Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777993

ANALYSIS OF IMPACT OF DIFFERENT SCALERS ON PERFORMANCE OF MACHINE LEARNING ALGORITHMS FOR PREDICTION OF TYPE 2 DIABETES

Kamal Shah[1], Suresh Rajpurohit[1], Rohini Patil[2], Anil Vasoya[1], Praniti Patil[3] and Sangeeta Mishra[4]*

[1]IT Department, Thakur College of Engineering and Technology, Mumbai – 400101, Maharashtra, India

[2]Computer Engineering Department, Terna College of Engineering, Navi Mumbai – 400706, Maharashtra, India

[3]AIML Department, Thakur College of Engineering and Technology, Mumbai – 400101, Maharashtra, India

*[4]EXTC Department, Thakur College of Engineering and Technology, Mumbai – 400101, Maharashtra, India

Abstract

Type 2 diabetes occurred due to unbalance in glucose consumption in body which eventually lead to disorders of the circulatory, nervous and immune systems. Many studies are done on prediction of this disease involving various clinical and pathological parameters and with advancement of technology many Machine Learning techniques are also incorporate for better predication accuracy. In this paper the idea of data preprocessing is explored and its effect on ML algorithms is Analyzed. For experimental set up two datasets PIMA which is from Kaggle and locally generated and validated dataset LS. Total 5 ML algorithms and 8 different scaling techniques are evaluated in the study. It is observed that without pre-processing of data with any of the scalar the accuracy of PIMA data set is from 46.99 to 69.88%, which improves with scalers up to 77.92 %. For LS dataset without scalers accuracy is as low as 78.67% which improves to 100% with two labels as the LS data set is small and controlled. Various Scalers have different impact on data pre-processing stage for PIMA and LS both datasets. With scalers introduced in pre-processing stage there is visible improvements in accuracy so depending on the data set selection of scalers surely going to improve the efficiency.

Keywords: Type 2 Diabetes ML algorithms, scaler for preprocessing, PIM dataset, LS dataset

INTRODUCTION

Diabetic mellitus (DM) is one of the most common non-communicable diseases globally. It is observed that, 46% of people with diabetes are not diagnosed at early stage. By the year of 2040, it is expected that the count may rise to 642 million all over the globe [1]. India contributes about 49% of world's burden. In Southeast Asia region, out of 88 million people with diabetes, India contributes 77 million people which is expected to increase to 134.2 in 2045 [2].

The number of people with diabetes in India increased from 26·0 million (95% UI 23·4–28·6) in 1990 to 65·0 million (58·7–71·1) in 2016 [3]. In Maharashtra, overall reported prevalence of diabetes in urban and rural area is 10.9% and 6.5% respectively [2].

Enormous data and increased complexities have led to rising interest in the use of machine learning (ML) in healthcare. It develops on existing statistical methods and finds patterns in the data. Ml uses different models for prediction of Type 2 diabetes. The accuracy of these models are of prime importance as analysis is directly impacting patient's life. The aim of this research is to design a predictive model for estimation of diabetes in healthy people with diverse age groups based on different life style related factors i.e. stress, food habit, smoking, profession and exercise. The impact of data pre-processing with different scalers on ML model performance is studied methodically to improve decision support systems for physician.

Some of the important pre-processing steps include data cleaning, pruning, feature selection, and scaling. Many researchers considered diverse ML algorithms along with feature selection [4],[5] few considered the effect of the data scaling process on overall model performance [6]. Thus, the primary purpose of this study is to evaluate the effect of different data scaling methods on different ML algorithms and develop a prediction model for healthy patients with early diabetes symptoms.

In the present study, five machine learning algorithms like - Logistic regression, K Neighbours (KNN), Gaussian Naïve Bias (GNB), Decision Tree (DT) and Random Forest (RF) and 7 data scaling methods like MinMaxScaler, Sandard scalar, RobustScaler, QuantileTransformer (QT), PowerTransformer (PT) and Normalizer are used together to find the best match for type 2 diabetes prediction. The effect of different data scaling techniques is observed using the UCI PIMA India dataset [7] and LS data set [8] where data is collected through survey in Indian environment.

LITERATURE SURVEY

Survey on Machine Learning Algorithm for Type-2 Diabetes

Contreas et al., [9] study focuses on all AI techniques for diabetic prediction and management. Study shows an AI is powerful tool applicable for prediction and prevention of complications due to diabetes. AI techniques are being progressively utilized in area of medicine containing complex sets of diagnostic and clinical information. This tool helps in improving quality of patient's life through predictive approach which helps for improving health outcomes.

Sneha and Tarun [10] proposed method for selecting the attributes which will be used in early detection of Diabetes and showed Random forest model and decision tree model has specificity of 98.00% and 98.20%

In Sisodia et al., [11] designed a system which can prognosticate the likelihood of diabetes in patients by achieving with higher accuracy. Dataset used in this study is PIDD from UCI repository. Experimentation done using weka tool by applying NB, DT and SVM classifier for early detection of diabetes. Model performance measured using accuracy, precision, and recall and F-score. As reported in the paper, NB achieved the best performance results, with a maximum accuracy of 76.3% and highest ROC value of 81.9.

In Mahabub et al., [12] designed a system which can prognosticate the diabetes by improving an accuracy. Used 11 classifiers as, NB, KNN, SVM, DT, RF, ANN, LR, GB, AdaBoosting etc. on PIMA dataset. Evaluations of all models are examined on various measures like accuracy, precision, F-measure and recall. Ensemble voting classifier developed using 3 best classifiers as SVM, MLP and KNN by applying hyper parameter tuning and cross validation. The proposed ensemble framework gives an accuracy of almost 86%.

Ahmed et al., [13] designed a system for predicting DM using ML algorithms, namely, DT, KNN, NB, RF, GB, LR, and SVM. Label–encoding and data normalization, are used for improving an accuracy. Two different datasets are used, PIMA and Tigga and Garg. PIMA dataset provides the highest accuracy for SVM and RF with 80.26% and for another dataset, the highest accuracy achieved by DT and RF with 96.81%. Developed a web app. Model is compared with other studies, and the findings reveal that, the suggested model can offer greater accuracy of 2.71% to 13.13%.

Survey for Application of different scalers for data pre-processing for ML Models.

There are many data scaling techniques available for ML algorithms and they have different impact on efficacy of the ML model [14,15].

study conducted by Ambarwari et al. (2020) showed that data scaling techniques such as Minimax normalization and standardization have also significant effects on data analysis [14]. The study was carried out using ML algorithms such as KNN, Naïve Bayesian, ANN, and SVM with RB . The result discovered that MinMax scaling with SVM performed better than other algorithms.

Another study conducted by Balabaeva et al., [16] addressed the effect of different scaling methods on heart failure patient datasets. Their study uses more robust ML algorithms such as XGB, LR, DT, and RF with scaling methods such as Standard Scaler, MinMax Scaler, Max Abs Scaler, Robust scaler, and Quantile Transformer. In their study, RF showed higher performance with Standard and Robust Scaler.

METHODOLOGY

To maintain integrity of research accurate data collection is necessary. To investigate efficacy of machine learning algorithms at the earlier stages of predicting risk of diabetes following datasets were used in this study

1] UCI repository diabetes dataset - PIMA Indian diabetes dataset

2] Self-collected questionnaire based dataset – LS_ diabetes dataset

PIMA Dataset [7]

Standard Dataset: P dataset is a UCI Repository dataset, consist of 768 records with female centric data. It contains both 268 diabetic instances and 500 non-diabetic instances. Dataset comprises of numeric-valued 8 attributes. Data contain both medical examination data as well as personal health data. Age (age), Body mass index (bmi), Diastolic blood pressure (pres),Number of times pregnant (preg) , 2-h serum insulin (insu),Plasma glucose concentration at 2 h in an oral glucose tolerance test (plas) , Skin fold thickness (skin) , Pedigree function (pedi), Class variable (class)

LS dataset- Self-collected Questionnaire-based Dataset for the Study [8]

The dataset was developed by web-based questionnaires. LS_diabetes dataset comprising of 374 people with of 35 features the questions were related to demographic information, dietary pattern, life style, pathological and stress related factors as shown in Fig 1.

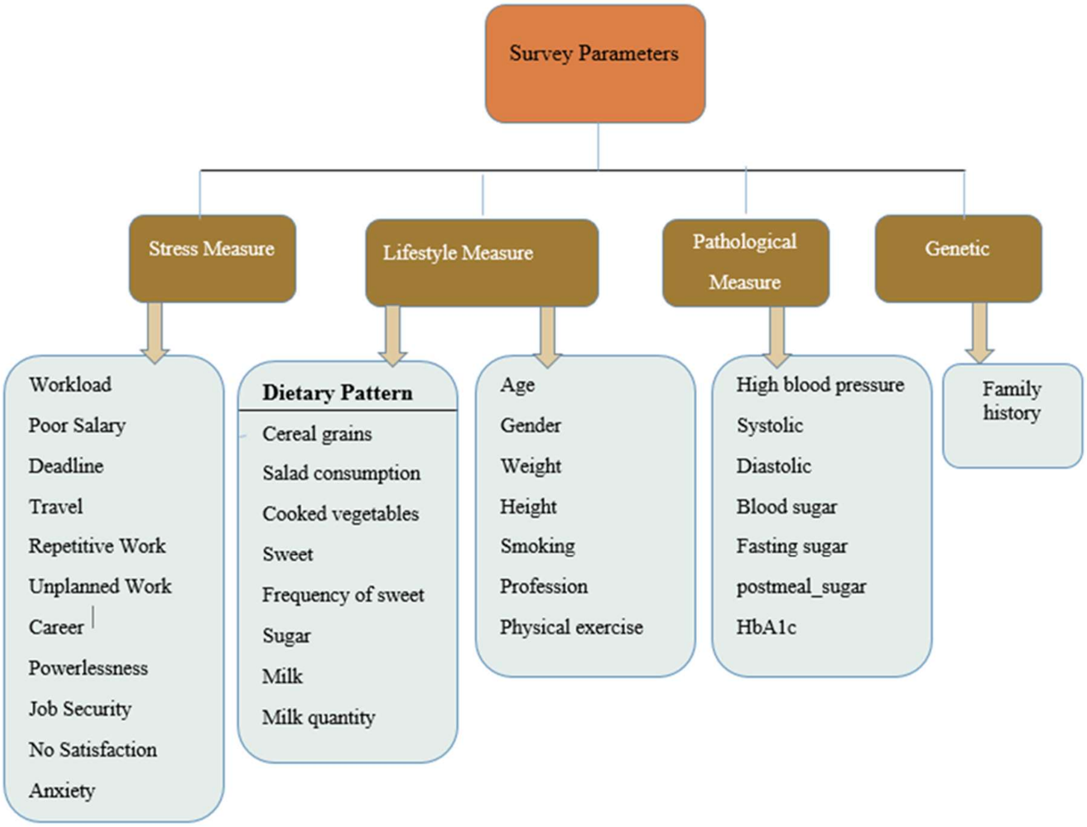

Fig 1: Features Collected Through Online Survey in Indian Environment for LS Dataset

Experimental Setup

The experiment was carried out by splitting the dataset into 80% and 20% for the training and testing set, respectively. The performance of the model was evaluated using 10-fold cross-validations, and the performance of the model is presented by averaging the outcomes of all 10 folds.

The results took place using the Anaconda modules with Python 3.7 and were run on an office-grade laptop with common specifications (Windows 11, AMD RYZEN 7 6800H, and 64 GB of RAM). Instead of developing different preprocessing steps, this study uses built-in preprocessing libraries provided by Scikit-learn tools: Normalization, Standardization, MinMax Scale, MaxAbs scale, Robust Scaler, Quantile Transformer

Following is a process flow where different scalers like Minimax Scaler, Standard Scaler, MaxAbs Scaler Robust Scaler, Quantile Transformer Scaler, Power Transformer Scaler and Normalizer Scale rare applied for data cleaning on both datasets with five diverse ML model like Logistic Regression, KNN, NB, DT and RF for predicting the type 2 diabetes. Fig 2 shows the process flow of the experiment

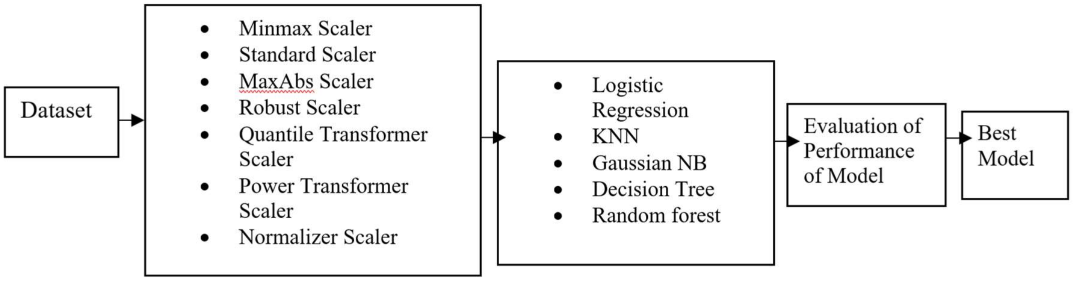

Fig 2: Process of Comparing Performance of ML Model with the Influence of Scalers as a Part of Preprocessing Stage

The Performance parameters of the model are calculated based on confusion matrix [17] as shown in Fig 3.

		Predicted Class	
Actual Class	Total Population	Positive Prediction	Negative Prediction
	Diabetes	True Positive (TP)	False Negative (FN)
	Non-diabetes	False Positive (FP)	True Negative (TN)

Fig 3: Confusion Matrix

The performance was evaluated based on accuracy, precision, recall, and F1 score.
The matrix outcomes are as follows:

- True positive (Tp) – TP denote the number of patients having diabetes and are predicted as diabetes individuals.
- False positive (Fp) –FP denote the number of patients having diabetes and are predicted as healthy individuals.
- True negative (Tn) –TN denote the number of patients not having diabetes and are predicted as healthy individuals.
- False negative (Fn) –FN denote the number of patients not having diabetes and are predicted as diabetes individuals.

The Accuracy, Precision and recall are defined as follows

$$Accuracy = \frac{Tp + Tn}{\mathrm{Tn + Tp + Fp + Fn}} \qquad (1)$$

$$Precision \; = \frac{Tp}{Tp+Fp} \qquad (2)$$

$$Recall/Sensitivity = \frac{Tp}{Tp+Fn} \qquad (3)$$

Characteristics of Different Scalers used for Preprocessing are

1) MinMax Scaler: It just scales all the data between 0 and 1.
   x_scaled = (x – x_min)/(x_max – x_min)
2) Standard Scaler- Z score
   For each feature, the Standard Scaler scales the values such that the mean is 0 and the standard deviation is 1
   x_ scaled = x – mean/std_dev
3) the MaxAbs scaler takes the absolute maximum value of each column and divides each value in the column by the maximum value. Range [-1,1]
4) Robust Scaler is not sensitive to outliers
   It removes the median from the data and scales the data by the InterQuartile Range(IQR)
   - Q1= First half of data and its median
   - Q2= Actual median
   - Q3= Second half of data and its median

   IQR = Q3 – Q1
   x_scaled = (x – Q1)/(Q3 – Q1)
5) Quantile Transformer Scaler is best to use this for non-linear data.
6) The Power Transformer actually automates this decision making by introducing a parameter called *lambda*. It decides on a generalized power transform by finding the best value of lambda
7) Normalizer: If we are using L1 norm, the values in each column are converted so that the sum of their absolute values along the row = 1

RESULTS AND DISCUSSION

The experiments were carried out on both datasets PIMA and LS with 80% training and 20% testing data. Following Table 1 to 5 shows the performance of PIMA dataset.

Table 1: Logistic Regression Model is Applied with Various Scalers in Data Preprocessing Stage PIMA Data Set

Sl. No.	Scaler Name	F1 Score	Precision	Recall	Accuracy (%)
1	No Scaler	0.45	0.47	0.43	46.99
2	Minmaxscaler	0.83	0.77	0.90	76.19
3	Sandard Scalar	0.84	0.80	0.88	77.92
4	Maxabsscaler	0.83	0.77	0.90	76.19

5	Robustscaler	0.84	0.80	0.88	77.92
6	Quantiletransformer	0.83	0.80	0.87	77.76
7	Powertransformer	0.82	0.79	0.85	75.76

Table 2: KNN Model is Applied with Various Scalers in Data Preprocessing Stage PIMA Data Set

Sl. No.	Scaler Name	F1 Score	Precision	Recall	Accuracy (%)
1	No scaler	0.68	0.73	0.63	69.88
2	MinMaxScaler	0.82	0.8	0.85	76.19
3	Sandard scalar	0.8	0.75	0.86	72.29
4	MaxAbsScaler	0.81	0.76	0.87	74.09
5	RobustScaler	0.8	0.76	0.85	72.73
6	QuantileTransformer	0.8	0.77	0.85	73.16
7	PowerTransformer	0.85	0.75	0.88	73.59

Table 3: GNB Model Is Applied with Various Scalers in Data Pre-Processing Stage PIMA Data Set

Sl. No.	Scaler Name	F1 Score	Precision	Recall	Accuracy (%)
1.	No scaler	0.49	0.49	0.48	49.4
2.	MinMaxScaler	0.82	0.79	0.85	75.76
3.	Sandard scalar	0.82	0.79	0.85	75.76
4.	MaxAbsScaler	0.82	0.79	0.85	75.76
5.	RobustScaler	0.82	0.79	0.85	75.76
6.	QuantileTransformer	0.82	0.81	0.83	76.62
7.	PowerTransformer	0.82	0.81	0.83	76.62

Table 4: DT Model Is Applied with Various Scalers in Data Preprocessing Stage PIMA Data Set

Sl. No.	Scaler Name	F1 Score	Precision	Recall	Accuracy (%)
1.	No scaler	0.55	0.58	0.53	57.23
2.	MinMaxScaler	0.76	0.75	0.76	68.04
3.	Sandard scalar	0.76	0.75	0.77	68.83
4.	MaxAbsScaler	0.78	0.76	0.81	70.56
5.	RobustScaler	0.76	0.74	0.77	67.97
6.	QuantileTransformer	0.76	0.75	0.77	68.04
7.	PowerTransformer	0.78	0.76	0.76	70.56

Table 5: RF Model is Applied with Various Scalers in Data Preprocessing Stage PIMA Data Set

Sl. No.	Scaler name	F1 score	Precision	Recall	Accuracy (%)
1.	No scaler	0.45	0.47	0.43	46.99
2.	MinMaxScaler	0.61	0.78	0.85	74.8
3.	Sandard scalar	0.62	0.79	0.84	74.89
4.	MaxAbsScaler	0.82	0.77	0.87	74.89
5.	RobustScaler	0.83	0.81	0.85	77.66
6.	QuantileTransformer	0.82	0.79	0.86	75.76
7.	PowerTransformer	0.81	0.78	0.84	74.46

It is observed that data cleaning and pre-processing with different scalers have remarkable impact on the accuracy of the ML models in PIMA dataset. In Logistic Regression model efficiency improved by 39.69 % with standard and robust scaler. In KNN Minmax scaler performed better than all other scalers and improve efficiency by 9.1 %. In GNB the accuracy is improved by 35.5 % and QT scaler performs better than other scalers. In DT the MaxAbs scaler performs better with improvement in efficiency by 18.89 % and in RF the efficacy improvement is 39.49 % with Robust scaler performing better. The performance in different ML model is based on data base characteristics and spread of data in features. Table 6 to 10 shows the performance of LS data set with respect to various scalers and different ML models.

Table 6: Logistic Regression Model is Applied with Various Scalers in Data Preprocessing Stage LS Data Set.

Sl. No.	Scaler name	F1 score	Precision	Recall	Accuracy (%)
1.	No scaler	0.59	0.8	0.47	85.33
2.	MinMaxScaler	1	1	1	99.87
3.	Sandard scalar	1	1	1	100
4.	MaxAbsScaler	1	1	1	100
5.	RobustScaler	1	1	1	100
6.	QuantileTransformer	1	1	1	98.92
7.	PowerTransformer	1	1	1	100

Table 7: KNN Model is Applied with Various Scalers in Data Preprocessing Stage LS Data Set

Sl. No.	Scaler name	F1 score	Precision	Recall	Accuracy (%)
1.	No scaler	0.43	0.55	0.35	78.67
2.	MinMaxScaler	1	1	1	100
3.	Sandard scalar	1	1	1	100
4.	MaxAbsScaler	1	1	1	100
	RobustScaler	1	1	1	100
5.	QuantileTransformer	1	1	1	100
6.	PowerTransformer	1	1	1	100

Table 8: GNB Model is Applied with Various Scalers in Data Preprocessing Stage LS Data Set

Sl. No.	Scaler Name	F1 Score	Precision	Recall	Accuracy (%)
1.	No Scaler	0.71	0.60	0.88	84.0
2.	Minmaxscaler	1	1	1	100
3.	Sandard Scalar	1	1	1	100
4.	Maxabsscaler	1	1	1	98.99
5.	Robustscaler	1	1	1	100
6.	Quantiletransformer	1	1	1	99.98
7.	Powertransformer	1	1	1	100

Table 9: DT Model is Applied with Various Scalers in Data Preprocessing Stage LS Data Set

Sl. No.	Scaler name	F1 score	Precision	Recall	Accuracy (%)
1.	No scaler	0.81	0.87	0.76	92.0
2.	MinMaxScaler	1	1	1	99.8
3.	Sandard scalar	1	1	1	100
4.	MaxAbsScaler	1	1	1	100

5.	RobustScaler	1	1	1	100
6.	QuantileTransformer	1	1	1	100
7.	PowerTransformer	1	1	1	99.46

Table 10: RF Model is Applied with Various Scalers in Data Preprocessing Stage LS Data Set

Sl. No.	Scaler name	F1 score	Precision	Recall	Accuracy (%)
1.	No scaler	0.83	1	0.71	93.33
2.	MinMaxScaler	1	1	1	100
3.	Sandard scalar	1	1	1	100
4.	MaxAbsScaler	1	1	1	100
5.	RobustScaler	1	1	1	100
6.	QuantileTransformer	1	1	1	100
7.	PowerTransformer	1	1	1	100

It is observed that data cleaning and preprocessing with different scalers have remarkable impact on the accuracy of the ML models in LS dataset. In Logistic Regression model efficiency improved by 14.67 %, in KNN accuracy improves by 21.33 %, In GNB efficiency improves by 16 %, in DT accuracy improves by 8% and RF accuracy improves by 6.66%. The variations in accuracy through different scalers is less in LS dataset as its controlled dataset collected by researchers so missing values are very less compared to PIMA dataset

CONCLUSION

As the stress and lifestyle parameters of individual are changing very rapidly towards negative curve the occurrences of diabetes at early age are expected in India and across the globe. For predicting the disease accurately at early stage, the experiment is done with Two datasets PIMA standard dataset and LS locally generated dataset. It is observed that data cleaning methods have high level of impact on accuracy of prediction in all models. Without scalar the accuracy of PIMA data set is from 46.99 to

69.88%, which improves with scalers upto 77.92 %. For LS dataset without sclares accuracy is as low as 78.67 which improves to 100% with two labels as the LS data set is small and controlled.

It is concluded that scaler have observable impact on the ML model prediction efficiency and as per the data spread if appropriate scaler is selected the accuracy of predication can be surely improving

Future work: The data scaling methods can be applied to other datasets in health care domain to improve the accuracy of prediction of decease to help mankind.

REFERENCES

- https://www.diabetes.co.uk/
- http://www.who.int/en/news-room/fact-sheets/detail/diabetes accessed on 21st Feb 2019
- Harris, M.L., Oldmeadow, C., Hure, A., Luu, J., Loxton, D., & Attia, J. (2017). Stress increases the risk of type 2 diabetes onset in women: A 12-year longitudinal study using causal modelling. PLoS One, 12(2):e0172126.doi:10.1371/journal. pone.0172126
- Kaur, H. & Kumari, V. (2018). (2022). Predictive Modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics, 12(½): 90-100.
- Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disorders, 19(1): 1–9. https://doi.org/10.1186/s12902-019-0436-6
- Srinivas, K., Rani, B.K. & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. Int. J. Comput. Sci. Eng., 2: 250–255.
- Pima Indians Diabetes dataset. Available from: http://archive.ics.uci.edu/ml/machine learning-databases/pima-indiansdiabetes/pima-indians-diabetes
- Patil, P. and Shah, L. (2019). Assessment of risk of type 2 diabetes mellitus with stress as a risk factor using classification algorithms. International Journal of Recent Technology and Engineering, 2019, 8(4): 11273-11277.
- Contreas, I, Vehi, J. (2018). Artificial intelligence for diabetes management and decision support: Literature review. J Med Internet Res, 20: e10775. doi: 10.2196/10775
- Sneha, N. and Gangil. T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection: J Big Data, 6:13 https://doi.org/10.1186/s40537-019-0175- 6
- Sisodia, D. and Sisodia, D. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132: 1578–1585.
- Mahabub, A., (2019). A robust voting approach for diabetes prediction using traditional machine learning techniques, SN Applied Sciences, 1:1667.
- Ahmeda, N., Ahammeda, R., Islama, M., Uddina, A., et al., (2021). Machine learning based diabetes prediction and development of smart web application, International Journal of Cognitive Computing in Engineering, 2: 229-241.
- Ambarwari, A. Adrian, Q.J., Herdiyeni, Y. (2020). Analysis of the Effect of Data Scaling on the Performance of the Machine Learning Algorithm for Plant Identification. J. Resti (Rekayasa Sist. Dan Teknol. Inf.) 4:117–122.
- Shahriyari, L. (2019). Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. Briefings Bioinform, 20, 985–994.
- Balabaeva, K. and Kovalchuk, S. (2019). Comparison of Temporal and Non-Temporal Features Effect on Machine Learning Models Quality and Interpretability for Chronic Heart Failure Patients. Procedia Comput. Sci. 156:87–96.

- Visa, S., (2011). Brian RamsaySentry Anca Ralesce Depa Esther van der Knaap “Confusion Matrix-based Feature Selection” Conference: Proceedings of the 22nd Midwest Artificial Intelligence and Cognitive Science Conference 2011, Cincinnati, Ohio, USA, April, 16-17.