קורס לימוד מכונה

פרויקט חלק ב

**תמונה שמכילה טקסט, אלקטרוניקה

התיאור נוצר באופן אוטומטי**

**מגישים**

|  |  |
| --- | --- |
| אופיר אזולאי | 319145959 |
| לירון מרינברג | 316266501 |

Smoking Data

תוכן עניינים

**הקדמה**---------------------------------------------------------------------------------------------------------------**2**

Decision Trees---------------------------------------------------------------------------------------------------2

**Neural network**---------------------------------------------------------------------------------------------------**7**

K Means-----------------------------------------------------------------------------------------------------------9

**השוואה בין מודלים**------------------------------------------------------------------------------------------------**10**

**נספחים**-------------------------------------------------------------------------------------------------------------**12**

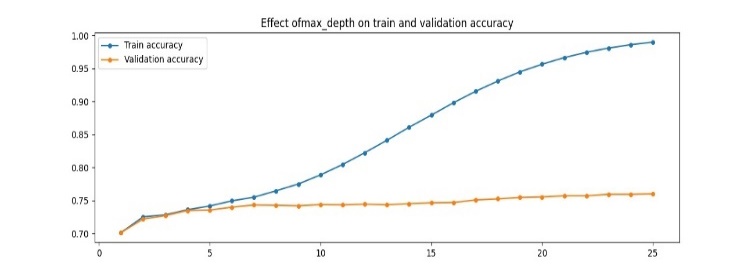
**הקדמה:** לפני שנאמן, נבחר ונבחן מודלי סיווג שונים על סט הנתונים שלנו נזכיר כי במסגרת חלק א' של הפרויקט ביצענו עיבוד מקדים לנתונים שכלל השלמת חוסרים והמרת נתונים לא הגיוניים לפי כללי החלטה שקיבלנו בהתאם לכל מאפיין לפי התפלגותו האינדיבידואלית או לפי להתפלגותו ביחס למאפיינים אחרים. כמו כן, במסגרת חלק א' ביצענו דיסקרטיזציה של משתנים אך כנקודת מוצא לבחינת המודלים בחלק זה, נבחר להסתכל על המשתנים בצורתם המקורית. כמו כן על אף הקשרים שמצאנו בחלק א' בין המשתנים שגרמו לנו להסיר משתנים מסוימים, בחרנו לבחון בנקודת המוצא את כלל המשתנים למודלים, ובהמשך לבחון הוצאה שלהם. כעת, בחלק זה של הפרויקט, על מנת ליישם מודלי סיווג על סט הנתונים, חילקנו את סט הנתונים באמצעות פונקציית train\_test\_split המסייעת ליצור שני סטים עם מאפיינים דומים לשני תתי סטים:  
**סט אימון –** סט הכולל 80% מהתצפיות (35610 תצפיות) שישמש אותנו לאימון המודל ומציאת הקונפיגורציה (כוונון הפרמטרים) הטובה ביותר.   
**סט בחינה -** סט הכולל 20% מהתצפיות (8903 תצפיות) שיסייע לנו בהשוואת ביצועי המודלים ובחירת המודל הטוב ביותר.   
אחוזי החלוקה נקבעו מתוך הצורך לשמור על כמות תצפיות גדולה משמעותית בסט האימון על מנת לאמן את המודלים באופן מקיף, ועל אף שהדבר פוגע בגודל סט הבחינה, משום שמספר התצפיות בסט הנתונים הוא גדול (כ-44,000 רשומות) סט הבחינה יהיה גדול דיו ויאפשר בחינה מהימנה של המודלים.

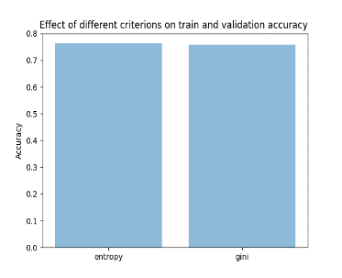
Decision Trees

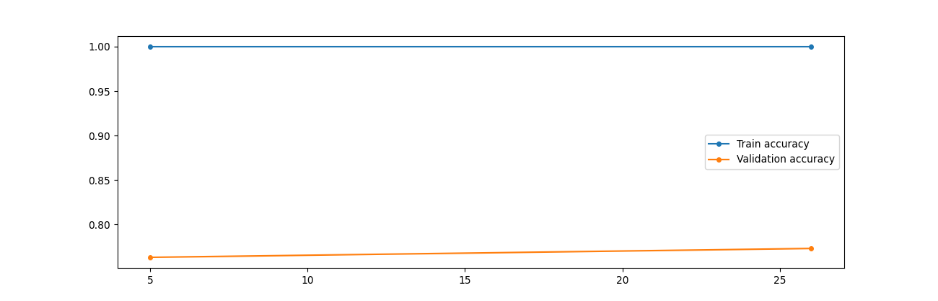
**עץ החלטה מלא -**תחילה בנינו עץ החלטה מלא עם כלל הערכים הדיפולטיביים של המודל ועם כלל התכונות. אחוזי הדיוק שהתקבלו על סט האימון הם 100% (דיוק מקסימלי שמעיד על Overfitting) ועל סט הבחינה הם77.39% (נספח 1 ). לאור ההתאמה המושלמת לנתוני האימון שנצפתה, אמנם המודל מסוגל להניב תוצאות מיטביות על סט נתוני האימון (סיווג מושלם) אך נוצר מצב לא רצוי בו המודל יתקשה לסווג נכונה נתונים מסט בחינה עתידי משום שאינו יצליח לבצע הכללה לאור העובדה שאינו גמיש לשינויים. כראיה ניתן להתבונן בתוצאות הדיוק שהתקבלו על סט הבחינה. בשל כך נרצה לכוונן את פרמטרי המודל בכדי להימנע מהתאמת יתר ולהתאים מודל עם יכולות מיטביות לסיווג סט בחינה עתידי.

**Hyperparameter Tuning-לכוונון פרמטרים** במודל, נשתמש ב-Grid Search שמאפשר לבחון באופן איטרטיבי מודלים עם קומבינציות שונות של היפר פרמטרים. במסגרת Grid Search נעשה שימוש בשיטת K-Fold cross-validation המאפשרת לבחון כל קומבינציית פרמטרים מספר פעמים (במקרה שלנו בחרנו K=10- ערך ברירת מחדל שנקבע על פי כלל אצבע שנמצא כערך המיטבי לאיזון בין סיבוכיות הריצה להטיה נמוכה באמידת ביצועי המודל). השיטה מתבססת על חלוקת סט הנתונים ל-K תתי סטים שווים, כאשר בכל פעם תת סט אחר מתפקד כסט אימות והשאר מתפקדים כסט אימון והיא מאפשרת לקבוע את הדיוק בהתאם לממוצע התוצאות של כלל האיטרציות. בנוסף, במהלך התהליך ביצענו מספר הרצות של Grid Search כמתואר בהמשך, ההרצות הראשונות בחנו אימונים של המודל על סט הנתונים הלא מאוזן, ולכן המדד שבחרנו למקסם הנו מדד F1 שמתאים לבחירת מודל שמתאמן על נתונים לא מאוזנים. בהמשך נראה שרמת הדיוק המתקבלת אינה משתפרת משמעותית כתוצאה משינויים שנבצע, ולכן נאזן את הנתונים. לאחר האיזון, המדד שנרצה למקסם יהיה מדד שכבר מתאים לסט אימון מאוזן - מדד הדיוק -Accuracy - אחוז התצפיות עבורן בוצע חיזוי נכון מתוך כלל התצפיות והקונפיגורציה שתיבחר היא זו שתניב את הדיוק הממוצע הגבוה ביותר על סט האימות.

**היפר-פרמטרים שבחרנו לכוונן– מוטיבציה ומשמעות:**

**Max Depth-**עומק העץ המקסימלי האפשרי. ככל שעץ עמוק יותר (מכיל יותר רמות) הוא יחטא ב Over-Fitting, ומנגד, עץ עם מספר רמות קטן מדי יצור מודל שלא מנצל כראוי את הנתונים שברשותו (Under-Fitting). לכן עלינו למצוא trade-off לעומק עץ שיניב דיוק מקסימלי על סט הבחינה. מרחב האפשרויות שניתן לעומק העץ הנו [0-26] כמספר המאפיינים בסט האימון. הגרף המוצג מתאר את הקשר בין עומקים שונים של עץ לדיוק המודל כאשר שאר הפרמטרים מקובעים ע"י הערכים האופטימליים שהתקבלו עם האלגוריתם:   
ניתן לראות התכנסות של מדד הדיוק לערכים הנעים סביב 75% עבור עומקי עץ הגדולים מ-7. על כן למרות שבדקנו מרחב אפשרויות רחב לעומקי העץ ניתן לראות כי ניתן לחסוך זמן ריצה משמעותי ולהגיע לתוצאות דומות של רמת דיוק גם עבור עומקי עץ קטנים יותר.

**Critirion**-קריטריון חישוב לפיו נבחר המאפיין לפיו יתבצע הפיצול בצומת. מרחב הערכים שנרצה לבחון הנו : [Entropy, Gini].המדדים משמשים להערכת איכות הפיצול בעץ ונבחר פיצול שממזער את ערכי המדדים.  
אנטרופיה - מדד לאי וודאות בדגימות המודד אקראיות. אם כל הדגימות בקבוצה שייכות לאותה מחלקה, האנטרופיה אפס. אם הדגימות מחולקות באופן שווה בין שתי מחלקות, האנטרופיה בערך המקסימלי שלה.  
Gini- מדד נוסף לאי ודאות שמוגדר כהסתברות לבחירה אקראית של שני דגימות מהסט שיהיו ממחלקות שונות. כמו אנטרופיה, אם כל הדגימות בקבוצה שייכות לאותה מחלקה, מדד ג'יני אפס. אם הדגימות מחולקות באופן שווה בין שתי מחלקות, מדד ג'יני נמצא בערך המקסימלי שלו.  
הגרף מתאר את אחוז הדיוק בהתאם לכל אחד מערכי הקריטריון עבור סט האימון וסט הולידציה כאשר שאר הפרמטרים מקובעים לפי ערכם המיטבי שנמצא בחיפוש הגרידי. ניתן לראות כי אין הבדל בערך הדיוק בהתאם לבחירה בין שני המדדים ולכן נשער כי סבירות הבחירה באחד מהם היא זהה ושניהם יפיקו עצים די דומים.

**Max\_features:** קריטריון המייצג את מספר המאפיינים מבניהם תיבחר הצומת הבאה בעץ. כלומר, הצומת הבאה לא תיבחר מבין כלל המאפיינים שטרם נבחרו, אלא רק ממספר מוגבל של מאפיינים בכל צומת. ככל שמספר המאפיינים יהיה קטן יותר ייתכן והמאפיינים המשפיעים לא יבחרו (under- fitting) ומנגד, במידה והבחירה תהיה מתוך מספר מאפיינים גדול הם יבחרו על פי סט האימון מתעורר חשש ל- over fitting . האפשרויות שנבחן: [ 'sqrt', 'log2', 'None] שיגדירו כמויות שונות של מאפיינים שיהיו מועמדים להבחר. (''None-ללא הגבלה, והשאר יחזירו ערכים כמותיים שמוגדרים כתוצאה של הפעולה המתמטית על מספר המאפיינים המלא בסט הנתונים).הגרף מתאר את אחוז הדיוק בהתאם לערכי הקריטריון עבור סט האימון וסט הולידציה כאשר שאר הפרמטרים מקובעים לפי ערכם המיטבי שנמצא בחיפוש הגרידי. ניתן לראות עליה קלה של הדיוק עבור סט הולידציה ככל שהמדד עולה.

1. **אימון עץ ההחלטה**

**אלגוריתם ה- Grid Searchהראשוני** שהרצנו על סט אימון הכולל את כלל התכונות (26 מאפיינים כולל תכונת ה-BMI שהוספנו בחלק א). תוצאות ה- Accuracyשהתקבלו על סט האימון הם: 99.18% (תזוזה מזערית ממצב overFittingהראשוני) ועל סט הבחינה 77.02% (**נספח 2**). ניתן לראות כי האלגוריתם לא הביא תוצאות ששיפרו את תוצאות העץ המלא (העץ שהאלגוריתם בחר כלל את כלל התכונות עם עומק עץ קטן רק ברמה אחת מהעץ המלא -25).

**בחירת תכונות** – לשיפור הדיוק, נאמן את המודל על סט עם מספר מצומצם של תכונות כך שנסיר תכונות שיש בניהן מולטי-קוליניאריות. נסיר את התכונות גובה ומשקל (נשתמש ב- BMI- משתנה שהגדרנו בחלק א) ונסיר את: LDL,HDL, triglyceride (בחלק א- איתרנו קשר ליניארי מובהק בין מאפיינים אלו לכולסטרול על פי הנוסחה ). ה- Accuracy שהתקבל על סט האימון: 97.33% (תזוזה מזערית ממצב overfittingבשל אי שימוש במאפיינים מסוימים) ועל סט הבחינה 76.32% (**נספח 3**). ניתן לצפות בהרעה (של 0.7 אחוז) שנבעה מהסרת המאפיינים אך הקלה בסיבוכיות הריצה של האלגוריתם.

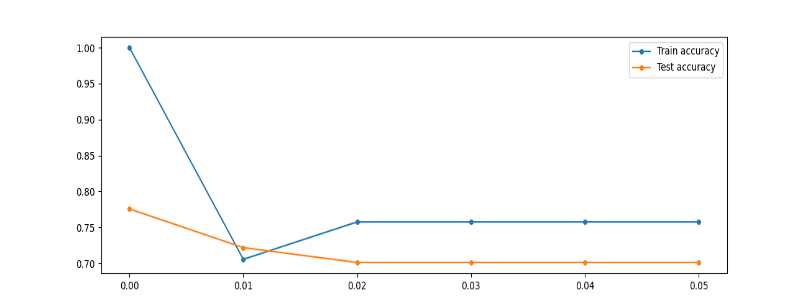
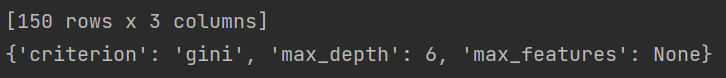
**נרמול הנתונים-** ראינו שהסרת המאפיינים לא הטיבה למודל, לכן בחרנו להחזיר את המאפיינים שהוסרו ונרמלנו את הנתונים. לאחר מכן הרצנו את החיפוש הגרידי לסט המנורמל. תוצאות ה- Accuracyשהתקבלו על סט האימון הם: 99.76% (שוב התקרבות למצב של overfitting) ועל סט הבחינה 75.61% (**נספח 4)** ניתן לראות שגם נרמול הנתונים לא הביא לשיפור בביצועי המודל ואף גרע מהם מעט.

הסבר הגיוני לתוצאות שהתקבלו שאינן מציגות שיפור ואף בוחרות מודל עם עומק עץ מקסימלי(26) הנו החוסר איזון שקיים בנתונים. לפי ההסתברויות הא-פריוריות שבחנו בחלק א' ראינו כי סט הנתונים שלנו אינו מאוזן:  
*P(Smoking = 0)=63.25% , P(Smoking = 1)=36.75%* . אימון מודלי סיווג על נתונים לא מאוזנים יכול לגרום להטיה בתוצאות הסיווג (למשל מקרה קיצון -סט של 100 תצפיות מתוכן 99 ממחלקה 1 ו-1 ממחלקה 2, אם המודל תמיד יסווג תצפיות למחלקה 1 דיוק המודל יהיה כ-99% - לכאורה משקף מודל מדויק אך המשמעות היא שלעולם המודל לא יזהה אלמנטים מהמחלקה השנייה, וכך נוצרת הטיה). לכן למרות שבחרנו למקסם את F1 שמתאים לסט נתונים לא מאוזן, המודלים לא שיפרו משמעותית את הדיוק על סט הבחינה. לכן נפעל באופן אגרסיבי יותר לטיפול בחוסר האיזון (לא רק ע"י הגדרת הקונפיגורציה לפי מקסום מדד F1 על סטי הוולידציה).  **לפיכך נרצה לאזן את הנתונים והאלטרנטיבות שעומדות בפנינו בכדי ליישם את משימת האיזון הנן:**דגימת יתר (Oversampling): כרוך בהוספת עותקים נוספים של מחלקת המיעוט למערך הנתונים כדי להפוך אותו למאוזן יותר. משום שכמות הנתונים שבידינו גדולה דיו, נעדיף שלא להשתמש בשיטה זו.  
תת דגימה (Undersampling): כולל הסרה של חלק ממחלקת הרוב כדי להפוך את מערך הנתונים למאוזן יותר. לא נרצה לאבד נתונים- ככל שהמודל מתאמן על כמות נתונים גדולה יותר הוא יהיה מוכשר יותר להתמודדות עם מרחב מקרים גדול יותר בעת מתן סיווג.  
דגימה ממושקלת (Weighted sampling): כולל התאמה של משקלי המחלקות במודל כדי להפוך את המודל לרגיש יותר למחלקת המיעוט. נשתמש בשיטה זו ע"י הוספת קונפיגורציה למודל: class\_weight='balanced'.

**Grid Search עם דגימה ממושקלת להתמודדות עם חוסר האיזון-** הרצת החיפוש הגרידיעם נתונים מאוזנים, כלל התכונות וללא נרמול ע"י מקסום מדד ה- Accuracy(שכבר מתאים לסט מאוזן) תוצאות ה- Accuracy על סט האימון: 99% (התקרבות ל- overfitting) ועל סט הבחינה 76.75% (**נספח 5).** ניתן לראות שאיזון הנתונים בשיטה זו לא שיפר את הביצועים מהעץ המקורי וכן ניצל את העומק עץ המקסימלי האפשרי.  
על כן נבצע איזון אגרסיבי שכרוך באיבוד נתונים ע"י שיטת Undersampling באמצעות פונקציית RandomUnderSampler() להסרת נתונים ממחלקת הרוב.

**Grid Search עם הסרת דגימות (Undersampling) להתמודדות עם חוסר האיזון-** לאחר האיזון התקבל סט נתוני אימון שמחציתו מעשנים ומחציתו לא מעשנים (**נספח 6**). כאשר אימנו את סט הנתונים המאוזן עם מקסום מדד הדיוק ה- Accuracy על סט האימון: 76.89% (התרחק מאוד מ- overfitting) ועל הבחינה 73.45%.(**נספח 6**) ניתן לראות ירידה של כ-3% - האיזון פגע בדיוק של סט הבחינה אך מנגד הביא להתרחקות מ- .OverFittingכמו כן, העץ המיטבי שנבחר כלל עומק של 6-(לראשונה לא נבחר עומק של 25) , הקריטריון הנבחר הוא GINI וסה"כ תכונות אופציונליות לבחירה בפיצול .None- קיבלנו עץ עם מורכבות הכי קטנה עד כה והתרחקות הכי משמעותית מ- OverFitting אך יחד עם זאת דיוק הכי נמוך. למרות הדיוק הנמוך, נבחר לאמן את המודל עם נתונים מאוזנים שכן ההתרחקות מ- OverFitting היא המיטבית וכן העץ מכיל מספר רמות מועט משמעותית ביחס לשאר העצים ולכן, נבחר במודל שמתאמן על נתונים מאוזנים.

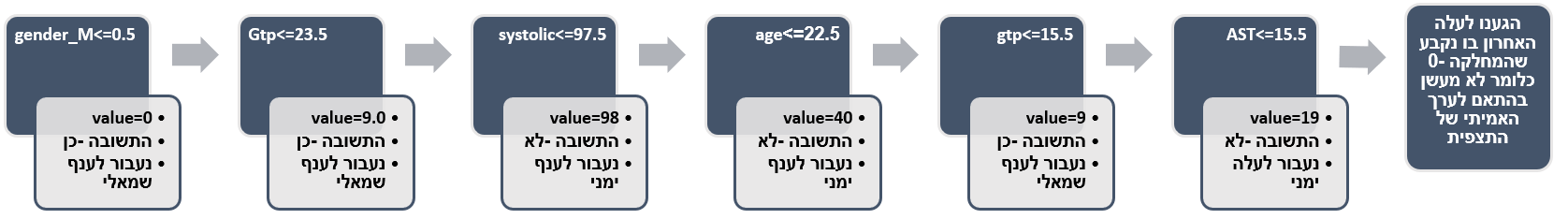
משום שאף אחת מהשיטות לא סייעה בשיפור המודל, נבחן שימוש **בטכניקת קיטום של עץ ההחלטה**- (pruning) - טכניקה המשמשת לשיפור יכולת ההכללה של המודל על ידי הפחתת מורכבות העץ. זה כולל הסרת צמתים מהעץ שאינם מספקים שיפור משמעותי בדיוק הניבוי. קיימות שתי גישות עיקריות לקיטום :  
קיטום מקדים (Pre-pruning): כרוך בעצירת העץ מלגדול מעבר לנקודה מסוימת. ניתן לעשות זאת על ידי הגדרת עומק עץ מקסימלי (פרמטר שכבר ביצענו עבורו כוונון) או מספר דגימות מינימלי הנדרש בצומת עלה.  
קיטום מאוחר (Post-pruning): כרוך בבניית העץ המלא ולאחר מכן הסרת צמתים שאינם משפרים את דיוק המודל. שיטה פופולריות לקיטום מאוחר היא אלגוריתם קיטום שגיאות מופחת (reduced error pruning algorithm), הכולל השוואת דיוק המודל עם ובלי צומת נתון והסרת הצומת אם זה לא משפר את הדיוק.

גם Pre-pruning וגם Post-pruning עשויים לשפר את יכולת ההכללה של העץ על ידי הפחתת התאמה יתר נבחן ביצוע של קיטום מאוחר ע"י כוונון היפר פרמטר נוסף **ccp\_alpha**  -מספק אפשרות נוספת לשלוט בגודל העץ. פרמטר זה מגדיר את עלות המורכבות של הצומת לפיה נצבע קיטום. ככל ש ccp\_alpha מקבל ערכים גדולים יותר כך אנו מגדילים את מספר הצמתים הקטומים. הגרף מתאר את הקשר בין ערכים שונים של ההיפר פרמטר לדיוק המודל.ניתן לראות שערך הדיוק המקסימלי מתקבל עבור ערך 0 של הפרמטר שבמילא זה הערך הדיפולטיבי של המודל ולכן לא נכוונן את פרמטר זה, ונשאר עם המודל האחרון שמצאנו- מודל המאומן על נתונים מאוזנים, עם התכונות:

**מצורף העץ שהתקבל (נספח 7).**

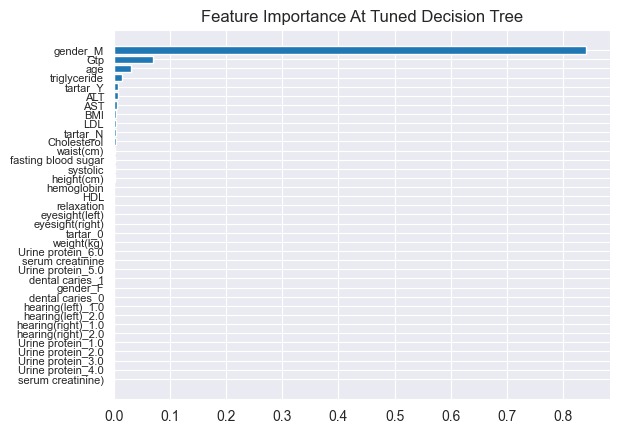
**סיווג רשומה לדוגמה מסט הבחינה-**נבחר אקראית את הרשומה הבאה מסט הבחינה( **נספח 8**)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **age** | **height(cm)** | **weight(kg)** | **waist(cm)** | **eyesight(left)** | **eyesight(right)** |
| **value** | 40 | 160 | 45 | 67 | 0.9 | 1.5 |
| **Feature** | **systolic** | **relaxation** | **fasting blood sugar** | **Cholesterol** | **triglyceride** | **HDL** |
| **value** | 98 | 70 | 79 | 208.0 | 40 | 79 |
| **Feature** | **LDL** | **hemoglobin** | **AST** | **ALT** | **Gtp** | **serum creatinine** |
| **value** | 121 | 12.9 | 19 | 15 | 9 | 0.6 |
| **Feature** | **BMI** | **gender** | **hearing(left)** | **hearing(right)** | **Urine protein** | **dental caries** |
| **value** | 17.578125 | F  (gender\_F=1) | 1  (hearing(left)\_1=1) | 1  (hearing(right)\_1=1) | 1  (Urine protein\_1=1) | 0  (dental caries\_0=1) |
| **Feature** | **tartar** | **smoking** | | | | |
| **value** | Y  (tartar\_Y=1) | **0 (לא מעשנת)** | | | | |

****

**תובנות ממבנה העץ-**מהתבוננות במבנה העץ, ניתן ללמוד מי המאפיינים שמסייעים לסווג בצורה הטובה ביותר. הסיווג בעץ ההחלטה לא מסתמך על תרומתו של כל מאפיין להסברת המשתנה המוסבר בנפרד, אלא בכל רמה המאפיין שנבחר הוא בהתאם למידת ההסברה שלו ביחס למאפיינים המופיעים ברמות מעליו. ניתן לראות ששורש העץ מפצל לפי מאפיין המגדר, ובהתבוננות כללית ניתן לראות שצד שמאל (ששייך למגדר M) מכיל יותר תאים כחולים – שמשמעותם יותר תצפיות של מטופלים מעשנים כלומר מגדר היא תכונה מרכזית בסיווג וניתן לראות שגברים יסווגו כמעשנים בסיכוי הרבה יותר גבוהה מנשים. כמו כן ניתן לראות שמשתני הגיל, GTP, Systolic, AST , ALT וTriglyceride נמצאים בעץ פעמים רבות במספר רמות שונות כלומר נבחרו מספר פעמים כתכונות מסווגות טובות בהינתן נתונים רפואיים אחרים על המטופל (הרמות מעל).

**Feature importance :**להלן תוצאות השימוש בפונקציית חשיבות המשתנים של המודל

ניתן לראות בפלט כי אכן המאפיינים שעל פיהם נסווג בעץ המכוונן הם המאפיינים שקיימת עבורם חשיבות בגרף. הסיווג הראשון בעץ נעשה על פי המאפיין מגדר ואכן גם בגרף ניתן לראות שהוא בעל החשיבות הרבה ביותר. לאחר מכן, GTP(אנזים כבד) הוא בעל חשיבות גבוהה בהתאם למשוער בחלק א' וניתן לראות שמסווג פעמיים רבות ברמות שונות בעץ . אחריו ניתן למצוא את Triglyceride שבהתאם להשערה בחלק א' שיערנו שיש קשר בינו לבין עישון (רמות גבוהות שלו סביר שיעידו על עישון). כמו כן, גם שאר המאפיינים שנמצאו עם חשיבות שאינה מתאפסת מתיישבים עם ההשערות מחלק א וניתן לראות שהם מאפיינים שמופיעים פעמים רבות בעץ וקובעים את הסיווג. כמו כן, ניתן לראות ש Systolic-קיבל חשיבות אפסית בגרף על אף שנבחר בעץ למסווג בצמתים רבים. ברמה הכללית מרבית התוצאות מתיישבות עם המסקנות מהסעיף הקודם, וכמובן שאין הפלט מעיד על חשיבותם הישירה של המאפיינים אצל משתנה המטרה, אלא כתלות בענף בו הסיווג מופיע.

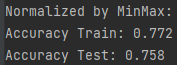
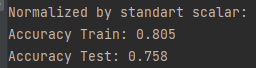
**דוגמאות של מאפיינים בעלי חשיבות שונה**  
**gender** מסווג בשורש העץ כלומר נבחר לבצע את הסיווג הראשוני- המשמעות שהמאפיין מלמד אותנו רבות על משתנה המטרה. תוצאות אלה מתיישבות עם האינטואיציה ועם המשוער בחלק א בו טענו כי לפי מחקרים, אחוז הגברים המעשנים מבין הגברים גבוה מאחוז הנשים המעשנות מבין הנשים וכן הצגנו תרשימים שתומכים בטענה. ניתן לראות כי מיפוי הצבעים בעץ מתבסס משמעותית על מאפיין זה כמתואר בסעיף הקודם.

**serum creatinine** -תפקוד כליות לקוי - נראה בגרף כי חשיבותו אפסית על אף ההשערה שהעלנו בחלק א' המתבססת על [מאמרים רבים הדנים בפגיעה של עישון בכליות שמכך הסקנו שתפקוד כליות לקוי יכול להעיד על סטטוס עישון חיובי](https://www.cancer.org.il/template/default.aspx?PageId=5795). ייתכן כי בהינתן מאפייני מטופל אחרים שנמצאו כמסבירים טובים יותר לעישון ונמצאו ברמות העץ, תפקוד הכליות מוסבר על ידיהם ולכן לא יגרור הבנה משמעותית על משתנה המטרה.

Neural network

# עבודה מקדימה- ראשית רצינו לבדוק באילו מהמשתנים התלויים שלנו יש להשתמש: מדד ה-BMI לעומת גובה ומשקל, ומדד הכולסטרול שבדקנו כי מורכב ממדי ה-LDL, HDL וטריגליצריד, הרצנו את מודל ברירת המחדל עבור כל אחת מארבעת השילובים האפשריים ביניהן וקבלנו את התוצאות הבאות: [(נספח](#מאפיינים) 10), לכן החלטנו להשאיר את מדדי ה-BMI וכולסטרול, מתוך הערכה של עלות מול תועלת- פגיעה בדיוק המודל לעומת שיפור זמן הריצה של האלגוריתם.

# **נרמול ערכים -** בדקנו 2 סוגי נרמול שונים למאפיינים (סטנדרטיזציה, MinMax) גם עבור סט הנתונים וגם עבור סט הבחינה על מנת ליצור אחידות בין המאפיינים ולמנוע הטיה של הנוירונים לכיוון מאפיינים בעלי תוחלת גבוהה:

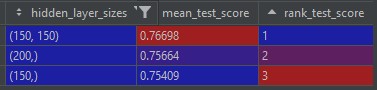
Not Normalized: Accuracy train-74.9%, Accuracy Test 74.3%  
Standard scalar: Accuracy train-80.5%, Accuracy Test 75.8%  
MinMax: Accuracy train-77.4%, Accuracy Test 75.8

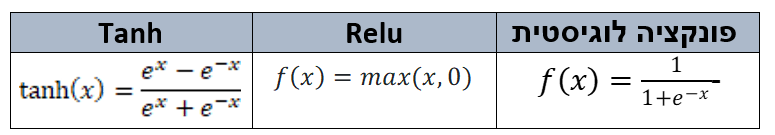
נראה שקבלנו תוצאות כמעט זהות על סט בחינה, ולכן נבחר בנרמול עם התפלגות נורמלית כיוון שזמן הריצה שלו מהיר יותר.

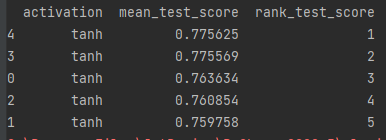
# ערכי ברירת המחדל- מצ"ב ערכי ברירת המחדל הנקבעים אוטומטית על ידי המודל: [(נספח 9)](#HYPER_PARAMETER_DEFAULT). הרצנו מודל רשת הנוירונים על פי ברירת המחדל, כאשר מספר הנוירונים בשכבת הכניסה זהה לכמות ערכי המאפיינים בסט הנתונים, 21, שכבה חבויה בודדת בעלת 100 נוירונים, פונקציית ההפעלה היא *'relu'-* פונקציית היחידה לינארית המתוקנת המוגדרת: f(x) = max(0, x).עבור מודל זה התקבלו התוצאות: 80.6% על סט האימון ו 75.8%על סט הבחינה- תוצאות הדיוק עבור סט האימון מראים כי לא הגענו ל- Over-fitting ולא נחשוש מהתאמה מושלמת של חיזויי המודל אך ורק על ערכים המתקבלים בסט האימון.

# Hyperparameters tuning-נבצע כוונון של הייפר-פרמטרים בעזרת GridSearchCV, עם K=10, מסיבות דומות לאלה הפורטו בעץ ההחלטה, במטרה למצוא שילובי קונפיגורציות שישפרו את התוצאות המתקבלות. את בחירת היפר-הפרמטרים נבצע בצורה חמדנית, משיקולי זמן ריצה, כלומר נבצע בחינה של כל פרמטר באיטרציה נפרדת, כאשר בכל איטרציה נוסיף למודל את הפרמטר שהכי משפר את דיוק המודל.

**1. שכבות חבויות Hidden layer size:** הפרמטר קובע את מספר השכבות החבויות ואת כמות הנוירונים הקיימים בכל שכבה, ככל שנוסיף כמות גבוהה יותר של שכבות ונוירונים בכל שכבה- כך נוכל להשיג תוצאות גבוהות יותר מבחינת הדיוק. כיוון שכך נוכל למצוא קשרים מורכבים ומסובכים יותר בסט הנתונים, אך כמות גבוהה מדי של שכבות ונוירונים עלולים להוביל ל- Over-fitting ממנה ננסה להימנע. תחילה ניסינו לבדוק מספר של קומבינציות אפשריות בעלות כמויות שונות של נוירונים (30-100) ושכבות (1-3).כפי שניתן לראות בתוצאות, הקומבינציה הטובה ביותר התקבלה עבור שלוש שכבות בעלות 100 נוירונים כל Table

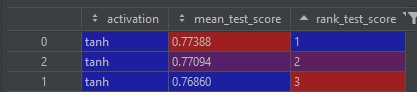
Description automatically generatedאחת. כאשר קונפיגורציה זו הובילה לאחוזי דיוק של **77.11%**. כיוון שראינו כי התוצאות הגבוהות ביותר התקבלו עבור 100 נוירונים בכל שכבה, הוספנו לבדוק אופציות נוספות בעלות כמות גבוהה יותר של נוירונים- אך שינוי זה אינו שיפר את המודל.

**2. פונקציית ההפעלה Activation Parameter:** משמעות פרמטר זה הוא הפעולה המתבצעת על כל נוירון בכל שכבה, כאשר התאמה של הפונקציה עשויה לתרום לבחינת אי הליניאריות של הבעיה, המוטיבציה לבדיקת ערך זה מתקבלת כיוון ששינויים בערך זה עשויים לגרום לשינויים גדולים בערכי החיזוי. עבור פרמטר זה בדקנו 3 פונקציות שונות:

****כפי שניתן לראות, התוצאה הטובה ביותר התקבלה עבור טנגנס היפרבולי אשר תוצאותיו היו נכונות ב77.55% מהמקרים (שיפור של **0.44%** מהתוצאה הקודמת).

**3**. **כמות איטרציות Max\_iter:** מספר האיטרציות המקסימלי עד לעצירת האימון של המודל, המודל יפסיק להתאמן כאשר יבצע max\_iter צעדים או כאשר יגיע להתכנסות. מוטיבציה לבדיקת פרמטר זה הייתה מציאת איזון בין זמן ריצה לבין תוצאות- הגבלה על כמות האיטרציות עשויה להקל על זמן הריצה בצורה משמעותית, אך באותה העת להרוס את ניבוי המודל- לכן רצינו לבדוק כיצד מתנהג המודל במספר שונה של איטרציות:

(0, 15),(1, 30), (2, 50), (3, 200), (4, 300).כפי שניתן לראות, התוצאה הגבוהה ביותר התקבלה, בהפרש קטן (0.0056%) כמות של 300 איטרציות.

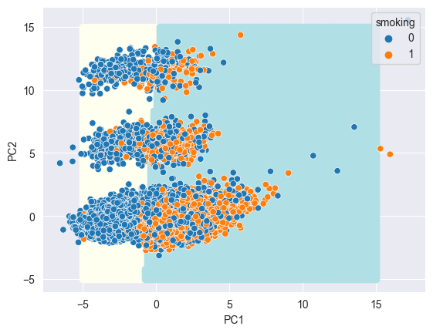
**4. קצב הלמידה ההתחלתי learning\_rate\_init:**הפרמטר קובע את קצב הלמידה ההתחלתי של המודל כתגובה לשגיאה בכל פעם שהמשקולות מתעדכנות, ערך ברירת המחדל של פרמטר זה הוא 0.001 – מקריאה באינטרנט גילינו כי ערך קטן מדי עלול לגרום לתהליך למידה ארוך- ממנו ננסה להימנע, בעוד שערך גדול מדי עלול לגרום ללמידה מהירה מדי של המשקולות המוביל לחיזויים לא נכונים. לכן בדקנו כ-3 ערכים של פרמטר זה: (0, 0.001), (1, 0.003), (2, 0.005).כפי שניתן לראות- החיזוי הטוב ביותר התקבל עבור ערך ברירת המחדל, איתו נבחר להישאר.

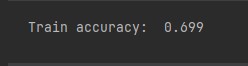
**הקונפיגורציה הנבחרת-** הערכים המתקבלים עבור הקונפיגורציה הטובה ביותר הם:מספר שכבות חבויות Hidden layer size: 3 שכבות בעלות 100 נוירונים כל אחת.פונקציית ההפעלה Activation Parameter: tanh, כמות איטרציות Max\_iter: 300, קצב למידה התחלתי learning\_rate\_init: 0.001. כל שאר הערכים נותרו כערכי ברירת המחדל. בנוסף, הערכים נורמלו סביב ההתפלגות הנורמאלית.

**אחוזי הדיוק המתקבלים -** אחוז הדיוק המתקבל על סט האימון הוא: **81.5%** ועל סט הבחינה הוא:**78.4%** . תוצאות אלה, כצפוי, טובות מהתוצאות אשר התקבלו לפני כיוונון הפרמטרים בכ-**2.6%** עבור סט הבחינה ובכ-**1%** עבור סט האימון. נסיק, כי הכיוונון בוצע בצורה טובה, אחוז הדיוק על סט הבחינה השתפר פי **2.6** יותר מהשיפור על סט האימון. בנוסף, ניתן לראות כי השינויים לא הובילו אותנו למצב של Over-fitting כיוון שתוצאות סט האימון לא מתקרבות לחיזוי מושלם של 100%, אלה נשארות ביחס קרוב לתוצאות סט הבחינה.

K Means

**1. ערכי ברירת מחדל ותוצאות טיב ההתאמה**-הגדרנו את מודל K Means עם ערכי ברירת המחדל. בנוסף קבענו 2 אשכולות ,עבור מעשנים ולא מעשנים, כמתואר בבעיית הסיווג איתה אנו מתמודדים. על מנת לייצר ייצוג גרפי של תוצאות החיזוי, נעזרנו באלגוריתם PCA, אלגוריתם לניתוח סטי נתונים המכילים כמות גבוהה של תכונות לכל רשומה, כאשר האלגוריתם מוריד את כמות התכונות לכל רשומה תוך שמירה על מקסימום המידע המתקבל מכל תצפית – כך הורדנו את הנתונים לשני מימדים. האלגוריתם חילק את הנתונים ל-2 אשכולות , כיוון שלא ניתן לדעת איזה מחלקה כל אשכול מייצג, נבחן את הסטרואיד של כל אשכול. בתור השוואה, בדקנו את הסטרואיד המייצג את החלוקה האמיתית- עבור מעשנים ולא מעשנים. בנוסף לשני גרפים אלה נעזר בגרף המציג את אותו חיזוי, רק שאת האשכולות המתקבלים על ידי האלגוריתם נציג כצבע הרגע של הגרף, ואת צבעי הנקודות נציג כערכי התצפיות האמיתיות מסט הנתונים:

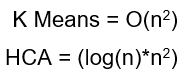


ניתן לראות, לפי גרף זה כי החיזוי ברובו נכון, אך ככל שנתקרב יותר אל מישור ההפרדה הסיכוי לחיזוי שגוי גדל.את טיב ההתאמה של המודל למצב למחלקות קבלנו על ידי ביצוע חישוב דיוק יחסי על סט הבחינה המלא על פי מדד הדיוק, Accuracy, אחוז ההתאמה המתקבל עמד על 69.9%. דיוק זה מעיד על כך שגם ללא משתנה מטרה ניתן להבחין בצורה יחסית טובה (בהשוואה לתוצאות אשר קבלנו בעצי ההחלטה וברשתות הנוירונים) על הבדלים המתקבלים בין המחלקות.

**2. K משתנים-**כעת, נאמן שמונה מודלים שונים, בניתוק מבעיית הסיווג הנתונה עם ערכי K הנעים בין 2-9 וננתח באמצעות מדדים שונים את התוצאות המתקבלות, מתוך מטרה להבין האם חלוקות סט הנתונים שלנו ל-K אשכולות עשוי להסביר את הנתונים בצורה שונה ואולי אף טובה יותר.

**מדדי ההשוואה:** בכדי לבחור מה כמות האשכולות הטובה ביותר עבור סט הנתונים שלנו, נרצה שהשונות בתוך כל אשכול תהיה מינימלית (homogeneity) ובנוסף לכך, נרצה שהשונות בין כל אשכול תהיו מקסימלית (separation) לכן, בחרנו בכ-4 מדדים שונים המתחשבים בפרמטרים אלה ([**נספח**](#bestKgraph) **11**):  
**אינרציה** - מדד זה מחושב באמצעות סכום ריבוי המרחקים של כל תצפית לנקודת המרכז שלה, לכן מודל טוב מאופיין עם מדד אינרציה נמוך. על מנת למצוא את כמות האשכולות האופטימלית נמצא את הנקודה שבה האינרציה מתחילה להאט  
**davies-bouldin** – מדד זה הוא מדד הערכה המחושב באמצעות כמויות וירושת תכונות של סט הנתונים- מדד זה נרצה למזער.  
**Silhouette-** מאפשר למדוד עד כמה חלוקת הנתונים במאגר המידע לקבוצות בעלות מאפיינים דומים – מדד זה נרצה למקסם  
**Calinski-Harabasz**- מדד זה הוא סכום הפיזור בין אשכולות והפיזור בתוך האשכול – מדד שנרצה למקסם.

לכן, בהתחשבות בכל המדדים נבחר את 4 להיות מספר האשכולות בבעיה שלנו- ערך זה מתקשר לסיפור המקרה שלנו ונתונים אשר יצאו באלגוריתמים שהרצנו. ראינו כי תכונת המין היא התכונה בעלת החשיבות הגבוהה ביותר לכן נעריך כי את המעשנים ניתן לחלק חלוקה פנימית- בין נשים וגברים, וכך נגיע ל-4 אשכולות.

**שיטת אשכול נוספת Agglomerative Clustering-**נבחן את החלוקה לארבע אשכולות גם באמצעות האלגוריתם Agglomerative Clustering המחלק לאשכולות באופן שונה מאלגוריתם ה-K Means. אלגוריתם זה הנקרא גם Hierarchial Clustering (HCA), היא שיטה הבונה אשכולות באמצעות הירארכיה, השיטה בא אנו בחרנו עובדת בצורה של "Buttom-up”, כלומר מלמטה למעלה כאשר כל תצפית בסט ה נתונים מתחילה כאשכול בפני עצמו וכל זוג אשכולות מתמזג לדרגה אחת גבוהה יותר במעלה ההירארכיה עד לקבלת כמות האשכולות הרצויה, בצורה זו נוצר אשכול ללא שימוש של הסנטרואידים של האשכולות וכתוצאה מכך גם יכולת ההפרדה שלו נפגעת. לעומת זאת, מבחינת זמני הריצה, הם פועלים בזמן דומה כאשר:



מבחינת יכולת ההשוואה של המודלים, ניתן לראות כי כמעט ולא ניתן להבחין בשינוי האשכולות המתקבלים מכל אחד מהאלגוריתמים אך יחד עם זאת, כאשר נוסיף את התרשים עבור המעשנים מסט הנתונים ניתן לראות את קווי הדימיון המתקבלים מאשכול "1" (צהוב) והנקודות המייצגות אנשים המעשנים.

**השוואה בין מודלים**

במשימת סיווג עלינו לסווג תצפיות נתונים למחלקות ידועות (במקרה שלנו מעשן/לא מעשן) כלומר עלינו לחזות לאיזו מחלקה של משתנה המטרה התצפיות שייכות. לכן במשימות סיווג נעדיף שימוש באלגוריתמים מונחים לעומת אלגוריתמים לא מונחים בהם אין ידע מקדים על המחלקות של משתנה המטרה והסיווג נעשה לפי חלוקה ל-clusters עפ"י דמיון בין רשומות. DT ו-NN הם מודלים מונחים בניגוד לK-MEANS שהינו מודל לא מונחה ולכן נעדיף למשימות סיווג להשתמש ב-DT ו-NN ולא K-MEANS . למרות שבעולם האמיתי לא ניתן לחשב דיוק לאלגוריתם K-MEANS כי המחלקות לא נתונות, עבור הבעיה שלנו קיימים סיווגים למחלקות בסט הבחינה ולכן נשתמש בסיווגים ובאמצעותם נחשב מדדים להשוואה בין המודלים המונחים למודל הלא מונחה.

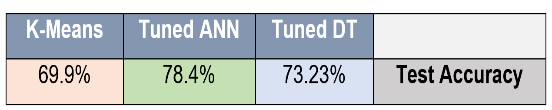
**על מנת להשוות בין המודלים נשתמש במדדים הבאים:**

**אחוז הדיוק על סט הבחינה ( accuracy ) -** מדד המשקף את אחוז החיזויים הנכונים מבין כלל החיזויים. **Precision ו- recall-** שני מדדים המשמשים בדרך כלל להערכת הביצועים של מסווג במיוחד כאשר המחלקות אינן מאוזנות (כלומר, מחלקה אחת נפוצה הרבה יותר מהשנייה).

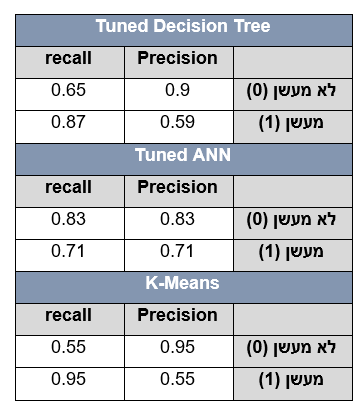
Precision - מדד לדיוק המסווג לסיווג למחלקה מסוימת. הוא מוגדר כמספר התחזיות הנכונות למחלקה מסוימת האמיתיות חלקי המספר הכולל של התחזיות למחלקה זו שנעשו על ידי המסווג.

recall - מדד ליכולתו של המסווג למצוא את כל המופעים של מחלקה מסוימת במערך הנתונים.

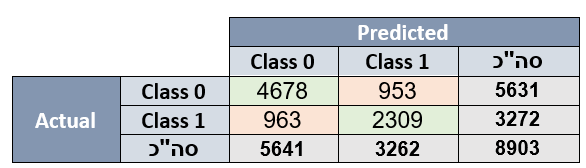
הוא מוגדר כמספר התחזיות הנכונות למחלקה מסוימת חלקי המספר הכולל של התצפיות מהמחלקה הזו.

משמעות המדדים-אם למסווג יש Precision גבוה אך recall נמוך, סביר שהוא מדווח באופן שגוי על מעשן לעיתים רחוקות, אך הוא גם מפספס הרבה מעשנים במערך הנתונים. מאידך, אם למסווג יש recall גבוה אך Precision נמוך, הוא מוצא את רוב המקרים של המעשנים, אבל הוא גם עושה הרבה תחזיות שווא לעישון.

תוצאות הדיוק על סט האימות שהתקבלו משלושת המודלים:

ניתן לראות כי עבור הרשת הנוירונלית התקבלו אחוזי הדיוק הגבוהים ביותר על סט הבחינה. מדובר באלגוריתם מורכב המבצע חישובים רבים ובזכותם לרוב מגיע לאחוזי דיוק גבוהים. השימוש באלגוריתם בא לרוב על חשבון זמן ריצה ארוך לעומת המודלים האחרים וכמו כן יש בו סיכון למצב של Over-Fitting והתאמה מדויקת עבור סט האימון אם ניתן לו להגיע להתכנסות. התוצאה השנייה הטובה ביותר התקבלה מאלגוריתם DT שהוא אלגוריתם פשוט וקל להבנה מבחינה ויזואלית ויישומית. גם אלגוריתם זה ללא כוונון פרמטרים יגיע בקלות למצב של Over-Fitting על סט האימון. האלגוריתם עם התוצאות הנמוכות ביותר הוא K-means -סיבה אפשרית לכך היא שמודל זה אינו בוחר את המאפיינים המשפיעים ביותר, אלא מקבץ את התצפיות לאשכולות על בסיס הדמיון בניהן. להלן מדדי recall ו- Precision לכלל המודלים (**נספח 12)**

* נראה כי עבור המעשנים- ANN מדווח דיווח שווא (דיווח מעשן ובפועל לא) על מעשנים במידה הכי מעטה (Precision גבוה ביותר) ו- K-Meansבמידה הכי גבוהה.
* נראה כי עבור המעשנים- האלגוריתם ANN מפספס דיווח במידה הכי גבוהה (recall נמוך ביותר) ו K-Meansבמידה הכי נמוכה.
* DT נמצא במקום השני מבחינת כל המדדים בשתי המחלקות.

**המודל הנבחר -** לטעמנו הצורך לא לפספס מעשנים חשוב כמו הצורך לא לסווג מטופל כמעשן למרות שהוא לא- ולכן בחירת המודל תיעשה על פי הדיוק- accuracy והמודל שנבחר הנו ANN מכוונן שגם לא מצוי בהתאמת יתר וגם מציג דיוק גבוה ביותר על סט הבחינה. מאפייני הקונפיגורציה שלו הם: מספר שכבות חבויות 3- שכבות בעלות 100 נוירונים כל אחת, פונקציית ההפעלהtanh , כמות איטרציות 300, קצב למידה התחלתי 0.001. שאר הערכים נותרו ברירת המחדל. בנוסף, הערכים נורמלו סביב ההתפלגות הנורמאלית. כפי שהוצג בפרק הרשתות של הדו"ח, קונפיגורציה זו נבחרה מפני שהשיגה את הדיוק הממוצע הטוב ביותר על סט הבחינה. לאחר ביצוע חיזוי על סט הבחינה התקבלה מטריצת מבוכה הבאה (**נספח 13).**

מתוך 5631 התצפיות שמייצגות לא מעשנים, ב-4678 חזינו שהמטופל לא מעשן, ואילו ב-953 האחרות חזינו שהוא מעשן. בהתאמה, מתוך 3272 התצפיות שמייצגות מעשנים, ב-2309 המודל חזה נכונה וב-963 טעה. נשים לב כי אחוז הטעויות בקרב לא מעשנים (P(FP|N)=0.169) – (1 פחות recall לקבוצה 0) נמוך מאחוז הטעויות בקרב מעשנים (P(FN|P)=0.294) (1 פחות recall לקבוצה 1), יתכן שהטיה זו נגרמת מחוסר האיזון בקבוצות בנתונים. כמו כן מבין המקרים שחזינו לא מעשן, הטעות היא 17% (1 פחות Precision לקבוצה 0) ובהתאם מבין המקרים שחזינו עישון, אחוז הטעות הוא 29.2%.(1 פחות Precision לקבוצה 1) מה שמתאים לנצפה בדוח המדדים שיצאו דומים מאוד בשתי הקבוצות (לא זהים). נסיק מכך שהמודל יותר מתקשה לזהות מעשנים אך אם הוא מדווח סביר שהוא צודק. נסיק מכך שהמידה שבה למודל קשה יותר לדייק בחיזוי של מעשנים, קרוב מאוד למידה שבה הוא מתקשה לחזות לא מעשנים, וכך גם הפספוסים של הלא מעשנים קרוב מאוד לאחוז הפספוסים של המעשנים.

# נספחים

**נספח 1**

עץ-מודל מלא

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי תמונה שמכילה טקסט

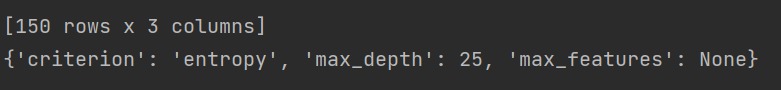
התיאור נוצר באופן אוטומטי

**נספח 2**

Grid search עם כל התכונות:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי



תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

**נספח 3**

Grid search על סט שאינו מכיל את 5 התכונות שיש ביניהן מולטיקולינאריות תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי



תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

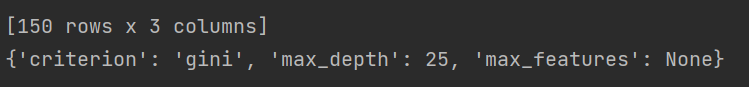
**נספח 4**

עבור נתונים מנורמלים עם כלל התכונות:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

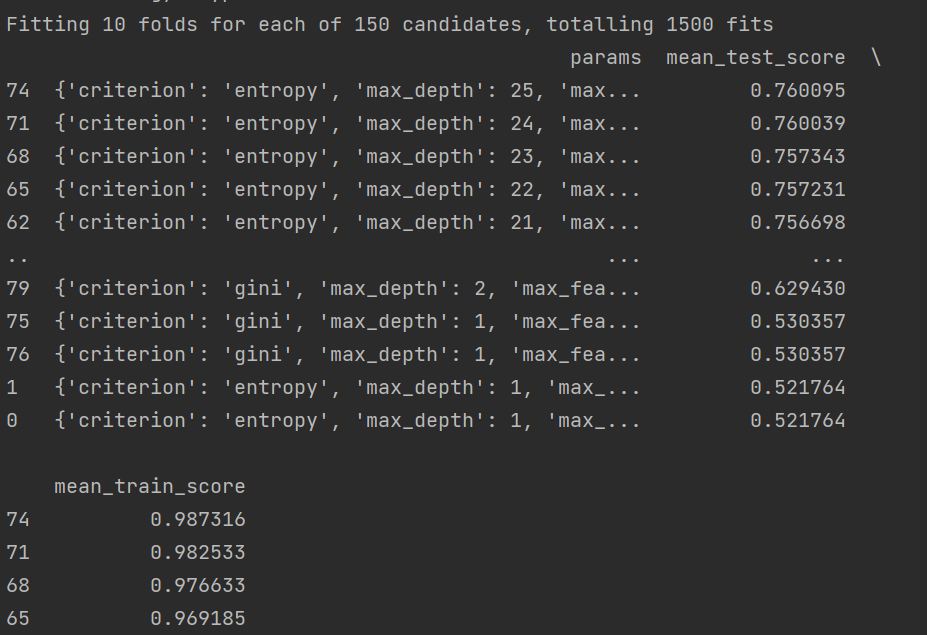


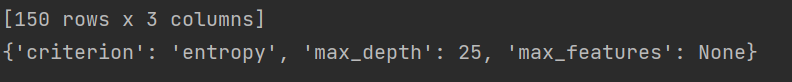
תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

**נספח 5**

איזון ממושקל עם כלל התכונות



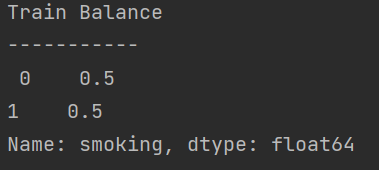


תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

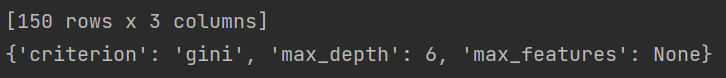
**נספח 6**

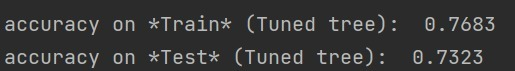
איזון ע"י הסרת תצפיות



תמונה שמכילה טקסט, לוחית

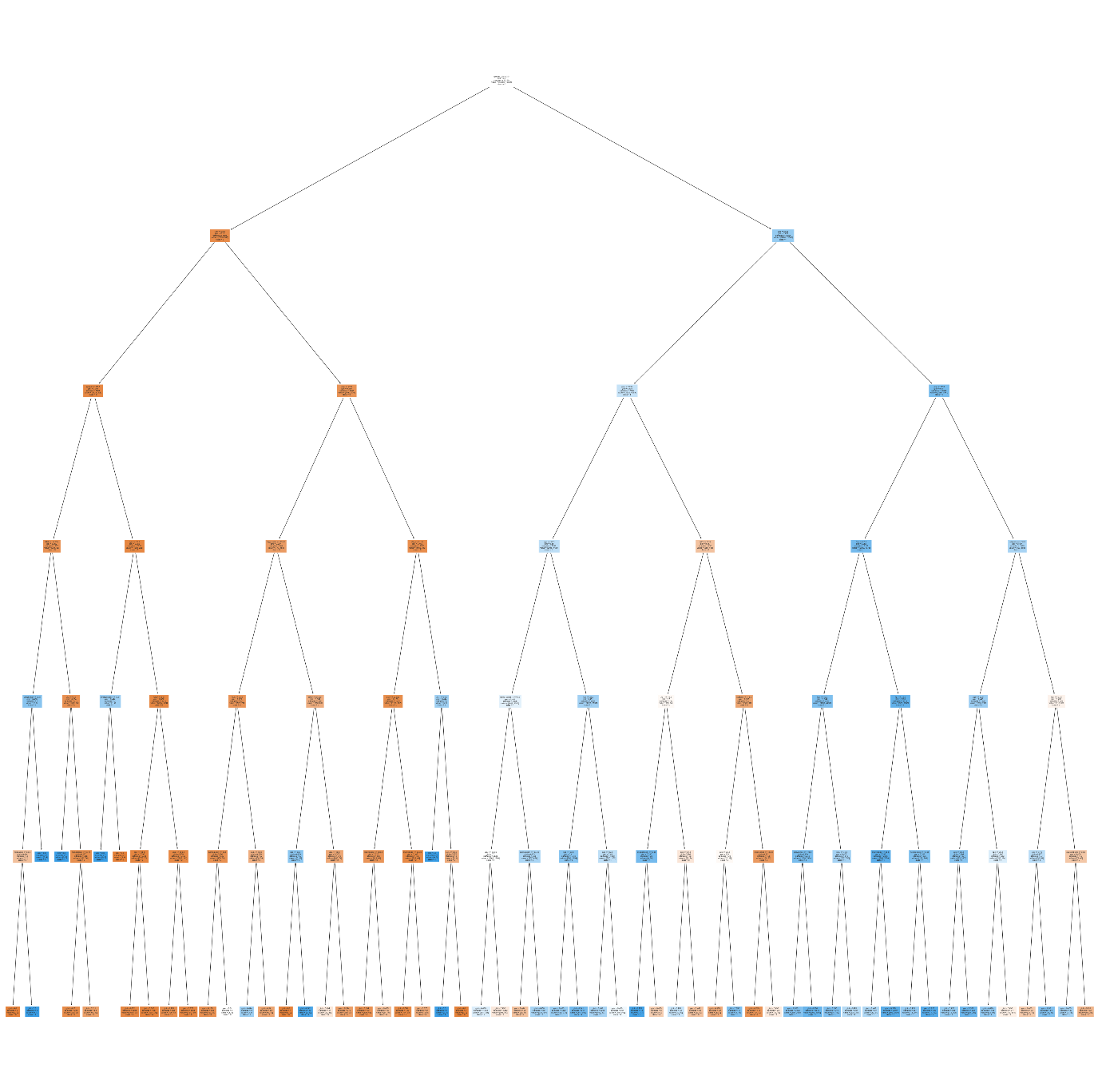
התיאור נוצר באופן אוטומטי



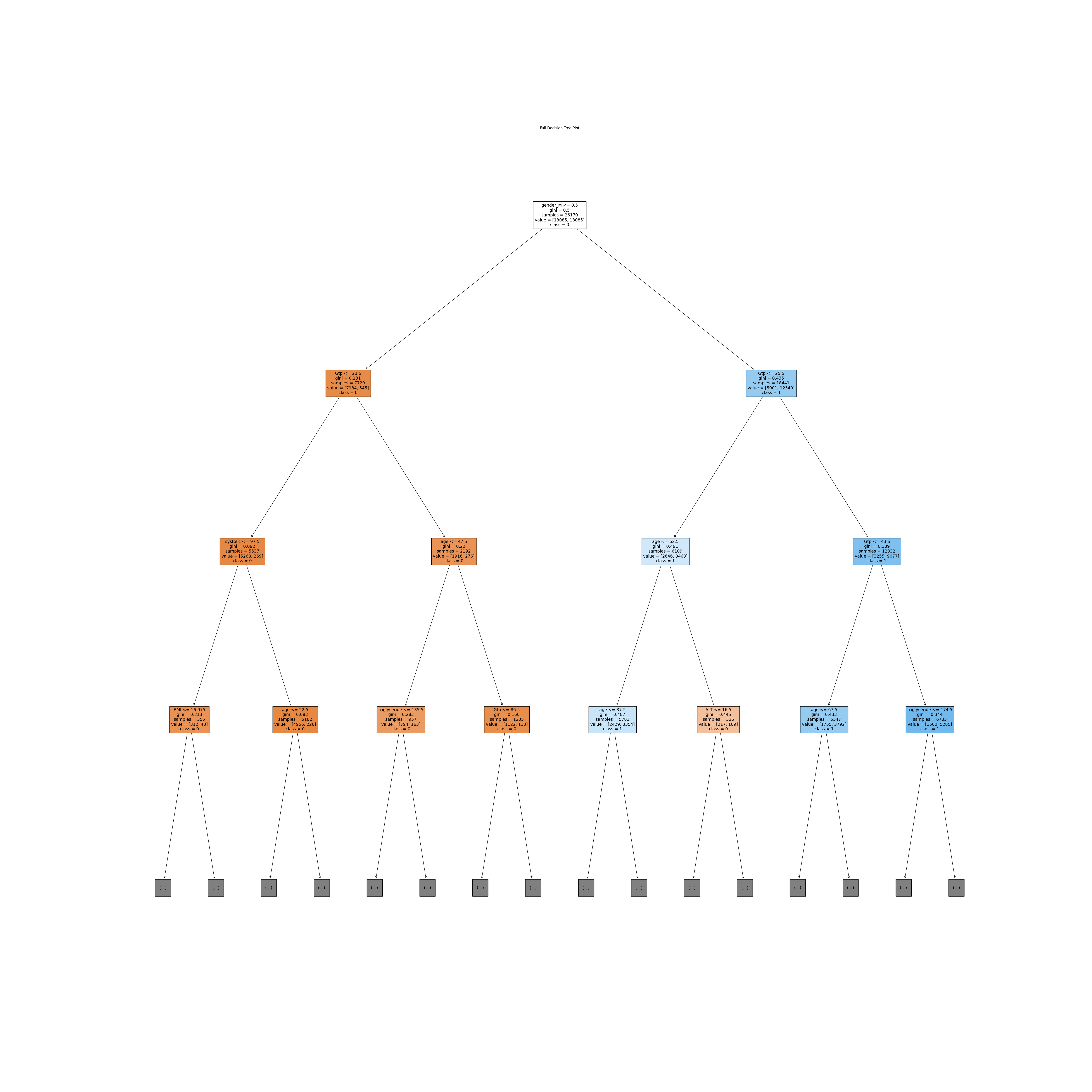


**נספח 7**

העץ הסופי שהתקבל



תקריב של עץ ההחלטה:



**נספח 8**

סיווג רשומה לדוגמא

**תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט, לוחית

התיאור נוצר באופן אוטומטי**

**נספח 9**

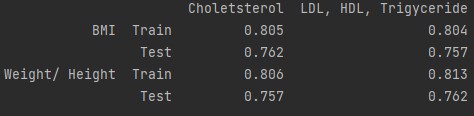
ערכי ברירת המחדל למודל MLP:

[Text

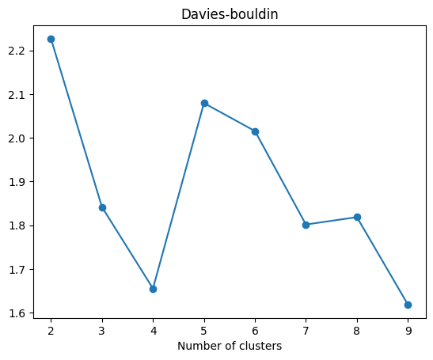
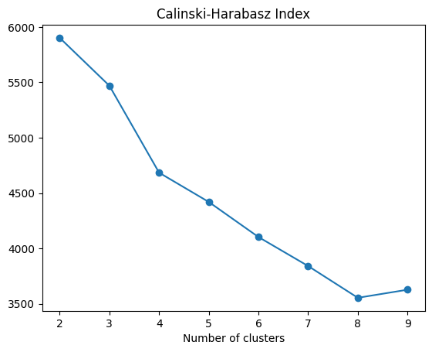
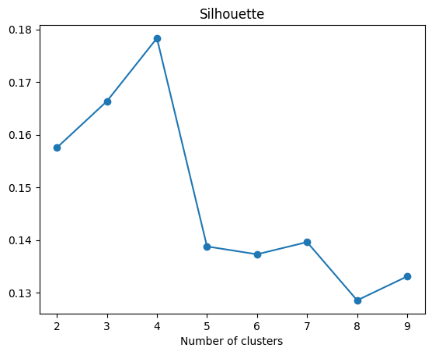
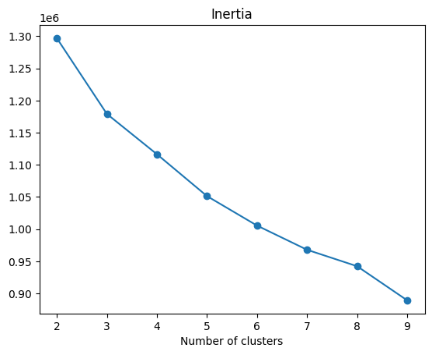
Description automatically generated](#HYPER_PARAMETER_DEFAULT)

**נספח 10**

תוצאות ההרצאה לבדיקת עדיפויות המאפיינים של חלק א':



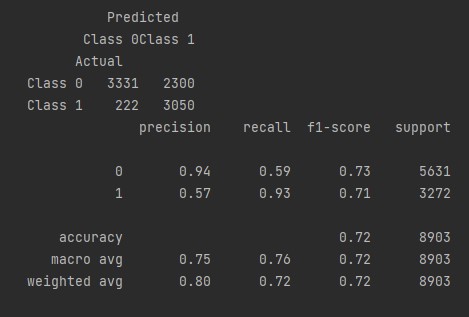
**נספח 11**



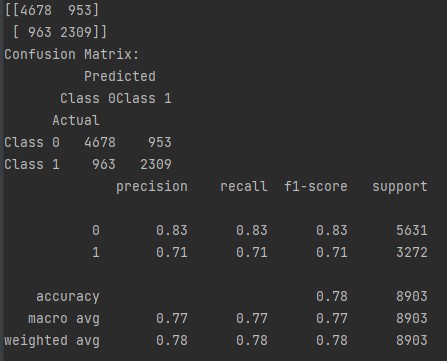
**גרפים של המדדים לבחירת K**

**נספח 12**

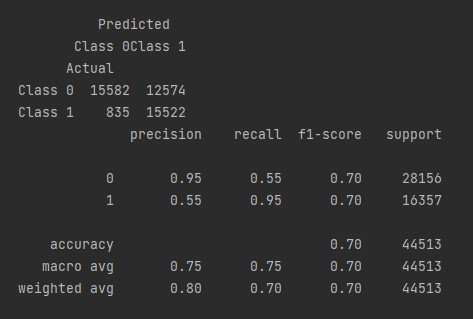
דוח על איכות המודל- עץ ההחלטה



דוח על איכות המודל- רשתות נוירונים



דוח על איכות המודל- K-MEANS



**נספח 13**

מטריצת מבוכה -ANN

