Data Mining
Homework # 6

Faiz Ali Shah
Reg.No # B55439

March 20, 2016

Question # 1

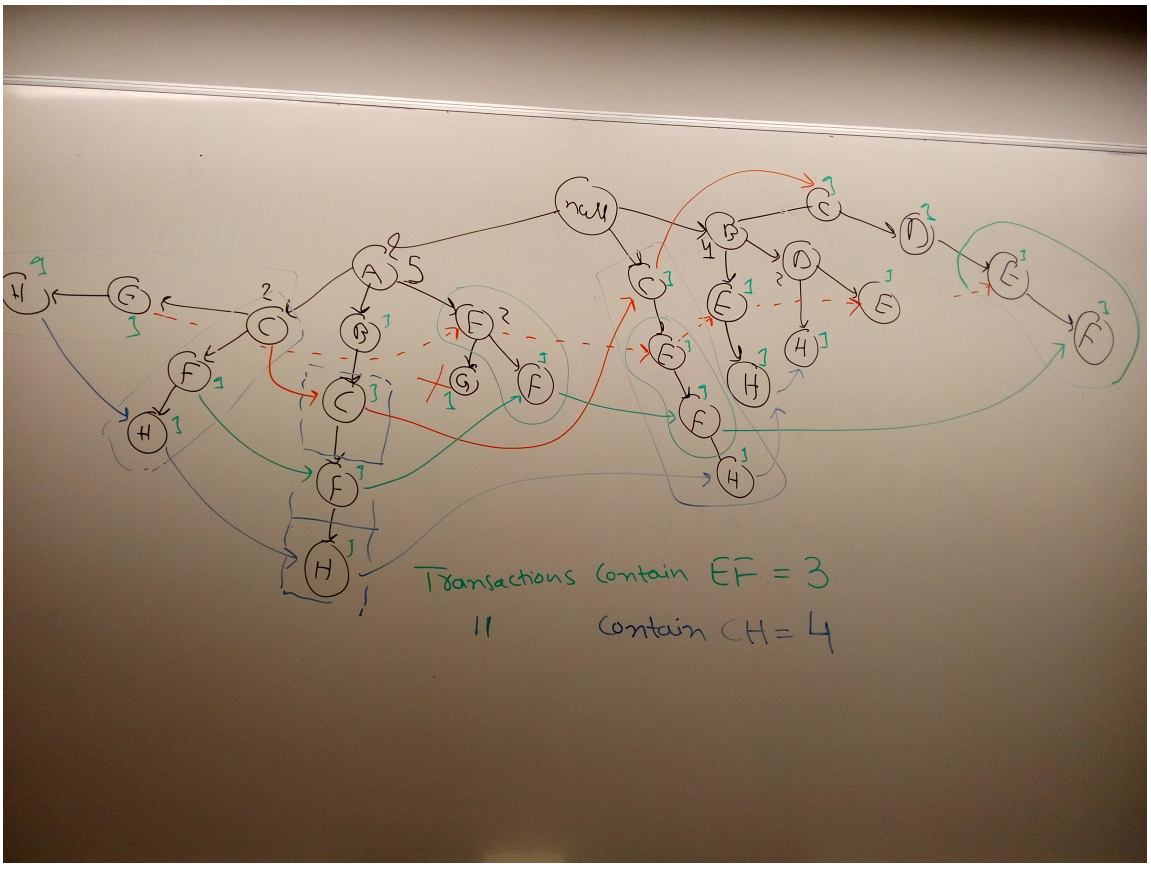

Figure 1: FP Tree

Support of EF = 3
Support of CH = 4

Question # 2

Below is the top 10 (2 x 2 tables) by objective measure function "conviction"

f11	f10	f01	f00	f1.p	f0.p	fp.1	fp.0	conviction
4074	1481	2471	2662	1998	3952	4660	4143	4469 7147.506273
7890	1479	2447	2972	1959	3926	4931	4451	4406 7069.046179
2374	1489	2330	2316	1980	3819	4296	3805	4310 7064.330472
8457	1481	2368	2377	1977	3849	4354	3858	4345 7062.459882
9693	1454	2433	1052	1986	3887	3038	2506	4419 7059.865598
3911	1479	2476	2373	1923	3955	4296	3852	4399 7026.674071
645	1487	2286	2153	1969	3773	4122	3640	4255 7022.797463
918	1466	2374	2577	1966	3840	4543	4043	4340 7020.050548
8525	1487	2450	1567	1917	3937	3484	3054	4367 7017.501633
1898	1460	2499	2968	1930	3959	4898	4428	4429 7016.571028

Similarly, the top 10 (2 x 2 tables) by objective measure function "Interest" are as follow:

f11	f10	f01	f00	f1.p	f0.p	fp.1	fp.0	IF
7023	1494	2073	1002	1933	3567	2935	2496	4006 1.091063111
9592	1463	2112	1041	1955	3575	2996	2504	4067 1.073903908
1926	1431	2040	1055	1926	3471	2981	2486	3966 1.069986392
4387	1476	2176	1064	1994	3652	3058	2540	4170 1.067688075
2317	1335	2024	1049	1990	3359	3039	2384	4014 1.066618830
885	1375	2146	1028	1980	3521	3008	2403	4126 1.061034660
6971	1491	2073	1096	1870	3564	2966	2587	3943 1.055982450
7044	1381	2016	1097	1930	3397	3027	2478	3946 1.053907177
122	1437	2015	1185	1970	3452	3155	2622	3985 1.048956795
5496	1331	2004	1055	1870	3335	2925	2386	3874 1.047095061

The top 10 2 x 2 confusion matrix sorted by objective measure function "Cosine" is as follow:

f11	f10	f01	f00	f1.p	f0.p	fp.1	fp.0	IS
7023	1494	2073	1002	1933	3567	2935	2496	4006 0.5006990326
9899	1484	2053	1004	1000	3537	2004	2488	3053 0.5002547062
2100	1487	2042	1024	1579	3529	2603	2511	3621 0.4995302386
1881	1481	2034	1021	1298	3515	2319	2502	3332 0.4994002336
1496	1494	2040	1095	1003	3534	2098	2589	3043 0.4939139582
8521	1498	2051	1102	1694	3549	2796	2600	3745 0.4931421675
8723	1446	2066	1008	1516	3512	2524	2454	3582 0.4925538978
6971	1491	2073	1096	1870	3564	2966	2587	3943 0.4910328886
1178	1452	2104	1020	1296	3556	2316	2472	3400 0.4897354105
9592	1463	2112	1041	1955	3575	2996	2504	4067 0.4889777984

Below is table presents 2 x 2 confusion matrix sorted by objective measure function "Jaccard"

f11	f10	f01	f00	f1.p	f0.p	fp.1	fp.0	Jaccard
7023	1494	2073	1002	1933	3567	2935	2496	4006	0.3269862114
9899	1484	2053	1004	1000	3537	2004	2488	3053	0.3268002643
2100	1487	2042	1024	1579	3529	2603	2511	3621	0.3265978476
1881	1481	2034	1021	1298	3515	2319	2502	3332	0.3264991182
1496	1494	2040	1095	1003	3534	2098	2589	3043	0.3227478937
8521	1498	2051	1102	1694	3549	2796	2600	3745	0.3220812728
6971	1491	2073	1096	1870	3564	2966	2587	3943	0.3199570815
8723	1446	2066	1008	1516	3512	2524	2454	3582	0.3199115044
9302	1484	2044	1143	1004	3528	2147	2627	3048	0.3177049882
1178	1452	2104	1020	1296	3556	2316	2472	3400	0.3173076923

Below is table presents 2 x 2 confusion matrix sorted by objective measure function "correlation"

f11	f10	f01	f00	f1.p	f0.p	fp.1	fp.0	correlation
7023	1494	2073	1002	1933	3567	2935	2496	4006	0.07924225046
9592	1463	2112	1041	1955	3575	2996	2504	4067	0.06334532906
1926	1431	2040	1055	1926	3471	2981	2486	3966	0.05979076809
4387	1476	2176	1064	1994	3652	3058	2540	4170	0.05773080839
2317	1335	2024	1049	1990	3359	3039	2384	4014	0.05397604962
885	1375	2146	1028	1980	3521	3008	2403	4126	0.05039447206
6971	1491	2073	1096	1870	3564	2966	2587	3943	0.04970728063
7044	1381	2016	1097	1930	3397	3027	2478	3946	0.04525434269
122	1437	2015	1185	1970	3452	3155	2622	3985	0.04153850576
5496	1331	2004	1055	1870	3335	2925	2386	3874	0.03946533188

Question # 3

All objective measure function shows a strong positive linear relationship with f11 as it is demonstrated in figure 2

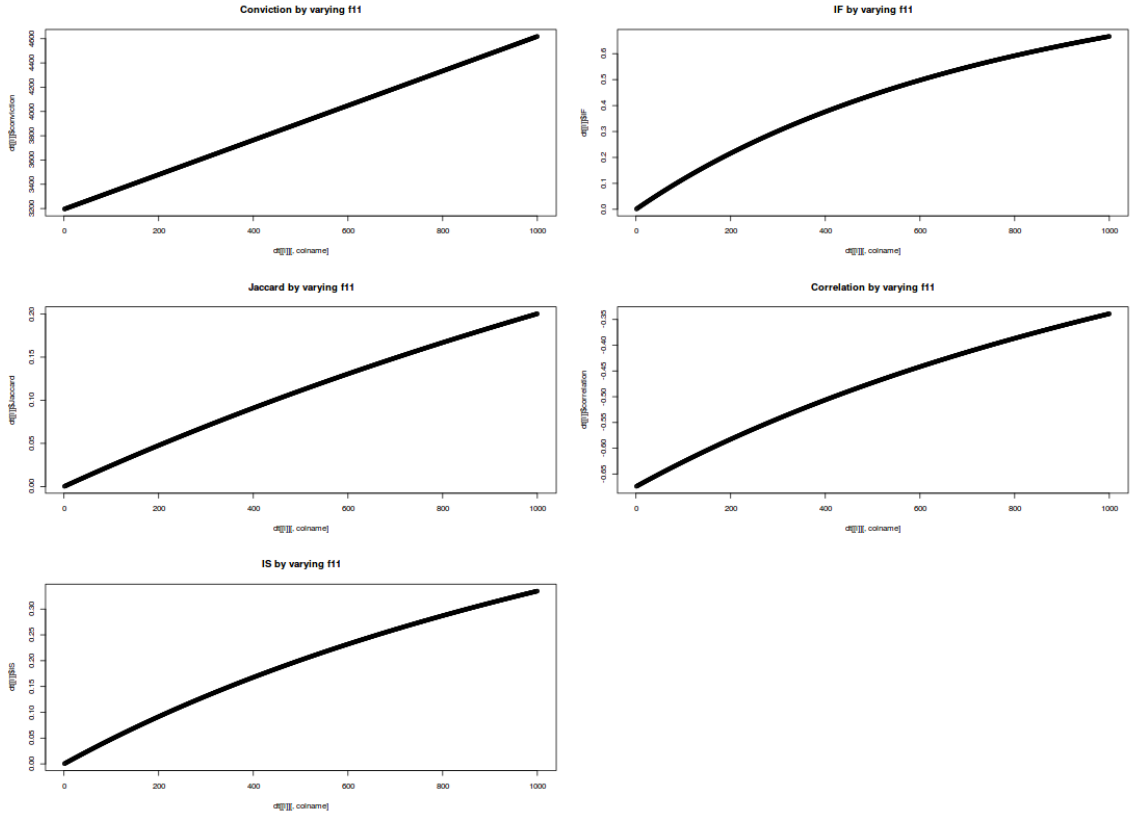

Figure 2: Plots of various interesting measures by varying F11

All objective measure functions in Figure 3 shows the negitive correlation. However, the results of conviction function is specially different as the function is zero if the False Positive value is greater than 25.

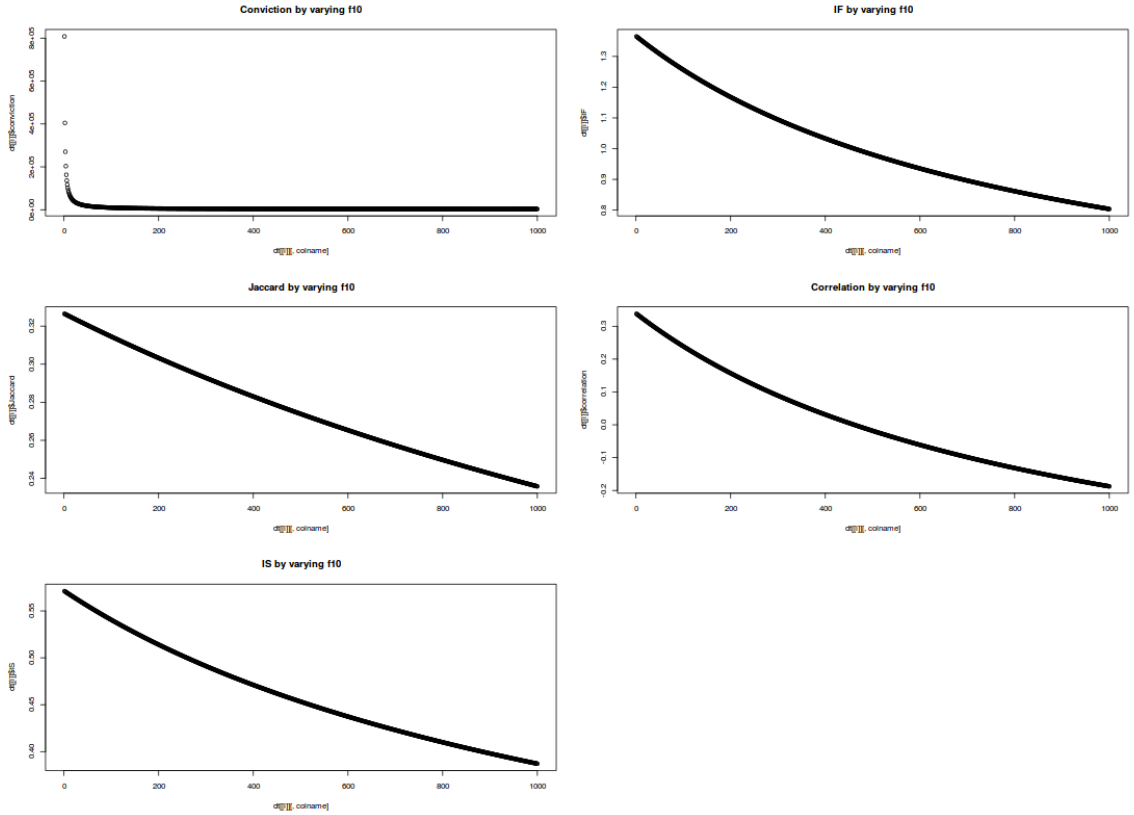

Figure 3: Plots of various interesting measures by varying F10

Figure 4 shows the plot of various objective measure function by varying the values of f01. It is evident from this plot that varying f01 gives a constant function. However, other measures show the negative correlation. By increasing the recall , the precision goes down that's the reason these functions shows negative correlation.

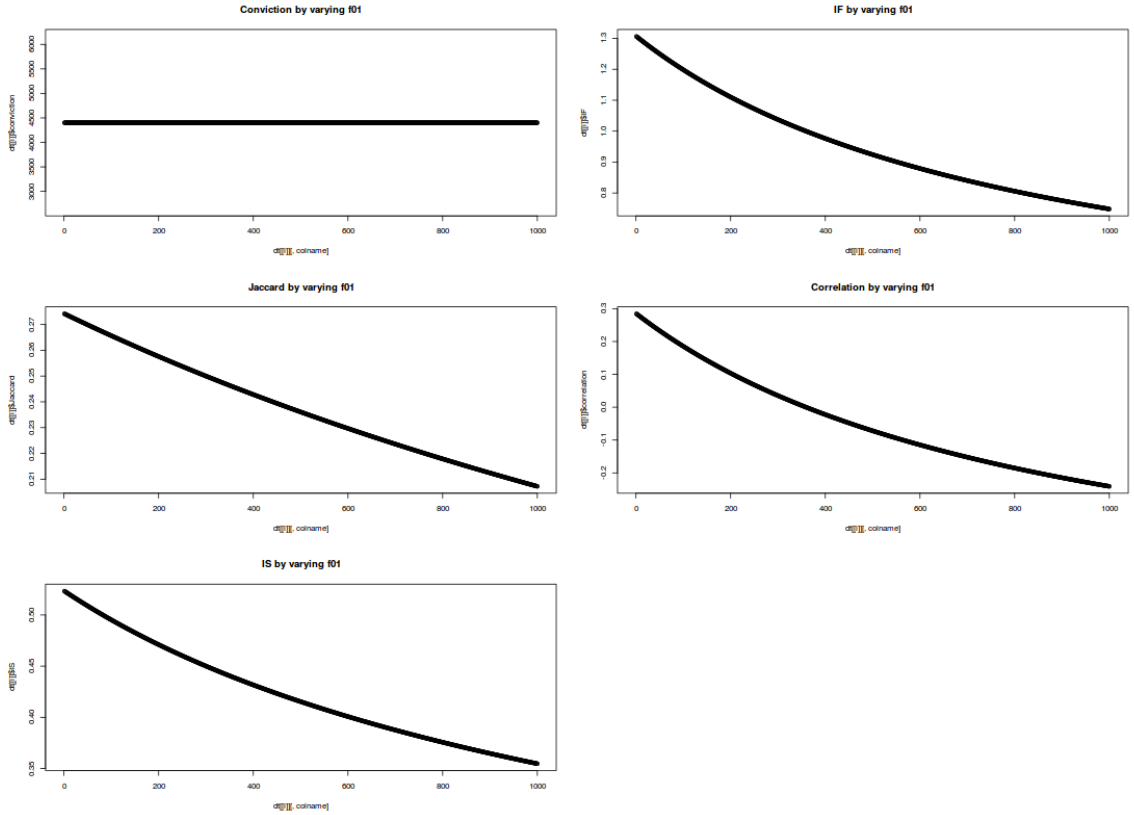

Figure 4: Plots of various interesting measures by varying F01

Figure 5 plot shows no impact on objective measure "Jaccard" and "IS" by varying the values of f00 because these functions do not use f00 in the formulas . Rest of the measures, shows the positive linear relationship.

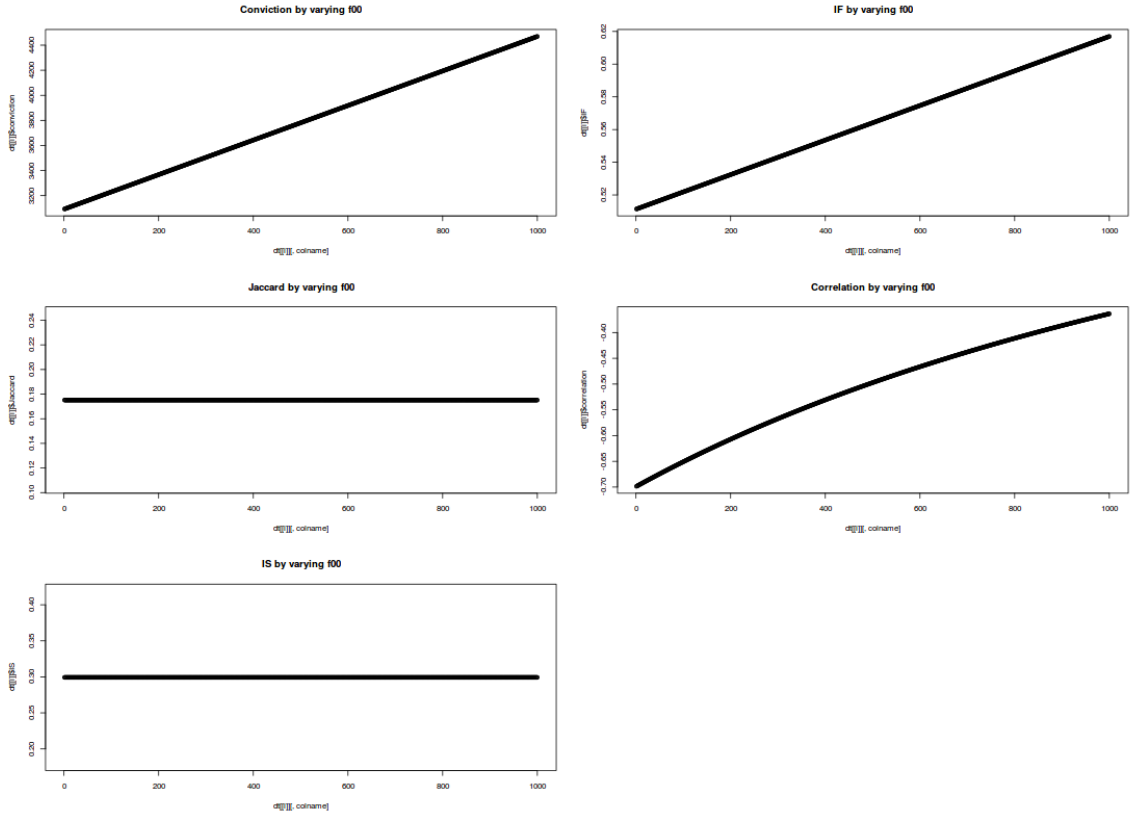

Figure 5: Plots of various interesting measures by varying F00

Question # 4

apriori function is provided in **arules package** in R to mine frequent items sets and association rules in a dataset.

```
rules = apriori(titanic)
```

if you just provide the data then the default parameters to mine rules are with support=0.1, confidence=0.8 , and k=10

```
inspect(rules)
```

The above command present all the rules as shown below :

```
  lhs                    rhs           support      confidence   lift
1 {}                  => {Age=Adult}   0.9504770559 0.9504770559 1.0000000000
2 {Class=2nd}         => {Age=Adult}   0.1185824625 0.9157894737 0.9635050820
3 {Class=1st}         => {Age=Adult}   0.1449341209 0.9815384615 1.0326798059
4 {Sex=Female}        => {Age=Adult}   0.1930940482 0.9042553191 0.9513699605
5 {Class=3rd}         => {Age=Adult}   0.2848705134 0.8881019830 0.9343749831
```

```
6  {Survived=Yes}                    => {Age=Adult}    0.2971376647 0.9198312236 0.9677574203
7  {Class=Crew}                      => {Sex=Male}     0.3916401636 0.9740112994 1.2384742172
8  {Class=Crew}                      => {Age=Adult}    0.4020899591 1.0000000000 1.0521032505
9  {Survived=No}                     => {Sex=Male}     0.6197183099 0.9154362416 1.1639948976
10 {Survived=No}                     => {Age=Adult}    0.6533393912 0.9651006711 1.0153855531
11 {Sex=Male}                        => {Age=Adult}    0.7573830077 0.9630271519 1.0132039968
12 {Sex=Female,Survived=Yes}         => {Age=Adult}    0.1435711040 0.9186046512 0.9664669394
13 {Class=3rd,Sex=Male}              => {Survived=No}  0.1917310313 0.8274509804 1.2222950388
14 {Class=3rd,Survived=No}           => {Age=Adult}    0.2162653339 0.9015151515 0.9484870213
15 {Class=3rd,Sex=Male}              => {Age=Adult}    0.2099045888 0.9058823529 0.9530817681
16 {Sex=Male,Survived=Yes}           => {Age=Adult}    0.1535665607 0.9209809264 0.9689670263
17 {Class=Crew,Survived=No}          => {Sex=Male}     0.3044070877 0.9955423477 1.2658513618
18 {Class=Crew,Survived=No}          => {Age=Adult}    0.3057701045 1.0000000000 1.0521032505
19 {Class=Crew,Sex=Male}             => {Age=Adult}    0.3916401636 1.0000000000 1.0521032505
20 {Class=Crew,Age=Adult}            => {Sex=Male}     0.3916401636 0.9740112994 1.2384742172
21 {Sex=Male,Survived=No}            => {Age=Adult}    0.6038164471 0.9743401760 1.0251064662
22 {Age=Adult,Survived=No}           => {Sex=Male}     0.6038164471 0.9242002782 1.1751385397
23 {Class=3rd,Sex=Male,Survived=No}  => {Age=Adult}    0.1758291686 0.9170616114 0.9648435022
24 {Class=3rd,Age=Adult,Survived=No} => {Sex=Male}     0.1758291686 0.8130252101 1.0337772891
25 {Class=3rd,Sex=Male,Age=Adult}    => {Survived=No}  0.1758291686 0.8376623377 1.2373790639
26 {Class=Crew,Sex=Male,Survived=No} => {Age=Adult}    0.3044070877 1.0000000000 1.0521032505
27 {Class=Crew,Age=Adult,Survived=No} => {Sex=Male}    0.3044070877 0.9955423477 1.2658513618
```

```
rules = apriori(titanic,appearance = list(rhs=c("Sex=Female","Sex=Male"),default="
    lhs"))
inspect(rules)
```

The above instructions present the rules where the right hand side of the rule only contains Sex = Male or Sex= Female. The results are shown as follows. As, it can be notice that no rule is shown with RHS contains Sex=Female because such rules have lower support as defined as a default parameter (i.e. 0.1)

```
 lhs                                rhs          support      confidence   lift
1 {Class=Crew}                   => {Sex=Male} 0.3916401636 0.9740112994 1.238474217
2 {Survived=No}                  => {Sex=Male} 0.6197183099 0.9154362416 1.163994898
3 {Class=Crew,Survived=No}       => {Sex=Male} 0.3044070877 0.9955423477 1.265851362
4 {Class=Crew,Age=Adult}         => {Sex=Male} 0.3916401636 0.9740112994 1.238474217
5 {Age=Adult,Survived=No}        => {Sex=Male} 0.6038164471 0.9242002782 1.175138540
6 {Class=3rd,Age=Adult,Survived=No}  => {Sex=Male} 0.1758291686 0.8130252101 1.033777289
7 {Class=Crew,Age=Adult,Survived=No} => {Sex=Male} 0.3044070877 0.9955423477 1.265851362
```

```
rules = apriori(titanic,parameter = list(minlen=2, supp=0.40, conf=0.6),appearance =
    list(rhs=c("Sex=Female", "Sex=Male"),default="lhs"))
inspect(rules)
```

The above-mentioned R code customizes the parameters for support (.40), confidence (.60), and size. By increasing the support, now we can notice a decrease in the number of rules.

```
 lhs                        rhs          support      confidence   lift
1 {Survived=No}           => {Sex=Male} 0.6197183099 0.9154362416 1.163994898
2 {Age=Adult}             => {Sex=Male} 0.7573830077 0.7968451243 1.013203997
3 {Age=Adult,Survived=No} => {Sex=Male} 0.6038164471 0.9242002782 1.175138540
```

Figure. 6 shows the visualization of the aforementioned 3 rules as a graph

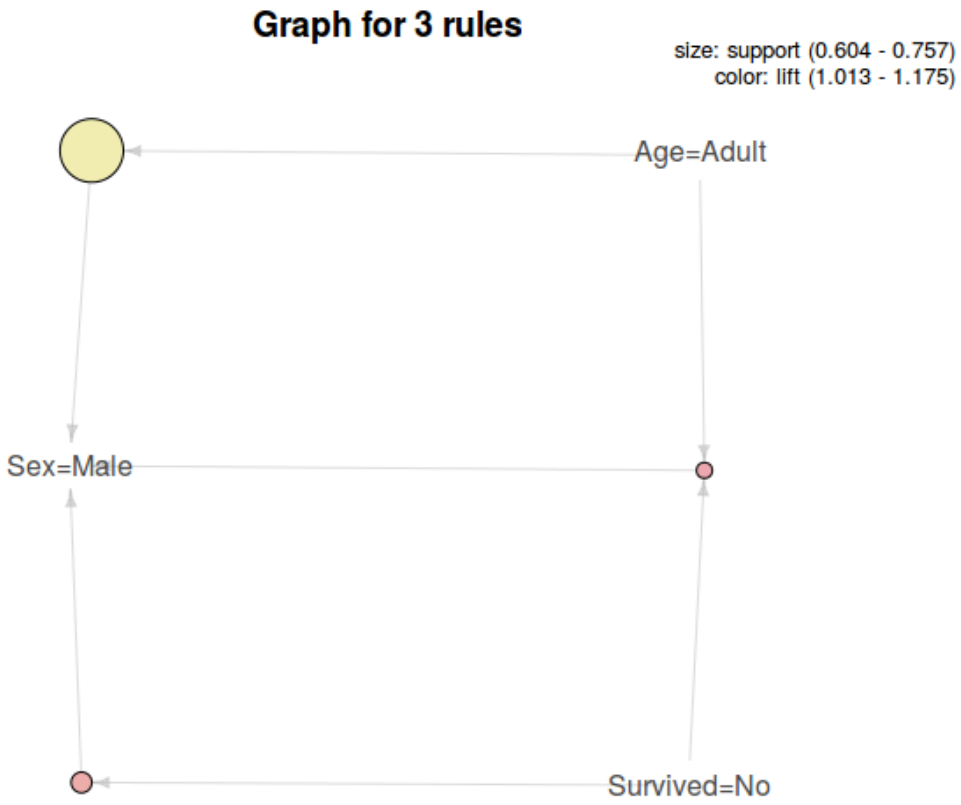

Figure 6: FP Tree

Question # 5

Below are some interesting rules based on objective measure "Interest/lift". The rules are statistical significant as they have high confidence and lift measures.

```
inspect(head(sort(rules,by ="lift"),10))
```

```
lhs                                rhs           support      confidence       lift
17 {Class=Crew,Survived=No}        => {Sex=Male}    0.3044070877 0.9955423477 1.265851362
27 {Class=Crew,Age=Adult,Survived=No} => {Sex=Male}    0.3044070877 0.9955423477 1.265851362
7  {Class=Crew}                    => {Sex=Male}    0.3916401636 0.9740112994 1.238474217
20 {Class=Crew,Age=Adult}          => {Sex=Male}    0.3916401636 0.9740112994 1.238474217
```

Question #6

Below is the confusion matrix of the rule (i.e. 17) with highest lift.

	Male	**Not Male**	
Class=Crew, Survived=No	670	3	673
NOT Class=Crew, Survived=No	1061	467	1528
	1731	470	2101

Below are the plots of different objective measure functions (symmetric and asymmetric) by varying the value of f11 from -400 to 400 with the increments of 10.

Figure 7, Figure 8, and Figure 9 presents the impact on symmetric objective measures, such as "Cosine","Interest", and "conviction" by varying the value of f11 from (-400 to 400 by increments of 10). It is clear from plot 9 that jaccard and cosine is linear. However, the Laplace interest is monotonically increasing but not that rapidly.

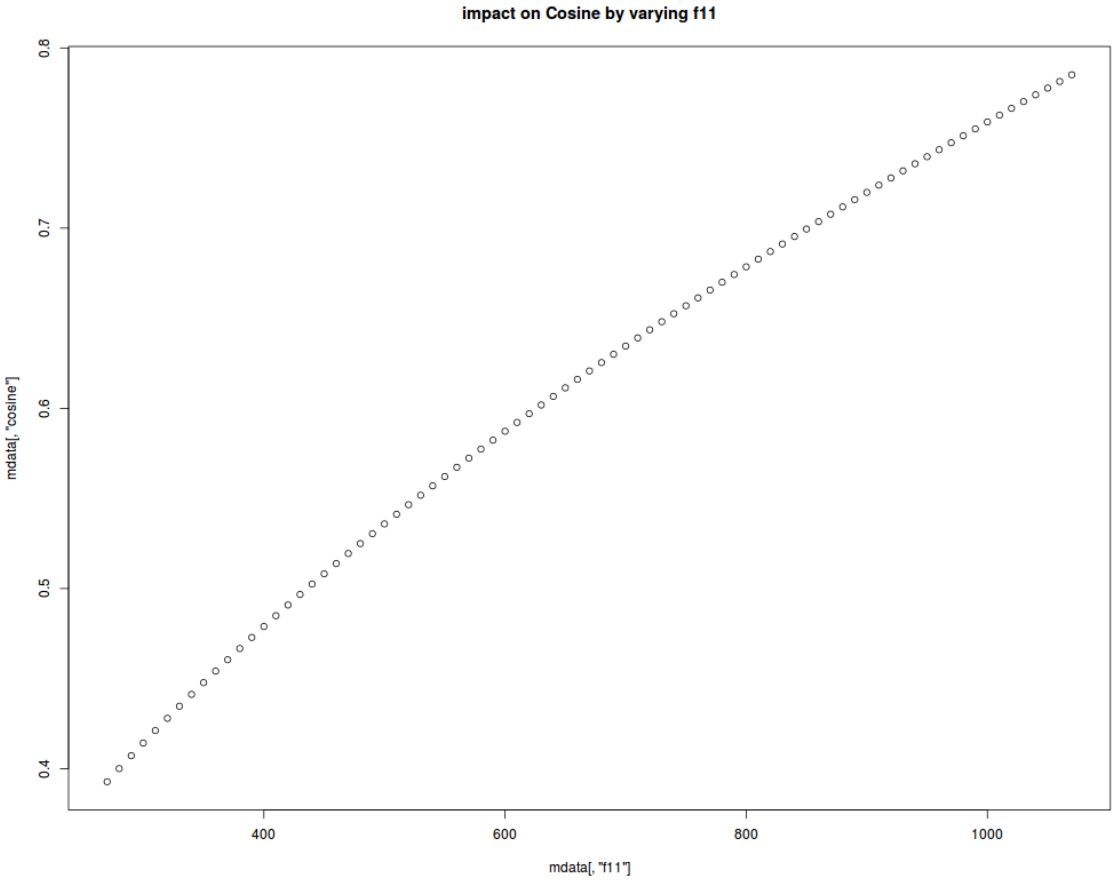

Figure 7: Impact on Cosine measure by varying f11

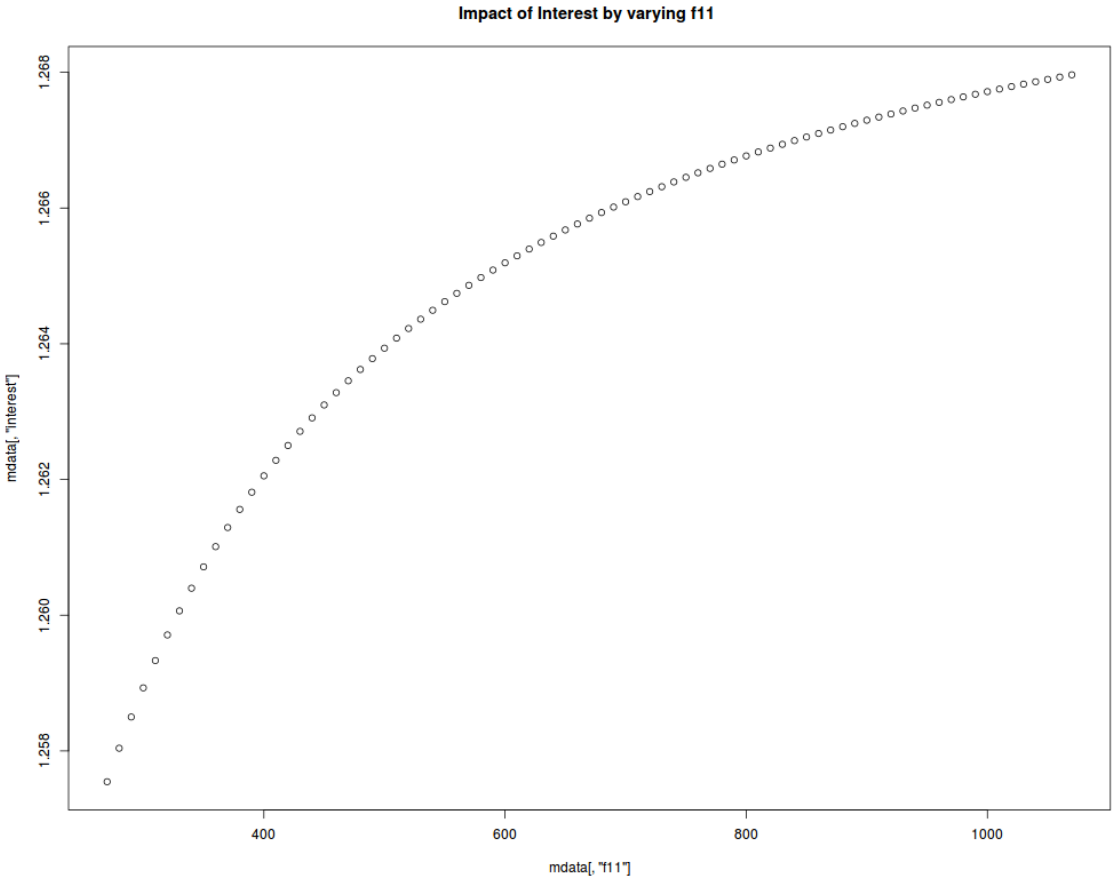

Figure 8: Impact on interest measure by varying f11

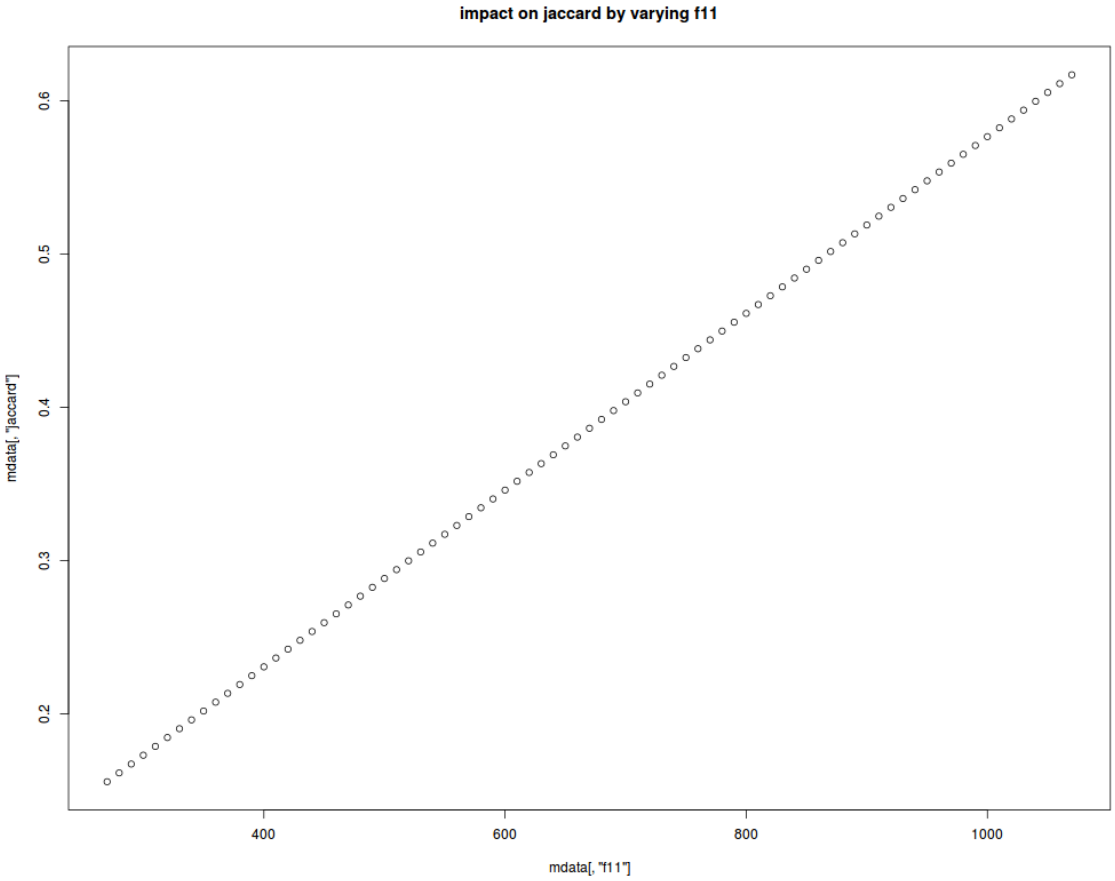

Figure 9: impact on jaccard measure by varying f11

Figure 10 and 11 presents the impact on asymmetric objective measures, such as "laplace" and "conviction" by varying the value of f11 from (-400 to 400 by increments of 10). It is clear from figure 11 that conviction is perfectly linear. However, the Laplace measure is monotonically increasing but not that rapidly.

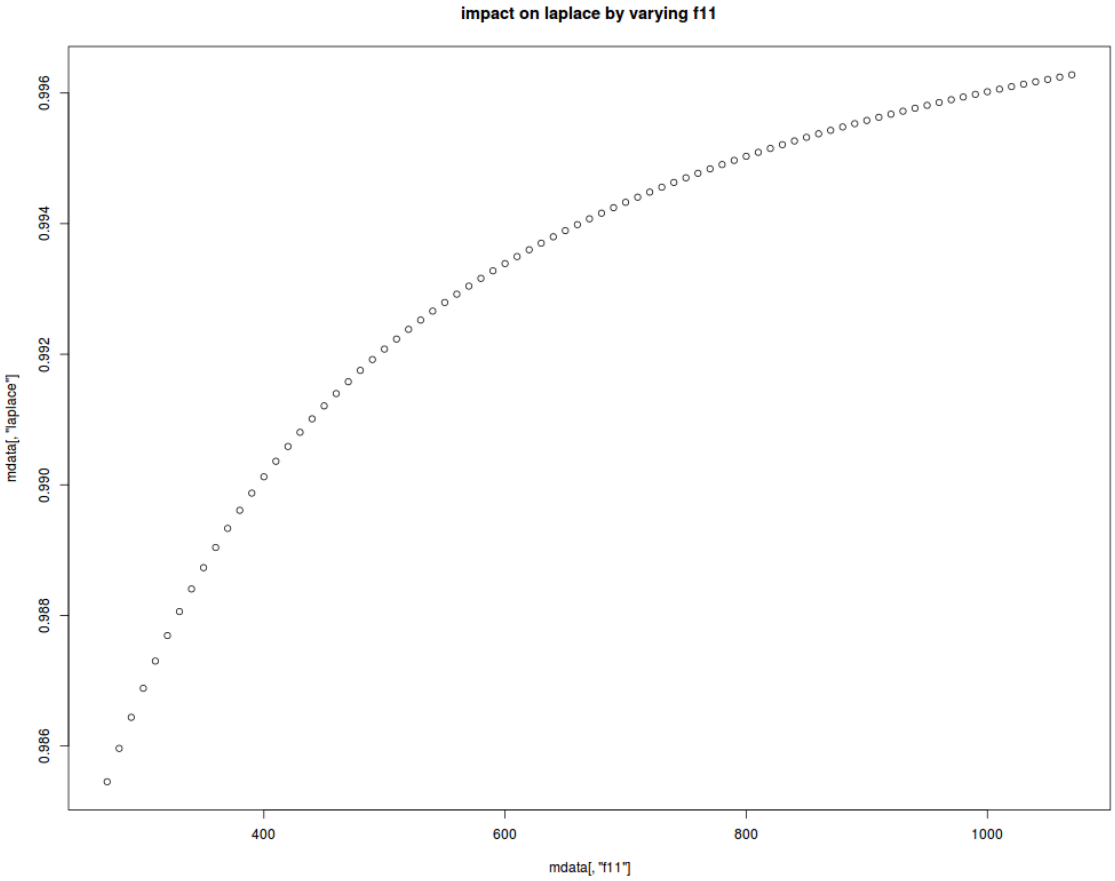

Figure 10: impact on Laplace measure by varying f11

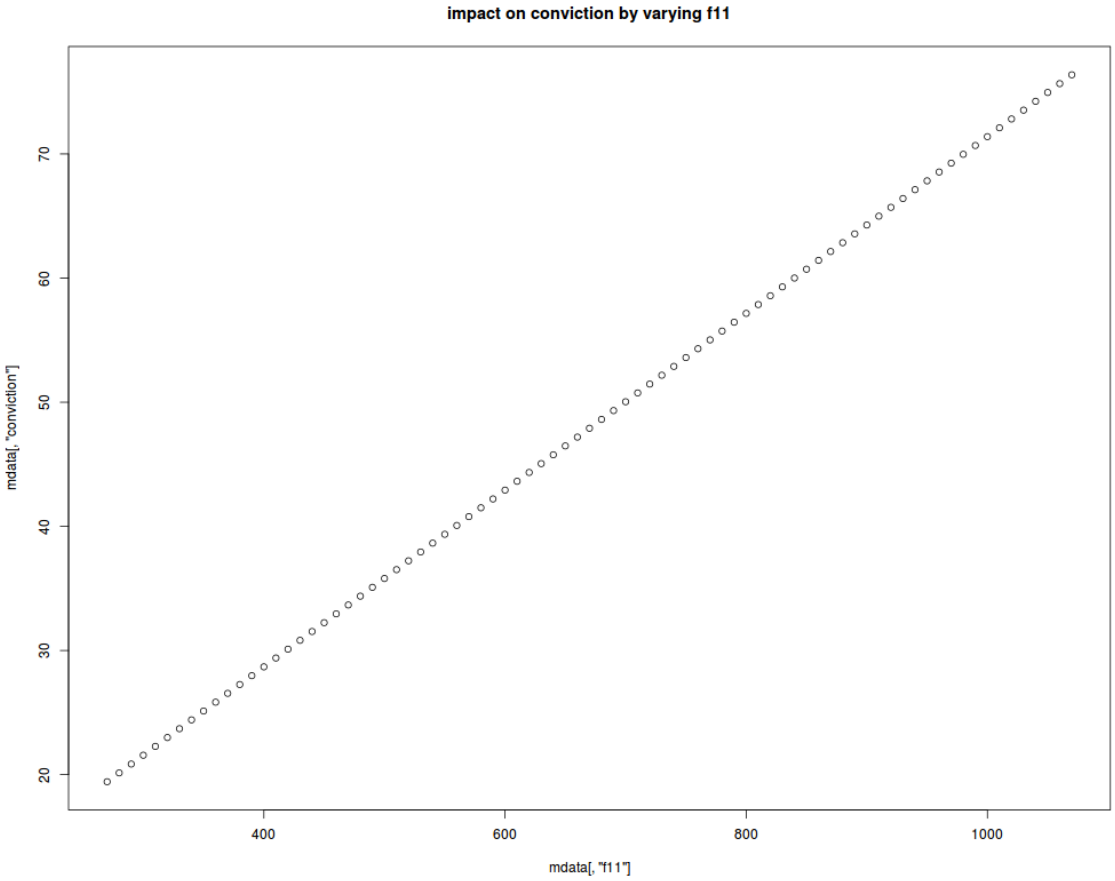

Figure 11: impact on conviction measure by varying f11

Conclusion : All symmetric/asymmetric objective measures are strongly correlated with the value of f11.