אוניברסיטת תל אביב
TEL AVIV UNIVERSITY

הפקולטה להנדסה ע״ש אייבי ואלדר פליישמן

המחלקה להנדסת תעשייה

דו"ח מסכם לפרויקט מבוא ללמידה מכונה

מרצה: ד"ר אייל קולמן

מתרגל: מר נעם תור

תקציר מנהלים:

בדו"ח זה נפרט אודות תהליכים שביצענו על מנת לסווג דגימות נתונות על פי סוגי פיצ'רים שונים.

בשלב הראשון, ביצענו אקספלורציה על הדאטה. חקרנו את התפלגות הפיצ'רים, רמת קורלציה בין פיצ'רים והן נתונים סטטיסטים.

בשלב השני, ביצענו עיבוד מקדים אודות הנתונים. מצאנו כי לא קיימים נתונים חריגים בדאטה, נרמלנו את הנתונים לפי Z-Score, הקטנו את ממדיות הבעיה בעזרת PCA וקורלציה וכן בנינו פיצ'רים חדשים (clustering) בעזרת מניפולציות מתמטיות לא ליניאריות כגון כפל, לוג ושורש.

בשלב השלישי, בנינו מודלים והרצנו עליהם את סט הנתונים המעובד. לכל מודל ביצענו למידה על חלק מסט הדגימות שניתן לנו ועל החלק שנותר ביצענו פרדיקציה כדי לבדוק את טיב המודל באמצעות cross validation. המטרה שעמדה לנגד עיננו הייתה לבדוק איזה מודל יעזור לנו לנבא את תוצאת הדגימות בצורה המדויקת ביותר.

בשלב הרביעי, הערכנו את טיב המודלים על ידי ניתוח תוצאות ה- confusion matrix והתבוננות על פערי ביצוע באמצעות סט הדגימות שחולק ל-Train ול-Validation.

לבסוף, לאחר התאמת נתוני ה-Test לקובץ ה-Train , השתמשנו במודל ANN הנבחר, על מנת לחזות את סיווג הדגימות בקובץ ה-Test.

אקספלורציה[1]:

בוצע לאחר שלב קריאת הקבצים. תחילה, בדקנו את נראות הדאטה על ידי הצגת חמשת השורות הראשונות שלה, וזאת על מנת לקבל רושם ראשוני ויזואלי של הדאטה. בנוסף, בדקנו את סיווג הדגימות ומצאנו כי רובן מסווגות כמספרים (סיווג נומרי) ומיעוטן מסווגות כאותיות (סיווג קטגוריאלי). לאור הבנת סוג הדגימות שבידנו, החלטנו להמיר חלק מהפיצ'רים הקטגוריאליים לערכים נומריים (לדוגמה: A→1,B→2 וכדומה..). כמו כן, ישנם שני פיצ'רים קטגוריאליים אותם בחרנו להוריד מהדאטה. בנוסף, ראינו כי מספר השורות אינו זהה בין הפיצ'רים (כלומר קיימים תאים בעלי הערך NaN). לאחר מכן, בדקנו את מספר הדגימות ומספר הפיצ'רים שבידנו וכן מהו אחוז הפיצ'רים הנומריים מול הקטגוריאליים. לאור המידע שבידנו, בנינו Box-Plot על מנת לראות מה הם הפיצ'רים החריגים. פיצ'ר מספר 12 התנהג כחשוד, אך לאחר בדיקה הכוללת את הרצת המודל כולו, איתו ובלעדיו, החלטנו להשאירו לאור המידע שהוא הוסיף. בשלב הבא ביצענו ניתוח סטטיסטי של הפיצ'רים (מקס', מינ', ממוצע וס"ת). לשם ויזואליזציה הדפסנו גרפים המתארים את אופן התפלגות הפיצ'רים. ניתן לראות שישנם פיצ'רים הנראים כמתפלגים בינארית, נורמלית ומעריכית. בעקבות ניתוח הנתונים, מצאנו לנכון להמיר את הדגימות בעלות הערך NaN לערך הממוצע של עמודת

[1] נספחים 1-3

הפיצ'ר הרלוונטית. כמו כן, הצגנו את ההתנהגות הקורלטיבית בין הפיצ'רים השונים בעזרת מטריצת קורלציה.

עיבוד מקדים[2]:

נתונים חריגים (Outliers):

לאור הצגת ה-Box-Plot נוכחנו לראות כי פיצ'ר 12 חריג לעומת שאר הפיצ'רים, על כן ביצענו בדיקה של הרצת המודל עם הפיצ'ר ובלעדיו. תוצאות הבדיקה הניבו כי פיצ'ר 12 תרם לדיוק התוצאות, לכן הוחלט להשאירו כחלק מהדאטה. הוצאת החריגים בוצעה לפי 3 ס"ת כך שדגימות שאינן בטווח הושמטו. כיוון שאופן התפלגות הפיצ'רים אינו גורם מכריע בדרך הוצאת החריגים, לא בדקנו בצורה ודאית כיצד הפיצ'רים מתפלגים.

בניית פיצ'רים חדשים (Clustering):

יצרנו חמישה פיצ'רים חדשים על ידי ביצוע פעולות מתמטיות לא לינאריות על הפיצ'רים הקיימים. פעולות אלו כללו בין היתר: שורש על סכום שני פיצ'רים, הכפלת שני פיצ'רים וביצוע לוג על פיצ'ר.

נרמול נתונים:

לאור ביצוע האקספלורציה הבחנו כי בכל פיצ'ר טווח הנתונים שונה, ונע בין מספרים בודדים לאלפים. ניתן לראות זאת הן בהצגת הגרפים של הפיצ'רים והן בהצגת גרף ה-Box-plot בו התקבלו ערכי פיצ'רים מגוונים. על כן, ניתן לראות שהנתונים לא מנורמלים בבסיסם, ואנו נדרשים לנרמל אותם. הסיבה לכך שהנתונים אינם מנורמלים נובעת מהעובדה כי מדובר בכמות נתונים גדולה אשר הגיעה ממקורות שונים ומדגימות שנבעו בתנאים שונים, על כן הדבר השפיע על התצפיות שנמדדו. אנו נרצה להקטין למינימום את ההבדלים הנובעים מגורמים אלו, ולצמצם הטיות לא רצויות בנתונים הן בין דגימות הפיצ'ר הספציפי והן בין פיצ'ר אחד לפיצ'ר אחר. נשים לב שישנם מודלים שמושפעים מהמרחקים בין הדגימות ,לדוגמה מודל PCA בו נשתמש בהמשך ועל כן נרצה לנרמל את הנתונים. השיטות בהן השתמשנו לנרמול הנתונים הן Min Max ו-Z-score ולאחר הרצת הנתונים והמודל עם שיטות אלו מצאנו לנכון ששיטת Z-score כשיטת נרמול הינה השיטה המניבה את ערך ה- AUC הגבוה ביותר עבור כלל המודלים.

ממדיות הבעיה:

ממדיות הבעיה הנתונה גדולה זאת מכיוון שקיימים מספר רב של פיצ'רים אשר מתארים את הנתונים ואיננו יודעים להסבירם. דרך לזיהוי ממדיות גדולה של הבעיה הינה חישוב המתאם (קורלציה) בין הפיצ'רים השונים. חישוב זה הניב כי ישנו מספר רב של פיצ'רים בעלי מתאם גבוה ועל כן ניתן להסיר את חלקם. בעיית הממדיות משפיעה עלינו בכמה מישורים. ראשית, ממדיות גדולה מוסיפה רעש לנתונים ומשפיעה לרעה על נתוני הקורלציה. שנית, כתוצאה משילוב של

[2] נספחים 4-7

ממדיות גדולה ומספר רב של דגימות, סיבוכיות הבעיה גדלה. על מנת להקטין את ממדיות הבעיה השתמשנו בשתי שיטות:

1. Correlations-Feature selection: לאחר בדיקת קורלציה בין פיצ'רים, מצאנו את זוגות פיצ'רים ביניהן קיימת התאמה הגבוהה מ-0.85. מכל זוג בחרנו פיצ'ר אחד אותו השמטנו מסט הנתונים. הסיבה להשמטה נובעת מהעובדה כי זוג הפיצ'רים מתאר בצורה דומה את הנתונים ולכן ניתן להשאיר רק פיצ'ר אחד מהשניים וכך להקטין את ממדיות הבעיה. בסה"כ השמטנו 8 פיצ'רים בדרך זו.
2. PCA- שיטה זו בוצעה לאחר השמטת הפיצ'רים מהשיטה לעיל. כעת ביכולתנו להשאיר אך ורק את הפיצ'רים שמסבירים 95% מהנתונים. על כן, בחרנו להשאיר 13 מסך הפיצ'רים אשר יספקו לנו את המידע הנדרש אודות הנתונים.

הרצת המודלים:

בנינו חמישה מודלים במטרה למצוא את המודל הטוב ביותר שיתאר את הנתונים ע"פ תוצאת AUC. תחילה, בנינו פונקציית K-Fold בעלת 10 folds שסייעה לנו בתהליך ה-cross Validation. השתמשנו בתהליך זה על מנת לאמן את המודל שלנו כל פעם על סט נתונים שונה כך שכל פעם נבחר חלק אחר מסט הנתונים שהמודל יתייחס אליו כ-Validation. בכל הרצה, התבצעה חלוקה של הנתונים ל-Train, Validation כאשר החלוקה משתנה מהרצה להרצה. כמו כן, ביצענו התאמה לנתוני ה-Train וניבוי לנתוני ה-Train, Validation לפי המודל הנבחר. לבסוף, ביצענו ממוצע על מדד ה-AUC בעבור כלל ההרצות לכל מודל נבחר. לכל מודל ביצענו הרצה בעזרת פונקציית Grid על מנת להעריך את ההיפר פרמטרים שימקסמו את ה-AUC של המודל.

<u>מודלים ראשוניים:</u>

1. KNN- בנינו פונקציית עזר על מנת להעריך את מספר השכנים האופטימלי במודל לפי כמות הפיצ'רים שנותרו. לאחר מכן, בעזרת פונקציית ה- Grid קיבלנו את ההיפר פרמטרים הבאים:

n_neighbors=$\left\lfloor\sqrt{num\ of\ rows}\right\rfloor$, weights="distance", metric="euclidean"

תוצאות המודל הניבו ערך AUC השווה ל- 0.886. גרף ה-ROC נראה חלק יחסית וללא רעש מיוחד.
2. Gaussian Naïve Bayes[3] - המודל מחשב בצורה ישירה את ההסתברות לקבלת תצפית מכל סוג. תוצאות המודל הניבו ערך AUC השווה ל- 0.840. במודל זה ניתן לראות קצת יותר רעש בגרף ה-ROC ביחס למודל ה-KNN.
3. Logistic Regression- היפר הפרמטרים שהתקבלו בעזרת פונקציית ה- Grid הינם:

penalty='l2', C=1.0, tol = 1e-06, max_iter = 100

[3] מודל נוסף בונוס

תוצאות המודל הניבו ערך AUC השווה ל- 0.883. גרף ה-ROC נראה חלק יחסית בדומה ל-KNN.

<u>מודלים מתקדמים:</u>

1. Decision Tree- היפר הפרמטרים שהתקבלו בעזרת פונקציית ה-Grid הינם:
criterion='gini',max_depth=None, min_samples_split=7,
max_features='sqrt', max_leaf_nodes=None,min_impurity_split=1e-07
תוצאות המודל הניבו ערך AUC השווה ל- 0.721. וזהו הערך הגרוע ביותר שקיבלנו בין המודלים הנבחנים. גרף ה-ROC אינו חלק וניתן לראות עיקול חד בעל רעש.
2. Neural Network- היפר הפרמטרים שהתקבלו בעזרת פונקציית ה- Grid הינם:
activation="relu", hidden_layer_sizes= (100,), alpha = 0.01,
solver = "sgd", learning_rate_init = 0.1, learning_rate = "invscaling",
power_t = 0.5, early_stopping = False, tol = 1e-4, batch_size = 10,
max_iter = 500, warm_start = False, random_state = 42
תוצאות המודל הניבו ערך AUC השווה ל- 0.893, וזוהי התוצאה הטובה ביותר מבין כלל המודלים הנבחנים. גרף ה-ROC נראה עקבי וחלק יחסית לעומת שאר המודלים.

<u>הערכת המודלים[4]:</u>

המודל בעל ערך ה-AUC המקסימלי הינו מודל רשת הנוירונים (ANN) ולכן הוא המודל הנבחר לביצוע החיזוי בסט הנתונים של קובץ ה-Test.

<u>Confusion Matrix:</u> בוצע על מודל ה-ANN, קיבלנו את המטריצה הבאה:

	Actual 1	Actual 0
Predicted 1	97	8
Predicted 0	389	1645

מטרת המטריצה היא להעריך את איכות החיזוי של המודל הנבחר לאור העיבוד המקדים שבוצע.

ניתן לראות כי חזינו ערך "0" גבוה מערך "1" (97+8<1645+389) דבר שתואם את גרף העוגה מחלק א' (אקספלורציה), המצביע על יחס החלוקה של הלייבל.

כמו כן, ניתן לראות כי סך הטעויות (389+8) נמוך משמעותית מסך החיזויים הנכונים (97+1645).

חישובים אודות המטריצה:

- אחוז המתאר את הדגימות אשר נחזו בצורה המתאימה לסיווג שלהן:
$$accuracy = \frac{97+1645}{97+8+389+1645} = 81.44\%$$

[4] נספחים 8-13

- אחוז המתאר את הדגימות שהן באמת "1" מסך כל הדגימות שנחזו "1":
$$Precision = \frac{97}{97+8} = 92.38\%$$
- אחוז המתאר את הדגימות שנחזו "1" מסך כל הדגימות שהן באמת "1" :
$$Sensitivity = \frac{97}{97+389} = 19.95\%$$
- אחוז המתאר את הדגימות שנחזו "0" מסך כל הדגימות שהן באמת "0" :
$$Specificity = \frac{1645}{1645+8} = 99.52\%$$

Overfitting: על מנת לבדוק האם המודל שלנו הוא overfitted חישבנו את ערך ה-AUC של ה-Train ושל ה-Validation וביצענו השוואה ביניהם. קיבלנו כי ההפרש ביניהם זניח ועומד על כ-0.0034, מה שמעיד על כך שהמודל שלנו אינו overfitted כיוון שערכי ה-AUC קרובים.

AUC value - Validation	0.8939
AUC value - Train	0.8973

עיבוד וחיזוי ה-Test:

ביצענו על קובץ ה-Test את העיבוד המקדים הכולל: מילוי תאי סט הנתונים הריקים מולאו בעזרת ממוצע של הפיצ'ר הרלוונטי. בוצעה הוספת חמישה פיצ'רים לסט הנתונים, נורמליזציה לפי Z-Score, הוצאת הפיצ'רים בהתאם לצורה שבוצעה בקורלציה וב-PCA (אותן עמודות שהוסרו בקובץ ה-Train הוסרו בקובץ ה-Test). יש לציין כי לא נמחקו רשומות חריגות (שורות מסט הנתונים של ה-Test) על מנת שנוכל לסווג את כלל הדגימות שקיבלנו. לאחר העיבוד המקדים, בוצע חיזוי לייבל על ידי המודל הנבחר (ANN) והודפס פלט החיזוי אל קובץ האקסל.

סיכום:

ביצענו ניתוח על סט נתונים שקיבלנו, הפעלנו מודלים שונים לצורך חיזוי שעזר לנו למצוא את המודל הטוב ביותר, ANN. מודל זה אפשר לנו לחזות בצורה המיטבית את ה-לייבלים של הדגימות השונות בקובץ ה-Test.

בשלב הרצת המודלים ביצענו הערכת טיב מודל לכלל המודלים שבחנו על מנת למצוא את המודל שחוזה בצורה האפקטיבית ביותר.

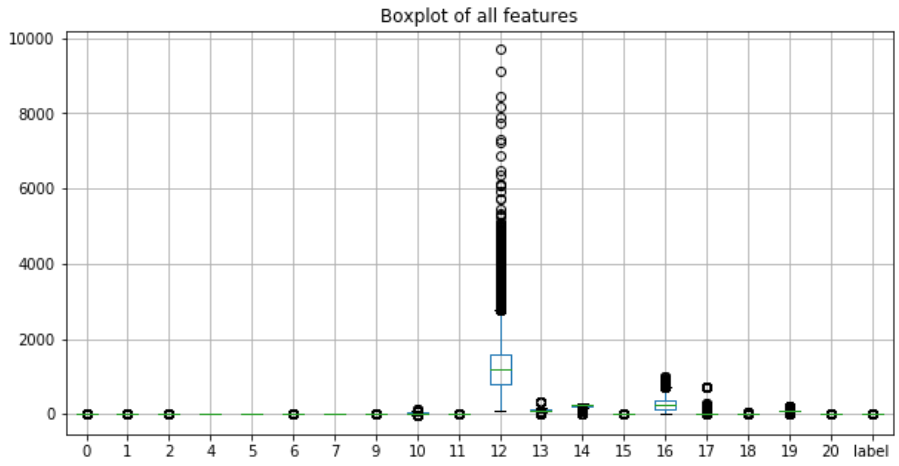
אנו מעריכות כי תוצאות מדד ה-AUC שהתקבלו היו טובות למדי לאור ביצוע עיבוד מקדים בצורה שהולמת את סט הנתונים שקיבלנו כבר בשלב הראשון.

בבדיקת ה-Overfitting הניבה כי המודל אינו מותאם יתר על המידה ל-Train בסט הנתונים.

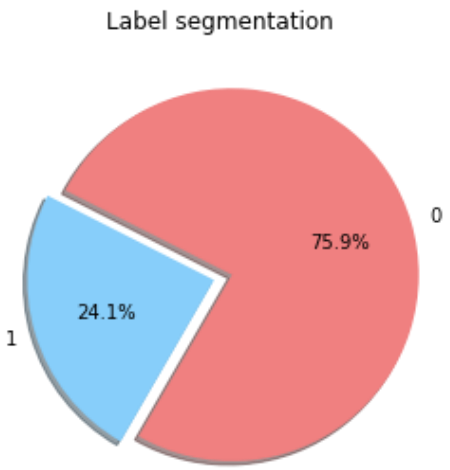
נספחים:

נספח 1 - גרפים המתארים את התפלגות הדגימות בכל פיצ'ר:

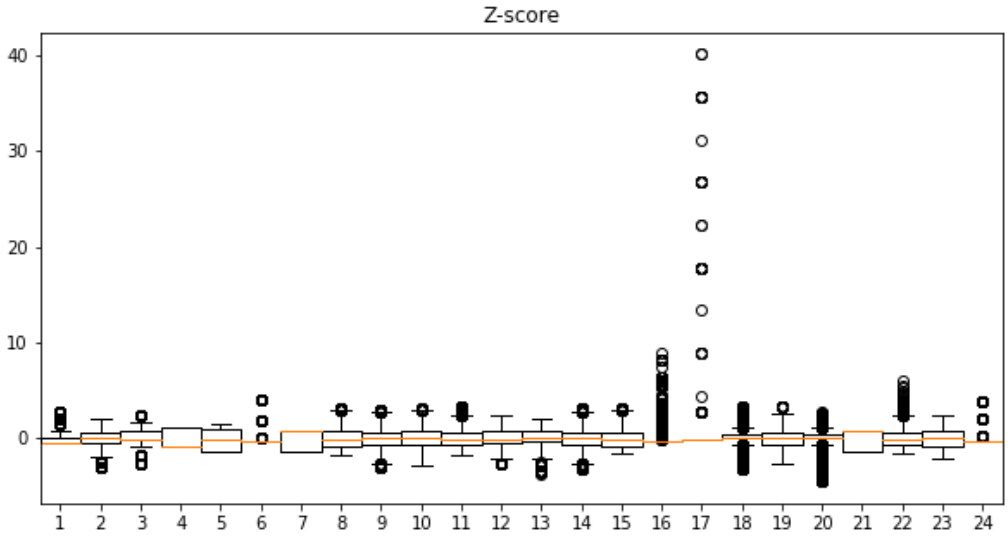
נספח 2 - גרף Box-Plot של הפיצ'רים המקוריים:



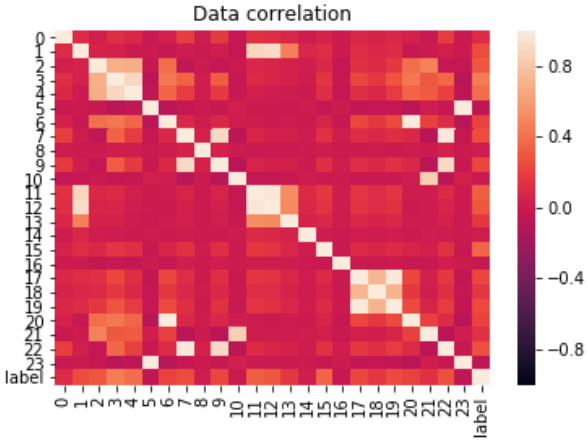
נספח 3 - גרף המתאר את התפלגות הלייבל:



נספח 4 - גרף Box-Plot המתאר את התפלגות הפיצ'רים לאחר נרמול לפי Z-Score:

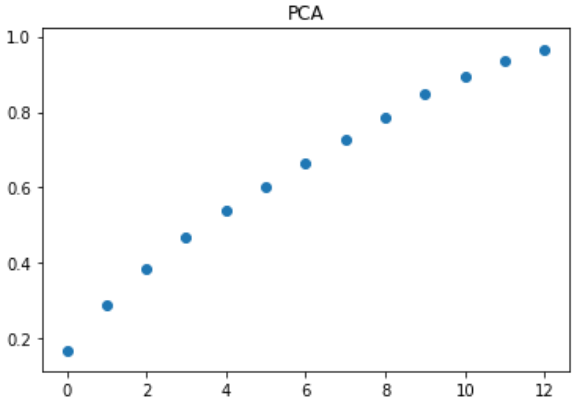


נספח 5 - גרף המתאר את הקורלציה לאחר הוספת פיצ'רים חדשים:

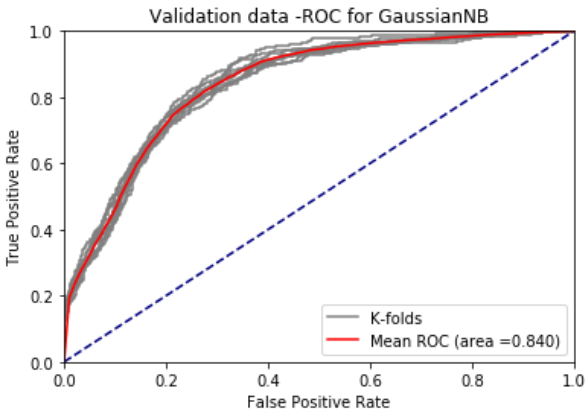


נספח 6 - רשימת הפיצ'רים שהסרנו בעקבות הקורלציה: 1, 4, 5, 6, 9, 11, 17, 22.

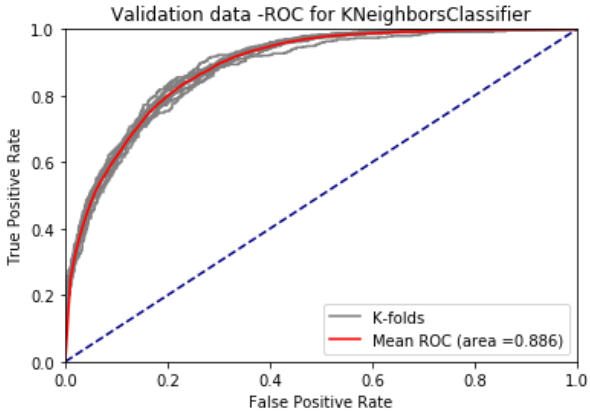
נספח 7 - גרף המתאר את הפיצ'רים שנשארו לאחר ה-PCA:



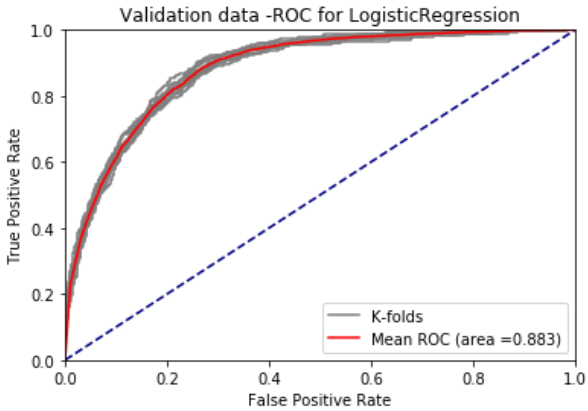
נספח 8 - גרף ROC עבור מודל Gaussian Naïve Bayes:



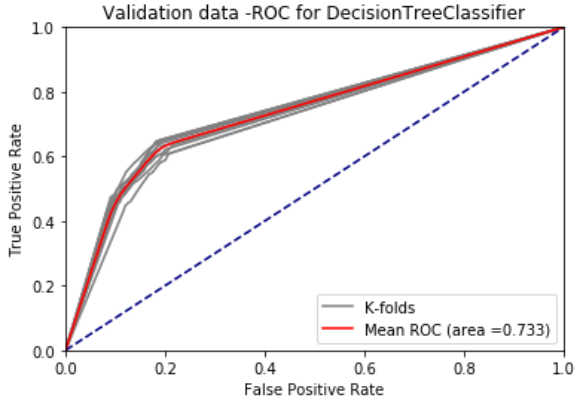
נספח 9 - גרף ROC עבור מודל KNN:



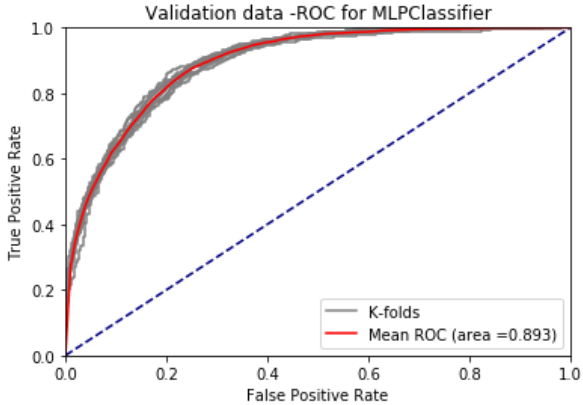
נספח 10 - גרף ROC עבור מודל Logistic Regression:



נספח 11 - גרף ROC עבור מודל Decision Tree:



נספח 12 - גרף ROC עבור מודל Neural Network:



נספח 13 - גרף ROC עבור ה-Train ועבור ה-Validation: