Q1)Provide brief answers (in about 50 words):

a) List considerations to be made while deploying a data mining model.

b) Compare cosine similarity and proximity-based similarity.

Answer:

**A.** For data mining, there are three phases to processing: **querying the source data, determining raw statistics, and using the model definition and algorithm to train the mining model**. The Analysis Services server issues query to the database that provides the raw data

**b.**

* Cosine similarity is the cosine of the angle between two n-dimensional vectors in an n-dimensional space. It is **the dot product of the two vectors divided by the product of the two vectors' lengths (or magnitudes)**.
* The Law of Proximity is **that closer objects are grouped together**. The Law of Similarity is that objects that are similar are grouped together.

**Q2.** You work in the Technology Division of a major hospital. The hospital management do not seem to fully understand value of the data. Prepare a detailed illustrated note explaining to the management how the hospital will benefit from mining their data.

**Answer 2**. As a senior support staff in the technology division of the major hospital of the city. I have created a detailed illustrated note explaining to the management that. how the hospital will be beneficial by adopting data mining.

- - - - - - - - - - - - - - - - - - - - - - - - - -

**What is "mining the data" means?**

In recent times the value of data has been highly increased. The word data means the raw information about the institution it is related to. For a hospital, their data can be the list of patients visiting each day, the medical devices purchased, and the amount of money used in the repairing of the machines and infrastructure.

The major process involved in data mining is Data integration, transformation, cleaning (removing the unwanted data from the system), presentation (adding minds to the data to be understandable), etc.

In the current scenario, data mining is cost-efficient by many software and the skills of data mining engineers. Our hospital can make multiple decisions but the possibility of better results can be highly improved if the decisions will be taken based upon the past analysis.

- - - - - - - - - - - - - - - - - - - - - - - - - -

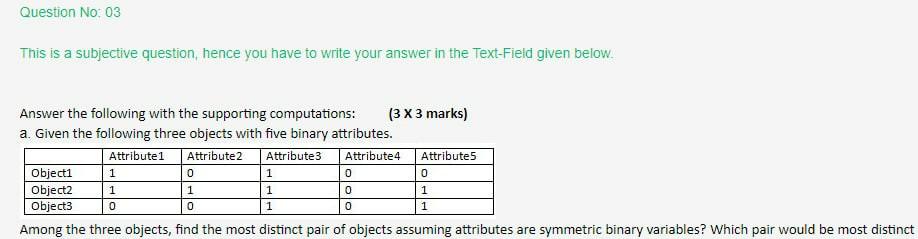
**How data mining can add value to the hospital?**

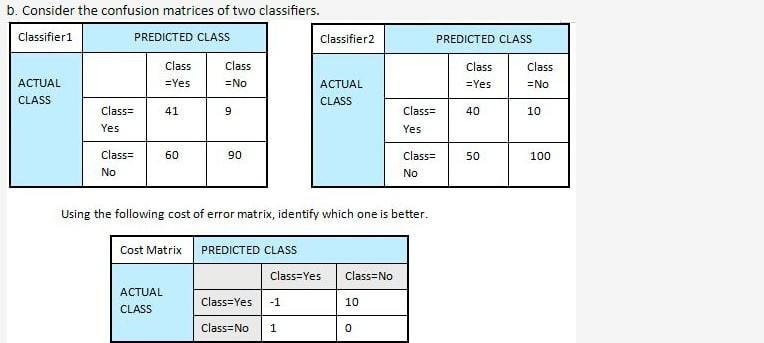
A hospital can have a large amount of data that can be of no use if it is not mined and later changed into valuable information. Mining the data helps in summarising the past faults, correct decisions, profit, loss, investment, increase in patients, and the value for money the hospital is providing to its customers. Data mining can analyze the various aspects of processed data for predicting the future and can also help in making future decisions for the betterment of our hospital. Further advantages are:

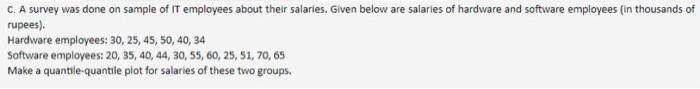
* Identifies hidden risks and probabilities.
* Helps in improving customer satisfaction.
* Provides market analysis.
* Helps in taking profitable decisions.

The customer responses can be mined to finally analyze the value we are providing to our patients, which will surely help us in the marketing campaigns and the reputation we represent of our hospital. It also increases brand loyalty by tracking and analyzing customer behavior.

**Q3.**







To check most distinct pairs we have to calculate Dissimilarity between Binary Variables.

**Dissimilarity** : Numerical measure of how different two data objects are

Suppose 1= positive & 0=Negative

So,

p= number of variables that is positive for one pair of objects

q= number of variables that is positive for ith attribute & negative for jth attribute for (i,j) pair of objects

r= number of variables that is negative for ith attribute & negative for jth attribute for (i,j) pair of objects

s=number of variables that is negative for one pair of objects

Now for symmetric matrix,

d(i,j)= (r+ s)/ (p+q+r+s) ; A/c to Matching Co-efficient rule

**For first pair:- Object 1 & Object 2**

**p= 2 , q= 0, r= 2, s=1**

**d(object1,object2)= 2 + 0 / 2+0+2+1 = 2/5=0.4**

**For second pair:- Object 2 & Object 3**

**p= 2 , q= 1, r= 0, s=1**

**d(object2,object3)= 0 + 1 / 2+1+0+1 = 1/4 = 0.25**

**For third pair:- Object 3 & Object 1**

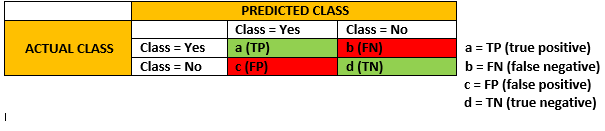
**p= 1 , q= 1, r= 0, s=2**

**d(object3,object1)= 0 + 2 / 1+1+0+2 = 2/4 = 0.5**

Hence,

Pair of Object 3 & Object1 is most distinct.

**Answer B)**



In confusion matrix, if both column and row is yes then it's true positive which means if actual class has the same value which we we predict correctly as **Yes** .

Similarly, If the class has actually different value and we predict wrong value as **No** then it's False negative.

Same way, it goes for FP & TN.

Now, based on error matrix, let's calculate the cost:

**Cost(classifier1)= 41(-1) + 9(10)+ 60(-1) + 90(10)= 889**

**Cost(classifier2)= 40(-1) + 10(10)+ 50(-1) + 100(10)= 1,010**

**We can see that cost of classifier 2 is higher then cost of classifier 1. So we' ll go with classifier 1.**

**Answer C)**

**Step:1**

From the given information we have samples of IT employees and their salaries

Hardware  
30 25 45 50 40 34

Software  
  20 35 40 44 30 55 60 25 51 70 65

we need to make quantile quantile plot

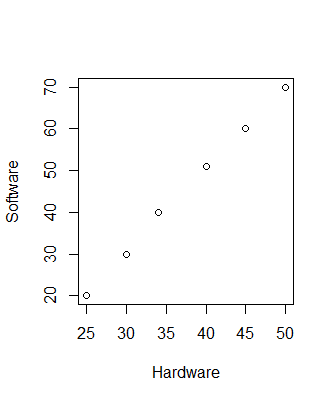
**Step:2**

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

By using R function qqplot() we can draw   quantile quantile plot

Hardware = c(30,25,45,50,40,34)  
> Software =c(20,35,40,44,30,55,60,25,51,70,65)  
> Hardware  
[1] 30 25 45 50 40 34

Software  
[1] 20 35 40 44 30 55 60 25 51 70 65

qqplot(Hardware,Software)  


Note:

In excel please follow the following steps

1. Enter and sort the data. Enter the following data into one column: ...
2. Step 2: Find the rank of each data value. ...
3. Step 3: Find the percentile of each data value. ...
4. Step 4: Calculate the z-score for each data value. ...
5. Step 5: Create the Q-Q plot.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data | Rank | Percentile | Z-Score | Data |
| 20 | 1 | 0.03 | -1.88951 | 20 |
| 25 | 2 | 0.09 | -1.3517 | 25 |
| 25 | 2 | 0.09 | -1.3517 | 25 |
| 30 | 4 | 0.21 | -0.82079 | 30 |
| 30 | 4 | 0.21 | -0.82079 | 30 |
| 34 | 6 | 0.32 | -0.45785 | 34 |
| 35 | 7 | 0.38 | -0.29931 | 35 |
| 40 | 8 | 0.44 | -0.14799 | 40 |
| 40 | 8 | 0.44 | -0.14799 | 40 |
| 44 | 10 | 0.56 | 0.147987 | 44 |
| 45 | 11 | 0.62 | 0.299307 | 45 |
| 50 | 12 | 0.68 | 0.457852 | 50 |
| 51 | 13 | 0.74 | 0.628904 | 51 |
| 55 | 14 | 0.79 | 0.820792 | 55 |
| 60 | 15 | 0.85 | 1.049131 | 60 |
| 65 | 16 | 0.91 | 1.351702 | 65 |
| 70 | 17 | 0.97 | 1.88951 | 70 |

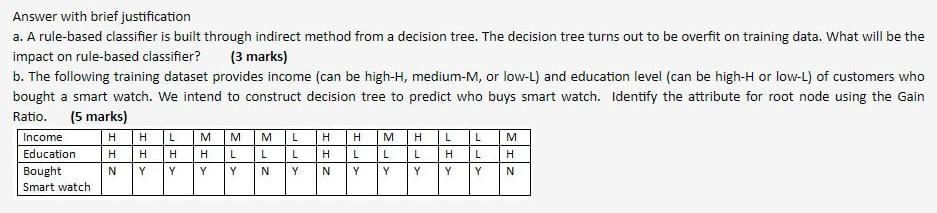
**Data Column:** Consist of both hardware and software salaries data

**Rank: =RANK(A2, $A$2:$A$11, 1)**

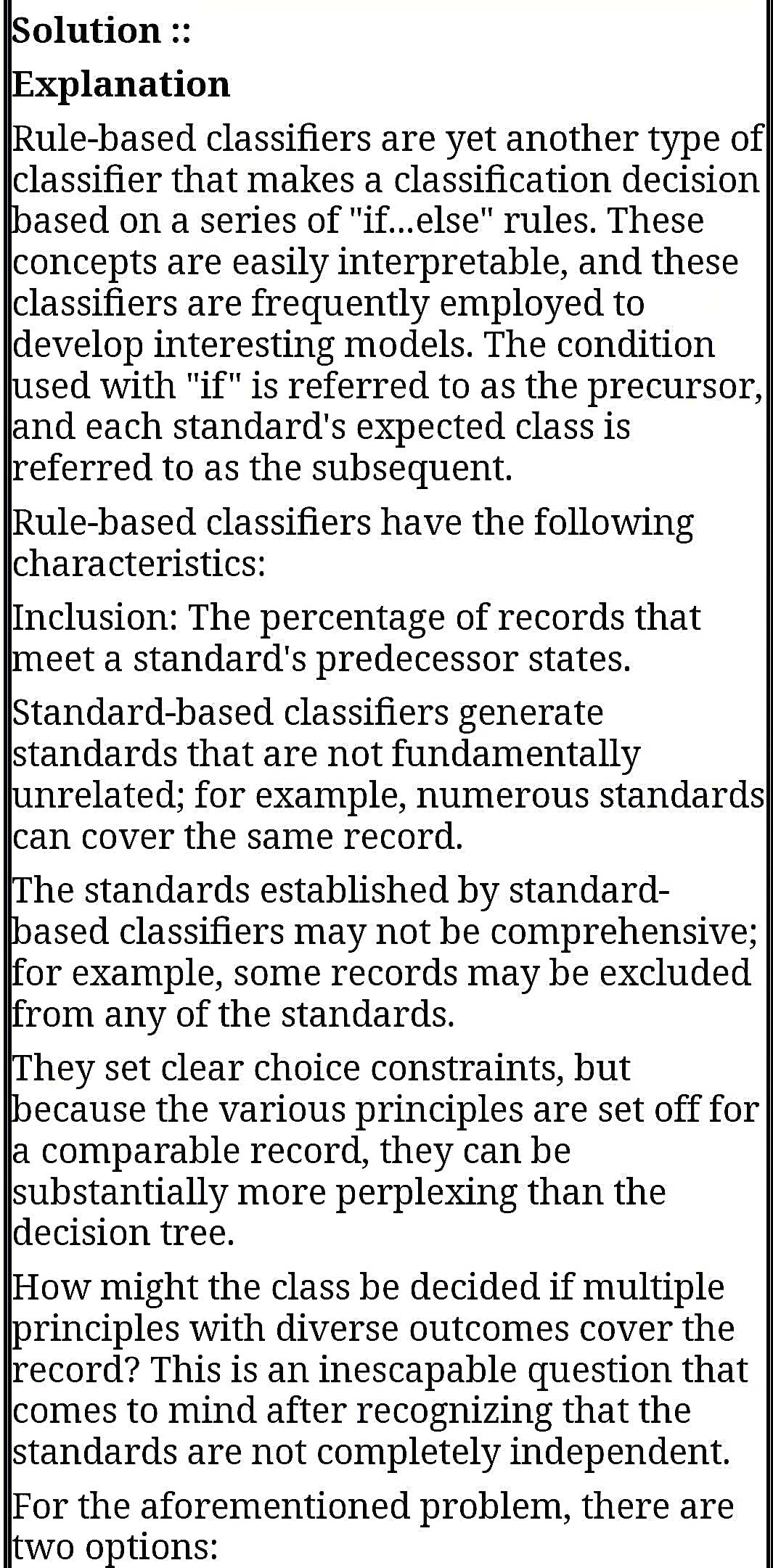
**Percentile: =(B2-0.5)/COUNT($B$2:$B$11)**

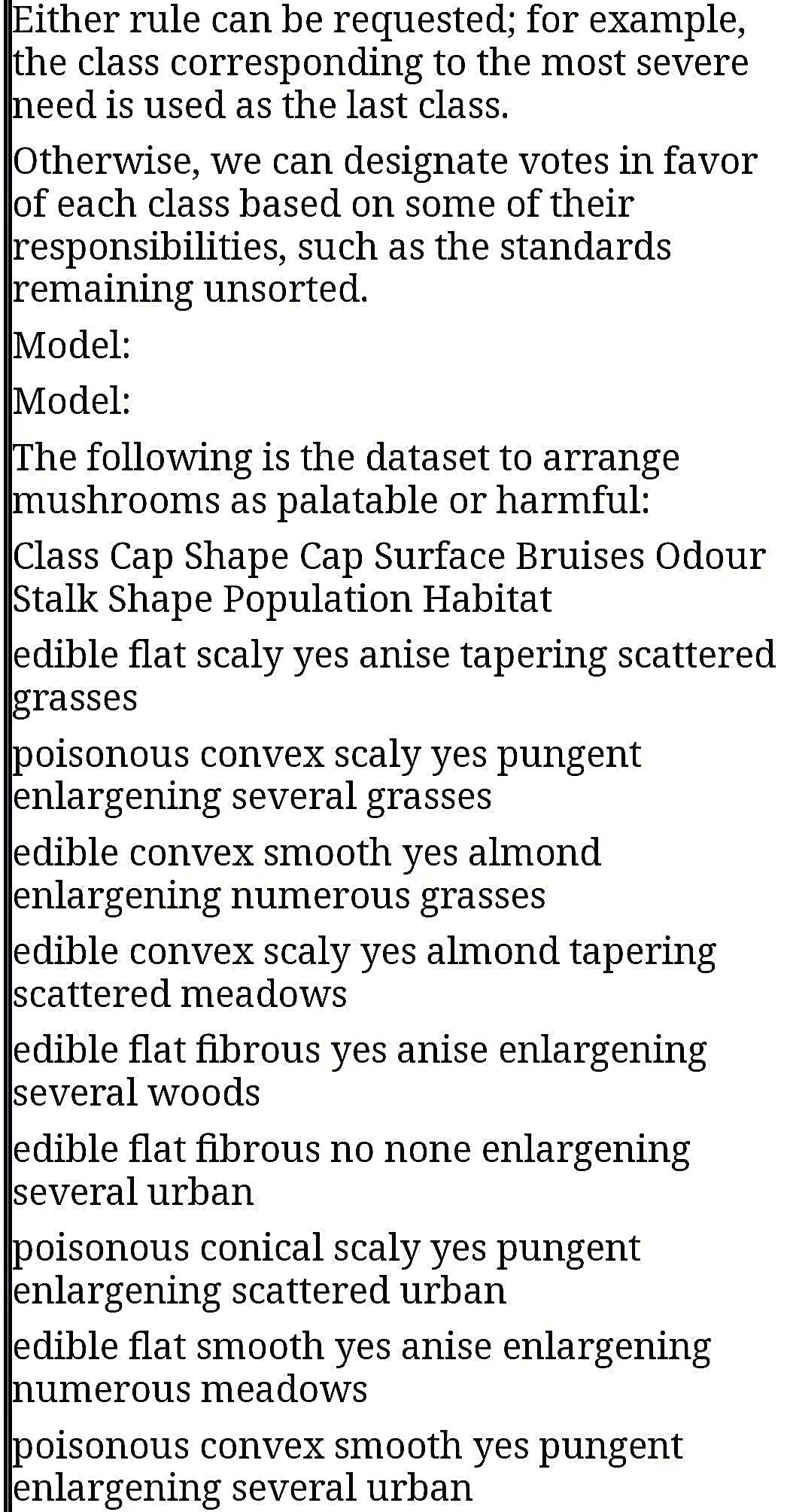
**Z-Score: =NORM.S.INV(C2)**

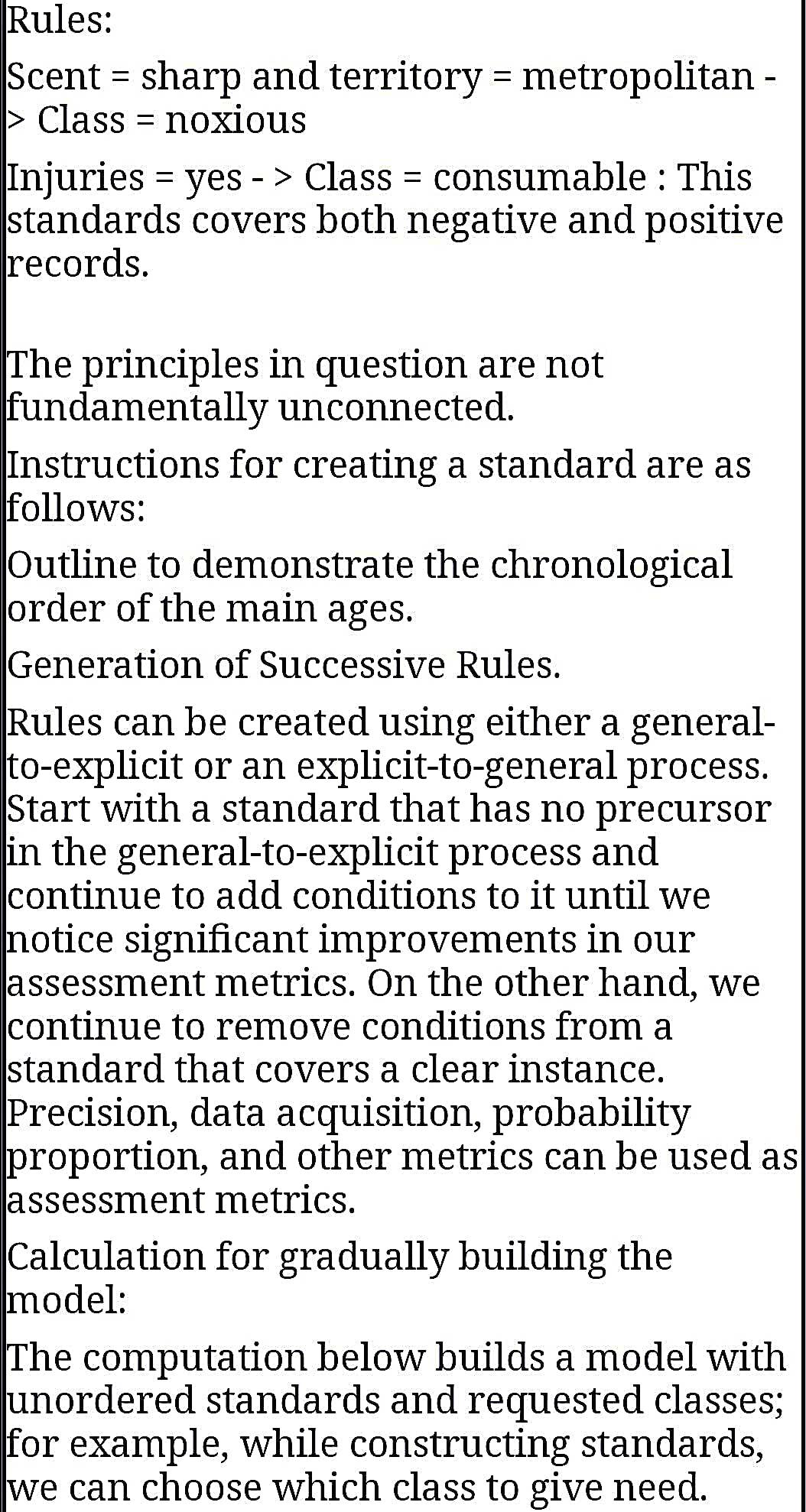
**Q4)**

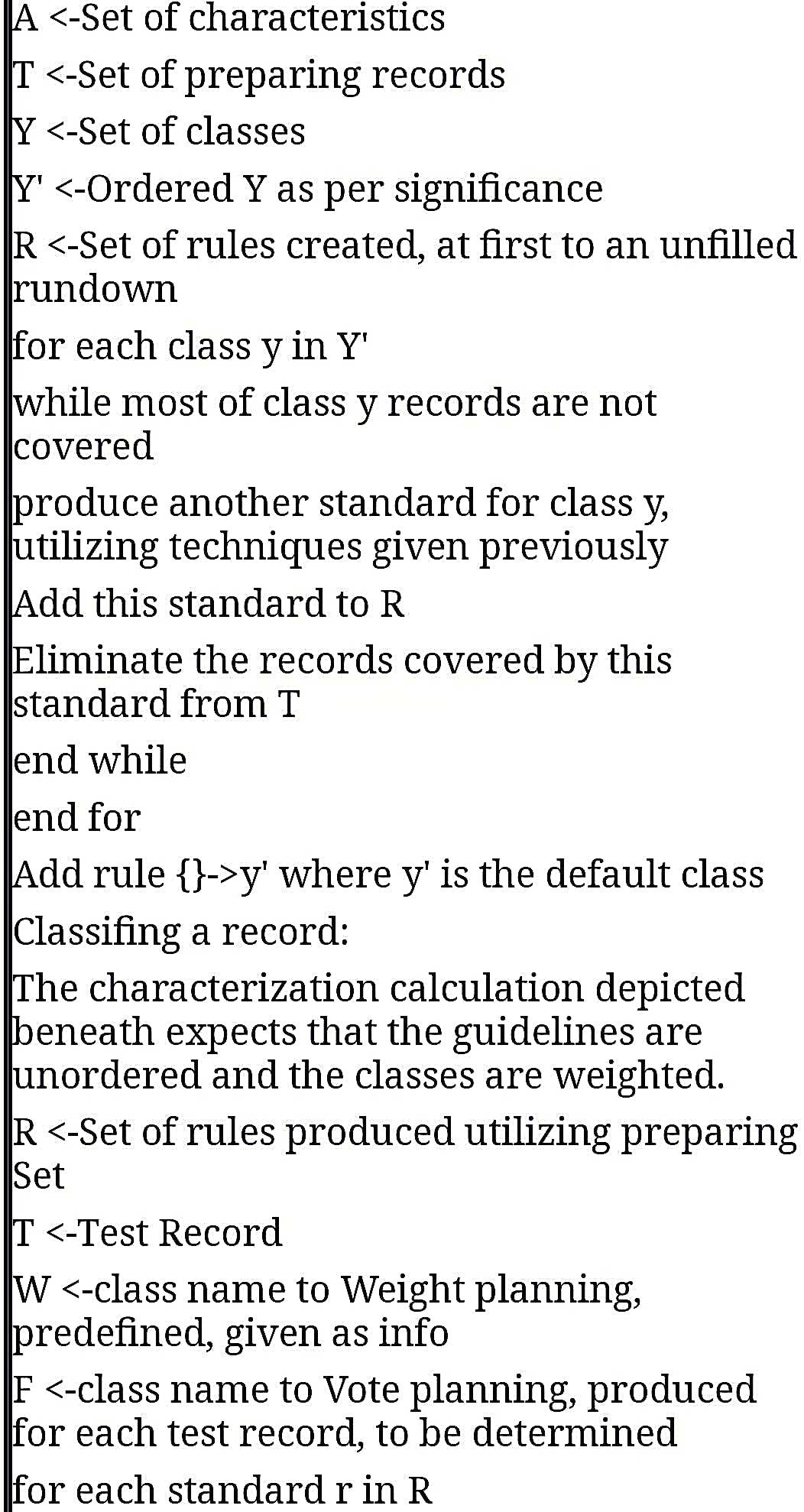


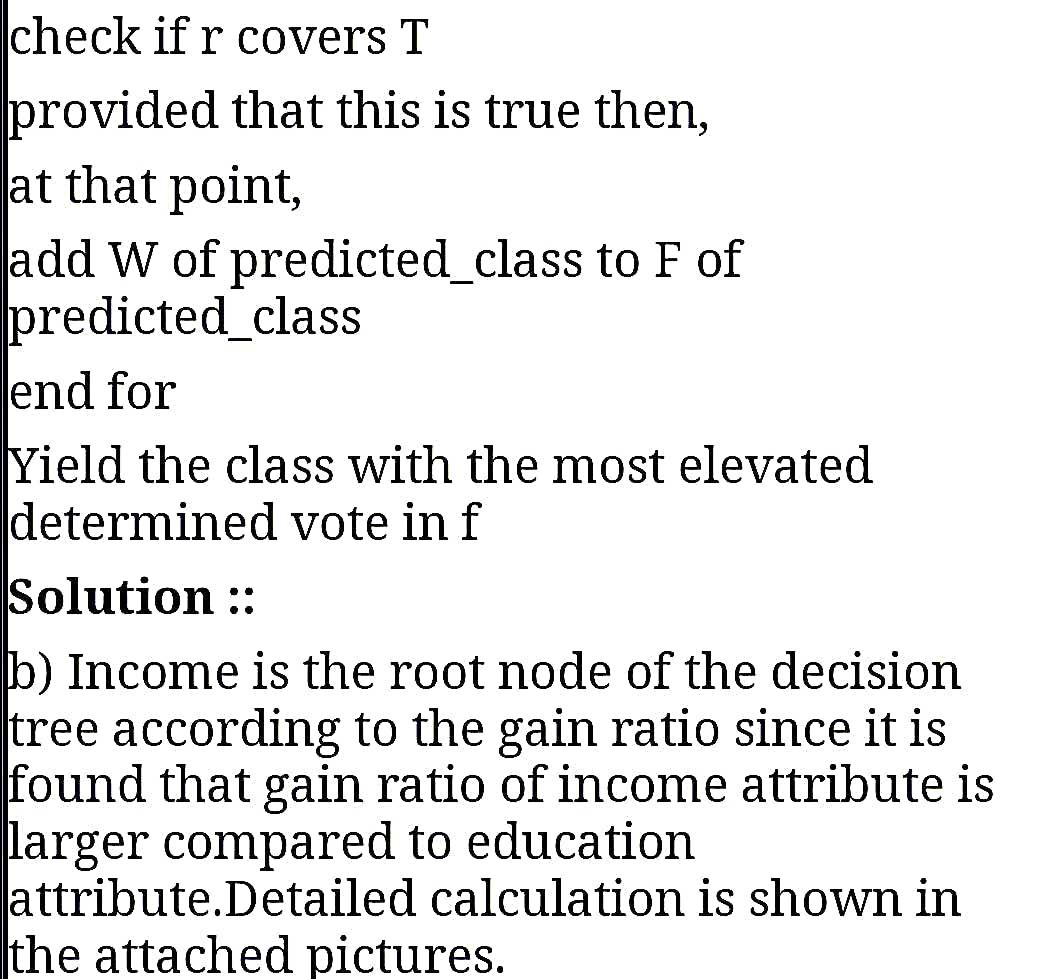
**Note: -** I AM Not sure for the answer

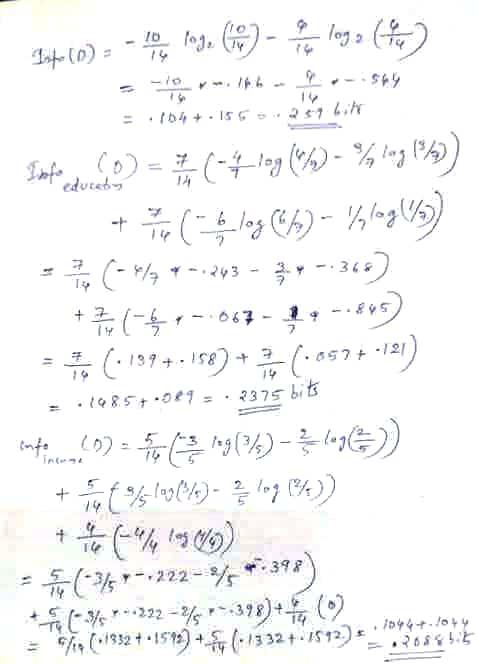


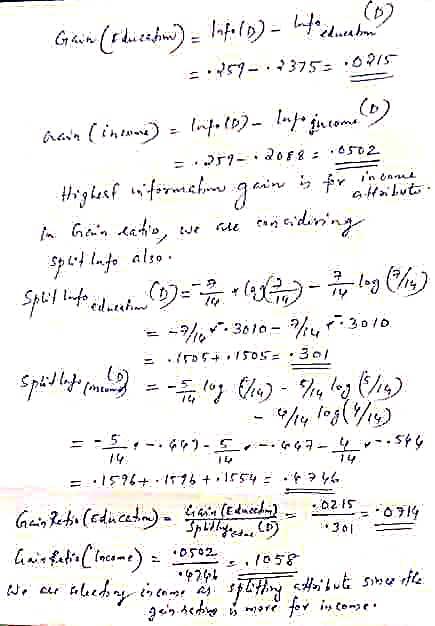












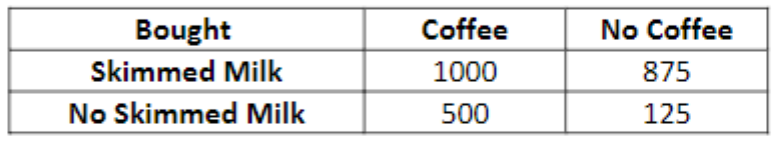
**Q 5. Answer the following:**

**a. Through some initial market study (captured in the contingency table below), a**

**marketing manager started promoting skimmed milk powder to the coffee buyers.**

**Do you agree with him? You answer should be justified with association mining**

**metrics. Assume thresholds for support = 30%, confidence = 60%.**



b. When do you prefer apriori algorithm over FP-tree for doing frequent pattern   
mining? (3 marks)

**Answer 5.**

From the given data

a) The total number of coffee buyers = 1000 + 500 = 1500 From the total 1500 purchases 1000 of them buy skimmed milk This constitutes to 66.7% of total coffee buyers.

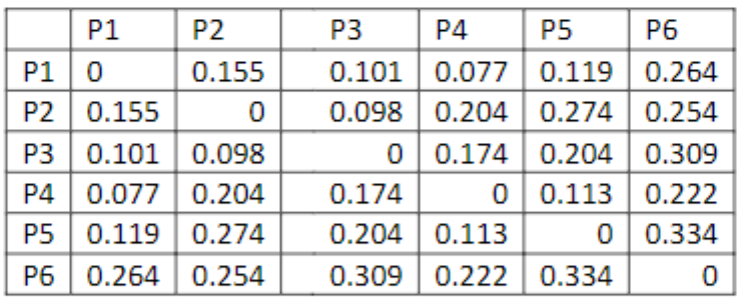
Hence, a marketing manager started promoting skimmed milk powder to the coffee buyers.

b) Apriori algorithm works more efficiently than FP-growth when the database consists of sparse/dense dataset.

In Apriori a generate candidate is required to get frequent itemsets. However FP-Growth generate candidate algorithm is not done because FP-Growth uses the concept of tree development in search of the frequent itemsets.

**Q 6) Answer the following:**

**a. We have the following distance matrix among 6 objects. Perform agglomerative hierarchical clustering with MIN approach and draw dendrogram. Show the intermediate steps. (4 marks)**



**b. You have been given a dataset of 800 objects. It is known that there are 8 natural clusters. You plan to perform K-means clustering starting with 8 random initial seeds? What is the probability that initial seeds come from distinct clusters? Comment on the results and consequences. [2+2 marks]**

**Answer 6)**

Agglomerative Hierarchical clustering using MIN distance measure

Distance between two clusters A and B:

d(A;B) = min{d(x; y) [1] x EA; y EB}

1. Dataset : {p1; p2; p3; p4; p5;p6}.

Initial clustering (singleton sets) C1: {p1}, {p2}, {p3}, {p4}, {p5},{p6}.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** |
| **P1** | 0 | 0.16 | 0.1 | **0.08** | 0.12 | 0.26 |
| **P2** | 0.16 | 0 | 0.1 | 0.2 | 0.27 | 0.25 |
| **P3** | 0.1 | 0.1 | 0 | 0.17 | 0.2 | 0.31 |
| **P4** | **0.08** | 0.2 | 0.17 | 0 | 0.11 | 0.22 |
| **P5** | 0.12 | 0.27 | 0.2 | 0.11 | 0 | 0.33 |
| **P6** | 0.26 | 0.25 | 0.31 | 0.22 | 0.33 | 0 |

In the above table, the minimum distance is the distance between the clusters {p1} and {p4}.

Also

d({p1}; {p4}) = 0.077

We merge {p1} and {p4}

Let us compute the distance of {p1; p4} from other clusters.

d({p1; p4}; {p2}) = min{d(p1; p2); d(p4; p2)} = min{0.155; 0.204} = 0.155

d({p1; p4}; {p3}) = min{d(p1; p3); d(p4; p3)} = min{0.101; 0.174} = 0.101

d({p1; p4}; {p5}) = min{d(p1; p5); d(p4; p5)} = min{0.119; 0.113} = 0.113

d({p1; p4}; {p6}) = min{d(p1; p6); d(p4; p6)} = min{0.264; 0.222} = 0.222

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **{p1,p4}** | **{p2}** | **{p3}** | **{p5}** | **{p6}** |
| **{p1,p4}** | 0 | 0.16 | 0.1 | 0.11 | 0.22 |
| **{p2}** | 0.155 | 0 | **0.098** | 0.27 | 0.25 |
| **{p3}** | 0.101 | **0.098** | 0 | 0.2 | 0.31 |
| **{p5}** | 0.113 | 0.27 | 0.2 | 0 | 0.33 |
| **{p6}** | 0.222 | 0.25 | 0.31 | 0.33 | 0 |

In the above table, the minimum distance is the distance between the clusters

{p2} and {p3} and d({p2}; {p3}) = 0.098

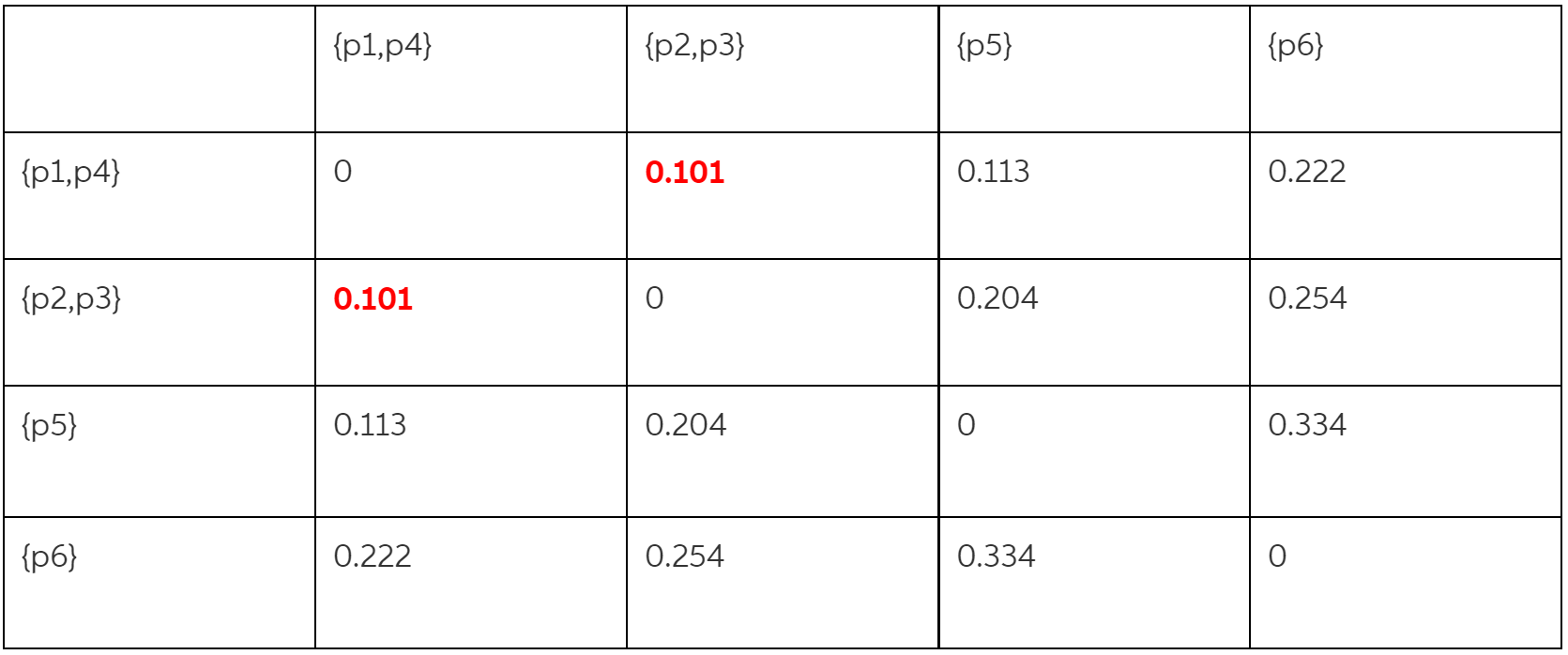
We merge {p2} and {p3}

Let us compute the distance of {p2; p3} from other clusters.

d({p2; p3}; {p1,p4}) = min{d(p2; p1), d(p2; p4), d(p3; p1),d(p3; p4} = min{0.155, 0.204, 0.101, 0.174} = 0.101

d({p2; p3}; {p5}) = min{d(p2; p5); d(p3; p5)} = min{0.274; 0.204} = 0.204

d({p2; p3}; {p6}) = min{d(p2; p6); d(p3; p6)} = min{0.254; 0.309} = 0.254



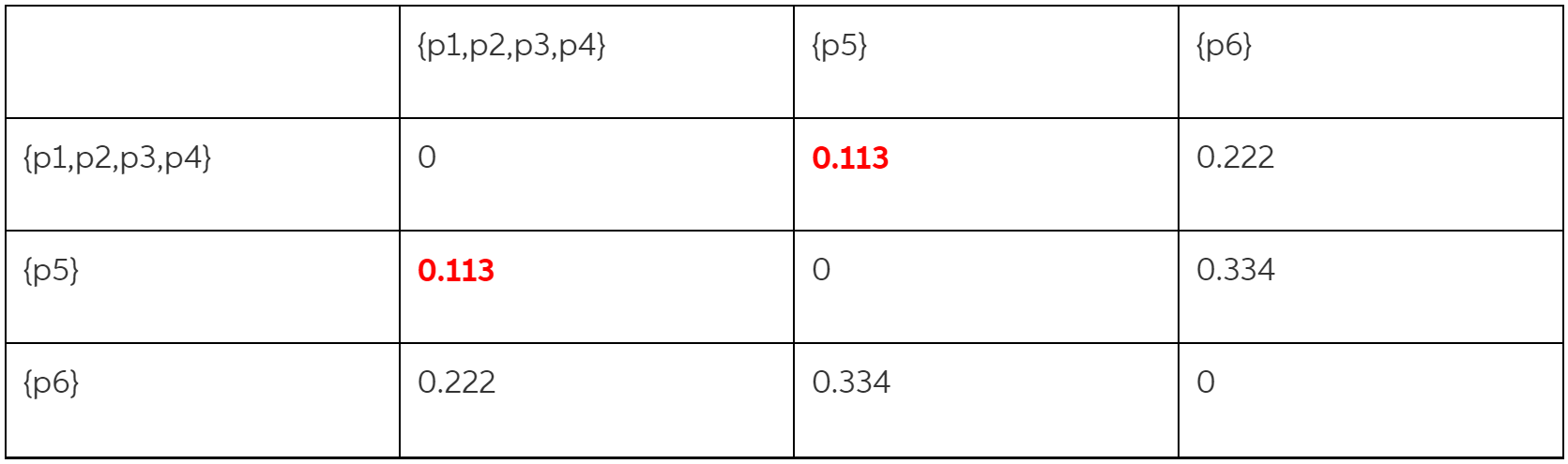
In the above table, the minimum distance is the distance between the clusters

{p1,p4} and {p2,p3} and d({p1,p4}; {p2,p3}) = 0.101 , merge them; {p1,p2,p3,p4}

Let us compute the distances

d({p1;p2; p3;p4}; {p5})=min{0.119,0.274,0.204,0.113}=0.113

d({p1;p2; p3;p4}; {p6})=min{0.264,0.254,0.309,0.222}=0.222



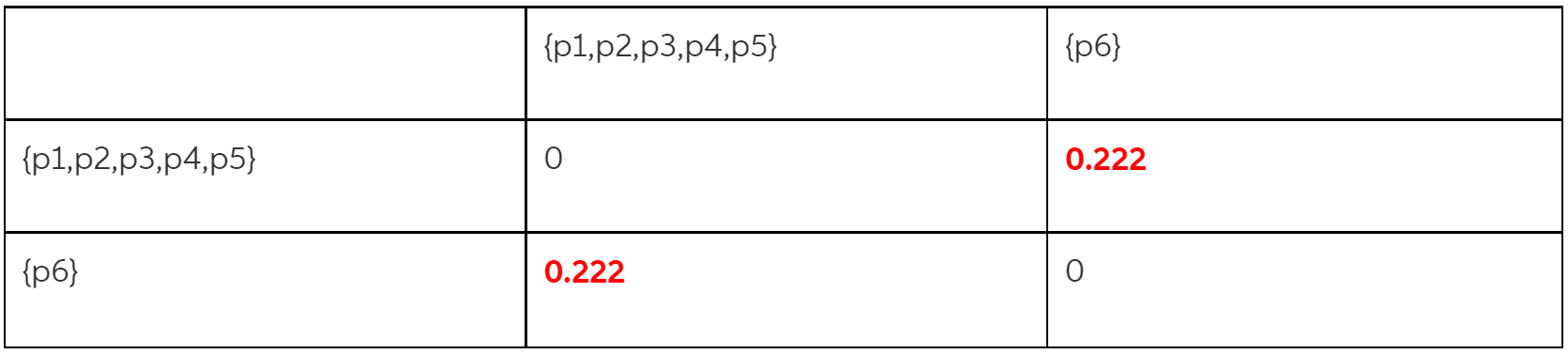
In the above table, the minimum distance is the distance between the clusters

{p1,p2,p3,p4} and {p5} and d({p1,p2,p3,p4},{p5})=0.113 ; merge them.

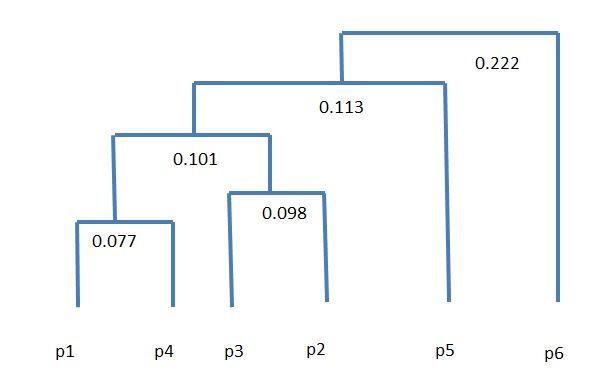
{p1,p2,p3,p4,p5}

Let us compute the distances

d({p1;p2; p3;p4,p5}; {p6})=min{0.264,0.254,0.309,0.222,0.334}=0.222



Only two clusters are left. We merge them form a single cluster containing all data points.

b)

Probability is about

8 Probability - 1.99 x 10-18 8000g