فاز 1 :  
1 تقسیم داده به **train/test** با حفظ نسبت کلاس‌ها.(80/20)

2 استانداردسازی داده‌ها با StandardScaler.

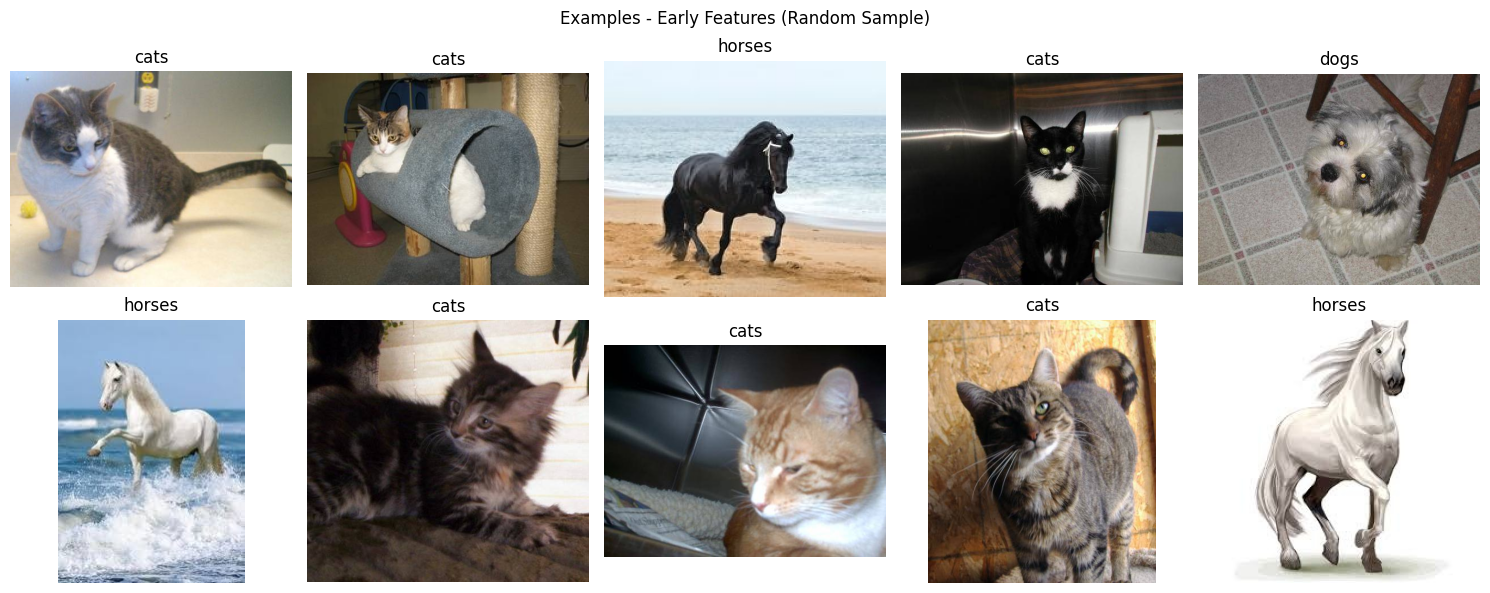
3 کاهش بعد با **PCA** (تا 95٪ واریانس حفظ شود)

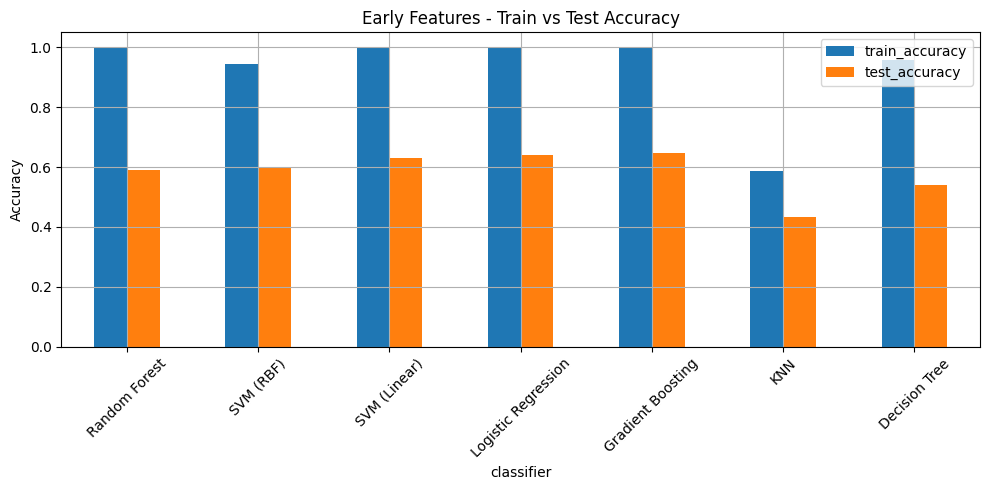
4 آموزش ۷ مدل طبقه‌بند مختلف:

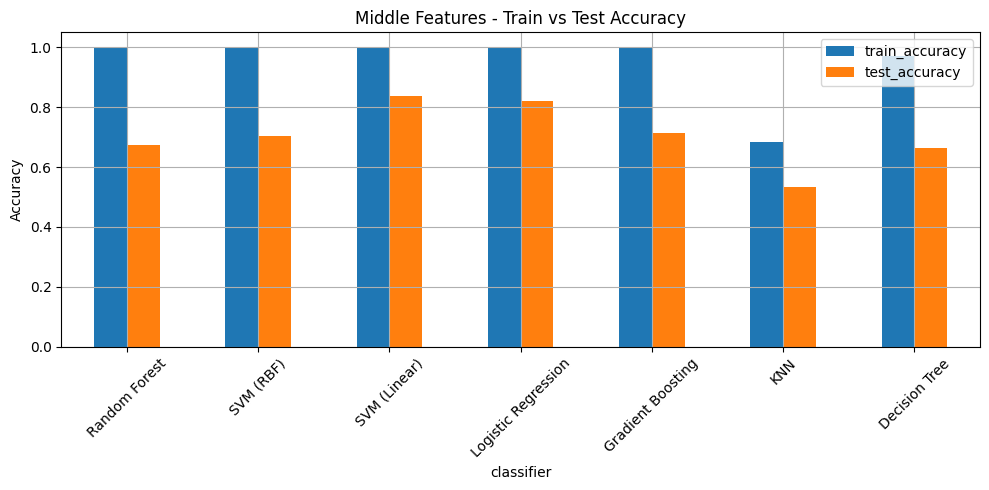
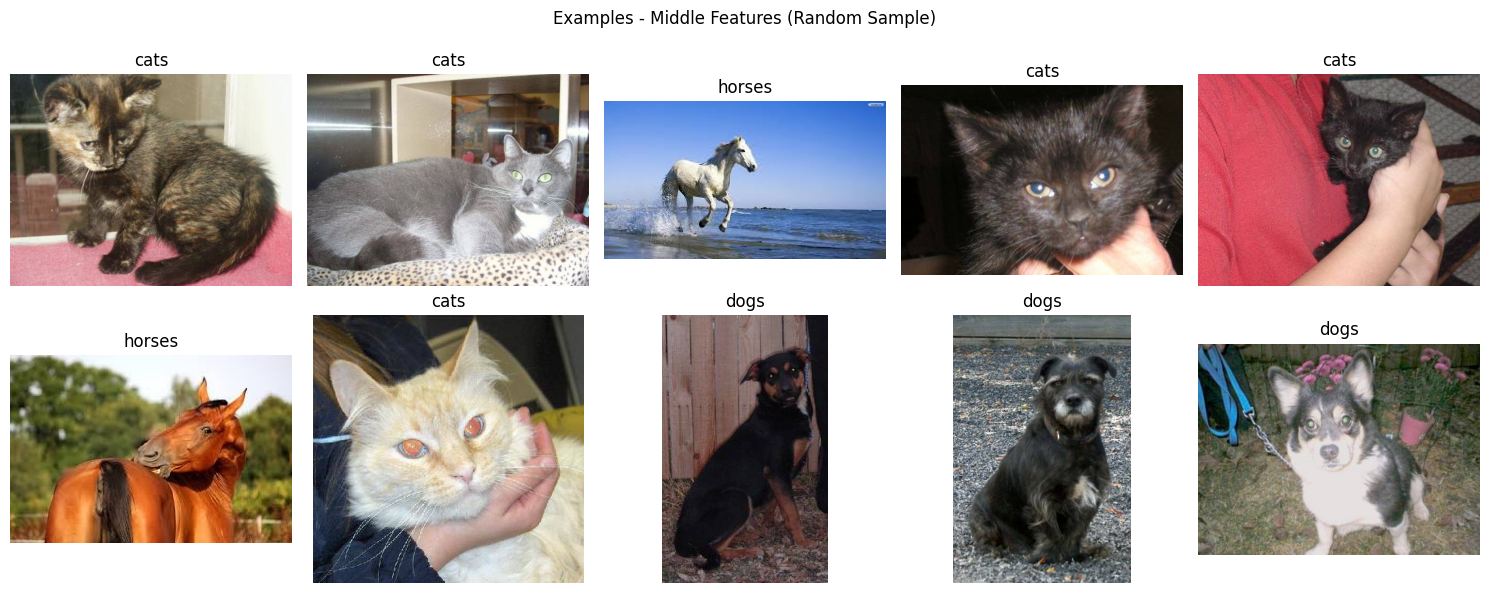
* Random Forest
* SVM (RBF/Linear)
* Logistic Regression
* Gradient Boosting
* KNN
* Decision Tree

5 ارزیابی روی train/test با:

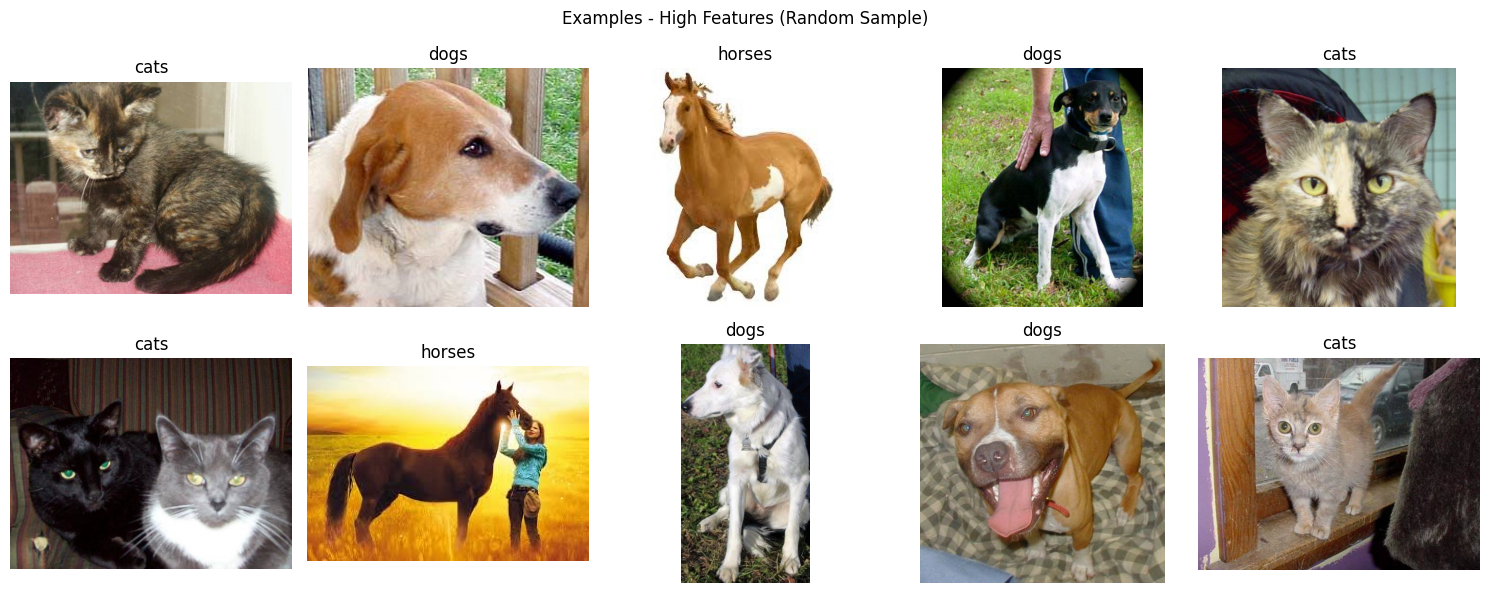
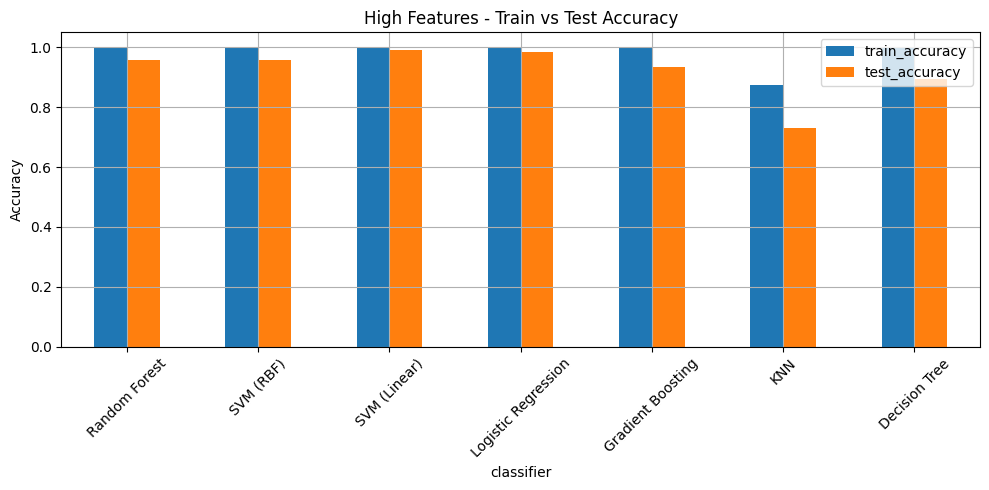
* دقت (accuracy)
* F1 macro
* Recall macro



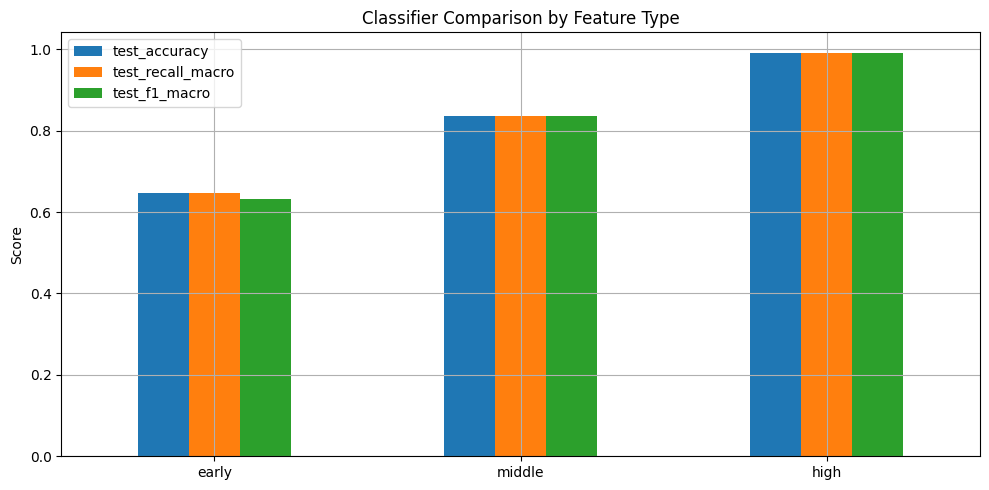
می توان دید که در همه classifier ها دقت پایین دارند و اختلاف train , test بالا شده است. دلیل اصلی این است که ویژگی های استخراج شده شامل ویژگی های ابتدایی می شود و بخش با اهمیت کمتر از شبکه ResNet را شامل می شود پس نمی توان به خوبی طبقه بندی کرد. همپچین تعداد عکس ها کم است و اگر بیشتر بود شایدoverfit، کمتری رخ می داد

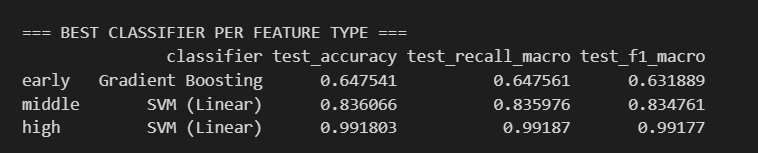


می توان دید که دقت در همه classifier ها بهتر شده است ولی هنوز و اختلاف train , test نسبتا بالا است. دلیل اصلی این است که ویژگی های استخراج شده شامل ویژگی های میانی می شود و بخش با همیت بیشتری از شبکه ResNet را شامل می شود پس می شود بهتر طبقه بندی کرد. همپچین تعداد عکس ها کم است و اگر بیشتر بود شایدoverfit، کتری رخ می داد



می توان دید که دقت در همه classifier ها بهتر شده است نزدیک به هم عمل می کنند . دلیل اصلی این است که ویژگی های استخراج شده شامل ویژگی های بالا می شود و بخش با اهمیت بیشتری از شبکه ResNet را شامل می شود پس می شود بهتر طبقه بندی کرد.





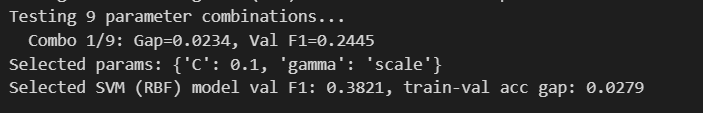
می توان دید که svm خطی به صورت کلی بهتر عمل می کند (شاید زدن pca در این تاثیر داشته است)

فاز 2:  
جلوگیری از Overfitting چگونه انجام شده؟

1. Regularization در تنظیمات مدل‌ها

Random Forest، Gradient Boosting، و Logistic Regression به‌صورت کنترل‌شده با پارامترهای منظم‌کننده مثل:

max\_depth، min\_samples\_split، min\_samples\_leaf برای محدود کردن درخت‌ها

C در SVM و Logistic برای کاهش پیچیدگی مدل

subsample و learning\_rate برای جلوگیری از یادگیری بیش از حد در Boosting

2. ارزیابی مدل با Cross-Validation (۵-لایه‌ای)

استفاده از StratifiedKFold برای تقسیم دقیق‌تر و متوازن داده‌ها.

امتیازدهی با f1\_macro به جای فقط accuracy (حساس‌تر به عدم تعادل کلاس‌ها).

بررسی تفاوت دقت آموزش و اعتبارسنجی (train-test gap) برای شناسایی overfitting.

3. محدود کردن انتخاب مدل به‌شرط Gap کم

شرط: فقط مدلی انتخاب شود که gap بین دقت آموزش و اعتبارسنجی کمتر از آستانه‌ی مشخص (مثلاً 0.05) باشد.

این باعث حذف مدل‌هایی می‌شود که آموزش را خوب یاد می‌گیرند اما در تست ضعیف هستند.

4. PCA برای کاهش بعد داده‌ها

اگر تعداد ویژگی‌ها زیاد باشد (بیش از 500)، از PCA با n\_components=0.85 استفاده می‌شود.

این یعنی فقط 85٪ از واریانس داده‌ها حفظ می‌شود.

کاهش بعد باعث کاهش پیچیدگی مدل و در نتیجه کاهش overfitting می‌شود.

5. ترکیب مدل‌ها با Stacking (Ensemble Learning)

استفاده از stacking باعث می‌شود خطاهای یک مدل توسط مدل‌های دیگر پوشش داده شود.

استفاده از meta-learner ساده (LogisticRegression با C=0.1) که خود منظم‌کننده است.

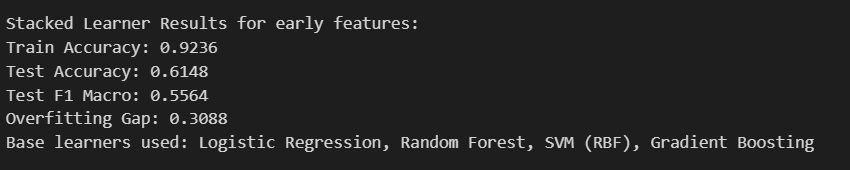
ترکیب خروجی پیش‌بینی احتمال‌ها (predict\_proba) از مدل‌های مختلف، نه فقط کلاس نهایی.

6. استفاده محدود از ویژگی‌های خام در Meta-Learner

فقط زمانی که تعداد ویژگی‌ها کم باشد، از ویژگی‌های اولیه همراه با خروجی مدل‌ها استفاده می‌شود (برای جلوگیری از overfitting در Meta-Learner).

راه های بهبود : Early Stoppingو Data Augmentation و Nested Cross-Validation

ویژگی های اولیه :

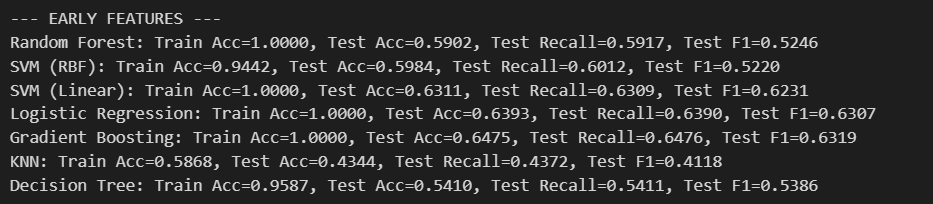


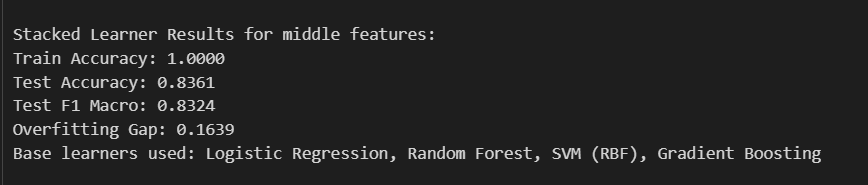
همان طور که می توان دید دقت کاهش یافته اما اختلاف acc بین train , test کاهش یافته است .

با افزایش دقت train ، Test نیز افزایش می یابد اما نسبت انها در درصد های اخر بسیار زیاد است .(train خیلی کم زیاد می شود)

طبقه بند ها به هم کمک نمی کنند و سیاست مقابله با overfitting به صورت سختگیرانه ایجاد شده که در درصد پایین ضعیف عمل می کند

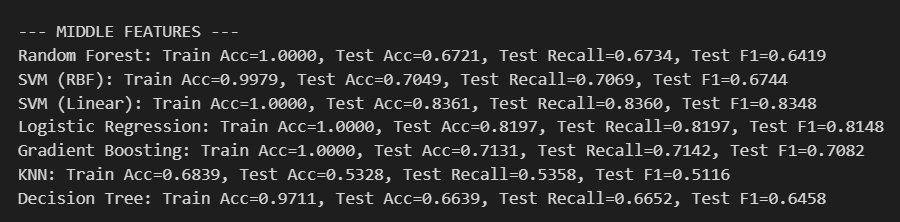
شاید با (Early Stoppingو Data Augmentation) می توانست آن را به حد خوبی کاهش داد اما دقت ممکن بود پایین تر بیاید (با توجه به این که ویژگی های ابتدایی است)

فاز 1 :

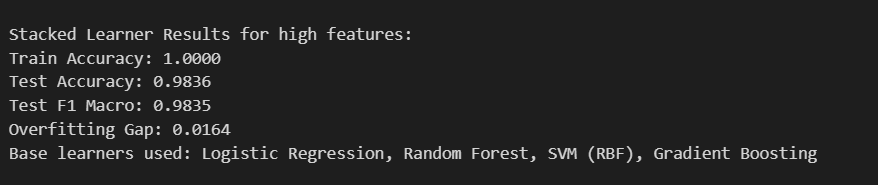
ویژگی های میانی :

می بینیم که نتیجه کمی بهتر شده است (در حالت قبل از بین این چهار تا logostic regression با 81% F1= بالاترین بود) چون از Stack learner استفاده می کنیم. ( طبقه بند ها مکمل هم دیگر)

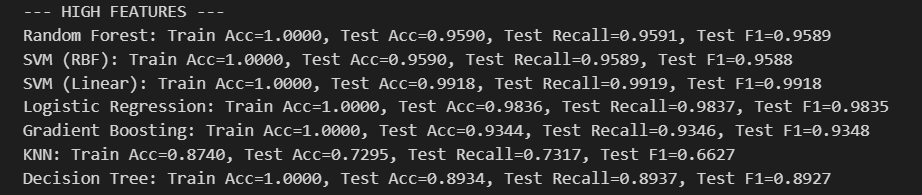
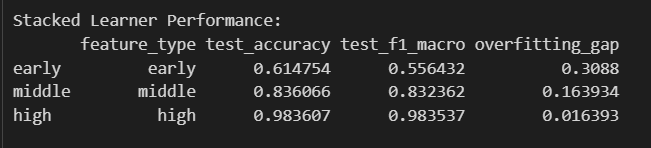
از فاز 1:



ویژگی های بالا :



در این مرحله f1 تغییری نمی کند.( طبقهبند ها به هم کمک می کنند و همپوشانی دارند و نقاط مختلف را درست تشخیص نمی دهند)

فاز 1 :