**Predict:**

**Why choose these features?**

**Why choose these models?**

**How do MSE and MAE help you to evaluate?**

**Classify students:**

**Why choose these features? (Needs improvement)**

Sum1: The total number of previous attempts at quiz and homework.

Sum2: The total sum of the highest scores in previous quiz and homework.

Score: Final score of course

The choice of these features suggests that the clustering is based on the cumulative performance across quizzes, homework, and final scores, providing a holistic view of the students' academic performance.

**Why choose these classifiers?**

KMeans Clustering:

The code applies KMeans clustering with different values of k (number of clusters) and uses the elbow method to determine the optimal value of k. KMeans is a popular centroid-based clustering algorithm that partitions the dataset into k clusters.

KMeans is chosen for its simplicity and efficiency, and it is widely used for clustering tasks.

Hierarchical Clustering:

The code uses hierarchical clustering with the Ward linkage method. Hierarchical clustering builds a hierarchy of clusters and is capable of capturing the nested structure in the data.

Ward linkage is known for its ability to minimize variance within clusters and is suitable for a variety of datasets.

DBSCAN:

DBSCAN is a density-based clustering algorithm that groups together data points that are close to each other and marks outliers as noise.

It does not require specifying the number of clusters beforehand and is suitable for datasets with irregular shapes and varying cluster densities.

The choice of these clustering algorithms suggests an exploration of different clustering strategies to understand the inherent structure of the data and identify groups of students with similar performance patterns. Each algorithm has its strengths and weaknesses, and using multiple methods helps in gaining a more comprehensive understanding of the dataset.

**How do you categorize students into four types?**

I**n Elbow Method plot, what is the meaning of inertia?**

In the context of the Elbow Method plot used for clustering analysis, inertia (also known as within-cluster sum of squares) is a measure of how compact the clusters are. It quantifies the sum of squared distances between each data point in a cluster and the centroid of that cluster. The inertia is calculated separately for each cluster, and the overall inertia for a clustering solution is the sum of these values across all clusters.

In the Elbow Method, inertia is plotted against the number of clusters (K). The idea is to observe how the inertia changes as the number of clusters increases. The Elbow Method is a heuristic approach to finding the optimal number of clusters for a dataset. The "elbow" in the plot is the point where adding more clusters does not significantly reduce the inertia.

**What is meaning of distance and cluster size in hierarchical clustering dendrogram?**

**What is the Silhouette Coefficient and how to use it to evaluate the model?**

The Silhouette Coefficient is a metric used to calculate the goodness of a clustering technique. It provides a measure of how well-separated the clusters are in a given dataset. The Silhouette Coefficient ranges from -1 to 1, where a high value indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters.

Near +1: Indicates that the sample is far away from the neighboring clusters. This is a good indication of a well-defined cluster.

Near 0: Indicates that the sample is on or very close to the decision boundary between two neighboring clusters.

Near -1: Indicates that those samples might have been assigned to the wrong cluster.

**What is the meaning of PCA component 1 and component 2 in K-Means clustering plot?**