

# Comparing Machine Learning Models to Determine Mushroom Edibility

Defne AYDIN, dga264

Timur GORDON, tbg252

<https://github.com/defneaydin/CS-UY-4563-Final-Project>

# Problem Statement

The project aims to determine the best machine learning algorithm to predict whether a mushroom species is poisonous or edible. The data will be taken from the UCI Machine Learning Repository: Mushroom Data Set.

The dataset contains information about 23 species of gilled mushrooms in the Agaricus and Lepiota family.

The goal is to determine the classifier with the highest precision.

Results falsely determined to be edible pose a great health risk.

*Which classification algorithm is best suited for this task?*



# Classifiers to Compare

- Logistic regression
- Support Vector Machine
- Decision Tree
- K Nearest Neighbors
- Random Forest



# Problem formulation

## Loss Functions

- Binary cross entropy

$$J(\bar{\mathbf{w}}) = \sum_{i=1}^n \ln(1 + e^{z_i}) - y_i z_i$$

- Mean Squared Error

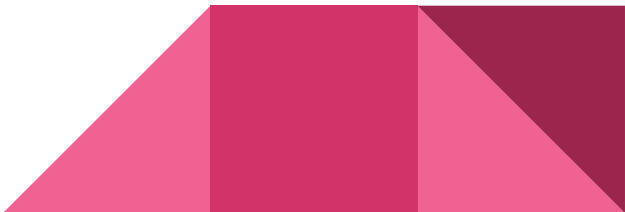
$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

## Metrics

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 = \frac{2 * precision * recall}{precision + recall}$$



# Dataset/training data

- 22 nominally valued attributes
- Encoded -> 98 binary valued attributes
- 8124 instances
- Classified as edible and poisonous

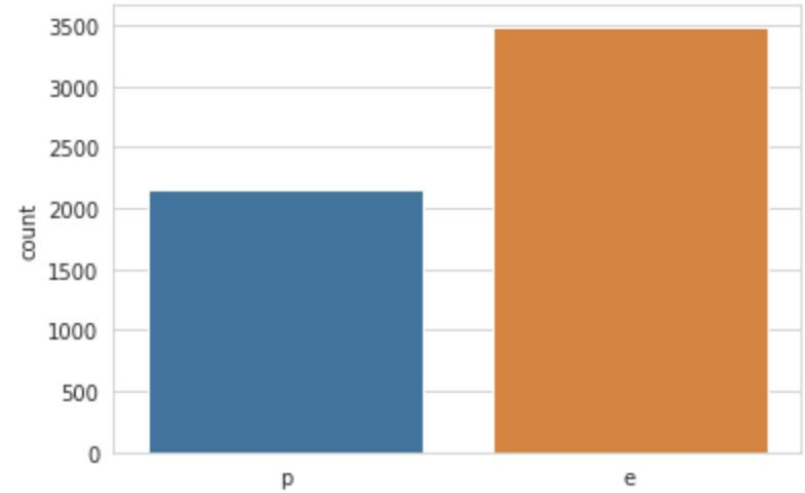


Figure 1: Class distribution



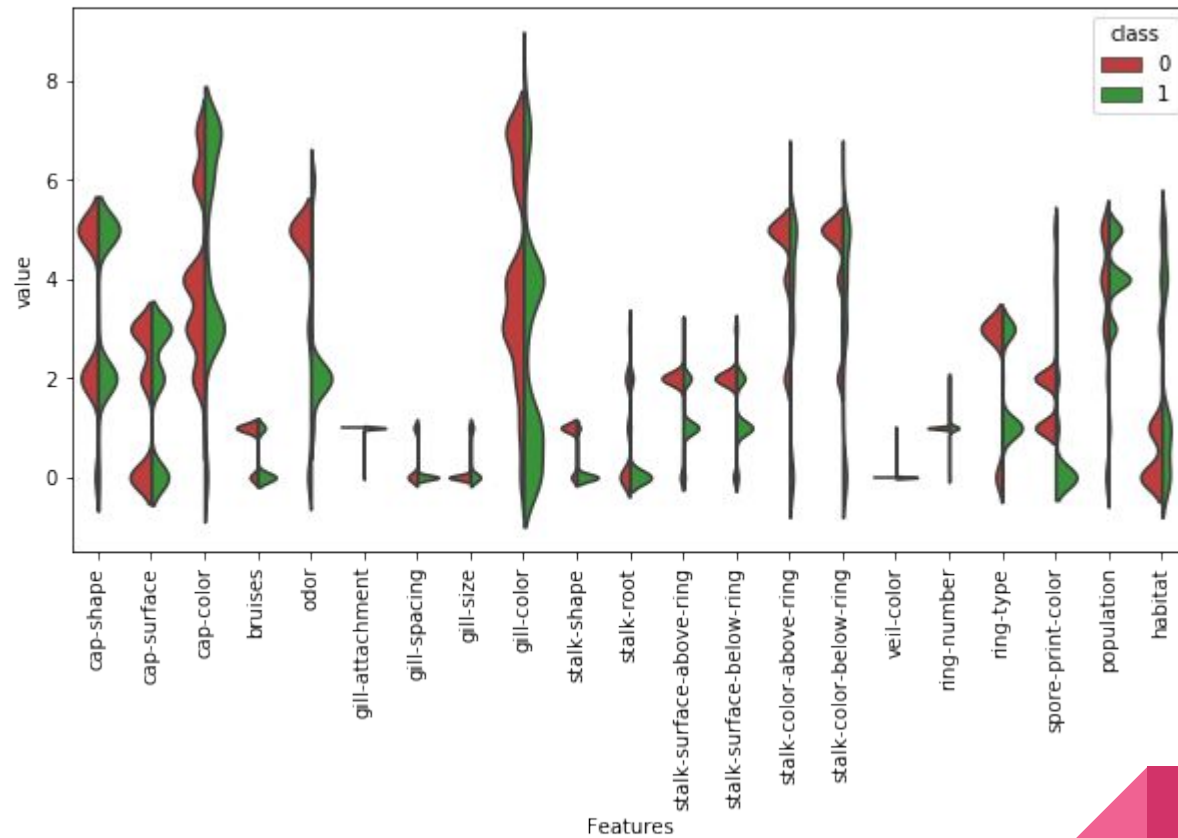


Figure 2: Class distribution per attribute

# Method

- When all attributes are used, classifiers yield near 100% accuracy.
- Chose a subset of attributes that yield approximately 90% accuracy for the logistic regression model.
- For comparison purposes, all classifiers were trained with this subset consisting of 9 attributes.

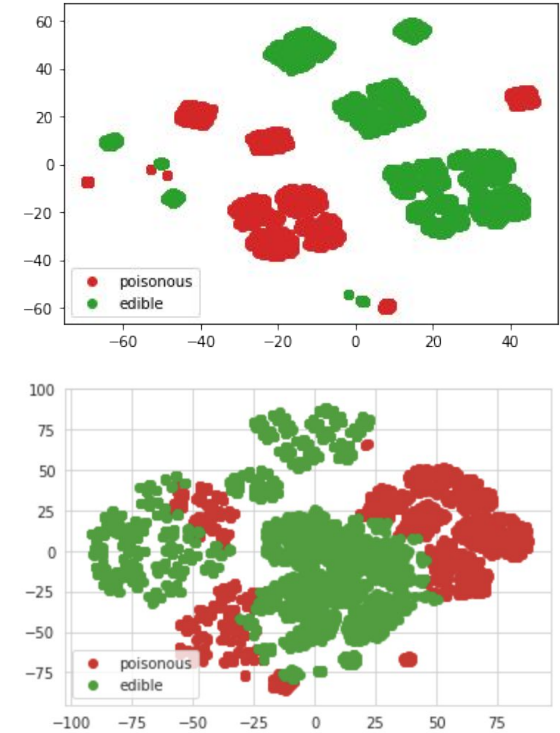


Figure 3: TSNE comparison for the initial data and a feature subset of 9 attributes

# Evaluation Results

- For all 5 classifiers, the f1-score, the accuracy, and the precision were recorded.
- Logistic regression has the lowest accuracy, f1-score, and precision.
- The decision tree classifier has the highest precision but lower accuracy than SVM and KNN.

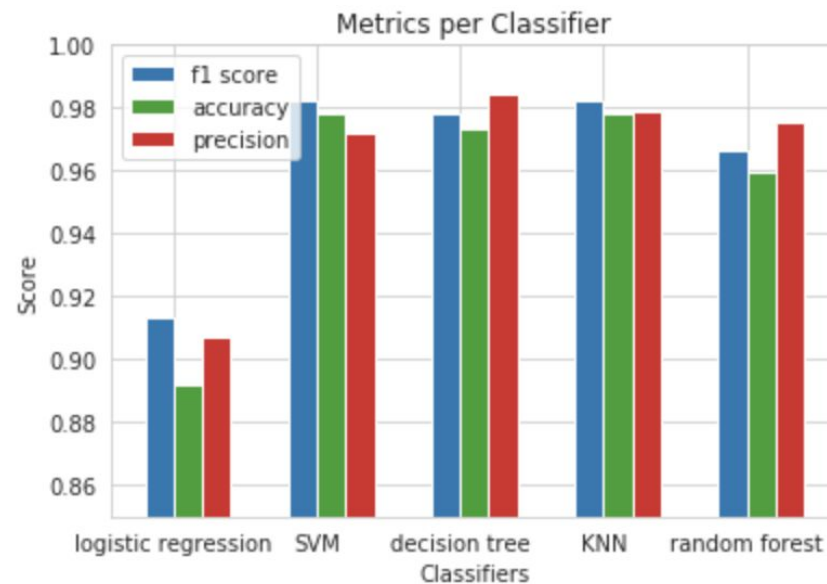


Figure 4: Classifier metrics comparison



# Confusion Matrices

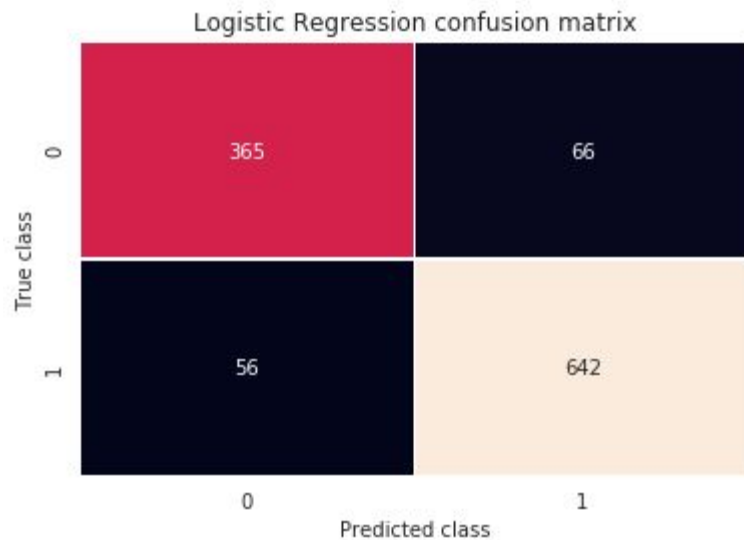


Figure 5: Logistic Regression confusion matrix

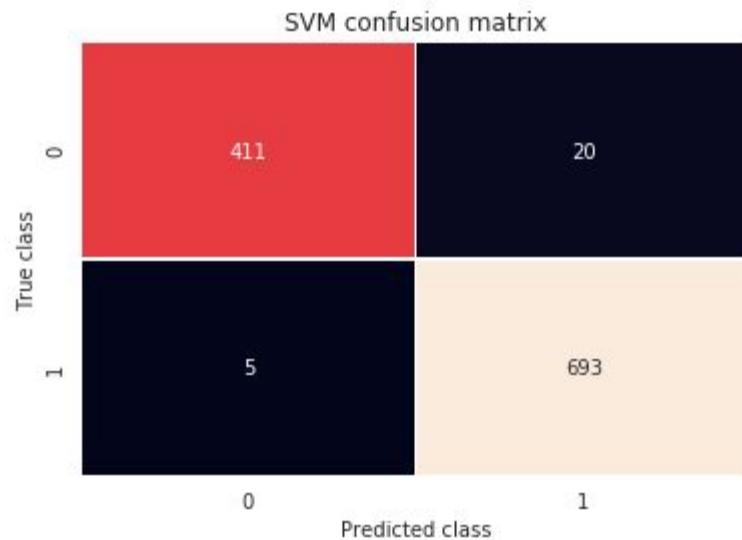


Figure 6: SVM confusion matrix

# Confusion Matrices

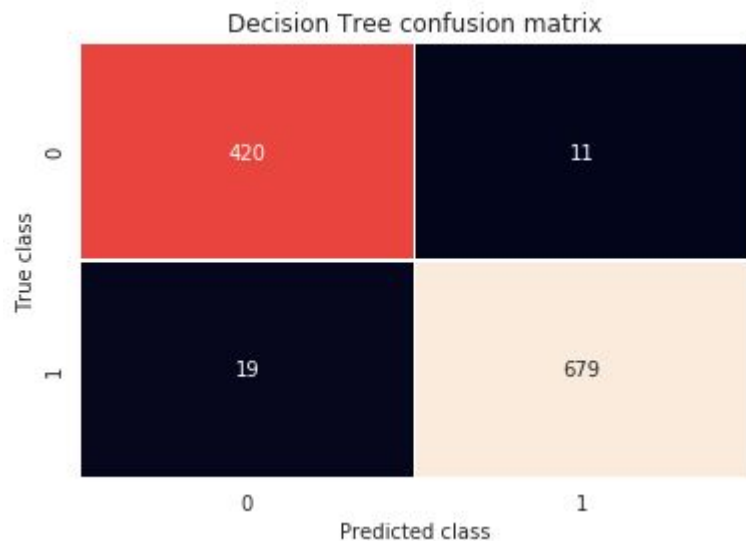


Figure 7: Decision Tree confusion matrix

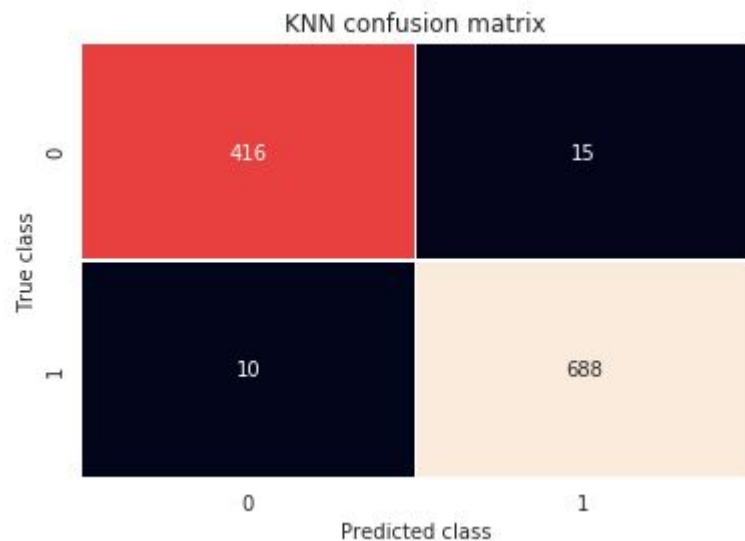


Figure 8: KNN confusion matrix

# Confusion Matrices

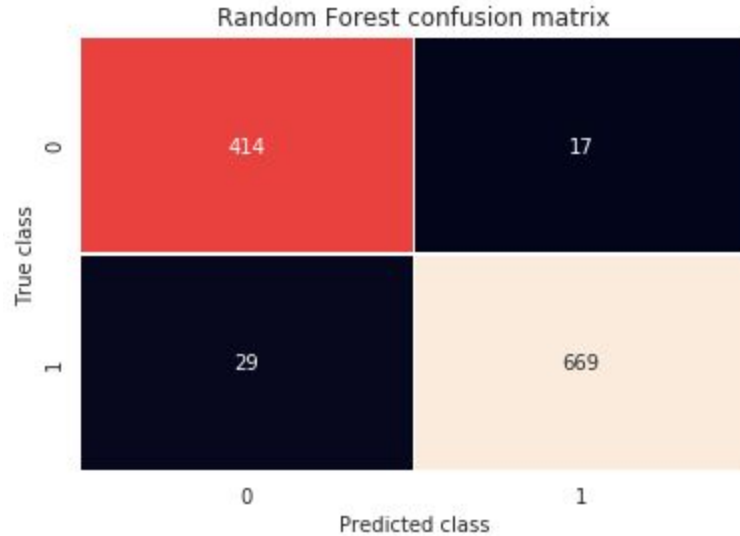


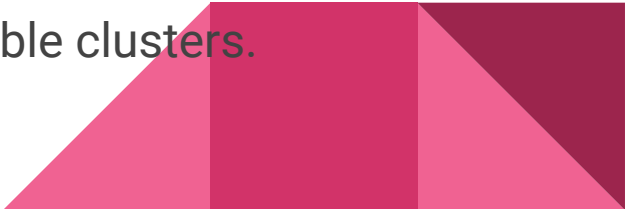
Figure 9: Random Forest confusion matrix

# False Positive Rates

- Logistic Regression = 0.153
- SVM = 0.046
- Decision Tree = 0.026
- KNN = 0.035
- Random Forest = 0.039



# Conclusion

- All five classifiers achieve perfect results when trained and tested on the complete dataset with 22 features.
  - When a subset of features are selected such that the logistic regression classifier achieves only 90% accuracy:
    - The SVM, Decision Tree and KNN outperform other classifiers.
    - The Logistic Regression classifier is the deadliest.
    - The Decision Tree classifier predicts the lowest false positives.
  - KNN and SVM are good candidates due to their high f1 score.
  - Decision Tree classifiers have the lowest false positive rate overall.
  - Logistic regression classifiers struggle with inseparable clusters.
- 

# References

[1]<https://www.kaggle.com/haimfeld87/analysis-and-classification-of-mushrooms>

[2]<https://archive.ics.uci.edu/ml/datasets/Mushroom>

[3] Schlimmer, J.S. (1987). Concept Acquisition Through Representational Adjustment (Technical Report 87-19). Doctoral dissertation, Department of Information and Computer Science, University of California, Irvine.

[4] Iba, W., Wogulis, J., & Langley, P. (1988). Trading off Simplicity and Coverage in Incremental Concept Learning. In Proceedings of the 5th International Conference on Machine Learning, 73-79. Ann Arbor, Michigan: Morgan Kaufmann.

[5] Duch W, Adamczak R, Grabczewski K (1996) Extraction of logical rules from training data using backpropagation networks, in: Proc. of the The 1st Online Workshop on Soft Computing, 19-30.Aug.1996, pp. 25-30.

[6] Duch W, Adamczak R, Grabczewski K, Ishikawa M, Ueda H, Extraction of crisp logical rules using constrained backpropagation networks - comparison of two new approaches, in: Proc. of the European Symposium on Artificial Neural Networks (ESANN'97), Bruges, Belgium 16-18.4.1997.