**Project Management Plan**

**Team Member Names and Emails**

Conner Alubowicz | calubowicz@arizona.edu

Benjamin Manninen | bmanninen@arizona.edu

Arjun Kumar | arjunkumar@arizona.edu

Drew Halloran | jahalloran0821@arizona.edu

**Team Name:**

Super Awesome Engineering Team

**Project Title:** Socioeconomic Factors Affecting Student Performance on Exams

Generative AI Tool Use Acknowledgement: Each team needs to document how and when they used a generative tool to accomplish any aspect of the project. \*I recommend an additional running doc that you can use to track use. Use of generative AI tools must be cited and referenced! Points will be deducted for inaccurate documentation, citation, or references. Entire projects may be rejected if no documentation is provided.

**Citation Style:** APA

**1.Define the Problem or Research Question**:

What outside factors most greatly influence a student’s standardized exam grades?

**Considerations:** By answering this question, we can better understand the effects of outside-of-school life and its effect on success in school. Understanding the effects of different socioeconomic factors can not only influence lawmakers who want education reform, but can also educate families, and help them better understand what they need to do in order to set their children up for success.

**2.Data:**

[Students Performance in Exams](https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams)

Kimmons, R. (n.d.). *Exam scores data generator*. Royce Kimmons. Retrieved November 12, 2024, from<http://roycekimmons.com/tools/generated_data/exams>

The above dataset contains 8 columns and 1000 rows, with features that include, gender, race/ethnicity, parental level of education, lunch type, test preparation, and exam scores for math, reading and writing.

The data comes from Kaggle, and is listed as public domain, so there is no need for special access to this data. The data originally comes from an Exam Scores Data Generator created by Dr. Royce Kimmons, a professor of Instructional Psychology and Technology at BYU. While the data is fictitious, the generator was carefully created to follow realistic patterns and distributions found in actual student data. This type of data is commonly used for educational and analytical purposes.

The dataset has a 10.0 usability score given by Kaggle, meaning that the data should be extremely usable and immediately ready to get started with.

**3. Data Exploration and Preprocessing:**

For cleaning the data, the first steps will be to search for any missing or duplicate data. If we find any rows that are missing essential data or identical, they will be removed from the dataset. Next, we will use data visualizations such as box plots to determine if there are any outliers in the data set. This will tell us if we need to perform any transformations. Then, we would have to create a new column in the data to categorize the scores as high, medium, or low. This will allow us to use a classification model to analyze the data.

The quality check plan will have each member review the data set, visuals, and code to confirm that everything is in order.

**Exploratory Data Analysis (EDA)**

**4. Expected Modeling:**

Since the data contains specific categories such as race or parental education, it would make the most sense to use a classification model to review this information. There are multiple classification model options using sklearn, which is a common package that has been present throughout this course.

**Model Selection:**

**5. Model Evaluation Plan:**

One of the most important evaluation tools will be the AUC and ROC scores. This is because we are using classes. For the gender category, this is a binary class, so we should see if the model can differentiate between these groups. I believe the dataset being proposed is fairly easy to interpret.

**6. Model Deployment or Implementation Plan:**

For this project, we are going to work in a Python environment using several different packages. These packages will include pandas, numpy, matplot, seaborn, and sklearn. These packages will be used to process, visualize, and predict outcomes based off of the original dataset.

This modeling process will be documented in two ways. The first will be through the github push history. This will keep a record of the different major changes that occurred to create the model. The second will be personal logs about our process in creating the model. This will ensure that all the ideas used to create the model will be visible.

**7. Communication:**

The team will communicate via Slack. We will stay in contact and update each other about our progress and issues. Once a week, we will meet on Monday to review the progress from the week, and look at the work ahead. This will keep us on track and informed.

8. **Risk Management:**

Potential Risks:

* Using race as a feature in tests results can lead to results that may be misinterpreted.
* Not getting full support from all team members.

Solutions:

* Values within race category are not labeled as actual races (white, black, hispanic, etc.), rather they are just listed as (race 1, race 2, race 3, etc.).
* Team has shown good communication skills in splitting up work that has not been completed by other members. This approach will continue with the rest of the project.

10. **Timeline and Milestones**

For this project, we have about a month between the project proposal and the final project presentations. So, we believe that we can break the project timeline into four main phases, and spend about a week on each starting on Wednesday, November 13.

* Week 1 - Data Preprocessing / Cleaning
* Week 2 - Data Visualization and Relationship Identification
* Week 3 - Data Modeling and Predicting
* Week 4 - Model Evaluation and Conclusion