Answer 1a

For a fixed x, we can predict c(x) = D or c(x) = $\overline{D}$, for each of these case we can incur 2 types of losses:

L(D,D) = 0

L($\overline{D}$,$\overline{D}$) = 0

L(D, $\overline{D}$) = 1

L($\overline{D}$, D) = 1

Answer 1b

The best classification strategy for this case would be the Bayes classifier because we know the conditional probabilities to effectively implement the Bayes classifier.

Bayes rule :

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

So, using the given formula we will calculate probabilities for both the classes D and $\overline{D}$ and whichever will be the greatest we will assign the variable to that class.

Answer 2a

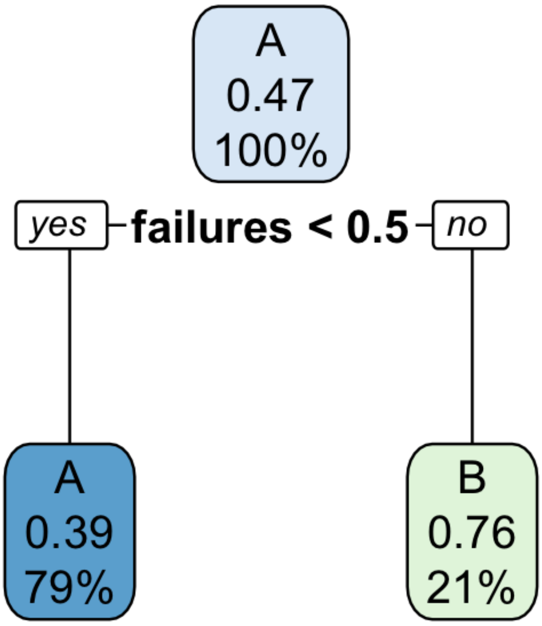
We have trained model using the training data.

Answer 2b

Confusion Matrix for training data is below:

```
  class_predict_2
      1    2    3
1    29   23  203
2    18   21  214
3   163  177 1652
```

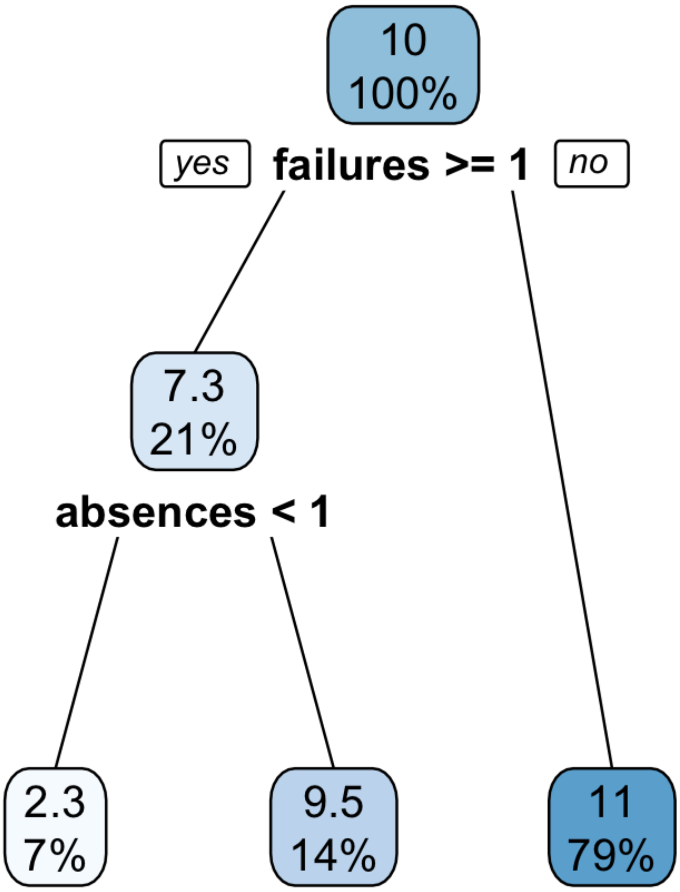
Answer 3a
We have created class A for G3>10 and class B for G3 <= 10.
Below is the decision tree created after pruning it from overfitting.



Answer 3b
The generalization error for this decision tree is 0.05158.
The error rate on training data can be found in the 4th column of the table generated by printcp function in R, which is equal to 0.78495.

Answer 3c
Since the generalization error rate after 1st cut is almost consistent we can say variable failure is the most important feature as we are making division on basis of failures > 0.5.

Answer 3d
Below is the decision tree generated when we used method = “anova”.



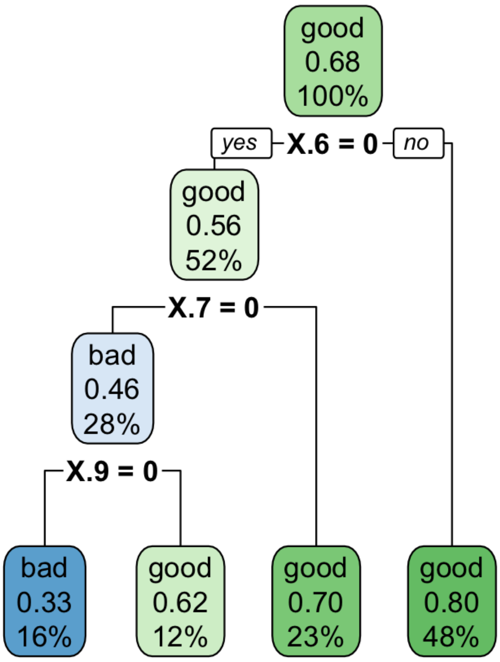
For regression problems, we calculate SSE, and from the predictions of the above regression tree our sum of squared error = 4923.775.

<u>Answer 4</u>

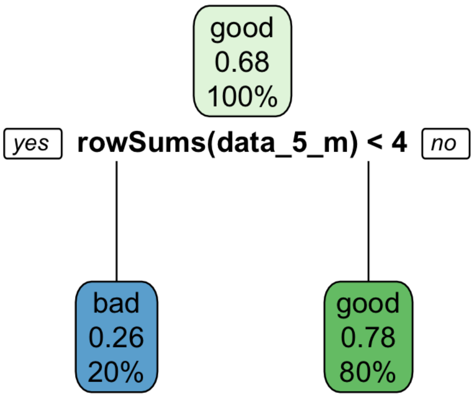
For the first case, we can see that we can separate circles from triangles using a straight line x = y. So, for the new feature if x > y for a particular point it will belong to the class triangle, and vice-versa it will belong to the class circle.

For the second case, we can see circles form a circle on the plane and triangles are outside that circle, from the graph we can assume that circle has a radius = 1and center at (0,0). So equation for this circle becomes, x^2 + y^2 = 1. If any point is inside the circle we will predict it as a class circle and for any point outside the circle, we will predict it as a class triangle. The new feature will be x^2 + y^2 <= 1, then the point belongs to the class circle, else, it belongs to the class triangle.

Answer 5



For this tree, accuracy is 72.5%.



The new feature is the sum of all the features and for this decision tree, the accuracy is 77.5%.

Answer 6a

Ans 6 a

Suppose we divide Node N with class probabilities P into N_l & N_r, with class probabilities P_l & P_r & proportions q_l & q_r

Average ~~that~~ entropy of this division will be

$$q_l H(P_l) + q_r H(P_r)$$

$$\Rightarrow q_l \log(1/P_l) + q_r \log(1/P_r) = q_l(-\log(P_l)) + q_r(-\log(P_r))$$

if we use equation $pK(x_1) + qK(x_2) \geq K(px_1 + qx_2)$ and put $K = -\log$, $p = q_l$, $x_l = P_l$, $q = q_r$ & $x_2 = P_r$

Our equation becomes:

$$-q_l \log P_l - q_r \log P_r \geq -\log(q_l P_l + q_r P_r)$$

$$\Rightarrow q_l \log P_l + q_r \log P_r \leq \log(q_l P_l + q_r P_r)$$

$$q_l P_l + q_r P_r = E(P)$$

$$\Rightarrow q_l \log P_l + q_r \log P_r \leq \log(E(P)) \quad - (1)$$

Now consider entropy at node N

$$H(P) = E(\log(1/P))$$

it is given that

$$E(\log(1/P)) \leq \log(E(1/P))$$

(1)

$$-E(\log(P)) \leq -\log(E(P))$$

which becomes

$$E(\log(P)) \geq \log(E(P)) \quad - (2)$$

from equation (1) & (2) we can clearly say that

$$q_l H(P_l) + q_r H(P_r) \leq H(P)$$

Hence, average entropy split is no greater than original entropy at node.

Answer 6b

Ans. 6b) $$H(C|T) = \sum_c \sum_t P(C_c, T_t) \log \frac{P(C_c, T_t)}{P(T_t)} \quad \text{— (1)}$$

$$\Rightarrow \sum_c \sum_t P(C_c, T_t) \log P(C_c | T_t)$$

Marginalise all over classes $\sum P(C_c) = 1$

$$\Rightarrow \sum_t P(T_t) \underbrace{P(C_c | T_t) \log (P(C_c | T_t))}$$

$$\Rightarrow \sum_t P(T_t) H(C_o | T_t)$$

$$\Rightarrow \sum_t P(T = t) H(C | T = t)$$

This is RHS of given equation to prone.

We started with LHS and prone that to be equal to RHS.

Given Equation:

$$H(C|T) = \sum_t P(T = t) H(C | T = t)$$

Answer 6c

Ans. 6c) To prove: $H(T,C) = H(C|T) + H(T)$

Bayes rule: $P(x,y) = P(y|x) + P(x)$

so let us write $H(T,C) = H(C|T) + H(T)$ in terms of probabilities and try to use bayes rule on it.

$$H(x,y) = \sum_{x} \sum_{y} P(x,y) \log \frac{1}{P(x,y)}$$ (formula for entropy)

$$= \sum_{x} \sum_{y} P(y|x) \cdot P(x) \cdot \log \frac{1}{P(y|x) \cdot P(x)}$$

$$\Rightarrow \sum_{x} \sum_{y} P(y|x) \cdot P(x) \left[\log\left(\frac{1}{P(y|x)}\right) + \log\left(\frac{1}{P(x)}\right) \right]$$

$$\Rightarrow \sum_{x} \sum_{y} P(y|x) \cdot P(x) \cdot \log\left(\frac{1}{P(y|x)}\right) + \sum_{x} \sum_{y} P(y|x) \cdot P(x) \cdot \log(1/P(x))$$

$$\Rightarrow \sum_{x} P(x) \cdot \sum_{y} P(y|x) \cdot \log\left(\frac{1}{P(y|x)}\right) + \sum_{x} P(x) \cdot \log\left(\frac{1}{P(x)}\right) \cdot \sum_{y} P(y|x)$$

$$\Rightarrow \sum_{x} P(x)\, H(Y|X=x) + \sum_{x} P(x) \cdot \log\left(\frac{1}{P(x)}\right) \cdot 1 \quad \text{— (1)}$$

as $\sum_{y} P(Y|x) = 1$.

so equation (1) could be written as

$$H(X,Y) = H(Y|X) + H(X)$$

Now replace, Y with C and X with T.

we get ~~H(C|T~~

$$H(T,C) = H(C|T) + H(T)$$

Hence, proved.