پاسخ به سوال اول پروژه یادگیری ماشین ۶۱۰۷۹۹۰۰۱ سید امیررضا موسوی مجد

توضیح کد نوشته شده:

ابتدا داده هایی که در مساله داده شده است را در سیستم لود میکنیم. همچنین دقت میکنیم که یک typo در فایل normal.counts.voom وجود دارد که آنرا رفع میکنیم(به کامنت کد رجوع کنید). در ادامه دیتافریمی تحت عنوان feature_vec میسازیم که ادغام شده از فایل های metadata و normal.counts.voom میباشد. در ادامه کار سه نوع feature selection انجام میدهیم. در حالت اول ۱۰ داده بیان را به صورت رندوم انتخاب میکنیم و SD_Rand نامگذاری اش میکنیم. در حالت دوم از principle component analysis بهره میبریم. به این صورت که بر کل دیتا counts_data متد pca را اعمال میکنیم و pc1-pc10 را انتخاب میکنیم. از این ۱۰ بردار یکه، ۱۰ تا ژنی که بیشترین قدر مطلق ضریب را در شکل گیری هر بردار یکه داشتند انتخاب کردم. بنابراین ۱۰۰ عدد ژن حاصل میشود که بیشترین اهمیت را دارند. اما برخی ژن ها duplicate هستند و در چند pc هستند. فلذا با اعمال تابع unique اضافات حذف میشوند و به ۶۳ ژن مهم پس از اعمال principle component analysis میرسیم. که به آن نام SD_PCA اطلاق میکنیم.

در روش سوم از Recursive Feature Elimination استفاده میکنیم. در ابتدا از تمامی ژنهای counts_data آنهایی که correlation بالا ۷۰ درصد داشتند را حذف میکنیم. سپس ۷۷۵۲ ژن باقی میمانند(از ۱۷۳۹۷). این تعداد

ژن را به یک مدل ژنریک Random Forest میدهیم تا بوسیله آن Recursive Feature Elimination انجام شود. مدل فوق پس از اینکه ساخته شد، Feature ها را براساس اهمیت مرتب میکند که ما ۱۰ تا ژن مهمتر را انتخاب میکنیم و SD_rf_rfe نامگذاری میکنیم.

این ۳ نوع فیچر معرفی شده اساس کار ما را برای train کردن ماشین های classification تشکیل میدهند.

بنابراین هم روش کاهش ابعاد و هم روش انتخاب ویژگی خواسته شده در صورت سوال انجام شده است.

لازم به ذکر است در برخی مواقع به علت crash نکردن rstudio و جواب دادن RAM کامپیوتر، داده های اضافی و بلااستفاده حذف شده اند.

در اینجا سه نوع متد هوش مصنوعی برای classification را پیاده سازی میکنم. KNN-SVM-Random Forest

با آن ۳ نوع فیچر و سه نوع متد هوش مصنوعی سرجمع به ۱۳ نوع ماشین کلاسیفیکشن میرسیم:

1. knn_SD_Rand, 2. knn_SD_PCA, 3. knn_SD_rf_rfe,
4. svm_radial_SD_Rand, 5. svm_linear_SD_Rand, 6. svm_radial_SD_PCA,
7. svm_linear_SD_PCA, 8. svm_radial_SD_rf_rfe, 9.svm_linear_SD_rf_rfe,
10. rf_SD_Rand, 11. rf_SD_PCA, 12.rf_SD_rf_rfe , 13. rf_rfe

لازم به یاداوری است که rf_rfe همان رندوم فورست اولیه ای بود که از ۷۷۵۲ ژن ساخته شده بود و موجب رسیدن ما به مجموعه ژن SD_rf_rfe شده است.

در ادامه عملکرد این ۱۳ ماشین بررسی شد که نتیجه را در فایل Explanation_A_Excel1.xlsx مشاهده میکنید.
در نهایت بهترین ماشین های ما، اینها هستند:

knn_SD_rf_rfe: Accuracy: 0.805; Kappa: 0.707

svm_radial_SD_rf_rfe: Accuracy: 0.822; Kappa:0.731

svm_linear_SD_rf_rfe: Accuracy: 0.807; Kappa:0.709

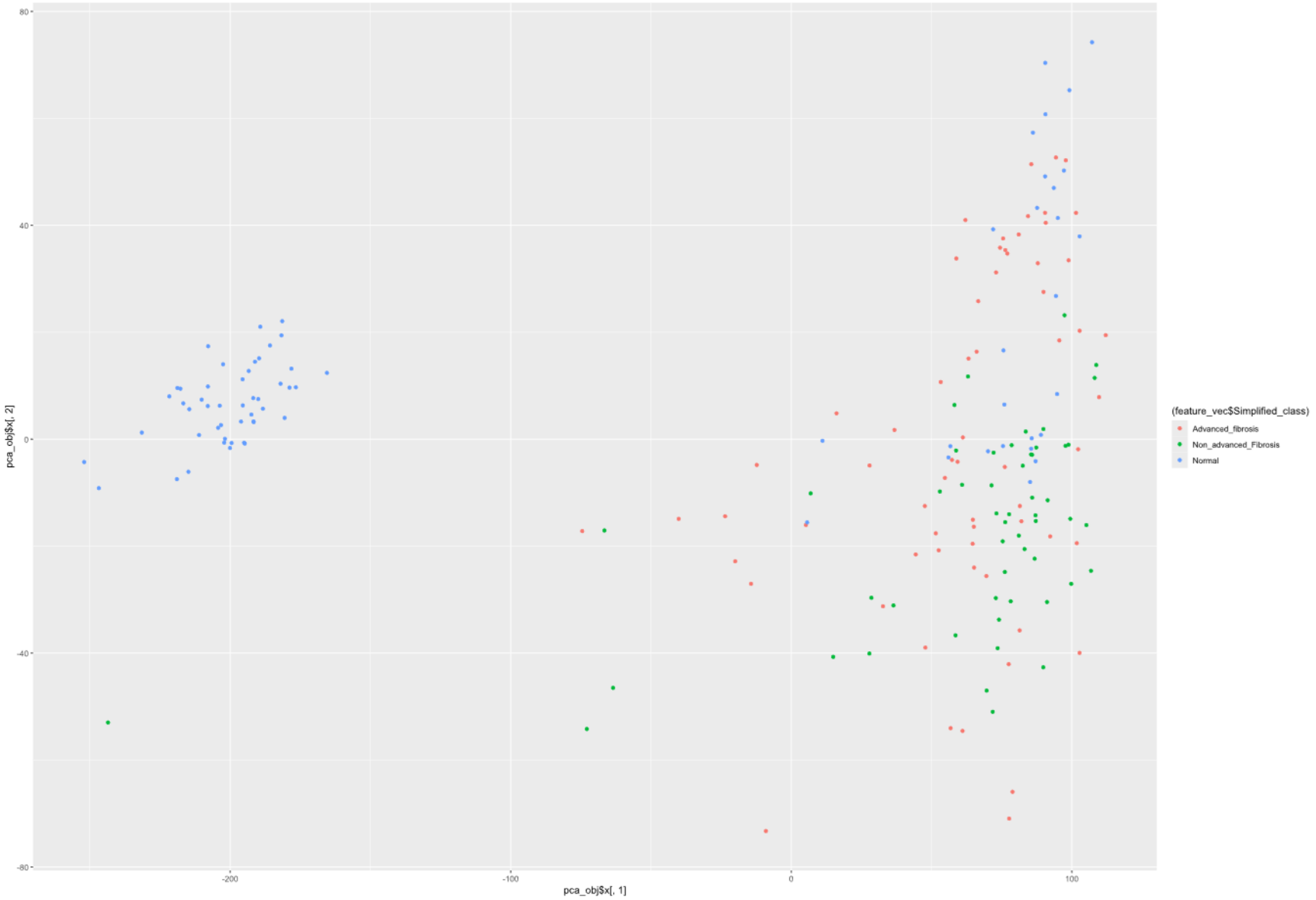
rf_SD_PCA: Accuracy: 0.802; Kappa: 0.700

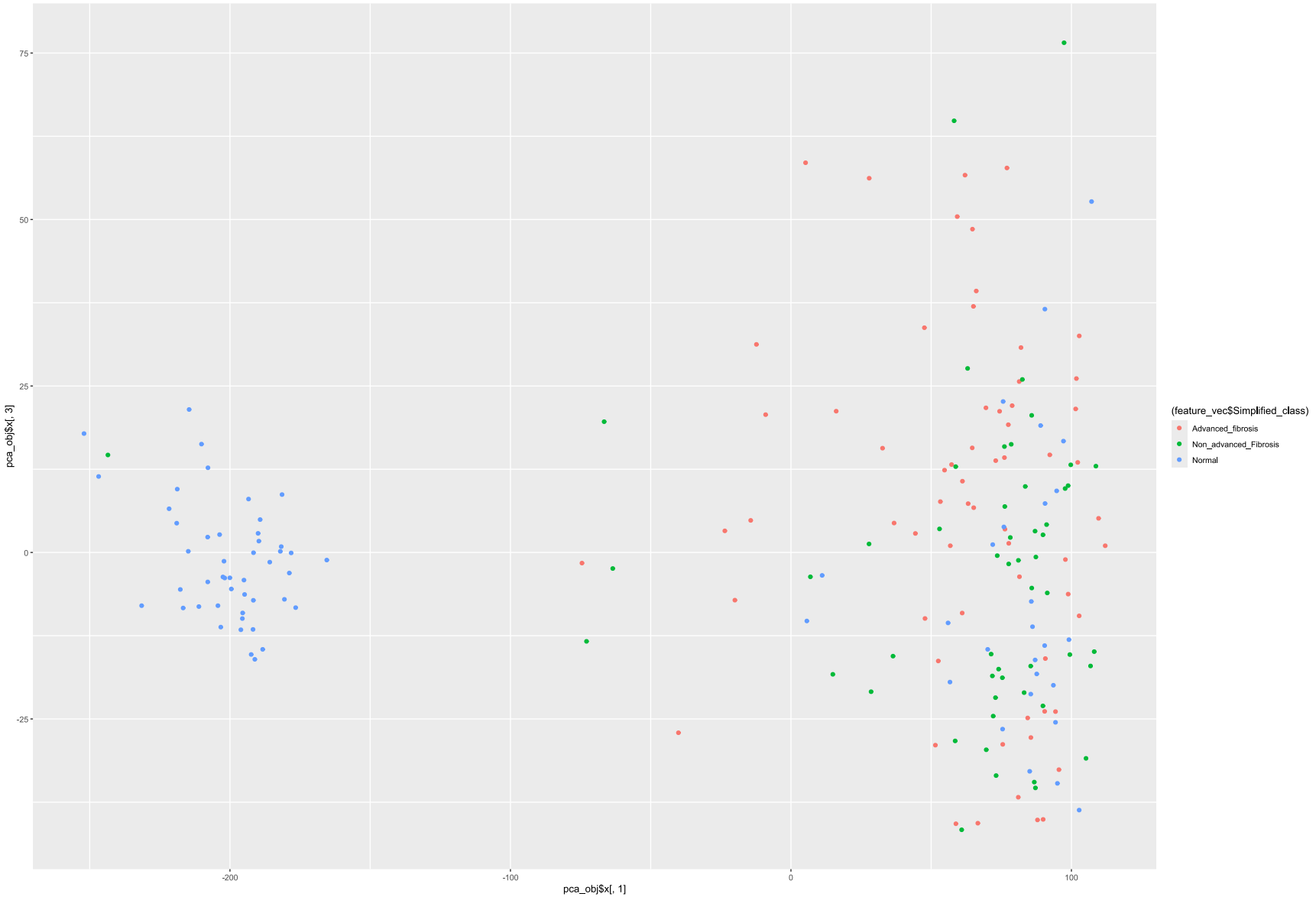
rf_SD_rf_rfe: Accuracy: 0.837; Kappa: 0.754

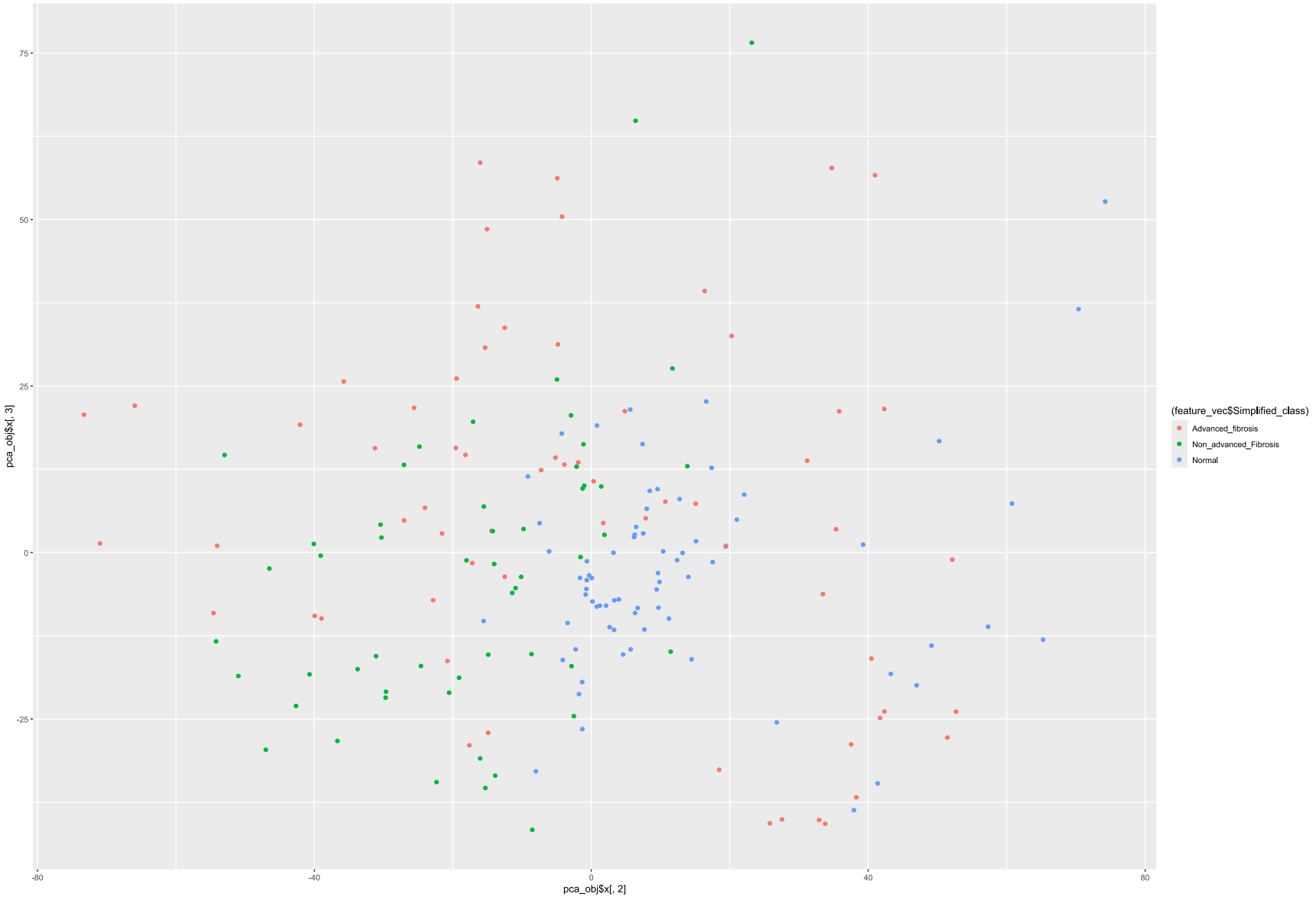
نتیجه میگیریم که بهترین مجموعه feature ها، SD_rf_rfe هست و بهترین متد ها svm_radial و random forest هستند که در ادامه کار، برای سوال b از آنها بهره میبریم.

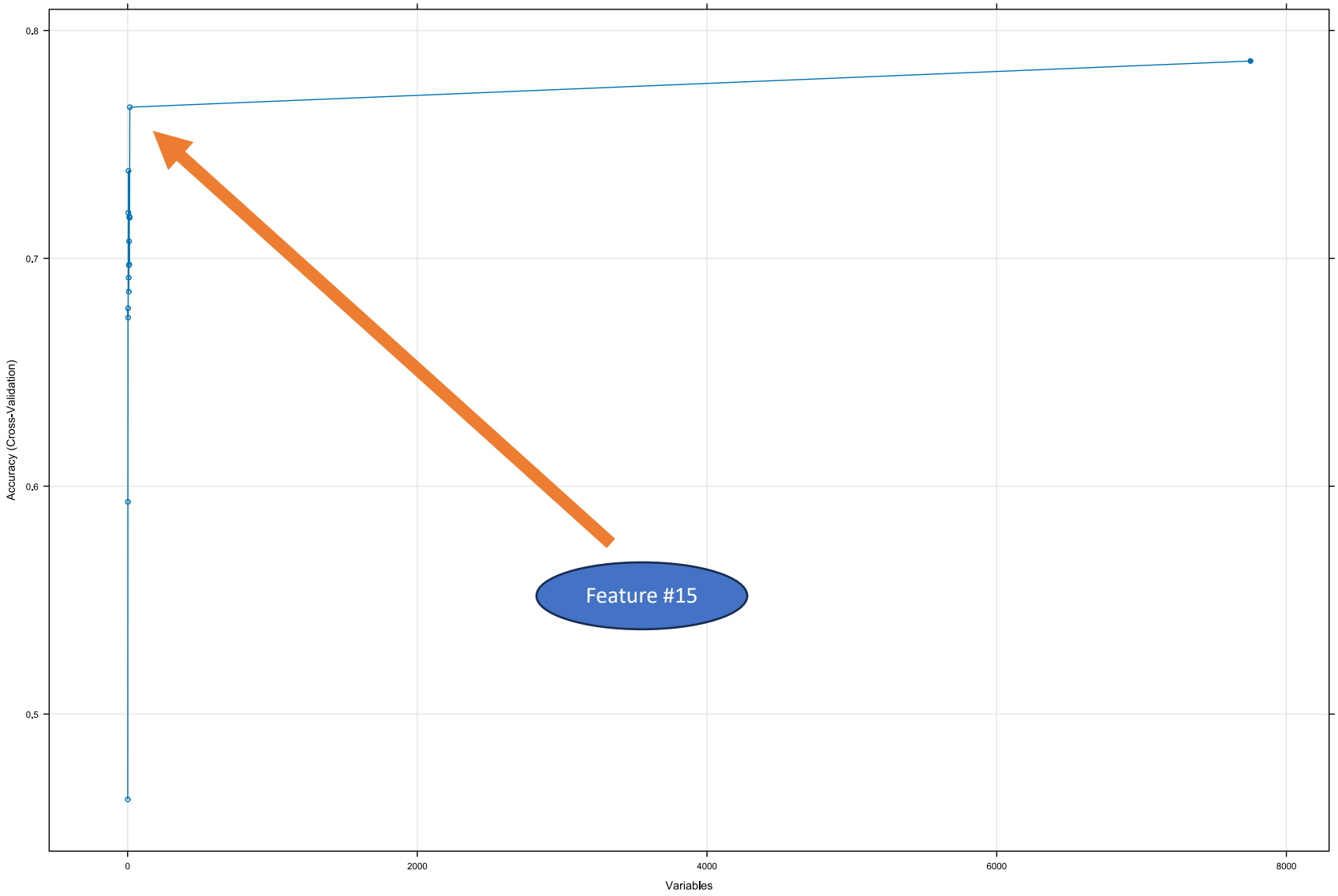
تمامی ماشین ها 10 fold cross validated هستند و preprocess شده اند.به فایل Explanation_A_Excel1 رجوع کنید. در ادامه کار ژن های SD_rf_rfe و SD_PCA (۱۰ تای اول) را در وبگاه NCBI بررسی کردیم تا ببینیم مطابقتی بین این ژنها و بیماری NAFLD مشاهده میشود یا خیر. که نتیجه در فایل Explanation_A_Excel2.xlsx آمده است.

در انتها، سه گراف برای سه بعد اولیه PCA و یک گراف برای Recursive Feature Selection آمده است:


80
40
0
-40
-80
-200
-100
0
100
pca_obj$x[, 2]
pca_obj$x[, 1]
(feature_vec$Simplified_class)
Advanced_fibrosis
Non_advanced_Fibrosis
Normal


75
50
25
0
-25
pca_obj$x[, 3]
-200
-100
0
100
pca_obj$x[, 1]
(feature_vec$Simplified_class)
Advanced_fibrosis
Non_advanced_Fibrosis
Normal


75
50
25
0
-25
pca_obj$x[, 3]
-80
-40
0
40
80
pca_obj$x[, 2]
(feature_vec$Simplified_class)
Advanced_fibrosis
Non_advanced_Fibrosis
Normal


0.8
0.7
0.6
0.5
Accuracy (Cross-Validation)
0
2000
4000
6000
8000
Variables
Feature #15