Data Research

Table of Contents

Introduction

The research questions at the core of this project are vital to uncovering how personalized education can support students more effectively. Key questions include: What learning styles and behaviors are most common among Ethiopian high school students? How do different learning styles and study habits influence academic performance? Can machine learning accurately predict student success and recommend personalized resources? Addressing these questions is essential to designing an educational system that not only supports academic improvement but also promotes inclusive learning—especially for underserved and remote communities. A thorough exploration of data is crucial to achieving this goal. By analyzing real-world and synthetic data from student profiles, quiz results, and in-app engagement, we can identify patterns in successful learning strategies, understand the impact of past academic performance and study habits, and train reliable machine learning models to drive effective and personalized recommendations. Without this foundational data analysis, any personalization efforts would rely on assumptions rather than evidence. Data research enables us to move beyond “one-size-fits-all” teaching and toward a student-centered system that empowers learners based on their individual needs and potential.

Organization

We organized our data research thematically into two categories: (1) a historical baseline dataset to establish patterns of student performance and learning behaviors, and (2) fresh user data to test and validate our Personalized Learning Management System (PLMS). This structure supports our goal of helping Grade 12 students pass the EUEE and understand subjects well, improving overall education quality in Ethiopia.

Data Description

Historical Baseline Dataset (Planned): We will collect real data from 50 Grade 12 students to create a historical baseline, capturing their past learning styles (visual, auditory, kinesthetic), past experiences (previous grades, 0-4 hrs/day social media, 20-100% assignment completion), and academic outcomes (e.g., past exam performance). Source: Surveys and interviews with

students, reflecting EUEE trends (e.g., 3.2% pass rate in 2023). Format: CSV, 50 rows, 10 columns (e.g., math_score, learning_style). Size: ~5 KB. We chose this to establish a reference for collaborative filtering, reflecting real Grade 12 student profiles. If more data is needed, we may adapt a Kaggle dataset (e.g., "Student Performance").

Fresh User Data (Planned): We will collect real data from 10-25 Grade 12 app users via a mock quiz (10 questions, EUEE-style) and surveys, capturing quiz scores, learning styles, and past experiences. Source: App interactions (quizzes, surveys). Format: CSV, 10-25 rows, 8 columns. Size: ~2 KB. This will test our system's personalization for Grade 12 students.

Local High School Data (In Progress): We are negotiating with a local high school to collect additional real Grade 12 student data (e.g., mock exam scores, study habits), which may expand our baseline dataset.

Study Materials: A library of 10 resources tailored for Grade 12 students preparing for the EUEE, including Khan Academy videos (e.g., "Grade 12 Algebra Video"), YouTube EUEE prep videos (e.g., "EUEE English Grammar Tutorial"), and local resources (e.g., Fetena quizzes), focusing on exam preparation and subject comprehension. Source: Public platforms (e.g., Khan Academy, YouTube, Fetena). Format: Text/CSV, 10 rows, 5 columns. Size: ~1 KB. These resources align with the Grade 12 curriculum and EUEE requirements.

Data Analysis and Insights

Historical Baseline Dataset (Expected Findings): Once collected, we expect the data to show patterns like: 60% of Grade 12 students with >2 hrs/day social media scoring below 50/100 in math, and visual learners with high assignment completion (80%+) averaging 20 points higher than auditory learners. These insights, to be derived via descriptive statistics (pandas), will guide our recommendations for exam success and subject understanding.

Fresh User Data (Planned): We anticipate similar trends in the 10-25 Grade 12 users' quiz results, such as lower English scores for high social media users, which we'll analyze using clustering (scikit-learn) to group users by performance and style.

Study Materials: Resources are Grade 12-specific and EUEE-focused, using Khan Academy and YouTube videos to address common challenges (e.g., English comprehension difficulties, a frequent EUEE failure point) while ensuring conceptual understanding for subjects like math and English.

Conclusion

Our data research uses real Grade 12 student data for both the historical baseline and fresh user data to establish performance patterns and test personalization. Expected insights will link learning styles and behaviors to academic outcomes, supporting our goal of helping Grade 12 students pass the EUEE and understand subjects well, using resources like Khan Academy and YouTube videos to improve education quality in Ethiopia. Ongoing negotiations with a local high school will further enhance our dataset.

References

[1] National Educational Assessment and Examinations Agency (NEAEA). (2023). EUEE Results Report.

[2] Tadesse, M., et al. (2020). Predicting Student Performance Using Machine Learning Techniques: A Case Study of Ethiopian Secondary Schools.