Breast Cancer data analysis

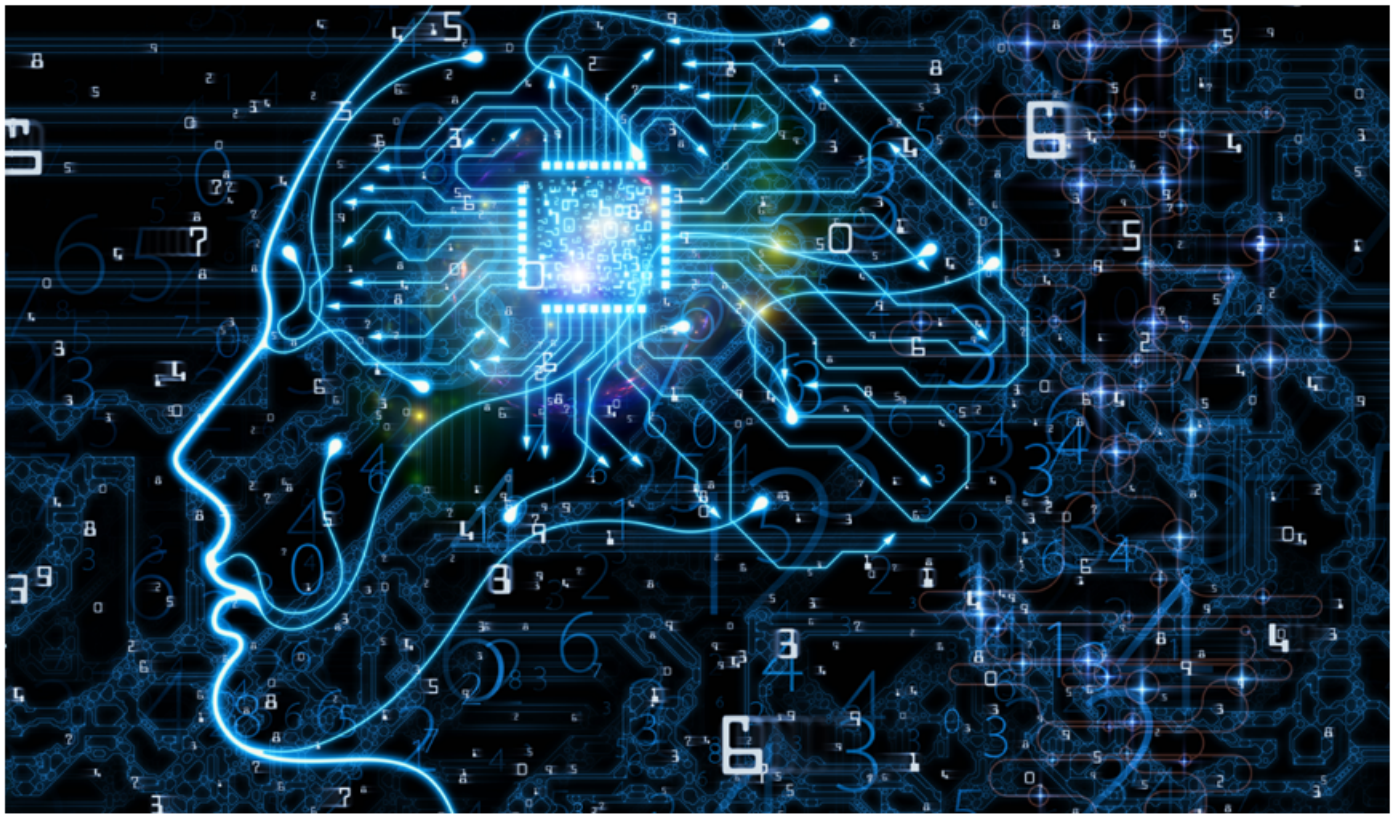
Hackathom IML 2022

Gabi Album – 316563949

Omri Wolf – 204867881

Guy Jascourt – 207090119

Ronel Charedim – 208917641



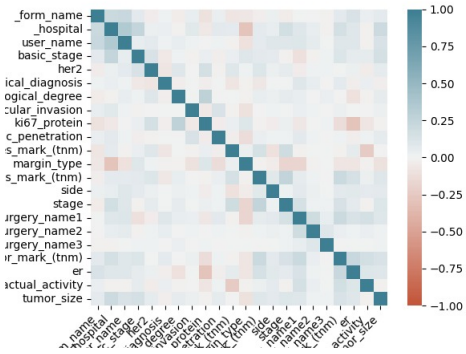
1 $Dataset\ description$:

קיבלנו מערך נתונים המכיל מידע על חולים בסרטן השד. הנתונים כללו מידע על גיל החולים, קצב התפשטות הגידול, תגובת הגידול לחומרים מסויימים, תאריכי ניתוחים ועוד.
כשניגשנו לטפל בנתונים ראשית ניסינו להבין מה המידע שכל פיצ'ר מכיל, באיזה טווח ערכים הוא מדורג, מה הוא מסמן ועוד. לאחר שהבנו מה כל נתון מסמן וכיצד נוכל להפיק ממנו את המירב בכדי שהאלגוריתם הלומד שלנו יוכל ללמוד בצורה מיטבית, התחלנו בניתוח הנתונים.
המידע אותו היינו צריכים לנתח היה מסומן בצורה אידה אלא היה מלא בהבדלים גסים דקים, נדהמנו מהעבודה עם דאטה אמיתי, וכמה הדאטה מגיע לא מסודר.
האתגר הראשון בו נתקלנו היה ניתוח הפיצ'ר $KI67$, שמסמן את קצב התפשטות הגידול. עוד לפני שחקרנו עליו הערכנו כי הוא יהיה פיצ'ר חשוב ואכן לאחר מחקר גילינו את חישובתו. לכן החלטנו לנתח את הנתונים בצורה המיטבית, תחילה היה עלינו להבין באיזו סקאלה ניתן לדרג את קצב התפשטות הגידול. לאחר מכן היינו צריכים למפות את הסטרינגים שבהם המידע נתון לכדי מספרים עליה נוכל להריץ אלגוריתם למידה. חלק מהסטרינגים נכתבו בהיפוך בין עברית לאנגלית חיובי. $= \text{jhuch}$
לאחר שהתקדמנו בפרוייקט הגענו לערך $f1\ score$ גבוה במיוחד - 0.9, חשבנו שאכן מצאנו את המודל שפותר את הבעיה בתצורה הטובה ביותר, אלך לאחר מעבד על הקוד ודיון הבנו כי מדדנו על דאטה בצורה שגויה, וכי ערכי ה ID של אותם החולים הופיעו בסט האימון וסט הטסט יחד. תיקנו את הבעיה בכך שדאגנו שהחלוקה תהיה לחולים עם מספרי ID שונים.

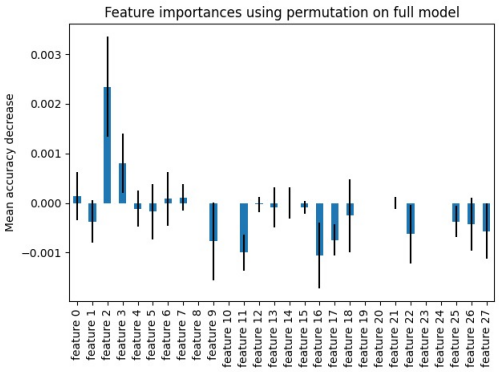
2 $Preprocessing$:

כתבנו מספר פונקציות $parse$ שעיבדו את הנתונים וחילצו את המספרים הרלוונטים, לאחר מכן דירגנו אותם לרמה הרצויה.
פיצ'רים נוספים סימנו בעזרת $fectorize$ שמתייגת כל קבוצה לתווית שונה.
במהלך התהליך ייצרנו פיצ'רים חדשים, כדוגמת פיצ'ר המכיל מידע על ההפרש בין הדיאגנוזה לניתוח הראשון.
היינו צריכים להבין מה לעשות עם ערכים שהיו מסומנים ב $nan\backslash unknown$, כיצד לתייג אותם ואיך להפיק מהם את המירב.

קורולציה בין הפיצ'רים:



חשיבות של הפיצ'רים:



3 *Learning systems design*:

לאחר שפירסרנו את הדאטה וניתחנו את הנתונים, ביצענו מספר בדיקות והרצות כדי להבין איך הדאטה שלנו מתנהג.
בדקנו מספר גישות וייצרנו גרף PCA בו ראינו כי הדאטה אינו לינארי, לכן החלטנו לבחור בפיתוח $baseline$ של $Desition\ tree$ במשימה הראשונה, במשימה השניה בחרנו להשתמש ב $Random\ forest$.
ניסינו להריץ Random Forrest/Adaboost/Decision Tree/KNN/Radius Neighbours,Decision Tree
התוצאות שהשגנו עם שיטות אלו החזירו לנו $score$ בטווח של $0.1-0.2$. מלבד $Desition\ tree$ שהשיג תוצאה גבוה יותר לכן בחרנו להמשיך איתו.
שיפרנו אותו בעזרת שיטת $Cross\ Validation$ שמחלקת את הדאטה לקבוצות ומאמנת את המודל בכל פעם על מספר קבוצות שונות. כך יכלנו למצוא את ה $hiper\ parameters$ המתאימים.

4 *Prediction of the generalization error*:

לאחר שהרצנו על המשימה השנייה הגענו לכך שה־ generalization error הוא 0.004 .
כלומר הגענו למודל שחוזה בצורה טובה את השגיאת הכללה.