PREDICTING THE RISK OF BEING A DIABETIC PATIENT USING STATISTICAL ANALYSIS AND DATA MINING

BPN Perera[1] , BTGS Kumara[2], and HACS Hapuarachchi[3]

[1,2]Department of Computing and Information Systems, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka

[3]Department of Sports Science and Physical Education, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka

[1]*pinipanisansala359@gmail.com*

Abstract-There is a vast and enormous amount of data available in hospitals and medical related institutions. But, the amount of knowledge obtained from such data is very little. Applying IT knowledge for healthcare is an emerging field of huge importance for providing prognosis as well as a deeper understanding of medical data. Diabetics is actually a disease which is affecting many people today Early prediction of diabetes is an extremely challenging task because of the complicated interrelationship between various factors. This research tried to diagnose diabetes which based on 12 risk factors using data from 200 people and applied data mining and statistical techniques to predict the risk of being a diabetic patient. Statistical model has been created using Minitab with the application of the binary logistic regression model. The created model provided the way of predicting the possibility of having diabetics for any person and identified the most suitable risk factors which are most relevant to the disease prediction. Through this identified risk factors, we clustered the data using k-means. An empirical study has proved the effectiveness of our proposed approach.

Keywords- Binary logistic regression, Data mining, K-means clustering

I. INTRODUCTION

Diabetes is a disease that occurs when the blood glucose, often known as blood sugar, is simply too high. Blood glucose is the prime source of strength and is derived with the foodstuff people eat. Further, diabetes is a disease that occurs just as the insulin manufacturing inside the body is insufficient or the body is not able to use the produced insulin in a correct manner, consequently, the aforementioned one leads to high blood glucose. Insulin, a hormone produced by the pancreas, helps sugar originating at food deal with the cells for use for strength and energy. Sometimes the body does not conduct enough or any insulin or does not use insulin properly. The body cells dismantle the food into glucose and this glucose should be shifted to all of the cells of the body. After that prevail person blood and does not reach his/her cells. Another word, the insulin is definitely the hormone that directs the glucose which is produced by breaking off the food into the body cells. Any change within the manufacturing of insulin results in an increase in the blood sugar levels and this can lead to damage to the tissues and failure of the organs (Kelly, 2006).

There are three main types of diabetes:

- Type 1: It occurs when the body doesn't manufacture insulin. The immune system of the body attacks and destroys the cells in the pancreas which make insulin. Though there are actually only around 10% of diabetes patients experience this type of Diabetes. The disease reveal as an autoimmune disease happening at a very young age of less than twenty years thus further called juvenile-onset diabetes. It is often diagnosed in teenagers and youthful adults, although it can present at any age.
- Type 2: It occurs when the body doesn't manufacture or use insulin properly. Type 2 is the commonest type of diabetes. It can happen and develop type 2 diabetes at any age, even all through childhood. This type accounts for nearly 90% of the diabetes cases

and generally known as the adult-onset diabetes or the non-insulin dependent diabetes. In this case, the several organs of the body turn into insulin resistant, and that increases the demand for insulin. At that case, pancreas doesn't manufacture the necessary amount of insulin. To keep this kind of diabetes at bay, the patients have to follow a strict diet, exercise routine and keep track of the blood glucose. Obesity, being overweight, being physically inactive can lead to type 2 diabetes.

- Gestational diabetes pursues some females when they are pregnant. Most of the time, this type of diabetes leave after the child is born. It is actually a type of diabetes that has a tendency to reveal in pregnant women because of the high sugar levels as the pancreas don't produce sufficient amount of insulin. Taking no treatment can result in complications for the duration of childbirth. Controlling the diet and taking insulin can regulate this kind of diabetes. However, if someone had gestational diabetes, then she has got a super chance of coming up type 2 diabetes thereafter in life (Your guide to diabetes : type one and type two, 2013).

Among these three types of diabetic, type 2 diabetic was selected for the research. According to International Diabetes Federation 382 million people are living with diabetes all over the world. By 2035, this would be doubled as 592 million. Diabetes led to 1.5 million deaths in 2012. Higher-than-optimal blood glucose was responsible for an extra 2.2 million deaths due to expanded risks of cardiovascular and different diseases, for a whole of 3.7 million deaths related to blood glucose levels in 2012. Many of these deaths (43%) occur below the age of 70. In 2014, 422 million people in the world had diabetes – a prevalence of 8.5% among the grownup population. The prevalence of diabetes has been unwaveringly increasing for the past 3 decades and is growing most rapidly in low- and middle-income countries. In 2011, 71.4 million people (8.3%) in South East Asia were affected by diabetes and 23.8 million people (2.8%) were affected by Impaired Glucose Tolerance (IGT). Numbers are expected to reach 120.9 million (10.2%) for diabetes and 38.6 million (3.2%) for IGT by 2030, according to the DASL. The diabetes prevalence of the people over two decades in Sri Lanka reveal that the urban population was 16% and among the rural population it was 8%. Under the age of 20 years it was 8.2%. In United States, 30.3 million people (9.4% of the population) had diabetes in 2015 (GLOBAL REPORT ON DIABETES, 2016).

II. RELATED WORKS

(Lowanichchai et al., 2006) proposed an approach using decision tree. The research used data originating at health description, Suranaree Army2 Hospital. It used screening data for diabetes risk which might be used for decision support in making plans the right treatment and relevant for individual patients. The result showed that the Random Tree model has the highest accuracy and the NB Tree model has minimal accuracy. Research paper (Yuvarani & Selvarani, 2016) shows the contrast of the three decision tree classification models for the UCI storehouse diabetes dataset and shows the tree structure formed enabling users to receive truthful decisions in accordance with the input parameters. Further, the genetic J48 model is found to be the most valuable and truthful when compared with any other two decision models in terms of time, efficiency and features.

(Huang et al., 2007) employed three data mining algorithms namely Naive Bayes, IB1 and C4.5 to predict diabetes on data gathered from Ulster Community and Hospitals Trust (UCHT) between 2000 and 2004. They were able to achieve an efficiency of 98%.

Research work (Rajesh & Sangeetha, 2012) has recorded a research study by the use of a variety of classification algorithms like ID3, C4.5, LDA, Naïve Bayes, K-NN for diagnosing diabetes for the given dataset. It has showed which C4.5 is definitely the most competitive algorithm with minor error rate of 0.0938 and further efficiency value of 91%.

(ferreira, oliveira & freitas, 2012) used the several classification algorithms like simplecart, j48, simple logistics, smo, naivebayes as well as bayesnet in order to get diagnosing neonatal jaundice in type1 diabetes. Among all algorithms, it found that Simple Logistics as the finest algorithm. Further, (Parthiban, Rajesh & Srivatsa, 2011) practiced Naïve Bayes approach to determine heart related problems that are occurring in diabetic patients.

However, compare to other research we used two approach to predict and analyse diabetic patients. First, we generated statistical model. Clustering approach is used as second method. In clustering approach, we clustered patients according to the risk level.

III. METHODOLOGY

In this research, Binary logistic regression is used for the creating the statistical model to calculate the possibility of being a diabetic patient. Next, Simple K-means clustering algorithm is used for predicting the level of diabetes.

A. Dataset Used

The data which is used in this research has records of 200 from the government hospital in Kiribathgoda, Sri Lanka. The dataset includes details of both diabetic patients and non-diabetic patients of all age group and Table 1 shows all the attributes used for this research work

Table 1. The attributes used in Data Set for this research

Attribute Name	Values
Age	Real Numbers
Sex	Male or Female
Sugar Consumption	0 or 1
BMI value	Real Numbers
Disease of Pancreas	0 or 1
Diastolic Blood Pressure	Real Numbers
Cholesterol Availability	0 or 1
Background (Family History)	0 or 1
Alcohol Consumption	0 or 1
Smoking	0 or 1
History of Heart Disease/ Stroke	0 or 1
Infection/Illness/skin rashes	0 or 1
No of Children	Real Numbers
Dark, thick skin around neck	0 or 1

B. Data Pre-processing

The Dataset used in this research is medical dataset which may have some inconsistencies. To remove those inconsistencies data pre-processing is done. In Data pre-processing the misclassified data is removed. In this process cleaning and filtering of the data is carried out with respect to the data and data mining algorithm employed so as to avoid the creation of deceptive or inappropriate rules or patterns.

C. Research Process

- Step 1: Applying Binary Logistic Regression to the data set for creating statistical model and identify most effective attributes.
- Step 2: Applying Simple K-Means algorithm to the dataset of most effective attributes for clustering the data into four clusters as normal, impaired fasting glucose, diabetic and high blood sugar.
- Step 3: Visualize and validate the statistical model and clustering results.

IV. RESULTS

A. Statistical Analysis

The pre-processed dataset was applied in to the Minitab tool in three steps as follows;

1) Creating the statistical model

2) Testing Attributes

3) Removing the unnecessary attributes from the model

4) Validating the result

By creating the statistical model, the most effective and less effective attributes can be identified using P-value. Table 2 shows the P- values of each attributes.

A small P-value (typically < 0.05) shows the strong evidence against the null hypothesis. It means they are mostly affected to the logistic model. A large P-value (>0.05) shows the weak evidence against the null hypothesis. It means they are not mostly affected to the logistic model. There are 11 significant attributes. Among them the most significant attributes which have the minimum P- values were identified. Following are the attributes.

- Sugar Consumption
- BMI value
- Disease of pancreas
- Alcohol Consumption
- Background (Family History)

Table 2. P- Values of the attributes

Source	P-Value
Regression	0
Age	0.023
Sex	0.812
Sugar Consumption	0
BMI value	0.002
Disease of Pancreas	0.001
Diastolic Blood Pressure	0.027
Cholesterol Availability	0.04
Background (Family History)	0.006
Alcohol Consumption	0.006
Smoking	0.032
History of Heart Disease/ Stroke	0.031
Infection/Illness/skin rashes	0.023
No of Children	0.344
Dark, thick skin around neck	0.938

As the result of applying the binary logistic regression, it built up a statistical model which consists of two mathematical equations. They generate a way to predict the possibility of being a diabetic patient.

Equation 1. Calculating Y'

Y' = -12.71 + 0.0474 Age + 1.951 Sugar Consumption + 0.1499 BMI value + 2.167 Disease of Pancreas + 0.0542 Diastolic Blood Pressure - 1.081 Cholesterol Availability + 1.395 Background (Family History) + 2.89 Alcohol Consumption + 0.983 Smoking + 1.659 History of Heart Disease/Stroke + 2.24 Infection/ Illness/Skin rashes

$$P(1) = \exp(Y')/(1 + \exp(Y'))$$

Equation 1 consists of most relevant attributes which are most affected to the statistical model. The attributes values should be replace to this equation and then the value of Y' should be calculated well. Then the calculated Y' value should be replace to the equation 2 and should calculate the value of P(1). Prediction of being a diabetic patient depends on this P(1) value.

If the P(1) value is equal or greater than 0.5, The result is close to 1 (one). It means the prediction risk is positive. In other words, the person has the risk of being a diabetic 00person.If the P(1) value is less than 0.5, The result is close to 0 (zero). It means the prediction risk is negative. In other words, the person has not the risk of being a diabetic person. Positive range is identified as the value which is equal or greater than 0.5 and the negative range is identified as the value which is less than 0.5.

Statistical model was validated using two steps. Step 1 was done using randomly selected 25 instances from the data set of 200 instances. Step 2 was done using new 50 instances which also gain from government hospital in Kiribathgoda, Sri Lanka. Table 3 shows the performance of validated the statistical result.

Table 3. Performance of validated the statistical result

Validating Steps	No of correct predicting	No of wrong predicting	Accuracy result (%)
Step 1	23	2	92
Step 2	42	8	88

B. Clustering Analysis

The identified most suitable attributes (Sugar Consumption, BMI value, Disease of pancreas, Alcohol Consumption, Background (Family History)) were applied in to the WEKA tool. K-means clustering algorithm was used to cluster these data into 4 clusters. Table 4 shows the number of cluster instances and percentage of each cluster.

Table 4. Number of cluster instances

Cluster Number	Clustered Instances	Percentage
Cluster 0	55	28%
Cluster 1	28	14%
Cluster 2	59	30%
Cluster 3	58	29%

Table 5 shows how the 200 data categorized into the four clusters. Here, the clusters were identified as follows.

- Cluster 3: fasting glucose is greater than 300 mg/dl.
- Cluster 0: fasting glucose is equal and greater than 126 mg/dl and lower than 300 mg/dl.
- Cluster 1: fasting glucose is equal and greater than 100 mg/dl and lower than 126 mg/dl (impaired level).
- Cluster 2: fasting glucose is greater than 70 mg/dl and lower than 100 mg/dl

Table 5. Final cluster centroids

Clusters	Precision (%)	Recall (%)	Accuracy (%)
Cluster 0	98.18	84.37	94.50
Cluster 1	83.33	71.42	94.00
Cluster 2	72.88	82.69	87.50
Cluster 3	86.20	83.33	91.00

Clustering result was validated using precision and recall criteria. Table 6 shows the precision, recall and the accuracy of each cluster. Experimental results show that our approach performs effectively in clustering the data set.

Table 6. Performance measures of clusters

Attribute	Full Data	Cluster 0 (55.0)	Cluster 1 (28.0)	Cluster 2 (59.0)	Cluster 3 (62.0)
BMI value	0.3234	0.3442	0.2725	0.2379	0.4152
Disease of Pancreas	0.1	0.1273	0	0	0.4828
Background (Family History)	0.43	0	1	0	1
Sugar consumption	0.565	1	0	0	1
Alcohol consumption	0.1	0	0	0	0.4483

Figure 1 shows the visualization of each cluster in Weka. Here, cluster 0 is represent in blue, cluster 1 is represent in red , cluster 2 is represent in light blue and cluster 3 is represent in ash color.

According to the clustering data discovery, there are several decisions regarding diabetic disease. When the BMI value is increasing the level of diabetic is also increasing. If someone has covered the main 5 attributes in diabetic, he/she can have fasting glucose level up to 300. And also if someone has not covered any of these main 5 attributes in diabetic, he/she is in the normal level in fasting glucose. But even though any person has not covered all the main attributes but he/she has the relatives who have diabetic, there is a possibility to be in the impaired level in fasting glucose.

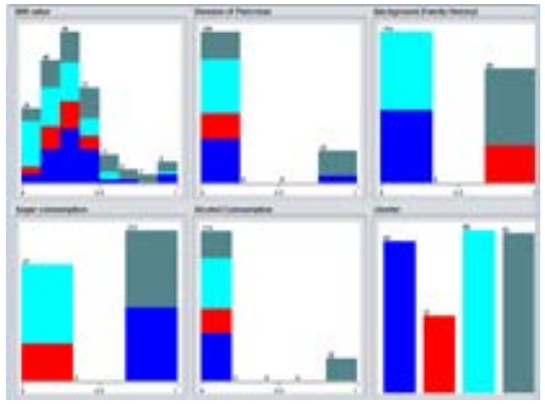

Figure 2. Visualizing the charts for most suitable attributes with fasting plasma glucose

Figure 2 shows how the most significant attributes (BMI value, Disease of Pancreas, Background (Family History), Sugar consumption and Alcohol Consumption) increased with the level of fasting plasma glucose.

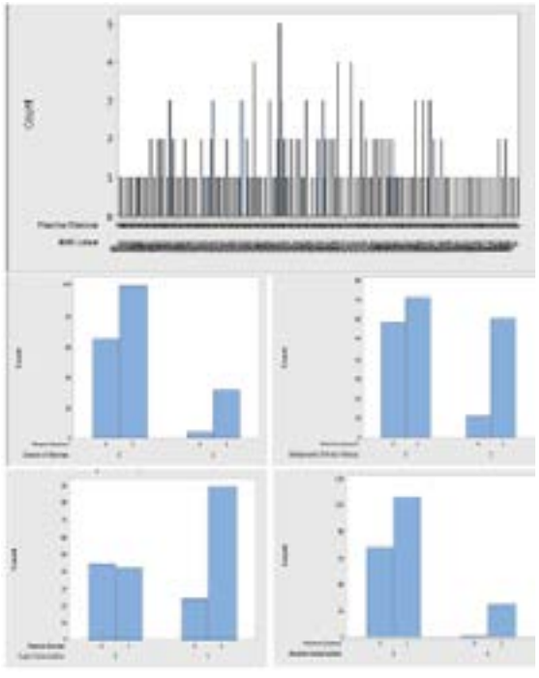

Figure 1. Visualization of Clusters

V. DISCUSSION AND CONCLUSION

The healthcare field is one of the most data enriched field in the world. As much as 30% of the entire world's stored data is generated in the health care environment. But unfortunately, the knowledge and learning getting from such kind of big data is very less. This research focused on those medical data. Every human being is once a patient who should deal with healthcare environment. So it is more important to holding a research actively with medical data. Aim is based on the value of predicting a disease and combining the computer system and information technology with the medical field. Predicting a disease is one of the major parts of dealing with medical data and the healthcare environment. Diabetic disease was selected here to predict in this research. Since diabetes is a chronic disease. An early prediction of the disease will save the patient life. Diabetes disease is the leading cause of death in the world over the past 10 years. In this research data mining was tremendously utilized to determine, analyse and further to visualize the useful information in diabetic. In this way data mining techniques are applied in medical data domain in order to predict diabetes and to find out efficient ways to treat them as well. Data mining is an interdisciplinary subfield of computer science is the computational process of discovering patterns in large data sets. It is also called as Knowledge Discovery in Databases which can help to support decision making in different fields including health care field. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. The discovery of knowledge from medical datasets is important in order to make effective medical diagnosis.

In this research, statistics, clustering and visualizing methods were selected for analysing tasks. Fasting plasma glucose test reports of 200 people were used with all related 14 attributes were considered. Among them 11 attributes were identified as the most affected attributes. Statistical model was created to check the possibility of being a diabetic patient using binary logistic regression in Minitab. K – means clustering was used to categorized the dataset in selected attributes. Each cluster belongs to the level of diabetic.

In our future work, we plan to conduct a thorough evaluation of our proposed cluttering approach and statistical model by comparing with existing works.

ACKNOWLEDGEMENT

This study was done by the department of Computing and Information Systems in faculty of Applied Sciences in Sabaragamuwa University of Sri Lanka. The medical information was gained from Government Hospital in Kiribathgoda, Sri Lanka. Authors would like to thanks to the doctors of the Government Hospital in Kiribathgoda, Sri Lanka, especially Mrs Anoma Gamage for the valuable support and help.

REFERENCES

Ferreira, D., Oliveira, A. and Freitas, A., (2012). Applying data mining techniques to improve diagnosis in neonatal jaundice. BMC medical informatics and decision making, 12(1), p.143.

Huang, Y., McCullagh, P., Black, N. and Harper, R., (2007). Feature selection and classification model construction on type 2 diabetic patients' data. Artificial intelligence in medicine, 41(3), pp.251-262.

http://apps.who.int. (2016). GLOBAL REPORT ON DIABETES. [ONLINE] Available at: http://apps.who.int/iris/bitstream/handle/10665 /204871/9789241565257_ eng.pdf;jsessionid =668AD6D16C59FD6CABA398693A61C579?sequence=1. [Accessed 10 July 2018].

https://www.niddk.nih.gov. (2013). Your guide to diabetes : type one and type two. [ONLINE] Available at: https://www.niddk.nih.gov/-/media/Files/Diabetes/YourGuide2Diabetes_508.pdf. [Accessed 10 July 2018].

Kelly, M. D. J. (2006). Diabetes. [Online] Available at: https://www.cdc.gov/media/presskits/aahd/diabetes.pdf. [Accessed 18 Jul. 2018].

Lowanichchai, S., Jabjone, S. and Puthasimma, T., 2006. Knowledge-based DSS for an Analysis Diabetes of Elder using Decision Tree. Faculty of Science and Technology Nakhon Ratchsima Raj abh at University, Nakhonratchasima, 30000.

Parthiban, G., Rajesh, A. and Srivatsa, S.K., (2011). Diagnosis of heart disease for diabetic patients using naive bayes method. International Journal of Computer Applications, 24(3), pp.7-11.

Rajesh, K. and Sangeetha, V., (2012). Application of data mining methods and techniques for diabetes diagnosis. International Journal of Engineering and Innovative Technology (IJEIT), 2(3).

Yuvarani, S. and Selvarani, R. (2016). An Analysis Of Decision Tree Models For Diabetes, International Research Journal of Engineering and Technology (IRJET), 3(11), 680 – 684.

pdf;jsessionid = 668AD6D16C59 FD6CA BA398693A61 C579?sequence =1. [Accessed 10 July 2018].

https://www.niddk.nih.gov. (2013). Your guide to diabetes : type one and type two. [ONLINE] Available at: https://www.niddk.nih.gov/-/media/Files/Diabetes/YourGuide2Diabetes_508.pdf. [Accessed 10 July 2018].

Kelly, M. D. J. (2006). Diabetes. [Online] Available at: https://www.cdc.gov/media/presskits/aahd/diabetes.pdf. [Accessed 18 Jul. 2018].

Lowanichchai, S., Jabjone, S. and Puthasimma, T., 2006. Knowledge-based DSS for an Analysis Diabetes of Elder using Decision Tree. Faculty of Science and Technology Nakhon Ratchsima Raj abh at University, Nakhonratchasima, 30000.

Parthiban, G., Rajesh, A. and Srivatsa, S.K., (2011). Diagnosis of heart disease for diabetic patients using naive bayes method. International Journal of Computer Applications, 24(3), pp.7-11.

Rajesh, K. and Sangeetha, V., (2012). Application of data mining methods and techniques for diabetes diagnosis. *International Journal of Engineering and Innovative Technology (IJEIT),* 2(3).

Yuvarani, S. and Selvarani, R. (2016). An Analysis Of Decision Tree Models For Diabetes, *International Research Journal of Engineering and Technology (IRJET),* 3(11), 680 – 684.