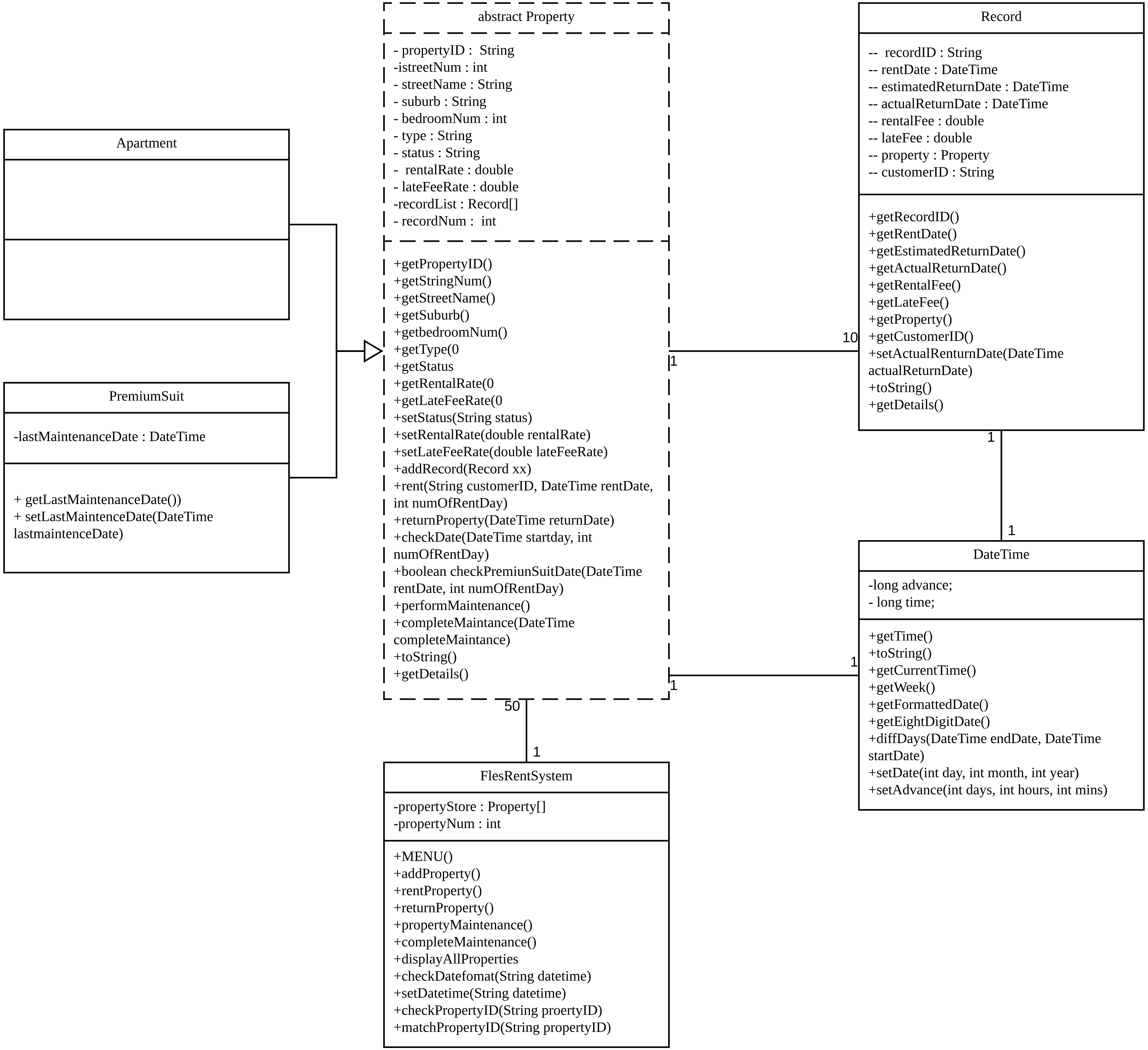


abstract Property

- propertyID : String
-istreetNum : int
- streetName : String
- suburb : String
- bedroomNum : int
- type : String
- status : String
- rentalRate : double
- lateFeeRate : double
-recordList : Record[]
- recordNum : int

+getPropertyID()
+getStringNum()
+getStreetName()
+getSuburb()
+getbedroomNum()
+getType(0
+getStatus
+getRentalRate(0
+getLateFeeRate(0
+setStatus(String status)
+setRentalRate(double rentalRate)
+setLateFeeRate(double lateFeeRate)
+addRecord(Record xx)
+rent(String customerID, DateTime rentDate, int numOfRentDay)
+returnProperty(DateTime returnDate)
+checkDate(DateTime startday, int numOfRentDay)
+boolean checkPremiunSuitDate(DateTime rentDate, int numOfRentDay)
+performMaintenance()
+completeMaintance(DateTime completeMaintance)
+toString()
+getDetails()

Record

-- recordID : String
-- rentDate : DateTime
-- estimatedReturnDate : DateTime
-- actualReturnDate : DateTime
-- rentalFee : double
-- lateFee : double
-- property : Property
-- customerID : String

+getRecordID()
+getRentDate()
+getEstimatedReturnDate()
+getActualReturnDate()
+getRentalFee()
+getLateFee()
+getProperty()
+getCustomerID()
+setActualRenturnDate(DateTime actualReturnDate)
+toString()
+getDetails()

Apartment

PremiumSuit

-lastMaintenanceDate : DateTime

+ getLastMaintenanceDate())
+ setLastMaintenceDate(DateTime lastmaintenceDate)

DateTime

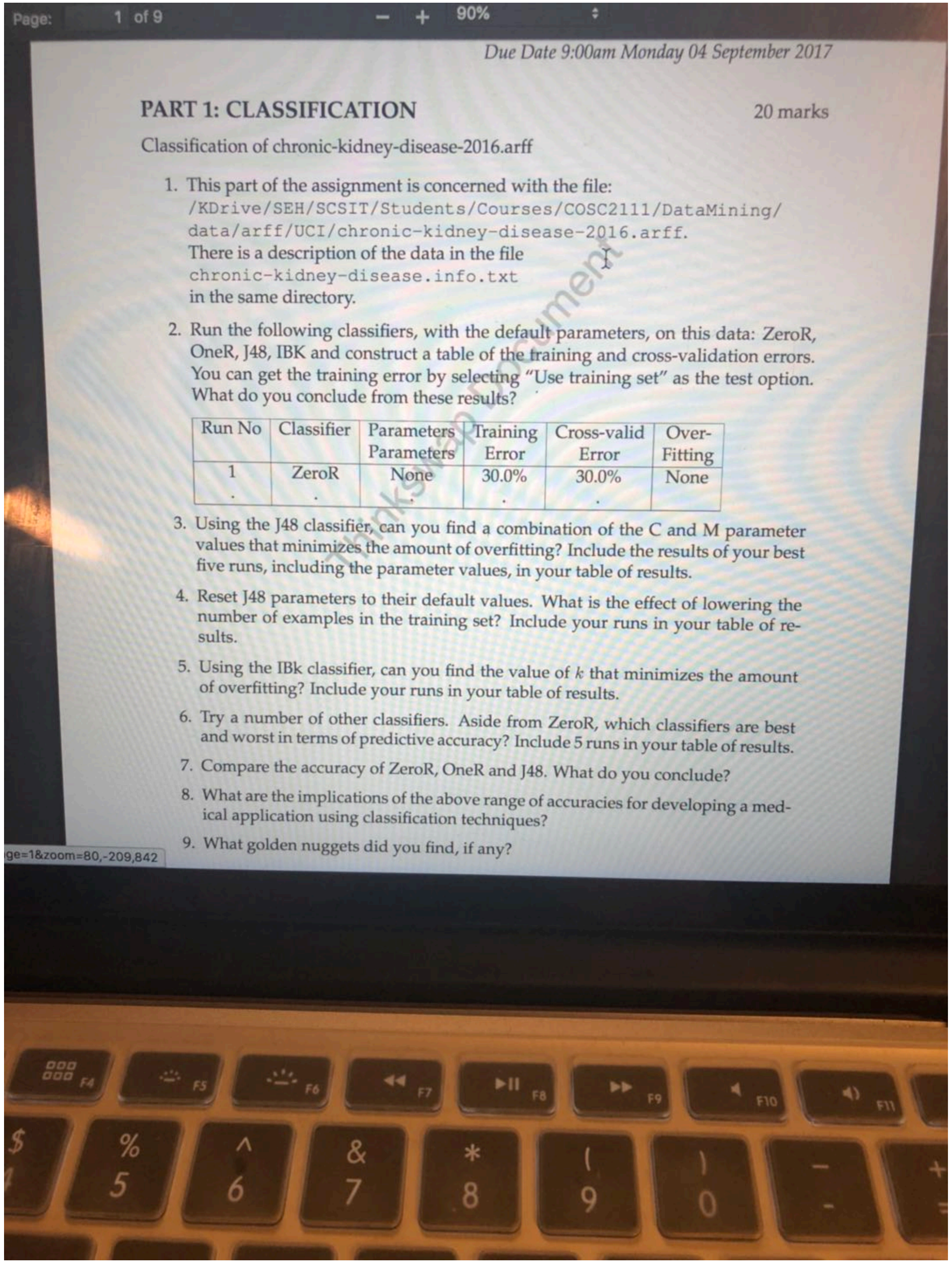
-long advance;
- long time;

+getTime()
+toString()
+getCurrentTime()
+getWeek()
+getFormattedDate()
+getEightDigitDate()
+diffDays(DateTime endDate, DateTime startDate)
+setDate(int day, int month, int year)
+setAdvance(int days, int hours, int mins)

FlesRentSystem

-propertyStore : Property[]
-propertyNum : int

+MENU()
+addProperty()
+rentProperty()
+returnProperty()
+propertyMaintenance()
+completeMaintenance()
+displayAllProperties()
+checkDatefomat(String datetime)
+setDatetime(String datetime)
+checkPropertyID(String proertyID)
+matchPropertyID(String propertyID)


Page: 1 of 9
90%
ge=1&zoom=80,-209,842

Due Date 9:00am Monday 04 September 2017

PART 1: CLASSIFICATION

20 marks

Classification of chronic-kidney-disease-2016.arff

1. This part of the assignment is concerned with the file:
`/KDrive/SEH/SCSIT/Students/Courses/COSC2111/DataMining/data/arff/UCI/chronic-kidney-disease-2016.arff`.
There is a description of the data in the file
`chronic-kidney-disease.info.txt`
in the same directory.

2. Run the following classifiers, with the default parameters, on this data: ZeroR, OneR, J48, IBK and construct a table of the training and cross-validation errors. You can get the training error by selecting "Use training set" as the test option. What do you conclude from these results?

Run No	Classifier	Parameters	Training	Cross-valid	Over-
		Parameters	Error	Error	Fitting
1	ZeroR	None	30.0%	30.0%	None
.	.	.	.	.	

3. Using the J48 classifier, can you find a combination of the C and M parameter values that minimizes the amount of overfitting? Include the results of your best five runs, including the parameter values, in your table of results.

4. Reset J48 parameters to their default values. What is the effect of lowering the number of examples in the training set? Include your runs in your table of results.

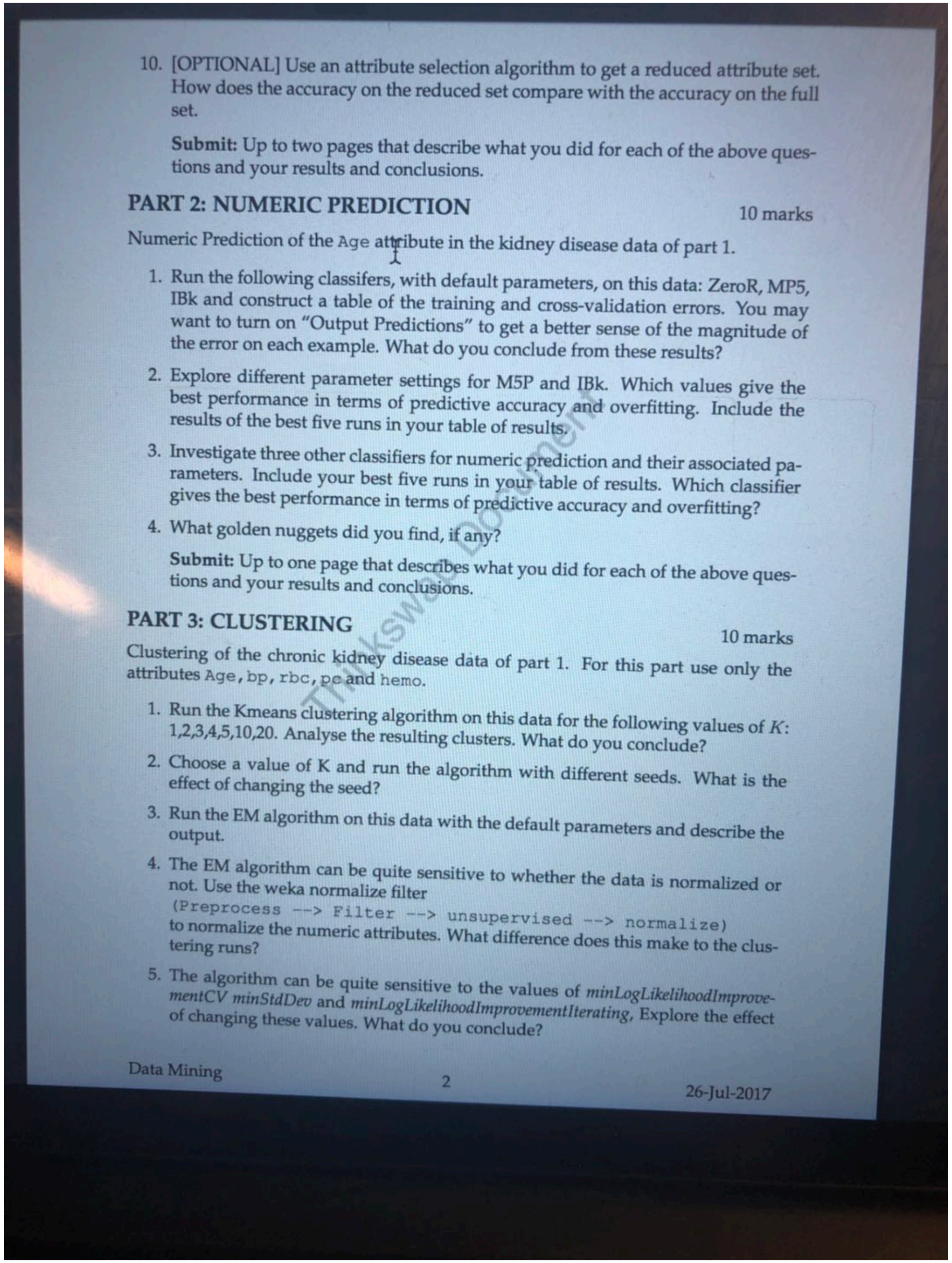
5. Using the IBk classifier, can you find the value of k that minimizes the amount of overfitting? Include your runs in your table of results.

6. Try a number of other classifiers. Aside from ZeroR, which classifiers are best and worst in terms of predictive accuracy? Include 5 runs in your table of results.

7. Compare the accuracy of ZeroR, OneR and J48. What do you conclude?

8. What are the implications of the above range of accuracies for developing a medical application using classification techniques?

9. What golden nuggets did you find, if any?



10. [OPTIONAL] Use an attribute selection algorithm to get a reduced attribute set. How does the accuracy on the reduced set compare with the accuracy on the full set.

    Submit: Up to two pages that describe what you did for each of the above questions and your results and conclusions.

PART 2: NUMERIC PREDICTION 10 marks

Numeric Prediction of the `Age` attribute in the kidney disease data of part 1.

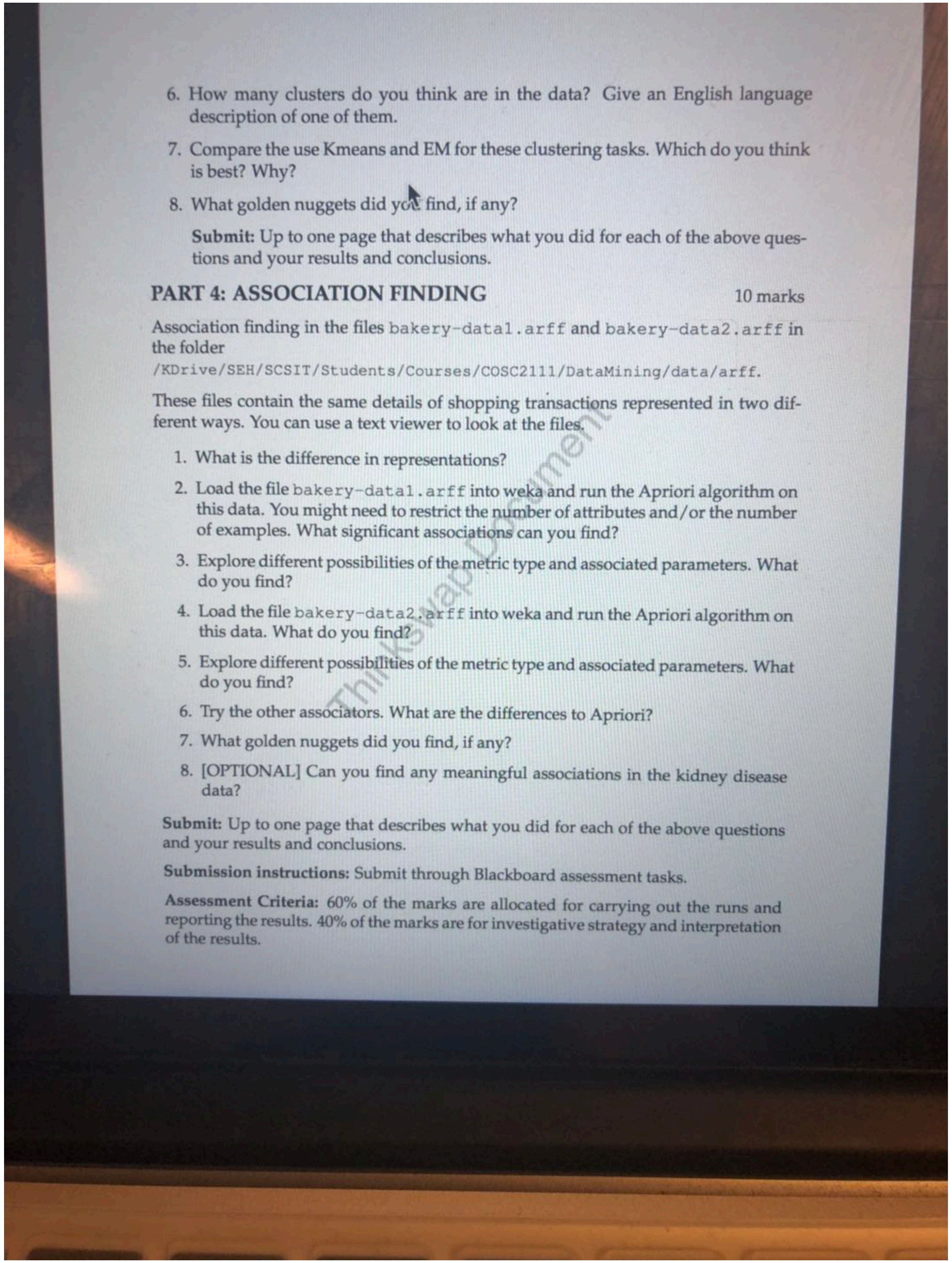
1. Run the following classifers, with default parameters, on this data: ZeroR, MP5, IBk and construct a table of the training and cross-validation errors. You may want to turn on "Output Predictions" to get a better sense of the magnitude of the error on each example. What do you conclude from these results?
2. Explore different parameter settings for M5P and IBk. Which values give the best performance in terms of predictive accuracy and overfitting. Include the results of the best five runs in your table of results.
3. Investigate three other classifiers for numeric prediction and their associated parameters. Include your best five runs in your table of results. Which classifier gives the best performance in terms of predictive accuracy and overfitting?
4. What golden nuggets did you find, if any?

   Submit: Up to one page that describes what you did for each of the above questions and your results and conclusions.

PART 3: CLUSTERING 10 marks

Clustering of the chronic kidney disease data of part 1. For this part use only the attributes `Age`, `bp`, `rbc`, `pc` and `hemo`.

1. Run the Kmeans clustering algorithm on this data for the following values of K: 1,2,3,4,5,10,20. Analyse the resulting clusters. What do you conclude?
2. Choose a value of K and run the algorithm with different seeds. What is the effect of changing the seed?
3. Run the EM algorithm on this data with the default parameters and describe the output.
4. The EM algorithm can be quite sensitive to whether the data is normalized or not. Use the weka normalize filter
   `(Preprocess --> Filter --> unsupervised --> normalize)`
   to normalize the numeric attributes. What difference does this make to the clustering runs?
5. The algorithm can be quite sensitive to the values of *minLogLikelihoodImprovementCV minStdDev* and *minLogLikelihoodImprovementIterating*, Explore the effect of changing these values. What do you conclude?



6. How many clusters do you think are in the data? Give an English language description of one of them.
7. Compare the use Kmeans and EM for these clustering tasks. Which do you think is best? Why?
8. What golden nuggets did you find, if any?

Submit: Up to one page that describes what you did for each of the above questions and your results and conclusions.

PART 4: ASSOCIATION FINDING 10 marks

Association finding in the files `bakery-data1.arff` and `bakery-data2.arff` in the folder
`/KDrive/SEH/SCSIT/Students/Courses/COSC2111/DataMining/data/arff`.

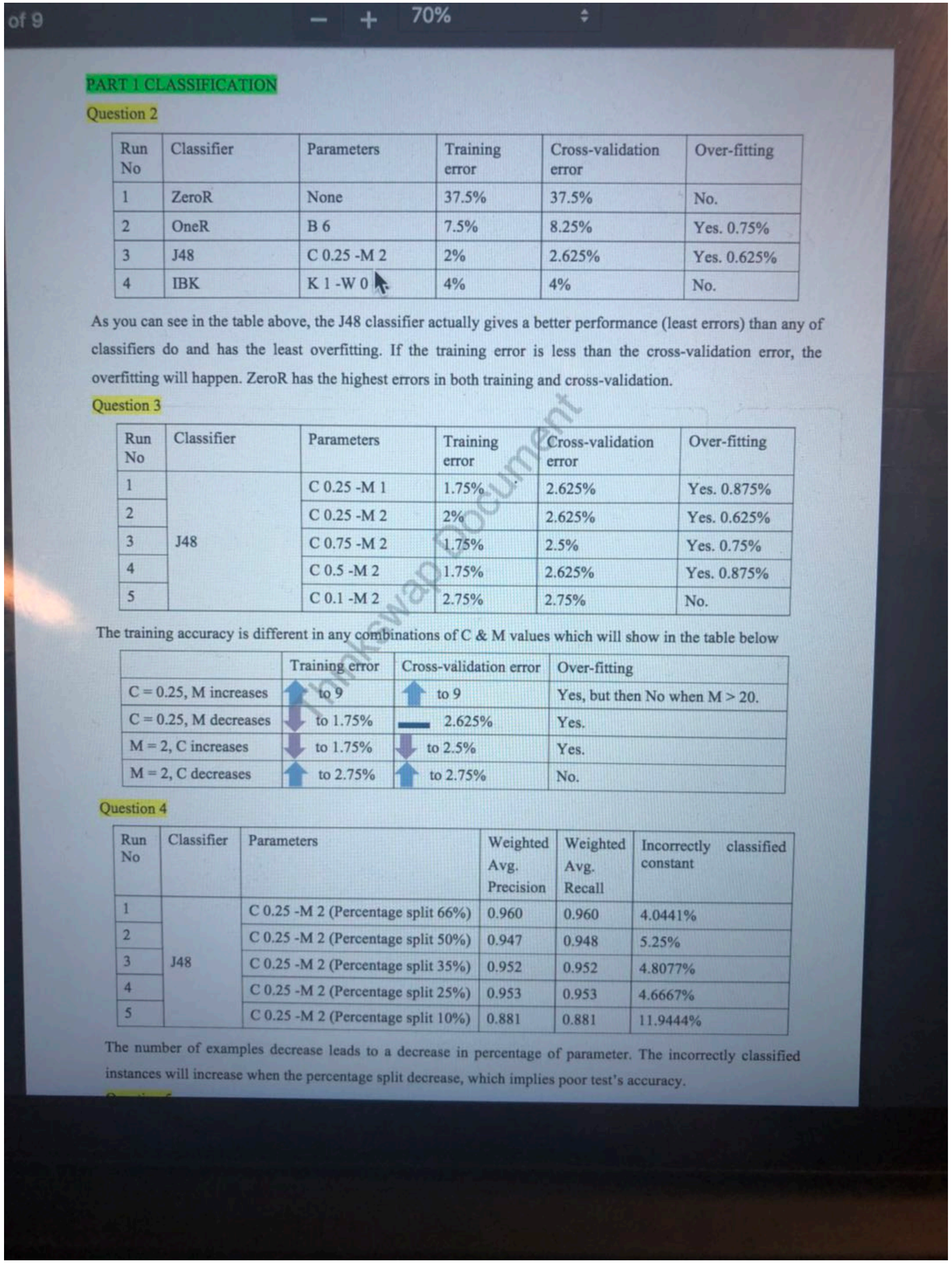
These files contain the same details of shopping transactions represented in two different ways. You can use a text viewer to look at the files.

1. What is the difference in representations?
2. Load the file `bakery-data1.arff` into weka and run the Apriori algorithm on this data. You might need to restrict the number of attributes and/or the number of examples. What significant associations can you find?
3. Explore different possibilities of the metric type and associated parameters. What do you find?
4. Load the file `bakery-data2.arff` into weka and run the Apriori algorithm on this data. What do you find?
5. Explore different possibilities of the metric type and associated parameters. What do you find?
6. Try the other associators. What are the differences to Apriori?
7. What golden nuggets did you find, if any?
8. [OPTIONAL] Can you find any meaningful associations in the kidney disease data?

Submit: Up to one page that describes what you did for each of the above questions and your results and conclusions.

Submission instructions: Submit through Blackboard assessment tasks.

Assessment Criteria: 60% of the marks are allocated for carrying out the runs and reporting the results. 40% of the marks are for investigative strategy and interpretation of the results.



PART 1 CLASSIFICATION

Question 2

Run No	Classifier	Parameters	Training error	Cross-validation error	Over-fitting
1	ZeroR	None	37.5%	37.5%	No.
2	OneR	B 6	7.5%	8.25%	Yes. 0.75%
3	J48	C 0.25 -M 2	2%	2.625%	Yes. 0.625%
4	IBK	K 1 -W 0	4%	4%	No.

As you can see in the table above, the J48 classifier actually gives a better performance (least errors) than any of classifiers do and has the least overfitting. If the training error is less than the cross-validation error, the overfitting will happen. ZeroR has the highest errors in both training and cross-validation.

Question 3

Run No	Classifier	Parameters	Training error	Cross-validation error	Over-fitting
1	J48	C 0.25 -M 1	1.75%	2.625%	Yes. 0.875%
2		C 0.25 -M 2	2%	2.625%	Yes. 0.625%
3		C 0.75 -M 2	1.75%	2.5%	Yes. 0.75%
4		C 0.5 -M 2	1.75%	2.625%	Yes. 0.875%
5		C 0.1 -M 2	2.75%	2.75%	No.

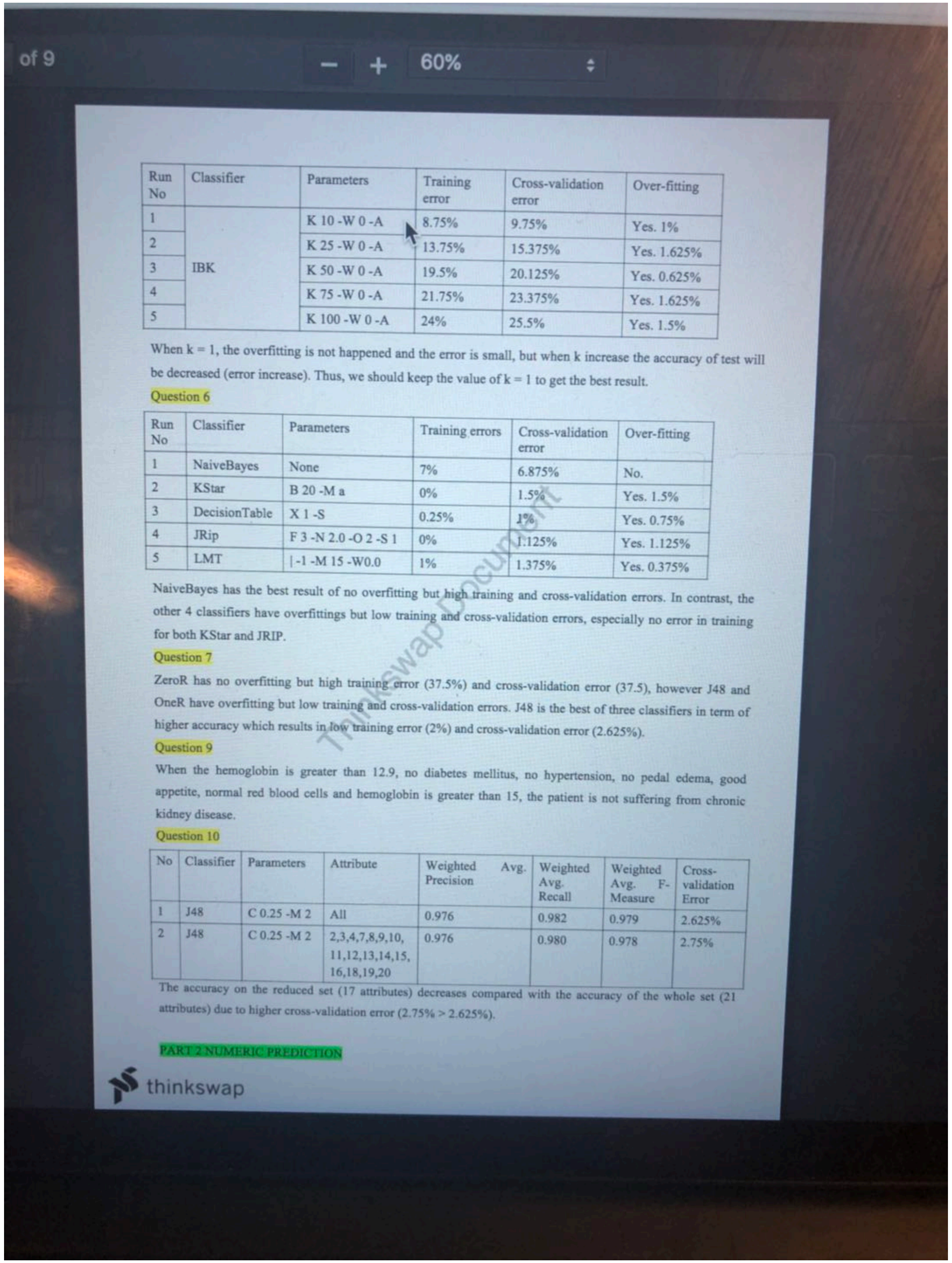
The training accuracy is different in any combinations of C & M values which will show in the table below

	Training error	Cross-validation error	Over-fitting
C = 0.25, M increases	⬆ to 9	⬆ to 9	Yes, but then No when M > 20.
C = 0.25, M decreases	⬇ to 1.75%	— 2.625%	Yes.
M = 2, C increases	⬇ to 1.75%	⬇ to 2.5%	Yes.
M = 2, C decreases	⬆ to 2.75%	⬆ to 2.75%	No.

Question 4

Run No	Classifier	Parameters	Weighted Avg. Precision	Weighted Avg. Recall	Incorrectly classified constant
1	J48	C 0.25 -M 2 (Percentage split 66%)	0.960	0.960	4.0441%
2		C 0.25 -M 2 (Percentage split 50%)	0.947	0.948	5.25%
3		C 0.25 -M 2 (Percentage split 35%)	0.952	0.952	4.8077%
4		C 0.25 -M 2 (Percentage split 25%)	0.953	0.953	4.6667%
5		C 0.25 -M 2 (Percentage split 10%)	0.881	0.881	11.9444%

The number of examples decrease leads to a decrease in percentage of parameter. The incorrectly classified instances will increase when the percentage split decrease, which implies poor test's accuracy.



Run No	Classifier	Parameters	Training error	Cross-validation error	Over-fitting
1	IBK	K 10 -W 0 -A	8.75%	9.75%	Yes. 1%
2		K 25 -W 0 -A	13.75%	15.375%	Yes. 1.625%
3		K 50 -W 0 -A	19.5%	20.125%	Yes. 0.625%
4		K 75 -W 0 -A	21.75%	23.375%	Yes. 1.625%
5		K 100 -W 0 -A	24%	25.5%	Yes. 1.5%

When k = 1, the overfitting is not happened and the error is small, but when k increase the accuracy of test will be decreased (error increase). Thus, we should keep the value of k = 1 to get the best result.

Question 6

Run No	Classifier	Parameters	Training errors	Cross-validation error	Over-fitting
1	NaiveBayes	None	7%	6.875%	No.
2	KStar	B 20 -M a	0%	1.5%	Yes. 1.5%
3	DecisionTable	X 1 -S	0.25%	1%	Yes. 0.75%
4	JRip	F 3 -N 2.0 -O 2 -S 1	0%	1.125%	Yes. 1.125%
5	LMT	| -1 -M 15 -W0.0	1%	1.375%	Yes. 0.375%

NaiveBayes has the best result of no overfitting but high training and cross-validation errors. In contrast, the other 4 classifiers have overfittings but low training and cross-validation errors, especially no error in training for both KStar and JRIP.

Question 7

ZeroR has no overfitting but high training error (37.5%) and cross-validation error (37.5), however J48 and OneR have overfitting but low training and cross-validation errors. J48 is the best of three classifiers in term of higher accuracy which results in low training error (2%) and cross-validation error (2.625%).

Question 9

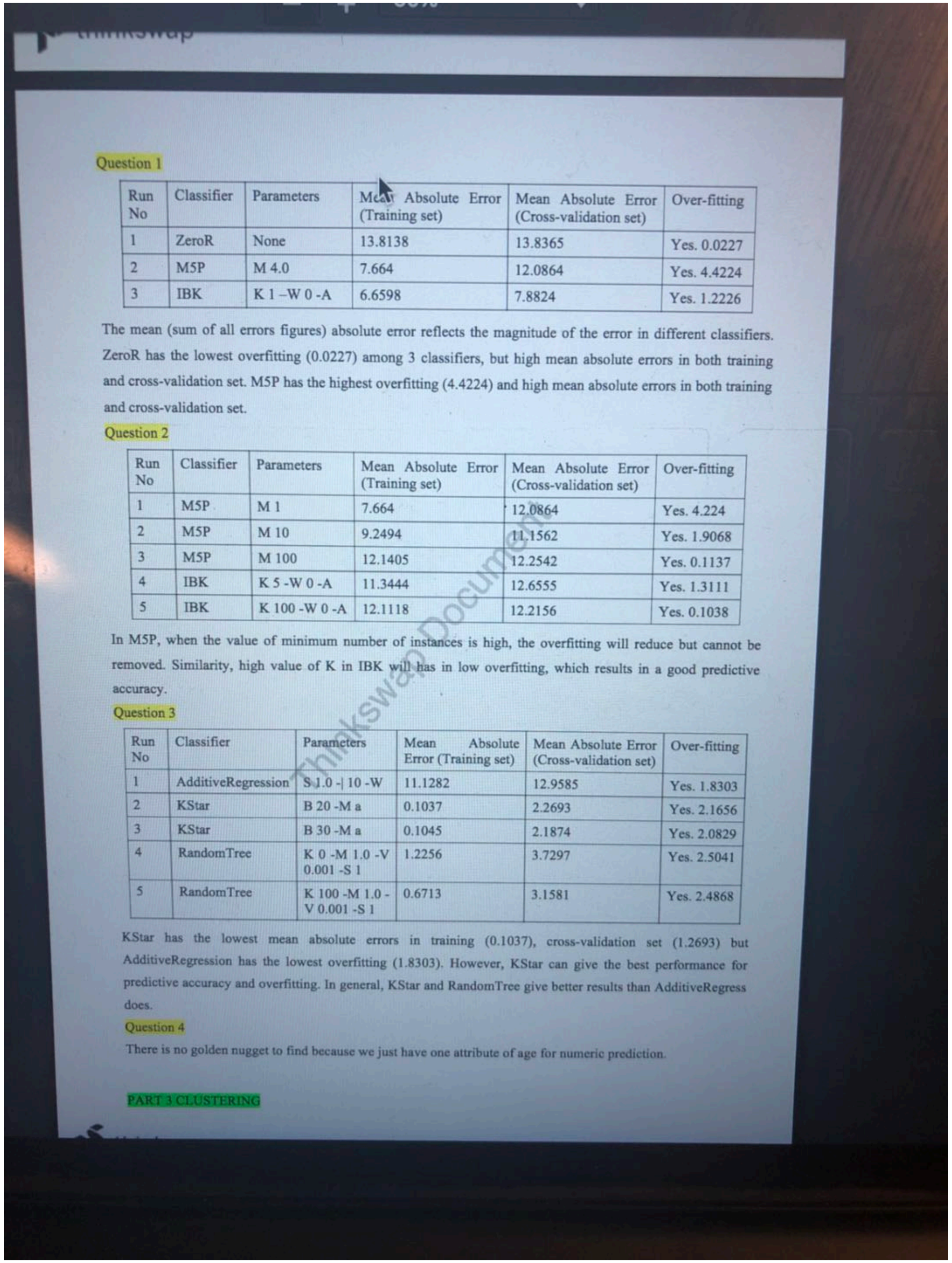
When the hemoglobin is greater than 12.9, no diabetes mellitus, no hypertension, no pedal edema, good appetite, normal red blood cells and hemoglobin is greater than 15, the patient is not suffering from chronic kidney disease.

Question 10

No	Classifier	Parameters	Attribute	Weighted Avg. Precision	Weighted Avg. Recall	Weighted Avg. F-Measure	Cross-validation Error
1	J48	C 0.25 -M 2	All	0.976	0.982	0.979	2.625%
2	J48	C 0.25 -M 2	2,3,4,7,8,9,10, 11,12,13,14,15, 16,18,19,20	0.976	0.980	0.978	2.75%

The accuracy on the reduced set (17 attributes) decreases compared with the accuracy of the whole set (21 attributes) due to higher cross-validation error (2.75% > 2.625%).

PART 2 NUMERIC PREDICTION



Question 1

Run No	Classifier	Parameters	Mean Absolute Error (Training set)	Mean Absolute Error (Cross-validation set)	Over-fitting
1	ZeroR	None	13.8138	13.8365	Yes. 0.0227
2	M5P	M 4.0	7.664	12.0864	Yes. 4.4224
3	IBK	K 1 –W 0 -A	6.6598	7.8824	Yes. 1.2226

The mean (sum of all errors figures) absolute error reflects the magnitude of the error in different classifiers. ZeroR has the lowest overfitting (0.0227) among 3 classifiers, but high mean absolute errors in both training and cross-validation set. M5P has the highest overfitting (4.4224) and high mean absolute errors in both training and cross-validation set.

Question 2

Run No	Classifier	Parameters	Mean Absolute Error (Training set)	Mean Absolute Error (Cross-validation set)	Over-fitting
1	M5P	M 1	7.664	12.0864	Yes. 4.224
2	M5P	M 10	9.2494	11.1562	Yes. 1.9068
3	M5P	M 100	12.1405	12.2542	Yes. 0.1137
4	IBK	K 5 -W 0 -A	11.3444	12.6555	Yes. 1.3111
5	IBK	K 100 -W 0 -A	12.1118	12.2156	Yes. 0.1038

In M5P, when the value of minimum number of instances is high, the overfitting will reduce but cannot be removed. Similarity, high value of K in IBK will has in low overfitting, which results in a good predictive accuracy.

Question 3

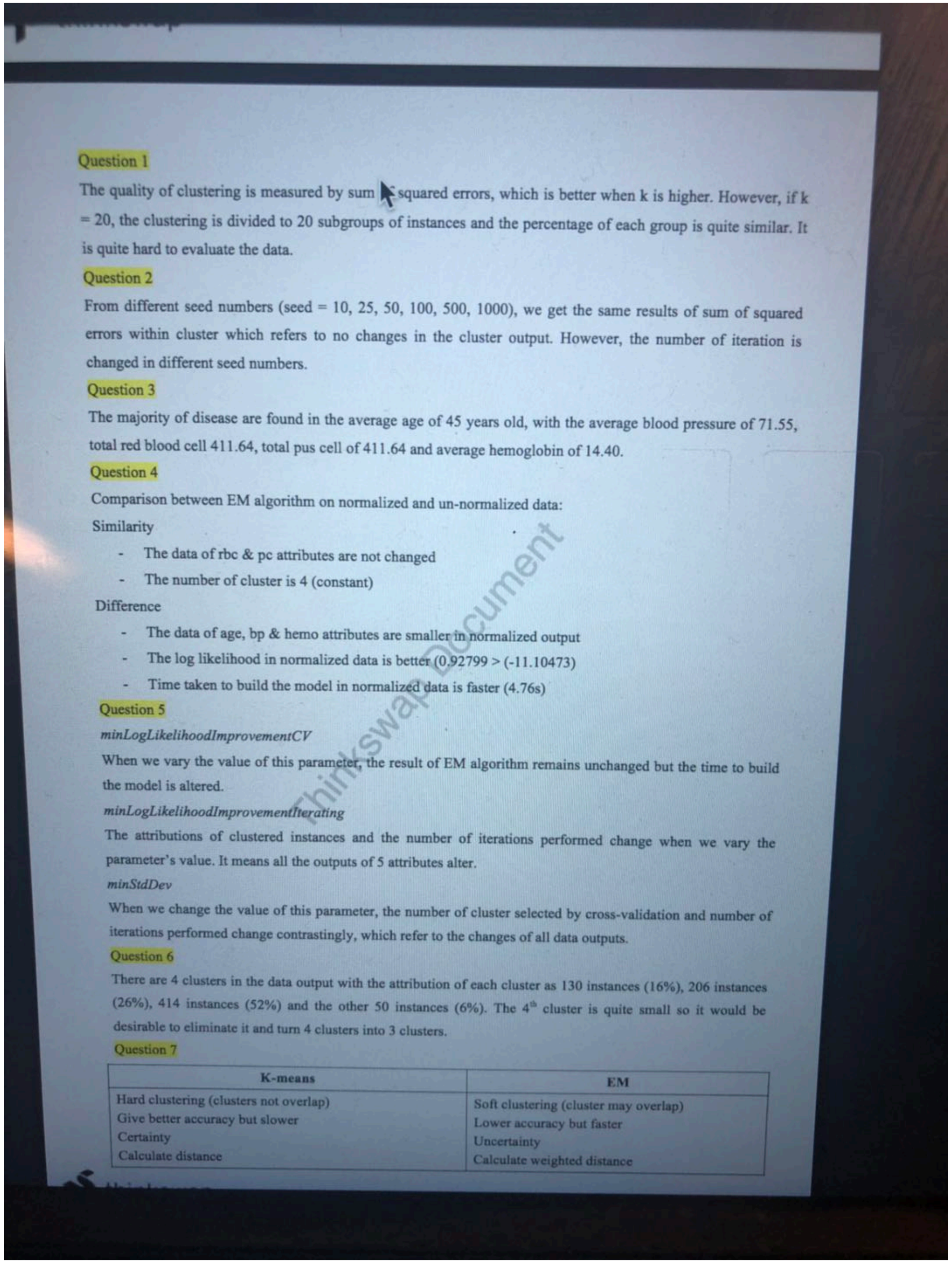
Run No	Classifier	Parameters	Mean Absolute Error (Training set)	Mean Absolute Error (Cross-validation set)	Over-fitting
1	AdditiveRegression	S 1.0 -I 10 -W	11.1282	12.9585	Yes. 1.8303
2	KStar	B 20 -M a	0.1037	2.2693	Yes. 2.1656
3	KStar	B 30 -M a	0.1045	2.1874	Yes. 2.0829
4	RandomTree	K 0 -M 1.0 -V 0.001 -S 1	1.2256	3.7297	Yes. 2.5041
5	RandomTree	K 100 -M 1.0 -V 0.001 -S 1	0.6713	3.1581	Yes. 2.4868

KStar has the lowest mean absolute errors in training (0.1037), cross-validation set (1.2693) but AdditiveRegression has the lowest overfitting (1.8303). However, KStar can give the best performance for predictive accuracy and overfitting. In general, KStar and RandomTree give better results than AdditiveRegress does.

Question 4

There is no golden nugget to find because we just have one attribute of age for numeric prediction.

PART 3 CLUSTERING



Question 1

The quality of clustering is measured by sum of squared errors, which is better when k is higher. However, if k = 20, the clustering is divided to 20 subgroups of instances and the percentage of each group is quite similar. It is quite hard to evaluate the data.

Question 2

From different seed numbers (seed = 10, 25, 50, 100, 500, 1000), we get the same results of sum of squared errors within cluster which refers to no changes in the cluster output. However, the number of iteration is changed in different seed numbers.

Question 3

The majority of disease are found in the average age of 45 years old, with the average blood pressure of 71.55, total red blood cell 411.64, total pus cell of 411.64 and average hemoglobin of 14.40.

Question 4

Comparison between EM algorithm on normalized and un-normalized data:

Similarity

- The data of rbc & pc attributes are not changed
- The number of cluster is 4 (constant)

Difference

- The data of age, bp & hemo attributes are smaller in normalized output
- The log likelihood in normalized data is better (0.92799 > (-11.10473)
- Time taken to build the model in normalized data is faster (4.76s)

Question 5

minLogLikelihoodImprovementCV

When we vary the value of this parameter, the result of EM algorithm remains unchanged but the time to build the model is altered.

minLogLikelihoodImprovementIterating

The attributions of clustered instances and the number of iterations performed change when we vary the parameter's value. It means all the outputs of 5 attributes alter.

minStdDev

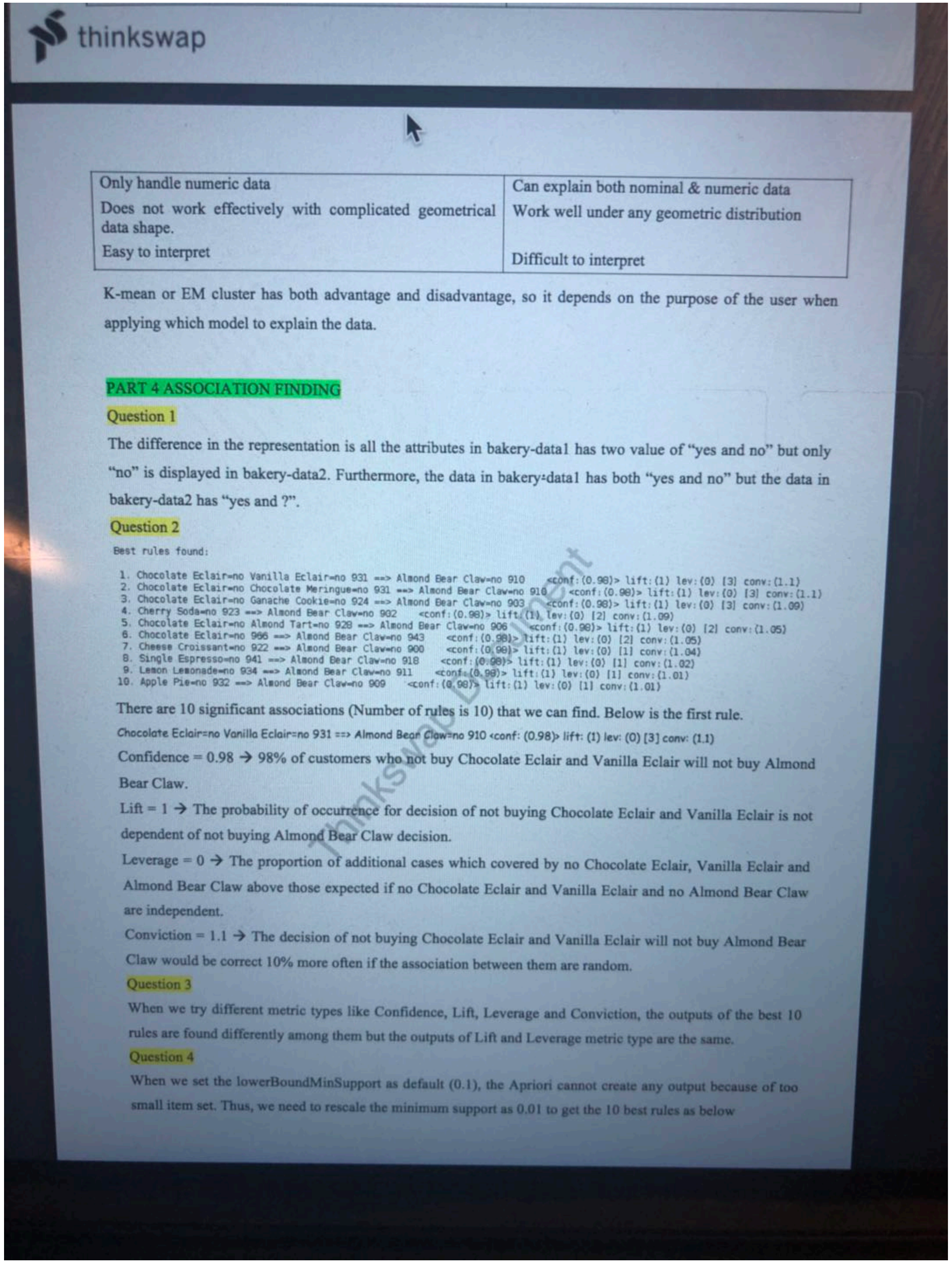
When we change the value of this parameter, the number of cluster selected by cross-validation and number of iterations performed change contrastingly, which refer to the changes of all data outputs.

Question 6

There are 4 clusters in the data output with the attribution of each cluster as 130 instances (16%), 206 instances (26%), 414 instances (52%) and the other 50 instances (6%). The 4th cluster is quite small so it would be desirable to eliminate it and turn 4 clusters into 3 clusters.

Question 7

K-means	EM
Hard clustering (clusters not overlap) Give better accuracy but slower Certainty Calculate distance	Soft clustering (cluster may overlap) Lower accuracy but faster Uncertainty Calculate weighted distance



Only handle numeric data	Can explain both nominal & numeric data
Does not work effectively with complicated geometrical data shape.	Work well under any geometric distribution
Easy to interpret	Difficult to interpret

K-mean or EM cluster has both advantage and disadvantage, so it depends on the purpose of the user when applying which model to explain the data.

PART 4 ASSOCIATION FINDING

Question 1

The difference in the representation is all the attributes in bakery-data1 has two value of "yes and no" but only "no" is displayed in bakery-data2. Furthermore, the data in bakery-data1 has both "yes and no" but the data in bakery-data2 has "yes and ?".

Question 2

```
Best rules found:

 1. Chocolate Eclair=no Vanilla Eclair=no 931 ==> Almond Bear Claw=no 910    <conf:(0.98)> lift:(1) lev:(0) [3] conv:(1.1)
 2. Chocolate Eclair=no Chocolate Meringue=no 931 ==> Almond Bear Claw=no 910    <conf:(0.98)> lift:(1) lev:(0) [3] conv:(1.1)
 3. Chocolate Eclair=no Ganache Cookie=no 924 ==> Almond Bear Claw=no 903    <conf:(0.98)> lift:(1) lev:(0) [3] conv:(1.09)
 4. Cherry Soda=no 923 ==> Almond Bear Claw=no 902    <conf:(0.98)> lift:(1) lev:(0) [2] conv:(1.09)
 5. Chocolate Eclair=no Almond Tart=no 928 ==> Almond Bear Claw=no 906    <conf:(0.98)> lift:(1) lev:(0) [2] conv:(1.05)
 6. Chocolate Eclair=no 966 ==> Almond Bear Claw=no 943    <conf:(0.98)> lift:(1) lev:(0) [2] conv:(1.05)
 7. Cheese Croissant=no 922 ==> Almond Bear Claw=no 900    <conf:(0.98)> lift:(1) lev:(0) [1] conv:(1.04)
 8. Single Espresso=no 941 ==> Almond Bear Claw=no 918    <conf:(0.98)> lift:(1) lev:(0) [1] conv:(1.02)
 9. Lemon Lemonade=no 934 ==> Almond Bear Claw=no 911    <conf:(0.98)> lift:(1) lev:(0) [1] conv:(1.01)
10. Apple Pie=no 932 ==> Almond Bear Claw=no 909    <conf:(0.98)> lift:(1) lev:(0) [1] conv:(1.01)
```

There are 10 significant associations (Number of rules is 10) that we can find. Below is the first rule.

Chocolate Eclair=no Vanilla Eclair=no 931 ==> Almond Bear Claw=no 910 <conf: (0.98)> lift: (1) lev: (0) [3] conv: (1.1)

Confidence = 0.98 → 98% of customers who not buy Chocolate Eclair and Vanilla Eclair will not buy Almond Bear Claw.

Lift = 1 → The probability of occurrence for decision of not buying Chocolate Eclair and Vanilla Eclair is not dependent of not buying Almond Bear Claw decision.

Leverage = 0 → The proportion of additional cases which covered by no Chocolate Eclair, Vanilla Eclair and Almond Bear Claw above those expected if no Chocolate Eclair and Vanilla Eclair and no Almond Bear Claw are independent.

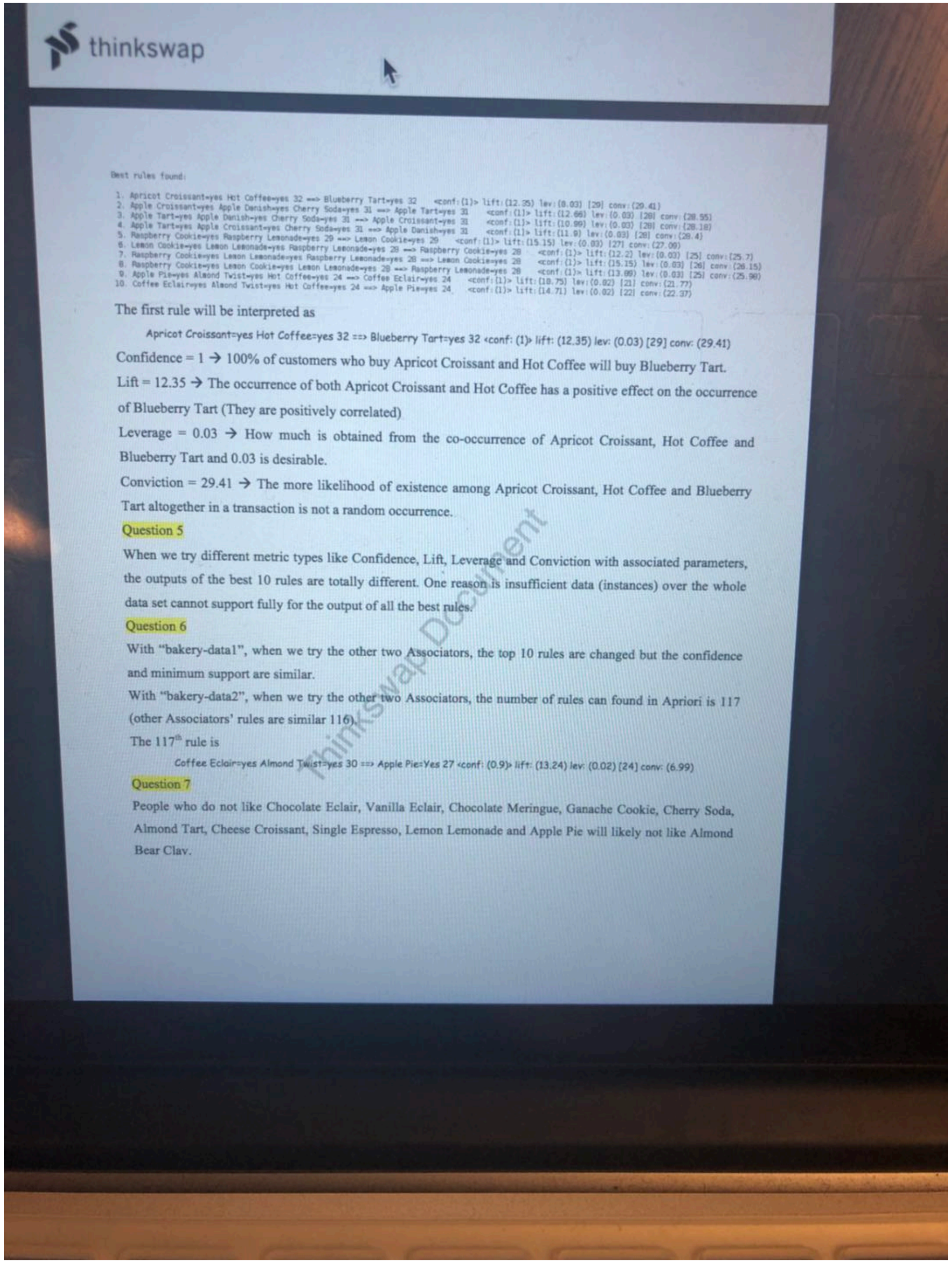
Conviction = 1.1 → The decision of not buying Chocolate Eclair and Vanilla Eclair will not buy Almond Bear Claw would be correct 10% more often if the association between them are random.

Question 3

When we try different metric types like Confidence, Lift, Leverage and Conviction, the outputs of the best 10 rules are found differently among them but the outputs of Lift and Leverage metric type are the same.

Question 4

When we set the lowerBoundMinSupport as default (0.1), the Apriori cannot create any output because of too small item set. Thus, we need to rescale the minimum support as 0.01 to get the 10 best rules as below



```
Best rules found:

 1. Apricot Croissant=yes Hot Coffee=yes 32 ==> Blueberry Tart=yes 32    <conf:(1)> lift:(12.35) lev:(0.03) [29] conv:(29.41)
 2. Apple Croissant=yes Apple Danish=yes Cherry Soda=yes 31 ==> Apple Tart=yes 31    <conf:(1)> lift:(12.66) lev:(0.03) [28] conv:(28.55)
 3. Apple Tart=yes Apple Danish=yes Cherry Soda=yes 31 ==> Apple Croissant=yes 31    <conf:(1)> lift:(10.99) lev:(0.03) [28] conv:(28.18)
 4. Apple Tart=yes Apple Croissant=yes Cherry Soda=yes 31 ==> Apple Danish=yes 31    <conf:(1)> lift:(11.9) lev:(0.03) [28] conv:(28.4)
 5. Raspberry Cookie=yes Raspberry Lemonade=yes 29 ==> Lemon Cookie=yes 29    <conf:(1)> lift:(15.15) lev:(0.03) [27] conv:(27.09)
 6. Lemon Cookie=yes Lemon Lemonade=yes Raspberry Lemonade=yes 28 ==> Raspberry Cookie=yes 28    <conf:(1)> lift:(12.2) lev:(0.03) [25] conv:(25.7)
 7. Raspberry Cookie=yes Lemon Lemonade=yes Raspberry Lemonade=yes 28 ==> Lemon Cookie=yes 28    <conf:(1)> lift:(15.15) lev:(0.03) [26] conv:(26.15)
 8. Raspberry Cookie=yes Lemon Cookie=yes Lemon Lemonade=yes 28 ==> Raspberry Lemonade=yes 28    <conf:(1)> lift:(13.89) lev:(0.03) [25] conv:(25.98)
 9. Apple Pie=yes Almond Twist=yes Hot Coffee=yes 24 ==> Coffee Eclair=yes 24    <conf:(1)> lift:(10.75) lev:(0.02) [21] conv:(21.77)
10. Coffee Eclair=yes Almond Twist=yes Hot Coffee=yes 24 ==> Apple Pie=yes 24    <conf:(1)> lift:(14.71) lev:(0.02) [22] conv:(22.37)
```

The first rule will be interpreted as

```
Apricot Croissant=yes Hot Coffee=yes 32 ==> Blueberry Tart=yes 32 <conf: (1)> lift: (12.35) lev: (0.03) [29] conv: (29.41)
```

Confidence = 1 → 100% of customers who buy Apricot Croissant and Hot Coffee will buy Blueberry Tart.

Lift = 12.35 → The occurrence of both Apricot Croissant and Hot Coffee has a positive effect on the occurrence of Blueberry Tart (They are positively correlated)

Leverage = 0.03 → How much is obtained from the co-occurrence of Apricot Croissant, Hot Coffee and Blueberry Tart and 0.03 is desirable.

Conviction = 29.41 → The more likelihood of existence among Apricot Croissant, Hot Coffee and Blueberry Tart altogether in a transaction is not a random occurrence.

Question 5

When we try different metric types like Confidence, Lift, Leverage and Conviction with associated parameters, the outputs of the best 10 rules are totally different. One reason is insufficient data (instances) over the whole data set cannot support fully for the output of all the best rules.

Question 6

With "bakery-data1", when we try the other two Associators, the top 10 rules are changed but the confidence and minimum support are similar.

With "bakery-data2", when we try the other two Associators, the number of rules can found in Apriori is 117 (other Associators' rules are similar 116).

The 117th rule is

```
Coffee Eclair=yes Almond Twist=yes 30 ==> Apple Pie=Yes 27 <conf: (0.9)> lift: (13.24) lev: (0.02) [24] conv: (6.99)
```

Question 7

People who do not like Chocolate Eclair, Vanilla Eclair, Chocolate Meringue, Ganache Cookie, Cherry Soda, Almond Tart, Cheese Croissant, Single Espresso, Lemon Lemonade and Apple Pie will likely not like Almond Bear Clav.