אוניברסיטת בן גוריון בנגב

הפקולטה להנדסה

המחלקה להנדסת תעשייה וניהול

קורס מערכות לומדות וכריית נתונים
דוח מסכם פרויקט

ALS Detection from EHR

מגישים:

גלעד ארז, 314898453

משה כהן, 316161694

מרצה: פרופ' בעז לרנר

תאריך הגשה: 23/02/2023

תקציר

פרויקט זה עסק בבעיית סיווג של חולי ALS מתוך אוכלוסיית מבוטחי מערכת הבריאות בישראל. זאת, בעזרת ניתוח נתוני הרשומה הרפואית (EHR) של המבוטחים תוך התמקדות במדדים הנאספים בבדיקות מעבדה לאורך השנים. בפרויקט ביצענו ניסוי למציאת טווחי הזמנים הרלוונטיים ביותר לאימון מודל סיווג והפקנו תובנות משלבי חקר הנתונים ועיבודם. בשלב הבנת הנתונים (Data understanding) זיהינו תופעות מעניינות שבאות לידי ביטוי בבדיקות המטופלים (חולים ובריאים) תחת חתכים של מגדר ומעמד סוציואקונומי. זהינו שלעיתים חיתוכים אלה שופכים אור על סיבות אפשריות להתקדמות ואופי המחלה, אך במקביל הצבענו גם על מספר בעיות שעלולות לנבוע מהסקה ויזואלית כזו לאור אופי הנתונים. האתגר המרכזי בו נתקלנו היה הכנת הנתונים (Data preparation) לצורך אימון המודל. בחרנו להשתמש במודל שאינו טמפורלי, אך תוך מתן ביטוי למימד הזמן, המשמעותי בבעיה זו. לשם כך, החלטנו לייצר מהרשומות הרפואיות מומנטים שיבטאו את התנהגות התכונות לאורך הזמן, ולאחר מכן להזין אותם למודל כמאפיינים שטוחים.

המסקנה העיקרית היא שקיים קושי במציאת חוקיות לגבי שיטת חיתוך טובה, אך ניתן להצביע על מספר צורות חיתוך שמביאות תוצאות טובות יותר או פחות בהשוואה לאחרות. כמו כן, האלגוריתם לבחינת איכות החיתוכים עשוי להיות רלוונטי בפרויקטים נוספים תוך התאמתו לנתונים המגיעים מעולם תוכן אחר.

תוכן עניינים

Business Understanding

מחלת ה-ALS (Amyotrophic Lateral Sclerosis) היא מחלת ניוון שרירים השייכת לקבוצת המחלות הניווניות הפוגעות בתאי העצב המוטוריים האחראיים על תנועת השרירים. המחלה היא נדירה וחשוכת מרפא שגורמת בהדרגה לשיתוק כל השרירים בגוף. המחלה בעלת אופי הטרוגני, כך שקיימים הרבה פרופילים לקצב התפתחות המחלה, ואין הרבה ידע מדעי הקושר גורמים ברורים המגדירים קבוצות סיכון ייעודיות. לכן, האבחון של המחלה נעשה באמצעות זיהוי סימפטומים, ללא בדיקה ייעודית. כיום, לא קיימת תרופה או דרכי טיפול לשיפור מצב החולה במחלה.

זיהוי מוקדם של המחלה או התרעה על סיכוי גבוה לחלות בה בעתיד, יכול לתרום רבות לתהליך הטיפול של החולה. זיהוי מוקדם יכול לקצר משמעותית את זמן האבחון ולשפר את איכות החיים של המטופל תוך שיפור סביבת המגורים שלו והתאמה מראש של ציוד רפואי, זאת יחד עם הכנה מנטלית לתקופה המאתגרת שהוא עתיד לעבור.

פרויקט זה יעסוק בניתוח הנתונים של הרשומה הרפואית הדיגיטלית (EHR) של מטופלים רבים, לצורך הבנה מהם מקטעי הזמן לאיסוף מידע האפקטיביים ביותר לצורך סיווג מבוטחים כחולים במחלת ה-ALS.

Data Understanding

<u>Electronic Health Records</u>

לצורך פרויקט זה קיבלנו רשומות רפואיות מקופת החולים מאוחדת. הרשומה הרפואית מספקת מידע נרחב על המטופלים וכוללת נתונים דמוגרפיים, אישיים, בדיקות מעבדה, אבחונים, מרשמים, ביקורים במרפאה ועוד. על מנת למקד את המחקר החלטנו להתמקד בנתונים הנאספים בבדיקות המעבדה. זאת, מתוך הבנה שבדיקות המעבדה מתבצעות לאחר ביקור המטופל אצל הרופא, ועל בסיס בדיקות אלה הרופא מאבחן את המטופל ונותן לו מרשמים במידת הצורך. כלומר, ביקורי רופא, אבחנות ומרשמים הן נגזרות של בדיקות המעבדה, ועל כן תוצאות הבדיקות עצמן מספקות מידע גולמי שדרכו ניתן ללמוד על התפתחות המחלה אצל המטופלים.

בפרויקט קיבלנו את הרשומות הרפואיות של 162 מטופלים שאובחנו במחלת ה-ALS. לכל מטופל הוצמדה קבוצת ביקורת של ארבעה אנשים נוספים, כך שהאנשים שנבחרו כביקורת למטופל מסוים - הם בעלי אותם מאפיינים אישיים (מגדר וגיל). ההצמדה נועדה לבודד את תכונת ה"חולי" מתכונות נוספות שעשויות להשפיע על הרשומה של שני אנשים, דבר שעלול לפגום ביכולת הלמידה של המודל (למשל, לא נרצה להשוות חולה בן 60 לבחורה בריאה בת 30). הנתונים שהתקבלו נאספו על המטופלים עד 16 שנים לפני האבחון שלהם במחלה.

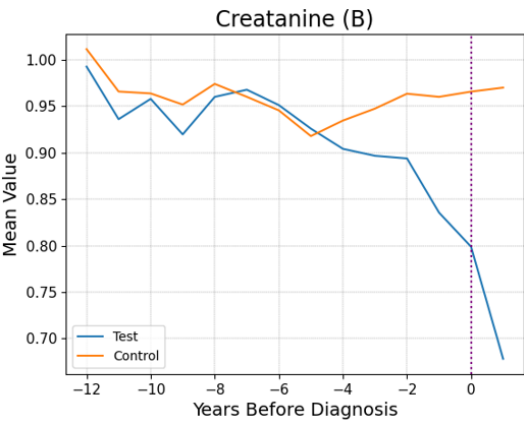
לבסיס נתונים זה שני אתגרים מרכזיים:

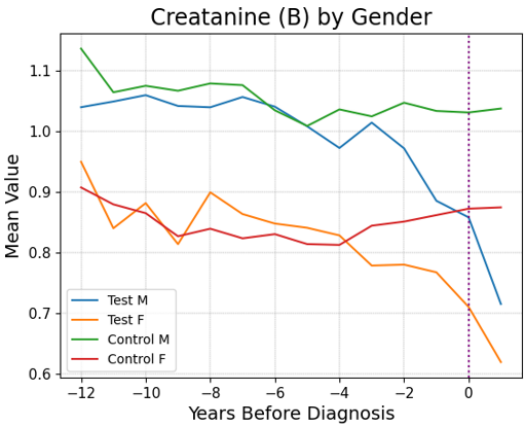
- סט הנתונים לא מאוזן - קיימת קבוצת מיעוט שאותה אנו מעוניינים לזהות.
- כמות בדיקות המעבדה והמרווח בין הבדיקות הם שונים עבור מטופלים שונים. השונות בתדירות הבדיקות מעלה בעיות שידרשו תשומת לב בשימוש במודלים טמפורליים המתקשים בהתמודדות עם מרווחי זמן שאינם קבועים. בעיה נוספת שיכולה להתעורר היא שלמטופל שנבדק יותר, עשויים להיות מומנטים יותר יציבים (אפקט של ריבוי data points בתחום זמן מסוים), שיצרו הבדל מוטה בין מטופלים בהתאם לתדירות הבדיקות (משתנה מתערב, confounding variable).

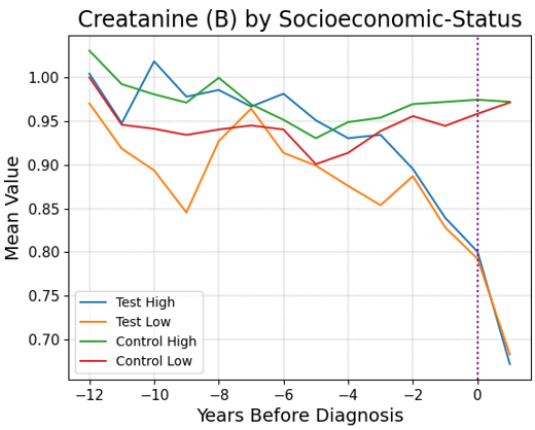
<u>התבוננות ויזואלית</u>

כשלב מקדים, חקרנו את הנתונים ברמה הכללית על מנת לזהות מגמות ומאפיינים שייתכן וימצאו כרלוונטים עבור אימון המודל. כמו כן, בדקנו אם ישנן אינטרקציות בין בדיקות המעבדה למאפיינים מגדר ומצב סוציואקונומי שייתכן ומהווים מידע אישי חשוב.

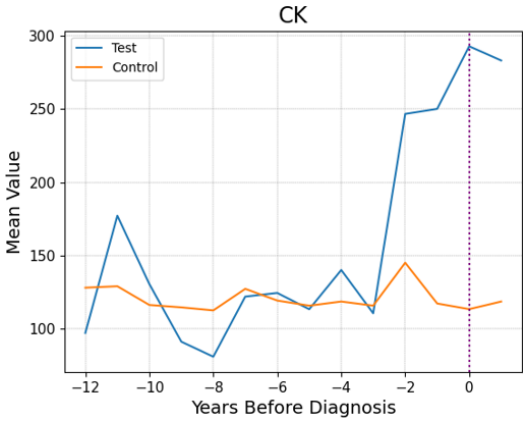
לדוגמה, עבור קריאטנין B (Creatanine B, מדד לתפקוד הכליות שירידה שלו עשויה להעיד על ירידה במסת שריר), ניתן לראות כי כשנתיים לפני האבחון יש ירידה משמעותית במדד לקבוצת החולים ללא תלות במגדר ובמצב הסוציואקונומי.

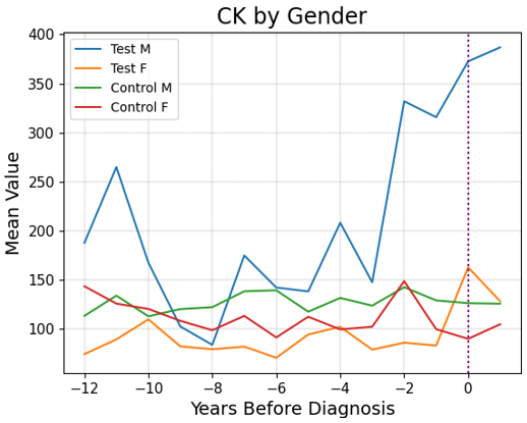


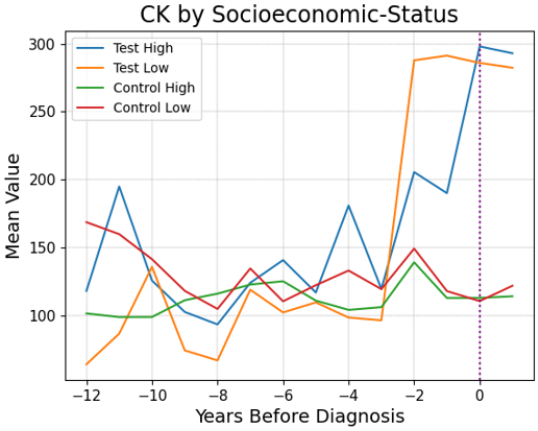




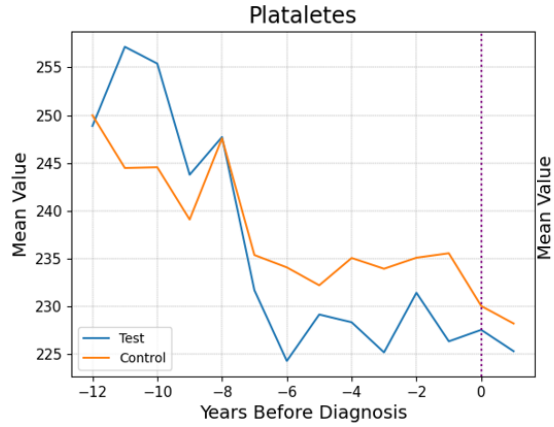
דוגמה נוספת היא האנזים קריאטין קינאז (Creatine Kinase) שדולף מהשרירים כשהם נפגעים ומתפרקים. ניתן לראות כי כשנתיים לפני האבחון יש עלייה משמעותית במדד אצל החולים. כמו כן, ההפרדה נובעת מעלייה משמעותית אצל הגברים החולים (המדד של הנשים החולות נותר זהה בהשוואה לנשים בריאות). ייתכן וניתן לייחס זאת למבנה הגוף הגברי השרירי יותר ממבנה הגוף הנשי. בנוסף, לא נצפה הבדל בולט בין הרמות הסוציואקונומיות.

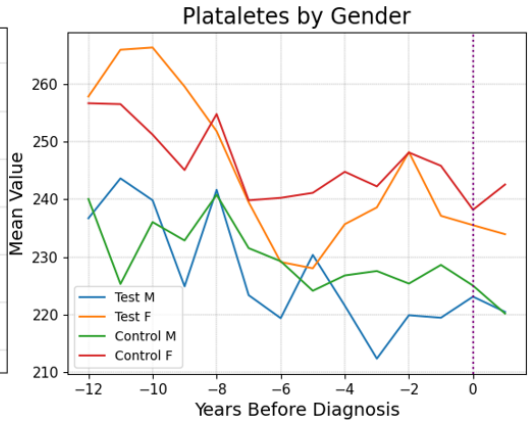


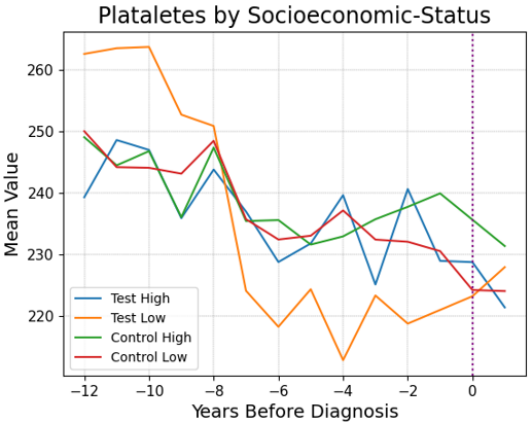




דוגמה נוספת היא מדד plataletes (טסיות דם) שערכו יורד ככל שמתבגרים. ניתן לראות כי יש ירידה עבור שתי הקבוצות, אך ירידה משמעותית יותר עבור החולים (כשבע שנים לפני האבחון). ככל הנראה ירידה זו נובעת מחולים בעלי מצב סוציאקונומי נמוך.

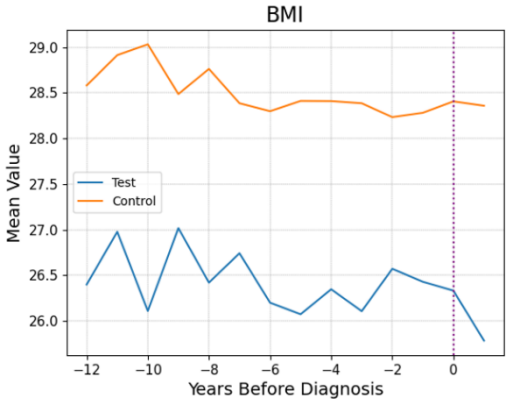


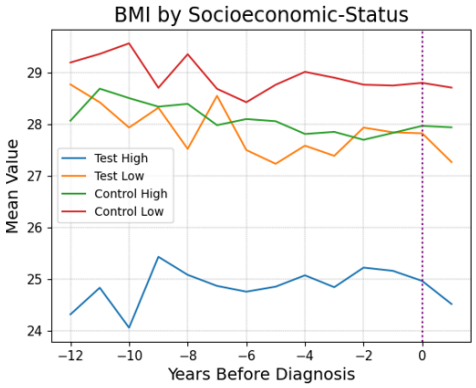


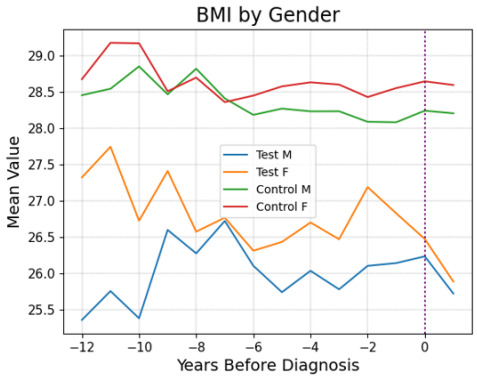


מדד מעניין נוסף הוא ה-BMI (שקלול של גובה ומשקל). ידוע שעם התפתחות המחלה חלה ירידה במשקל, אך לא ראינו ירידה בולטת כזו בתרשימים, מלבד בשנים הקרובות לאבחון. אנו משערים שהירידה במשקל חלה בשנים מאוחרות יותר מאלה שאנחנו בוחנים. כמו כן, ההפרדה במדד זה קיימת לאורך כל 12 השנים בין החולים לבריאים. הופתענו למצוא שההבדל בין האוכלוסיות נובע בעיקר מה-BMI הנמוך של חולים במעמד סוציואקונומי גבוה. השערה שעלתה בנושא היא שחולים במעמד סוציואקונומי נמוך "משלימים" את המשקל בעזרת תזונה עתירת קלוריות, שידועה כנפוצה יותר באוכלוסייה זו.

שאלה נוספת שהתעוררה כתוצאה מהתבוננות בגרפים האלה היא בנוגע להסקה על הסיבתיות. משום שההפרדה חזקה גם 12 שנים לפני האבחון, קיים קושי בהבחנה האם BMI נמוך הוא גורם סיכון ל-ALS או שזוהי תוצאה של דגירת המחלה אצל המטופל.

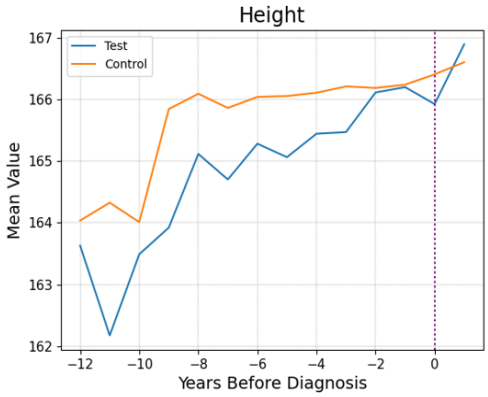


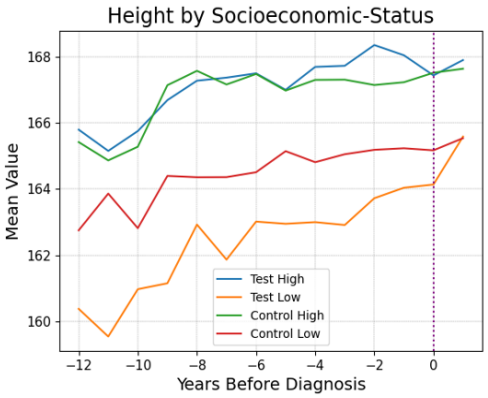


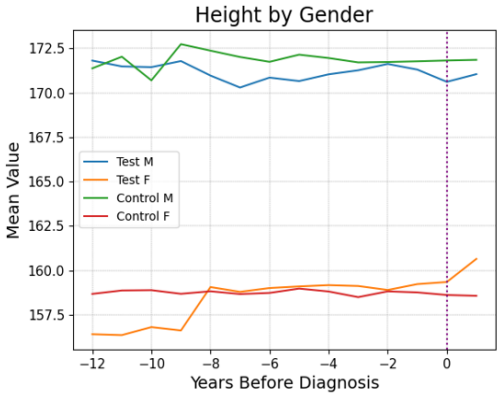


תוצאה שהפתיעה אותנו היא הגרף שמציג את מדד הגובה לאורך השנים. ראינו כי לאורך הזמן חלה עליה בגובה המבוטחים החולים והבריאים. כדי להסביר את העלייה הזו, בדקנו אם במדגם יש אנשים בגילאי ההתבגרות, אך ראינו שהגיל המינימלי הוא 21. לאחר מכן, בדקנו אם ישנן שגיאות מדידה חמורות שיכלו לגרום לעליה כזו, וראינו הבדלים של סנטימטרים בודדים בין מדידה למדידה (רעש אקראי) שאמורים לבטל זה את זה בעת המיצוע.

לבסוף, מצאנו את מקור העלייה - הוספה של נבדקים למדגם. משום שהנתונים של הנבדקים נאספו בתקופות שונות, המדגם ב-12 שנים קטן בהרבה מהמדגם לקראת האבחון. מסיבה כלשהי, ממוצע הגבהים של המטופלים ה"ותיקים" היה נמוך מהממוצע הכללי ומכאן נבעה העלייה. בעיה זו שמה בספק כל מסקנה שנסיק מהגרפים האלה, בעיקר בשנים המוקדמות – שכן המדגם שם הוא חלקי בלבד וגם נתון להשפעות רעש מוגברות.







סטטיסטיקה תיאורית זו תרמה בהבנת הנתונים לפני תהליך הכנת הנתונים וחילוץ המאפיינים המשמעותיים. עבור מדדים שונים נמצאו אינטראקציות עם המשתנים מגדר ומצב סוציואקונומי, ולכן בהמשך נתייחס עליהם כמאפיינים חשובים בעת אימון המודל.

כאמור, לגישה זו שני קשיים עיקריים עבור בסיס הנתונים עליו עבדנו בפרויקט. דרך גרפים אלו לא ניתן לדעת מהו כיוון ההשפעה של המאפיינים (הסיבתיות. לדוגמה, האם BMI נמוך גרם ל-ALS, או שה-ALS גרם לBMI נמוך). בנוסף, משום שהמדד מחושב ע"י ממוצע של כלל המטופלים בנקודת זמן מסוימת, גודל המדגם משתנה לאורך הגרף. ייתכן וחלק מהמגמות שזוהו בגרפים אלו מושפעות יתר על המידה משינוי בגודל המדגם ולכן יש להתייחס אליהן בקפדנות תוך חשיבה ביקורתית

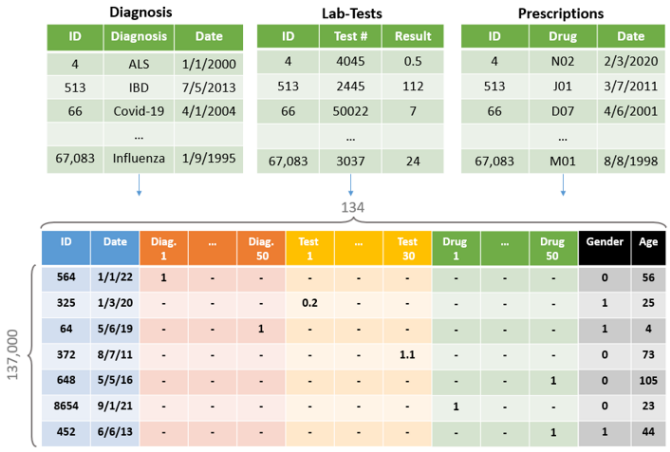
(לדוגמה, הגובה הממוצע גדל ככל שהזמן מתקדם - ייתכן ועם השנים נוספו לנתונים אנשים גבוהים שהגדילו את ממוצע הגובה במדגם).

Data Preparation

מאמץ רב הוקדש בהכנת הנתונים לקראת אימון אלגוריתם הסיווג. כעת, נמחיש את התהליך שהתבצע על הנתונים הכולל בחירת מאפיינים, ניקוי הנתונים והשלמת ערכים חסרים תוך הסבר השיקולים וההחלטות שהתקבלו במהלך התהליך.

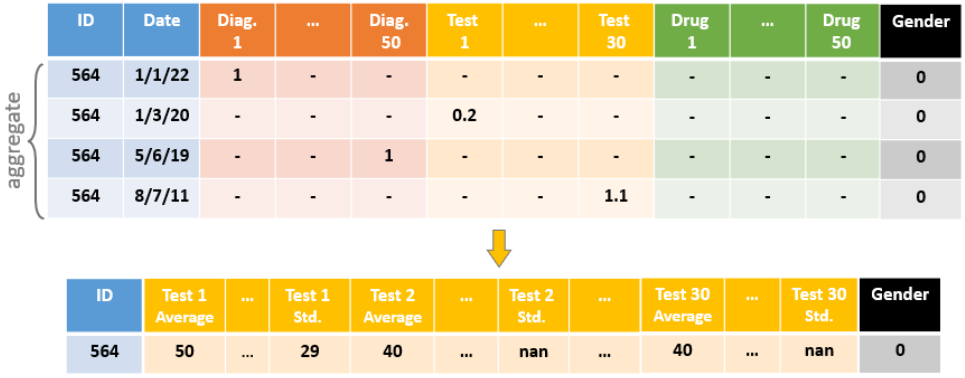
<u>שלב ראשון</u> - איחוד הנתונים מהמקורות השונים.

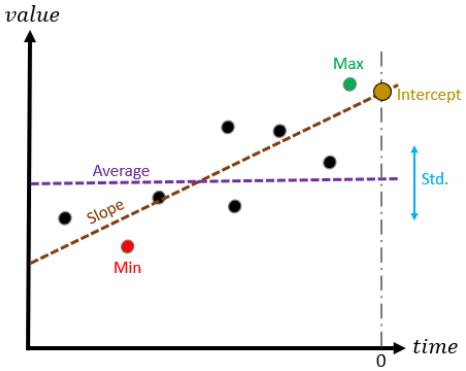
אוחדו הנתונים של הרשומות הרפואיות של המטופלים הכוללות בדיקות מעבדה, מרשמים ואבחונים. קיבלנו טבלה אחת מאוחדת שהמימדים שלה מאוד גדולים. כל רשומה מייצגת נקודת זמן מסוימת שבה התקבלו תוצאות בדיקת מעבדה, התקיים ביקור, נרשמה תרופה וכו. משום שלאורך השנים לכל מטופל התקיימו מאות פעולות רפואיות שכאלה, התקבלו כ-137,000 רשומות. בנוסף, משום שכל רשומה מייצגת פעולה בודדת, ובטבלה מאות עמודות, לאחר האיחוד התקבלו הרבה ערכים חסרים.



<u>שלב שני</u> - חילוץ מומנטים ואגריגציה.

בכדי לתת ביטוי לזמן, המהווה מרכיב חשוב בזיהוי התקדמות המחלה, עבור כל מאפיין של בדיקת מעבדה (Test) חולצו שישה מומנטים המבטאים את השינוי של המאפיין לאורך הזמן. המומנטים שנבחרו הם ממוצע, סטיית תקן, ערך מקסימלי, ערך מינימלי, שיפוע קו רגרסיה וחותך רגרסיה. בדרך זו ביצענו אגריגציה לכל הרשומות של כל מטופל.





<u>שלב שלישי</u> - ניקוי רשומות ומאפיינים דלילים.

עקב מחסור בבדיקות מעבדה למטופלים מסוימים, ודרישה למספר מינימלי של דרגות חופש עבור מומנטים ספציפיים (לדוגמה, שיפוע קו רגרסיה דורש לפחות שתי תצפיות של המאפיין בזמנים שונים) התקבלו רשומות ועמודות רבות בהם הרבה ערכים חסרים. החלטנו להסיר רשומות ועמודות כך שמספר הערכים החסרים בהם היה גבוה מרף שהוגדר. המטרה הייתה להכין את הנתונים לשלב השלמת הערכים, כך שלא יהיו רשומות שרוב התאים בהם יושלמו באופן מלאכותי שייתכן ויוביל להטיה של המודל.

<u>שלב רביעי</u> - השלמת ערכים חסרים (Imputation).

לאחר ביצוע האגריגציה צורפו לטבלה גם המשתנים הדמוגרפיים אותם בחרנו לבחון בשלב ה- Data Understanding (מגדר ומצב הסוציואקונומי), לאחר שראינו שהם מקיימים אינטרקציות מעניינות עם בדיקות המעבדה. השלמת הערכים החסרים

התבצעה על ידי אלגוריתם Iterative Imputer המבוסס על רגרסיה ליניארית. האלגוריתם משלים כל ערך חסר ע"י חיזוי רגרסיה, כשיתר הערכים ברשומה משמשים כמשתנים מסבירים.

ID	Test 1 M1	...	Test 1 M6	Test 2 M1	...	Test 2 M6	...	Test 30 M1	...	Test 30 M6	Gender	Socio. Status	Group	Control ID
564	50	...	29	40	...	nan	...	40	...	nan	0	7	1	564
253	48	...	25	63	...	43	...	40	...	33	1	3	0	243

ID	Test 1 M1	...	Test 1 M6	Test 2 M1	...	Test 2 M6	...	Test 30 M1	...	Test 30 M6	Gender	Socio. Status	Group	Control ID
564	50	...	29	40	...	44	...	40	...	35	0	7	1	564
253	48	...	25	63	...	43	...	40	...	33	1	3	0	243

Modeling

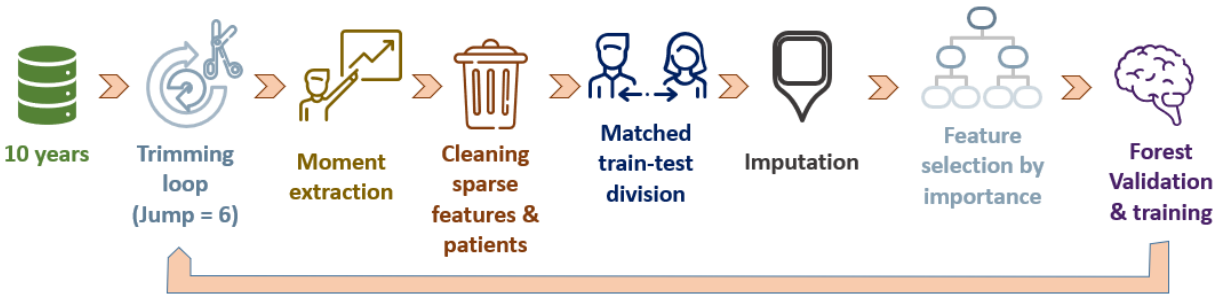
משום שעל נתוני ה-ALS בוצעו כבר מספר פרויקטים שהניבו תוצאות סיווג יפות, החלטנו (בעצתו ואישורו של בעז) לבחון את חשיבותו של מקטע הזמן שישמש ללמידת המודל. לשם כך, תכננו ניסוי שיבחן את ביצועיו של מודל יער החלטה רנדומלי, אליו יוזנו נתוני train מזמנים שונים בהתפתחות המחלה.

Trimming Test

כעת, כשניתן לאמן מודל הלומד את הנתונים לצורך סיווג תצפיות חדשות כחולים או בריאים, מטרתנו היא לבחון כיצד מקטעי זמן שונים של הנתונים משפיעים על דיוק המודל שיבנה. לצורך כך, ביצענו תהליך איטרטיבי שבו כל איטרציה משתמשת בנתונים מזמנים שונים (בקפיצות של חצי שנה, Jump=6). כלומר, בכל איטרציה נבנה מודל שאומן על נתונים שונים בהתאם למרווחי הזמן שהוגדרו (עד עשר שנים לפני האבחון).

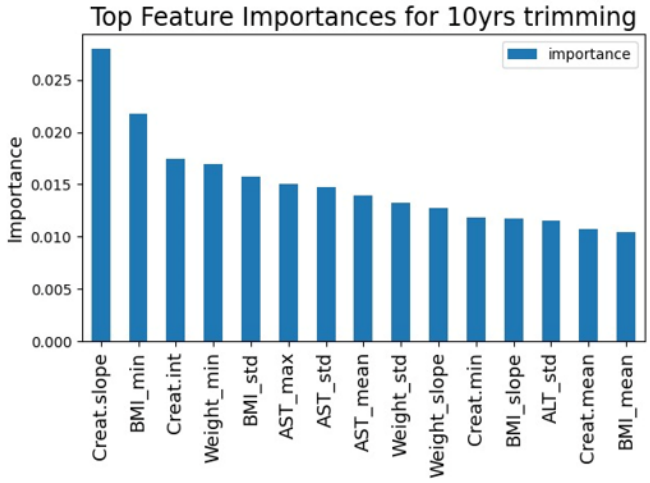
בכל איטרציה התבצע התהליך הבא: בחירת הרשומות הרלוונטיות לתקופת הזמן שנבחרה, חילוץ מומנטים, הסרת רשומות עם יותר מדי ערכים חסרים, פיצול הסט לאימון ובחינה (כך שמטופלים יצוותו לקבוצת ה-train או ה-test יחד עם הבריאים שהוצמדו להם), השלמת הערכים החסרים, ובחירת 50 המאפיינים החשובים ביותר על פי מדד features importance, של יער החלטה (random forest) שנבנה במיוחד לצורך זה בלבד.

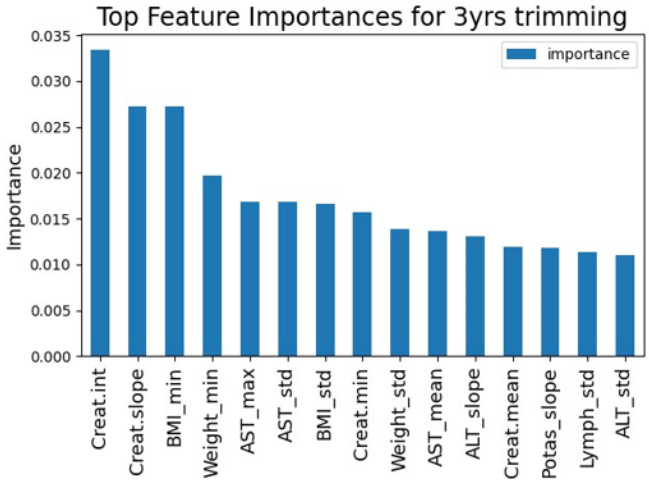
לאחר תהליך זה, אומן המודל על בסיס הנתונים והמאפיינים שנבחרו. זאת, לטובת ולידציה של יער החלטה שבסופו נבחר היער הטוב ביותר למקטע הזמן המסוים, וביצועיו נבחנו בהשוואה למקטעים אחרים. כל איטרציה בוצעה 5 פעמים, כך שגם תהליך הכנת הנתונים הושפע מרנדומליות. כלומר, מודלים שאומנו על אותו חיתוך זמן, עדיין קיבלו 5 סטים שונים (חלוקה רנדומלית לtrain, test, השלמת ערכים חסרים בסדר רנדומלי וכו...). בצורה זו הקטנו את השפעת המקריות בניסוי והתקרבנו לביטוי מדויק יותר של טיב החיתוכים השונים.



נראה מדוע קיימת חשיבות לביצוע תהליך בחירת המאפיינים החשובים (features importance) בכל איטרציה, לעומת בחירת מאפיינים כללית ללא תלות במקטע הזמן.

בגרפים הבאים ניתן לראות את החשיבויות של 15 המאפיינים החשובים ביותר, על פי יער ההחלטה ששימש לבחירת המאפיינים. בגרף משמאל מוצגים המאפיינים המשמעותיים בחיתוך של 10 שנים לפני האבחון, ובימני עבור חיתוך של 5 שנים. משום שהנתונים משתנים בכל איטרציה, דירוג המאפיינים החשובים גם הוא משתנה. ניתן לראות כי עבור מקטעי הזמן השונים, נבחרו מאפיינים שונים וכי גם הדירוג הפנימי של המאפיינים המשותפים שנבחרו הוא שונה. לכן, קיין צורך בבחירה מחודשת של המאפיינים החשובים בכל אחת מהאיטרציות.

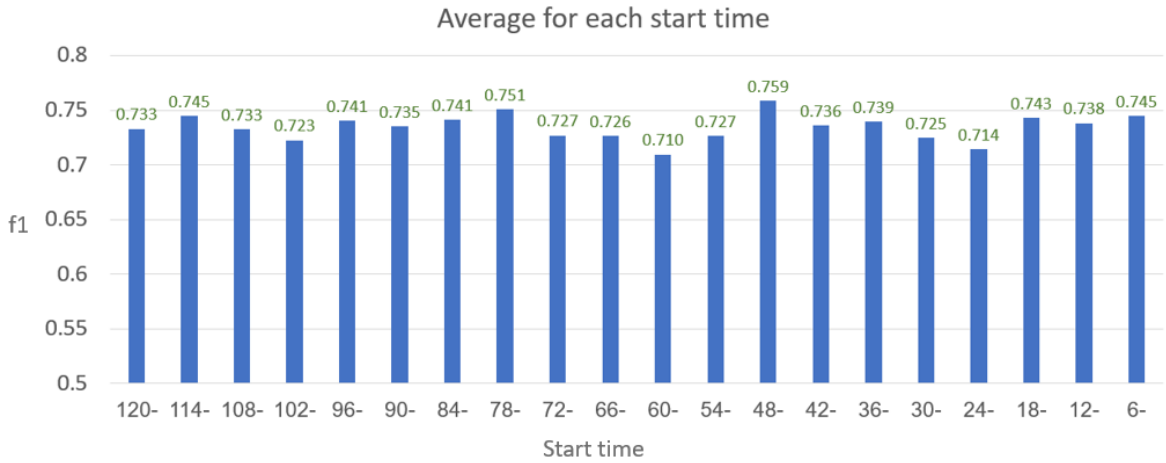




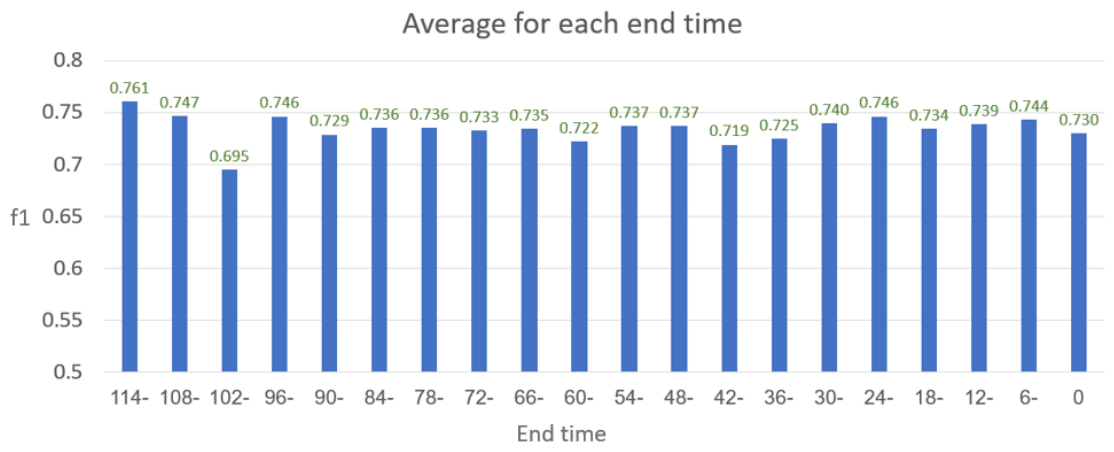
Evaluation

בעת ביצוע ה-Trimming Test נבדק ערך המדד f1 על סט בחינה. המדד מעריך את ביצועי המודל תוך התייחסות לprecision ול-recall, ולכן מתאים להערכת הביצועים על סט נתונים שאינו מאוזן. בשלושת הגרפים הבאים מוצגים ערכי הf1 עבור חיתוכים מסוימים, כאשר כל עמודה מייצגת את הערך הממוצע שהתקבל ממודלים להם תכונה משותפת המוגדרת בכותרת התרשים.

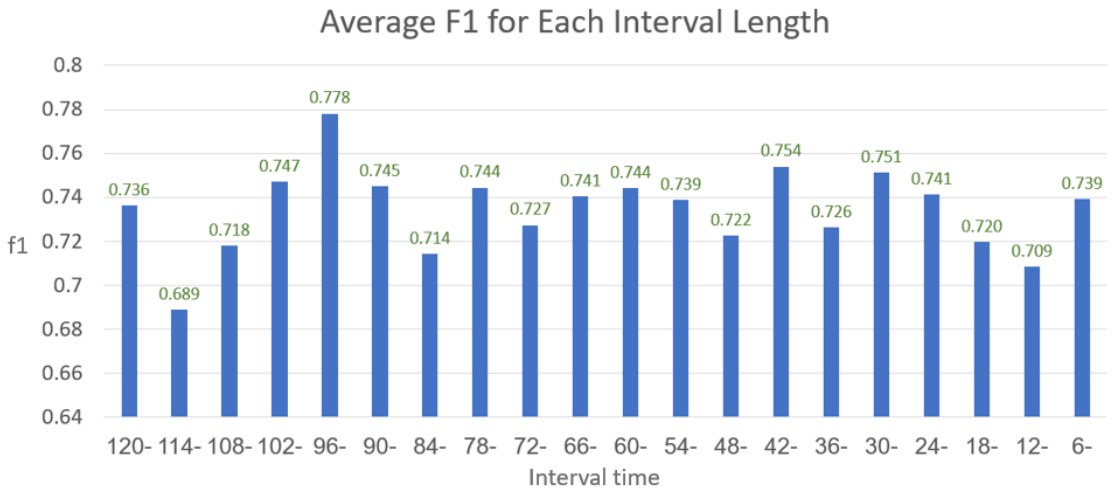
- <u>נקודת התחלה</u> (start time) - נבדק האם קיימת השפעה של נקודת זמן התחלתית (מספר חודשים לפני אבחון המחלה) ללא תלות בנקודת הסיום. כפי שניתן לראות בגרף, לא ניתן להצביע על מגמה, אך ניתן לראות נקודות התחלה מסוימות בהן אחוזי הדיוק היו גבוהים ביחס לשאר.



- <u>נקודת סיום</u> (end time) - נבדק האם קיימת השפעה של נקודת זמן הסיום (מספר חודשים לפני אבחון המחלה) ללא תלות בנקודת ההתחלה. כפי שניתן לראות בגרף, לא ניתן להצביע על מגמה, וגם פה קיימות נקודות התחלה מסוימות בהן אחוזי הדיוק היו גבוהים ביחס לשאר. בלטו לרעה החיתוכים שנקודת הסיום שלהם 9 שנים לפני אבחון המחלה. מלבד חריגה זו לא נראים הבדלים חריגים בין החיתוכים.



- <u>אורך המקטע</u> (interval length) - נבדק האם קיימת השפעה של אורך מקטע הזמן ממנו חולצו הנתונים ללא תלות בנקודת ההתחלה ובנקודת הסיום. כפי שניתן לראות בגרף, קשה גם כאן להצביע על מגמה ברורה, אם כי ניתן לומר ש-96 חודשים הוא מקטע הזמן שהניב את הערכים הגבוהים ביותר בממוצע. נסיק מכך שבמסגרת הנתונים שלנו וצורת העיבוד שעברו, 96 חודשים הם מרווח הזמן האידיאלי לשם בניית מודל יעיל ומדויק ככל הניתן.



סיכום, דיון, ומסקנות

מטרת הפרויקט היא להצביע על מרווח זמן אידיאלי לשם ניתוח הנתונים וסיווג המבוטחים לבריאים וחולים במחלת ה-ALS. מרווחי הזמן נבחנו בזוויות שונות, תוך הסתכלות על נקודות התחלה, נקודות סיום ואורך מרווחי זמן שונים. לאור התוצאות שקיבלנו, לא ניתן להצביע על מגמות ברורות ואחידות. עם זאת, חיתוכי זמן מסוימים הניבו תוצאות טובות מאחרים. למשל: חיתוכים באורך של 96 חודשים (8 שנים), חיתוכים שמתחילים 48 חודשים לפני האבחון (4 שנים), ועוד.

אנו מעריכים שהסיבה להיעדר המגמה היא הטרייד-אוף בין אחוז הערכים החסרים, לבין מובהקות ההבדלים בין מומנטים של חולים לבריאים. לדוגמה, אם ניקח מקטע ארוך של 10 שנים, נוכל להוציא את רוב המומנטים ולא נצטרך להשלים הרבה ערכים. מצד שני, כשמסתכלים על טווח כה ארוך, ההבדלים במומנטים מצטמצמים ונהיים דומים בעבור חולים ובריאים. לעומת זאת, בטווח של שנה, יהיו המון ערכים חסרים שהשלמה שלהם תיצור הטייה, אך נוכל לזהות הבדלים גדולים ומובהקים במומנטים של שתי המחלקות.

כפי שצפינו בשלב הבנת עולם התוכן ובחירת הנתונים, נתוני בדיקות מעבדה הן אמצעי טוב לסיווג המטופלים. דרך בדיקות אלו, ניתן להבחין בהתקדמות המחלה, ולאורך הזמן ניתן להפריד בין קבוצת הבריאים והחולים. בנוסף, מאפיינים דמוגרפיים כמו מגדר ומצב סוציואקונומי נמצאו כמשפיעים על המדדים, כתלות ברמות השונות שלהם.

בפרויקט הוקדשה חשיבה רבה בתהליך הכנת הנתונים. ראינו שיש חשיבות רבה לאופן עיבוד הנתונים, תוך התייחסות לאיכות הנתונים והשלמת ערכים חסרים בשיטות מורכבות ולא נאיביות. שינויים קטנים שביצענו בעיבוד הנתונים הובילו לשינויים משמעותיים בביצועי המודלים.
להערכתנו, תהליך הכנת הנתונים האיטרטיבי שבוצע ב-trimming test יכול להוות מתודולוגיה עבור בסיסי נתונים אחרים בפרויקטים שונים, לאחר ביצוע התאמות ושיפורים מסויימים תוך הבנה של עולם התוכן המתאים.

<u>רעיונות שיפור להמשך</u>
ניתן לבחון שינויים במודל, ואופן שימוש שונה בנתונים הגולמיים. להערכתנו, מימד הזמן הוא חלק מרכזי בבעיית סיווג זו, ולכן נמליץ על בחינה של מומנטים נוספים המבטאים מאפיין זה. בנוסף, יש לשקול בחינה של דרכים נוספות להשלמת הערכים החסרים (imputation), שכן מבדיקה ברשת, איכות השיטה בה בחרנו שנויה במחלוקת לצרכים של סיווג. שיפור אפשרי בהכנת הנתונים הוא הוצאת מומנטים בנפרד לכל מקטע זמן. שיטה זו תייתר את חיפוש המקטע הטוב ביותר, ותאפשר למודל לבחור את המומנטים החשובים עבור כל תקופת זמן (לדוגמה, השיפוע של טסיות הדם בשנה החמישית לפני האבחון, וה-BMI הממוצע שנה לפניו). שיטה זו תביא לריבוי מאפיינים, כך שלכל בדיקה יהיו מספר מומנטים ומספר תקופות. עם זאת, פתרון אפשרי הוא שימוש במודל שבוחר מאפיינים בעצמו (לדוגמה אלגוריתם מבוסס עצים כמו יער או XGBoost). שיפור נוסף שאפשר לבחון הוא שימוש במודל טמפורלי (temporal model) בכדי להביא לידי ביטוי את מימד הזמן בדרך מובנית. המגבלה העיקרית של מודל זה ביחס לנתונים אלו הוא שמרווחי הזמן בין הבדיקות המעבדה לא קבועים עבור המטופלים השונים. פתרון אפשרי הוא הוספת משתנים נוספים למודל שיציינו את המרווח בין בדיקות אלו.

לאור התוצאות, והאופי המעט שרירותי (ככל הנראה לעין) של איכויות החיתוכים בגרפים לעיל, אנו מעלים ספק בנוגע לתוקף החיצוני של הבדיקה שביצענו ב-trimming test. ייתכן שהתוצאות נכונות רק עבור הנתונים המסוימים האלה, ועבור שיטת עיבוד הנתונים בה בחרנו. לכן, מחקר המשך אפשרי הוא לבחון האם יתקבלו תוצאות דומות מנתוני קופות חולים אחרות ולאחר עיבוד שונה.

לסיכום, פתרון בעיית סיווג זו תוך בניית מסווג מדויק יכול לשמש כמערכת התראה והמלצה לרופאים, ולתרום רבות לחולים הפוטנציאליים במחלת ה-ALS. יש להמשיך ולשפר את המודל בכדי לשאוף לביצועים מקסימליים, בתקווה שהתרומה למטופלים תהיה המשמעותית ביותר.