**תרגיל למידת מכונה 3**

תרגיל זה יעסוק בחיזוי סיכויי הישרדותם של נוסעי ספינת הטיטאניק.

1. נתונים ותיאור הבעיה:

קובץ אימון וקובץ מבחן

<https://www.kaggle.com/c/titanic>

מדריך לאתר:

<https://www.kaggle.com/alexisbcook/titanic-tutorial>

1. תיאור הנתונים:

* גרפים.
* טבלאות.
* כל כלי רלוונטי אחר.
* התשתמשו גם באלגוריתמי ויזואליזציה שלמדתם.

1. הכנת הנתונים:
   * ארגון הנתונים והשלמת מידע.
2. למידת מכונה:
   * בחלק הזה עליכם למשחק עם המודלים שנלמדו (אפשר גם אחרים) כדי לקבל את תוצאות הסיווג הכי טובות.
   * חלק זה יכלול:
     1. שיטות מבוססות עצים.
     2. SVM.
     3. KNN
     4. Neural nets.
     5. PCA
     6. שיטה נוספת לבחירתכם.
     7. השתמשו באלגוריתם unsupervised clustering כלשהו והעזרו בתוצאות על מנת לשפר את אחוזי הדיוק.
3. דיון ומסקנות:
   * הציגו טבלה מסכמת המציגה את תוצאות כל המודלים (עפ"י התוצאות שבדקתם ב-KAGGLE)
   * הסבירו את תוצאות המודלים במילים.
   * תארו את הבעיות איתם התמודדתם ואת הפתרון (אופטימיזציה לפרמטרים, השלמת ערכים וכו').
   * סכמו את העבודה במצגת קצרה (מצגת מנהלים- עד שמונה שקפים) המסבירה את תהליך העבודה והתוצאות.

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.**
* **הקוד הוא הדו"ח שלכם. כל פעולה שאתם עושים צריכה להיות מוסברת וכתובה. כל התהליך אמור להיות מתועד היטב- לא רק בהערות קוד, אלא בהסברים ברורים.**

**הגשה:**

* מחברת פיתון (ipynb) אחת.
* קובץ html של הקוד כולל הפלטים.
* מצגת שמורה בקובץ PDF.
* נא לא להגיש קובץ ZIP אחד.
* כל קובץ נתונים אחר שהיה בו צורך.
* ההגשה היא בזוגות בלבד.
* תאריך הגשה: 26.6.2020 (זהו היום האחרון של הסמסטר- נא להיערך בהתאם- לא יינתנו הארכות).

**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…).

**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.