**תרגיל למידת מכונה 2**

תרגיל זה יעסוק בנתוני מערכות הבחירות האחרונות- אפיון, חיזוי והצגת המסקנות.

1. נתונים:

תוצאות שלושת הבחירות האחרונות:

* <https://votes23.bechirot.gov.il/>
* <https://votes22.bechirot.gov.il/>
* <https://votes21.bechirot.gov.il/>

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.

1. בעזרת שני (לפחות) אלגוריתמי CLUSTERING לבחירתכם, מצאו קשר בין נתוני ההצבעות למאפיינים דמוגרפים.

לדוגמא: כל הקיבוצים מצביעים למפלגה X.

מבין האלגוריתמים שהשתמשתם, ציינו מהו האלגוריתם המתאים ביותר.

באילו הנחות השתמשתם?

1. השתמשו ב-RANDOM FOREST ע"מ לחזות את כמות הקולות הכשרים בבחירות האחרונות ביישובים חיפה, איילת השחר, אילת, סכנין וקצרין. הציגו את התוצאות הרלוונטיות בטבלה ובגרף מתאימים.
2. השתמשו ב-ADABOOST ע"מ לחזות את כמות המצביעים למפלגות: ליכוד, הרשימה המשותפת וכחול לבן ביישובים ירושלים, בני ברק, סכנין, כרמיאל, ודאלית אל כרמל. הציגו את התוצאות בטבלה ובגרף מתאימים.
3. לאחר שתי מערכות בחירות, ועדת הבחירות מעוניינת לחזות באילו ישובים עליה להציב נציגים על מנת למזער את כמות הקולות הפסולים. השתמשו באלגוריתם מתאים לבחירתכם על מנת לעזור לועדת הבחירות. הציגו את התוצאות ב-CONFUSION MATRIX מתאימה.
4. 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 פיתון (ipynb) אחת.
* מצגת שמורה בקובץ PDF.
* נא לא להגיש קובץ ZIP אחד.
* כל קובץ נתונים אחר שהיה בו צורך.
* ההגשה היא בזוגות בלבד.
* תאריך הגשה: 16.5.2020.

**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.
* 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 ממכם לכתוב לפונקציות, לולאות או כל פתרון אחרון.
* הגדירו SEED בתחילת הקוד.
* השתמשו בקבועים היכן שניתן.
* 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 (ממליץ על אנגלית).

**לימוד עצמי:**

* כפי שציינתי כבר בעבר, 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 והמודלים יהיו זהים.

**שאלות ושעות קבלה:**

* שאלות שישלחו במייל לא יענו. נא לפרסם את השאלות בפורום במודל, לאחר שקראתם את ההודעות הקודמות.
* נא להגיע ממוקדים עם רשימת שאלות לשעות קבלה.