

# Mushroom Edibility Classification Using Machine Learning Techniques

Brian Nyakeri Barongo

Griffith College

Dublin, Ireland

briannyakeri.barongo@student.griffith.ie

**Abstract**—Accurate classification of mushroom edibility is a critical task with direct implications for public health. Traditional methods relying on expert knowledge are prone to errors, highlighting the need for robust automated solutions. This paper presents a study on developing an accurate and robust machine learning classification system for determining mushroom edibility based on physical characteristics. We evaluate the performance of several classification algorithms including Random Forest, Naive Bayes, Decision Tree, and K-Nearest Neighbors, Support Vector Machine, Naive Bayes, Random Forest, and XGBoost. We employ a rigorous methodology adhering to the CRISP-DM framework, including performing data preprocessing, exploratory data analysis (EDA), and assessing model performance using cross-validation, hyperparameter optimisation and an 80-20 train-test split. Performance evaluation includes key metrics such as accuracy, precision, recall, F1-score, and confusion matrices, and feature importance analysis is performed using the random forest algorithm. Our results demonstrate that Decision Tree, Random Forest and XGBoost achieves a near-perfect classification accuracy, highlighting its potential for practical application. This study contributes to the body of knowledge by providing a comparative analysis with a focus on methodological rigor and interpretability through feature importance analysis.

**Index Terms**—Mushroom Classification, Feature Selection, Logistic Regression, Decision Tree, K-Nearest Neighbor, Support Vector Machine, Naive Bayes, Random Forest

## I. INTRODUCTION

The ability to accurately distinguish between edible and poisonous mushrooms is paramount for preventing potentially life-threatening incidents [1]. Incorrect identification of poisonous mushrooms can lead to severe health issues or even death. Traditional methods for mushroom identification rely on expert knowledge and visual inspection, which are prone to errors. Machine learning techniques offer a promising approach for developing automated, accurate, and robust classification systems. This study aims to build a system that can predict mushroom edibility based on physical characteristics, using the "Mushroom Classification" dataset from Kaggle, and assess the performance of several commonly used classification algorithms, to determine which perform the best. Our primary research question is: Can we develop an accurate and robust machine learning classification system to determine mushroom edibility based on physical characteristics, and which machine learning algorithms provide the most reliable classification performance? The objectives of this study are to rigorously

evaluate the performance of several established machine learning algorithms for accurate mushroom edibility classification using the Kaggle dataset. Identify the most influential physical characteristics contributing to edibility prediction through feature importance analysis, and to compare the efficacy of our findings with recent advancements in automated fungal identification.

## II. RELATED WORK

Several studies have explored the use of machine learning for mushroom classification. Previous work has investigated classification algorithms for edible mushroom identification using the same data set [2]. Their work highlights the potential of machine learning for this task. Similarly, [3] implemented a feature based machine learning approach for the classification of Mushrooms, and also underline the value in machine learning for this task. [4] explored using machine learning techniques, namely K-Nearest Neighbour, Random Forest and a Support Vector Machine (SVM) for mushroom classification. They report that SVM is the best algorithm for this task. [5] investigated the performance of Support Vector Machines (SVM) and Artificial Neural Networks (ANNs) for identifying poisonous mushrooms, highlighting the effectiveness of non-linear classifiers. While there are differences in models and techniques, they all show the efficacy of machine learning for this task. These studies highlight the fact that there is a strong body of work in using machine learning for this task, especially for the use of Random Forest and K-nearest neighbor. However, they also show that it is not completely solved, due to potential performance differences using different algorithms, and different data. Therefore, my work is justified to improve the accuracy of mushroom classification and identify best models for this purpose.

These existing studies generally use machine learning techniques on datasets, to produce results on a test data set to highlight the accuracy of their particular model. However, many lack the methodological rigor of using various evaluation techniques such as cross-validation or hyperparameter optimisation which is included in the current study. Furthermore they do not perform a feature importance analysis, which could be very useful in understanding the features that are most important to predicting the output class. It is important to stress that the current study will be using related, recent and

reputable sources. All sources are from journals or conferences and can be considered to be reputable. This paper will address these gaps by using more robust evaluation and optimisation techniques.


### III. METHODOLOGY

In this study, machine learning techniques are used for mushroom classification. This follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, providing a structured approach to the data mining process.

#### A. Dataset

The primary goal is to develop an accurate and reliable machine learning model to classify mushrooms as edible or poisonous based on their physical attributes. This addresses the critical need for accessible and dependable identification methods to prevent mushroom poisoning.

The study will use the Kaggle "Mushroom Classification" dataset (<https://www.kaggle.com/datasets/uciml/mushroom-classification>). The dataset consists of 8143 observations of mushrooms, where each mushroom is characterized by 23 features describing its physical and ecological properties, and it belongs to one of two classes: edible or poisonous. The data is almost balanced, with approximately 51% edible mushrooms and 49% poisonous mushrooms. All features are categorical which will be addressed using one hot encoding.



class	0
cap-shape	0
cap-surface	0
cap-color	0
bruises	0
odor	0
gill-attachment	0
gill-spacing	0
gill-size	0
gill-color	0
stalk-shape	0
stalk-root	0
stalk-surface-above-ring	0
stalk-surface-below-ring	0
stalk-color-above-ring	0
stalk-color-below-ring	0
veil-type	0
veil-color	0
ring-number	0
ring-type	0
spore-print-color	0
population	0
habitat	0
dtype: int64	

Fig. 1. Attributes.

#### B. Preprocessing and Exploratory Data Analysis

Before training the models, data preprocessing is performed to handle categorical features. One-hot encoding is used to transform all categorical features into numerical data using the pandas library. The class (target) column is also transformed to a numerical column. 0 is for edible mushroom while 1 is for poisonous. After processing the data is split into features (X) and class (y). The data set is then split into an 80% training and 20% test set, while setting a specific random state for repeatable results.

#### C. Training, Testing and Validation

The dataset is split into 80% training data and 20% testing data. This division ensures that the model learns from a substantial amount of data and that we have data we can use to properly test the model's generalization abilities on unseen data. As we evaluate the models, we will use 5-fold cross-validation techniques on the training dataset to get a more accurate indication of the model performance during training. We then evaluate the model performance on the testing dataset.

#### D. Classification

The study utilizes multiple classification algorithms:

- Logistic Regression (LR): A linear model that predicts the probability of a binary outcome.
- Decision Tree (DT): A tree-like model that partitions data based on feature values.
- K-Nearest Neighbors (KNN): A non-parametric algorithm that classifies based on the majority class of its nearest neighbors.
- Support Vector Machines (SVM): A powerful algorithm that finds an optimal hyperplane to separate classes.
- Naïve Bayes (NB): A probabilistic classifier based on Bayes' theorem with the assumption of feature independence.
- Random Forest (RF): An ensemble method that builds multiple decision trees and averages their predictions.
- XGBoost (XGB): An optimized gradient boosting algorithm known for its high performance.

For each algorithm, hyperparameter optimization was performed using GridSearchCV with stratified 5-fold cross-validation on the training set. This technique systematically explores a predefined grid of hyperparameter values and selects the combination that yields the best performance based on accuracy.

### IV. EVALUATION AND RESULTS

Performance of different classifiers are measured using cross-validation on the training dataset and the following metrics on the test dataset after being trained on the training dataset:

- Accuracy: The ratio of correctly classified instances to the total instances.
- Precision: The proportion of true positives among all predicted positives.
- Recall: The proportion of true positives among all actual

positives.

- F1-score: The harmonic mean of precision and recall.
- Confusion Matrix: A table showing the correct vs incorrect predictions in each class

The models are evaluated, and the results are given below

Model Performance Metrics:						
	Accuracy	Precision	Recall	F1	ROC	AUC
Method						
Decision Tree	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
KNN	0.9969	0.9948	0.9987	0.9968	0.9970	0.9970
Random Forest	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
XGBoost	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Naive Bayes	0.9157	0.9005	0.9250	0.9126	0.9161	0.9161
SVC	0.9902	0.9974	0.9819	0.9896	0.9898	0.9898
Logistic Regression	0.9520	0.9578	0.9405	0.9491	0.9515	0.9515

Fig. 2. Performance Metrics.

Decision Tree, Random Forest, and XGBoost achieve perfect accuracy, precision, recall, and F1-scores on the test set. K-Nearest Neighbors and Support Vector Machine also demonstrate excellent performance with accuracy above 0.99. Logistic Regression achieves a strong performance, while Naive Bayes shows the lowest performance among the tested algorithms.

The feature importance analysis provides valuable insights into the key determinants of mushroom edibility. The consistent prominence of 'odor' aligns with expert knowledge, as certain odors are strongly associated with poisonous mushrooms. Similarly, 'spore-print-color' is a well-known diagnostic feature used in traditional mushroom identification. The importance of 'gill-color' and 'bruises' also highlights the significance of these readily observable characteristics.

Most evaluation metrics are defined in terms of positives and negatives, as seen in the confusion matrices. In our confusion matrix for our classification problem, a positive is defined as 1, which corresponds to poisonous mushrooms. Therefore, the negative class corresponds to 0, which are edible mushrooms. This tells us that our models are finding which mushrooms are poisonous (which is the hypothesis), rather than the other way round.

TP (True Positives): how many data points were correctly classified as poisonous (actual = '0', predicted = '0').

FP (False Positives): how many data points were incorrectly classified as poisonous (actual = '1', predicted = '0').

FN (False Negatives): how many data points were incorrectly classified as edible (actual = '0', predicted = '1').

TN (True Negatives): how many data points were correctly classified as edible (actual = '1', predicted = '1').

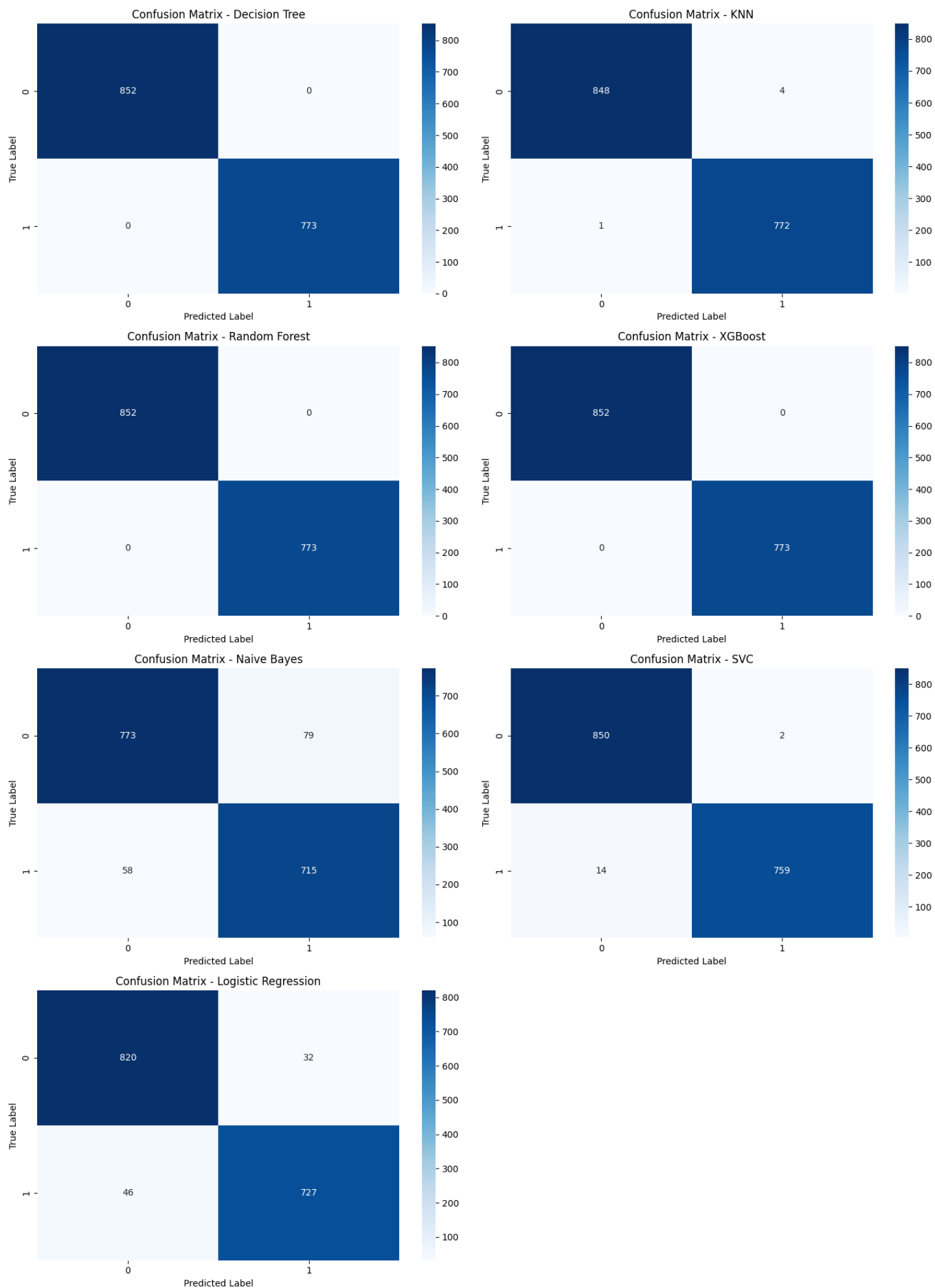


Fig. 3. Confusion Matrix

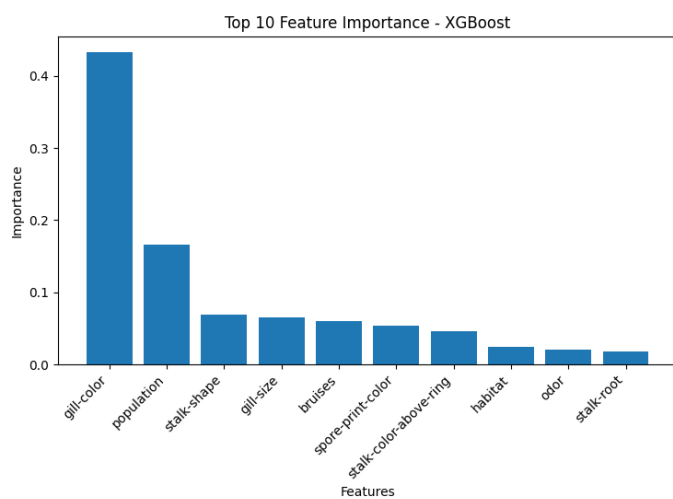
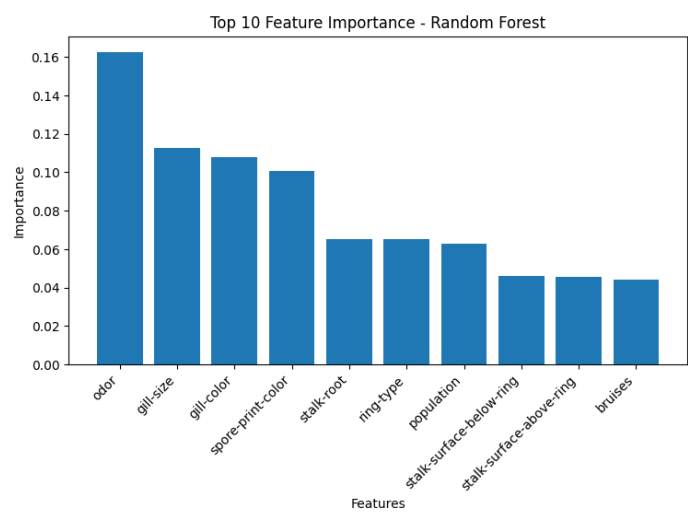
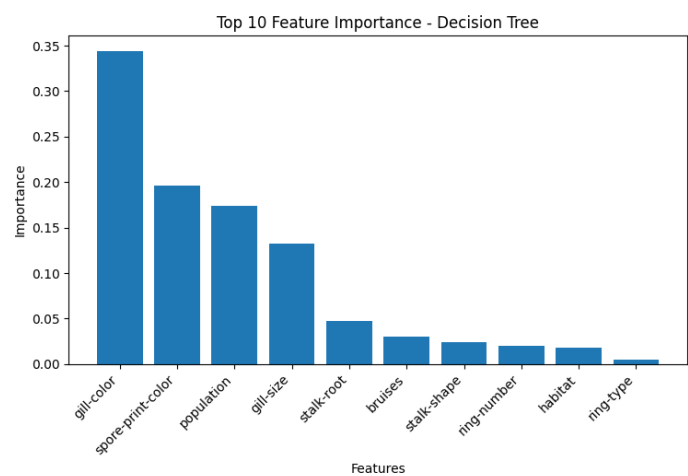


Fig. 4. Feature Importance

## V. CONCLUSION AND FUTURE WORK

The results demonstrate the efficacy of machine learning for mushroom edibility classification. Random Forest, Decision Tree and XGBoost perform very well giving a perfect prediction in the test set, and with perfect cross-validation scores during training.

The superior performance of ensemble methods (Random Forest and XGBoost) can be attributed to their ability to reduce overfitting and improve generalization by aggregating the predictions of multiple decision trees. The lower performance of Naive Bayes, despite its simplicity, likely stems from its assumption of feature independence, which may not hold true for this dataset.

All models could be further optimised with more detailed hyperparameter optimisation techniques. Future work could explore the application of different feature selection techniques, and investigate new ensemble techniques. Furthermore more data could be added to investigate how it might impact the results. Testing the models on newly collected data from diverse geographical locations and mushroom species would assess their generalizability in real-world scenarios. Creating a web application or mobile app based on these models could provide a valuable resource for the public to aid in mushroom identification.

## REFERENCES

- [1] Li H, Tian Y, Menolli N, et al. Reviewing the world's edible mushroom species: A new evidence-based classification system. *Compr Rev Food Sci Food Saf.* 2021; 20: 1982–2014. <https://doi.org/10.1111/1541-4337.12708>
- [2] S. K. Pal, R. Pant, R. Roy, S. Singh, L. Choudhary and S. Naaz, "Mushroom Classification Model to Check Edibility using Machine Learning," 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2023, pp. 214-217.
- [3] Maurya, P., Singh, N.P. (2020). Mushroom Classification Using Feature-Based Machine Learning Approach. In: Chaudhuri, B., Nakagawa, M., Khanna, P., Kumar, S. (eds) *Proceedings of 3rd International Conference on Computer Vision and Image Processing. Advances in Intelligent Systems and Computing*, vol 1022. Springer, Singapore.
- [4] O. Tarawneh, M. Tarawneh, Y. Sharrab and M. Husni, "Mushroom classification using machine-learning techniques," *AIP Conference Proceedings*, vol. 2979, no. 1, p. 030003, 2023, doi: 10.1063/5.0174721.
- [5] M. S. Ahmed et al., "Comparative Analysis of Interpretable Mushroom Classification using Several Machine Learning Models," 2022 25th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2022, pp. 31-36, doi: 10.1109/ICCIT57492.2022.10055555.