LONDON METROPOLITAN UNIVERSITY

islington college
(इस्लिङ्टन कलेज)

Module Code & Module Title

CU6051NI Artificial Intelligence

25% Individual Coursework

Submission: Milestone 1

Academic Semester: Autumn Semester 2025

Credit: 15 credit semester long module

Student Name: Digdarshan Bhattarai

London Met ID: 23049051

College ID: NP01CP4A230320

Assignment Due Date: 17/12/2025.

Assignment Submission Date: 17/12/2025

Submitted To: Er. Roshan Shrestha

GitHub Link	*https://github.com/DigdarshanB/mushroom-edibility-classification*

I confirm that I understand my coursework needs to be submitted online via MST Classroom under the relevant module page before the deadline for my assignment to be accepted and marked. I am fully aware that late submissions will be treated as non-submission and a mark of zero will be awarded.

Table of Contents

No table of figures entries found.

1. Introduction

Mushrooms are widely consumed across the globe, loved for being a delicious, healthy, food source. However, determining which ones to eat and which ones to avoid is a big hassle. Many varieties of poisonous species appear nearly the same shape, color, and size as the ones that can be consumed, and thus cannot be detected by merely visual inspection. Even a small mistake during identification can result into severe illness or even death.

Figure 1: Visual similarity between ***(a)*** **Chlorophyllum molybdites** *(poisonous) and* ***(b)*** **Macrolepiota procera** *(edible), highlighting the difficulty of identification based on appearance alone. (ResearchGate, 2025)*

The conventional methods of identifying mushrooms have relied on human experience, reference books, or visual examination of physical characteristics. Such approaches are generally subjective, slow and inaccurate. Fortunately, there is an increase in the availability of structured biological data, which opens the prospects of applying computational methods in enhancing accuracy and consistency in the process of classifying mushrooms.

As the amount of structured biological data increases, there is an excellent chance to use AI methods to assist in classifying mushrooms. AI systems are ideal because they can compare data in large amounts in a timely and objective manner and are therefore suitable in safety-sensitive decision-making tasks, such as determining the edibility of mushrooms.

1.1 Aim and Objectives of the Project

The overall goal of the project is to use and assess the suitability of supervised machine learning algorithms to determine whether mushrooms are edible or poisonous, based on an already existing dataset and the objectives are:

- To prepare and investigate the mushroom dataset to learn its features and adaptability to supervised classification.
- To use chosen supervised machine learning algorithms to the mushroom data to make the edibility classification.
- To compare the performance of the algorithms on the standard classification metrics to figure out which approach is most appropriate in this issue.

1.2 Overview of the Dataset and Problem Formulation

In this project we are addressing the problem of mushroom edibility and solving it as a binary classification problem - each mushroom is assigned with either edible or poisonous tag

The data is taken out of the UCI Machine Learning Repository a famous go-to resource in AI research benchmark data, which currently has and maintains 688 datasets, serving students of machine learning and AI. (UCI, 2025)

The dataset consists of 8,124 rows which consists various kinds of mushrooms and 22 columns, having attributes such as:

- Cap shape
- Cap colour
- Odour
- Gill size
- Gill colour
- Habitat

Each mushroom has been classified as either as edible or poisonous, hence the data samples are excellent at supervised learning. The attributes are diverse and rich which enables machine-learning models to identify subtle patterns.

1.3 AI Concepts Used

a) Supervised Machine Learning

Supervised ML, simply put means training a model on known dataset to predict outcomes for unknown data. First step is to train a model, then it can be used to predict new inputs (Verma, 2025).

Supervised learning is especially suitable to this problem as the correct classifications are already present in historical data. The model learns the relationship of feature and outcome and predicts correctly on new samples of the mushrooms, which are not visible.

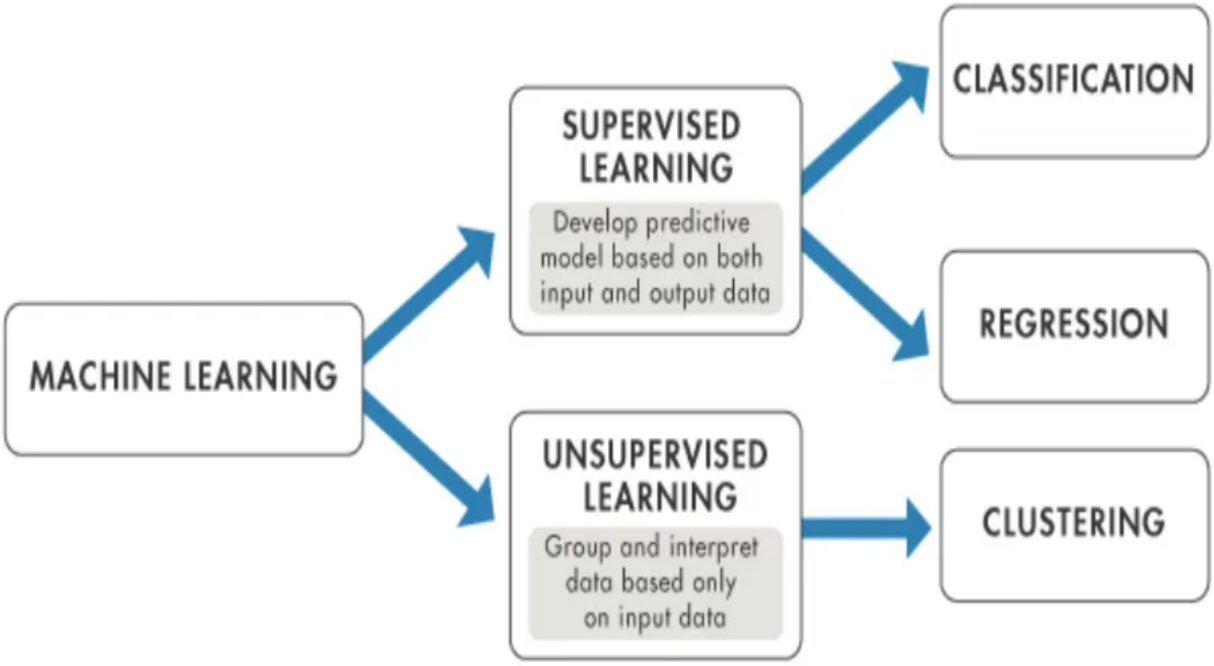

Figure 2: Types of Machine Learning (Verma, 2025)

1.4 Classification Algorithms

In this experiment, we have used and compared various supervised classification algorithms to determine which one is best used to classify a mushroom as being edible. The algorithms are:

a) Logistic Regression

Logistic Regression is used to predict binary outcomes.

b) Naïve Bayes

Naive Bayes predicts probability distribution over a set of classes.

c) Random Forest

Random Forest combines a collection of decision trees and is used to boost accuracy and reduce overfitting.

2. Background

2.1 Problem Context: Why Mushroom Identification is Hard

Correct identification of mushroom is crucial task because some deadly species resemble edible ones in shape, colour, and overall appearance closely just like the example above. People used to rely on expert knowledge, books and manuals which is both time consuming and difficult to apply consistently, especially for beginners.

The difficulty of this issue is especially the fact that this property is not defined by one obvious characteristic that makes the product edible. The own description of the dataset (and its UCI description) clearly points out that no single simple rule can be considered reliable to determine the mushroom edibility: rather, the choice relies on the combination of various traits. UCI machine learning repository.

This is precisely what kind of environment supervised learning might come in: a model can gain the ability to learn many relations between many attributes at once and generalize them.

2.2 Dataset Background and Origin

The dataset used in this coursework is the popular UCI mushroom dataset (Agaricus and Lepiota). It includes the description of hypothetical Herbs which represent 23 species of the gilled mushrooms of the family Agaricus and Lepiota which had been first present in The Audubon Society Field Guide to North American Mushrooms (1981).

Instances: 8,124

The predictor attributes include: 22 (all nominal/categorical).

Label: target: It is edible (e) or poisonous (p).

Distribution of class:

- edible = 4,208 (51.8%)
- poisonous = 3,916 (48.2%)

In the data set, it is also mentioned that species may be definitely edible, definitely poisonous, or unknown/not recommended, and the unknown/not recommended category is grouped in with the toxic category. It becomes significant as it influences us to perceive the label of poisonous in a certain way (this also involves cases of do not eat, but not necessarily being medically poisonous).

2.3 Research on Mushroom Classification

Mushroom Edibility prediction has been studied by researchers mainly as supervised classification problem, with the goal to predict edible vs unsafe mushroom from recorded traits. The UCI mushroom dataset is used for this purpose because it is stated that there is not any particular rule for finding out edibility, meaning many features should be considered together (UCI ML Repository, 1987) .

2.3.1 How mushrooms are identified usually

Traditionally, mushrooms have been characterised by observable characteristics including cap shape/colour, gill features, stalk features, spore print colour, habitat and smell. This method is effective with experts, though may be challenging with beginners since numerous edible and toxic mushrooms are similar. This is why scientists have considered the application of AI to help the process of the decision making based on learning patterns of the recorded mushroom characteristics.

2.3.2 Why AI is used for this problem

The edibility of mushrooms is an AI study that focuses on mushroom edibility as a supervised classification problem: in the training procedure, examples that specify the correct label (edible or poisonous) are provided. This is helpful since it is a combination of traits that classify edibility rather than an individual property. The description of the UCI data also indicates that it cannot be determined whether the item is edible using a single simple rule and because of this, multi-feature machine learning models can be used.

2.4 Review of Existing Work on Mushroom Edibility Prediction

a) Prior Use of Random Forest/Decision Trees on Mushroom Datasets

Tree-based methods are widely used because they represent step by step decisions using categorical features. To put it into perspective, if the odour of mushroom is x, then y, and so on.

In a study which compares Random Forest and REP Tree on the UCI mushroom dataset, the study reports high performance with Random Forest achieving 100% accuracy and REP Tree achieving slightly lover but still high results.

This suggests that for this UCI dataset, tree-based models often perform well because they capture interactions between traits. That said, the scores should still be interpreted carefully as real-world identification can involve uncertain or missing data too (Nepal Journals Online, 2022).

Nepal Journal of Mathematical Sciences (NJMS)
ISSN: 2738-9928 (online), 2738-9812 (print)
Vol. 3, No. 1, 2022 (February): 111-116
DOI: https://doi.org/10.3126/njmathsci.v3i1.44130

Research Article
Received Date: December 25, 2021
Accepted Date: February 24, 2022
Published Date: February 28, 2022

Mushroom Classification using Random Forest and REP Tree Classifiers

Nawaraj Paudel[1] and Jagdish Bhatta[2]

[1,2]*Central Department of Computer Science and IT, Tribhuvan University, Kathmandu, Nepal*
Email: [1]*nawarajpaudel@cdcsit.edu.np*, [2]*jagdish@cdcsit.edu.np*
Corresponding Author: Jagdish Bhatta

Abstract: *Mushroom is a reproductive structure produced by some fungi that has a high level of protein and a rich source of vitamin B. It aids in the prevention of cancer, weight loss, and immune system enhancement. There are numerous thousands of mushroom species within the world and a few are edible and a few are noxious due to noteworthy poisons on them. Hence, it is a vital errand to distinguish between edible and harmful mushrooms. This paper focuses on comparing the performance of two tree-based classification algorithms, Random Forest and Reduced Error Pruning (REP) Tree, for the classification of edible and poisonous mushrooms. In this paper, mushroom dataset from UCI machine learning repository has been classified using Random Forest and REP Tree classifiers. The evaluation of these two algorithms using accuracy, precision, recall and F-measure shows that the Random Forest outperforms REP Tree algorithm with value of 100% for accuracy, precision, recall and F- measure. The performance of Random Forest is 100% and is better with respect to REP Tree classifier.*

Keywords: *Mushroom Dataset, Random Forest, REP Tree, 10-fold cross-validation Confusion Matrix.*

Figure 3: Random Forest vs REP Tree (Poudel, Bhatta, 2022)

4. Experiments and Results

The two classification algorithms were executed on the mushroom dataset using 10-folds cross-validation for the classification of mushrooms based on their class labels. The table below shows confusion matrix of the classification report that has been obtained after testing Random Forest algorithm.

Table 1 Confusion Matrix of Random Forest Algorithm

		Predicted class		
		poisonous	edible	Total
Actual Class	Poisonous	3916	0	3916
	Edible	0	4208	4208
	Total	3916	4208	8124

The table below shows confusion matrix of the classification report that has been obtained after testing REPT Tree algorithm.

Table 2 Confusion Matrix of REP Tree Algorithm

		Predicted class		
		poisonous	edible	Total
Actual Class	Poisonous	3916	0	3916
	Edible	2	4206	4208
	Total	3918	4206	8124

Based on the classification reports shown in Table 1 and Table 2, the calculated summary performance result for the comparison of two algorithms applied on mushroom dataset is shown in the table below. The accuracy, precision, recall and F-measure value are shown is the average of precision, recall and F-measure for both categories.

Table 3 Performance result of two algorithms

Algorithm	Accuracy	Precision	Recall	F-measure
Random Forest	100%	100%	100%	100%
REP Tree	99.98%	99.95%	100%	99.97%

It is clearly seen that the accuracy, precision, recall, and F-measure values of Random Forest is 100% and that of REP Tree is 99.98%, 99.95%, 100%, and 99.97% respectively.

Figure 4: Result of the study (Paudel, Bhatta, 2022)

Advantages:

- Can interact with categorical features well.
- Can often attain extremely high benchmark accuracy on this dataset.
- Offer feature-importance style methods of information (in particular, with Random Forest).

Disadvantages:

- Random Forest is less interpretable than a single decision tree.
- Model behaviour may seem too good to be true on benchmark data and should be carefully examined.
- The behaviour of the models can vary.

b) Prior Use of Naïve Bayes/ Logistic Regression on Mushroom Datasets

Some works use Naïve Bayes and Logistic Regression, and these models are fast, easy to compare and provide a 'reference level' of performance which is why these models are very useful.

A study by Sulistianingsih and Martono (2025) that explains several supervised learning models that are used to classify mushrooms edibility using the UCI Mushroom dataset (8,124 instances and 22 attributes). Both Random Forest and one of their Stacking models are found to have 100% accuracy whereas Naive Bayes has significantly lower-performance (59.8% accuracy) due to the strong assumption of conditional independence by Naivete Bayes that the authors claim is very poor in capturing the interactional nature of mushroom traits.

In general, this work justifies the adoption of Logistic Regression and Naive Bayes as benchmarks and recommends the adoption of tree based / ensemble models to this field since they can represent the interactions between features. (ResearchGate, 2025)

J-INTECH (Journal of Information and Technology)
Accredited Sinta 4 Ministry of Higher Education, Science and Technology Republic of Indonesia SK No. 10/C/C3/DT.05.00/2025
E-ISSN: 2580-720X || P-ISSN: 2303-1425

J-INTECH Journal of Information and Technology

Analysis of the Effectiveness of Traditional and Ensemble Machine Learning Models for Mushroom Classification

Neny Sulistianingsih[1*], Galih Hendo Martono[2]

[1,2]Departement of Computer Science, Master Program, Universitas Bumigora, Ismail Marzuki St. Mataram, Indonesia

Keywords
Bagging; Ensemble Learning; K-Nearest Neighbors; Mushroom Classification; Random Forest; Stacking; Voting Classifier

***Corresponding Author:**
neny.sulistianingsih@universitasbumigora.ac.id

Abstract
The classification of edible versus poisonous mushrooms presents a critical challenge in the domains of applied biology and public health, particularly due to the serious implications of misidentification. This research employs the UCI Mushroom Dataset to evaluate and compare the effectiveness of several machine learning models, including traditional algorithms like Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors and Naïve Bayes, as well as advanced ensemble techniques such as Stacking and Voting Classifier. Notably, both Random Forest and Stacking achieved flawless accuracy, reaching 100%, underscoring the high predictive capacity of these models in complex categorical scenarios. Conversely, Naïve Bayes exhibited significantly weaker performance—achieving only 59.8% accuracy—likely due to its underlying assumption of feature independence, which does not hold for this dataset. The ensemble learning approaches, including the combination of Stacking and Bagging, not only preserved but also enhanced model robustness and generalization. These methods effectively leverage the complementary strengths of individual learners to yield more accurate and stable predictions while mitigating overfitting risks. Comparative analysis with previous research confirms the consistency of these findings and reinforces the viability of ensemble strategies for handling intricate classification tasks. Overall, this study highlights the importance of algorithm selection tailored to data characteristics and supports the use of ensemble learning to boost predictive reliability.

Figure 5: Traditional vs Ensemble models (Sulistianingsih & Martono, 2025)

Table 5. Classification Results with Traditional Methods

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.841	0.842	0.841	0.842	0.913
Decision Tree	0.998	0.998	0.998	0.998	0.998
Support Vector Machine	0.995	0.995	0.995	0.995	0.999
K-Nearest Neighbors	1.000	1.000	1.000	1.000	1.000
Naive Bayes	0.598	0.787	0.598	0.550	0.835

Figure 6: Result of Baselines vs ensembles (Sulistianingsih & Martono, 2025)

Advantages:

- Logistic Regression is easy and easily explained as a baseline.
- Naive Bayes is fast and is usually used to form the basis of categorical classification.
- Both yield a clear baseline with which the ensembles can be evaluated as adding meaningful contributions.

Disadvantages:

- A Logistic Regression can fare poorly when the boundary of the decision is far from linear.
- Naive Bayes can fail badly when the independence properties are broken.
- It has been found that this is the case in this dataset.

c) Prior Use of Multiple Classifiers on Mushroom Datasets

Besides individual model comparisons, several studies also compare multiple supervised classifiers on the UCI mushroom dataset. The feature selection employed by Tank (2021) is the Principal Component Analysis (PCA) and various models such as Logistic Regression, Decision Tree, KNN, SVM, Naive Bayes, and Random Forest were tested. The paper provides a comparison of model performance in ROC-based performance measure, which justifies the notion that prediction of mushroom edibility is generally considered a benchmark classification task where many algorithms are assumed to be tested simultaneously and then one algorithm decided to be applied.

MINIMAL DECISION RULES BASED ON THE APRIORI ALGORITHM †

María C. FERNÁNDEZ*, Ernestina MENASALVAS*, Óscar MARBÁN*
José M. PEÑA**, Socorro MILLÁN***

Based on rough set theory many algorithms for rules extraction from data have been proposed. Decision rules can be obtained directly from a database. Some condition values may be unnecessary in a decision rule produced directly from the database. Such values can then be eliminated to create a more comprehensible (minimal) rule. Most of the algorithms that have been proposed to calculate minimal rules are based on rough set theory or machine learning. In our approach, in a post-processing stage, we apply the Apriori algorithm to reduce the decision rules obtained through rough sets. The set of dependencies thus obtained will help us discover irrelevant attribute values.

Keywords: rough sets, rough dependencies, association rules, Apriori algorithm, minimal decision rules

Figure 7:Use of Multiple Classifiers on Mushroom Datasets (Fernandez, 2001)

Advantages:

- A fair big picture comparison.
- Facilitates easier justification of results due to benchmarking of performance both against simple control baselines and stronger models
- Allows to understand what algorithms are most consistent across evaluation measures.

Disadvantages:

- Preprocessing options can bias results (e.g. the choice of encoding algorithm and the use of PCA)
- Comparison can be unreliable with models not tuned equally
- Interpretability may be compromised by dimensionality-reduction techniques such as PCA,

2.5 Key Takeaways from Existing Works

Much of the previous research indicates that models of mushroom edibility can acquire clear patterns of choices based on categorical characteristics. Various studies identify a consistent theme which is that tree-based methods especially the random forest are frequently able to perform to an extremely high benchmark score on the UCI mushroom data. An example by Paudel and Bhatta (2022), where they contrasted Random Forest with a pruned decision tree (REP Tree) had 100 percent accuracy with Random Forest with a REP Tree only slightly lower, indicating that tree-based logic is well-polarized with the interactions between the mushroom properties and the forecast of edibility.

Simultaneously, Logistic Regression and Naive Bayes are often used as a baseline model in studies. These baselines assist in interpreting the results, in that, they indicate the presence of meaningful improvement in the more sophisticated procedures. According to Sulistianingsih and Martono, (2025) nothing can be more accurate than Random Forest/stacking, and much less accurate than Naïve Bayes, which confirms the fact that mushroom characteristics tend to be interactive and that models based on simplistic assumptions are possibly not appropriate.

3. Solutions

3.1 Proposed Solution

The proposed solution for the project is a supervised ML pipeline that predicts whether a mushroom is edible or not based on recorded categorical attributes. The solution is designed to be implemented in Jupyter Notebook using Python as the programming language. The solution follows a structured workflow to evaluate every step fairly.

Three classification algorithms have been used – Logistic Regression, Naïve Bayes, and Random Forest for comparative analysis. These models present different levels of complexity. Logistic Regression provides a basic linear baseline for the project. Naïve Bayes is probability-based baseline which is best suited for categorical data and Random Forest is an ensemble model that can capture feature interactions. The final model recommendation will be based on evaluations of the train/test data.

3.2 System Workflow

1. **Load and Structure Data**: .data file is loaded into a table and feature names are applied from .names file.
2. **Data Preprocessing**: Categorial attributes are converted into machine-readable form and unknown values are handled.
3. **Dividing the dataset**: Dataset is divided into training and testing sets.
4. **Model Training**: Train Logistic Regression, Naïve Bayes, and Random Forest on the training dataset.
5. **Prediction**: Use trained models to predict class labels on the test set.
6. **Evaluation and Comparison**: Compare models using standard classification metrics and justify the most suitable model.

3.3 Description of AI Algorithms used

a) Logistic Regression

Logistic Regression The algorithm is supervised learning algorithm that is used in the binary classification. It will learn a set of weights of the input features and create a probability score that a sample is of a particular class (e.g., poisonous). The purpose of using Logistic Regression as a baseline model in this project is that it is easy, quick to train and can be used as a reference point to other more complicated methods. It is also useful in demonstrating the level of improvement that is achieved in switching it to non-linear relationship models.

Advantages:

- Easy to describe and simple to understand making it a robust baseline model.
- Trains quickly and will be useful in cases where the classes are separable in a relatively linear manner.

Disadvantages:

- Fail to perform well when the relationship between features and class is either non-linear or feature interactions.
- Does not work well when the decisions involve a combination of traits, compared to tree-based models.

b) Naïve Bayes

Naive bayes is a probabilistic classifier which operates under the Bayes theorem. It uses the estimates of the probability of observed values of features to each class and determines the most probable class. The naive component involves the fact that features play a separate role in the eventual decision given the knowledge of the class. Although this does not necessarily apply to real data, Naive Bayes can be effective with categorical data that are structured and is computationally economical. It has been made part of this report since it is a common foundation on which the classification activity can be carried out and offers a handy comparison to Logistic Regression and Random Forest.

Advantages:

- It is extensively fast to train and predict, which is why it is helpful in the creation of a fast baseline.
- Works fairly well with limited training information and gives prediction in terms of probability.

Disadvantages:

- Usually difficult to assume, that features are independent over given the class.
- As feature interactions are important (as is the case with real classification tasks), performance may also fail.

c) Random Forest

Random Forest is an ensemble machine learning procedure, which creates a variety of decision trees and integrates the results of these trees to provide a resulting prediction. A tree comes up with a set of decision rules by partitioning the data according to the feature values, and the forest gets the predictions (usually by a majority vote) to enhance reliability. Random Forest is also included, as it can deal with complex patterns and interactions between features, as the situation in mushrooms identification is expected to be (edibility is often a combination of factors).

Advantages:

- Can interact with categorical features well.
- Can often attain extremely high benchmark accuracy on this dataset.

Disadvantages:

- Random Forest is less interpretable than a single decision tree.
- Model behaviour may seem too good to be true on benchmark data and should be carefully examined.

3.4 Pseudocode for the Overall Project

Input: Mushroom dataset (.data) and feature names (.names).

Output: Trained models, results evaluation, best model recommendation.

START

1. IMPORT essential libraries

2. LOAD the mushroom dataset from .data file

3. NAME columns using .names file

4. PRE-PROCESS dataset:

    4.1. Handle missing values

    4.2. TRANSFORM categorical attributes into numeric form

5. SPLIT data into features X and label y.

6.SPLIT dataset into training set and testing set.

7. TRAIN Logistic Regression model using training set.

8. TRAIN Naïve Bayes model using training set.

9. TRAIN Random Forest model using training set.

10. PREDICT labels on the test set using each model.

11. EVALUATE each model using standard classification metrics.

12. COMPARE model results and identify best performing model.

END

3.5 Pseudocode for Each Algorithm

3.5.1 Pseudocode – Logistic Regression

START

1. INPUT: X_train, y_train

2. INITIATE Logistic Regression Classifier

3. FIT classifier using (X_train, y_train)

END

3.5.1 Pseudocode – Naïve Bayes

START

1. INPUT: X_train, y_train

2. INITIATE Naïve Bayes Classifier

3. LEARN class probability distribution from y_train

4. LEARN conditional probabilities of each feature from X_train

5.OUTPUT: Trained Naïve Bayes Model

END

3.5.1 Pseudocode – Random Forest

START

1. INPUT: X_train, y_train

2. INIT Random Forest Classifier

3. FOR each tree in the forest:

    3.1. SELECT random sample of training records

    3.2. TRAIN one decision tree on sampled records

4. COMBINE all trees into a single model

5. OUTPUT: Trained Random Forest Model

END

3.6 Flowchart – Overall System

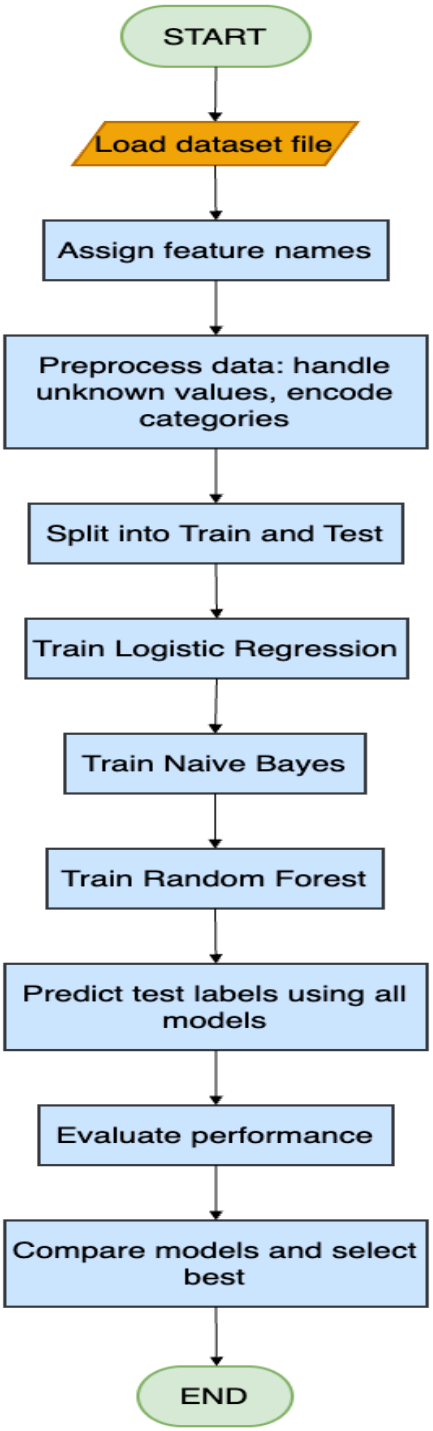

Figure 8: Flowchart of the Overall System

3.6.1 Flowchart - Logistic Regression

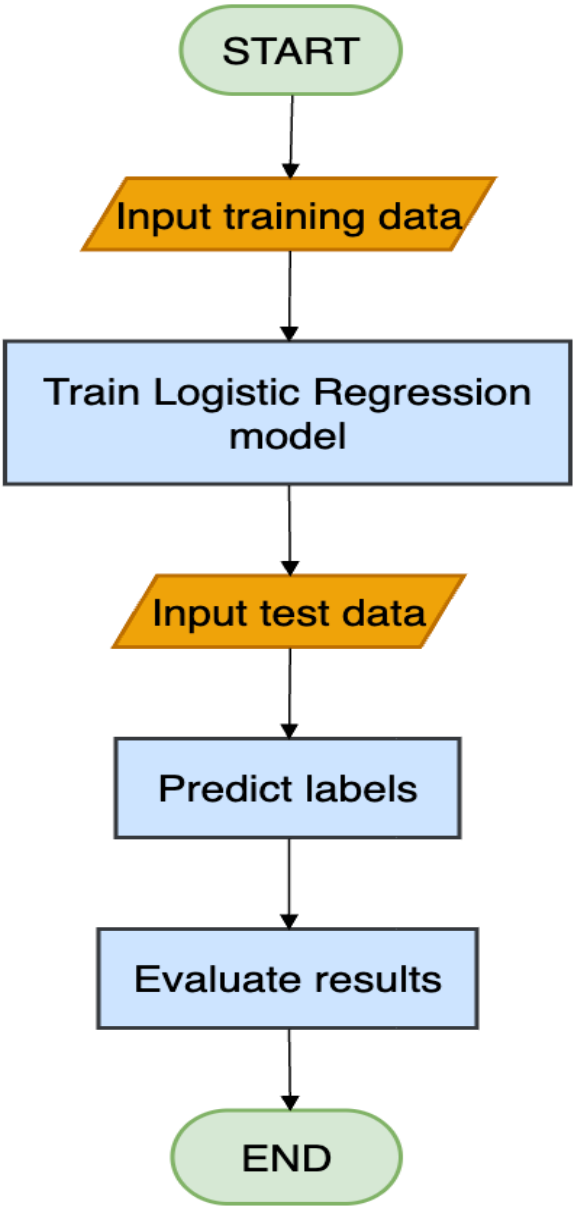

Figure 9: Flowchart for Logistic Regression

3.6.2 Flowchart – Naïve Bayes

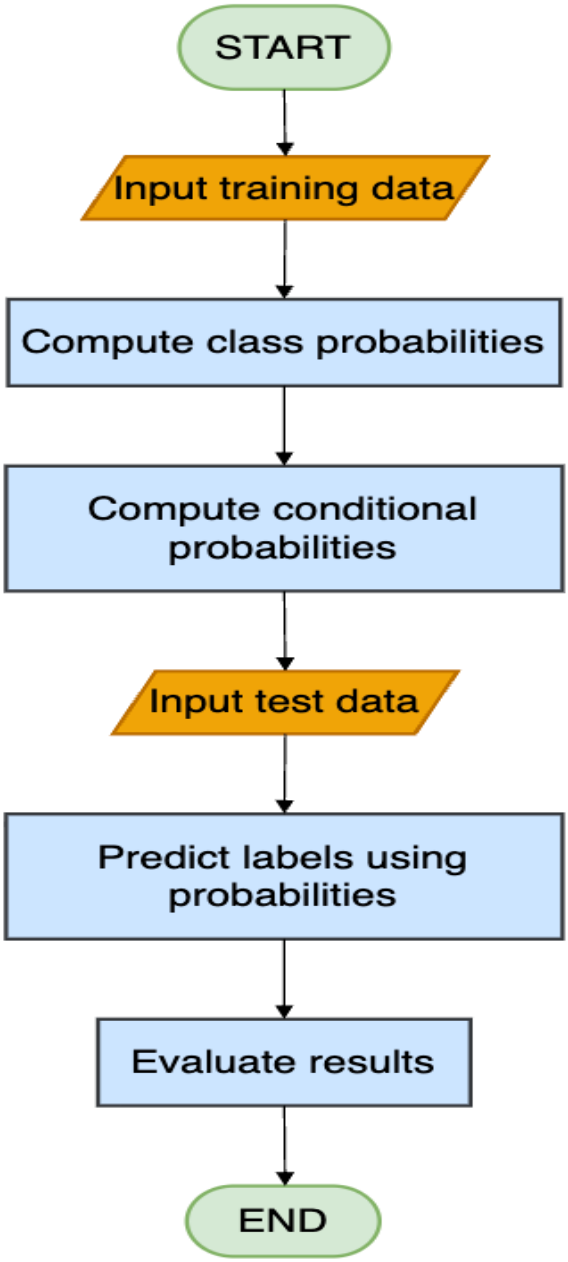

Figure 10: Flowchart for Naïve Bayes

3.6.3 Flowchart – Random Forest

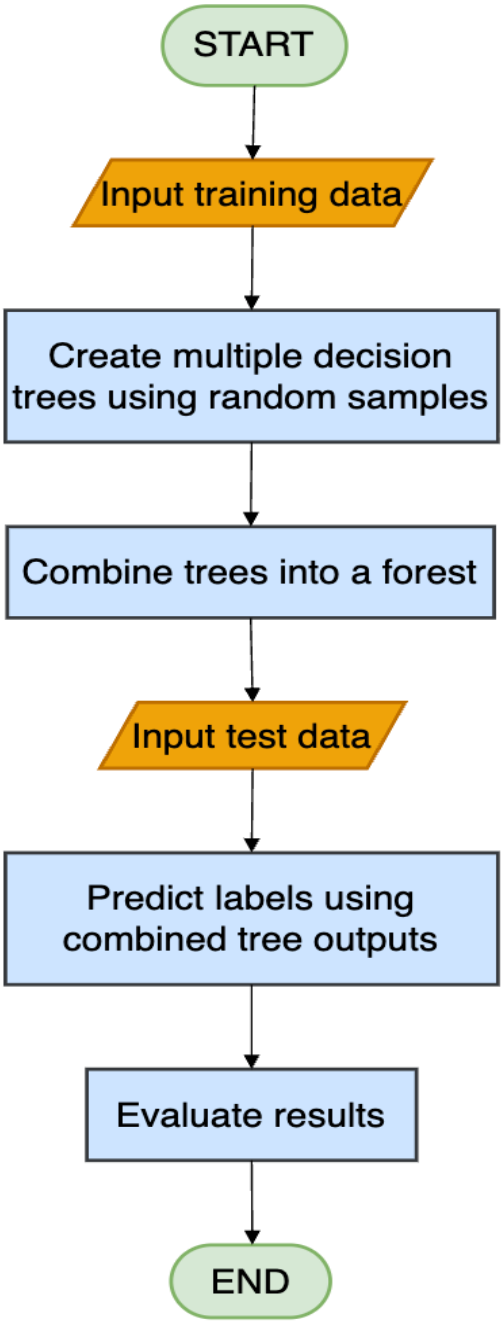

Figure 11: Flowchart for Random Forest

4. Conclusion

4.1 Analysis of the Work

In this work, a full implementation-complete plan of solving mushroom edibility prediction as a supervised learning classification problem was developed. It started with the real-world situation, which is that edible and poisonous mushrooms can be very similar in appearance and therefore no one criterion can help one differentiate between them, this is because checking several attributes is the only way one can be sure of what they are looking at. It was then analysed based on the chart and applicability to supervised learning, and it was observed that the dataset consists of categorical values of mushroom features that are easy to observe and a binary value on whether a mushroom will be edible or poisonous/unsafe. Another significant note by the dataset description is the fact that the values of unknown/not recommended are combined into the unsafe category, and thus the task may be best understood as trying to differentiate between the cases that are safe to consume and cases that are not safe to consume, which is again the aspect of safety-first approach. Subsequently, a study was conducted to inspect existing literature on the topic of mushroom classification that revealed that this issue is typically addressed as a benchmark supervised classification problem involving several models, being evaluated by the same evaluation plan. Based on this analysis three algorithms have been chosen reflecting different modelling strategies: two contrasted modelling strategies Logistic Regression and Naive Bayes to provide transparency, and a stronger ensemble modelling strategy, the Random Forest which is able to capture interaction between multiple categorical variables. This report then digested these research findings into a systematized solution design by clearly defining each implementation process within the Python language: loading the datasets files, giving the names to the features, doing the preprocessing (handling the unknown values and encoding nominal categories), splitting into training and testing sets, training all the models, making the predictions, and performance evaluation. To make sure that the solution is readable and repeatable, the solution was represented in system-level pseudocode and flow charts, one flowchart per algorithm, to allow the implementation to be systematically implemented in a Jupyter Notebook without confusion.

4.2 How the Solution Helps to Address Real-World Problems

Mushroom identification is safety-critical in the real world, since errors may prove disastrous, and very often identifications are made by non-experts in the state of uncertainty. In the suggested AI-based solution, safer decision-making is facilitated through learning trends on a range of observable characteristics and implementing such reasoning continuously, instead of utilizing trial-of-thumb opinion only. It can best be considered decision support, and the comparative design of the report can guarantee that the ultimate recommendation is not based on assumptions but is evidence based.

The main approaches the solution may contribute to solve the real-life problem:

Consistency: The logic of feature-combination is used consistently, minimizing variations.

Multi-trait thinking: The model does not rely on a single (rule) but rather a combination of evaluations, as is the case with the determination of edibility.

Evidence-based Model Choice: The Logistic Regression, Naïve Bayes and the Random Forest will be compared and assessed under an equal preprocessing and train/test split assuring that the comparison remains fair.

Safety-oriented Assessment: The inspection cannot be limited to overall performance but to identify high-risk errors (such as unsafe mushrooms as to be consumed).

Conservative Meaning of the Unsafe: Since the dataset is extended to include the unknown/not recommended category together with the poisonous category, the classification is consistent with a conservative do not eat unless you are sure attitude.

4.3 Further Work

• Once the project has been developed, the system might be tested on new mushroom data to determine whether it can generalise beyond the UCI benchmark data.

• The models could be generalized to deal with real world uncertainty, including missing feature information, noisy inputs, or missing attribute values as ideal identification is hardly ever the case in practice.

• More sophisticated models might be included in the comparison to determine whether they might provide any benefits over the three models applied in this project.

• An actual system that would allow users to input their mushroom attributes and give feedback in a clear manner as it is decision support would be handy.

• Lastly, the usability can be enhanced by displaying a confidence indicator and a warning-oriented design that can minimize the interpretation of predictions in an unsafe manner.

5. References

Bibliography

GeeksForGeeks, 2025. *Machine Learning.* [Online] Available at: https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/ [Accessed 11 December 2025].

Nepal Journals Online, 2022. *Article.* [Online] Available at: https://www.nepjol.info/index.php/njmathsci/article/view/44130 [Accessed 11 December 2025].

ResearchGate, 2025. *Figure.* [Online] Available at: https://www.researchgate.net/figure/Similar-looking-mushrooms-are-hard-to-differentiate-a-A-picture-of-Chlorophyllum_fig1_382904445 [Accessed 10 December 2025].

ResearchGate, 2025. *Publication.* [Online] Available at: https://www.researchgate.net/publication/393698893_Analysis_of_the_Effectiveness_of_Traditional_and_Ensemble_Machine_Learning_Models_for_Mushroom_Classification [Accessed 12 December 2025].

UCI ML Repository, 1987. *dataset.* [Online] Available at: https://archive.ics.uci.edu/dataset/73/mushroom [Accessed 11 December 2025].

UCI, 2025. *Home.* [Online] Available at: https://archive.ics.uci.edu/ [Accessed 10 December 2025].

Verma, A., 2025. *Solutions.* [Online] Available at: https://medium.com/@amitvsolutions/supervised-machine-learning-decoded-from-forecasts-to-decisions-4e57c6b1a0c4 [Accessed 11 December 2025].

Verma, A., 2025. *Solutions.* [Online] Available at: https://medium.com/@amitvsolutions/supervised-machine-learning-decoded-from-forecasts-to-decisions-4e57c6b1a0c4 [Accessed 11 December 2025].