**תרגיל למידת מכונה 2 – הגשה: 27.5.21**

תרגיל זה יעסוק בנתוני פשיעה המובאים ממשטרת ישראל – אפיון, חיזוי והצגת המסקנות.

1. נתונים:

נתוני הפשיעה על פי משטרת ישראל נכון לשנת 2019:

* [crimes](crimes.xlsx)

מאגר הישובים הממשלתי ונתוני היישובים:

* [קובץ היישובים 2018](https://www.cbs.gov.il/he/publications/doclib/2019/ishuvim/bycode2018.xlsx)
* [מקרא לקובץ היישובים.](https://www.cbs.gov.il/he/publications/doclib/2019/ishuvim/index%202018.xlsx)

1. תאר את הנתונים בקצרה:

* גרפים.
* טבלאות.
* מטריצת קורלציה.   
  לדוגמה - האם ניתן לראות קורלציה בין גודל יישוב לבין כמות הפשיעה בו?

1. בעזרת שני (לפחות) אלגוריתמיCLUSTERING **לבחירתכם**, מצאו קשר בין נתוני הפשיעה למאפיינים דמוגרפים.

לדוגמה: כל הקיבוצים מבצעים עבירות תנועה ביחס לגודל האוכלוסייה שלהם.

מבין האלגוריתמים שהשתמשתם, ציינו מהו האלגוריתם המתאים ביותר.

באילו הנחות השתמשתם?

1. השתמשו ב-RANDOM FOREST וחזו מה הוא סוג העבירה **השני** הנפוץ ביותר בשנת 2019 ביישובים הבאים: כרמיאל, נתניה, כפר יאסיף, ראש העין, מטולה, ביתר עילית.  
   לדוגמה: בחיפה ב-2019 סוג העבירה השני הנפוץ ביותר הוא עבירות סדר ציבורי.
2. השתמשו ב- ADABOOSTע"מ לחזות את מספר העבירות הכללי בשנת 2019 עבור היישובים: קריית אתא, ראש פינה, אילת, סכנין ובאר שבע. הציגו את התוצאות הרלוונטיות בטבלה ובגרף מתאימים.

|  |  |
| --- | --- |
| **סוג העבירה (x)** | **שעות עבודה בממוצע לעבירה (h)** |
| עבירות בטחון | 20 |
| עבירות כלכליות | 5 |
| עבירות כלפי המוסר | 15 |
| עבירות כלפי הרכוש | 10 |
| עבירות מין | 35 |
| עבירות מרמה | 25 |
| עבירות נגד אדם | 40 |
| עבירות נגד גוף | 30 |
| עבירות סדר ציבורי | 1 |
| עבירות רישוי | 1 |
| עבירות תנועה | 2 |

1. השתמשו בשנים 2014-2018 ועזרו למשטרת ישראל להחליט באילו יישובים עליה לתגבר/להפחית/לא לבצע שינוי במספר כוחותיה באמצעות שוטרים ומשאבים נוספים. מדידת המשאבים תתבצע לפי שעות עבודה.   
   נגדיר פונקציה המחשבת את שעות העבודה בהתאם לסוג העבירה:  
     
    , כאשר:

N – מספר העבירות סה"כ עבור אותו יישוב לשנה ספציפית.

X – סט העבירות עבור אותו יישוב לשנה ספציפית.

השתמשו באלגוריתם מתאים לבחירתכם על מנת לעזור למשטרת ישראל. הציגו את התוצאות ב-CONFUSION MATRIX ו-SCATTER PLOT מתאימים. (המלצה: נסו להראות ב-scatter plot את תשובת המודל אל מול הנתונים המקוריים עבור כל יישוב).

1. הריצו שוב את האלגוריתם שבו השתמשתם בסעיף 6 והיעזרו בתוצאות של אלגוריתם ה-CLUSTERING שבחרתם כמתאים ביותר בסעיף 3 (הגדירו אותן כ-feature נוסף).  
   תארו את השינוי בתוצאות במידה וקיים. האם השימוש בתוצאות ה-CLUSTERING משפיע על החלטת האלגוריתם?
2. סכמו את העבודה במצגת קצרה (מצגת מנהלים- עד שישה שקפים) המסבירה את תהליך העבודה, תוצאות (חווה דעתך על משמעות התוצאות) והמלצות. יקבע תאריך בו תדרשו להציג בקצרה את עבודתכם אל מול הכיתה.

**קווים מנחים:**

נא לקרוא היטב את החלק הבא, טיב העבודה והציונים יהיו תלויים בסעיפים הבאים:

* **בחנו את המידע בצורה גרפית לפני שאתם ניגשים לתהליכים מורכבים יותר.**
* **אל תפעילו תהליכים/ מודלים בצורה עיוורת. חישבו על הנתונים ועל שאלת המחקר.**

הגשה:

* מחברת Jupyter notebook בפורמט html.
* מחברת פיתון (ipynb) אחת.
* מצגת שמורה בקובץ PDF.
* כל קובץ נתונים אחר שהיה בו צורך.
* ההגשה היא בזוגות (פורום למציאת שותפים קיים במודל).
* תאריך הגשה: **27.5.2021**  
  הגשה באיחור (ללא סיבה מוצדקת שתאושר מראש) תגבה 3- נקודות עבור כל יום איחור.   
  לאחר חמישה ימי איחור, העבודות לא תיבדקנה.

נתונים:

* הנתונים מובאים בצורה גולמית, עליכם להפוך אותם לראויים לעבודה (נדרשת לא מעט עבודה).
* כתבו כהערה בקוד את רשימת המשתנים (כפי שמופיעים בקוד שלכם) והמשמעות שלהם.
* חלק מהנתונים רלוונטיים וחלקם לא, עליכם לברור את הנתונים התורמים למודלים (variable importance, variable selection…).
* חלקו את הנתונים לאימון, ולידציה ומבחן בצורה הגיונית.

קוד:

* הקוד צריך לרוץ ללא אזהרות או שגיאות.
* תיעוד טוב הוא **קריטי**.
* יש לפרט את התהליכים/פעולות שביצעתם בקוד – **עבור כל סעיף**.
* השתמשו בחבילות מוכרות ובעלות הסברים מפורשים.
* כאשר אתם מציירים גרפים חובה להוסיף את הדברים הבאים: שמות צירים ברורים, כותרת ומקרא אם יש צורך.
* נא לציין את סעיפי התרגיל גם בקוד.
* השתמשו בשמות איברים משמעותיים.
* אל תשמשו במילים שמורות.
* נדרש מכם לכתוב לפונקציות, לולאות או כל פתרון אחרון.
* השתמשו בקבועים היכן שניתן.
* הגדירו SEED בתחילת הקוד.
* לכל מודל תבוצע אופטימיזציה.
* עליכם לבחור את המדדים (gini, mse, entropy, acc…) המתאימים או אלו שיובילו לתוצאה הכי טובה בהינתן שאלת מחקר. שימו לב, עליכם ללמוד ולחפש את סוגי המדדים הרלוונטיים על מנת להגיע לתוצאה טובה.
* השתמשו בגרפים רלוונטיים וברורים היכן שניתן.

מצגת:

* המצגת צריכה להתאים לתוצאות הקוד. כלומר, ניתן יהיה להוכיח טענות או מספרים דרך הקוד.
* אל תכללו פרטים לא חשובים כמו שמות קבצים, גרפים שניתן לתאר בשתי מילים, קטעי קוד או גרפים גדולים מאוד.
* נקודת הנחה: קורא המצגת בעל רקע בתחום אך איננו מומחה.
* עליכם להסביר את התוצאות והתהליך שהוביל להשגתן.
* המצגת יכולה להיות באנגלית או בעברית (מומלץ באנגלית).

**ניקוד וציון:**

התרגיל ינוקד על פי הקריטריונים הבאים:

* נכונות
* בהירות
* יעילות המימוש
* אלגנטיות המימוש

לימוד עצמי:

* כפי שציינו בתחילת הקורס, לימוד עצמי הוא חלק חשוב מהקורס.
* התייחסו בזהירות ובצורה ביקורתית לכל מקור מידע.
* ניתן ומומלץ להתייעץ עם שאר הסטודנטים בקורס, אך כל זוג נדרש לכתוב את עבודותיו לבד.
* סביר להניח שלא כל התוצאות והאלגוריתמים יהיו זהים.

שאלות ושעות קבלה:

* נא לפרסם את השאלות **בפורום במודל**, לאחר שקראתם את ההודעות הקודמות (שאלות בנושא מקצועי שישלחו במייל לא יענו).
* במידה ומעוניינים לקבוע שעת קבלה עם אחד המתרגלים – **יש לשלוח לפני כן את השאלות במייל.**
* בכל מקרה אחר (שאלות אישיות, בקשה להארכה עם סיבה מוצדקת וכו') יש לשלוח מייל למתרגל/ת.

בהצלחה,

צוות הקורס.