JSCDM

http://penerbit.uthm.edu.my/ojs/index.php/jscdm

e-ISSN : 2716-621X

Journal of Soft Computing and Data Mining

Identifying Genes Related to Diabetes Mellitus Using Penalized Logistic Regression

Masithoh Yessi Rochayani[1*], Arief Rachman Hakim[1], Sugito[1]

[1]Department of Statistics,
Universitas Diponegoro, Semarang, 50275, INDONESIA

*Corresponding Author

DOI: https://doi.org/10.30880/jscdm.2023.04.02.003
Received 20 August 2023; Accepted 19 September 2023; Available online 04 October 2023

Abstract: Identification of genes associated with Diabetes Mellitus is important for early detection of this disease. This study tried to find some potential genes related to T2DM. The dataset used was GSE25462 and the method used was penalized logistic regression, specifically Lasso. The top eight selected genes were ABRA, EVX1, MIR7-3HG, SAYSD1, SLC26A1, SRGAP3, WFDC1, and 240244_at. The training data reaches the accuracy and kappa of 1 for the model with 8 genes. But, when the model is used for testing data the maximum accuracy is 0.9 and the maximum kappa is 0.615, obtained in models with 14 genes. This happened because the dataset lacked samples of the positive class. The use of ensemble learning methods is recommended to combine predictive results. The role of some genes we found in T2DM remains unclear. Biology researchers can further study the role of these genes in T2DM.

Keywords: Diabetes Mellitus, gene expression, penalized logistic regression, Lasso

1. Introduction

Diabetes Mellitus (DM) is a chronic disease indicated by high blood glucose. DM is classified as type 1 or type 2 based on the primary cause. Type 2 DM (T2DM) is the more common type of DM, manifested by insulin resistance or insulin secretion disorder [1]. Uncontrolled DM can cause serious health complications, including eye damage, skin problem, stroke, and heart disease.

DM is believed to be a public health problem worldwide as this number is predicted to grow sharply over time. There were around 450 million people diagnosed with diabetes worldwide in 2017 and predicted to increase to 693 million by 2045 [2]. To prevent this phenomenon from happening, it is necessary to do early detection, such as by monitoring symptoms from the start or predicting the possibility of diabetes in offspring with diabetes. Several studies have predicted DM disease with various methods. Joshi and Dhakal [3] predicted T2DM using logistic regression and classification trees. The data used is the Pima Indian dataset. Predictor variables used include glucose, insulin, BMI, blood pressure, age, skin thickness, and pregnancy. The classification tree modeling shows that glucose becomes a split root node. Sillagan and Fitriyani [4] predicted DM predict DM based on symptoms. The dataset used contains 15 nominal scale attributes, some of which are gender, polyuria status (a condition of excess urine), polydipsia status (a condition of being easily thirsty), and obesity status. They use the C4.5 algorithms to build the model. From the modeling results, it is found that polydipsia is a split root node.

The fast development of microarray technology has made it simpler to observe the expression data of very large number of genes. The goal of studying the genes associated with a disease is to serve as a targeted therapy or to study the possibility that the disease is genetically inherited. However, data generated from microarray experiments are high-

*Corresponding author: yessirochayani@lecturer.undip.ac.id

dimensional. Due to the complexity of dealing with thousands of genes, the selection of genes is a crucial topic in microarray data analysis. Therefore, preprocessing steps are required before further analysis.

Feature selection is a strategy for modeling high-dimensional data. Most researchers combine feature selection and machine learning to model diseases based on gene expression data. There are three main classes of feature selection approaches, namely wrapper, filter, and embedded. Embedded is preferred over filters and wrappers because it does not tend to overfit and has fast computation. In the current study, an embedded approach is applied to identify genes associated with T2DM.

2. Related Work

2.1 Feature Selection for Gene Expression Data

Finding genes associated with certain diseases has been carried out recently. Morais-Roudrigues et al. (2020) [5] identified genes related to breast cancer using modified logistic regression. Gumaei et al. (2021) [6] carried out gene selection to diagnose prostate cancer. They use the Correlated Feature Selection method to select genes associated with prostate cancer. Hamidi et al. (2021) [7] used Lasso and Elastic Net to find potential miRNA biomarkers for ovarian cancer. Gilani et al. (2022) [8] identify potential biomarkers for gastric cancer using Boruta feature selection.

Penalized logistic regression, including Lasso, is a popular method to select genes from microarray data. Penalized regression is included in the embedded approach. Penalized regression, Lasso, has been successfully used to determine genes associated with ovarian cancer [9] and breast cancer [10], where the genes obtained are in line with biological theory.

2.2 Exploring Genes Related to T2DM

Some research had also identified genes related to T2DM and modeled it. Bian et al. (2022) [11] modeled T2DM disease from gene expression data using Lasso. The dataset contains 20 samples of pancreas tissue from 10 donors with diabetic conditions and 10 donors with non-diabetic conditions. The results give 8 selected genes. The obtained model reached an AUC of 0.84 for the training data and 0.67 for the test data, showing that the Lasso might be a good method to investigate biological markers to diagnose T2DM. Meanwhile, Zhu et al. (2020) [12] conducted a differential expression analysis on the GSE26168 dataset. The dataset contains 60 samples and recorded expression of 24526 genes. The differential expression analysis yielded 301 upregulated genes and 680 downregulated genes.

3. Methods

3.1 Data Collection

The dataset used is GSE25462 which was downloaded from the GEO database. The samples of GSE25462 are obtained from skeletal muscle tissue from 10 subjects with T2DM and 40 subjects without T2DM (25 subjects have a family history of T2DM, and 15 subjects have no family history of T2DM). The dataset provides the expression of 54675 genes. The gene expression values had processed using MAS5.0 normalization.

3.2 Penalized Logistic Regression

A logistic model is a statistical model that models the relation between the categorical response variable and predictor variables. In logistic regression models, there is a linear combination of predictor variables with log odds of the probability of an event. In binary logistic regression, the dependent variable has a binary value, coded by "0" and "1", while the predictor variables can each be a binary (discrete) variable or a continuous variable (any real value). The binary logistic regression model is expressed by Equation (1).

$$\text{logit}(\pi(x_i)) = \log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{1}$$

where $\pi(\mathbf{x}_i)=P[Y=1|\mathbf{X}=\mathbf{x}]$ represent the probability of success, $1-\pi(\mathbf{x}_i)=P[Y=0|\mathbf{X}=\mathbf{x}]$ represent the probability of fail. Using exponential transformation and some algebraic manipulations, the expression in Equation (2) is obtained.

$$\pi(x_i) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}{1+\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)} \tag{2}$$

The vector of coefficients $\boldsymbol{\beta}$ is obtained by maximizing the likelihood function expressed in Equation (3).

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left(1-\pi(x_i)\right)^{1-y_i} \tag{3}$$

The estimated logistic regression coefficients are stated by Equation (4)

$$\hat{\boldsymbol{\beta}} = \arg\min \left[-\left(\sum_{i=1}^{n} y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) - \log\left(1+\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right) \right) \right] \tag{4}$$

Penalized regressions work by adding a penalty to optimize the objective function. Equation (5) expresses the Lasso penalty function,

$$p\left(\beta_j, \lambda\right) = \lambda \sum_{j=1}^{p} \left|\beta_j\right| \tag{5}$$

where $\lambda > 0$ is a tuning parameter. Given the logistic regression model (1), the negative log-likelihood with the Lasso penalty is expressed by Equation (6) [13].

$$J(\beta_j) = -\frac{1}{n}\left(\sum_{i=1}^{n} y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) - \log\left(1+\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right) \right) + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \tag{6}$$

By replacing the binary response variables {0, 1} in terms of sign variables {1, +1}, the penalized (negative) log-likelihood is stated by Equation (7).

$$J(\beta_j) = \frac{1}{n}\left(\sum_{i=1}^{n} \log\left(1+\exp\left[-y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right] \right) \right) + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \tag{7}$$

Coefficients of Lasso, Equation (8), are obtained by solving Equation (7).

$$\hat{\boldsymbol{\beta}}_{Lasso} = \arg\min \left[J\left(\beta_j\right) \right] \tag{8}$$

The optimum tuning parameter (λ) is determined by using the k-fold Cross Validation (CV). The k-fold CV divides the dataset into subsets of the same size. A total of *K*-1 subsets are used to train the model and a subset is used to validate the model. This process is performed k times and therefore each data subset will have a turn to become test data. The use of k=5 or k=10 is recommended by [13]. The regularized logistic regression model of each fold is evaluated using binomial deviance. The smaller the binomial deviance the better the model [14].

3.3 Evaluation Metrics

The performance of the classification model was evaluated using the confusion matrix. Table 1 presents the confusion matrix of a case involving two classes. TP and TN respectively denoted true positive and true negative, stating the sample quantity in positive class and negative class accurately classified. FN and FP stand for false negative and false positive respectively, stating the sample quantity in the positive class and negative class is inaccurately classified.

Table 1 - Confusion matrix

Actual	**Predicted**	
	Positive (T2DM)	**Negative (Normal)**
Positive (T2DM)	TP	FP
Negative (Normal)	FN	TN

Accuracy is a popular metric to measure classification performance. It states the percentage of data correctly classified, and range from 0 to 1. Accuracy is calculated using Equation (9). Besides using accuracy, Cohen's kappa coefficient is also used to measure the goodness of the model. Kappa is a statistic to measure the level of inter-rater agreement. The value of Kappa is in the range [0, 1]. Kappa is calculated using Equation (10).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{9}$$

$$\text{Kappa} = \frac{p_0 - p_e}{1 - p_e} \tag{10}$$

where

$$p_0 = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{11}$$

$$p_e = \frac{(\text{TP} + \text{FN})(\text{TP} + \text{FP}) + (\text{FP} + \text{TN})(\text{FN} + \text{TN})}{(\text{TP} + \text{FN} + \text{FP} + \text{TN})^2} \tag{12}$$

4. Results and Discussions

First, the dataset is split with a proportion of 80% for the training data and 20% for the test data. This ratio resulted in 40 observations in training data (8 DM and 32 normal) and 10 observations in test data (2 DM and 8 normal). The training data is used to make the classification model, and the test data is used to validate the model.

4.1 Gene Selection

The gene selection was executed utilizing glmnet. In this study, 100 iterations were used. The tuning parameter (λ) decreases exponentially as the iteration index increases as shown in Fig. 1. Suppose the iteration index is denoted by h, where h=1, 2, …, H. The glmnet fixed the $\lambda_H = 0.01 \times \lambda_1$. The ratio to decrease the tuning parameter can be written by Equation (13).

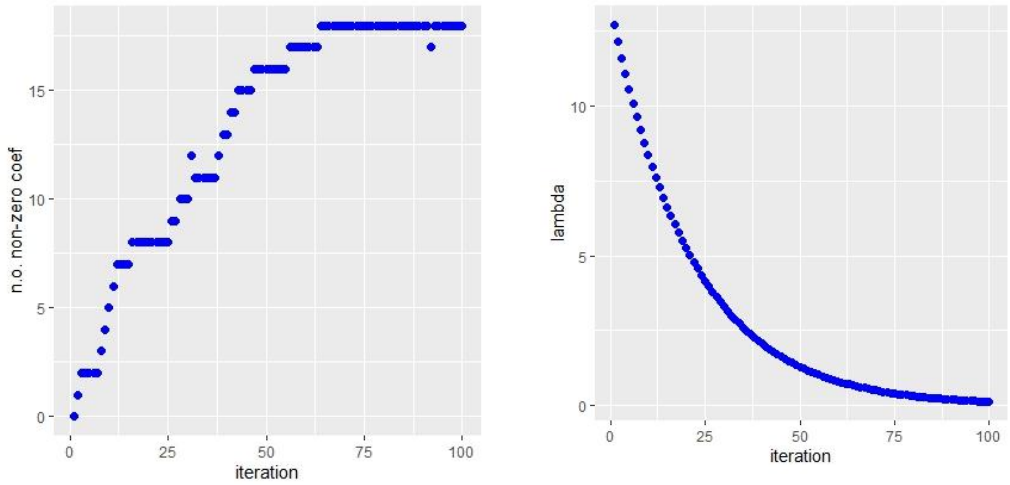
$$r = \left(\frac{\lambda_1}{\lambda_H}\right)^{\frac{1}{H-1}} \tag{13}$$

From the GSE25462 dataset, the λ in the first iteration is 12.7054 and in the 100th iteration is 0.1270542. Therefore, the ratio is

$$r = \left(\frac{12.7054}{0.1270542}\right)^{\frac{1}{99}} = 1.047616 \tag{14}$$

This ratio also applies to all datasets when 100 iterations (H = 100) are used, because glmnet uses $\lambda_{100} = 0.01 \times \lambda_1$. Or in general, can be written

$$r = \left(\frac{\lambda_1}{\lambda_H}\right)^{\frac{1}{H-1}} = \left(\frac{100\lambda_H}{\lambda_H}\right)^{\frac{1}{H-1}} = 100^{\frac{1}{99}} = 1.047616 \tag{15}$$



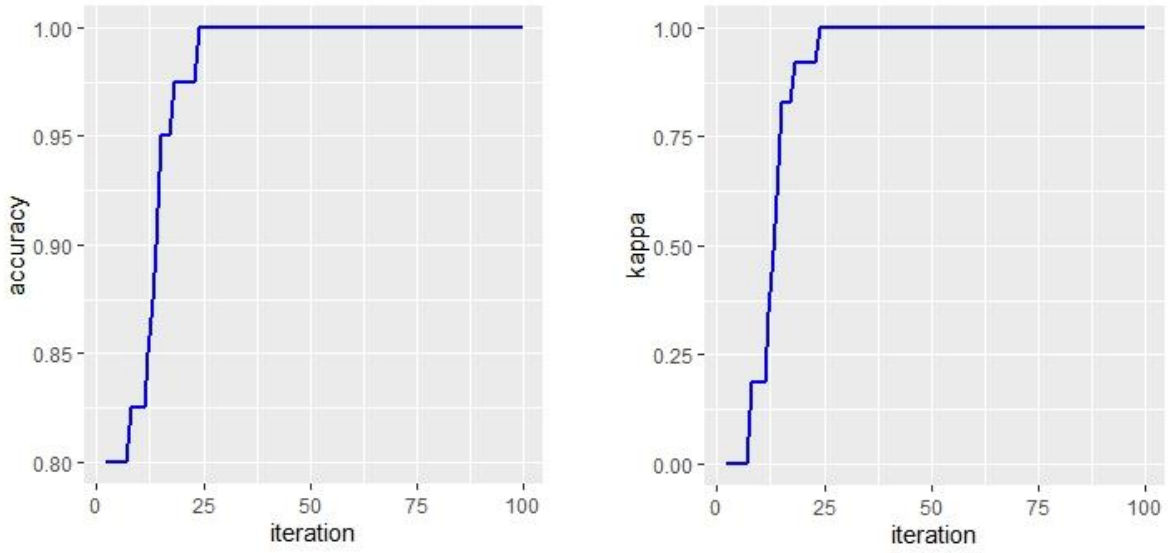
(a) (b)

Fig. 1 - The plot of (a) the number of non-zero coefficients and; (b) tuning parameter

The optimum λ is commonly estimated using the k-fold CV. However, the data used in this study were only 40 observations. Splitting the data into 5 folds only gives 8 observations on each fold. This size is too small to make a model. For this reason, the k-fold CV is not used to determine the optimum λ. In this work, the best model is determined by the evaluation of the model on training data and test data.

4.2 Exploring The Best Model

The model obtained is evaluated using accuracy and kappa. In the models involving a gene (2nd iteration) and 2 genes (3rd to 7th iteration), get an accuracy of 0.8. This value looks good, but no positive class (T2DM) is classified correctly. In other words, TP=0. Kappa is better to provide acceptable value. The kappa for those models is 0. For this reason, the iteration is continued. In the 8th iteration, 3 genes are obtained, in the 9th iteration 4 genes, in the 10th iteration 5 genes, and in the 11th iteration (6 genes). The accuracy and kappa of those models are 0.825 and 0.186, respectively. Those models only correctly classify an observation of the positive class. The iteration is continued. In the 24th iteration, the accuracy and kappa reach the values of 1, which means that all observed objects were classified correctly. The number of genes selected in the 24th iteration is 8 genes with λ= 4.3586. Accuracy and kappa values for training data can be seen in Fig. 2 and the list of 8 selected genes is presented in Table 2.



(a) (b)

Fig. 2 - The (a) accuracy and; (b) kappa of training data

The first 8 selected are then investigated. The first was the gene with probe ID 207914_x_at which was selected in the 2nd iteration. In the 3rd iteration, the gene 1552731_at was selected. The 3rd and 4th selected gene was 223973_at and 219478_at which were first to appear on the 8th and 9th iterations, respectively. The 5th gene which was first to appear on the 10th was 232992_at. The genes with probe ID 240244_at, 205058_at, and 209794_at firstly appear in the 11th, 12th, and 16th iterations respectively. Table 2 shows the selected genes for some iterations. Table 3 presents the Lasso coefficient and gene name of the 24th iteration, which resulted in 8 selected genes which gives accuracy and kappa of 1. The boxplot diagrams of those genes are shown in Fig. 3. Based on Fig. 3, the expression of the selected genes is higher in the T2DM condition. This boxplot diagram matches the results in Table 3 where the coefficient is positive which indicates the higher the gene expression, the higher the probability of an individual suffering from T2DM

In this study, 8 selected genes were obtained, but these genes were different from the 8 genes obtained by Bian et al. (2022) [11]. Because the datasets used are different and the genes that feature genes are also very large, up to tens of thousands, the possibility of getting the same results is very small. However, if these tuning parameters are continuously reduced, thousands of genes will be obtained and it is very possible to find overlapping genes. In Alhumaydhi's research (2022) [15] which carried out gene selection related to T2DM disease using 2 datasets, GSE25724 and GSE101931. Using differential expression analysis with a threshold of p-value <0.05, he found 287 overlapping genes. Because the genes he obtained were so many, he only mentioned a few genes. However, none of the genes mentioned are the same as the genes selected using the Lasso in this study.

Even though the genes we found were different from previous studies, some of the genes we found were related to diabetes. Petrie et al. (2014) [16] stated that ABRA expression increases in diabetic muscles. Mao et al. (2023) [17] reveal that the dysfunction of pancreatic β-cells leads to diabetes mellitus development. They conclude that the MIR7-3HG encodes a short peptide, and preserves pancreatic β-cells from dexamethasone-induced dysfunction. Lee et al.

(2023) [18] reported an increase in the expression of genes WFDC1 protruding in the platelets of the patient with periodontitis, and it showed more significant DEGs in the platelets of the patient with periodontitis and DM.

Table 2 - Lambda and the number of selected genes

Iteration	Lambda	Number of Selected Genes	Probe ID of Selected Genes
2nd	12.1279	1	207914_x_at
3rd	11.5767	2	207914_x_at, 1552731_at,
8th	9.1743	3	207914_x_at, 1552731_at, 223973_at
9th	8.7574	4	207914_x_at, 1552731_at, 223973_at, 219478_at
10th	8.3593	5	207914_x_at, 1552731_at, 223973_at, 219478_at, 232992_at
11th	7.9794	6	207914_x_at, 1552731_at, 223973_at, 219478_at, 232992_at, 240244_at
12th	7.6167	7	207914_x_at, 1552731_at, 223973_at, 219478_at, 232992_at, 240244_at, 205058_at
16th	6.3235	8	207914_x_at, 1552731_at, 223973_at, 219478_at, 232992_at, 240244_at, 205058_at, 209794_at
24th	4.3586	8	207914_x_at, 1552731_at, 223973_at, 219478_at, 232992_at, 240244_at, 205058_at, 209794_at

Table 3 - Lasso Coefficient and gene name of top eight genes for λ=4.3586

Gene Probe ID	Coefficient	Gene Name
207914_x_at	0.0018123	EVX1 (even-skipped homeobox 1)
1552731_at	0.0000839	ABRA (actin binding Rho activating protein)
223973_at	0.0001793	MIR7-3HG (MIR7-3 host gene)
219478_at	0.0004027	WFDC1 (WAP four-disulfide core domain 1)
232992_at	0.0019610	SAYSD1 (SAYSVFN motif domain containing 1)
240244_at	0.0000963	-
205058_at	0.0005908	SLC26A1 (solute carrier family 26 member 1)
209794_at	0.0004131	SRGAP3 (SLIT-ROBO Rho GTPase activating protein 3)

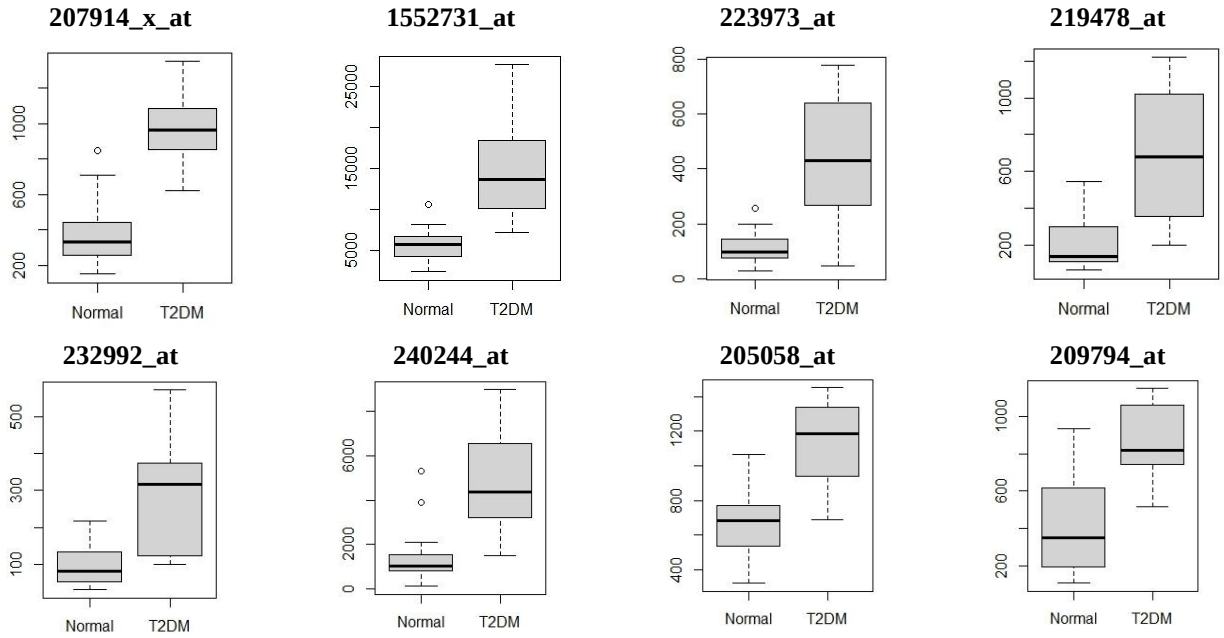
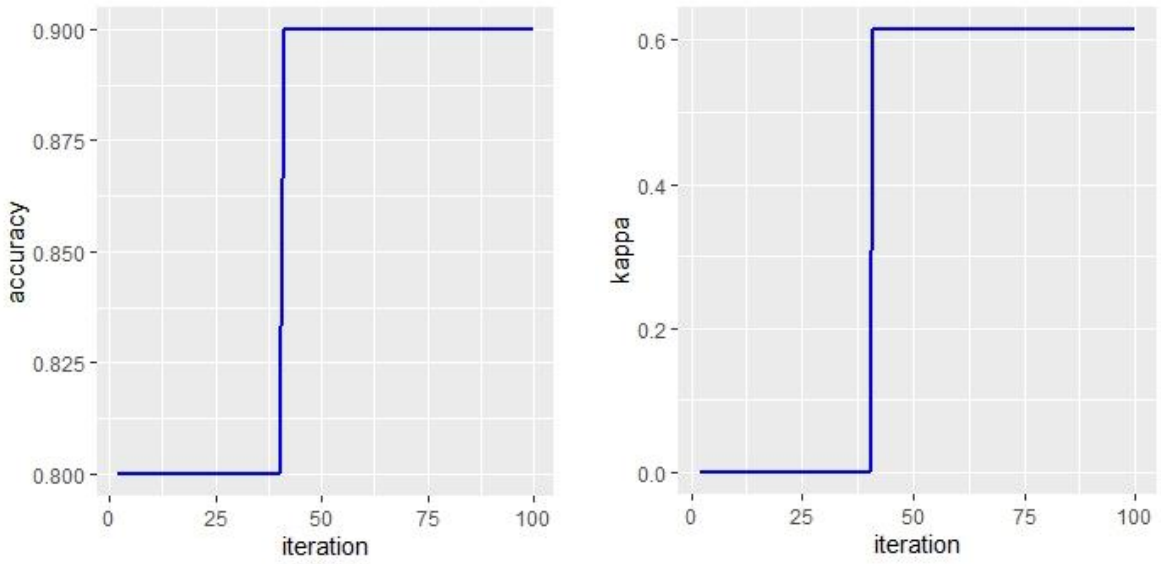


Fig. 3 - Boxplot diagram of the top eight selected genes

4.3 Model Validation

The model is then validated using test data. The model of the 24th iteration (λ=4.3586) gives an accuracy of 0.8 for the test data. But the kappa for this model is zero. This happens because there is no positive class (T2DM) that is classified correctly. Because of that, the iteration is increased. In the 41st iteration, an accuracy of 0.9 and a kappa of 0.615 is obtained with the number of selected genes being 14 genes and λ=1.9766. The probe ID of those genes are 1552731_at, 1554744_at, 1554923_at, 1563070_at, 205058_at, 207914_x_at, 209794_at, 216871_at, 219478_at, 219734_at, 224971_at, 232992_at, 238127_at, and 240244_at. Kappa is very small because in the confusion matrix, only an object from the positive class is classified correctly. In the test data, there are only 2 objects of the positive class (T2DM), one is classified correctly, and one is misclassified. It is necessary to have a large number of samples to make the classifier can learn more about the data patterns. The accuracy and kappa values for the training data are presented in Fig. 4.



(a) (b)

Fig. 4 - The (a) accuracy and; (b) kappa of test data

5. Conclusion

This study tried to find some candidate genes associated with T2DM using penalized logistic regression. The top eight selected genes were ABRA, EVX1, MIR7-3HG, SAYSD1, SLC26A1, SRGAP3, WFDC1, and 240244_at. The training data reaches the accuracy and kappa of 1 for the model with 8 genes. But, when the model is used for testing data the maximum accuracy is 0.9 and the maximum kappa is 0.615, obtained in models with 14 genes. This happened because the dataset lacked samples of the positive class. The number of samples is only 50 which is split into 40 training data and 10 test data. This sample size in test data is small enough to validate the model. For this reason, the use of ensemble learning methods such as bagging and random forest to multiply models and combine predictive results. We also validated several genes based on biological theory. But the role of other genes in T2DM remains unclear. Therefore, researchers in the field of biology can further study the role of these genes in T2DM.

Acknowledgement

The authors would like to thank the Department of Statistics, Universitas Diponegoro, for supporting this work.

References

[1] Schofield, C. J., & Sutherland, C. (2012). Disordered insulin secretion in the development of insulin resistance and Type 2 diabetes. *Diabetic medicine*, *29*(8), 972-979.

[2] Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. I. D. F. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes research and clinical practice, 138, 271-281.

[3] Joshi, R. D., & Dhakal, C. K. (2021). Predicting type 2 diabetes using logistic regression and machine learning approaches. International journal of environmental research and public health, 18(14), 7346.

[4] Siallagan, R. A. (2021). Prediksi penyakit diabetes mellitus menggunakan algoritma c4. 5. Jurnal Responsif: Riset Sains dan Informatika, 3(1), 44-52.
[5] Morais-Rodrigues, F., Silv́erio-Machado, R., Kato, R. B., Rodrigues, D. L. N., Valdez-Baez, J., Fonseca, V., & Dos Santos, M. A. (2020). Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression. Gene, 726, 144168.
[6] Gumaei, A., Sammouda, R., Al-Rakhami, M., AlSalman, H., & El-Zaart, A. (2021). Feature selection with ensemble learning for prostate cancer diagnosis from microarray gene expression. Health Informatics Journal, 27(1), 1460458221989402.
[7] Hamidi, F., Gilani, N., Belaghi, R. A., Sarbakhsh, P., Edgünlü, T., & Santaguida, P. (2021). Exploration of potential miRNA biomarkers and prediction for ovarian cancer using artificial intelligence. Frontiers in Genetics, 12, 724785.
[8] Gilani, N., Arabi Belaghi, R., Aftabi, Y., Faramarzi, E., Edgünlü, T., & Somi, M. H. (2022). Identifying potential miRNA biomarkers for gastric cancer diagnosis using machine learning variable selection approach. Frontiers in Genetics, 12, 779455.
[9] Sa'adah, U., Rochayani, M. Y., & Astuti, A. B. (2021). Knowledge discovery from gene expression dataset using bagging lasso decision tree. Indonesian Journal of Electrical Engineering and Computer Science, 21(2), 1151-1159.
[10] Rochayani, M. Y., Sa'adah, U., & Astuti, A. B. (2020). Two-stage Gene Selection and Classification for a High-Dimensional Microarray Data. Jurnal Online Informatika, 5(1), 9-18.
[11] Bian, Q., Li, H., Wang, X., Liang, T., & Zhang, K. (2022). Multiomics Integrated Analysis Identifies SLC24A2 as a Potential Link between Type 2 Diabetes and Cancer. Journal of Diabetes Research, 2022.
[12] Zhu, H., Zhu, X., Liu, Y., Jiang, F., Chen, M., Cheng, L., & Cheng, X. (2020). Gene expression profiling of type 2 diabetes mellitus by bioinformatics analysis. Computational and Mathematical Methods in Medicine, 2020.
[13] Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity: the lasso and generalizations. CRC press.
[14] Agresti, A. (2012). Categorical data analysis (Vol. 792). John Wiley & Sons.
[15] Alhumaydhi, F. A. (2022). Integrated computational approaches to screen gene expression data to determine key genes and therapeutic targets for type-2 diabetes mellitus. Saudi Journal of Biological Sciences, 29(5), 3276-3286.
[16] Petrie, M. A., Suneja, M., Faidley, E., & Shields, R. K. (2014). A minimal dose of electrically induced muscle activity regulates distinct gene signaling pathways in humans with spinal cord injury. PloS one, 9(12), e115791.
[17] Mao, X., Zhou, J., Kong, L., Zhu, L., Yang, D., & Zhang, Z. (2023). A peptide encoded by lncRNA MIR7-3 host gene (MIR7-3HG) alleviates dexamethasone-induced dysfunction in pancreatic β-cells through the PI3K/AKT signaling pathway. Biochemical and Biophysical Research Communications, 647, 62-71.
[18] Lee, H., Joo, J. Y., Kang, J., Yu, Y., Kim, Y. H., & Park, H. R. (2023). Single-cell analysis of platelets from patients with periodontitis and diabetes. Research and Practice in Thrombosis and Haemostasis, 7(2), 100099.