Emergency calls in the U.S.A

<u>המגישים :</u>

איריס קוטל - 208575043

אמיר כהן - 205665086

עומר ארואטי - 204332126

מבוא

מוקד 911 הוא מוקד טלפוני המשרת את תושבי אמריקה הצפונית במקרי חירום.
בארצות הברית מתקיימות כ-240 מיליון שיחות למוקד 911 מדי שנה, כאשר ב-80% מהמקרים מדובר בקריאות ממכשירים ניידים. באזורים בהם פועל מוקד ה-911 השיחה מנותנת באופן סלקטיבי למוקד אזורי של 911 המטפל באזור הספציפי ממנו מתקבלת השיחה.

הפרויקט עוסק בניתוח וכריית מידע מתוך מסד נתונים של מוקד החירום "911" הפועל במחוז מונטגומרי שממוקם במדינת פנסילבניה בארה"ב. מסד הנתונים מכיל כ-600,000 מקרי התקשרות אל המוקד בין השנים 2014-2020.
לגבי כל שיחה נתון מיקום המתקשר (באמצעות מיקוד, כתובת וקווי אורך ורוחב), חותמת זמן של השיחה (שכוללת תאריך ושעה מדויקת), קטגוריה של מקרה החירום (אם מדובר במקרה מסוג FIRE,EMS או TRAFFIC) והסבר קצר על מקרה החירום.

כידוע, מטרתו של מוקד החירום היא לספק עזרה למתקשר בצורה הכי מהירה שמתאפשרת.
השאלה העסקית הנשאלת היא כיצד ניתן לתת את השירות הטוב ביותר בצורה היעילה והחסכונית ביותר.
מטרת הפרויקט היא ליצור תחזית מסוימת לאירועי חירום עתידיים באמצעות שימוש בקורלציות שקיימות בין העמודות השונות שבמסד הנתונים. חיזוי עתידי כזה של אירועים יעניק למוקד ה-911 אפשרות לנצל באופן יעיל יותר את משאבים הנתונים לרשותו וכתוצאה מכך ייווצר חסכון עלותי.

מבחינה טכנית, האתגר העיקרי היה להבין את האופן בו הכי כדאי לנקות ולהפחית את הנתונים הקיימים. המטרה הייתה להביא את הנתונים למצב שיאפשר הפעלה פשוטה של האלגוריתמים בכדי להגיע לתחזית.
הקושי היה המספר הרב של הנתונים שדרשו טיפול ובנוסף תהליך הפחתת המידע דרש הצלבה של מידע ממקורות שלא היו במסד הנתונים הנבחר.

ערך המטרה שנבחר להתמקדות עבור הפרויקט הוא השיחה המתקבלת למוקד הינה בנושא "שריפה" או לא.
שדה זה הינו משתנה בינארי שערכו הוא 0 או 1.
0 – עבור מקרה בו השיחה אינה בנושא "שריפה".
1 – עבור מקרה בו השיחה הינה בנושא "שריפה".

ניתוח ערך מטרה זה ותוצאותיו יכול להוות מידע שימושי עבור המוקד. בצורה זו, המוקד יוכל להתנהל בצורה יעילה יותר מול המשאבים הנתונים לרשותו ובכך להביא לייעול השירות שמתקבל.

קישור למאגר הנתונים מתוך אתר Kaggle : https://www.kaggle.com/mchirico/montcoalert

תהליך ה-Preprocessing והשיטות שבוצע בהן שימוש בפרויקט

לאחר בחינת מאגר הנתונים המכיל כ-600,000 רשומות, הוחלט לצמצם את טווח ההסתכלות לתקופה שבין החודשים ינואר עד מרץ של שנת 2016, מדובר בסה"כ 18,200 רשומות.

Data Reduction

העמודות שהוסרו ממסד הנתונים כחלק מתהליך "ניקוי המידע" :

לאחר סקירת כלל העמודות במסד הנתונים, נמצאו מספר עמודות שהיוו כפילויות של עמודות אחרות:

Longitude, Latitude, Township ו- General address כולם היוו כפילויות אחת של השנייה ולכן הוחלט להסירן.

בנוסף, היו מספר עמודות שכללו מידע שלא היה רלוונטי לערך המטרה ולכן הוחלט להסירן:

עמודת Description of emergency כללה נתונים שתיארו בצורה קצרה את אודות האירוע.

מידע זה לא היה רלוונטי בחיפוש אחר הקורלציה הנדרשת למימוש מטרת הפרויקט ולכן הוחלט להסיר עמודה זו.

עמודת ה-Index column לא כללה מידע שקשור לערך המטרה, מדובר בעמודה שערכה תמיד 1.

העמודות שבוצע בהן שימוש למטרת הפרויקט :

עמודת ה-Zip Code שכללה את המיקוד של האדם שביצע את השיחה למוקד.

עמודת ה-Title of emergency שכללה את קטגוריית האירוע EMS\FIRE\TRAFFIC, ובנוסף תיאור בצמד מילים על האירוע לדוגמא: "Fire: GAS-ODOR/LEAK".

עמודת ה-Time Stamp שכללה חותמת זמן של תאריך ושעת האירוע בצורה הבאה: 2015-12-10 18:32:25

Data Transformation

בכדי לקבל ראייה מעמיקה יותר על הנתונים, הוחלט לפצל את שדה ה-Time Stamp אשר מורכב מתאריך ושעה ולבצע שימוש רק בשעה, מפני שהתאריך אינו רלוונטי עבור ערך המטרה.

בוצעה דיסקריטיזציה על ערך השעה מתוך חותמת הזמן ושעות היום חולקו ל-6 קטגוריות באופן הבא:

לפנות בוקר – בין השעות 00:00-04:59

בוקר – בין השעות 05:00-11:59

צהריים – בין השעות 12:00-13:59

אחר צהריים – בין השעות 14:00-16:59

ערב – בין השעות 17:00-21:59

לילה – בין השעות 22:00-23:59

בנוסף, הוחלט על הוספת עמודת "Week End" שכוללת ערכים בינאריים (0 או 1) ומשמעות הערך הוא:

0 – אם מדובר ביום חול ו- 1 – אם מדובר בסוף שבוע.

והוספת עמודת "Mid Week" שגם היא כוללת ערכים בינאריים (0 או 1) ומשמעותם:

0 – אם מדובר ביום במהלך הסוף שבוע ו- 1 – אם מדובר ביום במהלך השבוע.

בעמודת ה-Title of emergency בוצע (באמצעות קוד Python) חיתוך של המידע הלא רלוונטי מתוך כל אחד מהערכים בעמודה. לדוגמא, עבור ערך "Fire: GAS-ODOR/LEAK" הערך המתקבל לאחר החיתוך יהיה "Fire".

עמודת ה-Zip Code כללה בתוכה מספר רב מאוד של מיקודים ובוצעה הפחתת מידע גם על עמודה זו.
באמצעות שימוש באתר אינטרנטי שהציג את מפת פנסילבניה בצירוף כל אזורי המיקודים השונים של מדינת פנסילבניה, בוצעה המרה של כל מיקוד ל"קוד אזורי". באופן זה, מאות מיקודים שונים מוינו ל-8 קטגוריות בלבד של קודים אזוריים. קישור לאתר האינטרנטי שבאמצעותו בוצע סיווג המיקודים השונים לקודים אזוריים:
https://www.unitedstateszipcodes.org/pa/

לאחר ביצוע העבודה המקדימה על סט הנתונים, הוא יראה כך:

Fire	Area Code	Time Category	Weekend	Midweek
1	215	Afternoon	Yes	No
0	484	Night	No	Yes
1	267	Early morning	No	Yes
1	610	Noon	Yes	No
0	215	Afternoon	No	Yes
1	484	Evening	Yes	No
0	610	Morning	Yes	No
1	267	Noon	No	Yes
0	215	Night	Yes	No

בנוסף לפעולות אלו, בשילוב עם עבודה בקוד Python בוצע שימוש **בקידוד "One Hot"** על כל המשתנים הבלתי תלויים (העמודות הבאות – Area Code, Time Category, Weekend).
מטרתו של הקידוד היא לייצג משתנים קטגוריאליים כווקטורים בינאריים וזאת על מנת לאפשר לאלגוריתם לתפקד.

Techniques Used

במהלך הפרויקט נבדקו ערכי ה-ROC, ה-Recall של האלגוריתמים הבאים:

1. עץ החלטה
2. KNN model
3. Naïve Bayesian

באופן כללי רצוי שערך ה-ROC יהיה קרוב ככל המתאפשר ל-1.

ערך ה-ROC מבטא את השטח שמתחת לגרף ה- True\False Positive Rate ולכן ככל שערכו מתקרב ל-1 המשמעות הינה דיוק טוב יותר.

מבחינת ערך ה-Recall, רצוי שגם ערכו יהיה קרוב ככל המתאפשר ל-1.

המשמעות לכך היא שהאלגוריתם יזהה את האירועים העתידיים בקירוב של 100%, ובכך יתאפשר למוקד להתנהל בצורה חכמה יותר שתביא לחסכון בעלויות מיותרות.

להלן השימוש שנעשה בכל אלגוריתם והתוצאות שהתקבלו בכל אחד מהשימושים:

1. אלגוריתם עץ החלטה – Decision tree

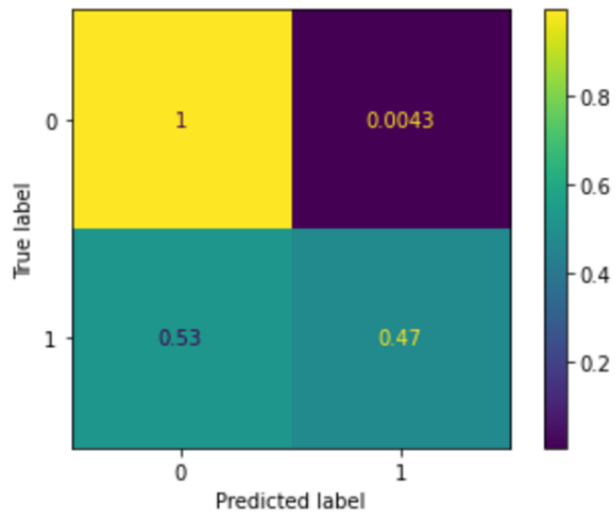
עץ החלטה הוא עץ בינארי מלא המורכב מצמתי החלטה שבכל אחד מהם נבדק תנאי מסוים על מאפיין מסוים של התצפיות ו"עלים" המכילים את הערך החזוי עבור התצפית המתאימה למסלול שמוביל אליהם בעץ.

במהלך הפרויקט בוצע שימוש באלגוריתם זה בכדי לזהות תבניות לפיהן יהיה ניתן להסיק האם שיחות שיגיעו אל המוקד יהיו בנושא "שריפה" או בנושא אחר.

להלן ה- Confusion Matrix של אלגוריתם עץ ההחלטה

כפי שניתן לראות באמצעות המטריצה, ערך ה-Recall של אלגוריתם עץ ההחלטה הוא 47.34%

```
accuracy: 0.9120879120879121
recall: 0.4734133790737564
precision: 0.9550173010380623
```



2. אלגוריתם ה-KNN Model

KNN Model הוא אלגוריתם מבוסס מופעים, שמסווג נתון ומחפש למי הנתון "דומה" מבין "השכנים".

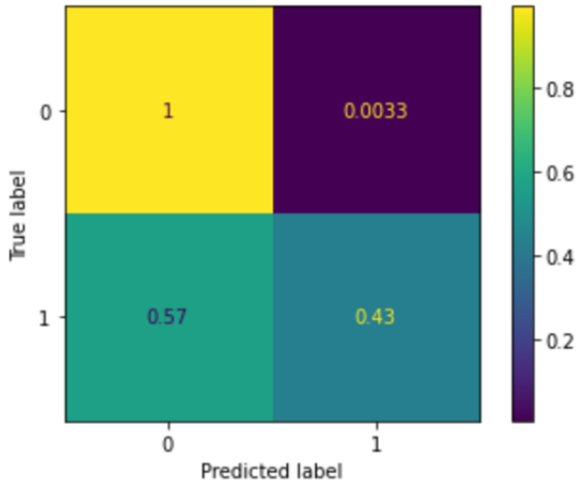
פעולה זו מתבצעת עד לקבלת סיווג של הנתון (כלומר מציאת הדמיון הגדול ביותר).

בפרויקט זה, בוצע שימוש באלגוריתם על מנת לחפש דמיון בין שיחות חדשות המגיעות למוקד לבין השיחות הישנות ("השכנים") וזאת על מנת למצוא חיזוי לערך המטרה.

להלן ה- Confusion Matrix של KNN Model ←

כפי שניתן לראות באמצעות המטריצה, ערך ה-Recall של אלגוריתם ה-KNN Model הוא 43.17%

```
accuracy: 0.8978021978021978
recall: 0.4317111459968603
precision: 0.9649122807017544
```



3. אלגוריתם ה- Naïve Bayesian

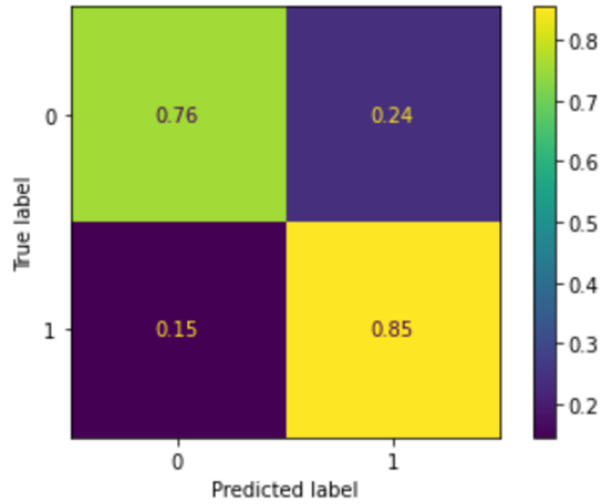
אלגוריתם זה מבוסס על תורת ההסתברות בדומה לאלגוריתם הבייסיאני, ההבדל הוא שבאלגוריתם זה ישנה הנחה "נאיבית" שהמשתנים אינם תלויים אחד בשני.

בפרויקט זה, השימוש באלגוריתם זה היה עבור סיבות סטטיסטיות. מדובר באלגוריתם המבוסס על סטטיסטיקה שיכולה לתת חיזוי מדויק למקרי חירום עתידיים.

להלן ה- Confusion Matrix של Naïve Bayesian ←

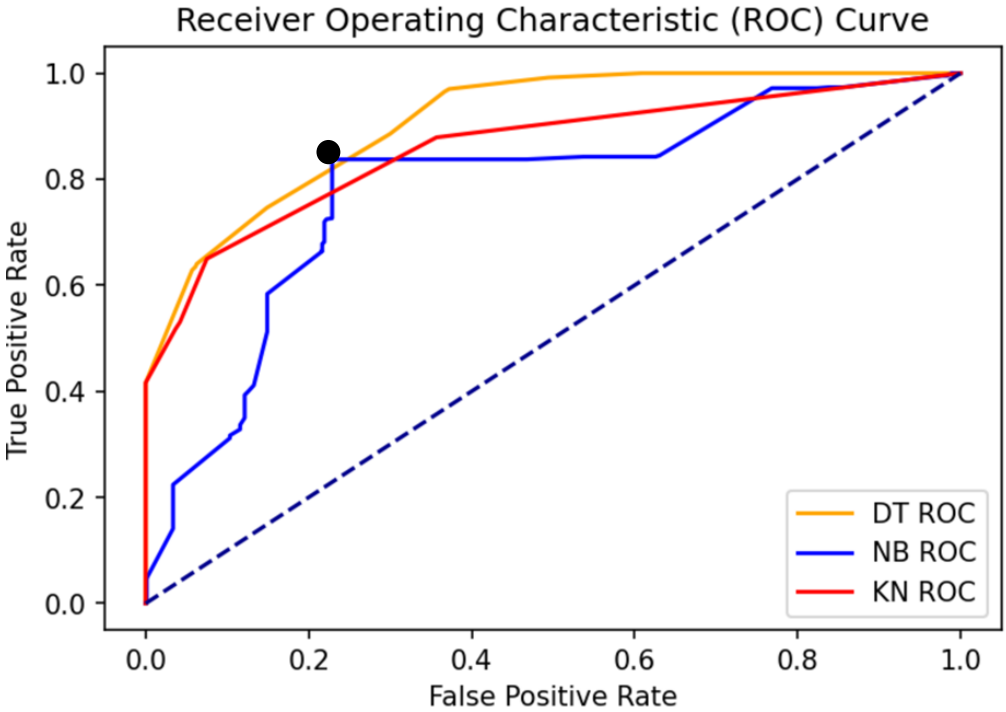
כפי שניתן לראות, ערך ה-Recall עבור אלגוריתם זה הינו 85.44%

```
accuracy: 0.7728021978021978
recall: 0.8544520547945206
precision: 0.40209508460918614
```



תוצאות

להלן גרף ה-ROC המכיל שלושה גרפים, כאשר כל גרף מייצג תוצאות של כל אחד מהאלגוריתמים :



על פי גרף זה ניתן להסיק כי האלגוריתם שקיבל את הציון הגבוה ביותר הוא אלגוריתם עץ ההחלטה.

כפי שניתן לראות, הגרפים אינם חופפים במיוחד מה שמעיד על השוני בתוצאות של כל אחד מהאלגוריתמים.

במסגרת הפרויקט ישנה עדיפות לבחור את האלגוריתם בעל ערך ה-Recall הגבוה ביותר וזאת מפני שכאשר מדובר באירועי חירום מסוג "שריפה" שמהווים סכנת חיים ממשית, נרצה דיוק גבוה ככל המתאפשר.

על ידי השוואת מטריצות הבלבול של כל אחד מהאלגוריתמים ניתן לראות כי אמנם אלגוריתם עץ ההחלטה ואלגוריתם ה-KNN הם בעלי ערך ה-Precision הגבוה ביותר, אך ה-Recall שלהם נמוך יחסית לעומת ערך ה-Recall באלגוריתם ה- Naïve Bayesian.

לכן, האלגוריתם הנבחר הינו אלגוריתם ה-Naïve Bayesian ונקודת העבודה היא כמסומנת בתרשים ובה ישנו חיזוי מוצלח של מעל ל-80% מהמקרים בהם יש אירועי חירום בנושא שריפות, עם שגיאה של 20%.

שימוש באלגוריתם ה-Naïve Bayesian ייתן למוקד חיזוי מדויק באמצעות שימוש בהסתברויות ובכך תינתן למוקד האפשרות להיערך בצורה טובה יותר לאירועי חירום עתידיים, מה שיביא לצמצום בעלויות מיותרות ולשיפור שירות המוקד.

מסקנות

ההמלצה הנובעת מהתהליך המחקרי הינה ביצוע הערכה מחודשת לגבי אופן השימוש במשאבים הנמצאים לרשות המוקד.

באמצעות התחזית לגבי סבירות השיחה המתקבלת להימצא בקטגוריית "שריפה", ניתן יהיה להקים מחלקה ייעודית לטיפול בנושא שריפות בה יענו מוקדנים בעלי הכשרה מקצועית ובעלי ידע מתאים לטיפול בפניות בנושאים אלו.

התהליך יתנהל באופן הבא - כאשר שיחה תגיע אל המוקד, היא תעבור ניתוב למוקד המתאים לטיפול בפנייה. הניתוב יתבצע ע"י שימוש בתחזית שנבנתה במהלך הפרויקט ובאמצעותה ניתן יהיה לסווג את קטגוריית הפנייה בצורה מיידית עוד לפני קבלת האינדיקציה של המדווח.

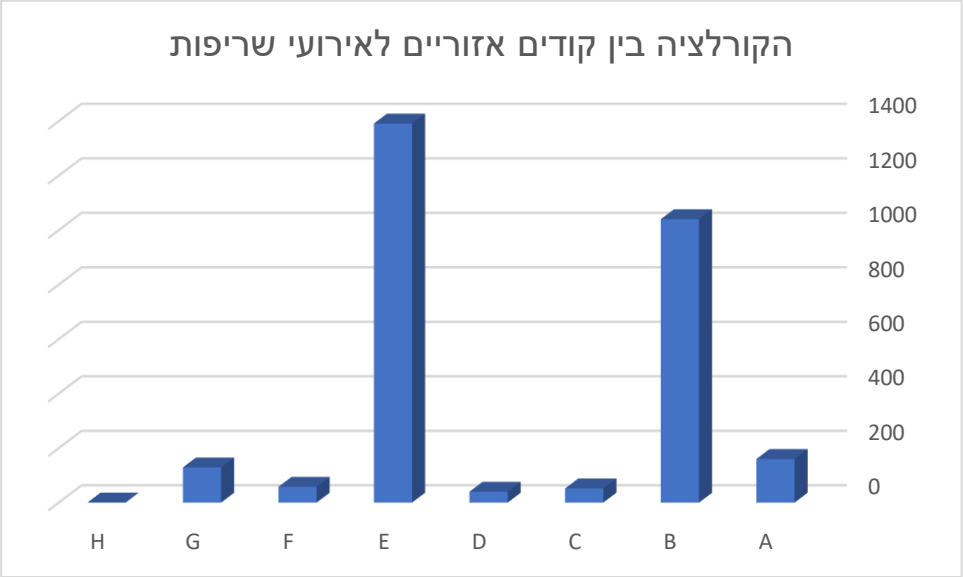
כלומר אם למשל תגיע שיחה אל המוקד ועל פי התחזית לשיחה זו יש סיכוי גבוה מספיק להיות בקטגוריית "שריפה", שיחה זו תנותב ישירות אל המוקד הייעודי לטיפול בשריפות.

המוטיבציה בתהליך מסוג זה נובעת מתוך ההבנה שבאירועים מסוג "שריפה" ישנן פעולות בסיסיות שגם אנשים חסרי הכשרה מקצועית יכולים לבצע בכדי לתרום לטיפול באירוע החירום.
(לעומת אירועי חירום מסוג EMS בהם לרוב מדובר במקרים בהן נדרשת הכשרה מקצועית בתחום הרפואי)
למשל, בדיקת קריטריונים מסוימים שיכולים לסייע לצוות ההצלה להבין מול איזה סיטואציה הם עומדים להתמודד.

לכן, הקמת מוקד ייעודי בנושא זה יכול לתרום רבות בהגברת מקצועיות המענה של מוקד החירום.
כמו כן, כתוצאה מהקמת המוקד הייעודי והפניית פניות אליו, סך הפניות המועברות למוקד הכללי יפחת מה שיאפשר עבודה טובה ויעילה יותר.

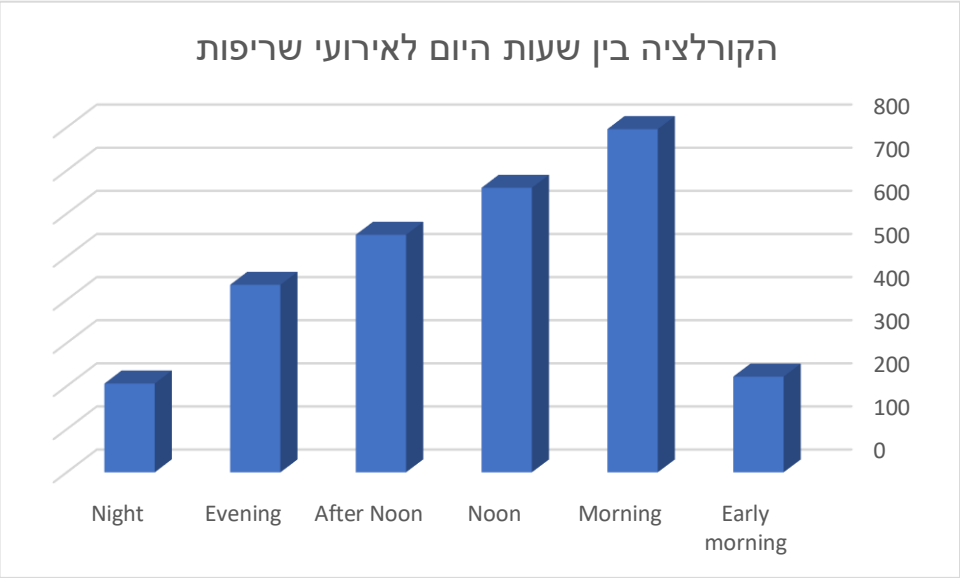
נספחים

הקורלציה שבין קודים אזוריים לבין כמות אירועי שריפות:



למען הנוחות הקריאה בוצעה המרה של הקודים האזוריים השונים לאותיות.
כפי שניתן לראות, ישנם שני מוקדים עיקריים בהם התרחשו רוב השריפות ואלו הם אזורים B ו-E.

הקורלציה שבין שעות היום לכמות אירועי שריפות:



כפי שניתן לראות, השעה העיקרית בה מדווח על אירועי שריפה היא שעת הבוקר (בין 05:00 עד 11:59).
אך באופן כללי ניתן לומר כי לאורך כל שעות היום ישנו דיווח על שריפה.