Classification & Prediction

Q. Consider training data set below, for binary classification problem.

(a) What are information Gains of a1 & a2 for given data?

(b) Which attribute should be chosen as first splitting attribute (between a1 & a2), according to classification error ~~rate~~?

(c) Which attribute should be chosen as first splitting attribute (between a1 & a2), according to Gini index?

Instance	a1	a2	a3	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	F	F	7.0	+
9	T	F	5.0	−

→ (a) Using <u>information gain</u> as impurity measure. ⟹

contigency table for a1

Class	a1 = T	a1 = F
+	3	1
−	1	4

contigency table for a2

Class	a2 = T	a2 = F
+	2	2
−	3	2

Overall information needed to classify a tuple D is.

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j).$$

Where $Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$

$$Gain(A) = Info(D) - Info_A(D).$$

Overall information needed to classify a tuple.

$$Info(D) = -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} =$$

$$= -0.444(-1.169) - 0.555(-0.847) = 0.989$$

Expected Information needed to classify a tuple D acc. to "a1"

$$Info_{a1}(D) = \frac{4}{9}\left[-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right] + \frac{5}{9}\left[-\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}\right]$$

$$= \frac{4}{9}\left[-0.75\times(-0.415) - 0.25\times(-2)\right] + \frac{5}{9}\left[-0.2\times(-2.321) - 0.8\times(-0.321)\right]$$

$$= 0.444[0.811] + 0.555[0.721]$$

$$= 0.760$$

Information Gain

$Gain(a_1) = Info(D) - Info_{a_1}(D) = 0.989 - 0.760 = 0.229$

Expected information needed to classify tuple D acc. to a_2

$$Info_{a_2}(D) = \frac{4}{9}\left[-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right] + \frac{5}{9}\left[-\frac{2}{4}\log_2\frac{2}{2} - \frac{2}{4}\log_2\frac{2}{4}\right]$$

$$= 0.444\left[-0.4\times(-1.321) - 0.6\times(-0.736)\right]$$

$$+ 0.555\left[-0.5\times(-1) - 0.5\times(-1)\right]$$

$$= 0.444\left[0.97\right] + 0.555\left[1\right] = 0.985$$

Information Gain

$Gain(a_2) = Info(D) - Info_{a_2}(D) = 0.989 - 0.985 = 0.004$

As gain is more for a_1, it will be selected as first splitting attribute.

Using Gain Ratio as measure ⇒

C.45 uses gain ratio.

$$Split\ Info_A(D) = -\sum_{j=1}^{v}\frac{|D_j|}{|D|}\times\log_2\left(\frac{|D_j|}{|D|}\right)$$

$$\cancel{Split\ Info_{a_1}(D) = -\left(\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9}\right)}$$

$$\cancel{= -0.444(-1.169) - 0.555(-0.847) = 0.989}$$

$$\cancel{Split\ Info_{a_2}(D)} = Gain\ Ratio(A) = \frac{Gain(A)}{Split\ Info_A(D)}$$

$$Split\ Info_{a_1}(D) = -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} = 0.989$$

$$Split\ Info_{a_2}(D) = -\frac{5}{9}\log_2\frac{5}{9} - \frac{4}{9}\log_2\frac{4}{9} =$$

$$= -0.555\times(-0.847) - 0.444\times(-1.169) = 0.989$$

$$Gain\ Ratio(a_1) = \frac{Gain(a_1)}{Split\ Info_{a_1}(D)} = \frac{0.229}{0.989} = 0.231$$

$$Gain\ Ratio(a_2) = \frac{Gain(a_2)}{Split\ Info_{a_2}(D)} = \frac{0.004}{0.989} = 0.004$$

The attribute with maximum gain ratio is selected as splitting attribute. Hence "a_1" will be selected as first splitting attribute.

(a) Gini Index ⟹

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

Overall Gini before splitting.

$$G(D) = 1 - \left(\frac{4}{9}\right)^2 - \left(\frac{5}{9}\right)^2 = 1 - 0.197 - 0.308 = 0.495$$

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

$$\Delta Gini(A) = Gini(D) - Gini_A(D).$$

for attribute "a_1"

$$Gini_{a_1}(D) = \frac{4}{9}\left[Gini(D_1)\right] + \frac{5}{9}\left[Gini(D_2)\right]$$

$$= \frac{4}{9}\left[1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2\right] + \frac{5}{9}\left[1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2\right]$$

$$= 0.444\,[0.375] + 0.555\,[0.32] = 0.3441$$

$$\Delta Gini(a_1) = 0.495 - 0.3441 = 0.151.$$

for attribute "a_2"

$$Gini_{a_2}(D) = \frac{5}{9} Gini(D_1) + \frac{4}{9} Gini(D_2)$$

$$= \frac{5}{9}\left[1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right] + \frac{4}{9}\left[1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right]$$

$$= 0.555\,(0.48) + 0.444\,(0.5) = 0.4884$$

The attribute that maximizes the reduction in impurity (or equivalently has the minimum Gini index) is selected as splitting attribute.

$$\Delta Gini(a_2) = 0.495 - 0.488 = 0.007.$$

Hence, a_1 will be selected as splitting attribute.

(b) Classification Error ⟹

$$Error(D) = 1 - \max P(i|t) \qquad i = \text{class}.$$

Error before splitting on a_1

$$Error_{orig} = 1 - \max\left(\frac{4}{9}, \frac{5}{9}\right) = 1 - \max(0.444, 0.555)$$

$$= 1 - 0.555 = 0.444$$

Error after splitting on a_1.

$$Error_{a_1} = 1 - \frac{4}{9}\left[1 - \max\left(\frac{3}{4}, \frac{1}{4}\right)\right] - \frac{5}{9}\left[1 - \max\left(\frac{1}{5}, \frac{4}{5}\right)\right]$$

$$= 1 - \frac{4}{9}\left[1 - \max(0.75, 0.25)\right] - \frac{5}{9}\left[1 - \max(0.2, 0.8)\right]$$

$$= 1 - 0.444\,(1 - 0.75) - 0.555\,(1 - 0.8)$$

$$= 1 - 0.111 - 0.111$$

$$= 0.778.$$

~~Δ Error = Error(D) − Error$_A$(D)~~
~~= 0.444~~

Error ~~after~~ before splitting on "q_2"

Error $_{org}$ = $1 - \max\left(\frac{5}{9}, \frac{4}{9}\right) = 1 - \max(0.555, 0.444) = 1 - 0.555 = 0.445$

Error after splitting on "q_2"

$$\text{Error}_{q_2} = 1 - \frac{5}{9}\left[1 - \max\left(\frac{2}{5}, \frac{3}{5}\right)\right] - \frac{4}{9}\left[1 - \max\left(\frac{2}{4}, \frac{2}{4}\right)\right]$$

$$= 1 - 0.555\left[1 - \max(0.4, 0.6)\right] - 0.444\left[1 - \max(0.5, 0.5)\right]$$

$$= 1 - 0.555\left[1 - 0.6\right] - 0.444\left[1 - 0.5\right]$$

$$= 1 - 0.222 - 0.222 = 0.556.$$

The attribute that minimizes classification error, will be selected as splitting attribute, Hence, q_2 will be selected as splitting attribute.

Q. find the GINI index for contents of the table shown below.

Target Value →	South Indian	North Indian	Continental	Other Cuisine	Overall
Count	20	35	75	130	260
Percentage.	8	13	29	50	100

→ $$\text{Gini}(D) = 1 - \sum_{i=1}^{m} p_i^2$$

$$\text{Gini}(D) = 1 - \left(\frac{20}{260}\right)^2 - \left(\frac{35}{260}\right)^2 - \left(\frac{75}{260}\right)^2 - \left(\frac{130}{260}\right)^2$$

$$= 1 - 0.005 - 0.018 - 0.083 - 0.25$$

$$= 0.644.$$

Q. Consider training data set below for binary classification problem. For a_3, compute information gain for every possible split.

Instance	a_1	a_2	a_3	Target
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	−

→ Arrange a_3 is descending order of values.

a_3	Class Label	Split Point Mid Point	Entropy	Info gain
1.0	+	2.0	0.8474	0.142
3.0	−	3.5	0.9864	0.003
4.0	+	4.5	0.9143	0.075
5.0	−	5.0		
5.0	−	5.5	0.9823	0.007
6.0	+	6.5	0.9713	0.018
7.0	+	7.0		
7.0	−	7.5	0.888	0.101
8.0	−	8.5		

$$\text{Info}(D) = -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} = 0.989$$

$$\text{Info}_{a_3 \le 2.0}(D) = \frac{0}{9}\left[-\frac{0}{0}\log_2\frac{0}{0} - \frac{0}{0}\log_2\frac{0}{4}\right] + \frac{9}{9}\left[-\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9}\right]$$

$$= 0 + [-0.444 \times (-1.169) - 0.555 \times (-0.847)] = 0.989$$

$$\text{Info}_{a_3 \le 3.5}(D) = \frac{2}{9}\left[-\frac{1}{2}\log_2\frac{1}{2} - \frac{8}{9}\log_2\frac{8}{9}\right] + \frac{8}{9}\left[-\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8}\right]$$

$$= \cancel{\frac{1}{9}[0 - 0.884 \times (-0.169)] + 0.888[-0.5 \times (-1) - 0.5 \times (-1)]}$$

$$= \cancel{0.016 + 0.888 = 0.904}$$

$$\text{Info}_{a_3 \le 2.0}(D) = \frac{1}{9}\left[-\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log\frac{0}{1}\right] + \frac{8}{9}\left[-\frac{3}{8}\log\frac{3}{8} - \frac{5}{8}\log\frac{5}{8}\right]$$

$$= 0.111[0 - 0] + 0.888[-0.375 \times (-1.415) - 0.625 \times (-0.678)]$$

$$= 0 + 0.8474 = 0.8474$$

$$\Delta\text{gain} = \text{Info}(D) - \text{Info}_{a_3 \le 2.0}(D) = 0.989 - 0.8474 = 0.142$$

$$\text{Info}_{a_3 \le 3.5}(D) = \frac{2}{9}\left[-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right] + \frac{7}{9}\left[-\frac{2}{7}\log_2\frac{2}{7} - \frac{5}{7}\log_2\frac{5}{7}\right]$$

$$= 0.222\left[0.5\times(-1) - 0.5(-1)\right] + 0.777\left[-0.428\times(-1.222) - 0.571\times(-0.807)\right]$$

$$= 0.222 + 0.764 = 0.9864$$

$$\text{Again} = 0.989 - 0.986 = 0.003$$

$$\text{Info}_{a_3 \le 4.5}(D) = \frac{3}{9}\left[-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right] + \frac{6}{9}\left[-\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6}\right]$$

$$= 0.333\left[0.666\times(-0.58) - 0.333(-1.584)\right]$$

$$+ 0.666\left[-0.333\times(-1.584) - 0.666\times(-0.584)\right]$$

$$= 0.3042 + 0.610 = 0.9143$$

$$\text{Again} = 0.989 - 0.914 = 0.075$$

Similarly compute Info for $a_3 \le 5.5$, $a_3 \le 6.5$, $a_3 \le 7.5$.

$$\text{Info}_{a_3 \le 5.5}(D) = \frac{5}{9}\left[-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right] + \frac{4}{9}\left[-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right]$$

$$= 0.9823$$

$$\text{Again} = 0.989 - 0.982 = 0.007$$

$$\text{Info}_{a_3 \le 6.5}(D) = \frac{6}{9}\left[-\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}\right] + \frac{3}{9}\left[-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right] = 0.971$$

$$\text{Again} = 0.989 - 0.971 = 0.018$$

$$\text{Info}_{a_3 \le 7.5}(D) = \frac{8}{9}\left[-\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8}\right] + \frac{1}{9}\left[-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right] = 0.8$$

$$\text{Again} = 0.989 - 0.888 = 0.101$$

As info gain is highest when split is carried out at $a_3 = 2$ it will be selected as splitting condition.

Q. You are working in ABC corp. As a part of profile creation the firm captures various important details about employees. Using below sample dataset, use Baysian classification to classify following tuple either "High", "Middle" or "Low" income.

$X = \{$Sr.No. = 7, Age = "Above-40", Gender = "M", Occupation = "Soft Engg"

Sr.No.	Age.	Gender	Occupation	Income Bracket.
1	Above-40	F	Software Eng	High.
2	Between 31-40	M	Unemployed	Low
3	Below-30	M	Marketing Exe	Middle
4	Below-30	M	Data Scientist	High.
5	Betwy-31-40	F	Software Eng	High.
6	Below-30	F	Unemployed	Low.

→ Naive Bayseian classifier predicts that tuple X belongs to class C_i if & only if

$$P(C_i|X) > P(C_j|X) \text{ for } 1 \le j \le m,\ j \ne i$$

Thus needs to maximize $P(C_i|x)$, for $i = 1, 2, 3 \ldots$

$$P(C_i|X) = \frac{P(X|C_i)\,P(C_i)}{P(X)}$$

Prior probability of each class, i.e. $P(C_i)$.

$P(\text{Income} = \text{High}) = 3/6 = 0.5$

$P(\text{Income} = \text{Middle}) = 1/6 = 0.17$

$P(\text{Income} = \text{Low}) = 2/6 = 0.33$

Conditional probabilities $P(X|C_i)$ for $i = 1, 2, 3$.

$P(\text{Age} = \text{"Above-40"} \mid \text{Income} = \text{High}) = 1/3 = 0.33$

$P(\text{Age} = \text{"Above-40"} \mid \text{Income} = \text{Middle}) = 0$

$P(\text{Age} = \text{"Above-40"} \mid \text{Income} = \text{Low}) = 0$

$P(\text{Gender} = \text{"M"} \mid \text{Income} = \text{High}) = 1/3 = 0.33$

$P(\text{Gender} = \text{"M"} \mid \text{Income} = \text{Middle}) = 1/1 = 1$

$P(\text{Gender} = \text{"M"} \mid \text{Income} = \text{Low}) = 1/2 = 0.5.$

$P(\text{Occupation} = \text{"Soft Engg"} \mid \text{Income} = \text{High}) = 2/3 = 0.67$

$P(\text{Occupation} = \text{"Soft Engg"} \mid \text{Income} = \text{Middle}) = 0$

$P(\text{Occupation} = \text{"Soft Engg"} \mid \text{Income} = \text{Low}) = 0$

$P(X \mid \text{Income} = \text{High}) = 0.33 \times 0.17 \times 0.67 = 0.03$

$P(X \mid \text{Income} = \text{Middle}) = 0 \times 1 \times 0 = 0$

$P(X \mid \text{Income} = \text{Low}) = 0 \times 0.5 \times 0 = 0$

To find class, C_i, that maximizes $P(X \mid C_i)\, P(C_i)$, we compute

$P(X \mid \text{Income} = \text{High})\, P(\text{Income} = \text{High}) = 0.03 \times 0.5$

$P(X \mid \text{Income} = \text{Middle})\, P(\text{Income} = \text{Middle}) = 0 \times 0.17 = 0$

$P(X \mid \text{Income} = \text{Middle})\, P(\text{Income} = \text{Low}) = 0 \times 0.33 = 0$

As $P(X \mid \text{Income} = \text{High})\, P(\text{Income} = \text{High})$ has maximum value, tuple X will be classified as "High" income bracket.

* Using Laplacian correction to avoid computing probability values of zero.

* Linear Regression -

Q Following table shows house area (in sq mtr) & price (in Lacks)

(a) Assuming linear regression, use method of least squares get equation for prediction of price.

(b) Predict price of house with 80 sq mtr area.

Area	72	50	81	74	94	86
Price	84	63	77	78	90	75

$y = \alpha + \beta x$, α & β are coeff, $x \rightarrow$ area, $y \rightarrow$ price.

$$\beta = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$\bar{x} = \text{Avg}(\text{area}) = \frac{457}{6} = 76.16$

$\bar{y} = \text{Avg}(\text{price}) = \frac{467}{6} = 77.83$

$\alpha = \bar{y} - \beta \bar{x}$

$$\beta = \frac{(72-76.16)(84-77.83) + (50-76.16)(63-77.89) + (81-76.16)(77-77.83) + (74-76.16)(78-77.83) + (94-76.16)(90-77.83) + (86-76.16)(75-77.83)}{(72-76.16)^2 + (50-76.16)^2 + (81-76.16)^2 + (74-76.16)^2 + (94-76.16)^2 + (86-76.16)^2}$$

$$= \frac{547.16}{1144.8} = 0.477$$

$\alpha = 77.83 - 0.477 \times 76.16 = 41.50$

$y = 41.50 + 0.477\, x$

price $= 41.50 + 0.477 \times$ area

price $= 41.50 + 0.477 \times 80$

$= 79.66$ lacks.

Q. Chi square statistics - Correlation analysis with nominal data.

χ^2 (Chi-square test) $\chi^2 = \frac{(\text{Original} - \text{Expected})^2}{\text{Expected}}$

Q. (10, 0.5) & (25, 1.20) are actual purchase amount & tip amount pairs. Predict the amount of tip on the purchase of amount 55 using linear regression.

→ $y = \alpha + \beta x$.

$$\beta = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$\bar{x} = \text{avg}(x)$

$\bar{y} = \text{avg}(y)$.

$\alpha = \bar{y} - \beta \bar{x}$

$\bar{x} = \frac{10+25}{2} = 17.5$

$\bar{y} = \frac{0.5+1.2}{2} = 0.85$

$$\beta = \frac{(10-17.5)(0.5-0.85) + (25-17.5)(1.2-0.85)}{(10-17.5)^2 + (25-17.5)^2}$$

$= 0.046$

$\bar{\alpha} = 0.85 - 0.046 \times 17.5 = 0.045$.

$y = \alpha + \beta x$

$= 0.045 + 0.046 x$

$= 0.045 + 0.046 (55)$

$= 2.575$.

Q. Training data set

RID	age	income	student	credit-rating	buys-computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31..40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31..40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31..40	medium	no	excellent	yes
13	31..40	high	yes	fair	yes
14	>40	medium	no	excellent	no.

Classify following sample using Naive Bayses classifier.

$X = \{$ age = "<=30", income = "medium", student = "yes"
credit-rating = "fair" $\}$.

→ Need to maximize $P(X|C_i)$ for all classes.

$P(\text{buys-computer} = \text{yes}) = 9/14 = 0.643$
$P(\text{buys-computer} = \text{no}) = 5/14 = 0.357$

To Compute $P(X|C_i)$, compute conditional probabilities for each attribute for each available class.

$P(\text{age} = \text{"<=30"} \mid \text{buys-comp} = \text{"yes"}) = 2/9 = 0.222$
$P(\text{age} = \text{"<=30"} \mid \text{buys-comp} = \text{"no"}) = 3/5 = 0.600$
$P(\text{income} = \text{"medium"} \mid \text{buy-comp} = \text{"yes"}) = 4/9 = 0.444$
$P(\text{income} = \text{"medium"} \mid \text{buy-comp} = \text{"no"}) = 2/5 = 0.400$
$P(\text{student} = \text{"yes"} \mid \text{buys-comp} = \text{"yes"}) = 6/9 = 0.667$
$P(\text{student} = \text{"yes"} \mid \text{buys-comp} = \text{"no"}) = 1/5 = 0.200$
$P(\text{credit-rating} = \text{"fair"} \mid \text{buys-comp} = \text{"yes"}) = 6/9 = 0.667$
$P(\text{credit-rating} = \text{"fair"} \mid \text{buys-comp} = \text{"no"}) = 2/5 = 0.400$.

Using above probabilities,

$P(X \mid \text{buys-comp} = \text{"yes"}) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$
$P(X \mid \text{buys-comp} = \text{"no"}) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019$

$P(X \mid \text{buys-comp} = \text{"yes"})\, P(\text{buys-comp} = \text{"yes"}) = 0.044 \times 0.643 = 0.028$
$P(X \mid \text{buy-comp} = \text{"no"})\, P(\text{buys-comp} = \text{"no"}) = 0.019 \times 0.357 = 0.007$

Hence, Bayesian classifier predicts buys-comp = "yes" for sample X.

Naive Bayes Classifier -

Q. The table gives data set about stolen vehicles. Using Naive Bayes classifier, classify new data (Red, SUV, Domestic) as "stolen" or "not".

→.

Color	Type	Origin	Stolen
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes.

Compute probabilities -

$P(\text{stolen} = \text{yes}) = 5/10 = 0.5$

$P(\text{Stolen} = \text{no}) = 5/10 = 0.5$

Compute conditional probabilities w.r.t. new data point to be classified.

$P(\text{color} = \text{Red} \mid \text{stolen} = \text{yes}) = 3/5 = 0.6$

$P(\text{color} = \text{Red} \mid \text{stolen} = \text{no}) = 2/5 = 0.4$

$P(\text{Type} = \text{SUV} \mid \text{stolen} = \text{yes}) = 1/5 = 0.2$

$P(\text{Type} = \text{SUV} \mid \text{Stolen} = \text{no}) = 3/5 = 0.6$

$P(\text{Origin} = \text{Domestic} \mid \text{stolen} = \text{yes}) = 2/5 = 0.4$

$P(\text{Origin} = \text{Domestic} \mid \text{Stolen} = \text{no}) = 3/5 = 0.6$

Using above probabilities -

$P(X \mid \text{Stolen} = \text{yes}) = 0.6 \times 0.2 \times 0.4 = 0.048$

$P(X \mid \text{Stolen} = \text{no}) = 0.4 \times 0.6 \times 0.6 = 0.144$

$P(X \mid \text{Stolen} = \text{yes})\, P(\text{stolen} = \text{yes}) = 0.048 \times 0.5 = 0.024$ — ①

$P(X \mid \text{Stolen} = \text{no})\, P(\text{Stolen} = \text{no}) = 0.144 \times 0.5 = 0.072$ — ②

As ② > ①, Bayesian classifier predicts Stolen = "no" for sample record.

k- nearest neighbour classification

Q Classify the new data record ($x_1 = 3$, $x_2 = 7$) using nearest neighbour with $k=3$ using following data set.

x_1	7	7	3	1
x_2	7	4	4	4
y	Bad	Bad	Good	Good.

→ Need to find out distance between new data point & each of existing data point using ~~Manhattan~~ Euclidian dist^n formula.

$$d[(7,7),(3,7)] = \sqrt{(7-3)^2 + (7-7)^2} = 4$$
$$d[(7,4),(3,7)] = \sqrt{(7-3)^2 + (4-7)^2} = 5$$
$$d[(3,4),(3,7)] = \sqrt{(3-3)^2 + (4-7)^2} = 3$$
$$d[(1,4),(3,7)] = \sqrt{(1-3)^2 + (4-7)^2} = 3.6$$

The nearest three neighbours ($k=3$) of new data point (3,7) are (~~3,4~~), (3,4) & (1,4), (7,7)

~~Bad~~ ⇓ Good ⇓ Good ⇓ Bad

Using majority poll over the nearest neighbours, new point can be classified as "Good" as it holds majority in nearest 3 neighbours class labels.

Using Manhattan distance as dist^n formula -

$$d[(7,7),(3,7)] = |7-3| + |7-7| = 4+0 = 4$$
$$d[(7,4),(3,7)] = |7-3| + |4-7| = 4+3 = 7$$
$$d[(3,4),(3,7)] = |3-3| + |4-7| = 0+3 = 3$$
$$d[(1,4),(3,7)] = |1-3| + |4-7| = 2+3 = 5.$$

The nearest three neighbours (as $k=3$) of new data point are (3,4), (7,7) & (1,4)

⇓ Good ⇓ Bad ⇓ Good.

Using majority poll over the nearest neighbours, new point can be classified as "Good" as it holds majority in nearest 3 neighbours class labels.

Q. Consider one dimensional data set. Classify data point $x = 5.0$ according to its 3- & 5- nearest neighbours. Use majority vote.

x	0.6	3.2	4.5	4.6	4.9	5.2	5.6	5.8	7.1	9.5
y	−	−	+	+	+	−	−	+	−	−

→ distⁿ betⁿ $x = 5.0$ & each x value given in data set.

$d(x=5.0, x=0.6) = 4.4$
$d(x=5.0, x=3.2) = 1.8$
$d(x=5.0, x=4.5) = 0.5$
$d(x=5.0, x=4.6) = 0.4$
$d(x=5.0, x=4.9) = 0.1$
$d(x=5.0, x=5.2) = 0.2$
$d(x=5.0, x=5.6) = 0.6$
$d(x=5.6, x=5.8) = 0.8$
$d(x=5.0, x=7.1) = 2.1$
$d(x=5.0, x=9.5) = 4.5$

Using 3- nearest neighbours method, 3 closest point to the point $x = 5.0$ are 5.2, 4.9, 4.6.
Corresponding classes are −, +, +.
Using majority vote, class of $x = 5.0$ will be +.

Using 5- nearest neighbours method, 5 closest point to the point $x = 5.0$ are 4.5, 4.6, 4.9, 5.2 & 5.6 with corresponding classes as +, +, +, −, −.
Using majority vote, class of $x = 5.0$ will be +.

Q.8

	Outlook	Temp	Humidity	Windy	Class.
1	Sunny	71..80	61..70	true	Play
2	Sunny	71..80	81..90	true	Don't
3	Sunny	81..90	81..90	false	Don't
4	Sunny	81..80	91..100	true	Don't
5	Sunny	61..70	61..70	false	Play
6	Overcast	71..80	81..90	true	Play
7	Overcast	81..90	71..80	false	Play
8	Overcast	61..70	61..70	true	Play
9	Overcast		71..80	false	Play
10	Rain	81..90	71..80	true	Don't
11	Rain	71..80	61..70	true	Don't
12	Rain	61..70	71..80	false	Play
13	Rain	71..80	71..80	false	Play
14	Rain	61..80	91..100	false.	Play
		61..80			

→ 1. Overall entropy before splitting

$$E_{orig} = -\frac{9}{14}\ln\frac{9}{14} - \frac{5}{14}\ln\frac{5}{14} = 0.284 + 0.367 = 0.651$$

Contigency tables for Outlook, for Temp.

Outlook →	Sunny	Over..	Rain
Play	2	4	3
Don't	3	0	2

Temp →	61..70	71..80	81..90
Play	2	3.	2
Don't	1	3	1

~~Info gain after splitting on outlook~~

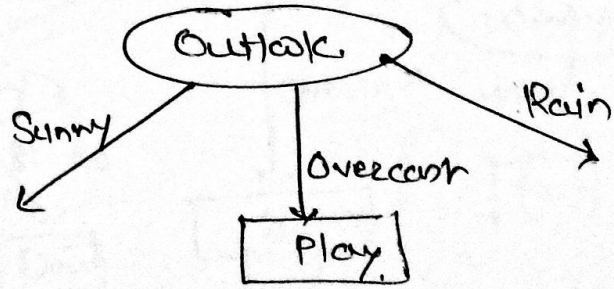
for Humidity

Humidity	61..70	71..80	81..90	91..100
Play	3	4	1	1
Don't	1	1	2	1

for windy

Windy →	True	false.
Play	3	6
Don't	4	1

As one pure node is obtained while splitting on attribute Outlook it can be taken as first attribute for splitting.



Outlook = Sunny

for Temp.

Temp →	Outlook = Sunny		
	61..70	71..80	81..90
Play	1	1	0
Don't	0	2	1

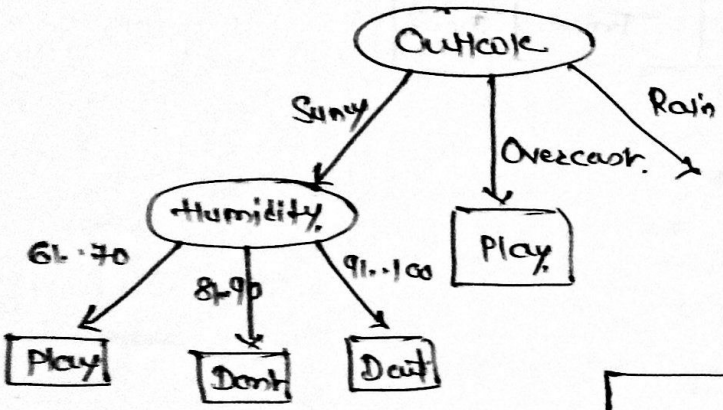
for Humidity

H →	Outlook = Sunny			
	61..70	71..80	81..90	91..100
Play	2	0	0	0
Don't	0	0	2	1

for windy

W →	Outlook = Sunny	
	true	false
Play	1	1
Don't	2	1

As pure nodes are obtained for attribute humidity, it can be taken as second attribute for splitting



When Outlook = Rain

for Temp

T →	Outlook = Rain		
	61..70	71..80	81..90
Play	2	1	1
Don't	1	1	0

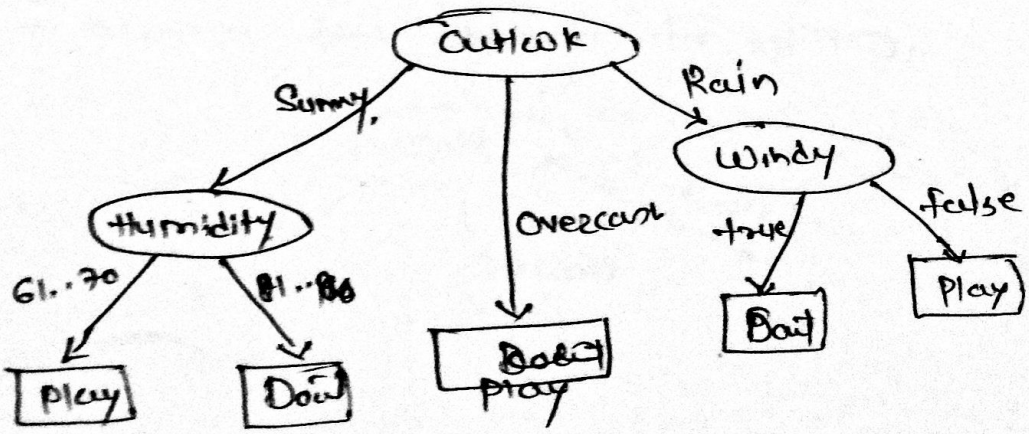
for Humidity.

H →	Outlook = Rain			
	61..70	71..80	81..90	91..100
Play		3	0	1
Don't	1	1	0	0

for windy.

W →	Outlook = Rain	
	true	false
Play	0	4
Don't	2	0

Pure nodes are obtained when splitting on attribute windy.



R1 : Outlook = Sunny & humidity 61..70 → Play

R2 : Outlook = Sunny & humidity 81..100 → Don't

R3 : Outlook = overcast → Play

R4 : Outlook = Rain & windy = false → Play

R5 : Outlook = Rain & windy = true → Don't.

Outlook	Temp	Humidity	Windy	Class
Rain	65	80	false	Play
Sunny	90	75	True.	Don't