Classification of diabetes events using discriminant analysis

W Rahayu¹*, V M Santi¹ and B S Putri²

¹Program studi Statistika, FMIPA Universitas Negeri Jakarta, Jakarta, Indonesia
²Program studi Matematika, FMIPA, Universitas Negeri Jakarta, Jakarta, Indonesia

*wrahayu@unj.ac.id

Abstract. This study aims to classify diabetes events accurately, because it can be used as early prevention before complications occur. Based on linear discriminant analysis, it gets that someone who have more weight, lower age, and more cholesterol level will make s/he classified into diabetes patient. Then, based on APER test, it gets results the percentage of misclassification is 14%. Therefore, classification of diabetes case using discriminant analysis can be used for the classification of diabetics, because the accuracy has a reasonably high result. Classification with discriminant analysis is expected to be applicable to other diseases datasets.

1. Introduction
Diabetes is a chronic disease where glucose (pure sugar) in the blood becomes high because the body cannot produce enough insulin. Diabetes can be caused by many things, such as heredity, unhealthy lifestyle, rarely exercising, being overweight (obese), etc. Increasing age, diabetics can also experience complications of the disease, such as blindness, kidney failure, heart disease, etc. Someone if you have blood sugar levels which slightly exceed the reasonable limit cannot be said as people with diabetes. Such conditions can still be noted as prediabetes sufferers.

Classification between prediabetes and diabetes patients is needed because it can be used as early prevention before complications occur and can also be obtained significant variables that influence the changes in a person's blood sugar levels. A classification can be done using discriminant analysis to get accurate classification results and also variables that can influence the formation of diabetes.

Discriminant analysis is a technique that is used to analyze data when the dependent variables is categorical and the independent variables is interval. The objective of discriminant analysis is to develop discriminant functions based on the linear combination of independent variables that will discriminate between the categories of the dependent variable in a perfect manner. Linear discriminant analysis is one type of discriminant analysis that focuses on the separation of observations where each population has a multivariate normal distribution and has the same variant covariance matrix (homogeneity).

The decision-making steps with discriminant analysis are: identify the dependent variable and the independent variable, specifies the method for creating a discriminant function, test the significance of the discriminant function using Wilks’s Lambda, calculates the classification accuracy, interpretation of the discriminant function.

Sayed applied discriminant analysis in differentiating between the signal patterns of healthy subjects and those of individuals with specific heart conditions based on diagnosis of ECG signals [1]. An approach for classifying multivariate ECG signals based on discriminant and wavelet analyses was proposed. Bian studied the variable selection criterion for linear discriminant rule and its optimality in
high dimensional and large sample data [2]. Asamoah Fishers Linear Discriminant Function (FLDF) was derived to provide maximum separation between Type 2 and Type1 diabetes patients based on identified risk factors [3]. Kubokawa investigating the performance of the multiple discriminant analysis (MDA) to the differential diagnosis of microcytic anemia [4]. Banu classification of hypothyroid disease using Linear Discriminant Analysis (LDA) [5].

In this paper diabetes disease is to be classification using discriminant analysis to achieve better accuracy. The dataset used for the study on diabetes is taken from Pesanggrahan District Health Center, South Jakarta City, DKI Province.

2. Research methods

2.1. Data and variables

Data in this study were obtained from the Pesanggrahan District Health Center, South Jakarta City, DKI Jakarta Province, December 2016 - January 2017. The total number of respondents (patients) in this study was 172 people, with details of 87 pre-diabetes sufferers and 85 diabetics. The variables used in this study are:

| Variable | Description | Type         |
|----------|-------------|--------------|
| Y        | Diabetes incidence (1 or 2) | Categorical |
| X₁       | Age (Year)  | Continuous   |
| X₂       | Weight (kg) | Continuous   |
| X₃       | Cholesterol level (mg/dl)     | Continuous  |
| X₄       | Triglyceride levels (mg/dl)   | Continuous  |
| X₅       | Systolic blood pressure (mmHg)| Continuous |
| X₆       | Diastolic blood pressure (mmHg)| Continuous |
| X₇       | Uric acid level (mg/dl)       | Continuous  |

2.2. Discriminant analysis

Discriminant analysis is used to classify individuals into one or two groups or more. If the classification consists of only two groups, simple linear discriminant analysis is used. But if the classification consists of three or more groups, the technique used is multiple discriminant analysis. [3]. Three assumptions must be fulfilled in discriminant analysis: normality, homogeneity, and there is no correlation between independent variables.

2.2.1. Discriminant function of two groups. Suppose that \( f₁(x) \) and \( f₂(x) \) are functions of probability density for population groups 1 (\( ω₁ \)) and group 2 (\( ω₂ \)). An observation must be included in population 1 (\( ω₁ \)) or population 2 (\( ω₂ \)). The discriminant function \( y \) can be written as:

\[
y = a₀ + a₁x₁ + a₂x₂ + \ldots + aₖxₖ = x'ₜ = a'x
\]

where

\[
a' = (a₀, a₁, a₂, \ldots, aₖ)
\]

\[
x' = (1, x₁, x₂, \ldots, xₖ)
\]

\[
y = (\bar{x}_1 - \bar{x}_2)' \cdot S^{-1}_pooled \cdot \bar{x}'
\]

Let in sample 1 have \( n₁ \) observations and sample 2 have \( n₂ \) observations with \( k \) variables, then obtained:

\[
x₁ = (x₁₁, x₁₂, \ldots, x₁n₁)
\]

\[
x₂ = (x₂₁, x₂₂, \ldots, x₂n₂)
\]

\[
\bar{x}_₁ = \frac{1}{n₁} \sum_{i=1}^{n₁} x₁i \text{ and } \bar{x}_₂ = \frac{1}{n₂} \sum_{i=1}^{n₂} x₂i
\]
\[ S_1 = \frac{1}{n_{1-1}} \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)(x_{1i} - \bar{x}_1)' \quad \text{and} \quad S_2 = \frac{1}{n_{2-1}} \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)(x_{2i} - \bar{x}_2)'
\]
\[ S_{\text{pooled}} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \]

Then \( x_0 \) can be classified into population \( \omega_1 \) if it satisfies:
\[ (\bar{x}_1 - \bar{x}_2)' S_{\text{pooled}}^{-1} x_0 \geq \frac{1}{2} (\bar{x}_1 - \bar{x}_2)' S_{\text{pooled}}^{-1} (\bar{x}_1 + \bar{x}_2) \]

2.2.2. Evaluation of classifications. APER (Apparent Error Rate) is a measure of classification accuracy. From classification table, as follows:

| Actual | Prediction | Total Observation |
|--------|-----------|------------------|
| 1      | 1         | \( n_{11} \)     |
| 1      | 2         | \( n_{12} \)     |
| 2      | 1         | \( n_{21} \)     |
| 2      | 2         | \( n_{22} \)     |
|        |           | \( n_1 \)        |
|        |           | \( n_2 \)        |

\[ APER = \left( \frac{n_{12} + n_{21}}{n_1 + n_2} \right) \times 100\% \]

3. Results and discussion
The analysis was carried out on 172 diabetic patients who were divided into 87 prediabetes patients and 85 diabetic patients with seven explanatory variables. Results of the analysis are obtained in Table 3.

3.1. Data exploration

| Independent Variable      | Prediabetes | Diabetes |
|---------------------------|-------------|----------|
| Mean                      | Std. Dev    | Mean     | Std.Dev  |
| Age                       | 62.70       | 6.356    | 53.62    | 8.826 |
| Weight                    | 59.63       | 9.745    | 65.59    | 12.313 |
| Cholesterol level         | 191.67      | 41.308   | 203.61   | 36.336 |
| Triglyceride              | 138.84      | 64.829   | 157.68   | 71.445 |
| Systolic blood pressure   | 125.29      | 11.085   | 125.06   | 13.150 |
| Diastolic blood pressure  | 80.11       | 6.378    | 76.82    | 6.937  |
| Uric acid                 | 6.84        | 1.598    | 5.55     | 1.373  |

In Table 3 showed that in the diabetes group the weight, cholesterol, and triglyceride groups had higher mean values than the prediabetes group.

3.2. Test assumption
The normal test results of the independent variables with p-p plot are as Figure 1.
Based on above figure, the normality test of the independent variables is met with dots spreading around the diagonal line. By calculating the VIF value of the independent variable against the non-independent variable, the following results in Table 4.

### Table 4. VIF of independent variables.

| Variable                  | VIF  |
|---------------------------|------|
| Age                       | 1.261|
| Weight                    | 1.095|
| Cholesterol level         | 1.070|
| Triglyceride              | 1.063|
| Systolic blood pressure   | 1.559|
| Diastolic blood pressure  | 1.604|
| Uric acid                 | 1.193|

Based on Table 4, all VIF of independent variables are <10 which means there is no multicollinearity between the independent variables.

3.3. Discriminant analysis

The aim is to reduce the dimensions of a set of independent variables to be smaller, through the following stages:

3.3.1. Canonical correlation. Measuring the closeness of the relationship between the magnitude of variability that can be explained by the independent variable on the dependent variable.

### Table 5. Canonical correlation value.

| Function | Eigenvalue | % of Variance | Cumulative % | Canonical Correlation |
|----------|------------|---------------|--------------|-----------------------|
| 1        | .861*      | 100.0         | 100.0        | .680                  |
Based on Table 5 obtained an eigenvalue of 0.861 and canonical correlation value is 0.680 which indicates the closeness of the independent variables to the status of diabetics.

### 3.3.2. Average similarity between groups
The average vector test between groups 1 (prediabetes) and group 2 (diabetes) using Wilk’s Lambda Test obtained Wilks’ Lambda value is 0.537 with p-value < 0.05 so there is a significant difference between patients.

### 3.3.3. Average similarity between independent variables
This test aims to see the independent variables that significantly influence the formation of groups. If there are > 50% significant independent variables, discriminant analysis can be continued.

#### Table 6. Average similarity independent variables.

| Independent Variables       | F   | p-value |
|-----------------------------|-----|---------|
| Age                         | 60.119 | .000 |
| Weight                      | 12.405 | .001 |
| Cholesterol level           | 4.048 | .046 |
| Triglyceride                | 3.284 | .072 |
| Systolic blood pressure     | .015  | .902  |
| Diastolic blood pressure    | 10.500 | .001 |
| Uric acid level             | 31.951 | .000 |

If α = 0.05, the significant independent variables are age, weight, cholesterol level, diastolic blood pressure, and uric acid levels.

### 3.3.4. Model of discrimination function
The results of discriminant function (Y) with 4 independent variables were obtained:

\[
Y = 0.681 X_1 - 0.644 X_2 + 0.586 X_6 + 0.564 X_7
\]

Variables that influence the grouping of diabetic patients or prediabetes are \(X_1\): age, \(X_2\): weight, \(X_6\): diastolic blood pressure, and \(X_7\): uric acid level.

### 3.3.5. Accuracy of classification
The classification results can be seen in Table 7.

#### Table 7. Classification results.

| Prediabetes/ Diabetess | Predicted | Total |
|------------------------|-----------|-------|
|                        | 1         | 2     |       |
| Original Count 1       | 78        | 9     | 87    |
|                        | 15        | 70    | 85    |
| %                      | 89.7      | 10.3  | 100.0 |
|                        | 17.6      | 82.4  | 100.0 |

The classification accuracy is APER = 14.0% which shows the number of the accuracy of grouping is 86% quite high, so the accuracy of classification model is quite high.
3.3.6. **Diabetics classification graph.**

From above it is seen that diabetics have a lower age and their uric acid levels are lower than prediabetes.

From figure 3. it is seen that diabetics have a lower age and have a higher body weight than prediabetes.

From figure 4. it is seen that diabetics have a higher body weight and lower uric acid levels than prediabetes.

4. **Conclusion**
Based on the results of the above study it can be concluded that there are significant differences in symptoms of diabetics and prediabetes. Variables that have a substantial effect on grouping prediabetes and diabetics are age, weight, systolic blood pressure, diastolic blood pressure, and uric acid levels. With discriminant analysis based on the independent variables, the accuracy of the classification of diabetics and prediabetes is quite high, which is 86%.
References

[1] R S E Sayed 2018 Linear Discriminant Analysis for An Efficient Diagnosis of Heart Disease via Attribute Filtering Based on Genetic Algorithm *Journal of Computer*.

[2] J Bian et. al. 2018 Early detection of diseases using electronic health records data and covariance-regularized linear discriminant analysis *IEEE EMBS BHI*.

[3] M Asamoah-Boaheng 2014 Application of Discrimination and Classification on Diabetes Mellitus Data *International J. of Applied Science and Technology* 4(6).

[4] M Hyodo and T Kubokawa 2014 A Variable Selection Criterion for Linear Discriminant Rule and its Optimality in High Dimensional Setting *Journal of Multivariate Analysis* 123, 364–379.

[5] R Banu 2016 Predicting Thyroid Disease using Linear Discriminant Analysis (LDA) Data Mining Technique *Communications on Applied Electronic Journal*. 