

# פרויקט למידת מכונה

מרצה : חן חנגן'

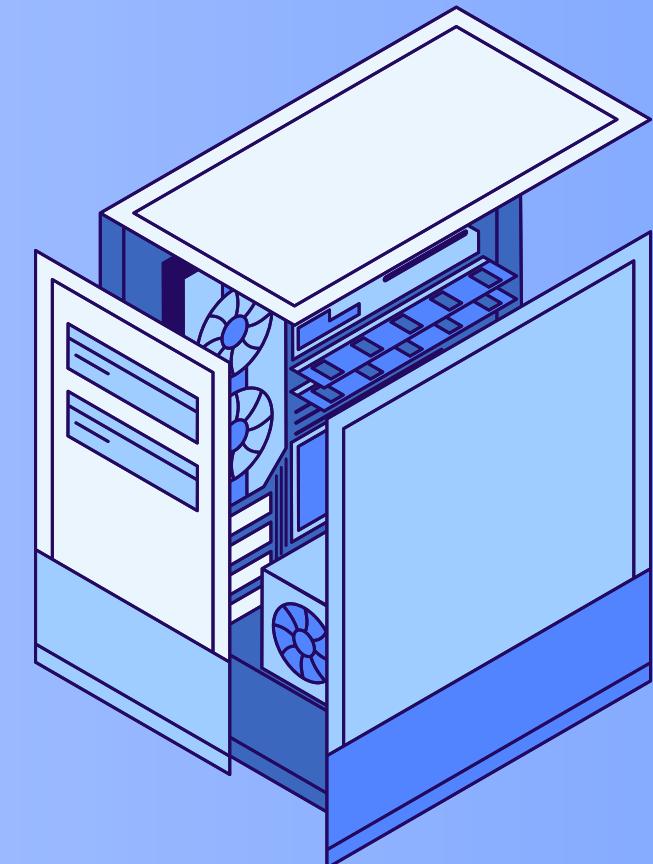
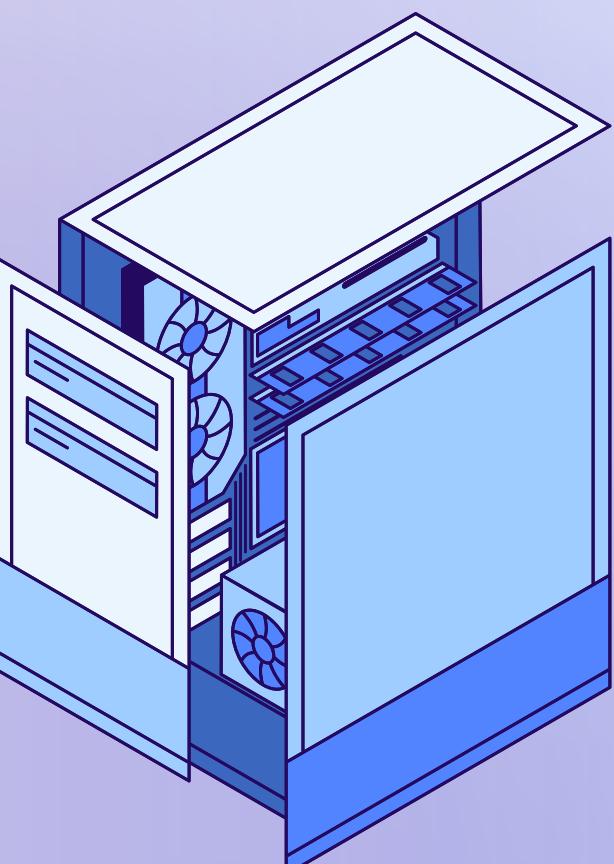
gal plg, orhab zr



# על הדאטה

בפרויקט ניתחנו>Data של משטרת אוסטין (טקסס) המתעד  
אירועי פשע سنאה בין השנים 2017–2025. הנושא על  
חשיבות רבה – ניתוח הנתונים יכול לעזור לרשויות אכיפת החוק  
ולקהילה להבין היכן, מתי וכלפי מי מתרחשים האירועים ואולי  
איפלו למנוע את האירוע הבא.

המטרה המרכזית של הפרויקט היא לפתח מודל למידת מכונה  
המסוגל לחזות את המנייע העומד מאחוריו פשע השנה, כגון  
גזענות, דת או נטייה מינית בהתבסס על מאפייני האירוע  
המדווחים בדאטה



# גישה לניתוח פשיי שנה

שימוש בסטטיסטיקה תיאורית לנתח  
מגמות ודפוסים היסטוריים

סיווג יידי של אירועים על בסיס חוקים  
והגדרות משפטיות

ניתוח איקוני על ידי מומחים מגופי אכיפת  
החוק

**שיטות אלו מוגבלות, דורשות זמן רב ועלולות להיות מושפעות מהטויות אנושיות**

למידת מכונה מציעה חלופה מבוססת-נתונים המאפשרת זיהוי דפוסים מורכבים, ניתוח וחיזוי מניעי פשיעה  
באופן עקבי ויעיל.

# DATA FEATURES

Month	Incident Number	Date of Incident	Day of Week	Number of Victims under 18	Number of Victims over 18	Number of Offenders under 18	Number of Offenders over 18	Race/Ethnicity of Offenders	Offense(s)	Offense Location	Bias	Zip Code	APD Sector
0	Jan 2017-241137	01/01/2017 12:00:00 AM	Sun	0	1	0	1	White/Not Hispanic	Aggravated Assault	Park/Playground	Anti-Black or African American	78704.0	Henry
1	Feb 2017-580344	02/01/2017 12:00:00 AM	Wed	0	1	0	1	Black or African American/Not Hispanic	Aggravated Assault	Highway/Road/Alley/Street/Sidewalk	Anti-White	78702.0	Charlie
2	Mar 2017-800291	03/21/2017 12:00:00 AM	Tue	0	0	0	0	Unknown	Destruction	Highway/Road/Alley/Street/Sidewalk	Anti-Jewish	78757.0	Ida
3	Apr 2017-1021534	04/12/2017 12:00:00 AM	Wed	0	0	0	0	White/Unknown	Simple Assault	Air/Bus/Train Terminal	Anti-Jewish	78723.0	Ida

Feature Name	מה הפיצ'ר מייצג	הסבר
Total Victims	מספר הקורבנות באירוע	חיבור קורבנות מתחת ומעלה גיל 18
Total Offenders	מספר התוקפים באירוע	חיבור תוקפים מתחת ומעלה גיל 18
Victim-Offender Ratio	יחס קורבנות-תוקפים	חישוביחס עם smoothing למניעת חלקה באפס
Day of Week	היום בשבוע של האירוע	חילוץ מהתאריך (0–6)
Weekend Indicator	האם האירוע התרחש בסופ"ש	בינארי: 1 = שבת/ראשון, 0 = יום חול
Season	עונה השנה של האירוע	Winter / Spring / Summer / Fall החודש
Zip Code Frequency	שכיחות אירועים באזור	ספרית מופיעים לכל Zip Code
Bias	מניע פשע השנה	אוחד לארבע קבוצות-על: מוצא אתני, נטייה מינית, דת, ואחר.
Offense Location Group	סוג מקום האירוע	מקומות > 5 מופיעים אוחדו ל-'Other_Location'

# METHODOLOGY

## Unsupervised Techniques

### Clustering

זהה קבוצות של אירועים דומים וחושף דפוסים הקשורים למנייע, מיקום או מאפייני קורבן ותוקף, ומספק תובנות נוספות למידה מונחית.

### Anomaly / Outlier Detection

מאתר אירועים חריגים, כגון מביעים נדירים או מספר קיצוני של קורבנות, עשויים להעיד על מקרים בעלי סיכון גבוה או חריגות מיוחדת.

### Supervised Models

#### Logistic Regression

מודל פשוט המאפשר להעריך את ההשפעה של כל פיצ'ר על סוג המנייע

#### Decision Tree

löcid קשרים לא-לינאריים בין פיצ'רים כגון מיקום, יום וסוג העבירה, ויוצר חוקים המפרידים בין סוגי מניעים שונים.

#### Random Forest

מודל של עצים רבים שטפחית התאמת יתר ומשפר את דיוק הסיווג.

#### Gradient Boosting

זהה דפוסים עדינים בנתונים המבדילים בין סוגי מניעים, ומתאים במיוחד למצבי שבהם ההבדלים קטנים בין הקטגוריות

#### Support Vector Machine (SVM)

יעיל בהפרדת סוגים מניעים כאשר קיימות אינטראקציות מורכבות בין פיצ'רים, באמצעות יצרת גבולות החלטה אופטימליים במרחב רב-ממדי.

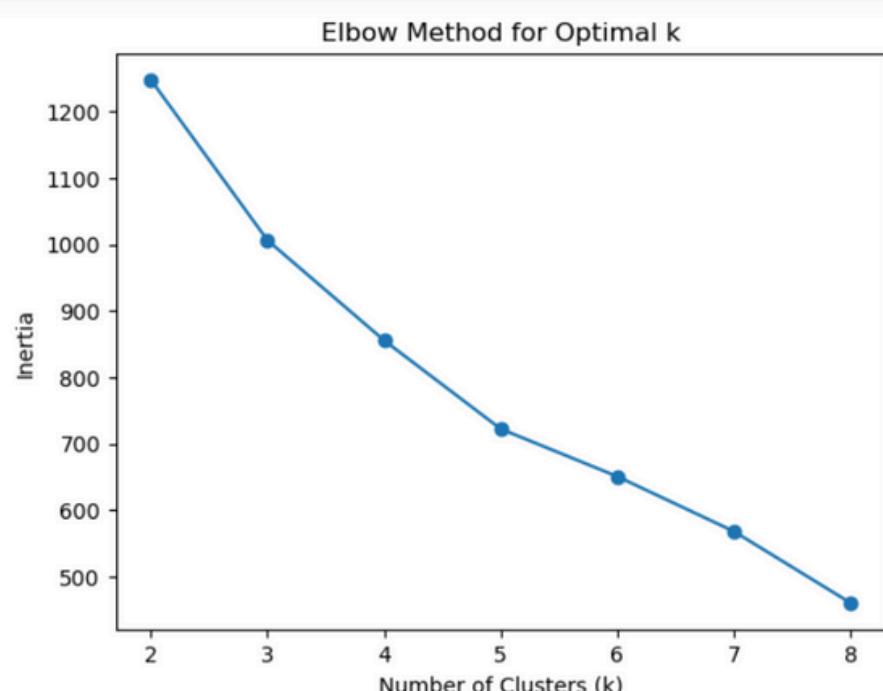


# ניסויים והערכת ביצועים



## למידה לא-מנוחית מערך הניסוי

- נבחרו פיצ'רים המציגים מאפיינים קורבן/תיקף, זמן ומקום Total Victims, Total Offenders, Victim-Offender Ratio, Zip Code Frequency, Is (Weekend, Is Offender Known)
- ערכים חסרים הושלמו ב-0 וונעשה שימוש ב-StandardScaler עלקב רגישות - K-Means לגדיל הפיצ'רים.
- מספר האשכולות נבחר באמצעות שיטת המרפק (Elbow Method for Optimal k), ונמצא כי  $k=5$  מספק איזון בין פירוט לאינטראטיביות. יושם Clustering K-Means, ובוצעה הפחיתת ממדים באמצעות PCA לצורך ויזואлизציה.



## מערך הניסוי

הדאטה חולק לשיט אימון וסט בדיקה. אומנו מספר מודלים מונחים לצורכי סיווג מניע פשע השנאה (Bias Super-Class). כל המודלים הוערכו תחת אותו מערך ניסויים לצורכי השוואת הוגנת.

## טכניות הערכה

- Accuracy – מדידת דיוק כולל.
  - Precision, Recall, F1-score
- הערכת איכות הסיווג ברמת המחלקות.

## שיטות לא-מנוחיות

- Clustering – ניתוח מבנה ודפוסים בתונונים.
- Outlier Detection – זיהוי אירועים חריגים וקיצוניים.

# RESULTS & CONCLUSION

## Supervised Models

דיק המודל המפוקח הגבואה ביותר שהושג היה 0.63 למרות שהתוצאה הזאת לא הכי גבוהה מדובר בתוצאה משמעותית בהתחשב בפייזור הרחב והמשתנה של הנתונים, ולא באיכות ירודה של המודל. גם לאחר מיזוג הנתונים ל-4 קטגוריות מרכזיות אומנם החיזוי עלה אך גודל הדאטה נשאר קטן יחסית ביחס לדפוסים המורכבים. ולכן השגת רמת חיזוי זו בתנאים אלו משמעותית ומשתפרת באופן משמעותי פניו סיכויי אקראי, ומוכיחה שהמודל הצליח לזהות את האותות הבסיסיים בנתונים.

1]:

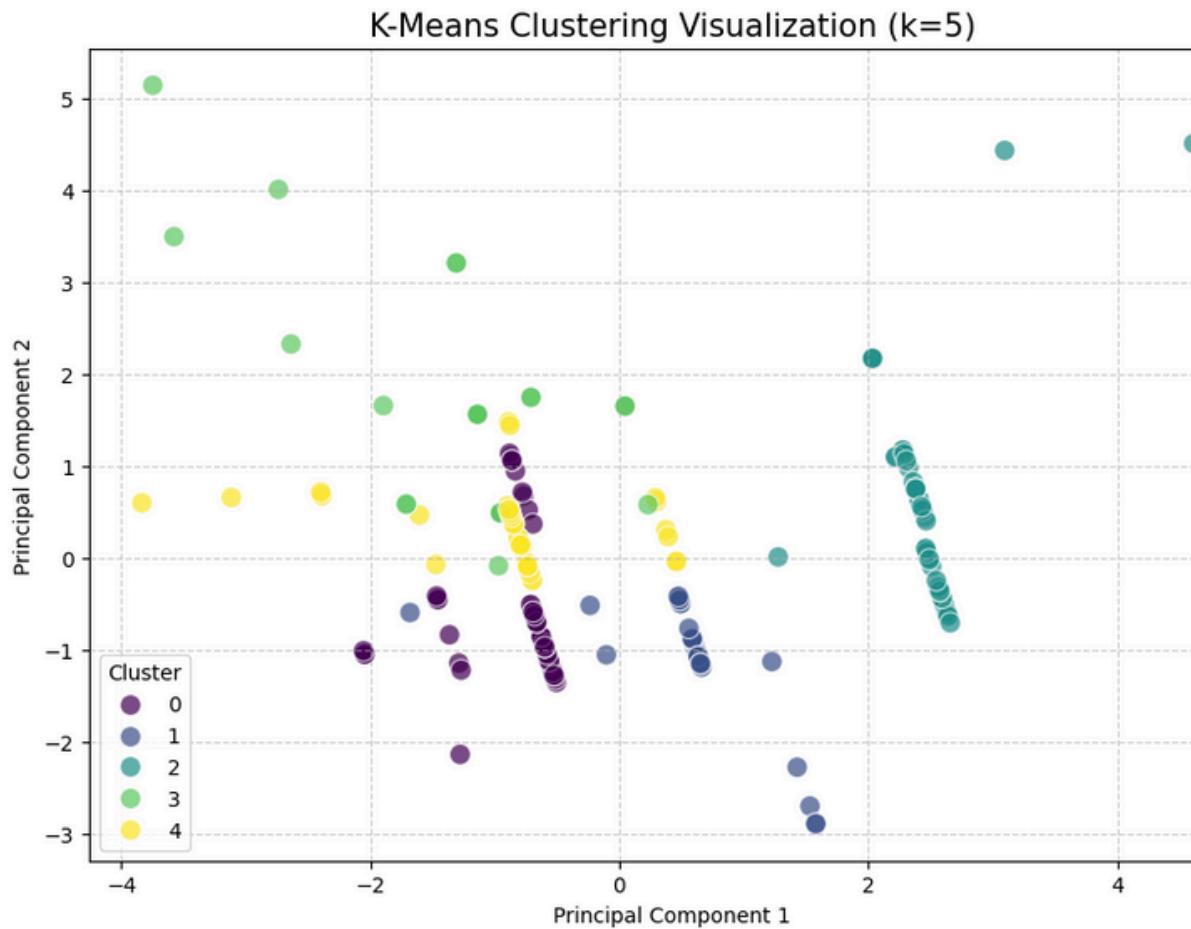
	Model	Accuracy	Weighted F1-score
0	Logistic Regression	0.48	0.49
1	Decision Tree	0.57	0.57
2	Random Forest (Tuned)	0.63	0.63
3	Gradient Boosting	0.57	0.57
4	SVM	0.52	0.52

# RESULTS

## & CONCLUSION

Unsupervised models:

### Clustering



**פשעים שנאה אינטלקטואליים:** קיימת דפוסיות לפי זמן (שבוע/סוף שבוע), מיקום (אזורים בערים מול רגעים) וסוג המקרה (אנונימי/ונדלים מול אירוע קבוצתי). המאפיינים שנבחנו מאפשרים זיהוי דפוסים נסתרים ומשלים את המודלים המפוקחים להבנת מגמות הפשעים.

Cluster	Total_Victims	Total_Offenders	Victim_Offender_Ratio	Zip_Code_Frequency	Is_Weekend	Is_Offender_Known
0	1.075472	1.122642	1.028302e+00	13.838095	0.000000	1.000000
1	0.833333	1.000000	7.750000e-01	11.400000	0.033333	0.000000
2	1.078125	0.000000	1.078125e+06	17.062500	0.343750	0.015625
3	1.162791	1.348837	9.399225e-01	47.372093	0.581395	0.860465
4	1.035088	1.228070	9.669591e-01	12.719298	1.000000	0.894737

**0-Cluster - פשע يومي ידוע:** מעשים יומיומיים ביום השבוע, לרובם בין אנשים שכיריהם זה זהה, עם קורבן אחד ופושע ידוע.

**1-Cluster - אירועים אוניברסליים שגרתיים ביום השבוע:** אירועים יומיימים ביום השבוע ללא פורע ידוע, לרוב מקרים עבירה קטנה או דיווחים בלי חשוד.

**2-Cluster – אוניברטיים/ונדליזם:** עבירות לא-אישיות או דיווחיות (כגון ונדליזם), ללא מעורבות ישירה של פורע וקורבן.

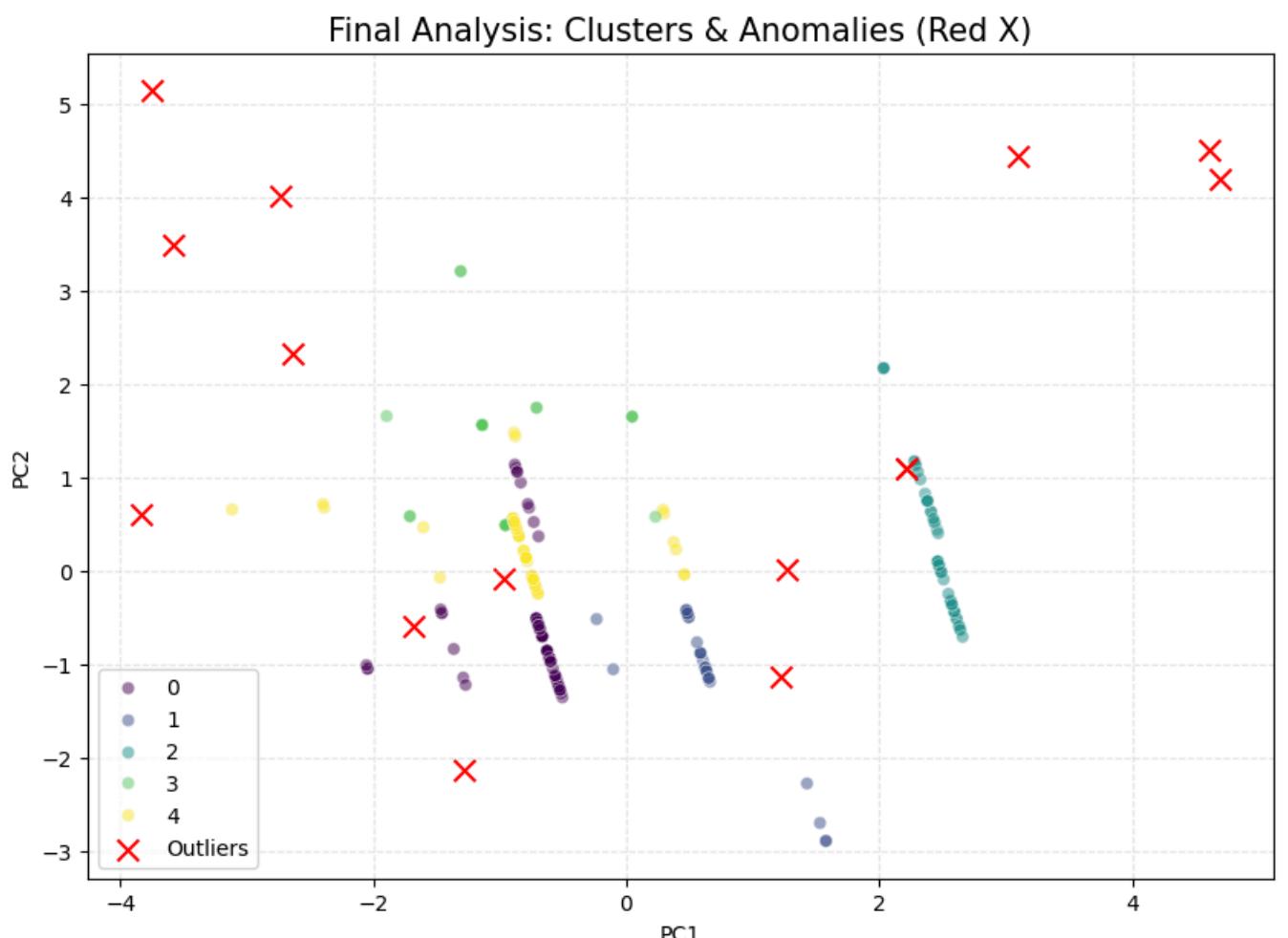
**3-Cluster- מוקדי פשע צפופים עם נתיחה לסופי שבוע:** מוקדי פשע צפופים, בעיקר בסופי שבוע, עם פורעים מזוהים לעתים קרובות.

**4-Cluster – אירועי עימות בסופי שבוע:** אירועים תוקפניים בסופי שבוע, עם מעורבות ישירה של פורעים וזהויות גבוהות שלהם.

# RESULTS & CONCLUSION

Unsupervised modrls:

## Outlier Detection



Total_Victims	Total_Offenders	Victim_Offender_Ratio	Super_Bias_Class
21	1	4	0.25 RACE_ETHNICITY
22	0	2	0.00 RACE_ETHNICITY
37	2	4	0.50 SEXUAL_GENDER
80	3	0	3000000.00 RACE_ETHNICITY
159	2	0	2000000.00 SEXUAL_GENDER

נקודות רגילהות: 283  
חריגות: 15

- חלק מהאירועים כוללים מספר חריג של עבריינים (למשל 4), דבר נדיר בפשע שנאה.
- אירועים עם 0 עבריינים אך קורבות גביה יוצרים יחסים קורבן-ערביין קיצוני, לרוב מקרים אוניבימיים כמו ונדייז או גרפייטי.
- מקרה אחד ללא קורבות מצביע על פשע שנאה ללא מגע ישיר, גם הוא כנראה גרפייטי או נזק לרכוש.

# FUTURE DIRECTIONS

הגדלת אוסף הנתונים תאפשר למודלים ללמידה מגוון רחב יותר של מקרים ולשפר את דיוק התוצאות

התמקדות באזוריים עירוניים עם ריכוז גבוה של פשעים ובמקרים המתרחשים בסופי שבוע יכולה לשיע בгибוש אסטרטגיות מניעה מוקדמות.