Final Project Proposal
CS-370 Python for Data Science
Spring 2025

Due: 5/2
Project Title: Academic Performance Factors
Team leader: Tatum Good
Other team members: Brooke Proctor, Courtney St Onge, Suzanne Gunderson

1. Data and descriptions (source, number of files, format, number of records in each file, etc.)
    a. Source: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors
    b. Number of files: 1
    c. Format: CSV
    d. Number of records in each file: 6,607
    e. Description: This dataset provides a comprehensive overview of various factors affecting student performance in exams. It includes information on study habits, attendance, parental involvement, and other aspects influencing academic success.
2. The project goal, tentative algorithms and approaches
    a. Exploratory Data Analysis to identify trends and relationships
    b. Linear regression for classifying performance levels
    c. Machine learning models for prediction
    d. Statistical analysis to find correlations between features?
3. Research questions or hypothesis
    a. What relationship exists between sleep hours and exam scores? What outliers/leverage points can be identified in this relationship, if any?
    b. What is the relationship between past and present exam scores?
    c. How do study habits and attendance impact the final exam scores? Does one of these variables impact final exams scores more than the other?
    d. How do socioeconomic factors (family income, school type, parental education level, access to resources, access to internet, etc.) affect academic performance?
    e. What are the interactions between motivation, sleep hours, and final exam scores?
    f. What makes an “ideal” student? What can our criteria be for that - ideally an even mix of traits like study habits and access to resources etc.?
4. Forecast/model techniques.
    a. New data visualization techniques (violinplot etc.)
    b. Machine learning
    c. Random Forests
5. New skills or knowledge need to learn
    a. Data Visualization in Python
    b. Data Cleaning/Wrangling in Python

    c. Potentially some statistical analysis
    d. Machine Learning
6. Foreseeable roadblock or difficulties
    a. Final exam scores do not seem to reach the top mark (101) so this could affect some results if we were to play with that column.
    b. Might need more datasets to find more patterns if 6,607 entries is not enough (listed above in the data description portion)
    c. Interpretations of data may be difficult due to a lack of initial data description (e.g. grade of students involved, state of collected data, other background knowledge).
7. A rough timeline and task division (if applicable)
    a. Week 7 - Exploratory Data Analysis, Increase Domain Knowledge
        i. Tatum
    b. Week 8 - Begin Modelling, exploratory as well
        i. Suzanne and Courtney
    c. Week 9 - SSRD abstract due (based on initial modeling)
        i. Brooke
    d. Week 10 - Refine research scope - create more models, advance current models
        i. Tatum, Courtney, Suzanne, Brooke
    e. Week 11 - Begin interpretations and potential conclusions
        i. Tatum, Courtney, Suzanne, Brooke
    f. Week 12 - Finish up interpretations, add additional changes here, finalize slideshow.
        i. Tatum
    g. Week 13 - SSRD Presentation
        i. Tatum, Courtney, Suzanne, Brooke
    h. Week 14 - Finish and update slideshow
        i. Tatum, Courtney, Suzanne, Brooke
    i. Week 15 - Final Presentations
        i. Tatum, Courtney, Suzanne, Brooke
8. Other information
    a. A lot of these are subject to change but these are many of our initial ideas
    b. As the project evolves and some exploratory data analysis reveals potential trends, some research questions may pivot and become more applicable to the data itself.