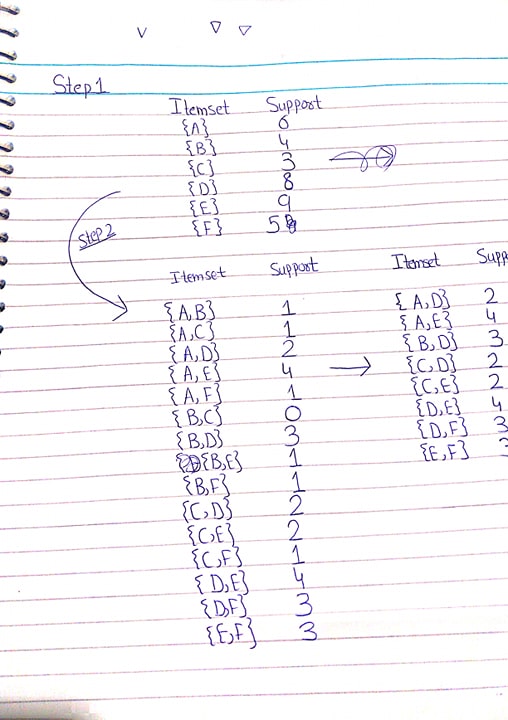
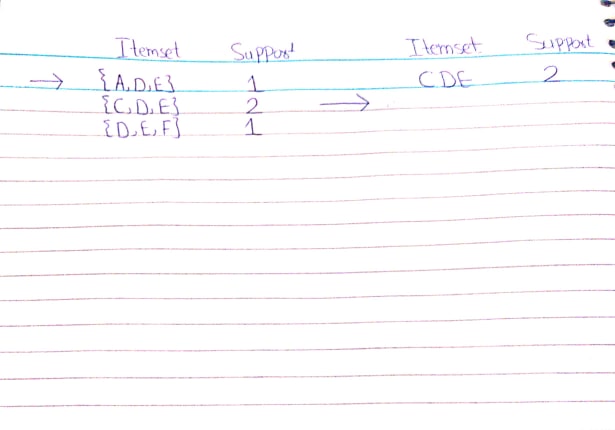
**Report**

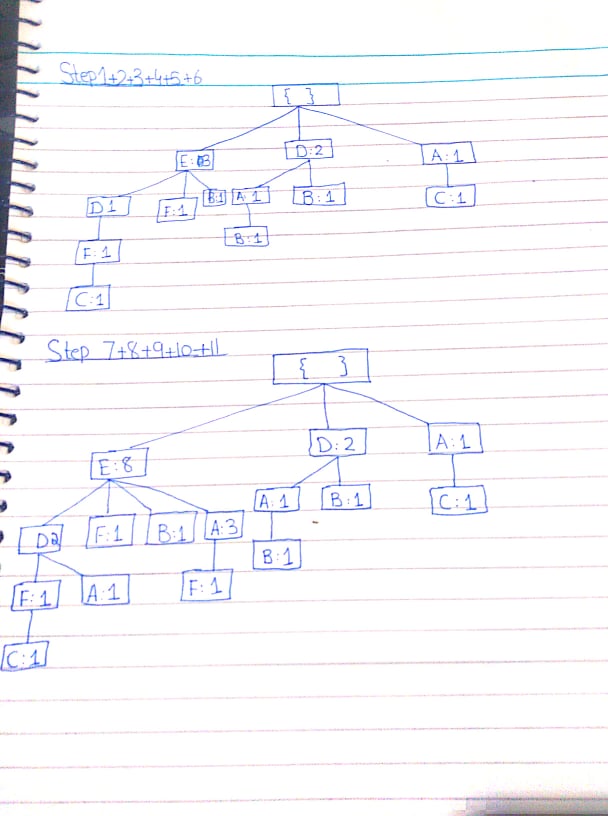
**Roll-Number:**2020-10-0004 **Name:** Dyass Khalid

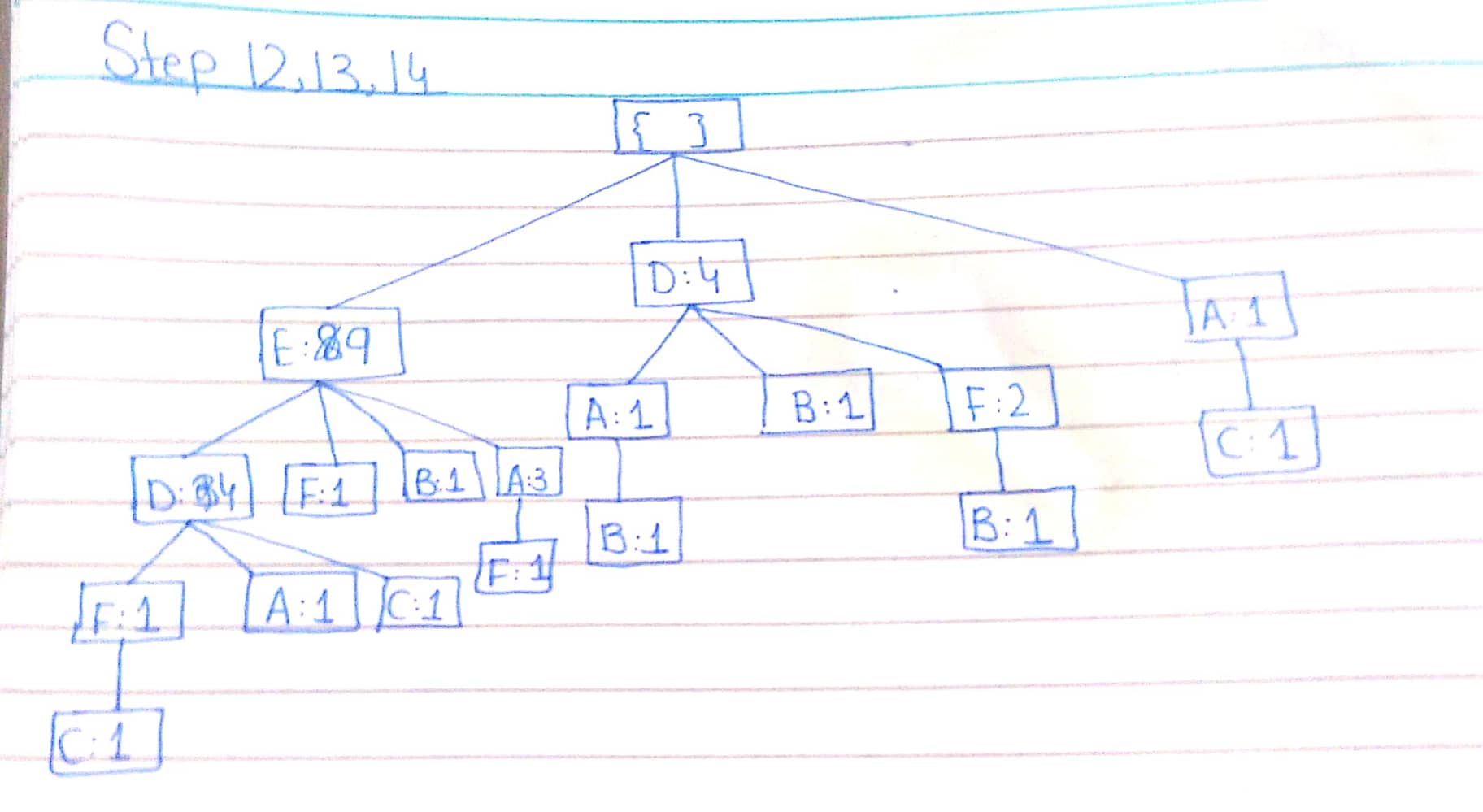
Question Number 1:



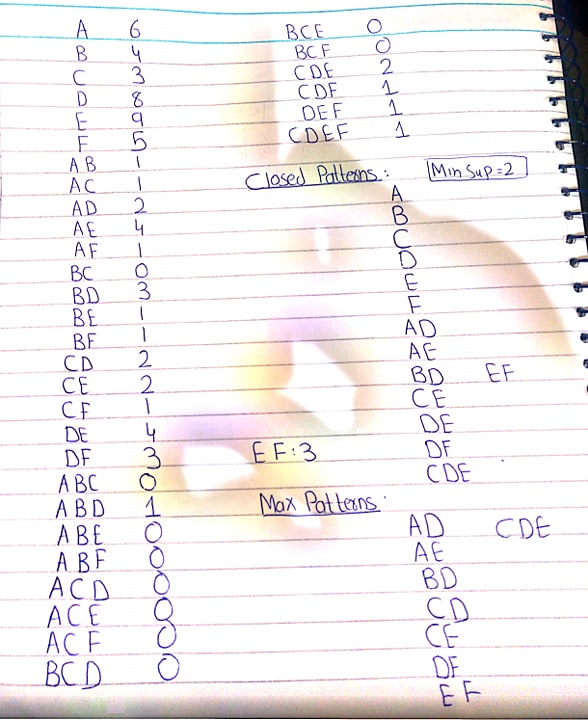


Part(b):

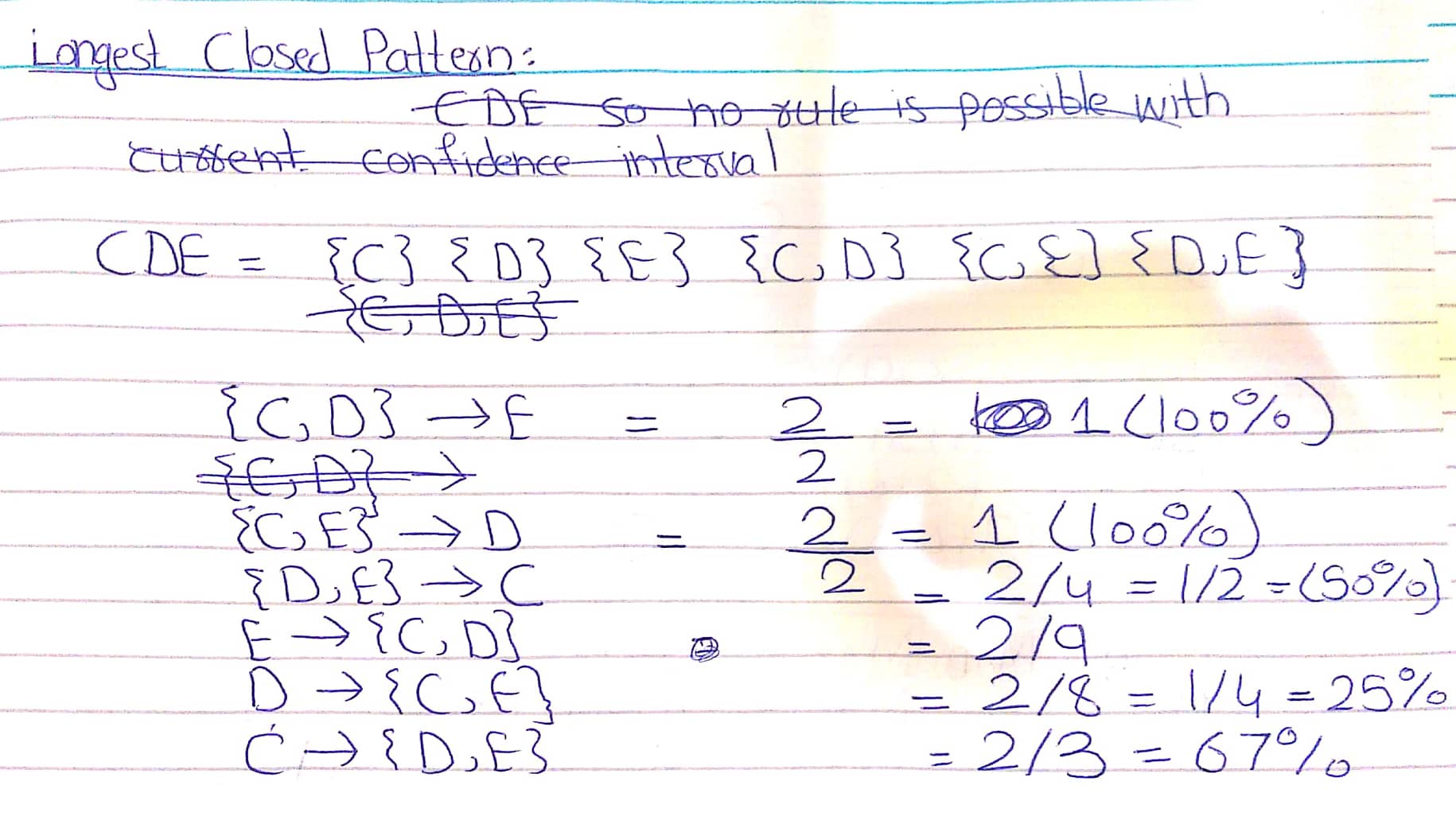




Part(c):



Part(d):



Question Number 2:

Part (a):

**1**:age=range2 [0 - 22.500] 5897 ==> marital-status=Never-married 5475 <conf:(0.93)>

If age belong to bin2 then with 93% confidence we can say that the marital status is never married.

**2**: 1. age=range2 [0 - 22.500] 5897 ==> marital-status=Never-married 5475 <conf:(0.93)>

If age belongs to bin 2 then with 93% confidence we can say that the marital status is never married

2. relationship=Own-child 7581 ==> marital-status=Never-married 6750 <conf:(0.89)>

If relationship is own child that with 89% confidence we can say that the marital status is never married

**3:** 1. workclass=? 2799 ==> occupation=? 2799 <conf:(1)>

Missing values in workclass and occupation are correlated

2. occupation=? 2809 ==> workclass=? 2799 <conf:(1)>

Missing values in workclass and occupation are correlated

3. age=range2 [0 - 22.500] relationship=Own-child 3955 ==> marital-status=Never-married 3883 <conf:(0.98)>

If age belongs to bin2 and relationship is own child then with 98% confidence we can say that the marital status is never married.

4. age=range2 [0 - 22.500] gender=Female 2794 ==> marital-status=Never-married 2616 <conf:(0.94)>

If age belongs to bin2 and gender is female then with 94% confidence marital status is never married.

5. age=range2 [0 - 22.500] 5897 ==> marital-status=Never-married 5475 <conf:(0.93)>

If age belong to bin2 then with 93% confidence marital status is never married.

Part (b):

1: The following interesting results are found out of many which can be seen in the rapid miner result files:

|  |  |
| --- | --- |
| 0.897 | United-States |
| 0.855 | White |
| 0.761 | <=50K |

There is a very high chance about 90% that the person in the data set belong to United states

There is a high chance about 85.5% that the person in the data set is white.

There is a chance bout 76.1% that the person income is <=50K

2:

|  |  |
| --- | --- |
| 0.700 | range3 [24.750 - 49.500] |
| 0.694 | Private |
| 0.668 | Male |

There is a chance of 70% that the person in data set belong to age bin 3

There is a chance of 69.4% that the educational institution is private.

There is a chance of 66.8% that the person in data set is male.

3:

|  |  |  |
| --- | --- | --- |
| 0.553 | <=50K | range3 [24.750 - 49.500] |
| 0.543 | <=50K | Private |
| 0.465 | <=50K | Male |

There is 55.3% chance that person income is less than 50k and he belong to age range 3

There is 54.3% chance that person income is less than 50k and his educational institution is private.

There is 46.5% chance that person icome is <=50K and he is male.

**C:**

No rule contains class attribute income on right side. This can be seen by running the result from rapid miner files.

**D:**

Different set of rules will be generated if we just change the bin max and min:

1. age=range2 [0 - 25] relationship=Own-child 5230 ==> marital- status=Never-married 5074 conf:(0.97)

2. relationship=Own-child 7581 ==> marital-status=Never-married 6750 conf:(0.89)

3. age=range2 [0 - 25] 9627 ==> marital-status=Never-married 8229 conf:(0.85)

4. marital-status=Never-married relationship=Own-child 6750 ==> age=range2 [0 - 25] 5074 <conf:(0.75)>

**Part 3:**

**1:**The code can be seen in the R code file where correlation matrix for each bin is calculated separately.

**Word freq\_address and word\_freq\_make are negatively coorelated in bin 1,2,4 whule it is positively correlated in bin 3.**

This behavior can be seen by comparing values in the correlation matrices generated in the R code.

**2:** If two values are highly correlated than we can simply take one and remove the other because they are representing the same predictive label such as date of birth and age of a person. High correlation simply means that we can take.

If you have two or more factors with a high VIF, remove one from the model. Because they supply redundant information, removing one of the correlated factors usually doesn't drastically reduce the R-squared.

We can simply use the PCA to cut down the number of highly correlated. This techniques helps in cutting the number of dimensions.