DIABETES PREDICTION USING SUPPORT VECTOR MACHINE AND GRADIENT DESCENT ALGORITHM

Yonando Pratama B
Program Studi Teknik Informatika Fakultas Ilmu Komputer, Universitas Katolik Soegijapranata
Yonando17@gmail.com

ABSTRACT (DIABETES PREDICTION USING SUPPORT VECTOR MACHINE AND GRADIENT DESCENT ALGORITHM)

Diabetes is a health problem that can be deadly if not treated early.An early prediction of diabetes can prevent so many health problems in someones life.Using machine learning and algorithm to predict whether a person has diabetes or not can be the best solution when it comes to diabetes problems.Support Vector Machine can classify a data point into positive or negative value,it can help with predicting whether a person has diabetes or not using the diabetes factor that a person has.Using this factor Support Vector Machine can give value to the data point of a person and decided in which side of the margin it lies to,positive or negative value.The dataset itself contains 2000 patients with diabetes factor such as Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, Diabetes, and Age with Outcome as the predictor.The final results shown that Support Vector Machine is a good approach when it comes to predicting patients with diabetes.it shown a high accuracy score of >70% in accuracy,which means the Support Vector Machine model will most likely predict 7 out of 10 patients to be correct when it comes to Diabetes Disease.

Keywords: *Diabetes, SVM, Machine_Learning*

CHAPTER 1 (INTRODUCTION)

2.1 Background

In various medical fields, a classification algorithm was developed to categorize diseases based on information about each individual.Topic includes diabetes, a disease that affects the body's ability to produce the hormone insulin, changing the body's normal carbohydrate metabolism and raising glucose levels on blood. A person with diabetes typically has high blood sugar, which worsens their hunger and thirst. Frequent urination is one of the symptoms of high blood sugar. Because it is impossible to control the amount of sugar material in the body, diabetes is regarded as a serious health issue. Diabetes is influenced by a number of factors, including body mass index (BMI), hereditary factors (Diabetes Pedigree Function), and insulin, but the main factor is the body's blood sugar level (or glucose level). Early detection based on the factor above is the most effective method to handle this disease.

A support vector machine (SVM) is a type of supervised learning algorithm used in machine learning. Using supervised learning, the support vector machine (SVM) can solve regression and

classification problems, this includes classifying diabetes patient. This algorithm can assist in determining if a person has diabetes or not.

A patient's initial diagnosis of diabetes can be identified by utilizing the SVM to identify a patient with the diabetes factors of bmi, diabetes pedigree function, insulin, hereditary, and blood glucose levels. Based on the factors mentioned above, this algorithm will help classify and predict whether a patient has a diabetes condition or not by making a hyperplane and classifying the data of a person into data point that lands in positive or negative side of the hyperplane.

2.2 Problem Formulation

Can this SVM model determine and classify whether or not a person has diabetes?

How is the accuracy of the SVM model based on the train and test data?

How is the SVM model compared to the SVM library in python?

2.3 Scope

This project only used dataset from diabetes patient and only implementing the algorithm in the diabetes dataset,it uses 2000 dataset of patients with diabetes factor.This project also compare the SVM model that used the gradient descent algorithm with the SVM library in python.This Project only focus on the accuracy of the SVM model and whether or not this model can predict a patient has diabetes or not.

2.4 Objective

The purpose of this project is to predict and classify using SVM whether a person had diabetes disease or not,this project also test the accuracy of the SVM model based on the train and test data of diabetic patient dataset.

CHAPTER 2 (RESEARCH METHODOLOGY)

2.1 Dataset Collection

In this step the dataset that are going to be used is collected throught website called kaggle and its from https://www.kaggle.com/datasets/johndasilva/diabetes/data. The dataset has 2000 rows and 9 columns of diabetic patients.The column consist of: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, Diabetes, Age and Outcome.

2.2 Dataset Cleaning

The dataset that have been aquired need to be re-process before its able to be used in the Classifier.The column Outcome (0 and 1) needs to be separated from the other 8 column which is going to become the input for the Support Vector Machine Classifier.The

data in the 8 columns beside Target is going to go through a standarization process and into the Gradient Descent algorithm for updating the weight and bias.

2.3 Data Augmentation

2.3.1 Splitting and Standardizing Input Data

The dataset that has been cleaned and has been split into Input data and Target data is going to be processed by the Gradient Descent Algorithm.First the Input data need to be standardized so it can be process by the algorithm.In this Standardizing Data Input,the Input Data needs to be scale into a value that can be process by the machine,the value such as Glucose and BloodPressure that are too high in value needs to be standarized into a new value so that it can be process,the high value like 50 or 100 will be converted into decimal like 0,5… and 0,6…,making it a new value that are in the same scale.

2.3.2 Split Train Data and Test Data

The Standarized data will be split again into Train Data and Test Data,the train data is going to be the data that the machine learning model used to train its algorithm so that it can improve the model and test data will be the data testing of the machine learning model.

2.3.3 Updating Weight and Bias with Gradient Descent Algorithm

Gradient Descent is an algorithm that optimized machine learning model,its an optimization algorithm used for minimising cost function.This algorithm also used the Hinge Loss,a type of Loss Function used for maximising margin in classification models.

$$\ell(y) = max(0,1 - \gamma_i \quad (w.x + b)) \quad (1)$$

3. Figure 3.3 Hinge Loss Function

In the function (1) , $\ell(y)$ is the raw output of loss function, 0 and 1 represent whether the classification is correct or not, γ_i is the index of input data in a single row,w is for weight, x is the input data from the dataset, while b is bias.This whole function (1) represent the Hinge Loss function that are going to be the condition for the Gradient Descent Algorithm.

$$(\gamma_i \,.\, (w.x + b) \geq 1) : \quad (2)$$

Figure 3.4 Hinge Loss Function in the correct side of margin

$$\frac{dJ}{dw} = 2\lambda w \quad (3)$$

$$\frac{dJ}{db} = 0$$

Figure 3.5 Formula for changing Weight and Bias in the correct side of margin

$$(\gamma_i \,.\, (w.x + b) \leq 1) : \quad (4)$$

Figure 3.6 Hinge Loss Function in the wrong side of margin

$$\frac{dJ}{dw} = 2\lambda w - \gamma_i \,.\, x_i \quad (5)$$

$$\frac{dJ}{db} = \gamma_i$$

Figure 3.7 Formula for changing Weight and Bias in the wrong side of margin

For Gradient Descent can be applied first it needs the condition to be fulfilled which is using the Hinge Loss Function condition.As show in function (2),when the input data of x lies in the correct side of the margin (≥ 1),The formulas for changing the weight and bias will be as shown in the function (3).When the input data of x lies in the wrong side of the margin (≤ 1) as shown in function (4),the formula of changing weight and bias will be different as shown in the function (5).For this function: d is derivative, J is cost function, λ is lambda parameter, γ_i is our rows index of data input, x_i is our input data, w is weight and b is bias.In the Gradient Descent Algorithm the purpose is to find the best w and b for the machine learning model to make correct prediction by reaching Global Minimum of cost function.

$$w2 = w1 - L \cdot \frac{dJ}{dw} \quad (6)$$

$$b2 = b1 - L \cdot \frac{dJ}{db}$$

Figure 3.8 Formula for updating Weight and Bias using Gradient Descent Algorithm

The Gradient Descent will find the global minimum cost function for the machine learning model by changing the weight and bias,after changing the weight and bias the new updated weight and bias will be our best weight and bias when it comes to training and testing the machine learning model.To update the weight and bias the formula will be shown in function (6), w2 is the new weight, w1 is the old weight, b2 is the new bias, b1 is the old bias, L is our machine learning model learning rate, $\frac{dJ}{dw}$ derivative of cost function based on weight (w) and $\frac{dJ}{db}$ is derivative of cost function based on bias (b).This function (6) is how Gradient Descent minimising the Cost Fucntion of our model to Global Minimum by finding the best weight and bias for the machine learning model.

2.4 Support Vector Machine Implementation

The SVM implementation will be conducted after the Gradient Descent Algorithm has been applied to the machine learning model.Support Vector Machine will classify a data point lies to which side of the margin of the hyperplane.

$$\gamma_i = (w.x_i + b) \quad (7)$$

Figure 3.9 Support Vector Machine hyperplane formulas

As shown in the function (7),the Support Vector Machine will make a hyperplane and determine which side is the data point lies to,whether its in the positive or negative side of the margin. γ_i is the data index of a person in a single row and became a label for that data,w is weight and b is bias that has been updated by the Gradient Descent Algorithm and x_i is the data input such as Pregnancies, Glucose, BloodPressure, SkinThickness,

Insulin, BMI, Diabetes, and Age.This function (7) is the formulas of Support Vector Machine that will classify based on the data input whether a person has diabetes or not.

2.5 Predicting Diabetes

Using the Machine Learning model that has been implemented with Support Vector Machine and Gradient Descent the model now can predict whether a person has diabetes or not based on the input data such as : Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, Diabetes, and Age.Inputing the data into the Machine Learning Model will make the model made prediction whether the person has diabetes or not based on which side of the margin the point data lies to.

2.6 Testing Accuracy of Support Vector Machine Model

The testing of the machine learning model will be conduct to test the accuracy of prediction from the machine learning model based on the test data and train data.The bigger percentage of the machine learning accuracy will make the machine learning model more accurate when it comes to predicting whether a person has diabetes or not.

2.7 Model Evaluation

The machine learning model will be evaluated by comparing the Support Vector Machine that used the Gradient Descent Algorithm with the built in Support Vector Machine in python library.

CHAPTER 3 (IMPLEMENTATION AND RESULTS)

3.1 Experiment Setup

This research is conducted using a laptop with specification of AMD A9-9425 Radeon R5 processor with 8 GB of RAM and using Python 3.0 as the programming language to finish the research of Diabetes Prediction.

3.2 Results

After conducting this research on predicting diabetes using support vector machine and testing the accuracy score of the classifier model it is showing a good result from the Support Vector Machine Classifier that are boosted by Gradient Descent Algorithm.Without the Gradient Descent Algorithm the pure SVM only shown 34% and 34,5% accuracy score respectively on Train Data and Test Data with 30% (600) of the dataset (2000) is used in the Test Data.Compared to the SVM library that has > 70% accuracy score the pure SVM models cannot be compared to it and needs the boost of Gradient Descent Algorithm for it to boost the perfomance,when the SVM is boosted using Gradient Descent Algorithm it shown a huge jump in the accuracy score of the models.The accuracy score of the SVM models shown 78% and 77,5% respectively on Train Data and Test Data using the same percentage on the Test Data which is 30% (600).The SVM models that are boosted by Gradient Descent Algorithm is on par with the SVM library in python with more than 70% accuracy score and it shown that the models will most likely predict 7 to 8 out of 10 prediction correct when it comes to Diabetes Disease.Different percentage of train and test data is also conducted in the testing and it shown result as below.

Train & Test	Train 70%	Test 30%	Train 50%	Test 50%	Train 30%	Test 70%
SVM Gradient	78%	77.50%	78.70%	77.20%	79.60%	76.60%
SVM Library	77.20%	77.80%	78.10%	76.90%	80%	76.50%

Table 3.1 Table Accuracy Score

3.3 Discussion

This research shown a good result in predicting whether a person has diabetes or not,it also shows good accuracy score of above 70% when it comes to correct prediction of the datasets based on the Outcome in Y_train and Y_test.The models can have that accuracy because of the predict function that predict the Outcome of the X_train and X_test data and comparing it to the Y_train and Y_test using accuracy_score function,its comparing the X_train and X_test predicted labels of 0 and 1 from the SVM_Classifier.predict function with the Y_train and Y_test that stored all the 0 and 1 values from the datasets of Outcome.The accuracy_score gives a result of 78% for the X_train_data and 77,5% for the X_test_data when comparing it to the Y_train and Y_test.The SVM_Clasiffier models can predict whether a person has diabetes or not from the Support Vector Machine Formulas as shown in Case Study for finding the value of positive or negative using the manually Equation,Elimination and Cross Multiplication to find each w1,w2,w3…w8 and then to

find the b or bias value then to multiply and to addition and subtraction of all the value to find postive or negative value that determines which side of the margin that data point lies to,whether that data point is less than equal to 1 or more than equal to 1 (minus or postive) and gives the data point label of 0 and 1 and then determining whether that person data point is Not Diabetic or Diabetic.

CHAPTER 4 (CONCLUSION)

From this research it can be concluded based on the problem formulation in chapter 1 that Support Vector Machine can determine and classify that a person has diabetes or not because SVM can give either 0 or 1 output by processing the value that been given from the diabetes factors.It can classify either positive or negative and determine a person has diabetes or not.The SVM model also gives a pretty high accuracy rate of more than 70% based on the correct correct prediction that the model gives and comparing it with the True Prediction from the dataset itself.This SVM model that are boosted by Gradient Descent Algorithm also on par with the Built in SVM from python library because of its accuracy that is pretty much the same or above with the score of >70%.To improve this research,deep learning approach can be implemented because this research only focused on the machine learning algorithm which is Support Vector Machine for Diabetes.

DAFTAR PUSTAKA

[1] Mujumdar A, Vaidehi V. Diabetes Prediction using Machine Learning Algorithms. Procedia Computer Science. 2019 Jan 1;165:292–9.

[2] Hasan MdK, Alam MdA, Das D, Hossain E, Hasan M. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. IEEE Access. 2020;8:76516–31.

[3] Mahboob Alam T, Iqbal MA, Ali Y, Wahab A, Ijaz S, Imtiaz Baig T, et al. A model for early prediction of diabetes. Informatics in Medicine Unlocked. 2019 Jan 1;16:100204.

[4] Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. ICT Express. 2021 Dec 1;7(4):432–9.

[5] Ayon SI, Islam MM. Diabetes prediction: a deep learning approach. International Journal of Information Engineering and Electronic Business. 2019 Mar 1;12(2):21.

[6] Yahyaoui A, Jamil A, Rasheed J, Yesiltepe M. A decision support system for diabetes prediction using machine learning and deep learning techniques. In2019 1st International informatics and software engineering conference (UBMYK) 2019 Nov 6 (pp. 1-4). IEEE.

[7] Saru S, Subashree S. Analysis and Prediction of Diabetes Using Machine Learning [Internet]. Rochester, NY; 2019 [cited 2023 Oct 12]. Available from: https://papers.ssrn.com/abstract=3368308

[8] Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM. Classification and prediction of diabetes disease using machine learning paradigm. Health information science and systems. 2020 Dec;8:1-4.

[9] Birjais R, Mourya AK, Chauhan R, Kaur H. Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Appl Sci. 2019 Aug 28;1(9):1112.

[10] Tripathi G, Kumar R. Early prediction of diabetes mellitus using machine learning. In2020 8th international conference on reliability, Infocom technologies and optimization (trends and future directions)(ICRITO) 2020 Jun 4 (pp. 1009-1014). IEEE.