UNIVERSITY OF CALIFORNIA, SAN DIEGO

COGS 118A Supervised Machine Learning - Fall 2025

PROFESSOR:

Zhuowen Tu

Comparison of Classification Systems Amongst a Diverse Dataset

AUTHOR :

Joanne Chen

ABSTRACT

Throughout this course, we learned about a range of supervised learning models, from decision trees to ensemble approaches. In this project, we apply that knowledge by comparing three common classifiers, logistic regression, linear SVM, and random forest, on three UCI datasets. These datasets differ in size, feature types, topic, and overall difficulty. Rather than trying to perform predictive analysis on any one dataset, we use them to see how modeling choices, especially train test partitioning, affect results across classifiers and to identify which model performs best based on accuracy. Following the course experiment format, we evaluate three train test splits (20/80, 50/50, 80/20), repeat each setting three times, and use cross-validation to choose hyperparameters before reporting training, validation, and test performance. Overall, all models performed almost perfectly on Mushroom, performed moderately on Students, and found Bank Marketing to be the most challenging dataset, especially when comparing F1 score rather than accuracy.

TABLE OF CONTENTS

1 INTRODUCTION

In machine learning, different models learn patterns in different ways, so a model that performs well on one dataset may not perform as well on another. Because of this, it is useful to compare the same set of models across multiple datasets, which is also the general idea behind the large comparison study by Caruana and Niculescu-Mizil. In this project, we follow that approach by comparing logistic regression, linear SVM, and random forest across three diverse datasets, while keeping the experiment setup consistent.

1.1 Dataset Descriptions

Mushroom is a classification task that predicts whether a mushroom is edible or poisonous based on physical characteristics. It contains 8,124 instances and 22 features, and most of the features are categorical. This dataset was the easiest because the classes were easily separable. The Students' Dropout and Academic Success dataset uses student background and academic information to predict dropout versus non-dropout. It contains 4,424 instances and 36 features. Labels were combined to create a two class setting for this project. This dataset was medium difficulty. Bank Marketing is a larger business dataset based on phone marketing campaigns, where the goal is to predict whether a client subscribes to a term deposit. It contains 45,211 instances and 16 features, with a mix of categorical and numeric variables. This dataset was the hardest because the positive class is relatively rare.

Table 1: Summary of Dataset Features

	Problem	#Attr	Train Size	Test Size	%Poz
0	Mushroom	22/117	1624	6500	48.2%
1	Mushroom	22/117	4062	4062	48.2%
2	Mushroom	22/117	6499	1625	48.2%
3	Students	36	884	3540	32.1%
4	Students	36	2212	2212	32.1%
5	Students	36	3539	885	32.1%
6	BankMarketing	16/51	904	3617	11.5%
7	BankMarketing	16/51	2260	2261	11.5%
8	BankMarketing	16/51	3616	905	11.5%

2 METHOLOGY

2.1 Supervised Learning Methods

Logistic regression and linear SVM are widely used classifiers that are often chosen for large comparison studies because they are straightforward to train and can give solid performance, although how well they rank can depend a lot on the dataset and on which metric is used. In the Caruana and Niculescu-Mizil study, these linear models are included as standard reference methods across many datasets, while tree-based ensemble models such as random forests are usually the most consistently strong performers overall and typically do better than a single decision tree.

2.2 Experimental Procedure

To set a control for the experiments, we standardized the procedures for each dataset. After loading the data, we separated it into a feature matrix X, which contains the input information for each example, such as mushroom traits, and a label vector Y, which contains the outcomes we want to predict. All features were converted to binary values. For Mushrooms, Y represents whether a mushroom is poisonous. For the Students dataset, we treated “Dropout” as the positive class and combined the remaining categories into a single negative class. For Bank Marketing, Y comes from whether the campaign outcome is “yes” or “no.” As for preprocessing, categorical features in X were converted into numeric columns using one hot encoding, and numeric features were standardized.

For evaluation, each dataset was tested using three train test splits, 20/80, 50/50, and 80/20, to see how results change as more training data is available. Each split was repeated three times with different random seeds to reduce the impact of any single lucky split. Within each trial, we used cross validation on the training set to choose hyperparameters, such as the regularization strength for logistic regression and linear SVM and tree settings for random forest. We then fit the best model on the full training set and reported training, cross validated validation, and held out test performance. Along with accuracy, we reported F1 score to better capture performance on the positive class, especially for Bank Marketing where positive outcomes are relatively rare.

3 EXPERIMENTAL RESULTS

The final results tables show that performance depends strongly on which dataset we are working with. On the Mushroom dataset, all three classifiers produced almost perfect accuracy and F1 scores across every train test split. Though these results do not do a great job of highlighting differences between models because even the simpler approaches already perform extremely well.

For the Students dataset, the results were more complicated, but still fairly consistent across models. Test accuracy stayed around the middle range of around 0.86 and test F1 was generally in the 0.78 to about 0.80 range depending on the split. The models did differ slightly, but the differences were small compared to the overall gap between this dataset and the others. This suggests that there is a clear pattern to learn, but not one that allows models to reach near perfect performance.

The Bank Marketing dataset showed the most noticeable difference between accuracy and F1. Random forest achieved the highest accuracy, close to 0.90, while logistic regression and linear SVM were closer to about 0.84 to 0.85. However, when we looked at F1 score, the linear models performed better, with test F1 values around 0.55, while random forest's F1 was lower. This likely happens because the positive "yes" outcome is rare in this dataset. A model can get high accuracy by predicting the more common "no" class most of the time, but that does not necessarily mean it is good at finding the smaller positive class. Including F1 score is important, since it gives a clearer picture of how well each model handles the minority outcome.

Table 2: Full Experimental Results on Each Dataset

	Dataset	Classifier	Train Size	Random State	Train Acc.	Val Acc.	Test Acc.	Best Param.	Train F1	Test F1	Trial
0	Mushroom	LogReg	0.2	100	1.000000	0.998769	1.000000	{'clf__C': 10.0, 'clf__solver': 'lbfgs'}	1.000000	1.000000	1
1	Mushroom	LinearSVM	0.2	100	1.000000	0.998769	1.000000	{'clf__C': 1.0}	1.000000	1.000000	1
2	Mushroom	RandomForest	0.2	100	1.000000	0.998769	1.000000	{'clf__n_estimators': 200, 'clf__max_depth': N...	1.000000	1.000000	1
3	Mushroom	LogReg	0.2	101	1.000000	0.998150	0.999077	{'clf__C': 1.0, 'clf__solver': 'lbfgs'}	1.000000	0.999042	2
4	Mushroom	LinearSVM	0.2	101	1.000000	0.998150	0.999385	{'clf__C': 0.1}	1.000000	0.999361	2
...	...	...	...	...	...	...	...	...	...	...	...
76	BankMarketing	LinearSVM	0.8	101	0.851139	0.850448	0.843194	{'clf__C': 10.0}	0.560346	0.551266	2
77	BankMarketing	RandomForest	0.8	101	1.000000	0.903451	0.902024	{'clf__n_estimators': 400, 'clf__max_depth': N...	1.000000	0.451733	2
78	BankMarketing	LogReg	0.8	102	0.845637	0.845029	0.841093	{'clf__C': 1.0, 'clf__solver': 'lbfgs'}	0.554180	0.549106	3
79	BankMarketing	LinearSVM	0.8	102	0.850448	0.849840	0.846511	{'clf__C': 1.0}	0.559205	0.554557	3
80	BankMarketing	RandomForest	0.8	102	1.000000	0.903064	0.901802	{'clf__n_estimators': 400, 'clf__max_depth': N...	1.000000	0.446384	3

Table 3: Summarized Testing Accuracy + F1 Scores for Training & Testing

20/80 split:

		Mushroom-acc	Mushroom-testf1	Mushroom-trainf1	Students-acc	Students-testf1	Students-trainf1	Bank-acc	Bank-testf1	Bank-trainf1
0	LOGREG	0.999	0.999	**1.000**	0.859	0.785	0.824	0.843	0.550	0.548
1	RANFOR	**1.000**	**1.000**	**1.000**	0.860	0.760	**0.987**	**0.899**	0.370	**1.000**
2	SVM	0.999	0.999	**1.000**	**0.861**	**0.786**	0.824	0.849	**0.555**	0.552

50/50 split:

		Mushroom-acc	Mushroom-testf1	Mushroom-trainf1	Students-acc	Students-testf1	Students-trainf1	Bank-acc	Bank-testf1	Bank-trainf1
0	LOGREG	**1.000**	**1.000**	**1.000**	0.861	0.788	0.801	0.844	0.551	0.554
1	RANFOR	**1.000**	**1.000**	**1.000**	**0.868**	0.774	**0.984**	**0.901**	0.424	**1.000**
2	SVM	**1.000**	**1.000**	**1.000**	0.865	**0.793**	0.808	0.850	**0.557**	0.560

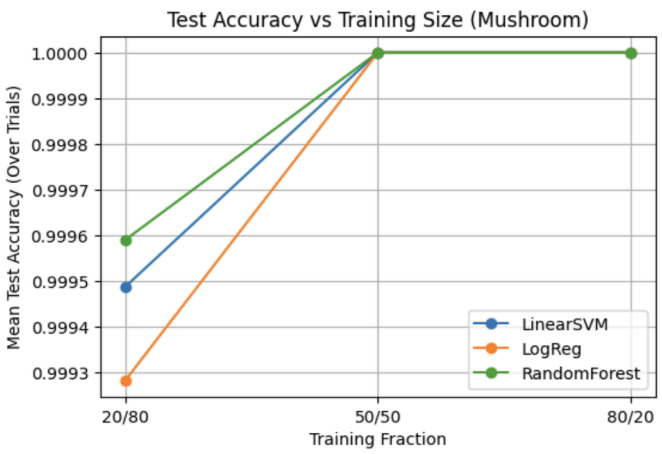
80/20 split:

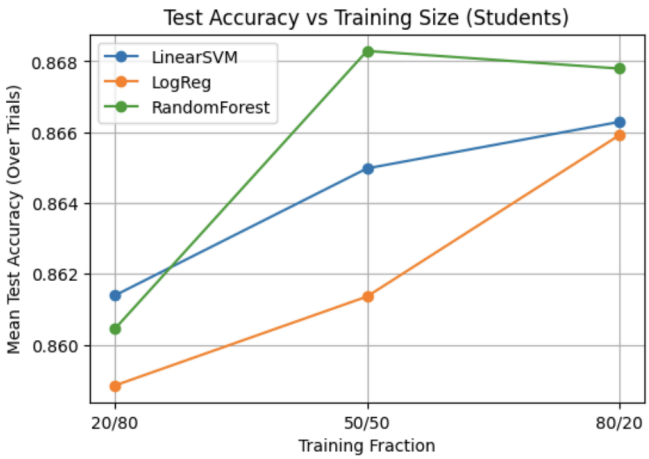
		Mushroom-acc	Mushroom-testf1	Mushroom-trainf1	Students-acc	Students-testf1	Students-trainf1	Bank-acc	Bank-testf1	Bank-trainf1
0	LOGREG	**1.000**	**1.000**	**1.000**	0.866	**0.795**	0.805	0.843	0.552	0.553
1	RANFOR	**1.000**	**1.000**	**1.000**	**0.868**	0.774	**0.976**	**0.902**	0.448	**1.000**
2	SVM	**1.000**	**1.000**	**1.000**	0.866	0.794	0.804	0.849	**0.558**	0.559

4 PLOT INTERPRETATION

4.1 Test Accuracy vs Training Size Plot

The training size versus test accuracy plot helps us check whether the models improve when they are given more training data. In general, we expect test accuracy to go up as we move from a smaller training split to a larger one, since the model has more examples to learn from. In our results, this trend is easiest to see for the harder datasets, especially Bank Marketing and Students. For Mushroom, the accuracy is already almost perfect even with only 20 percent training data, so the line stays mostly flat.





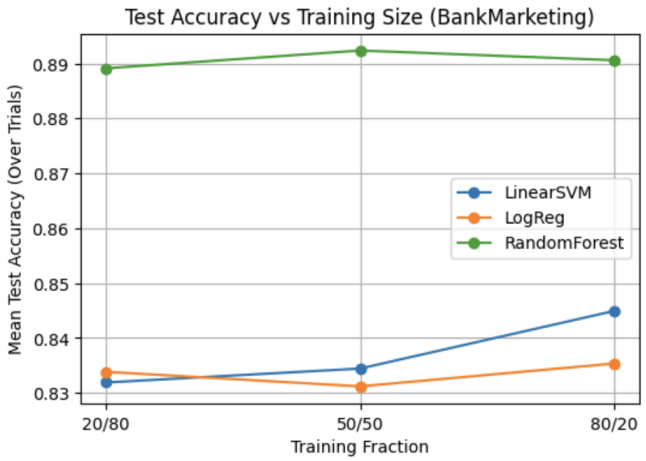
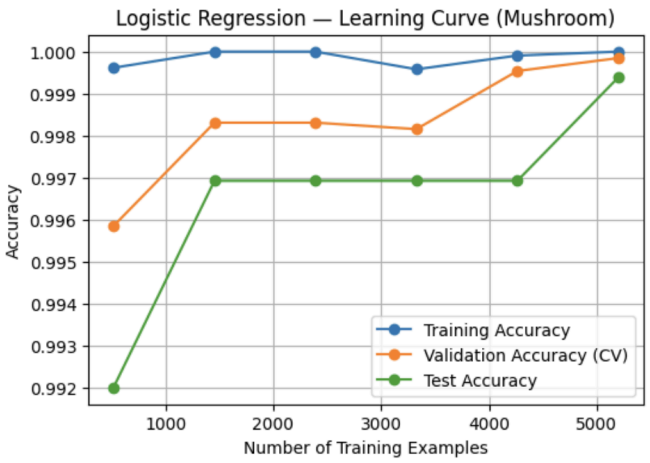
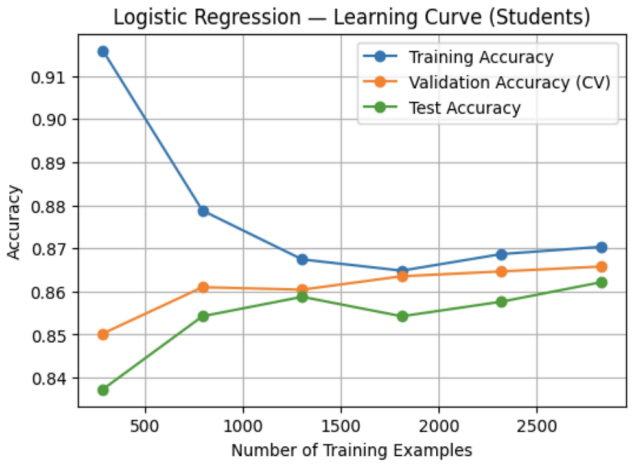


Figure 1-3: Testing Accuracy vs Training Size Plots for Each Dataset

4.2 Learning Curves

The learning curve plot gives a clearer picture of how a model is learning. It compares performance on the training data to performance on validation data as the training set grows. If the training accuracy stays much higher than the validation accuracy, it suggests the model may be fitting the training data too closely and not performing as well on new data. If the two curves are closer together, it suggests the model is learning better patterns. For Students and Bank Marketing, the learning curves explain why they don't show near perfect performances, since these datasets are likely harder and have more overlap between the two classes.





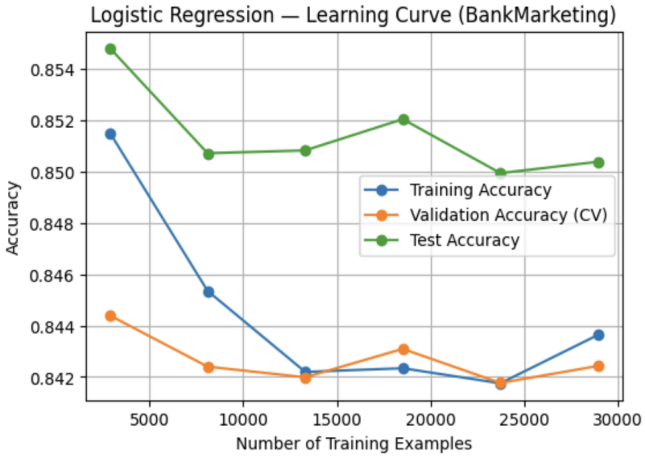
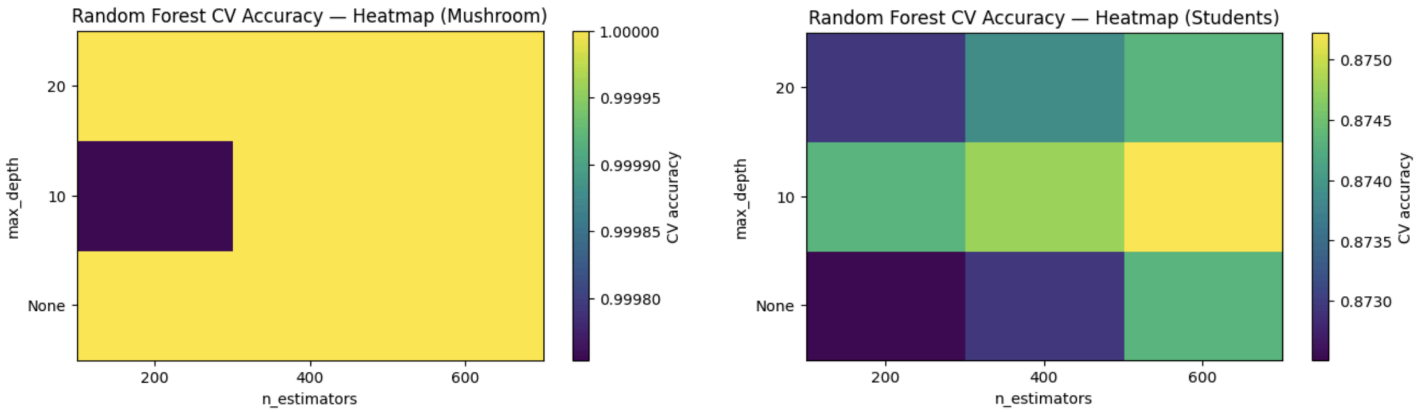


Figure 4-6: Learning Curve Plots for Each Dataset

4.3 Heatmaps

The hyperparameter heatmaps help show how much the results depend on the model settings we choose. If a large area of the heatmap shows similar performance, it means the model is not very sensitive and many settings work about equally well. If only a small region performs well, it means the model depends more on choosing the right settings. In our experiments, these heatmaps reinforce why cross validation is useful, because they show that different hyperparameter choices can noticeably change validation performance.



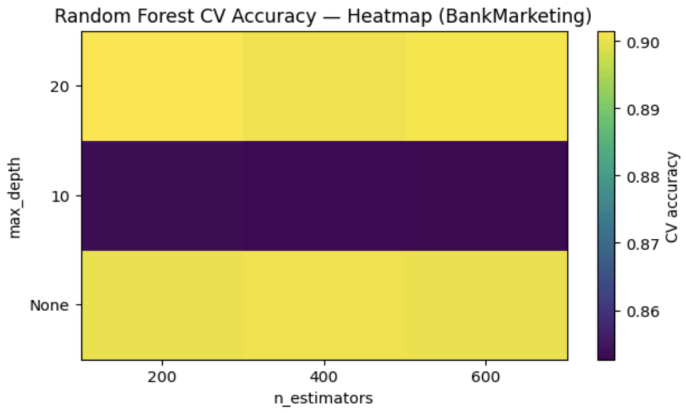


Figure 7-9: Heatmap Plots for Each Dataset

5 ERROR ANALYSIS

There are two main reasons our results could be inconsistent: possible issues in the code and limitations of the dataset choices. Even if the code runs, small undetected problems could still affect performance, such as mistakes in how the positive class was defined when creating labels or computing F1. Differences in settings like random seeds, class weighting, or how thoroughly each model's hyperparameters were tuned could also unintentionally favor one classifier. In addition, the datasets themselves may not be ideal for comparing models.

6 CONCLUSION

Overall, the Mushroom dataset was very easy for all three classifiers, so there were almost no meaningful differences between them. The Students dataset was more challenging, but the three models ended up with fairly similar accuracy and F1 scores, with only small gaps between them. Bank Marketing was the most difficult dataset and it clearly showed why accuracy by itself can be misleading. Although random forest had the highest accuracy, logistic regression and linear SVM achieved better F1 scores, suggesting they were more effective at catching the smaller positive class when the data is imbalanced. Finally, using multiple train test splits, repeating trials, and using cross validation helped make the results more dependable and kept the evaluation consistent with the course guidelines.

7 REFERENCES

[1] Caruana, R. *An Empirical Comparison of Supervised Learning Algorithms*.

[2] Krishnan, A., Chen, J., & et. al. *COGS 108 - The Influence of Race on Readmission Rates.*https://github.com/COGS108/Group070-FA24/blob/master/FinalProject_Group070-FA24.ipynb.

[3] Freund, Y., & Schapire, R. E. “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.” *Journal of Computer and System Sciences*, vol. 55, no. 1, Aug. 1997, pp. 119–139, https://doi.org/10.1006/jcss.1997.1504.

[4] Freund, Y., & Schapire, R. E. “Improved Boosting Algorithms Using Confidence-Rated Predictions.” *Machine Learning*, vol. 37, no. 3, Dec. 1999, pp. 297–336, https://doi.org/10.1023/a:1007614523901.

[5] Wikipedia . “Logistic Regression.” *Wikipedia*, Wikimedia Foundation, 12 Apr. 2019, en.wikipedia.org/wiki/Logistic_regression.