Visual Analysis of Student Performance and Social Factors

VA 2025 Project Proposal

Leonardo Ricca
Sapienza University of Rome
Matricola: 2211129
ricca.2211129@studenti.uniroma1.it

Federico Turrini
Sapienza University of Rome
Matricola: 2175431
turrini.2175431@studenti.uniroma1.it

I. Dataset & Context

The project focuses on the **"Student Performance Data Set"**, retrieved from the UCI Machine Learning Repository.

A. Context

The data approaches student achievement in secondary education of two Portuguese schools. It includes student grades, demographic, social, and school-related features. This domain is critical for educational data mining, as predicting student failure is a key task for improving the efficiency of educational systems.

B. Characteristics

The dataset contains approximately **649 instances** (rows) and **33 attributes** (columns).

- **Target variables:** G1, G2, G3 (grades for first/second period and final grade).
- **Features:** Alcohol consumption (weekend/workday), absences, study time, internet access, parents' jobs, etc.

The dimensions respect the "AS Index" rule required for the exam ($N \times M \approx 21,417$, which falls within the 10,000–50,000 range).

II. Intended User

The system is explicitly designed for **School Counselors** and **Educational Policymakers**. Currently, these actors often rely on simple grade sheets to evaluate students. Our tool aims to provide a holistic view, enabling them to:

- Identify "at-risk" students not just by grades, but by risk factors (e.g., high alcohol consumption combined with low study time).
- Plan timely interventions before the final exam failure occurs.

III. Proposed Solution

The application will be a web-based dashboard featuring **Coordinated Views**: user interactions on one graph will automatically filter and update the data displayed in all other views.

A. Visual Part

- **Main View (Scatter Plot):** Maps the Final Grade (G3) on the Y-axis and the Number of Absences on the X-axis. This immediately highlights outliers (e.g., students with many absences but high grades).
- **Multidimensional View (Parallel Coordinates):** Displays the complex profile of selected students across 5-6 dimensions (Age, StudyTime, Health, Walc, Dalc).
- **Context View (Bar Charts):** Aggregated view for categorical social factors (Internet Access, Romantic Relationships, Family Status).

B. Analytics Part

The system goes beyond simple filtering.

- **Dimensionality Reduction (Mandatory):** A Principal Component Analysis (PCA) will project the 33-dimensional student profile into a 2D space, allowing users to detect clusters of students with similar behaviors.
- **On-demand Statistics:** Upon selecting a subset of students, the system dynamically computes the *mean deviation* of the selected group compared to the general population.

IV. The Visual Analytics Cycle

1) **Data Stage:** Raw CSV data is pre-processed. Categorical values (e.g., "yes/no") are encoded into numerical values for the PCA algorithm.
2) **Mapping & Visualization:** The processed data is mapped to visual variables (position, color) in the Scatter Plot and Parallel Coordinates.
3) **User Interaction:** The user filters the data by selecting specific ranges of grades or absences directly on the charts.
4) **Models & Analytics:** The interaction triggers the recalculation of the PCA projection and the statistical summary for the selected subset only.
5) **Knowledge:** The user gains insight (e.g., "Students with high weekend alcohol consumption tend to fail regardless of study time") and decides on an intervention.

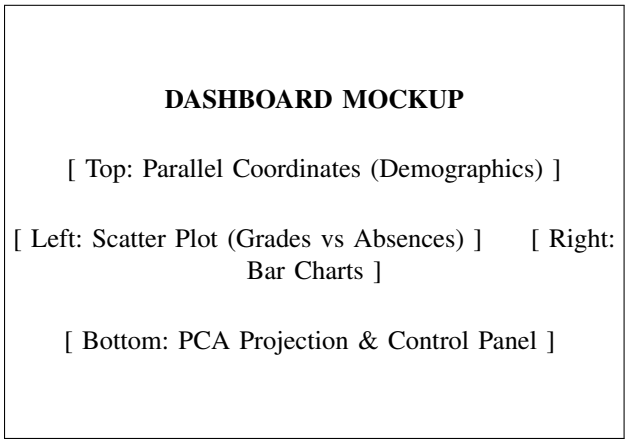

Fig. 1. Draft of the proposed UI. The top view handles multidimensional filtering, while the scatter plot highlights performance correlation.

V. Mockup of the User Interface

The interface (Fig. 1) is divided into three areas to facilitate the exploration flow. The top section allows for broad demographic filtering, while the central area focuses on the specific academic performance correlation.