The Iby and Aladar Fleischman
Faculty of Engineering
Tel Aviv University

מבוא למערכות
חיפוש למידה והמלצה

פרויקט מסכם

מרצה: ד״ר לירון יצחקי אלרהנד

עוזר הוראה: מר משה לאופר

מוגש על-ידי:

תמר שנירר 206699639

לין בורג 203839527

תקציר מנהלים

- מטרת הדו"ח הינה לתעד ולתאר את תהליך העבודה בעיבוד סט נתונים אודות מטופלי לימפומה מסוג B, על-מנת לייצר מודל לניבוי של תאים חוץ בלוטיים (גרורות).
- המודל שנבנה הצליח להגיע לתוצאה של 82.4 במדד AUC, באמצעות מודל Random Forest, ולאחר הורדת תצפיות חריגות באמצעות שיטת MCD.
- שלבים עיקריים שבוצעו במהלך העבודה:
  - **Data Exploration & Pre-Processing**
    - בחינת סט הנתונים משלב האימון העלתה כי תצפיות המתויגות כ- 0 (היעדר גרורות) מהווה 34% מהנתונים, לעומתן תצפיות שתויגו כ- 1 (לפחות גרורה אחת) מהוות 41%, והיתר- תצפיות ללא סיווג מהוות 25% מכלל הדאטה סט.
    - נבנו ארבעה סוגי עיבוד עבור הפיצ'רים (אשר יפורטו בהמשך הדו"ח – ראה איור 2):
      1. עיבוד ראשוני – הורדת עמודות שאינן רלוונטיות לניבוי, ויצירת לייבל בינארי.
      2. טרנספורמר נומרי – השלמת ערכים חסרים עבור משתנים נומריים, באמצעות חישוב הערך החציוני.
      3. טרנספורמר קטגוריאלי – השלמת ערכים חסרים עבור משתנים קטגוריאליים, באמצעות הערך "Unknown", וביצוע OHE.
      4. טרנספורמר IPI – תיקון ערכי IPI שהוזנו באופן שגוי, וחישוב סטטיסטים (ממוצע, מינימום ומקסימום).
    - כמו כן, בוצעו השלמת נתונים חסרים ונרמול הנתונים לטווח [0,1], במטרה לתת חשיבות זהה עבור כל פיצ'ר.
  - **Feature Selection** – נבחנו שתי שיטות להורדת ממד: Permutation Importance ו- Random Forest Feature Importance. לבסוף נבחרו הפיצ'רים שדורגו גבוה בשיטת ה- Permutation Importance, נוכח מהימנות השיטה על-פני השנייה, שמוטה לטובת משתנים רציפים.
  - **Modeling** – נבנו ארבעה מודלים לחיזוי הסיווג:
    - KNN
    - Random Forest
    - Logistic Regression
    - Neural Network - MLP

    המודלים נבחנו בהשוואה למודל בייסליין בסיסי, אשר בנוי מעץ החלטה הכולל חמישה פיצ'רים בלבד.

    כל אחד מהמודלים שנבדקו נבחנו על שני סטים של קבוצות משתנים:
    1. פיצ'רים משיטת ה - Permutation Importance (סה"כ 15 פיצ'רים)
    2. סט הפיצ'רים המלא (סה"כ 23 פיצ'רים)

    לבסוף, נבחרה השיטה שהחזירה לנו בשלב המידול את התוצאות האופטימליות על סט הווליד‌ציה באמצעות מדד ה- AUC.
  - **Model Evaluation** –
    - היות והתהליך היה איטרטיבי מבחינת בחירת הפיצ'רים, בדו"ח זה הוצגו התוצאות לכל מודל על סט הפיצ'רים שהחזיר את התוצאות הגבוהות ביותר.

    להלן תיאור תוצאות המודלים במדד AUC:

$Model\ (num\ of\ features)$	$Train\ AUC\ Score$	$Test\ AUC\ Score$
Baseline: Decision Tree (5)	0.807	0.752
MD Recommendation system: Decision Tree (15)	0.774	0.770
KNN (15)	0.559	0.634
Logistic Regression (15)	0.667	0.699
Random Forest (15)	**0.776**	**0.824**
MLP (23)	0.638	0.720

- ניכר כי מודלים מבוססי-עץ מתפקדים בצורה הטובה ביותר על סט הנתונים, ולכן הוחלט להשתמש במודל ה- Random Forest לצורך ביצוע התחזית על סט הבחינה.

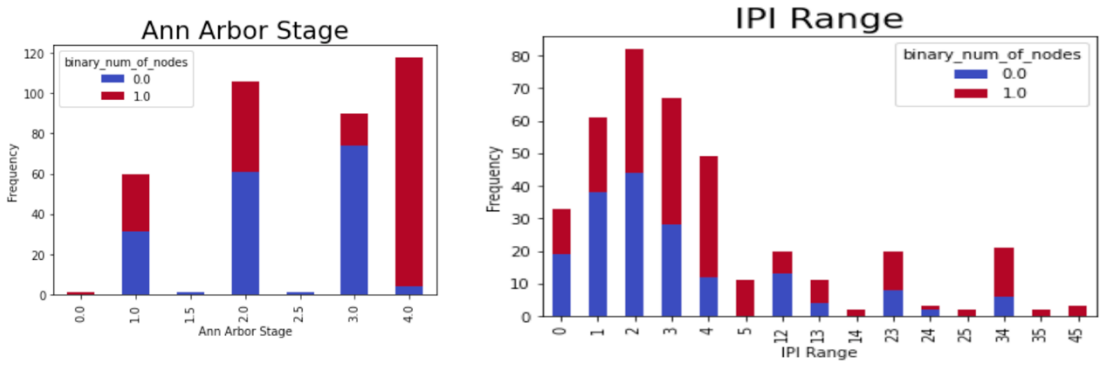
- לטובת הצוות הרפואי פותחו שני תוצרים: מערכת המלצה מבוססת חוקים לחיזוי (זהו מודל עץ בסיסי בעל ה- explainability הגבוה ביותר, בנספח ו'), וממשק נוח המאפשר להזין את נתוני הנבדק ולקבל את ההסתברות לקיומן של גרורות (מחושב על-ידי המודל המורכב) (נספח ז').

רקע – לימפומה גרורתית - חוץ בלוטית

לימפומה היא קבוצה של מספר סוגי סרטן בקשרי-לימפה (Boffetta, 2011). כחלק מפגיעתה בגוף, היא מבטאת התרבות ממאירה של לימפוציטים שהינם תאי דם לבנים המהווים חלק ממערכת החיסון של הגוף. מטרת תאים אלה היא לשמור על מאזן החלבונים והנוזלים ברקמות הגוף, שמסייעים בהגנה מפני חיידקים ווירוסים. בלימפומה גרורתית (חוץ בלוטית), תאי הלימפומה המעורבים הם תאים מסוג B, אשר אחראיים לייצור תאי הדם הלבנים (Armitage, Gascoyne, Lunning & Cavalli, 2017). לימפומה, לסוגיה, באה לידי ביטוי בסימפטומים של חום ממושך, ירידה במשקל, הזעות לילה, והחולים בה עשויים לחוש בקשרי לימפה בולטים בגופן, המאופיינים למשל בהתנפחות באזור הצוואר ובבית השחי). למול זאת, לימפומה המערכת תאי B עלולה להתפתח במקומות נוספים בגוף, שהינם מחוץ לבלוטות הלימפה – ומכאן שמה: לימפומה חוץ בלוטית.

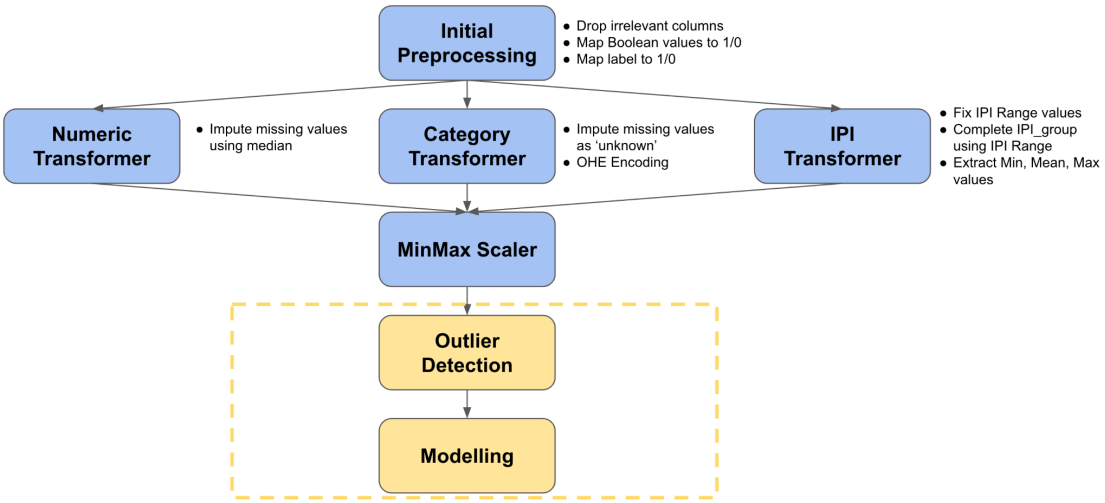
שלבים א' וב' - *Data Exploration & Pre Processing*

רשימת הפיצ'רים המקורית בדאטה סט מפורטת בנספח א' (רשימה זו כוללת עבור כל פיצ'ר הסבר אודותיו, וציון הטרנספורמר שהוחל עליו).
בבחינת הדאטה סט נמצא איזון באשר להתפלגות המשתנה התלוי (binary_num_of_nodes), כאשר 174 (34%) תצפיות קיבלו את הערך 0 (אין גרורות), 213 (41%) תצפיות קיבלו את הערך 1 (קיימת לפחות גרורה אחת) ועבור 129 (25%) תצפיות היה חסר ערך בעמודה זו.
בנוסף, נבחנו כמה פיצ'רים שמבחינה ראשונית של הדאטה נדמה שיכלו לסייע בניבוי המשתנה התלוי: תחילה, נראה כי עבור המשתנה Ann Arbor Stage, נמצא כי החל מהערך 3, עולה ההסתברות לנוכחות של גרורות בקרב הנבדק (איור 1). שנית, בחינת משתנה IPI Range העלתה כי קיימות שגיאות בדאטה. למעשה, טווח הערכים של משתנה זה נע בין 0 ל-5, ואילו בדאטה סט נמצאו ערכים כמו 14, 23, או 24. עיון בספרות הרפואית הרלוונטית והיוועצות בגורם רפואי מוסמך, העלו את ההבנה שמדובר בשגיאות הזנה, ולאחר שלב עיבוד הנתונים המקדים, ערכים אלה הוצגו כטווחים (כלומר הערך 14 מבטא את הטווח 1-4, והערך 23 מבטא את הטווח 2-3). איור 1 מתאר את התפלגות הערכים של משתנה זה טרם תיקונם, וניתן להיווכח כי ככל שהטווח גבוה יותר, כך גדלה ההסתברות לנוכחות של גרורות בקרב הנבדק.



איור 1:
מימין: היסטוגרמת הלייבל (0 בכחול, 1 באדום) עבור עמודת Ann Arbor Stage. ניתן לראות כי החל מהערך 3, עולה ההסתברות לנוכחות של גרורות בקרב הנבדק.
משמאל: היסטוגרמת הלייבל (0 בכחול, 1 באדום) עבור עמודת IPI Range טרם תיקון הנתונים. ניתן לראות כי ככל שהטווח גבוה יותר, כך גדלה ההסתברות לנוכחות של גרורות בקרב הנבדק.

שלב ה – Feature Engineering בוצע באמצעות שני טרנספורמרים מותאמים אישית: 1. Initial_pre_processing_transformer, ו-2. FeatureTransformer, שכלל שלושה תת-טרנספורמרים נוספים: numeric_transformer, cat_transformer & ipi_values_Transformer. זאת תוך שימוש ב– TransformerMixin class של ספריית sklearn.



איור 2: הפייפליין של המודל. הדאטה עובר דרך הטרנספורמרים (בכחול), לאחר מכן עובר נורמליזציה ולבסוף עובר לשלב הוצאת החריגים לקראת המידול.

1. Initial_pre_processing_Transformer:

טרנספורמר זה בוצע על כל הדאטה סט, והורדו בו עמודות שהיו בלתי-רלוונטיות לחיזוי המשתנה התלוי (מפורט בנספח א'). לאחר מכן יצרנו לייבל בינארי תוך מיפוי עמודת ה-Number of Extranodal Sites, כאשר הלייבל 0 מתקבל אם לנבדק אין כל גרורה, או 1 אם קיימת לנבדק גרורה אחת לפחות. לבסוף, חולק הדאטה לסט אימון (train) ולסט מבחן (test), והופעלו עליו שלושה תתי טרנספורמרים נוספים שנכללו תחת הטרנספורמר FeatureTransformer.

2. FeatureTransformer:

כאמור, טרנספורמר זה הורכב משלושה תתי-טרנספורמרים:

1. numeric_transformer – רלוונטי עבור פיצ'רים בעלי ערכים נומריים, והשלים עבורם חוסרים לפי הערך החציוני שלהם בסט האימון.
2. cat_transformer – רלוונטי עבור פיצ'רים בעלי ערכים קטגוריאליים, והשלים עבורם ערכים חסרים באמצעות הערך "Unknown". בנוסף, הטרנספורמר פיצל את הפיצ'ר למספר dummy features, כמספר הערכים בכל פיצ'ר, כפי שהופיעו בסט האימון.
3. ipi_values_Transformer – רלוונטי לפיצ'רים IPI Group ו- IPI Range. בוצע תיקון בהתאם למתואר מעלה, ובנוסף הושלמו ערכים חסרים ב- IPI Group בהתאם לערך ה- IPI Range (ערכים חסרים שתאמו את הטווח 0-1 סווגו כ- "low", בהתאם הטווח 2-3 סווג כ- "intermediate" והטווח 4-5 כ- "high").

לסיכום, יישום הטרנספורמרים על סט האימון הוביל ל – 23 פיצ'רים בסך הכל. לאחר מכן כל פיצ'ר נורמל לסקאלה שבין 0-1. חשיבות הנרמול תתבטא בעיקר בהמשך במודל ה-KNN, בו המרחק מחושב באמצעות מרחק אוקלידי, במטרה לתת חשיבות זהה לכל פיצ'ר. נציין כי אין בהכרח צורך לנרמל עבור מודל ה– Random Forest, אך גם במודל זה נעשה שימוש בנתונים המנורמלים מטעמי נוחות. הדבר חל גם על יתר המודלים, אך הנרמול מסייע בהתכנסות מהירה יותר, ולכן הוחלט להשתמש בסט נתונים מנורמל גם עבורם.

שלב ג' - *Outlier Removal*

נעשה שימוש בשלוש שיטות להוצאת חריגים, כאשר מכל אחת נבנתה "מסיכה", כלומר וקטור בינארי שקבע איזה מהתצפיות בסט האימון הן חריגות ואיזה לא. לאחר סינון הדאטה ע"י המסיכה אומנו כל אחד מהמודלים שנבחנו (לרבות מסווג ה- Random Forest שנבחר) ואז נעשה חיזוי על ה- Test (כמובן ללא הוצאת החריגים מה- Test). המסיכה שימשה אותנו גם בשלב מציאת המודל האופטימלי.

השיטה הראשונה נשענה על ההנחה שתצפיות לא חריגות נוצרות על-ידי התפלגות של הסתברות מסוימת, ולפיה נקודות עם צפיפות הסתברות נמוכה סווגו כחריגות. עבור נתונים המגיעים מהתפלגות נורמלית, ניתן לעשות זאת על ידי חישוב מרחק מהלנוביס מכל נקודה לממוצע, והגדרת חריגות כנקודות עם מרחק העובר מרחק סף מסוים. מרחק מהלנוביס דורש את הפרמטרים של ההתפלגות (ממוצע ושונות משותפת). מכיוון שאלה אינם ידועים, יש להעריך אותם מהנתונים.

הערכת הפרמטרים וחישוב מרחק מהלנוביס בוצעו על-ידיי תת-ספרייה של sklearn הנקראת MinCovDet (או בקיצור MCD) תוך שימוש במתודה fit ו- mahalanobis על ה- Train, ואז על-ידי לולאת for בדקנו החל מאיזה אחוזון מרחק (מ-50 עד 100) ניתן להגדיר את הנתונים כ-"לא חריגים" על מנת לקבל את המסווג האופטימלי שיעלה ככל הניתן את ניקוד ה-AUC על סט הבחינה (התוצאות מוצגות בנספח ג). כך לדוגמא, עבור סף של 90, נקבע כי 90% מהדאטה ב-Train אינו חריג - כלומר יש לו מרחק מהלנוביס שנמוך מהערך באחוזון ה-90. אימנו את המודל על הדאטה המנורמל בלבד וחזינו על ה- Test.

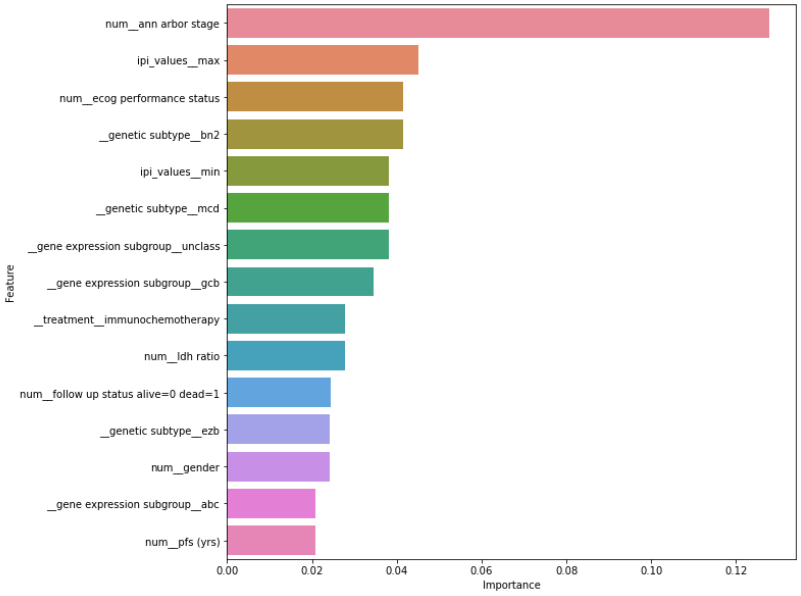
התוצאה שהביאה את ניקוד ה-AUC הגבוה ביותר (0.793) על ה- Test קבעה כי עלינו לסווג 6% מהתצפיות בסט האימון כחריגות.

שיטה שנייה שבה נעשה שימוש היא Isolation Forest לזיהוי חריגים (Liu, Ting & Zhou, 2008). שיטה זו משתמשת בלמידה בלתי מונחית (unsupervised) ודומה ל – Random Forest במנגנון בניית העצים שלה, ופיצול לפי פיצ'רים. תצפיות חריגות הן כאלה שעבורן נדרשת בממוצע כמות פיצולים נמוכה באופן יחסי (אילוסטרציה מצורפת בנספח ב').
במודל הנוכחי נבדק בכל פעם איזה אחוז contamination (כלומר, איזה אחוז מהדאטה צריך לסווג כתצפיות חריגות) על-מנת שיינתן ניקוד ה-AUC הגבוה ביותר על ה- Test. לבסוף, בהתאם לתוצאה שהתקבלה נמצא כי יש לסווג 11% מהדאטה כחריג על-מנת להגיע לערך AUC של 0.808. (בנספח ד' ניתן לראות כיצד הניקוד משתנה כתלות באחוז ה- contamination בטווח בין 0% עד 20%).

שיטה שלישית לזיהוי חריגים היא ע"י אלגוריתם DBSCAN, שהינו אלגוריתם clustering לא-פרמטרי מבוסס צפיפות (Ester, Kriegel, Sander & Xu, 1996). בהינתן קבוצת נקודות במרחב מסוים, יתקבל cluster של נקודות בהתאם להיפר-פרמטרים הבאים: אפסילון, שקובע עד כמה שתי תצפיות צריכות להיות קרובות כדי להשתייך לאותו cluster; ומספר נקודות מינימליות, שקובעות את המספר המינימלי של תצפיות קרובות שמגדירות cluster. הנקודות שאינן משתייכות ל- cluster נחשבות כ"רעש", והוגדרו במודל כחריגות. האלגוריתם נבדק עבור ערכי אפסילון של 1 עד 14 וערכי מספר נקודות מינימלי של 2 עד 4. התקבל שניקוד ה-AUC האופטימלי (0.802) מושג כאשר המרחק (האוקלידי) המקסימלי בין שתי נקודות הינו 3, ומספר הנקודות המינימלי להגדרת cluster הינו 2. כך, 19.65% מהנקודות בדאטה הוגדרו כחריגות.

שלב ד' - *Feature Selection*

נבדקו שני סטים שונים של פיצ'רים: האחד כלל את 23 הפיצ'רים שנוצרו לאחר החלת הטרנספורמרים, והשני כלל סט מצומצם של פיצ'רים שנבחרו באמצעות *Permutation Importance* . על-מנת לצמצם את סט הפיצ'רים, נבחנו שתי שיטות שונות להורדת ממד: 1) *Random Forest Importance*, שבה הורץ אלגוריתם *Random Forest*; האלגוריתם החזיר את החשיבות של כל פיצ'ר באמצעות מדד *Gini* לחישוב טיב הפיצולים. מכיוון ששיטה זו מוטה לטובת ערכים רציפים/בעלי קרדינאליות גבוהה (עם סוגי ערכים שונים רבים), בחרנו בסופו של דבר להשתמש בשיטה נוספת - *Permutation Importance*, שבה מערבלים בכל פעם פיצ'ר אחד, ובודקים בכמה השתנה הציון של המודל. אם הציון ירד באופן משמעותי (באופן יחסי ליתר הפיצ'רים), זו אינדיקציה לכך כי הפיצ'ר המעורבל חשוב עבור המודל. באמצעות שיטה זו הגענו לסט המונה 15 פיצ'רים (רשימת הפיצ'רים שנתקבלו בשיטה זו נמצאים באיור 3).



איור 3: 15 רשימת הפיצ'רים שהתקבלו לאחר שימוש בשיטת Permutation Importance.

בהמשך בחנו כל אחד מסט הפיצ'רים לעיל, בקומבינציה של כל אחת משיטות הוצאת החריגים, על כל אחד מארבעת המודלים שנבחרו, ולבסוף השתמשנו בשיטה שהחזירה לנו את התוצאות הטובות ביותר על סט הווליד‌ציה (לאחר ביצוע אופטימיזציה לפרמטרים בשיטת K-Fold Cross Validation).

בניית מודלים ראשוניים והערכתם – *KNN & Logistic Regression*

- מודל בייסליין – **עץ החלטה** – על-מנת שנוכל להשוות את טיב המודלים שיפורטו מטה, בנינו תחילה מודל בסיסי המורכב מעץ החלטה המכיל את חמשת הפיצ'רים שנמצאו כחשובים בשיטת ה- Feature Selection. מודל "צעצוע" זה הוא בעל explainability גבוה, וניתן לראות כיצד פועל בנספח ה.
- עבור סט האימון – טיב מודל זה באמצעות מדד *AUC* הינו: 0.807
- עבור סט הווליד‌ציה – טיב מודל זה באמצעות מדד *AUC* הינו: 0.752

- **KNN** – ראשית נבחנו 4 אפשרויות שונות ל- k, כולן מספרים אי-זוגיים (3, 5, 7, 9), זאת על-מנת למנוע מצבים של שוויון מבחינת הסיווגים. לאחר מכן ניסינו למצוא התאמה מיטבית ל- k, עבור כל אחת משני סטי הפיצ'רים שתוארו לעיל. פרמטר נוסף שנבדק הוא weights, והוא מהווה את המשקול שניתן לכל תצפית מתוך שתי אפשרויות קיימות: משקול אחיד או מתן משקל גבוה יותר לתצפיות שקרובות זו לזו.

לבסוף, שילוב הפרמטרים הטובים ביותר על כל אחד מסטי הפיצ'רים, נבחר באמצעות חישוב מדד AUC.

- סט הפיצ'רים שנתן את התוצאה הטובה ביותר עבור המודל הינו זה המצומצם (15), לפי שיטת הוצאת החריגים של DBSCAN, כאשר $k = 9$ ומשקול התצפיות הינו יחסי למרחק.
  - עבור סט האימון – טיב מודל זה באמצעות מדד AUC הינו: 0.559
  - עבור סט הבחינה – טיב מודל זה באמצעות מדד AUC הינו: 0.634

- **Logistic Regression** – ראשית נבחנו 5 אפשרויות שונות לפרמטר הרגולריזציה, C: 0.01, 0.1, 1, 100, 1000, כאשר C בעל ערך גבוה יותר, משמעו פחות רגולריזציה (קנס). הפרמטר הבא הינו פונקציית הקנס: l1 או l2. ההבדל בין שתי פונקציות קנס אלה הוא ש- l1 מנסה למזער ככל הניתן פיצ'רים בעלי חשיבות נמוכה, על-ידי איפוסם. פיצ'רים נוספים שנבדקו, עבורם הגדרנו פרמטרים, הם מספר האיטרציות המקסימאלי (max_iter), והאלגוריתם לפתרון (solver), ששווה ל- liblinear, אשר נחשב אופטימאלי עבור דאטה סטים קטנים יחסית.
- סט הפיצ'רים שנתן את התוצאה הטובה ביותר עבור המודל הינו זה המצומצם (15), לפי שיטת הוצאת חריגים של DBSCAN, כאשר C=1, max_linear=100, ופונקציית הקנס שווה ל- l1.
  - עבור סט האימון – טיב מודל זה באמצעות מדד AUC הינו: 0.667
  - עבור סט הבחינה – טיב מודל זה באמצעות מדד AUC הינו: 0.699

בניית מודלים מתקדמים - $Random\ Forest\ \&\ MLPClassifier$

- **Random Forest** – הגדרנו מסווג RF (Random Forest), שקיבל את סט האימון על שתי תצורותיו (כמפורט לעיל ב- Feature selection), ואת וקטור הלייבלים. נבחנו שלוש אפשרויות למספר העצים ביער (51, 101, 151). בנוסף הגדרנו את כמות הדגימות המינימאלית (2, 3, 4) במטרה לייצר פיצול עבור כל צומת. הגדרנו גם את עומק העץ באמצעות שלוש אפשרויות (12, 25, או ללא הגבלה) ואת כמות הפיצ'רים בכל עץ (שורש מספר הפיצ'רים בקלט, או ללא הגבלה) ולבסוף בחרנו במיטבית.
- סט הפיצ'רים שנתן את התוצאה הטובה ביותר עבור המודל הינו זה המצומצם (15), לפי שיטת הוצאת חריגים של MCD, כאשר bootstrap = True, העומק המקסימלי של העץ הוא בלתי מוגבל, כמות הדגימות המינימאלית לפיצול היא 4, כמות הפיצ'רים המקסימלית בעץ היא בלתי מוגבלת ומספר העצים הוא 51.
  - עבור סט האימון – טיב מודל זה באמצעות מדד AUC הינו: 0.776
  - עבור סט הבחינה – טיב מודל זה באמצעות מדד AUC הינו: 0.824

- **MLPClassifier** – פרמטרים להם ביצענו אופטימיזציה הם ארכיטקטורת הרשת, פונקציית האקטיבציה (logistic/relu) וכמות האיטרציות המקסימלית (100, 120, 150, 200). באשר לארכיטקטורת הרשת, התאמנו לכל סט פיצ'רים ארכיטקטורה אחת לפחות, שגודל השכבה הראשונה שלה, היא ככמות הפיצ'רים.
- סט הפיצ'רים שנתן את התוצאה הטובה ביותר עבור המודל הינו זה המורחב (23), לפי שיטת הוצאת חריגים של DBSCAN, כאשר פונקציית האקטיבציה שנבחרה היא relu, הארכיטקטורה של הרשת (כמות הנוירונים בכל שכבה) היא 23 (כמספר הפיצ'רים), ומספר האיטרציות המקסימאלי הוא 120.
  - עבור סט האימון – טיב מודל זה באמצעות מדד AUC הינו: 0.638
  - עבור סט הבחינה – טיב מודל זה באמצעות מדד AUC הינו: 0.720

לסיכום, נמצא כי הקומבינציה הכי מוצלחת של אלגוריתם מסווג, אלגוריתם לזיהוי חריגים ומספר הפיצ'רים במודל היא: Random Forest, MCD ו-15 פיצ'רים שנתנו לנו ניקוד AUC של 0.824.

תיוג הלא מתוייגים - Semi Supervision

כמתואר 25% מהנתונים בעמודת המשתנה התלוי (binary_num_of_nodes) היו חסרים. מכיוון שאחוז זה מתוך הדאטה עשוי להיות רלוונטי לשיפור יכולת החיזוי של הלייבל, הוחלט לבחון את האפשרות לייצר בצורה "חכמה" תיוג עבור תצפיות אלה. תיוג התצפיות הלא מתויגות בוצע בשיטת Semi Supervision, באופן האיטרטיבי הבא: תחילה אומן המסווג Random Forest על נתוני ה- Train הנבחרים. לאחר מכן, נעשה שימוש במסווג זה על על התצפיות הלא-מתויגות במטרה לחזות אותן. בשלב הבא חוברו התצפיות "החדשות" עליהן התקבל סיווג ברמת אמינות גבוהה יחסית לסט ה- Train של המודל, ולבסוף נעשה שימוש בתצפיות המאומנות ליצירת חיזוי עבור סט ה- Test של הדאטה - וכך שוב עד אשר לא קיימים סיווגים ברמת אמינות גבוהה.
אולם, לאור הממצאים, הוספת החיזוי עבור התצפיות הלא מתויגות לסט האימון, לא שיפרה באופן מספק את יכולת המודל לחזות את המשתנה התלוי.

תוצר סופי

התוצר הסופי הוא תוכנה שנכתבה בפייתון בספריית tkinter ובשאיפה תוכל לסייע לרופאים ולאחיות בקביעה האם לחולה יש גרורות. מתוך הבנה שלחולי לימפומה המגיעים לבית החולים אין בהכרח את המשתנים שקשורים ל-follow-up, ל-progression ול-survival analysis (שכן לא כל החולים בלימפומה כלולים במחקר), בחרנו שלא להציג למשתמש בקשה למלאם. את המשתנים המספריים ניתן להקליד כטקסט חופשי ואת המשתנים קטגוריאליים יש לבחור מתוך תפריט. לאחר לחיצה על submit מתקבלת ההסתברות שלחולה לימפומה יש גרורות. בנספח ז' ניתן לראות תוצאה של חולה המתקבלת לאחר מילוי המשתנים.

ביבליוגרפיה

Armitage, J. O., Gascoyne, R. D., Lunning, M. A., & Cavalli, F. (2017). Non-hodgkin lymphoma. *The Lancet*, 390(10091), 298-310.

Boffetta, P. (2011). I. Epidemiology of adult non-Hodgkin lymphoma. *Annals of oncology*, 22, iv27-iv31.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*. Vol. 96, No. 34, 226-231.

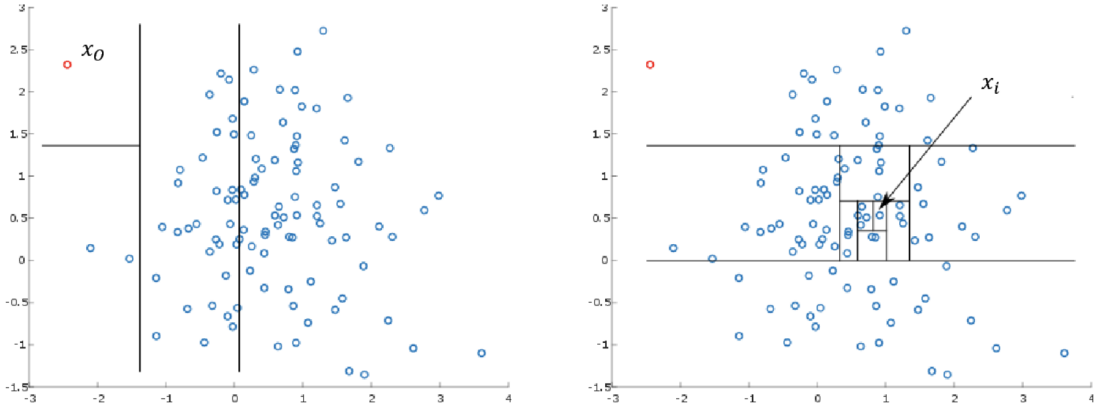
Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining,* 413-422.

נספח א' – רשימת הפיצ'רים המקוריים בדאטה סט:

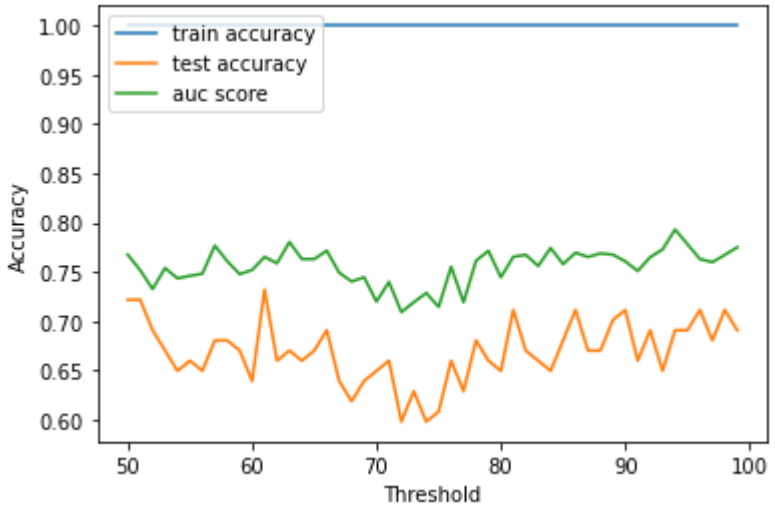
FieldName	*Type*	*Description*	*Transformer Handler*
dbGaP submitted subject ID	object	ערך מזהה ייחודי עבור כל נבדק	Not used for prediction
dbGaP accession	object	מזהה ייחודי שניתן לרשומת DNA או רצף חלבונים, ע"מ לאפשר מעקב אחר גרסאות שונות של רשומת הרצף הגנטי	Not used for prediction
Diagnosis	object	סוג הסרטן הספציפי לנבדק	Not used for prediction
Gene Expression Subgroup	object	תת סוג הלימפומה – כל תת סוג מושפע באופן שונה מטיפול כימותרפיה	cat_Transformer
Genetic Subtype	object	סוג הגן שגרם ללימפומה	cat_Transformer
Biopsy Type	object	סוג הביופסיה: הערך "Pre-treatment" מעיד שדגימת הביופסיה נלקחה טרם טיפול בנבדק; הערך "Relapse" פירושו שהדגימה נלקחה לאחר אבחון הנדבק עם לימפומה	cat_Transformer
Treatment	object	סוג הטיפול שניתן לנבדק	cat_Transformer
Gender	object	מגדר הנדבק	numeric_Transformer
Age	float64	גיל הנבדק (בשנים)	numeric_Transformer
Ann Arbor Stage	float64	מערכת לסיווג חומרת הלימפומה מ: 1 (הקלה ביותר) עד 4 (הקשה ביותר)	numeric_Transformer
LDH Ratio	float64	רמת הלקטט דהידרוגנאז (LDH) בדם	numeric_Transformer
ECOG Performance Status	float64	מתאר את רמת התפקוד של הנבדק במונחים של יכולתו לטפל בעצמו מ: 0 (תפקוד טוב) 5 (תפקוד לקוי)	numeric_Transformer
Number of Extranodal Sites	float64	מספר גרורות של סרטן בלוטות מעבר לגבולות הקפסולה של בלוטת לימפה לרקמות סמוכות	Mapped to binary label: $0 \rightarrow 0,\ 1 + \rightarrow 1$
IPI Group	object	הצגה קטגוריות של כלי קליני שפותח על ידי אונקולוגים, למען חיזוי הפרוגנוזה של חולים עם לימפומה אגרסיבית שאינה הודג'קין: 0-1 (נמוך); 2-3 (בינוני); 4-5 (גבוה)	ipi_values_Transformer
IPI Range	int64	הצגת הטווח בין 0-5 של כלי קליני שפותח על ידי אונקולוגים, למען חיזוי הפרוגנוזה של חולים עם לימפומה אגרסיבית שאינה הודג'קין	ipi_values_Transformer
Follow up Status Alive=0 Dead=1	int64	מציין האם האדם חי או מת במהלך המחקר	numeric_Transformer
Follow up Time (yrs)	float64	נמדד מזמן אפס (מתחילת המחקר או מהנקודה בה המשתתף נחשב בסיכון), ועד למות המשתתף או לסיום המחקר	numeric_Transformer

PFS Status No Progress=0 Progress=1	int64	האם האדם הראה התקדמות מאז מתן הטיפול	numeric_Transformer
PFS (yrs)	float64	משך הזמן במהלך ואחרי הטיפול במחלה שחולה חי עם המחלה אך היא אינה מחמירה	numeric_Transformer
Included in Survival Analysis	object	האם האדם נכלל בניתוח ההישרדות	Not used for prediction

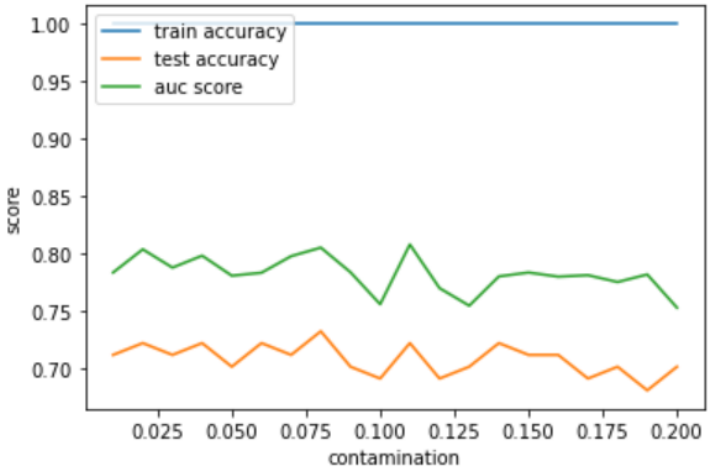
נספח ב' – Isolation Forest - אילוסטרציה:



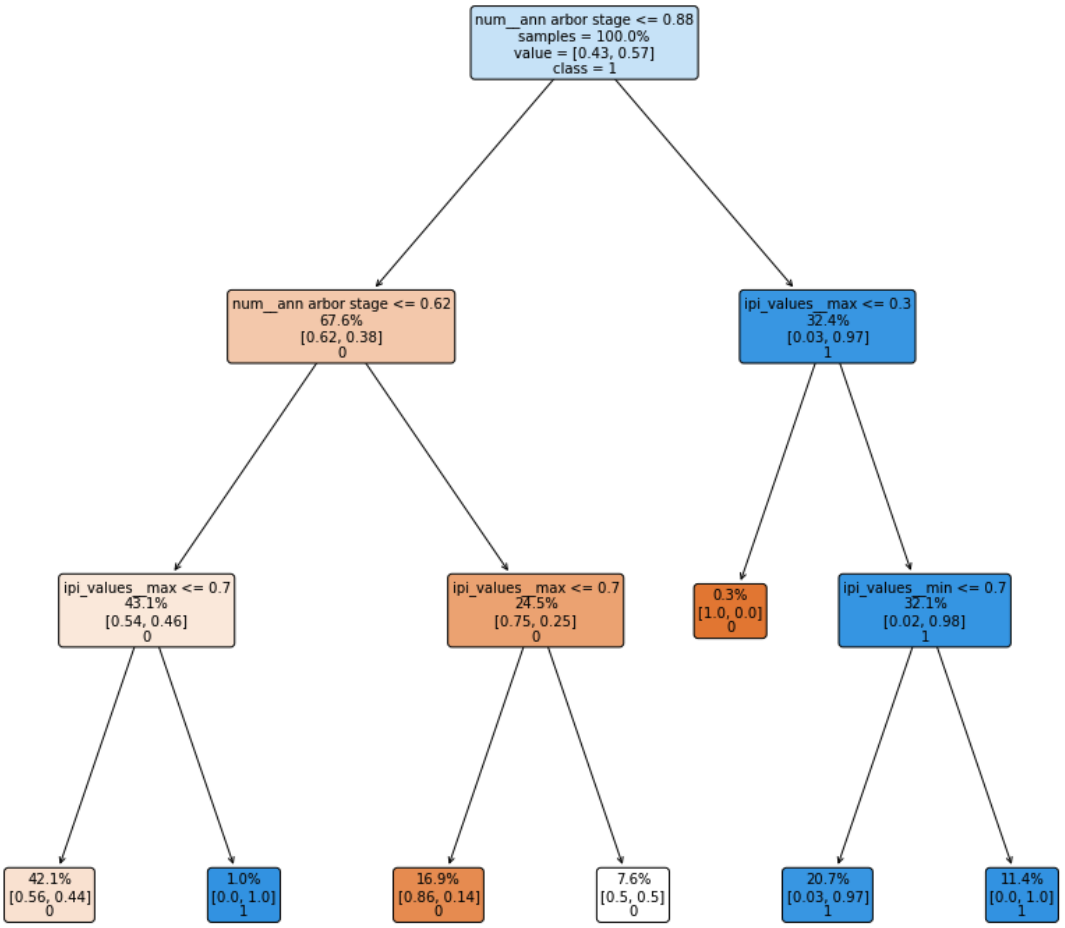
נספח ג' - גרף המתאר את ביצועי המודל על כל אחד מספי המרחק הקובעים את האנומליה



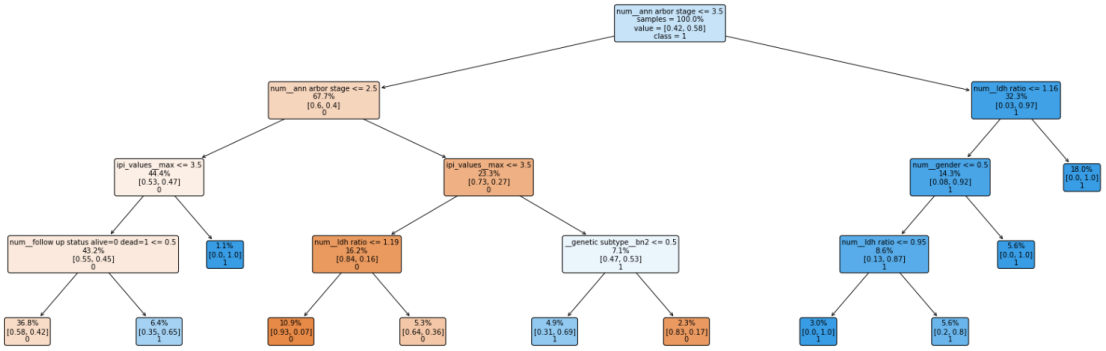
נספח ד' - גרף המתאר את ביצועי המודל על כל אחד מספי המרחק הקובעים את האנומליה.



נספח ה – מודל הבייסליין – עץ החלטה כאשר הדאטה מנורמל:



נספח ו' – מודל עץ בסיסי לרופאים - בעלת explainability גבוה:



נספח ז' - חלון תוכנת מערכת ההמלצה לצוות הרפואי, המאפשרת הזנת נתונים עבור חולה וקבלת ההסתברות לקבלת גרורות עבור החולה