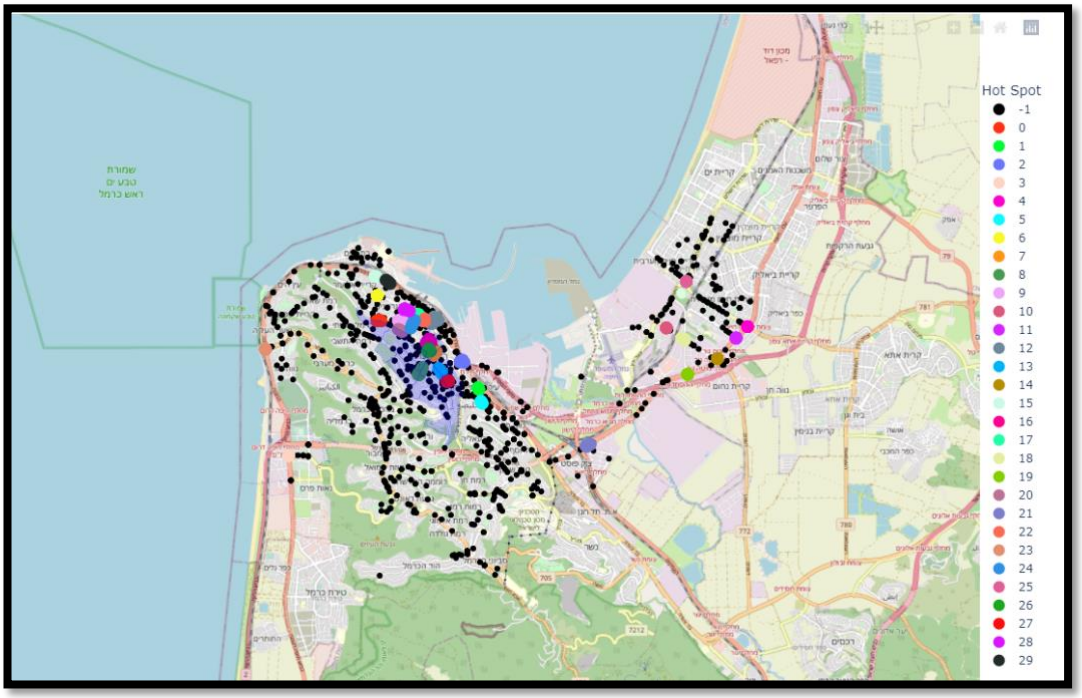
זיהוי וניתוח Hot Spots בחיפה

מערכות נבונות אינטראקטיביות 096235

דו"ח פרויקט סופי



מגישים:

משה עבאדי 324658939

רועי רימר 314828732

קישור ל-GitHub של הפרויקט:

https://github.com/moshinhoabadi/Hotspot-Identification-and-Analysis

מבוא:

תאונות דרכים הן דבר שללא ספק היינו רוצים למנוע. על פי נתוני ארגון הבריאות העולמי, בכל שנה כמיליון איש נהרגים ועשרות מיליונים נפצעים מתאונות דרכים ברחבי העולם. בישראל ישנה מגמת ירידה בכמות ההרוגים בתאונות דרכים בשנים האחרונות, אך עדיין מאות אנשים נהרגים בשנה. בשנת 2020 נהרגו 302 בני אדם, ובסך הכול מאז הקמת המדינה נהרגו יותר מ-30 אלף איש בתאונות דרכים.

ישנם כבישים ומקטעי דרך שעלולים להיות מועדים יותר לתאונות דרכים עקב תנאי מערכת הכבישים (למשל צומת מסוכן), או בגלל תנאי השטח (תאורה לקויה, מחסור בתמרורים ועוד). בפרויקט שלנו, התמקדנו בנתונים אודות תאונות דרכים בעיר חיפה, ובפרט בשכונת הדר, וזיהינו קטעי דרך מסוכנים. משימה זו נקראת בספרות Hot/Black Spot Identification. כמו כן, ניסינו להצביע על מאפיינים בתנאי השטח שעשויים להיות קשורים לריבוי היחסי של מקרי התאונות ב-Hot Spots שהתגלו, או ב-Hot Spot ספציפי.

אנו השתמשנו בנתונים מתוך אתר הלמ"ס על תאונות דרכים ברחבי הארץ בין השנים 2017-2019. נתונים אלו מכילים פרטים על מיקום התאונה, סוג התאונה, דרגת חומרתה, המהירות המותרת באותו כביש, תמרורים שהיו בסביבה ועוד.

ישנן גישות שונות להתמודדות עם משימת ה-Hot Spot Identification. הגישה אותה בחרנו ליישם היא Clustering באמצעות שימוש באלגוריתם DBSCAN, כאשר האלגוריתם מקבל כקלט את הקואורדינטות של מיקומי התאונות. האלגוריתם מזהה את האזורים המסוכנים על ידי איתור מקומות בעלי צפיפות גבוהה של אירועי תאונות דרכים.
בנוסף, השתמשנו במודל Association Rule אשר קיבל נתונים על תנאי השטח והכבישים בכדי לזהות מאפיינים המתקשרים ל-Hot Spot נקודתי שזוהה על ידי DBSCAN, או לכלל ה-Hot Spots.

על סמך התוצאות והתובנות מן המודל, בנינו Dashboard אינטראקטיבי אשר מציג את תוצאות ה-Clustering על מפה, עם אפשרות לסנן את התאונות שמוצגות לפי דרגת חומרתן ולפי סוג התאונה, כאשר ישנה גם אפשרות להציג תאונות שקרו בהדר בלבד. כמו כן, ב-Dashboard מוצגים ה-Association Rules שנמצאו עבור התאונות שנמצאות ב-Hot Spots, כאשר בחירת Hot Spot ספציפי תציג את החוקים והנתונים הרלוונטיים עבורו.

בעזרת אלגוריתם DBSCAN מצאנו 30 אזורים שזוהו כ-Hot Spots ברחבי העיר חיפה, כאשר 8 מתוכם נמצאים בתחומי הדר או גובלות בה. בהשוואה עם בדיקה שביצעה עמותת אור ירוק עבור 15 הצמתים המסוכנים ביותר בחיפה בשנים 2017-2019, 12 מן 15 הצמתים (80%) זוהו כ-Hot Spot גם על ידי האלגוריתם שהפעלנו. בנוסף, ייצרנו חוקים עבור ה-Hot Spots יחד ולחוד, כמו גם עבור אוסף התאונות שלא שויכו לאף Hot Spot.
את ה-Dashboard שיצרנו ניתן יהיה לשלב כשכבה על גבי ה-GIS של החמ"ל החברתי, והוא יוכל לסייע למקבלי ההחלטות באיתור וניתוח האזורים המסוכנים בהדר מבחינת בטיחות בדרכים.

סקירת ספרות:

קיימות בספרות מספר גישות לפתרון בעיית ה-Hot Spot Identification. כאשר המידע הזמין אינו כולל נתונים כמו נפח התנועה אלא רק את מיקומי, כמויות וחומרת התאונות (בדומה לנתונים שזמינים עבורנו), ישנן שיטות מבוססות דירוג המתחשבות רק בכמות התאונות שהתרחשו בהן ובחומרתן [1]. כל שיטה מתייחסת לכל אחד משני הגורמים האלו באופן שונה. כך למשל שיטת Crash Frequency מדרגת את האוזרים תוך כדי התייחסות אך ורק לכמות התאונות שהתרחשו בכל אזור, ואילו שיטה כמו Equivalent Property Damage Only נותנת ניקוד שונה לכל סוג תאונה על פי רמת חומרתה, ועל סמך ניקוד זה נבנה הדירוג של האזורים השונים.

גישה אחרת להתמודדות עם בעיית Hot Spot Identification מתבוסת על איתור אזורים (בעיקר צמתים ומקטעי רחובות) בעלי צפיפות גבוהה של תאונות ביחס לכמות השנים שבמאגר הנתונים, זאת על ידי שימוש באלגוריתם clustering מבוסס צפיפות כמו DBSCAN [2]. אלגוריתם זה איננו מקבל מספר מוגדר של clusters מראש, ואיננו משייך בהכרח כל תצפית ל-cluster, אלא מאגד ל-clusters רק קבוצות של תצפיות שנמצאות באזורים בעלי צפיפות גבוהה יחסית.
ההגדרה לצפיפות גבוהה נקבעת על ידי פרמטר של DBSCAN בשם epsilon. פרמטר זה מגדיר את המרחק המקסימלי בו תצפיות שסמוכות לתצפיות מסוימת נחשבות כשכנות שלו. מכיוון שלפרמטר זה השפעה חיונית על תוצאות האלגוריתם, ישנתן התייחסויות בספרות לאופן בחירתו על סמך הנתונים, למשל ב-[3]. השיטה במאמר זה מבוסת על התייחסות למרחק של כל תצפית מהשכן ה-k הקרוב ביותר שלב (עבור k כלשהו), מיון מרחקים אלו בסדר עולה והצגתם בגרף, ומציאת ערך סף של epsilon על סמך שיפוע הגרף.

ב-[4] נֶעשה שימוש באלגוריתם DBSCAN שקיבל כקלט את קואורדינטות ה-GPS של התאונות כדי לאתר Hot Spots. במאמר מצוין שלשימוש בגישה זו ישנו יתרון על פני גישות שמתבוסות על דירוג אזורים מוגדרים מראש כמו רחובות, מכיוון שמרבית התאונות העירוניות מתרחשות בצמתים. גישה שקובעת את האזורים המועמדים מראש עלולה לפספס את הצמתים האלו ולא להתייחס אליהן כאל אזור מסוכן, עקב כך שהתאונות שהתרחשו בצומת כלשהו עלולות להתחלק בין האזורים השונים שנפגשים בצומת.

ישנן שיטות לשערוך גישות של Hot Spot Identification, המאפשרות לבדוק את רמת האמינות והעקביות של גישה מסוימת אל מול גישה אחרת [5]. חלק מהשיטות מערבות ידע אמיתי של מומחים על אלו אזורים הם אכן Hot Spots ואלו לא. עם זאת, קיימות גם שיטות שלא מסתמכות על כך, ועיקרן הוא השוואה בין שני קטעי זמן (למשל בין שנה כלשהי לשנה העוקבת לה) והשוואה בין ה-Hot Spots שהמודל גילה בשתי נקודות הזמן האלו, תוך כדי הנחה שלא נעשו שינויים משמעותיים בתנאי הדרך בין שתי תקופות הזמן האלו.

בעבודה שלנו, נרצה בין היתר לנסות ולזהות גורמים שעלולים להשפיע על ריבוי תאונות הדרכים באזורים בחיפה שנזהה כ-Hot Spots. ישנם כל מיני גורמים שקשורים בתנאי הכביש הסביבה שעלולים להשפיע על רמת הסיכון לתאונת דרכים באזור מסוים. למשל, בניתוח שנעשה עבור נתונים על תאונות דרכים בבלגיה [6], נמצא שתאונות נוטות להתרחש בצמתים בעלי פנייה שמאלה. כמו כן,

נעשתה השוואה בניתוח זה בין תאונות שקרו ב-Hot Spots לעומת כאלו שהתרחשו מחוץ לאזורים אלו, ונמצא שמאפיינים כמו תנאים ליליים ומשטח רטוב אפיינו תאונות שהתרחשו ב-Hot Spots יותר מאשר תאונות שהתרחשו באזורים אחרים, כנראה עקב יכולת מופחתת ב-Hot Spots להתמודד עם תנאים של חשיכה או כביש רטוב מבחינת התשתית הקיימת.

מודל ה-Association Rule [7] מאפשר לזהות ולאפיין קשרים בין ערכים שונים בסט הנתונים. למודל זה מגוון יישומים, למשל עבור Market Basket Analysis [8], כלומר עבור חקירת ומציאת קשרים בין מוצרים שנוטים להופיע יחד בסלי קניות של לקוחות.
למודל ה-Association Rule יש יישומים בספרות גם בנושא של אפיון Hot Spots. ב-[9] נותחו נתונים של תאונות מהאזור Flanders שבבלגיה, ובפרט של חלקים ב-Flanders שמוגדרים כ- Hot Spots. בתהליך של Data mining נוצרו חוקים עבור תאונות שהתרחשו ב-Hot Spots, ועבור תאונות שהתרחשו באזורים אחרים. עבור חוקים שהיו משותפים לשני המקרים נעשתה השוואה כדי לבדוק האם ניתן לומר שחוקים אלו נכונים יותר ל-Hot Spots ולכן מאפיינים אותם יותר לעומת תאונות מחוץ ל-Hot Spots.
בניתוח דומה שנעשה עבור העיר Elabuga שברוסיה [10], אותרו שני Hot Spots בעלי כמות רבה של תאונות, זאת למרות שהקיבולת שלהן מבחינת כמות מכוניות היא יחסית נמוכה (איננה חוצה את ה-35%). לאחר מכן, יוצרו Association Rules עבור כל אחד משני ה-Hot Spots הללו. כך זוהו למשל מקרים לא מועטים של תאונות בהן ירד גשם באותו הזמן, או שהתרחשו בחורף בתנאים לקויים (למשל שלג שלא פונה כראוי).

שיטות:

סט הנתונים בו השתמשנו נטען מתוך קובץ המכיל נתונים שלקוחים מאתר הלמ"ס אודות 1158 תאונות שהתרחשו בעיר חיפה בין השנים 2017-2019. נתונים אלו כוללים פרטים על מיקום התאונה, סוג התאונה, דרגת חומרתה, מהירות מותרת באותו כביש, תמרורים שהיו בסביבה ועוד. עם זאת, בנינו את הפרויקט באופן שאיננו תלוי בסט הנתונים הספציפי הזה, כך שניתן לטעון קובץ אחר (בעל מבנה עמודות זהה), באמצעות הזנת נתיב הקובץ הרצוי במהלך הרצת הקוד דרך ה-command line (הסבר בנספח שבסוף המסמך). כברירת מחדל ייטען הקובץ בו אנו השתמשנו.

הפרויקט שלנו מורכב משלושה חלקים עיקריים:

1. <u>מציאת Hot Spots</u> על ידי שימוש באלגוריתם DBSCAN + בחירת פרמטרים מתאימים באופן אוטומטי לאלגוריתם על סמך סט הנתונים.
2. <u>מציאת קשרים בין משתנים בסט הנתונים לבין תאונות שמתרחשות ב-Hot Spots</u>, ובפרט עבור תאונות שהתרחשו ב-Hot Spot ספציפי, זאת על ידי שימוש במודל Association Rule.
3. <u>הצגת התוצאות בדאשבורד אינטראקטיבי</u> המאפשר סקירה וניתוח של הממצאים.

המימוש של שלושת החלקים נעשה בשפת python. לצורך מימוש אלגוריתם DBSCAN השתמשנו בספריית sklearn, כאשר מימשנו פונקציות לבחירה האוטומטית של הפרמטרים (תוצג בהמשך).

לצורך שימוש במודל ה-Association Rules השתמשנו בספרייה mlxtend, והכנסנו למודל משתנים להם ערכים שקשר בינם לבין אזור מועד לתאונות עשוי להיות מעניין. לצורך יצירת הדאשבורד השתמשנו בספרייה plotly.

נפרט כעת על כל אחד משלושת החלקים, ועל השיטות בהן השתמשנו במהלכן:

1. מציאת Hot Spots

את זיהוי ה-Hot Spots בסט הנתונים שלנו ביצענו על ידי שימוש באלגוריתם DBSCAN. נתאר את פעולת האלגוריתם במילים, ולאחר מכן נציג גם pseudo-code מתאים:

DBSCAN מקבל כקלט סט של וקטורים (התצפיות), פרמטר בשם MinPts אשר קובע את מספר התצפיות המינימלי האפשרי בקלאסטר, ואת הפרמטר epsilon שקובע את רדיוס ה"שכונה" של תצפיות מסוימת.
האלגוריתם בוחר תצפית P באופן שרירותי, ובודק האם לתצפית זו יש לפחות MinPts שכנים (כולל עצמה) בשכונה שלה, שמוגדרת על ידי רדיוס של epsilon מן התצפית. במידה ולא, תצפית זו תסווג כרעש (אם כי זה עשוי להשתנות בהמשך הריצה אם תימצא קרובה מספיק לקלאסטר שייווצר).
במידה ולתצפיות נמצאו לפחות MinPts שכנים, כל התצפיות האלו ביחד יוצרות cluster חדש. כעת, נעשה המעבר הבא על כל השכנים של התצפית P שהאלגוריתם טרם עבר עליהם במהלך ריצתו: אם לאחד השכנים האלו (נסמן ב-P') יש גם כן לפחות MinPts שכנים (כולל עצמו), אז כל שכנים אלו מתווספים גם כן ל-cluster ולרשימת התצפיות שייעשה עליהם המעבר שלעיל.
כאשר נעשה המעבר על כל התצפיות עליהן הוא היה צריך להתבצע, נבחרת תצפיות שרירותית P חדשה בה עוד לא ביקרו במהלך ריצת האלגוריתם, והתהליך שלעיל מתבצע מחדש עליה. כך ממשיכים עד אשר האלגוריתם מבקר בכל התצפיות שבסט הנתונים.

איור 1 מציג pseudo-code של ריצת אלגוריתם DBSCAN:

```
DBSCAN(D, epsilon, MinPts):
    C = 0
    for each unvisited point P in dataset D:
        mark P as visited
        NeighborPts = regionQuery(P, epsilon)
        if sizeof(NeighborPts) < MinPts:
            mark P as NOISE
        else:
            C = next cluster
            expandCluster(P, NeighborPts, C, epsilon, MinPts)

expandCluster(P, NeighborPts, C, epsilon, MinPts):
    add P to cluster C
    for each point P' in NeighborPts:
        if P' is not visited:
            mark P' as visited
            NeighborPts' = regionQuery(P', epsilon)
            if sizeof(NeighborPts') >= MinPts:
                NeighborPts = NeighborPts joined with NeighborPts'
        if P' is not yet member of any cluster:
            add P' to cluster C

regionQuery(P, epsilon):
    return all points within P's epsilon-neighborhood (including P)
```

איור 1 - pseudo-code של ריצת אלגוריתם DBSCAN.

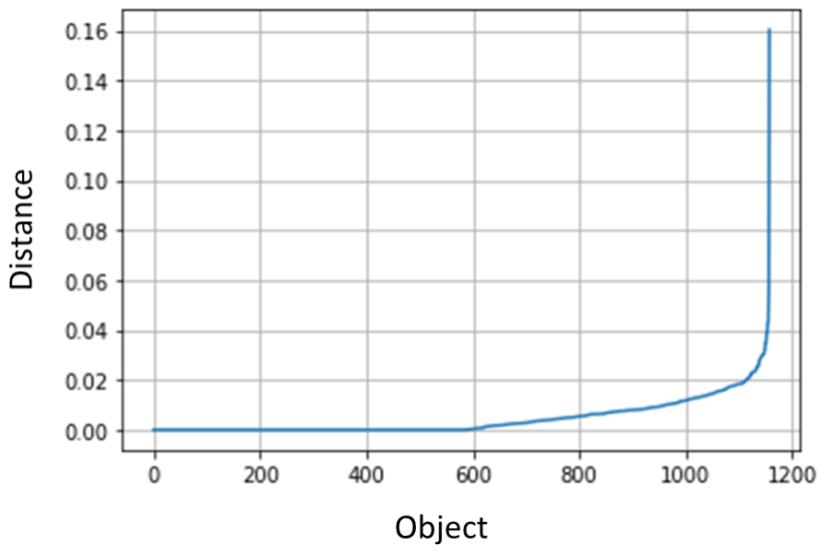
בחרנו להשתמש באלגוריתם DBSCAN כאלגוריתם ה-clustering משתי סיבות עיקריות. ראשית, אלגוריתם זה איננו צריך לקבל כקלט מראש את מספר ה-clusters שאמורים להיווצר. איננו יכולים לקבוע מראש כמה Hot Spots יש בסט הנתונים, כך שזהו גורם חשוב המבדיל את DBSCAN מכמה אלגוריתמי clustering אחרים דוגמת K-Means.
בנוסף, DBSCAN לא משייך כל תצפית בסט הנתונים ל-cluster, אלא מזהה גם תצפיות שהן "רעש", כלומר אינן שייכות לאף cluster. זוהי גם תכונה חשובה עבור צרכינו, מכיוון שאיננו רוצים שכל תאונה שהתרחשה תשתייך ל-Hot Spot. אנו רוצים למצוא רק אזורים ממוקדים שאכן מועדים לסכנה מבחינת ריבוי התאונות היחסי בהם. נציין שתכונה זו גורמת לאלגוריתם להיות חסין ל-outliers, וזהו דבר חיובי גם כן.

החיסרון העיקרי באלגוריתם זה הוא הצורך בקביעה של הפרמטרים MinPts, epsilon על ידי המשתמש. לשני הפרמטרים האלו השפעה רבה על תוצאות האלגוריתם, ולכן האלגוריתם רגיש לערכים שהוא מקבל כקלט עבור פרמטרים אלו. בעוד MinPts הוא פרמטר שלמשתמש יותר קל לבחור על פי צרכיו, בחירת הערך לפרמטר epsilon היא פחות אינטואיטיבית. נציין כעת את הנתונים שהזנו לאלגוריתם DBSCAN, כמו גם את תהליך בחירת ערכי MinPts ו-eps:

לאלגוריתם DBSCAN הזנו כקלט את הקואורדינטות (המנורמלות) של מיקומי כל התאונות שבסט הנתונים.
את הפרמטר MinPts (מספר התצפיות המינימלי לתחילת יצירת קלאסטר) אנו בוחרים כ-$2n + 1$ כאשר n הוא מספר השנים הנסקרות בסט הנתונים. ישנן בספרות שלל הגדרות ל-Hot Spot, כאשר ההגדרה משתנה גם לפי גודל האזור. רוב ההגדרות האלו נעות סביב טווח של 3-4 תאונות באותה שנה, או כ-10 תאונות בטווח של שלוש שנים. מכיוון שאנו רוצים ליצור Hot Spots קטנים וממוקדים, ברמה של צמתים ומקטעי רחובות (רוב ההגדרות בספרות מתייחסות לאזורים נרחבים יותר), החלטנו שזו תהיה ההגדרה המתאימה עבורנו.

את הקביעה של הפרמטר epsilon אנו מבצעים גם כן באופן אוטומטי המבוסס על סט הנתונים. פרמטר זה הוא בעל השפעה רבה על תוצאות האלגוריתם מבחינת גדלי ה-clusters וצפיפותם, ולכן יש לבחור אותו באופן שמתאים לרצון שלנו במציאת Hot Spots ממוקדים.
האופן בו אנו בוחרים את epsilon מבוסס על האמור ב-[3], אם כי הגדרנו תנאי עצירה שונה בעקבות ביצוע של כמה ניסויים, מכיוון שתנאי העצירה המובא במאמר גרר במקרה שלנו קבלת clusters גדולים מדי לצרכינו. אופן בחירת epsilon נעשה באופן הבא:

- ראשית, חישבנו עבור כל תצפית בסט הנתונים את המרחק מהשכן הקרוב ביותר אליה (שאינו היא עצמה).
- לאחר מכן, מיינו ערכים אלו ויצרנו מהם גרף, כאשר ציר x הוא המספר הסידורי של הערך ברשימה הממוינת (1, 2 וכן הלאה), וציר ה-y הוא הערך עצמו. איור 2 מציג דוגמא ליצירת גרף כזה על סמך סט הנתונים שלנו:

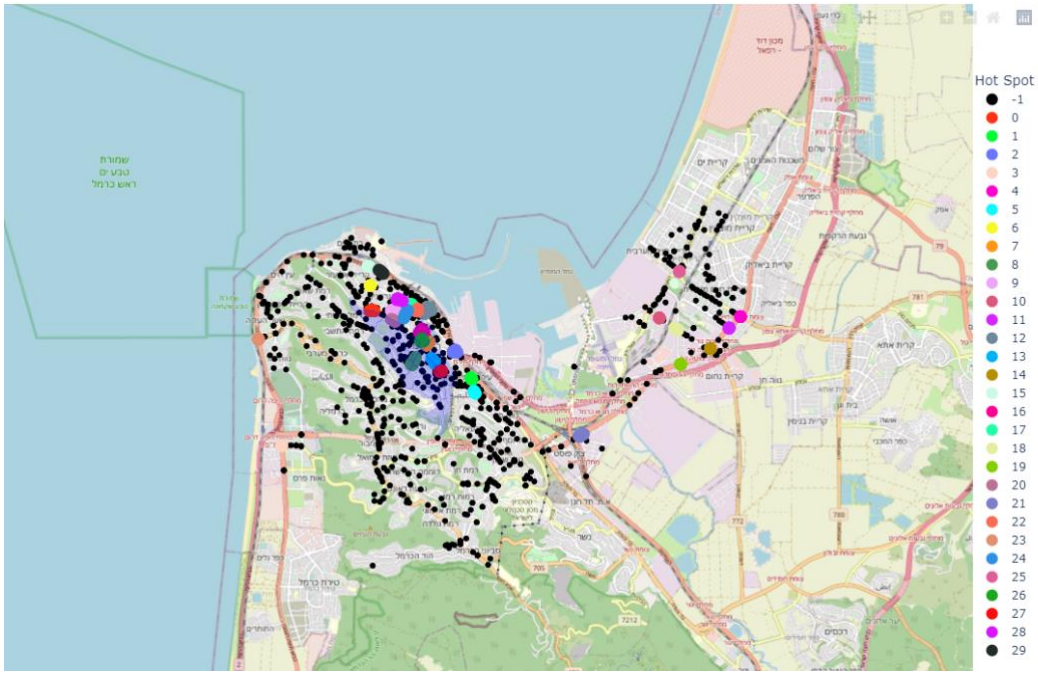


איור 2 – גרף ערכי מרחקים מהשכן הקרוב ביותר.

- חישבנו את השיפוע בין כל שתי נקודות סמוכות בגרף, ושמרנו את השיפועים לפי סדרם בגרף.
- עברנו על השיפועים לפי סדרם, ועצרנו כאשר נמצא רצף של שישה ערכי שיפועים אשר לפחות שלושה מהם היו גדולים ביותר מאחוז אחד לעומת ערך השיפוע הקודם להם. במקרה זה התמקדנו בשיפוע הראשון באותו רצף, הסתכלנו על הנקודה השנייה בין שתי הנקודות הסמוכות בגרף ששימשו לחישוב שיפוע זה, וערך ציר ה-y שלה נבחר להיות ערכו של epsilon.

זהו תנאי עצירה שונה ומורכב יותר מן התנאי שמצוין ב-[3], אך גם הוא מבוסס על הגידול היחסי בשיפוע הגרף ומתאים יותר לטווח ערכי epsilon שאנו מחפשים לאלגוריתם DBSCAN עבור המטרה שלנו. זאת הסקנו על סמך ניסויים שביצענו עם תנאי עצירה דומים.

לאחר שהוגדרו כל הקלטים של אלגוריתם DBSCAN עבור סט הנתונים שלנו, הרצנו אותו וקיבלנו את ה-Hot Spots שאותרו על ידי האלגוריתם, כמו גם את התאונות שלא נחשבות שייכות לאף Hot Spot. איור 3 מציג את תוצאת האלגוריתם על חלק ממפת חיפה (פירוט על התוצאות בהמשך המסמך):



איור 3 – ויזואליזציה חלקית של תוצאות אלגוריתם DBSCAN על מפת העיר חיפה.

2. מציאת קשרים בין משתנים בסט הנתונים לבין תאונות שמתרחשות ב-Hot Spots

לאחר מציאת ה-Hot Spots, השלב הבא הוא לנסות ולאתר משתנים בסט הנתונים, להם יש קשר לתאונות שהתרחשו ב-Hot Spots או לתאונות שהתרחשו ב-Hot Spot ספציפי. כדי לעשות זאת, בחרנו לבצע Association Rule Learning, כלומר שימוש במודל ה-Association Rule לצורך יצירת "חוקים" שמתרחשים בתדירות ובסבירות גבוהה מספיק בסט הנתונים. חוקים אלו עשויים באופן פוטנציאלי לסייע למצוא גורמים בסט הנתונים להם קשר עם התאונות שהתרחשו ועם גורמים אחרים, ובכך לסייע למקבלי ההחלטות בבואם לנתח את הנתונים.

נגדיר כעת את מודל ה-Association Rule, ונתאר כיצד השתמשנו בו:
ניתן לפרק כל משתנה קטגורי בסט הנתונים לאוסף של משתנים בינאריים, כמספר הערכים האפשריים למשתנה זה. נגדיר את סט כל המשתנים הבינאריים הללו, עבור כל המשתנים הקטגוריים יחדיו, כ- $I = \{i_1, i_2, \dots, i_n\}$ (בהנחה ויש n כאלו). סט הנתונים D מורכב מאוסף T של רשומות (התאונות במקרה שלנו), כאשר פרטי כל רשומה הם תת קבוצה של הפריטים ב-I.
חוק מוגדר ככלל גרירה מהצורה הבאה: $X \Rightarrow Y: X, Y \subseteq I$. אנו בחרנו ליצור רק חוקים בהם $X, Y \in I$.
כדי ליצור על סמך הקבוצה I חוקים מעניינים, נרצה לבחור חוקים בעלי משתנים שאכן הופיעו יחד פעמים רבות בסט הנתונים, ושאכן ביחס גבוה של פעמים החוק היה נכון (כלומר ש-X אכן הופיע יחד עם Y). נרצה גם לוודא שאכן ישנה תלות בין הופעת X להופעת Y יחדיו בסט הנתונים. לצורך כל אלו, נגדיר את ההגדרות הבאות:

- **Support** - אחוז הפעמים שסט משתנים נתון הופיע בסט הנתונים:

$$Support(X) = \frac{|\{X \subseteq T\}|}{|T|}$$

נרצה להתחשב בחוק כלשהו $X \Rightarrow Y$ אך ורק אם $Support(X \cup Y)$ הוא מעל ערך סף כלשהו, כלומר ש-X ו-Y אכן הופיעו מספיק פעמים יחדיו בסט הנתונים.

- **Confidence** – אחוז הפעמים שחוק כלשהו אכן התקיים:

$$Confidence(X \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)}$$

נרצה להתייחס רק לחוקים להם ערך Confidence גבוה מערך סף מסוים, כלומר שבאחוז מספיק גדול מהרשומות, כש-X הופיע אז Y הופיע יחד איתו. לא נדרוש רמת Confidence גדולה מאוד, שכן לא נוכל לצפות ששני משתנים כלשהם יופיעו יחד תמיד, אך כן נצפה לרמת ביטחון מסוימת כדי שנקבל את החוק.

- **Lift** – יחס ה-Support של חוק מסוים לעומת מה שהיה מצופה אם X ו-Y היו בלתי תלויים:

$$Lift(X \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X) * Support(Y)}$$

אם לחוק כלשהו יש Lift שקרוב מאוד ל-1, נסיק כי הסתברויות ההופעה של X ו-Y הן בלתי תלויות זו מזו, ולכן לא ניתן ליצור חוק על סמך שתי קבוצות משתנים אלו. לעומת זאת, אם ה-Lift שונה מ-1, נוכל להסיק כי יש תלות בין הופעות שתי קבוצות משתנים אלו. הקשר ייחשב חזק יותר ככל שה-Lift גדול מ-1 אם היחס בין X ל-Y הוא ישר, וככל שה-Lift מתקרב ל-0 אם זהו יחס הפוך.

אנו נחפש חוקים אשר עוברים ערך סף מסוים של Support על ידי שימוש באלגוריתם Apriori, ולאחר מכן נבחר מתוך חוקים אלו רק חוקים שעוברים ערך סף מסוים של Confidence. עבור כל חוק שנמצא, נציג את ערכי Support, Confidence, Lift שלו. נמיין את סדר הופעת החוקים בסדר יורד לפי ערך ה-Lift.

- ערך הסף שבחרנו ל-Support הוא 0.05.
- ערך הסף שבחרנו ל-Confidence הוא 0.3.

בחרנו מראש את המשתנים שיהיו בקבוצה I כך שיהיו משתנים שלדעתנו עשויים לעניין את מקבלי ההחלטות במידה ויופיעו בחוקים שיימצאו עבור ה-Hot Spots. משתנים אלו הם למשל סוגי התאונה השונים, רמות איכות שונות של סימוני הכביש והתמרורים, איכות התאורה בדרך, חומרת התאונה, האפשרויות השונות לרוחב המסלול ועוד. בנוסף, בחרנו להראות רק חוקים אשר מכילים ערך כלשהו של סוג התאונה ו/או חומרת התאונה, זאת כדי להציג חוקים שעשויים להיות רלוונטיים ומעניינים יותר.

אנו נייצר Association Rules על סמך סט הנתונים שמכיל את כל התאונות שאלגוריתם DBSCAN שייך ל-Hot Spot כלשהו. נייצר גם Association Rules לסט הנתונים שמכיל את כל התאונות שסווגו כ"רעש" (לא שייכים לשום Hot Spot) על ידי DBSCAN.

נוסף על האמור לעיל, ניסינו גם למצוא משתנים הקשורים לריבוי התאונות ב-Hot Spot ספציפי. עקב המספר המועט (יחסית) של תאונות ב-Hot Spot ספציפי, לא נוכל לסמוך על אמינותם של חוקים שייווצרו על סמך הנתונים באותו Hot Spot. במקום זאת, ניסינו למצוא ערכי משתנים יחידים שנטו להופיע בחלק ניכר מן התאונות באותו Hot Spot. כלומר, אם ב-Hot Spot ספציפי היה משתנה שה-Support שלו על סט התאונות שקרו באותו Hot Spot היה לפחות 30 אחוז, בחרנו להציג אותו. נבהיר כי עקב המספר המועט של תאונות ב-Hot Spot, קשה לקבוע מבחינת מובהקות סטטיסטית שיש תלות בין משתנים אלו לריכוז התאונות באזור. עם זאת, הגדרנו את סט המשתנים I במקרה זה להיות משתנים שהופעה חוזרת שלהם, גם אם רק 2-3 פעמים באותו Hot Spot, עשויה לעניין את מקבלי ההחלטות. משתנים כאלו הם למשל רמות שונות של תקינות הכביש ואיכות התאורה.

3. הצגת התוצאות בדאשבורד אינטראקטיבי

את הדאשבורד אנו יוצרים באמצעות חבילת plotly המותאמת לבנייה של דאשבורדים אינטראקטיביים. הרעיון מאחורי עיצוב הדאשבורד שלנו הוא ויזואליזציה של שני חלקי הפרויקט הראשונים על גבי דאשבורד אחד, בצורה שתאפשר למקבלי ההחלטות לנתח את הנתונים בצורה נוחה. יש לאפשר להתמקד ב-Hot Spots ספציפיים, כמו גם באזור שכונת הדר בלבד מכיוון שזהו האזור המרכזי מבחינת עניין עבור החמ"ל החברתי.
לכן, בחרנו לעצב את הדאשבורד כך שתופיע בו מפת חיפה, ועל גביה יופיעו מיקומי התאונות מתוך סט הנתונים. תאונות שאינן שייכות לאף Hot Spot מופיעות בצבע שחור, ותאונות שכן שייכות ל-Hot Spot מסוים צבועות בצבעים אחרים, כאשר צבע אחר ניתן לכל Hot Spot. לצד המפה, תופיע אפשרות לסנן ולהציג רק תאונות שקרו בתוך שכונת הדר או ב-Hot Spots ספציפיים. אפשרויות נוספות לסינון שבחרנו הן הצגת תאונות לפי דרגת חומרתן, זאת במקרה ומקבלי ההחלטות ירצו להתמקד למשל רק בתאונות קשות או בתאונות קטלניות בהן נהרג אדם, כמו גם הצגת תאונות לפי סוג התאונה (פגיעה בהולך רגל או אחר).

מתחת למפה ולאפשרויות הסינון, יוצגו נתונים אודות התאונות המופיעות על המפה באותה העת ואודות Association Rules שנמצאו וקשורים לתאונות. בדאשבורד תוצג טבלה המפרטת מאפיינים שונים של התאונות, אשר משתנה בהתאם לסינונים שהופעלו. בנוסף, בחירת Hot Spot ספציפי תוביל לכך שבטבלה יוצגו רק נתוני תאונות מאותו Hot Spot. לצד טבלה זו, תופיע טבלה שתציג את החוקים שנמצאו עבור תאונות שהיו שייכות ל-Hot Spot, כאשר בחירה של Hot Spot ספציפי תוביל לכך שבטבלה יוצגו המשתנים הקשורים לתאונות באותו Hot Spot. במידה וייבחרו התאונות שלא שייכות לאף Hot Spot, טבלה זו תציג החוקים שנמצאו על סמך תאונות אלו. טבלה זו תשתנה באופן דינמי בעקבות הפעלת סינונים, אם כי לא כל סינון ישפיע עליה.
שתי הטבלאות יעזרו לצופי הדאשבורד להבין ולנתח את התאונות המופיעות במפה, ובפרט את אלו שהתרחשו ב-Hot Spots ו/או בתחומי שכונת הדר. סכמה של הדאשבורד ניתן לראות באיור 4.

איור 4 – סכמה של הדאשבורד.

תוצאות:

נציג כעת את תוצאות חלקי הפרויקט שלנו באמצעות סקירת תוצאות אלגוריתם DBSCAN והתבוננות בחוקים שהתקבלו ממודל Association Rule וברכיבי הדאשבורד:

לאחר הפעלת אלגוריתם DBSCAN, קיבלנו 30 אזורים שזוהו כ-Hot Spots (תצוגה חלקית באיור 3), כאשר כל Hot Spot מכיל לפחות 7 תאונות. סך הכול, 285 מתוך 1158 התאונות שבסט הנתונים זוהו כתאונות שקרו בתוך Hot Spot. יתר 837 התאונות שאינן שייכות ל-Hot Spot כלשהו מוצגות על המפה בצבע שחור ובגודל קטן יותר מהאחרות.
בנוסף, מוצג במפה פוליגון בצבע כחול בהיר (איור 3) המתאר את השטח של הדר, כך שתאונות ו-Hot Spots שנמצאו בתוך אזור זה הם כאלו שקרו בתחומי שכונת הדר. מתוך ה-Hot Spots שזוהו, 8 נמצאים בתוך תחומי הדר או גובלים בו.

כדי לקבל מושג על טיב תהליך ה-Hot Spot Identification שביצענו, השווינו בין ה-Hot Spots שמצאנו לבין ניתוח שביצעה עמותת אור ירוק [11], בו היא מצאה את 15 הצמתים המסוכנים ביותר על פי כמות נפגעים באותו טווח שנים כמו בסט הנתונים שלנו (2017-2019). מתוך 15 הצמתים שהופיעו בניתוח, 12 אותרו גם על ידי אלגוריתם DBSCAN. האלגוריתם מצא בנוסף Hot Spots נוספים שלא מופיעים בבדיקת אור ירוק, אך ייתכן והם היו מופיעים בה אילו היו בוחרים להציג רשימה רחבה יותר של צמתים מסוכנים. נזכיר כי מספר ה-Hot Spots לא מוגדר לאלגוריתם מראש, אלא ה-Hot Spots נבחרים על פי צפיפות התאונות באזורים.
נעיר כי לעיתים שמות ה-Hot Spots בדאשבורד שיוצג בהמשך לא יחפפו במדויק לשמות הצמתים בבדיקה של עמותת אור ירוק, אך תהיה חפיפה של לפחות שם רחוב אחד של הצומת עם שם ה-Hot Spot המתאים.

לאחר שימוש במודל Association Rule, קיבלנו חוקים שונים שיוצרו עבור כלל ה-Hot Spots, עבור התאונות שלא שייכות לאף Hot Spot, ול-Hot Spots ספציפיים. טבלה 1 מציגה חלק מן החוקים שהתקבלו עבור כלל ה-Hot Spots (אלו בעלי מדד Lift הגבוה ביותר):

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(מהירות מותרת_70 קמ"ש)	(סוג תאונה_התנגשות חזית אל צד)	0.200	0.540	0.158	0.789	1.461
(חומרת תאונה_קשה)	(סוג תאונה_פגיעה בהולך רגל)	0.137	0.305	0.053	0.385	1.260
(סוג תאונה_פגיעה בהולך רגל)	(מהירות מותרת_עד 50 קמ"ש)	0.305	0.744	0.281	0.920	1.236
(מהירות מותרת_עד 50 קמ"ש)	(סוג תאונה_פגיעה בהולך רגל)	0.744	0.305	0.281	0.377	1.236
(רוחב הכביש_יותר מ-14)	(סוג תאונה_התנגשות חזית אל צד)	0.126	0.540	0.084	0.667	1.234
(חד מסלולית_חד סיטרי)	(סוג תאונה_פגיעה בהולך רגל)	0.225	0.305	0.084	0.375	1.228
(רב מסלולית_מיפרדה בנויה ללא גדר בטיחות)	(סוג תאונה_התנגשות חזית אל צד)	0.305	0.540	0.200	0.655	1.212
(סוג תאונה_התנגשות חזית אל צד)	(רב מסלולית_מיפרדה בנויה ללא גדר בטיחות)	0.540	0.305	0.200	0.370	1.212

טבלה 1 – החוקים עם ה-Lift הגבוה ביותר שנמצאו עבור תאונות ששייכות ל-Hot Spots.

החוקים מאפשרים לנו למצוא קשרים בין גורמים שונים בסט הנתונים שמתרחשים עבור תאונות שהיו שייכות ל-Hot Spots . כך למשל ניתן לראות מהחוק הראשון שבטבלה 1 שבכ-79 אחוז מהמקרים (Confidence של 79 אחוז), התאונות שקרו בכביש בו המהירות המותרת היא עד 70 קמ"ש היו תאונות מסוג התנגשות חזית של רכב אל צד של רכב אחר. כדוגמא נוספת, ניתן לראות (החוק הרביעי בטבלה) שרק כ-38 אחוזים מן התאונות שקרו בכביש בעל מהירות של עד 50 קמ"ש (כביש עירוני) היו תאונות מסוג פגיעה של רכב בהולך רגל. מכאן שבמרבית המקרים, תאונות בכביש עירוני היו מסוג של התנגשות של רכב ברכב אחר או בחפץ דומם (אלו מרבית סוגי התאונה האחרים שקיימים בסט הנתונים). חוקים מעין אלו עשויים לעזור בניתוח הגורמים לתאונות ב-Hot Spots שנמצאו.

בטבלה 2 מוצגים חוקים שנמצאו עבור התאונות שלא שייכות לאף Hot Spot:

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
חומרת תאונה_קשה	סוג תאונה_פגיעה בהולך רגל	0.132	0.326	0.068	0.513	1.572
רוחב הכביש_יותר מ-14	סוג תאונה_התנגשות חזית אל צד	0.119	0.376	0.061	0.510	1.356
חד מסלולית_לא קיים	סוג תאונה_התנגשות חזית אל צד	0.425	0.376	0.191	0.450	1.198
סוג תאונה_התנגשות חזית אל צד	חד מסלולית_לא קיים	0.376	0.425	0.191	0.509	1.198
חד מסלולית_חד סיטרי	סוג תאונה_פגיעה בהולך רגל	0.221	0.326	0.086	0.389	1.190
רב מסלולית_מיפרדה בנויה ללא גדר בטיחות	סוג תאונה_התנגשות חזית אל צד	0.336	0.376	0.149	0.444	1.181
סוג תאונה_התנגשות חזית אל צד	רב מסלולית_מיפרדה בנויה ללא גדר בטיחות	0.376	0.336	0.149	0.396	1.181
חומרת תאונה_קשה	רוחב הכביש_5 עד 7	0.132	0.332	0.052	0.391	1.178

טבלה 2 – החוקים עם ה-Lift הגבוה ביותר שנמצאו עבור תאונות שלא שייכות לאף Hot Spot.

ניתן לראות חפיפה מסוימת בין החוקים שבטבלה 1 לאלו שבטבלה 2. כך למשל, החוק השני שנמצא בטבלה 1 זהה לחוק הראשון שבטבלה 2. פירוש הדבר הוא שיש חוקים שנמצאו גם עבור תאונות

שהתרחשו ב-Hot Spots וגם עבור תאונות שקרו במקומות בהן תאונות מתרחשות באופן פחות תדיר. עם זאת, נשים לב שעשויים להיות קיימים הבדלים במדדי ה- Lift, Confidence, Support עבור חוקים משותפים לשתי הטבלאות, כך שייתכנו הבדלים במאפייני החוקים.

למרבית החוקים שמצאנו התקבלו ערכי Lift יחסית נמוכים. אמנם ערכי ה-Lift במרבית החוקים היו שונים מ-1, אך ייתכן שבחלק מהמקרים הם אינם מספיק שונים מ-1 מכדי שנוכל לדחות אפשרות של אי תלות בין המשתנים המשתתפים בחוק.

נציג לדוגמא גם חוקים שהתקבלו עבור תאונות שהתרחשו ב-Hot Spots כאשר הופעל סינון שהותיר רק תאונות שאינן מסוג של פגיעה בהולך רגל (טבלה 3):

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
סוג תאונה: התנגשות חזית אל חזית	רב מסלולית: לא קיים	0.061	0.480	0.051	0.833	1.737
סוג תאונה: התנגשות חזית אל חזית	מהירות מותרת: עד 50 קמ"ש	0.061	0.667	0.056	0.917	1.375
חומרת תאונה: קשה	רוחב הכביש: 7 עד 10.5	0.121	0.379	0.056	0.458	1.210
מהירות מותרת: 60 קמ"ש	סוג תאונה: התנגשות חזית אל צד	0.061	0.778	0.056	0.917	1.179
מהירות מותרת: 70 קמ"ש	סוג תאונה: התנגשות חזית אל צד	0.258	0.778	0.227	0.882	1.134
חומרת תאונה: קשה	מהירות מותרת: עד 50 קמ"ש	0.121	0.667	0.091	0.750	1.125
חומרת תאונה: קשה	חד מסלולית: לא קיים	0.121	0.520	0.071	0.583	1.121
רוחב הכביש: יותר מ-14	סוג תאונה: התנגשות חזית אל צד	0.141	0.778	0.121	0.857	1.102

טבלה 3 - החוקים עם ה-Lift הגבוה ביותר שנמצאו עבור תאונות ששייכות ל-Hot Spots ואינן מסוג פגיעה בהולך רגל.

נראה כי אין שוני משמעותי בערכי ה-Lift לעומת החוקים שבטבלאות 1 ו-2, אף על פי שכעת החוקים יוצרו רק עבור תאונות שאינן מסוג פגיעה בהולך רגל. עם זאת, מרבית החוקים שהתקבלו בטבלה 3 שונים מאלו שהתקבלו עבור כלל התאונות שהתרחשו ב-Hot Spots (טבלה 1), בה חצי מהחוקים שמוצגים בטבלה 1 כללו את הערך "סוג תאונה: פגיעה בהולך רגל" עם זאת, יש שני חוקים אחרים שגם הופיעו בטבלה 1 אך אינם מופיעים בטבלה 3. אם כן, הסינון גרם לשינוי בחוקים הבולטים ביותר שנמצאו.

נסתכל כעת על ערכי משתנים שהופיעו בניתוח עבור Hot Spots ספציפיים וננסה לנתח אותם. בטבלה 4 מוצגות כדוגמא שתי טבלאות שהתקבלו עבור שני Hot Spots שונים: עבור חטיבת גולני/אבן גבירול (טבלה 4a, 7 תאונות בסט הנתונים) ועבור חלוצי התעשיה/הגומא (טבלה 4b, 8 תאונות).

a.

support	itemsets
1.000	סוג תאונה: התנגשות חזית אל צד
1.000	תקינות הדרך: אין ליקוי
1.000	סימון ותמרורים: אין ליקוי
0.857	חומרת תאונה: קלה
0.857	מזג אוויר: בהיר
0.857	פני הכביש: יבש
0.571	תאורה: אור יום רגיל
0.429	תאורה: לילה פעלה תאורה

b.

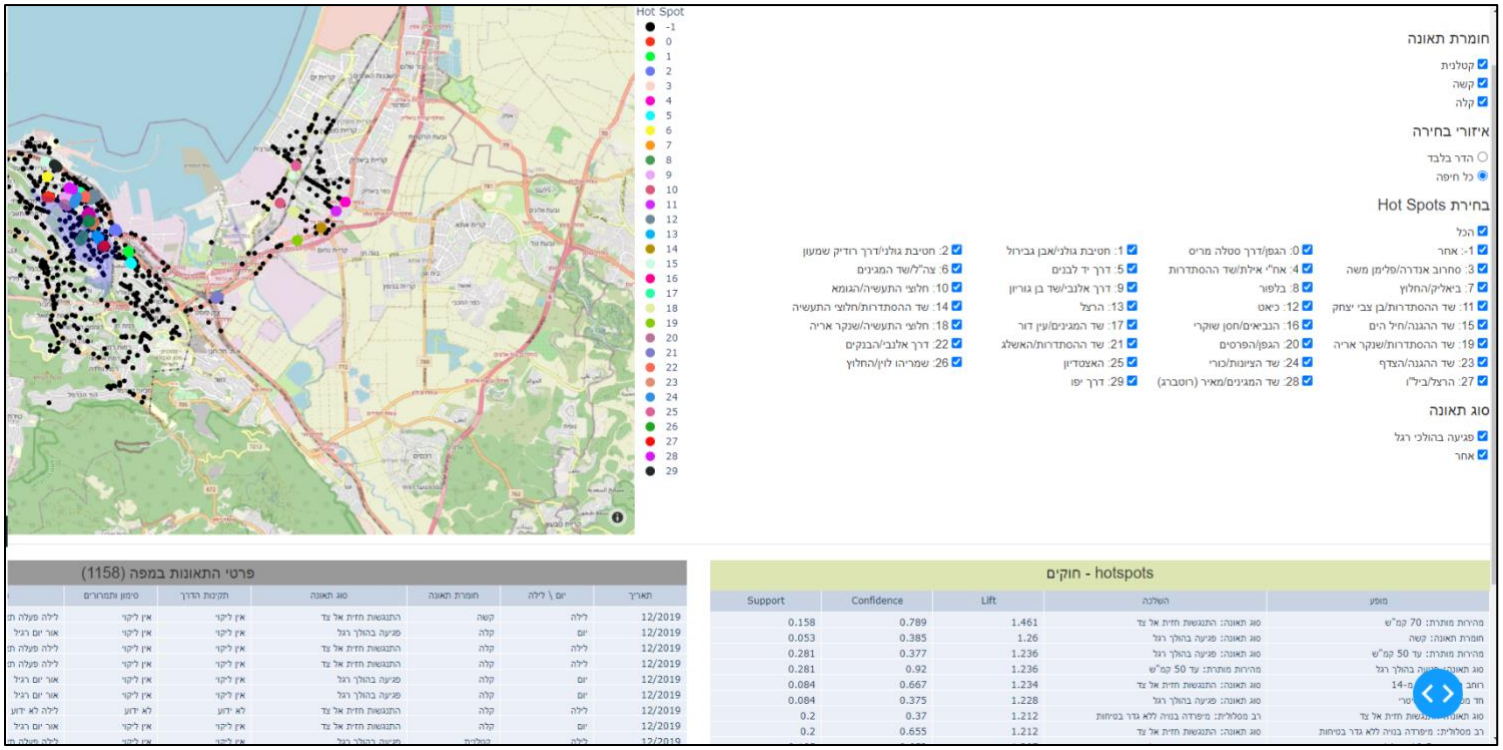
support	itemsets
0.875	מזג אוויר: בהיר
0.875	פני הכביש: יבש
0.750	חומרת תאונה: קלה
0.750	תקינות הדרך: אין ליקוי
0.750	סימון ותמרורים: אין ליקוי
0.375	יום בשבוע: שישי
0.375	סוג תאונה: התנגשות חזית אל צד
0.375	תאורה: אור יום רגיל
0.375	תאורה: לילה פעלה תאורה

טבלה 4 – משתנים נפוצים שהתקבלו עבור Hot Spots ספציפיים.
(a) - חטיבת גולני/אבן גבירול (b) - חלוצי התעשיה/הגומא

עבור ה-Hot Spot שהתגלה בחטיבת גולני/אבן גבירול (טבלה 4a), ניתן לראות כי כל התאונות שהתרחשו בו היו מסוג חזית אל צד, כלומר רק התנגשויות בין רכבים. כמו כן, לא היו ליקויים בתקינות הדרך, בסימונים ובתמרורים. עם זאת, אחת מתוך 7 התאונות שהיו באזור לא הייתה תאונה קלה, ושלוש מן התאונות התרחשו בלילה, כך שייתכן ויש קושי בנסיעה באזור זה בשעות הלילה על אף התאורה הקיימת בו. בנוסף, אחת התאונות התרחשה כאשר פני הכביש לא היו יבשים.

עבור ה-Hot Spot שהתגלה בחלוצי התעשייה/הגומא (טבלה 4b), ניתן להסיק על פי היחס המשלים כי ב-2 מתוך 8 תאונות שהתרחשו בו הדרך לא הייתה תקינה והיו ליקויים בסימוני הדרך והתמרורים. נראה אם כך כי היו מקרים בהם תנאי הכביש במקום היו בעייתיים. בנוסף, 3 מתוך 8 תאונות הדרכים קרו ביום שישי, כך שייתכן שביום זה יש עומס מיוחד שעשוי להגדיל את הסיכוי להתרחשות תאונה. ניתן גם לראות מהטבלה למשל כי בשני מקרים חומרת התאונה לא הייתה קלה, ושאחת התאונות התרחשה כשפני הכביש לא היו יבשים.
באמצעות התבוננות וניתוח של הטבלאות של Hot Spots ספציפיים (כמו אלו שבטבלה 4), ניתן להסיק מסקנות ולהפיק לקחים באופן פרטני עבור אזורים מועדים לתאונות בחיפה.

כעת, נסתכל על התצורה הסופית של הדאשבורד, המכילה ויזואליזציה של שני חלקי הפרויקט האחרים (איור 5):



איור 5 – צילום מסך של הדאשבורד.

הדאשבורד נבנה בהתאם לסכמה שתוכננה מראש (איור 4). בפינה השמאלית העליונה מופיעה מפת העיר חיפה, כאשר עליה מוצגות התאונות וה-Hot Spots שנמצאו על פי תוצאות אלגוריתם DBSCAN. לצד המפה, מופיעות אפשרויות לסינון הנתונים על פי דרגת חומרת התאונה, שייכות לאזור הדר, סוג התאונה ושייכות ל-Hot Spot.

בפינה השמאלית התחתונה מוצג מידע על התאונות המופיעות במפה בהתאם לסינונים שפועלים באותה נקודת זמן. בפינה הימנית התחתונה מוצגים חוקים שנמצאו על ידי מודל Association Rule עבור התאונות בסט הנתונים. לכל חוק מופיעים הערכים המשתתפים בחוק, כמו גם ה-Support, ה-Confidence וה-Lift של החוק. טבלה זו מתעדכנת בהתאם לסינונים של בחירת Hot Spots וסוג תאונה. פירוט על אופן השימוש בדאשבורד ניתן למצוא בנספח שבסוף המסמך.

דיון:

בפרויקט שלנו ניתחנו מידע אודות תאונות דרכים שהתרחשו בחיפה ובפרט בשכונת הדר. הניתוח כלל זיהוי של Hot Spots, כלומר אזורים מועדים לפורענות מבחינת כמות תאונות דרכים, ומציאת גורמים ומשתנים שעשויים לעזור בניתוח אותם Hot Spots יחד ולחוד. כמו כן, בנינו דאשבורד שמאפשר הצגה נוחה של הנתונים והממצאים בפני המשתמשים בבואם לנתח את מקרי תאונות הדרכים בחיפה.

על סמך סט הנתונים שלרשותנו, איתרנו 30 אזורים מסוכנים בחיפה על סמך השנים 2017-2019, רובם המוחלט הם צמתים, שם נראה כי הסיכוי לתאונה גובר עקב הצטלבויות כבישים. מניתוח של החוקים שהתקבלו עבור כלל התאונות שהתרחשו ב-Hot Spots לא מצאנו חוקים בעלי ערך Lift

גבוה, כלומר שייתכן שהחוקים שמצאנו אינם בהכרח מעידים על תלות בין המשתנים המשתתפים בחוק. עם זאת, חוקים אלו עשויים לעזור לעורר את תשומת ליבם של הגורמים האחראים לגבי מאפיינים שונים של התאונות והאזורים המסוכנים שמצריכים מעקב או התבוננות מעמיקה יותר. המשתנים התדירים שנמצאו עבור Hot Spots ספציפיים עשויים לשפוך אור במקרים מסוימים על הסיבות לתאונות שהתרחשו באזור מסוים, או על קשרים בין התאונות לגורמים מסוימים. לדוגמא, בניתוח הטבלה שמתקבלת עבור צומת חלוצי התעשיה/הגומא (טבלה 4b) ניתן לזהות שמתוך 8 תאונות, שתיים התרחשו במקרה בו היה ליקוי בתקינות הדרך, כך שייתכן ותקינות הדרך נפגמת לעיתים קרובות יחסית. בנוסף, שלוש מן התאונות התרחשו דווקא ביום שישי. ייתכן כי הדבר מצביע על כך שביום זה התאונה בצומת ערה מן הרגיל. באמצעות ניתוחים מעין אלו, ושימוש במערכת לצורך ניטור וניתוח של חוקים חריגים (בפרט כאלו בעלי ערך Lift גבוה), אנו מאמינים כי המערכת תוכל לסייע באיתור אזורים מסוכנים, זיהוי הבעיה בהן ופתירתה באופן יעיל.

במהלך הפרויקט למדנו והתנסינו בבניית מערכת הכוללת המנתחת ומציגה נתונים, כאשר חלקי המערכת נבנים אחד על גבי השני. ראשית, למדנו על אלגוריתם DBSCAN לצורך זיהוי ה- Hot Spots, והתנסינו בבחירת הפרמטרים של האלגוריתם (MinPts, epsilon) לצורך קבלת Hot Spots ממוקדים שתופסים צומת או מקטע של רחוב. בנוסף, למדנו על מודל Association Rule, אותו לא הכרנו קודם, ועל כיצד להשתמש בו כדי לאתר ולנתח גורמים מעניינים מתוך סט הנתונים. בשיחת זום שביצענו עם בת אל מן החמ"ל החברתי הבנו שחשוב לנסות ולהבין מה מאפיין את האזורים המסוכנים. לכן בחרנו להתמקד ביצירת החוקים ובויזואליזציה שלהם על גבי הדאשבורד באופן שיאפשר בחינה של החוקים המתקבלים עבור התאונות, ובפרט עבור סוגי תאונה ו-Hot Spots ספציפיים.

במידה והיה לנו זמן נוסף, היינו מוסיפים על גבי הדאשבורד אפשרות להזנת פרמטרים לאלגוריתם ה-DBSCAN באופן ידני למקרה והבחירה האוטומטית של הפרמטרים תניב תוצאות לא רצויות. ייתכן למשל, שבחירת epsilon האוטומטית לא תהיה מדויקת ותוביל לאיתור Hot Spots גדולים מדי. במקרה זה ניתן יהיה להזין באופן ידני והמערכת תוכל להריץ את האלגוריתם מחדש, לחשב את החוקים ולעדכן את הדאשבורד בהתאם לתוצאות. כמו כן, ניתן להרחיב את הפרויקט בעבודה משותפת עם החמ"ל על ידי הוספת פילטרים נוספים לדאשבורד שרלוונטיים לדעתם, כדי לפשט עבורם עוד יותר את תהליך ניתוח הנתונים.

ביבליוגרפיה:

[1] R. Bandyopadhyaya and S. Mitra, “Comparative Analysis of Hotspot Identification,” *Road Saf. Simul.*, 2011.

[2] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” 1996.

[3] N. Rahmah and I. S. Sitanggang, “Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra,” in *IOP Conference Series: Earth and Environmental Science*, 2016, vol. 31, no. 1, doi: 10.1088/1755-1315/31/1/012012.

[4] S. Szénási and P. Csiba, “Clustering algorithm in order to find accident black spots identified by GPS coordiantes,” *Int. Multidiscip. Sci. GeoConference Surv. Geol. Min. Ecol. Manag. SGEM*, vol. 1, no. 2, pp. 497–504, 2014, doi: 10.5593/sgem2014/b21/s8.063.

[5] W. Cheng and S. Washington, “New criteria for evaluating methods of identifying hot spots,” *Transp. Res. Rec.*, no. 2083, pp. 76–85, 2008, doi: 10.3141/2083-09.

[6] K. Geurts, I. Thomas, and G. Wets, “Understanding spatial concentrations of road accidents using frequent item sets,” *Accid. Anal. Prev.*, vol. 37, no. 4, pp. 787–799, 2005, doi: 10.1016/j.aap.2005.03.023.

[7] R. Agrawal, T. Imieliński, and A. Swami, “Mining Association Rules Between Sets of Items in Large Databases,” *ACM SIGMOD Rec.*, vol. 22, no. 2, 1993, doi: 10.1145/170036.170072.

[8] F. Kurniawan, B. Umayah, J. Hammad, S. M. S. Nugroho, and M. Hariadi, “Market Basket Analysis to Identify Customer Behaviours by Way of Transaction Data,” *Knowl. Eng. Data Sci.*, vol. 1, no. 1, p. 20, 2017, doi: 10.17977/um018v1i12018p20-25.

[9] K. Geurts, G. Wets, T. Brijs, and K. Vanhoof, “Profiling High Frequency Accident Locations,” *Transp. Res. Rec.*, vol. 32, no. 0, pp. 1–18, 2003, doi: 10.3141/1840-14.

[10] I. Makarova, G. Yakupova, P. Buyvol, E. Mukhametdinov, and A. Pashkevich, “Association rules to identify factors affecting risk and severity of road accidents,” 2020, doi: 10.5220/0009836506140621.

[11] אתר חי פה - https://haipo.co.il/item/291759

נספח:

בנספח זה נתאר את אופן ההתקנה והשימוש בדאשבורד שבנינו:

<u>התקנת המערכת:</u>

1. ראשית, התקינו את anaconda3 64 bit – הגרסה הכי חדשה שזמינה.
   קישור להורדה: https://www.anaconda.com/products/individual
   באמצעות פלטפורמה זו נוכל להתקין ספריית python "בעייתית" בהמשך (סעיף 4).

2. הורידו את תיקיית הפרויקט מ-GitHub וחלצו אותה. התיקייה נמצאת בקישור:
   https://github.com/moshinhoabadi/Hotspot-Identification-and-Analysis
3. פתחו command line מתוך התיקייה של הפרויקט, והזינו את הפקודה:

```
pip install -r requirements.txt
```

פקודה זו תתקין את ספריות ה-python הנחוצות לשם הפעלת המערכת. שימו לב כי ההתקנה עשויה לקחת כמה דקות.

4. הזינו ב-command line את הפקודה:

```
conda install shapely
```

ועקבו אחרי ההתקנה. פקודה זו תתקין ספרייה נחוצה נוספת באמצעות anaconda3.

<u>הפעלת המערכת:</u>

1. ראשית, יש להריץ את הקובץ main.py דרך ה-command line. פתחו את ה-command line מתוך תיקיית הפרויקט, והריצו את הפקודה הבאה:

```
python3 main.py
```

הערה: כדי לטעון קובץ נתונים אחר מקובץ ברירת המחדל, יש להזין את הנתיב לקובץ במהלך כתיבת הפקודה באופן הבא (דוגמא עבור קובץ בשם example.csv):

```
python3 main.py -file example.csv
```

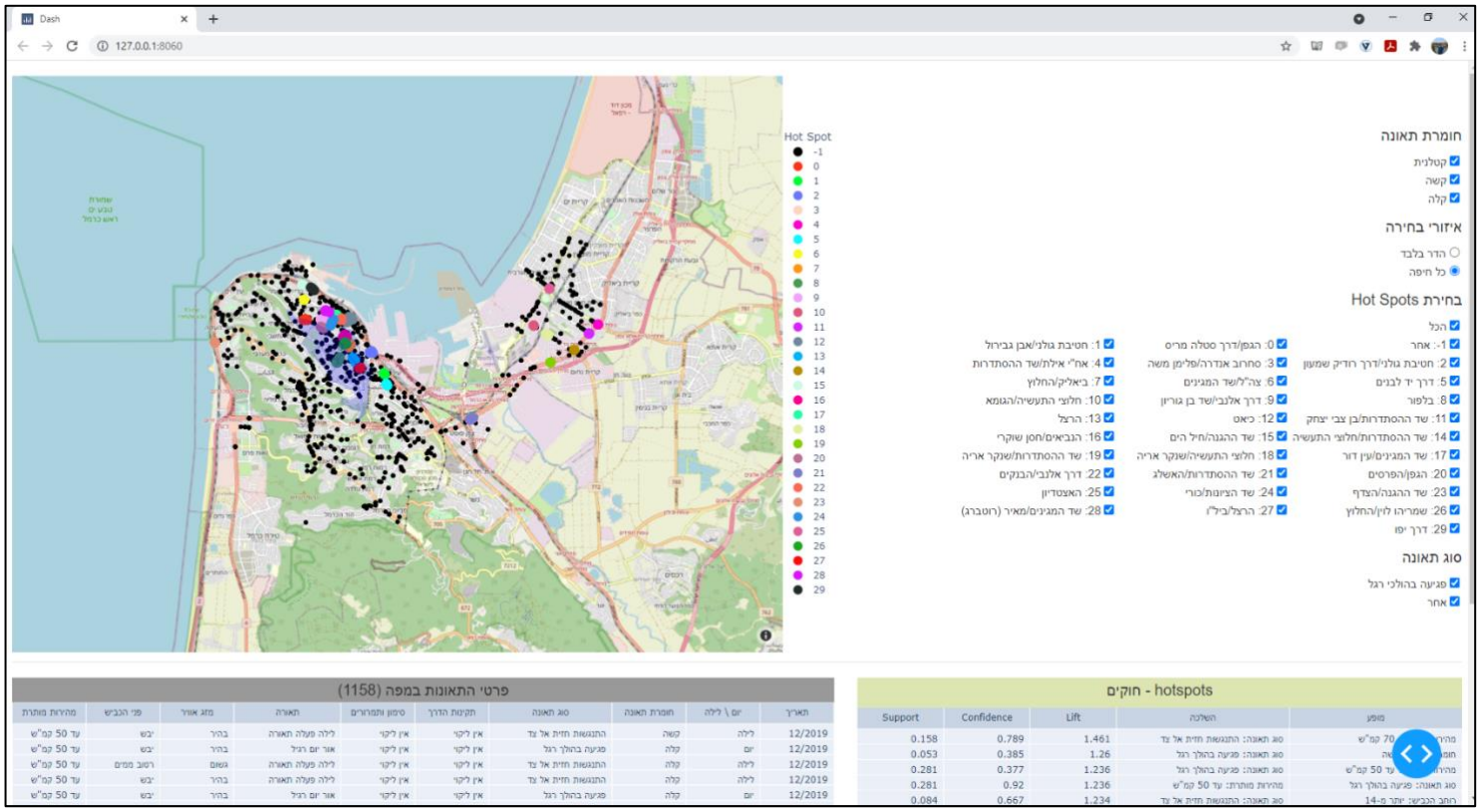
2. כעת הקוד יתחיל לרוץ ולבצע את הניתוחים על קובץ הנתונים. יש לחכות עד אשר תופיע ההודעה:

```
Dash app running on http://127.0.0.1:8060/
```

הערה: בטעינת קובץ שאינו קובץ ברירת המחדל, ריצת הקוד עד לשלב זה עלולה לקחת כמה דקות.

3. לאחר שההודעה מופיעה, פתחו כרטיסייה בדפדפן, והזינו את הנתיב http://127.0.0.1:8060/

4. לאחר כמה שניות, הדאשבורד ייטען (איור 6). מומלץ להשתמש במסך גדול ככל האפשר לצורך תצוגה מיטבית (אנחנו השתמשנו במסך בגודל 21 אינץ').



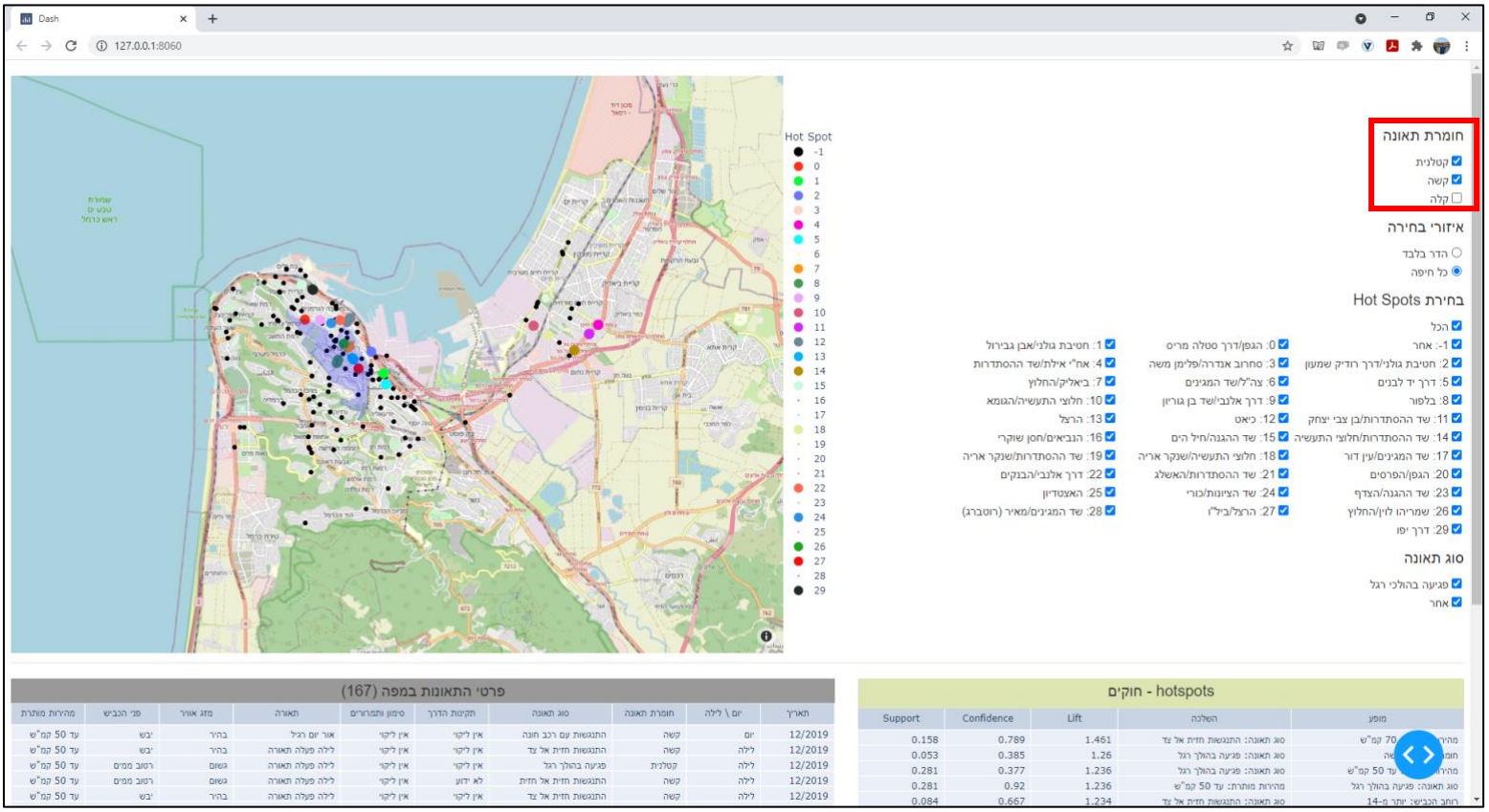
איור 6 – צילום מסך של הדאשבורד מתוך הדפדפן.

שימוש במערכת:

הדאשבורד מכיל את סימון התאונות על מפת חיפה, כאשר ה-Hot Spots מסומנים בצבעים, ושאר התאונות מיוצגות על ידי עיגולים שחורים. מתחת למפה נמצאת טבלה המציגה את פרטי התאונות שמופיעות במפה. בכותרת טבלה זו רשום בסוגריים מספר התאונות המוצגות במפה באותו רגע. בפינה הימנית התחתונה נמצאת טבלת החוקים שיוצרו על ידי מודל Association Rule. במידה ובאחת הטבלאות מספר שורות גבוה, תופיע אפשרות לגלול בה שורות על ידי הצבת העכבר על הטבלה ושימוש בגלגלת.

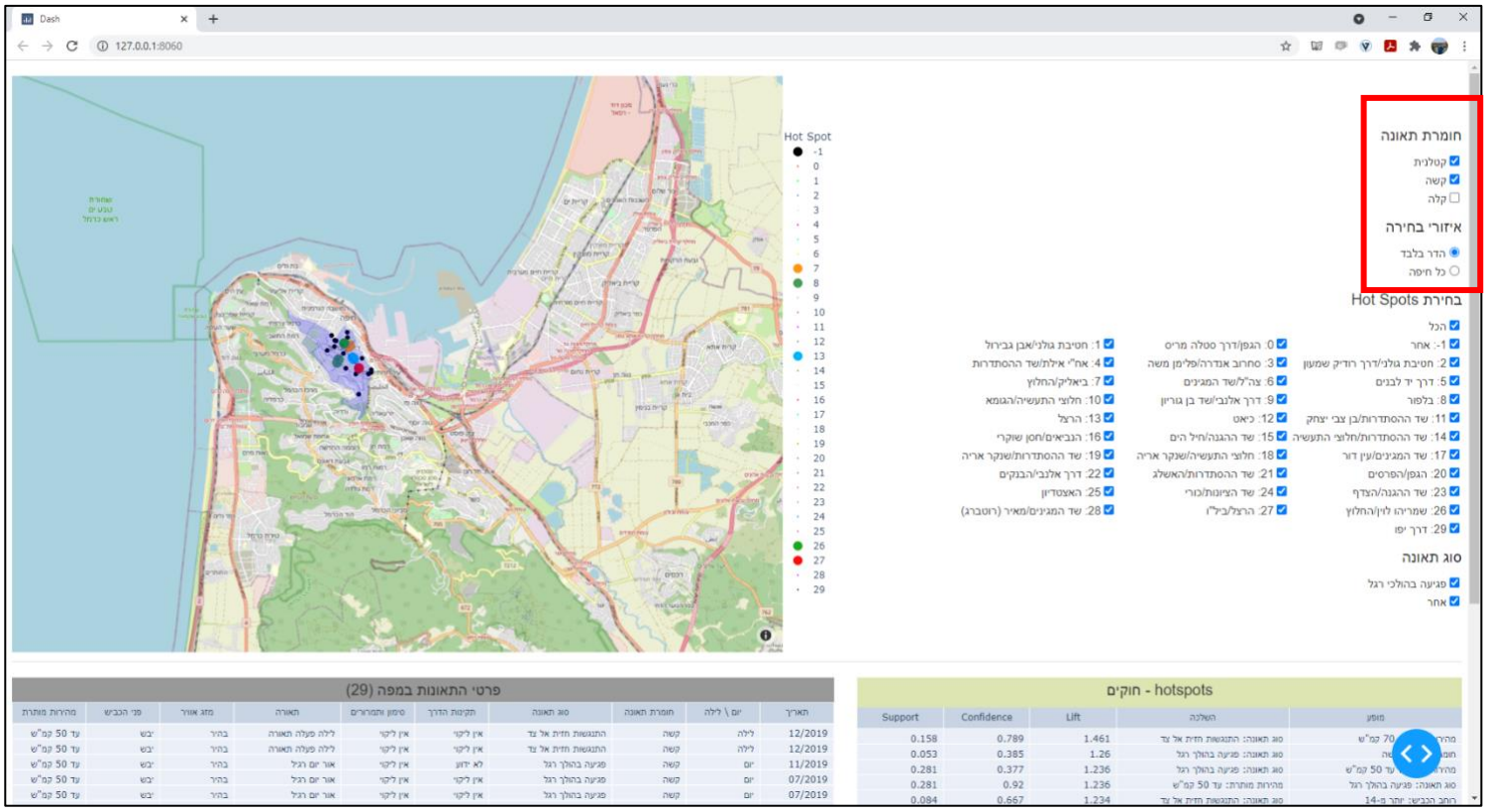
האינטרקציה עם הדאשבורד באה לידי ביטוי באפשרויות הסינון בו. נעבור עליהן כעת (החל מהעמוד הבא) ונציג דוגמא למצב הדאשבורד עבור כל סוג סינון:

- **חומרת תאונה** – סינון התאונות המוצגות במפה על פי דרגת החומרה של התאונות: קלה, קשה או קטלנית. ניתן לסמן כל קומבינציה של אפשרויות אלו. סינון זה משפיע גם על טבלת פרטי התאונות, בה יופיעו כעת רק תאונות בעלי דרגת חומרה המתאימה לאפשרויות שסומנו. באיור 7 ניתן לראות את מצב הדאשבורד לאחר שבחרנו להציג רק תאונות מדרגת חומרה קשה או קטלנית.



איור 7 - מצב הדאשבורד לאחר הפעלת סינון תאונות בעלות דרגת חומרה קלה.

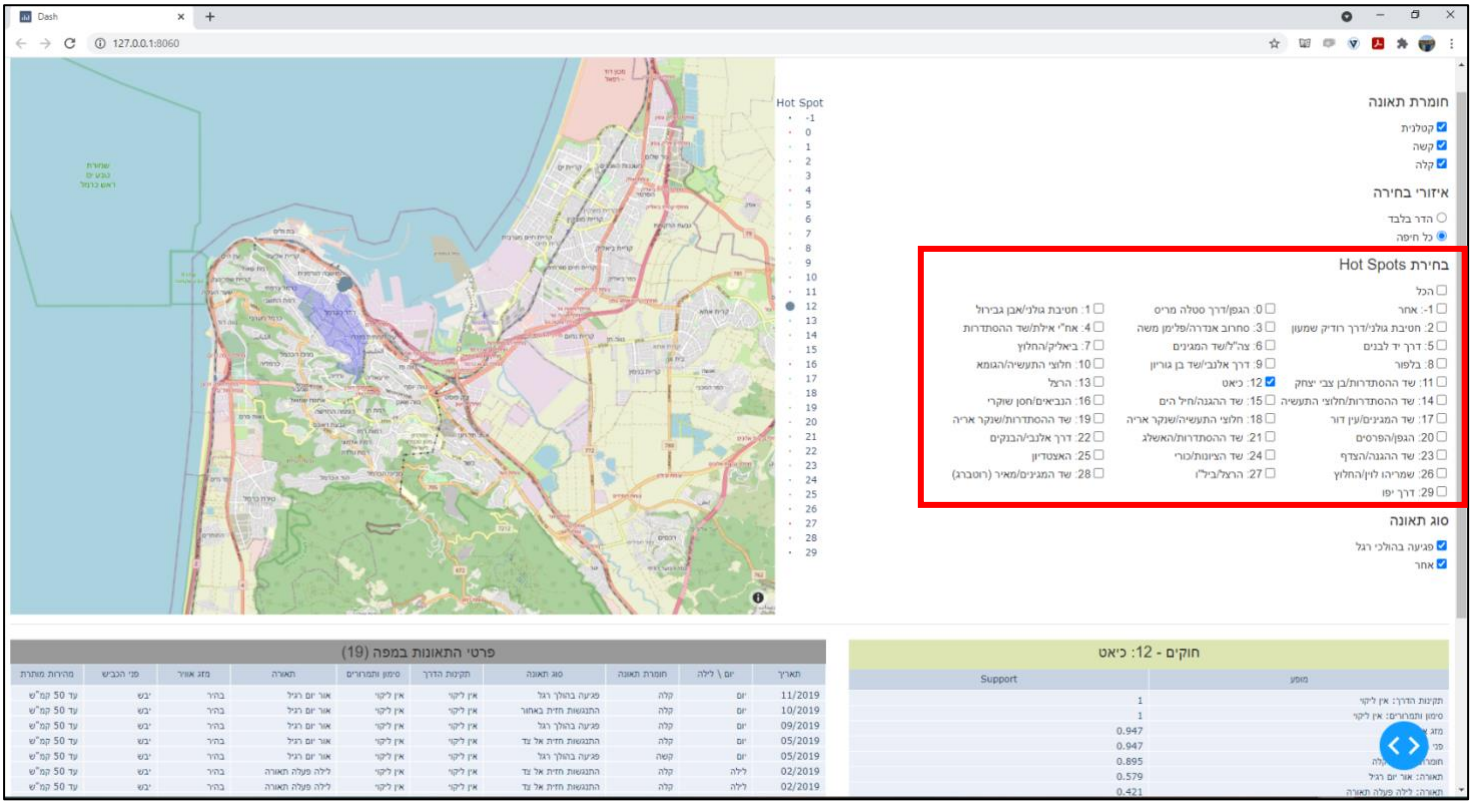
- **איזורי בחירה** – ניתן לבחור להציג את כלל התאונות שהתרחשו בחיפה, או רק תאונות שהתרחשו בתחומי הדר. המפה וטבלת פרטי התאונות תתעדכנה בהתאם (איור 8).



איור 8 – הפעלת סינון "הדר בלבד" על הדאשבורד מאיור 7.

- **בחירת Hot Spots** – ניתן לבחור אלו Hot Spots יוצגו. האינדקס המספרי שליד שמות ה- Hot Spots מתאים למפתח הצבעים שמופיע לצד המפה. הסמן "אחר" מתייחס לכל התאונות שלא סווגו לאף Hot Spot. באמצעות הסמן "הכל" ניתן לבצע בחירה/ביטול של כל הסמנים האחרים בקטגוריה זו.

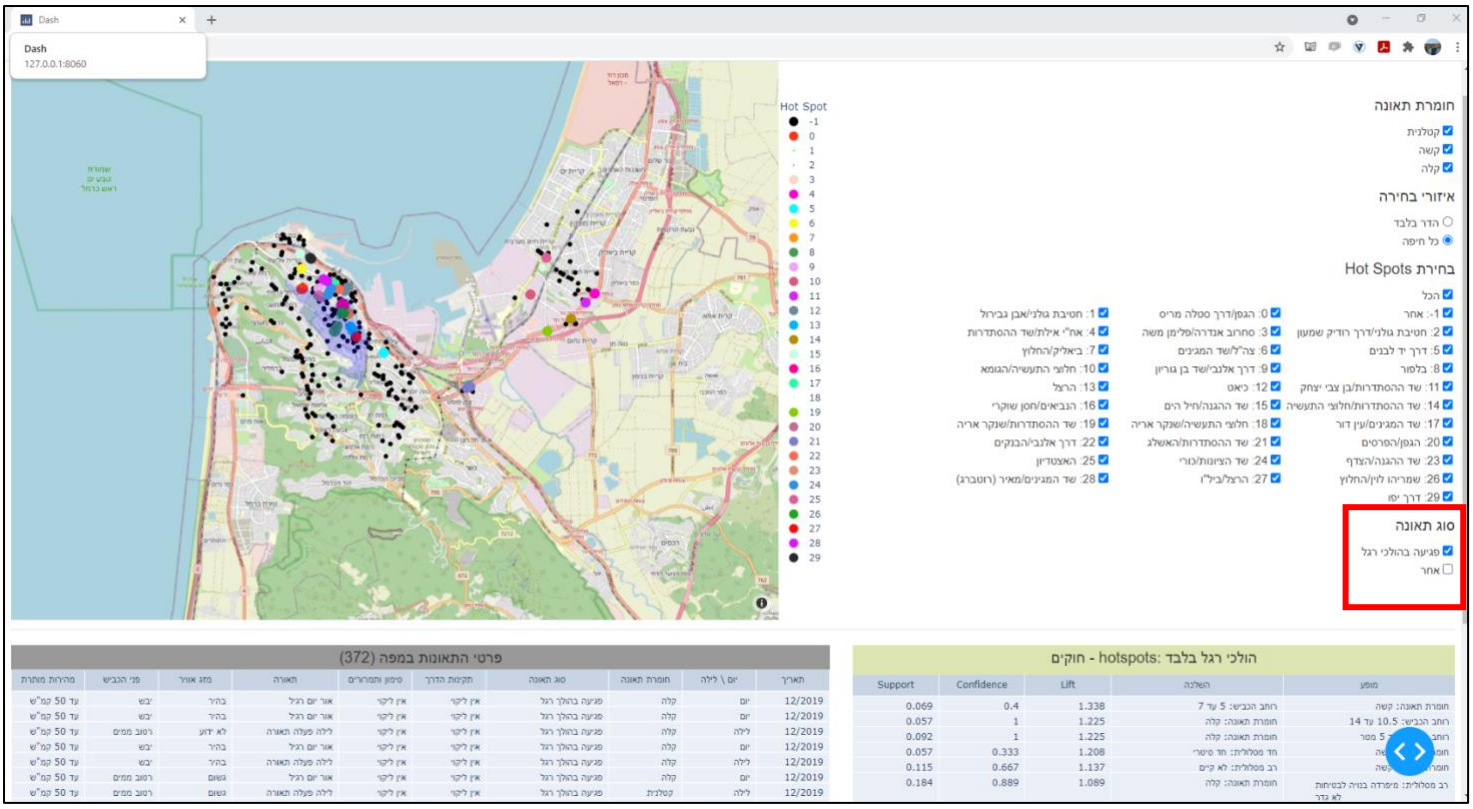
מלבד המפה וטבלת פרטי התאונות, סינון זה משפיע גם על טבלת החוקים. אם מסומן רק Hot Spot אחד, הטבלה תציג את המשתנים התדירים שנמצאו בו. אם מסומן רק "אחר", הטבלה תציג את החוקים שנמצאו עבור התאונות שאינן שייכות לאף Hot Spot. בכל מקרה אחר, הטבלה תציג את החוקים שנמצאו עבור כלל התאונות ששייכות Hot Spot כלשהו. איור 9 מציג דוגמא בה סומן Hot Spot יחיד:



איור 9 – מצב הדאשבורד לאחר סימון ה-Hot Spot שנמצא בכיאט בלבד.

- **סוג תאונה** – ניתן להציג תאונות מסוג פגיעה בהולך רגל, תאונות מסוג אחר, או את שתי האפשרויות יחד. המפה, טבלת פרטי התאונות, וטבלת החוקים ישתנו בהתאם לסינון שנבחר. עם זאת, טבלת החוקים תשתנה רק בנתונים שהיא מציגה עבור כלל ה-Hot Spots והתאונות שלא שייכות לאף Hot Spot. החוקים שיוצגו עבור Hot Spots ספציפיים יישארו ללא שינוי.

איור 10 מציג את מצב הדאשבורד לאחר בחירת תאונות מסוג פגיעה בהולכי רגל בלבד:



איור 10 – מצב הדאשבורד לאחר בחירת תאונות מסוג "פגיעה בהולכי רגל" בלבד.