A picture containing logo

Description automatically generated

**Fundamental of Data Science**

**Assignment Four**

Reyad Melies

Part A: Association Rules:

1. Find all frequent item sets in database X.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | E | F | G | SUM |
| A | 5 | 3 | 3 | 4 | 1 | 2 | 1 | 19 |  |
| B | 3 | 4 | 2 | 2 | 0 | 1 | 2 | 14 |  |
| C | 3 | 2 | 5 | 4 | 1 | 2 | 3 | 20 |  |
| D | 4 | 2 | 4 | 6 | 1 | 4 | 3 | 24 |  |
| E | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 5 |  |
| F | 2 | 1 | 2 | 4 | 0 | 4 | 2 | 15 |  |
| G | 2 | 2 | 3 | 3 | 1 | 2 | 5 | 18 |  |
| SUM | 20 | 14 | 20 | 24 | 5 | 15 | 17 | 115 |  |

|  |  |
| --- | --- |
| A | 5 |
| B | 4 |
| C | 5 |
| D | 6 |
| E | 1 |
| F | 4 |
| G | 5 |

b)Find strong association rules for database:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  | support |
| A | 5 | 0.625 |
| B | 4 | 0.5 |
| C | 5 | 0.625 |
| D | 6 | 0.75 |
| E | 1 | 0.125 |
| F | 4 | 0.5 |
| G | 5 | 0.625 |

|  |  |
| --- | --- |
| A | 5 |
| B | 4 |
| C | 5 |
| D | 6 |
| F | 4 |
| G | 5 |

|  |
| --- |
| 0.25\*8=2 |

Records>=2 :Remove E

|  |  |
| --- | --- |
| A,B | 3 |
| A,C | 3 |
| A,D | 4 |
| A,F | 2 |
| A,G | 2 |
| B,C | 2 |
| B,D | 2 |
| B,F | 1 |
| B,G | 2 |
| C,D | 4 |
| C,F | 2 |
| C,G | 3 |
| D,F | 4 |
| D,G | 3 |
| F,G | 2 |

|  |  |
| --- | --- |
| A,B | 3 |
| A,C | 3 |
| A,D | 4 |
| A,F | 2 |
| A,G | 2 |
| B,C | 2 |
| B,D | 2 |
| B,G | 2 |
| C,D | 4 |
| C,F | 2 |
| C,G | 3 |
| D,F | 4 |
| D,G | 3 |
| F,G | 2 |

Records>=2

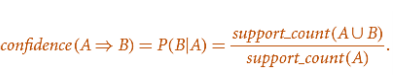
Remove {B,F}

|  |  |
| --- | --- |
| A,B,C | 1 |
| A,B,D | 2 |
| A,C,D | 3 |
| A,C,F | 1 |
| A,B,G | 1 |
| A,C,G | 1 |
| A,D,F | 2 |
| A,D,G | 1 |
| A,F,G | 0 |
| B,C,D | 1 |
| B,C,G | 1 |
| B,D,G | 0 |
| C,D,F | 2 |
| C,D,G | 2 |
| C,F,G | 1 |
| D,F,G | 2 |

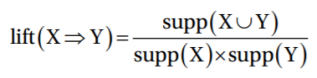
|  |  |
| --- | --- |
| A,B,D | 2 |
| A,C,D | 3 |
| A,D,F | 2 |
| C,D,F | 2 |
| C,D,G | 2 |
| D,F,G | 2 |

Remove { A,B,C },{ A,C,F},{ A,B,G },{ A,C,G },{ A,D,G },{ A,F,G },{ B,C,D},

{ B,C,G },{ B,D,G },{ C,F,G }

Rule Generation:

|  |  |  |
| --- | --- | --- |
| A,B->D | "2/3" | 66.6667 |
| A,D->B | "2/4" | 50 |
| B,D->A | "2/2" | 100 |
| A->A,B | "2/5" | 40 |
| B->A,D | "2/4" | 50 |
| D->A,B | "2/6" | 33.3333 |
| A,C->D | "3/3" | 100 |
| A,D->C | "3/4" | 75 |
| C,D->A | "3/4" | 75 |
| A->C,D | "3/4" | 75 |
| C->A,D | "3/5" | 60 |
| D->A,C | "3/6" | 50 |
| C,D->F | "2/4" | 50 |
| C,F->D | "2/2" | 100 |
| F,D->C | "2/4" | 50 |
| C->D,F | "2/5" | 40 |
| D->C,F | "2/6" | 33.333 |
| F->C,D | "2/4" | 50 |
| C,D->G | "2/4" | 50 |
| C,G->D | "2/2" | 100 |
| G,D->C | "2/3" | 66.66667 |
| C->D,G | "2/5" | 40 |
| D->C,G | "2/6" | 33.3333 |
| G->C,D | "2/5" | 40 |
| D,F->G | "2/4" | 50 |
| D,G->F | "2/3" | 66.6667 |
| F,G->D | "2/2" | 100 |
| D->F,G | "2/6" | 33.333 |
| F->D,G | "2/4" | 50 |
| G->D,F | "2/5" | 40 |
| A,D->F | "2/4" | 50 |
| A,F->D | "2/2" | 100 |
| DF->A | "2/4" | 50 |
| A->F,D | "2/5" | 40 |
| F->A,D | "2/4" | 50 |
| D->A,F | "2/6" | 33.3333 |



Choose confidence >=60

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Confidence | Lift |
|  |  |  |  |
| A,B->D | "2/3" | 66.6667 | 0.889 |
| B,D->A | "2/2" | 100 | 1.6 |
| A,C->D | "3/3" | 100 | 1.3 |
| A,D->C | "3/4" | 75 | 1.2 |
| C,D->A | "3/4" | 75 | 1.2 |
| A->C,D | "3/5" | 60 | 1.2 |
| C->A,D | "3/5" | 60 | 1.2 |
| C,F->D | "2/2" | 100 | 1.3 |
| C,G->D | "2/3" | 66.67 | 0.8889 |
| G,D->C | "2/3" | 66.666667 | 1.0667 |
| D,G->F | "2/3" | 66.6667 | 1.3 |
| F,G->D | "2/2" | 100 | 1.3 |
| A,F->D | "2/2" | 100 | 1.3 |

We have 13 rules driven.

1. Analyze misleading associations for the rule set obtained in (b).

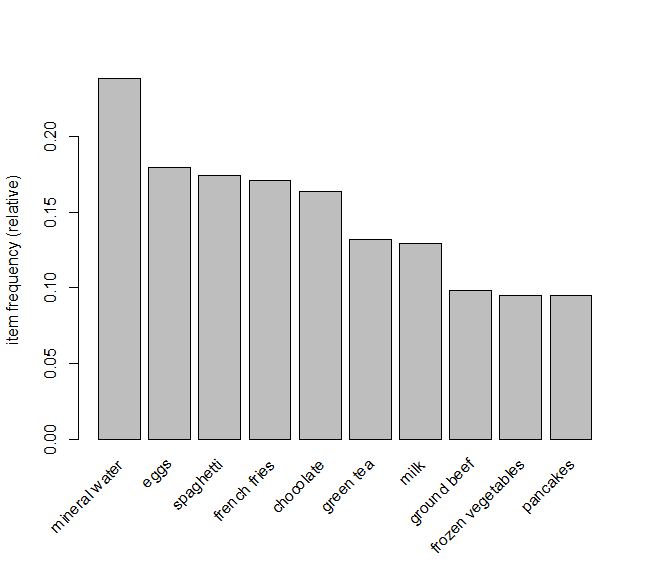
Therefore, we calculated lift. Events with high probability is likely to happen events related to event happening or not and can be even negatively associated. As Support & confidence are insufficient at filtering out uninteresting rules.

Calculating lift values which are less than 1 are misleading values.

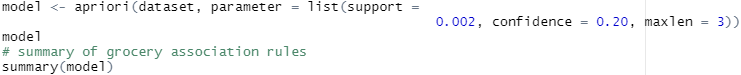
|  |  |  |  |
| --- | --- | --- | --- |
| A,B->D | "2/3" | 66.6667 | 0.889 |
| C,G->D | "2/3" | 66.67 | 0.8889 |

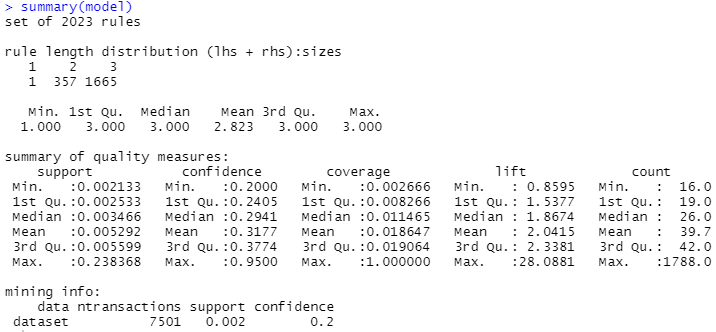
The rest are valid rules.

II. a) Generate a plot of the top 10 transactions.



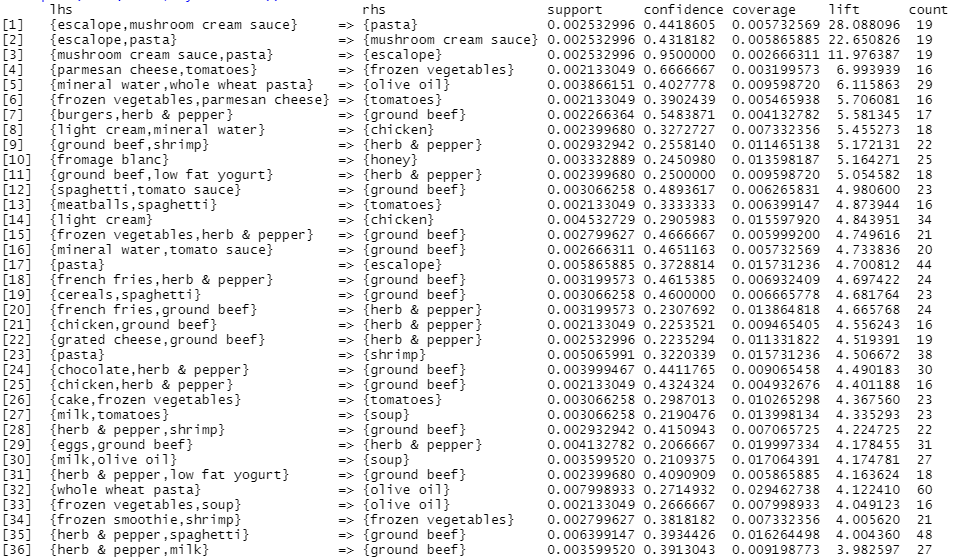
b) • Generate association rules using minimum support of 0.002, minimum confidence of 0.20, and maximum length of 3.





Display the rules, sorted by descending lift value.





Select the rule from Q1 with the greatest lift.



//////////////////////////////////////////////////////



///////////////////////////////////////////////////////////////////////////////

maximum length of 2.



Select the rule from Q1 with the greatest lift and length of 2:



///////////////////////////////////////////////////////////////////////////////



///////////////////////////////////////////////////////////////////////////////

1. Which rule has the better lift?

Rule 1 maxlen equal 3

Rule 1 has the better lift with value = 28.0881.

1. Which rule has the greater support?
2. Rule 2 maxlen equal 2

Rule 2 has the greater support with value = 0.003332889.

1. If you were a marketing manager, and could fund only one of these rules, which would it be, and why?

Rule 1 with maxlen equal 3 as it has better lift and confidence.

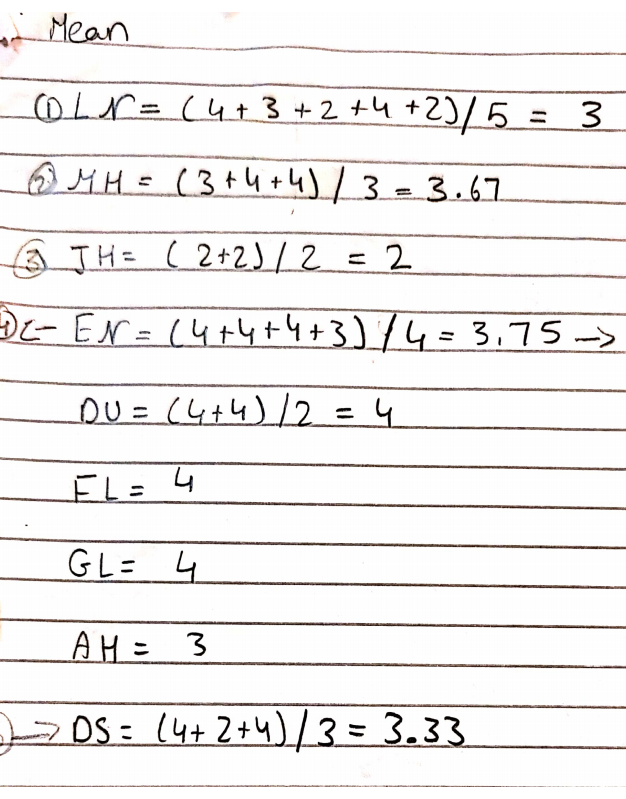
Lift gives better correlation measure ­­that judge how many times more often events occur together.

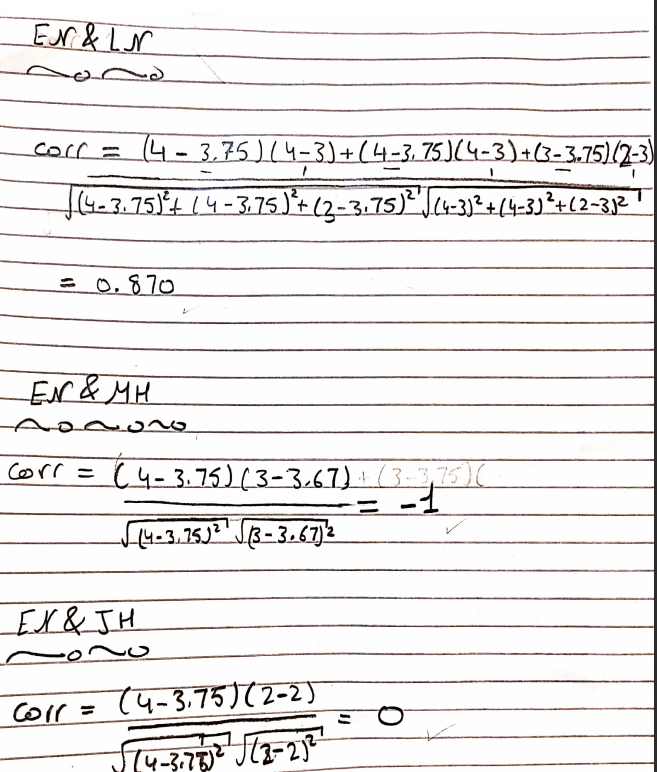
Confidence is a measurement of the predictive power and accuracy. So, the higher the better.

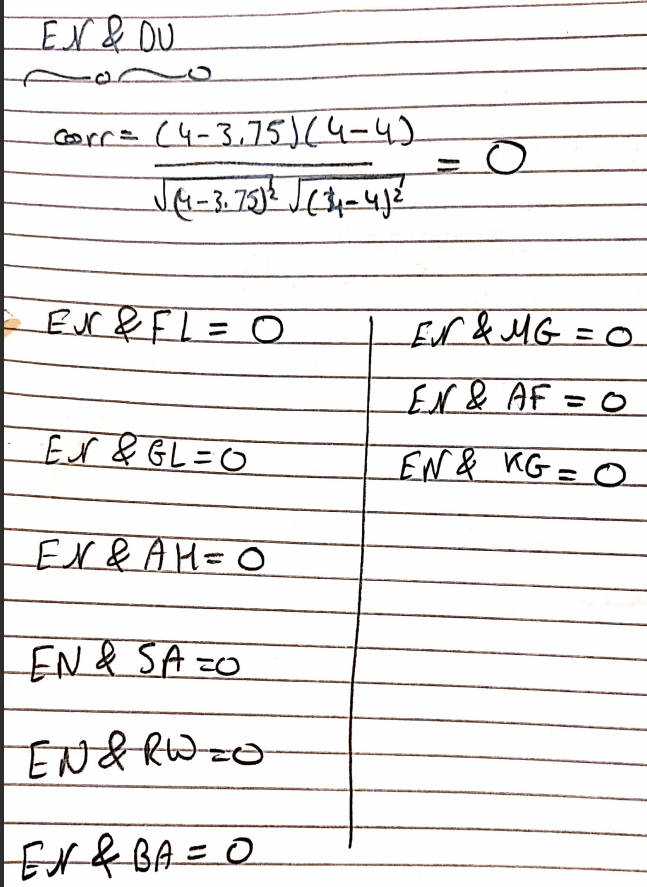
While confidence measures how frequently an itemset occurs the data set and due to the count is higher in rule 2 which will lead to better probability to happen together but not mean that they are reasons for happening together

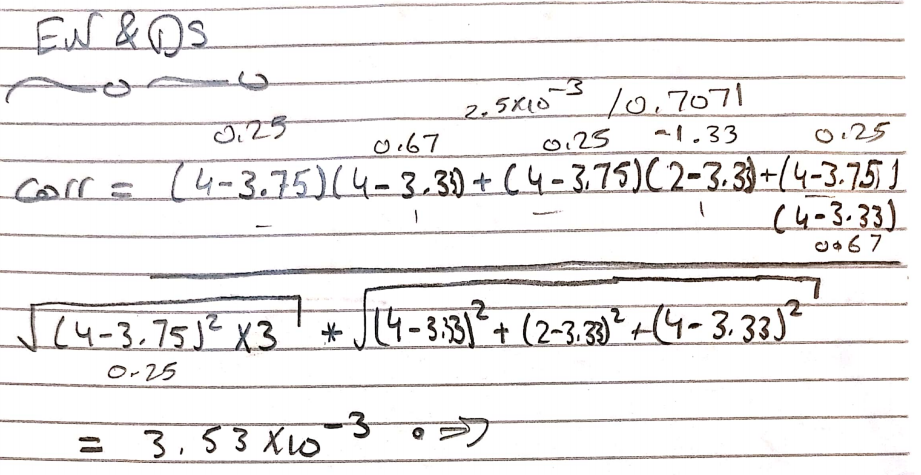
Part B: Course Recommender System using Collaborative Filtering:

First consider a user-based collaborative filter. This requires computing correlations between all student pairs. For which students is it possible to compute correlations with E.N.? Compute them.







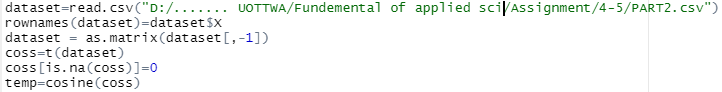


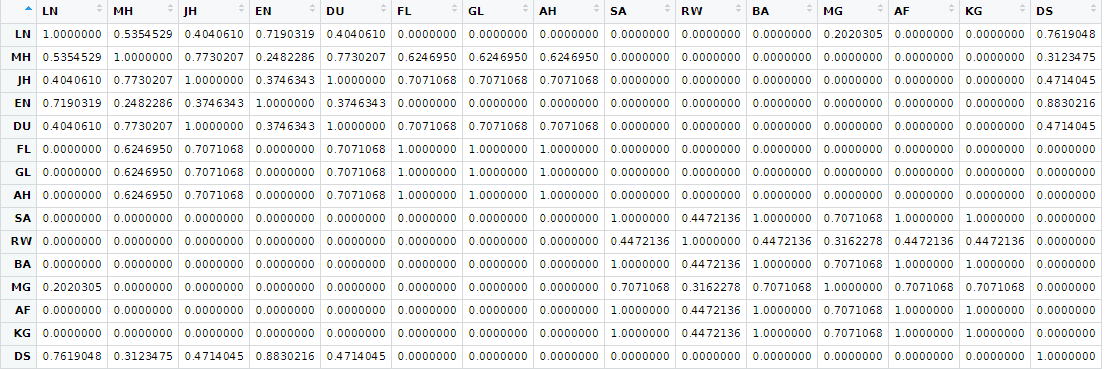
2) Based on the single nearest student to E.N., which single course should we recommend to E.N.? Explain why.

E.N and L.N had the highest correlation with value 0.870.

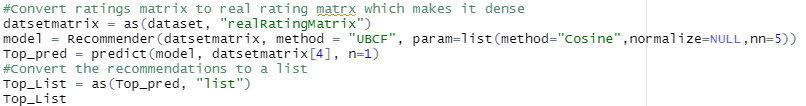
We have 2 choices Python or forecast but we should recommend Python as it had higher ratting.

3) Use R to compute the cosine similarity between users.





4) Based on the cosine similarities of the nearest students to E.N., which course should be recommended to E.N.?





5) Apply item-based collaborative filtering to this dataset

