# Introduction To Machine Learning

# Hackathon 2022

Michael Hashkes, Tomer Rozenstine, Elinoy Gidon, Eden Ofek

## שלבי העבודה

### 1 Preprocessing

ראשית, הפרדנו את סט האימון לקבוצת אימון ולקבוצת טסט ולאחר מכן התחלנו את תהליך עיבוד המידע. תחילה, סיננו רשומות שתיארו להבנתנו ביקורים זהים של אותה המטופלת. הליך זה צמצם את כמות הדגימות ל-8,749 מדידות וסייע לנו להימנע מזליגת מידע בין ה-test ל-train. בהמשך, כתבנו קוד שמטרתו לסדר את הפיצ׳רים כך שיצליחו לתאר את סט האימון בצורה המיטבית. במסגרת כך, הפכנו פיצ׳רים קטגוריאליים לרציפים; חילצנו מלל חופשי מפיצ׳רים ויצרנו קטגוריות רציפות בהתאם להיגיון הרפואי שמאחורי כל פיצ׳ר (ובהתבסס על הערכים שנמצאים בtrain בלבד, גם עבור הtest); ומילאנו ערכים ריקים בהתאם להתנהגות של כל פיצ׳ר. כמו כן, פיצלנו את ווקטור הרספונס שהכיל רשימות של אזורי התפשטות ל-11 וקטורים בינאריים נפרדים.

### 2 EDA

בסיום שלב עריכת המידע התפננו לחקור את הפיצ׳רים. בדקנו את הקורלציה של הפיצ׳רים עם עצמם (גרף 5) ושל הפיצ׳רים עם כל ווקטוריי הרספונס השונים (גרף 3). כמו כן, בדקנו את הקורלציה של כל פיצ׳ר עם הרספונס של המשימה השנייה – גודל הגידול (גרף 4).

### 3 Baseline

עבור כל אחת מהשאלות יצרנו מודל Baseline שהתבסס על שלושה פיצ׳רים שעברו את שלב ה-preprocessing והציגו קורלציה יחסית גבוהה לרספונסים. עבור שאלת הקלסיפיקציה, הרצנו מודל Random-Forest. עבור שאלת הרגרסיה, הרצנו מודלי Linear-Regression ו--Polynomial Fitting (שאת דרגתו בחרנו בעזרת K-Fold CV). בשני המודלים הללו קיבלנו על הסט ה-tset ערכי Loss של 1.62.

### 4 Model Selection

<u>עבור מודל הקלסיפיקציה</u> (המשימה הראשונה): כשהרצנו את מודל ה-Random-Forest קיבלנו שהניקוד של סט האימון היה כמעט 1 ולכן הסקנו שיש לנו Over-Fitting על ה-train. על מנת לפתור את זה, הפעלנו רגולריזציה על המודל וכן ניסינו להריץ KNN עם בחינה של כמות השכנים האופטימלית. בסופו של דבר התוצאה הטובה ביותר התקבלה עבור מודל KNN עם $k = 2$ (גרף מס׳ 1).

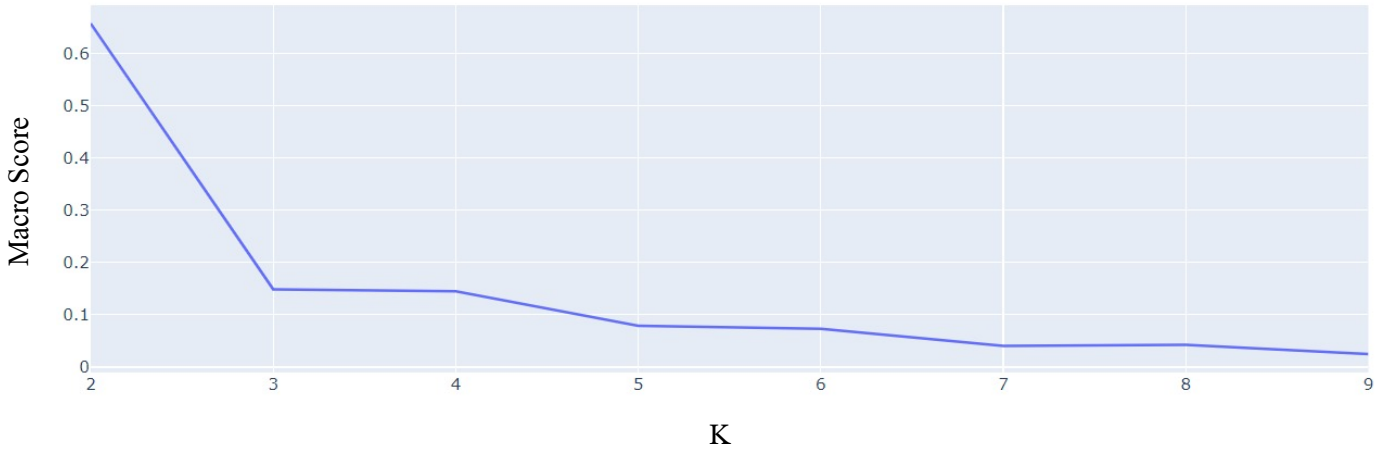
<u>עבור מודל הרגרסיה</u> (המשימה השנייה): ראשית, הבנו שקיימת בעיה במודל שאימנו שכן ה-Loss של ה-train היה פחות טוב מזה של ה-test. ניסינו לשפר את הלמידה של המודל במספר דרכים: סינון פיצ׳רים בהתאם לקורלציה שלהם עם ווקטוריי הרספונס; בחירת פיצ׳רים עם מודל ERF; הרצת רגרסיה על תוצאות ריצת PCA. כמו כן, בדקנו את המודלים עם פרמטרי רגולריזציה שונים (שבחרנו בעזרת K-Fold CV) והתוצאה הטובה ביותר שהתקבלה ניתנה על ידי רגרסיה לינארית עם רגולריזציית Ridge בעלת מקדם $\lambda = 52.2$. עם זאת, ה-Loss על ה-train נותר גבוה מעל ה-test.

בהמשך, ניסינו לשפר את הלמידה של סט האימון על ידי אימון מודלים מורכבים יותר כגון Decision-Tree-Regressor ובהמשך, Random-Forest-Regressor. צעד זה אכן תרם לשיפור

תוצאות ה-Loss על הtrain אך החמיר את ה-Loss על ה-test. על מנת לאזן זאת, הרצנו K-Fold על העץ על מנת למצוא את העומק האופטימלי. עם השיטה הזאת הגענו לתוצאת ה-Loss הטובה ביותר עד כה - 0.84, עבור מודל Random-Forest בעומק 4 (גרף מס' 2).

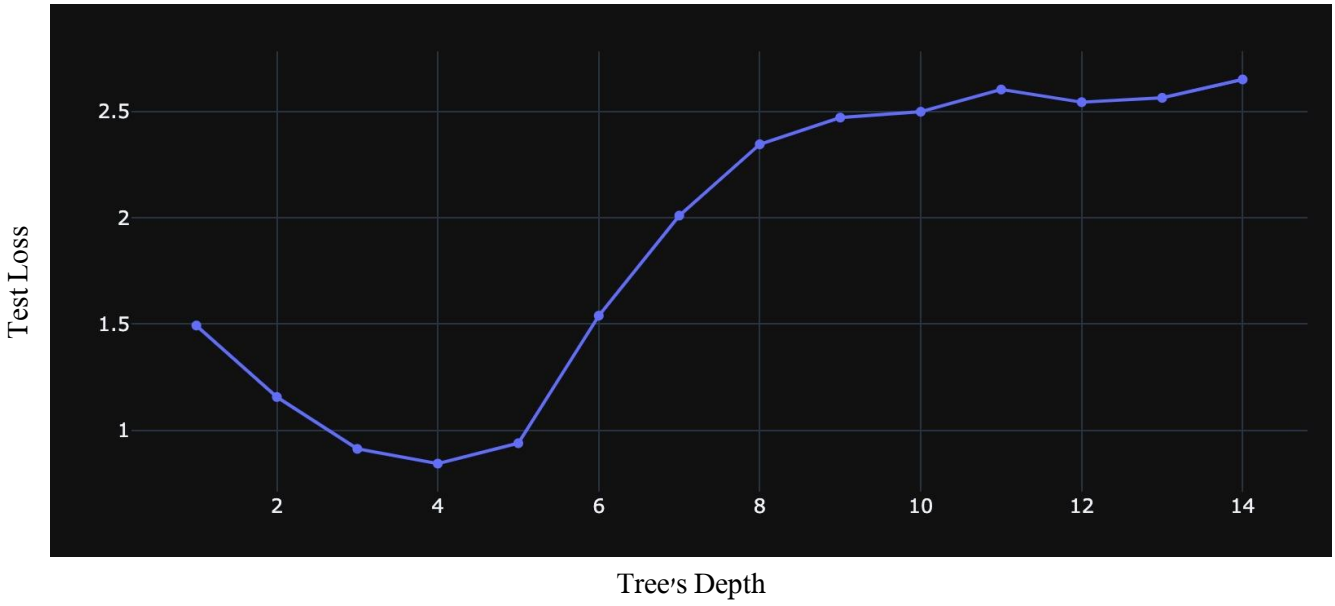
## גרפים

**גרף 1** Macro score עבור מודל KNN כפונקציה של מספר השכנים K



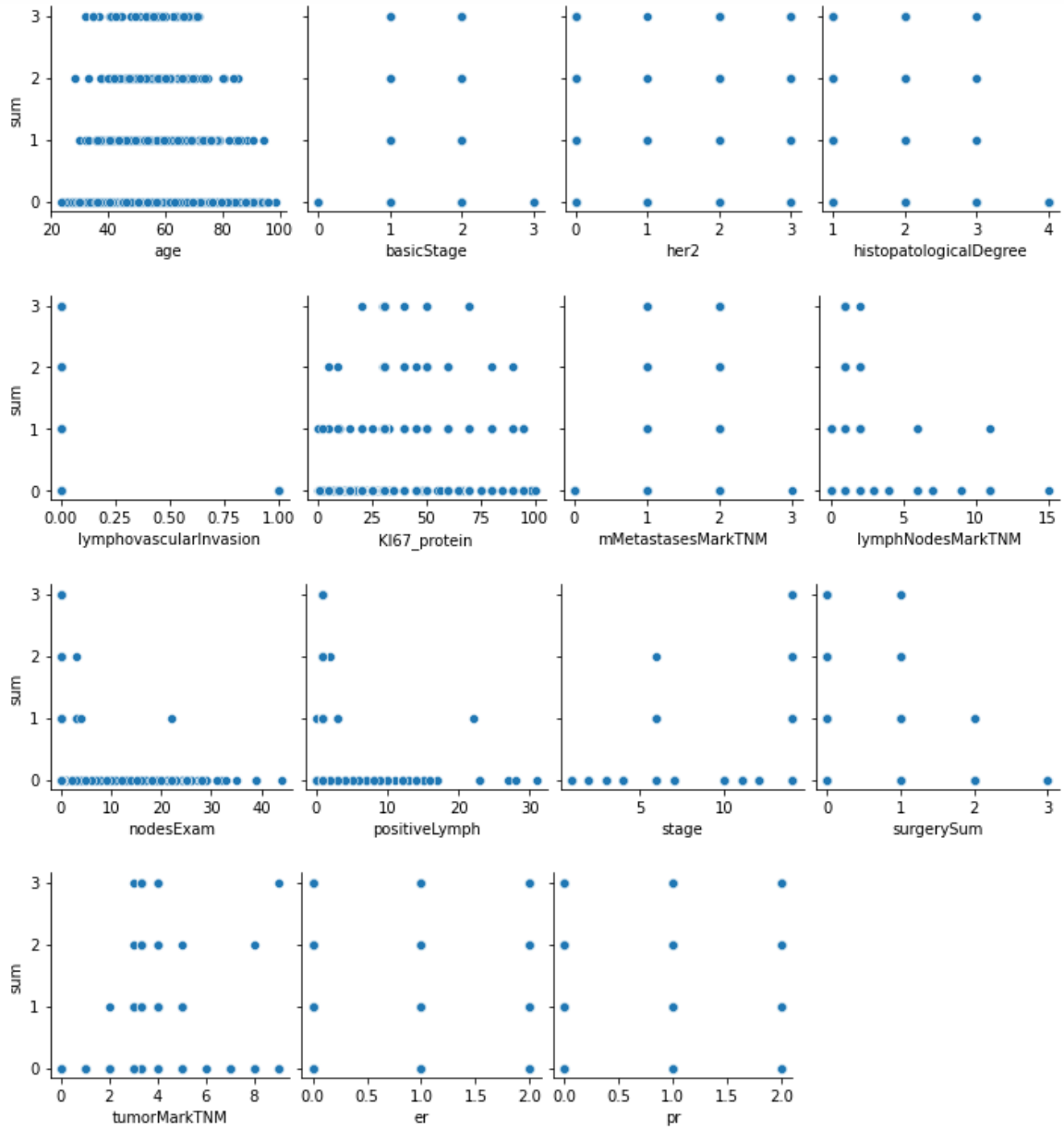

בגרף ניתן לראות את הניקוד שניתן למודל Multilabel KNN על ידי פונקציית Macro Score F1, כתלות במספר השכנים. כפי שניתן לראות, הניקוד הגבוה ביותר התקבל עבור מודל עם $k = 2$.

**גרף 2** Test Loss כפונקציה של עומק ה-Random Forest



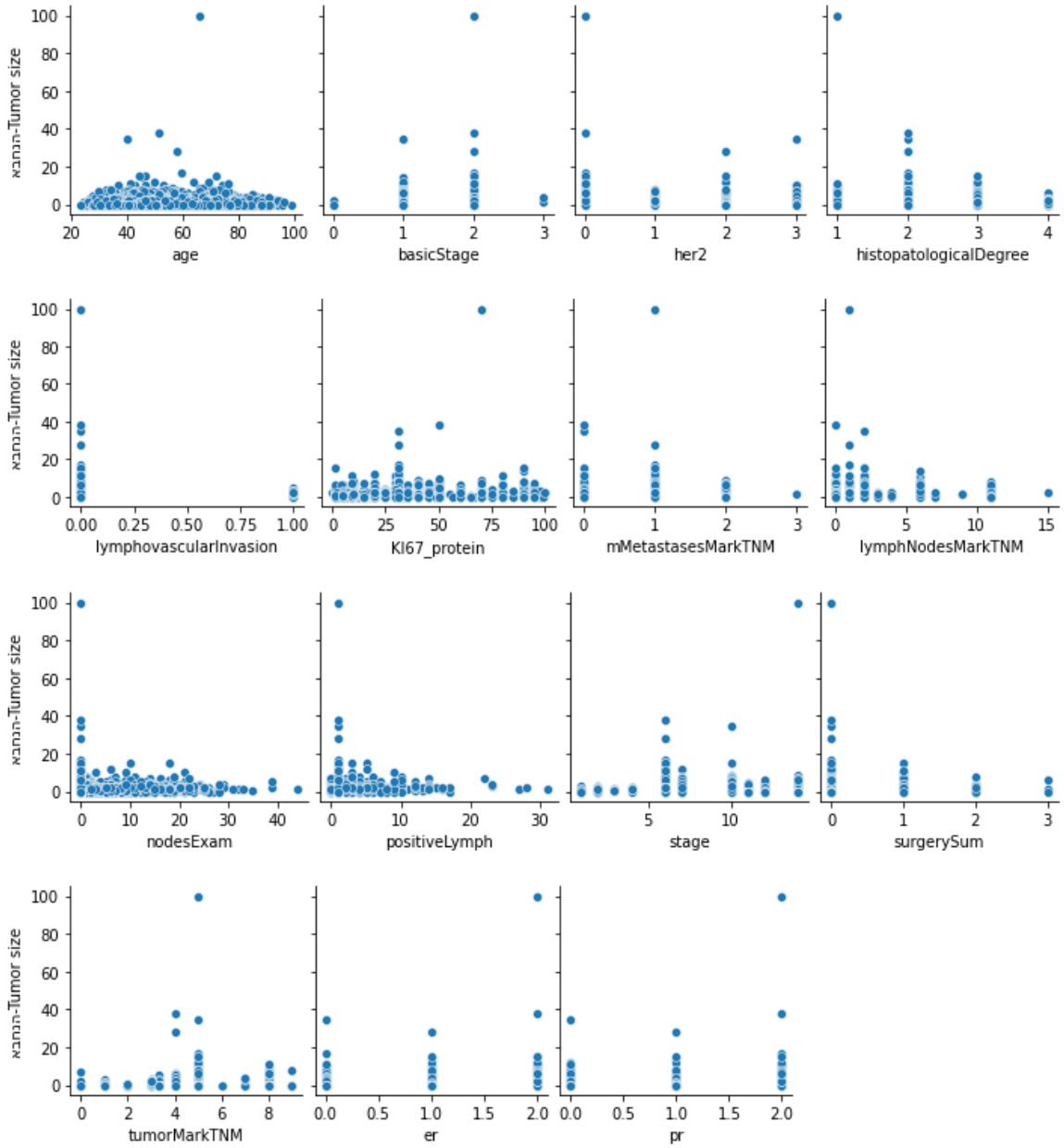

בגרף ניתן לראות את ערכי ה-Loss של ה-test כפונקציה של עומק ה-Random Forest. ניתן להבחין שהשגיאה יורדת ככל שעומק היער גדל עד לרמת עומק 4 ולאחר מכן השגיאה מתחילה לעלות שוב. כפי שלמדנו התנהגות זו היא תולדה של Bias-Variance Tradeoff. עומק נמוך מדי משמעותו מודל פשוט שלא מצליח ללמוד מספיק את סט האימון. מנגד, עומק גבוה מדי משמעותו מודל מורכב אשר עושה Over-fitting על סט האימון. בשני המקרים, טעות ההכללה תהיה גדולה כפי שניתן לראות בגרף.

**גרף מס׳ 3** הקורלציות של הפיצ׳רים עם כמות הגרורות

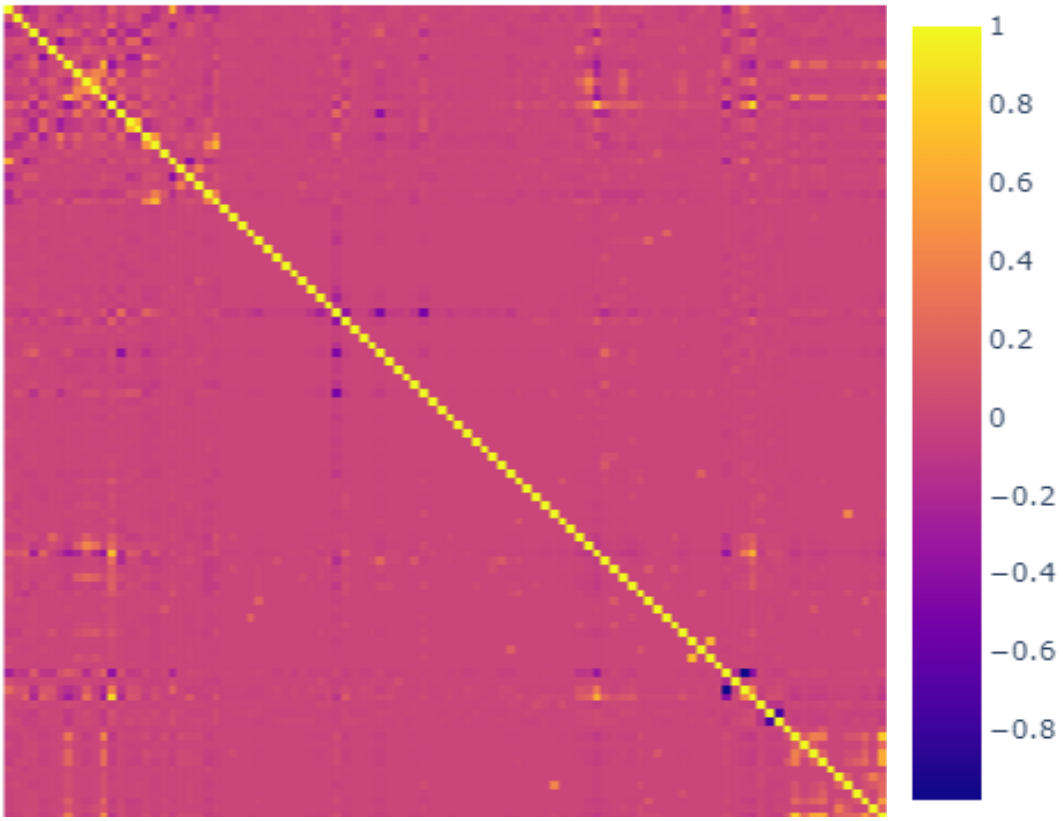



בגרפים ניתן לראות את התנהגות הפיצ׳רים לעומת התנהגות הלייבלים של סט האימון (משמאל) ואת התנהגות הפיצ׳רים לעומת גודל הגידול (מימין). נתמקד למשל בפיצ׳רים stage ו-surgerySum. הפיצ׳ר stage מתאר את השלב של המחלה בהתאם למדדי הרפואה. בגרף המתאים ניתן לראות שדגימות בעלות ערכי stage גבוהים יותר (אשר מתארים בדאטא שלנו שלבים קריטיים יותר של המחלה) תויגו עם מספר גרורות גבוה יותר. הפיצ׳ר surgerySum מתאר את מספר הניתוחים שעבר החולה (0-3). בגרף המתאים ניתן לראות שדגימות אשר צוין שלא עברו ניתוחים כלל או שעברו ניתוח 1 בלבד תויגו עם מספר גרורות גבוה יותר מאחרות.

**גרף מס' 4** הקורלציות של הפיצ'רים עם גודל הגידול



**גרף מס׳ 5** מפת חום של קורלציות הפיצ׳רים והלייבלים




במפת החום ניתן לראות את הקורלציה בין כל הפיצ׳רים וסוגי הלייבלים. הצירים בנויים כך שלאורכם נמצאים הפיצ׳רים ואחריהם הלייבלים (משמאל-לימין ומלמטה-למעלה). התמקדות בפינה הימנית התחתונה מראה שקיימת קורלציה יחסית גבוהה בין הלייבלים השונים (כלומר בין האזורים אליהם יכול להתפשט הסרטן). כמו כן, מהתמקדות בפינה הימנית העליונה ניתן לראות שני פסים כתומים אשר מעידים על קורלציה גבוהה. אחד מהפסים מתלכד עם השורה של הפיצ׳ר stage והעמודות של הלייבלים השונים. ואכן, כמו שראינו בגרף מס׳ 3 הפיצ׳ר stage ערכי למשימה הראשונה.