הגשה פרוייקט למידת מכונה:

שם המשתתפים: דביר סעדון

תיאור של המאגר: המאגר מתאר מידע לגבי יין אדום מסוג "Vinho Verde"

המאגר מכיל 12 פיצ'רים: הם מכילים בינהם כאלה שמתארים כמות של סוגי חומצה שונים שנמצאים ביין כמו גם כמו גם סוכר ומלח. בנוסף, כמות הגופרית הדו חמצנית ביין, ריכוז האלכוהול, צפיפות היין ו-PH והמשתנה שארצה לחזות – האיכות של היין מ0-10.

השאלות שעליהם עניתי: עניתי על השאלה של חיזוי (classification) של הדירוג איכות של היין (0-10). ניסיתי גם למצוא את המשוואה לregression של האיכות. ניסיתי לענות על אילו הם הפיצרים הכי טובים כדי לענות על השאלה הראשונה.

בנוסף, ניסיתי לבדוק אם אוכל לייצר תוצאות טובות יותר אם אבצע קלסיפיקציה בין קבוצות של תוצאות (1-3) – יין רע, טוב וטוב מאוד.

הטכניקות בהם השתמשתי: Random Forest, KNN, Adaboost וגם Decision tree.

בכדי לענות על השאלה השנייה, השתמשתי בLogistic regression. בנוסף, ניסיתי להריץ PCA ולראום אם אני אוכל לשפר את המודל כך.

אתגרים: האתגר הכי גדול שנתקלתי בו היה לשפר את המודל מ accuracy של כ55% לכ-88%.

בנוסף, היה לי אתגר לענות על אחת השאלות שתכננתי ("אנסה גם למצוא את המשוואה לregression של האיכות") ולקח לי זמן רב להבין איך אני מוצא את המשוואה.

טכניקות שלא עבדו/ עבדו פחות טוב: טכניקה אחת שעבדה פחות טוב היא Adaboost . המודלים האחרים שהרצתי נתנו תוצאה טובה הרבה יותר ממנה. אני חושב שהסיבה הגדולה ביותר לזה היא שהמודל עושה overfitting ולכן נותן תוצאות פחות טובות.

בנוסף, PCA לא הצליח לשפר את המודל. אני חושב שהסיבה לזה היא שהוצאתי את הפיצ'רים הפחות חשובים וכל האלה שנשארו הם חשובים מספיק למודל כך שניתן לראות שכשהורדנו עוד פיצ'רים המודל נהיה פחות טוב.

טכניקות שהצליחו: המודל הכי טוב שהורץ הוא Random Forest Classifier (~88%) אך גם KNN וגם Decision tree נתנו תוצאות יחסית טובות בקלסיפיקציה. המודל שנתן את התוצאה ברגרסיה הוא Logistic regression (Score = 0.63)