IJARSCT

ISSN (Online) 2581-9429

IJARSCT
Impact Factor: 6.252

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 2, Issue 6, June 2022

Diabetes Prediction Model Comparison between XgBoost and SVM Algorithms

Harsh Vardhan[1], Harasis Singh[2], Amit Mithal[3], Himanshi Kabra[4], Aditya Sharma[5]

Students, Department of Computer Science and Engineering[1,2,4,5]
Assistant Professor, Department of Computer Science and Engineering[3]
Jaipur Engineering College and Research Centre, Jaipur, Rajasthan, India
harshvardhan.cse22@jecrc.ac.in[1], harasissingh.cse22@jecrc.ac.in[2], amitmithal.cse@jecrc.ac.in[3]
himanshikabra.cse22@jecrc.ac.in[4], adityasharma1.cse22@jecrc.ac.in[5]

Abstract: *Diabetes is a global health epidemic. It increases the danger of cardiovascular disease by fourfold in women and around twice in men. 'Diabetes' is an umbrella term for a number of different subtypes of the disease. The most common are Type 1 Diabetes Mellitus (T1DM) and Type 2 Diabetes Mellitus (T2DM). Compared to men, women are also at a greater risk of retinopathy and neuropathy from diabetes. Pregnancy may worsen pre-existing conditions and lead to significant blindness. It also aggravates pre-existing kidney diseases. Elderly women with type 2 diabetes mellitus (T2DM) and end-stage renal disease have a significantly higher risk of death than men with similar diseases. Women with diabetes have higher chances of suffering a stroke in comparison to women without it. Women are also more likely to develop depression compared to men. The modeling of support vector machines may additionally be a promising classification technique for identifying women among the population with common diseases like polygenic disorder and pre-diabetes. We use different algorithms for classification, XGBoost based on SVM with GridSearchCV predict results with 83.5% accuracy.*

Keywords: Machine Learning, Support Vector Machine, Algorithms, Medical Diagnosis, Classification.

I. INTRODUCTION

Diabetes mellitus is a growing health problem in India, but also around the world. When it comes to managing diabetes and stemming the epidemic, Indian doctors and researchers are increasingly taking the lead. Being a chronic, potentially life-threatening disease, diabetes affects millions of people worldwide. More than 3 million people die from the disease every year. Diabetes affects one out of five adults in certain Indian urban cultures. "Virtual and physical Al are the two main branches of the technology. The virtual component is Machine Learning (also called Deep Learning), which is a set of algorithms that improve learning through experience. Machine learning algorithms fall into three categories: (1) unsupervised (finding patterns). (ii) Supervised (classification and prediction algorithms that depend on previous examples), and (iii) reinforcement learning (use of sequences of rewards and punishments to make way for operation throughout a particular downside space). Firstly, Al has led to and continues to have a positive impact on advances in genetic science and molecular medicine, via machine learning algorithms and data management. [3]. With the help of machine learning algorithms and certain diagnostic measures included in the dataset, we can diagnose whether a patient has diabetes. In this study, we use the SVM algorithm and several other methods of machine learning to differentiate between women without diabetes and women with pre-diabetes or undiagnosed diabetes. Research problem formulation: The analysis of expectations is an approach to forecasting future results through the use of current data. Women with diabetes should make sure their glucose levels, blood pressure, insulin and body mass index (BMI) are kept within their goal range targets, neither too high nor too low. They should learn when and what to eat between meals and snacks. This paper is organized as follows: In section II, the formulation of the research problem is described. In section III, the various classification techniques used for predicting diabetes are discussed. A brief overview of some basic SVM principles is presented in Section IV followed by a presentation of the kernel functions with precision measurements in Section V, and a discussion of the experimental results in Section VI. Finally, a discussion of the potential application of the technique is presented in Section VII.

DOI: 10.48175/IJARSCT-5002

II. LITERATURE REVIEW

A. Bamnote, M.P., G.R., (2014) Using the Diabetes dataset from the UCI repository, Genetic Programming (GP) was used to train and assess the info for predicting polygenic disease. Results obtained from GP have the optimum exactness relative to alternative techniques introduced. By taking less time for the classifier generation, there's conjointly a serious increase in accuracy. It seems to be a helpful model for the predicting.

B. Aishwarya 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 theday, the model proposed gives the safest and most successful outcome.

C. 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.

D. 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.

E. 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 diabetescluster-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.

F. 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 and XGBoost 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 algorithm and XGBoost and then 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)

The SVM is a standard set of supervised machine learning models based on classification. SVMs aim to find the best separation hyperplane between two classes, given two training samples. "In SVM, the empirical classification error and geometric margin are simultaneously minimized, thus the term Maximum Margin Classifier. In SVM, the principle of structural risk minimization is applied, which is a general algorithm, which supports guaranteed risk bounds of statistical learning theory.By implicitly transforming their inputs into high-dimensional feature spaces by using the kernel trick, SVMs can perform non-linear classification efficiently. It is possible to construct the classifier without explicitly knowing the feature space by using the kernel trick. "[7]. Using the SVM technique, the ideal separation hyperplane is found by optimizing the space between the two decision limits. Maximizing the distance between the hyperplanes is established mathematically by $wtx + b = -1$, this distance is equally equal to $\mathbf{2}/\|\boldsymbol{w}\|$. This means we want to solve a limit of $\mathbf{2}/\|\boldsymbol{w}\|$.

Equally, we want a minimum of $\|\boldsymbol{w}\|/\mathbf{2}$. Also, the SVM should properly classify all x (i), which means $yi\ (wtxi + b) \geq 1$, $\forall i \in \{1, N\}$.

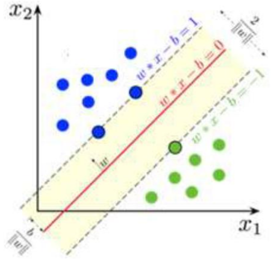

Fig 1: “Maximum-margin hyperplane and margins for an SVM trained with samples from two classes.” [11]

3.1.2 Radial Basis Kernel Function (RBF)

Radial basis kernels seek out non-linear classifiers or regression curves. For our regression or classification line, the kernel's primary purpose is to calculate quadratic, cubic, or polynomial equations in any d-dimensional space where d > l in order to get a quadratic, cubic, or polynomial equation of greater degree. Since a Radial basis kernel uses exponent and we know that the expansion of exgives a polynomial equation of infinite power, so making use of this kernel, we can get a curve that can satisfy any complex dataset. The kernel output depends on the Euclidean distance of ($\boldsymbol{xi}$,$\boldsymbol{xj}$) from (among these one will be a support vector and the other will be a testing data point) and $\boldsymbol{\sigma}$is a free parameter.

$$\boldsymbol{K}\ (\boldsymbol{xi}, \boldsymbol{xj}) = \mathbf{exp}\ (-\|\boldsymbol{xi} - \boldsymbol{xj}\|/\mathbf{2}\boldsymbol{\sigma}^2)$$

3.1.3 XGBoost

XGBoost is an implementation of Gradient Boosted decision trees. This algorithm involves creation of decision trees in sequential form. Weights have a vital role to play in XGBoost. All the independent variables are assigned with weights which are then sent as input into the decision tree for predicting results. The weight of variables which are wrongly predicted by the decision tree increases and then these variables are sent as input to the second decision tree. These individual classifiers/predictors then ensemble to provide a robust and accurate model. It will work on regression, classification, ranking, and user-defined prediction issues.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \epsilon \mathcal{F}$$

3.1.4 GridSearchCV

GridSearchCV is a library function which is a member of Scikit-learn machine learning library. It allows us to loop through predefined hyper parameters and properly fit our estimator (model) on our training dataset, so we can handpick the best parameters from the listed hyper parameters. Additionally, it allows us to specify the number of times we want to cross-validate each set of hyper parameters. Below are the few basic hyper parameters we need to define:

1. **estimator**: estimator object we created
2. **params_grid:** the dictionary object that holds the hyper parameters we want to try
3. **scoring:** evaluation metric that we want to use, we can simply pass a valid string/ object of evaluation metric
4. **cv**: number of cross-validation we want perform for hyper parameters
5. **verbose:** set it to 1 to get the detailed print out while we fit the data to GridSearchCV
6. **n_jobs**: number of processes we wish to run in parallel for this task. If it’s -1, it will use all available processors.

3.2 Dataset Used

“The planned work is conducted on polygenic disease Dataset named Pima Indians Diabetes Dataset (PIDD) that is taken from the UCI Repository. This dataset contains medical details of 768 female patients. The dataset additionally comprises of numeric-valued eight attributes” [12]. “And at intervals the last column wherever the price of one category ’0’ treated as tested negative for diabetes and category ’1’ is treated as tested positive for diabetes [9]. Dataset description is outlined by Table-1 and Table-2 represents Attributes descriptions” [13]

.

Database	No. of columns	No. of rows
PIDD	9	768

Table 1: Dataset Description

S. No.	Attribute Name	Attribute Description
1	Pregnancies	Amount of Pregnancy Cycles
2	Glucose	Concentration of 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’

Table 2: Dataset Features

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 concealed information, the better forecasts and insights it can offer. The SVM algorithm with its RBF kernel and XGBoost is employed during 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.

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)

Table 3: Measures of Accuracy

3.4 Experimental Procedure

The steps are the following:

Step 1- Obtain the dataset.

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

Step 3- Normalizing the features- “Normalization may be a technique typically applied as a part of information preparation for machine learning. The goal of standardization is to vary the values of numeric columns within the dataset to use a typical scale, while not distorting variations in the ranges of values or losing information” [5].

Step 4- Applying SVM with totally RBF kernel and XGBoost on the testing sample and checking for the accuracy score.

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

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

Step 7- Preciseness and recall to gauge the performance of classification.

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

3.5 Model Diagram

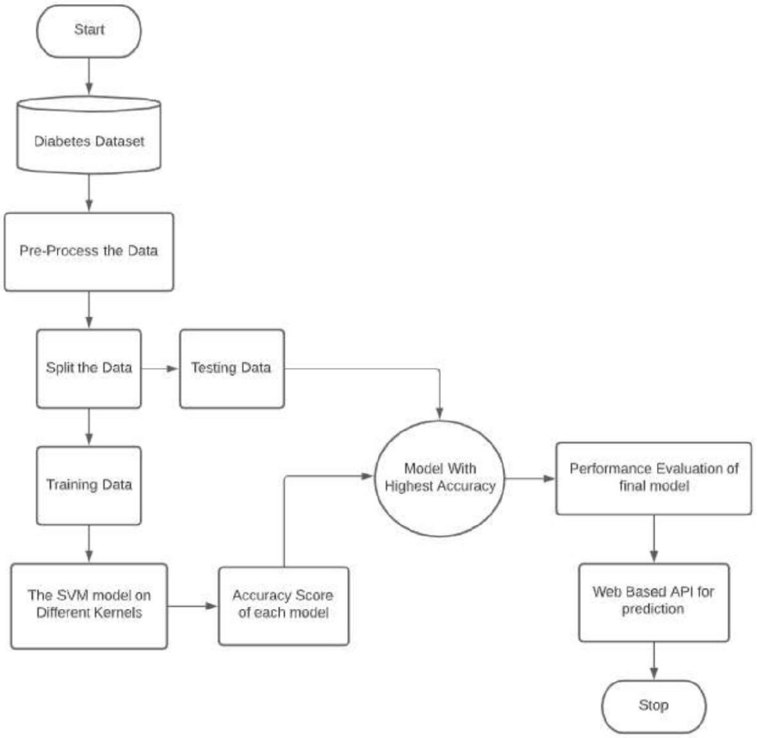
Start
Diabetes Dataset
Pre-Process the Data
Split the Data
Testing Data
Training Data
The SVM model on Different Kernels
Accuracy Score of each model
Model With Highest Accuracy
Performance Evaluation of final model
Web Based API for prediction
Stop

Fig 2: Experimental Diagram

IV. RESULTS

Table-4 represents output values with the SVM algorithm for RBF kernel and XGBoost 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 with SVM and the XGBoostafter the GridSearchCV can predict the chances of diabetes with more precision.

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. ArtifIntell 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-modulereference/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

DOI: 10.48175/IJARSCT-5002