FACULTY OF COMPUTING
UTM Johor Bahru

SECB3203-01(PROGRAMMING FOR BIOINFORMATIC)

Semester 01 2025/2026
Section 01

Project Proposal

Faculty of Computing

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 3

1.1 Problem Background 3

1.2 Problem Statement 4

1.3 Objectives 4

1.4 Scopes 5

1.5 Conclusion 6

1.6 References 6

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, definitely poisonous, or of unknown edibility and not recommended. 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.

Although many studies use this dataset for machine learning, fewer focus on simple exploratory data analysis (EDA). EDA is important because it helps us understand the data, see how many mushrooms are edible or poisonous, and observe how their features vary. Visual charts can also help reveal patterns more clearly.

The main goal of this project is to perform a detailed EDA on the Mushroom Classification dataset. The analysis will look at the distribution of edible versus poisonous mushrooms, identify useful patterns in their features, and present the results with visualizations. This will help build a better understanding of the dataset and support future work in mushroom identification and safety.

1.2 Problem Statement

Many people enjoy exploring forests and nature, but not everyone has the knowledge to safely identify mushrooms. For first time forest visitors or inexperienced foragers, poisonous mushrooms can easily be mistaken for edible ones because they often look very similar. This can lead to dangerous situations, including accidental touching, smelling, or eating the poisonous mushrooms which may cause serious health problems or even death.

To help study these differences, researchers commonly use the Mushroom Classification dataset, which contains information about various mushroom features such as cap shape, color, odor, and gill size. Important details such as how many mushrooms are edible or poisonous and how their features different are often not clearly explained.

Therefore, there is a need for a detailed exploratory data analysis of the Mushroom Classification dataset. By exploring the data and examining key features differences, this aims to help understanding of mushroom characteristics and identification where it is safe or not, especially for people with little or no experience in the forest.

1.3 Objectives

1. To analyze the distribution of edible and poisonous mushrooms.

2. To explore and compare mushroom features such as cap shape, odor, gill size to identify the patterns and differences between edible and poisonous classes.
3. To generate visualization that clearly presents relationships patterns within the dataset.
4. To provide understanding that guides future work on mushroom classification and ensuring safety research.

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 project focuses only on EDA and does not include machine learning modeling, and 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, the exploratory data analysis of the Mushroom Classification dataset revealed important patterns distinguishing edible and poisonous mushrooms. By examining features such as cap shape, color, gill structure, and odor, the analysis showed that while some traits are strongly associated with edibility, many characteristics overlap between classes, making manual identification challenging for non-experts. Visualizations helped highlight these patterns and clarified feature relationships across the dataset.

Overall, this study emphasizes the importance of data-driven approaches for mushroom classification. Understanding feature distributions and class distinctions provides a foundation for building accurate predictive models and promotes safer practices in mushroom identification, helping to prevent accidental poisoning.

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