|  |  |
| --- | --- |
| **P170M109 Computational Intelligence and Decision** | **Date: 30 May 2018** |
| Individual Task | **Student: Mattia Fucili** |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



**Introduction**

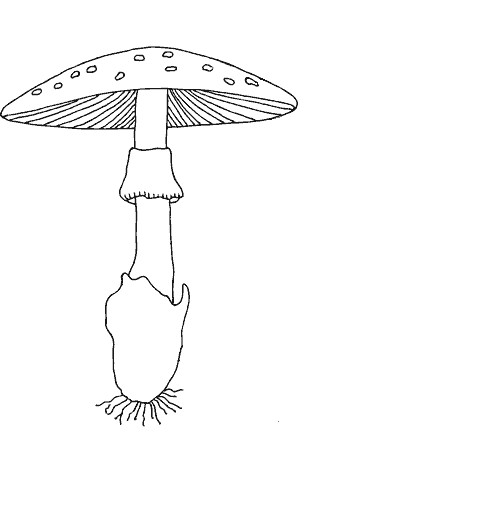
This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in two different Families Mushroom from North America. Each row of the dataset contains a specific mushroom labelled either as definitely edible or definitely poisonous.

The challenge to face is to determine if a mushroom is poisonous or not, obviously with an accuracy as higher as possible!

The structure of the dataset is summarized in this table

|  |  |  |
| --- | --- | --- |
| Name | Type | Description |
| class | String | Class label of a mushroom. edible, poisonous |
| cap-shape | String | Shape of the cap. Examples: bell, conical, … |
| cap-surface | String | Surface of the cap. Examples: grooves, smooth, … |
| cap-color | String | Color of the cap. Examples: brown, red, … |
| bruises | Numeric | Bruises in the mushroom. present, absent |
| odor | String | Odor of the mushroom. Examples: spicy, fishy, … |
| gill-attachment | String | Attachment type of the gill. Examples: attached, free, … |
| gill-spacing | String | Spacing between gills. Examples: close, distant, … |
| gill-size | Numeric | Size of the gill. broad, narrow |
| gill-color | String | Color of the gill. Examples: black, green, |
| stalk-shape | Numeric | Shape of the stalk. enlarging, tapering |
| stalk-root | String | Root shape of the stalk. Examples: bulb, rooted, … |
| stalk-surface-above-ring | String | Surface of the top stalk. Examples: scaly, smooth, … |
| stalk-surface-below-ring | String | Surface of the bottom stalk. Examples: scaly, smooth, … |
| stalk-color-above-ring | String | Color of the top stalk. Examples: brown, white, … |
| stalk-color-below-ring | String | Color of the bottom stalk. Examples: brown, white, … |
| veil-type | String | Type of the veils. partial, universal |
| veil-color | String | Color of the veils. Examples: white, yellow, … |
| ring-number | Numeric | Number of rings in the stalk. 0, 1, 2 |
| ring-type | String | Type of the rings in the stalk. Examples: none, pendant, … |
| spore-print-color | String | Color of the spores. Examples: black, brown, … |
| population | String | Number of individuals. Examples: abundant, solitary, … |
| habitat | String | Habitat. Examples: grass, woods, … |

Visual explanation of the features in the dataset



type

color

shape

root

surface above ring

surface below ring

color above ring

color below ring

number

type

spore print color

shape

surface

color

cap

veil

stalk

ring

gills

attachment

spacing

size

color

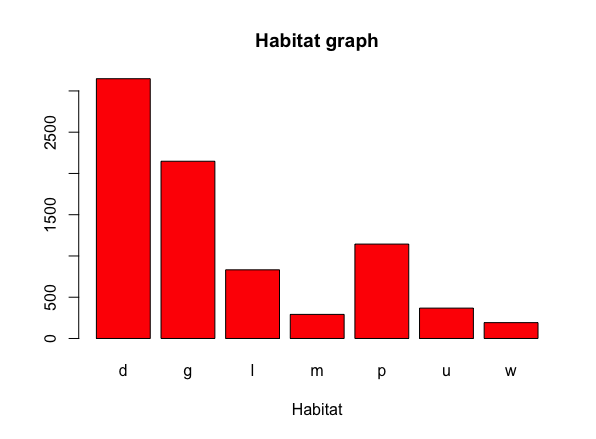
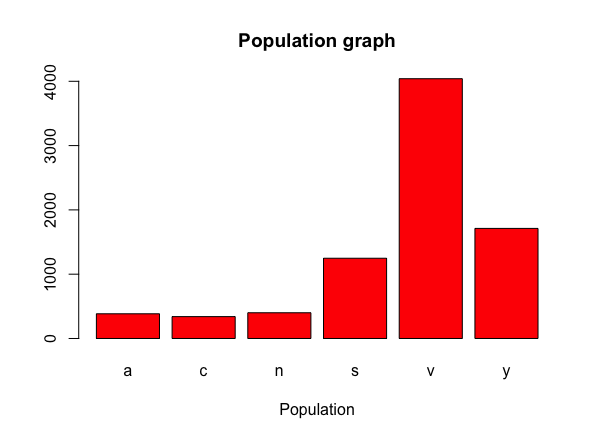
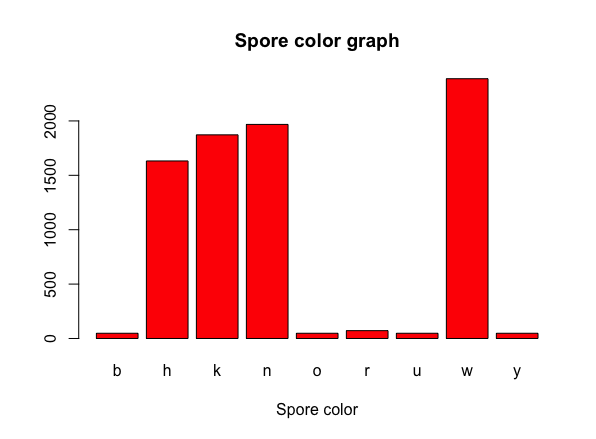
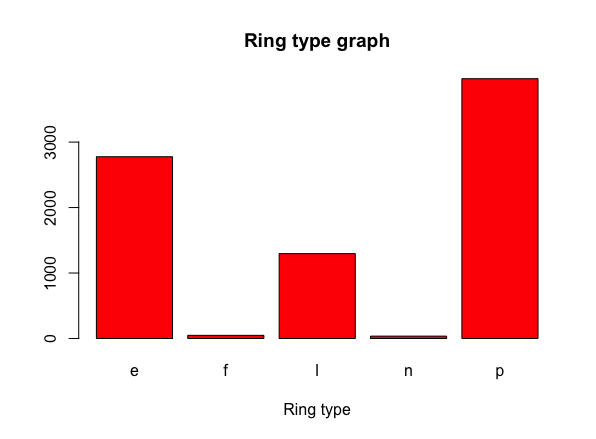
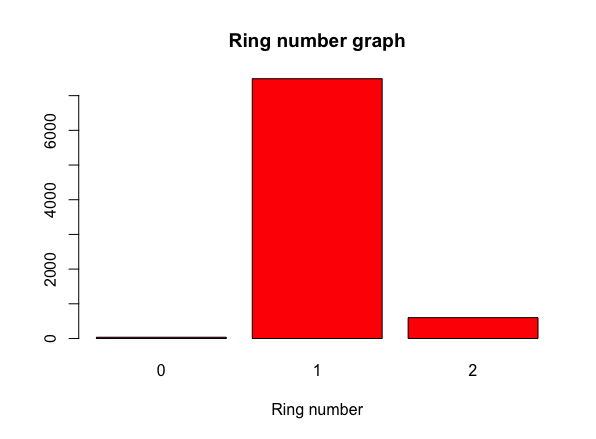
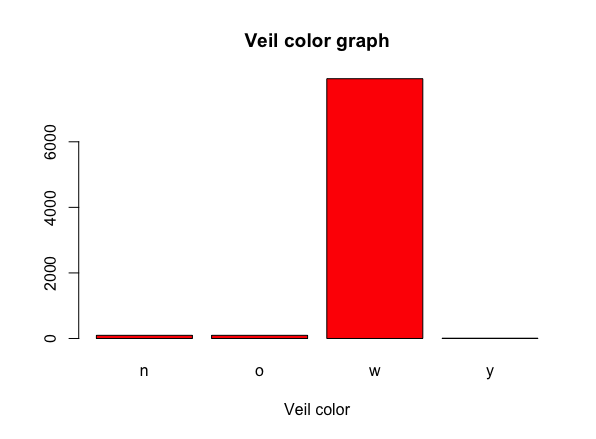
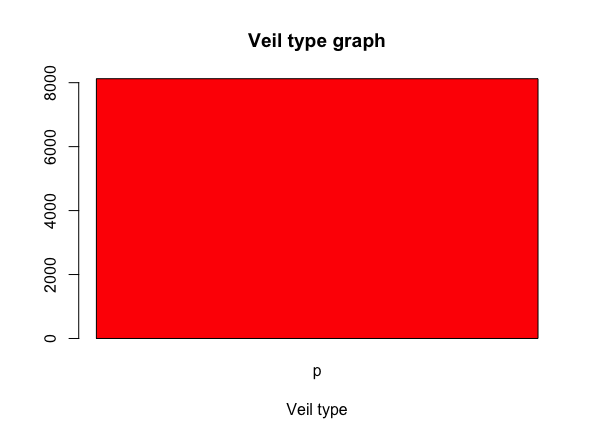
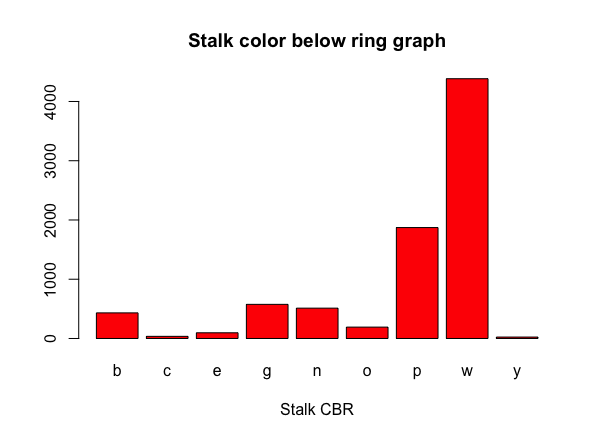
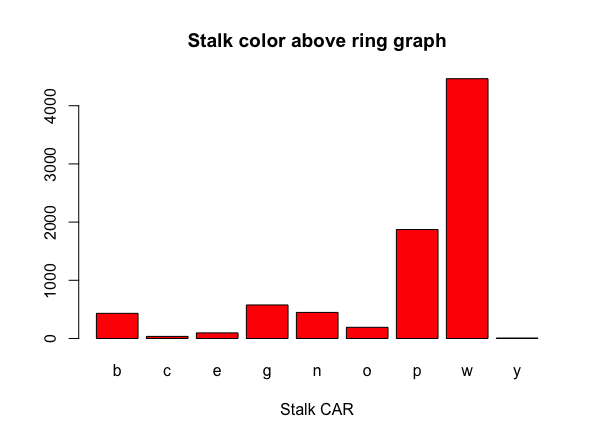
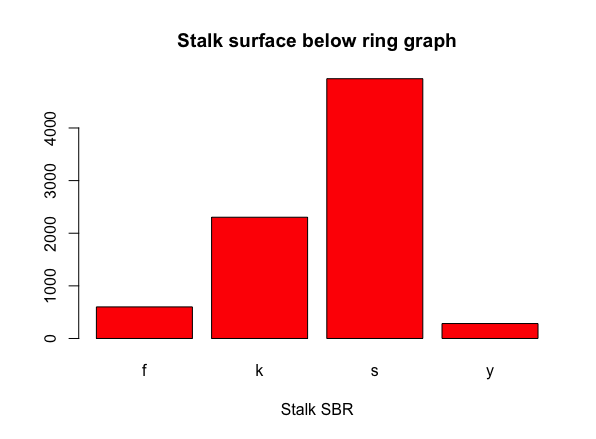
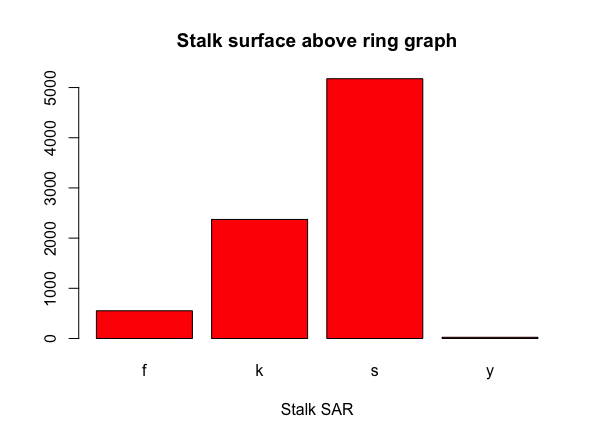
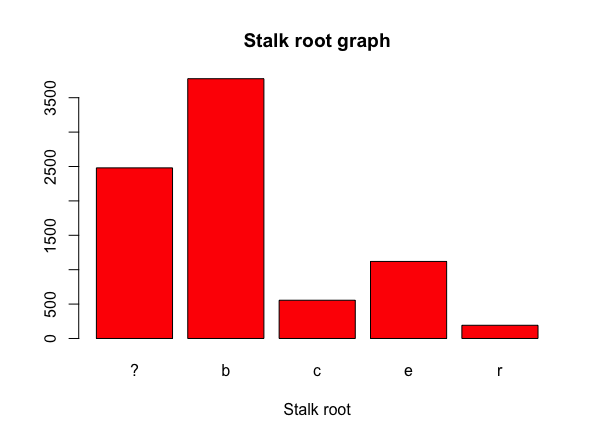
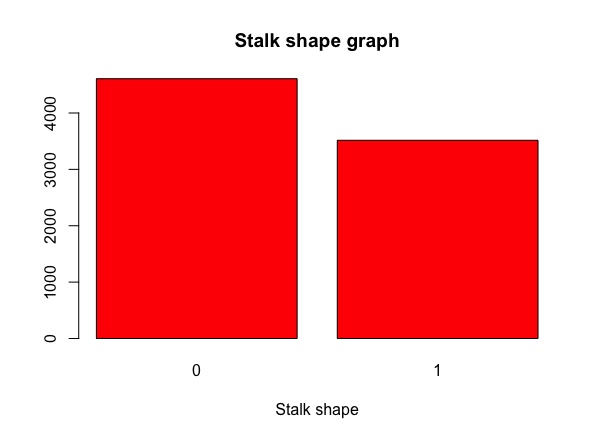
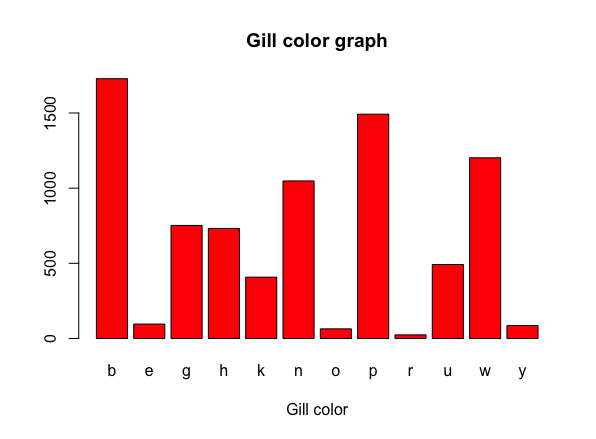
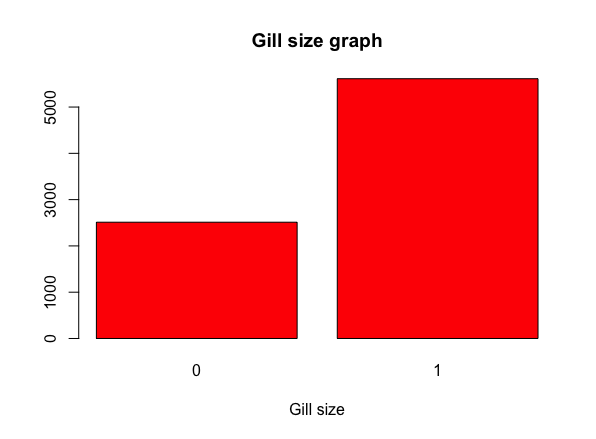
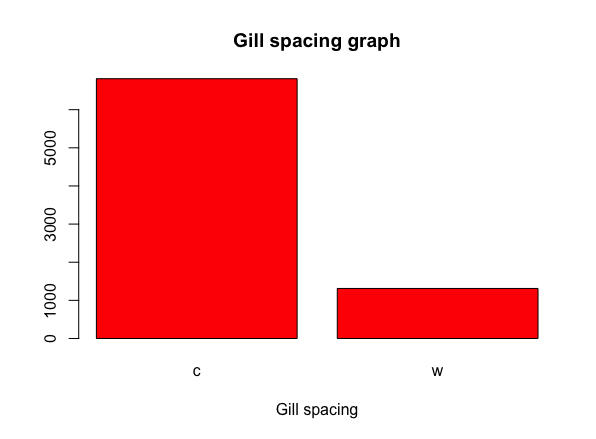
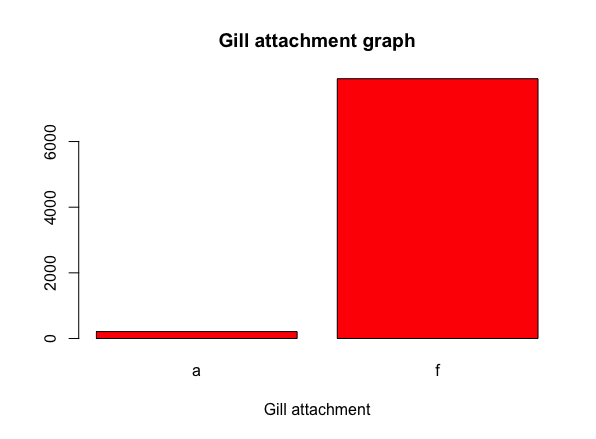
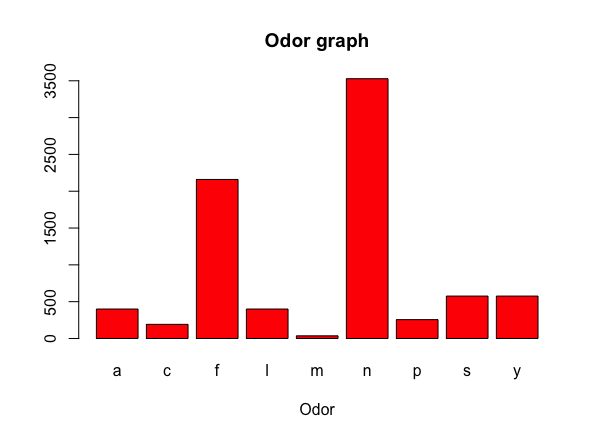
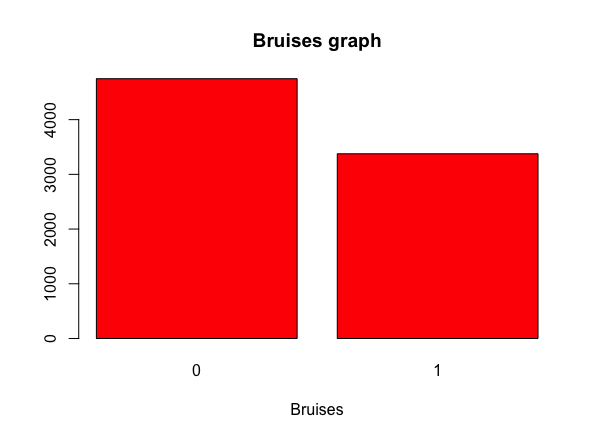
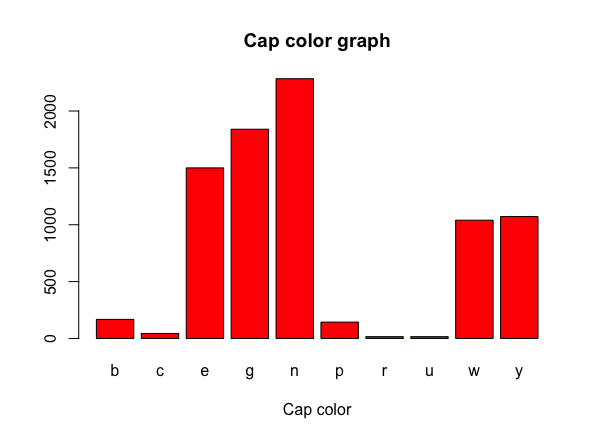
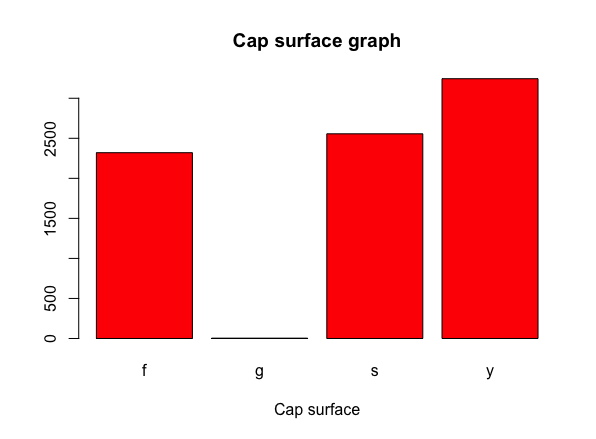
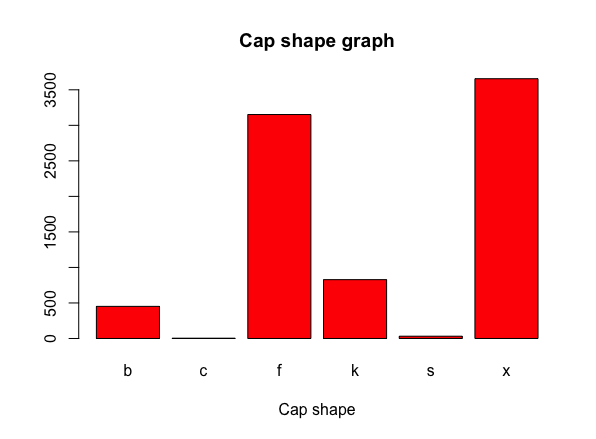
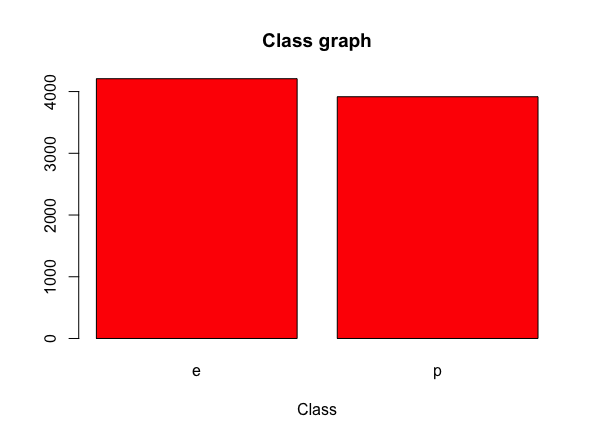
**Data Analysis**

The following table summarizes the analysis of the features in the dataset

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Count | % Missing | Card | Mode | Mode freq | mode  % | 2nd mode | 2nd mode freq | 2nd mode % |
| class | 8124 | 0% | 2 | e | 4208 | 52% | p | 3916 | 48% |
| cap-shape | 8124 | 0% | 6 | x | 3656 | 45% | f | 3152 | 39% |
| cap-surface | 8124 | 0% | 4 | y | 3244 | 40% | s | 2556 | 31% |
| cap-color | 8124 | 0% | 10 | n | 2284 | 28% | g | 1840 | 23% |
| bruises | 8124 | 0% | 2 | 0 | 4748 | 58% | 1 | 3376 | 42% |
| odor | 8124 | 0% | 9 | n | 3528 | 43% | f | 2160 | 27% |
| gill-attachment | 8124 | 0% | 2 | f | 7914 | 97% | a | 210 | 3% |
| gill-spacing | 8124 | 0% | 2 | c | 6812 | 84% | w | 1312 | 16% |
| gill-size | 8124 | 0% | 2 | 1 | 5612 | 69% | 0 | 2512 | 31% |
| gill-color | 8124 | 0% | 12 | b | 1728 | 21% | p | 1492 | 18% |
| stalk-shape | 8124 | 0% | 2 | 0 | 4608 | 57% | 1 | 3516 | 43% |
| stalk-root | 8124 | 0% | 5 | b | 3776 | 47% | ? | 2480 | 31% |
| stalk-surface-above-ring | 8124 | 0% | 4 | s | 5176 | 64% | k | 2372 | 29% |
| stalk-surface-below-ring | 8124 | 0% | 4 | s | 4936 | 61% | k | 2304 | 28% |
| stalk-color-above-ring | 8124 | 0% | 9 | w | 4464 | 55% | p | 1872 | 23% |
| stalk-color-below-ring | 8124 | 0% | 9 | w | 4384 | 54% | p | 1872 | 23% |
| veil-type | 8124 | 0% | 1 | p | 8124 | 100% | - | - | - |
| veil-color | 8124 | 0% | 4 | w | 7924 | 98% | n/o | 96 | 0,01% |
| ring-number | 8124 | 0% | 3 | 1 | 7488 | 92% | 2 | 600 | 0,07% |
| ring-type | 8124 | 0% | 5 | p | 3968 | 49% | e | 2776 | 34% |
| spore-print-color | 8124 | 0% | 9 | w | 2388 | 29% | n | 1968 | 24% |
| population | 8124 | 0% | 6 | v | 4040 | 50% | y | 1712 | 21% |
| habitat | 8124 | 0% | 7 | d | 3148 | 39% | g | 2148 | 26% |

As we can see there aren’t any missing values in the features, but some of them are not well distributed. An example is veil-color where the value corresponding to the first mode is the 98% of the total values.

A graphical representation of the data is given below.



As we can see from the table and the graphs, the target feature is well distributed over the dataset. There is also a feature, veil-type, that has cardinality of 1 indicating that every row has the same value. This mean that this feature should be removed. The features stalk-color-above-ring and stalk-color-below-ring are very similar looking to the graphs but differs just a bit in some categories, so they can’t be merged. The same analysis could be done for stalk-surface-above-ring and stalk-surface-below-ring. Moreover, we cannot merge some values in other features, even if they are less present, because they mean totally different. An example is ring-type where we can see that f and n are the shorter bins, but one means flaring and the other none, so they cannot be merged.

The table that summarize the final ABT is the following

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Count | % Missing | Card | Mode | Mode freq | mode  % | 2nd mode | 2nd mode freq | 2nd mode % |
| class | 8124 | 0% | 2 | e | 4208 | 52% | p | 3916 | 48% |
| cap-shape | 8124 | 0% | 6 | x | 3656 | 45% | f | 3152 | 39% |
| cap-surface | 8124 | 0% | 4 | y | 3244 | 40% | s | 2556 | 31% |
| cap-color | 8124 | 0% | 10 | n | 2284 | 28% | g | 1840 | 23% |
| bruises | 8124 | 0% | 2 | 0 | 4748 | 58% | 1 | 3376 | 42% |
| odor | 8124 | 0% | 9 | n | 3528 | 43% | f | 2160 | 27% |
| gill-attachment | 8124 | 0% | 2 | f | 7914 | 97% | a | 210 | 3% |
| gill-spacing | 8124 | 0% | 2 | c | 6812 | 84% | w | 1312 | 16% |
| gill-size | 8124 | 0% | 2 | 1 | 5612 | 69% | 0 | 2512 | 31% |
| gill-color | 8124 | 0% | 12 | b | 1728 | 21% | p | 1492 | 18% |
| stalk-shape | 8124 | 0% | 2 | 0 | 4608 | 57% | 1 | 3516 | 43% |
| stalk-root | 8124 | 0% | 5 | b | 3776 | 47% | ? | 2480 | 31% |
| stalk-surface-above-ring | 8124 | 0% | 4 | s | 5176 | 64% | k | 2372 | 29% |
| stalk-surface-below-ring | 8124 | 0% | 4 | s | 4936 | 61% | k | 2304 | 28% |
| stalk-color-above-ring | 8124 | 0% | 9 | w | 4464 | 55% | p | 1872 | 23% |
| stalk-color-below-ring | 8124 | 0% | 9 | w | 4384 | 54% | p | 1872 | 23% |
| veil-color | 8124 | 0% | 4 | w | 7924 | 98% | n/o | 96 | 0,01% |
| ring-number | 8124 | 0% | 3 | 1 | 7488 | 92% | 2 | 600 | 0,07% |
| ring-type | 8124 | 0% | 5 | p | 3968 | 49% | e | 2776 | 34% |
| spore-print-color | 8124 | 0% | 9 | w | 2388 | 29% | n | 1968 | 24% |
| population | 8124 | 0% | 6 | v | 4040 | 50% | y | 1712 | 21% |
| habitat | 8124 | 0% | 7 | d | 3148 | 39% | g | 2148 | 26% |

Since all data are categorical is meaningless to compute outliers and cross correlation.

It was decided to use information gain to have more information about data and, moreover, to know which data are most predictive. After analyzing the information gain of all features, the ones that could be considered relevant are the following: odor (0.63), spore-print-color (0.33), gill-color (0.29), ring-type (0.22) and stalk-surface-above-ring (0.20).

**Methods applied**

First of all, it has been decided to split the entire dataset into train and test sets. To do that the 80% of the dataset has been used as training set and the remaining 20% has been used as testing set.

Below is described, in a table, the proportion of class target feature in training and test set

|  |  |  |
| --- | --- | --- |
|  | train set | test set |
| % edible | 51% | 53% |
| % poisonous | 49% | 47% |

**Decision tree**

The first method applied is decision tree, in particular, CART algorithm has been used. The method was applied in python using sklearn library and in order to use decision tree algorithm data must be transform in numerical features. To do so has been used another library from sklearn specific for this task with the method LabelEncoder().

After various tests using different train and test sets, due to random method used to obtain them, is possible to say that the accuracy varies between 99,8% to 100%. AUC result is between 0,998 to 1,000.

Here an example of confusion matrix

|  |  |  |
| --- | --- | --- |
|  | edible | poisonous |
| edible | 831 | 0 |
| poisonous | 1 | 818 |

**SVM**

The other method that has been applied in this dataset is Support Vector Machine. For this method has been used python and sklearn as well. Using the same library used in decision tree step, data has been transformed in numerical features.

After various tests using different train and test sets, due to random method used to obtain them, is possible to say that the accuracy varies between 99,4% to 99,9%. AUC result is between 0,994 to 0,998.

Here an example of confusion matrix

|  |  |  |
| --- | --- | --- |
|  | edible | poisonous |
| edible | 804 | 0 |
| poisonous | 5 | 813 |

**Conclusion**

Tests between decision tree and SVM are made with the same train and test set, and as we can see from the results, a CART algorithm performs slightly better than SVM. For this delicate task to predict if a mushroom is edible or poisonous this small difference is very important, so I would suggest to use a decision tree algorithm.