Problem 1:

P(C1|N): P(class=yes) = 8/14

P(C1|N): P(class=no) = 6/14

Entropy(N) = -6/14 x log(6/14) – 8/14 x log(8/14) = 0.15777+0.138 = 0.2965

Entropy(Medium) = -1/5 x log(1/5) – 4/5 x log(4/5) = 0.217

Entropy(Low) = -4/6 x log(4/6) – 2/6 x log(2/6) = 0.276

Entropy(High) = -3/3 x log(3/3) – 0/3 x log(0/3) = 0

EntropyA1(N) = 5/14 x 0.217 + 6/14 x 0.276 = 3/14 x 0 = 0.0775 + 0.118 + 0 = 0.195

Entropy(Mild) = -3/5 x log(3/5) – 2/5 x log(2/5) = 0.292

Entropy(Cool) = -2/4 x log(2/4) – 2/4 x log(2/4) = 0.301

Entropy((Hot) = -3/5 x log(3/5) – 2/5 x log(2/5) = 0.292

EntropyA2(N) = 5/14 x 0.292 + 4/14 x 0.301 + 5/14 x 0.292 = 0.104 + 0.086 + 0.104 = 0.294

Entropy(East) = -5/9 x log(5/9) – 4/9 x log(4/9) = 0.298

Entropy(West) = -3/5 x log(3/5) – 2/5 x log(2/5) = 0.292

EntropyA3(N) = 9/14 x 0.298 + 5/14 x 0.292 = 0.2958

Gain(N, A3) = Entropy(N) – EntropyA3(N) = 0.2965 – 0.2958 = 0.00064

Problem 2:

If I don't want to concern about the time spent to build the model, I would definitely choose to use the Decision Tree model, because it has approximately 4% more chances to correctly predict the income data. The Bayes Naive model is built faster, but the prediction is less accurate.

=== 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: 0.03 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