Review on analysis of data mining tools for diabetes Using PIMA

R. Rifat Ameena[1], Research Scholar, Department of Computer science, M.V. Muthiah Government Arts College for Women, Dindigul & India, rifat.ameena7869@gmail.com

B. Ashadevi[2], Assistant Professor, Department of Computer science, M.V. Muthiah Government Arts College for Women, Dindigul & India, asharajish2005@gmail.com

Abstract- Diabetes Mellitus is a chronic disease associated with abnormally high levels of the blood sugar levels over a prolonged period. Due to the raised blood sugar levels Patients with diabetes are affected with infections, heart disease, Nephropathy, Neuropathy, Retinopathy, Polycystic Ovarian Syndrome, Gastro paresis and depression. So, diabetes is one of the serious health issues even in developed countries. Usually diabetes is diagnosed by physical examination and blood sugar test. But it takes more cost, time and it does not give accurate results. To overcome this problem various machine learning techniques such as classification, clustering and regression are implemented using New and sophisticated Big Data Analytics tools. These tools in data mining helps to reduce cost and increase the efficiency of treatment for diabetes. In this paper the main focus is to make detailed survey of data mining tools in the field of diabetes research, sources start from 2013 onwards. The aim of this paper is to analyze and compare different data mining tools that are used to predict diabetes. The research presented here is a survey focused mainly on data mining tools such as Weka, Rapid Miner, R Studio, Tanagra, MATLAB, Python and sharper light. The dataset chosen for experimental simulation is PIMA Indian Diabetes Dataset. PIMA are people of Indian American origin. The dataset has taken 768 instances from PIMA Indian Dataset to determine the accuracy of the data mining tools used for prediction of diabetes. Hence, this research paper concentrates on the overall survey of various datamining tools that are used to Detect and Prevent the complications of diabetes at the early stage.

Keywords —PIMA, Diabetes, machine learning, Data mining, Big Data Analytics Tools, Tanagra.

I. INTRODUCTION

Diabetes is a chronic disease resulting due to acquired deficiency in the production of insulin by the pancreas. such a deficiency causes increased amount of glucose in the blood, which in turn damage the blood vessels and nerves. It causes retinopathy, neuropathy, cardiovascular disease, kidney failure, diabetic foot disease and it also give rise to several problems during pregnancy. Worldwide, diabetes is a leading public health concern which affects over 422 million people. World Health Organization (WHO) believes that nearly 8.8 percent of adult population worldwide has diabetes. This may rise to 9.9 percent by the year 2045 [26]. Diabetes is more prevalent in developing countries and disadvantaged minorities. Diabetes is a non-communicable that is a threat to health and human development [25]. Due to lack of insulin production in the body patient with diabetes are treated with insulin injection with tablets. The aim of this study is to investigate the performance of different classification methods using WEKA, Rapid miner, R Studio, TANAGRA, python, sharper light and MATLAB tool on Pima Indian Diabetes Dataset. Data mining is a valuable asset for diabetic research because it can find knowledge from a huge amount of diabetic related data. Modern technology like Big Data and Analytics is helping in the control and treatment of diseases such as diabetes. There is 'Big Data' huge data that cannot be analyzed using conventional statistical tools. These can aid immensely in accumulation of unexplored medical knowledge for treatment of diabetes.

A major problem in medical science is to attain the correct diagnosis of certain important information. For the diagnosis, generally many tests are done that involve clustering or classification of large scale data [19]. All of these test procedures are needed to reach the ultimate diagnosis. However, on the other hand, too many tests could complicate the main diagnosis process and lead to the difficulty in obtaining the end results, particularly in the case where many tests are performed. This kind of difficulty could be resolved with the aid of machine learning which could be used directly to obtain the end result with the aid of several artificial intelligence techniques. The value of machine learning in healthcare is its ability to process huge datasets beyond the scope of human capability, and then

 DOI : 10.18231/2454-9150.2019.0027

reliably convert analysis of that data into clinical insights that aid physicians in planning and providing care, ultimately leading to better outcomes, lower costs of care, and increased patient satisfaction. Prediction of diabetes using data mining tools is an important task because it allows doctors to see which attributes are more important for diagnosis such as age, blood glucose, weight, symptoms etc. This will help the doctors diagnose the disease more efficiently.

1.1 TYPES OF DIABETES

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. Hyperglycaemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels.

A. Type 1 diabetes

Type 1 diabetes (previously known as insulin-dependent, juvenile or childhood-onset) is characterized by deficient insulin production and requires daily administration of insulin. The cause of type 1 diabetes is not known and it is not preventable with current knowledge. Symptoms include excessive excretion of urine (polyuria), thirst (polydipsia), constant hunger, weight loss, vision changes, and fatigue [21].

B. Type 2 diabetes

Type 2 diabetes (formerly called non-insulin-dependent, or adult-onset) results from the body's ineffective use of insulin. Type 2 diabetes comprises the majority of people with diabetes around the world, and is largely the result of excess body weight and physical inactivity. Symptoms may be similar to those of type 1 diabetes, but are often less marked. Until recently, this type of diabetes is seen only in adults but it is now also occurring increasingly frequently in children.

C. Gestational diabetes

Gestational diabetes is hyperglycaemia with blood glucose 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 and their children are also at increased risk of type 2 diabetes in the future.

II. DATA MINING TECHNIQUES

The advancement in the field of Information technology has lead to large number of databases in various areas. Due to this, there is a need to store and manipulate important data which can be used later for decision making. Data Mining is the process of extracting useful information and patterns from huge amount of data [23]. Data Mining includes collection, extraction, analysis and statistics of data. It yields improved predictions. One of the most important task in Data Mining is to select the correct data mining tools and techniques. A generalized approach has to be used to improve the accuracy and cost effectiveness of using data mining techniques. There are basically seven main Data Mining techniques namely Clustering, Visualization, Decision Tree, CART, Neural Networks, Association and Classification.

A. Clustering

Clustering is one of the oldest techniques used in Data Mining. Clustering is the process of finding data that are similar to each other. This helps to understand the differences and similarities between the data. This is also called as segmentation. The most popular clustering algorithm is K Nearest Neighbor [2]. K Nearest neighbor technique is very similar to clustering. Each of the separated clusters contains number of samples that may belong to either first class which is (diabetic) or to the second one which is non-diabetic [16].

B. Visualization

Visualization is the most useful technique which is used to discover data patterns [20]. This technique is used at the beginning of the Data Mining process. visualization is a technique used for finding hidden patterns. In data mining. It converts Poor data into good data.

C. Decision Tree

A decision tree is a predictive model that looks like a tree structure. In this technique, each branch of the tree is considered as classification question and the leaves of the trees are considered as partitions of the dataset related to that particular classification. This technique can be used for exploration analysis, data pre-processing and prediction work [11]. Hence, selection of decision tree will be an appropriate choice for early prediction of Diabetes symptoms [21].

D. CART

CART which stands for Classification and Regression Trees is a data exploration and prediction algorithm which selects the questions in a more complex way [18]. It chooses one best question which is used to split the data into two or more segments.

E. Neural Network

This technique is mostly used in the starting stages of the data mining technology. Artificial neural network was formed out of the community of Artificial intelligence [5]. Neural networks are very easy to use as they are automated to a particular extent and because of this the user is not expected to have much knowledge about the work or

 DOI : 10.18231/2454-9150.2019.0027

database. Multi-Layer Perception is one of the commonly used Neural Network algorithms [6]. Deep learning neural network filters the data accordingly [30].

F. Association

Association technique helps to find frequent items that appears together in a dataset. It discovers the hidden patterns in the data sets and it is used to identify the variables and the frequent occurrence of different variables that appear with the highest frequencies.

G. Classification

Classification is a machine learning based data mining technique. It is used to classify each information in a set of data into one of predefined set of groups or classes. It makes use mathematical techniques such as decision trees, linear programming, neural network and statistics to classify the data into different groups. Newly developed classification techniques provide more intelligent methods for effective prediction of diseases [4]. Different types of classification techniques include Support vector machine [10], discriminant analysis, naive based, decision trees, linear and non-linear regression.

III. DATA MINING TOOLS

Data mining tools makes Big Data more manageable. It helps to analyze large amount of data that is generated every second. Hence Data mining tools helps to explore the unknown patterns such as cluster analysis, anomaly detection and association rule mining. These tools rely on computer algorithms for effective decision making. The Diabetes patients are prone to various diseases. hence early detection and prediction is important to control the disease using various data mining tools.

A. WEKA

The Waikato Environment for Knowledge Analysis (WEKA) is an open source software and machine learning toolkit introduced by Waikato University, New Zealand. WEKA supports several standard data mining tasks like data pre-processing, clustering, classification, regression, visualization and feature selection New algorithms can also be implemented using WEKA with existing data mining and machine learning techniques. WEKA has various sources for loading data, including files, URLs and databases [1]. It supports file formats include WEKA"s own ARFF format, CSV, Lib SVMs format, and C4.5's format. In WEKA, many evaluation criteria are also provided such as confusion matrix, precision, recall, true positive and false negative, etc. Some of the benefits of WEKA tool includes Open source, platform independent and portable, graphical user interface and contains very large collection of different data mining algorithms. Some algorithms were selected to establish the prediction model for type 2 diabetes [3]. As per the data given in the Table.1, Gaganjot Kaur and Amit Chhabra [1] compared the accuracy and error rate of various data mining algorithms such as Naive Bayes, MLP, Random forest, Random Tree and Modified J48 using WEKA and MATLAB. Among these Modified J48 classifier produced a higher accuracy of 99.87%.

SNO	AUTHOR	METHOD	ACCURACY
1	Gaganjot Kaur &Amit Chhabra, [1]	Modified J48 algorithm	99.87%
2	Han Wu, shengqi Yang, zhangqin hunag, jian he, xiaoyi Wang, [3]	Improved k means algorithm and logistic regression algorithm	90%
3	Tuba pala & Ali Yilmaz Camurcu, [4]	Logistic Regression Model	98.60%
3	Rashedur, M.Rahuman, Farhana Afroz, [6]	J48graft classifier	81.66%
4	Uswa Ali zia, naeem khan, [7]	Decision tree J48 graft	94.44%
5	Arwa Al-Rofiyee,Maram AlNowiser,Nasebi h Al-Mufadi, [9]	Multi-layer perception	97.61%
6	Dr. D. Ashok Kumar and R. Govindasamy, [13]	SVM regression naïve Bayes decision table	79.81%
7	c.s kanimozhi selvi, s.v kogilavani,s.malli ga,d.jayaprakash, [18]	Naive Bayes algorithm	74.5%
8	Deepti Sisodia,Dilip Singh Sisodia, [19]	Naive Bayes algorithm	76.30%
9	Saman Hina,Anita Shaikh and Sohail Abul Sattar, [20]	multi-layer perception	81.82%

Table.1 Accuracy rate produced by using WEKA tool in Pima Indian Diabetic Dataset.

B. RAPID MINER

RAPID MINER (RM) is open source software which provides a good environment for data mining processes. It is used to construct the dataflow with the help of drag-and-drop facility. It supports various file formats. In rapid miner Regression, classification and clustering tasks can be performed easily with different learning algorithms [5]. Rapid Miner supports a large number of the classification and regression algorithms, decision trees, association rules, clustering algorithms, and many features are available for data pre-processing, normalization, filtering and data

 DOI : 10.18231/2454-9150.2019.0027

analysis. It can import data from different traditional and standard databases. It is one of the best predictive analysis systems. It was developed by the company with the same name as the Rapid Miner. It is written in JAVA programming language [4]. It provides an integrated environment for deep learning. This tool can be used for business applications, commercial applications, training, education, health care etc. It has a client/server model as its base. Rapid Miner comes with template based frameworks. Also, it enables speedy delivery with a reduced number of errors. According to the data given in the Table.2. Tuba pala &Ali Yilmaz camurcu using the diabetes dataset have separated them in to two different clusters diabetic and non-diabetic using k means clustering algorithm. after this clustering process, clustering operation was done on the remaining data using support vector machine, Naïve Bayes, Decision Trees, Artificial Neural Networks, Multilayer perceptron and Logistic Regression was used in the process of classifying in Rapid miner. Among this Multilayer perceptron gave highest accuracy rate of 99.44%.

SNO	AUTHOR	METHOD	ACCURACY
1	Tuba pala &Ali Yilmaz camurcu, [4]	Multi-Layer Perception(MLP)	99.44%
2	S.Priya,R.R.Rajalaxmi,[5]	Neural Network	97.93%
3	Basharat Naqvi et.al [17]	decision tree ID3	84.94%

Table.2 Accuracy rate produced by using Rapid miner in Pima Indian Diabetic Dataset.

C. R STUDIO

R, sometimes called 'the superstar of free data mining', is a free, open source software [15] easy to use for people with little to no previous experience with programming. It can run on a wide variety of platforms including Mac and Windows. It allows to Manipulate the data, Visualize the data, Analyze the data. R has thousands of packages to perform a statistical computation and analysis [2]. In accordance with the data in the Table.3, David valls lanaquera [15] using SVM (support vector machine) technique produced a highest accuracy rate of 94%.

SNO	AUTHOR	METHOD	ACCURACY
1	Aanurag kumar sri vastava,chandankumar,neha mangla,[2]	KNN Algorithm	79%
2	David valls lanaquera[15]	SVM	94%
3	Aakansha Rathore,Simran Chauhan,Sakshi Gujral,[21]	SVM Classifier	82%

Table.3 Accuracy rate produced by using R Studio in Pima Indian Diabetic Dataset.

D.TANAGRA

TANAGRA is a free open source data mining software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. TANAGRA is more powerful, it contains some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and non-parametric statistics, association rule, feature selection and construction algorithms. The main purpose of Tanagra project is to give researchers and students an easy-to-use data mining software, conforming to the present norms of the software development in this domain (especially in the design of its GUI and the way to use it), and allowing to analyze either real or synthetic data [6]. As per the data given in the table. 4 Rashedur, M.Rahuman, Farhana Afroz [6] detect the used various data mining tools such as WEKA,MATLAB AND TANAGRA to detect diabetes. In TANAGRA, Naïve Bayes classifier provides accuracy of 100% with training time 0.001 seconds. They perceived that TANAGRA machine learning tool is the best compared to WEKA and MATLAB.

SNO	AUTHOR	METHOD	ACCURACY
1	Rashedur, M.Rahuman, Farhana Afroz, [6]	Naïve Bayes classifier	100%
2	Dr.V.Karthikeyani et.al,[11]	PLS-LDA	74%
3	P. Radha, Dr. B. Srinivasan, [23]	C4.5 Decision Tree	86%

Table.4 Accuracy rate produced by using TANAGRA in Pima Indian Diabetic Dataset.

E.MATLAB

When doing data mining, a large part of the work is to manipulate data. MATLAB has a lot of toolboxes for data mining [16] and it makes the coding short. when manipulating data, MATLAB is definitely better. It is normal since it is done to work with matrices (Matrix Laboratory). As per the data given in the Table.5 Mustafa s.kadhm, Ikhlas watan Ghindawi,Duaa Enteeshamhawi[16] used K-means clustering to get best separation points and

 DOI : 10.18231/2454-9150.2019.0027

accurate results. The experiment result achieved by them using k -means clustering is 98.7%.

SNO	AUTHOR	METHOD	ACCURACY
1	Rashedur,M.Rahuman , Farhana Afroz[6]	adaptive neuro fuzzy inference system(ANFIS)	78.79%
2	V. Anuja Kumari, R.Chitra[10]	SVM (RBF kernel)	78%
3	Thirumalaimuthu Thirumalaiappan Ramanathan,Dharmendra Sharma[12]	SVM fuzzy set	96%
4	Ankita Parashar Kavita Burse Kavita Rawat, [14]	Support vector machine(SVM)	75.65%
5	Mustafa s.kadhm, Ikhlas watan Ghindawi,Duaa Enteeshamhawi,[16]	K means clustering	98.7%
6	Vaishali Jain,Supriya Raheja,[22]	Fuzzy verdict mechanism	87.2%

Table.5 Accuracy rate produced by using MATLAB in Pima Indian diabetic dataset.

F. PYTHON

Python is a free and open source language, it is most often compared to R for ease of use. python can connect to data base systems and It can do complex analysis in minutes [24]. As per the data given in the table.6 Shaksham Kapoor and Krishna priya [24] used artificial neural network with hyper parameter tuning. This method gave an accuracy of 98.70%.

SNO	AUTHOR	METHOD	ACCURACY
1	Shaksham Kapoor,Krishna priya,[24]	Artificial neural network	98.70%
2.	Akm Ashiquzzaman et.al [30]	neural network	88.41%

Table.6 Accuracy rate produced by using Python in Pima Indian Diabetic Dataset.

G. SHARPER LIGHT

IV. DATASET

Sharper Light Reporting Software creates complex and interactive reports. It connects to multiple sources by using dynamic data access technique [8]. As per the data given in table.7 Veena Vijayan, V. Aswathy Ravi Kumar used Amalgam KNN algorithm which gave an accuracy greater than 80%.

SNO	AUTHOR	METHOD	ACCURACY
1	Veena Vijayan. V. Aswathy Ravi Kumar, [8]	Amalgam KNN algorithm	>80%

Table.7 Accuracy rate produced by using Sharper Light in Pima Indian Diabetic Dataset

The proposed work makes use of PIMA Indian Diabetes Data-set. It is published by National Institute of Diabetes and Digestive and Kidney Diseases. This Data set is primarily concerned with the women health. Here, 768 instances of almost 21 years of age of women is collected and various parameters are defined [29]. The 9 attributes that are defined in this dataset are

1. Number of times pregnant 2. Plasma glucose concentration 2 hours in an oral glucose tolerance test 3. Diastolic blood pressure (mm Hg) 4. Triceps skin fold 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) 9. Class variable (0 or 1).

 DOI : 10.18231/2454-9150.2019.0027

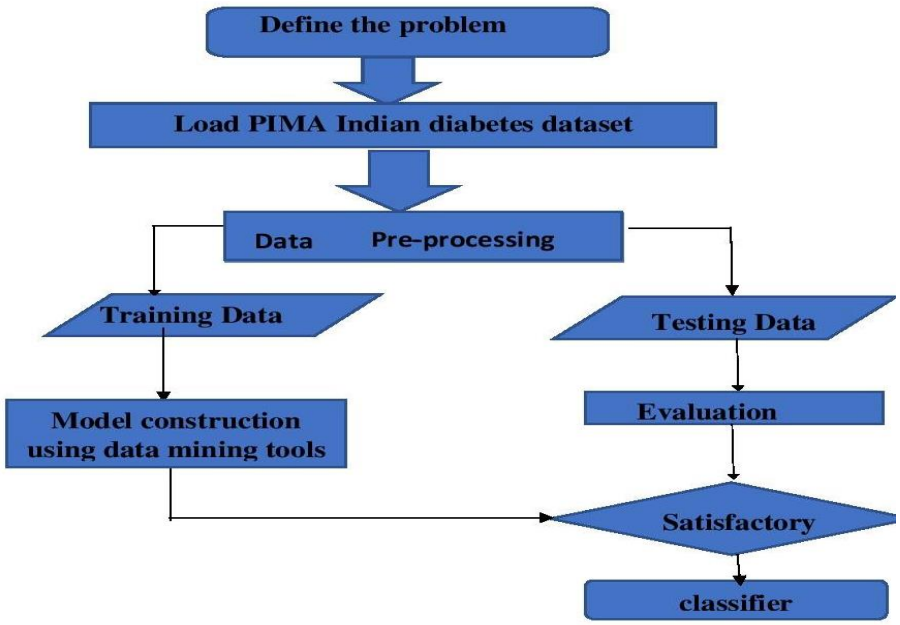

Fig.1 Framework for diabetes prediction

Fig.1 shows about the various process that are followed for diabetes prediction. In the first step define the problem and then load the dataset into the data mining tool. Then the next step is to carry out the preprocessing with the obtained dataset. After that the data is split for training and testing and then the data is evaluated. once after evaluating the data, finally the algorithm is applied on different data mining tools to predict whether the patient is diabetic or non-diabetic.

V. COMPARITIVE ANALYSIS OF DATA MINING TOOLS FOR DIABETIC PATIENTS

SNO	DATA MINING TOOLS	AUTHOR & YEAR	DATA MINING TECHNIQUES	REMARKS
1.	**WEKA & MATLAB**	Gaganjot Kaur &Amit Chhabra, [2014]	Naive Bayes, MLP, Random Tree, REP tree, RAD, Random Forest, J48, Modified J48 Classifier	Modified J48 algorithm produced an accuracy of 99.87%The data mining tool WEKA is used as application programming interface(API) of MATLAB.
2.	**R Studio**	Aanurag kumar sri vastava,chandan kumar,neha mangla,[2016]	KNN Algorithm	KNN algorithm was used to make the prediction model by using R tool. By using this model, they found the number of sample predictions made correctly. KNN algorithm produced an accuracy of 79% in this dataset.
3.	**WEKA**	Han Wu, shengqi Yang, zhangqin hunag, jian he, xiaoyi Wang, [2018]	HPM, AMMLP, J48, Hybrid model, MLP, Logistic, SGD, ELM, Naive Bayes, Cart, KNN	This paper is aimed to establish a prediction model for type 2 diabetes patients. The model is comprises of improved k means algorithm and logistic regression algorithm. The Pima Indian Diabetes Dataset and WEKA environment were used to compare the results. It produced an accuracy of 90%
4.	**Weka and Rapid miner**	Tuba pala &Ali Yilmaz camurcu, [2014]	Decision tree, naïve Bayes, artificial neural network, SVM, LR, multilayer perception	Logistic Regression model gave the accuracy percentage of 98.60% in WEKA. While Multi-layer perception algorithm showed a higher accuracy rate of 99.10% in Rapid miner tool.
5.	**Rapid miner**	S.Priya, R.R.Rajalaxmi, [2013]	C4.5 classifier, neural network	A hybrid prediction model has been developed by using k-means clustering and C4.5 classifier-score normalization is applied on the dataset. A model is built by using neural network in Rapid miner.it produced a accuracy of 97.93%
6.	**WEKA, TANAGRA and**	Rashedur, M.Rahuman, Farhana Afroz [2013]	MLP, Bayes net, Naïve Bayes, J48 graft, fuzzy Lattice Reasoning, Jrip,	Nine selected classification algorithms are implemented in Weka, Tanagra and MATLAB. The best algorithm in WEKA is J48graft classifier with an accuracy of 81.33%. In

 DOI : 10.18231/2454-9150.2019.0027

	MATLAB		fuzzy inference system, adaptive neuro fuzzy inference system	TANAGRA, Naïve Bayes classifier provides accuracy of 100% and in MATLAB, adaptive neuro fuzzy inference system(ANFIS) showed an accuracy of 78.79% accuracy.
7.	**WEKA**	Uswa Ali zia, naeem khan, [2017]	K- nearest neighbor, naïve Bayes, decision tree J48, after boot strapping	The WEKA software was employed as mining tool. Bootstrapping resampling technique is used to enhance the accuracy of classifiers. Decision tree J48 graft produced a accuracy of 94.44% after boot strapping.
8.	**Sharper light**	Veena Vijayan. V,Aswathy Ravi kumar,2014,[8]	EM, KNN, K-means, amalgam KNN and ANFIS algorithm	Amalgam KNN algorithm that combines both the features of KNN and K means produced an accuracy rate greater than 80% in this dataset.
9.	**WEKA**	Arwa Al-Rofiyee, Maram AlNowiser, Nasebih Al-Mufadi, 2013 [6]	MLP	Multi-layer perception(MLP) is implemented in WEKA tool. It produced an accuracy of 97.61%
10.	**MATLAB**	V. Anuja Kumari, R.Chitra (2013)	SVM (RBF kernel)	SVM with Radial basis function kernel (RBF)is used for classification. The performance of SVM with RBF showed an accuracy of 78%.
11.	**TANAGRA**	Dr.V.Karthikeyani et.al (2013)	C4.5 Decision tree, Modified SVM, KNN, PNN, BLR, MLR, CRT, CS-CRT, PLS-DA and PLSLDA	The PLS-LDA was the best algorithm in this paper for automatic classification with an accuracy of 74%.
12.	**MATLAB**	Thirumalaimuthu Thirumalaiappan Ramanathan, Dharmendra Sharma (2015)	SVM Fuzzy system	For Diagnosis of diabetes mellitus (Type 2 diabetes) Fuzzy reasoning is used to classify the level of risks from data Pima diabetes dataset is trained and tested by using SVM and fuzzy system respectively. The experiments from the model showed an accuracy of 96%.
13.	**WEKA**	Dr. D. Ashok Kumar and R. Govindasamy (2015)	SVM REGRESSION BAYES NET NAÏVE BAYES DECISION TABLE	The results obtained in this paper indicates that decision table algorithm perform better than all other techniques with the help of feature selection with an accuracy of 79.81%.
14.	**MATLAB**	Ankita Parashar Kavita Burse Kavita Rawat (2014)	Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Feed Forward Neural Network (FFNN)	The experimental results obtained in this paper shows that SVM gives better performance as compared to Feed Forward NN. It produced an accuracy of 75.65%.
15.	**R studio**	David valls lanaquera[2018]	SVM	Support vector machine produced higher accuracy of 94% due to its reliability, sample size and redundancy.
16.	**MATLAB**	Mustafa s.kadhm, Ikhlas watan Ghindawi,Duaa Enteeshamhawi, [2018]	K mean clustering	By experiments, the k- nearest neighbor algorithm achieved high classification result of 98.7%. It uses K Mean cluster for eliminating the undesired data, thus reducing the processing time.
17.	**Rapid Miner**	Basharat Naqvi,Arshad liMuhammad Adam Hashmi, and Muhammad Atif,[2018]	Random forest, Decision tree ID3, Decision stump	This work found that the decision tree ID3 with an accuracy of 84.94% is the best technique for prediction of disease in diabetic patients by considering evaluation metrics such as accuracy, precision and recall.
18.	**WEKA With Hive**	c.s kanimozhi selvi, s.v kogilavani, s.malliga,d.jayaprakash,[2018]	Naïve Bayes, Decision tree, Decision stump, k star and Random forest	Big data tools such as Hadoop and Hive is used in this paper. Classifiers are trained efficiently in a supervised learning setting. Naive Bayes algorithm showed an accuracy of 74.5%
19	**WEKA**	Deepti Sisodia,Dilip Singh Sisodia[2018]	Naïve Bayes, SVM, Decision Tree	Results obtained in this work shows that Naive Bayes outperforms with the highest accuracy of 76.30% than other algorithms. Those results are verified using Receiver Operating Characteristic (ROC) curves in a proper and systematic manner.

 DOI : 10.18231/2454-9150.2019.0027

20	**WEKA**	Saman Hina,Anita Shaikh and Sohail Abul Sattar,[2017]	Naïve Bayes, MLP, J48, Zero R, Random forest, Logistic Regression	In terms of performance, it was found that multi-layer perception function is most effective with few error rates. It showed an accuracy of 81.812%
21	**R studio**	Aakansha Rathore,Simran Chauhan,Sakshi Gujral,[2017]	SVM, Decision Tree	Diabetes Detection and prediction is done in R studio with SVM Classifier. It produced 82% accuracy.
22	**MATLAB**	Vaishali Jain,Supriya Raheja,[2015]	Fuzzy Sets	Fuzzy verdict mechanism in Fuzzy Logic based Diabetes Diagnosis System (FLDDS) which is proposed for the diagnosis of diabetes. This method indicates that their performance is increased with the increase in the number of parameters. This method gave 87.2% of accuracy.
23	**TANAGRA**	P. Radha, Dr. B. Srinivasan, [2014]	C4.5, SVM, K-NN, PNN, and BLR	In this paper five classification techniques (C 4.5, SVM, K-NN, PLR, and BLR) are applied to predict the diabetes disease in patients. High computing time with improved accuracy and error rate is produced by C4.5 with an accuracy of 86%.
24	**Python**	Shaksham Kapoor,Krishna priya,[2018]	KNN, GBC, Optimized AN	The purpose of this study is to emphasize the importance of hyperparameter tuning in improving the accuracy of data mining techniques. After hyper parameter tuning Artificial neural network produced an accuracy of 98.70%.
25	**Python**	Akm Ashiquzzaman, Abdul Kawsar Tushar, Md. Rashedul Islam, and Jong-Myon Kim, [2018]	neural network with Multi-layer preceptor(MLP).	In this study, issue of overfitting is minimized by using the dropout method. Deep learning neural network produced an accuracy of 88.41%

VI. RESULTS AND DISCUSSION

Seven data mining tools such as Weka, Rapid miner, R Studio, Tanagra, MATLAB, python and Sharper light are analyzed and their performances are compared on PIMA Indian Diabetes data set using various algorithms. The results obtained from the given dataset classified into two classes i.e. patients with diabetes and without diabetes using various data mining techniques. The accuracy to predict the diabetes disease using different techniques is shown in graphical representation in the fig.2. Based on the results demonstrated, Naïve Bayes classifier provides highest accuracy 100% to predict the diseases using Tanagra tool. The performance of the algorithm is calculated using the Accuracy. Here, True positive and True Negative, False positive and False Negative parameters are taken to evaluate the equation. Gaganjot Kaur and Amit Chhabra compared classification techniques and found the modified J48 algorithm gives better accuracy 99.87% in prediction using Weka and MATLAB. Tuba pala &Ali Yilmaz camurcu using Rapid miner tool to predict the disease with 97.61% accuracy. The classification accuracy of data mining tools discovered by various authors on Pima Indian Diabetes data set is highlighted in Table.8.

Tools	Highest Accuracy produced by data mining tools	Precision	Recall
Weka	99.87%	0.759	0.763
Rapid miner	99.10%	0.735	0.724
R Studio	94%	0.684	0.582
Tanagra	100%	1	1
MATLAB	99.87%	0.644	0.548
Python	98.70%	0.9933	0.9867
Sharper light	80%	0.804	0.806

Table.8 Classification Accuracy of data mining tools on Pima Indian Diabetes data

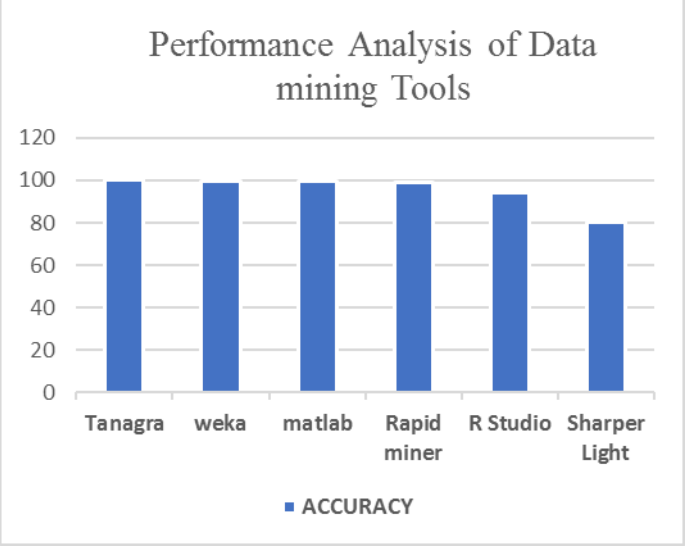

Fig.2 Performance analysis of data miming tools

 DOI : 10.18231/2454-9150.2019.0027

VII. CONCLUSION

The main objective of this survey paper is to compare the data mining tools on the basis of their classification accuracy. A careful survey of various data mining tools was carried in this study and it has been observed that different data mining tools are furnishing different results on Pima Indian Diabetes data set. Among the data mining tools used by various researchers it could be concluded that Tanagra tool showed higher accuracy of 100% with training time 0.001 seconds using Naïve Bayes classifier. Thus, Machine learning tools and techniques has a wide range of scope for extracting the hidden knowledge in prediction of diabetes. These data mining tools has the ability to predict whether the patient has diabetes or not. In future, more dataset other than PIMA can be loaded and also more parameters such as DNA gene analysis, previous history of diabetes, thirst, fatigue. frequency of urination can be included for the prediction of diabetes.

ACKNOWLEDGMENT

We would like to thank god almighty and all the authors that provide significant help for prediction of diabetes using various data mining tools.

REFERENCES

[1] Gaganjot Kaur, Amit Chhabra, "Improved J48 Classification Algorithm for the prediction of Diabetes", *International Journal of Computer Applications (0975-8887) vol.98 No.22, July 2014.*

[2] Aanurag Kumar Srivastava1, Chandan Kumar and Neha Mangla" Analysis of Diabetic Dataset and Developing Prediction Model by using Hive and R", *Indian Journal of Science and Technology, Vol 9(47), DOI: 10.17485/ijst/2016/v9i47/106496, December 2016.*

[3] Han Wu, shengqi yang, zhangqin hunag, jian he , xiaoyi Wang," Type 2 Diabetes mellitus prediction model based on data mining", *Informatics in Medicine Unlocked journal homepage: www.elsevier.com/locate/imu,2018.*

[4] Tuba PALA and Ali Yilmaz Camurcu, "Evaluation of data mining classification and clustering technique for diabetes", *Malaysian Journal of Computing Vol. 2, Issue 1, 2014.*

[5] S.Priya, R.R.Rajalaxmi Professor, "An Improved Data Mining Model to Predict the Occurrence of Type-2 Diabetes using Neural Network", *International Conference on Recent Trends in Computational Methods, Communication and Controls (ICON3C 2012) Proceedings published in International Journal of Computer Applications® (IJCA)*

[6] Rashedur M. Rahman, Farhana Afroz" Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis", *Journal of Software Engineering and Applications, 2013, 6, 85-97 http://dx.doi.org/10.4236/jsea*

[7] Uswa Ali Zia, Dr. Naeem Khan," Predicting Diabetes in Medical Datasets Using Machine Learning Techniques", *International Journal of Scientific & Engineering Research Volume 8, Issue 5, May-2017 ISSN 2229-5518.*

[8] Veena vijayan, Aswathy Ravikumar, "Study of Data mining algorithms for prediction and diagnosis of Diabetes Mellitus", *International Journal of Computer Applications (0975-8887) vol. 95-No.17, June 2014.*

[9] Arwa Al-Rofiyee, Maram Al-Nowiser, Nasebih Al-Mufad, Dr. Mohammed Abdullah AL-Hagery, "Using Prediction Methods in Data mining for Diabetes Diagnosis", *http://www.psu.edu.sa/megdam/sdma/Downloads/Posters.*

[10] V. Anuja Kumari, R.Chitra," Classification Of Diabetes Disease Using Support Vector Machine".

[11] Dr.V.Karthikeyani. Parvin Begum," Comparison a Performance of Data Mining Algorithms (CPDMA) in Prediction of Diabetes Disease", *International Journal on Computer Science and Engineering (IJCSE).*

[12] Thirumalaimuthu Thirumalaiappan Ramanathan, Dharmendra Sharma," An SVM-Fuzzy Expert System Design for Diabetes Risk Classification", *(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3), 2015, 2221-2226.*

[13] Dr. D. Ashok Kumar#1 and R. Govindasamy," Performance and Evaluation of Classification Data Mining Techniques in Diabetes", *(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2), 2015, 1312-1319.*

[14] Ankita Parashar, Kavita Burse, Kavita Rawat," A Comparative Approach for Pima Indians Diabetes Diagnosis using LDA-Support Vector Machine and Feed Forward Neural Network", *International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 11, November 2014.*

[15] David valls lanaquera," Machine Learning Classification case (NN, SVM, k-NN, Logistic Regression", *Rpubs,2018.*

[16] Mustafa S. Kadhm, Ikhlas Watan Ghindawi,Duaa Enteesha Mhawi," An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach", *International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 6 (2018) pp. 4038-4041 © Research India Publications. http://www.ripublication.com.*

[17] Basharat Naqvi, Arshad Ali, Muhammad Adnan Hashmi and Muhammad Atif "Prediction Techniques for Diagnosis of Diabetic Disease: A Comparative Study", *IJCSNS International Journal of Computer Science and Network Security, VOL.18 No.8, August 2018.*

[18] Dr.C.S.Kanimozhi Selvi, Dr.S.V.Kogilavani, Dr.S.Malliga, D.Jayaprakash ," *Classification and Prediction of Diabetics using Weka and Hive Tool", International Journal of Advance Engineering and Research Development Volume 5, Issue 04, April -2018.*

[19] Deepti Sisodia, Dilip Singh Sisodia," Prediction of Diabetes using Classification Algorithms", *International Conference on Computational Intelligence and Data Science (ICCIDS 2018),elseiver Procedia Computer Science 132 (2018) 1578–1585.*

 DOI : 10.18231/2454-9150.2019.0027

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

[21] Aakansha Rathore, Simran Chauhan, Sakshi Gujral," Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women", *International Journal of Advanced Research in Computer Science Volume 8, No. 5, May-June 2017.*

[22] Vaishali Jain, Supriya Raheja," Improving the Prediction Rate of Diabetes using Fuzzy Expert System*", I.J. Information Technology and hComputer Science, 2015, 10, 84-91 Published Online September 2015 in MECS (http://www.mecs-press.org/).*

[23] P. Radha, Dr. B. Srinivasan," Predicting Diabetes by cosequencing the various Data Mining Classification Techniques", *IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014.*

[24] Shaksham Kapoor, Krishna Priya S," Optimizing Hyper Parameters for Improved Diabetes Prediction", *International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 05 | May-2018.*

[25] http://www.idf.org/diabetesatlas.

[26] *https://WWW.Who.int/mediacentre/factsheets/fs138/en/*

[27] http://archivehealthcare.financialexpress.com

[28] *"The Importance of Early Diabetes Detection", ASPE, 2018. [Online]. Available: https://aspe.hhs.gov/report/diabetes-national-planaction/importance-early-diabetes-detection. [Accessed: 09- May- 2018].*

[29] *"Pima Indians Diabetes Database | Kaggle", Kaggle.com, 2018. [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetesdatabase. [Accessed: 09- May- 2018.*

[30] Akm Ashiquzzaman,, Abdul Kawsar Tushar, Md. Rashedul Islam, and Jong-Myon Kim*, "Reduction of Overfitting in Diabetes Prediction Using Deep Learning Neural Network", conference paper, Lecture notes in Electrical Engineering book series(LNEE, volume 449)*

 DOI : 10.18231/2454-9150.2019.0027