**Case study – Predicting program choice during admissions**

**Introduction**

This case study focuses on predicting program choices in the context of student admissions using Linear Discriminant Analysis (LDA). The analysis aims to predict whether a student will enroll in the BBA, MBA, or B.Tech program. By analyzing students' academic performance data, the LDA model can help in understanding and potentially forecasting program enrollment.

**Objective**

To perform Linear Discriminant Analysis (LDA) for predicting student program choices (BBA, MBA, Btech).

**Dataset link**

<https://drive.google.com/file/d/1_O-bgbAaVYvHOkkrx_x8anj_o5aNN21Q/view?usp=sharing>

**Understanding the data**

The dataset contains the information of 200 students. Following are the description of columns:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Column |  |  |  | Description | Values |
| id |  |  |  | Roll number of the student | Unique integer identifier for each student |
| Gender |  |  |  | Gender of the student | 0 = Male, 1 = Female |
| race |  |  |  | Racial background of the student | 1 = North, 2 = South, 3 = East, 4 = West |
| ses |  |  |  | Socio-economic status | 1 = Low, 2 = Medium, 3 = High |
| schtyp |  |  |  | School type | 1 = Public School, 2 = Private School |
| prog |  |  |  | Academic program enrolled in | 1 = BBA, 2 = MBA, 3 = B.Tech |
| read |  |  |  | Reading marks obtained | Numerical value representing marks |
| write |  |  |  | Writing marks obtained | Numerical value representing marks |
| math |  |  |  | Mathematics marks obtained | Numerical value representing marks |
| science |  |  |  | Science marks obtained | Numerical value representing marks |
| socst |  |  |  | Social Science marks obtained | Numerical value representing marks |

**Procedure for coding**

* Import necessary libraries
* Load dataset
* Check the number of rows and columns
* Check for missing values
* Define independent (X) and dependent (y) variables
* Standardize independent variables
* Split data into training and test sets
* Train LDA model
* Make predictions
* Evaluate model performance (confusion matrix & accuracy)
* Display actual vs predicted program choices

**Code File Link**

<https://github.com/Ishita2003M/Predicting-Program-choice-during-Admissions/blob/main/prog_LDA.ipynb>

**Interpretation and conclusion**

1. The purpose of this analysis is to predict a student's enrolled academic program—BBA, MBA, or B.Tech—based on their academic performance across core subjects, using Linear Discriminant Analysis (LDA). This method seeks to identify the linear combinations of subject scores that best differentiate between the three programs.
2. The dataset includes information on 200 students, featuring both demographic and academic performance variables. Key attributes are:
   * Demographics: Gender, race, socio-economic status, school type
   * Target Variable: prog – academic program (1 = BBA, 2 = MBA, 3 = B.Tech)
   * Predictors (X): Academic scores in five subjects:
     + Reading
     + Writing
     + Mathematics
     + Science
     + Social Science
3. The LDA model was trained using only the academic subject scores as predictors (i.e., cognitive performance factors), excluding demographic variables. The target variable was the program choice (prog), a categorical variable with three classes.
4. Model Performance:

The confusion matrix is as follows:

[[ 1 8 7] → Actual BBA

[ 0 18 9] → Actual MBA

[ 1 5 11]] → Actual B.Tech

* + The model struggles to correctly classify BBA students, with only 1 out of 16 correctly predicted.
  + MBA predictions were most accurate, with 18 correctly classified.
  + The B.Tech group also faced notable misclassification, especially confused with MBA.
  + Overall Accuracy: The model achieved 50.0% accuracy, meaning only half of the total predictions matched the true program choices.

1. Insights:
   * The moderate accuracy indicates that subject scores alone may not be sufficient to reliably distinguish between all three academic programs.
   * Overlap in performance profiles, especially between MBA and B.Tech, likely contributed to misclassifications.
   * Including demographic or behavioral variables (e.g., socio-economic status, school type, personal interests) may enhance model performance by capturing additional context influencing program choice.