Artificial Intelligence Protocol

Introduction:

For our project we wanted to create an Artificially trained model. This model could decide based on the provided dataset (dataset would regard mushrooms and their characteristics) whether the mushroom is edible or poisonous. We also wanted to provide some statistically proven data which could help with understanding the dataset and overall decision making.

Statistical analysis:

We provide statistics based on provided information.

Class:

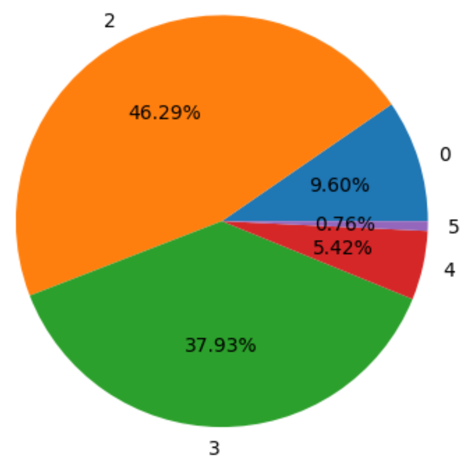
This class is a response column which decides whether a mushroom is edible or poisonous. In our dataset we have 4208 edible mushrooms and 3916 poisonous mushrooms. Which is relatively balanced data that can be further explored.

Cap-shape:

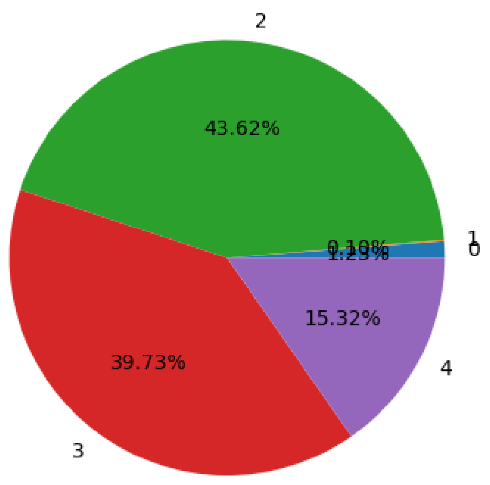
There are six types of cap shapes in total. However, most mushrooms, both edible and poisonous tend to have mostly convex(2) (3656 units) and flat(3) shapes (3152 units). Both of these characteristics are approximately equally divided between edible and poisonous mushrooms. This being said, this column will not tell us much about how to differentiate betweens edible and poisonous mushrooms. Mushrooms with bell(0) cap shape are more likely to be edible, while mushrooms with knobbed(4) cap-surface are more likely to be poisonous. Mushrooms with sunken(5) cap shape are definitely edible, there are 48 of them in this dataset. Conical(1) cap shaped mushrooms are definitely poisonous, but it is only 4 units.

Legend: 0-bell, 1-conical, 2-convex, 3-flat, 4-knobbed, 5-sunken.

Distribution of cap shapes between edible mushrooms



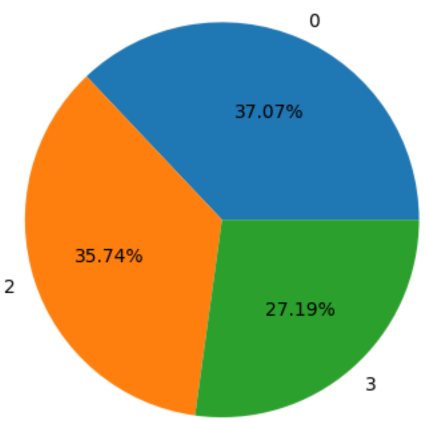
Distribution of cap shapes between poisonous mushrooms



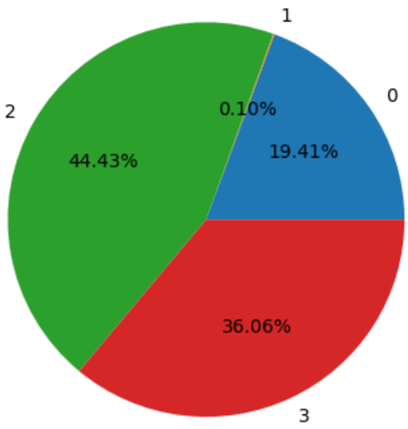
Cap-surface:

This column has four categories in total. We can surely say that if a mushroom's cap has grooves, it’s definitely poisonous. However, there are only four mushrooms like that in a dataset, so this information is not very valuable. Edible mushrooms tend to have more fibrous(0) cap surface. Poisonous mushrooms tend to have more scaly(2) and smooth(3) cap surface than edible ones. All in all, this attribute does not really help with solving our task.
Legend: 0-fibrous, 1-grooves, 2-scaly, 3-smooth

Distribution of cap surface in edible mushrooms



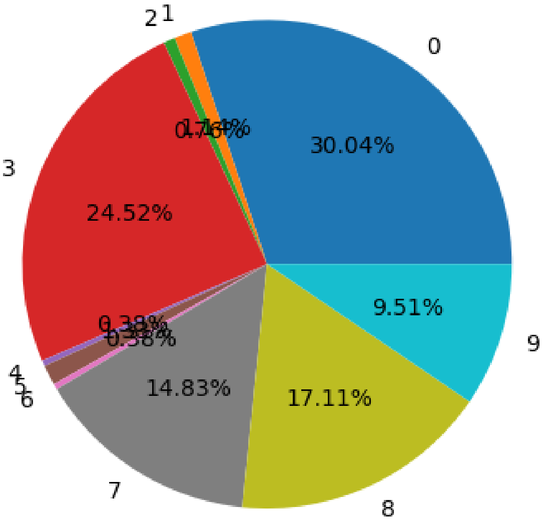
Distribution of cap surface in poisonous mushrooms



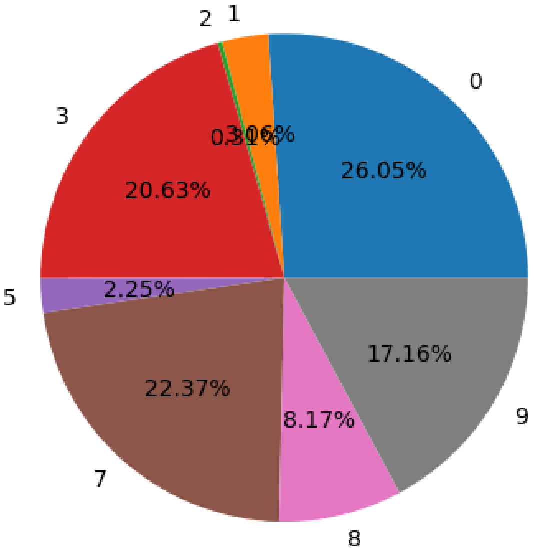
Cap-color:

Most mushrooms have brown and gray cap color. We can say, that mushrooms with green(4) and purple(6) cap color are edible, but it is only a small part (32 units) of the dataset. The rest of the cap colors are distributed evenly, without big differences between both classes.
Legend: 0-brown, 1-buff, 2-cinnamon, 3-gray, 4-green, 5-pink, 6-purple, 7-red, 8-white, 9-yellow

Distribution of cap color in edible mushrooms

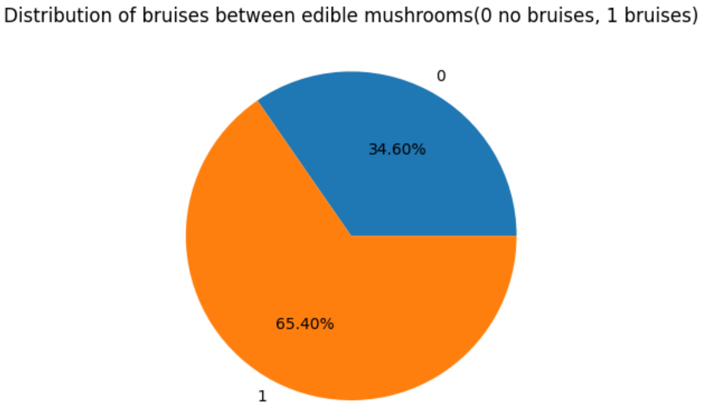


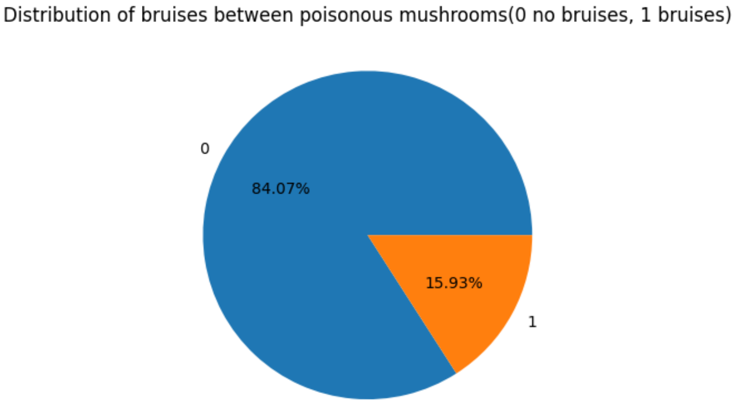
Distribution of cap color in poisonous mushrooms



Bruises:

What we can learn from this column is that poisonous mushrooms tend to have more bruises, the ratio between no bruises and bruises is about 85:15. The ratio between edible mushrooms is 35:65 for no bruises and bruises. So the ratio is almost inverted.

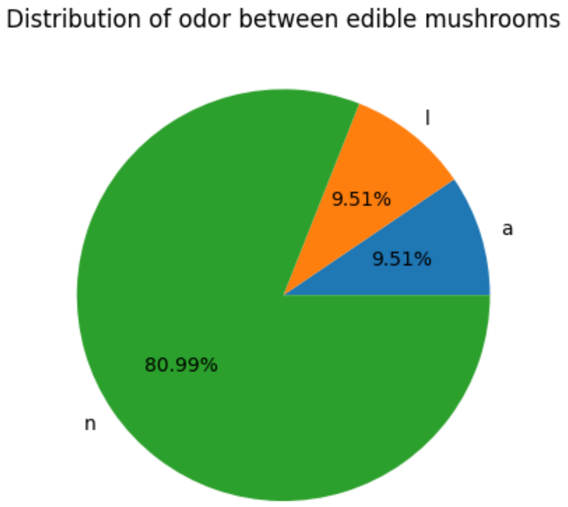


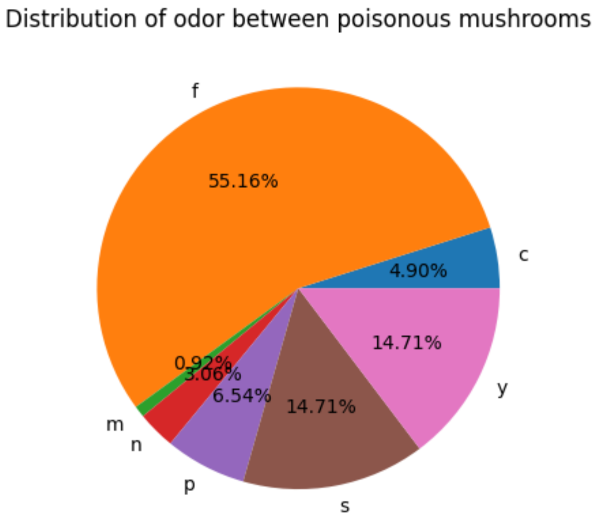


Odor:

This is probably the most important column. Because there is only one small part that both classes have in common, we could decide this problem just from knowledge of this column with relatively high precision. Mushrooms with almond(a) and anise(l) odor are definitely edible. If a mushroom has no odor, there is a small chance that it might be poisonous. Mushrooms with rest of the odors are poisonous.

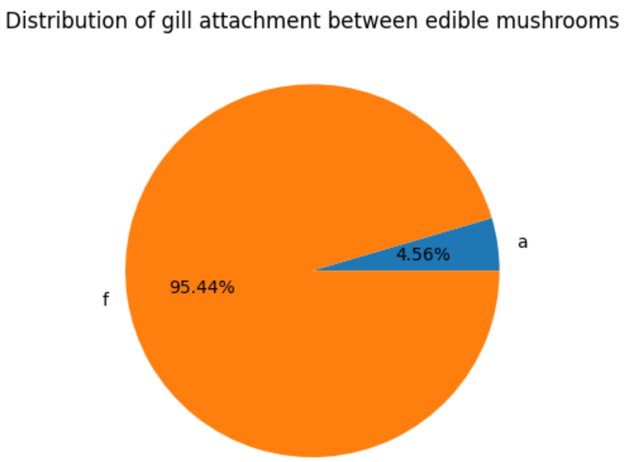
Legend: a-almond, l-anise, c-creosote, y-fishy, f-foul, m-musty, n-none, p-pungent, s-spicy.

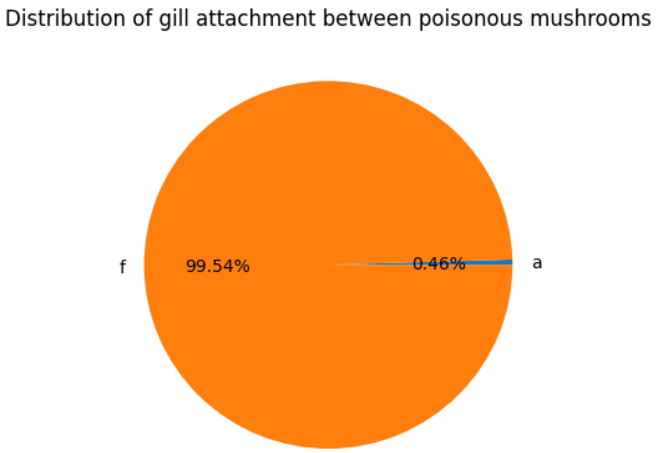




Gill-attachment:

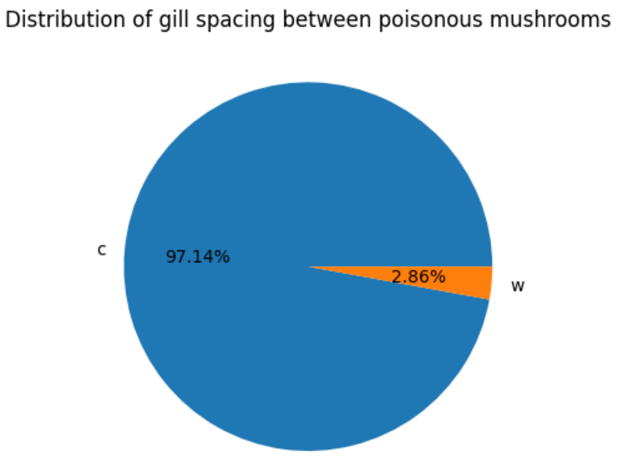
From this column we can not tell much, because the vast majority of mushrooms, both edible and poisonous, have a free type of gill attachment.
Legend: a-attached, d-descending, f-free, n-notched.

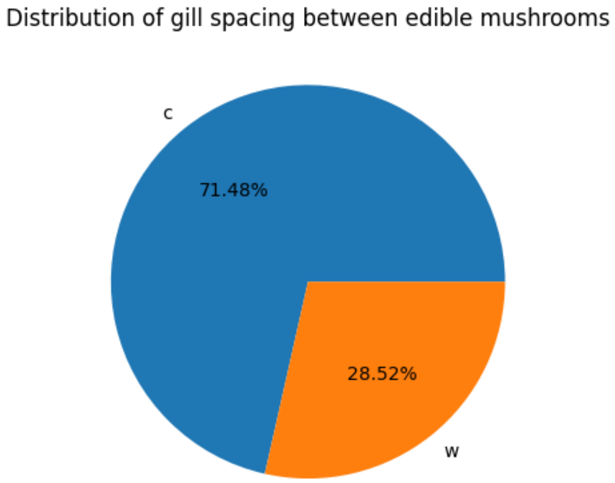

Distribution of gill attachment between edible mushrooms


Distribution of gill attachment between poisonous mushrooms

Gill-spacing:

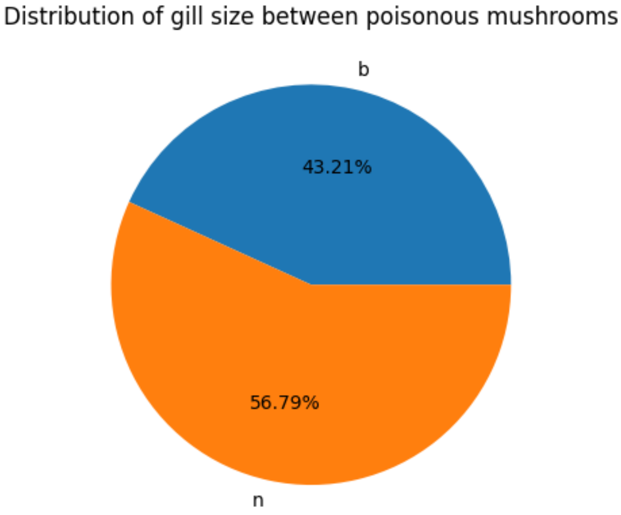
Only a small part of poisonous mushrooms has crowded(w) gill spacing, so a mushroom with this attribute will be more likely edible. Most mushrooms, both edible and poisonous, tend to have close gill spacing, so it will not help us much with the class differentiating.
Legend: c-close, w-crowded, d-distant.

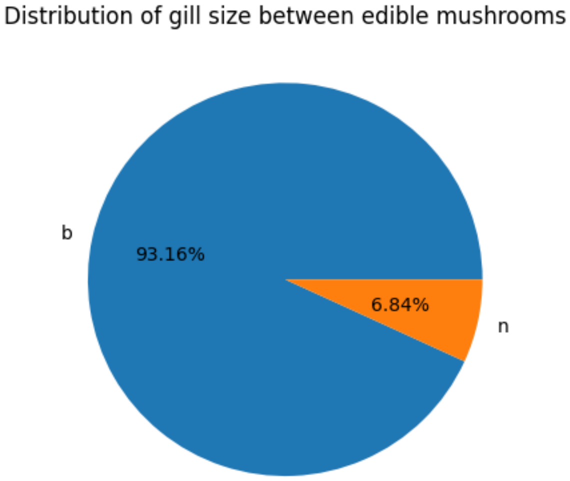

Distribution of gill spacing between poisonous mushrooms


Distribution of gill spacing between edible mushrooms

Gill-size:

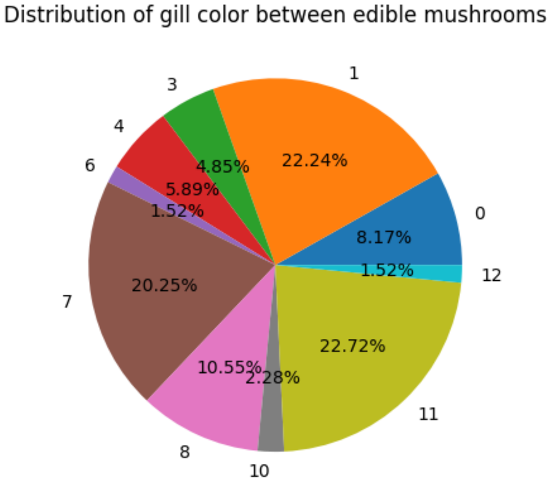
Majority of edible mushrooms have a broad gill size. Gill size between poisonous mushrooms is more evenly distributed. So if a mushroom has a narrow gil size, it is more likely going to be poisonous.
Legend: b-broad, n-narrow.

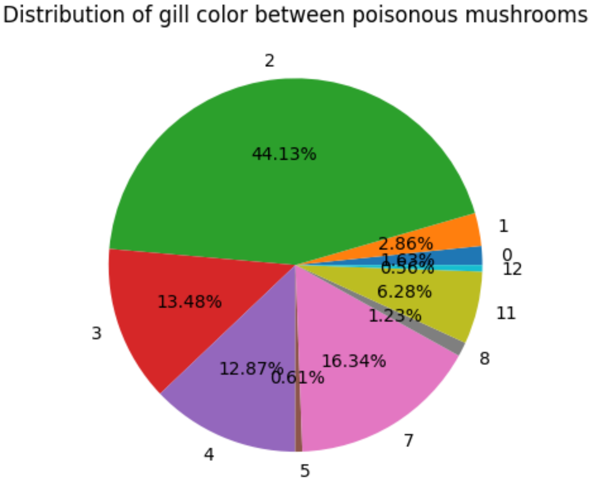

Distribution of gill size between poisonous mushrooms


Distribution of gill size between edible mushrooms

Gill-color:

We can say that a mushroom with buff(2) gill color is definitely poisonous. This is useful information, because this attribute takes part of 44% of poisonous mushrooms in the dataset. We can be sure about other attributes as well, but it is not as big part of the dataset. If a mushroom has orange(6) gill color, it is an edible mushroom, but this attribute has only 1.52% of edible mushrooms.
Legend: 0-black, 1-brown, 2-buff, 3-chocolate, 4-gray, 5-green, 6-orange, 7-pink, 8-purple, 9-red, 10-white, 11-yellow, 12-undefined.


Distribution of gill color between edible mushrooms


Distribution of gill color between poisonous mushrooms

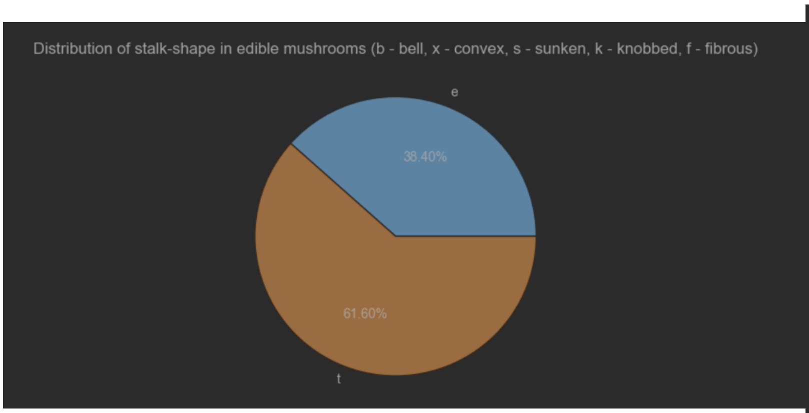
Stalk-shape:

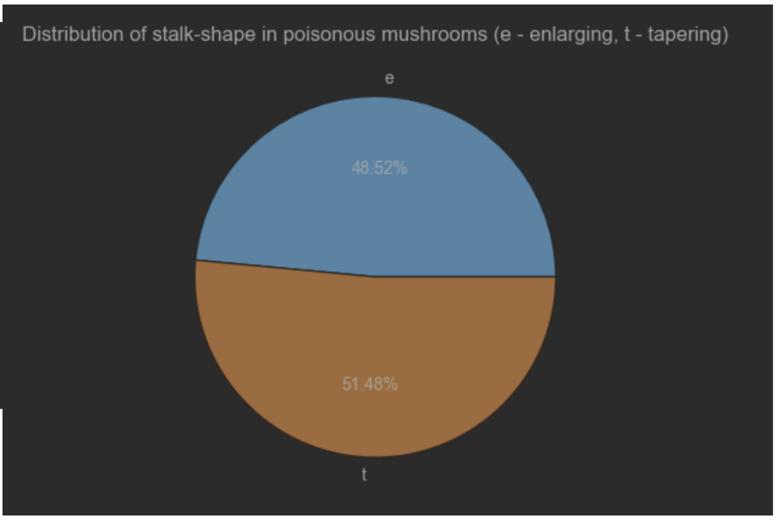
Stalk-shape can enlarging or tapering.

There is 3516 mushrooms with enlarging stalk shape from which 1616 is edible and 1900 poisonous.

There is 4608 mushrooms with tapering stalk shape from which 2592 is edible and 2016 poisonous.

So we can say that enlarging are more likely to be poisonous and tapering are more likely to be edible.





Stalk-root:

Can be bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?

The mushrooms with a bulbous is the most exactly 3776 from which 1920 is edible and 1856 that means that the probability is almost the same, only slightly higher that it will be edible.

The stalk root on which we can say with great probability whether a mushroom is edible or poisonous is club which has 556 mushrooms and only 44 of them are poisonous.

There are no mushrooms with cup and rhizomorphs Stalk-root in dataset.

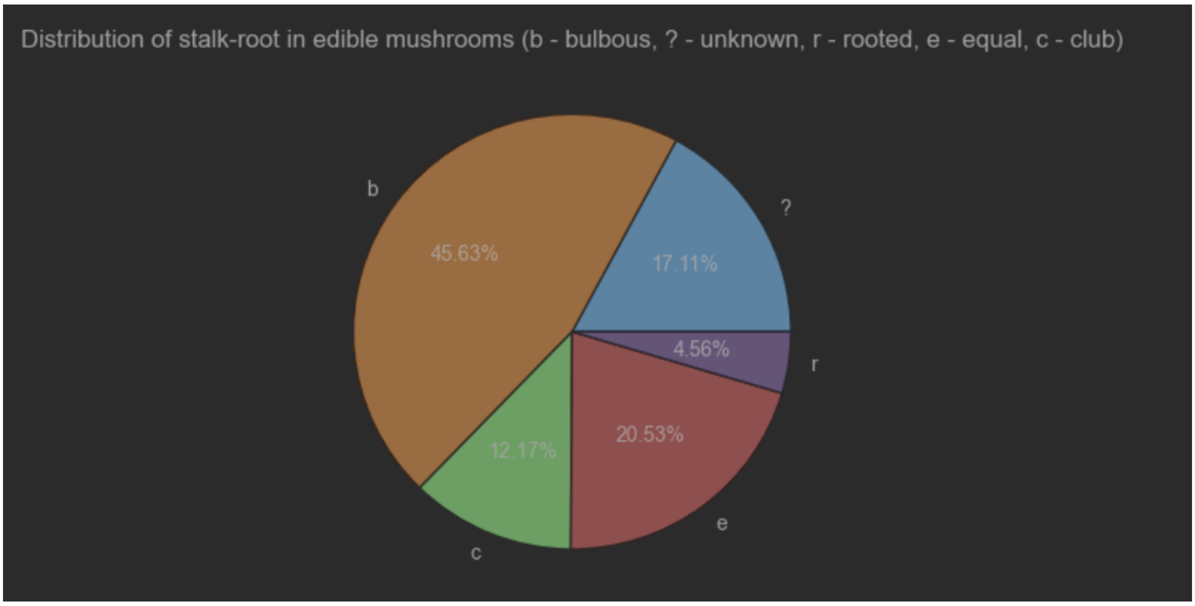
There are 1120 equal mushrooms in the dataset of which 256 are poisonous.

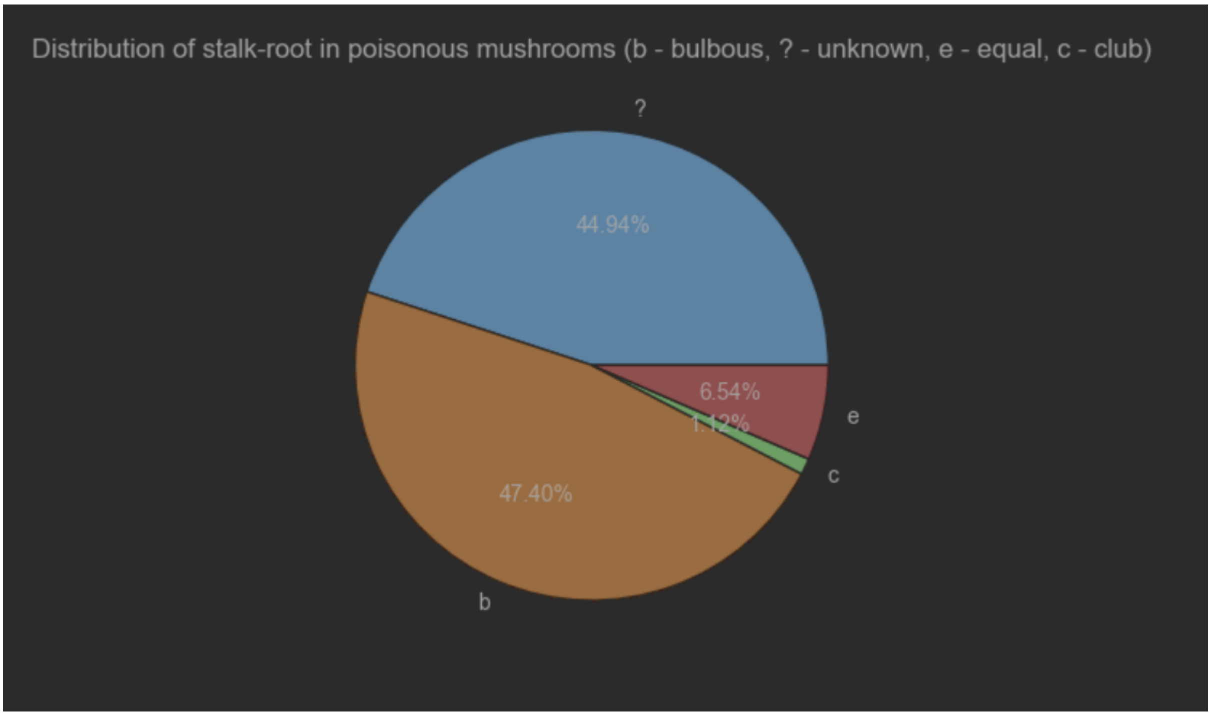
So we can say that equal are more likely to be edible.

The stalk root about which we can say one hundred percent if is poisonous or edible is rooted which has 192 mushrooms and all of them are edible.

There are 2480 mushrooms with missing stalk root in the dataset of which 720 are edible.

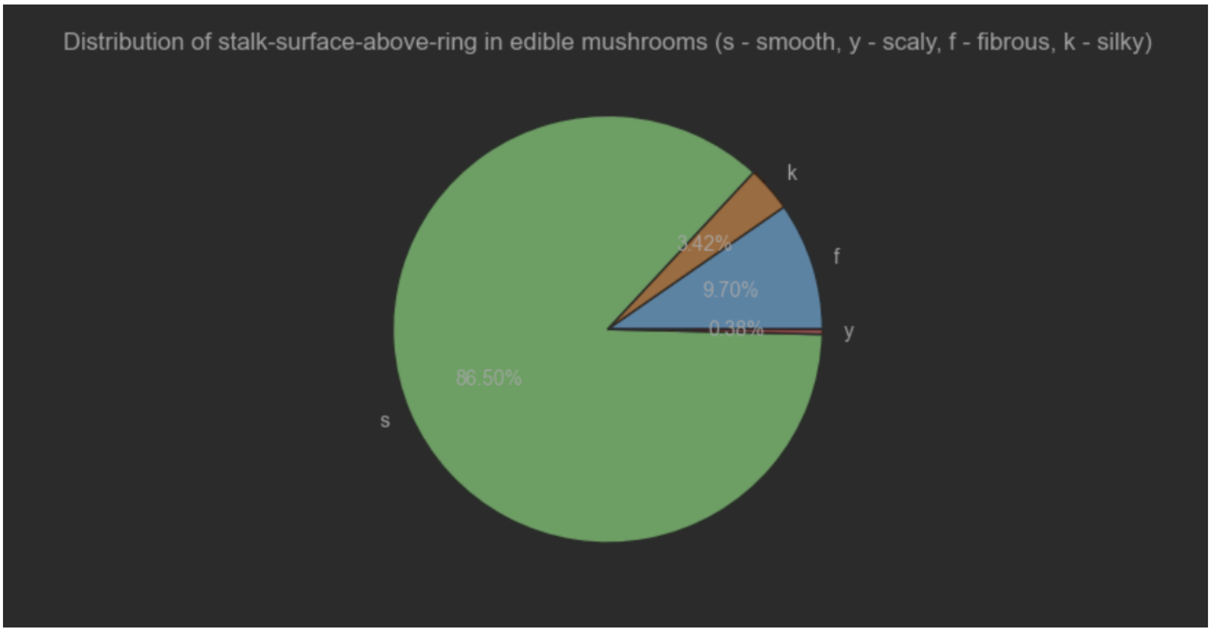
So we can say that mushrooms with missing stalk root are more likely to be poisonous.

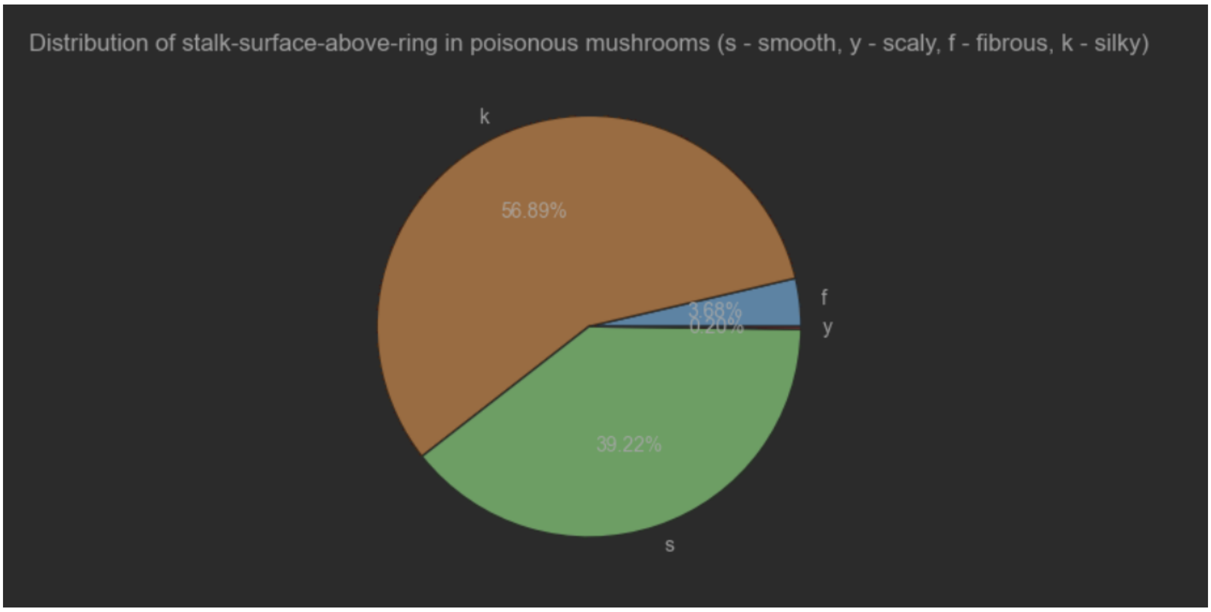




Stalk-surface-above-ring:

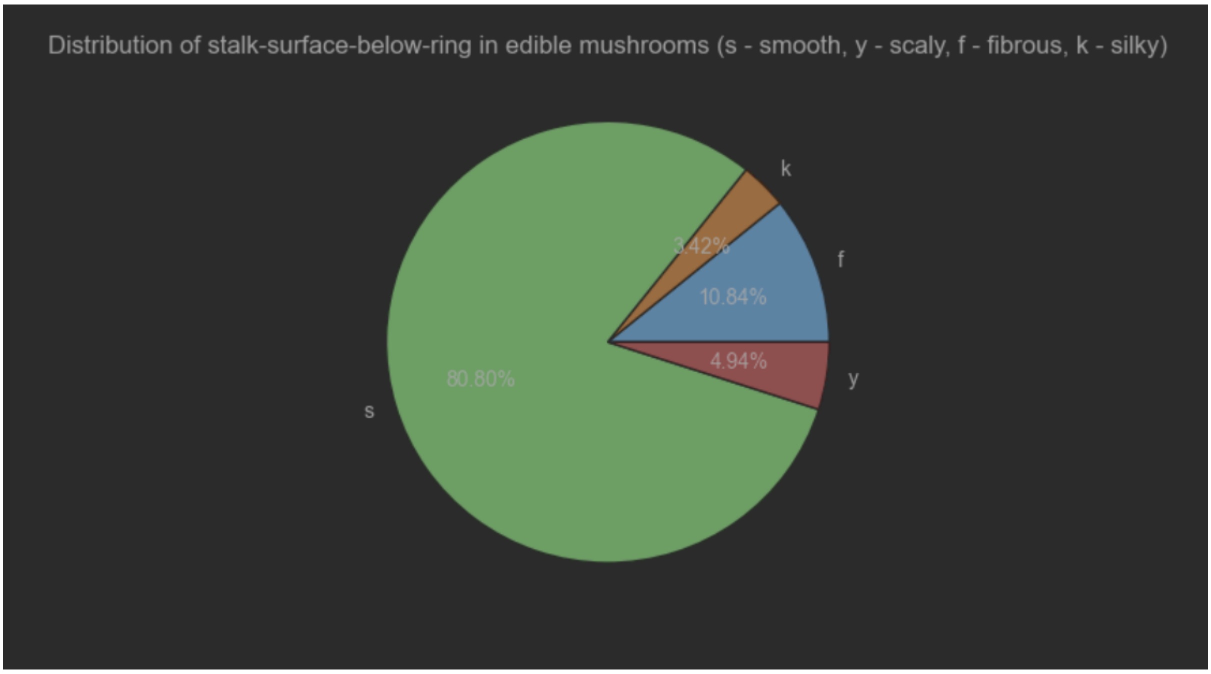
Can be fibrous=f, scaly=y, silky=k, smooth=s.
There are 552 fibrous mushrooms in the dataset of which 144 are poisonous.
There are only 24 scaly mushrooms in the dataset of which 8 are poisonous.
So we can say that fibrous and scaly are more likely to be edible.
The surface on which we can say with great probability whether a mushroom is edible or poisonous is silky which has 2372 mushrooms and only 144 of them are edible.
So we can say that silky are most likely to be poisonous.
Mushrooms most often have the smooth surface in total 5176 of which there are 1536 poisonous. So we can say that smooth are more likely to be edible.

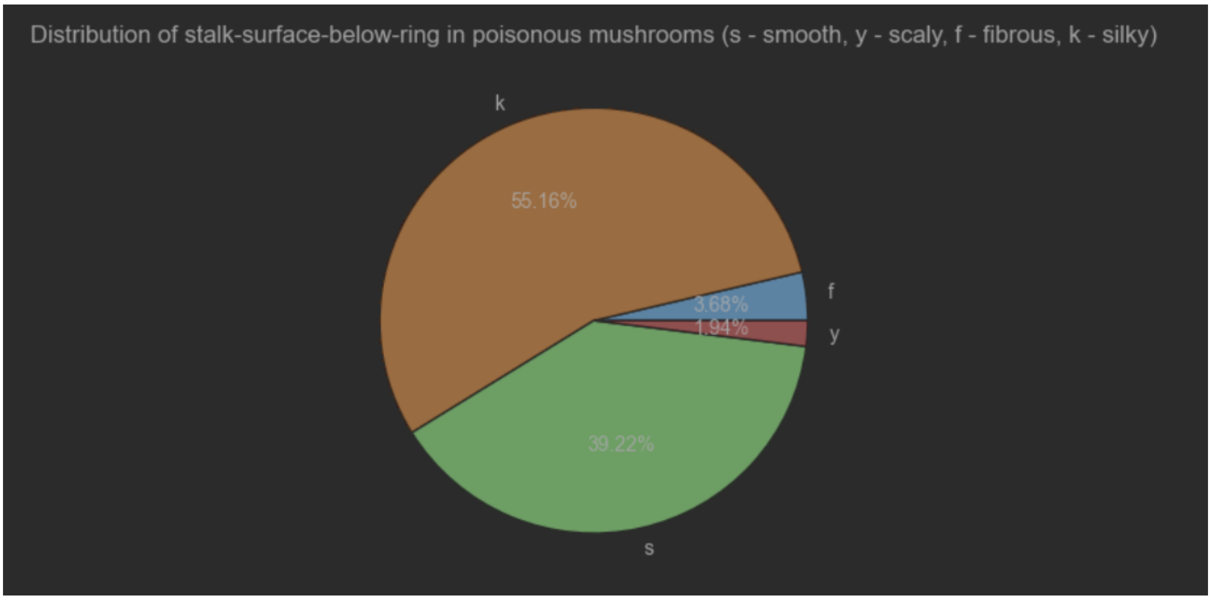




Stalk-surface-below-ring:

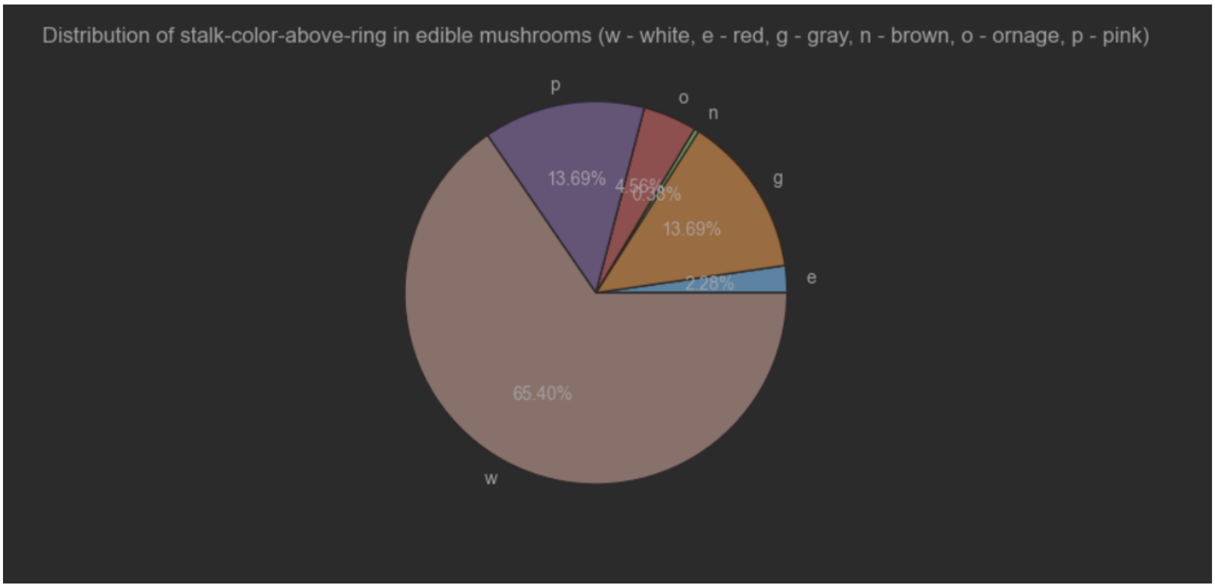
Can be fibrous=f, scaly=y, silky=k, smooth=s.
There are 600 fibrous mushrooms in the dataset of which 146 are poisonous.
There are 284 scaly mushrooms in the dataset of which 76 are poisonous.
So we can say that fibrous and scaly are more likely to be edible.
The surface on which we can say with great probability whether a mushroom is edible or poisonous is silky which has 2304 mushrooms and only 144 of them are edible.
So we can say that silky are most likely to be poisonous.
Mushrooms most often have the smooth surface in total 4936 of which there are 1536 poisonous. So we can say that smooth are more likely to be edible.

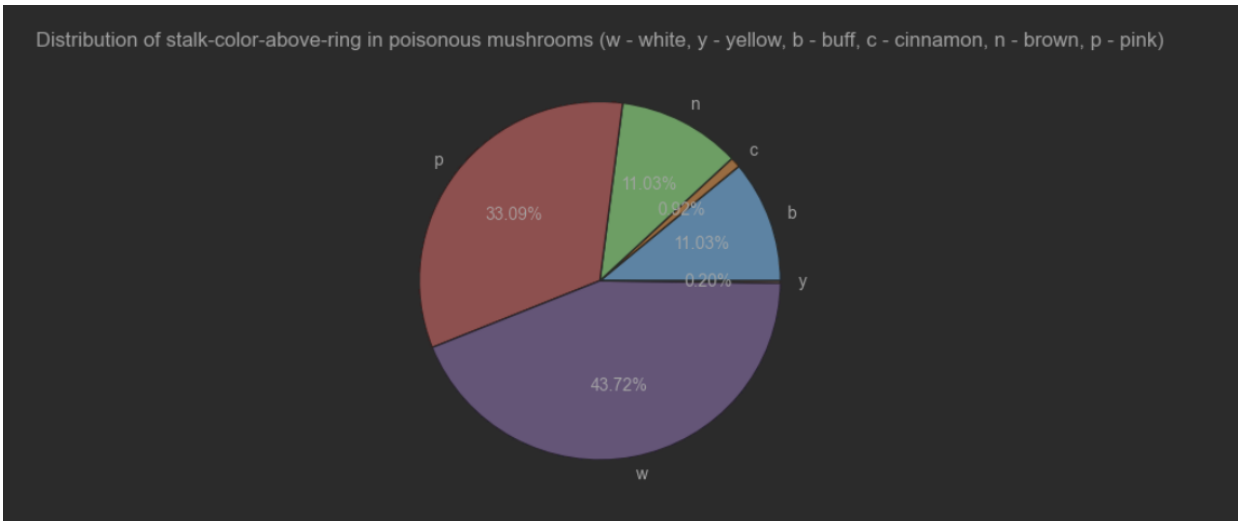




Stalk-color-above-ring:

Can be brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
In dataset i found 576 gray, 192 orange, 96 red and 8 yellow mushrooms and i found out that gray, orange, red and yellow are always edible.
In dataset i found only 432 buff and 36 cinnamon mushrooms and i found out that all of these mushrooms with this color are only poisonous.
There are a total of 1,340 of these mushrooms in the dataset for which we can 100% determine their class.
There are a total of 448 brown mushrooms in the dataset and only 16 of them are edible
But the most mushrooms are white, a total of 4464 mushrooms, of which only 2752 are edible. So we cannot decide with high probability whether the mushroom is edible or not.
There are no mushrooms with pink stalk-color-above-ring in dataset.





Stalk-color-below-ring:

Can be brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y

In dataset i found 432 buff, 36 cinnamon and 24 yellow mushrooms and i found out that buff, cinnamon and yellow are always poisonous.

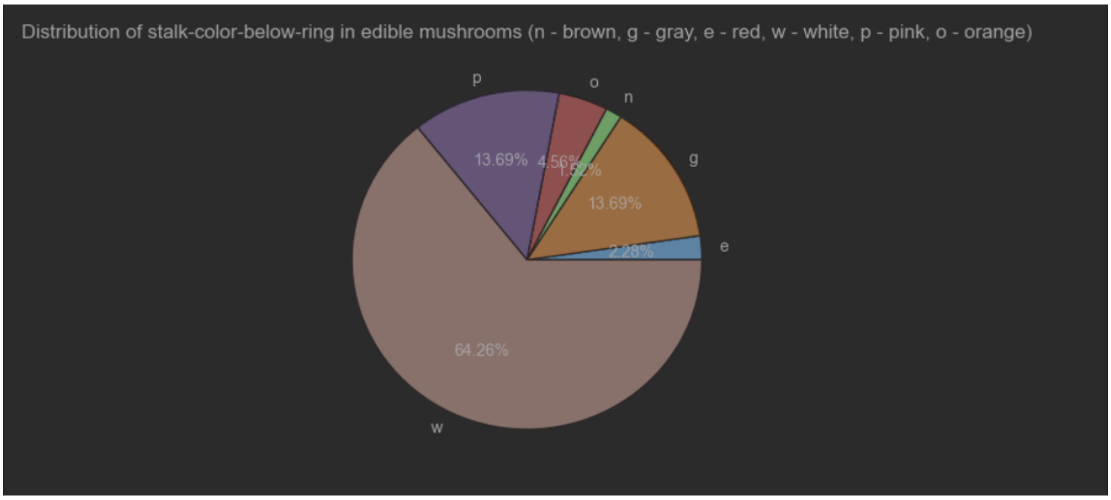
In dataset i found 576 gray, 192 orange and 96 red mushrooms and i found out that all of these mushrooms with this color are only edible.

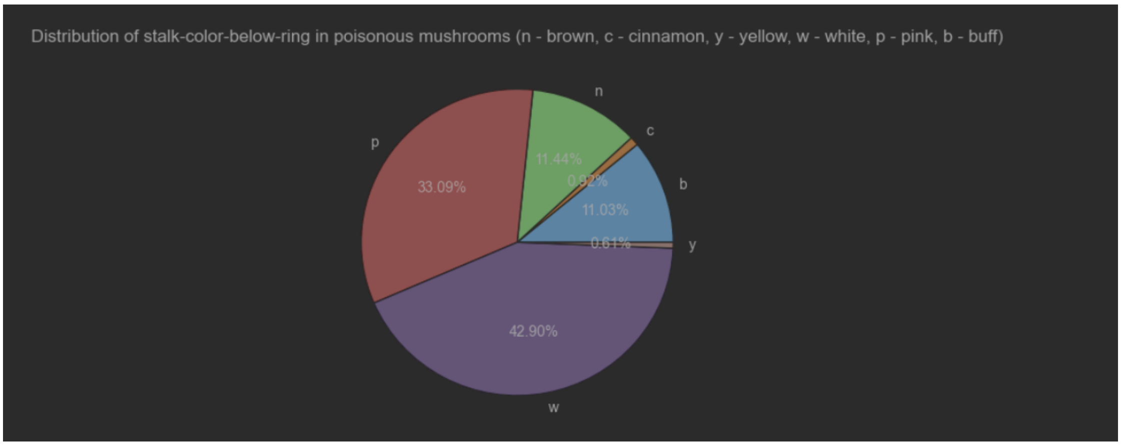
There are a total of 1,356 of these mushrooms in the dataset for which we can 100% determine their class.

This time, the dataset also has a pink color and a total of 1872 mushrooms have this color, of which only 576 mushrooms are edible So we cannot decide with high probability whether the mushroom is edible or not.

There are a total of 512 brown mushrooms in the dataset and only 64 of them are edible. Thanks to this, we can assume that the mushroom will be poisonous, but not with great probability.

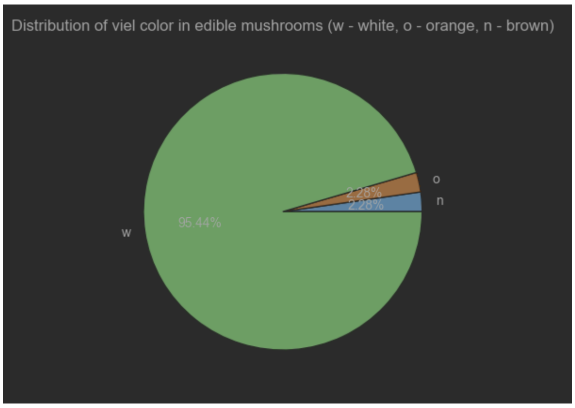
But the most mushrooms are again white, a total of 4384 mushrooms, of which only 2704 are edible. So we cannot decide with high probability whether the mushroom is edible or not.

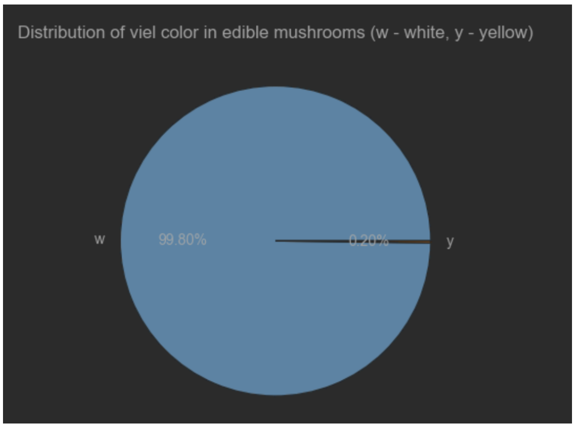




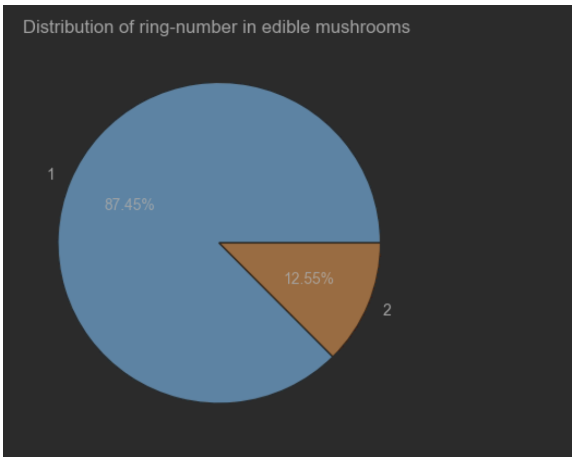
Veil-type: This characteristic of the dataset is constant for the whole dataset. For either edible or poisonous mushrooms veil type is partial. Therefore this column is not useful in decision-making and can be deleted.

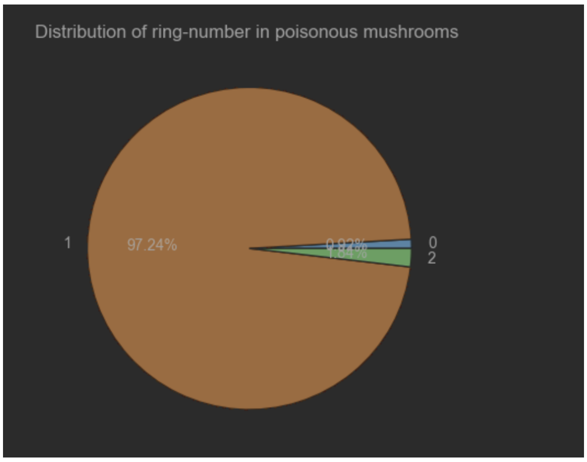
Veil-color: Veil color can help in distinguishing some mushrooms. In total there were 4 veil types - brown, orange, white, yellow. All mushrooms with either brown or orange veil type are edible. However in the whole dataset there are only 192 mushrooms with these 2 colors of veil. Also there are only 8 yellow veil mushrooms which are all poisonous. Other than that there are 7924 mushrooms with white veil - 4016 edibles and 3908 poisonous. Overall this characteristic can help in distinguishing between some mushrooms but it doesn't help in huge dataset of mushrooms.



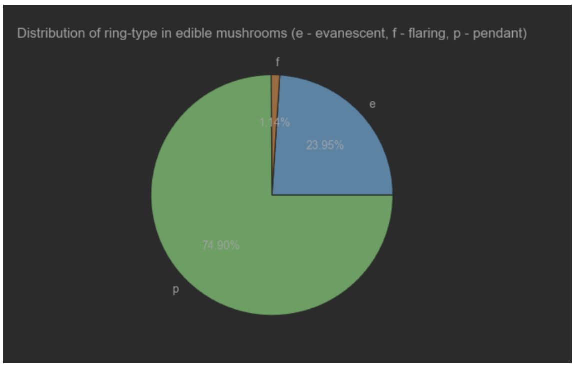


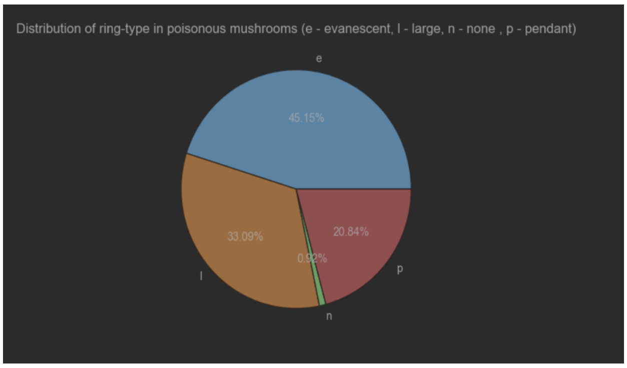
Ring-number: Regarding Ring-number characteristics and decision making. All mushrooms have either one, two or none ring numbers. All mushrooms without ring can be classified as poisonous. Furthermore most of mushrooms with two rings can be classified as edible (there are 528 edible and 72 poisonous mushrooms with 2 rings). Majority of mushrooms have one ring, however these datasets are balanced in both cases. There are 3808 poisonous and 3680 edible mushrooms with one ring. These data characteristics are similar to Veil-color. Both of these characteristics can distinguish some mushrooms and their similarities however they can not most of the dataset correctly.





Ring-type: Ring type is characteristic of a dataset with many different attributes. In total there were 5 different types of attributes - evanescent, flaring, pendant, large and none. Large ring types are helpful. These data show data all mushrooms with large rings are poisonous. In total there are 1296 mushrooms with large ring type (poisonous mushrooms). Regarding other attributes - flaring ring type is certain to be edible however there are only 72 of these mushrooms. Also all 36 mushrooms without a ring are poisonous. Regarding evanescent ring type - there are 1008 edible and 1768 poisonous mushrooms with this type of ring. We can not decide accurately based on this information whether mushroom is poisonous or edible. Last type of ring type that we encountered in our dataset is pendant - there were 3152 poisonous and 816 edible mushrooms. These data can incline us towards classifying mushroom as poisonous rather than edible. Overall these ring type is rather helpful attribute which can help us in decision making.



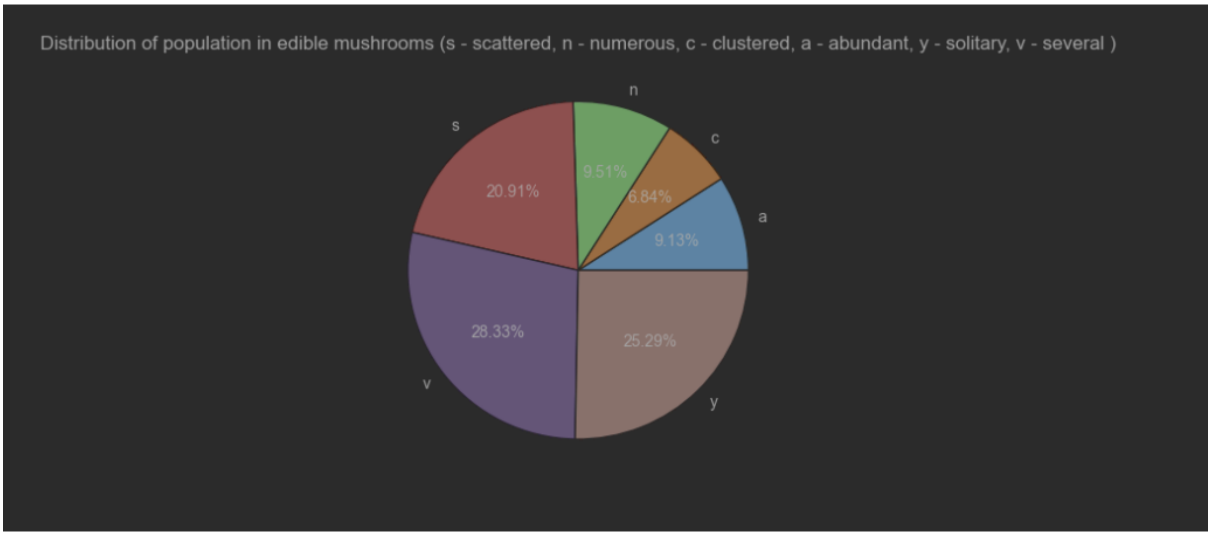


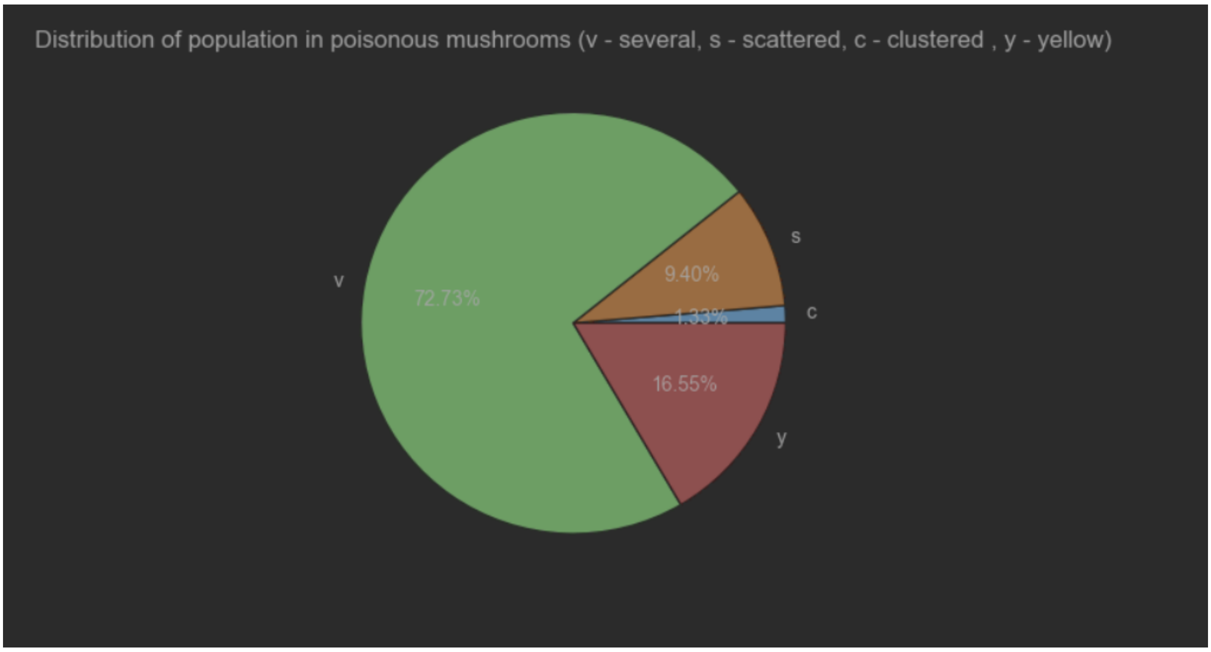
Spore-print-color:

If we ignore the odor, this is the most important column.

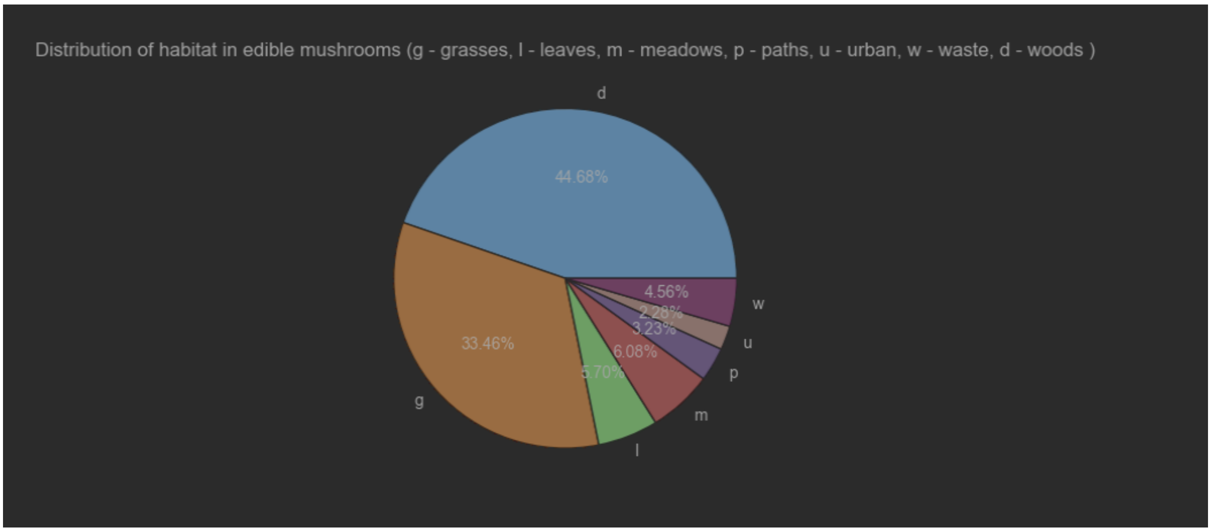
There are 9 different colors here. If the color is black or brown it will be 80 percent edible. If color is buff, orange, yellow or purple its edible on 100 percent but there is only 192 of those. If color is chocolate its edible only on 3 percent. If color is green its poisonous 100 percent. If color is white its edible on 30 percent.

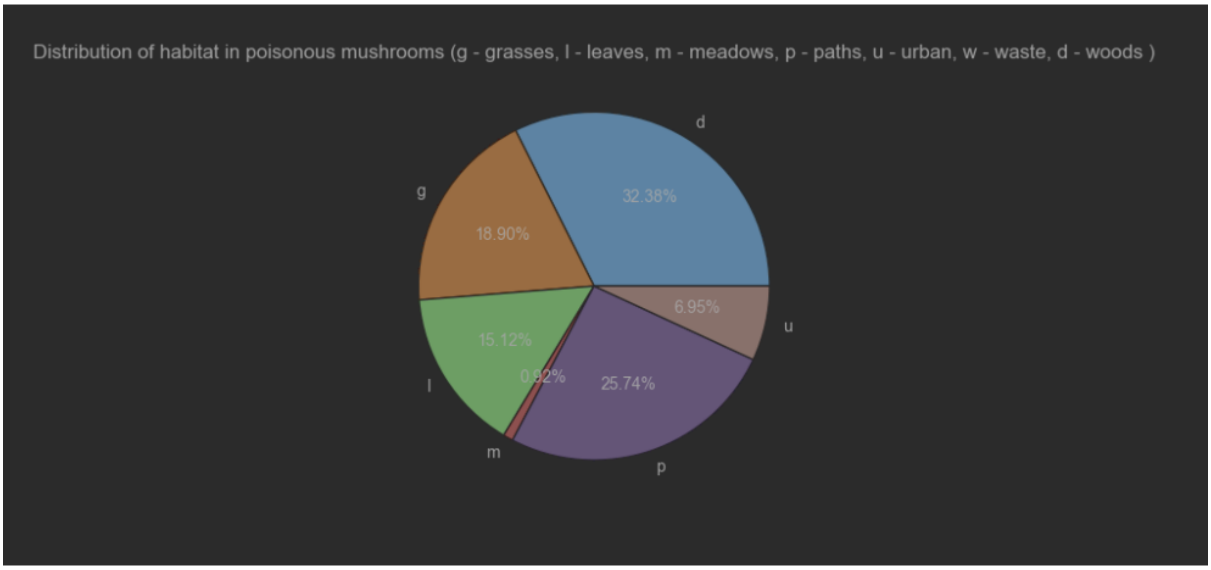
Population: This attribute contains many attributes which can lead to deeper understanding of dataset. Most important of all are mushrooms which live in abundant or numerous populations. All of these mushrooms are edible. Other than that our dataset provided us with clustered and sheathing populations. These can decide with accuracy of approximately 25% whether the mushroom provided is edible or poisonous. (288 edible to 52 poisonous mushrooms and 880 edible to 368 poisonous mushrooms). Other than that we were provided with a solitary mushrooms population and several mushrooms population. Both of these attributes have approximately 50% edible and poisonous population which does not help us in further decision making (Out of all solitary mushrooms there are 1064 edible to 648 poisonous mushrooms and out of all several mushrooms populations there are 1192 edible and 2848 poisonous).


Distribution of population in edible mushrooms (s - scattered, n - numerous, c - clustered, a - abundant, y - solitary, v - several)


Distribution of population in poisonous mushrooms (v - several, s - scattered, c - clustered , y - yellow)

Habitat: All mushrooms characteristic provided in these section are approximately divided into same sized groups (the only exception being paths mushrooms which have ratio 1008 poisonous to 136 and woods 192 woods mushrooms which are all edible). Rations that we were provided with are 1880 edible woods to 1268 poisonous woods mushrooms, 1408 edible grasses to 740 poisonous grasses mushrooms, 240 edible leaves mushrooms compared to 592 poisonous leaves mushrooms, 256 edible meadows mushrooms compared to 36 edible meadows mushrooms, 96 edible urban mushrooms compared to 272 poisonous urban mushrooms.





Observations:

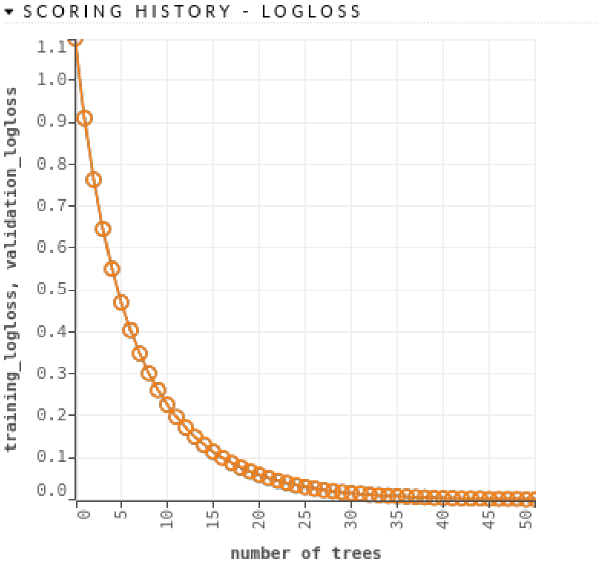
The most important is column C6.n, which refers to the odor = none and second C12.c which refers to stalk-root = club. If we ignore this column, then the most important column becomes C13.k which refers to stalk-surface-above-ring = silky and second c9.n which refers to gill-size = narrow.

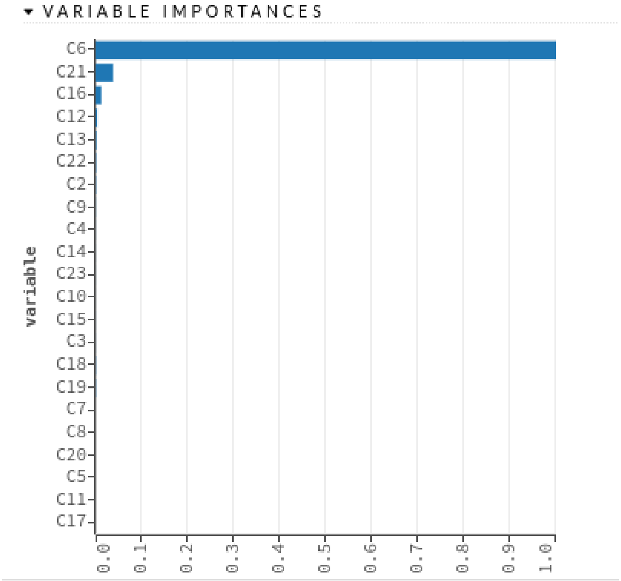
Success rate was 100%.

Training model:

We chose h2o machine learning library for solving this problem because of it's user friendly environment. This library offers a variety of algorithms from which we sorted those suitable for our task. We were looking for supervised classification algorithms. After studying these algorithms, we trained models on our mushrooms.csv dataset with the same parameters. We compared metrics of the models like MSE(Mean Square Error), RMSE (Root Mean Square Error) and r^2 and from this comparison came two winners: Gradient Boosting Machine(GBM) and XGBoost.

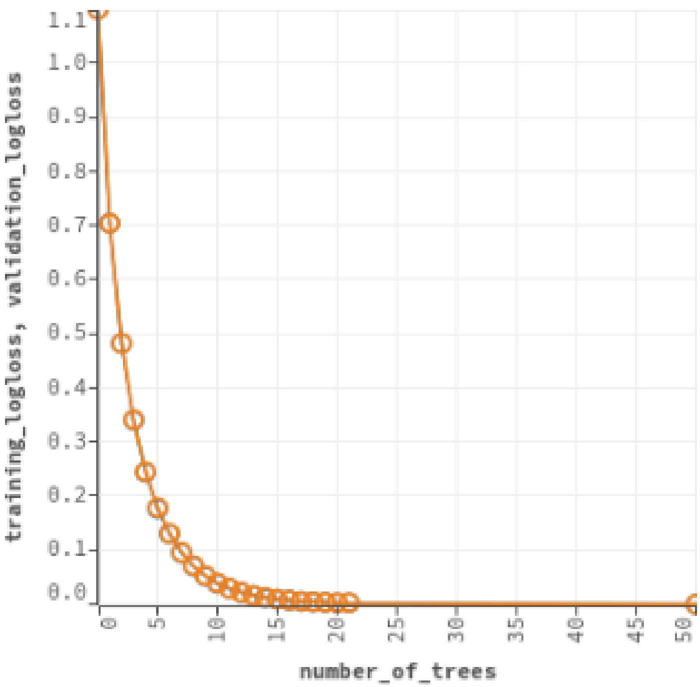
GBM:



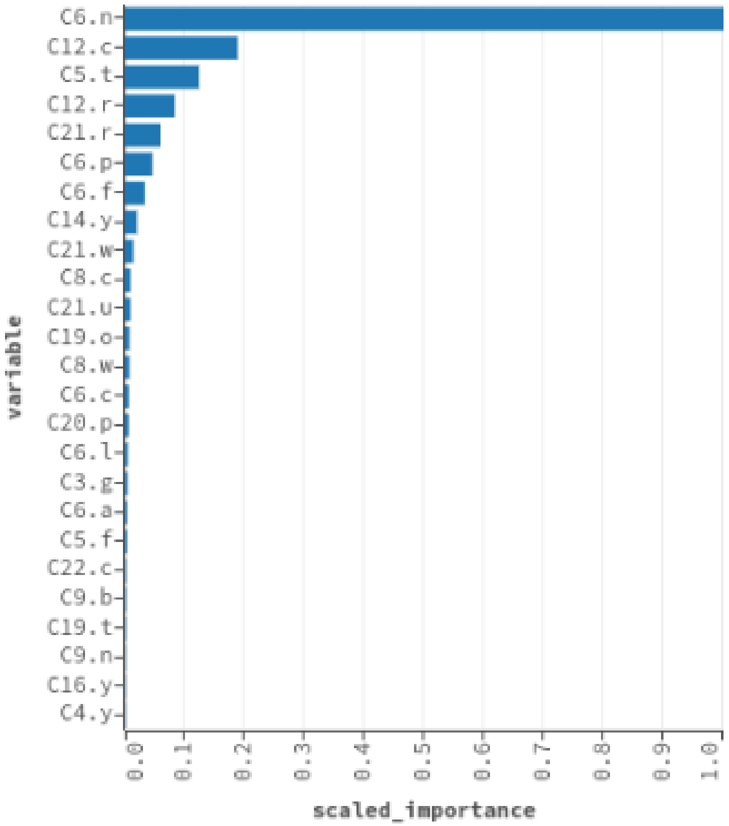


XGBoost

▾ SCORING HISTORY - LOGLOSS



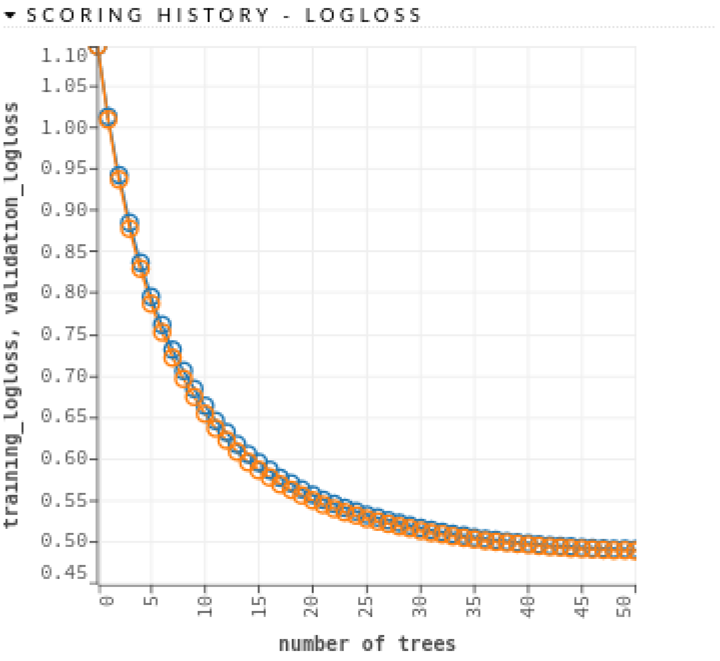
▾ VARIABLE IMPORTANCES

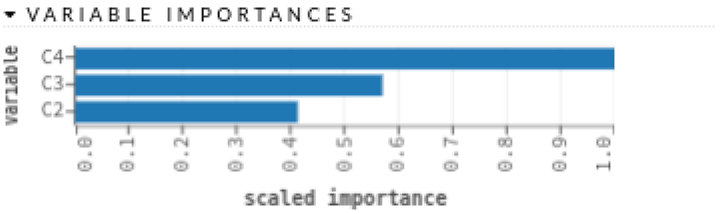


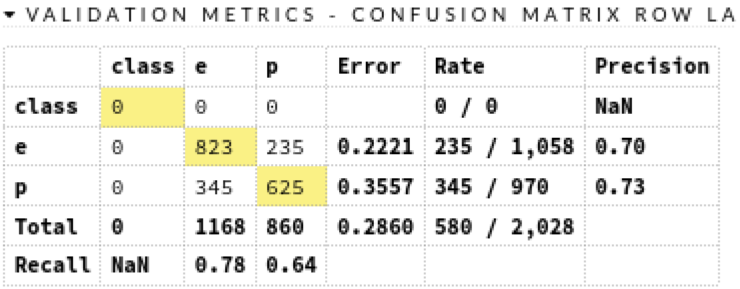
Experiments:

Training a model with only cap related columns

We wanted to know if training a model with only cap related columns will be sufficient. It turned out that it was not a good idea, because the precision of the model was only about 70%. This is not a good result, compared with other models we trained.





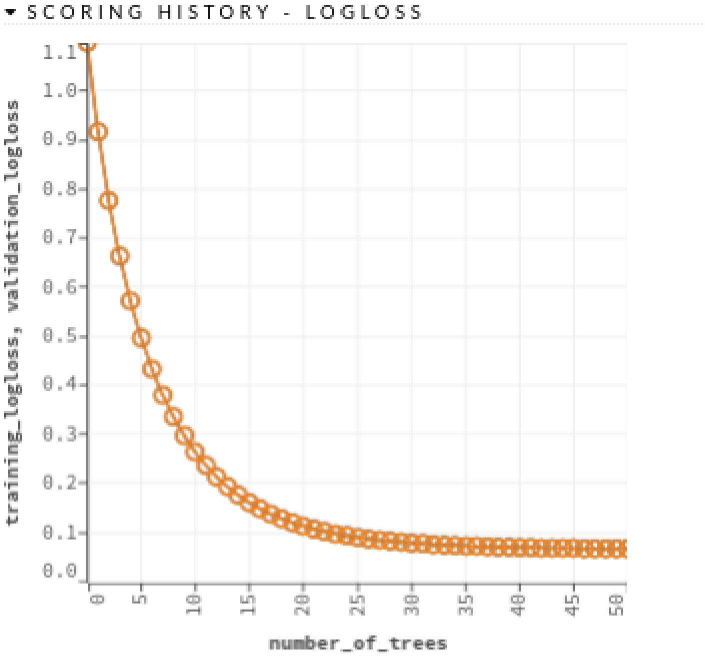

VALIDATION METRICS - CONFUSION MATRIX ROW LA

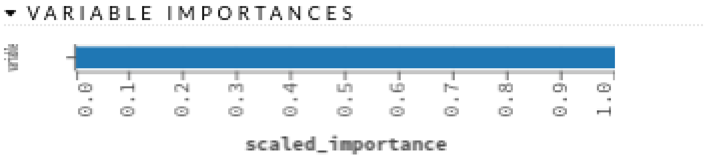
	class	e	p	Error	Rate	Precision
class	0	0	0		0 / 0	NaN
e	0	823	235	0.2221	235 / 1,058	0.70
p	0	345	625	0.3557	345 / 970	0.73
Total	0	1168	860	0.2860	580 / 2,028	
Recall	NaN	0.78	0.64			

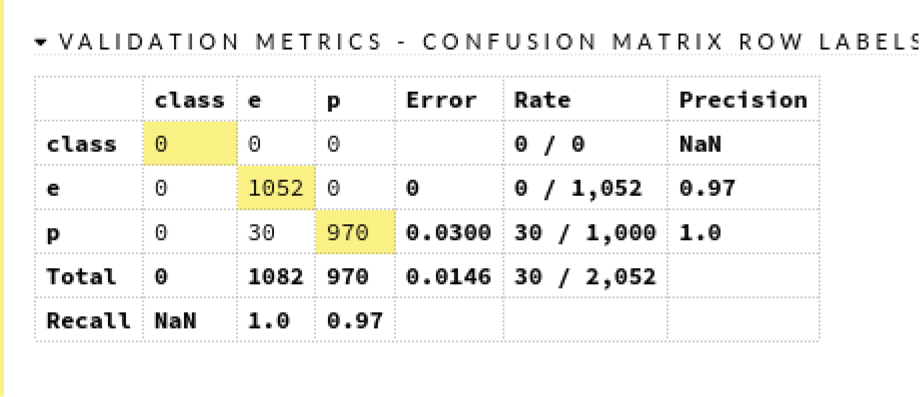
We can see that the most important column was C4 that stands for cap color. The precision when it comes to edible mushrooms is 0.70 and 0.73 when it is a poisonous mushroom. We can say that roughly every fourth prediction by this model is wrong. In conclusion, we can not surely solve this problem only by looking at the cap of a mushroom, we need to consider other attributes.

Training a model with only odor

We can determine with 97% precision if a mushroom is edible or not considering odor only. If we have a closer look at this confusion matrix, we can see the model prediction was wrong in only 30 cases. Unfortunately the model predicted that these mushrooms were edible, when in fact they were poisonous. Having said that, this model is not good for a practical usage. It would be better, if the model predicted that the mushrooms were poisonous and they were in fact edible.





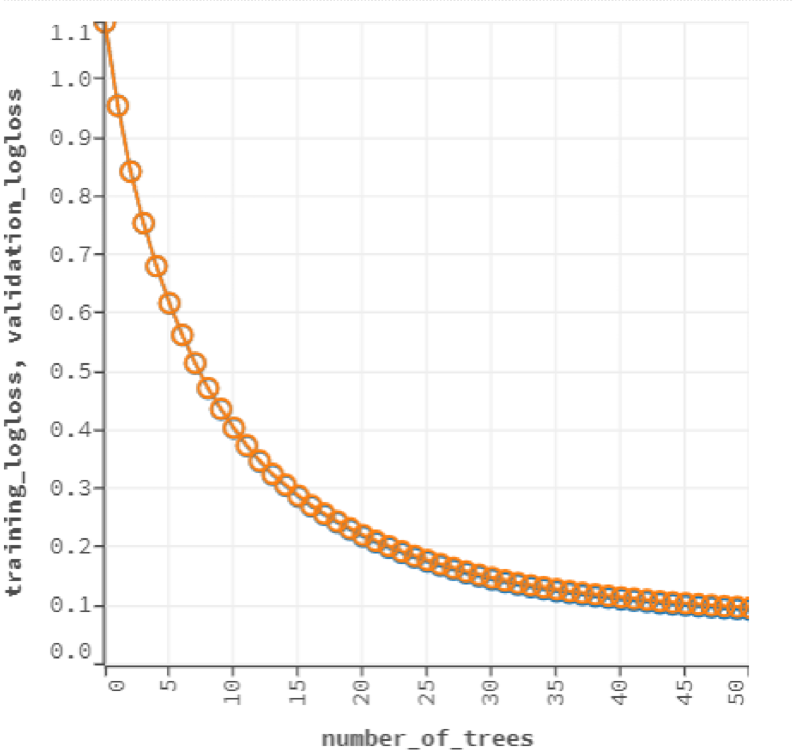


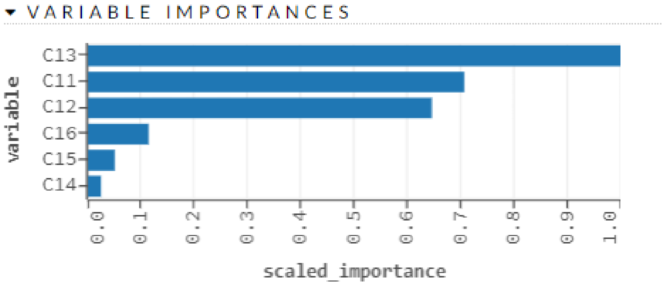
▾ VALIDATION METRICS - CONFUSION MATRIX ROW LABELS

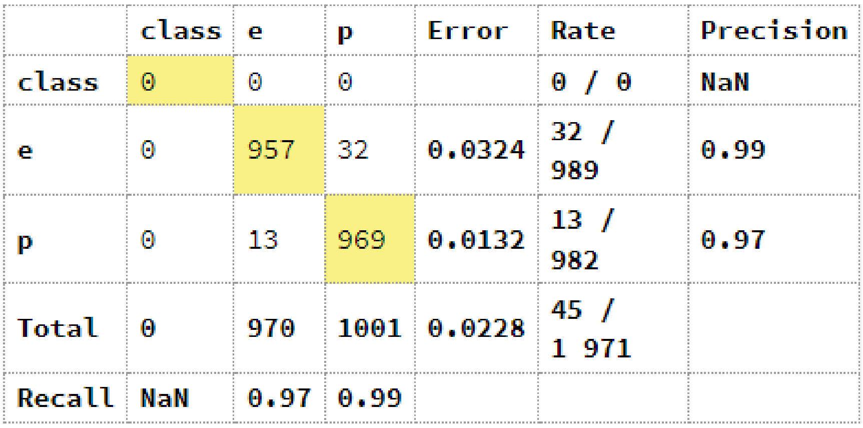
	class	e	p	Error	Rate	Precision
class	0	0	0		0 / 0	NaN
e	0	1052	0	0	0 / 1,052	0.97
p	0	30	970	0.0300	30 / 1,000	1.0
Total	0	1082	970	0.0146	30 / 2,052	
Recall	NaN	1.0	0.97			

Training a model with stalks only

We tried to train model with stalks only that means.
stalk-shape, stalk-root, stalk-surface-above-ring, stalk-surface-bellow-ring, stalk-color-above-ring, stalk-color-below-ring.
So, if you don't want to risk your life, then according to the stalkers, you better not try it, but if you are very hungry, you can risk it





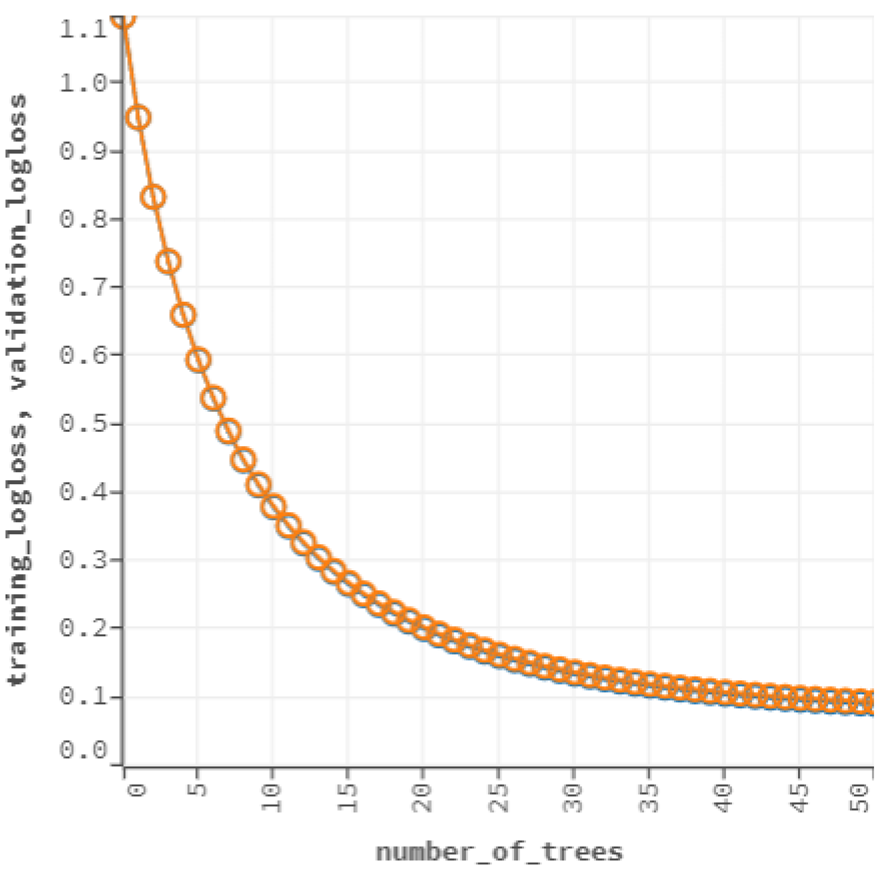


	class	e	p	Error	Rate	Precision
class	0	0	0		0 / 0	NaN
e	0	957	32	0.0324	32 / 989	0.99
p	0	13	969	0.0132	13 / 982	0.97
Total	0	970	1001	0.0228	45 / 1 971	
Recall	NaN	0.97	0.99			

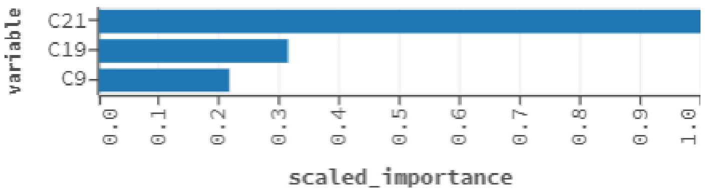
We tried to train the model based only on stalks and it was 98 percent accurate. Usually changed the mushroom that was edible for an poisonous mushroom, which is a better case.

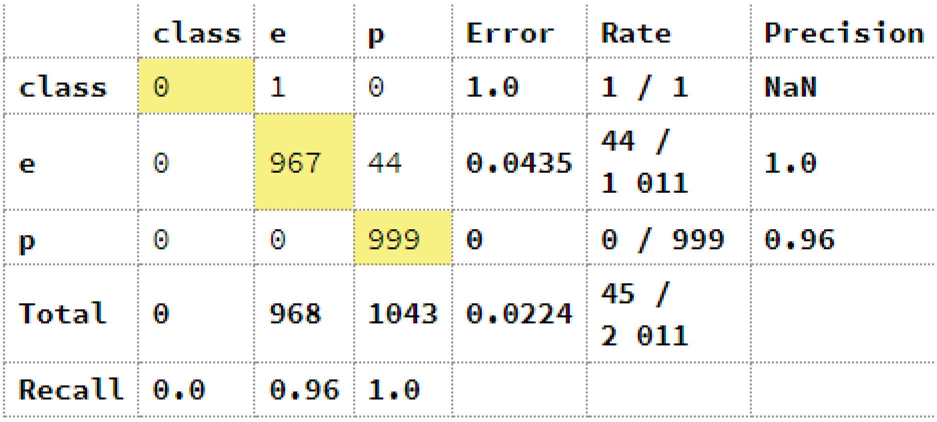
Lowest number of used columns to know if mushroom is edible on 100 percent

We wanted to find out what is the lowest number of dataset attributes required to determine all edible mushrooms. We found out that if you use only columns C9=gill-size, C19=ring-number and C21=spore-print-color you can get accuracy of 96%. With 4% mistakes being mushrooms which were edible but were predicted as poisonous. This is required behavior since we would rather predict mushroom to be poisonous rather than edible (we want to make sure that we do not eat poisonous mushroom).



▾ VARIABLE IMPORTANCES





	class	e	p	Error	Rate	Precision
class	0	1	0	1.0	1 / 1	NaN
e	0	967	44	0.0435	44 / 1 011	1.0
p	0	0	999	0	0 / 999	0.96
Total	0	968	1043	0.0224	45 / 2 011	
Recall	0.0	0.96	1.0			