תוצאות (שימוש ב random forest):

תוצאות ראשוניות (ללא over sample, רק הורדה של age\_60\_and\_above):

Text

Description automatically generated

Chart, treemap chart

Description automatically generated

ניתן לראות בבירור כי האלגוריתם "מעדיף" לתייג דגימות כחיוביות בגלל החוסר איזון במדגם. מספר ה false negative מאוד גבוה. במקרה שלנו (איבחון קורונה) מאוד חשוב להפחית את ה – false negative כי לא נרצה שחולי קורונה יסתובבו ברחובות ויגרמו להתפרצות.

תוצאות עם oversampling למדגם שבו corona\_result=1 (ה minority). Ratio = 1:

Text

Description automatically generated

Chart

Description automatically generated

בגלל שהמדגם לא מאוזן אז המודל פשוט מתייג את רוב הדוגמאות כ – negative ובגלל זה מקבלים accuracy ו – precision גבוהים. ברגע שנאזן את המדגם נראה ירידה משמעותית של precision כי יהיו fp ואין הרבה tp כי פשוט אין הרבה דגימות שהן positive.

תוצאות עם הוספת מאפיין – test\_week שמתאר באיזה שבוע בשנה נלקחה הבדיקה:

Text

Description automatically generated with low confidence

Chart

Description automatically generated

ניתן לראות במחברת כי הפילוג שמתקבל (הסיכוי להיות חולה בהינתן השבוע) הוא הטרוגני ואכן ניתן לראות שהפיצר שיפר את התוצאות.

הרצת האלגוריתם עם threshold אופטימאלי שנמצא באמצעות ROC curve:

Text

Description automatically generated

Chart

Description automatically generated

בגלל שהמדגם לא מאוזן יש "יתרון" לקבוצה של הרוב (הדגימות השליליות), כלומר האלגוריתם יעדיף לתייג דגימות כשליליות (לתת הסתברות גבוהה לשלילי ונמוכה לחיובי), ולכן נרצה לשנות לו את ה threshold. נרצה להקטין אותו כי כך יתקבלו תגיות חיוביות עבור הסתברויות נמוכות יותר וכך יופחת ה false negative.

תוצאות Xgboost:

לפני אופטימיזציה של הפרמטרים:

A picture containing table

Description automatically generated

Chart, treemap chart

Description automatically generated

לאחר סיבוב ראשון של GridSearchCV:

הפרמטרים:



תוצאות:

Text

Description automatically generated

Chart

Description automatically generated

לאחר סיבוב שני:

פרמטרים:



תוצאות:

Text

Description automatically generated

Chart, treemap chart

Description automatically generated