**עבודה שנייה – סדנא למדען המידע**

**תהליך העבודה:**

**עיבוד מקדים:**

בתחילה בחנו את הפיצ'רים עצמם ואת הנתונים על מנת להחליט כיצד לפעול.

בתחילה, בחרנו לא לוותר על הרשומות הריקות אך כן לוותר על שולשה פיצ'רים – Name, Cabin ו-Ticket.

לאחר הסרת שלושת הפיצ'רים, נותרו רשומות ריקות תחת הפיצ'רים: Age ו-Embarked.

פיצ'ר Age – בחרנו להזין את הגיל הממוצע ברשומות עם המידע החסר.

פיצ'ר Embarked – בחרנו להזין את הערך בעל מספר המופעים הרבים ביותר, ברשומות הריקות.

כעת, עבדנו לפי ההנחה שכדי להרחיב את כמות הפיצ'רים ועל כן חילקנו את הפיצ'רים SibSp,Parch ,Pclass ,Sex , Embarked לפיצ'רים בינאריים – כל ערך הפך להיות עמודה והעמודות הוזנו ע"י 0 ו-1 (לדוגמא, אם מדובר בגבר, הרשומה תקבל 1 תחת העמודה של גבר ו-0 תחת העמודה של אישה).

לאחר מכן הרצנו בדיקות עם מספר מודלים שגרמו לנו להבין שאת שני הפיצ'רים (SibSp ו - Pclass) כדי לנו להשאיר כעמודה יחידה בפני עצמה.

את כלל הפיצ'רים הפכנו ל-Factor.

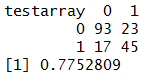
\* נציין כי בקוד השארנו את החלק בו בחנו את כלל השיטות המכונה אך החלק של ה-TEST הוא זה שבו ייצרנו את הקבצים בעבור התחרות ב-Kaggle.

**שיטות המכונה**:

מימשנו חמישה אלגוריתמים:

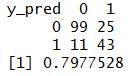
1. **Decision tree** – במבחן התוצאה, הוא זה שסיפק לנו ראשון את התוצאה הטובה ביותר ב-Kaggle (0.7799). (בהמשך עקפנו אותה עם האנסמבל).

**ייצוג גרפי של התוצאות**:



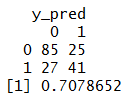
1. **SVM** – לבסוף, לאחר כלל הניסיונות שביצענו וציינו לעיל, גם הוא סיפק תוצאה זהה לעץ ההחלטה.

**ייצוג גרפי של התוצאות**:



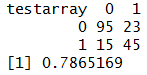
1. **KNN** – לאחר מימוש השיטה, ניסינו לבצע כיוונונים עם מספר ה-K אך לבסוף מצאנו ש-3 הוא המספר המיטבי. למרות זאת, הוא לא סיפק ציון טוב יותר ממה שהשיגו עץ ההחלטה וה-SVM.

**ייצוג גרפי של התוצאות**:



1. **Logistic regression** – מימוש השיטה לא סיפק תוצאה טובה יותר מזו שהשיגו עץ ההחלטה וה-SVM.

**ייצוג גרפי של התוצאות**:



1. **Ensemble (Caret)** – השיטה הזו דרשה מאיתנו כיוונונים רבים אך לבסוף הצלחנו להריץ אותה. בשל שילוב שני אלגוריתמים (GLM,RPART) השיטה שיפרה את התוצאות שלנו (ניתן לראות בעמוד הבא).

**[למודל זה לא ביצענו הערכה מקדימה בשלב ה-Train]**

**GitHub**:

https://github.com/Liadna/ex2---Titanic-Solve-Ready

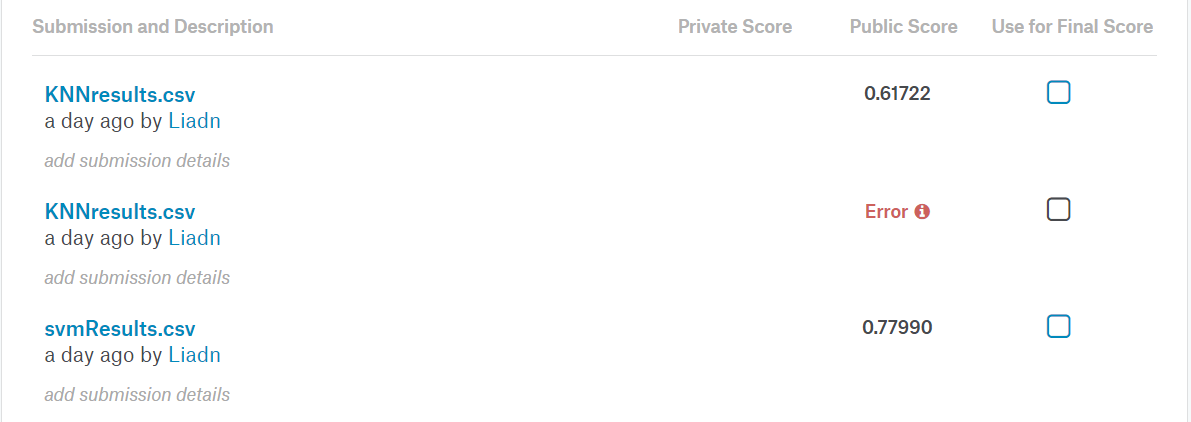
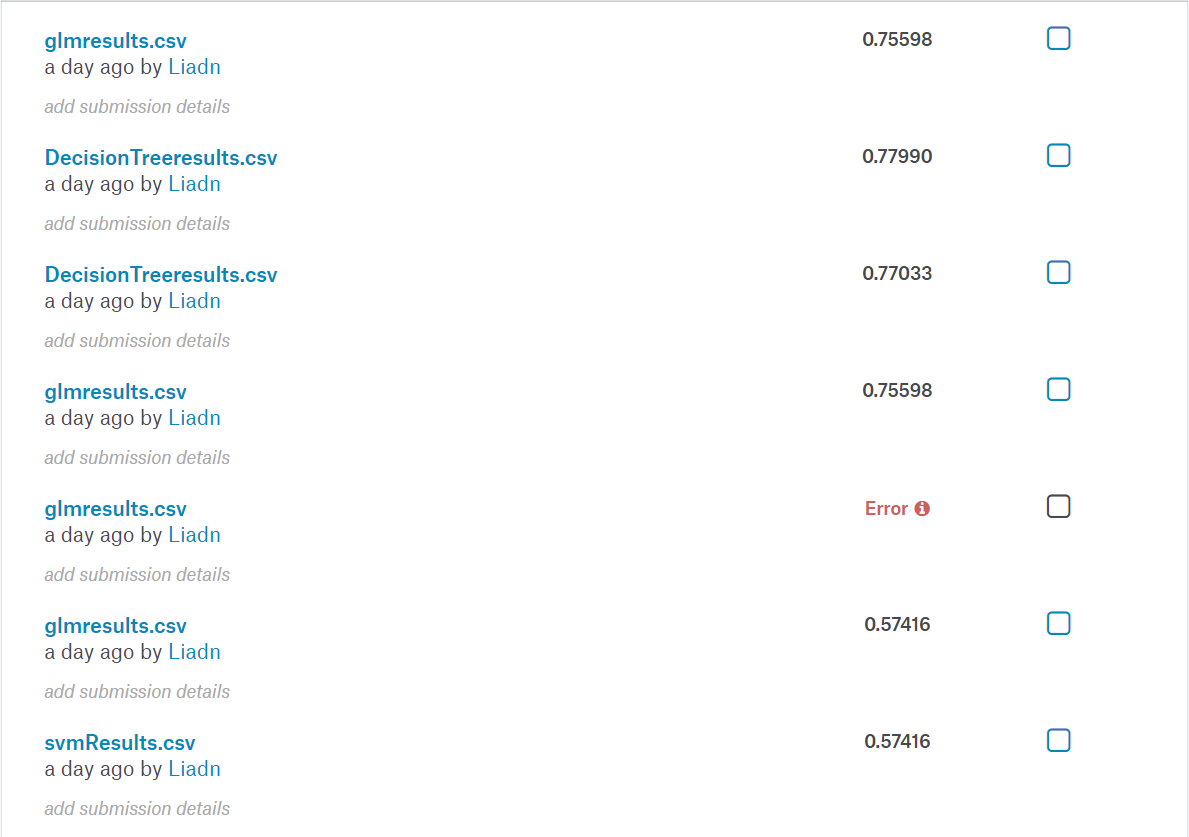
**הגשות ל-kaggle:**

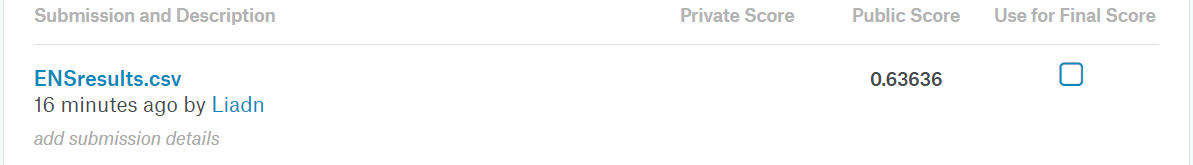
**המקום הטוב ביותר אליו הגענו: 2736**

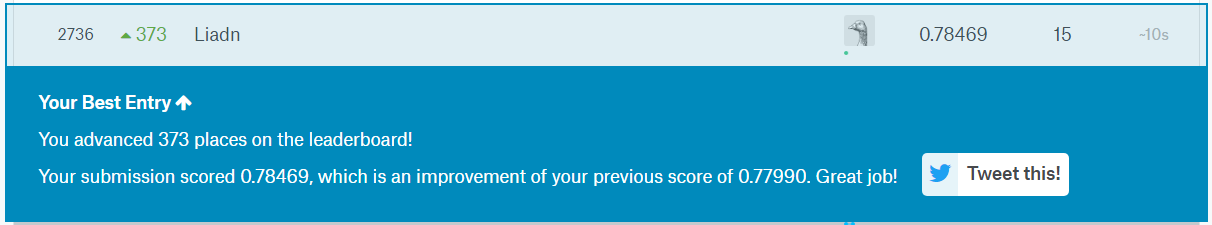
**שיטה: אנסמבל של GLM ו-RPART.**

**דיוק: 0.78469**

**18.04.17:**



**19.04.17:**

**23.04.17:**