# CLASSIFICATION ALGORITHM FOR EDIBLE MUSHROOM IDENTIFICATION GENERATES MUSHROOM BUSINESS ENHANCEMENT IN THE GLOBAL MARKET

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

CERTIFIED **PROFESSIONAL**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**GRAHAM G**

A close up of a sign

Description automatically generated

**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**DECEMBER 2019**

**Table of Contents**

[**1.  Problem Definition**](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Problem-Definition)

[1.1  Project Overview](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Project-Overview)

[1.2  Problem Statement](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Problem-Statement)

**2. Introduction**

[**3. Dataset Analysis**](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Dataset-Analysis)

[3.1  Importing Python libraries](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Importing-Python-libraries)

[3.2  Importing the data](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Importing-the-data)

[3.3  Data Description and Exploratory Visualisations](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Data-Description-and-Exploratory-Visualisations)

3.3.1 *overview*

3.3.2 *numerical features overview*

[3.4 Target distribution](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Feature-distribution-by-target-attribute)

[3.5 Feature distribution by target attribute](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Feature-distribution-by-target-attribute)

[3.5.1 ***Cap feature****s*](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Education)

[*3.5.1.1 Cap shape*](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Gender)*vs class*

[*3.5.1.2*](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Marital-Status) *Cap surface vs class*

[*3.5.1.3 Cap color vs class*](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Distance-from-Home)

[3.5.2***Gill features***](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Department)

[*3.5.2.1 Gill attachment vs class*](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Role-and-Work-Conditions)

[*3.5.2.2*](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Years-at-the-Company)*Gill spacing vs class*

[*3.5.2.3 Gill size vs class*](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Years-With-Current-Manager)

[*3.5.2.4*](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Work-Life-Balance-Score)*Gill color vs class*

[*3.5.2.5*](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Pay/Salary-Employee-Information)*Spore print color vs class*

[3.5.3***Stalk features***](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Employee-Satisfaction-and-Performance-Information)

3.5.3.1 *Stalk shape vs class*

3.5.3.2 *Stalk root vs class*

3.5.3.3 *Stalk surface above ring vs class*

3.5.3.4 *Stalk surface below ring vs class*

3.5.3.5 *Stalk color above ring vs class*

3.5.3.6 *stalk color above ring vs class*

3.5.4 ***Miscellaneous Features***

3.5.4.1 odor vs class

3.5.4.2 Bruises vs class

3.5.4.3 Habitat vs class

3.5.4.4 Population vs class

[3.6 Correlation](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Correlation)- statistical technique

[3.7 EDA Concluding Remarks](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#EDA-Concluding-Remarks)

[**4.  Pre-processing Pipeline**](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Pre-processing-Pipeline)

[4.1  Encoding](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Encoding)

[4.2  Splitting data into training and testing sets](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Splitting-data-into-training-and-testing-sets)

[**5 . Building Machine Learning Models**](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Building-Machine-Learning-Models)

[5.1  Logistic Regression](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Logistic-Regression)

5.2 SVM

5.3 DECISION TREE

5.4 RANDOM FOREST CLASSIFIER

[**6.  Concluding Remarks**](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Concluding-Remarks)

[6.1  Risk Category](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Risk-Category)

[6.2  Strategic Plan](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Strategic-Retention-Plan)

Abstract

*Mushrooms belong to fungi kingdom and they contain essential nutrients such as proteins, vitamins, minerals, amino acids, antibiotics and antioxidants. There are numerous health benefits of mushrooms. All species of mushrooms are not edible. So before consuming, it should be checked for edibility. Accurate determination and proper identification of species are the only safe way to ensure edibility, and safeguard against possible accident of consuming poisonous one. This Project aims in classifying edible and poisonous mushrooms based on certain attributes like shape, size, color, etc. We use different Supervised learning Algorithms like Logistic Regression, Support Vector Machine (SVM), Random Forest classifier and Decision tree classifier for training. This classification model is saved and then used later to predict whether a given mushroom is edible or not*

**1. Problem Definition**

**1.1 Project Overview**

This project focuses on the opportunities of mushroom industry in the global market with inputs of machine learning algorithms. Mushroom industry is growing day by day because of the nutrition benefits of mushrooms. Global mushroom market is expected to grow at a CAGR of 7.9% by 2026 as by the forecasting report. So before consuming it should be checked for edibility. Accurate determination and proper identification of species are the only way to ensure edibility. So critical determination can’t be applied based on human analysis. As we all know there are different species of mushroom in the world but there is a very limited study on determining edible or poisonous mushroom. Classification process of poisonous mushroom or not will be easily conducted by learning machine using mining data as one of the ways to extract computer assisted knowledge. Currently, there are four comparisons of the best classification algorithms in data mining, namely: Logistic Regression, support vector machine (SVM), Random Forest classifier and Decision tree classifier.

**1.2 Problem Statement**

Problem Statement: To make a suitable machine learning algorithm to predict if the mushroom is poisonous or not (e or p)

**OBJECTIVES**

In this study, we will attempt to identify:

* The **likelihood** of a mushroom to be edible or poisonous thereby contributes in business enhancement.
* The key **indicators** of an edible and poisonous mushroom identification.
* The categorization of a **prospective** mushroom as edible or poisonous one.

## Domain Knowledge

As anticipated, we are going to gain some basic domain knowledge about mushrooms.

**2. INTRODUCTION**

The global mushroom market consumption was 12.74 million tons (MT) in 2018 and is projected to reach 20.84 million tons (MT) by 2026, exhibiting a CAGR of 6.41% in the forecast period.

Mushrooms are edible fungi that have been classified as vegetables, however technically they are not considered as plants. Mushrooms are rich in riboflavin, potassium, vitamin D, selenium, and other ingredients that are beneficial for human health. Results of the clinical and preclinical studies on edible mushroom consumption suggest that their consumption may support healthy immunity, weight management, and enhancement of overall health. Additionally, mushroom consumption can potentially reduce the risk of diseases, such as prostate cancer and breast cancer.

Agaricus bisporus, also called as common mushroom or button mushroom, is widely cultivated and consumed mushroom type in the world. Mushroom is gaining gradual recognition and is becoming a crucial part of the diet not only in developed markets but also in developing markets.

With the expansion of the global mushroom industry, the commercial production and distribution of mushroom have improved significantly. Positive trade scenario for mushroom, coupled with initiation of cultivation of number of mushroom varieties across the major mushroom producing countries/regions, is expected to fuel the growth of mushroom market in the forthcoming years.

Organization of sales distribution channel for the sale of processed and fresh mushroom across several regions where demand for the product is high is anticipated to further propel the sales of mushroom. Furthermore, the high medicinal value of the mushroom and increasing awareness about it among people is expected to contribute to the growth of the mushroom market during the foreseeable years.

## MARKET DRIVERS

***“Mushrooms are Nutrition-Dense Food – Rising Health Consciousness is leading to Increased Demand for Food that has High Nutritional Value”***

Growing demand for food that is low in fat and cholesterol content and is packed with nutrients such as selenium, potassium, vitamins, and other micronutrients is expected to escalate the growth of the global mushroom market. Rapidly expanding the foodservice industry is boosting the demand for mushroom across hotels, restaurants, and cafeterias. Increasing adoption of mushrooms as a substitute for meat, coupled with increasing number of vegan populations, is anticipated to further drive the growth of the market during the foreseeable years. Growing demand for processed mushroom especially across western countries is generating export opportunities for major mushroom producing Asian countries.

***“Increasing investment in R&D to Improve Commercial Cultivation of Mushroom Can Offer Growth Opportunities to the Global Mushroom Market.”***

The governments across several countries are investing in the production of high-quality mushroom varieties that are not only safe for human consumption but can also be exported to other countries to generate revenue. Adoption of the advanced pest management system to control damage caused due to pest infection in mushroom can further result in higher mushroom yields, which will ultimately aid the market growth of mushrooms.

## SEGMENTATION

### By Type Analysis

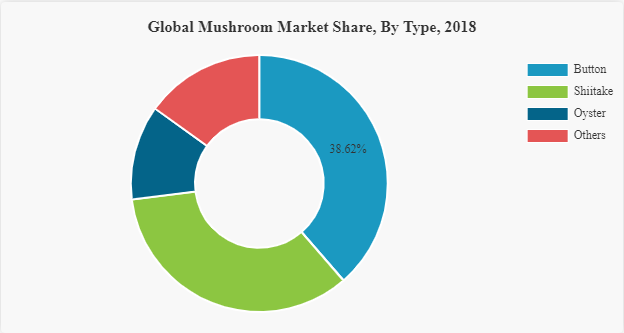
***“Button Mushroom to Hold the Largest Market Share”***

The button mushroom is the widely consumed mushroom type across the world and can potentially offer a number of health benefits. The mushroom type is commercially grown in almost all the major mushroom producing countries such as China, USA, Italy, UK, Germany, Taiwan, and others. Increased R&D on white button mushroom to find out about its potential in preventing cancer, along with its availability at affordable prices as compared to specialty mushroom varieties is projected to escalate the demand for button mushrooms during the forecast period. The demand for Shiitake mushroom is expected to grow in the upcoming years as the consumption of Shiitake in Southeast Asian countries and European countries is increasing at a decent pace.

### By Form Analysis

***“The Demand for Fresh Mushroom is Witnessing Rapid Increase”***

The demand for fresh mushroom is increasing across the globe, especially across U.S. and other developed economies, due to the growing demand for healthier food products in which no preservatives are added. The U.S. is one of the largest producers of fresh mushrooms. The demand for fresh mushroom is projected to witness a steep rise across developing economies of Asia Pacific and in European countries due to rising consumer inclination towards organic food.



In countries, such as South Korea, India, and others, where particular mushroom varieties such as button mushrooms are grown in abundance, the cost of fresh mushroom is comparatively lower. Low price of fresh mushroom as compared to certain processed mushroom forms may push the sales of fresh mushroom across the aforementioned countries.

## REGIONAL ANALYSIS

***“The Asia Pacific is the Leading Region in the Global Market; Europe Market is Projected to Witness Fastest Growth”***

The per capita consumption of mushroom in China, which is the largest producer of mushroom in the world, is higher than any other country. The consumption of mushroom in Asian countries such as Japan, India, and others is increasing at a significant pace attributed by increasing production along with the increased import of mushrooms in these countries. Adoption of advanced and modern mushroom cultivation techniques by mushroom farmers of the region is expected to fuel the mushroom market of the region in the forthcoming years. The volumetric consumption of mushroom in the Asia Pacific was 10.25 Million Ton in 2018 and is anticipated to grow at a rapid pace during the forecast years.

Mushroom production and consumption in Europe is growing rapidly due to increasing demand for nutrition-rich food in the region. In countries such as the U.K., Netherlands, Germany, Spain, and others the demand for mushroom for medicinal purposes is witnessing decent growth and the trend is expected to continue during the upcoming years. Utilization of advanced mushroom processing techniques by mushroom producers and distributors in Europe is predicted to contribute to the growth of the processed mushroom sales across the region.

The mushroom market trends in North America, especially in the U.S., is expanding significantly due to the higher demand for the product attributed by abundance of health benefits offered by mushrooms. The U.S. is one of the largest producers of button mushrooms in the world. Key market players in North America are adopting advanced cultivation technology along with merger and acquisition activities to increase their production capacity and maintain their prominence in the regional market.

## INDUSTRY KEY PLAYERS

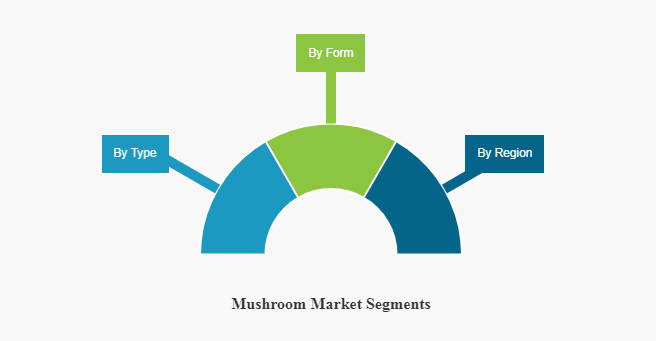
***“Increasing Focus on Merger & Acquisition Activities by International Market Players”***

The mushroom market in many developed and developing economies is proliferating at a significant pace, however, their systematic distribution is still a challenge across many countries and regions. Investment by industry players as well as the government to improve the distribution system of fresh and processed mushroom can bring immense growth opportunities for the sales of mushroom across the world. A number of key market players are focusing on merger & acquisition strategies to remain prominent in the global market.

### List Of key Companies Profiled

* Agro Dutch Industries Ltd.
* [Weikfield Foods Pvt. Ltd.](https://weikfield.com/products/fresh-white-button-mushrooms/)
* BioFungi GmbH
* California Mushroom Farms Inc.
* Monaghan Mushrooms
* Highline Mushrooms
* Scelta Mushrooms BV
* [Monterey Mushrooms, Inc.](https://www.montereymushrooms.com/)
* Nasza Chata
* Banken Champignons B.V.
* Bonduelle
* Ecolink Baltic

The increasing demand for a mushroom which is leading to increasing investment by key players and government to step up the cultivation of edible mushroom is a significant growth engine of the global mushroom market share. Mushroom is commercially grown at large scale in countries, such as China, U.S., and several European countries. The consumption of mushroom across European and Asian countries is projected to flourish significantly during the forecast period due to increased production as well as the positive trade scenario of mushroom across the world.



### Mushrooms Basics Concepts

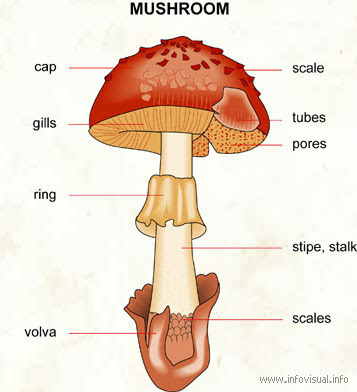
A mushroom, or toadstool, is the fleshy, spore-bearing fruiting body of a fungus, typically produced above ground on soil or on its food source.

The standard for the name “mushroom” is the cultivated white button mushroom, Agaricus bisporus; hence the word “mushroom” is most often applied to those fungi (Basidiomycota, Agaricomycetes) that have a stem (stipe), a cap (pileus), and gills (lamellae, sing. lamella) on the underside of the cap. “Mushroom” also describes a variety of other gilled fungi, with or without stems, therefore the term is used to describe the fleshy fruiting bodies of some Ascomycota. These gills produce microscopic spores that help the fungus spread across the ground or its occupant surface.

Forms deviating from the standard morphology usually have more specific names, such as “bolete”, “puffball”, “stinkhorn”, and “morel”, and gilled mushrooms themselves are often called “agarics” in reference to their similarity to Agaricus or their order Agaricales. By extension, the term “mushroom” can also designate the entire fungus when in culture; the thallus (called a mycelium) of species forming the fruiting bodies called mushrooms; or the species itself.

Identifying mushrooms requires a basic understanding of their macroscopic structure. Most are Basidiomycetes and gilled. Their spores, called basidiospores, are produced on the gills and fall in a fine rain of powder from under the caps as a result. At the microscopic level the basidiospores are shot off basidia and then fall between the gills in the dead air space. As a result, for most mushrooms, if the cap is cut off and placed gill-side-down overnight, a powdery impression reflecting the shape of the gills (or pores, or spines, etc.) is formed (when the fruit body is sporulating). The color of the powdery print, called a spore print, is used to help classify mushrooms and can help to identify them. Spore print colors include white (most common), brown, black, purple-brown, pink, yellow, and creamy, but almost never blue, green, or red.

Mushrooms are used extensively in cooking, in many cuisines (notably Chinese, Korean, European, and Japanese). Separating edible from poisonous species requires meticulous attention to detail; there is no single trait by which all toxic mushrooms can be identified, nor one by which all edible mushrooms can be identified. Many mushroom species produce secondary metabolites that can be toxic, mind-altering, antibiotic, antiviral, or bioluminescent. Although there are only a small number of deadly species, several others can cause particularly severe and unpleasant symptoms. Toxicity likely plays a role in protecting the function of the basidiocarp: the mycelium has expended considerable energy and protoplasmic material to develop a structure to efficiently distribute its spores.



### Mushroom Features Glossary

* Cap (Pileus): the expanded, upper part of the mushroom; whose surface is the pileus
* Cup (Volva): a cup-shaped structure at the base of the mushroom. The basal cup is the remnant of the button (the rounded, undeveloped mushroom before the fruiting body appears). Not all mushrooms have a cup.
* Gills (Lamellae): a series of radially arranged (from the center) flat surfaces located on the underside of the cap. Spores are made in the gills.
* Mycelial threads: root-like filaments that anchor the mushroom in the soil.
* Ring (Annulus): a skirt-like ring of tissue circling the stem of mature mushrooms. The ring is the remnant of the veil (the veil is the tissue that connects the stem and the cap before the gills are exposed and the fruiting body develops). Not all mushrooms have a ring.
* Scale: rough patches of tissue on the surface of the cap (scales are remnants of the veil).
* Stalk (or Stem, or Stape): the main support of the mushroom; it is topped by the cap. Not all mushrooms have a stalk (stem).

Another feature to consider when identifying mushrooms is whether they bruise or bleed a specific color. Certain mushrooms will change colors when damaged or injured. Cutting into a mushroom and observing any color changes can be very important when trying to determine what it is.

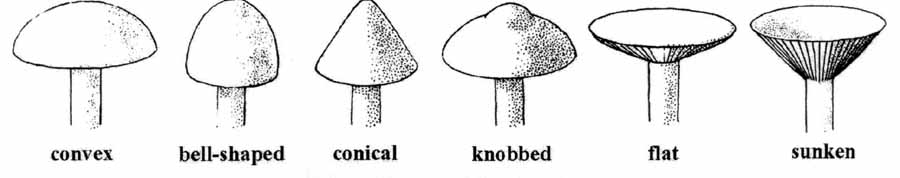
A universal veil is a temporary membranous tissue that fully envelops immature fruiting bodies of certain gilled mushrooms. The developing Caesar’s mushroom (Amanita caesarea), for example, which may resemble a small white sphere at this point, is protected by this structure. The veil will eventually rupture and disintegrate by the force of the expanding and maturing mushroom, but will usually leave evidence of its former shape with remnants. These remnants include the volva, or cup-like structure at the base of the stipe, and patches or “warts” on top of the cap.

A partial veil (also called an inner veil, to differentiate it from the “outer” veil, or velum is a temporary structure of tissue found on the fruiting bodies of some basidiomycete fungi, typically agarics. Its role is to isolate and protect the developing spore-producing surface, represented by gills or tubes, found on the lower surface of the cap. A partial veil, in contrast to a universal veil, extends from the stem surface to the cap edge. The partial veil later disintegrates, once the fruiting body has matured and the spores are ready for dispersal. It might then give rise to a stem ring, or fragments attached to the stem or cap edge. In some mushrooms, both a partial veil and a universal veil may be present.

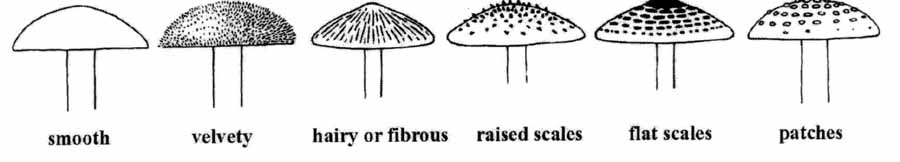
### Mushroom Features by pictures

As shown by, some pictures outline basic mushroom features as they can be found within our dataset.

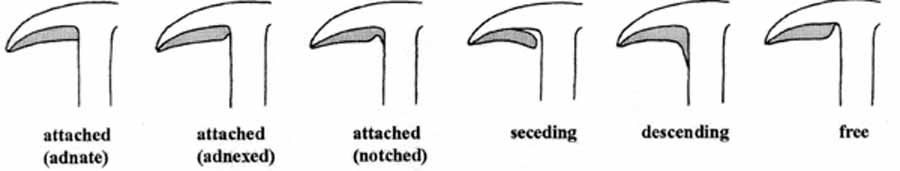
Mushroom cap shape

[](https://datascienceplus.com/wp-content/uploads/2018/02/mushroom-cap-shape.jpg)

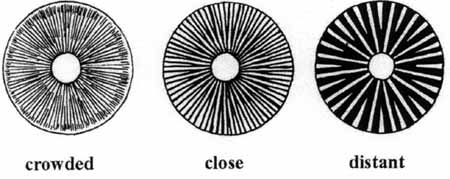
Mushroom cap surface

[](https://datascienceplus.com/wp-content/uploads/2018/02/mushroom-cap-surface.jpg)

Mushroom gill attachment

[](https://datascienceplus.com/wp-content/uploads/2018/02/mushroom-gill-attachment.jpg)

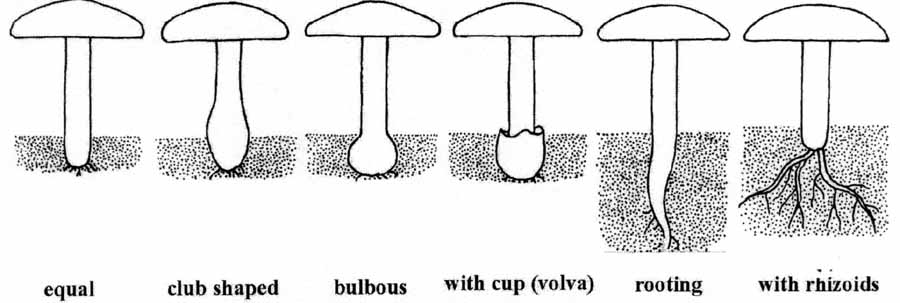
Mushroom gill spacing

[](https://datascienceplus.com/wp-content/uploads/2018/02/mushroom-gill-spacing.jpg)

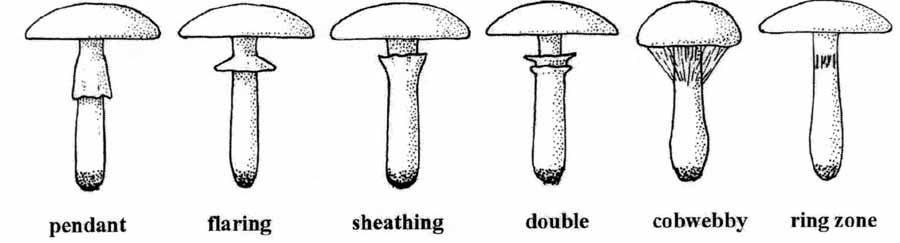
Mushroom gill tissue arrangement

[](https://datascienceplus.com/wp-content/uploads/2018/02/mushroom-gill-tissue-arrangement.jpg)

Mushroom stalk type

[](https://datascienceplus.com/wp-content/uploads/2018/02/mushroom-stalk.jpg)

Mushroom ring type

[](https://datascienceplus.com/wp-content/uploads/2018/02/mushroom-ring-type.jpg)

Given that we have data on mushrooms, this is a **standard supervised classification problem** where the label is a binary variable, 0 (edible mushroom), 1 (poisonous mushroom). In this study, our target variable Y is the probability of a mushroom is edible and poisonous.

**BENEFITS AND OPPURTUNITIES OF MUSHROOM BUSINESS**

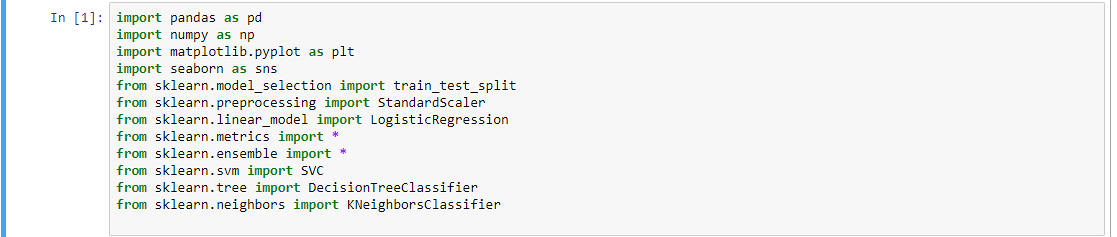
* **The Asia Pacific is the Leading Region in the Global Market** .The global mushroom market consumption was 12.74 million tons (MT) in 2018 and is projected to reach 20.84 million tons (MT) by 2026, exhibiting a CAGR of 6.41% in the forecast period.
* Mushroom consumption can potentially reduce the risk of diseases, such as prostate cancer and breast cancer.
* Rapidly expanding the foodservice industry is boosting the demand for mushroom across hotels, restaurants, and cafeterias.
* Growing demand for processed mushroom especially across western countries is generating export opportunities for major mushroom producing Asian countries.
* Adoption of the advanced pest management system to control damage caused due to pest infection in mushroom can further result in higher mushroom yields, which will ultimately aid the market growth of mushrooms.

In countries, such as South Korea, India, and others, where particular mushroom varieties such as button mushrooms are grown in abundance, the cost of fresh mushroom is comparatively lower. Low price of fresh mushroom as compared to certain processed mushroom forms may push the sales of fresh mushroom across the aforementioned countries.

**3. Data Analysis**

In this study, Mushroom Dataset is selected which contains data for 8124 rows with various information about the mushrooms. We will use this dataset to predict whether mushrooms are edible or poisonous and to find out the key indicators for edible and poisonous mushroom identification

**3.1 Importing Python libraries**



**3.2 Importing the data**

**3.2 Data source**

The type of documentation data is being kept in (i.e. whether it is paper-based, Excel spreadsheets, databases, etc) has a massive impact on the accuracy and the ease of access to the mushroom data.

Let's import the dataset and make of a copy of the source file for this analysis.  
The dataset contains 8124 rows and 23columns.

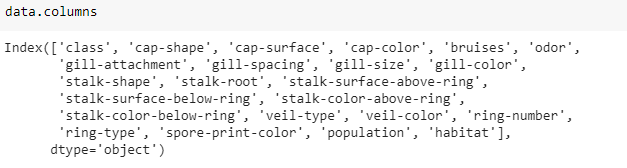


**3.3 Data Description and Exploratory Visualizations**

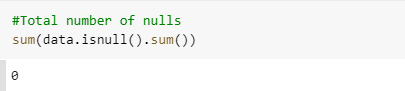
The dataset contains only categorical columns providing information about edible and poisonous mushrooms.

3.3.1**Overview**

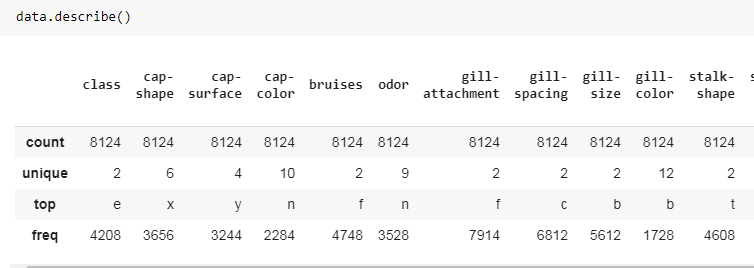
#dataset columns



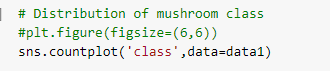
#null identification

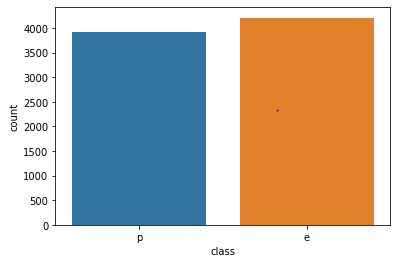


3.3.1 **Numerical features Overview**



[3.4   Target distribution](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Feature-distribution-by-target-attribute)





The above graph looks like a fairly balanced data set with an almost equal number of poisonous and edible mushrooms.

**Checking the data balance**

In the mushroom dataset the target variable contains two classes “edible “and “poisonous”. By plotting the class distribution of the data we can find that the dataset is balanced as it contains almost equal participation from both the classes. The above figure illustrates it.

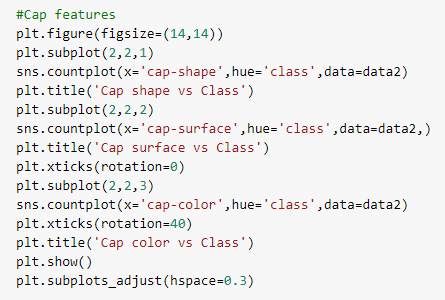
Now, I wanted to see how each feature affects the target. To do so, for each feature, I made a bar plot of all possible values separated by the class of mushroom.

[**3.5 Feature distribution by target attribute**](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Feature-distribution-by-target-attribute)

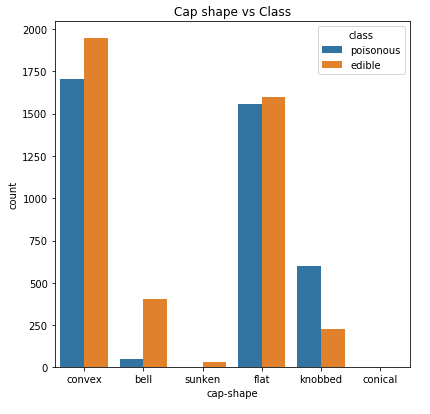


For feature distribution by target attribute each alphabets in the corresponding columns are replaced with its full names

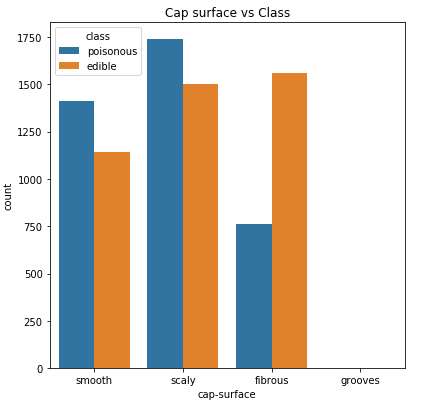
**3.5.1 cap features vs class**



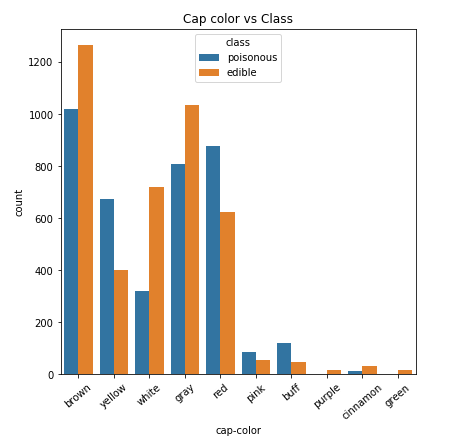
**3.5.1.1 Cap shape vs class**



**3.5.1.2 Cap surface vs class**



**3.5.1.3 Cap color vs class**

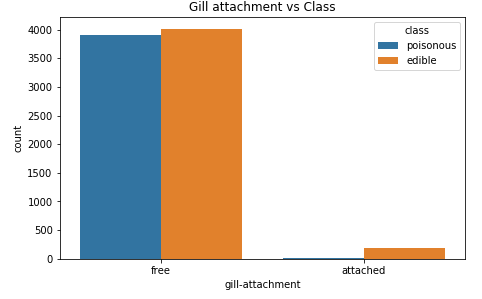


* Mushroom with sunken cap shape is edible  (but the number of data points is less)
* Cap surface doesn’t give much idea about the mushroom class.
* Purple and green cap color mushrooms are edible (less data points).

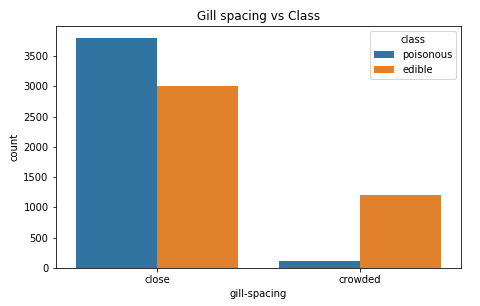
**3.5.2 Gill features vs class**



**3.5.2.1 gill attachment vs class**



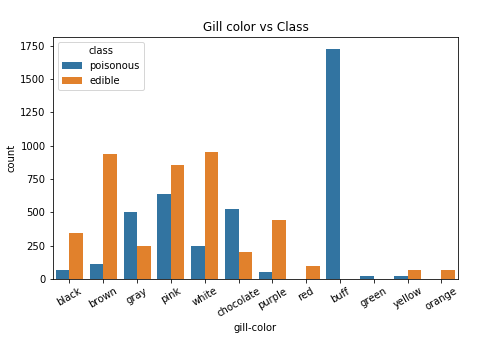
**3.5.2.2 gill spacing vs class**



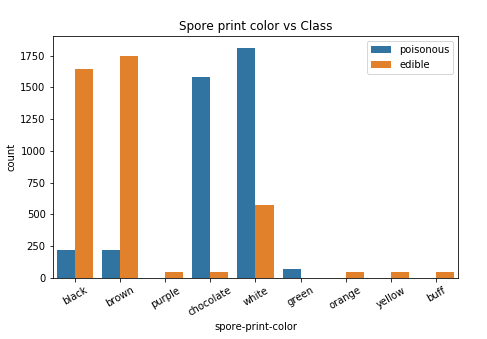
**3.5.2.3 gill size vs class**



**3.5.2.4 gill color vs class**



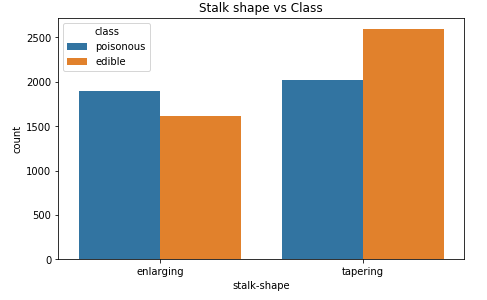
* + - 1. **Spore print color vs class**



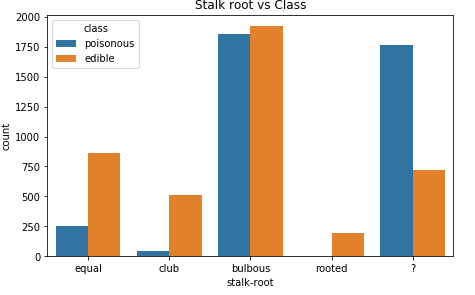
* Attached gill attachments are mostly edible.
* Crowded gill spacing are mostly edible.
* Narrow gill size is mostly poisonous.
* Red and orange gill color mushrooms are edible.
* Buff and green gill color mushrooms are poisonous.
* Mushroom with purple, orange, yellow and buff spore print color are edible.
* Mushroom with green spore print color is poisonous. Even in other spore print color the distribution is heavily skewed towards one class.
  + 1. Stalk **features**



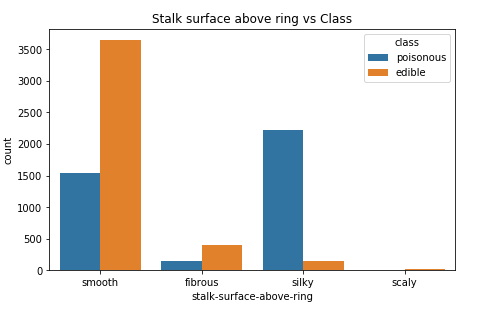
**3.5.3.1 stalk shape vs class**



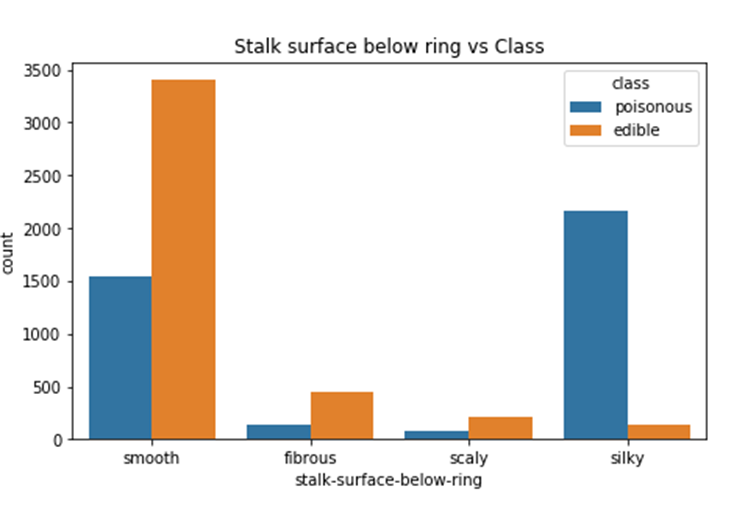
**3.5.3.2 Stalk root vs class**



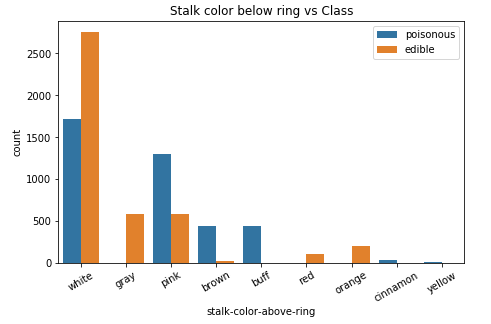
**3.5.3.3 Stalk surface above ring vs class**



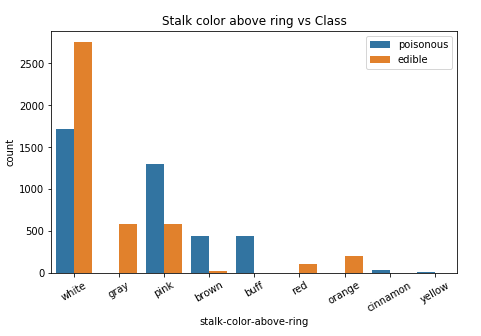
**3.5.3.4 Stalk surface below ring vs class**



**3.5.3.5 Stalk color below ring vs class**



**3.5.3.5 Stalk color above ring vs class**



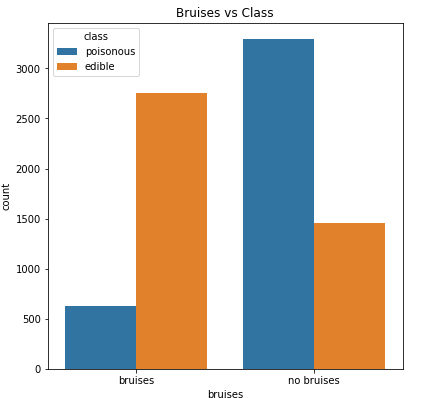
* Stalk shape in itself doesnt give an idea about mushroom class.
* Rooted stalk root mushrooms are edible.
* Club and equal stalk root mushrooms are mostly edible.
* Smooth stalk surface are mostly edible and silky ones are mostly poisonous. We can also see that stalk surface above and below ring have similar distribution.
* Stalk color above and below ring are also similar. Grey, red and orange stalk colors means the mushroom is edible. Buff, cinnamon and yellow stalk color mean the mushroom is poisonous.  
  + 1. **Miscellaneous features**



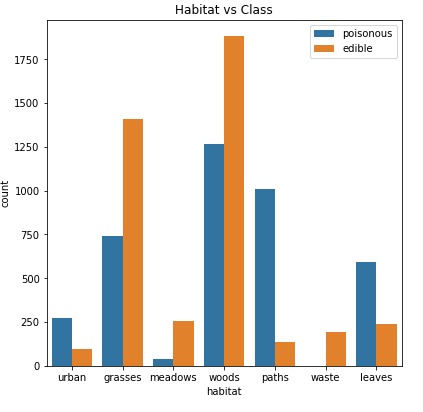
**3.5.4.1 odor vs class**

## 

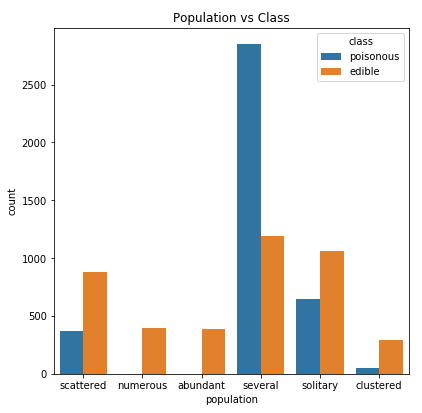
**3.5.4.2 Bruises vs class**



**3.5.4.3 Habitat vs class**



**3.5.4.4 Population vs class**



* + Odor can be used to determine the class of the mushroom except for mushrooms with none smells.
  + Bruises in itself can’t be used to predict class.
  + Mushrooms growing in waste habitat are edible!!
  + Mushroom displaying numerous and abundant population characteristic are edible.

**3.6 CORRELATION**

In this mushroom dataset almost all features are categorical so normal correlation technique is not possible, instead of that we can go for chi-square test in order to check for the significative relationship between mushroom features and their classification as edible or poisonous.

The**Chi Square Test** is a test that involves the use of parameters to test the statistical significance of the observations under study. The task of the chi square test is to test the statistical significance of the observed relationship with respect to the expected relationship. The chi square statistic is used by the researcher for determining whether or not a relationship exists.

The calculation of the statistic in the chi square test is done by computing the sum of the square of the deviation between the observed and the expected frequency, which is divided by the expected frequency.

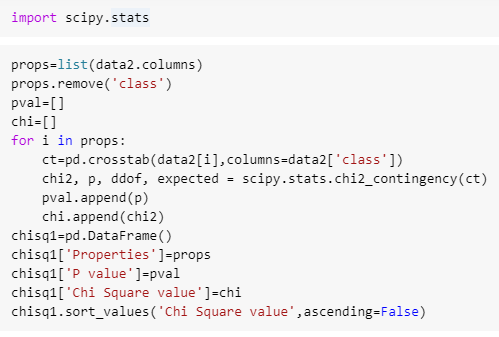
http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTesting-ChiSquare/lessonimages/equation_image7.gif

O=Observed frequency

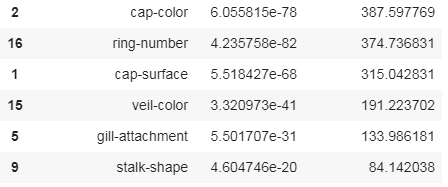
E=Expected frequency

The researcher should know that the greater the difference between the observed and expected cell frequency, the larger the value of the chi square statistic in the chi square test.

In order to determine if the association between the two variables exists, the probability of obtaining a value of chi square should be larger than the one obtained from the chi square test of cross tabulation.



## 



From the above table it is clear that the chi square value for the feature odor is highest and the value is 7659.72. The next features are spore print color, gill-color, ring-type and so on

**3.7 EDA CONCLUDING REMARKS**

* + only poisonous mushrooms have convex cap-shape; only edible mushrooms have sunken cap-shape
  + only poisonous mushrooms have cap-surface with grooves
  + only edible mushrooms have green or purple cap-color
  + odor is strongly indicative of what mushrooms are (edible/poisonous)
  + only poisonous mushrooms have buff or green gill color
  + only edible mushrooms have red or orange gill color
  + only edible mushrooms have rooted stalk root
  + stalk color above ring and stalk color below ring are relevant features for out classification problem
  + only edible mushrooms have brown veil color
  + only poisonous mushrooms have yellow veil color
  + only poisonous mushrooms do not have rings
  + only edible mushrooms have flaring ring type
  + only poisonous mushrooms have none ring type
  + only edible mushrooms have black, orange, purple or yellow spore print color
  + only poisonous mushrooms have green spore print color
  + only edible mushrooms have abundant or numerous population
  + only edible mushrooms have waste type habitat

[**4.  Pre-processing Pipeline**](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Pre-processing-Pipeline)

[**4.1 Encoding**](https://www.kaggle.com/hamzaben/employee-churn-model-w-strategic-retention-plan#Encoding)

There are too many types mushroom species present in the world. Many of them are deadly poisonous. Our main objective is to identify the edible mushrooms. We build a classifier using machine learning techniques for this purpose. Mainly *logistic regression, SVM, decision tree and random forest* models are considered.

**Checking the data balance and missing values**

The dataset contains 8124 rows with 22 features. The target class contains two classes “edible “and “poisonous”. By plotting the class distribution of the data we can find that the dataset is balanced as it contains almost equal participation from both the classes. The following figure illustrates it. As the data is balanced it is the ideal case for classification. No need of applying any advanced sampling techniques.



Fig: data balance

In the dataset we can see that there are no missing values present and we can make use of all the data for our training and testing. But found some entries in the ‘stalk root’ column marked with ‘?’.A total of 2480 entries are marked with ‘?’ and it is not a good idea to drop it. We consider it as a label for classification. The ‘veil\_type’ feature contains only one value ‘p’. This maybe dropped as it will not affect the classification for our dataset.

**Converting categorical data to numerical**

The dataset contains categorical values and it is important to convert the values into numerical for performing the machine learning tasks. Here the target column contains “e” and “p” for edible and poisonous categories. We can convert it to 0 or 1 using either replacing or by using label encoder as it contains only two classes.

We here first perform label encoding on the dataset so that the categorical values are converted into numerical values. (y –target variable) contains 1 for poisonous class and 0 for edible class.

The machine learning algorithms will give more importance to higher numerical values and can have false correlations, hence we have to perform one hot encoding to avoid this biasing in the independent variables (X). But one hot encoding will increase the number of columns.

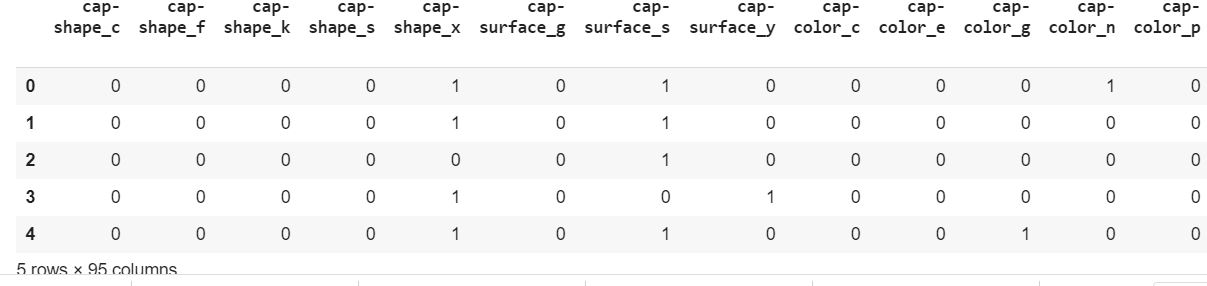
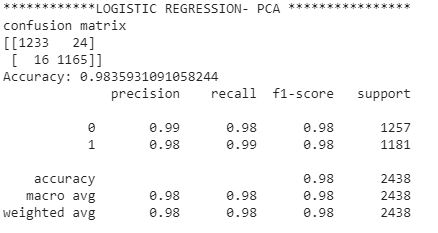


Fig :one hot encoding

**Train test split and Dimensionality reduction**

Now we split the data into test and training set and apply scaling on both test and training sets. Splitting the dataset into train and test set will help to evaluate the models on new data that has not seen before i.e. the test data. Here random state is set as 42 for reproducing results later.

After the one hot encoding the dataset contains around 95 columns. We here tried the PCA method to study the feasibility of dimensionality reduction. In our experiment with logistic regression on PCA applied data it did not work well as for achieving 98% accuracy the number of components required was as large as 20. So we continue with all the features for prediction which gave perfect results. The following figure shows performance of logistic regression model with PCA applied data for number of components 15.



**Fig:** logistic regressionon PCA data

**Models**

For the mushroom classification we mainly tried logistic regression model, SVM, Decision tree and Random forest classifiers on the dataset to select the best performing one.

It is very important to evaluate the predictions made by our models. We here make use of the following measures.

* **Accuracy**: It is the proportion of total number of correct results. High accuracy is nice to have for a model.
* **Precision**: Proportion of correct positive results out of all predicted positive results. We can make use of this measure when having false negative is not an issue.
* **Recall or Sensitivity**: Proportion of actual positive cases. This measure is used when cost or risk of false negative is high.
* **F1 score**: Combined measure of both precision and accuracy. It is more practical.
* **Specificity**: Proportion of actual negative cases.

It is important to adjust the threshold value for the classification. In our case having poisonous mushrooms can cause death. So we cannot afford at least one false positive. So we adjust the prediction probability accordingly reducing the risk factor.

**AUC ROC CURVE**

Using too many evaluation measures jointly such as precision recall etc. to find the best model is confusing. So it is good to use ROC, AUC metrics by which we will get a single score for evaluating the model. ROC stands for receiver operating characteristics and AUC stands for area under curve. It is simply a plot with the true positive rate on Y axis and false positive rate on the X axis. Area under the ROC curve is an indication of how good our model is.

**Logistic regression**

Logistic regression model can be used to predict the probability of an outcome. This model can be used for both binary and multiple class problems. Here in our problem based on the feature values, the model can predict whether it belongs to edible or poisonous category. It is determined by applying a threshold value.

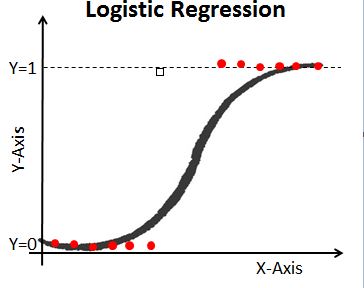


Fig:logistic regression

If we want to be more sure about the certainty of an event, we can increase or decrease the threshold value. Here for our problem the implications are much severe. Eating poisonous mushroom can even cause death.

The following figure shows the classification performance of logistic regression model on our data.

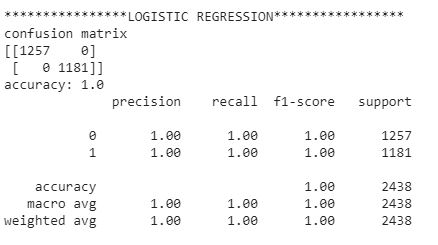


Fig: logistic regression model metrics

The above result shows that the model is performing very well on the data. The model correctly classified without any false positive or false negative.

The ROC curve is shown below and we can see that it is a perfect model.

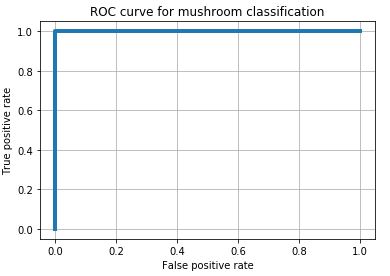


Fig : ROC logistic regression

Even though the model is perfect we cannot take a risk since consuming poisonous mushroom can cause death. For that we adjusted the probability associated with prediction i.e. the threshold value for the classification to edible and poisonous class. For our dataset the model is so confident it can correctly classify poisonous mushroom with probability of 0.98.

**SVM- Support vector machine**

The basic idea behind the model support vector machine is a hyperplane separating the two distinct classes. This supervised model can be used for both regression and classification problems. Here the hyperplane is a maximum distance line from the nearest distance data points of the two separate classes. We here find the magnitude of vector to data points and compare its magnitude for classifying the data point to corresponding classes. There is a possibility that the data may not be linearly separable, in such cases we have to create another plane to classify the data. Different kernel functions are there such as RBF, Polynomial, Sigmoid etc. for the nonlinear data points.

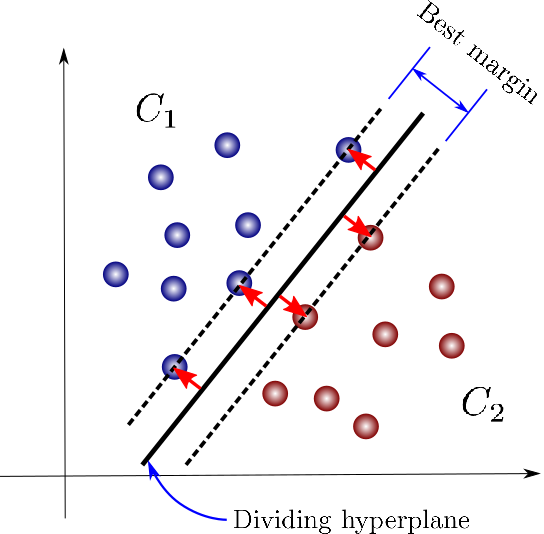


Fig : SVM hyperplane

We got the following result for the mushroom dataset. SVM also made perfect classification.

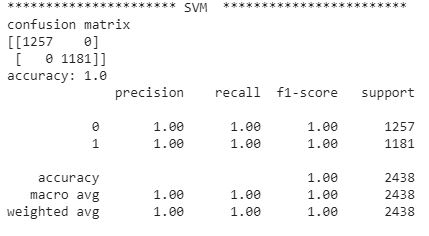


Fig : SVM Evaluation metrics

**Decision Tree**

The decision tree model is also a supervised method. Basically it is a decision support tool that uses a tree like graph or model of decisions and their possible consequences. This can be used in the case of both categorical as well as continuous variables. This model is useful when there many features present in determining the outcome. In such cases each feature will be having a cumulative effect. The tree goes on splitting till it gets a pure subset. For a new data traversing through the tree can determine the classification.

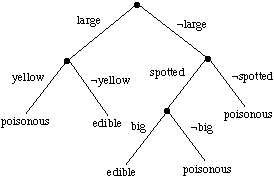


Fig : decision tree example

The mushroom dataset contains features such as color, shape, odor, habitat etc. which are suitable for making decisions. The performance of decision tree model is shown below. This model also made perfect classification.

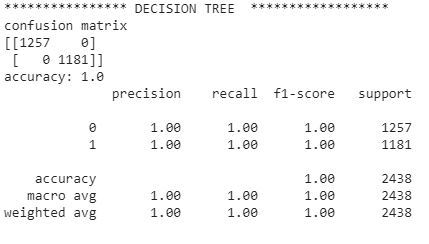


Fig : Decision Tree evaluation metric

The following table shows the sorted feature importance in decision tree model. We can find which features are mainly contributing to the target class.

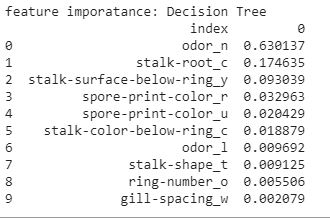
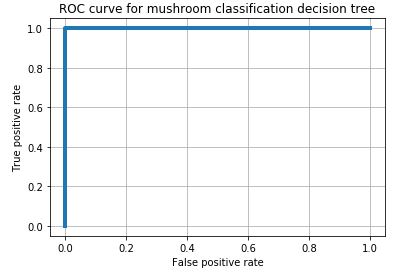
.

Table: Feature importance Decision Tree

The roc curve for the model is shown below.



Fir : ROC curve Decision tree

**Ensemble learning and random forests**

Random forest model is an improvement in the original concepts of decision tree. We all know that collective wisdom is higher than the individual intelligence. This concept is used in the case of ensemble learning. Here we generate group of base learners and combined result gives higher accuracy.

There are mainly two major ensemble learning methods

* Bagging

Here we split the data to create bag for different base learners which are built in parallel. All models vote to give the final prediction.

* Boosting

The decision trees are trained in sequence so that a tree can learn from the previous one by focussing on the incorrect observations with higher weight.

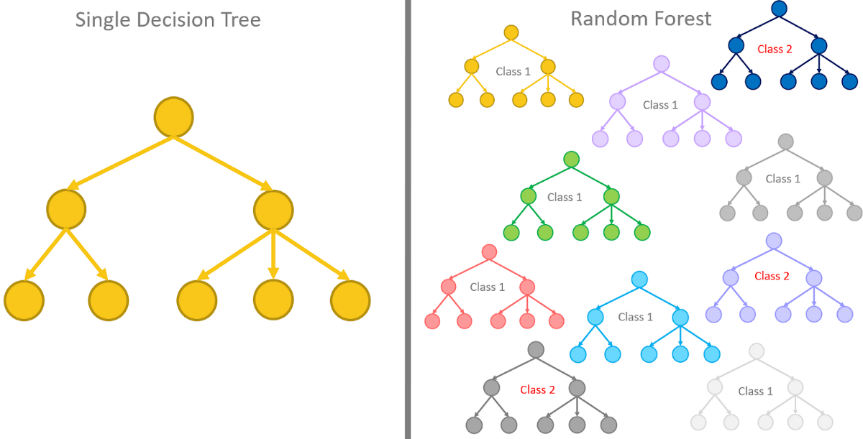


Fig: Random forest example

Random forest model performs better on most situations on datasets as it is an ensemble model. For the mushroom dataset also the model is perfect. The following figure shows evaluation metrics of the model.

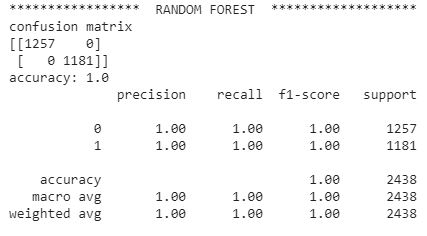


Fig : random forest evaluation metrics

The feature importance for the random forest model is shown in the table below.

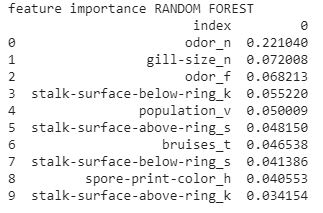


Table :feature Importance random forest

The ROC curve for the model is shown in the following figure.

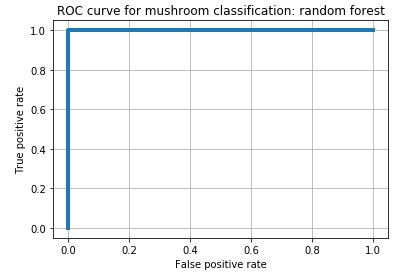


Fig:Roc curve random forest

Almost all the models made perfect predictions with the dataset even without tuning the parameters. We tried different ratios for the test and train data split and found that the models are able to make good predictions. Even with 0.2 percentage of data used for training the models gave good predictions accurately. This implies that the features we use for classification are having good effect on the target class. When we applied strategies like dropping different columns, reducing the number of features etc. the models were able to make efficient predictions without degradation.

**Final predictions**

By combining all the results from above we get the the final predictor framework:

1. If odor is almond or anise, then mushroom is edible.
2. If odor is pungent, foul, spicy, musty, fishy, creosote then mushroom is poisonous.
3. If odor is none and spore print color is green, then mushroom is poisonous.
4. If odor is none and spore print color is white, then we go to habitat and population properties.
5. If population is numerous, scattered or solitary, then mushroom is edible.
6. If population is several and habitat is leaves or path, then mushroom is edible.
7. If population is clustered and habitat is waste, then mushroom is edible.
8. If population is clustered and habitat is leaves, then mushroom is poisonous.
9. In all other cases of odor being none then mushroom is edible.

**6. CONCLUSION**

From the project It can be concluded that almost all classification algorithms are performing similar. The edible mushrooms have green or purple cap color, red or orange gill color, rooted stalk root, brown veil color , flaring ring type, black, orange, purple or yellow spore print color, abundant or numerous population and waste type habitat. From the algorithms we have understood that the edible mushrooms belong to the category of 2 % and poisonous mushrooms belongs to remaining 98%. So the risky category can be eliminated. These classification algorithms can be used in companies for predicting the nature of prospective mushrooms. Based on the same technique companies can distribute and export edible mushrooms in successful manner.

**REFERENCE**

Schlimmer,J.S. (1987). Concept Acquisition Through Representational Adjustment (Technical Report 87-19). Doctoral disseration, Department of Information and Computer Science, University of California, Irvine.  
  
Iba,W., Wogulis,J., & Langley,P. (1988). Trading off Simplicity and Coverage in Incremental Concept Learning. In Proceedings of the 5th International Conference on Machine Learning, 73-79. Ann Arbor, Michigan: Morgan Kaufmann.  
  
  
Duch W, Adamczak R, Grabczewski K (1996) Extraction of logical rules from training data using backpropagation networks, in: Proc. of the The 1st Online Workshop on Soft Computing, 19-30.Aug.1996, pp. 25-30,   
  
Duch W, Adamczak R, Grabczewski K, Ishikawa M, Ueda H, Extraction of crisp logical rules using constrained backpropagation networks - comparison of two new approaches, in: Proc. of the European Symposium on Artificial Neural Networks (ESANN'97), Bruge, Belgium 16-18.4.1997.

# Classification algorithm for edible mushroom identification

**https://ieeexplore.ieee.org/xpl/conhome/8343062/proceeding**