TEL AVIV UNIVERSITY אוניברסיטת תל-אביב

משימה 2

הפקולטה להנדסה
המחלקה להנדסת תעשייה וניהול
מבוא למדעי הנתונים (Introduction to Data Science)
-סמסטר א' תשע"ז

מבוא:

עיקרי מטרות משימה זו הם הכרה ותרגול של שני המרכיבים הנוספים של מודל CRISP-DM: modelling ו- evaluation.

במשימה זו, יעשה שימוש ב-3 בסיסי נתונים: Carvana[1] , Diabetes[2] , ו- Movies[3] (מהמשימה הראשונה).
כל בסיסי הנתונים נמצאים באתר הקורס במודל.

- בסיס הנתונים Carvana-פורסם במסגרת תחרות שפורסמה ב-Kaggle.com עם פרס בסכום של 5,000$.. בסיס נתונים זה מכיל נתונים אודות רכבים המוצעים למכירה במכירות פומביות, משתנה המטרה הוא האם הקנייה של רכב זה היא טובה/לא טובה. מידע נוסף אודות התחרות אפשר למצוא בקישור מטה[4] .
- בסיס הנתונים Diabetes- מכיל נתונים רפואיים של נשים מעל גיל 21 שמגיעות מרקע הודי, משתנה המטרה הוא האם האישה סובלת/לא סובלת מסכרת
- בסיס הנתונים Movies- מוכר לכם מהמשימה הקודמת

קווים מנחים:

נקודות תורדנה על חוסר הצמדות לקווים המנחים:

- ☒ השתמש בקובץ קוד המופיע במודל. **אין להסיר ממנו תגובות!!!**
- ☒ שאלות ייענו בפורום **בלבד.**
- ☒ בבדיקה יינתן דגש על יעילות הקוד אז שימו לב.
- ☒ הקוד להיות מלווה בהסברים לצד כל שורה.
- ☒ הניקוד יופיע בתחילתה של כל שאלה בסוגריים והוא יוכל לספק אינדיקציה לרמת הקושי שלה (כך שאם אתם נתקעים, נסו להמשיך הלאה).
- ☒ הציון הסופי שלכם יושפע גם מ'קריאות' הקוד שלכם, אז שימו לב לבצע זאת באופן קצר ומתומצת והשתמשו בשמות משמעותיים עבור המשתנים.

הגשה:

מועד אחרון להגשה: 16/02/2017
קבוצות: 2 תלמידים
אופן ההגשה: קובץ R שיועלה למודל על ידי **אחד** מחברי הקבוצה. לא לשכוח להכניס ת.ז של כל חברי הקבוצה בראש הקוד
ניקוד: 1-100 , קוד שלא ירוץ יזכה ב-0 נקודות
מייל עוזר הוראה: omermiran@gmail.com

טיפים מומלצים:

- ☒ במהלך התרגול נדרש להתקין חבילות רבות לאורך כתיבת הקוד. בכדי להתקין חבילה ב R יש להשתמש בפקודה `install.packages` (חפשו את הפקודה בגוגל)
- ☒ לקבלת עזרה על פונקציה מסויימתיש להוסיף לשם הפונקציה "?" לפניה, לדוגמא: `?nameOfFunction` , בתחתית עמוד העזרה (help page) שייפתח תוכלו למצוא בדרך כלל דוגמאות טובות לשימוש הפונקציה. תוכלו להשתמש בדוגמאות אלו ולשנות בהתאם לצורכיכם.
במידה ואין לכם את החבילה של אותה פונקצייה, לא תראו עמוד עזרה לפונקצייה ותצטרכו להתקין את החבילה כפי שהוסבר בסעיף הקודם.

בהצלחה!!!

1- Taken from: https://www.kaggle.com/c/DontGetKicked/data
2 -Taken from: https://www.kaggle.com/uciml/pima-indians-diabetes-database
3 -Taken from: https://vincentarelbundock.github.io/Rdatasets/doc/ggplot2/movies.html (from previews assignment)
4 -more info about the data and the competition: https://www.kaggle.com/c/DontGetKicked

TEL AVIV UNIVERSITY אוניברסיטת תל-אביב

משימה 2

הפקולטה להנדסה
המחלקה להנדסת תעשייה וניהול
מבוא למדעי הנתונים (Introduction to Data Science)
-סמסטר א' תשע"ז

Preparations:
הגדר את ספריית (סביבת) העבודה שלך להיות תיקיית המשימה

 # CARVANA #

1. PREPERATION

(0) **1.a.** טענו את הנתונים בפורמט המתאים למשתנה מסוג Data Frame וקרא למשתנה 'data.c'

(1) **1.b.** קבעו ערך אקראי (random seed) לערך מסויים כדי שבבדיקת המודל התוצאות לא יושפעו מרנדומליות

(2) **1.c.** מכיוון שבסיס הנתונים Carvana גדול ויגרום למודלים לרוץ במשך זמן רב, קחו דגימה של 30,000 שורות ושמרו אותם לאותו משתנה בשם 'data.c'

(3) **1.d.** פצלו את הדגימות לשני משתנים 'train.c' ו- 'test.c' כאשר במשתנה האימון 70% מהדגימות ובמשתנה הבדיקה 30% מהדגימות. פצלו את הנתונים תוך התחשבות במשתנה המטרה.

2. FEATURE SELECTION AND CORRELATION

(2) **2.a.** המירו את כל הפיצ'רים הפקטוריאלים למשתנים נומריים (אכן, צעד זה יכול לגרום לטעויות במודלים, אך לשם הפשטות בצעו המרה רגילה ממשתנה פקטוריאלי לנומרי).

(1) **2.b.** המירו את משתנה המטרה 'IsBadBuy' למשתנה פקטוריאלי עם 2 רמות

(4) **2.c.** הציגו את ה- correlation plot של המשתנים, וודאו שהגרף ברור לעין.

(2) **2.d.** מצאו את הפיצ'רים שהמתאם ביניהם מעל 0.65.

(1) **2.e.** שמרו 2 משתנים מסוג Data Frame אשר יחזיקו את הנתונים שנמצאים במשתנים 'train.c' ו- 'test.c' ללא הפיצ'רים בעלי המתאם הגבוה וקראו להם באותם שמות רק עם הסיומת noHightCor. (ראה קובץ R).

3. KNN

(4) **3.a.** בעזרת המשתנים החדשים (ללא הפיצ'רים עם המתאם הגבוה), חזו את התוצאות למשתנה 'test.c' בעזרת מודל ה-KNN עם k=1.

(2) **3.b.** הציגו את ה- confusion matrix

(5) **3.c.** בעזרת שימוש ב-cross-validation אמנו את מודל ה-KNN. השתמשו בפרמטרי קלט בכדי לבצע center ו scale והריצו את המודל על כמה k-ים (לא יותר מ-3). זמן הריצה לא יעלה על 2 דק'.

(2) **3.d.** השתמשו במודל שאימנתם בכדי לחזות בשנית את התוצאות למשתנה 'test.c'

4. ROC

(6) **4.a.** הציגו את עקומת ה-ROC של המודל שאימנתם.

הפקולטה להנדסה
המחלקה להנדסת תעשייה וניהול
מבוא למדעי הנתונים (Introduction to Data Science)
-סמסטר א' תשע"ז

5. PCA

(3) **5.a.** השתמשו במשתנה 'train.c' בכדי למצוא מרכיבים עיקריים (principal components). שימו לב למשתני הקלט.

(2) **5.b.** הציגו את הירידה בשונות המוסברת על ידי ה .PC's.

(6) **5.c.** בעזרת ה-PC's שיצרתם קודם לכן, צרו שני משתנים חדשים מסוג Data Frame בשמות 'train.c.pca' ו 'test.c.pca' כשבתוכם הפיצ'רים יהיו ה PC's שיצרתם. (פעולה זו מאתגרת אז חפשו דוגמאות באינטרנט)

(3) **5.d**. בעזרת שלושת ה PC's הראשונים בלבד- התאימו מודל KNN (כמו בסעיף 3) בעזרת K=7

(2) **5.e**. הציגו את ה confusion matrix (השתמשו באותו קוד מהחלק הראשון של הניתוח רק עם הנתונים החדשים).

 # DIABETES #

6. PREPERATION

(0) **6.a.** טענו את הנתונים בפורמט המתאים למשתנה מסוג Data Frame וקרא למשתנה 'data.c'

(2) **6.b.** אפשר להבחין כי להרבה מהדגימות יש ערך 0 בפיצ'ר 'SkinThickness'. נתון זה אינו הגיוני במיוחד ואפשר להניח שמודר בערך חסר (missing value). הפכו את כל ערכי 0 בפיצ'ר זה ל-NA.

(2) **6.c.** השלימו את הערכים החסרים מהסעיף הקודם באמצעות התוחלת (mean)

7. LOF

(5) **7.a.** הציגו את הצפיפות של תוצאות ה-LOF באמצעות כל הפיצ'רים (density plot)

(4) **7.b.** בהתבסס על התצוגה מהסעיף הקודם, הוציאו את כל הנקודות השקריות (outliers) מעל רמה מסוימת של תוצאת LOF.

8. SVM

(1) **8.a.** פצלו את הדגימות לשני משתנים 'train.d' ו- 'test.d' כאשר במשתנה האימון 70% מהדגימות ובמשתנה הבדיקה 30% מהדגימות. פצלו את הנתונים תוך התחשבות במשתנה המטרה.

(4) **8.b.** צרו מודל SVM עם הפיצ'רים הנראים לכם רלוונטיים. ציינכם בסעיף זה ייקבע על ערך השגיאה. השתמשו במודל שיצרתם על מנת לנבא את משתנה המטרה ב'test.d'. שמרו את התחזיות במשתנה בשם 'res'.

(1) **8.c.** חשבו את ערך השגיאה.

(6) **8.d**. כוונו את מודל הSVM: השתמשו בעד 5 עלויות (costs), 5 גמות (gammas) ו-5 CV. נקודות מלאות יינתנו באם השגיאה תהיה מתחת ל-23%

(3) **8.e**. הציגו את המודל הטוב ביותר שהושג (הפרמטרים שלו) והשתמשו בו בכדי לחזות בשנית את משתנה המטרה ב'test.d' ושמרו את תחזיות אלו במשתנה 'res2'

(1) **8.f**. הראו האם ערך השגיאה שופר לאחר כיוונון המודל.

הפקולטה להנדסה
המחלקה להנדסת תעשייה וניהול
מבוא למדעי הנתונים (Introduction to Data Science)
-סמסטר א' תשע"ז

9. RANDOM FOREST

(6) **9.a.** צרו מודל Random Forest עם הפיצ'רים הנראים לכם רלוונטיים. (הסתכלו על הנתונים ובחרו בחוכמה). השתמשו במודל בעד 2000 עצים

(1) **9.b.** השתמשו במודל בכדי לחזות את משתנה המטרה ב'test.d'. שמרו את התחזיות במשתנה 'resForest'

(1) **9.c.** חשבו את ערך השגיאה.

(3) **9.d**. מצאו פונקציה שמראה את החשיבות של כל אחד מהפיצ'רים שנבחרו ליצירת העצים. הגרף שיוצרת הפונקציה מראה את ההשפעה של הוצאת כל משתנה על הדיוק ומדד הgini.

 # MOVIES #

10. KMEANS

(0) **10.a.** טענו את הנתונים בפורמט המתאים למשתנה Data Frame וקרא למשתנה 'data.m'.

(4) **10.b.** בעזרת הפיצ'רים: 'rating', 'year', 'votes', 'length' הריצו מודל Kmeans באמצעות 6 סנטרואידים.

(3) **10.c.** צרו תצוגה של הקלאסטרים.

(2) **10.d**. הציגו את מיקומי הסנטרואידים. האם התוצאות הגיוניות?