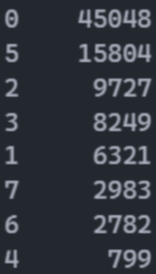
עבודה 1

תיאור כללי של ה-pipeline:

ניתוח מקדים

הכנת נתונים

1. החלפת ערכים חסרים- עבור משתנים רצפים החלפנו בממוצע, ועבור משתנים בדידים החלפנו בערך הנפוץ ביותר.
2. המרת משתנים מסוג STRING למשתנים נומריים.
3. הורדת פיצ'רים שאינם רלוונטים (עמודות המהוות ids)
4. ניקוי המידע-
   1. תחילה ביצענו חילוץ של טווח ההפרש בין הרבעון הראשון Q1 לרבעון שלישי Q3 עבור עמודת גיל.
   2. לאחר מכן סיננו את כל הרשומות אשר ערך הגיל שלהן גדול מהרבעון השלישי ועוד הטווח שחילצנו כפול משקל מסוים.
   3. ביצענו שינויים בנוסחה ובתנאים עד שקיבלנו את הנוסחה שהיתה לנו הכי הגיונית מבחינת התוצאות שלה, והיא-age > Q3 + 0.4\*IQR.
   4. קיבלנו כי:
      1. סה"כ נזרקו 7471.
      2. טווחי הגילאים שנזרקו נע בין 84-89.
      3. מתוך 7471 רשומות שנזרקו, ישנם 1042 מקרי ממוות (שהינם מהווים בערך 13.165% מתוך סך כלל הנפטרים ב-dataset).
   5. בסופו של דבר, קיבלנו החלטה שלא לבצע ניקוי של המידע ולא להשתמש בסינון הזה מאחר ואנחנו מאבדים אחוז יחסית גבוה של נפטרים, וגיל הוא בעיננו אחד הפיצ'רים המשמעותיים שיכול להוות סיבה הגיונית לסיווג או פיצול וקביעת מותו של חולה שהגיע למיון (הגיוני שככל שגיל המטופל עולה כך הסיכוי למותו גדל).
5. הוספת פיצ'רים-
   1. Hospital\_visit\_counts- כמות המבקרים סה"כ באותו בית חולים.
   2. Hospital\_death\_count- כמות המתים סה"כ באותו בית חולים.
   3. Precent\_of\_death\_in\_hospital- אחוז המתים באותו הבית חולים, חושב באמצעות שני הפיצ'רים מעלה (לאחר חישוב זה הם נמחקו).
   4. K-mean column- הוספנו פיצ'יר ע"י שימוש באלגוריתם זה, ככה שכמות הקלסטרים היא לוג בבסיס 4 של כמות הרשומות הקיימות (יצא 8 קלסטרים). להלן התפלגות הפיצ'ר (משמאל- מספר הקלסטר, מימין- כמות הרשומות שמופו לאותו הקלסטר).  
      
6. בחירת פיצ'רים- השתמשנו במבחן ANOVA F value ובפונקציה קיימת SelectKBest של sklearn המאפשרת בחירת פיצ'רים. סיננו את כל הפיצ'רים שה-p\_value שלהם גדול מ-0.01 . סה"כ נשארנו עם 196 פיצ'רים.

מידול אבלואציה

5 שיטות הקלספיקציה השונות שבחרנו הן:

1. Neural Network
2. Logistic Regression
3. Naive Bayes
4. Random Forest
5. KNN

ביצוע של 10-fold-cross-validation. לכל fold בוצעו השלבים הבאים:

1. דסקרטיזציה- לכל פיצ'ר נומרי בחרנו את כמות הbins (המינימום בין 10 לכמות הערכים הייחודים) ואת שיטת החלוקה עומק שווה.
2. אימון על גבי סט האימון (תחת כל fold בנפרד) של כל אחד מלחמשת סוגי המודלים.
3. הערכת המודל בעזרת סט המבחן.