Lucas Unterberger (01325438), Sophie Rain (01425316), Peter Stroppa (01326468) 11.10.2019
Group 20

Report Datasets

Why did those datasets catch our eye?

We were looking for datasets that we could relate to personally. Therefor food and drinks seemed suitable. Also, the classification of this datasets is something one could use in the real world. Additionally, the number of classes varies from one to the other. The wine dataset has 11 classes, whereas the mushroom dataset has only 2. What is more, is that the number of attributes is quite diverse (11 vs. 22). We know that the number of samples is similar, hence we are considering taking only the red wines into account within the wine quality dataset. Then we have 1599 and 8124 samples for wine and mushrooms, respectively. Missing values occur in the mushroom data only.

Overview

Dataset: Wine-Quality http://archive.ics.uci.edu/ml/datasets/Wine+Quality

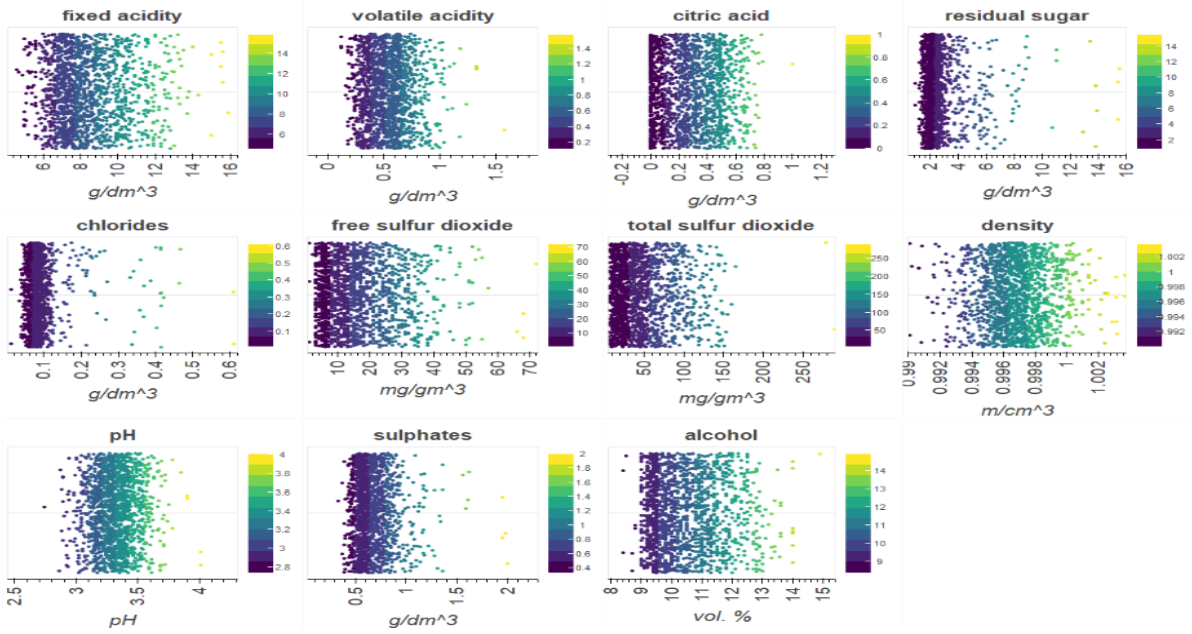
The wine quality dataset consists of 1599 red wine and 4898 white wine samples of the Portuguese “Vinho Verde” wine, as described in: http://www.vinhoverde.pt/en/ or the reference [Cortez et al., 2009]. Those samples were classified according to their quality. This classification is purely based upon sensory data obtained by the median of 3 evaluations made by wine experts. I.e. the experts were tasting the wine.

Dataset: Mushroom http://archive.ics.uci.edu/ml/datasets/Mushroom

The mushroom dataset consists of 8124 hypothetical samples of gilled mushrooms in the Agaricus and Lepiota Family. They are classified into 2 classes: edible or poisonous, samples of unknown edibility are classified as poisonous as well.

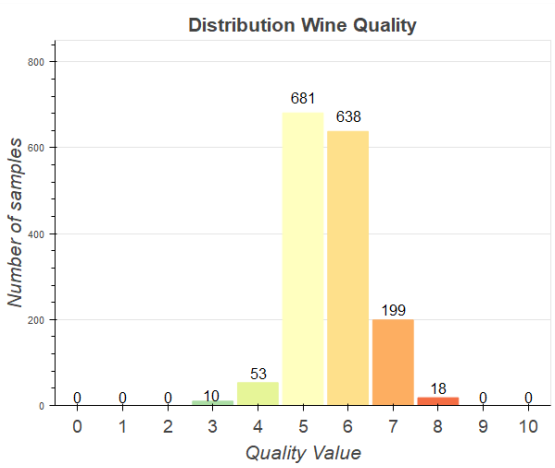
Wine Quality Dataset (Reds only):

We chose this dataset for its big number of classes (11) and its quite small number of samples (1599 for the reds only, 6497 in total). Also, we liked the kind of attributes it has. They are exclusively numeric and should technically be reliable, since they are measurable facts. 9 of the 11 attributes are



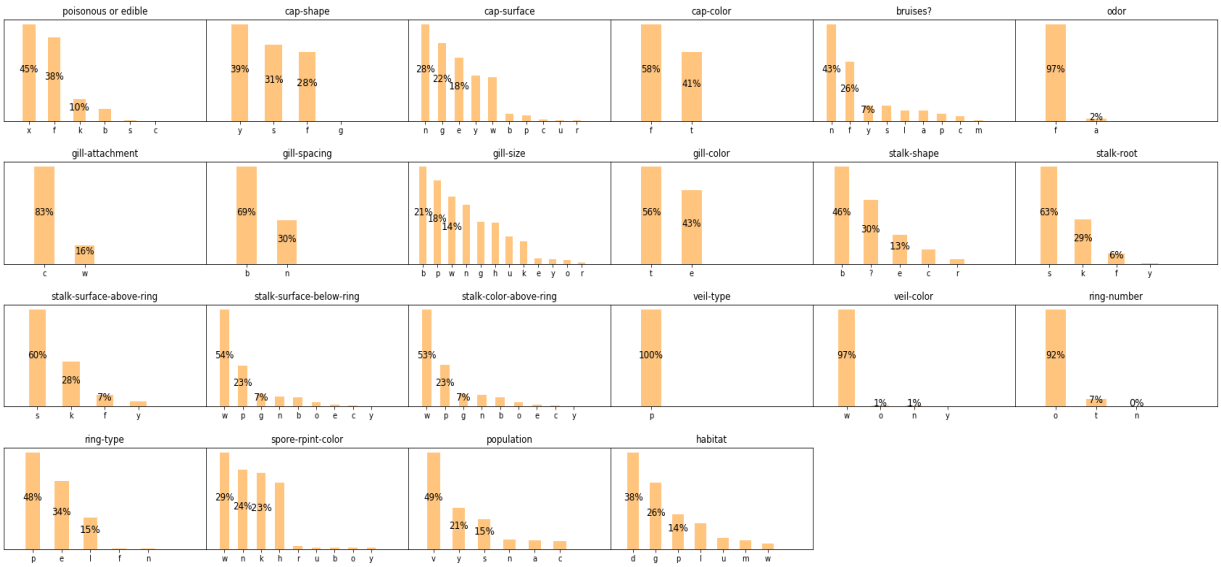
chemical values, i.e. values of the form (m)g/dm³, hence they are positive real numbers, given with at most 5 decimal places. The others are alcohol in percent (1 decimal place) and the ph-value (2 decimal places).

Both, density and the ph-value seem to have gaussian distribution. The citric acid could have a gaussian distribution with a cut-off at 0. All the others are likely to have a more or less steep Chi-square-distribution. One can also get a good intuition about mean and median values looking at the plots. It will be interesting to find out whether the outliers in the sinlge attributes tend to result in better or worse wine quality.



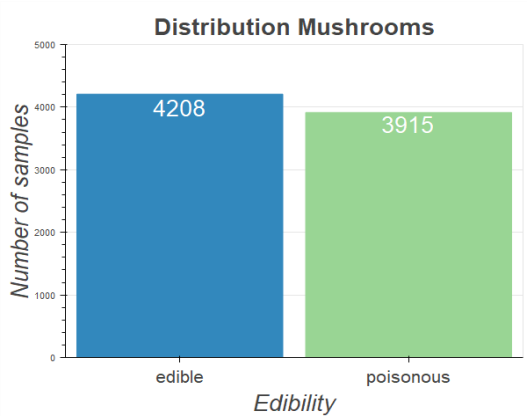
The target class ranges from 0 (bad quality) to 10 (excellent quality). The more important classes are the low ones (0-4) and the high ones (8-10), since most wines are fair and only very few are poor and even less are excellent. This can be seen easily in the plot. The plot is similar to Gaussian distribution, which strengthens our point. We have figured out two use cases. First the consumer looking for a really good wine. This person focusses on the highly rated wines. The second consumer looking for some average wine. This one wants to avoid the very poorly rated wines.

Mushroom Dataset:



The reasons for choosing this dataset were driven by the idea of finding a second dataset which is different from the first one. Each sample (8124 in total) in the mushroom dataset has 22 corresponding attributes, every single one being a nominal. The attributes have 2 to 10 different values. Moreover, it has missing values as well. The target class is either e=edible or p=poisonous

and the others are encoded as described in detail in the addendum. The kind of attributes is very different from the numeric, chemical attributes of the wine dataset. They are describing properties such as shape and odour. Which is to a certain extend subjective and hence interesting for modelling purposes. An interesting property of the attributes can be seen by looking at the plot above. There is no attribute having a similar number of samples for every value.



Since there are only 2 classes ‘edible’ and ‘poisonous’, both are of equal importance. Also, in the dataset there are similarly many edible- as poisonous-classifications. Whenever you have a new sample, you want to predict which class it belongs to, although one can argue that getting the information edible is more valuable, because you want to be completely sure before you try it. Hence the assignment of the class edible should be very conservative. Also, a possible use case is to decide edibility for mushroom collectors, who brought some from the forest themselves.

Addendum:

Attributes in detail Wine Quality:
- fixed acidity (g(tartaric acid)/dm^3)
- volatile acidity (g(tartaric acid)/dm^3)
- citric acid (g/dm^3)
- residual sugar (g/dm^3)
- chlorides (g(sodium chloride/dm^3)
- free sulfur dioxide (mg/dm^3)
- total sulfur dioxide (mg/dm^3)
- density (g/cm^3)
- pH (0-14)
- sulphates (g(potassium sulphate)/dm^3)
- alcohol (vol.%)
- output variable: quality (score between 0 and 10)

Attributes in detail Mushrooms:
-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