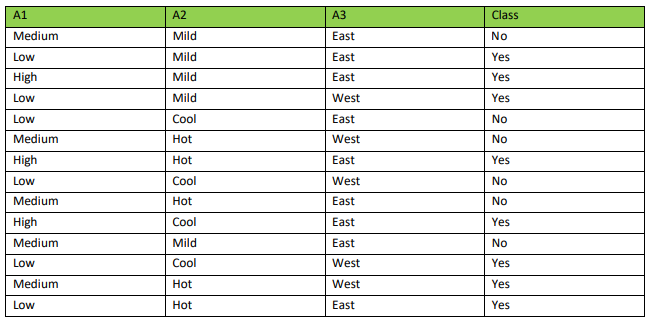
Data



**1.a**

For class we have 2 possible values (c1 = Yes, c2 = No).

Entropy(t) = - P(c1 | t) \* log(p(c1 | t)) - P(c2 | t) \* log(p(c2 | t))

= - 8/14 \* log(8/14) – 6/14 \* log(6/14)

= 0.29658

**1.b**

A3 column has 2 categories (N1 = East, N2 = West)

Entropy(N1) = - 5/9 \* log(5/9) – 4/9 \* log(4/9) = 0.991

Entropy(N2) = - 3/5 \* log(3/5) – 2/5 \* log(2/5) = 0.97

Entopy\_Avg(N) = 9/14 \* Entropy(N1) + 5/14 \* Entropy(N2)

= 9/14 \* 0.991 + 5/14 \* 0.97 = 0.9835

Gain(N, A3) = Entropy(N) – Entropy\_Avg(N) = 0.985 – 0.9835 = 0.0015

**Prob.2 solution**

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: breast-cancer

Instances: 286

Attributes: 10

age

menopause

tumor-size

inv-nodes

node-caps

deg-malig

breast

breast-quad

irradiat

Class

Test mode: 10-fold cross-validation

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

J48 pruned tree

------------------

node-caps = yes

| deg-malig = 1: recurrence-events (1.01/0.4)

| deg-malig = 2: no-recurrence-events (26.2/8.0)

| deg-malig = 3: recurrence-events (30.4/7.4)

node-caps = no: no-recurrence-events (228.39/53.4)

Number of Leaves : 4

Size of the tree : 6

Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 216 75.5245 %

Incorrectly Classified Instances 70 24.4755 %

Kappa statistic 0.2826

Mean absolute error 0.3676

Root mean squared error 0.4324

Relative absolute error 87.8635 %

Root relative squared error 94.6093 %

Total Number of Instances 286

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.960 0.729 0.757 0.960 0.846 0.339 0.584 0.736 no-recurrence-events

0.271 0.040 0.742 0.271 0.397 0.339 0.584 0.436 recurrence-events

Weighted Avg. 0.755 0.524 0.752 0.755 0.713 0.339 0.584 0.647

=== Confusion Matrix ===

a b <-- classified as

193 8 | a = no-recurrence-events

62 23 | b = recurrence-events

**Findings: in naïve bays the correctly classified instances percentage was 71%, while in thing classification algorithm the percentage is 75%. Therefore I would use decision tree classification for this problem space.**