ע"י דוד פרנס, אריאל נובומינסקי ואילי ששון
ת.ז. 337977045, 32228974, 322995358

הגדרת המשימה: בהינתן תחנה ויום כלשהו, בנינו מודל אשר מנבא כמה גשם ירד בתחנה זו ביום זה.

חילקנו את סט הנתונים שלנו באקראי בחלוקה של 70% אימון 30% ולידציה. על חלוקה זאת, בדקנו את כל המודלים שלנו (לכן הצלחה או אי הצלחת מודל אינו תלוי בחלוקה המקרי.) גם נציין שלמרות שבמחברת רואים רק חלוקה אחת, התוצאות היו דומות לחלוקות רנדומליות שונות.

כיון שלא גילינו קשר ליניארי (גם לא מוכלל), לא בדקנו מודל של רגרסיה ליניארית (פשוטה או מוכללת.) אלא מצאנו קשרים בין תכוניות שביחד מרמזים על המשקעים שירדו באותו יום. לכן, לקחנו מודל של עץ החלטה אשר מסוגל למצוא קשרים לא ליניאריים. בדקנו גם RandomForest וגם GBT ומצאנו כי GBT מניב תוצאות יותר טובות. לכן, בחרנו במודל זה.

נציין כי הpredictions יצאו בטווח קרוב לממוצע של כל תחנה. זה צפוי כי בפרויקט חתרנו למזער RMSE. לפי הגדרה, $RMSE = \left(\frac{1}{n}\Sigma(real - predicted)^2\right)^{0.5}$.

ידוע כי ממוצע ממזער מרחק אוקלידי שלפי הגדרה הוא:

$$euclidean\ distance = (\Sigma(real - predicted)^2)^{0.5}$$

קל לראות שמזעור RMSE גורר שאיפה לממוצע.

לכן, הRMSE הגבוה הגיוני. כי RMSE יכוון את המודל לממוצע אבל הראנו בשלב הניתוח שכמות משקעים ביום בעלת שונות רבה. (לפחות במדינות שלנו).