Educational Data Mining

Natalie Souaid, Dr. Bruce Shapiro, Dr. Carol Shubin
California State University, Northridge

Abstract

This is an 'Educational Data Mining' (EDM) project for which we investigate the impact of taking mathematics courses in the final year of high school on student success in college. This sort of inquiry is achieved by looking at vast amounts of data - using data mining techniques - on a student admission and records database from California State University, Northridge (CSUN) to pull out relevant information for analysis. Specifically, we hope to identify a relationship between students who took a math class in their senior year of high school with their scores from either the Scholastic Aptitude Test (SAT) or the Entry Level Mathematics (ELM) placement exam. Using these results we can draw conclusions and continue to investigate these effects to better understand what can be done to improve learning outcomes for underrepresented students in the STEM fields in college.

Figure 1. Software Logos

Introduction to Educational Data Mining

The challenges that face higher education today are numerous but improving student learning outcomes is still the goal of higher learning institutions. This is where Educational Data Mining stems from. This is achieved through use of a relational repository/database that contains years of information about a given institution. Education Data Mining is an attempt to explore the vast data that is involved in the educational systems that exist today. This allows us to gain better insights about the trajectory of our schools and leverage existing information to make informed decisions.

Our goal is to identify trends about high school students and their performance in the STEM fields, specifically in mathematics. This can have a lasting impact on their success at the college-level which ultimately impacts their career choices. Educational data mining has the potential to drive significant change in the areas of improving student learning outcomes for improved confidence and ability to be successful in higher education. This may help identify what support is needed to keep students active and encourage students (especially those of whom are underrepresented in STEM) to stay included in the higher learning community.

Materials

The data information came from a repository of California State University, Northridge student information from years 2014 and 2016. There are several database management systems (DBMS) that we utilized to extract relevant information. The two main DBMS were MySql Workbench and DB Browser for Sqlite. We used the Structured Query Language (SQL) to do these tasks.

For the data manipulation we utilized the Jupyter Notebook platform to facilitate our use of Python code for more efficient testing, and many Python libraries, including OpenCV and Numpy.

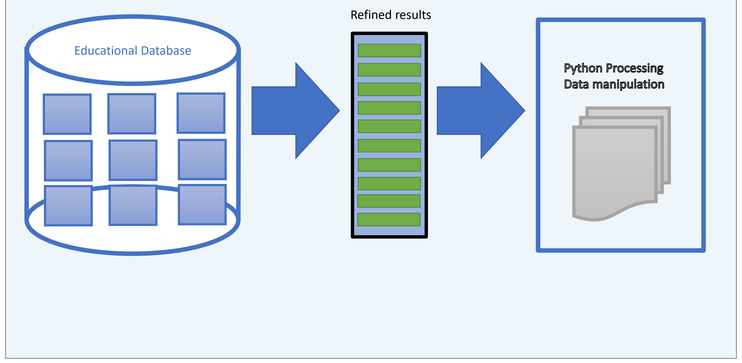

Figure 2. Diagram of Methods

Table 1. Results

Results

Elm values
Score: 35 Percentage: 0.6294840294840295
Score: 40 Percentage: 0.511145856600402
Score: 45 Percentage: 0.3485369667187849
Score: 50 Percentage: 0.26522224704042885
Score: 55 Percentage: 0.14451641724368997

SAT values:
Score: 450 Percentage: 0.9975728155339806
Score: 500 Percentage: 0.9975728155339806
Score: 550 Percentage: 0.9902912621359223

Methods

The first step in this project was to perform data preprocessing and data cleaning. This is achieved by formatting the raw data in a structured way. There are many fields and attributes for every entity so ensuring that the data has proper relationships allows for a more intuitive and seamless data extraction process. The next part of the process is performing SQL queries on the database. We use programming logic based on set theory to form valid queries. This is a consuming exercise as there are a lot of constraints that we must be aware of when developing them. Once we have the results from the queries, the data is in a form that needs to be parsed. The refined results are still not yet clearly broken down into its meaningful parts. This is achieved by using Python and OpenCV. Our goal for this task was to determine the percentages of students who took a math course in their terminal year of high school and compare either their ELM or SAT results with students who did not take a mathematics course that year.

```
#  store a score from each sid
#  there are multiple, just pick one "randomly" to compare the others with
for i, r in distinct.iterrows():
   if r['exam'] == 'SAT':
      mydictionary[r['sid']] = r['score']

#  go through each sid and check if there is a score in the dictionary.
#  compare with that score and keep the highest score
for i, r in tests.iterrows():
   if r['exam'] == 'SAT':
      if isinstance(mydictionary.get(r['sid']), int):
         if(r['score'] > mydictionary.get(r['sid'])):
            mydictionary['sid'] = r['score']
```

Figure 3. Code Snippet for SAT

Discussion

The results represent the students who have taken a math class in the 12th grade. Each score accounts for all students scores that reach that value threshold. As we continue our research, we hope to compare these with those students who have not taken a mathematics class in the 12th grade. For future work, we also look to specify the types of math classes that are taken during that terminal year. In order to gain more clarity on the significance of these results it would be interesting to further look into the demographics of the students. Longer term we hope to use more years from the repository for more accurate insights. The years 2014 and 2016 were valuable but including more years would be improve accuracy and outlook.

Contact natalie.souaid.402@my.csun.edu
bruce.e.shapiro@csun.edu
shubincarolann@gmail.com

References

Angotti, Robin, and Karen Rosenberg. "Strategic Collaboration for Richer Assessment: Educational Data Mining to Improve Learning Centers." The Learning Assistance Review, vol. 23, no. 2, 2018, p. 115.

Bakhshinategh, Behdad, et al. "Educational Data Mining Applications and Tasks: A Survey of the Last 10 Years." Education and Information Technologies, vol. 23, no. 1, 2018, pp. 537–553.

Tsiakmaki, Maria, et al. "Implementing AutoML in Educational Data Mining for Prediction Tasks." Applied Sciences, vol. 10, no. 1, 2020, p. 90.