UTM UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF COMPUTING
UTM Johor Bahru

SECB3203-01(PROGRAMMING FOR BIOINFORMATIC)

Semester 01 2025/2026
Section 01

Project Proposal

Faculty of Computing

Group 7

Dataset:
https://www.kaggle.com/datasets/uciml/mushroom-classification

---

Student Name	Matric Number
NURUL NATALIA BINTI ROSNIZAM	A23CS0165
ADLINA NARISYA BINTI ISMAIL	A23CS0033
NAZATUL NADHIRAH BINTI SABTU	A23CS0144

Table of Contents

1.0 Introduction

Mushrooms are widely consumed as a nutritious food source, but many wild species are poisonous and pose serious health risks when misidentified. The dataset used in this project is derived from *The Audubon Society Field Guide to North American Mushrooms (1981)* and contains descriptions of hypothetical samples representing 23 species of gilled mushrooms belonging to the *Agaricus* and *Lepiota* families.

Each species in the dataset is classified as definitely edible or definitely poisonous. For safety purposes, the "unknown" category has been merged with the "poisonous" class. Importantly, the Field Guide emphasizes that there is no simple universal rule to determine the edibility of a mushroom unlike plants such as Poison Ivy, where simple identification rules exist ("leaflets three, let it be"). It is because edible and poisonous mushrooms often share similar morphological features such as cap color, odor, gill structure, and stalk characteristics, manual identification is difficult for non-experts. This makes mushroom classification a challenging biological problem, highlighting the need for computational approaches to support safe identification and prevent accidental poisoning.

1.1 Problem Background

Mushrooms are eaten worldwide for food, nutrition, and sometimes medicine, but not all mushrooms are safe. Many poisonous mushrooms look very similar to edible ones, which often causes people to misidentify them. This can lead to serious health problems, including poisoning or even death. Because of this, it is important to understand the features that separate safe mushrooms from dangerous ones.

Researchers use data to study these differences. One useful dataset is the Mushroom Classification dataset, which contains information on more than 8,000 mushrooms. Each mushroom is labeled as edible or poisonous and includes details such as cap shape, color, odor, and gill size. These are features normally used to identify mushrooms in real life.

While many wild mushrooms are edible, a large number are toxic, with some being deadly. Distinguishing between edible and poisonous species is challenging and often requires expert knowledge. Moreover, misidentification can lead to severe illness or even death. Therefore, there is a critical need for a reliable and accessible tool to help individuals accurately classify mushrooms and ensure their safety

1.2 Problem Statement

The primary challenge in mushroom classification is the high risk associated with human error in prediction, as manual identification tend to make mistakes due to identical visualization between mushroom species. This problem always occurs in real-life situations which lead to life-threatening scenarios. On the other hand, there is a computational challenge where the classification fully depends on multiple categorical features that often overlap between classes, occurs when samples from different classes share similar characteristics. Thus, this project aims to address both human and computational challenges by developing an automated machine learning model that can classify mushrooms accurately and safely based on their characteristics. Therefore, reducing the risk of poisoning.

1.3 Objectives

1. To prepare the Mushroom Classification dataset by conducting data wrangling, including removing the useless features.
2. To perform exploratory data analysis to identify the most significant features for machine learning training.
3. To develop and train multiple models including Random Forest, Decision Tree, and Logistic Regression.
4. To evaluate and compare the performance of these models based on accuracy, AUC, and f1-score to find the optimal one.

1.4 Scopes

1. **Data Used:**

   This project uses the Mushroom Classification dataset from Kaggle, which contains 8,124 samples of mushrooms with categorical features and a class label indicating whether each mushroom is edible or poisonous.

2. **Techniques to Be Used:**

   The project applies basic exploratory data analysis, descriptive statistics, and simple data visualizations to understand the distribution and patterns in the dataset.

3. **Methodology:**

   The study involves loading and inspecting the dataset, cleaning the data, exploring each feature, visualizing important patterns, and summarizing the findings in a clear report.

4. **Limits of the Research:**

   The analysis is limited to the features provided in the dataset, which may not represent all mushroom species in real life.

5. **Expected Outcomes:**

   The project aims to provide a clear understanding of the class balance between edible and poisonous mushrooms, identify key feature patterns, and present meaningful visualizations that highlight important insights from the data.

1.5 Conclusion

In conclusion, this project demonstrates the effectiveness of machine learning techniques in addressing the critical problem of mushroom classification. By utilizing the Mushroom Classification dataset, the study successfully applied data preprocessing, exploratory data analysis (EDA), and multiple classification models to distinguish between edible and poisonous mushrooms. The EDA phase provided valuable insights into class distribution and feature relevance by analyzing key attributes such as odor, cap color, gill size, and stalk characteristics, which helped guide feature selection and improve model reliability.

The results show that machine learning models can achieve high classification accuracy despite the complexity and similarity of mushroom features. Through model comparison and performance evaluation, this project highlights the potential of computational approaches to reduce human error and improve safety in mushroom identification. Although limited to the available dataset, the study provides meaningful insights and establishes a strong foundation for future research and real-world applications in automated biological classification.

1.6 References

Chiraag K V. *Mushroom Classification Project part 2 — Exploratory Data Analysis (EDA)*. Medium. June 2, 2021. https://chiraagkv.medium.com/mushroom-classification-project-part-2-exploratory-data-analysis-eda-cc0d96da574c (chiraagkv.medium.com)

Amit. *Exploratory Data Analysis (EDA) — Datasets*. Medium. https://medium.com/@amit25173/exploratory-data-analysis-eda-datasets-7f5115864b7d