**פרויקט בלמידת מכונה לדיאגנוסטיקה קלינית**

**סידור הנתונים והרצת המודלים:**

**קריאת נתונים והכנתם ללמידה:**

ראשית הורדנו ידנית את השורות הלא רלוונטיות בתחילת הקובץ AMI\_GSE66360\_series\_matrix כדי להשאר רק עם הנתונים הרלוונטים.

אח"כ עברנו לעבוד בגוגל קולב, שם קראנו את הטבלה, הורדנו שורות שחוזרות על עצמן ושמנו את שמות הגנים ככותרות לעמודות.

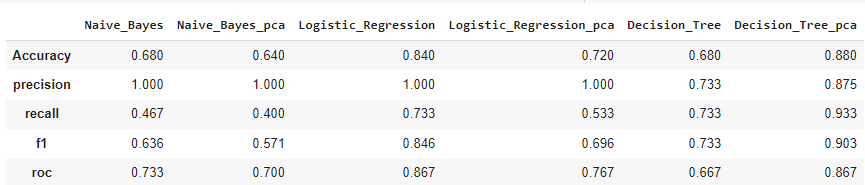
שחלפנו את הטבלה וסדרנו אינדקסים טובים.

כדי ליצור את הtarget יצרנו קובץ csv חדש המכיל את מספר הפציינט וסיווגו לחולה או בריא. את הקובץ הזה קראנו לתוך y.

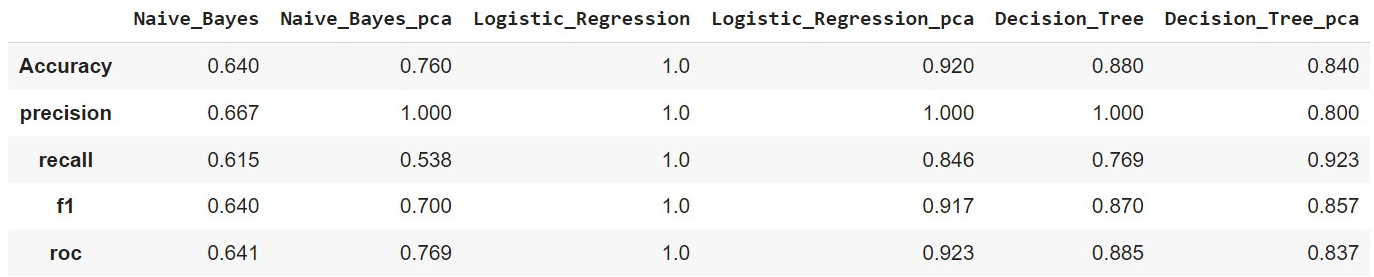
הכנסו לפונקציית train\_test\_split את הטבלה המשוחלפת (x) ואת y כדי לקבל חלוקה. חילקנו את הדאטא ל 25% test ו 75% train עם רנדומליות 25. (הגענו למסקנה שערכים אלו עדיפים לאחר שהרצנו מודלים עם יחסי חלוקות שונים ורנדומליות שונה)

הצגת נתוני ההרצות שלנו לבחירת הפרמטרים לאופן חלוקת הנתונים:

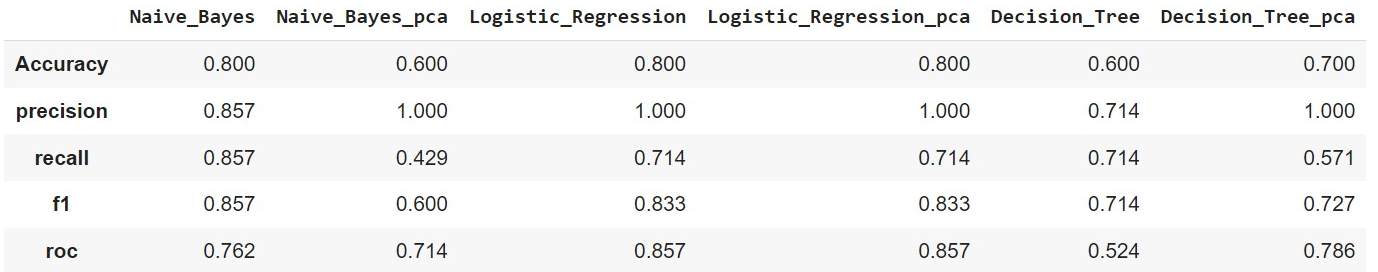
רנדומליות 25 train=75%, test=25%:

****

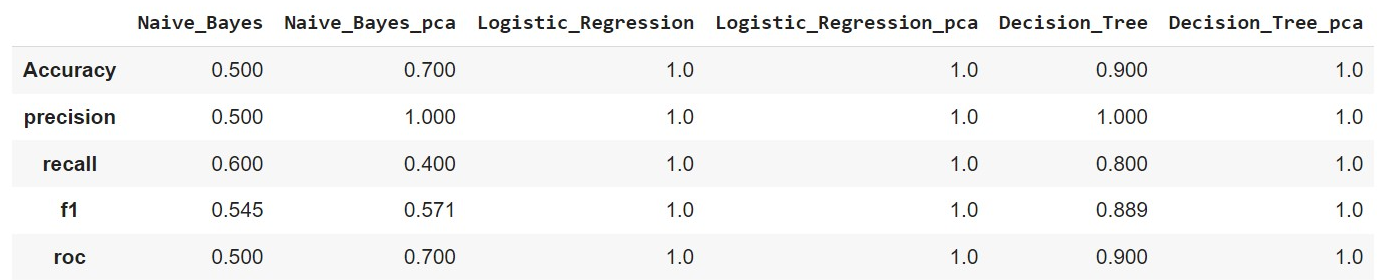
רנדומליות 0 train=75%, test=25%:

****

רנדומליות 25 train=90%, test=10%:

****

רנדומליות 0 train=90%, test=10%:

****

אמנם נראה כי לפי הפרמטרים של רנדומליות 0 train=90%, test=10% אנו מקבלות את התוצאות הטובות ביותר, אבל כיוון שאם לא ניקח רנדומליות אנו מסתכנות בover fitting- שהמודלים יתאימו את עצמם יותר מידי לדאטא שלנו, לכן אנו מעדיפים חלק את הדאטא לפי רנדומליות 25 train=75%, test=25% שנותן את הערכים הטובים ביותר מבין האפשרויות עם הרנדומליות.

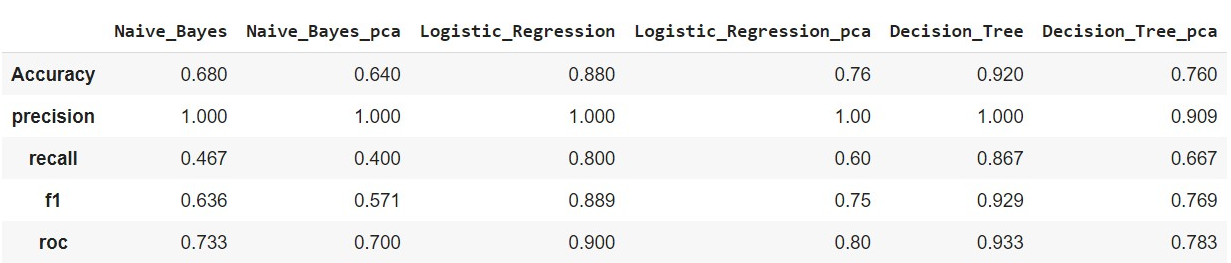
**הורדת מימדים:**

בצענו את הורדת המימדים באמצעות pca כאשר חיפשנו כמה מימדים יסבירו לנו 90% מהשונות של הנתונים. קיבלנו כי 42 מימדים (גנים) מסבירים כ 90% מהשונות ולכן בצענו pca כאשר n\_components

=42.

כדי לראות אם הורדת המימדים משפרת את ביצועי המודלים הרצנו מספר מודלים על הנתונים לפני ואחרי הורדת המימדים. סה"כ ראינו תוצאות טובות יותר במודלים שהורצו על הנתונים ללא הורדת המימדים.

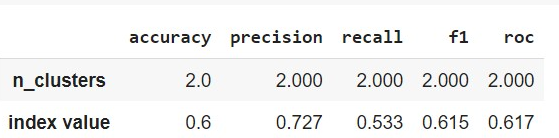
תוצאות ההשוואה:



לכן ניסינו לבצע הורדת מימדים באמצעות k-means, תחילה חיפשנו באמצעות הפונקציה GridSearchCV

לאיזה מספר מימדים להוריד בין 2-10, קיבלנו כי קליסטור הנתונים ל2 מרכזים נותן את התוצאות הטובות ביותר.

טבלה המתארת את תוצאות GridSearchCV כאשר כל פעם המדד הוא index value:



בצענו k-means עם n\_clusters=2 והוצאנו את המרכזים. אך לא הצלחנו להמשיך להריץ את המודלים על המרכזים הללו.

לכן את ניתוח התוצאות נבצע באמצעות הרצת המודלים על הנתונים ללא הורדת המימדים.

נבחר את המודלים decision tree ו- logistic regression ואותם ננתח:

**Decision Tree**

תחילה בחרנו את ההיפר פרמטרים הנותנים את התוצאות הטובות ביותר באמצעות ()GridSearchCV וקיבלנו כי ההיפר פרמטרים הטובים ביותר הם:

criterion: entropy

max\_depth: 20

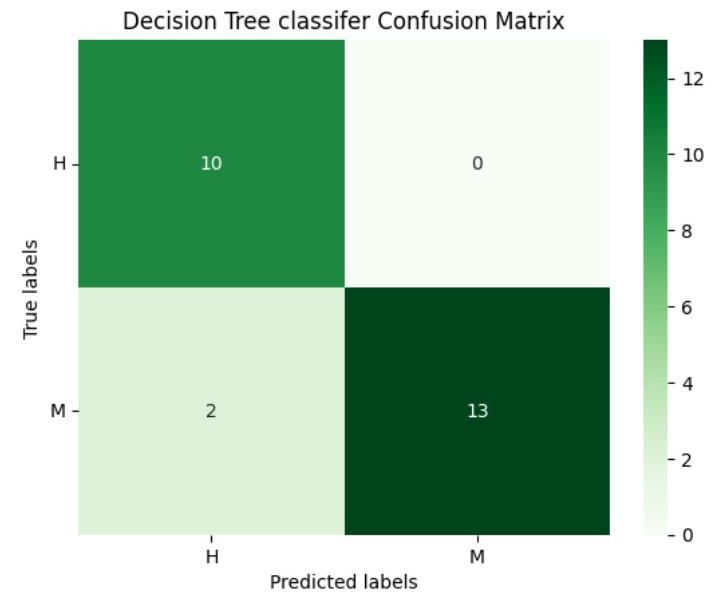
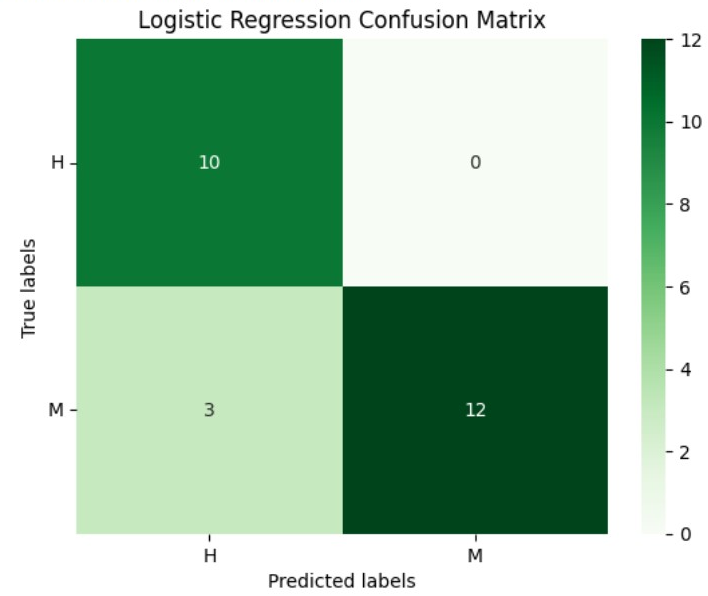
splitter: random

**Logistic Regression**

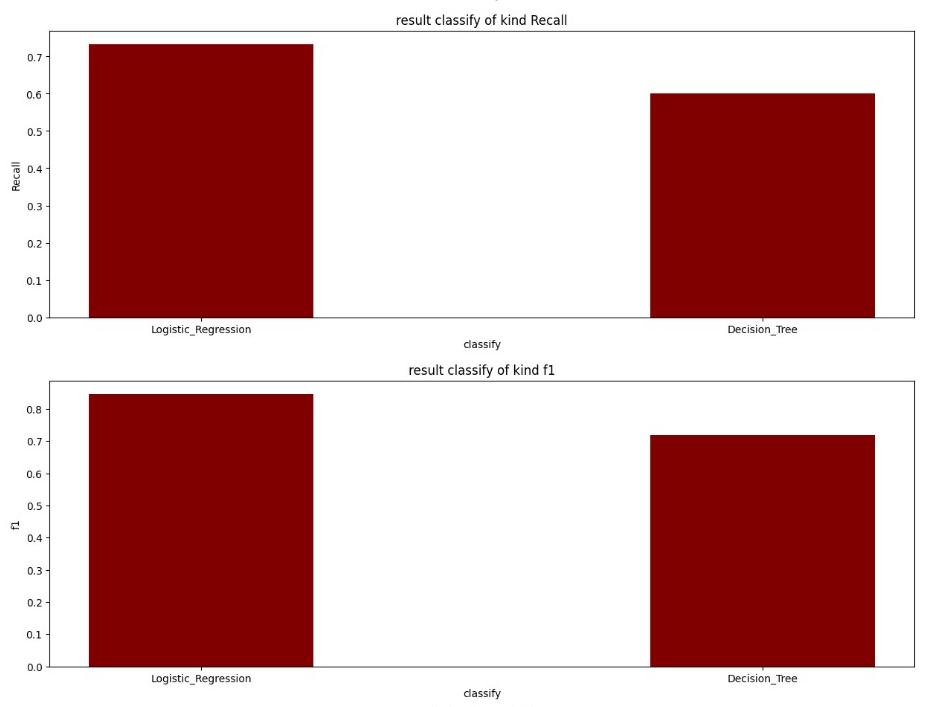
גם כאן בחרנו את ההיפר פרמטרים באמצעות ()GridSearchCV וקיבלנו כי ההיפר פרמטרים הנותנים את התוצאות הטובות ביותר הם:

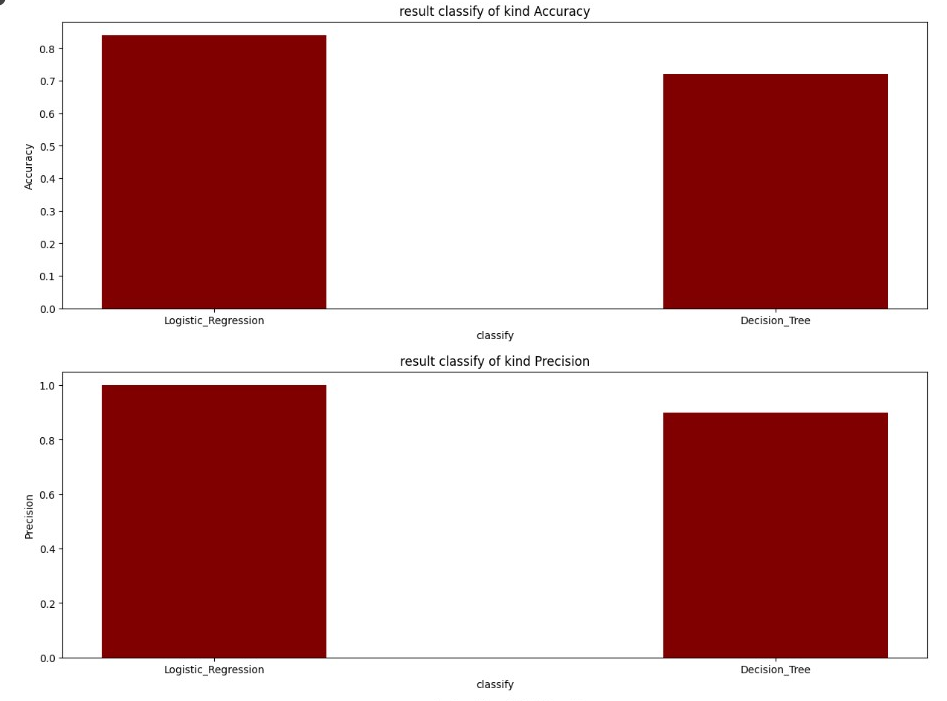
C: 0.1

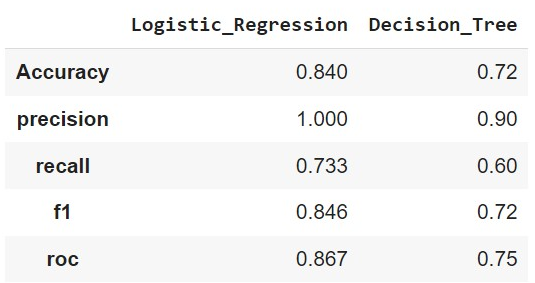
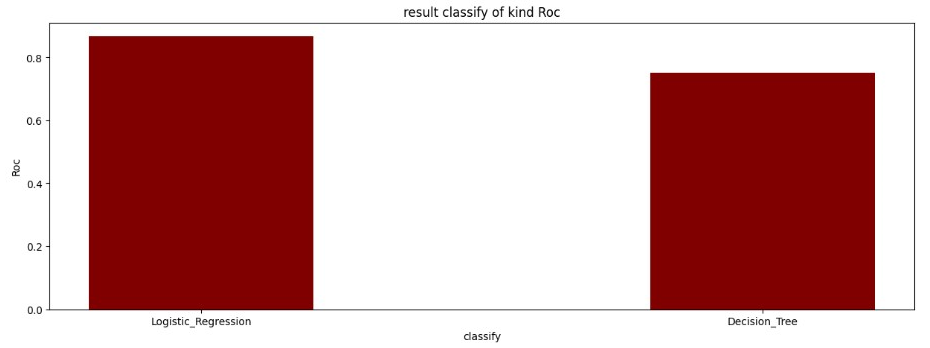
penalty: l1

**הצגת התוצאות ומסקנות:**

ניתן לראות כי בשני המודלים החיזוי נכון למדי. בlogistic regression עשרים ושתיים מתוך 25 נתונים חוזו נכון, וב decision tree עשרים ושלוש מתוך 25 חוזו נכון.

**ניתוח תוצאות לפי המדדים accurancy, recall, precision, f1, ו roc:**





לפי כל המדדים המוצגים בגרפים לעיל ובטבלה, ניתן לראות כי logistic regression נותן את החיזוי הטוב ביותר, אך גם decision tree נותן חיזוי לא רע.

**מסקנות:**

בגדול נראה כי ניתן לחזות סכנה להתקף לב באחוזים די גבוהים (האחוזים הנתונים בטבלה לעיל, ע"פ מדדים שונים) ע"פ רמת ביטוי הגנים של 54,628 הגנים שעליהם ביצענו את המודלים, אך זה לא פרקטי ולא סביר. בכדי שניתן יהיה לחזות סכנה להתקף לב, צריך להבין מי הם הגנים המרכזיים המשחקים כאן תפקיד, וזאת ניתן ע"י כך שנמצא מספר גנים שאם נריץ רק עליהם את המודלים הנ"ל התוצאות לא ישתנו או תהיינה טובות יותר, כיוון שנוריד הרבה רעש מיותר. חיפשנו הרבה איך להצליח לרדת למספר מימדים קטן יותר, (ניסינו כמעט את כל האפשרויות המפורטות בקובץ ההדרכה לתרגיל) הצלחנו לרדת ל42 מימדים כפי שפורט לעיל, אך התוצאות לא היו טובות יותר מאשר ללא הורדת המימדים, אבל יכול להיות שקיימת הורדת מימדים אחרת שתסייע או היפר פרמטרים אחרים שיתנו תוצאות טובות יותר, ודרכם נוכל למצוא את מספר הגנים הקטן ביותר שיתנו תוצאות מקסימליות. אם נגיע למספר גנים קטן מספיק, ניתן יהיה לבדוק את רמת הביטוי שלהם באמצעות בדיקת דם פשוטה, ולפי זה ניתן יהיה לחזות מהי מידת הסכנה של הפציינט לחוות התקף לב.