Indian Journal of Science and Technology, Vol 9(4), DOI: 10.17485/ijst/2016/v9i4/87028, January 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

# Developing a Modified Logistic Regression Model for Diabetes Mellitus and Identifying the0 Important Factors of Type II DM

**M . Nirmala Devi[1*], Appavu alias Balamurugan[2] and M. Reshma Kris[1]**

[1]IT Department, Thiagarajar College of Engineering, Madurai - 625005, Tamil Nadu, India; nirmaladevi2004@gmail.com, reshmakris.rk@gmail.com
[2]Department of IT, KLN College of Information Technology, Sivaganga - 630612, Tamil Nadu, India; app_s@yahoo.com

## Abstract

**Background/Objectives:** Different methods can be applied to create predictive models for the clinical data with binary outcome variable. This research aims to explore the process of constructing the modified predictive model of Logistic Regression (LR). **Method/Statistical Analysis:** To improve the accuracy of prediction, the Distance based Outlier Detection (DBOD) is used for pre-processing and Bipolar Sigmoid Function calculated using Neuro based Weight Activation Function is used in Logistic Regression instead of Sigmoid Function. Datasets were collected from clinical laboratory of AR Hospital in Madurai for the three years 2012, 2013 and 2014 are used for analysis. Data pre-processing is done to avoid the existence of insignificant data in the dataset. The detected outliers, using DBOD method are treated using a method closest to the normal range. A comparative study among different distance measures likes Euclidean and Manhattan etc. are done for DBOD method. The pre-processed data finally is fed as input to the Logistic Regression model. Maximum likelihood estimation is used to fit the model. Logistic Model is built from the Sigmoid Function using the Regression Coefficients. The accuracy of the model is evaluated by 10 fold cross validation. **Findings:** Logistic Model is built from the Sigmoid Function using the Regression Coefficients, produces the accuracy of 79%. The Sigmoid Function calculated using Random Weight Function provides the prediction accuracy of 84.2% and the Bipolar Sigmoid Function calculated using Neuro based Weight Activation function provides the prediction accuracy of 90.4%. On comparison, Bipolar Sigmoid Function calculated using Neuro weight activation function outperforms well than the Sigmoid Function calculated using regression coefficients. **Improvements/Applications:** The accuracy of Logistic Regression is improved from 79% to 90.4%. The most important factors: Erythrocyte Sedimentation Rate (ESR) and Estimation of Mean blood Glucose are identified from positive subjects of Diabetes Mellitus. The analysis is done for the 31 Diabetes Disease attributes of three years dataset.

**Keywords:** Bipolar Sigmoid Neuro-Weight Activation Function, Distance based Outlier Detection Method, Logistic Regression, Random Weight Function, Sigmoid Activation Function, Type 2 Diabetes Risk Factors

## 1. Introduction

The burden of Diabetes Mellitus (DM) is expanding internationally, especially in developing nations. The reasons are a complex, yet the expansion is in vast part because of fast increments in overweight, including obesity and physical idleness. Despite the fact that there is great proof that an expansive extent of instances of diabetes and its confusions can be anticipated by a solid eating regimen, general physical movement, keeping up a typical body weight and maintaining a strategic distance from tobacco, this confirmation is not broadly executed.

* Author for correspondence

Composed global and national strategies are expected to lessen presentation to the known danger elements for diabetes and to enhance access to and nature of consideration[1–3].

About 347 million people worldwide have diabetes. There is an emerging global epidemic of diabetes that can be traced back to rapid increases in overweight, including obesity and physical inactivity. Diabetes is predicted to become the 7th leading cause of death in the world by the year 2030. Total deaths from diabetes are projected to rise by more than 50% in the next 10 years. There are two major forms of diabetes. Type 1 diabetes is characterized by a lack of insulin production and Type 2 diabetes results from the body's ineffective use of insulin. A third type of diabetes is gestational diabetes. This type is characterized by hyperglycemia or raised blood sugar with values above normal but below those diagnostic of diabetes, occurring during pregnancy. Women with gestational diabetes are at an increased risk of complications during pregnancy and at delivery. They are also at increased risk of Type 2 diabetes in the future. Type 2 diabetes is much more common than Type 1 diabetes. Type 2 accounts for around 90% of all diabetes worldwide. Reports of Type 2 diabetes in children - previously rare – have increased worldwide. In some countries, it accounts for almost half of newly diagnosed cases in children and adolescents. Cardiovascular disease is responsible for between 50% and 80% of deaths in people with diabetes. Diabetes has become one of the major causes of premature illness and death in most countries, mainly through the increased risk of Cardiovascular Disease (CVD). In 2012 diabetes was the direct cause of 1.5 million deaths. 80% of diabetes deaths occur in low- and middle-income countries. In developed countries most people with diabetes are above the age of retirement, whereas in developing countries those most frequently affected are aged between 35 and 64. Diabetes is a leading cause of blindness, amputation and kidney failure. Lack of awareness about diabetes, combined with insufficient access to health services and essential medicines, can lead to complications such as blindness, amputation and kidney failure. Type 2 diabetes can be prevented. Thirty minutes of moderate-intensity physical activity on most days and a healthy diet can drastically reduce the risk of developing Type 2 diabetes. Type 1 diabetes cannot be prevented[1–3].

Over 30 million have now been diagnosed with diabetes in India. The CPR (Crude Prevalence Rate) in the urban areas of India is thought to be 9 per cent. In rural areas, the prevalence is approximately 3 per cent of the total population. The population of India is now more than 1000 million: this helps to give an idea of the scale of the problem. The estimate of the actual number of diabetics in India is around 40 million. This means that India actually has the highest number of diabetics of any one country in the entire world. IGT (Impaired Glucose Tolerance) is also a mounting problem in India. The prevalence of IGT is thought to be around 8.7 per cent in urban areas and 7.9 per cent in rural areas, although this estimate may be too high. It is thought that around 35 per cent of IGT sufferers go on to develop Type 2 diabetes, so India is genuinely facing a healthcare crisis. In India, the type of diabetes differs considerably from that in the Western world. Type I is considerably more rare and only about 1/3 of Type 2 diabetics are overweight or obese. Diabetes is also beginning to appear much earlier in life in India, meaning that chronic long-term complications are becoming more common. The implications for the Indian healthcare system are enormous[4–6]. Data mining plays a major role in the field of medicine. The prediction model also helps the doctors in the process of diagnosis. Preventing the diabetes disease of diabetes is an ongoing area of interest to the health care community[7].

## 2. Previous Work

Predictive analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behavior patterns. Predictive analytics also focuses on distilling insight from data, but its main purpose is to explicitly direct individual decisions. Medical professionals need a reliable prediction methodology to diagnose diabetes. It is very important for the clinicians and as well as patients to know about the future holds of diabetes for planning better treatment. Predictive models have been developed to make predictions about the particular disease. Predictive modeling technique consists of tasks like classification, regression and time series analysis as it helps in improving the patient outcomes[7,8]. An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Outlier detection has many applications, such as data cleaning, fraud detection and network intrusion. The existence of outliers can indicate individuals or groups that exhibit a behavior that is very different from most of the individuals of the dataset. Detecting outliers, instances in a database with unusual

properties, is an important data mining task. People in the data mining community got interested in outliers after Knorr and Ng proposed a non-parametric approach to outlier detection based on the distance of an instance to its nearest neighbors. There are several methods for detecting multivariate outlier: Robust statistical-based Outlier Detection outlier detection by clustering, Outlier Detection by neural networks, distance-based outlier detection and density-based local Outlier Detection. In this research distance-based Outlier Detection is taken for data cleaning. The distance based Outlier Detection method outperforms than other methods. The problem of detecting outliers has been extensively studied in the statistics community. Typically, the user has to model the data points using the statistical distribution and points are determined to be outliers depending on how they appear in the relation to the postulated model. The problems with these approaches are that in a number of situations, the user might simply do understand the underlying data distribution[9–14].

In Data Mining there are many Prediction Algorithms like Support Vector Machine, Logistic Regression. The Logistic Regression well suits for the Health Care Domain[15–17]. Logistic Regression measures the relationship between a categorical dependent variable and one or more independent variables, which are usually continuous, by using probability scores as the predicted values of the dependent variable. The prediction model is developed using Logistic Regression Model. Logistic Regression estimates the probability of an event occurring. The logit function is the inverse of the Sigmoidal Logistic Function or Logistic Transform used in mathematics, especially in statistics. Logit is defined as the logarithm of the ratio of frequencies of two different categorical and mutually exclusive outcomes such as healthy and sick. The logit of success is then fit to the predictors using Linear Regression analysis. The predicted value of the logit is converted back into predicted odds. The regression coefficients are usually estimated using maximum likelihood estimation. It is the procedure of finding the value of one or more parameters for a given statistic which makes the known likelihood distribution a maximum. The input to the model is the high dimensional dataset and the output of this model is the list of predicted diseases that diabetic patients had the chances of occurrence. The Neuro based Weight Activation Function is sometimes described as squashing function. It is the nonlinear characteristic of the sigmoid function. The sigmoid function will only produce positive numbers between 0 and 1. The Sigmoid Activation function is most useful for training data that is also between 0 and 1. Because the Sigmoid Activation function has a derivative, it can be used with gradient descent based training methods[18–21].

Logistic Regression is a non-linear regression technique for prediction for dichotomous (binary) dependent variables in terms of independent variables (covariates). The dependent variable represents the status of the patient. It is a modern technique mostly used in the field of medical research. Researchers have been using several data mining techniques in the diagnosis of heart disease[22,23]. Using Diabetic Diagnosis, the system exhibited good accuracy and predicts the attributes such as age, sex, blood pressure and blood sugar and the chances of diabetic patient getting various disease. Erisi Mafuratidze et al. had done cross-sectional study was done on 239 patients who have been analyzed as Type 2 diabetics for over 12 months, who routinely go to the Parirenyatwa Diabetic Clinic. Serum test were utilized for urea and creatinine investigation. The pervasiveness of impeded renal capacity found in patients going to Parirenyatwa Diabetic facility was roughly 27.2%. All patients with impeded renal capacity were hypertensive. Guys had a more noteworthy rate of raised urea and creatinine levels contrasted with females. Age and period on treatment were observed to be essentially connected with debilitation of renal capacity. The important factors of diabetes were also analyzed[24].

If the important factors are well understood for a disease, it is possible to educate the public to reduce their risk by avoiding the important factors that they can control. These are publicly available data stores which document factors facing the general public in United States. Nine Years of existing data from 2002 to 2010 is used to build the logistic model. The generated model shows promise in modeling diabetes as function of risk factor and correctly identified the important factors of diabetes[25].

## 3. DBOD for Pre Processing

The datasets with insignificant data cause an issue in the Prediction Model and the result will not be accurate. The Data Preprocessing with Outlier Detection is used to

remove those insignificant data. The input of this method is Excel sheet with 31 attributes. The initial phase is the cleaning of missing data. The missing data is replaced by the standard method called replacement by mean and mode. The next phase is the detection of outliers. In this phase DBOD method is used. The distances for all the data points are calculated. There are two threshold values in this method. One is minimum threshold and other is maximum threshold. The setting process of the threshold values is based on the continuous range values of each attributes. The individual calculated distances are compared with the threshold values. If the calculated distance value is less than the minimum threshold value or else if it is greater than the maximum threshold value, it is categorized as outlier value. It is replaced by nearest normal range method. The algorithm is as follows:

### 3.1 Proposed DBOD Method - Outlier Detection

```
Data Set D= {t1, t2, t3……., tn}
t =>Tuple= {A1, A2, A3……, Am}
a =>Attributes
m =>No of Attributes
n =>No of Data points
s =>set of outliers,s=ϕ
Tmin = >minimum threshold
Tmax =>maximum threshold
E = >Euclidean distances= {e 1, e2, e3…...,en}
for i = 0 to n
 if (e[i]<Tmin || e[i]>Tmax)
            {
             s = s ∪ ti
             }
end
```

**Pseudo-code 1.** Pseudo-code of DBOD algorithm for pre-processing.

### 3.2 Proposed Replacement by Nearest Normal Range (NNR) Method - Outliers Treatment

The normal range of each attribute value is known as $d_1$ to $d_2$ where $d_1$ is the starting value of the normal range and $d_2$ is the ending value of the normal range. For instance, the distance calculated for a data point p is r.

```
Begin
Result 1 = |d1-r|
Result 2 = |d2-r|
            if (Result 1 > Result 2)
                        {
Replace the value by Result2
                        }
            else
                        {
Replace the value by Result1
                        }
end.
```

**Pseudo-code 2.** Pseudo-code of replacement by Nearest Normal Range (NNR) method - Outliers Treatment.

### 3.3 Distance Measures

There are many different distance metrics to measure similarity/dissimilarity between two images of same size i.e. Manhattan Distance, Euclidean Distance, Minkowski Distance (MD) of order p between two points is given below:

$$P = (X_1, X_2, X_3)$$

$$Q = (Y_1, Y_2, Y_3)$$

$$MD(X,Y) = \sum (|X_i - Y_i|^p)^{\frac{1}{p}}$$

**Equation 1:** Minkowski Distance Model.

## 4. Logistic Regression Model for Predictive Analytics

In the existing Logistic Regression, the value of Logistic Function (Sigmoid Function) is calculated using Regression Coefficient β and constant term α. Therefore, the calculation of Sigmoid Function is done using the most frequently used Sigmoid Activation function in neuro based weight activation function.

### 4.1 Neuro based Weight Activation Function

In the Neuro based Weight Activation Function, the value of Sigmoid Function is calculated as follows:

$$y = b_0 + b_1 x$$

**Equation 2.** Target Variable (Y) calculation.

$$b_0 = \bar{y} - b_1 \bar{x}$$

**Equation 3.** Bias Calculation.

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

**Equation 4.** Weight Calculation.

Where,

$y_i$: Class Label (0 or 1)
$x_i$: Attributes in the dataset
$\bar{y}$: Mean of Class Label
$\bar{x}$: Mean of Attributes values

$$z = \Sigma X_k W_{jk}$$

**Equation 5.** Sigmoid Function Calculation.
Where
$X_k$ = >The Value of each Attribute.
$W_{jk}$ = >The Weight associated with each attribute.
Therefore,

$$z = \Sigma X_k W_{jk}$$

$$f(z) = \frac{1}{1+e^{-z}}$$

**Equation 6.** Logistic Regression Logit Function Calculation.

## 4.2 Proposed Logistic Regression Pseudo Code

$X$ = >Attributes in the dataset.
$W$ = >Weight of each Attributes.
$Z$ = >Value returned by Neuro Weight Activation Function.
$f(z)$ = > Predicted value of sigmoid function.
= > Probability that the class label is 1.
= > Probability that the class label is 0.
Begin

Load the Data Set into the system;

for i = 0 to instance.size()
for i = 0 to (x. length-1)

Calculate$^{Z}$
Return$^{Z}$
if (class label == 1)
{
Calculate $f(z)$
$P = z$;
}
else
{
(1-$P$) = 1 - $z$;
}
Calculate$^{P(X)}$
Return$^{P(X)}$
End.

**Pseudo-code 3.** Proposed Logistic Regression Pseudo code.

# 5. Discussion and Evaluation

The data set which was collected for the three years of 2012, 13 and 14 is taken for evaluation of results. The dataset consists of 739 patients' details with 31 attributes. The following Table 1 describes the attributes and the range of values to decide the presence of attributes.

**Table 1.** Attributes and its range of values

| ATTRIBUTES | NORMAL RANGE OF ALUES |
|---|---|
| AGE | 3 -100 |
| GENDER | m/f or 0/1 |
| BSR | 70-140mg/dl |
| BSF | 70-110 mg/dl |
| BSPP | 100-160 mg/dl |
| CHO | 130-200 |
| UREA | 10-50 mg/dl |
| CREATININE | 0.4-1.4 mg/dl |
| T4 | 4.6-12 ug/dl |
| FT4F | 0.03-0.005% |
| FT4 | 0.7-1.9 ng/dl |
| THBR | 0.9-1.1 |
| HDL | more than 45 |
| LDL | less than 100 |
| VLDL | lessthan 30 |
| EOSINOPHIL | 01-06% |
| TRICGLYCERIDE | less than 130 |
| HBA1C | <6 non diabetes 6-8 diabetes with good control |
| ESTIMATION_OF_MEAN_BLOOD_GLUCOSE | 70-140(normal) |
| GLUCOSE_TOLERANCE_TEST(GTT) | mg/dl |
| TOTAL COUNT | 4000-11000mm |
| HAEMOGLOBIN | 13.5-17 |
| POLYMORPHS | 40-70gms% |
| LYMPOCYTES | 20-50% |
| EOSINOPHILS | 01-06% |
| ESR | 10-20mm |
| SODIUM(SERUM) | 128-160mmol/l |
| POTASSIUM(SERUM) | 3.2-5.7 mmol/l |
| CHLORIDE(SERUM) | 95-115 mmol/l |
| BICARBONATE(SERUM) | 20-28 mmol/l |

The attribute descriptions are given below:

- **BSR** : Blood Sugar Random.
- **BSF** : Blood Sugar Fasting.
- **BSPP** : Blood Sugar Post Prandial.
- **Hb** : Hemoglobin.
- **CHO** : Cholesterol.
- **TC** : Total Count.
- **T3**: Tri-iodoThyronine.
- **T4**: Thysonine.
- **TSH** : Thyroid Stimulating Hormone.
- **DC** : Differential Count.
- **ESR** : Erythrocyte Sedimentation Rate.
- **PC** : Platelets Count.
- **GTT** : Glucose Tolerance Test.
- **HDL** : High Density Lipoprotein.
- **LDL**: Low Density Lipoprotein.
- **VLDL**: Very Low Density Lipoprotein.
- **HbA1c**: Glycated Hemoglobin.
- **T4:** Serum Thyroxine.
- **FT4F:** Free Thyroxine Fraction.
- **FT4 :** Free Thyroxine.
- **THBR:** Thyroid Hormone Binding Ratio.

The dataset in the format of Excel sheet is given as an input to the Distance based Outlier Detection Algorithm built using Euclidean Distance measures with various *p* values. The output resulted with the outliers and those Outliers are treated using Closest to Normal Range Values.

By varying the values of p, in Table 2, different distances measure is taken for analysis. The performance of the algorithm can be devised on either the input data or output data. The time and space complexity taken by all the distances are same. Their performance is measured based on the parameter - number of outlier detected.

**Table 2.** Performance analysis of various distance measures

| Distance measure | Number of outlier |
|---|---|
| P = 1 ,Manhattan Distance | 146 |
| P = 2, Euclidean Distance | 165 |
| P = 3 | 120 |
| P = 4 | 100 |
| P = 5 | 100 |

On comparing the five Distance Measures (Reference Table 2) the space and time taken by the distances are same**.** Euclidean Distance outperforms all other distances based on the measure - Number of Outliers Detected.

The result of Outlier Analyses will be the pre-processed dataset in the notepad, which was fed as the input for the predictive model built using Sigmoid Logistic Regression with the Weight function of Neuro based Activation Function.

**Table 3.** Overall prediction accuracy using maximum likelihood estimation

| Observed | Predicted | | |
|---|---|---|---|
| | classlabel | | Percentage |
| | 0 | 1 | Correct |
| Step 1 classlabel 0 | 279 | 72 | 79.5 |
| Overall Percentage | 83 | 305 | 78.6 |
| | | | 79.0 |

**Table 4.** Selection of right measure by performance analysis of various accuracy measures

| Accuracy Measure | Accuracy Achieved |
|---|---|
| Hosmor-Lemeshow Goodness of Fit | 59.1 % |
| Maximum Likelihood Estimation | 79 % |

On comparing the two Accuracy Measures, the Maximum Likelihood Estimation Outperforms than the Hosmor-Lemeshow Goodness of Fit. (Reference Table 4).

So, the Maximum Likelihood Estimation accuracy measure was used in improving the Predictive model by using Sigmoid Function (Reference Table 3), which results 84.2%. In order to improve the predictive model accuracy, the improved Sigmoid Function with Neuro Based Weight Function was used , which results 90.4 %. (Reference Table 5.)

**Table 5.** Performance analysis of Logistic Regression model for different weight functions

| Name of the Function | Accuracy Achieved |
|---|---|
| Sigmoid Function | 79% |
| Sigmoid Function using Random weight fucntion | 84.2% |
| Improved Bipolar Sigmoid fucntion using Neuro weight Function | 90.4% |

On comparing the prediction Accuracy of Logistic Regression (79%) with Neuro Weight Activation based Bipolar Sigmoid Function; it provides an accuracy of 90.4%.

The attributes which are involved for the presence of diabetes are analyzed. From that the high percentage of involvement of each attribute is calculated .They are ranked based on the involvement percentage in the following table. (Reference Table 6.)

**Table 6.** Finding high risk factors of Diabetes Mellitus based of the percentage of involvement in the performance analysis of Logistic Regression model for different weight functions

| ATTRIBUTE | 2012 DATASET | 2013 DATASET | 2014 DATASET |
|---|---|---|---|
| BSR | 142 - 22 % | 151.4 - 60 % | 149 -70% |
| BSF | 137 - 45 % | 135.0 - 80 % | 139 -85% |
| BSPP | 194 - 35 % | 189 - 65 % | 198 -69% |
| LDL | 88.8 - 39 % | 131 - 90 % | 135 -90% |
| VLDL | 40.7 - 54 % | 41 - 90% | 43 -92% |
| TRIGLYCERIDES | 138.2 - 40 % | 151.4 - 90 % | 155 -85% |
| HbA1C | 7.9 - 55 % | 15 - 90 % | 14 -90% |
| ESR | 18.2 – 95% | 22.72 – 97 % | 21.6-98% |
| Estimation of Mean blood Glucose | 166.5 – 55% | 222 – 96 % | 202-97% |

**Table 6.** Finding high risk factors of Diabetes Mellitus based of the percentage of involvement in the performance analysis of Logistic Regression model for different weight functions

Referring to the Table 6, the most important factors of Type 2 diabetes are identified as or ESR and Estimation of Mean blood Glucose.

## 6. Conclusion

The distance based Outlier Detection method is used to identify the insignificant data in the dataset. The Detection Technique is implemented using various distances such as Manhattan distance; Euclidean distance etc. On comparison with various distance measures Euclidean distance outperforms other distances based on the number of outliers detected. The preprocessed patient data is subjected to the Logistic Regression model that is implemented using Sigmoid Function with regression coefficients and provides accuracy of 79%, but improved Sigmoid Function implemented using Neuro based weight activation function, provides accuracy of 90.4%. On comparison Sigmoid Function calculated using Neuro based weight activation function outperforms well than Sigmoid Function using regression coefficients, in addition the developed web service "Dia Care" help the diabetic patients to know the chances of them getting other related diseases, that is implemented using Association Rule. The future work can be extended by implementing the predictive model for other diseases like cancer disease, eye disease etc. The model is provided as a web service in our project. The web service gives the diseases that the diabetic patients can get affected based on their attribute values. And also it can be extended for all other diseases and on other platforms like android etc.

## 7. References


1. WHO | Facts and figures about diabetes. 2015.Available from: http://www.who.int/features/factfiles/diabetes/facts/en/
2. Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes-estimates for the year 2000 and projections for 2030. PubMed Diabetes Care. 2004 May; 27(5):1047–53.
3. Whiting DR, Guariguata L, Weil C, Shawj J. IDF Diabetes atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. PubMed Diabetes Research Clin Pract. 2011 Dec; 94(3):311–21.
4. Joshi SR, Parikh RM. India-diabetes capital of the world: Now heading towards hypertension. J PubMed Assoc Physicians India. 2007 May; 55:323–34.
5. Kumar A, Goel MK, Jain RB, Khanna P, Chaudhary V. India towards diabetes control: Key issues. PubMed Australas Med J. 2013; 6(10):524–31.
6. Diabetes in India. Available from: http://www.diabetes.co.uk/global-diabetes/diabetes-in-india.html
7. El-Sappagh SH, El-masri S, Raid AM, Elmogy M. Data mining and knowledge discovery: Applications, techniques challenges and process models in health care. IJERA. 2013 May-Jun; 3(3):900–6.
8. Milovic B, Milovic M. Prediction and decision making in health care using data mining. International Journal of Public Health Science. 2012 Dec; 1(2):69–78.
9. Bay S, Schwabacher M. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2003. p. 29–38.
10. Knorr E, Ng R, Tucakov V. Distance-based outliers: Algorithms and applications. VLDB Journal: Very Large Data Bases. 2000 Feb; 8(3-4):237–53.

11. Hadi AS, Imon AHMR, Werner M. Detection of outliers. Computational Statistics. 2009 Jul-Aug; 1(1):57–70.
12. Angiulli F, Fassetti F. Detecting distance-based Outliers in streams of data. In Proceedings of IKM '07; 2007 Nov. p. 811–20.
13. Aggarwal S, Singh J. Outlier detection using K-mean and hybrid distance technique on multi-dimensional data set. International Journal of Advanced Research in Computer Engineering and Technology. 2013 Sep; 2(9):2626–31.
14. Chandvanya JR, Aluvalu R. Ranking with distance based Outlier Detection Techniques: A survey. International Journal of Computer Applications. 2014; 89(6):8–11.
15. Logistic Regression - An Overview. Available from: http://www.math.umt.edu/graham/stat452/logistic.pdf
16. Goteti VS, Pannucci M, Zhang J. Logistic Regression Analysis of the occurrence of diabetes in pima Indian Women. 2004. Available from: http://www.rci.rutgers.edu/~cabrera/587/pima.pdf
17. Samanta B, Kuijpers M, Zimmerman RA, Jarvik GP, Wernovsky G, Clancy RR, Licht DJ, Gaynor JW, Nataraj C. Prediction of per ventricular leukomalacia. Part I: Selection of hemodynamic features using logistic regression and decision tree algorithms. Artificial Intelligent Medicine. 2009 Jul; 46(3):201–15.
18. Dong G, Lai K, Yen J. Credit score based on logistic regression with random coefficients. Procedia Computer Science. 2010 May; 1(1):2463–8.
19. Van der Heijden H. Decision support for selecting optimal logistic regression model. Expert Systems with Applications. 2012 Aug; 39(10):8573–83.
20. Maalouf M, Siddiqi M. Weighted Logistic Regression for large scale imbalanced and rare events data. Knowledge-Based Systems. 2014 Mar; 59(1):142–8.
21. Park J, Edington DW. A sequencial Neural Network Model for diabetes Prediction. Artifical Intelligence in Medicine. 2011 Nov; 23(3):277–93.
22. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH, Steinbach M, Hand DJ, Steinberg D. Top 10 algorithms in data mining. Journal Knowledge and Information Systems. 2007; 14(1):1–37.
23. Gorunescu F. Data mining concepts, models and techniques. Intelligent Systems Reference Library. Berlin Heidelberg, Springer-Verlag: 2011. p. 256–60.
24. Mafuratidze E, Chako K, Phillipo H, Zhou DT. Over 27% of Type 2 diabetic patients studied at Parirenyatwa Diabetic Clinic in Zimbabwe have evidence of impaired renal function. International Journal of Scientific and Technology Research. 2014 Mar; 3(3):14–8.
25. Pederson J, Liu F, Alfarraj F, Ngondo H. Examining disease risk factors by mining publicly available information. Procedia Computer Science. 2010; 17:48–53.