Diabetes Prediction Using Different Kernels of SVM Classification Algorithm

[1] Harasis Singh,[2] Richa Sharma

[1] B.Tech Research Scholar,[2] Assistant Professor,
[1,2] Department of Computer Science , Jaipur Engineering College and Research Centre ,Jaipur ,Rajasthan.

Abstract : Diabetes increases the danger of heart condition by about fourfold in women but only around twice in men. After a heart attack, the outcomes are severe in women too. Women also are at higher risk of other diabetes-related complications like blindness, renal disorder, and depression. The modeling of support vector machines may also be a promising classification method for identifying people within the population with common diseases such as diabetes and pre-diabetes. We illustrate different SVM techniques to detect women with diabetes based on the sample of the women population. For the classification of patients with diabetes and without diabetes based on the set of diabetes-related variables, Compared to other kernels used for SVM, the RBF kernel SVM algorithm can predict the chances of diabetes with 83 percent accuracy. To demonstrate a user-friendly and platform-independent application that allows for person or community assessment with a configurable, user-defined threshold, the Docker-based web tool for Diabetes Classifier was developed. Using common variables, this method can be further explored and updated for other complex diseases.

***IndexTerms* – Diabetes Prediction , Machine Learning , SVM .**

I. Introduction

“Diabetes may be a chronic, metabolic disease characterized by elevated levels of blood sugar (or blood sugar), which leads over time to serious damage to the guts, blood vessels, eyes, kidneys, and nerves. the foremost common is type 2 diabetes, usually in adults, which occurs when the body becomes immune to insulin or doesn't make enough insulin. within the past three decades, the prevalence of type 2 diabetes has risen dramatically in countries of all income levels. Type 1 diabetes, once referred to as type I diabetes or insulin-dependent diabetes, maybe a chronic condition during which the pancreas produces little or no insulin by itself. For people living with diabetes, access to affordable treatment, including insulin, is critical to their survival. there's a globally agreed target to halt the increase in diabetes and obesity by 2025” [1].

“About 422 million people worldwide have diabetes, the bulk living in low-and middle-income countries, and 1.6 million deaths are directly attributed to diabetes annually. Both the number of cases and therefore the prevalence of diabetes are steadily increasing over the past few decades” [1].

“Virtual and Physical are the 2 main branches of the appliance of AI. The virtual component is represented by Machine Learning, (also called Deep Learning) that's represented by mathematical algorithms that improve learning through experience. There are three kinds of machine learning algorithms: (i) unsupervised (ability to hunt out patterns), (ii) supervised (classification and prediction algorithms supported previous examples), and (iii) reinforcement learning (use of sequences of rewards and punishments to form a way for operation during a selected problem space). First, AI has boosted and remains to boost discoveries in genetics and molecular medicine by providing machine learning algorithms and knowledge management” [3]. “An example of successes in medicine is the unsupervised protein–protein interaction algorithms that led to novel therapeutic target discoveries” [2].

To diagnostically predict whether or not a patient has diabetes, we can use machine learning algorithms, assisted by certain diagnostic measures included in the dataset. This research is based on using the SVM algorithm with its different kernels of machine learning to separate women without any of these conditions from women with either undiagnosed diabetes or pre-diabetes.

Research problem formulation: The analysis of expectations is the technique that can forecast future results from the current data. They should work out when women have diabetes to ensure that they keep glucose, blood pressure, insulin, body mass index (BMI) levels within their goal range targets, not too high, not too low. That means finding out when and what to eat for snacks and meals

This paper is structured as follows: The formulation of the research problem is defined in Section II. The related work of various classification techniques for the prediction of diabetes is outlined in section III. Section IV briefly reviews some basic SVM principles and thus the selection of the kernel function with precision measurements. In Section V, the experimental results are given. Finally, Section VI ends the paper with information on the potential scope being given by Section VII.

II. Literature Review

Bamnote, M.P., G.R., (2014) Using the Diabetes dataset from the UCI repository, Genetic Programming (GP) was used to train and evaluate the database for predicting diabetes. Results obtained using GP have the optimum precision relative to other techniques introduced. By taking less time for the classifier generation, there is also a major increase in accuracy. It turns out to be a useful model for forecasting

Aiswarya Iyer, et.al (2015) As a real-world body, it determines the diagnosis of diabetes. Diabetes diagnosis at an early stage is the key point of this. For the diagnosis of diabetes, the decision tree and naïve Bayes techniques are used this. And at the end of the day, the model proposed gives the safest and most successful outcome.

Orabi et al. (2016) a diabetes prediction system has been developed, with the main objective of predicting a candidate's diabetes at a particular age. The proposed framework is constructed using a decision tree algorithm, based on the notion of machine learning.

The developed method works well at a selected age in predicting diabetes events, using the Decision tree algorithm with greater precision.

Sajida et al.(2016)discuss the role of ensemble machine learning methods such as Ada-boost and Bagging using the decision tree since diabetes risk factors were supported by the basis for classifying diabetes and patients as diabetic or non-diabetic. Results obtained after the experiment prove that the technique of Ada-boost outperforms the techniques of bagging and decision tree.

P. Suresh Kumar, et.al (2017)a model was proposed to solve all problems such as clustering and classifications from the current framework by applying the Data Mining technique. This approach is to diagnose the type of diabetes from patient data obtained. For the purpose of the investigation, all the data obtained from the 650 patients was included in the paper and its results were described. To cluster the entire dataset, the K-means algorithm is used. It is split into three datasets, such as gestational diabetes cluster-0, type 1 diabetes cluster-1 and cluster-2 for type-2 diabetes, a former clustered dataset was used in the classification model as an input for the classification process, such as the patient's risk levels of diabetes as mild, moderate, and extreme. Finally, the output of each classification algorithm is evaluated based on the obtained outcome.

Han Wu, et.al (2018)model was suggested to predict type 2 diabetes mellitus (T2DM) and to increase the prediction model's accuracy. The two-part model supported a series of pre-processing techniques, and the K-means algorithm or logistic regression algorithm was improved in the second step.For the Information Research toolkit, the Pima Indians Diabetes Dataset and the Waikato Ecosystem were used to contrast the effects of various techniques. Compared to other models, the proposed model shows improved accuracy and also provides adequate consistency for the dataset.

.

III. RESEARCH METHODOLOGY

To classify the data into a certain number of classes, the SVM classifier is used. It is very difficult to apply machine learning and data mining in any single research study in order to evaluate diabetes. Using various kernels, we'll evaluate SVM algorithms and add them to the dataset. In terms of precision scores, the outcomes of these various methods will be compared.

3.1 Brief Overview of Used Algorithms :

3.1.1 Support Vector Machine (SVM):

SVM is one of the standard set of classification-based supervised machine learning models. A support vector machine aims to find the best high-margin separation hyperplane between the two classes, given a two-class training sample. "SVM simultaneously minimizes the empirical classification error and maximize the geometric margin. So SVM is called Maximum Margin Classifiers. SVM is a general algorithm supported guaranteed risk bounds of statistical learning theory i.e. the so-called structural risk minimization principle. SVMs can efficiently perform non-linear classification using what's called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. The kernel trick allows constructing the classifier without explicitly knowing the feature space " [7].

By optimizing the space between the 2 decision limits, the SVM finds the ideal separation hyperplane. Mathematically, maximizing the distance amoung the hyperplane is established by $w^t x + b = -1$,This distance is equally equal to $\frac{\mathbf{2}}{||\boldsymbol{w}||}$. This means we want to solve a limit of $\frac{\mathbf{2}}{||\boldsymbol{w}||}$. Equally, we want a minimum of $\frac{||\boldsymbol{w}||}{\mathbf{2}}$. Also, the SVM should properly classify all x(i), which means $y^i(w^t x^i + b) \geq 1, \forall i \ \in \{1, N\}$.

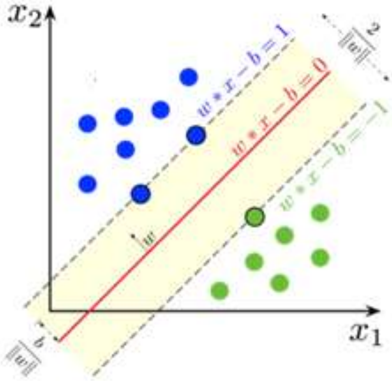

Fig 1: "Maximum-margin hyperplane and margins for an SVM trained with samples from two classes." [11]

3.1.2 Linear Kernel Function:

The linear kernel is used for linear data separation. It is the kernel function that is most widely used. The inner product of <x, y> plus an optional constant c is given. It is often used when a dataset, such as in Text Classification, has large numbers of features.

$$\boldsymbol{k(x, y) = x^t y + c}$$

3.1.3 Polynomial Kernel Function:

The similarity of training samples over a feature space over polynomials of the original variables represents the polynomial function, enabling the training of non-linear models. It is a kernel that is non-stationary. For problems where all training samples are normalized, polynomial kernels are well equipped.

$$\boldsymbol{k(x, y) = (\propto x^t + c)^d}$$

The slope **α**, the constant term **c** and the polynomial of degree **dadjustable parameters**.

3.1.4 Radial Basis Kernel Function (RBF):

The Radial Basis Kernel function is employed to seek out a non-linear classifier or a regression curve. The main motive of the kernel is to do calculations in any d-dimensional space where $d > 1$ so that we can get a quadratic, cubic, or any polynomial equation of higher degree for our classification or regression line. Since a Radial basis kernel uses exponent and as we know the expansion of e^xgives a polynomial equation of infinite power, so using this kernel, we can get a curve fitting any complex dataset.

The kernel output depends upon the Euclidean distance of $(\boldsymbol{x_i}, \boldsymbol{x_j})$ from (among these one will be a support vector and the other will be a testing data point) and $\boldsymbol{\sigma}$is a parameter which is free.

$$\boldsymbol{K(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2\sigma^2})}$$

3.1.5 Sigmoid Kernel Function:

We may use it as a neural network proxy. SVM with a Sigmoid kernel is like a 2-layer perceptron. Equation is

$$\boldsymbol{k(x, y) = \tanh(\alpha x^t y + c)}$$

3.2 Dataset Used:

“The proposed work is conducted on Diabetes Dataset named Pima Indians Diabetes Dataset (PIDD), which is taken from the UCI Repository. This dataset comprises of medical detail of 768 subjects who are female patients. The dataset is also comprised of numeric-valued 8 attributes” [12]. “and in the last column where the value of one class ’0’ treated as tested negative for diabetes and the value of another class ’1’ is treated as tested positive for diabetes[9]. Dataset description is defined by Table-1 and Table-2 represents Attributes descriptions” [13].

Table 1: Database Prediction

Database	No. of Columns	No. of Rows
PIDD	9	768

Table-2: Attribute Description

S. No.	Attribute Name	Attribute Description
1	Pregnancies	Amount of Pregnancy Cycles
2	Glucose	Concentrationof plasma glucose for 2 hours in an oral glucose tolerance procedure
3	Blood Pressure	Diastolic blood pressure (mm Hg)
4	Skin Thickness	Triceps skin fold thickness (mm)
5	Insulin	2-Hour serum insulin (mu U/ml)
6	BMI	Body mass index (weight in kg/(height in m)^2)

7	Diabetes Pedigree	Diabetes pedigree function
8	Age	Age (years)
9	Class	‘0’ and ‘1’

3.3 Measures of Accuracy

Machine Learning model accuracy is the estimation used to figure out which model is good at identifying patterns and relationships between the various features in a database depending on the input or the training data. The better a model can sum up to concealed information, the better forecasts and insights it can offer.

The SVM algorithm with its different kernels is used in this review work. Accuracy Score, Recall, Precision, and F1 Score measures are utilized for the arrangement of this work. Table 3 defines these measures below.

Table-3: Measures of Accuracy

S. No.	Measure	Definition	Formula
1	Accuracy Score	the closeness of a measured value to a true value	**A = (TP + TN)/ (TP + FN + TN + FP)**
2	Precision	the closeness of two or more measurements to each other	**P = TP / (FP + TP**
3	Recall	**recall** means **the percentage of a certain class correctly identified**	**R = TP / (FN + TP)**
4	F1 Score	It is the harmonic mean between precision and recall	F = 2*(P*R) / (P+R)

3.4 Experimental Procedure

The steps are the following:
7
Step 1- Getting the dataset.

Step 2- Splitting the dataset into training and testing samples.

Step 3- Normalizing the features- “Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information” [5].

Step 4- Applying SVM with different kernels on the testing sample and check for the accuracy score of SVM with each kernel.

Step 5- Instantiating the best model with the highest accuracy score.

Step 6- Check for the accuracy of the testing sample.

Step 7- Precision and recall to evaluate the performance of classification.

Step 8- Use the model with a Docker-based web API for Flask to predict results for individual women patients.

3.5 Model Diagram

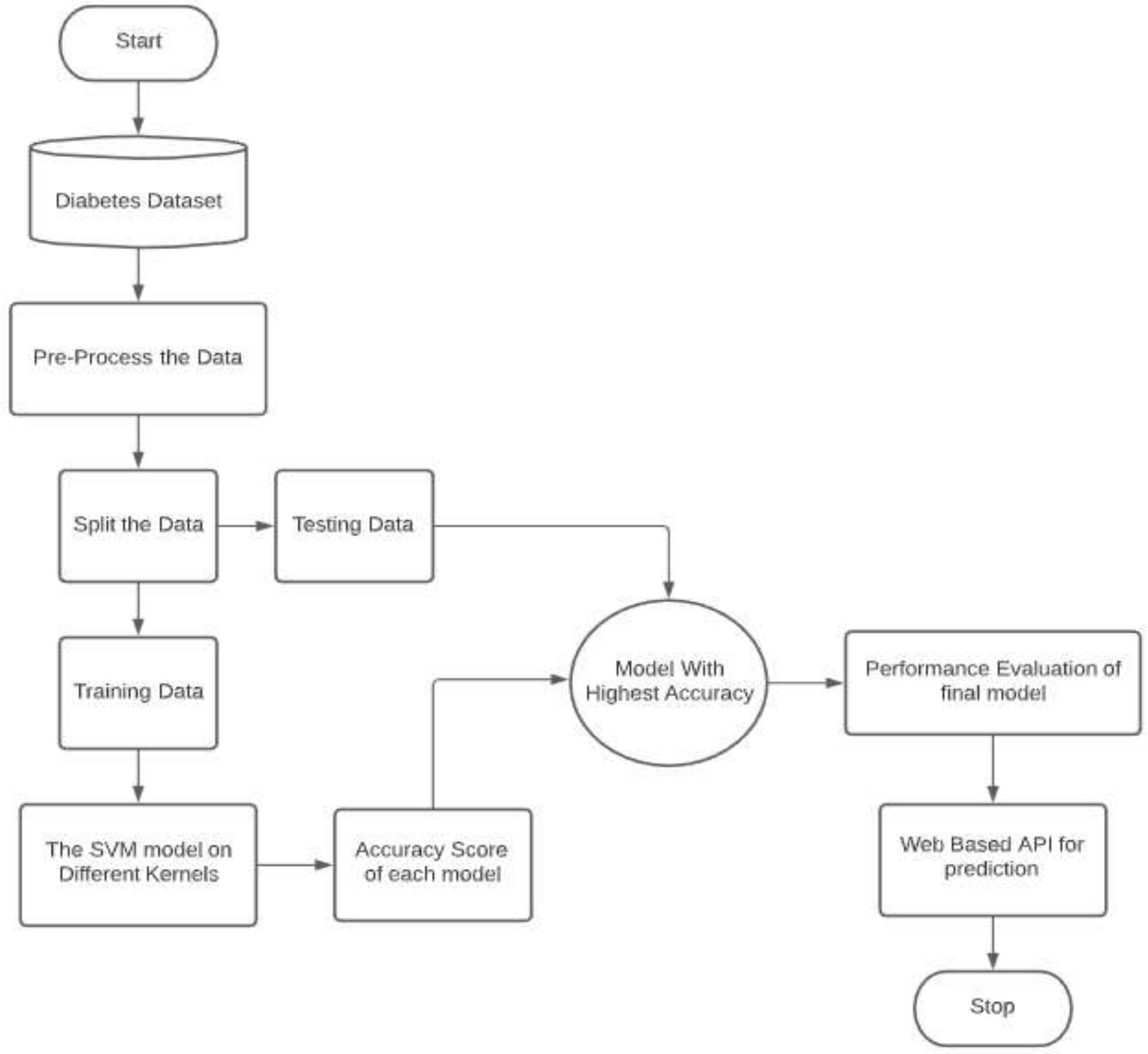
Start
Diabetes Dataset
Pre-Process the Data
Split the Data
Testing Data
Training Data
The SVM model on Different Kernels
Accuracy Score of each model
Model With Highest Accuracy
Performance Evaluation of final model
Web Based API for prediction
Stop

Fig 2 Modal Diagram

IV. Results

Table-4 represents the different SVM algorithm output values with the various kernels determined on the basis of the Accuracy Ranking. From Table-4 it is analyzed that Radial Basis (RBF) kernel is showing the maximum accuracy on the training data.So, compared to others, the SVM with the RFB kernel machine learning classifier can predict the chances of diabetes with more precision.

Table-4: Comparative Performance of Kernels with SVM Algorithm on Accuracy Scores based on training data

S. No.	Kernels	Accuracy %
1.	Linear	78.12
2.	Polynomial	79.51
3.	RBF	83.15
4.	Sigmoid	68.75

By seeing the comparison of accuracy scores of different kernels we instantiate the SVM algorithm with the RBF kernel that gives the highest accuracy score on training data.
So we used this classifier on the diabetes testing data and Table-5 shows the results.

Table-5: Accuracy Measures on testing diabetes data with RBF as a kernel with SVM algorithm

Algorithm	Accuracy Score %	Precision	Recall	F1-Score
SVM with RBF Kernel	72.91	0.72	0.73	0.73

V. Conclusion

Diabetes is a chronic condition worldwide that causes massive casualties. So, one of the most significant medical issues in the real world is identifying diabetes at its early stage. Methodical efforts are being made during this study to design a classifier that ends up predicting diabetes in female patients. During this work, the support vector machine (SVM) algorithm is studied and evaluated on different measures with its different kernels. Experiments on the Diabetes Dataset of the UCI Repository called Pima Indians Diabetes Dataset (PIDD). The adequacy of 83.15 percent on training data and 72.91 percent on testing data using the Radial Basis Function (RBF) kernel with the SVM classification algorithm is calculated by experimental results.

VI. Future Scope

In the future, the same Docker-based web API may be used to deploy the designed system with advanced machine learning classification algorithms.This research is only targeted towards women patients with diabetes. More models can be created using different algorithms and datasets to predict or diagnose other diseases for every patient.

References

1. https://www.who.int/health-topics/diabetes#tab=tab_1
2. Theofilatos K, et al. Predicting protein complexes from weighted protein–protein interaction graphs with a novel unsupervised methodology: evolutionary enhanced Markov clustering. Artif Intell Med 2015;63(3):181–9
3. Pavel Hamet, Johanne Tremblay Centre de recherche, Centre hospitalier de l'Université de Montréal (CRCHUM), Montréal, Québec, Canada, H2X 0A9 Department of Medicine, Université de Montréal, Montréal, Québec, Canada, H3T 3J77, Canada, Artificial intelligence in medicine,
4. International Journal for Research in Engineering Application & Management (IJREAM) ISSN : 2454-9150 Vol-05, Issue-02, May 2019 Prediction of Diabetes Using Support Vector Machine
5. https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/normalizedata#:~:text=Normalization%20is%20a%20technique%20often,of%20values%20or%20losing%20information.
6. Sisodia, D., Shrivastava, S.K., Jain, R.C., 2010. ISVM for face recognition. Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010554–559doi:10.1109/CICN.2010.109
7. V. Anuja Kumari, R.Chitra / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 3, Issue 2, March -April 2013, pp.1797-1801 1797 | P a g e Classification Of Diabetes Disease Using Support Vector Machine V. Anuja Kumari1, R.Chitra2
8. https://en.wikipedia.org/wiki/Radial_basis_function_kernel
9. International Conference on Computational Intelligence and Data Science (ICCIDS 2018) Prediction of Diabetes using Classification Algorithms Deepti Sisodia, Dilip Singh SisodiaNational Institute of Technology, G.E Road, Raipur and 492001, India National Institute of Technology, G.E Road, Raipur and 492001, India
10. http://www.ijera.com/
11. https://hal.sorbonne-universite.fr/
12. Deepti Sisodia, Dilip Singh Sisodia. "Prediction of Diabetes using Classification Algorithms" , Procedia Computer Science, 2018
13. https://www.ijitee.org/