

# K-Means & Naïve-Based Algorithms

**Group Members:** Muhammad Ali (2022-CS-724)

Muhammad Talha (2022-CS-630)

**Supervisor:** Ma’am Namra

**Course Name:**  Artificial Intelligence

**Course Code:** CS-371

# University of Engineering and Technology, Lahore

**AI Project Documentation**

**1. Introduction**

This project leverages machine learning techniques to analyze educational data and extract meaningful insights. It focuses on clustering and classification to identify patterns and predict outcomes within student-related datasets. The aim is to provide data-driven recommendations to improve educational outcomes and resource allocation.

**2. Objectives**

1. Analyze educational data to identify trends and patterns.
2. Perform clustering to group students with similar characteristics.
3. Use classification models to predict student outcomes based on input features.
4. Provide actionable insights for stakeholders in the education sector.

**3. Features**

**3.1 Data Preprocessing**

1. Handle missing data and inconsistencies.
2. Standardize and normalize data for better model performance.
3. Encode categorical variables for compatibility with machine learning algorithms.

**3.2 Clustering**

1. Apply K-Means clustering to segment students based on their performance, demographics, and other features.
2. Visualize clusters to aid understanding of group characteristics.

**3.3 Classification**

1. Use the Naive Bayes algorithm to predict outcomes such as graduation likelihood, performance tiers, or dropout risk.
2. Evaluate models using metrics like accuracy, precision, recall, and F1-score.

**3.4 Reporting and Visualization**

1. Generate detailed reports on data insights and model performance.
2. Provide visualizations, such as scatter plots and confusion matrices, to illustrate findings.

**4. Data Overview**

**4.1 Dataset Details**

* 1. **Source**:

The dataset contains fields such as marital status, admission grades, attendance, curricular activities, and financial details.

* 1. Key Features
  2. Admission grade
  3. Attendance
  4. Curricular units (enrolled, approved, grades)
  5. Factors (e.g., parent qualifications, occupation)

**4.2 Preprocessing Steps**

1. Handle missing values by imputing with mean/median or dropping rows/columns as necessary.
2. Standardize numerical features using z-score normalization.
3. Encode categorical variables using label encoding or one-hot encoding.

**5. Machine Learning Workflow**

**5.1 Clustering (K-Means)**

1. Objective: Group students with similar profiles to identify potential areas for intervention or support.
2. Steps:
   * 1. Determine optimal cluster count using the elbow method.
     2. Fit the K-Means model on preprocessed data.
     3. Visualize clusters using scatter plots.

**5.2 Classification (Naive Bayes)**

1. Objective: Predict student outcomes based on historical data.
2. Steps:
   * 1. Split data into training and testing sets (e.g., 80-20 split).
     2. Train the Gaussian Naive Bayes model on the training data.
     3. Evaluate the model using testing data and metrics like accuracy, confusion matrix, and classification report.

**6. Technology Stack**

1. **Programming Language:**
   1. Python
2. **Libraries**
3. **Data Manipulation:**
   1. Pandas
   2. NumPy
4. **Machine Learning:**
   1. scikit-learn
5. **Visualization**:
   1. Matplotlib
   2. Seaborn

**7. Results and Insights**

1. Clustering revealed distinct student groups with unique characteristics, such as high performers, at-risk students, and average performers.
2. Classification achieved an accuracy of **60%** in predicting student outcomes.
3. Visualizations highlighted key trends, such as the impact of socioeconomic factors on performance.

**8. Conclusion**

The AI-based Educational Data Analysis project provides valuable insights into student performance and behavior. By leveraging clustering and classification, it enables stakeholders to make informed decisions aimed at improving educational outcomes.