בס״ד

מעבדה 1 - חיזוי אלח דם - Spesis prediction

גיא חדד 313465890
נעם שמיר 316299098

1. תקציר

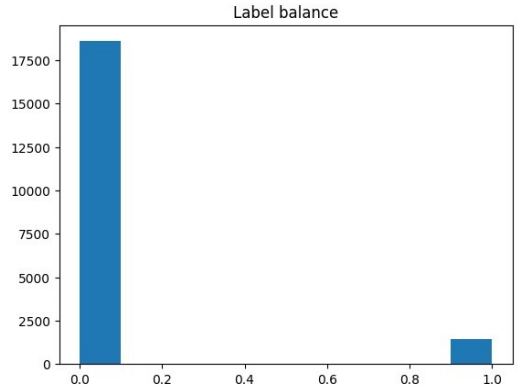
- תהליך: טענו את הדאטה מתוך הקבצים שקיבלנו. עיבדנו את המידע בשלב ראשוני כך שמחקנו את כל העמודות שהופיעו 6 שעות טרם אלח הדם (עבור מטופלים שאובחנו כבעלי אלח דם). כך למעשה השארנו רק את המידע שזמין ועל בסיסו נדרש להסיק במשימת החיזוי.
- עבור כל מטופל, כיווצנו את המידע לשורה אחת שהכילה נתונים אגרגטיביים שמייצגים את הנתונים שלו. המידע האגרגטיבי כלל: משך הזמן הכולל (או עד לאבחון אלח הדם פחות 6 שעות), לייבל: אובחן עם אלח דם / לא, ובנוסף מדדים סטטיסטיים עבור כלל המדדים הנומריים: מינימום, מקסימום, הערך הראשון, אחרון, ממוצע וסטיית התקן.
- לאחר שהדאטה סודר ברמה טבלאית כך שכל שורה מייצגת מטופל - הפעלנו 4 מודלים (XGBoost, LightGBM, Logistic Regression, CatBoost) ובחרנו את ה3 הטובים מביניהם. ביצענו משימת פרדיקציה בינארית - לחיזוי אלח דם.
  השתמשנו בטכניקת Resampling להתמודדות על חוסר האיזון בדאטה בין המחלקות.
  איפטמנו את הפרמטרים של המודל ע״י grid search ובעזרת optuna.
  f1 score המקסימלי שקיבלנו במודל CatBoost הוא 0.747.
- את ה df כיווצנו לחסכון בזיכרון ושיפור היעילות מה שהקטין את הנפח של הדאטה בכ70%.
- בחנו feature importance ע״י מודל SHAP.

2. אקפסלורציה ואנליזה של הדאטה

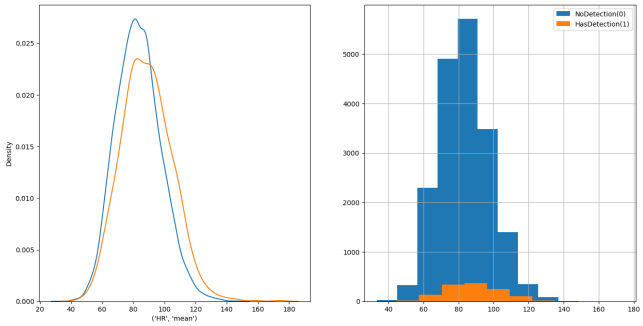
הפיצ׳רים שזמינים בדאטהסט הם: לכל מטופל שהגיע לחדר מיון יש עבור כל שעה עשרות מדדים נומריים שמתארים נתונים פיזיים שנמדדו. המידע מכיל ערכי null רבים שמייצגים נתון שלא נמדד. בנוסף החל מ6 שעות טרם אבחון אלח דם (אם היה) מסומן לייבל 1 ואחרת 0. בין הלייבלים המרכזיים יש מדדים על דופק, טמפרטורה וקצב נשימות שהם מאפיינים משמעותיים באבחון של אלח דם.
בחרנו את **כל** הפיצ׳רים שמתארים מצב גופני כי כולם היו רלוונטיים עבורנו וכי המודלים מספיק חזקים בשביל להכיל את כל הדאטה באופן יעיל. בבדיקות שבהם ניסינו לבחור תת קבוצה של פיצ׳רים (למשל, בחירת 30 הפיצ׳רים החשובים ביותר) קיבלנו תוצאות פחותות ולכן התמדנו בגישה של שימוש בכלל הפיצ׳רים הזמינים. עבור כל מטופל, כיווצנו את המידע לשורה אחת ובה מינימום, מקסימום, טווח, הערך הראשון, אחרון, ממוצע וסטיית התקן של כל פיצ׳ר נומרי. לכל מטופל הוספנו את הלייבל המתאים לפי הלוגיקה שניתנה בהנחיות. את הפיצ׳ר של מגדר פירקנו ל2 עמודות כ - dummy variable.
פיצ׳רים שאין בהם שינוי (למשל גיל המטופל, unit וכדו׳) המכונים בהוראות demographic values - נותרו כפי שהם.

במחברת הצגנו תחת פרק EDA מדדים כללים לכל עמודה (describe), ויצרנו גרפים של היסטוגרמות וטבלאות לפיצ׳רים וטבלת heatmap לקורולציה. ניתחנו qqplot עבור משתנים רלוונטים. נתאר כאן מבחר של דוגמאות עבור פיצ׳רים מעניינים ומרכזיים:

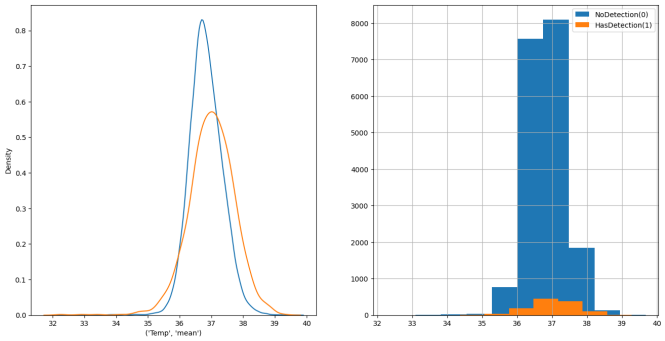
- הדאטה אינו מאוזן. היחס הוא 7% אלח דם לעומת 93% ללא.



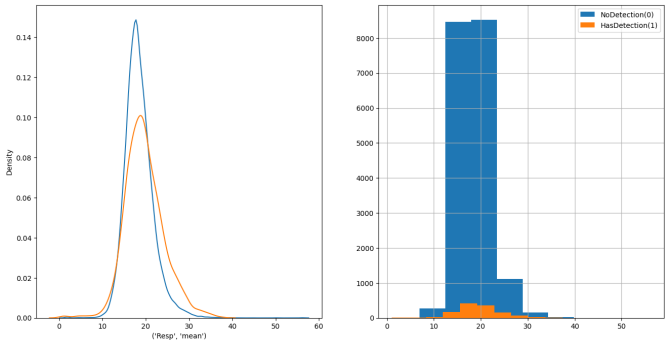
- ניתן לראות למשל שערכי ממוצע הדופק, הטמפרטורה וקצב הנשימה גבוהים יותר בקרב בעלי אלח דם. להלן היסטוגרמה, פונקציית הצפיפות בכל מחלקה, וערכי מבחן ההשערות עבור בדיקות pearson correlation, spearman correlation להשוואת התפלגות ממוצעי הפי״צרים. מדדי הקורולציה מציגים את הקשר בין הנתון לבין אלח דם, כאשר פירסון מציג את הקשר הלינארי, בעוד שספירמן מציג את הקשר המונוטוני (אם קיים). בכל המדדים שנציג להלן, הקשר קיים אך חלש מאוד. המובהקות של הקשר נתונה ע״י ערכי p-value קטנים מאוד (ב3 הדוגמאות הבאות הערך הוא 10^-24 לכל היותר, כלומר מובהק).
  - דופק (כתום - אלח דם, כחול - ללא)
    - Pearson correlation: 0.08 (p-value < 1%)
    - Spearman correlation: 0.07 (p-value < 1%)



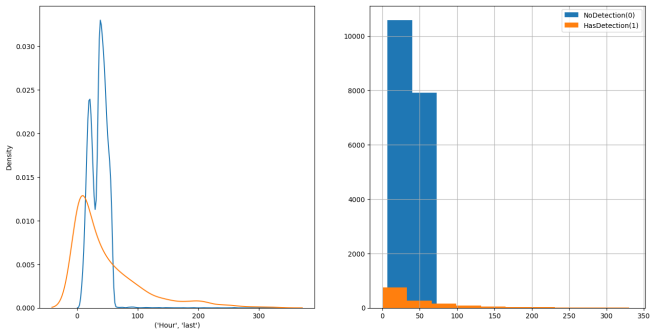
  - טמפרטורה (כתום - אלח דם, כחול - ללא)
    - Pearson correlation: 0.07 (p-value < 1%)
    - Spearman correlation: 0.07 (p-value < 1%)



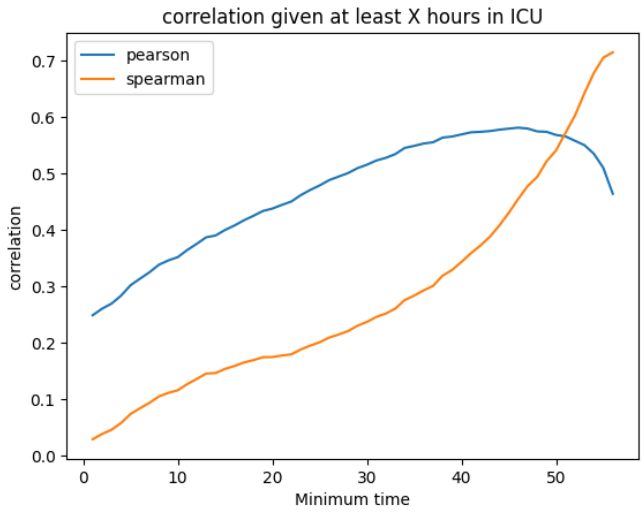
- קצב נשימה (כתום - אלח דם, כחול - ללא)
- Pearson correlation: 0.08 (p-value < 1%)
- Spearman correlation: 0.07 (p-value < 1%)



- ניתן לראות שמטופלים שהיו זמן רב בחדר מיון נטו יותר להיות מאובחנים כבעלי אלח דם.



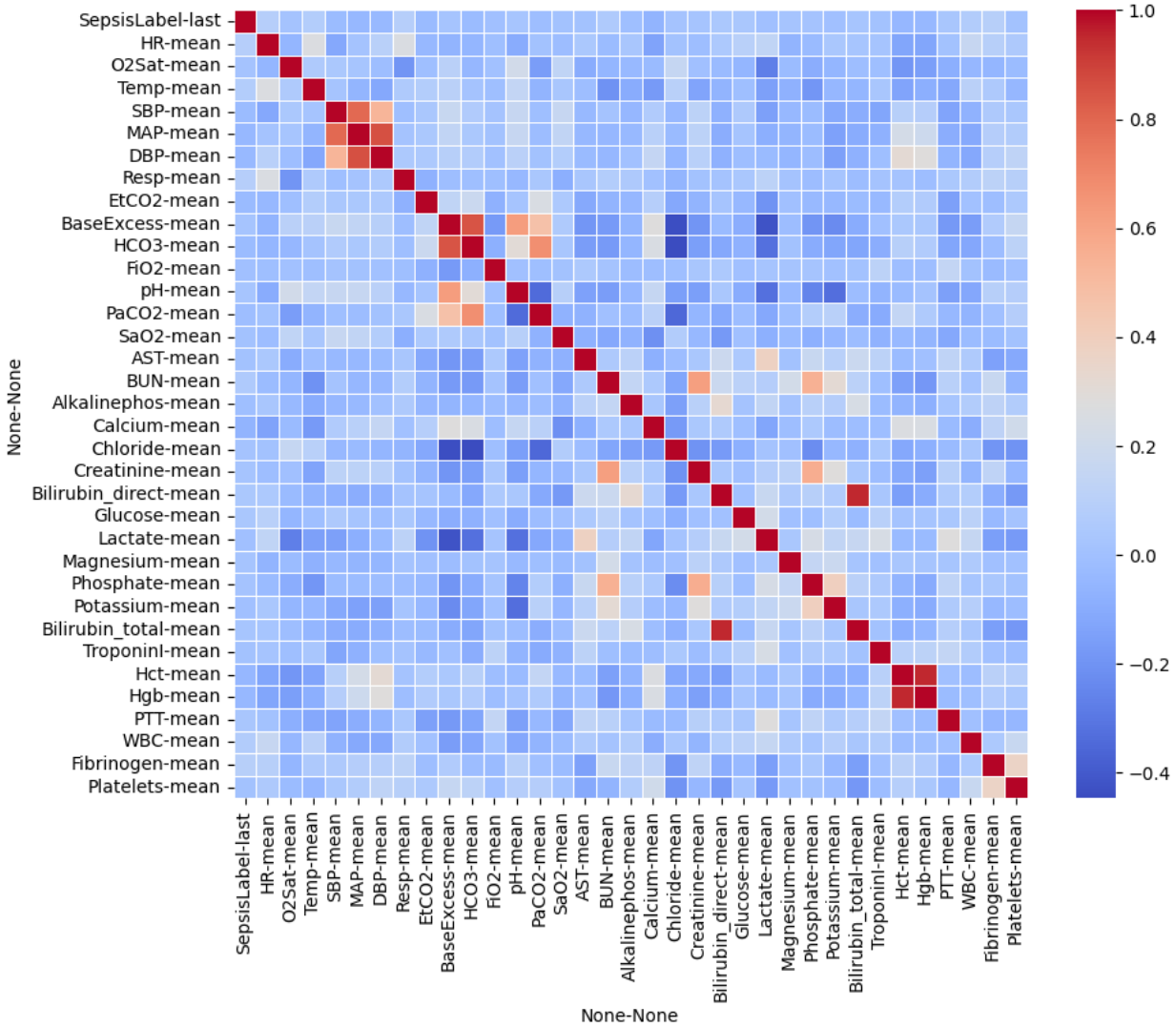
בהמשך לגרף המוצג לעיל - נראה שעבור האוכלוסיה שהייתה זמן רב בחדר ניתוח - סיכויי האבחון לאלח דם גדלים. במילים אחרות - בהינתן שהזמן בחדר ניתוח גדל, סיכויי אלח הדם עלו.
רצינו לבחון את הממצא הזה, ולכן בחנו עבור ערכים שונים של זמנים בחדר ניתוח - מה מידת הקורולציה עם אלח דם.
להלן גרף שמציג את ערכי pearson correlation ו- spearman correlation כתלות בזמן המינימלי בחדר טיפול נמרץ. עבור האוכלוסיה שנשארה לפחות x שעות - הקורולציה של תת קבוצה זו להיות מאובחנת עם אלח דם - גדלה.



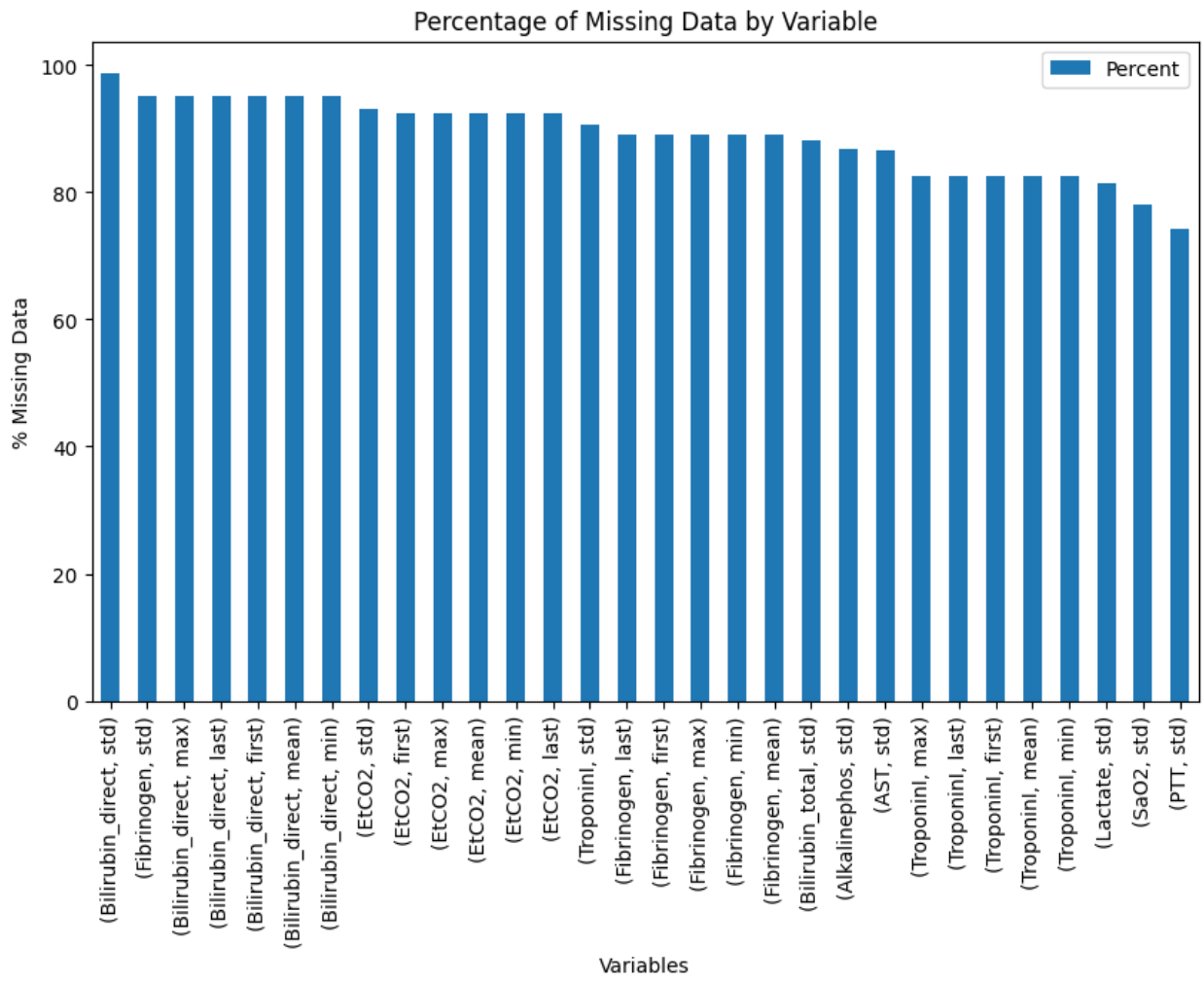
הגרף מוצג עד 57 שעות ולא יותר, שכן החל מנקודת זמן זו ערכי p-value נעשים לראשונה גדולים מ0.05, כלומר **כל** הקורולציות המוצגות בגרף הן ברמת מובהקות של 5%. כמובן שככל שהזמן המינימלי גדל (ימינה בציר הx)- גודל האוכלוסיה קטן ולכן המובהקות נפגעת.

הערה בעניין bias במדידה זו: המידע הנתון עבור בעלי אלח דם מרגע הכניסה ועד לסיום בחדר טיפול נמרץ קוצץ ב6 שעות. לכאורה - עובדה זה מייצרת הטייה ולא מאפשרת את ההשוואה לעיל, אלא שהעובדה הזו דווקא מחזקת את הטענה באופן מאוד משמעותי: זמן השהיה בחדר טיפול נמרץ של בעלי אלח דם ארוך יותר, <u>אפילו</u> אם מקצצים 6 שעות.

ב heatmap בחרנו בממוצע של התכונות על פני השעות שבהן המטופל היה בחדר טיפול נמרץ. ניתן לראות שאף תכונה כשעומדת לבדה אינה קורולטיבית במיוחד עם Spesis (בשורה העליונה).



השלמת נתונים חסרים: ראשית - עבור כל מטופל עשינו אגרגציה של ערכים כמו max,min ועוד - שהן פונקציות שמסתדרות מצוין כאשר רק חלק מהנתונים חסרים. הן מתייחסות לנתונים הקיימים בלבד.
לאחר האגרגציה - ייתכן שישארו ערכי null. עבור xgboost ו lightGBM עשינו נסיון של הסרת פיצ׳רים בעלי 25% null או יותר וראינו שלא הושג שיפור. ניסינו גם להפעיל imputation מסוג KNN עם 3 שכנים ללא שיפור.
בחנו גם השלמה עם ממוצע וחעם חציון, אך לא הושג שיפור. מסיבה זו - בחרנו שלא להשלים את הנתונים החסרים מתוך ידיעה שהמודלים שבחרנו מסתדרים היטב עם ערכי null.
בנוסף - מנגנון החסרות הוא לא missing at random אלא missing not at random. העובדה שמידע מסוים קיים או חסר מלמדת המון. אם הרופא בחר למדוד או לא למדוד ערך של תכונה כלשהי - עובדה זו כשלעצמה מלמדת משהו על מצב המטופל ועל הדעה של הרופא על מצבו.
כמות ערכי הnull היא גדולה מאוד, להלן דוגמה של top 30 פיצ׳רים בעלי כמות מירבית של nulls:



3. הנדסת פיצ׳רים Feature Engeering

בחרנו את כל הפיצ׳רים שקיימים. את המגדר (זכר/נקבה) הפכנו ל2 עמודות, כdummy variable מכיוון שזהו משתנה קטגורי. ליתר הפיצ׳רים הנומריים חישבנו ערכי Min, Max, Std, First, Last, Mean עבור כל מטופל.
הערכנו שערך התחלתי (בהגעה לטיפול נמרץ), הערך האחרון שנמדד וזמין (last), ממוצע, ערך מינימלי ומקסימלי - כל אלו יחד יכולים לתפוס את מירב המידע מתוך הטבלה.
כטרנספורמציה בחרנו לחשב את הטווח של המדד הפיזי: max-min.
באופן כללי למודלים מבוססי עצים פחות דואגים מרגישות לסקאלה ולכן לא הכרחי לנרמל את הערכים.

4. חיזוי - מודלים

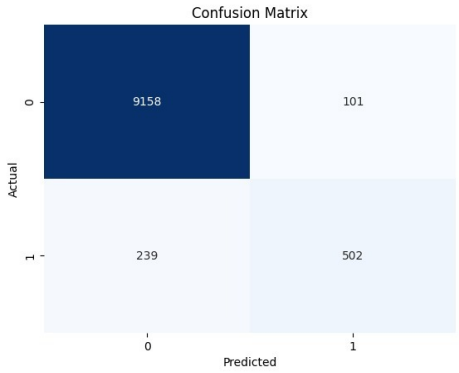
Xgboost

בחרנו במודל זה כיוון שהוא נחשב טוב מאוד לדאטה טבלאי.
הסבר על אופן פעולת המודל:
Extreme gradient boosting algorithm,
מודל ensemble שמשלב בוסטינג על עצי החלטה.
מאוד יעיל וגמיש.

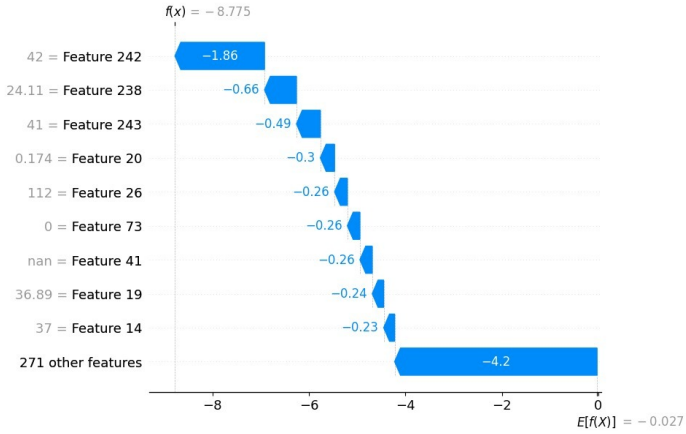
בחירת היפרפרמטרים: בחרנו בעזרת grid search את הפרמטרים:
max_depth=15, n_estimator=100, learning_rate=0.3,
objective= Binary Logistic
הגבלת עומק העץ משמשת כרגולריזציה.
ביצועים על train-set: המודל משיג fit על הtrain, כלומר accuracy=1
ביצועים על test-set: מדד f1 השיג 0.73

	precision	recall	f1-score	support
0	0.97	0.99	0.98	9259
1	0.87	0.64	0.74	741
accuracy			0.97	10000
macro avg	0.92	0.82	0.86	10000
weighted avg	0.96	0.97	0.96	10000

אנליזה:
מטריצת בלבול: ניתן לראות שהמודל לעיתים מפספס אלח דם (טעות מסוג שני), אך לעיתים רחוקות יותר (תוך התחשבות גם בחוסר האיזון במחלקות) המודל מחזיר false positive (טעות מסוג ראשון).



בחינת הפיצ׳רים החשובים על פי מודל shap:



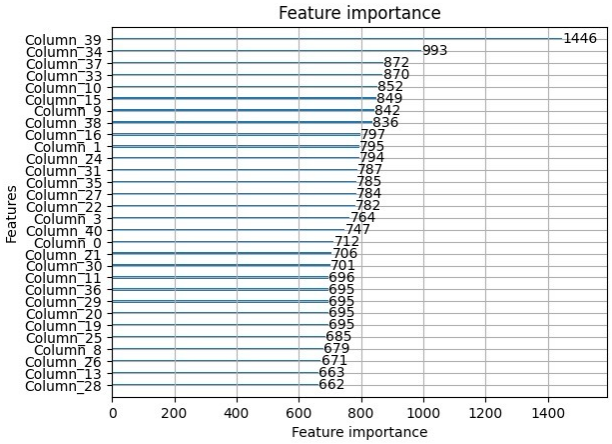
מודל shap לקוח מתורת המשחקים ומעריך את התרומה של כל פיצ׳ר. בגרף מוצגות התכוניות החשובות ביותר על פי shap בתצוגת waterfall. בשורה הראשונה, כלומר הפיצ׳ר החשוב ביותר הוא ICULOS כלומר ICU Length Of Stay. כפי שזיהינו באנליזה את חשיבותו - על פי shap זהו הפרמטר החשוב ביותר.
לאחר מכן בחשיבות לפי הסדר הבא:
גיל המטופל

סטיית תקן של טמפרטורה
Systolic Blood Pressure
ריכוז חמצן Fio2 - fraction of inspired oxygen
סטיית תקן של DBP.
הטמפרטורה האחרונה שנמדדה.
טמפרטורה מקסימלית.

LightGBM

בחרנו במודל זה כיוון שהוא נחשב טוב מאוד לדאטה טבלאי.
הסבר על אופן פעולת המודל:
gradient boosting algorithm,
מודל ensemble שמשלב בוסטינג על עצי החלטה. בשונה מXGBoost משתמש בהיסטוגרמה של הערכים ולא בתבנית קבועה כדי להחליט על אופן הsplit. בנוסף משתמש בחישוב מקבילי - מה שמשפר מאוד את זמן הריצה שלו.
בחירת היפרפרמטרים: בחרנו בעזרת optuna ובנוסף תיקונים שלנו את הפרמטרים:
'num_leaves': 150,
'learning_rate': 0.02,
'feature_fraction': 0.9378537303583372,
'bagging_fraction': 0.9990828605870183,
'bagging_freq': 2,
'min_child_samples': 31,
'num_iterations': 400}
ביצועים על train-set: המודל משיג fit על הtrain, כלומר accuracy=1
ביצועים על test-set: מדד f1 השיג 0.72

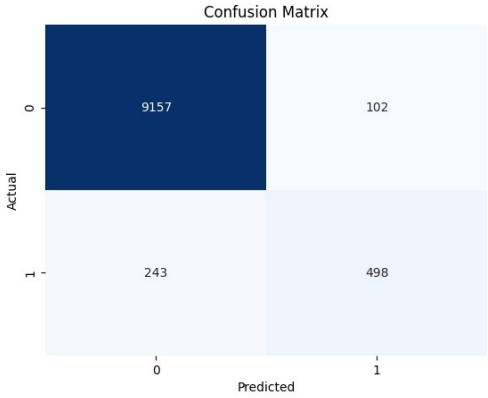
במודל זה קיבלנו feature importance שניתן ע״י החבילה של lightGBM. ניתן לראות שיש פיצ׳רים שונים מאוד מאלו שהתקבלו ב xgboost.



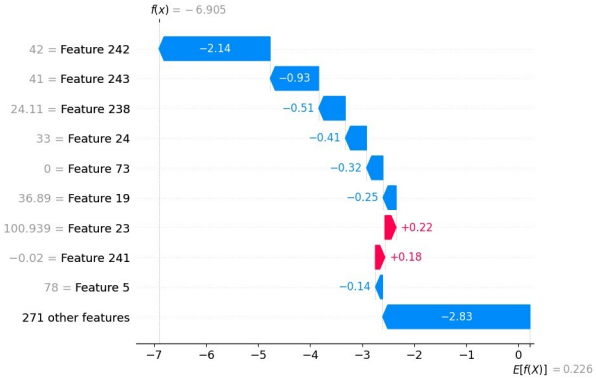
עבור המודל השלישי - עשינו נסיון של מודל בסיסי - logistic regression. בחרנו במודל זה כדי שנוכל להסיק מהם הפיצ׳רים שתרמו יותר להחלטה ולהסיק מסקנות של interpretabiility.
השגנו f1-score של כ30% שהוא מתחת לסף הדרוש ולכן חזרנו למודלים החזקים מבוססי boosting על desiction trees.
מודל שלישי:

CatBoost

בחרנו במודל זה כיוון שהוא נחשב טוב מאוד לדאטה טבלאי.
הסבר על אופן פעולת המודל:
gradient boosting algorithm,
מודל ensemble שמשלב בוסטינג על עצי החלטה ויודע לקבל כקלט גם משתנים קטגוריים.
בחירת היפרפרמטרים: השתמשנו בהיפרפרמטרים של ברירת המחדל שהשיגו תוצאות טובות מאוד
ביצועים על train-set: המודל משיג fit על הtrain, כלומר accuracy=1
ביצועים על test-set: מדד f1 השיג 0.74
אנליזה:
מטריצת בלבול: ההתפלגות נראית דומה למטריצת הבלבול של xgboost.



ניתוח feature importance ע״י SHAP:



גם כאן ניתן לראות שקיים דמיון בפיצ׳רים הגדולים אך גם שוני לעומת XGBoost.

5. סיכום

התוצאות הטובות ביותר הושגו ע״י CatBoost, עם f1-score של 0.74.
חוסר האיזון בדאטה דרש התייחסות מיוחדת כדי להתמודד איתו, כמו גם המבנה של כמות רשומות רבה ומשתנה לכל מטופל. אנו מעריכים שהצלחנו באגרגציה שכללה כיווץ המידע מציר הזמן לרשומה אחת עבור כל מטופל.
הפונקציות האגרגטיביות השונות (ממוצע, ראשון, אחרון, מינימום, מקסימום, סטיית תקן וטווח) הצליחו לתפוס את המאפיינים של הפיצ׳רים.
באנליזת ההתפלגות, במבחן ההשערות לבדיקת קורולציה ובמודל shap - כולם הראו את חשיבות הזמן בחדר טיפול נמרץ כאינדיקציה לאלח דם. נראה שאלח דם זו מחלה שמאופיינת בכך שניתן (לא תמיד) להיות בטיפול נמרץ במשך שעות רבות. זאת בשונה ממחלות אחרות שדורשות טיפול נמרץ שבהן משך השהייה קצר יותר באופן ניכר.