1 (a) Business Problem

Healthcare professionals have a problem as to how to identify patients to detect chronic kidney disease (CKD) due to two reasons: (i) the disease does not present symptoms in its early stage but left undetected, could lead to a gradual loss of kidney function; and (ii) regular screening for CKD in patients who do not show symptoms or risk factors is also not recommended.

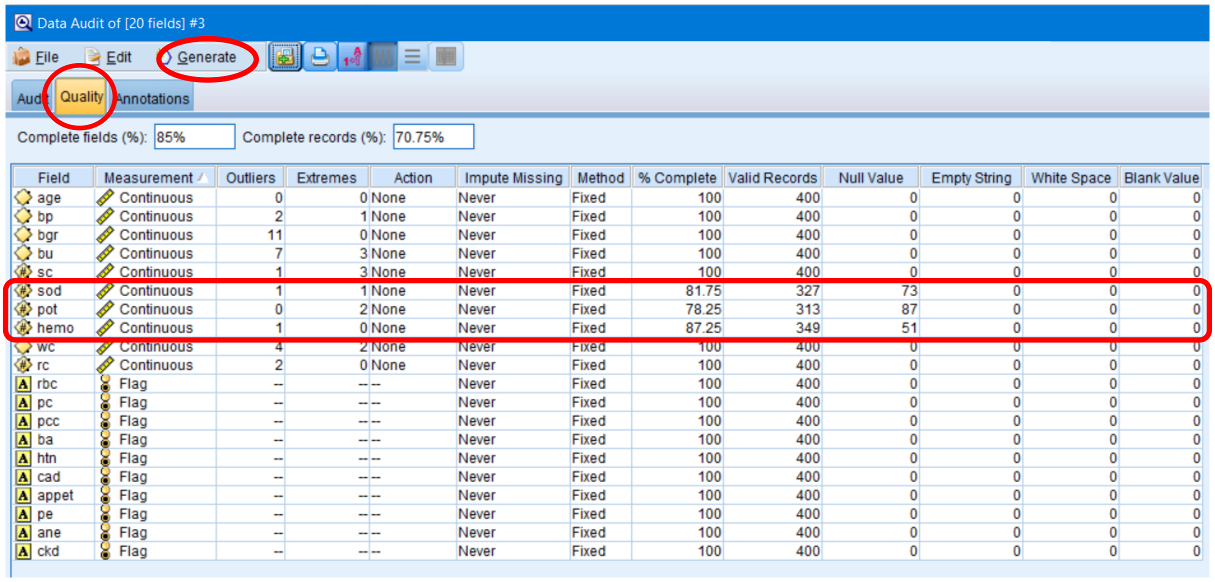
The healthcare organisation needs to be able to derive patients' profiles that could help doctors identify patients who are at risk of developing CKD. Once they can identify this group of at-risk patients, they can then perform screening tests to detect CKD. This will help to facilitate treatment at an earlier stage of the disease.

Data Mining Objective

Using the K-means clustering method, the data mining objective is to use segmentation analysis applied on patients' demographics (age) and various health screening-tests measures to help identify groups of patients who have high risk of being diagnosed with CKD. The screening-tests measures to be used are mentioned in the following sections of this paper.

1(b) Data Preparation

Figure 1: Table of % of valid records for each field generated from Data Audit Node

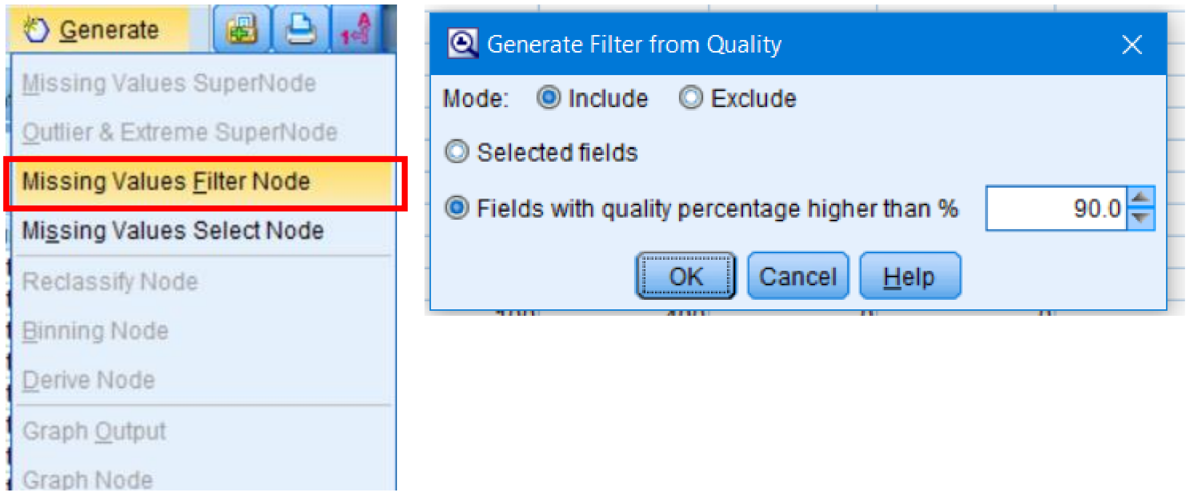


Complete fields (%): 85% Complete records (%): 70.75%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
age	Continuous	0	0	None	Never	Fixed	100	400	0	0	0	0
bp	Continuous	2	1	None	Never	Fixed	100	400	0	0	0	0
bgr	Continuous	11	0	None	Never	Fixed	100	400	0	0	0	0
bu	Continuous	7	3	None	Never	Fixed	100	400	0	0	0	0
sc	Continuous	1	3	None	Never	Fixed	100	400	0	0	0	0
sod	Continuous	1	1	None	Never	Fixed	81.75	327	73	0	0	0
pot	Continuous	0	2	None	Never	Fixed	78.25	313	87	0	0	0
hemo	Continuous	1	0	None	Never	Fixed	87.25	349	51	0	0	0
wc	Continuous	4	2	None	Never	Fixed	100	400	0	0	0	0
rc	Continuous	2	0	None	Never	Fixed	100	400	0	0	0	0
rbc	Flag	--	--	--	Never	Fixed	100	400	0	0	0	0
pc	Flag	--	--	--	Never	Fixed	100	400	0	0	0	0
pcc	Flag	--	--	--	Never	Fixed	100	400	0	0	0	0
ba	Flag	--	--	--	Never	Fixed	100	400	0	0	0	0
htn	Flag	--	--	--	Never	Fixed	100	400	0	0	0	0
cad	Flag	--	--	--	Never	Fixed	100	400	0	0	0	0
appet	Flag	--	--	--	Never	Fixed	100	400	0	0	0	0
pe	Flag	--	--	--	Never	Fixed	100	400	0	0	0	0
ane	Flag	--	--	--	Never	Fixed	100	400	0	0	0	0
ckd	Flag	--	--	--	Never	Fixed	100	400	0	0	0	0

To remove any field if 10% of its records are invalid (in red box) in Figure 1, click on the Quality tab, then the Generate tab to select "Missing Value Filter Node". From the "Generate Filter from Quality" Node, click the radio button "include". In the "Fields with quality percentage higher than %", key in "90" (see Figure 2).

Figure 2: Generate Filter from Quality Tab



Generate Filter from Quality

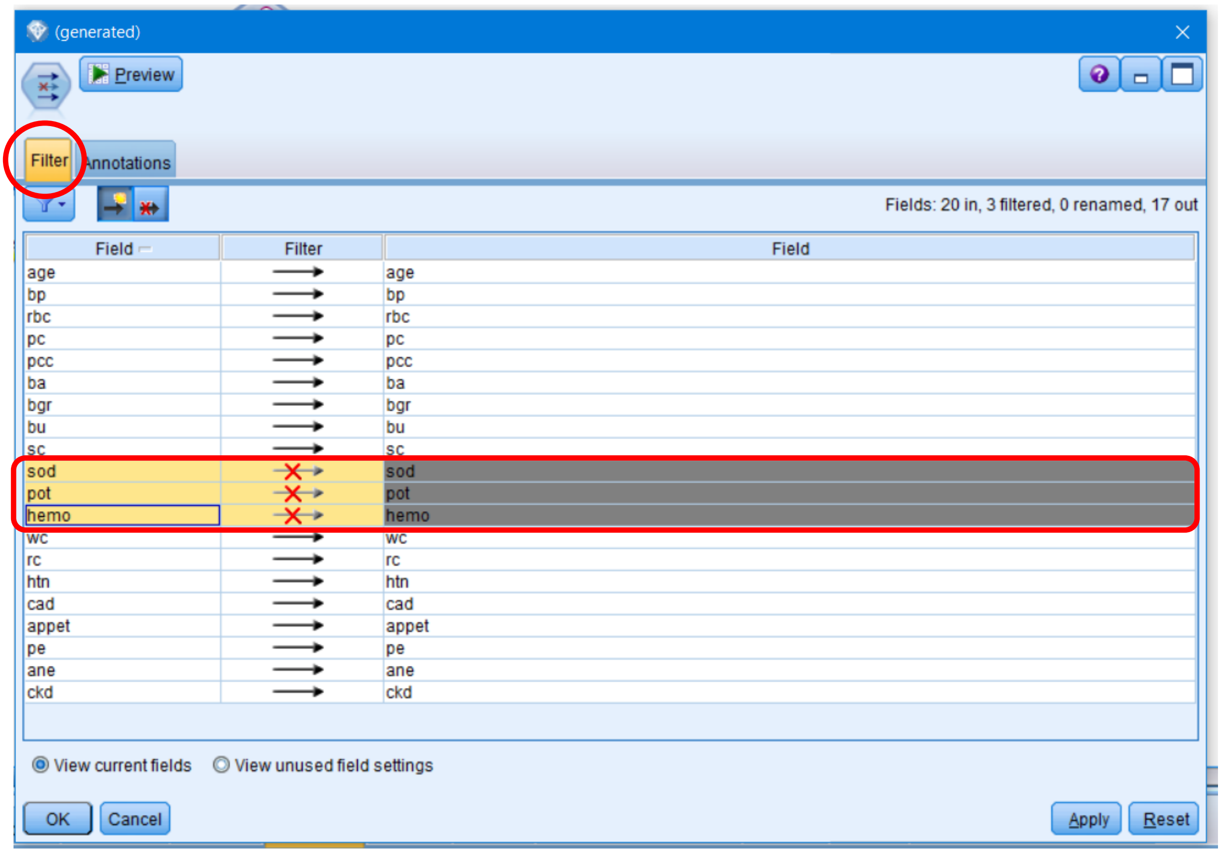
Mode: ● Include ○ Exclude

○ Selected fields

● Fields with quality percentage higher than % 90.0

Connecting the generated node to the source node and click “Run”, we have the following Figure 3 showing the “Filter Tab” with invalid fields removed.

Figure 3: Filter Tab Showing Invalid Fields Removed



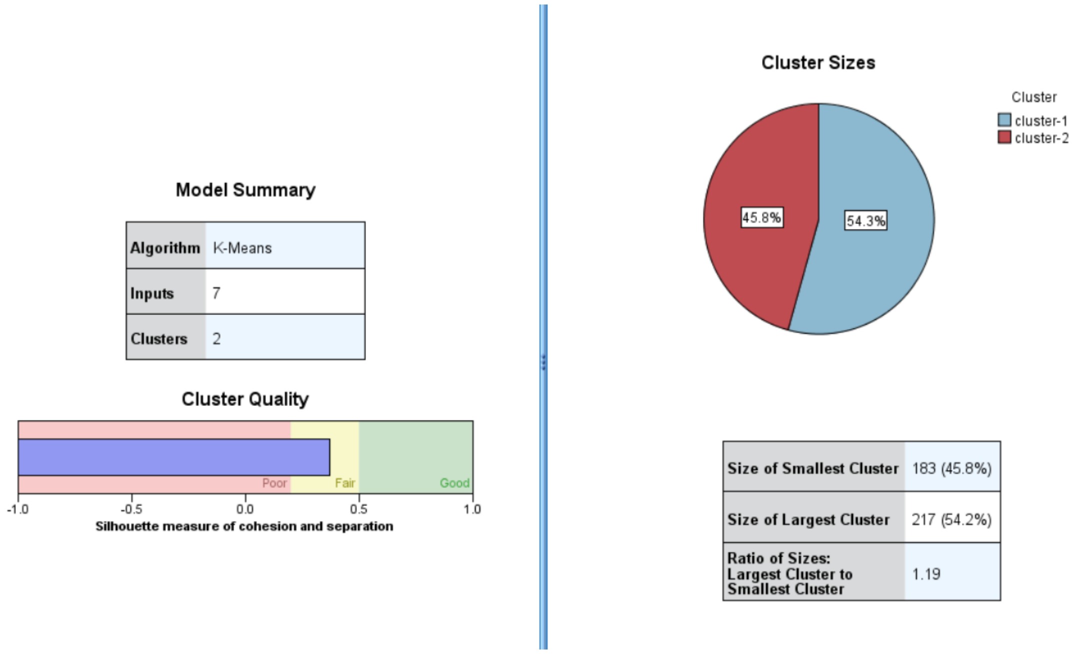
Fields: 20 in, 3 filtered, 0 renamed, 17 out

Field	Filter	Field
age	→	age
bp	→	bp
rbc	→	rbc
pc	→	pc
pcc	→	pcc
ba	→	ba
bgr	→	bgr
bu	→	bu
sc	→	sc
sod	✗	sod
pot	✗	pot
hemo	✗	hemo
wc	→	wc
rc	→	rc
htn	→	htn
cad	→	cad
appet	→	appet
pe	→	pe
ane	→	ane
ckd	→	ckd

● View current fields ○ View unused field settings

1(c) Cluster Model A (K = 2)

Figure 4A: Cluster Summary (K =2)



Model Summary

Algorithm	K-Means
Inputs	7
Clusters	2

Size of Smallest Cluster	183 (45.8%)
Size of Largest Cluster	217 (54.2%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	1.19

Figure 4B: Cluster Comparison (K = 2)

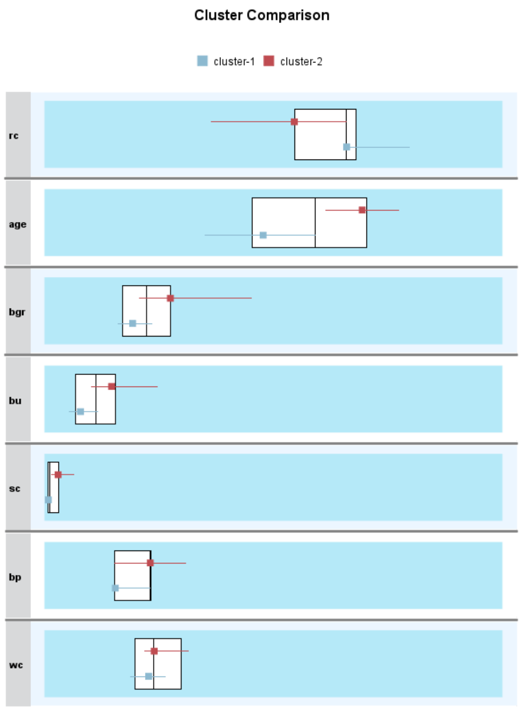
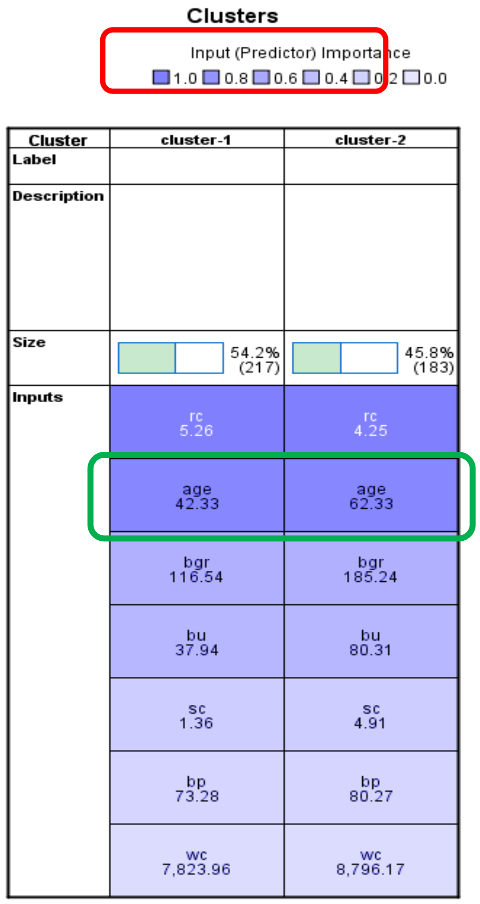


Figure 4C: Cluster Profile (K =2)



Cluster	cluster-1	cluster-2
Label		
Description		
Size	54.2% (217)	45.8% (183)
Inputs	rc 5.26	rc 4.25
	age 42.33	age 62.33
	bgr 116.54	bgr 185.24
	bu 37.94	bu 80.31
	sc 1.36	sc 4.91
	bp 73.28	bp 80.27
	wc 7,823.96	wc 8,796.17

Profile of Clusters 1 and 2

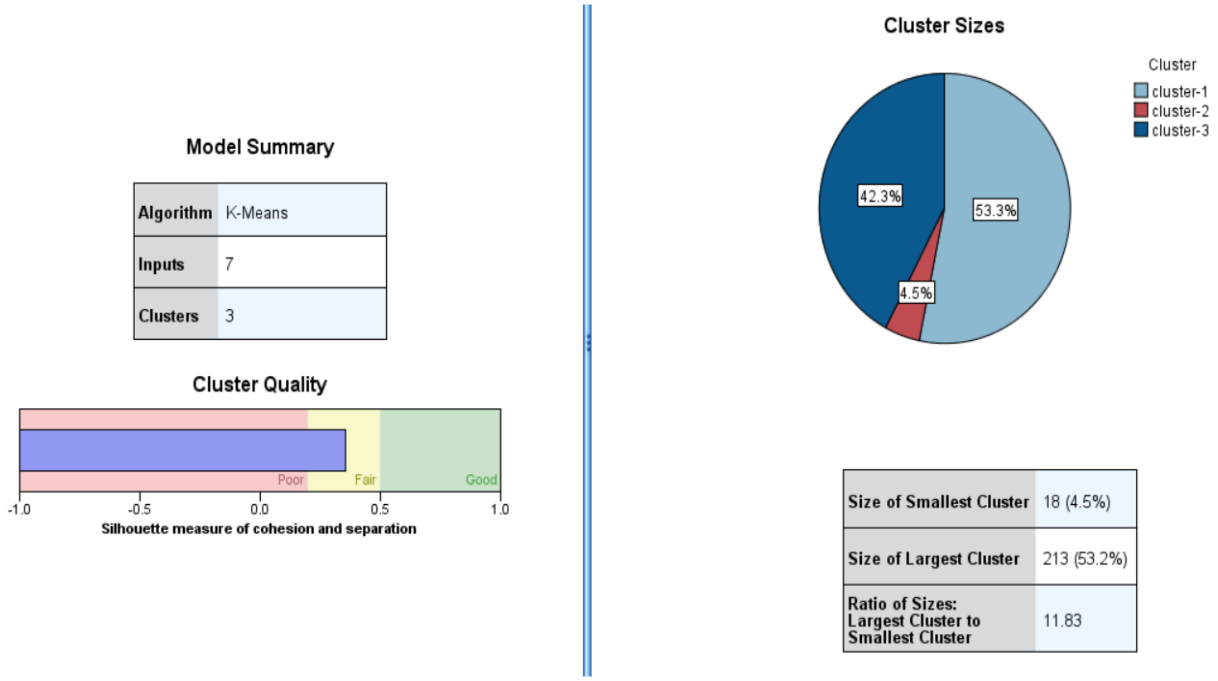
Cluster 1 is the larger cluster of the two, comprising 54.2% of the dataset. From the cluster results shown in Figures 4B and 4C, Cluster 1 comprises a relatively younger group of patients of mean age 42 years, with all their screening test results scoring lower than patients in Cluster 2, except for the red blood cell count (rc).

The results show that older patients of mean age 62 (mostly in Cluster 2) registered lower red blood cell count (rc) than younger patients, but higher in the other measures such as blood glucose random (bgr), blood urea (bu), serum creatinine (sc), blood pressure (bp) and white blood cell count (wc).

In terms of predictor importance (measures in darker shades of colour in Figure 4C), the four most important measures, in order of importance, are red blood cell count (rc), age, blood glucose random (bgr), and blood urea (bu). White blood cell count (wc) ranked the least important.

1(c) Cluster Model B (K = 3)

Figure 5A: Cluster Summary (K =3)



Model Summary

Algorithm	K-Means
Inputs	7
Clusters	3

Size of Smallest Cluster	18 (4.5%)
Size of Largest Cluster	213 (53.2%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	11.83

Figure 5B: Cluster Comparison (K = 3)

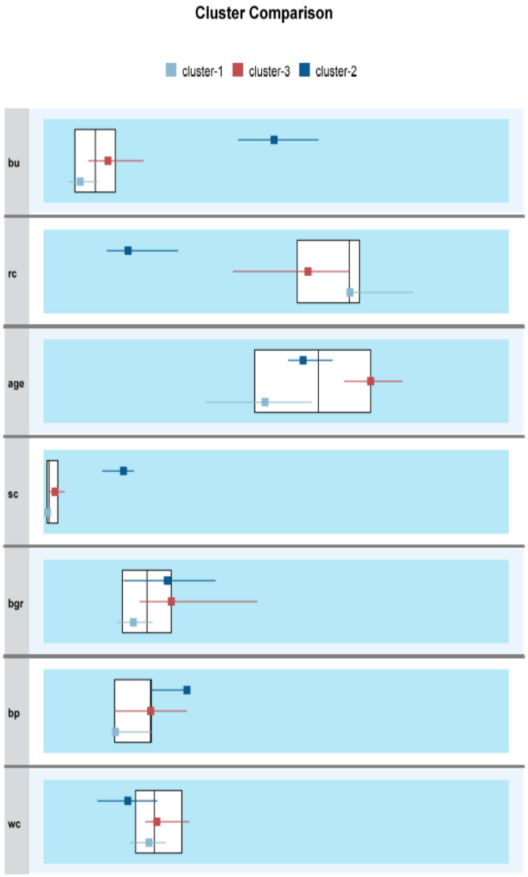
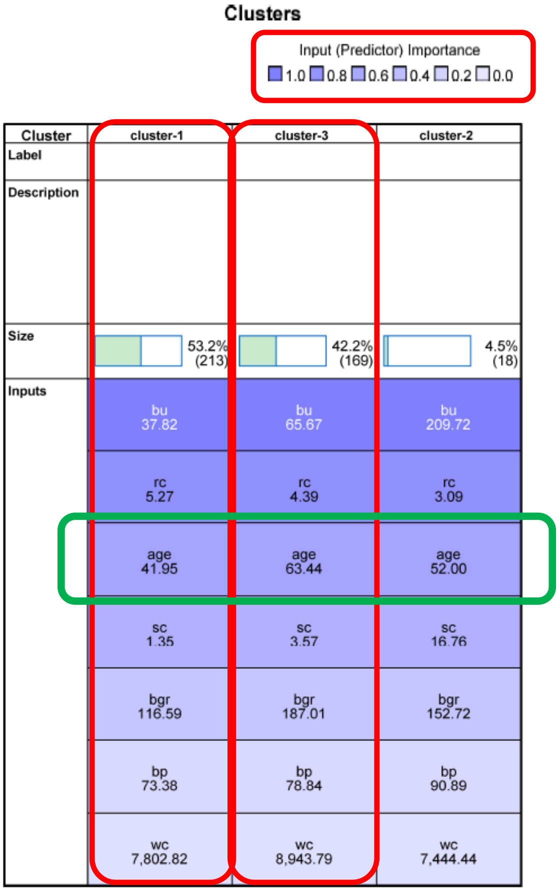


Figure 5C: Cluster Profile (K =3)


Clusters

Input (Predictor) Importance
▇1.0 ▇0.8 ▇0.6 ▇0.4 ▇0.2 □0.0

Cluster	cluster-1	cluster-3	cluster-2
Label			
Description			
Size	53.2% (213)	42.2% (169)	4.5% (18)
Inputs	bu 37.82	bu 65.67	bu 209.72
	rc 5.27	rc 4.39	rc 3.09
	age 41.95	age 63.44	age 52.00
	sc 1.35	sc 3.57	sc 16.76
	bgr 116.59	bgr 187.01	bgr 152.72
	bp 73.38	bp 78.84	bp 90.89
	wc 7,802.82	wc 8,943.79	wc 7,444.44

Profile of Clusters 1, 2 and 3

The clusters comprise three age groups of patients:

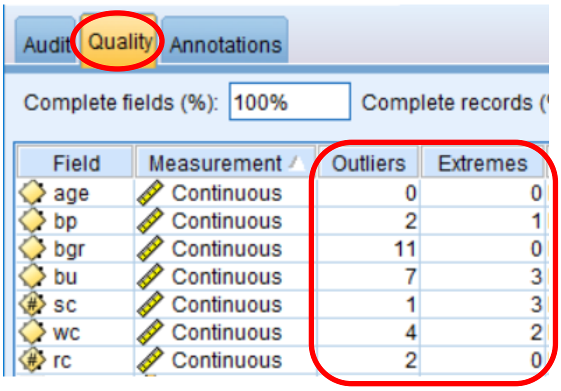
Table 1: Model B - Cluster Age Group

Cluster 1 (53%)	Cluster 2 (4.5%)	Cluster 3 (42%)
Grouped as Younger patients Mean Age: 42 years	Grouped as Mid-age patients Mean Age: 52 years	Grouped as Older patients Mean Age: 63 years

Results of both Clusters 1 and 3 (see summary in Figure 5C) resemble very closely the profile of patients identified in the two clusters in Model A: Older patients (in Cluster 3) have a lower red blood cell count (rc) than younger patients (in Cluster 1) but scored higher in the rest of the screening-test measures. The cluster sizes of both models are also very similar.

Cluster 2, who are the mid-age group of patients (mean 52 years), interestingly, do not show similar results as Clusters 1 and 3. The results of patients in this cluster tend to fall very far from the mean measure, showing extreme high scores (outside the mean) in blood urea (bu), serum creatinine (sc) and blood pressure (bp) and low scores in red blood cell count (rc) and white blood cell count (wc). This Cluster appeared to have captured records of outliers (more than 3 standard deviations from the mean) and extreme values (more than 5 standard deviations from the mean) which may skew the overall results. The quality tab of the data audit node confirms the existence of outlier and extreme records (see Figure 5D below). Cluster 2 has also the smallest cluster size of 18 records, or 4.5% of dataset, which is not ideal. Cluster sizes should have a minimum of 5% to 10% of the dataset size for the results to be meaningful. These records should either be removed or normalised to derive better cluster results.

Figure 5D: Outliers and Extremes Records from Data Audit Node



Complete fields (%): 100% Complete records ('

Field	Measurement	Outliers	Extremes
age	Continuous	0	0
bp	Continuous	2	1
bgr	Continuous	11	0
bu	Continuous	7	3
sc	Continuous	1	3
wc	Continuous	4	2
rc	Continuous	2	0

In terms of predictor importance, as shown in Cluster Summary in Figure 5C (measures in darker shades of colour), the four most important measures, in order of importance, are blood urea (bu), red blood cell count (rc), age and serum creatinine (sc). White blood cell count (wc) is the least important.

1(d) Comparing Model A and Model B

Both models in (c) have similar cluster quality, with the silhouette measure of cohesion and separation in the "fair" range (see Figures 4A and 5A). Both also identified red blood count (rc) and age as among the top three important predictors, though not in the same order, and blood pressure (bp) and white blood cell count (wc) were the least important predictors.

Table 2: Ranking of Predictor Importance

	Ranking of Predictor Importance ["1" Highest and "7" Lowest]	
	Model A (K = 2)	**Model B (K = 3)**
Blood Urea (bu)	**4**	**1**
Red Blood Cell Count (rc)	**1**	**2**
Age	**2**	**3**
Serum Creatinine (sc)	5	4
Blood Glucose Random (bgr)	**3**	**5**
Blood Pressure (bp)	6	6
White Blood Cell Count (wc)	7	7

The two models also have similar mean screening-test results as seen in Table 3:

Table 3: Summary of Mean of Screening-Test Results for Models A & B by Cluster

	Mean of screening-test results				
Inputs	**Model A -** Cluster 1 (Younger)	**Model B –** Cluster 1 (Younger)	**Model A –** Cluster 2 (Older Age)	**Model B –** Cluster 3 (Older Age)	**Model B –** Cluster 2 (Mid-age)
rc (mlns/cmm)	5.26	5.27	4.25	4.39	3.09
age (years)	42.33	41.95	62.33	63.44	52.00
bgr (mgs/dl	116.54	116.59	185.24	187.01	152.72
bu (mgs/dl)	37.94	37.82	80.31	65.67	209.72
sc (mgs/dl)	1.36	1.35	4.91	3.57	16.76
bp (mm/Hg)	73.28	73.38	80.27	78.84	90.89
wc cells/cumm	7,823.96	7,802.82	8,796.17	8,943.79	7,444.44

Clusters 1 of both Model A and Model B contain similar profile of younger patients with mean age of 42 years while Clusters 2 and 3 of Model A and Model B respectively profile older patients of mean age 63 years, with most of the mean screening-test results with fairly close numbers, except for blood urea (bu) and serum creatinine (sc).

In conclusion, we are of the view that both models are just as important as they help to confirm the results that:

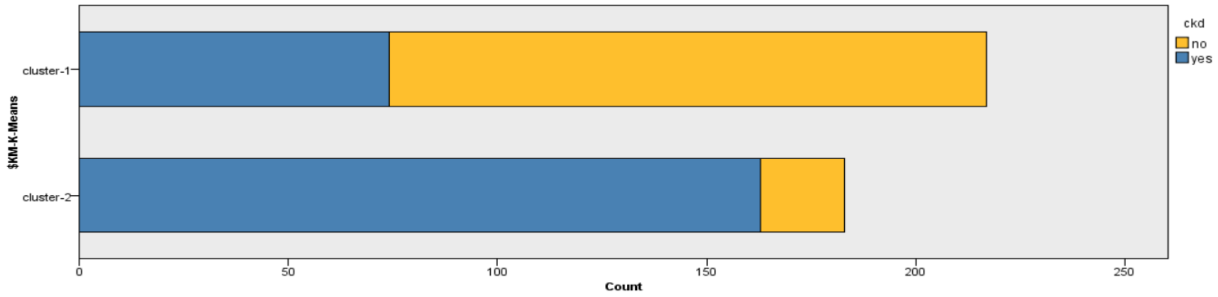
(i) older people have lower red blood cell count and higher readings in all the other screening-test measures; and

(ii) Age and red blood cell count are consistently the more important predictors than the other measures.

For this exercise, as both models generated fairly similar results, the preference will be to use Model A (with 2 clusters) as it is easier to work with fewer clusters.

(e) Using Model A’s results, we use the Graph Node to generate the distribution chart below.

Figure 6: Distribution of Patients with CKD in Clusters 1 and 2 (Model A: K=2)



The distribution chart confirms the findings of part (d), which is that older patients as profiled in Cluster 2 are likely to suffer from CKD, compared to younger patients profiled in Cluster 1. Older patients in Cluster 2 form a higher proportion of patients with CKD than younger patients in Cluster 1. We can summarise the results from part (c) and (e) that: Older patients of mean age 62 years whose test results showed low red blood cell count (rc), high blood urea (bu) and blood glucose random (bgr) are at higher risk of being diagnosed with CKD, meeting the data mining objective defined in part (a).

(f) Deployment plan of data mining results

In deployment planning, the healthcare organisation makes use of the data mining results to make decisions related to daily business operations. For the deployment plan to succeed, the organisation must appoint a lead medical staff who is trained and able to supervise the plan, supported by a team of medical staff. Each staff member must be trained and briefed on his/her role and tasks in the plan. At the start of the plan, the team must ensure it has adequate resources, such as financial, manpower, medical supplies and technical support services (for example, laboratory services).

The deployment plan will involve:

(a) Identifying patients to perform screen test. Here, they may choose patients from the older age group (62 years and older) to test as the model indicated this group at higher risk of CKD.

(b) Decide on the screening tests to perform. In the model, we have identified three screening tests which are strong predictors for CKD. They are screening for low red blood cell count, high blood urea and high blood glucose random. The other screening tests (for high blood pressure, serum creatinine and white blood cell count) may need to be carried out if the first three test results are negative.

(c) Obtaining patients’ consent for the screening tests.

(d) Scheduling patients for screening tests.

(e) Perform the screening tests on patients by trained medical staff.

(f) Record the results. There should be proper systems and procedures to record results, ideally one that is integrated with the patient database system that contains patients’ personal details and demographics, health profile and any existing medical conditions. If the systems are integrated, it will be easier for the organisation to evaluate the results to check the model accuracy.

(g) Evaluation of test results. If one or more test results indicate patients are at risk, the hospital should test for CKD to confirm if the relationship is positive, as predicted by the model. If patient tests positive for CKD, they could then be referred to the correct healthcare unit for treatment.

(h) For patients whose screening-test results do not show them at risk of CKD, there should be procedures to schedule them for periodic testing , as symptoms may appear later.

(i) A metrics should be maintained to observe the success (i.e. accuracy) rate so that the data used to build the model and its parameters can be validated for continuing use by the health organization.

There must be continuous tracking and monitoring of model performance to provide assurance to the health organization that the model is fit for its purpose. If validation results fail, or if there are changes such as emergence of new screening tests, symptoms and medical research findings on CKD, the health organization should recalibrate parameters, reassess assumptions or tweak the model design. Data mining is an iterative and interactive process and may have to move between and among the various stages to achieve better results. Monitoring model performance will provide information to the healthcare organization if the model is meeting its business objectives.

Finally, the project team should write a report to document the project from start to finish. This would include a summary, the key take-away from the project and the data mining results.

Q2(a) **Data Mining Objective**

The data mining objective is to use Apriori algorithm to determine if one or more health symptoms present below co-occur frequently among patients with CKD, which may help doctors correlate existence of symptoms with the existence of CKD. Translating it into an association rule:

Presence of one or more Symptoms [Antecedent]	→	Existence of CKD [Consequent]

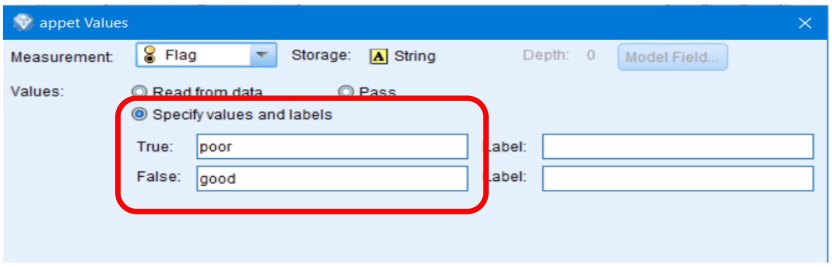
The symptoms recorded here are the presence of :

1	abnormal red blood cells (rbc)	2	Abnormal pus cells (pc)	3	Pus cell clumps (pcc)
4	Coronary artery disease (cad)	5	Bacteria (ba)	6	Hypertension (htn)
7	Poor appetites (appet)	8	Pedal edema (pe)	9	Anemia (ane)

Q2(b) **Values of Flag Fields**

All the flag fields have been specified as shown in the field settings in the Source Node in **Figure 8**, with the “True Value” shown in the red box under the column “Values”. If the specification is incorrect, we can double-click on the relevant flagged symptom and change values by selecting “Specify values and labels” and change the description under “True ” and “False” fields, shown in example below [for the symptom “appetite” (appet)].

Figure 7: Flag Fields



appet Values

Measurement: Flag Storage: String Depth: 0 Model Field...

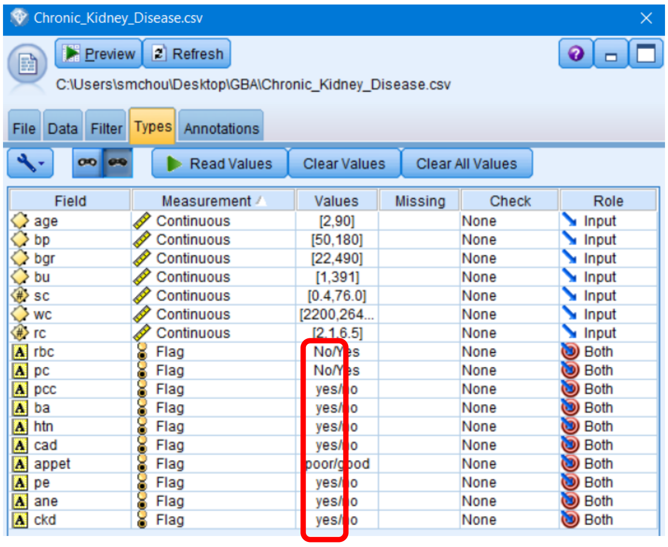
Values: ○ Read from data ○ Pass

● Specify values and labels

True: poor Label:

False: good Label:

2(c) `Figure 8: Field settings of the Source Node:



Chronic_Kidney_Disease.csv

Preview Refresh

C:\Users\smchou\Desktop\GBA\Chronic_Kidney_Disease.csv

File | Data | Filter | Types | Annotations

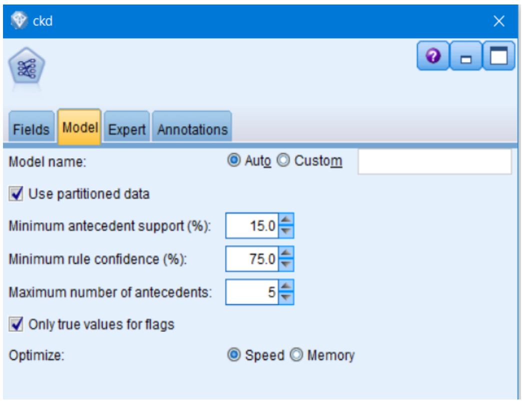
Read Values Clear Values Clear All Values

Field	Measurement	Values	Missing	Check	Role
age	Continuous	[2,90]		None	Input
bp	Continuous	[50,180]		None	Input
bgr	Continuous	[22,490]		None	Input
bu	Continuous	[1,391]		None	Input
sc	Continuous	[0.4,76.0]		None	Input
wc	Continuous	[2200,264...		None	Input
rc	Continuous	[2.1,6.5]		None	Input
rbc	Flag	No/Yes		None	Both
pc	Flag	No/Yes		None	Both
pcc	Flag	yes/no		None	Both
ba	Flag	yes/no		None	Both
htn	Flag	yes/no		None	Both
cad	Flag	yes/no		None	Both
appet	Flag	poor/good		None	Both
pe	Flag	yes/no		None	Both
ane	Flag	yes/no		None	Both
ckd	Flag	yes/no		None	Both

True Values of Flag Fields

Q2(d) Below are the settings in the Model Node:

Figure 9: Settings in Model Node



ckd

Fields | Model | Expert | Annotations

Model name: ● Auto ○ Custom

☑ Use partitioned data

Minimum antecedent support (%): 15.0

Minimum rule confidence (%): 75.0

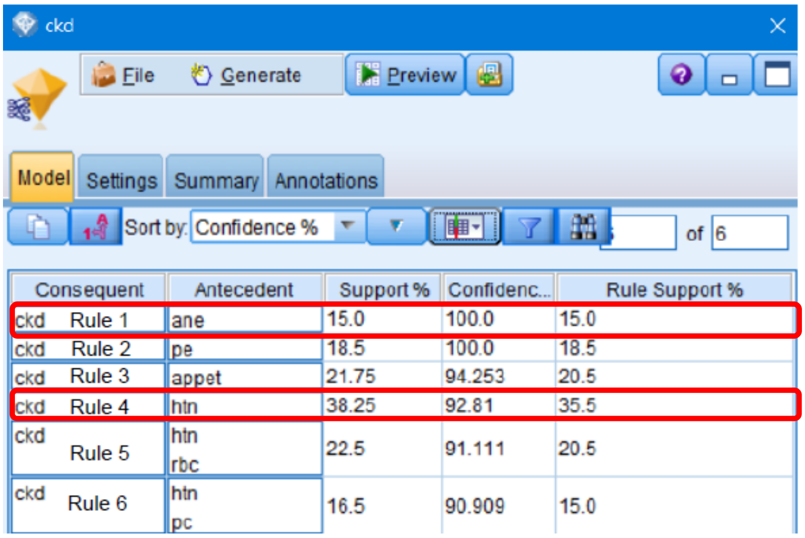
Maximum number of antecedents: 5

☑ Only true values for flags

Optimize: ● Speed ○ Memory

2(e) In Figure 10 below, there are 6 association rules generated with “ckd”:

Figure 10: Association Rules Generated



Consequent		Antecedent	Support %	Confidenc...	Rule Support %
ckd	Rule 1	ane	15.0	100.0	15.0
ckd	Rule 2	pe	18.5	100.0	18.5
ckd	Rule 3	appet	21.75	94.253	20.5
ckd	Rule 4	htn	38.25	92.81	35.5
ckd	Rule 5	htn rbc	22.5	91.111	20.5
ckd	Rule 6	htn pc	16.5	90.909	15.0

Two association rules with “ckd” are of interest here:

Rule 4:	Presence of hypertension (htn)	→	Existence of ckd
Rule 1:	Presence of anemia (ane)	→	Existence of ckd

Rule 4 is of interest to us here because it showed hypertension is a significant indicator of CKD, As shown in Figure 10, this rule has:

(i) a strong confidence of 93%, implying a high likelihood CKD occurs when hypertension is present; the highest support of 38%, which is the frequency hypertension occurs in the dataset and the highest rule support of 35.5%, which is the percentage of times hypertension and CKD existed together in the dataset.

(ii) hypertension as a symptom appears in half of the rules generated. Besides Rule 4, it co-occurred with abnormal red blood cells (rbc) in Rule 5, and in Rule 6 with abnormal pus cells (pc). Both Rules 5 and 6 have high confidence of over 90%.

Interestingly, hypertension, which is a condition of high blood pressure based on common medical knowledge (https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410), is also one of the predictors of CKD from our cluster analysis earlier. It ranked the 6th most important predictor in Models A and B (refer to Table 2). The cluster analysis also identified older patients (62 years and older) tend to have higher blood pressure (bp) than younger patients.

Rule 1 is of interest as it has a confidence of 100%, implying the likelihood of CKD is certain when anemia is present. Although the support of 15% and rule support of 15% are not as high as hypertension, doctors should not ignore this symptom because of the high confidence. Anemia is a condition of deficiency of red blood cells or haemoglobin that carry oxygen in the human body (https://www.webmd.com/a-z-guides/understanding-anemia-basics). It relates back to the cluster analysis results which identified older patients tend to have low red blood cell count (rc), one of the strongest predictors of CKD (see table 2). Rule 5 which shows abnormal red blood cells (rbc) co-occurring with hypertension lends further support to anemia being an important symptom of CKD.

We can summarise the key findings derived from cluster and association analysis in Table 4 below, which will help the healthcare organisation identify profile of patients at risk of CKD, meeting both data mining objectives in 1(a) and 2(a). Healthcare professionals can now solve the problem of deciding which group of patients to screen to diagnose CKD from the findings below:

Table 4: Profile of Patients at Risk of CKD

Presence of Symptoms (Derived from association analysis)	**Characteristics of Screen-Test Results (Derived from Cluster Analysis)**
Hypertension (Supported by Rules 4, 5 & 6)	High blood pressure
Anemia (Supported by Rule 1) Abnormal red blood cells (in Rule 5, lends support to Rule 1)	Low red blood cell count
	Patient Age Group: 62 years and above

References

Dr James Tan Swee Chuan, A/P Lee Pui Mun, Prof Koh Hian Chye, Dr Wang Deliang (2019) Singapore University of Social Sciences, School of Business, ANL303 Fundamentals of Data Mining Study Guide.

Website of Mayo Clinic, Weblink: https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410

Website of Webmd, Weblink: https://www.webmd.com/a-z-guides/understanding-anemia-basics