Diabetes Prediction – Machine Learning (WebAPP)

TEAM MEMBERS:

Aditya Nerusu

Sai Krishna Vaibhav

Nihaal Gooty

Korrapati Sai Sujan

Manikanta Rayani

Chetan Krishna

Sanjay

*Abstract*— The objective of this project is to predict a person's risk of having diabetes by utilizing Support Vector Machine (SVM) algorithms in an intuitive web application interface. This application attempts to provide accurate and reasonable predictions by using input health parameters (number of pregnancies, blood pressure, glucose level, insulin level, age, skin thickness, diabetes pedigree function, etc.) that users provide via a graphical user interface (GUI). By combining the power of SVM with user-friendly web technology, the project endeavors to enhance accessibility to predictive healthcare tools. The seamless integration of machine learning into a web application facilitates a simple and effective method for diabetes prediction, which could aid people in making accurate choices regarding their health. By promoting preventive measures and giving people early awareness, this initiative hopes to support proactive healthcare.

Keywords— Diabetes Prediction, Machine Learning, Support Vector Machine, Graphical User Interface, Web Application using Streamlit, Health Sector.

# Introduction

Globally, the prevalence of diabetes, a chronic metabolic disease, is steadily increasing and presents serious health risks. Diabetes arises from various factors including age, sedentary habits, familial predisposition, hypertension, psychological factors like depression and stress, and unhealthy dietary choices. Diabetes puts a person at risk of heart disease, kidney disease, stroke, eye problems, blood vessel damage, nerve damage, and other conditions making the body incapable of producing insulin. Proactive management and early detection are essential to reducing its negative effects on people's health.

Symptoms of Diabetes are:

* Blurred Vision
* Unexplained Weight Loss
* Mood Swings
* Recurring Infections
* Increased Thirst
* Frequent Urination
* Increased Thirst and Hunger
* Slowly Healing Wounds
* Fatigue
* Tiredness or Sleepiness

The International Diabetes Federation reports [9] that globally, 382 million individuals are afflicted with diabetes, with projections indicating a rise to 592 million by 2035. Each day, numerous individuals are affected by this condition, with many unaware of their status. It predominantly impacts individuals aged between 25 and 74 years. Failure to detect and treat diabetes can result in a range of complications.

One of the most significant aspects of artificial intelligence is machine learning, which enables the development of computing devices with the capability to learn from past experiences without requiring programming in each instance. It is believed that machine learning is an immediate necessity for the present scenario of events to enable automation with the least amount of possible flaws in order to eliminate human work. Present-day laboratory tests like oral glucose tolerance and fasting blood glucose are used to detect diabetes. Yet, this process takes a lot of time.

As a result, this project presents a novel method of predicting diabetes by utilizing machine learning, specifically Support Vector Machine (SVM) algorithms with four types of kernels: polynomial, sigmoid, RBF, and linear. The model is trained using data from both diabetic and nondiabetic instances (PIMA Indian Dataset) and is integrated into an easy-to-use web application interface using the Streamlit library in Python. The combination of machine learning algorithms and user interface (GUI) allows people to simply enter their information and receive personalized predictions about whether or not they have diabetes. The application's incorporation of Support Vector Machines (SVM) facilitates the examination of data submitted by users, including medical history, lifestyle choices, and demographic data. Through empowering people to take control of their health, this initiative seeks to close the gap between technology and healthcare and promote a proactive approach.

# LITERATURE REVIEW

Numerous studies have been conducted to automatically predict diabetes through the use of ensemble and machine learning techniques. The majority of these projects used the publicly available Pima Indian dataset [1]. The following article briefly discusses some of these papers on automatic diabetes prediction using the PIMA Indian dataset.

A smart home health monitoring system was developed by Goyal and colleagues [4] for tracking diabetes. The researchers also sued PIMA Native American records during their study. They used SVM, KNN, and decision trees to predict diabetes and decision-based decision-making to predict blood pressure status. In comparison, SVM produced better results, with an accuracy rate of 75%.

Mohan and Jain adopted the SVM algorithm in a recent paper [3] to examine and identify diabetes using the Pima Indian Diabetes Dataset. This work used four different types of kernels: polynomial, sigmoid, RBF, and linear, to identify diabetes in the machine learning platform. Between 0.69 and 0.82, the authors' accuracies varied depending on the kernel. The maximum accuracy of 0.82 was achieved by the SVM method using a radial basis kernel function.

A study by Kumar [2] created a system that can rapidly and accurately predict diabetes employing the random forest algorithm. Initially, the authors employed standard preprocessing methods for data, such as reduction, integration, and cleaning. The accuracy level was 90% when using the random forest algorithm, which is much higher than other algorithms.

Kim [6] proposed an early prediction system for type 2 diabetes based on machine learning. The authors used a private dataset with over 253,000 volunteer data points from a nearby Korean hospital for six years. The data imbalance issue is addressed by using algorithms for synthetic oversampling, SMOTE, and undersampling. Several machine learning techniques are applied. In the end, both the SVM and random forest classifiers had the best results.

Jackins [5] used machine-learning methods and the Pima Indian dataset to develop a diabetes prediction model. The authors claim that with accuracy increments of 0.43%, the Naive Bayes method outperformed the random forest technique.

Birjais in his study [10] conducted experiments on the PIMA Indian Diabetes (PID) dataset, comprising 768 instances and 8 attributes, accessible via the UCI machine learning repository. Their focus was on diabetes diagnosis, recognized by the World Health Organization (WHO) in 2014 as one of the fastest-growing chronic diseases worldwide. The study employed gradient boosting, logistic regression, and naive Bayes classifiers to predict diabetes occurrence, achieving accuracies of 86%, 79%, and 77% respectively.

To create an automatic diabetes prediction system, Pranto [7] used a private dataset from a hospital located nearby in Bangladesh along with Pima Indian. Using the Pima Indian dataset, multiple machine-learning techniques were trained in this work. On the private dataset, K-Nearest Neighbor and decision tree models yielded 81.2% and 79.2% accuracy, respectively.

This study by Nazin [8] focuses on developing effective machine-learning classifiers to detect diabetes using clinical data. Various algorithms including Decision Tree, Naive Bayes, k-nearest neighbor, Random Forest, Gradient Boosting, Logistic Regression, and Support Vector Machine are trained and evaluated. Pre-processing techniques such as label encoding and normalization are employed to enhance model accuracy. Feature selection methods are applied to identify significant risk factors. The models are tested on multiple datasets, outperforming previous studies by 2.71% to 13.13% depending on the dataset and algorithm. The most accurate algorithm is selected for further development, and the model is integrated into a web application using Python Flask. Overall, the findings demonstrate the potential of preprocessing and machine learning classification in accurately predicting diabetes from clinical data.

Sadhu, A. and Jadli A. conducted experiments using a diabetes dataset sourced from the UCI repository, containing 520 occurrences and 16 attributes in their study [11]. Their focus was on early-stage diabetes prediction. Seven classification techniques, including k-NN, logistic regression, SVM, naive Bayes, decision tree, random forests, and multilayer perceptron, were applied to the validation set of the dataset. Results indicated that the random forests classifier performed the best, achieving an accuracy score of 98%. This was followed by logistic regression at 93%, SVM at 94%, naive Bayes at 91%, decision tree at 94%, random forests at 98%, and multilayer perceptron at 98%, based on training multiple machine learning models.

Xue [12] conducted experiments on a diabetes dataset from the UCI repository, comprising 520 patients and 17 attributes. Their focus was on early diabetes detection, training on data from 520 diabetic and potential diabetic patients aged 16–90 using supervised machine learning techniques including SVM, naive Bayes classifiers, and LightGBM. SVM exhibited superior performance in terms of classification and recognition accuracy, achieving the highest accuracy rate of 96.54%. The naive Bayes classifier, the most commonly used algorithm, attained an accuracy of 93.27%, while LightGBM demonstrated lower accuracy at 88.46%. These findings suggest that SVM is the optimal classification algorithm for diabetes prediction.

Agarwal [13] also utilized the PID dataset consisting of 738 patients in their research. To assess the dataset's efficacy in identifying diabetic patients, the authors employed various models including SVM, k-NN, NBC, ID3, C4.5, and CART. Among these, SVM and LDA algorithms emerged as the most accurate, achieving an accuracy rate of 88%.

In their study, Saravananathan and Velmurugan [14] evaluated the performance of J48, CART, SVM, and k-NN algorithms on a medical dataset. They conducted comparisons based on accuracy, specificity, sensitivity, precision, and error rate. Their findings revealed that J48 algorithms exhibited the highest accuracy at 67.15%, followed by SVM at 65.04%, CART at 62.28%, and k-NN at 53.39%.

Kavakiotis's study [15] utilized NBC, RFC, k-NN, SVM, DT, and LR methods to forecast diabetes. These algorithms were implemented utilizing a ten-fold cross-validation technique. The study found that SVM exhibited the highest accuracy among all approaches, achieving a score of 84%.

Rawat [16] conducted research on the classification of "Diabetes Prediction" based on eight attributes. The study introduced five ML algorithms—AdaBoost, LogicBoost, RobustBoost, naive Bayes, and bagging—for analyzing and predicting diabetic patients. The strategies were tested on a group of diabetic PIMA Indians, yielding highly accurate results with classification accuracies of 81.77% and 79.69% for the bagging and AdaBoost techniques, respectively. Consequently, the proposed DM prediction algorithms were deemed notably attractive, effective, and efficient.

Rathore utilized classification techniques such as SVM and DTs for predicting diabetes mellitus, utilizing the PID dataset for their analysis. The study focused on women's health, particularly in the context of PIMA India. SVM achieved an accuracy rate of 82% in this prediction task. [17]

In [18] Hassan made us of classification methods including DT, k-NN, and SVM to predict diabetes mellitus. Among these approaches, SVM demonstrated superior performance compared to DT and KNN, achieving a maximum accuracy of 90.23%.

Nai-Arun and Moungmai [19] designed a web application centered on diabetes prediction accuracy. They assessed various prediction methods such as DTs, NNs, LR, NBC, RFC, bagging, and boosting. Their investigation highlighted RFC as the top performer in terms of accuracy and ROC score, achieving an accuracy level of 85.558% and an ROC value of 0.912.

In this study [20] Jyoti aimed to develop a system capable of performing early prediction of diabetes for patients with higher accuracy by combining the results of different machine learning techniques. Various algorithms including K-nearest neighbor, Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree were utilized. The accuracy of each algorithm was evaluated, and the one demonstrating superior accuracy was selected as the model for diabetes prediction. Experiments were conducted on the John Diabetes Database, with results indicating the adequacy of the designed system, achieving an impressive accuracy of 99% using the Decision Tree algorithm.

Branimir in his study [21] introduced a system to address two primary challenges: the heterogeneity observed in previous techniques and the lack of transparency in features. Employing the PRISMA methodology, the study conducted comparisons among 18 different models, focusing on algorithms based on trees. The findings highlighted KNN and SVM as the primary choices for prediction tasks.

# RESEARCH GAP AND OBJECTIVES

The research exploring diabetes prediction through machine learning tools has shown significant promise, yet a considerable gap exists in accessible predictive healthcare tools. Although the model mentioned above works well, some issues need to be addressed.

There is a lack of readily available tools for diabetes prediction, despite advances in machine learning for healthcare. Widespread adoption may be impeded by the lack of user-friendly interfaces in many of the current models.

Research involving the PIMA Indian Dataset has been the focus of most researchers, potentially leading to differences. It's possible that the model's applicability to a broader range of demographics than just the Pima Indian community was limited by the dataset's lack of diversity in population representation.

The model may lose its efficacy over time if fresh data are not introduced to it regularly or if it is not adjusted to reflect the evolving health trends.

To address these gaps and limitations, the objectives are twofold:

To enhance accessibility through a user-friendly interface, ensuring easy input of health parameters for diabetes prediction, and improving model generalizability by exploring methods that account for diverse populations.

Furthermore, implementing strategies for continuous learning and updates to the model, along with fostering user engagement and feedback mechanisms, aims to enhance the web application's usability and efficacy in predicting diabetes across diverse demographics, thus filling the existing gaps and mitigating inherent limitations.

# METHODOLOGY

Numerous machine learning algorithms have been developed, including Naive Bayes, Decision Trees, Linear Regression, K Nearest Neighbors, Random Forest, Support Vector Machines, and Logistic Regression. In this paper, we employ support vector machines (SVM) with four different kernel types: sigmoid, polynomial, RBF, and linear to identify diabetes and assess each case's accuracy. Our project's sole advantage is that it features a web interface that collects users' medical information to accurately predict whether the user is diabetic. The methodology for each step is as follows:

## Importing libraries and Data Collection

Library Imports: Utilize Python libraries like Pandas for data alteration, NumPy for computational operations, and Scikit-learn for machine learning tools.

Dataset Loading: Access the PIMA Indian Diabetes dataset from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) website or repository and load it into a data frame using the imported Pandas library.

The dataset consists of features like number of pregnancies, blood pressure, age, glucose level, diabetes pedigree function, insulin level, skin thickness, etc. These attributes serve as the foundation for predicting whether the user is diabetic or not.

## Data Preprocessing and Standardizing

Data Cleaning: Handle missing values, outliers, and inconsistencies within the dataset.

Feature Standardizing: Normalize or scale features to ensure all have a similar impact during modeling.

## Data Splitting

Using an 80:20 ratio, split the pre-processed dataset into training and testing sets. Model training will take place on the training set (80%), and model performance evaluation will take place on the testing set (20%).

## Training Predictive Model

Machine Learning models are trained using Support Vector Machines (SVM). Support Vector Machine (SVM) is a powerful supervised learning algorithm capable of handling both non-linear and linear data used for both regression and classification tasks. It works by figuring out which hyperplane divides the classes in a dataset the best.

In a binary classification scenario, SVM aims to find a hyperplane that maximizes the margin between two classes (either diabetic or non-diabetic), effectively creating a linear separator. It aims to classify data points by their position relative to this hyperplane.

SVM can utilize various kernels to handle complex datasets that are not linearly separable in their original feature space. Here are the four SVM kernels that have been used in our study. The hyperparameters governing the regularization parameter (C), the kernel type, and the kernel coefficient were systematically optimized to maximize the model's performance.

**Linear Kernel:** The linear kernel calculates the dot product between data points, suitable for datasets where classes can be distinctly divided by a straight line or plane. It operates efficiently for large datasets and tends to be less susceptible to overfitting due to its simplicity.

**Polynomial Kernel:** The polynomial kernel transforms data into higher dimensions using polynomial functions. This kernel is beneficial when datasets demand more complex boundaries beyond linear separation. By introducing non-linearity through higher dimensions, it accommodates more intricate relationships between data points.

**RBF Kernel:** The Radial Basis Function (RBF) kernel, a highly adaptable choice, measures the similarity of data points to landmarks in a higher-dimensional space. Widely utilized due to its versatility, the RBF kernel excels in capturing complex relationships within datasets that lack easily definable boundaries, offering a robust solution across various data complexities.

**Sigmoid Kernel:** The sigmoid kernel employs hyperbolic tangent functions to map features into higher dimensions. While less computationally intensive compared to other kernels, it might be more sensitive to feature scaling. It serves as an alternative for specific datasets where other kernels might not perform optimally.

Ultimately, the accuracy is computed for each of the four implementations.

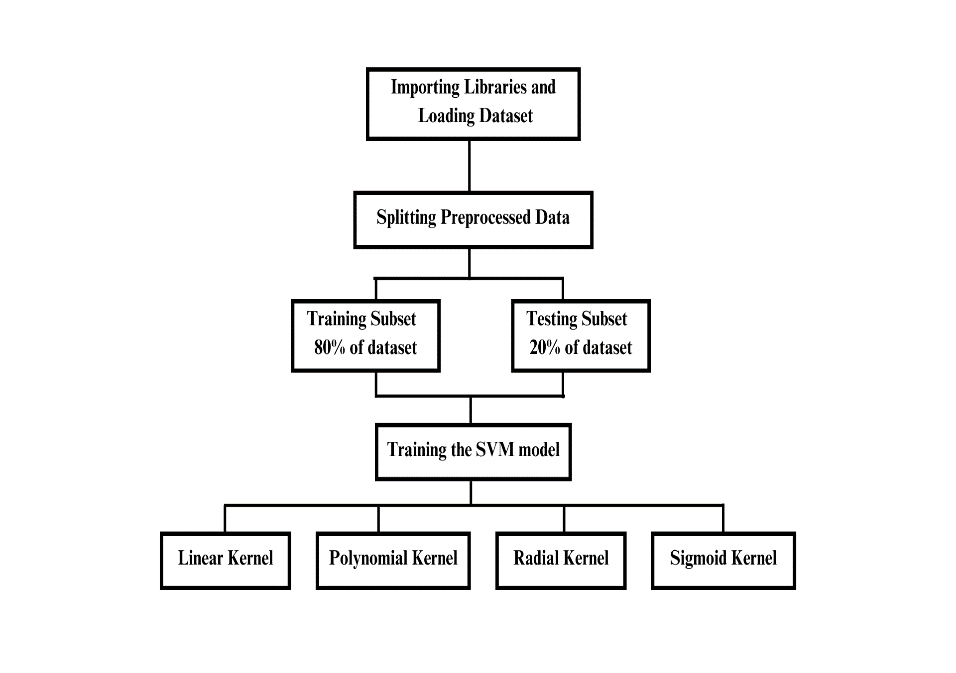
## Web Application – Streamlit Library

By importing the pickle library from Python, we load our trained SVM models in binary format and use the Streamlit library to create an intuitive web interface.

Streamlit is a popular open-source Python library specifically designed to simplify and accelerate the process of creating dynamic web applications for machine learning and data projects. It enables developers and data scientists to build user-friendly interfaces with minimal code, allowing for rapid prototyping and deployment of data-driven applications.

Incorporate input fields in the web application to collect user medical data, passing it through the loaded SVM model will display the predicted diabetic status of the user.

This methodology combines data collection, preprocessing, model training, evaluation, and deployment into a web application using Python libraries like Scikit-learn for machine learning, Pickle for model serialization, and Streamlit for building the user interface.



# DISCUSSION AND RESULTS

The Pima Indian Diabetes dataset typically includes information about various health parameters for individuals, such as glucose levels, blood pressure, BMI, age, and whether the person is diagnosed as diabetic or not. The classification of whether a person is diabetic or not is often represented as a binary outcome in the dataset, with values like 0 (indicating non-diabetic) and 1 (indicating diabetic).

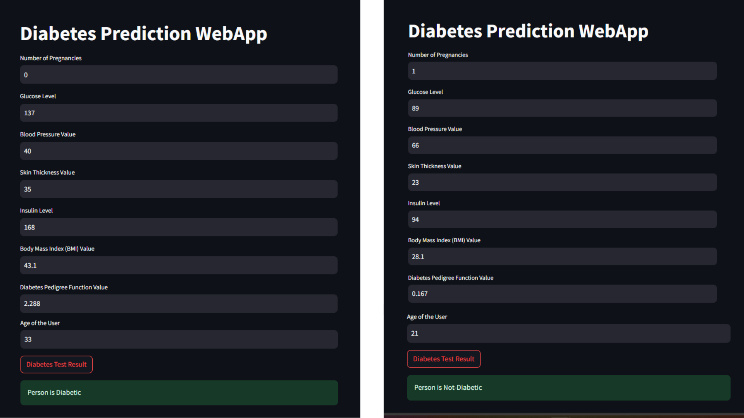
The table below illustrates a trend that reveals the standard range of parameters that indicate the probability that a user has diabetes based on the values of the 789 instances of the dataset that are present. A user is more likely to have diabetes if their records are higher than the standard values, as shown in the table below.

|  |  |
| --- | --- |
| Features | Standard Values |
| Number of Pregnancies | 4 |
| Glucose Level | 120.89 |
| Blood Pressure | 69.10 |
| Skin Thickness | 20.53 |
| Insulin Level | 79.79 |
| BMI | 31.99 |
| Diabetes Pedigree Function | 0.47 |
| Age | 34 |

It's important to note that determining whether a person is diabetic or not solely based on ranges of individual health parameters (like glucose levels, blood pressure, etc.) is not straightforward and might vary based on multiple factors. Healthcare professionals typically consider a combination of factors and conduct specific tests to diagnose diabetes accurately.

The following are the accuracies in each case that were obtained after training our SVM model using the dataset involving the four kernels:

|  |  |  |
| --- | --- | --- |
| Kernel Type | Train Accuracy | Test Accuracy |
| Linear Kernel | 0.7654 | 0.8246 |
| Polynomial Kernel | 0.7850 | 0.7792 |
| RBF Kernel | 0.8469 | 0.7922 |
| Sigmoid Kernel | 0.6840 | 0.7402 |



Considering these results, the RBF Kernel emerges as the most suitable choice for the model. Despite a slightly lower test accuracy compared to the Linear Kernel, its robust performance on both training and test datasets signifies a good balance between complexity capture and generalization. The RBF Kernel's adaptability to diverse data patterns, coupled with its competitive accuracy metrics, positions it as the optimal kernel choice for the diabetes prediction model on the PIMA Indian Diabetes dataset.

This model, equipped with the RBF Kernel, offers a reliable predictive tool for users seeking to assess their risk of diabetes. Streamlit's interactive features will enable easy integration of the model, allowing users to receive predictions regarding their diabetic status based on their input parameters, fostering proactive health management and informed decision-making. Deploying this model through Streamlit will ensure accessibility, usability, and accuracy, empowering users to make informed health-related choices based on predictive insights.

The application now seamlessly incorporates patient details, providing precise predictions on their diabetic status. Users can input their information effortlessly, receiving instant insights into whether they have diabetes or not. With this accessible and accurate tool, individuals can take proactive steps towards managing their health, fostering informed decision-making and promoting a healthier lifestyle.

# CONCLUSION AND FUTURE WORK

In conclusion, the development and evaluation of the diabetes prediction model utilizing Support Vector Machine (SVM) kernels on the PIMA Indian Diabetes dataset have yielded valuable insights. Among the kernels tested, the RBF Kernel emerged as the optimal choice, demonstrating robust predictive performance with a training accuracy of 84.69% and a test accuracy of 79.22%.

This model, poised to be deployed within a Streamlit-based web application, stands as a valuable tool for individuals seeking proactive health management. The integration of Streamlit's user-friendly interface with the RBF Kernel-equipped model offers a user-centric approach, allowing individuals to input their medical details and receive predictions regarding their diabetic status, fostering informed decision-making and proactive health measures.

Moving forward, there are several avenues for further exploration and enhancement of this project:

Firstly, expanding the dataset to encompass a more diverse demographic would improve the model's generalizability.

Additionally, fine-tuning the model parameters and exploring ensemble methods might further boost predictive accuracy.

Incorporating real-time data updates and continuous model learning mechanisms would ensure the model's relevance and adaptability to evolving health trends.

Furthermore, enhancing the web application's functionalities by integrating features such as personalized health recommendations based on predictions and incorporating additional health parameters could provide a more comprehensive health assessment for users.

Collaborations with healthcare professionals to validate the model's predictions and ensure alignment with clinical diagnoses would bolster its reliability and applicability in real-world healthcare settings.

Overall, the project lays a solid foundation for predictive healthcare tools, and future endeavors could focus on refinement, scalability, and increased accuracy to better serve individuals in proactive health management.

Top of Form

##### References

1. <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database> This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.
2. Vijiya Kumar, K., Lavanya, B., Nirmala, I., Caroline, S.S.: Random forest algorithm for the prediction of diabetes. In: International Conference on System, Computation, Automation and Networking, pp. 1–5 (2019).
3. Mohan, N., Jain, V.: Performance analysis of support vector machine in diabetes prediction. In: *International Conference on Electronics*, Communication and Aerospace Technology, pp. 1–3 (2020)
4. Chatrati, S.P., Hossain, G., Goyal, A., et al.: Smart home health monitoring system for predicting type 2 diabetes and hypertension. *J. King Saud Univ. Comput. Inf. Sci.* 34(3), 862–870 (2020)
5. Jackins, V., Vimal, S., Kaliappan, M., Lee, M.Y.: AI‐based smart prediction of clinical disease using random forest classifier and Naive Bayes. *J. Supercomput.* 77, 5198–5219 (2021).
6. Deberneh, H.M., Kim, I.: Prediction of type 2 diabetes based on a machine learning algorithm. *Int. J. Environ. Res. Public Health* 18, 1–14 (2021)
7. Pranto, B., et al.: Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. *Information* 11, 1–20 (2020).
8. Nazin Ahmed, Rayhan Ahammed, Md. Manowarul Islam, Md. Ashraf Uddin, Arnisha Akhter, Md. Alamin Talukder, Bikash Kumar, Machine learning based diabetes prediction and development of smart web application, International Journal of Cognitive Computing in Engineering 2 (2021)
9. Firdous, Shimoo1; Wagai, Gowher A.2, Sharma, Kalpana1. A survey on diabetes risk prediction using machine learning approaches. Journal of Family Medicine and Primary Care, November 2022.
10. R Birjais, AK Mourya, Chauhan R, Kaur H Prediction and diagnosis of future diabetes risk - A machine learning approach. SN Appl Sci 2019
11. Sadhu A, Jadli A Early-stage diabetes risk prediction - A comparative analysis of classification algorithms. Int Adv Res J Sci Eng Technol (IARJSET).
12. Xue J, Min F, Ma F Research on diabetes prediction method based on machine learning. J Phys Conf Ser 2020.
13. Agrawal P, Dewangan AK A brief survey on the techniques used for the diagnosis of diabetes-mellitus. Int Res J Eng Tech IRJET 2015.
14. Saravananathan K, Velmurugan T Analyzing diabetic data using classification algorithms in data mining. Indian J Sci Technol 2016
15. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017.
16. Rawat V, Suryakant S A classification system for diabetic patients with machine learning techniques. Int J Math Eng Manag Sci 2019.
17. Rathore A, Chauhan S, Gujral S Detecting and predicting diabetes using supervised learning:An approach towards better healthcare for women. Int J Adv Res Comput Sci 2017
18. Hassan AS, Malaserene I, Leema AA Diabetes mellitus prediction using classification techniques. Int J Innov Technol Explor Eng 2020
19. Nai-Arun N, Moungmai R Comparison of classifiers for the risk of diabetes prediction. Procedia Comput Sci 2015
20. KM Jyoti Rani, Diabetes Prediction Using Machine Learning, Department of CSE, IMG Engineering College, Ghaziabad, Uttar Pradesh, July-August-2020.
21. B. Ljubic, A. A. Hai, M. Stanojevic et al., “Predicting complications of diabetes mellitus using advanced machine learning algorithms,” Journal of the American Medical Informatics Association 2020.