הפקולטה להנדסה
ע"ש איבי ואלדר פליישמן
אוניברסיטת תל אביב

אוניברסיטת תל אביב
TEL AVIV UNIVERSITY

מבוא ללמידת מכונה

פרויקט מסכם

קבוצה 35

יהונתן לייטנר - 312474646

נתנאל שאשא - 313298473

בראל לנציאנו - 206082794

תקציר מנהלים:

בפרויקט זה נעסוק בבעיית סיווג בינארי , בה יש לסווג רשומות לשתי קטגוריות על סמך הפיצ'רים הקיימים בסט הנתונים.
בחלק הראשון של הפרויקט נבצע אקספלורציה, שבה נחקור את הנתונים.
נציג נתונים סטטיסטיים על הפיצ'רים, את אופי התפלגותם ונחשב קורלציות בין הפיצ'רים השונים.
בחלק השני של הפרויקט נבצע עיבוד מקדים. נזהה נתונים חסרים ונסיר נתונים חריגים, כמו כן, נקטין את ממדיות הבעיה באמצעות שימוש בניתוח גורמים ראשיים (PCA).
בחלקים הבאים, נחיל על סט הנתונים שבידנו ארבעה מודלי חיזוי: Gaussian Naïve Bayes, KNN, Decision Tree Artificial Neural Network.
נבצע הערכות למודלים שהחלנו על פי מספר מדדים ונבחן את פערי הביצוע.

בחלקו האחרון של הפרויקט, נבחן את מודל בעל מדד ה- AUC הגבוה ביותר, ונבצע פרדיקציה על נתונים חדשים אותם המודל לא פגש.

חלק ראשון: Exploration of data

1. *התמודדות עם חוסרים*: תחילה בדקנו עבור כל עמודה מהו אחוז הנתונים החסרים. גילינו 80% חוסרים בעמודה 19, ו20% חוסרים בעמודה 14. מלבדם בכל שאר הפיצ'רים אחוז החוסרים היה קטן מאוד.
לאחר מכן, ביצענו ויזואליזציה של "מידע חסר" – הוויזואליזציה של המידע החסר מסייעת לנו להבין האם יש קורלציה בין מידע חסר של משתנים שונים והאם יש חזרתיות מסוימת באופן בו המידע חסר. זיהינו כי בסט הנתונים שלנו אין חזרתיות מיוחדת. בהמשך נשקול להשלים או להסיר את העמודות עם אחוז גבוה של חוסרים תוך הבנה שאנו מסתכנים בעיוות של הנתונים והכנסת רעש שייפגע בקבלת מסקנות מהימנות ויחליש את תהליך הלימדה אליו אנו מקוונים.
בPre-processing. ,עבור השלמה הנתונים החסרים ,נשלים את המידע באופן הבא – משתנה קטגוריאלי עם השכיח, ומשתנה נומרי עם הממוצע.
2. התפלגות הנתונים: כדי ללמוד על התפלגות הנתונים ביצענו היסטוגרמות למשתנים הקטגוריאליים ולמשתנים הנומריים. בדקנו האם קיימת התפלגות ידועה למשתנים בדגש על התפלגות נורמלית, גילנו שפיצ'רים 3,5,20 אכן ממתפלגים נורמלית. מההיסטוגרמות של הפיצ'רים הקטגוריאליים ראינו מהן הרמות השונות של כל פיצ'ר.
3. קורלציה: ביצענו בדיקת קורלציה ע"י "heat map" , ראינו שיש שני זוגות משתנים בעלי קורלציה גבוה – 5 ו 17 , 7 ו 9 . עבורם הכנו גרף כדי לראות בצורה מיטבית את סוג הקורלציה. זיהינו קורלציה חיובית בין שני זוגות המשתנים .

חלק שני: Pre processing

1. *זיהוי חוסרים נוספים:* החלפנו בכל סט הנתונים את המקומות בהן הופיע ערך "unknown" או "?" בערך של "None" כיוון שמודבר בעינינו בחוסרים. בעמודה 2 השמטנו את הסיומת "d" המופיעה בכל הערכים והמרנו את ערכי התא לנומריים.

2. *השמטת עמודות*: בשלב זה ביצענו השמטה של עמודות: השמטנו את עמודה "19" בעקבות אחוז החוסרים הגבוה, וכן את עמודות 17 ו9 כיוון שכפי שראינו בחלק הקודם ישנה קורלציה גבוהה עם עמודות אחרות ולכן אנו מניחים שהן לא מוסיפות מידע חדש. השמטנו גם את עמודה 15 כיוון שראינו בהתפלגות כי כמעט כל הנתונים מסווגים כ-0. כך שעמודה זו לא מוסיפה מידע חדש.

3. *הוצאת חריגים:* outliers הינן תצפיות שקיבלו ערכים קיצוניים החורגים מהערכים שתועדו בשאר התצפיות, ועשויות להצביע על שונות חריגה, טעות בשלב הזנת המדידות וכן על בעיות בתכנון המדידות, תקלה במדידות, ועוד. כדי לגלות מיהם הנתונים החריגים ביצענו "Box plot" . מהגרף גילינו שיש מספר נתונים חריגים גבוה בפיצ'ר 4. לאחר מכן החלטנו להסיר מהעמודות הנומריות את כל את התאים החורגים בלפחות 3 סטיות תקן מהממוצע. בחרנו ב-3 סטיות תקן כיוון שזהו המספר שנחשב כסטנדרטי על פי

המקורות שבידנו (נלמד בתרגול). נדגיש שכיוון ואנו לא יודעים מאיזה נושא לקוחים הפיצ'רים, לא יכולנו להגדיר "חריגות" באופן ספציפי.

4. *Scaling:* בקובץ ישנם נתונים רבים בסדרי גודל שונים ובמגוון יחידות מידה שלא ידועות לנו. טווח הערכים שאינו מנורמל עלול להיות רחב מאוד. מצב זה יכול להשפיע לרעה על פעולות של פונקציות האקטיבציה השונות שאנו מפעילים באלגוריתמי החיזוי. בחרנו להשתמש בשיטת standardization לצורך ביצוע נרמול של הנתונים הנומריים. משמעות הנרמול באופן זה היא יצירת התפלגות עם ממוצע 0 וסטיית תקן 1.בשאר סוגי הנתונים נטפל באופן אחר בהמשך .
5. *השלמת חוסרים:* כיוון שלא זהינו באקספלורציה התפלגויות מוכרות מהן ניתן להשלים את הנתונים, השלמנו את הנתונים החסרים עם השכיח והממוצע כפי שהוסבר בחלק הקודם.
6. *נתונים בינאריים:* בעמודות 1 ו12 זיהנו נתונים בינאריים (y/n ,a/b) , לכן המרנו אותם ל1 ו 0 באופן הבא : {a: 1 , b: 0} {n: 1 , y: 0}.
7. multi variant columns: השתמשנו ב"one hot encoding" כדי לפרק את הרמות השונות של כל עמודה קטגוריאלית באופן שממיר כל רמה לווקטור בינארי (dummy variables). לצורך כך השתמשנו בפונקציית get dummies והפרדנו את עמודות 6,10,13,16,18 על פי מספר הרמות בכל פיצ'ר. ולאחר מכן הסרנו את העמודות שפיצלנו כמובן.
8. *ממדיות הבעיה:* בתום התהליכים הנ"ל קיבלנו סט נתונים כולל המכיל 61 פיצ'רים. כדי לבחון את הממדיות של הבעיה בנינו מפת קורלציות לכל סט הנתונים, תוך הבנה שאם ישנה קורלציה גבוה להרבה זוגות נתונים זה יכול להעיד על בעיית ממדיות. לאחר בחינת מפת הקורלציה גילינו שאין קורלציה גבוהה בין המשתנים. בנוסף, בדקנו האם מתקיים יחס של n פיצ'רים ל-n^2 רשומות. לנו יש 21,931 רשומות מספר גדול בהרבה מn^2 . לסיכום, אנו מניחים שממדיות הבעיה תקינה.
9. *הורדת ממדיות PCA* : למרות שהנחנו שלא קיימת בעיית ממדיות בחרנו לבצע תהליך PCA להקטנת ממדיות הבעיה. לאחר ניסוי וטעיה הגענו ל-20 פיצ'רים המסבירים כ- 94% מהשונות.

חלק שלישי : הרצת המודלים

- טרום הרצת המודלים ביצענו תהליך של GRID SERCH במטרה לגלות מהם ההיפר-פרמטרים שימקסמו את ההצלחה של כל מודל. בתהליך זה מבוצעת סריקה של הערכים האפשריים על פרמטרים שונים שמוגדרים לפונקציה – ומוחזרים הערכים האופטימליים עבור המודל. את תהליך זה ביצענו לכל המודלים שיש להם היפר-פרמטרים.

מודלים בסיסיים:

1. **Gaussian Naive Bayes Classifier:**

הסבר: המודל מתבסס על חוק בייס ועל ההנחה הנאיבית שאין תלות בין המשתנים מאפיינים את תכונות סט הנתונים. לא היה צורך במימוש פונקציית Grid Search עבור מודל זה כיוון שהוא חסר היפר-פרמטרים. בנוסף, המודל מניח התפלגות נורמלית של הנתונים – הנחה שלא מתקיימת עבור רוב הפיצ'רים שלנו ולכן נסיק שמודל זה לא יניב תוצאות טובות.

2. **KNN:**

הסבר: אלגוריתם KNN משייך כל תצפית לקבוצה המתאימה על פי קבוצת החלוקה הנפוצה ביותר מבין K השכנים הקרובים ביותר לתצפית. לכן פעילות המודל תלויה גם בבחירת K.

Grid search function: מימשנו את הפונקציה למציאת ההיפר פרמטרים האופטימליים, וקיבלנו:
K =140, wights =distance, metric = Euclidean, algorithm = brute.

כלומר, נבחן על 140 השכנים הקרובים, נשתמש בחישוב מרחק אוקלידי נבצע אופטימיזציה באמצעות אלגוריתם BRUTE force הבסיסי שרץ בזמן של $O[DN^2]$ עבור D משתנים וN רשומות.

מודלים מורכבים:

3. **<u>רשת נוירונים NN:</u>**

הסבר: רשת נוירונים היא מודל מתמטי שפותח בהשראת תהליכים מוחיים המתרחשים ברשת עצבית טבעית ומשמש במסגרת למידת מכונה.

Grid search function: מימשנו את הפונקציה למציאת ההיפר פרמטרים האופטימליים, וקיבלנו:

פונקציית אקטיבציה – רגרסיה לוגיסטית, שכבה חבויה בגודל 50,50 . קצב אימון = 0.01. מספר איטראציות מקסימלי – 1500 . גודל batch=50. בנוסף התקבל מדד learning rate השווה ל"constant " ,מדד זה קובע את מידת התאמת המשקולות ברשת.

4. **<u>Decision Tree:</u>**

הסבר: למידה בעץ החלטה משתמשת בעץ החלטה כמודל חיזוי כדי לעבור מתצפיות על פריט המיוצג בענפים, למסקנות לגבי ערך היעד של הפריט-המיוצג בעלים.

Grid search function: Criterion = entropy, מדד זה קובע את אופן הפיצול.
Max depth = 6, קיבלנו עומק עץ של 6.
min samples split = 2, מספר הדגימות המינימלי הנדרש לפיצול – ערך קטן עלול להוביל ל-OVERFITTING
max features = 'sqrt' , מספר הפיצ'רים המקסימלי לבדיקה בזמן כל סיווג, במקרה שלנו שורש ריבועי של מספר המשתנים.
min impurity decrease = 1e-7, מדד זה מעריך את הסף לעצירה מוקדמת של העץ כך שצומת יפוצל אם נקבל ערך הגבוה מהרף הנקבע.

<u>חלק רביעי: הערכת המודלים</u>

כעת נרצה לבחור את המודל המתאים ביותר לסט הנתונים שלנו.
ראשית נבנה Confusion Matrix עבור אחד המודלים שבחנו. המודל שבחרנו להציג עבורו הוא מודל NN. עבורו התקבלה המטריצה הבאה:

$$\begin{bmatrix} 177 & 322 \\ 5102 & 762 \end{bmatrix}$$

כך ש: True positive = 322
False positive = 177
False negative = 762
True negative =5102

- חלוקת סט ה- train באמצעות K-fold : השתמשנו בערך הסטנדרטי 5. חילקנו את הנתונים ל-5 -k folds כך שנוכל לאמן כל מודל על 5 חלקים שונים מסט הנתונים ולבחון עבור 5 חלקים שונים של validation. נשים לב שבחירת ערך נמוך של K מקטינה את השונות ויכולה להוביל ל BIAS גדול יותר ומאידך בחירה של K גבוה מייצרת עומס חישובי אך מגדילה את השונות ומקטינה את ה-BIAS.
את החלוקה ביצענו באמצעות פונקציה שבנינו " KFold_func" ,מלבד חלוקת סט הנתונים, הפונקציה מייצרת עבור כל מודל גרף ROC וכן מחשבת את השטח תחת העקומה – AUC. את חישוב מדד ה-AUC הממוצע חישבנו גם עבור הבחינה על סט ה- train וגם עבור סט ה-Validation.
- כדי לבחור את המודל המתאים ביותר לסט הנתונים חיפשנו את המודל שעוברו מדד ה- AUC הוא הטוב ביותר. להלן התוצאות:

	Average Auc train	Average Auc validation
Gaussian Naive Bayes Classifier	0.82	0.82
KNN	1	0.84
רשת נוירונים NN	0.85	0.85
Decision Tree	0.70	0.69

- לאור התוצאות הנ"ל בחרנו במודל "רשת נוירונים" לביצוע הפרדיקציה שלנו. בחרנו במודל זה כל על פי מדד ה- Average Auc validation.

בדיקת OVERFITTING: כדי לבחון זאת, ביצענו השוואה בין מדד ה-AUC המתקבל על ה validation לזה המתקבל מהרצת המודל על ה-train. לאחר בחינה במקורות מידע שונים זיהינו כי עבור הבדל הגדול מ 3% בין התוצאות המתקבלות נחשוד ב-OVERFITTING. בנינו לולאה הבוחנת את הבדלים אלו עבור על מודל ומודיעה האם מתקיים חשד. לשמחתנו, המודל הנבחר NN לא נחשד ב-overfitting. עם זאת נשים לב כי המודלים, KNN , DT חשודים כבאלו.

כדי להגדיל את יכולת ההכללה של המודל השתמשנו בשיטת K-fold כפי שציינו קודם. בנוסף, עבור המודל הנבחר השתמשנו בשיטת "עצירה מוקדמת". שיטה זו תוחמת את מספר האיטראציות המקסימלי, כף שניתן לעצור את המודל טרם הגעה למצב של OVERFITTING.

<u>חלק חמישי : פרדיקציה</u>

בשלב זה ביצענו חיזוי בעזרת המודל הנבחר- ANN , על סט ה- test .
יצרנו קובץ CSV חדש אשר כולל את תחזיות ההסתברות לקלסיפיקציה (1).

<u>סיכום:</u>
בפרויקט זה עסקנו בניתוח בעיית Binary Classification.

קיבלנו סט נתונים המכיל כ-20,000 רשומות על 21 פיצ'רים. עבור סט נתונים זה קיבלנו גם את הלייבל כך שיכולנו להשתמש בנתונים לצורך אימון מודלים שונים.

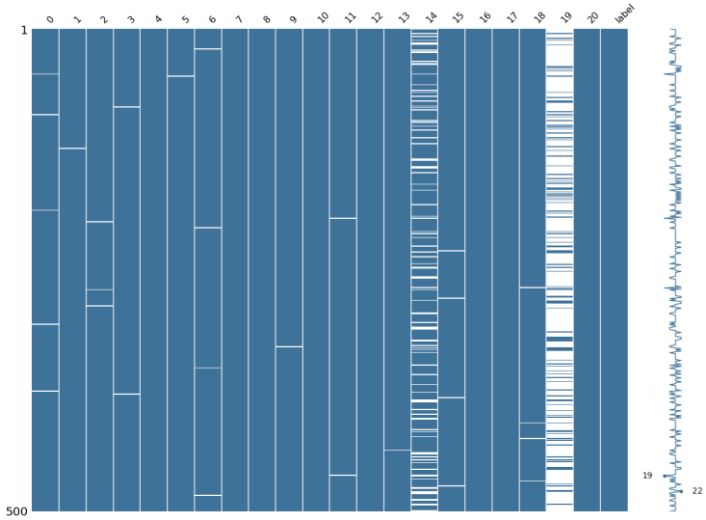
תחילה, ביצענו אקספלורציה במטרה להבין את התפלגות הנתונים השונים, וכן לחפש מידע שאינו גלוי במבט ראשוני. בהמשך החלנו עיבוד מקדים על הנתונים כך שיתאימו להרצת המודלים השונים. בתהליך זה השמטנו עמודות בעלות קורלציה גבוהה/מספר גדול של נתונים חסרים, הסרנו נתונים שהיו חריגים בעיננו, ביצענו נרמול, וכן השתמשנו בתהליך PCA להקטנת הממדיות של הבעיה. כל אלו אפשרו את תהליך הרצת המודלים.

לאחר מכן, בחרנו 4 מודלים ללמידת מכונה. שניים בסיסיים: **KNN**,**Gaussian Naive Bayes Classifier** ושניים מורכבים: **רשת נוירונים NN**,**Decision Tree** .טרם תהליך הלמידה של כל המודלים השתמשנו בפונקציית GRID SEARCH למציאת ההיפר-פרמטרים הטובים ביותר הידועים לנו. ולאחר מכן בעת תהליך הלמידה נעזרנו בתהליך ה-K-fold לצמצום יצירת OVER-fitting. לאחר שאימנו את כל המודלים השתמשנו במדד הAUC לבחירת המודל הטוב ביותר. מודל שהשיג את התוצאות הגבוהות ביותר עבור מדד זה נבחר על ידנו כמודל הטוב ביותר לצורך ביצוע הפרדיקציה. נדגיש שאת ערך ה-AUC קיבלנו לאחר בניית גרף ROC וחישוב השטח תחת עקומה זו. המודל הטוב ביותר שהתקבל הוא מודל ANN עבורו התקבל מדד AUC השווה ל-0.85.

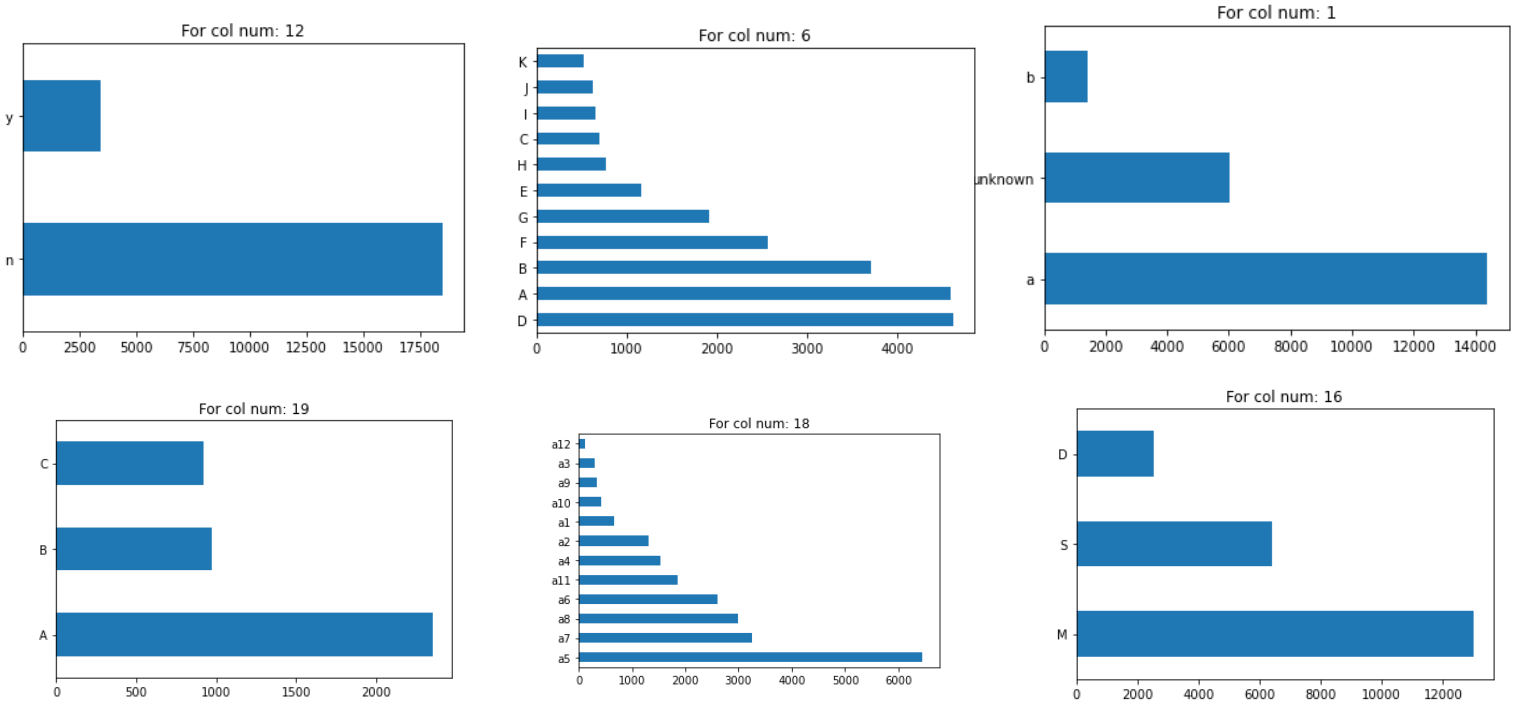
כדאי לבחון את אמינות המודל, בחנו האם המודל מבצע over-fitting . עשינו זאת באמצעות בחינת הפער בין מדד ה-AUC המתקבל עבור סט ה- TEST לעומת מדד זה עבור סט ה- VALIDATION. קיבלנו שעבור המודל הנבחר אין חשד ל-overfitting. אולם מתקיים חשד כזה על מודל KNN. לבסוף ביצענו חיזוי לסט ה-TEST ואת התוצאות ייצאנו לקובץ אקסל המצורף למסמך זה.

נספחים:

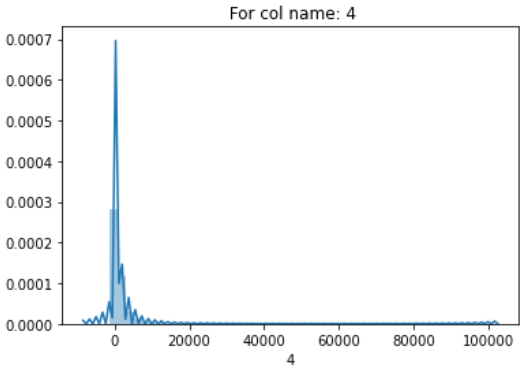
1. בדיקת חוסרים:

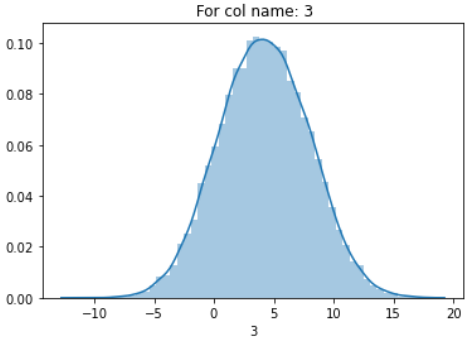


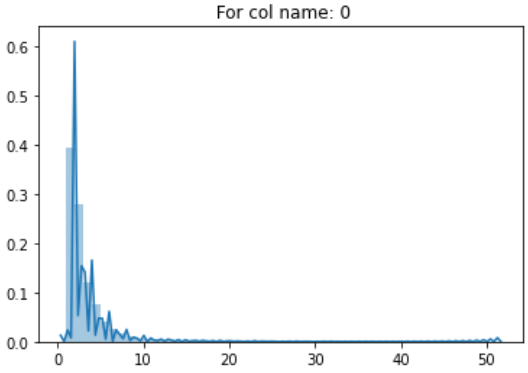
2. התפלגויות קטגוריאליות:

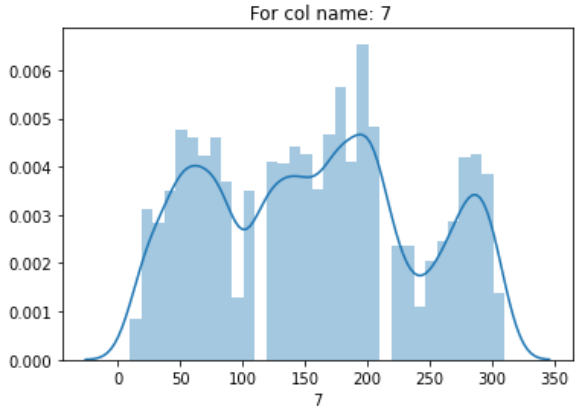


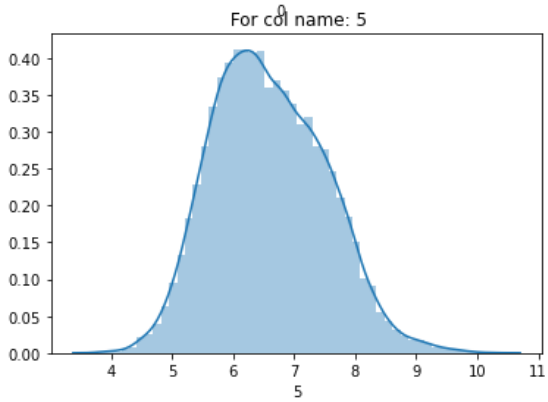
3. התפלגות נומריות:

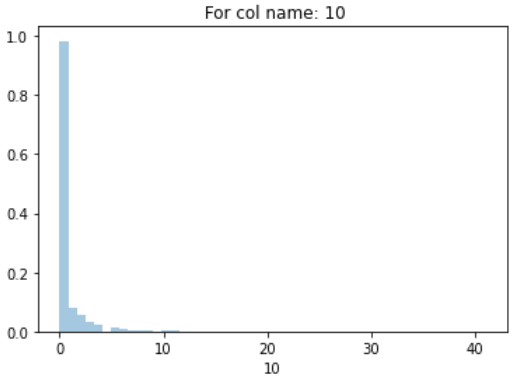


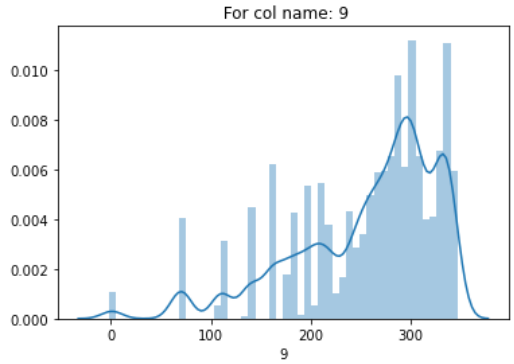


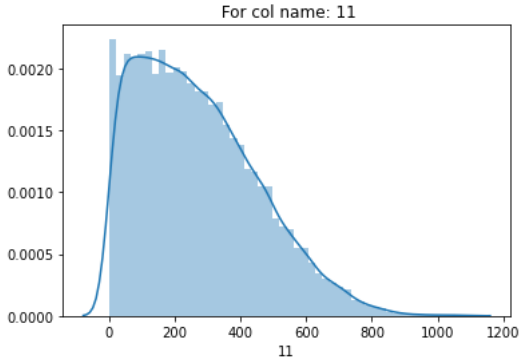


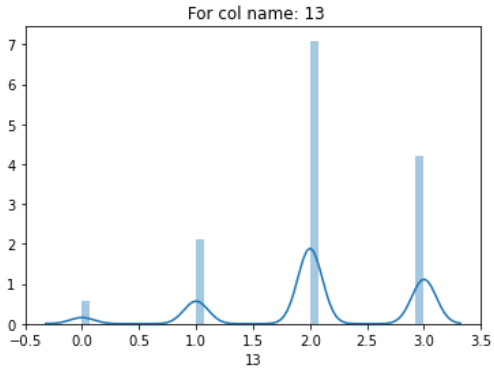


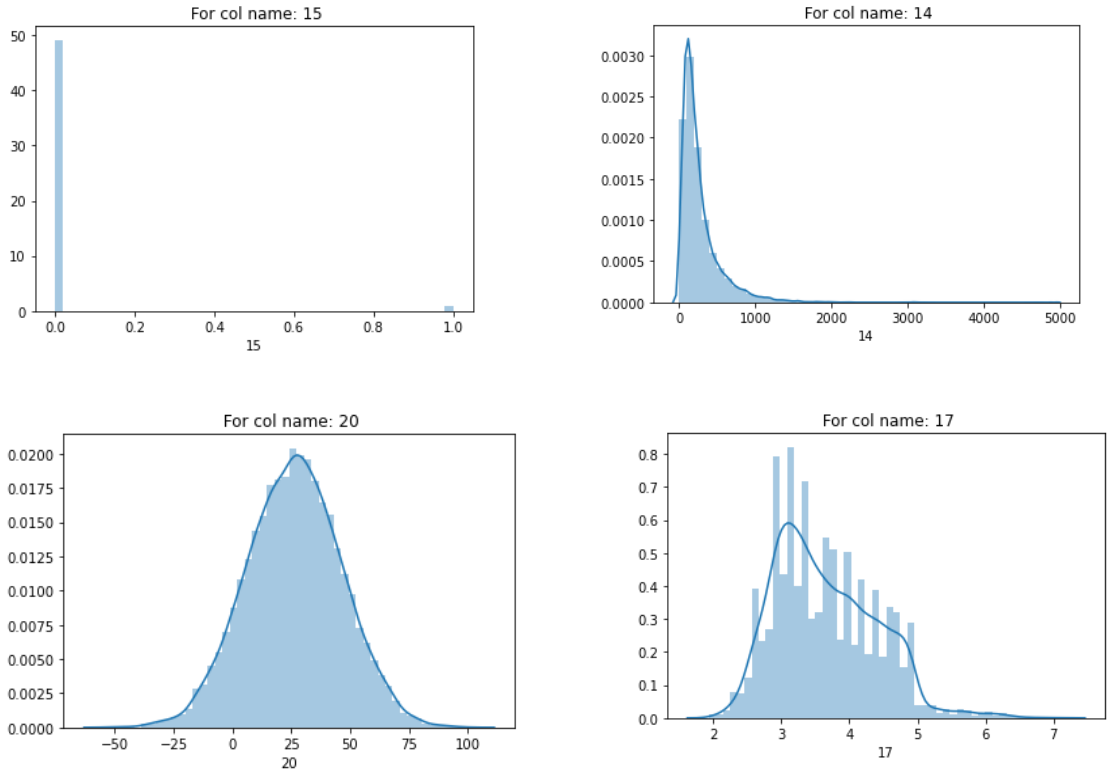




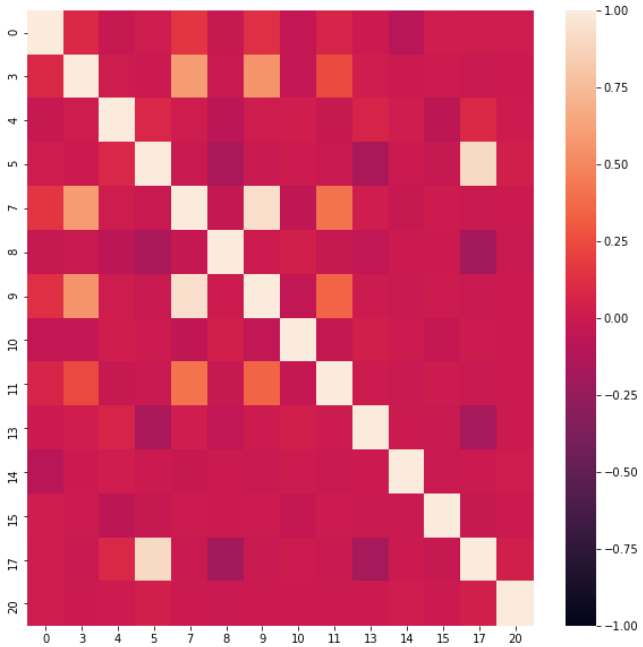




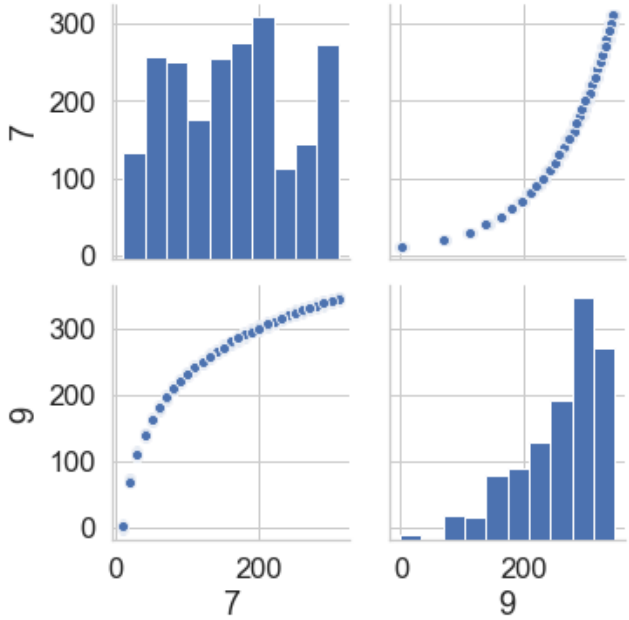


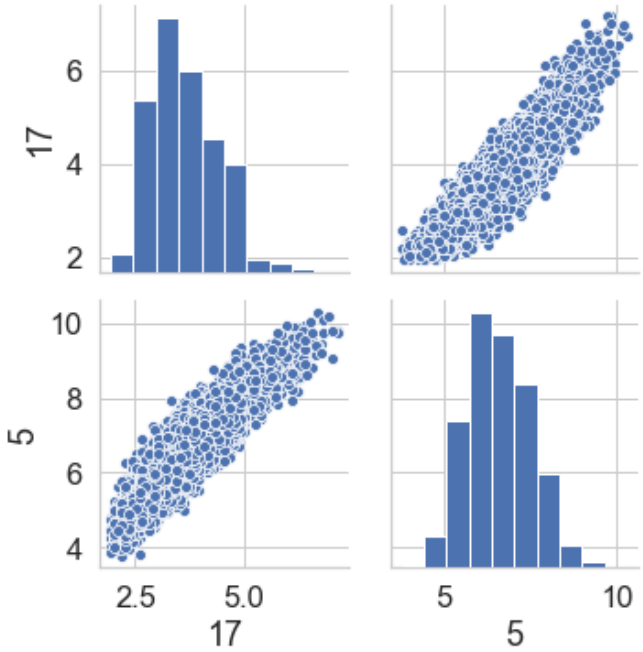


4. מפת קורלציה:

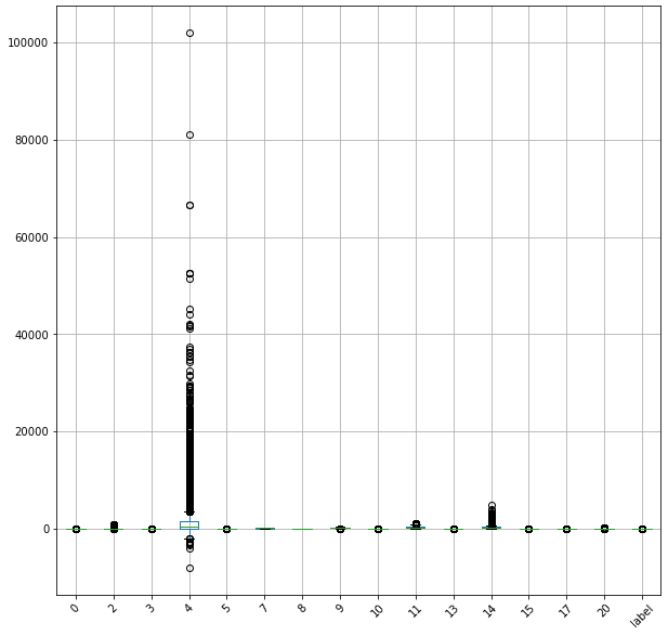


5. קורלציה בין זוגות משתנים:



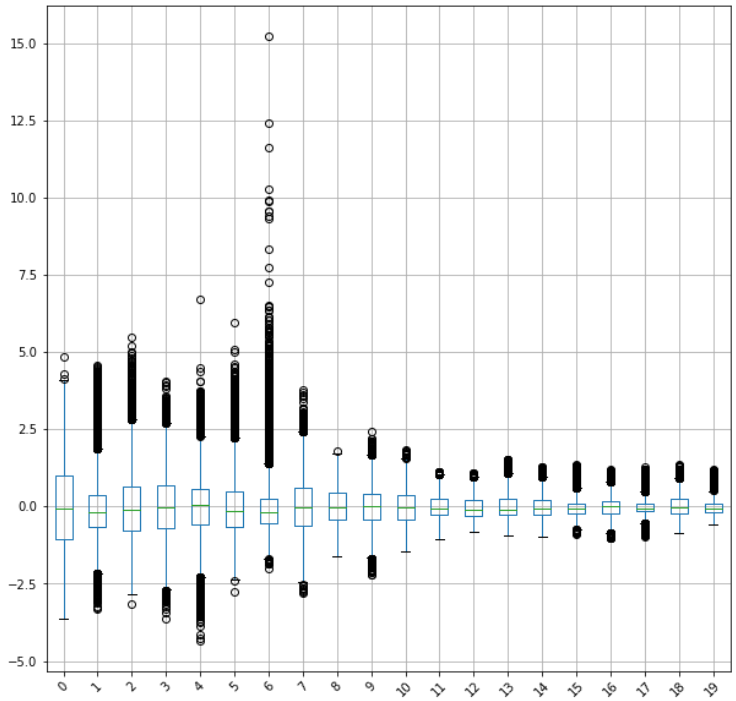


6. Box plot למציאת חריגים:



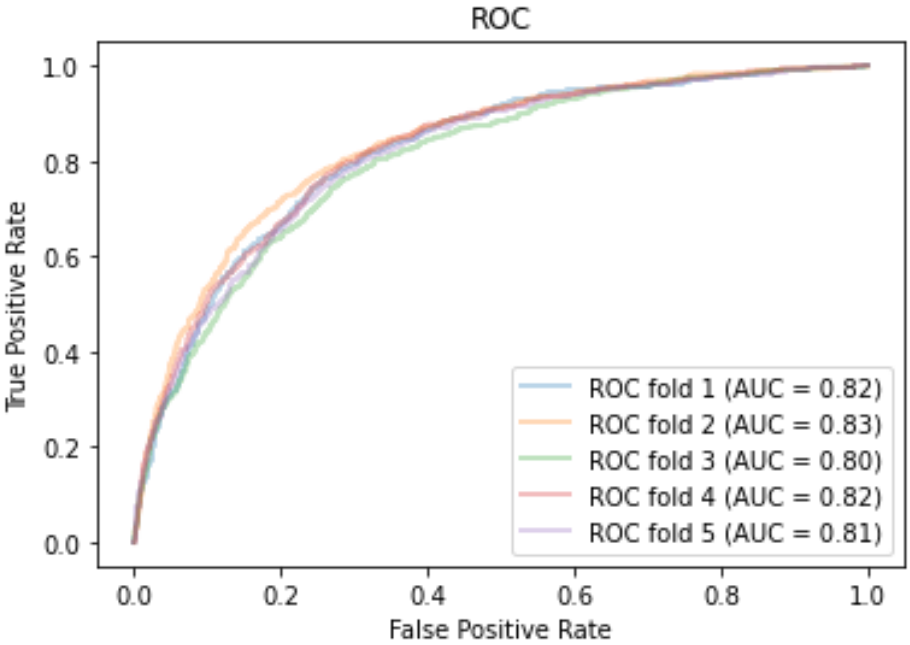
7.. box plot אחרי הוצאת חריגות ונרמול

.

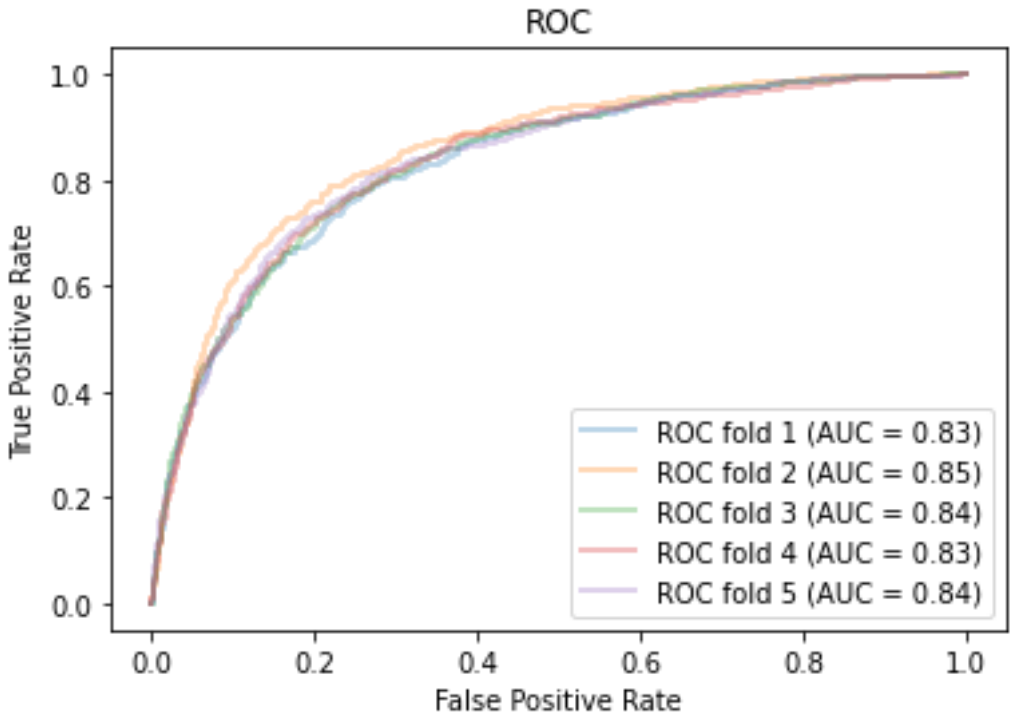


8. גרפי ROC עבור כל אחד מהמודלים:

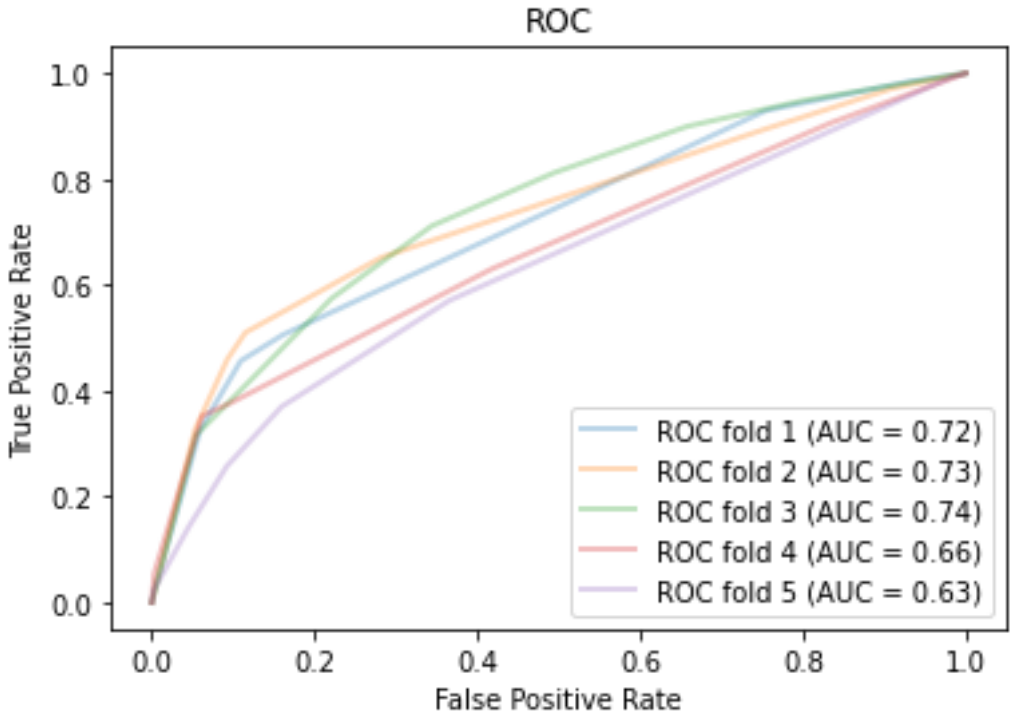
8.1 Gaussian



K NeighborsClassifier 8.2



MLP Classifier 8.3



DT CLASSIFIER 8.4