SE-488

Assignment-1

Question-1

Training	Fever	Vomiting	Diarrhea	Shivering	classification
D_1	No	no	no	no	healthy (H)
D_2	avg	no	no	no	influenza (I)
D_3	high	no	no	yes	influenza (I)
D_4	high	yes	yes	no	S P (S)
D_5	avg	no	yes	no	S P (S)
D_6	no	yes	yes	no	B I (B)
D_7	avg	yes	yes	no	B I (B)

1.1 Calculate entropy of the dataset

$$E(S) = \sum_{i=1}^{c} - P_i \log_2 P_i$$

Total no. of classification - 7
for Healthy - 1 (H)
for Influenza - 2 (I)
for Salmonella poisoning - 2 (S)
for Bowel Inflamation - 2 (B)

Now By applying the formula of Entropy

$E(D) = \sum_{i=1}^{c} - P_i \log_2 P_i$, we get

$$E(D) = -\left(\left(\frac{1}{7}\right)\log_2\left(\frac{1}{7}\right) + \left(\frac{2}{7}\right)\log_2\left(\frac{2}{7}\right) + \left(\frac{2}{7}\right)\log_2\left(\frac{2}{7}\right) + \left(\frac{2}{7}\right)\log_2\left(\frac{2}{7}\right)\right)$$

$$\boxed{= 1.950212}$$ Entropy of Dataset

1.2 Calculate the information gain (IG) for each of the attribute

Total no. of attributes - 4

Fever, vomiting, Diarrhea, Shivering

- First we need to calculate entropy of each attribute and we'll also calculate weighted entropy with

→ $\text{Entropy}(x, y) = \sum_{c \in y} P(c)\, E(c)$

1 E(classification, fever) → no, avg, high
2 E(classification, Vomiting) → no, yes
3 E(classification, diarrhea) → no, yes
4 E(classification, Shivering) → no, yes

① For Fever $E(\text{fever}) = P(\text{no})E(\text{no}) + P(\text{avg})E(\text{avg}) + P(\text{high})E(\text{high})$

for $E(\text{no}) = -\left(\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right)\right)$

$= 1$

for $E(\text{avg}) = -\left(\left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right)\right)$

$= 1.585$

For $E(\text{high}) = -\left(\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right)\right)$

$= 1$

		H	I	class	S	B	
	No	1	0		0	1	– 2
Fever	avg	0	1		1	1	– 3
	high	0	1		1	0	– 2

For fever

$$\text{Entropy(class, Fever)} = \left(\left(\frac{2}{7}\times 1\right) + \left(\frac{3}{7}\right)\times 1.585 + \left(\frac{2}{7}\times 1\right)\right) = \boxed{1.2507}$$

→ for E(vomiting) = P(no) E(no) + P(yes) E(yes)

$$E(no) = -\left(\left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) + \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) + \left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right)\right)$$

$$= 1.5$$

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

$$= 0.9183$$

		H	I	S	B	
	No	1	2	1	0	– 4
Vomiting	Yes	0	0	1	2	– 3

$$\text{Entropy(class, Vomit)} = \left(\left(\frac{3}{7}\times 0.9183\right) + \left(\frac{4}{7}\times 1.5\right)\right)$$

$$= 1.2507$$

→ for E(Diarrhea) = P(no) E(no) + P(yes) E(yes)

$$E(no) = -\left(\left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) + \left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right)\right)$$

$$= 0.9183$$

$$E(yes) = -\left(\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) + \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right)\right)$$

$$= 1$$

		H	I	S	B	
Diarrhea	Yes	0	0	2	2	– 4
	No	1	2	0	0	– 3

$$E(\text{class, Diarrhea}) = \left(\left(\frac{3}{7} \times 0.9183\right) + \left(\frac{4}{7} \times 1\right)\right)$$

$$= 0.9649$$

→ for $E(\text{Shivering}) = P(\text{no})\, E(\text{no}) + P(\text{yes})\, E(\text{yes})$

$$E(\text{no}) = -\left(\left(\frac{1}{6}\right)\log_2\left(\frac{1}{6}\right) + \left(\frac{1}{6}\right)\log_2\left(\frac{1}{6}\right) + \left(\frac{2}{6}\right)\log_2\left(\frac{2}{6}\right) + \left(\frac{2}{6}\right)\log_2\left(\frac{2}{6}\right)\right)$$

$$= 1.9183$$

$$E(\text{yes}) = -\left(\left(\frac{1}{1}\right)\log_2\left(\frac{1}{1}\right)\right)$$

$$= 0$$

		H	I	S	B	
Shivering	Yes	0	1	0	0	– 1
	No	1	1	2	2	– 6

$$\text{Entropy(class, Shivering)} = \left(\left(\frac{6}{7} \times 1.9183\right) + \left(\frac{1}{7} \times 0\right)\right)$$

$$= 1.6442$$

→ calculating Information Gain (IG) for each Attribute

$$\text{Gain}(x, y) = \text{Entropy}_x(x) - \text{Entropy}(x|y)$$

→ Gain(class, fever) = E(class) − E(class, fever)

= 1.9502 − 1.2507

= 0.6995

→ Gain(class, Vomiting) = E(class) − E(class, Vomiting)

= 1.9502 − 1.2507

= 0.6995

→ Gain(class, Diarrhea) = E(class) − E(class, Diarrhea)

= 1.9502 − 0.9649

= 0.9853

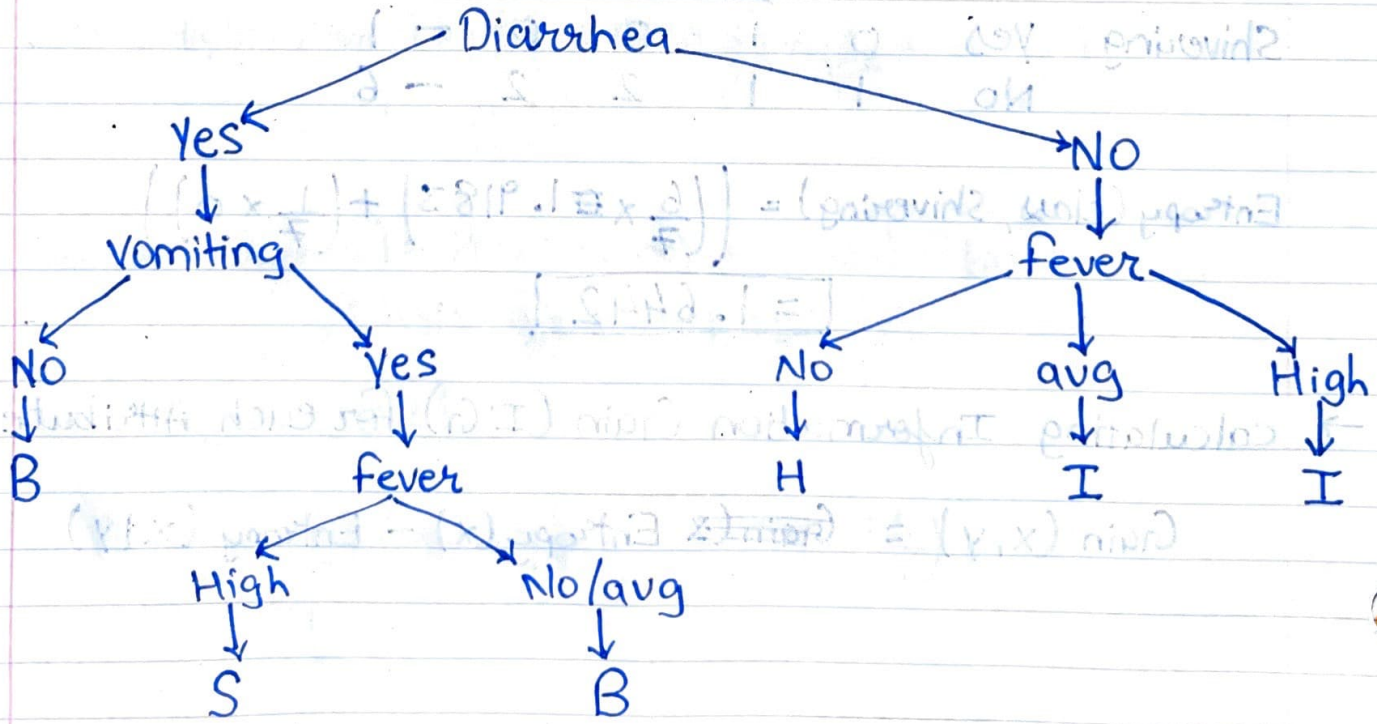
→ Gain(class, Shivering) = E(class) − E(class, Shivering)

= 1.9502 − 1.6442

= 0.3060

1.3 → Highest(I G) is of Diarrhea, So that Diarrhea will be root node

1.4



Question 2

Instance	a_1	a_2	a_3	Target class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

2.1 → Else calculate the entropy of dataset

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Total no. of target class / Instance = 9
for + = 4
for − = 5

By applying the formula, we get.

$$E(D) = -\left(\left(\frac{4}{9}\right)\log_2\left(\frac{4}{9}\right) + \left(\frac{5}{9}\right)\log_2\left(\frac{5}{9}\right)\right)$$

$$= 0.9910$$ Entropy of Dataset

2.2 Calculate Information Gain (IG) of a_1 & a_2
a_1 − T F
a_2 − T F
We'll also calculate weighted entropy

Entropy $(x, y) = \sum_{c \in y} P(c)\, E(c)$

1 E(Target, a_1) - T F
2 F(Target, a_2) - T F

$E(a_1) = P(T)\, E(T) + P(F)\, E(F)$

$E(T) = -\left(\left(\frac{3}{4}\right) \log_2 \left(\frac{3}{4}\right) + \left(\frac{1}{4}\right) \log_2 \left(\frac{1}{4}\right)\right)$

$= 0.8112$

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

$= 0.7219$

		+	−	
a_1	T	3	1	4
	F	1	4	5

Entropy (Target, a_1) $= \left(\left(\frac{4}{9} \times 0.8112\right) + \left(\frac{5}{9} \times 0.7219\right)\right)$

$= 0.7615$

For $E(a_2) = P(T)\, E(T) + P(F)\, E(F)$

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

$= 0.9709$

$E(F) = -\left(\left(\frac{2}{4}\right) \log_2 \left(\frac{2}{4}\right) + \left(\frac{2}{4}\right) \log_2 \left(\frac{2}{4}\right)\right)$

$= 1$

		+	−
a_2	T	2	3 − 5
	F	2	2 − 4

Entropy(Target, a_2) $= \left(\frac{5}{9} \times 0.9709\right) + \left(\frac{4}{9} \times 1\right)$

$= 0.9838$

- Calculating Information Gain (IG)

2.2 ~~Entropy~~ (Target, a_1) = E(Target) − E(Target, a_1)

Gain $= 0.9910 - 0.7615$

$IG = 0.2295$

~~Entropy~~(Target, a_2) = E(Target) − E(Target, a_2)

Gain $= 0.9910 - 0.9838$

$IG = 0.0072$

2.3 Based on IG (higher) the best split will be with a_1

Question 3

ID	time	gender	area	Risk
1	1-2	M	Urban	low
2	2-7	M	Rural	high
3	>7	F	Rural	low
4	1-2	F	Rural	high
5	>7	M	Rural	high
6	1-2	M	Rural	high
7	2-7	F	Urban	low
8	2-7	M	Urban	low

3.1 Create a decision tree

- first we need entropy of the dataset
- Second we need Information gain to create the best split for the decision tree

→ calculating entropy of the dataset

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Total no. of risk = 8
for High = 4
for low = 4

By applying the formula, we get

$$E(R) = -\left(\left(\frac{4}{8}\right)\log_2\left(\frac{4}{8}\right) + \left(\frac{4}{8}\right)\log_2\left(\frac{4}{8}\right)\right)$$

$$= 1$$ Entropy of dataset.

Now, we need to find the entropy of the attributes.
E(Risk, area) - Urban rural
E(Risk, gender) - Male, female
E(Risk, time) - 1-2, 2-7, >7

$E(area) = P(urban)\,E(urban) + P(rural)\,E(rural)$

$E(urban) = -\left(\left(\frac{3}{3}\right)\log_2\left(\frac{3}{3}\right)\right)$

$= 0$

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

$= 0.7219$

	high	low		
area urban	0	3	-	3
rural	4	1	-	5

$Entropy(Risk, area) = \left(\left(\frac{3}{8}\times 0\right) + \left(\frac{5}{8}\times 0.7219\right)\right)$

$= 0.4511$

$E(gender) = P(M)\,E(M) + P(F)\,E(F)$

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

$= 0.9709$

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

$= 0.9182$

		Risk		
		High	low	
gender	M	3	2	- 5
	F	1	2	- 3

$E(\text{~~area~~ Risk, gender}) = \left(\left(\frac{5}{8} \times 0.9709\right) + \left(\frac{3}{8} \times 0.9182\right)\right)$

$= 0.9511$

$E(\text{time}) = P(1\text{-}2)\,E(1\text{-}2) + P(2\text{-}7)E(2\text{-}7) + P(>7)\,E(>7)$

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

$= 0.9182$

$E(2\text{-}7) = -\left(\left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) + \left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right)\right)$

$= 0.9182$

$E(>7) = -\left(\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right)\right)$

$= 1$

		high	low	
Time	1-2	2	1	- 3
	2-7	1	2	- 3
	>7	1	1	- 2

$\text{Entropy (Risk, time)} = \left(\left(\frac{3}{8} \times 0.9182\right) + \left(\frac{3}{8} \times 0.9182\right) + \left(\frac{2}{8} \times 1\right)\right)$

$= 0.9386$

→ Calculating Information gain (IG)

$\text{Gain (Risk, area)} = E(\text{Risk}) - E(\text{Risk, area})$

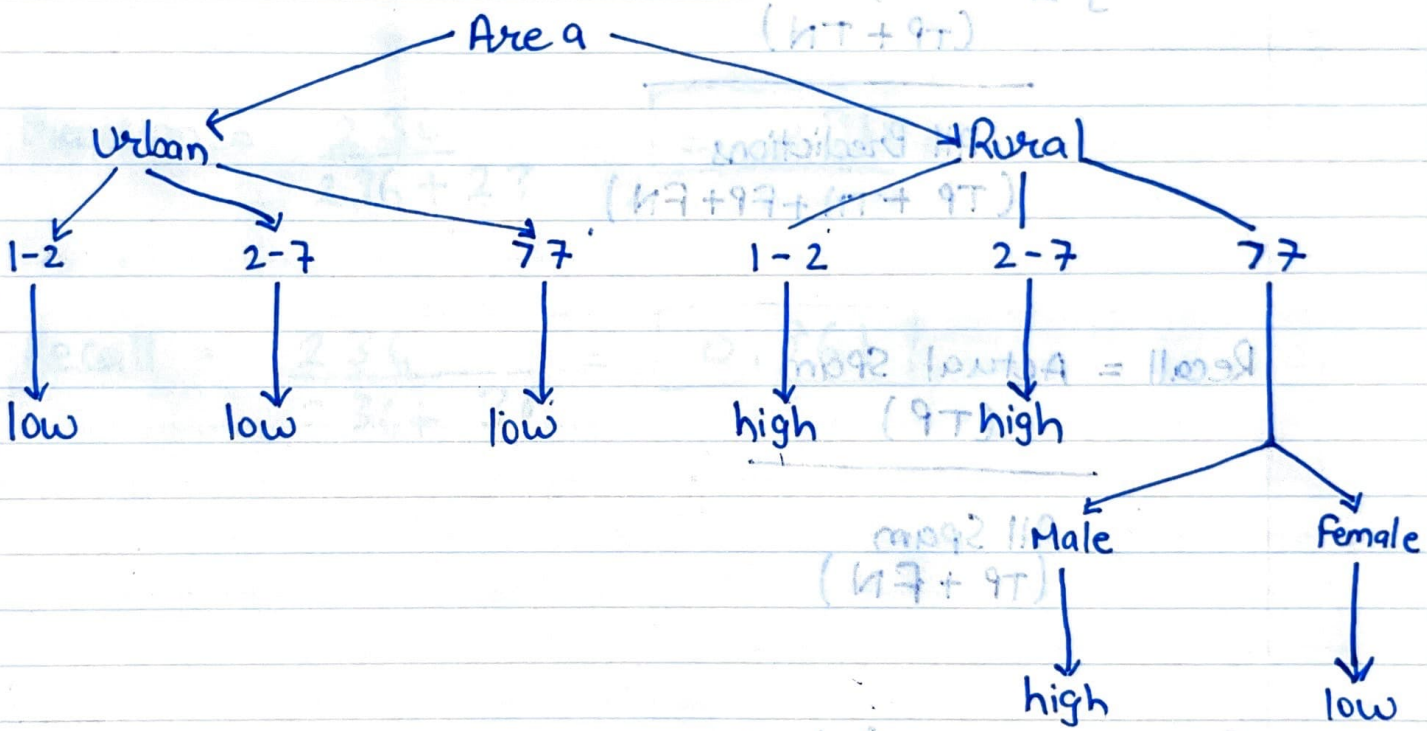
$= 1 - 0.4511$

$= 0.5489$

Gain(Risk, gender) = E(Risk) − E(Risk, gender)
= 1 − 0.9511
= 0.0489

Gain(Risk, time) = E(Risk) − E(Risk, time)
= 1 − 0.9386
= 0.0614

3.1 Now, we can make the decision tree with highest IG as root node:



3.2 Predict result:

ID	time	Gender	area	Risk (Predicted)
A	1-2	F	rural	High
B	2-7	M	urban	low
C	1-2	F	urban	low

Question 4

4.3

		Predicted: positive	Predicted: negative
Actual	positive	True Positive	False negative
	negative	False Positive	true negative

Accuracy = $\frac{\text{true predictions } (TP + TN)}{\text{All Predictions } (TP + TN + FP + FN)}$

Recall = $\frac{\text{Actual Spam } (TP)}{\text{All Spam } (TP + FN)}$

Precision = $\frac{\text{Actual Spam } (TP)}{\text{Predicted Spam } (TP + FP)}$

		Predicted	
		a	b
actual	a	236	38
	b	23	303

$$\text{Accuracy} = \frac{236 + 303}{236 + 38 + 23 + 303}$$

$$= 0.8983 \qquad \boxed{= 89.833\%}$$

$$\text{Precision} = \frac{236}{236 + 23} \qquad \boxed{= 0.911}$$

$$\text{Recall} = \frac{236}{236 + 38} = \boxed{0.861}$$

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **J48** -C 0.25 -M 2

Test options

- Use training set
- Supplied test set — Set...
- Cross-validation — Folds 10
- Percentage split — % 66

More options...

(Nom) pep

Start | Stop

Result list (right-click for options)

21:09:17 - trees.J48

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     bankinfo
Instances:    600
Attributes:   12
              id
              age
              sex
              region
              income
              married
              children
              car
              save_act
              current_act
              mortgage
              pep
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

children <= 1
|   children <= 0
|   |   married = NO
|   |   |   mortgage = NO: YES (48.0/3.0)
|   |   |   mortgage = YES
|   |   |   |   save_act = NO: YES (12.0)
|   |   |   |   save_act = YES: NO (23.0)
|   |   married = YES
|   |   |   save_act = NO
|   |   |   |   mortgage = NO
|   |   |   |   |   income <= 21506.2
|   |   |   |   |   |   age <= 41: NO (11.0/1.0)
|   |   |   |   |   |   age > 41: YES (5.0/1.0)
|   |   |   |   |   income > 21506.2: NO (20.0)
|   |   |   |   mortgage = YES: YES (25.0/3.0)
|   |   |   save_act = YES: NO (119.0/12.0)
|   children > 0
|   |   income <= 15538.8
|   |   |   age <= 41: NO (22.0/2.0)
|   |   |   age > 41: YES (2.0)
|   |   income > 15538.8: YES (111.0/5.0)
children > 1
|   income <= 30404.3: NO (124.0/12.0)
|   income > 30404.3
|   |   children <= 2: YES (51.0/5.0)
|   |   children > 2
|   |   |   income <= 44288.3: NO (19.0/2.0)
|   |   |   income > 44288.3: YES (8.0)

Number of Leaves  :     15

Size of the tree :      29


Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===
```

Status

OK

Log | x 0

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **J48** -C 0.25 -M 2

Test options

- Use training set
- Supplied test set — Set...
- Cross-validation — Folds 10
- Percentage split — % 66

More options...

(Nom) pep

Start | Stop

Result list (right-click for options)

21:09:17 - trees.J48

Classifier output

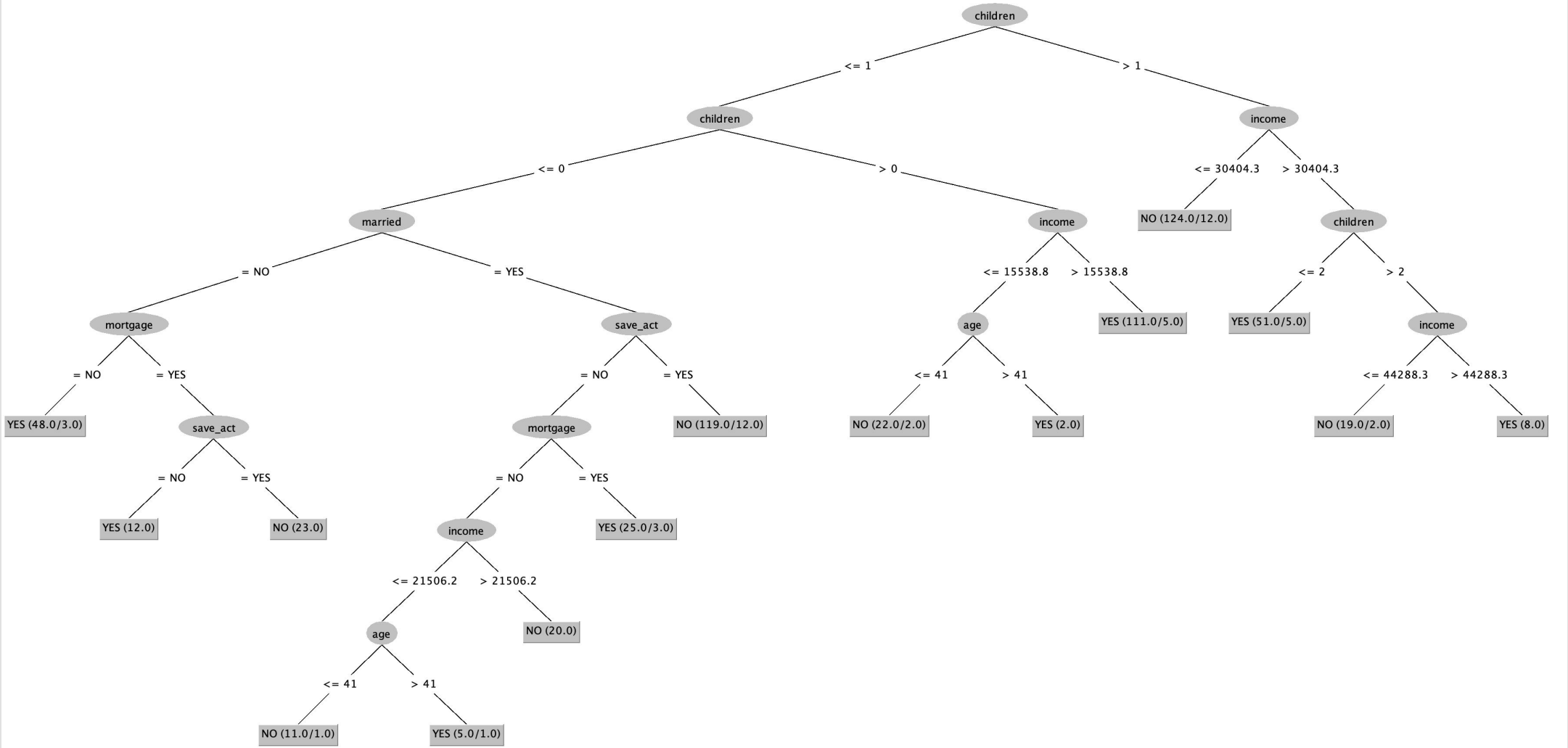
```
------------------

children <= 1
|   children <= 0
|   |   married = NO
|   |   |   mortgage = NO: YES (48.0/3.0)
|   |   |   mortgage = YES
|   |   |   |   save_act = NO: YES (12.0)
|   |   |   |   save_act = YES: NO (23.0)
|   |   married = YES
|   |   |   save_act = NO
|   |   |   |   mortgage = NO
|   |   |   |   |   income <= 21506.2
|   |   |   |   |   |   age <= 41: NO (11.0/1.0)
|   |   |   |   |   |   age > 41: YES (5.0/1.0)
|   |   |   |   |   income > 21506.2: NO (20.0)
|   |   |   |   mortgage = YES: YES (25.0/3.0)
|   |   |   save_act = YES: NO (119.0/12.0)
|   children > 0
|   |   income <= 15538.8
|   |   |   age <= 41: NO (22.0/2.0)
|   |   |   age > 41: YES (2.0)
|   |   income > 15538.8: YES (111.0/5.0)
children > 1
|   income <= 30404.3: NO (124.0/12.0)
|   income > 30404.3
|   |   children <= 2: YES (51.0/5.0)
|   |   children > 2
|   |   |   income <= 44288.3: NO (19.0/2.0)
|   |   |   income > 44288.3: YES (8.0)

Number of Leaves  :     15

Size of the tree :      29


Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         539               89.8333 %
Incorrectly Classified Instances        61               10.1667 %
Kappa statistic                          0.7942
Mean absolute error                      0.167
Root mean squared error                  0.305
Relative absolute error                 33.6511 %
Root relative squared error             61.2344 %
Total Number of Instances              600

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.861    0.071    0.911      0.861    0.886      0.795    0.883     0.847     YES
                 0.929    0.139    0.889      0.929    0.909      0.795    0.883     0.863     NO
Weighted Avg.    0.898    0.108    0.899      0.898    0.898      0.795    0.883     0.856

=== Confusion Matrix ===

   a   b   <-- classified as
 236  38 |   a = YES
  23 303 |   b = NO
```

Status

OK

Log | x 0

Tree View


children
<= 1
> 1
children
income
<= 0
> 0
<= 30404.3
> 30404.3
married
income
NO (124.0/12.0)
children
= NO
= YES
<= 15538.8
> 15538.8
<= 2
> 2
mortgage
save_act
age
YES (111.0/5.0)
YES (51.0/5.0)
income
= NO
= YES
= NO
= YES
<= 41
> 41
<= 44288.3
> 44288.3
YES (48.0/3.0)
save_act
mortgage
NO (119.0/12.0)
NO (22.0/2.0)
YES (2.0)
NO (19.0/2.0)
YES (8.0)
= NO
= YES
= NO
= YES
YES (12.0)
NO (23.0)
income
YES (25.0/3.0)
<= 21506.2
> 21506.2
age
NO (20.0)
<= 41
> 41
NO (11.0/1.0)
YES (5.0/1.0)