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.