Final Presentation

Emmanuel Paalam, Charles Smith, Everett Williams

Overview

- High school dropout remains a significant concern, often viewed as a cumulative process influenced by a wide range of academic, psychosocial, and sociodemographic factors.
- This study utilizes data from the Educational Longitudinal Study of 2002 to explore whether a high-dimensional and domain-diverse set of these factors can be used to naturally generate distinct profiles of students who drop out versus those who do not between their sophomore and senior years.

Summary

- An unsupervised clustering approach was employed on a sample of 14,654 students using 39 exploratory variables recorded during participants' high school years.
- To address a significant class imbalance in dropout status, the Density-Based Spatial Clustering of Applications with Noise algorithm was selected for clustering.
- The data underwent a rigorous preprocessing pipeline, including iterative imputation for missing values, mixed-data encoding, and Principal Component Analysis for dimensionality reduction.

Conclusion

- The key takeaway is that based on the selected explanatory variables we could not find a pattern to effectively isolate dropouts.
- The majority of students, regardless of dropout status, were grouped into a single massive cluster. This result strongly suggests that the exploratory variables used, while logical and in line with the literature, do not contain strong enough signals to naturally separate students into groups that align with their dropout status using this unsupervised method.