IML Hackathon

הדאטא והאתגרים בו:

הדאטא היה מאתגר ממספר בחינות, ראשית, איתרנו כפילויות רבות. עבור אותה חולה ואותו ביקור למעשה הופיעו מספר סמפלים שונים. כמו כן, היו חסרים נתונים רבים ואתגר משמעותי היה להחליט איך למלא את הפרטים החסרים.

ניקוי ופרה-פרוססינג:

בשלב הראשון, איחדנו את הכפילויות שהופיעו בדאטא, ומספר סמפלים שונים שהיו חלק מאותו ביקור אוחדו לאחד. לאחר מכן החלפנו משתנים קטגורים בone-hot והחלפנו תאים ריקים בערכים בעלי משמעות ואיחדנו מהויות דומות בפיצ'ארים מסויימים. בנוסף, התמודדנו עם עמודות שבהן היו ערכים שונים שהוקלדו ללא בקרה והפכנו אותן לבעלות סטנדרט אחיד. לאחר מכן יצרנו מספר פיצ'רים שונים בעלי משמעות כגון, יחס בין כמות קשרי לימפה חיוביים לבדיקה וקשרי לימפה שנבדקו, מספר ימים מהניתוח האחרון ופיצ'רים נוספים.

בניית המודל

ראשית, יצרנו מודל בייס-ליין על מנת שנוכל להשוות אליו מודלים מתקדמים יותר. עבור הלייבלים של מיקומי הגרורות, פיצלנו את העמודה לעמודה אחת עבור כל מיקום אפשרי (11 מיקומים אפשריים) של גרורה ואימנו מודל random-forest עם הפרמטרים הדיפולטים של sklearn עבור כל עמודה. הפרדיקציה הסופית היא איחוד של כל המיקומים שכל מודל חזה בנפרד. שגיאה:

```
INFO:root:Micro f1 = 0.6827118644067797
 Macro f1 = 0.5636112828782252
INFO:root:DONE
```

בעבור המשימה השניה, חיזוי גודל גידול, השתמשנו כבייס-ליין ברגרסיה לינארית של sklearn עם פרמטרים דיפולטים. שגיאה:

```
Std: 2.3460966697527867
RMSE: 2.1234245636583324
```

לאחר מכן השתמשנו בבחירת פיצ'רים של sklearn וגילינו כי שימוש ב36 פיצ'רים הטובים ביותר נותן את התוצאות הטובות ביותר עבור רגרסיה:

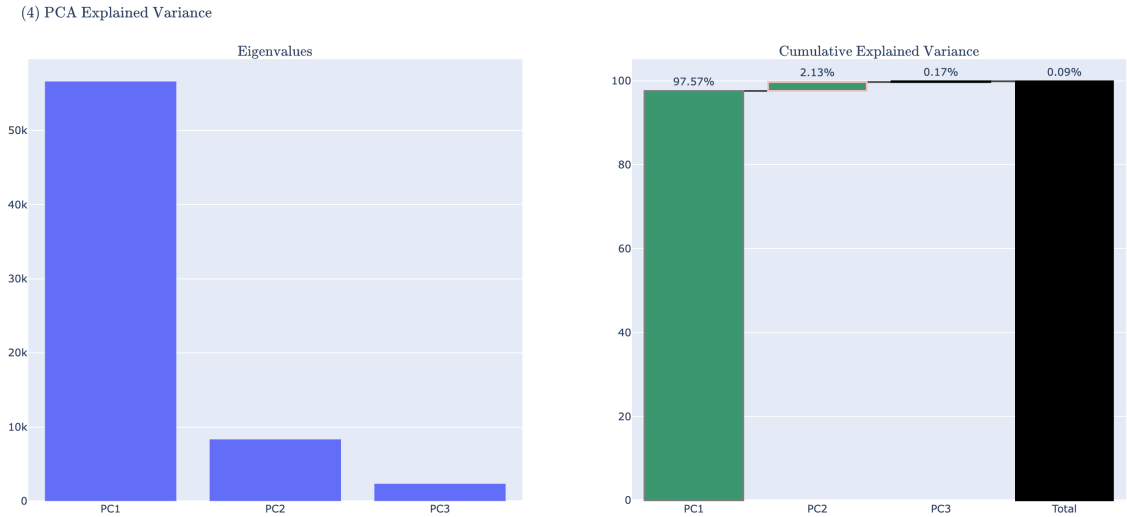
```
Num of features: 53
Best error: 1.6973002037913358 is for 36 num of features
```

במשימה השנייה, של חיזוי גודל הגידול, מכיוון שמדובר ברגרסיה, קיבלנו גם גדלים שליליים. על כן ביצענו ReLU לערכים, כלומר לקחנו עבור כל ערך את המקסימום בין 0 לבין הערך.

לכל אחת מהמשימות, החלטנו לבצע comitee של מספר מודלים שונים, ולקחנו את הרוב/הממוצע (בהתאמה לסוג החיזוי הרצוי).

במהלך ניתוח הפיצ׳רים ניסינו להבין מי הם הפיצ׳רים הקורלטיביים ביותר, עשינו את הניתוח הזה בשתי דרכים

1) ויזואלית



2) בדיקת קורלציה באמצעות corr = cov / std_feature / std_y כפי שעשינו בתרגיל של ה-house proce prediction:

```
Correlation between feature Age and label OTH - Other is: -0.00042417258706185 44
Correlation between feature KI67_protein and label ADR - Adrenals is: 1.6125459015436185e-18
Correlation between feature KI67_protein and label BRA - Brain is: 6.066084305521186e-18
Correlation between feature N_lymph_nodes_mark_(TNM) and label MAR - Bone Marrow is: -7.85062098825597e-05
Correlation between feature Surgery_date3_diff and label ADR - Adrenals is: -0.00014194438881014578
Correlation between feature Surgery_date3_diff and label SKI - Skin is: -0.0004710985296755377
Correlation between feature Surgery_date3_diff and label PLE - Pleura is: -0.00034781026239476 22
Correlation between feature Surgery_date3_diff and label BRA - Brain is: -0.0003757037298307953
Correlation between feature Surgery_date3_diff and label PER - Peritoneum is: -0.00034781026239476243
Correlation between feature Surgery_date3_diff and label OTH - Other is: -0.00034781026239476254
Correlation between feature Surgery_date3_diff and label MAR - Bone Marrow is: -0.00014194438881014 58
Correlation between feature QUADRANTECTOMY_surgery_1 and label ADR - Adrenals is: -0.00033515628728820416
Correlation between feature QUADRANTECTOMY_surgery_1 and label MAR - Bone Marrow is: -0.0003351562872882043
Correlation between feature OOPHORECTOMY_surgery_1 and label ADR - Adrenals is: -0.00013678019166957897
Correlation between feature OOPHORECTOMY_surgery_1 and label SKI - Skin is: -0.00045395910133836195
Correlation between feature OOPHORECTOMY_surgery_1 and label PLE - Pleura is: -0.00033515628728820427
Correlation between feature OOPHORECTOMY_surgery_1 and label BRA - Brain is: -0.00036203493923219154
```