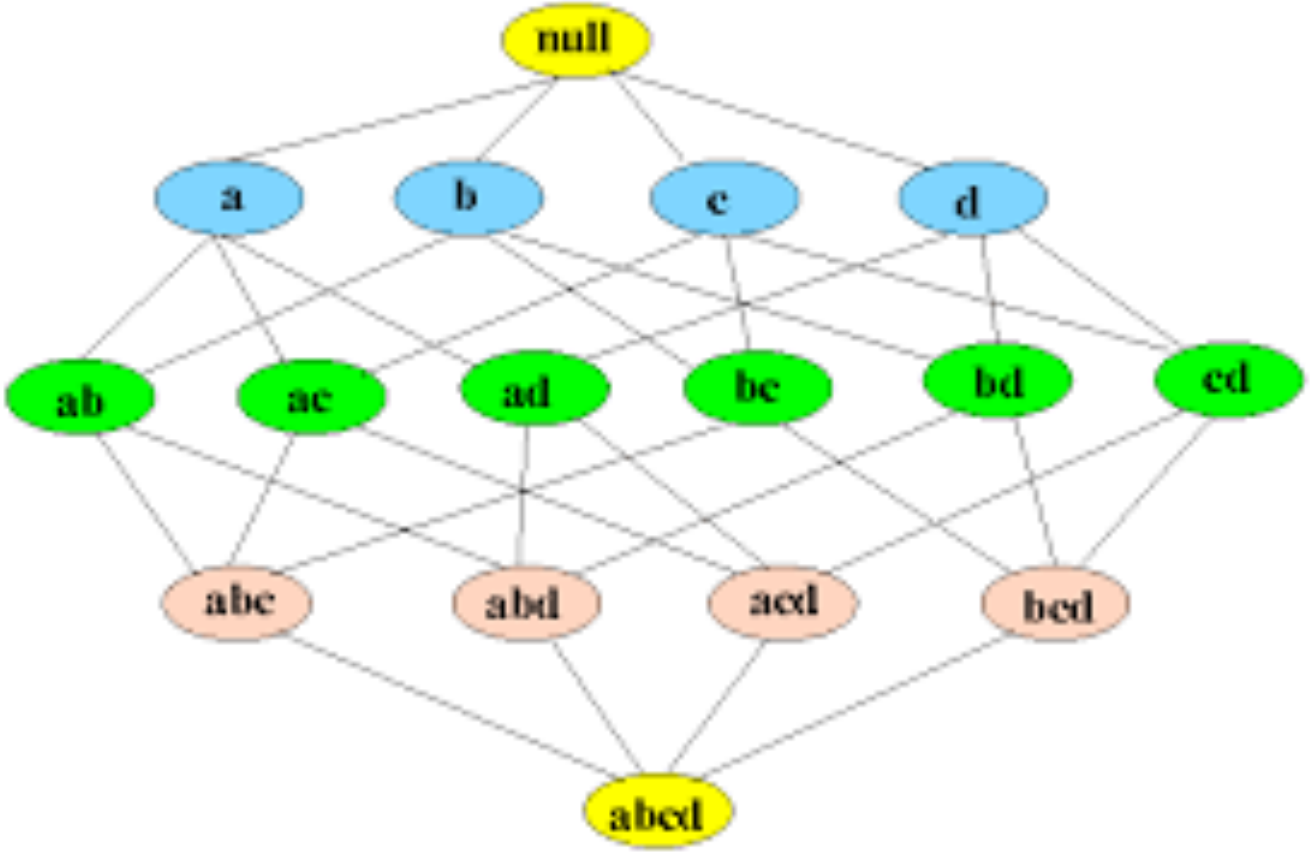
The impact of failing English/Academic Literacy: market basket analysis algorithms and applications to modules data.

Mr. Stanley Lekata & Dr. Elize Venter

Sefako Mokgatho Health Science University (SMU)

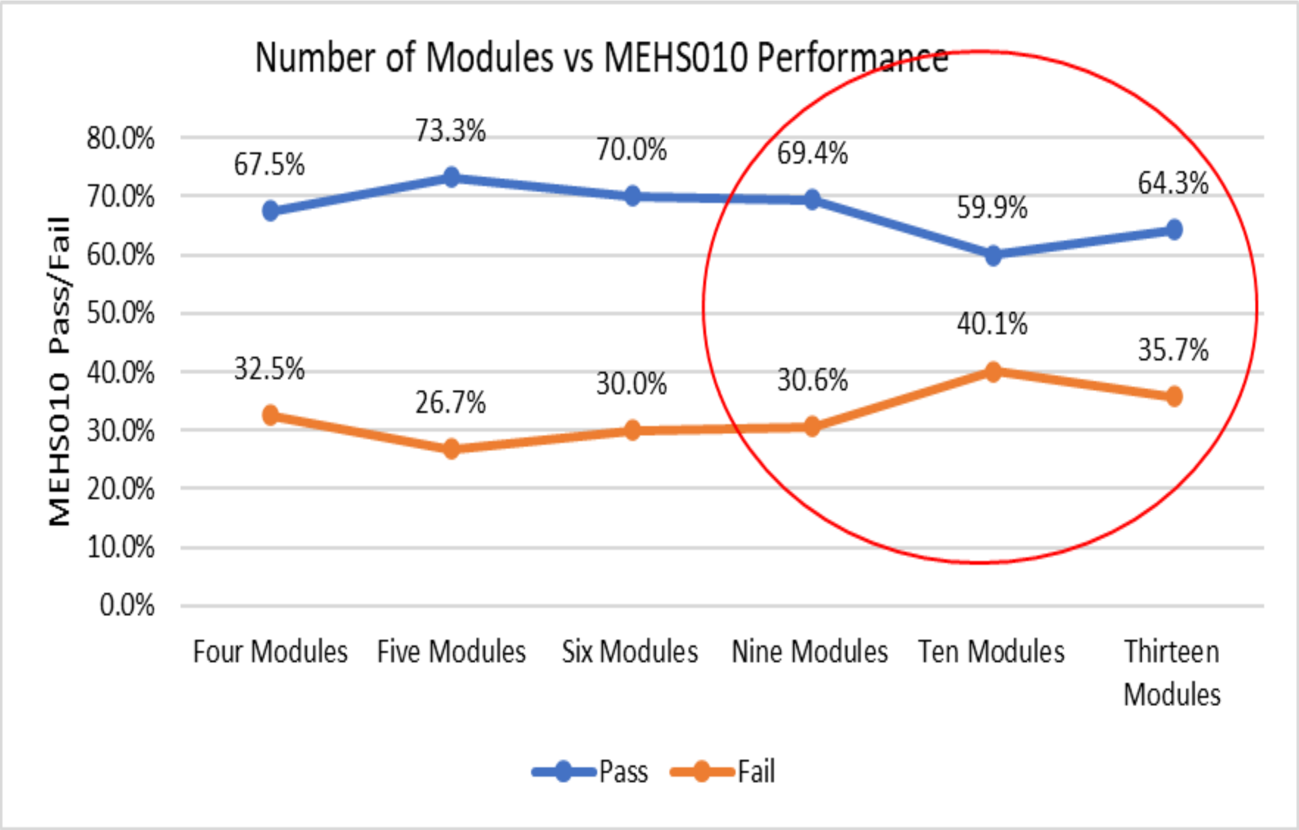


Previous work....

Relationship between success in English and Number of Modules enrolled

Number of Enrolled Modules vs MEHS010 Pass/Fail (%)

	MEHS010 _Pass/Fail		
Number of Modules	Pass	Fail	Grand Total
Four Modules	67.5%	32.5%	100.0%
Five Modules	73.3%	26.7%	100.0%
Six Modules	70.0%	30.0%	100.0%
Nine Modules	69.4%	30.6%	100.0%
Ten Modules	59.9%	40.1%	100.0%
Thirteen Modules	64.3%	35.7%	100.0%
Grand Total	71.0%	29.0%	100.0%



Number of Enrolled Modules vs MEHS010 Pass/Fail (%)

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	25.772a	5	<.001
Likelihood Ratio	24.827	5	<.001
N of Valid Cases	3619		

Data view on excel...Example

Student Num	English Pass/Fail	N Modules Enrolled
1	Pass	six
2	Fail	Ten
3	Pass	Eight
4	Fail	Eleven
5	Pass	six
6	Pass	seven
7	Fail	Thirteen

Reject the null hypothesis

Conclusion: there is a relationship between number of modules enrolled and performance in English.

Therefore, high number of Modules enrolled leads to higher chances of failing English.

Previous work....

Can performance in English Affect performance in other modules?

H_0: There is independence between Passing/Failing English and Performance in other modules

Variable 1 categories

X < 50% - Less than 50% modules passed

X>= 50% - 50% or modules passed

Performance in English categories

Pass

Fail

Data view on excel...Example

Student Num	English Pass/Fail	N Modules Passed	Narrative
1	Pass	X > 50%	Passed English, & passed over 50% of modules
2	Fail	X < 50%	Failed English, & Failed over 50% of modules
3	Pass	X > 50%	Passed English, & passed over 50% of modules
4	Fail	X < 50%	Failed English, & Failed over 50% of modules
5	Pass	X > 50%	Passed English, & passed over 50% of modules
6	Pass	X < 50%	Passed English, & passed over 50% of modules
7	Fail	X > 50%	Failed English, & but passed over 50% of modules

How English affects other Modules

MEHS010	X < 50%	X >= 50%	Grand Total
Pass	1.3%	98.7%	100.0%
Fail	37.3%	62.7%	100.0%
Grand Total	9.2%	90.8%	100.0%

Reject the Null hypothesis.

Conclusion: There is a relationship between Passing/Failing English and performance in other Modules.

When English is Passed, chances of passing other modules is higher.

Chi-Square Results

	Value	df	Asymptotic Significance	Exact Sig.(2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	963.831a	1	<.001		
Continuity Correction	959.539	1	<.001		
Likelihood Ratio	784.778	1	<.001		
Fisher's Exact Test				<.001	<.001
N of Valid Cases	3631				

‘When students fail English/Academic Literacy, which modules are they likely to fail?
When they pass English which modules are they likely to pass?’

The impact of failing English/Academic Literacy: a market basket analysis algorithms and applications to modules data.

Association rule & Apriori Algorithm

- Data mining algorithms studied extensively by database and data mining community.
- First proposed by Rakesh Agrawal, Tomasz Imielinski, and Arun Swami in 1993
- Assumes that all data is categorical
- Not good for numeric data
- Initially used for market ***Basket Analysis*** to find how items purchased by customers are related.
- **Motivation:** finding regularities in data
    - What kind of products are often purchased together?

Compliments & Substitute Goods

Complementary goods are products that are typically used together. They are goods that people tend to buy at the same time because they go well together or enhance each other's use.

Examples – Goods (Compliments)

- Smartphones and protective cases
- Printer and ink cartridges
- Cereal and milk
- Laptops and laptop cases

Examples – Academic Majors (Compliments)

- Mathematics and physics
- Biology and chemistry
- English and French
- Mathematics, physics and chemistry
- Statistics and economics
- History and development studies

Substitute goods are similar products that a customer may use for the same purpose.

Examples – Goods (substitutes)

- Butter and margarine
- Laptop computers and desktop computers
- Bottled spring water and bottled purified water
- Coffee and tea

Association rule

- Let $I = \{i_1, i_2, \dots, i_m\}$ be an itemset
    - Supermarket Example: $I = \{T - shirt, Trousers, Belt, Jacket, Gloves, Sneakers\}$, which is unique items in the store
    - Modules Example: $I = \{"MBPC010", "MCHM010", “MINM010", "MEHS010", MBLC010, MBEH010\}$, which is all failed first year dentistry modules.
- Each transaction T_n comprised of items $\{i_1, i_2, i_3 \dots i_n\}$, such that $T \subset I$, and each transection is a non-empty set.

- ………..
- Let X be a set of items. Then T_n is the transaction that is said to contain X if X⊆ T.
- Then, an association rule is defined as an implication of the form

$$X \Rightarrow Y, where\ X \subset I, Y \subset I\ and\ X \cap Y = 0$$

- In simple terms association rule is a relationship where $\{i_1, i_2\} \implies i_3$, such that the purchase of the antecedents

  implies the likely purchase of the consequence.
- For Example:
  - Purchase of **t-shirt & trousers** implies the likely purchase of **belt**:

    {T-shirt, Trousers} ⇒{Belt}

- Failing Mathematics implies likely failure of physics

  {Mathematics} ⇒{physics}

Transections – Modules Failed Example:

Trans.	Dentistry Students	Aca. Year	First Year Dentistry Modules Failed
t1	2023145651	2023	"**MEHS010**", "**MBEH010**","**MBLC010**"
t2	2023385655	2023	"**MBLC010**", "MINM010", **"MEHS010"**
t3	2023745654	2023	"MBPC010", "MCHM010”, “MINM010", "MICL010"
t4	2023748658	2023	"MBPC010", "MCHM010”, “MINM010", **"MEHS010"**
t5	2023345657	2023	**"MEHS010"**, "**MBEH010","MBLC010"**
t6	2023145700	2023	"**MBLC010**"
t7	2023845631	2023	"**MBEH010**", "**MBLC010**", "**MEHS010**"
t8	2023645663	2023	"MBPC010", "**MEHS010**”, “MICL010"
t9	2024345711	2023	"MCHM010", "MINM010", "**MBEH010"**
t10	2023385456	2023	"MBPC010", "**MEHS010**”, “MINM010"

Transactions: PnP a clothing store

Transaction	Items
t1	{T-shirt, Trousers, Belt}
t2	{T-shirt, Jacket}
t3	{Jacket, Gloves}
t4	{T-shirt, Trousers, Jacket}
t5	{T-shirt, Trousers, Sneakers, Jacket, Belt}
t6	{Trousers, Sneakers, Belt}
t7	{Trousers, Belt, Sneakers}

Support

- Support is an indication of how frequently the item set appears in the data set.
- In other words, it's the number of transactions(cases) with both X and Y divided by the total number of transactions.
- Examples - supermarket data
  - $supp(T - shirt \Rightarrow Trousers) = \frac{3}{7} = 43\%$
  - $supp(Trousers \Rightarrow Belt) = \frac{4}{7} = 57\%$
  - $supp(\{T - shirt, Trousers\} \Rightarrow \{Belt\}) = \frac{2}{7} = 28\%$
- Example - Modules data
  - Total Number of cases (students), N, is 10.

- $supp(MEHS010 \Rightarrow MBLC010) = \frac{4}{10} = 40\%$
- $supp(MBLC010 \Rightarrow MBEH010) = \frac{3}{10} = 30\%$
- $supp(\{MEHS010 \Rightarrow MBLC010\} \Rightarrow \{MBEH010\}) = \frac{3}{10} = 30\%$

Confidence (A => B)

- Confidence refers to the likelihood that an item B is also bought if item A is bought.
- It can be calculated by finding the number of transactions where A and B are bought together, divided by total number of transactions where A is bought.
- It is commonly depicted as

$$Confidence\ (A \Rightarrow B) = \frac{Transactions\ containing\ both\ (A\ and\ B)}{(Transactions\ containing\ A)} = \mathrm{P(B|A)} = \frac{Support\ (A \cap B)}{Support\ (A)}$$

- Example – Supermarket items
  - $conf(Trousers \Rightarrow Belt) = \frac{\left(\frac{4}{7}\right)}{\left(\frac{5}{7}\right)} = \frac{4}{7} * \frac{7}{5} = \frac{4}{5} = 80\%$
  - $conf(Trousers \Rightarrow Belt) = \frac{\left(\frac{4}{7}\right)}{\left(\frac{5}{7}\right)} = \frac{4}{7} * \frac{7}{5} = \frac{4}{5} = 80\%$
- Example – Modules items
  - $conf(MEHS010 \Rightarrow MBEH010) = \frac{\left(\frac{3}{10}\right)}{\left(\frac{7}{10}\right)} = \frac{3}{10} * \frac{10}{7} = \frac{3}{7} = 42.86\%$
  - $conf(\{MEHS010, MBEH010\} \Rightarrow \{MBLC010\}) = \frac{\left(\frac{3}{10}\right)}{\left(\frac{3}{10}\right)} = \frac{3}{10} * \frac{10}{3} = \frac{1}{1} = 100\ \%$

Lift (A => B)

- $Lift\ (A\ \ =>\ \ B)$ refers to the increase in the ratio of sale of B when A is sold.
- $Lift\ (A\ \ =>\ \ B)$ can be calculated by dividing Confidence (A -> B) by Support(B).
- Mathematically it can be represented as:

$$Lift(A => B) = \frac{Support\ (A \cap B)}{Support\ (A) * Support(B)} = \frac{Confidence\ (A => B)}{(Support\ (B))}$$

- A Lift of 1 means there is no association between products A and B.
- Lift of greater than 1 means products A and B are more likely to be bought together.
- Finally, Lift of less than 1 refers to the case where two products are unlikely to be bought together.

- Example supermarket data
  - $lift(T - shirt \Rightarrow Trousers) = \frac{\left(\frac{3}{7}\right)}{\left(\frac{4}{7}\right)*\left(\frac{5}{7}\right)} = 1.05$
  - lift ({T-shirt, Trousers} ⇒{Belt})= $\frac{\left(\frac{2}{7}\right)}{\left(\frac{3}{7}\right)*\left(\frac{4}{7}\right)}$ = 1.17
- Example Modules data
  - $lift(MEHS010 \Rightarrow MBEH010) = \frac{\left(\frac{3}{10}\right)}{\left(\frac{7}{10}\right)*\left(\frac{4}{10}\right)} = 1.071$
  - lift $(\{MEHS010, MBEH010\} \Rightarrow \{MBLC010\}) = \frac{\left(\frac{3}{10}\right)}{\left(\frac{3}{10}\right)*\left(\frac{5}{10}\right)}$ = 2

Apriori Algorithm

```
Apriori(T, ε)
    L₁ ← {large 1 - itemsets}
    k ← 2
    while Lₖ₋₁ is not empty
        Cₖ ← Apriori_gen(Lₖ₋₁, k)
        for transactions t in T
            Dₜ ← {c in Cₖ : c ⊆ t}
            for candidates c in Dₜ
                count[c] ← count[c] + 1

        Lₖ ← {c in Cₖ : count[c] ≥ ε}
        k ← k + 1

    return Union(Lₖ)

Apriori_gen(L, k)
    result ← list()
    for all p ∈ L, q ∈ L where p₁ = q₁, p₂ = q₂, ..., pₖ₋₂ = qₖ₋₂ and pₖ₋₁ < qₖ₋₁
        c = p ∪ {qₖ₋₁}
        if u ∈ L for all u ⊆ c where |u| = k-1
            result.add(c)
    return result
```

Apriori Algorithm

Trans.	items bought
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

C_1

Itemset	count
A	6
B	7
C	6
D	2
E	2

Min support count = 2

L_1

Itemset	count
A	6
B	7
C	6
D	2
E	2

C_2

Itemset
A, B
A, C
A, D
A, E
B, C
B, D
B, E
C, D
C, E
D, E

Itemset	Count
A, B	4
A, C	4
A, D	1
A, E	2
B, C	4
B, D	2
B, E	2
C, D	0
C, E	1
D, E	0

L_2

Itemset	Count
A, B	4
A, C	4
A, E	2
B, C	4
B, D	2
B, E	2

C_3

Itemset
A, B, C
A, B, D
A, B, E
A, C, D
A, C, E
A, D, E
B, C, D
B, C, E
B, D, E
C, D, E

Itemset	count
A, B, C	2
A, B, D	1
A, B, E	2
A, C, D	0
A, C, E	1
A, D, E	0
B, C, D	0
B, C, E	1
B, D, E	0
C, D, E	0

L_3

Itemset	count
A, B, C	2
A, B, E	2

C_4

Itemset
A, B, C, D
A, B, C, E
A, B, D, E
A, C, D, E
B, C, D, E

Itemset	count
A, B, C, D	0
A, B, C, E	1
A, B, D, E	0
A, C, D, E	0
B, C, D, E	0

L_4 IS EMPYT

Now rules can be build based on set L3, and accepted based on given confidence

Rules	Confidence = 50%	Rule Selected?
{A} ⟹ {B, C}	33.3%	x
{B} ⟹ {A, C}	28.6%	x
{C} ⟹ {A, B}	33.3%	x
{A, B} ⟹ {C}	50.0%	✓✓✓
{A, C} ⟹ {B}	50.0%	✓✓✓
{B, C} ⟹ {A}	50.0%	✓✓✓
{E} ⟹ {A, B}	100.0%	✓✓✓
{A} ⟹ {B. E}	33.3%	x
{B} ⟹ {A, E}	28.6%	x
{A, B} ⟹ {E}	50.0%	✓✓✓
{A, E} ⟹ {B}	100.0%	✓✓✓
{B, E} ⟹ {A}	100.0%	✓✓✓

Rules	Confidence = 50%	Rule Selected?
{A, B} ⟹ {C}	50.0%	✓✓✓
{A, C} ⟹ {B}	50.0%	✓✓✓
{B, C} ⟹ {A}	50.0%	✓✓✓
{E} ⟹ {A, B}	100.0%	✓✓✓
{A, B} ⟹ {E}	50.0%	✓✓✓
{A, E} ⟹ {B}	100.0%	✓✓✓
{B, E} ⟹ {A}	100.0%	✓✓✓

Lift and Conviction can be calculated on final rules.....

Real data. ....

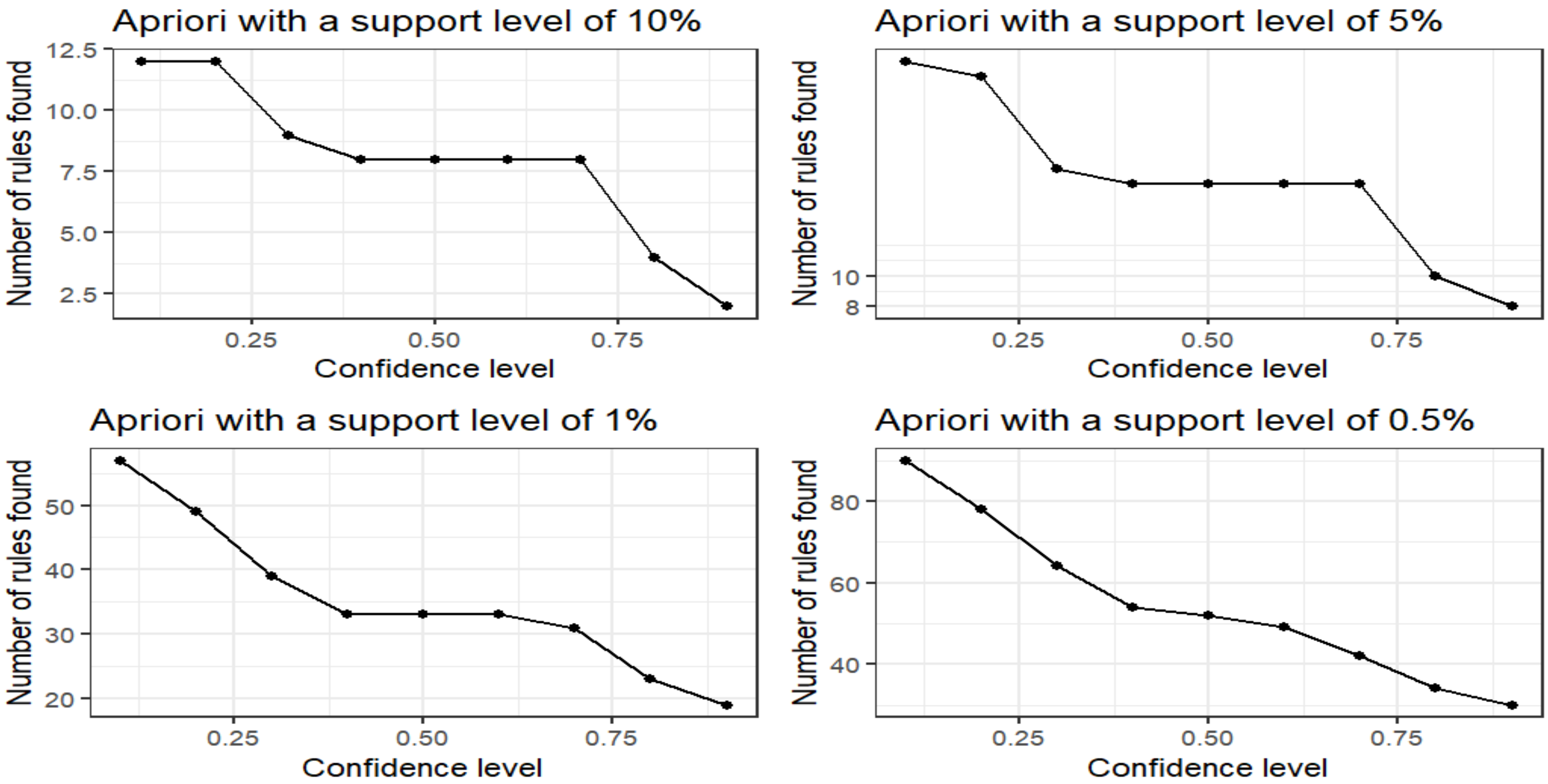
Number Of Rules at Different Support and Confidence Levels

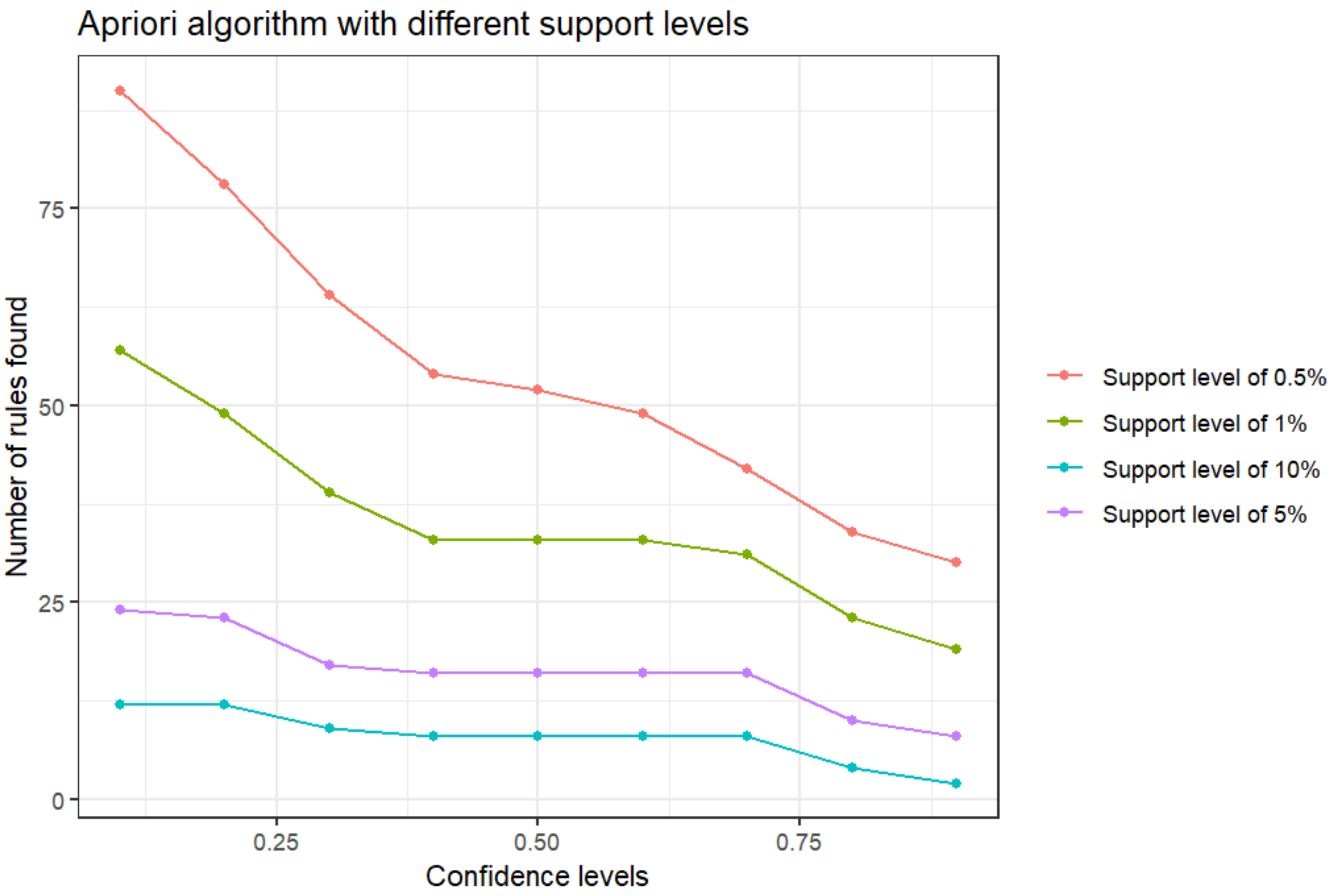
Support	Confidence	Rules	Comment
Not specified	Not specified	4	
0.1 (10%)	0.8 (80%)	4	too restrictive / strict
0.5 (50%)	0.5 (50%)	1	too restrictive / strict
0.01(1%)	0.5(50%)	33	Preferred
0.005 (0.5%)	0.005 (0.5%)	105	less restrictive / strict
0.01 (1%)	0.01 (1%)	66	preferred
0.01 (1%)	0.1 (10%)	57	preferred

First Year Modules used

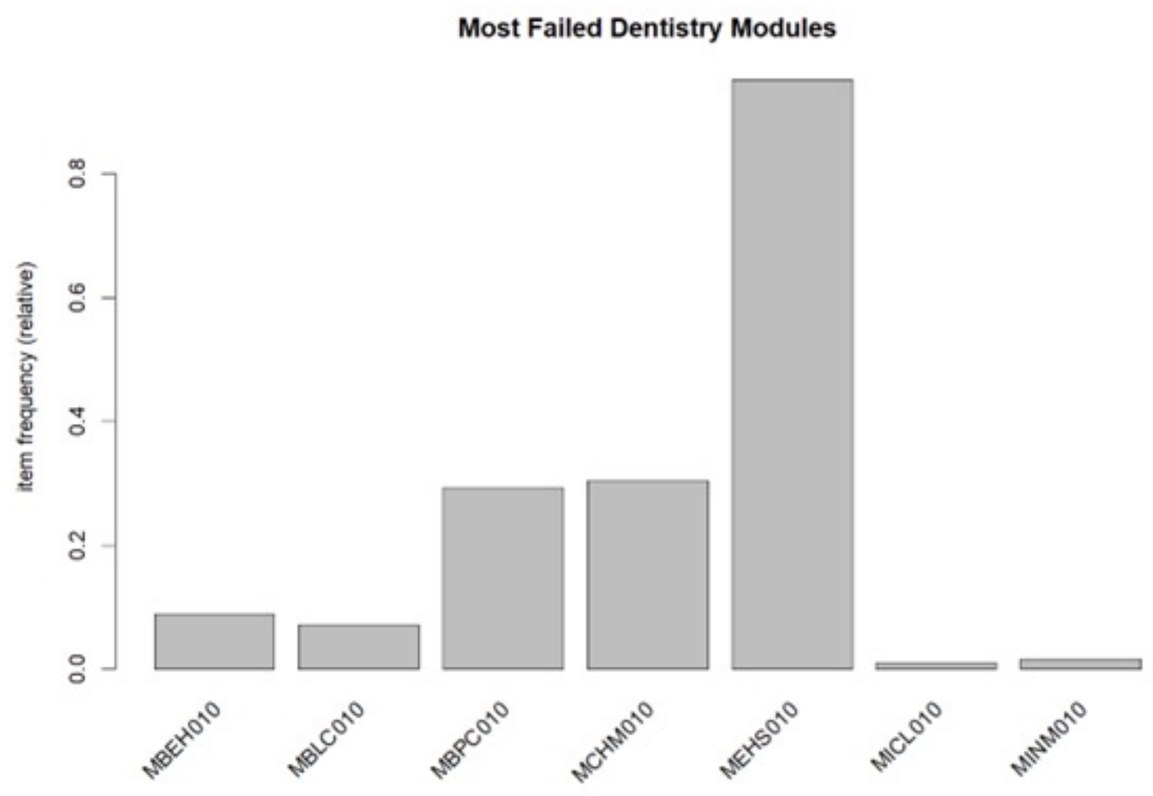
Module Code	Module Name
MEHS010	English for Health Sciences
MBEH010	Behavioural Sciences
MBLC010	Biology I
MBPC010	Biophysics I
MCHM010	Chemistry IA
MINM010	Introduction to Microbiology
MICL010	Integrated Clinical Dentistry I

Support & Confidence versus Number of rules




Apriori algorithm with different support levels
Number of rules found
75
50
25
0
0.25
0.50
0.75
Confidence levels
Support level of 0.5%
Support level of 1%
Support level of 10%
Support level of 5%

Most failed Dentistry Modules – Support = 1%

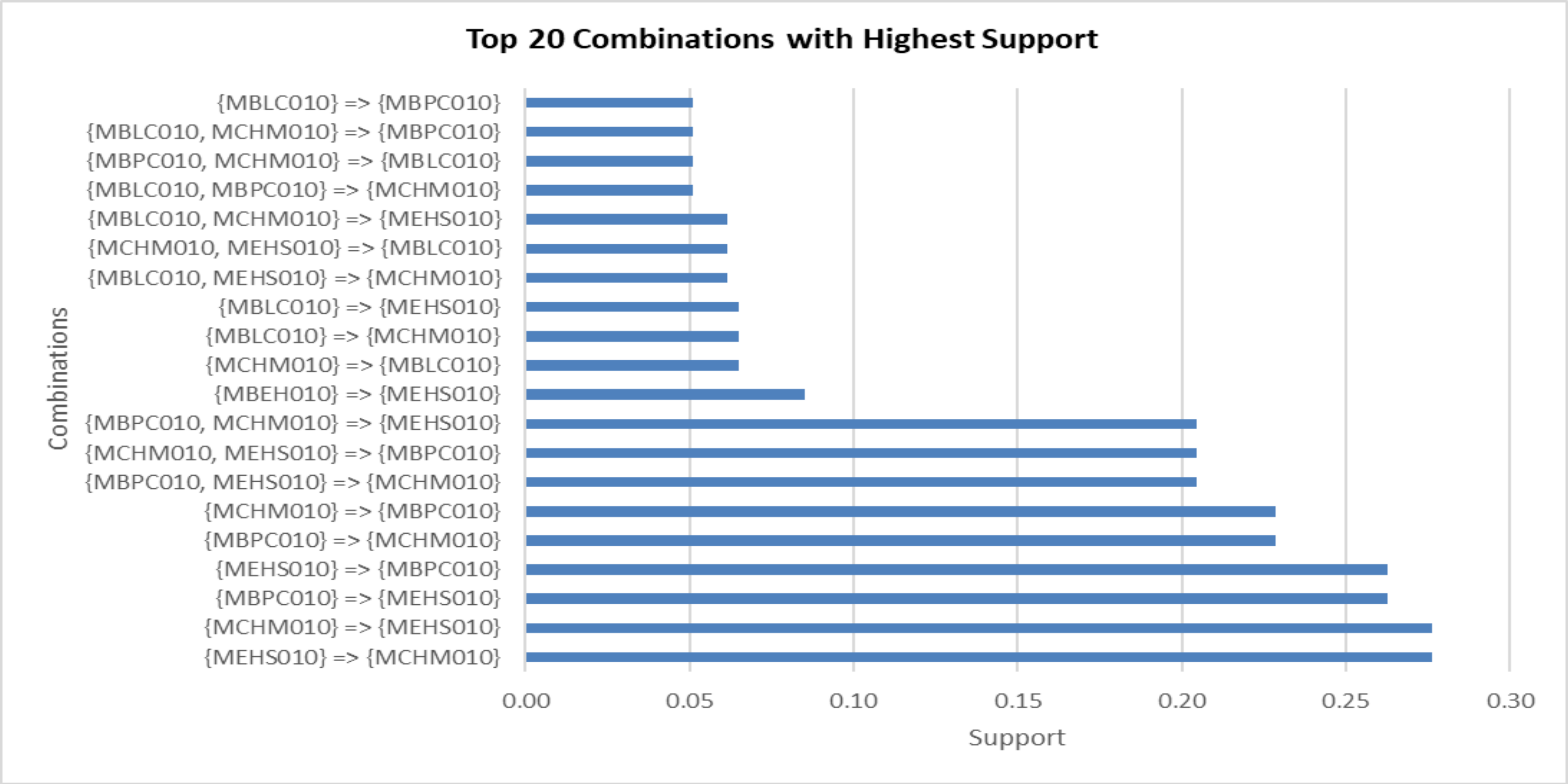


itemFrequencyPlot(df, support = 0.01, main = "Most Failed Dentistry Modules")

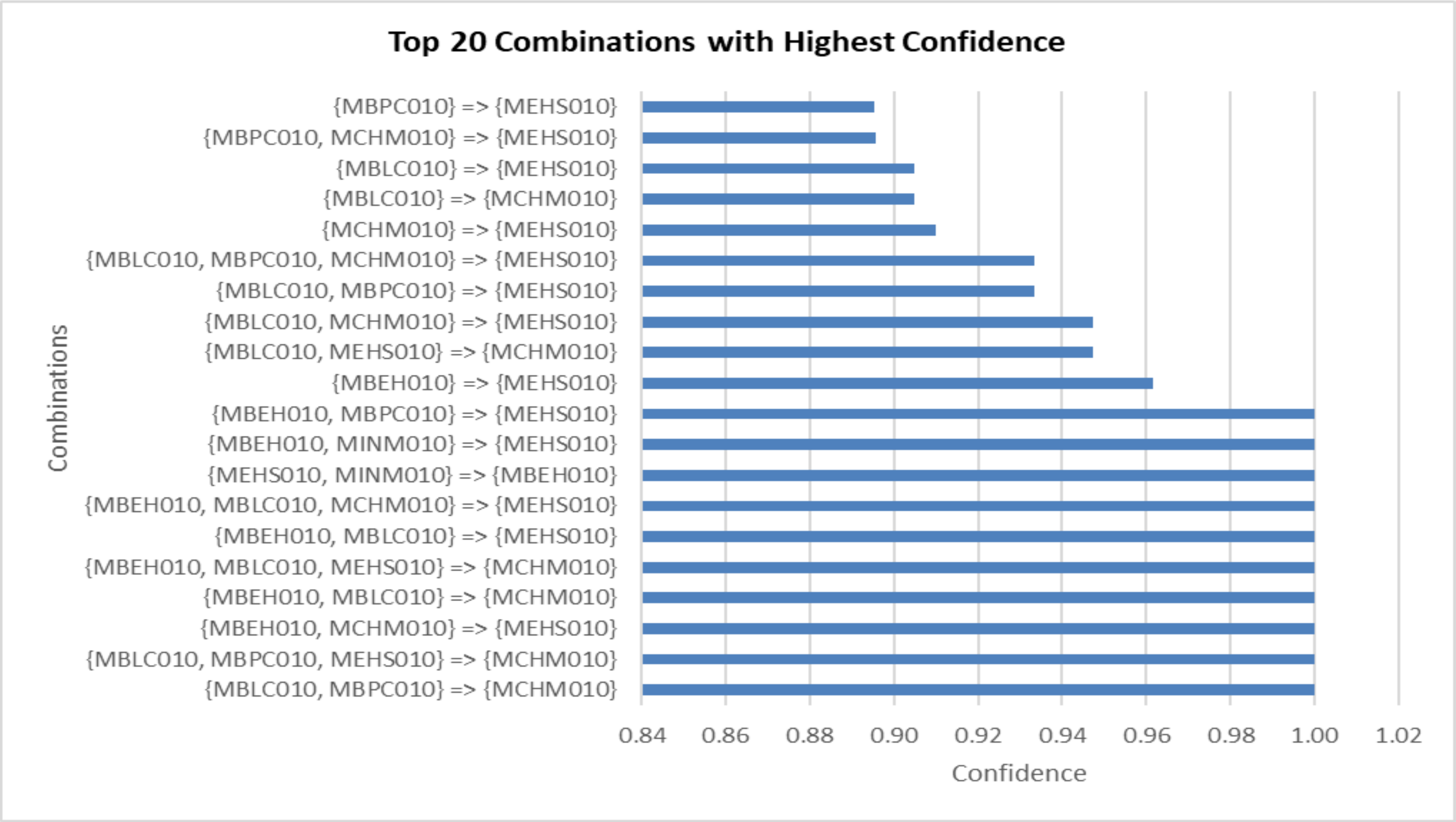
Section of Model Results – Sorted by Support

Row	lhs		rhs	support	confidence	coverage	lift	count
1	{MEHS010}	=>	{MCHM010}	0.2765	0.2903	0.9522	0.9558	81
2	{MCHM010}	=>	{MEHS010}	0.2765	0.9101	0.3038	0.9558	81
3	{MBPC010}	=>	{MEHS010}	0.2628	0.8953	0.2935	0.9403	77
4	{MEHS010}	=>	{MBPC010}	0.2628	0.2760	0.9522	0.9403	77
5	**{MBPC010}**	**=>**	**{MCHM010}**	**0.2287**	**0.7791**	**0.2935**	**2.5648**	**67**
6	**{MCHM010}**	**=>**	**{MBPC010}**	**0.2287**	**0.7528**	**0.3038**	**2.5648**	**67**
7	**{MBPC010, MEHS010}**	**=>**	**{MCHM010}**	**0.2048**	**0.7792**	**0.2628**	**2.5653**	**60**
8	**{MCHM010, MEHS010}**	**=>**	**{MBPC010}**	**0.2048**	**0.7407**	**0.2765**	**2.5237**	**60**
9	{MBPC010, MCHM010}	=>	{MEHS010}	0.2048	0.8955	0.2287	0.9405	60
10	{MBEH010}	=>	{MEHS010}	0.0853	0.9615	0.0887	1.0098	25
11	{MCHM010}	=>	{MBLC010}	0.0648	0.2135	0.3038	2.9786	19
12	{MBLC010}	=>	{MCHM010}	0.0648	0.9048	0.0717	2.9786	19
13	{MBLC010}	=>	{MEHS010}	0.0648	0.9048	0.0717	0.9502	19
14	{MBLC010, MEHS010}	=>	{MCHM010}	0.0614	0.9474	0.0648	3.1189	18
15	{MCHM010, MEHS010}	=>	{MBLC010}	0.0614	0.2222	0.2765	3.1005	18
16	{MBLC010, MCHM010}	=>	{MEHS010}	0.0614	0.9474	0.0648	0.9949	18
17	{MBLC010, MBPC010}	=>	{MCHM010}	0.0512	1.0000	0.0512	3.2921	15
18	{MBPC010, MCHM010}	=>	{MBLC010}	0.0512	0.2239	0.2287	3.1237	15
19	{MBLC010, MCHM010}	=>	{MBPC010}	0.0512	0.7895	0.0648	2.6897	15
20	{MBLC010}	=>	{MBPC010}	0.0512	0.7143	0.0717	2.4336	15

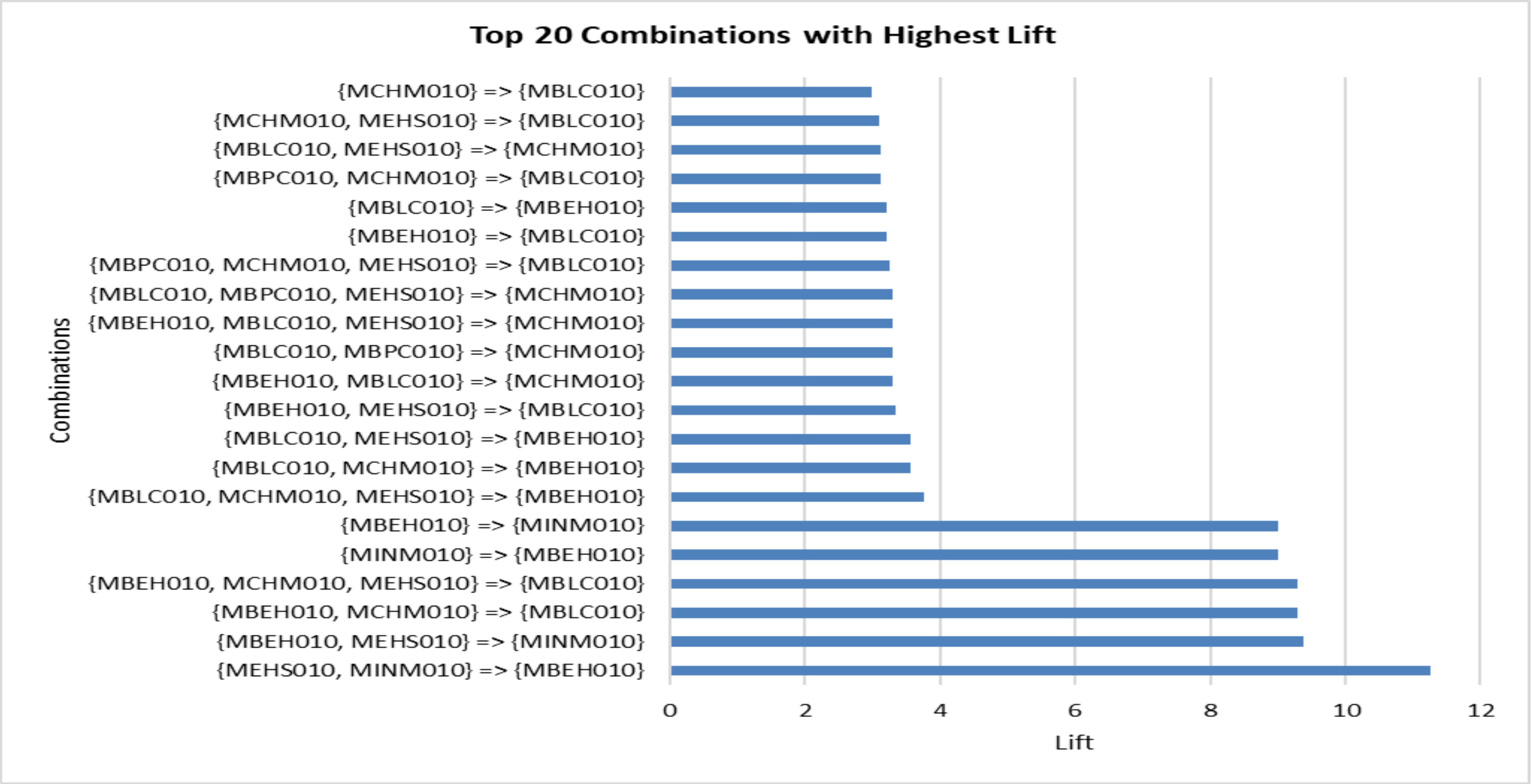
t7: The failure of Biophysics I & English {MBPC010, MEHS010} drives the failure of Chemistry IA {MCHM010}, and this shows in 20.48% of the data (Support), with likelihood of 78.9% (Confidence), and there is a positive relationship between failing {MBPC010, MEHS010} and {MCHM010} (support), and this combination is found in 60 transactions (Count)



English {MEHS010} and Chemistry 1A {MCHM010} have the highest support. This means that they are the most failed combination at Dentistry department first-year level



- The likelihood of failing English {MEHS010} when Biophysics I {MBPC010} is failed, is 89.53%
- The likelihood of failing ng English MEHS010 when Biophysics & Chemistry IA {MBPC010, MCHM010} is failed, is 89.55%



- English & Introduction to Microbiology {MEHS010, MINM010} together with Behavioural Sciences {MBEH010} have the highest lift. This implies they have the highest positive relationship

- When students fail English & Introduction to Microbiology, there are high chances of failing Behavioural Sciences

Conclusions....

- There is a relationship between number of modules enrolled and performance in English.

  High number of Modules enrolled leads to higher chances of failing English.

....

- There is a relationship between Passing/Failing English and performance in other Modules.

  When English is Passed, chances of passing over 50% of other modules is higher.

Conclusion… Failing English

lhs		rhs	support	confidence	coverage	lift	count
{MEHS010}	=>	{MCHM010}	0.277	0.290	0.952	0.956	81
{MEHS010}	=>	{MBPC010}	0.263	0.276	0.952	0.940	77
{MBPC010, MEHS010}	=>	{MCHM010}	0.205	0.779	0.263	2.565	60
{MBEH010}	=>	{MEHS010}	0.085	0.962	0.089	1.010	25

Example…

The failure of English {MEHS010} drives the failure of Chemistry IA {MCHM010}, and this shows in 27.7% of the data (Support), with likelihood of 29% (Confidence), and there is an inverse between failing {MEHS010} and {MCHM010} (lift), and this combination is found in 81 transactions (Count)

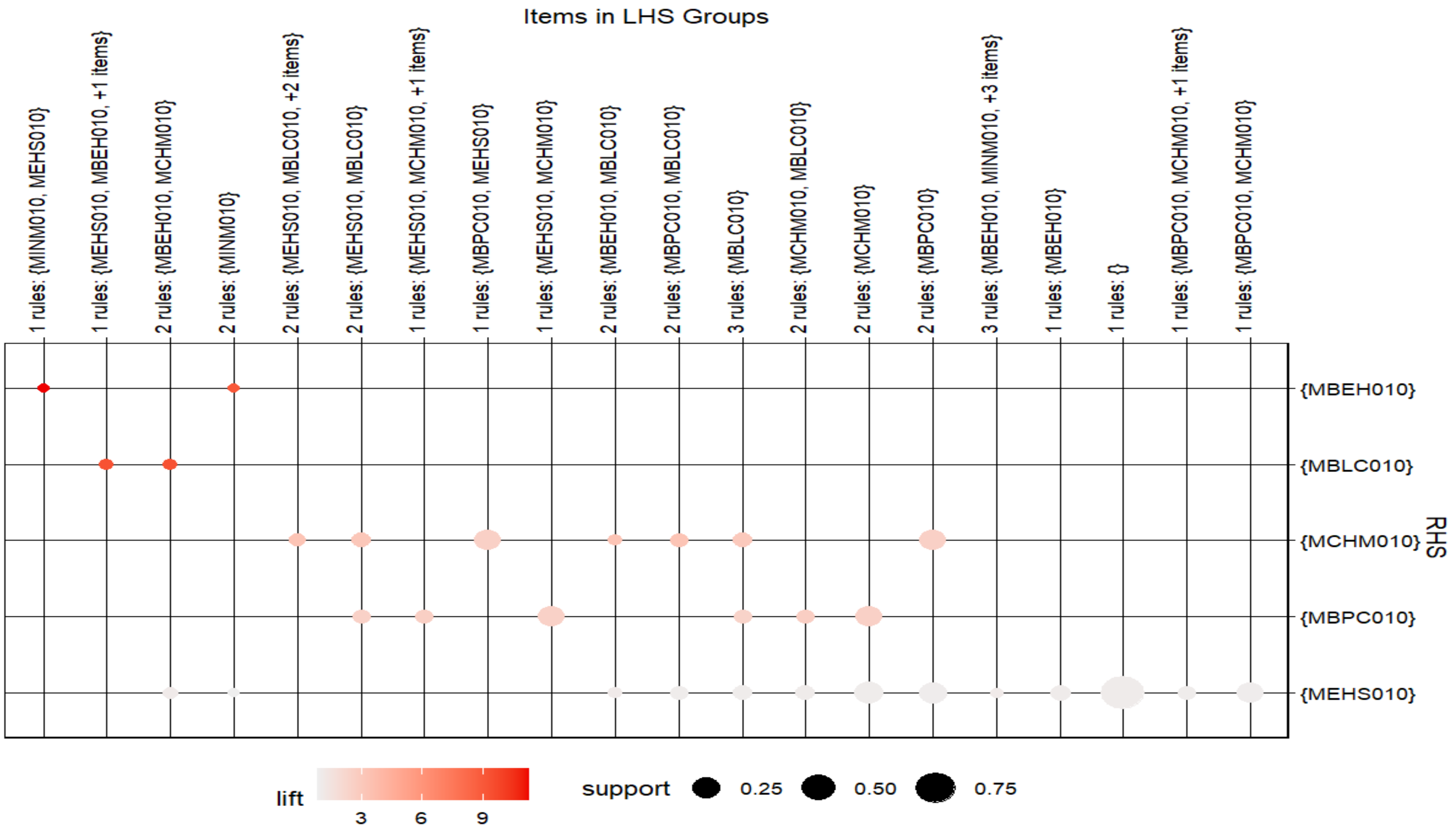
Conclusion… Passing English

lhs		rhs	support	confidence	coverage	lift	count
{MEHS010}	=>	{MCHM010}	0.421	0.460	0.915	1.013	528
{MEHS010}	=>	{MBPC010}	0.420	0.459	0.915	1.006	527
{MCHM010, MEHS010}	=>	{MBPC010}	0.419	0.996	0.421	2.180	526

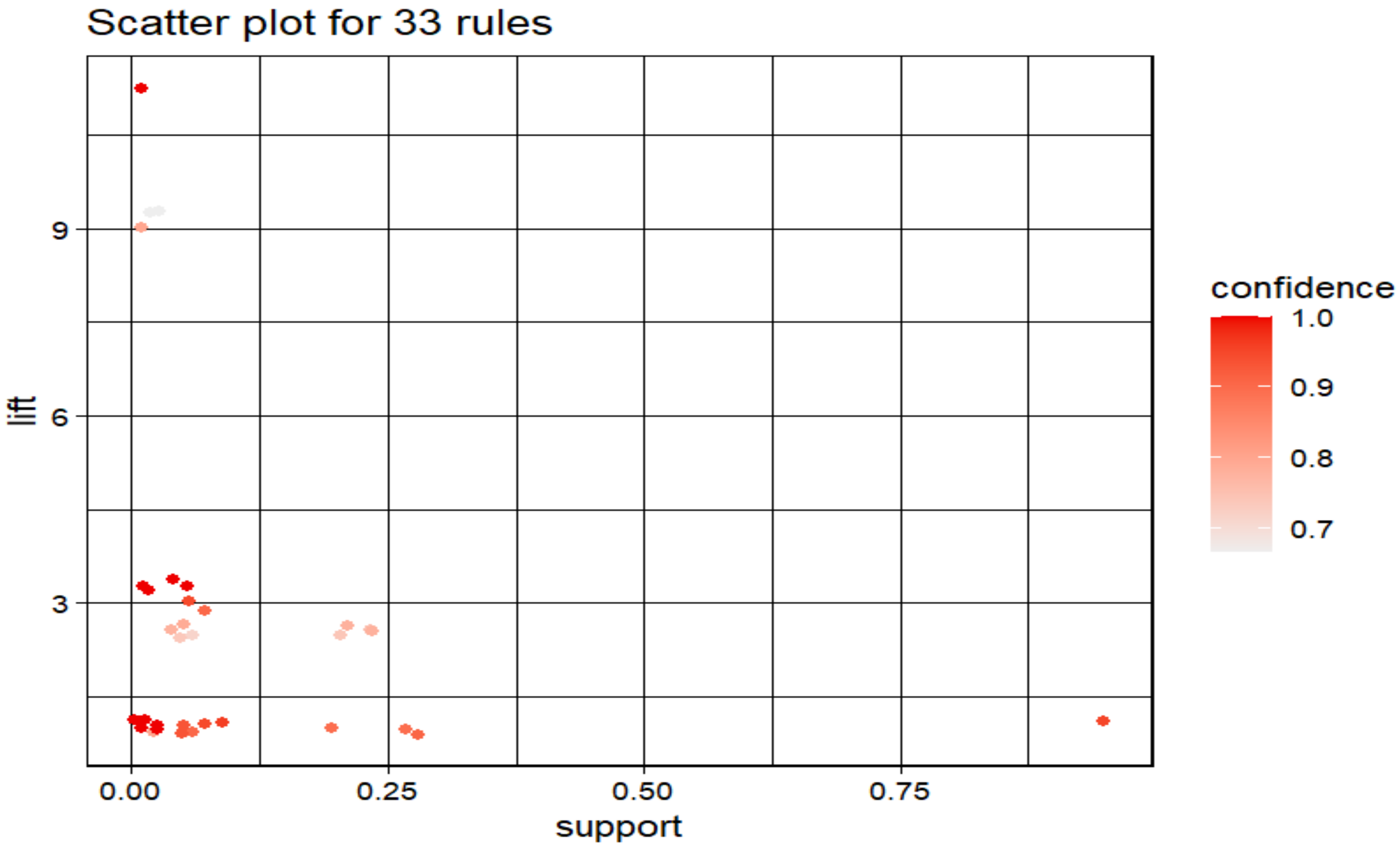
Example…

Passing English {MEHS010} drives chances of passing {MCHM010}, and this shows in 42.1% of the data (Support), with likelihood of 46% (Confidence), and there is a positive relationship between passing {MEHS010} and {MCHM010} (support), and this combination is found in 528 transactions (Count)

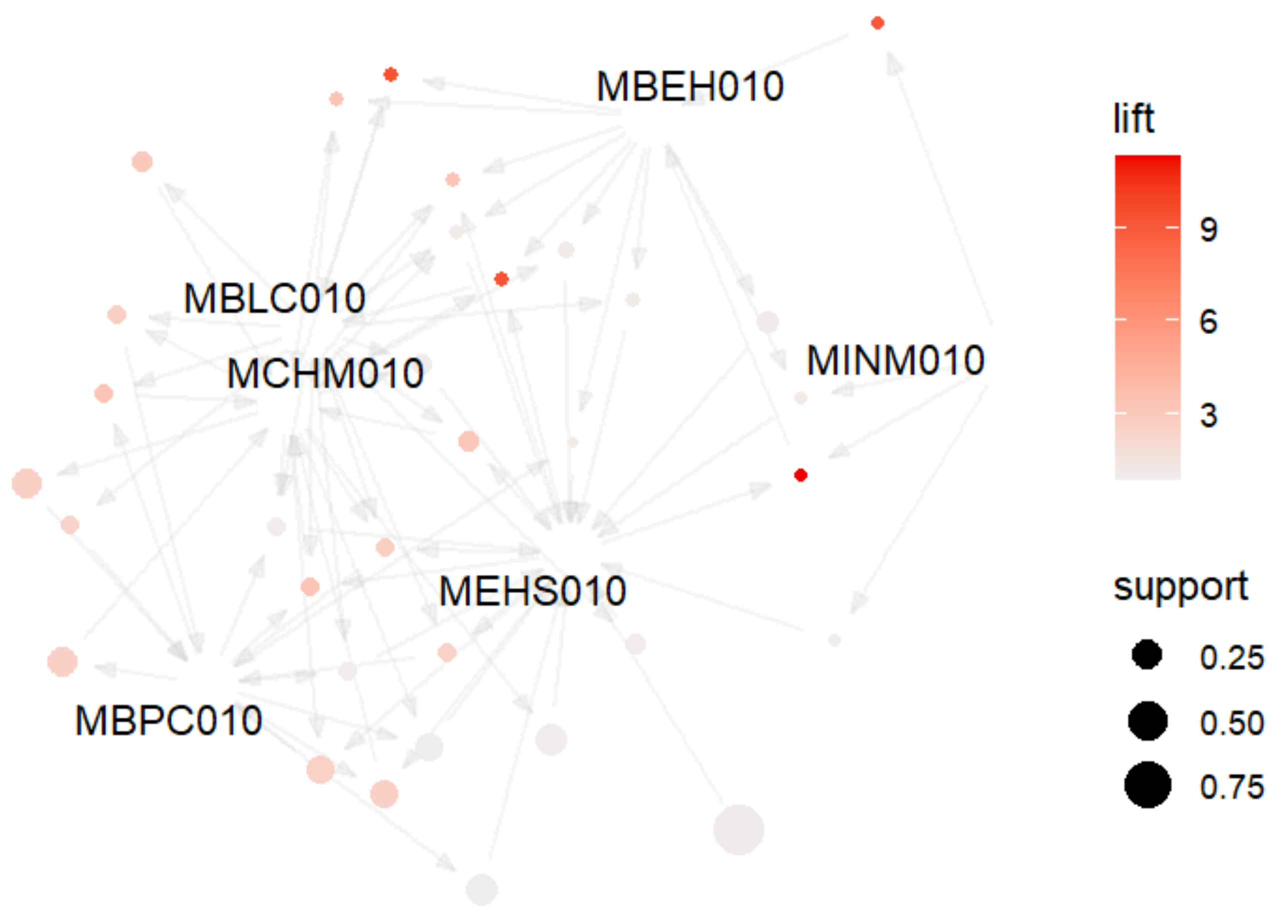
APPENDIX 1A: Visualizations When English is failed…



APPENDIX 1B: Visualizations When English is failed…



APPENDIX 1C: Visualizations When English is failed…



APPENDIX 2: Some Modules Passed when English is passed

lhs		rhs	support	confidence	coverage	lift	count
{MCHM010}	=>	{MBPC010}	0.439	0.967	0.455	2.116	551
{MBPC010}	=>	{MCHM010}	0.439	0.962	0.457	2.116	551
{MEHS010}	=>	{MCHM010}	0.421	0.460	0.915	1.013	528
{MCHM010}	=>	{MEHS010}	0.421	0.926	0.455	1.013	528
{MEHS010}	=>	{MBPC010}	0.420	0.459	0.915	1.006	527
{MBPC010}	=>	{MEHS010}	0.420	0.920	0.457	1.006	527
{MCHM010, MEHS010}	=>	{MBPC010}	0.419	0.996	0.421	2.180	526
{MBPC010, MEHS010}	=>	{MCHM010}	0.419	0.998	0.420	2.196	526
{MBPC010, MCHM010}	=>	{MEHS010}	0.419	0.955	0.439	1.044	526
{MEHS010}	=>	{MBEH010}	0.192	0.210	0.915	0.934	241
{MBEH010}	=>	{MEHS010}	0.192	0.855	0.225	0.934	241
{MICL010}	=>	{MBEH010}	0.163	0.923	0.176	4.105	204
{MBEH010}	=>	{MICL010}	0.163	0.723	0.225	4.105	204
{MICL010}	=>	{MEHS010}	0.142	0.805	0.176	0.881	178
{MEHS010}	=>	{MICL010}	0.142	0.155	0.915	0.881	178
{MEHS010, MICL010}	=>	{MBEH010}	0.142	1.000	0.142	4.447	178
{MBEH010, MICL010}	=>	{MEHS010}	0.142	0.873	0.163	0.954	178
{MBEH010, MEHS010}	=>	{MICL010}	0.142	0.739	0.192	4.191	178
{MINM010}	=>	{MICL010}	0.115	1.000	0.115	5.674	144
{MICL010}	=>	{MINM010}	0.115	0.652	0.176	5.674	144
{MINM010}	=>	{MBEH010}	0.108	0.944	0.115	4.200	136
{MICL010, MINM010}	=>	{MBEH010}	0.108	0.944	0.115	4.200	136
{MBEH010}	=>	{MINM010}	0.108	0.482	0.225	4.200	136
{MBEH010, MINM010}	=>	{MICL010}	0.108	1.000	0.108	5.674	136
{MBEH010, MICL010}	=>	{MINM010}	0.108	0.667	0.163	5.806	136
{MICL010}	=>	{MBPC010}	0.104	0.588	0.176	1.287	130
{MBPC010}	=>	{MICL010}	0.104	0.227	0.457	1.287	130