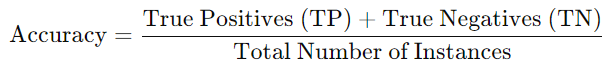
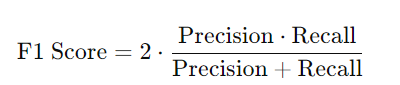
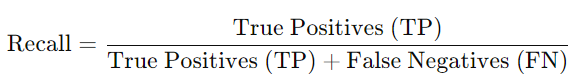
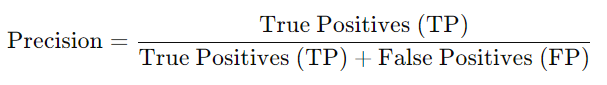
**داده کاوی - علی پاکدامن**

آنالیز و پیشبینی سرطان ریه در دیتاست پزشکی بر اساس شاخص های باینری در پایتون با تمرین و تست الگوریتم های داده کاوی

منبع دیتاست و کد های jupyter notebook:

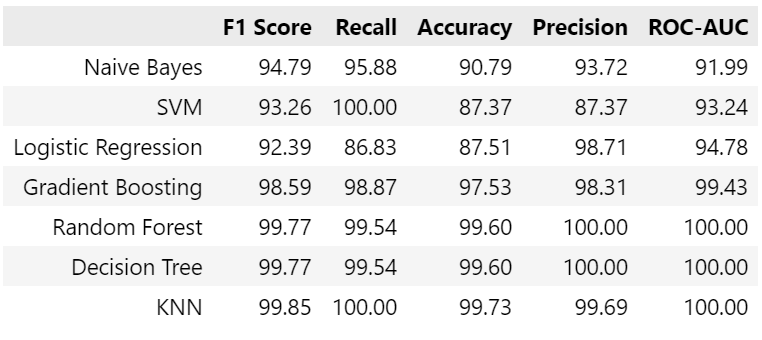
<https://github.com/alipakdamangh/lung_cancer_analysis>

**فرمول معیار های استفاده شده**

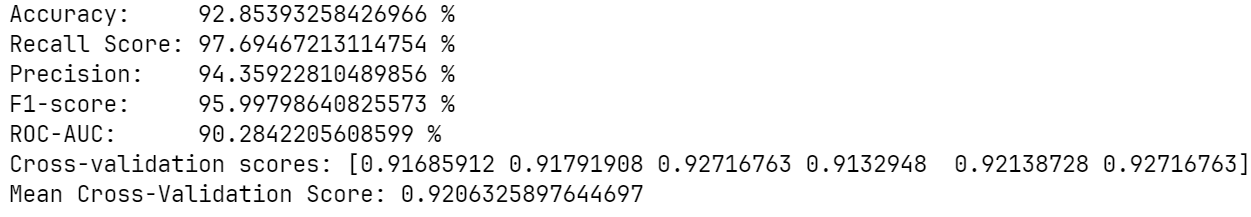


معیار Roc-Auc میزان Recall را با معکوس آن در Threshold های متفاوت مقایسه می‌کند تا با اندازه گیری trade off بین کلاس ها یک مقدار ارائه کند.

**مدل های استفاده شده و معیار ها**

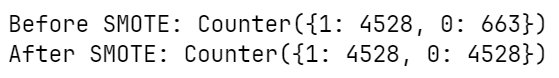
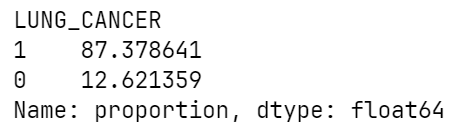


**معیار های مدل Logistic Regression**

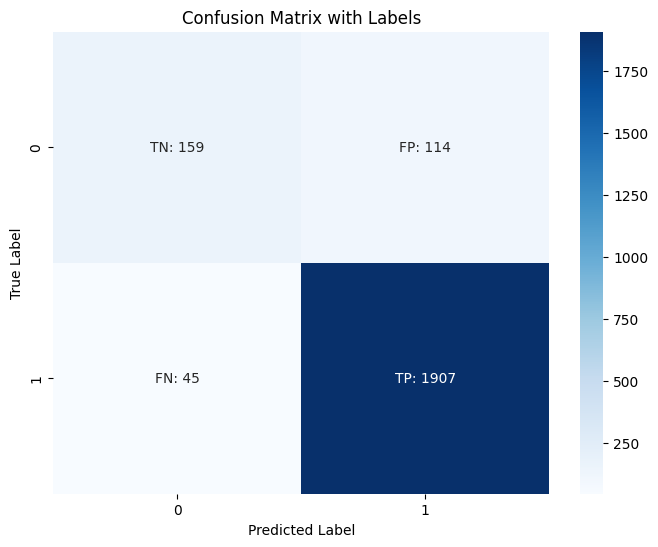
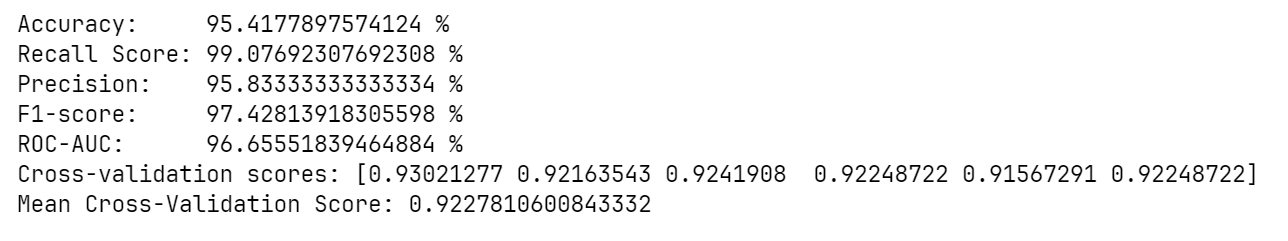
****

**بهبود معیار ها**

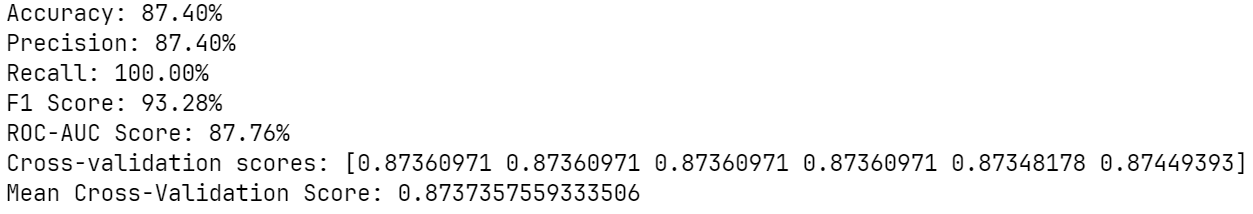
بالانس دیتا را که بیشتر متمرکز روی مقادیر False است با روش SMOTE به هر طرف 50 درصد می‌رسانیم



ستون هایی که تاثیر کمتری بر نتایج دارند را حذف کرده تا مدل با دیتاهای مهمتری آموزش و تست را انجام دهد

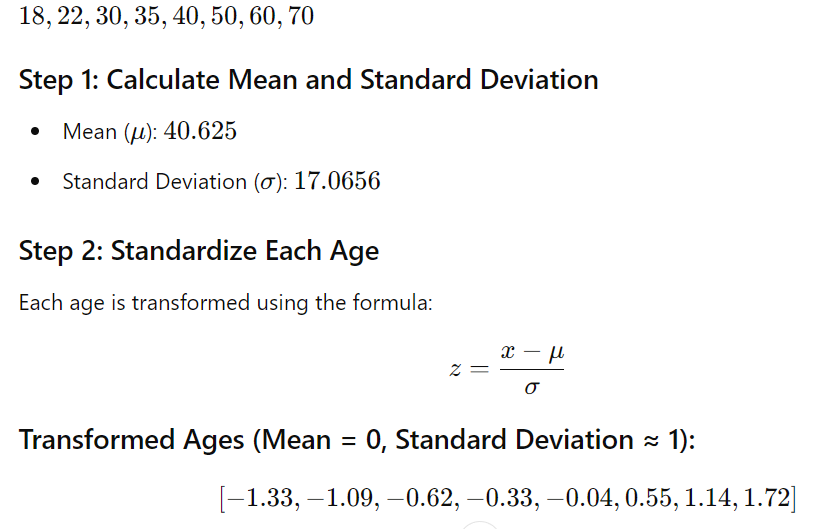
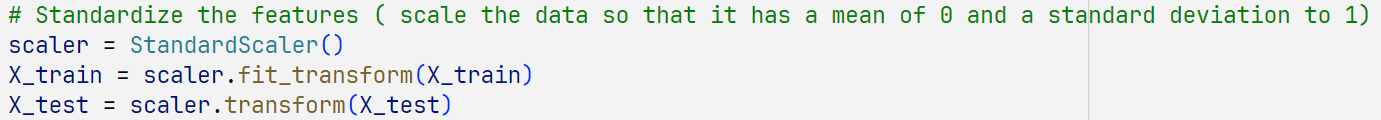
**نتایج بهبود مدل**

**معیار های مدل Support Vector Machine - SVM**

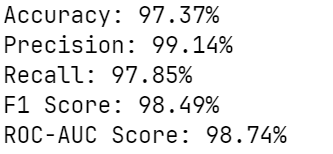
****

**بهبود معیار ها**

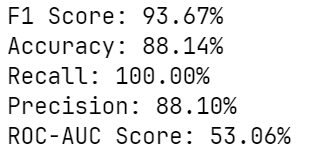
این الگوریتم نیازمند مقادیر باینری true وfalse است که ستون سن جزو این مقادیر نیست و باید با استاندارد سازی آن را به مدل بفهمانیم تا به صورت بهینه از آن در آموزش استفاده شود (فرمول استاندارد سازی در پایین شکل مثال زده شده)

****

**نتایج بهبود مدل**

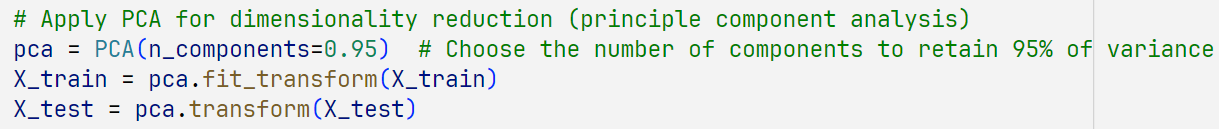


**معیار های مدل Naive Bayes**

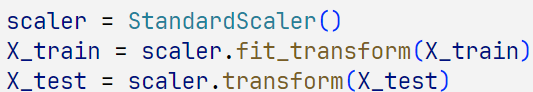
****

**بهبود معیار ها**

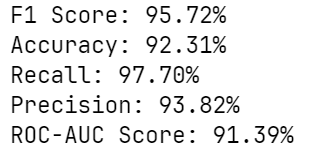
برای کاهش ابعاد و نگه داشتن پارامتر های مهم از PCA استفاده می‌‍‌کنیم (درصد 95 را برای نگه داشتن حداکثر مقدار اصلی انتخاب می‌کنیم تا تفاوت زیادی بین دیتاست پیش فرض نباشد)

****

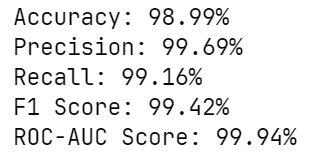
همانند SVM سطون سن را استاندارد سازی می‌کنیم تا آموزش روی دیتای باینری انجام شود

****

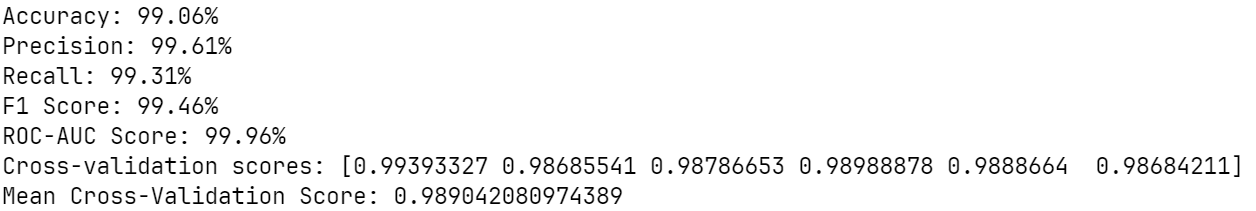
**نتایج بهبود مدل**

****

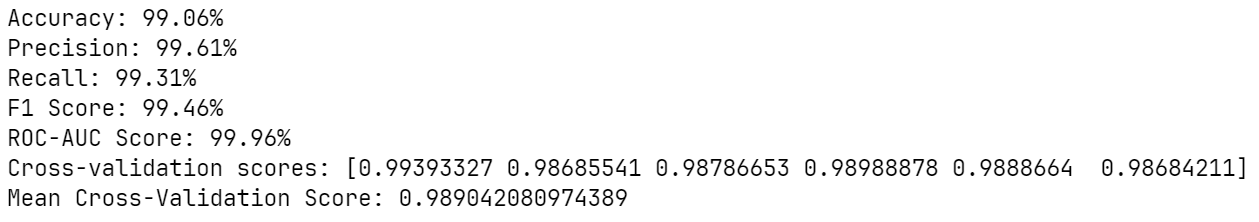
**معیار های مدل KNN - K-Nearest Neighbors**

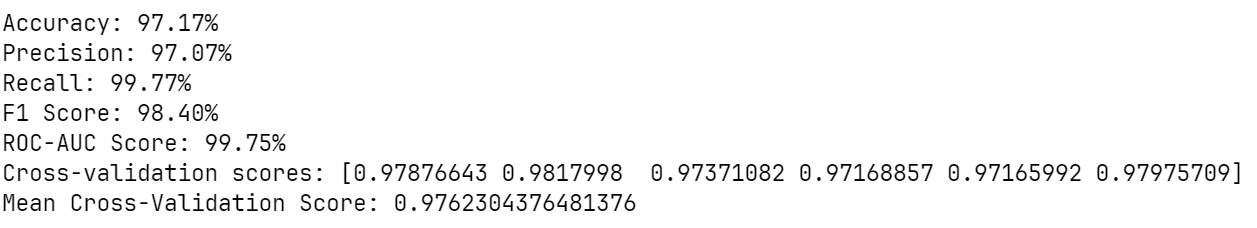
****

**معیار های مدل Decision Tree Classifier**

****

**معیار های مدل Random Forest**

**معیار های مدل Gradient Boosting**

****