Visual Analysis of Student Performance and Social Factors

Visual Analytics Course Project - Fall 2025

Sapienza University of Rome

Leonardo Ricca (ID: 2211129)
Federico Turrini (ID: 2175431)

Contents

CONTENTS

Visual Analysis of Student Performance and Social Factors

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

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

I. Introduction

The academic success of secondary school students is a multifaceted phenomenon, influenced not only by school-related factors but also significantly by demographic, social, and family backgrounds. Predicting student performance and identifying those at risk of failure is a critical task for educational institutions. While grades provide a direct, quantitative measure of academic achievement, they are often "lagging indicators"—by the time a student fails a final exam, it is often too late for intervention. The root causes of underperformance—such as high alcohol consumption, lack of family support, or health issues—are latent variables that require early detection.

Educational Data Mining (EDM) has emerged as a discipline dedicated to extracting actionable knowledge from academic data. However, traditional EDM approaches often rely on black-box machine learning models to predict pass/fail outcomes. While accurate, these models often lack transparency and interpretability, failing to build trust with domain experts such as school counselors and educational policymakers. These stakeholders need more than a prediction; they need to understand the *why* and *how* behind the data to design effective intervention strategies.

Visual Analytics (VA) offers a powerful solution to problem. By combining automated analysis methods with interactive visualizations, VA enables human reasoning at scale. It allows experts to explore the data, test hypotheses (e.g., "Does high alcohol consumption affect grades more than study time?"), and discover unexpected patterns that a purely statistical model might miss.

In this project, we present a web-based Visual Analytics dashboard designed to explore the **Student Performance Data Set** from the UCI Machine Learning Repository. Our system integrates coordinated multiple views with advanced analytics to support three main tasks:

1) **Multidimensional Profiling:** Understanding the complex interplay of social and academic variables.
2) **Correlation Analysis:** Identifying dynamic relationships between lifestyle choices and academic outcomes.
3) **Cluster Identification:** Detecting groups of students with similar behaviors using dimensionality reduction techniques.

The remainder of this paper is organized as follows: Section II discusses related work in the field of Visual Analytics for education. Section III describes the dataset and the domain. Section IV briefly outlines data preparation. Section V details the design and implementation of our solution. Section VI presents the insights discovered using the tool, and Section VII concludes the work.

II. Related Work

The application of data mining and visualization to educational contexts is a well-established field. Our work draws inspiration from foundational studies in Educational Data Mining and recent advances in Visual Analytics.

A. Foundations of Educational Data Mining

The seminal survey by Romero and Ventura [1] establishes the framework for EDM. They categorize educational mining tasks into statistics, visualization, clustering, outlier detection, association rule mining, and sequential pattern mining. A key takeaway from their work is the critical role of visualization for non-technical users. They argue that while algorithms can detect patterns, visualization is the bridge that allows educators to interpret these patterns and translate them into pedagogical actions. Our project directly addresses this need by providing a visual interface for exploration rather than a static prediction report.

B. Predictive Modeling on Student Performance

Cortez and Silva [2] are the original authors of the dataset used in this project. In their work, they applied various Data Mining techniques—including Decision Trees, Random Forests, and Neural Networks—to predict student grades. Their findings highlighted that previous grades ($G1, G2$) are the strongest predictors of the final grade ($G3$), but they also noted the significance of social variables like alcohol consumption ($Walc$) and mother's education ($Medu$). **Comparison:** While Cortez and Silva focused on maximizing predictive accuracy (a "black box" approach where the goal is the output score), our work focuses on the exploratory process (a "white box" approach). We aim to visualize the *distributions* and *correlations* that their models exploited, allowing users to verify these relationships interactively.

C. Visual Analytics Systems for Education

Several dedicated VA tools have been proposed. Govaerts et al. [3] introduced the "Student Activity Meter" (SAM), a visualization tool designed to track student time-tracking data and resource usage in online learning environments. SAM focuses on helping both teachers and students reflect on their activity through temporal visualizations. **Comparison:** The primary difference lies in the nature of the data and the analytical goal. SAM deals with *temporal* activity streams (time spent on tasks), whereas our dashboard focuses on *multidimensional static profiling* of demographic and social attributes. Our challenge is not visualizing time, but visualizing the high-dimensional correlation between lifestyle and performance.

D. Integration of Dimensionality Reduction

The challenge of visualizing high-dimensional data is central to our project. Sacha et al. [4] proposed a "Knowledge Generation Model" for Visual Analytics, advocating for the tight integration of automated analysis methods (like dimensionality reduction) into the visual exploration loop. They argue that users should be able to interact with the model's parameters and see the results instantly. **Comparison:** We adopt this methodology by integrating Principal Component Analysis (PCA) directly into the dashboard. Unlike static PCA plots found in many reports, our PCA view is coordinated with the other views, allowing users to filter a cluster in the PCA projection and immediately see the demographic profile of those students in the Parallel Coordinates view.

III. Dataset and Domain

A. The Dataset

We utilized the **Student Performance Data Set** (Math course) obtained from the UCI Machine Learning Repository. The data was collected using questionnaires and school reports from two Portuguese schools.

- **Volume:** 395 Instances (Students).
- **Dimensionality:** 33 Attributes per student.
- **Target Variable:** Final Grade ($G3$, scale 0–20).

B. Domain Description

The attributes provide a holistic view of the student's life, categorized as follows:

- **Demographics:** Sex, Age, Address (Urban/Rural), Family Size ($famsize$), Parent's cohabitation status ($Pstatus$).
- **Socio-Economic:** Mother's and Father's education ($Medu, Fedu$) and job ($Mjob, Fjob$).
- **Academic History:** Number of past class failures ($failures$), School absences ($absences$), and grades for the first and second periods ($G1, G2$).
- **Lifestyle:** Weekly study time, Free time after school, Going out with friends ($goout$), Workday alcohol consumption ($Dalc$), Weekend alcohol consumption ($Walc$), and current Health status.

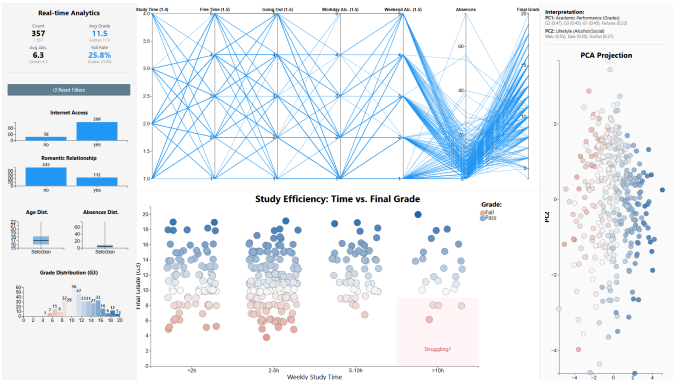
IV. Data Preprocessing

Although the UCI dataset is structured, it required specific preprocessing steps to be suitable for visual analysis. To ensure a seamless user experience, we implemented these steps directly within the client-side application logic (`index.js`), removing the need for offline Python scripts.

- **Parsing and Cleaning:** The raw CSV file utilized semicolons (;) as delimiters, which is non-standard for many web parsers. We implemented a custom parsing routine using `d3.dsvFormat` to correctly structure the data objects.
- **Type Conversion:** Many attributes were encoded as categorical strings (e.g., numeric levels "1" to "5" stored as text). We systematically converted these to JavaScript integers to enable mathematical operations such as calculating means, correlations, and PCA projections.
- **Filtering:** We verified data integrity by filtering out any rows with missing or invalid target variables ($G3$), ensuring that our performance analysis remains accurate.

V. Our Solution

We designed a web-based Visual Analytics dashboard built with **D3.js** for visualization and **Webpack** for efficient module bundling. The interface follows the "Overview first, zoom and filter, details on demand" mantra.



A. Visual Design Rationale

1) 1. Multidimensional Profiling (Parallel Coordinates): Located at the top-center, this view handles the complexity of the dataset. We specifically selected 7 key dimensions for the axes to balance complexity with clarity: **Study Time**, **Free Time**, **Going Out**, **Workday Alcohol** ($Dalc$), **Weekend Alcohol** ($Walc$), **Absences**, and **Final Grade** ($G3$). **Design Choice:** Parallel coordinates are ideal for spotting correlations (parallel lines) and inverse correlations (crossing lines) across multiple variables simultaneously. We implemented interactive "brushing" on axes, allowing users to filter the population by defining ranges (e.g., selecting only students with high study time).

2) 2. Performance Analysis (Scatter Plot): The bottom-center view focuses on the relationship between effort and outcome. It maps **Weekly Study Time** (X-axis) against **Final Grade** (Y-axis) to visualize study efficiency. **Visual Encoding:**

- **Color (Semantic):** We implemented a custom linear color scale designed to separate passing from failing students. The scale maps grades < 10 (Fail) to **Pink** and grades ≥ 10 (Pass) to **Blue**, with intermediate colors representing the transition between these states. This allows users to instantly perceive academic performance levels across the distribution.
- **Jittering:** Since grades (0-20) and study time categories (1-4) are discrete integers, standard scatter plots suffer from overplotting. We implemented a jittering technique, adding random noise to the coordinates to spread the points. This allows density estimation, revealing clusters (e.g., many students studying 2-5 hours) without distorting the data trends.

3) 3. Cluster Analysis (PCA Projection): To satisfy the requirement for dimensionality reduction, we implemented Principal Component Analysis (PCA) using the `pca-js` library. This view (right panel) projects the 33-dimensional student vectors into a 2D space. **Rationale:** PCA allows us to see "clusters" of students who share similar holistic profiles. Our analysis reveals that PC1 (explaining 21.2% of variance) largely correlates with **Academic Performance**, while PC2 (explaining 13.1% of variance) captures **Social and Lifestyle Behaviors**. The chart explicitly displays these variance percentages to aid interpretation.

4) 4. Context Filters (Bar Charts, Histograms, Box Plots): The left panel serves as a control center, utilizing standard statistical charts to control the complex views.

- **Bar Charts:** Visualize categorical variables. We chose **Internet Access** and **Romantic Relationships** as key binary filters. Clicking a bar filters the entire dashboard. We implemented dynamic counts that appear above the bars to show the exact number of selected students in each category.
- **Dynamic Box Plots:** We integrated dynamic box plots for numerical distributions, specifically **Age** and **Free Time**. These allow users to see the median and quartiles of the selected subset, providing a quick statistical summary of the active population.
- **Histogram:** Shows the distribution of Final Grades ($G3$). This view is critical for assessing the "shape" of performance for a selected subgroup (e.g., does the curve shift left for heavy drinkers?).

B. Interaction and Real-Time Analytics

A static visualization is insufficient for complex analysis. Our system features tight **Brushing & Linking**: a selection in any chart immediately propagates to all others. Crucially, this interaction triggers **Real-Time Analytics**, displayed in the stats panel.

The system dynamically recalculates and displays the following key metrics for the active selection:

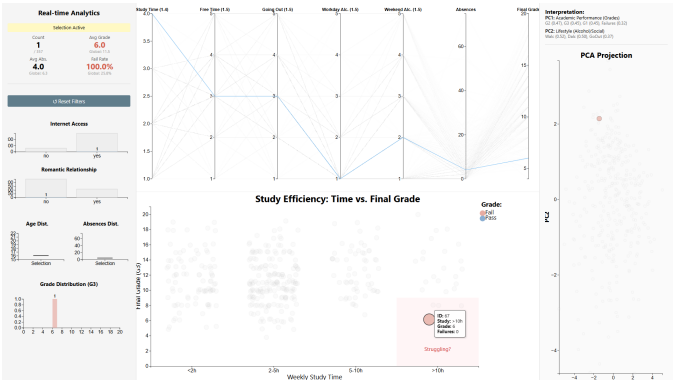
- **Count:** The total number of students currently selected.
- **Failure Rate:** The percentage of selected students with a failing grade ($G3 < 10$). This provides an immediate, quantifiable risk metric.
- **Average Absences:** The mean number of school absences for the group.
- **Average Grade:** The mean final grade ($G3$) for the group.

VI. Case Studies and Insights

To validate the analytical power of the system, we conducted an analysis focusing on the relationship between student effort (weekly study time) and academic outcome (Final Grade $G3$). The "Study Efficiency" scatter plot, combined with the coordinated views, revealed three significant insights that challenge simple linear assumptions.

A. Insight A: The "Inefficient Effort" Paradox

Hypothesis: We initially hypothesized that students studying more than 10 hours weekly would consistently achieve higher grades.



Visual Evidence: By observing the Scatter Plot, we identified a cluster of students in the bottom-right quadrant: those reporting **>10 hours** of weekly study time but achieving **failing grades (<10/20)**.

Interaction & Analysis:

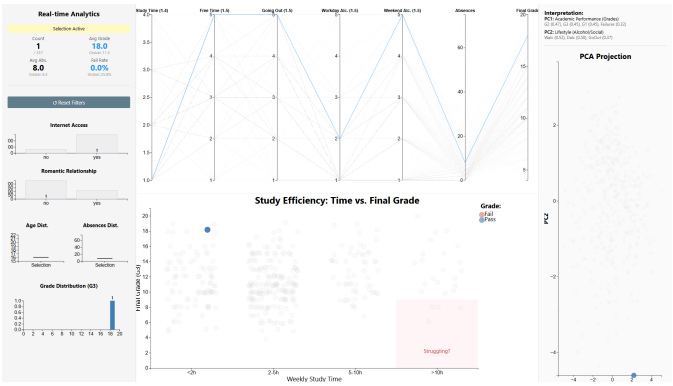
1) We probed this cluster by selecting representative students individually.
2) The coordinated Parallel Coordinates view highlighted that these individuals generally maintain **low to medium alcohol consumption** and **regular school attendance** (*Absences* are not high), ruling out behavioral negligence as the primary cause.
3) Interestingly, the PCA projection showed these students clustered away from the "Lifestyle/Alcohol" axis (PC2), further indicating that social distractions are *not* the primary cause of their failure.

Conclusion: These students are "inefficient learners." They possess the motivation (high study time) and discipline (attendance), but likely lack foundational skills or effective learning

strategies. A generic intervention like "study more" would be ineffective; they specifically require remedial tutoring or method coaching.

B. Insight B: The "Gifted Underachiever" Risk

Hypothesis: Students with minimal study time (<2 hours) are expected to fail.



Visual Evidence: The Scatter Plot revealed a counter-intuitive group in the top-left quadrant: students studying **<2 hours weekly** yet achieving **excellent grades (>15/20)**.

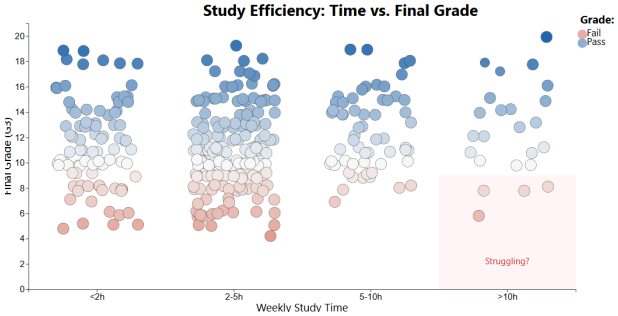
Interaction & Analysis:

1) By selecting individual high-performers within this group, we inspected their detailed profiles.
2) Cross-referencing with the Parallel Plot revealed a recurring behavioral pattern: significantly higher frequency of going out ($goout$) compared to the typical high-achieving student.

Conclusion: While these students are currently succeeding due to natural aptitude, the system flags them as high-risk for future academic stages (e.g., University) where raw intelligence without work ethic is often insufficient. The visualization allows the school counselor to identify "coasting" students who need engagement challenges rather than academic support.

C. Insight C: The Diminishing Returns of Study Volume

Hypothesis: A linear correlation exists between study hours and grades.



Visual Evidence: By comparing the density of grades across the categorical X-axis of the Scatter Plot ($< 2h$, $2-5h$, $5-10h$, $> 10h$), we observed a "saturation point."

Analysis: The average grade for the $5-10h$ group is often equal to or marginally higher than the $> 10h$ group.

Conclusion: The data suggests a law of diminishing returns. Students pushing beyond 10 hours do not see proportional grade increases, potentially due to burnout or stress (correlated via the $health$ variable in the Parallel Plot). This insight suggests that school policy should focus on study *quality* rather than enforcing higher study *volume*.

VII. Conclusion

In this project, we developed a comprehensive Visual Analytics system to support educational decision-making. By moving beyond black-box predictions and offering an interactive, exploratory environment, we empower domain experts to understand the multidimensional risks affecting students. The successful integration of dimensionality reduction (PCA) and real-time statistical monitoring provides a depth of analysis that static reports cannot match. Future work could focus on integrating temporal data to track student trajectories over time, further enhancing the tool's predictive capabilities.

References

[1] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601-618, Nov. 2010.

[2] P. Cortez and A. Silva, "Using Data Mining to Predict Secondary School Student Performance," *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*, pp. 5-12, Porto, Portugal, April 2008.

[3] S. Govaerts, K. Verbert, E. Duval and A. Pardo, "The student activity meter for awareness and self-reflection," *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, pp. 869-884, 2012.

[4] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis and D. A. Keim, "Knowledge Generation Model for Visual Analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1604-1613, Dec. 2014.