(Assignment-1) (Neha Verma) · (AIT).

Ans 1). a).
Ans 2). c).
Ans 3). d).
Ans 4). b).
Ans 5). c).
Ans 6). c).
Ans 7). b).
Ans 8). d).
Ans 9). c).
Ans 10). c).

Q. Exp. overfitting in Decision tree using Depth as parameter?

A:- The reason behind overfitting in Decision tree due to the increasing ↑ no. of tree. We can solve it by pruning, Hyperparameter turning, max depth limit.

Then exp. how bagging & Random forest address this problem?

→ In Bagging, we train on multiple models & on different samples of Data.

→ In Random forest, we usually combine multiple Decision tree in one random forest to get more accurate result.

→ Due to this, the chances of overfitting will reduced.

Q:- Exp. Random forest working in Details?

A:- (a) Bootstrap Sampling.
(b) Random feature selection
(c) Majority voting

DOMS

A:- Basically, Random forest is that which use multiple decision tree / combine multiple Decision tree to get accurate result.
Due to this, the chances of overfitting will reduces.
→ Random forest uses bootstrap Sampling & Aggregation in which, we have some kind of Data, Divide random forest to Multiple Decision tree, the step before dividing tree is called Bootstrap & After assinging the value means we got 0.1, 0.9 anything that lead to aggregation / Sampling.
→ Majority Voting is used in classification in which we assume two trees are giving us 1 or one tree giving us 0 val. So, there will be majority voting means we got 1 as va value.

Q:- A fraud Detection model produced the following results :-

	Predicted fraud	not Predicted fraud.
Actual fraud	120	30
not fraud	50	800.

(a) for calculating Accuracy we use → Accuracy_score.

```
from sklearn.metrics import accuracy_score
print("\nAccuracy Score\n", accuracy_score(y_test, y_pred))
```

(b) for calculating precision, recall & f1 score. → we use Confusion matrix.

```
from sklearn.metrics import accuracy_score, confusion_matrix
```

TP = 120 FP = 50
FN = 30 TN = 800
Total = 1000

Accuracy → $\frac{TP+TN}{Total} \Rightarrow \frac{120+800}{1000} = \frac{920}{1000} = 0.92$

Accuracy → 92%.

Precision → $\frac{TP}{TP+FP} = \frac{120}{120+30} = \frac{120}{170} = 0.705$

Precision → 70%.

Recall → $\frac{TP}{TP+FN} = \frac{120}{120+30} = 0.80$

Recall → 80%.

F1score → $\frac{2 \times Precision \times recall}{Precision + recall} = \frac{2 \times 0.705 \times 0.80}{0.705 + 0.80} = \frac{1.128}{1.505}$

fscore = 0.75

F1score → 75%.

⇒ Model Acceptable Because
High Recall (80%) → most fraud cases.
Better Accuracy we got (92%).

Q:- What happen if max_ feature none?
A:- All features considered at every split..

Q:- Will this model overfit more or less than single Decision tree? Why?

DOMS

(b) This model will overfit less than single Decision tree because we combine multiple Decision tree over there because of that we got good Accuracy. But in single Decision tree, model will definately overfit because the max depth of tree is reached.

→ we can use Bootstrap Sampling.
→ Majority voting.
→ Random feature selection.

(C). Increase no. of tree if :-

n_estimators → no. of tree(200).

Bias → Slightly Increases chances.

Variance → will be Decrease.

Because in Random forest we have :-

{ low Bias, low variance } required.