Nome: Ayush Jain
SAP ID: 60004200132
Div: B2
Computer Engineering

DMW - Assignment 2

Q. 1 $X = <$ Young, Myope, Yes, Reduced $>$

→ Let A → Age
S → Spectacle Prescription
B → Astigmatism
T → Tear Production rate
L → class lable lens

X → (Young, Myope, Yes, Reduced)

(i) $P(\text{Young} / \text{Non-contact}) = \dfrac{P(\text{Non Contact} / \text{Young}) \cdot P(\text{Young})}{P(\text{Non-contact})}$

$$= \frac{\frac{4 \times 8}{12} \times \frac{4}{8} \times \frac{8}{20}}{12/20} = \frac{4}{12} = \frac{1}{3}$$

$\therefore P(\text{Young} / \text{Non-contact}) = \frac{1}{3}$

(ii) $P(\text{Myope}/\text{Non-contact}) = \dfrac{P(\text{Non-contact}/\text{Myope}) \cdot P(\text{Myope})}{P(\text{Non-contact})}$

$$= \frac{7/12 \cdot 12/20}{12/20} = \frac{7}{12}$$

$\therefore P(\text{Myope} / \text{Non-contact}) = \frac{7}{12}$

(iii) $P(Yes / Non-contact) = \dfrac{P(Non-contact / Yes) \cdot P(Yes)}{P(Non-contact)}$

$$= \frac{7/12 \cdot 12/20}{12/20} = \frac{7}{12}$$

$\therefore P(Myope / Non-contact) = \dfrac{7}{12}$

(iv) $P(Reduced / Non-contact) = \dfrac{P(Non-contact / Reduced) \cdot P(Reduced)}{P(Non-contact)}$

$$= \frac{10/10 \cdot 10/20}{12/20} = \frac{5}{6}$$

$\therefore P(Reduced / Non-contact) = \dfrac{5}{6}$

(v) $P(x / Non-contact) = \dfrac{1}{3} \times \dfrac{7}{12} \times \dfrac{1}{2} \times \dfrac{5}{6}$

$$= \frac{35}{432}$$

$$= 0.081$$

$\therefore P(x / Non-contact) = 0.081$

(vi) $P(Young / Soft\ contact) = \dfrac{P(Soft\ contact / Young) \cdot P(Young)}{P(Soft\ contact)}$

$$= \frac{2/8 \times 8/20}{4/20} = \frac{1}{2}$$

$\therefore P(Young / Soft\ contact) = \dfrac{1}{2}$

(viii) $P(\text{Myopic} / \text{soft contact}) = \dfrac{P(\text{soft contact} / \text{Yes}) \cdot P(\text{Yes})}{P(\text{soft contact})}$

$$= \frac{\frac{0}{10} \times 10/20}{4/20} = 0$$

$\therefore P(\text{Yes} / \text{soft contact}) = 0$

(ix) $P(\text{Young} / \text{Hard Contact}) = \dfrac{P(\text{Hard Contact} / \text{Young}) \cdot P(\text{Young})}{P(\text{Hard contact})}$

$$= \frac{2/8 \times 8/20}{4/20} = \frac{1}{2}$$

(x) $P(\text{Myopic} / \text{Hard contact}) = \dfrac{P(\text{Hard contact} / \text{Myopic}) \cdot P(\text{Myopic})}{P(\text{Hard contact})}$

$$= \frac{3/12 \times 12/20}{4/20} = \frac{3}{4}$$

$\therefore P(\text{Myopic} / \text{Hard contact}) = \frac{3}{4}$

(xi) $P(\text{Yes} / \text{Hard contact}) = \dfrac{P(\text{Hard Contact} / \text{Yes}) \cdot P(\text{Yes})}{P(\text{Hard contact})}$

$$= \frac{4/10 \cdot 10/20}{4/20} = 1$$

$P(\text{Yes} / \text{Hard contact}) = 1$

(xii) $P(\text{Reduced / Hard contact}) = \frac{P(\text{Hard contact / Reduced}) \cdot P(\text{Reduced})}{P(\text{Hard contact})}$

$$= \frac{0/10 \times 10/20}{4/20} = 0$$

(xiii) $P(x / \text{Hard contact}) = 0$

$\therefore P(x / \text{Non contact}) > P(x \mid \text{Soft contact}) = P(x / \text{Hard contact})$

$\therefore 0.081 > 0 > 0$

$\therefore$ Prediction of <Young, Myope, yes, Reduced> → Non-contact.

Q. 2)

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

- The count of each tuple must be integrated into the calculation of attribute selection measuring
- Take the count into consideration to determine the most common class among the tuples.

<b> $P = 113$

$N = 52$

$$DE = \frac{-113}{52} \log_2 \left(\frac{113}{11} \right) - \frac{52}{165} \log_2 \left(\frac{52}{165} \right)$$

$DE = 0.8990$

Department	P_i	N_i	Entropy
Sales	80	30	0.8454
System	23	8	0.8238
Marketing	4	10	0.8631
Secretory	6	4	0.9709

Entropy for department: $\frac{80+30}{105}(0.8454) + \frac{23+8}{165}(0.8238)$

$+ \frac{4+10}{165}(0.8631) + \frac{4+6}{165}(0.9709)$

$= 0.8504$

Gain $= 0.99 - 0.8504$

$= 0.049$

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

Age Entropy $= \frac{44+35}{165}(0.9906)$

$= \frac{79}{165} = 0.4743$

$\therefore$ Gain $= 0.8990 - 0.4743$

$= 0.4247$

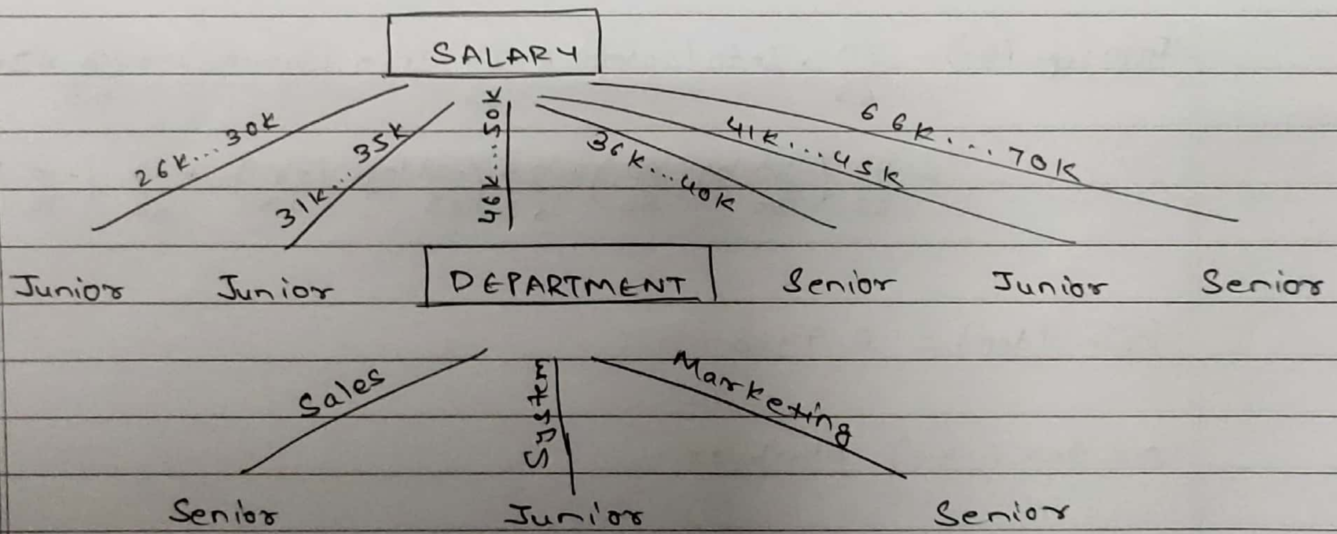
SALARY	P	N	Entropy
26K...30K	46	0	0
31K...35K	40	0	0
36K...40K	0	4	0
41K...45K	4	0	0
46K...50K	23	40	0.9468
66K...70K	0	8	0

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

$= 0.3615$

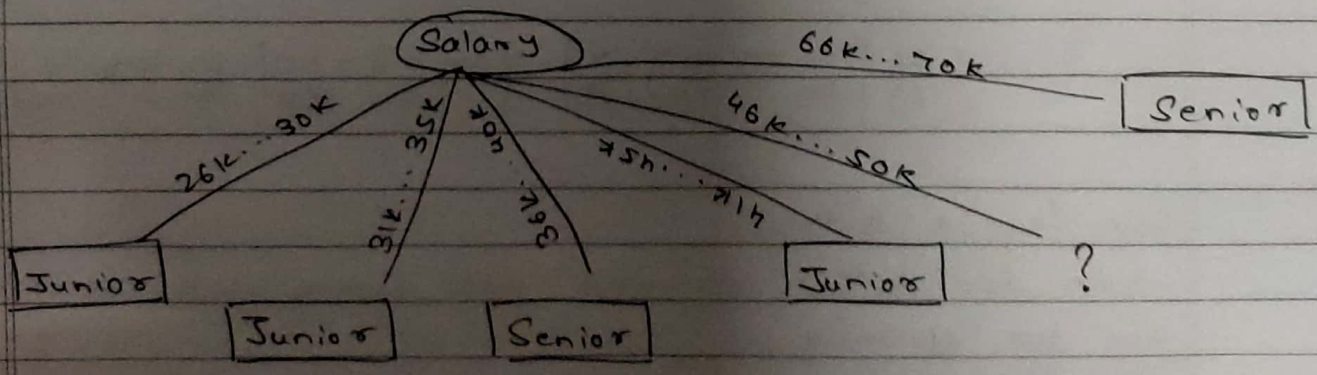
$\therefore$ Gain = 0.5375

DECISION TREE:



Now, Gain (Salary) > Gain (Age) > Gain (department)

$\therefore$ Salary is selected as splitting attribute



Now, we consider data tuples having salary 40K ... 50K and find the splitting attributes for it.

$$\text{Info}(D_1) = -\frac{40}{63}\log_2\frac{40}{63} - \frac{23}{63}\log_2\frac{23}{63}$$

$$= 0.9468$$

Consider department attribute:

$$\text{Info}_{dept}(D_1) = \frac{30}{63} \times \text{Info}(\text{sales}) + \frac{23}{63}\text{Info}(\text{Systems}) + \frac{10}{63} \times \text{Info}(\text{Marketing})$$

$$= \frac{30}{63}\left(-\frac{30}{30}\log_2\frac{30}{30}\right) + \frac{23}{63}\left(-\frac{23}{23}\log_2\left(\frac{23}{23}\right)\right) + \frac{10}{63}\left(-\frac{10}{10}\log_2\frac{10}{10}\right)$$

$$\text{Gain}(dep) = 0.9468$$

Consider (age) attribute:

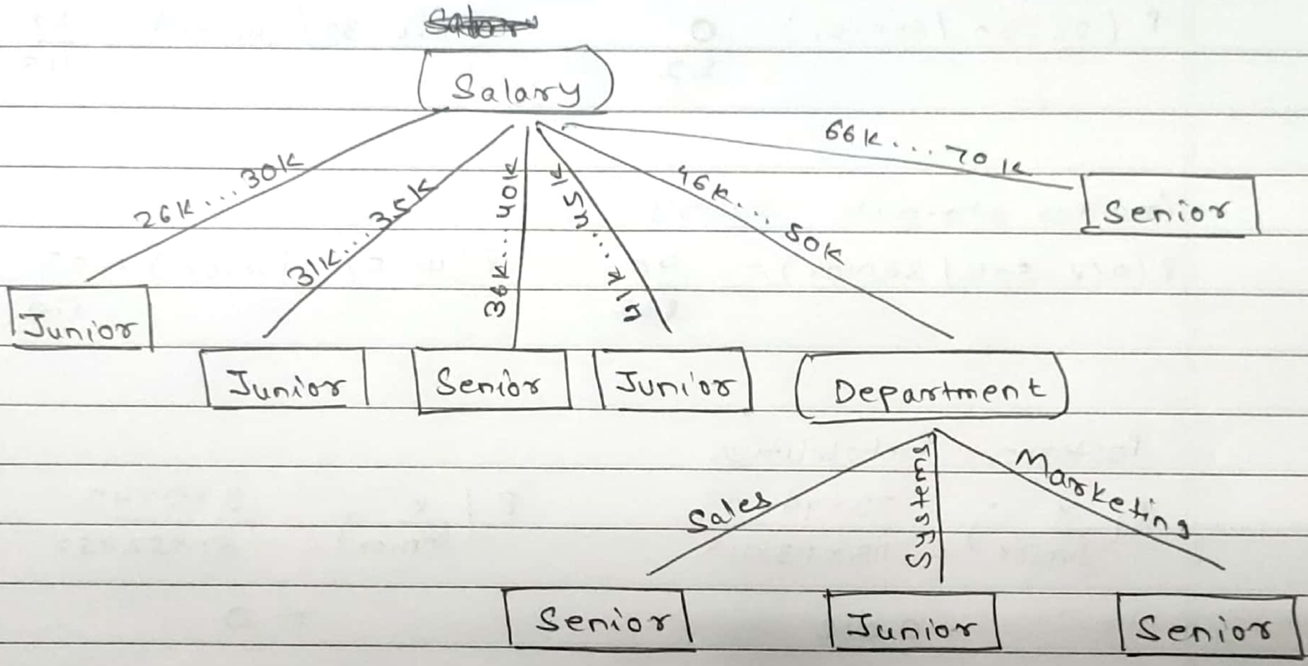
$$\text{Info}_{age}(D_1) = \frac{20}{63}\text{Info}(21\text{-}25) + \frac{3}{63}\text{Info}(26\text{-}30) + \frac{30}{63}\text{Info}(31\text{-}25)$$
$$+ \frac{10}{63}\text{Info}(36\text{-}40)$$

$$= \frac{20}{63}\left[-\frac{20}{20}\log_2\frac{20}{20}\right] + \frac{3}{63}\left[-\frac{3}{3}\log\frac{3}{3}\right] + \frac{30}{63}\left[-\frac{30}{30}\log_2\frac{30}{30}\right] + \frac{10}{63}\left[-\frac{10}{10}\log_2\frac{10}{10}\right]$$

$$\text{Gain}(age) = 0.9468$$

∴ Gain for both attributes are same. We can choose any for splitting.

∴ Decision Tree :



Above tree is the final decision tree from Information Gain (ID3) algorithm)

(c) Prior Probabilities :

$P(\text{Senior}) = \frac{52}{165}$ $\quad$ $P(\text{Junior}) = \frac{113}{165}$

Consider attribute department :

$P(\text{System} / \text{senior}) = 8/52$

$P(\text{System} / \text{junior}) = 23/113$

Consider attribute age

$P(26-30 / senior) = \frac{0}{52}$ $P(26-30 / junior) = \frac{49}{113}$

Consider attribute salary

$P(46k-50k / senior) = \frac{40}{52}$ $P(46-50k / junior) = \frac{23}{113}$

∴ Posterior Probabilities:

$P\left(\frac{X}{junior}\right) = \frac{23 \times 49 \times 23}{113 \times 113 \times 113}$ $P\left(\frac{x}{senior}\right) = \frac{8 \times 0 \times 40}{52 \times 52 \times 52}$

$= 0.018$ $= 0$

$\therefore P\left(\frac{junior}{x}\right) = \frac{113}{165} \times 0.018$ $P\left(\frac{senior}{x}\right) = 0 \times \frac{52}{165}$

$= 0.0123$ $= 0$

$\therefore 0.0123 > 0$

∴ X = (systems, 26:::30, 46:::50k) will be predicted as status = junior.

Q. 3

→ ① Data Based Handling:

(i) Over Sampling: In over sampling we increase the no. of minority class keeping the number of majority class till balance is achieved.

example: Intially gradulent : 10

Non- gradulent : 990 ^{00}

If we increase gradulent to 10 times the original, then gradulent = 100, non-gradulent = 990. We can increase number of gradulent cases till balance is achieved.

(ii) Under Sampling: In under sampling we decrease the no. of majority class members keeping the number of minority class members till balance is achieved.

example: Intially Fraud: 10, Non- Fraud : 990

If we decrease no. of non-fraud cases by 50%. then fraud: 10 and non-fraud: 495 ; we can decrease the number of non-fraud cases till balance is achieved.

(iii) Hybrid Sampling: In hybrid sampling we increase number of minority class members and decrease no. of majority class members till balance of data is achieved.

example: Fraud: 10, Non-Fraud: 990

we increase fraud cases by 10 times and decrease non fraud cases by 50% then fraud: 100, non-fraud: 495, we can do so till balance b/w data is achieved.

② Algorithmic Based:

(i) Ensemble learning: Imbalanced data can be handled modifying existing classification algorithms to make them appromiate for handling imbalanced data. Ensemble learning involves constructing several stage classifiers and aggregates their predictions, which gives higher performance than single classifiers.

BAGGING: It generates 'n' different training samples with replacement, and training the algorithm on each bootstrapped samples and then aggregating their predictions. This helps to balance the data and make predictions.

BOOSTING: It is a technique to combine weak learners to create a strong learner and make accurate predictions. At each iteration the new classifier places more weight on the cases which were predicted incorrect by previous classifier, thus giving more accurate predictions.

The above mentioned methods can be used to induce a quality classifier for the bank that wants to guard against gradulent transcations where number of frauds <<< number of non-frauds.

Q. 4)

→ RHS = sensitivity * $\frac{P}{P+N}$ + specificity * $\frac{N}{P+N}$

now, sensitivity = $\frac{TP}{P}$, specificity = $\frac{TN}{N}$

Substituting in RHS = $\frac{TP}{P} \times \frac{P}{P+N} + \frac{TN}{N} \times \frac{N}{P+N}$

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

$= T$ —— (i)

consider, LHS = accuracy = $\frac{TP+TN}{P+N}$

$= T$ —— (2)

From (i) and (2) LHS = RHS

∴ accuracy = sensitivity * $\frac{P}{P+N}$ + specificity * $\frac{N}{P+N}$

Q. 5)

→ **Bagging:** It is used when the goal is to reduce the variance of a decision tree classifier. Here the objective is to create several subsets of data from training sample chosen randomly from sample. Each collection of subset data is used to train their decision trees.

Bagging Steps:

- Suppose there are N observations and M features in training data set. A sample from training data set is taken randomly with replacement.

- A subset of M features are selected randomly and whichever features gives the best split is used to split the node iteratively.
- The tree is grown to the largest.
- Above steps are repeated n times and prediction is given based on the aggregation of predictions from n no. of trees.

Advantages:

- Reduces over-fitting on the model.
- Handles higher dimentionality data very well.
- Maintains accuracy for missing data.

Disadvantages:

- Since final prediction is based on the means prediction from subset trees, it won't give precise values for the classification and regression model.

BOOSTING: Boosting is used to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analyzing data for errors. Consecutive trees are fit and at every step, the goal is to improve the accuracy from the prior tree. When an input is misclassified by a hypothesis, its weight is increased so that the next hypothesis is more likely to classify it correctly.

Boosting Steps:

- Draw a random subset of training samples d1 without replacement from the training set D to train a weak learner C1
- Draw a second random training subset D2 without replacement from the training set and add 50% of the samples that were previously falsely classified to train a weak learner C2.
- Find the training sample d3 in the training set D on which C1 and C2 disagree to tain a weak learner C3.
- Combine all the weak learners via majority voting.

Advantages:

- Supports different loss functions.
- works well with interactions.

Disadvanteges:

- Prone to over-fitting.
- Requires carefu timing of different hyper-parameters.

Similarities:

- Both are ensemble methods to get N learners from 1 learner.
- Both generate several training data sets by random sampling. Both make the final decision by averaging the N learners.
- Both are good at reducing variance and provide ~~higher~~ higher stability.

Bagging	Boosting
1) The simplest way of combining predictions of same type.	1) A way of combining predictions that belong to different types.
2) Aims to decrease variance, not bias.	2) Aims to decrease bias, not variance.
3) Each model receives equal weight.	3) Models are weighted according to their performance.
4) Different training data sets are randomly taken w/o replacement from entire training dataset	4) Every new subset contains the element that were misclassified by previous models.
5) example: Random forest	5) example: Adaboost.