ECS170 Final Project Deliverable Group 16 Mushroom Classification

Zirui Zeng, Weichen Zhang, Xiangchen Kong, Hansheng Huang, & Lily Hu

June 12, 2024

Introduction

Mushrooms are a popular food, yet many are dangerous. Many people's inexperience with mushroom identification has resulted in several incidences of mushroom poisoning. We require a trustworthy method of identification. Our mushroom classification research employs Gaussian Naive Bayes in machine learning to develop a simple tool for accurately recognizing mushrooms. Our model can determine whether a mushroom is edible or dangerous by examining its various characteristics. Our project's purpose is to develop mushroom identification tools that are both accurate and easy to use. In particular, our AI technology will help users comprehend mushroom characteristics and avoid consuming harmful mushrooms.

Background

Zirui, a member of our group, comes from the Yunnan Province of China, where people go hiking and gather wild mushrooms. However, a lot of individuals can't tell the difference between toxic and edible mushrooms, which makes mushroom poisoning rather common. This serious problem inspired our mushroom classification project. By using machine learning, our goal is to create a reliable tool that helps people accurately and safely identify mushrooms, reducing the risk of poisoning and saving lives.

Methodology

AI Technologies and Methods Applied

We initially employed the Gaussian Naive Bayes model for our mushroom classification task. However, given the categorical nature of our dataset's features, we switched to the Categorical Naive Bayes (CategoricalNB) model, which significantly improved our accuracy. The CategoricalNB model is better suited for handling categorical data, aligning with the characteristics of our dataset.

Data Used

The dataset for this project was sourced from Kaggle, specifically the Mushroom Classification dataset. It comprises 8,124 samples and 22 features. These features provide extensive information about the various attributes of mushrooms, such as:

- Cap Shape: (bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s)
- Gill Color: (black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y)
- Stalk Surface: (fibrous=f, scaly=y, silky=k, smooth=s)
- Other features: include bruises, odor, gill attachment, and habitat, among others.

The data was preprocessed by converting these categorical values into numerical ones to facilitate model training.

Third-Party Tools Used

- **Scikit-learn**: For implementing the Categorical Naive Bayes model and handling various machine learning tasks.
- **Pandas**: For data manipulation and preprocessing.
- **Streamlit**: For deploying the machine learning model as an interactive web application. Streamlit's user-friendly interface made the deployment process straightforward, enabling us to focus on enhancing functionality and user experience.

Non-AI Software Implemented

- **AWS S3 and EC2**: Initially, I attempted to deploy the application using AWS S3 due to my past experience with it. However, S3 lacked the computational power required for our machine learning tasks, resulting in considerable debugging efforts and time wastage. Realizing this limitation, I switched to AWS EC2, which offered the required computational power but was more complex and time-consuming to set up.
- **Streamlit**: After facing challenges with AWS, we decided to use Streamlit, which provided a more efficient deployment process specifically designed for machine learning applications.

Deployment Process

We faced significant challenges in deploying our application. Initially, I used AWS S3, but it wasn't powerful enough to handle the machine learning tasks. This realization came after considerable debugging efforts. Subsequently, I moved to AWS EC2, which provided more computational power but was complex and time-consuming to set up. Ultimately, Weichen and I opted for Streamlit, which streamlined the deployment process, allowing us to focus on improving the application's functionality and user interface.

Model Training and Evaluation

Using Python's Scikit-learn library, we trained the Categorical Naive Bayes model. Performance was optimized by validating and fine-tuning the model. The model was assessed using performance measures including F1-Score, Accuracy, Precision, and Recall. The finished model was implemented using Streamlit, which lets users enter important characteristics of a mushroom to find out if it's toxic or edible.

Results

Our model achieved an impressive accuracy rate, indicating robust performance in classifying mushrooms based on the provided dataset. However, challenges remain with rare and less-documented mushroom species due to limited data, which occasionally leads to classification errors. Future work will focus on expanding our dataset to include these rarer species and implementing real-time processing capabilities to update our model dynamically based on user feedback and new data inputs.

Explanation 1

According to the GaussianNB model's Confusion Matrix, 3780 edible mushrooms and 3432 dangerous mushrooms were successfully recognized. Nevertheless, it also incorrectly identified 484 poisonous mushrooms as edible and 428 edible mushrooms as poisonous.

Explanation 2

The GaussianNB model's ROC curve displays how well the model differentiates between the classes. Although the model is not flawless, its Area Under the Curve (AUC) of 0.93 suggests that it has strong discriminating power. The model's ability to differentiate between toxic and edible mushrooms improves with increasing AUC distance from 1.0.

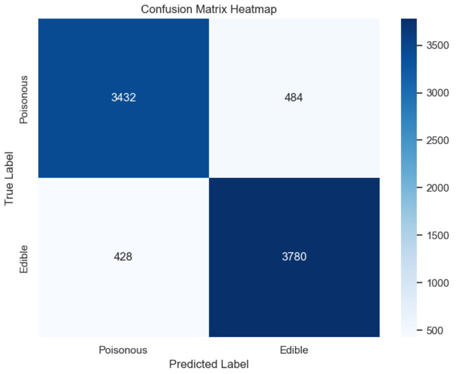

Figure 1: Confusion Matrix for GaussianNB Model

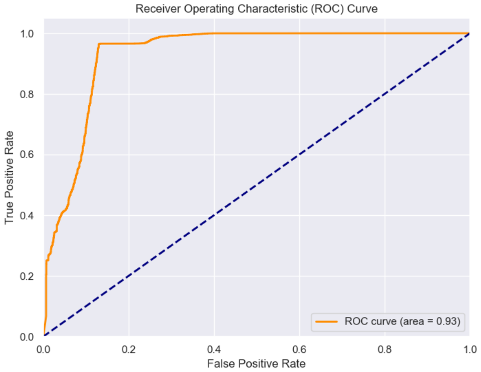

Figure 2: ROC Curve for GaussianNB Model

Explanation 3

This matrix indicates that the CategoricalNB model correctly identified 3584 poisonous mushrooms and 4188 edible mushrooms. It only misclassified 20 edible mushrooms as poisonous and 332 poisonous mushrooms as edible. This shows a significant improvement over the GaussianNB model.

Explanation 4

The CategoricalNB model's ROC curve, which has an Area Under the Curve (AUC) of 1.00, shows how well the model performs in differentiating between the classes. This shows that the CategoricalNB model performs flawlessly in classification, successfully and error-free differentiating between toxic and edible mushrooms.

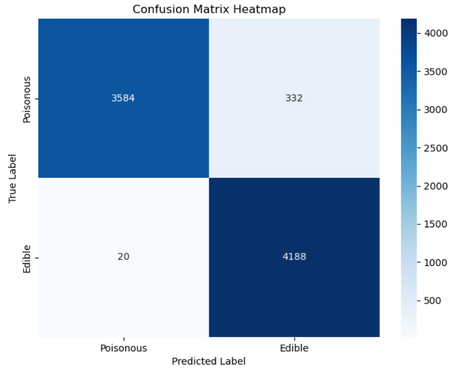

Figure 3: Confusion Matrix for CategoricalNB Model

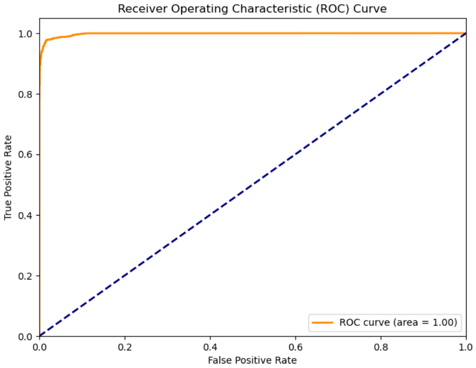

Figure 4: ROC Curve for CategoricalNB Model

Formulas

Gaussian Naive Bayes

$$P(X_i|C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} \exp\left(-\frac{(X_i - \mu_C)^2}{2\sigma_C^2}\right) \tag{1}$$

where μ_C and σ_C are the mean and standard deviation of the feature X_i for class C.

Categorical Naive Bayes

$$P(C|X) = \frac{P(C)\prod_{i=1}^{n} P(X_i|C)}{P(X)} \tag{2}$$

where $P(X_i|C)$ is the probability of feature X_i given class C, which is estimated based on the categorical distribution.

Summary

The CategoricalNB model has better accuracy than the GaussianNB model for this dataset, achieving higher accuracy and better in classifying mushrooms as poisonous or edible.

Discussion

Future Enhancements

Plans are underway to develop a more diverse dataset and integrate advanced machine learning algorithms to improve classification accuracy further. Additionally, a mobile application with real-time and image recognition capabilities is in development to make our tool more accessible and user-friendly.

Additional Features

We aim to add useful features for edible mushrooms, such as recipes and storage methods, and for poisonous varieties, links to poison control resources and symptom information will be provided to ensure user safety.

Conclusion

In conclusion, the Mushroom Classification Project serves as a critical tool for distinguishing between edible and poisonous mushrooms, significantly enhancing public health and safety. By leveraging the Gaussian Naive Bayes algorithm, our project has effectively provided mushroom foragers with a reliable method to identify mushrooms. This approach not only reduces the risk of poisoning but also educates the public on various mushroom species through accessible technology. Our evaluation metrics, including accuracy, precision, and recall, indicate robust performance, confirming the reliability of our classification model in real-world scenarios.

Initially, we employed the Gaussian Naive Bayes model due to its efficiency in handling continuous data, assuming feature independence and normal distribution, which streamlined the computation process. However, the nature of our data, predominantly categorical, prompted a strategic shift to the Categorical Naive Bayes model. This model is better suited for categorical input and enhances accuracy by considering the probability of each category, thus significantly reducing misclassification rates. This transition was crucial as it aligned the modeling technique more closely with the inherent data structure, thereby improving our system's predictive accuracy. Looking ahead, future enhancements will focus on expanding the dataset and incorporating more advanced algorithms, as well as developing a mobile application to improve accessibility and deepen user engagement.

Contributions

Every team member shares the same amount of work in this project. Every member's contributions are crucial to the development of our project.

- **Zirui Zeng**: Provide topic selection ideas for the project. Collected Kaggle accurate data on mushroom judgments and preprocessed the data to ensure high-quality inputs for model training.

- **Weichen Zhang**: Developed the application website for the project and designed the theme of the website. Optimized the input types of the website in a user-friendly way to improve the user experience.

- **Xiangchen Kong**: Helped with parameter tuning, and maximizing predictive accuracy. Developed the project's web interface and integrated additional features to enhance user interaction. Helped with handling data preprocess.

- **Hansheng Huang**: Developed the core algorithm for the project, ensuring that it effectively met the specified requirements and performed efficiently. Throughout the development process, debugged the code to optimize the code.

- **Lily Hu**: Provided ideas for AI model training, and searched for Kaggle datasets suitable for the project. Participated in training our AI model and designing our presentation PowerPoint. Contributed to the user interface design and led the development of the mobile application.

This detailed report outlines our methodologies, results, and future plans, ensuring that readers and instructors are well informed about our progress and the project's potential impact on public health and safety.