Tikrit Journal of Administrative and Economic Sciences, Vol. 19, No. 62, Part (2): 401-411
Doi: www.doi.org/10.25130/tjaes.19.62.2.22

Tikrit Journal of Administrative and Economics Sciences
مجلة تكريت للعلوم الإدارية والاقتصادية
ISSN: 1813-1719 (Print)

Estimating the impact of Pregnancy, Systolic and Age on Diabetes for Women by Using Support Vector Regression Model (SVR)

Rawa Saman Maaroof*[A], Shamazad Rahim[B], Shahla Othman Salih[C], Hindreen Abdullah Taher[D]

[A] Department of Tourism, College of Commerce, University of Sulaimani
[B] Department of Finance and Banking College of Commerce, University of Sulaimani
[C] Department of Statistics and Informatics, College of Administration Economic University of Sulaimani
[C] Department of Information Technology, College of Commerce, University of Sulaimani

Keywords:
Regression Model, Support vector regression, kernel functions.

ARTICLE INFO

Article history:
Received 18 May. 2023
Accepted 19 Jun. 2023
Available online 30 Jun. 2023

*Corresponding author:
Rawa Saman Maaroof
Department of Tourism, College of Commerce, University of Sulaimani

Abstract: In this paper we have 623 cases of diabetes patients, the data partitioned in to training dataset (500 observations) and testing dataset (123 observations), and the aim is to estimate the impact of pregnancy duration in weeks, systolic and age as factors on diabetes of the women patients for this purpose SVR has been used. According to the results radial kernel function gave highest performance compared to the other kernel functions, the R2 = 83% this implies the factors capable of explaining 83% of diabetes variable with MSE and RMSE of (0.000958 and 0.030956) respectively. And p-values of the three aforementioned variables are less than the significant level of 0.01, implying that the three factors have a statistically significant impact on the response variable. Where Pregnancy duration in weeks has an impact of 0.401 on the patient, that means if duration increase by one week, then diabetes will increase by 0.401 units, also both Systolic and age have a significant positive effect on the response variable, and the amount of impact is (0.621 and 0.557) respectively.

Tikrit Journal of Administrative and Economic Sciences, Vol. 19, No. 62, Part (2): 401-411
Doi: www.doi.org/10.25130/tjaes.19.62.2.22

تقدير تأثير الحمل والضغط الدم والعمر على داء السكرى عند النساء باستخدام نموذج انحدار ناقل الدعم (SVR)

رەوا سامان معروف
قسم السياحة، كلية التجارة،
جامعة السليمانية

شەم ئازاد رحيم
قسم المالية والمصرفية كلية التجارة ة
جامعة السليماني

شهلا عثمان صالح
قسم الإحصاء والمعلوماتية، كلية الإدارة الاقتصادية
جامعة السليمانية

هندرين عبدالله طاهر
قسم تقنية المعلومات، كلية التجارة
جامعة السليمانية

المستخلص

في هذا البحث، لدينا 623 حالة لمرضى السكري، تم تقسيم البيانات إلى مجموعة بيانات تدريبية (500 مشاهدة) ومجموعة بيانات اختبار (123 مشاهدة)، والهدف هو تقدير تأثير مدة الحمل في الأسابيع، وضغط الدم والعمر كعوامل على مرض السكري للنساء، لهذا الغرض تم استخدام SVR. وفقًا للنتائج، أعطت دالة RB أعلى أداء مقارنة بدوال كرنلات الأخرى، وكانت R^2 = 83٪ وهذا يدل على ان العوامل القادرة على تفسير 83٪ من متغير السكري مع MSE وRMSE بقيمة (0.000958 و0.030956) على التوالي. وقيم p-values للمتغيرات الثلاثة المذكورة أعلاه أقل من المستوى المعنوي 0.01، مما يعني أن العوامل الثلاثة لها تأثير معنوي ذو الدلالة الإحصائية على متغير الاستجابة. حيث أن مدة الحمل في الأسابيع لها تأثير بمقدار 0.401 على المريضة، أي إذا زادت المدة بمقدار أسبوع واحد، فإن مرض السكري سيزداد بمقدار 0.401 وحدة، كما أن كلا من ضغط الدم والعمر لهما تأثير معنوي على متغير الاستجابة، ومقدار التأثير (0.621 و0.557) على التوالي.

الكلمات المفتاحية: نموذج الانحدار، دعم متجه الانحدار المتجه، ودوال كرنال.

1-1. Introduction

It is estimated that at least 5% of pregnant women will develop diabetes. Pregnancy-related diabetes is the name for this condition. There are several hormones at work in the body that maintain healthy levels of glucose in the blood. Hormonal fluctuations, however, make it harder for the body to maintain steady blood sugar levels throughout pregnancy. Blood glucose (sugar) levels can be managed with diet and medical treatment. The placenta secretes a hormone that reduces the body's sensitivity to insulin, so pregnant women typically require a higher insulin dose. Blood sugar levels rise as a result of this. A high sugar level in the blood may cause problems for the mother and child, such as increasing the possibility of needing a cesarean section, Premature birth, severe breathing difficulties, obesity, type 2 diabetes later in life, and stillbirth. Risks can be reduced if detected and

managed early, so in this paper, we will try to estimate the impact of pregnancy, systolic, and age on diabetes for women, and support vector regression has already demonstrated its ability to address such issues. The support vector regression (SVR) model is a machine learning algorithm used for regression analysis, which includes a prediction of continuous numerical values. It also falls under the Support Vector Machine (SVM) algorithm, a well-known algorithm for classification and prediction tasks. The support vector regression algorithm aims to best fit a regression line that separates the data into two classes: the target variable and the residual errors. This regression line represents the upper level that increases the margin between the target variable and residual errors. Margin is defined as the distance between the hyper plane and the nearest data points, known as support vectors. It deals with nonlinear relationships between independent and dependent variables and can also manage outliers effectively. This makes support vector regression a valuable algorithm for various predictions. In summary, the support vector regression algorithm is a powerful and flexible machine learning algorithm, mainly when the data contains a nonlinear relationship or outliers.

1-2. Literature Review: Lama (2017), to classify thermography images into normal or abnormal categories for the detection of canine bone cancer disease, canine anterior cruciate ligament rupture, and feline hyperthyroid disease, employed SVM models as binary classifiers using gray level co-occurrence matrix texture features extracted from the thermographs. Also based on parallel factor analysis coupled with support vector regression (SVR), Gu and Sun (2019, 218, 27-32) designed a probe-based fluorescence spectroscopy for the quick detection of lysed and oxidized chemicals (i.e., acids, aldehydes, alcohols, ketones, hydrocarbons, etc.) in frying palm oil. Characteristic fluorescence peaks were identified using loading scores at relevant components with the help of the parallel factor analysis technique. Then, a variety of preprocessing algorithms were combined with the SVR algorithm. Grid search performed better than the other three methods in a regression test using four distinct SVM models. The final SVR models' performance was evaluated using the following metrics: R^2 = 0.9753, P = 0.9724, MSE = 0.0089, and P = 0.0088 for the calibration and prediction sets, respectively. And Gu et al. (2020,13(11), p2080-2086) invented probe-based three-dimensional fluorescence spectroscopy using parallel factor

analysis and support vector regression (SVR) to identify, discriminate, and quantify dissolved organic materials in frying oil. Compared to time-consuming and expensive chemical procedures, the proposed methodology improved the rapid assessment of frying oil quality and other high-oil food and beverages. Considering time and model robustness, parallel factor analysis combined with analysis of characteristic peaks data may be better for model creation· Mohamed H. Ibrahim and Asmaa G. Jaber (2022, 28,p 132) They believed that the financial markets are constantly changing, making it difficult to predict their trends. To classify the stock data using five variables, the support vector machine and CART regression tree algorithms were used. The results showed that the SVM algorithm was the best when compared with the CART algorithm, using the Classification Error and MSE criteria. It was found that the SVM algorithm obtained the best results when compared with the CART algorithm, using all criteria. Waleed Dhhan and Thaera Alameer (2018) They proposes a robust variable selection technique for the SI-SVR model, RESI-SVR, to simultaneously outlier detection and dimension reduction. The key to success is the use of the FP-SVR to detect and minimize outliers and leverage points. Comparative results show superiority over existing ENSI-SVR.

2. Methodology

2-1. Support Vector Regression (SVR): Linear regression is the most statistical model used in practical applications because these types of models are linearly dependent on their unknown parameters. This can be fitted much more easily than the other models which response have a non-linear relationship with their unknown parameters and because the properties of statistical estimators are easier to explain. But the assumption of OLS method cannot be achieved easily (Ahmed, Taher, 2018,50) & (Taher & Ahmed, 2023, 2087). Support vector machines (SVMs) are well-suited to generalizing on unseen data due to their statistical learning or Vapnik-Chervonenkis (VC) foundations (Vapnik,2000, 314). Kernels, sparse solutions, VC margin, and SVR control are similar to categorization. SVR estimates real-valued functions better than SVM, despite its lesser fame. SVR's loss function punishes over and under-estimates equally during training. SVR is supervised learning. In his e-insensitive technique, Vapnik builds a flexible tube with a minimal radius symmetrically around the estimated function to reject absolute values of errors below a predefined

Doi: www.doi.org/10.25130/tjaes.19.62.2.22

threshold e in both the upper and lower regions of the estimate. This approach affects the region above and below the function but not the tube (Vapnik, 2000: 314) & (Astuti W., Adiwijaya, 2018: 971) & (Naik, G. R,2018). (SVR's computational complexity is independent of input space size, which is a significant benefit. It can predict and generalize well. Thus, this chapter will cover SVR and Bayesian regression. An adjusted SVR can be used to avoid underestimating a function. Figure (1) depicts a one-dimensional problem that can be viewed geometrically to help establish the best formulation for an SVR problem. Equation(1(provides a convenient form for approximating continuous-valued functions. Simplifying the mathematical terminology, we may construct the multivariate regression from Equation (2) by increasing x by one and adding b to the w vector.

$$y = f(x) = <w, x> + b = \Sigma_{j-1}^{m} w_j \; x_j + \text{b, y,b } \epsilon \text{ R,x, w } \epsilon \; R^m \qquad (1)$$

$$f(x) = \begin{bmatrix} w \\ b \end{bmatrix}^r \begin{bmatrix} x \\ 1 \end{bmatrix} = w^T \; x + b \; x, w \; \in R^{m+1} \qquad (2)$$

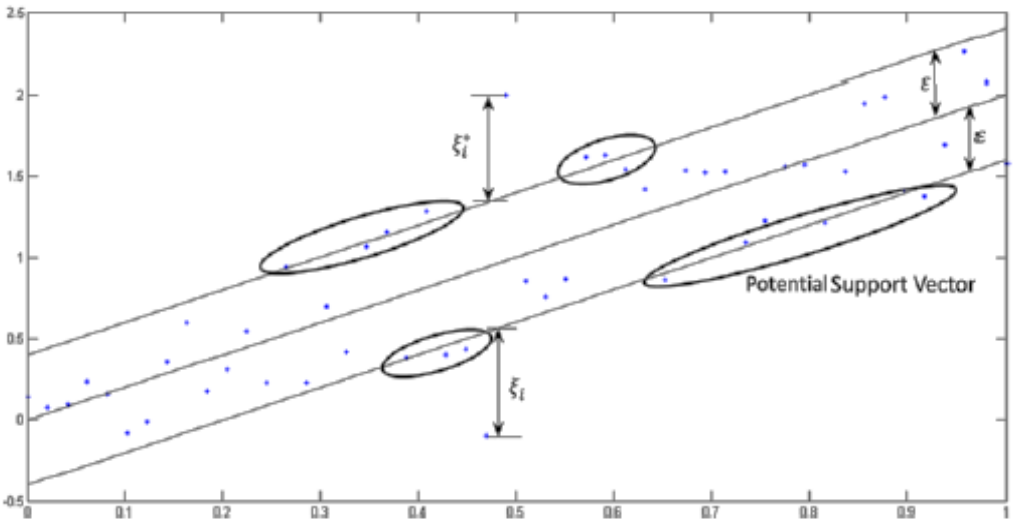

Figure (1) shows one-dimension SVR

By framing the work as an optimization issue (Ibrahim, M. H., & Jaber, A. G. (2022, 132) support vector regression (SVR). Seeks to find an approximation for a function that minimizes the prediction error, or the difference between the expected and the intended outputs, where ||w|| is the magnitude of the normal vector to the surface being approximated, the objective function that is (estimating the impact of the aforementioned

Doi: www.doi.org/10.25130/tjaes.19.62.2.22

explanatory variables on the response variable which is diabetes with minimum residuals) is given by Equation (3):

$$min_w \frac{1}{2}|w|^2$$

Here's an illustration of how the sum of the weights might be used as a proxy for levelness:

$$f(x,w) = \sum_{t-1}^{M} w_t\, x^t\,, x \in R, w \in R^M \qquad (3)$$

There is a clear indication of the approximate polynomial's order, M. As the size of the vector w grows. The horizontal line stands for a significantly off-ideal 0^{th} order polynomial solution. While the 1^{st}-order polynomial linear function better approximates portions of the data, it still doesn't produce a satisfactory match to the training data as a whole. The 6^{th}-order solution provides an acceptable compromise between function flatness and prediction error. The immense complexity of the highest-order solution means it will likely over fit the answer on unseen data even though it has zero error. The size of the regularizing term w determines the extent to which the flatness of the solution can be manipulated in an optimization problem. The constraint is to restrict the value of the function to be as close as possible to the expected value for a particular input (Chowdhury, 2017,15(8), p28-32) & (Arık OA (2020).13. p415-425) The SVR algorithm penalizes predictions that are more than e distant from the target value by using a loss function that does not consider e. In addition to affecting the number of support vectors and, by extension, the sparsity of the solution, the value of e determines the width of the tube. A smaller value suggests a lower tolerance for error. Figure (1) provides a visual representation of this latter concept (Mechelli, A., Vieira, S. (Eds.), 2019) If e is lowered, the tube's boundary creeps inward. Since there are more data points near the boundary, more support vectors exist. Increased e also reduces the number of locations near to international borders. The model's resilience is enhanced by the e-insensitive zone, which makes it less vulnerable to perturbations in the data. Equations (4, 5, and 6) illustrate the linear, quadratic, and Huber e loss functions, respectively, and can be applied. The Huber loss function, as seen in Figure 4-3, is more lenient

Doi: www.doi.org/10.25130/tjaes.19.62.2.22

on minor deviations from the desired output than linear and quadratic loss functions, but it still penalizes any and all outliers. Which loss function to employ depends on the available computational resources for training, the desired degree of model sparsity, and a priori knowledge of the noise distribution affecting the data samples.

Symmetric and convex loss functions are provided. To ensure that the optimization problem has a unique solution that can be found in a finite number of iterations, the loss function used to correct for under- or over-estimation must be convex. To begin deriving the topics of this chapter, we will use Equation (4)'s linear loss function.

$$L_{\delta}\big(y, f(x,w)\big) = \begin{cases} 0 & |y - f(x,w)| \le \delta; \\ |y - f(x,w)| - \delta otherwise, \end{cases} \quad (4)$$

$$L_{\delta}\big(y, f(x,w)\big) = \begin{cases} 0 & |y - f(x,w)| \le \delta; \\ (|y - f(x,w)| - \delta)^2 otherwise, \end{cases} \quad (5)$$

$$L\big(y, f(x,w)\big) = \begin{cases} c|y - f(x,w)| - \frac{c^2}{2} |y - f(x,w)| > c \\ \frac{1}{2}|y - f(x,w)|^2 |y - f(x,w)| \le c \end{cases} \quad (6)$$

2-2. Kernel SVR and Different Loss Functions: Before, we assumed that *f(x)*was linear and focused on data in the feature space. When dealing with nonlinear functions, it is possible to improve classification accuracy by mapping the data into a higher-dimensional space (called kernel space) using kernels that satisfy Mercer's condition[8,10] (Blanco, V., Puerto, J., Rodriguez-Chia,2020,21(14)) & (Chowdhury, 2017, 15(8), 28-32). Substituting $k(x_i, x_j)$ for x in Equations (1and 8) results in the fundamental formulation illustrated in Equation (9). Where $\Phi(x_i)$ denotes the transformation from feature to kernel space. The reformulated weight vector, in terms of the original input, is defined by Equation (10). Equation (11) represents the dual problem.

$$min \frac{1}{2}\|w\|^2 + C\sum_{t-1}^{N} \xi_i + \xi_i^n \quad (7)$$

Doi: www.doi.org/10.25130/tjaes.19.62.2.22

Subject to

$$y_t - w^T\Phi(x_i) \le \xi_i + \xi_i^n i = 1, \dots\dots, N$$
$$w^T\Phi(x_i) - y_i \le \xi + \xi_i i = 1, \dots\dots, N$$
$$\xi_i, \xi_i^n \ge 0 \qquad i = 1 \dots\dots N$$

$$w = \sum_{i-1}^{Nsv}(\alpha_i^n - \alpha_i)\phi(x_i\ ) \tag{8}$$

$$max_{n,m} - \varepsilon\sum_{i=1}^{Nsv}(\alpha_i + \alpha_i^n) + \sum_{i=1}^{Nsv}(\alpha_i^n - \alpha_i)y_i - \frac{1}{2}\sum_{i-1}^{Nsv}\sum_{j-1}^{Nsv}(\alpha_i^n - \alpha_i)\left(\alpha_j^n - \alpha_j\right)k(x_i, x_{j)} \tag{9}$$

$$\alpha_i, \alpha_i^n \in [0, C], i = 1, \dots\dots N_{sv}, \sum_{i-1}^{Nsv}(\alpha_i^n - \alpha_i) = 0$$

$$f(x) = \sum_{i=1}^{Nsv}(\alpha_i^n - \alpha_i)k(x_i, x) \tag{10}$$

$$k(x_i, x) = \emptyset(x_i).\emptyset(x) \tag{11}$$

3. Applications

3-1. Data Description: The data of our study has been collated at Hawler Diabetes centre, the cases are about 623 cases, and all cases are females. Diabetes has been chosen as a response variable, and the others as explanatory variables (pregnancy duration in weeks, systolic and age).

Table (1): Represents the description of the factors.

	Diabetes	**Pregnancy duration in weeks**	**Systolic**	**Age**
Athematic Mean	121.89	4.50	72.48	34.38
Standard deviation	30.88	3.21	11.95	11.88
Coefficient of variation	25%	71%	16%	35%

The above table demonstrates the arithmetic mean and standard deviation of the whole variables in the study. The coefficient of variation tells that the most variation exists in the pregnancy duration in weeks, reaching 71%, and the minimum variation is recorded in the systolic variable Azad Rahim and et al. 2023, 10(3S), 4099-4104).

Doi: www.doi.org/10.25130/tjaes.19.62.2.22

Table (2): Shows the performance of SVR for each kernel functions.

Kernel	number of SVR	R^2	MSE	RMSE
Linear	421	74%	0.001162	0.034088
Polynomial	422	65%	0.002023	0.044978
Radial	384	83%	0.000958	0.030956
Sigmoid	499	27%	1.575737	1.255284

Summing up to the table-2, which represents the application of the epsilon support vector regression model with selecting the best kernel function, it is clear that the radial kernel function has the highest performance among the other kernel functions. Furthermore, the R^2 of the best kernel is equal to 83% with minimum MSE and RMSE (0.000958 and 0.030956), respectively (Azad Rahim and et al. 2023, 10(3S), 4099-4104).

Table (3): Displays the test of the estimators and their impacts.

Explanatory variables	Estimated	S.E	P -Value
Pregnancy duration in weeks	0.401	0.041	0.000
Systolic	0.621	0.036	0.000
Age	0.557	0.044	0.000

The table-3 clarifies the estimated values of the parameters for the features and their test to check whether they have an impact on diabetes in pregnant women or not. The three values of the p-value column are less than the significant level of 0.01, implying that the three factors have a statistically significant impact on the response variable. Where Pregnancy duration in weeks has an impact of 0.401 on the patient, that means if duration increase by one week, then diabetes will increase by 0.401 units. Also both Systolic and age have a significant positive effect on the response variable, and the amount of impact is (0.621 and 0.557) respectively Azad Rahim and et al. 2023, 10(3S), 4099-4104).

Doi: www.doi.org/10.25130/tjaes.19.62.2.22

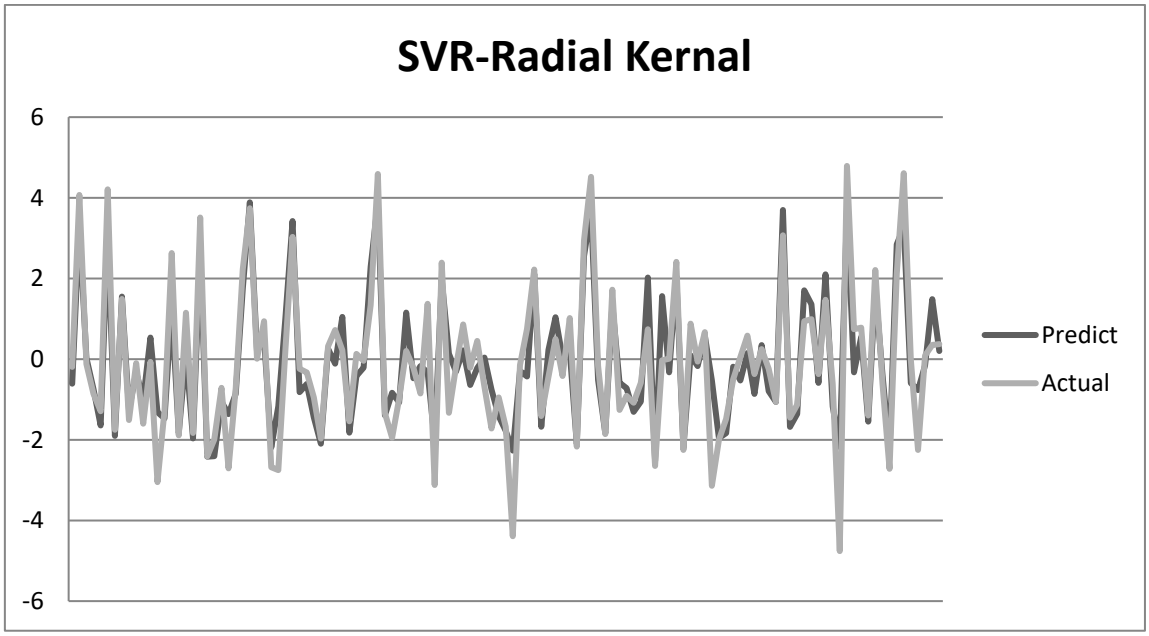

Figure (2): The above figure shows the line graph between the actual and predicted values of the best kernel function for the test dataset (Aziz, A. and et al., 2023, 10(3S), 4931-4937).

4. Conclusions: This paper estimated the impact of pregnancy duration in weeks, systolic and age on diabetes of 623 diabetes patients using radial kernel functions. Results showed that the three factors had a statistically significant impact on the response variable. And systolic has highest impact.

References

1. Lama, N., (2017). Optimized Veterinary Thermographic Image Classification using Support Vector Machines and Noise Mitigation (Doctoral dissertation, Southern Illinois University at Edwardsville).
2. Gu, H., & Sun, Y., (2019). Enhancing the fluorescence spectrum of frying oil using a nanoscale probe. Spectrochemical Acta Part A: Molecular and Biomolecular Spectroscopy, 218, 27-32.
3. Gu, H., Huang, X., Chen, Q., & Sun, Y., (2020). Rapid Assessment of Total Polar Material in Used Frying Oils Using Manganese Tetraphenyl porphyrin Fluorescent Sensor with Enhanced Sensitivity. Food Analytical Methods, 13(11), 2080-2086.
4. Ahmed, N. M., & Taher, H. A., (2018). Multi-response Regression Modeling for an Agricultural Experiment. Journal of University of Human Development, 4(2), 46-52.
5. Taher, H. A., & Ahmed, N. M., (2023). Using Bayesian Regression Neural Networks Model to Predict Thrombosis for Covid-19 Patients. resmilitaris, 13(1), 2077-2087.
6. Vapnik, V., (2000). The nature of statistical learning theory. Springer, 314. doi: https://doi.org/10.1007/978-1-4757-3264-1
7. Mechelli, A., Vieira, S. (Eds.), (2019). Machine learning: methods and applications to brain disorders. Academic Press. doi: https://doi.org/10.1016/C2017-0-03724-2

8. Blanco, V., Puerto, J., Rodriguez-Chia, A. M. (2020). On lp-Support Vector Machines and Multidimensional Kernels. Journal of Machine Learning Research, 21 (14). Available at: https://jmlr.org/papers/volume21/18-601/18-601.pdf
9. Astuti, W., Adiwijaya (2018). Support vector machine and principal component analysis for microarray data classification. Journal of Physics: Conference Series, 971, 012003. doi: https://doi.org/10.1088/1742-6596/971/1/012003
10. Chowdhury, U. N., Rayhan, M. A., Chakravarty, S. K., Hossain, M. T. (2017). Integration of principal component analysis and support vector regression for financial time series forecasting. International Journal of Computer Science and Information Security (IJCSIS), 15 (8), 28–32.
11. [17] Naik, G. R. (Ed.) (2018). Advances in Principal Component Analysis. Springer. doi: https://doi.org/10.1007/978-981-10-6704-4
12. Arık OA (2020) Comparisons of metaheuristic algorithms for unrelated parallel machine weighted earliness/tardiness scheduling problems. Evol Intel 13:415–425.
13. Ibrahim, M. H., & Jaber, A. G. (2022). The Use of the Regression Tree and the Support Vector Machine in the Classification of the Iraqi Stock Exchange for the Period 2019-2020. journal of Economics and Administrative Sciences, 28(132).
14. Dhhan, W., & Alameer, T. (2018). Robust Variable Selection Technique for Single Index Support Vector Regression Model.
15. Aziz, A. A., Mahmood, H. O. F., Rahim, S. A., Maaroof, R. S., & Taher, H. A. (2023). Using Optimizing Parameters Support Vector Regression Model to Predict Potassium Ratio in Carb Fish. Journal of Survey in Fisheries Sciences, 10(3S), 4931-4937.
16. Azad Rahim, S., & Taher, H. A. (2023). Postulating Support Vector Regression Model to Measure the Effect of Protein, Carbohydrate and Fats on the Weight of Carb Fish. Journal of Survey in Fisheries Sciences, 10(3S), 4099-4104.