MBA טכניון: פרויקט בקורס למידת מכונה

(מגישים: יחזקאל רבינוביץ , אלחי רפואה)

הקדמה:

פרויקט זה מהווה סיכום למידה של הקורס. בחרנו לחקור dataset של נתוני אלצהיימר מאתר Kaggle. אלצהיימר היא מחלה ניוונית של תאי העצבים במוח, הגורמת לירידה מתמשכת בזיכרון, בכישורי השפה ובתפקוד הקוגניטיבי. המחלה מחמירה בהדרגה בטווח של 3 עד 8 שנים עד למוות של החולה. בעולם יש כ 45 מיליון חולים (בישראל כ 100 אלף חולים). המחלה מהווה אתגר משמעותי בתחום הבריאות הציבורית והבנת גורמי הסיכון וההתקדמות שלה הוא קריטית לצורך זיהוי מוקדם וטיפול מותאם אישית לחולה. (הצעת הפרויקט ותיאור מורחב של ה dataset נמצאים בנספח 2.)

1. קריאת הקובץ:

1.1. טעינת הדאטה:

כצעד מקדים, טענו את כל ספריות פייתון הרלוונטיות לפרויקט (ראה בנספח 3), לאחר מכן ביצענו טעינה של ה dataset באמצעות פקודה מובנת של ספריית pandas , לאחר מכן השמטנו מספר עמודות לא רלוונטיות מהdataset הגולמי, לבסוף טענו את ה dataset למבנה נתונים שנקרא בשם קיצור df .

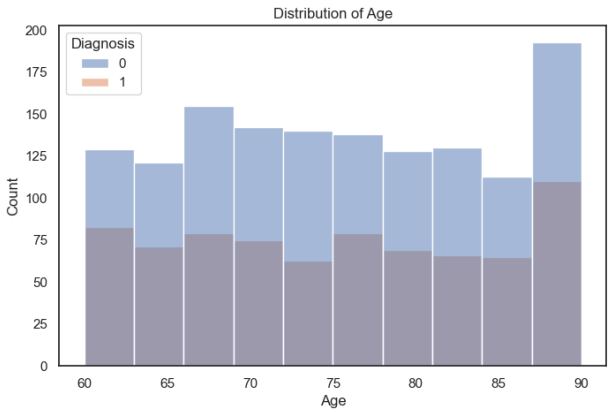
1.2. חקירת הדאטה:

ביצענו בדיקה ראשונית של הנתונים, חילקנו את הפיצ'רים לשתי קבוצות: נתונים נומריים (רציפים) ונתונים קטגוריאליים (בדידים). לאחר מכן ביצענו חקירה ראשונית של הפיצ'רים באמצעות שימוש בפקודות מובנות של pandas: df.head().T ו df[columns_numerical].describe().T .

2. עיבוד מקדים ויזואליזציה:

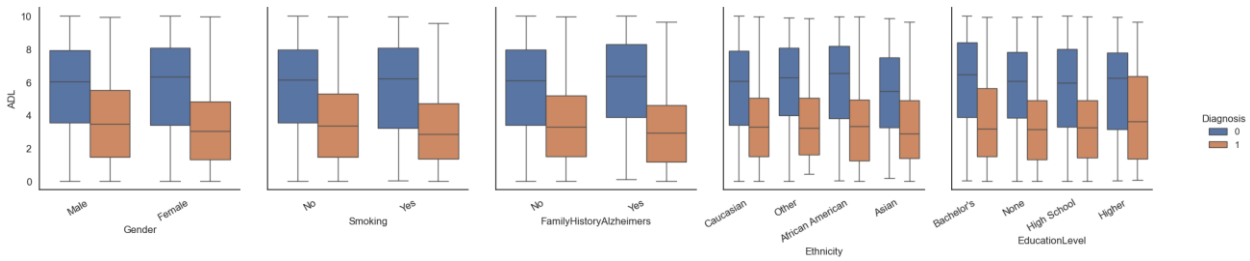
בחלק זה ביצענו מספר ויזואליזציות של הנתונים, ולאחר מכן עיבוד מקדים שלהם כהכנה לשלב הבא:

2.1. היסטוגרמה של הגילאים:



(תרשים מספר 1 : היסטוגרמה של הגילאים)

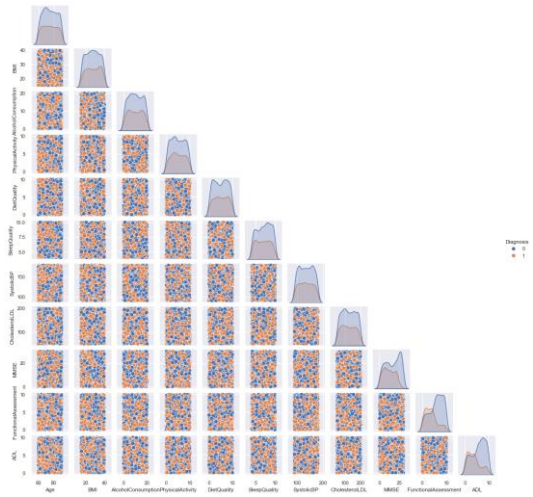
2.2. התפלגות לפי מאפייני אוכלוסייה שונים:



(תרשים מספר 2 : גרף boxplot של מדד ADL)

במדד תפקודי של פעולות היומיום ADL (קישור להסבר על המדד) ניתן לראות שלחולי אלצהיימר יש ציון נמוך בצורה די מובהקת מאשר לאנשים הבריאים, ללא קשר לחלוקה הפנימית בפיצ'רים השונים. כך שניתן לומר באופן ראשוני שמדד ADL יכול להוות פיצ'ר חשוב לניבוי מחלת אלצהיימר.

2.3. ביצוע בדיקה האם יש תלות בין הפיצ׳רים השונים:



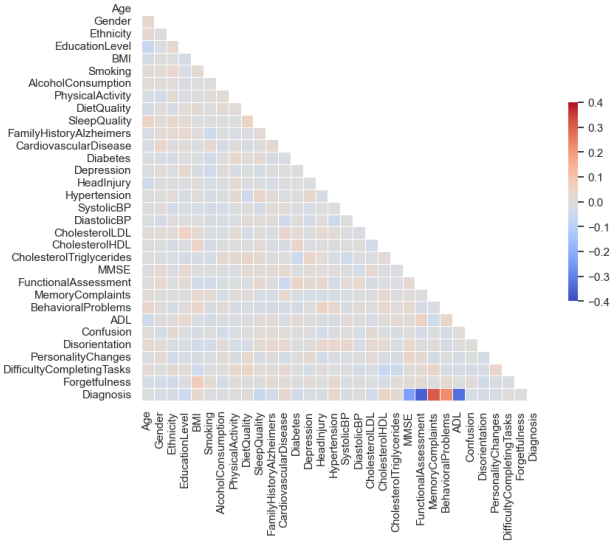
(תרשים מספר 3 : גרף pairgrid של פיצ׳רים נומריים)

ניתן לראות שאין תלות בין הפיצ׳רים הנומריים השונים, כך שניתן להשתמש בכולם.

2.5. נרמול של הפיצ׳רים

ביצענו נרמול של הפיצ׳רים הנומריים באמצעות MinMaxScaler() שנמצא בספריית sklearn.

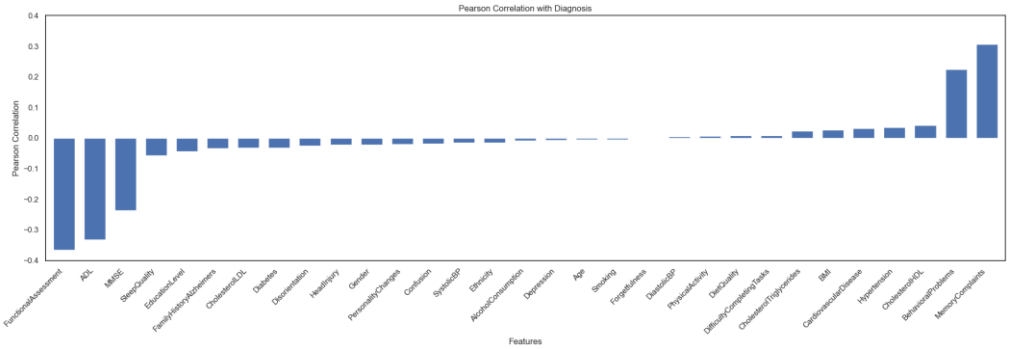
2.6. זיהוי ראשוני של פיצ׳רים מרכזיים:



(תרשים מספר 4 : גרף heatmap של פיצ׳רים נומריים)

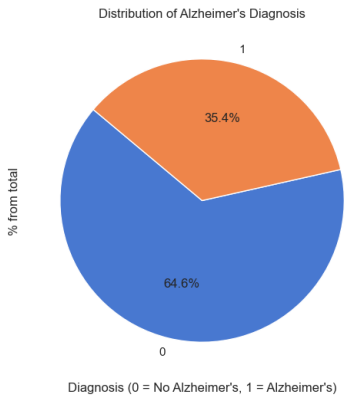
ניתן לראות בצורה ויזואלית שיש חמישה פיצ׳רים מרכזיים התורמים בצורה הגבוהה ביותר לניבוי של המחלה.

2.7. כימות ראשוני של פיצ׳רים מרכזיים:



(תרשים מספר 5 : גרף עמודות של כימות פיצ׳רים נומריים)

2.8. בדיקת איזון של הדאטה:



(תרשים מספר 6 : גרף עוגה של אחוז החולים בנתונים)

ניתן לראות שמתוך הנבדקים כ 65% היו בריאים וכ35% היו חולים. (נשתמש בכך בהמשך באמצעות השוואת המודלים השונים ל Dummy classifier). יש לציין, שרוב הנתונים ב dataset הוא של אנשים בריאים , כך שיש לשים לב שה dataset מוטה במידה נמוכה מסוימת.

2.9. פיצול של הדאטה ל X ו Y :

X מכיל את הפיצ'רים השונים ו Y הוא העמודה של 'Diagnosis' שהוא משתנה בינארי.

3. פיצול ל train ו test:

ביצענו פיצול ל x_train, x_test, y_train, y_test באמצעות הפונקציה המובנת בספריית sklearn.

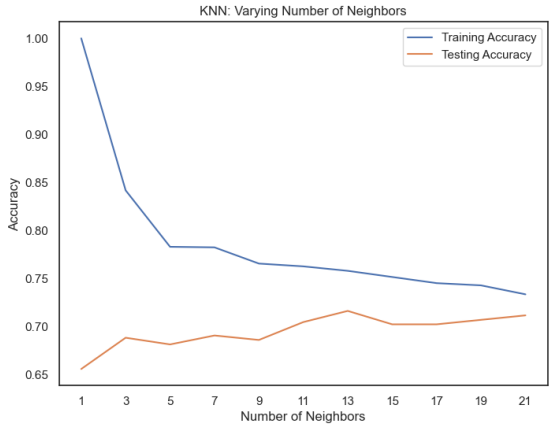
למען הפשטות בקוד, וגם בגלל ש dataset קטן יחסית והרצון לשמור כמות סבירה של נתונים עבור ה test, חילקנו את ה dataset לשני חלקים כאשר 80% הוא train ו 20% הוא test (בהמשך נבצע גם פיצול ל 3 חלקים : train , validation , test עבור המודל הטוב ביותר)

4. בחירת ראשונית של מודלים:

התחלנו עם מספר מודלים פשוטים :

במודל Dummy Classifier כצפוי לפי החלוקה היחסית בין בריאים לחולים (ראה סעיף 2.8) התקבל : accuracy=0.64.

במודל KNN ביצענו כיוונון של היפר-פרמטר של מספר השכנים :

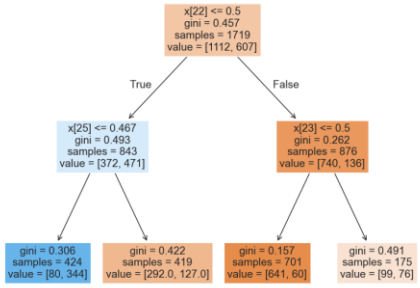


(תרשים מספר 7 : גרף accuracy של נתוני אימון ונתוני מבחן במודל KNN)

התקבל ש K=7 הוא האופטימלי. ו accuracy=0.70 (לא משהו בכלל, רק קצת יותר מה Dummy)

(את כל מדדי הדיוק השונים של המודלים השונים נציג בהמשך בטבלה מרכזת).

מודל נוסף שבדקנו הוא מודל עץ החלטה (Decision Tree) השתמשנו בעומק עץ שדה ראשוני של 2 והתקבל accuracy=0.76 (בהמשך ניישם פעם נוספת את המודל הזה עם כיוונון של ההיפר-פרמטר), העץ שמתקבל הוא :

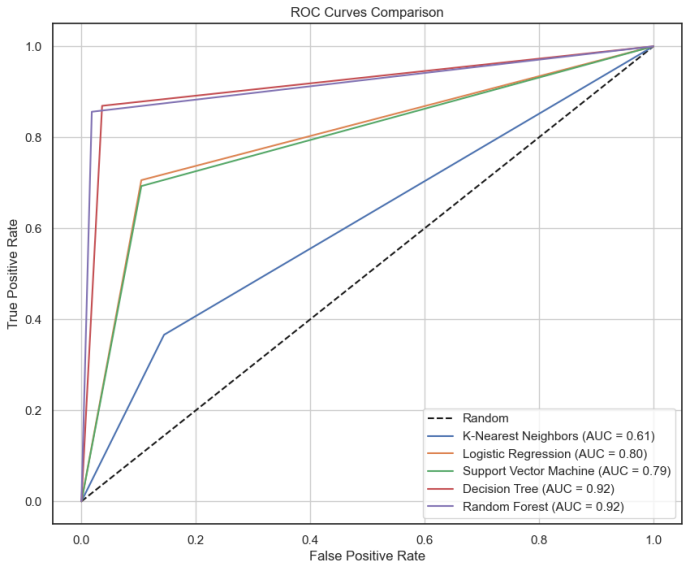


(תרשים מספר 8 : מודל Decision Tree, עומק עץ שווה ל 2)

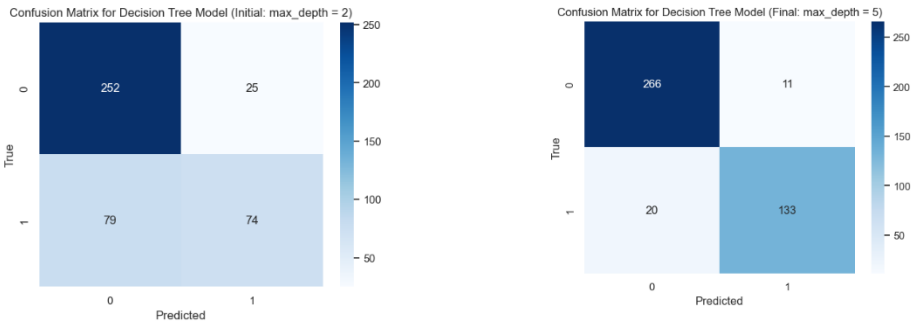
5. בחירת מודלים נוספים והשוואה בין המודלים:

לאחר מכן יישמנו מספר מודלים נוספים תוך תהליך מובנה של כיוונון היפר-פרמטרים שונים באמצעות Grid Search . (לערכי ההיפר-פרמטרים שנבדקו ראה צילום מסך בנספח). תוך כדי ריצת הקוד של ה Grid Search חושבו גם מדדי הדיוק השונים והתבצעה בניית גרף ROC. מצורפת טבלה מרכזת למדדי הדיוק במודלים השונים :

	Accuracy	Precision	Recall
Dummy Classifier	0.64	0.00	0.00
KNN	0.68	0.58	0.37
Decision Tree (max_depth=2)	0.76	0.75	0.48
Logistic Regression	0.83	0.79	0.71
Support Vector Machine (SVM)	0.82	0.79	0.69
Decision Tree (max_depth=5)	0.93	0.93	0.87
Random Forest	0.94	0.96	0.86



(תרשים מספר 9 : גרף ROC להשוואת המודלים השונים)



(תרשים מספר 10 : השוואת confusion matrix למודל Decision Tree)

כסיכום, ניתן לראות מגרף ה ROC את התוצאות הבאות: מודל KNN נמצא בתחתית הרשימה, בטווח הביניים נמצאים המודלים Logistic Regression ו SVM , והמודלים הטובים ביותר הם Decision Tree (לאחר מציאת היפר-פרמטר אופטימלי) ומודל Random Forest.

עבור מודל Decision Tree הצגנו גם בצורה ויזואלית את confusion matrix עבור שני ערכי היפר-פרמטר שנבדקו. בהשוואה של שני ה confusion matrix (תרשים מספר 10) ניתן לראות שגם בעומק עץ ששווה ל2 וגם בעומק עץ ששווה ל 5 , ה TN הוא יחסית דומה, אבל כיוונון של ההיפר-פרמטר במודל מפחית בצורה משמעותית מאוד את ה FN וה FP.

6. ביצוע cross validation ובדיקות נוספות:

מגרף הROC ניתן לראות שמודל Decision Tree (עם עומק עץ של 5) הוא המודל האופטימלי. על מודל זה ביצענו שלוש בדיקות נוספות:

הבדיקה הראשונה הייתה Cross validation עם KFold=10 . התקבל accuracy=0.94.

כיוון שמדובר ב dataset עם מספר נתונים קטן יחסית ביצענו גם Leave one out (LOO), גם כאן התקבל accuracy=0.94.

לבסוף ביצענו Shuffle Split שמפצל את הנתונים ל train , validation , test . וקיבלנו גם כאן accuracy=0.94.

(קישור להסבר על Shuffle Split , קישור לאנימציה של train-test-validation)

7. נקודות לדיון נוסף:

ישנם מספר נושאים נוספים שאפשר לחשוב עליהם כהרחבה עתידית לפרויקט חקר זה:

7.1. האם יש מודלים טובים אפילו יותר?

בחיפוש ב Kaggle מצאנו דוגמת קוד עם תוצאת accuracy=0.96 , הוא עשה שימוש Binary Classification לאחר סינון של חלק מהפיצ'רים ושימוש רק ב 10 הפיצ'רים המרכזיים. – קישור לקוד: Kaggle : Binary Classification (96.6% acc)

7.2. האם חסרים פיצ'רים מרכזיים בנתונים:

בשנים האחרונות במחקר הרפואי בנושא עלתה השערה שהמנגנון העומד בבסיס המחלה כולל שקיעה של חלבון בשם עמילואיד-ביתא (Aβ) ברקמת המוח ושל חלבון נוסף בשם טאו. שקיעת חלבון זה היא אירוע מוקדם מאוד בשרשרת התהליכים המובילים למחלת האלצהיימר, וזו מתרחשת עשור או יותר לפני הופעת הסימפטומים הקליניים הראשונים. הגן הידוע ביותר שמגביר את הסיכון לכך הוא הגן .ApoE (גן זה מקודד לחלבון הקשור לנשיאת שומנים וכולסטרול בדם). – קישור להצגת מחקר בנושא.

לפי תיאוריה זאת הוספת פיצ'ר של גן ApoE ופיצ'ר נוסף של מדידת הצטברות עמילואיד-ביתא (Aβ) , יכולה להוות אולי ניבוי טוב בהרבה של מחלת אלצהיימר (ועוד מספר שנים לפני שמתרחשת ירידה במדדי התפקוד כמו ה ADL שנמצאים ב dataset)

7.3. האם יש שיטות נוספת לניבוי אלצהיימר:

שיטה אפשרית היא שימוש בסריקות MRI לזיהוי המחלה ,
כך למשל מצאנו ב Kaggle דאטה סט (Kaggle: MRI and Alzheimers) שעושה שימוש בנתוני פרויקט OASIS:

The Open Access Series of Imaging Studies (OASIS) is a project aimed at making MRI data sets of the brain freely available to the scientific community. By compiling and freely distributing MRI data sets, we hope to facilitate future discoveries in basic and clinical neuroscience.

שיטה נוספת אפשרית היא שימוש בניתוח כתב יד: Kaggle: Handwriting Data to Detect Alzheimer's Disease

נספחים:

נספח 1:

הוראות לטעינת ה dataset לקובץ הקוד:

יש להוריד את ה dataset מאתר Kaggle, Alzheimer's Disease Dataset. את קובץ ה CSV יש למקם באותה תיקייה של מחברת הקוד.

נספח 2:

מסמך הצעת הפרויקט:

הקדמה:

אלצהיימר היא מחלה ניוונית של תאי העצבים במוח, הגורמת לירידה מתמשכת בזיכרון, בכישורי השפה ובתפקוד הקוגניטיבי. המחלה מחמירה בהדרגה בטווח של 3 עד 8 שנים עד למוות של החולה. בעולם יש כ 45 מיליון חולים (בישראל כ 100 אלף חולים). המחלה מהווה אתגר משמעותי בתחום הבריאות הציבורית והבנת גורמי הסיכון וההתקדמות שלה הוא קריטית לצורך זיהוי מוקדם וטיפול מותאם אישית לחולה.

תיאור ה dataset:

מקור הנתונים הוא:

https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset/data

הנתונים כוללים 2149 נבדקים, עבור כל אחד מהם קיימת הבחנה האם הוא חולה (כן או לא), קיימים סך הכל 32 פרמטרים שונים לגבי כל אחד מהנבדקים לפי הפירוט הבא: (הפרמטרים הם מסוגים שונים: רציף, בדיד, בינארי, נומינלי, אורדינלי).

- מדדים דמוגרפיים: גיל, מגדר, מוצא, רמת השכלה.
- מדדי אורח חיים: מדד BMI, האם מעשן, כמות צריכת אלכוהול, רמת פעילות גופנית, רמת התזונה, איכות השינה.
- מדדי היסטוריה רפואית: היסטוריה משפחתית של אלצהיימר, רקע של: מחלות לב וכלי דם, סוכרת, דיכאון, פגיעת ראש, יתר לחץ דם.
- מדדים רפואיים: שני מדדי לחץ דם (SystolicBP, DiastolicBP), ארבעה מדדי כולסטרול (CholesterolTotal, CholesterolTriglycerides, CholesterolLDL, CholesterolLDL).
- מדדים קוגניטיביים: ציון קוגניטיבי, ציון הערכה תפקודית, בעיות זיכרון, בעיות התנהגות, מדד ADL.
- מדדי סימפטומים: בלבול, חוסר התמצאות, שינוי באישיות, קושי בביצוע משימות, שכחה.

שאלות המחקר:

1. classification: האם ניתן לסווג שאדם חולה באלצהיימר בהינתן ערכי פרמטרים מסוימים?
2. Logistic regression: מהו הקשר בין הפרמטרים להסתברות להיות חולה במחלה?
3. אחוז הדיוק במודלים: מהו אחוז הדיוק לאבחון המחלה במודלים שונים של למידת מכונה?
4. מודלים חלקיים: האם הוצאת פרמטרים מסוימים (כמו למשל עבור כלי ניבוי ראשוני באוכלוסייה כללית ללא מדדי סימפטומים מוקדמים) מפחית בצורה משמעותית את דיוק אבחון המחלה? ואם כן בכמה?

נקודות נוספות:

- הפרויקט יכלול שימוש במודלים שיילמדו במהלך הקורס, יחד עם השוואה למודלים שפותחו והועלו לאתר kaggle על ידי משתמשים אחרים.
- דיון סיכום האם חסרים ב dataset פרמטרים חשובים שיאפשרו אבחנה מדויקת יותר?
- שימוש בכלי ויזואליזציה של פייתון להצגת קשרים בין הפרמטרים, מדדי פיזור שלהם ולבסוף תוצאות השוואה בין מודלים שונים.