day=13 **Assessment -1**

Answers

(1) a) Large K with noisy data

(2) (c) They overfit

(3) (c) Reduce variance

(4) c) All features are considered at each split

(5)(c) Features are independent

(6)(c) Sigmoid.

(7) (d) $F_1$-score

(8) (d) Overfitting

9) a) to reduce bias

10) c) Logistic Regression

11) Overfitting in Decision Trees :

→ Overfitting is a condition when a model start cramming a ~~data~~ training data rather than learning patterns, logic and relationship b/w target & features. and can't perform well on unseen data.

→ Decision trees are unstable learners because in most cases and depend upon data they prone to overfitting means low bias but high variance

→ generally, ~~depth~~ max-depth is a hyperparameter in decision trees which decides that how deep the splits are gone, if we select the high value for it the model can be prone to overfitting because all the focus will biased to splitting maximum as possible and which cause overtraining on training data cause overfitting

→ solution, use bagging techique → Random Forest
It solves the problem of overfitting (low bias high variance) to (low brais, low variance) by traning n-estimators → no. of decision trees parallely on n-samples of data by using technique called as bootstraping which ~~split~~ send sample of the original training data by replacing rows and features depend on us.

→ and n-estimators learns on diff samples of traning data and produce outputs

→ and we use technique like voting for classificat and average for numerical outputs of diff trees and agregate their result and produce best model from no. of week learners.
that's how bagging resolve the decision tree problem

Ans:12) Random Forest Working:

Random Forest is ~~a~~ Ensemble learing bagging technique, which train no. of decision tree on samples of data called weak learners and ~~using~~ aggregate their outputs to produce best model.

→ Bootstrap sampling: It is a sampling data techniques used by the random forest to train no. of decision trees on diff samples of same training data.
it uses row sampling and column samping to produce a data in a way that not same data will repeat for all decision trees

like with row replacement → some rows can be repeat in single sample
similarly about column replacement

Eg: Row id

A
B
C
D
E

} → Bootstrapping → Sample(A)
A, B C, C, E

some rows can be repeat or some can never be included ≈ 68% Rows are selected and 32% can be taken as for unseen data generally.

or without replacement → If a row is included in a bag it will considered out of bag means can't repeat now.

→ Random Feature selection: Similarly some features can be included ~~can~~ or can't be included for different samples, features are selected randomly

→ Majority Voting: It is technique when we have classification problem output select that one for which most of the models produce maximum same output

Max Voting

Eg:

Decision Tree 1 —output→ class 0

DT 2 → Class 1 ✓    1 = 4

DT 3 → Class 1 ✓    0 = 2

DT 4 → Class 1 ✓

DT 5 → Class 0

DT 6 → Class 1 ✓

∴ Best Model will be output → Class 1 selected as max weak learns produce class 1

Ans → 13

→ (a) Accuracy : $\frac{T.P + T.N}{T.P + T.N + F.N + F.P} = \frac{120 + 800}{120 + 800 + 50 + 30}$

$= \frac{92\cancel{0}}{100\cancel{0}} = 0.92$

b) Precision : $\frac{T.P}{T.P + F.P} = \frac{120}{120 + 50} = \frac{12\cancel{0}}{17\cancel{0}} = 0.70$

c) Recall : $\frac{T.P}{T.P + F.N} = \frac{120}{120 + 30} = \frac{120}{15\cancel{0}} = 0.8$

(d) $F_1$-score → $\frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * 0.70 * 0.8}{0.70 + 0.8}$

$= \frac{1.12}{1.5} = 0.74$

Actually ~~these~~ accuracy about 92 %, but ac to confusion matrix actually 150 are actually fraud out of 120 are positively classified as fraud by model, 30 are miss classified,

So model can be considered but can be improved for fraud detection.

Ans → 14

(i) if max-feature = None. → all features are considered at each split

(ii) model will overfit more

iii) If we choose n-estimators = 200, the model will perform well on traing data and reduce overfitting which mean, [illegible]

it will accept low variance and low bias → bias,
variance trade-off.