

Classification of Mushrooms (Edible/Poisonous)

**Artificial Intelligence
CS-617B**

Curious 4 AI



**Sacred Heart University
School of Computer Science & Engineering
The Jack Welch College of Business & Technology**

**Submitted To:
Dr. Reza Sadeghi**

Fall - 2022

Table of Contents

Introduction of Team members.....	3
1. Introduction.....	5
1.1. Artificial Intelligence.....	5
1.2. Mushrooms.....	7
1.3. Research Question.....	8
1.4. Github Repository.....	8
2. Dataset Description.....	8
2.1. URL of Dataset.....	8
2.2. When, where and how the data is collected.....	8
2.3. Name, definition and characteristics of features.....	9
3. Related work.....	10
3.1. Comparison of previous works.....	10
3.2. Source website.....	10
4. Project plan.....	10
5. Data Exploration.....	11
5.1. Data analysis.....	11
5.2. Research observations.....	12

Classification of Mushrooms (Edible/Poisonous)

Team: **Curious 4 AI**

Name	Contact
Sarvani Konda (Lead)	kondas2@mail.sacredheart.edu
Venkata Prasanth Pinaka	pinakav@mail.sacredheart.edu
Srikar Singam	singams2@mail.sacredheart.edu
Jayadev Varma Sri Kakarlapudi	srikakarlapudij@mail.sacredheart.edu

Introduction Of Team Members:

Sarvani Konda:

I have done my undergraduate in Information Technology and started my career in 2016. I've worked in the technology sector for over 5 years, throughout my career I have worked on Openstack, Kubernetes, Python, Golang, and other technologies related to cloud. I am an enthusiastic coder and enjoy using my skills to contribute to open-source projects. In my spare time, I love to travel, watch TV series and learn new skills.

Venkata Prasanth Pinaka:

I have completed my graduation in the stream of Computer Science and Engineering. I worked over 6 years as a developer in technologies like Oracle PLSQL and Spring boot in JAVA in a few prime organizations. I like collaborating with my team and solving the problem at hand. I enjoy researching topics and gaining knowledge.

Srikar Singam:

Enthusiastic and Driven engineer to learn and explore new things. Graduate student in Computer science at Sacred Heart University.

Jayadev Varma Sri Kakarlapudi:

I am Jayadev Varma Sri Kakarlapudi, undergraduate in Computer science engineering and worked with ADP as a software developer for almost 4 years. My interest to learn and explore more in my field of work led me here to pursue my graduation. I would like to work with people with great passion and dedication towards their work just like the bunch of people in this team.

Why we chose Sarvani Konda as team lead:

Sarvani has a proactive and enthusiastic personality. Her calm, collected and problem solving nature along with her acute knowledge on AI has led us to choose her as our Team lead. She is a teamplayer and we trust her to take the right steps for our project.

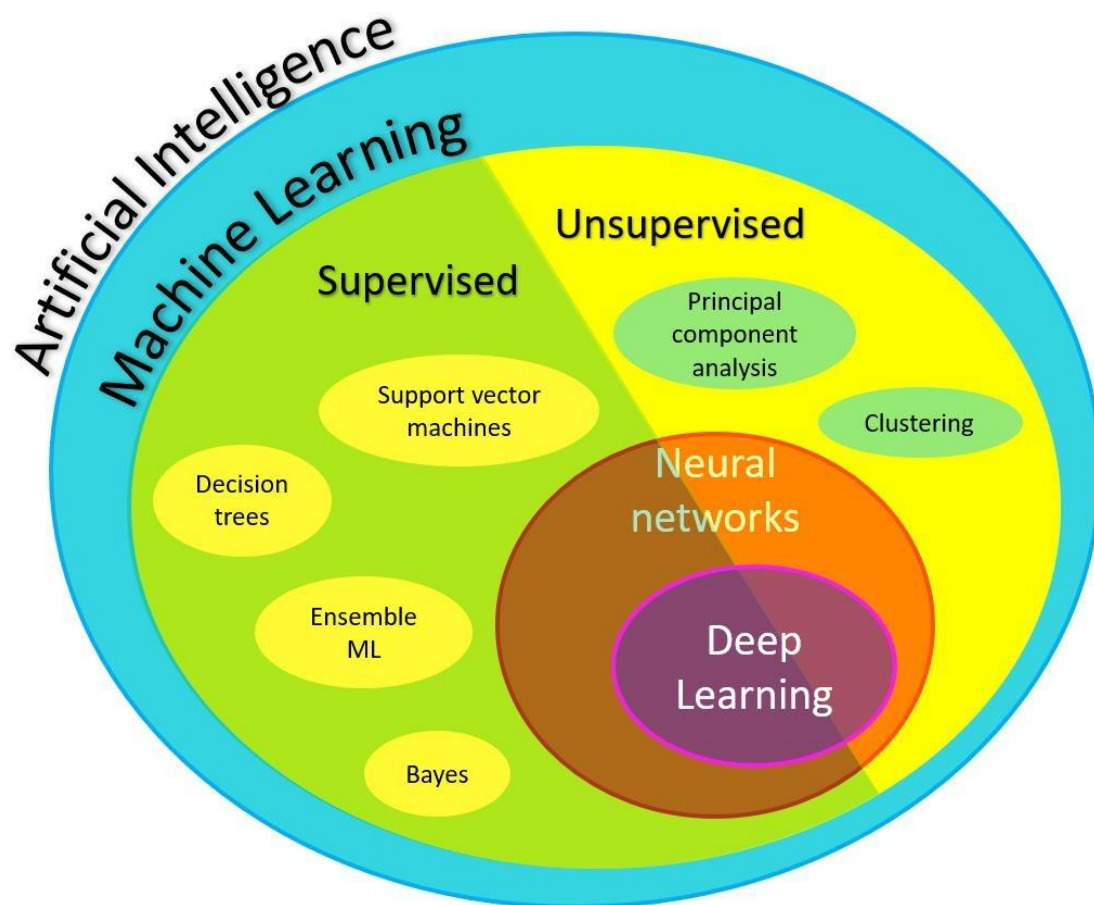
1.0 Introduction

1.1 Artificial Intelligence:

Definition of AI:

It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but A.I. does not have to confine itself to methods that are biologically observable.

(Source: <https://hai.stanford.edu>)



(Source: <https://www.physicamedica.com>)





Few Applications of AI:

- Healthcare.
- Manufacturing.
- Banking and Finance.
- Law
- Retail.
- Real Estate.
- Digital Marketing.

Types of AI:

Types of AI

The emergence of artificial superintelligence will change humanity, but it's not happening soon.
Here are the types of AI leading up that new reality.

Reactive AI	Limited memory	Theory of mind	Self-aware
<ul style="list-style-type: none">• Good for simple classification and pattern recognition tasks• Great for scenarios where all parameters are known; can beat humans because it can make calculations much faster• Incapable of dealing with scenarios including imperfect information or requiring historical understanding	<ul style="list-style-type: none">• Can handle complex classification tasks• Able to use historical data to make predictions• Capable of complex tasks such as self-driving cars, but still vulnerable to outliers or adversarial examples• This is the current state of AI, and some say we have hit a wall	<ul style="list-style-type: none">• Able to understand human motives and reasoning. Can deliver personal experience to everyone based on their motives and needs.• Able to learn with fewer examples because it understands motive and intent• Considered the next milestone for AI's evolution	<ul style="list-style-type: none">• Human-level intelligence that can bypass our intelligence, too
			

SOURCE: DAVID PETERSSON; ICONS: MIKHEY/GETTY IMAGES

©2020 TECHTARGET. ALL RIGHTS RESERVED TechTarget

Advantages of AI:

- ❖ Good at detail-oriented jobs.
- ❖ Reduced time for data-heavy tasks.
- ❖ Delivers consistent results.

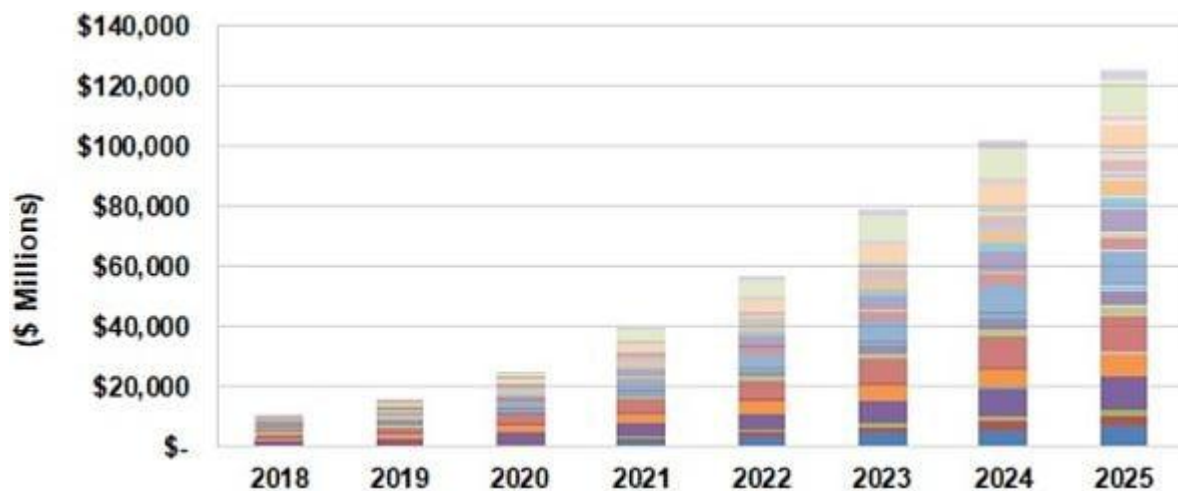
- ❖ AI-powered virtual agents are always available.

Disadvantages of AI:

- ❖ Expensive
- ❖ Requires deep technical expertise
- ❖ Limited supply of qualified workers to build AI tools
- ❖ Only knows what it's been shown

(Source: <https://www.techtarget.com>)

Trend of AI:



(Source: <https://www.newark.com>)

1.2 Mushrooms:

Mushroom is the fleshy body of a fungus, typically produced above ground, on soil, or on its food source. The terms "mushroom" and "toadstool" go back centuries and were never precisely defined. During the 15th and 16th centuries, the terms mushroom is used. Identifying mushrooms requires a basic understanding of their macroscopic structure. While modern identification of mushrooms is quickly becoming molecular, the standard methods for identification are still used by most and have developed into a fine art harking back to medieval times and the Victorian era, combined with microscopic examination. Many species of mushrooms seemingly appear overnight, growing or expanding rapidly. Raw brown mushrooms are 92% water, 4% carbohydrates, 2% protein and less than 1% fat. They have

minimal or no vitamin C and sodium content. Mushrooms can be used for dyeing wool and other natural fibers. Some mushrooms are used in folk medicine.

1.3 Research Question:

Mushrooms are of several types, there are over 50,000 species of mushrooms only in North America, and some of them are extremely toxic in nature and can lead to death when consumed. It can be hard to identify the edibility of them. To classify mushrooms into edible or poisonous in an easier way different machine-learning/AI models can be used. In this project, we will examine the data and build different ML/AI models that will detect the edibility of a mushroom into edible or poisonous by its specifications like gill color, cap shape, etc using classifiers.

1.4 GitHub Repository:

[Here](#) is the link for our project GitHub repository.

2.0 Dataset Description:

2.1 URL of the Dataset:

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like "leaflets three, let it be" for Poisonous Oak and Ivy.

(Source: <https://archive.ics.uci.edu/ml/datasets/mushroom>)

2.2 When, Where and How the Dataset is collected:

These Mushroom records were taken in 1981 by G. H. Lincoff (Pres.).

Origin: Mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knopf

Donor: Jeff Schlimmer (Jeffrey.Schlimmer '@' a.gp.cs.cmu.edu)

2.3 Name, Definition and characteristics of features:

Here we have 8124 samples in the given dataset with 22 attributes(features). The datatype of all the features is String.

1. **cap-shape:** bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
2. **cap-surface:** fibrous=f,grooves=g,scaly=y,smooth=s
3. **cap-color:**
brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w,yellow=y
4. **bruises?:** bruises=t,no=f
5. **odor:** almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
6. **gill-attachment:** attached=a,descending=d,free=f,notched=n
7. **gill-spacing:** close=c,crowded=w,distant=d
8. **gill-size:** broad=b,narrow=n
9. **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
10. **stalk-shape:** enlarging=e,tapering=t
11. **stalk-root:** bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
12. **stalk-surface-above-ring:** fibrous=f,scaly=y,silky=k,smooth=s
13. **stalk-surface-below-ring:** fibrous=f,scaly=y,silky=k,smooth=s
14. **stalk-color-above-ring:**
brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
15. **stalk-color-below-ring:**
brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
16. **veil-type:** partial=p,universal=u
17. **veil-color:** brown=n,orange=o,white=w,yellow=y
18. **ring-number:** none=n,one=o,two=t
19. **ring-type:**
cobwebby=c,evanescent=e,flaring=f,large=l,none=n,pendant=p,sheathing=s,zone=z
20. **spore-print-color:**
black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y
21. **population:** abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
22. **habitat:** grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

3.0 Related Work:

3.1 Comparison of previous works:

In the paper of Schlimmer, J.S. (1987), using Representational Adjustment where a Bayesian weighting method was used that was capable of acquiring simple ideas but while implementing the four methods he faced limitations in the concept of learning tasks which lead him to Boolean chunk learning.

In 'Duch W, Adamczak R, Grabczewski K (1996)' paper the problem of extracting rules from neural networks has a natural geometrical interpretation. Logical description of data is possible if input spaces are correctly separated but this may become arbitrarily accurate by increasing the number of variables but this leads to disadvantage of rules becoming very large in number.

In the paper 'Iba, W., Wogulis, J., & Langley, P. (1988)' given a set of observations, humans acquire concepts that organize those observations and use them in classifying future experiences. This carries search through a space of possible hierarchies.

3.2 Source website(s):

<https://escholarship.org/uc/item/48r6d4z0>

https://www.researchgate.net/publication/2633347_Extraction_of_Logical_Rules_From_Training_Data_Using_Backpropagation_Networks

<https://www.sciencedirect.com/science/article/abs/pii/0004370289900465>

4.0 Project Plan:

By applying different Classification models on the mushroom dataset, we will be classifying mushroom edibility(edible or poisonous). First step of the project is to clean up, transform and preprocess the data, so it can be applied on different classifier models to estimate the classification of the mushrooms.

5.0 Data Exploration:

In the data exploration phase, step one is to check for missing, duplicate and null values from the selected dataset (mushrooms) and use strategies like imputation and interpolation to fill the missing data.

From our research, it is explored that there are some missing values for the feature “stalk-root”, (missing data represented by “?”). The missing values were replaced by performing an imputation strategy and estimated the value to be of “b” as it had the most frequency.

After data cleaning, the next step is to transform the data. We have used label-encoder to transform the feature values from string to integer as Scikit-learn’s algorithms cannot be powered by objects or string values. It must be either in float or in integer which the Label-Encoder method provides. Next step is to analyze the data.

5.1 Data Analysis:

- **Univariate Analysis:** One of the methods to analyze the dataset is Univariate Analysis, Univariate analysis is a process of exploring each variable in a data set, separately. It provides the summary statistics for each feature in the data set or summary only on one variable Ex: PDF, CDF, etc. In this project, Histogram is used on each feature of the mushroom dataset to understand and analyze the data.
 - **Descriptive Analysis:** As the name suggests, the main goal of descriptive analysis is to describe and summarize the data, different types of measures like frequency, central tendency, dispersion and etc of the dataset can be known using this method.
 - **Distribution Analysis:** In distribution analysis, using different statistics, graphs and density estimates of data, which will help us to identify whether the distribution of the data is a normalized distribution or if the data is skewed. We can say that the data is skewed when an asymmetric curve is seen on a graph, skewed data has a ‘tail’ on either side of the graph and a

symmetrical or a bell shaped graph is classified as a normalized data.

- **Bivariate Analysis:** This is another method to analyze more of the dataset by finding the relationship between each feature of the dataset and the target variable or a relationship between any 2 features. Ex: Pair-plot, Scatter plot, etc.
 - **Pearson Correlation:** It is a method for numerical variables, it assigns a value between -1 to 1 for all the features, where 0 signifies no relationship, 1 is total positive correlation, and -1 is total negative correlation. A correlation value of 1 between two variables would indicate that a significant and positive relationship exists between the two, for example if feature A increases, then B will also increase, whereas if the value of the correlation is negative, then if A increases, B decreases.
 - **Pair Plot:** It allows us to plot pairwise relationships between features within a dataset and helps us understand the data by summarizing a large amount of data in a single figure. We can understand the linear and non-linear relationships among features.

5.2 Research observations:

Using the above analysis, Histograms are plotted for each feature and a box plot is plotted between the target “class” and feature “stalk-color-above-ring” (selected stalk-color-above-ring feature as it has the highest mean value, calculated using descriptive analysis), from the plots(in Jupyter Notebook) it can be observed that the data is skewed, as plots shown are asymmetric and outliers are detected.

Pearson correlation values are calculated using a heatmap plot, it is observed that the feature veil-type is assigned a 0 value indicating that there is no relationship between veil-type and any other feature, and veil-color is highly correlated with the feature gill-attachment with the correlation value of a positive 0.9 which means if veil-color increases gill-attachment also increases.

Using a Pair plot it is observed that cap-color and gill color are non linearly related when compared to other features.