BASIC DEFINITIONS

Support:

How often a rule appears in the database being mined.

- X → Y, support is the percentage of transactions that contain X and Y.
- $Support = |\{i \mid \{X, Y\} Ti\}|$

  E.g., $Support(Chicken, Clothes \rightarrow Milk) = 3/7 = 42.84\%$

Confidence:

The amount of times a given rule turns out to be true in practice.

- $Confidence = |\{i \mid \{X, Y\} Ti\}| \, |\{j \mid X \; Tj\}|$

  E.g., $Confidence(A \rightarrow B) = \frac{Support\ (A \cup B)}{Support(A)}$ …OR… $Confidence(B \rightarrow A) = \frac{Support\ (B \cup A)}{Support(B)}$

APRIORI ALGORITHM

Question: Given is the transaction table apply apriori algorithm. $MinimumSupport\ (minsup) = 50\%$ and $MinimumConfidence\ (minconf) = 75\%$.

TRANSACTION TABLE		
TransactionID	**Items**	**ItemID**
1	1-Bread, 2-Cheese, 3-Egg, 4-Juice	1,2,3,4
2	1-Bread, 2-Cheese, 4-Juice	1,2,4
3	1-Bread, 5-Milk, 6-Yogurt	1,5,6
4	1-Bread, 4-Juice, 5-Milk	1,4,5
5	2-Cheese, 4-Juice, 5-Milk	2,4,5

Solution:

Step1: Write all items in table form with frequency, percentage and min-sup qualification value (yes/no).

Step#1 – Items/ItemID, Frequencies, Percentages, Minimum Support Qualification				
ItemID	**Items**	**Frequencies**	**Percentages**	**Percentage ≥ min-sup?**
1	Bread	4/5	80%	80% ≥ 50% == YES
2	Cheese	3/5	60%	60% ≥ 50% == YES
3	Egg	1/5	20%	20% ≥ 50% == NO
4	Juice	4/5	80%	80% ≥ 50% == YES
5	Milk	3/5	60%	60% ≥ 50% == YES
6	Yogurt	1/5	20%	20% ≥ 50% == NO

- $Final\ Itemset1 = \{Bread, Cheese, Juice, Milk\}$ OR {1,2,4,5}

Step2: Create a new table having sets of two item following the Lexi-Order (no backward patching/set making) and repeat step1 for the obtained table.

Step#2 – Table of Having Sets of Two Items in Lexi-Order and Repeating Step#1				
ItemID	**Items**	**Frequencies**	**Percentages**	**Percentage ≥ min-sup?**
1,2	Bread, Cheese	2/5	40%	40% ≥ 50% == NO
1,4	Bread, Juice	3/5	60%	60% ≥ 50% == YES
1,5	Bread, Milk	2/5	40%	40% ≥ 50% == NO
2,4	Cheese, Juice	3/5	60%	60% ≥ 50% == YES
2,5	Cheese, Milk	1/5	20%	20% ≥ 50% == NO
4, 5	Juice, Milk	2/5	40%	40% ≥ 50% == NO

- $Final\ Itemset2 = \{\{Bread, Juice\}, \{Cheese, Juice\}\}\ OR\ \{(1,4), (2,4)\}$

Step3: Find $Confidence\ (A \rightarrow B)$ and $Confidence\ (B \rightarrow A)$ for both the sets in $Final\ Itemset2$.

FOR {Bread, Juice} OR {1,4}

$$Conf(Bread \rightarrow Juice) = \frac{Sup(Bread \cup Juice)}{Sup(Bread)} = \frac{\left(\frac{3}{5}\right)}{\left(\frac{4}{5}\right)} = \frac{3}{5} \times \frac{5}{4} = \frac{3}{4} = 75\%$$

$$Conf(Juice \rightarrow Bread) = \frac{Sup(Juice \cup Bread)}{Sup(Juice)} = \frac{\left(\frac{3}{5}\right)}{\left(\frac{4}{5}\right)} = \frac{3}{5} \times \frac{5}{4} = \frac{3}{4} = 75\%$$

FOR {Cheese, Juice} OR {2,4}

$$Conf(Cheese \rightarrow Juice) = \frac{Sup(Cheese \cup Juice)}{Sup(Cheese)} = \frac{\left(\frac{3}{5}\right)}{\left(\frac{3}{5}\right)} = \frac{3}{5} \times \frac{5}{3} = 1 = 100\%$$

$$Conf(Juice \rightarrow Cheese) = \frac{Sup(Juice \cup Cheese)}{Sup(Juice)} = \frac{\left(\frac{3}{5}\right)}{\left(\frac{4}{5}\right)} = \frac{3}{5} \times \frac{5}{4} = \frac{3}{4} = 75\%$$

MULTIPLE MINIMUM SUPPORT

Question: Solve with Multiple Minimum Support with the given data.

$MinimumSupport\ (minsup) = 50\%\ , MinimumConfidence\ (minconf) = 75\%\ and\ \varphi = 20\%$

TRANSACTION TABLE		
TranID	**Items**	**ItemID**
1	Bread, Egg, Juice	1,3,4
2	Cheese, Egg, Milk	2,3,5
3	Bread, Cheese, Egg, Milk	1,2,3,5
4	Cheese, Milk	2,5

MINIMUM ITEM SUPPORT (MIS)		
ItemID	**Items**	**MIS**
1	Bread	50%
2	Cheese	50%
3	Egg	50%
4	Juice	20%
5	Milk	50%

Solution:

Step#1: Sort the table based on MIS value, add their frequencies (support) and qualification values (yes/no). After doing all these, create a FinalSet and a Candidate set following the Lexi-Order.

	Step1: Sorting, Frequencies, Support, Qualifications				
ItemID	**Items**	**Freq**	**MIS**	**Support**	$\geq$ ***Min*(MIS)== ?**
4	Juice	1	20%	1/4 = 25%	$25\% \geq 20\% == YES$
1	Bread	2	50%	2/4 = 50%	$50\% \geq 20\% == YES$
2	Cheese	3	50%	3/4 = 75%	$75\% \geq 20\% == YES$
3	Egg	3	50%	3/4 = 75%	$75\% \geq 20\% == YES$
5	Milk	3	50%	3/4 = 75%	$75\% \geq 20\% == YES$

- $FinalSet1 = \{4,1,2,3,5\}$
- $DataSet1 = \{(4,1), (4,2), (4,3), (4,5), (1,2), (1,3), (1,5), (2,3), (2,5), (3,5)\}$.

Step#2: Start applying the formula on the pair that you just made and pass them to next step only if the qualify.

Formula = $MAX(Sup(i)) > MIN(MIS)\ \ AND\ \ |MAX(Sup(i)) - MIN(Sup(i))| < \varphi$

Step2: Formulating and Qualifying		
Sets	**Items**	$\boldsymbol{MAX(Sup(i)) > MIN(MIS)}$ ***AND*** $\boldsymbol{|MAX(Sup(i)) - MIN(Sup(i))| < \varphi}$
4,1	Juice, Bread	$50\% > 20\%\ \&\ |50\% - 25\%| = 25\% \ldots < \boldsymbol{\varphi} == NO$
4,2	Juice, Cheese	$75\% > 20\%\ \&\ |75\% - 25\%| = 50\% \ldots < \boldsymbol{\varphi} == NO$
4,3	Juice, Egg	$75\% > 20\%\ \&\ |75\% - 25\%| = 50\% \ldots < \boldsymbol{\varphi} == NO$
4,5	Juice, Milk	$75\% > 20\%\ \&\ |75\% - 25\%| = 50\% \ldots < \boldsymbol{\varphi} == NO$
1,2	Bread, Cheese	$75\% > 20\%\ \&\ |50\% - 25\%| = 25\% \ldots < \boldsymbol{\varphi} == NO$
1,3	Bread, Egg	$75\% > 20\%\ \&\ |75\% - 50\%| = 25\% \ldots < \boldsymbol{\varphi} == NO$
1,5	Bread, Milk	$75\% > 20\%\ \&\ |75\% - 50\%| = 25\% \ldots < \boldsymbol{\varphi} == NO$
2,3	Cheese, Egg	$75\% > 20\%\ \&\ |75\% - 75\%| = 0\% \ldots < \boldsymbol{\varphi} == YES$
2,5	Cheese, Milk	$75\% > 20\%\ \&\ |75\% - 75\%| = 0\% \ldots < \boldsymbol{\varphi} == YES$
3,5	Egg, Milk	$75\% > 20\%\ \&\ |75\% - 75\%| = 0\% \ldots < \boldsymbol{\varphi} == YES$

- $FinalSet2 = \{(2,3), (2,5), (3,5)\}$

Start generalization – joining and pruning. A rule where the last digits of two sets (or more) are different but rest digits are the same are kept, and sets, that do not follow this rule are discarded.

- (2,3) and (2,5) are the two sets, whose last digits are the same '5' and rest are different '2' and '3' so (2,3) and (2,5) are considered and (3,5) is discarded.
- $(2,3) \rightarrow (2,5) \rightarrow (2,3,5)$
- $DataSet2 = \{(2,3,5)\}$

Step#3: Rule generation is done in this part on the finalized $DataSet2 = \{(2,3,5)\}$.

	$\boldsymbol{Confidence = \frac{Support(A \cup B)}{Support(A)} > minconf}$
$\boldsymbol{2,3 \rightarrow 5}$	$\left(\frac{2}{4}\right) \div \left(\frac{2}{4}\right) = \frac{2}{4} \times \frac{4}{2} = 1 = 100\% \quad \ldots > minconf\ == YES$
$\boldsymbol{2,5 \rightarrow 3}$	$\left(\frac{2}{4}\right) \div \left(\frac{3}{4}\right) = \frac{2}{4} \times \frac{4}{3} = \frac{2}{3} = 66.66\% \quad \ldots > minconf\ == YES$
$\boldsymbol{3,5 \rightarrow 2}$	$\left(\frac{2}{4}\right) \div \left(\frac{2}{4}\right) = \frac{2}{4} \times \frac{4}{2} = 1 = 100\% \quad \ldots > minconf\ == YES$
$\boldsymbol{5 \rightarrow 2,3}$	$\left(\frac{2}{4}\right) \div \left(\frac{3}{4}\right) = \frac{2}{4} \times \frac{4}{3} = \frac{2}{3} = 66.66\% \quad \ldots > minconf\ == YES$

$3 \rightarrow 2,5$	$\left(\frac{2}{4}\right) \div \left(\frac{3}{4}\right) = \frac{2}{4} \times \frac{4}{3} = \frac{2}{3} = 66.66\%$ $\quad \ldots > minconf == YES$
$2 \rightarrow 3,5$	$\left(\frac{2}{4}\right) \div \left(\frac{3}{4}\right) = \frac{2}{4} \times \frac{4}{3} = \frac{2}{3} = 66.66\%$ $\quad \ldots > minconf == YES$

NAIVE BAYES CLASSIFICATION

Question: Apply the Bayesian classification on the following dataset.

DAY	OUTLOOK	TEMP	HUMIDITY	WIND	PLAY
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Solution:

Step#1: Calculating some basic things.

$Total\ ROWS = 14$	
$Total\ YES = 9$	$Total\ NO = 5$
$Probability\ (YES) = 9/14$	$Probability\ (NO) = 5/14$

Step#2: Make a table for every attribute. (4 attributes, 4 tables)

Attribute and Table#1: OUTLOOK		
	YES	NO
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5
Total	**9/9**	**5/5**

Attribute and Table#2: TEMPERATURE		
	YES	NO
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5
Total	**9/9**	**5/5**

Attribute and Table#3: HUMIDITY		
	YES	NO
High	3/9	4/5
Normal	6/9	1/5
Total	**9/9**	**5/5**

Attribute and Table#4: WIND		
	YES	NO
Strong	3/9	3/5
Weak	6/9	2/5
Total	**9/9**	**5/5**

Predict(X), X= Sunny, Cool, High, Strong

- Take positive for all and multiply them
- Take negative for all and multiply them
- Compare both, the greater value is the prediction

E.g.,

- Probability (YES for X) = $\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = 0.005$

- Probability (NO for X) = $\frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = 0.02$
- $0.02 > 0.005$ so the label is NO.

DECISION TREE

Question: Make the decision tree of the dataset.

AGE	HAS JOB	OWNS HOUSE	CREDIT RATING	LOAN APPROVAL
Young	False	False	Fair	**No**
Young	False	False	Good	**No**
Young	True	False	Good	**Yes**
Young	True	True	Fair	**Yes**
Young	False	False	Fair	**No**
Middle	False	False	Fair	**No**
Middle	False	False	Good	**No**
Middle	True	True	Good	**Yes**
Middle	False	True	Excellent	**Yes**
Middle	False	True	Excellent	**Yes**
Old	False	True	Excellent	**Yes**
Old	False	True	Good	**Yes**
Old	True	False	Good	**Yes**
Old	True	False	Excellent	**Yes**
Old	False	False	Fair	**No**

Solution: We need to remember to count total rows, total YES(Positive), NO(Negative) and some fundamental formulae for calculating decision tree.

Step#1: Calculating the initial steps.

$Total\ ROWS = 15$	
$Total\ YES = 9$	$Total\ NO = 6$

- Impurity (Entropy) in the dataset = $I(Yes, No) = I(Positive, Negative) = I(9,6) =$
- $I(9,6) = \sum_{i=1}^{c} -p_i \log_2(p_i) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$
  - $= -\frac{9}{9+6}\log_2\left(\frac{9}{9+6}\right) - \frac{6}{9+6}\log_2\left(\frac{6}{9+6}\right)$
  - $= -\frac{9}{15}\log_2\left(\frac{9}{15}\right) - \frac{6}{15}\log_2\left(\frac{6}{15}\right)$
  - $= -0.6(-0.73) - 0.4(-1.32)$
  - $= 0.97$

Step#2: Keep making tables for the counts of total positives, negatives for a particular value of an attribute along with their entropies and information gain.

Attribute and Table#1: AGE			
	POSITIVE	**NEGATIVE**	**I(AGE)**
YOUNG	2	3	$I(2,3) = 0.97$
MIDDLE	3	2	$I(3,2) = 0.97$
OLD	4	1	$I(4,1) = 0.72$
Total	**9**	**6**	

- $= \sum_{i=1}^{total} I(v_1, v_2)_i \left(\frac{p+n}{total}\right)_i$
  - $= \sum 0.97\left(\frac{2+3}{15}\right) + 0.97\left(\frac{3+2}{15}\right) + 0.72\left(\frac{4+1}{15}\right)$
  - $= \sum 0.97\left(\frac{5}{15}\right) + 0.97\left(\frac{5}{15}\right) + 0.72\left(\frac{5}{15}\right)$
  - $= 0.886$
- $Gain = I(Dataset) - \sum(Age) = 0.97 - 0.88 = 0.09$

Attribute and Table#2: HAS JOB			
	POSITIVE	**NEGATIVE**	**I(AGE)**
TRUE	5	0	$I(5,0) = 0$
FALSE	4	6	$I(4,6) = 0.97$
Total	**9**	**6**	

- $= \sum_{i=1}^{total} I(v_1, v_2)_i \left(\frac{p+n}{total}\right)_i$
  - $= \sum 0\left(\frac{5+0}{15}\right) + 0.97\left(\frac{4+6}{15}\right)$

- $= \sum 0 + 0.97\left(\frac{10}{15}\right)$
- $= 0.646 = 0.65$

- $Gain = I(Dataset) - \sum(HAS\ JOB) = 0.97 - 0.65 = 0.32$

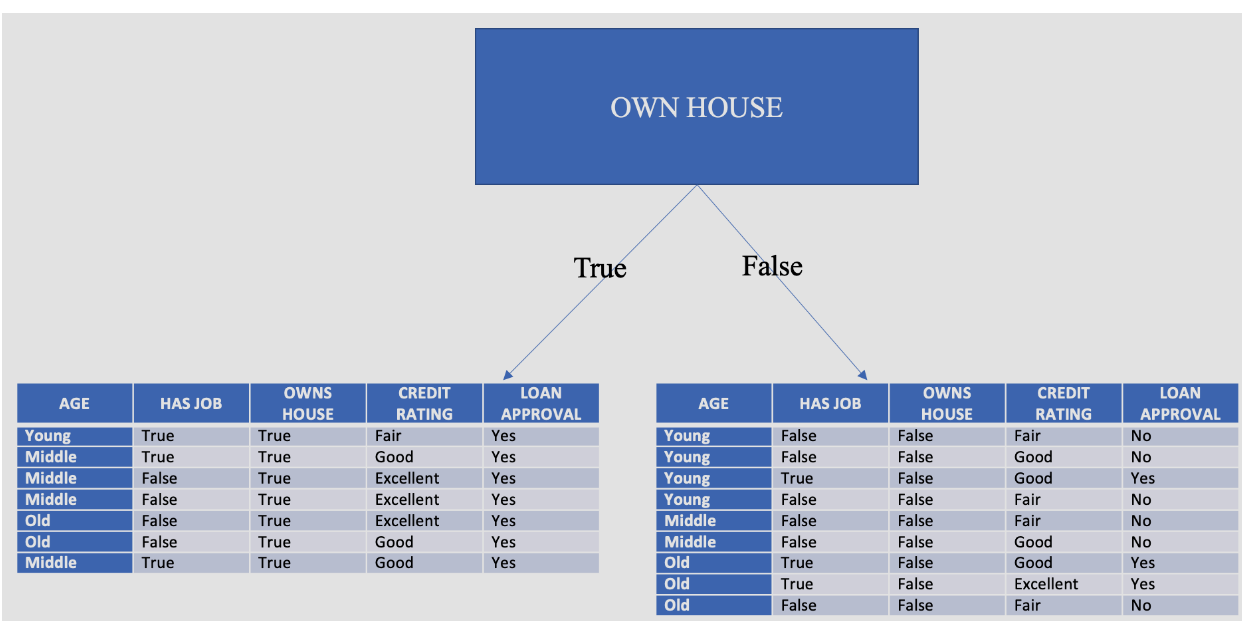
Attribute and Table#3: OWNS HOUSE			
	POSITIVE	**NEGATIVE**	**I(AGE)**
TRUE	6	0	$I(6,0) = 0$
FALSE	3	6	$I(3,6) = 0.92$
Total	**9**	**6**	

- $= \sum_{i=1}^{total} I(v_1, v_2)_i \left(\frac{p+n}{total}\right)_i$
    - $= \sum 0\left(\frac{6+0}{15}\right) + 0.92\left(\frac{3+6}{15}\right)$
    - $= \sum 0 + 0.92\left(\frac{9}{15}\right)$
    - $= 0.55$
- $Gain = I(Dataset) - \sum(HAS\ JOB) = 0.97 - 0.55 = 0.42$

Attribute and Table#4: CREDIT RATING			
	POSITIVE	**NEGATIVE**	**I(AGE)**
FAIR	1	4	$I(1,4) = 0.72$
GOOD	4	2	$I(4,2) = 0.92$
EXCELLENT	4	0	$I(4,0) = 0$
Total	**9**	**6**	

- $= \sum_{i=1}^{total} I(v_1, v_2)_i \left(\frac{p+n}{total}\right)_i$
    - $= \sum 0.72\left(\frac{1+4}{15}\right) + 0.92\left(\frac{4+2}{15}\right) + 0\left(\frac{4+0}{15}\right)$
    - $= \sum 0.72\left(\frac{5}{15}\right) + 0.92\left(\frac{6}{15}\right) + 0$
    - $= 0.24 + 0.368 = 0.608$
- $Gain = I(Dataset) - \sum(HAS\ JOB) = 0.97 - 0.608 = 0.36$

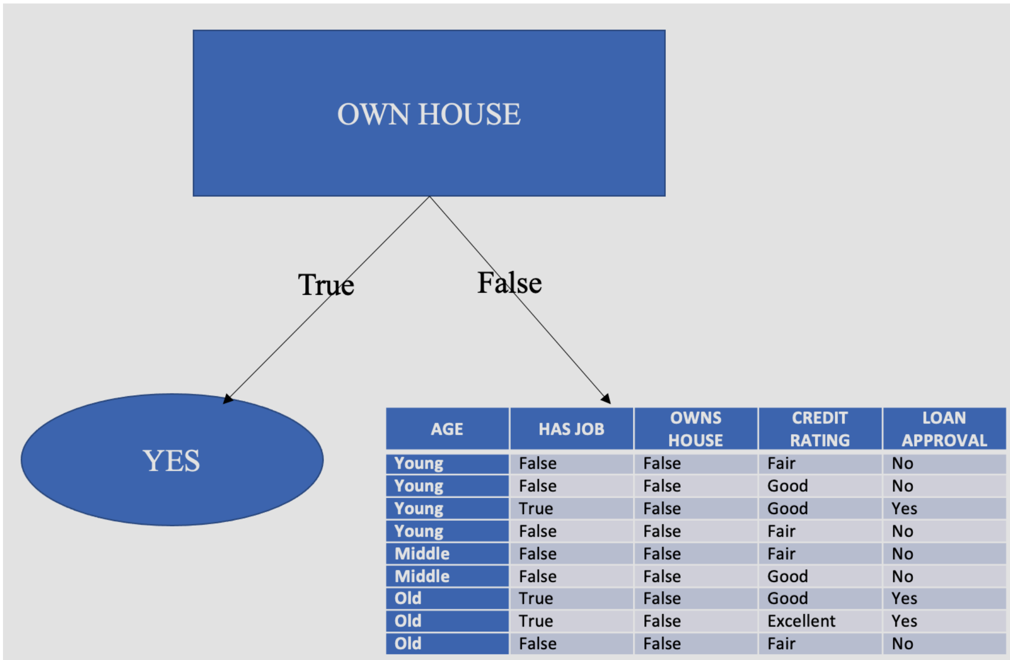
NOTE: Since the gain of OWNS HOUSE is highest amongst all, it is going to be the root node.



AGE	HAS JOB	OWNS HOUSE	CREDIT RATING	LOAN APPROVAL
Young	True	True	Fair	Yes
Middle	True	True	Good	Yes
Middle	False	True	Excellent	Yes
Middle	False	True	Excellent	Yes
Old	False	True	Excellent	Yes
Old	False	True	Good	Yes
Middle	True	True	Good	Yes

AGE	HAS JOB	OWNS HOUSE	CREDIT RATING	LOAN APPROVAL
Young	False	False	Fair	No
Young	False	False	Good	No
Young	True	False	Good	Yes
Young	False	False	Fair	No
Middle	False	False	Fair	No
Middle	False	False	Good	No
Old	True	False	Good	Yes
Old	True	False	Excellent	Yes
Old	False	False	Fair	No

- Owning a house is always giving a 'YES' label, so we will make it a leaf node and will perform the same operations for the remaining dataset (giving on the right).



- The remaining dataset is given below.

AGE	HAS JOB	OWNS HOUSE	CREDIT RATING	LOAN APPROVAL
Young	False	False	Fair	No
Young	False	False	Good	No
Young	True	False	Good	Yes
Young	False	False	Fair	No
Middle	False	False	Fair	No
Middle	False	False	Good	No
Old	True	False	Good	Yes
Old	True	False	Excellent	Yes
Old	False	False	Fair	No

Step#2: Repeat the same procedure for the remaining dataset.

$Total\ ROWS\ =\ 9$	
$Total\ YES\ =\ 3$	$Total\ NO\ =\ 6$

- Impurity (Entropy) in the dataset = $I(Yes, No) = I(Positive, Negative) = I(3,6) =$
- $I(3,6) = \sum_{i=1}^{c} -p_i \log_2(p_i) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$
  - $= -\frac{3}{3+6}\log_2\left(\frac{3}{3+6}\right) - \frac{6}{3+6}\log_2\left(\frac{6}{3+6}\right)$
  - $= -\frac{3}{9}\log_2\left(\frac{3}{9}\right) - \frac{6}{9}\log_2\left(\frac{6}{9}\right)$
  - $= -0.33(-1.584) - 0.66(-0.584)$
  - $= 0.908 = 0.91$

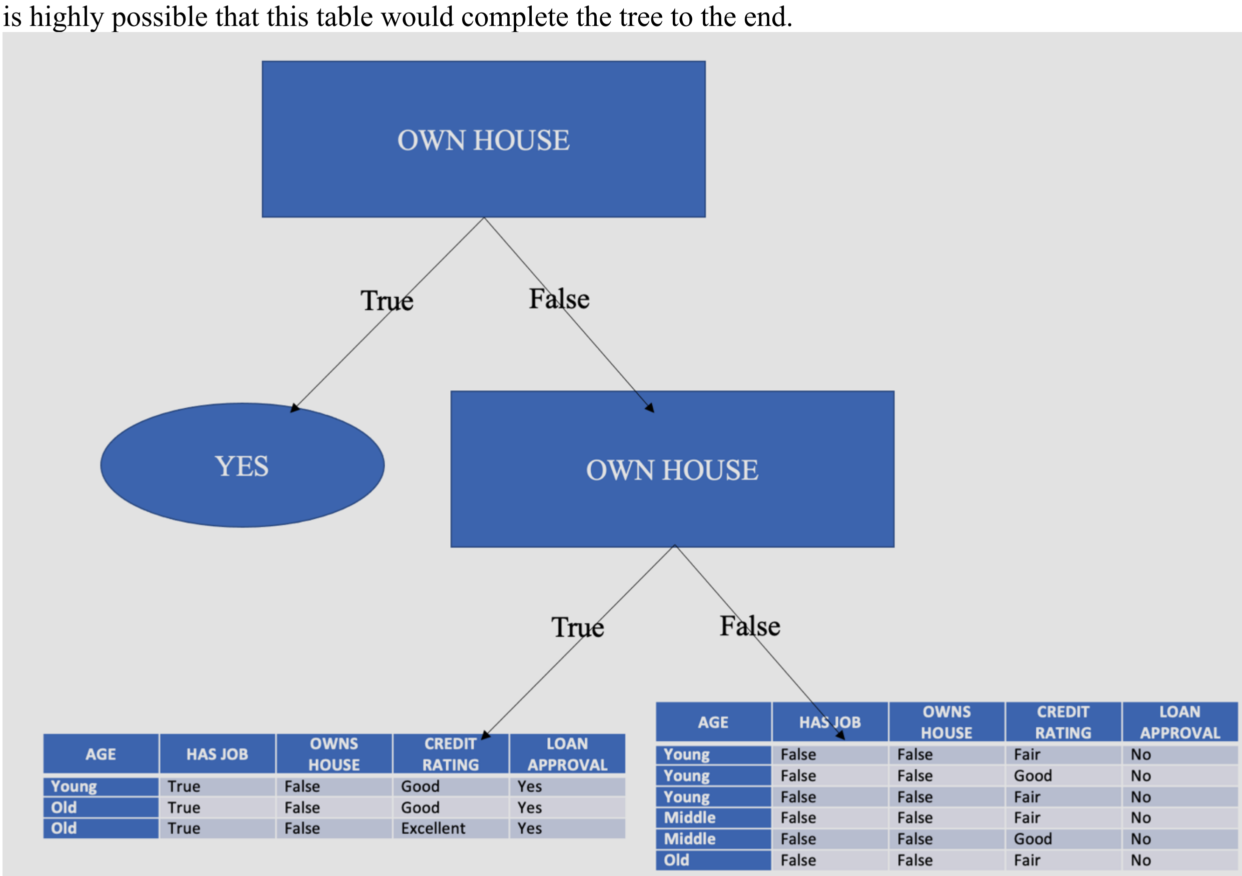
Attribute and Table#1: AGE			
	POSITIVE	**NEGATIVE**	**I(AGE)**
YOUNG	1	3	$I(1,3) = 0.811$
MIDDLE	0	2	$I(0,2) = 0$
OLD	2	1	$I(2,1) = 0.918$
Total	**3**	**6**	

- $= \sum_{i=1}^{total} I(v_1, v_2)_i \left(\frac{p+n}{total}\right)_i$
  - $= \sum 0.811\left(\frac{1+3}{9}\right) + 0\left(\frac{0+2}{9}\right) + 0.918\left(\frac{2+1}{9}\right)$
  - $= \sum 0.811\left(\frac{4}{9}\right) + 0\left(\frac{2}{9}\right) + 0.918\left(\frac{3}{9}\right)$
  - $= 0.30 + 0 + 0.306\ = 0.606$
- $Gain = I(NewDataset) - \sum(Age) = 0.91 - 0.60 = 0.31$

Attribute and Table#2: HAS JOB			
	POSITIVE	**NEGATIVE**	**I(AGE)**
TRUE	3	0	$I(3,0) = 0$
FALSE	0	6	$I(0,6) = 0$
Total	**3**	**6**	

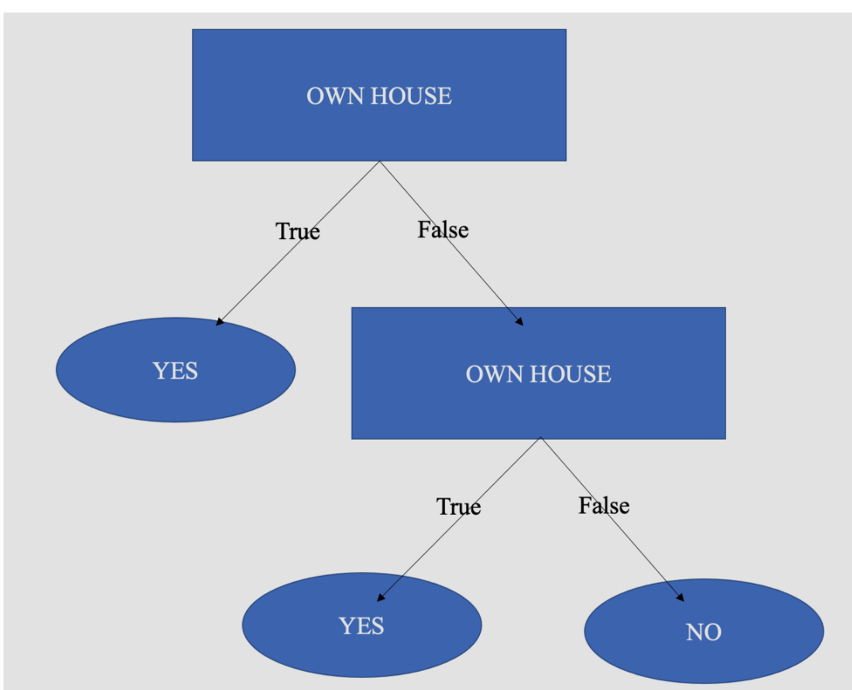
- $= \sum_{i=1}^{total} I(v_1, v_2)_i \left(\frac{p+n}{total}\right)_i$
  - $= \sum 0 \left(\frac{3+0}{9}\right) + 0 \left(\frac{0+6}{9}\right)$
  - $= \sum 0 + 0$
  - $= 0$
- $Gain = I(Dataset) - \sum(HAS\ JOB) = 0.97 - 0 = 0.97$

NOTE: You can keep going forward but since the Entropy/Impurity is ‘0’ and gain is extremely high, so it is highly possible that this table would complete the tree to the end.



AGE	HAS JOB	OWNS HOUSE	CREDIT RATING	LOAN APPROVAL
Young	True	False	Good	Yes
Old	True	False	Good	Yes
Old	True	False	Excellent	Yes

AGE	HAS JOB	OWNS HOUSE	CREDIT RATING	LOAN APPROVAL
Young	False	False	Fair	No
Young	False	False	Good	No
Young	False	False	Fair	No
Middle	False	False	Fair	No
Middle	False	False	Good	No
Old	False	False	Fair	No



- This is how the final decision tree must look like.