سید امیررضا موسوی مجد ۶۱۰۷۹۹۰۰۱ حل سوال دوم پروژه یادگیری ماشین

در خصوص سوال دوم مطرح شده:

در قسمت اول سوال دیدیم که مجموعه ژنی تحت عنوان SD_rf_rfe با عملکرد ماشین SVM یا Random Forest عملکرد بهینه ای جهت classification دارد. ما در این قسمت، از همان SD_rf_rfe که لیست آن در فایل Explanation_A_Excel1 هم موجود هست استفاده میکنیم و چهار رکورد از داده meta data را هم به آن اضافه میکنیم و مجددا الگوریتم ها را تکرار میکنیم:

- svm_radial_SD_rf_rfe = svm radial with SD_rf_rfe genes
- svm_radial_SD_rf_rfe_extended = svm radial with SD_rf_rfe genes AND SEX, Age, Diabet, BMI_surge records.
- rf_SD_rf_rfe = random forest with SD_rf_rfe genes
- rf_SD_rf_rfe_extended = random forest with SD_rf_rfe genes AND SEX, Age, Diabet, BMI_surge records

مشاهده میشود که روش های extended حداکثر ۱ درصد بهبود در accuracy دارند (اگر اصلا تاثیر داشته باشند) و در هر چهار ماشین ما، Accuracy بین ۸۰ الی ۸۴ درصد میباشد.

این افزایش ناچیز بهبود دقت در اثر اعمال feature های بیشتر در ماشین های extended هست، (یعنی ۱۰ به ۱۴) و این اثر لزوما ناشی از اهمیت ۴ ویژگی SEX, Age, Diabet, BMI_surge **نیست.**

برای این که باز هم مطمئن بشویم یک ماشین هوش مصنوعی پنجمی را train میکنیم. این ماشین که rf_rfe_extended نام دارد، ۱۰ ژن SD_rf_rfe و ۴ داده متادیتا را به عنوان فیچر لحاظ میکند و یک ماشین مصنوعی Random Forest مطابق با Recursive Feature Elimination میسازد. نتیجه کار این هست که الگوریتم مربوطه feature ها را براساس اهمیت مرتب میکند. و همیشه میبینیم که چهار دیتا SEX, BMI_surge, Age, Diabet در انتها قرار میگیرند و کمترین اهمیت را دارند.

فلذا پاسخ به این سوال منفی هست. در صورت استفاده از داده مجموعه ژنی SD_rf_rfe ۴ رکورد از داده های متادیتا اهمیتی ندارند و دانش جدیدی اضافه نمیکنند. همان مجموعه ژن کافی هست.