Student Dropout versus Non-dropout Profiling via DBSCAN on the Base-Year to First Follow-up ELS:2002 Cohort

Emmanuel Paalam[1], Charles Smith[2], Everett Williams[1]

[1]Department of Computer Science, University of West Florida, FL, USA
[2]Department of Cybersecurity, University of West Florida, FL, USA

E-mail: ejp25@students.uwf.edu, cts38@students.uwf.edu, ebw13@students.uwf.edu

Abstract

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. An unsupervised clustering approach was employed on a sample of 14,654 students using 39 exploratory variables recorded during a sector of the survey which occurred 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. The DBSCAN analysis produced one large, dominant cluster that contained the vast majority of both dropout and non-dropout students, with a smaller group of outliers. This result suggests that the selected literature-informed variables, while diverse, do not contain strong enough signals to naturally separate students into groups that align with their dropout status using this unsupervised method.

I. Introduction, Literature Review

While a reduction in national dropout rates has been seen in the last decade (OECD, 2010), the issue of dropout itself remains a force to be reckoned with for academics, primarily for the sake of minimizing not only dropout rates itself for the sake of the workforce but in-vain spending of financial and general resources for educators and their program administration.

Spady (1971) and successively Tinto and Cullen (1973) and Pascarella (1980) have proposed different dropout process models which view the occurrence as a *temporal, modular sequence*, generally composed of various conditions influencing a student's dropout decision. Considered conditions may include but are not limited to student background characteristics, institutional

factors, and social experiences, suggesting the importance of high-dimensional profiling of students regarding investigation of dropout.

While these texts provide models of domain knowledge to consider emphasizing when developing methods that can assess or predict dropout, they remain limited in their applicability as college-level analyses. Nevertheless, literature on high school dropout mirrors these papers' views on dropout causation: Dupéré et al. (2015) in fact critique Tinto's model among various others in favor of one with more emphasis on precipitating and contextual long-term factors, and Kearney and Phillip (2016) present a proportional relationship between lower socioeconomic background and dropout rates. Generally, dropping out is observed as a cumulation of various factors that can play into academic disengagement, ranging from personal resources to academic performance to sociodemographical factors (Alexander et al., 1997).

III. Problem Description

The Educational Longitudinal Study of 2002 (ELS:2002) provides trend data about critical transitions experienced by students as they proceed through high school and into postsecondary education or their careers. A 2002 sophomore cohort was followed and surveyed at multiple intervals to collect policy-relevant data about educational processes and outcomes, including a variety of academic and non-academic information. The study was conducted in several waves, including a base year collection of data in 2002 and first follow-up resurvey of students, who were then either high school seniors or had left school. Our project leverages this dataset to explore the factors associated with early high school withdrawal and to identify distinct characterizing profiles of students who leave or stay in school between their sophomore and senior years.

The research goal for this project is to focus on dropout at the secondary level of education, particularly through an attempt at analysis and cluster modeling of students who do versus don't drop out of schooling at this level. We distinguish the two with the column F1EVERDO, which reports a student's dropout status as of the 2004 follow-up as a binary variable (depicted below). This approach allows us to use various other factors reported at base-year to understand the different circumstances and characteristics associated with leaving education during or after completing high school. Observing F1EVERDO as this study's "class" or "split" variable, secondary-level dropouts make up approximately 7.35% of the survey sample.

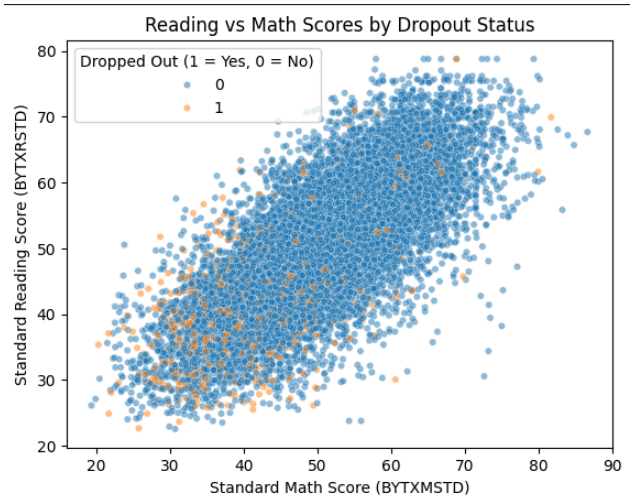
There are 4,012 attributes recorded in the ELS:2002 overall, many of which either come from irrelevant follow-ups conducted in the study or aren't valuable information, such as unnecessary ID variables or attributes that compare other attributes to scales external to the study. Thus, a subset of 42 variables—F1EVERDO included—were selected for this project in order to encapsulate the range of "life domains" emphasized by the literature while moderating the project's computational cost.

Life Domain/Purpose	Variable Name(s)
ID variable (not used in profiling)	STU_ID
1st follow-up dropout status (target)	F1EVERDO
1st follow-up panel weight	F1PNLWT
Academic Background and Performance	BYTXRSTD, BYTXMSTD, BYSCHPRG, BYGRDPT, BYS33A, BYS33D, BYS33E, BYS33H
Student Behavior and Engagement	BYHMWRK, BYXTRACU, BYS24A, BYS24B, BYS24C, BYS38C, F1S31
Psychosocial & Attitudinal Factors	BYSTEXP, F1STEXP, BYSES1, BYS89A, BYS89I, BYS89E, BYS89N, BYS20A, BYS20B, BYS20J, BYS20K
Peer Group Factors	BYS90B, BYS90D, BYS90F, BYS90L, BYS91
Family, Home, & Financial Factors	BYINCOME, F1MOTHED, F1FATHED, F1OCCUM, F1OCCUF, BYFCOMP, F1FCOMP, BYHOMLNG

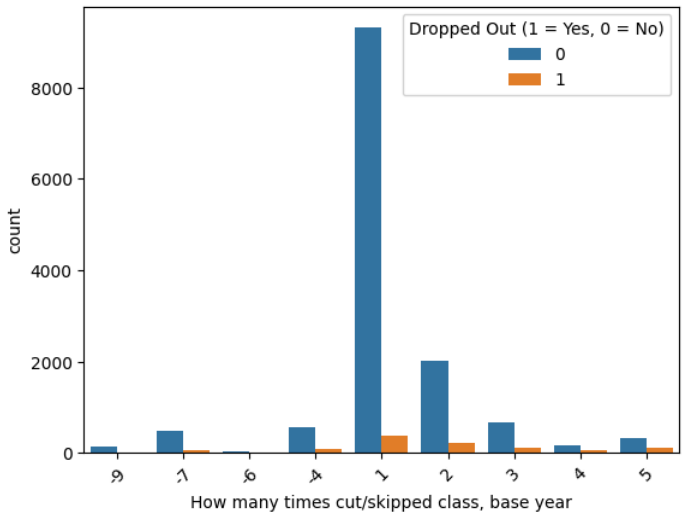
Table 1. *Variables selected for study. Individual documentation for each can be found here.*

With these variables in hand, we aim to test the reliability of utilizing a diverse surplus of student information in generating distinct dropout and nondropout student profiles at the secondary level. Profiling is distinct from the concept of classification, which is generally used to refer to supervised methods of predicting pre-defined categories or labels; neither was this project's intention to discern the most powerful predictors of student dropout or nondropout by the end of high school. Rather, the intention was to take the selected variables and determine how well they are able to form distinct groups, ideally groups which differentiate between dropouts and non-dropouts *without* the guidance of pre-defined labels—this is synonymous with unsupervised clustering, which is the approach taken here.

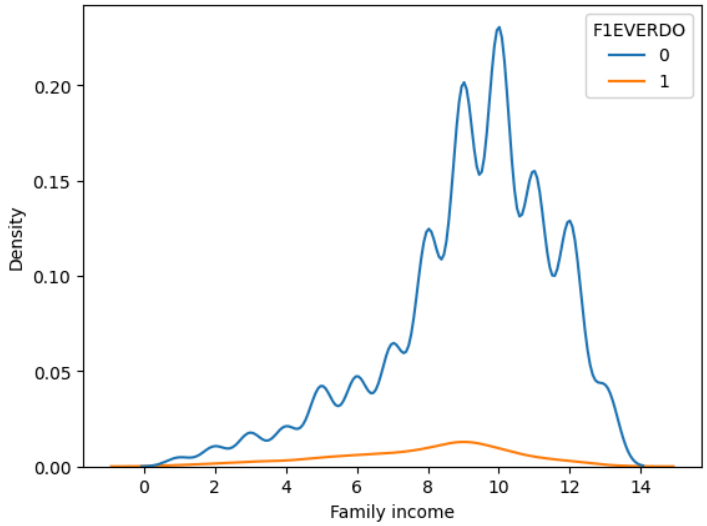
IV. EDA



This scatterplot compares students' math and reading scores highlighting their dropout status Most students cluster around the mid-range scores, showing a positive correlation between math and reading performance. Students who eventually dropped out appear more concentrated in the lower score ranges, while students who remained typically had higher scores. This suggests that lower academic performance may be associated with a higher likelihood of dropping out.



This bar chart illustrates how often students skipped classes during the base year and how the behavior relates to dropout status. The majority of students reported never skipping class, showing a strong peak in that category. Whereas dropouts while there are considerably less of them are still proportional to non dropouts as they follow similar patterns.



This density plot shows the distribution of family income food students based on their dropout status. Non-dropouts have a much higher density, with a peak around income levels 8-10, suggesting that most students who stayed in school came from moderate to higher-income households. Dropouts appear consistently lower in density but they still follow the same pattern as non-dropouts.

V. Methodology

Different preprocessing steps were taken beyond subset selection in order to prepare data for analysis and modeling. To begin, student records were removed from the raw data, along with the significant cut down on columns, in order to ensure the sole analysis on participants who were present at both base year and first follow-up data collections. This decreased our number of student records from 16,197 to 14,654, with secondary-level dropouts now making up approximately 6.74% of the survey sample.

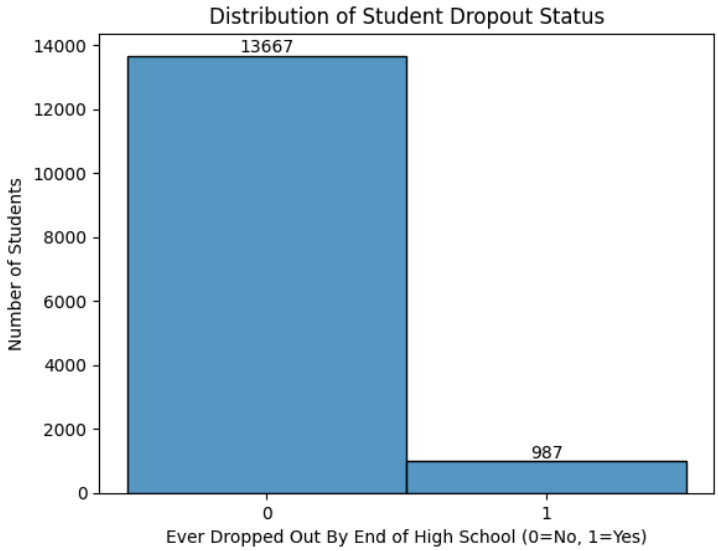


Fig. 1. *Histogram of dropouts vs. non-dropouts among sub-sample of BY to F1 present students*

Beyond attribute and sample filtering, the appropriate preprocessing steps were taken in order to prepare the data for the chosen clustering method. As F1EVERDO was used as the "target" variable for this project, it can be treated as the variable by which "actual" clusters are defined; for this reason, certain clustering techniques may not be efficient for modeling due to cluster size differences (Jain, A. K., 2010). Ester et al. (1996) provide strong support for selecting Density-Based Spatial Clustering of Applications with Noise (DBSCAN) over center-based methods when dealing with clusters of varying sizes. DBSCAN effectively identifies entire clusters regardless of their size.

To prepare the data for DBSCAN, the first issue to be addressed was the presence of missing values, which were originally stored as select negative values. Due to the incorporation of mixed data types (numerical and categorical), standard imputation methods like mean or mode replacement were avoided in favor of sklearn's IterativeImputer, which predicts missing values in rows based on the row's non-missing features—this imputer is based on the Multiple Imputation by Chained Equations principle (van Buuren & Groothuis-Oudshoorn, 2011).

After imputation, the data was encoded to account for the fact that DBSCAN depends on the mathematical distance value between variables, which cannot be clearly deduced from raw mixed data. All categorical variables were first converted into a numerical format using ordinal and one-hot encoding; numerical variables were also standardized (Pedregosa et al., 2011). However, this encoding process significantly increased the data's dimensionality, which can impair the effectiveness of distance-based clustering. Principal Component Analysis (Pearson, 1901; Shlens, 2005) was applied to reduce the feature space from 96 columns to 36 while retaining 95% of the original data's variance, ensuring a more robust and computationally efficient clustering process.

VI. Results

A common heuristic for high-dimensional data is to require twice the number of features in the data being clustered to form a dense region or cluster. After imputation and before preprocessing, there were 39 features in the data to use for modeling, excluding F1EVERDO, STU_ID, and F1PNLWT; thus, core points were required to have at least 78 points, including itself, within its epsilon (ϵ) or maximum radius within which another point can be considered a “neighboring” point to it.

ϵ itself was estimated using a k-distance graph: each point’s distance from its 78th nearest neighbor was calculated, and these distances were plotted on a graph, with the distance at the graph’s “elbow” or area where its curve indicates increasing distance between points used to determine a useful ϵ value. Two curves are actually seen in the k-distance graph resulting from the clustering data. Given that the graph plots the distances for all points on its horizontal axis, the rightmost elbow would represent the threshold separating the dense core points (the flatter part of the curve) from the sparse noise points (the steep part of the curve). The epsilon value at this elbow, which can be found on the rightmost axis, was used for DBSCAN, which was approximated to be around 5.75.

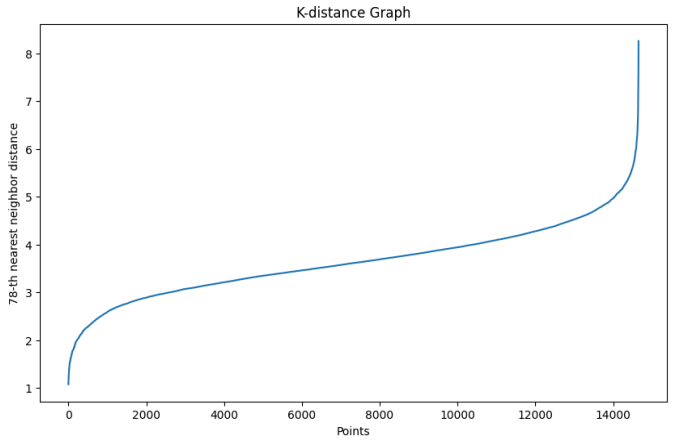


Fig. 2 *K-distance graph representing density of clustering data*

With the selected ϵ and required number of neighbors, DBSCAN produced one large primary cluster (labeled 0 by default) and a smaller group of outliers or "noise" points (labeled -1 by default). A weighted crosstabulation—utilizing the row-by-row values from F1PNLWT—was used to analyze the estimated population distribution. Looking at the primary cluster, 13,466 students from the ELS:2002 sample who didn't drop out at the secondary level were grouped with 945 who did; 201 non-dropouts and 42 dropouts were left out of the cluster and treated as noise. This accounts, when estimating clustering performance for the 2002 sophomore population, for 3,023,941 non-dropouts and 300,187 dropouts being clustered together with 48,958 non-dropouts and 15,376 dropouts being excluded, based on the provided data to formulate clusters off of.

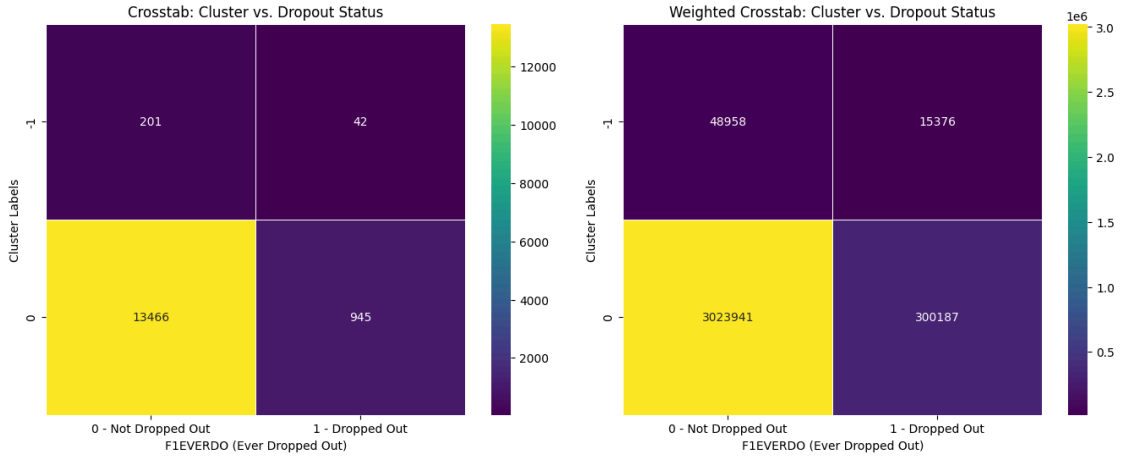


Fig. 3 *Confusion matrices (unweighted, weighted) demonstrating cluster performance*

The key takeaway is that the DBSCAN algorithm, based on the selected exploratory variables, 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 suggestions made by literature to include a variety of academic and non-academic factors, do not contain strong enough signals to naturally separate students into groups that align with their dropout status using this unsupervised method.

VII. Group Reflections

a) Everett Williams

Contributions

Much like everyone else my initial contributions come from our discussions in deciding on our chosen problem. This process was very short as Emmanuel had mentioned a dataset he has had previous experience with, and the thought of looking at drop rates interested me due to my love of education. My work started with the creation of the github project board and creating broad cards based on the rubric of the project. The split of work was between modeling and data analysis. Charles and I took on data analysis and graph generation. Shortly after our first meeting I took some time to look through our chosen variables, and the early notebooks from class and decide on what graphs to make. I landed on a scatterplot comparing standard math and reading scores with the hue selected as the dropout variable, a violin plot built off of how often a student skipped class, and a graph comparing family income of dropouts and non-dropouts. Later on the violin plot was tossed in favor of another. Concluding the data analysis, I created the slideshow from the report we had finished and eventually recorded the video presentation alongside Charles.

Key Lessons Learned

My initial thoughts on the project based on just personal experience and anecdotes, was that it would be rather easy to characterize a drop out from a non-dropout based on the variables selected. While creating the graphs for the data analysis, I noticed three things. Dropouts typically tested lower in standard tests, which is what I would have expected. Next, while there wasn't a major difference in how often dropouts skipped class, there was a noticeable increase in them not having an answer for that question, for one reason or another. Finally that family income distribution was no different for each group. Despite dropouts making up a significantly less amount of the group their family incomes were relatively the same, which goes against the typical stereotype I had in mind. Going through the modeling work next, the algorithm chosen (DBSCAN) produced one massive cluster with very few outliers. Suggesting to us that the variables chosen were not enough to characterize someone as dropout or non-dropout. This was a fascinating find for me, as I initially thought this would be a relatively easy thing to be able to classify.

b) Emmanuel Paalam

Contributions

Having previously worked with the ELS:2002 dataset, I felt that it would be ideal for the sake of progressing through the final project for us as a group to work with alongside my experience and understanding of the variables it contains as context for our findings. I also had ideas on what we could do with the dataset itself, knowing that it contained particular variables that we could test hypotheses and presumptions with. With this in mind, I talked with my group and solidified a plan to grab a particular range of student attributes based on literary recommendations from preexisting studies and theories, hence why a range of academic and non-academic factors was preserved in data subset selection. After confirming that my group felt good about my work, we continued on with our research interest in student dropout (particularly at the high school level) and its relationship with the attributes chosen for our project.

I particularly worked on the data modeling portion of the project. Because our project was focused on the potential of our chosen attributes in generating some distinction between dropouts and non-dropouts, I felt that clustering was the best means of modeling in order to determine this potential and explored various algorithms to determine which can be utilized most effectively and confidently in the context of the data being used. Due to the reality of imbalance in representation between high school level dropouts and non-dropouts, I determined that Density-Based Spatial Clustering of Applications with Noise was the most appropriate algorithm to configure and report findings off of as it is able to handle differently-sized clusters.

Key Lessons Learned

The main finding, which is that the selected attributes fail to create a sufficient natural distinction between dropouts and non-dropouts, surprised me in the context of what one might expect based on the literary claims reviewed for this paper. I and my teammates were expecting at least some ability to distinguish, even if not sufficiently, between dropouts and nondropouts even without explicit labelling or through supervised means, and I as well found it surprising that the patterns and clustering results actually indicate a complete indistinguishability between the two, at least in respect to the observed population.

In the meantime, however, I also was able to learn a lot about clustering algorithms and how they work, particularly from the comfort of working with a dataset I already understood very well. With this ease, I had the chance to really reference my own notes and notebooks taken from this course and apply them to my portion of the project. This provided me the opportunity to easily come into practical contact with intelligent means of unsupervised learning, which was not a means of modeling I've had to work with in previous analyses of the ELS:2002 dataset.

c) Charles Smith

Contributions

My initial contribution began during our first group meeting, where we discussed how to approach the project and divided up responsibilities. Like Everett, I was assigned data analysis, working to explore our selected variables and create visualizations that would help illustrate patterns between dropouts and non-dropouts. My first major task was reviewing the graphs Everett created, as well as exploring the early exploratory notebooks, to determine which visualizations would best communicate our findings.

A key part of my work was refining our approach to visualizing the survey question about how often students skipped classes. While Everett initially created a violin plot for this variable, I realized that a bar graph would be a more effective way to clearly show the frequency of skipped classes for dropouts versus non-dropouts. I created and implemented this updated visualization, which ultimately replaced the violin plot in our final analysis. This bar graph provided a more straightforward comparison of student behavior and highlighted how missing or unreported values differed between the two groups. In addition to graph creation, I also contributed to writing the Exploratory Data Analysis section of the final paper, detailing the observations we made from the visualizations and how they related to our research question. Toward the end of the project, I helped record the video presentation alongside Everett, ensuring that our analysis and findings were presented clearly and cohesively for the final submission.

Key Lessons Learned

This project reinforced the idea that data analysis often challenges initial assumptions. At the start, I expected that variables like test scores, family income, and class-skipping behavior would clearly separate dropouts from non-dropouts. But as our analysis progressed, it became clear that many of the variables we had assumed would be highly predictive didn't offer as much insight as we thought. For example, family income was nearly the same across both groups, and skipping class didn't differ dramatically either. This taught me that choosing the right variables is just as important as how you analyze them, assumptions alone aren't enough to guide meaningful analysis.

From a skills perspective, I also learned how essential it is to match the right visualization with the data. Swapping out the violin plot for a bar graph made our message easier to understand and helped us better communicate our findings to others. Working on both the EDA and the final presentation helped me develop a better sense of how to turn complex analysis into a clear and compelling story, which is a skill I'll carry forward into future projects.

References:

1. OECD (2010), *Education at a Glance 2010: OECD Indicators*, OECD Publishing, Paris, https://doi.org/10.1787/eag-2010-en

2. Spady, W. (1971). Dropouts from higher education: towards an empirical model. Interchange, 2(3), 632-644. https://doi.org/10.1007/BF02282469

3. Tinto, V., & Cullen, J. (1973). Dropout in higher education: a review and theoretical synthesis of recent research. Office of Education (DHEW), Contract OEC-0-73-1409, pp. 99. Retrieved from https://files.eric.ed.gov/fulltext/ED078802.pdf

4. Pascarella, E. T. (1980). Student-faculty informal contact and college outcomes. Review of Educational Research, 50(4), 545-595. https://doi.org/10.3102/00346543050004545

5. Dupéré, Véronique, et al. "Stressors and Turning Points in High School and Dropout: A Stress Process, Life Course Framework." *Review of Educational Research*, vol. 85, no. 4, 2015, pp. 591–629. *JSTOR*, http://www.jstor.org/stable/24753024. Accessed 19 July 2025.

6. Alexander, Karl L., et al. "From First Grade Forward: Early Foundations of High School Dropout." *Sociology of Education*, vol. 70, no. 2, 1997, pp. 87–107. *JSTOR*, https://doi.org/10.2307/2673158. Accessed 19 July 2025.

7. Kearney, Melissa S., and Phillip B. Levine. "Income Inequality, Social Mobility, and the Decision to Drop Out of High School." *Brookings Papers on Economic Activity*, 2016, pp. 333–80. *JSTOR*, http://www.jstor.org/stable/43869027. Accessed 19 July 2025.

8. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

9. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, & U. M. Fayyad (Eds.), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)* (pp. 226-231). AAAI Press.

10. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

12. Pearson, K. (1901). LIII. *On lines and planes of closest fit to systems of points in space*. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

13. Shlens, J. (2005). *A tutorial on principal component analysis*. Salk Institute for Biological Studies.