**Question 2:**

Download *the Original Pima Indians Diabetes1* dataset <https://www.kaggle.com/kumargh/pimaindiansdiabetescsv>

which has 8 numerical attributes and a binary class attribute (1 indicates that the person is assumed to have diabetes), indicating the following information:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

5 Examples in the Pima Indians Diabetes Dataset:

6,148,72,35,0,33.6,0.627,50,1

1,85,66,29,0,26.6,0.351,31,0

8,183,64,0,0,23.3,0.672,32,1

1,89,66,23,94,28.1,0.167,21,0

0,137,40,35,168,43.1,2.288,33,1

Question 2 Sub-Tasks

Apply the following exploratory data analysis techniques **using Weka or Orange** (or other tools of your preference) to the original/cleaned Pima Indian Diabetes Dataset; if not otherwise stated, use the original dataset:

1. Compute the mean value and standard deviation for attributes 2, 3, 6, and 7**.** Remove 0’s that do not make sense prior to computing these statistics. Compute the covariance matrix for four attributes next, compute the correlations for each of the 6 pairs of the 4 attributes. Interpret the statistical findings!

Remove 0’s that do not make sense prior to computing the covariance matrix and correlations.

1. Create histograms for attributes 2 and 4. Then create the same histograms for the 2 attributes for the instances of class 1 and for the instances of class 0; interpret the obtained 6 histograms.
2. Create box plots for the 6nd and 8th attribute; one for the whole dataset and one each for the instances of the two classes. Remove 0’s that do not make sense prior to computing the box plots. Interpret and compare the obtained 6 boxplots!
3. Write a conclusion (at most 13 sentences!) summarizing the most important findings of the assignment—what did we learn about the dataset? In particular, discuss the findings obtained related to predicting diabetes.
4. Calculate the row-wise Euclidean distance matrix of all attributes and plot in Matrix

Data Mining

By *Ansif Arooj*

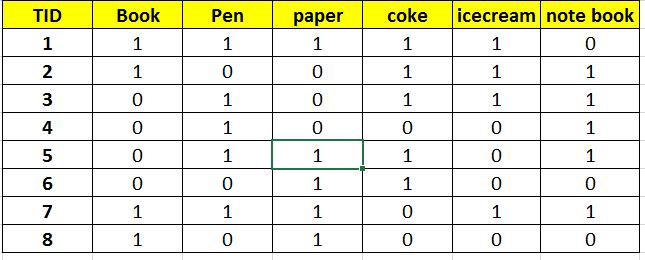
**Question 3:**

Download the **Wine, Forest Fire, and Adult datasets** from UCI Machine Learning Repository. Read the datasets’ instructions and report the following:

1. The types of the attributes (continuous [interval, ratio], categorical [nominal, ordinal]). Also identify which attribute(s) are input attribute(s) and which are class attribute(s) (if any).
2. Compute the five-number summary for the continuous attributes (you may create boxplots for this). Compute the mode for categorical attributes.
3. Compute the mean and standard deviation for the continuous attributes.
4. Generate the quantile (percentile) plots for two key attributes in each dataset.
5. Generate the histogram or distribution plot for each of the two attributes selected in (d).
6. Generate the scatter plots for the two attributes selected in (d). (in MS. Excel)
7. Compute and visualize correlation matrices for the continuous attributes.
8. Comment on the results regarding characteristics of this database. (one Paragraph)

**Question 4:**

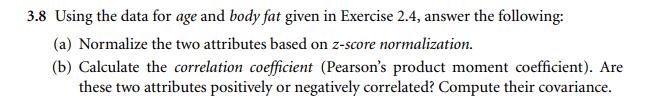
**Find the Support & Confidence according to the following rule**

****

1. {Paper,Book,Pen}{Notebook,icecream}
2. {Book,Pen}{notebook}
3. {coke}{Notebook,icecream}

**Use your Reference Book to solve these questions.**

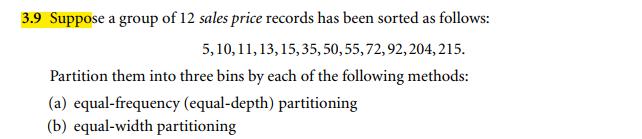
**Question 5:**

****

Data Mining

By *Ansif Arooj*

**Question 6:**

****

**Question 7:**

Consider the one-dimensional data; 10, 12, 15, 16, 20, 20, 21, 21, 22, 22, 25, 25, 25, 25, 30, 33, 34, 35, 35, 35, 36, 36, 40, 45, 46, 52, 70.

1. Use binning (by bin means and by bin boundaries) to smooth the above data, using a bin depth of
   1. Illustrate your steps, and comment on the effect of these techniques for the given data.
2. Perform the following:

Normalize the value **35** to the range [0.0, 1.0]. Compute the z-score of value 35.

Normalize the value 35 by decimal scaling.

Comment on which method you would prefer to use for the given data (assuming it is age data), giving reasons why.