

هدف پروژه:

مدلسازی هزینه تعمیرات و نگهداری (متغیر وابسته) با استفاده از سن و میزان کارکرد ماشین آلات (متغیرهای مستقل)

دیتا شامل 4247 رکورد که به 61 گروه (کلاستر) تقسیم می شوند و هدف اینه با استفاده از حلقه For در ابتدا outlier هر کلاستر مشخص و حذف گردد، سپس 0.2 هر کلاستر به عنوان دیتای test کنار گذاشته شود و با 0.8 باقی مانده دیتا هر کلاستر با استفاده از Bootstrap مدل Train داده شود و با استفاده از Best tune بدست آمده از هر مدل صحت برای دیتاهای Test محاسبه شود.

مدلهای مورد نظر:

1-KNN 2-SVM 3-MARS(Multivariate adaptive regression splines) 4-M5 Tree

پکیج یا library مورد استفاده برای مدلهای CARET میباشد که استفاده از مدلهای بسیار راحت کرده است

موارد قابل ذکر:

\* هر چهار مدل ذکر شده بصورت کامل مدل نویسی شده است و با استفاده از bootstrap و حلقه for برای تمامی کلاسترها train مدل انجام می شود و نتیجه با استفاده از besttune نمایش داده می شود (فایلهای knn,tree,svm,mars)

\* همچنین دیتاها به درستی به داده train & test تقسیم می شوند (فایل split main)

\* کدنویسی outlier detection نوشته شده است که البته به نادرست برای همه دیتا زده می شود(outlier without for) و نیاز است برای هر کلاستر جداگانه زده شود (فایل outlier with for) که این فایل ناتمام است.

خواسته اصلی از پروژه:

برای هر یک از چهار مدل یک فایل کدنویسی R یکپارچه داشته باشیم بصورتی که به ازای هر کلاستر در ابتدا outlier detection اجرا شود و داده های پرت حذف شوند، سپس دیتاهای باقیمانده به داده های train&test تقسیم شوند(80-20) و با استفاده از داده های train مدل آموزش ببیند و بعد از مشخص شدن Best tune هر مدل برای هر کلاستر، صحت و دقت مدل با استفاده از داده های test ارزیابی شود.