The Milestone Report of Capstone Project 1

1. The Aiming Problem

According to the American Association Poison Control Center(AAPCC)'s annual report, there were over 6000 people poisoned by poisonous mushroom per year over the past decade in U.S. The tricky thing about poisonous mushroom is, there is no simple rule to identify them from the non-poisonous ones by their outlook. Hence even experienced mushroom hunters may mistake a poisonous mushroom from the wild.
The objective of the project is to find out if it is possible to identify the poisonous mushroom by the outlook.

2. The Clients

The aiming clients of the project is professional/amateur mushroom hunters, survivalist and anyone who would like to put wild mushroom onto their dining tables but not sure about its edibilty.

3. The Data Set

It will be using the mushroom data set from UCI's machine learning repository.
The dataset has 8124 observations, 22 features. Here is the details:
Attribute Information:
(classes: 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

4. Data Wrangling

The mushroom data set is quite clean already, thus not much cleaning is required. The only missing data is within the '*stalk-root'* variable, there are 2480 of missing values, which would be replaced with NaN or other guessing value if needed(will be illustrated in later process).

5. Initial Findings

The chi-square test is run between the target variable(*'Class'*) and all the input variables. It turns out that the target variable has stronger association with some of the input variables, namely, the 'odor', 'spore-print-color', 'gill-color', and 'ring-type'. And the target variable has zero association with the 'veil-type' variable, for all the observations fall into one veil type, partial type. Hence the 'veil-type' won't help distinguish a mushroom's edibility.