המחלקה להנדסת תעשייה
אוניברסיטת תל אביב

דו"ח מסכם - פרויקט למידת מכונה

קבוצה 05

מרצה

ד"ר אייל קולמן

מתרגלת

גב. שי עובד

<u>מגישים</u>:

אביהו מנחם

טל אילון

עמיחי כלב

תקציר מנהלים

בפרויקט זה נעסוק בבעיית סיווג בינארית, בה עלינו לסווג רשומות לשתי קטגוריות על סמך מספר פיצ'רים בסט הנתונים. העבודה התחלקה לארבעה חלקים עיקריים.

בחלק הראשון, בוצע חקר מעמיק של סט נתוני האימון וסט נתוני ה test לחיזוי. נבחנו התנהגויות הפיצ'רים, התפלגותם, סוגם, כמות הערכים החסרים בהם, שכיחות הערכים בהם והקשר ביניהם לבין הפיצ'רים האחרים. לאחר ניתוח הפיצ'רים באופן כולל ואינדיבידואלי הוסקו מסקנות רלוונטיות.

בחלק השני, שלב העיבוד המקדים, יושמו המסקנות והעיבוד הותאם לכל פיצ'ר אינדיבידואלית. בפרט, נבחן כיצד להשלים את הערכים החסרים במידה וקיימים וכן הוצאו ערכים חריגים. כמו כן, נבחנה האפשרות להוסיף פיצ'רים חדשים באמצעות פעולות מתמטיות וכן נבחן האם הנתונים מנורמלים, ובהתאם לכך נבחר האם לנרמל אותם או לא. לבסוף, נבחנה הממדיות של הבעיה ובוצעה הקטנת ממדים בשיטת PCA.

בחלק השלישי, בוצעה בניית מודלים והרצתם. נבחנו 2 מודלים ראשוניים ו2 מודלים מתקדמים. נבחנו מספר אפשרויות להיפר פרמטרים ונבחר סט היפר פרמטרים מיטבי עבור כל מודל.

בחלק הרביעי, בוצעה השוואת המודלים. ההשוואה בוצעה ע"י הרצת 4 המודלים על סט הנתונים לאחר שעבר את העיבוד המקדים, ובאמצעות שיטת K-Fold Cross Validation חולק סט נתוני האימון ל5 מחיצות עבור כל מודל, כאשר כל מחיצה חולקה לTrain ול Validation. לכל מודל התקבלו מדדי דיוק ומדדי AUC על הValidation ועל ה Train, ונבחן הפער בין מדדי הAUC הללו לטובת מציאת overfitting אפשרי. באמצעות מטריקת AUC נבחר המודל בעל הביצועים הגבוהים ביותר.

נמצא כי המודל Multi-Layer Perceptron (ANN) הינו בעל הביצועים הגבוהים ביותר, ולכן הוא נבחר להיות זה שסביר להניח ינבא בצורה הטובה ביותר את הנתונים.

<u>חלק ראשון - אקספלורציה</u>

בחלק זה נחקר סט הנתונים לאימון המודלים וכן סט נתוני ה test. סט הנתונים לאימון כולל 22,161 רשומות ו26 פיצ'רים, מתוכם 19 פיצ'רים נומריים, 6 פיצ'רים קטגוריאליים ופיצ'ר נוסף שהוא עמודת ה label. סט נתוני ה test כולל 7387 רשומות.

ראשית, נבחנו כל הפיצ'רים באופן ראשוני באמצעות פונקציית describe השייכת לספריית pandas במטרה לבחון את סוגי הטיפוסים – int/float/object. בנוסף הוצג:

- **גרף היסטוגרמה** עבור הפיצ'רים הנומריים
- **מטריצת פיזור** תוך התייחסות לפיצ'ר עצמו והשוואתו אל מול הפיצ'רים האחרים
- טבלה המציגה את מספר **הערכים החסרים בכל פיצ'ר**

מהתבוננות בממצאים הראשוניים, נראה כי פיצ'ר 14 הינו פיצ'ר נומרי שערכיו מסתיימים ב"mm". בנוסף, פיצ'ר 9 המכיל ערכים בטווח 1-12 וכן פיצ'ר 22 המכיל ערכים בטווח 2010-2012 סווגו כברירת מחדל כפיצ'רים נומריים, על אף שהם נראים קטגוריאליים מהגרפים. לכן סיווג הפיצ'רים הנ"ל תוקן על מנת שיתאים לסוגיהם. לאחר התיקון, מספר הפיצ'רים עודכן ל17 פיצ'רים נומריים, 8 פיצ'רים קטגוריאליים ופיצ'ר label. לאחר מכן, נבחן כל פיצ'ר באופן אינדיבידואלי. לשם כך נבנתה פונקציה ייעודית המציגה את הנתונים עבור כל פיצ'ר תוך התייחסות לסוגו:

- **<u>פיצ'רים נומריים</u>** – עבור פיצ'רים אלו הוצגו נתונים סטטיסטיים, שכיחות הערכים שלהם בתצוגה גרפית, הערכים החריגים הפוטנציאליים בגרף boxplot, שכיחות התיוגים "0" או "1" בגרפים נפרדים ובגרף מאוחד וכן מספר הערכים החסרים.
- **<u>פיצ'רים קטגוריאליים/בינאריים</u>** – עבור פיצ'רים אלו נבחנו הוצגו מספר הערכים של כל קטגוריה, מספר הערכים הייחודיים, הקטגוריה השכיחה ביותר, שכיחות התיוגים "0" אל מול שכיחות התיוגים "1" עבור כל קטגוריה וכן מספר הערכים החסרים.

לבסוף, הוצגו גרפים נוספים שמספקים מידע נוסף על הפיצ'רים ועל הקשרים ביניהם:

- **מפת חום של הקורלציה בין הפיצ'רים**, במטרה לבחון האם קיים קשר בין הפיצ'רים.
- **מפת חום של הערכים החסרים בכל פיצ'ר**, במטרה לבחון האם מדובר בפיזור אחיד של ערכים חסרים או במקומות ספציפיים.

<u>הממצאים העיקריים</u>:

1. התגלה כי בין מספר פיצ'רים קיימת קורלציה חיובית הגבוהה מ0.9 (נספח 1). על אף קורלציה גבוהה זו בחרנו שלא להסיר בשלב זה פיצ'רים, אלא לבחון את הסוגיה בהמשך בחלק העיבוד המקדים.
2. קיימים פיצ'רים קטגוריאליים בעלי מספר זהה של ערכים ייחודיים (נספח 2). בפרט, עבור פיצ'רים 5,18,19 נמצאה התנהגות דומה של שכיחויות הערכים.
3. למספר פיצ'רים הייתה התנהגות דומה לזו של התפלגות נורמלית והתפלגות לוג-נורמלית (נספח 3).
4. למספר פיצ'רים הייתה שכיחות גבוהה של ערכים חסרים, ול8 פיצ'רים לא היו ערכים חסרים (נספח 4).
5. קרוב ל80% מהרשומות מתוייגות כ"0".

<u>חלק שני – עיבוד מקדים</u>

ראשית נציין כי הפעולות הסופיות שנבחרו לביצוע, נבחרו לאחר תהליך ממושך של ניסוי וטעייה. בנוסף, **כל הפעולות בוצעו הן על סט נתוני האימון והן על סט נתוני ה test** במקביל, למעט הוצאת חריגים שבוצעה על סט נתוני האימון בלבד. בחלק זה בוצעו החלקים הבאים:

1. **<u>התמודדות עם ערכים חסרים</u>** – בחלק זה, נבחנו הפיצ'רים הקטגוריאליים והנומריים באופן אינדיבידואלי.

א. **<u>עבור הפיצ'רים הקטגוריאליים</u>** בחלק מהקטגוריות ראינו שכיחות משמעותית של נתונים חסרים (נספח 4), ולכן בחרנו להמיר במקרה זה את הערכים החסרים לקטגוריה בפני עצמה, במטרה למזער ככל האפשר איבוד מידע. במקרים נוספים נמצאו פיצ'רים עם שכיחות נמוכה של ערכים חסרים (נספח 5), ואלו אוחדו עם הקטגוריה הפחות שכיחה. לאחר מכן נבחנו מספר שיטות לקידוד הקטגוריות:

1. שימוש ב one-hot encoding
2. שימוש ב one-hot encoding ולקיחת 5 הקטגוריות השכיחות ביותר (נספח 6)
3. הפיכת הפיצ'ר לבינארי, בהתאם לשכיחות הערכים ובפרט במקרה בו יש ערך דומיננטי

כפי שנמצא בחלק האקספלורציה, פיצ'רים 5,18 ו19 בעלי שכיחות נתונים דומה יחסית. לאחר ניסוי וטעייה, מצאנו כי השארת פיצ'ר 18 והסרת פיצ'רים 5 ו19 הביאו לביצועים טובים יותר. בנוסף, עבור פיצ'ר 9 נמצא כי לקיחת 5 הקטגוריות השכיחות ביותר הביאה אף היא לביצועים טובים יותר.
ספציפית עבור פיצ'ר 13 הבינארי, הערכים החסרים בו שסומנו תחילה כ"unknown" הושלמו ב-0.

ב. **<u>עבור הפיצ'רים הנומריים</u>**, הושלמו הערכים החסרים לפי החציון ולפי הממוצע. הבחירה להשלמת הערכים באמצעות החציון או הממוצע נבחרה לאחר בחינה מדוקדקת של כלל הפיצ'רים, לרבות הממוצע והחציון בכל אחד מהם. בפרט, הבחנו כי בין פיצ'ר 14 לפיצ'ר 13 נמצא קשר לידי ביטוי באופן הבא: בפיצ'ר 14 היכן שהערך ברשומה גדול ממש מ1, הערך בפיצ'ר 13 הוא 1, ואחרת הוא 0. לכן, באופן פרטני, בחרנו להפוך את פיצ'ר 14 לבינארי, תוך בחינת האפשרות למחוק את אחד מהפיצ'רים ואף את שניהם, או להשאיר את שניהם. לאחר ניסוי וטעיה, מצאנו כי הביצועים הגבוהים ביותר התקבלו כאשר פיצ'ר 14 מקודד לפיצ'ר בינארי כאשר הערכים החסרים בו הושלמו להיות 1.
לאחר סיום חלק זה וביצוע הקידודים השונים, ישנם 75 מאפיינים כולל ה label.

2. **<u>הוצאת חריגים</u> –** רלוונטי עבור הפיצ'רים הנומריים בלבד. כאמור, החריגים הוצאו רק ברשומות שבסט נתוני האימון. כל פיצ'ר נומרי נבחן באופן אינדיבידואלי באמצעות שיטת הטווח הבין-רבעוני (Interquartile Range). בפיצ'רים שהתקבלו עבורן קבוצות גדולות של ערכים חריגים, בחרנו לא לבצע הסרת חריגים במטרה לא לאבד מידע. בנוסף, מהתנהגות הפיצ'רים ראינו כי מתקיימת בחלקם התנהגות של התפלגות לוג-נורמלית (נספח 3). בעקבות ההתפלגות הזו, אנו סבורים כי ערכים שאינם בהכרח חריגים סווגו ככאלה בשיטת הטווח הבין רבעוני שמניחה התנהגות של התפלגות נורמלית.
לכן במטרה להפחית את כמות הערכים החריגים, בחרנו להפעיל log transform. הפעלת הלוג הופכת את הפיצ'ר להתפלגות המזכירה התפלגות נורמלית, וראינו כי בפעולה זו כמות הערכים החריגים קטנה וכן השתפרו ביצועי המודלים. לבסוף, עבור הפיצ'רים שהתנהגו בדומה להתפלגות נורמלית, בחרנו כן להסיר את הערכים החריגים. בסה"כ הוסרו 178 ערכים חריגים, ולאחר הסרתם נשארו בסט האימון 21,982 רשומות. נציין שוב כי הסרת הערכים החריגים לא חלה על סט נתוני ה test.

3. **<u>מניפולציות מתמטיות ו-Clustering</u>** – יצרנו פיצ'רים חדשים באמצעות פעולות מתמטיות בין העמודות, ובפרט בחנו יצירה של פיצ'רים שיש בהם תלות לא ליניארית במטרה לנסות לשפר את ביצועי המודלים, אך לא נמצא כי פיצ'רים אלה הביאו לשיפור בביצועי המודלים, ולכן נבחר לא להוסיפם לנתונים.

4. **<u>נרמול</u>** – סט הנתונים הן של האימון והן של הtest אינם מנורמלים, כפי שנבחן בשלב האקספלורציה. טווח הערכים של כל פיצ'ר משתנה ומגוון, בין הפיצ'רים וגם בפיצ'רים עצמם, מצב שעשוי להעיד בין היתר על אופי הנתונים וגם על טעויות הקלדה, הזרמה כפולה של הנתונים ועוד. יש חשיבות רבה לנרמול, בעיקר כדי להגיע למצב שבו הנתונים מדברים באותה השפה, כלומר להגיע למצב בו הפיצ'רים בעלי מאפיינים קרובים תוך מזעור ההטיות בין הנתונים השונים עד כמה שאפשר.
לאחר הפרדת סט הנתונים הן של האימון והן של הtest ל17 פיצ'רים נומריים ו57 בינאריים, נבחנו עבור כל סט 2 שיטות נרמול:
   א. **<u>סטנדרטיזציה – Z-Score normalization</u>** – בשיטה הזו הנתונים מנורמלים כך שכל פיצ'ר ינורמל בקירוב להתפלגות נורמלית בעלת ממוצע 0 וס"ת 1.
   ב. **<u>שיטת MinMax Scaling</u>** – בשיטה זו הנתונים מנורמלים כך שטווח הערכים יהיה נתון (0-1) בכל פיצ'ר.
כדי לבחון איזו שיטת נרמול מתאימה יותר לסט הנתונים, בחנו את ביצועי המודלים בכל שיטה, עם הפיצ'רים החדשים שנוצרו בשלב המניפולציות המתמטיות ובלעדיהם. <u>תחת ההנחה</u> כי הביצועים של המודל טובים יותר ככל שמדד הAUC **הסופי** (כלומר בשלב ביצוע הערכת המודל בחלק 4) גבוה יותר, מצאנו כי שיטת הסטנדרטיזציה ללא הפיצ'רים החדשים מביאה לביצועים הגבוהים ביותר. חשוב לציין כי הנרמול בוצע על הפיצ'רים הנומריים בלבד הן בסט האימון והן בסט נתוני ה test.

5. **<u>ממדיות הבעיה והקטנתה</u>** – ממדיות נתונים גדולה מידי בעייתית ממספר סיבות, כאשר העיקרית שבהן היא העובדה שככל שכמות הפיצ'רים גדולה יותר, כך נדרשות הרבה דגימות, וכתוצאה מכך גם כמות הרעש גדולה יותר וגם סיבוכיות החישוב גדולה יותר. בפרט, בבעיות סיווג, כמות מאפיינים גדולה וכמות נתונים גדולה מקשה על האבחנה בין תצפיות חריגות לבין רעש. כמו כן, לפי כלל האצבע שנלמד בכיתה, עבור N מאפיינים נדרשות N^2 דגימות, ולכן ככל שיש יותר מאפיינים כך נדרשות הרבה יותר דגימות, מה שעלול לגרום ל Overfitting.
בפרויקט זה עבור 21,982 הרשומות שנותרו לאחר הסרת החריגים, ממדיות הבעיה תהיה גדולה מידי כאשר יהיו למעלה מ$\sqrt{21982} \approx 149$ פיצ'רים. לאחר שבוצעו שלבי עיבוד הנתונים הנ"ל, קיימים 74 פיצ'רים (לא כולל ה label), ולכן היא אינה גדולה מידי. עם זאת, כפי שנצפה בשלב האקספלורציה, קיימים פיצ'רים שיש ביניהם קורלציה גבוהה, ולכן השתמשנו בשיטת PCA לאיתור התלויות הלינאריות בין הפיצ'רים וע"י כך להקטנת הממדיות. בהמשך לחלוקת הפיצ'רים לנומריים ובינאריים, ביצענו PCA על 17 הפיצ'רים הנומריים לאחר שעברו נרמול כדי למצוא כמה קומפוננטות מסבירות לפחות 95% מהשונות. אנו שואפים להסבר רב של השונות מכיוון שאיננו יודעים בוודאות מה מייצגים הנתונים.
מצאנו כי 9 קומפוננטות מתוך 17 הפיצ'רים מסבירות לפחות 95% מהשונות בנתונים (נספח 7). בנוסף, בדקנו האם ביצוע PCA על הפיצ'רים הבינאריים והקטנת הממדיות בפיצ'רים אלה משפר את ביצועי המודלים, ומצאנו כי אמנם 46 קומפוננטות מתוך 57 מסבירות לפחות 95% מהשונות, אך דווקא שימוש ב46 הפיצ'רים הביא הן לביצועים נמוכים יותר והן להגדלה משמעותית בסיבוכיות החישוב ובזמן הריצה. לכן, בחרנו לבצע PCA רק על הפיצ'רים הנומריים.
לאחר איחוד הפיצ'רים הנומריים שעליהם בוצע PCA עם הפיצ'רים הבינאריים, נותרנו עם 66 פיצ'רים.

כשלב אחרון ביצענו בדיקת קורלציה נוספת בין הפיצ'רים, ומצאנו כי אמנם אין פיצ'רים בעלי קורלציה מובהקת ביניהם, אך כן נצפו פיצ'רים עם קורלציה שגבוהה משמעותית מ0 (נספח 8). עם זאת, לא נמצאו פיצ'רים בעלי קורלציה גבוהה במיוחד (מעל 85%) ולכן לא הוסרו פיצ'רים נוספים בשלב זה. כמו כן, לכל פיצ'ר הוצג גרף שכיחות התיוג "1" לעומת תיוג "0" (נספח 9).

חלק שלישי – הרצת המודלים

המודלים שנבחרו להרצה הם:

- מודלים ראשוניים – Naïve Bayes, Logistic Regression
- מודלים מתקדמים – Multi-Layer Perceptron (ANN), Adaptive Boosting (AdaBoost)

עבור כל מודל נבחנו מספר אפשרויות להיפר-פרמטרים באמצעות GridsearchCV. האפשרויות נבחרו מתוך ניסוי וטעייה. בחירת סט האפשרויות להיפר פרמטרים נוצרה מתוך ניסוי וטעייה, תוך התחשבות במגבלת זמן הריצה של הפרויקט. את בחירות ההיפר-פרמטרים מתוך סט האפשרויות שניתן עבור כל מודל ניתן לראות בטבלה 1. ההסבר על ההיפר-פרמטרים ניתן למצוא בנספח 10.

המודל	היפר פרמטרים אפשריים	החלופה הנבחרת	היפר פרמטרים שנותרו כברירת מחדל
Naïve Bayes	'priors' : [None], 'var_smoothing' : [1e-9, 1e-7, 1e-5, 1e-3, 0.1, 1, 3]	{'priors': None, 'var_smoothing': 0.1}	--
Logistic Regression	'penalty' : ['l1', 'l2'], 'C' : [0.001, 0.01, 0.1, 0.5, 1, 10, 100], 'tol' : [0.1, 0.01, 0.001], 'max_iter' : [2000], 'solver' : ["liblinear"]	{'C': 0.1, 'max_iter': 2000, 'penalty': 'l2', 'solver': 'liblinear', 'tol': 0.001}	dual, fit_intercept, intercept_scaling, class_weight, multi_class, verbose, warm_start
Multi-Layer Perceptron (ANN)	'activation' : ["relu"], 'hidden_layer_sizes' : [(10,), (20,), (50, 50), (20, 20, 10, 10, 10),(100,)], 'learning_rate_init' : [0.01, 0.001], 'random_state' :[100], 'max_iter': [2000]	{'activation': 'relu', 'hidden_layer_sizes': (10,), 'learning_rate_init': 0.001, 'max_iter': 2000, 'random_state': 100}	solver, alpha, batch_size, learning_rate, power_t, shuffle, verbose, warm_start, momentum, beta_1, beta_2, epsilon, early_stopping
Adaptive Boosting (AdaBoost)	'n_estimators':[500,1000], 'learning_rate': [0.01,0.1,0.3], 'random_state' :[100]	{'learning_rate': 0.1, 'n_estimators': 1000, 'random_state': 100}	base_estimator, algorithm

טבלה 1: פירוט ההיפר-פרמטרים שנבחרו עבור כל מודל

החלופה הנבחרת נבחרה לפי מטריקת AUC באמצעות הפונקציה GridsearchCV.

חלק רביעי – הערכת המודלים

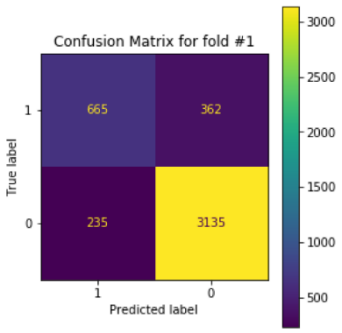
עבור כל המודלים שנבחרו בחלק השלישי התבצעה הערכה באמצעות K-Fold Cross Validation, וטיב ביצועי המודל נבחן על סמך מדד ה-AUC שהתקבל מביצועיהם. בכל מודל ההרצה התבצעה ע"י חלוקת סט נתוני האימון ל5 מחיצות, כאשר בכל מחיצה יש חלוקה בין ה Train לבין הValidation. נבחר גרעין אקראיות random_state=100 לטובת עקביות התוצאות. מספר המחיצות נקבע להיות 5 משיקולי סיבוכיות חישובית וכמות הנתונים. לכל מודל נמדד גם זמן הריצה שלו.

<u>עבור כל מודל</u>:

- חושב מדד דיוק עם משקל זהה לכל הדגימות (מסומן כמדד דיוק 1)
- חושב מדד דיוק משוקלל עם עלות סיווג שגוי של דגימה בעלת תיוג "1" חמורה פי 5 מעלות סיווג שגוי של דגימה בעלת תיוג "0" (מסומן כמדד דיוק 2)
- נבחן המודל על הtest ועל ה train, וחושב הפער בין הAUC שהתקבל על ה train לבין הAUC שהתקבל על הtest. ככל שהפער גדול יותר כך הסיכוי ל overfitting גדול יותר. אנו הגדרנו שהפרש גדול מ0.1 מעיד על overfitting.

- הוצג גרף עם הROC הממוצע שכולל את מדד הAUC הממוצע.
- הוצגה מטריצת הטעות (Confusion Matrix) עבור כל מחיצה שנבחנה.

מוצגת להלן מטריצת טעות מדגמית על מחיצה 1 במודל רשתות נוירונים.
משמעות הערכים בתוך המטריצה:
665 דגימות סווגו בצדק כ-1 (TP), 3135 דגימות סווגו בצדק כ-0 (TN),
235 דגימות סווגו בטעות כ-1 (FP) ו362 דגימות סווגו בטעות כ-0 (FN).
הדיוק שהתקבל במחיצה זו הוא $\frac{665+3135}{235+362+665+3135} = 86.4\%$



סיכום ההשוואה בין המודלים מוצג בטבלה 2.

המודל	AUC ממוצע על ה Test	AUC ממוצע על ה Train	מצב overfitting	מדד דיוק 1	מדד דיוק 2
Multi-Layer Perceptron (ANN)	**0.902485**	**0.928690**	**אין**	**0.864798**	**0.729191**
Adaptive Boosting (AdaBoost)	0.887208	0.899197	אין	0.855245	0.739983
Logistic Regression	0.881546	0.884323	אין	0.849831	0.724014
Naïve Bayes	0.871117	0.871846	אין	0.840597	0.779071

טבלה 2: פירוט המדדים שהתקבלו עבור כל מודל

<u>מסקנות והערות</u>:

- בנספח 11 ניתן לראות סיכום ממצאים מפורט שהתקבל עבור כל מודל.
- ניתן לראות כי במדד הדיוק השני, דווקא המודל Naïve Bayes היה בעל ערך הדיוק הגבוה ביותר.
- זמן הריצה הקצר ביותר היה של המודלים הראשוניים, והארוך ביותר במודלים המתקדמים.
- עבור כל המודלים שנבחנו, ההפרש בין הAUC הממוצע על סט נתוני האימון לבין סט נתוני ה-test קטן מ0.1, כלומר לא נמצא overfitting. הדבר מעיד כי ההיפר-פרמטרים שנבחרו התאימו ללמידת המודלים. בפרט, ניכר לראות שבמודל Naïve Bayes, הפער בין הAUC על הtest לבין הAUC על הtrain היה קטן מאוד. באם היה נמצא overfitting, היינו מנסים מספר פעולות לטובת הגדלה של יכולת ההכללה שלו, וביניהן הוספת נתונים באופן מלאכותי, הסרת מאפיינים והפחתת מורכבות המודל באמצעות רגולריזציה (Regularization).

המודל **רשתות נוירונים (Multi-Layer Perceptron (ANN** מוביל בביצועים. **לפי מטריקת AUC, בחרנו במודל זה לביצוע החיזוי** על סט נתוני ה test. בסט נתוני האימון, כ23.4% מהדגימות מתוייגות כ"1" והשאר כ"0". כשהרצנו את המודל על סט נתוני ה-test, התקבל כי 21.6% מהדגימות תויגו כ"1" והשאר כ"0". מכאן שהתוצאה שקיבלנו הינה בגדר הסביר. על סמך הממצאים הנ"ל, אנו חוזים כי בממוצע, הדיוק של מודל רשתות הנוירונים על סט נתוני ה-test הינו כ86%.

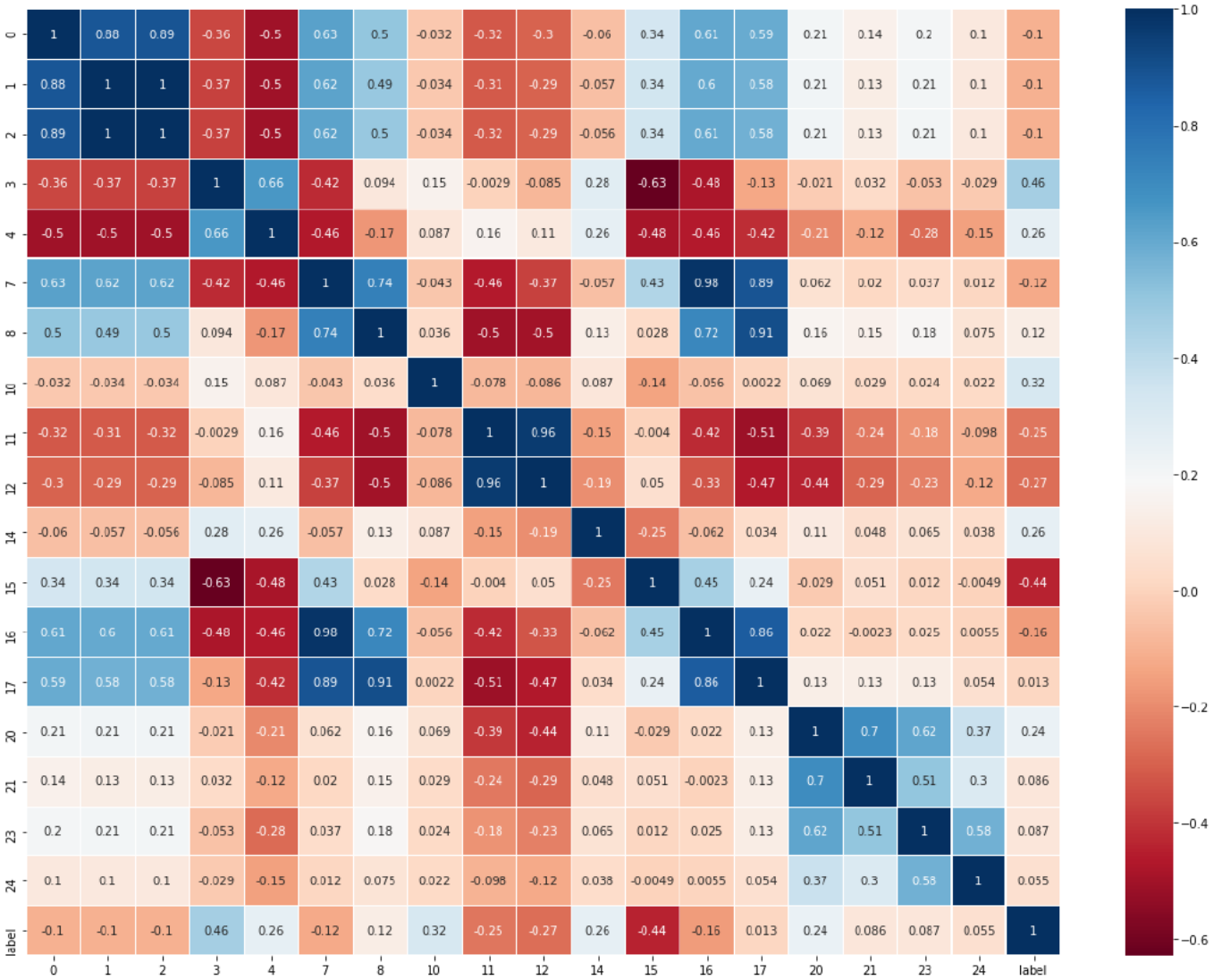
<u>סיכום</u>

בפרויקט זה בוצעה הערכה על סט נתונים. בשלב האקספלורציה, בחנו לעומק את התנהגויות הפיצ'רים ובפרט את התפלגותם. לצורך הבדיקות, נעשה שימוש רב בויזואליזציה. ככלל, כל פיצ'ר נבחן לגופו, ובהתאם התקבלו החלטות אינדיבידואליות. לאחר מכן בוצע עיבוד מקדים על הנתונים, תוך ההבנה כי לעיבוד המקדים יש השפעה קריטית על ביצועי המודלים. בשלב זה בוצעה התמודדות עם ערכים חסרים, הוסרו ערכים חריגים, בוצע נרמול ונבחנה ממדיות הבעיה. לאחר העיבוד המקדים, נבחרו ונבחנו 4 מודלים: 2 מודלים התחלתיים Naïve Bayes, Logistic Regression ו2 מודלים מתקדמים Multi-Layer Perceptron(ANN), Adaptive Boosting (AdaBoost). התאמת ההיפר פרמטרים ובחינת המודלים כללה ניסוי וטעייה וכן חזרה מרובה אל שלב העיבוד המקדים לצורך בחינת שיפור ביצועי המודלים. ניכר לראות את ההבדלים בין המודלים ההתחלתיים, שהיו עם זמן ריצה קטן מאוד וכן צרכו מעט מאוד משאבים, לבין המודלים המתקדמים, שזמן הריצה היה ארוך יותר ונדרשו משאבים רבים יותר לפעולתם. כמו כן, ניכר לראות כי מבין המודלים שנבחנו, המודל בעל הביצועים הנמוכים ביותר היה Naïve Bayes והמודל בעל הביצועים הגבוהים ביותר היה רשתות נוירונים, ולכן רשתות נוירונים נבחר להיות המודל לחיזוי הסיווגים על סט נתוני ה test.

נספחים

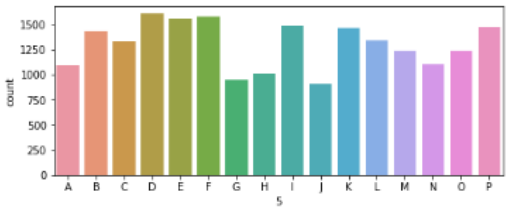
נספח 1

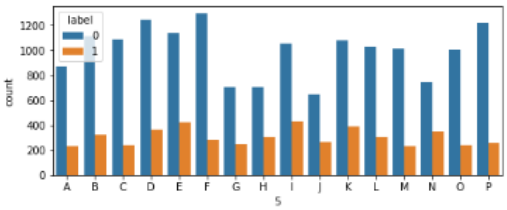
בחינת הקורלציה בין הפיצ'רים בשלב האקספלורציה:



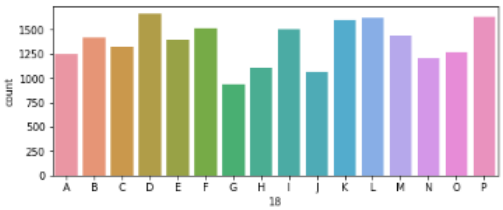
נספח 2 – פיצ'רים 5,18,19

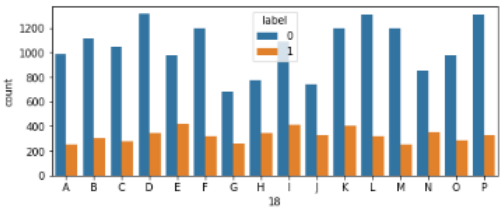
```
Feature: 5
------------------------
count      20812
unique        16
top            D
freq        1604
Name: 5, dtype: object
Total nulls: 1349
```



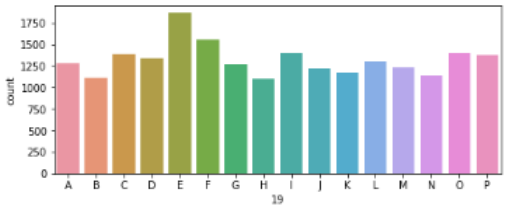


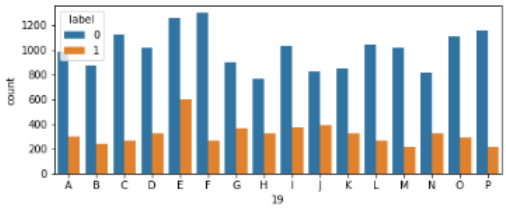

```
Feature: 18
------------------------
count      21951
unique        16
top            D
freq        1659
Name: 18, dtype: object
Total nulls: 210
```






```
Feature: 19
------------------------
count      21141
unique        16
top            E
freq        1861
Name: 19, dtype: object
Total nulls: 1020
```

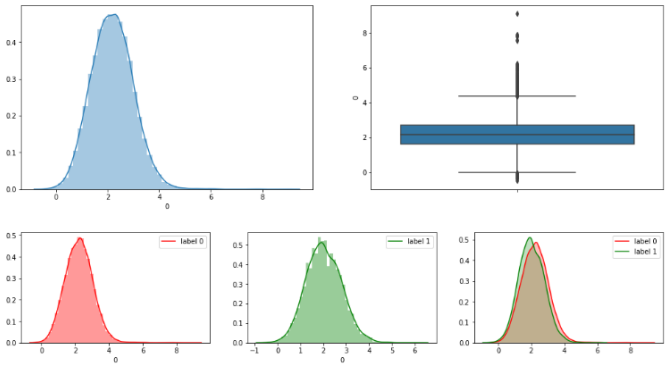




נספח 3

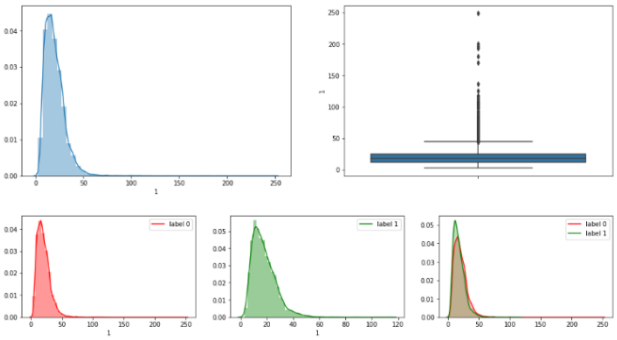
פיצ'ר 0 בעל התנהגות של התפלגות נורמלית:

```
Feature: 0
-----------------------
count    22161.000000
mean         2.185958
std          0.815080
min         -0.490607
25%          1.622068
50%          2.167701
75%          2.720341
max          9.092011
Name: 0, dtype: float64
Total nulls: 0
```



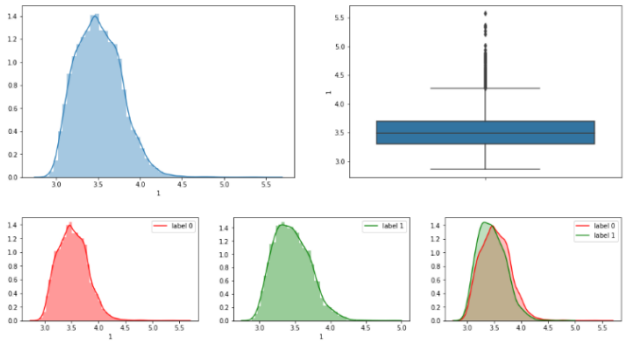
פיצ'ר 1 בעל התנהגות של התפלגות לוג-נורמלית:

```
Feature: 1
-----------------------
count    22161.000000
mean        19.797754
std         10.763614
min          2.437300
25%         12.268371
50%         17.833216
75%         25.196446
max        248.877854
Name: 1, dtype: float64
Total nulls: 0
```



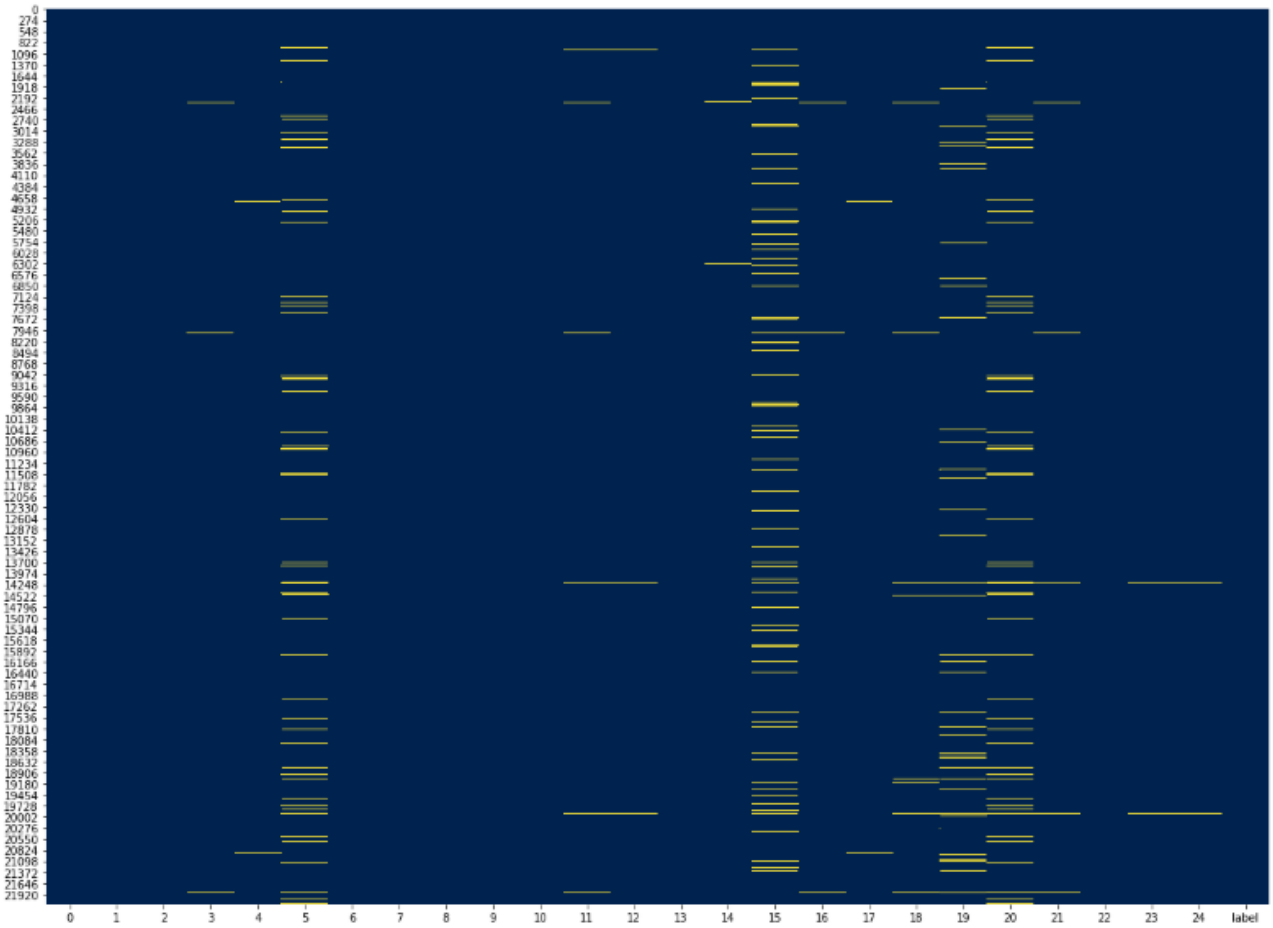
פיצ'ר 1 לאחר הפעלת log-transform (מקור לשיטה) מקבל התנהגות של התפלגות נורמלית:

```
Feature: 1
-----------------------
count    22161.000000
mean         3.510274
std          0.272544
min          2.858612
25%          3.305727
50%          3.491441
75%          3.693779
max          5.575486
Name: 1, dtype: float64
Total nulls: 0
```



נספח 4

מפת חום של הערכים החסרים:



טבלה שמסכמת את מספר הערכים החסרים בכל פיצ'ר:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	label
0	0	0	0	59	104	1349	0	7	18	0	0	106	113	0	81	1871	28	52	210	1020	1345	99	0	100	100	0

שכיחות משמעותית של ערכים חסרים בפיצ'ר 5:

```
D      1604
F      1575
E      1558
I      1485
P      1475
K      1469
B      1432
NaN    1349
L      1337
C      1330
O      1241
M      1239
N      1098
A      1097
H      1010
G       952
J       910
Name: 5, dtype: int64
```

נספח 5

שכיחות נמוכה של ערכים חסרים בפיצ'ר 13:

```
0          16906
1           5174
unknown       81
Name: 13, dtype: int64
```

נספח 6

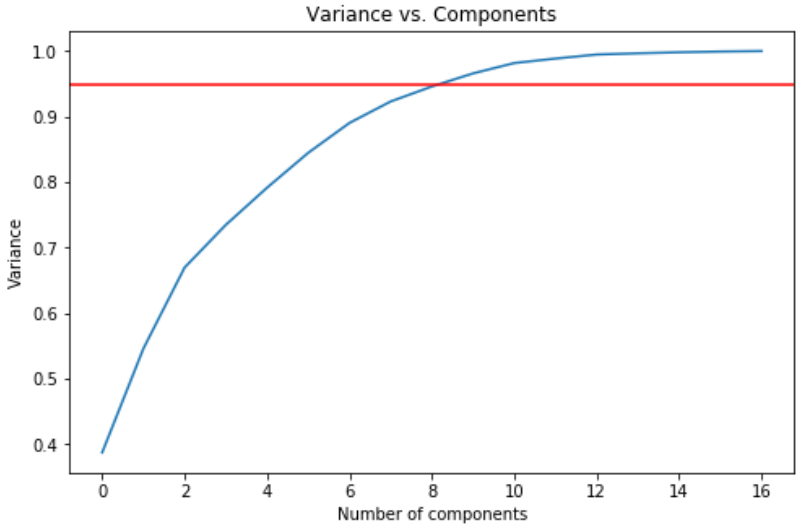
לקיחת 5 הקטגוריות השכיחות ביותר בוצע בפיצ'ר 9 (מקור לשיטה).

קטגוריה	כמות מופעים
5	2016
8	1992
7	1987
6	1974
3	1966
1	1963
9	1949
10	1943
11	1932
2	1828
12	1316
4	1295

Name: 9, dtype: int64

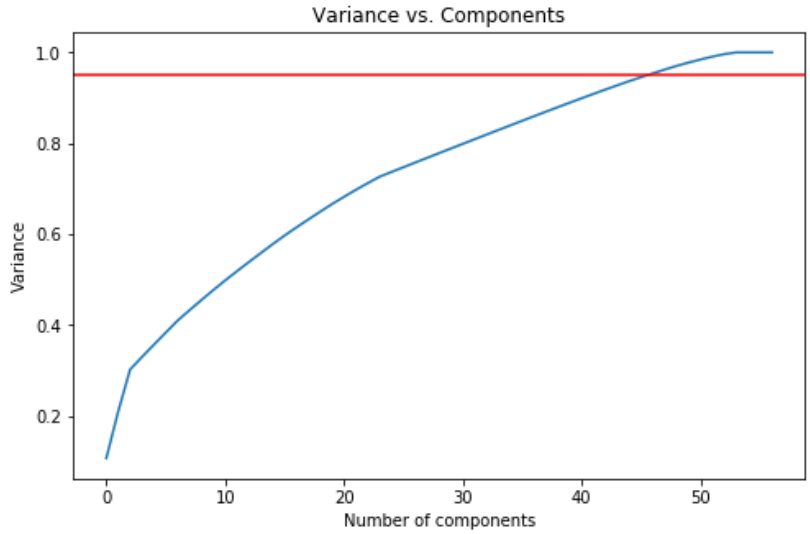
נספח 7

גרף PCA עבור הפיצ'רים הנומריים:




```
[9] components explains at least 95 percent of the variance in the data
```

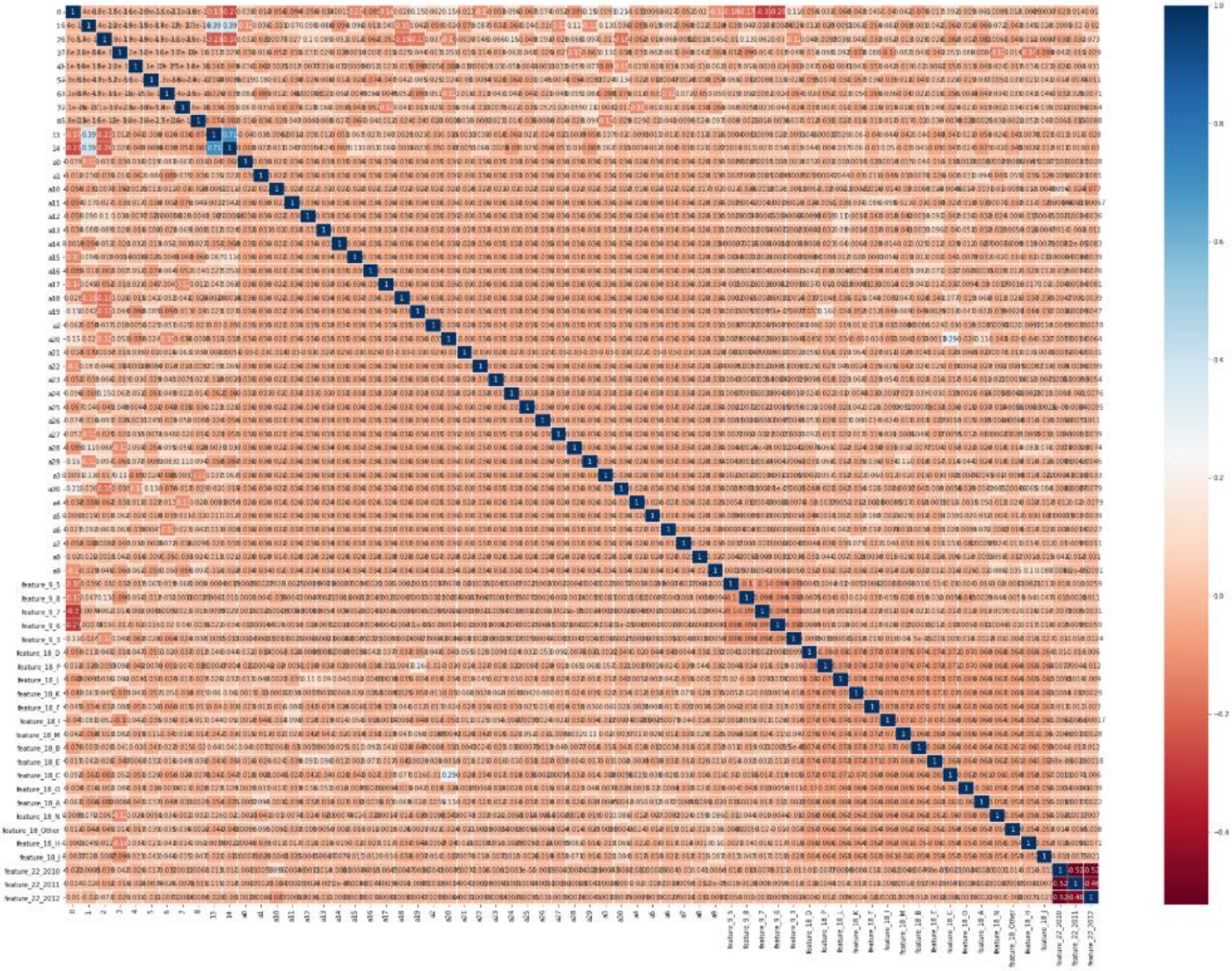
גרף PCA עבור הפיצ'רים הבינאריים:




```
[46] components explains at least 95 percent of the variance in the data
```

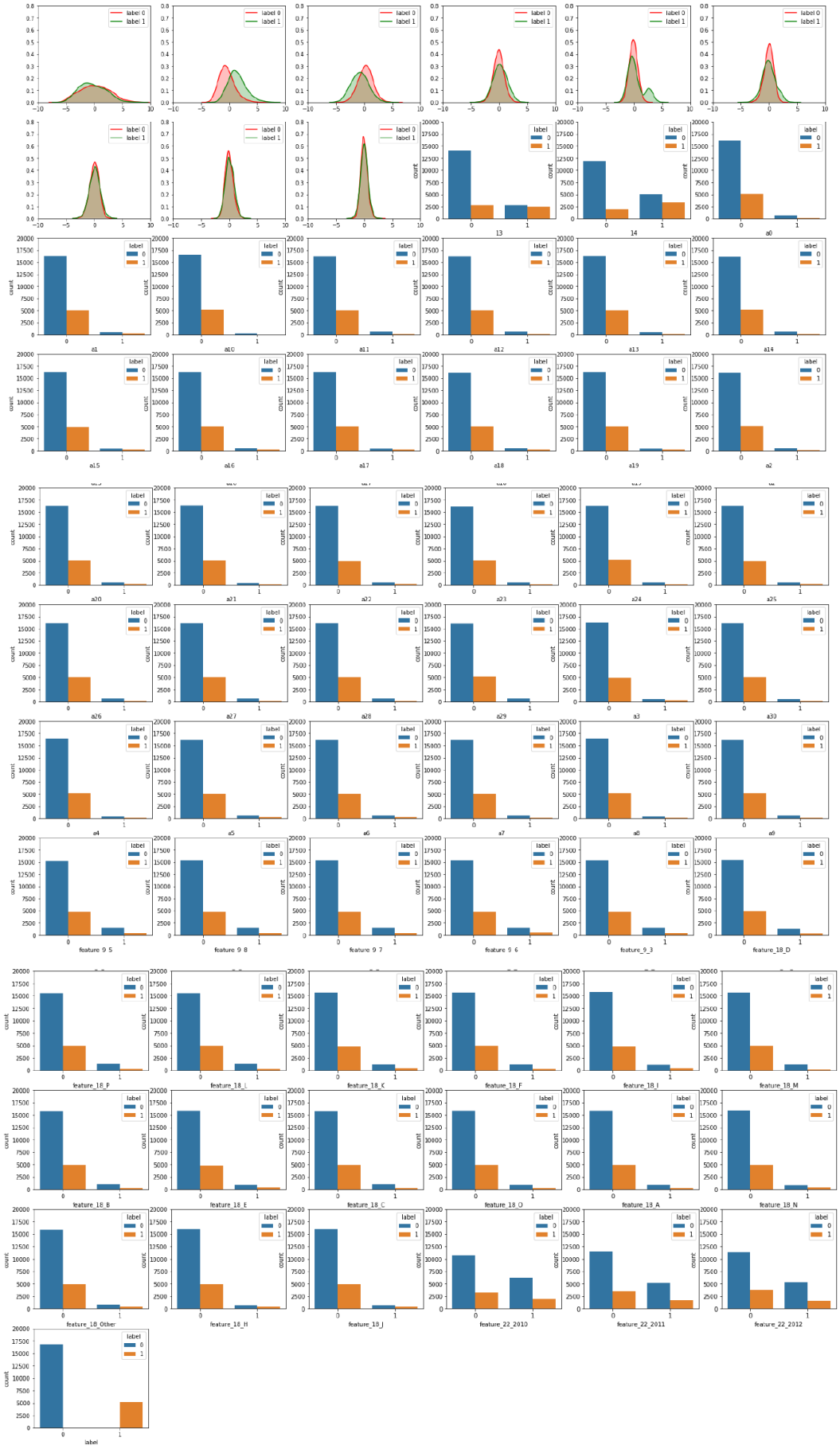
נספח 8

קורלציה בין הפיצ'רים לאחר העיבוד המקדים ולאחר ביצוע PCA על הפיצ'רים הנומריים:



נספח 9

שכיחות התיוגים 0 לעומת 1 בפיצ'רים לאחר העיבוד המקדים ולאחר ביצוע PCA על הפיצ'רים הנומריים:

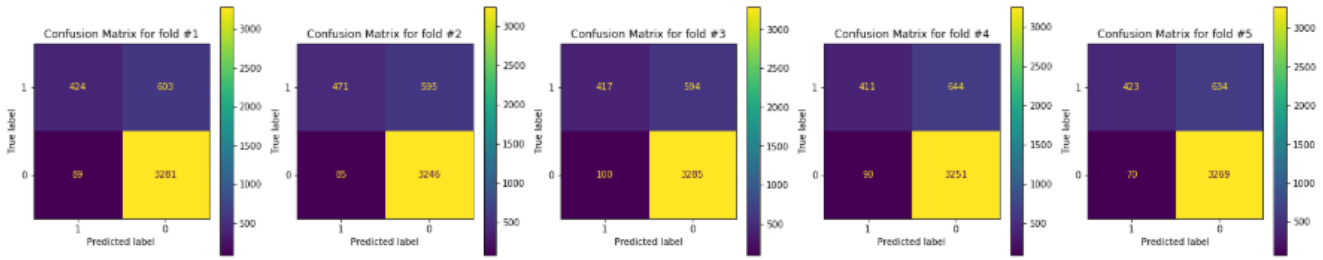


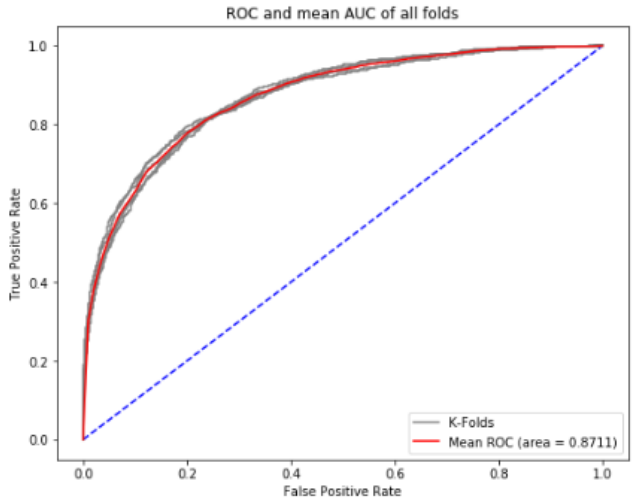
נספח 10

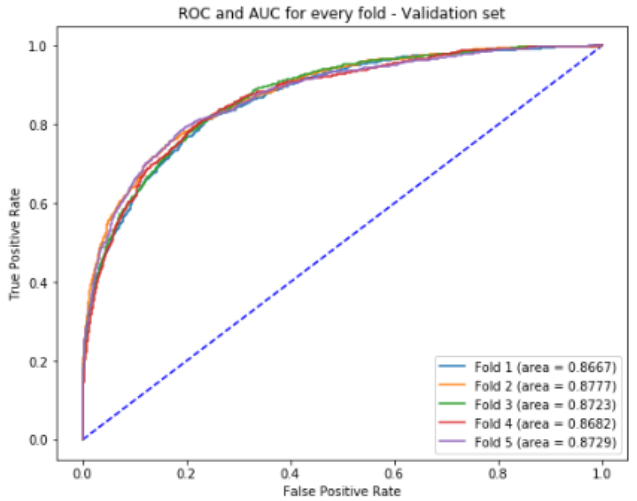
המודל	היפר פרמטר	הסבר
Gaussian Naïve Bayes	priors	ההסתברויות הפריוריות לסיווגים
	var_smoothing	תוספת של השונות הגדולה ביותר מבין הפיצ'רים לכל הפיצ'רים
Logistic Regression	penalty	פונקצית הקנס עמה נשתמש ברגולשריזציה: l1 – מרחק בערך מוחלט, l2 – מרחק בריבוע.
	tol	תנאי עצירה להתכנסות
	C	פרמטר הרגולריזציה ההופכי $\left(\frac{1}{\lambda}\right)$, חייב להיות ערך חיובי. ככל שC גדול יותר יש פחות רגולריזציה
	solver	האלגוריתם בו משתמשים בבעית האופטימיזציה
	max_itr	הגבלת איטרציות ההתכנסות למספר מקסימלי כלשהו.
	random_state	גרעין אקראי לטובת תוצאות קבועות
Multi Layer Perceptron (ANN)	activation	פונקצית אקטיבציה שתפעל בין כל הנוירונים ברשת
	hidden_layer_sizes	כמה נוריונים חבויים יש בכל שכבה, ארכיטקטורת הרשת
	alpha	פרמטר רגולריזציה (λ).
	solver	התהליך האיטרטיבי של מציאת משקולות, למשל stochastic gradient descent.
	learning_rate_init	קצב הלמידה, גודל הקפיצה.
	learning_rate	איך קצב הלמידה מעדכן את עצמו (קצב קבוע, אדפטיבי ובהתאם לpower_t)
	power_t	אם הוגדר שמקטנים את קצב הלמידה מepoch ל epoch אז היפר פרמטר זה קובע בכמה מקטנים.
	early_stopping	עצירת האימון אם אין שיפור ע"י שימוש בסט validation פנימי, במידה ומסומן כ True. False כברירת מחדל.
	batch_size	מספר הרשומות בכל קבוצה. כברירת מחדל – גודל קבוצת רשומות הוא המינ' מבין 200 לבין מספר הרשומות הכולל.
	warm_start	אם אנחנו מאמנים את הרשת בפעם נוספת, האם להתחיל מהאימון האחרון או לא.
	max_itr	הגבלת איטרציות ההתכנסות למספר מקסימלי כלשהו.
	random_state	גרעין אקראי לטובת תוצאות קבועות
Adaptive Boosting (AdaBoost)	n_estimators	מספר העצים
	learning_rate	קצב למידה
	algorithm	האלגוריתם לאופטימיזציה
	random_state	גרעין אקראי לטובת תוצאות קבועות

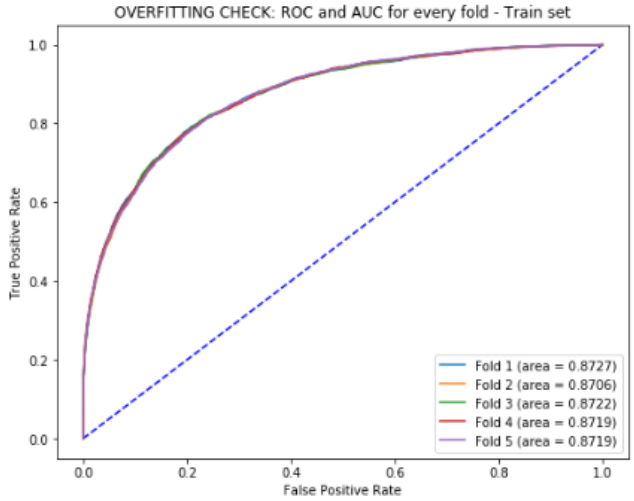
נספח 11

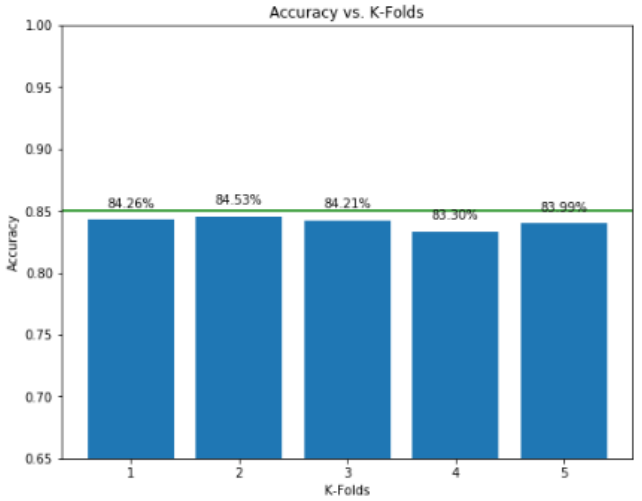
סיכום ממצאים עבור Naïve Bayes:





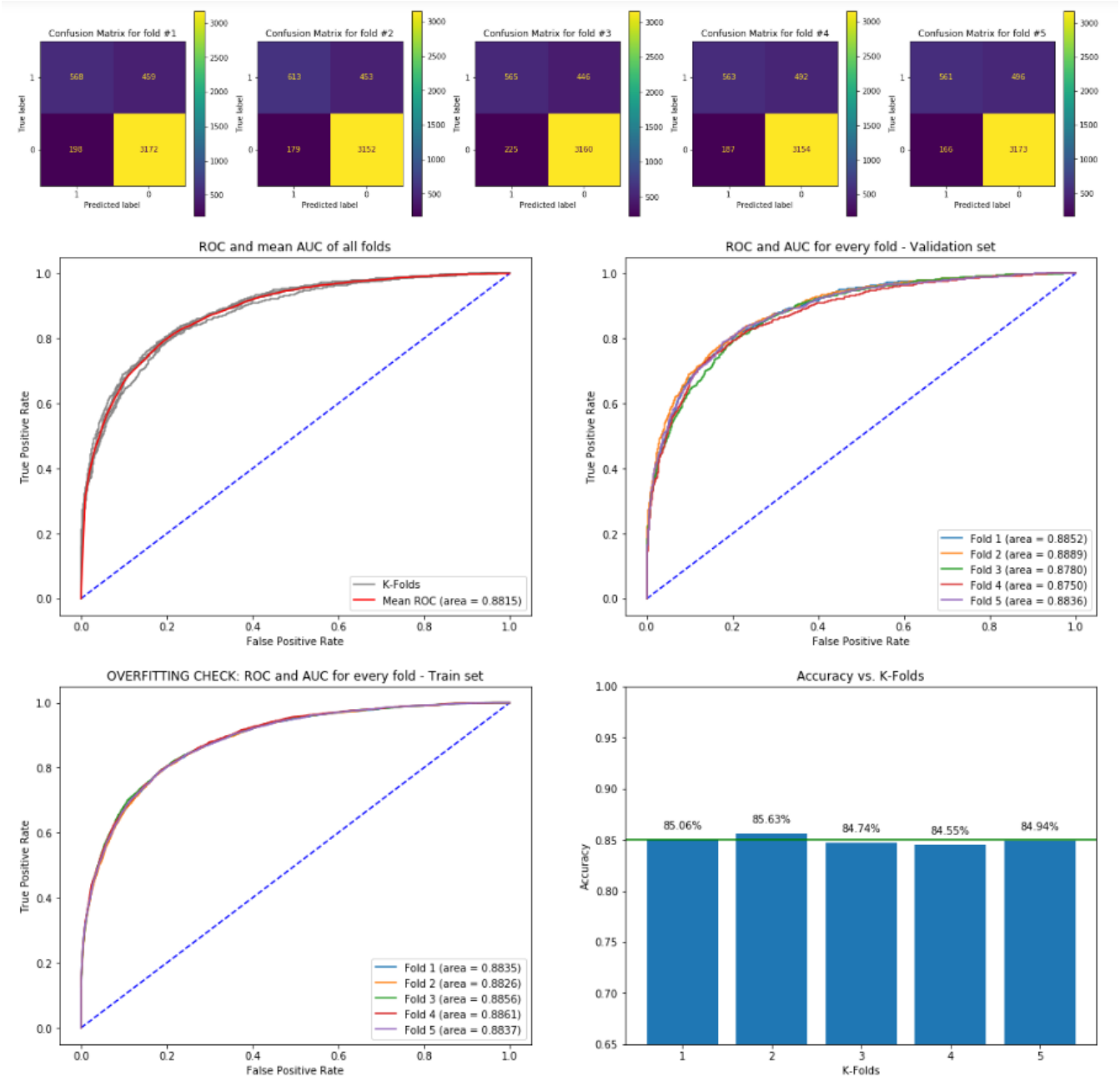






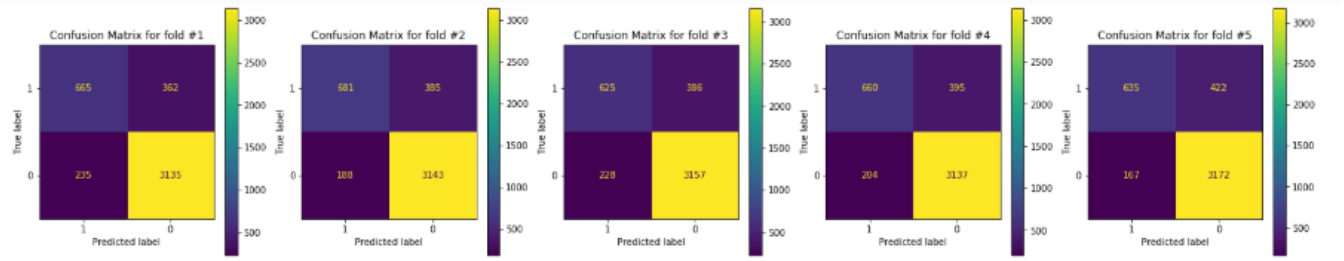

```
Mean Accuracy: 0.84059654
Mean accuracy, where classifing wrongly a 1 target is 5 times more severe than classifying wrongly a 0 target: 0.77907075
Mean AUC Test: 0.87111697
Mean AUC Train: 0.87184609
Difference between AUC: 0.00072912
```

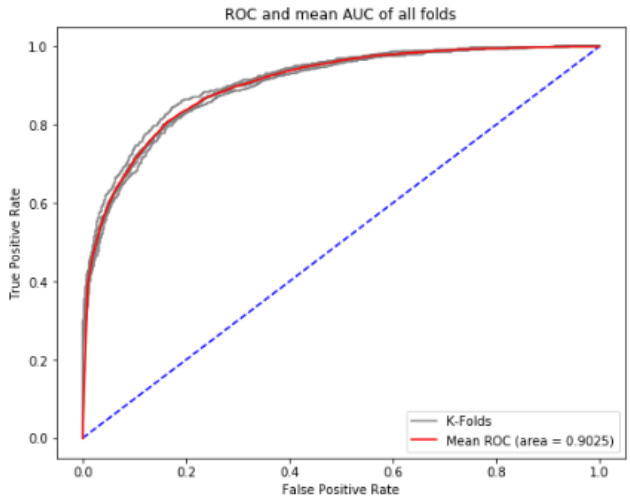
סיכום ממצאים עבור Logistic Regression:

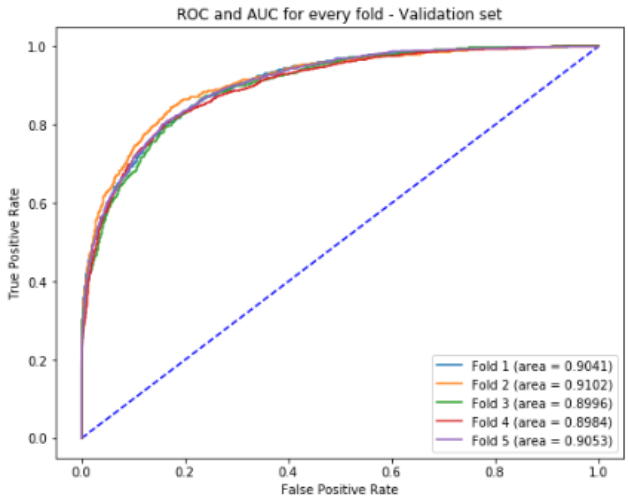


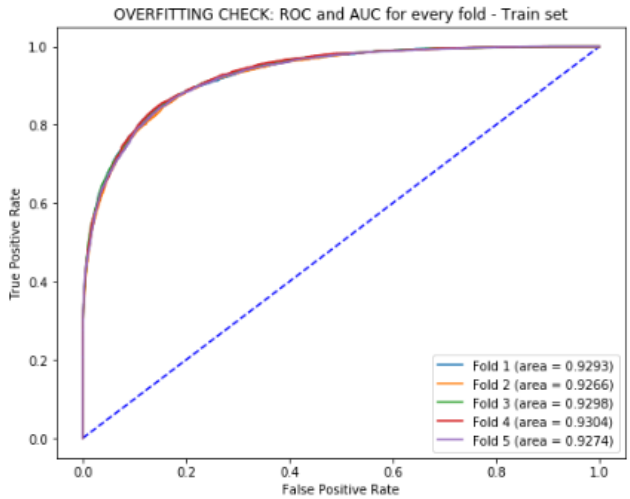

```
Mean Accuracy: 0.84983135
Mean accuracy, where classifing wrongly a 1 target is 5 times more severe than classifying wrongly a 0 target: 0.72401364
Mean AUC Test: 0.88154591
Mean AUC Train: 0.88432259
Difference between AUC: 0.00277668
```

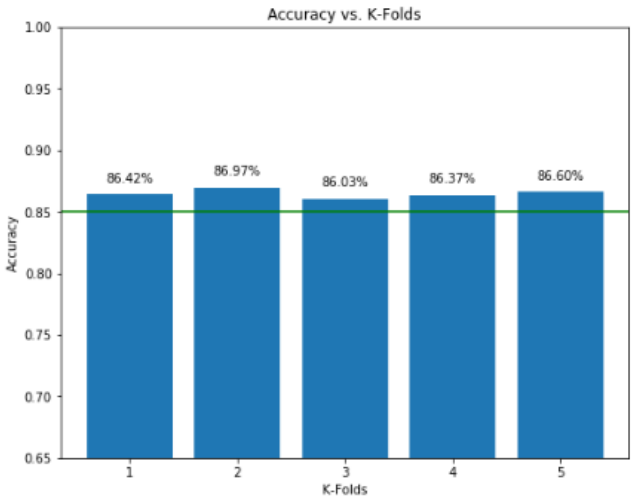
סיכום ממצאים עבור Multi-Layer Perceptron (ANN):





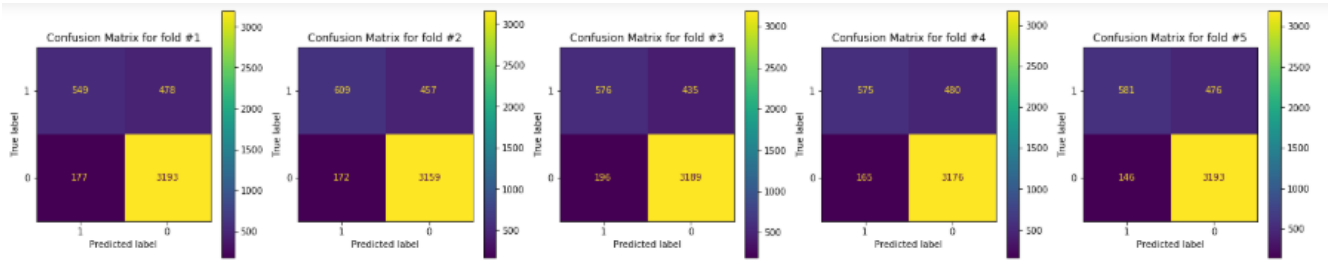


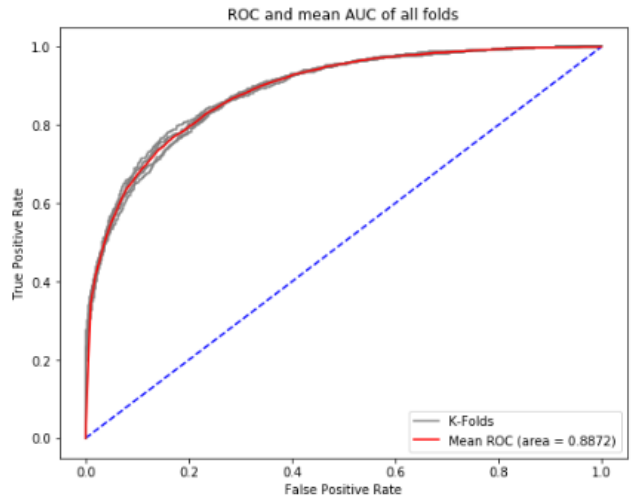


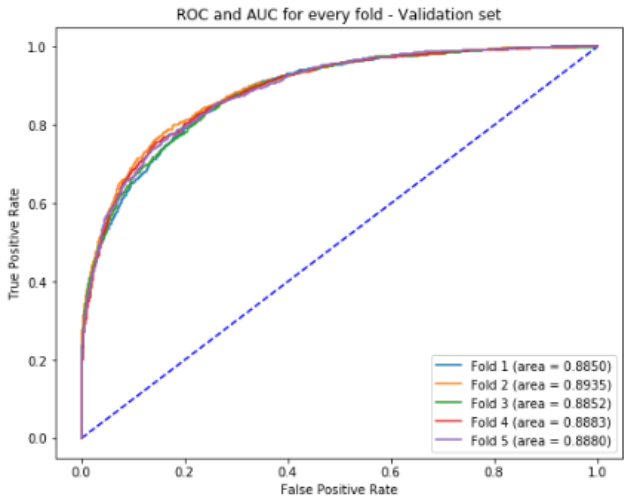


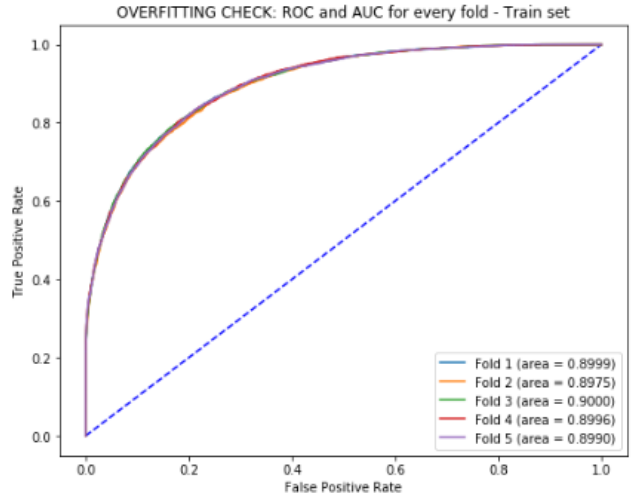

```
Mean Accuracy: 0.86479828
Mean accuracy, where classifing wrongly a 1 target is 5 times more severe than classifying wrongly a 0 target: 0.72919064
Mean AUC Test: 0.90248486
Mean AUC Train: 0.92869046
Difference between AUC: 0.02620560
```

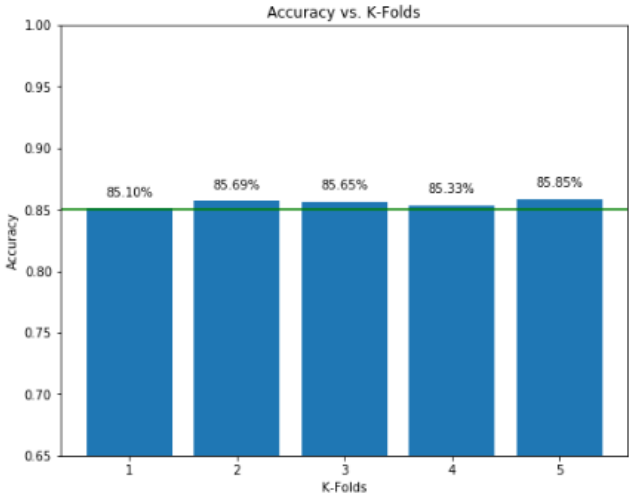
סיכום ממצאים עבור Adaptive Boosting (AdaBoost):












```
Mean Accuracy: 0.85524531
Mean accuracy, where classifing wrongly a 1 target is 5 times more severe than classifying wrongly a 0 target: 0.73998268
Mean AUC Test: 0.88720830
Mean AUC Train: 0.89919722
Difference between AUC: 0.01198892
```

סיכום ממצאים מאוחד עבור המודלים שנבחנו: