Computer Science Capstone – C964 Project

Analyzing and Projecting Student Performance

Feng Parra

Table of Contents

[A1. Letter of Transmittal 2](#_Toc80032793)

[A2. Project Recommendation 3](#_Toc80032794)

[Problem Summary 3](#_Toc80032795)

[Application Benefits 3](#_Toc80032796)

[Application Description 4](#_Toc80032797)

[Data Description 4](#_Toc80032798)

[Objective and Hypotheses 4](#_Toc80032799)

[Methodology 4](#_Toc80032800)

[Funding Requirements 5](#_Toc80032801)

[Stakeholders Impact 5](#_Toc80032802)

[Data Precautions 5](#_Toc80032803)

[Developer’s Expertise 5](#_Toc80032804)

[B. Project Proposal 6](#_Toc80032805)

[Problem Statement 6](#_Toc80032806)

[Customer Summary 6](#_Toc80032807)

[Existing System Analysis 6](#_Toc80032808)

[Data 7](#_Toc80032809)

[Project Methodology 7](#_Toc80032810)

[Project Outcomes 7](#_Toc80032811)

[Implementation Plan 8](#_Toc80032812)

[Evaluation Plan 8](#_Toc80032813)

[Resources and Costs 9](#_Toc80032814)

[Timeline and Milestones 9](#_Toc80032815)

[D. Post-Implementation Report 10](#_Toc80032816)

[Project Purpose 10](#_Toc80032817)

[Datasets 10](#_Toc80032818)

[Data Product Code 10](#_Toc80032819)

[Hypothesis Verification 10](#_Toc80032820)

[Effective Visualizations and Reporting 10](#_Toc80032821)

[Accuracy Analysis 10](#_Toc80032822)

[Application Testing 10](#_Toc80032823)

[Application Files 10](#_Toc80032824)

[User’s Guide 10](#_Toc80032825)

[Summation of Learning Experience 10](#_Toc80032826)

[E. Sources 10](#_Toc80032827)

# A1. Letter of Transmittal

June 6, 2021

Mary Douglas, Superintendent of Public Instruction

Montana Office of Public Instruction

3215 16th Ave.

Mt. Ana, MT 59342

Dear Ms. Douglas,

I have been contacted by a member of your staff about how I could help further your public school system. Specifically, the staff member wanted to know if I could assess and use the exam performances of your students to better adjust the curriculum.

My proposal is to take your existing set of data pertaining to student scores and then use them to help shape the work and attention needed for your classrooms. Using this existing set of data, our dedicated analysis application here at Ki Co. can help determine the performance of a student before the test is even taken.

Knowing these projected scores can drastically improve the performance of your students as well as the teachers. Teachers or department heads can use these scores to adjust the course load of their classrooms to better suit the students within them. This could help relieve stress for everyone involved in the classroom as their needs will be better catered to.

The objectives are simple:

1. Analyze current scores to create a predictive formula.
2. Use the predictive formula to project future test scores.

To have this project come to life, the initial application will need funding. As of now, the project will need $5,000 to cover the cost of a single developer, $2,000 for the time needed to test the application and $10,000 for maintenance of the project past the point of its inception.

Our suggested developer has worked with several educational institutions around the world. These institutions have seen a rise in student performance while using our proposed application.

The developer that is to be assigned to the project has worked on similar applications alongside other educational institutions. These institutions have seen a rise in student performance while using our proposed method.

With all that in mind, I look forward to hearing from you.

Sincerely,

Feng Parra, Ki Co. CEO

# A2. Project Recommendation

## Problem Summary

The problem is that exam scores are currently not being used efficiently. Some students need more attention than others and the curriculum does not usually change to reflect those needs. If we could reliably predict the scores for an upcoming exam, the teacher could determine whether to ramp up their teaching methods, scale back or simply keep everything the same.

## Application Benefits

Classrooms will benefit greatly using our application. Teachers can use our scores to assess not just the performance of the class, but their own. For example, changing a curriculum on the fly can do more damage than good. Yet, what if the teacher were provided with enough information to warrant the shift, perhaps the benefits would then outweigh the risks. The instructors using our application can use our projected scores to better adjust for the next exam.

More than anything, the students will benefit most. A student can be at ease knowing that their learning experience will be better suited to their needs. Teachers as much as the students will feel a sense of relief when using our application. Projecting performance removes the uncertainty of how a student has and will perform on future exams. Students and teachers will both be better prepared throughout the year.

## Application Description

The application overall is simple. First, it will take a set of data that is to be analyzed. After the data analysis is complete, the application will then project upcoming exam scores. Our application will be written in Python in conjunction with a few libraries.

The Python libraries our team used included pandas to read the data, matplotlib and seaborn to create the visuals, and numpy and sklearn to apply our methodology. Voila was also used to create a dashboard that was more user-friendly.

## Data Description

Our data is stored within a comma-separated values file (.csv) consisting of over 2,000 entries. These entries house the scores of many students, specifically a range of two different scores, along with a student ID to help differentiate between entries.

The data is saved as a .csv file as it can be easily read by our Python libraries. When opening the file within a program such as excel, you can see that there are other attributes associated with each entry. A school ID, male or female, etc... These attributes will not be used in our application but could be used in future iterations.

The analysis and predictions of our application rely on a set of pre and post exam scores. In order to provide an accurate performance for the future, the application needs an initial before and after example that will judge the progression of a student. The pre and post exam columns will serve as the initial set.

## Objective and Hypotheses

The objective is to provide an accurate enough prediction that can help shape the curriculum of a classroom throughout the year. Our hypothesis is if we can evaluate the performance of a student based on exam A to exam B, then we can determine how that student will perform on exam C.

## Methodology

There are several methodologies to pick from, but due to the size of the team involved, the project will be implementing the waterfall methodology. This is a method that must clearly define the various stages of a project that will be kept along a track. It inherently promotes strong planning in the beginning that will allow the team to easily move along the track (Stephens).

We also want to make sure the project is finished in time for the next school term. Due to this, development speed will be important. The waterfall methodology can develop a product rather quickly as there is no going back once you advance along the track. Any roadblocks or potential need for changes can be noted in the planning phase and added to the track before the project begins.

Our application will output complex results, but the application itself is simple. This method is easy to follow and works best with a small team and project, so it is the ideal choice for our project.

## Funding Requirements

The project does not require much funding. The most important tools we are using to develop the application are free to use. This includes the Python language and the associated libraries.

Beyond the tools, funding is mostly required to pay our team and to keep the application operational after it has been released. An upfront cost of $5,000 will be needed to pay the developer as he completes the application. Once the code is written, testing will be required to make sure that the application functions as it should. $2,000 will suffice for the additional time needed to test the system. After the applications is released, it will need continuous updates as more data comes in. The incoming data will help refine the application and allow for more features in the future. This comes at the price of $10,000 to secure a long-term maintenance cycle.

Overall, the cost of the project as anticipated will be $17,000.

## Stakeholders Impact

The stakeholders understand that the projects success can lead to more contracts in the future, expanding the business and strengthening our finances. They want to get more involved with state and federal governments as they believe that will provide the steadiest stream of income for the company.

They are fully behind the success of this project, and if we succeed, Ki Co. can then use the application’s functionality in other aspects of business.

## Data Precautions

There are no links to real people within our data. We use this data to suffice the statistical criteria that we need to meet in order to provide a more accurate projection. Since we are using the data in a statistical capacity, we are not violating any terms set out by its publisher. The website itself prohibits data and services that violate any regulations set by HIPPA and the like, so we can be sure that our initial dataset does not violate any terms of that nature.

## Developer’s Expertise

The developer assigned to the project has a bachelor’s degree in computer science and has written numerous projects in the Python language. He also has experience in working with data analysis.

It should be noted that this is not the first time he has developed an application similar to this one. He has worked on similar applications in the past with other educational institutions. Although he is new to the company, these institutions have vouched for him and his credibility.

He has all the technical skills and characteristics needed to successfully complete this project.

# B. Project Proposal

## Problem Statement

Every school system uses testing to measure the success of their students, but exam scores are rarely used in a way to help students actually succeed. Periodically teachers hand out exams, pass or fail the students and then move onto the next topic. This method does not suit the needs of the student and can hinder the progress of several students at a time.

Instead, what we can do is propose a solution using data analysis. Students will still take exams, but the scores will be used more efficiently. These exam scores can be used to project how a student will perform on an upcoming exam through predictive analysis.

Using the Python programming language alongside libraries such as Numpy, Matplotlib, Sklearn as well as a few others, an application can be designed that accurately predicts a student’s performance. This will be done using a statistical modeling process known as linear regression.

## Customer Summary

The application caters to educational institutions that want to improve the performance of their students. It is a tool that can improve the efficiency of a classroom and help relieve some stress for the students and teachers alike.

The backbone of the product is to provide the user with accurate projections. This makes the application simple to use as the calculations run in the background with little to no user interference.

## Existing System Analysis

The application does not require much to be completed. Using Jupyter notebooks and the Python libraries, much of what is needed is covered by these two resources. Since the application will be run from a remote location in our offices, any hardware needed has already been set aside for this project.

## Data

The data is stored within a comma-separated values file (.csv) consisting of over 2,000 rows. It was submitted by a user named Kwadwo Ofosu and hosted on the website Kaggle (<https://www.kaggle.com/kwadwoofosu/predict-test-scores-of-students>). The data has several attributes including a school ID, the area associated with the school, its type, classroom details and other miscellaneous details. Most importantly, it contains a large amount of test scores associated with student IDs.

There are two sets of scores, a pretest score, and a protest score. This is the data to be focused on and the data that will determine our projections. When these scores are read by the application, the linear regression model will analyze the scores and a relationship will be realized between the two scores. When this relationship is identified, the application will create a formula that can predict a future score.

## Project Methodology

The application will be developed using the Waterfall methodology. Although Agile was under consideration, Waterfall makes more sense due to the size of the team working on the project. Our workflow process is presented as follows:

1. Requirement Gathering and Analysis:

* The team will meet with the client to outline the application and gather any valid information needed to complete the project.

1. System Design:

* The team will use any information gathered from the client to determine if there are any hardware or system specifications to be met.

1. Implementation:

* The foundations of the application will be developed here. Functionality will be created and then tested before moving on to the next phase.

1. Integration and Testing:

* All the functions and separate components of the application will be connected here in a presentable manner. Further testing will be done to make sure everything works together as intended.

1. Deployment of System:

* This is where the application will be released to the client.

1. Maintenance:

* All application issues will be fixed in this phase and any additional features can be implemented here too.

## Project Outcomes

The project itself is small in scale, so there will be few deliverables. Some of the deliverables will be:

* Presenting a project timeline to the client.
* Reports every time the project moves onto the next phase that detailing what was accomplished, and any pitfalls encountered.
* The final version of the application displaying the functionality and effectiveness of the program

## Implementation Plan

The application will be released to the client once all testing has been completed and the team officially arrives at the deployment phase of our timeline. Once the application is given to the client, it is assumed that an administrator or team will be assigned to use the application. All information regarding the use of the application will be outlined within the documentation. Ki Co. is also on standby in case there are any questions regarding the use of the program and there will be an initial training session provided by the developer as well.

Alternatively, Ki Co. can administrate the application for specified amount of time. Negotiations regarding the handling of the program will take place at this point if the client does not have the resources in place to use the product efficiently.

Ki Co. will continue to monitor the application for several months as updates will be implemented monthly. For circumstances that require urgent fixes then our team will prioritize the development of a fix and delivered as quickly as possible.

Additional features will also be added to the program over time, adding functionality that further improves the collection of data and accuracy of results. Data collecting will be vital to the success of the project and Ki Co. will need that data collected by the client in order to improve the accuracy of the application.

## Evaluation Plan

Ki. Co will monitor the progress of the application for several months, verifying that the application works as intended. Results will be gathered and tested for accuracy. If the results gathered are inaccurate or require some tuning, then the team will administer fixes to improve the functionality of the program.

Updates will be needed from the client, assuming Ki Co. is not administrating the application, to make sure that the application is functioning as it is supposed to.

Overall, the application will be monitored to make sure that the accuracy rate is initially at least 80%. This accuracy rate is based on how the students on performing on exams and comparing them to the projections set out by the program.

Student improvement on exams will also be documented to better see if the application is fulfilling its purpose. The overall objective of the program is to improve the performance of students, so it is important to Ki Co. to evaluate the progress of students as the application is used.

## Resources and Costs

* Programming Environment:
  + Python along with libraries such as Numpy, Matplotlib, Sklearn, Seaborn and Pandas are all free and will run on Windows 10 computers. Programming will be done within Jupyter Notebook within Chrome browser. All these things will be available on our developer’s work computer within the office.
  + Cost: $0.00
* Environment Costs:
  + Ki Co. offices will be used as the environment for the work to be completed. Our company provides everything the team will need at no additional cost to the client.
  + Cost: $0.00
* Human Resource Requirements:
  + The only expense to be incurred on the client will be the wage of the developer. The developer will create the program, test, and maintain it.
    - Application development cost: $5,000.00
    - Testing before deployment cost: $2,000.00
    - Product maintenance cost: $10,000.00
  + Cost: $17,000.00

## Timeline and Milestones

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Task** | **Start Date** | **End Data** | **Duration** | **Dependencies** | **Resources Assigned** |
| Requirement Gathering and Analysis | 6/28/2021 | 7/2/2021 | 1 Week | N/A | Client |
| System Design | 7/5/2021 | 7/9/2021 | 1 Week | Requirements and Analysis | 4/5/2021 |
| Implementation | 7/12/2021 | 7/23/2021 | 2 Weeks | System Design | 4/16/2021 |
| Integration and Testing | 7/26/2021 | 8/6/2021 | 2 Weeks | Implementation | 4/20/2021 |
| Deployment of System | 8/16/2021 | 8/20/2021 | 1 Week | Integration and Testing | 4/23/2021 |
| Maintenance | 8/30/2021 | 2/1/2022 | 6 Months | Deployment | 4/26/2021 |

# D. Post-Implementation Report

## Project Purpose

The education system can be flawed at times and in this case, we consider the examination process. It is safe to say that exams can be scary for students, you either pass or fail and then move onto the next topic. A lot of times students are ignored, and their needs are put to the side after taking these tests. What we can do is repurpose the use of these assessments and paint them in a better light for the student and the teacher.

This project was designed to help schools assess the progress of their students and eliminating the lack of effectiveness that exams currently provide. The idea is that if you can consider how a student will perform on a quiz or exam leading up to a final assessment, then you can accurately predict how that student will perform. If you can accurately project what a specific student, or an entire class, will receive as a grade on a particular exam, then you can better prepare them for that final test.

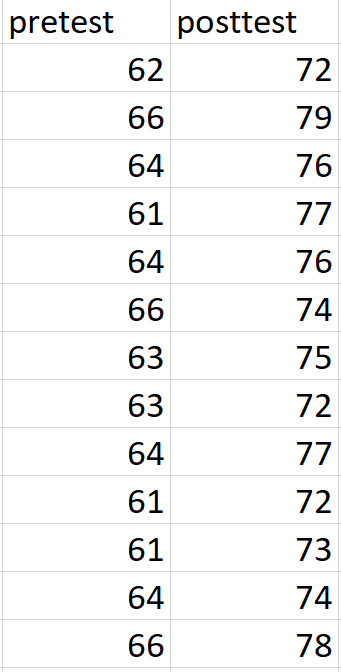
Teachers can adjust their curriculum on the fly every time a score is fed into the machine. They will receive real time data that does not have to be manually computed. Our project considers the entire performance history of a student, not just the most recent score. This is very important since it is one of the most reliable ways of assessing performance and ultimately predicting accurate and relevant scores.

## Datasets

The dataset is stored in a comma-separated values file which consists of various test scores. This data was gathered from Kaggle (<https://www.kaggle.com/kwadwoofosu/predict-test-scores-of-students>). In the file, there are multiple attributes associated with each data entry. The characteristics of the data we are most interested in are the pretest and posttest scores. These scores are vital to accurately predicting future test performance. Below are examples of the dataset in a readable form:



**Fig. the dataset in spreadsheet form**



**Fig. part of the dataset to be focused on**

The other columns within the dataset are not needed and instead ignored completely by the application. Our focus lies solely on the exam scores. The only objective of the data is to create projections and the other attributes are not relevant to this factor.

We considered using the other attributes but adding other factors could mislead the application into giving an inaccurate prediction. Then there was the case that some of the columns were entirely irrelevant to the purpose of our project.

The relationship between the pretest and posttest scores is very important to determining our projection. By establishing how a student performed from one exam to another, the application can apply that increase or decrease in performance to predict another score.

## Data Product Code

The code itself is rather simple. What we were trying to achieve was fully capable with a limited amount of code. There is room for improvement which can be addressed in future updates, but overall, the purpose is fulfilled.

The application was coded with the capability to display the data in a more user-friendly fashion. This was done by creating three different visualizations that shows the relationship between pre and post test scores. We accomplished this by using the Python library, matplotlib, in conjunction with the seaborn library.

Our descriptive and predictive methods were represented throughout the code using the sklearn Python library. Both methods were based off the Linear Regression model which sklearn provides as a built-in function. It allowed us to quickly sort and split the data along with the numpy library. It also provided functionality to create the projections and score the accuracy of our program.

## Hypothesis Verification

We stated in our hypothesis that if we had a pre and post test score, A and B, could determine another score, C, based on scores A and B. By focusing our scope on just predicting a score using these two variables, our application is fully capable of projecting a score.

It is important to note that there are other characteristics we could use to improve the accuracy of our projected scores. At the end of the day, our application is simply outputting predictions. It is possible that a predicted performance will be farther off from the real score. If we were to introduce more variables into the mix, such as time spent studying, then we could generate a more accurate projection.

## Effective Visualizations and Reporting

Our team chose to represent our data in three different forms. All the visualizations focused on every pre and post exam score and not simply limited to a specific range. Although we focused on the same range of data, it helps to visualize this data in different ways so that a user can understand what they are working with and assess the values themselves.

The data was first represented in a scatter plot diagram. And this was the most effective way to showcase the relationship between scores. With a scatter graph, we can draw a best fit line that showcases what our linear regression model is actually doing. This also helped represent any outliers that might be found within our data. Fortunately. there were not many, and a clean line was drawn across a substantial number of points in our data.

We also represented the scores in a histogram chart and kernel density estimate (KDE) plot. Both charts are closely related and showcased where the scores mostly lied separate from each other. These visualizations were important because they show the user how the scores improved.

## Accuracy Analysis

Our system seems to be fairly accurate. It is difficult to tell because our team used such a limited amount of data. This was mostly due to focusing the scope of what we wanted to achieve and whether we were capable of proving our hypothesis. As stated before, it is highly plausible that the expected score turns out to be far different than the actual score. Some outlying factors that were not considered could be gender, race, the district itself, all of which could provide a more accurate projection.

## Application Testing

Fortunately for our team, the application was created in an environment that provided us with errors on the spot. The simplicity of the application itself also made for little errors to be encountered.

The code was tested by running the program after every change to make sure that Jupyter did not throw back an error. The libraries chosen provided us with an abundance of functions to help visualize and represent the data, so it was important that our environment could help guide us through the error process.

Due to the simplicity of the product and the ease of use in the environment, any formal testing was not required. Simply running the source code repeatedly was enough to make sure the application came out error free. It is not to say that errors encountered were not noted down.

The team recorded any errors encountered to keep the client informed. Our team was also very small, in fact, our team was just one person, so it was easier to keep track of errors and fix them. Communication was utilized and it was key to the applications completion.

## Application Files

The application only needs four files to function as intended. These files will be submitted to the client, but they will also be hosted on GitHub.

The files included are:

* Main.ipynb – This is the Jupyter Notebook file that contains the source code
* Main.py – This file contains a copy of the source code
* test\_scores.csv – Contains our dataset
* requirements.txt – Simple text file with all the packages needed to properly launch the notebook

## User’s Guide

1. Go to <https://mybinder.org/>
2. Under ‘GitHub repository name or URL’ paste the following link:

<https://github.com/Fengkp/WGU-Capstone-Project>

1. Click ‘launch’ and wait for the new page to load
2. The new page will have a ‘jupyter’ logo in the top left and contain a listing of all the files within the repository
3. Click on ‘Main.ipynb’ and a new page will open containing our jupyter notebook and the source code
4. On the toolbar there will be a button labeled as ‘Voila’, click on that button
5. A new page will open once again presenting the dashboard for our application

You can now navigate through the dashboard. It contains several representations of the data including tables, charts, and graphs. Further down the dashboard, you can test functionality of the application by doing the following:

1. Underneath the table of the projections, there will be to text boxes
2. In the first text box labeled ‘Test Score:’ type in a score between 0 and 100
3. Press enter/return on your keyboard after you have typed in a number and a predicted score will be generated

## Summation of Learning Experience

There were a few concepts throughout the project that I have encountered in previous courses. In particular, a few projects required that I look into the different software development models. This made choosing the appropriate methodology a lot easier.

This project was very challenging in comparison to my previous assignments. Some of them were written in Python as well, but not as extensively as for this application. There were several topics regarding the IDE and libraries I chose to work with that required extensive review. This should come in handy throughout my career.

More importantly, data analysis was a completely new topic to me. I have delved into machine learning for a previous project, but not in this manner. Once again, I had to do my research to better understand some of the concepts I had to encounter.

I learned a lot from different videos and websites, but most importantly, the instructor was incredibly helpful to get the project started. All the concepts and practice will be invaluable and will surely be used down the line.

# E. Sources

References

Stephens, Rod. “Chapter 12.1 Waterfall.” *Beginning Software Engineering*, by Rod Stephens, Wrox, 2015.