כריית נתונים ב- R // תרגיל בית מספר 4

בתרגיל זה נשתמש במסד הנתונים Accidents.csv.

להלן תיאור המשתנים במסד:

TABLE 11.5	DESCRIPTION OF VARIABLES FOR AUTOMOBILE ACCIDENT EXAMPLE
ALCHL_I	Presence (1) or absence (2) of alcohol
PROFIL_I_R	Profile of the roadway: level (1), other (0)
SUR_COND	Surface condition of the road: dry (1), wet (2), snow/slush (3), ice (4), unknown (9)
VEH_INVL	Number of vehicles involved
MAX_SEV_IR	Presence of injuries/fatalities: no injuries (0), injury (1), fatality (2)

מטרתנו תהיה לחזות את מספר הנפגעים בתאונה.

שאלות:

1. פצלו את המסד ל training (**60%**) ו validation. השתמשו ב seed = 1.
2. (6 נק') התאימו ארבעה מודלי סיווג **בינאריים** (כאשר אחד מהם עץ החלטה ואחד רגרסיה לוגיסטית) לחיזוי קיום נפגעים - **מספר נפגעים (injury or fatality) גדול מאפס, או לאו**. דווחו על המודל הטוב ביותר שמצאת, וביצועיו לפי מדד AUC. הדפיסו לפלט את עץ ההחלטה שהתקבל (ממודל עץ סיווג) ותארו אותו (מה המשמעות של n, y והמספרים על הקשתות)
3. (2 נק') צרו confusion matrix של נתוני החיזוי של המודל הטוב ביותר שהתקבל בסעיף הקודם (ערך cutoff = 0.5). דווחו לפלט את ערכי מדדי ההערכה השונים, Sensitivity, Specificity, Accuracy. הסבירו מה המשמעות של כל אחד מהמדדים בהינתן positive class = 0.
4. (2 נק') שנו את ערך ה cutoff לערך 0.6, דווחו לפלט את ערכי מדדי ההערכה אותם דיווחתם בסעיף הקודם. הסבירו (במונחי confusion matrix) כיצד שינוי ה cutoff השפיע על ערכי המדדים.

<u>אופן הגשה:</u>

- ✓ הגשה דרך אתר למידה
- ✓ הגשה בזוגות או ביחידים (רק אחד מבני הזוג צריך להגיש. על הקובץ יופיעו השמות של שני המגישים)
- ✓ יש להגיש קובץ R **מתועד**, וקובץ **פלט** המכיל את התשובות לשורות המסומנות בצהוב