College of Basic Education Research Journal

www.berj.mosuljournals.com

Identifying Factors that Affect Diabetes through Regression Analysis

Hussein A. Hashem
Dept. of Mathematics, College of Science, University of Duhok, Kurdistan Region, Iraq
Azad A. Shareef
Dept. of Statistics, College of Administration and Economics, University of Duhok, Kurdistan Region, Iraq
Sherzad M. Ajeel
Dept. of Mathematics, College of Science, University of Duhok, Kurdistan Region, Iraq

Article Information

Article history:

Received: October 1.2024

Reviewer: November 11.2024

Accepted: November 11.2024

Key words: Diabetes data, SCAD, LASSO

Correspondence:

Abstract

Globally, diabetes is a significant cause of both death and morbidity, affecting a wide range of populations irrespective of age, gender, or region. The purpose of this paper is to draw attention to this urgent public health issue and to increase awareness of it among world leaders and policymakers. The current study looks into the Collins, relationship between diabetes and age, gender, height, and weight—four important characteristics. The Azadi Hospital in Dohuk provided the study's data. Finding out whether gender, age, weight, or height are related to high prevalence of diabetes as the main goal of study. To achieve this, the researchers utilized a variety of multiple regression analysis techniques, including Least Absolute Shrinkage and Selection Operator (LASSO), Ordinary Least Square (OLS), Minimax Concave Penalty (MCP) Regression, Quantile Regression (QR), and Smoothly Clipped Absolute Deviation (SCAD) penalty. Age, weight, and diabetes were found to be significantly correlated by data analysis

ISSN: 1992 – 7452

تحديد العوامل المؤثرة على مرض السكري من خلال تحليل الانحدار

حسين عبد الرحمن هاشم
قسم الرياضيات، كلية العلوم، جامعة دهوك، إقليم كردستان، العراق

ازاد عادل شريف
قسم الإحصاء، كلية الإدارة والاقتصاد، جامعة دهوك، إقليم كردستان، العراق

شيرزاد محمد عجيل
قسم الرياضيات، كلية العلوم، جامعة دهوك، إقليم كردستان، العراق

مستخلص البحث

قد ثبت عالميًا أن مرض السكري هو سبب مهم للوفاة والمرض، حيث يؤثر على مجموعة واسعة من السكان بغض النظر عن العمر أو الجنس أو المنطقة. والغرض من هذا البيان هو لفت الانتباه إلى هذه القضية الصحية العامة العاجلة وزيادة الوعي بها بين قادة العالم وصناع السياسات. وتبحث الدراسة الحالية في العلاقة بين مرض السكري والعمر والجنس والطول والوزن – أربع خصائص مهمة. قدم مستشفى آزادي في دهوك بيانات الدراسة. وكان الهدف الرئيسي للدراسة هو معرفة ما إذا كان الجنس أو العمر أو الوزن أو الطول مرتبطًا بالانتشار المرتفع لمرض السكري. ولتحقيق هذه الغاية، استخدم الباحثون مجموعة متنوعة من تقنيات تحليل الانحدار المتعدد، بما في ذلك الاختيار المطلق الأقل(انحدارلاسو(LASSO))، وطريقة المربعات الصغرى (OLS)، وانحدار (MCP)، والانحدار الكمي(QR) ، وانحدار (SCAD) وقد وجد أن العمر والوزن ومرض السكري مرتبطان بشكل كبير من خلال تحليل البيانات.

1. Introduction

One of the most prevalent illnesses in humans, diabetes has grown to be a major global public health issue. In 2017, there were over 1.37 million fatalities worldwide from diabetes, which affected about 450 million people with a diagnosis (Cho et al., 2018). In 2020, diabetes was the seventh greatest cause of mortality in the US, affecting over 100 million persons. Diabetes already affects 10% of US people, and if the current trend continues, up to 30% of US adults may develop the disease by 2050 . Patients with diabetes are more likely to experience health issues that can cause malfunction and long-term tissue damage, including kidney failure, vision loss, heart disease, stroke, early mortality, and amputation of the feet or legs (Krasteva et al., 2011).Furthermore, there are significant financial expenses linked to the illness. In the United States, the overall anticipated cost of diabetes diagnoses climbed from USD 188 billion in 2012 to USD 237 billion in 2017. Over the same time period, the excess medical expenses per person related to diabetes rose from USD 8417 to USD 96001. Additionally, workers with diabetes may experience a decrease in productivity.

It's possible that a person with a high risk of diabetes is unaware of the risk factors. Researchers are interested in identifying the most prevalent risk factors for diabetes because of its great prevalence and severity, which could be caused by a number of variables.Early diabetes prediction and risk factor identification have proven essential in lowering the financial burden (Habibi et al., 2015) and problems (Alghamdi et al., 2017) associated with diabetes, and they are advantageous from the standpoints of public health and clinical practice (Ryden et al., 2007). In a similar vein, research indicates that screening high-risk individuals helps identify the demographic groups who will benefit most from diabetes prevention efforts (Tuso, 2014). Early action are crucial in creating successful preventative strategies and may aid in avoiding difficulties and enhancing quality of life (Gregg et al., 2001). A rising body of research suggests that changing one's lifestyle can either prevent or postpone type 2 diabetes (Knowler et al., 2002). An poor diet, aging, family history, ethnicity, obesity, sedentary lifestyle, and a history of gestational diabetes are thought to be the primary risk factors for diabetes (Habibi et al., 2015). According to earlier research, diabetes is also linked to sex, body mass index (BMI), pregnancy, and metabolic condition (Engelgau et al., 2000).

To determine the appropriate clinical therapy for patients, prediction models can test for pre-diabetes or individuals at higher risk of getting the disease. To estimate the risk variables of incident diabetes, a variety of predictive equations have been proposed (Naz and Ahuja, 2020). For example, Razavian et al.,(2015) created logistic regression-based prediction models for the occurrence of type 2 diabetes, while Heikes et al. (2008) investigated a tool to forecast the likelihood of diabetes in the US using undiagnosed and pre-diabetic data. In order to identify those who are at a high risk of developing diabetes, these models also assist in screening. In Luzhou, China, Zou et al. (2018) employed machine learning techniques to forecast diabetes, and they validated the models using a five-fold cross-validation. Using deep learning algorithms, Nguyen et al. (2019) forecast the development of diabetes, indicating that advanced techniques could enhance model performance. However, a number of other studies

have demonstrated that, for the purpose of predicting illness risk, logistic regression performs no better than machine learning techniques. In a similar vein, Anderson et al. (2016) discovered that the logistic regression model had a greater accuracy when combined with machine learning methods. These mostly rely on evaluating diabetes risk factors, such as personal and household traits; yet, the absence of an impartial and objective assessment remains a problem (Collins et al., 2011). Furthermore, there is rising worry that such predictive models are underdeveloped as a result of small sample sizes, missing data, improperly stated statistical models, and inadequate covariate selection (Mikolajczyk et al., 2008). Because of this, very few risk prediction algorithms have been applied consistently in clinical settings. Geographical location, data accessibility, and ethnicity all significantly affect the quality and dependability of these prediction tools and equations (Nguyen et al. 2019, Sherzad et al. 2023, Azad et al. 2024). Risk factors for one ethnic group might not apply to others; for instance, the Pima Indian community is said to have a greater prevalence of diabetes. In order to determine whether a person is at risk of acquiring diabetes depending on particular diagnostic variables, this study analyzes the Pima Indian dataset (Bennett et al. 1971).

This paper's main goal is to investigate the correlation between diabetes and gender, age, weight, and height utilising data from Azadi Hospital in Dohuk.

The remainder of the document is structured as follows: The methods work are discussed in Section 2. The data and the methodology are presented in Sections 3 and 4., and Section 5 wraps up the experiment as a whole.

2. Methods

2.1. Multiple Linear Regression

The multiple linear regression model is the most popular type of regression model. The weighted sum of the explanatory (independent) variables $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{p}$ (let's say) with unknown parameters $\beta_1, \beta_2, \ldots, \beta_p$ can be considered the response (dependent) variable y in this model. The multiple linear regression model can generally be expressed as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \, , i = 1,2, \ldots, n$$

where n is the number of observations and p is the number of explanatory variables.

The model's matrix representation is $y = X\beta + \varepsilon$

$y = (y_1, \ y_2 \, , \ \ldots, y_n)'$ is the vector of the response variable, $\beta = \left(\beta_0, \ \beta_1 \, , \ \ldots, \beta_p\right)'$ is the vector of the unknown parameters,

$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ & \vdots & & \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}$ is an $n \times \ (p \ + \ 1)$ matrix of explanatory variables and

$\varepsilon' = (\varepsilon_1, \ \varepsilon_2 \, , \ \ldots, \varepsilon_n)'$ is the vector of random errors.

Numerous presumptions must be met in this case, and they can be summed up as follows:

a. The variance σ^2 of the error terms is constant and they are independent.
b. With a variance-covariance matrix and a mean vector of zero, the error terms are normally distributed $\varepsilon \sim N(0, \sigma^2 I_n)$.
c. The matrix X is of full column rank .

The most often used estimating technique is the ordinary least squares method, which minimizes the sum of squared deviations $\varepsilon' \varepsilon$ where (Drapper and Smith,1981).

$$\varepsilon' \varepsilon = (y - X\beta)'(y - X\beta)$$

The solution of this equations is $\beta_{ols} = (X'X)^{-1}X'y$

2.2. Minimax Concave Penalty (MCP) method

Regression analysis uses the Minimax Concave Penalty (MCP), a statistical method, to solve the problem of bias in sparse models. Zhang (2010) established the MCP technique, which uses the MCP penalty function to choose variables in linear regression models. Since it lessens the issue of inconsistent variable selection that LASSO frequently has, it is thought to be an improvement over LASSO. The following formula can be used to produce the MCP estimator.

$$\hat{\beta}_j^{MCP} = \underbrace{argmin}_{p} [\|Y - X\beta\|^2 + \sum_{j=1}^{p} P_{\lambda,\gamma}^{MCP}]$$

Where: $\sum_{j=1}^{p} P_{\lambda,\gamma}^{MCP}$ the MCP penalty function.

The MCP function has the following format:

$$p_\lambda(\lceil\beta\rceil) = \begin{cases} \lambda\left(|\beta| - \frac{|\beta|^2}{2\lambda\gamma}\right), & \lceil\beta\rceil < \lambda\gamma \\ \frac{\lambda^2\gamma}{2}, & \lceil\beta\rceil \geq \lambda\gamma \end{cases}$$

Where: $\gamma > 1$

Numerous penalty functions are used in statistical regression models, and these functions are usually concave. Typically, they rely on a tuning parameter (λ) and furthermore incorporate another tuning parameter (γ) that regulates the penalty's level of concavity. The penalty's rate of decline depends critically on this parameter (γ).

The adaptive LASSO (least absolute shrinkage and selection operator) and MCP approaches enable the estimated coefficients to develop faster than with the classic LASSO approach, especially for nonzero coefficients. While the goal of most of these methods is to shrink coefficients towards zero, the adaptive LASSO and MCP methods reduce the amount of shrinkage that is applied to nonzero coefficients, suggesting reduced bias in the estimate

process. Numerous penalty functions are used in statistical regression models, and these functions are usually concave.

Typically, they rely on a tuning parameter (λ) and furthermore incorporate another tuning parameter (γ) that regulates the penalty's level of concavity. The penalty's rate of decrease is largely dependent on this parameter (γ).

The MCP (minimally concave penalty) function has an intriguing property in that it covers a range of values where all estimations stay constant. Interestingly, the estimates are exactly the same inside this range as when derived with the least squares regression technique.

2.3. SCAD

The best way to comprehend the SCAD penalty, as presented by Fan and Li (2001), is to look at its first derivative.

$$p'_\lambda(\beta) = \lambda\left\{I\{\beta \le \lambda\} + \frac{(a\lambda-\beta)+}{(a-1)\lambda} I\{\beta > \lambda\}\right\} \text{ for some } a \text{ and } \beta > 0,$$

The letter I stands for the indicator function in the SCAD technique, β for a vector of unknown parameters, and λ for the regularization parameter.

The approach described in this response involves setting all less significant variables to zero in order to produce simpler, easier-to-manage models. Fan and Li (2001) demonstrated that the SCAD penalty may produce estimates with the required oracle quality. This suggests that non-zero coefficient estimate is as accurate as it would be if the correct model had been known beforehand. Moreover, a true parameter is roughly equal to zero, and its probability is very guaranteed when it is zero.

Numerous search criteria, such as the Bayesian Information Criterion (BIC), generalized cross-validation, and cross-validation, can be used to locate the two tuning parameters (λ,a).

According to Fan and Li (2001), a = 3.7 should be selected as a suitable value for one of the tuning parameters.

According to Fan and Li (2001), a good penalty function should yield an estimator with three crucial characteristics. The first feature is unbiasedness, which ensures that there is no unnecessary modeling bias by avoiding an excessive penalty of big parameters in the final estimator. Additionally, the estimator should indicate sparsity by automatically setting unimportant parameters to 0. Last but not least, continuity is essential because if the final estimate can demonstrate continuity in the data, model prediction instability may be prevented.

2.4. LASSO

The Lasso penalty was introduced by Tibshirani (1996) as a regularization method for simultaneous variable estimation and selection in large datasets.The definition of the Lasso estimate $\hat{\beta}$ is:

$$\hat{\beta}_{lasso} = \min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \sum_j \beta_j x_{ij}\right)^2 + \lambda\sum_j|\beta_j|\right\}, \qquad \text{or}$$

$$\hat{\beta}_{lasso} = \min_{\beta} \|y - x\beta\|_2^2 + \lambda\|\beta\|_1$$

The trade-off between minimizing the residual sum of squares (RSS) and the penalty term (the sum of the absolute values of the coefficients) is determined by a parameter known as lambda.

The Lasso technique minimizes the sum of squared residuals while ensuring that the absolute values of all the coefficients are smaller than a constant. Regression models with a lot of variables and limited data frequently employ it. Lasso's main goal is to select the variables while fitting the regression line to the data. This is accomplished by setting some coefficients to zero and decreasing the value of others. Lasso adds a penalty to the optimization objective in order to conduct L1 regularization.

2.5. Quantile Regression

Ordinary least squares (OLS) regression is used to estimate the average response based on predictor variables. However, another method known as least absolute deviation (LAD) regression is used to estimate the median function. Dealing with response outliers and heavy-tailed errors is made easier by LAD regression's increased robustness.

As an extension of LAD regression, Koenker and Bassett (1978) developed quantile regression (QR). QR offers a comprehensive comprehension of the conditional distribution of the response variable by calculating the response's conditional quantile function. When everything is taken into account, QR maintains the advantages of LAD regression while producing a model that is generally more informative.

Quantile regression offers a flexible and all-encompassing way to describe the connection between predictors and response variables by adjusting the quantile parameter θ. Remarkably, when θ = 0.5, quantile regression is equivalent to least absolute deviation regression (also known as median regression), which is renowned for its resilience to outliers. This approach is selected in these situations because it is known to be robust against outliers and can estimate the conditional quantiles of a response variable.

LAD and median regression are almost the same since they both seek to minimize the absolute differences between the expected and actual values in the data. For this reason, median regression is another name for least absolute deviation (LAD) regression. LAD regression minimizes the sum of the absolute values of the residuals (differences between predicted and actual values), whereas median regression searches for the line (or hyperplane in higher dimensions) that minimizes the absolute departures of the data points from a central point, the median.

The strong value of quantile regression is a substantial benefit when the assumptions of least squares regression are not satisfied or when a deeper understanding of the connection between variables over several areas of the conditional distribution is needed. Its computational components and interpretation, however, require significant thought. In reality, the coefficients

may be accurately determined by solving a minimization problem, which produces reliable parameter estimates for a range of quantiles of interest.

$\min_{\beta} \sum_{i=1}^{n} \rho_{\theta}(y_i - x_i^T \beta)$

Where $\rho(.)$ An outlier-resistant loss function that is utilized in a variety of optimization tasks is referred to as the goal function. Because of its exceptional resilience in the face of outliers, it is an invaluable tool for machine learning and regression applications.

$$\rho_{\theta}(t) = \begin{cases} \theta t & if \quad t \geq 0 \\ -(1-\theta)t & if \quad t < 0 \end{cases}, \text{ where } 0 < \theta < 1$$

Koenker (2004) pioneered the use of regularisation in quantile regression.. The LASSO penalty was first developed in this seminal study to control random effects inside a mixed-effect quantile regression framework. The goal was to reduce the random effects to zero by utilizing the regularization characteristics of the LASSO approach. This new method represented a significant development in the field by providing a novel way to control model complexity and enhance estimation accuracy in mixed-effect quantile regression models.

3. Data and Methodology

For a study on the relationship between diabetes levels and four parameters (gender, age, weight, and height), data were gathered from Azadi Hospital in Dohuk. Examining the effects of these factors on diabetes levels was the aim of the study. There are 5 columns and 109 rows in the dataset.

• Gender(X_1)

• Age (X_2)

• Weight(X_3)

• Height(X_4)

• Diabetes levels (y)

The following techniques were used to assess the five variables in the data set: Least Absolute Shrinkage and Selection Operator (LASSO), Minimax Concave Penalty (MCP) Regression, Quantile Regression (QR), Smoothly Clipped Absolute Deviation (SCAD) penalty, and Ordinary Least Square (OLS).

4. Interpretation and Discussion of Results

The primary focus of this work is on data analysis methods and associated technology. Three R packages, MASS, ncvreg, and quantreg, were utilized to do the necessary computations and estimate parameters. The results of our analysis are displayed in the table that follows. Our main objective is to establish a connection between various clinical markers

and diabetes levels.Table:

OLS, MCP, SCAD, LASSO and QR Results

	Factors			
OLS Results	X_1	X_2	X_3	X_4
Estimate	-10.69	2.10	1.30	-0.14
z-value	-0.85	4.23	2.19	-0.27
p-value	0.40	0.0	0.03	0.79
MCP Results				
Estimate	-10.62	2.09	1.29	-0.13
z value	-0.85	4.11	2.22	-0.28
p-value	1.00	0.00	0.28	1.00
SCAD Results				
Estimate	-3.41	2.13	1.23	0.00
z- value	-0.76	4.18	2.12	0.00
p-value	1.00	0.00	0.33	-
LASSO Results				
Estimate	-4.80	1.93	1.02	0.00
z- value	-0.82	4.27	2.21	0.00
p-value	1.00	0.00	0.28	-
QR Results				
Estimate	-20.21	1.96	1.60	0.08
z- value	-1.51	5.76	3.98	0.08
p-value	0.13	0.00	0.00	0.94

5. Conclusion and Recommendation

After analyzing the data using a range of regression techniques, such as Ordinary Least Square (OLS), Minimax Concave Penalty (MCP), Smoothly Clipped Absolute Deviation (SCAD), Least Absolute Shrinkage and Selection Operator (LASSO), and Quantile Regression (QR), we have found a significant relationship between age and weight with diabetes levels (y). Because fat can lead to diabetes, it is thus recommended that all governmental levels educate the public on the need of reducing fat intake.

Acknowledgements

The authors would like to express their gratitude to Azadi Hospital in Dohuk for providing the data that helped to enhance the quality of this work.

REFERENCES

- Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J. and Sakr, S.(2017) Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project. 12, e0179805.
- Ajeel, S. M., Haji,J.A. , and Jahwar, B.H.(2023). Using Multinomial Logistic Regression to Identify Factors Affecting Platelet, Journal of University of Duhok., Vol. 62, No.2.
- Anderson, A.E., Kerr, W.T., Thames, A., Li, T., Xiao, J. and Cohen, M.S.(2016). Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study. J. Biomed. Inform. 60, 162–168.
- Bennett, P., Burch, T. and Miller, M.(1971). Diabetes mellitus in American (Pima) indians. Lancet . 298, 125–128.
- Cho, N., Shaw, J., Karuranga, S., Huang, Y., da Rocha Fernandes, J., Ohlrogge, A. and Malanda, B. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res. Clin. Pract. 138, 271–281.
- Collins, G.S., Mallett, S., Omar, O.and Yu, L.M.(2011). Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting. BMC Med. 9, 1–14.
- Drapper, N. R. and Smith, H.(1981). Applied Regression Analysis , Second Edition, John Wiley and Sons, New York.
- Engelgau,M.M., Narayan, K. and Herman,W.H.(2000). Screening for type 2 diabetes. Diabetes Care 23, 1563–1580.
- Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties,Journal of the American Statistical Association 96(456), 1348–13.
- Gregg, E.W., Geiss, L.S., Saaddine, J., Fagot-Campagna, A., Beckles, G., Parker, C., Visscher, W., Hartwell, T., Liburd, L. and Narayan, K.V.; et al.(2001). Use of diabetes preventive care and complications risk in two African-American communities. Am. J. Prev. Med. 21, 197–202.
- Habibi, S., Ahmadi, M. and Alizadeh, S. (2015).Type 2 diabetes mellitus screening and risk factors using decision tree: Results of data mining. Glob. J. Health Sci. 7, 304.
- Heikes, K.E., Eddy, D.M., Arondekar, B. and Schlessinger, L.(2008). Diabetes Risk Calculator: A simple tool for detecting undiagnosed diabetes and pre-diabetes. Diabetes Care . 31, 1040–1045.
- Krasteva, A., Panov, V., Krasteva, A., Kisselova, A. and Krastev, Z. (2011). Oral cavity and systemic diseases—Diabetes mellitus. Biotechnol. Biotechnol. Equip. 25, 2183–2186

- Ryden, L., Standl, E., Bartnik, M., Van den Berghe, G., Betteridge, J., De Boer, M.J., Cosentino, F., Jönsson, B., Laakso, M. and Malmberg, K.; et al.(2007) Guidelines on diabetes, pre-diabetes, and cardiovascular diseases: Executive summary: The Task Force on Diabetes and Cardiovascular Diseases of the European Society of Cardiology (ESC) and of the European Association for the Study of Diabetes (EASD). Eur. Heart J. 28, 88–136.

- Tuso, P. (2014). Prediabetes and lifestyle modification: Time to prevent a preventable disease. Perm. J. 18, 88.
- Knowler,W.C., Barrett-Connor, E., Fowler, S.E., Hamman, R.F., Lachin, J.M.,Walker, E.A.and Nathan, D.M.(2002). Diabetes Prevention Program Research Group . Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N. Engl. J. Med., 346, 393–403.
- Koenker, R. and G. W. Bassett (1978). Regression quantiles. Economerrica 46, 33–50.
- Koenker, R. (2004). Quantile regression for longitudinal data. Journal of Multivariate Analysis, 91(1), 74–89.
- Mikolajczyk, R.T., DiSilvesto, A. and Zhang, J.(2008). Evaluation of logistic regression reporting in current obstetrics and gynecolog literature. Obstet. Gynecol. 111, 413–419.
- Naz, H. and Ahuja, S.(2020) Deep learning approach for diabetes prediction using PIMA Indian dataset. J. Diabetes Metab. Disord. 19, 391–403.
- Nguyen, B.P., Pham, H.N., Tran, H., Nghiem, N. , Nguyen, Q.H., Do, T.T., Tran, C.T. and Simpson, C.R.(2019). Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Comput. Methods Programs Biomed. 182, 105055.
- Razavian, N., Blecker, S.,Schmidt, A.M., Smith-McLallen, A., Nigam, S. and Sontag, D.(2015) Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data . 3, 277–287.
- Shareef, A. A., Ajeel, S. M. and Hashem, H.A. (2024). Utilizing Multinomial Logistic Regression For Determining The Factors Influencing Blood Pressure. Science Journal of University of Zakho , Vol. 12, No.3, 367–374 .
- Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society, Series B 58, 267–288.
- Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics 38, 894–942.
- Zou, Q., Qu, K., Luo, Y., Yin, D. and Ju, Y.(2018) Tang, H. Predicting diabetes mellitus with machine learning techniques. Front. Genet. 9, 515.