ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.5

Diabetic Data Analysis using Neural Network

Jyoti Kataria[1], Dr. Sunita Dhingra[2], Babita kumari[3]
M.Tech Student[1, 3], Assistant Professor[2]
Department of Computer Science &Engineering
UIET, M. D. University, Rohtak, India

Abstract:
Diabetic facts evaluation is a place of studies which comes underneath Analytics. Analytics is a subject of records to the diploma that we study uncooked records via deploying computational techniques then we shape experience out of this raw facts this is known as assessment. Advancing enterprise of healthcare circulates towards processing large fitness information, and to achieve those for analysis and set into motion substantially will increase the complexities. Because of the developing unformed nature of Big Data from fitness organization, it's far crucial to form and emphases its size into nominal price with viable answer. Healthcare industry faces many worrying conditions that formulate us to recognize the importance to increase the records analytics. Diabetic Mellitus (DM) is one of the Non Communicable Diseases (NCD), is an outstanding health hazard in growing international locations along with India. The intense nature of DM is associated with eternal complications and masses of health issues. The paper proposes to look at uncooked data from exclusive belongings and evaluate it towards an educated machine to be expecting patterns in reports of patients which reasons diabetes. Depending on the analysis, the machine affords a nicely deliberate manner to treatment and care the sufferers with powerful outcomes peer to availability and affordability.

Keywords: Diabetic data analysis, data set, modeling, diabetic mellitus, machine learning, healthcare industry.

I. INTRODUCTION

In recent years, the problem of "Predictive Diabetic Data" has been eying attention. Using appropriate mechanisms and techniques, the vast amount of data generated in the medical field can be processed into information to support strategic decision making [2]. The possibility of a 30-70 year antique Indian loss of life from the 4 primary non-communicable ailments - diabetes, most cancers, stroke and respiration illnesses - is 26 percent at gift, consistent with the World Health Organization. According to the Global Status Report, Non-Communicable Diseases (NCDs) would possibly claim almost fifty million lives globally via the 12 months 2030. Nearly eight.Five million human beings died of NCDs sicknesses in the WHO's South-East Asia Region in 2012. In India, NCDs are expected to have accounted for 60 percent of all deaths in 2014, whilst 26 percent most of the a while of 30-70 years had a possibility of succumbing to the 4 illnesses. Diabetic data evaluation using tool studying dreams to perceive and extract applicable styles from datasets of file from Patients suffering from specific diabetes .In order to attain this we are going to categorize records. A wide range of machine learning algorithms were employed. In general, 85% of those used were characterized by supervised learning approaches and 15% by unsupervised ones, and more specifically, association rules. Support vector machines (SVM) arise as the most successful and widely used algorithm. Concerning the type of data, clinical datasets were mainly used. The title applications in the selected articles project the usefulness of extracting valuable knowledge leading to new hypotheses targeting deeper understanding and further investigation in Diabetes Mellitus (DM) [3].

II. MODELLING

1.) Input Diabetic Data Set

Pattern discovery - For diabetic treatment it is critical to test the patterns like, plasma glucose focus, serum insulin, diastolic blood strain, diabetes pedigree, Body Mass Index(BMI), age, kind of times pregnant. The sample discovery of predictive evaluation need to encompass the following 5 subjects:

- Association rule mining- Association among diabetic kind and pages considered (e.g. Laboratory consequences)
- Clustering - clustering of similar patterns of utilization, and so forth.
- Classification - Classification of health danger fee by using the extent of affected character health scenario.
- Usage of statistics Application of pre-described deductive guidelines throughout facts. This is all known as information units [5].

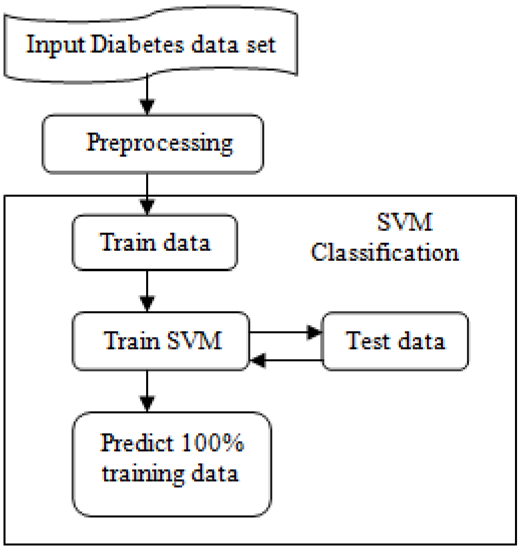


Figure.1. Shows process of Diabetic data Analysis on product reviews

2.) Data Preprocessing

Masters of the respective domain understand the metadata and what it means. This is the phase where a data mining expert decides the source of his data and starts collecting it .This is a vital phase because the data collected here will decide the outcome of an important decision [10].

3.) Content Parsing

This is the phase where domain experts build the Model of data for modeling process. They collect clean and cleanse the data for example. A data from excel sheets usually starts from the 2nd row. This is not true for a data from CSV file because it is neither in block f like the excel sheet data and neither does it start from the 2nd row. Proper steps have to be taken to prepare the data manually or through computational techniques.

4.) Analysis and Scoring

This is the phase where we decide on a technique to be used to analyze data. It could be KNN where we come to a conclusion by plotting a graph using a threshold and seeing which way the overall data is leaning towards [11]. We also do linear pattern plotting to match the pattern. This is how we conclude if something is fitting the curve, which makes it relevant otherwise it makes it irrelevant.

4.1) Evaluation

Data mining specialists evaluate the version [13]. If the version does no longer satisfy their expectations, they move again to the modeling phase and rebuild the model via changing its parameters until most appropriate values are performed. When they may be in the long run glad with the model, they can extract business employer explanations and compare the following questions:

- Does the model achieve the business objective?
- Have all business issues been considered?
- At the end of the evaluation phase, the data mining experts decide how to use the data mining results.

4.2) Deployment

Data mining experts use the mining results by exporting the results into database tables or into other 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 – Negative Diabetic data)/Total Diabetic data. 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 decide a threshold and see which way the most number of tweets is leaning towards. In this hassle, you'll review the crucial elements of the algorithms we have learned approximately in magnificence for the reason that midterm. For each algorithm indexed inside the tables on the next pages, ll out the entries beneath every column in keeping with the subsequent pointers. Do no longer ll out the grayed-out cells. Turn on your completed table along with your trouble set.

III. LITERATURE REVIEW

Dr. **Saravan Kuma**r's paper on predictive methodology for diabetic data helped me to get a clear picture of Diabetics. This helped me to identify and define the problem better [1]. I learnt a lot about how to model the whole thing and go the whole 9 yards. This way we were able to make this project happen. **Abdullah A. Alijumah** and team of King Sud University did an extensive research in this domain. This is what motivated us to take it one step ahead and reengineer it and contribute to it. [2]. various different methods and algorithms for machine learning were studied and opinion of many researchers was taken in the journal of Science Direct to get the right approach [3]. We gained more insight from reading **Andre W. Kushniruk**;s paper on Predictive data Analytics and forecasting in Health Care. In Medicare and other areas, prediction is most useful when that knowledge is transferrable into action. The willingness to intervene is the key to supersede the power of historical and real-time data [4]. Aishwarya R's paper on Classification method which uses machine learning technique was one of many papers that was studied to do data classification which is one of many step required for data analysis [5]. We learned about various challenges and solutions for ensuring timely and appropriate response require extensive linkage and support to enhance the availability of trained workforce, investigated facility and drug [8]. Krzysztof Jędrzejewski, MauriceZamorski, wrote about the use of KNN in data mining for plotting purposes. Data analytics is nothing without data visualization [9]. IBM's CRISP DM process model helped us understand the flow of code in data mining. Our code is based on the code flow of this model [11]. Algorithms that analysis data and recognize patterns. Training is done for categorizing and linear regression analysis. We learned the use of SVM and KNN using blogs from research gate and Math works .There examples are more than enough to get a start and start building algorithms which can work with data sets [13]. **John Dillard** wrote about using different techniques which we can use to do data analysis [14]. We went through all the pros and cons of various techniques like mean, regression etc. We compared them to see which technique is better and how is it better. But most of all we learned how to perform analysis of data. We decided to go for Neural Networks because it takes inputs and mimics the data to produce an output this means we don't have to go through a fixed set of formulae because there are not any when it comes to studies of patterns. Once trained it makes it easy to do read through diabetes [15]. SVM is one the data mining technique, SVM is one the supervised learning model with learning Sanjaya De Silva's article on use of machine learning algorithms talks about the various challenges we have to face during analysis of diabetic data and how different algorithms give different level of accuracy [16]. We also learned how to increase accuracy in our research.

IV. COMPARISION OF MACHINE LEARNING ALGORITHMS

In this problem, you will review the important aspects of the algorithms we have learned about in class. For every algorithm listed in the two tables on the next pages, Table II out the entries under each column according to the following guidelines. Turn in your completed table with your problem set. [$\frac{1}{2}$ point per entry].

Guidelines:

1. Explanatory or general description { Choose either \generative" or \discriminative"; you may write \G" and \D" respectively to save some writing.

2. Loss Function {Write either the name or the form of the loss function optimized by the algorithm (e.g., \exponential loss"). For the clustering algorithms, you may alternatively write a short description of the loss function.

3. Decision Boundary {Describe the shape of the decision surface, e.g., \linear". If necessary, enumerate conditions under which the decision boundary has different forms.

4. Estimate or prediction of algorithm {Name or concisely describe an algorithm for estimating the parameters or predicting the value of a new instance. Your answer should t in the provided box.

5. Reducing complexity of model {Define the method for putting model complexity and preventing over thing.

Table.1. Comparison of Classification Algorithms

Learning Method	Generative or Discriminative?	Loss Function	Parameter Estimation Algorithm	Predictio n Algorithm	Model Complexity Reduction
Bayes Nets	Generative	log P(X;Y)	MLE	V ariable Elimination	MAP
Hidden Markov Models	Generative	log P(X;Y)	MLE	Viterbi or Forward-Backward, depending on prediction task	MAP
Neural Networks	Discriminative	Sum-squared error	Back-Propagation	Forward Propagation	Reduce number of hidden layers, regularization, early stopping

Table.2. Clustering Algorithm

Learning Method	Loss Function	Number of clusters: Predetermined or Data-Dependent	Cluster shape: isotropic Or anisotropic?	Parameter Estimation Algorithm
K-means	Within-class squared distance from mean	Predetermined	Isotropic	K-means
Gaussian Mixture Models (identity covariance)	log P (X), (equivalent to within-class squared distance from mean)	Predetermined	Isotropic	Expectation Maximization (EM)
Single-Link Hierarchical Clustering	Maximum distance between a point and its nearest neighbor within a cluster	Data-dependent	Anisotropic	Greedy agglomerative clustering
Spectral Clustering	Balanced cut	Predetermined	Anisotropic	Run Laplacian s Eigenmaps followed by K-means or thresholding eigenvector signs

6. Clusters in order and manner {Select \predetermined" or \data-dependent"; you may write \P" and \D" to save time.

7. Shape of cluster {Select \isotropic" (i.e., spherical) or \anisotropic"; you may write \I" and \A" to save time.

V. CHALLENGES IN DIABETIC DATA ANALYSIS

India will want to also plan for the care of the super quantity of people with diabetes, with a purpose to save you and reduce morbidity because of complications. A fitness machine strengthening technique with necessities of care in any respect ranges, nationally famous manipulate protocols and regulatory framework can help in tackling this undertaking [8]. Diabetes management stays a task for advanced and growing worldwide places alike.

The implementation of evidence-based absolutely suggestions and restructuring of clinical care employer has yielded profits in a few international locations.

There have been severe tries in developing nations as properly to generate possible and effective care systems. These projects and initiatives keep promise however an awful lot relies upon on the re-orientation of the overall health machine for powerful and sustainable care [12].

In India, as in distinctive international places, the health device has historically been designed to cater to acute contamination and maternal and toddler health concerns.

The burgeoning load of diabetes is a actual danger in India, underscored by using the restrictions of the fitness device in terms of manpower and capability. The need for lengthy-time period care, for non-communicable sicknesses, is a notably new health state of affairs, and personnel and infrastructure areas but not geared to face this task.

Workable techniques for ensuring nicely timed and appropriate management require extremely good linkage and manual for reinforcing the provision of knowledgeable manpower, investigational centers and drugs. Primary prevention through selling of healthy life and threat discount is diagnosed as the most charge-powerful intervention in useful resource-terrible settings.

VI. COMPARISON OF DIFFERENT TECHNIQUES

1.) Mean

The mathematics mean, greater normally known as "the average," is the sum of a listing of numbers divided by way of the variety of objects at the listing. The advice is useful in identifying the general style of an information set or offering a fast picture of your data. Another gain of the Suggest is that it's very clean and brief to calculate [14].

Disadvantages:

Taken on my own, endorse is a risky device. In a few facts devices, the advocate is likewise carefully Related to the mode and the median (unique measurements close to the average).However, in a statistics set with a excessive amount of outliers or a skewed distribution, the suggest really doesn't offer the accuracy you want for a nuanced choice.

2.) Standard Deviation

The popular deviation, regularly represented with the Greek letter sigma, is the degree of an expansion of records across suggest. Excessive famous deviations means that data is unfold extra widely from suggest, in which a low general deviation signals that more records align with the advocate. In a portfolio of statistics analysis methods, the equal vintage deviation is useful for quickly determining dispersion of facts.

Disadvantages:

Just much like imply, the same antique deviation is devious if taken by myself. For instance, if the data have a completely brilliant pattern including a non-everyday curve or a massive amount of outliers, then the same old deviation won't come up with all of the facts you want.

3.) Regression

Regression fashions the relationships between established and explanatory variables, which might be normally charted on a scatter plot. The regression line additionally designates whether or not those relationships are strong or weak. Regression is commonly taught in high college or college information guides with applications for science or business in determining trends through the years.

Disadvantages:

Regression is not very nuanced. Sometimes, the outliers on a scatter plot (and the motives for them) remember considerably. For example, an outlying statistics point might also represent the input from your maximum vital supplier or your maximum promoting product. The nature of a regression line, however, tempts you to ignore those outliers. As an illustration, observe a photo of Anscombe's quartet, in which the statistics sets have the precise same regression line however encompass extensively unique information points.

4.) Sample Size Determination

When measuring a large data set or population, like a workforce, you don't always need to collect information from every member of that population – a sample does the job just as well. The trick is to determine the right size for a sample to be accurate. Using proportion and standard deviation methods, you are able to accurately determine the right sample size you need to make your data collection statistically significant.

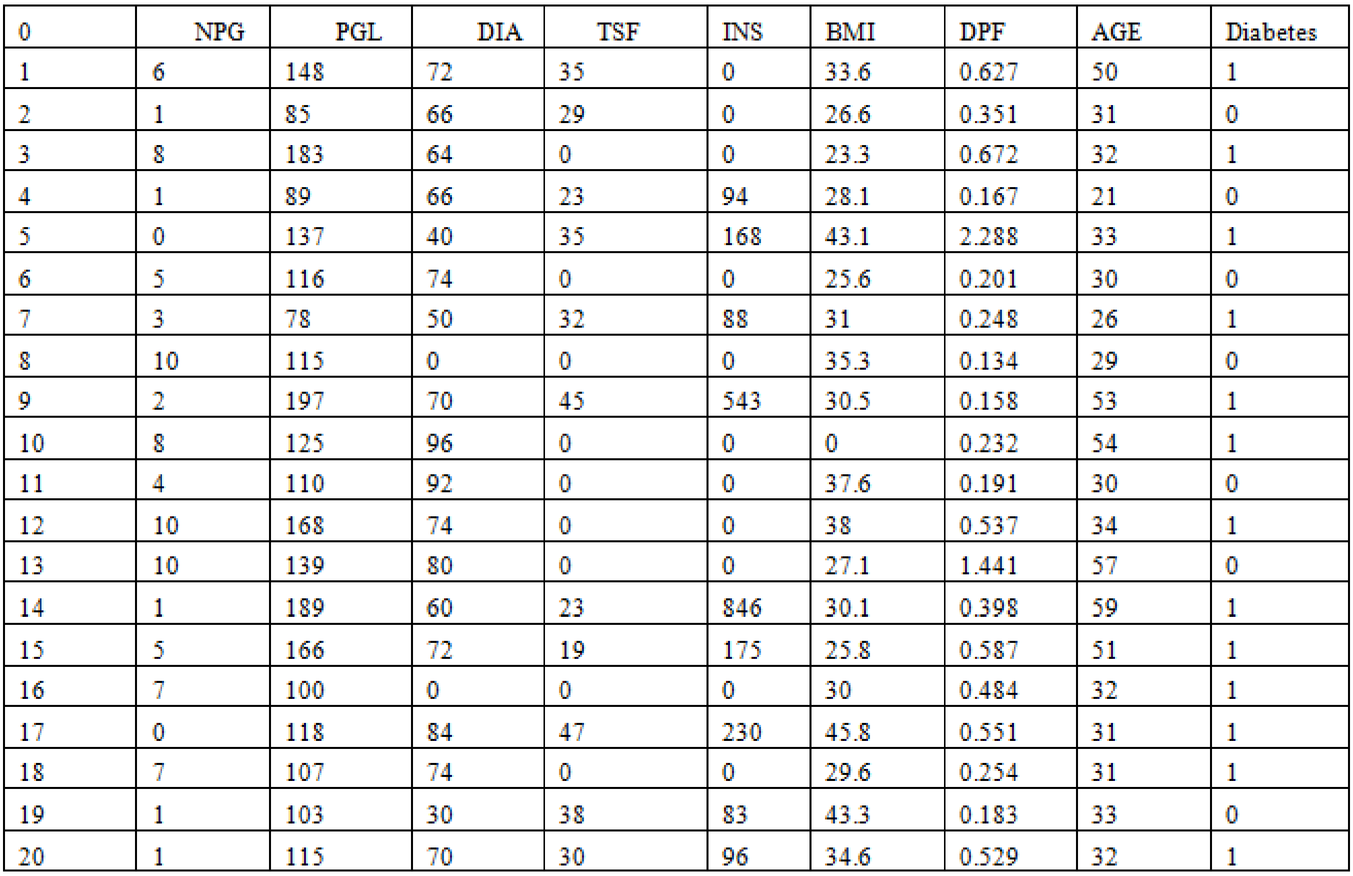
Disadvantages:

When studying a new, untested variable in a population, your proportion equations might need to rely on certain assumptions. However, these assumptions might be completely inaccurate. This error is then passed along to your sample size determination and then onto the rest of your statistical data analysis.

5.) Hypothesis Testing

Also commonly referred to as t sorting out, speculation checking out assesses if a nice premise is surely true on your information set or population. In information analysis and facts, you keep in mind the stop result of a speculation check statistically massive if the outcomes couldn't have took place with the resource of random hazard. Hypothesis exams are applied in the entirety from era and research to enterprise and monetary.

Table.3. Details of patients



0	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

Disadvantages:
To be thorough, speculation checks want to take a look at out for common errors. For example, the placebo effect takes area while members falsely anticipate a superb result and then perceive (or in reality benefit) that stop result. Another commonplace mistakes is the Hawthorne effect (or observer effect), which takes place even as members skew results because they recognize they may be being studied. Overall, those methods of information analysis upload some of notion in your choice - making portfolio, in particular if you've in no manner analyzed a technique or records set with records earlier than. However, avoiding the common Disadvantages related to every method is just as critical. Once you grasp the ones essential strategies for statistical facts evaluation, then you definitely prepared to growth to extra powerful facts evaluation gear.

VII. RESULTS

Data Set Information:

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. ADAP is an adaptive learning routine that generates and executes digital analogs of perception - like devices.

Attribute Information:
1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4 .Triceps skin folds thickness (mm)
5. 2 - Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)

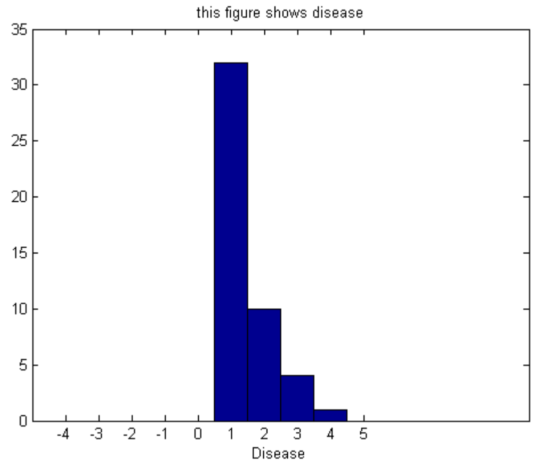


Figure.4. Result of Disease Was Performed On the Report Data Set of 768 Patients.

Abbreviations used in data analysis of patients given in the following table 3:

- NPG- No. of Times Pregnant
- PGL- Plasma Glucose
- DIA- Diastolic Blood Pressure
- TSF- Triceps Skin Fold Thickness
- INS- 2 hours Insulin
- BMI- Body Mass Index
- DPF- Diabetes Pedigree Function

VIII. APPLICATIONS

1.) Chronic care of diabetes comes with large amount of data concerning the self and clinical management of the disease.

2.) Better medical treatments can be devised based upon the pre information of diabetics.

3.) Predictive analysis can help us predict if someone have diabetes in future.

4.) Diabetes can be identified at earlier stage and there is not an ounce of that presentation is better than cure.

5.) Detecting the stage of diabetes can help doctors to understand the patient better, in this way a better care plan can be provided which would tried and tested upon the solid facts of statics.

IX. CONCLUSION

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. Thus we conclude that machine learning can be sued to categories and find diabetes. If the model is further improved we can find the cause of diabetes which would obviously required input from more variables like eating patterns, stress levels etc.

X. REFERENCES

[1]. Dr. Saravana Kumar “Predictive Methodology for Diabetic Data analysis”, *science direct.*

[2]. Abdullah A. Aljumah, Mohammed GulamAhamad, Mohammad Khubeb Siddiqui, (2012), “Application of data mining: Diabetes health care in young and old patients”, *Journal of King Saud University – Computer and Information Sciences*, vol. 25, pp. 127–136.

[3]. “Machine Learning and Data Mining Methods in diabetes research”, *science direct.*

[4]. Andre W. Kushniruk, (2008), “Predictive Analytics and Forecasting in Health Care: Integrating Analytics with Electronic Health Records”, *SAS Institute Inc*.

[5]. Aishwarya R. “Method of classification using machine learning technique “, *International Journal of Engineering and Technology (IJET).*

[6]. K. Rajesh, V. Sangeetha, (2012) “Application of Data Mining Methods and Techniques for Diabetes Diagnosis” in *International Journal of Engineering and Innovative Technology (IJEIT)* VOL 2(3).

[7]. Sadhana, Savitha Shetty, (2014) “Analysis of Diabetic Data Set Using Hive and R”, *International Journal of Emerging Technology and Advanced Engineering*, Vol. 4(7).

[8]. Kavita Venkataraman, (*2009*), “Challenges in Diabetes management in INDIA”, *International Journal Diabetes Dev CtiresV.29,* JULY-AUGUST.

[9]. Krzysztof, Jędrzejewski, Maurycy Zamorski, “Performance of K-Nearest Neighbors Algorithm”, *Foundations of Computing and Decision Science*, Vol-38, No.22.

[10]. Jenn Riley, “Understanding Meta Data”, NISO Primer.

[11]. Mohamed Baddar, (11 February, 2015) “A Framework for Text Classification using IBM SPSS Modeler”, *IBM Learning Center,.*

[12]. Madhura A.Chinchmalatpure “Review of Big data Challenges in Healthcare Applications”, *IOSR-JCE.*

[13]. Read Microsoft Excel, xls read, Spreadsheet file, math works.

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

[15]. Primoz Potocnik, “Neural Networks”, University of Ljubljana, LASIN.

[16]. Sanjaya De Silva,(30 Dec., 2016) “Predicting Diabetes using a Machine Learning Approach”, *Big Data Zone*.