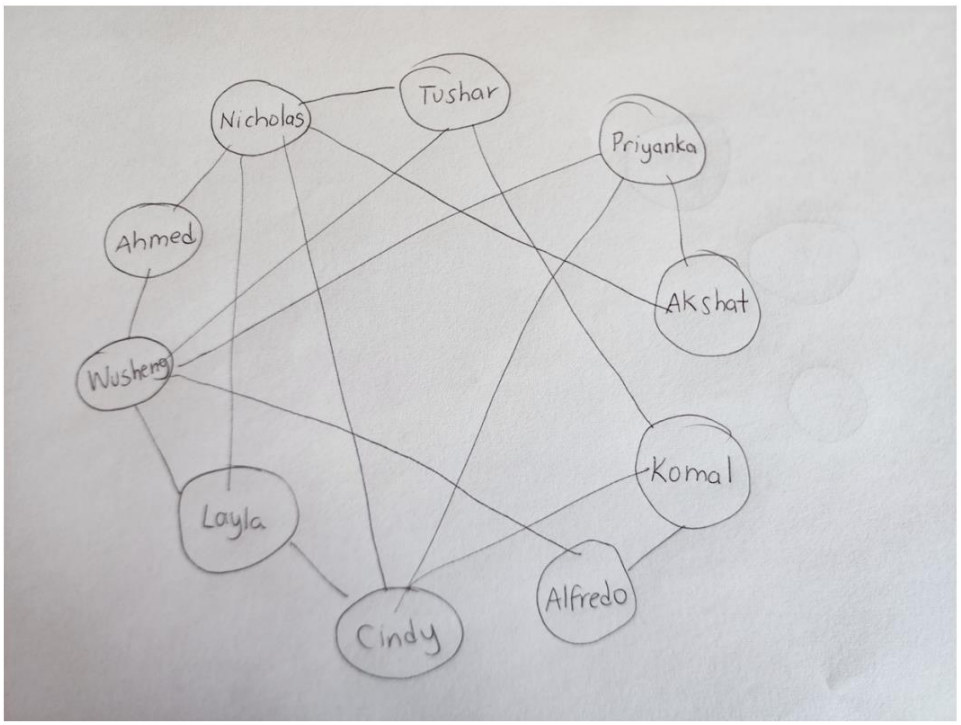
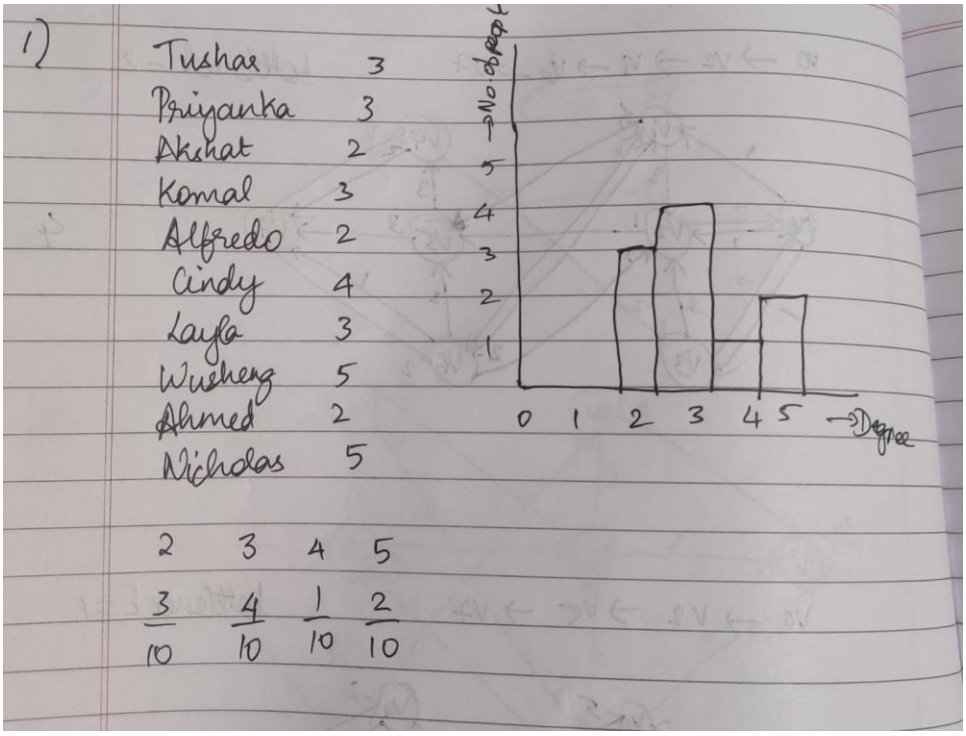
Online Social Network Analysis – HW – 1 Pranathy - A20555875

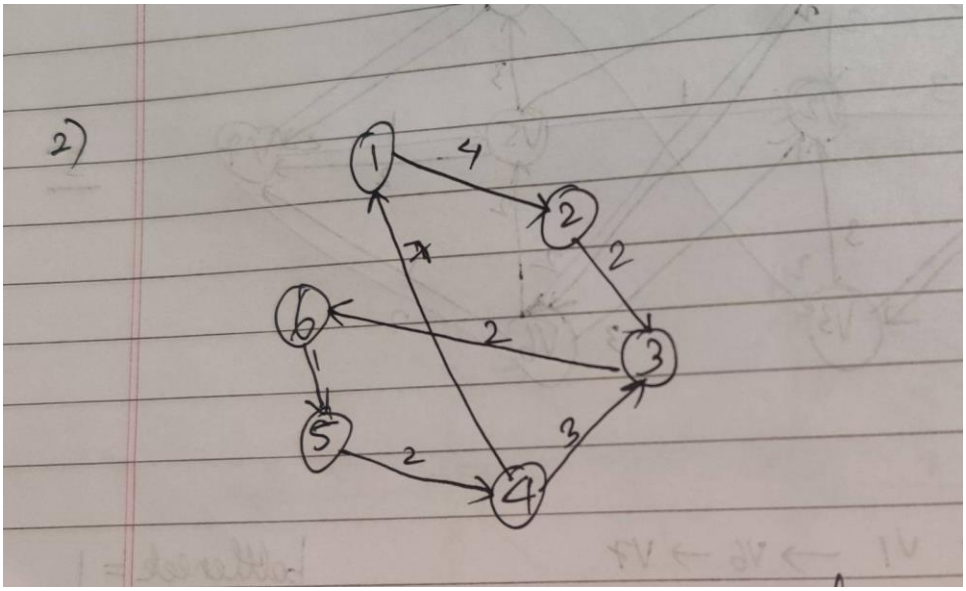
1. Given the friendship graph below, calculate and plot the degree distribution of the graph. Be sure to label the plot axes.





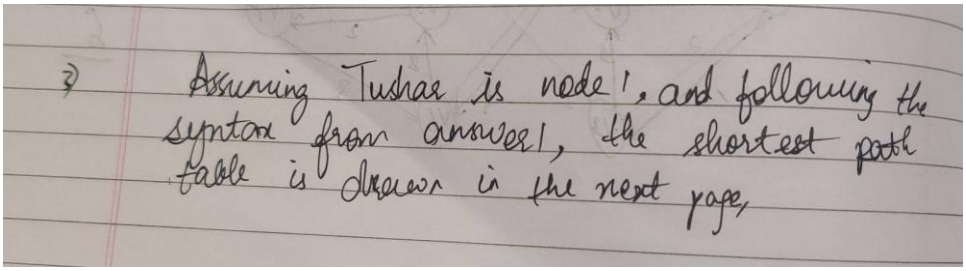
2. Draw the graph specified in the adjacency matrix below. Is this graph connected? If yes, is it weakly connected or strongly connected?

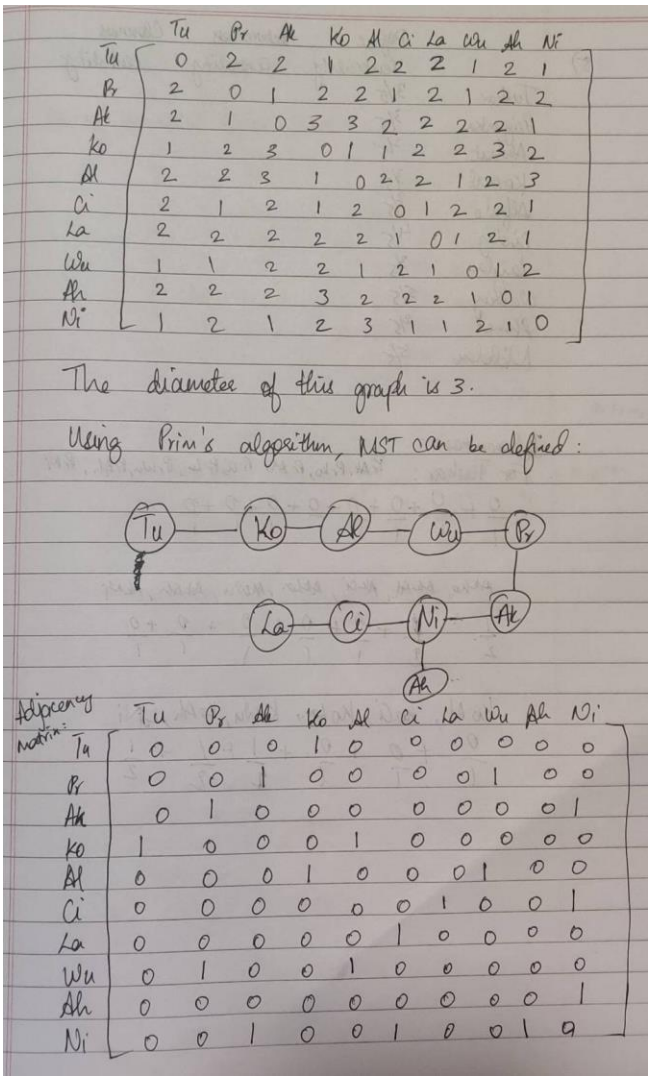
$$\begin{bmatrix} 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \\ 7 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$



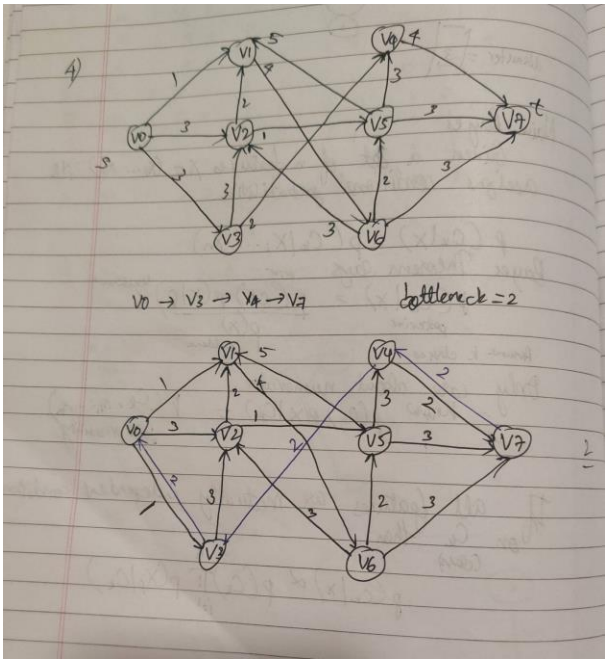
Strongly connected.

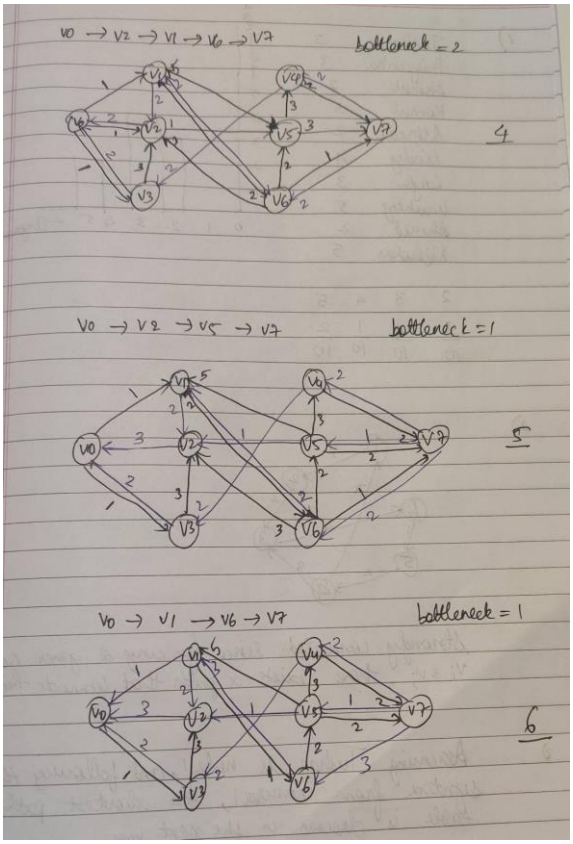
3. Use Dijkstra's or Prim's Algorithm to create a shortest path table for the friend graph from problem 1. What is the diameter of this graph? Show a minimum spanning tree of this graph as an adjacency matrix.



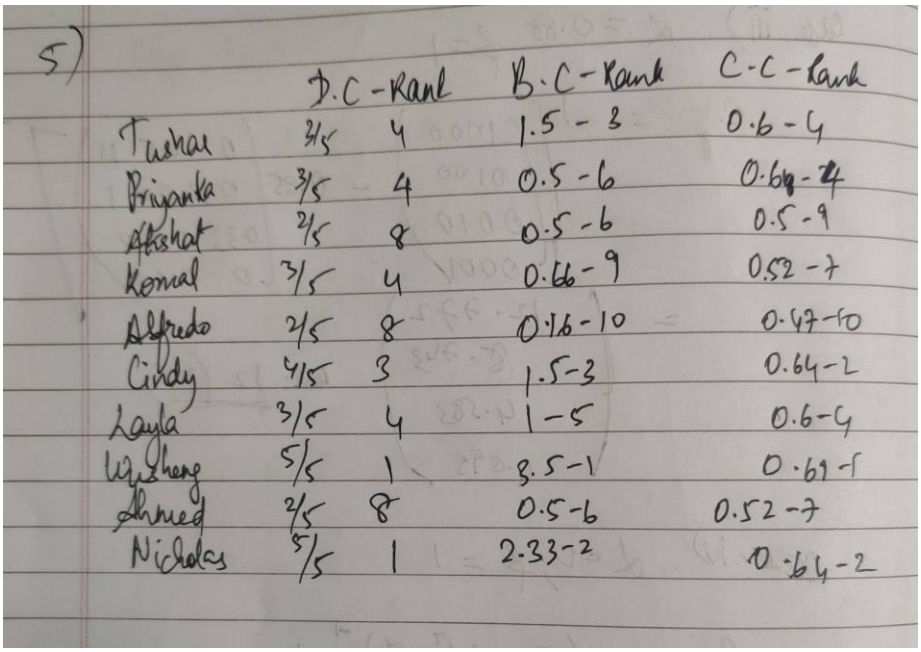


4. Given the following flow network from v0 (source) to v7 (sink), use the Ford-Fulkerson algorithm to determine the maximum flow. Provide the resulting flow and draw and label the flow network and residual network. Edge list:
{(v0,v1,1),(v0,v2,3),(v0,v3,3),(v1,v6,4),(v2,v1,2),(v2,v5,1),(v3,v2,3),(v3,v4,2),(v4,v7,4),
(v5,v1,5),(v5,v4,3),(v5,v7,3),(v6,v2,3),(v6,v5,2),(v6,v7,3)}


4)
V0 → V3 → V4 → V7
bottleneck = 2


V0 → V2 → V1 → V6 → V7
bottleneck = 2
4
V0 → V2 → V5 → V7
bottleneck = 1
5
V0 → V1 → V6 → V7
bottleneck = 1
6

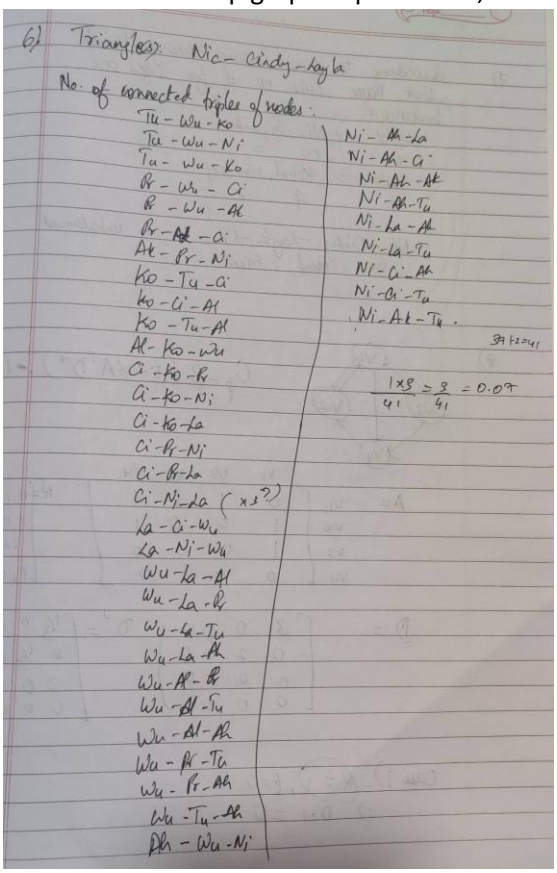
5. For the friendship graph in problem 1, calculate the degree centrality, betweenness centrality and closeness centrality for each node. Provide a table showing the rank of each node for each measure.



5)

	D.C - Rank		B.C - Rank	C.C - Rank
Tushar	3/5	4	1.5 - 3	0.6 - 4
Priyanka	3/5	4	0.5 - 6	0.64 - 4
Akshat	2/5	8	0.5 - 6	0.5 - 9
Komal	3/5	4	0.66 - 9	0.52 - 7
Alfredo	2/5	8	0.16 - 10	0.47 - 10
Cindy	4/5	3	1.5 - 3	0.64 - 2
Layla	3/5	4	1 - 5	0.6 - 4
Wusheng	5/5	1	8.5 - 1	0.69 - 1
Ahmed	2/5	8	0.5 - 6	0.52 - 7
Nicholas	5/5	1	2.33 - 2	0.64 - 2

6. For the friendship graph in problem 1, what is the clustering coefficient?



6) Triangles: Nic - Cindy - Layla

No. of connected triples of nodes:

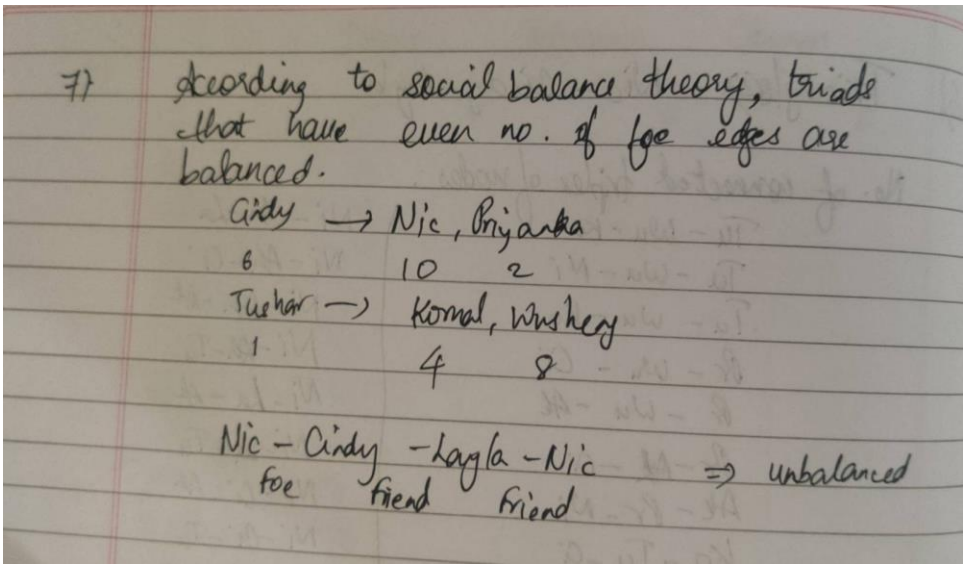
Tu - Wu - Ko
Tu - Wu - Ni
Tu - Wu - Ko
Pr - Wu - Ci
Pr - Wu - Al
Pr - Al - Ci
Al - Pr - Ni
Ko - Tu - Ci
Ko - Ci - Al
Ko - Tu - Al
Al - Ko - Wu
Ci - Ko - Pr
Ci - Ko - Ni
Ci - Ko - La
Ci - Pr - Ni
Ci - Pr - La
Ci - Ni - La (x3?)
La - Ci - Wu
La - Ni - Wu
Wu - La - Al
Wu - La - Pr
Wu - La - Tu
Wu - La - Al
Wu - Al - Pr
Wu - Al - Tu
Wu - Pr - Tu
Wu - Pr - Al
Wu - Tu - Al
Al - Wu - Ni

Ni - Al - La
Ni - Al - Ci
Ni - Al - Al
Ni - Al - Tu
Ni - La - Al
Ni - La - Tu
Ni - Ci - Al
Ni - Ci - Tu
Ni - Al - Tu

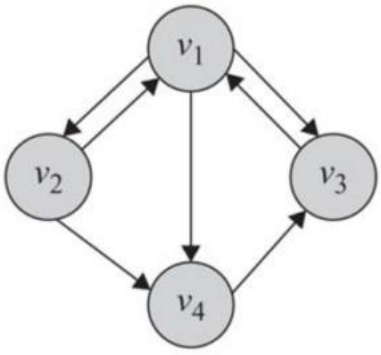
39 + 2 = 41

$\frac{1 \times 8}{41} = \frac{8}{41} = 0.07$

7. For the friendship graph in problem 1, assume that Cindy is a foe of Nicholas and Priyanka, Tushar is a foe of Wusheng and Komal and all the other edges in the graph represent friendship. According to social balance theory, is this new friend/foe graph balanced?



8. Calculate PageRank values for the graph below when • a=1, b=0 • a=b=0.3 • a=0.85, b=1 • a=0, b=1 Discuss the effects of different values of a and b for this particular problem.



Networkx's pagerank function output that doesn't factor β:

```
PageRank for each node: alpha:  1  beta:  0
Node: 1 PageRank: 0.33333260221557337
Node: 2 PageRank: 0.11111141117268056
Node: 3 PageRank: 0.33333369598839135
Node: 4 PageRank: 0.22222229062335402
PageRank for each node: alpha:  0.3  beta:  0.3
Node: 1 PageRank: 0.25876513
Node: 2 PageRank: 0.20087667499999998
Node: 3 PageRank: 0.2792185175
Node: 4 PageRank: 0.2611396775
PageRank for each node: alpha:  0.85  beta:  1
Node: 1 PageRank: 0.31409230948584155
Node: 2 PageRank: 0.1264931095706823
Node: 3 PageRank: 0.32540314534572534
Node: 4 PageRank: 0.23401143559775067
PageRank for each node: alpha:  0  beta:  1
Node: 1 PageRank: 0.25
Node: 2 PageRank: 0.25
Node: 3 PageRank: 0.25
Node: 4 PageRank: 0.25
```

Custom pagerank function that implements α and β:

```
PageRank for each node: alpha = 1 , beta = 0
Node: 1 PageRank: 0.33333333333333304
Node: 2 PageRank: 0.11111111111111127
Node: 3 PageRank: 0.3333333333333336
Node: 4 PageRank: 0.22222222222222215
PageRank for each node: alpha = 0.3 , beta = 0.3
Node: 1 PageRank: 0.25171554611551344
Node: 2 PageRank: 0.2273507066416142
Node: 3 PageRank: 0.26258067151376524
Node: 4 PageRank: 0.25835307572910704
PageRank for each node: alpha = 0.85 , beta = 1
Node: 1 PageRank: 0.25270984543118913
Node: 2 PageRank: 0.22182022457443407
Node: 3 PageRank: 0.265940267242289
Node: 4 PageRank: 0.25952966275208783
PageRank for each node: alpha = 0 , beta = 1
Node: 1 PageRank: 0.25
Node: 2 PageRank: 0.25
Node: 3 PageRank: 0.25
Node: 4 PageRank: 0.25
```

9. You have been tasked to design a classifier that decides whether students will be admitted to a CS graduate program. Applications to the program are received from students all around the world. Application contain student name, address, mobile phone number, final grade point average in undergraduate program and transcript. Describe what information from the application you would use as input to your classifier. For each piece of information, what (if anything) would need to be done to clean or transform the information into input data. For each of the transformed data input, identify the type (nominal, ordinal, interval or ration). Give at least 3 other pieces of information that would be helpful and describe why you think they would help.

Inputs:

1. Undergraduate GPA: Interval data, no transformation needed

2. Undergraduate major: Nominal data, Encode majors into numeric categories

3. Undergraduate university ranking: Ordinal data, Assign numeric ranks

4. GRE quantitative score: Interval data, no transformation needed

5. GRE verbal score: Interval data, no transformation needed

6. Number of CS courses taken: Ratio data, no transformation needed

The other fields like student name, address, and phone number need not be considered while deciding.

Additional useful inputs:

7. Letters of recommendation: Text content, extract key phrases about student qualities

8. Personal statement: Text content, analyze writing quality and stated interests

9. Years of work experience: Numeric, relevant industry experience indicates preparedness

The GPA, GRE scores, and CS coursework give quantitative measures of academic ability and preparation that would likely correlate with success in a graduate CS program.

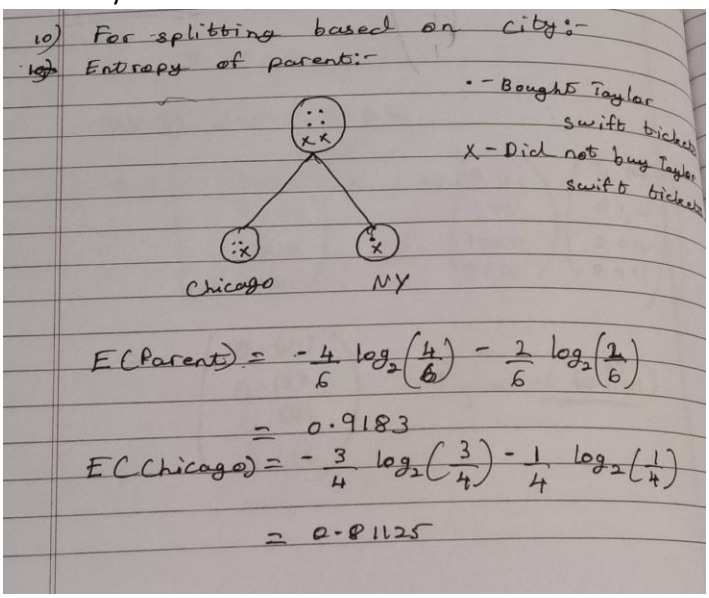
The additional qualitative inputs help gauge interests, communication skills, recommendations, and experience - other helpful factors for evaluating applicants.

The inputs would be fed into a classification algorithm like logistic regression, SVM, or neural network to predict admissions chance.

10. You are given the following set of data

Name	City	Likes Beyoncé	In a relationship	Age	Number of concerts per year	Bought ticket to see Taylor Swift
Kate	Chicago	Yes	No	23	8	Yes
Joe	New York	No	No	36	4	Yes
Mena	New York	Yes	Yes	43	20	No
Pat	Chicago	No	Yes	19	2	No
Tim	Chicago	Yes	No	20	14	Yes
Tina	Chicago	Yes	Yes	54	7	Yes

Using entropy as a measure of purity, design a decision tree to predict whether someone bought a ticket to see Taylor Swift or not. Show how each decision node was selecte



$E(N,Y) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$

$= 1$

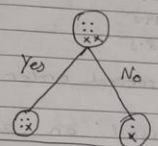
$E(City) = \frac{4}{6} \times 0.81125 + \frac{2}{6} \times 1$

$= 0.8741$

$IG_{City} = 0.9183 - 0.8741$

$= 0.0442$

For splitting based on liking Beyoncé:-



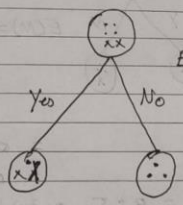
$E(Yes) = 0.81125$

$E(No) = 1$

$E(Beyoncé) = 0.8741$

$IG_{Beyoncé} = 0.0442$

For splitting based on relationship status:



$E(Yes) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)$

$= 0.3899 + 0.5283$

$= 0.9183$

$E(No) = -\frac{3}{3}\log_2\left(\frac{3}{3}\right) = 0$

$\therefore IG_{relationship} = \frac{6}{6}0.9183 - \left(\frac{3}{6} \times 0.9183 + \frac{3}{6} \times 0\right)$

$IG = 0.45913 \rightarrow$ Maximum

For splitting based on age:-

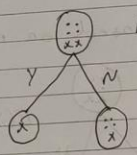
Since it is a continous numerical value, the avg. of pairs of values was taken.

19 → 19.5
20 → 21.5
23 → 29.5
36 → 39.5
43 → 48.5
54

↑ In sorted order

splitting based on avg. age range

For ≤ 19.5



$E(Y) = -1\log_2(1) = 0$

$E(N) = -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right)$

$= 0.4642 + 0.2578$

$= 0.72172$

$E(\leq 19.5) = 0 + \frac{5}{6} \times 0.72172 = 0.60143$

$IG_{\leq 19.5} = 0.31687$

For ≤ 21.5

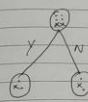
$E(Y) =$ 0 since 1

$E(N) = 0.81125$

$E(\leq 21.5) = 0.8741$

$IG_{\leq 21.5} = 0.0442$

For ≤ 29.5



$E(Y) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$

$= 0.9183$

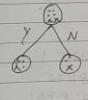
$E(N) = 0.9183$

$E(\leq 29.5) = \frac{3}{6}(0.9183) + \frac{3}{6}(0.9183)$

$= 0.9183$

$\therefore IG_{\leq 29.5} = 0$

For ≤ 39.5



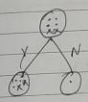
$E(Y) = 1$

$E(N) = 0.81125$

$E(\leq 39.5) = 0.8741$

$IG_{\leq 39.5} = 0.0442$

For ≤ 48.5



$E(Y) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log$

$= 0.5284 + 0.4421$

$= 0.97086$

$E(N) = 0$

$E(\leq 48.5) = 0.80905$

$IG_{\leq 48.5} = 0.1092$

For splitting based on No. of contacts ∵ it is a continous, we take avg. values.

2 → 3
4 → 5.5
7 → 7.5
8 → 11
14 → 17
20

For ≤ 3:-

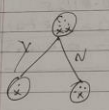
$E(Y) = 0$

$E(N) = 0.72172$

$E(\leq 3) = 0.60143$

$IG_{\leq 3} = 0.31687$

For ≤ 5.5:-



$E(Y) = 0.81125$ 1

$E(N) = 0.81125$

$E(\leq 5.5) = 0.8741$

$IG_{\leq 5.5} = 0.0442$

For ≤ 7.5:-

$E(Y) = 0.9183$

$E(N) = 0.9183$

$E(\leq 7.5) = 0.9183$

$IG_{\leq 7.5} = 0$

For ≤ 11:-

$IG_{\leq 11} = 0.0442$

For ≤ 17:-

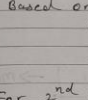
$IG_{\leq 17} = 0.31687$

Out of all the Information gain values, we get the max for In a relationship.

Based on this, we consider:-

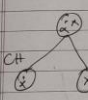
Mena
Pat
Tina.

For 2nd iteration:-



$E(\text{Parents}) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$

For city:-



$E(CH) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2$

$= 1$

$E(NY) = 0$

$E(\text{city}) = 0.5$ 0.66

$IG_{\text{city}} = 0.2583$

For liking Beyoncé :-

$IG_{Beyoncé} = 0.2583$

for age (Taking avg.) :-

19 → 31
43 → 48.5
54

For ≤31 :-

$IG_{\leq 31} = 0.2583$

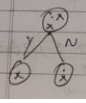
For ≤48.5 :-

$IG_{\leq 48.5} = 1$ → Maximum

for No. of concerts :- (Taking avg.)

2 → 4.5
7 → 15.5
20

for ≤4.5 :-



$IG_{\leq 4.5} = 0.2583$

For ≤15.5 :-

$IG_{\leq 15.5} = 0.2583$

The 2nd split is done using Age

Final Decision tree :-