תרגול 10

שלבים של :pre processing לעשות one hot encoding על משתנים קטגוריים, להוריד משתנה ייחודי לכל שורה כמו תעודת זהות, לבדוק קורלציה ולהוריד תכונות עם קורלציה גבוהה, השלמת נתונים, לבצע נורמליזציה למספרים בטווח מסוים, לדוגמה בין 0 ל1.

להוסיף לדף נוסחאות לכל אלגוריתם אם יש משהו מיוחד שצריך לעשות כדי להשתמש בו

לדוגמה לSVM נדרש לעשות multiclass, SVM לא עובד עם תכונות קטגוריות ולכן לעשות one hot encoding.

אלגוריתמים של עצים יכולים לעבוד עם משתנים קטגוריים. עצים לא יודעים לעבוד עם עצים חסרים ולכן נדרש לעשות imputation כדי להשלים ערכים חסרים.

KNN עובד יותר טוב עם תכונות שעשו עליהם scaling

אם בתהליך האימון המודל מאוד מדויק ועל הסט מבחן הוא ממש לא מדויק. הסיבה לכך יכולה להיות שהנתונים לא מאוזנים- לעשות preprocessing לlabel, או overfitting- צריך להשתמש בהיפר פרמטרים אחרים.

כאשר יש הרבה מאוד פיצ'רים יש זמן ריצה ארוך מאוד, יכול להיות שיש הרבה פיצ'רים זבל, לכן מחפשים קורלציות על מנת להוריד פיצ'רים או להשתמש בPCA על מנת להוריד ממדים.

gridsearchCV בודק את כל השילובים האפשריים בין פרמטרים- לוקח המון זמן, אם חשוב יותר הדיוק אז נשתמש בgridSearch

RandomizedSeachCV לוקח שילובים אפשריים רנדומליים, לא עובר על כל השילובים, מאפשר זמן ריצה קצר יותר במקרים שבהם הדיוק פחות חשוב והזמן ריצה יותר חשוב.

Confusion Matrix

להכניס לחומר פתוח: false positive ,false negative, את המשוואות שלהם

Sensitivity- בודק כמה טוב אנחנו מזהים משהו חיובי (לדוגמה אם מגדירים זיופים כחיובי אז נרצה לבדוק את הsensitivity).

צריך סט ולידציה כדי לעשות cross validation של המודל על מנת למצוא את ההיפר פרמטרים שבונים את המודל הכי טוב. לאחר מכן בודקים את התוצאה של המודל הכי טוב על סט המבחן.

בSVM אם ניתן לclass מסוים משקל גדול יותר כך נגדיל את החשיבות של לסווג קלאס מסוים.

לדוגמה יותר חשוב לזהות קלאס של טפסים שמסווגים כמזויפים.

עצי החלטה

כאשר אומרים לנו לבנות עץ החלטה נדרש לחלק את הגרף לקווים על פי הקלאסים , כל קו הוא עוד עומק בעץ ובונים את העץ לפי זה.