













A: The key assumption behind the naive Bayes algorithm is that all features are independent of each other given in the class label i.e., feature independence.





Answer :

The types of Naive Bayes classifiers are:

1. Gaussian Naive Bayes – assumes features follow a normal distribution.
2. Multinomial Naive Bayes – used for discrete counts e.g., text classification.
3. Bernoulli Naive Bayes – for binary/Boolean features.







Answer :

Naive Bayes is considered as a good choice for high-dimensional datasets because of its computationally efficient and fast, requires less training data, and handles high-dimensional input well due to its simplicity and independence assumption.









Answer :

The main purpose of using k-fold cross-validation is to assess the model’s performance more reliably by reducing variance due to a single train-test split.







Answer :

Instead of one split, data is split into k parts and the model is trained and validated k times, improving robustness in k-fold cross-validation.







Answer :

The advantage of using stratified k-fold cross-validation is that it ensures that each fold has a similar class distribution as the overall dataset, especially important for imbalanced datasets.





Answer :

K is the number of folds.

* Larger k → lower bias, higher variance, longer training.
* Smaller k → higher bias, lower variance, faster training.







Answer :

k-fold cross-validation help prevent overfitting by testing the model on multiple subsets, it helps ensure the model generalizes well rather than just performing well on one split.







Answer :

k-fold cross-validation provides a reliable estimate of performance, helping choose the best model.





Answer :

Disadvantage of k-fold cross-validation is computationally expensive because the model is trained k times.









Answer :

Confusion matrix summarizes the performance of a classification model by showing actual vs predicted classifications.







Answer :

* True positives (TP) : Correctly predicted positives.
* True negatives (TN): Correctly predicted negatives.







Answer :

* False positives (FP): Incorrectly predicted as positive (actually negative).
* False negatives (FN): Incorrectly predicted as negative (actually positive).





Answer :

Accuracy = (TP + TN) / (TP + TN + FP + FN)







Answer :

Precision = TP / (TP + FP)

It measures how many predicted positives are actually correct.







Answer :

Recall = TP / (TP + FN)

It measures how many actual positives were correctly predicted.







Answer :

F1 = 2 \* (Precision \* Recall) / (Precision + Recall)

It balances precision and recall.







Answer :

High False Positive rate indicates that the model is predicting too many positives, increasing the risk of incorrectly labeling negatives.







Answer :

Low False Negative rate indicates that the model is successfully identifying most actual positives, which is important in sensitive domains (like disease detection).







Answer :

Accuracy may not be a good metric for an imbalanced dataset because it can be misleading—a model predicting only the majority class may still achieve high accuracy despite poor performance on the minority class.

