

**SafeRoads**

**Business Understanding Report**

**רום קדוש – 207691841**

**סנדרה זייגרמכר -206987521**

**תוכן עניינים**

1. **רקע ............................................................................. 3-8**
2. **יעדים עסקיים וקריטריונים להצלחה ................................... 8-10**
3. **הערכת מצב .....................................................................11**
4. **מלאי משאבים ...............................................................12-14**
5. **דרישות, הנחות ואילוצים .................................................14-17**
6. **סיכונים ומצבים ................................................................18**
7. **טרמינולוגיה ..................................................................18-19**
8. **עלויות ויתרונות .............................................................19-20**
9. **מטרות מדעי הנתונים וקריטריונים להצלחה ........................20-23**
10. **תוכנית הפרויקט ...........................................................23-25**
11. **כלים וטכניקות ................................................................25**
12. **רקע**
    1. **קביעת יעדים עסקיים**

בעצם, עיקר המטרה של המערכת שלנו היא לצמצם את תאונות הדרכים בישראל. אנחנו שואפות לבנות מודל אשר ידע לזהות מוקדי תאונות דרכים. המערכת תתבסס על נתונים ממאגרים ציבוריים וכוללת מידע על מיקום, זמן, סוג רכב, תנאי הדרך, ומספר נפגעים. בעזרת ניתוח הנתונים המערכת תוכל לזהות מוקדים מסוכנים ולהציע פתרונות לשיפור הבטיחות, כמו תיקון תשתיות, הוספת תמרורים, ושיפור צמתים. המטרה היא לספק כלים למוסדות אחראיים כמו הרשות לבטיחות בדרכים, משרד התחבורה והרשויות המקומיות, כדי שיבצעו פעולות לשיפור הבטיחות בדרכים ולהפחית את מספר התאונות. נכון להיום, לא ידוע לנו על פלטפורמה כזו שמסייעת באיתור מוקדי סיכון, ובעצם האנליזות שמסופקות ברשתות לא מהוות אינדיקציה מספיק טובה לסיבות מהן נגרמות מרבית תאונות הדרכים. השאיפה שלנו להפוך את המערכת הזו לכלי עזר חיוני אשר כל אחד ממשתמשי הקצה יוכל להשתמש בה בקלות, ביעילות ולהפיק ממנה תובנות מצילות חיים. הנתונים עליהם נתבסס הינם נתונים ממשלתיים אשר מפורסמים בשתי אתרים עיקריים, data gov והלשכה המרכזית לסטטיסטיקה, משמע הם מבוססים על מקור מהימן ועם שיתוף פעולה של כוחות ההצלה (משטרה, מד״א).

* 1. **רקע עסקי**

על פי מחקרים והערכות עולמיות בנושא, תאונות הדרכים מהוות כיום את סיבת המוות המרכזית בקרב אוכלוסיות צעירות בגילאי 5-29. ביחס לכלל האוכלוסייה זוהי סיבת המוות התשיעית בחומרתה, שצפויה לגדול עד כדי סיבת המוות החמישית בחומרתה עד לשנת 2030. אם מגמה זו אכן תתברר כנכונה ולא תבוצע מניעה אגרסיבית כלפי הגורמים העיקריים לתאונות הדרכים, ייהרגו כבר בעשור הבא יותר אנשים בהתנגשות מכונית מאשר ממחלות האיידס והשחפת גם יחד. אחד הגורמים לעלייה בכמות ההרוגים והפצועים במצב קשה היא שימוש מוגבר מאוד במכוניות, במדינות שאזרחיהן בעלי הכנסה בינונית אך שתשתיות הדרכים שלהן עדיין יחסית מיושנות. הנתונים הללו תואמים את המתרחש בישראל באופן עקרוני, אם כי שילוב בין תכנון עירוני מוכוון לתחבורה ציבורית, אמצעי בטיחות מתקדמים ומיתון התנועה צפויים להביא לצמצום מהותי. התאונות גורמות לפגיעות רבות בנפש ורכוש, והן גם נושאות עלויות כלכליות כבדות.

**משאבים זמינים**

**כוח אדם:** מדעני נתונים אשר יבצעו ניתוח נתונים ופיתוח מודלים חיזויים, בעלי מיומנויות כמו Python, למידת מכונה, Big Data. **מפתחים ומתכנתים:** לפיתוח המערכת והממשקים בעלי מיומנויות כמו JavaScript, Python, אינטגרציה עם APIs. **אנשי תפעול** לתחזוקת המערכת ושילובה עם מערכות קיימות. כוח אדם מקצועי וחומרה מתקדמת מבטיחים עיבוד נתונים מדויק, פיתוח מודלים חיזויים, ומתן המלצות יישומיות לשיפור הבטיחות בדרכים.

**בעיות:**

* הנתונים הגולמיים המגיעים ממקורות שונים עשויים לכלול ערכים לא תקינים, ערכים חסרים, או מידע לא אחיד, לכן נדרשת עבודה רבה על ניקוי והכנה של הנתונים, כולל המרה לערכים מספריים או קטגוריים בהתאם לדרישות המודלים. תיתכן הקצאת נתונים לא נכונה לקטגוריות, דבר שעלול להוביל לפגיעה בביצועי המודל.
* עלולים לצוץ פערים אשר אנחנו לא בעלות השפעה עליו כמו תשתיות הכבישים אשר משתנות באופן דינמי – פרויקטים לשיפור כבישים, שינוי צמתים, הוספת נתיבים או שינוי תמרורים – עשויים להשפיע על תחזיות המודל. שינויים אלו עלולים לגרום לכך שהנתונים הקיימים לא יהיו רלוונטיים לתחזיות עתידיות.
* שינויי מזג אוויר ותנאי דרך: נתונים על תאונות דרכים כוללים גם פרטים על תנאי מזג האוויר בעת התאונה. חוסר מידע מדויק או השתנות מזג האוויר בצורה קיצונית עלולים להשפיע על המודל ולהקטין את דיוק התחזיות בהתאם.
* זיהוי מוקדי סיכון בהתבסס על המודל ללא תובנות פרקטיות לשיפור. לדוגמה: ייתכנו מקרים בהם מוקד סיכון נמצא באזור שאינו מאפשר שינוי תשתיתי משמעותי או שידרוש עלויות גבוהות מאוד.

**מטרות:**

* הנחלת המערכת ברמה ארצית עבור כל בעלי העניין המטרה העיקרית של הפרויקט היא להטמיע את מערכת SafeRoads בכל רחבי הארץ, כך שתהפוך לכלי מרכזי ומשמעותי עבור כלל בעלי העניין המעורבים בתחום התחבורה והבטיחות בדרכים. המערכת תוכל לשמש את הרשויות המקומיות, משרד התחבורה, הרשות הלאומית לבטיחות בדרכים, וגם גופים פרטיים כמו חברות ביטוח. באמצעות הכלים המתקדמים של המערכת, כל אחד מבעלי העניין יוכל לקבל מידע מדויק, להבין את מוקדי הסיכון בדרכים שבאחריותו, ולבצע פעולות יזומות לשיפור הבטיחות ולהפחתת הסיכון לתאונות.
* לטווח הרחוק: לראות שינוי ושיפור במגמת תאונות הדרכים בארץ: המטרה היא לראות שיפור אמיתי וסטטיסטי במגמת תאונות הדרכים בישראל. על ידי שימוש במערכת, אנו שואפות להפחית את מספר התאונות באופן משמעותי ולהפוך את הכבישים לבטוחים יותר. המערכת תנתח את מוקדי הסיכון, תספק המלצות לפעולה ותסייע לגורמים השונים לפעול באופן יעיל וממוקד במטרה למנוע תאונות וגם לצמצם את חומרתן, ובכך להפחית את מספר הנפגעים וההרוגים בכבישים. מדובר בתהליך הדרגתי ומתמשך של שיפור שמבוסס על למידה מתמדת וניתוח נתונים, שיאפשר למדינה להיות עם כבישים בטוחים יותר לאורך השנים.
* **התאמת המודל למדינות אחרות ברחבי העולם** המטרה הבאה של SafeRoads היא להפוך את המודל לישים גם במדינות אחרות ברחבי העולם. תאונות דרכים הן אתגר גלובלי, והטכנולוגיה שפותחה בפרויקט יכולה להוות פתרון למדינות רבות נוספות שמתמודדות עם בעיות דומות. אנו שואפים לפתח את המערכת כך שתוכל להיות מותאמת לאזורים שונים בעולם תוך התאמה לצרכים ולמאפיינים הייחודיים של כל מדינה.

**משימה 1 – קביעת המבנה הארגוני**

מנהל הפרויקט (Project Manager)

* נמצא בראש המבנה הארגוני, אחראי על ניהול הפרויקט כולו ותיאום בין הצוותים השונים.

צוות מדעי הנתונים (Data Science Team)

* תת-קבוצה שמתמקדת בניתוח נתונים וחיזויים, תחתיה יש:
  + אנליסטים עסקיים (Business Analysts): אחראים לניתוח צרכים עסקיים ותובנות.
  + מדעני נתונים (Data Scientists): עוסקים בפיתוח מודלים לחיזוי ובניתוח מתקדם של נתונים.

צוות הנדסה (Engineering Team)

* מתמקד בפיתוח המערכת ובתמיכה הטכנולוגית, תחתיו יש:
  + מפתחים (Developers): אחראים על פיתוח התוכנה והמערכת.
  + מנהלי IT (IT Administrators): אחראים על תחזוקת המערכות ותשתיות ה-IT.
  + בודקי איכות (QA Testers): מבצעים בדיקות על מנת להבטיח את איכות המערכת.

שיווק וקשרי לקוחות (Marketing and Customer Relations)

* מתמקדים בקידום המערכת ובהתקשרות עם לקוחות וגורמים חיצוניים.

**משימה 2 – תיאור אזור הבעיה**

אזור הבעיה העיקרי של הפרויקט הוא תחום הבטיחות בדרכים. חשוב להדגיש כי פרויקט SafeRoads הוא מיזם חדש, ואין כל מערכת קיימת הדומה לו כיום. המשמעות היא שכל הפיתוח נעשה מאפס, ללא כל בסיס או מודל קיים שניתן להסתמך עליו. האתגר המרכזי הוא יצירת מערכת שתוכל להפיק תובנות מתוך נתונים גולמיים שאינם מעובדים. הנתונים נאספים ממקורות שונים כמו דו"חות תאונות, תנאי דרך, ומאפייני רכבים, ונדרש לבצע ניקוי, הכנה, וניתוח מתקדם של נתונים אלו כדי להגיע לתובנות מועילות. כיוון שמדובר בפרויקט חדש, האתגר הוא לא רק לפתח את המערכת אלא גם להטמיע אותה בקרב גורמים שונים במערכת הציבורית. כיום, אין שימוש נרחב במדעי הנתונים בתחום הבטיחות בדרכים, ולכן נדרש ליצור אמון בקרב מקבלי ההחלטות ולהראות את הערך המוסף של המערכת. יהיה צורך בהצגת המערכת והיתרונות שלה באמצעות מצגות, הדגמות, וחומרי הסברה, כדי לוודא שהמערכת תתקבל ותהיה בשימוש על ידי הגורמים הרלוונטיים, כמו רשויות מקומיות, משרד התחבורה והרשות הלאומית לבטיחות בדרכים.

**משימה 3 - הפתרון הנוכחי**

* דיווח ותיעוד תאונות באופן היסטורי: הפתרון הנוכחי כולל תיעוד תאונות דרכים על ידי המשטרה ורשויות התחבורה. נתוני תאונות נאספים ומתועדים בקבצים סטטיסטיים, כמו אלו המפורסמים על ידי הלשכה המרכזית לסטטיסטיקה, ומאגרים כמו דאטה גוב.
* אכיפה משטרתית מוגברת: הרשויות עושות שימוש באכיפה משטרתית מוגברת באזורים שנחשבים למסוכנים על פי נתוני תאונות עבר. אכיפה זו כוללת שימוש במצלמות מהירות, תמרורים, ונוכחות משטרתית להגברת ההרתעה.
* פרויקטים לשיפור תשתיות באופן נקודתי: נעשים פרויקטים מקומיים לשיפור תשתיות כבישים, כגון הרחבת נתיבים, הוספת מעברי חצייה ותאורה, בהתאם לדיווחים חוזרים על בעיות תשתית או אזורים עם עומסי תנועה מסוכנים.
* קמפיינים ציבוריים להגברת המודעות: בנוסף, ישנם קמפיינים ציבוריים להגברת המודעות לסכנות הנהיגה ולחשיבות הנהיגה הזהירה.

**יתרונות הפתרון הנוכחי:**

* תיעוד ואיסוף נתונים: הנתונים על תאונות נאספים בצורה מסודרת לאורך שנים ומאפשרים הבנה כללית של מוקדי תאונות ותדירותן.
* פעולות אכיפה ושיפור תשתיות: האכיפה המשטרתית והפרויקטים לשיפור תשתיות הביאו לשיפורים נקודתיים באזורים מסוימים, והוספת תמרורים או תאורה במוקדים מסוכנים הובילה לשיפור הבטיחות במקומות אלו.
* הגברת מודעות הציבור: קמפיינים ציבוריים מסייעים בהגברת המודעות לסכנות הנהיגה ומעודדים נהגים לנקוט באמצעי זהירות.

**חסרונות הפתרון הנוכחי:**

* חוסר חיזוי ופוקוס על מוקדים מסוכנים: הפתרון הנוכחי מתבסס על נתונים היסטוריים בלבד, ולא כולל יכולת חיזוי מוקדי תאונות עתידיים. דבר זה מקשה על הרשויות לפעול מראש ולמנוע תאונות, אלא מגיב רק לאחר שמתרחשות תאונות חוזרות במקום מסוים.
* פעולות שיפור מוגבלות והיעדר תכנון ארוך טווח: הפעולות שנעשות הן לרוב נקודתיות ולא מתוכננות בצורה הוליסטית. אין תכנון ארוך טווח שמבוסס על נתונים מתקדמים או על חיזוי של מוקדי סיכון פוטנציאליים.
* מחסור בניתוח מתקדם: הנתונים שנאספים אינם מנותחים בצורה מספקת כדי לייצר תובנות מעמיקות ואסטרטגיות להתמודדות עם הסיכונים. השימוש בנתונים מוגבל לדיווחים סטטיסטיים בסיסיים.
* הסתמכות על אכיפה בלבד: האכיפה היא מרכיב חשוב, אך הסתמכות מוגזמת עליה ללא שיפור כולל בתשתיות, עלולה להוביל רק להפחתת זמנית של התאונות ולא למיגור הבעיה לטווח הארוך.

**רמת הקבלה של הפתרון בתוך הארגון:**

הפתרון הנוכחי לא עונה על הציפיות כיוון שהוא מבוסס על תהליכים קיימים ומוכרים, ויש צורך בשינוי משמעותי בתפיסות או בתקציבים על מנת למגר את התופעה באופן מוחלט, ולא לטפל בה באופן נקודתי. בנוסף, גופים רבים מכירים בכך שהפתרון הקיים מוגבל ביכולתו להביא לשינוי רחב ומשמעותי, ולכן יש פתיחות לשימוש בטכנולוגיות חדשות ובמדעי הנתונים כדי להרחיב את היכולת לחזות ולטפל במוקדי תאונות בצורה פרואקטיבית.

הפרויקט "SafeRoads" נועד לענות על החסרונות הללו, ולספק מערכת מתקדמת שתוכל להתמודד בצורה טובה יותר עם האתגרים הקיימים בשיפור הבטיחות בדרכים.

**2. יעדים עסקיים וקריטריונים להצלחה**

2.1**. מטרות עסקיות**

המטרה העיקרית של הפרויקט שהיא הבסיס והכיוון המנחה לאורך תהליך העבודה כולו היא להפחית את מספר תאונות הדרכים בישראל ואת מספר הנפגעים. את המטרה שלנו נשיג באמצעות זיהוי ומיפוי מוקדי תאונות בעלי רמת סיכון גבוהה, בהתבסס על ניתוח נתוני תאונות היסטוריים ותנאי דרך. ניתוח זה יסייע לרשויות לתעדף פעולות לשיפור התשתיות והאכיפה באזורים אלו. מטרת העל היא להפחית את מספר התאונות ב-15% תוך שנתיים מרגע ההטמעה של המערכת.

**תיאור הבעיה שברצוננו לפתור:**

הבעיה המרכזית שהפרויקט "SafeRoads" שואף לפתור היא הפחתת תאונות הדרכים בישראל על ידי זיהוי מוקדי סיכון, חיזוי תאונות עתידיות, ומתן המלצות לשיפור תשתיות בטיחות. באמצעות ניתוח נתונים היסטוריים של תאונות, הפרויקט ישפר את היכולת לזהות דפוסים חוזרים ותנאים שמובילים לתאונות, ויאפשר למקבלי ההחלטות לפעול במקומות הנכונים על מנת לשפר את תחום הבטיחות בדרכים.

**ציון השאלות העסקיות:**

1. מה הם המקומות בישראל בהם יש סיכון מוגבר לתאונות דרכים על פי הנתונים הקיימים?
2. אילו גורמים עיקריים משפיעים על הסיכון לתאונות בכל אחד מהמוקדים הללו (למשל, תנאי דרך, סוגי רכבים, עומס תנועה)?
3. מהן הפעולות המומלצות לשיפור תשתיות ואכיפה במוקדים המסוכנים, כך שניתן יהיה להפחית את הסיכון לתאונות?
4. איך ניתן להגדיל את היעילות של תקציב האכיפה והשיפורים בתשתיות, ולהפנות משאבים לאזורים המועדים ביותר לסכנה?

**קביעת דרישות עסקיות אחרות:**

* המערכת חייבת לספק תובנות שניתן ליישם בצורה פרקטית, כלומר חשוב להבטיח שהנתונים המסופקים יהיו מדויקים, ושהמערכת תיתן המלצות ברמת אמון גבוהה כך שמקבלי ההחלטות יוכלו לפעול על פיהן בביטחון.
* המערכת צריכה להיות נגישה ומשתמשים מכל הרשויות השונות (משרד התחבורה, רשויות מקומיות, חברות ביטוח) צריכים להיות מסוגלים להבין ולנתח את התובנות המתקבלות כדי לקבל החלטות מושכלות, את התובנות נציג בצורה אינטראקטיבית וקלה על מנת לספק חווית משתמש מקסימלית.

**יתרונות צפויים:**

* הפחתת מספר תאונות הדרכים: צמצום מספר התאונות ב-15% בתוך שנתיים יוביל להפחתת הנפגעים בגוף ובנפש.
* חיסכון כלכלי לחברות הביטוח: על ידי התאמת הפרמיה בהתאם לרמת הסיכון האזורי, ניתן להוזיל את עלויות הביטוח לנהגים באזורים בטוחים יותר ולהפך. חברות הביטוח יוכלו לבצע התאמות עסקיות על ידי שימוש במערכת.
* ייעול השימוש במשאבי אכיפה: שימוש במידע מבוסס נתונים יאפשר הפניית אמצעי אכיפה לאזורים המועדים ביותר לסכנה, וכתוצאה מכך ניצול טוב יותר של תקציב האכיפה ושיפור הבטיחות באזורים המועדים לפורענות.

**2.2. קריטריוני הצלחה**

**מבחינה אובייקטיבית:**

|  |  |
| --- | --- |
| **יעד עסקי** | **קריטריון הצלחה** |
| הפחתת תאונות במוקדי סיכון. | ירידה של לפחות 15% במספר תאונות הדרכים במוקדים שזוהו על ידי המערכת, במהלך שנתיים מרגע הטמעת המערכת לאחר ניתוח הנתונים. |
| זיהוי מוקדי סיכון עתידיים בצורה מדויקת. | השגת רמת דיוק של לפחות 85% בתחזיות חיזוי מוקדי הסיכון. |
| שיפור תשתיות בטיחות בדרכים באזורים המוגדרים כמסוכנים. | ביצוע לפחות 20 פרויקטים לשיפור תשתיות בהתאם להמלצות המערכת בתוך ארבע שנים. |

**מבחינה סובייקטיבית:**

|  |  |
| --- | --- |
| **יעד עסקי** | **קריטריון הצלחה** |
| להבטיח שמקבלי ההחלטות במשרד התחבורה, ברשויות המקומיות, וברשות לבטיחות בדרכים מרוצים מהמערכת ומאמינים באיכות התחזיות: | דיווח על שביעות רצון גבוהה מצד מקבלי ההחלטות (נמדד באמצעות שאלונים או דיווחים תקופתיים) הכולל התייחסות לאיכות התחזיות וליישום ההמלצות המערכתיות. |
| העלאת מודעות הציבור לסכנות בדרכים ושיפור הזהירות האישית. | קבלת משוב חיובי מהציבור דרך סקרים או תגובות למדיה חברתית על השימוש במערכת והמודעות לסיכונים. |
| הבטחת שיתוף פעולה פורה בין משרד התחבורה והרשויות המקומיות. | מספר הפגישות המשותפות, היקף היישום של המלצות המערכת שנעשה בשיתוף פעולה בין הגופים השונים, והשתלבות תובנות המערכת בתהליכי קבלת החלטות. |

**3. הערכת מצב**

את נתוני תאונות דרכים לקחנו ממאגרי דאטה גוב, הנתונים האלו מכילים אינפורמציה על מיקום התאונה, סוג הרכב המעורב, מספר נפגעים, תנאי מזג האוויר, זמן התאונה ועוד. בנוסף יש לנו גישה לנתונים מרשות הבטיחות בדרכים, שם יש ניתוחים סטטיסטיים על תאונות, פעולות אכיפה, מסעות הסברה, תשתיות בטיחות ומאפיינים דמוגרפיים באזורים שונים, אלו מהווים את מכלול הנתונים הרלוונטיים לניתוח ולפילוח.

כוח האדם הדרוש לנו כולל צוות מקצועי בתחומים שונים כמו:

* מדעני נתונים: אחראים על ניתוח הנתונים, פיתוח מודלים לחיזוי, ויצירת תובנות מבוססות נתונים.
* מהנדסי תשתיות תחבורה: אחראים על בדיקת התאמות תשתיתיות וליישום ההמלצות בשטח.
* מתכנתים ומפתחים: אחראים על פיתוח המערכת עצמה, בניית ממשקים והטמעת המודלים.

גורמי הסיכון העיקריים אשר עלולים לפגוע בתהליך בניית המערכת נובעים בעיקר מנתונים באיכות נמוכה או לא מספקים. בחלק מהנתונים עלול להיות חוסרים (כמו ערכי null) או חוסר דיוק, דבר שעלול לפגוע באיכות המודלים ובתחזיות המתקבלות. בנוסף, חוסר שיתוף פעולה מצד בעלי העניין ביניהם - רשויות מקומיות, משרד התחבורה, וחברות עלול להוביל לאי התאמה מבחינת דרישות תפעוליות ופונקציונליות, מה שעלול להאריך את זמן יישום והטמעת המערכת ובזבוז משאבים. מורכבויות טכניות כמו שילוב של נתונים ממקורות שונים ועיבוד כמות נתונים גדולה מאוד דורש משאבי מחשוב גבוהים ויכולות טכניות מורכבות, לכן יש להעריך את זמני העבודה בהתאם. במקרה וניתקל באחד מגורמי הסיכון הנ״ל נבצע התאמה לנתונים באמצעות טכניקות שונות, כמו קבלת נתונים משלימים ממקורות אחרים או שימוש בטכניקות של השלמת נתונים חסרים. במקרה של חוסר שיתוף פעולה נדאג לבצע הסברה אל מול הגורמים השונים, כולל מצגות שמסבירות את היתרונות של המערכת והשפעתה על הבטיחות בדרכים. כמו כן, נדאג למנות בעל תפקיד אשר ידאג לטפח את קשרי הלקוחות ותחום אחריותו יכלול את שיתוף הפעולה עם הגורמים השונים. מבחינה טכנית אנו שואפות להשתמש במשאבי מחשוב ענן על מנת להתמודד עם עומסי נתונים. כמו כן, שימוש בטכנולוגיות מתקדמות לעיבוד נתונים מקבילי, שמקטינות את העומס על המערכת, כך אנחנו גם נדאג לבקרת עומסים וגם נהיה מתואמות מבחינת חדשנות וקידמה טכנולוגית.

**4. מלאי משאבים**

**משימה 1 - משאבי חומרה, אחסון, רשת ואבטחת מידע**

**משאבי חומרה**

יש צורך בשרתים חזקים או בשימוש בתשתיות ענן לצורך עיבוד כמויות גדולות של נתונים. העיבוד יכלול ניתוח נתוני תאונות, ניתוח דפוסים, וחישובי למידת מכונה. שימוש במחשוב ענן כמו AWS, Google Cloud, או Azure, אשר יאפשר גמישות ועיבוד נתונים מהיר, ובמיוחד עבור מודלים לחיזוי שדורשים כוח עיבוד משמעותי, כך גם נוכל להשתמש בשירותי הגיבוי של אותן חברות ונוכל לבצע ניהולי גרסאות. גם נדרש שימוש במחשבים חזקים עבור צוותי הפיתוח והחוקרים, אשר יאפשרו ביצוע ניתוחים, הרצת מודלים וכתיבת קוד בצורה הממוטבת ביותר מחשבים עם מעבדים חזקים ו-GPU יהיו חיוניים להרצה מהירה של מודלים מבוססי למידת מכונה.

**משאבי אחסון**

נדרש אחסון גדול עבור כמויות הנתונים שמגיעות ממקורות שונים (נתוני תאונות, עומסי תנועה, תנאי דרך, וכדומה). רוב ספקי האחסון מגבילים את כמות האחסון, כיוון שאנחנו מתבססות על נתוני אמת אשר מתעדכנים מעת לעת אנחנו נשתמש בפתרונות אחסון בענן שמאפשרים גידול דינמי, גמישות, עיבוד מהיר, ואחסון בלתי מוגבל של נתונים, לצד אפשרויות גיבוי מאובטח של נפח האחסון בהתאם לצרכים המשתנים.

**רשת ואבטחת מידע:**

יש לדאוג לתשתיות רשת מהירות ובטוחות, כולל הגנות סייבר ואבטחת מידע, על מנת להבטיח גישה חלקה לנתונים ושמירה על פרטיות ואבטחתם.

יש להקים חומת אש (Firewall) ומנגנוני הצפנה כדי להבטיח שהנתונים לא ייפגעו ושלא תהיה גישה לא מורשית ובנוסף תמיכה בפרוטוקולי הצפנה מתקדמים (SSL/TLS) על מנת לשמור על שלמות הנתונים.

**משימה 2 - זיהוי מקורות נתונים ומאגרי ידע**

ֿהפרויקט מבוסס על יכולת ניתוח והפקת תובנות מנתונים גולמיים ומורכבים במטרה לזהות מוקדי סיכון בכבישים ולחזות תאונות דרכים עתידיות. כדי לממש את החזון של הפרויקט, אנחנו נדרשות להסתמך על מקורות נתונים מגוונים ומורכבים, באיכות גבוהה. מקורות הנתונים הם הליבה של המערכת, לכן אנחנו בבקרה על האופן שבו נאספים, מאוחסנים, ומופצים הנתונים.

**נתונים ממאגרי מידע ציבוריים**

הנתונים אשר נתבסס עליהם לקוחים מהאתר Data Gov והאתר של הלשכה המרכזית לסטטיסטיקה. הנתונים הללו מספקים מידע חיוני הכולל כוללים עשרות משתנים המפרטים את פרטי התאונה, מיקומה, נסיבותיה ותוצאותיה. בין המשתנים ניתן למצוא מזהה תאונה ייחודי, סוג הטיפול בתאונה, אזור גיאוגרפי, סוג הדרך, יישוב ומידע על רחובות הסמוכים לאירוע. בנוסף, הנתונים מכילים מידע כרונולוגי כמו שנה, חודש ושעה, יחד עם נתונים על חומרות התאונה, סוג התאונה ומיקום צמתים. כמו כן, מצוינות המהירות המותרת וקואורדינטות גיאוגרפיות של המקום. הגישה לנתונים הללו הינה חופשית, הקבצים נמצאים באתרים תחת פורמט csv. השימושים האפשריים כוללים חיזוי תאונות עתידיות, זיהוי אזורים מסוכנים, שיפור אכיפת תנועה, תכנון תשתיות מותאמות, והערכת סיכונים עבור חברות ביטוח, מה שהופך את הנתונים לכלי קריטי לשיפור בטיחות בדרכים ותשתיות.

**משימה 3 - זיהוי משאבי כוח אדם**

הפרויקט זקוק לגישה למומחים מגוונים על מנת לממש את מטרותיו בצורה מיטבית:

* **מומחים עסקיים:**
  + יועצים בתחום התחבורה ובטיחות בדרכים שיספקו הבנה מעמיקה של הבעיות והאתגרים הקיימים.
  + אנשי מקצוע מהרשויות המקומיות, משרד התחבורה והרשות הלאומית לבטיחות בדרכים.
* **מומחי נתונים:**
  + מדעני נתונים (Data Scientists) בעלי ניסיון בפיתוח מודלים לחיזוי ותובנות מנתונים גולמיים.
  + מהנדסי נתונים (Data Engineers) המסוגלים להתמודד עם תהליכי ניקוי, הכנה ואיחוד של מאגרי נתונים גדולים ומגוונים.
  + אנליסטים עסקיים שיתרגמו את התובנות הטכניות לשפה עסקית מובנת ויציעו פתרונות מעשיים לשיפור הבטיחות.
* **מנהלי מסדי נתונים (DBA):**
  + צוות מנהלי מסדי נתונים יהיה אחראי על הקמה, תחזוקה, ואבטחה של בסיסי הנתונים הרלציוניים והלא-רלציוניים. תפקידם יכלול ניהול גיבויים, שיפור ביצועים, ופתרון תקלות בזמן אמת.
* **צוותי תמיכה טכנית:**
  + **מפתחים (Developers)** אחראים לפיתוח הממשק של המערכת, הכולל נראות ידידותית ושימוש קל למשתמשים.
  + **מומחי IT** נדרשים לניהול תשתיות המערכת, הגנה על שרתים, ואבטחת מידע. בנוסף, הם יספקו תמיכה טכנית שוטפת למשתמשים במערכת.
  + **בודקי תוכנה (QA Testers)** יוודאו את איכות המערכת, בדיקות תקינות וחוויית משתמש מיטבית.
* **צוות הדרכה והטמעה:**
  + ייעודו של צוות זה הוא לספק הדרכות ולתמוך במשתמשים הראשונים של המערכת, כדי להבטיח תהליך הטמעה חלק ונכון.
* **מנהלי פרויקט:**
  + מנהלי פרויקט יובילו את תהליך הפיתוח, תוך תאום בין הצוותים השונים, עמידה בלוחות זמנים וניהול המשאבים באופן מיטבי.

1. **דרישות, הנחות ואילוצים**

**משימה 1 – קביעת דרישות**

1. **אבטחת נתונים**:

הפרויקט מטפל בנתונים רגישים כמו מיקומי תאונות, נתונים אישיים פוטנציאליים, ומידע גיאוגרפי מדויק, ולכן נדרש להבטיח אבטחת מידע ברמה גבוהה. יש להחיל הצפנה על כל שלבי הטיפול בנתונים, הן בזמן העברתם (Data in Transit) והן בזמן אחסונם (Data at Rest), באמצעות פרוטוקולים מתקדמים כמו SSL/TLS. בנוסף, על שמירת אבטח מידע גבוהה יש להגדיר הרשאות מבוססות תפקידים (RBAC) אשר יסייעו למנוע גישה בלתי מורשית לנתונים שלנו.

1. **רגולציה ותקנות**:
   * עמידה בתקנות פרטיות מידע כמו - General Data Protection GDPR Regulation או תקנות מקומיות המגנות על פרטיות המשתמשים.
   * אנונימיזציה של נתונים אישיים: על כל מידע אישי (PII) חייב לחול עיבוד כך שלא ניתן יהיה לזהות אדם פרטי, לדוגמה, באמצעות הסרת מזהים או הצפנה.
   * הפרויקט יכלול תהליך ביקורת קבוע על מנת לוודא עמידה בדרישות החוקיות.
2. **פרסום תוצאות**:

אנחנו מבטיחים כי התוצאות שיופקו יעמדו בתקנות אתיות וחוקיות. לדוגמה, לא ניתן לחשוף מידע רגיש כמו זיהוי מדויק של נפגעי תאונות, ישנו חיסיון מוחלט של מידע על הנפגעים באירוע.

**תיאום דרישות תזמון הפרויקט**

1. **תזמון שלבי הפרויקט:**

הפרויקט יתבצע בשלבים ברורים: איסוף נתונים, ניקוי ועיבוד נתונים, בניית מודלים לחיזוי, יצירת ממשק משתמש, והטמעת המערכת. לכל שלב יוקצו לוחות זמנים ברורים עם אבני דרך (Milestones) לבחינת ההתקדמות. זמנים אלו יהיו מיושרים עם זמינות הצוותים הרלוונטיים תוך שאיפה למיקסום זמנים. מקורות הנתונים, ואילוצים חיצוניים, כמו עיכובים אפשריים בקבלת אישורי רגולציה או נתונים ממקורות חיצוניים יילקחו בחשבון גם הם.

1. **תקשורת ותיאום:**

כל צוות בפרויקט, כולל מדעני הנתונים, המפתחים, ומנהלי הפרויקט, יפעלו בתיאום מלא. אנו נקיים פגישות עדכון שוטפות על מנת להבטיח יישור קו.

1. **ניהול סיכונים:**

אנו נזהה מבעוד מועד סיכונים פוטנציאליים המשפיעים על תזמון הפרויקט, כמו עיכובים באיסוף נתונים או תקלות טכנולוגיות, ונגדיר תוכנית גיבוי שתאפשר התמודדות עם הסיכונים הללו.

**משימה 2 - הבהרת והנחות**

1**. גורמים כלכליים שעלולים להשפיע על הפרויקט:**

**דמי ייעוץ:** ייתכן שיהיה צורך לשכור יועצים חיצוניים בתחומים כמו למידת מכונה, הנדסת תשתיות, או מומחי תחבורה לצורך ייעוץ או פיתוח ספציפי.

**מוצרים מתחרים:** אין מערכת חיזוי מתקדמת דומה בישראל נכון להיום, אך מערכות קיימות לטיפול בתאונות או בניתוח תחבורה עשויות להוות תחרות בעתיד, בעיקר אם יתפתחו במקביל.

**תקציב מוגבל:** מגבלות תקציביות עשויות להשפיע על רכישת תשתיות טכנולוגיות כמו שרתים חזקים או שירותי ענן מתקדמים.

**רגישות כלכלית:** הפרויקט עשוי להיות מושפע מתקציבי רשויות מקומיות או משרד התחבורה, במיוחד אם לא יראו תוצאות מהירות.

2. **הנחות איכות נתונים**

הנתונים ההיסטוריים מתאונות דרכים המגיעים מדאטה גוב ומהלשכה המרכזית לסטטיסטיקה נחשבים מקיפים, אך ייתכנו נתונים חסרים או שגויים (למשל, דיווחים חלקיים על תאונות, חוסר במידע גיאוגרפי מדויק או בתנאי מזג האוויר). חלק מהנתונים עלולים להיות לא מעודכנים, במיוחד נתוני תאונות או תשתיות. לכן, יתכן צורך בעיבוד נוסף של הנתונים לצורך שמירה על פרטיות, מה שעשוי להפחית את הדיוק של חלק מהניתוחים. ניקח בחשבון כי ייתכנו הטיות במידע (למשל, נתוני תאונות נוטים לדווח בעיקר על אזורים עירוניים ופחות על אזורים כפריים).

3. **ציפיות נותן החסות/צוות הניהול:**

נותני החסות מעוניינים בעיקר בתוצאות ברורות ומעשיות, כגון זיהוי מוקדי סיכון, הפחתת תאונות, והמלצות לשיפור תשתיות ואכיפה. בעוד שהצוות המקצועי עשוי להתעניין בפירוט הטכני של המודלים (כדי לוודא אמינות ושקיפות), מקבלי ההחלטות ירצו גישה לתובנות ברמה גבוהה (High-Level Insights) ולא בהכרח להבין את הפרטים הטכניים של המודל.

**משימה 3 - אימות אילוצים**

הפרויקט משתמש בנתונים הזמינים בפלטפורמת דאטה גוב, שברובם זמינים לציבור הרחב ללא צורך בסיסמאות מיוחדות. חלק מנתוני הלשכה המרכזית לסטטיסטיקה דורשים הרשאות גישה, כולל חתימה על הסכמים לשימוש בנתונים, במקרה שלנו הגישה אליהם היא חופשית (נכון להיום).

**אילוצים חוקיים על השימוש בנתונים**

* + 1. **רגולציות ותקנות פרטיות:**

השימוש בנתונים כפוף לחוקי הגנת הפרטיות בישראל. אנו נוודא שהנתונים יהיו אנונימיים ושכל עיבוד נתונים יעשה בהתאם לחוקי הגנת המידע.

לדוגמה, נתוני תאונות הכוללים מיקום גיאוגרפי חייבים לעבור תהליך של השמטת פרטים מזהים כדי למנוע פגיעה בפרטיות.

* + 1. **הסכמי שימוש:**

אנחנו נוודא שכל תנאי השימוש של פלטפורמות הנתונים, כמו דאטה גוב והלמ"ס, יתקיימו במלואם. חלק מהמאגרים עשויים להטיל מגבלות על הפצה או שימוש מסחרי, לכן אנחנו נערך לכך במידת הצורך.

* + 1. **אחריות משפטית:**

מבחינה משפטית נדאג שהמערכת או המלצותיה לא יהיו חשופות לתביעה משפטית עבור הגורמים המשתמשים בתוצאותיה, כמו הרשויות או חברות הביטוח.

**אילוצים כספיים ותקציב:**

כל האילוצים הכספיים הקשורים לפיתוח המערכת, כגון רכישת תשתיות מחשוב, אחסון ענן, ייעוץ מקצועי וניתוח נתונים, נלקחו בחשבון במסגרת התקציב. עלויות העסקת מדעני נתונים, מתכנתים, ומהנדסים נכללות בתקציב, אך במידה ותהיה התרחבות בצוות או צורך ביועצים חיצוניים, ייתכן שיהיה צורך להגדיל את התקציב.

**מגבלות תקציביות:**

במידה והפרויקט ייתקל במגבלות תקציביות, יש לתעדף רכיבים קריטיים כמו עיבוד נתונים ודיוק המודלים, על פני רכיבים שאינם הכרחיים מידית.

1. **סיכונים ומצבים**

|  |  |  |
| --- | --- | --- |
| **סוג הסיכון** | **תיאור** | **תוכנית מגירה** |
| **תזמון** | הפרויקט עלול להימשך זמן רב מהצפוי עקב הערכת זמנים לא נכונה, תקלות במהלך פיתוח המודלים הנדרשים. | תיעדוף משימות קריטיות ודחיית רכיבים שאינם דחופים לשלב מאוחר יותר. |
| **פיננסי** | בעיות תקציביות מצד נותן החסות עלולות להוביל לקיצוצים בפרויקט. | בחינה מחודשת של התקציב ותיעדוף משימות קריטיות.  חיפוש מימון נוסף משותפים פוטנציאליים כמו חברות ביטוח או קרנות ממשלתיות. |
| **נתונים** | איכות הנתונים או הכיסוי שלהם עלולים להיות ירודים (חוסר נתונים, דיוק לא מספק). | חיפוש מקורות נתונים משלימים.  פיתוח מודלים עמידים לשינויים באיכות הנתונים. |
| **תוצאות** | התוצאות הראשוניות עלולות להיות פחות דרמטיות מהצפוי או לא לספק תובנות המשך (תחזיות לא מדויקות מספיק). | פיתוח גרסה ראשונית המתמקדת במוקדים בעלי נתונים איכותיים.  יישור ציפיות עם נותני החסות באמצעות תקשורת שקופה. |

7**. טרמינולוגיה**

**מוקדי סיכון:** אזורים גיאוגרפיים בהם קיימת הסתברות גבוהה לתאונות דרכים, בהתבסס על נתונים היסטוריים ודפוסים שנמצאו בניתוח המידע.

**תחזיות חיזוי:** תהליך של שימוש במודלים סטטיסטיים או מבוססי למידת מכונה כדי לקבוע את ההסתברות לאירועים עתידיים, כגון תאונות דרכים במקומות מסוימים.

**מפת חום (Heat Map):** ייצוג גרפי של נתונים המציג רמות סיכון שונות באזורים גיאוגרפיים באמצעות צבעים, כאשר צבעים כהים יותר מסמנים סיכון גבוה יותר.

**API (ממשק תכנות יישומים):** ממשק המאפשר לתוכנות שונות לתקשר ולהחליף נתונים, כמו משיכת נתונים ממקורות כמו שירותי תחבורה או מזג אוויר.

**סקלאביליות (Scalability):** היכולת של מערכת להתמודד עם גידול בכמות הנתונים או הבקשות, מבלי לפגוע בביצועים.

**רגולריזציה:** טכניקה המשמשת במודלים של למידת מכונה כדי למנוע התאמת יתר (Overfitting) של המודל לנתוני האימון.

**מוצר MVP - Minimum Viable Product:** גרסה ראשונית של המערכת עם מינימום פונקציות הכרחיות, שמטרתה לבדוק את הרעיון ולקבל משוב מוקדם לשיפורים.

**רשימת משימות**

* + 1. שמירת רשימה של מונחים:

יצירת רשימה דינמית של מונחים, שתתעדכן על בסיס שאלות או קשיים שעולים במהלך העבודה.

* + 1. שיתוף המונחים:

פרסום הרשימה במערכת ארגונית, מייל או כחלק מתיעוד הפרויקט, כך שתהיה נגישה לכל חברי הצוות.

* + 1. שילוב מילון המונחים בתהליכי הדרכה:

הצגת המונחים החדשים כחלק מפגישות הצוות או חומרים כתובים, כדי להבטיח שכולם מבינים את השפה המשותפת.

8**. עלויות ויתרונות**

**עלויות משוערות**: נתונים מדאטה גוב והלמ"ס: חינם ברובם, אך ייתכן שיהיו עלויות עבור עיבוד.

**עיבוד נתונים ראשוני:** משאבים לעיבוד וניקוי נתונים צפויים לעלות כ-30,000 ש"ח (עלויות צוות ותשתית).

**פיתוח מערכת:** עלויות פיתוח, כולל בניית ממשקים ידידותיים למשתמש והטמעת מודלים, כ-150,000–200,000 ש"ח.

**פריסה בענן:** תשתית מחשוב ענן להפקת תובנות בזמן אמת – כ-10,000 ש"ח לחודש.

**עלויות תפעול:**

* תחזוקה שוטפת: צוות קטן לניהול ותחזוקה של המערכת – כ-20,000 ש"ח לחודש.
* עדכון מודלים: שדרוג ותחזוקה של המודלים והנתונים – כ-50,000 ש"ח בשנה.
* הדרכות ותקשורת: הכשרת משתמשים והסברה מול רשויות – כ-30,000 ש"ח.

**יתרונות הפרויקט**

המטרה העיקרית המושגת היא הפחתת תאונות דרכים. הפחתת תאונות ב-15% צפויה לחסוך חיי אדם ולצמצם פציעות, ובמקביל לחסוך למשק כ-100 מיליון ש"ח בשנה בעלויות רפואיות, ביטוחיות ומשפטיות.

**תובנות נוספות הנובעות מחקירת נתונים**

* זיהוי דפוסים חדשים: הבנת קשרים בין תנאי דרך, מזג אוויר והתנהגות נהגים, שיכולים לתרום לפיתוח תשתיות תחבורה מתקדמות יותר.
* הכוונת תקציבים: ייעול חלוקת תקציבים לשיפור תשתיות באזורים שבהם ההשפעה תהיה הגדולה ביותר.

**יתרונות מהבנת נתונים טובה יותר**

* תכנון ארוך טווח: תובנות עמוקות שיאפשרו למשרד התחבורה ולרשויות המקומיות לתכנן פרויקטים חכמים וממוקדים יותר.
* שיתוף פעולה בין-ארגוני: שיפור עבודת הרשויות המקומיות, חברות הביטוח והמשטרה באמצעות כלי ניתוח משותפים.
* תועלת לחברות ביטוח: התאמת פרמיות ביטוח לפי רמת סיכון אזורי, מה שיגדיל את שביעות הרצון של הלקוחות.

הפרויקט "SafeRoads" צפוי לעלות כ-1.5–2 מיליון ש"ח בשנה הראשונה (כולל הקמה ותפעול), אך היתרונות הפוטנציאליים, בעיקר בהפחתת תאונות וחיסכון למשק, עשויים להחזיר את ההשקעה תוך שנה עד שנתיים. יתר על כן, המערכת תיצור בסיס לשיפורים מתמשכים בבטיחות בדרכים לאורך זמן.

**9. מטרות מדעי הנתונים וקריטריונים להצלחה**

**9.1. סוג הבעיה והאלגוריתם המתאים לה במדעי הנתונים**

אשכולות (Clustering): קיבוץ מוקדי סיכון דומים בהתבסס על גורמים כמו מיקום, סוגי תאונות, ותנאי דרך.

סיווג (Classification): סיווג אזורים לפי רמת סיכון (נמוכה, בינונית, גבוהה) בהתבסס על נתונים היסטוריים.

רגרסיה (Regression): חיזוי מספר תאונות צפוי באזורים שונים על פי משתנים כמו תשתיות, עומסי תנועה, ותנאי מזג אוויר.

2. **יעדים טכניים**

פיתוח מודל חיזוי המספק תחזיות לשינויים ברמת הסיכון לתאונות עבור פרק זמן של בערך כשלושה חודשים.

**ניתוח התנהגות נהגים ותשתיות:** יצירת ניתוחים חודשיים של גורמי סיכון ספציפיים (כמו עומסי תנועה ומהירות מופרזת) וההשפעה שלהם על הסיכון.

**זמן חישוב קצר:** המערכת תפיק דוחות ותחזיות בתוך פחות מ-24 שעות מרגע הזנת הנתונים.

3. **מספרים ממשיים עבור התוצאות הרצויות:**

**דיוק תחזיות:** דיוק של לפחות 85% בזיהוי מוקדי סיכון.

**הפחתת תאונות:** זיהוי אזורים בהם ניתן להפחית תאונות ב-15% באמצעות שיפורים בתשתיות או אכיפה ממוקדת.

**זמן פריסה:** השלמת הפריסה הראשונית של המערכת בתוך 6 חודשים, עם יכולת לספק תובנות ב-90% מהמקרים בהם נתונים הוזנו.

**קריטריונים להצלחה**

**תפעול אפקטיבי של המערכת**: המערכת תפעל ללא תקלות ותספק תחזיות מדויקות לנתוני אמת.

**יישום תובנות על ידי הרשויות:** לפחות 50% מההמלצות שיוצגו על ידי המערכת יבוצעו על ידי הרשויות המקומיות בתוך שנה.

**משוב חיובי ממשתמשים:** שביעות רצון גבוהה מצד מקבלי ההחלטות (ציון של לפחות 8 מתוך 10 בסקר משתמשים).

**חיסכון מוכח:** חיסכון מצטבר של כ-10 מיליון ש"ח בשנה הראשונה בעקבות הפחתת תאונות ושיפור תשתיות.

**9.2. קריטריוני הצלחה במדעי הנתונים**

תיאור השיטות להערכת המודל:

**דיוק (Accuracy):** שיעור התחזיות הנכונות של המודל מתוך כלל התחזיות.

**Recall (שליפה):** מדידת אחוז המקרים בפועל שבהם זוהו מוקדי סיכון על ידי המודל.

**Precision (דיוק התחזיות החיוביות):** אחוז התחזיות שסומנו כמוקדי סיכון ואכן היו נכונות.

**AUC-ROC (איכות ההבחנה בין קטגוריות):** מדד לשיפור היכולת של המודל להבדיל בין אזורי סיכון גבוה לנמוך.

**זמן חישוב:** פרק הזמן הדרוש להפקת תובנות ותחזיות מרגע הזנת הנתונים.

**הגדרת אמות מידה להערכת הצלחה עם מספרים ספציפיים:**

* **דיוק התחזיות:** המודל יעמוד בדרישה של 85% דיוק לפחות בזיהוי מוקדי סיכון.
* **Recall**: לפחות 80% מהמוקדים בעלי הסיכון הגבוה יסומנו נכון על ידי המודל.
* **Precision**: לפחות 75% מהתחזיות למוקדי סיכון גבוה יתבררו כנכונות.
* **זמן הפקת תובנות:** הפקת תובנות בתוך פחות מ-24 שעות לאחר הזנת נתונים חדשים.
* **אזורי סיכון שנחקרו:** כיסוי של לפחות 90% מאזורי הכבישים המועדים לתאונות לפי נתונים היסטוריים.

**מדידות סובייקטיביות**

* רמת שביעות רצון גבוהה מצד הרשויות המקומיות, משרד התחבורה, ומשטרת התנועה (ציון של לפחות 8 מתוך 10 בסקרים).
* אמון בשימוש במערכת מצד חברות הביטוח והמשתמשי קצה.

**פוסק ההצלחה:**

הצוות אשר מוביל את הפרויקט בשיתוף פעולה עם נציגי משרד התחבורה והרשויות המקומיות יהיו אלו אשר אחראים על קביעת הצלחתו הסופית של הפרויקט.

**תכנון פריסה מוצלחת של המודל**

* 1. שלבים בפריסה:
* פריסה ראשונית באזורים עירוניים עם נתונים איכותיים.
* בדיקות חיות ותיקוני באגים במהלך פריסת הפיילוט.
* הרחבת המערכת לכלל הכבישים בישראל.
  1. תמיכה בפריסה:
* תוכנית הדרכה למשתמשים ברשויות המקומיות ובחברות הביטוח.
* תמיכה טכנית רציפה במהלך השימוש הראשוני במערכת.
  1. מדדים להצלחת הפריסה:
* לפחות 50% מהרשויות המקומיות יישמו את תובנות המערכת תוך 6 חודשים מפריסתה.
* לפחות 70% מהמשתמשים יעניקו ציון חיובי על השימוש במערכת.

**משימות נוספות להבטחת הצלחה מיטבית**

**מעקב מתמיד אחר ביצועי המודל:** יצירת לולאת פידבק אוטומטית לשיפור מתמשך של המודל בהתבסס על נתוני אמת.

**התאמת המודל לשינויים בסביבה:** שיפור מתמיד של המודל כך שיתאים לנתונים חדשים ותנאים משתנים (למשל, מזג אוויר קיצוני או שינויים בתשתיות).

**הטמעת תהליך בקרה לאחר יישום ההמלצות:** ניטור ההשפעה של ההמלצות המיושמות, כמו שיפור תשתיות או אכיפה, על צמצום התאונות.

**10. תוכנית הפרויקט**

המטרה אשר מנחה אותנו לאורך כל התהליך היא לוודא שכל המשתתפים ובעלי העניין כולל מדעני הנתונים, צוותי הפיתוח, נציגי משרד התחבורה, רשויות מקומיות, וחברות הביטוח, מבינים את מטרות הפרויקט ואת שלביו בשביל להפיק ולבצע את הנדרש על הצד הטוב ביותר, גם מבחינה מערכתית וגם מבחינה הישגית.

**משימות:**

* עריכת פגישות סקירה עם כל הגורמים המעורבים.
* תיעוד סיכומי הפגישות וההחלטות שהתקבלו.
* הצגת גרסה ראשונית של התוכנית וקבלת משוב לשיפור.

**הערכות זמן לכל שלב או פעילות**

**המטרה:** לספק מסגרת זמנים ברורה שתאפשר מעקב יעיל אחרי התקדמות הפרויקט.

**משימות:** הערכת זמן עבור שלבים עיקריים: איסוף נתונים, עיבוד נתונים, פיתוח מודלים, פריסה, ותחזוקה.יצירת לוח זמנים מפורט לכל פעילות עם תאריכי התחלה וסיום מתואמים, סימון תאריכי אבני דרך (Milestones) לסקירת התקדמות של הפרויקט.

**מאמץ ומשאבים לפריסת תוצאות ופתרון עסקי**

**המטרה:** להבטיח שיש משאבים זמינים לפריסה חלקה של המערכת ותמיכה לאחר מכן.

**משימות:** הערכת עלויות משאבים (תשתיות, כוח אדם, הדרכות). לבצע תכנון מערך תמיכה טכנית לפריסה ולתפעול. בנוסף, אנו ניצור מסמך הנחיות למשתמשים לצורך יישום מהיר של התובנות.

**נקודות החלטה ובקשות עיון**

**המטרה:** להדגיש שלבים קריטיים בהם יש לקבל החלטות אסטרטגיות ולהשיג אישורים.

**משימות:** סימון נקודות שבהן תידרש החלטה, כמו בחירת המודל המתאים ביותר או אישור פריסה.

הגדרת גורמי מפתח שיקבלו החלטות (פוסק הצלחה או נציגי משרד התחבורה). לבצע תיעוד של כל ההחלטות לצורך בקרה ושקיפות.

סימון שלבים עם חזרות מרובות (Iteration):

**המטרה:** לתכנן מספיק זמן עבור כל פעילות ולהקצות משאבים עבור שלבים כמו עיבוד נתונים, בהם מתרחשות חזרות לצורך שיפור מתמיד.

**משימות:** תיעוד תהליך הפיתוח המערכתי (מודלים, עיבוד נתונים).לבצע תכנון לולאות פידבק בין תוצאות ביניים לשיפורים עתידיים ומתמידיים. אנו נבטיח זמינות ונקצה זמן ומשאבים עבור שיפור המודלים על בסיס נתונים חדשים.

**מבנה תוכנית הפרויקט**

**מבוא:** תיאור כללי של הפרויקט: מטרות עסקיות, בעיות לפתרון, והתועלת הצפויה.

**לוח זמנים:** הכנת טבלת שלבים מרכזיים ותאריכים הכוללת את הפעולות הנדרשות כמו - איסוף נתונים, עיבוד, פיתוח מודלים, בדיקות, ופריסה.

**משאבים נדרשים:**

* כוח אדם: תפקידים מרכזיים וצוותים תומכים.
* חומרה ותוכנה: מערכות אחסון, מחשוב ענן, כלים טכנולוגיים.
* תקציב: פירוט של עלויות לפי שלבים.

**תכנון שלבים חוזרים:**

סימון אזורים בהם נדרשות חזרות (לדוגמא – אזור סיכון אשר עבר שדרוג מבחינת תשתיות או לסירוגין מערכות טכנולוגיות משטרתיות כמו מצלמות מהירות וכו׳) ועל האזורים האלו נפעיל מניפולציות של למידת מכונה כמו בניית מודלים או בדיקות. נדאג ליצור לוחות זמנים גמישים לאזורים אלו כיוון שעבודות ושיפורים מסוג זה הינן דינמיות ועלולות להימשך זמן רב מהצפוי.

**מדדי הצלחה:**

אנו נשתמש בקריטריונים כמותיים וסובייקטיביים להערכת הצלחת הפרויקט מכל הבחינות.

**סיכונים ותוכניות מגירה:**

נדאג שיהיה לנו פירוט הסיכונים האפשריים ותוכניות פעולה במקרה של בעיות העלולות לצוץ תוך משך הפרויקט ואפילו בשלבים מוקדמים שלו, אנחנו נבצע הערכות סיכונים שקולות וחכמות על מנת שהקמת הפרויקט תעבור בצורה החלקה ביותר.

11**. הערכה ראשונית של כלים וטכניקות**

בחרנו ב-Python ככלי המרכזי לניהול וביצוע תהליכי מדעי הנתונים בפרויקט. Python מספקת גמישות וכלים מתקדמים שמאפשרים לנו לטפל בכל שלב בתהליך. נוכל לבצע עיבוד מקיף של הנתונים וטיפול בערכים חסרים, כדי להבטיח שהמודל יתבסס על נתונים איכותיים ומדויקים. בנוסף, Python מאפשרת לנו לפתח מודלים חכמים ומתקדמים, כמו מודלים לחיזוי מוקדי סיכון ותאונות עתידיות. בשלב האימון נשתמש בטכניקות מתקדמות כדי להבטיח שהמודלים שלנו יהיו מדויקים ועמידים בפני שינויים. כמו כן, נערוך בדיקות חוזרות להערכת ביצועי המודל ונבצע אופטימיזציות כדי להגיע לתוצאות הטובות ביותר.

לאחר שלב בניית המודל, נציג את התוצאות בצורה ברורה ואינטראקטיבית באמצעות לוחות מחוונים. לשם כך, נשלב את Python עם כלים נוספים כמו Power BI, שיאפשרו הצגת נתונים חזותיים ברמה גבוהה. הלוח המחוון יציג נתונים גיאוגרפיים, כמו מפות חום ומוקדי סיכון, ויספקו למשתמשים חוויה ידידותית ומידע נגיש לקבלת החלטות.

**Data Understanding Report**

* 1. **איסוף נתונים**
  2. **מקורות נתונים**

הנתונים אשר נתבסס עליהם בפרויקט שלנו נאספו ממאגרים ציבוריים הכוללים את Data.gov.il .

**Data.gov**

הנתונים מכילים מידע כמו מיקומי תאונות (יישוב, רחובות), נתוני זמן (שנה, חודש, שעה), תנאי דרך (מהירות מותרת, סימון ותמרור, תאורה), סוג התאונה וחומרתה, ומאפייני הכביש. כמו כן, הנתונים כוללים מידע על האזור המנהלי בו אירעה התאונה ומאפיינים סביבתיים כמו מזג האוויר ופני השטח. הנתונים מגיעים בקובץ CSV אשר מכיל מידע מפורט על תאונות עירוניות ובין-עירוניות בשנים 2022-2023.

**נתונים שנרכשו:**

בפרויקט לא נעשה שימוש בנתונים שנרכשו ממקורות חיצוניים, כגון נתוני דמוגרפיה. אם יתעורר צורך בהבנת פרופיל הנהגים או התאונות לפי מאפיינים דמוגרפיים (כמו גיל, מגדר), ניתן יהיה לשקול שימוש בנתונים משלימים ממקורות כמו סקרים או מאגרים פרטיים.

**נתונים נוספים:**

במידה ויתברר שהמידע הקיים אינו מספיק או שיעלה צורך בנתונים משלימים שאין ברשותנו, נוכל להוסיף נתונים נוספים על ידי שיתוף פעולה עם גורמים חיצוניים כמו המשטרה, מד"א והרשויות המקומיות. בנוסף נוכל להשתמש בנתונים מהלשכה המרכזית לסטטיסטיקה (הלמ"ס).

* 1. **בדיקה ראשונית של הנתונים**

פירוט של קובץ הנתונים מאתר data gov.

|  |  |  |  |
| --- | --- | --- | --- |
| מס׳ | מאפיין | תיאור מאפיין | סינון ראשוני לעיבוד נתונים בהמשך |
| 1 | PK\_TEUNA\_FIKT | מספר תאונה פיקטיבי ייחודי |  |
| 2 | SUG\_TIK | סוג התיק, אופן הטיפול בתאונה |  |
| 3 | THUM\_GEOGRAFI | האזור הגיאוגרפי בו התרחשה התאונה, 1 – ישראל ללא יהודה ושומרון, 2 -יהודה והשומרון. |  |
| 4 | SUG\_DEREH | סוג דרך – 1 - עירוני בצומת, 2 - עירוני לא בצומת, 3 - לא עירוני בצומת, 4 - לא עירוני לא בצומת. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 5 | SEMEL\_YISHUV | סמל ישוב – מספרים שמסמנים את היישובים בארץ ישראל. מגיע עם קידוד מתאים במילון עזר. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 6 | REHOV1 | רחוב 1 – הראשי, מילון רחובות לפי מספר מקבלים את הרחוב בו התרחשה התאונה. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 7 | REHOV2 | רחוב 2 – המשני במידה והתאונה קרתה בצומת, מילון רחובות לפי מספר מקבלים את הרחוב בו התרחשה התאונה. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 8 | MISPAR\_BAIT | מספר בית. |  |
| 9 | ZOMET\_IRONI | צמתים עירוניים (המילון עזר לא נמצא). |  |
| 10 | KVISH1 | דרך לא עירונית – 1. |  |
| 11 | KVISH2 | דרך לא עירונית – 2. |  |
| 12 | KM | ק״מ. |  |
| 13 | ZOMET\_LO\_IRONI | צמתים לא עירוניים (המילון עזר לא נמצא). |  |
| 14 | YEHIDA | יחידה משטרתית, 11 – מרחב חוף (חיפה), 12 – מרחב גליל, 14 – מרחב עמקים, 20 – מרחב תל אביב, 33 – מרחב אילת, 34 – מרחב הנגב, 36 – מרחב שמשון (עד 1999), 37 – מרחב שמשון (החל מ2004), 38 – מרחב לכיש, 41 – מרחב שומרון, 43 – מרחב יהודה, 51 – מרחב השרון, 52 – מרחב השפלה, 61 – מחוז ירושלים. |  |
| 15 | SHNET\_TEUNA | שנת תאונה – 2022 /2023. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 16 | HODESH\_TEUNA | חודש תאונה – לפי מס׳ חודשים לועזי 1 -12. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 17 | YOM\_BE\_HODES | יום בחודש (המילון עזר לא נמצא). |  |
| 18 | SHAA | שעה (המילון עזר לא נמצא). |  |
| 19 | SUG\_YOM | סוג יום, 1 – חג, 2 – ערב חג, 3 – חול המועד, 4 – יום אחר. |  |
| 20 | YOM\_LAYLA | יום/לילה – יום 1, לילה 5 | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 21 | YOM\_BASHAVUA | יום בשבוע – כמס׳ הימים בשבוע 1 -7 | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 22 | RAMZOR | מרומזר/לא מרומזר – לא זה 0, כן זה 1, לא ידוע זה 9. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 23 | HUMRAT\_TEUNA | חומרת התאונה – לפי דרגות, 1 – זה קטלנית, 2 – קשה, 3- קלה. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 24 | SUG\_TEUNA | סוג תאונה – 1 – פגיעה בהולך רגל, 2 – התנגשות חזית אל צד, 3 - התנגשות חזית באחור, 4 – התנגשות צד בצד, 5 – התנגשות חזית אל חזית, 6 – התנגשות עם רכב שנעצר ללא חנייה, 7 – התנגשות עם רכב חונה, 8 – התנגשות עם עצם דומם, 9 – ירידה מהכביש/ עלייה למדרכה, 10 – התהפכות, 11- החלקה, 12 – פגיעה בנוסע בתוך כלי רכב, 13 – נפילה ברכב נע, 14 – שריפה, 15 – אחר, 17 – התנגשות אחור אל חזית, 18 – התנגשות אחור אל צד, 19 – התנגשות עם בעל חיים, 20 – פגיעה ממטען של רכב. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 25 | ZURAT\_DEREH | צורת הדרך, 1 – כניסה למחלף, 2 – ביציאה ממחלף, 3 – מ.חנייה/ת דלק, 4 – שיפוע תלול, 5 – עקום חד, 6 – על גשר מנהרה, 7 – מפגש מסילת ברזל, 8 – כביש ישר/ צומת, 9 – אחר, 10 – תחנת אוטובוס, 11 – נתיב תחבורה ציבורית. |  |
| 26 | HAD\_MASLUL | דרך חד מסלולית (חלופה לדרך רב מסלולית), 1 – חד סטרי, 2 – דו סיטרי + קו הפרדה רצוף, 3 – דו סיטרי אין קו הפרדה רצוף, 4 -אחר, 9 – לא ידוע מספר מסלולים. |  |
| 27 | RAV\_MASLUL | דרך רב מסלולית (חלופה לדרך חד מסלולית), 1 – מיפרדה מסומנת בצבע, 2 – מיפרדה עם גדר בטיחות, 3 – מיפרדה בנויה ללא גדר בטיחות, 4 – מיפרדה לא בנויה, 5 – אחר. |  |
| 28 | MEHIRUT\_MUTERET | מהירות מותרת (בקמ״ש) – 0 – לא ידוע, 1 – עד 50, 2 – 60, 3 – 70, 4 – 80, 5 – 90, 6 – 100, 7 – 110. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 29 | TKINUT | תקינות הדרך – 0 – לא ידוע, 1 – אין ליקוי, 2 – שוליים גרועים, 3- כביש משובש, 4 – שוליים גורעים בכביש משובש. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 30 | ROHAV | רוחב הכביש, 0 – לא ידוע, 1 – עד חמש מטר, 2 – חמש עד שבע מטר, 3 – שבע עד עשר וחצי, 4 – עשר וחצי עד ארבע עשרה, 5 – יותר מארבע עשרה. |  |
| 31 | SIMUN\_TIMRUR | סימון/תימרור – 1 – סימון לקוי/ חסר, 2 – תמרור לקוי / חסר, 3 – אין ליקוי, 4 – לא נדרש תמרור, 5 – לא ידוע. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 32 | TEURA | תאורה – 1 – אור יום רגיל, 2 – ראות מוגבלת עקב מזג אוויר (עשן, ערפל), 3 – לילה פעלה תאורה, 4 – קיימת תאורה בלתי תקינה/ לא פועלת, 5- לילה לא קיימת תאורה, 6 – לילה לא ידוע, 7 – לילה תאורה תקינה עם ראות מוגבלת, 8 – לילה תאורה לא תקינה עם ראות מוגבלת, 9 – לילה ללא תאורה עם ראות מוגבלת, 10 – דמדומים, 11 – יום לא ידוע. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 33 | BAKARA | בקרה בצומת – רק בעירוני בצומת או לא עירוני בצומת. 1 – אין בקרה, 2 – רמזור תקין, 3 – רמזור מהבהב צהוב, 4 – רמזור לא תקין, 5 – תמרור עצור, 6- תמרור זכות קדימה, 7 – אחר. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 34 | MEZEG\_AVIR | מזג אויר – 1 – בהיר, 2 – גשום, 3 – שרבי, 4 – ערפילי, 5 – אחר, 9 – לא ידוע. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 35 | PNE\_KVISH | מצב פני הכביש – 1 – יבש, 2 – רטוב ממים, 3 – מרוח בחומר דלק, 4 – מכוסה בבוץ, 5 – חול או חצץ על הכביש, 6 – אחר, 9 – לא ידוע. | כללי ברזל - בלוק - Block - מרכז הגנת סייבר לאזרחים |
| 36 | SUG\_EZEM | סוג העצם (רק בסוג תאונה 8 – התנגשות עם עצם דומם), 1 – עץ, 2 – עמוד חשמל/ תאורה /טלפון, 3 – תמרור ושלט, 4 – גשר סימניו ומגניו, 5 – מבנה, 6 – גדר בטיחות לרכב, 7 – חבית, 8 – אחר, 9 – לא ידוע. |  |
| 37 | MERHAK\_EZEM | מרחק העצם (רק בסוג תאונה 8 – התנגשות עם עצם דומם), 1 – עד מטר, 2 – אחד עד שלוש מטר, 3 – על הכביש, 4 – על שטח הפרדה, 9 – לא ידוע. |  |
| 38 | LO\_HAZA | לא חצה ( רק בסוג תאונה 1, חלופה לאופן חצייה, תאונה מסוג 1 זה פגיעה בהולך רגל), 1 – הלך בכיוון התנועה, 2 – הלך נגד, 3 – שיחק על הכביש, 4 – עמד על הכביש, 5 – היה על אי הפרדה, 6 – היה על שוליים/ מדרכה, 7 – אחר, 9 – לא ידוע. |  |
| 39 | OFEN\_HAZIYA | אופן חצייה (רק בסוג תאונה 1, חלופה לאופן חצייה, תאונה מסוג 1 זה פגיעה בהולך רגל), 1 -התפרץ אל הכביש, 2 – חצה שהוא מוסתר, 3 – חצה רגיל, 4 – אחר, 8 – לא חצה, 9 – לא ידוע אם חצה. |  |
| 40 | MEKOM\_HAZIYA | מקום חצייה (רק בסוג תאונה 1, תאונה מסוג 1 זה פגיעה בהולך רגל), ), 1 – לא במעבר חצייה ליד צומת, 2 - לא במעבר חצייה לא ליד צומת, 3 – במעבר חצייה בלי רמזור, 4 – במעבר חצייה עם רמזור, 9 – לא ידוע מקום חצייה. |  |
| 41 | KIVUN\_HAZIYA | כיוון חצייה (רק בסוג תאונה 1, תאונה מסוג 1 זה פגיעה בהולך רגל), 1 – מימין לשמאל, 2 – משמאל לימין, 9 – לא ידוע. |  |
| 42 | MAHOZ | מחוז (רק בסוג דרך 1,2 – 1 זה עירוני בצומת, 2 זה עירוני לא בצומת) המילון עזר לא נמצא. |  |
| 43 | NAFA | נפה (רק בסוג דרך 1,2 – 1 זה עירוני בצומת, 2 זה עירוני לא בצומת) המילון עזר לא נמצא. |  |
| 44 | EZOR\_TIVI | אזור טבעי (רק בסוג דרך 1,2 – 1 זה עירוני בצומת, 2 זה עירוני לא בצומת) המילון עזר לא נמצא. |  |
| 45 | MAAMAD\_MINIZIPALI | מעמד מוניציפלי (רק בסוג דרך 1,2 – 1 זה עירוני בצומת, 2 זה עירוני לא בצומת) המילון עזר לא נמצא. |  |
| 46 | ZURAT\_ISHUV | צורת יישוב (רק בסוג דרך 1,2 – 1 זה עירוני בצומת, 2 זה עירוני לא בצומת) המילון עזר לא נמצא. |  |
| 47 | IGUN\_MEKUBAZ | איכות עיגון, 1 – עיגון מדויק, 2 – מרכז יישוב, 3 – מרכז דרך, 4 – מרכז קילומטר, 9 – לא עוגן. |  |
| 48 | X\_KORDINAT | קורדינטה איקס. |  |
| 49 | Y\_KORDINAT | קורדינטה וואי. |  |

הנתונים שנאספו כוללים מידע רחב ומפורט על תאונות דרכים משנים 2022 ו- 2023, בחרנו להתמקד בשנים האלו כיוון שיש המון רשומות ותיעוד על תאונות דרכים וגם הן עדכניות. כך נוכל לדייק את התחזיות באופן המקסימלי. בנוסף, הנתונים בdataset שלנו מכילים מספר רב של תכונות המאפשרות לבצע ניתוחים סטטיסטיים והסקת מסקנות. מאפיינים כמו שעת התאונה, חומרת התאונה, סוג הדרך, תנאי מזג האוויר ומיקום גיאוגרפי מספקים בסיס מצוין לבניית מודלים חיזויים ולזיהוי מגמות בתאונות דרכים. עם בסיס נתונים הכולל מידע רחב על שנתיים מלאות, ניתן להפיק תובנות משמעותיות וליישם מודלים לחיזוי תאונות עתידיות. בנתונים המקוריים נכללו 49 עמודות (מאפיינים), אך לא כולן רלוונטיות לניתוח וחיזוי תאונות דרכים. לאחר סינון והסרת העמודות הלא רלוונטיות, נותרו 18 עמודות שמתארות את מאפייני התאונה בצורה ממוקדת ואפקטיבית. מספר זה נחשב מתאים ליישום מודלים חיזויים, תוך שמירה על איזון הנתונים ומיצוי מלא של המודל.

* על פי הניתוח הראשוני וההבנה שלנו על הנתונים בdataset, החלטנו להוסיף עמודה חדשה אשר בה סימנו בוי ירוק את המאפיינים שלפי דעתנו ישמשו אותנו בהמשך כבר בעיבוד הנתונים נציין כי זה סינון ראשוני וייתכן ונשנה את הסיווג של המאפיין.
* במהלך הטיפול בנתונים, זוהו עמודות מסוימות המכילות ערכים חסרים, כפי שנראה במידע הגולמי. ערכים חסרים בדאטה עשויים להופיע עקב טעויות באיסוף נתונים או העדר מידע באירועים מסוימים, כמו מיקום גיאוגרפי (X, Y), מספר בית, או רחובות חסרים. הטיפול בערכים החסרים לא בוצע בדאטה שלנו, יש פשוט תאים ריקים בנתונים עצמם, במידת הצורך, בהמשך הפרויקט נבצע טיפול בערכים חסרים על מנת לעבד את הנתונים באופן המקסימלי.

1. תיאור הנתונים
   1. כמות הנתונים:

בשנת 2022 ישנם מספר תצפיות כולל של [10,404].

בשנת 2023 ישנם מספר תצפיות כולל של [8,832].

מספר המאפיינים (עמודות): הדאטה מכיל 49 מאפיינים שמתארים את התאונות ואת נסיבותיהן.

* 1. סוגי הערכים:

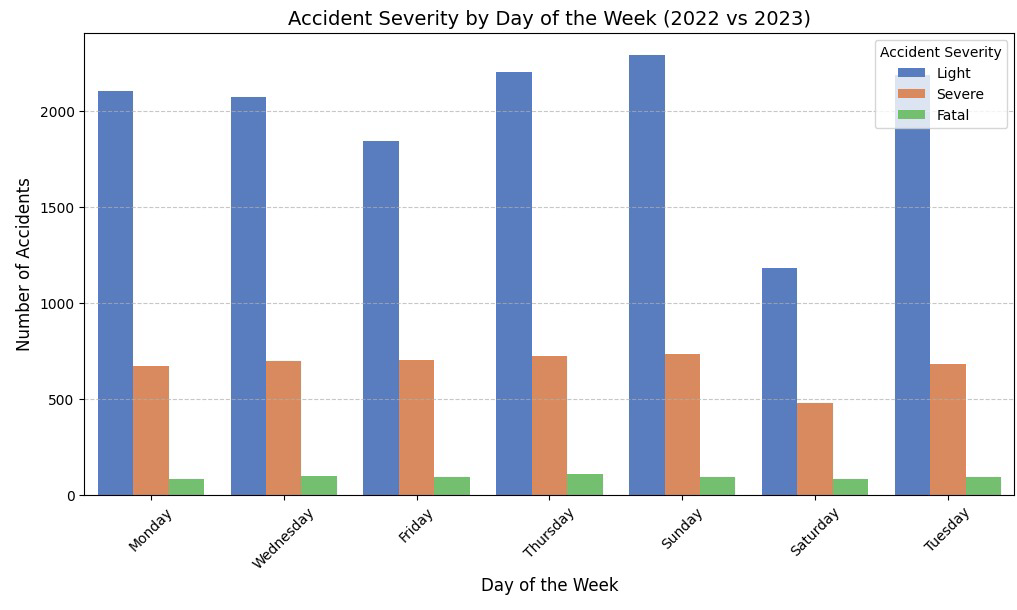
כל הנתונים בדאטה שלנו הם מסוג מספרי (Numeric), כולל ערכים שמייצגים מספרים רציפים כמו קואורדינטות או ערכים שלמים המייצגים קטגוריות מקודדות, כמו סוג יום, חומרת תאונה, ומאפייני דרך. הקטגוריות המקודדות מובילות למילון עזר שבו על פי הנתון המספרי נקבל נתון טקסטואלי מתאים בערכו.

* 1. סכמות קידוד:

בנתונים שלנו נעשה שימוש בסכמות קידוד מספריות עבור מאפיינים קטגוריאליים, כאשר ערכים מספריים מייצגים קטגוריות שונות. לדוגמה, עמודות כמו SUG\_YOM (סוג יום) ו-SUG\_TEUNA (סוג תאונה) מקודדות באמצעות מספרים שלמים כדי לייצג מצבים או קטגוריות שונות, כגון יום חול, סוף שבוע, או סוגי תאונות שונים.

סכמות קידוד אלו מאפשרות עבודה נוחה עם המידע במודלים סטטיסטיים ולמידת מכונה. לנוחות ולהבנה טובה יותר, צירפנו למעלה טבלה המסבירה את המשמעות של כל קוד בעמודות השונות, כפי שנמצא בנתונים.

1. חקירת נתונים
   1. במהלך חקירת הנתונים, השתמשנו בגרפים ובכלים שונים בשפת פייתון באמצעות Google Colab כדי להמחיש ולהבין את הקשרים בין המשתנים השונים. העבודה כללה יצירת גרפים שמטרתם הייתה לזהות אם קיימת השפעה הדדית בין המשתנים, ולבדוק את מידת השפעתם של משתנים אחד על השני. באמצעות הגרפים, ניתחנו את המידע בצורה ויזואלית, מה שסייע לנו להבין את הדינמיקה בין הנתונים ולזהות מגמות, קשרים ויחסים פוטנציאליים בין המשתנים השונים.

ניתוח גרף ראשון: חומרת תאונות לפי יום בשבוע (השוואה בין 2022 ל-2023)

מטרת הגרף היא לבחון את הקשר בין יום בשבוע לבין כמות התאונות בכל רמות החומרה (קלות, קשות וקטלניות) לאורך השנים 2022 ו-2023. בנוסף, הגרף נועד לזהות תבניות והתנהגויות הקשורות לימים מסוימים שבהם מתרחשות יותר תאונות, וכיצד חומרתן משתנה בין טווח השנים.

בחרנו להציג את הגרף בצורה זו כיוון שהוא מאפשר השוואה ברורה בין שתי השנים עבור כל יום בשבוע. הוא מראה גם את כמות התאונות הכללית וגם את ההתפלגות לפי חומרת התאונות.

מסקנות מהגרף:

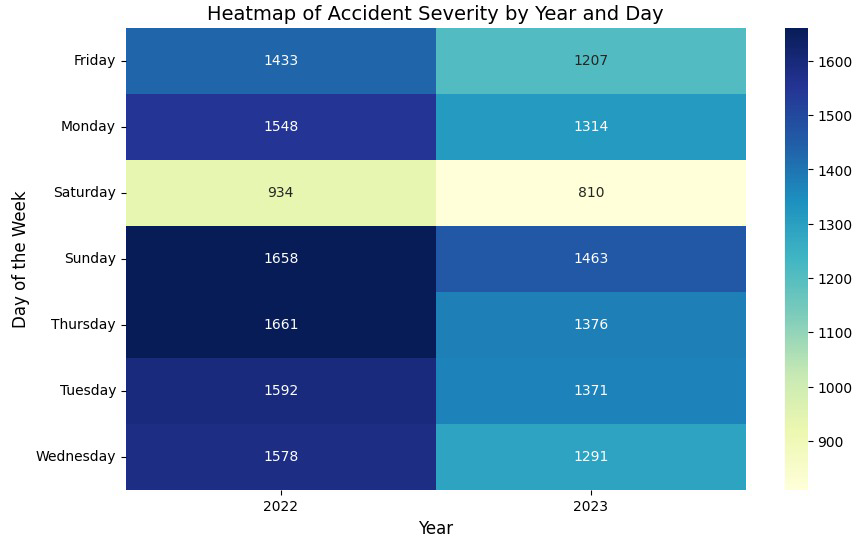
* ימים כמו ראשון ושלישי בולטים בכמות תאונות גדולה, מה שעשוי להעיד על דפוסי תנועה מסוימים, כמו עומסים או התנהגויות נהיגה.
* רוב התאונות הן קלות, עם שיעור קטן של תאונות קשות וקטלניות. עם זאת, ייתכן שתאונות קשות בולטות יותר בימים עם פחות תנועה.
* שינוי בין שנים - ייתכנו שינויים בכמות התאונות או בהתפלגות החומרה בין 2022 ל-2023, מה שעשוי להצביע על גורמים כמו שינויים במדיניות התנועה, תנאי מזג אוויר, או התנהגות נהגים.

כיצד הגרף תורם לפרויקט:

* קבלת החלטות מבוססת נתונים: הגרף מספק בסיס להשוואה בין שנים, ומאפשר בחינה האם צעדים מסוימים שננקטו (כגון הגברת האכיפה) השפיעו על הפחתת התאונות.
* תחזיות עתידיות: המידע יכול לתרום לבניית מודלים לחיזוי תאונות לפי יום בשבוע, מה שיאפשר היערכות טובה יותר.

מסקנות וצעדים עתידיים:

* כדאי לבחון את הימים עם כמות תאונות גבוהה יותר באופן מפורט ולזהות האם יש מאפיינים ייחודיים (שעות מסוימות, מיקומים ספציפיים).
* שילוב נתוני מזג אוויר או עומסים עשוי לתת פרספקטיבה רחבה יותר על הסיבות לעלייה בכמות התאונות בימים מסוימים.

ניתוח גרף שני: מפת חום של חומרת תאונות לפי יום ושנה

מטרת מפת החום היא להוות כלי ויזואלי עוצמתי להצגת קשרים ודפוסים במערכת נתונים מרובת משתנים. במקרה זה, מפת החום נועדה להמחיש את הקשר בין ימי השבוע, השנים (2022 ו-2023), וכמות התאונות בכל יום. הצבעים במפה מייצגים את כמות התאונות, כאשר צבעים כהים יותר מייצגים ערכים גבוהים וצבעים בהירים מייצגים ערכים נמוכים יותר.

בחרנו להציג את הגרף בצורה זו כיוון שמפת חום מאפשרת זיהוי חזותי מהיר, במקרה שלנו של ימים מסוימים בשבוע שבהם כמות התאונות גבוהה יותר, והשוואה פשוטה בין השנים, הצבעים מקלים על הבנת הריכוזיות וההתפלגות של התאונות לאורך השבוע.

מסקנות מהגרף:

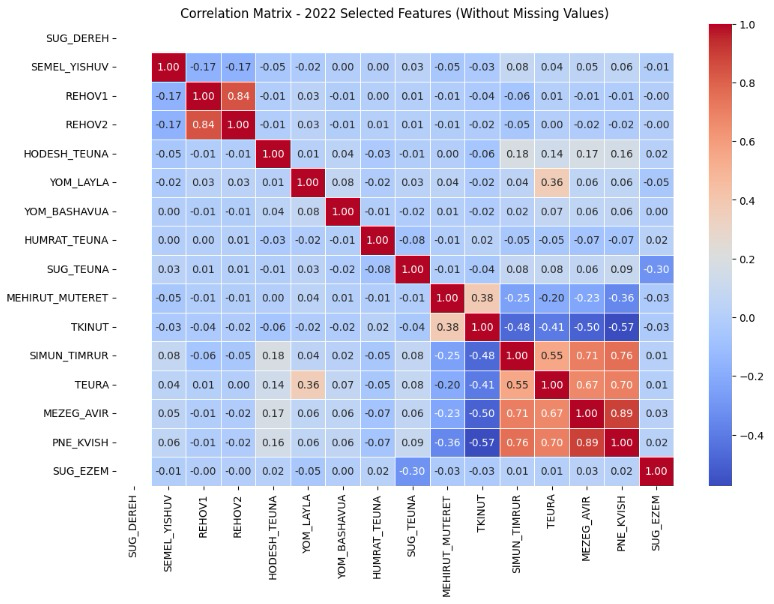
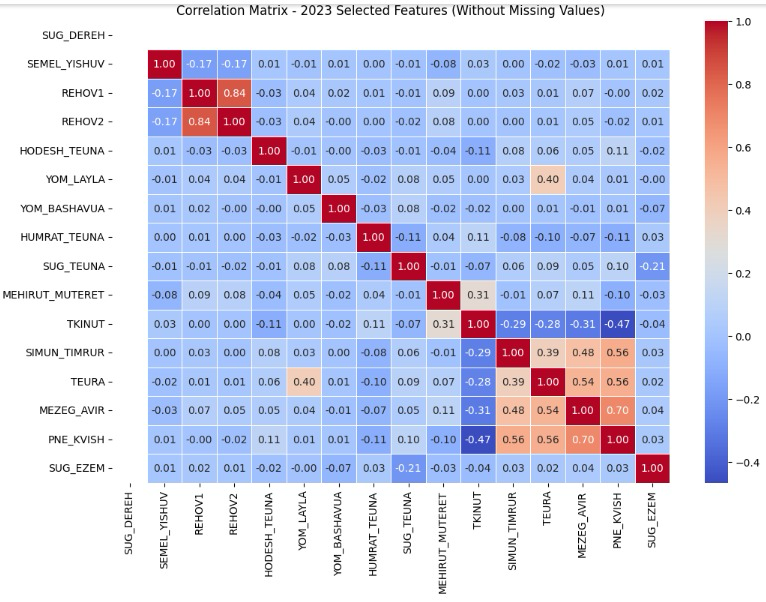
* ימי חמישי וראשון נראים כבולטים במיוחד בכמות התאונות לשנת 2022. לעומת זאת, ב-2023 ניתן לראות ירידה מסוימת בימים אלו.
* ירידה או עלייה בכמות תאונות: לדוגמה, יום שישי רשם ירידה בכמות התאונות בין 2022 ל-2023, מה שמרמז על אפשרות של שינוי בהתנהגות הנהגים או עומסים שונים ביום זה.
* זיהוי ימים פחות מסוכנים: יום שבת הוא יום שבו כמות התאונות היא הנמוכה ביותר בשני השנים. תובנה זו יכולה להועיל בקביעת מדיניות או פעולות אכיפה מוגברת ביום זה.

ֿ כיצד הגרף תורם לפרויקט:

* תובנות מעשיות: ניתן להשתמש בתוצאות על מנת להמליץ על פעולות מניעה ממוקדות, כמו הגברת אכיפה בימים מועדים לפורענות.
* השוואה רחבה: מאפשרת זיהוי תבניות כלליות במערכת נתונים בצורה פשוטה וויזואלית.
* הנחיות להמשך ניתוח: תובנות ממפת החום יכולות להנחות לניתוח מעמיק נוסף, כמו השפעת תנאי מזג אוויר או שעות עומס בימים שנמצאו כבעייתיים.

מסקנות וצעדים לעתיד:

* ניתן לבצע ניתוח נוסף של הקשרים בין שעות עומס ותנאי מזג אוויר לימים שבהם כמות התאונות גבוהה.
* כחול כהה מצביע על ימי שיא בתאונות, שבהם נדרשת תשומת לב מוגברת ואולי גם התערבות.
* צהוב בהיר מצביע על ימים רגועים יותר, שבהם ניתן ללמוד על גורמים שמפחיתים תאונות, כגון תנועה דלילה יותר (לדוגמה, מנוחה בשבת).

ניתוח גרף שלישי : מטריצת מתאם - השוואת משתנים נבחרים בין השנים 2022 ל-2023

מטריצת המתאם (Correlation Matrix) היא כלי סטטיסטי חשוב בניתוח נתונים, שמטרתה למדוד את הקשר הלינארי בין משתנים שונים. המתאם נע בטווח שבין -1 ל-1, כאשר:

* מתאם חיובי גבוה (קרוב ל-1) מציין קשר ישיר בין המשתנים, כלומר כשמשתנה אחד עולה גם השני נוטה לעלות.
* מתאם שלילי גבוה (קרוב ל--1) מציין קשר הפוך בין המשתנים, כלומר כשמשתנה אחד עולה השני נוטה לרדת.
* מתאם קרוב ל-0 מציין שאין קשר לינארי משמעותי בין המשתנים.

מטרת השימוש במטריצת מתאם:

1. זיהוי קשרים חזקים בין משתנים - מאפשר לזהות משתנים שיכולים להיות תלויים אחד בשני או להשפיע זה על זה.
2. סינון משתנים לא רלוונטיים - בעבודה עם מודלים של Machine Learning, משתנים עם מתאם גבוה ביניהם עשויים להיות מיותרים (Multicollinearity).
3. בחירת משתנים חשובים - עוזר לבחור משתנים שיכולים להסביר טוב יותר את התופעה הנחקרת.

משמעות הצבעים במטריצת המתאם:

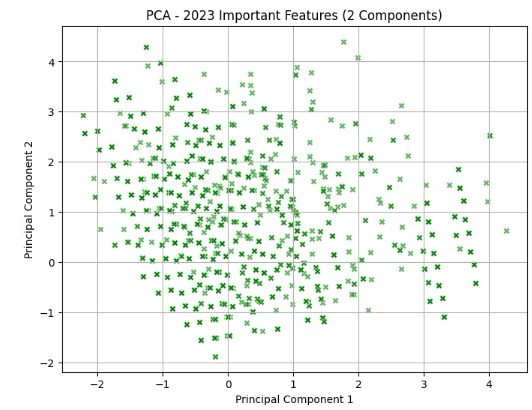
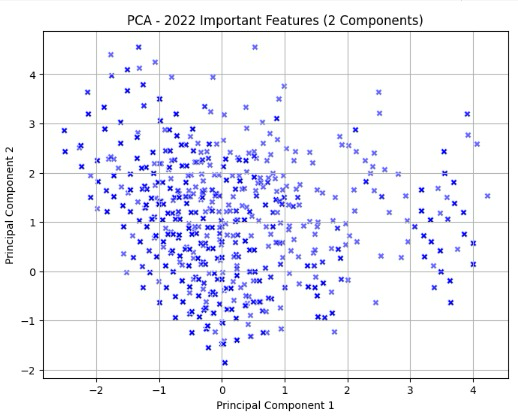
* אדום כהה: מתאם חיובי חזק מאוד (קרוב ל-1).
* כחול כהה: מתאם שלילי חזק מאוד (קרוב ל--1).
* צבעים בהירים: מתאם חלש או לא קיים (קרוב ל-0).

ניתוח המטריצה :

* הבדלים בין השנים: במטריצה של 2022 ושל 2023, ניתן לראות קשרים שונים בין משתנים, מה שמספק תובנות לגבי שינויים דינמיים בתאונות.
* משתנים עם קשר חזק: ניתן לראות, למשל, קשרים חיוביים חזקים בין REHOV1 ל-REHOV2 בשתי השנים. הדבר מצביע על כך שהמשתנים הללו עשויים להיות קשורים גאוגרפית או מבניים.
* משתנים עם קשר שלילי: הקשר השלילי בין מהירות מותרת (MEHIRUT\_MUTERET) לבין משתנים אחרים, כמו חומרת התאונה (HUMRAT\_TEUNA), עשוי לרמוז על השפעת המהירות על חומרת התאונה.

מסקנות וצעדים לעתיד:

* באמצעות המטריצה נוכל לזהות משתנים אשר משפיעים משמעותית על תאונות הדרכים.
* בבוא העת נוכל לסנן משתנים שאינם מוסיפים מידע חדש.
* נצליח להבין קשרים בין משתנים אשר בעתיד יוכלו לשפר את ניתוחי המודל שלנו.

 ניתוח גרף 4: PCA לשנים 2022 ו-2023

גרפי ה-PCA לשנים 2022 ו-2023 ממחישים את הפיזור של תאונות הדרכים בהתבסס על מאפיינים עיקריים שנבחרו מתוך הדאטה. מיפוי הרכיבים איפשר לנו לצמצם את המידע הרב לשני צירים עיקריים (רכיב עיקרי 1 ורכיב עיקרי 2) כך שניתן לנתח את הנתונים בצורה ויזואלית ברורה.

מטרת גרף ה-PCA:

* הבנת הקשרים המורכבים בין המשתנים השונים ותיאורם בצמצום ממדים.
* זיהוי תבניות דומיננטיות או התנהגויות יוצאות דופן בנתונים.
* השוואה בין השנים - באמצעות מיפוי בשני גרפים נפרדים ניתן לזהות שינויים בהתנהגות התאונות.

מה ניתן ללמוד מהגרפים:

* פיזור נתונים: הפיזור המרוכז במרכז מצביע על דפוסי תאונות דומים בין רוב התאונות, בעוד הנקודות הרחוקות ממרכז הפיזור עשויות להצביע על תאונות יוצאות דופן עם מאפיינים חריגים.
* השוואה בין שנים: ניתן לראות אם פיזור הנתונים השתנה בין השנים ולזהות מגמות. לדוגמה, אם ב-2023 קיימת התפשטות רחבה יותר של הנקודות, הדבר עשוי להעיד על שינוי במאפייני התאונות.
* חיזוי: אם קיימות קבוצות ברורות של נקודות (Cluster), ניתן יהיה להשתמש במידע זה כדי לפתח מודלים שינתחו את הקבוצות השונות.

המשמעות של ה-PCA בפרויקט:

* חידוד משתנים: גרף ה-PCA מאפשר להעריך האם הבחירה במאפיינים שנבחרו באמת מייצגת את הנתונים בצורה מיטבית.
* בסיס לניתוחים מתקדמים: התוצאות יכולות לשמש לבניית מודלים חיזויים, כמו קלסיפיקציה של סוג תאונה.
* ניטור שינויים בין שנים: הבדלים במבנה הפיזור יכולים להצביע על שינוי מגמות, מדיניות או נסיבות סביבתיות בין השנים.

מה רואים בגרפים בפועל?

שנת 2022:

רוב הנקודות קרובות יותר למרכז הפיזור, מה שמעיד על כך שרוב התאונות בשנת 2022 היו קרובות לממוצע מבחינת התכונות שנבחרו (חומרת התאונה, סוג הדרך, מזג האוויר, ועוד).

עם זאת, ישנם ערכי X חזקים ורחוקים מהמרכז, המייצגים תאונות קיצוניות או חריגות מבחינת המאפיינים שלהן.

שנת 2023:

בגרף ה-PCA לשנת 2023, הנקודות נראות מפוזרות באופן דומה, אך יש נטייה של חלק מהנקודות להיות מרוכזות בצורה צפופה יותר באזורים מסוימים.

גם כאן ניכרים ערכי X חזקים בקצוות, שמרמזים על מקרים יוצאי דופן, ייתכן שתאונות בתנאים חריגים או שילובים פחות שכיחים של מאפיינים.

גרף חמישי: השפעת המהירות המותרת על חומרת התאונה (2022 ו-2023)

גרף התיבה (Boxplot) מציג את ההשפעה של מהירות מותרת על חומרת התאונה. זהו גרף סטטיסטי המאפשר לנו להבין את התפלגות הנתונים עבור כל קטגוריה של מהירות, באמצעות תיאור גרפי של הערכים המרכזיים, הפריסה של הנתונים, וחריגים.

אלמנטים בולטים בגרף:

* ציר ה-X: מייצג את קטגוריות המהירויות המותרות (למשל עד 50 קמ"ש, 60 קמ"ש).
* ציר ה-Y: מייצג את רמות חומרת התאונה (קטלנית, קשה, קלה).
* תיבות (Box): מתארות את אחוזוני הנתונים (25%-75%).
* קו בתוך התיבה: מייצג את החציון (Median), שהוא הערך המרכזי.
* קווי תוחלת (Whiskers): מראים את טווח הנתונים ללא חריגים.
* נקודות בודדות: מייצגות חריגים - מקרים שבהם הנתונים חורגים מהטווח הצפוי.

מגמת ההתפלגות:

* כל קטגוריות המהירות מציגות תיבה רחבה יחסית, דבר המצביע על התפלגות רחבה של חומרת תאונות עבור כל קטגוריה.
* רוב התאונות הן בדרגת חומרה "קלה" (Light), אך ישנן גם תאונות בדרגת "קשה" ו"קטלנית" בכל מהירויות.

קטגוריות ספציפיות:

* עבור קטגוריות "Up to 50 km/h" ו-"Unknown", ההתפלגות מציגה דפוס דומה מאוד, דבר שעשוי להעיד על דמיון בין המקרים שבהם המהירות אינה ידועה לבין מהירויות נמוכות.
* בקטגוריות מהירויות גבוהות יותר (90 קמ"ש ו-110 קמ"ש), ניתן לראות שעדיין יש אחוז גבוה של תאונות בדרגה קלה, אך ייתכן שיש חריגים נוספים בדרגת קשה.

ערכים חריגים:

* נקודות חריגות המייצגות תאונות קטלניות (Fatal) מופיעות בכל הקטגוריות, דבר שמעיד על כך שחומרת התאונה אינה תלויה אך ורק במהירות.

מסקנות וצעדים לעתיד:

* השפעה מוגבלת של מהירות מותרת על חומרת התאונה - למרות שמהירות גבוהה נחשבת למסוכנת יותר, רוב התאונות הקלות מתרחשות גם במהירויות נמוכות. ייתכן שהגורם המשמעותי הוא תנאי דרך ולא המהירות בלבד.
* צורך בניתוח נוסף - ייתכן שמהירות לבד אינה מספיקה להסיק מסקנות משמעותיות, וצריך לשלב פרמטרים נוספים כמו תנאי מזג אוויר או סוג הכביש.
* העדר תלות ברורה במהירות - הגרף מדגיש את היעדר קשר חד משמעי בין המהירות המותרת לבין חומרת התאונה, מה שאומר כי יש גורמים נוספים המשפיעים על תוצאות התאונה.

יתרונות הגרף

* ויזואליזציה של התפלגות: הגרף מאפשר לראות את ההתפלגות המלאה של חומרת התאונות בכל קטגוריית מהירות.
* הצגת חריגים: נקודות חריגות מבליטות מקרים יוצאי דופן שדורשים תשומת לב מיוחדת.
* השוואה ברורה בין קטגוריות: ניתן לראות את הדמיון והשוני בין קטגוריות המהירויות השונות.

1. איכות הנתונים

מאגר הנתונים שלנו מכיל בין היתר חוסר עקביות - ערכים חסרים, שגיאות שונות וערכים לא ידועים, כל אלו מקשים על תהליך הניתוח. בעוד שמדובר באחד הכלים החשובים ביותר להבנת תופעות מורכבות כמו תאונות דרכים וזיהוי דפוסים, איכות הנתונים היא קריטית. כאשר נתונים גולמיים אינם עוברים תהליך ניקוי ובקרה, הדבר עלול להוביל למסקנות שגויות, מודלי חיזוי לא מדויקים ואף לפספוס של תובנות קריטיות. לכן, אנו בוחנות באופן יסודי את הנתונים הזמינים, מזהים בעיות פוטנציאליות ומתמודדים עימן לפני תחילת עבודת הקמת המודלים והניתוחים המתקדמים. בעיות נפוצות שעלולות לצוץ כוללות נתונים חסרים, שגיאות הקלדה, חוסר עקביות בקידוד, מטא-נתונים לא מדויקים ושגיאות מדידה הן דברים שקיימים כמעט בכל מאגר נתונים, לכן אנחנו סוקרות אותו באופן מלא.

**נתונים חסרים:**

* קואורדינטות גיאוגרפיות (X, Y): מיקום התאונה חסר או לא נרשם במדויק.
* שמות רחובות: רחוב אחד או יותר אינם מצוינים בקובץ.
* סוג התאונה: ערכים ריקים בסיווג סוג התאונה – למשל תאונה חזיתית או פגיעת הולך רגל.
* נתונים כרונולוגיים: שדות כמו שעה, חודש או שנה עלולים להכיל ערכים חסרים.

איך נתמודד:

* איתור: שימוש בכלים כמו Python או Excel לזיהוי ערכים חסרים פקודותisnull()  או סינון ערכים ריקים.
* השלמה: ניתן להחליף נתונים חסרים בערכים ממוצעים (בשדות מספריים כמו מהירות מותרת).
* חיזוק המודל: להתעלם מנתונים חסרים בשדות לא קריטיים או לשלב טכניקות חיזוי להשלמת ערכים חסרים.

**חוסר עקביות בקידוד:**

חוסר עקביות בקידוד הוא בעיה נפוצה בשדות קטגוריים. לדוגמה:

* מגדר: שימוש ב-"M" וב-"male" לציון גברים או ב-"F" וב-"female" לציון נשים.
* תאורה: במקום ערך אחיד כמו תאורת רחוב יש גם רחוב מואר.

איך נתמודד:

* איחוד ערכים: להמיר את כל הערכים הלא אחידים למערכת קידוד אחידה.
* הגדרת קטגוריות ברורות: שימוש בטבלאות חיפוש או כלים אוטומטיים להסדרת הקידוד.