Mushroom classification

MD FAIZ KHAN

Contents

Abstract

The Audubon Society Field Guide to North American Mushrooms contains descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom (1981). Each species is labeled as either definitely edible, definitely poisonous, or maybe edible but not recommended. This last category was merged with the toxic category. The Guide asserts unequivocally that there is no simple rule for judging a mushroom's edibility,such as "leaflets three, leave it be" for Poisonous Oak and Ivy. The main goal is to predict which mushroom is poisonous & which is edible.

Introduction

1. Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

- ✓ Present all of the design aspects and define them in detail
- ✓ Describe the user interface being implemented
- ✓ Describe the hardware and software interfaces
- ✓ Describe the performance requirements
- ✓ Include design features and the architecture of the project
- ✓ List and describe the non-functional attributes like:
  - Maintainability
  - Portability
  - Reusability
  - Application compatibility
  - Resource utilization
  - Serviceability

2. Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

3. Definitions

Database

IDE AWS

Description

Collection of all the information monitored by this system

Amazon Web Services

2. General Description

2.1 Product Perspective

The Mushroom classification is a machine learning-based classification modelwhich will help us to predict whether the mushrooms are edible or poisonous.

2.2. Problem statement

The Audubon Society Field Guide to North American Mushrooms contains descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom (1981). Each species is labeled as either definitely edible, definitely poisonous, or maybe edible but not recommended. This last category was merged with the toxic category. The Guide asserts unequivocally that there is no simple rule for judging a mushroom's edibility,such as "leaflets three, leave it be" for Poisonous Oak and Ivy. The main goal is to predict which mushroom is poisonous & which is edible.

2.3. PROPOSED SOLUTION

Based on the dataset, predict the future then we might want to consider using SVM. However, drawing a baseline in the form of some Machine Learning algorithm would behelpful. Why make a baseline model important? Well, to compare the performance of our actual model, let say SVM in this case, is very important to ascertain that we are in the right direction as if performance of SVM is not better than the baseline model then there is no point of using SVM. 1. Baseline Model: Logistic Regression, since this is a classification problem. 2. Actual model: SVM

2.4 Technical Requirements

This document addresses the requirements for predicting the edible mushroom andpoisoning mushroom at early stages and recommending the necessary and rapid action to avoid intake.

2.5 Data Requirements

Data requirements completely depend on our problem statement.

- We need images data that is balanced and must have at least 1000images.
- We require at least 30- 40 images for each class label with annotation.
- An image is nothing more than a two-dimensional array of numbers(pixels)
- Pixel value ranging between 0 to 255
- It is defined by the mathematical function (x, y), the value off(x, y) at anypoint is giving the pixel value at that point of an image
- Original image is in the format of (width, height, no of RGB channels).

There are numerous image file formats out there so it can be hard to know which file type best suits your image needs (on your requirement).

TIFF — Tagged image file format
BMP — Bitmap image file form
JPEG - Joint photographic experts' groups GIF
- graphics interchange format
PNG — portable network graphics

EPS — encapsulated postscript

RAW image files

- Tiffs are great for printing. These are lossless image files meaning they don't need to compress or lose any image quality or information. These format images are high quality images.
- bmp format developed by Microsoft for windows. There is no compression or information loss; this format is generally recommended for high quality scans.
- JPEG is a lossy format meaning that the image is compressed to make a smaller file but this loss is not noticeable.
- JPEG is a very popular format for digital cameras.
- GIFs are widely used for web graphics because they are limited to only 256 colours, can allow for transparency and can be animated. These types of files are typically small in size and very portable.
- PNG are a lossless image format; these files are able to handle up to 16 million colours unlike the 256 colours supported by GIF.
- EPS is a common vector type file.

RAW images that are unprocessed that have been created by a camera orscanner. Digital cameras can shoot in raw, mostly used in photography.

2.6 Tools used

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, TensorFlow, Keras and Roboflow are used to build the whole model.

- PyCharm is used as IDE.
- For visualization of the plots, Matplotlib, Seaborn and Plotly are used.
- AWS is used for deployment of the model.
- Tableau/Power BI is used for dashboard creation.
- MySQL/MongoDB is used to retrieve, insert, delete, and update the database.
- Front end development is done using HTML/CSS
- Python Django is used for backend development.
- GitHub is used as version control system.

1. ROS (Robotic Operating System)

Robot Operating System is an open-source robotics middleware suite. Although ROS is not an operating system but a collection of software frameworks for robot software development, it provides services designed for a heterogeneous computer cluster such as hardware abstraction, low-level device control, implementation of commonly used functionality, message-passing between processes, and package management.

1. **Constraints**

We will only be selecting a few of the mushrooms

1. **Assumptions**

The main objective of the project is to implement the use cases as previously mentioned (2.2 Problem Statement) for new dataset that comes through images. Machine Learning model is used for predicting the above-mentioned use cases based on the input data. It is also assumed that all aspects of this project have the ability to work together in the way the designer is expecting.

Design Details

1. Process Flow

For identifying the different types of anomalies, we will use a deep learning basemodel. Below is the process flow diagram as shown below.

3.1. Data Description

Mushroom dataset is the biggest publicly available dataset. This dataset contains8314 rows and 23 columns.

3.2 Data collection and preparation

Collect a dataset of mushrooms along with their corresponding labels (edible or poisonous). Preprocess the data to ensure that it is clean and consistent.

3.3. Data Pre-processing

Data Pre-processing steps we could use are Null value handling, stop words removal, punctuation removal, Tokenization, Lemmatization, TFIDF, Imbalanced data set handling, Handling columns with standard deviation zero or below a threshold, etc.

3.4 Feature extraction

Extract features from the images using techniques such as edge detection, colorhistograms, and texture analysis. These features should capture the important characteristics of the mushrooms that will be used for classification.

3.5 Feature selection

Use feature selection techniques to reduce the number of features while retainingas much relevant information as possible. This helps to simplify the classification task and avoid overfitting.

3.6 Model selection and training

Select an appropriate machine learning model such as SVM (Support VectorMachines), Random Forest, or Neural Networks. Train the model on the preprocessed and reduced feature dataset using a training set of labeled examples.In this project we used SVM model.

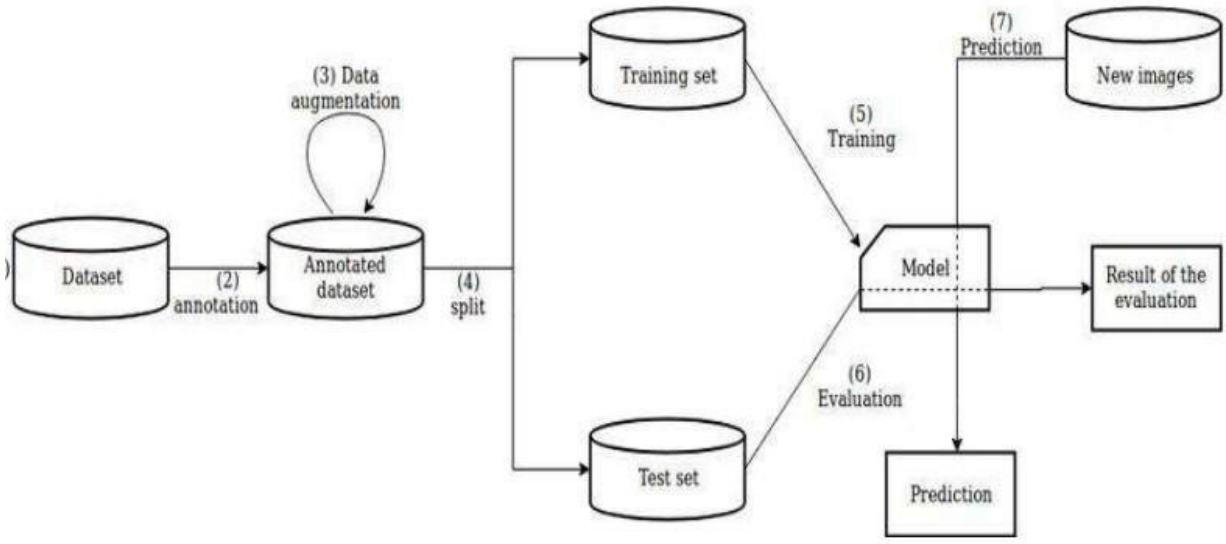
3.7 Model evaluation and fine-tuning

Evaluate the model's performance on a separate testing set of labeled examples.Fine-tune the model's hyperparameters and architecture to optimize its performance.

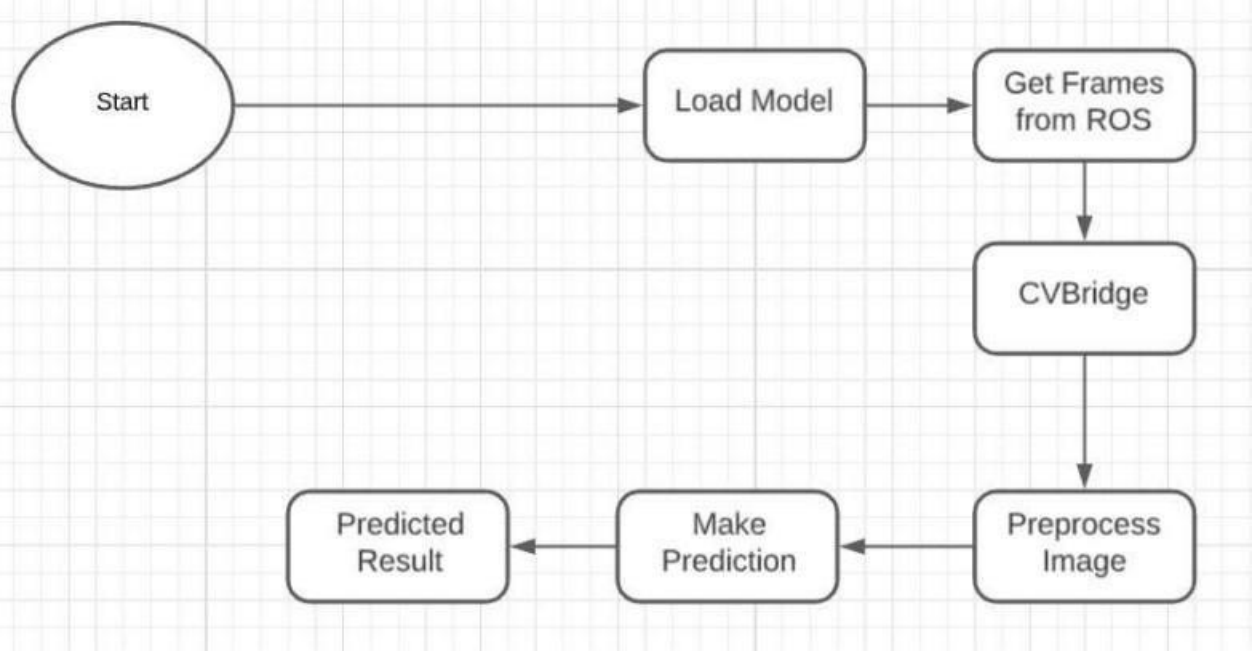
3.8 Prediction and deployment

Use the trained model to predict the class of new mushroom images that were notused during the training and testing phases. Deploy the model in a production environment for use in mushroom classification tasks.

Model Training and Evaluation:



Deployment Process:



Event log

The system should log every event so that the user will know what process is running internally.

Initial Step-By-Step Description:

1. The System identifies at what step logging required
2. The System should be able to log each and every system flow.
3. Developers can choose logging methods. You can choose database logging/ Filelogging as well.
4. System should not hang even after using so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

Error Handling

Should errors be encountered, an explanation will be displayed as to what went wrong? An error will be defined as anything that falls outside the normal and intendedusage.

Performance

The mushroom classification is used for prediction of edible mushrooms or poison. It will inform concerned authorities and take necessary action, so it should be as accurate as possible. So that it will not mislead the concerned authorities.Also, model retraining is very important to improve the performance.

Reusability

The code written and the components used should have the ability to be reused withno problems.

Application Compatibility

The different components for this project will be using Python as an interface between them. Each component will have its own task to perform, and it is the job of the Python to ensure proper transfer of information.

Resource Utilization

When any task is performed, it will likely use all the processing power available until that function is finished.

Deployment

AWS
Google Cloud Microsoft
Azure

Conclusion

The Designed Mushroom classification will predict edible or poisonous mushrooms based on various data used to train our algorithm, so we can identify the intake in earlystages and can take necessary action to stop them immediately.