CS 4412 - Data Mining Project Proposal

Pattern Discovery in Student Academic Performance Data

By Cesar Arevalo Colocho
Computer Science
Kennesaw State University
careval3@students.kennesaw.edu

***Abstract*—This project applies data mining techniques to explore patterns in a student performance dataset. The analysis focuses on discovering natural groupings of students and frequent associations among academic, social, and lifestyle attributes. Clustering and association rule mining are used to identify meaningful structures in the data, emphasizing pattern discovery and interpretation rather than outcome prediction.**

I. Dataset Description

Dataset Name: Student Performance Dataset
Source: UCI Machine Learning Repository
URL: https://archive-beta.ics.uci.edu/dataset/320/student+performance
Size: 395 instances, 33 attributes

Description: The Student Performance dataset contains demographic, academic, social, and lifestyle information about secondary school students enrolled in two distinct subjects: Mathematics and Portuguese language. The data was collected from two Portuguese schools and captures a wide range of factors including family background, study habits, school engagement, alcohol consumption, and academic outcomes. Academic performance is represented through three grade attributes rather than a single target label, making the dataset suitable for exploratory pattern discovery.

Key Attributes:

- **Demographic:** age, sex, family size, parental education
- **Academic behavior:** study time, absences, failures, school support
- **Social and lifestyle:** free time, alcohol consumption, internet access
- **Performance indicators:** G1, G2, G3

Data Quality Considerations:
The dataset contains no missing values but includes a mix of categorical and numerical attributes that require preprocessing. The structure of the dataset shows imbalances in some features, requires transformation of categorical data, and includes repeated information within grade-related attributes.

II. Discovery Questions

The following questions are focused on pattern discovery rather than prediction:

Question 1 - Do family and socioeconomic attributes form recognizable patterns linked to academic consistency?

This question is valuable because it helps to uncover social structured effects on student performance, where we can discover the relationship such as between a *low parental education vs. higher performance*, or *stable family relationship vs. consistent grades.*

Question 2 - How do study habits and absences interact to shape academic performance patterns?

This question is valuable because it helps to explore interaction effects that can reveal behavioral patterns that can discover students groups such as *high study time vs. high absences* or vise versa, etc

Question 3 - What natural clusters of students emerge when combining demographic, family, and academic features?

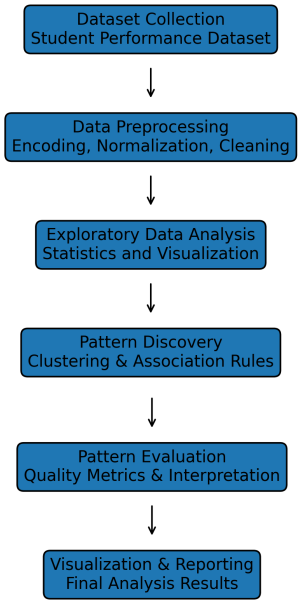
This question is valuable because it aims to identify latent student profiles using clustering techniques, which can help to uncover student profiles without predefined labels

III. Planned Techniques

I intend to use the following two data mining techniques:

- **Clustering (K-Means, Hierarchical Clustering)**

  Clustering techniques will be used to identify natural groupings of students based on academic behavior, lifestyle factors, and social attributes. By grouping students with similar characteristics, clustering supports the discovery of latent student profiles without relying on predefined labels. These techniques directly address the first discovery question by revealing whether distinct behavioral or engagement-based groups exist within the



dataset.

Therefore, K-Means clustering will be applied to numerical attributes after normalization, while hierarchical clustering will be used to explore cluster relationships and validate group structure

- **Association Rule Mining (Apriori, FP-Growth)**

  Association rule mining will be employed to discover frequent co-occurring attributes among students, such as combinations of study habits, family background, and lifestyle factors. This technique addresses the second discovery question by identifying common patterns and relationships between categorical attributes. Measures such as support, confidence, and lift will be used to evaluate the significance of discovered rules.

  For instance, FP-Growth may be applied as an alternative to Apriori to improve efficiency when mining larger itemsets.

IV. Preliminary Timeline

This is my rough plan for M2, M3 and M4: