Final Report

Allison Moore

620007444

STAT 689: SPTP- Statistical Data Science

May 3, 2023

Background

Ovarian cancer is a significant issue in women's health, with more than 19,000 cases expected to occur in 2023 (American Cancer Society 2023). Early detection of ovarian cancer is incredibly difficult; however, protein mass spectra and wavelet transforms offer a unique method for detecting the presence of this cancer. Below, I will discuss the methods I used to apply machine learning techniques to evaluate these spectra for the presence of cancer. My results will be tabulated and recorded in appendices to produce a concise report. Lastly, I will discuss the results of my models and provide recommendations for future studies.

The Data and Methodology

To develop a model to predict ovarian cancer, I used a simple three step methodology, as shown in figure 1. As the data was provided in a clean format, data collection and data wrangling were not necessary. The provided data consisted of 253 observations from 29 protein mass spectra wavelet transformed slopes which were calculated using the standard method (American National Cancer Institute Internet Repository, 2022). Given this dataset, I verified there were no missing values and then conducted thorough exploratory data analysis (EDA).

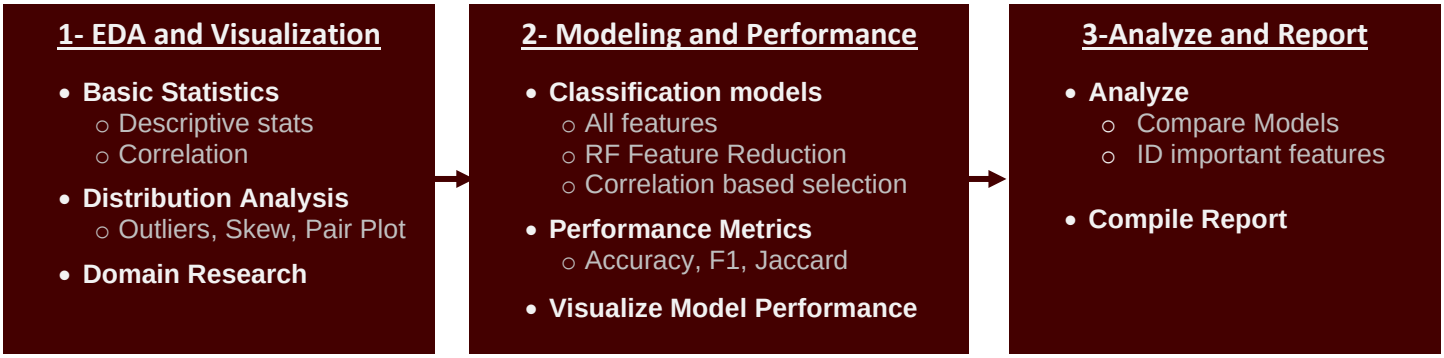

Figure 1. Machine Learning methodology. A large portion of time was spent conducting EDA and additional domain-based research to better understand the dataset.

Under the presumption that outliers are likely key indicators of abnormal or cancerous protein mass spectra, no outliers were removed from the datasets. During EDA and visualization, I uncovered several insights which will be discussed in detail in the next section.

Following EDA, I developed several user defined functions to train and test multiple machine learning models. These functions contribute to the overall organization of the notebook in Appendix C by facilitating condensed sections for the display and analysis of all model results. Three sets of models were developed: the base models using all data, models using a dataset narrowed by Random Forest feature selection, and models using a dataset with features reduced by correlation analysis (top 5 positive and top 5 negative correlations with the binary response variable, 'state'). To compare the performance of the tested models, I utilized a combination of the predicted test Accuracy, the F-1 Score and the Jaccard Index. After this analysis, the report was developed, and recommendations were made.

Results and Recommendations

Initial EDA demonstrated the insights detailed below. Refer to Appendix A for graphical representations of this information.

1. Slopes 16, 17, and 26 demonstrate substantially wider interquartile ranges (IQRs) for cancerous spectra than healthy spectra.

2. Slopes 17 and 22 exhibit substantial median shifts between healthy and cancerous spectra.

3. In pair plot analysis, Slopes 17 and 18 demonstrate clear separation between healthy and cancerous spectra when compared to other slopes.

4. Cancerous spectra tend to be more negative on average with the largest disparity between healthy and cancerous spectra amongst Slopes 17, 25, and 26.

As a result of these insights, EDA led to the expectation of finding significance in higher ordered slopes (in particular, Slope 17) in regard to model feature importance.

Table 1 summarizes the models tested, the parameters applied during a hyperparameter grid search, and their resultant performance metrics. The top three important features are also included in the 'Top 3 Slopes' column. Of note, these are only reported for the base model. For feature importance for narrowed and reduced data, refer to Appendix C.

Table 1.
Summary of all models tested. All parameters included in a cross-validation grid search are listed in the far-right column. Appendix B provides graphical representations of these results. The highest accuracy occurs under Correlation Reduction, Random Forest. For more information about best model parameters, refer to Appendix C under 'Models.' Scaled LR performed nearly identically to regular LR, likely due to the data existing on a short-range scale already.

Model	Metric	Base	RF Narrowed	Correlation Reduction	Top 3 Slopes	Parameters Tested
Logistic Regression	Accuracy	95.39%	92.13%	92.31%	17	C: np.logspace(-4, 4, 50)
	F1	0.9538	0.9204	0.9207	26	penalty: l1, l2 , elasticnet
	Jaccard	0.9119	0.8540	0.8545	25	solver: lbfgs, liblinear, newton-cg, sag
Scaled Logistic Regression	Accuracy	95.39%	94.08%	92.31%		C: np.logspace(-4, 4, 50)
	F1	0.9542	0.9403	0.9015		penalty: l1, l2 , elasticnet
	Jaccard	0.9119	0.8882	0.8214		solver: lbfgs, liblinear, newton-cg, sag
Normalized Logistic Regression	Accuracy	92.76%	90.79%	92.31%		C: np.logspace(-4, 4, 50)
	F1	0.9278	0.9074	0.9024		penalty: l1, l2 , elasticnet
	Jaccard	0.8650	0.8313	0.8214		solver: lbfgs, liblinear, newton-cg, sag
K Nearest Neighbors	Accuracy	93.12%	93.14%	88.98%		n_neighbors: 1-31
	F1	0.9312	0.9312	0.8898		weights: uniform, distance
	Jaccard	0.8716	0.8716	0.8014		algorithm: auto, ball_tree, kd_tree, brute
Random Forest	Accuracy	92.31%	Used for Feature Selection	96.15%	17	criterion: gini, entropy
	F1	0.9115		0.9214	18	max_depth: 2-20
	Jaccard	0.8378		0.8537	22	n_estimators: 1-30
Tree and Prune	Accuracy	81.58%	84.62%	84.31%	17	criterion: gini, entropy
	F1	0.8171	0.8382	0.8443	26	max_depth: 1-20
	Jaccard	0.6889	0.7175	0.7288	18	min_sample_split: 2-10
Support Vector Classification	Accuracy	67.42%	67.42%	67.42%	17	C: 0.1, 1, 10, 100, 1000
	F1	0.5429	0.5429	0.5429	18	gamma: 1, 0.1, 0.01, 0.001, 0.0001
	Jaccard	0.5085	0.5085	0.5085	26	kernel: linear, rbf

As expected, the vast majority of important features originated from higher slope values, with Slope 17 offering the highest contribution to the models. Although Random Forest (RF) had the highest Accuracy, Logistic Regression (LR) appears to be a more well-rounded model due to more consistently high performance metrics. Furthermore, LR generated the best ROC Curve and demonstrated high rates of learning, as shown in the graphical results recorded in Appendix B. As recommended by Vimalajeewa, Bruce, and Vidakovic (2022, 10), the final model to select should use a threshold that corresponds to the Youden Index.

In response to these results, I recommend further analysis which considers only the most impactful spectra slopes and incorporates additional independent variables, such as the presence of risk factors (e.g., family history of ovarian cancer, use of tobacco products, age given birth, etc.) (American Cancer Society, 2023). Furthermore, more advanced machine learning algorithms may yield even more advantageous results.

Conclusion

Several models were developed in an effort to identify the best method for identifying the presence of ovarian cancer in a set of protein mass spectra. Predicted test accuracies indicate that Logistic Regression may be the best method for this task. However, this was not an exhaustive test. Further research should be conducted using more advanced machine learning models and should include additional independent variables which are unrelated to slope values. Furthermore, expanding parameter hyper grids could produce more ideal results, but will require increased levels of computing power.

Bibliography

American Cancer Society. (2023). Cancer Facts & Figures 2023. *American Cancer Society,* 21-22.

American National Cancer Institute Internet Repository. (2022).Clinical Proteomics Program. *Dataset.* https://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp. (Accessed April 27, 2023).

Vimalajeewqa, D., Bruce, S. A., and Vidakovic, B. (2022). Early Detection of Ovarian Cancer by Wavelet Analysis of Protein Mass Spectra. arXiv:2207.07028**.**

Appendix A: Data Visualization Excerpts

Refer to Appendix C for the source code of this plot.

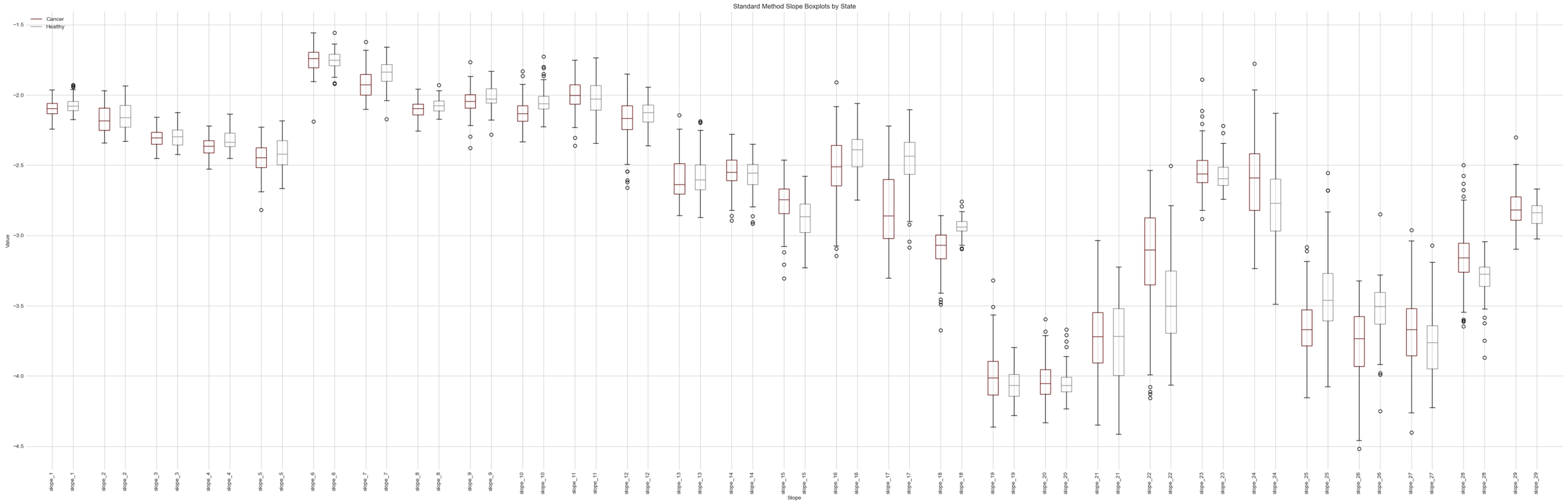

Figure A-1. Standard Method Box Plots comparing cancerous and healthy slopes. This plot yields the insights listed below:

* *Slopes 16, 17 and 26 have substantially wider IQRs for cancerous than healthy spectra. Interestingly, only Slope 16's cancerous spectra contain outliers, while only Slope 17's healthy spectra contain outliers.*
* *Slopes 17 and 22 exhibit substantial median shifts between cancerous and healthy spectra.*
* *Slope 18's cancerous medians are roughly equal to the lower end of the healthy IQRs.*

Appendix B: Selected Results

Refer to Appendix C for the source code of these plots. Selected model results are reported in the order tested.

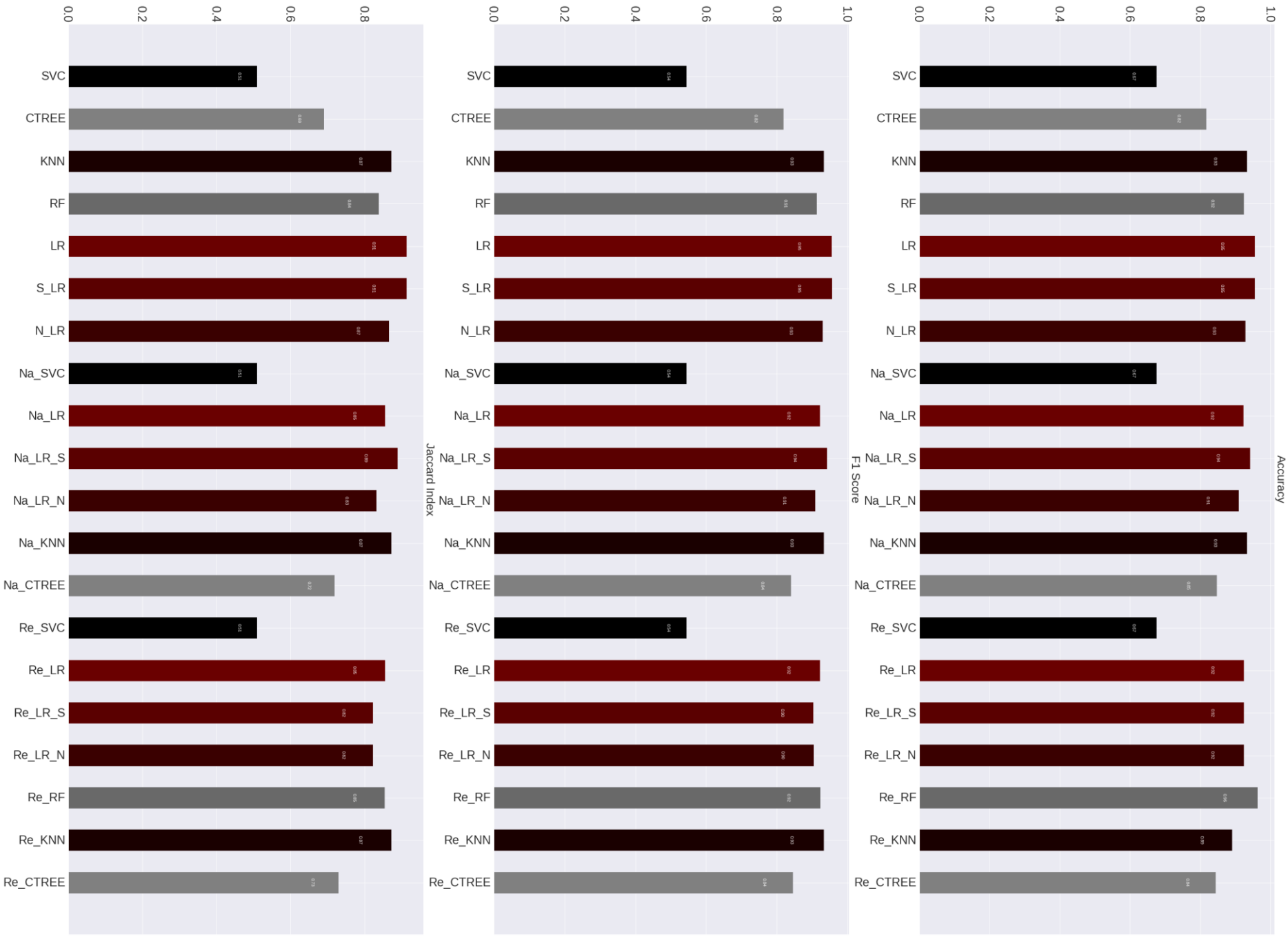

Figure B-1. All performance metrics for all models. 'Na' indicates the data was narrowed by RF, while 'Re' indicates the features were reduced by correlation.

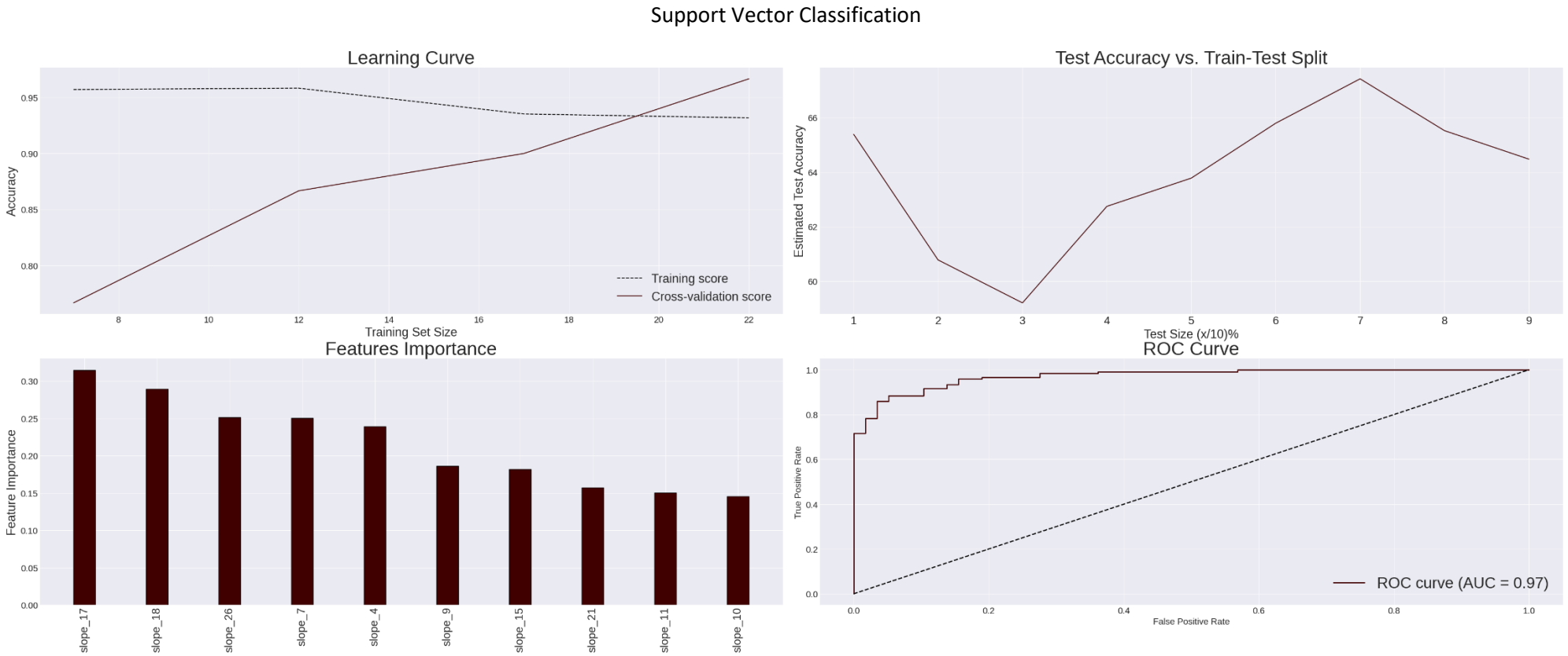

Figure B-2. Support Vector Classification Results- Base Model. The learning curve indicates the model is learning; however, the predicted test accuracy is consistently low across the test sizes. Of note, all versions of SVC provided the same results. SVC consistently provides the worst performance of all models trained and tested.

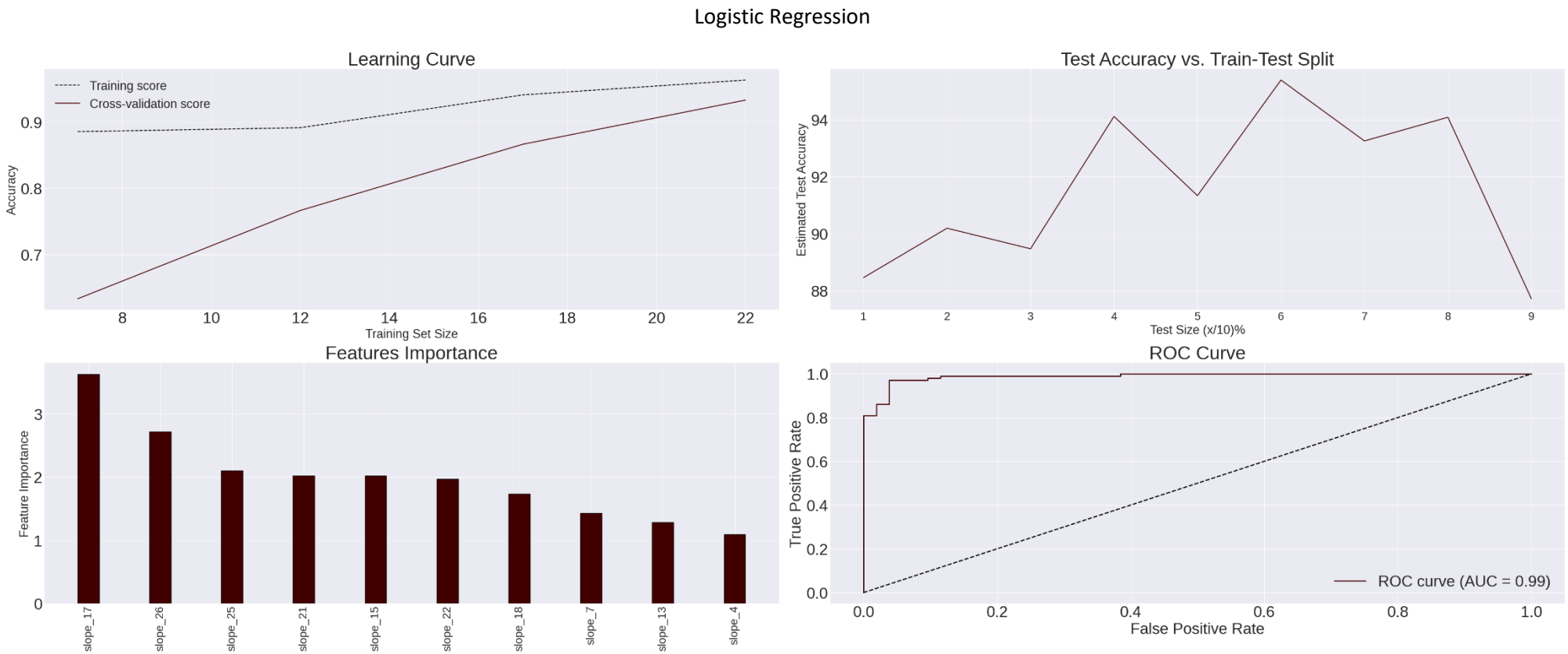

Figure B-3. Logistic Regression Results- Base Model. The learning curve indicates the model is learning at a consistent rate and the estimated test accuracy is consistently high. The ROC curve also displays a high true positive rate at lower false positive rates, as compared to other models. Scaled Logistic Regression provided similar results to the base LR model. Normalized LR provided noticeably lower performance.

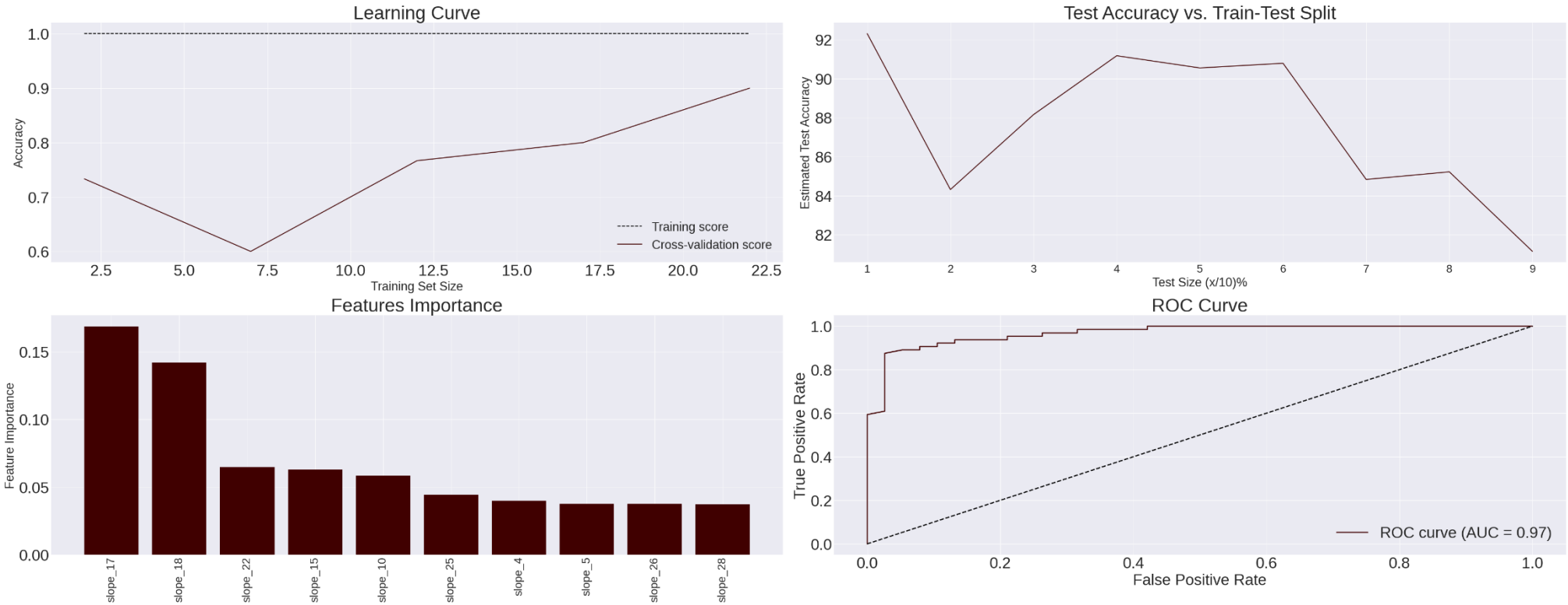

Figure B-4. Random Forest Results- Base Model. The learning curve indicates a slow learning rate, and the estimated test accuracy demonstrates large variation across test sizes. The ROC curve indicates relatively poor performance compared to other models. Of note, the correlation-based reduction of features resulted in the highest accuracy of any model; however, the learning curve, ROC curve, and Jaccard Index indicate the model does not perform as well as the Accuracy metric implies.

K Nearest Neighbors

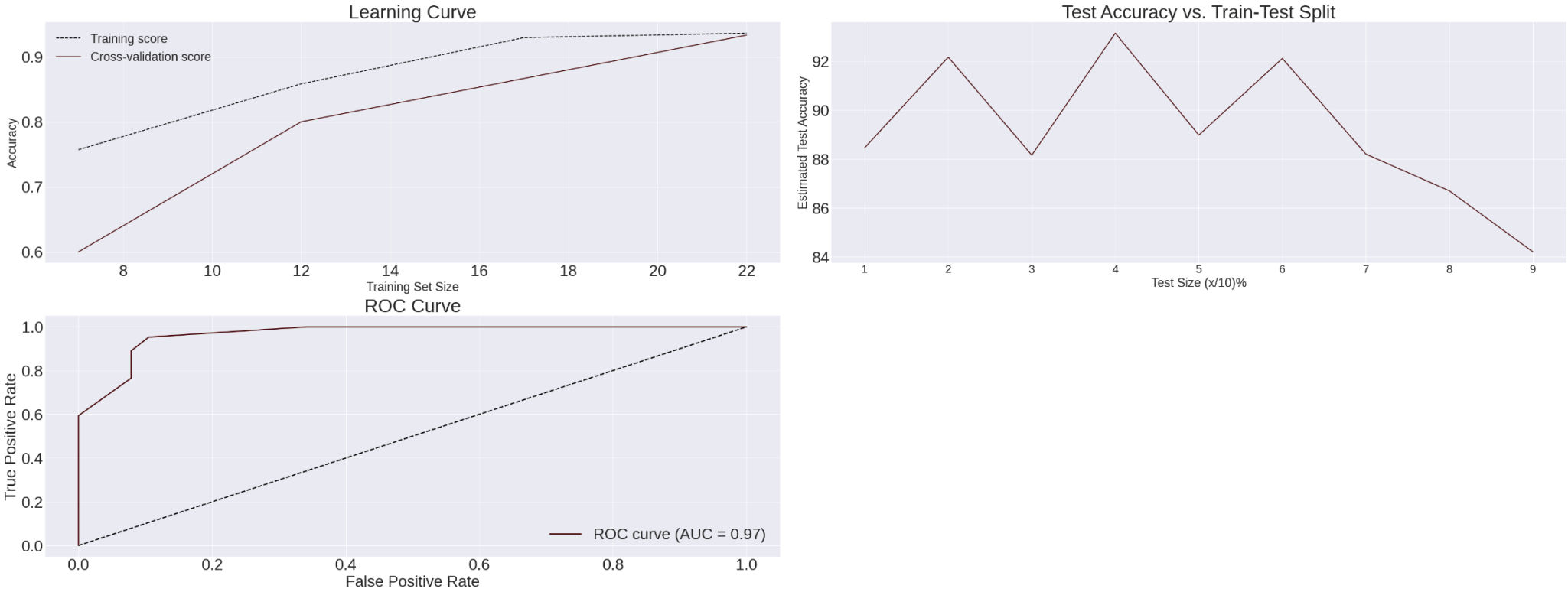

Figure B-5. K Nearest Neighbors Results- RF Narrow Model. The learning curve indicates the model is quickly learning; however, estimated test accuracy is rather low across test sizes and the ROC curve displays higher false positive rates than other models. Of note, a features importance plot is not included for this method as feature importance is not provided in the 'KNeighborsClassifier()' function. The base model KNN provided similar results; however, the correlation-based model demonstrated significantly reduced performance.

Basic Classification Tree with Pruning

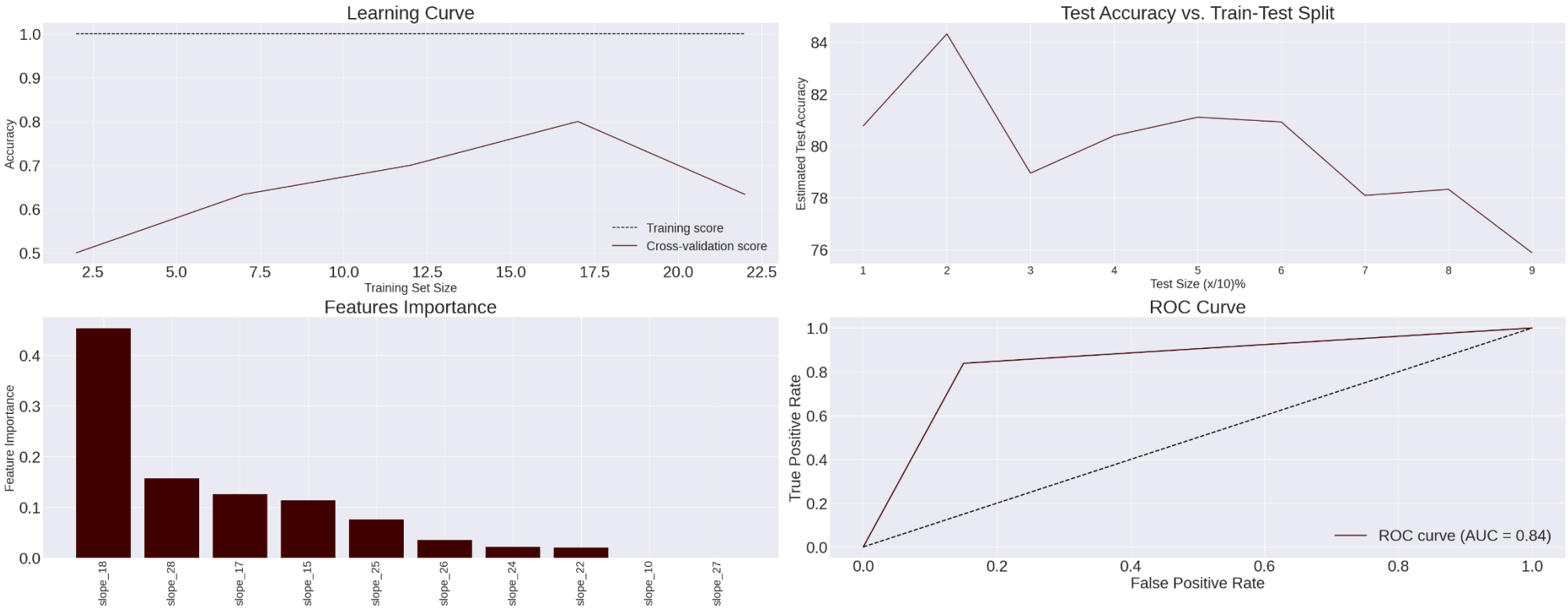

Figure B-6. Classification Tree and Pruning Results- Correlation Feature Reduction Model. The learning curve indicates moderate learning occurs to a point and the estimated test accuracy is generally inconsistent. The ROC Curve is the worst of the tested models; however, this model demonstrates strong indications of Slope 17's importance in predicting the occurrence of ovarian cancer in the spectra. Of note, this is the only model which truly experienced the best model under the correlation-based feature reduction method.