CSCI 6316 Final Applications Project

Emma Glass and Gabe Hanson

Dataset

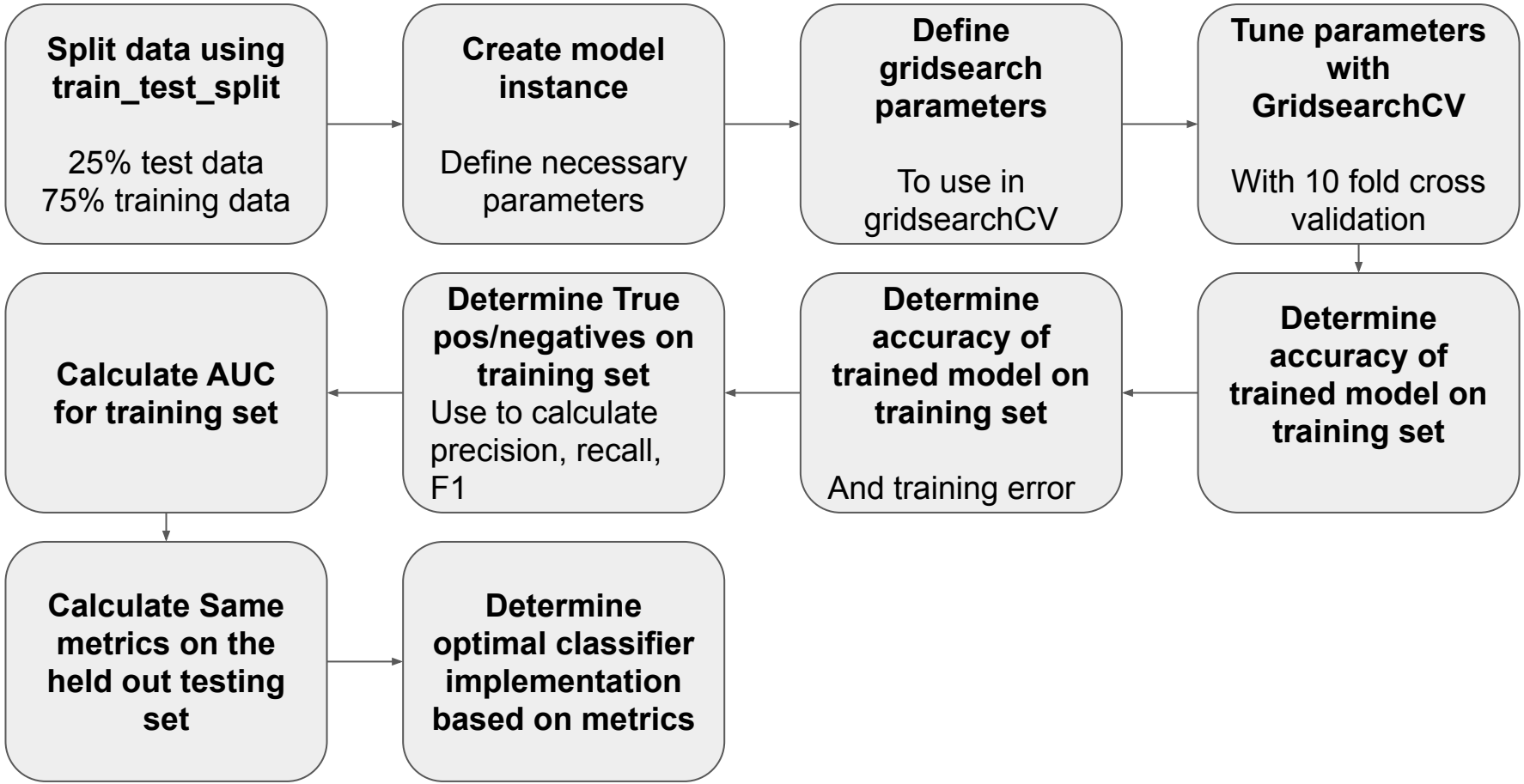
Dataset 1: 569 observations, 31 features

Dataset 2: 462 observations, 10 features

Pre-processing:

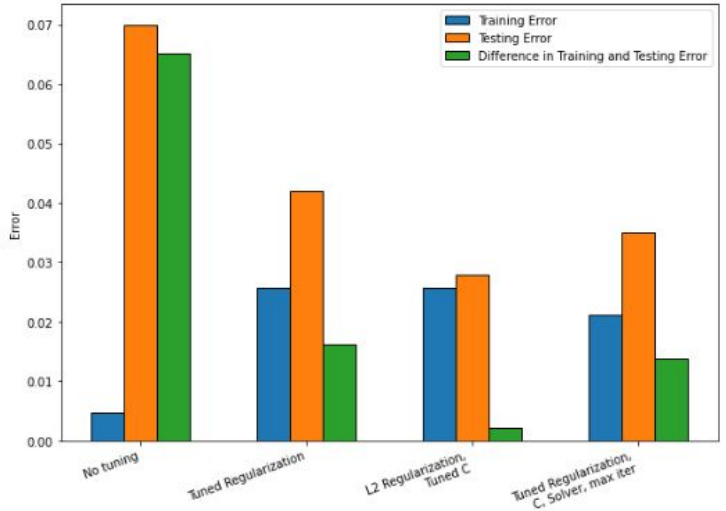
1. Feature columns extracted
    a. Data normalized with min-max normalization
2. Label column extracted

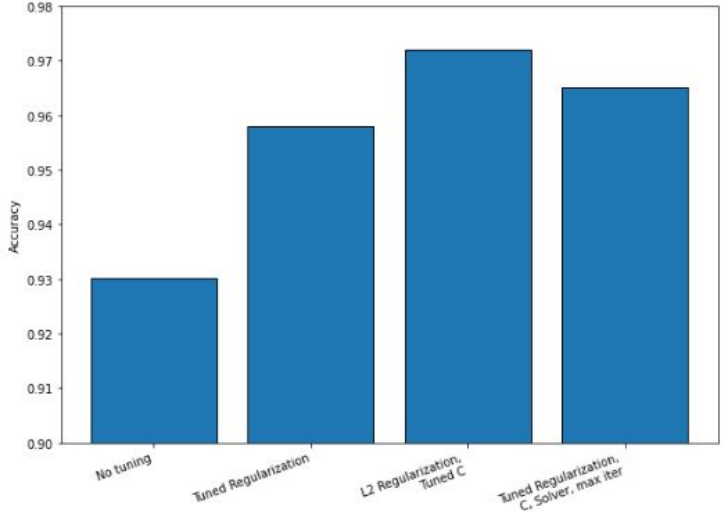
General Workflow



Logistic Regression - **Dataset 1**

- Trained 4 classifiers with different tuned parameters sets
- Best classifier: <u>L2 regularization with Tuned C parameter</u>
  - Highest accuracy (**97.2%**), lowest testing error, lowest bias, highest variance





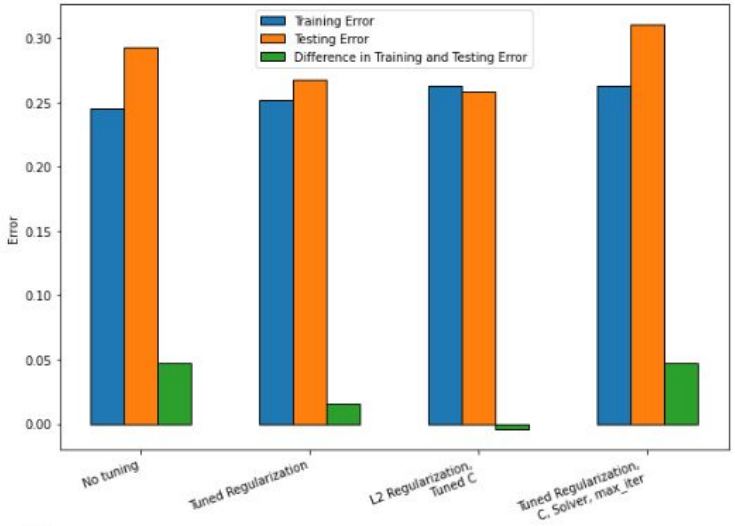
Best Logisitic Regression Dataset 1 Classifier Metrics

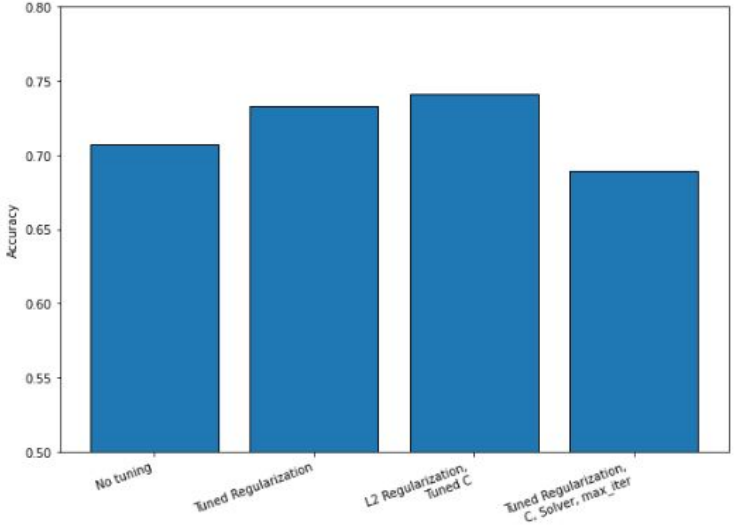
	Accuracy	Precision	Recall	F1 Measure	Training/Testing Error	AUC
Results of 10-fold Cross Validation on Training Set	0.974	0.994	0.939	0.965	0.0258	0.967
Results on held out test set	0.972	0.978	0.9358	0.958	0.0280	0.964

Best C Parameter: 1.481

Logistic Regression - **Dataset 2**

- Trained 4 classifiers with different tuned parameters sets
- Best classifier: <u>L2 regularization with Tuned C parameter</u>
  - Highest accuracy (**74.1%**), lowest testing error, lowest bias, highest variance



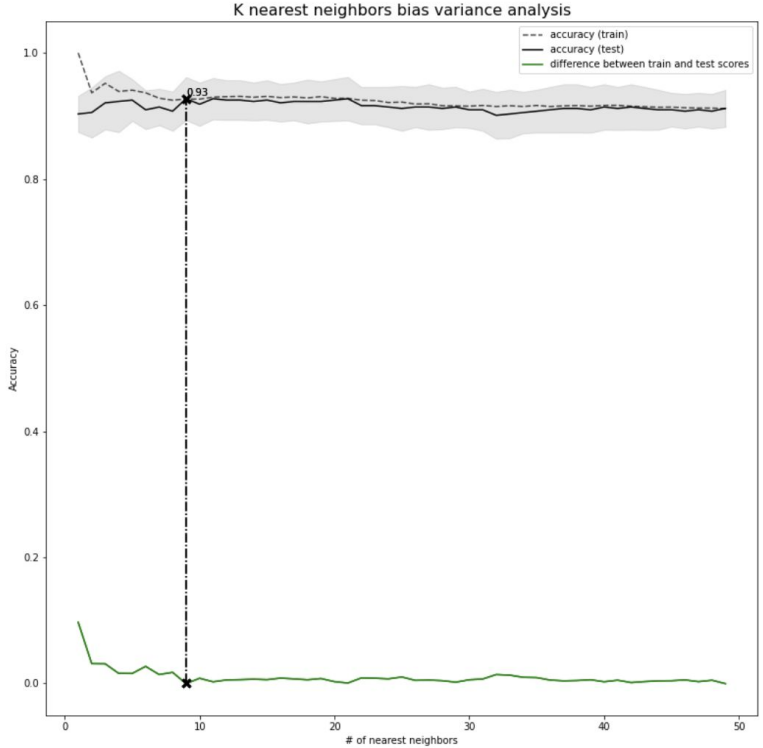


Best Logisitic Regression Dataset 2Classifier Metrics

	Accuracy	Precision	Recall	F1 Measure	Training/Testing Error	AUC
Results of 10-fold Cross Validation on Training Set	0.737	0.692	0.388	0.497	0.263	0.650
Results on held out test set	0.741	0.75	0.477	0.583	0.259	0.690

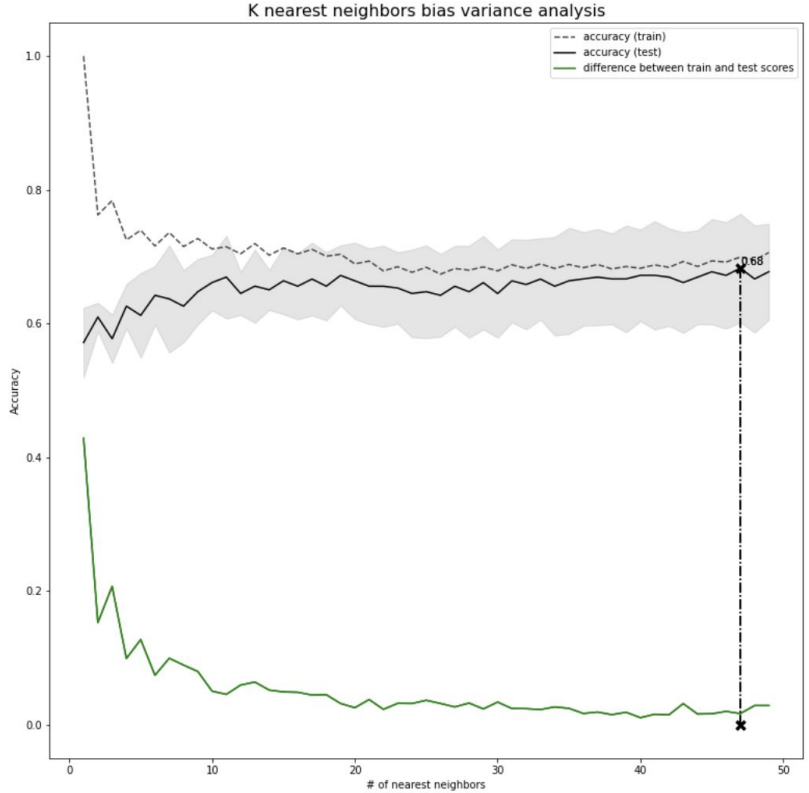
Best C Parameter: 0.251

K Nearest Neighbors - **Dataset 1**



	Accuracy	Precision	Recall	F1	AUC
Mean score of 10 fold CV on training set	0.92744	0.945313	0.854779	0.894091	0.967267
Test Scores	0.947368	1	0.869565	0.930233	0.934783

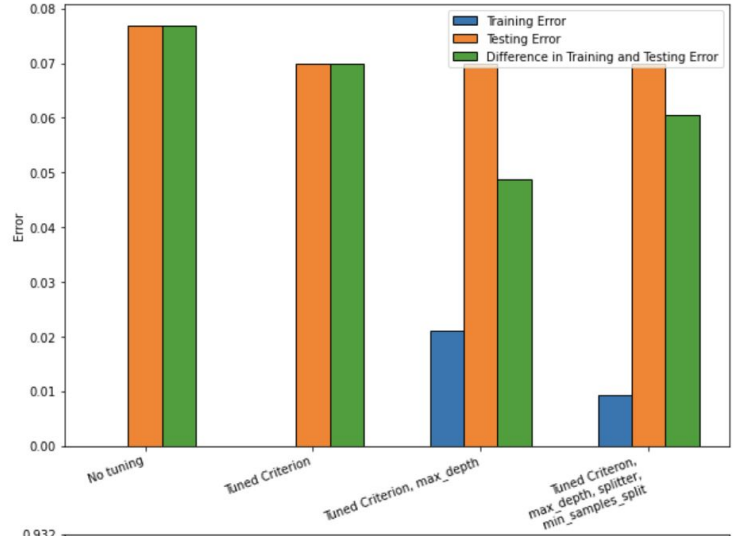
K Nearest Neighbors - **Dataset 2**

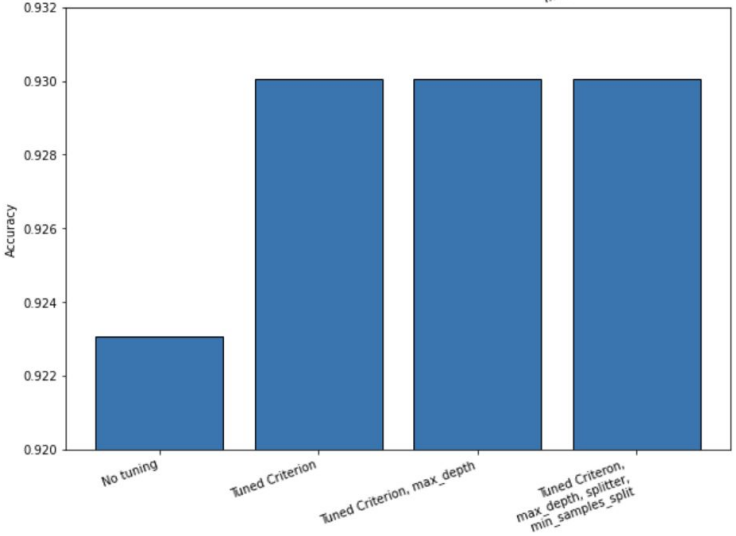


	Accuracy	Precision	Recall	F1	AUC
Mean score of 10 fold CV on training set	0.682883	0.64246	0.315385	0.410354	0.728511
Test Scores	0.688172	0.411765	0.269231	0.325581	0.559989

Decision Tree - **Dataset 1**

- Trained 4 classifiers with different tuned parameters sets
- Best classifier: <u>tuned citerion/max_depth</u>
  - Highest accuracy (**93.0%**), lowest testing error, lowest bias, highest variance





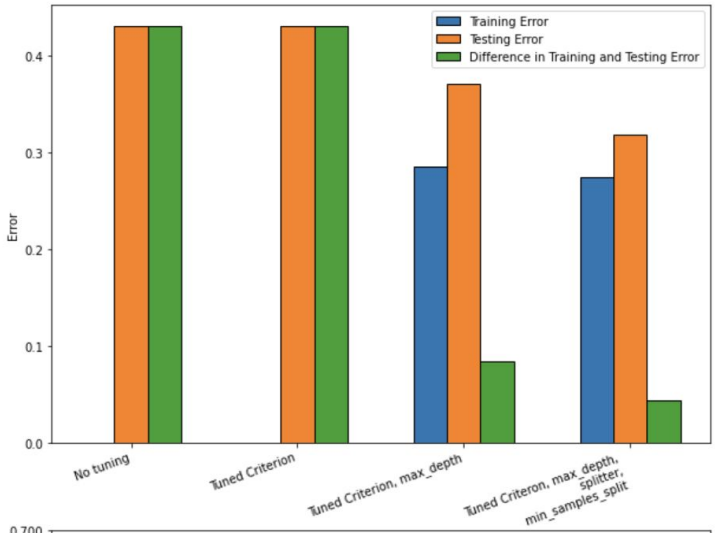
Best Decision Tree Dataset 1 Classifier Metrics

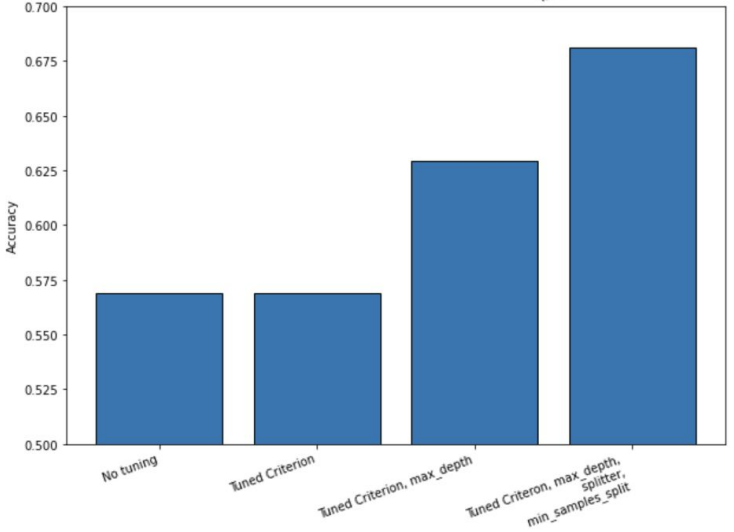
	Accuracy	Precision	Recall	F1 Measure	Training/Testing Error	AUC
Results of 10-fold Cross Validation on Training Set	0.979	1.0	0.944	0.972	0.0211	0.972
Results on held out test set	0.930	0.882	0.912	0.9	0.0699	0.927

Best Criterion: entropy **Best max_depth:** 4

Decision Tree - **Dataset 2**

- Trained 4 classifiers with different tuned parameters sets
- Best classifier: <u>tuned citerion, max_depth, splitter, min_samples_split</u>
    - Highest accuracy (**68.1%**), lowest testing error, lowest bias, highest variance





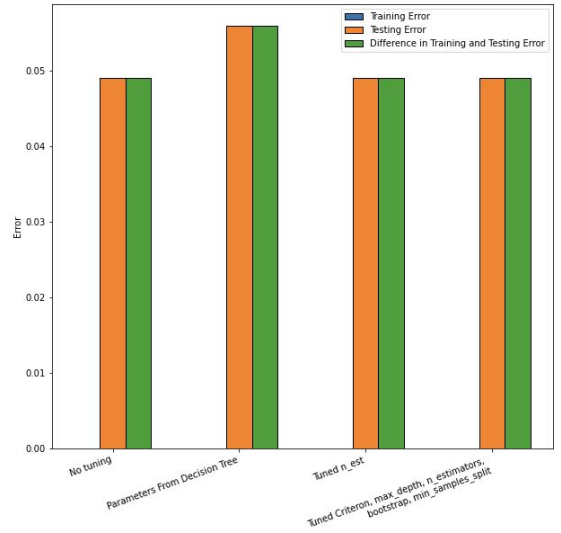
Best Decision Tree Dataset 2 Classifier Metrics

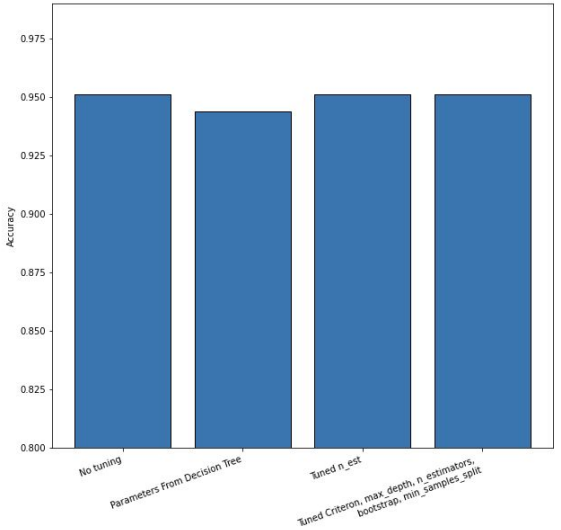
	Accuracy	Precision	Recall	F1 Measure	Training/Testing Error	AUC
Results of 10-fold Cross Validation on Training Set	0.725	0.590	0.595	0.592	0.274	0.693
Results on held out test set	0.681	0.600	0.477	0.532	0.319	0.641

Best criterion: entropy **Best max_depth:** 5 **Best splitter:** random **Best min_samples_split:** 2

Random Forest - **Dataset 1**

- Trained 4 classifiers with different tuned parameters sets
- Best classifier: three way tie between default parameters, tuned n_estimator parameter, or tuned criterion, max_depth, n_estimators, bootstrap, min_samples_split
  - All reported the same metrics. Highest accuracy (**95.1%**), lowest testing error, lowest bias, highest variance





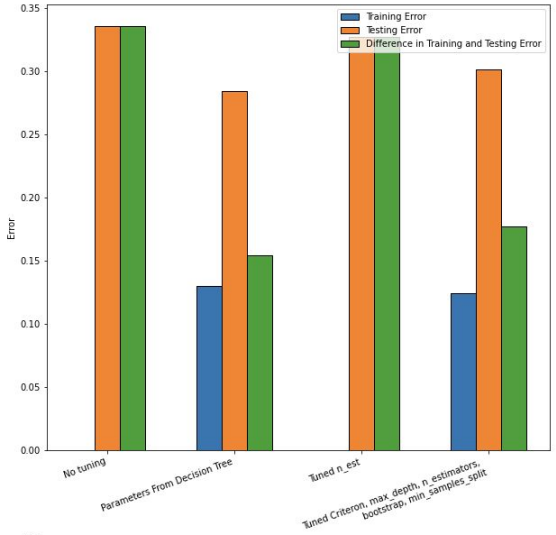
Example of Best Random Forest Dataset 1 Classifier Metrics (since there was a 3-way tie)

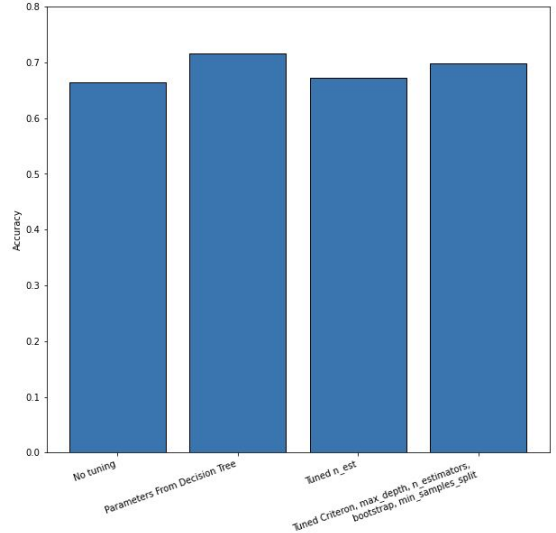
	Accuracy	Precision	Recall	F1 Measure	Training/Testing Error	AUC
Results of 10-fold Cross Validation on Training Set	1.0	1.0	1.0	1.0	0.0	1.0
Results on held out test set	0.951	0.920	0.939	0.929	0.0490	0.948

Best n_estimators: 50 **Best criterion:** gini **Best max_depth:** 10 **Best bootstrap:** False
Best min_samples_split: 5

Random Forest - **Dataset 2**

- Trained 4 classifiers with different tuned parameters sets
- Best classifier: parameters from decision tree (criterion=entropy, max_depth = 5, min_samples_split = 2)
  - Highest accuracy (**71.6%**), lowest testing error, lowest bias, highest variance



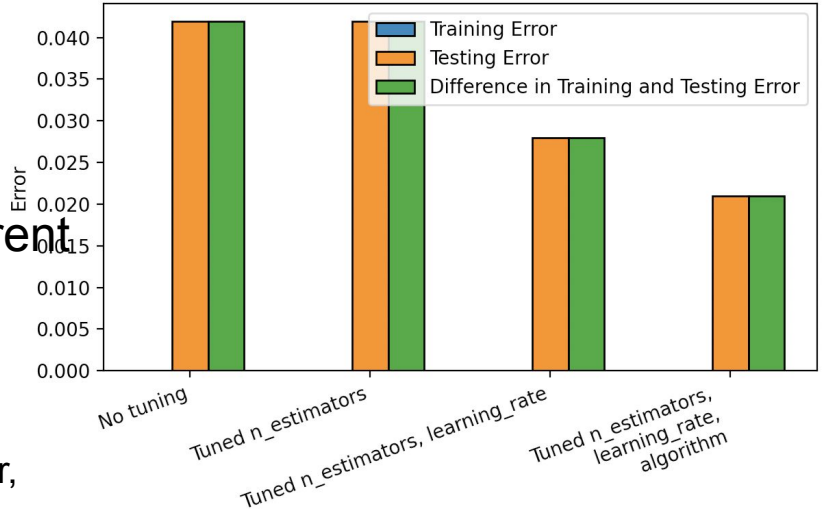


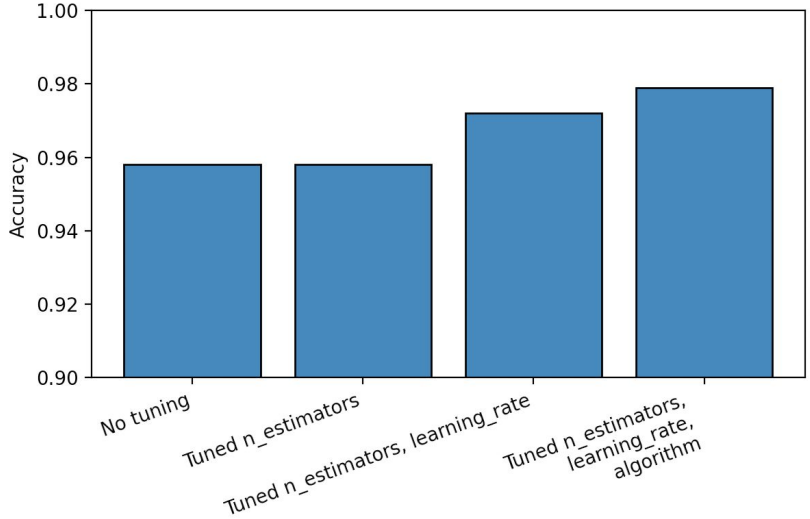
Example of Best Random Forest Dataset 2 Classifier Metrics

	Accuracy	Precision	Recall	F1 Measure	Training/Testing Error	AUC
Results of 10-fold Cross Validation on Training Set	0.870	0.938	0.655	0.772	0.130	0.817
Results on held out test set	0.716	0.690	0.455	0.548	0.284	0.665

Boosting - **Dataset 1**

- Trained 4 adaboost classifiers with different tuned parameters sets
- Best classifier: <u>tuned n_estimators, algorithm, and learning rate</u>
  - Highest accuracy (**97.9%**), lowest testing error, lowest bias, highest variance





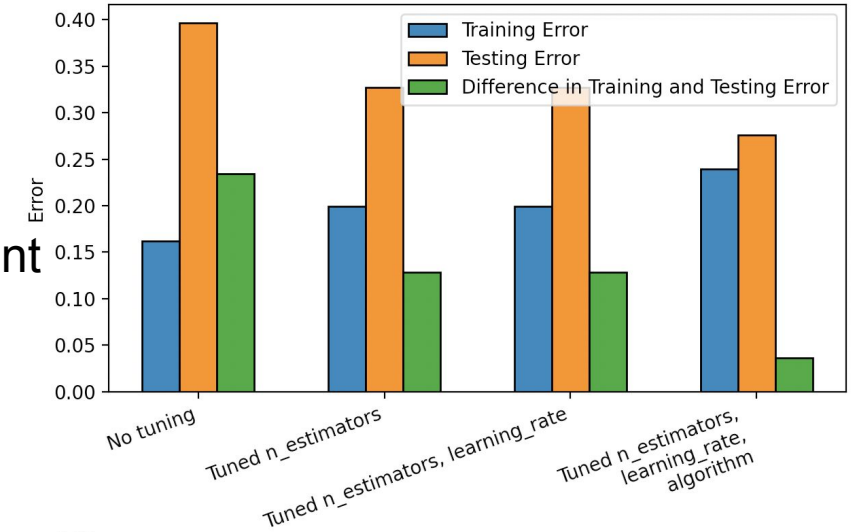
Best Adaboost Dataset 1 Classifier Metrics

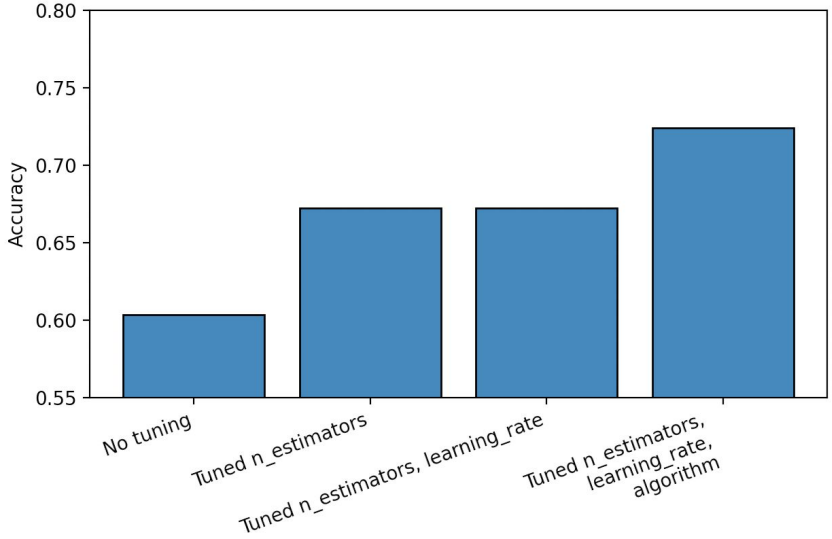
	Accuracy	Precision	Recall	F1 Measure	Training/Testing Error	AUC
Results of 10-fold Cross Validation on Training Set	1.0	1.0	1.0	1.0	0.0	1.0
Results on held out test set	0.979	0.96	0.980	0.970	0.0210	0.979

Best n_estimators: 150 **Beest learning_rate:** 1.2 **Best algorithm:** SAMME

Boosting - **Dataset 2**

- Trained 4 adaboost classifiers with different tuned parameters sets
- Best classifier: <u>tuned n_estimators, algorithm, and learning rate</u>
  - Highest accuracy (**72.4%**), lowest testing error, lowest bias, highest variance



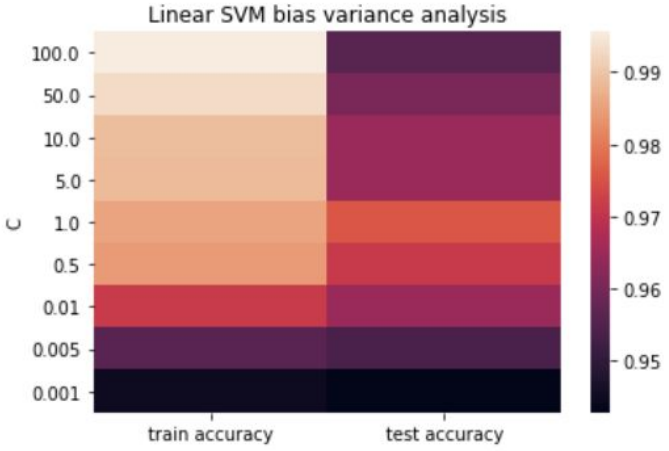


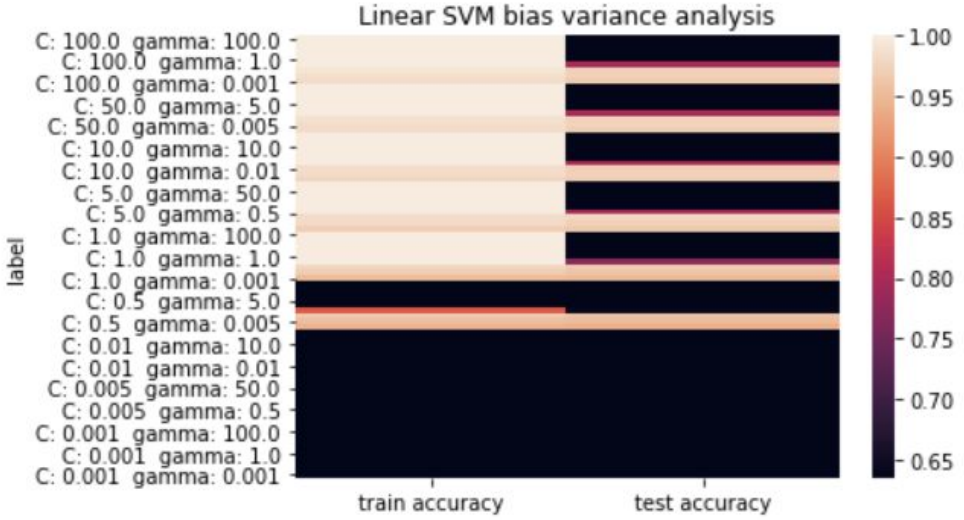
Best Adaboost Dataset 1 Classifier Metrics

	Accuracy	Precision	Recall	F1 Measure	Training/Testing Error	AUC
Results of 10-fold Cross Validation on Training Set	0.760	0.746	0.431	0.546	0.240	0.679
Results on held out test set	0.724	0.731	0.432	0.543	0.276	0.667

Best n_estimators:8 **Beest learning_rate:**1.5 **Best algorithm:** SAMME

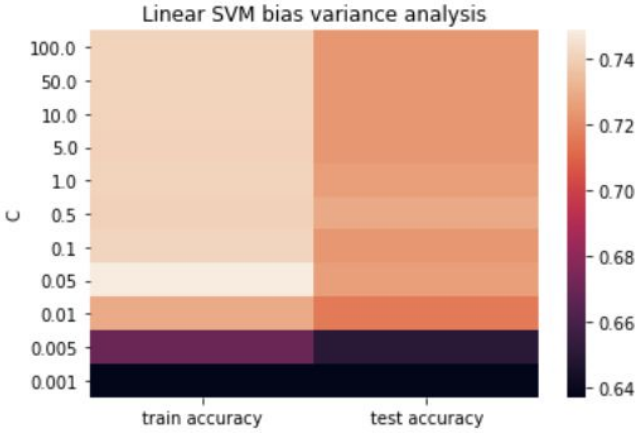
Support Vector Machines - **Dataset 1**

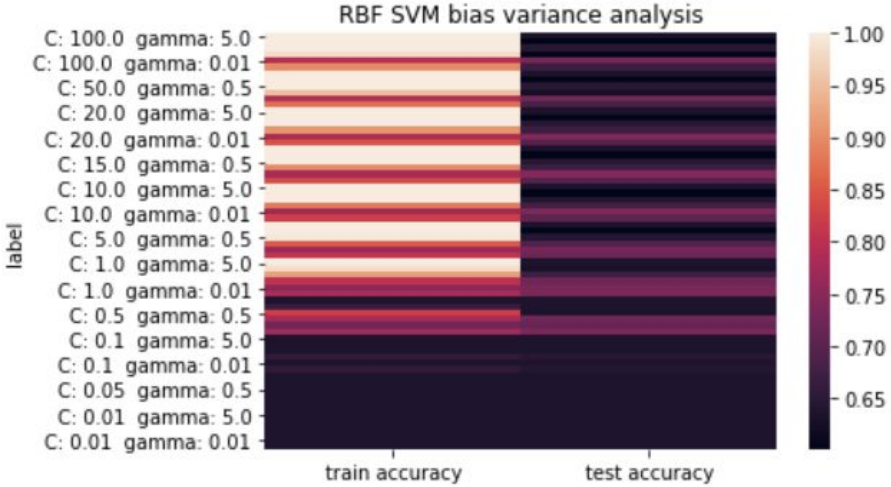




	Accuracy	Precision	Recall	F1	AUC
Mean score of 10 fold CV on training set	0.980145	0.988194	0.957353	0.971457	0.993645
Test Scores	0.973684	0.977778	0.956522	0.967033	0.970908

Support Vector Machines - **Dataset 2**





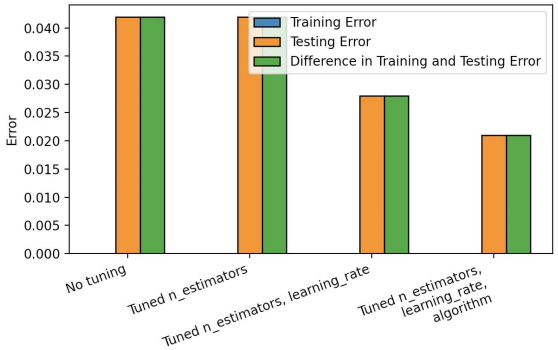
	Accuracy	Precision	Recall	F1	AUC
Mean score of 10 fold CV on training set	0.74527	0.692114	0.537912	0.600444	0.760844
Test Scores	0.731183	0.518519	0.538462	0.528302	0.672216

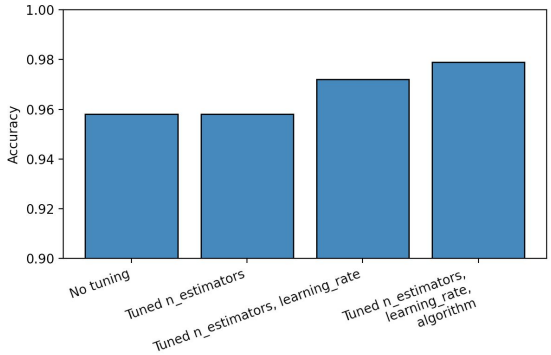
Best Overall Classifier - **Dataset 1**

The best overall classifier for dataset 1 was boosting with a test accuracy of 97.9% and a training accuracy of 100%

	Accuracy	Precision	Recall	F1 Measure	Training/Testing Error	AUC
Results of 10-fold Cross Validation on Training Set	1.0	1.0	1.0	1.0	0.0	1.0
Results on held out test set	0.979	0.96	0.980	0.970	0.0210	0.979

Best n_estimators:150 **Beest learning_rate:**1.2 **Best algorithm:** SAMME





Best Overall Classifier - **Dataset 2**

The best overall classifier for dataset 2 was an rbf kernel SVM with a testing accuracy of 73.11% and a training accuracy of 74.5%

	Accuracy	Precision	Recall	F1	AUC
Mean score of 10 fold CV on training set	0.74527	0.692114	0.537912	0.600444	0.760844
Test Scores	0.731183	0.518519	0.538462	0.528302	0.672216

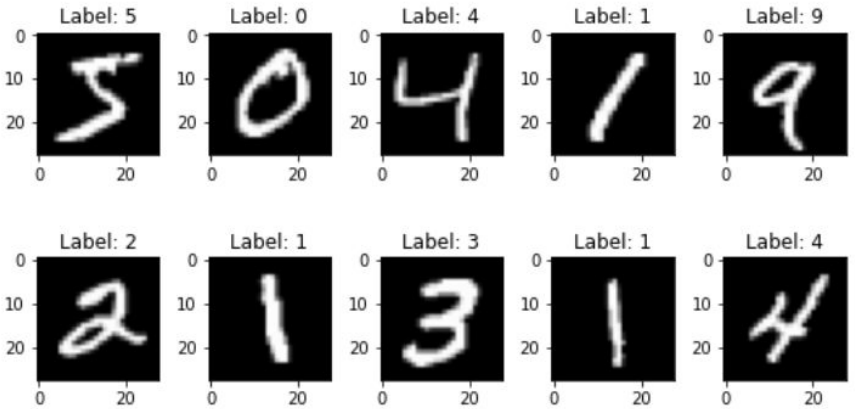
MNIST Dataset

MNIST (Modified National Institute of Standards and Technology) database

Images of handwritten numbers commonly used to train and benchmark various image processing methods.

- Each image is 28x28
- 50,000 training images
- 10,000 test images
- 10,000 validation images (not used in this analysis)

Goal: Train a deep learning model to classify the 10 handwritten digits



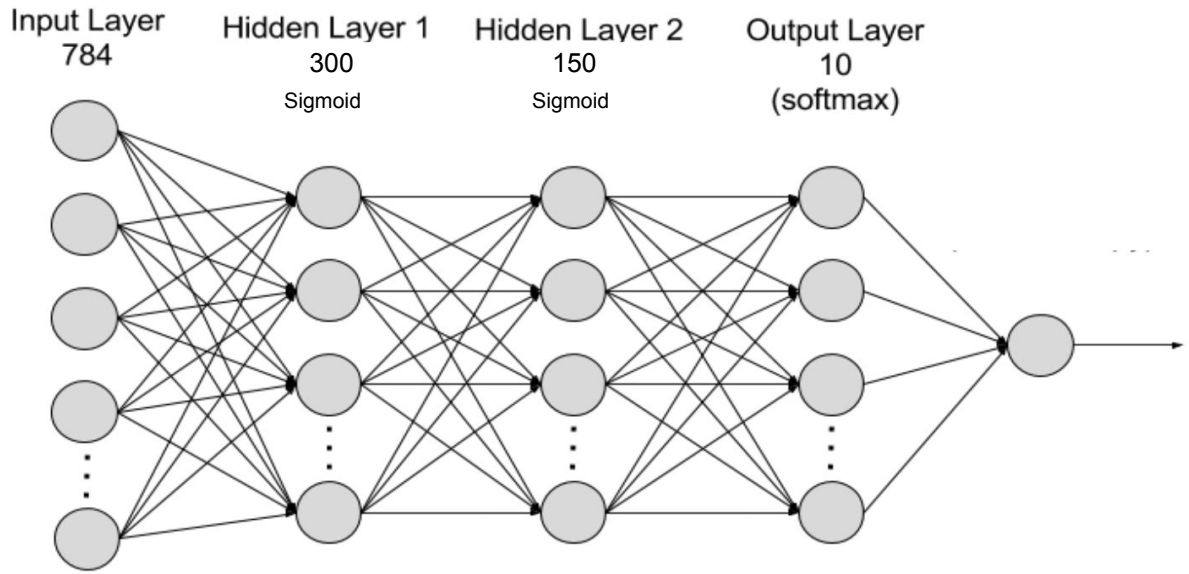
Deep Learning model - architecture

As its input, this model takes a flattened normalized image vector

It has 2 hidden layers using sigmoid activation functions

The number of hidden units were chosen after several rounds of parameter selection

It has a softmax output layer which generates a prediction.



Deep Learning model: Results

Hyperparameters tested:

- # of hidden units (first layer, second layer)
  (2,2), (10,10), (70,70), (150,150), (300,300), (300,150), (150,300)
- Learning rates
  1E-1, 1E-2, 5E-3, 1E-3, 5E-4, 1E-4, 1E-5, 1E-6, 5E-7
- Initialization weights:
  Random uniform, random normal, glorot uniform, glorot normal, truncated normal, ones, zeros

Parameters chosen for final model:

hidden units = (300,150)

Learning rate = 1E-3

Initialization weights = Glorot Uniform

For all models i trained

```
Model: "sequential_354"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 hidden_layer_1 (Dense)      (None, 300)               235500

 hidden_layer_2 (Dense)      (None, 150)               45150

 predictions (Dense)         (None, 10)                1510

=================================================================
Total params: 282,160
Trainable params: 282,160
Non-trainable params: 0
_________________________________________________________________
```

```
Fit model on training data
Epoch 1/2
5000/5000 [==============================] - 12s 2ms/step - loss: 0.3135 -
accuracy: 0.9064 - mean_squared_error: 0.0140 - auc_348: 0.9931 - val_loss:
0.1480 - val_accuracy: 0.9573 - val_mean_squared_error: 0.0067 - val_auc_34
8: 0.9974
Epoch 2/2
5000/5000 [==============================] - 11s 2ms/step - loss: 0.1223 -
accuracy: 0.9627 - mean_squared_error: 0.0057 - auc_348: 0.9981 - val_loss:
0.1021 - val_accuracy: 0.9679 - val_mean_squared_error: 0.0049 - val_auc_34
8: 0.9986
```

```
: results_final = model.evaluate(X_test, y_test)

313/313 [==============================] - 1s 1ms/step - loss: 0.1021 - accuracy: 0.9679 -
mean_squared_error: 0.0049 - auc_348: 0.9986
```