Aman Chauhan 180S158

1) a) <u>Capabilities</u> :- i) Compared to other algorithms decision trees requires less effort for data preparation during pre-processing.

ii) A decision tree does not require normalization of data.

iii) A decision tree does not require scaling of data as well.

iv) Missing values in the data also do not affect the process of building a decision tree to any considerable extent.

v) A decision tree model is very intuitive and easy to explain to technical teams as well as stakeholders.

<u>Disadvantages</u> :- i) A small change in the data can cause a large change in the structure of the decision tree causing instability.

ii) For a decision tree sometimes calculation can go far more complex compared to other algorithms.

iii) Decision tree often involves higher time to train the model.

iv) Decision tree training is relatively expensive as the complexity and time has taken are more.

v) The decision tree algorithm is inadequate for applying regression and predicting continuous values.

Aman Chauhan : 1805158

b) Now in this question we have total samples = 14 [8+, 6-]

$$Entropy(S) = -P_{\oplus} \log_2 P_{\oplus} - P_{\ominus} \log_2 P_{\ominus}$$
$$= -\frac{8}{14} \log_2 \frac{8}{14} - \frac{6}{14} \log_2 \frac{6}{14} = 0.99$$

Now, we have to find the gain to decide the root node.
Thus, we need to find entropy of each attribute.

$$Entropy(Fever_{yes}) = -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8} = 0.81$$

$$Entropy(Fever_{No}) = -\frac{2}{6} \log_2 \frac{2}{6} - \frac{4}{6} \log_2 \frac{4}{6} = 0.91$$

Now, $Gain(S, Fever) = E(S) - \sum_{v \in \{Yes, No\}} \frac{|S_v|}{|S|} Entropy(S_v)$

$$= 0.99 - \frac{8}{14} \times 0.81 - \frac{6}{14} \times 0.91 = 0.13$$

$$\boxed{Gain(S, Fever) = 0.13}$$

$$Entropy(Cough_{yes}) = -\frac{5}{10} \log_2 \frac{5}{10} - \frac{5}{10} \log_2 \frac{5}{10} = 0$$

$$Entropy(Cough_{No}) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.811$$

Now, $Gain(S, Cough) = E(S) - \sum_{v \in \{Yes, No\}} \frac{|S_v|}{|S|} Entropy(S_v)$

$$= 0.99 - \frac{10}{14} \times 0 - \frac{4}{14} \times 0.81 = 0.04$$

$$\boxed{Gain(S, Cough) = 0.04}$$

$$Entropy(BI_{yes}) = -\frac{7}{8} \log_2 \frac{7}{8} - \frac{1}{8} \log_2 \frac{1}{8} = 0.54$$

Aman Chauhan 1805158

$$\text{Entropy}(BI_{No}) = -\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6} = 0.64$$

$$\text{Now, Gain}(S, BI) = E(S) - \sum_{v \in \{Yes, No\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$= 0.99 - \frac{8}{14} \times 0.54 - \frac{6}{14} \times 0.64 = 0.40$$

$$\boxed{\text{Gain}(S, BI) = 0.40}$$

Since the feature Breathing Issue have the highest Gain, thus it is root node.

Now, we have to find which feature comes under Yes (left branch) or which comes under No (right branch). We will take subset of dataset with Yes values in Breathing Issues

Fever	Cough	Breathing Issue	Infected
Yes	Yes	Yes	Yes
Yes	No	Yes	Yes
Yes	Yes	Yes	Yes
Yes	No	Yes	Yes
Yes	No	Yes	Yes
No	Yes	Yes	Yes
No	Yes	Yes	Yes
No	Yes	Yes	No

$$E(S_a) = -\frac{7}{8}\log\frac{7}{8} - \frac{1}{8}\log\frac{1}{8} = 0.54$$

$$\text{Entropy}(\text{Fever}_{Yes}) = -\frac{5}{5}\log_2\frac{5}{5} - \frac{0}{5}\log_2\frac{0}{5} = 1$$

$$\text{Entropy}(\text{Fever}_{No}) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.90$$

$$\text{Now, Gain}(S_{ar}, \text{Fever}) = E(S_a) - \sum_{v \in \{Yes, No\}} \frac{|S_v|}{|S|} E(S_v)$$

Scanned with CamScanner

Aman Chauhan 1805158

$$= 0.54 - \frac{5}{8} \times 1 - \frac{3}{8} \times 0.90 = 0.20$$

$$\boxed{Gain(S_{BY}, Fever) = 0.20}$$

$$Entropy(Cough_{Yes}) = -\frac{4}{5} \log_2 \frac{4}{5} - \frac{1}{5} \log \frac{1}{5} = 0.72$$

$$Entropy(Cough_{No}) = \frac{3}{3} \log \frac{3}{3} - \frac{0}{3} \log \frac{0}{3} = 1$$

Now, $Gain(S_{BY}, Cough) = E(S_{BY}) - \sum_{v \in \{Yes, No\}} \frac{|S_v|}{|S|} E(S_v)$

$$= 0.54 - \frac{5}{8} \times 0.72 - \frac{3}{8} \times 1 = 0.09$$

$$\boxed{Gain(S_{BY}, Cough) = 0.09}$$

We can clearly seen that Gain of Fever is more, i.e. Fever is comes under Yes (left branch) of the root node and Cough will come under NO (right branch) respectively.

Now we have to find the leaf nodes. For that we will take subset of dataset with Fever & Breathing Issue value Yes.

Fever	Cough	Breathing Issues	Infected
Yes	Yes	Yes	Yes
Yes	NO	Yes	Yes
Yes	Yes	Yes	Yes
Yes	No	Yes	Yes
Yes	No	Yes	Yes

We can easily seen that all the values in the

Aman Chauhan 1805158.

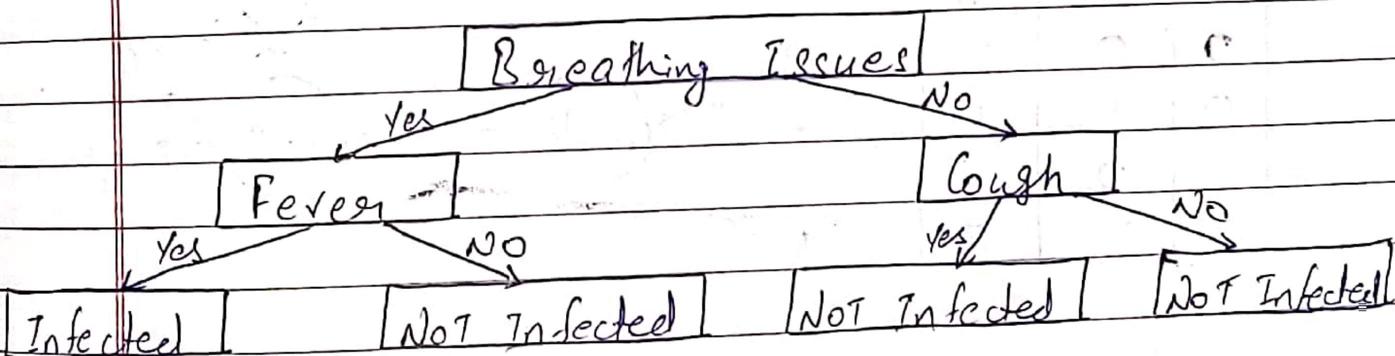
target column are Yes, ∴ we can say that left node of Fever is Infected and right node is Not Infected.

At last we only left with left and right node of Cough. for that take a subset of dataset with Breath Issue value No and Cough Yes.

Fever	Cough	Breathing Issue	Infected
Yes	Yes	No	No
No	Yes	No	No
Yes	Yes	No	Yes
No	Yes	No	No
Yes	Yes	No	No

Most values in target are No when Cough is Yes ∴ left Node is Not Infected.

When Cough value is No and Breathing Issue is also no, then also target is No ∴ Its both Nodes are Not Infected



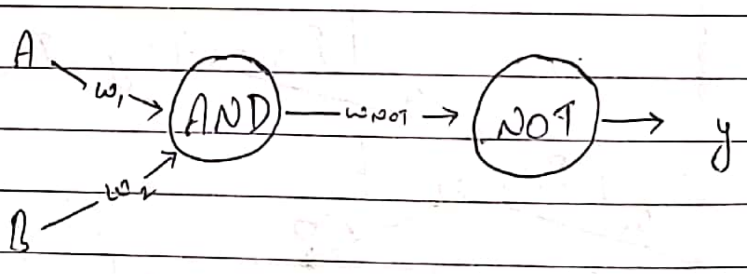
Aman Chauhan 1805158

(Q) a) Decision Tree: Decision trees have an easy to follow natural flow. Then are also easy to program for computer systems with If, then, Else statements. We can see that the top node in the tree is the most influential piece of data that affects the response variable in the model. Because these trees are so easy to understand, they are very useful as modeling techniques and provide visual representation of data.

Artificial Neural Networks: It is not so easy to understand from the visual representation. It is very difficult to create computer systems from them, and almost impossible to create an explanation from the model. Neural Networks can handle binary data better than decision trees but cannot handle categorical values.

b) Suppose the two inputs are A & B and weight associated with inputs are w_1 & w_2

A	B	y
0	0	1
0	1	1
1	0	1
1	1	0



Aman Chauhan 1805158

c)

	Standard Gradient Descent	Stochastic Gradient Descent
i)	It computes gradient using the whole training sample.	i) It computes gradient using a single training sample.
ii)	Slow and computationally expensive algorithm.	ii) Faster and less computatio expensive than Batch GD.
iii)	No suggested for huge training samples.	iii) Can be used for large training samples.
iv)	Deterministic in Nature.	iv) Stochastic in Nature.
v)	Gives optimal solution given sufficient time to converge.	v) Gives good solution but not ~~go~~optimal.
vi)	No random shuffling of points are required.	vi) The data sample should be in a random order and this is why we want to shuffle the training set for every epoch.
vii)	Cant escape shallow local minima easily	vii) ~~St~~ It can escape shallow local minima easily.
viii)	Convergence is slow.	viii) Reaches the convergence much faster

d) Adaline which stands for adaptive Linear Neuron, is a network having a single linear unit. Some important points are-

- It uses bipolar activation function.
- It uses delta rule for training to minimize the Mean-Squared Error between the actual output and the desired/target output.

Aman Chauhan 1805158

- The weights and the bias are adjustable.

Architecture: The basic structure of Adaline is similar to perceptron having an extra feedback loop with the help of which the actual output is compared with the desired/target output. After comparison on the basis of training algorithm, the weights and bias will be updated.