JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Comparative Performance Analysis of Data Mining in Diabetes

Aisha[1*], K. Solanki[2], S.Dalal[3], A. Dhankhar[4]

[1,2,3,4]Department of Computer Science and Engineering, Maharshi Dayanand University, Rohtak, India

[*]*Corresponding Author: aisha.dhankhar623@gmail.com, Tel.: +91-7015191462*

DOI: https://doi.org/10.26438/ijcse/v7i6.8894 | Available online at: www.ijcseonline.org

Accepted: 12/Jun/2019, Published: 30/Jun/2019

***Abstract*—** The technique used for mining the vital data from the pre-existent record known as data mining. It is used for diseases detection at an early stage in medical services. In medical issues, diabetes is a major worldwide problem from various deadliest diseases. "Around 422 million people worldwide are suffering from diabetes". The purpose of this research is to determine a prototype which can prophesy the possibility with a maximum accuracy of diabetes in patients. So to identify pre-diabetes using two (decision tree and naïve bayes) classification algorithms. The next main focus is to analyze the outcomes and ascertain which technique is more effective and superior from both of them. This paper (pinpointed on) is comparing data mining algorithms which are used for diabetes prognosticate.

***Keywords*—**Data Mining, Diabetes, Decision Tree, Naïve Bayes, WEKA

I. Introduction

A. *Data Mining:*

"Data mining is the process of extracting useful information. Basically, it is the process of discovering hidden patterns and information from the existing data". Data mining methods used in many fields for information collection. In health care industry data mining plays a crucial part in diseases prediction. It can reduce the risk of various deadliest diseases like heart diseases, diabetes, cancer, etc. In medical issues, diabetes is a major worldwide problem from various deadliest diseases.

It has wide applications such as CRM (customer relationship management), HR (Human Resource), GIS (Geographic Information System), Medical care, Brokerage and securities trading, Credit analysis, Government and defence, Computer hardware and software, Banking, Insurance.

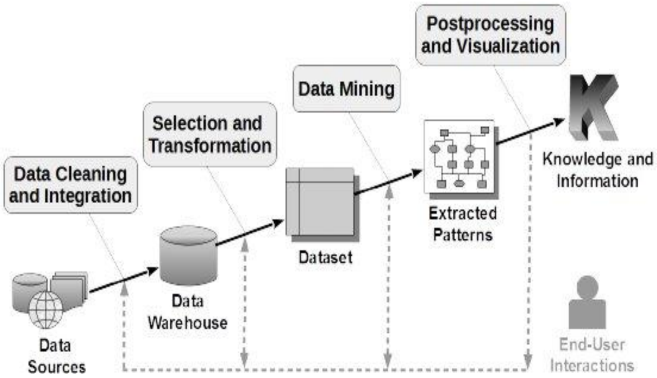

Postprocessing and Visualization
Data Mining
Selection and Transformation
Data Cleaning and Integration
Knowledge and Information
Extracted Patterns
Dataset
Data Warehouse
Data Sources
End-User Interactions

Fig. 1: Data mining process

B. *Diabetes:*

As per the WHO report (World Health Organization, April 2019), "India has close to 62 million people living with the diseases and is projected to have close to 70 million diabetics by 2025". In Asia (headed by India and China) has reached pandemic proportions. As per north worthy report found that risk of an untimely death has increased substantially, especially among adult and women. Nowadays, diabetes high-rise in India and China globally. "Throughout Asia, more than 230 million people are living with diabetes. More than one million individuals were followed for an average of 12.6 years". The researchers found that the risk of an untimely death has increased substantially in diabetes' patients.

Diabetes raises the danger of death, approximately twofold for all reasons. As per the study, the risk of an untimely death has increased substantially in Asia due to less availability of diabetes care. So, Asia populations require immediate implementation of diabetes controlling events.

Types of diabetes:

Type 1: This type caused due to lack of insulin production in the body, also known as "insulin dependent or childhood-onset diabetes."

Type 2: Enough insulin is not produced for maintaining the levels of glucose (sugar) in the body. Also known as "non-insulin-dependent or adult-onset diabetes." Inactiveness and overweight are the main reasons for this type.

Gestational: When the body doesn't produce enough insulin during pregnancy (for extra requirement) cause gestational diabetes. This creates a health risk for both mother and child. Both the mother and child can have high risk.

1. Diabetes in Women:

"Difficulty in conceiving, miscarriages, malformed babies, and an overall poor outcome in-pregnancies, are the major impacts of diabetes in women". One out of 10 women has diabetes, as per the IDF report (International Diabetes Federation). A significant number of whom don't approach medicinal services and don't aware of the ailment. 1 out of 7 infants has gestational diabetes — is a considerably more concerning issue is diabetes in pregnancy or uncontrolled diabetes before pregnancy. "Diabetes is also the ninth leading cause of death in women globally, causing 2.1 million deaths per year."

C. Classification:

It is a technique which divides objects into groups or classes on the bases of their resemblance and dissimilarities. It is like arrange books into library according to subject courses. Classification algorithms predict result according to a given input.

The training phase helps to find out the relationship between targets' values and predictors' values by using classification. Predicted and targeted values are compared for testing the phase. The given dataset is commonly divided into two parts: one is called training data (set), which is used to train data, and other is called testing data, which is unknown value.

Spam email analysis, disease diagnosis analysis, customer relation, and text classification, etc. are the application of classification in data mining. This research work show contrast between various classification techniques in data mining implements on recorded information.

1. Decision Tree:

A decision tree is a tree-like structure model. Just like the tree has various branches decision tree divides the database into subgroups. It has a decision node and a leaf node. Classes are leaf nodes. A top node called root in the tree. The root node has to splits into two branches, other nodes either have a branch or no branch. The decision tree has numerical and categorical both data.

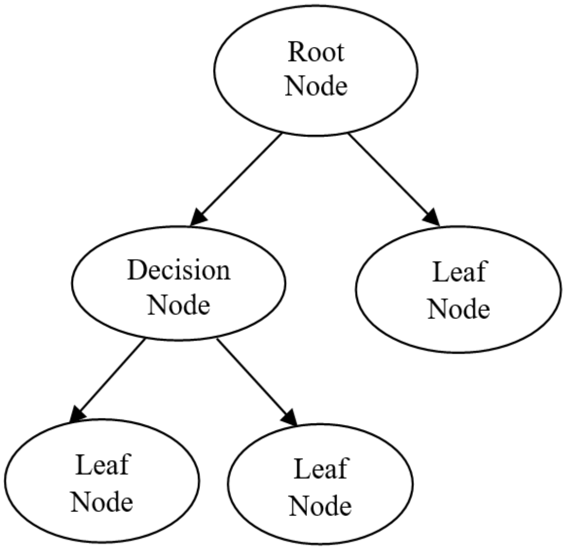


Fig. 2: Structure of the decision tree

Algorithm for decision tree:

Step 1: Attributes which is used for training data should not in continuous value, i.e. have to in discrete value. Put all of the data (training set) into a single tree node.

Step 2: End the tree if all of the instances are of the same class.

Step 3: Choose the attribute which has optimal value for decision node.

Step 4: Based on attributes, value divides the node.

Step 5: End if following constraints encountered otherwise go to step 3:

a) If attributes belong to the same class after splitting and no more fork required.

b) If no attributes remained for further division.

ID3:

In 1986 J. Ross Quinlan developed iterative dichotomise 3 (ID3) algorithm. This use entropy and information gain for attribute selection. This algorithm generates the smallest decision tree. The amount of information calculation in an attribute known as entropy.

Algorithm for ID3:

1. Entropy is calculated:

E(S) = - $\sum_{j=1}^{c} \mathbf{Pj} * \mathbf{log\ Pj}$

- S: a set of instances.
- j: class from 1 to m.
- c: number of classes in S.
- P_i: no. of instances in class u / no. of instances in S

Entropy used to calculate the quantity of information in measures the amount of information in an attribute.

2. Information Gain G(Y) is calculated:

G(Y) = $\mathbf{E(S)} - \sum_{i=1}^{n}(\mathbf{Ti} / \mathbf{T}) * \mathbf{E(Si)}$

- S_i: the subset of S with the same instance attribute i.
- T: total no. of instances in S.
- T_i: no. of objects in subset S_i.
- n: total no. of subsets in S.

An attribute which has maximum gain value is selected for splitting.

Limitations of this algorithm are not suitable for a large amount of domain. Such attributes perform no prediction because entropy for such is very less.

C4.5:

The limitations of the ID3 algorithm is reduced by using the C4.5 algorithm. Features of C4.5 are:

- It takes both discrete and continues values.
- The missing information is managed.
- It deals with dissent costs.
- It uses pruning after tree generation.

J48:

In WEKA for generating a decision tree by using the C4.5 algorithm known as J48 classifier. This algorithm is highly used in mining research.

2. Naive Bayesian

It is based on Bayes' theorem. This algorithm is easy to create. It has no complex parameters and because of this, it is useful for the large sized database. In spite of its ease, the Naive Bayesian classifier gives amazingly good results. It is used globally and overtook other advanced algorithms.

Bayesian classifier use hypothesis to defining the class for given data. After that, find out the probability of the hypothesis is true.

P(d|w) = (P(w|d) * P(d)) / P(w)

Where

P(d|w) = P(w_1|d) * P(w_2|d) …… P(w_n|d) * P(w|d) * P(d)

- P(d|w): posterior probability of target class when predictor (attribute) given.
- P(d): the prior probability of class.
- P(w|d): the posterior probability of predictor when class is given.
- P(w): the prior probability of predictor.

II. Related Work

Many researchers stated and bared that data mining is a blessing in the field of medical science, Sneha, Gangil (2019) and Indoria, Rathore (2018) and Balpande, Wajgi (2017). Juliet and Bhavadharani (2019) improve mining models' exactness by making methods versatile for many records (dataset). The dataset was pre-processed by using K-means clustering method. They subsequently apply the K-means clustering method. The suggested method provides a superior outcome.

According to Ghorbania, Ghousia (2019) heart disease, diabetes and breast cancer are high risks bearing in healthcare. They use data mining application in healthcare. Surya proposed prediction of diabetes mellitus by utilizing big data analytics. They used the various dataset for prediction. For the calculation, authors used decision tree algorithm and the result is pretty good. It is conceivable to future improve the diabetes mellitus to utilize any AI techniques (for the effectiveness of the analysing). Use ANN in light of the fact that it indicates a better outcome when contrast with SVM.

Martin *et.*al (2018) told that artificial intelligence (AI) is a powerful tool for estimate and inhibition of problems related to diabetes. Outcomes demonstrations that AI approaches are being gradually recognized as appropriate in day-to-day practice use in clinical, for the self-management of diabetes as well.

Explained machine learning structure for predicting diabetes, observing and application (DPMA). They functioned at five most significant AI grouping procedures were considered for anticipating diabetes. The performance of the classification techniques is examined by using different estimation measures, Mahmud *et.*al 1(2018), Singh *et.*al (2017).

Hina *et.*al (2017) authors proposed that diabetes caused 1.5 million worldwide passing every year around the world. Diabetes mellitus put fourth among non-transferable disease NCDs. The development in digitization has raised various difficulties, particularly with regards to robotized content investigation and to utilize some AI procedures to help humanity for anticipating the non-transferable sicknesses like diabetics.

III. Methodology

A. WEKA:

It stands for Waikato Environment for Knowledge Analysis. It is based on java language. WEKA is a data mining tool which has built-in algorithms like classifiers, clustering algorithms, tree algorithms, etc. which is used for data analysis. It is easy to use and give a clear result.

Features of Weka are:

- It is freely available.
- It is portability, as written in java language and it runs on any system because written in java language.
- Information pre-processing and displaying strategies gathered at one platform.
- It has a GUI which makes it convenient.

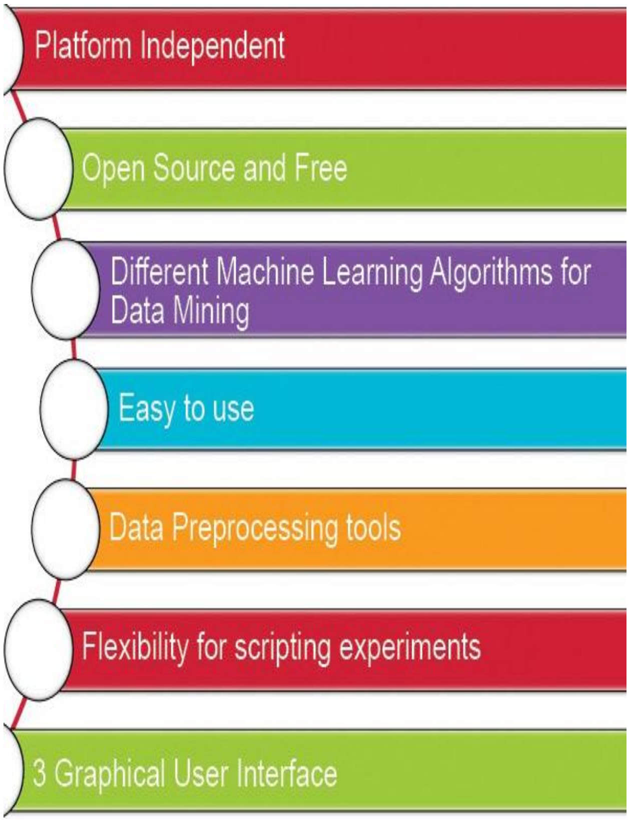
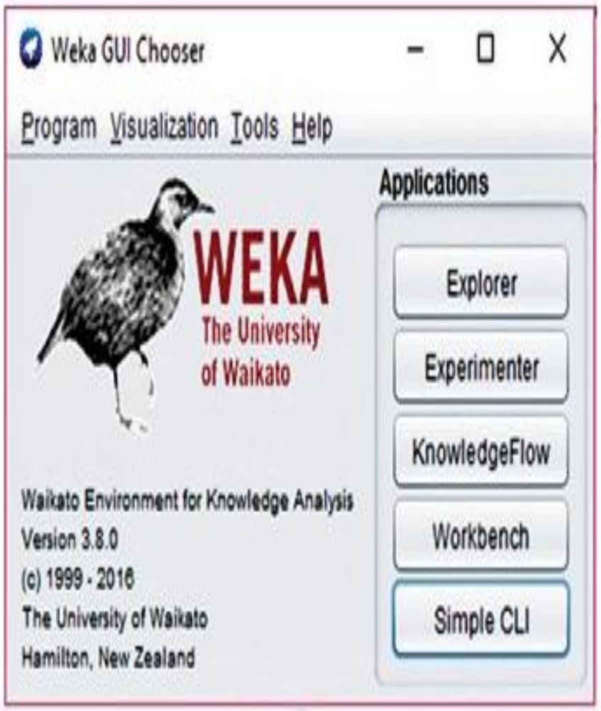

Platform Independent
Open Source and Free
Different Machine Learning Algorithms for Data Mining
Easy to use
Data Preprocessing tools
Flexibility for scripting experiments
3 Graphical User Interface

Fig. 3: Features of WEKA tool.



Waikato Environment for Knowledge Analysis
Version 3.8.0
(c) 1999 - 2016
The University of Waikato
Hamilton, New Zealand

Fig. 4: Opening screen of WEKA.

Table 1. Dataset's Meta information

Sr.No.	**Meta Data**		
	Attribute Name	***Description***	***Type***
1.	Pregnancies	Number of times pregnant	Numeric
2.	Glucose	Plasma glucose concentration a 2 hour in an oral glucose tolerance test	Numeric
3.	Blood pressure	Diastolic blood pressure (mm Hg)	Numeric
4.	Skin thickness	Triceps skin fold thickness(mm)	Numeric
5.	Insulin	2 hour Serum insulin (mu u/ml)	Numeric
6.	BMI	Body mass index (weight in kg /(height in m)^2)	Numeric
7.	Diabetes pedigree function	Diabetes extraction function	Numeric
8.	Age	Patient's age in years	Numeric
9.	Outcome	Diagnosis of diabetes (output)	Categorical: 0 or 1

The above table shows the description of diabetes database consisted of 800 instances and 9 attributes. All patients are here females at least 21 years old. The database is taken from kaggle (pima indian diabetes database).

IV. RESULTS AND DISCUSSION

When the classification model is trained and tested by J48 tree algorithm, the following results are obtained:

Confuse Matrix resulting from J48 algorithm applied to the dataset:

```
   a      b   <-- classified as
 107    13 |   a = 0
  22    35 |   b = 1
```

In a confusion matrix, instance on the left of the top and right from the bottom are correctly classified instances and instances on the right of the top and left of the bottom are inaccurately classified instances.

As shown, 107 are correctly classified instances as 0, 35 are correctly classified instances as 1, 13 are incorrectly classified as 1 when actually they are 0 and 22 are incorrectly classified as 0 when actually they are 1.

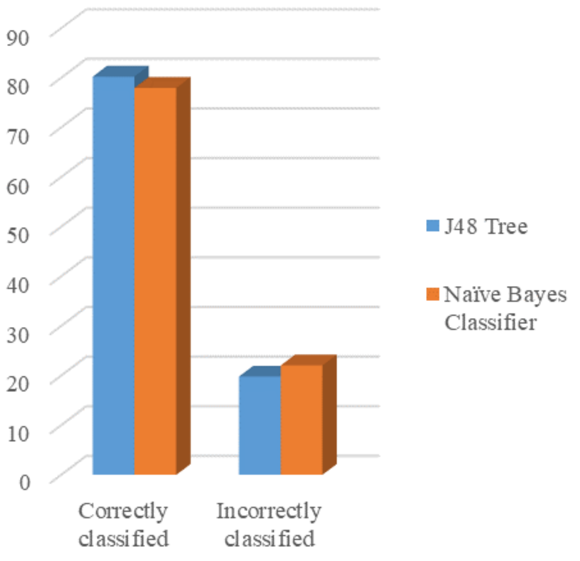
When the classification model is trained and tested by the naïve bayes algorithm, the following results are obtained:

Confuse Matrix resulting from naïve bayes algorithm applied on the dataset:

```
  a      b   <-- classified as
 101    19 |   a = 0
 20     37 |   b = 1
```

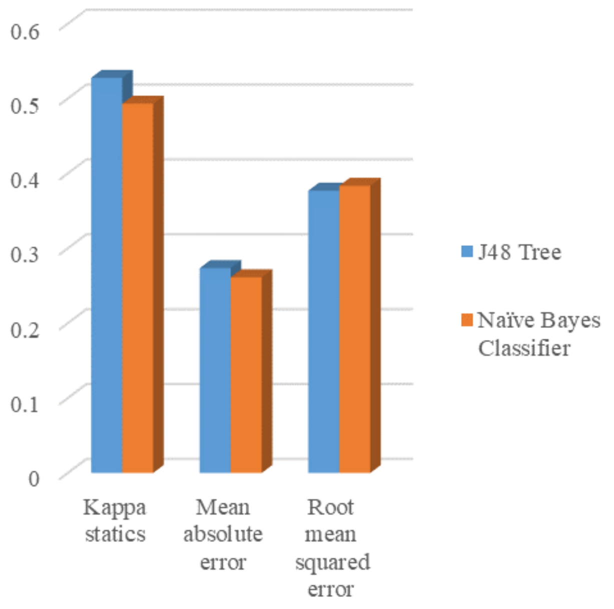
In a confusion matrix, instance on the left of the top and right from the bottom are correctly classified instances and instances on the right of the top and left of the bottom are inaccurately classified instances. As shown, 101 are correctly classified instances as 0, 37 are correctly classified instances as 1, 19 are incorrectly classified as 1 when actually they are 0 and 20 are incorrectly classified as 0 when actually they are 1.

Graph 1: Correctly and incorrectly classified instances of algorithms.



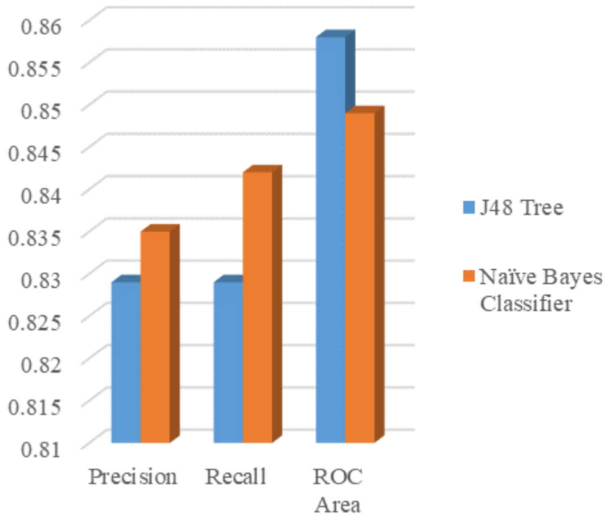
As shown in graph 1, 80.226% intances are correctly classified by J48 tree, whereas naïve bayes classifier classified 77.9661% instances correctly. 19.774 are incorrectly classified by J48 tree whereas naïve bayes classifier classified 20.339 instances incorrectly.

Graph 2: Kappa statics, mean absolute error, root mean squared error



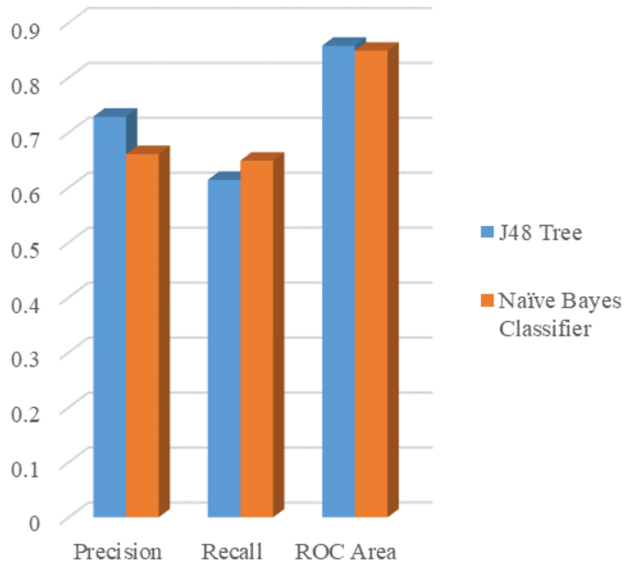
As shown in graph 2, Kappa statics from J48 tree is 0.5276, whereas 0.4931 from naïve bayes classifier. Mean absolute error from J48 tree is 0.2735 whereas 0.2614 from naïve bayes classifier. Root mean squared error from J48 tree is 0.3769 whereas 0.3834 from naïve bayes classifier.

Graph 3: Precision, Recall, ROC area for class 0



As shown in graph 3, the Precision value from J48 tree is 0.829, whereas 0.835 from naïve bayes classifier. Recall value from J48 tree is 0.829 whereas 0.842 from naïve bayes classifier. ROC area value from J48 tree is 0.858 whereas 0.849 from naïve bayes classifier.

Graph 4: Precision, Recall, ROC area for class 1



As shown in graph 4, the Precision value from J48 tree is 0.729, whereas 0.661 from naïve bayes classifier. Recall value from J48 tree is 0.614 whereas 0.649 from naïve bayes classifier. ROC area value from J48 tree is 0.858 whereas 0.849 from naïve bayes classifier.

Above graphs, results show that J48 tree algorithm is better than the naïve bayes classifier algorithm.

V. Conclusion and Future Scope

In this paper, two classification algorithms are explained and applied on data set required for diabetes (patients have diabetes and don't have diabetes) which shows how data mining helps in overcoming those challenges faced by diabetes and the classification model is trained and further tested. It will help them in choosing better data mining technique. Here Naïve Bayes and Decision Trees algorithms are used for extracting the information. For future result evaluation data refining is required as a large amount of data is exist that having the patients' record. First and foremost Naïve Bayes technique performance is evaluated, secondly J48 decision trees algorithm performance is evaluated, afterward ascertain the performance of which technique is more appropriate. Decision tree using J48 method is more efficient than naïve Bayes algorithm, concluded after performing both the techniques. The accuracy achieved through decision tree (J48 algorithm (80.226%)) is more than Naïve Bayes (77.9661%). So, it is concluded that the J48 decision tree algorithm is preferable than the Naïve Bayes algorithm.

References

[1] N. Sneha and Tarun Gangil,"Analysis of diabetes mellitus for early prediction using optimal featuresselection",(2019).https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0175-6

[2] Mrs. P. Laura Juliet1, T. Bhavadharani,"An Improved Prediction Model For Type 2 Diabetes Mellitus Disease Using Clustering And Classification Algorithms", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 06 Issue: 02, Feb 2019.

[3] Ramin Ghorbania and Rouzbeh,"Ghousia Predictive data mining approaches in medical diagnosis: A review of some diseases prediction", International Journal of Data and Network Science 3 (2019) 47–70.

[4] S.R.Surya,"Literature Survey On Diabetes Mellitus Using Predictive Analytics Of Big Data", International Journal of Advance Engineering and Research Development Volume 5, Issue 02, February -2018.

[5] Clare Martin, Antonio Martinez-Millana, Andrew Stranieri, Klerisson Paixao, Maurice Mulvenna, and Francisco Nuñez-Benjumea , "Artificial Intelligence for Diabetes Management and Decision Support: Literature Review", J Med Internet Rest 2018 May; 20(5): e10775.Published online 2018 May 30.

[6] S. M. Hasan Mahmud ,Md Altab Hossin,Md. Razu Ahmed, Sheak Rashed Haider Noori and Md Nazirul Islam Sarkar, "Machine Learning Based Unified Framework for Diabetes Prediction",Proceedings of the 2018 International Conference on Big Data Engineering and Technology Pages 46-50 ISBN: 978-1-4503-6582-6.

[7] Priyanka Indoria,Yogesh Kumar Rathore, "A Survey: Detection and Prediction of Diabetes Using Machine Learning Techniques", International Journal of Engineering Research & Technology (IJERT) Vol. 7 Issue 03, March-2018.

[8] Dr.D. Asir Antony Gnana Singh, Dr. E. Jebamalar Leavline, B. Shanawaz Baig, "Diabetes Prediction Using Medical Data", Journal of Computational Intelligence in Bioinformatics ISSN 0973-385X Volume 10,Number 1(2017) pp.1-8.

[9] Saman Hina, Anita Shaikh and Sohail Abul Sattar, "Analyzing Diabetes Datasets using Data Mining", Journal of Basic & Applied Sciences, 2017, 13, 466-471

[10] Vrushali Balpande, Rakhi Wajgi, "Review on Prediction of Diabetes using Data Mining Technique",(2017) International Journal of Research and Scientific Innovation (IJRSI) | Volume IV, Issue IA, January 2017 | ISSN 2321–2705.

[11] Nirmal Kaur, Gurpinder Singh, "A Review Paper On Data Mining And Big Data", International Journal of Advanced Research in Computer Science Volume 8, No. 4, May 2017.

[12] Shuja Mirza, Dr. Sonu Mittal and Dr. Majid Zaman,"A Review of Data Mining Literature", International Journal of Computer Science and Information Security(IJCSIS), Vol. 14, No. 11, November 2016.

[13] Ashish Kumar Dogra and TanujWala, "A Review Paper on Data Mining Techniques and Algorithms",International Journal of Advanced Research in Computer Engineering &Technology (IJARCET) Volume 4 Issue 5, May 2015.

[14] GyorgyJ.Simon,Pedro J.Caraballo,Terry M. Therneau,Steven S. Cha, M. Regina Castro and Peter W.Li "Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus," IEEE Transactions on Knowledge and Data Engineering,vol 27, No.1,January 2015.

[15] Sukhdev Singh Ghuman,"A Review of Data Mining Techniques", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.4, April- 2014, pg. 1401-1406

(2014).

[16] Dr.Zuber khan, shaifali singh and Krati Sexena,“Diagnosis of Diabetes Mellitus using K- Nearest Neighbor Algorithmin”, proceeding of International Journal of Computer Science Trends and Technology, vol.2 , July-Aug 2014.

[17] Mukesh kumari and Dr. Rajan Vohra,“Prediction of Diabetes Using Bayesia network,” in proceeding of International Journal of Computer Science and Information Technology vol. 5 , 2014.

[18] Dr. Pramanand Perumal and Sankaranarayanan, “Diabetic prognosis through Data Mining Methods and Techniques,” in proceeding of International Conference on Intelligent Computing Applications, vol.2, 2014.

[19] Satyanarayana Gandi and Amarendra Kothalanka,“An Efficient Expert System For Diabetes By Naïve Bayesian Classifier,” in proceeding of International Journal of Engineering Trends and Technology,2013.

[20] Shankaracharya, Devang Odedra, Subir Samanta, and Ambarish S. Vidyarthi1 et.al, “Computational Intelligence in Early Diabetes Diagnosis: A Review”,(2010). The Review of Diabetic Studies · January 2010 DOI: 10.1900/RDS.2010.7.252 · Source: PubMed.

[21] Ramkrishnan Shrikant and Rakesh Agrawal,“Fast Algorithms for mining association rule,”in proceeding of IEEE International Conference on Data Engineering,vol.16,2007.

[22] Chris Fleizach and Satoru Fukushima, “A naïve bayes classifier on 1998 KDD cup”,(1998).

[23] [Online]Available:Machine Learning Group at the University of Waikato. Weka 3: Data Mining software in Java. Retrieved September 4, 2016, from http://www.cs.waikato.ac.nz/ml/weka/.

[24] [Online]Available:https://www.thehindu.com/sci-tech/health/focus-on-women-and-diabetes/article 20393636.ece.

[25] [Online]Available:https://www.kaggle.com/uciml /pima-indians-diabetes-database.

[26] [Online]Available:https://www.ndtv.com/health/diabetes-indian-women-at-high-death-risk-from-diabetes-finds-study-2027180.

[27] [Online]Available:https://www.apollopharmacy.in/blog/indian-women-diabetes/.

[28] [Online]Available:http://www.searo.who.int/india/topics/diabetes_mellitus/en/.

[29] [Online]Available:https://www.womenshealth.gov/a-z-topics/diabetes.

[30] [Online]Available:https://en.wikipedia.org/wiki/Weka (machine_learning).

[31] [Online]Available:https://courses.cs.washington.edu/courses/csep521/07wi/prj/leonardo_fabricio.pdf