A Comparison of Naïve Bayes and Random Forest on Mushroom Poisonous

Ho Yin, Tam

Description and motivation:

- Mushrooms have become an integral part of the daily meals. However, eating toxic mushrooms may cause vomiting, organ failure or even death [1]. Therefore, distinguishing if a mushroom is edible or poisonous is important.
- Build Naïve Bayes model and random forest model to classify mushroom poisonous under the approach of 80% training set and 20% test set splitting and 10-fold cross-validation on the training set.
- Contrast and evaluate the model performance of Naïve Bayes and random forest including accuracy, precision, recall, specificity, F1 score, area under curve (AUC) value, training time and test time.

Exploratory data analysis:

- The data extracted from the UCI Machine Learning Repository consists of 8124 rows and 23 columns which 22 columns are features and 1 column is the target label [2].
- The features are the mushroom's physical characteristics, such as the cap color, gill size, stalk shape, ring number, and habitats, while the target is to classify if a mushroom is edible or poisonous. Both the features and the target label are categorical variables.
- Feature 'stalk-root' is removed as 2480 observations have missing values. Also, the feature 'veil-type' is dropped as there is only one outcome.
- The pie chart (figure 1) shows that 4208 observations are edible, and 3916 observations are poisonous which accounts for 51.8% and 48.2% respectively. It counts as a balanced dataset.
- The Cramer's V heatmap (figure 2) provides an overview of the correlation within the features and between the features and the label. The 'odor' and 'spore print color' seem to have a strong correlation with the mushroom poisonous as the correlation coefficient is 0.94 and 0.57 respectively.
- The bar chart (figure 3) effectively visualizes the frequency of outcomes of 20 features as they are all categorical variables.

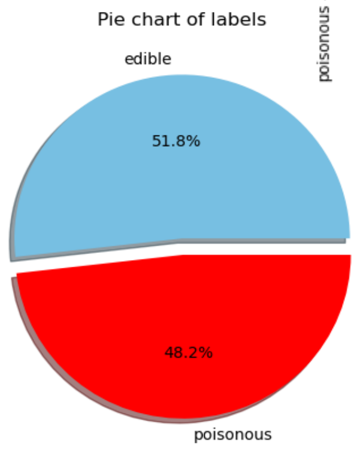

Figure 1

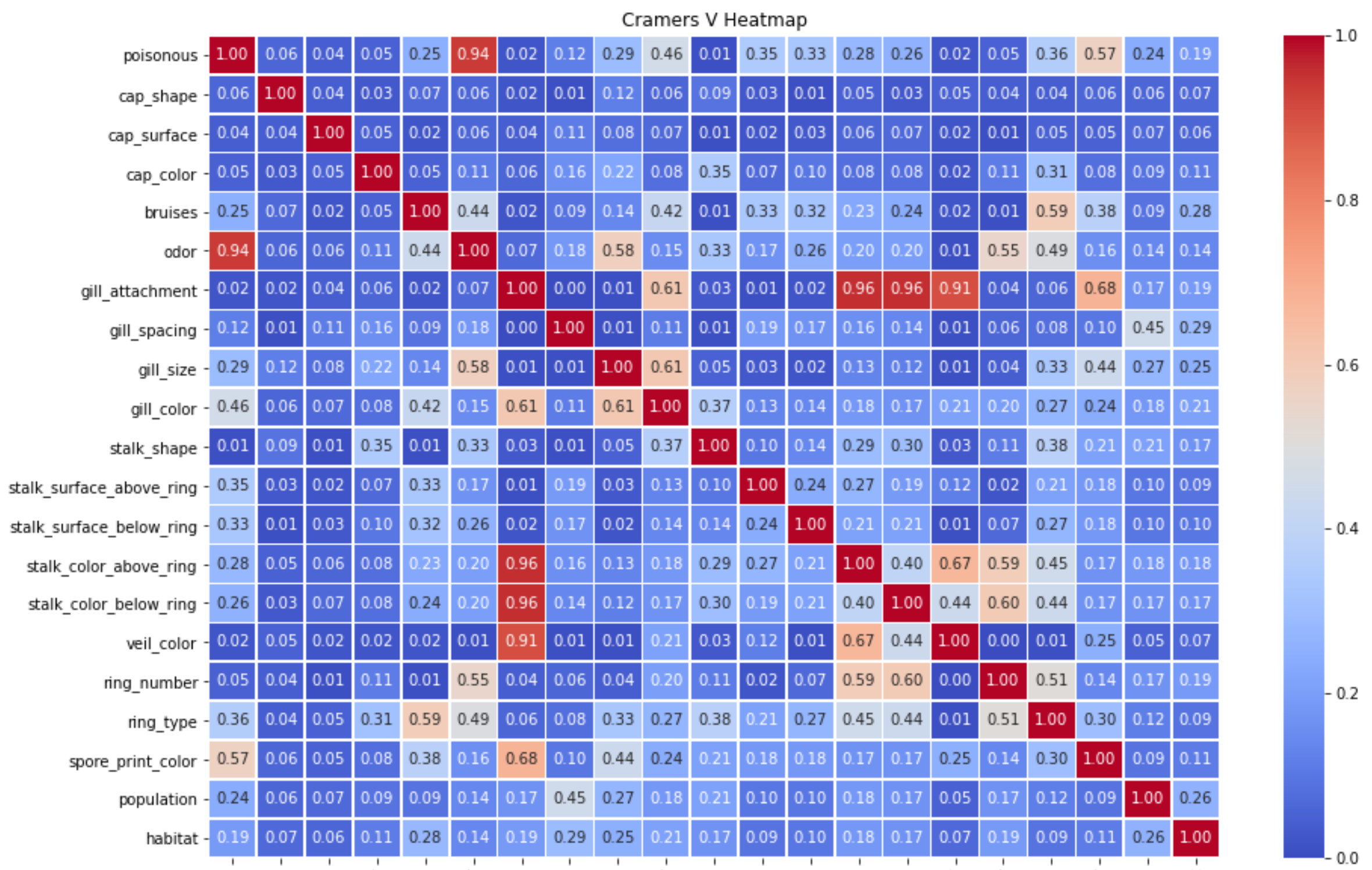

Figure 2

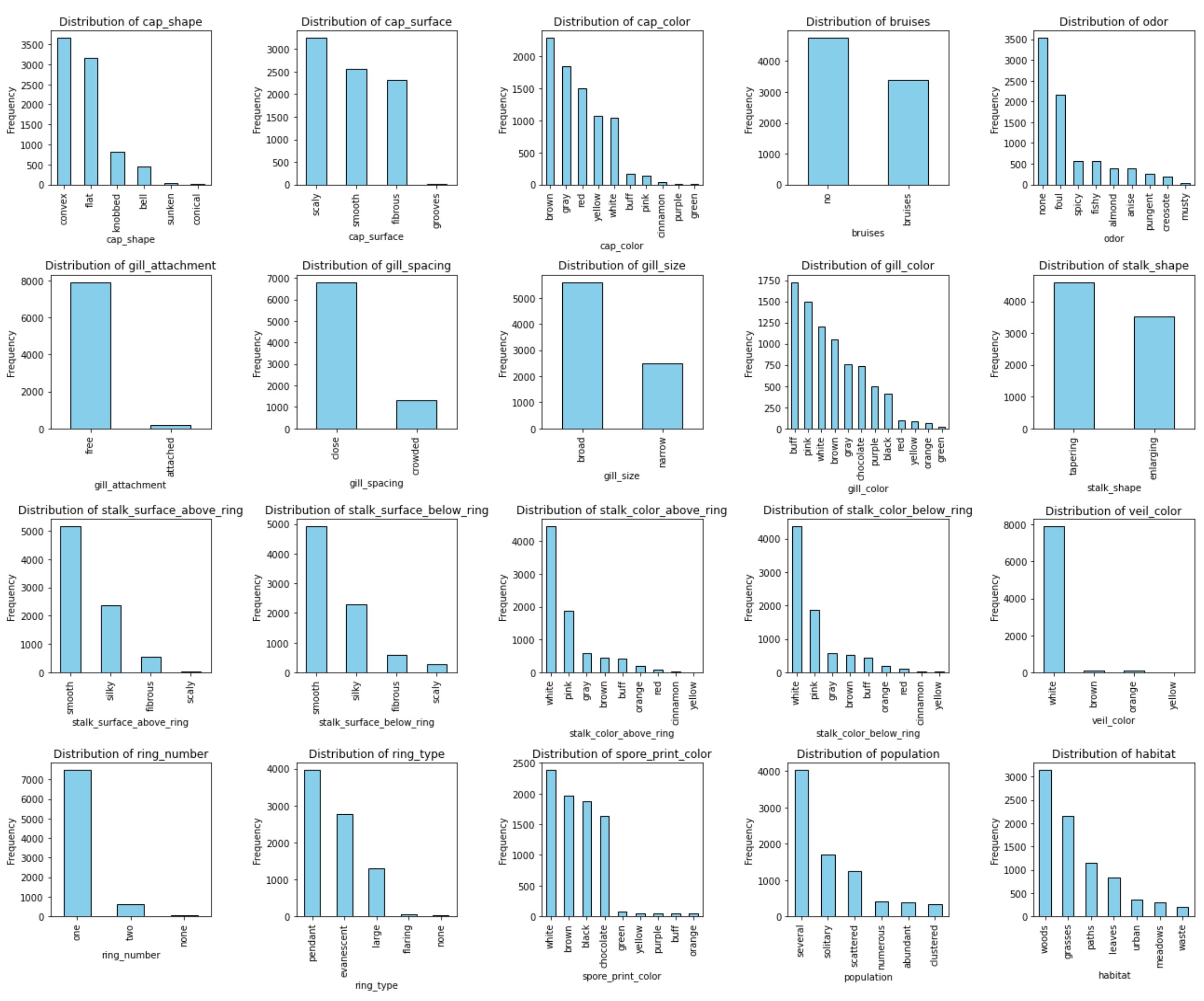

Figure 3

Hypothesis statement:

- In the paper of comparison of different classification algorithms, Naïve Bayes is outperformed by random forest and other classification methods in terms of performance metrics [3].
- The random forest has a longer training time than the Naïve Bayes.

Description of the choice of training and evaluation methodology:

- The dataset is split into an 80% training set (6500 observations) and a 20% test set (1624 observations). The test set remains unseen until the model is trained.
- Ten-fold cross-validation is applied to the training set (6500 observations) in which the training set is further divided into 10 folds. 9 folds of training set (5850 observations) and 1 fold of validation set (650 observations) are in each iteration and repeat the process for 10 times until all folds are validated.
- The confusion matrix for the cross-validation is shown. Then the average performance metrics (accuracy, precision, recall, specificity, F1 score, and AUC value), and average training time are calculated to evaluate the performance of the models. Also, the receiver operating characteristic (ROC) curve with the AUC value for 10 folds is plotted.
- Test the two trained models with the test set.
- The confusion matrix, performance metrics, test time, and the ROC curve with the AUC value of the test set are evaluated.

Choice of parameters and experimental result:

Naïve Bayes:

Choice of parameters:

- Since all features are categorical variables, a multivariate multinomial Naïve Bayes model is used.
- Addictive smoothing is included to avoid zero-frequency problems.

Experiment results:

- The performance metrics of the validation set and test set are shown in Table 1 and Table 2 respectively, such as the average accuracy of the 10 validation sets is 0.96 while the test accuracy is 0.95.
- The average AUC value of the validation set and the AUC value of the test set are 0.9979 and 0.9966 respectively.
- The total training time is 0.2910 seconds while the test time is 0.0434 seconds.

Random Forest:

Choice of parameters:

- Using grid search, the optimal number of trees is 15 and the optimal number of features to select in each node is 9.

Experiment results:

- The performance metrics of the validation set and test set are shown in Table 1 and Table 2 respectively. All the performance metrics including accuracy, precision, recall, specificity, and F1 score of both the validation set and the test set is 1.
- Both the average AUC value of the validation set and the AUC value of the test set are 1.
- The hyperparameter tuning time is 1879 seconds, the model training time is 5.3739 seconds, and the test time is 0.3245 seconds.

Validation Set		
	Naïve Bayes	Random Forest
Average Accuracy	0.9600	1.0000
Average Precision	0.9932	1.0000
Average Recall	0.9234	1.0000
Average Specificity	0.9332	1.0000
Average F1 Score	0.9570	1.0000
Average AUC value	0.9979	1.0000
Total Training time (seconds)	0.2910	5.3739 + 1879 (Hyperparameter tuning time) = 1884.3739

Table 1

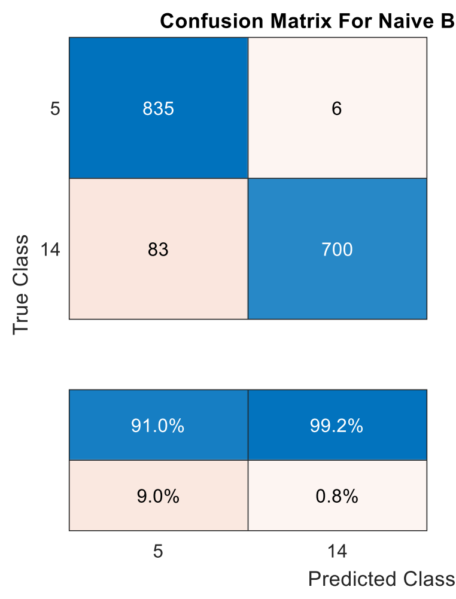

Figure 4

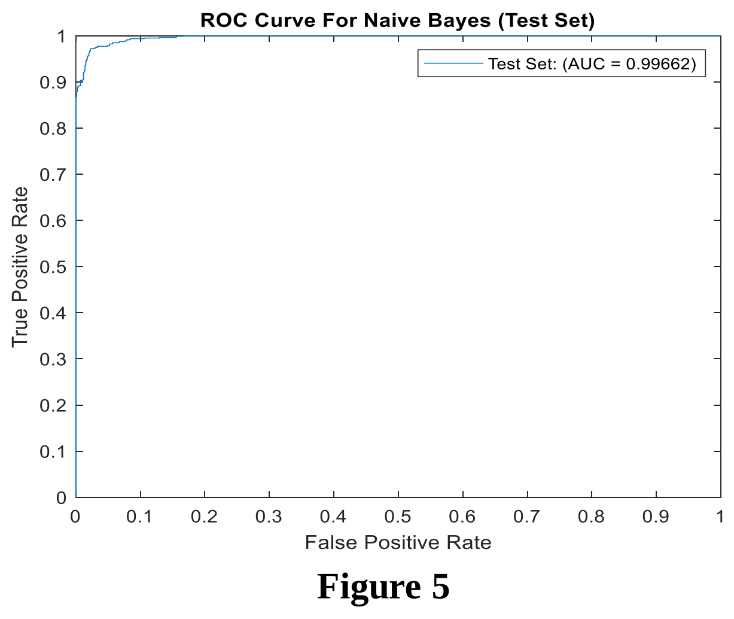

Figure 5

Test Set		
	Naïve Bayes	Random Forest
Test Accuracy	0.9452	1.0000
Test Precision	0.9915	1.0000
Test Recall	0.8940	1.0000
Test Specificity	0.9096	1.0000
Test F1 Score	0.9402	1.0000
Test AUC value	0.9966	1.0000
Test Time (seconds)	0.0434	0.3245

Table 2

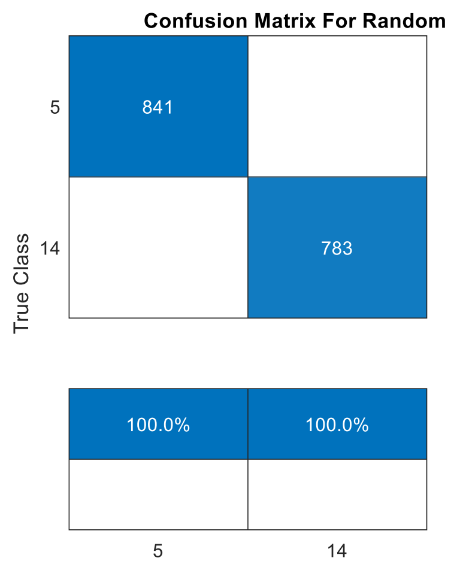

Figure 6

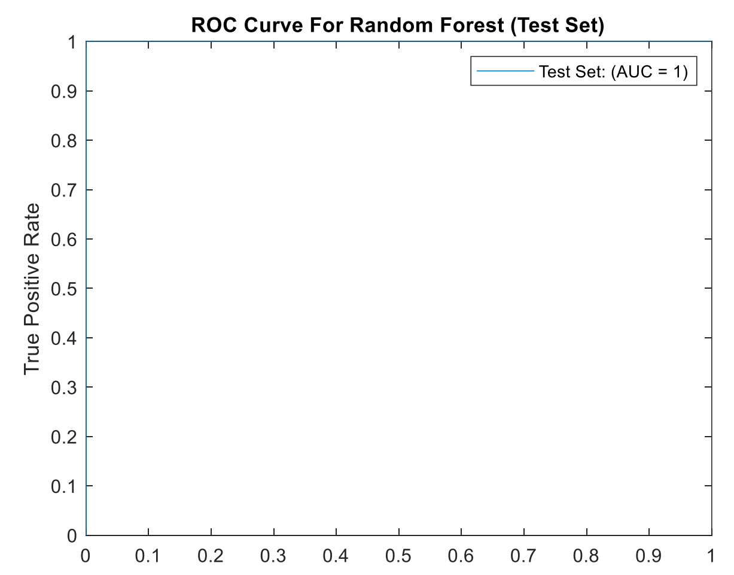

Figure 7

Summary of the two machine learning methods with their pros and cons:

Naive Bayes:

- Naïve Bayes is a supervised learning algorithm for classification based on Bayes' Theorem.
- Naïve Bayes is also named Idiot Bayes and Simple Bayes as the algorithm assumes all the features or attributes are conditionally independent of each other [4].
- Naïve Bayes can be applied for the mushroom poisonous. The class of edible or poisonous with a higher probability will be chosen.

Pros:

- Simple and fast.
- High accuracy even with less training data.
- Robust to noise.

Cons:

- Naïve Bayes assumed that all attributes are conditionally independent. However, from the Cramer's V heatmap, it is shown that the variable of 'stalk color above ring' and 'gill-attachment', and 'stalk color below ring' and 'gill-attachment' have a strong correlation.
- Test data with an attribute that never appears in the training data will be classified with zero probability.

Random Forest:

- Random forest is a supervised learning algorithm for both classification and regression and is an ensemble of decision trees [5].
- Random Forest can be applied to classify if a mushroom is edible or poisonous. The class with the highest vote from most trees is chosen for the classification task while the average value of all trees is calculated for the regression task.

Pros:

- Risk of overfitting is significantly reduced.
- Boosted performance compared with a single decision tree.

Cons:

- Longer time to train.
- Loss of interpretability.

Analysis and critical evaluation of results:

- The overall experiment results show that random forest outperformed Naïve Bayes in terms of classification performance metrics (accuracy, precision, recall, specificity, F1 score), AUC value, training time, hyperparameter tuning time and test time. This supports the first hypothesis statement.
- In both the validation set and test set, Naïve Bayes performs well and has a value of over 0.9 in all performance metrics while random forest performs extremely well and has a value of 1 in all performance metrics. Both models give surprisingly great and satisfactory results.
- Higher classification performance metrics are preferred in health affairs and other issues that have a high cost of error. As eating toxic mushrooms will cause vomiting, organ failure or even death, having higher performance metrics implies that an edible or poisonous mushroom can be accurately and precisely classified. The miserable consequence can then be avoided.
- As the dataset has a balanced target label, accuracy is a crucial measure of the number of correctly predicted observations out of the total number of observations. The test accuracy of random forest is 5.8% higher than Naïve Bayes. The average accuracy of validation sets, and the accuracy of the test set for random forest are 1 in which the observations are 100% accurately predicted.
- The test AUC value of random forest is 0.34% higher than Naïve Bayes. The average AUC value of the validation sets, and test AUC value of random forest is 1 in which the observations are 100% correctly classified.
- The total training time of Naïve Bayes is over 6475 times faster than random forest as random forest takes more than 30 minutes for hyperparameter tuning. It indicates that Naïve Bayes is simpler and has fewer hyperparameters while random forest is more complicated and has more hyperparameters.
- Even if the hyperparameter tuning process is omitted, the model training time of Naïve Bayes is still 18 times faster than random forest. This supports the second hypothesis statement.

Lesson learned:

- The Naïve Bayes model and random forest model are two powerful machine learning models. The higher value of classification performance metrics is preferred when the task is related to health affairs or other issues in which the cost of error is high. In this case, the higher values of classification metrics are preferred as eating toxic mushrooms will cause serious consequences or even lead to death.
- The Naïve Bayes has fewer hyperparameters and leads to a shorter training time and test time, while the random forest has more hyperparameters, such as the number of trees and the number of features to select in each node, causing a longer time to train the model and test.

Future work:

- Split the dataset into other percentages of training set and test set, such as 70% training set and 30% test set, and evaluate the performance metrics and compare with the current results.
- Other hyperparameters of the random forest model, such as the maximum depth of trees, maximum number of leaf nodes and minimum number of observations present in leaf nodes, can be tuned to evaluate the impact of the performance metrics.

Reference:

[1] Tran HH, Juergens AL. Mushroom Toxicity. [Updated 2023 Aug 7]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK537111/
[2] Mushroom. (1987). UCI Machine Learning Repository. https://doi.org/10.24432/C5959T.
[3] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning, pages 161–168, 2006
[4] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.
[5] Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York, Springer.