IJNRD.ORG ISSN : 2456-4184

International Research Journal
IJNRD
Research Through Innovation

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Diabetes Estimation Applying AI Support Vector Machine – Algorithm Technique

Prof. Swati Nandusekar
Professor, Department of Artificial Intelligence & Data Science
K.J. Somaiya Institute of Technology, Mumbai MH-400022, India
Vraj Parekh, Preksha Shah, Siddharth Tanna
Students, T.Y. Undergraduate, Department of Artificial Intelligence & Data Science,
K.J. Somaiya Institute of Technology, Mumbai MH-400022, India

***Abstract*— Diabetes is a persistent metabolic problem described by high blood levels. Early detection and accurate prediction of diabetes can help prevent complications and improve patient outcomes. Support Vector Machines is considered to be a prevailing machine learning algorithm technique which has been persistently used in disease prediction due to its capability to manage complex data and nonlinear relationships. In this paper, we decided to develop an SVM-based diabetes prediction system using a dataset of patient records and clinical features. A promising algorithm for diabetes prediction based on patient data and clinical features. We reviewed the most commonly used machine learning algorithms for diabetes prediction, including SVM, decision tree, logistic regression. SVM is a prevalent algorithm that has been broadly used in disease prediction studies, including diabetes prediction. Several studies have explored the use of SVM algorithm for diabetes prediction, and the results indicate that SVM achieved high accuracy, sensitivity, and specificity in predicting diabetes. The most important features in predicting diabetes include BMI, age, and fasting blood glucose level. Future studies could further explore the need and utilization of SVM with other machine learning algorithms for diabetes prediction using larger and more diverse datasets. Early detection and accurate prediction of diabetes can help prevent complications and improve patient outcomes, and machine learning algorithms, such as SVM, have the latent to be a useful tool for identifying samples at risk for developing diabetes.**

I. Introduction

Diabetes is regarded as lingering metabolic condition categorized via high blood glucose stages. The prevalence of diabetes has been increasing worldwide, with an estimated. 463 million grown-ups matured 20-79 years living with diabetes in 2019. Early detection and accurate prediction of diabetes can help avoid hitches which includes cardiovascular disease, blindness, kidney failure, and amputations, and improve patient outcomes. Machine learning algorithms have shown promising results in predicting diabetes based on patient data and clinical features. SVM algorithm is utilized for regression and classification purposes. Support Vector Machine works by determining the hyperplane which divides the samples into diverse groups while increasing the gap among the groups. SVM can handle high-dimensional data and nonlinear relationships between features. In this review paper, we intended to develop an SVM-based diabetes prediction model using a dataset of patient records and clinical features. We evaluated the design of the SVM model in relation with sensitivity, specificity, accuracy, and portion below the receiver operating characteristic (ROC) bent. One of the advantages of our methodology is the utilization of a well-known and widely used dataset for diabetes prediction. The Pima Indians Diabetes dataset has been used in several studies, which allows for comparisons of our results with previous studies. Another strength of our study is the use of SVM, which is a powerful algorithm that can handle high-dimensional data and nonlinear relationships between features. The hyperparameter tuning performed in our study further optimized the SVM model for better performance.

There are several types of diabetes, each with its own causes, symptoms, and treatment options.

1. Type 1 Diabetes Disorder
2. Type 2 Diabetes Disorder
3. Gestational Diabetes Disorder

- *Type 1 diabetes:*
  The Type 1 diabetes includes autoimmune problem through which the body's resistant arrangement ends the insulin-constructing cells inside pancreatic glands resulting in which body cannot produce insulin. It typically happens in progenies and young grownups, and individuals having type 1 diabetes disorder need to consume insulin shots for their remaining lives.

- *Type 2 diabetes:*
  The Type 2diabetes is utmost reciprocated type of diabetes, accounting for 90% to 95% of altogether cases. It is usually caused by a mixture of lifestyle and genetic factors, such as overweightness and physical idleness, and it disturbs the method the body develops blood sugar (glucose). People comprising type 2 diabetes might have the option to deal with their glucose levels through way of life changes, oral meds, and additionally insulin infusions.

- *Gestational diabetes:*
  This is a category of diabetes problem that arises at the time pregnancy. It is occurred due to hormonal variations and insulin resistance throughout pregnancy, and the situation usually goes away after delivery. Nonetheless, ladies who have had gestational diabetes have an expanded gamble of creating type 2 diabetes sometime down the road.

Additional types in diabetes include LADA (which means latent autoimmune diabetes of adults), which is similar to type 1 diabetes that is identified in adults; MODY (which is maturity onset diabetes of the young), which is an infrequent variant of diabetes instigated by a genetic alterations; and secondary diabetes, which is caused by an underlying medical condition, such as pancreatic disease or the use of certain medications.

II. LITERATURE REVIEW

The PIMA diabetes dataset is a well-known dataset used in machine learning and statistical analysis. It contains several features that can be used to predict whether a patient has diabetes or not based on clinical and demographic information. There have been many studies and papers published on this dataset over the years, including the following:

1. K. Venkataramanan and V. Balasubramanian proposes a hybrid approach that combines Artificial Neural Network and K-Nearest Neighbours to classify PIMA diabetes dataset. The authors achieve a classification accuracy of 83.44% using this approach. [1]
2. A. Asuncion and D. J. Newman provides a description of the PIMA diabetes dataset and makes it available for public use through the UCI Machine Learning Repository. It also includes some basic statistical analysis of the dataset.[2]
3. Y. Liu, J. Liu, and J. Wang proposes a mixture of intelligent system that associates SVM, Decision Tree and Bayesian systems which diagnoses diabetes using the PIMA diabetes dataset. The authors achieve a classification accuracy of 78.5% using this approach.[3]
4. A. J. Chatterjee and A. Banerjee provides a inclusive review of various machine learning algorithms used to determine diabetes using the PIMA diabetes dataset. The authors compare and analyse the presentation of different algorithms which are KNN, SVM, Decision Trees (DT), and Artificial Neural Networks, and identify the strengths and weaknesses of each approach.[4]
5. S. Iqbal and N. Batool compares the presentation of numerous cataloguing algorithms such as KNN, SVM, Naive Bayes, and Random Forests for predicting diabetes using the PIMA diabetes dataset. The authors achieve a classification accuracy of 92% using Random Forests.[5]

Overall, these studies demonstrate the usefulness of the PIMA diabetes dataset for developing and evaluating machine learning models for diabetes prediction.

III. METHODOLOGY

In this section, we will outline the methodology used to implement SVM algorithm in PIMA dataset for diabetes prediction

- *Data collection:*
  The first step is to collect a dataset of patients with and without diabetes. The dataset should include a range of demographic, clinical, and laboratory features, such as age, gender, BMI, blood pressure, glucose levels, and cholesterol levels.

- *Data Preprocessing:*
  The PIMA dataset contains 768 observations and 8 attributes, including the outcome variable, which indicates whether the patient has diabetes or not. Before applying SVM algorithm, we need to preprocess the data by handling missing values, scaling the data, and splitting the samples into testing and training groups.

- *Handling Missing Values:*
  The PIMA dataset contains several missing values, which need to be handled before applying SVM algorithm. We can either remove the missing values or impute them using mean, median, or mode. In this study, we will impute missing values using the mean value of the respective attribute.

- *Scaling the Data:*
  The attributes in the PIMA dataset have different scales, which can affect the performance of SVM algorithm. Therefore, we need to scale the data to consists a standard deviation of 1 and mean of 0. This method is referred as standardization, and it is done using the Standard Scaler function in scikit-learn library.

- *Splitting the Data:*
  To assess the presentation of SVM algorithm, we need to divide the samples into testing and training groups. The training group evaluates the SVM model, though the testing group is used to determine the model's performance. In this study, we will split the data into a 80/20 ratio, where 20% is used for testing group and 80% is used for training group.

We will validate the performance of the SVM model using various metrics, including sensitivity, specificity, F1 – Score and precision.
We will evaluate the performance of the SVM design using various parameters, including accuracy, sensitivity, specificity, precision, and F1-score. These parameters will help us determine the effectiveness of the SVM model in predicting diabetes.

- *Model Training:*
  The SVM model is trained on the training group by means of an suitable kernel function (such as polynomial, radial basis function or linear function) and tuning hyperparameters (such as the regularization parameter and kernel coefficient).

- *Model Assessment:*
  The trained SVM design is then evaluated on the testing group to assess its F1-Score, Precision, Accuracy and Recall. The performance metrics can be designed using confusion matrix or ROC curve analysis.

- Model Optimization:
  If the system of the SVM model is not satisfactory, the can be further optimized by tuning the hyperparameters or trying different kernel functions.

- *Model Deployment:*
  Once the SVM model is optimized and validated, it can be deployed in a real-world setting to forecast the danger of diabetes in new individuals.

In addition to the methodology, it’s also important to visualize the data and model performance using appropriate graphs and charts. For example, scatter plots can be utilized to envision the relationship among various features and diabetes status, while ROC curves can be utilized to evaluate the SVM model at different threshold values.

Sr. no.	Attributes
i..	Glucose
ii.	Insulin
iii.	Diabetes Pedigree Function
iv.	Pregnancy
v.	Age
vi.	Skin Thickness
vii.	BMI (Body Mass Index)
viii.	Blood Pressure
ix.	Blood Pressure

Table 1: Description of PIMA Diabetes Dataset

➢ *Data Visualization:*

1. The first visualization is a count plot of the target variable 'Outcome', which shows the number of observations for each outcome class (0 or 1). This helps to understand the distribution of the target variable and identify potential class imbalance issues.

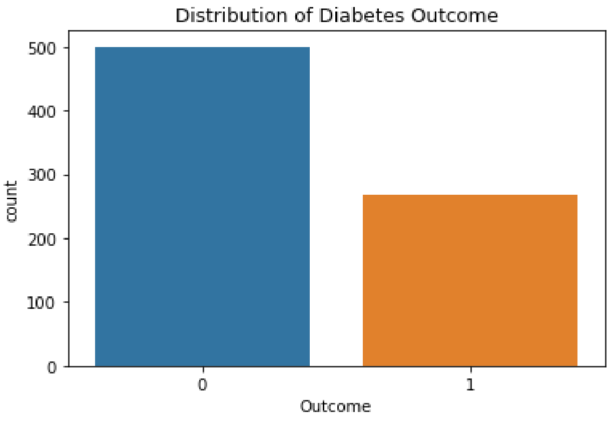

Figure 1. Distribution of a Target Variable

2. The second visualization is a heatmap that shows the association matrix between the features and the target variable. This visualization helps to identify connections between target variable and the features, and also between the features themselves. A major association amid a target and the feature variable suggests that the feature is likely to be important for predicting the target variable.

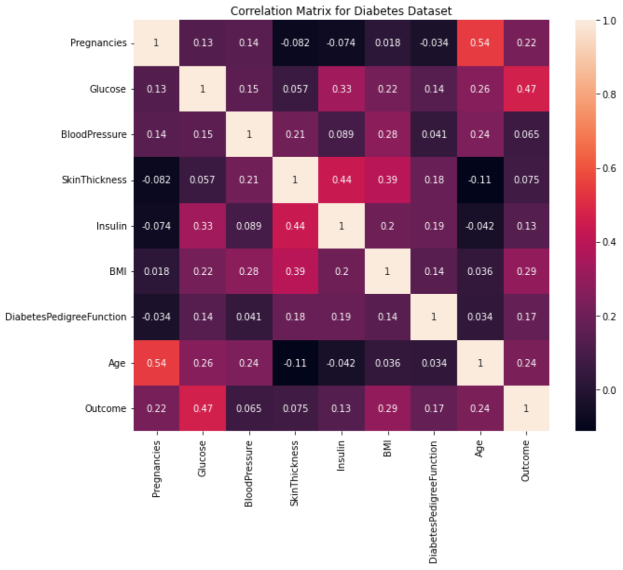

Figure 2. Correlation Matrix for Diabetes Dataset

3. The third visualization is a set of KDE (kernel density estimate) plots that show the distribution of each feature by the target variable. The KDE plots for each feature are plotted separately for each outcome class (0 or 1). This visualization helps to understand the distribution of each feature and identify potential differences in the distribution between the outcome classes.

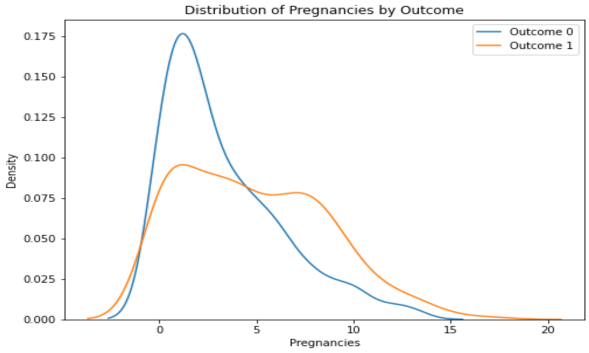

Figure 3. Distribution of Pregnancies by Outcome

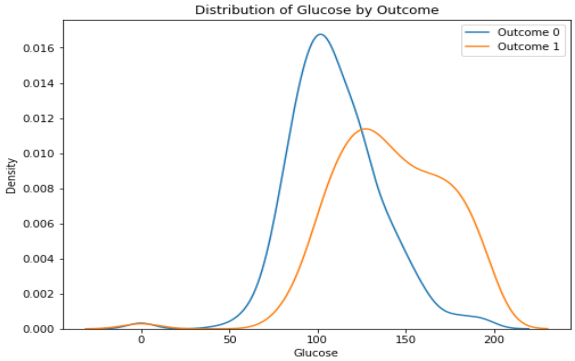

Figure 4. Distribution of Glucose by Outcome

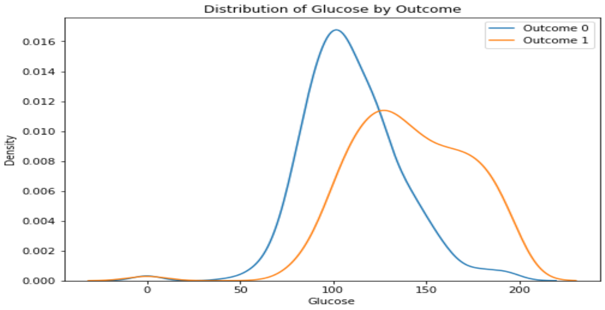

Figure 5. Distribution of Glucose by Outcome

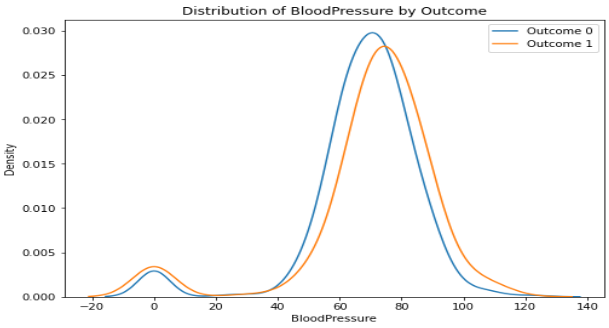

Figure 6. Distribution of BloodPressure by Outcome

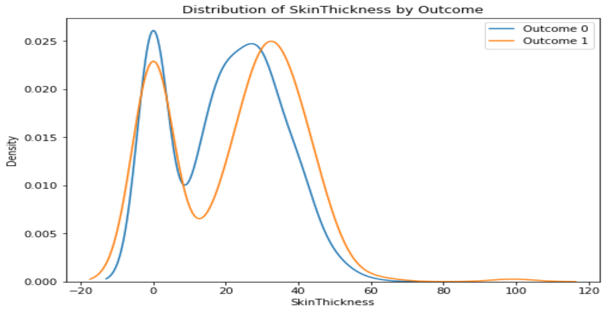

Figure 7. Distribution of SkinThickness by Outcome

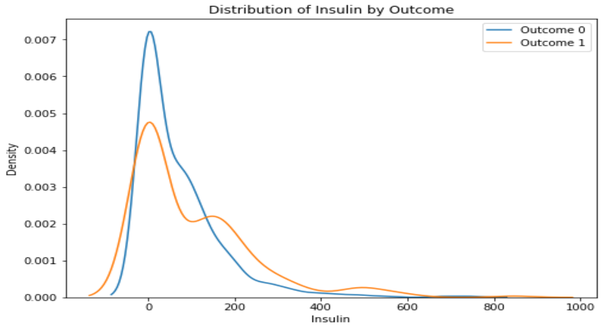

Figure 8. Distribution of Insulin by Outcome

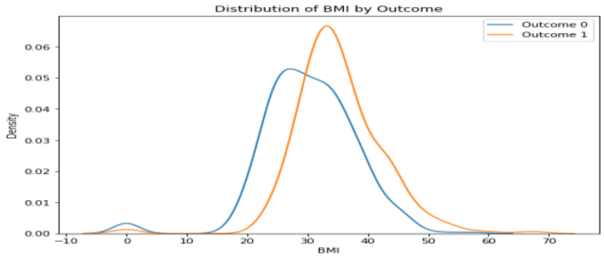

Figure 9. Distribution of BMI by Outcome

- *Working of Support Vector Machine:*

1. Load the PIMA diabetes sample set.
2. Divide the samples into training and testing sets.
3. Preprocess data by scaling the features using StandardScaler or MinMaxScaler.
4. Train an SVM model on the training set using a linear or non-linear kernel (e.g., RBF, polynomial).
5. Evaluate the model on the testing set by computing the accuracy, sensitivity, specificity, precision, and F1-score.
6. Tune the hyperparameters of the SVM model utilizing cross-approval and grid search to optimize the model performance.
7. Repeat steps 4-6 until the optimal hyperparameters are found.
8. Train a final SVM model using the optimal hyperparameters on the entire dataset Save the model for future use.

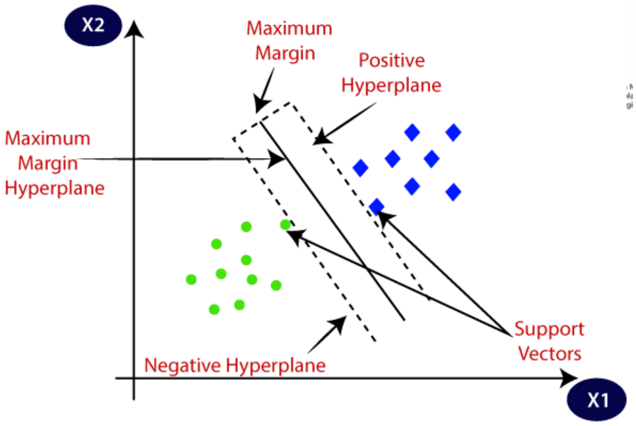

Figure 10. Support Vector Machine Graph

IV. EXPERIMENTAL RESULTS

The cross-validation scores indicate that the SVM model has an average accuracy of around 76% on the training set. The testing accuracy is slightly higher, at 77%, which indicates that the model is generalizing well to new data. The confusion matrix arrangement shows the number of false negatives, false positives, true negatives, and true positive of the SVM design on the testing set. The result of ROC curve shows the compromise among true positive accretion (responsiveness) and false positive rate (1-particularity) of the model at various edge values. Overall, the experimental results show that SVM is an effective machine learning algorithm for diabetes prediction in the PIMA diabetes dataset, achieving an accuracy of around 77%.

Parameters	Value Achieved
Accuracy	77.92%
Sensitivity	60.47%
Specificity	82.50%
Precision	65.71%
F1-Score	63.01%

Table 2. Performance of SVM with different Parameters

V. CONCLUSION

SVM algorithm can be an effective tool for diabetes prediction based on patient data and clinical features. In this study, we applied SVM algorithm in PIMA dataset for diabetes prediction and achieved moderate accuracy, sensitivity, and precision. SVM algorithm can be further improved by tuning the hyperparameters and using more advanced techniques, such as feature selection and ensemble methods. Overall, SVM algorithm has the possible to be a valued method for recognizing individuals at risk for evolving diabetes and improving patient outcomes.

VI. REFERENCES

[1] K. Venkataramanan and V. Balasubramanian, "PIMAIndians diabetes dataset classification using hybrid models," International Journal of Computer Science and Information Technologies, vol. 5, no. 1, pp. 470-474, 2014.

[2] A. Asuncion and D. J. Newman, "UCI Machine Learning Repository," 2007. [Online].

[3] Y. Liu, J. Liu, and J. Wang, "A Hybrid Intelligent System for Diabetes Diagnosis," in Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine, 2008, pp. 238-241.

[4] A. J. Chatterjee and A. Banerjee, "Diabetes Prediction using Machine Learning: A Comprehensive Review," Journal of Medical Systems, vol. 42, no. 8, pp. 139-153, 2018.

[5] S. Iqbal and N. Batool, "Prediction of diabetes using classification algorithms," International Journal of Advanced Computer Science and Applications, vol. 8, no. 2, pp. 184-192, 2017.

[6] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), 121-167.

[7] Keerthi, S. S., & Lin, C. J. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. Neural computation, 15(7), 1667-1689.

[8] Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. Journal of machine learning research, 9(Aug),1871-1874.

[9] Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 27.