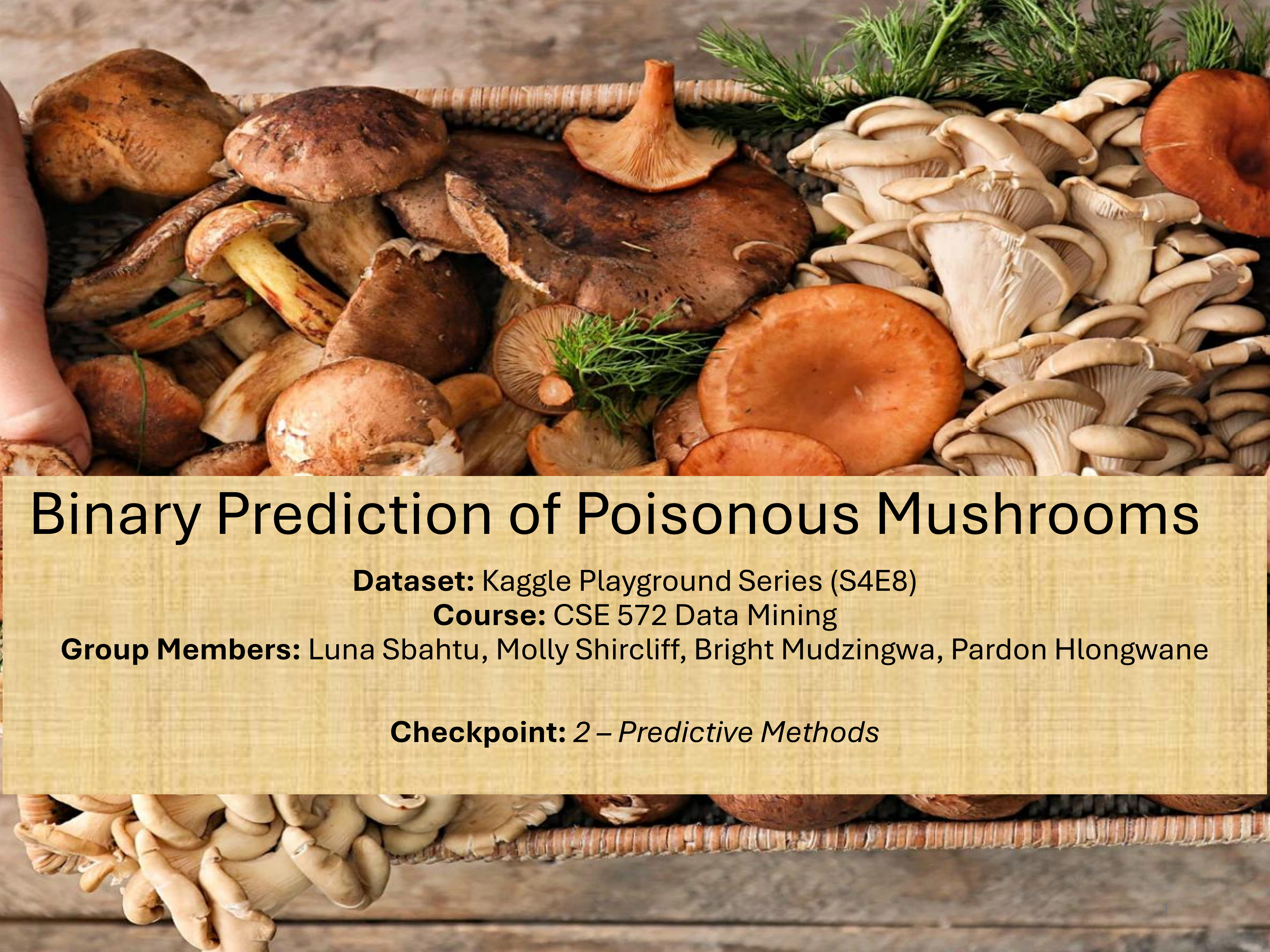


Binary Prediction of Poisonous Mushrooms

Dataset: Kaggle Playground Series (S4E8)
Course: CSE 572 Data Mining
Group Members: Luna Sbahtu, Molly Shircliff, Bright Mudzingwa, Pardon Hlongwane

Checkpoint: *2 – Predictive Methods*

Checkpoint 2: Overview

- Predictive Methods
- Results & Issues
- Improvements Proposed & Results
- Future Steps

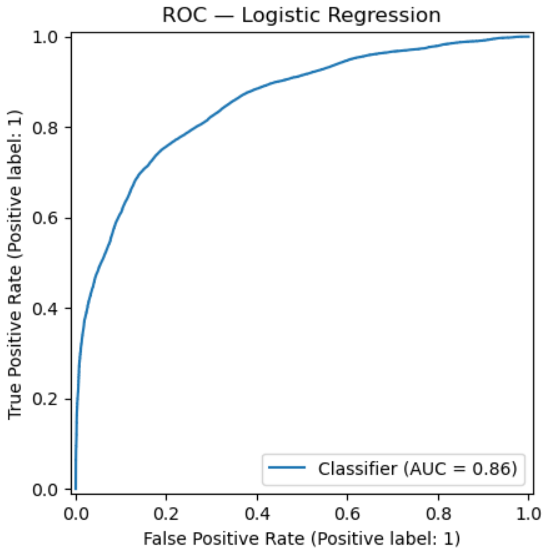
Predictive Methods

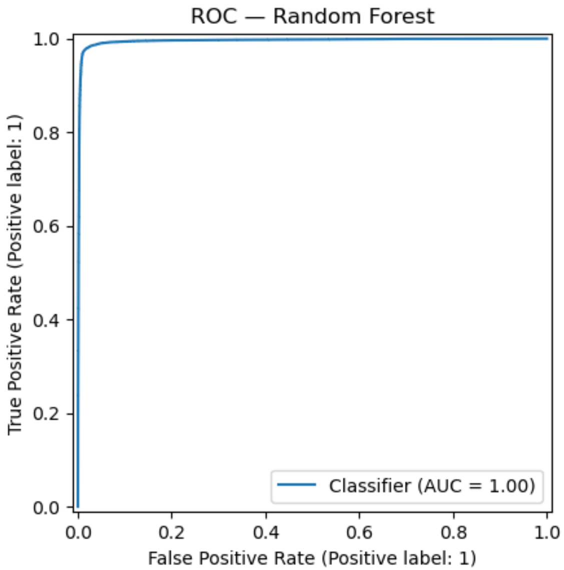
- Chose to perform 3 separate models:
  - Logistic Regression Classifier
    - Use this model as a baseline. Is there a linear relationship between classes?
  - Random Forest Classifier
    - This is preferred if there is nonlinear relationship. It is better than Decision Tree to check multiple split combinations
  - AdaBoost
    - This is a further improvement on random forest, but slower. Will it be worth the computational tradeoff?

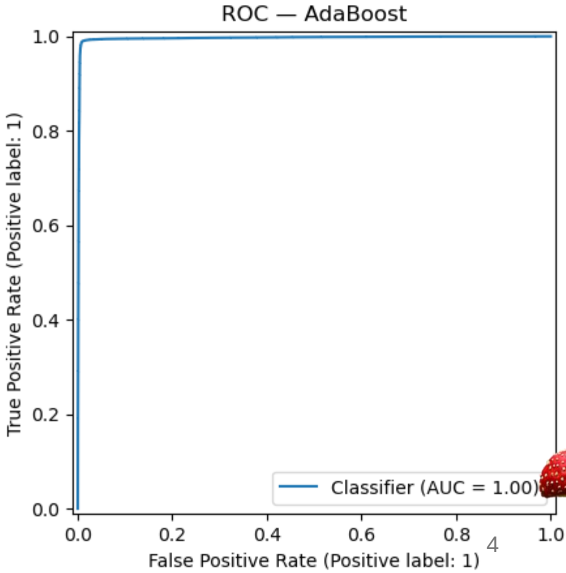
Model Baseline Results

Due to the large dataset (> 3m rows), models were created with 10% random sampling with replacement. In order to keep class proportions balanced, a stratified split was also applied when creating the train-test split for modeling.

Model	Accuracy	F1	ROC
Logistic Regression	78.02%	0.773	0.8551
Random Forest	98.06%	0.981	0.995
AdaBoost	98.99	0.9899	0.996







Baseline Model Discussion

✗ **Logistic Regression** performed worst (78.02% accuracy)

- Potential causes:
  - Non-linear boundary between edible & poisonous class
  - Feature interactions

✓ **Random Forest & AdaBoost** performed best, with AdaBoost performing slightly better (0.9% improvement)

Decision:

- ➢ Choose Random Forest & AdaBoost to perform improvements on

Improvements Proposal

5-fold cross validation

Ensure that accuracy values are correct and not overfitted

Grid search & hyperparameters analysis

Determine best parameters for better error metrics

5-Fold Cross Validation

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avg
Random Forest	97.85	97.88	98.18	97.97	98.1	98.05%
AdaBoost	98.90	98.94	98.94	98.95	98.94	98.93%

➢ The average of both models matches their individual predictions. **Thus, both models are well-fitted to both train & test data.**

Grid Search & Hyperparameters Analysis: Random Forest

Model	Hyperparamters Baseline	Hyperparameters Attempted

Random Forest	`n_estimators=100,`
	`max_depth=14,`
	`min_samples_leaf=10,`
	`max_features='sqrt'`
		`n_estimators={64,100,200}`
		`max_features={2,3,4}`
		`bootstrap = {True,False}`
		`oob_score = {True,False}`

- **Goal:** try different number of trees, number of features used to consider best split, try utilizing the whole dataset to build tree
- **Results:**
  - with this grid search, could only achieve accuracy of 75%.
  - **It appears that *max_features* has the most effect on accuracy for this dataset. Utilizing the *square root* (features) returns best accuracy.**
  - Can achieve 99% accuracy if max_depth & min_samples_leaf are set to default. However, this takes too long computationally. There is tradeoff between the time and model accuracy.

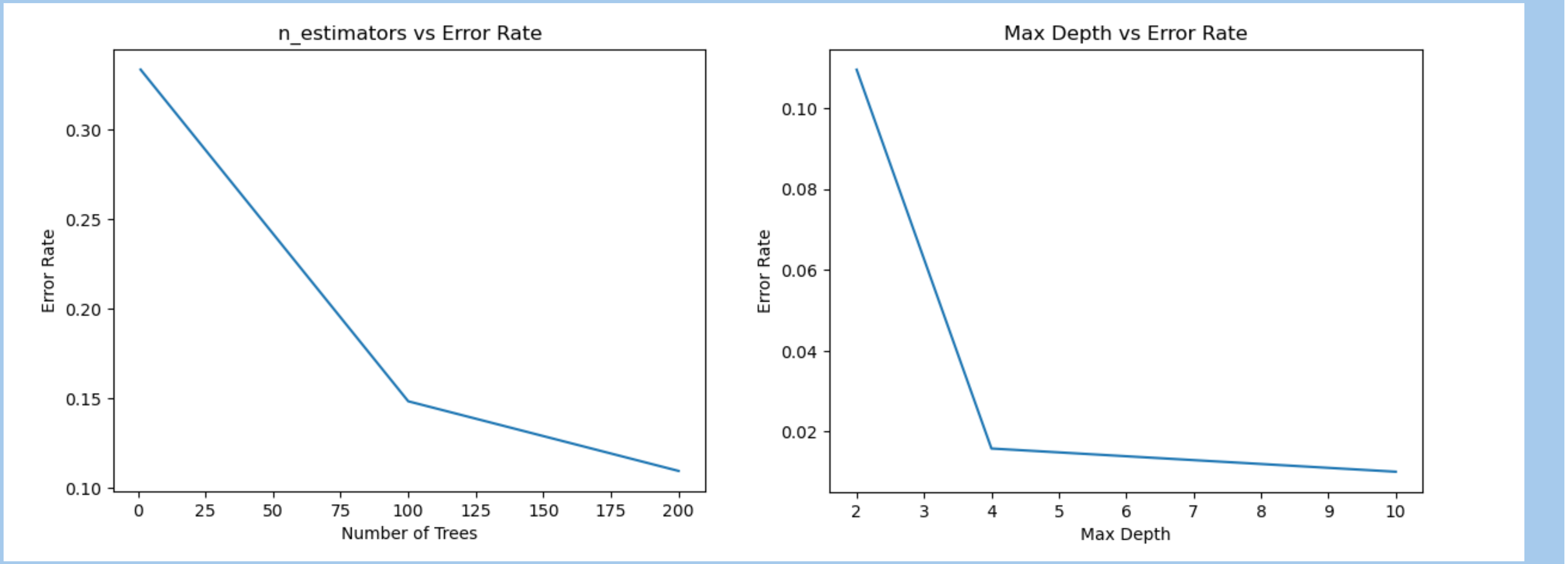
Grid Search & Hyperparameters Analysis: AdaBoost

Model	Hyperparamters Baseline	Hyperparameters Attempted

AdaBoost	`n_estimators=1,`
	`max_depth=2,`
	`learning_rate=1`
		`n_estimators={1,100,200}`
		`max_depth={2,4,10}`
		`learning_rate = {0.5,1,2}`

- **Goal:** try different number of estimators for ensemble learning, different tree depths, and the weight (learning rate) applied to each classifier

Grid Search & Hyperparameters Analysis: **AdaBoost**



- Results: with the baseline hyperparameters, accuracy was ~75%. **200 estimators performed best, with accuracy at 89% for depth = 2. Accuracy continued to improve as depth increased.**

Future Work

- Try SVM model to compare with random forest & AdaBoost
  - This will provide us with 4 models of comparison: **logistic reg, random forest, AdaBoost, and SVM**
- Is there any further feature engineering that we can do to the dataset? Will it improve accuracy?

Conclusion

Random forest and AdaBoost baseline models return the best accuracy & fit. AdaBoost performs slightly better.

In order to achieve such high accuracy, hyperparameters for random forest & AdaBoost were tuned.

Will create SVM model for further comparison.

Determine best overall model and further fine tune as able.

Links

- Jupyter Notebook Codes:

  https://github.com/molls1889/DataMiningCheckpoint2

- Video Link:

https://github.com/molls1889/DataMiningCheckpoint2