**School of Computing & Informatics**

**CSC 442 Knowledge Discovery and Data Mining**

3rd March 2020

*To be submitted by 9th march 2020 or earlier*

**Task Theme area: Clustering**

***This comprises of 3 parts – they should be relatively easy to attempt***

**Part A:**

You are provided with the following dataset about some foods.

Using two different clustering Algorithms, suitable cluster the foods below using the nutrients and omitting the label.

**Questions**

1. Have you identified and outliers? If so, which ones have you identified?
2. Check for relevant attributes. Which ones are they?
3. Do your clusters agree with the labels?

**Data**

|  |
| --- |
| **NAME; Calories; Protein; Fat; Calcium; Iron; Label** |
| BEEF BRAISED;340;20;28;9;2.6;1 |
| HAMBURGER;245;21;17;9;2.7;1 |
| BEEF ROAST;420;15;39;7;2;1 |
| BEEF STEAK;375;19;32;9;2.6;1 |
| BEEF CANNED;180;22;10;17;3.7;1 |
| CHICKEN BROILED;115;20;3;8;1.4;2 |
| CHICKEN CANNED;170;25;7;12;1.5;2 |
| BEEF HEART;160;26;5;14;5.9;3 |
| LAMB LEG ROAST;265;20;20;9;2.6;1 |
| LAMB SHOULDER ROAST;300;18;25;9;2.3;1 |
| SMOKED HAM;340;20;28;9;2.5;1 |
| PORK ROAST;340;19;29;9;2.5;1 |
| PORK SIMMERED;355;19;30;9;2.4;1 |
| BEEF TONGUE;205;18;14;7;2.5;1 |
| VEAL CUTLET;185;23;9;9;2.7;1 |
| BLUEFISH BAKED;135;22;4;25;0.6;2 |
| CLAMS RAW;70;11;1;82;6;3 |
| CLAMS CANNED;45;7;1;74;5.4;3 |
| CRABMEAT CANNED;90;14;2;38;0.8;2 |
| HADDOCK FRIED;135;16;5;15;0.5;2 |
| MACKEREL BROILED;200;19;13;5;1;2 |
| MACKEREL CANNED;155;16;9;157;1.8;3 |
| PERCH FRIED;195;16;11;14;1.3;2 |
| SALMON CANNED;120;17;5;159;0.7;3 |
| SARDINES CANNED;180;22;9;367;2.5;3 |
| TUNA CANNED;170;25;7;7;1.2;2 |
| SHRIMP CANNED;110;23;1;98;2.6;3 |

**Part B:**

You have been provided with a two dimension dataset - clustering dataset 2D set 1 (xls)

1. Are there any outliers? If there are, what did you do about them?
2. Establish the optimal number of clusters for this dataset using two different clustering algorithms

**Part C:**

You have been provided with a high dimension dataset - clustering dataset HD set 2 (txt)

1. What is the dimensionality of the dataset?
2. Are there any outliers? If there are, what did you do about them?
3. Are there any irrelevant features
4. Establish the optimal number of clusters for this dataset using two different clustering algorithms