

MATH 189 Project Proposal

Leena Kang, Ethan Lin, Krystal Qiu, Mackenzie May, Samuel Mai

In an ideal world, every student has an equal chance to be successful in their educational pursuits. Though we see that high school dropouts in the United States have declined in the past decade, we can leverage data to analyze the leading causes of dropouts in order to make better informed solutions to ultimately reduce dropout rates. With this project, we hope to identify which factors are significantly correlated with academic success. These findings, if the factors are controllable, would be able to provide future students (and/or) parents with guidance on positive actions that can be taken to help level the playing field. In addition, the findings can help to reduce school dropout rates, by influencing education policy makers to make adjustments to policy, and informing parents how to identify if their children are at the risk of dropping out of school.

The data can be sourced from higher education institutions, which provide information about students' academic paths and their academic performance throughout their enrollment. For our project, we decided to use a dataset adopted from the [UCI Machine Learning Repository](#) that includes data containing the academic path, demographics, and socio-economic status of each student along with their academic performance. Each row represents a student, and records their academic status at the end of semester as dropout, enrolled, or graduate. With this, we intend to investigate whether any discrepancies between 'successful' and 'unsuccessful' students are statistically significant. Our aim is to determine whether we can anticipate students who are at risk of poor academic performance early on based on academic, demographics, and socio-economic factors, and potentially suggest data-driven, optimal solutions to reduce dropout rates. We also plan on performing various hypothesis tests to identify any potential biases in the dataset itself in order to address potential drawbacks/imbalances of our data. Because of the socioeconomic factors that follow education, we would expect the data to suggest that children from lower income families have a higher proportion of dropout students, however, hypothesis testing would be necessary to determine whether these differences are significant. Additionally, as the dataset was created in Portugal, we would expect to address potentially ethnic imbalances of the data that could affect our findings/conclusions.