תרגיל 3 - Data Sconce in the Industry

1. **פצלו את קובץ הנתונים מעבודה 1 ל-train ו-test.**

פיצלנו את סט הנתונים באמצעות השיטה:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

2. **בחרו את המודל הטוב ביותר שאימנתם בעבודה 1 (כולל ה-Pre-processing) ואמנו את המודל מחדש לפי קובץ ה-train מסעיף קודם.**

ביצענו שוב את תהליך ה- Pre-Process כפי שבוצע בעבודה 1:

- הסרנו עמודות לא רלוונטיות עם ערכים קבועים
- הסרנו עמודות עם מעל 80% ערכים חסרים
- השלמת ערכים חסרים: הגדרנו עמודות קטגוריאליות ונומריות- בנומריות בדקנו אם הערכים מתפלגים נורמליות או לא והשלמנו ערכים חסרים בחציון או בממוצע בהתאמה. בעמודות קטגוריאליות החלפנו ערכים חסרים בערך הנפוץ ביותר.
- נירמול
- חלוקה ל-bins

בחרנו במודל של Random Forest שהציג לנו תוצאות טובות בעבודה 1 והוא כמובן מבוסס אנסמבל של עצי החלטה ולכן יתכתב עם מה שאנו רוצים לעשות בעבודה זו. הגדרנו:

```
random_forest = RandomForestClassifier(n_estimators=100, max_depth=20)
```

ביצענו את ההערכה גם על סט האימון וגם על סט הבדיקה.

המדדים שבחרנו לבדוק הינם:

F1 macro

Average precision

Roc AUC

וקיבלנו תוצאות ריצה של:

		F1 macro	Roc AUC	precision
1	Train model	0.9209128385067662	0.8723925410872313	0.7667968035864839
2	Test model	0.6891458198893654	0.6391316080778671	0.2771593434786628

3. <u>אמנו מודל בייסליין להשוואה: השתמשו ב-Random Forest בגודל של 100 עצים וערכי הייפר-פרמטרים דיפולטים על קובץ ה-train תוך ביצוע Pre-processing מינימלי.</u>

אימנו מודל Random Forest מסוג Baseline. ביצענו עיבוד נתונים מינימלי שכלל רק החלפת ערכים חסרים והסרת עמודות עם ערכים קבועים ועמודות עם מעל 80% ערכים חסרים.

4. <u>הציגו בצורה מסודרת את ביצועי הדיוק של שני המודלים על קבצי ה- train וה-test.</u>

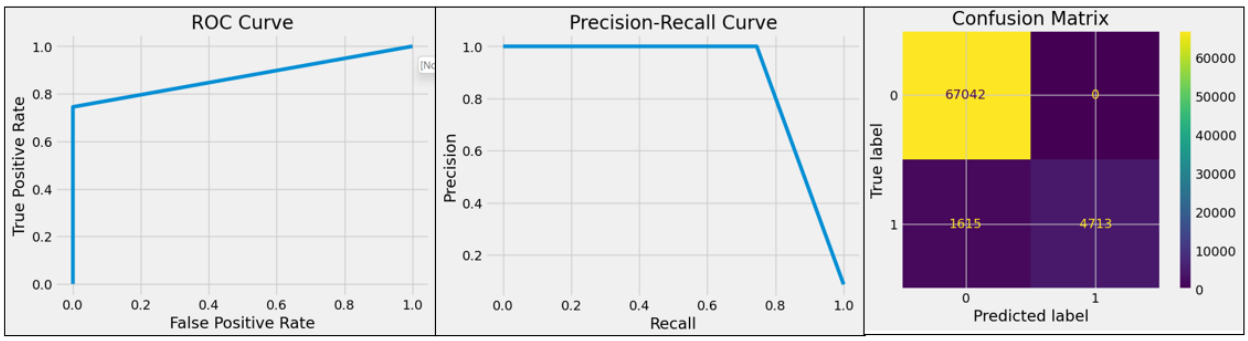
חילקנו את הבדיקות עבור סט האימון וסט הבדיקה והצגנו את המדדים שבחרנו עבור המודל הטוב ביותר בכל אחד מהם.

- סיכום התוצאות מובא בטבלה הבאה:

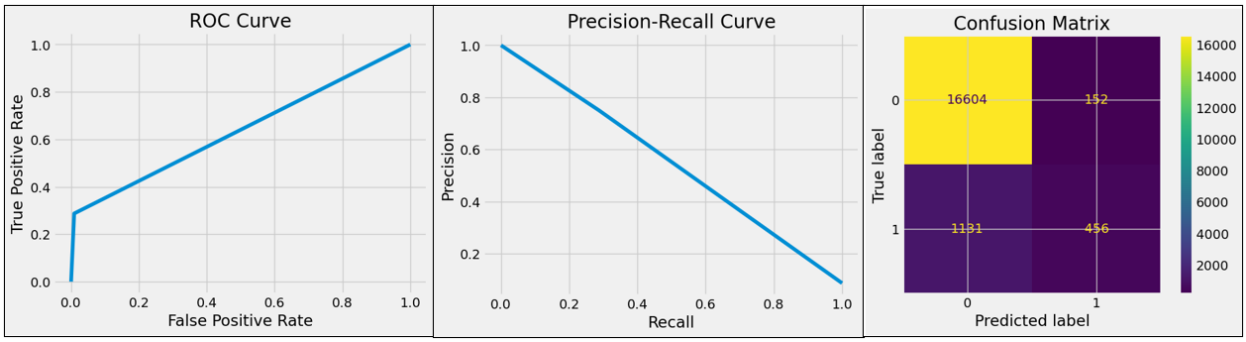
		F1 macro	Roc AUC	precision
1	Train model	0.9209128385067662	0.8723925410872313	0.7667968035864839
2	Test model	0.6891458198893654	0.6391316080778671	0.2771593434786628
3	Train baseline model	0.9998702642724786	0.9997629582806574	0.9995668052079006
4	Test baseline model	0.6926959769023562	0.6424480662665129	0.28128656740343005

<u>הצגת התוצאות:</u>

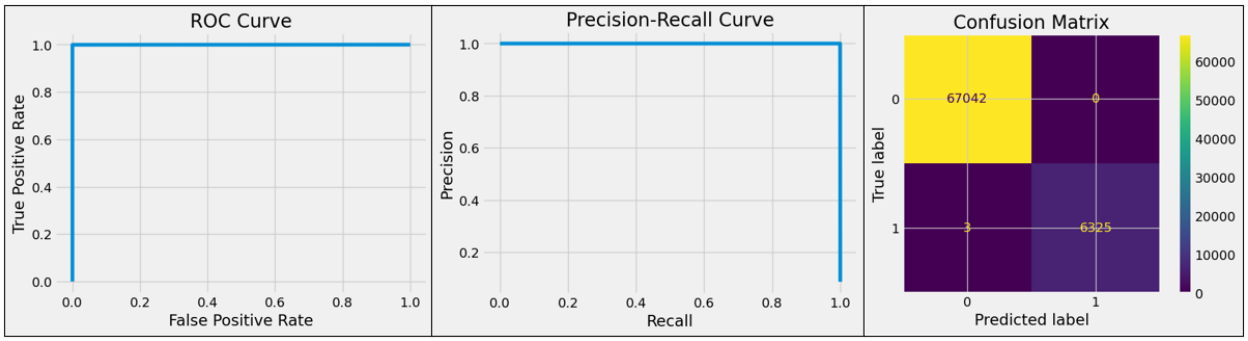
<u>1:</u>



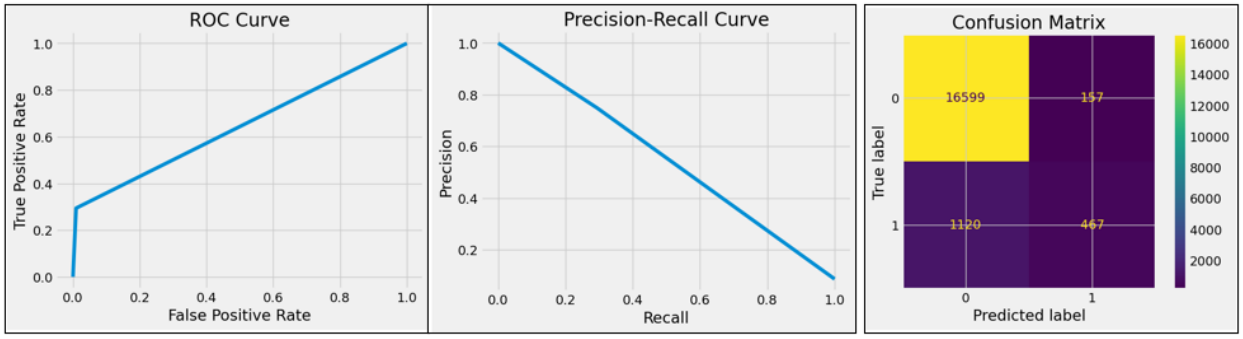
<u>-2</u>


ROC Curve
True Positive Rate
False Positive Rate
Precision-Recall Curve
Precision
Recall
Confusion Matrix
True label
Predicted label
16604
152
1131
456

-3


ROC Curve
True Positive Rate
False Positive Rate
Precision-Recall Curve
Precision
Recall
Confusion Matrix
True label
Predicted label
67042
0
3
6325

-4


ROC Curve
True Positive Rate
False Positive Rate
Precision-Recall Curve
Precision
Recall
Confusion Matrix
True label
Predicted label
16599
157
1120
467

<u>ניתוח תוצאות:</u>

- שני המודלים מצוינים במערך האימון, אך מודל ה-baseline משיג את התוצאות הכי טובות, עם דיוק גבוהה ביותר בכלל המדדים

- ניתן לראות שתוצאות ההערכה של מודלי ה-Train גבוהים באופן משמעותי יותר מתוצאות ההערכה של מודלי ה-Test. זאת מכיוון שמודלי ה-Train מותאמים בדיוק לנתוני האימון, בעוד מודלי ה-Test נבדקים על נתונים חדשים ולא ראו אותם בזמן האימון, מה שמוביל אותנו לנקודה הבאה-

- ישנה בעיה של overfitting של שני המודלים. ניתן לראות זאת מההפרש הגדול בין התוצאות שהושגו בין הסטים של האימון והבדיקה. המודל מצליח להתמודד בהצלחה עם הנתונים שכבר ראה, אך כאשר מגיעים נתונים חדשים (כמו בסט הבדיקה), הוא מתקשה לחזות באופן מדויק את התוצאה. במידה והיינו חוזרים עוד לעבודה על המודל, היינו מנסים למנוע ooverfitting על ידי שימוש בסט ולידציה נוסף, או שינוי במבנה של המודל עצמו.

- כאשר מסתכלים על מודל ה-Baseline אל מול-random forest, ניתן לראות שתוצאות ההערכה בעלות קורלציה חיובית ודומות באופן מוחלט.

- במבט כללי ואובייקטיבי, ניתן לראות ששני המודלים אינם מספיק טובים לשימוש בפרקטיקה, מכיוון שהם מציגים precision מאוד נמוך בסט הבדיקה, דבר שעשוי להוביל להרבה חיזויים שגויים.

.5 <u>הפעילו את שיטת SAHP למתן הסברים על המודל מסעיף 2.</u>

המודל הינו random forest השתמשנו ב-TreeShap .

- תוצאות לסעיף זה מוצגות בסעיף 7

.6 <u>הפעילו את שיטת SAHP למודל ה-בייסליין מסעיף 3.</u>

המודל הינו random forest השתמשנו ב-TreeShap ובהמשך גם ב-KarnelShap

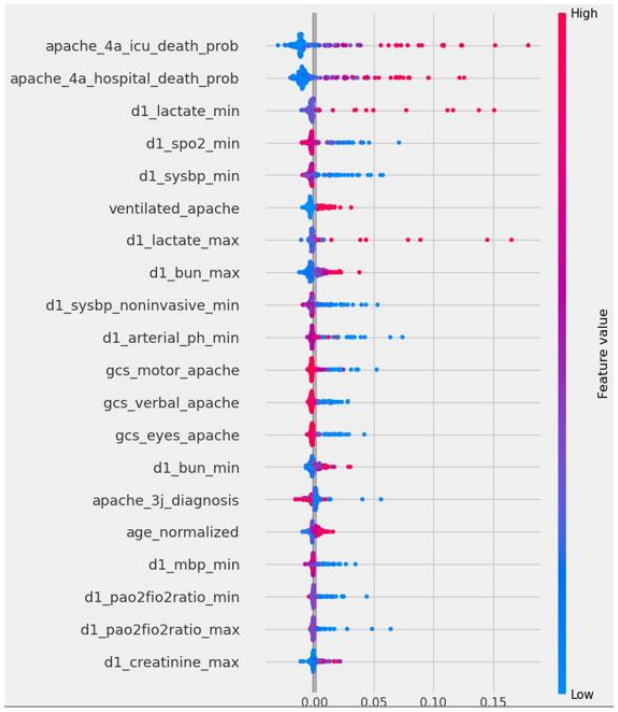
- תוצאות לסעיף זה מוצגות בסעיך 7

7. <u>הציגו את תוצאות ההסברים שיצרתם – עבור כל אחד משלושת ההפעלות של SHAP (פעם אחת עבור המודל מסעיף 2, ופעמיים עבור המודל מסעיף 3). בסעיף זה יש להתמקד בתרשימי ה- Global Feature Importance, ותרשימי ה-Summary Plot</u>

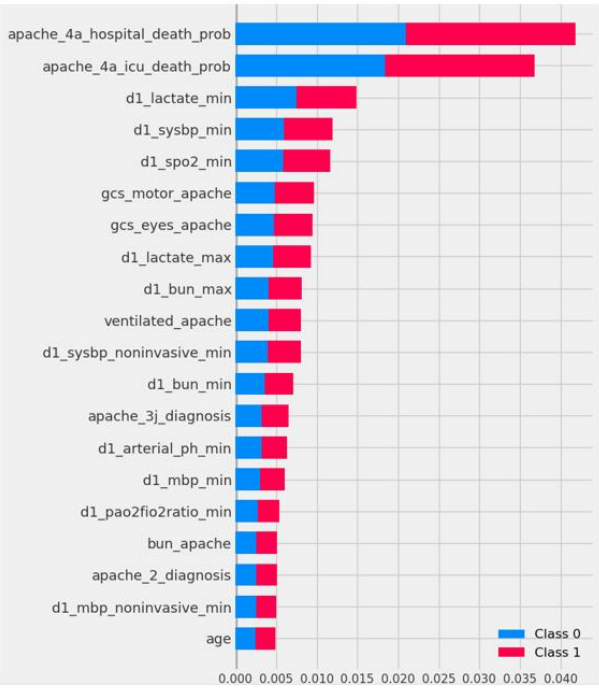
<u>תוצאות סעיף 5:</u>

TreeShap - Train

<u>Summery plot</u>



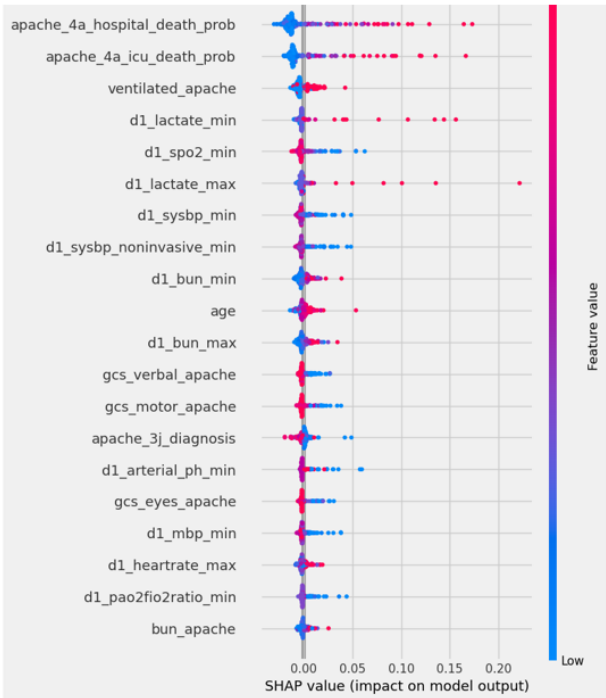
<u>Global Feature Importance</u>



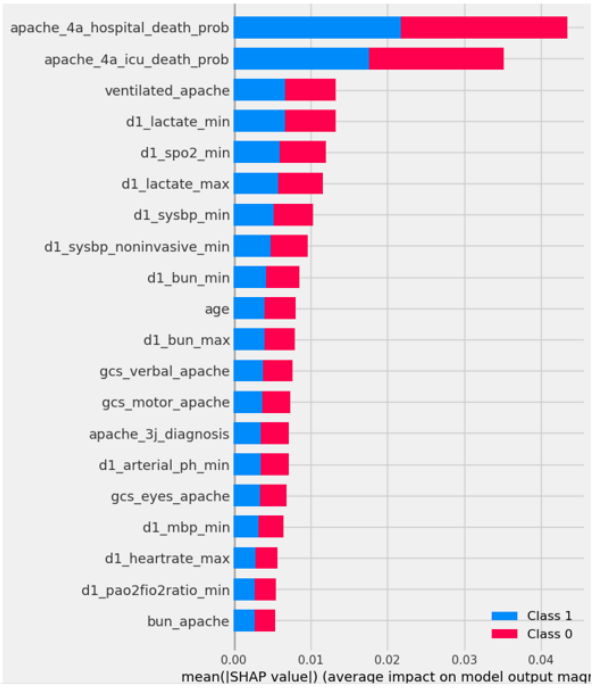
תוצאות סעיף 6:

TreeShap - Baseline Train

Summery plot

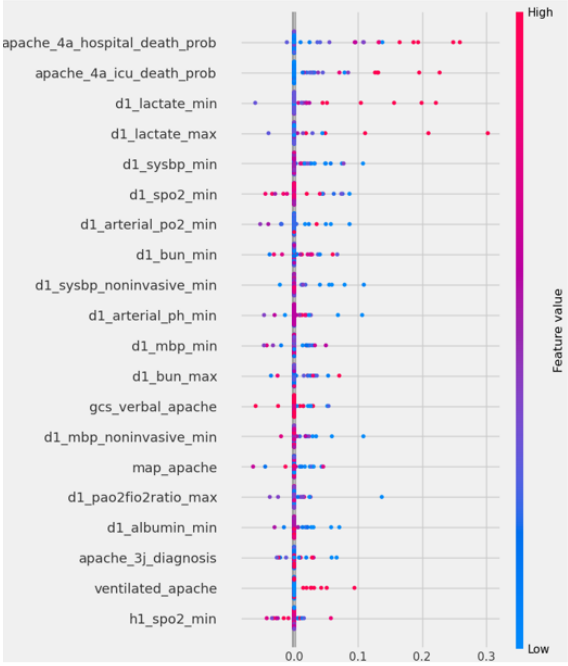


Global Feature Importance

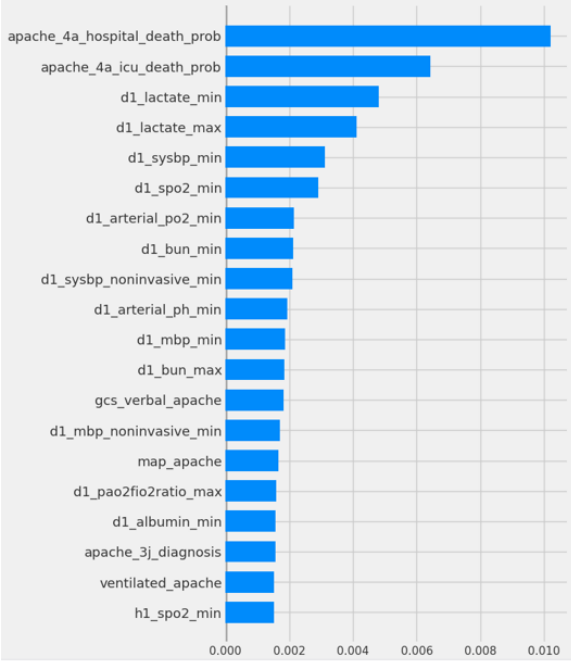


BaseLine Train – KarnelShap

Summery plot



Global Feature Importance



ניתוח התוצאות:

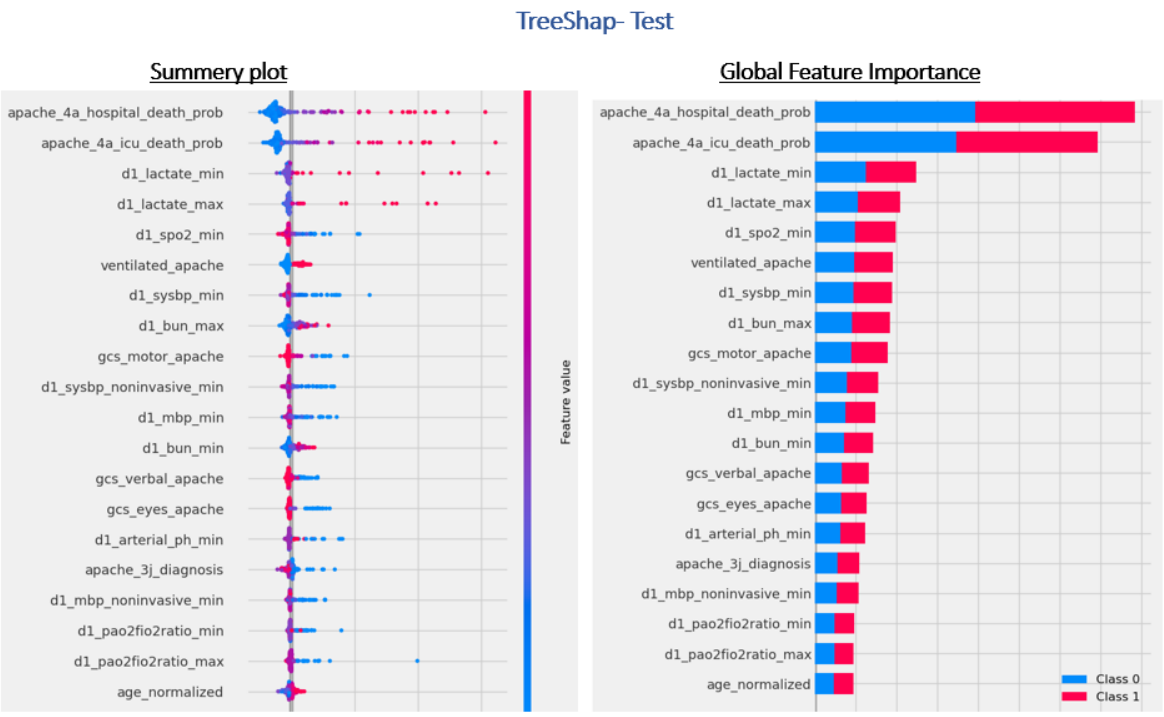
Summery Plots:

- המאפיינים החשובים במודלים שונים עשויים להשתנות. זאת אומרת, מודלים שונים מתמקדים במאפיינים שונים כדי לבצע חיזויים.
- יש קשר ברור בין ערך המאפיינים שנמצאו כמשמעותיים ביותר לבין ה-target (ערך 1 המייצג מוות). זאת אומרת, ככל שהערך במאפיינים הללו גדול, כך גם הסיכוי למוות מרבה.
- אמנם סט הפיטצ'רים המשמעותיים ביותר אינו זהה בדיוק בין המודלים, והחשיבות של כל אחד מהם משתנה בין כל מודל ומודל- אך ישנה הסכמה שסט הפיטצ'רים המשמעותיים ביותר הינם:
apache_4a_hospital_death_prob, apache_4a_icu_death_prob, d1_lactate_min, d1_lactate_max, d1_sysbp_min, d1_spo2_min, ventilated_apache
בנוסף באופן גורף וחד משמעי שני הפיטצ'רים:
apache_4a_hospital_death_prob, apache_4a_icu_death_prob, d1_lactate_min
התקבלו כמשמעותיים ביותר בכל המודלים.

Global Feature Importance:

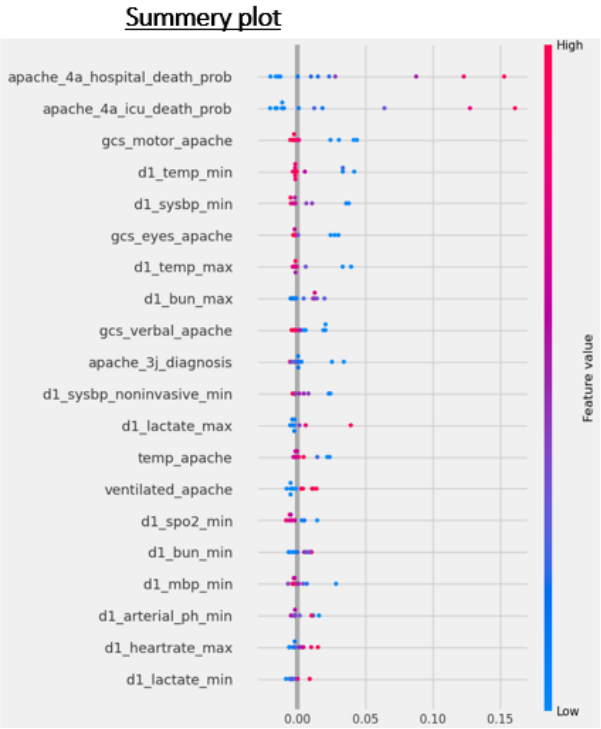
- קיים פער משמעותי בין הפיטצ'ר החשוב ביותר לבין אלו שבאים אחריו. כלומר חשיבותו גבוהה באופן מובהק לעומת השאר.
- באופן זהה ל-summery plot- סט הפיטצ'רים המשמעותיים ביותר אינו זהה בדיוק בין המודלים, והחשיבות של כל אחד מהם משתנה בין כל מודל ומודל- אך ישנה הסכמה שסט הפיטצ'רים המשמעותיים ביותר הינם:
apache_4a_hospital_death_prob, apache_4a_icu_death_prob, d1_lactate_min, d1_lactate_max, d1_sysbp_min, d1_spo2_min, ventilated_apache
בנוסף באופן גורף וחד משמעי שני הפיטצ'רים:
apache_4a_hospital_death_prob, apache_4a_icu_death_prob, d1_lactate_min
התקבלו כמשמעותיים ביותר בכל המודלים.

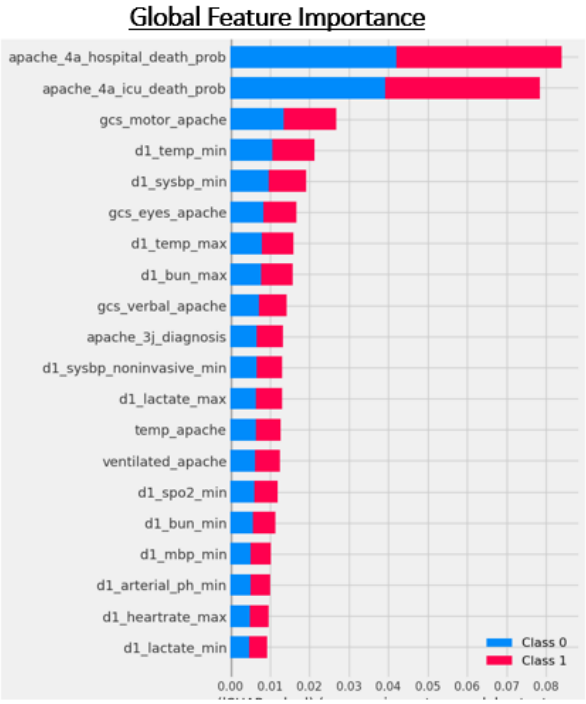
8. <u>תתמקדו בניתוח הסברים שניתנים לדוגמאות חדשות כאשר המודל כבר בשימוש בפועל. הפעילו את שיטת SHAP על המודל מסעיף 2, אבל הפעם על קובץ ה-test. הציגו את תרשימי Global Feature Importance, ו-Summary Plot.</u>



- ניתן לראות כי התוצאות כמעט עקביות אל מול ההפעלה על קובץ ה- train מסעיף 2: בשתי הריצות קיבלנו מאפיינים כמעט זהים, אם כי קיימים הבדלים דקים, מבחינת גודל החשיבות שלהם. 3 המאפיינים הראשונים: apache_4a_hospital_death_prob, apache_4a_icu_death_prob, d1_lactate_min זהים בשני הריצות. אך למשל- d1_mbp_noninvasive_min , age_normelized בעלי הבדל גדול בין התוצאה שקיבלו ב- train לזו שקיבלו כעת, שניהם ירדו בחשיבותם בעוד ש-d1_lactate_max ו-gcs_motor_apache עלו בחשיבותם.

9. <u>חיזרו על הניתוח מסעיף קודם, אבל הפעם יש לסנן את קובץ ה-test, שיכיל רק רשומות שהסיווג שלהם היה לא נכון ע"י המודל.</u>



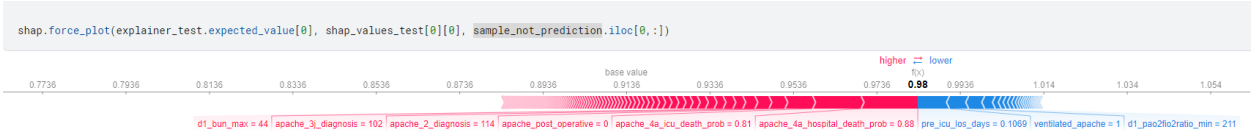


- מסקנות אל מול הריצה בסעיף 8:
  - שני הפיטצ'רים המשמעותיים נשארו כפי שהיו בכל שאר המודלים. כמו כן, השפעתם מובהקת באופן בולט אל מול השאר גם במקרה זה.
  - מדד ה-gcs_motor_apache מראים חשיבות רבה גם באופן פרטי וגם באופן יחסי לשאר הריצות בהם לא הוצגו בעלי חשיבות כה גבוהה. אם כי גם בסעיף 8 שיפר את חשיבותו אל מול התוצאות הקודמות.
  - ישנו פיזור יותר אחיד של התוצאות באופן גורף בכל המאפיינים

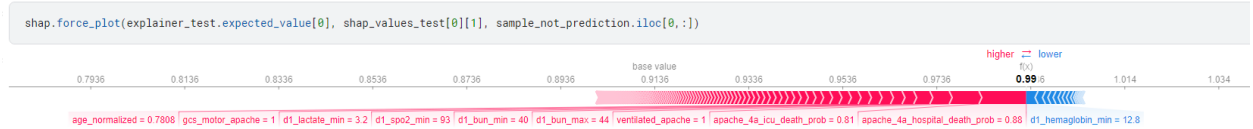
10. <u>תתמקדו ביכולות של SHAP לספק הסברים ברמת הרשומה הבודדת (שקף 58). בחרו מקובץ ה-test, 2 דוגמאות שסווגו נכון ו-2 דוגמאות שסוגו לא נכון ע"י המודל הראשי שלכם.</u>

2 דוגמאות שסווגו נכון:

.1



.2

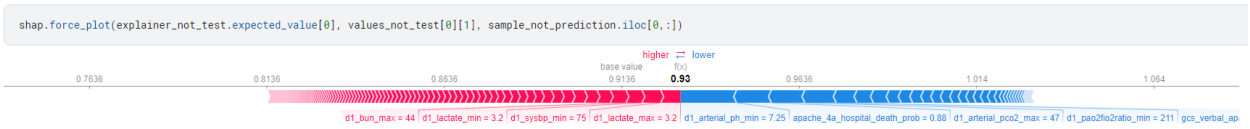


- דוגמאות מעניינות ברמת הרשומה:

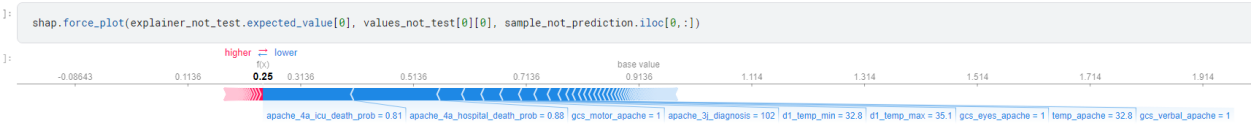
  ניתן לראות את התוצאות שהקודמות באות בתרשים זה לידי ביטוי- המאפיינים apache_4a_hospital_death_prob, apache_4a_icu_death_prob בעלי השפעה מרובה, גם מבחינתם היותם המשפיעים ביותר וגם מבחנות ה-"נפח" של ההשפיה שהם מייצרים. כמו כן, כמות המאפיינים בעלי ההשפעה לחיזוי 0 הוא נמוך באופן ניכר, יחד עם זאת, ישנם לא מעט מאפיינים כאלו.

2 דוגמאות שסווגו לא נכון:

.1



.2



- דוגמאות מעניינות ברמת הרשומה:

  בדוגמא הראשונה ניתן לראות כי יש מאפיינים שמתנהגים באופן שונה מזה שאנחנו מצפים לו, לדוגמא- apache_4a_hospital_death_prod בעל חיזוי 0 בשונה למה שאנחנו מצפים.

  בדוגמא השניה ניתן לראות רוב של מאפיינים שחוזים פרדיקציה 0 בניגוד למה שאנחנו מצפים. אחוז המאפיינים הללו גדול באופן יחסי גם אל מול התוצאות שקיבלנו עד כה.