2. What type of learning model is KNN?

KNN is a supervised machine learning model, where the model is trained using labeled data to make predictions on new, unseen data.

3. How does KNN classify a new data point?

KNN classifies a new data point by finding the 'k' closest data points to it in the dataset and assigning the majority class among them to the new point.

4. What is the role of the 'n_neighbors' parameter in KNN?

The 'n_neighbors' parameter defines the number of nearest neighbors to consider for classifying a new data point.

5. What library is used in Python to implement KNN in this assignment?

Scikit-learn, a popular machine learning library in Python, is used to implement KNN.

6. How is the dataset split for model training and testing?

The dataset is split into training and testing sets using the `train_test_split` function from scikit-learn, with 80% for training and 20% for testing.

7. What does setting 'random_state' in `train_test_split` ensure?

It ensures reproducibility by generating the same train-test split each time the code runs.

8. Why do we use stratified splitting in `train_test_split`?

Stratified splitting maintains the proportion of each class (diabetes vs. no diabetes) in both training and test sets, improving the model's generalizability.

12. How is 5-fold cross-validation implemented in this assignment?

By using `cross_val_score` from scikit-learn with `cv=5`, the dataset is split into 5 groups, and the model is trained and tested on each, providing an average accuracy score.

13. Why is cross-validation preferred over a single train-test split?

Cross-validation is more robust as it tests the model on multiple subsets, giving a better estimate of its performance on unseen data.

14. What is hypertuning, and how is it applied to KNN in this assignment?

Hypertuning is the process of optimizing model parameters. Here, `GridSearchCV` is used to test different values of 'n_neighbors' to find the best-performing value for KNN.

15. What does `GridSearchCV` do in KNN tuning?

`GridSearchCV` iterates over a range of values for ‘n_neighbors’ and selects the value that yields the highest accuracy based on cross-validation.

16. What was the optimal ‘n_neighbors’ value found through `GridSearchCV`?

The optimal ‘n_neighbors’ value determined in this assignment is 14.

17. How did hypertuning improve the model's accuracy?

Hypertuning improved model accuracy by over 4% by selecting the best value for ’n_neighbors’.

18. What is the ROC-AUC curve, and why is it important?

The ROC-AUC curve shows the trade-off between sensitivity and specificity for different thresholds. The area under the curve (AUC) indicates the model's ability to distinguish between classes.

19. What does an ROC-AUC score of 1 represent?

An ROC-AUC score of 1 represents perfect classification, where the model correctly classifies all instances.

21. What are the predictor variables used in this diabetes dataset?

Predictor variables include the number of pregnancies, BMI, insulin level, age, glucose level, diastolic blood pressure, and skin thickness.

22. What is the purpose of the ‘Outcome’ variable in the diabetes dataset?

The ‘Outcome’ variable indicates whether a patient has diabetes (1) or not (0), serving as the target variable for classification.

23. How is the target variable separated from the features in this dataset?

The target variable (diabetes) is separated by using `drop` to remove it from the features (X), while it is stored in a separate variable (y).

25. What is the impact of choosing a high value for ‘k’ in KNN?

A high value for ‘k’ can make the model more generalized but may also reduce sensitivity to local patterns, leading to a loss of accuracy.

28. Why might KNN be chosen over SVM for this diabetes prediction task?

KNN is simpler to implement, requires less tuning in basic cases, and is effective for small to medium-sized datasets with clear class separation.

30. What conclusion was drawn in this assignment about the KNN model?

The assignment concluded that KNN could successfully predict diabetes, achieving an accuracy improvement with cross-validation and hypertuning, although the model still had room for further optimization.