1.(a)

$$
\begin{aligned}
GainRatio(A) &= \frac{Gain(A)}{SplitInfo(A)} \\
&= \frac{Info(D) - Info_A(D)}{SplitInfo(A)} \\
&= \frac{-\sum_{i=1}^{m} p_i log_2(p_i) - \sum_{j=1}^{v} \frac{|D_j|}{D} \times Info(D_j)}{-\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2(\frac{|D_j|}{|D|})}
\end{aligned}
$$

, where the discrete random variable Y takes m distinct values $\{y_1, y_2, \dots, y_m\}$, H(Y) for Y distributed as $[p_i]$, D are split into v partitions using A, and $Info(D_j)$ could be calculated in the same way as $Info(D)$.

(b)

$$
\begin{aligned}
\delta gini(A) &= gini(D) - gini_A(D) \\
&= (1 - \sum_{j=1}^{n} p_j^2) - \sum_{i=1}^{k} \frac{|D_k|}{|D|} gini(D_k) \\
&= 1 - \sum_{j=1}^{n} p_j^2 - \sum_{i=1}^{k} \frac{|D_k|}{|D|} gini(D_k)
\end{aligned}
$$

,where the attribute A provides the smallest reduction in impurity, a dataset D contains examples from n classes, p_j provides the relative frequency of class j in D, and $gini(D_k)$ could be be calculated in the same way as $gini(D)$.

(c)

I will choose Patrons as the attribute to split the tree.

Reasons:

$$Info(D) = I(6,6) = -\tfrac{6}{12} log_2(\tfrac{6}{12}) - \tfrac{6}{12} log_2(\tfrac{6}{12}) = 1$$

$$
\begin{aligned}
Info_{Patrons} &= \frac{2}{12} I(2,0) + \frac{4}{12} I(0,4) + \frac{6}{12} I(4,2) \\
&= \frac{2}{12}(-log_2 1) + \frac{4}{12}(-log_2 1) + \frac{6}{12}(-\frac{2}{3} log_2 \frac{2}{3} - \frac{1}{3} log_2 \frac{1}{3}) \\
&\approx 0.4591479170272448
\end{aligned}
$$

$$
\begin{aligned}
SplitInfo(Patrons) &= -\frac{2}{12} log_2(\frac{2}{12}) - \frac{4}{12} log_2(\frac{4}{12}) - \frac{6}{12} log_2(\frac{6}{12}) \\
&\approx 1.4591479170272448
\end{aligned}
$$

$$
\begin{aligned}
GainRatio(Patrons) &= \frac{Gain(Patrons)}{SplitInfo(Patrons)} \\
&= \frac{Info(D) - Info_{Patrons}(D)}{SplitInfo(Patrons)} \\
&\approx \frac{1 - 0.4591479170272448}{1.4591479170272448} \\
&\approx 0.371
\end{aligned}
$$

$$
\begin{aligned}
Info_{Type} &= \frac{2}{12}I(1,1) + \frac{2}{12}I(1,1) + \frac{4}{12}I(2,2) + \frac{4}{12}I(2,2) \\
&= \frac{2}{12}(-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2}) \times 2 + \frac{4}{12}(-\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4}) \times 2 \\
&= 1
\end{aligned}
$$

$$
\begin{aligned}
SplitInfo(Types) &= -\frac{2}{12}log_2(\frac{2}{12}) - \frac{2}{12}log_2(\frac{2}{12}) - \frac{4}{12}log_2(\frac{4}{12}) - \frac{4}{12}log_2(\frac{4}{12}) \\
&\approx 1.9182958340544896
\end{aligned}
$$

$$
\begin{aligned}
GainRatio(Type) &= \frac{Gain(Type)}{SplitInfo(Type)} \\
&= \frac{Info(D) - Info_{Types}(D)}{SplitInfo(Types)} \\
&= \frac{1 - 1}{1.9182958340544896} \\
&= 0
\end{aligned}
$$

$GainRatio(Patrons) \approx 0.371 > 0 = GainRatio(Types)$, therefore, based on the Gain Ratio, I will choose Patrons as the attribute to split the tree.

(d)

I will choose Patrons as the attribute to split the tree.

Reasons:

$gini(D) = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = 0.5$

$$
\begin{aligned}
gini_{Patrons}(D) &= \frac{2}{12}gini(2,0) + \frac{4}{12}gini(0,4) + \frac{6}{12}gini(4,2) \\
&= \frac{2}{12}(1 - 1^2) + \frac{4}{12}(1 - 1^2) + \frac{6}{12}(1 - (\frac{4}{6})^2 - (\frac{2}{6})^2) \\
&\approx 0.2222222222222224
\end{aligned}
$$

$$
\begin{aligned}
SplitInfo(Patrons) &= -\frac{2}{12}log_2(\frac{2}{12}) - \frac{4}{12}log_2(\frac{4}{12}) - \frac{6}{12}log_2(\frac{6}{12}) \\
&\approx 1.4591479170272448
\end{aligned}
$$

$$
\begin{aligned}
GiniReductionRatio(Patrons) &= \frac{GiniReductionInImpurity(Patrons)}{SplitInfo(Patrons)} \\
&= \frac{gini(D) - gini_{Patrons}(D)}{SplitInfo(Patrons)} \\
&\approx \frac{0.5 - 0.2222222222222224}{1.4591479170272448} \\
&\approx 0.190
\end{aligned}
$$

$$
\begin{aligned}
gini_{Types}(D) &= \frac{2}{12}gini(1,1) + \frac{2}{12}gini(1,1) + \frac{4}{12}gini(2,2) + \frac{4}{12}gini(2,2) \\
&= \frac{2}{12}(1 - \frac{1}{2}^2 - \frac{1}{2}^2) \times 2 + \frac{4}{12}(1 - (\frac{2}{4})^2 - (\frac{2}{4})^2) \times 2 \\
&= 0.5
\end{aligned}
$$

$$
\begin{aligned}
SplitInfo(Types) &= -\frac{2}{12}log_2(\frac{2}{12}) - \frac{2}{12}log_2(\frac{2}{12}) - \frac{4}{12}log_2(\frac{4}{12}) - \frac{4}{12}log_2(\frac{4}{12}) \\
&\approx 1.9182958340544896
\end{aligned}
$$

$$
\begin{aligned}
GiniReductionRatio(Types) &= \frac{GiniReductionInImpurity(Types)}{SplitInfo(Types)} \\
&= \frac{gini(D) - gini_{Types}(D)}{SplitInfo(Types)} \\
&\approx \frac{0.5 - 0.5}{1.9182958340544896} \\
&= 0
\end{aligned}
$$

$GiniReductionRatio(Patrons) \approx 0.190 > 0 = GiniReductionRatio(Types)$, therefore, based on the Gini Reduction Ratio, I will choose Patrons as the attribute to split the tree.

2.(a)

According to Bayes Rule, $p(H_i|X) = \frac{p(X|H_i)p(H_i)}{p(X)}$, where X is a data sample or evidence, H_i belongs to hypothesis set $\{H_1, H_2, \ldots, H_m\}$

$$
\begin{aligned}
p(Candy_1 = lime) =& p(h_1)p(Candy_1 = lime|h_1) + p(h_2)p(Candy_1 = lime|h_2) \\
&+ p(h_3)p(Candy_1 = lime|h_3) \\
&= \frac{1}{4} \times 0 + \frac{1}{2} \times \frac{1}{2} + \frac{1}{4} \times 1 = 0.5
\end{aligned}
$$

$p(h_1|Candy_1 = lime) = \frac{p(Candy_1 = lime|h_1)p(h_1)}{p(Candy_1 = lime)} = \frac{\frac{1}{4} \times 0}{1} = 0$

$p(h_2|Candy_1 = lime) = \frac{p(Candy_1 = lime|h_2)p(h_2)}{p(Candy_1 = lime)} = \frac{\frac{1}{2} \times 0.5}{0.5} = 0.5$

$p(h_3|Candy_1 = lime) = \frac{p(Candy_1=lime|h_3)p(h_3)}{p(Candy_1=lime)} = \frac{\frac{1}{4}\times 1}{0.5} = 0.5$

(b)

According to Bayes Rule, $p(H_i|X_1, X_2) = \frac{p(H_i|X_1)p(X_2|H_i)}{p(X_2|X_1)}$, where X_1, X_2 are conditionally independent given H_i, H_i belongs to hypothesis set $\{H_1, H_2, \dots, H_m\}$

Assume constant $\frac{1}{k} = p(Candy_2 = lime|Candy_1 = lime)$

$$
\begin{aligned}
p(h_1|Candy_1 = lime, Candy_2 = lime) &= \frac{p(h_1|Candy_1 = lime)p(Candy_2 = lime|h_1)}{p(Candy_2 = lime|Candy_1 = lime)} \\
&= \frac{0 \times p(Candy_2 = lime|h_1)}{\frac{1}{k}} \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
p(h_2|Candy_1 = lime, Candy_2 = lime) &= \frac{p(h_2|Candy_1 = lime)p(Candy_2 = lime|h_2)}{p(Candy_2 = lime|Candy_1 = lime)} \\
&= \frac{0.5 \times \frac{49}{99}}{\frac{1}{k}} \\
&= \frac{49}{198}k
\end{aligned}
$$

$$
\begin{aligned}
p(h_3|Candy_1 = lime, Candy_2 = lime) &= \frac{p(h_3|Candy_1 = lime)p(Candy_2 = lime|h_3)}{p(Candy_2 = lime|Candy_1 = lime)} \\
&= \frac{0.5 \times 1}{\frac{1}{k}} \\
&= 0.5k
\end{aligned}
$$

$$
\begin{aligned}
&p(h_1|Candy_1 = lime, Candy_2 = lime) + p(h_2|Candy_1 = lime, Candy_2 = lime) \\
&\qquad + p(h_3|Candy_1 = lime, Candy_2 = lime) = \frac{49}{198}k + 0.5k = 1
\end{aligned}
$$

Therefore, $k = \frac{99}{74}$

$p(h_1|Candy_1 = lime, Candy_2 = lime)$=0

$p(h_2|Candy_1 = lime, Candy_2 = lime) \approx 0.332$

$p(h_3|Candy_1 = lime, Candy_2 = lime) \approx 0.668$

3.(a)

x =$\{x_1(Size), x_2(Color), x_3(Shape)\}$

$y = Yes$ or No

We need 7 independent parameters:

$$\begin{gathered} P(y = Yes), \\ P(x_1 = Large|y = Yes), P(x_2 = Red|y = Yes), P(x_3 = Circle|y = Yes), \\ P(x_1 = Large|y = No), P(x_2 = Red|y = No), P(x_3 = Circle|y = No) \end{gathered}$$

(b) Bayes requires each conditional probability to be non-zero, or the predicted probability will be zero. Therefore, we can use Laplacian correction to calculate the probability.

$P(y = Yes) = \frac{5}{10} = 0.5$

$P(x_1 = Large|y = Yes) = \frac{count(Large,Yes)+1}{count(Yes)+2} = \frac{3+1}{5+2} \approx 0.5714$

$P(x_2 = Red|y = Yes) = \frac{count(Red,Yes)+1}{count(Yes)+2} = \frac{5+1}{5+2} \approx 0.8571$

$P(x_3 = Circle|y = Yes) = \frac{count(Circle,Yes)+1}{count(Yes)+2} = \frac{5+1}{5+2} \approx 0.5714$

$P(x_1 = Large|y = No) = \frac{count(Large,No)+1}{count(No)+2} = \frac{3+1}{5+2} \approx 0.5714$

$P(x_2 = Red|y = No) = \frac{count(Red,No)+1}{count(No)+2} = \frac{0+1}{5+2} \approx 0.1429$

$P(x_3 = Circle|y = No) = \frac{count(Circle,No)+1}{count(No)+2} = \frac{2+1}{5+2} \approx 0.4286$

(c)

Assume

$$\begin{aligned} P(y = Yes|x) = k \times P(x_1 = Large|y = Yes) \times P(x_2 = Green|y = Yes) \\ \times P(x_3 = Circle|y = Yes) \times P(y = Yes) \\ = \frac{4}{7} \times (1 - \frac{6}{7}) \times \frac{4}{7} \times \frac{1}{2}k = \frac{8}{343}k \end{aligned}$$

and

$$\begin{aligned} P(y = No|x) = k \times P(x_1 = Large|y = No) \times P(x_2 = Green|y = No) \\ \times P(x_3 = Circle|y = No) \times P(y = No) \\ = \frac{4}{7} \times (1 - \frac{1}{7}) \times \frac{3}{7} \times \frac{1}{2}k = \frac{36}{343}k \end{aligned}$$

,where k is a constant between 0 and 1.

$P(y = Yes|x) + P(y = No|x) = \frac{8}{343}k + \frac{36}{343}k = \frac{44}{343}k = 1$

Therefore, $k = \frac{343}{44}$

$P(y = Yes|x) = \frac{2}{11} \approx 0.1818$

$P(y = No|x) = \frac{9}{11} \approx 0.8182$

$P(y = Yes|x) < P(y = No|x)$. Therefore, the naive Bayes will classify $x = (Large; Green; Circle)$ as $\{GoodApple = No\}$

4.(a)

the tree depth d:

Tree depth is numerically the number of nodes from the root to the farthest leaf. Each Random Forests algorithm have maximum tree depth. The lower d is, the faster computing costs.

the number of attributes randomly selected as candidates for splits m:

When we are constructing training tree, we select m number of attributes randomly from all attributes to ensure the robustness. Then we choose the best attributes from these m attributes and use it as a rule to split all its child tree or child nodes into the left tree or the right tree.

the total number of trees T:

Random Forests construct T independent trees in the same way, which votes for the final result equally.

(b)

5.(a)Sensitivity. True positive recognition rate.

$Sensitivity = \frac{TP}{P} = \frac{a}{a+b}$

$Sensitivity_{M_1} = \frac{2588}{2588+412} \approx 0.8627$

$Sensitivity_{M_2} = \frac{2999}{2999+1} \approx 0.9997$

(b)Specificity. True negative recognition rate.

$$Specificity = \frac{TN}{N} = \frac{d}{c+d}$$

$$Specificity_{M_1} = \frac{6954}{6954+46} \approx 0.9934$$

$$Specificity_{M_2} = \frac{6999}{6999+1} \approx 0.9999$$

(c)Accuracy. Percentage of test set tuples that are correctly classified.

$$Accuracy = \frac{TN+TP}{ALL} = \frac{a+d}{a+b+c+d}$$

$$Accuracy_{M_1} = \frac{2588+6954}{10000} = 0.9542$$

$$Accuracy_{M_2} = \frac{2999+6999}{10000} = 0.9998$$

(d)Precision. What percentage of tuples that classifier labeled as positive are actually positive.

$$Precision = \frac{TP}{TP+FP} = \frac{a}{a+c}$$

$$Precision_{M_1} = \frac{2588}{2588+46} \approx 0.9825$$

$$Precision_{M_2} = \frac{2999}{2999+1} \approx 0.9997$$

(e)Recall. What percentage of positive tuples did the classifier label as positive.

$$Recall = \frac{TP}{TP+FN} = \frac{a}{a+b}$$

$$Recall_{M_1} = \frac{2588}{2588+412} \approx 0.8627$$

$$Recall_{M_2} = \frac{2999}{2999+1} \approx 0.9997$$

(f)F1 score. Assuming β=1 harmonic mean of precision and recall.

$$F1score = \frac{2P \times R}{P+R} = \frac{2a}{2a+b+c}$$

$$F1score_{M_1} = \frac{2 \times 2588}{2 \times 2588+412+46} \approx 0.9187$$

$$F1score_{M_2} = \frac{2 \times 2999}{2 \times 2999+1+1} \approx 0.9997$$

4.(b)

Train: RFs Forest-RI bagging with decision trees as base models. It uses a subset of training data by sampling with replacement for each tree.(putting-back pick) At each node, it uses a random subset of attributes as candidates and split by the best attribute among them, which ensures its high diversity. During classification, each tree votes and the most popular class is returned. The final output is judged by each tree equally. The cart methodology is used to grow the trees to maximum size.

Test: RFs Forest-RI uses a subset of training data from the original data, which is chosen randomly. Therefore, there should be some data out of the bag which could be used to test the accuracy of RFs Forest-RI. The more the number of the event that the output of RFs Forest-RI is the same as the given of data out of bag, the better the RFs Forest-RI is. The testing way is called OOB error (out of bag error).

(c)

By sampling with replacement, some observation may be repeated in each D_i, if size n'=n, then for large n the set D_i is expected to have the fraction (1-1/e) of the unique examples of D, the rest being duplicates

n*(1-(n-1/n)^n)=n-n/e

(d)

I agree with the Professor.

Reasons:

When m=1, there will be only one random feature for building a tree, which improves the diversity and decreases the correlation between trees. Therefore, the accuracy is increased.

What's more, the variance of the trees increases so that the random forests can better represents the whole dataset. Therefore, the variance of the random forests' predictions are lower.

2.(b)

$$
\begin{aligned}
p(h_1|Candy_1 = lime, Candy_2 = lime) &= \frac{p(Candy_2 = lime, Candy_1 = lime|h_1)p(h_1)}{p(Candy_2 = lime, Candy_1 = lime)} \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
p(h_2|Candy_1 = lime, Candy_2 = lime) &= \frac{p(Candy_2 = lime, Candy_1 = lime|h_2)p(h_2)}{p(Candy_2 = lime, Candy_1 = lime)} \\
&= \frac{0.5 \times \frac{49}{99} \times 0.5}{\frac{37}{99}} \\
&\approx 0.33108
\end{aligned}
$$

$$
\begin{aligned}
p(h_3|Candy_1 = lime, Candy_2 = lime) &= \frac{p(Candy_2 = lime, Candy_1 = lime|h_3)p(h_3)}{p(Candy_2 = lime, Candy_1 = lime)} \\
&= \frac{1 \times 1 \times 0.25}{\frac{37}{99}} \\
&\approx 0.66892
\end{aligned}
$$