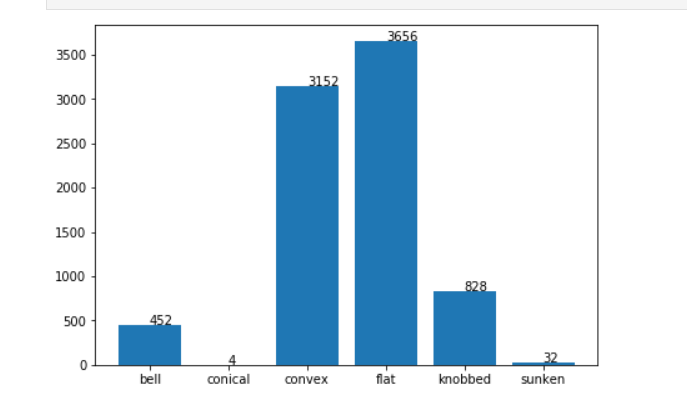
**2.data visualization** (hw1\_1.ipynb,hw1\_2.ipynb)**:**

hw1\_1.ipynb因為要畫圖過多，所以有warning，但是圖都有畫出來，在我的電腦上跑也只須等個幾秒就好。

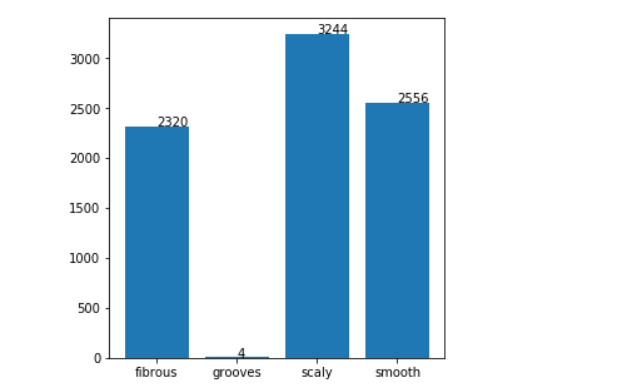
Mushroom dataset:

Value frequency of every feature:

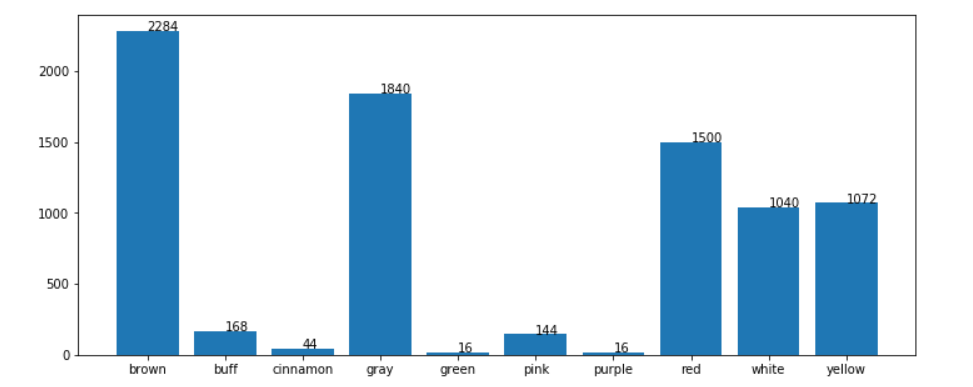
Cap\_shape



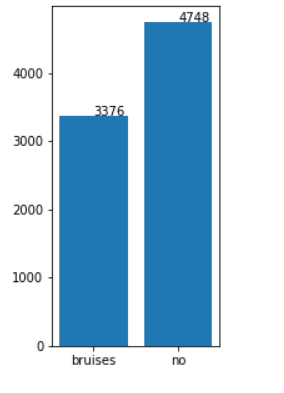
Cap\_surface:



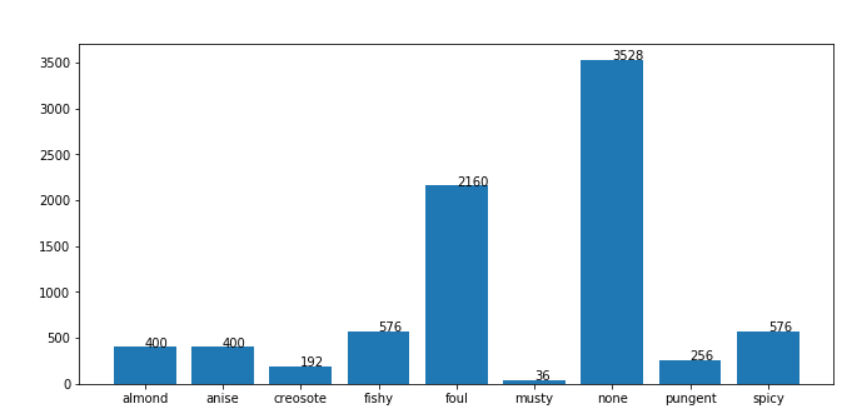
Cap\_color:



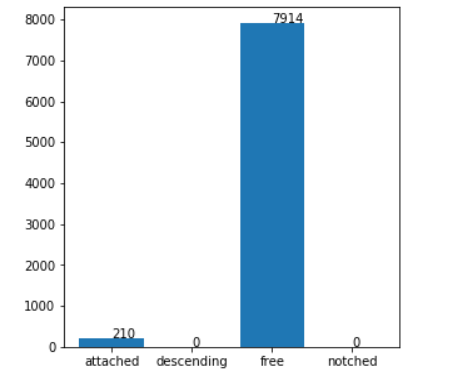
Bruises:



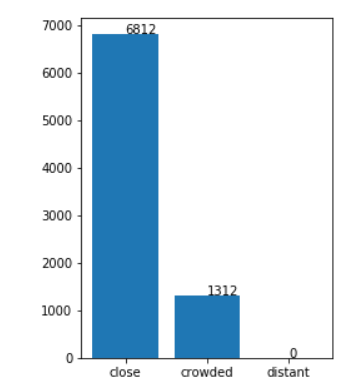
Odor:



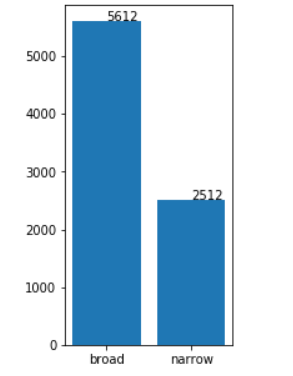
Gill\_attachment:



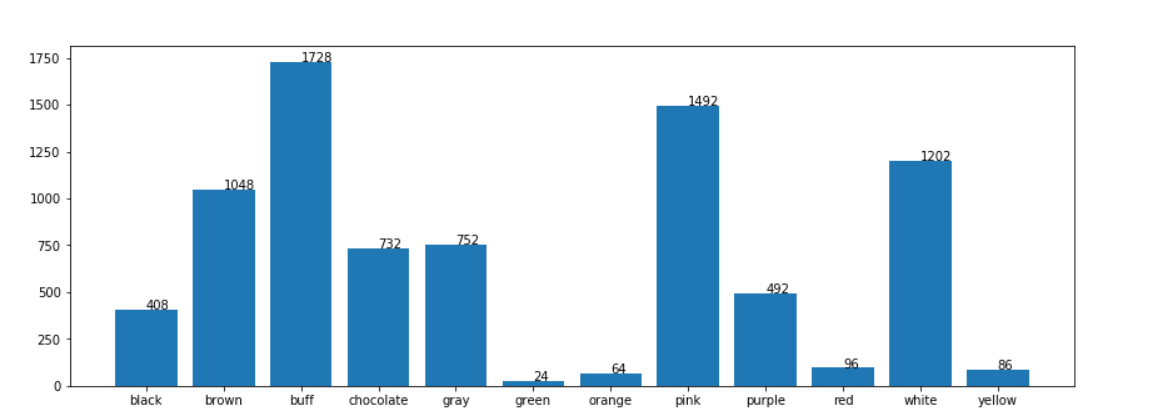
gill-spacing:



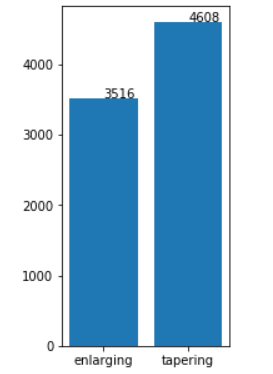
gill-size:



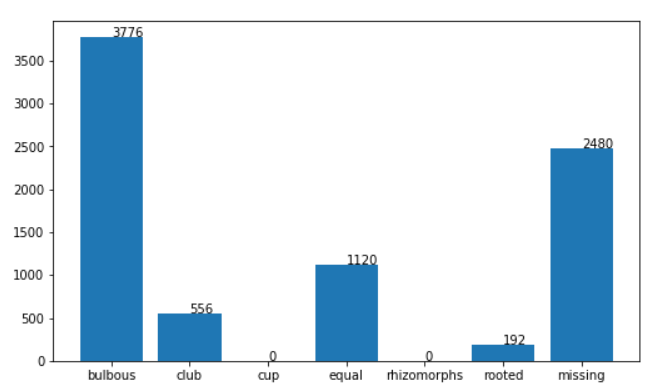
gill-color:



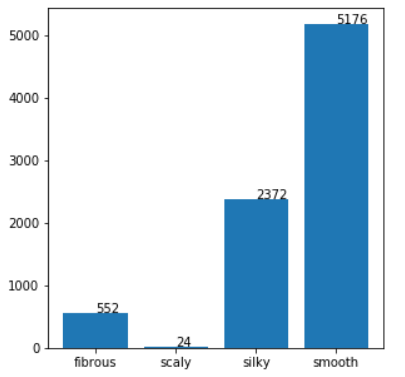
stalk-shape:



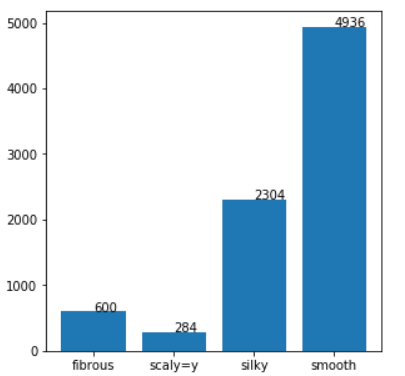
stalk-root:



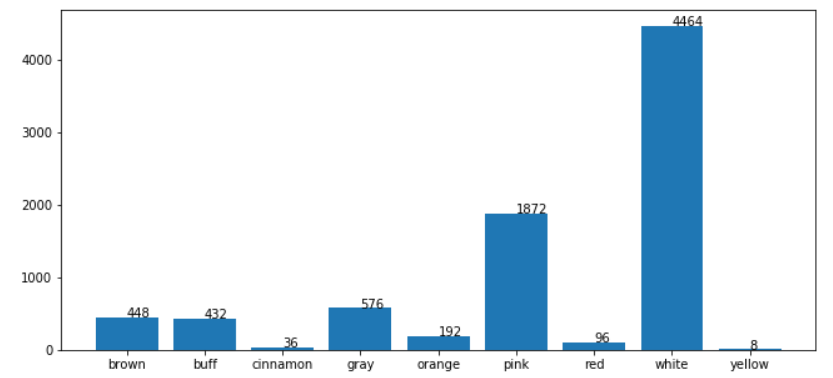
stalk-surface-above-ring:



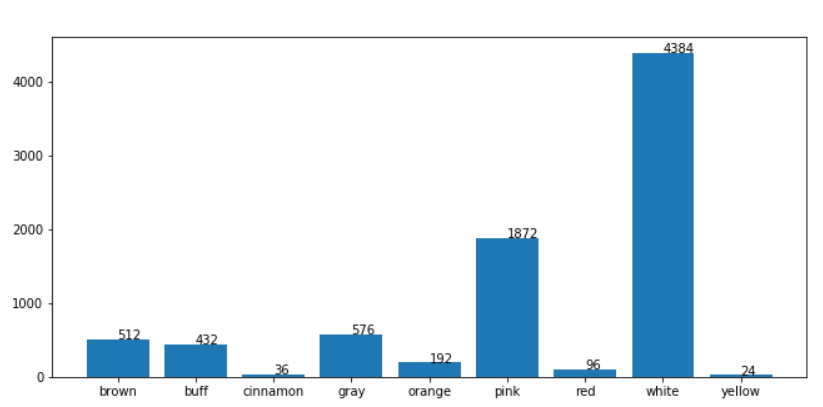
stalk-surface-below-ring:



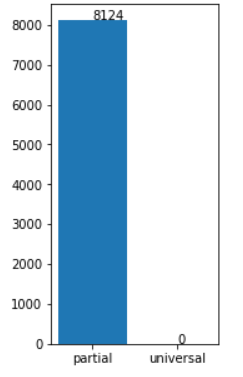
stalk-color-above-ring:



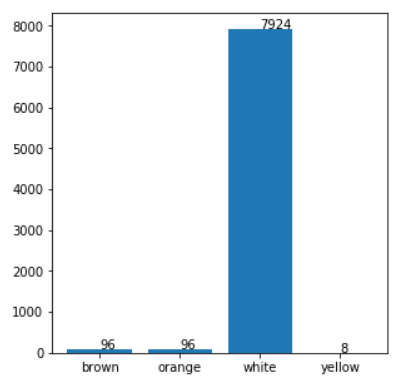
stalk-color-below-ring:



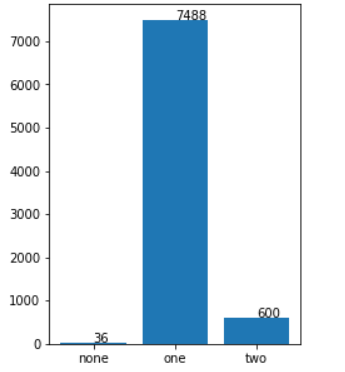
veil-type:



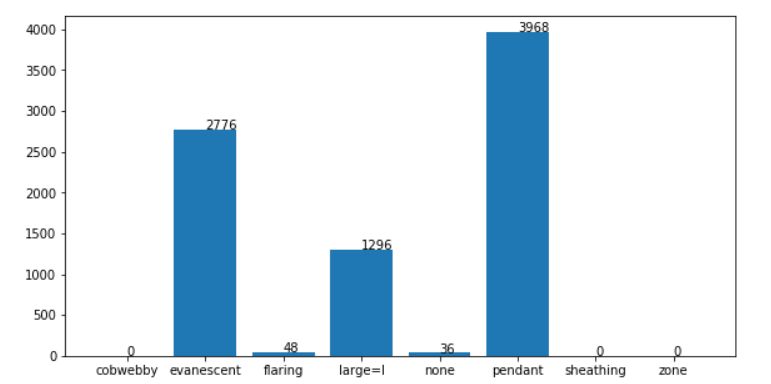
veil-color:



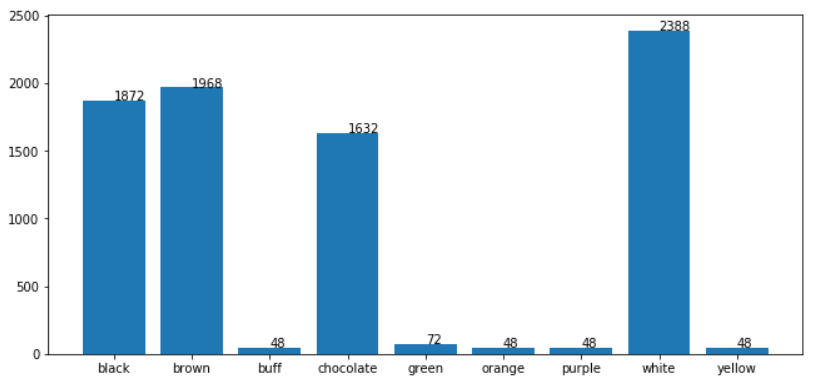
ring-number:



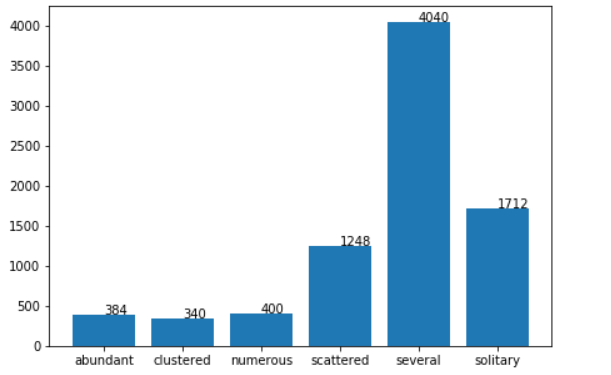
ring-type:



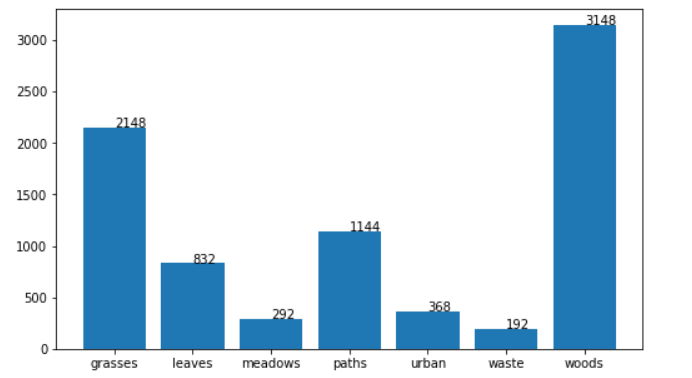
spore-print-color:



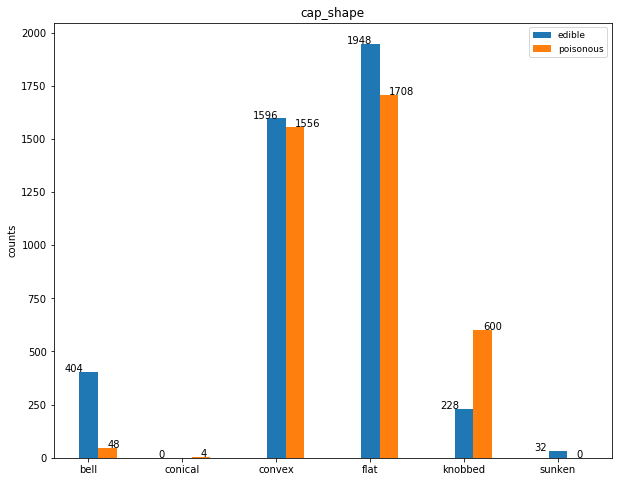
population:

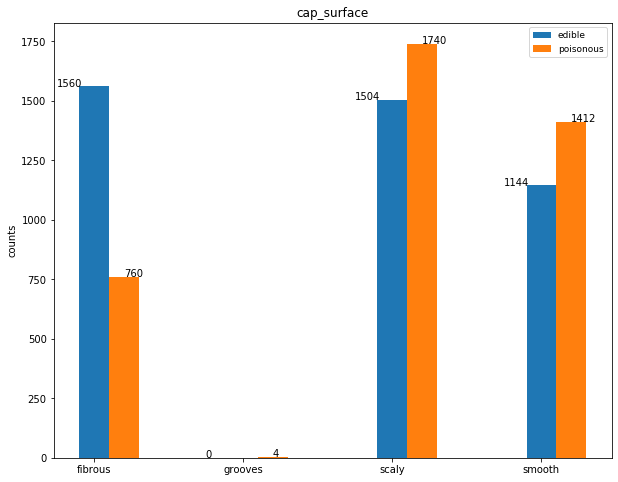


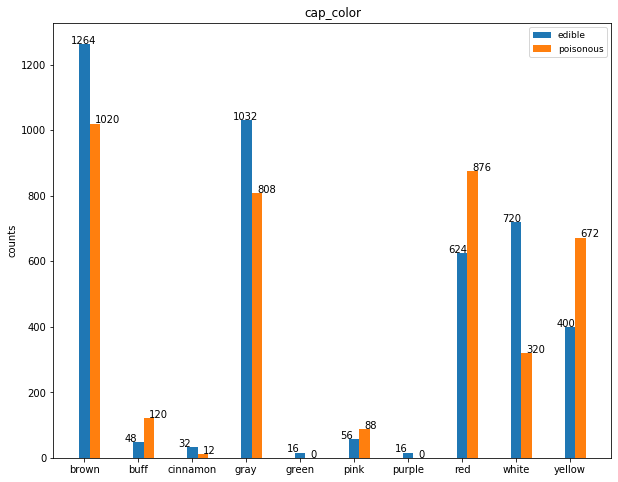
habitat:

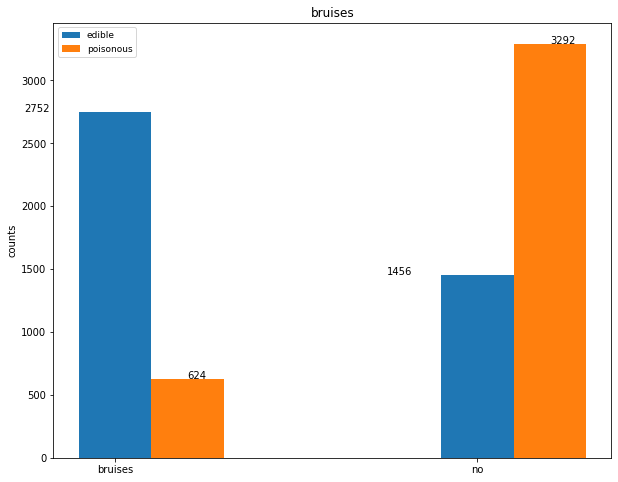


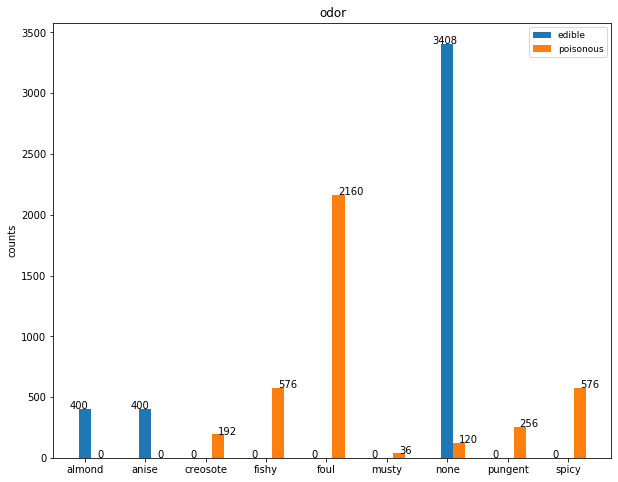
Mushroom :split data based on their labels and show the data distribution again:

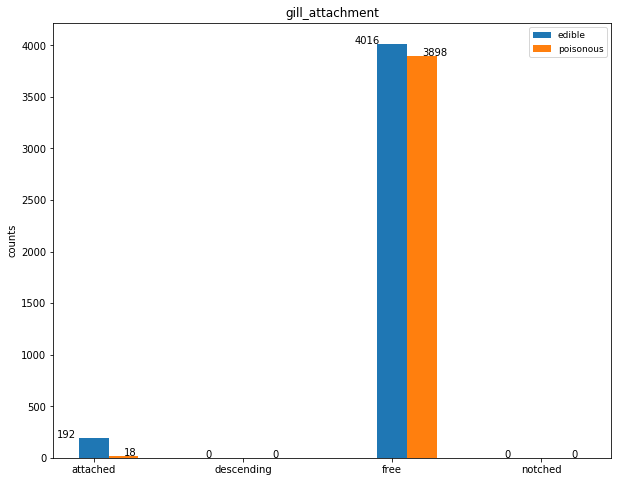


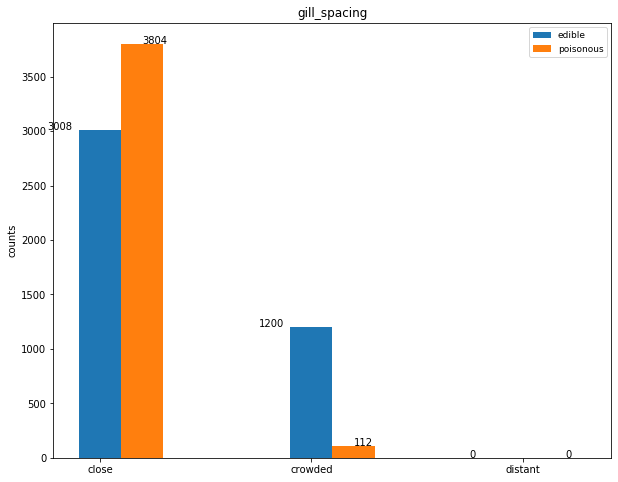


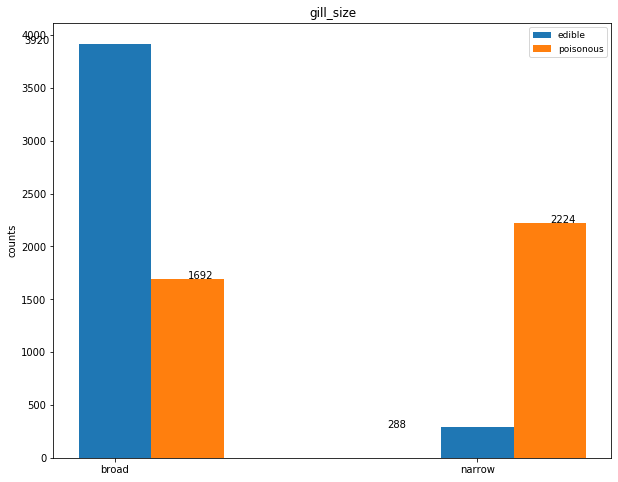


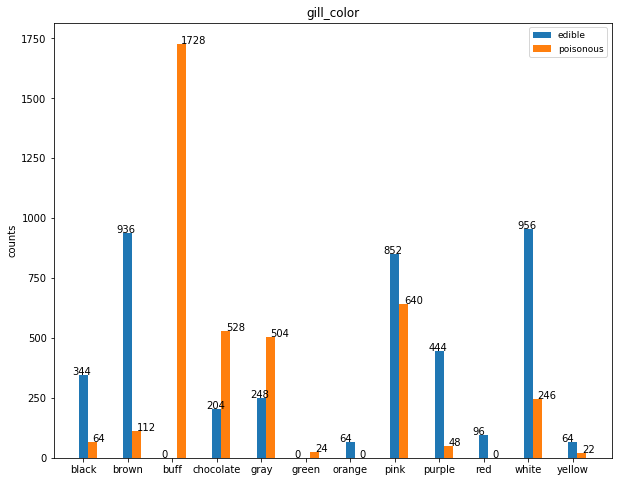


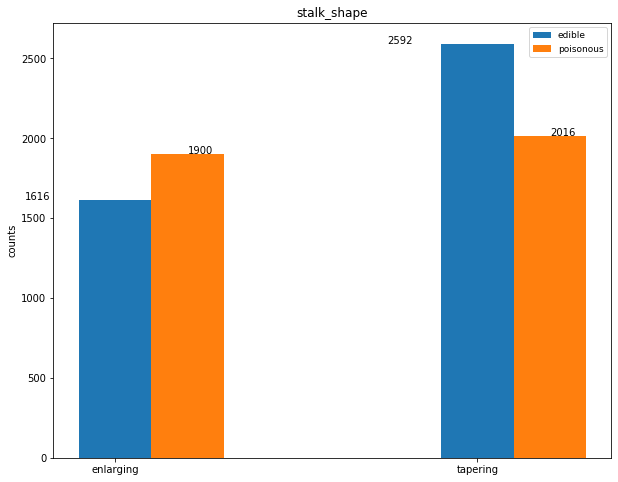


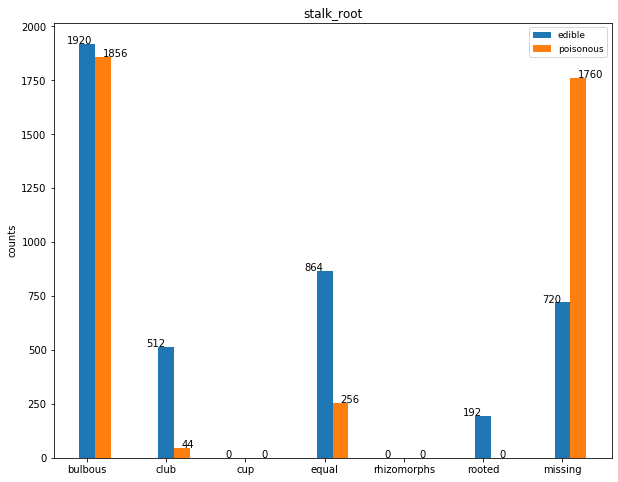


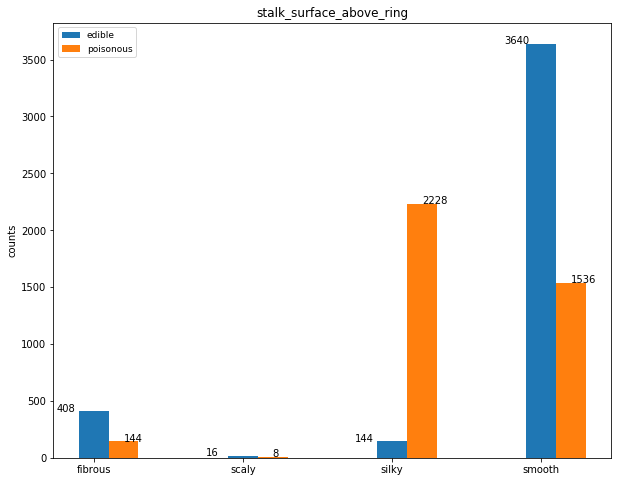


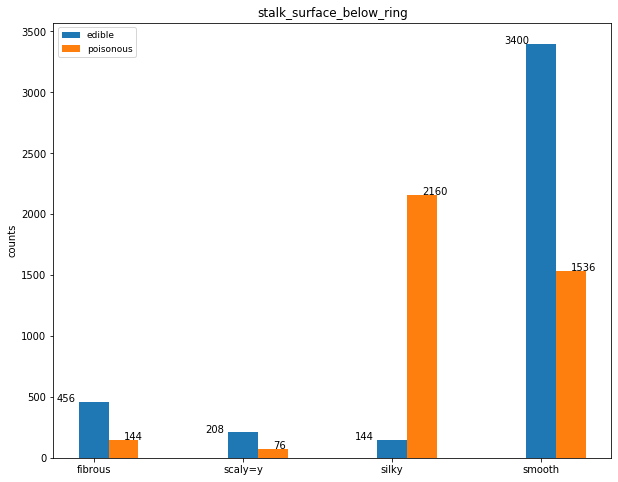


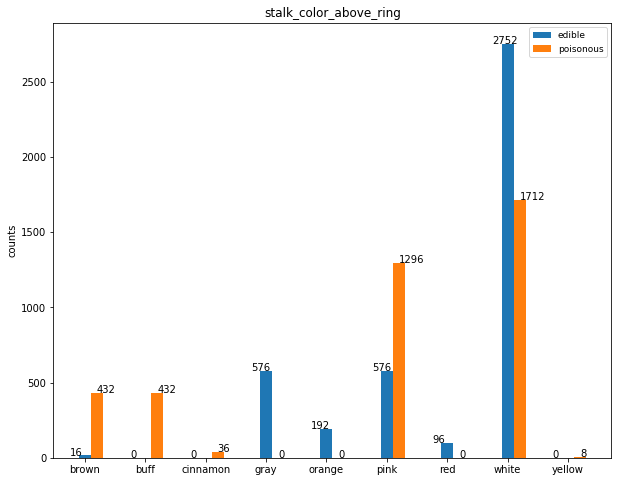


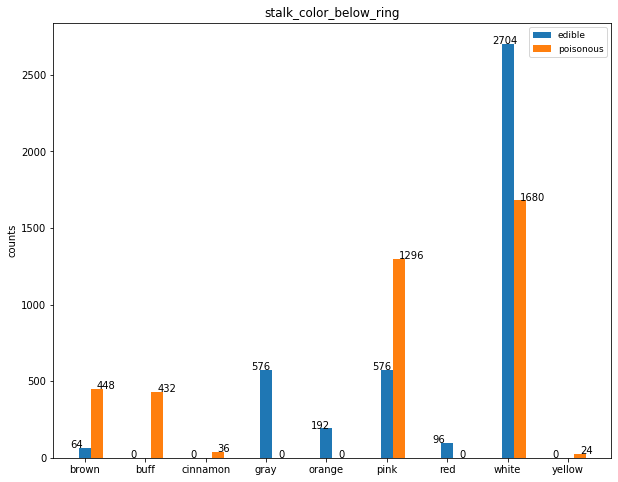


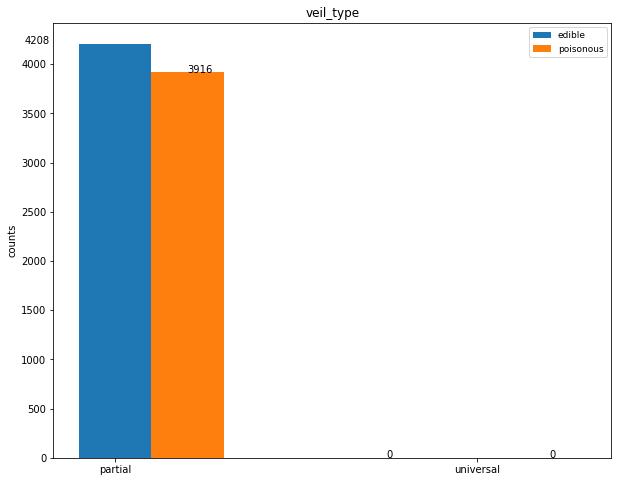


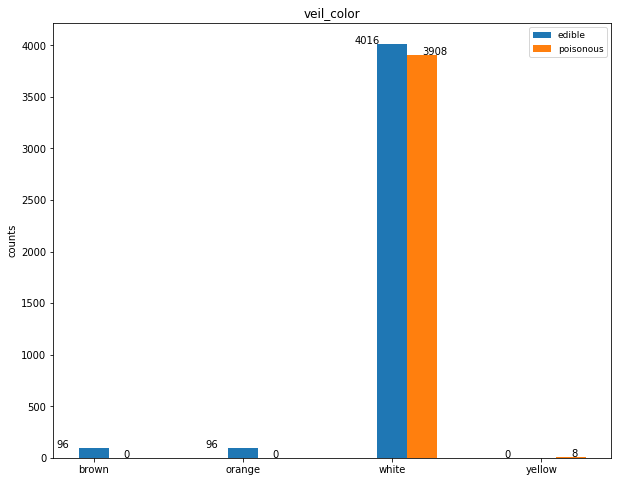


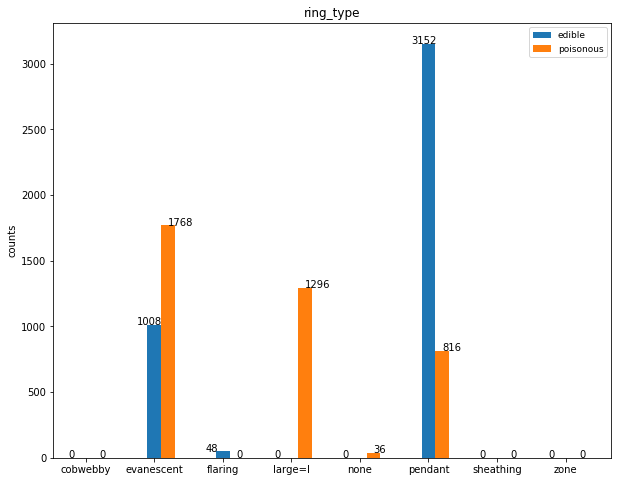
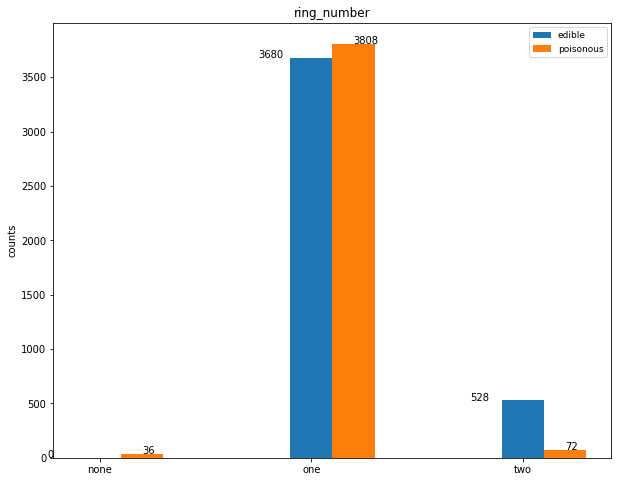


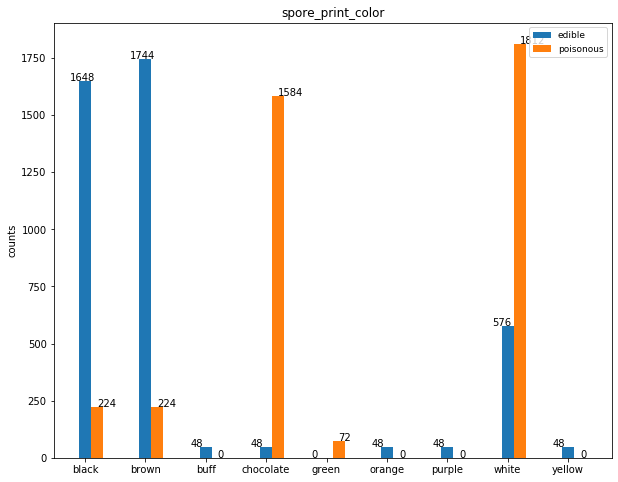


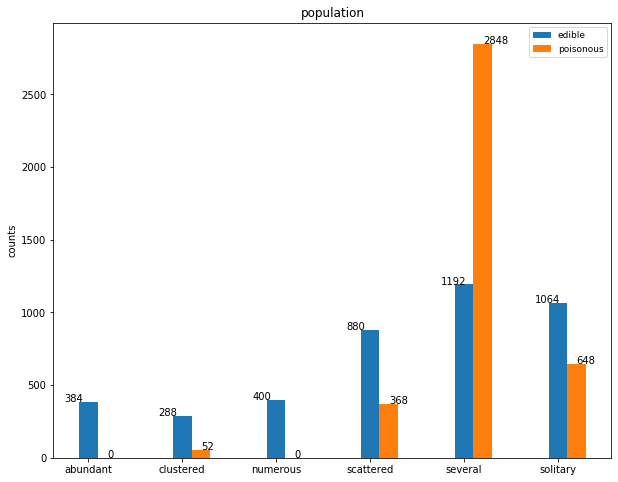


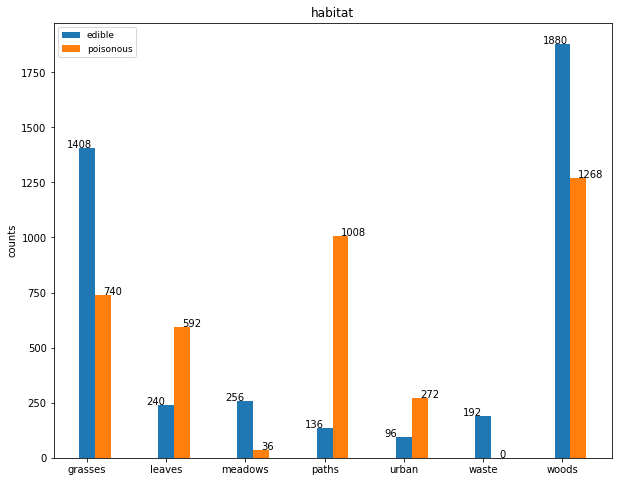






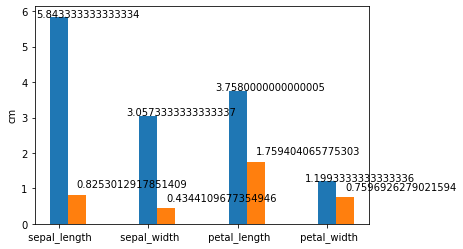




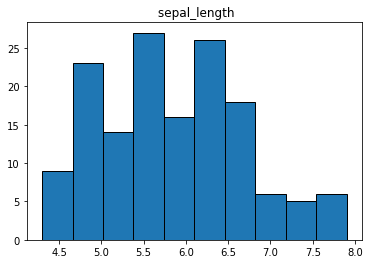


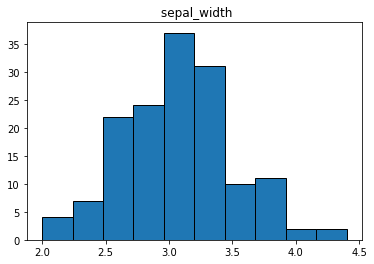
**Iris dataset:**

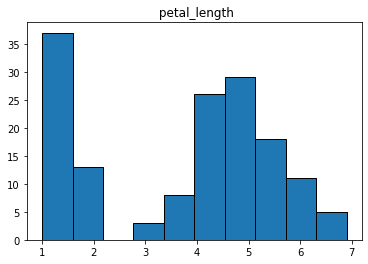
**Blue** : average orange: **standard deviation**

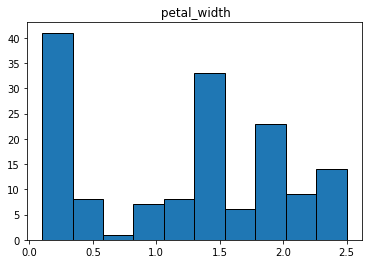


Value frequency of each feature:







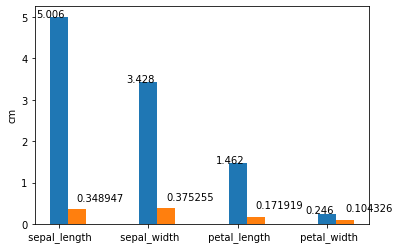


Iris dataset with split data based on their labels:

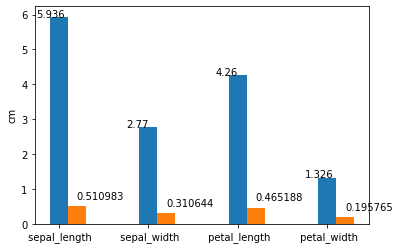
**average, standard deviation:**

**Blue** : average orange: **standard deviation**

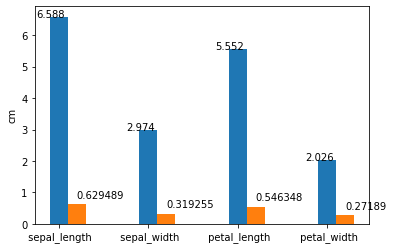
**Iris-setosa:**



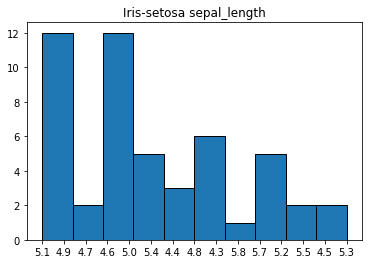
Iris-versicolor:

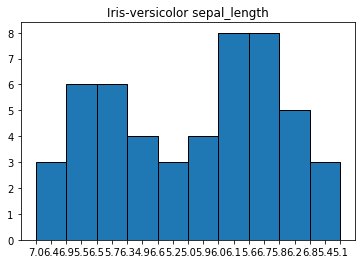


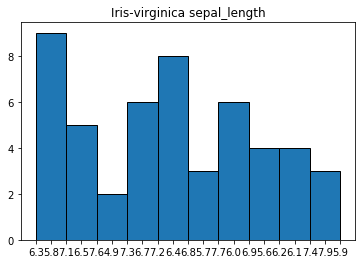
Iris-virginica

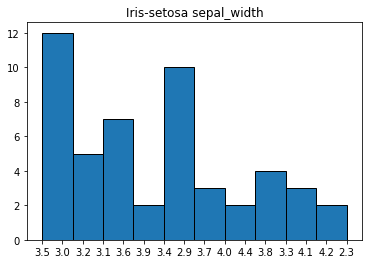


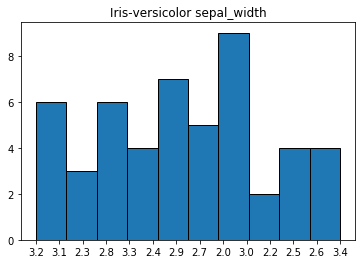
Each label’s frequency:

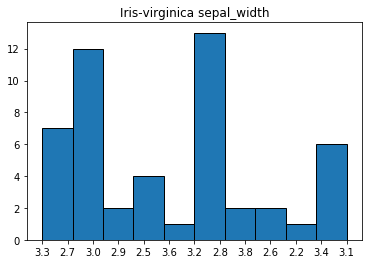


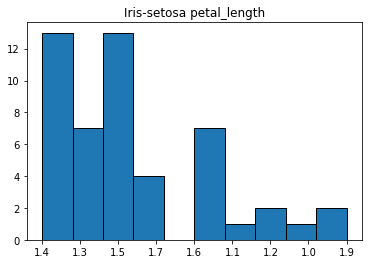


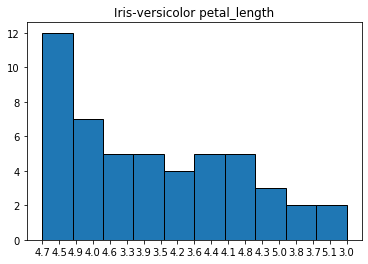


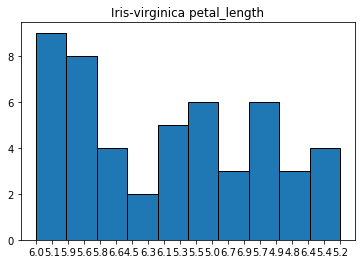


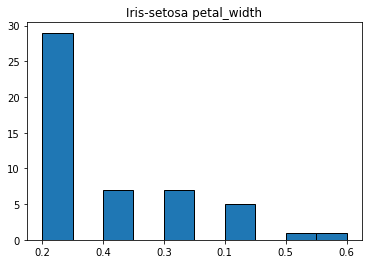


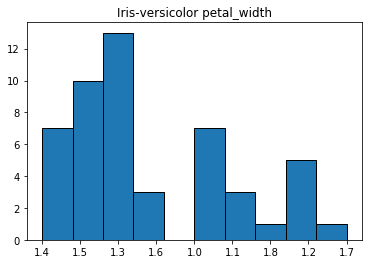


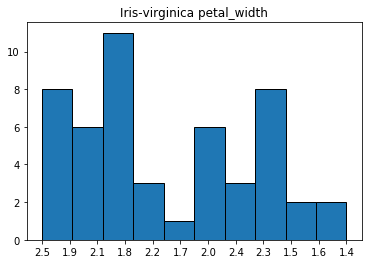












**3. Data preprocessing:**

Drop feature:

don’t use stalk\_root to bulid classfier in my code because it has missing value.

Shuffle the data:

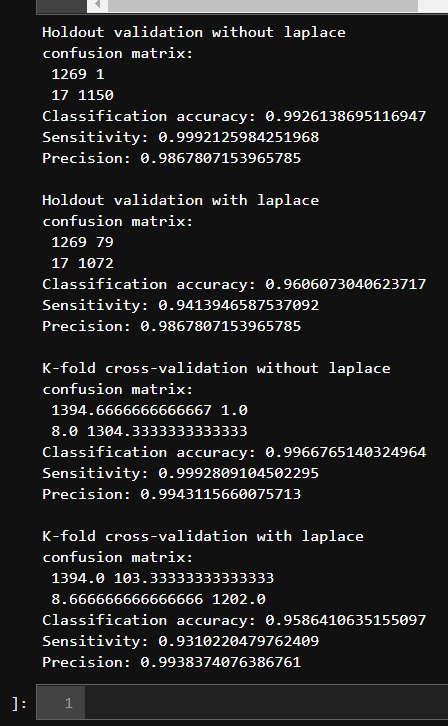
Use random.shuffle() in python to shuffle the data.

3.Data Preprocessing+4.Model Construction

+5.Train-Test-Split in code +6.Results hw1\_3.ipynb,hw1\_4.ipynb.

**6 result:**

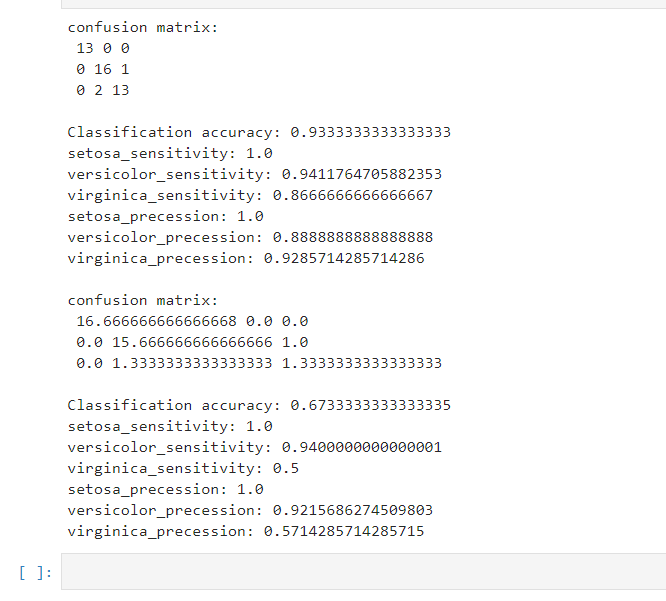
Mushroom:



Iris:

Above is [Holdout validation](https://en.wikipedia.org/wiki/Cross-validation_(statistics)#Holdout_method) with the ratio 7:3

Below is [K-fold cross-validation](https://en.wikipedia.org/wiki/Cross-validation_(statistics)#k-fold_cross-validation) with 𝐾=3



7.Comparison & Conclusion:

如果有做laplace smoothing 準確度會下降，但是可以處理train data feature中有項目為0的情況，[K-fold cross-validation](https://en.wikipedia.org/wiki/Cross-validation_(statistics)#k-fold_cross-validation)跟[Holdout validation](https://en.wikipedia.org/wiki/Cross-validation_(statistics)#Holdout_method) 起來多了小數點，因為它是多次平均綜合，感覺結果比較讓人信服。