ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Research Paper on Diabetic Data Analysis

Jyoti Kataria
M. Tech. Student, Department of CSE
U.I.E.T., M.D. University, Rohtak
Haryana, India

Babita Kumari
M. Tech. Student
U.I.E.T., M.D. University, Rohtak
Haryana, India

Abstract: Diabetic Data Analysis is a field of research which comes under analytics. Analytics is a subject of statistics to the extent that we read raw data by using computational techniques and then we make sense out of this raw data this is called analysis. An essential function in data mining and analytics is the Data Classification. A machine learning tool known as neural network is capable to perform various tasks in diabetic data analysis. Today, healthcare industries having large amount of data and to access that data analysis process is required, so there arise many complexities. Medicare industries face different kind of challenges, so it is very important to develop data analytics. In this paper an integrated approach is used to predict diabetes from neural network. Neural network can be taken as ubiquitous indicator. From various resources raw data has been collected and compare it to a tool that can be a trained machine for the prediction of diabetes patients. Main aim of integrating approach in neural network is to increase the accurate results in the prediction of diabetic patients. Big data is an approach to resolve the problem in an enhanced manner. A modeling structure is used in this paper.

Keywords: Neural network, big data, data classification, diabetic data analysis, data mining, modeling.

I. INTRODUCTION

Data mining is concerned as a knowledge discovery database (KDD) with a huge amount of data sets. Cluster analysis and association rules are the main aspects of data mining. Different type of techniques, tools, methods and algorithms are used in data mining that play a vital role to take exotic facts under the light. In recent years, the problem of “Predictive Diabetic Data” has been eying attention. Using suitable mechanisms and strategies, the big quantity of facts generated within the scientific area may be processed into records to guide strategic selection making. The opportunity of a thirty to seventy years vintage Indian loss of existence from the four critical non-communicable sicknesses - diabetes, most cancers, stroke and breathing illnesses - is 26 percent at gift, in line with the World Health Organization. On the report of Global Status Report, noninfectious diseases may additionally want to declare nearly trillion resides universally globally via the year 2030. Nearly eighty five million people died of noninfectious ailments inside the WHO's South-East Asia Region in 2012. In India, the non-transmissible diseases are envisioned to have deemed for sixty percentages of all deaths in 2014, at the equal time as 26 percentages a number of the sometime of thirty to seventy years old had a danger of capitulating to the 4 diseases. Diabetic facts evaluation using machine analyzing pursuits to find out and extract applicable patterns from datasets of record from Patients affected by brilliant diabetes .In order to accumulate this we are going to categorize data and. A huge type of system studying algorithms was hired. In trendy, 85% of these used were characterized by using way of supervised getting to know strategies and 15% by means of way of unsupervised ones, and extra especially, association tips. Support vector machines (SVM) rise up because the maximum a success and extensively used set of hints. Concerning the kind of facts, scientific datasets have been specifically used. They become

aware of applications inside the determined on articles mission the usefulness of extracting precious know-how maximum crucial to new hypotheses focused on deeper knowledge and in addition research in DM [12].

1.) MACHINE LEARNING MODEL:

Learning process in machine learning model is divided into two steps as:

□ Training
□ Testing

In training process, samples in training data are taken as input in which features are learned by learning algorithm or learner and build the learning model. In the testing process, learning model uses the execution engine to make the prediction for the test or production data. Tagged data is the output of learning model which gives the final prediction or classified data.

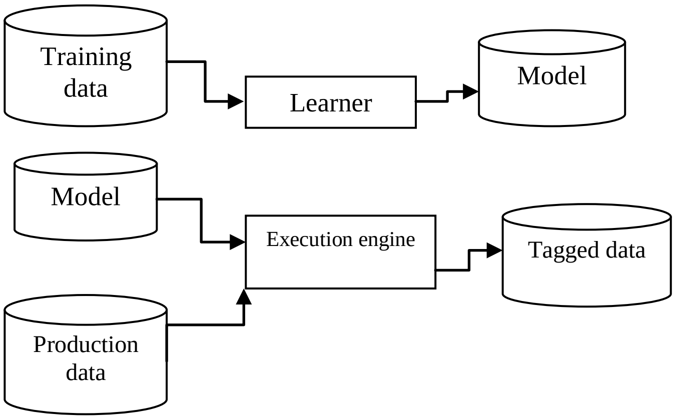


Figure 1: Operational model of machine learning

II. MACHINE LEARNING TECHNIQUES:

The techniques of Machine learning are grouped in 3 main parts, due to the nature of the learning demonstration or assessment convenient for a learning terminology as follows:

A. Supervised learning:

Supervised studying has informed using categorized examples, together with a center in which the popular output is idea. Supervised studying offers dataset inclusive of each abilities and labels. For instance, a piece of gadget wants to have education records elements categorized both as F (failed) rather as R (runs). The undertaking of supervised studying is to accumulate an approximation that is expert to assume the description at any point in the stated group of competencies. The studying set of guidelines receives a tough and fast of skills as inputs on the hassle of the corresponding accurate outputs, and the set of guidelines learns via usage of collating its definite outcome with exact returns to note out mistakes. Then it adapted the version consequently. This version isn't always preferred so long as the inputs are available, however if some of the enter values are missing, it isn't always possible to infer a few element approximately the outputs. Supervised studying is commonly implemented in applications wherein historic facts predict probable destiny activities [3]. For example, it is able to anticipate on the same time as credit rating score rating card transactions are probable to be deceitful or which coverage customer is probably to record a plea. Another utility is predicting the species of it is given a difficult and fast of computation of its blossom. Other greater complicated examples includes reputation device as particular a rainbow photo of a thing along a telescope, determine whether or no longer or now not that item is a movie celebrity, a quasi-stellar source, or a constellation, or stated a listing of films someone has observed and their private status of the movie, propose a group of films they will love. Supervised studying responsibilities are divided into instructions as elegance and relapse. In type, the label is distinct, at the same time as in throwback, the description is non-stop. By illustrating, in astrometry, the venture of figuring out in case or not an item is a celebrity, a constellation, or a quasi-stellar source is a class hassles wherein the description is against 3 extremely good commands. On the alternative hand, in regression problem, the label (age) is a non-prevent amount, as an instance, locating the age of an item based mostly on observations [14]. Supervised reading model is given in determine 2 which suggests that set of guidelines makes the distinction maximum of the raw decided information X this is training records which can be text, document or image and some label given to the version in some unspecified time inside the future of education. In the way of training, supervised learning algorithm builds the predictive model. Besides guidance, the matched configuration will try to estimate the best possible description for latest groups of specimen Z in test data. Determined by the substance of the function x, this method can be classified as follows:

- ➢ If y has values in an attached set of grouping outcomes (represented by integers) the effort to predict y is called classification.

If y has floating point values (e.g. to epitomize a price, a temperature, a size...), the predict y is called regression.

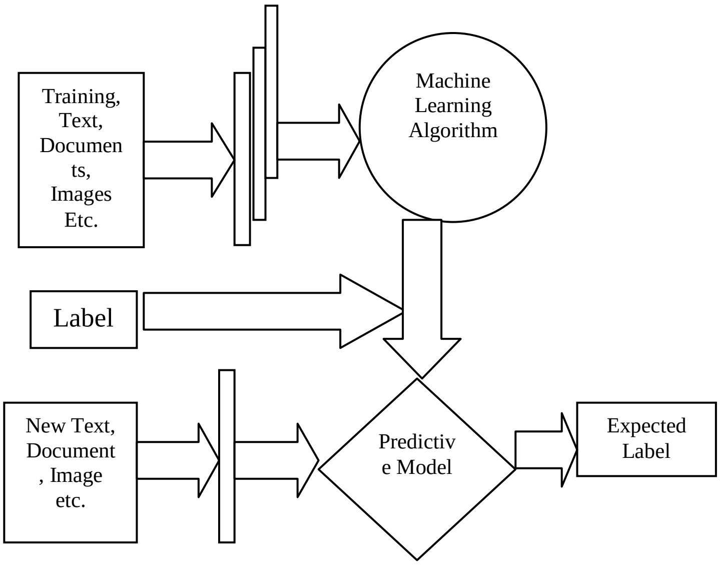

Training, Text, Documents, Images Etc.
Machine Learning Algorithm
Label
New Text, Document, Image etc.
Predictive Model
Expected Label

Figure 2: Supervised learning model

B. Unsupervised learning:

This method operates on the facts that have no historical documentation and the purpose is to review the information and to get similarities between the objects. It is the technique of detect labels from the data itself. That learning works well on transactional data such as discover the components of consumers by same aspects through they considered, that can be viewed identically in trading offensives. It can search the main aspects that divide consumer components from each other [5].

Other unsupervised learning complications are:

- Given detailed observations of distant galaxies, determine which features or combinations of features are most important in distinguishing between galaxies.

- Given a mixture of two sound sources for example, a person talking over some music, separate the two which is called the blind source separation problem.

- Given a video, isolate a moving object and categorize in relation to other moving objects which have been seen.

Typical unsupervised task is clustering where a deck of attributes is distributed into association, unlike in classification; the groups are not known before. Popular unsupervised approach contains self-regulation charts, nearest-neighbor planning, k-means clustering and single volume dissolution. The algorithms are also applied for segmenting the narrative keynotes, suggest modules and discover exceptions in the data.

The unsupervised model has given in figure 3 which shows that this approach uniquely applies on a distinct group of considerations Z by m examples with n samples and k properties that doesn’t handle many type of descriptions. In the training process, unsupervised learning algorithm builds the predictive model which will try to able its boundary so as to perfectly outline the consistencies construct in the information.

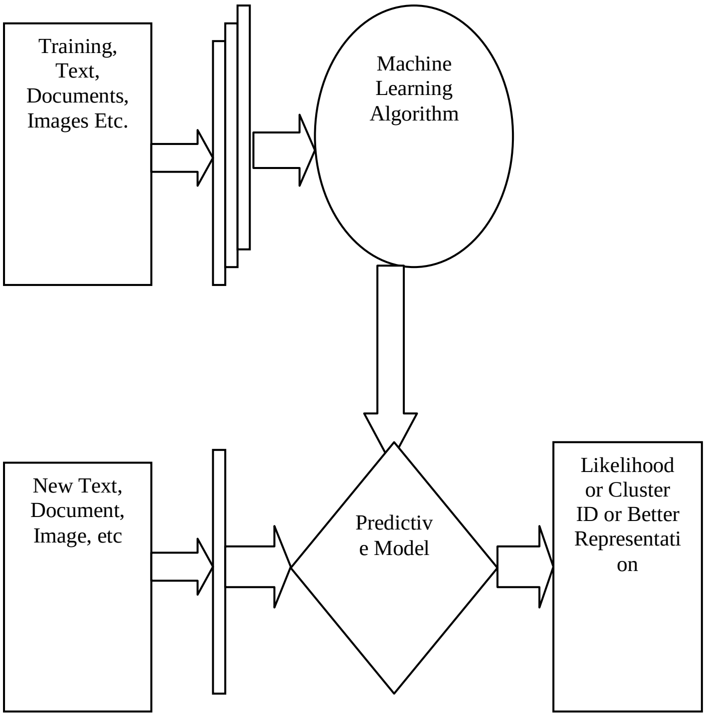

Training, Text, Documents, Images Etc.
Machine Learning Algorithm
New Text, Document, Image, etc
Predictive Model
Likelihood or Cluster ID or Better Representation

Figure 3: Unsupervised learning model

C. Semi-supervised Learning:

In Many sensible gaining knowledge of region which includes textual content processing, video indexing, bioinformatics, there may be large supply of unlabeled statistics but restricted labeled data which may be high priced to generate .So this learning studying is cast-off for the identical requisition guides as supervised analyzing however it uses each marked and unmarked statistics for schooling. There is a preferred prediction problem however the version needs to have a take a look at the structures to set up the statistics further to make predictions. Semi-supervised learning is beneficial whilst the charge related with labeling is simply too excessive to permit for a totally categorized education technique. This kind of reading could use with techniques together with categorization, backsliding and forecasting. Prior specimen of this encompasses to locate the face of a person on an internet squint. Example algorithms are extensions to precise bendy strategies that make assumptions approximately a way to version the unlabelled data [15].

D. Reinforcement Learning:

It is frequently pre-owned for robotics, gaming and seamanship. It is the reading method which keep in touch with a dynamic environment in which it need to carry out a pleasant purpose without a trainer explicitly telling it whether or not or no longer or not or not or not it has come close to its goal. By reinforcement learning, the set of regulations find via trial and mistakes which move to provide the greatest advantage. So inside the chess gambling, reinforcement studying learns to play an activity thru gambling in opposition to an opponent which performs trial and mistakes movements to win.

III. LITERATURE REVIEW

Dr. Saravan Kumar’s paper on predictive approach for diabetic statistics helped me to get a clean picture of Diabetics. This helped me to perceive and define the hassle better [1].

I learnt masses approximately the way to version the whole problem and bypass the whole 9 yards .This manner we had been able to make this venture take location.

Abdullah A. Alijumah and institution of King Suad university did an intensive studies in this domain .This is what endorsed us to take it one step earlier and reengineer it and make contributions to it [2].

We won greater belief from analyzing Andre W. Kushniruk’s paper on Predictive facts Analytics and forecasting in Health Care. In healthcare and outstanding industries, prediction is maximum beneficial while that statistics can be transferred into movement. The willingness to interfere is the vital detail to harnessing the power of historic and real-time statistics [4].

We found about diverse disturbing conditions and Workable strategies for ensuring well timed and suitable manage require extensive linkage and help for reinforcing the supply of educated manpower, investigational facilities and tablets[6].

Krzysztof JĘDRZEJEWSKI, Maurice ZAMORSKI, wrote approximately the use of KNN in information mining for plotting talents. Data analytics is not anything without statistics visualization.

IBM’s CRISP DM manner model helped us understand the flow of code in records mining. Our code is primarily based at the code go with the flow of this model [9].

SVM is only the machine analyzing expertise SVM is one the supervised mastering version with gaining knowledge of algorithms that survey records and grant styles. Guidance is completed for type and backsliding evaluation [11].

We found out using SVM and KNN the usage of blogs from studies gate and Math works .There examples are extra than enough to get a start and begin constructing algorithms which can work with data sets[16].

IV. METHODOLOGY

We are using the data set of PIMA to train our system to recognize Diabetic data report. This data set is collected based on the report of diabetic patients. It contains a lot of information relevant to the study of analysis like BMI(Body mass index), Age, DIA, TSF [11].A pattern is read based on the collected data, this pattern is plotted using curve fitting of Linear plots and a conclusion is achieved. It is only the capability model that analyzes information and allows outlining. Instructions are finished after categorization and reverting inspection. Basic facts take the resources and then foresee the outcome that stand on the prime details. At most 2 viable groups are assuming to be feasible. By a group of practicing, every one spotted a part for 1 of 2 groups, and the terminology forms a structure that allocates recent instances in single class or in different. An SVM model is a description of the illustrations that position in infinity and planned so that the illustrations of the different classes are disjoined by a direct space that is as open as feasible. Recent illustrations are then plotted inside exact area and foresee to stand to a class pointed on which part of the section they decline in computation of execution of continuous categorization.

V. PROPOSED WORK

To gather and analyze report of diabetic patients and conclude if diabetics can be detected in early stages so that proper steps can be taken to counter it and neutralize it. To cover the research work it is necessary to determine the work related to flow which is properly explained and maintained in the research flow in section.

A. Modelling:

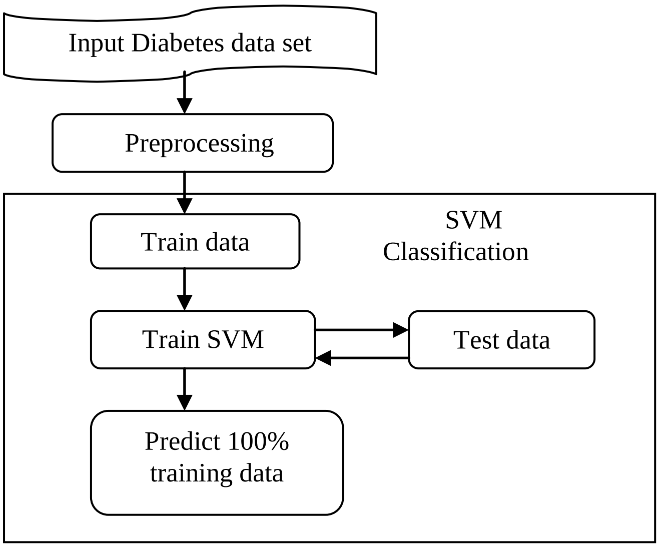


Figure 1 shows the process of Diabetic data Analysis on product reviews:

B. Input Diabetes Data Set:

Pattern popularity or Data Mimicry: For diabetic remedy its miles critical to trial the design like, glucose interest of plasma (PC), insulin, blood strain, diabetes pedigree, Body Mass Index, age, quantity of period pregnant. The sample discovery of predictive statistics assessment need to encompass the following [4]:

- Association rule mining- Relationship among diabetic kind as well as pages seemed (e.g. Lab effects).
- Clustering – It is interchangeable types of utilization, and so on.
- Classification - Classification of fitness is a chance to rate through the use of the amount of affected man or woman health scenario.
- Practice states appeal of pre-described deductive hints at some point of facts. This is all referred to as records sets.

C. Data processing:

Masters of the respective domain recognize the metadata and what it manner. This is the segment wherein a information mining expert makes a selection the supply of his records and starts off developed amassing it .This is a critical phase because the data gathered proper right here will determine the final outcomes of an vital desire[10].

D. Content parsing:

This is the section wherein masters of respective vicinity assemble the Model of information for technique modeling. They acquire smooth and cleanse the records as an example. A facts set from excel sheets normally starts from the 2ndrow.This is not actual for a records from CSV file because of the truth its far neither in block form just like the excel sheet information and neither does it start from the 2nd row. So right steps need to be taken to put together the information manually or through computational strategies.

E. Analysis and scoring:

This is the segment in which we determine on a way to be used to investigate records. It might be KNN wherein we come to a give up thru manner of plotting a graph using a threshold and seeing which way the overall facts is leaning in the direction of[11].We additionally do linear sample plotting to match the sample. This is how we conclude if something is becoming the curve, which makes it relevant otherwise it makes it beside the factor.

- **Evaluation**:

Masters of Data Mining judge the model [13]. If the model does not perform as expected, then modeling phase is revisited and model is rebuilt by altering its VARIABLES until optimum values are attained. When they are finally satisfied with the model, they can remove business clarification and estimate the following questions:

Does the model accomplish the business impersonal? Have all business problems been taken into consideration?

In the unceasing of the analysis stage, the data mining experts determine how to utilize the data mining outcome.

- **Deployment**:

Masters of data mining employ the reaction of mining by transmitting the upshot into tables or into dissimilar applications, for example, spreadsheets for diabetic data analysis. We are using the concept of SVM, Neural Networks and K-NN (Nearest Neighbors):

a) Supervised learning method: We are using stored sentences in an excel sheet. The sentences are stored in such a way that each important phrase in the sentence is extracted and stored separately. We have given these individual phrases some weight we have another matrix called the weight matrix which we are using as training sets [10].

b) We are then comparing the sample test sets with our weight matrix to calculate the net diabetic data ratio

NR = (Positive Diabetic data's – Negative Diabetic data's)/Total Diabetic data's

We are then using Neural Networks [8] to train our system against these data and phrases to assign prediction values to our system .This way when we read data from excel sheets, reports etc we can do predictive analysis of data sets and use our trained system and conclude if the result of the data set is that of a diabetic patient etc are positive, negative or neutral. We finally use the concept of KNN [7] to calculate the overall response of a set or data. We plot the NSR values of a 768 patients and plot it on a graph .We decides a threshold and see which way the most number of tweets is leaning towards.

Table 1: Detail of Patients

Sr. No.	*NPG*	*PGL*	*DIA*	*TSF*	*INS*	*BMI*	*DPF*	*AGE*	*Diabetes*
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1
16	7	100	0	0	0	30	0.484	32	1
17	0	118	84	47	230	45.8	0.551	31	1
18	7	107	74	0	0	29.6	0.254	31	1
19	1	103	30	38	83	43.3	0.183	33	0
20	1	115	70	30	96	34.6	0.529	32	1

VI. RESULTS

Various restrictions were pointed according to the choice of this case against huge amount of data. Specially, here the age of all female patients are at least 21 years old according to PIMA Indian heritage. An adaptive learning routine(ADAP) that initiates/completes digital analogs of sensation-namely appliance.

The information about all the attributes is given below:
1. Count the time a female get pregnant.
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3.Blood pressure in diastolic form i.e. counted in (mm Hg)
4.Triceps skin fold thickness (mm)
5.Insulin serum counted in two-hours(mu U/ml)
6.B.M.I. (weight in kg/(height in m)^2)
7. Function of Pedigree in diabetes Age (years)

VII. ABBREVATIONS:

D.M.-Diabetes Mellitus
N.R.-Net Results
N.P.G. - number of time a female get pregnant
P.G.L.-Plasma glucose concentration a 2 hours in an oral glucose tolerance test
D.I.A.-Diastolic Blood Pressure
T.S.F.-Triceps Skin Fold thickness
I.N.S. - Two hours Insulin
B.M.I.-Body Mass Index
D.P.F.-Diabetes Pedigree function

Table 1 Comparison of machine learning algorithms

Learning Method	Decision Trees	Support-vector machine(with stack variable, no kernel)	Gaussian Naïve Baves	K-nearest Neighbors	Logistic Regression
Generative or discriminative	Discriminative	Discriminative	Generative	Discriminative	Discriminative
Function lost	Either log M(Z kX) or zero-one loss	Hinge loss: k1 z(u^T v)k+	Log M(O; P)	Zero-one loss	Log M (P,Ok)
Decision boundary	Axis-aligned partition of feature space	Linear(depends on Kernel)	For equal variance: lower boundary, for unequal variance: quadratic boundary	Arbitrarily complicated	Linear
Parameter estimation algorithm	Many algorithms: CART, C4.5, ID3	Solve quadratic program to find boundary that maximum margin	Estimate ^,^^{2} and M(P) using maximum likelihood	Must store all training data to classify new points. Choose A using cross	No closed form estimate. Optimize objective functio

				validation	n using gradient descent
Model complexity reduction	Limit tree depth/prune tree	Decrease the value of C	Place prior on parameters and use MAP Estimator	Increase in the value of K	L$_2$ regularization

VIII. PROSPOSED METHODOLOGY TESTING

Mostly known as *t* testing, the proposed methodology testing evaluates if a definite value is really accurate for our input group/inhabitants. In detail study and census, we examine the outcome of proposed methodology analysis stats is outstanding if all the outcomes have not happened by the arbitrary possibilities. Proposed methodology criterion is applied aggregately through their field/fact-finding due to the circumstances or by fiscal policy.

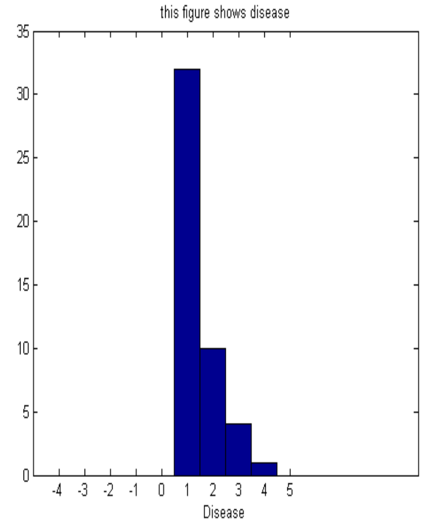
By going through, proposed methodology measure require an obtainable view to find out the ordinary mistakes. By illustrating, the placebo result attained when members speciously forecast an exact outcome after that interpret (or literally achieve) the effect. One more tiresome mistake is the Hawthorne effect which also means the noticeable result that locates when members bios conclusion is achieved as you understand that you are being studied.

IX. CONCLSION

Analysis was done on Pima India Diabetic data base. Different level of disease showing severity of diabetes was found on as many as 768 patients and results were plotted using KNN (K- Nearest Neighbors).

There reports were studied for various parameters like NPG, PGL, DIA, TS, INS, BMI, DPF and age.

Result was plotted for degree of disease against percentage of the patient which has them while a very high percentage of patients had level 1 diabetes a very low percentage had level-2 diabetes. And a very low percentage had type for diabetes which is incurable.



Thus we conclude that machine learning can be used to categories and find diabetes. As all the patients with level one or type diabetes can be advised to take better care of themselves before it becoming unmanageable. If the model is further improved we can find the cause of diabetes which would obviously require input from more variables like eating patterns, stress levels etc.

I. REFERENCES

[1] Dr. N.M. Saravana Kumar, “Predictive Methodology for Diabetic Data Analysis in big data”, Procedia Computer Science, volume 50, pages 203-238, 2015

[2] Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, “Application of data mining: Diabetes health care in young and old patients”, Journal of King Saud University – Computer and Information Sciences, volume 25, issue 2, pages 127–136, July 2013

[3] Olga Tsave, Nicos Magleveras “Machine Learning and Data Mining Methods in diabetes research” Computational and Structural Biotechnology Journal, volume 15, pages 104-116, 2017

[4] Andre W. Kushniruk, “Survey on Predictive Analysis of Diabetes in Young and Old Patients”, International Journal of Advance Research in Computer Science and Software Engineering, volume 5, issue 10, October, 2015

[5] Aishwarya R., Gayathri P., N. Jaisankar “A Method for Classification Using Machine Learning Technique”, International Journal of Engineering and Technology, volume 5, no. 3, June-July, 2013

[6] K. Rajesh, V. Sangeetha, “Application of Data Mining Methods and Techniques for Diabetes Diagnosis”, International Journal of Engineering and Innovative Technology volume 2, issue 3, September 2012

[7] Sadhana, Savitha Shetty, “Analysis of Diabetic Data Set Using Hive and R”, International Journal of Emerging Technology and Advanced Engineering, volume 4, issue 7, July 2014

[8] Kavita Venkataraman, “Challenges in Diabetes management in INDIA”, International Journal Diabetes Dev CtiresV.29, Jul-Aug/2009

[9] Krzysztof, Jędrzejewski, Maurycy Zamorski, “Performance of K-Nearest Neighbors Algorithm in Opinion Classification”, Foundations of Computing and Decision Sciences, volume 38, issue 2, 2013

[10] Jenn Riley, “Understanding Meta Data”, National Information Standard Organization (NISO) Primer, suite 302, MD 21211

[11] Mohamed Baddar, “A Framework for Text Classification using IBM SPSS Modeler”, IBM Business Analytics Proven Practices, 11 February, 2015

[12] Madhura A. Chinchmalatpure, Dr. Mahender P. Dhore, “Review of Big Data Challenges in Healthcare Applications” IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p-ISSN: 2278-8727, PP 06-09, 2016

[13] Loren Shure, “Read Microsoft Excel Spreadsheet File”, File exchange, Math works, 12 July, 2014

[14] John Dillard, “Most important Methods for Statistic analysis”, Big-Sky-Associates, issue no. 356764, 05 April, 2013

[15] Primoz Potocnik, “Neural Networks: MATLAB Examples”, University of Ljubljana, LASIN, June 2015

[16] Thiyagarajan C, Dr. K. Anandha Kumar, Dr. A. Bharathi, “A Survey on Diabetes Mellitus Prediction Using Machine Learning Technique”, volume 11, issue 3, pages 1810-1814, 2016