DMBI Assignment 1

1. Mean $\bar{x} = \sum_{i=1}^{2} n_i/n = 809/27 = 29.96 = 30$

The median (middle value of the ordered set, as the number of values in the set is odd) of the data is 25.

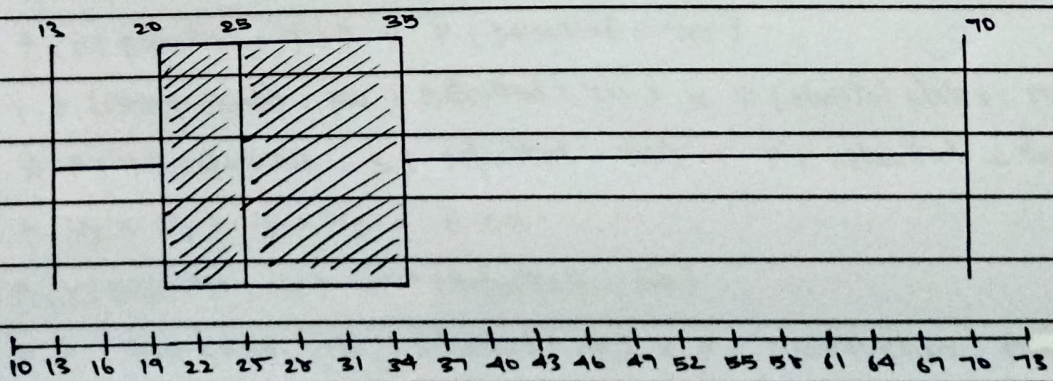
The data set has two values that occur with the same highest frequency and is, therefore bimodal. The mode (value) occurring with greatest frequency) of the data are 25 and 35.

The mid-range (average of the largest and smallest values in the data set) of the data is $(70+13)/2 = 41.5$

Five-point summary :

* Minimum value of the data = 13
* The first quartile Q_1 (corresponding to the 25th percentile) of the data = 20
* The median of the data = 25
* The third quartile Q_3 (corresponding to the 75th percentile) of the data = 35
* Maximum value of the data = 70

$\therefore$ (13, 20, 25, 35, 70)



2. From the above table, it is clear that there are 3 tuples classified as Defaulted = Yes and 7 tuples classified as Defaulted = No.

From the given training data, we estimate

P(Defaulted = Yes) = 3/10

P(Defaulted = No) = 2/10

Probabilities associated with attributes is given in table below:

Attribute	Value	Count		Probabilities	
		Yes	No	Yes	No
Homeowner	Yes	1	3	1/3	3/7
	No	2	4	2/3	4/7
Marital status	Single	1	2	1/3	2/7
	Married	1	4	1/3	4/7
	Divorced	1	1	1/3	1/7
Job experience (in years)	2	2	1	2/3	1/7
	3	1	2	1/3	2/7
	4	0	3	0	3/7
	5	0	1	0	1/7

Unknown tuple is t = <Homeowner : No, Marital status : Married, Job experience : 3>

P(t | Defaulted = Yes) x P(Defaulted = Yes)

= P(Home owner : No | Defaulted = Yes) x P(Marital status : Married | Defaulted = Yes)

x P(Job experience : 3 | Defaulted = Yes) x P(Defaulted = Yes)

= 1/3 x 1/3 x 1/3 x 1/3 = 0.011

P(t | Defaulted = No) x P(Defaulted = No)

= P(Homeowner : No | Defaulted = No) x P(Marital status : Married | Defaulted = No)

x P(Job experience : 3 | Defaulted = No) x P(Defaulted = No)

= 4/7 x 4/7 x 2/7 x 7/10 = 0.032

Based on above probabilities, we classify the new tuple as Defaulted = No because it has the highest probability in Naive Bayes model.

Using Decision Tree:

Defaulted = No Probability = 7/10

Defaulted = Yes Probability = 3/10

Information = $I(7,3)$

$= -7/10 \log_2 7/10 - 3/10 \log_2 3/10 = 0.3602 + 0.5211$

$= 0.8813$

Homeowners: $I_{homeowners} = 4/10 I(3,1) + 6/10 I(4,2)$

$= 4/10 \times 0.43627 + 6/10 + 0.918$

$= 0.7253$

∴ Information gain = $0.8813 - 0.7253 = 0.156$

Marital status: $I_{marital\ status} = 3/10 I(2,1) + 5/10 I(4,1) + 2/10 I(1,1)$

$= 0.83635$

∴ Information gain = $0.8813 - 0.83635 = 0.04495$

Job experience: $I_{job\ experience} = 3/10 I(2,1) + 3/10 I(2,1) + 4/10 I(4,0)$

$= 0.5508$

∴ Information gain = $0.8813 - 0.5508 = 0.3305$

The information gain for job experience was the highest, we select that attribute as the root node. The tuple we want to classify has job experience = 3.

∴ For job experience = 3,

Homeowner	Marital status	Defaulted
Yes	Single	No
No	Married	Yes
No	Married	Yes

These last two records adds to the ambiguity in the default dataset and make it hard to classify.

Handling ambiguous data in Decision Tree :

During the algorithm, when class prediction is undecided -

* Approach is to predict the majority outcome class for the branch.
* Find more information on attributes if possible OR
* Remove the ambiguous tuples.

Hence, we choose the majority class of the branch 'job experience = 3, which is defaulted = No. Hence the tuple will be classified as Defaulted = No.

3. The 2x2 confusion matrix is denoted as :

		Predicted class	
		Yes	No
Actual class	Yes	TP	FN
	No	FP	TN

Given confusion matrix -13

Cancer (class)	Yes	No	Total
Yes	90	210	300
No	140	9560	9700
Total	230	9770	10000

Comparing the above matrices, we get

TP = 90 , FN = 210 , FP = 140 and TN = 9560

Accuracy = TP+TN / TP+TN+FP+FN = 90 + 9560 / 90 + 9560 + 210 + 140

= 0.965 = 96.5%

Precision = TP / TP+FP

= 90/90+140 = 0.3913 = 39.13 %

Recall = TP / TP+FN = 90/90+210

= 0.3 = 30 %

4. Information Gain : It is the reduction in the entropy and measures how well a given feature classifies the target classes. The factor within the highest information gain is selected as the best one. Entropy (D) is measure of disorder and is given by

$$D = \sum_{i=1}^{n} P_i \log_2 1/P_i$$

∴ Information gain = D − D(A) where A is a selected attribute.

Gain ratio : It is an alteration of information gain that reduces its favouritism on high-branch attributes. It considers number of branches and size of branches when it select attribute to split.

Intrinsic information = Entropy of distribution of instances into branches

∴ Intrinsic Info (S, A) = $- \sum |S_i|/|S| \log_2 |S_i|/|S|$

∴ Gain ratio = $\frac{\text{Information Gain (S, A)}}{\text{Entropy / Intrinsic Info (S, A)}}$

Gini Index : It computes the degree of probability of a specific variable that is wrongly being classified. We select the attribute as the node of the tree where Gini is the lowest.

Gini (Attribute = value) = $1 - \sum_{i=1}^{m} (P_i)^2$ where m = no. of classes.

Different types of attributes used in Data Mining used are :

★ Nominal attribute : It is a qualitative attribute. The values of a nominal attribute are names of things or some symbols. these values represent some category or state thus these attributes are also called categorical attributes. there is no order or rank among values of such attribute.

★ Binary attribute : It is a qualitative attribute. Binary data has only 2 possible values or states. Symmetric binary attribute is when both values are equally important and

asymmetric binary attribute is when both values are not equally important.

★ Ordinal attribute : It is a qualitative attribute. The ordinal attribute contain values that have a meaningful sequence or ranking between them, but the magnitude between values is not actually known.

★ Numeric attribute : It is a quantitative attribute as it is a measurable quantity. There are two types :

Interval scaled attribute : These have values whose differences are interpretable but they don't have a reference point. Data can be added or subtracted but not multiplied or divided.

Ratio scaled attribute : These have numeric values with a fixed zero point. We can say that a value is a multiple of the other. These values are ordered and we can draw a five point summary for them.

★ Discrete attribute : It is a quantitative attribute. These attributes have finite or countably infinite set of values.

★ Continous attribute : It is a quantitative attribute. It can take any value between two specified values.