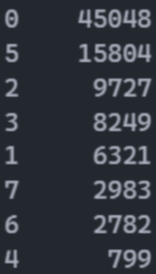
עבודה 1

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

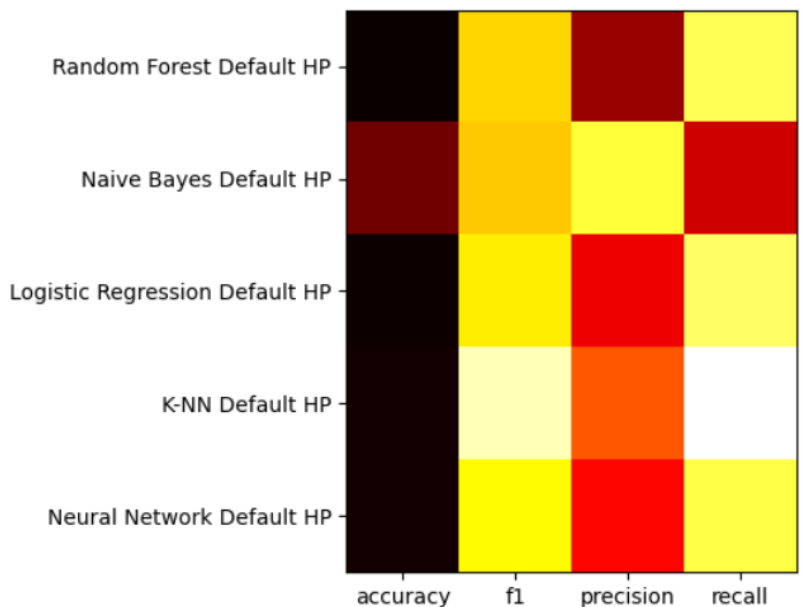
ביצוע של 10-fold-cross-validation:

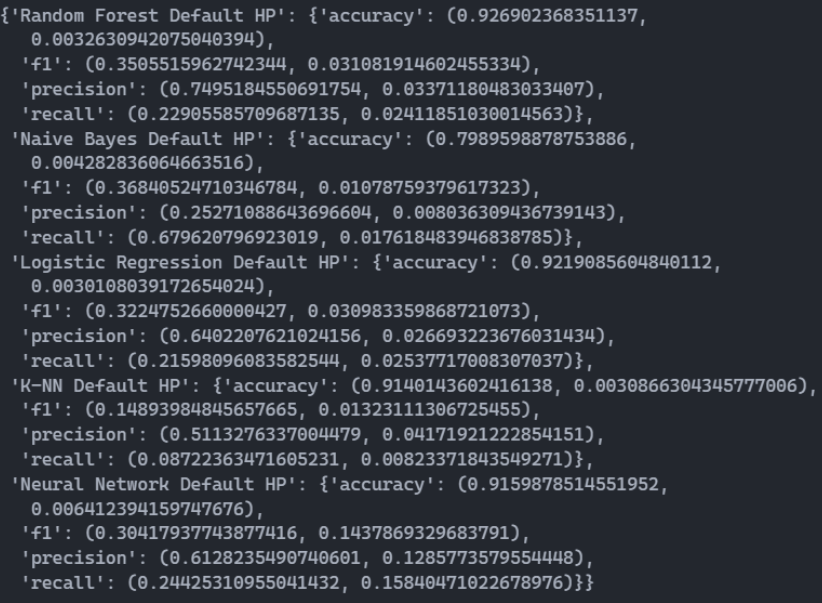
הגדרנו מחלקה הנקראת FOLD. המחלקה מייצגת fold ספציפי מתוך העשרה ומופע שלה מכיל את סט האימון והמבחן וטווחי ה-binnings. לבסוף יצרנו קובץ Pickle שישמור את אותם הfolds כדי שבמידה ונרצה להרי את הקוד מהתחלתו, כל החלק של הדסקרטיזציה (שהיה ארוך מאוד) יחסך ונקרא פשוט את הקובץ פיקל עם האובייקטים המייצגים את הfolds.

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 מחמשת סוגי המודלים.
3. הערכת המודל בעזרת סט המבחן.

תחילה ביצענו את השלבים המתוארים מעלה עבור היפר-פרמטרים דיפולטיביים עבור כל אחד מהמודלים שבחרנו. סה"כ לאחר חישוב משוקלל של מדדי הערכה על גבי כל הfolds תחת כל אלגוריתם הניבו את התוצאות הבאות:

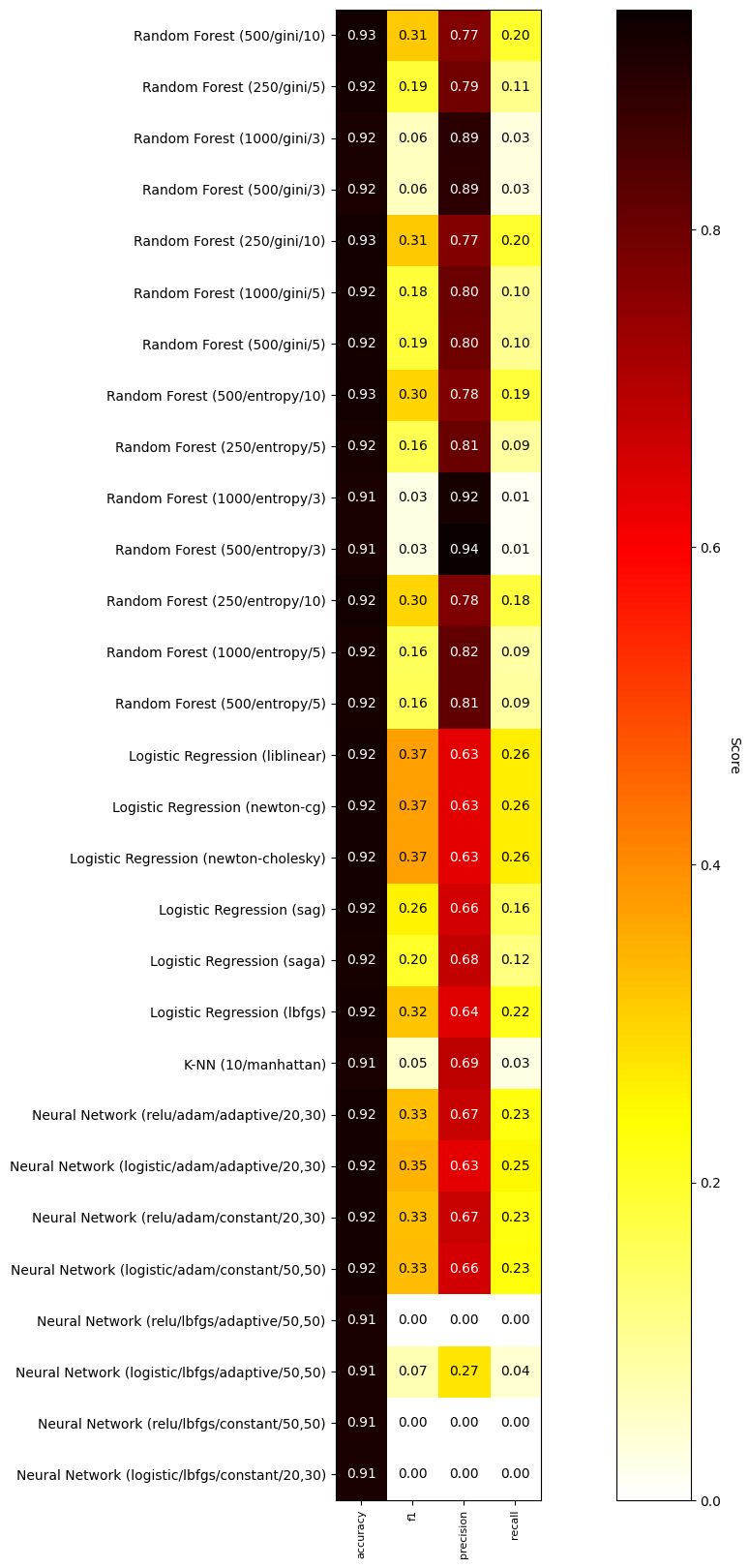




אפשר לראות שמדד הדיוק הוא מאוד גבוה בכל המדדים בזמן שמדדים אחרים די חלשים יחסית.

נראה כי Naïve base יוצא דופן- הוא היחיד שמדד הrecall שלו גבוה יבחס לשאר, ומדד הprecision שלו נמוך ביחס לשאר.

לאחר הרצה ראשונית, התחלנו "לשחק" עם ההיפר-פרמטרים של כל מודל.



בחירת האלגוריתם:

נראה כי מבחינת דיוק, לכלל המודלים יש ביצוע דומה וכי מדד הדיוק גבוה בכולם.

בהיבט של recall אל מול precision, לדעתנו מדד הprecision הינו חשוב יותר למודל שנבחר מאחר וחשוב לנו יותר לזהות את אלו שישארו בחיים לא פחות מאשר לזהות את אלו שימותו. נסביר- נסתכל על מקרה הקיצון בו על כל חולה נחזה שימות. כלומר משמעות מקרה זה הוא recall גבוה. בהנחה והמודל שלנו יעזור לרופאים במיון להחליט איך לחלק את תשומת הלב שלהם, עדיף להפנות את המשאבים למקרים בהם הסיכויים לחיות גדולים יותר. לכן מקרה הקיצון המתואר פחות עדיף במקרה זה, ונרצה שהמודל פחות יחמיר ויוכל לחזות גם במקרים בהם החולה לא ימות.

תחילה נבחר עבור כל מודל את ההיפר-פרמטרים האופטימלים:

1. Neural Network:
   1. הפרמטר solver עם הערך lbfgs הניב את התוצאות הכי נמוכות (0 בכל המדדים).
   2. שאר תוצאות ההרצה די דומות, ולכן ניקח את קומבינציית הפרמטרים שממקסמת את הprecision (על חשבון הrecall).
   3. קומבינציית הפרמטרים הסופית- relu, adam, adaptive, 20, 30
2. Logistic Regression:
   1. קומבינציית הפרמטרים הסופית- saga
3. Naive Bayes:
   1. קומבינציית פרמטרים הסופית-
4. Random Forest:
   1. קומבינציית פרמטרים הסופית- 500, entropy, 10
5. KNN:
   1. קומבינציית פרמטרים הסופית-

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ACC | F1 | PRECISION | RECALL |  |
| ANN | 0.92 | 0.33 | 0.67 | 0.23 |  |
| LR | 0.92 | 0.2 | 0.68 | 0.12 |  |
| NB |  |  |  |  |  |
| RF | 0.93 | 0.3 | 0.78 | 0.19 |  |
| KNN |  |  |  |  |  |