The Exploring of Student’s School Majoring Data at Madrasah Aliyah Negeri in Samarinda Using Linear Discriminant Analysis Models

N A Rizki*, Wasono, Y N Nasution
Mathematics Department, Faculty of Mathematics and Natural Sciences, Mulawarman University, Samarinda, Indonesia
*Corresponding author: nanda.arista.r@gmail.com

Abstract. Curriculum 2013, which applied to Madrasah Aliyah Negeri (MAN) requires that majors are implemented since the 10th grade. The majors are supposed to relate to the study plan when they study at University. The aims of this research are to describe correspondence between current majors and desired future majors, to observe distribution pattern of student exam data based on their boxplot, to classify student majors data using the Linear Discriminant Analysis (LDA) models, and to get the best LDA model to describe the characteristics of majors in MAN. The LDA models used in this research were Fisher’s Linear Discriminant Analysis (FLDA), Diagonal Linear Discriminant Analysis (DLDA), Shrunken Linear Discriminant Analysis (SLDA), Maximum-uncertainty Linear Discriminant Analysis (MLDA), and Factor-model Linear Discriminant Analysis (RFLDA). The best LDA model was chosen based on the classification accuracy produced by resampling with $n = 1000$ and $n = 5000$ for training data and testing data with the probability of each data being drawn were 60:40, 70:30, 80:20, and 90:10.

1. Introduction
Currently, madrasah has changed the paradigm become like public school. The majority of Samarinda Islamic citizens have known madrasah as a religious educational institution provided for students who want to deepen Islamic learning. Curriculum in Madrasah Aliyah (MA) regulated that there were four majors for students, i.e. natural sciences (Ilmu Pengetahuan Alam, or IPA), social sciences (Ilmu Pengetahuan Sosial, or IPS), languages (Bahasa), and religion (Agama). Curriculum 2013 set the majors for students starting from 10th grade. The lack of preparedness and understanding of these majors, will have an impact on the role of the student and the teacher. [1]

Madrasah Aliyah prepares its alumni to become prospective students in various higher education, such as universities. The major that students have taken at MA will have an impact on the selection of majors while studying at university. In this case, it is expected that there will be no misclassification in the 10th grade student majors. That is, students who have the potential for a particular department will be placed in the right direction. [1]

The process of finding patterns from majors data can be accomplished using data mining. One of the data mining that leads to the classification method is discriminant analysis. Discriminant analysis method aims to find a linear combination of characteristics between classes. The properties of this method are homogeneous in the same class, but heterogeneous between different classes. By using this method, the results obtained can separate the characteristics between departments that are inherent in 10th grade students.
Linearity of model in LDA becomes primary choice for high interpretation capabilities possessed by the model in analyzing the data. In addition, the linear case is more easily understood and visualized. Therefore, the linear discriminant analysis which is also known as Fisher discriminant, is a very popular technique. This models is quite sensitive to the presence of outliers. In the process, discriminant analysis makes a class separator and classifies observations into corresponding classes. Discriminant analysis involves a combination of two or more independent variables that will form the best separator between classes.

Discriminant analysis is often used to determine the relationship of dependence between one response variable and two or more independent variables. However, it is different from variance analysis and regression analysis, that categorical discriminant analysis response variables have been grouped into several classes. Therefore, discriminant analysis aims to classify an observation into a class. Classification results are homogeneous for each observation in one class, but are heterogeneous between the classes. [3]

**Figure 1.** Separation of two classes by a dividing line.

**Figure 2.** The projection of each class to a determining line.

Suppose the observation sample will be classified into two classes as shown in Figure 1. Each class was separated by a dividing line, which was formed from the determining line so that each observation in within class was homogeneous but heterogeneous between the classes. The basic idea is projecting each class of data to a determining line as in Figure 2, so the projection results from each class produced the largest difference of average and the smallest variance for each class. Thus, a dividing line was formed that can classify an observation into a particular class as in Figure 1.

Suppose that variable $\mathbf{Y}$ which consists of $k$ class, will be classified based on variable $\mathbf{X}$ which consists of $d$ variables with $n$ observations. The matrix $\mathbf{X}$ was transformed by the projection matrix $\mathbf{W}$ into matrix $\tilde{\mathbf{Y}}$ which was the $k$ possibilities of a linear combination of variable $\mathbf{X}$.

$$\tilde{\mathbf{Y}} = \mathbf{W}'\mathbf{X}. \quad (1)$$

The determining line used as a linear combination to classify it is one column of the matrix $\mathbf{W}$, such that

$$y_j = \mathbf{W}'\mathbf{X} = w_1^* x_{1j} + w_2^* x_{2j} + \cdots + w_d^* x_{dj}, \quad (2)$$

for $j = 1, 2, \ldots, n$. The determining line in Equation 2 was measured by the score function written in Equation 3.

$$\lambda = \frac{\mathbf{W}'\mathbf{S}_B \mathbf{W}}{\mathbf{W}'\mathbf{S}_W \mathbf{W}}, \quad (3)$$
which

\[ S_B = \sum_{i=1}^{k} d(x_i - \bar{x})(x_i - \bar{x})' \]  
\[ S_W = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' \].

In this case \( W, S_B, S_W \) were respectively the projection matrix, between-class covariance matrix, and within-class covariance matrix.

Furthermore, from Equation 3, was obtained

\[ W'S_B W = \lambda W'S_W W \]
\[ S_W^{-1} S_B W = \lambda W \],

because \( \lambda_i \) was scalar for \( i = 1, 2, \ldots, k \). So that the optimal projection vector of the matrix \( W \) is the column that has the largest eigen value. [5]

In addition to the analysis introduced by Fisher, several methods of development were also carried out, namely Diagonal Linear Discriminant Analysis (DLDA), Shrunken Linear Discriminant Analysis (SLDA), Maximum-uncertainty Linear Discriminant Analysis (MLDA), and Factor-model Linear Discriminant Analysis (RFLDA) which available in an R package called HiDimDA. These models are not only used for high-dimensional data, but also even for data whose number of variables exceeds the number of observations. The difference among these four models is how to estimate the inverse of sample within-class covariance matrix \( S_W^{-1} \). DLDA models uses the diagonal matrix of inverse sample variances. SLDA models uses inverse of an optimally shrunken for covariance matrix. MLDA models uses a regularized inverse covariance that deemphasizes the importance given to the last eigen vectors of the sample covariance, determined by maximum-uncertainty of LDA. RFLDA uses a factor model estimate of the true inverse covariance matrix. [4]

This research was a preliminary study in exploring the characteristics of inter-classes for Madrasah Aliyah Negeri (MAN) students in Samarinda. We tried to apply what had been done by Dewi et al [2] about the LDA was applied to the data placement of students of MAN. According to Dewi et al [2], if the covariance variance matrix is homogeneous then the linear model of discriminant analysis is better than the quadratic model. So the purpose of this study was to establish a model of linear discriminant analysis on majors grouping students of MAN in Samarinda. Therefore, we want to compare the Fisher’s LDA model with other LDA models such as Diagonal Linear Discriminant Analysis (DLDA), Shrunken Linear Discriminant Analysis (SLDA), Maximum-uncertainty Linear Discriminant Analysis (MLDA), and Factor-model Linear Discriminant Analysis (RFLDA).

2. Methodology

The data as an object in this research were students of Madrasah Aliyah Negeri (MAN) in Samarinda and in 2018. Some madrasah schools will be taken as samples. Secondary data were obtained from school test scores of all subjects in SMP/MTs of students who studying at 10th grade in academic year 2018-2019, and their majors. The number of students majoring in IPA, IPS, Agama, and Bahasa were 292, 133, 124, and 40 respectively. Then based on the formula for determining the number of Slovin’s samples with a confidence level of 0.95, that the number of samples used for the LDA models from IPA, IPS, Agama, Bahasa were 118, 54, 50, and 16 respectively. To describe the characteristics of MAN students in Samarinda, we dig up information based on previous school (SMP/MTs), questionnaire regarding future majors, school test scores, and their majors in 10th grade. Furthermore, data analysis was performed using a linear discriminant analysis model to classify students based on majors in 10th grade using R version 3.1.3.
In this research, the variables used were dependent variables ($Y$) and independent variables ($X$). The dependent variable was madrasah student majors, namely IPA, IPS, Bahasa dan Agama. The independent variables used were the national exam scores held by the school of origin

$$x_1 = \text{bahasa indonesia (INA)}$$
$$x_2 = \text{bahasa inggris (ING)}$$
$$x_3 = \text{matematika (MATH)}$$
$$x_4 = \text{ilmu pengetahuan alam (IPA)}$$

and school exam scores when they graduated from SMP/MTs, which includes

$$x_5 = \text{pendidikan agama islam (PAI)}$$
$$x_6 = \text{pendidikan kewarganegaraan (PKN)}$$
$$x_7 = \text{bahasa indonesia (INA.1)}$$
$$x_8 = \text{bahasa inggris (ING.1)}$$
$$x_9 = \text{matematika (MAT)}$$
$$x_{10} = \text{ilmu pengetahuan alam (IPA.1)}$$
$$x_{11} = \text{ilmu pengetahuan sosial (IPS.1)}$$

3. Analysis and Discussion
Before performing LDA, it should be better to know the general description of the data using descriptive statistics. Based on previous school, that the majority of MAN students came from Islamic schools, namely, Madrasah Tsanawiyah (MTs). It can be seen in Figure 3, that students whose schools originate from MTs by 70.59% of the samples taken. The classes used as the dependent variable if sorted by the greatest percentage were IPA, IPS, Agama, and Bahasa, respectively for 49.58%, 22.69%, 21.01% and 6.72%, which can be seen in Figure 4.

The next analysis was to describe research instruments such as questionnaires. The instrument was designed so that students can choose one of three options plan on college majors, i.e. Universitas/Politeknik Jurusan Sains & Teknologi (IPA), Universitas/Politeknik Jurusan Sosial Humaniora (IPS), dan Sekolah Tinggi/Institut Agama Islam. The results are presented in Table 1. Majority of students would still choose majors relevant to current majors. Percentage of choosing the appropriate majors for majors in IPA, IPS, Agama, Bahasa was 88.14%, 74.07%, 84%, and 93.75%, respectively.
Table 1. Conformity Between Current Majors and Future Majors

| Current majors | Future majors |
|----------------|---------------|
|                | Not answer   | Agama | IPA | IPS |
| Agama          | 2            | 42    | 4   | 2   |
| Bahasa         | 1            | 15    | 0   | 0   |
| IPA            | 1            | 9     | 104 | 4   |
| IPS            | 2            | 11    | 1   | 40  |

Distribution pattern of student exam data can be observed based on their boxplot a.k.a. box and whisker diagram which can be seen in the Figure 5. Exams for mathematics and natural science were still considered scourge by students because their mathematics and natural science scores were lower than the four subjects with a median value of less than 60. Furthermore, based on the school exam scores data when they graduated from previous school in Figure 6, that students’ abilities for each subject were almost same. Scores for each subject ranged from 80 to 90, because median for each lesson was in interval from 80 to 90.

Figure 5. Boxplot for national exam scores

Table 2. Shapiro-Wilk Normality Test

| SW   | P-value  |
|------|----------|
| 0.81136 | 2.778e-16 |

When data was normally distributed, the ADL models can guarantee to minimize expected errors. If not, then the ADL models can also be applied with the consequence that the model error did not reach
their minimum. Based on Table 2, the data obtained were not multivariate normal distribution. This can be proven by looking at the p-value of the Multivariate Shapiro Wilk statistic that is far below 0.05. If seen from Table 3, the homogeneity for the covariance matrix still has not met because the p-value of Box’s M statistic is $1.092 \times 10^{-16}$ which is still far below 0.05.

![Boxplot for school exam scores](image)

**Figure 6.** Boxplot for school exam scores

**Table 3.** Box’s M-test for Homogeneity of Covariance Matrices

| Chi-Sq (approx.) | df | P-value   |
|------------------|----|-----------|
| 351.58           | 198| 1.092e-10 |

Next analysis was to form LDA models. Based on the results of R software, the LDA models for FLDA, DLDA, RFLDA, MLDA, and SLDA respectively are

$$FLDA(X) = 0.083176968x_1 + 0.011158506x_2 + 0.015385221x_3 - 0.000173923x_4 + 0.007513118x_5 + 0.007082483x_6 + 0.015511028x_7 - 0.054347935x_8 + 0.102776971x_9 + 0.02673583x_{10} + 0.02017088x_{11}$$

$$DLDA(X) = 0.053694589x_1 + 0.031653691x_2 + 0.021360765x_3 + 0.029423752x_4 + 0.005537252x_5 + 0.046527838x_6 + 0.04410738x_7 + 0.053360294x_8 + 0.049627917x_9 + 0.034153103x_{10} + 0.026155849x_{11}$$

$$RFLDA(X) = 0.017056648x_1 + 0.049272752x_2 - 0.00525529x_3 - 0.017209506x_4 - 0.117596549x_5 + 0.019862538x_6 + 0.027489129x_7 + 0.063839171x_8 - 0.012239339x_9 - 0.03093756x_{10} - 0.002838729x_{11}$$

$$MLDA(X) = -0.045300678x_1 - 0.026981695x_2 - 0.006064194x_3 - 0.002511572x_4 + 0.061252538x_5 - 0.025185093x_6 - 0.024367394x_7 - 0.031484691x_8 - 0.037788711x_9 - 0.004926048x_{10} - 0.009244758x_{11}$$
SLDA(\(X\)) = 0.06034749x_1 - 0.03378179x_2 + 0.0200284x_3 + 0.10221992x_5
- 0.01324947x_6 - 0.05395315x_7 - 0.08899704x_8 + 0.07680448x_9
+ 0.04182166x_{10} + 0.01574709x_{11}.

Table 4. Accuracy results for the LDA models are based on the number of bootstraps and the proportion of training data

| Bootstrap | train:test | FLDA  | DLDA  | RFLDA | MLDA  | SLDA  |
|-----------|-----------|-------|-------|-------|-------|-------|
| n = 1000  | 60%:40%   | 0.8376757 | 0.7139188 | 0.7936239 | 0.7964977 | 0.8348012 |
|           | 70%:30%   | 0.8340527 | 0.7092372 | 0.7885911 | 0.7908666 | 0.8312543 |
|           | 80%:20%   | 0.8326919 | 0.7073641 | 0.7873742 | 0.7889228 | 0.8305157 |
|           | 90%:10%   | 0.8319291 | 0.7052797 | 0.7865000 | 0.7883172 | 0.8297123 |
| n = 5000  | 60%:40%   | 0.8378988 | 0.7139398 | 0.7934125 | 0.7956703 | 0.8345192 |
|           | 70%:30%   | 0.8345966 | 0.7093999 | 0.7894745 | 0.7913066 | 0.8317797 |
|           | 80%:20%   | 0.8332141 | 0.7076065 | 0.7881837 | 0.7896593 | 0.8308669 |
|           | 90%:10%   | 0.8322374 | 0.7057506 | 0.7865326 | 0.7881956 | 0.8297566 |

The next step was to determine the best model based on classification accuracy. We divided the sample data into training data and testing data with probability taken of each data were 60 : 40, 70 : 30, 80 : 20, dan 90 : 10. This study used a bootstrap technique with a number of resampling tested was \(n = 1000\) and \(n = 5000\). The results of accuracy measure for each model were presented in Table 4.

Based on Table 4, accuracy decreases with increasing proportions for training data. The low of accuracy values can be caused the data was not homogeneous in variance, so it did not guarantee that it will produce the optimal separator.

![Figure 7. Projection result for FLDA model](image-url)
The best model was FLDA. However, in estimating inverse covariance matrices, SLDA models were better than the DLDA, RFLDA and MLDA models, because it approaches the sample matrix obtained from the FLDA models. The projection for FLDA models can be seen in Figure 7. Visually, there are differences between the characteristics of IPA and IPS majors. Characteristics of Bahasa majors resembled IPA majors, while Agama major’s characteristics resemble IPS majors.

Based on the results of our research, that our data was not normally distributed. Errors of LDA models obtained were not guaranteed to be minimal. This means that for the same data, it was possible that non-linear models have smaller errors than the linear models. These were also explained by the Box’s M test. Because of non-homogeneous covariance matrix, non-linear models were better than the linear models. As a suggestion for further research if you want to use this data, it might be better when applied to non-linear discriminant analysis models or other classification methods for data that was not multivariate normally distributed.

Acknowledgement
This research was supported by BOPTN Faculty of Mathematics and Natural Sciences Mulawarman University Samarinda. The authors are also immensely grateful to the anonymous reviewers for their comments on an earlier versions of the manuscript, although any errors are our own and should not tarnish the reputations of these esteemed professionals.

References
[1] At-Taubany T I B, Suseno H 2017 Desain Pengembangan Kurikulum 2013 di Madrasah (Depok: Kencana)
[2] Dewi Y S, Wati P I, Hadi A F 2013 Analisis Diskriminan Linier dan Kuadratik Majalah Ilmiah Matematika dan Statistika 13 (1): pp. 1–10 http://repository.unej.ac.id/handle/123456789/56818
[3] McLachlan G J 2004 Discriminant Analysis and Statistical Pattern Recognition (New York: John Wiley & Sons)
[4] Silva A P D 2015 HiDimDA: High Dimensional Discriminant Analysis http://CRAN.R-project.org/package=HiDimDA
[5] Tharwat A, Gaber T, Ibrahim A, and Hassanien A E 2017 Linear Discriminant Analysis: A Detailed Tutorial Ai Communications 30(2): pp. 169–190 DOI: 10.3233/AIC-170729