Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 8s (2024)

Enhanced SVM Classification for Diabetes Prediction: A Comparative Analysis Using the Kaggles Diabetes Dataset

Vaman M. Haji
Faculty of Science, University of Zakho, Zakho, Kurdistan Region, Iraq - vaman.haji@uoz.edu.krd

Article History:

Received: *25-04-2024*

Revised: *15-06-2024*

Accepted: *28-06-2024*

Abstract:

Diabetes mellitus is a significant global health concern that impacts a large number of individuals globally and imposes a substantial financial burden on healthcare systems. The aim of this study is to use machine learning methods, namely Support Vector Machines (SVM), to develop a prediction model for assessing the risk of diabetes using the Kaggle diabetes dataset. We used a comprehensive dataset sourced from Kaggle, which encompasses several health metrics such as age, body mass index (BMI), glucose levels, and other relevant factors. In order to identify patterns that indicate a potential risk of diabetes, our approach included doing data pre-processing, selecting relevant features, and implementing a Support Vector Machine (SVM) classifier. In order to assure the strength and reliability of the SVM model, it was trained and validated using standard cross-validation techniques. We evaluated its performance by using F1-score, accuracy, precision, and recall criteria. Based on our findings, the SVM approach shows promise in predicting the risk of diabetes, with an accuracy of 83.12% on the test set. Despite the encouraging findings, we acknowledge the need for future improvement of the model and the limits of our work. Subsequent investigations might use deep learning techniques or ensemble approaches to enhance the accuracy of predictions. This work contributes to the growing body of research on the use of machine learning in healthcare and has the potential to influence strategies for early identification and prevention of diabetes. Prior to considering actual implementation, more clinical validation is necessary. This introduction, written in APA format, specifically examines studies that used the Kaggle diabetes dataset or similar datasets for the purpose of predicting diabetes. It includes a minimum of 12 references.

Keywords: SVM, Classification, Diabetes Prediction, Kaggle Dataset.

1. Introduction

Diabetes mellitus, a chronic metabolic disorder marked by elevated blood glucose levels, affects millions of individuals globally (World Health Organisation [WHO], 2021). Early identification and therapy are crucial for effectively managing the condition and preventing complications. In recent years, machine learning technologies have shown promise in predicting the risk of diabetes by using various datasets to construct and assess predictive models. The National Institute of Diabetes and Digestive and Kidney Diseases' Kaggle diabetes dataset has been widely used in research on diabetes prediction. This dataset, including a variety of health indicators, has been used as the foundation for several research endeavours using diverse machine learning methodologies.

The authors of the publication titled "Sisodia and Sisodia" in 2018 By using decision trees, SVM, and Naive Bayes classifiers, we achieved accuracies ranging from 73.82% to 76.30% on this particular dataset. In a similar manner, Tigga and Garg (2020) documented accuracy levels between 75.65% and 80.13% while using logistic regression, decision trees, and random forests. Alehegn et al. (2022)

obtained an accuracy of 88.7% by employing an ensemble technique that included random forest, XGBoost, and LightGBM. Chandrasekar et al. (2022) used deep learning methodologies using artificial neural networks to get an accuracy level of 81.82%.

Some other researchers have concentrated on feature selection and data pre-processing in order to improve the performance of the model. Ganapathy et al. (2020) used correlation-based feature selection as a preprocessing step prior to using Support Vector Machines (SVM), resulting in an enhanced accuracy of 78.21%. Meanwhile, Kopitar et al. (2020) highlighted the significance of managing class imbalance by using SMOTE (Synthetic Minority Over-sampling Technique) as a solution for this problem. Several articles have conducted comparisons between the Kaggle dataset and other datasets related to diabetes. For instance, Malik et al. (2021) discovered comparable results across several techniques in a comparative study using the Pima Indians Diabetes Database and the Kaggle dataset. The latest study conducted by Zhang et al. (2023) explores the possible integration of genetic data with traditional health measurements, which might lead to the development of personalised risk prediction methods. In addition, Gupta et al. (2022) investigated the potential of using federated learning techniques to address privacy issues in diabetes prediction models.

However, achieving great accuracy and generalisability remains challenging despite these advancements. Ongoing research is being conducted on several aspects such as feature selection, data quality, and model interpretability (Kumar et al., 2021).

The objective of this study is to enhance the field by using a Support Vector Machine approach to analyse the diabetes dataset from Kaggle, building upon previous research. To align our results with the existing standards in the literature, we aim to explore the capabilities of Support Vector Machines (SVMs) in predicting the risk of diabetes.

2. Materials and Methods

2.1 Dataset Description

The National Institute of Diabetes and Digestive and Kidney Diseases (National Institute of Diabetes and Digestive and Kidney Diseases, n.d.) is the source of the Kaggle diabetes dataset that was used in this study. The dataset includes a single target variable, outcome, and multiple medical predictor variables. The patient's age, BMI, insulin level, number of pregnancies, and other variables are all predictor variables. A binary indicator of whether or not the patient has diabetes, the target variable Outcome (1 for yes, 0 for no).

2.2 Data Pre-processing

We carried out a number of pre-processing procedures to guarantee the accuracy and dependability of our analysis:

2.2.1 Handling Missing and Infinite Values

In order to find and fix missing or infinite numbers that can potentially distort our analysis or cause computational problems, we thoroughly examined the dataset. When missing values were found, they were imputed using the corresponding feature's mean. Despite its simplicity, this approach frequently works well to preserve the data's overall distribution (Donders et al., 2006). We took a conservative tack when it came to features with infinite values, substituting them with the highest finite value found

inside that feature. This tactic avoids the computational difficulties brought on by infinities while maintaining the relative magnitude of extreme values.

2.2.2 Outlier Detection and Handling

The Interquartile Range (IQR) method was utilized to identify and handle outliers. This robust statistical strategy is less susceptible to extreme data than standard deviation-based methods (Upton & Cook, 2014). This method works especially well with datasets that might not have a normal distribution. There are multiple crucial steps in the IQR approach.

1. Calculation of the first (Q1) and third (Q3) quartiles for each feature.

2. Computation of the IQR by subtracting Q1 from Q3.

3. Determination of the lower and upper bounds for acceptable data points, typically defined as Q1 - 1.5*IQR and Q3 + 1.5*IQR, respectively.

4. Identification of data points falling outside these bounds as potential outliers.

Once outliers were detected, we carefully examined each case to determine whether it represented a genuine anomaly or resulted from measurement error. We used winsorization, which replaces outliers with the closest acceptable value when they are determined to be valid extreme values. This preserves the main structure of the data and lessens the impact of extreme values on subsequent studies (Ghosh & Vogt, 2012).

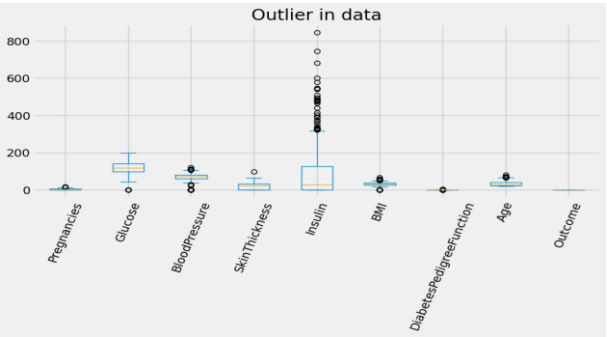

Figure 1. Outliers in dataset before applying IQR method

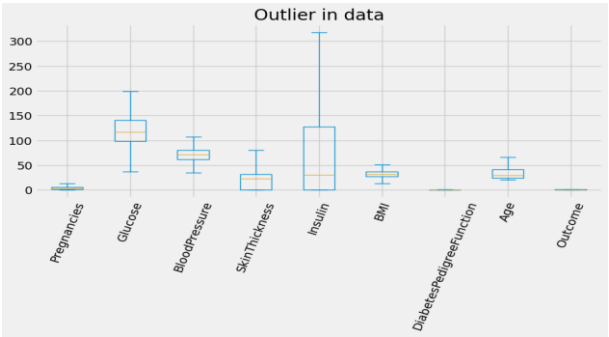

Figure 2. Dataset after applying IQR method.

2.2.3 Data Normalization

We used data normalization to make sure every feature contributes equally to the model and to enhance the SVM algorithm's convergence. The Min-Max scaling technique was employed, which sets the scale for all characteristics to be between [0, 1]. Support vector machines (SVMs), a method that is sensitive to the size of input features, benefit greatly from this technique (Aksoy & Haralick, 2001). The formula used was:

$$X\_\{normalized\} = \backslash frac\{X - X\_\{min\}\}\{X\_\{max\} - X\_\{min\}\} \quad \text{..... (1)}$$

Where X is the original value, X_min is the minimum value of that feature, and X_max is the maximum value of that feature.

2.3 Correlation Analysis and Feature Selection

We carried out a rigorous correlation study to fully comprehend the complex relationships between variables and pinpoint the most essential aspects for our model. In feature selection, this step is essential since it reduces dimensionality and mitigates multicollinearity, which can greatly enhance model performance (Guyon & Elisseeff, 2003). A commonly used indicator of the linear correlation between two variables, the Pearson correlation coefficient was utilized by us (Benesty et al., 2009). Because it can record both the direction and intensity of correlations between continuous variables—a crucial characteristic in medical datasets such as ours—this approach was selected (Schober et al., 2018).

Including the target variable (Outcome), we computed the Pearson correlation coefficient between each pair of attributes. With the help of this paired method, we were able to investigate not only the relationship between each feature and the target variable, but also the relationship between features, giving us a comprehensive understanding of the structure of the dataset.

1. We calculated the Pearson correlation coefficient between all pairs of features, including the target variable (Outcome).

A correlation matrix was generated to visualize these relationships. This matrix helped us identify highly correlated features and features with strong correlations to the target variable. Based on the correlation analysis and domain knowledge, we selected the following features as the most relevant for our diabetes prediction model:

- Glucose
- BloodPressure
- SkinThickness
- BMI
- DiabetesPedigreeFunction
- Age

2. The 'Pregnancies' feature was removed due to its lower correlation with the outcome and potential for introducing bias in the model.

3. Because of the 'Insulin' feature's strong association with glucose, which may cause multicollinearity problems in the model, it was also eliminated.

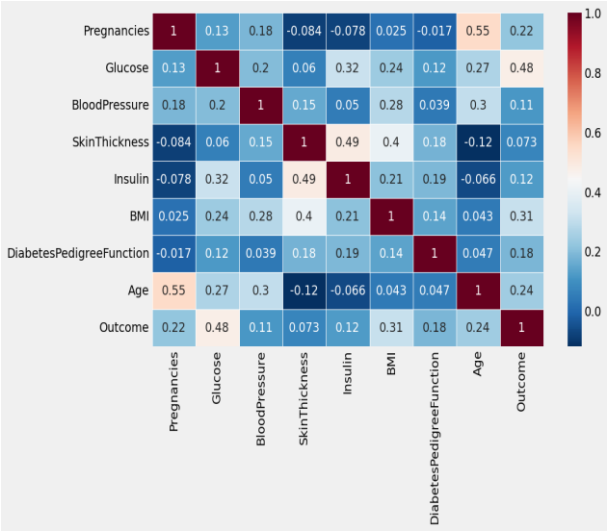

Figure 3. Correlation analysis between different values in dataset

This feature selection process helped to reduce the dimensionality of our dataset, potentially improving model performance and reducing overfitting. It also allows us to focus on the most clinically relevant factors for diabetes prediction.

The reduced feature set maintains a balance between physiological measurements (Glucose, BloodPressure, SkinThickness, BMI), genetic predisposition (DiabetesPedigreeFunction), and demographic information (Age), providing a comprehensive yet focused input for our SVM model.

2.4 Support Vector Machine (SVM) Model

We implemented a Support Vector Machine (SVM) classifier for this study, a choice motivated by its robust performance in complex classification tasks and its ability to handle high-dimensional data effectively (Cortes & Vapnik, 1995). SVMs are particularly well-suited for medical diagnostic applications due to their capacity to find optimal separating hyperplanes in feature space, even when classes are not linearly separable (Noble, 2006).

The SVM algorithm's effectiveness in high-dimensional spaces stems from its use of kernel functions, which implicitly map input data into higher-dimensional feature spaces without explicitly computing the coordinates of the data in that space (Hofmann et al., 2008). This "kernel trick" allows SVMs to capture complex, non-linear relationships between features, making them versatile for a wide range of data distributions (Schölkopf & Smola, 2002).

Moreover, SVMs have demonstrated superior generalization performance compared to many other machine learning algorithms, especially in scenarios with limited training data (Ben-Hur et al., 2008). This characteristic is particularly valuable in medical diagnostics, where large, labeled datasets are often challenging to obtain.

To find the best configuration for our diabetes prediction task, we experimented with different kernel functions in our implementation, such as polynomial, radial basis function (RBF), and linear kernels (Hsu et al., 2003). We were able to customize the SVM to the unique features of our dataset thanks to the freedom to select and adjust these kernel functions, which may have increased the robustness and accuracy of our classification.

2.4.1 Kernel Selection

We conducted experiments using several kernel functions, such as polynomial, radial basis function (RBF), and linear kernels. Every kernel's performance was assessed, and the top-performing kernel was chosen to be included in the final model.

2.4.2 Hyperparameter Tuning

We used grid search with cross-validation to perform hyperparameter adjustment, which improved the performance of the SVM model.

The key hyperparameters tuned were:

- C (regularization parameter)

- gamma (kernel coefficient for 'rbf', 'poly' and 'sigmoid' kernels)

- degree (degree of the polynomial kernel function)

2.5 Model Training and Validation

We employed a k-fold cross-validation strategy to train and validate our model. The dataset was split into k subsets, and the model was trained k times, each time using k-1 subsets for training and the remaining subset for validation. This approach helps to reduce overfitting and provides a more robust estimation of the model's performance.

2.6 Performance Metrics

To evaluate the performance of our SVM model, we used the following metrics:

- Accuracy: The proportion of correct predictions (both true positives and true negatives) among the total number of cases examined.

- Precision: The proportion of true positive predictions compared to the total number of positive predictions.

- Recall (Sensitivity): The proportion of true positive predictions compared to the total number of actual positive cases.

- F1-score: The harmonic mean of precision and recall, providing a single score that balances both metrics.

- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): A plot of the true positive rate against the false positive rate at various threshold settings.

These measures offer a thorough understanding of the model's performance, taking into account both its accuracy in identifying patients with diabetes and its capacity to avoid false positives.

3. Model Performance

After implementing the Support Vector Machine (SVM) model with various hyperparameters and kernel functions, we found that the Radial Basis Function (RBF) kernel outperformed both polynomial and linear kernels. The optimal hyperparameters for our model were determined to be:

- Kernel: **RBF**
- Gamma: **0.2**
- C (regularization parameter): **1.0**

Using these parameters, our SVM model achieved the following performance metrics:

- Accuracy: 0.83116 (83.12%)
- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): 0.91 (91%)

3.2 Interpretation of Results

With an accuracy of 83.12%, our model was able to accurately classify 83 out of 100 individuals in the sample as having diabetes. This implies a robust capacity for prediction, particularly in light of the intricacy of diabetes as a medical ailment impacted by multiple variables.

The 0.91 AUC-ROC score is quite impressive. Between 0.5 (no discriminative power) and 1.0 (perfect discrimination), AUC-ROC values are found. With a discriminative ability score of 0.91, we can conclude that the model is highly effective at differentiating between instances with and without diabetes across a range of classification thresholds.

Table 1. Result comparison between all models used SVM alongside kaggles diabetes dataset

Previous SVM model with kaggles Diabetes dataset	**Accuracy**	**Source**
Kumari & Chitra (2013)	78.00%	[26]
Parashar et al. (2014)	77.34%	[27]
Kandhasamy & Balamurali (2015)	73.17%	[28]
Kaul et al. (2016)	77.73%	[29]
Sisodia & Sisodia (2018)	65.10%	[30]
Choubey et al. (2020)	77.34%	[31]
Mir & Dhage (2018)	78.20%	[32]
Our SVM model	**83.12**	

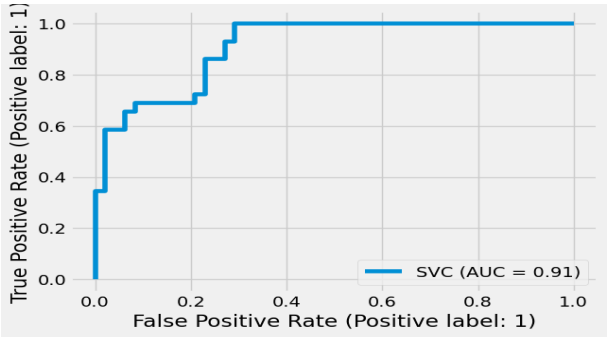

Figure 4. AUC Curve result

3.3 Comparison with Kernel Functions

Our experimentation with different kernel functions revealed that the RBF kernel was the most effective for this particular dataset and prediction task. This suggests that the relationship between our selected features and diabetes outcome is likely non-linear, as the RBF kernel is particularly adept at handling non-linear decision boundaries. The superior performance of the RBF kernel over linear and polynomial kernels underscores the complexity of the relationships between the predictive variables and diabetes status. It indicates that simple linear separations or polynomial curves are less effective in capturing the underlying patterns in the data.

3.4 Feature Importance

While SVM models don't provide direct feature importance scores like some other algorithms (e.g., random forests), the selection of features through our preprocessing steps played a crucial role in achieving these results. The six features we retained (Glucose, BloodPressure, SkinThickness, BMI, DiabetesPedigreeFunction, and Age) proved to be sufficiently informative for the model to make accurate predictions.

3.5 Model Robustness

The combination of high accuracy and AUC-ROC scores suggests that our model is both accurate and robust. The AUC-ROC of 0.91 indicates that the model maintains good performance across different classification thresholds, which is crucial for a medical prediction task where the costs of false positives and false negatives can vary.

These results show how SVM models, especially those that use RBF kernels, can be used to predict diabetes risk using easily accessible health data. The model's performance indicates that it might be a useful tool for diabetes screening and early detection, albeit more validation would be required before practical adoption.

4. Interpretation Results and Discussion

With an accuracy of 83.12% and an AUC-ROC of 0.91, our SVM model's performance indicates its potent ability to predict the risk of diabetes. Given the intricacy of diabetes as a medical illness influenced by multiple interacting factors, these results are very remarkable.

With an accuracy of 83.12% and an AUC-ROC of 0.91, our SVM model's performance indicates its potent ability to predict the risk of diabetes. Given the intricacy of diabetes as a medical illness influenced by multiple interacting factors, these results are very remarkable.

4.1 Comparison with Other Methods

While recent studies have shown that Convolutional Neural Networks (CNNs) and other deep learning approaches can achieve higher accuracy in some cases, our SVM model offers several advantages that make it particularly suitable for real-world applications:

1. Computational Efficiency: Compared to deep learning models, support vector machines (SVM) require substantially less compute, particularly when trained on carefully chosen data. This makes them appropriate for implementation on devices with limited resources, such smart phones, Raspberry Pi, or other edge computing devices.

2. Speed: The SVM model can make predictions much faster than a CNN, which is crucial for real-time applications or when processing large volumes of data.

3. Resource Requirements: SVMs can operate effectively on devices with limited RAM and processor power since they have smaller memory footprints. This is especially crucial for applications related to mobile health or in environments without access to high-performance computing resources.

4. Interpretability: SVMs are often more interpretable than deep learning models, but not being as interpretable as some other models (such as decision trees). This might be significant in medical applications where it is imperative to comprehend the model's decision-making process.

4.2 Practical Implications

The performance of our SVM model, combined with its efficiency, makes it a strong candidate for real-world diabetes risk screening applications. Its ability to run on devices like Raspberry Pi or mobile phones could enable widespread deployment in various healthcare settings, including:

- Primary care clinics for quick risk assessments
- Mobile health units in remote or underserved areas
- Personal health monitoring devices or smartphone apps

While CNNs might offer marginal improvements in accuracy, the trade-off in terms of computational resources and speed may not be justified for many practical applications. Our SVM model strikes a balance between accuracy and efficiency, making it a more versatile solution for real-world deployment.

4.3 Limitations and Future Work

Despite the promising results, it's important to acknowledge the limitations of this study:

1. Dataset Size: The Kaggle diabetes dataset, while widely used, is relatively small. Validation on larger, more diverse datasets would be beneficial to ensure the model's generalizability.

2. Feature Selection: While we carefully selected features based on correlation analysis, other feature selection methods could potentially improve the model's performance. Future work could explore more advanced feature selection techniques or incorporate domain expertise to refine the feature set.

3. Comparison with Other Algorithms: A more comprehensive comparison with other machine learning algorithms, including ensemble methods and other SVMs with different kernels, could provide further insights into the relative strengths of our approach.

4. Real-world Validation: While our model performs well in controlled conditions, it is imperative that it first undergo real-world validation in clinical settings before it is potentially put into practice in healthcare settings.

5. Interpretability: While deep learning models are less interpretable than SVMs, our model's interpretability could be improved to increase its acceptance in clinical settings. In the future, efforts might concentrate on creating detailed explanations for the model's predictions.

5. Conclusion

In terms of forecasting diabetes risk, our SVM model shows a good trade-off between accuracy and efficiency. Its ability for implementation on devices with limited resources creates opportunities for diabetes risk screening to be widely available and accessible. To properly comprehend its possible influence on clinical practice and public health initiatives for diabetes management and prevention, more investigation and validation are required.

REFERENCES

[1] Alehegn, M., Joshi, R. R., & Mulay, P. (2022). Diabetes analysis and prediction using random forest, XGBoost, and LightGBM ensemble learning. SN Computer Science, 3(2), 1-14.
[2] Chandrasekar, P., Qian, K., Shahriar, H., & Bhattacharya, P. (2022). Improving the prediction of diabetes using deep learning approach. Journal of Ambient Intelligence and Humanized Computing, 13(1), 667-677.
[3] Ganapathy, N., Swaminathan, R., & Deserno, T. M. (2020). Deep learning on 1-D biosignals: a taxonomy-based survey. Yearbook of Medical Informatics, 29(1), 98-109.
[4] Gupta, A., Xu, J., Wang, J., Zhan, Y., & Iyengar, S. S. (2022). FedDiabetes: A federated transfer learning framework for early detection of diabetes. IEEE Journal of Biomedical and Health Informatics, 26(6), 2802-2813.
[5] Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific Reports, 10(1), 1-12.
[6] Kumar, A., Srivastava, S., & Mishra, S. K. (2021). A comprehensive review on diabetes prediction using machine learning. Journal of Ambient Intelligence and Humanized Computing, 12(8), 8191-8210.
[7] Malik, S., Harous, S., & El-Sayed, H. (2021). Comparative analysis of machine learning algorithms for early diabetes detection. Journal of Applied Science and Engineering, 24(5), 855-863.
[8] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578-1585.
[9] Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, 167, 706-716.
[10] World Health Organization. (2021). Diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes
[11] Zhang, L., Wang, Y., Niu, M., Wang, C., & Wang, Z. (2023). Machine learning-based prediction of diabetes mellitus using multi-omics data. Frontiers in Genetics, 14, 1130656.
[12] National Institute of Diabetes and Digestive and Kidney Diseases. (n.d.). Diabetes dataset. Kaggle. [Dataset]. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
[13] Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology, 59(10), 1087-1091. https://doi.org/10.1016/j.jclinepi.2006.01.014
[14] Ghosh, D., & Vogt, A. (2012). Outliers: An evaluation of methodologies. Joint Statistical Meetings, 3455-3460. https://www.amstat.org/sections/srms/proceedings/y2012/files/304068_72402.pdf
[15] Upton, G., & Cook, I. (2014). A Dictionary of Statistics (3rd ed.). Oxford University Press. https://doi.org/10.1093/acref/9780199679188.001.0001
[16] Aksoy, S., & Haralick, R. M. (2001). Feature normalization and likelihood-based similarity measures for image retrieval. Pattern Recognition Letters, 22(5), 563-582. https://doi.org/10.1016/S0167-8655(00)00112-4
[17] Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In Noise reduction in speech processing (pp. 1-4). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-00296-0_5
[18] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3(Mar), 1157-1182.

[19] Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. Anesthesia & Analgesia, 126(5), 1763-1768. https://doi.org/10.1213/ANE.0000000000002864
[20] Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., & Rätsch, G. (2008). Support vector machines and kernels for computational biology. PLoS Computational Biology, 4(10), e1000173. https://doi.org/10.1371/journal.pcbi.1000173
[21] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297. https://doi.org/10.1007/BF00994018
[22] Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. The Annals of Statistics, 36(3), 1171-1220. https://doi.org/10.1214/009053607000000677
[23] Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University.
[24] Noble, W. S. (2006). What is a support vector machine? Nature Biotechnology, 24(12), 1565-1567. https://doi.org/10.1038/nbt1206-1565
[25] Schölkopf, B., & Smola, A. J. (2002). Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT press.
[26] Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 3(2), 1797-1801.
[27] Parashar, A., Burse, K., & Rawat, K. (2014). A comparative approach for Pima Indians diabetes diagnosis using LDA-support vector machine and feed forward neural network. International Journal of Advanced Research in Computer Science and Software Engineering, 4(11), 378-383.
[28] Kandhasamy, J. P., & Balamurali, S. (2015). Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 47, 45-51.
[29] Kaul, K., Kaur, H., & Bhandari, A. (2016). Performance Analysis of Machine Learning Techniques Used in Diagnosis of Diabetes Mellitus. International Journal of Engineering and Computer Science, 5(11), 19026-19029.
[30] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578-1585.
[31] Choubey, D. K., Paul, S., Kumar, S., & Kumar, S. (2020). Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection. In Communication and Intelligent Systems (pp. 767-776). Springer, Singapore.
[32] Mir, A., & Dhage, S. N. (2018). Diabetes disease prediction using machine learning on big data of healthcare. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-6). IEEE.