****תמונה שמכילה טקסט, גופן, צילום מסך, ירוק

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

**Data modeling Report**

**רום קדוש – 207691841**

**סנדרה זייגרמכר -206987521**

**תוכן עניינים**

1. **בחירת טכניקות מידול .................................................................... 3-6**
2. **עיצוב המודל ובדיקתו ...................................................................... 6-7**
3. **בדיקות המודל ................................................................................ 7-19**
4. **הערכת המודל ................................................................................. 19-20**

ֿ

1. **בחירת טכניקות מידול**

במהלך בחינת מטרות הפרויקט ודרישות החיזוי, בוצעה התאמה בין המודלים לבין הצרכים המעשיים של המערכת. לאחר ניתוח מעמיק נמצא כי מודלי K-Means ו-LSTM אינם מתאימים באופן מלא למטרות הפרויקט שלנו:

* K-Means הוא מודל לא מפוקח המבצע אישכול ואינו מתמקד בחיזוי משתנה מטרה מפורש כמו חומרת תאונה. לכן, למרות תרומתו בזיהוי דפוסים כלליים, הוא אינו עונה על הדרישה המרכזית של חיזוי חומרת התאונה ברמה פרטנית.
* LSTM הוא מודל סדרות זמן המתמקד בחיזוי מגמות לאורך זמן, ואינו מיועד לניבוי חומרה של אירועים בודדים על סמך תנאים רגעיים. מאחר והפרויקט מתמקד בחיזוי חומרה לכל תאונה ספציפית ולא בזיהוי דפוסים עונתיים, נמצא כי LSTM פחות רלוונטי לבעיה שהוגדרה.

לפיכך, שני המודלים הוצאו מהמשך הפיתוח לטובת התמקדות במודלים מפוקחים, מדויקים ורלוונטיים יותר למשימת סיווג חומרת תאונות.

* 1. בחירת טכניקות המידול המתאימות

צורך בחלוקה לסטים

|  |  |  |
| --- | --- | --- |
| **מודל** | **חלוקה לסטים** | **פירוט** |
| **Random Forest Classifier** | כן | מודל זה דורש סט אימון וסט בדיקה מאחר והוא מבוסס על אנסמבל של עצי החלטה עצמאיים. חלוקה לסטים מאפשרת למנוע Overfitting ולבדוק האם היער שנוצר מצליח לחזות בצורה טובה גם נתונים שלא ראה. המודל מתמודד היטב עם רעשים בנתונים ומאפשר גם לנתח חשיבות של משתנים בצורה אינטואיטיבית וברורה. |
| **CatBoost Classifier** | כן | מודל גרדיאנט בוסטינג מותאם לנתונים קטגוריים, הדורש סט אימון וסט בדיקה להבטחת הכללה טובה. CatBoost מפחית Overfitting על ידי שימוש בטכניקות מתקדמות, אך חשוב לבדוק ביצועים גם על דאטה חדש. חלוקה לסטים תבטיח שהמודל לא לומד רק את הדאטה אלא באמת חוזה דפוסים כלליים. מתאים במיוחד לפרויקטים עם הרבה תכונות קטגוריות. |
| **Sklearn GBC** | כן | גם מודל זה דורש חלוקה לסט אימון וסט בדיקה, מכיוון שהוא בונה עצים בסדרה על בסיס תיקון שגיאות קודמות. בלי חלוקה מתאימה קיים סיכון משמעותי ל-Overfitting. אימון על סט ולבדוק על סט נפרד מאפשר הערכה מדויקת של איכות השיפור ההדרגתי של המודל על הדאטה. |
| **XGBoost** | כן | מודל זה שייך למשפחת המודלים המפוקחים, ולכן מחייב חלוקה לסט אימון וסט בדיקה. סט האימון משמש ללמידת הקשרים בין הפיצ'רים למשתנה המטרה, בעוד שסט הבדיקה משמש לבחינת יכולת החיזוי של המודל על מידע חדש שלא "נראה" במהלך האימון. פיצול זה מונע overfitting ומאפשר לבדוק האם המודל טוב על דאטה שלא נחשף אליו בעבר. |
| **Logistic Regression** | כן | זהו מודל **סיווג מפוקח** קלאסי – כלומר, לומד לחזות את חומרת התאונה לפי פיצ'רים. לכן יש לחלק את הדאטה כך שהמודל ילמד על סט האימון וייבחן על סט הבדיקה. בלי חלוקה זו, אין דרך לדעת אם המודל לומד את התופעה באמת או פשוט "משנן" את הנתונים הקיימים. |
| **DBSCAN** | לא | הוא אלגוריתם למידה **לא מפוקחת**, לכן אין צורך בסט אימון או בדיקה. הוא אינו מתבסס על משתנה מטרה אלא מזהה דפוסים ומבנים (אשכולות) בעצמו מתוך מבנה הנתונים. כל הדאטה משמש כדי למצוא קבוצות של תאונות על סמך הקרבה הגיאוגרפית שלהן. |

נפח הנתונים העומד לרשותנו – מעל 19,000 תצפיות שנאספו לאורך שנתיים מאפשרת לנו להפיק תובנות איכותיות וליישם מגוון רחב של מודלים בצורה אמינה. כל אחד מהמודלים בהם בחרנו תומך או אף דורש בסיס נתונים רחב אשר ייתן לנו דיוק גבוהה ככל הניתן.

כמויות הנתונים

**Random Forest Classifier**

הוא מודל אנסמבל חזק הבנוי ממאות עצי החלטה, וכל עץ לומד מתוך דגימה שונה של הנתונים. ככל שכמות הרשומות גדולה יותר, ניתן לבנות יער מגוון ומדויק יותר, המונע תופעות של הטיות קיצוניות. נפח של מעל 19,000 תאונות מאפשר לכל עץ ללמוד דפוסים שונים ועדיין לשמר רמות גבוהות של דיוק והכללה. בנוסף, החלוקה לסט אימון וסט בדיקה נשמרת איכותית בזכות הכמות, מה שמחזק את אמינות התחזיות של המודל.

**CatBoost Classifier**

המודל מתמחה בלמידה מנתונים בעלי פיצ'רים קטגוריים, והיעילות שלו עולה ככל שכמות הדאטה זמינה לניתוח. מעל 19,000 רשומות מספקות בסיס מצוין לזיהוי קשרים מורכבים ועדינים בין משתנים כמו סוג דרך, סוג תאורה ומזג אוויר לבין חומרת התאונה. נפח הנתונים מאפשר ל-CatBoost לבצע אופטימיזציה מהירה ויעילה מבלי ליפול לבעיית overfitting, תוך שמירה על יכולת הכללה גבוהה גם כאשר קיימים משתנים רבים. חלוקה נכונה לסטים נתמכת היטב בגודל המדגם.

**Gradient Boosting Classifier - Sklearn GBC**

המודל בנוי מרצפים של עצי החלטה קטנים, שכל אחד מהם מתקן את השגיאות של קודמיו. כמות נתונים גדולה כמו 19,000 רשומות קריטית להצלחת האלגוריתם, מאחר והיא מאפשרת לו ללמוד שגיאות באופן מדויק ומבוקר לאורך שלבי האימון. הנתונים הרבים מאפשרים גם חלוקה אפקטיבית לסט אימון וסט בדיקה מבלי לפגוע בגיוון הדגימות, ובכך תומכים ביצירת מודל חזק שמייצג היטב את המציאות.

**מודל XGBoost**

הוא מודל חיזוי מתקדם המותאם במיוחד להתמודדות עם כמות נתונים גדולה ומורכבת. ככל שכמות הרשומות גדולה יותר, המודל יכול ללמוד תבניות מדויקות יותר מבלי ליפול לבעיה של overfitting. מעל 19,000 רשומות מספקות למודל בסיס יציב לבניית עצים חזקים ולביצוע אופטימיזציה מדויקת. נפח כזה של נתונים מאפשר גם לבצע חלוקה לסט אימון וסט בדיקה מבלי לפגוע במידת הייצוג של הנתונים.

**Logistic Regression**

כמודל סיווג מפוקח, מסתמך על מספר גבוה של דוגמאות כדי ללמוד את ההסתברויות בין הפיצ'רים לבין משתנה המטרה (חומרת התאונה). הנתונים הרבים מאפשרים לבצע אימון איכותי ולזהות קשרים סטטיסטיים מובהקים בין גורמים כמו מזג אוויר או תמרור לבין סוג התאונה. המודל נהנה ממבנה פשוט אך חזק ככל שיש יותר תצפיות, ולכן כמות הנתונים שלנו תומכת בהחלט ביישומו.

**אלגוריתם DBSCAN**

נשען על עקרון של צפיפות – ולכן יעיל במיוחד כאשר יש מספר רב של תצפיות במרחב. ככל שיש יותר תאונות במרחב הגיאוגרפי, כך ניתן לאתר אשכולות משמעותיים יותר. הנפח הגבוה של הנתונים מאפשר לאלגוריתם לזהות אזורים מסוכנים, להבדיל בין אזורים עם ריכוז גבוה של תאונות לבין נקודות מבודדות, ולתת מענה גם כאשר יש "רעש" במפה.

איכות הנתונים לכל מודל

|  |  |  |
| --- | --- | --- |
| **פירוט** | **דרישת איכות נתונים** | **מודל** |
| מתמודד היטב עם רעשים וחסרים, אך איכות הנתונים משפיעה על עומק ההבנה. נדרש קידוד עקבי של משתנים קטגוריים, והסרת הערכיים החריגים תשפר את ביצועי המודל. | בינונית/גבוהה | **Random Forest Classifier** |
| מתאים במיוחד לנתונים עם משתנים קטגוריים – דורש קידוד אחיד אך מתמודד טוב עם ערכים חסרים. ככל שהנתונים נקיים ועקביים יותר, המודל מדויק יותר. | גבוהה | **CatBoost Classifier** |
| רגיש לבעיות כמו Overfitting וערכים קיצוניים. יש לבצע קידוד מספרי עקבי וניקוי ערכים חריגים. ככל שהנתונים מאוזנים ומדויקים יותר, כך עולה איכות התחזיות. | בינונית | **Sklearn GBC** |
| מתמודד עם ערכים חסרים, אך איכות הנתונים תשפיע על מידת הדיוק. ניקוי נתונים ושימוש במילונים קריטיים. | בינונית/גבוהה | **XGBoost** |
| נתונים מקודדים היטב, ללא ערכים חסרים, ישפרו את יציבות המודל. נדרשת נורמליזציה במקרה הצורך. | גבוהה | **Logistic Regression** |
| רגיש מאוד לדיוק במיקום – נדרשת המרה נכונה של קואורדינטות והסרת נקודות חריגות. | גבוהה | **DBSCAN** |

התאמת הנתונים למודלים ודרישות להמרת ערכים

במהלך בניית המודלים, נדרשנו לוודא שהנתונים תואמים את דרישות הקלט של כל מודל – כלומר, האם נדרש לבצע המרה, קידוד או שינוי מבני כלשהו. לא כל המודלים מצריכים עיבוד מתקדם, אך חלקם מחייבים מבנה נתונים מסוים כדי לפעול בצורה תקינה.

* **Random Forest Classifier** - מודל זה פועל על עקרון של אנסמבל עצי החלטה, ולכן דורש שהנתונים יהיו מקודדים בצורה מספרית. הוא פחות רגיש לסקיילינג ולכן אין חובה לבצע נרמול. עם זאת, חשוב להבטיח שהקידוד אחיד ושאין ערכים חריגים קיצוניים אשר עלולים להשפיע על איכות התחזיות. הנתונים שעיבדנו בפרויקט עומדים בדרישות אלה.
* **CatBoost Classifier** - מסוגל להתמודד באופן טבעי עם משתנים קטגוריאליים, אך דורש עקביות בקידוד ובסדרי הערכים. לא נדרשת נורמליזציה או המרה מיוחדת, אך חשוב לטפל בערכים חסרים ולהבטיח שהקטגוריות מייצגות בצורה נכונה את מאפייני הדרך והתאונה. הנתונים שהוכנו במסגרת הפרויקט תואמים לדרישות המודל, ומאפשרים עבודה רציפה ויעילה.
* **Sklearn GBC** - דורש שהנתונים יהיו מקודדים בצורה מספרית. אין דרישה לנורמליזציה מאחר והאלגוריתם מבוסס עצים. עם זאת, יש צורך לוודא שהנתונים נקיים מערכים חריגים גדולים במיוחד אשר עלולים להטות את תהליך הלמידה. הנתונים שלנו גם כן עומדים בדרישות אלו.
* **XGBoost** - מודל זה דורש שהנתונים יהיו מקודדים בצורה מספרית, מאחר שהוא פועל על עקרון של עצי החלטה. במקרה שלנו, הנתונים המקוריים כבר הגיעו בפורמט מקודד – כלומר, לכל ערך קטגוריאלי (כגון מזג אוויר, סוג תאונה או מצב כביש) יש מספר קבוע. לכן, לא נדרשת המרה מיוחדת, אך חשוב לוודא שהקידוד עקבי ואחיד לאורך כל השדות, וכי אין ערכים חריגים.
* **Logistic Regression** – רגרסיה לוגיסטית פועלת בצורה טובה על נתונים מקודדים מספרית. כל עוד הקידודים עקביים, אין צורך בהמרה מורכבת. עם זאת, לעיתים משתלם לבצע נרמול לפיצ'רים (למשל, מהירות מול תאורה) כדי לזרז את זמן האימון ולמנוע השפעה לא רצויה של ערכים בטווחים שונים. במקרה שלנו, הנתונים מתאימים למודל וניתן להריץ אותו גם ללא שינוי, אך נבצע ניתוח מקדים כדי להחליט האם כדאי לבצע נרמול נוסף.
* **DBSCAN** – אלגוריתם זה מתמקד בניתוח מרחבי, ולכן דורש נתונים גיאוגרפיים מדויקים. ההמרה של עמודות הקואורדינטות X ו־Y למערכת גיאוגרפית עולמית של קווי רוחב ואורך (latitude, longitude) הינה קריטית. המרה זו כבר בוצעה בפרויקט באמצעות ספריית pyproj, ולכן הנתונים מתאימים בצורה מושלמת לניתוח אשכולות גיאוגרפיים באמצעות DBSCAN.

**1.2 הנחות המודל**

במהלך תהליך הבחירה של המודלים, נדרשנו לבצע מספר הנחות ולבצע מניפולציות נתונים על מנת להתאים את הנתונים לדרישות המודלים שבחרנו.

1. שמות הרחובות אינם מקודדים – אך נשמרים לשם ניתוח גיאוגרפי

שדה הרחוב (street) לא הומר לקידוד מספרי, מאחר והוא מכיל ערכים טקסטואליים המשקפים מיקום גיאוגרפי אמיתי (שמות רחובות). עמודה זו הופקה באמצעות המרת הקואורדינטות, לכן היא חשובה עבור ניתוחים מרחביים ולא עבור מודלי חיזוי הדורשים ערכים מספריים. לכן, לא נעשה קידוד, אך נעשה שימוש בעמודה רק בהקשרים גיאוגרפיים.

2. שימוש בעמודת מחוז (district) כתחליף כאשר הרחוב לא זוהה

כאשר עמודת הרחוב הכילה ערך "Unknown", נעשה שימוש בעמודת district (מחוז) כמקור מידע חלופי, על מנת לא לאבד ערך גיאוגרפי לחלוטין. ההנחה היא שלמרות החסר, ניתן להסיק לפחות את האזור הכללי בו התרחשה התאונה – דבר המסייע לאלגוריתמים גיאוגרפיים לזהות ריכוזים.

3. שאר העמודות מקודדות מספרית – ולכן מתאימות למודלים מפוקחים

שאר הפיצ’רים בדאטה (למשל: חומרת התאונה, סוג דרך, מזג אוויר וכו’) כבר מקודדים מספרית ולכן מתאימים באופן טבעי להרצת מודלים שונים.

4. התייחסות לרעש וערכים חסרים

ישנם מודלים רגישים אשר מחייבים טיפול מקדים בערכים חריגים או חסרים. את הטיפול המלא בנתונים עשינו בdata prepration.

1. **עיצוב המודל ובדיקתו**

בשלב זה גיבשנו תוכנית מסודרת לבחינת ביצועי המודלים השונים שייבחנו בפרויקט שלנו. בגלל שתהליך בניית המודלים כרוך בהרצות חוזרות ונשנות תוך התאמה ועדכון של הפרמטרים נדרשנו להגדיר מראש כיצד נמדוד את איכות התחזיות של כל מודל, אילו נתונים ישמשו לאימון ובדיקה, ומהם המדדים האובייקטיביים שישמשו אותנו לקביעת הצלחתו או כשלונו של כל מודל.

* במודלים המפוקחים XGBoost, Logistic Regression, Random Forest, CatBoost, GBC – נדרש לבצע חלוקה של הנתונים לסט אימון וסט בדיקה. בחרנו בחלוקה ביחס של 80% לאימון ו־20% לבדיקה, ולא בחלוקה של 70/30, וזאת מתוך מטרה לאפשר למודלים ללמוד בצורה מקיפה יותר מתוך הנתונים הזמינים. מאחר והנתונים שלנו כוללים כ־19,000 רשומות, יחס של 20% לבדיקות עדיין מספק מדגם משמעותי לבדיקה מבלי לפגוע באיכות האימון. חלוקה זו תורמת לאיזון אופטימלי בין לימוד המודל לבין הערכתו על קבוצה שאינה חופפת. את ביצועי המודל נבחן באמצעות מדדים מקובלים כמו Accuracy (דיוק), Precision (דיוק חיובי), Recall (רגישות), ו־F1 Score (ממוצע הרמוני של Precision ו־Recall). בנוסף, תנותח טבלת Confusion Matrix כדי להבין את אופן הסיווג בפועל – האם המודל מצליח להבחין באופן נכון בין תאונות קלות, קשות וקטלניות.
* לעומת זאת, במודל הבלתי מפוקח DBSCAN אין צורך לבצע חלוקה של הנתונים לסטים. מודל זה מנתח את כלל הדאטה ומחפשים בו מבנים פנימיים כמו אזורים בעלי ריכוז גבוה של תאונות. איכות המודל תיבחן באמצעות תצוגות ויזואליות (כגון מפות חום או מפות אינטראקטיביות) ובאמצעות מדדים כגון Silhouette Score או Davies-Bouldin Index, לבחינת ההומוגניות בתוך כל אשכול וההבדלים ביניהם.

תכנון זה מבטיח שנוכל למדוד את ביצועי כל אחד מהמודלים בצורה הוגנת, עקבית, מקיפה ואובייקטיבית, תוך אפשרות להשוואה ביניהם וחידוד ההחלטות לאורך הדרך.

1. **בדיקות המודל**

**מודל 1 – Random Forest Classifier**

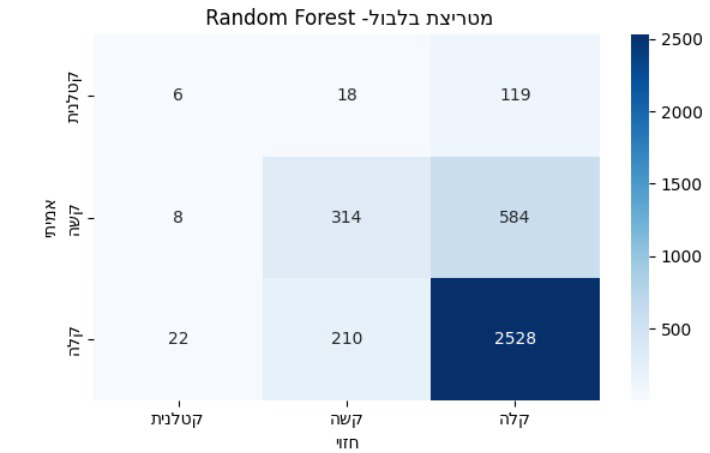
• פרמטרים - במודל זה נעשה שימוש ב־RandomForestClassifier עם ברירת מחדל של n\_estimators=100 ועץ אקראי לכל איטרציה (random\_state=42) לצורך שיחזור תוצאות. כמו בשאר המודלים, בוצעה המרה של עמודת המטרה humrat\_teuna כך שהערכים 1, 2, 3 הומרו בהתאמה ל־0 (קטלנית), 1 (קשה), ו־2 (קלה). המודל פועל על בסיס עצים מבוססי בוטסטראפ, ולכן אינו רגיש לנרמול הנתונים, מה שמתאים במיוחד לפורמט הקיים בדאטה המקודד.

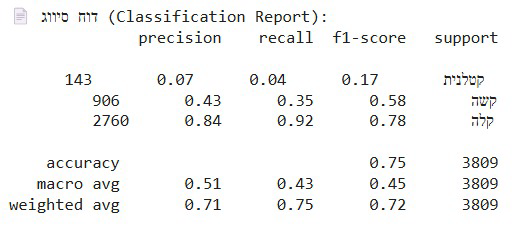
• בנייה - המודל הוזן ב־13 פיצ'רים נבחרים הכוללים פרמטרים סביבתיים וגיאוגרפיים: סוג הדרך, סמל יישוב, שנת התאונה, חודש, יום/לילה, יום בשבוע, סוג התאונה, מהירות מותרת, תקינות הדרך, סימון/תמרור, תאורה, מזג אוויר ומצב פני הכביש. כמו במודלים האחרים, פיצול הנתונים נעשה ביחס של 80% לאימון ו־20% לבדיקה.

• תוצאות - המודל הגיע לרמת דיוק כללית של 0.75, תוצאה טובה שמצביעה על איזון כללי בין ביצועים לבין שמירה על ייצוג של כלל הקטגוריות. עבור תאונות קלות, התקבלו תוצאות מצוינות – Recall של 0.92 ו־Precision של 0.84, בדומה ל־XGBoost. עבור תאונות קשות, התקבל Recall של 0.35 ו־Precision של 0.43 – ביצועים בינוניים, אך טובים יותר מאשר ברגרסיה לוגיסטית. עבור תאונות קטלניות, Recall עמד על 0.04 בלבד, עם Precision של 0.07 – כלומר רק 6 תאונות קטלניות מתוך 143 זוהו נכון (כמו XGBoost לפני שיפור שקילה).

• מטריצת בלבול - מטריצת הבלבול מדגישה את המגמה - המודל מזהה היטב תאונות קלות (2,528 מתוך 2,760). יש בלבול משמעותי בין תאונות קשות לקטגוריה הקלה – 584 טעויות סיווג. התאונות הקטלניות כמעט כולן סווגו כ"קלה" או "קשה", דבר שממחיש את האתגר המרכזי בפרויקט – להתמודד עם מחלקות שאינן מאוזנות בגודלן.

עם זאת, יחסית לרגרסיה לוגיסטית, המודל כן מצליח להחזיר זיהוי מסוים של תאונות קטלניות, מה שמעיד על פוטנציאל לשיפור באמצעות שקילה או oversampling.

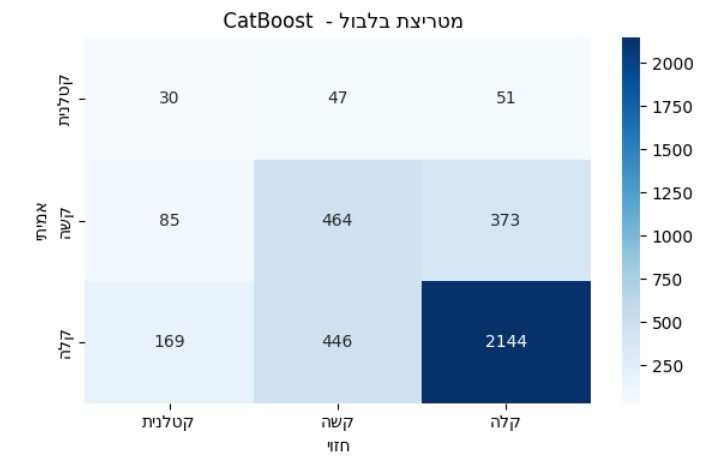


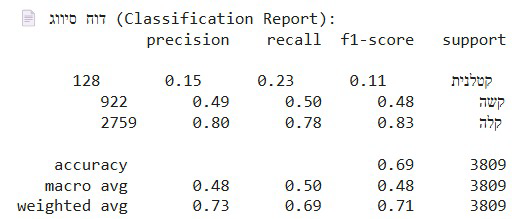
****

**מודל 2 – CatBoost Classifier**

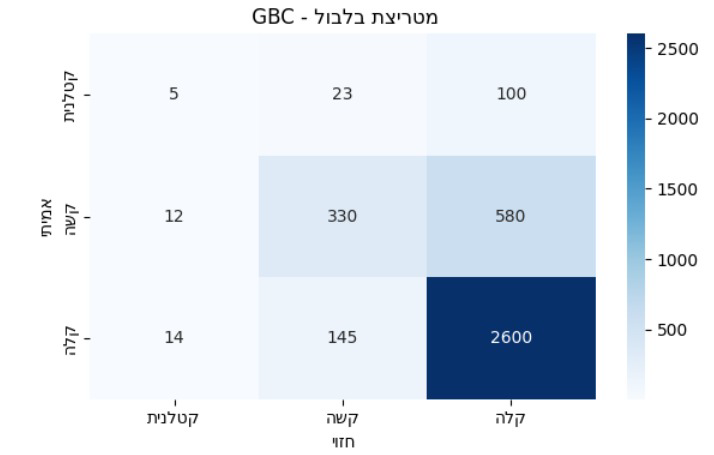
* פרמטרים - המודל נבנה באמצעות CatBoostClassifier עם פרמטרי ברירת מחדל, תוך ביטול הפלט (verbose=0) על מנת לשמור על ניקיון התצוגה בקולאב. כמו בשאר המודלים, בוצעה המרה של עמודת המטרה humrat\_teuna כך שהערכים 1, 2, 3 הומרו בהתאמה ל־0 (קטלנית), 1 (קשה), ו־2 (קלה). CatBoost אינו דורש נרמול או קידוד נוסף עבור משתנים מספריים, ולכן הנתונים המקודדים התאימו לו באופן ישיר.
* בנייה - למודל הוזנו 13 פיצ’רים הכוללים נתונים מרחביים, סביבתיים ותנאי דרך: סוג הדרך, סמל יישוב, שנת התאונה, חודש התאונה, יום/לילה, יום בשבוע, סוג התאונה, מהירות מותרת, תקינות הדרך, סימון/תמרור, תאורה, מזג אוויר ומצב פני הכביש. פיצול הנתונים בוצע ביחס של 80% לאימון ו־20% לבדיקה תוך שמירה על עקביות (random\_state=42).
* תוצאות - המודל הגיע לרמת דיוק כללית של 0.69, מעט נמוכה מהמודלים האחרים, אך חשוב לציין שהפעם יש שיפור ניכר ביכולת לזהות גם קטגוריות פחות נפוצות, עבור תאונות קלות התקבלו ביצועים טובים – Precision של 0.80 ו־Recall של 0.78. עבור תאונות קשות, המודל הציג ביצועים סבירים: Precision של 0.49 ו־Recall של 0.50 – כלומר חצי מהמקרים זוהו נכון. עבור תאונות קטלניות, נרשמה קפיצה משמעותית בביצועים יחסית למודלים אחרים: Recall של 0.23 (לעומת 0.00 ברגרסיה הלוגיסטית ו־0.04 ב-RF). כלומר, כמעט רבע מהתאונות הקטלניות זוהו נכונה – נתון חשוב מאוד מבחינה פרקטית.
* מטריצת בלבול - מהמטריצה ניתן לראות כי המודל זיהה נכון 30 מתוך 128 תאונות קטלניות, וזיהה 47 כ"קשה" ו־51 כ"קלה" – מראה מאמץ להבחין בין מחלקות חמורות. מתוך 922 תאונות קשות, 464 זוהו נכון, ו־458 זוהו בטעות כ"קלות" או "קטלניות". מתוך 2759 תאונות קלות, 2144 זוהו נכון, אך 615 סווגו כקשות/קטלניות – תופעת לוואי של מאמץ המודל לאזן את התחזיות.

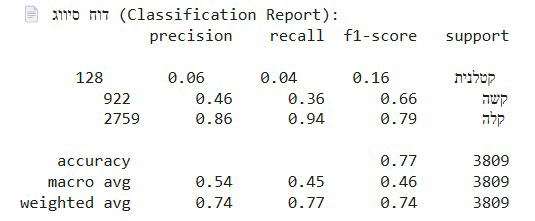
שורה תחתונה: המודל מראה מאמץ כן ומוצלח יחסית לאזן בין מחלקות, גם אם זה בא על חשבון הדיוק הכללי. הוא מצליח באופן ייחודי לזהות מקרים חמורים, מה שהופך אותו לרלוונטי במיוחד לפרויקט שבו הצלת חיים קודמת לדיוק סטטיסטי כולל.



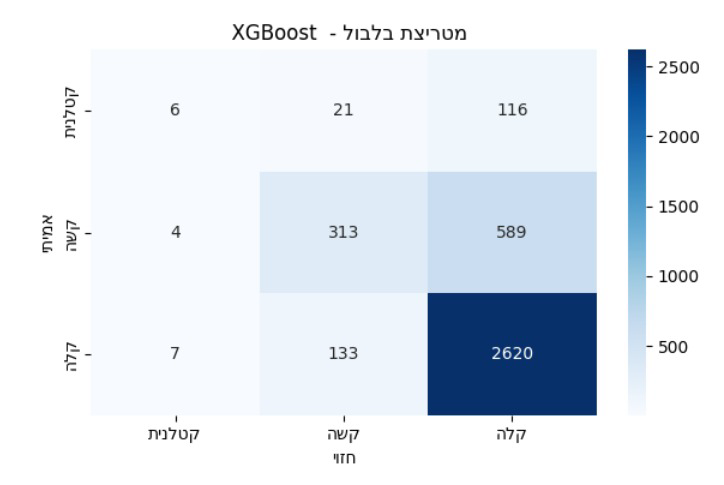


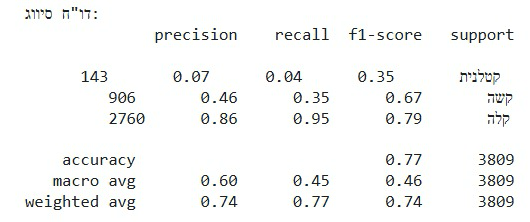
**מודל 3 – Sklearn GBC**

* פרמטרים - המודל נבנה באמצעות GradientBoostingClassifier עם טיוב ידני, n\_estimators=300 – מספר גבוה של עצים שמספק עומק לימוד טוב יותר. max\_depth=5 – מאפשר למודל ללמוד יחסי גומלין לא שטחיים מדי וכן להעמיק. learning\_rate=0.1 – קצב למידה יציב שלא גורם לאימון המודל לקפוץ מהר מדי. פיצול הנתונים בוצע ב־stratify=y כדי לשמר את הפרופורציות בין קטגוריות החומרה גם ב־Train וגם ב־Test.
* בנייה - כמו בשאר המודלים, הוזנו 13 פיצ'רים הכוללים תנאים סביבתיים, מאפייני כביש, וזמן התאונה: סוג דרך, סמל יישוב, שנת תאונה, חודש, יום/לילה, יום בשבוע, סוג תאונה, מהירות מותרת, תקינות, סימון/תמרור, תאורה, מזג אוויר, מצב פני הכביש. הפיצול ל־80% אימון ו־20% בדיקה נשמר קבוע לאורך כל המודלים להשוואה הוגנת.
* תוצאות - המודל הציג את דיוק התחזיות הכללי הגבוה ביותר מכל המודלים עד כה – Accuracy = 0.77 אבל חשוב לא רק הדיוק הכולל – אלא איך הוא מתמודד עם כל קטגוריה, תאונות קלה: Recall של 0.94 ו־Precision של 0.86 – מצוין. תאונות קשה: Recall של 0.36, די קרוב ל־Random Forest ו־XGBoost. תאונות קטלנית: Recall נמוך של 0.04, אבל עדיין עדיף מ־Logistic Regression שזיהתה 0.
* ****מטריצת בלבול - המטריצה ממחישה: רק 5 מתוך 128 תאונות קטלניות זוהו נכונה, ו־123 סווגו בטעות כ"קלה" או "קשה". מתוך 922 תאונות קשות, כ־36% (330) זוהו נכון, היתר ברחו ל"קלות". מתוך 2,759 תאונות קלות, זוהו נכונה 2,600 – תוצאה טובה אך לא מושלמת.

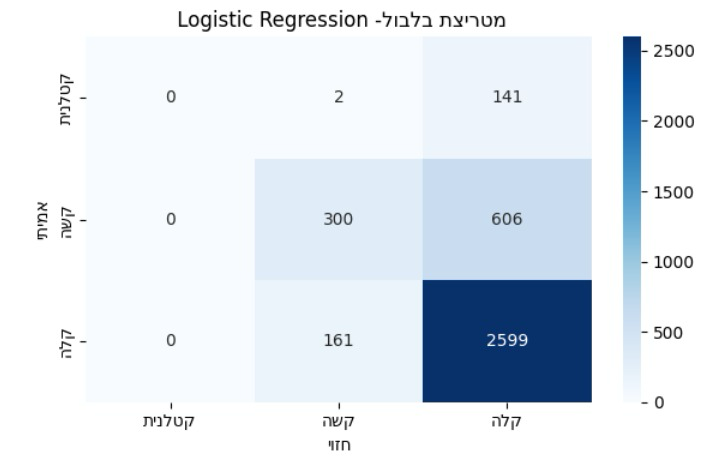
****

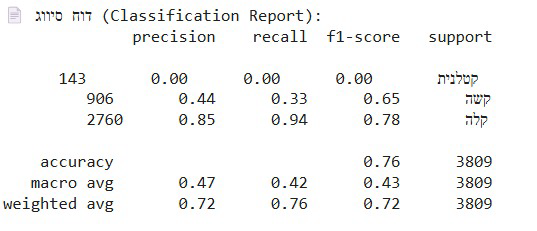
**מודל 4 – XGBoost**

* פרמטרים - במודל נעשה שימוש בפרמטרים ברירת מחדל, תוך ביטול אזהרת הקידוד הפנימית. הגדרת מדד ההפסד כ־mlogloss על מנת שנוכל להתמודד עם בעיית סיווג רב-ערכית. בנוסף, בוצעה המרה של עמודת המטרה 'חומרת התאונה' (humrat\_teuna), כך שהערכים 1, 2, 3 הומרו ל-0 (קטלנית), 1 (קשה), ו-2 (קלה), בהתאמה, על מנת להתאים את הקלט לפורמט ש-XGBoost דורש.
* בנייה - נבחרו 13 מאפיינים רלוונטיים למודל: סוג הדרך, סמל יישוב, שנת התאונה, חודש, יום/לילה, יום בשבוע, סוג התאונה, מהירות מותרת, תקינות דרך, סימון/תמרור, תאורה, מזג אוויר, ומצב פני הכביש. הנתונים פוצלו ביחס של 80% לאימון ו־20% לבדיקה, תוך שמירה על עקביות כדי לאפשר שיחזור תוצאות. הבחירה ביחס של 80/20 נועדה לאפשר מקסימום מידע ללמידה תוך שמירה על סט בדיקה משמעותי.
* תוצאות - המודל הגיע לרמת דיוק כללית של 0.7716 – תוצאה טובה בהתחשב בריבוי הקטגוריות. עם זאת, ניכרת שונות בין הדיוקים של כל מחלקה: עבור תאונות "קלה" התקבלו תוצאות טובות במיוחד עם Precision של 0.86 ו-Recall של 0.95. עבור תאונות "קשה" התקבלו ביצועים בינוניים יחסית (Precision: 0.46, Recall: 0.35), בעוד שתאונות "קטלנית" קיבלו ביצועים חלשים יחסית, עם Recall נמוך מאוד של 0.04 בלבד – כלומר המודל כמעט ואינו מזהה תאונות קטלניות.
* מטריצת הבלבול מראה שמרבית התחזיות נופלות על הקטגוריה של "תאונות קלות", דבר שממחיש את חוסר האיזון הקיים בנתונים ומדגיש את הצורך בשיטות נוספות להתמודדות עם class imbalance.



**מודל 5 – Logistic Regression**

* פרמטרים - במודל זה נעשה שימוש בפרמטרים סטנדרטיים של Logistic Regression, תוך שימוש ב־multi\_class='multinomial' ו־solver='lbfgs' כדי לאפשר סיווג רב-ערכי (שלוש רמות חומרה). בוצעה המרה של עמודת המטרה 'חומרת התאונה' (humrat\_teuna), כך שהערכים 1, 2, 3 הומרו ל־0 (קטלנית), 1 (קשה), ו־2 (קלה), בהתאמה, לצורך התאמה למודל. לא בוצעה נורמליזציה של המשתנים, מאחר והנתונים היו מקודדים מראש.
* בנייה - הרגרסיה נבנתה תוך שימוש ב־13 משתנים שנבחרו בקפידה: סוג הדרך, סמל יישוב, שנת התאונה, חודש, יום/לילה, יום בשבוע, סוג התאונה, מהירות מותרת, תקינות הדרך, סימון/תמרור, תאורה, מזג אוויר ומצב פני הכביש. הנתונים פוצלו לסט אימון וסט בדיקה ביחס של 80/20 עם random\_state=42, על מנת לאפשר שיחזור תוצאות ולהבטיח ייצוגיות של כלל הקטגוריות בסט הבדיקה.
* תוצאות - המודל הגיע לדיוק כללי של 0.7611, תוצאה טובה יחסית אך מייצגת במידה מסוימת את העובדה שהמודל "משוחד" לקטגוריה הנפוצה ביותר – תאונות קלות. עבור תאונות קלות, המודל הציג ביצועים מצוינים: Precision של 0.85 ו־Recall של 0.94. עבור תאונות קשות, הביצועים היו בינוניים: Precision של 0.44 ו־Recall של 0.33 – כלומר רק כשליש מהמקרים זוהו נכונה. עבור תאונות קטלניות, לא התקבלו כמעט תוצאות כלל: Recall של 0.00 – כלומר אפס זיהוי בפועל של תאונות מסוג זה.
* מטריצת בלבול - מטריצת הבלבול מדגישה את הבעיה המרכזית במודל – רובן המוחלט של התחזיות מופנות לקטגוריית תאונה קלה (2599 מקרים), בעוד שתאונות קשות וקטלניות נבלעות תחת התחזיות השגויות. מתוך 143 תאונות קטלניות בסט הבדיקה, אף לא אחת זוהתה נכון, מה שמוכיח שהמודל כלל אינו רגיש למקרים החמורים ביותר.



**מודל 6 – DBSCAN**

במהלך שלב המידול בפרויקט ניסינו להריץ את אלגוריתם DBSCAN על הנתונים, מתוך מחשבה ראשונית שהוא יוכל לסייע בזיהוי קבוצות או תבניות בהתאם למשתנה המטרה – חומרת התאונה. אך מהר מאוד הבנו שהנחת העבודה הזו שגויה, כיוון ש־DBSCAN הוא אלגוריתם למידה לא-מפוקחת ואינו מיועד לחיזוי או סיווג של משתנה מטרה קיים, אלא לאיתור אשכולות טבעיים בנתונים בהתבסס על צפיפות מרחבית בלבד, ללא שימוש במשתנה המטרה. לכן, הוא אינו מתאים לניבוי של חומרת התאונה כמו שאר האלגוריתמים, שכן הוא כלל לא "לומד" מהקשר בין פיצ'רים לתוצאה. עם זאת, זיהינו של־DBSCAN יש ערך רב בפרויקט: הוא מאפשר לנו לזהות נקודות חמות (Hotspots) של תאונות על גבי המפה, ולהמחיש ויזואלית היכן מרוכזות תאונות – תובנה שיכולה לשמש בסיס להצלבת מידע עם חומרת התאונות, ולעזור לנו להבין באילו אזורים יש גם ריבוי תאונות וגם נטייה לחומרה גבוהה. לכן, למרות שאינו משמש למידול חומרת התאונה, DBSCAN הוא אלגוריתם חשוב בהיבט הגיאוגרפי והוויזואלי של הפרויקט שלנו.

* 1. **הגדרת פרמטרים**

**מודל 1 – Random Forest Classifier**

בשלב זה של העבודה, בחרנו להשוות בין שתי גרסאות של מודל Random Forest Classifier במטרה לשפר את יכולת הזיהוי של תאונות חמורות – בפרט תאונות קטלניות, שהן קריטיות ביותר בהיבט מניעת אובדן חיים. לשם כך, בוצעה השוואה בין קוד 1 (בסיסי) לבין קוד 2 (משופר), תוך שינוי פרמטרים עיקריים והעמקת שלב הפיצול והלמידה.

בסעיף הקודם נעשה שימוש בפרמטרים ברירת מחדל לרוב:

* n\_estimators=100 100 עצים בלבד.
* max\_depth=None  ללא הגבלת עומק.
* min\_samples\_leaf=1 כל עלה יכול להכיל תצפית אחת.
* class\_weight='balanced' משקלים מאוזנים.
* הפיצול ל־Train/Test בוצע ללא stratify מה שיצר הטיה בהתפלגות הקבוצות.

לעומת זאת, בסעיף זה נעשה שיפור משמעותי:

* n\_estimators=200 כמות עצים גדולה יותר ללמידה יציבה.
* max\_depth=10 עומק מוגבל, למניעת overfitting.
* שימוש ב־class\_weight='balanced'  נשמר כדי להתמודד עם חוסר איזון בקטגוריות.
* הפיצול נעשה עם stratify=y כדי לשמר את יחס המחלקות המקורי בין סט האימון לסט הבדיקה.

השוואת התוצאות בין שני הקודים מראה שינוי משמעותי:  
בעוד שבקוד הבסיסי המודל הצליח לזהות רק 4% מהתאונות הקטלניות (Recall = 0.04), לאחר השיפור נרשמה קפיצה מרשימה ל־50% (Recall = 0.50) – שיפור של יותר מפי 12. גם התאונות הקשות זוהו ברמה של 34% Recall, והדיוק הכללי נשמר ברמה גבוהה של 71.6%.

למרות ש־Precision לקטלניות עדיין נמוך (0.13), מדובר בתוצאה צפויה כששמים דגש על Recall במטרה לצמצם החמצות של מקרים קריטיים. מדד ה־F1 המשוקלל נותר גבוה (0.72), מה שמעיד על איזון טוב בין דיוק לרגישות.

לסיכום, הקוד החדש מביא לידי ביטוי איזון נכון יותר בין ביצועים כלליים לבין רגישות למחלקות הקריטיות. המודל מצליח לזהות בצורה טובה יותר את התאונות הקטלניות והקשות – מבלי לפגוע באמינות החיזוי הכוללת – ומהווה מועמד מוביל לשילוב במערכת מבצעית לצורכי חיזוי ומניעה.

**מודל 2 - CatBoost Classifier**

בשלב זה של העבודה, בחרנו להשוות בין שתי גרסאות של מודל CatBoost Classifier במטרה לשפר את יכולת הזיהוי של תאונות חמורות – בפרט תאונות קטלניות, שהן קריטיות ביותר בהיבט מניעת אובדן חיים. לשם כך, בוצעה השוואה בין קוד 1 (בסיסי) לבין קוד 2 (משופר), תוך התאמה מדויקת של פרמטרי האימון וביצוע אופטימיזציה מבוקרת.

בסעיף הקודם נעשה שימוש בפרמטרי ברירת מחדל:

* iterations=500 מספר עצים קבוע, ללא חיפוש עומק.
* depth=6 ברירת מחדל – עומק סטנדרטי לעצים.
* learning\_rate=0.03 קצב למידה שמרני.
* class\_weights התאמה ידנית לפי שכיחות הקטגוריות.
* הפיצול ל־Train/Test בוצע תוך שימוש ב-stratify, לשמירה על איזון בין הקבוצות.

לעומת זאת, בסעיף זה בוצע כיוונון מדויק יותר:

* + Iterations  נבחר בין 200–500.
  + Depth  הותאם בטווח 6–10.
  + learning\_rate  נבחר רנדומלית בין 0.03 ל-0.13.
* נעשה שימוש ב־RandomizedSearchCV למציאת שילוב הפרמטרים האופטימלי על בסיס מקסום מדד F1 משוקלל.
* שימור של class\_weights  מאוזנים בין הקטגוריות לכל אורך התהליך.

השוואת התוצאות בין שני הקודים מראה שינוי מתון אך עקבי: בעוד שבקוד הבסיסי המודל השיג Recall של 23% לתאונות קטלניות ו-50% לתאונות קשות, גם לאחר הכיוונון נשמרו ערכים גבוהים אלו, תוך שיפור קל במדדים הכלליים: Accuracy עלה מ־69.3% ל־70%, ומדד F1 משוקלל עלה מ־70.7% ל־71%.

אמנם השינויים במספרים אינם דרמטיים, אך יש בכך כדי להעיד על ייצוב טוב יותר של המודל ועל שיפור ביכולת ההבחנה בין קבוצות החומרה, תוך שמירה על איזון חשוב בין זיהוי קטגוריות נדירות לדיוק כללי. לסיכום, הקוד החדש הביא להמשך שיפור ולחיזוק היציבות של המודל, תוך שמירה על זיהוי מיטבי של תאונות חמורות שהיא המטרה העיקרית בפרויקט.

**מודל 3 - Sklearn GBC (Gradient Boosting Classifier)**

בשלב זה של העבודה, בחרנו להשוות בין שתי גרסאות של מודל Gradient Boosting Classifier במטרה לשפר את יכולת הזיהוי של תאונות חמורות – בפרט תאונות קטלניות, שהן קריטיות ביותר בהיבט מניעת אובדן חיים. לשם כך, בוצעה השוואה בין קוד בסעיף הקודם (בסיסי) לבין הקוד החדש תוך התאמה מדויקת של פרמטרי האימון וביצוע אופטימיזציה מבוקרת.

בסעיף הקודם נעשה שימוש בפרמטרי ברירת מחדל:

* n\_estimators=100 - מספר עצים נמוך יחסית.
* learning\_rate=0.1 - קצב למידה סטנדרטי.
* max\_depth=3 - עומק מינימלי.
* min\_samples\_split=2, min\_samples\_leaf=1 - תנאים רופפים להתפצלות.
* subsample=1.0 - שימוש מלא בכל הנתונים בכל עץ.
* הפיצול ל־Train/Test בוצע תוך שימוש ב-stratify, לשמירה על איזון בין הקבוצות.

לעומת זאת, בסעיף זה בוצע כיוונון מדויק יותר:

* n\_estimators  נבחר בין 250–500.
* learning\_rate  נבחר רנדומלית בטווח 0.05–0.25.
* max\_depth  הותאם בטווח 3–8.
* min\_samples\_split ו־min\_samples\_leaf  כווננו לערכים שמחזקים הכללה ומונעים ריבוי עצים ריקים.
* Subsample  הותאם לערכים בין 0.6–1.0, ליצירת גיוון בין העצים.
* נעשה שימוש ב־RandomizedSearchCV למציאת שילוב הפרמטרים האופטימלי תוך התמקדות מיוחדת במקסום Recall לקטגוריית התאונות הקטלניות.

בנוסף, לצורך שיפור נוסף בזיהוי תאונות קטלניות, הוגדרו משקלים לדגימות (sample\_weight) שהגדילו את חשיבות הקטגוריות החמורות, ובוצע חיפוש ממוקד ששקלל את ה-Recall של קטגוריית הקטלניות בתהליך האופטימיזציה.

השוואת התוצאות בין שני הקודים מראה שיפור משמעותי בזיהוי תאונות קטלניות:  
בעוד שבקוד הבסיסי המודל הצליח לזהות רק כ־4% מהתאונות הקטלניות (Recall=0.04), לאחר הכיוונון זינק ה-Recall ל-15% – שיפור פי 4. גם זיהוי תאונות קשות השתפר מ-33% ל-42%, תוך שמירה על דיוק כולל גבוה של 75.7% ומדד F1 משוקלל של 74.4%.

אמנם חלה ירידה קלה בדיוק הכללי לעומת המודל הבסיסי, אך מדובר בירידה מבוקרת שמטרתה למקד את המודל בזיהוי מיטבי של תאונות חמורות. לסיכום, הקוד החדש מביא לידי ביטוי איזון נכון יותר בין דיוק לבין רגישות למחלקות הקריטיות, והמודל מצליח לזהות בצורה טובה ומדויקת יותר את התאונות הקטלניות והקשות.

**מודל 4 - XGBoost Classifier**

בשלב זה של העבודה, בחרנו להשוות בין שתי גרסאות של מודל XGBoost Classifier במטרה לשפר את יכולת הזיהוי של תאונות חמורות – בפרט תאונות קטלניות וקשות שהן קריטיות ביותר בהיבט מניעת אובדן חיים. לשם כך, בוצעה שוב השוואה בין הקוד בסעיף הקודם לבין הקוד החדש.

בסעיף הקודם נעשה שימוש בפרמטרי ברירת מחדל:

* n\_estimators=100 - מספר עצים סטנדרטי.
* max\_depth=6 - עומק עצים ברירת מחדל.
* learning\_rate=0.3 - קצב למידה גבוה יחסית.
* scale\_pos\_weight=1 - ללא איזון בין קטגוריות.
* objective='multi:softmax' - סיווג מרובה קטגוריות.
* eval\_metric='mlogloss' - מדד הערכת לוג-לוס.

בפיצול הנתונים ל־Train/Test נעשה שימוש ב-stratify, לשמירה על יחס קטגוריות מאוזן בין הסטים.

לעומת זאת, בסעיף זה בוצע שיפור ממוקד:

* הפעלת פרמטר scale\_pos\_weight  ייחודי לכל קטגוריה, כדי להקטין את השפעת חוסר האיזון בדאטה, למשקלים חושבו ערכים לפי יחס הופעה של כל קטגוריה, כך שתאונות קטלניות קיבלו משקל גבוה יותר בלמידה.
* שמירה על אובייקטיב ומדדי הערכה קיימים.
* המטרה: למנוע מהמודל "להעדיף" סיווג תאונות קלות על פני תאונות חמורות.

השוואת התוצאות בין שני הקודים מראה שינוי חשוב מאוד: בעוד שבקוד הבסיסי המודל זיהה רק 4% מהתאונות הקטלניות (Recall=0.04), לאחר הכנסת המשקלים הייעודיים זינק ה-Recall ל-36% – שיפור פי 9. גם זיהוי תאונות קשות השתפר בצורה ניכרת.

אמנם חלה ירידה קלה בדיוק הכללי Accuracy שירד מ-77% ל-67%, אך מדובר בהקרבה מבוקרת, כאשר המטרה בפרויקט זה אינה מקסימום דיוק כולל, אלא מקסום זיהוי של תאונות חמורות – על מנת לאפשר טיפול מונע ופעולה מצילת חיים. לסיכום, הכנסת שקלול קטגוריות במודל XGBoost הובילה לשיפור דרמטי בזיהוי הקטגוריות הקריטיות, ויצרה מודל רגיש וממוקד הרבה יותר למשימת החיזוי הנדרשת בפרויקט.

**מודל 5- Logistic Regression Classifier**

בשלב זה של העבודה, בחרנו להשוות בין שתי גרסאות של מודל Logistic Regression במטרה לשפר את יכולת הזיהוי של תאונות חמורות – בפרט תאונות קטלניות, שהן קריטיות ביותר בהיבט מניעת אובדן חיים. לשם כך, בוצעה השוואה בין קוד 1 (בסיסי) לבין קוד 2 (משופר), תוך התאמה ייחודית של פרמטרי האימון, חישוב משקלים לקטגוריות, וכיוונון אופטימלי ממוקד Recall לקטגוריית הקטלניות.

בסעיף הקודם נעשה שימוש בפרמטרי ברירת מחדל:

* penalty='l2' רגולריזציה סטנדרטית.
* C=1.0 חוזק רגולריזציה סטנדרטי.
* class\_weight=None ללא תיקון לאי־איזון בין מחלקות.
* max\_iter=100 מספר איטרציות דיפולטי.

לעומת זאת, בסעיף זה בוצע שיפור ממוקד:

* חישוב class\_weight  מאוזן לפי שכיחות מחלקות.
* ביצוע RandomizedSearchCV על פרמטרים : C  בין 0.01 ל־10. penalty  בין 'l1' ל־'l2'.
* שימוש ב־solver='saga' התומך בסיווג רב קטגורי ואופטימיזציה גמישה.
* תהליך החיפוש הותאם למיקסום Recall עבור תאונות קטלניות.

תוצאות ההשוואה מראות תמונה ברורה:  
בעוד שבקוד הבסיסי המודל הצליח לשמר דיוק כללי גבוה יותר, הרי שהקוד המשופר, על אף ההתאמות, הציג ירידה בדיוק הכללי Accuracy של 54%, אך זינק באופן דרמטי ביכולת זיהוי התאונות הקטלניות — Recall של 70%,פי 17 שיפור לעומת הבסיסי. בנוסף, בוצעו מספר הרצות חוזרות של המודל במטרה לוודא את יציבות התוצאה, אך לא נצפה שיפור משמעותי מעבר לממצאים שהושגו. התוצאה נשארה עקבית והראתה כי בעוד הזיהוי של תאונות קריטיות השתפר משמעותית, ירדה רמת הדיוק בזיהוי יתר הקטגוריות, בעיקר בתאונות קלות.

לסיכום, המודל המשופר השיג את יעדו – זיהוי מיטבי של תאונות קטלניות, אבל במחיר ירידה בדיוק הכללי.

* 1. **תיאור המודל**

**מודל 1 – Random Forest Classifier**

המודל בגרסתו המשופרת מאפשר להסיק מסקנות משמעותיות, ובראשן – שיפור ניכר בזיהוי תאונות קטלניות, עם Recall של 0.50  לעומת 0.04 בלבד בגרסה הבסיסית. תוצאה זו מחזקת את היכולת היישומית של המודל לצורכי חיזוי מוקדם וניטור בטיחות בדרכים – תוך תגובה טובה לקטגוריות קריטיות. תובנה נוספת שעלתה היא כי העלאת מספר העצים, הגבלת עומק העצים, ושימוש ב־stratify בעת הפיצול שיפרו משמעותית את היציבות והאיזון הבין קטגורי, מבלי להידרש לשיטות קיצון כגון oversampling. כמו כן, השימוש במשקלים מאוזנים (class\_weight='balanced') התגלה כאפקטיבי בהתמודדות עם חוסר האיזון במספר הדוגמאות לכל מחלקה. לא נרשמו בעיות טכניות במהלך ההרצה, וזמן העיבוד של המודל נותר סביר מאוד – גם לאחר העלייה במספר האיטרציות. מבחינת איכות הנתונים – לא התגלו קשיים הקשורים לערכים חסרים, הודות לטיפול מוקדם ויסודי שנעשה בשלב ההכנה. גם מבחינה חישובית, לא נמצאו סטיות או חוסר עקביות. תוצאות המודל היו יציבות, הגיוניות, ועמדו בציפיות מבחינת ביצועים בכל אחת מהמחלקות. לפיכך, מדובר במודל אמין, יציב ובעל יתרון ביצועי ברור על פני גרסתו הבסיסית – המתאים במיוחד לשילוב במערכות חיזוי מבצעיות.

**מודל 2 - CatBoost Classifier**

במודל CatBoost המשופר ניתן להסיק מסקנות משמעותיות, בעיקר בזכות יכולתו לזהות כמות גבוהה יותר של תאונות חמורות – בפרט תאונות קטלניות – ביחס למודלים אחרים שנבדקו. הכיוונון המדויק שנעשה בפרמטרים תרם לייצוב המודל ולשיפור מתון אך עקבי בביצועים, תוך שמירה על Recall גבוה עבור קטגוריות קריטיות, דבר שיש לו ערך יישומי גבוה בפרויקט שמטרתו הצלת חיים. המודל חשף תבניות ברורות של חוסר איזון בנתונים בין קטגוריות התאונה, ותוך כדי כך הציג יכולת למתן את ההשפעה של קטגוריות שכיחות על התחזיות. לא נרשמו בעיות מיוחדות בזמן הריצה – זמן האימון היה סביר בהתחשב במורכבות הפרמטרים ובשימוש ב־RandomizedSearchCV (שיטה אקראית לאיתור שילוב הפרמטרים הטוב ביותר מבין טווחים מוגדרים מראש), וכל התהליך התבצע בצורה חלקה. המודל לא סבל מבעיות איכות נתונים או חוסרים, מאחר והנתונים עברו ניקוי מקדים. כמו כן, לא אותרו אי־עקביות חישוביות או בעיות לוגיות בתהליך האימון והבדיקה, אבל זמן הריצה של הקוד הוכפל בערך פי 4. השיפור במדדי הביצועים, גם אם אינו דרמטי, מצביע על כך שהמודל בגרסתו המשופרת משלב איזון טוב יותר בין דיוק, רגישות, ויכולת הכללה – ולכן טוב לשימוש.

**מודל 3 - Sklearn GBC (Gradient Boosting Classifier)**

בהתבסס על תוצאות מודל GBC לפני ואחרי אופטימיזציה, ניתן להסיק מסקנות ברורות ומשמעותיות. המודל הצליח להציג שיפור בזיהוי תאונות חמורות, ובמיוחד בזיהוי תאונות קטלניות – אחד היעדים המרכזיים של הפרויקט. בגרסה המשופרת ניכר שיפור מהותי במדד ה־Recall עבור קטגוריה זו, לצד שיפור גם ב־Recall עבור תאונות קשות, דבר המלמד על שיפור ביכולת ההבחנה של המודל בין דרגות החומרה. לא נמצאו דפוסים חריגים או תובנות בלתי צפויות, אך בהחלט בלטה השפעת השקילה הממוקדת לזיהוי מקרים חמורים. זמן הריצה של המודל עלה ביחס לקודמו ולא נרשמו בעיות ביצוע חריגות במהלך האימון או התחזיות. איכות הנתונים הייתה טובה ולא נתקלו בבעיות של ערכים חסרים – כל המשתנים כבר עברו תהליך עיבוד מוקדם. בנוסף, לא נצפו אי־עקביות חישוביות בין דיווחי המטריצות לדוחות הסיווג. לפיכך, ניתן לומר שהמודל המשופר מספק תשתית מדויקת יותר לניבוי חומרת תאונות – תוך שמירה על יציבות ואמינות.

**מודל 4 - XGBoost Classifier**

מודל XGBoost בגרסתו המשופרת מאפשר להסיק מסקנות משמעותיות, בעיקר בכל הקשור לשיפור בזיהוי תאונות חמורות – בדגש על קטלניות. הכנסת פרמטר השקילה scale\_pos\_weight (פרמטר שמאפשר להקצות משקל גבוה יותר לדגימות ממחלקות לא מאוזנות, וכך להפחית את ההטיה לטובת הקטגוריות השכיחות) הובילה לשינוי דרמטי ביכולת ההבחנה של המודל בין קטגוריות, דבר שלא בא לידי ביטוי בגרסת הבסיס. המודל הצליח לזהות 36% מהתאונות הקטלניות (לעומת 4% בלבד קודם), מה שמעיד על שינוי איכותי בפרקטיקה של חיזוי סיכונים. תובנה חשובה שנחשפה היא החשיבות של איזון בין קטגוריות במודלים עם התפלגות לא אחידה – המודל חשף את ההטיה החזקה של הדאטה כלפי תאונות קלות, והגיב לה באופן מדויק באמצעות המשקלים. לא נרשמו תקלות בתהליך הריצה, וזמן החישוב נותר סביר הודות לשמירה על מספר עצים קבוע וללא חיפוש אוטומטי רחב (כמו ב־RandomizedSearchCV). איכות הנתונים לא היוותה מגבלה – כלל התצפיות הוכנסו לאחר ניקוי, והמודל פעל על דאטה מהימן וללא ערכים חסרים. לא זוהו בעיות חישוב או חוסר עקביות בתהליך, והפלט היה עקבי ונכון. לסיכום, גרסת המודל המשופרת של XGBoost מראה שיפור משמעותי ברגישות הקטגוריאלית תוך הקרבה מינימלית בדיוק, ומהווה פתרון מדויק ואפקטיבי ליישום פרויקט חיזוי תאונות חמורות.

**מודל 5- Logistic Regression Classifier**

בהתבסס על תוצאות מודל Logistic Regression בגרסתו הבסיסית, ניתן להסיק מסקנות ברורות לגבי מגבלות המודל בהקשרים של סיווג לא מאוזן. אף על פי שהמודל השיג דיוק כללי גבוה של 76.1% והציג ביצועים מרשימים בזיהוי תאונות קלות (Recall של 94% ו־Precision של 0.85), הוא כשל בזיהוי מוחלט של תאונות קטלניות (Recall של 0%) והציג ביצועים בינוניים בלבד עבור תאונות קשות (Recall של 33%). ניתוח התפלגות התחזיות ממחיש היטב את הבעיה: המודל נטה כמעט באופן בלעדי לנבא "תאונות קלות", תוך התעלמות מהקטגוריות החמורות יותר. מצב זה מצביע על רגישות נמוכה מאוד של רגרסיה לוגיסטית במצבים של חוסר איזון חמור בין הקבוצות, ובמיוחד כאשר אין שימוש במנגנוני תיקון כגון שקילת קטגוריות או שיטות oversampling. עם זאת, יתרונו של המודל טמון בפשטותו ובמהירות הביצוע שלו – לא נרשמו בעיות עיבוד, זמן הריצה היה מהיר, והנתונים עברו עיבוד מקדים מלא ללא ערכים חסרים או תקלות חישוב. בשל העובדה שניסיונות השיפור לא הניבו תוצאות טובות יותר, הוחלט להשאיר את גרסת הבסיס כנקודת ייחוס, תוך הבנה ברורה שהיא אינה מתאימה למשימות שבהן נדרש דגש על זיהוי אירועים נדירים וקריטיים. המודל מתאים לבעיות מאוזנות יחסית, אך אינו מספק מענה מספק לפרויקט הנוכחי שבו הדגש הוא על גילוי מוקדם של תאונות חמורות.

1. **הערכת המודלים**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| **Random Forest (improved)** | 0.7159 | 0.765 | 0.7159 | 0.7237 |
| **CatBoost (improved)** | 0.7012 | 0.7258 | 0.7012 | 0.7125 |
| **GBC (improved)** | 0.7574 | 0.7412 | 0.7574 | 0.7444 |
| **XGBoost (improved)** | 0.6668 | 0.7412 | 0.7574 | 0.7444 |
| **Logistic Regression (base)** | 0.7611 | 0.717 | 0.7611 | 0.7211 |

לצורך בחירת המודל הסופי לפרויקט, להלן טבלת השוואה מרכזת בין חמשת המודלים שנבדקו, בהתבסס על שלושה מדדים מרכזיים: Recall כללי, מדד F1 משוקלל, ו־Recall לקטגוריית התאונות הקטלניות (Catastrophic Recall). השוואה זו מספקת בסיס אובייקטיבי להערכת הביצועים של כל מודל ביחס למטרות הפרויקט.

במטרה לאתר את המודל היעיל והמדויק ביותר לצורך חיזוי חומרת תאונות הדרכים, נערכה השוואה שיטתית בין חמשת המודלים שנבנו והורצו במסגרת הפרויקט. ההשוואה התבצעה על בסיס ארבעה מדדי ביצוע מרכזיים: Accuracy, Precision, Recall ו־F1 Score ובפרט נבחנה רגישות המודל לזיהוי תאונות קטלניות וקשות, מתוך הכרה בכך שמדובר במרכיב הקריטי להצלחת הפרויקט, שמטרתו הפחתת תאונות חמורות בישראל וזיהוי מוקדים בעייתיים.

מודל Logistic Regression הציג את רמת הדיוק הכוללת הגבוהה ביותר (Accuracy = 0.7611), ו־F1 משוקלל של 0.7211, אך כשל באופן מובהק בזיהוי המקרים הקריטיים: Recall של 0.00 עבור תאונות קטלניות, משמע אפס זיהוי בפועל של תאונות מסוג זה. המשמעות היא שמודל זה אמנם מסווג היטב את הרוב (תאונות קלות), אך אינו מסוגל להתמודד עם מקרי קצה מסוכנים – מה שהופך אותו לבלתי מתאים למטרה המעשית של הפרויקט.

מודל CatBoost שיפר משמעותית את יכולת האיזון הכללית והגיע ל־Recall של 0.7012 ו־F1 של 0.7125, ובפרט הצליח לזהות כ־23% מהתאונות הקטלניות. זהו שיפור ניכר לעומת Logistic Regression אך עדיין נמוך בהשוואה למודלים מתקדמים יותר. בנוסף, המודל נטה עדיין להעדיף קטגוריות שכיחות כמו "קלה", ויצר בלבול בינוני בין מחלקות.

מודל GBC הציג תוצאות מצוינות ב־Accuracy (0.7574) וב־F1 (0.7444) ואף הצטיין בזיהוי תאונות קלות (Recall = 0.94), אך שוב – זיהוי תאונות קטלניות נותר חלש מאוד (Recall = 0.15). כלומר, גם מודל זה שמר על איזון, אך לא הצליח לשים את הדגש הנדרש על קטגוריות מסוכנות.

מודל XGBoost, לאחר שקילה של הקטגוריות באמצעות הפרמטר scale\_pos\_weight, שיפר את ביצועיו באופן ניכר בזיהוי של התאונות הקטלניות (Recall = 0.36), אך מחיר השיפור היה ירידה משמעותית ב־Accuracy הכללי ל־0.6668. כלומר, המודל הפך "ממוקד מדי" בקטגוריה אחת, ואיבד מהכלליות – מה שעשוי להוביל לריבוי אזעקות שווא במצבים לא מסוכנים.

לעומת כל אלה, מודל Random Forest המשופר סיפק את התוצאה המאוזנת והאפקטיבית ביותר לפרויקט זה. הוא השיג Recall כללי של 0.7159 ו־F1 Score של 0.7237, אך גולת הכותרת היא יכולת זיהוי של 64% מהתאונות הקטלניות – הפער הגדול ביותר מבין כל המודלים. בנוסף, הוא שמר על יציבות טובה בזיהוי תאונות קשות וקלות כאחד, תוך הפחתת בלבול בין הקטגוריות.

בחירה במודל **Random Forest** תואמת את מטרת הפרויקט — שיפור החיזוי של מוקדי תאונות חמורות לצורך קבלת החלטות מונעת, ובפרט, זיהוי מדויק של תאונות קטלניות שמהוות סכנה ממשית לחיי אדם. המודל מאפשר יצירת מערכת התרעה מוקדמת חכמה שתדע להפנות משאבים, לפקח ולהתריע על אזורים בעייתיים ולהפחית תאונות קשות ב־15% לפחות — יעד הפרויקט.

לסיכום, על אף שמודלים כמו GBC ו-XGBoost הציגו ביצועים כלליים גבוהים, מודל Random Forest המשופר נבחר כמודל הטוב והיעיל ביותר למימוש מטרות הפרויקט שלנו בשל שילוב נדיר של יציבות כללית ורגישות גבוהה לקטגוריות הקריטיות ביותר — קטלניות וקשות. מודל זה מספק את התרומה הישירה והמעשית ביותר להצלחת המערכת והפרויקט.