תרגיל 2 – with R Machine Learning

בתרגיל זה תתנסו ביישום אלגוריתמי למידת מכונה תוך שימוש בשפת R לטובת משימת סיווג.

עליכם להשתתף בתחרות Machine Learning from Disaster שבאתר Kaggle.

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

קרוב למחצית מנוסעי הטיטאניק נהרגו במהלך טביעתה של הספינה לאחר שהתנגשה בקרחון. מאחר ובתהליך הפינוי לסירות ההצלה נכנסו שיקולים כמו מין הנוסע וגילו, נראה שניתן להשתמש בנתונים שונים אודות הנוסעים כדי לנבא האם ישרדו. התחרות עוסקת בחיזוי מי מבין נוסעי הטיטאניק שרד את טביעתה.

השתמשו בסט האימון שבאתר, המכיל פרטים שונים אודות הנוסעים, וסווגו כל נוסע כשורד (כן/לא). לרשותכם תרגיל מודרך (HW2.rmd) שיסייע לכם בפתרון השלבים הראשונים של התרגיל. התחילו בפתרון התרגיל המודרך ומשם התקדמו ונסו ללמוד מודלים חכמים יותר בכוחות עצמכם.

מבנה הציון:

* 5% מיקומכם היחסי בטבלת התוצאות.
* 95% שלמות ותקינות התהליך המתואר בדוח המסכם.

**מבנה הדוח המסכם:**

* תעדו את הקוד בו השתמשתם בדוח מסכם הכתוב בפורמטRMD , הממוקם בgithub שלכם וכולל את קטעי הקוד שבהם השתמשתם.
* עליכם להגיש לאתר kaggle לפחות 3 ניסיונות סיווג, ביניהם:
  + לפחות אלגוריתם בסיסי אחד
  + לפחות שיטת אנסמבל אחת תוך שימוש בחבילה caret
* ציינו מהו מיקומכם בטבלת הleaderboard, ציינו שם משתמש באתר kaggle וצרפו צילום מסך של לוח התוצאות המראה את מיקומכם בעת ההגשה.
* לפני כל הרצה כווננו פרמטרים חשובים (למשל k ב KNN, מספר עצים ב random forest)
* על הדוח המסכם לכלול עבור כל הגשה שבוצעה לאתר kaggle:
  + תיאור העיבוד המקדים (במידה ובוצע)
  + תיאור האלגוריתם בו השתמשתם
  + תיאור תהליך כוונון פרמטרים חשובים (כולל תיאור גרפי)
  + הפניה לקובץ שהוגש (שישב גם הוא בgithub)
  + צילום מסך של הדירוג שניתן לכם עבור ההגשה (צלמו את המסך שהתקבל לאחר ההגשה לאתר kaggle)

**דגשים:**

* כל זוג יצור חשבון משתמש ב kaggle כדי להוריד את הנתונים הרלוונטיים ולהשתתף בתחרות. ההסברים המלאים נמצאים באתר.
* תוכלו להיעזר ברעיונות מתוך פורומים המלווים את התחרות, אך יודגש כי אין להעתיק פתרונות מהאינטרנט.
* שימו לב לכך למגבלותיהם של מימושים שונים של אלגוריתמים בR : יתכן שתידרשו להמיר מאפיינים נומינליים למספריים, לטפל בערכים ריקים, להמיר משתנה מטרה מספרי לנומינלי ועוד.
* כמו כן שימו לב לכך שתהליך כוונון הפרמטרים דורש לפצל את סט האימון לסט אימון וסט ולידציה, כדי להעריך תוצאות לפי תוצאות החיזוי על סט הוולידציה. בסט המבחן נשתמש רק כדי לחזות ולהגיש לאתר (אין לנו את תוצאות האמת עבור סט המבחן).

**יש להגיש את התרגיל בזוגות עד 26/11/2017.**

תרגילים שיוגשו באיחור "יקנסו" באופן הבא: 5 נקודות יורדו מציון התרגיל לכל יום איחור.

ההגשה תתבצע בתיקיה המתאימה באתר הקורס באופן הבא:

1. על כל אחד מבני הזוג לבצע הגשה באופן הבא:

2. יש להגיש קובץ ובו קישור למאגר ה Githubשל התרגיל וכן שמות השותפים לתרגיל. שם הקובץ יהיה מהסוג ID1\_ID2.txt (תעודות הזהות של המגישים מופרדות באמצעות קו תחתון).

בהצלחה ובהנאה!