MUSHROOM CLASSIFICATION PROJECT PROPOSAL

PREPARED BY

Sergio Robledo

Mushrooms Inc.

1. Project Overview

Foraging for wild mushrooms, or mushroom picking, is the pastime of many people, and even though it does not seem like a dangerous hobby, consumption of a misidentified mushroom could cause death; it is imperative for the person foraging for mushrooms to know whether the mushrooms they have picked are edible or poisonous. However, many people do not have the time required to become mushroom foragers, because knowing the difference between an edible and fatal mushroom might be in the slightest detail. In an effort to not only give peace to the forager but to possibly save their lives, a reliable machine learning model capable of accurately identifying mushrooms is needed; below I will propose the steps required for such application to come to fruition.

2. Data Acquisition

The mushroom dataset I will use to build such model comes from the UCI Machine Learning Repository and can be found here. The dataset includes 22 categorical features ranging from odor to gill characteristics; it contains 8124 observations labeled as either edible or poisonous. We will preprocess and explore the dataset before building the machine learning model.

3. Plan of Action

We will begin by cleaning the dataset which will mainly consists of handling missing values and transforming it into a format that could be passed to the model; in order to convert the dataset to a compatible format, we will apply one-hot encoding to the categorical features. Once the dataset has been cleaned and preprocessed, we can begin the exploration and visualization part of the project. Exploration and visualization will give us a better idea of what features are important and which ones are not; the insights gained from exploration will help us decide how to proceed with the project. For example, if it is found that a given feature does not provide useful information about the mushrooms' edibility, we can decide to remove it from the dataset because keeping it would only increase the amount of noise passed on to the machine learning model; noise makes it harder for the model to differentiate between important and unimportant trends by increasing the amount of data it needs to filter out before acquiring an acceptable solution. Finally, the last step will be the creation of the predictive model; this step consists of evaluating the performance of multiple

initial models, selecting the best performing one and fine tuning the final model until it achieves an accuracy of 100%. The data will be split into a training and a testing set, and the testing set will not be used in any part during the creation of the machine learning model and only in the final testing phase; making the machine learning model predict whether or not a mushroom is edible or not using never before seen data ensures the observed performance is valid. Further, the reason we need to achieve such a high accuracy score is due to the possibility of poisonous intoxication that could occur from an incorrect edibility classification. The following section will describe the deliverables that will be produced from the conclusion of the project once the desired model performance is reached.

4. Deliverables

Upon the conclusion of the project, two milestone reports, a slide deck, a final report and the code used will have been delivered. The milestone reports will communicate the project's progress; the first milestone report will be created after the cleaning and exploration phase of the project, and the second milestone report will be created after the machine learning model has been created. The final report will summarize the project from start to finish, and the slide deck will be a presentation of the final report. Finally, the code will include all the code used from the project's start to finish; this includes the cleaning, exploration and visualization code as well as a functioning machine learning model with the best performance achieved.