CIND 110 Data Organization for Data Analysts
Assignment 3
Michael McAllister – 501133880

Question 1.1

Apply the Apriori algorithm on this dataset.
Note that, the set of items is {Meat, Potato, Onion, Noodle, Spinach, Eggs, Salt}.
You may use 0.3 for the minimum support value.

Step 1

Trans ID	Items Purchased	Eggs	Meat	Noodle	Onion		Potato	Salt	Spinach	
2001	Meat, Potato, Onion			1		1		1		
2002	Meat, Noodle			1	1					
2003	Noodle, Spinach				1					1
2004	Meat, Potato, Onion			1		1		1		
2005	Onion, Potato, Noodle				1	1		1		
2006	Eggs, Spinach		1							1
2007	Eggs, Noodle		1		1					
2008	Meat, Potato, Salt, Onion			1		1		1	1	
2009	Salt, Spinach								1	1
2010	Meat, Potato			1				1		
Frequency			2	5	4	4		5	2	3
Support			0.2	0.5	0.4	0.4		0.5	0.2	0.3

Calculate frequency and support, min support level is 0.3 as given and add them to a list L1.

Step 3
Itemset
{Meat, Noodle}
{Meat, Onion}
{Meat, Potato}
{Meat, Spinach}
{Noodle, Onion}
{Noodle,Potato}
{Noodle, Spinach}
{Onion, Potato}
{Onion, Spinach}

Step 4
Support Calc
{Meat, Noodle} 0.125
{Meat, Onion} 0.5
{Meat, Potato} 0.6666666667
{Meat, Spinach 0
{Noodle, Onion 0.1428571429
{Noodle,Potatc 0.125
{Noodle, Spina 0.1666666667
{Onion, Potato 0.8
{Onion, Spinac 0

Step 5
L2
{Meat, Potato} 0.6666666667
{Onion, Potato} 0.8
{Meat, Onion} 0.5

Step 6
Repeat Loop
{Meat, Potato, Noodle} 0
{Meat, Potato, Onion} 0.5
{Meat, Potato, Spinach} 0
{Onion, Potato, Noodle} 0.125
{Onion, Potato, Spinach} 0
{Meat, Onion, Noodle} 0
{Meat, Onion, Spinach} 0

Step 7
L3
{Meat, Potato, Onion} 0.5

Step 8

{Meat, Potato, Onion, Noodle} 0
{Meat, Potato, Onion, Spinach} 0

Terminate Algorythm

Create a set of all permutations for possible combinations of items over the 0.3 support. Calculate the support for each set and add them to a new list L(k+1) if they're over the 0.3 support level. Check items in L2 with all remaining items from L1 (there is no need to use the original list because of antinomicity) and calculate the support. Alorythm ends with one 3 item set of {Meat, Potato, Onion}.

1.2
Show the rules that have a confidence of 0.8 or greater for an itemset containing three items.

{Meat, Potato} => Onion	sup('Meat', potatoes, 'Onion')/sup{'Meat', potatoes}	0.75	
{Potato, Onion} => Meat	sup('Meat', potatoes, 'Onion')/sup{potatoes, Onion}	0.625	
{Meat,Onion} => Potato	sup('Meat', potatoes, 'Onion')/sup{'Meat', Onion}	1	{Meat,Onion} => Potato is over .8

The rule that has a confidence over 0.8 is (Meat, Onion) => Potato.

Question 2
Assuming, that the class attribute is Profile, apply a classification algorithm to this dataset.

ID	Age	City	Gender	Education	Profile
101	20-30	NY	F	College	Employed
102	31-40	NY	F	College	Employed
103	51-60	NY	F	College	Unemployed
104	20-30	LA	M	High School	Unemployed
105	41-50	NY	F	College	Employed
106	41-50	NY	F	Graduate	Employed
107	20-30	LA	M	College	Employed
108	20-30	NY	F	High School	Unemployed
109	20-30	NY	F	College	Employed
110	51-60	SF	M	College	Unemployed

Age		Gender	
20-30	5	F	7
31-40	1	M	3
41-50	2		
51-60	2		

City		Education	
NY	7	High School	2
LA	2	College	7
SF	1	Graduate	1

Entropy | Profile | = -0.6log2|0.6|-0.4log2|0.4| 0.9709505945

Gain in Information

Age 0.9709505945

	Employed	Unemployed	
20-30	3	2	
31-40	1		All employed
41-50	2		All employed
51-60		1	Unemployed

Information gain 0.9709505945 -0.5(-0.6log2|.6|-.4log2|.4)-0.1(-1log2|1|-0log2|0|)
0.4854752972 ← Highest Information gain

City 0.9709505945

	Employed	Unemployed
NY	5	2
LA	1	1
SF		1

Information gain 0.9709505945 -0.7(-(5/7)log2|5/7|-.(2/7)log2|2/7)-0.2(-1log2|1|-1
0.1667661965

Gender 0.9709505945

	Employed	Unemployed
M	1	2
F	5	2

Information gain 0.9709505945 -0.3(1/3log2|1/3|-2/3log2|2/3|)-0.7(5/7log2|1|-2/7l
0.0912774462

Education 0.9709505945

	Employed	Unemployed
Highschool	0	2
College	5	2
Graduate	1	

Information gain 0.9709505945 -0.2(-0log2|0|-.1log2|1|)-0.7(5/7log2|5/7|-2/7log2|
0.3667661965

The highest information gain was age, this became the root node of the decision tree.

Step 2.

Age Node

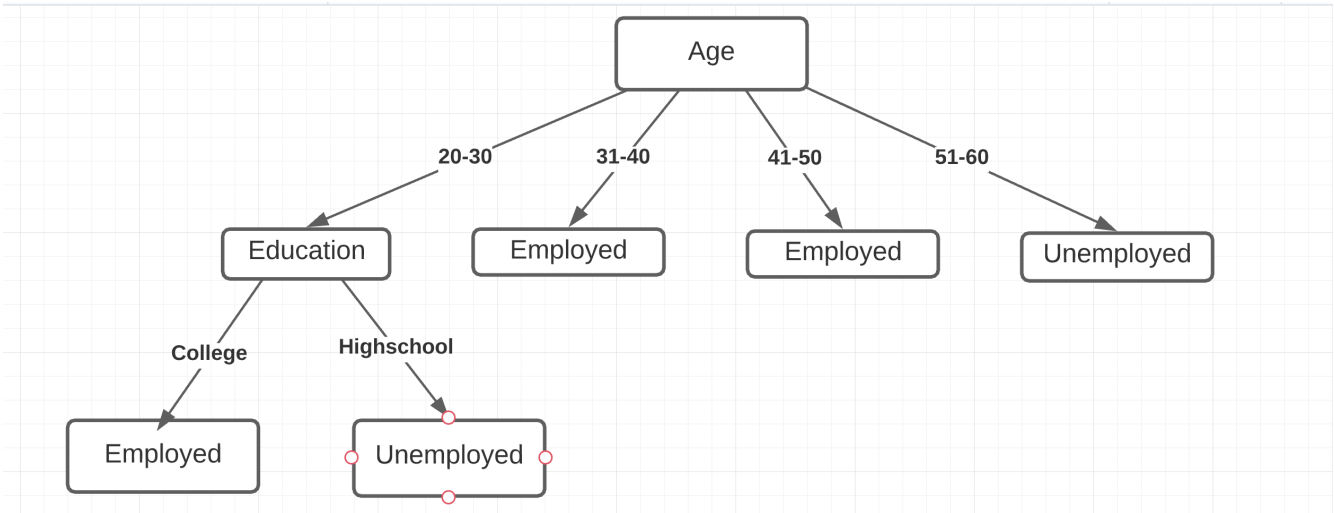
Age	20-30	Entropy	-0.6log2|0.6|-0.4log2|0.4|	0.9709505945

City		0.9709505945	
	Employed	Unemployed	
NY	2	1	
LA	1	1	
Information gain	0.9709505945	-3/5(-2/3log2|2/3|-1/3log2|1/3|)-2/5(1/2log2|1/2|-1	
	0.419973094		

Gender		0.9709505945	
	Employed	Unemployed	
F	2	1	
M	1	1	
Information gain	0.9709505945	-3/5(-2/3log2|2/3|-1/3log2|1/3|)-2/5(1/2log2|1/2|-1	
	0.419973094		

Education		0.9709505945	
	Employed	Unemployed	
Highschool	0	2	
College	3	0	
Information gain	0.9709505945	-3/5(-2/3log2|2/3|-1/3log2|1/3|)-2/5(1/2log2|1/2|-1	
	0.9709505945	**← Highest Information gain**	

The highest information gain was then education, and the algorythm ends here and no further nodes are needed. The final decision tree is:



Question 3.

3.1

Use the K-means algorithm to cluster this dataset. You can use a value of 2 for K and can assume that the records with RIDs 103, and 104 are used for the initial cluster centroids.

Question 3

RID	Age	Years of Service
101	30	5
102	50	25
103	50	15
104	25	5
105	30	10
106	55	25

Iteration 1

RID	Age	Years of Service	Cluster 1 Distance (103)	Cluster 2 Distance (104)	Cluster Allo		
101	30	5	22.360679775	5	2	x	51.666666667 28.333333333
102	50	25	10	32.0156211872	1	y	21.666666667 6.6666666667
103	50	15	0	26.9258240357	1		
104	25	5	26.9258240357	0	2		
105	30	10	20.6155281281	7.0710678119	2		
106	55	25	11.1803398875	36.0555127546	1		

Iteration 1

RID	Age	Years of Service	Cluster 1 Distance	Cluster 2 Distance	Cluster		
101	30	5	27.3353657781	2.357022604	2	x	51.666666667 28.333333333
102	50	25	3.7267799625	28.3823106099	1	y	21.666666667 6.6666666667
103	50	15	6.8718427094	23.213980462	1		
104	25	5	31.4466037735	3.7267799625	2		
105	30	10	24.6080384337	3.7267799625	2		
106	55	25	4.7140452079	32.3608130649	1		

Final iteratation, no change to mean or cluster allocation

3.2

What is the difference between describing discovered knowledge using clustering and describing it using classification?

Classification and clustering are two methods of pattern identification used in machine learning. Although both techniques have certain similarities, the difference lies in the fact that classification uses predefined classes in which objects are assigned, while clustering identifies similarities between objects, which it groups according to those characteristics in common and which differentiate them from other groups of objects. These groups are known as "clusters". Classification involves supervised learning allocates into already defined classes and clusters involves unsurpervised learning and relies on similarities between data items.