Name: Deep Salunkhe
Roll No: 2110LA00016
Sub: ML

D D M M Y Y Y Y

Assignment 5

Q1] You are tasked with developing a predictive model for patient outcomes in a healthcare setting. Explain the importance of k-fold cross-validation in evaluating the performance of your model. Demonstrate how you would implement k-fold cross validation and interpret the results

⇒ **Importance of k-fold cross-validation**

K-fold cross-validation is a crucial technique for evaluating the performance of an predictive model due to several reason.

- Preventing Overfitting : By splitting the data into multiple folds we ensure the model is evaluated on data it hasn't seen.
- Estimatising Generalization Error: It provides more reliable estimate of the model's performance on unseen data
- Hyperparameter tunning : It can be used to optimize model hyperparameter
- Robustness.

Implementation of k-Fold cross validation

- Data preparation : Collect and preprocess patient data, including relevant features and target variable
- Handle missing values, outliers and feature scaling as need

- splitting Data : Randomly divide the data into k equal sized folds

- cross - validation - loop:

for each fold :
  - Use k-1 folds as the training set to build the model
  - Use the remaining fold as the test set to evaluate the model's performance.
  - Calculate performance matrix

- performance Evaluation

Calculate the average performance metrix accross all folds to obtain a reliable estimate of the model's performance.

Q2] Compare the performance of a basic decision tree model with an XGBoost model and how it improves over a single decision tree. Evaluate the model using appropriate performance metrics

⇒ Decision tree and XGBoost are both popular machine learning algorithms for classification tasks like employee prediction. However, XGBoost is generally considered more powerful due to its ensemble-based approach.

XGBoost improves upon a single decision tree by:

- Ensemble learning: Combining multiple decision tree to reduce variance and improve accuracy.

- Gradient Boosting: Sequentially building trees, with each tree correcting the errors of previous trees.

- Regularization: Preventing overfitting through technique like L1 & L2 regularization.

- Handling missing values: Built-in mechanism for handling missing data

Performance Evaluation metrics

Accuracy : proportion of correctly predicted attrition cases
Precision : proportion of predicted attrition cases that are actually true.
Recall : proportion of actual attrition cases that are correctly predicted.
F1-score : Harmonic mean of precision and recall.
AUC-ROC : Area under the receiver operating characteristic curve, meaning the model's ability to distinguish between positive and negative classes.