Process Mining and Simulation

Assignment-1

Name: Abdullah Daoud

Roll No.: 22I-2626

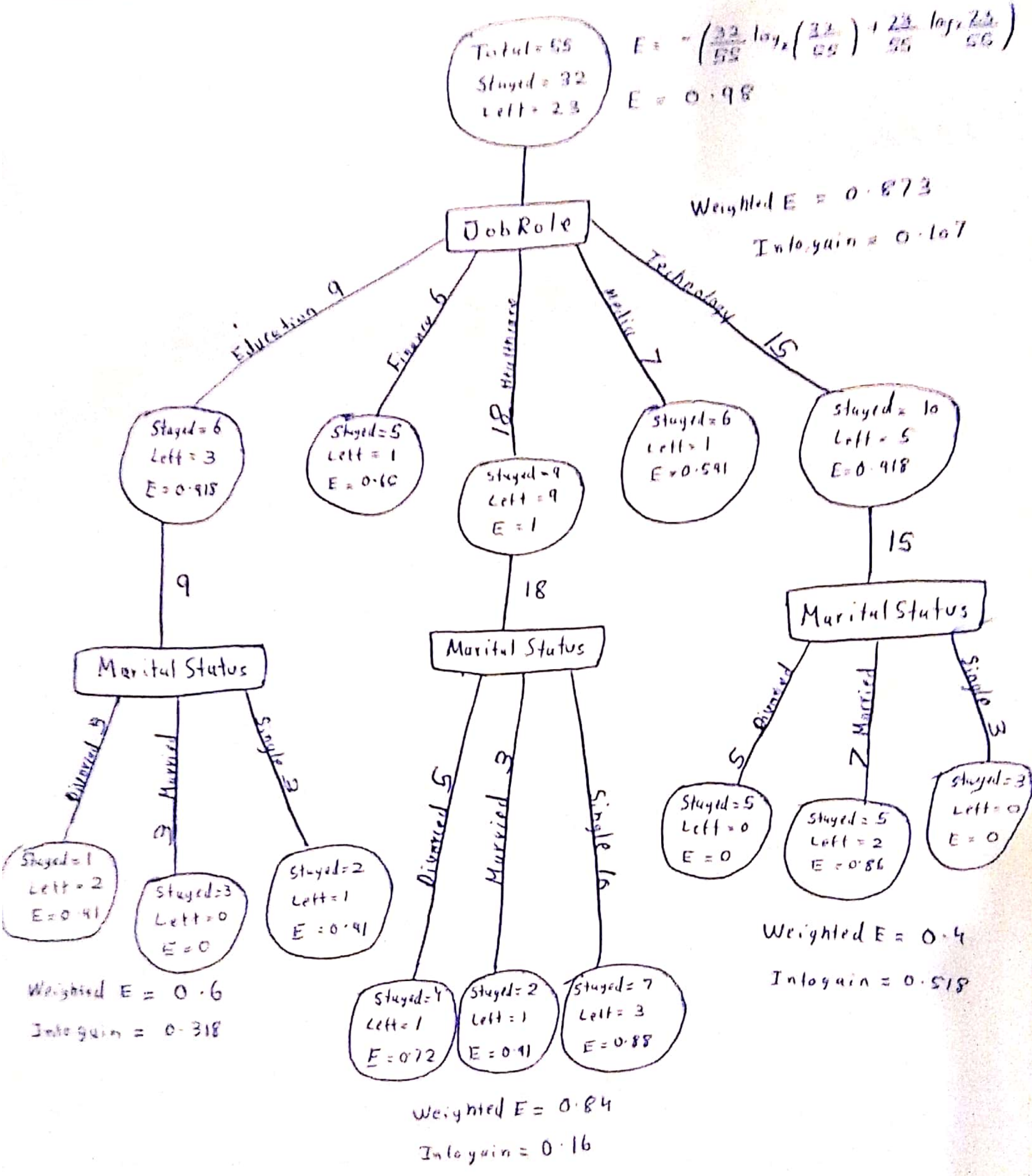
Section: SE-E

Submitted to: Dr. Behjat Zuhaira

Selecting Strong Attributes:

- The rows provided to me for the dataset are 1261 - 1315
- These rows include a total of 55 records and 24 variables.
- The Response variable for this dataset is Attrition which has 2 possible values Stayed and Left
- Strong attributes are those predictor variables that have the most effect on the Response variable
- To identify the strong attributes from the given dataset, we need to calculate the information gained after a split based on specific variables.
- To make the process intuitive, we will set the threshold of Information Gain to be 0.03 meaning that those variables whose information gain is 0.03 or above after one split will be considered as strong attributes.
- After performing rigorous calculations, here are the strong attributes for the 1261-1315 dataset
- Job Role = 0.107 information gain
- Marital Status = 0.102 information gain
- Company Size = 0.0852 information gain
- Job Level = 0.05 information gain
- Monthly Income = 0.05 information gain
- Based on these strong attributes, 3 decision trees will be created

Tree 1: JobRole → Marital Status



→ Depth of tree is 2

→ To calculate information gain of the whole tree, we will first find out the weighted Average Entropy of the whole tree and then subtract it from the root nodes Entropy

Weighted Average Entropy $= 3/55 \times 0.91 + 3/55 \times 0.91 + 6/55 \times 0.65$

$+ 5/55 \times 0.72 + 3/55 \times 0.91 + 10/55 \times 0.88$

$+ 7/55 \times 0.591 + 7/55 \times 0.86$

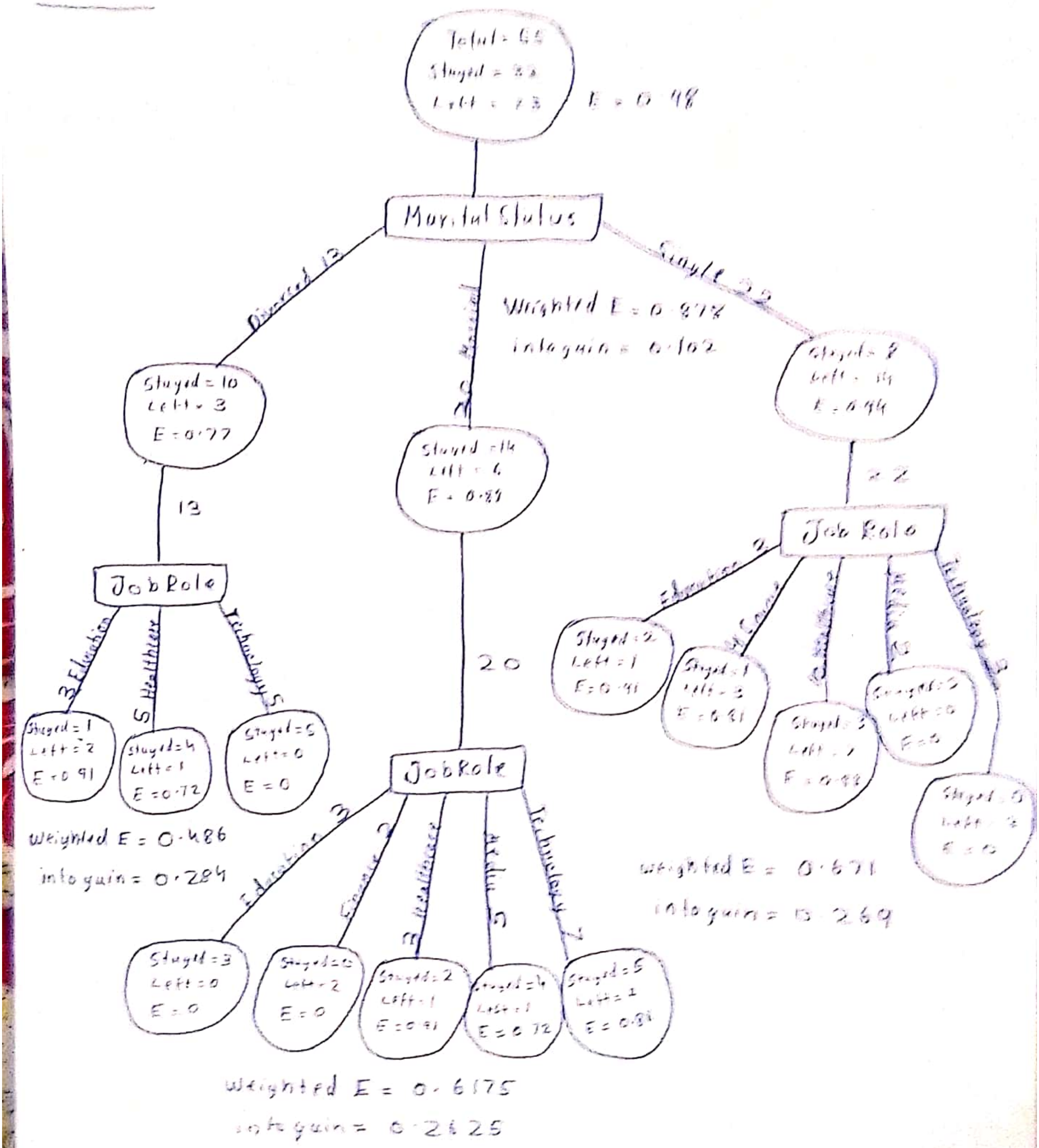
$= 0.629$

Information Gain $= 0.98 - 0.629 = 0.351$

→ So information gain of Tree1 $= 0.351$

(P.T.O)

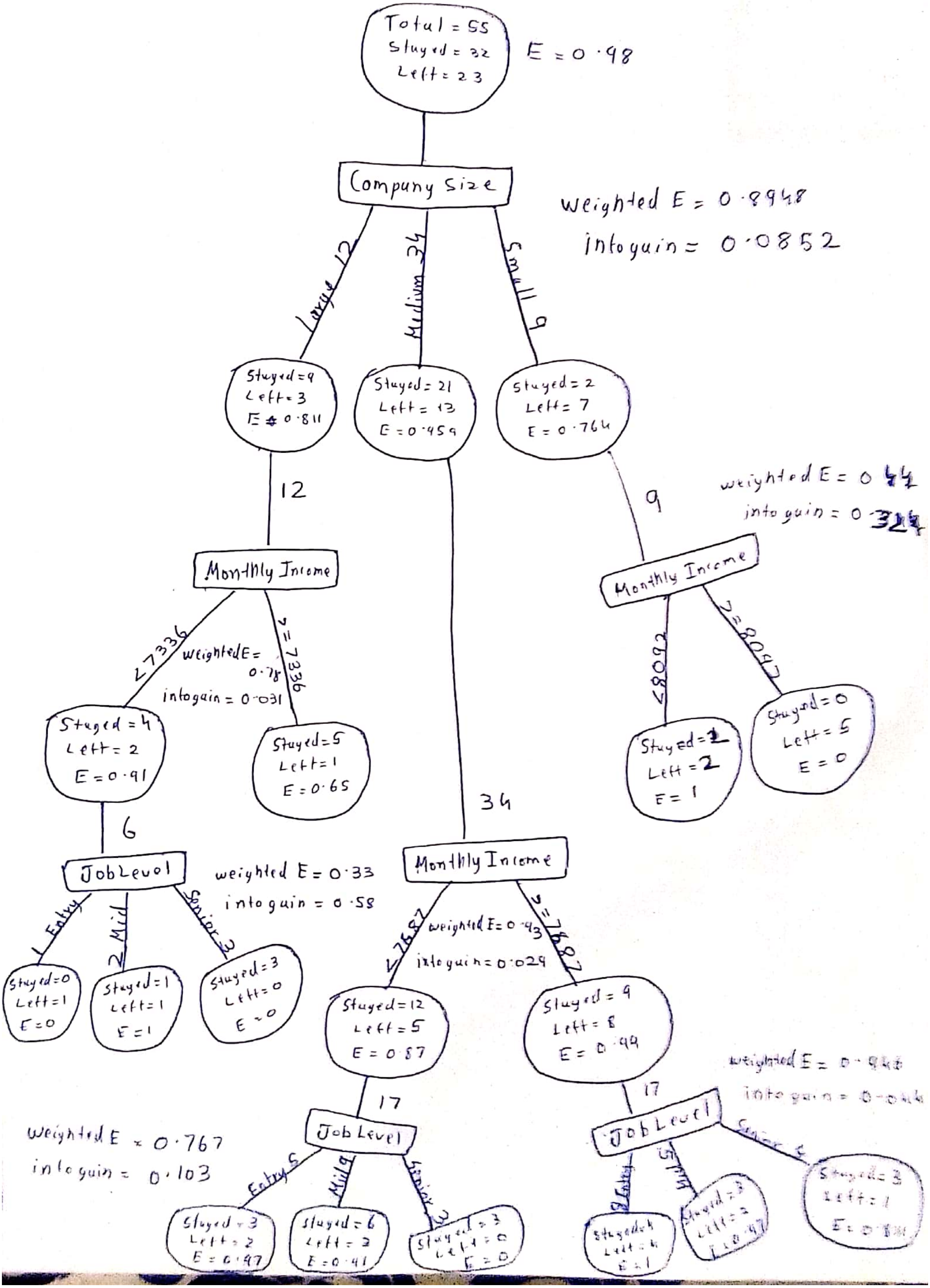
Tree 2: Marital Status → Job Role



weighted Average Entropy = 3/55 × 0.91 + 5/55 × 0.72 + 3/55 × 0.91
+ 5/55 × 0.72 + 7/55 × 0.86 + 3/55 × 0.91
+ 4/55 × 0.81 + 10/55 × 0.88
= 0.60

Information Gain = 0.98 − 0.60 = 0.38

Tree 3: Company Size → Monthly Income → Job Level



→ Depth of tree is 3

Weighted Average Entropy $= 2/55 \times 1 + 6/55 \times 0.65 + 5/55 \times 0.97$
$+ 9/55 \times 0.91 + 8/55 \times 1 + 5/55 \times 0.97$
$+ 4/55 \times 0.811 + 4/55 \times 1$
$= 0.709$

Information Gain $= 0.98 - 0.709 = 0.271$

Choosing Optimal Tree:

→ The information gains for the 3 decision trees are:

Tree 1 $= 0.351$

Tree 2 $= 0.38$

Tree 3 $= 0.271$

→ The most optimal tree is the one which gives the highest information gain

→ High information gain means that tree or specific node has better prediction capabilities

→ Since Tree 2 has the highest information gain of all 3 trees, therefore

Tree 2 is the most Optimal Tree.