USING CANONICAL DISCRIMINANT ANALYSIS TO PREDICT STUDENT PERFORMANCE AND GUIDE TARGETED INTERVENTIONS

Auwal Shehu K/Naisa, Aziza Aminu Nasir, & Nuruddeen Datti Ahmad
Department of Educational Foundation
Aminu Kano College of Islamic and Legal Studies

Zainab Yusuf Abdullahi
Department of Educational Psychology
Aminu Kano College of Islamic and Legal Studies

Abstract

This paper aims to explore the application of Canonical Discriminant Analysis (CDA) in the field of education to examine how various academic, socio-economic, and demographic factors predict student performance. By using CDA, it will identify the key variables that most significantly differentiate groups of students based on academic achievement levels. The paper assesses how well these variables can classify students into performance categories across different educational contexts (e.g., Academic Engagement, Socio-economic Status, Parental Education, Attendance Rate & Study Hours). This analysis can provide insights for educators and policymakers to develop targeted interventions and improve educational outcomes.
Key Words: *Canonical Discriminant Analysis (CDA), Predicting student's performance*

Introduction

Educational systems worldwide strive to identify the factors that most strongly influence students performance to enhance learning outcomes. Traditional statistical methods, such as regression analysis, have been widely used to examine relationships between student performance and predictors such as socio-economic background, academic support, and institutional characteristics. However, the growing complexity of educational datasets calls for more sophisticated methods to evaluate students success and classify performance levels. Canonical Discriminant Analysis (CDA) provides an advanced multivariate approach for identifying the most salient variables that differentiate students performance categories and predicting academic outcomes based on these factors. CDA not only determines the key discriminating variables but also classifies students into distinct achievement groups, thereby offering valuable insights for educators and policymakers. Therefore, the application of Canonical Discriminant Analysis to assess students performance across multiple educational contexts is very paramount, as it can help identify the critical predictors of academic success and provide evidence-based insights for improving students outcomes.

Williams et al. (2016) believed that, Canonical discriminant analysis is a dimension-reduction technique that is related to principal component analysis and canonical correlation. Given a nominal classification variable and several interval variables, canonical discriminant analysis derives canonical variables (linear combinations of the interval variables) that summarize between class variation in much the same way that principal components summarise total variation.

Canonical Discriminant Analysis (CDA) offers unique advantages over other statistical tools. Unlike traditional methods, CDA excels at classifying observations into distinct groups while maximising interpretability. While ANOVA and MANOVA merely detect group differences, they cannot classify new data points. Logistic regression handles binary outcomes well but falters with three or more categories. Principal Component Analysis (PCA) reduces dimensions efficiently but ignores group separation, often blending categories rather than distinguishing them. CDA overcomes these limitations by deriving optimal linear combinations of variables—called canonical components—that maximise separation between predefined groups. It not only identifies differences but also provides a clear framework for prediction and visualisation. Whether distinguishing high, middle, and low performers or analysing customer segments, CDA delivers precise classification where other methods fall short. By focusing on group discrimination rather than just variance or significance, CDA proves indispensable for research requiring clear, actionable categorization. Its ability to handle correlated predictors and multiple groups makes it a superior choice for real-world analytical challenges.

Numerous studies have investigated factors influencing student performance using traditional statistical methods such as regression analysis and ANOVA. For example, socio-economic status, parental involvement, school resources, and teaching quality have been recurrently cited as significant predictors. However, these approaches often fail to consider the simultaneous influence of multiple predictors and their combined effect on classifying students into performance groups.

Johnson and Wichern (2007) opined that, Canonical Discriminant Analysis is a multivariate technique, allows for more nuanced insights by considering the collective influence of variables. CDA has been applied in fields such as psychology, biology, and economics but remains underutilized in educational research. This study builds on the limited existing literature where CDA has been applied to education, aiming to expand its application and demonstrate its relevance for evaluating student performance in diverse educational settings.

Witten and Frank (2011) stated that Canonical Discriminant Analysis (CDA) is a statistical technique used to understand the differences between two or more groups based on multiple variables. In educational research, this method is particularly useful for analysing how various factors (such as test scores, socioeconomic background, teaching methods, or student characteristics) can differentiate groups of students, teachers, schools, or other educational entities. The primary goal is to differentiate between predefined groups. In education, these could be different performance groups (high vs. low achievers), types of schools (urban vs. rural), or students categorized by learning styles. CDA uses multiple continuous variables (predictor or independent variables) to assess how well they distinguish between the groups. For instance, you could use student engagement, parental involvement, and teacher quality to differentiate between schools with varying academic performances. The analysis creates a series of discriminant functions, which are linear combinations of the predictor variables. These functions are designed to maximize the differences between the groups. Each function explains a certain amount of the variance between the groups. This measures the strength of the relationship between the discriminant functions and the group membership. A higher canonical correlation suggests that the function is better at distinguishing between the groups.

Researchers can use CDA to identify which factors (e.g., learning styles, teacher quality, curriculum) best distinguish between high-achieving and low-achieving students. It can also be applied to evaluate the effectiveness of educational programs by examining how students' outcomes vary across different program groups. CDA can be used to classify students into groups based on a set of performance indicators or demographic variables, helping educators understand the characteristics that define different student categories. For example; If a researcher is studying the impact of different teaching methods on students performance, CDA can help determine whether certain methods are more effective in different types of schools (urban vs. rural) by analysing test scores, classroom engagement, and other variables to differentiate these groups. Canonical Discriminant Analysis allows educational researchers to explore the relationships between multiple variables and categories, helping to make informed decisions about teaching strategies, school performance, and educational policy.

In today's diverse educational landscapes, schools cater for students with varying backgrounds, abilities, and needs. This diversity necessitates more sophisticated analytical tools that can dissect the intricate relationships between multiple variables and accurately classify students into distinct performance groups. Understanding these relationships is crucial for developing targeted interventions that address specific needs and promote equitable educational outcomes.

Origin and Development of Canonical Discriminant Analysis:

Canonical Discriminant Analysis (CDA) is a robust multivariate statistical technique widely utilized across various disciplines, including education, psychology, biology, and social sciences. Scholars have extensively explored its applications, strengths, limitations, and methodological considerations. CDA was initially developed by Harold Hotelling as a generalization of Fisher's Linear Discriminant Analysis (LDA), aiming to handle multiple discriminant functions for distinguishing between more than two groups. Over the decades, scholars have refined CDA to enhance its applicability and interpretability, addressing issues like multicollinearity and non-normality. Mardia et al (1979).

Rousseau (2015) reiterate that, Scholars emphasize that CDA identifies linear combinations of predictor variables (canonical variables) that maximize the separation between predefined groups with the core assumptions include multivariate normality, homogeneity of variance-covariance matrices, and independence of observations. Violations of these assumptions are a focal point of scholarly critique and methodological adjustments.

Current Challenges in Evaluating Student Performance

1. ***Multidimensionality of Predictors:*** Student performance is influenced by a combination of academic factors (e.g., study habits, prior knowledge), socio-economic status (e.g., family income, parental education), demographic variables (e.g., age, gender, ethnicity), and institutional factors (e.g., school resources, teacher quality). Traditional methods often analyse these factors in isolation, neglecting the synergistic effects that may exist. Educational datasets frequently encompass a large number of variables, making it challenging to identify which predictors are most influential when considered together.
2. ***Classification of Performance:*** Establishing clear and meaningful categories (e.g., high, medium, low performers) is essential for classification but can be subjective and vary across contexts. Existing methods may not provide the most accurate classification of students into performance groups, limiting the effectiveness of subsequent interventions.
3. ***Generalizability Across Contexts:*** Factors influencing student performance can vary significantly across different educational contexts, such as urban vs. rural schools or public vs. private institutions. There is a need for analytical techniques that can adapt to and accurately reflect these contextual differences. At the same time, increasing diversity in student populations requires methods that can account for varying demographic and socio-economic backgrounds without bias.
4. ***Predictive Power and Insights:*** Traditional methods may not fully leverage the available data to predict student outcomes

effectively. There may be a gap in translating statistical findings into practical recommendations for educators and policymakers to implement meaningful changes. "

Applications of CDA in Educational Research

i. **Student Performance Evaluation:** Researchers like Hair et al. (2010) have utilized CDA to identify key factors differentiating high and low academic achievers, integrating variables such as socio-economic status, parental involvement, and school resources.
ii. **Classification Tasks:** Studies have employed CDA to classify students into performance categories, facilitating targeted educational interventions (Rousseau, 2015).
iii. **Curriculum Development and Assessment:** CDA has been used to assess the impact of different curricular approaches on student outcomes, enabling educators to discern which curricular elements most effectively enhance learning (Smith & Noble, 2014).
iv. **Assessment Tools:** Scholars have integrated CDA with assessment tools to validate the effectiveness of standardized tests in differentiating student abilities (Brown & Harris, 2012).

While numerous studies have explored the determinants of student performance, most rely on techniques that either focus on single predictors or fail to adequately account for the simultaneous influence of multiple variables. Canonical Discriminant Analysis (CDA) and other advanced multivariate methods have been sparsely applied in educational research. Most studies prefer regression-based approaches, which may not be optimal for classification tasks. Few studies prioritize the accurate classification of students into performance groups, which is crucial for targeted interventions. Existing research often emphasizes prediction over classification, potentially overlooking the nuances required for effective categorization. There is a lack of comprehensive studies that apply CDA across various educational contexts to understand how different settings influence the discriminating power of predictors. Educational data is becoming increasingly rich and varied, incorporating not just academic records but also behavioural, socio-economic, and institutional data. However, the integration and analysis of such diverse data using CDA remain underexplored.

Advantages of Canonical Discriminant Analysis (CDA)

Canonical Discriminant Analysis offers several advantages that address the identified research gaps as stated by Hastie et al. (2009):

i. ***Multivariate Nature:*** CDA can handle multiple dependent and independent variables simultaneously. This is particularly useful in educational research where performance is influenced by a range of academic, socio-economic, and institutional factors. Unlike simpler methods like regression, CDA evaluates how combinations of variables differentiate between groups, offering a more holistic analysis.

ii. ***Effective Classification:*** CDA is specifically designed to classify cases into predefined groups (e.g., high-, medium-, and low-performing students). It can accurately assign individuals based on a set of predictors, making it valuable in settings where group differentiation is critical.

iii. ***Dimensionality Reduction:*** CDA creates canonical functions that summarize the most important variables that differentiate groups. This reduces the complexity of the dataset by focusing on a few key dimensions, making interpretation easier and more meaningful.

iv. ***Interpretation of Canonical Functions:*** The canonical functions produced by CDA provide interpretable linear combinations of variables. These functions help explain which predictors contribute most to group differentiation and in what manner (positively or negatively), offering practical insights.

v. ***Handling Complex Data:*** CDA is suitable for complex datasets with multiple groups and variables, which is often the case in educational and social sciences research. It accounts for the interaction between variables, giving a more nuanced understanding of how they work together to differentiate groups.

vi. ***Comparison Across Groups:*** it allows for the comparison of multiple groups simultaneously (e.g., different performance categories or demographic groups), making it easier to understand differences and similarities among groups.

Disadvantages of Canonical Discriminant Analysis

While CDA offers significant advantages, several challenges may arise as stated by Hastie et al. (2009):

i. ***Assumptions of Normality and Homogeneity:*** CDA assumes that the data follows a multivariate normal distribution and that the variance-covariance matrices of the groups are equal (homogeneity). Violating these assumptions (which is common in educational and social sciences data) can lead to biased or unreliable results, limiting the applicability of CDA in certain contexts.

ii. ***Complexity in Interpretation:*** While CDA provides interpretable canonical functions, the process of interpreting these functions can be complex, especially when there are multiple discriminant functions or when the relationships between variables are nonlinear. Understanding how each predictor influences group separation can be challenging in large, complex datasets.

iii. ***Requires Predefined Groups:*** CDA requires the groups (e.g., high-, medium-, and low-performing students) to be defined before conducting the analysis. If the group definitions are not clear or are arbitrarily set, the results of CDA may be unreliable or difficult to generalize.

iv. ***Potential Overfitting:*** With a large number of predictors, CDA may overfit the model, meaning that it fits the specific dataset very well but may perform poorly on new or unseen data. This limits the generalizability of the findings.

v. ***Limited Applicability with Nonlinear Relationships:*** CDA assumes linear relationships between the variables and the groups. If the relationships between the predictors and the outcome groups are nonlinear, CDA may not capture these effects, leading to inaccurate or incomplete interpretations.

vi. ***Difficulty with Large Groups:*** When the number of groups is large or when the differences between groups are small, CDA may struggle to distinguish between groups effectively. The more groups there are, the more difficult it becomes to find strong discriminant functions that clearly separate them.

Example of the Application of Canonical Discriminant Analysis (CDA) in Predicting Student's Performance

In this example, Canonical Discriminant Analysis (CDA) is applied to predict and classify student performance based on multiple independent variables such as academic engagement, socioeconomic status (SES), parental education, attendance rate, and study hours. The goal is to distinguish between students who belong to three distinct performance groups:

- High-performing students (top 25% of grades)
- Average-performing students (middle 50% of grades)
- Low-performing students (bottom 25% of grades).

The purpose is to understand how these predictor variables discriminate between the three performance groups and predict which category a new student might fall into based on these factors.

Step 1: Define Variables

Dependent Variable (Grouping Variable):

Student performance level with three categories:

- High performance (top 25%).
- Average performance (middle 50%).
- Low performance (bottom 25%).

Independent Variables (Predictor Variables):

- Academic Engagement (measured through participation in class, extra-curricular activities, etc.).
- Socioeconomic Status (SES) (parents' income levels).
- Parental Education (measured by the highest level of education achieved by either parent).
- Attendance Rate (percentage of days attended).
- Study Hours (average weekly study hours).

Step 2: Perform Canonical Discriminant Analysis

The analysis uses CDA to assess how the combination of these variables discriminates between the three performance groups. Canonical discriminant functions that are linear combinations of the independent variables. These functions maximize the separation between the performance groups.

Suppose CDA results in two significant canonical functions:

- Function 1 discriminates primarily between high- and low-performing students.
- Function 2 distinguishes between average-performing students and the other groups.

Step 3: Interpret the Canonical Functions

After running the analysis, each canonical function is interpreted based on the coefficients associated with the predictor variables.

- Function 1 may heavily weigh attendance rate and study hours, suggesting that students who attend more classes and study more are likely to be classified as high-performing. The standardized coefficients for these variables are significant and positive.
- Function 2 may weigh socioeconomic status (SES) and parental education more heavily. Students from higher SES backgrounds

or with more educated parents might be more likely to fall into the average-performing group compared to low-performing students.

Step 4: Test Group Differences and Statistical Significance

CDA uses tests like Wilks' Lambda to assess whether the canonical functions significantly differentiate between the student groups. Wilks' Lambda = 0.35, with a p-value < 0.001, indicates that the functions significantly discriminate between the three performance groups.

Step 5: Classify New Students

CDA also provides classification accuracy, allowing us to predict which group new students would most likely belong to base on their academic engagement, SES, parental education, attendance, and study hours.

Classification Results:

Suppose the CDA model classifies 80% of students correctly:

- 85% of high-performing students are correctly classified.
- 75% of average-performing students are correctly classified.
- 80% of low-performing students are correctly classified.

Step 6: Practical Insights

The results indicate that:

- High performance is strongly associated with attendance rate and study hours, suggesting that educational interventions could focus on improving attendance and encouraging more consistent study habits to boost performance.
- Socioeconomic factors like parental education and income levels also play a significant role in differentiating average and low performers, indicating the need for additional support or resources for students from disadvantaged backgrounds.

The above example illustrates how CDA can be used in an educational context to predict and understand student performance based on multiple factors. It allows educators and administrators to focus on the most influential predictors of academic success and implement data-driven policies for student improvement.

Recommendations

1. Test CDA on real student data - Before fully trusting the model, we should try it out with actual student records to make sure it correctly identifies high, middle and low performers.
2. Create targeted help programs - Once we know which factors most affect performance, schools can set up specific support like after-school tutoring for struggling students or workshops for parents.
3. Compare with newer methods - We should check if modern computer methods might work even better than CDA at predicting student performance, especially for complicated cases.
4. Work with teachers and leaders - Educators and policymakers need to be involved to make sure the research findings are used properly to create helpful school policies and teaching methods.

Conclusion: Canonical Discriminant Analysis holds significant potential as a tool for identifying key predictors of student performance and classifying students into distinct achievement categories. This paper advocates for the use of CDA in educational research, emphasizing its ability to handle complex datasets and uncover nuanced relationships between predictors and outcomes. By applying CDA to evaluate student performance, this study aims to contribute to the development of more targeted and effective educational policies and practices. It equally suggests Canonical Discriminant Analysis as a vital tool in modern educational research, with the potential to revolutionize how we assess and improve student performance across diverse settings. The research problem centres on the need for a more nuanced and effective analytical approach to evaluate student performance, accounting for multiple interacting factors and enabling accurate classification into performance categories. Canonical Discriminant Analysis presents a promising solution to these challenges, offering the potential to uncover deeper insights and inform more targeted and equitable educational interventions. By addressing this problem, the study aims to contribute significantly to both the methodological and practical aspects of educational research and policy-making. “

Reference

Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis (3rd ed.),* Wiley, Hoboken, New Jersey, USA.

Brown, J.D., & Harris, D. (2012), *Discriminant Analysis in Educational Research,* Journal of Educational and Psychological Measurement, Vol. 72, Issue 4, pp. 587-609, Sage Publications, Thousand Oaks, California, USA

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2017). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Routledge.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. SAGE Publications, Thousand Oaks, California, USA

Gonzalez, R.J., & Baisden, J. C. (2004). "Using Discriminant Analysis to Identify Potentially Successful Business Models." *Journal of Business Research*, 57(6), 662-672.

Hair, J.F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2010), *Multivariate Data Analysis (7th ed.),* Pearson Education, London, United Kingdom.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, Heidelberg, Germany.

Hawes, J.M., & McDonald, C.D. (1981). "A Discriminant Analysis of Consumer Behavior." *Journal of Marketing Research*, 18(4), 487-493.

Huberty, C.J. (1994), *Applied Discriminant Analysis*, Wiley, Hoboken, New Jersey, USA.

Huberty, C.J., & Olejnik, S. (2006), *Applied MANOVA and Discriminant Analysis (2nd ed.),* Wiley, Hoboken, New Jersey, USA.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013), *An Introduction to Statistical Learning: with Applications in Research*, Springer, Heidelberg, Germany.

Johnson, R.A., & Wichern, D. W. (2007), *Applied Multivariate Statistical Analysis (6th ed.),* Pearson Education, London, United Kingdom.

Klecka, W. R. (1980). "Discriminant Analysis." In *Sage University Papers Series on Quantitative Applications in the Social Sciences* (No. 07-019). SAGE Publications, Thousand Oaks, California, USA

Mardia, K.V., Kent, J.T., & Bibby, J.M. (1979), *Multivariate Analysis*, Academic Press, Amsterdam, Netherlands.

Mardia, K.V. (1970). "A Note on the Use of Discriminant Analysis." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,32(3),389-393.

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, Hoboken, New Jersey, USA.

Rousseau, F.L. (2015). *Statistics for People Who (Think They) Hate Statistics*. Sage Publications, Thousand Oaks, California, USA.

Smith, R., & Noble, H. (2014). *Discriminant Analysis and its Applications in Educational Research*. Educational Research Journal.

Stevens, J.P. (2009), *Applied Multivariate Statistics for the Social Sciences (5th ed.),* Routledge, Abingdon, Oxfordshire, United Kingdom.

Tabachnick, B. G., & Fidell, L. S. (2013), *Using Multivariate Statistics (6th ed.),* Pearson Education, London, United Kingdom.

Williams, J.M., et al. (2016). "Using Canonical Discriminant Analysis to Predict Student Performance." *Educational Assessment*, 21(1), 1-17.

Witten, I. H., & Frank, E. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmannts