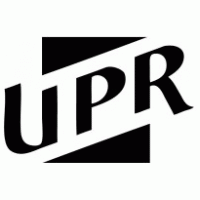
University of Puerto Rico - Mayagüez

Department of Computer Science and Engineering

**RUM Analytics Platform**

Progress Report

|  |  |  |
| --- | --- | --- |
| Irixa Vales Torres  Antonio Lugo Alicea | 802-15-8660  802-13-4140 | CIIC4151  Senior Design Project  2020/10/20  Prof. Wilson Rivera |

[**1. Introduction**](#_ht59ku6v2jq2) **3**

[1.1 Problem Statement](#_163cb5emq0og) 3

[1.2 Project Objectives](#_ktn66bb3k99g) 3

[1.3 Summary of Solution](#_fbf4uupc0wtm) 4

[1.3.1 Stakeholders / Clients](#_ts8g7yavnbwz) 5

[1.3.2 Benefits and Limitations](#_jg84ly3stxm9) 5

[1.3.3 Commercial potential](#_pl7g93tcngiz) 5

[1.3.4 Market Analysis](#_k75zcxkjgsv9) 5

[**2. Technical Description**](#_yws70vadapl0) **7**

[2.1 System architecture](#_gx9kxsegobuv) 7

[2.2 Modules to implement](#_z9oswfm5cq9o) 7

[2.3 Schemas/ER Diagrams](#_4cb0vdj1bpmo) 7

[2.4 Wire-frame for UI](#_uczu4iliic59) 8

[2.5 Realistic engineering constraints](#_nzhlv1n0tbrm) 9

[2.6 Engineering standards to be used](#_r2rj78tvslac) 9

[**3. Project plan**](#_qs6ntdnfrk5o) **10**

[3.1 Milestones](#_89y80wyf0e4b) 10

[3.2 Deliverables](#_46s5ypfd7f4b) 11

[3.3 Schedule](#_lhr9v6e7bnbw) 11

[3.4 Team Roles](#_y88jm0alvr49) 12

[**4. Performance Report**](#_u5iwy88gjxbh) **12**

[4.1 Performance Review](#_qx3lh1frv48r) 12

[4.2 Data Visualizations](#_k8hmzhnlr800) 13

[4.3 Future Performance](#_2m47te3e0lmg) 14

[**5. Project Documentation**](#_ofvu2zz4ni44) **15**

[5.1 Code repository](#_ob9z6r5hrsn5) 15

[5.2 Project Web Page](#_v2w7xt1vvf1x) 15

[5.3 Table Formats](#_kfdn2ilryigo) 15

[**6. References**](#_32ow0bqn4zi) **16**

# 1. Introduction

The lack of analytical tools in the academic sector means that identifying emerging patterns is a slow process. A student’s academic performance may be affected by events outside of the university. There have been a lot of major events in Puerto Rico in the past few years, including but not limited to, hurricanes Irma and Maria, university strikes, power outages, earthquakes, and the COVID-19 pandemic. These phenomena have impacted the island as a whole, but their effects on university students and universities themselves have scarcely been analyzed beyond an economic viewpoint. Our intention with this project is to provide an analytical tool that helps the University of Puerto Rico at Mayaguez understand how these events directly or indirectly affect its student community by analyzing data such as grades and retention rates throughout these past years. As the COVID-19 pandemic continues and online teaching becomes the norm, more data such as study habits or teaching methods can be analyzed to see what aids student performance.

## 1.1 Problem Statement

This project aims to find any relation between external events that affect students and their academic performance. What we hope to find when analyzing gpa and Twitter data is some anomaly at the time of a significant event or no change in data trends. This will also give us an insight into how the university is managing the academic affairs in such cases. This tool could be used as a resource for students to understand what kind of circumstances have been conducive to academic success for them and their peers.

## 1.2 Project Objectives

* Get access to student data for a reasonable population of UPRM students
  + Obtain information about at least 10% of enrolled students.
  + Get records from university through OPIMI dating back to 2010 until current year, and store them in a database for our use.
* Design and create algorithms to analyze data and draw conclusions useful to UPRM administration
  + Create a model that predicts how external events affect students with at least an 80% accuracy.
* Create a data visualization platform that includes reports and dashboards which showcase the analytics
  + Create a series of dashboards and reports using a visualization tool that allows users to see clear correlations in the dataset.
  + Allow platform users to view their own individual analytics, including data since the year they enrolled or since 2010 (whichever is more recent).

## 1.3 Summary of Solution

The first step for our solution is the data acquisition. We will need the university administration to provide us with the student data. Once we have access to the data we will work on cleaning and organizing it in the format that we need to analyze it. Before setting up the database, we will work to find any inconsistencies within the dataset and fix any potential anomalies. Once we know we have the data we need, we will set up the correct datatypes and add all the records in a database following the proposed schema (part 3.2). We will have a schema for a student with basic information about that student, noticing that since the data is anonymized we will not have any names or information that could expose the identity of each student. We will use a unique identifier to distinguish between them. We will also have relations from students to the courses that they have taken (including the semester they were taken and the grade they got), their gpa each semester, and program changes if any.

In the case that the data cannot be provided to us, we will proceed by creating an anonymous survey, so that a smaller population of the university’s students can fill and provide us with some data.

We also plan to include data acquisition from social media applications as part of the solution. Using the Twitter REST API, we will collect tweets that contain keywords for each of the events in our timeline and classify them by frequency (amount of times people tweeted about it and for how long were people tweeting about it) and the connotation of the messages (how negatively did it affect people).

The next step is the exploratory data analysis. We will use tools such as Pandas, Numpy, Matplotlib, and Seaborn to find patterns within the dataset and build an intuition towards possible correlations. This involves a number of methods from dimension reduction to frequency analysis in order to find correlations.

Once we have some concrete conclusions about the data, the final step is to create a series of reports and dashboards using a visualization tool, such as PowerBI, to showcase our findings. We will make this platform available to the target audience (for now, the university administration), by hosting a website that they can access. The dashboards need to be straightforward and clearly communicate the data that we want to convey. The website will also be interactive to allow the users to view correlations between the data based on input parameters.

We will also provide a platform for students to view their individual analytics.

### 1.3.1 Stakeholders / Clients

University of Puerto Rico Mayaguez Administration

University of Puerto Rico Mayaguez Students

### 1.3.2 Benefits and Limitations

*Benefits:*

* The ability to use analytics to gleam information from available student data.
* Identification of university issues affecting university staff and students.
* Identification of trends in student GPA per major.
* Identification of enrollment trends.
* Student self analysis capability via voluntary enrollment.

*Limitations:*

* Due to privacy concerns, this solution must be limited to anonymized or volunteered student data.
* Information that could personally identify someone cannot be stored.
* Finding a correlation does not equal finding a causal relationship.
* Twitter data is not standardized and instances of university issues discussions might get overlooked if search parameters don’t match.

### 1.3.3 Commercial potential

With the advent of Big Data Analytics around the world there is a growing need for Data Analysis that has not been met in Puerto Rico’s academic sector. This tool aims to fill this empty role for the university. The replication of this tool for other universities is a potential commercial goal. A tool for analyzing anonymous student data could aid universities identify and tackle the issues their students are facing. Additionally it could help identify opportunities for growth in the universities’ systems via frequency analysis of Twitter data about the university.

### 1.3.4 Market Analysis

We have found various projects involving the research of students’ academic performance at different universities. An article [4] used the methods of Classification and Association Rule Discovery to analyze the factors that are affecting students taking into account two important factors from students’ feedback questionnaires: the courses themselves that are taken by the students and the course instructor. Another study [6] used students’ backgrounds, including information such as gender, place of precedence, and level of education. Both research studies come up with a model that identifies different variables affecting students. However, no study was found that takes into consideration the phenomena that is outside the institute’s control and might affect students. The correlation between those phenomena, whether socio political events or climatic disaster, and the academic performance of students is what we intend to discover. Our focus will be on the University of Puerto Rico at Mayaguez. Market research shows there are no analytics tools available in the PR market. As such, there are no local competing services and we would be leading in this market. Additionally, no other analytics tool has provided a platform for students with an individual analysis of their data.

This technology generally sparks privacy concerns but appropriate handling of the data accounts for these. In order to properly handle this data, only anonymous or volunteered information will be stored. This helps with the second concern when handling personal data, security. Since information will be anonymized and unable to lead to a specific person, or in the case of Twitter data, publicly available, security will not be an issue.

The technologies needed for this product would be:

* Somewhere to host the analytic platform. We are taking into consideration either a cloud hosting provider or using the servers from the university.
* A database for Twitter data
* A database for anonymized student data

Intellectual property will include copyright to the code for the analytics platform itself and its associated webpages, but not the analytics held within. As per section 107 of the Copyright Act (1976), use of the platform is allowed for fair use.

# 2. Technical Description

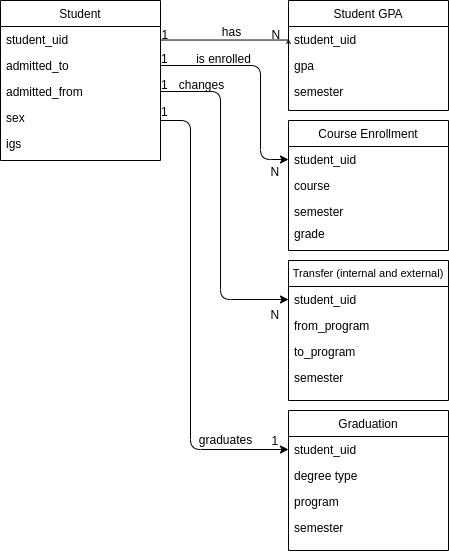
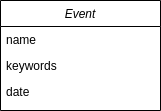
## 2.1 System architecture

## 

## 2.2 Modules to implement

1. Data collection / Data sources
   1. University records
   2. Twitter
2. Data management / Analytics Store
   1. Data storage
   2. Data formatting
3. Data analysis
   1. Analytic design
   2. Analytic storage
   3. Analytic creation
4. Data visualization / UI
   1. Dashboards and reports

## 2.3 Schemas/ER Diagrams



## 2.4 Wire-frame for UI

This is an example UI using the PowerBI [3] platform.



## 2.5 Realistic engineering constraints

1. **Maintainability** - We have chosen simple data types to save so the database can be easily maintained.
2. **Scalability** - When possible, the implementation that allows for the most scalability will be implemented. This means storing data as efficiently and standardized as we can to allow for more data influx and storage.
3. **Learnability** - UI design will be easy to use and easy to learn, with a built-in tutorial.
4. **Social** - Acquiring sufficient data might be a constraint for the project. We are in need of students who are willing to volunteer and fill anonymous surveys.
5. **Ethical** - We have chosen to use anonymized data to avoid disclosing personal information regarding the people being part of the study. When analyzing the data, results will be generalized to avoid this problem as well.
6. **Time** - Due to the semester nature of the project, it will not include machine learning methods whose learning curve is too high. This is also a possible scalability opportunity due to machine learning’s power when it comes to analytics.

## 2.6 Engineering standards to be used

* Abide by the Data Science Association’s Code of Professional Conduct [2].
* Follow the guidelines set forth in the CITI Program Course, *Data or Specimens Only Research - Basic/Refresher*.
* After thorough research, no more standards pertaining to Data Science have been found.

# 3. Project plan

The timeline for the project has been divided into sprints that are two weeks long. Each sprint will have its own milestones in order to track the progress of the project. Following an agile-like approach, we will have sprint planning meetings to allocate our time budgets appropriately and to review the past sprint for possible improvements. Additionally, we will have meetings at least once a week to update each other on progress and blockers for the project. Milestones will serve as the product backlog from which user stories are drawn. The next part details what we expect to accomplish for each sprint.

## 3.1 Milestones

- **Sprint 1**: Data acquisition & Data preparation

* Obtain access to the data sources (Currently communicating with administration to obtain access to anonymous student data).
* Set up databases for the data.
* Conduct surveys of students via social media and, if possible, institutional email to obtain data on student perceptions per semester and study habit development.
* Twitter data acquisition is a possibility.

- **Sprint 2**: Exploratory data analysis

* Apply univariate analysis, bivariate analysis, correlation analysis
* Find any inconsistencies within the data and fix them
* Identify and handle duplicate/missing data using external Python packages such as Pandas, Numpy, Matplotlib, Seaborn, etc.
* Transform the data into the correct format

- **Sprint 3**: Data modeling

* Conceptual, logical, and physical data modeling.
* Create Entity Relationship Diagrams for data.

- **Sprint 4**: Visualization and communication

* Generate reports and dashboards that show our findings using a visualization tool
* Develop a web-based User Interface for the reports and dashboards

- **Sprint 5**: Deploy and documentation

* This last sprint will focus on finalizing documentation for the project and preparing the capstone demo.

## 3.2 Deliverables

At the end of the project we expect to accomplish the following deliverables:

* Webpage that allows visualization of data.
* Analytics Platform - Allows the filtering of data via user input or pre-written analytics.
* Series of dashboards and reports showcasing the data
* Show a clear correlation between students’ performance and major social, academic, and national events

## 3.3 Schedule

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **week of:**  **(mm/dd-mm/dd)** | **sunday** | **monday** | **tuesday** | **wednesday** | **thursday** | **friday** | **saturday** |
| **09/17-09/19** |  |  |  |  | proposal due |  |  |
| **09/20-09/26** |  | sprint planning meeting |  | sync-up |  |  |  |
| **09/27-10/03** |  |  |  | sync-up |  |  |  |
| **10/04-10/10** |  | sprint planning meeting |  | sync-up |  | stand-up |  |
| **10/11-10/17** |  | sync-up |  | demo prep start |  | stand-up |  |
| **10/18-10/24** |  | demo finished | progress report and demo |  | sprint planning meeting |  |  |
| **10/25-10/31** |  | sync-up |  | stand-up |  | stand-up |  |
| **11/01-11/07** |  | sync-up |  | stand-up |  | stand-up |  |
| **11/08-11/14** |  | sprint planning meeting |  | stand-up |  | stand-up |  |
| **11/15-11/21** |  | sync-up |  | stand-up |  | stand-up |  |
| **11/22-11/28** |  | sync-up |  | stand-up | thanksgiving day |  |  |
| **11/29-12/05** |  | sprint planning meeting |  | stand-up |  | sync-up |  |
| **12/06-12/07** |  | final report and demo |  |  |  |  |  |

## 3.4 Team Roles

The roles for each team member on each specific sprint will be assigned on the corresponding sprint planning. The following table shows the task division for sprint 1.

|  |  |  |
| --- | --- | --- |
|  | **Antonio Lugo** | **Irixa Vales** |
| Sprint 1 | Create a survey for students to fill | Talk with administration to acquire student data |
| Set up databases | Acquire data from Twitter using REST API |
| Sprint 2 | Complete CITI HSR course | Complete CITI HSR course |
| Format data to database | Explore data to find patterns |
| Clean data anomalies | Analyze Machine Learning algorithms |

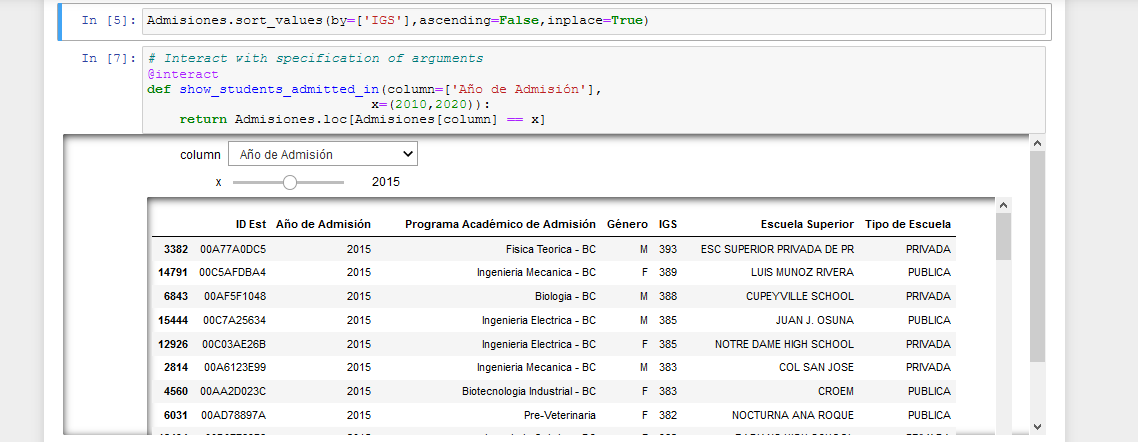
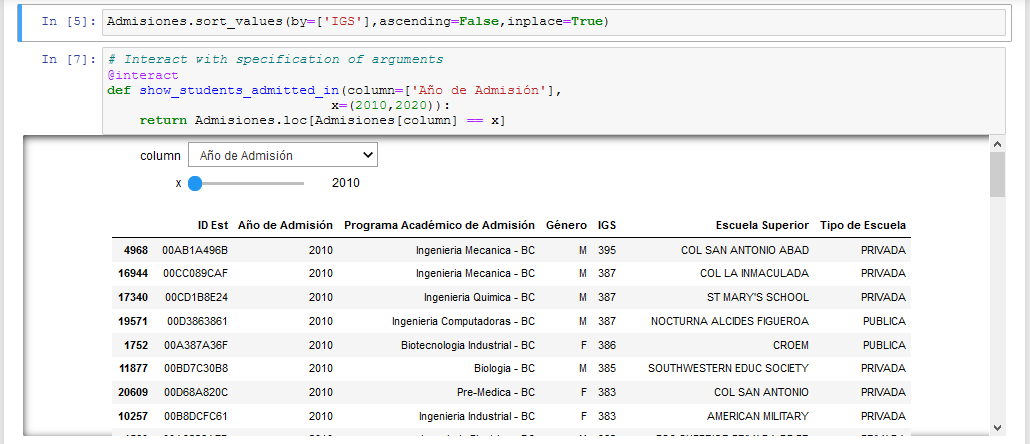
# 4. Performance Report

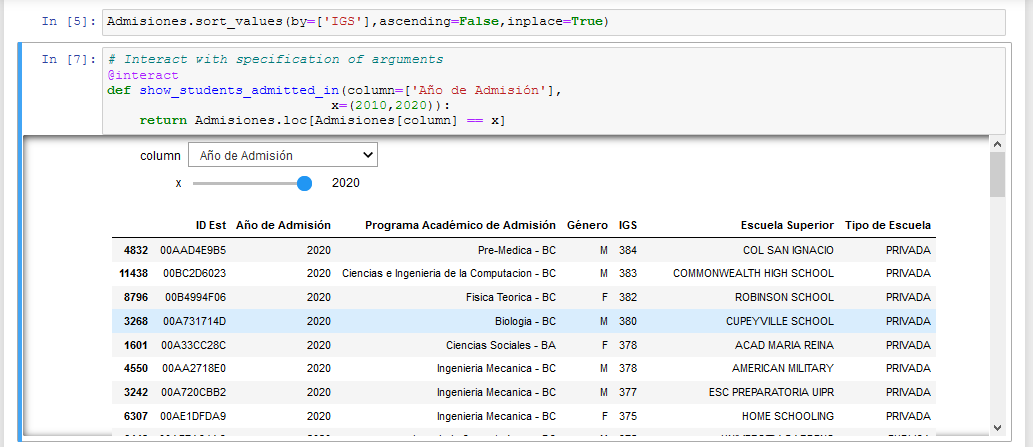
## 4.1 Performance Review

Performance for the first sprint was below expectations. We underestimated the time it would take to acquire the data we needed and learned about requirements we needed to fulfill in order to disseminate our survey. Nevertheless, the data that we requested from the institution has been successfully acquired. The Office of Planning, Research and Institutional Improvement - *Oficina de Planificación, Investigación y Mejoramiento Institucional (OPIMI)* - has sent tables for student admissions, gpa per semester, grades per class, internal transfers, and external transfers. The last two were received on October 19th, 2020. This development caused sprint progress to slow down since data is such an integral part of this project. We are now in the process of studying the data with tools such as pandas and numpy, to get a sense of how the raw data behaves. Once we understand the structure of the data we can proceed to choose the best “Classification” algorithm to apply to it. We have also received approval to disseminate the survey, so we will start to collect that data in order to integrate it.

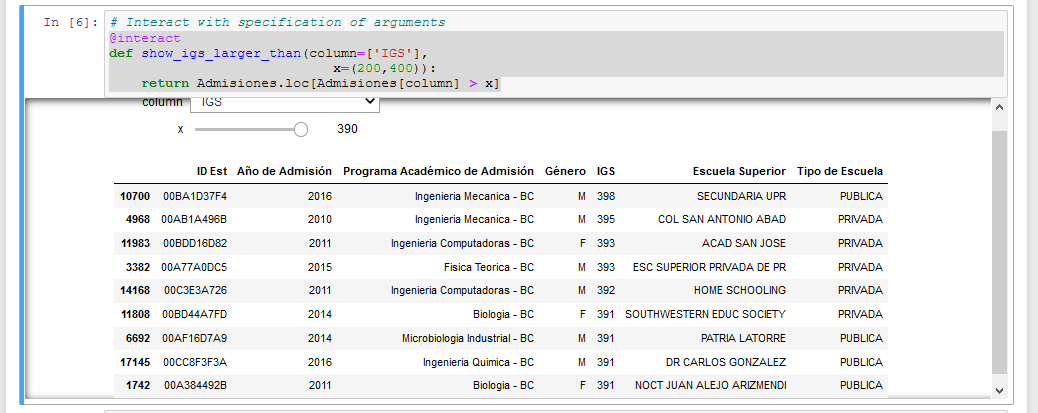
## 4.2 Data Visualizations

While graphs have not yet been implemented, by just reordering and filtering the data we have obtained it is possible to see some interesting trends. For example, here we see the top GPAs for 2010, 2015, and 2020.



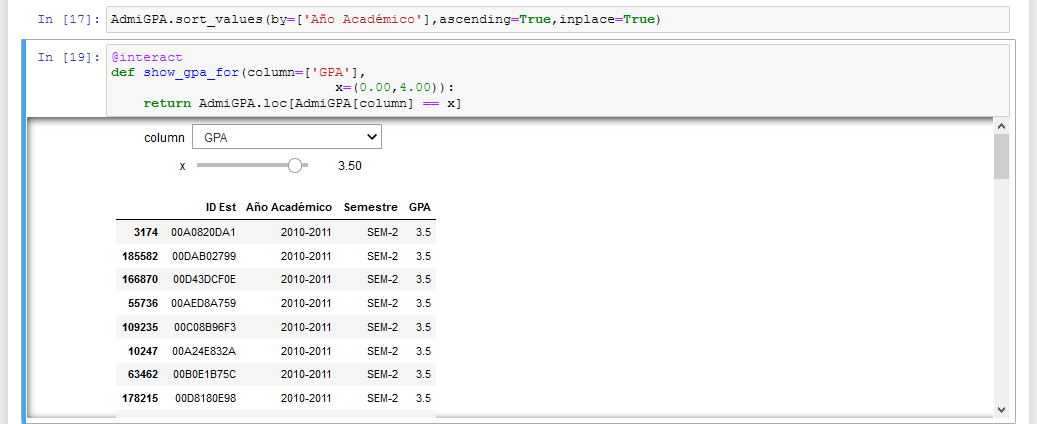


We can even look at the top GPAs admitted from 2010 to 2020.



## 4.3 Future Performance

Moving forward, we’ll be able to groom the data into more usable formats, like changing Academic Year in some tables to an integer instead of a string, so it can be transformed into a vector that will then be used to feed the model.



Changing the data formats will allow us to aggregate results like these into more meaningful formats like line charts through semesters of avg gpa per major.

# 5. Project Documentation

## 5.1 Code repository

<https://github.com/Manathonys/Capstone>

## 5.2 Project Web Page

<https://manathonys.github.io/Capstone/>

## 5.3 Table Formats

**Admissions**

**SID:** Unique ID following hex values

**AdmissionYear:** Int value between 2010-2020

**Program:** String with name and type of program the student enrolled in

**Gender:** Gender student identified as upon admission (only has M-F values, doesn’t account for non-binaries)

**IGS:** Int value between 200-400

**HighSchool:** String with name of school

**SchoolType:** String with type of school (PUBLICA or PRIVADA), could be a boolean or 0/1 given lack of homeschool data

**Grades**

**SID:** Unique ID following hex values

**AcademicYear:** {int - (int+1)}, can be simplified to first int and/or mixed with Semester for more granular searching

**Semester:** (SEM1) or (SEM-2), could be boolean or 1 / 2

**Course:** String with course code {CODE 0000}

**Grade:** Grade obtained, {A,B,C,D,F,W}

**GPA**

**SID:** Unique ID following hex values

**AcademicYear:** {int - (int+1)}, can be simplified to first int and/or mixed with Semester for more granular searching

**Semester:** (SEM1) or (SEM-2), could be boolean or 1 / 2

**GPA:** real value with range [0.00-4.00]

**Graduations**

**SID:** Unique ID following hex values

**AcademicYear:** {int - (int+1)}, can be simplified to first int and/or mixed with Semester for more granular searching

**Semester:** (SEM1) or (SEM-2), could be boolean or 1 / 2

**Program:** String with name and type of program the student graduated from

# 6. References

1. Abaidullah, Anwar Muhammad, et al. “Identifying Hidden Patterns in Students’ Feedback through Cluster Analysis.” International Journal of Computer Theory and Engineering, vol. 7, no. 1, 2014, pp. 16–20.
2. Datascienceassn.org. 2020. *Code Of Conduct | Data Science Association*. [online] Available at: <https://www.datascienceassn.org/code-of-conduct.html> [Accessed 16 September 2020].
3. "Data Visualization | Microsoft Power BI". Powerbi.Microsoft.Com, 2020, <https://powerbi.microsoft.com/en-us/>.
4. Nahid Hajizadeh, Marzieh Ahmadzadeh. Analysis of Factors That Affect Students’ Academic Performance - Data Mining Approach. 31 Aug. 2014.
5. Sachin Bhoite, Suresh Pathare. Students’ Academic Performance Analysis Using Data Mining Techniques in Higher Education: A Case Study of College. 25 Feb. 2010.
6. Sana, Isma Farrah Siddiqui, Qasim Ali Arain. ANALYZING STUDENTS’ ACADEMIC PERFORMANCE THROUGH EDUCATIONAL DATA MINING. 17 May 2019.