**תרגיל למידת מכונה 3 –**

**הגשה: 18.6.21 בשעה 12:00**

תרגיל זה יעסוק בחיזוי שביעות רצון של נוסעים עבור חברות טיסה שונות (Airline Passenger Satisfaction).

1. נתונים ותיאור הבעיה:

קובץ אימון וקובץ מבחן

<https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction>

* + קיימים רפרנסים שונים תחת לשונית code – שם תוכלו לקבל **השראה** לניקוי נתונים, ויזואליזציה, קורלציה והפעלת מודלים שונים לשאלות מחקר שונות.

אנו ממליצים לעיין במחברות אלו טרם ביצוע משימותיכם.

* + Kaggle מספקת הרבה meta data לגבי כל אחד מה-datasets. תוכלו לראות זאת בעמוד הראשי של הקישור המצורף. באמצעות הנתונים הללו – תוכלו לבנות לעצמכם סדר פעולות שיש לבצע עבור שלב ה-pre processing (האם נדרש ניקוי נתונים? האם יש צורך לבצע scaling? וכו').

1. תיאור הנתונים:

* גרפים
* טבלאות

ניתן לשלב בין שני הסעיפים הללו – כלומר ניתן להציג ויזואליזציות עבור הנתונים הגולמיים ויזואליזציות עבור הנתונים לאחר ארגונם.

* מטריצת קורלציה
* השתמשו גם באלגוריתמי ויזואליזציה שלמדתם

1. הכנת הנתונים:
   * ארגון הנתונים
   * השלמת מידע חסר
   * ביצוע טרנספורמציה על הנתונים הגולמיים
   * שימוש בשיטות נוספות שנלמדו במידת הצורך
2. חברות התעופה מעוניינות לדעת אילו פיצ'רים משפיעים בעיקר על שביעות רצונם של הלקוחות. עליכם להשתמש בשיטות שונות על מנת לבסס את תשובתכם, באמצעות הכלים הבאים:

* אלגוריתמי אשכול (clustering)
* אלגוריתמי וויזואליזציה: PCA, t-SNE, SOM – יילמד בהמשך (למעוניינים, קיימות דוגמאות רבות למימוש ב-kaggle ואתרים נוספים).
* בניית פיצ'רים חדשים/ביצוע טרנספורמציה לפיצ'רים קיימים (לדוגמא: איגוד בין פיצ'רים קיימים)  
  ועוד...

1. השתמשו במודלי קלסיפיקציה שונים שלמדנו על מנת למקסם את בעיית הסיווג של שביעות הרצון (ה-label הינו הפיצ'ר האחרון ב-dataset הגולמי – satisfaction):
   * + שיטה מבוססות עצים.
     + ADABOOST

בחרו 3 מן השיטות הנ"ל

* + - SVM.
    - Neural nets.
    - שיטה אחת נוספת לפחות שלא למדנו (לבחירתכם).

יש להציג את תוצאותיכם בצורות ויזואליות שונות (למשל: confusion matrix).

1. **א.** נרצה להגדיר label חדש שמייצג את מידת שביעות הרצון של הלקוחות בטווח של 0-5.   
   לשם הגדרת ה-label – נבצע ממוצע לתת קבוצה (בין 3-5) מהפיצ'רים הבאים לבחירתכם (כאשר למעשה הממוצע של ערכי הפיצ'רים הללו יהווה את מידת שביעות הרצון של הלקוחות):

* Inflight wifi service
* Departure/Arrival time convenient
* Ease of Online booking
* Gate location
* Food and drink
* Online boarding
* Seat comfort
* Inflight entertainment
* On-board service
* Leg room service
* Baggage handling
* Checkin service
* Inflight service
* Cleanliness

**ב.** השתמשו ב- 2 מודלי רגרסיה שונים (לפחות) על מנת לחשב את **מידת** שביעות הרצון של הלקוחות (הממוצע יהיה ה-label החדש שהגדרנו בסעיף א). בצעו השוואה בין המודלים השונים באמצעות מדדים שונים שקיימים בספריית sklearn והציגו אותה בצורה גרפית.

**דאגו להוריד את הפיצ'רים שבהם השתמשתם לצורך הגדרת ה-label.**(כלומר הפיצ'רים אותם בחרתם להגדרת ה-label **לא** יהיו חלק מסט האימון וסט המבחן שלכם)**.**

**ג.** הריצו את המודלים מסעיף ב' **ביחד עם הפיצ'רים שהגדרתם איתם את ה-label**.   
(כלומר הפיצ'רים אותם בחרתם להגדרת ה-label **כן** יהיו חלק מסט האימון וסט המבחן שלכם).  
כיצד זה השפיע על התוצאות מסעיף קודם? בצעו השוואה ונמקו.   
טיפ: השתמשו בכלים כגון: feature importance, histogram וכו'.

1. דיון ומסקנות:
   * הציגו טבלה מסכמת המציגה את תוצאות כל המודלים ע"פ ההשוואות שביצעתם.
   * הסבירו את תוצאות המודלים במילים.
   * תארו את הבעיות איתן התמודדתם ואת הפתרון (אופטימיזציה לפרמטרים, השלמת ערכים וכו').
   * סכמו את העבודה במצגת קצרה (מצגת מנהלים- עד שמונה שקפים) המסבירה את תהליך העבודה, התוצאות (חווה דעתך על משמעות התוצאות) וההמלצות.

**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 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: **18.6.21 בשעה 12:00 (שימו לב זהו היום האחרון לסמסטר – לא יינתנו הארכות נוספות).**  
  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…).

קוד:

* הקוד צריך לרוץ ללא אזהרות או שגיאות.
* תיעוד טוב הוא **קריטי**.
* יש לפרט את התהליכים/פעולות/מסקנות 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
* הצגת תוצאות ומסקנות עבור כל סעיף

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.