Jahangirnagar University Journal of Statistical Studies
ISSN 1022-4734

http://www.juniv.edu/department/stat

JUJSS, Vol. 37, 2023, pp.45-56

Predicting Diabetes Status using Machine Learning Algorithms with Hyperparameter Tuning: Evidence from Bangladesh Demographic and Health Survey 2017-2018 Data

Sonjit Mondol*
Department of Statistics, Jahangirnagar University, Savar, Dhaka, Bangladesh
and
Ajit Kumar Majumder
Department of Statistics, Jahangirnagar University, Savar, Dhaka, Bangladesh

Abstract

Diabetes is a metabolic disorder in which the body fails to produce enough insulin to keep blood sugar levels stable. It can cause heart disease, renal failure, nerve damage, and blindness if left undiagnosed and untreated. So, early identification of diabetic disorders is crucial for living a healthy life. However, it might be difficult to ascertain the diabetes status of people who live in remote areas or other places where testing or exploring diabetes is challenging. Furthermore, it takes time and money to diagnose diabetes in urban areas because a doctor's appointment, as well as a diagnostic facility visit, are required. In this situation, machine learning can be used to handle these concerns as there are numerous algorithms available to deal with this type of categorization problem. The objective of this work is to design an algorithm and optimize its hyperparameter so that it can properly identify diabetes based on a patient's early symptoms without conducting a diagnostic test. As a consequence, on the Bangladesh Demographic and Health Survey (BDHS) 2017–18 dataset, three machine learning algorithms—Decision Tree (DT), Neural Network (NN), and Support Vector Machine (SVM), as well as a grid search hyperparameter tuning strategy—are employed. Evaluation metrics—accuracy, sensitivity, specificity, kappa, and ROC curve—are used to evaluate the performances of these algorithms. In comparison to the other two algorithms, the NN has the highest accuracy, at 71.71%, with a 4.70% improvement brought on by hyperparameter optimization.

Keywords: Diabetes; Decision Tree; Neural Network; Support Vector Machine; Hyperparameter Tuning; Accuracy.

1. Introduction

Diabetes is a life-threatening metabolic condition that can damage every organ in the body. Although it remains incurable, it can be effectively controlled through treatment and the use of medication. It is influenced by several variables including height, weight, hereditary factors, and insulin, but sugar concentration is the main component (Vijayan & Anjali, 2016). Diabetes can be classified into three types. Insulin-Dependent Diabetes Mellitus (IDDM) is the medical term for type 1 diabetes. This type of diabetes is caused by the body's inability to produce enough insulin. In this case, the patient must inject insulin. Non-Insulin-Dependent Diabetes Mellitus (NIDDM) is another name for Type 2. This type of Diabetes occurs when body cells are unable to properly use insulin. Type-3 Gestational Diabetes raises a pregnant woman's blood sugar

* Corresponding author: sonjit214ju@gmail.com

level (Mujumdar & Vaidehi, 2019). This occurs when diabetes is not detected early. According to WHO (2021) statistics, diabetes kills about 1.6 million people worldwide each year and affects about 8.5% of people ages 18 and up. Although most developing countries saw a decline in the rate of diabetes-related fatalities between 2000 and 2010, the numbers spiked between 2010 and 2016 (N. Ahmed et al., 2021)**.** It is undeniably a serious threat in a developing country like Bangladesh, where a large portion of the population is unaware of its harmful effects (S. Ahmed et al., 2017). The prevalence of diabetes has risen rapidly in Bangladesh as a result of remarkable demographic, socioeconomic, and epidemiological changes over the last few decades (Khan et al., 2011). Although approximately 8 million people in Bangladesh are already affected by this global health issue, unfortunately, the availability of basic technologies, medicine, and necessary procedures for diabetes monitoring in primary health care is extremely limited (Pranto et al., 2020). Thus, this research holds substantial importance as it enables the detection of diabetes at an early stage.

Due to its global spread and increased incidence, it is currently regarded as an epidemic (Kaur & Kumari, 2022). Undiagnosed diabetes or persistent blood glucose elevations can lead to life-threatening conditions affecting the cardiovascular system, kidneys, vessels, teeth, eyes, and nerves, eventually leading to death (Kandhasamy & Balamurali, 2015). However, by employing necessary treatments and adopting a healthy lifestyle at an early stage, the symptoms and long-term complications of diabetes can be controlled. For diagnosing early-stage diabetes mellitus Machine Learning (ML) algorithms come as a blessing cause different types of ML algorithms like Logistic Regression, SVM, NN, Naïve Bayes, etc. (Nai-arun & Moungmai, 2015**;** Islam et al., 2020) are available for solving this kind of medical problem.

In this work, three machine learning classification algorithms (DT, NN, and SVM) are used on the BDHS 2017–18 dataset to see how effectively they can predict respondents' diabetes status. The experimental results of these algorithms are compared to one another on various metrics like accuracy, sensitivity, specificity, kappa, and ROC curve and revealed acceptable accuracy. Finally, tuning the model hyperparameters to improve the chosen algorithms' predictive ability, displaying the accuracy improvement as a result of tuning, and identifying the algorithm with the best accuracy for detecting patients with early-stage diabetes mellitus.

The rest of this paper is structured as follows: In Section II, the related study of several classification algorithms for diabetes prediction is briefly reviewed. Section III provides an explanation of the methodologies as well as a quick summary of the dataset that is used. Section IV covers the results and discussions, while Section V concludes the research work.

2. Related Works

In the medical sector, Machine learning has become a more popular and useful technique, and different types of ML-based research works are available. Researchers applied different types of ML algorithms to predict diabetes. Pranto at el. (2020) used decision tree (DT), K-nearest neighbor (KNN), random forest (RF), and Naïve Bayes (NB) on the PIMA dataset to train and then applied these train algorithms on unseen parts of PIMA and collected dataset from Kurmitola General Hospital, Dhaka. They found both random forest and NB classifiers performed well on both datasets, with accuracy levels exceeding 72% (Pranto et al., 2020). On the Pima Indians Diabetes dataset, Thirumal and Nagarajan (2015)employed the NB, C4.5, KNN and SVM for the early diagnosis of diabetes and discovered that C4.5 has a greater accuracy of 78.25% than other algorithms(Thirumal & Nagarajan, 2015). In order to categorize diabetic patients, Alehegn and Joshi (2017) used KNN, RF, J48, and NB. They concluded that the decision tree performed better than the other three algorithms in terms of diabetes prediction(Alehegn et al., 2017). For detecting early-stage diabetes mellitus, Sneha and Gangil (2019) reported that the KNN has a minimum accuracy rate of 63.04% and the support vector machine has a rate of 77.73% accuracy(Sneha & Gangil, 2019). Choudhury and Gupta (2019) employed SVM, logistic regression (LR), NB, RF, DT, and artificial neural networks (ANN) to identify diabetes early, andthe findings demonstrate that LR, with a precision of 0.7761, more reliably predicts diabetes disease(Choudhury & Gupta, 2019). In (Khanam & Foo, 2021)Khanam and Foo (2021) used NB, SVM, LR, Adaboost, RF, KNN, DT, and NN with different hidden layers for predicting diabetes status on the PIMA dataset, and they found LR and SVM perform well with 76.85% and 76.82% accuracy, respectively. Ahamed et al. (2021) used different datasets to train DT, NB, KNN, RF, Gradient Boosting, LR, and SVM models, and they used efficient pre-processing techniques such as label-encoding and normalization to improve the accuracy of the models. They found SVM outperforms others (N. Ahmed et al., 2021).All of the aforementioned research provided a performance comparison of various machine learning algorithms. Some of these employed cross-validation, normalization, and preprocessing techniques to enhance classification performance. To improve their performance, however, no one used the hyperparameter tuning procedure. To increase the model's accuracy, we are concentrating on a grid search hyperparameter tuning strategy in this study.

3. Methodology

3.1 Data Source and Study Variables

The nationally representative BDHS 2017–18 dataset served as the source of the study's data. The survey's execution in Bangladesh was overseen by the National Institute for Population, Research, and Training (NIPORT), Ministry of Health and Family Welfare, which received

funding from USAID. The data set can be downloaded after authorization from the DHS Program website at https://dhsprogram.com/data/availabledatasets.com. A two-stage stratified sampling technique was developed by the BBS for the 2011 population census of the People's Republic of Bangladesh, and it is used by the BDHS to sample the entire country. In the first stage, 675 clusters—425 from rural and 250 from urban strata—were randomly selected. The second stage involved selecting 30 randomly selected families from each of the clusters, and all mothers in those families who had ever been married and were between the ages of 15 and 49 were interviewed. The report of the BDHS 2017–18 contains information about the sampling process in detail.

There are eight separate files, each with its own set of requirements. However, because our primary goal is to predict diabetes status, we only look at the household characteristics file (BDHR7RSV). The variables that are related to this study are described in Table A.1 (appendix), which is found at the end of the study. The numerical form of certain variables, such as blood pressure (normal: systolic<120mmHg & diastolic<80mmHg, prehypertension: 120mmHg ≤systolic ≤139mmHg & 80mmHg ≤diastolic ≤89mmHg, and hypertension: systolic≥140mmHg & diastolic ≥90mmHg), BMI (underweight: BMI < 18.5, normal: 18.5 ≤ BMI < 25, overweight: 25 ≤ BMI < 30, and obese: BMI ≥ 30), and glucose level, is converted into a categorical form before use for simplicity. A plasma blood glucose level of 7 or higher indicates diabetes, while one below 7 is regarded as normal. Chi-square analysis shows that 10 predictor factors show a significant association with diabetes status out of the 21 variables extracted from the BDHS 2017–18 dataset. Employment status, residence, age, wealth index, occupation, consumption of caffeinated drinks, smoking, the highest level of education, blood pressure, and weight status are some of these variables. Our final dataset consists of 850 instances and ten attributes, with 369 positive cases and 481 negative cases, after taking into account the information on glucose levels that is currently available and excluding unexpected observations (such as system missing, not present, refuse, or others).

3.2 Used Algorithms

Decision Tree: Decision Tree algorithm(Ramezankhani et al., 2014) is created in a tree structure, where the algorithm progressively constructs the tree from the root, selecting informative features using equation 1, passing through internal nodes, and ending with outcome-reflected leaf nodes (Marsland, 2015). Our research employs the popular C5.0 algorithm, an updated version of C4.5 by J Ross Quinlan, which extends the ID3 technique(Lantz, 2019).

$$Gain(S,F) = entropy(S) - \sum_{f \in values(F)} \frac{|S_f|}{|S|} entropy(S_f) \quad (1)$$

where $|S_f|$ is a count of the number of members of S that have value f for feature F, S is the set of instances, F is a potential feature out of all conceivable ones, and $entropy(S) = \sum_{i=1}^{c} p_i log_2(p_i)$.The C5.0 technique computes information gain for each feature using equation (1) and then chooses the feature with the highest information gain. After selecting the best feature at each level, the process is then performed recursively on the remaining features.

Neural Network: In general, an artificial neural network (Park & Edington, 2001)is consisting of multiple layers and a network function. These layers include an input layer, a hidden layer, and an output layer. The input layer consists of neurons (nodes) that represent the attribute values associated with the model. The hidden layer receives input from the input neurons and uses an activation function (f_i)to generate output for the output neurons(Komi et al., 2017). Equation 2 determines the output (y_i) of a specific output neuron i.

$$y_i = f_i\left(w_{0i} + \sum_{j=1}^{k} w_{ji}x_{ji}\right) \quad (2)$$

here $x_{ji}$is the input values, w_{ji} represents the weight of the connection from i to j, and w_{0i} stands for the bias node. In this work, the widely used neural network architecture called multilayer per ceptron is employed, utilizing the sigmoid function as the activation function(f_i).

Support Vector Machine: In the SVM algorithm(Sanakal & Jayakumari, 2014), we must find the best hyper-plane that separates all data from one class from data from the other (Savas & Dovis, 2019). With input data set $(x_1, x_2, \dots, x_N)$ and N_S support vectors in the set S, equation 3 represents the decision function of SVM for binary classification tasks.

$$\hat{y}(x_i) = sign\left(\sum_{i=1}^{N_S}[\,\hat{a}_i t_i K(x_j, x_i)] + b_0\right) \quad (3)$$

where t_i stands for the target values, b_0 is a parameter of the ideal hyperplane, a_i is the lagrangian multipliers with $a_i \geq 0$.$K(x_i, x_j)$ denotes the kernel function and we utilize the Gaussian kernel, $K(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) = exp\left(-\gamma\|x_i - x_j\|^2\right)$ where σ denotes the width of the kernel and a suitable value needs to be selected for the σ.

3.3 Tuning Hyperparameter

Machine learning algorithms offer flexibility in adjusting hyperparameters, and more advanced models provide extensive options for customization. The tuning hyperparameters for the selected predictive models in our study are listed in Table 1.Rather than picking arbitrary values for each of the model's parameters, which is not only time-consuming but also unscientific, it is

preferable to run a search across a large number of possible parameter values to identify the optimal combination (Lantz, 2019). Bayesian optimization, Grid search, and Random search can all be used to improve the values of hyperparameters (Probst et al., 2019). In this work, the model's hyperparameters are tuned using the Grid search technique.

Table 1. Tuning Hyperparameters of DT, NN, and SVM

Model	**Hyperparameters**
Decision Tree	model, trials, winnow
Neural Network	Hidden layers and nodes per layer
Support Vector Machine	kernel, C

Grid search: The simplest method for hyperparameter tuning is grid search(Elgeldawi et al., 2021). With this method, we simply create a model for every potential combination of the supplied hyperparameter values, evaluate each model, and choose the design that yields the best results.

3.4 Evaluation Metrics

Table 2's confusion matrix is used to construct the accuracy, sensitivity, specificity, and kappa metrics (DROTÁR & SMÉKAL, 2014) that we utilize in this research to assess the performance of selected classifiers.

Table 2. Confusion Matrix for Two Classes

	Predicted class		
Actual class	Positive	Negative	Total
Positive	True positive (TP)	False negative (FN)	p
Negative	False positive (FP)	True negative (TN)	n
Total	p'	n'	N

Now,

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (4)$$

$$Sensitivity = Recall = \frac{TP}{TP+FN} \quad (5)$$

$$Specificity = \frac{TN}{TN+FP} \quad (6)$$

The Kappa statistic, defined by equation 8, is computed by combining the observed ($observed\ accuracy = \frac{TP+TN}{N}$) and expected($expected\ accuracy = \frac{\frac{p \times p'}{N} + \frac{n \times n'}{N}}{N}$) accuracy.

$$Kappa = \frac{observed\ accuracy - expected\ accuracy}{1 - expected\ accuracy} \quad (7)$$

Receiver Operating Characteristics (ROC) curve: By plotting the false positive rate on the x-axis and the true positive rate on the y-axis, the ROC curve visually evaluates classifier performance. A better classifier is located near the top-left corner, and Area Under the Curve estimates a classifier's overall performance, with 1 being the ideal classifier(Alpaydin, 2014).

3.5 Computational Tools

In this study, data is analyzed using R statistical software (version 4.1.1) and IBM SPSS (version 22), where R is employed for the implementation of machine learning algorithms, while SPSS is used specifically for bivariate analysis.

4. Results and Discussion

After being prepared for this study, the dataset is divided into two groups using random seed (123): train and test dataset. 90% of the extracted data is utilized to construct the training dataset, and the remaining 10% is used to create the test dataset. The training dataset is employed to train the three machine learning algorithms we have chosen: Decision Tree, Neural Network, and Support Vector Machine. After training, the trained models are evaluated on the test dataset using accuracy, sensitivity, specificity, and kappa. The values of these metrics, as determined by equations 4, 5, 6, and 7 are presented in Table 3. According to Table 3, the neural network has the best performance with an accuracy of 67.06%, kappa of 0.3418, sensitivity of 0.6923, and AUC of 0.723, but its specificity (0.6522) is lower than that of the SVM (0.6739). SVM accuracy is 65.88% with a kappa of 0.3143, sensitivity of 0.6410, and AUC of 0.741 (greater than NN), making it the second-best algorithm. With an accuracy of 62.35%, the decision tree is the lower-performing algorithm in our study compared to the other two.

Table 3. Performance Evaluation of DT, NN, and SVM (Without Hyperparameters Tuning)

Algorithms	Accuracy	Kappa	Sensitivity	Specificity	AUC
Decision Tree	62.35%	0.2360	0.5385	0.6957	0.670
Neural Network	67.06%	0.3418	0.6923	0.6522	0.723
SVM	65.88%	0.3143	0.6410	0.6739	0.741

Then, in order to improve the performance of these algorithms, we use hyperparameter tuning, and the hyperparameters of the selected algorithms are optimized using the Grid search method.

Model, winnow, and trails are the three parameters that make up a decision tree. The caret package is used to determine the best values for these parameters, and the findings are listed in Table A.2 (appendix). Based upon the accuracy and kappa value, DT's specified parameters are model = rules, winnow = TRUE, and trails = 20. Table A.3 (appendix) lists the accuracy and kappa values for the different layers of the neural network with sigmoid activation function, each with several hidden nodes. Based on maximum accuracy and kappa, the number of hidden layers for our customized NN is selected as (7, 4), i.e., two layers, one with 7 nodes and the other with 4 nodes. SVM has two hyperparameters Kernel and Cost of constraints violation (C). To provide better classification results, both are chosen based on high accuracy and kappa. Since the kappa and accuracy values for the Gaussian kernel and cost equal to 0.80 are the highest in Table A.4 (appendix), use those values as the parameter values in our modified SVM algorithm. Once again, the customized DT, NN, and SVM are trained using the training dataset, and their performance is assessed using the test dataset. Table 4 summarizes the performance metrics value for these changed classifiers.

Table 4. Performance Evaluation of DT, NN, and SVM (With Hyperparameters Tuning)

Algorithms	Accuracy	Kappa	Sensitivity	Specificity	AUC
Decision Tree	67.06%	0.3341	0.6154	0.7174	0.729
Neural Network	71.76%	0.4336	0.7179	0.7174	0.752
SVM	69.41%	0.3743	0.5641	0.8043	0.747

Our study demonstrates that the neural network algorithm consistently surpassed other algorithms both before and after fine-tuning. After fine-tuning, the accuracy of the NN algorithm enhanced to 71.76%, along with a kappa value of 0.4336, sensitivity of 0.7179, specificity of 0.7174, and an AUC value of 0.752. The support vector machine (SVM) algorithm ranked second in terms of performance, attaining an accuracy of 69.41%. It obtained a kappa value of 0.3743, sensitivity of 0.5641, specificity of 0.8043, and an AUC value of 0.747. Lastly, the decision tree (DT) algorithm displayed the most inferior performance, with an accuracy of 67.06%.

Table 5. Accuracy Improvement Due to Tuning

Algorithms	Accuracy (without tune)	Accuracy (with tune)	Improvement
Decision Tree	62.35%	67.06%	4.71%
Neural Network	67.06%	71.76%	4.70%
SVM	65.88%	69.41%	3.53%

All performance indicators show a considerable improvement following hyperparameter modification, according to a comparison of Table 3 and Table 4 findings. Table 5 lists an

improvement in the classifier's accuracy. The improvement of the DT, NN, and SVM along with their AUC value is depicted visually in Figures 1, Figure 2, and Figure 3 respectively.

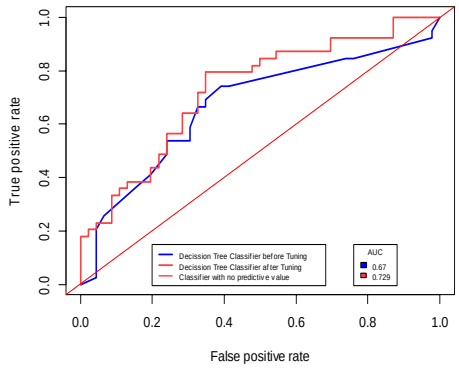

Figure 1: ROC Curve (DT)

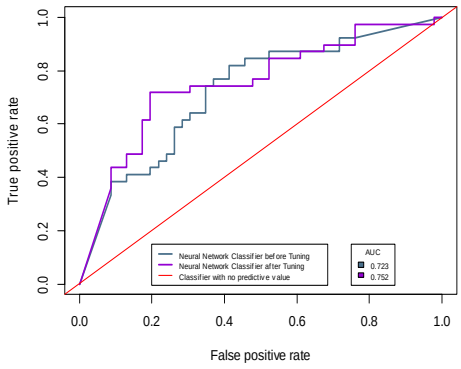

Figure 2: ROC Curve (NN)

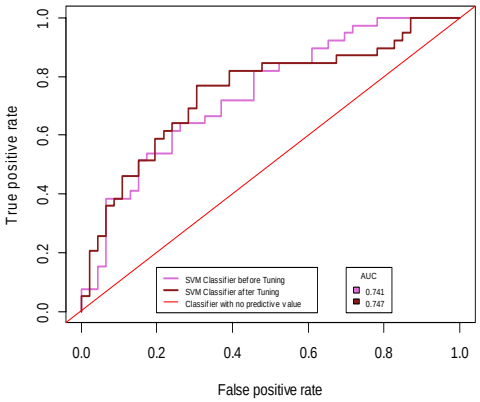

Figure 3: ROC Curve (SVM)

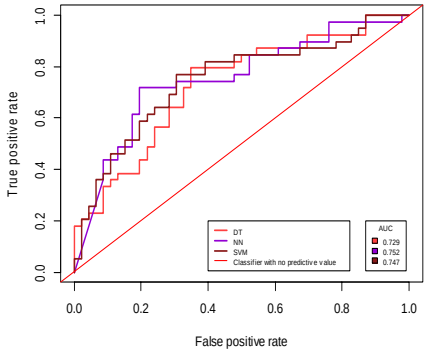

Figure 4: ROC Curve (DT, NN, and SVM)

The results shown in Table 5 show that modifying the hyperparameters considerably improved the performance of DT (4.71%), NN (4.70%), and SVM (3.53%) algorithms in terms of accuracy. Kappa, sensitivity and specificity also indicate the improvement of the algorithms due to tuning. The findings in Table 4 show that neural network outperform SVM and DT classifiers in terms of accuracy, sensitivity, specificity, kappa, and AUC. As a result, this classifier is more accurate than the other two in determining if a person has diabetes. The classification performance of DT, NN, and SVM are represented visually in Figure 4. The neural network has the highest AUC (0.752) value in this image, which implies that it can classify diabetes status more precisely.

5. Conclusion

As earlier mentioned, after adjusting the hyperparameters of selected algorithms, all assessment metrics (accuracy, kappa, sensitivity (except SVM), specificity, and AUC) greatly increased, meaning that classifier classification performance improved significantly. In terms of accuracy, the Decision Tree's classification performance improved by 4.71 percent from 62.23%. Meanwhile, Neural Network and Support Vector Machine grew by 4.70 and 3.53 percent from 67.06% and 65.88%, respectively and the remaining **evaluat**ion metrics (by comparing theirs values utilizing Table 3 and Table 4) also indicate their improvement. Furthermore, with a prediction accuracy of 71.76%, Neural Network outperforms Decision Tree and Support Vector Machine, whereas Decision Tree has 67.06% and Support Vector Machine has 69.41% accuracy. So, to assess diabetes status based on early symptoms of patients without performing diagnostic tests, Neural Network can assist us to minimize cost and time among the selected algorithms with more precision.

In this work, the Grid search technique is employed to improve hyperparameters throughout the tuning phase. In the future, one can apply random search or Bayesian optimization tuning processes to improve the classification performance of these algorithms. One also can use ensemble algorithms to get better performance.

Limitations of the study

This study focused only on machine learning algorithms' prediction performance and improving their performance through hyperparameter tuning, hence no emphasis is given to calculating summary statistics of the dataset. Furthermore, during the NN tuning process, only a sigmoid activation function is used, though there is the availability of other kinds of activation functions such as hyperbolic tangent and rectified linear unit.

Acknowledgements

The authors would like to express the most profound gratitude to Dr. Rumana Rois and Dr. Mohammad Alamgir Kabir for their suggestions and assistance.

References

Ahmed, N., Ahammed, R., Islam, M., Uddin, A., Akhter, A., Talukder, A., & Kumar, B. (2021). Machine learning based diabetes prediction and development of smart web application. *International Journal of Cognitive Computing in Engineering*, *2*(December), 229–241. https://doi.org/10.1016/j.ijcce.2021.12.001

Ahmed, S., Ahmed, T., Sharmin, T., Mohammad, S., & Quddus, R. (2017). Impact of type 2

Diabetes Mellitus for developing severe health complications in Bangladeshi population. *Asian Journal of Medical and Biological Research*, *3*(2), 152–157. https://doi.org/10.3329/ajmbr.v3i2.33562

Alehegn, M., Joshi, R., & Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. *International Research Journal of Engineering and Technology*, *4*(10), 426–436. www.irjet.net

Alpaydin, E. (2014). *Introduction to Machine Learning* (3rd ed.). The MIT Press.

Choudhury, A., & Gupta, D. (2019). A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques. *Advances in Intelligent Systems and Computing*, *740*. https://doi.org/10.1007/978-981-13-1280-9_6

DROTÁR, P., & SMÉKAL, Z. (2014). Comparative Study of Machine Learning Techniques for Supervised Classification of Biomedical Data. *Acta Electrotechnica et Informatica*, *14*(3), 5–10. https://doi.org/10.15546/aeei-2014-0021

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. *Informatics*, *8*(4), 1–21. https://doi.org/10.3390/informatics8040079

Islam, M., Rahman, J., & Chandra, D. (2020). Automated detection and classification of diabetes disease based on Bangladesh demography and health survey data , 2011 using machine learning approach. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(3), 217–219. https://doi.org/10.1016/j.dsx.2020.03.004

Kandhasamy, J. P., & Balamurali, S. (2015). Performance analysis of classifier models to predict diabetes mellitus. *Procedia Computer Science*, *47*(C), 45–51. https://doi.org/10.1016/j.procs.2015.03.182

Kaur, H., & Kumari, V. (2022). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*, *18*(1–2), 90–100. https://doi.org/10.1016/j.aci.2018.12.004

Khan, M. H., Krämer, A., Khandoker, A., Prüfer-krämer, L., & Islam, A. (2011). *Trends in sociodemographic and health-related indicators in Bangladesh , 1993 – 2007 : will inequities persist ?February*, 583–592. https://doi.org/10.2471/BLT.11.087429

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, *7*(4), 432–439. https://doi.org/10.1016/j.icte.2021.02.004

Komi, M., Li, J., Zhang, X., & Xianguo, Z. (2017). Application of Data Mining Methods in Diabetes Prediction Messan. *In 2017 2nd International Conference on Image, Vision and Computing Application*, 8 (8), 1006–1010. https://doi.org/10.1109/ICIVC.2017.7984706

Lantz, B. (2019). *Machine Learning with R* (3rd ed.). Packt Publishing.

Marsland, S. (2015). *Machine Learning: An Algorithmic Perspective* (2nd ed.). CRC Press.

Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Aishwarya Mujumdar Diabetes Prediction using Machine Learning Aishwarya Mujumdar Aishwarya. *Procedia Computer Science*, *165*, 292–299. https://doi.org/10.1016/j.procs.2020.01.047

Nai-arun, N., & Moungmai, R. (2015). Comparison of Classifiers for the Risk of Diabetes Prediction. *Procedia - Procedia Computer Science*, *69*, 132–142. https://doi.org/10.1016/j.procs.2015.10.014

Park, J., & Edington, D. W. (2001). A sequential neural network model for diabetes prediction. *Artificial Intelligence in Medicine*, *23*(3), 277–293. https://doi.org/10.1016/S0933-3657(01)00086-0

Pranto, B., Mehnaz, S. M., Mahid, E. B., & Mahmud, I. (2020). Evaluating Machine Learning Methods for Predicting Diabetes among Female Patients in Bangladesh. *Information*. https://doi.org/10.3390/info11080374

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), 1–15. https://doi.org/10.1002/widm.1301

Ramezankhani, A., Pournik, O., Shahrabi, J., Khalili, D., Azizi, F., & Hadaegh, F. (2014). Applying decision tree for identification of a low risk population for type 2 diabetes: Tehran Lipid and Glucose Study. *Diabetes Research and Clinical Practice*, *105*(3), 391–398. https://doi.org/10.1016/j.diabres.2014.07.003

Sanakal, R., & Jayakumari, S. T. (2014). Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine. *International Journal of Computer Trends and Technology*, *11*(2), 94–98. https://doi.org/10.14445/22312803/ijctt-v11p120

Savas, C., & Dovis, F. (2019). The impact of different kernel functions on the performance of scintillation detection based on support vector machines. *Sensors (Switzerland)*, *19*(23), 1–16. https://doi.org/10.3390/s19235219

Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0175-6

Thirumal, P. C., & Nagarajan, N. (2015). Utilization of data mining techniques for diagnosis of diabetes mellitus - A case study. *ARPN Journal of Engineering and Applied Sciences*, *10*(1), 8–13.

Vijayan, V. V., & Anjali, C. (2016). Prediction and diagnosis of diabetes mellitus - A machine learning approach. *2015 IEEE Recent Advances in Intelligent Computational Systems, RAICS 2015*, *December*, 122–127. https://doi.org/10.1109/RAICS.2015.7488400