© 2018 IJSRST | Volume 4 | Issue 2 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Diabetes Mellitus Detection using Support Vector Machine

Qamar Rayees Khan[1], Mohammed Asger[2]

[1]Department of Computer Sciences, Baba Ghulam Sahah Badshah University, Rajouri (J&K), India

[2]Department of Computer Sciences, Baba Ghulam Shah Badshah University, Rajouri (J&K), India

ABSTRACT

With the advancement in technology, diagnosis of diseases is performed using Artificial Intelligence and Machine Learning. Diabetes is one of the chronic diseases that is spread all over the world. Diagnosing Diabetes early can save millions of lives. In this paper Support Vector Machine is used to diagnose a patient whether it is diabetic or Non-Diabetic. PIMA dataset is collected the public repository KAGGLE. For training machine learning algorithm 70% of data is used while as 30% is used for testing. The results showed that Support Vector Machine has 86.2% Accuracy with Precision of 83%. In future other machine learning algorithms can be used for performing the said task.

Keywords : Artificial Intelligence, Machine Learning, Support Vector Machine, PIMA

I. INTRODUCTION

Data mining has been evolved as one of the most pleasing areas by which one can predict hidden information. With artificial intelligence development, most of the researchers gained interest in Artificial intelligence and Data Mining. These days Machine learning approaches are used for prediction as well as classification tasks. Machine Learning has shown promising results in every field, be it medical or any other area. The force and adequacy of these methodologies are gotten from proportionate strategies to extricate designs and make models from the data. The previously mentioned reality is incredibly huge in the enormous information period, mainly when the dataset can arrive at terabytes or petabytes of information. Thus, the wealth of knowledge has fortified extensively information arranged science examination. In a particularly crossover field, perhaps the main exploration applications is forecast and analysis identified with human-undermining or potentially life quality lessening illnesses. One such sickness is diabetes mellitus. Diabetes mellitus (D.M.) is one of the most extensive diseases the world has confronted, particularly in developed or developing countries. Diabetes mellitus is a cluster of diseases that has symptoms of hyperglycaemia [1]. The D.M. is a metabolic disease that causes high blood sugar. The hormone insulin transfers sugar from the blood into cells for energy utilization. With Diabetes, the body doesn't make sufficient insulin. The various types of Diabetes are Type 1, Type 2, Prediabetes, and Gestational [2]. Some of the significant symptoms of Diabetes are the increase in Hunger, Feeling Thirsty, Weight Loss, Frequent Urination, Blurry Vision, Extreme Fatigue and Sores that do not Heal. Diagnosing Diabetes as early as possible is the need of an hour because it can lead to various complications. Some of the complications that are faced due to Diabetes are Heart attack, hearing loss, neuropathy, and depression.

Applying machine learning and data mining methods in D.M. research is crucial for the utilization of large volume diabetes-related data for extracting knowledge. The severe social impact of the specific disease renders D.M. one of the main priorities in medical science research, which inevitably generates vast amounts of data [3]. Undoubtedly, machine learning and data mining approaches in D.M. area are of great concern for diagnosis, management and other related clinical administration aspects. In this work, the machine learning algorithm, Support Vector Machine, is used to perform binary classification that whether the patient is diabetic or non-diabetic. The paper consists of V sections. Section I provides the basic Introduction about Diabetes, its symptoms and complications. In section II, a detailed background is being provided about the work done in this area. Section III discusses the proposed Methodology for performing the binary classification. Section IV provides the results, and Section V concludes the work.

II. Literature Review

Diabetes Mellitus is an undeniable common illness that occurs due to our body's inability to produce glucose [4]. Discovering the sickness at the early stage lessens clinical expenses. The danger of patients having more complicated wellbeing problems is an undeniably common persistent sickness described by the body's powerlessness to process glucose. Discovering the infection at the early stage decreases clinical expenses and patients' danger of more complicated medical conditions. Wilson et al. [5] developed a D.M. prediction model that predicts middle-aged American adults' risk. Logistic Regression is being used for performing this task. The factors considered to predict the risk were parental history of D.M., obesity, high blood pressure, low levels of high-density lipoprotein cholesterol, elevated triglyceride levels, and impaired fasting glucose. About 3140 Samples were taken, and the results showed that the developed model gave an accuracy of 85%. The evaluation of the proposed algorithm was performed by [11] on the Canadian population in which about 4403 samples were taken, and after performing experiments, the algorithm showed 78.6% accuracy. [12] used various machine learning algorithms to predict prediabetes. Some of the algorithms used in [12] work are Decision Tree, Logistic Regression and Artificial Neural Network. In their work, there were about 735 patients who were diagnosed as Diabetic and 752 as Non-diabetic. The results showed that the decision tree outperformed than other two algorithms by showing an accuracy of 77.3%. [6,7,8,9,10,11] used Naïve Bayes algorithm, K- Nearest Neighbour and Random Forest, to perform the classification task of Diabetes and Non-Diabetes. The results showed Naïve Bayes showed better results than other algorithms. In this work, we are using a Support Vector Machine for performing binary classification.

III. Methodology

The proposed methodology consists of five phases it starts with data collection, Pre-processing, Feature Selection, and Classification. Figure 1 shows the basic architecture of the proposed methodology.

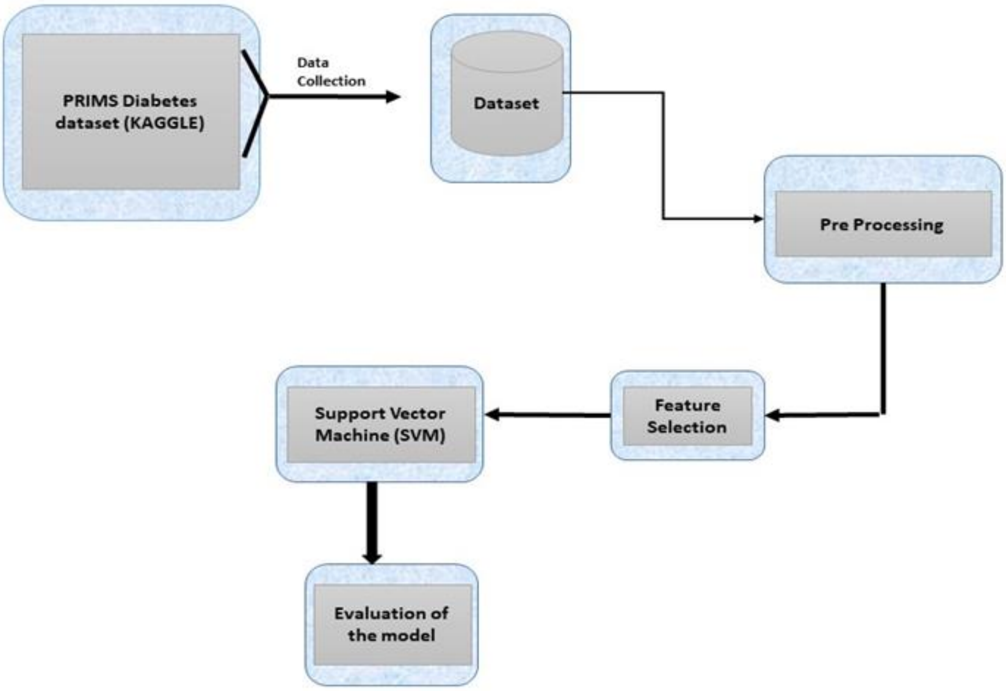
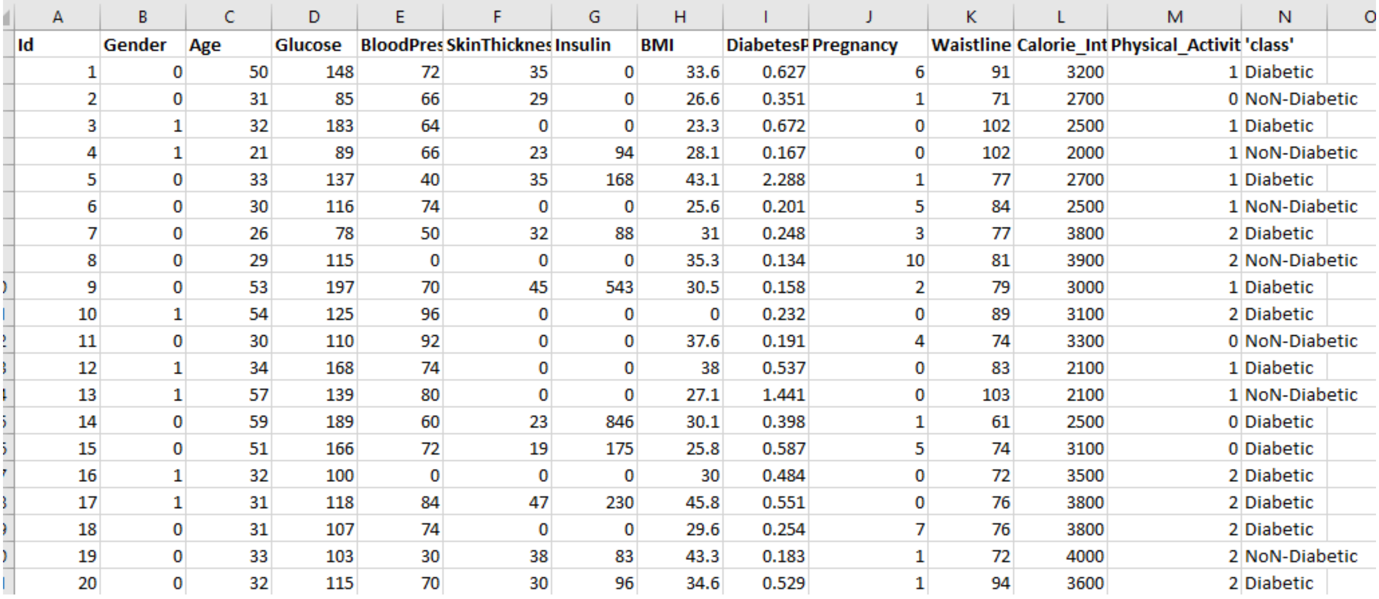


Fig.1. Proposed Methodology

Data Collection

For performing machine learning the first and most important step is the availability of Dataset. We collected PIMA dataset from a public data repository KAGGLE. More attributes are added in our dataset. Figure 2. Is the Screenshot of the dataset. The data set consists of 14 attributes, in which 14th attribute is the class label.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Id	Gender	Age	Glucose	BloodPres	SkinThicknes	Insulin	BMI	DiabetesP	Pregnancy	Waistline	Calorie_Int	Physical_Activit	'class'	
1	0	50	148	72	35	0	33.6	0.627	6	91	3200	1	Diabetic	
2	0	31	85	66	29	0	26.6	0.351	1	71	2700	0	NoN-Diabetic	
3	1	32	183	64	0	0	23.3	0.672	0	102	2500	1	Diabetic	
4	1	21	89	66	23	94	28.1	0.167	0	102	2000	1	NoN-Diabetic	
5	0	33	137	40	35	168	43.1	2.288	1	77	2700	1	Diabetic	
6	0	30	116	74	0	0	25.6	0.201	5	84	2500	1	NoN-Diabetic	
7	0	26	78	50	32	88	31	0.248	3	77	3800	2	Diabetic	
8	0	29	115	0	0	0	35.3	0.134	10	81	3900	2	NoN-Diabetic	
9	0	53	197	70	45	543	30.5	0.158	2	79	3000	1	Diabetic	
10	1	54	125	96	0	0	0	0.232	0	89	3100	2	Diabetic	
11	0	30	110	92	0	0	37.6	0.191	4	74	3300	0	NoN-Diabetic	
12	1	34	168	74	0	0	38	0.537	0	83	2100	1	Diabetic	
13	1	57	139	80	0	0	27.1	1.441	0	103	2100	1	NoN-Diabetic	
14	0	59	189	60	23	846	30.1	0.398	1	61	2500	0	Diabetic	
15	0	51	166	72	19	175	25.8	0.587	5	74	3100	0	Diabetic	
16	1	32	100	0	0	0	30	0.484	0	72	3500	2	Diabetic	
17	1	31	118	84	47	230	45.8	0.551	0	76	3800	2	Diabetic	
18	0	31	107	74	0	0	29.6	0.254	7	76	3800	2	Diabetic	
19	0	33	103	30	38	83	43.3	0.183	1	72	4000	2	NoN-Diabetic	
20	0	32	115	70	30	96	34.6	0.529	1	94	3600	2	Diabetic	

Fig.2. Screenshot of Dataset

Data Preprocessing

The data that is collected consists many missing values, Null Values and Noise. To overcome these hindrances preprocessing of the data is being performed. In this step average value is being put in place of Missing and NULL values. The redundancy of data is being removed by removing the duplicates in the dataset.

Feature Selection

The Dataset consists of 14 different features which are as follows:
Id,Gender,Age,Glucose,BloodPressure,SkinThickness,Insulin,BMI,DiabetesPedigreeFunction,Pregnancy,Waistline,Calorie_Intake,Physical_Activity_Level and 'class'. The last attribute 'Class' represents the label (Diabetic/Non-Diabetic).

Classification

Support Vector Machine is fine-tuned for performing this binary classification. The SVM is Fine Tuned as:
SV type: eps-svr, epsilon = 0.1, cost C = 1, Sigma = 0.0520722411952879, Number of Support Vectors : 574,Objective Function Value : -319.0824, Training error : 0.513793, Laplace distr. width : 0.338132

Results

For performing experiments/implementation Workstation of Configuration 8GB RAM and 1TB hard Drive was used. The dataset was first splitted into 60:40 that is 60% of data was used for training the machine learning model and 40% of data was used for testing. After evaluating the performance it was concluded that the model showed 70% of Accuracy. To improve the accuracy of the proposed model data was split into 70:30, 70% of data is used for training the Support Vector Machine and 30% is used to test the model. After evaluating the model the accuracy of 86% is achieved. Table I shows the Classification Report of our Support Vector Machine (SVM).

Table I. Classification Report of SVM in percentage

Precision	Recall	F-Score	Accuracy
83	65	72.9	86.2

After generating the classification report it was found that the proposed model showed precision of 83%, Recall of 65%, F-Score of 72.9% and overall accuracy of 86.2%.

Figure 3 shows the comparison of various evaluation metrics like True Positive, False Positive, F-Score, Precision, Recall etc.

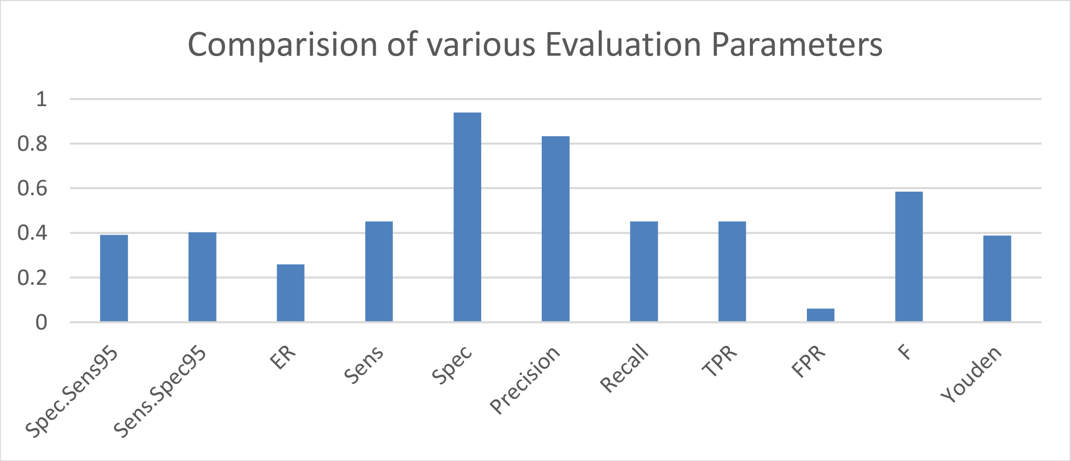


Fig.3. Comparison of Various Parameters.

Figure 4, shows the Evaluation report of Support Vector Machine Graphically

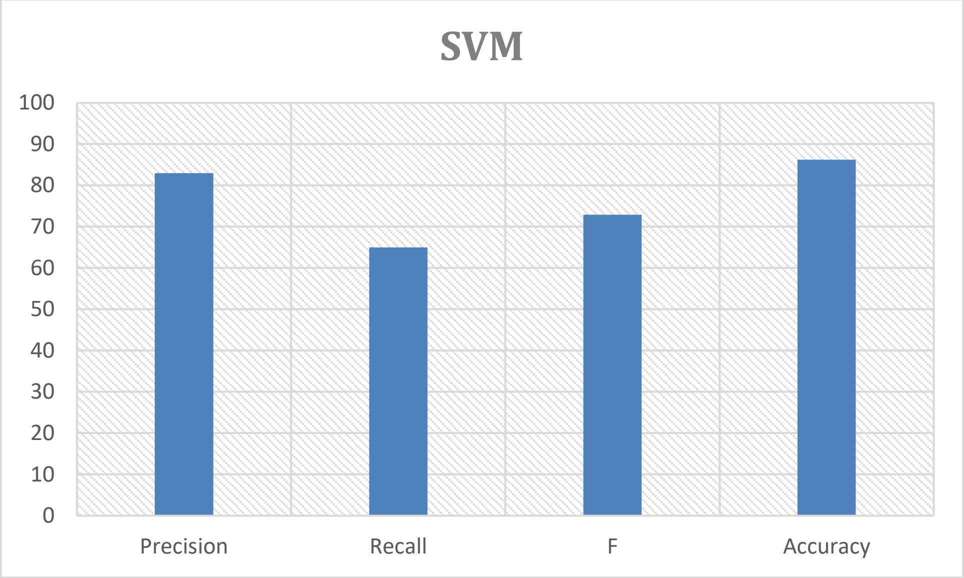


Fig. 4. Evaluation Report of SVM

IV. CONCLUSION

Diabetes Mellitus is one of the chronic disease which is spreading at enormous pace around the globe. In this work, Support Vector Machine is used to diagnose a patient whether he/she is Diabetic or non-diabetic. PIMA Dataset is collected from a public data repository KAGGLE. Data Preprocessing and Feature Selection is performed using various techniques. 70% of data is used for training the classifier and 30% is used for testing the classifier. The results showed the accuracy of 86.2% with the precision of 83%. More data can be supplied to the proposed model for improving the performance in terms of Precision and Recall. In future other Machine Learning algorithms may be used to perform this binary classification.

V. REFERENCES

[1]. Griffin, M. E., Coffey, M., Johnson, H., Scanlon, P., Foley, M., Stronge, J., ... & Firth, R. G. (2000). Universal vs. risk factor-based screening for gestational diabetes mellitus: detection rates, gestation at diagnosis and outcome. Diabetic Medicine, 17(1), 26-32.

[2]. Harris, M. I., & Eastman, R. C. (2000). Early detection of undiagnosed diabetes mellitus: a US perspective. Diabetes/metabolism research and reviews, 16(4), 230-236.

[3]. Zhang, B., & Zhang, D. (2013). Noninvasive diabetes mellitus detection using facial block color with a sparse representation classifier. IEEE transactions on biomedical engineering, 61(4), 1027-1033.

[4]. Kreis, R., & Ross, B. D. (1992). Cerebral metabolic disturbances in patients with subacute and chronic diabetes mellitus: detection with proton MR spectroscopy. Radiology, 184(1), 123-130.

[5]. Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, et al. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham offspring study. Arch Intern Med. 2007;167:1068–74

[6]. Iyer A, Jeyalatha S, Sumbaly R. Diagnosis of Diabetes using classification mining techniques. Int J Data Min Knowl Manage Process (IJDKP). 2015; 5(1):1–14.

[7]. . Ioannis K, Olga T, Athanasios S, Nicos M, et al. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15: 104–16.

[8]. Jayalakshmi T, Santhakumaran A. A novel classification method for diagnosis of diabetes mellitus using artificial neural networks, International conference on data storage and data engineering, India; 2010. p. 159–63.

[9]. Kahn HS, Cheng YJ, Thompson TJ, Imperatore G, Gregg EW. Two riskscoring systems for predicting incident diabetes mellitus in U.S. adults age 45 to 64 years. Ann Intern Med. 2009;150:741–51.

[10]. Kandhasamy JP, Balamurali S. Performance analysis of classifier models to predict diabetes mellitus. Procedia Comput Sci. 2015;47:45–51.

[11]. Mashayekhi M, Prescod F, Shah B, Dong L, Keshavjee K, Guergachi A. Evaluating the performance of the Framingham diabetes risk scoring model in Canadian electronic medical records. Can J Diabetes. 2015;39(30):152–6.

[12]. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors.