CS 598 Foundations of Data Curation Project Proposal

Usman Asghar (uasgh2@illinois.edu)
Hasham Ul Haq (huhaq2@illinois.edu)
Jiaqing Mo (jiaqing7@illinois.edu)

September 15, 2025

Overview

The overall goal of this project is to curate a dataset that explores the relationship between student well-being and academic performance in college. Student well-being has become an important focus in higher education, as stress, sleep, social life, and mental health can all influence how students perform academically. This dataset will provide a structured foundation for studying how different aspects of well-being connect to academic outcomes.

The use case is to enable researchers, educators, and institutions to better understand the role of well-being in student success. With this curated dataset questions such as How do sleep patterns relate to GPA? Does physical activity or social engagement impact academic performance? Which well-being factors have the strongest correlation with outcomes? can be addressed. Ultimately the project will contribute to building a clearer picture of how student lifestyle and well-being shape educational performance.

Plan

This project will follow the full data lifecycle model, broken down into the following stages:

1. **Ethics & Legal Handling**: We will make sure all data is used responsibly, following all licenses and terms of service. Since the studentlife dataset contains sensitive data we will make sure it's anonymous to protect participant privacy.

2. **Acquisition / Collection**: Gather data about student well-being factors like sleep, stress and activity level, along with demographic statistics, and academic performance from the Dartmouth StudentLife dataset portal.

3. **Modeling**: Develop a data schema that maps well-being indicators (e.g., sleep hours, stress levels) to academic performance metrics (e.g., GPA, attendance). Variables will be typed and annotated (categorical, continuous, ordinal), and consistent identifiers (e.g., participant ID, time period) will be used for relational modeling.

4. **Quality Assessment**: Perform data validation to check for out-of-range values, missing data patterns, and schema violations. Quality metrics (e.g., missingness ratio, data drift) will be computed, and issues will be logged and prioritized for correction.

5. **Cleaning & Integration**: Standardize formats, handle missing data and align well-being measures with academic performance outcomes.

6. **Automation & Provenance**: Create an automated workflow that handles the whole pipeline from raw data to curated dataset. Track dataset versions using Git and checksums for transparency.

7. **Metadata & Documentation**: Write a data dictionary and structured metadata file following standards such as schema.org or DataCite. Provide clear README instructions to make the dataset reproducible and easy to use.

8. **Dissemination & Reproducibility**: Package everything together into a GitHub repository. Users will be able to regenerate the curated dataset with a single command.

Datasets

The dataset we plan to use is Dartmouth StudentLife Dataset Portal (https://studentlife.cs.dartmouth.edu/dataset/), which is a collection of curated datasets and metadata resources that will support the project by providing examples of documentation practices and helping maintain consistency across the data lifecycle and integration process.

Licensing and usage policy:

1. This dataset does require citation - we will cite the relevant paper in our github repository.
2. To keep it simple and avoid any further licensing challenges, we will not publish the data as part of our submission. Users can download the data themselves. We will be publishing the details of the dataset, scripts, and metadata - without the actual copy of the dataset.

Team Members & Roles

Member	Role	Responsibilities
Usman	Project Manager & Modeling Lead	Coordinate team communication and deliverables Define schema, data model (relational + time-series) Map dataset to lifecycle stages and track milestone progress
Jiaqing	Data Acquisition & Ethics Lead	Identify and request dataset (StudentLife) Document ethical/legal considerations Summarize data source, license, and PII handling
Hasham	Data Cleaning & Workflow Automation Lead	Assess data quality (missingness, duplication, etc.)

		Implement cleaning and transformation scripts Set up reproducible workflow with Git/Jupyter

Project Timeline

(subject to change)

Date Range	Task	Owner(s)
9/10 – 9/15	Select dataset, define project scope, submit proposal	Jiaqing, Usman, Hasham
9/16 – 9/30	Acquire dataset, review ethics, draft data model	Jiaqing, Usman, Hasham
10/1 – 10/14	Explore and assess data quality, clean & transform data	Jiaqing, Usman, Hasham
10/15 – 10/26	Build automated workflow	Jiaqing, Usman, Hasham
10/27	Submit progress report	Usman (with support from Jiaqing & Hasham)
10/28 – 11/15	Metadata creation, schema diagram, data dictionary	Jiaqing, Usman
11/16 – 11/30	Package final repo (GitHub, scripts, README, docs)	Hasham
12/1 – 12/10	Final review, write narrative report, submit final project	Usman (compile inputs from Jiaqing & Hasham)

Constraints

1. Licensing Restrictions:

   - The Dartmouth StudentLife dataset cannot be redistributed. This means users must independently download it before using your pipeline.
   - May limit who can run the full workflow if they're unfamiliar with data acquisition or citation practices.

2. Data Sensitivity:
   - Although anonymized, the dataset includes behavioral data collected from students.
   - Any reuse must ensure ethical compliance and appropriate handling.
3. Heterogeneous Data Sources:
   - Integration of datasets from different domains (e.g., sensor data vs. demographic/economic data) may lead to challenges in normalization, alignment, and schema mapping.

Gaps / Areas Requiring Input

1. Ground Truth for Academic Performance:
   - The overall issue of student performance is much broader, and is impacted (severely) by other factors not covered in this study.
2. Lack of Demographic Detail:
   - The Dartmouth dataset may lack demographic diversity (race, gender, SES), which could limit the generalizability of our findings.