→ DMW Assignment 2

Q1 $P(C_i)$ - Prior Probability of each class

$P(\text{lenses} = \text{'Non soft contact'}) = \frac{13}{23}$ [Adding 1 to avoid 0 available problem]

$P(\text{lenses} = \text{'soft contact'}) = \frac{5}{23}$

$P(\text{lenses} = \text{'Hard Contact'}) = \frac{5}{23}$

$X = \langle \text{Young}, \text{Myope}, \text{Yes}, \text{Reduced} \rangle$

$P = (\text{Age} = \text{'Young'} \mid \text{lenses} = \text{'Non Contact'}) = \frac{5}{13}$

$P = (\text{Age} = \text{'Young'} \mid \text{lenses} = \text{'Soft Contact'}) = \frac{3}{5}$

$P = (\text{Ages} = \text{'Young'} \mid \text{lenses} = \text{'Hard Contact'}) = \frac{3}{5}$

$P = (\text{Spectacle Pres} = \text{"Myope"} \mid \text{lenses} = \text{'Non contact'}) = \frac{8}{13}$

$P(\text{Spectacle Pres} = \text{"Myope"} \mid \text{lenses} = \text{'Soft Contact'}) = \frac{3}{5}$

$P(\text{spectacle Pres} = \text{"Myope"} \mid \text{lenses} = \text{'Hard Contact'}) = \frac{4}{5}$

$P(\text{Astigmatic} = \text{'Yes'} \mid \text{lenses} = \text{"Non contact"}) = \frac{7}{13}$

$P(\text{Astigmatic} = \text{"Yes"} \mid \text{Lenses} = \text{"Soft Contact"}) = \frac{1}{5}$

$P(\text{Astigmatic} = \text{"Yes"} \mid \text{lenses} = \text{"Hard Contact"}) = \frac{5}{5}$

$P(\text{Test Production Rate} = \text{"Reduced"} \mid \text{Lenses} = \text{'Non Contact'}) = \frac{11}{13}$

$P(\text{Test Production Rate} = \text{"Reduced"} \mid \text{Lenses} = \text{'Soft contact'}) = \frac{1}{5}$

$P(\text{Test production Rate} = \text{"Reduced"} \mid \text{Lenses} = \text{'Hard Contact'}) = \frac{1}{5}$

$P(x \mid \text{Lenses} = \text{'Non contact'}) = \frac{5}{13} \times \frac{8}{13} \times \frac{7}{13} \times \frac{11}{13} = 0.108$

$P(x \mid \text{Lenses} = \text{"soft contact"}) = \frac{3}{5} \times \frac{4}{5} \times \frac{5}{5} \times \frac{1}{5} = 0.096$

$P(x \mid \text{Lenses} = \text{'Hard contact'}) = \frac{3}{5} \times \frac{4}{5} \times \frac{5}{5} \times \frac{1}{5} = 0.096$

$P(\text{Lenses} = \text{'Non contact'} \mid X) = \frac{13}{23} \times 0.0924 = 0.0522$

$P(\text{Lenses} = \text{'soft contact'} \mid X) = \frac{5}{23} \times 0.0944 = 0.0031$

$P(\text{Lenses} = \text{'hard contact'} \mid X) = \frac{5}{23} \times 0.096 = 0.0209$

X belongs to the class (Lenses = 'Non Contact'

2a) The basic decision tree algorithm should be modified as follows to take up into consideration the count of each generalized data tuple

→ The count of each tuple must be integrate into the calculation of the attribute selection measure (such as the information gain)

→ Take the count into consideration to determine the most common class among the tuples

b) No of Junior = 113
No of Senior = 52

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

$$Info(D) = -\frac{113}{165}\log_2\left(\frac{113}{165}\right) - \left(\frac{52}{165}\right)\log_2\left(\frac{52}{165}\right)$$

$$= 0.8990$$

Department

	P_i	N_i	Entropy
Sales	80	30	0.8454
System	23	8	0.8238
marketing	4	10	0.8631
Secretary	6	4	0.9710

$$\text{Entropy for department} = \frac{110}{165} \times 0.8454 + \frac{31}{165} \times 0.8238 + \frac{14}{165} \times 0.8631 + \frac{10}{165} \times 0.9710$$

$$= 0.8505$$

$$\text{Gain} = 0.899 - 0.8505$$

$$= 0.0485$$

Age	P	N	Entropy
21..25	20	0	0
26..30	49	6	0
31..35	44	35	0.9906
36..40	0	0	0
41..45	0	0	0
46..50	0	0	0

Age entropy = $\frac{44+35}{165}$

$= \frac{79}{165} = 0.4247$

Gain = 0.8990 − 0.4743

= 0.4247

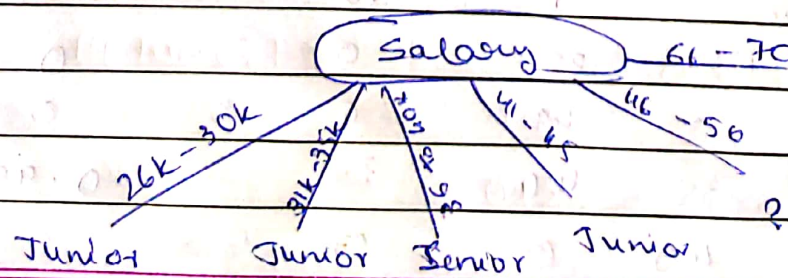
Salary	P	N	Entropy
21 − 30K	46	0	0
31 − 35K	40	0	0
36 − 40K	0	4	0
41 − 45K	4	0	0
46 − 50K	23	40	0.9468
66 − 70K	0	8	0

Info (Salary) Entropy = $\frac{23+40}{165} \times 0.9468$

= 0.5375

Now gain (salary) > gain (age) > Gain (dept)

∴ Salary attribute is selected as splitting attribute.



$\therefore$ Info $D_1 = \frac{-40}{63}\log\left(\frac{40}{63}\right) - \frac{23}{63}\log\left(\frac{23}{63}\right)$

$= 0.9168$

Consider dept attribute

$Info = \frac{30}{63}\log_2(\text{sales}) + \frac{23}{63} \times Info(\text{system})$

$+ \frac{10}{63} \times Info(\text{marketing})$

$= 0$

Gain $= 0.9468$

Consider age attribute

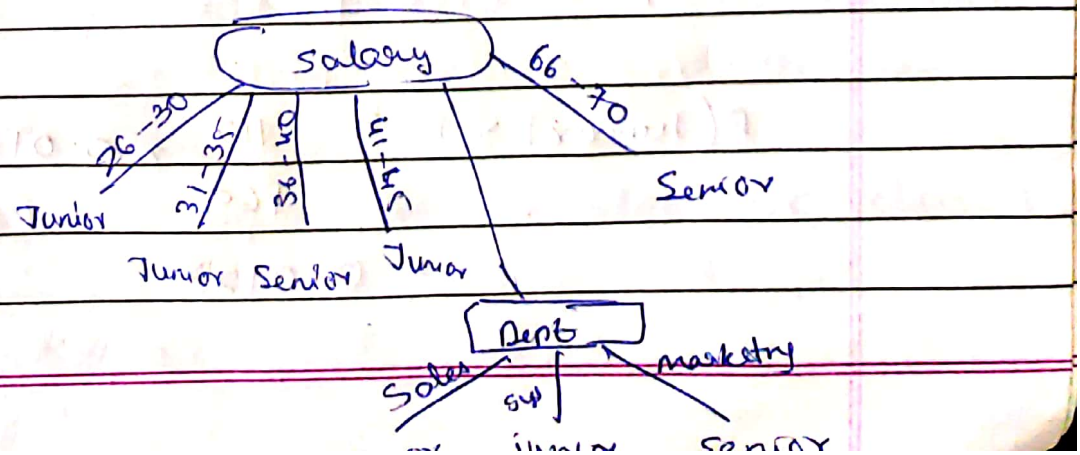
$Info(D_1) = \frac{20}{63} Info(21-25) + \frac{3}{63} Info(26-30)$

$+ \frac{30}{863} Info(31-35) + \frac{10}{63} \times Info(36-40)$

$= 0$

Gain $= 0.9468$

$\therefore$ Gain for both are same.



3) Prior probabilities

$$P(\text{senior}) = \frac{52}{165}$$

$$P(\text{junior}) = \frac{113}{165}$$

Consider attribute department

$$P(\text{system} \mid \text{senior}) = \frac{8}{52}$$

$$P(\text{systems} \mid \text{junior}) = \frac{23}{113}$$

Posterior Probabilities

$$P(X \mid \text{junior}) = \frac{23 \times 49 \times 23}{113 \times 113 \times 113} = 0.018$$

$$P(X \mid \text{senior}) = \frac{6 \times 0 \times 40}{52 \times 52 \times 52} = 0$$

$\therefore$ Bayes theorum $P\left(\frac{c_i}{x}\right) = \frac{P(c_i) \times P(x \mid c_i)}{P(x)}$

$$P(\text{Junior} \mid x) = \frac{113}{165} \times 0.018$$

$$= 0.0123$$

$P(\text{senior} \mid X) = 0 \times \frac{52}{165} \times 20$

$\therefore 0.43 > 0$

$X = (\text{Systems}, 26 \ldots 30, 46 \ldots 50k)$ will be predicted as status junior

3 To develope a quality classifier to guard against fraudulent credit and transactions with limited fraudulent cases, we can employ various techniques -

(1) Data collection - Gather a large data set of non-fraudulent transaction. Collect a small but diverse set of fraudulent transactions. They must ensure that various types of frauds are represented.

(2) Data preprocessing
Clean and preprocess data by handling missing values and scaling features. There is imbalance in the dataset due to uneven number of fraudulent transaction. We must try and oversample fraudulent or undersample non fraudulent to balance

(3) Classifier Selection
Choose appropriate ML algo for classifier such as random forest or gradient boosting.

④ Model Training.
Train classifier based on preprocessed balanced dataset so that it learns to distinguish between fraudulent and non fraudulent transaction.

⑤ Evaluation metrics.
Deploy suitable metrics such as AUC-ROC and precision, to evaluate the models performance.

⑥ Combine multiple models using bagging and boosting to improve classification accuracy

⑦ Feedback mechanism must be set up so that in case a fraudulent transaction is skipped by the model, feedback can be provided inorder to improve and update the model.

4 Accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

$\Rightarrow$ accuracy $= \frac{TP + TN}{P + N}$

TP – True Positive

TN – True Negative.

Sensitivity or true positive Rate $= \frac{TP}{P}$

Sensitivity or True negativity Rate $= \frac{TN}{N}$

P x Sensitivity = TP — (4)

N x Specificity = TN — (5)

Put (4), (5) in (1)

$$\frac{P}{P+N} \times \text{Sensitivity} + \frac{N}{P+N} \times \text{Specificity} = \text{accuracy}$$

Hence, accuracy of a classifier is a function of sensitivity and specificity.

5 Compare & contrast Bagging & Boosting

	Bagging	Boosting
(1)	The original dataset is divided into multiple subsets, selecting observations with replacements.	The new subset contains the components misstrained by the previous model.
(2)	This method combines predictions that belong to the same type.	This method combines predictions that belong to different types.
(3)	Bagging decreases variance	Bagging decreases bias
(4)	Base classifiers are trained parallely	Base classifiers are trained sequentially.
(5)	The models are created independently	The model creation is dependent on the previous on