**Architecture Document**

**for Mushroom Classifier Application**

Revision date: 19/07/2024

**1. Introduction**

1.1 Project Overview

**2. Data Acquisition and Preprocessing**

2.1 Data Source

2.2 Data Description

2.3 Data Cleaning

**3. Exploratory Data Analysis (EDA)**

3.1 Understanding Feature Distributions (visualizations)

**4. Model Selection and Training**

4.1 Choosing a Classification Algorithm

4.2 Model Training Process

**5. Conclusion**

8.1 Summary of Findings and Achievements

**1. Introduction**

1.1: Project Overview

This is a project whose aim is to predict which mushroom is poisonous or not using Machine learning and its description (e.g. cap-shape colour etc).

**2. Data Acquisition and Preprocessing**

2.1: Data Source

Donated to the UCI Machine Learning Repository in 1987.

2.2: Data Description

This dataset details hypothetical samples of 23 gilled mushroom species (Agaricus and Lepiota families) based on The Audubon Society Field Guide (1981). The dataset classifies each sample as edible, poisonous, or of unknown edibility.

2.3: Data Cleaning

To clean the data, we need to import the csv file using pandas. We also need to check if there are any duplicate rows or any empty cells.

**3. Exploratory Data Analysis (EDA)**

3.1: Understanding Feature Distributions (visualizations)

For Visualisations, we use matplotlib to plot bar charts. It helps us to get an understanding of how data is spread out

**4. Model Selection and Training**

4.1: Choosing a Classification Algorithm

Since machine learning models can't handle text data, we used OneHotEncoding (scikit-learn) to convert categories into numbers. For binary classification (edible vs poisonous), Logistic Regression was chosen.

4.2: Model Training Process

After training the model and was trained on 80% of the data, achieving nearly 99.8% accuracy.

**5. Conclusion**

5.1: Summary of Findings and Achievements

This is document contains the basic way to implement a ML Algorithm