IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.6, Issue.10, Oct. 2018 E-ISSN: 2347-2693

Diabetes Mellitus Prediction on Obese Adult Ladies Using Data Mining Techniques

Manumol Thomas[1*], Chandra J[2]

[1]Dept. of Computer Science, Christ University, Bengaluru, India
[2]Dept. of Computer Science, Christ University, Bengaluru, India

[*]*Corresponding Author:* manumol.thomas22@gmail.com

Available online at: www.ijcseonline.org

Accepted: 10/Oct/2018, Published: 31/Oct/2018

Abstract—The research deals with prediction of Diabetes Mellitus in Pregnant and Non pregnant obese adult ladies using the data mining classification algorithms such as J48, Naïve Bayes, and SVM. Comparative study of classification algorithm is done by using Naïve Bayes, J48 and SVM. The results obtained by J48 and Naïve Bayes were found to be more satisfactory when compared to SVM classification. Naïve Bayes and J48 can be used as a predictor classifier for doing the diabetes mellitus prediction.

Keywords-Medical data mining, Support Vector Machine, J48, Naïve Bayes, Diabetes mellitus

I. INTRODUCTION

Diabetes is a very severe health issue, when the level of blood glucose becomes unregulated. Glucose acts as fuel of human body. When the body uses glucose as a fuel, insulin is essential to get the glucose into cells. Diabetes happens because either production of insulin is insufficient or the cells don't react accurately to insulin or both. The symptoms are increased thirst and hunger [8]. There are three types of diabetes mellitus. Type 1 Diabetes is known as "Insulin dependent diabetes mellitus" (IDDM). It is also known as "Juvenile Diabetes". IDDM is very common among children. Type 1 diabetes causes the body's failure to produce sufficient insulin. Type 2 Diabetes known as "Non-insulin dependent diabetes mellitus" or "Adult onset diabetes" is a condition caused by producing insufficient insulin to the body demands and it does not respond to the same. Type 3 Diabetes is gestational diabetes is caused by the development of high blood glucose level during pregnancy due to undiagnosed diabetes [1].

In general Diabetes is seen in all kinds of age group so it is often known as queen of all diseases. The research focuses on identifying standard set of parameters that directly contribute to diabetes and selecting an efficient algorithm would greatly aide patients to identify their possibility of getting diabetes without the need for doctors or other expensive equipment.

Data mining is process of identifying valuable information from large amounts of heterogeneous data and identifying and determining patterns and rules. Currently health organizations produce huge amount of data mainly in the area of cancer which are very difficult to analyses. Medical data mining helps extract hidden patterns, thereby opening the door to an enormous source of analysis of medical data including classification, clustering, regression, and so on. Data mining Classification is an information technique which is used to do lot of prediction. Classification is a technique which is every so often used for predicting valuable patterns and it helps to reduce the time required to identify these patterns. It also helps to extract some efficient rules from the proposed dataset [3].

Prediction on diabetes can be done with the help of three classification algorithms such as J48, SVM and Naïve Bayes. J48 builds a classification model by using Decision Tree from a set of labeled training data with the concept of information entropy. The Naïve Bayes Algorithm is based on conditional probabilities [4].

Few studies have focused on comparing data mining classifiers such as Support vector machine (SVM), Regression, BayesNet, Naïve Bayes and Decision Table for the grouping of diabetic patient dataset. Data mining techniques is utilized as a part of healthcare field for diagnosis of diabetes and treatment [1].The healthcare associations use Data mining techniques like classification, clustering and association to build their ability for decision making regarding patient health. A patient classified as "high risk" and "low risk" on the basis of the seriousness of the sickness[4].The researchers' concentrates on recognizing the

© 2018, IJCSE All Rights Reserved

best classification algorithm for the combining of data that functions better on various data sets and they have observed the correctness of the tool may vary depending on the data set has been used. The results of the few studies shows that the execution of a classifier depends upon the data set, especially when there are number of attributes which are used in the data set [6].Most of the researchers have done comparison of classification algorithms by using different data set. Researchers highlight the application challenges and future issues of data mining in the field of health and care filed and Data mining methods are also used in management of healthcare for diagnosis and treatment, Healthcare resource management and Customer relationship Management. According to the author's experimental work it is found that the Naïve Bayes and decision tree are the best and efficient Data mining classifiers.

The remaining of the paper is formulated as follows; Section 2 discusses the methodology of research, Brief description about dataset and classification algorithms. In section 3 Result and discussion compares the accuracy of three different classification algorithms with PIMA Indian dataset. Finally the paper concludes in section 4.

II. METHODOLOGY

The Decision Tree and Naïve Bayes are the most popular classifiers in the field of Data mining. The current research focuses on Diabetes Mellitus prediction in Pregnant and Non pregnant obese adult ladies using the data mining classifiers such as J48, Naïve Bayes, and SVM. So these different classification algorithms were taken into consideration for experimental comparison.

2.1. Decision Tree (DT)

DT classifier is an efficient Data mining classification algorithm. Quinlan's ID3, C4.5, J48 and CART are some of the popular DT algorithms. As the name suggests DT is a tree which helps to make powerful decisions and it consists of root node, branches and leaf nodes. Internal node contains questions. Each branch denotes the result of a test and each leaves represents decision. The main aim of applying decision tree is to anticipate the estimation of target attribute based on input values. Usually tree creates from top to bottom.

To implement DT classifier there are different steps involved: they are,

1. Given dataset S, select an attribute as class Label (Dependent variable) to splitting tuples in partitions.
2. To decide a splitting criterion to generate a partition in which all tuples have a place with a single class and select best split to create a node
3. Repeat above steps until complete tree is grown.

DT Classifier helps to extract IF-THEN rules, thus it is also known as rule based classifier.

2.2. J48 Algorithm

The algorithm J48 has two inputs T and A where T is the training record and A is the Attributes. It expands the leaf node until criteria is met and works recursively selecting the best node to split the data.

Step 1 : BuildTree (T, A)
Step 2 : IF End_Cond (T,A)=false THEN
Step 3 : Root=CNode ()
Step 4 :Root.TCondition=findbest-partition (T, A);
Step 5 : Let X = {v}//v is the possible outcome of Root.TCondition
Step 6 : For each v ε X do
Step 7 : Tv = {t} // Root.TCondition (t) =v and t ε
Step 8 : Leaf= BuildTree (Tv, A)
Step 9 : Add leaf as descendent of root node
Step 10 : End for loop
Step 11 : End IF
Step 12 : Return root
Step 13 : ELSE
Step 14 : Child=CNode ();
Step 15 : Child. Label=Classify (T); Return child.
End_Cond:

This function terminates the build tree process by checking whether all the records have the child label or the same attribute values

CNode ():
This function creates the new node
Findbest-partition (T, A):
This function selects the best attribute for splitting the records
Classify ();
This function assigns the Class label to the child node [9].

2.3. Naïve Bayes

Naïve Bayes classification algorithm is a probability based classifier. Naive Bayes predicts the future activities based on some historical data. Naïve Bayes gives high accuracy and speed when it applied on large data set.

Algorithm mainly based on three concepts. They are ,
Prior => all the information from day to day and past experiences
Likely hood =>possibility of information
Posterior =>predicting some particular information based on the information

2.3.1. Algorithm

1. Assume D is a training set of data along with its associated class labels and each tuple represented by T. Each tuple consists of n dimensional attributes T= (T_1, T2….$T_{n)}$.

© 2018, IJCSE All Rights Reserved

2. Suppose there are M classes C_1, C2……C_M. Given a tuple T, the classifier will predict that T belongs to the class having the highest posterior probability conditioned on T.
   $P(C_i \mid T) > P(C_j \mid T)$

3. As P (T) is constant for all classes only $P(T \mid C_i)\ P(C_i)$ need to be maximized.

4. Dataset D with many attribute would be computationally expensive, so compute $P(T \mid C_i)$ based on equation 1

   $$P(T \mid C_{i)} = P(T_1 \mid C_{i)}.\ P(T_2 \mid C_{i)} \ldots\ldots\ P(T_n \mid C_i) \text{ --------(1)}$$

5. Compute the maximized posterior hypothesis as in equation 2
   $$P(C_i \mid T) = P(T \mid C_{i)} P(C_{i)} / P(T) \text{ ----------(2)}$$

2.4. Support Vector Machine (SVM)

SVM Classifier can be applied on linear and non-linear data. It transforms the original data set into larger dimensions by making use of non-linear mapping and searches for a linear separable hyper plane within the new dimensions. This hyperplane is a decision boundary separates tuples of different classes. Data from two classes (for example, data of patients with and without diabetes) can always be separated by a hyper plane. Support vectors and margins are used order to find the hyper plane. Support vectors are subsets of the actual training tuples. SVM is used for both prediction and classification.

Maximum marginal hyper plane separates two classes correctly. Hyper planes with larger margins will be more accurate as compared to those with smaller margins. A hyper plane will always be equally distant from both sides of the margin. Separating hyper plane can be written as:

$$Wv * X + B = 0 \text{ ----------(3)}$$

In equation 3 where,

Wv: weight of the vector

B: Biase

X: Training tuple

III. RESULTS AND DISCUSSION

3.1. Dataset Description

The data set collected from UCI machine learning repository. It consists of 768 instances and 9 Attributes. All patients here are females at least 21 years old of Pima Indian heritage.

	A	B	C	D	E	F	G	H	I
1	pregnant	plasma	bp	skin	insulin	mass	pedegree	age	class
2	3	126	88	41	235	39.3	0.704	27	tested_negative
3	3	158	76	36	245	31.6	0.851	28	tested_positive
4	3	180	64	25	70	34	0.271	26	tested_negative
5	4	129	86	20	270	35.1	0.231	23	tested_negative
6	3	171	72	33	135	33.3	0.199	24	tested_positive
7	1	120	70	30	135	42.9	0.452	30	tested_negative
8	3	170	64	37	225	34.5	0.356	30	tested_positive
9	4	154	62	31	284	32.8	0.237	23	tested_negative
10	1	136	74	50	204	37.4	0.399	24	tested_negative
11	1	153	82	42	485	40.6	0.687	23	tested_negative
12	3	148	66	25	0	32.5	0.256	22	tested_negative
13	6	134	70	23	130	35.4	0.542	29	tested_positive
14	5	139	80	35	160	31.6	0.361	25	tested_positive
15	5	158	84	41	210	39.4	0.395	29	tested_positive
16	4	148	60	27	318	30.9	0.15	29	tested_positive
17	1	138	82	0	0	40.1	0.236	28	tested_negative
18	3	162	52	38	0	37.2	0.652	24	tested_positive
19	1	142	86	0	0	44	0.645	22	tested_positive
20	4	122	68	0	0	35	0.394	29	tested_negative
21	1	171	72	0	0	43.6	0.479	26	tested_positive
22	2	146	76	35	194	38.2	0.329	29	tested_negative
23	3	141	0	0	0	30	0.761	27	tested_positive
24	1	128	64	42	0	40	1.101	24	tested_negative

Figure 1. Screenshot of Pima Indian Dataset

The Figure 1 depicts an overview of the sample Pima Indian Dataset. The attributes are Number of times pregnant, Plasma glucose concentration a two hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), and Body Mass Index, Age (years) etc.

The figure creates a mining structure which excludes some of the fields of Dataset, in favor of a model that is filtered on particular attributes such as Plasma greater than 120 , BMI greater than 40 who belongs to the target group of 20-30 again dataset divides into two based of pregnant and Non-pregnant ladies. The performance of classifiers has been analyzed and prediction has been done on the basis of possibility of getting diabetes in pregnant and non-pregnant ladies. The dataset is used to train and test the diabetic data set by dividing training data and test data using 90-10 ratio. The experiment results with J48, Naïve Bayes, and SVM show the improvement in accuracy.

© 2018, IJCSE All Rights Reserved

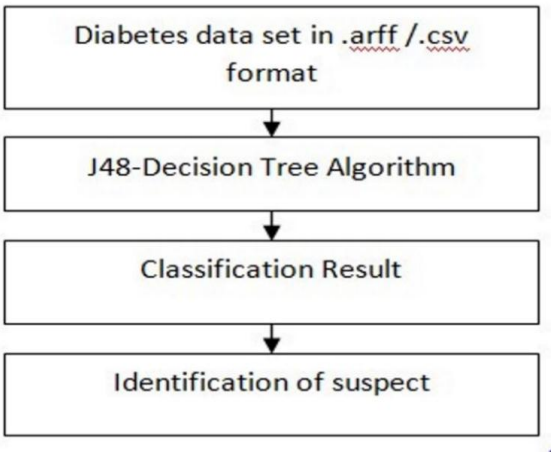

Figure 2. Block Diagram of J48

The Figure 2 depicts the block diagram of j48 classifier. After collecting the dataset the implementation is done with J48 Decision Tree. Then decision rules are extracted for predicting the diabetes in pregnant and non-pregnant ladies separately.J48 is a kind of decision tree which generates some relevant rules by proposed dataset. The J48 decision tree for diabetes diagnosis in pregnant ladies using Pima Indian diabetes dataset is depicted in Figure 3.

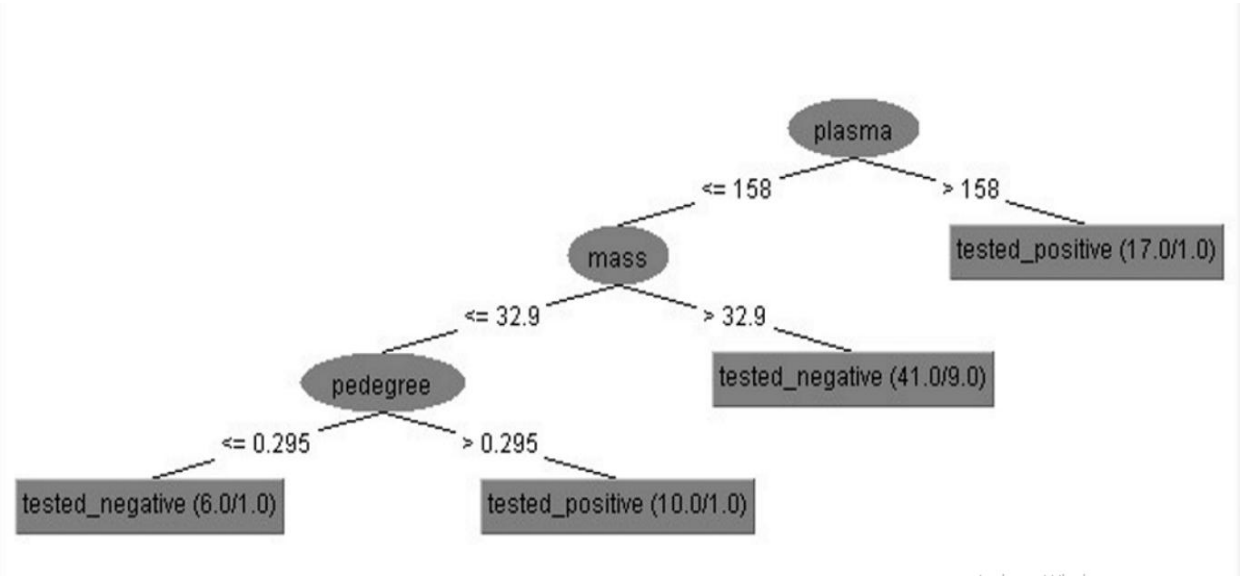

Figure 3.The J48 decision tree for diabetes in pregnant ladies

Rule 1. If plasma> 158 then Class=> Tested Positive
Rule 2. If plasma<= 158 & BMI >32.9 then Class=> Tested Negative
Rule 3. If plasma<= 158 & BMI <= 32.9 & Pedegree > 0.295 then Class=>Tested Positive
Rule 4. If plasma<= 158 & BMI <= 32.9 & Pedegree <= 0.295 then Class=>Tested Negative

Certain rules have been extracted while executing J48 decision tree algorithm for pregnant ladies. If a person has plasma concentration above 158 then the person will be a diabetic patient. If plasma concentration is less than 158 and BMI greater than 32.9 then that particular person will not be a diabetic patient. If plasma is less than or equal to 158 and BMI is less than or equal to 32.9 and pedegree greater than 0.295 then the patient will be tested positive, but if pedegree is less than or equal to 0.295 then that person will be non-diabetic for the same.

© 2018, IJCSE All Rights Reserved

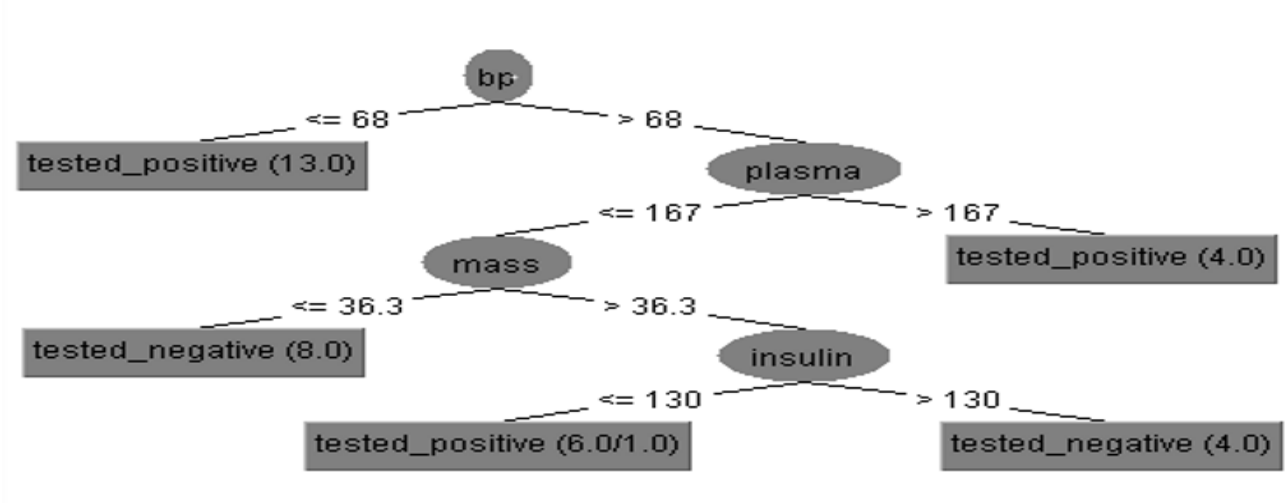

Figure 4. The J48 decision tree for diabetes in non- pregnant ladies

The Figure 4 depicts the visualization of rule based tree which is extracted by J48 with proposed dataset related to non-pregnant ladies. The extracted IF-THEN rules follow,

Rule 1. If bp<= 68 then Class=>Tested Positive
Rule 2. If bp> 68 & plasma> 167 then Class=>Tested Positive
Rule 3. If bp> 68 & plasma<= 167 & mass <= 36.3 then Class=>Tested Negative
Rule 4. If bp> 68 & plasma<= 167 & mass > 36.3 & Insulin>130 then Class=>Tested Negative
Rule 5. If bp> 68 & plasma<= 167 & mass > 36.3 & Insulin<=130 then Class=>Tested Positive

From Rule 1 to Rule5, it is achieved that if the blood pressure is less than or equal to 68 then the particular patient will be diabetic. If the blood pressure is greater than or equal to 68 and plasma is less than or equal to 167, Body mass index is greater than 36.3 then the particular patient will not be diabetic.

Table. 1: Performance result of classifiers

	Classifier	Correctly classified instances	Incorrectly classified instances	Time (sec)	Kappa statistic	Mean absolute error	Root mean squared error
Pregnant Ladies	Naïve Bayes	85.71 %	14.28 %	0.01	0.72	0.2221	0.2778
	J48	85.71 %	14.28 %	0.01	0.72	0.2175	0.3464
	SVM	42.85%	57.14 %	0.01	0	0.5714	0.7559
Non-Pregnant Ladies	Naïve Bayes	75 %	25 %	0.01	0.3846	0.2608	0.4555
	J48	68.75 %	31.25 %	0.02	0.3103	0.3542	0.5408
	SVM	66.66%	33.33 %	0.00	0	0.3333	0.5774

The Table 1 shows the performance of Naïve Bayes, J48 and SVM based on correctly classified instances, incorrectly classified instances, computing time, Kappa Statistic, Mean absolute error and Root mean squared error. Naïve Bayes and J48 have more accuracy compared to that of the SVM. The table above shows that the SVM classifier is not suitable for this particular dataset. Percentage split is used as the test option.90 % of data used for training and the remaining 10 % for testing. Graphical representation of above table is given below in Figure5.

© 2018, IJCSE All Rights Reserved

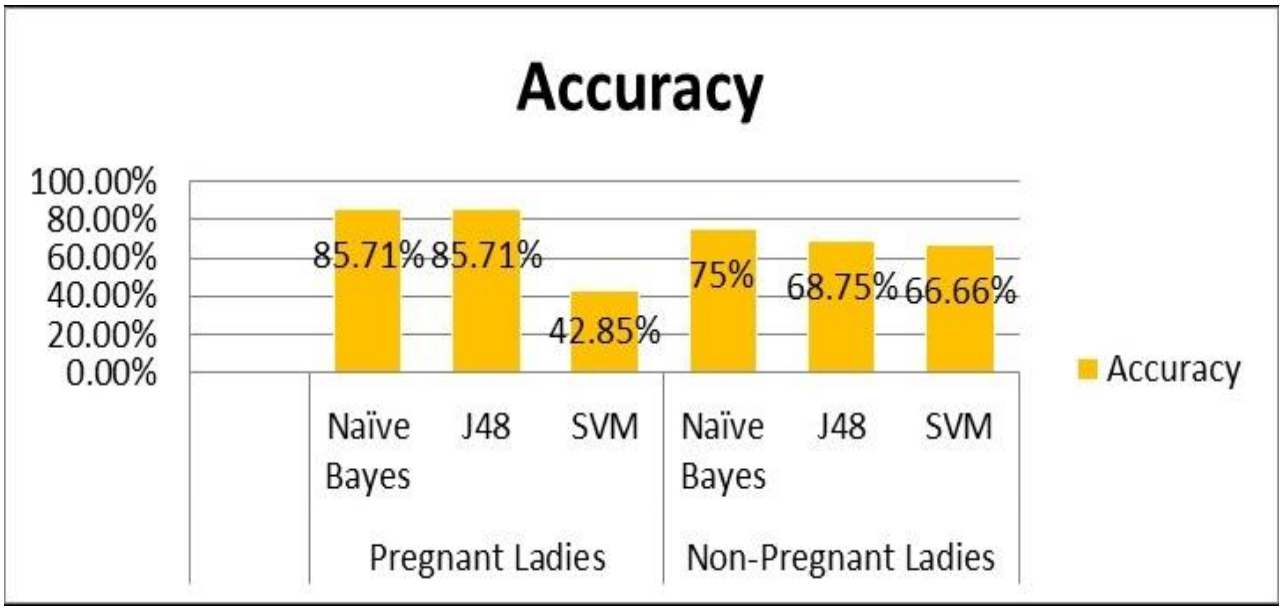

Figure 5. Performance evaluation on Naïve Bayes, J48 and SVM

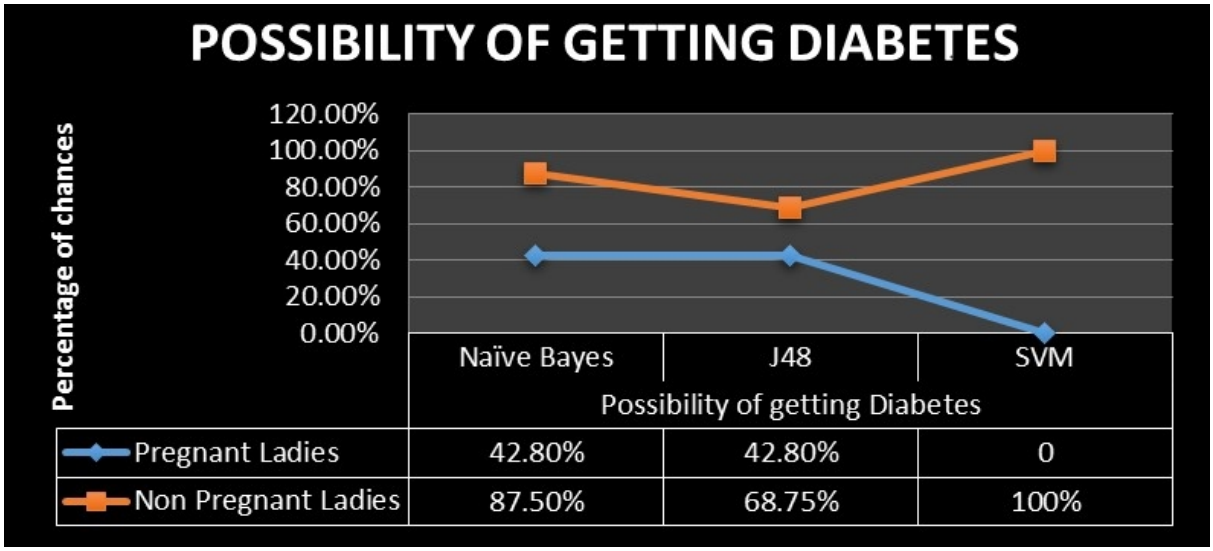

Figure 6. Diabetes prediction in pregnant and non-pregnant ladies

The Figure 6 depicts the possibility of getting diabetes in pregnant and non-pregnant ladies with the help of three classifiers. Naïve bayes classifier predicts that, 42.8% chances is there for getting diabetes in pregnant ladies but in the case of non-pregnant ladies the possibility is more, that is 87.5 %.while running J48 classifier possibility of getting diabetes in pregnant ladies is 42.8% but in non-pregnant ladies it is 68.75%.In the case of SVM chances of getting diabetes in pregnant ladies is 0% but in non-pregnant ladies it is 100%.

© 2018, IJCSE All Rights Reserved

IV. CONCLUSION

In this work, we discussed the possibility of getting diabetics in pregnant and non-pregnant adult obese ladies based on some relevant attributes such as Age, BMI and Plasma concentration. Through this research work we can conclude that Naïve Bayes and J48 as an efficient predictable classifier than SVM for diabetes data set and also predicted the chances of getting diabetes in Non-Pregnant Ladies are greater than that for pregnant Ladies.

ACKNOWLEDGMENT

First and foremost we would like to express our deep gratitude to Prof. Joy Paulose, Head of the Department Computer Science, Christ University, for constantly monitoring and providing us with constructive feedback.

REFERENCES

[1] D. A Kumar and R. Govindasamy," Performance and Evaluation of Classification Data Mining Techniques in Diabetes", International Journal of Computer Science and Information Technologies, Vol. 6 (2), 2015.

[2] Chandra. J, Nachamai.M, and A.S Pillai, "Predicting Cervical Carcinoma Stages Identification Using SVM Classifier", International Journal of Computer Trends and Technology, Volume (22) Number(3), 2015.

[3] C.Shah, A.G. Jivani," Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction", IEEE – 31661

[4] D.Tomar and S.Agarwal ,"A survey on Data Mining approaches for Healthcare", International Journal of Bio-Science and Bio-Technology,Vol(5), No(5), pp. 241-266,2013.

[5] G.Kaur and A.Chhabra,"Improves J48 classification algorithm for prediction of diabetes", International journal of computer Applications,Volume(98) ,No(22), July 2014.

[6] G.K.M.Nookala, B.K.Pottumuthu, N.Orsu, and S.B. Mudunuri," Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification", International Journal of Advanced Research in Artificial Intelligence,Vol(2), No(5), 2013.

[7] M.A. Khaleel, S. K. Pradham, G.N. Dash," A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases", International Journal of Advanced Research in Computer Science and Software Engineering, Volume(3), Issue(8), August 2013.

[8] M.kumari, R.Vohra ,Anshularora," Prediction of Diabetes Using Bayesian Network", International Journal of Computer Science and Information Technologies, Volume(5), Number(4) , pp(5174-5178), 2014.

[9] N.Nandkumarsakhare, s.joshy, "Classification of Criminal Data Using J48 Decision", International journal forum of researchers students and Academician,2012

[10]V.Karthikeyan, I.ParvinBegum, K.Tajudin, I.S.Begam, "comparative of data mining classification algorithm(CDMCA)in diabetes disease prediction",International journal of computer applications(0975-8887) Volume(60),No.(12),December 2012.

Authors Profile

Ms.Manumol Thomas pursed Bachelor of Science from University of Kannur, Kerala in 2014 and Master of Computer Application from Christ University in year 2017. Currently working as Assistant Professor in Department of Computer Science, at St.pius X college Rajapuram,Kerala since 2017.

Mrs. Chandra J pursed Bachelor of Science and Master of Computer Application from University of Bharathidasan University, Tamilnadu in year 1996.Completed MPhil from Vinayaka Missions University in 2009 And completed Phd from Hindustan University in 2016. She is currently working as Assistant Professor in Department of Computational Sciences at Christ University, Bangalore,Karnataka.

© 2018, IJCSE All Rights Reserved