VIETNAMESE – GERMAN UNIVERSITY

FACULTY OF ENGINEERING

COMPUTER SCIENCE DEPARTMENT

Data Analysis in High Dimensions

<DISCRIMINANT ANALYSIS>

Module: Data Analysis in High Dimensions

1. <Nguyễn Khắc Hoàng – 18230>
2. <Mai Nguyễn Vy – 17647>
3. <Hà Quách Phú Thành – 18840>
4. <Thái Quang Nam – 18770>

Lecturer: Prof. Christina Andersson

Abstract

High-dimension data is a complicated area with various methods developed to cope with the need for big data analysis. One of the methods is Discriminant Analysis. This survey was conducted to explore whether discriminant analysis can decide the potential academic performance of a student, given several input variables about their lifestyles and study habits. Previous research has proved the usefulness of this method in various scientific fields to classify objects but has not focused much on its application to a human target. As for the research method, we sent out a questionnaire asking about the GPA and lifestyles of our peers in the form of a Google form to collect primary data, which was later applied to discriminant analysis in R Studio to classify respondents into 2 groups of academic performances: High and Low. Finally, the performance of the model was moderately high with some configurations in terms of the predictor variables, and the most influencing lifestyle factors on academic performance were also revealed.

I. Introduction

Discriminant analysis is one of the useful methods researchers use to analyze multivariate data. Its application can be seen in a wide variety of previous research papers in all areas, such as biology, archeology, biochemistry, and so on. For instance, research was carried out on Swiss Bank notes to determine if a banknote is real or counterfeit ('Multivariate Statistics', 2015, Wolfgang Karl Härdle & Zdeněk Hlávka). In the study, six measures, including length, right-hand width, left-hand width, top margin, bottom margin, and diagonal across the printed area, were taken from two populations of notes, genuine and counterfeit. After applying discriminant analysis, a banknote of unknown origin can be detected as real or not.

Similarly in this project, we would like to answer the question "*Can discriminant analysis predict student achievements based on their lifestyles and study habits?*". This paper argues that taking into account multiple information about an individual's lifestyle, discriminant analysis can predict whether he or she belongs to the high-performance student group. It first presents the methodology, then thoroughly describes how we processed data, utilized the model to extract interesting findings, discusses them, and lastly summarizes the suitability of the method in this case.

II. Methodology

Population and Sampling Technique of Survey

The survey encompassed students from Vietnamese-German University, comprising 78 individuals across various majors and intake years. However, for this study, these two factors will not be taken into account. This decision is rooted in the recognition that considering different majors and intake years could introduce unnecessary complexity to the Discriminant Analysis process, considering the small volume of data collected.

Survey Procedure

The questionnaire was initially developed around the premises of this study, e.g. factors, independent variables, and objectives. Subsequently, the questionnaire underwent a meticulous review during a consultation session with a professor. Upon approval, it was distributed to all Bachelor's students via Gmail. Initially, 63 participants responded within two weeks. However, upon conducting a more in-depth data analysis, it became apparent that the dataset needed to be expanded for meaningful analysis. Consequently, the survey was reissued through the university's Gmail system, ultimately garnering responses from a total of 78 participants.

Discriminant Analysis

A. Introduction

Discriminant Analysis is a powerful descriptive and classificatory technique to distinguish groups and classify cases from pre-existing groups based on similarities between that case and the other cases belonging to the group. The primary mathematical objective of discriminant analysis is to

weigh and linearly combine information from a set of p-dependent variables in a way to maximize the separation between k-groups (‘Discriminant analysis and clustering’, 1988, National). This paper offers a comprehensive and in-depth exploration of what the investigators want to know and how to properly apply discriminant analysis to problems.

There are various types of discriminant analysis: Linear Discriminant Analysis, Quadratic Discriminant Analysis, Regularized Discriminant Analysis, Multinomial Discriminant Analysis, Fisher’s Linear Discriminant, Flexible Discriminant Analysis, Canonical Discriminant Analysis. Each type of discriminant analysis depends on how the data analyzer wants their data and results to be displayed.

B. Strengths and Weaknesses

Discriminant Analysis (DA) is known for its strengths since it is a valuable tool for data analysis. On top of it, it mainly focuses on the ability to efficiently separate multiple classes or groups (‘Linear discriminant analysis: A detailed tutorial’, 2017, Tharwat). In such scenarios, it emphasizes the practical application of DA in discriminating more than 2 classes or groups. It is suitable for a diverse range of classes.

Secondly, its ability to enhance class separation is what makes it special for specific types of separation problems. By creating linear combinations between features among the datasets, DA significantly improves the distinguishing process in finding the accurate separation. Thirdly, data transformation is crucial when using DA as a tool. DA transforms original data into testing and training data for better prediction of the model which makes it more reliable when it comes to discriminating classification.

Outliers in data have always been a problem to any discriminating analysis model since outliers can make the model overly influenced by extreme values, potentially leading to incorrect predictions and estimation.

Overfitting for Small Sample Sizes is possible to be another problem occurring if given data is not enough for discriminant analysis to work which lowers the possibilities to get a correct answer for the problem. Since overfitting occurs when the given data contains noisy data or random fluctuations in the training data instead of underlying the true patterns, the mode may not generalize well to new observations, leading to poor performance on the real data.

C. Linear DA and Quadratic DA

There are two primary methods of Discriminant Analysis relevant to this study: Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Before going into their application, it's crucial to explore the distinctions between these two techniques, considering their characteristics and the potential impact on our decision-making based on the acquired dataset.

The key differences between LDA and QDA lie in their decision boundaries and their assumptions about the data. Both methods are parametric, with LDA assuming a common covariance matrix across classes, while QDA assumes that each class possesses its covariance matrix. In practical terms, this means that LDA is constrained to learning only linear boundaries, whereas QDA has the flexibility to learn quadratic boundaries. This distinction is visually evident in the example provided in Figure 1 below:

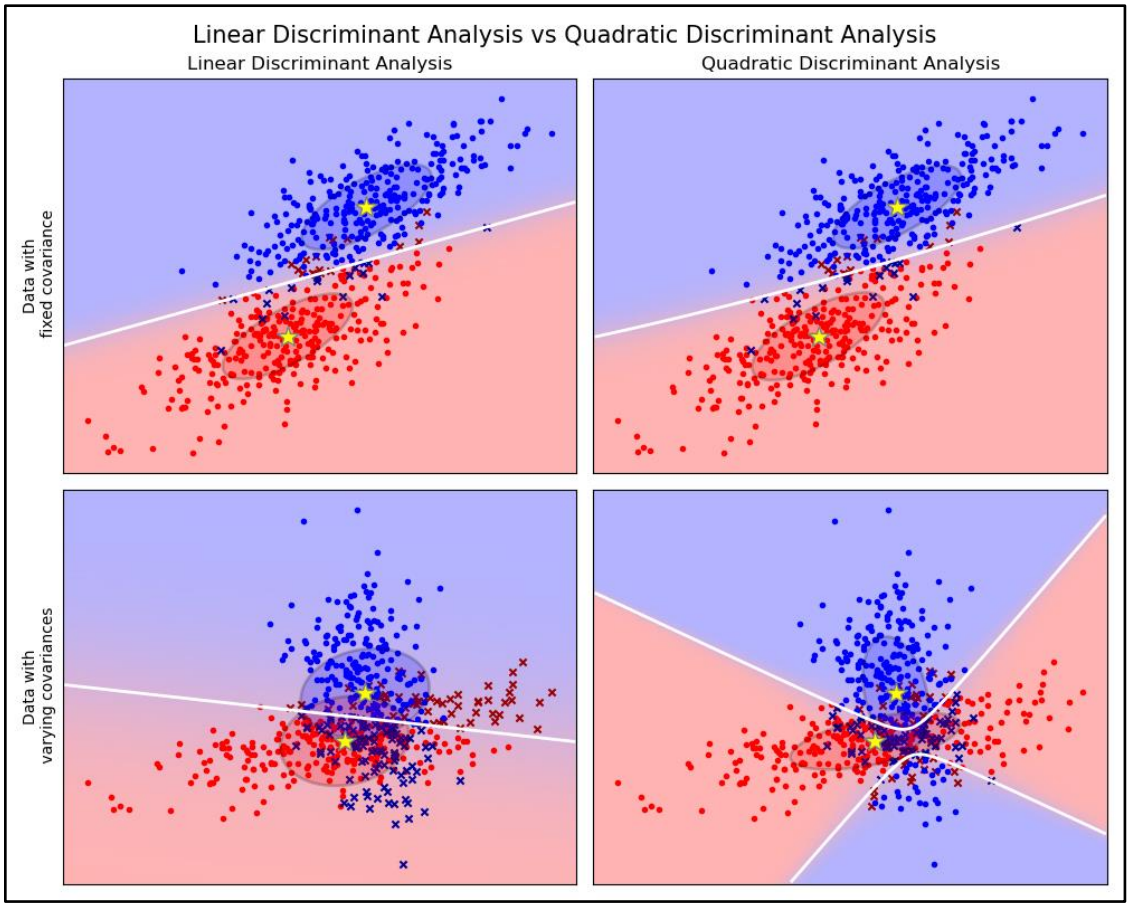

Figure 1: LDA vs QDA (Scikit learn: 1.2. Linear and Quadratic Discriminant Analysis)

In a dataset featuring two predictor variables (x and y) aimed at discriminating between the "red" and "blue" classes, distinct outcomes arise when employing Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). The upper row of the dataset shows effective discrimination with a linear boundary when assuming a fixed covariance matrix. However, in the lower row, where each class is assigned a different covariance matrix, LDA demonstrates a notable decline in performance compared to QDA.

In this scenario, QDA proves more efficient, as it can draw a quadratic boundary that effectively captures and discriminates between the two classes. Conversely, LDA is restricted to establishing a linear boundary, resulting in suboptimal performance. This example underscores the general flexibility of QDA over LDA, particularly in situations where QDA does not assume a fixed covariance matrix, allowing for more effective decision boundaries.

While there are other distinctions between QDA and LDA, such as computational cost and considerations of efficient sample sizes, our study places primary emphasis on the assumptions made about the data. This section has established the fundamental differences between QDA and LDA, setting the stage for a subsequent section where we will delve into detailed analysis to determine which method is more suitable for our specific study.

Data Analysis and Exploration

Data Re-formatting

The collected data has undergone several preprocessing steps before being inputted into the Linear Discriminant Analysis (LDA) model, including cleaning and transforming.

Firstly, a few data points were deleted. *'Study_Hour'* was added to the survey at the beginning because we thought that studying day or night may make a difference. However, many answers lasted from day till night, or from the evening till early morning the next day, thus were very hard to be classified, such as 'Day', 'Afternoon', or 'Night'. Therefore, we decided to remove the *'Study_Hour'* column. The second step was removing unnecessary columns, which recorded emails and timestamps (code, lines 26-27). Also, a respondent with an 8.5 IELTS, whose academic performance the model could not predict as there were no tuples with 8.5 IELTS in the train data for the model to learn. Another answer with disturbing data, which was believed to be from a playful peer, was removed, as well (code, lines 33-34).

As for data transformation, the GPAs, originally in number format, were converted into *'Performance_Type'*, High, and Low, which would be the levels for our target variable. We divided the GPAs into two intervals, 2.5-1 would fall into High and 4-2.6 into Low. The final columns after cleaning, renaming, and transforming were *"GPA"*, *"Performance_Type"*, *"Ielts_Score"*, *"Attendance"*, *"Groupstudy_Likelyness"*, *"Sleep_Duration"*, *"Selfstudy_Duration"*, *"Extracurricular_Duration"*, *"Leisure_Duration"*, *"Distance"* (code, line 38-43), in which the second one is the target variable and the rest, except for *"GPA"*, were predictors. All elements about duration were measured in hours. *"Attendance"*, and *"Groupstudy_Likelyness"* were in percentage. One of the most important steps was to format the variables. R requires all input variables to be numeric or factor in order to be applied to the model. Nevertheless, they were seen as *"char"*, character type, hence, must be reformatted to "numeric" (code, line 48-53). Furthermore, the "%" sign was truncated from *"Attendance"*, and *"Groupstudy_Likeliness"*, and divided by 100 to become float numbers (code, line 56-60). Lastly, the output element, *"Performance_Type"* was changed into factor levels (code, line 63-66).

All in all, there is one target variable, *"Performance_Type"*, which depends on 8 predictor variables. The cleaned dataset is now ready for the next stage.

Performance_Type	Ielts_Score	Attendance	Groupstudy_Likelyness	Sleep_Duration	Selfstudy_Duration	Extracurricular_Duration	Leisure_Duration	Distance
High	7.0	0.15	0.40	6.0	2.5	4.00	4.0	0.0
Low	7.5	0.95	1.00	5.0	4.0	1.00	1.5	0.0
High	6.5	0.90	0.80	10.0	1.0	1.00	3.0	0.0
High	6.0	0.75	0.75	10.0	1.0	0.00	3.0	0.0
High	7.0	0.30	0.20	8.0	3.0	1.00	5.0	0.0
High	6.5	0.80	0.85	8.0	2.0	1.00	2.0	3.0
High	7.5	0.90	0.10	8.0	4.0	0.50	1.0	0.0
High	6.5	0.60	0.60	7.0	3.0	1.00	3.0	0.0
High	7.5	0.50	0.10	7.0	0.5	3.00	5.0	0.0
Low	6.0	0.50	0.60	6.0	1.0	1.00	2.0	0.0
Low	6.5	1.00	0.80	7.0	1.0	0.00	2.0	15.0

Figure 2: Data frame for clean dataset

Data Characteristics

It is crucial to analyze your dataset to get a better understanding of its main characteristics, how it may affect the overall model's performance, and from there, what appropriate methods to use for the data.

1. Data Distribution

The first important thing is to test how data is distributed throughout the process of putting the dataset into the DA model since there is a consideration of which Discriminant Analysis model to use later on. Firstly, we plot the data onto histograms (code, lines 101-108) to see how our data is distributed across the predictor variables.

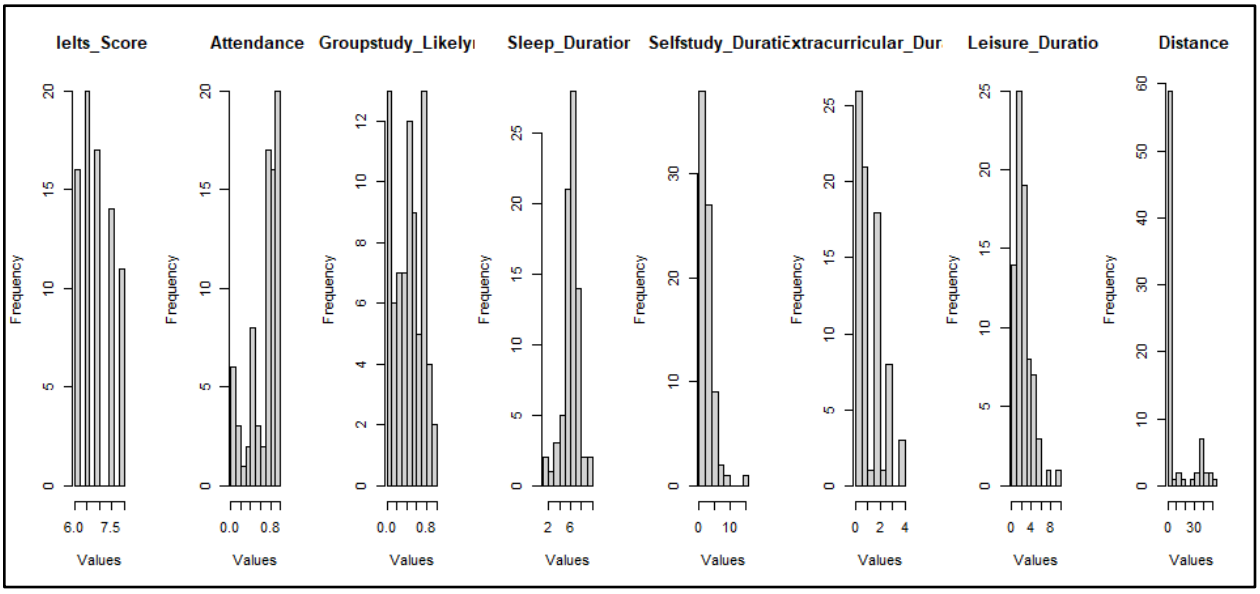

Figure 3: Data histogram for each predictor variable

As can be seen from Figure 3, our data is not normally distributed. However visual inspection is not sufficient.

So we then performed a statistical test for further demonstration of normality. The Shapiro-Wilk test is a reliable method for doing this. After performing the Shapiro-Wilk test (code, line 86-99), we obtained the following data frame:

	Variable	W	p_value
1	Ielts_Score	0.8966147	1.107120e-05
2	Attendance	0.8262440	3.764017e-08
3	Groupstudy_Likelyness	0.9543947	7.115413e-03
4	Sleep_Duration	0.8998209	1.501856e-05
5	Selfstudy_Duration	0.7658328	8.150101e-10
6	Extracurricular_Duration	0.8966102	1.106655e-05
7	Leisure_Duration	0.8763458	1.794903e-06
8	Distance	0.5789244	1.288788e-13

Figure 4: Data frame of Shapiro-Wilk test for each predictor variable

From Figure 4, we can see that although the *w* values, which indicate the probability of normality for each predictor variable, are high, which would likely create a normal distribution, each corresponding *p-value* is lower than 0.05 so we should reject the null hypothesis that the data has not been sampled from a normal distribution.

So based on both the visual and statistical test, we have provided evidence that none of the predictor variables resemble a normal distribution and that normality is not to be found in our data.

2. Data Correlation and Collinearity

The next important concepts to take into consideration are the correlation between predictor variables and the potential presence of collinearity within our dataset. These factors hold

significance, especially in the context of a predictive model like Discriminant Analysis, as high correlation and collinearity have the potential to significantly impact the model's performance. Understanding the interplay of these variables is crucial for assessing and mitigating potential challenges that may arise during the modeling process.

To assess the correlation between every possible pair of predictor variables, three types of correlation — Pearson, Spearman, and Kendall — can be calculated. Notably, the Pearson correlation is a parametric test assuming normality and linearity. Given the evidence against normality presented in the previous section, opting for the Pearson correlation might introduce issues. Therefore, we have chosen to utilize the Spearman correlation for its non-parametric characteristics in this analysis.

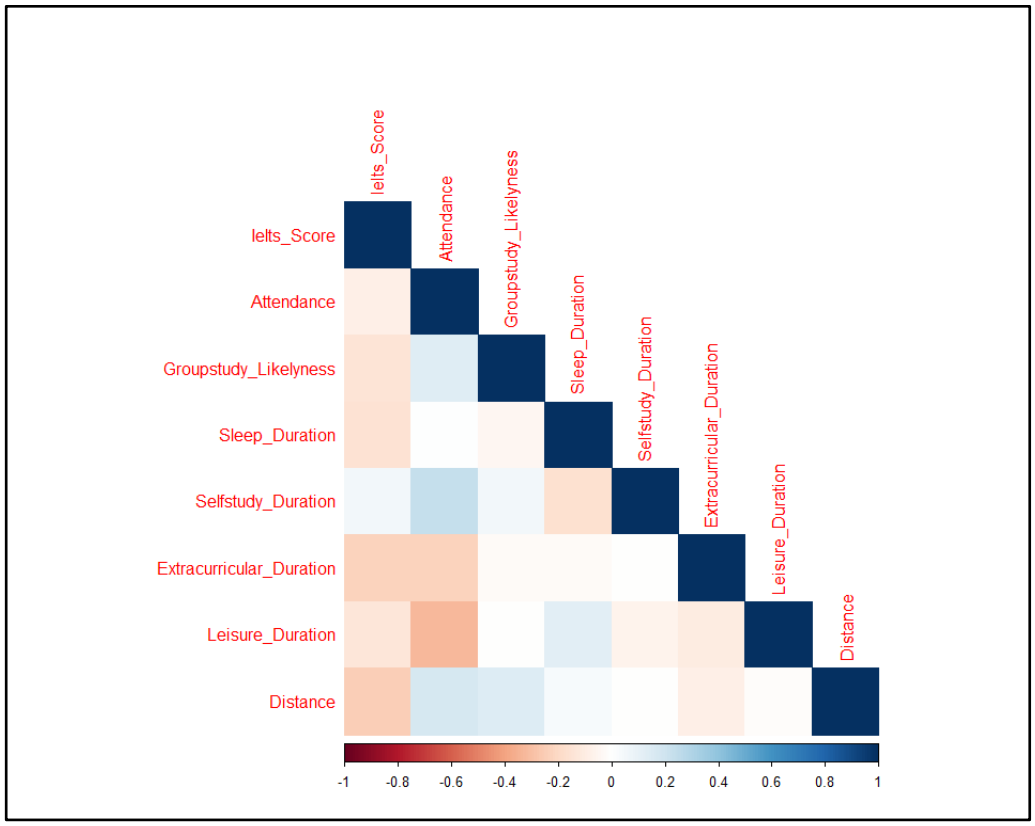

Figure 5: Correlogram of every pair of predictor variables

Figure 5 displays the plot generated by calculating Pearson correlation values for every possible pair of predictor variables (code, lines 112-123). In general, the colors representing correlation values appear pale with a lack of saturation, indicating relatively weak correlations. The only potential exception lies in the correlation value between "*Leisure_Duration*" and "*Attendance*", where a slight increase in color saturation is observed. However, these correlations are deemed insignificant and can be safely disregarded for the current analysis.

To assess the presence of collinearity in our dataset, we employ a statistical measure known as the Variance Inflation Factor (VIF). This metric is essential in calculating the multicollinearity by estimating the extent to which the variances of individual predictor variables are inflated due to collinearity.

After putting in the dataset through the VIF function provided by R (code, line 127-136), the following result is obtained in Figure 6:

Ielts_Score	Attendance	Groupstudy_Likelyness	Sleep_Duration
1.174888	1.453529	1.050001	1.134465
Selfstudy_Duration	**Extracurricular_Duration**	**Leisure_Duration**	**Distance**
1.127086	1.303878	1.176815	1.061786

Figure 6: VIF value for each predictor variable

While these values may not currently represent specific outcomes, a widely accepted rule of thumb to follow is that a *VIF* value equal to 1 indicates no collinearity and a *VIF* value larger than 5 or 10 suggests a high or severe collinearity. As it is clear from the result obtained above, all the *VIF* values for the predictor variables are just slightly above 1. Because of this, we can say with confidence that collinearity can not be found in our dataset.

3. Summary

Following the formatting and cleaning of our data, a critical next step is to thoroughly understand the characteristics of the data and the sample. Knowing the data's characteristics can be crucial and beneficial for the sequential stage in fitting the data into our models with appropriate methods.

We have presented compelling evidence indicating that our data does not follow a normal distribution. Consequently, it is required to reconsider and conduct further analysis on any parametric methods and statistical tests intended for use. Furthermore, we have demonstrated that our predictor variables do not exhibit significant correlations, and collinearity is absent in our dataset. With these considerations in mind, our clean data stands ready for modeling without the necessity of additional formatting or configuration.

LDA and QDA Testing

The distinctions between Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), along with an illustrative example, have been previously outlined and discussed, setting the foundation for this section. Herein, our focus is specifically devoted to the comparison and assessment of different methods, aiming to ascertain which approach is better suited for our study.

1. Performance Comparison

The initial test to conduct involves a performance comparison between Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), even though it is not sufficient on its own, it is still a benchmark that serves as an efficient means for estimating the relative efficacy of the two Discriminant Analysis methods. Our reasoning for this comparison testing is that the method exhibiting superior performance may be considered more suitable for our study.

The R code (lines 145-202) outlines a comprehensive pipeline for the performance comparison between Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). The sequence involves: first splitting the data into training data and testing data with a predefined seed, then fitting them through a normalization pipeline and formatting, after that, fitting them through new instances of an LDA and a QDA model, then calculating the accuracy on the testing data of each model, subsequently, storing the accuracy values into a data frame for each model. After a pre-defined number of runs, two data frames containing the historical values of the accuracy of both models, are calculated to get the mean of each one. The final two means of accuracy are then compared to each other.

```
> testing_da_models(data.transformed, 10000)
[1] "Average LDA Accuracy ( 10000 runs): 0.660006666666667"
[1] "Average QDA Accuracy ( 10000 runs): 0.678246666666667"
```

Figure 7: Average accuracy of LDA and QDA after 10000 runs

Figure 7 illustrates the ultimate mean accuracy values obtained after 10000 runs for each model. Despite the marginal superiority of QDA over LDA, this result positions QDA as the more suitable model for our study. While this comparison alone may not be enough to determine the optimal model, it serves as a valuable test that, when combined with other assessments, strengthens the evidence in favor of QDA as a significantly better model for our specific case.

2. Covariance Matrices Testing

Another way to provide evidence for the preference for QDA over LDA in our case is that we can examine the indication that our data does not share the same covariance matrices across the classes. This method is both legitimate and sufficient to furnish further evidence for our decision.

To visually assess the spread of our data, Figure 8 presents pairs of plots for all predictor variables (code, line 206). The observed variability in the data spread across different pairs suggests that the covariance matrices may not be consistent or fixed across classes.

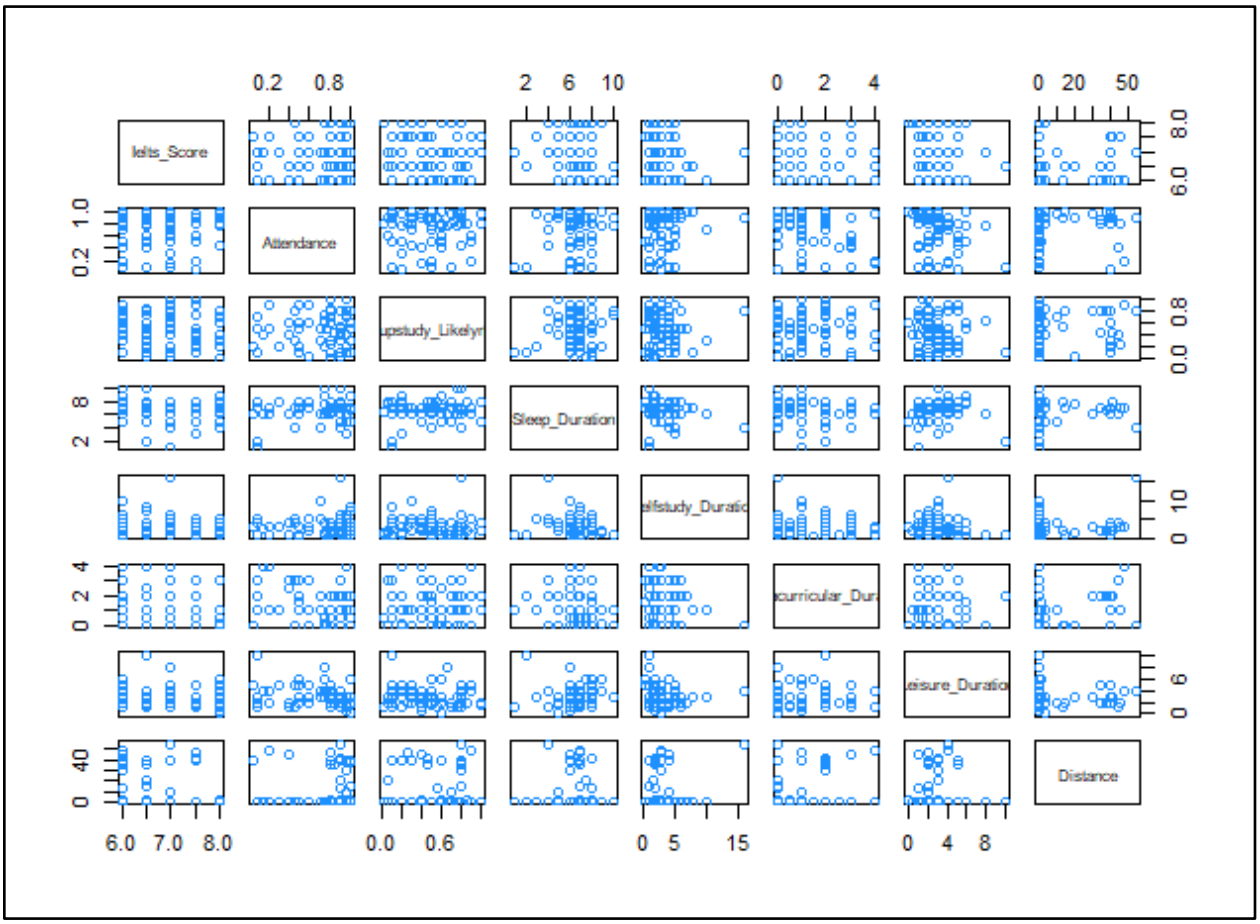

Figure 8: Pairs plot for all pairs of predictor variables

However, a visual interpretation of covariance matrices is not sufficient. That is where the Box's M test comes in. Box's M is a multivariate statistical test designed to offer evidence, with a certain level of confidence, regarding whether a dataset possesses fixed covariance matrices. The Box's

M test is conducted on our data, as indicated in the code (lines 208-214), and the resulting outcome is as follows:

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  data.transformed[, 2:9]
Chi-Sq (approx.) = 51.353, df = 36, p-value = 0.04668
```

Figure 9: Box's M-test result

In Figure 9, the test statistic for fixed covariance matrices, denoted as *Chi-Square*, is reported as 0.51 with a corresponding p-*value* of 0.046. With a p-*value* below the common significance level of 0.05, it is reasonable to reject the null hypothesis that our data has a fixed covariance matrix for the two classes.

3. Conclusion

After a detailed exploration of the distinctions between Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) and their potential implications on our results, we conducted two testing methods: a performance comparison and covariance matrix testing. Both methodologies consistently favored QDA over LDA.

Firstly, the performance comparison indicated that QDA outperformed LDA, albeit by a small margin, in terms of mean accuracy after numerous runs. Subsequently, the covariance matrix testing, incorporating both visual inspection and statistical analysis using Box's M test, provided compelling evidence that our data does not conform to the assumption of fixed covariance matrices required by LDA.

Conclusively, the results from both testing methods align, leading to the decision to employ Quadratic Discriminant Analysis (QDA) instead of Linear Discriminant Analysis (LDA) for our study.

Model Pipeline and Procedure

This section will go through briefly the transforming pipeline and the procedure of data, before fitting into our model.

The methodology comprises five key steps: data transformation into *z-scores* for clarity, splitting data into train and test sets, generating variable coefficients for the DA model, and conducting prediction tests for accuracy assessment for analysis.

Data Normalization

In this experiment, Discriminant Analysis relies on data normalization for consistency, utilizing Z-score normalization.

```
# Normalized data #
data.norm <- data.transformed %>% mutate_each_(list(~scale(.) %>% as.vector),
                                  vars = c("Ielts_Score", "Attendance", "Groupstudy_Likelyness",
                                           "Sleep_Duration", "Selfstudy_Duration",
                                           "Extracurricular_Duration", "Leisure_Duration", "Distance"))
```

Figure 10: Z-score normalization in R

Figure 10 illustrates the R code used to transform original variable values into *z-scores* using the 'dplyr' package and 'scale' function in R. This process standardizes predictor variables within the initial dataset ('data.transformed'), enhancing manageability for Discriminant Analysis modeling.

	Performance_Type	Ielts_Score	Attendance	Groupstudy_Likelyness	Sleep_Duration	Selfstudy_Duration	Extracurricular_Duration
1	High	0.1528285	-2.06133329	-0.30091363	-0.38253867	-0.24034141	2.4663009
2	Low	0.8978676	0.81173344	1.90988040	-1.05608306	0.39874825	-0.2694868
3	High	-0.5922105	0.63216677	1.17294905	2.31163888	-0.87943107	-0.2694868
4	High	-1.3372496	0.09346675	0.98871622	2.31163888	-0.87943107	-1.1814160
5	High	0.1528285	-1.52263328	-1.03784497	0.96455010	-0.02731152	-0.2694868
6	High	-0.5922105	0.27303342	1.35718189	0.96455010	-0.45337130	-0.2694868
7	High	0.8978676	0.63216677	-1.40631065	0.96455010	0.39874825	-0.7254514
8	High	-0.5922105	-0.44523326	0.43601771	0.29100572	-0.02731152	-0.2694868
9	High	0.8978676	-0.80436660	-1.40631065	0.29100572	-1.09246096	1.5543717
10	Low	-1.3372496	-0.80436660	0.43601771	-0.38253867	-0.87943107	-0.2694868
11	Low	-0.5922105	0.99130011	1.17294905	0.29100572	-0.87943107	-1.1814160
12	High	0.8978676	0.81173344	0.98871622	0.29100572	-0.02731152	0.6424425
13	Low	-1.3372496	-1.88176662	1.54141473	0.29100572	-0.02731152	2.4663009

Figure 11: Z-score values

Figure 11 demonstrates the conversion of actual values to *z-values* by subtracting the mean and dividing by the standard deviation. *Z-scores* of 0 align with the mean, while positive and negative scores indicate positions above or below the mean in terms of standard deviations.

Data Partition

In preparing for the Discriminant Analysis (DA) model, the dataset is split into training and testing subsets. Typically, 80% of the data is allocated for training, while 20% is reserved for accuracy assessment, with randomness ensuring fair assignment to these subsets.

```r
# Partition data #
partition_data <- function(data) {
  # Create an index-based for training data #
  training.samples <- data$Performance_Type %>%
    createDataPartition(p = 0.8, list = FALSE)

  # Split the data according to the index
  train.dt <- data[training.samples, ]
  test.dt <- data[-training.samples, ]

  # Estimate pre-processing parameters #
  preproc.param <- train.dt %>%
    preProcess(method = c("center", "scale"))

  # Transform the data using the estimated parameters #
  train.transformed <- preproc.param %>% predict(train.dt)
  test.transformed <- preproc.param %>% predict(test.dt)

  # Return both transformed dataset #
  return(list(train = train.transformed, test = test.transformed))
}
```

Figure 12: Data partition in R

test_data	15 obs. of 9 variables
train_data	63 obs. of 9 variables

Figure 13: Examination of data partition

Figure 12 displays the R code implementing this partitioning using the 'caret' package, generating training indices and dividing data into '*train.dt*' and '*test.dt*' subsets. This process maintains a near 20-80% ratio, as seen in Figure 13, with 15 observations for testing and 63 for training. Randomness ensures an equal likelihood for each observation to belong to either set, ensuring unbiased data partitioning across runs.

Model Training

The application of the Quadratic Discriminant Analysis (QDA) model in R involves establishing algorithms to discern between two distinct classes, particularly the '*Performance_Type*.'

```
model <-  qda(Performance_Type~., data = train.transformed)
model
```

Figure 14: Training QDA model in R

Figure 14 showcases the R code utilizing the 'MASS' library to implement QDA. This code fits the QDA model to the training dataset, enabling classification based on other variables and facilitating an evaluation of its classification process.

```
Prior probabilities of groups:
     High       Low
0.7460317 0.2539683

Group means:
     Ielts_Score  Attendance Groupstudy_Likelyness Sleep_Duration Selfstudy_Duration
High   0.1078253 -0.04290497            -0.1267760      0.1328558        -0.01776886
Low   -0.3167368  0.12603336             0.3724044     -0.3902638         0.05219603
     Extracurricular_Duration Leisure_Duration
High              -0.007877152      -0.06189898
Low                0.023139133       0.18182825
```

Figure 15: QDA model statistic

Figure 15 illustrates the insights derived from the QDA model, including prior probabilities between classes and mean variable values within each class. These insights guide the model in predicting the class of the test data, aiding in its classification process.

Model Prediction

In this stage, the test data is used with the model to compute accuracy. To obtain the most accurate value, we calculate and store the model's accuracy for each run. This process is illustrated in Figure 16 below.

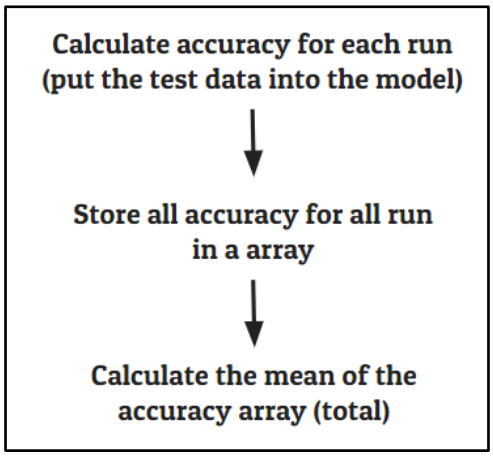

Figure 16: Accuracy calculation process

```
for (i in 1:n) {
  qda_predictions <- predict(qda_fit, newdata = test_data)
  qda_accuracies[i] <- mean(qda_predictions$class == test_data$Performance_Type)
}
# Calculate the average prediction accuracy #
average_qda_accuracy <- mean(qda_accuracies)
```

Figure 17: Accuracy calculation implemented in R

Figure 17 demonstrates the QDA model's application in R, predicting '*Performance_Type*' for the test data. It evaluates model accuracy by comparing predicted and actual values, storing results in '*qda_accuracies*'. After completing iterations, it computes the mean from these stored values. This method gauges the QDA model's average performance in classifying the target variable.

III. Findings

Base Model Performance

At the end phase of this experiment, the evaluation of the Discriminant Analysis models revealed compelling outcomes. This section, dedicated to the QDA base model performance, dives into the aftermath of examining academic performances through the lens of eight key predictor variables. These predictors encompass "*Ielts_Score*", "*Attendance*", "*Groupstudy_Likelyness*", "*Sleep_Duration*", "*Selfstudy_Duration*", "*Extracurricular_Duration*", "*Leisure_Duration*" and "*Distance*".

For clarification, the selected model for this analysis is the QDA, as extensively discussed in the section dedicated to comparing the Quadratic Discriminant Analysis (QDA) and Linear Discriminant Analysis (LDA).

After performing the analysis and data reformatting, QDA was now run for multiple different numbers of iterations to get the average mean (code, lines 220-248), and here is the result obtained:

```
> testing_qda_models(data.transformed, 1)
[1] "Average QDA Accuracy ( 1 runs): 0.733333333333333"
> testing_qda_models(data.transformed, 100)
[1] "Average QDA Accuracy ( 100 runs): 0.671333333333333"
> testing_qda_models(data.transformed, 1000)
[1] "Average QDA Accuracy ( 1000 runs): 0.676866666666667"
> testing_qda_models(data.transformed, 10000)
[1] "Average QDA Accuracy ( 10000 runs): 0.67894"
```

Figure 18: Result for the base model

As depicted in Figure 18, the model's performance tends to converge to an accuracy of around 0.67 as the number of iterations increases. Therefore, it is plausible and justifiable to say that the basic

Quadratic Discriminant Analysis model, devoid of any specific configurations related to predictor variables and algorithms, can adeptly classify and predict a student's academic performance based on lifestyle routines with an accuracy of up to 67% when being given our dataset.

It's important to acknowledge that while this analysis provides valuable insights into accuracy ranges, it might not encapsulate the complete spectrum of performance variability among all models given how small and incomplete our dataset is. Nevertheless, it underscores the inherently dynamic nature of predictive outcomes within the context of this research. This assessment suggests that the expected average accuracy for the base model stands around 67%.

Models with Predictor Variable Combinations

In previous experiments, a key inquiry emerged: "Is it possible to enhance the prediction accuracy rate through the elimination of variables?" This section presents the result of attempts to elevate the mean accuracy prediction rate through the random selection of variable combinations from the original set of eight decisive variables.

To arrive at a conclusive answer, a comprehensive list comprising combinations of predictor variables for Discriminant Analysis was generated. This list encompassed combinations scaling from 2 up to 8 key variables (the setup of the Base Model) resulting in a total of 247 unique combinations. Each combination's performance was assessed using statistics representing the average LDA accuracy rate, average QDA accuracy rate, and the combined sum of both average LDA and QDA accuracy rates. Intriguingly, the analysis revealed significant variations in the accurate prediction rates across different variable combinations. Contrary to expectations, the Base Model displayed considerably lower average prediction accuracies for both LDA and QDA models as depicted in Figure 19 below.

Rank	Total var	combination	Average LDA	Average QDA	LDA + QDA
236	8	Attendance, Distance, Extracurricular_Duration, Groupstudy_...	0.6582600	0.6780533	1.336313
237	5	Attendance, Extracurricular_Duration, Groupstudy_Likelynes...	0.7063467	0.6292267	1.335573
238	6	Attendance, Distance, Extracurricular_Duration, Ielts_Score, S...	0.6724400	0.6619800	1.334420
239	6	Attendance, Distance, Extracurricular_Duration, Groupstudy_...	0.6662200	0.6657800	1.332000
240	6	Attendance, Distance, Extracurricular_Duration, Groupstudy_...	0.6768067	0.6544200	1.331227
241	6	Attendance, Extracurricular_Duration, Groupstudy_Likelynes...	0.6747800	0.6555867	1.330367
242	7	Attendance, Distance, Extracurricular_Duration, Groupstudy_...	0.6785467	0.6516800	1.330227
243	6	Attendance, Distance, Extracurricular_Duration, Groupstudy_...	0.6942933	0.6359133	1.330207
244	4	Distance, Groupstudy_Likelyness, Selfstudy_Duration, Sleep_...	0.6921667	0.6361467	1.328313
245	5	Extracurricular_Duration, Groupstudy_Likelyness, Ielts_Score,...	0.6864933	0.6386400	1.325133
246	7	Attendance, Distance, Extracurricular_Duration, Groupstudy_...	0.6690467	0.6538600	1.322907
247	5	Attendance, Distance, Groupstudy_Likelyness, Selfstudy_Dur...	0.6882333	0.6342200	1.322453

Figure 19: Ranking table of variables combination performance through 10000 runs

As depicted in Figure 19, the base model is near the top 10 models with the worst prediction performance. Considering the sum of average LDA and QDA accuracy rates in the column LDA + QDA, the base model is standing in 236 of 247 models with the sum between 2 model accuracy LDA(0.658) and QDA(0.678) is around 1.336. To put it in comparison, the highest LDA average accuracy is 0.732, achieved by the models of combination between 2 variables of Attendance and Leisure Duration (Figure 20). In the case of QDA average accuracy, the highest record is 0.744 held by the combination of 4 variables (Attendance, Ielts Score, Leisure Duration, and Sleep Duration) (Figure 21). The highest average sum between QDA and LDA is 1.471 belonging to the combination of Distance and Extracurricular Duration variable (Figure 22).

Rank	Total var	combination	Average LDA
1	2	Attendance, Leisure_Duration	0.7329267
2	2	Attendance, Groupstudy_Likelyness	0.7317733
3	2	Distance, Leisure_Duration	0.7313800
4	2	Distance, Extracurricular_Duration	0.7313733
5	2	Attendance, Distance	0.7311400
6	2	Extracurricular_Duration, Leisure_Duration	0.7297933
7	3	Distance, Extracurricular_Duration, Leisure_Duration	0.7294133
8	2	Attendance, Selfstudy_Duration	0.7293267
9	3	Attendance, Distance, Leisure_Duration	0.7292467
10	2	Extracurricular_Duration, Groupstudy_Likelyness	0.7291400

Figure 20: Ranking table of top 10 variables combination LDA performance 10000 runs

Rank	Total var	combination	Average QDA
1	4	Attendance, Ielts_Score, Leisure_Duration, Sleep_Duration	0.7444400
2	5	Attendance, Groupstudy_Likelyness, Ielts_Score, Leisure_Dur...	0.7419800
3	2	Distance, Extracurricular_Duration	0.7397867
4	2	Leisure_Duration, Sleep_Duration	0.7379133
5	5	Attendance, Distance, Ielts_Score, Leisure_Duration, Sleep_D...	0.7353667
6	4	Attendance, Distance, Ielts_Score, Leisure_Duration	0.7342267
7	4	Distance, Extracurricular_Duration, Leisure_Duration, Sleep_...	0.7342067
8	3	Attendance, Ielts_Score, Sleep_Duration	0.7342067
9	6	Attendance, Distance, Groupstudy_Likelyness, Ielts_Score, Le...	0.7336867
10	3	Attendance, Leisure_Duration, Sleep_Duration	0.7331533

Figure 21: Ranking table of top 10 variables combination QDA performance 10000 runs

Rank	Total var	combination	LDA + QDA
1	2	Distance, Extracurricular_Duration	1.471160
2	2	Attendance, Groupstudy_Likelyness	1.453940
3	2	Leisure_Duration, Sleep_Duration	1.453173
4	3	Distance, Extracurricular_Duration, Groupstudy_Likelyness	1.452880
5	2	Ielts_Score, Leisure_Duration	1.452340
6	3	Attendance, Distance, Extracurricular_Duration	1.451373
7	2	Attendance, Leisure_Duration	1.449880
8	2	Attendance, Ielts_Score	1.449687
9	2	Attendance, Distance	1.447760
10	3	Attendance, Ielts_Score, Leisure_Duration	1.446567

Figure 22: Ranking table of top 10 variables combination QDA performance 10000 runs

The statistics presented in Figures 20, 21, and 22 highlight a consistent trend observed across all experiments. Notably, after conducting 10,000 runs, the ranking tables consistently maintain the same relative positions for each combination of variables pertaining to each performance type. It is noteworthy that the top five combinations of variables consistently hold their positions across all three tables, remaining unchanged throughout all the experiments conducted thus far.

In summary, the experimentation revealed that eliminating certain predictor variables resulted in numerous variable combinations showcasing superior performance compared to the base model across three key metrics (average LDA, average QDA, and combined LDA + QDA). It is crucial to highlight that these findings are specific to the observations derived from this particular experiment's dataset.

IV. Discussion

This section aims to dive into the discoveries we've uncovered and offer insights into them. Despite the relatively modest size of our dataset, consisting of 78 respondents, and the potential need for

reevaluation and fine-tuning of the data processing procedures for this particular study, the findings and results remain conclusive to some extent, serving as a basis for future investigations.

Our initial observation revolves around the performance of our baseline Quadratic Discriminant Analysis (QDA) model, which achieved an accuracy of approximately 67% on the test data. While this accuracy might appear respectable, it is important to note that it falls below general expectations. An AI agent, whether human or algorithmic, could randomly guess and predict student performance based on lifestyle factors and still achieve at least 50% accuracy (for 2 classes). Therefore, the QDA model's 17% accuracy improvement over random guessing is not entirely persuasive and requires further analysis, both in terms of the dataset and the model itself. Our second finding heavily suggests that the problem actually lies somewhere in the dataset, specifically among the predictor variables.

The second noteworthy discovery furnishes evidence that the QDA model exhibits improved performance when certain predictor variables are excluded. This implies that some variables were deemed insignificant and contributed noise to our dataset. This outcome is further reinforced by thorough analyses assessing the correlation between variables, and an absence of collinearity in our dataset.

The exemption of predictor variables in the case of a Quadratic Discriminant Analysis (QDA) model reveals that the combination of "*Attendance*," "*Ielts_Score*," "*Leisure_Duration*," and "*Sleep_Duration*" produces the highest average accuracy, reaching up to 74.4%. These specific predictor variables consistently appear in the top-ranking combinations, particularly with "*Attendance*" and "*Ielts_Score*". This observation strongly implies that these four predictor variables play a pivotal role in determining a student's academic performance compared to other variables.

The prominence of "*Attendance*" and "*Ielts_Score*" in the top combinations suggests their significant impact on academic outcomes. Intuitively, this aligns with our expectation- regular attendance often correlates with better performance, and high proficiency in English is indicative of enhanced research and study capabilities, offering a broader exposure to English-centric academic content. Additionally, effective time management in terms of leisure activities and sleep emerges as a key factor, reflecting positively on academic grades. Ultimately, we believe that these identified predictor variables should be the primary focus factors for future study - Lecture attendance, English background and time management.

However, it is essential to acknowledge that this feature's importance may be unique to our dataset. A larger dataset with more diversity could unveil additional significant factors, potentially altering our perspective. Thus, the identified predictors serve as valuable insights for further exploration and should be considered in the context of the dataset's characteristics and scope.

Despite our analysis indicating that Linear Discriminant Analysis (LDA) is inferior to Quadratic Discriminant Analysis (QDA), exploring the same predictor variable combinations for LDA models provides some insightful findings. The most consistent predictor variables in the top-ranking combinations for LDA models are "Attendance" and "Distance." Interestingly, certain LDA models can achieve accuracy levels of up to 73%, which may initially seem promising. However, it's important to recognize the limitations of these results.

Firstly, as established through various testing methods, LDA is not considered suitable for our study. Second, the inclusion of "Distance" as a predictor variable poses challenges, given the limited variability within our dataset where most respondents reside on the VGU campus. Thus, while LDA appears to achieve relatively high accuracy, this outcome is misleading. It underscores

our dataset's limited variability and specific characteristics, emphasizing the importance of careful consideration when interpreting model performance in the context of dataset constraints.

V. Conclusion

The Final Answer

Student lifestyle serves as an important determinant in shaping academic success. This exploration delved deeply into the impact of various lifestyle factors on academic performance. In light of our investigation, the research question posed in this paper regarding the correlation between lifestyle and academic success is addressed. The application of Discriminant Analysis revealed noteworthy evidence suggesting that our Quadratic Discriminant Analysis (QDA) might effectively predict students' academic performance based on lifestyle routines. And that this predictive capability is dependent upon a distinct combination of predictor variables, rather than incorporating all eight original variables.

The Future Hope

However, while our findings offer valuable insights, they represent a preliminary step in understanding this relationship. To draw more conclusive and comprehensive results, our research underscores the necessity for a larger dataset, potentially encompassing multiple colleges. This broader scope would afford a deeper exploration into the interplay between lifestyle and academic success, increasing our understanding of these dynamics and paving the way for more definitive conclusions.

Ultimately, our aspiration is for this research to catalyze further exploration and development within educational institutions. We aim to empower institutions to guide students toward successful college experiences. We hope this endeavor sparks further investigations, providing students with a roadmap to navigate and optimize their academic lives for enduring success.

VI. Appendices

Appendix of Survey

Our survey was conducted by using a Google Form since it is a reliable and fastest way to connect with VGU students' email. The questions in the survey were required to type a specific number instead of dividing into categories. Here is the list of 10 questions that were released:

1. What is your GPA?
2. What is your most recent IELTS score?
3. How often do you attend lectures and classes? (percent)
4. How likely do you study in a group (percent)
5. On average, how many hours a day do you sleep? (hours/day)
6. At what time of the day do you usually study outside of class?
7. On average, how many hours a day do you spend doing your academic homework? (hours/day)
8. On average, how many hours a day do you spend doing extracurricular activities? (hours/day)
9. On average, how many hours a day do you spend doing leisure activities? (hours/day)
10. What is the distance (kilometers) from your home to your campus (0 km if you stay on campus)

VII.References

Härdle, W., & Hlávka, Z. (2007). Multivariate statistics. *Barlin and Praue*.

Tharwat, Alaa & Gaber, Tarek & Ibrahim, Abdelhameed & Hassanien, Aboul Ella. (2017). Linear discriminant analysis: A detailed tutorial. Ai Communications. 30. 169-190,. 10.3233/AIC-170729.

National, R. C., Division, O. E. A. P. S., Commission, O. P. S. M. A., Board, O. M. S., Committee, O. A. A. T. S., Classification, A. C., & Panel, O. D. A. (1988). *Discriminant analysis and clustering*. National Academies Press.

scikit-learn. (n.d.). Linear and Quadratic Discriminant Analysis. Retrieved March 11, 2023, from https://scikit-learn.org/stable/modules/lda_qda.html

Contribution Table

Num	Task	Name	State	Note
1	Create the survey	All	Done	
2	Clean, and prepare data in R	Mai Nguyen Vy Thai Quang Nam	Done	
3	Analyze (correlation, collinear data in R	Ha Quach Phu Thanh Nguyen Khac Hoang	Done	
4	LDA vs QDA comparison testing in R	Ha Quach Phu Thanh	Done	
5	Experiments and Findings in R	Nguyen Khac Hoang	Done	
6	Write an abstract and introduction in the report	Mai Nguyen Vy	Done	
7	Write survey procedures and techniques in the report	Mai Nguyen Vy	Done	
8	Write the discriminant analysis method introduction in the report	Thai Quang Nam	Done	
9	Write data distribution in the report	Thai Quang Nam	Done	
10	Write data correlation, collinearity, and summary in the report	Ha Quach Phu Thanh	Done	
11	Write LDA vs QDA comparison testing in the report	Ha Quach Phu Thanh Mai Nguyen Vy	Done	
12	Write DA process in the report	Nguyen Khac Hoang	Done	
13	Write Experiment and Findings in report	Nguyen Khac Hoang	Done	
14	Write Discussion in the report	Ha Quach Phu Thanh	Done	
15	Write Conclusion in the report	Nguyen Khac Hoang	Done	