Classification of Diabetes Mellitus Based on Machine Learning Classifier Techniques

Pydipala Laxmikanth, Bhramaramba Ravi

Abstract: ***In recent days diabetes is recorded as fastest growingst common disease among several diseases in the world. It became one of the major health problems in several states and countries. This occurs mainly when the normal human body is incapable to produce the sufficient amount of insulin in order to adjust the quantity of sugar levels in the body. This improper maintenance of sugar levels may lead to other diseases like heart disease, kidney disease, blindness, nerve damage and blood vessels damage. every one knows that there are mainly two general reasons for diabetes: One reason is the pancreatic gland not able make sufficient insulin or the body not produce make enough insulin. This type of symptom is mainly found on 5-10 % of citizens with diabetes and they come under Type-1 Diabetes. Another reason is cells do not respond to the insulin that is produced and this type of symptom people come under Type-2 Diabetes. In recent days, machine learning as well as DM techniques have been considered to design automatic diagnosis system for diabetes. In this proposed paper we aim to use the SVM, a ML method as a classifier for identification of diabetes data set. Here we applied the data cleaning techniques to handle the incomplete by fill in absent values, smoothening the noisy data, identified and removed the inconsistencies. By performing the data cleaning activities for verifying all the fields in the data set are properly arranged or not. Once after the data cleaning is completed then we try to apply SVM and Naive Bayes for classifying the diabetes dataset and we try to compare the both classifiers and find which classification techniques are efficient by comparing the both classifiers based on the results we observed. Our experimental results clearly tell that SVM can be effectively used for identifying diabetes disease. In this proposed application we try to analyze the diabetes based on location wise i.e. diabetes is differently affected by various people in and around the various corners of the city. In this proposed application we take sample data set of Visakhapatnam city with four area's data like NORTH, SOUTH, and WEST & EAST. So by applying the SVM, we will try to analyze the different reasons which differ for each and every diabetes patients based on location wise. some area people mostly suffer with food, pollution, eating habits, daily habits, sleeping habits and other reasons. So based on each and every individual region diabetes differ one with other person in and around the city. So we are going to classify a set of patients data based on region wise and classify the cause of affecting diabetes for them and try to provide a counter measure and pre-cautions for the patients.***

Index Terms**: Data Mining, SVM, Diabetes, Naïve Bayesian Classification, Diagnosis**

I. INTRODUCTION

Diabetes is one of the fast-growing diseases. It is an unceasing disease and it can be termed as Diabetes mellitus (DM). Insulin is a hormone which regulates the blood glucose in the human body. The pancreatic gland was unable to release insulin or the released insulin not able to utilize properly by the body. Due to this, a person can be affected by diabetic disease are categorized as type1, type 2, gestational predicting. According to the latest statistics released by WHO. It is given 8th rank globally[1][2].In the South-East Asia region, it is ranked 9th.In India, a recent survey done between 2015 and 2019 noted the diabetes disease prevalence is 11.8% in all ages. According to the WHO estimated 72.96 million cases noted in India. Especially in South India it is very rapidly growing [3][4]. In Andhra Pradesh and its districts, the prevalence is alarming. One of the largest districts of Andrapradesh in Visakhapatnam. In the district Visakhapatnam having a population of 4.3 million both rural and urban areas. This city having major industries like steel plant, port trust, HPCL etc. Presently the city facing pollution. Recently the survey proved that the citizens of Visakhapatnam facing several health issues in which diabetes are one among the diseases[3][4]. It is mostly affecting the children's and aged people. In 2014 AMC and KGH established Diabetes child society to deal with this disease.

With an increased rate the people affected diabetes disease. In Visakhapatnam, tons of diabetic data was generating in private and government health centres. Requirement for computer-based analysis to predict diabetes, prediabetes of the citizens of Visakhapatnam. Early detection of the disease can reduce the effect and also can handle the Bayes secondary diseases related to diabetes.

Here, we are using a machine learning, the supervised learning support vector machine algorithm along with another machine learning algorithm Naïve Bayes classifier also called probabilistic classifier used for classifying and probabilistic to predetermine the patients diabetic, pre-diabetic or not. The collected patient's data from various health centres of Visakhapatnam. it can be analyzed by using the two classifications among the two algorithms.

The remaining paper is organized as Section –II dataset description, Section –III Navie Bayes Implementation, Section-iv SVM implementation, Section – V experimentation results. Section –vi conclusion and future work.

Revised Manuscript Received on January 27, 2020

Pydipala Laxmikanth, Pursuing Ph.D Degree in GIT, GITAM DEEMED TO BE UNIVERSITY, Visakhapatnam

Dr. Bhramaramba Ravi, Professor, Dept. of CSE, GIT, GITAM, Visakhapatnam.

II. DATASET DESCRIPTION

Dataset consists of 500 hundred samples collected from the various health centers across the Visakhapatnam city. While collecting the samples, taken the permission and maintain the privacy and anonymity for patient's data. The attributes of the dataset are Gender (Male or Female), Age(age of the patient), BMI (Body mass index), Skin Thickness(skin fold thickness), Blood sugar value (glucose levels), Postprandial Blood Glucose (glucose level after 2 hours), HBA1C% (average level of blood sugar over the past 2 to 3 months), Diabetes pedigree function (a *function* which gives chance of getting *diabetes* based on family history), Outcome (indicates the class variable either 0 or 1, 0 indicates no diabetes and 1 indicates diabetes). The boundary values of the attributes are mentioned in the table below.

Table 1. Description of the Data set and their Boundary values:

Attribute Name	Description of the Attribute	Boundary values
Gender	Patients Male or Female	
AGE	Person Age	
BMI	Body mass index	15 – 60 BMI (kg/m^2)
Skin Thickness	Skin fold thickness of the upper arm	Varies for men and women, age
Blood sugar values	Glucose levels	Normal- 70-99 mg/dl, (3.9 to 5.5 mmol/L) With diabetes -80-130 mg/dl, (4.4 to 7.2 mmol/L)
Postprandial Blood Glucose	Glucose level after 2 hours	Normal <140 mg/dl, (7.8 mmol/L). With diabetes <140 mg/dl(7.8 mmol/L).
HBA1C%	Average level of blood sugar over the past 2 to 3 months	Normal - < 5.7%, With diabetes <7.0%
Diabetes pedigree function	a *function* which gives chance of getting *diabetes* based on family history	
Outcome	(indicates the class variable either 0 or 1, 0 indicates no diabetes and 1 indicates diabetes)	0 indicates no diabetes, 1 indicates diabetes

III. NAIVE BAYESIAN CLASSIFICATION

Navie Bayes is used for building classifiers. it is also called Bayesian classifier. Generally it follows the conditional independence based on the assumption [5]. Generally, Bayesian classifiers are considered the best classifier when the positive assumption holds true. In most cases it is proved and Bayesian classifier believes that for any given class the effect of an attribute values on that class is independent on other attribute values. Once the model is build us which classify and assigns class labels based on the feature of the attribute[6]. All the records are classified accordingly within the predetermined class labels based on features then the classification happened otherwise the unseen record needs to assign a new class label. It is mainly used for predicting the problems.
It works as given below:

1. Each and every data sample is represented for an n-dimensional attribute vector,
   X=(x1, x2, x3,……,xn)
   --- (1)
2. The proposed modeled classifier

Based on the conditions the Bayesian classifier assigns a class label to the unseen record and also predicts that the variable belongs to highest posterior probability. even though there are m classes. the below is the condition

$$P(Ci/X) > P(Cj/X) \quad \text{for } 1 \le j \le m, j \ne i. \qquad \text{--- (2)}$$

for the given class Ci for which P(Ci/X) will be the maximum posterior hypothesis. By the Bayes theorm : In our example, the probability that a given record should be classifies in the class C, can be formulated as

$$P(Ci/X) = \frac{P(X/Ci)\,P(Ci)}{P(X)} \qquad \text{--- (3)}$$

IV. PROPOSED DIABETES DISEASE PREDICTION USING SVM CLASSIFICATION ALGORITHM

In this section we will mainly discuss about the proposed disease prediction based on SVM classification algorithm.

MOTIVATION

SVM algorithm is a set of related supervised learning method used in medical diagnosis for classification and regression [7]. It is mainly used in minimizing the most common classification errors that occur in the data set and try to utilize the geometric margin, that is the reason why SVM is also known as Maximum Margin Classifiers(MMC). This is a general algorithm based on the risk bounds of statistical learning theory i.e. the so called structural risk minimization principle[8][9]. This proposed algorithm can effectively perform non-linear classification using a principle like kernel trick, by mapping all the inputs implicitly to the high dimensional vector spaces. The use of kernel trick mainly helps the end user to construct the classifier without having the knowledge about the feature space explicitly[10][11]. For example, we try to take a set of points belonging to either one of the two classes, an SVM classifier try to find out the hyper plane which is having the largest possible fraction of points of the same class on the same plane [12][13].This hyper plane which is used for separating the two fractions is known as optimal separating hyper plane (OSH),

which is used to maximize the distance between the two parallel hyper planes and can minimize the risk of misclassifying examples of the test dataset

For example:

Initially we try to take a labeled training data as data points of the form

$$M=\{(x1,y1),\ldots\ldots,(xn,yn)\} \quad \text{--(4)}$$

Where yn = 1/-1, a constant that denotes the class to which the point xn is mapped to

n = Represents the number of samples taken

xn = reprsents the real time dimensional vector

The SVM classifier first maps the input vectors into a decision value, and then performs the classification using an appropriate threshold value.

V. PROPOSED ARCHITECTURE

In the Proposed architecture we considered two machine learning namely SVM and Navies Bayes. The collected instances divided into two parts training and test data and build the model of classifiers using Matlab and all the records under gone the classification activity by the both classifiers. And compared the performance of the two classifiers based on the accuracy levels.

Table 2: Representing the attribute considered and their values

GENDER	AGE	BMI	SKIN THICKNESS	BLOOD SUGAR VALUE(mg/dl)	POSTPRANDIAL BLOOD GLUCOSE	HBA1C%	DPF	OUTCOME
F	46	33.6	35	219	323	7.6	0.627	1
M	52	26.6	29	480	593	12.6	0.351	1
M	50	23.3	25	109	112	6.2	0.672	1
M	58	28.1	23	165	337	7.8	0.167	1
M	52	43.1	35	149	186	7	2.288	1
M	50	25.6	33	143	295	8	0.201	1
M	49	31	32	173	313	8.6	0.248	1
M	49	35.3	22	176	313	8.6	0.134	1
M	65	30.5	45	100	142	6.3	0.158	1
M	66	0	25	189	391	9.3	0.232	1
M	47	37.6	28	107	216	6.4	0.191	1
M	55	38	34	220	314	7.4	0.537	1
M	49	27.1	33	285	316	9.6	1.441	1
F	44	30.1	23	209	338	9.5	0.398	1
F	53	25.8	19	169	328	7.7	0.587	1
F	64	30	25	136	159	7.1	0.484	1
M	64	45.8	47	115	152	5.8	0.551	1
F	33	29.6	31	124	153	6.2	0.254	1
M	52	43.3	38	107	237	7.6	0.183	1
F	24	34.6	30	237	300	7.5	0.529	1
M	64	39.3	41	218	398	12.2	0.704	1
M	57	35.4	32	78	135	5.6	0.388	0
F	27	39.8	34	82	111	5.2	0.451	0
M	60	29	35	125	201	6.6	0.263	1
M	44	36.6	33	140	213	9	0.254	1
M	39	31.1	26	121	190	7.4	0.205	1
M	47	39.4	28	219	318	7.6	0.257	1
F	65	23.2	15	191	326	9.6	0.487	1
F	56	22.2	19	131	166	7.5	0.245	1
M	50	34.1	29	145	314	6.5	0.337	1
F	38	36	30	90	102	5.5	0.546	0
F	33	31.6	36	153	187	7.8	0.851	1
F	50	24.8	11	302	423	13.5	0.267	1
M	54	19.9	22	216	346	10.5	0.188	1
M	40	27.6	31	290	452	12.2	0.512	1
M	36	24	33	153	321	7.6	0.966	1
F	75	33.2	25	127	193	6.2	0.42	1
F	50	32.9	37	155	198	8	0.665	1
M	57	38.2	42	144	205	8.3	0.503	1
M	57	37.1	47	144	205	8.3	1.39	1
M	48	34	25	291	368	9.2	0.271	1
M	66	40.2	33	97	125	5.4	0.696	0
F	65	22.7	18	234	274	8.2	0.235	1
F	55	45.4	24	119	218	6.9	0.721	1
M	41	27.4	29	142	218	8.6	0.294	1
M	53	42	39	141	244	9.1	1.893	1
F	55	29.7	40	117	152	6.1	0.564	1
F	62	28	27	230	400	11.6	0.586	1
M	52	39.1	32	269	413	10.9	0.344	1
M	45	42	35	119	216	7.2	0.305	1
M	72	19.4	11	175	269	7.3	0.491	1
M	40	24.2	15	134	254	7.6	0.526	1
M	55	24.4	21	141	173	6.6	0.342	1
M	38	33.7	34	155	286	7.9	0.467	1
F	43	34.7	42	109	168	8.5	0.718	1
M	37	23	10	179	218	8.7	0.248	1
M	28	37.7	39	190	251	10.9	0.254	1
	38	46.8	60	334	401	20.2	0.962	1

Table 3: Representing the classification of Diabetes based Gender

GENDER	Total	DIABETES AFFECTED	DIABETES NOT AFFECTED
MALE	315	295	20
FEMALE	185	170	15
TOTAL	500	465	35

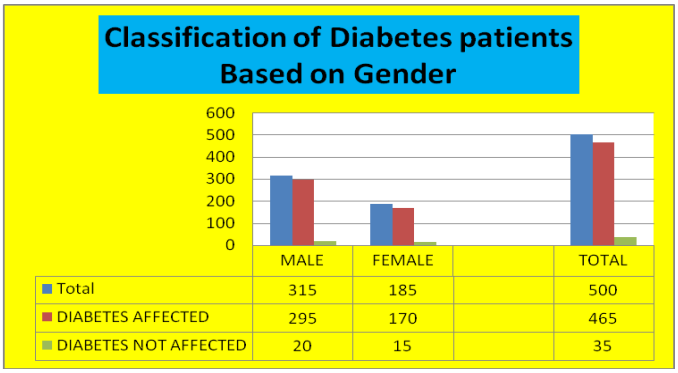

Fig 1. Figure represents the statistics of Diabetes Patients Based on Gender

VI. EXPERIMENT RESULTS

Before building the model, analyzing the data and performing the classification by using the sample data by using methodology developed. The reliability of the classification can be measured by the outcome of the two methods SVM and naive Bayes. if both the classifiers were done the classification then the record resulted as positive or negative. The results displayed here is the mean value of the Navie Bayes classifier the accuracy is 96% and the SVM classifier accuracy is 98.4%. This witness the high reliability and stability of classifiers.

Table 4 Classification Accuracy Using Svm And Naive Bayes

Classifier	No of Samples	Training Data	Test Data	Attributes	No.of Samples correctly classified	No of Samples Incorrectly classified	Classification Performance Mean Value
Naïve Bayes	500	300	200	9	480	20	96.0%
SVM	500	300	200	9	492	8	98.4%

The performance of the classification is also predicted in terms of other multi-class classification such as recall and precision. The two measures are calculated with the below two formulas.

The following are the Out Come which we can observer after performing SVM classification on these two data sets.

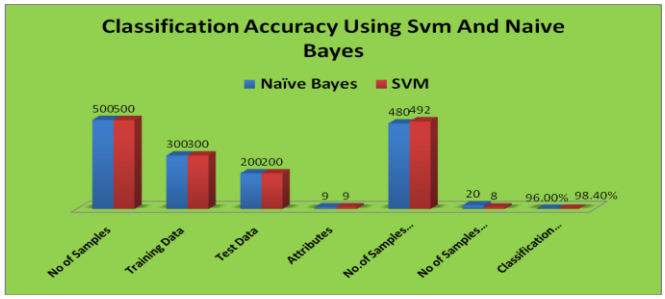

Fig 2. Classification Accuracy Using SVM and NAÏVE Bayes algorithms

The performance of the classification is also predicted in terms of other multi-class classification such as recall and precision. The two measures are calculated with the below two formulas.

The following are the Out Come which we can observer after performing SVM classification on these two data sets.

$$\text{Precision}_y = \frac{TPy}{TPy + FPy} \quad \text{--(5)}$$

$$\text{Recall}_y = \frac{TPy}{TPy + FNy} \quad \text{--(6)}$$

Table 5. Precession and Recall

Classifier	Recall	Precesion	Recall	Precesion
Naïve Bayes	0.941	0.971	0.923	0.813
SVM	0.965	0.96	0.877	0.893

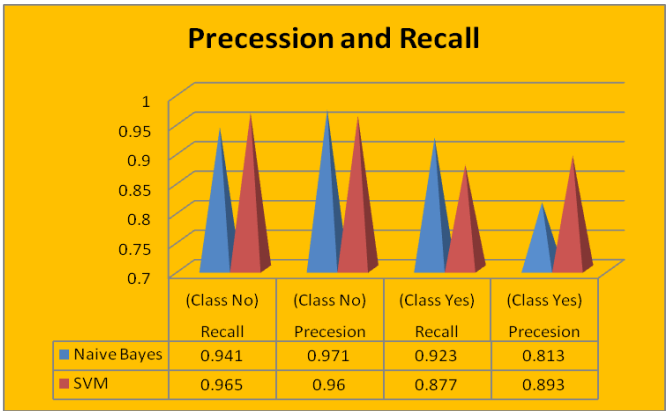

Fig 3. Precession and Recall

VII. CONCLUSION

In this paper, we used to find the best accuracy of classification by using the machine learning algorithms SVM and Naive Bayes on the diabetes data set and find the accurate result by taking a sample of several patient records. In this proposed application we take sample data set of Visakhapatnam city with four area's data like NORTH, SOUTH, and WEST & EAST. So by applying the SVM and Naive Bayes, we proved a cause and report which differ for every diabetes patients based on location wise.

REFERENCES

1. D. Deng and N. Kasabov, " On-line pattern analysis by evolving self-organizing maps", In Proceedings of the fifth biannual conference on artificial neural networks and expert systems (ANNES), 2001, pp. 46-51.
2. Balakrishnan Sarojini, Narayanasamy Ramaraj and Savarimuthu Nickolas, "Enhancing the Performance of LibSVM Classifier by kernel F-Score Feature Selection", Contemporary Computing, 2009, Volume 40, Part 10, pp. 533-543.
3. Yue, et al. " An Intelligent Diagnosis to Type 2 Diabetes Based on QPSO Algorithm and WLSSVM," International Symposium on Intelligent Information Technology Application Workshops, IEEE Computer Society, 2008.
4. Christopher J.C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery", Springer, 2(2), pp.121-167, 1998. [9] T.Mitchell, Machine Learning, McGrawHill, New York, 1997.
5. Quinlan, J.R. "C4.5: programs for machine learning", San Mateo, Calif., Morgan Kaufmann Publishers, 1993.
6. Van Gerven M. A.J., Jurgelenaite R., Taal B. G., Heskes., Lucas P. J.F., "Predicting carcinoid heart disease with the noisythreshold classifier". Artificial Intelligence in Medicine, 2007, vol.40, 45-55.
7. Smith, J.W., J. E. Everhart, et al.- "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus", Proceedings of the Symposium on Computer Applications and Medical Care (Washington, DC). R.A. Greenes. Los Angeles, CA, IEEE Computer Society Press, 1988, pp. 261-265.
8. S.Sahan, K.Polat, H. Kodaz, and S. Gunes, " The medical applications of attribute weighted artificial immune system (awais): Diagnosis of heart and diabetes diseases", in ICARIS, 2005,p. 456-468.
9. I.Tsoulos, D. Gavrilis, E. Glavas,- "Neural network construction and training using grammatical evolution", Science Direct Neurocomputing Journal, Vol.72, Issues 1- 3, December 2008,pp. 269-277.
10. V. Vapnik, "The Nature of Statistical Learning Theory." NY: Springer-Verlag. 1995.
11. Park, J. and Sandberg, I. W., "Universal approximation using radial basis function networks", Neural Computation, vol.3, pp.246-257, 1991.

12. P. Venkatesan and S. Anitha, "Application of a radial basis function neural network for diagnosis of diabetes mellitus" , Current Science, vol. 91, no.9,pp.1195- 1199, 10 November 2006.
13. UCI repository of bioinformatics Databases, Website: http://www.ics.uci.edu/~mlearn/MLReposit ory.html.

AUTHORS PROFILE

P.Laxmikanth, is pursing Ph.D from GIT, GITAM DEEMED TO BE UNIVERSITY, and Visakhapatnam. He is received M.Tech from GITAM COLLEGE OF ENGINEERING, affiliated to Andhra University, Visakhapatnam, India in 2007 and M.Sc from SIR C.R.Reddy P.G.college Eluru, affiliated to Andhra University, India. He has published 10 papers in various reputable national/International journals and conferences. He is an active member of CSI .His research interest are Data mining, image processing.

Dr. Bhramaramba Ravi, obtained her Ph.D from JNTUH in the year 2011. She has about 19 yrs of experience in engineering colleges and is currently working as a professor in the Dept. of CSE, GIT, GITAM, and Visakhapatnam. She has about 32 publications in reputed Journals. Her area of interest is Data Mining and Bioinformatics.