**תרגיל 2 – Machine Learning With R**

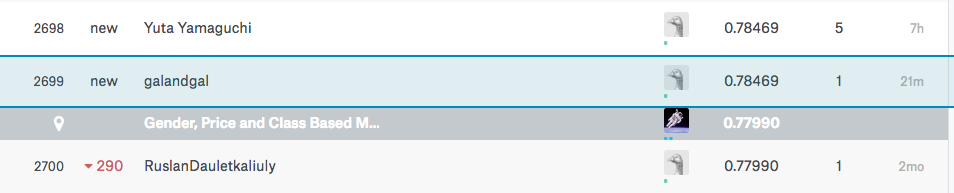
דו״ח מסכם עבור ניסיונות החיזוי

גל בר 200462133

גל שטימברג 201253572

**שם משתמש ב-Kaggle:** galandgal

**דירוג נוכחי ב-Kaggle:** 2699



**שלב 1 - חיפוש משתנים משמעותיים:**

נרצה קודם לעבוד על הדאטה כדי לחלץ את המשתנים החשובים ביותר. כלומר, את אותם הפרמטרים אשר השפעתם על יכולת החיזוי היא המשמעותית ביותר. נבחר את הפיצ׳רים האלה כדי שנוכל להפעיל את האלגוריתמים השונים על הפיצ׳רים הטובים ביותר.

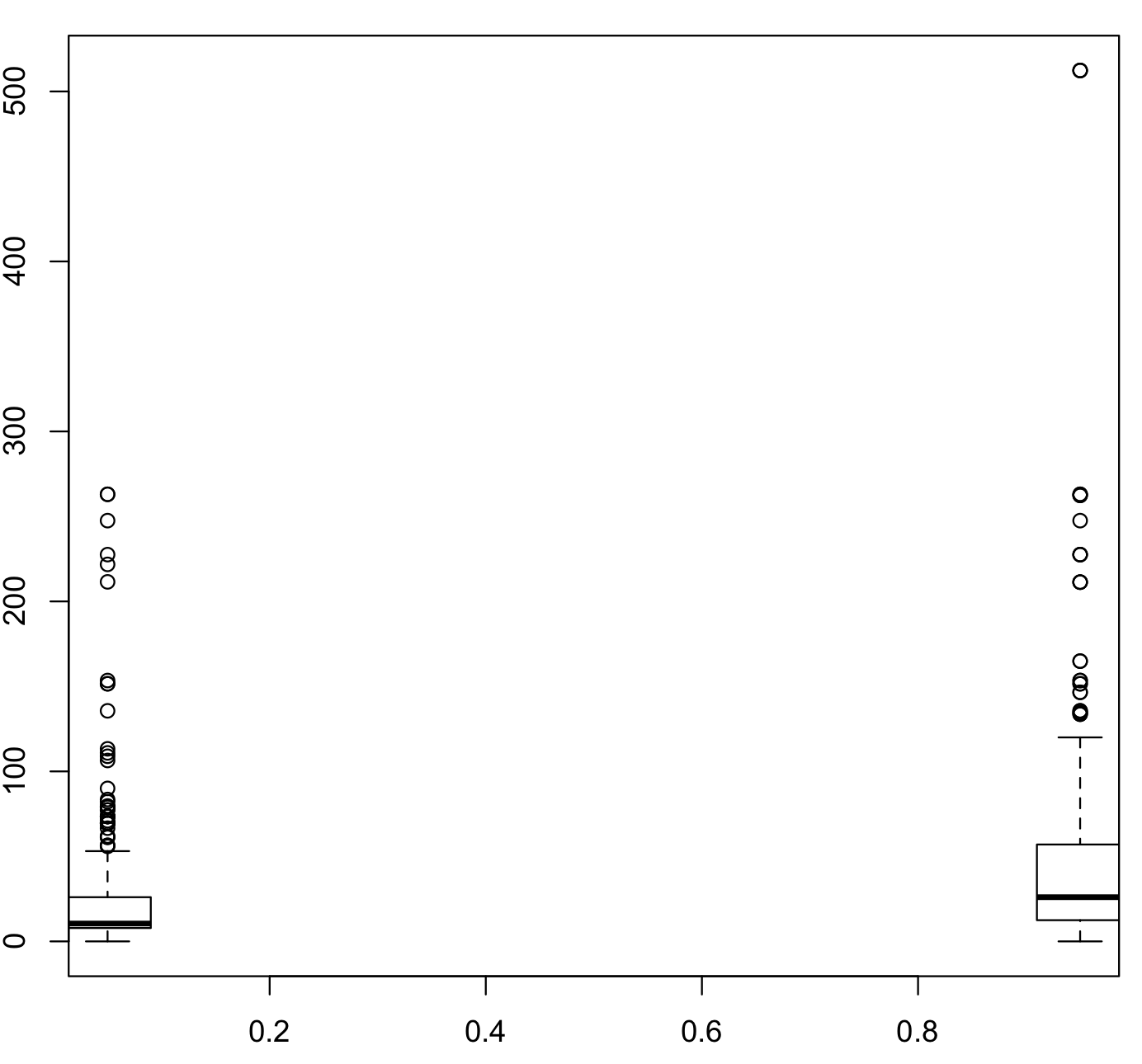
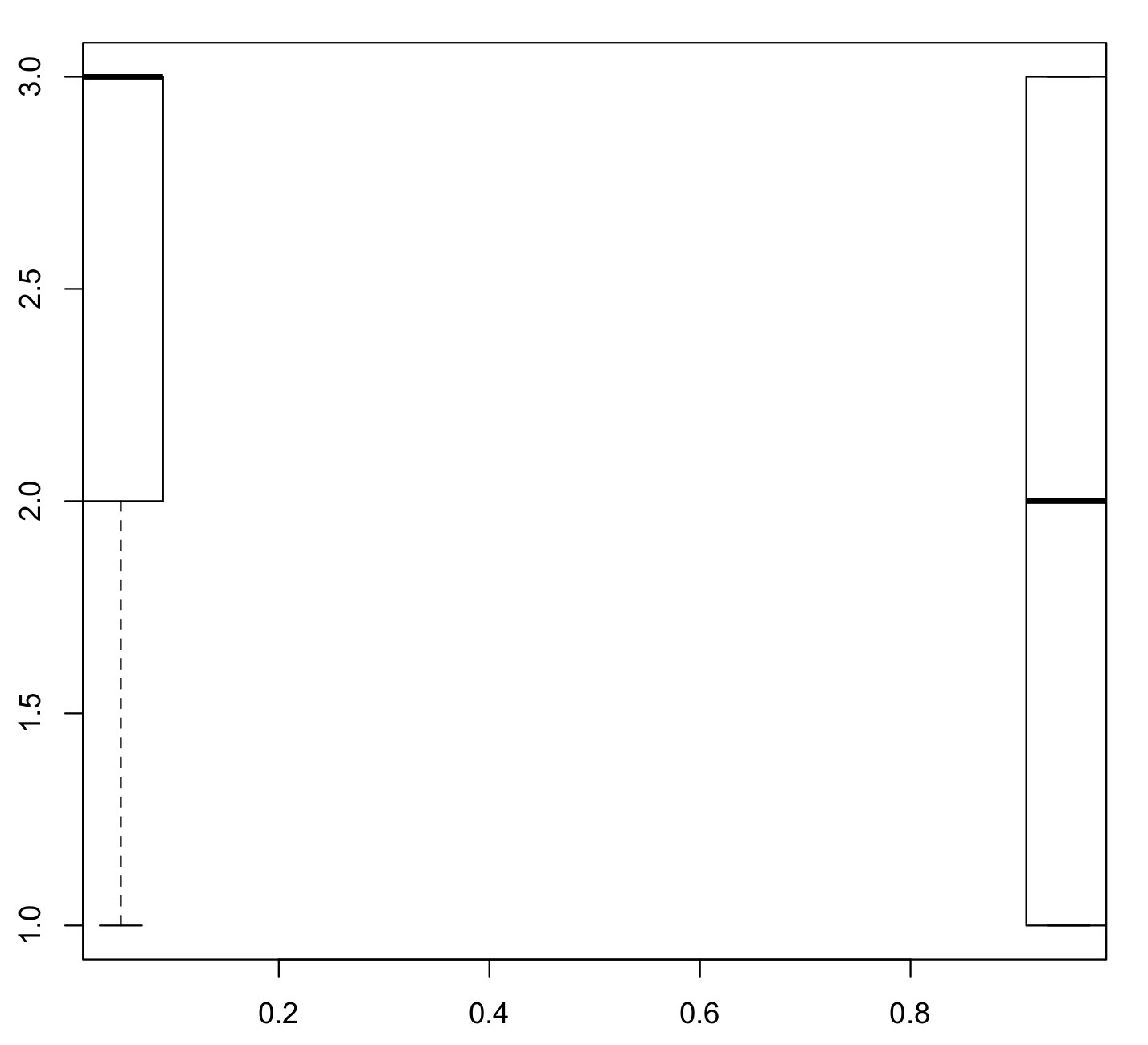
התחלנו לעשות זאת בדף עזר שניתן לנו, על-ידי יצירת גרפים עבור הפרמטרים השונים. לטובת בדיקות ההמשך רצינו לבחון עוד כל מיני משתנים ולכן הוצאנו עוד גרפים.

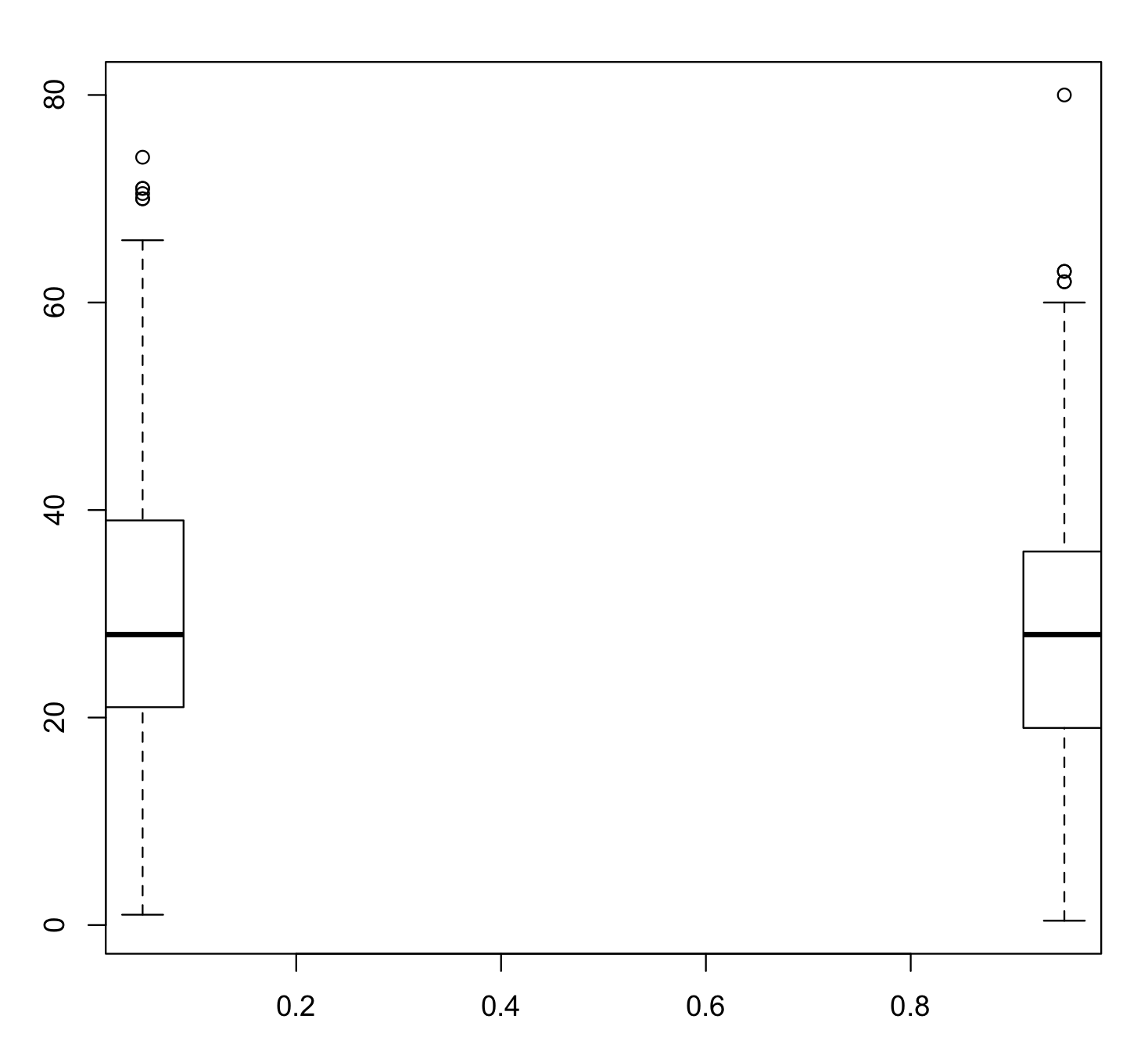
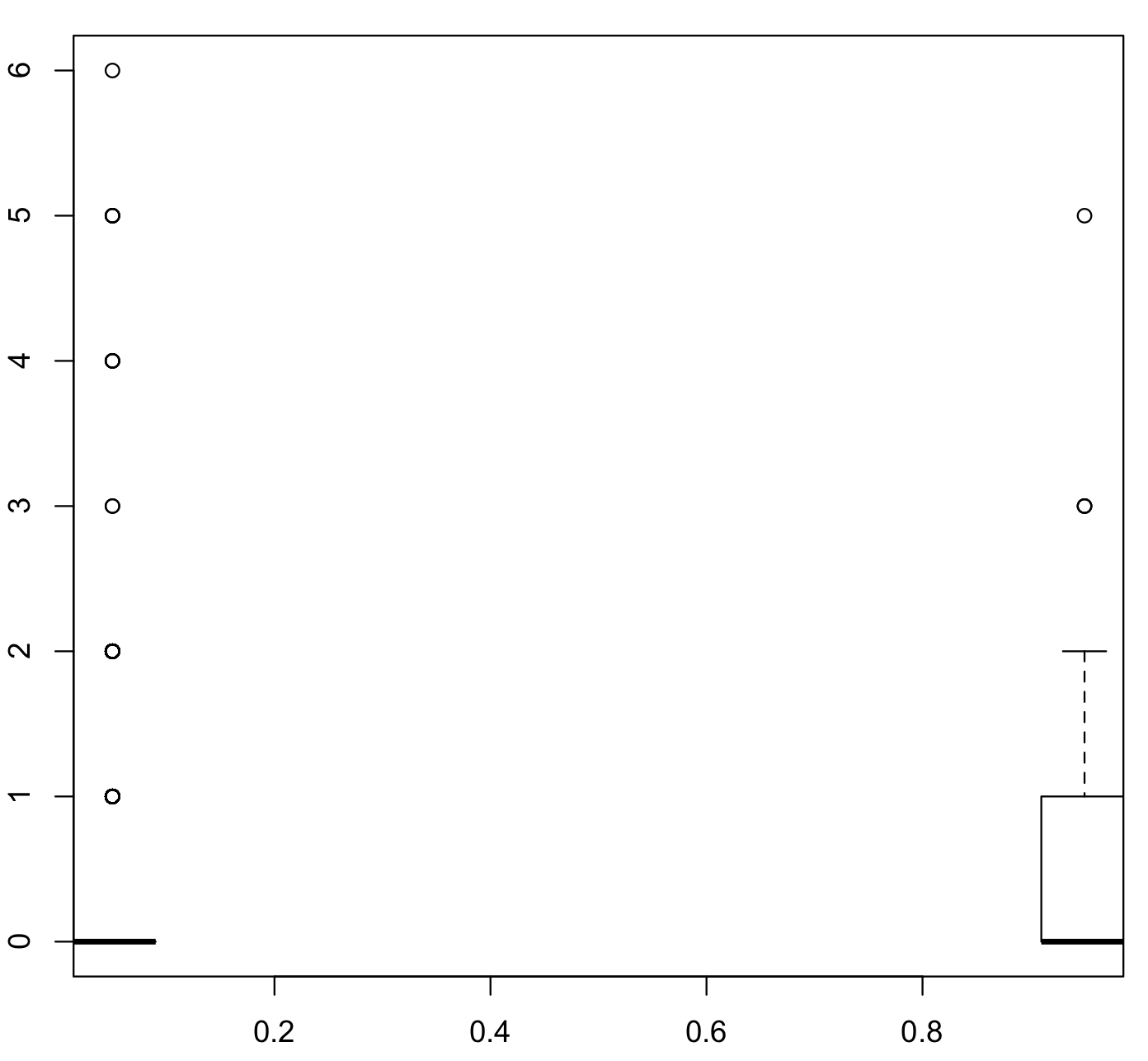
ביצענו את בדיקת הפיצ׳רים על –ידי בדיקה ידנית של המשתנים והערכים שלהן מהקובץ train.csv. נשתמש בספריה fields בשביל לצייר גרפים שיוכלו לתת לנו קצת יותר מידע על הפיצ׳רים הטובים יותר והטובים פחות.

כמו כן השתמשנו במידע שהוצג בגרפים במסמך הקדמה לתרגיל שהועלה באתר מודל.

*ניתן לראות שאין הבדל גדול למדי בין גילאי השורדים/ניצולים בפני עצמו*

להלן חלק הגרפים שהתקבלו עבור חלק מהמשתנים שבדקנו:





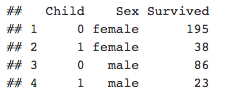
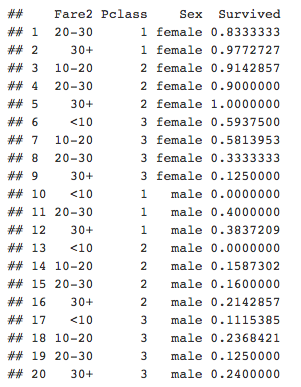
Parch-Survived

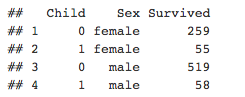
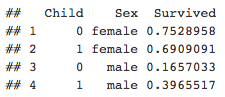
Pclass-Survived

Fare-Survived

Age-Survived

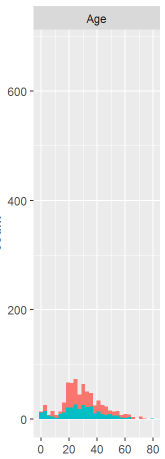
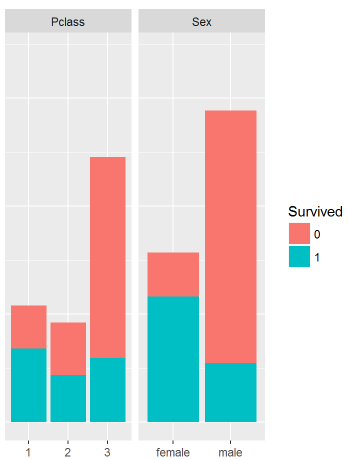
1. ניסיון הגשה ראשון (אלגוריתם בסיסי ללא שימוש בחבילת caret):
   * בניסיון הגשה זה המשכנו לחקור את המידע משלב החקר המקדים.
   * במהלך החקירה (המפורטת בהמשך) גילינו את המסקנות הבאות:
     1. נשים שורדת באופן עקרוני יותר מגברים
     2. נשים שהיו בclass גבוהה ושילמו מחיר יקר לכרטיס בדרך כלל לא שרדו
     3. לילדים סיכוי נמוך יותר להינצל
   * לפי מסקנות אלו כתבנו את התוצאות לניסוי הראשוני והפשטני.





**ניתן להבחין כי ככל שהמחלקה יותר גבוהה ושילמו יותר עבור כרטיס, הסיכוי להינצל יורד, והסיכוי לנשים תמיד גבוהה מהגברים**

**ניתן להבחין כי הסיכוי להינצל הוא מרבי עבור נשים ואז ילדות ואז גברים ואז ילדים**



test $Survived <- 0

test $Child[test $Age<18] <- 1

test $Survived[test $Sex == 'female'] <- 1

test $Survived[test $Sex == 'female' & test $Pclass == 3 & test $Fare >= 20] <- 0

בהתחלה הגדרנו שאף אחד לא ניצל.

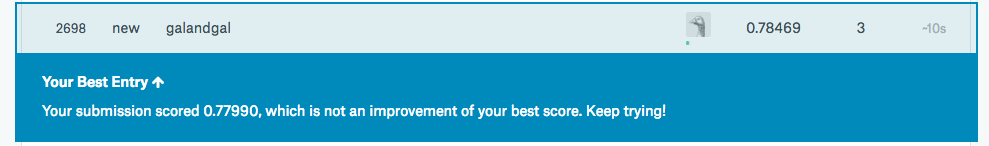
לאחר החקר שעשינו הגדרנו פעם אחת שכל הילדים ניצלים ופעם שניה הגדנו שבכל הנשים ניצלים מלבד הנשים שיושבות במחלקה 3 והכרטיס שקנו היה יקר מ-20.

לאחר מכן כתבנו את התוצאות לקובץ csv אותו העלנו לאתר.

קישור לתוצאות ב-github:

<https://github.com/galst/ex1/blob/master/zeroTry.csv>

צילום מסך מאתר Kaggle:



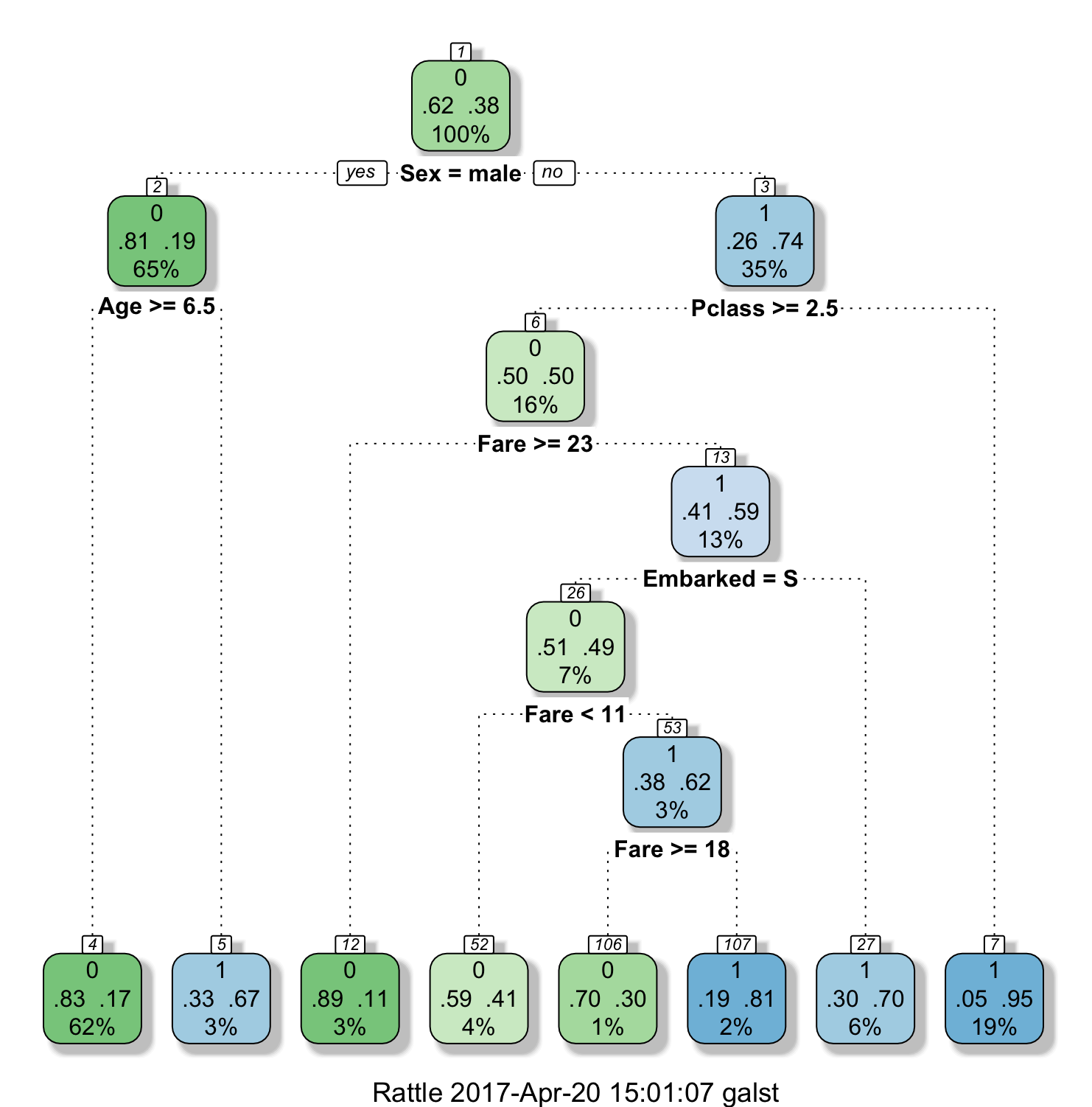
1. ניסיון הגשה שני:

* השתמשנו באלגוריתם CART שנמצא בחבילת rpart שמובנית בתוך R.
* לפי העיבוד המקדים שעשינו, גילינו את הפרמטרים שהם המשמעותיים ביותר עבור חיזוי התוצאות ואימנו מודל בהתאם למשתנים הנ״ל – Pclass, Sex, Age, Parch, Fare, Embarked
* תיאור האלגוריתם:

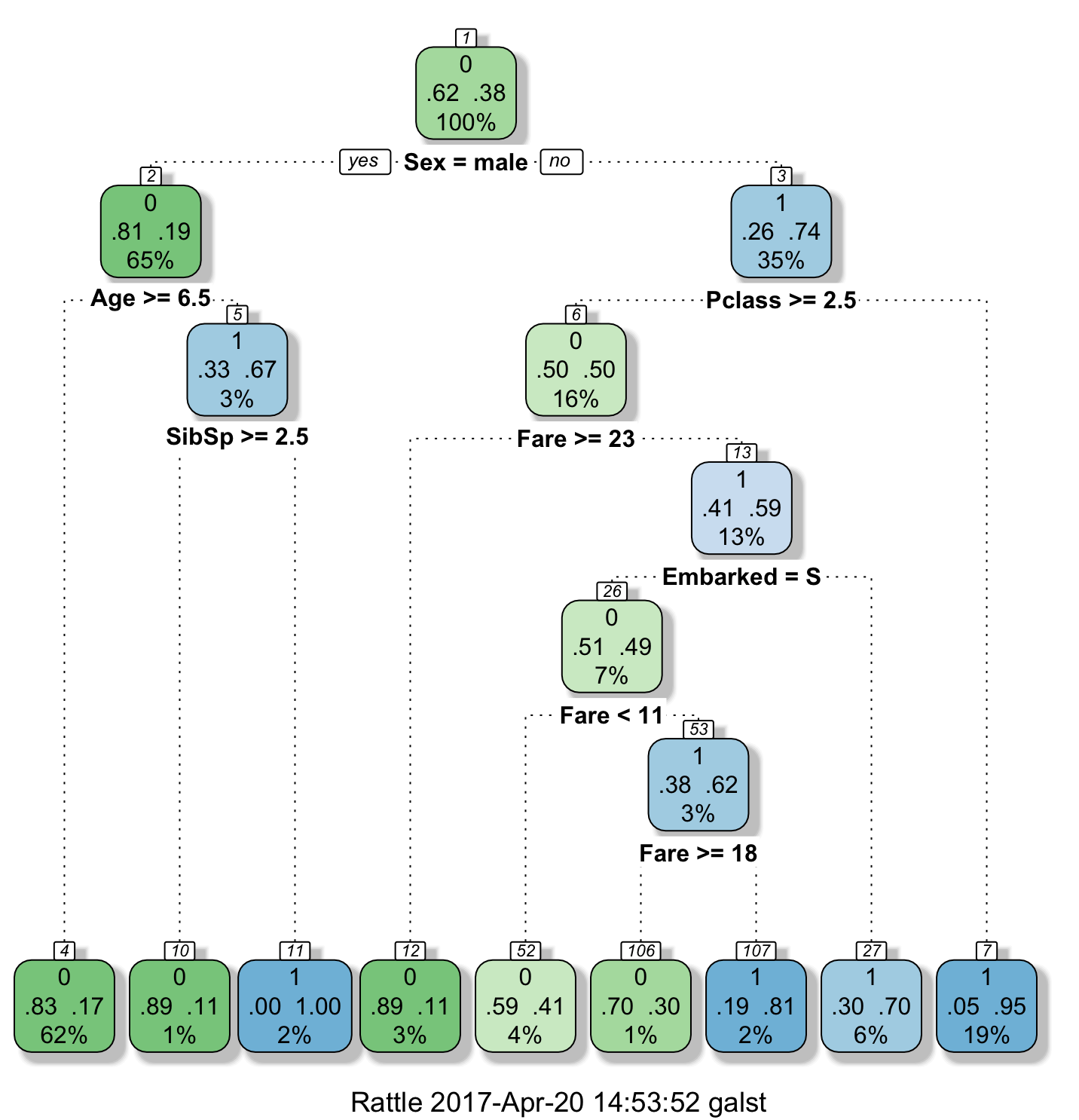
עץ החלטה CART הוא [עץ בינארי](https://he.wikipedia.org/wiki/%D7%A2%D7%A5_%D7%91%D7%99%D7%A0%D7%90%D7%A8%D7%99) מלא המורכב מצמתי החלטה שבכל אחד מהם נבדק תנאי מסוים על מאפיין מסוים של התצפיות ועלים המכילים את הערך החזוי עבור התצפית המתאימה למסלול שמוביל אליהם בעץ. סוגים של עצי החלטה הם עצי רגרסיה שבהם מותאם ערך רציף לכל תצפית ועצי סיווג שבהם מותאם ערך בדיד או מחלקת סוג לכל תצפית.

עצי החלטה מסוג CARTמשלבים את שני סוגי החיזוי.

* השתמשנו בספריה חיצונית שמצאנו באינטרנט בשביל לצייר את הגרף הבא:



מהתרשים ניתן לראות את האופן בו עץ ההחלטה פועל. ניתן לראות כי התכונה העיקרית לפיה העץ החלטה נבנה הוא לפי מין הנוסע. ניתן לראות שלנשים יש סיכוי גבוה יותר להינצל לפי שחזינו. כמו כן, שוב לפי מה שחזינו במבוא, הגיל וה-class של הנוסע משפיע על סיכוייו להינצל. לאחר מכן, ניסינו לשחק עם משתנים נוספים שחשבנו שיוכלו לשפר את הביצועים. מצאנו שאם מוסיפים לעץ ההחלטה את המשתנה SibSp נקבל עץ החלטה טוב יותר והמודל שבנינו הי טוב יותר. במקרה הזה הגרף היה נראה באופן הבא:

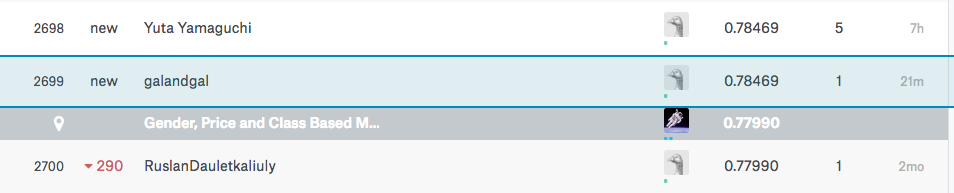


ניתן להבחין כי נוסעים שה-SibSp שלהם גדול מ-2.5 סיכויים להינצל קטן יותר וזהו גורם שהשפיע על קבלת ההחלטה בעץ.

קישור לתוצאות ב-github:

<https://github.com/galst/ex1/blob/master/firstTry.csv>

צילום מסך מאתר Kaggle:



1. ניסיון הגשה שלישי:
   * השתמשנו באלגוריתם Random Forest שהוא חלק מחבילת Caret.
   * חשוב להדגיש כי היינו צריכים להמיר את השדה Survived ל-factor כדי שהאלגוריתם יבנה מודל קלסיפיקציה ולא מודל רגרסיה.

* לפי העיבוד המקדים שעשינו, גילינו את הפרמטרים שהם המשמעותיים ביותר עבור חיזוי התוצאות ואימנו מודל בהתאם למשתנים הנ״ל – Pclass, Sex, Age, Parch, Fare, Embarked
* תיאור האלגוריתם:

אלגוריתם אשר משתמש במספר עצי החלטה, המוסבר מעלה, על מנת לשפר את דירוג הסיווג.

את הפרמטרים השונים כווננו גם לפי ניסוי וטעיה וגם לפי קרירה באינטרנט. לדוגמה, את פרמטר מספר העצים בחרנו על-יד מספר ניסויים ובחינת התוצאה:

for (ntree in c(1000, 1500, 2000, 2500)) {

set.seed(seed)

fit <- train(Survived~Pclass + Sex + SibSp + Embarked + Parch + Fare, data=train,method="rf",ntree=ntree)

key <- toString(ntree)

modellist[[key]] <- fit

}

בבדיקה הזאת יצרנו רשימה של ארבעה מודלים. כל מודל מורכב ממספר עצים שונה ביער. עשינו זאת ע״י יצירת מפה של ארבע ערכים שונים של ntree ושימוש בלולאה על מנת לעבור על כל אחד מהערכים הנ״ל וליצור עבורו מודל.

התוצאה הטובה ביותר התקבלה כאשר מספר העצים היה 2000.

קישור לתוצאות ב-github:

<https://github.com/galst/ex1/blob/master/secondTry.csv>

צילום מסך מאתר Kaggle:

