













Ans : The Key assumption is that all feature are independent of each other which simplifies the calculation.





Ans : 1) Gaussian Naïve Bayes – Assumes features follow a normal distribution

2) Multinomial Naïve bayes – Used for discrete counts

3) Bernoulli Naïve Bayes – used for binary features







Ans: Naïve bayes is computationally efficient and performs well with high dimensional data because it doesn’t require feature interactions and less prone to overfitting due to its simplicity









Ans: To evaluate the model performance more reliably by reducing variance due to specific train test split and to make better use of limited data.







Ans: In traditional train test split the data is divided into training and testing sets . In k fold cross validation the data is split into k subsets and the model is trained an evaluated k times each time using a different fold as the test set and the rest for training.







Ans : It ensures that each fold has appropriately the same distribution of target classes as the original dataset, which is especially important for imbalance data.







Ans : K represents the number of folds or splits .

* Increasing k generally leads to lower bias and more reliable evaluation but increases computation
* Decreasing k is faster but may increase variance and reduce reliability .







Ans : 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.







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





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









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







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







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





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







Ans: Precision = TP / (TP + FP)

It measures how many predicted positives are actually correct.







Ans : Recall = TP / (TP + FN)

It measures how many actual positives were correctly predicted.







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

It balances precision and recall.







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







Ans: Low False Negative rate indicates that the model is successfully identifying most actual positives, which is important in sensitive domains .







Ans: 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.

