

۱ نتایج فاز اول: استخراج ویژگی و طبقه‌بندی ابتدایی

در این بخش، عملکرد مدل‌های طبقه‌بندی مختلف بر پایه سه سطح متفاوت از ویژگی‌های استخراج شده از مدل ResNet18 مورد ارزیابی قرار گرفت. ویژگی‌ها به سه دسته تقسیم شدند: ویژگی‌های سطح پایین (تا لایه maxpool)، ویژگی‌های میانی (تا پایان layer2) و ویژگی‌های سطح بالا (تا لایه avgpool). برای هر سطح، شش مدل طبقه‌بندی مختلف شامل Logistic، Random Forest، KNN، SVM، Gaussian NB و Extra Trees، Regression آزمایشی بررسی شد.

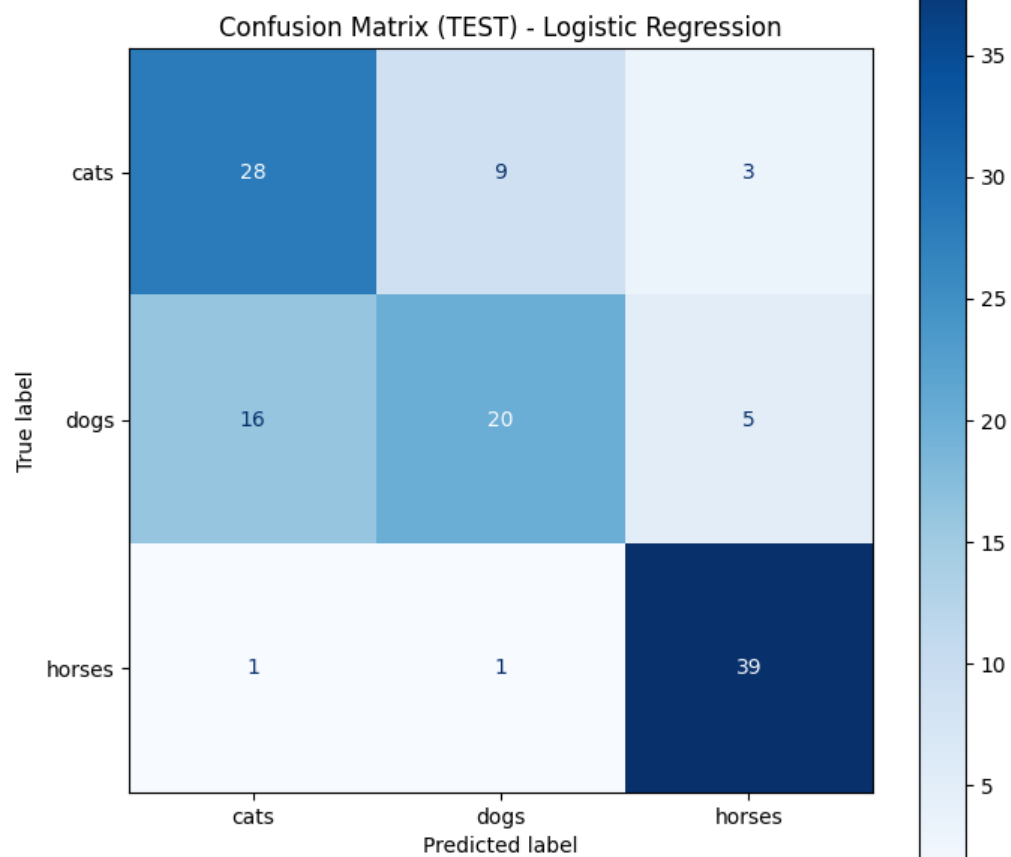
۱. ویژگی‌های سطح پایین (Low-Level Features):

در این سطح، ویژگی‌ها شامل لبه‌ها، الگوهای ابتدایی و اطلاعات بافتی پایه هستند. نتایج نشان دادند که این ویژگی‌ها در تفکیک کلاس‌ها محدودیت دارند. دقت بهترین مدل (Logistic Regression) برابر با 71% روی داده‌های آزمایشی بود. کلاس horses به طور قابل توجهی بهتر از cats و dogs طبقه‌بندی شد که نشان‌دهنده تمایز بیشتر این کلاس در ویژگی‌های سطح پایین است.

Evaluating SVM...					Evaluating Logistic Regression...				
SVM CV folds: 100% ██████████ 5/5 [19:28<00:00, 233.79s/it]					Logistic Regression CV folds: 100% ██████████ 5/5 [02:14<00:00, 26.81s/it]				
	precision	recall	f1-score	support		precision	recall	f1-score	support
cats	0.57	0.80	0.67	162	cats	0.62	0.71	0.66	162
dogs	0.61	0.43	0.50	161	dogs	0.64	0.54	0.58	161
horses	0.85	0.75	0.80	161	horses	0.79	0.79	0.79	161
accuracy			0.66	484	accuracy			0.68	484
macro avg	0.68	0.66	0.66	484	macro avg	0.68	0.68	0.68	484
weighted avg	0.68	0.66	0.66	484	weighted avg	0.68	0.68	0.68	484
Evaluating KNN...					Evaluating Extra Trees...				
KNN CV folds: 100% ██████████ 5/5 [00:07<00:00, 1.45s/it]					Extra Trees CV folds: 100% ██████████ 5/5 [00:15<00:00, 3.15s/it]				
	precision	recall	f1-score	support		precision	recall	f1-score	support
cats	0.42	0.86	0.57	162	cats	0.61	0.72	0.66	162
dogs	0.51	0.31	0.39	161	dogs	0.58	0.50	0.53	161
horses	0.89	0.32	0.47	161	horses	0.82	0.80	0.81	161
accuracy			0.50	484	accuracy			0.67	484
macro avg	0.61	0.50	0.47	484	macro avg	0.67	0.67	0.67	484
weighted avg	0.61	0.50	0.47	484	weighted avg	0.67	0.67	0.67	484
Evaluating Random Forest...					Evaluating Gaussian NB...				
Random Forest CV folds: 100% ██████████ 5/5 [00:38<00:00, 7.78s/it]					Gaussian NB CV folds: 100% ██████████ 5/5 [00:05<00:00, 1.08s/it]				
	precision	recall	f1-score	support		precision	recall	f1-score	support
cats	0.59	0.68	0.63	162	cats	0.61	0.72	0.66	162
dogs	0.57	0.50	0.53	161	dogs	0.55	0.50	0.52	161
horses	0.81	0.79	0.80	161	horses	0.76	0.69	0.72	161
accuracy			0.66	484	accuracy			0.64	484
macro avg	0.66	0.66	0.66	484	macro avg	0.64	0.64	0.63	484
weighted avg	0.66	0.66	0.66	484	weighted avg	0.64	0.64	0.63	484
🏆 Best model based on Accuracy: **Logistic Regression** with Accuracy = 0.6799									

🚀 FINAL TEST REPORT for **Logistic Regression**

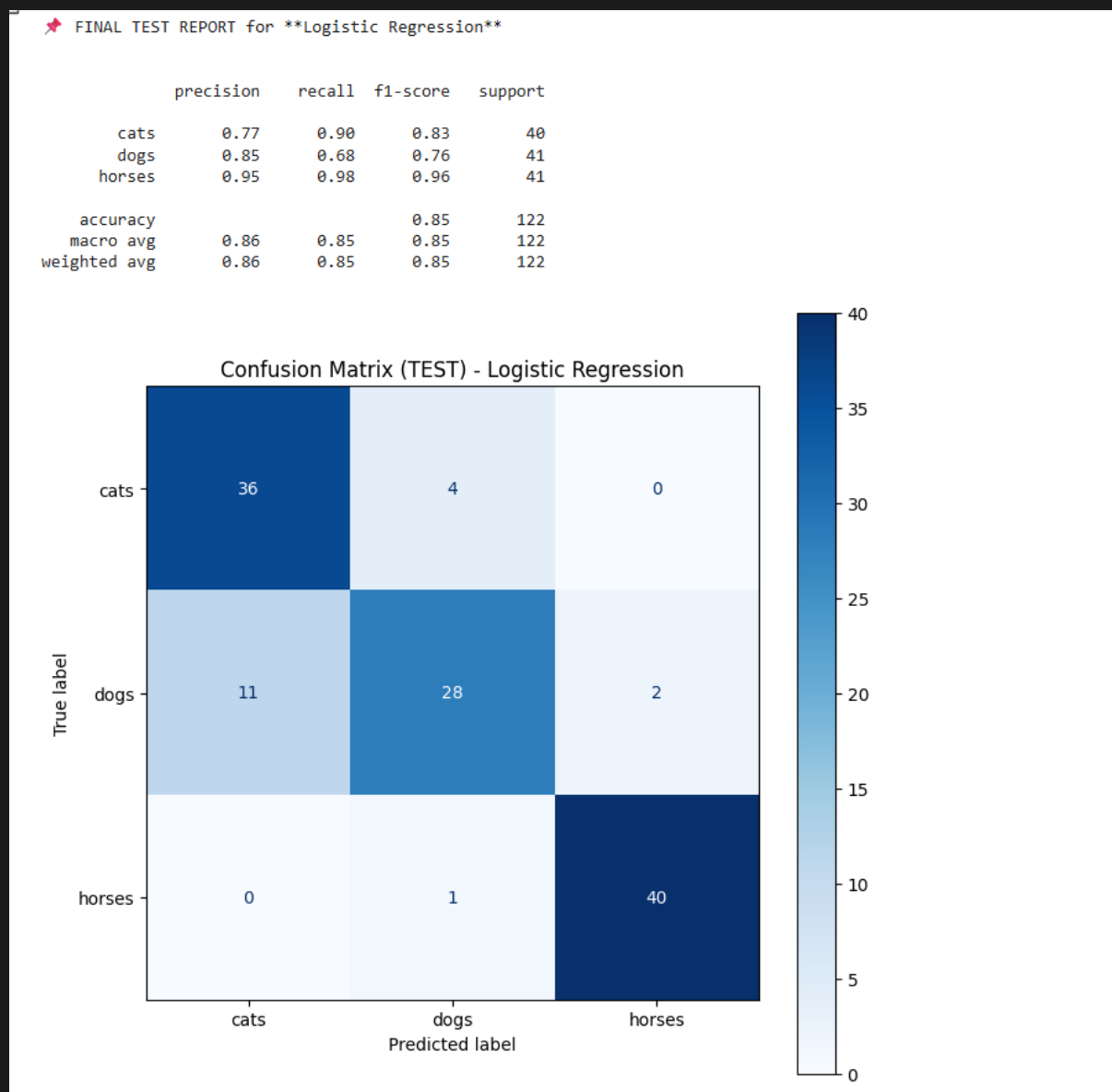
	precision	recall	f1-score	support
cats	0.62	0.70	0.66	40
dogs	0.67	0.49	0.56	41
horses	0.83	0.95	0.89	41
accuracy			0.71	122
macro avg	0.71	0.71	0.70	122
weighted avg	0.71	0.71	0.70	122



۲. ویژگی‌های میانی (Mid-Level Features):

با استفاده از لایه‌های layer1 و layer2، مدل قادر به استخراج ویژگی‌های ترکیبی و ساختاریافته‌تری شد. در این مرحله، دقت مدل Logistic Regression به حدود 85% افزایش یافت و کلاس‌ها به صورت متوازن‌تری طبقه‌بندی شدند. نسبت به مرحله‌ی قبل، کلاس dogs که پیش‌تر عملکرد ضعیفی داشت، بهبود محسوسی در دقت و recall نشان داد. این موضوع حاکی از قدرت بیشتر ویژگی‌های میانی در بازنمایی ساختارهای مفهومی‌تر تصویر است.

Evaluation for: mid-Level Features									
Evaluating SVM...					Evaluating Logistic Regression...				
SVM CV folds: 100% [██████████] 5/5 [10:43<00:00, 128.61s/it]					Logistic Regression CV folds: 100% [██████████] 5/5 [02:14<00:00, 26.81s/it]				
	precision	recall	f1-score	support		precision	recall	f1-score	support
cats	0.78	0.79	0.74	162	cats	0.62	0.71	0.66	162
dogs	0.66	0.68	0.67	161	dogs	0.64	0.54	0.58	161
horses	0.96	0.80	0.87	161	horses	0.79	0.79	0.79	161
accuracy			0.76	484	accuracy			0.68	484
macro avg	0.77	0.76	0.76	484	macro avg	0.68	0.68	0.68	484
weighted avg	0.77	0.76	0.76	484	weighted avg	0.68	0.68	0.68	484
Evaluating KNN...					Evaluating Extra Trees...				
KNN CV folds: 100% [██████████] 5/5 [00:03<00:00, 1.46it/s]					Extra Trees CV folds: 100% [██████████] 5/5 [00:15<00:00, 3.15s/it]				
	precision	recall	f1-score	support		precision	recall	f1-score	support
cats	0.53	0.21	0.30	162	cats	0.61	0.72	0.66	162
dogs	0.47	0.42	0.44	161	dogs	0.58	0.50	0.53	161
horses	0.46	0.78	0.58	161	horses	0.82	0.80	0.81	161
accuracy			0.47	484	accuracy			0.67	484
macro avg	0.49	0.47	0.44	484	macro avg	0.67	0.67	0.67	484
weighted avg	0.49	0.47	0.44	484	weighted avg	0.67	0.67	0.67	484
Evaluating Random Forest...					Evaluating Gaussian NB...				
Random Forest CV folds: 100% [██████████] 5/5 [00:20<00:00, 4.14s/it]					Gaussian NB CV folds: 100% [██████████] 5/5 [00:05<00:00, 1.08s/it]				
	precision	recall	f1-score	support		precision	recall	f1-score	support
cats	0.60	0.67	0.64	162	cats	0.61	0.72	0.66	162
dogs	0.55	0.53	0.54	161	dogs	0.55	0.50	0.52	161
horses	0.87	0.79	0.83	161	horses	0.76	0.69	0.72	161
accuracy			0.67	484	accuracy			0.64	484
macro avg	0.67	0.67	0.67	484	macro avg	0.64	0.64	0.63	484
weighted avg	0.67	0.67	0.67	484	weighted avg	0.64	0.64	0.63	484
					🏆 Best model based on Accuracy: **Logistic Regression** with Accuracy = 0.6799				



۳. ویژگی‌های سطح بالا (High-Level Features):

در این حالت، از تمام شبکه (به جز لایه FC نهایی) برای استخراج ویژگی استفاده شد. این ویژگی‌ها شامل نمایش‌های انتزاعی از مفاهیم بصری مانند "چهره گربه" یا "پیکربندی بدن اسب" هستند. مدل SVM در این مرحله با دقت 100% روی داده‌ی تست بهترین عملکرد را نشان داد. تمامی کلاس‌ها بدون خطا شناسایی شدند. این نتیجه تأییدی بر قدرت بالای نمایش‌های سطح بالا در مدل‌های یادگیری عمیق برای تفکیک دقیق بین کلاس‌ها است.

🔍 Evaluation for: High-Level Features

Evaluating SVM...
SVM CV folds: 100% | ██████████ | 5/5 [00:01<00:00, 4.69it/s]

	precision	recall	f1-score	support
cats	0.98	0.96	0.97	162
dogs	0.96	0.98	0.97	161
horses	1.00	1.00	1.00	161
accuracy			0.98	484
macro avg	0.98	0.98	0.98	484
weighted avg	0.98	0.98	0.98	484

Evaluating KNN...
KNN CV folds: 100% | ██████████ | 5/5 [00:00<00:00, 44.63it/s]

	precision	recall	f1-score	support
cats	0.95	0.97	0.96	162
dogs	0.97	0.94	0.96	161
horses	0.99	1.00	1.00	161
accuracy			0.97	484
macro avg	0.97	0.97	0.97	484
weighted avg	0.97	0.97	0.97	484

Evaluating Random Forest...
Random Forest CV folds: 100% | ██████████ | 5/5 [00:05<00:00, 1.17s/it]

	precision	recall	f1-score	support
cats	0.99	0.94	0.97	162
dogs	0.94	0.99	0.96	161
horses	0.99	1.00	1.00	161
accuracy			0.98	484
macro avg	0.98	0.98	0.98	484
weighted avg	0.98	0.98	0.98	484

Evaluating Logistic Regression...
Logistic Regression CV folds: 100% | ██████████ | 5/5 [00:04<00:00, 1.17it/s]

	precision	recall	f1-score	support
cats	0.97	0.96	0.97	162
dogs	0.96	0.98	0.97	161
horses	1.00	1.00	1.00	161
accuracy			0.98	484
macro avg	0.98	0.98	0.98	484
weighted avg	0.98	0.98	0.98	484

Evaluating Extra Trees...
Extra Trees CV folds: 100% | ██████████ | 5/5 [00:02<00:00, 2.21it/s]

	precision	recall	f1-score	support
cats	0.99	0.93	0.96	162
dogs	0.94	0.99	0.96	161
horses	1.00	1.00	1.00	161
accuracy			0.98	484
macro avg	0.98	0.98	0.98	484
weighted avg	0.98	0.98	0.98	484

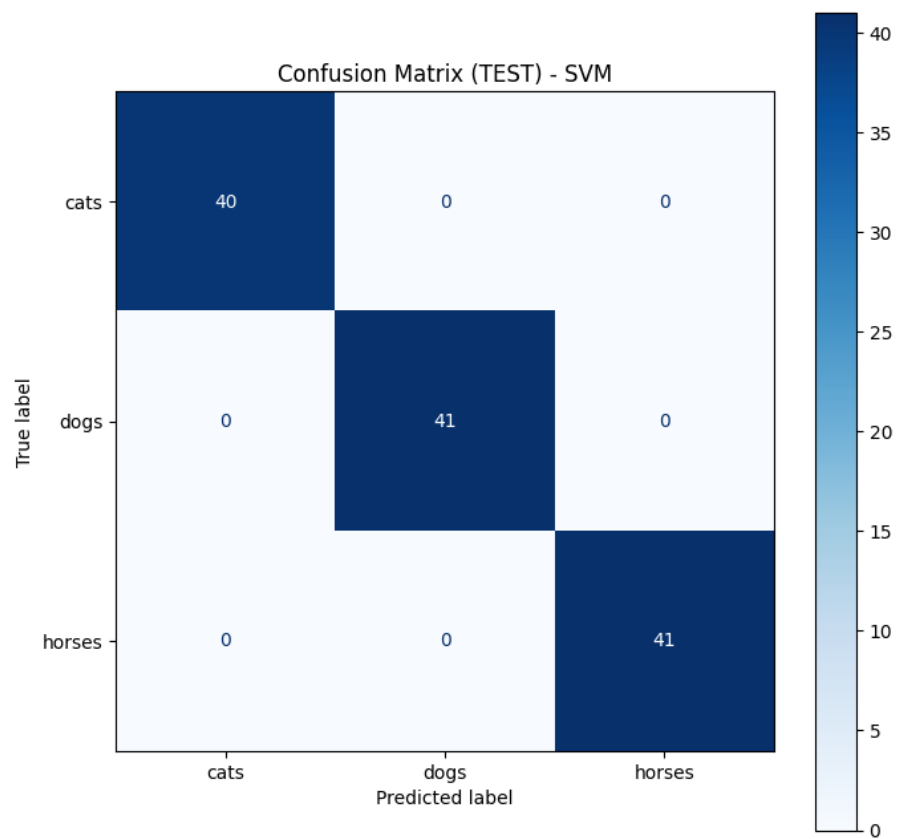
Evaluating Gaussian NB...
Gaussian NB CV folds: 100% | ██████████ | 5/5 [00:00<00:00, 75.96it/s]

	precision	recall	f1-score	support
cats	0.98	0.93	0.96	162
dogs	0.93	0.98	0.96	161
horses	1.00	1.00	1.00	161
accuracy			0.97	484
macro avg	0.97	0.97	0.97	484
weighted avg	0.97	0.97	0.97	484

🏆 Best model based on Accuracy: **SVM** with Accuracy = 0.9793

FINAL TEST REPORT for **SVM**

	precision	recall	f1-score	support
cats	1.00	1.00	1.00	40
dogs	1.00	1.00	1.00	41
horses	1.00	1.00	1.00	41
accuracy			1.00	122
macro avg	1.00	1.00	1.00	122
weighted avg	1.00	1.00	1.00	122



جمع‌بندی:

- با افزایش عمق ویژگی‌های استخراج‌شده، دقت و کیفیت طبقه‌بندی نیز به‌طور قابل توجهی افزایش یافت.
- مدل Logistic Regression در دو سطح اول بهترین عملکرد را داشت، در حالی که در سطح سوم، مدل SVM با اختلاف واضحی بهترین بود.
- کلاس horses در تمامی سطوح نسبت به دو کلاس دیگر بهتر طبقه‌بندی شد، که می‌تواند به تفاوت‌های بصری بارزتر این کلاس نسبت داده شود.
- نتایج نشان می‌دهند که حتی بدون fine-tuning شبکه، استفاده از ویژگی‌های استخراج‌شده از ResNet18 می‌تواند عملکرد بالایی در طبقه‌بندی تصاویر داشته باشد.

۲ نتایج فاز دوم

در این فاز، هدف ما استفاده از روش Stacking برای ترکیب خروجی مدل‌های مختلف است تا با بهره‌گیری همزمان از سطوح مختلف ویژگی‌ها (Low، Mid، High) و مدل‌های متنوع یادگیری ماشین، عملکرد نهایی طبقه‌بندی را بهبود ببخشیم. انتظار داریم این روش در مقایسه با فاز قبلی (استفاده مجزای هر مدل بر روی یک سطح از ویژگی‌ها)، نتایج بهتری به همراه داشته باشد.

برای پیاده‌سازی استک‌لرنر، دو رویکرد متفاوت مورد استفاده قرار گرفت:

۱. روش مبتنی بر کتابخانه (استفاده از StackingClassifier)

در این روش، از کلاس StackingClassifier در کتابخانه scikit-learn استفاده شد. مدل‌های پایه به‌صورت زیر تعریف شدند:

• SVM

• Random Forest

• Logistic Regression

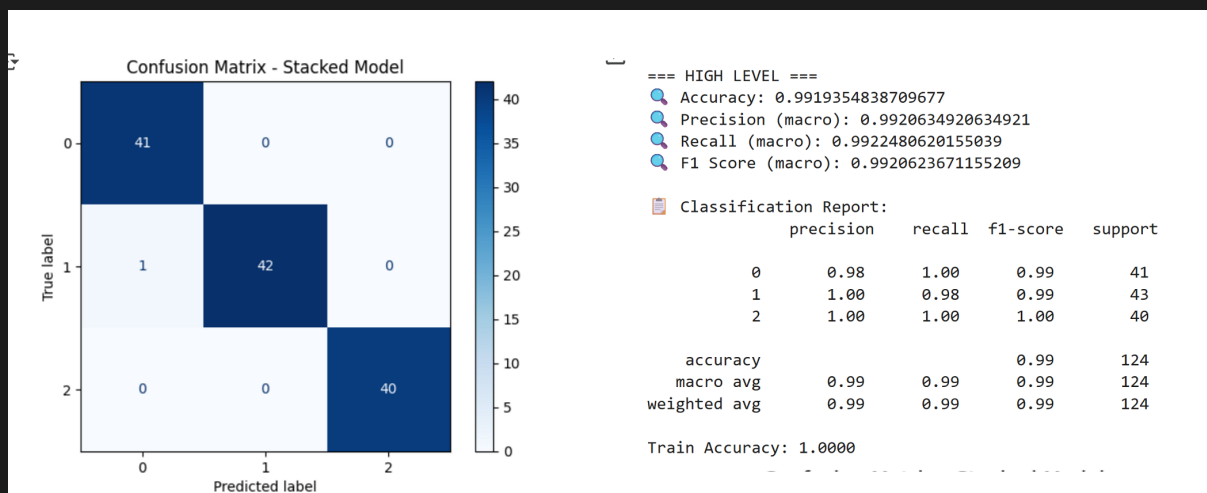
• Naive Bayes

• MLP

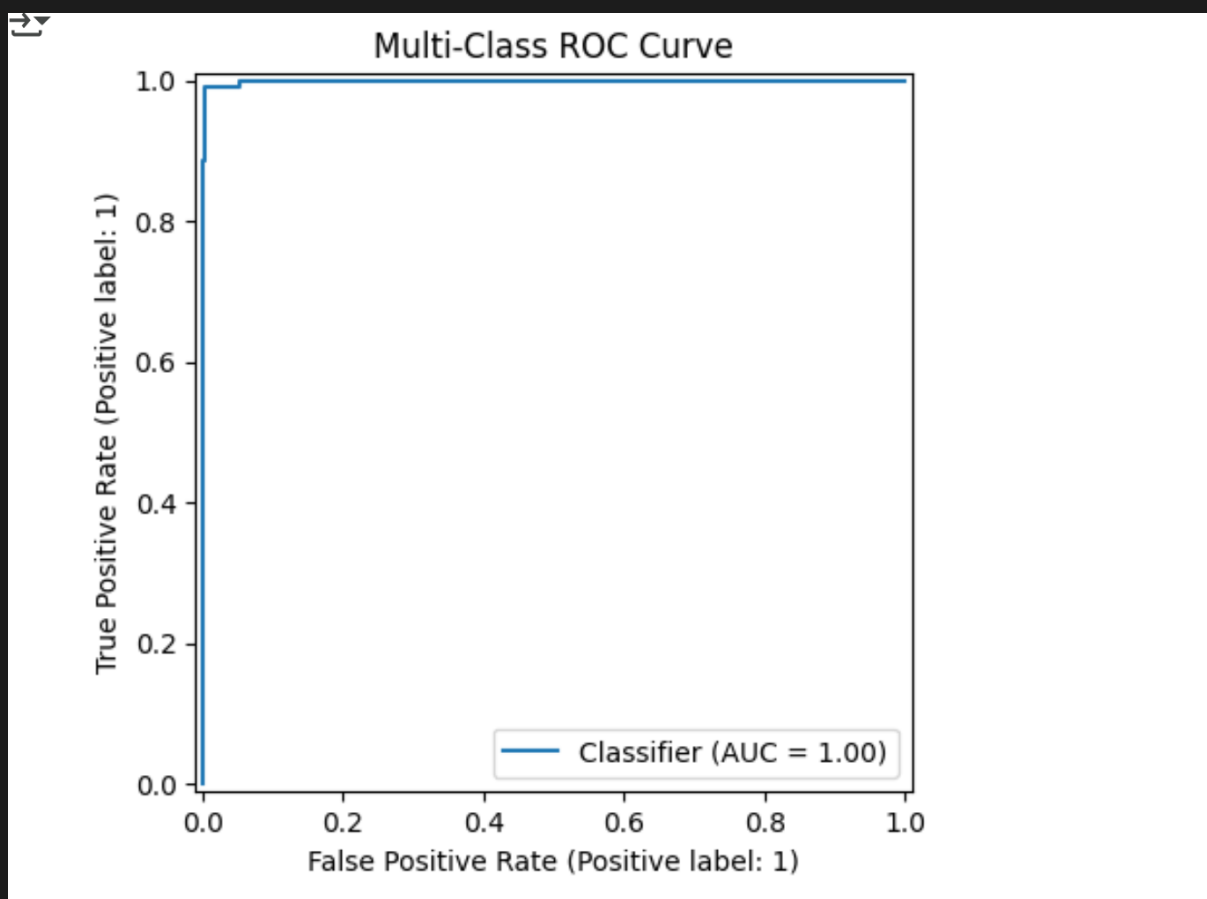
با تنظیم هایپرپارامترها و آزمون مدل‌های مختلف، سعی کردیم بهترین ترکیب ممکن را برای استک بیابیم. همچنین، حذف و اضافه کردن مدل‌های مختلف نیز انجام شد. این آزمایش‌ها نشان دادند که افزودن مدل‌های بیشتر به‌صورت قابل توجهی دقت را افزایش نمی‌دهد و صرفاً موجب افزایش زمان آموزش می‌گردد.

با اینکه دقت آموزش بسیار بالا بود، اما در داده‌های آزمون شاهد افت بودیم. این مسئله نشان‌دهنده وقوع Overfitting بود. برای رفع این مشکل، از Data Augmentation استفاده شد که تا حدی در کاهش اورفیت مؤثر واقع شد، اما به سطح رضایت‌بخشی نرسید.

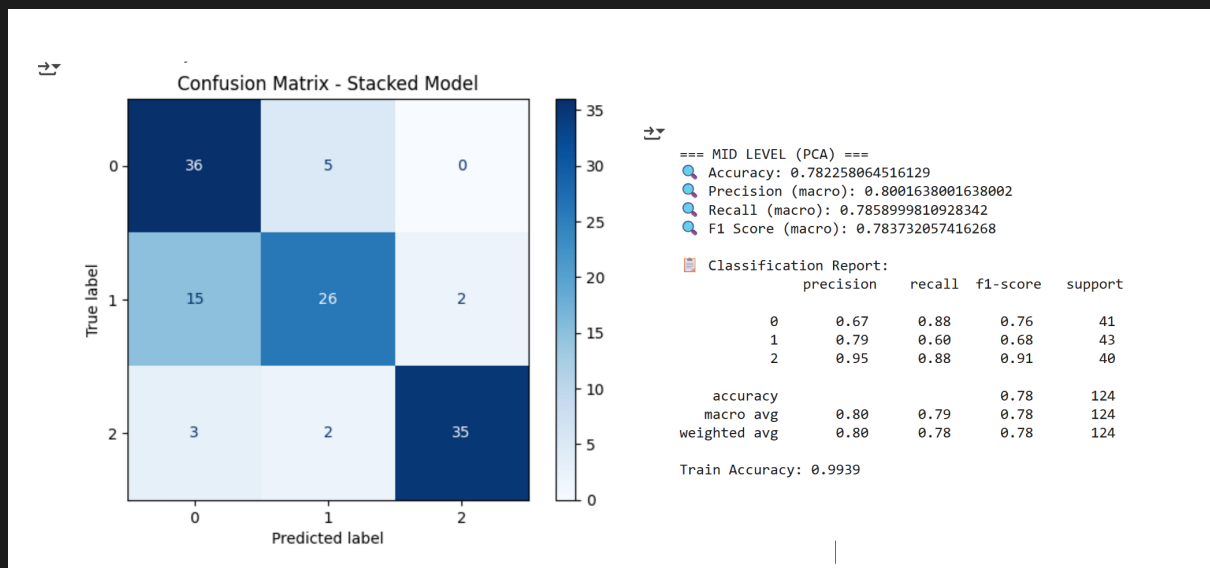
بدون اضافه کردن داده:



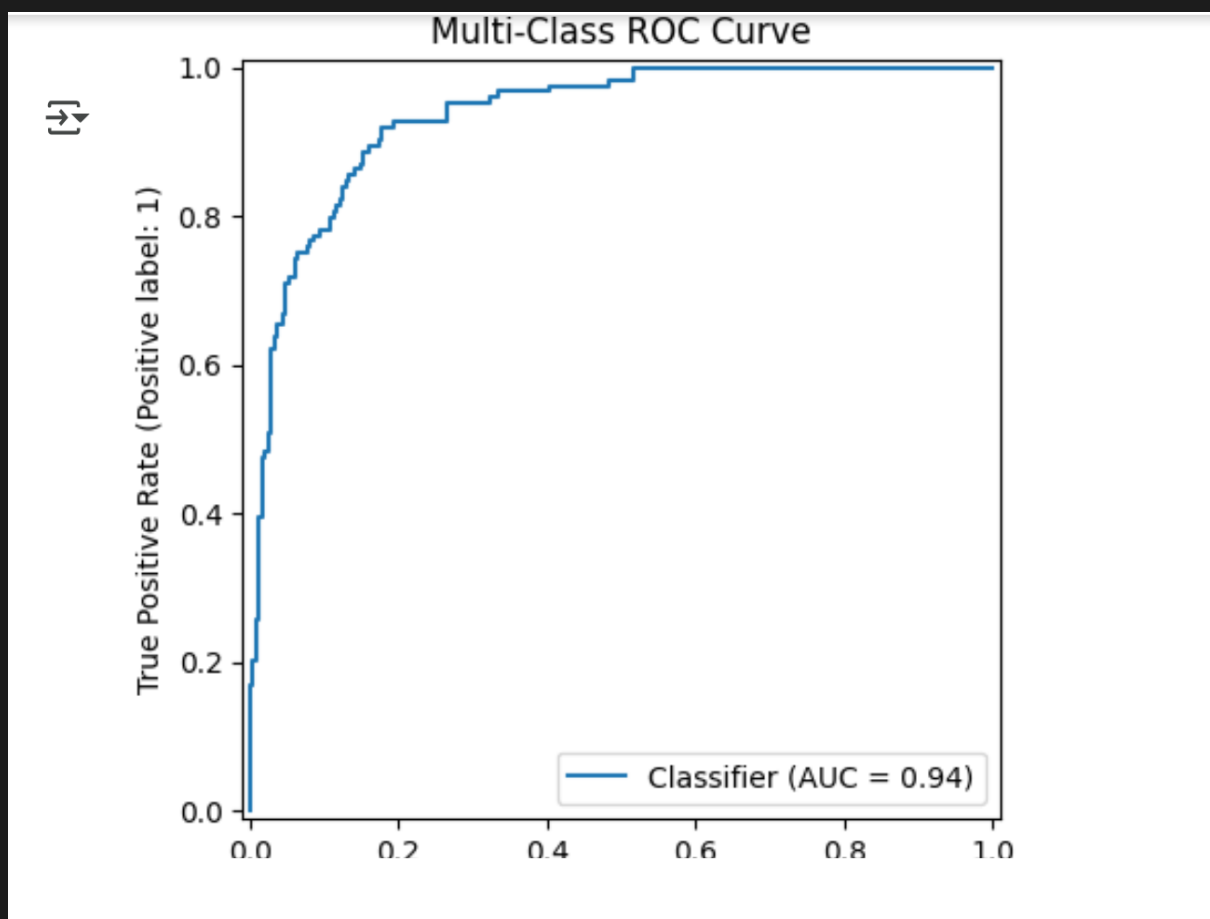
خروجی مدل استک روی ویژگی‌های سطح بالا بدون داده‌افزایی



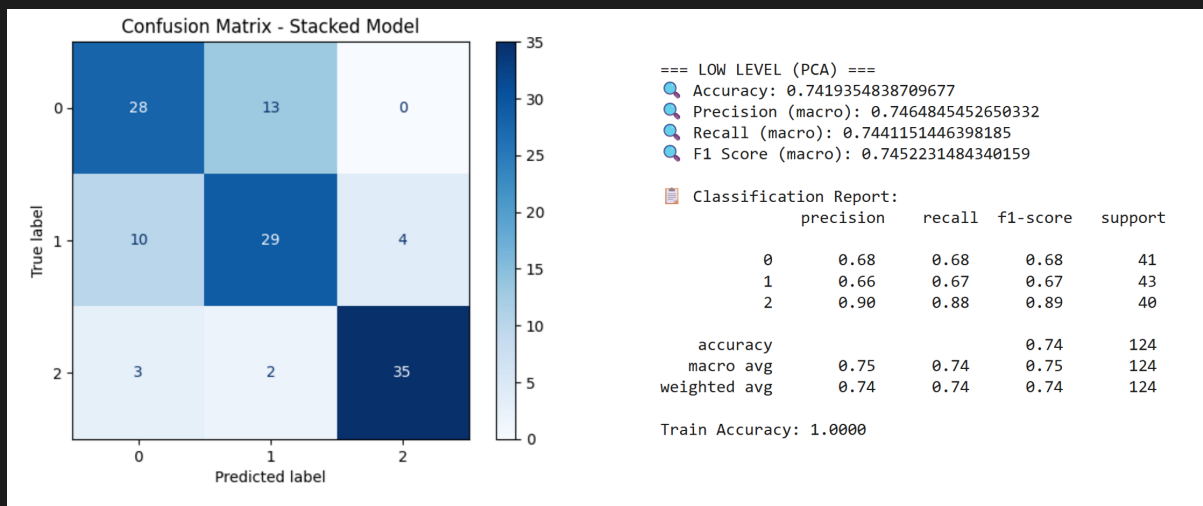
نتایج جزئی مدل استک روی ویژگی‌های سطح بالا بدون داده‌افزایی



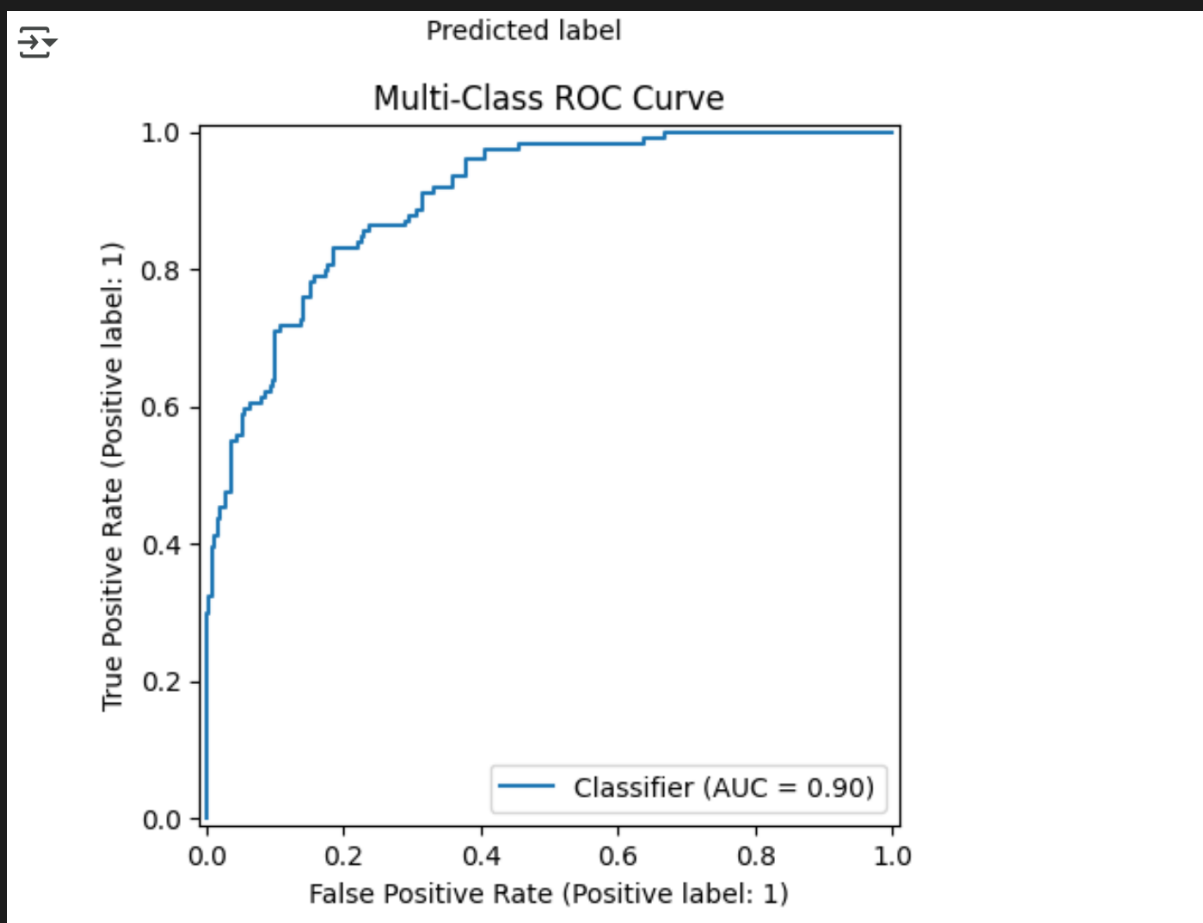
خروجی مدل استک روی ویژگی‌های سطح میانی بدون داده‌افزایی



نتایج جزئی مدل استک روی ویژگی‌های سطح میانی بدون داده‌افزایی



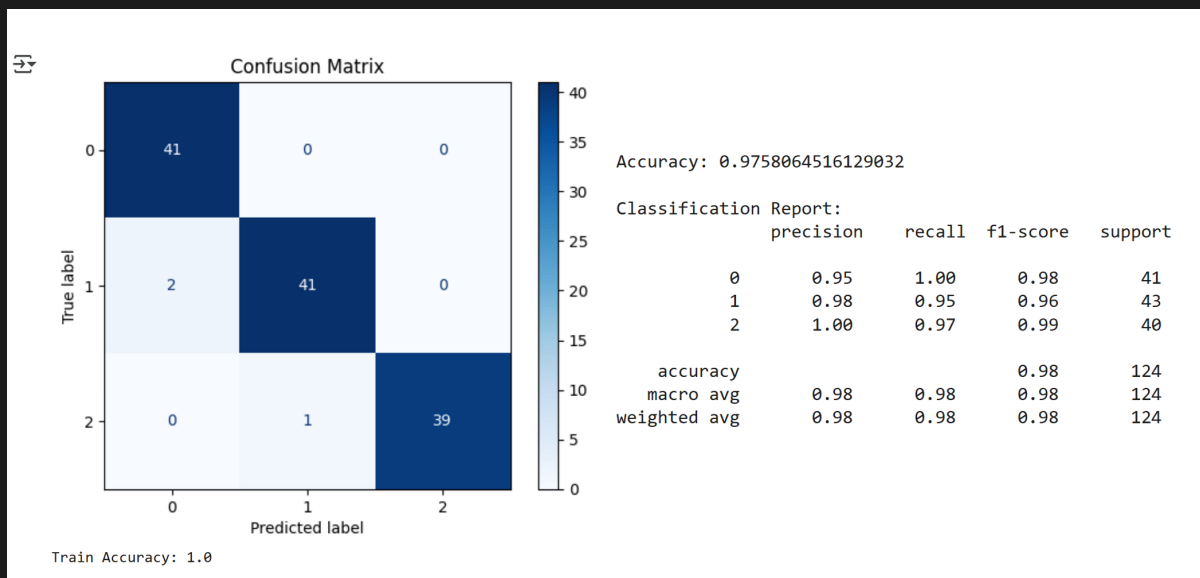
خروجی مدل استک روی ویژگی‌های سطح پایین بدون داده‌افزایی



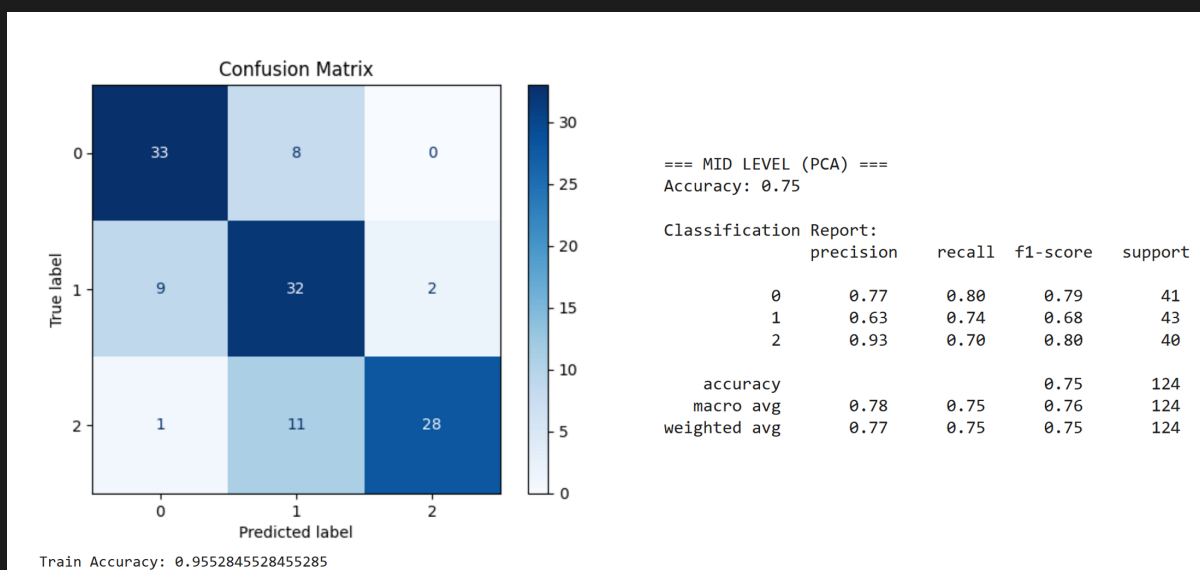
نتایج جزئی مدل استک روی ویژگی‌های سطح پایین بدون داده‌افزایی

نتایج نشان می‌دهد ویژگی High همچنان عملکرد بهتری نسبت به سایر ویژگی‌ها دارد. در دو ویژگی دیگر، اورفیت به وجود آمده است. در تمام ویژگی‌ها کلاس ۲ (اسب) بهتر از بقیه عمل کرده که احتمالاً

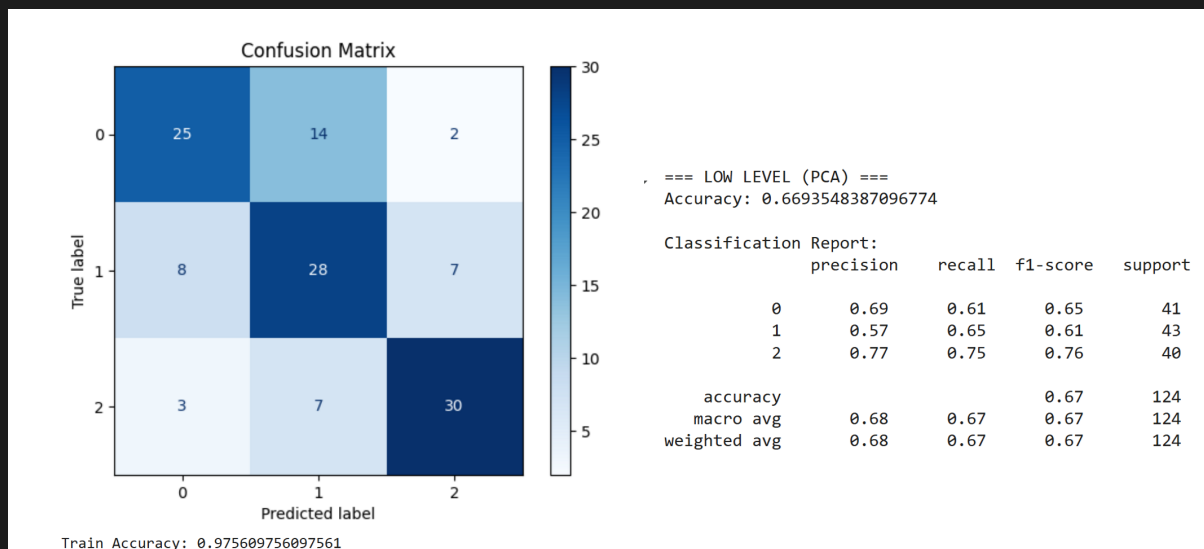
به دلیل وجود ویژگی‌های متمایز بیشتر نسبت به دو گروه دیگر است.
اضافه کردن داده:



خروجی مدل استک روی ویژگی‌های سطح بالا با داده‌افزایی



خروجی مدل استک روی ویژگی‌های سطح میانی با داده‌افزایی



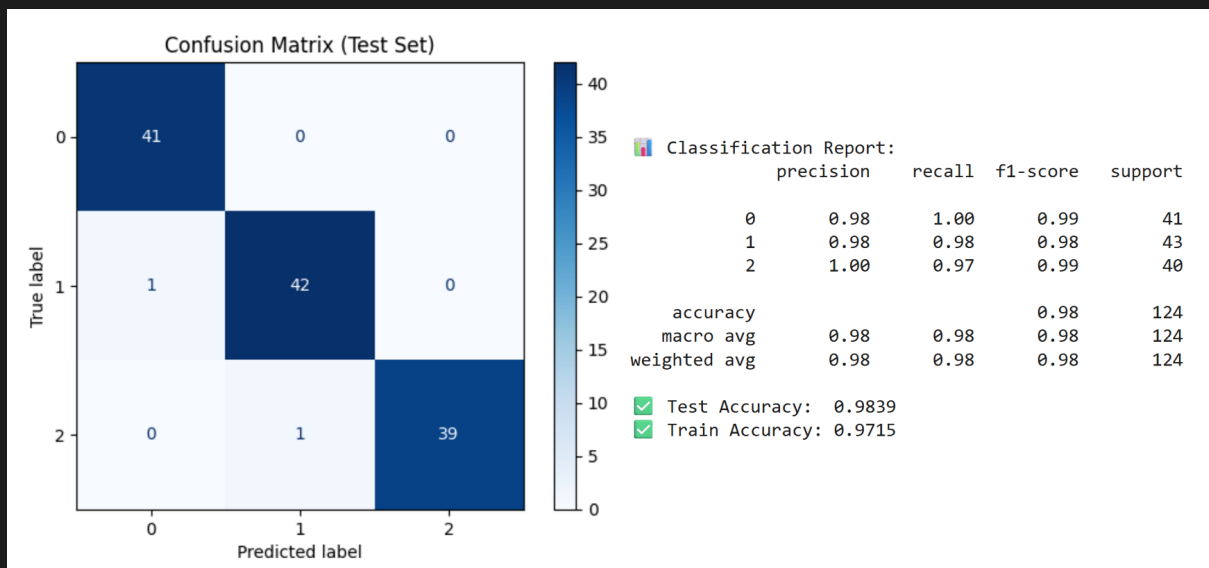
خروجی مدل استک روی ویژگی‌های سطح پایین با داده‌افزایی

نتایج نشان می‌دهد اضافه کردن داده به‌طور چشم‌گیری دقت را افزایش نمی‌دهد ولی در کاهش اورفیت مؤثر بوده است.

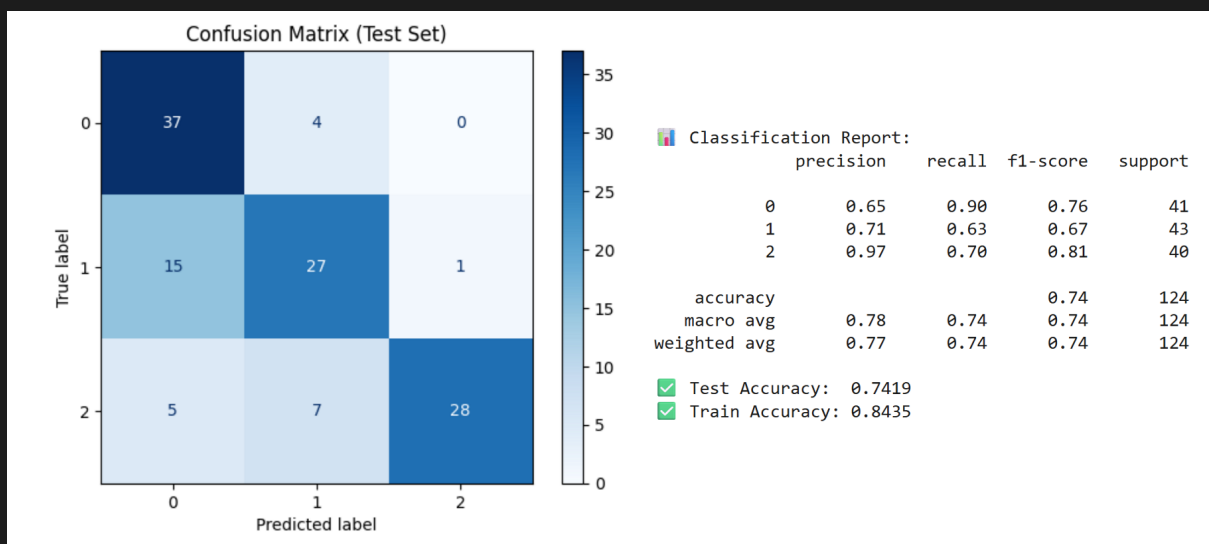
۲. روش دستی (ساخت استک با کنترل بیشتر)

در این روش، برخلاف رویکرد کتابخانه‌ای، فرآیند استکینگ به‌صورت دستی پیاده‌سازی شد. این امر کنترل بیشتری در تنظیم و انتخاب مدل‌ها به ما داد. در این ساختار، یک یا چند مدل پایه انتخاب شد و خروجی‌های احتمالاتی آن‌ها با استفاده از `cross_val_predict` استخراج شده و به مدل نهایی داده شد.

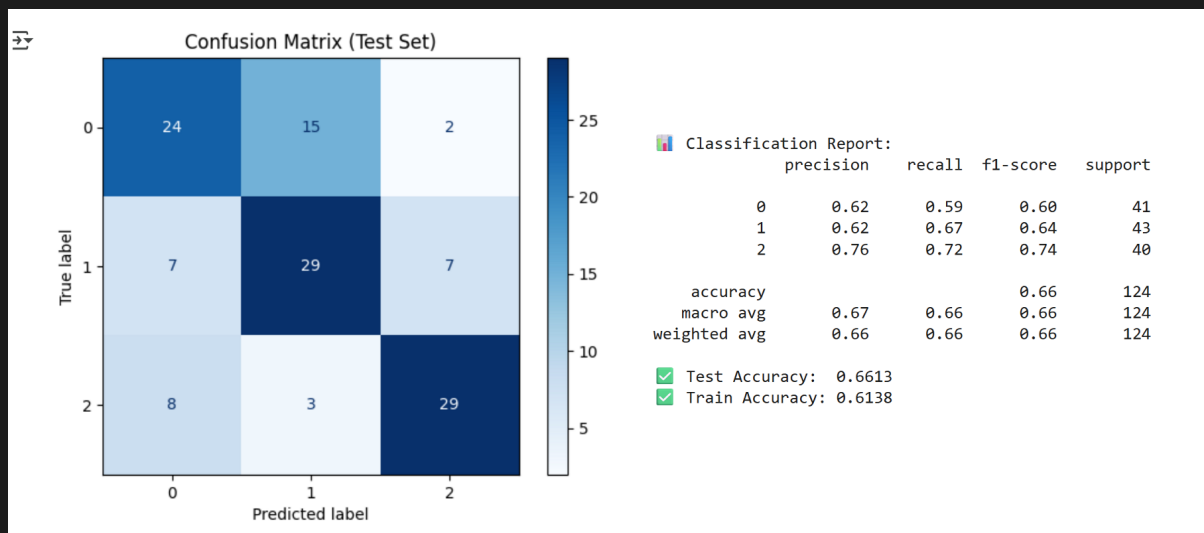
در این روش، تنظیمات مدل‌ها دقیق‌تر و هدفمندتر انجام شد. حتی بدون استفاده از داده‌افزوده‌شده نیز دقت تست بالاتری نسبت به روش قبلی حاصل شد. ساختار طراحی‌شده موجب کاهش اورفیت و افزایش قابلیت تعمیم مدل نهایی شد.



نتایج استک دستی روی ویژگی‌های سطح بالا



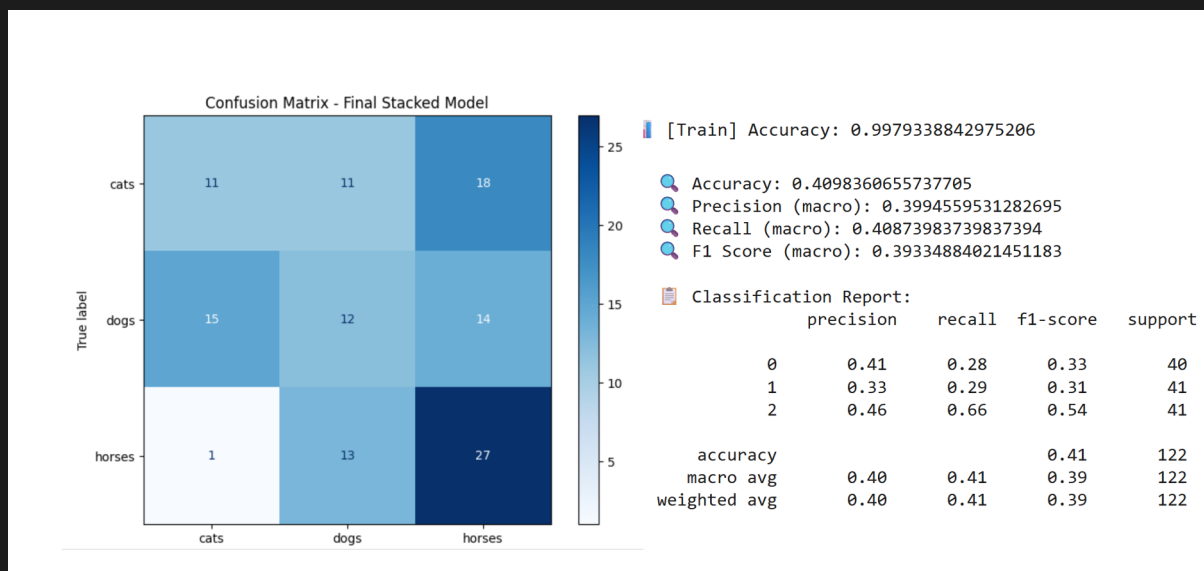
نتایج استک دستی روی ویژگی‌های سطح میانی



نتایج استک دستی روی ویژگی‌های سطح پایین

در اینجا در جلوگیری از اورفیت عملکرد بسیار بهتری مشاهده شد. در ویژگی‌های High، تمام کلاس‌ها درصد مناسبی دارند. در ویژگی Mid، کلاس ۰ و در ویژگی Low، کلاس‌های ۱ و ۲ عملکرد بهتری داشته‌اند. این نتایج نشان می‌دهد کنترل دستی در مهار اورفیت مؤثرتر بوده است.

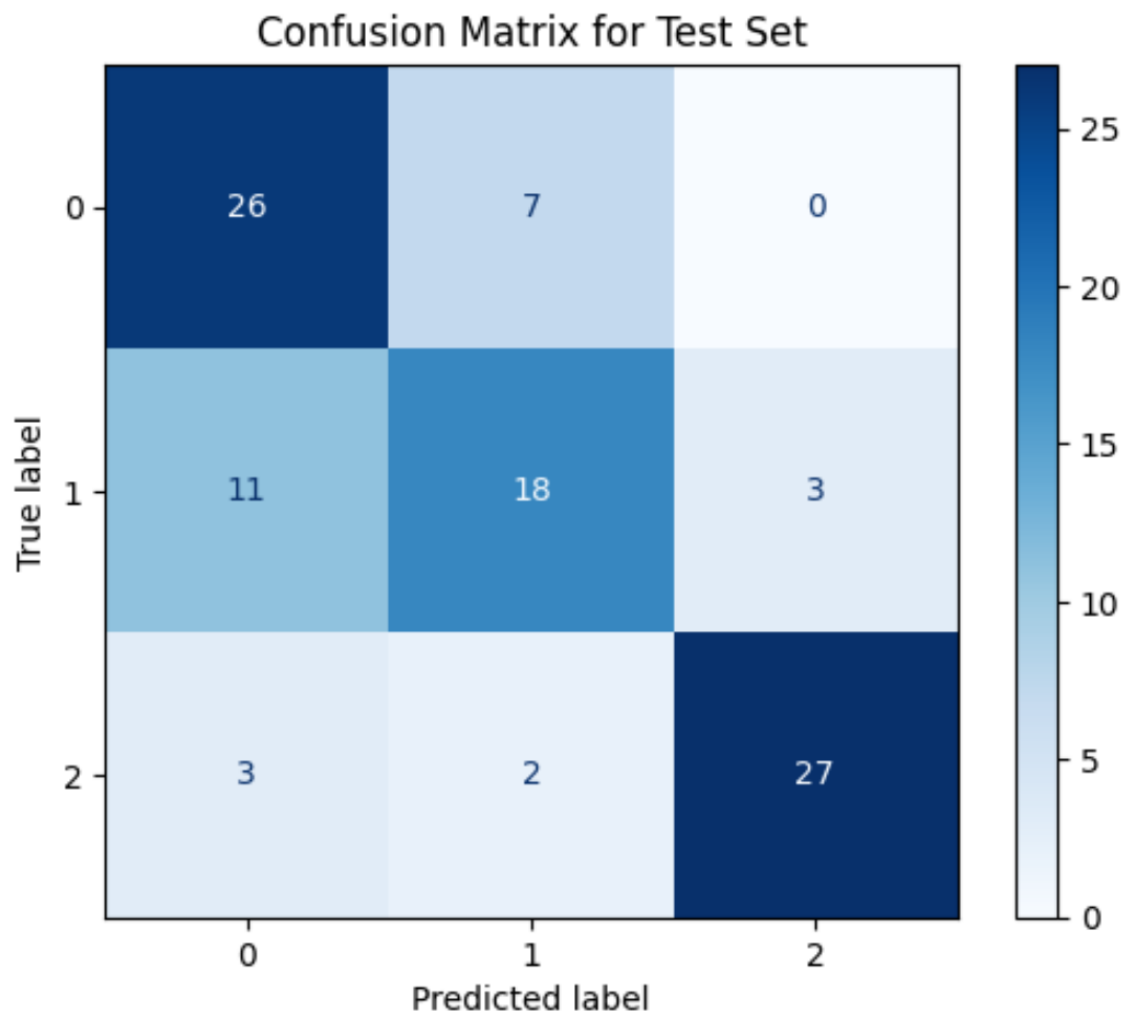
در انتهای این فاز، دو حالت دیگر نیز برای ساخت استک مورد بررسی قرار گرفت:
استفاده از تمام سطوح ویژگی‌ها به‌طور هم‌زمان به‌عنوان ورودی به مدل نهایی:



استفاده از تمام سطوح ویژگی‌ها به‌صورت هم‌زمان در مدل استک

ساخت استک ترکیبی بر اساس نتایج فاز اول، به‌گونه‌ای که از هر سطح ویژگی، فقط بهترین مدل طبقه‌بند (با توجه به دقت فاز قبلی) انتخاب شده و در استک نهایی استفاده شد:

Test Accuracy: 0.7320
Train Accuracy: 0.7209



ساخت استک ترکیبی با استفاده از بهترین طبقه‌بندها در هر سطح ویژگی

نتایج نشان دادند که حالت دوم (ترکیبی و انتخابی)، ساده‌تر، سریع‌تر و با عملکردی مشابه یا بهتر نسبت به حالت اول عمل کرده است.

چرا ویژگی High-Level عملکرد بهتری داشت؟

چون ویژگی‌های High-Level در لایه‌های نهایی شبکه هستند و اطلاعات انتزاعی و معنایی تصویر را بهتر درک می‌کنند (مثل شکل، کلاس، بافت‌های متمایز). در حالی که Low و Mid بیشتر ویژگی‌های اولیه مثل لبه‌ها و رنگ‌ها را استخراج می‌کنند که برای تفکیک کلاس‌های پیچیده کافی نیست. برای کاهش اورفیت نیز از دو روش استفاده شد: افزایش داده‌ها (Data Augmentation) و کاهش ابعاد ویژگی‌ها با PCA که باعث شد ویژگی‌ها به صورت کلی‌تر عمل کنند و وابستگی کمتری به داده‌های آموزشی داشته باشند.

مقایسه با فاز یک:

در استفاده از استک مدل‌های مختلف، دیدگاه‌های متنوعی نسبت به داده‌ها دخیل می‌شوند. مثلاً:

- SVM برای داده‌های با مرزهای واضح بهتر است،
- Random Forest مناسب داده‌های نویزی و پیچیده است،
- Naive Bayes بر پایه فرض استقلال ویژگی‌ها عمل می‌کند.

استکینگ این دیدگاه‌ها را ترکیب کرده و تصمیم نهایی هوشمندانه‌تری اتخاذ می‌کند. همچنین هر مدل پایه ممکن است در تشخیص برخی کلاس‌ها یا نمونه‌ها ضعف داشته باشد. وقتی خروجی چند مدل به مدل نهایی داده می‌شود، مدل متا (مثل Logistic Regression) یاد می‌گیرد که چه زمانی به کدام مدل بیشتر اعتماد کند.

استکینگ با استفاده از cross-validation روی مدل‌های پایه، از اورفیت روی یک مدل خاص جلوگیری می‌کند. به جای استفاده از یک مدل، از خروجی چند مدل استفاده شده و مدل نهایی تصمیم می‌گیرد کدام خروجی‌ها اهمیت بیشتری دارند.

همچنین می‌توان از سه سطح ویژگی مختلف (Low/Mid/High) به صورت ترکیبی استفاده کرد: مثلاً استفاده از MLP روی ویژگی‌های High، SVM روی Mid و Logistic Regression روی Low و ترکیب همه خروجی‌ها.

در مجموع، Stacking اغلب از هر مدل پایه به تنهایی عملکرد بهتری دارد؛ به ویژه در دیتاست‌های پیچیده و چندکلاسه. در این پروژه نیز نتایج فاز دوم (به ویژه با تنظیم دستی) از تمام نتایج فاز اول بهتر بوده‌اند. در ویژگی‌های High همیشه عملکرد مناسبی وجود داشته و در دو ویژگی دیگر نیز نسبت به بهترین حالت فاز اول یا برابر یا بهتر بوده است.