פרוייקט גמר למידת מכונה

מגישים: ישי לוי ותהילה עבאדי

הפרוייקט עוסק במידע מתוייג על אנשים והאפשרות שלהם לחוות שבץ.

על ידי שימוש בלמידת מכונה, למדנו את המאפיינים של אנשים שחלו בשבץ וכאלה שלא, וניסינו לסווג קבוצה נוספת של אנשים לקטגוריות אלו.

Data:

הדאטה של הפרוייקט שלנו מתעסק באנשים שחלו בשבץ וכאלה שלו, ומכיל עשרה פרמטרים בנוסף לתגית הזיהוי.

המאפיינים:

id

hypertension

heart_disease

age

ever_married

avg_glucose_level

work_type

bmi

gender

smoking_status

ניתן לראות שחלק מהמאפיינים רפואיים ומדעיים לחלוטין (כמו bmi ורמת גלוקוז) וחלק נראים שרירותיים לגמרי כמו סוג עבודה או האם אי פעם התחתן.

preprocessing of the data:

תחילה, המרנו את כל הנתונים שניתנו לנו כמחרוזת לנתונים מספריים כיוון שרוב האלגוריתמים לא מסוגלים לעבוד עם מחרוזות.

לדוגמא:

```
df["gender"].replace({"Male":0,"Female":1,"Other":2}, inplace=True)
df["ever_married"].replace({"Yes":0,"No":1}, inplace=True)
```

לאחר מכן, על ידי שימוש ב MinMaxScalar נרמלנו את כל הדאטה לפרמטרים בין 0-1 כדי לתת לאלגוריתמים ביצועים טובים יותר.

Feature Selection:

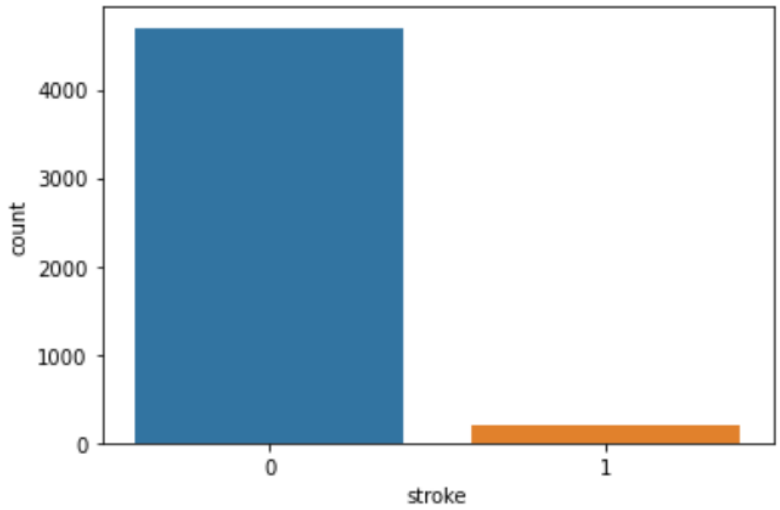
כדי להתחיל לעבוד עם הדאטה הנתון, יש צורך בבחירת הפרמטרים הרלוונטיים ביותר לתהליך גם כדי לא לבצע בטעות overfitting וגם כדי לא להשתמש בפרמטרים שלא משנים כמעט בכלל.על ידי שימוש בSelectKBest נתנו ניקוד לכל פרמטר בהתאם לקורולציה שלו עם התיוג, ואלה הנתונים שהתקבלו:

hypertension	90.543821
heart_disease	88.779204
age	38.452374
ever_married	35.384105
bmi	18.814363
avg_glucose_level	17.198932
work_type	11.740406
gender	0.099242
smoking_status	0.026097

באופן מאוד מפתיע גילינו לדוגמא שלשאלה האם אדם היה נשוי אי פעם או לא, יש הרבה יותר קשר לשבץ מאשר האם הוא מעשן.

כך בחרנו את חמשת הפיצ'רים המרכזיים לעבודה בפרוייקט שלנו.

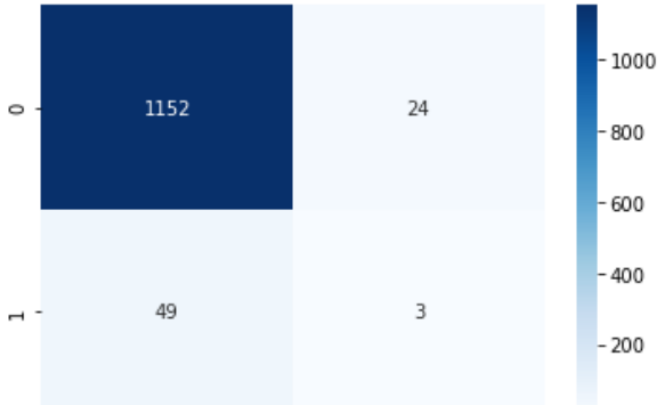
Imbalanced Data:



בתמונה זו ניתן לראות את החלוקה של הדאטה המתוייג. כמעט כולו שייך לאלו שלא חלו בשבץ.

כמעט כל האלגוריתמים שניסינו להריץ תייגו את כל המידע לתיוג 0, כך שאחוז הדיוק הכללי היה גבוה מאוד, אבל התיוג הפרטני של קבוצת 1 היה מאוד מאוד נמוך :

	precision	recall	f1-score	support
0	0.96	0.98	0.97	1176
1	0.11	0.06	0.08	52
accuracy			0.94	1228
macro avg	0.54	0.52	0.52	1228
weighted avg	0.92	0.94	0.93	1228



1: נסיון ראשון ללא איזון הדאטה

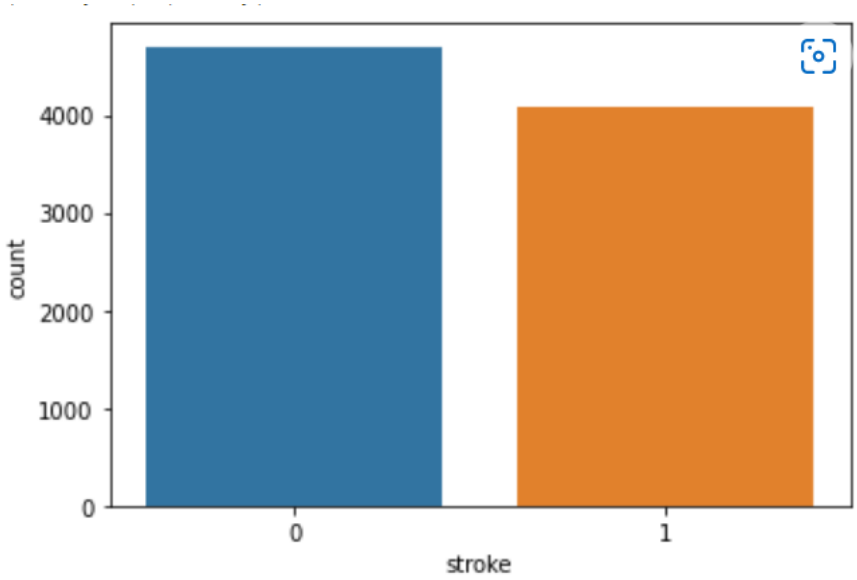
הבנו שהמידע שלנו מאוד לא מאוזן ויש צורך למצוא שיטה שתאזן אותו ותאפשר לאלגוריתמים למצוא דרך לסווג גם את הקבוצה המתוייגת ב-1.

השתמשנו בresample .

תחילה חילקנו את הדאטה לשתי קבוצות, קבוצת 'הרוב' וקבוצת 'המיעוט'

אחר כך גם את קבוצת המיעוט חילקנו לשתי קבוצות של שני שליש ושליש (minority1 and minority2)

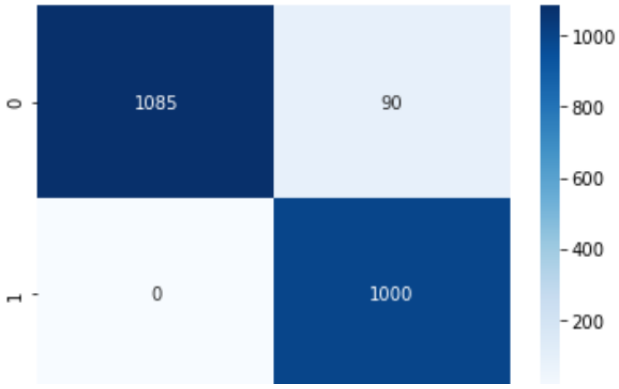
את הקבוצה השניה השארנו כמו שהיא ושמרנו אותו לקבוצת הtest, ואת הקבוצה הראשונה שכפלנו מספר רב של פעמים ולבסוף איחדנו בין כל הרשימות. הדאטה החדש שהתקבל:



ניתן בקלות לראות כי עכשיו המידע הרבה יותר מאוזן והגיוני להריץ עליו את האלגוריתמים.

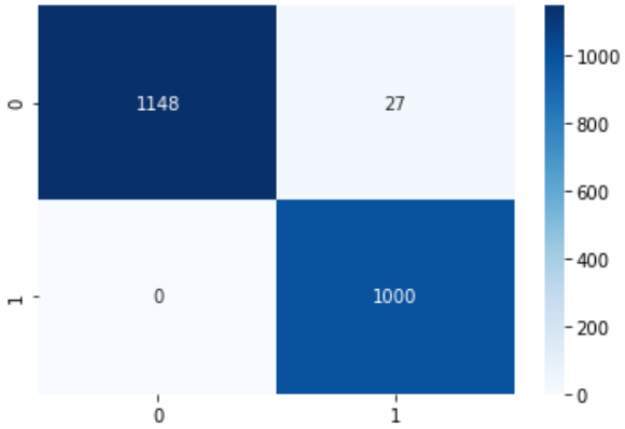
התחלנו עם אלגוריתם Knn כמו שנלמד בכיתה אך עם מימוש הספרייה sklearn והתוצאות שקיבלנו היו טובות מאוד, אבל כל אלגוריתמי העצים נתנו תוצאות טובות יותר:

	precision	recall	f1-score	support
0	1.00	0.92	0.96	1175
1	0.92	1.00	0.96	1000
accuracy			0.96	2175
macro avg	0.96	0.96	0.96	2175
weighted avg	0.96	0.96	0.96	2175



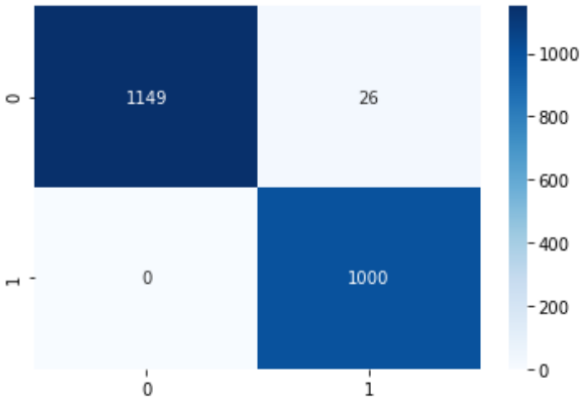
KNN

	precision	recall	f1-score	support
0	1.00	0.98	0.99	1175
1	0.97	1.00	0.99	1000
accuracy			0.99	2175
macro avg	0.99	0.99	0.99	2175
weighted avg	0.99	0.99	0.99	2175



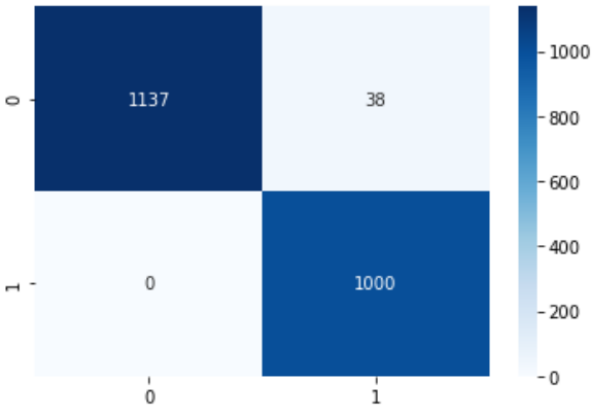
RandomForestClassifier

	precision	recall	f1-score	support
0	1.00	0.98	0.99	1175
1	0.97	1.00	0.99	1000
accuracy			0.99	2175
macro avg	0.99	0.99	0.99	2175
weighted avg	0.99	0.99	0.99	2175



ExtraTreesClassifier

	precision	recall	f1-score	support
0	1.00	0.97	0.98	1175
1	0.96	1.00	0.98	1000
accuracy			0.98	2175
macro avg	0.98	0.98	0.98	2175
weighted avg	0.98	0.98	0.98	2175



XGBClassifier