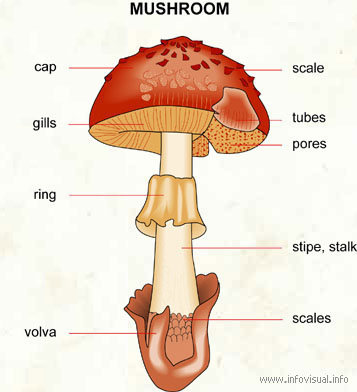
**Part 2) Machine Learning (10%)**

This exam is a classification task to identify whether a mushroom is edible or poisonous.

A picture containing fungus, table, sitting, man

Description automatically generated

Please download data in the link below:

<https://github.com/kaopanboonyuen/Python-Data-Science/raw/master/Dataset/hed2020_dataset.csv>

You have to submit the following items to MyCourseVille:

* File.ipynb – a source code
* File.docx – a capture screen of each question. It cannot be graded if we cannot map your answer to the question.
* All files must be renamed as “{student\_id}\_{firstname}\_Part2”, e.g., 6030133421\_Chaiyatad\_Part2.ipynb

# Introduction

This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981).

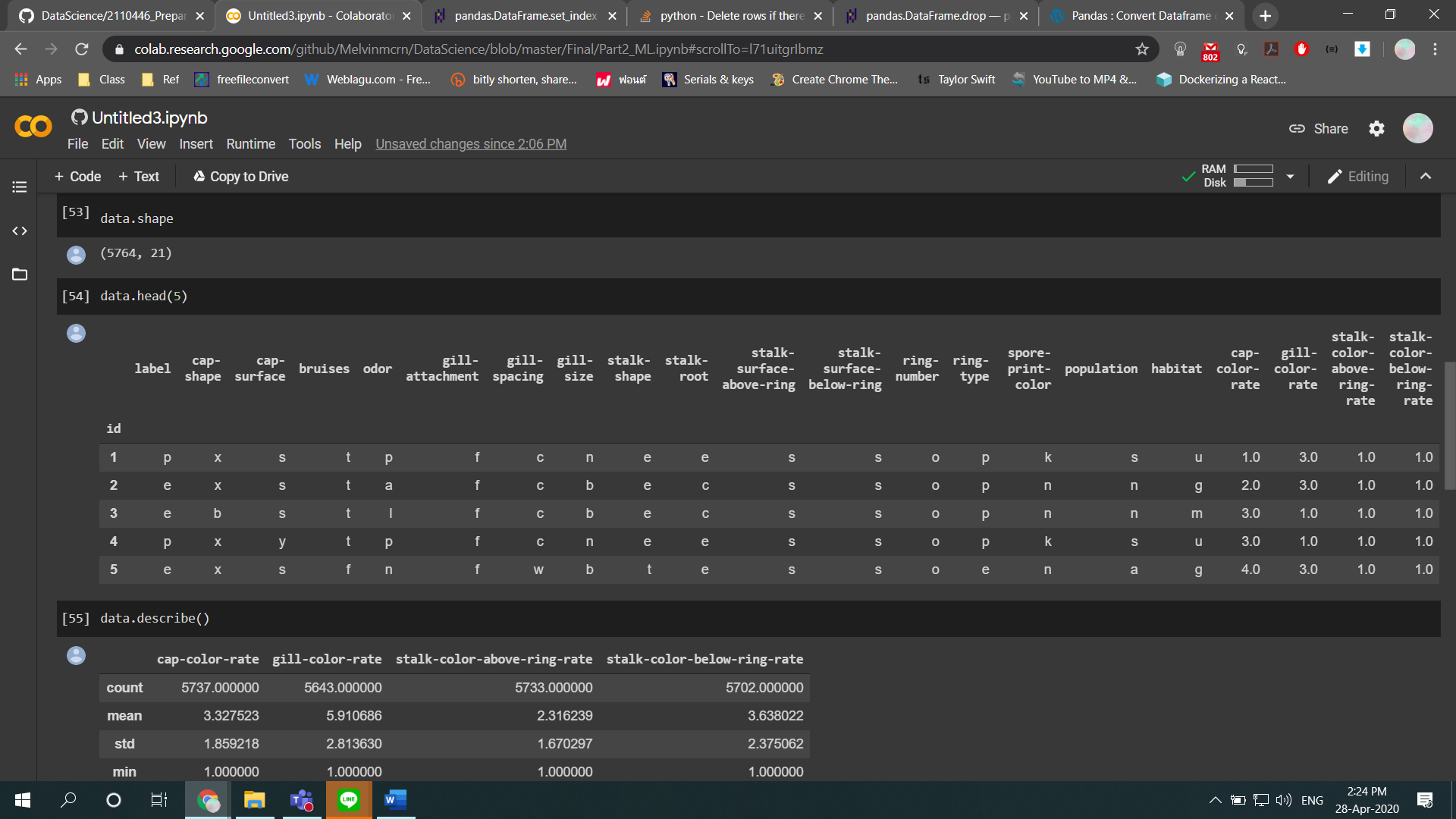
Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the credibility of a mushroom; no rule like "leaflets three, let it be'' for Poisonous Oak and Ivy.

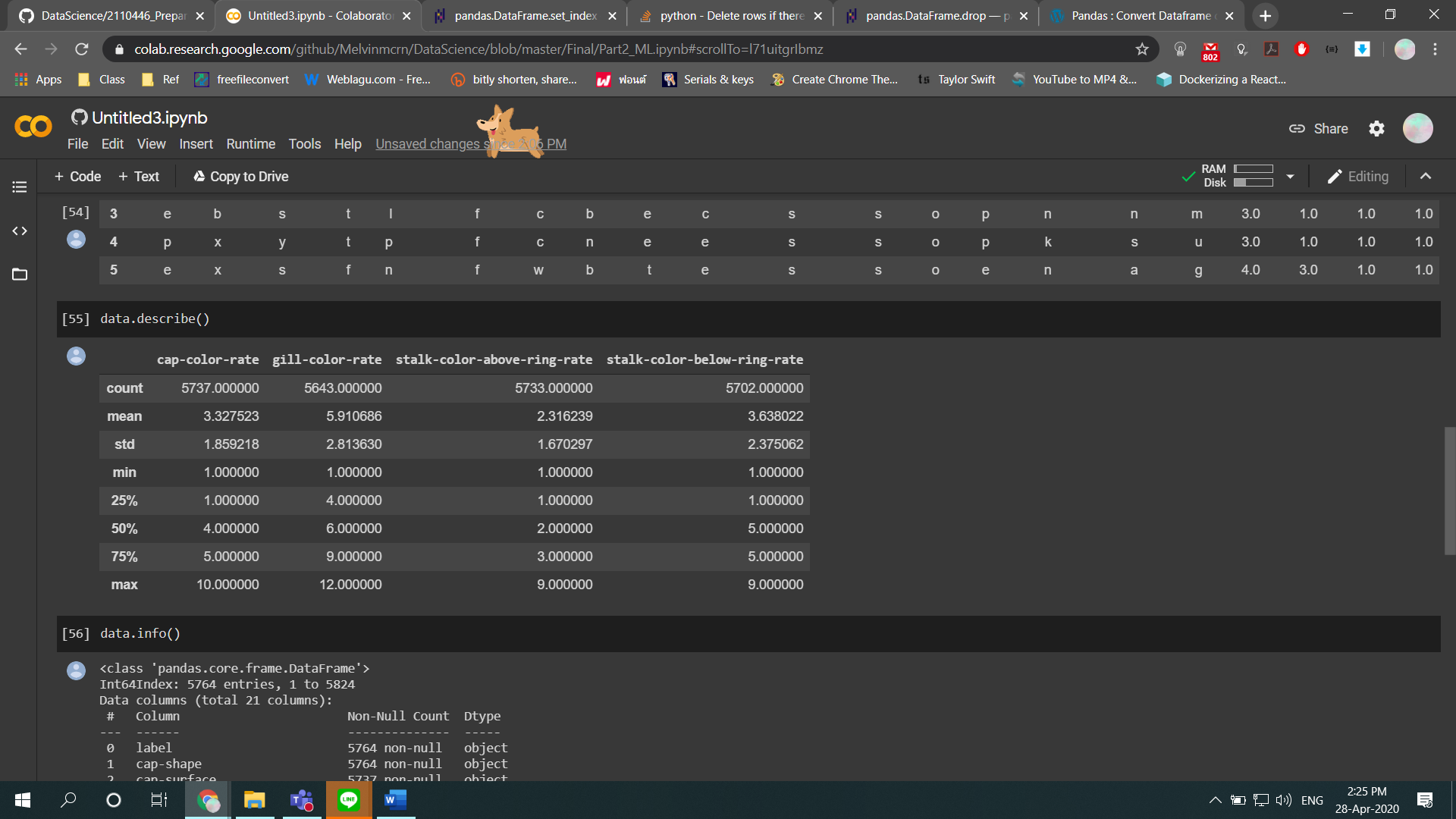
# Attribute Information

* label: edible=e, poisonous=p
* cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
* cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
* cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w,yellow=y
* bruises: bruises=t,no=f
* odor: almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m,none=n,pungent=p,spicy=s
* gill-attachment: attached=a,descending=d,free=f,notched=n
* gill-spacing: close=c,crowded=w,distant=d
* gill-size: broad=b,narrow=n
* gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e,white=w,yellow=y
* stalk-shape: enlarging=e,tapering=t
* stalk-root: bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r,missing=?
* stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
* stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
* stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
* stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
* veil-type: partial=p,universal=u
* veil-color: brown=n,orange=o,white=w,yellow=y
* ring-number: none=n,one=o,two=t
* ring-type: cobwebby=c,evanescent=e,flaring=f,large=l,none=n,pendant=p,sheathing=s,zone=z
* spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y
* population: abundant=a,clustered=c,numerous=n,scattered=s,several=v,solitary=y
* habitat: grasses=g,leaves=l,meadows=m,paths=p,urban=u,waste=w,woods=d

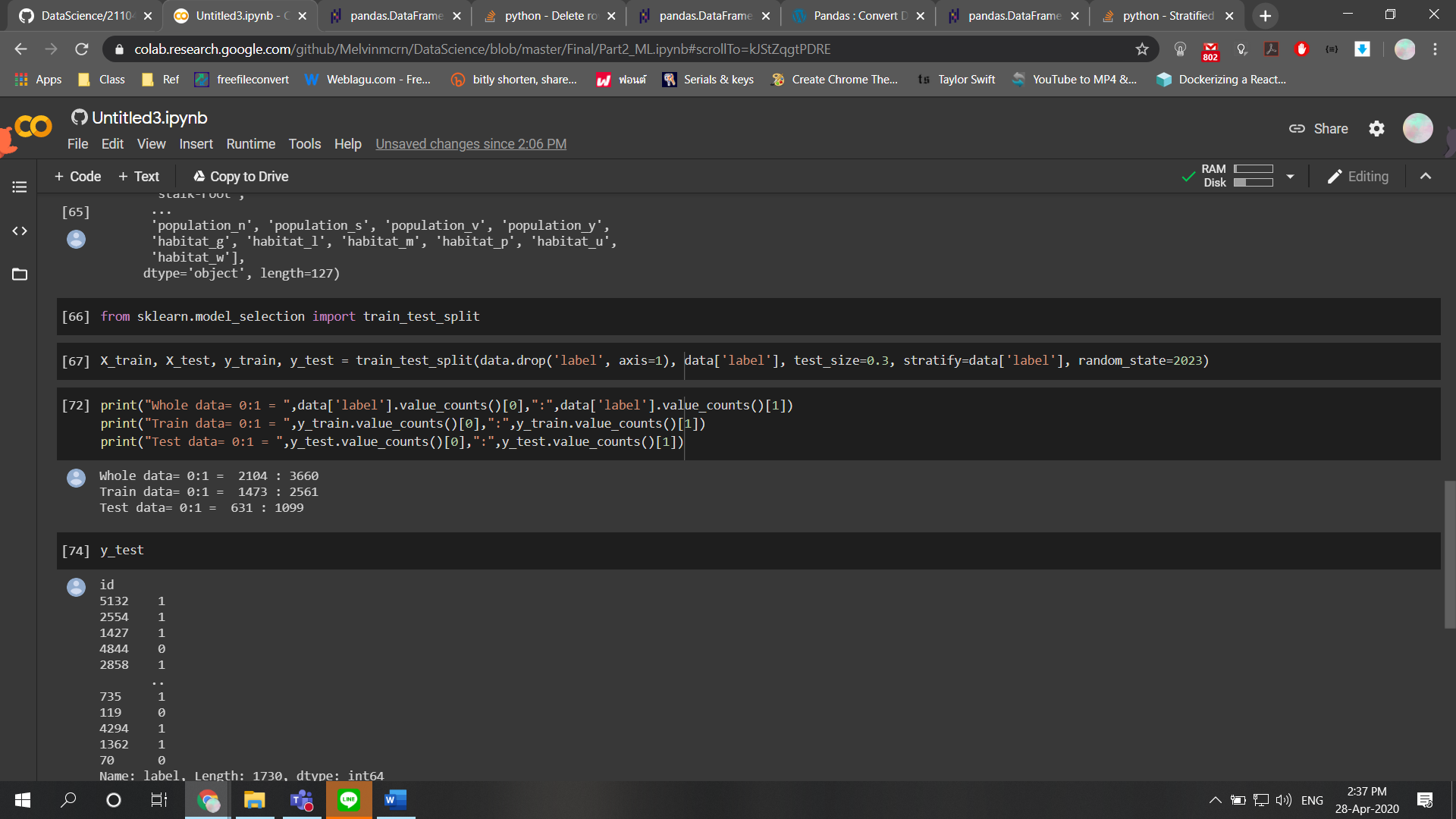
# Tasks

1. Import data and preprocess the data [2 points]
   1. Set “id” to be index
   2. Remove missing target
   3. Remove unused variables “prefix with veil”
   4. (Screen capture) Preview top 5 rows, show data statistics of **all** variables (Hint: describe(), info())

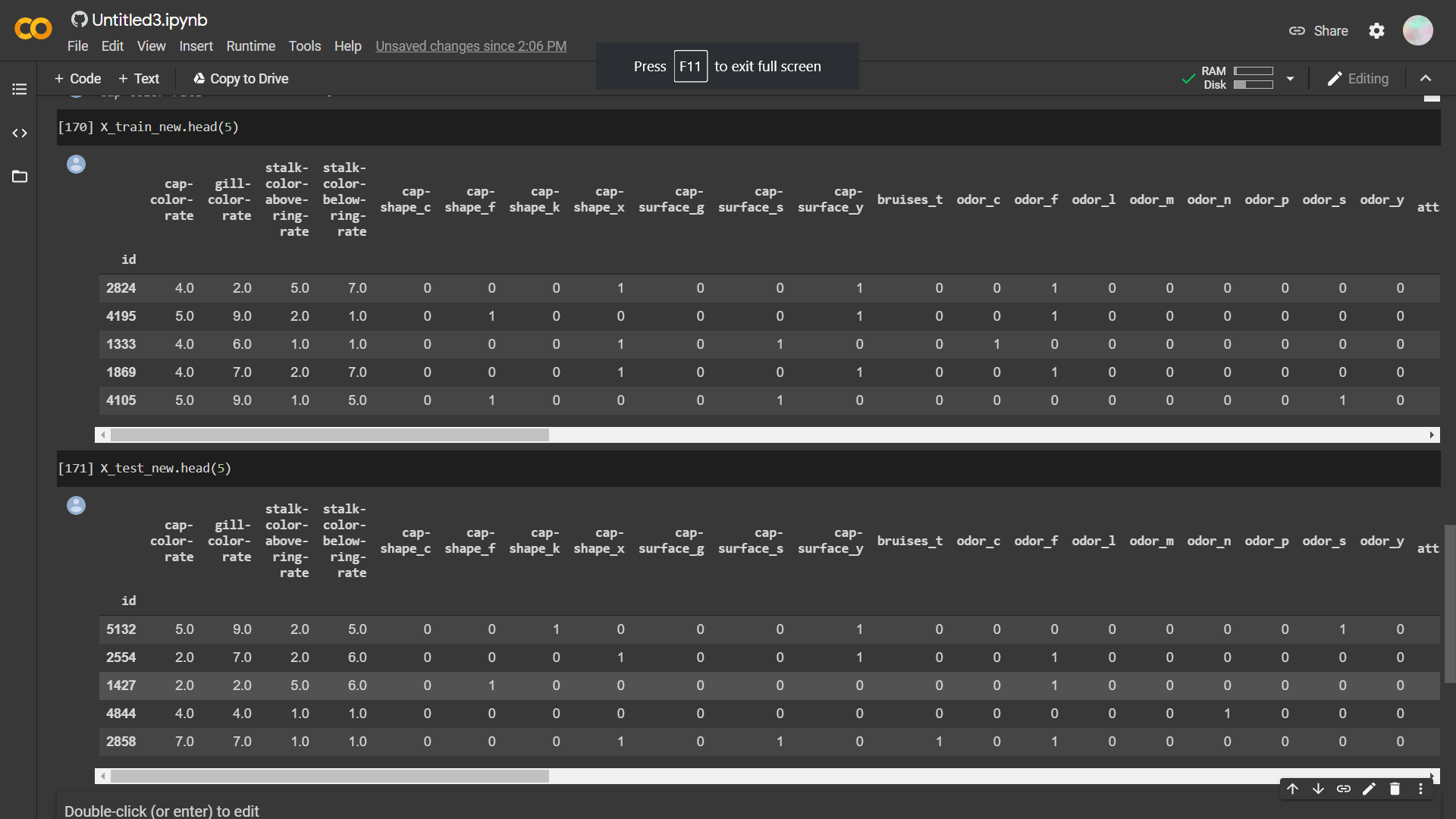


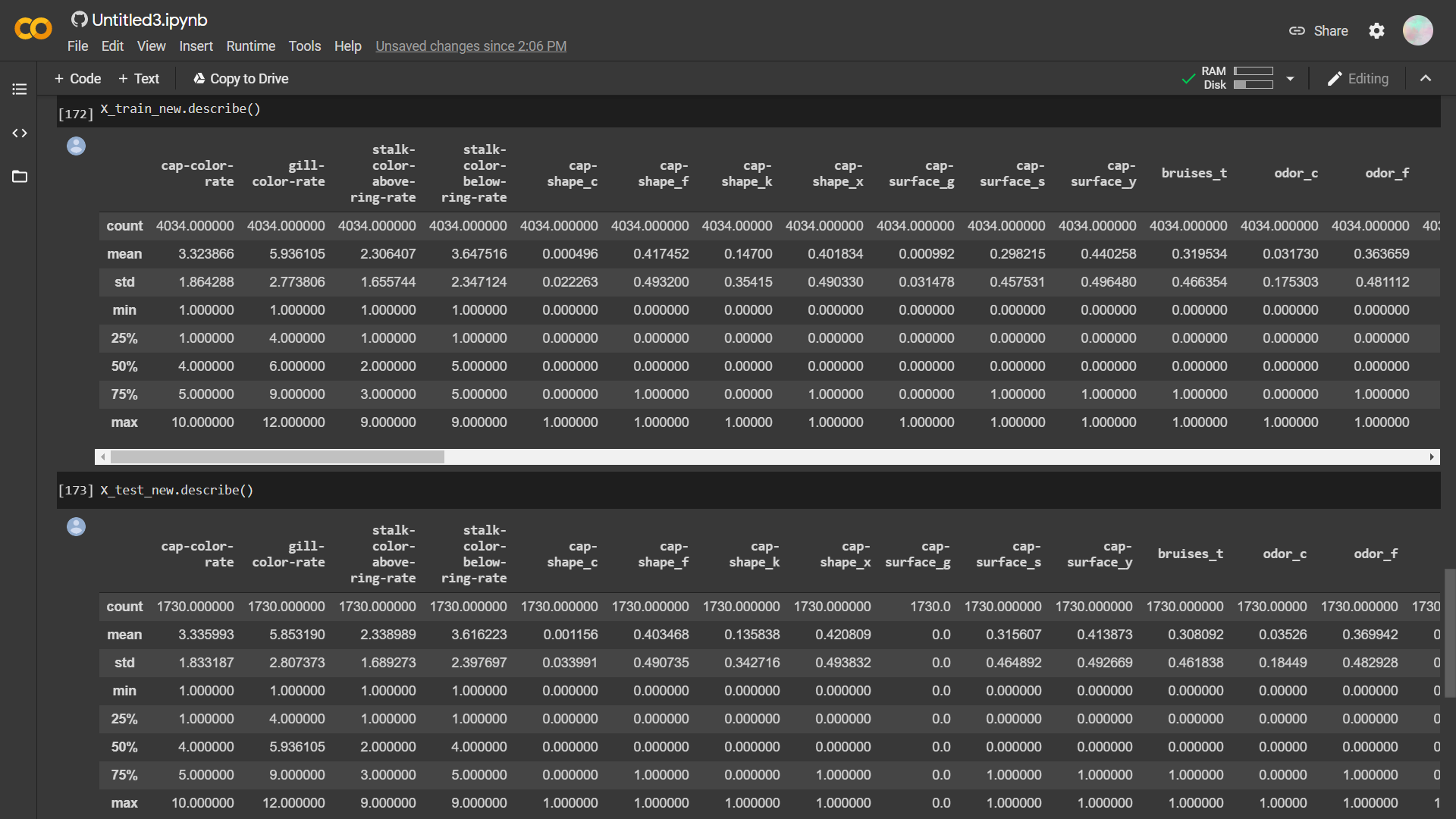


1. Transform the data [2 points]
   1. Replace the target variable using 0 as edible and 1 as poisonous.
   2. All categorical inputs must convert to be dummy codes with option to remove the first level.
   3. Split train/test with stratification by setting testing size to be 30% and random seed to be 2023.
   4. (Screen capture) Show the proportion between each target class in the whole data, train, test

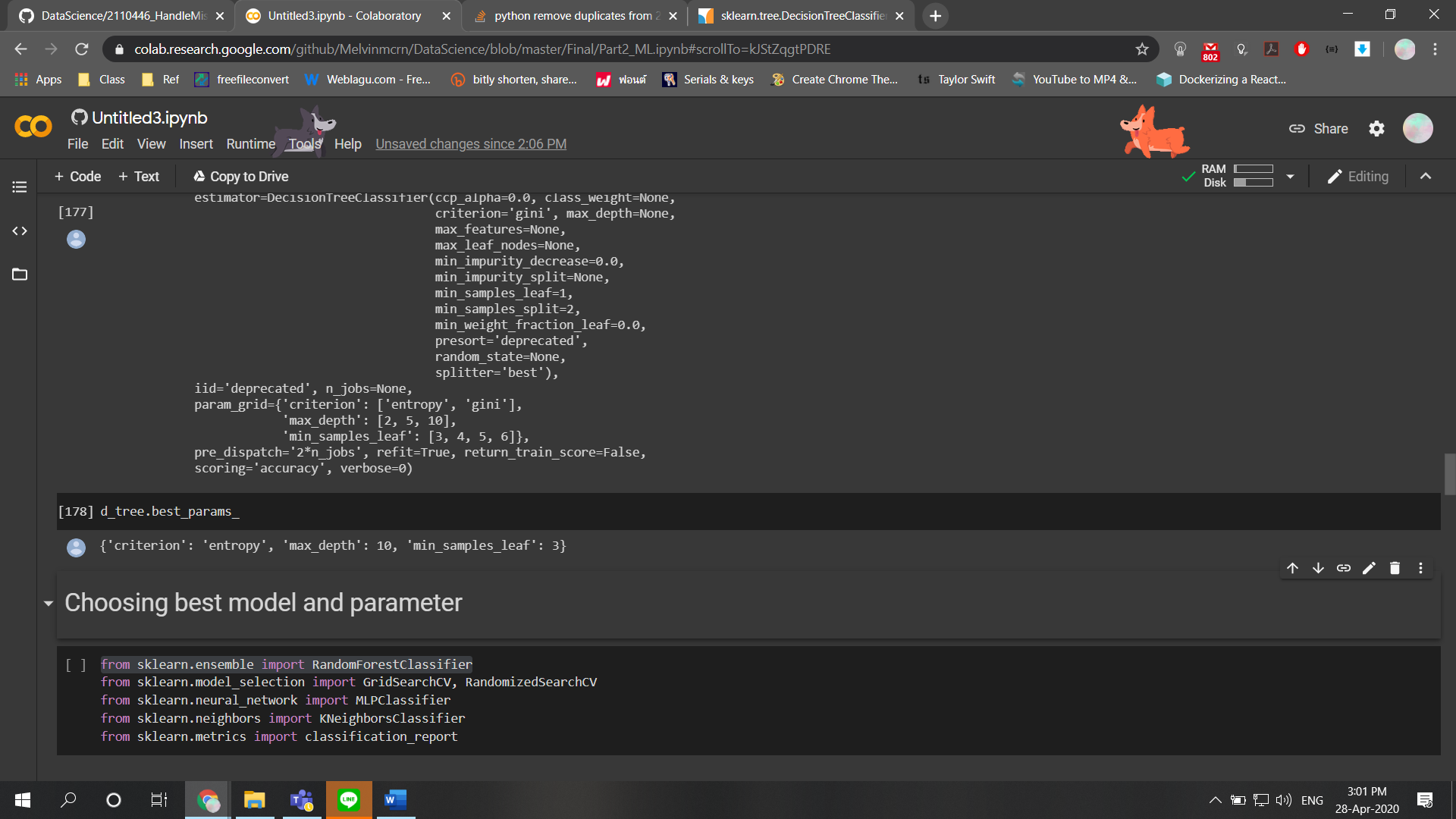


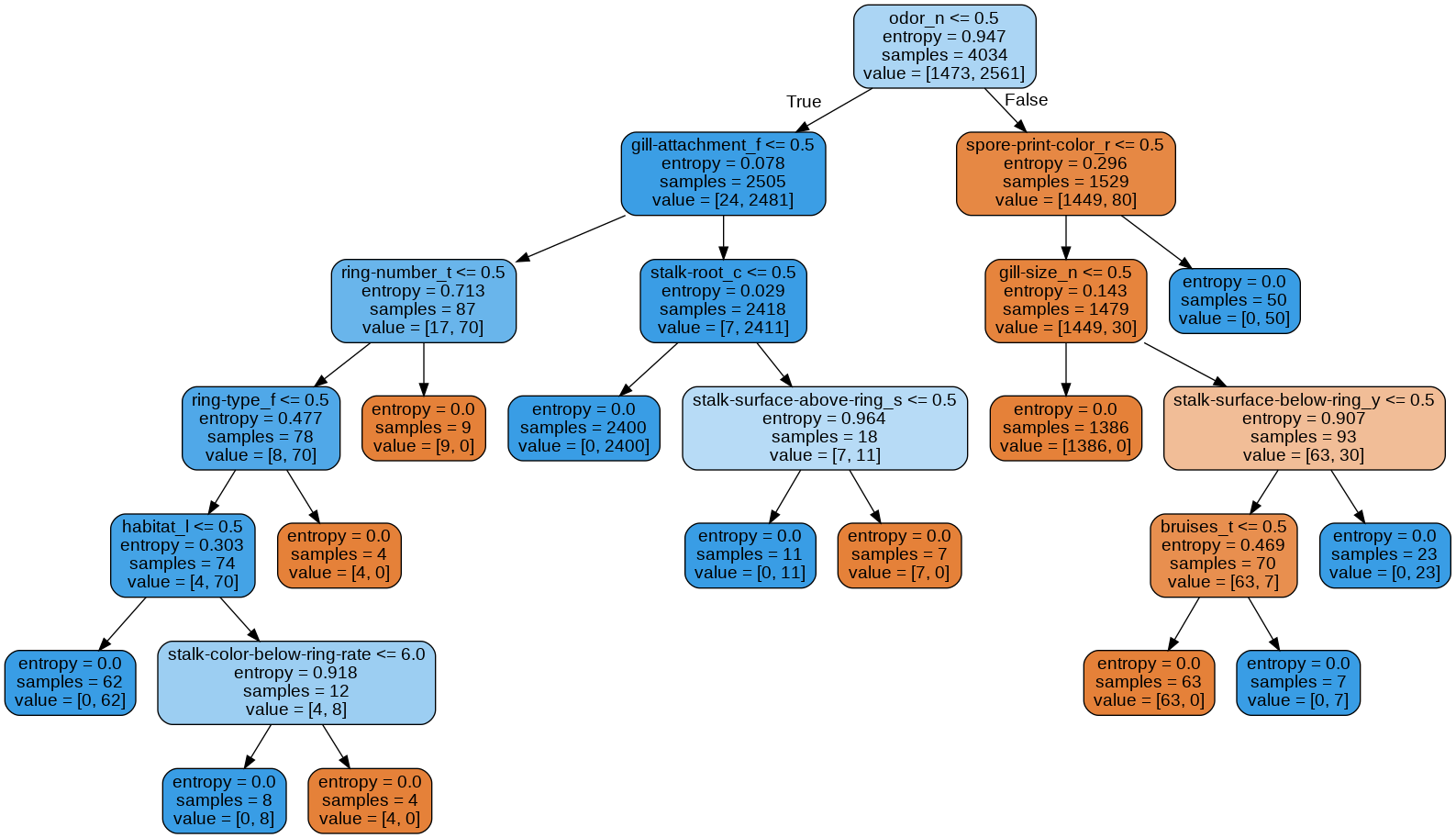
1. Continue Transform the data [2 points]
   1. Use the statistics from “training data” to impute missing values in train & test data sets
   2. (Screen capture) Preview top 5 rows, show data statistics of **all** variables (Hint: describe(), info())





1. Create a prediction model [2 points]
   1. Use a decision tree model with a grid search with 5CV as follows
      * The minimum number of examples at leaf node should be {3, 4, 5, 6}.
      * The maximum tree depth is {2, 5, 10}.
      * The splitting criteria is {entropy, gini}.
   2. (Screen capture) print the best parameters and **the figure** of the decision tree models





1. Evaluation on **the test data** [2 points]
   1. (Screen capture) confusion matrix, classification report
   2. (Screen capture) predict the case id= 1012 along with its true label