P170M109 Computational Intelligence and Decision

Individual Task

Date: 30 May 2018

Student: Mattia Fucili

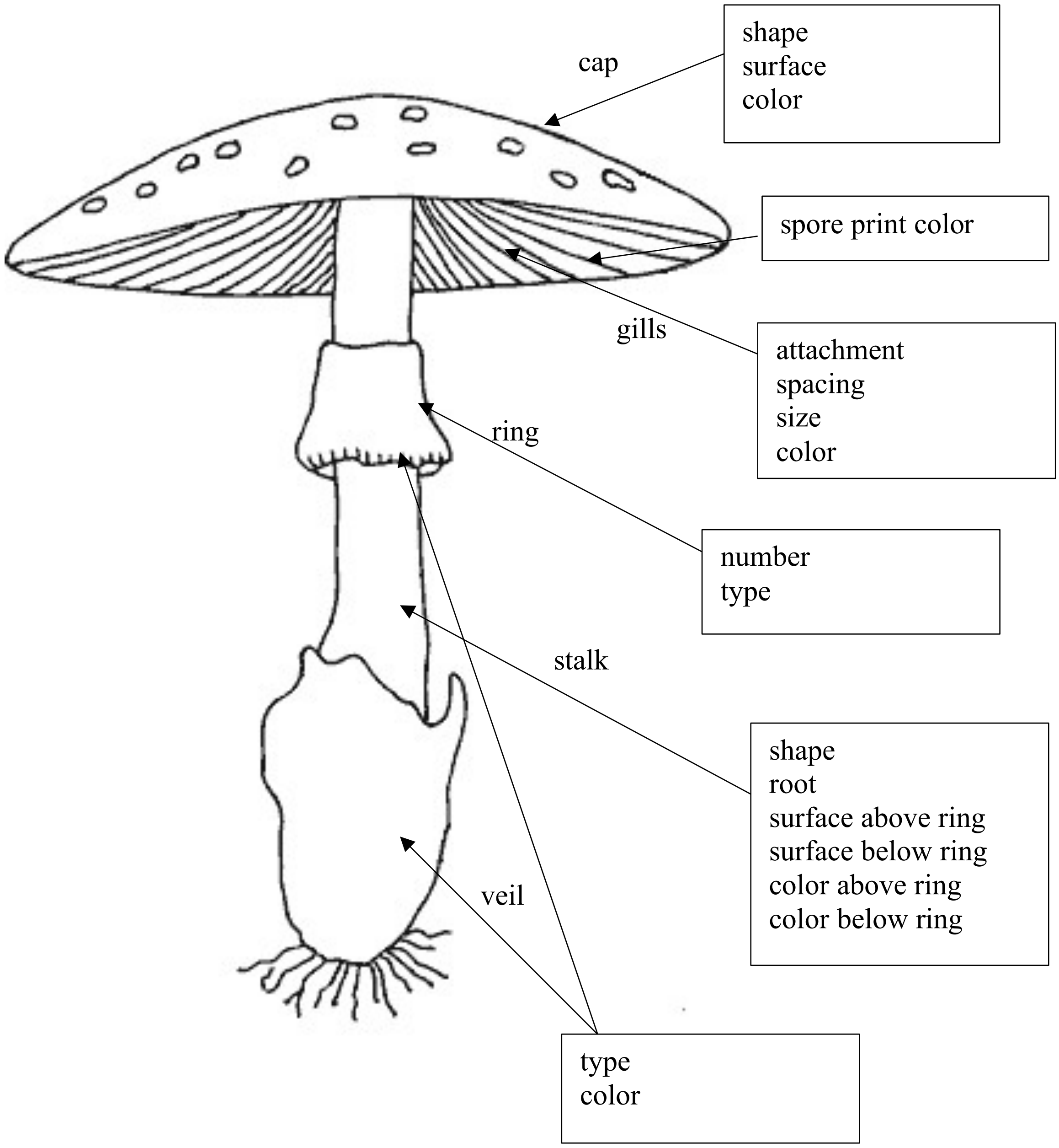
Introduction

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



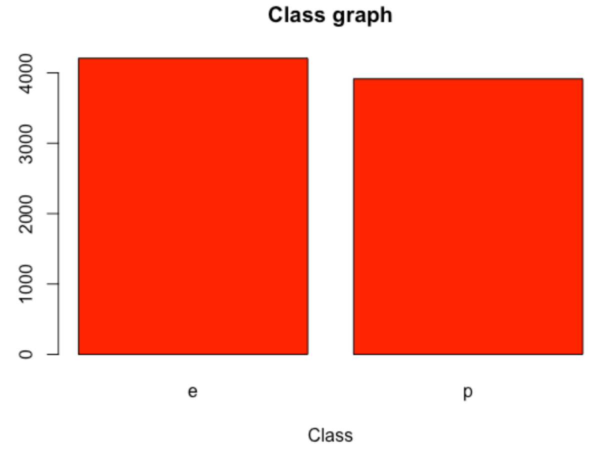
Data Analysis

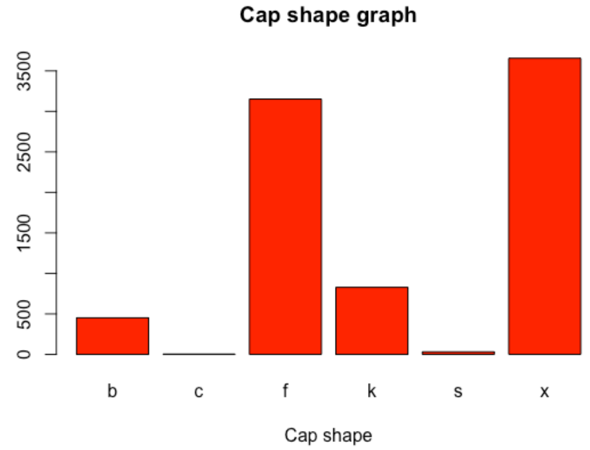
The following table summarizes the analysis of the features in the dataset

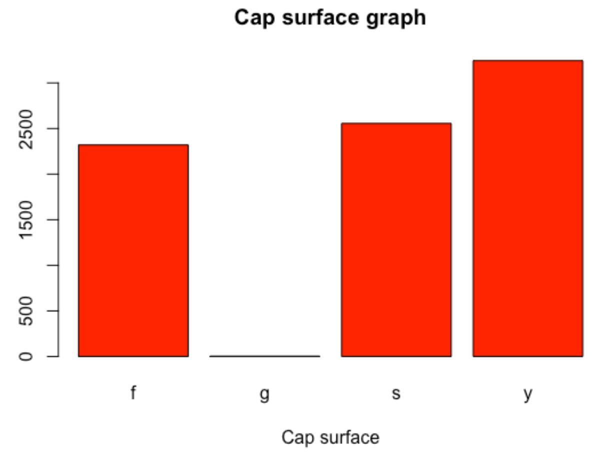
FEATURE	COUNT	% MISSING	CARD	MODE	MODE FREQ	MODE %	2ND MODE	2ND MODE FREQ	2ND MODE %
CLASS	8124	0%	2	e	4208	52%	p	3916	48%
CAP-SHAPE	8124	0%	6	x	3656	45%	f	3152	39%
CAP-SURFACE	8124	0%	4	y	3244	40%	s	2556	31%
CAP-COLOR	8124	0%	10	n	2284	28%	g	1840	23%
BRUISES	8124	0%	2	0	4748	58%	1	3376	42%
ODOR	8124	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%

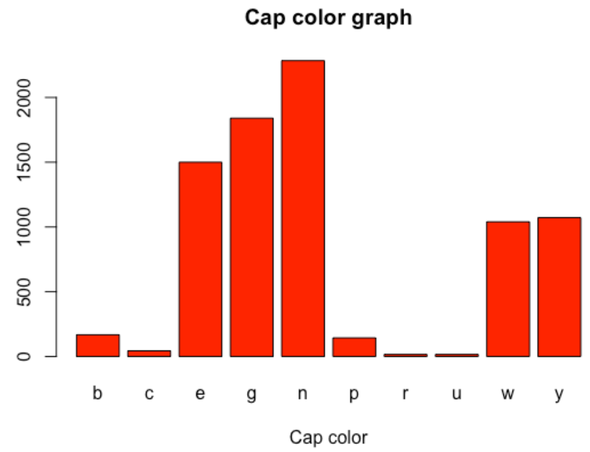
As we can see there aren't any missing values in the features, but some of them are not well distributed. An example is `veil-color` where the value corresponding to the first mode is the 98% of the total values.

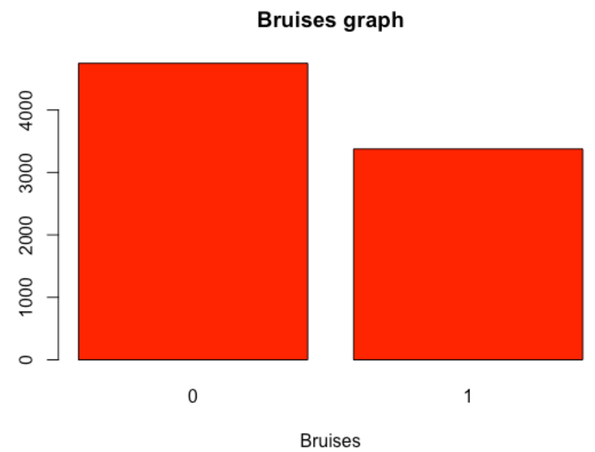
A graphical representation of the data is given below.

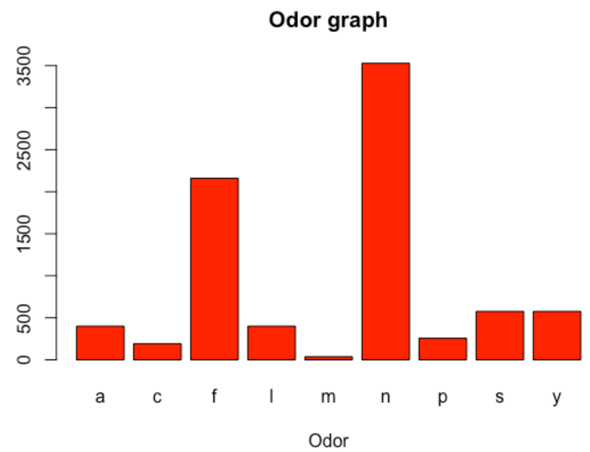


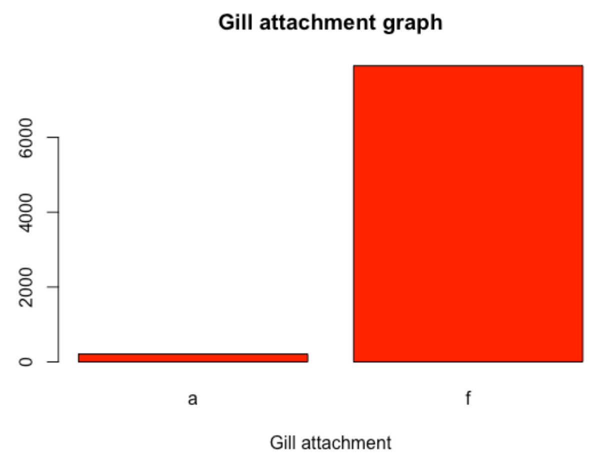


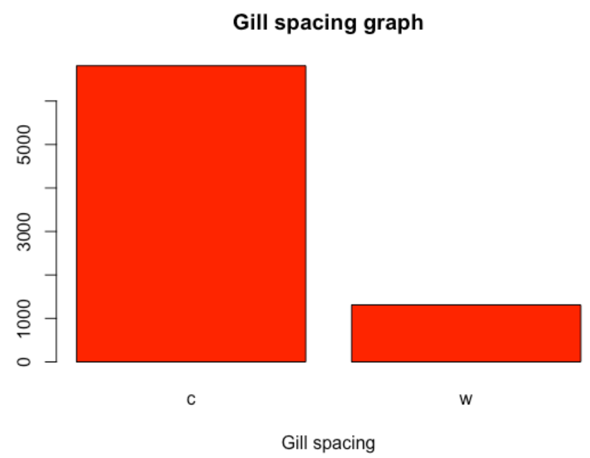


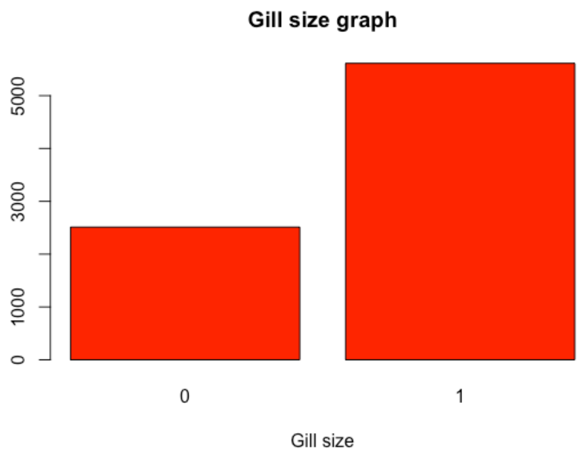


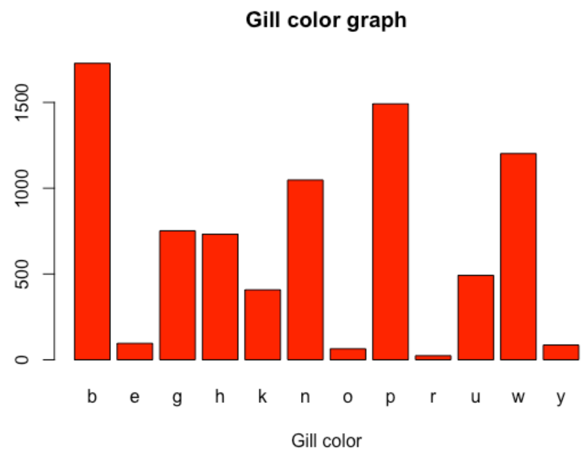


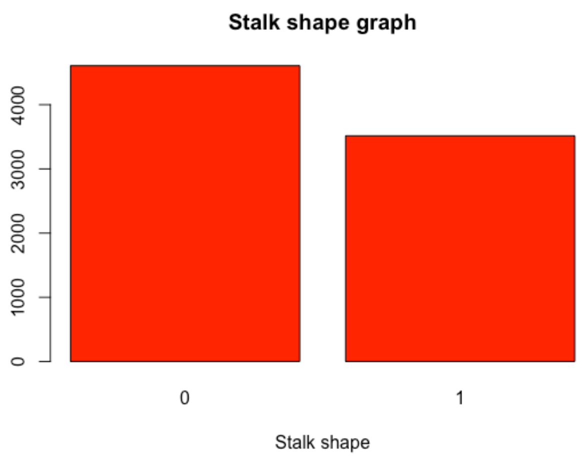


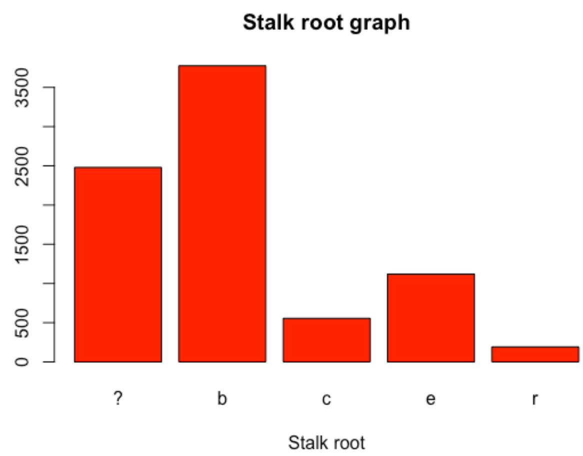


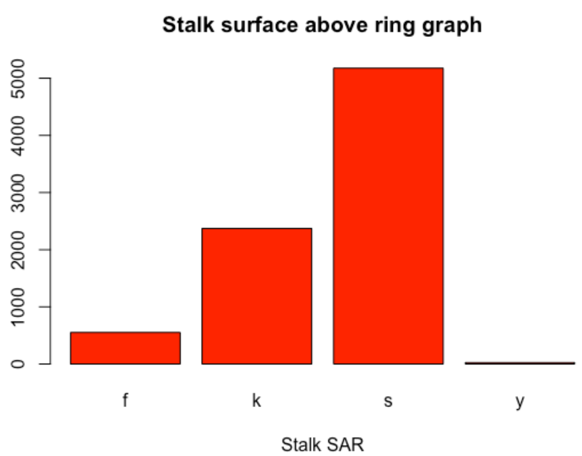


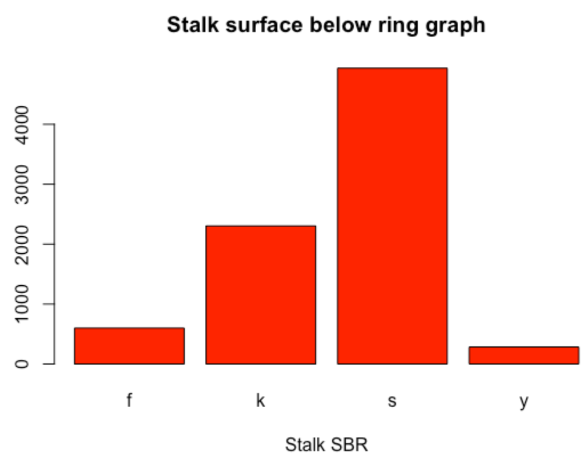

Gill size graph
0
1000
3000
5000
0
1
Gill size

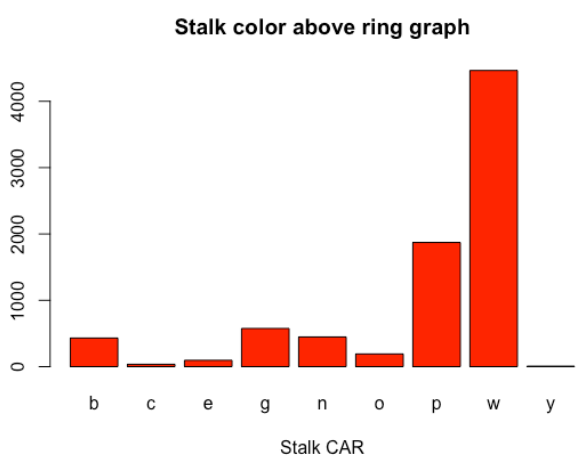

Gill color graph
0
500
1000
1500
b e g h k n o p r u w y
Gill color

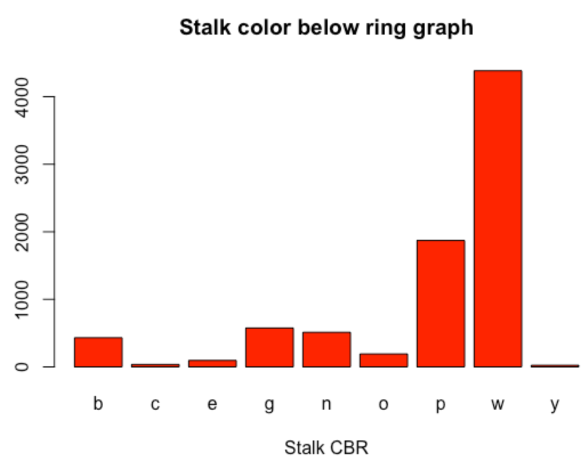

Stalk shape graph
0
1000
2000
3000
4000
0
1
Stalk shape

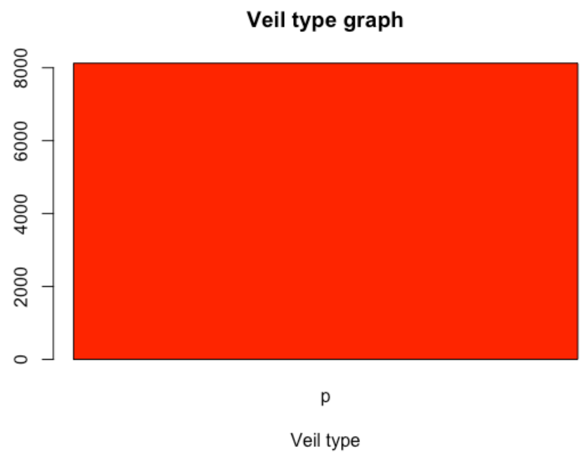

Stalk root graph
0
500
1500
2500
3500
? b c e r
Stalk root

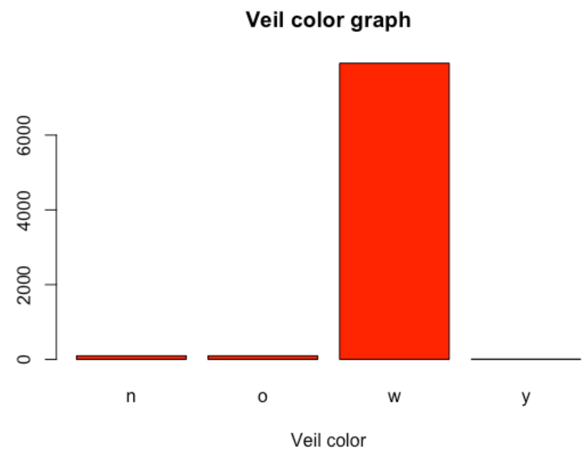

Stalk surface above ring graph
0
1000
2000
3000
4000
5000
f k s y
Stalk SAR

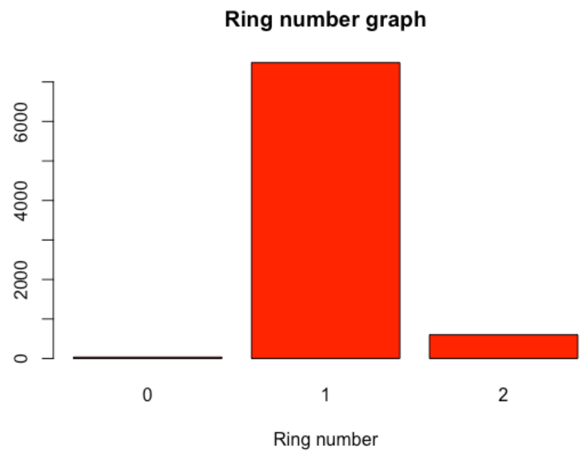

Stalk surface below ring graph
0
1000
2000
3000
4000
f k s y
Stalk SBR

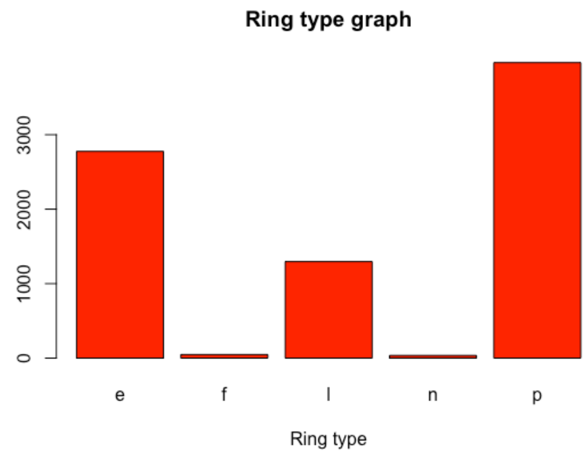

Stalk color above ring graph
0
1000
2000
3000
4000
b c e g n o p w y
Stalk CAR

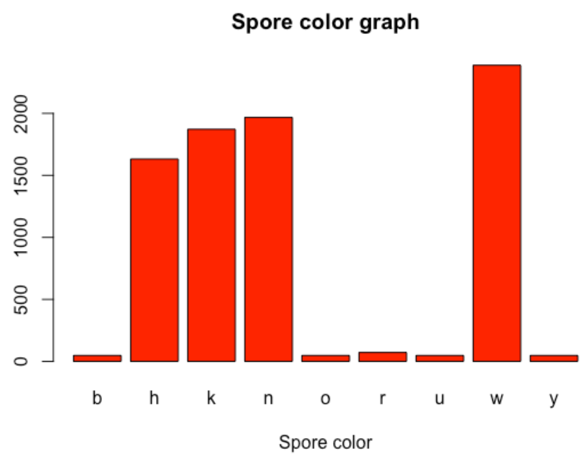

Stalk color below ring graph
0
1000
2000
3000
4000
b c e g n o p w y
Stalk CBR

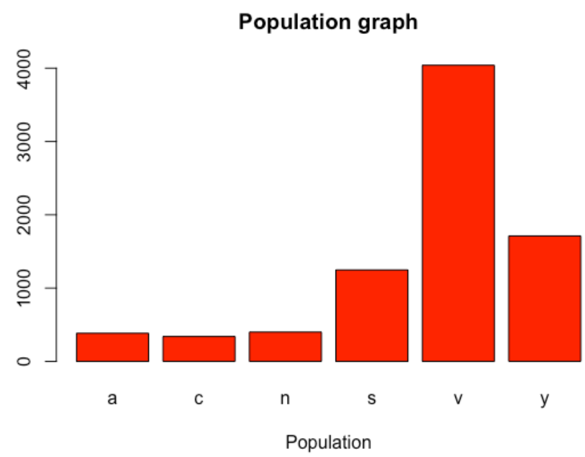

Veil type graph
0
2000
4000
6000
8000
p
Veil type

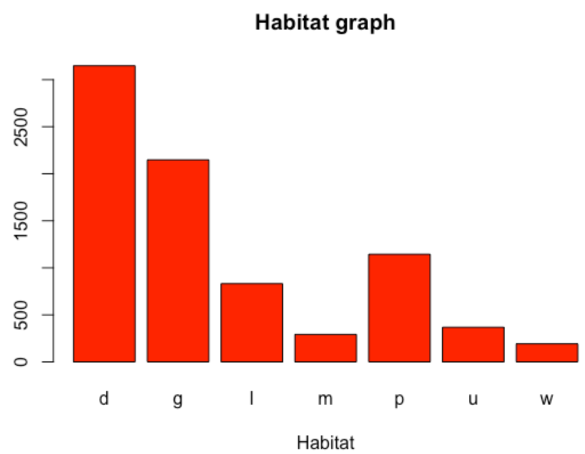

Veil color graph
0
2000
4000
6000
n
o
w
y
Veil color


Ring number graph
0
2000
4000
6000
0
1
2
Ring number


Ring type graph
0
1000
2000
3000
e
f
l
n
p
Ring type


Spore color graph
0
500
1000
1500
2000
b
h
k
n
o
r
u
w
y
Spore color


Population graph
0
1000
2000
3000
4000
a
c
n
s
v
y
Population


Habitat graph
0
500
1500
2500
d
g
l
m
p
u
w
Habitat

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.