו בקצבי הלמידה האלה לאחר שמצאנו כי מומלץ לבחון את קצבי הלמידה בסולם לוגריתמי. |
| Activation function | לבחירת פונקציית אקטיבציה יש תפקיד מכריע בביצועים וביכולות של מודל MLP. לפונקציות אקטיבזציה שונות יש מאפיינים שונים והן מתאימות לסוגים שונים של בעיות. על ידי כוונון פונקציית האקטיבציה נוכל לשפר את יכולת הלמידה של המודל ולהתאים אותו טוב יותר לבעיה ולנתונים שלנו. | אין משמעות להקטנה או הגדלה של פונקציית אקטיבציה. לכל אחת מפונקציות האקטיבציה יש נוסחה מתמטית והתנהגות שונה. | טווח הערכים שבחרנו לכיוונון הוא הפונקציות הבאות :  [logistic, tanh, relu]  וזאת לאחר שבדקנו את ביצועי המודל עבור סט האימון והוולידציה עבור הפונקציות השונות. |

נציג גרפים שמציגים את ערכי ההיפר-פרמטרים שנבחנו כפונקציה של אחוז הדיוק על סט הנתונים והולידציה :

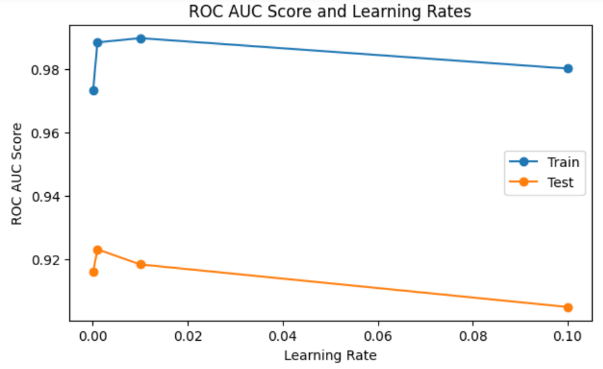
**מספר שכבות חבויות ומספר נוירונים בכל שכבה :**

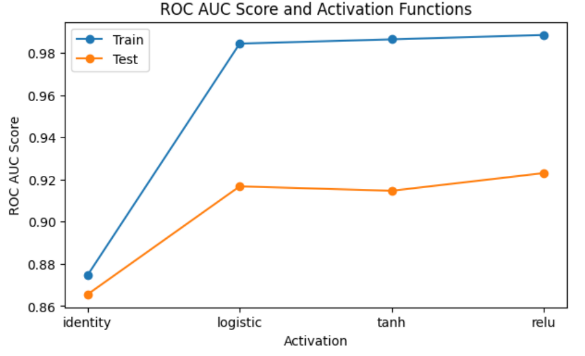
עבור שכבה אחת : עבור שתי שכבות :



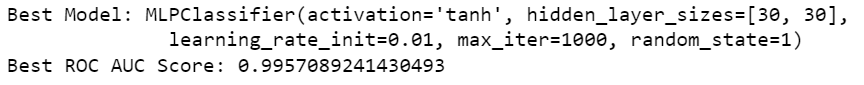
עבור שלוש שכבות חבויות :

מהגרפים ניתן לראות כי אחוזי הדיוק עבור סט האימון דומים עבור כל הקומבינציות של מספר שכבות ומספר נוירונים. עבור סט הוולידציה יש הבדל משמעותי בין שכבה חבויה אחת לבין הקומבינציות של מספר נוירונים עם 2 ו3 שכבות חבויות. נוכל להסיק כי סביר שהמודל האופטימלי ייבחר ב-2 או 3 שכבות חבויות ובנוסף נשים לב כי ההתייצבות על אחוזי הדיוק המקסימליים מופיעה באזור של 30 נוירונים בשכבה.

**קצב למידה -** מהגרף שבוחן את קצב הלמידה כפונקציה של אחוז הדיוק ניתן לראות שאחוז הדיוק המירבי על סט הולידציה מתקבל בקרבת האפס לכן הגיוני וסביר שמבין סט הערכים שבדקנו קצב הלמידה האופטימלי יהיה 0.001.

**פונקצית אקטיבציה -** מהגרף ניתן לראות כי עבור הפונקציה Identity אחוזי הדיוק לא טובים הן עבור סט האימון והן עבור סט הוולידציה. עבור שלושת הפונקציות האחרות נזהה אחוזי דיוק דומים עבור סט האימון וסט הוולידציה כאשר המובילה מביניהן על פי הגרף היא relu. נוכל לבחון רק את שלושת הפונקציות בכיוונון הפרמטרים וסביר שנקבל את אחת מהן כאופטימלית.

לאחר בחינת ההיפרפרמטרים השונים אל מול אחוזי הדיוק צמצמנו את מרחב האפשרויות שלנו כדי למנוע זמני ריצה ארוכים והכנסנו את מרחב הערכים שנותר לאלגוריתם Grid search.

המודל הטוב ביותר שהתקבל הוא :

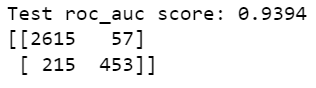
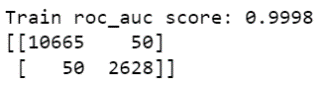
**אחוזי הדיוק המתקבלים על סט האימון וסט הוולידציה, עבור המודל הטוב ביותר הם:**

עבור סט האימון : 0.9998 , עבור סט הולידציה : 0.9394

**מסקנות מאחוזי הדיוק של המודל -** המודל למד בצורה טובה מאוד את נתוני האימון וכן הצליח לסווג בצורה טובה את נתוני הוולידציה.ניכר כי המודל מתמודד בצורה טובה עם נתונים שלא ראה קודם לכן וכי לא מדובר במצב של התאמת יתר לסט האימון.

**ההבדל בין תוצאות המודל הדיפולטיבי למודל הנבחר -** תוצאות הדיוק של המודל הנבחר חושבו בתהליך של קרוס ולידציה בשיטת K-FOLD. תהליך הקרוס ולידציה מחשב את ממוצע תוצאות הדיוק על פני מספר איטרציות אימון של המודל מה שגרם לכך שתוצאות הדיוק עבור המודל הנבחר נמוכות מהתוצאות שקיבלנו עבור המודל הדיפולטיבי (עבורו לא ביצענו תהליך של קרוס ולידציה). קוניפוגרציית ההיפרפרמטרים של המודל הנבחר שונה מזו של המודל הדיפולטיבי והיא מהווה קונפיגורציה מיטבית שמטרתה לשפר את תוצאות הדיוק של המודל.

אימנו רשת עבור הקונפיגורציה שנבחרה באמצעות אלגוריתם Grid search וקיבלנו את מטריצות המבוכה הבאות :

******עבור סט האימון:** **עבור סט הוולידציה:**

נבחן את מטריצת המבוכה שהתקבלה עבור סט הוולידציה כדי להסיק על התנהגות המודל הטוב ביותר :

ערכי ה TP וה – TN במטריצת המבוכה גבוהים, מציגים רוב של סיווגים נכונים של המודל ומעידים על כך שביצועי המודל טובים. באמצעות נתוני המטריצה נוכל לחלץ מדדים שונים להערכת ביצועי המודל :

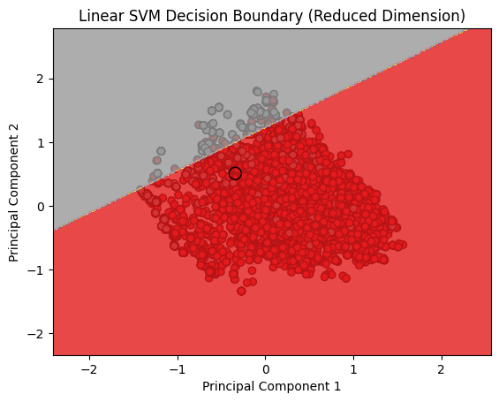
על פי מדד הדיוק נסיק כי המודל מצליח לסווג נכונה מודעות ב – 91% מהמקרים.

ה- Precision מעיד שב98% מהמקרים כאשר המודל אומר שמודעה היא אמיתית הוא אכן צודק.

מדד ה – Recall מעיד שב -92% מהמקרים המודל מזהה את המודעות האמיתיות .

# SVM

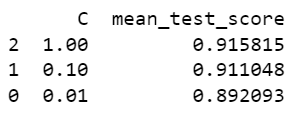
לשם אימון מודל SVM ניסינו תחילה לבצע הורדת מימד באמצעות PCA.

התקבל מישור ההפרדה הבא:

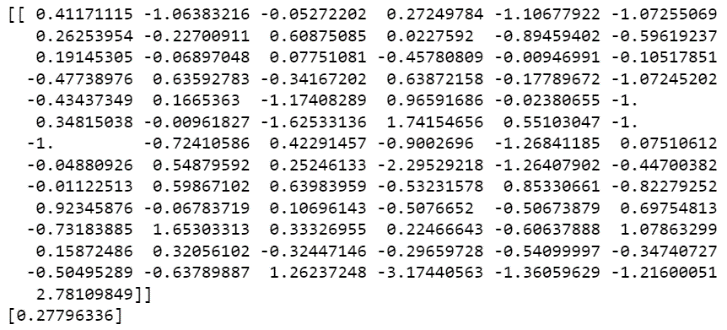
תוצאות הדיוק עבור המודל לאחר הורדת המימד היו 57.7% על סט הבחינה לכן פנינו לאימון מודל ללא הורדת מימד.

אימנו מודל SVM תוך ביצוע Hyperparameter Tuning באמצעות אלגוריתם Grid search למציאת הקונפיגורציה המיטבית (השתמשנו רק בKERNEL לינארי).

אחוזי הדיוק של המודל הנבחר הם : **עבור סט האימון:** 0.918 **עבור סט הולידציה:** 0.893

מצורפת טבלה שמסכמת את תוצאות תהליך הכיוונון כאשר כיוונו את הפרמטר C בלבד:

כפי שניתן לראות הפרמטר הנבחר הוא C =1.

**מצורפים החותך וטבלת המקדמים של משוואת המישור המפריד:**

מקדמי המשתנים במשוואת הישר מייצגים את המשקל של כל פיצ'ר כלומר , המקדמים קובעים את חשיבות כל אחת מהתכונות בגבול ההחלטה שמפריד בין המחלקות השונות.

על פי טבלת המקדמים הפיצ'ר בעל המקדם המשמעותי ביותר הוא :

company\_profile\_length

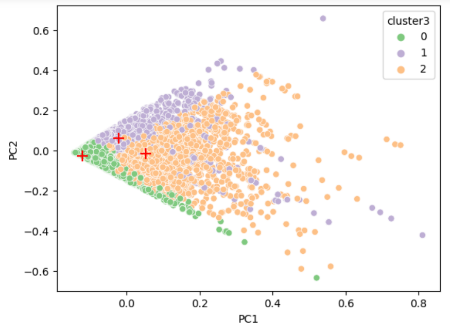
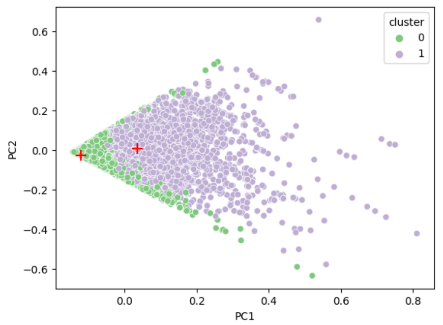
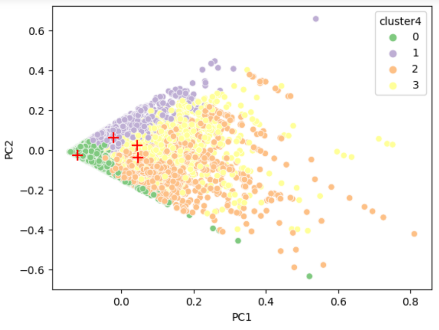
כאשר בחנו את חשיבות הפיצ'רים באמצעות פונקציית החשיבות של מודל עץ ההחלטה מצאנו כי זהו אכן הפיצ'ר המשמעותי ביותר, כלומר המשמעות של המקדמים תואמת את המסקנות שלנו מעץ ההחלטה לגבי חשיבות הפיצ'רים ותרומתם למודל.

# **Unsupervised Learning - Clustering**

**ההבדל בין אלגוריתם האשכול למודלים הקודמים -** אלגוריתמי אישכול משתייכים לקבוצת האלגוריתמים של למידה לא מונחית. בלמידה לא מונחית האלגוריתם מנסה ללמוד דפוסים ולהגדיר קבוצות נתונים מבלי שעומדים לרשותו הלייבלים של הדאטה. המודלים הקודמים שהרצנו משתייכים לקבוצת הלמידה המונחית בה האלגוריתמים משתמשים בתגיות לשם אימון המודל. עבור אלגוריתם האישכול נשתמש באותו סט נתונים שמכיל את כל הפיצ'רים שחילצנו אך ללא תגיות הסיווג שמציינות האם מודעה היא אמיתית או מזוייפת. המטרה של משימת הלימוד באלגוריתמים של למידה מונחית היא לאמן את המודל כך שיוכל לסווג באופן המדויק ביותר מודעות חדשות כמזויפות או אמיתיות בעוד שהמטרה באלגוריתם האישכול היא לבחון את מבנה הנתונים ולזהות קבוצות דומות של מודעת דרושים על סמך התכונות שלהן.

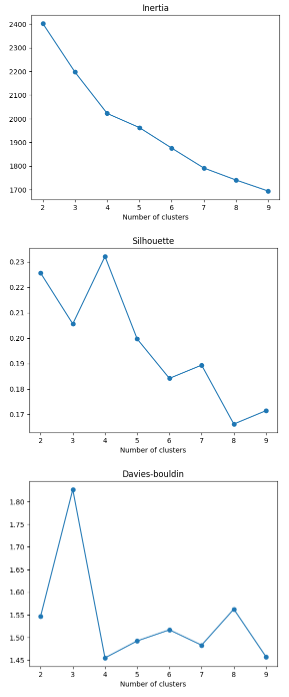
**הרצת אלגוריתם K-medoids - סט הפיצ'רים שבחרנו, מטריקת מדידת המרחק וקלאסטרים צפויים:**

לשם ההרצה של האלגוריתם K-medoids בחרנו את ארבעת הפיצ'רים שמייצגים את אורך השדות השונים במודעה (אורך פרופיל החברה, אורך תיאור המשרה, אורך הדרישות ואורך ההטבות) וביצענו הורדת מימד באמצעות PCA ל – 2. כל הפיצ'רים האלה הם נומריים ומנורמלים בין 0 ל-1 לכן נוכל לבחור במרחק מנהטן להיות מטריקת מדידת המרחק שלנו. מרחק מנהטן הוא מדד מרחק שמחשב את סכום ההפרשים בערך מוחלט בין ערכי הפיצ'רים של שתי נקודות נתונים ובאמצעותו ניתן לכמת את השונות או הדמיון בין נקודות הנתונים. בהתבסס על הפיצ'רים שבחרנו ובהתחשב במטריקת מדידת המרחק נצפה לראות קלסטרים של מודעות דרושים שחולקות דמיון מבחינת אורכי התיאורים במודעה. הרצנו את האלגוריתם עבור K = [2,3,4] מצורפים הגרפים המציגים את הסיווגים עבור כל אחד מערכי ה - K :

****

**השוואת תוצאות האשכול עבור ערכי K שונים :**

**קריטריונים להשוואה:**

* **אינרציה – Inertia** – מדד האינרציה מחשב את סכום המרחקים בריבוע של כל סאמפל למרכז האשכול הקרוב ביותר אליו. סכום ריבועים נמוך יותר יעיד על אשכול הדוק ומרוכז יותר.
* **מדד סילואט – Silhouette** – מודד את רמת הלכידות וההפרדה של האשכולות כלומר מציין עד כמה דומה נתון מסוים לנתונים האחרים באשכול אליו הוא משויך (לכידות) ועד כמה הוא שונה מהנתונים באשכולות הסמוכים (הפרדה). המדד נע בין -1 ל- 1 כאשר ערכים גבוהים מצביעים על אשכולות מוגדרים ומופרדים היטב וערכים נמוכים מעידים על אשכול לקוי.
* **מדד דייויס-בולדין – Davies-Bouldin** - מודד את הדימיון הממוצע בין האשכולות. הוא לוקח בחשבון גם את הלכידות של אשכול מסוים וגם את השוני וההפרדה בין האשכולות. המדד שואף למצוא איזון בין הגודל של קלאסטר מסויים לבין המרחק שלו מקלאסטרים אחרים. ציון נמוך אומר שהאשכול קטן יחסית למרחק שלו מאשכול אחר, ולכן ייחשב כמוגדר היטב.

**מספר מחלקות נבחר -** על מנת לבחור את מספר המחלקות האופטימלי נבחן את ערכי שלושת המדדים שבחרנו - אינרציה, סילואט ודיויס בולדין. עבור כל מדד הפקנו גרף שמציג את ערך המדד כפונקציה של מספר האשכולות, ניעזר בגרפים לבחירת מספר האשכולות האופטימלי על פי הכללים הבאים שהוצגו במעבדה 3: עבור קריטריון אינרציה ועל פי כלל המרפק נבחר את ערך ה-K שנמצא בנקודת ה"מרפק" של הגרף. עבור קריטריון סילוואט נבחר את ערך K שבו מוצגת הירידה המשמעותית ביותר שיש בערך המדד. ועבור מדד דיוויס בולדין נרצה לבחור את מספר האשכולות בו יש את העלייה המשמעותית ביותר בערך המדד.

מדד אינרציה – קשה לזהות בגרף מרפק מובהק ניתן לומר

שהוא מופיע עבור 4 אשכולות.

עבור מדד סילואט ניתן לראות ירידה משמעותית עבור 3 ו-5 אשכולות.

עבור מדד דיויס בולדין ישנה עלייה משמעותית ב3 קלאסטרים.

בהסתכלות כוללת על שלושת המדדים נבחר 3 מחלקות כמספר המחלקות האידיאלי.

**משמעות הקלאסטרים :**

כיוון שהשתמשנו בארבעה פיצ'רים המייצגים את אורך הטקסט בתיאורים שונים במודעה סביר כי כל אחד מארבעת הקלאסטרים מייצג קבוצת מודעות בעלת אורכים דומים של התיאורים השונים.

# השוואה בין המודלים - Evaluation

|  |  |  |
| --- | --- | --- |
| **ROC AUC Score - Training** | **ROC AUC Score - Validation** | **Model** |
| 0.9995 | 0.8622 | Decision Tree |
| 0.9998 | 0.9394 | MLP |
| 0.918 | 0.893 | SVM |

בהתבסס על התוצאות שקיבלנו נבחר במודל MLP.

מודל ה MLP מציג את הביצועים הטובים ביותר עבור סט הוולידציה מה שמצביע על כך שהוא בעל היכולות הטובות ביותר לסווג מודעות דרושים כאמיתיות או מזויפות.

# שיפור המודל הנבחר – Improvements

**נקודות עיקריות שנרצה לשפר :**

* Feature Engineering – בתהליך חילוץ הפיצ'רים שביצענו לא השתמשנו בטכניקות ואלגוריתמים מקובלים, ייתכן שפיספסנו קשרים או תכונות חשובות שהיו יכולים לתרום למשימת הסיווג שלנו. כעת נרצה לבחון האם קיימות תכונות רלוונטיות נוספות שנוכל לחלץ מהפיצ'רים הקיימים כך שיכולת הסיווג של המודל תשתפר.
* Model architecture – על אף שביצענו תהליך של כיוונון פרמטרים עבור כל אחד מהמודלים עדיין יש חשש כי המודל שבחרנו לא קיבל את הקונפיגורציה המיטבית עבורו שהרי לא כיוונו את כל ההיפרפרמטרים של המודל ובדקנו כמות ערכים מוגבלת כדי להמנע מזמני ריצה אסטרונומיים. כדי להעלות את רמת הביטחון במודל שלנו ולשפר את הביצועים שלו נשתמש בשיטות אנסמבל.

**שיפור בנתונים - Polynomial Features** השיפור שביצענו על הנתונים היה הוספת פיצ'רים פולינומיאליים. זוהי טכניקה שיוצרת תכונות חדשות משילובים פולינומיים של הפיצ'רים המקוריים. הטכניקה הזו מאפשרת למודל זהות קשרים לא לינאריים. לשם ביצוע השיפור השתמשנו בפונקציה PolynomialFeatures והוספנו פיצ'רים פולינומיאליים עד דרגה 2, את המודל הטוב ביותר שלנו אימנו כעת על סט הפיצ'רים החדש ולא קיבלנו שיפור בתוצאת הROC AUC על סט הוולידציה. ייתכן והוספת הפיצ'רים הפולינומיאליים יצרה פיצ'רים מיותרים וגרמה למודל להיות יותר מורכב ופחות מתאים למשימת הסיווג .

**שיפור במודל – Bagging Classifier -** לשם שיפור המודל השתמשנו בשיטת אנסמבל שנקראת bagging. זו שיטה שמפרקת את סט האימון לתתי קבוצות באופן אקראי, מאמנת את המודל הבסיסי שהגדרנו על כל תת קבוצה ולבסוף מתכללת את תוצאות הסיווג של כל מודל לכדי סיווג סופי.השימוש ב – bagging מפחית את הסיכוי להתאמת יתר ואמור לשפר את דיוק המודל. לאחר יצירת מופע של BaggingClassifier ואימון המודל על הפיצ'רים המורחבים נתוני הדיוק חזרו להיות כפי שהיו קודם.

**תוצאות הדיוק של המודל לאחר השיפורים : **

# הגשת חיזויים סופיים

על קובץ הנתונים X\_Test בצענו עיבוד מקדים באופן זהה לזה שביצענו קודם (ללא הסרת כפילויות) ואז ביצענו חיזוי באמצעות המודל הסופי שבחרנו לאחר הוספת השיפורים.