Rigre Garciandia
RRG190004
CS 4375.002

NOTE: Corresponding plots are in attached html file

[1]

- There are 155 entities in this dataset
- Types of each attributes
    - Class: Nominal
    - AGE: Ratio
    - SEX: Nominal
    - STEROID: Nominal
    - ANTIVIRALS: Nominal
    - FATIGUE: Nominal
    - MALAISE: Nominal
    - ANOREXIA: Nominal
    - LIVER BIG: Nominal
    - LIVER FIRM: Nominal
    - SPLEEN PALPABLE: Nominal
    - SPIDERS: Nominal
    - ASCITES: Nominal
    - VARICES: Nominal
    - BILIRUBIN: Ratio
    - ALK PHOSPHATE: Ratio
    - SGOT: Ratio
    - ALBUMIN: Ratio
    - PROTIME: Ratio
    - HISTOLOGY: Nominal
- There are missing values, but there are no duplicates
- The distribution of the bilirubin attribute seems to be skewed to the left, with most values being between 0.5 and 1. Although the age attribute ranges from 20 to 70, most of the participants are aged 30-50, with the mean being right under 40. The SGOT attribute seems to have no relation to age or gender. Most of the entities are female, which may skew the quality of the dataset. Most of the entities lived, which may also skew the quality of the dataset.
- While fatigue by itself does not suggest that the patient will die, patients with ages 40-50 with fatigue were the most likely to die. Patients who took steroids were more likely to have lower levels of Bilirubin. Female participants also had higher levels of Bilirubin in general. Participants with anorexia had higher average levels of SGOT.

[2]

It seems that the the most pronounced effect that sex had was that the bilirubin levels were higher for female records. Alk Phosphate levels seemed to be lower for female records, but the difference is smaller (to scale). Albumin levels seemed almost identical.

[3]

Looking at a summary of the distances table, we can see that the data seems to somewhat follow standard distribution. Objects 20 and 6 seem to be very close, as well as 20 and 13, all of which are female.

[4]

The amount of duplicated objects in the selected random variable varies very little from trial to trial, but it tends to stick around 80. This makes sense because 80 is also the amount of items we have in the original clean.hepatitis dataframe, which is where the random sampling is being drawn from. These numbers being so close makes sense because we picked out 500 samples out of a dataset of 80, which means that it is highly likely that each record will be picked at least twice.

[5]

PC1 and PC2 are the best choice for the new dimensions since they include the biggest proportion of the variance, while PC3 only accounts for .1880 of the total variance. Reducing these three attributes to two might increase the effectiveness of an ML model trained on this dataset.

[6]

For the frequency method, the intervals ended up being the following [20,32) [32,38.5) [38.5,49.2) [49.2,72]. Respectively with 18, 22, 20, and 20 objects in those intervals. For the interval method, the intervals ended up being the following [20,33) [33,46) [46,59) [59,72], respectively the amount of objects in each of those intervals was: 22, 32, 20, 6. We can see that with the frequency method we ended up with relatively equal amounts of objects in each interval, while with the interval method we ended up with equal width intervals.

[7]

Using the pearson correlation matrix with age and alk_phosphate of 50 randomly sampled objects, we can see that the variables do not really have a correlation since their pearson correlation is 0.057. As expected, the correlation between a variable and itself is 1.

[8]

See R script submitted alongside this assignment

[9]

A. If there are duplicate objects in the dataset and the function is only returning 0 for each of those, then those objects will be stuck at the top of the sorted distances list. This means that among the K nearest neighbors, the function will always return the objects that are identical, which is probably not what the intended result is
B. This problem could be fixed by making the distance assigned to identical objects infinity

[10]

A. Binary. Qualitative
B. Continuous. Quantitative
C. Discrete. Qualitative//people will likely give descriptive words like “Bright” or “Dim”
D. Continuous. Quantitative
E. Discrete. Qualitative
F. Continuous. Quantitative
G. Discrete. Quantitative
H. Continuous (calculated using a mathematical formula). Quantitative
I. Discrete. Qualitative
J. Discrete. Qualitative
K. Continuous. Quantitative
L. Continuous. Quantitative
M. Continuous. Quantitative