Name - Yash Soni

1 a) Prediction using Naive Bayes and KNN where $k = 2$.

Naive Bayes :

Probabilities of the whole dataset

$P(\text{spam}) = \frac{6}{10} = 0.6$

$P(\text{ham}) = \frac{4}{10} = 0.4$

For spam class :

$P(\text{contains link} = \text{Yes} \mid \text{Spam}) = \frac{4}{6} = 0.67$

$P(\text{contains money words} = \text{No} \mid \text{spam}) = \frac{1}{6} = 0.17$

$P(\text{length} = \text{long} \mid \text{spam}) = \frac{4}{6} = 0.67$

$P(X \mid \text{Spam}) = 0.67 \times 0.17 \times 0.67$
$= 0.076$

$P(\text{Spam} \mid X) = P(X \mid \text{spam}) \times P(\text{Spam})$
$= 0.076 \times 0.6$
$= \underline{0.0456}$

For Ham class

$P(\text{contains link} = \text{Yes} \mid \text{Ham}) = \frac{1}{4} = 0{\cdot}25$

$P(\text{contains money words} = \text{No} \mid \text{Ham}) = \frac{4}{4} = 1$

$P(\text{length} = \text{long} \mid \text{ham}) = \frac{1}{4} = 0{\cdot}25$

$P(\text{Ham} \mid x) = P(x \mid \text{Ham}) \times P(\text{Ham})$

$P(X \mid \text{Ham}) = 0{\cdot}25 \times 1 \times 0{\cdot}25$
$= 0{\cdot}0625$

$P(\text{Ham} \mid x) = 0{\cdot}0625 \times 0{\cdot}4$
$= \underline{0{\cdot}025}$

Comparing both the probabilities :

$P(\text{Spam} \mid x) > P(\text{Ham} \mid x)$

This test example is classified as Spam.

KNN where $k = 2$

It relies on distance calculations so we will convert data to numbers. For instance, Yes as 1 & No as 1 Long as 1 and short as 0, Spam as 1 and Ham as 0.

Eudidean distance $= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$

For given test sample (contains link = Yes, contains money words = No, length = long) or $(1, 0, 1)$ to compute distance.

ID = 1 $(1, 1, 1)$ Spam class
Distence calculation $= \sqrt{(1-1)^2 + (0-1)^2 + (1-1)^2} = 1$

ID = 2 $(0, 0, 0)$ = Ham class
Distance $= \sqrt{(1-0)^2 + (0-0)^2 + (1-0)^2} = 1.41$

ID = 3 $(1, 0, 1)$ Spam class
Distance $= \sqrt{(1-1)^2 + (0-0)^2 + (1-1)^2} = 0$

ID = 4 $(0, 1, 0)$ Spam class
Distance $= \sqrt{(1-0)^2 + (0-1)^2 + (1-0)^2} = 1.73$

ID = 5 $(1, 1, 0)$ Spam class
Distance $= \sqrt{(1-1)^2 + (0-1)^2 + (1-0)^2} = 1.41$

ID = 6 $(0, 0, 1)$ Ham class
Distance $= \sqrt{(1-0)^2 + (0-0)^2 + (1-1)^2} = 1$

ID = 7 $(1, 0, 0)$ Ham class

Distance $= \sqrt{(1-1)^2 + (0-0)^2 + (1-0)^2} = 1$

ID = 8 $(0, 1, 1)$ Spam class

Distance $= \sqrt{(1-0)^2 + (0-1)^2 + (1-1)^2} = 1.41$

ID = 9 $(1, 1, 1)$ Spam class

Distance $= \sqrt{(1-1)^2 + (0-1)^2 + (1-1)^2} = 1.0$

ID = 10 $(0, 0, 0)$ Ham class

Distance $= \sqrt{(1-0)^2 + (0-0)^2 + (1-0)^2} = 1.41$

After sorting the distance in ascending order, we get ID 3 = 0 as Spam and ID 1 = 1 as spam.

This test example is classified as Spam.

2a)

True positives and false positives of the dataset are as follows :

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Based on the formulas, the below table & values are calculated

Threshold	TP	FP	TN	FN	TPR	FPR
0·95	39	4	74	33	0·541	0·051
0·90	46	5	73	26	0·638	0·064
0·85	51	5	73	21	0·708	0·064
0·80	54	5	73	18	0·750	0·064
0·75	55	6	72	17	0·763	0·076
0·70	58	6	72	14	0·805	0·076