ITCS 3156

Evaluating Mushroom with ML Algorithms

Muxin Feng

1. Introduction

This is the ITCS 3156 individual final project. This study aims to train two models to be used as a foundation for machine learning-based mushroom species differentiation using the online resource available at the UC Irvine Machine Learning Repository. The code for the Jupyter Notebook can be found on Github: https://github.com/muxxxxin-feng/3156-Final-Project.git

1.1. Problem Statement

To maintain safety and prevent any health hazards, it is essential that mushrooms be classified as either edible or potentially hazardous. Mistaking a poisonous mushroom for a food source may result in serious health repercussions such as poisoning.

Machine learning provides a strong approach for automating the categorization of mushrooms because of the wide range of useful data. The goal of this study is to use machine learning techniques to create models that can reliably identify mushrooms as either edible or poisonous based on their category and physical characteristics. This project compares the performance of K-Nearest Neighbors (KNN) and Naive Bayes (NB) models in order to identify the best method for this classification problem while learning more about the structure and significance of the features in the dataset.

1.2. Motivation

One of the most essential elements of our natural surroundings is mushrooms. They play an essential part in ecosystems by communication and nutrient sharing among plants as well as the breakdown of organic materials, which replenishes the soil with nutrients. Humans also like mushrooms as a nutrient-dense food source since they are abundant with vitamins and minerals. In scientific terminology, mushrooms are also known as fungi, and there are numerous kinds worldwide. Some of these species are edible, while others are extremely poisonous. Although mushrooms are a very common food source in western southern China, thousands of people are poisoned by them because they are not properly cooked or eaten because locals are not aware of how to distinguish between a variety of mushrooms.

This dataset is a useful tool for the task of discriminating different species of mushrooms because edible mushrooms are difficult to identify and most mushrooms have similar characteristics. Despite their identical appearance, the majority of mushrooms can be identified by certain traits like color, cap shape, etc. As a result, it is crucial to group different kinds of mushrooms according to particular traits.

1.3. Summary

This study investigates the use of machine learning algorithms to classify mushrooms as either edible or poisonous. K-Nearest Neighbors (KNN) and Naive Bayes (NB) are two different models that were constructed using the UCI Mushroom Dataset, which has a range of categorical characteristics. During the preprocessing stage, categorical data had to be encoded into numerical values and divided into training and testing sets. The main distinctions between edible and poisonous mushrooms were revealed by analyzing feature distributions to comprehend the links between attributes and class labels.

The KNN model achieved higher classification accuracy than the Naive Bayes model, which indicates feature independence. Evaluation measures that emphasized each model's advantages and disadvantages included classification reports and accuracy. Confusion matrices and feature distribution diagrams are two examples of visualization tools that brought additional insight into the dataset's structure and model performance.

2. Data

2.1.Introduce the data

The "Secondary Mushroom" dataset was discovered in the UCI Data Repository. 61069 fictitious mushrooms with caps based on 173 species (353 mushrooms per species) are included in this dataset. The data included three quantitative variables, seventeen nominal variables, and one binary class. Every mushroom is classified as either definitely poisonous, or definitely edible (Wagner).

2.2. Perform basic (visual) analysis

Figure 1: Example of edible and poisonous mushrooms (Emily Stearn)

While there are many various kinds of mushrooms, it is important to know the difference between poisonous and edible mushrooms. Online resources like wikiHow, for instance, state that while most edible mushrooms have tan or brown gills, brightly colored mushrooms, such as red ones, are frequently dangerous (Burba). (Figure 1)

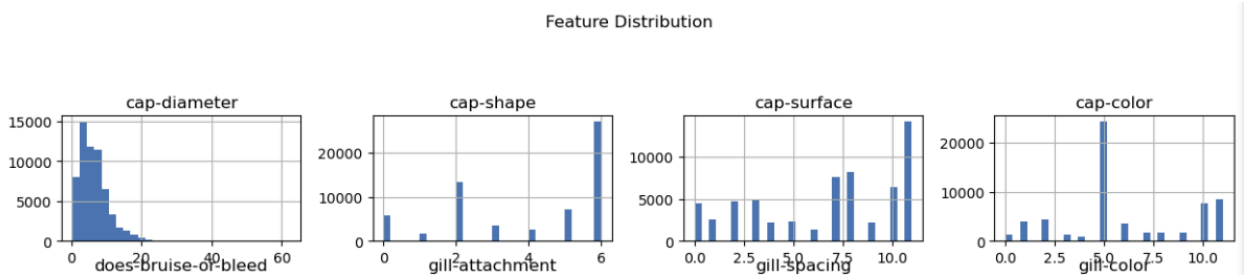

Figure 2: Naive Bayes Model

Figure 2 is an example of the usage of the Naive Bayes Model. There are several categories of the characteristics of mushrooms, such as the cap diameter, cap shape, cap surface, and cap color throughout the dataset. The feature distribution table also clearly stated the number of mushrooms within the dataset for each characteristic. For instance, most of the mushrooms are under 20 cm cap-diameter, and a small number of mushroom has a cap diameter greater than 20 cm.

2.3. Preprocessing

Before applying machine learning models to the dataset, several preprocessing steps need to be performed to prepare the data for analysis. According to the instructions of the UCI Data Repository, this dataset needs to be imported into the Jupiter Notebook, and make sure to install ucimlrepo package. Furthermore, the dataset needs to be fetched based on the ID that the website provided.

3. Methods

This project includes the usage of two models: the Naive Bayes Model and the K-nearest neighbor(kNN) model. The goal of using these models is to identify which model is the better fit for the dataset by checking the accuracy of the model and also to see how the diagrams are plotted by categorizing the characteristics of mushrooms according to the dataset. Both models are similar but much shorter than the homework, and algorithm of the models is based on the previous homework.

3.1. Naive Bayes Model

The Naive Bayes model assumes features are conditionally independent given the target output according to the class slide (Lee and Benedict). It is based on Bayes' theorem as Figure 3

showed below. Throughout the model training, it seemed like it was easy to implement and worked very well with the large dataset. The waiting time was as fast as it could be.

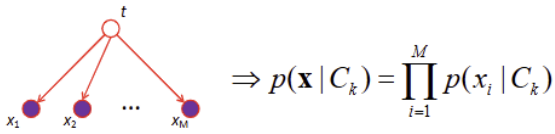

Figure 3: Naive Bayes Model

3.2. KNN Model

The KNN model, known as the k-nearest neighbor model, uses the k neighbors to make a decision instead of a single nearest neighbor (Lee and Benedict). As for processing the dataset, it seems that the KNN model is simple and easy to implement, with no training required, the running time can be an issue when it comes to a large dataset. The waiting time for plots or diagrams is extremely slow.

4. Results

4.1. Explanation of Setup

Classifying mushrooms as edible or poisonous based on their physical and chemical characteristics is the goal of this project. Since many poisonous mushroom species are difficult to identify because they closely resemble edible mushrooms, this classification process is crucial for maintaining public safety. Two distinct class labels are used to train machine learning models and compare their performance: edible as e, which is labeled data that has been identified in the dataset as edible mushrooms; and poisonous as p, to determine which mushrooms are poisonous.

4.2. Test Methods and Results

To evaluate the effectiveness of the machine learning models, the dataset was divided into training and testing subsets using a split. The training set was used to train the models, while the testing set served as unseen data to measure the models' ability to generalize.

For the Naive Bayes Model, the model is implemented using the Gaussian Naive Bayes, which assumes features are conditionally independent given the class label (Lee and Benedict). As for the KNN model, used 5 neighbors as the default hyperparameter in the model to get us to start on the project. Both models calculated significant performance measures, including accuracy, confusion matrix, and classification report as an output so that it easily identified the difference between the edible and poisonous mushrooms.

Both models result in different accuracy throughout the project: Naive Bayes Model surprisingly has 60.42% accuracy showed lower performance, possibly because of the model assume the

features in the dataset are independent. On the other hand, the KNN model has an accuracy of 99.96% by correctly identifying most edible and poisonous mushrooms.

4.3. Observations and Analysis

The Naive Bayes model was faster and simpler to implement overall and the running time for the result was essentially faster than the KNN model, making it suitable for scenarios with limited computational resources, but the accuracy was much lower than expected. However, KNN outperformed the Naive Bayes in terms of accuracy and precision, but the disadvantage was the running time for the outcome was extremely long.

5. Conclusions

This research demonstrates how well machine learning works to solve classification issues with categorical datasets. When it came to distinguishing between poisonous and edible mushrooms, the KNN model proved to be the more effective algorithm. However, Naive Bayes proved to be a faster alternative with acceptable accuracy and was suitable for simpler implementations.

The main takeaways from this study were the significance of appropriate preprocessing in getting categorical data ready for machine learning and the advantages and disadvantages of models, such as the KNN model, in comparison to the Naive Bayes model. Furthermore, while learning how to use the method, the visualization's use in analyzing data and model performance was also essential.

6. Acknowledgment

The Figure 1 image in the paper is from an online website that indicates the difference between the mushrooms. Figure 2 image was the visual output for the Naive Bayes model in the project. Figure 3 image is from the previous class slide. Also utilized AI components such as ChatGPT to help establish the visual representation like the feature distribution diagrams, which clearly present the dataset in different categories (cap diameter, cap shape, etc)

References

Burba, Kira. “Mushrooms: Vital to Ecosystems - Smithsonian Gardens.” *Mushrooms: Vital to Ecosystems*, 8 Sept. 2020, gardens.si.edu/learn/blog/mushrooms-vital-to-ecosystems-and-a-culinary-delight/.

Emily Stearn, Health Reporter For Mailonline. “Guide to the Mushrooms You Can Safely Eat - and the Poisonous Ones You Must Avoid.” *Daily Mail Online*, Associated Newspapers, 9 Aug. 2023, www.dailymail.co.uk/health/article-12384153/Guide-mushrooms-safely-eat-poisonous-ones-avoid.html.

Lee , Minwoo, and Aileen Benedict. “Matrix Operations and K-Nearest Neighbors.”

Lee, Minwoo, and Aileen Benedict. “Probability and Naïve Bayes.”

Wagner, Dennis, D. Heider, and Georges Hattab. "Secondary Mushroom." UCI Machine Learning Repository, 2021, https://doi.org/10.24432/C5FP5Q.