המכללה האקדמית של עמק יזרעאל
החוג למערכות מידע

DATA MODELING REPORT

פרויקט גמר

חיזוי רמת חומרת פציעה בתאונות דרכים

שם הסטודנטיות:
מאי יוסף 318608072
אלמוג יבדייב 305325417

שם המנחה:
יואב זיו

תוכן עניינים

1. בחירת טכניקות מידול

1.1. בחירת טכניקות מידול נכונות

בשלב המידול, לאחר שביצענו ניקוי, טיוב ועיבוד של הנתונים, הגענו לשלב בו נדרש לבחור את המודלים המתאימים ביותר למשימת הסיווג. בפרויקט שלנו אנו מתמודדות עם אתגר של חיזוי חומרת הפציעה בתאונות דרכים, כאשר המודל נדרש להבחין בין חמש קטגוריות שונות של רמות פציעה: No Apparent Injury : ללא פציעה נראית לעין, Possible Injury : פציעה אפשרית, Suspected Minor Injury : חשד לפציעה קלה, Suspected Serious Injury : חשד לפציעה חמורה, Fatal Injury : פציעה קטלנית.

מדובר במשימת סיווג רב-קטגורי, על בסיס עשרות תכונות כמותיות וקטגוריאליות, רבות מהן כוללות ערכים חסרים. עמודת המטרה אינה מאוזנת: מרבית המקרים מתארים תאונות ללא פציעה, בעוד שמקרי המוות נדירים ומהווים שיעור מזערי מכלל הרשומות. בנוסף, נפח הדאטה גדול מאוד (כ־188,277 רשומות), מה שמאפשר שימוש במודלים מתקדמים אך מחייב עיבוד יעיל, יכולת התמודדות עם מורכבות ושונות גבוהה בנתונים.

בהתאם למאפיינים אלה, נבחנו מספר טכניקות מידול שונות, תוך התאמה לאתגרים כמו חוסר איזון, ריבוי משתנים קטגוריאליים, ערכים חסרים ודרישות ביצועיות. נבחנו גם שיקולים של פרשנות המודלים, זמן חישוב והתאמה למשתנים מסוגים שונים. כל מודל שנבחר נבחן הן לפי מדדים כמותיים (כגון Precision, Recall ו־F1) והן לפי עמידות לרעש, יכולת פרשנות והתאמה לדאטה בפועל.

בחרנו חמישה מודלים עיקריים, המייצגים שיטות מגוונות מעולם הסיווג: החל ממודלים סטטיסטיים פשוטים, דרך אלגוריתמים מבוססי עצים, ועד לשיטות Boosting מתקדמות:

Logistic Regression- מודל סטטיסטי בסיסי ומוכר, אשר מעריך הסתברויות באמצעות פונקציית לוגיט. הוא מתאים במיוחד כאשר קיימת הנחה (או קירוב) של קשר לינארי בין הפיצ'רים ללוג-הסתברות של התוצאה. רגרסיה לוגיסטית נבחרה לשמש כקו בסיס (baseline), שכן יתרונותיה המרכזיים כוללים פשטות, מהירות חישוב ויכולת להסביר במדויק את תרומתה של כל תכונה לתוצאת החיזוי. עם זאת, למודל מגבלות משמעותיות: הוא מתקשה לייצג קשרים מורכבים ולא ליניאריים, ורגיש במיוחד להבדלים בקני מידה בין משתנים - מה שחייב אותנו לבצע סטנדרטיזציה מלאה. כמו כן, הוא דורש קידוד נומרי לכל משתנה קטגוריאלי.

Random Forest- מודל Ensemble מבוסס עצי החלטה, הפועל על עיקרון של בניית אוסף עצים (Forest) המאומנים על תתי-מדגמים שונים של הדאטה ושל הפיצ'רים. כל עץ "מצביע" על תוצאה, והתחזית הסופית מתקבלת לפי רוב. המודל מצטיין בהתמודדות עם רעש, משתנים מיותרים, ודאטה לא ליניארי, והוא נחשב לגמיש, יציב ועמיד יחסית ל־Overfitting. בנוסף, הוא מאפשר הפקת מדדים של חשיבות פיצ'רים בצורה אינטואיטיבית.
עם זאת, Random Forest עשוי להיות כבד חישובית בדאטה גדול, קשה לפרשנות בהשוואה למודלים פשוטים, ונוטה להעדיף את קטגוריות הרוב כאשר קיימת בעיית חוסר איזון בין המחלקות, כפי שקיים בדאטה שלנו. חשוב לציין כי בגרסתו בספריית scikit-learn, המודל אינו תומך בערכים חסרים ודורש טיפול מקדים, כגון השלמה או השמטה של רשומות לא שלמות.

XGBoost- מודל Boosting מתקדם, המבוסס על רצף של עצי החלטה אשר כל אחד מהם מתקן את שגיאות העץ הקודם. המודל מצטיין בהתמודדות עם דאטה רועש או לא ליניארי, ומשלב מנגנוני רגולריזציה (L1 ו־L2) למניעת Overfitting. הוא תומך בערכים חסרים באופן מובנה, דורש ייצוג נומרי בלבד (ולכן נדרש קידוד לכל המשתנים הקטגוריאליים), ומציע ביצועים גבוהים במגוון רחב של בעיות חיזוי.
במקרה של סיווג בינארי, XGBoost כולל את הפרמטר scale_pos_weight, אשר מאפשר להתמודד עם חוסר איזון בין הקטגוריות ולחזק את הקבוצה הנדירה (למשל - פציעות חמורות). עם זאת, בפרויקטים של סיווג רב־קטגורי, פרמטר זה אינו רלוונטי, ויש להשתמש באלטרנטיבות כמו sample_weight או טכניקות Oversampling כגון SMOTE.
יחד עם יתרונותיו, חשוב לציין כי XGBoost דורש כוונון פרמטרים מדויק וזמן חישוב ארוך יחסית בדאטה רחב, אך התמורה בביצועים לרוב מצדיקה זאת.

LightGBM- הוא מודל Boosting מתקדם, המתאפיין בביצועים מהירים ובחיסכון בזיכרון, במיוחד בעבודה עם דאטה גדול או בעל ממדיות גבוהה. המודל משתמש בטכניקות חדשניות כמו GOSS (Gradient-based One-Side Sampling) ובנייה בסגנון leaf-wise, שמאפשרות לו לבנות עצים עמוקים ביעילות גבוהה.
בניגוד למודלים אחרים, LightGBM תומך במשתנים קטגוריאליים ללא צורך בהמרה לנומריים, כל עוד הם מוגדרים כ־category. תמיכה זו מקצרת את שלבי ההכנה וחוסכת משאבים. עם זאת, המודל רגיש ל-Overfitting אם לא מכוונים אותו כראוי, במיוחד כאשר יש משתנים עם קטגוריות רבות או נדירות.

CatBoost- מודל חדשני מבית Yandex אשר מותאם במיוחד לעבודה עם דאטה הכולל משתנים קטגוריאליים. יתרונו העיקרי הוא בכך שהוא מבצע encoding פנימי ואוטומטי למשתנים קטגוריאליים, ללא צורך בהמרות ידניות כמו Label Encoding או One Hot Encoding. המודל שומר על יציבות החיזוי גם עם דאטה רועש או לא מאוזן, ומספק ביצועים תחרותיים מאוד, לעיתים אף ללא כוונון פרמטרים משמעותי. עם זאת, החסרונות של CatBoost כוללים מהירות מעט איטית יותר בהשוואה ל-LightGBM, וקושי בפרשנות בשל המבנה המורכב שלו.

בעת בחירת המודלים נבחנו מספר שיקולים מהותיים שהתבססו הן על דרישות אלגוריתמיות והן על מאפייני הדאטה שלנו. כל המודלים שנבחנו דרשו חלוקה של הנתונים לסט אימון ובדיקה, ולכן בוצע פיצול ל־Training ו־Test תוך שימוש ב־stratified split, אשר שמר על התפלגות הקטגוריות הלא מאוזנת גם בסט הבדיקה.
נפח הדאטה הגדול (כ־188,277 רשומות) אפשר הפקת תוצאות מהימנות, הן ברמת האימון והן בבדיקת ביצועים על סט נפרד. הנתונים אפשרו לנו גם ליישם טכניקות איזון כגון SMOTE לקבוצות מיעוט, מבלי להסתכן באובדן מייצוגיות.

מודלים כמו Logistic Regression רגישים יותר לאיכות ולפורמט הנתונים, ודורשים קלט נומרי מלא ונרמול של המשתנים הכמותיים. לעומתם, מודלים מבוססי עצים כמו Random Forest, XGBoost, LightGBM ו־CatBoost אינם רגישים לסקאלת המשתנים, ומרביתם כוללים גם מנגנוני התמודדות פנימיים עם ערכים חסרים פרט ל־Random Forest, אשר דורש השלמה מוקדמת של ערכים חסרים.
הנתונים כללו מספר רב של משתנים קטגוריאליים, שחייבו קידוד שונה בהתאם לדרישות כל מודל. מאחר שלרוב המודלים נדרש קלט נומרי בלבד, המרה לפורמט מספרי הייתה הכרחית לצורך עיבוד תקין. בהתאם לכך, בוצע קידוד של המשתנים הקטגוריאליים באמצעות טכניקות שונות, כגון Label Encoding, One-Hot Encoding, או .cat.codes - בהתאם למודל הספציפי שבו נעשה שימוש.
בנוסף, עבור Logistic Regression, שהוא המודל היחיד שרגיש לסקאלת המשתנים, ביצענו סטנדרטיזציה (Z-score) למשתנים הכמותיים.

השילוב בין עיבוד מותאם אישית לבין בחירת מודלים מגוונים אפשר לנו לבחון את בעיית הסיווג מכיוונים שונים, להפיק תובנות מדויקות ולשפר את ביצועי התחזיות באופן משמעותי.

1.2 הנחות המודל

במהלך תהליך המידול, נדרשנו לבצע מספר מניפולציות על הנתונים כדי להתאים אותם לדרישות הספציפיות של כל מודל שנבחר. לכל אחד מהמודלים שבהם השתמשנו קיימות הנחות שונות בנוגע לסוגי הנתונים שהוא מסוגל לעבד, לצורתם ולרמת ההכנה הנדרשת.

משתנים קטגוריאליים- הנתונים הכילו מספר רב של משתנים קטגוריאליים, שחייבו קידוד שונה לפי המודל. מודלים כמו Logistic Regression, Random Forest ו־XGBoost אינם תומכים בקטגוריות לא מקודדות, ולכן השתמשנו עבורם ב־Label Encoding, One Hot Encoding או .cat.codes.
גם במקרים שבהם נעשה שימוש במודלים שתומכים במשתנים מסוג category (כמו CatBoost ו־LightGBM), כאשר השתמשנו בטכניקת האיזון SMOTE/SMOTENC נדרשנו להמיר את המשתנים לייצוג נומרי באמצעות cat.codes, שכן אלגוריתמי האיזון אינם תומכים בעמודות קטגוריאליות לא מקודדות.

משתנים כמותיים ונרמול- מודלים כמו Logistic Regression רגישים להבדלים בסקאלות של משתנים כמותיים, ולכן ביצענו סטנדרטיזציה (נרמול Z-Score) רק עבור מודל זה. בשאר המודלים, בהם Random Forest, XGBoost, LightGBM ו־CatBoost, אין צורך בנרמול, שכן עצי החלטה אינם מושפעים מסקאלות המשתנים.

ערכים חסרים- הדאטה כלל שיעור גבוה של ערכים חסרים. התאמנו את הטיפול בהם לפי התמיכה של כל מודל: CatBoost, LightGBM ו־XGBoost כוללים מנגנוני טיפול פנימיים בערכים חסרים, ולכן יכולים להתמודד איתם ישירות.
Logistic Regression דורש קלט מלא בלבד, ולכן בוצעה השלמה מוקדמת של הערכים.
Random Forest אינו תומך רשמית בערכים חסרים ב־scikit-learn, ולכן גם עבורו נדרש מילוי מוקדם.
החלטנו לבצע השלמה לערכים חסרים לפני ביצוע כל המודלים - השלמנו ערכים חסרים באמצעות: חציון למשתנים כמותיים, ולמשתנים קטגוריאליים ביצענו השוואה בין שלוש שיטות שונות (מילוי לפי שכיח, Other, ושילוב-חלק מהעמודות מילאנו לפי שכיח וחלק לפי Other), ונמצא כי גישת Other הייתה הטובה מבניהן ולכן בחרנו בה.

איזון בין מחלקות- מאחר ועמודת המטרה הייתה בלתי מאוזנת באופן מובהק, נעשה שימוש בטכניקת SMOTE להגדלת קבוצות מיעוט. בנוסף, במודלים שתומכים בכך, הוגדרו משקלים מותאמים לקטגוריות נדירות.

פיצול ל־Train/Test- כל המודלים דרשו חלוקה לסט אימון וסט בדיקה. ביצענו פיצול סטטיסטי מבוקר באמצעות StratifiedSplit, ששמר על התפלגות הקטגוריות גם לאחר הפיצול ובכך אפשר הערכה מייצגת וביצועים מהימנים.

2. עיצוב בדיקה

לפני בניית המודלים בפועל, הקדשנו שלב ייעודי לתכנון אופן בדיקת הצלחתם. עיצוב בדיקה שיטתי ואחיד מאפשר לא רק להעריך את המודלים בצורה הוגנת, אלא גם להבין באילו תנאים הם מצליחים או נכשלים, ולהחליט מתי נכון להפסיק ניסוי ולהמשיך בגישה חלופית. תכנון זה כלל הגדרה ברורה של קריטריונים לאיכות המודל, ובחירה של מערך הנתונים שעל בסיסו יתבצעו ההשוואות.

מטרתנו הראשונית הייתה לפתח מודל שמסוגל לזהות במדויק את חומרת הפציעה בתאונות דרכים, תוך דגש מיוחד על איתור מקרי פציעה חמורים או קטלניים, מקרים בעלי השלכה תפעולית גבוהה.

בשל חוסר האיזון המובהק בין המחלקות, יושמה טכניקת SMOTE בסט האימון בלבד, לצורך איזון קבוצות המיעוט- זאת מבלי להוסיף מידע חדש, אלא על ידי יצירת דוגמאות סינתטיות מבוססות על תצפיות קיימות.

<u>קריטריונים למדידת הצלחה</u>
מכיוון שכל המודלים שנבחרו הם מסוג Supervised Learning, הערכתם התבצעה על בסיס השוואה ישירה בין תחזיות המודל לבין תוויות האמת. כל מודל נבחן הן כמודל עצמאי והן בהשוואה למודלים אחרים, תוך מתן חשיבות מיוחדת לביצועים על הקטגוריות הקריטיות (Severe Injury, Fatal Injury).
לצורך כך השתמשנו במדדי הערכה מגוונים:

- Accuracy - מדד כללי להצלחת הסיווג
- Precision - מידת הדיוק עבור כל קטגוריה
- Recall - זיהוי נכון של כלל המקרים האמיתיים (בדגש על קטגוריות חמורות)
- F1-Score - ממוצע הרמוני בין Precision ו־Recall
- Confusion Matrix - טבלת שגיאות המאפשרת ניתוח מפורט של התחזיות, כולל שיעורי זיהוי נכונים (True Positives/Negatives) ושגויים (False Positives/Negatives) לפי כל קטגוריה.

דגש מיוחד הושם על Recall לפציעות חמורות וקטלניות, שכן בפרויקט זה, המטרה היא צמצום הסיכון לפספוס מקרי פציעה. לפיכך, בוצעה הורדת סף סיווג (Threshold) מ־0.5 ל־0.3 (ולעיתים ערכים נוספים), כדי להגביר את הרגישות של המודלים על חשבון דיוק מסוים (Precision).

לסיכום, פעלנו בגישה איטרטיבית אך ממוקדת, לכל מודל הוקצו בין 2-4 סבבי ניסוי עם שינוי פרמטרים או גישות חלופיות, לפני מעבר למודל או אסטרטגיה אחרת.

<u>איטרציות וטיוב המודלים</u>
תהליך העבודה שלנו היה איטרטיבי ומתמשך. לאחר כל הרצה של מודל, ניתחנו את המדדים שקיבלנו, ובמיוחד את הביצועים על הקטגוריות הקריטיות. בהתאם לכך:
כיוונו Threshold לסף החלטה רגיש יותר לקבוצות מיעוט, שינינו היפר-פרמטרים בהתאם לביצועים הקודמים (למשל learning_rate, depth, n_estimators), שיפרנו את טכניקת האיזון (כמו שימוש ב-SMOTE ממוקד על חלק מהקטגוריות), והוספנו class-weight במודלים שתומכים בכך. כמו כן בחלק מהמודלים ניסינו להשתמש ב- grid search על מנת לכוונן את הפרמטרים של המודל בצורה המיטבית.

<u>אסטרטגיות חיזוי חלופיות</u>
נוסף למודלים שסיווגו את כל חמש הקטגוריות המקוריות, בחנו גם אסטרטגיות סיווג חלופיות, במטרה לשפר את ביצועי המודלים, בעיקר עבור קטגוריות הפציעה הקטלניות והחמורות. תחילה בנינו מודלים שהתמודדו עם כל חמש קטגוריות הפציעה, אך הביצועים ובעיקר מדדי ה־Recall עבור הקטגוריות החמורות והקטלניות לא היו מספקים.
לכן, בחנו אסטרטגיות חלופיות שכללו קיבוץ מחדש של הקטגוריות למספר גרסאות, כגון: חיזוי עם שלוש קטגוריות: "אין פציעה", "פציעה קלה", ו"פציעה חמורה/קטלנית" (איחוד שתי הקטגוריות החמורות).
בנוסף, יישמנו גם מודלים רב-שלביים שבנו תחזית מדורגת: מודלים דו־שלביים, לדוגמא: מודלים שבהם השלב הראשון ניבא האם קיימת פציעה, ובשלב השני בוצע חיזוי של סוג הפציעה (קלה, חמורה, קטלנית). בנוסף גם מודל תלת־שלבי, שבו: שלב ראשון חזה "פציעה"/"אין פציעה", אם יש פציעה אז שלב שני חזה "פציעה קטלנית"/"לא קטלנית", ומתוך הלא קטלניות, שלב שלישי הבחין בין "פציעה קלה"/"פציעה חמורה".
לבסוף, בחנו חיזוי בינארי פשוט: "פציעה" מול "ללא פציעה". נקטנו בגישה זו מתוך הבנה כי עדיף לא לפספס פציעות כלל, גם אם קלות, מאשר להסתכן בזיהוי חסר של פציעות חמורות.

מטרת הגישות הללו הייתה להקל על המודלים להתמודד עם חוסר האיזון ולהעלות את הרגישות (Recall) עבור כלל הפציעות בייחוד עבור המקרים הקריטיים מבחינה תפעולית. בסעיף הבא נרחיב על הפעולות שביצענו.

3. תיאור המודל

בשלב המידול, לאחר תהליך יסודי של ניקוי, קידוד והשלמת ערכים חסרים, בוצע תהליך נרחב של בניית והשוואת מודלים לצורך חיזוי חומרת הפציעה בתאונות דרכים. תחילה התאמנו מודלים למשימת סיווג רב־קטגורי (חמש רמות פציעה), ובהמשך נוסו גישות חלופיות, ובהן סיווג לשלוש רמות, מודלים דו־שלביים ותלת-שלביים, וכן חיזוי בינארי פשוט ("יש פציעה" / "אין פציעה").

כל שלב נבנה תוך התחשבות במגבלות הדאטה, בשיקולים של חוסר איזון, ובצורך במדדי הערכה מותאמים. הדגש המרכזי לאורך כל התהליך היה השגת Recall גבוה לפציעות חמורות או קטלניות, מתוך מטרה שלא לפספס אירועים קריטיים עבור שירותי חירום ורשויות האכיפה. עם זאת, ניתן משקל גם לדיוק התחזיות (Precision), מתוך הבנה שכמות גבוהה מדי של התראות שווא עשויה להפחית את האפקטיביות של מודל החיזוי ולפגוע באמון בו.

במהלך התהליך נעשה שימוש במגוון מודלים מתקדמים:
Logistic Regression, Random Forest, XGBoost, LightGBM ו־CatBoost, תוך שילוב טכניקות של קידוד משתנים, נרמול, SMOTE, שינוי ספים והוספת משקלים למחלקות מיעוט.

המודלים נוסו במספר תצורות:
חיזוי חמש קטגוריות (שלב ראשוני): התחלנו בחיזוי של המשתנה התלוי המקורי, הכולל חמש קטגוריות של פציעה :

```
Injury Severity
no apparent injury          154541
possible injury              18625
suspected minor injury       13380
suspected serious injury      1559
fatal injury                   172
Name: count, dtype: int64
```

המודלים שהופעלו בשלב זה לא הצליחו להניב ביצועים מספקים, בעיקר עקב חוסר איזון חמור בין הקטגוריות (ניתן לראות בתמונה), אשר פגע באופן משמעותי במדדי Recall- במיוחד עבור הקטגוריות הקריטיות ביותר. מכיוון ש־Recall גבוה חשוב במיוחד עבור קבלת החלטות בזמן אמת (למשל, הקצאת צוותים רפואיים או ניידות משטרה), הביצועים הנמוכים הובילו אותנו לשקול גישות נוספות:

- <u>מיפוי ל־3 קטגוריות:</u> בשלב הבא, מיפינו את המשתנה התלוי לשלוש קטגוריות עיקריות: אין פציעה, פציעה קלה, פציעה קשה/קטלנית. גישה זו הובילה לשיפור מסוים ב־precision וב־recall של חלק מהקטגוריות, אך עדיין לא פתרה את הבעיה המרכזית: המודלים נטו להחמיץ מקרים של פציעות חמורות - מצב בלתי רצוי בפרויקט שנועד לשמש גופים.
- <u>גישה מדורגת (מודלים רב-שלביים):</u> חיזוי בשלבים: ניסינו גם לפשט את המשימה בעזרת מודל רב-שלבי במסגרתו בוצעו שלושה חיזויים עוקבים: שלב ראשון - חיזוי האם התרחשה פציעה או לא, שלב שני - אם אכן זוהתה פציעה, ניבוי האם מדובר בפציעה קטלנית או לא ושלב שלישי - אם הפציעה אינה קטלנית, סיווג נוסף האם מדובר בפציעה קלה או פציעה חמורה.
גישה זו נועדה לצמצם את המורכבות של המשימה על ידי חלוקה הדרגתית, אך בפועל לא הביאה לשיפור מספק בביצועים, ובייחוד לא במקרים של פציעות חמורות - שם השאיפה הייתה להגיע ל־Recall גבוה ככל האפשר. בנוסף בשלב השני התוצאות העלו סימן ל-overfitting.
כמו כן ניסינו 2 מודלים דו-שלביים שכללו: שניהם בשלב הראשון חזו- יש פציעה/אין פציעה, אך בשלב השני: אחד המודלים חזה- פציעה קלה/קשה-קטלנית ומודל נוסף חזה 3 קטגוריות- פציעה קלה/חמורה/קטלנית.
אך גם הניסיונות האלה לא הועילו ולא הביאו לשיפור מספק בביצועים.
- <u>חיזוי בינארי ("יש פציעה" / "אין פציעה"):</u> לאור האתגרים שצוינו, החלטנו להתמקד בגישה זו בשלב האחרון. היא אומנם מצמצמת את רזולוציית התחזית, אך אפשרה שיפור משמעותי ב־Recall עבור אירועים עם פציעות - יעד מרכזי בפרויקט, נעשה חיזוי בינארי תוך שימוש ב־Random Forest, CatBoost, XGBoost , ו- LightGBM , החלטה זו התקבלה מתוך הבנה שעדיף לסווג כל פציעה כאירוע הדורש התייחסות, מאשר לפספס פציעות חמורות. בשלב זה נצפו ביצועים יותר טובים.

כל המודלים שנבחנו דורשים פיצול ל־Training ו־Test, ולכן בוצע פיצול מתאים מראש. הנתונים הומרו לקלטים מספריים בלבד באמצעות טכניקות קידוד שונות (Label Encoding, category.cat.codes ו־One-Hot Encoding),חוץ מ-CatBoost שלא הצריך קידוד אלא רק המרה לcategory, בהתאם לצרכי המודל. כמו כן, טיפלנו בבעיית חוסר האיזון באמצעות טכניקת SMOTE וכן על-ידי הוספת משקלים מוגברים לדגימות של פציעות חמורות.

עם זאת, חוסר האיזון בנתונים נותר אתגר מרכזי. SMOTE אומנם סייע במעט, אך אינו מוסיף מידע חדש, אלא רק מייצר וריאציות על סמך דוגמאות קיימות. לכן, כאשר מדובר באירועים נדירים ובעלי חשיבות קריטית כמו פציעות חמורות, הפער במידע המקורי ממשיך להשפיע לרעה על יכולת המודל לחזות אותן בדיוק גבוה.

3.1. הגדרת פרמטרים

בשלב ראשון, רוב המודלים נוסו עם פרמטרי ברירת מחדל לצורך הערכה ראשונית. בהמשך, בוצעו שיפורים ממוקדים בכמה מהמודלים המרכזיים - תוך כוונון ידני של פרמטרים וניסויים עם Grid Search (בפרט עבור XGBoost ו- LightGBM).
להלן הפרמטרים עם הביצועים הכי טובים שנבחרו במודלים העיקריים שלנו:

- **<u>LightGBM-</u>** מודלים בינאריים (יש פציעה / אין פציעה):
n_estimators=500, max_depth=10, learning_rate=0.03, class_weight='balanced', random_state=42, threshold=0.2473, בוצע קידוד One-Hot-Encoding למשתנים קטגוריאליים, וביצענו איזון באמצעות טכניקת SMOTE לאחר הפיצול לקבוצת האימון, הדאטה מחולקת ל 70% Train ו- 30% Test.
- **<u>XGBoost-</u>** מודלים בינאריים (יש פציעה / אין פציעה):
n_estimators= 350 ,max_depth = 6 , learning_rate = 0.05 , subsample = 0.8 , colsample_bytree = 0.8 ,threshold =0.208 ,random_state = 42 ,encoding=One-Hot Encoding ,SMOTE על נתוני האימון, הדאטה מחולקת ל 70% Train ו- 30% Test.
- **<u>CatBoost-</u>** מודל בינארי (יש פציעה / אין פציעה):
iterations=500, learning_rate=0.05, depth=6, loss_function='Logloss', random_state=42, verbose=0, threshold=0.25, הקידוד לעמודות הקטגוריאליות נעשה על-ידי שינוי טיפוס העמודות ל category באמצעות astype('category'). בהתאם לאופן שבו CatBoost מתמודד באופן מובנה עם משתנים קטגוריאליים, נעשה שימוש

ב- SMOTENC גרסה של SMOTE שתומכת בעמודות קטגוריאליות על קבוצת האימון, הדאטה מחולקת ל 70% Train ו- 30% Test.

- **Random Forest-** מודל בינארי (יש פציעה / אין פציעה):
n_estimators=300, max_depth=15, class_weight="balanced", threshold=0.48, random_state=42, encoding = One-Hot Encoding, SMOTE רק על נתוני האימון, הדאטה מחולקת ל 70% Train ו- 30% Test.

במהלך שלב המידול נבחנו מגוון רחב של מודלים ווריאציות שונות, שכללו התאמות פרמטרים, שיטות איזון וגישות קידוד שונות, במטרה לזהות את הפתרון המדויק והיעיל ביותר לחיזוי. בסעיף זה מוצגים רק המודלים שהשיגו את התוצאות הטובות ביותר בחיזוי יש פציעה / אין פציעה: LightGBM, CatBoost, XGBoost ו־Random Forest.
בנוסף, נוסו גם מודלים נוספים כגון Logistic Regression, וכן גרסאות אלטרנטיביות של המודלים (למשל חיזוי של 5 קטגוריות, חיזוי 3 קטגוריות, מודלים דו-שלביים ותלת-שלביים), אך אלו לא נכללו בדוח הנוכחי מאחר שביצועיהם היו נמוכים יותר ביחס למודלים הבינאריים שנבחרו להצגה. לכן בחרנו להציג את מודלים הבינאריים, המציגים תוצאות יחסית יציבות וכי נעדיף לא לפספס פציעה בכלל מאשר לקחת סיכון בזיהוי חסר של פציעות חמורות.

התאמות נוספות שבוצעו:

- SMOTE הופעל על קבוצת האימון בלבד, במטרה להתמודד עם חוסר איזון בין המחלקות.
- שינוי סף החיזוי (Threshold) בוצע בחלק מהמודלים (כגון LightGBM, XGBoost, CatBoost ו־Random Forest) במטרה לשפר את ה־Recall לקבוצת הפציעה. לדוגמה: ב־CatBoost הוגדר threshold = 0.25, ב־LightGBM וב-XGBoost נעשה שימוש בבחירת סף אופטימלי באמצעות F2-score (בחירת סף לפי אופטימיזציית מדד F2- בדגש על Recall) וב-Random Forest הוגדר threshold = 0.48.
- שיטות הקידוד השתנו בהתאם למודל: עבור CatBoost השתמשנו ב־.astype('category') בהתאם לתמיכה המובנית של המודל. בעוד שעבור LightGBM, XGBoost ו־Random Forest בוצע One-Hot Encoding.

3.2. תיאור המודלים

בהתבסס על הגישות שנוסו, נמצא כי חיזוי בינארי של "יש פציעה / אין פציעה" הניב את התוצאות היציבות והמדויקות ביותר. בניגוד לגישות הקודמות (חמש קטגוריות, שלוש קטגוריות או מודלים רב-שלביים), שהתקשו בזיהוי מקרי פציעה חמורים, גישה זו אפשרה שיפור ניכר ב־Recall של קבוצת הפצועים, מדד מרכזי בפרויקט זה.

Recall גבוה חשוב במיוחד בהקשרים תפעוליים כמו תיעדוף חירום והקצאת משאבים שם פספוס של פציעה חמורה עלול לגרום לנזק חמור. על-מנת להתמודד עם בעיית האי איזון בין הקבוצות, בוצעו מספר התאמות מקדימות: החלת SMOTE ליצירת איזון בסט האימון, הגדרת משקלים עבור קטגוריות המיעוט, והתאמת סף החיזוי (Threshold) בהתאם למודל ולתוצאותיו..

ארבעת המודלים המרכזיים שהוערכו בשלב הסופי היו LightGBM, XGBoost, Random Forest ו-CatBoost. להלן סקירה ממוקדת של מודלים אלו:

LightGBM- המודל אפשר להסיק מסקנות משמעותיות, שכן הצליח להבחין בצורה אפקטיבית בין מקרים של פציעה לבין מקרים ללא פציעה. ה-Recall לקבוצת הפצועים עמד על 0.95, מה שמעיד על רגישות גבוהה- יעד מרכזי בפרויקט שמטרתו למנוע פספוס של פציעות ובפרט חמורות וקטלניות.
לא זוהו תובנות חדשות או דפוסים חריגים, וההתפלגות התנהגה כמצופה. עם זאת, המודל הניב גם שיעור לא מבוטל של False Positives (6,328 מקרים), נתון שמחזק את הצורך באיזון בין רגישות לדיוק.
מבחינה טכנית, ההרצה בוצעה בצורה תקינה, ללא תקלות או שגיאות. זמן העיבוד היה סביר, והמבנה המודולרי של הקוד תרם ליעילות האימון. הנתונים עברו הכנה מוקדמת שכללה טיפול בערכים חסרים, כך שלא נרשמו קשיים מיוחדים באיכות הנתונים.לא אותרו חוסר עקביות בחישובים או התנהגויות לא צפויות במהלך ההרצות. המודל שמר על יציבות גבוהה בין הרצות שונות.

XGBoost- המודל הציג יכולת איתור גבוהה מאוד של מקרים עם פציעה, עם Recall של 0.973 לקבוצת הפצועים, ממצא שמתיישב עם היעד המרכזי של הפרויקט: צמצום פספוס של מקרים חמורים. עם זאת, Precision נמוך יחסית של 0.567 הביא לכך ש-7,515 מקרים סווגו כשגויים מסוג False Positive, כלומר המודל התריע על פציעה כאשר בפועל לא הייתה פציעה, תוצאה זו מעידה על הטיה ברורה לרגישות על חשבון הדיוק, תוצאה הצפויה בגישות המכוונות לזיהוי מרבי של מקרים קריטיים, אך יש בה גם חסרון תפעולי, שכן ריבוי התראות שווא עלול להוביל לשחיקה באמון המשתמשים או לבזבוז משאבים.
המודל לא חשף דפוסים חריגים או תובנות לא צפויות, ההתפלגות הייתה עקבית, עם שיעור גבוה של זיהוי פציעות, לצד שיעור טעויות גבוה יותר בקבוצת הלא פצועים.
מהבחינה הטכנית, תהליך ההרצה היה יציב וללא תקלות. זמן האימון היה סביר, אם כי ארוך מעט יותר בהשוואה ל־LightGBM. המבנה המודולרי של הקוד סייע ביעילות בתהליך הפיתוח.
בנוסף, תהליכי ההכנה המקדימים שכללו ניקוי נתונים, טיפול בערכים חסרים, וכן איזון הקטגוריות באמצעות smote, תרמו לכך שהמודל פעל בצורה חלקה וללא חריגות בהתנהגותו.

CatBoost- מודל זה הציג תוצאות דומות מאוד לאלו של LightGBM, עם Recall גבוה של 0.96, נתון שמבטא רגישות גבוהה לזיהוי פצועים, בהתאם למטרה המרכזית של הפרויקט: הפחתת פספוס של פציעות, בדגש על פציעות חמורות וקטלניות.
גם ה-Precision עמד על 0.59, קרוב מאוד ל־0.60 שהתקבל ב-LightGBM, כך שההבדלים בין המודלים היו מזעריים מבחינת הביצועים. התרשמנו שהמודל הצליח לשמור על יחס סביר בין זיהוי פצועים לבין כמות התראות השווא, ובכך ענה היטב על הצרכים התפעוליים של המערכת. בחרנו ב-CatBoost כמודל המועדף, בשל יתרונותיו המובהקים בעבודה עם נתונים קטגוריאליים : הוא תומך ישירות בעמודות מסוג category, מבלי לדרוש המרות כמו One-Hot Encoding, שכן תכונה זו השתלבה היטב עם מבנה הנתונים שלנו, שכלל מספר רב של משתנים קטגוריאליים, מה שהקל על תהליך ההכנה, חסך זמן ועזר לשמור על מבנה הנתונים המקורי.
מבחינת תפעולית, המודל פעל ביציבות מלאה, השלים את האימון בהצלחה, וזמן העיבוד היה סביר, גם אם מעט ארוך יותר מזה של LightGBM.
השילוב של דיוק גבוה, התאמה טכנית לנתונים קטגוריאליים ויציבות, הפך את CatBoost לבחירה המרכזית בפרויקט.

Random Forest- מודל זה נבחן במסגרת השוואת הביצועים בשלב החיזוי הבינארי ("יש פציעה / אין פציעה") תוך שימוש בסף מותאם (Threshold = 0.48). הוא הציג Recall של 0.94, המעיד על רגישות גבוהה לזיהוי מקרי פציעה - יעד מרכזי בפרויקט. עם זאת, Precision של 0.58 הוביל ל־6,928 חיזויי שווא (False Positives).
נוסף על כך, כמות הפציעות שלא זוהו בפועל (False Negatives) הייתה גבוהה יחסית: 657 מקרים לעומת 429 בלבד ב־CatBoost. לאור מטרת-העל של המערכת, צמצום פספוס של פציעות, במיוחד חמורות וקטלניות - נמצא שהמודל אינו נותן מענה מספק, ולכן לא נבחר למימוש בפועל.
מהבחינה הטכנית, המודל פעל ביציבות, ללא תקלות או חריגות. זמן האימון היה מהיר, והשלבים המקדימים (טיפול בערכים חסרים, One-Hot Encoding ו־SMOTE) בוצעו בהצלחה. עם זאת, האיזון בין Recall ל־Precision לא ענה לצרכים התפעוליים, ולכן המודל שימש כבסיס להשוואה בלבד.

סיכום:
בהתבסס על ניתוח תוצאות המודלים השונים, נמצא כי גישת החיזוי הבינארית ("יש פציעה / אין פציעה") הניבה את התוצאות היציבות והמדויקות ביותר, בשונה מגישות מרובות קטגוריות או מודלים שלביים, היא אפשרה שיפור ניכר ב-Recall של קבוצת הפצועים, מדד קריטי בפרויקט שמטרתו העיקרית היא זיהוי מלא של פציעות חמורות וקטלניות לצורכי תיעדוף חירום, הקצאת משאבים, ומניעת פספוס מסוכן.

כלל המודלים: LightGBM, XGBoost, CatBoost ו־Random Forest, עברו תהליך אחיד של הכנה שכלל טיפול באי־איזון באמצעות (SMOTE), קידוד משתנים קטגוריאליים (כאשר נדרש), והתאמת סף תחזית (threshold) במטרה לשפר Recall תוך שמירה על Precision תפעולי.

המודלים LightGBM ו־CatBoost הציגו איזון טוב בין Recall ל־Precision, בעוד XGBoost הדגים רגישות מקסימלית במחיר של ירידה משמעותית בדיוק. מודל Random Forest הציג ביצועים נמוכים יותר בזיהוי מקרי פציעה, נקודת חולשה מהותית בפרויקט זה.

אף שהמודלים לא הציגו בעיות ביצוע חריגות או זמני ריצה יוצאי דופן, ניתן להניח כי XGBoost ו־CatBoost דרשו מעט יותר זמן עיבוד לעומת LightGBM ו־Random Forest, בשל מאפייני האימון שלהם והיקף הנתונים המעובדים.

המסקנה הסופית היא כי המודל הנבחר בפרויקט הוא CatBoost בגישת חיזוי בינארית, אשר סיפק את השילוב הטוב ביותר בין ביצועים חזקים, רגישות גבוהה, והתאמה טכנולוגית מלאה למבנה הנתונים המכיל רוב של פיצ'רים קטגוריאליים.
פתרון זה תומך הן בקבלת החלטות מיידיות בשטח (כגון תיעדוף טיפול רפואי והקצאת כוחות חירום), והן בניתוח רחב היקף של תאונות לצורכי תכנון מדיניות תחבורתית מבוססת נתונים.

4. הערכת מודלים

בשלב זה, נערכה הערכה שיטתית ומעמיקה של כלל המודלים שנבחנו במסגרת הפרויקט, במטרה לזהות את שיטות החיזוי האפקטיביות ביותר לחיזוי חומרת פציעות בתאונות דרכים. ההערכה התבצעה במספר רמות סיווג שונות: חיזוי חמש קטגוריות מקוריות, חיזוי בשלוש קטגוריות מאוחדות, חיזוי בינארי (יש פציעה / אין פציעה), וכן מודלים רב-שלביים (דו-שלביים ותלת-שלביים) שנועדו להתמודד עם מורכבות וחוסר איזון בנתונים.

לכל אחת מהמשימות נבחנו מספר מודלים מתקדמים, ביניהם: Logistic Regression, Random Forest, XGBoost, LightGBM ו־CatBoost. כל מודל הוערך בהתאם לקריטריונים שנגזרו מהמטרות התפעוליות של המערכת: רגישות לזיהוי פציעות (Recall) נבחרה כמדד המרכזי, בשל החשיבות הגבוהה שבזיהוי מירבי של פציעות חמורות וקטלניות – מקרים שבהם פספוס עלול להוביל לסיכון חיים. דיוק חיובי (Precision) נבחן לצורך בקרת כמות ההתראות השגויות, אשר עלולות להוביל להקצאת יתר של משאבים. מדד F1 שימש לאיזון בין דיוק לרגישות ולהשוואה כללית בין מודלים. בנוסף, הוצג גם דיוק כללי (Accuracy) להשלמת התמונה, אך הוא נלקח בעירבון מוגבל נוכח חוסר האיזון המשמעותי בין הקבוצות – שכן מודל עשוי להשיג דיוק גבוה גם אם יפספס את מרבית מקרי הפציעה.

בנוסף למדדים המספריים, נלקחו בחשבון גם שיקולים איכותיים כגון: רגישות לחוסר איזון בין מחלקות, נוחות פרשנות, עמידות לרעש, והתאמה לסוגי המשתנים (כמותיים וקטגוריאליים). בכל אחד מהחיזויים שנבדקו נבנתה טבלה מסכמת של תוצאות המודלים, ולבסוף גובשו מסקנות באשר למודל המומלץ לכל תרחיש.

להלן פירוט של הגישות השונות שנוסו במהלך הפרויקט, כולל תיאור רמות הסיווג, המודלים שנבחנו, והערכת הביצועים שנמדדו בכל אחת מהן:

חיזוי חמש קטגוריות של חומרת פציעה:

בשלב זה בוצעה השוואה מקיפה בין ארבעה מודלים: Logistic Regression, Random Forest, XGBoost ו־LightGBM- במסגרת משימת סיווג רב קטגורית של חומרת פציעה לחמש הקטגוריות המקוריות:
Fatal Injury, Suspected Serious Injury, Suspected Minor Injury, Possible Injury, ו־No Apparent Injury.

הטבלה שלהלן מציגה את ביצועי המודלים לפי שלושת מדדי הדיוק המרכזיים: Precision, Recall ו־F1-score - עבור כל אחת מהקטגוריות. בנוסף, מוצגים גם מדדי סיכום כלליים: דיוק כולל (Accuracy), ממוצע מאוזן (Macro Avg) ו־ממוצע משוקלל (Weighted Avg).

CatBoost			Logistic Regression			Random Forest			XGBoost			LightGBM			
Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	
0.39	1	0.57	0.42	0.96	0.58	0.78	0.77	0.78	0.81	0.83	0.82	0.76	0.81	0.79	fatal injury
1	0.83	0.9	0.98	0.84	0.9	0.95	0.9	0.93	0.92	0.95	0.93	0.91	0.96	0.93	no apparent injury
0.3	0.53	0.38	0.3	0.49	0.37	0.36	0.53	0.43	0.41	0.4	0.41	0.42	0.38	0.4	possible injury
0.27	0.36	0.31	0.25	0.26	0.25	0.3	0.31	0.3	0.35	0.23	0.28	0.38	0.21	0.27	suspected minor injury
0.11	0.59	0.18	0.08	0.58	0.14	0.3	0.09	0.14	0.24	0.22	0.23	0.26	0.19	0.22	suspected serious injury
0.76			0.76			0.82			0.84			0.84			*accuracy*
0.41	0.66	0.47	0.41	0.63	0.45	0.54	0.52	0.52	0.55	0.53	0.53	0.55	0.51	0.52	*macro avg*
0.87	0.76	0.8	0.85	0.76	0.8	0.84	0.82	0.83	0.82	0.84	0.83	0.82	0.84	0.83	*weighted avg*

בהתבסס על מטרות הפרויקט זיהוי מקסימלי של פצועים, במיוחד פציעות חמורות וקטלניות, תוך הימנעות מהתראות שווא שיפגעו באפקטיביות המערכת, בוצעה השוואה בין חמישה מודלים בסיווג לחמש קטגוריות חומרת פציעה. XGBoost ו-LightGBM הציגו את האיזון הטוב ביותר בין Recall גבוה לקטגוריות הקריטיות לבין Precision סביר. XGBoost זיהה כ־83% ממקרי המוות (Recall = 0.83) בדיוק של 0.81, והגיע ל-Recall של 0.22 עבור פציעות חמורות. LightGBM הציג תוצאה כמעט זהה, עם Recall של 0.81 לקטלניות ו־0.19 לפציעות חמורות, אך עם מעט פחות דיוק בקטגוריות אלו. Random Forest הצליח בזיהוי פציעות אפשריות (Recall = 0.53), אך נכשל בזיהוי פציעות חמורות (Recall = 0.09) והציג F1 נמוך מאוד בקטגוריה זו (0.14), מה שמגביל את התאמתו למקרי קצה. Logistic Regression הפתיע עם Recall גבוה מאוד בקטגוריית "fatal injury" (0.96) ו־0.58 לפציעות חמורות, אך Precision נמוך מאוד (0.08 ב־"serious" ו־0.42 ב־"fatal") הפך אותו למודל שמזהה הרבה, אך מייצר כמות רבה של תחזיות שווא. CatBoost זיהה את כל מקרי המוות ללא יוצא מן הכלל (Recall = 1.00) וגם זיהה 59% מהפציעות החמורות, הנתון הגבוה ביותר מכל המודלים, אך עשה זאת עם Precision נמוך מאוד (0.39–0.11), מה שעלול לגרום לעומס תפעולי בלתי סביר בשטח.

לאור כלל הנתונים, והתוצאות הנמוכות והלא מספקות בעיקר בקטגוריות הקריטיות של פצועים חמורים וקטלניים, החלטנו לשנות גישה ולהמשיך לבחון חלופות נוספות, כגון סיווג לשלוש קטגוריות בלבד, מודלים רב-שלביים ועוד, במטרה לשפר את רמת הזיהוי בקטגוריות עבור קבלת החלטות בזמן אמת.

חיזוי בשלוש קטגוריות:

בשלב זה בוצע חיזוי מחודש, לאחר מיפוי מחדש של עמודת המטרה לשלוש קטגוריות מרכזיות בלבד:

- No Injury (ללא פציעה)
- Minor Injury (פציעה קלה- כולל "Possible Injury" ו־"Suspected Minor Injury")
- Severe Injury (פציעה חמורה – כולל "Suspected Serious Injury" ו־"Fatal Injury")

מהלך זה נועד להתמודד עם חוסר האיזון הקיצוני והמורכבות הרבה שעמדו בבסיס מודל החיזוי של חמש הקטגוריות המקוריות ובעיקר עם הקושי בזיהוי אמין של פציעות חמורות.
באמצעות איחוד הקטגוריות לצמצום מספר הקבוצות, שאפנו לשפר את יכולת ההבחנה של המודלים בין סוגי פציעות מרכזיים, תוך שמירה על מובהקות תפעולית של זיהוי מקרי פציעה חמורים בזמן אמת.

הטבלה להלן מציגה את ביצועי ארבעת המודלים המובילים: CatBoost, Random Forest, XGBoost ו־LightGBM- בהתאם למדדי Precision, Recall ו־F1-score עבור כל אחת מהקטגוריות, לצד ממוצעים כוללים (Macro Avg, Weighted Avg, ו־Accuracy).

CatBoost			Random Forest			XGBoost			LightGBM			
Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	
0.52	0.81	0.63	0.6	0.85	0.7	0.53	0.88	0.66	0.64	0.73	0.68	minor injury
0.99	0.84	0.91	0.97	0.89	0.93	0.99	0.84	0.91	0.95	0.92	0.93	no injury
0.13	0.59	0.22	0.5	0.14	0.22	0.18	0.49	0.26	0.39	0.26	0.31	severe injury
0.83			0.88			0.84			0.88			*accuracy*
0.55	0.75	0.59	0.69	0.63	0.62	0.57	0.74	0.61	0.66	0.64	0.64	*macro avg*
0.91	0.83	0.86	0.9	0.88	0.88	0.91	0.84	0.86	0.89	0.88	0.89	*weighted avg*

המטרה הייתה לאזן בין זיהוי פצועים חמורים לבין יציבות מערכתית. מהשוואת ביצועי המודלים עולה כי כל המודלים הציגו ביצועים טובים מאוד בקטגוריית "no injury" (F1 בין 0.91 ל־0.93) ותוצאות סבירות גם לפציעות קלות (Recall עד 0.88), אך ניכר קושי משמעותי בזיהוי נכון של פציעות חמורות, הקטגוריה הקריטית ביותר מבחינה תפעולית. שיעורי ה-Recall

לפצועים חמורים נעו בין 0.14 ל-0.5 בלבד, וה-Precision היה נמוך אף יותר (0.13-0.50), מה שמעיד על שיעור גבוה של התרעות שווא או פספוסים.
למרות ש-CatBoost בלט ברגישות לפציעות חמורות (Recall = 0.59), הוא סבל מדיוק נמוך מאוד (Precision = 0.13). לעומתו, LightGBM הציג את האיזון הטוב ביותר בין רגישות לדיוק, עם Macro F1 הגבוה ביותר (0.64) וביצועים יציבים בכלל הקטגוריות (Weighted F1 = 0.89), ולכן נבחר כמודל המועדף בגישה זו.

עם זאת, התוצאות הכוללות לא היו מספקות, בעיקר בשל הקושי בזיהוי מדויק של פציעות חמורות, ולכן הוחלט להמשיך ולבחון גישות מתקדמות יותר, ביניהן מודלים דו-שלביים ותלת-שלביים, במטרה לשפר את הביצועים בקטגוריות הקריטיות ולהתאים טוב יותר את המערכת לצרכים התפעוליים בשטח.

חיזוי על ידי גישה הדרגתית:
לאחר ניסויים עם מודלים שונים בשיטות סיווג ישיר, עברנו לבחון גם גישה הדרגתית, המבוססת על חיזוי בשלבים עוקבים. מטרת הגישה הייתה להתמודד בצורה יעילה יותר עם חוסר האיזון המשמעותי בנתונים ובמיוחד עם האתגר בזיהוי פציעות חמורות או קטלניות, שהן קריטיות בהיבט התפעולי ובמענה למצבי חירום. גישה זו מתבססת על עיקרון של פירוק המשימה המורכבת למספר שלבים פשוטים יחסית, במטרה לשפר את ביצועי המודל במקרים קריטיים ולהפחית את שיעור פספוסי הפציעה.

גישה תלת-שלבית: חיזוי מדורג של חומרת הפציעה
בשלב זה יושמה גישת חיזוי הדרגתית בשלושה שלבים, במטרה להתמודד עם חוסר האיזון הקיצוני ולשפר את הביצועים עבור מקרים חמורים. התהליך כלל:

- Stage 1 – חיזוי בינארי: האם קיימת פציעה (Yes/No).
- Stage 2 – עבור התצפיות שסווגו כבעלות פציעה: חיזוי האם מדובר בפציעה קטלנית או לא.
- Stage 3 – עבור התצפיות שסווגו כלא קטלניות: הבחנה בין פציעה קלה לפציעה חמורה.

הטבלה שלהלן מסכמת את תוצאות החיזוי עבור כל שלב, תוך השוואה בין שני מודלים מובילים: LightGBM ו־XGBoost. הנתונים כוללים את כלל המדדים עבור כל קטגוריה נבחנת (Precision, Recall, F1, Accuracy וכו'):

XGBoost			LightGBM			Label Description	
Precision	Recall	F1-Score	Precision	Recall	F1-Score		
1	0.82	0.9	0.98	0.87	0.92	no injury	Stage 1
0.54	0.99	0.7	0.62	0.93	0.74	Injury	
0.85			0.88			*accuracy*	
0.77	0.9	0.8	0.8	0.9	0.83	*macro avg*	
0.92	0.85	0.86	0.92	0.88	0.89	*weighted avg*	
1	1	1	1	1	1	Non-Fatal Injury	Stage 2
0.68	0.94	0.79	0.73	0.94	0.82	Fatal Injury	
1			1			*accuracy*	
0.84	0.97	0.89	0.87	0.97	0.91	*macro avg*	
1	1	1	1	1	1	*weighted avg*	
0.97	0.92	0.94	0.96	0.97	0.96	light Injury	Stage 3
0.17	0.34	0.23	0.25	0.23	0.24	serious Injury	
0.89			0.93			*accuracy*	
0.57	0.63	0.59	0.6	0.6	0.6	*macro avg*	
0.93	0.89	0.91	0.93	0.93	0.93	*weighted avg*	

בהתבסס על תוצאות השלב התלת־שלבי, ניתן לראות כי גם XGBoost וגם LightGBM הציגו שיפור משמעותי בזיהוי פצועים לעומת גישת הסיווג הישיר- במיוחד בקטגוריות הקשות יותר. בשלב הראשון, שמטרתו הייתה להבחין בין מקרים עם פציעה לבין מקרים ללא פציעה, שני המודלים הציגו ביצועים גבוהים במיוחד ב־Recall (0.99 ב-XGBoost ו-0.93 ב-LightGBM), מה שמעיד על יכולת טובה לזהות מקרים עם פציעה. עם זאת, כבר בשלב זה ניתן היה להבחין בפערים בין ערכי ה־Recall ל-Precision, בייחוד ב-XGBoost, שם זוהו רמות רגישות גבוהות אך במחיר של דיוק נמוך (Precision = 0.54), מה שמעיד על שיעור גבוה יחסית של התראות שווא.

בשלב השני, שמטרתו הייתה לסווג בין פציעות קטלניות ללא קטלניות, התקבלו תוצאות כמעט מושלמות בשני המודלים שהציגו ערכיRecall של 1.00 ו-0.94, לפציעות לא קטלניות וקטלניות בהתאמה. עם זאת, ערכים כה גבוהים מעלים חשש ממשי לOverfitting, במיוחד לאור גודל קבוצת הפציעות הקטלניות בדאטה והשימוש ב־SMOTE. תוצאה זו עשויה להעיד על למידה מדויקת מדי של דפוסי הדאטה המאומן, מצב שבו המודל מאבד מיכולת ההכללה שלו לתרחישים חדשים.

בשלב השלישי, שבו נדרש סיווג בין פציעה קלה לפציעה חמורה, נרשמה ירידה בביצועים, בפרט בזיהוי פציעות חמורות – הקטגוריה הקריטית ביותר ברמה התפעולית. שני המודלים הציגו Recall נמוך (0.34 ב-XGBoost ו־0.23 ב-LightGBM), לצד F1 נמוך במיוחד (0.23-0.24), מה שמעיד על קושי ממשי של המודלים להתמודד עם סיווג מדויק בשלב זה, למרות שהמערכת כבר צמצמה את ההתמקדות למקרים שבהם ישנה פציעה.

לסיכום, הגישה התלת־שלבית אפשרה התמודדות ממוקדת יותר עם המורכבות של חיזוי חומרת הפציעה, והביאה לשיפור בזיהוי פצועים (בשלבים הראשונים), אך עדיין נדרשת עבודה נוספת לשיפור הזיהוי של פציעות חמורות, תוך שמירה על איזון בין Recall ל־Precision. החשד ל Overfitting בשלב האמצעי חיזק את ההבנה שיש להיזהר מביצועים גבוהים מדי בדאטה מאומן ולבחון היטב את יכולת ההכללה של המודלים. לאור ממצאים אלו, החלטנו להמשיך ולבחון גישות נוספות.

גישה דו־שלבית: זיהוי פציעה ולאחר מכן חיזוי פציעות קלות מול חמורות/קטלניות

במטרה להתמודד עם חוסר האיזון הקיצוני ולהתמקד בזיהוי מדויק של פציעות חמורות, נוסתה גישה דו־שלבית, בה החיזוי בוצע בשני שלבים:

- Stage 1 – סיווג בינארי: האם קיימת פציעה (Injury) או לא (No Injury).
- Stage 2 – בקרב תצפיות שסווגו כבעלות פציעה: הבחנה בין פציעה קלה (Minor Injury) לבין פציעה חמורה או קטלנית (Severe/Fatal Injury).

גישה זו גובשה לאור מגבלות הגישה התלת-שלבית, שבה נרשמו סימני Overfitting בשלב הביניים וביצועים חלשים במיוחד לזיהוי מקרים חמורים.

הטבלה שלהלן מסכמת את ביצועי המודל (LightGBM) בכל שלב, על פי מדדי דיוק, Recall, Precision ו־F1 לכל תווית, וכן ממוצעים Macro ו־Weighted:

Precision	Recall	F1-Score	Label Description		
0.98	0.87	0.92	no injury	Stage 1	LightGBM
0.62	0.93	0.74	Injury		
	0.88		*accuracy*		
0.8	0.9	0.83	*macro avg*		
0.92	0.88	0.89	*weighted avg*		
0.97	0.88	0.92	Minor Injury	Stage 2	
0.19	0.51	0.28	Severe/Fatal Injury		
	0.86		*accuracy*		
0.58	0.7	0.6	*macro avg*		
0.93	0.86	0.89	*weighted avg*		

בחרנו לנסות גישה דו-שלבית שנועדה לשפר את רמת הזיהוי של פצועים, תוך הפחתת העומס החישובי והתפעולי.

בשלב הראשון סיווגנו את מקרי התאונה לשתי קבוצות בלבד: "אין פציעה" לעומת "יש פציעה". המודל השיג Recall של 0.93 לקבוצת הפצועים- נתון גבוה במיוחד, המעיד על רגישות מרשימה לזיהוי פציעות, ו-Precision של 0.62, רמת דיוק סבירה. ה־F1 בקטגוריה זו עמד על 0.74, והדיוק הכללי (accuracy) בשלב זה עמד על 88%.

בשלב השני נבנה מודל נוסף שהתמקד אך ורק בתצפיות שסווגו כ"יש פציעה", וניסה להבחין בין "פציעה קלה" לבין "פציעה חמורה/קטלנית". כאן נרשמה ירידה מסוימת ב-Recall של קבוצת הפציעות הקשות ל-0.51, מדובר בזיהוי של מחצית מהמקרים החמורים בלבד. יחד עם זאת, Precision בקטגוריה זו נותר נמוך (0.19), מה שמצביע על שיעור לא מבוטל של תחזיות שווא. ה־F1-score לקבוצה זו עמד על 0.28 בלבד, מה שמעיד על תוצאה לא מספקת.

שיטה זו אפשרה לנו למקד את תהליך הסיווג, תוך שמירה על רגישות גבוהה בשלבים הראשונים, ומתן הבחנה נוספת בשלב המתקדם. יחד עם זאת, לאור הביצועים המוגבלים בקטגוריות הקריטיות, ובעיקר בשל רמת הדיוק הנמוכה בזיהוי פציעות חמורות וקטלניות, מצאנו כי יש צורך להמשיך ולבחון אסטרטגיות חיזוי נוספות, במטרה למקסם את רמת החיזוי של פציעות קריטיות תוך שמירה על איזון תפעולי.

גישה דו־שלבית: זיהוי פציעה ולאחר מכן סיווג לשלוש קטגוריות

בשלב זה יושמה גישה דו-שלבית נוספת, שנועדה לשלב בין זיהוי יעיל של מקרים עם פציעה לבין סיווג מדויק לרמות חומרה:

- Stage 1- חיזוי בינארי של "יש פציעה / אין פציעה".
- Stage 2- עבור תצפיות שסווגו כבעלות פציעה: סיווג נוסף לשלוש קטגוריות נפרדות- פציעה קלה (Light Injury), פציעה חמורה (Serious Injury) ופציעה קטלנית (Fatal Injury).

גישה זו נועדה לשלב בין דיוק בסיסי בזיהוי פציעות לבין יכולת הבחנה בין דרגות חומרה שונות. הטבלה שלהלן מסכמת את תוצאות המודלים העיקריים (LightGBM, XGBoost, Random Forest ו-CatBoost) בכל אחד מהשלבים.

CatBoost			Random Forest			XGBoost			LightGBM				
Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Label Description	
0.99	0.85	0.92	1	0.79	0.88	0.99	0.84	0.91	1	0.82	0.9	no injury	Stage 1
0.59	0.96	0.73	0.51	1	0.67	0.57	0.97	0.72	0.55	0.99	0.71	Injury	
0.87			0.83			0.86			0.85			*accuracy*	
0.79	0.91	0.82	0.75	0.89	0.78	0.78	0.91	0.81	0.77	0.9	0.8	*macro avg*	
0.92	0.87	0.88	0.91	0.83	0.85	0.92	0.86	0.87	0.92	0.85	0.87	weighted avg	
0.74	0.82	0.78	0.72	0.76	0.74	0.71	0.76	0.74	0.72	0.86	0.79	Fatal Injury	Stage 2
0.96	0.92	0.94	0.96	0.98	0.97	0.95	0.99	0.97	0.96	0.94	0.95	Light Injury	
0.11	0.2	0.15	0.16	0.08	0.11	0.52	0.03	0.06	0.17	0.26	0.2	Serious Injury	
0.89			0.94			0.95			0.91			*accuracy*	
0.6	0.65	0.62	0.61	0.61	0.61	0.73	0.6	0.59	0.62	0.69	0.65	*macro avg*	
0.92	0.89	0.9	0.92	0.94	0.93	0.93	0.95	0.93	0.93	0.91	0.92	*weighted avg*	

בהתבסס על תוצאות המודלים בגישת החיזוי הדו-שלבית: שלב ראשון: "יש פציעה/אין פציעה", ושלב שני: סיווג לפציעה קלה, חמורה או קטלנית, ניתן להסיק כי הגישה השיגה שיפור מסוים בזיהוי מקרי פציעה, אך עדיין הציגה אתגרים בשלב השני.

בשלב הראשון, כל המודלים הציגו ביצועים טובים בזיהוי פצועים, עם Recall גבוה במיוחד: למשל, Random Forest הגיע ל־1.00, LightGBM ל־0.99, ו־CatBoost ל־0.96. עם זאת, ערכי ה־Precision היו נמוכים יותר – לדוגמה, 0.51 ב- Random Forest ו־0.59 ב-CatBoost. ה-F1-score לפצועים עמד בטווח של 0.67-0.73, כאשר CatBoost הציג איזון יחסי טוב (Precision = 0.59, Recall = 0.96, F1 = 0.73).

בשלב השני, שבו נותחו רק תצפיות שסווגו כ"יש פציעה", נצפו תוצאות נמוכות ב-Recall של קטגוריית הפציעה החמורה לדוגמה: 0.03 בלבד ב-XGBoost, 0.2 ב-CatBoost, ו-0.26 ב-LightGBM. גם ה-Precision בקטגוריה זו נותר נמוך מאוד (0.11-0.52), ו־F1-score נע בין 0.06 ל־0.2 בלבד. מה שמעיד על קושי משמעותי בזיהוי פציעות חמורות. לעומת זאת, הפציעות הקלות זוהו היטב, עם Recall של 0.92-0.99 ו־F1-score גבוה מאוד (מעל 0.94) בכל המודלים. קטגוריית הפציעה הקטלנית הציגה תוצאות סבירות, עם F1 של כ־0.74-0.79.

לסיכום, הגישה הדו-שלבית הצליחה לזהות היטב פצועים בשלב הראשון, אך לא הצליחה לספק זיהוי אמין ומדויק לפציעות חמורות בשלב השני. דבר שפוגע ביכולתה לשמש ככלי תומך החלטה לצוותים בשטח, ולכן נזנחה לטובת גישת החיזוי הבינארית.

<u>חיזוי בינארי: יש פציעה / אין פציעה – המודל הסופי שנבחר</u>

לאחר שנוסו מספר גישות מורכבות לחיזוי חומרת הפציעה כולל: סיווג לחמש קטגוריות מקוריות, שלוש קטגוריות מאוחדות, וכן מודלים דו-שלביים ותלת-שלביים. מצאנו כי הביצועים בקטגוריות הקריטיות ובפרט פציעות חמורות לא היו מספקים. מדדי ה-Recall וה-Precision היו נמוכים כאחד, מה שפגע באמינות ובאפקטיביות של המודלים ככלי תומך החלטה. בהתאם לכך, בחרנו לבחון גישה פשוטה יותר אך ממוקדת יותר מבחינה תפעולית: חיזוי בינארי של "יש פציעה" לעומת "אין פציעה". גישה זו נמצאה כאפקטיבית ביותר בפרויקט, לאור שילוב של Recall גבוה (שמעיד על זיהוי מירבי של פצועים), יציבות כללית בתוצאות, ודיוק מספק. על כן, היא נבחרה כבסיס למודל הסופי.

הבחירה בגישה זו התבססה על מספר שיקולים: ראשית, פספוס של פציעה מכל סוג גם אם קלה, עלול לעכב תגובה רפואית ולהוביל להחמרה. שנית, שירותי החירום שואפים לרגישות גבוהה ככל הניתן, במטרה שלא לפספס אף אירוע שמצריך תגובה בשטח. יחד עם זאת, היינו מודעות לכך שרגישות גבוהה במיוחד עשויה להוביל לעלייה בהתראות שווא, ולכן בחנו את המודלים גם לפי מדד ה־Precision וביקשנו לאזן בין שני המדדים. מעבר לכך, הגישה הבינארית סיפקה תוצאות יציבות, ברורות ומשכנעות יותר מכל שאר הגישות שנבדקו, שבהן נצפו ביצועים נמוכים וחוסר עקביות בזיהוי חלק מהקטגוריות. לכן, חיזוי בינארי נמצא כמתאים ביותר לצורך תפעולי, גם בהיבט של ביצועים וגם מבחינת מהימנות.

הטבלה הבאה מסכמת את ביצועי ארבעת המודלים המרכזיים: LightGBM, XGBoost, CatBoost ו־Random Forest.

CatBoost			Random Forest			XGBoost			LightGBM			Label Description
Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
0.99	0.85	0.92	0.98	0.85	0.91	0.99	0.84	0.91	0.99	0.86	0.92	no injury
0.59	0.96	0.73	0.58	0.94	0.71	0.57	0.97	0.72	0.6	0.95	0.74	Injury
0.87			0.87			0.86			0.88			*accuracy*
0.79	0.91	0.82	0.78	0.89	0.81	0.78	0.91	0.81	0.8	0.91	0.83	*macro avg*
0.92	0.87	0.88	0.91	0.87	0.88	0.92	0.86	0.87	0.92	0.88	0.89	*weighted avg*

במבחן הביצועים של גישת החיזוי הבינארי: זיהוי האם קיימת פציעה או לא, נצפו תוצאות טובות ויציבות בכלל המודלים, תוך שיפור משמעותי ביחס לגישות המורכבות שנוסו קודם לכן. מודל CatBoost, שנבחר לבסוף כמודל המרכזי בפרויקט, הציג Recall גבוה במיוחד בקבוצת הפציעה (0.96), לצד Precision של 0.59 ו-F1-Score של 0.73. ה-F1 המשוקלל (Weighted Avg) הגיע ל-0.88 מה שמעיד על איזון סביר בין זיהוי פציעות לבין מניעת התראות שווא. מעבר לתוצאות, הבחירה ב-CatBoost הושפעה גם מהתאמתו הטבעית לדאטה עתיר משתנים קטגוריאליים, ללא צורך בהמרות או קידוד מורכב.
יחד עם זאת, חשוב לציין כי LightGBM הניב ביצועים דומים: Recall של 0.95, Precision של 0.60 ו-F1 של 0.74 בקבוצת הפציעה. מבחינת ביצועים מדובר בהבדל קל ולא משמעותי, שלא הצדיק ויתור על היתרונות התפעוליים של CatBoost.
מודלים נוספים שנבדקו כללו את XGBoost, אשר השיג Recall של 0.97 (הגבוה ביותר מבין כל המודלים), אך Precision נמוך יותר של 0.57, שהוביל ל-F1 של 0.72. המודל נטה להעדיף זיהוי יתר של פציעות, דבר התורם לרגישות גבוהה אך הוביל ליותר התראות שווא. Random Forest השיג ביצועים מעט חלשים יותר מהשאר, עם Recall של 0.94, Precision של 0.58 ו־F1 של 0.71, נתונים טובים אך מעט פחות משאר המודלים.

לסיכום, כלל המודלים סיפקו תוצאות טובות בגישת החיזוי הבינארי, אך הבחירה ב-CatBoost נעשתה מתוך שיקול של יציבות, רגישות גבוהה לזיהוי פציעות, והתאמה למאפייני הדאטה. העדפנו לזהות כמה שיותר פציעות, גם במחיר של התראות שווא מתונות, תוך שמירה על דיוק תפעולי סביר שמונע הצפה של מערכת קבלת ההחלטות.

בשלב זה על סמך הביצועים במודלים ובגרסאות החיזוי השונים החלטנו לשנות את המטרה הראשונית שלנו: מחיזוי רמות הפציעה, עברנו להתמקד בחיזוי **האם יש פציעה** או לא. החלטה זו נבעה מהעדפה להתייחס גם לפציעות קלות כמו לפציעות חמורות, כדי לא לפספס מקרים מסוכנים, וגם משום שהתוצאות שהתקבלו במודל הבינארי היו מדויקות ויציבות יותר.

לאחר בחירת מודל CatBoost בגישת חיזוי בינארית כמודל הסופי, בחנו האם ניתן להפיק ממנו מידע נוסף על חומרת הפציעה בפועל באמצעות ניתוח ההסתברויות שחוזה המודל, במטרה לשפר את התרומה התפעולית של המערכת.

<u>ניסיון לשילוב הסתברויות המודל לזיהוי חומרת הפציעה במודל הבינארי הנבחר - CatBoost :</u>

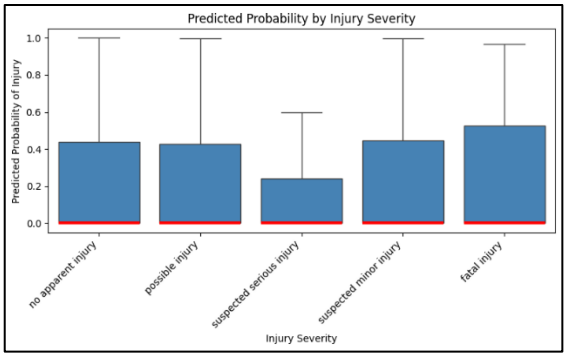
בשלב זה ביקשנו לבחון האם ניתן להרחיב את החיזוי הבינארי (יש פציעה / אין פציעה) ולהפיק ממנו מידע נוסף על חומרת הפציעה בפועל, מבלי לפתח מודל נפרד. מאחר שמודל CatBoost מחזיר לכל תצפית הסתברות משוערת להשתייכות לקטגוריית "יש פציעה", החלטנו לבחון האם ערך ההסתברות יכול לשמש כאינדיקציה עקיפה לדרגת החומרה. כלומר, האם תצפיות עם הסתברות גבוהה במיוחד משויכות בפועל למקרי פציעה חמורה או קטלנית. הרעיון שעמד בבסיס הבדיקה היה, כי ייתכן שקיים קשר בין רמת הביטחון של המודל בזיהוי פציעה לבין מידת החומרה בפועל, כך שניתן יהיה להשתמש בהסתברות כמדד משלים המספק סדר עדיפויות או התרעה מוגברת במקרי פציעות קשות יותר.

לצורך כך, חישבנו את הסתברויות המודל עבור כל תצפית בסט הבדיקה, יצרנו טבלת סיכום של ממוצע, חציון וסטיית תקן לכל קטגוריה של חומרת פציעה בפועל, והצגנו את ההתפלגות באמצעות תרשים Boxplot. כמו כן, בחנו האם קיימת הפרדה ברורה בין קטגוריות הפציעה על פי ההסתברויות.

תוצאות הבדיקה הראו כי החציון בכל הקטגוריות כמעט אפס (≈0.00), והממוצעים דומים מאוד בין כל רמות החומרה (טווח של 0.17–0.24), כולל עבור פציעות קטלניות. בנוסף, בתרשים ההתפלגות נצפתה חפיפה משמעותית בין כל הקטגוריות, ללא מגמה ברורה של עלייה בהסתברות במקרי פציעות חמורות. ממצא זה מעיד כי המודל אינו מצליח, במסגרת החיזוי הבינארי הקיים, להבחין בין דרגות חומרה על בסיס ערך ההסתברות בלבד.

מסקנתנו היא כי השיטה אינה מספקת ערך מוסף מהותי ואינה מתאימה ליישום תפעולי לצורך דירוג חומרת פציעה. ולכן נישאר עם החיזוי הבינארי שמביא תוצאות מספקות למטרת זיהוי האם יש פציעה או אין.

להלן הגרף והטבלה שעליהם התבססנו:



	Injury Severity	mean	median	std
0	fatal injury	0.237555	0.000003	0.360536
1	no apparent injury	0.209914	0.000004	0.327767
2	possible injury	0.206904	0.000004	0.326615
3	suspected minor injury	0.209363	0.000004	0.327100
4	suspected serious injury	0.176943	0.000003	0.307364

סיכום:

לאורך שלב המידול בפרויקט הושקעה מחשבה רבה בבחירת טכניקות מתאימות, ניסוי של גישות מגוונות והשוואה שיטתית בין מודלים, במטרה לזהות את הפתרון היעיל, המאוזן והאמין ביותר לחיזוי חומרת פציעות בתאונות דרכים.

מסקירת כלל הגישות שנבדקו עולה כי חיזוי של חמש קטגוריות פציעה לא הניב תוצאות מספקות, בשל חוסר איזון בולט בנתונים, בעיקר בקטגוריות הקשות כמו פציעות חמורות וקטלניות. חוסר איזון זה גרם ל־Recall נמוך במקרים הקריטיים ביותר. מעבר למודל של שלוש קטגוריות מאוחדות הביא לשיפור חלקי בביצועים, אך עדיין לא נתן מענה הולם לבעיית זיהוי הפצועים הקשים. הגישות הרב־שלביות, הן הדו־שלבית והן התלת־שלבית, אפשרו פירוק לוגי של הבעיה והובילו לשיפורים מתונים בזיהוי פציעות, בעיקר פציעות קטלניות, אך גם הן התקשו לדייק בקבוצות המאתגרות ביותר לזיהוי, ובראשן קטגוריית הפציעות החמורות שאינן קטלניות. לעומת זאת, החיזוי הבינארי, שמבחין בין "יש פציעה" לבין "אין פציעה", נמצא כגישה היעילה והמאוזנת ביותר. הוא השיג תוצאות מרשימות הן מבחינת Recall והן מבחינת יציבות כללית של המודל, ולכן נבחר כפתרון הסופי והמומלץ בפרויקט זה.