Q) Test data has 45 instances

error rate = 6.67 %.

⇒ accuracy = 93.33 %.

The accuracy of a random variable is a Gaussian distribution from Central Limit Theorem.

Now, data I have-

mean = 0.0667 = p

variance = $\sqrt{p(1-p)}$ =

By Bernoulli's Theorem,

Test interval = $p \pm z \times \sqrt{\frac{p(1-p)}{n}}$

$= 0.0667 \pm 1.96 \times \sqrt{\frac{0.0667 \times 0.933}{45}}$

$= 0.0667 \pm 1.96 \times 0.037$

$= 0.0667 \pm 0.0728$

$= 0.1395, -0.0061$

∴ Interval : $[-0.0061, 0.1395]$

Q2 Test data has 45 instances.

$\Rightarrow n = 45$

Hypothesis	Error (%)	Mean	Standard Deviation
h_1	6.67	0.0667	0.0372
h_2	8.89	0.0889	0.0424
h_3	13.30	0.1330	0.0506

Standard Deviation $= \sqrt{\frac{p(1-p)}{n}}$; p: Mean

Δh_{21} :

Mean $= 0.0222$

Standard Deviation $= \sqrt{0.0372^2 + 0.0424^2} = 0.0564$

$$\text{Area} = \frac{\text{Mean}}{\text{Standard Deviation}} = \frac{0.0222}{0.0564}$$

$$= 0.39$$

$\therefore$ Confidence $\simeq 30\%$

$\therefore h_2$ is underperforming compared to h_1.

Δh_{31} :

Mean $= 0.0663$

Standard deviation $= \sqrt{0.0872^2 + 0.0506^2} = 0.0628$

$$\text{Area} = \frac{\text{Mean}}{\text{Standard Deviation}} = \frac{0.0663}{0.0628}$$

$$= 1.0557$$

$\therefore$ Confidence $\simeq 68\%$.

$\therefore h_3$ underperforms compared to h_1.

Q3

Error Rates

CV fold	Favorite Algo	Decision Tree	Δ(Diff.)
1	8.89	9.30	0.41
2	9.52	9.48	-0.04
3	8.13	9.12	0.99
4	9.48	9.13	-0.35
5	10.12	9.98	-0.14
6	10.23	11.01	0.78
7	8.56	9.02	0.46
8	9.12	8.56	-0.56
9	9.23	9.23	0
~~10~~	~~9.11~~	~~9.08~~	~~-0.03~~
			3.73

$$\text{Mean} = \frac{3.73}{9} = 0.4144 = p$$

$$\text{Standard deviation} = \sqrt{p(1-p)} = \sqrt{\frac{0.4144 \times 0.5855}{9}}$$

$$= 0.255$$

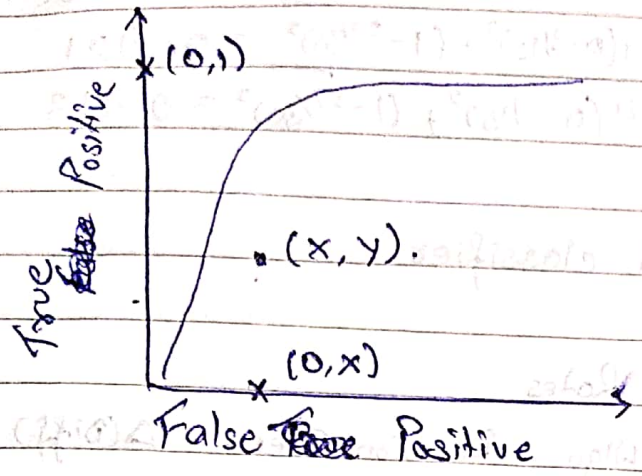
$$\text{Area} = \frac{\text{Mean}}{\text{Standard Deviation}} = \frac{0.4144}{0.255}$$

$$= 1.62 = Z$$

From table,

confidence ≃ 90%

Q4 ROC curve is a trade off b/w sensity (True positives) & fallout (False positives).



For ideal classifier the euclidean distance must be low i.e. point should lie close to (0,1).

For a given confusion matrix,

True Positive: Predicted class - P
Actual class - P

False Positive: Predicted class - P
Actual class - N

	Hypothesis		
	h_1	h_2	h_3
TP	$\frac{29}{30}$	$\frac{29}{30}$	$\frac{27}{30}$
FP	$\frac{1}{14}$	$\frac{3}{15}$	$\frac{3}{15}$

Euclidean distance $= \sqrt{(0-x)^2 + (1-y)^2}$

(a) Equal cost for false positive & negative

h_2: Distance $= \sqrt{(0 - 3/15)^2 + (1 - 29/30)^2} = 0.041$, closest to (0,1).

h_1: Distance $= \sqrt{(0 - 1/14)^2 + (1 - 29/30)^2} = 0.079$

h_3: Distance $= \sqrt{(0 - 3/15)^2 + (1 - 27/30)^2} = 0.050$

$\therefore h_2$ is ideal classifier

(b) False positive cost 4 times false negative

closest to

h_1 : Distance = $\sqrt{4(0-8/14)^2 + (1-29/30)^2}$ = 0.1467

h_2 : Distance = $\sqrt{4(0-3/15)^2 + (1-29/30)^2}$ = 0.401

h_3 : Distance = $\sqrt{4(0-3/15)^2 + (1-27/30)^2}$ = 0.412

$\therefore h_1$ is ideal classifier