Data Mining Algorithms in Knowledge Management for Predicting Diabetes After Pregnancy by Using R

Dr Shruti Traymbak

Associate Professor, Department of Management,Jagannath International Management School, New Delhi India
jeanetrmbk@gmail.com

Ms Neha Issar

Assistant Professor, Department of Management, Lloyd Business School, Greater Noida, India
neha.issar@lloydcollege.in

Abstract

Data mining is considered one of the important phenomena in Knowledge Discovery in Database (KDD) process. In healthcare sector ample amount of data are available which is very essential to convert these data into useful and valuable. In this context data mining classifier algorithms play an important role to discover knowledge from big data.The objective of the present study is tocompare accuracy, sensitivity, specificity, and receiver operating characteristics of five classifier algorithms like Linear Discriminant Analysis (LDA), k-Nearest Neighbour (kNN), Support Vector Machine (SVM), Random Forest (RF), and Adaboostto predict diabetes.R is used to process data analysis. The SVM was recorded the best classifier with the accuracy of 0.7386, sensitivity0.8300 and specificity 0.5660, followed by LDA algorithm with accuracy of 0.732, sensitivity 0.8800 and specificity 0.4528 to predict diabetes.

***Keywords*:Knowledge Discovery in Database, Diabetes, Linear Discriminant Analysis, Random Forest, Support Vector Machine,Adaboost, data mining, R-programming**

1.Introduction

In the era of digital technology huge amount of structured and unstructured data result into big data which had become one of the important resources of information or knowledge for organizations. According to Komando tech[26]in current scenario people are producing 2.5 quintillion bytes of data per day. The big data industry will be valued approx. $77 billion by 2023 [26]. It is one of the greatest challenges for organizations to discover knowledge and meaningfulness information from big data.Most of the companies don't have data experts to analyse and identify meaningful pattern, and valuable knowledge from huge data. Thereare various data mining algorithms which have the potential to find meaningful patterns, correlations, and causal relationships in healthcare sector.According to [49], datamining relies on a process of using algorithms to extract valuable knowledge from large datasets and data mining is an essential element of knowledge management and in Knowledge Discovery in Database (KDD).

Knowledge Discovery in Database (KDD) is one of the important processes in knowledge management and data mining is considered to be one of the important steps in KDD. KDD and data mining differ in terms. KDD is the overall process of discovery of valuable knowledge from data. Data mining involves discovering new models from a large amount of data by applying various algorithms to fetchesvaluable knowledge[12].This combination of data mining and knowledge management assumes greater significance in knowledge management, as we move from being a data poor to the data rich economy. Knowledge is an expensive product; it must be managed properly to become an asset.According to Dalkir, K.[8] the Knowledge Management is a process which focuses on knowledge streams and the process of knowledge creation, sharing and dissemination. Although the term 'knowledge management' has been incorporated into popular usage in the late 1980's. By the early 1960s, Drucker was the first to use the term 'knowledge worker' [9]. Technologies plays an important role in knowledge management and knowledge insemination. Thus, Knowledge Management needs technologies to facilitate communication, collaboration, and content to capture, share, disseminate and apply knowledge.

Diabetes is considered one of the dreadful diseases if not diagnosed and treated on time and result into critical problems. In healthcare sector diabetes is one of the important concerns among researchers which generate a lot of information. In this context data mining algorithms is found to be important techniques to extract knowledge from ample amount of data are available in healthcare sector.The present study used Fayyad et al. [8]KDD concept to predictdiabeteswith help of data mining algorithms using R tool as shown in Figure2. They found

that data mining algorithms transformed data into meaningful and useful information for various sectors especially for healthcare sectors. For example, previousresearchers used data mining classifier algorithms to predict diabetes. [10], [24], [33], [34] [20], [13], [17][40].

The objective of the study is to compare accuracy, sensitivity, and specificity of Linear Discriminant Analysis (LDA), k-Nearest Neighbour (kNN), Support Vector Machine (SVM), Random Forest (RF), and Adaboost topredictdiabetes. For example, in the domain of healthcare three categories of data sets can be considered – health care providers, the out-patient healthcare database and the medical status data sets [28]. Another source of data relates to hospital medical records [15]. According to India Brand Equity Foundation [18] healthcare industry is supposed to reach US$193.83 billion by 2020 and estimated to grow US $372 billion by 2022. In today's digital world it not only became easier to collect big data but also to generate comprehensive healthcare reports and convert them into meaningful insights and knowledge to predict diseases.

2 Related Work

2.1 Data Mining and Knowledge Management

In healthcare sector ample amount of data are available which is very essential to convert these data into useful and valuable. Data mining is seen as the best way to manage knowledge. When data mining first came into existence in the 1990's, it was used to group, classify and predict in order to make effective decisions. But in current scenario data mining and knowledge discovery are used in interchangeable way.According to Chen et al. [4] and Fayyad et al. [12], data mining is one of the effective tools to discover knowledge from large database and it is also called Knowledge Discovery in which means to discover meaningful and useful patterns from large databases. Knowledge Discovery in Databases (KDD) term came into existence at the first KDD workshop in 1989to focus on "knowledge" as theend product of a data-driven discovery, which has been popularized by artificial intelligence and machine learning. KDD is a process of exploring valuable and useful knowledge from data whereas data mining refers to a specific step in KDD process or data mining can play an important to discover knowledge from database as shown in diagram Figure1.

Data mining is the application of specific algorithms to retrieve patterns from the database. According to Chen and Liu [5] data mining is one of the important components of knowledge management and also familiar as a strong business intelligence tool for extracting knowledge. Hwang et al. [15] data mining's algorithmsenforce in the creation of knowledge which supports decision making in various fields especially in healthcare sectors to build Knowledge Management System to support clinical practice and to improve the quality of treatment. Similarly, according to Liao, Chen and Wu [29]in case of other sectors like retail sectors data mining provides suggestions and solutions to the companies for product line and extensions of brands by discovering knowledge of customer's purchase behaviour pattern in order to fulfil the customer's behaviour pattern. In financial or banking sector data mining can help banking institutions make decisions and exchange knowledge with a view to classification of corporate bond Cheng, H., Lu, Y. and Sheu, C. [6] .Small business-like food businesses and food supply chain there were two methods and processes for generating knowledge resources: knowledge seeding refers to information and knowledge regarding the problems and knowledge- cultivating means the process to extract the knowledge from knowledge seedling. According to Li, X., Zhu, Z. and Pan, X. [30], integrated data mining and knowledge management helps in better decision making.Cantú, F.J. & Ceballos, and H.G [7] in entrepreneurial science,data mining enforced for knowledge extraction and helping managers in formulating strategies on knowledge focussed organization competition

Data mining can perform various functions like data analysis, descriptive modelling and predictive modelling. It can be broadly categorized as- classification, clustering, association rule mining and time series analysis among which classification is one of the most widely applied methodologies in different areas. The major functions of data mining are clustering, regression, classification, deviation detection, and dependency modellingto manage knowledge management process. With the help of data mining tools and techniques it helps in to detect and diagnose many diseases in health sector like diabetes Breault J.L. et al [2],heart diseases El-Sappagh S.H et al [11], K. Srinivas et al. [21], M A. Jabbar et al [32], Olatubosun Olabode et al. [35]and kidney,Kusiak A et al. [23].

Data mining is to be considered one of the extracting hidden predictive intelligences from major databases. Shearer C. [43] refers data mining as a CRISP-DM model refers to Cross Industry Standard Process for the data mining model and helps organizations to achieve better and faster results through data mining. CRISP-DM organizes process of data mining into six steps - Business Understanding which is one of the important phase of data mining. It comprises to understand business objectives, analysing the current situation, identifying data mining goals and generating plan**.** After cataloguing existing data resources, the data should be prepared for data exploration. The preparations consist of the selection, cleaning, construction, integration and formatting of the data. It is likely that these tasks will be performed more than once, and not necessarily in the prescribed order. Modelling consists in selecting modelling techniques, producing test plans and building. Conceptualizing a

model is an iterative process and try several models, and modelling techniques before finding best model and evaluation in which once the models are selected, ready to assess how data mining works which can assist to achieve business objectives. Final stage is deployment in which a successful model is applied to new data to make predictions. This might be relatively simple if done within the data-mining software (and Modeler allows the user to easily score new data), or more complex if the model is to be applied directly against an existing.

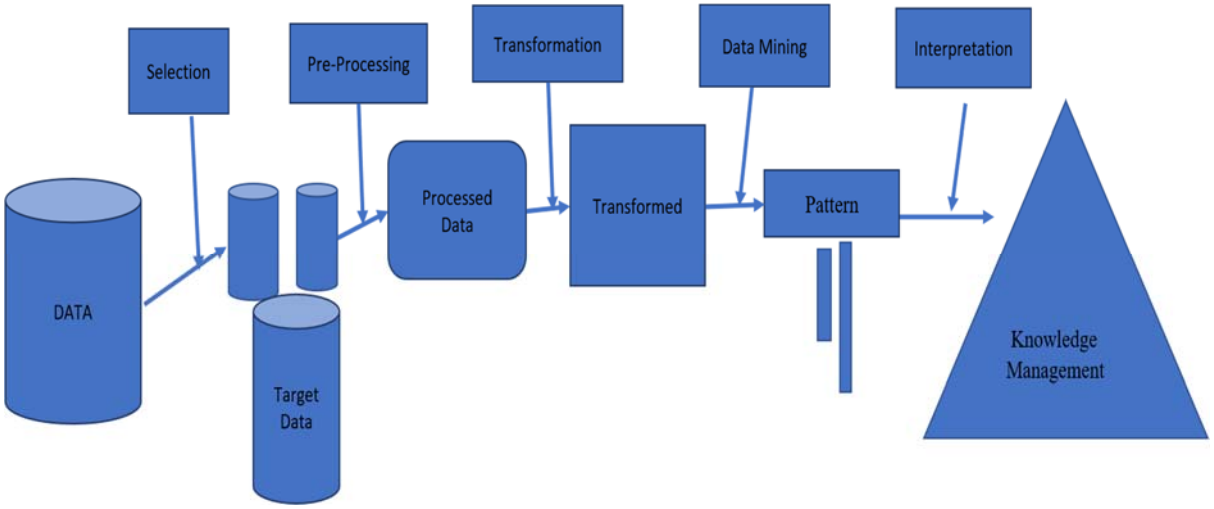

Fig1: Conceptual Framework of Knowledge Discovery in Database adapted from Fayyad et al. (1996)

2.2*Data Mining Algorithms and Knowledge Management in Healthcare Sector*

Srinivas et al. [39]have stated that healthcare sector is considered to be information rich and knowledge poor. Due to lack of powerful tools, these sectors are unable to fetch valuable information. There are multiple data mining algorithms used in healthcare sector to collect useful knowledge topredict diseases. Fang Yeet al. [14] explored risk factors of infant anaemia with the help of Chi-squared automatic interaction detection (CHAID) decision tree analysis. Similarly, Arnold [3] used CHAID decision tree algorithm for Tuberculosis Patients for relapse treatment to provide help, advice, and appropriate caution to at-risk TB patients.sSathyadevi [38]employed decision trees algorithm such as C4.5, ID3 and CART algorithms for diagnosing the disease of hepatitis. According to Kourou et al. [25], Pan et al. [36] and Anderin [1] machine learning models help in prediction and diagnosis of different diseases.

The use of data mining algorithms reveals an innovative way to predict diabetes from different sources of healthcare data sets as shown in Table 1.For example,Wie Yiu et al.[50], Priya et al. [37] and Huang et al. [16] also found that support vector machine learning model helped in to predict diabetes disease. Karegowda et al.[27] Artificial Neural Network with Genetic Algorithm, Khan et al.[22] and Thangaraju et al.[48] used Naïve bayes classifier to predict diabetes disease.

Prior researchers like D. Senthil Kumar et al. [42] found CART algorithm performed better in terms of accuracy as compared to ID3and C4.5 algorithms. K Gomathi and D. Shanmuga Priya [24] found 100% accuracy of algorithm J48 followed by Naïve Bayes 77.6% in case of diabetes prediction. Meng et al.[33] compared three algorithms like decision tree (C5.0), logistics regression and ANN model and they found decision tree (C5.0) has highest accuracy among three algorithms. Kandhasamy and Balamurali [20] compared classifiers like J48 Decision Tree, K-Nearest Neighbors, Random Forest and Support Vector Machine with noisy data sets and without noisy datasets and found that J48 Decision Tree had highest accuracy among all four algorithms. Fikirte Girma Woldemichael, et.al [13] found back propagation is highest in accuracy as compared to other algorithms. Sathya S & Rajesh. A [45] found J48 more accurate classifier algorithms, S Selvakumar et al. [40] found k-Nearest Neighbour to predict diabetes. Sisodia and Sisodia [41] used three algorithms such as Decision Tree, SVM and Naive Bayes and they found Naïve Bayes is most suitable algorithms to predict disease. Others researchers like Saravananatha and Velmurugan [47] found J48 algorithm to predict diabetes more accurately. On the basis of previous research, it is interesting to know that usage and accuracy of J48 algorithm is very high. Naveen Kishore G et al. (34) compared five classifier algorithms and they found random forest showed more accuracy to predict diabetes.

Researchers	Knowledge	Applications of	Knowledge	Algorithms Accuracy

	Sources	Data Mining	Management to predict / diagnose disease	
D.Senthil Kumar et al. [42]	Healthcare database	decision trees C4.5 algorithm, ID3 algorithm and CART algorithm	Diabetes, heart diseases and hepatitis	CART Algorithm-83.2%, ID3 Algorithm-64.8% and C4.5 Algorithm-71.4%
K Gomathi and D. Shanmuga Priya [24]	Healthcare database	Naïve Bayes and J48.	Diabetes, heart diseases and breast cancer	case of Diabetes, J48 - 100% and Naïve Bayes-77.6%
Meng et al. [33]	Healthcare Database	Neural Networks, Logistic Regression and Decision Tree C5 model.	Diabetes	decision tree (C5.0)-77.87%, logistic regression model - 76.13% and ANN model accuracy of 73.23%.
Kandhasamy and Balamurali [20]	Healthcare	J48 Decision Tree, KNN, Random Forest, and SVM	Diabetes	J48 Decision Tree-73.8%, KNN- 70.8%, SVM- 73.38% and Random Forest 71.74%
Fikirte Girma Woldemichael and S. Menaria. [13]	Healthcare	J48, naïve bayes, back propagation and support vector machine	Diabetes	back propagation-83.11%, SVM-81.69%, Naïve Bayes- 78.9%,J48-78.26%
Sathya S & Rajesh. A [45]	Healthcare	J48 classifier, the Random tree and Naïve Bayes	Diabetes	J48, 99.0521, the Random tree 95.5766 and Naïve Bayes with 93.8389
Mahmoudineja d Dezfuli S A[31]	Healthcare	Decision tree, weighted k-nearest neighbour, logistic regression, and the ensemble method	Diabetes	Ensemble method 80.60%, logistic regression 79.30%, weighted k-nearest Neighbor 77.30%, decision tree 77.00%,
S Selvakumar et al. [40]	Healthcare	the Logistic Regression, Multilayer Perceptron and KNN	Diabetes	k-NN-80%, Multilayer perceptron 71% and Logistics Regression 69%
Hashi et al. [17]	Healthcare	Decision tree, K-Nearest Neighbor and Support Vector Machine	Diabetes	Support Vector Machine, 87.01%, K-Nearest Neighbor, 86.36% and Decision tree- 81.17%,
Sisodia and Sisodia [41]	Healthcare	Decision Tree, SVM and Naive Bayes	Diabetes	Naive Bayes 76.30%, Decision Tree 73.82 % and SVM 65.10%
Saravananatha n and Velmurugan [47]	Healthcare	J48, CART, SVM,and KNN	Diabetes	J48 67.15%, CART 62.28%, SVM 65.04 %and k-NN 53.39% .
Naveen Kishore G et al. [34]	Healthcare	SVM, Decision Tree, KNN, Logistic Regression, Random Forest	Diabetes	SVM 73.43%, Decision Tree 72.91%, KNN 71.3%, Logistic Regression 72.39%, Random Forest 74.4%

Table 1: Major Role of Data Mining Algorithms of diabetes prediction

2.3 *Research Gap*

Most of the researchers have used SVM and k-Nearest Neighbour. To predict diabetes disease. A very few researchers have used Linear Discriminant Analysis, Random Forest and adaptive boosting to predict and diagnose diabetes.

2.4 *Objective of the Study*

The present study used Linear Discriminant Analysis, SVM, k-Nearest Neighbour (kNN), Random Forest and adaptive boosting (AdaBoost). Adaptive boosting (AdaBoost) is a machine learning meta-algorithm formulated by iteration our boosted classifier is a linear combination of the weak classifiers. In this study R -programming tool is used for data analysis. The objective of the present study is to compare all five classifier algorithms in terms of performance of accuracy topredict diabetes before pregnancy.Also, to classify diseases such as mild, moderate, and severe depends upon patients' various factors. This suggests doctors to reduce the pace of their treatment

3. Research Methodology

3.1 *Dataset*

The dataset contains information about diabetes samples 768 observations of 9 variables from Pima Indian Diabetes Data Set (PIDD)as shown in Table 2.The 80% of the training dataset were used for training the models while the remaining 20% were used for testing the models. The purpose of the dataset is to determine the most suitable algorithms to predict whether a patient has diabetes or not after pregnancy. The datasets consist of several medical predictors/features and one target/response variable named as Outcome. In this paper R programming tool is used for conducting the analysis.

3.2 *Process of Statistical Analysis*

In model selection, the system is trained to predict disease by data mining algorithms and represent discovered knowledge at a high level of accuracy.Data mining is one of the important steps inKDD process that applied five algorithms in a PIDD data sets as shown in Figure 5 to convert data into meaningful pattern. Five algorithms were chosen for the study -LDA, kNN, SVM, Random Forest and Adaboost. The present study determined performance of algorithms in terms of Accuracy (ACC), Specificity (Spec), Sensitivity (SEN) and Receiver Operating Characteristics (ROC).

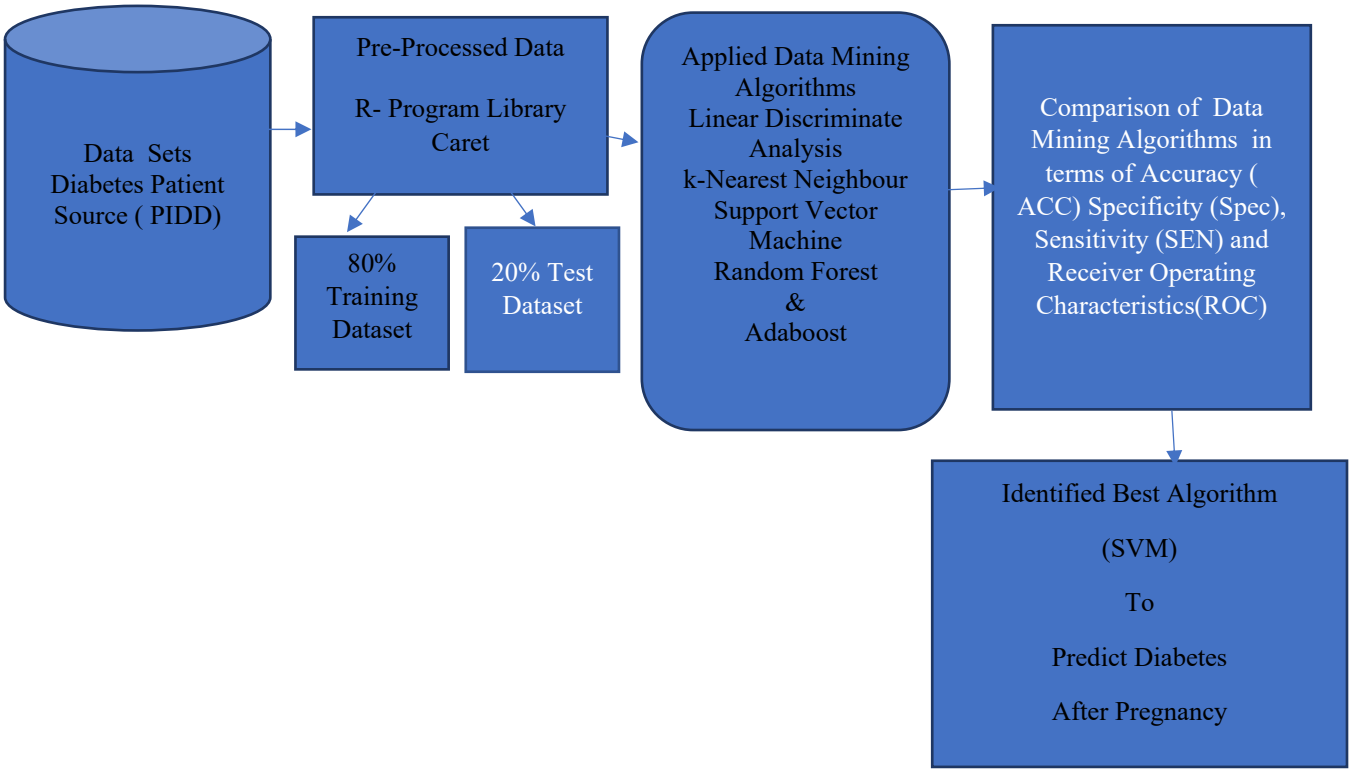

Fig 2: Conceptual Framework of Knowledge Database in Discovery (KDD) adapted from Fayyad et al. (1996)

3.3 *Data Analysis*

3.3.1R Programming Tool

R is a programming language commonly used for statistical decision making and graphical representation of diabetic dataset. The functions used are structure and summary of the dataset, different algorithms used by calling library (caret). The caret package refers to abbreviation of Classification and Regression training. Caret estimates model performance from a training set.

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Pedigree Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	**1**
1	85	66	29	0	26.6	0.351	31	**0**
8	183	64	0	0	23.3	0.672	32	**1**
1	89	66	23	94	28.1	0.167	21	**0**
0	137	40	35	168	43.1	2.288	33	**1**
5	116	74	0	0	25.6	0.201	30	**0**

Source: PIDD (Pima Indian Diabetes Data Set)

Table 2: 768 Observations and 9 Variables

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Pedigree Function	Age	Outcome
9	89	62	0	0	22.5	0.142	33	**0**
10	101	76	48	180	32.9	0.171	63	**0**
2	122	70	27	0	36.8	0.34	27	**0**
5	121	72	23	112	26.2	0.245	30	**0**
5	126	60	0	0	30.1	0.349	47	**1**
1	93	70	31	0	30.4	0.315	23	**0**

Table 3: 768 Observations and 9 Variables

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Pedigree Function	Age	Outcome
Min	0	0	0	0	0	0	0.078	21	**0-500**
1st Qu	1	99	62	0	0	27.3	0.2437	24	**1-268**
Median	3	117	72	23	30.5	32	0.3725	29	
Mean	3.845	120.9	69.11	20.54	79.8	31.99	0.4719	33	
3rd Qu.	6	140.2	80	32	127.2	36.6	0.6262	41	
Max	17	199	122	99	846	67.1	2.42	81	

Table 4: Descriptive Statistics-Summary of the Data

The outcome variable is specified as integer. It is better to represent categorical variables as factors in R.

```
diab$Outcome<-factor(diab$Outcome)
 Class (diab$Outcome)
[1] "Factor"
```

3.4 *Relation of Diabetes and Pregnancies*

The relationship between occurrence of diabetes disease and age of the subjects with the pregnancies. Diabetic outcome is given as binary, where "0" refers to norm.

3.4.1 *Scatter Plot*

The Scatter plot shows the details about pregnancies and its distribution across the age of the subjects with diabetes outcome.

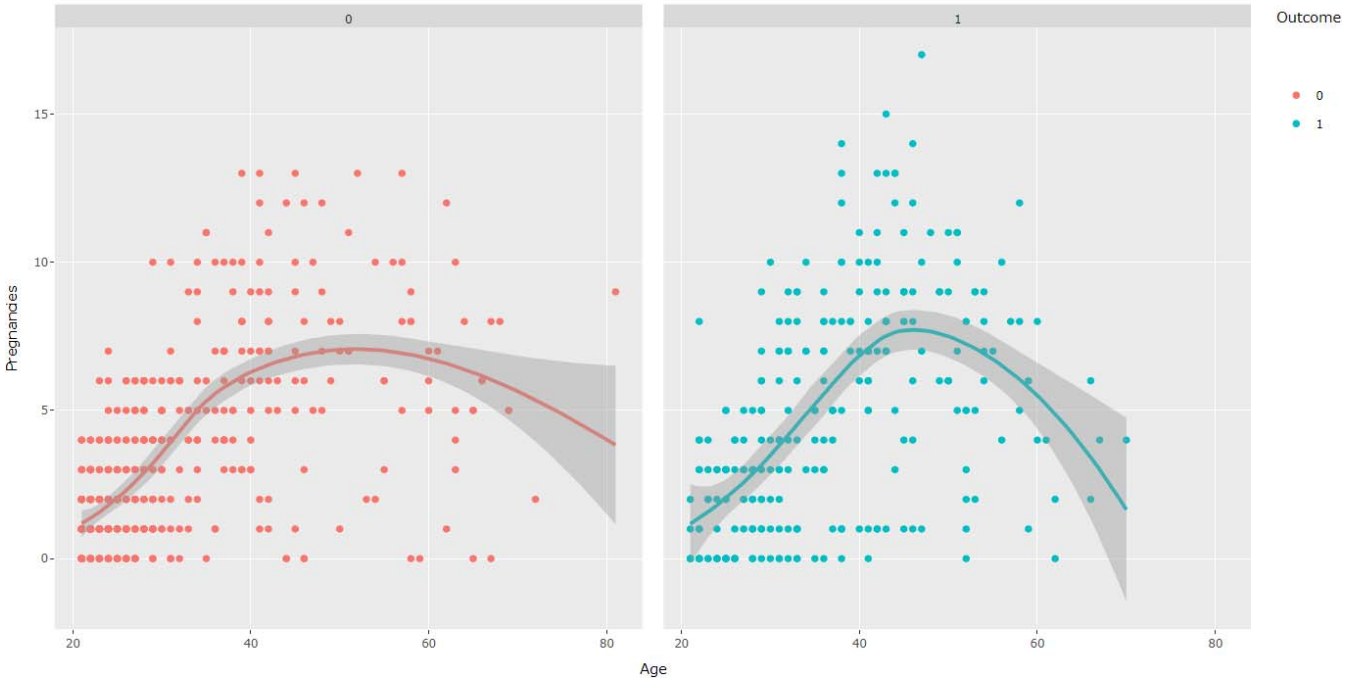

Fig 3: show the details about pregnancies and its distribution across the age of the subjects with diabetes outcome.

3.4.2 *Boxplot*

The plot shows the details about pregnancies and its distribution across the age of the subjects with diabetes outcome.

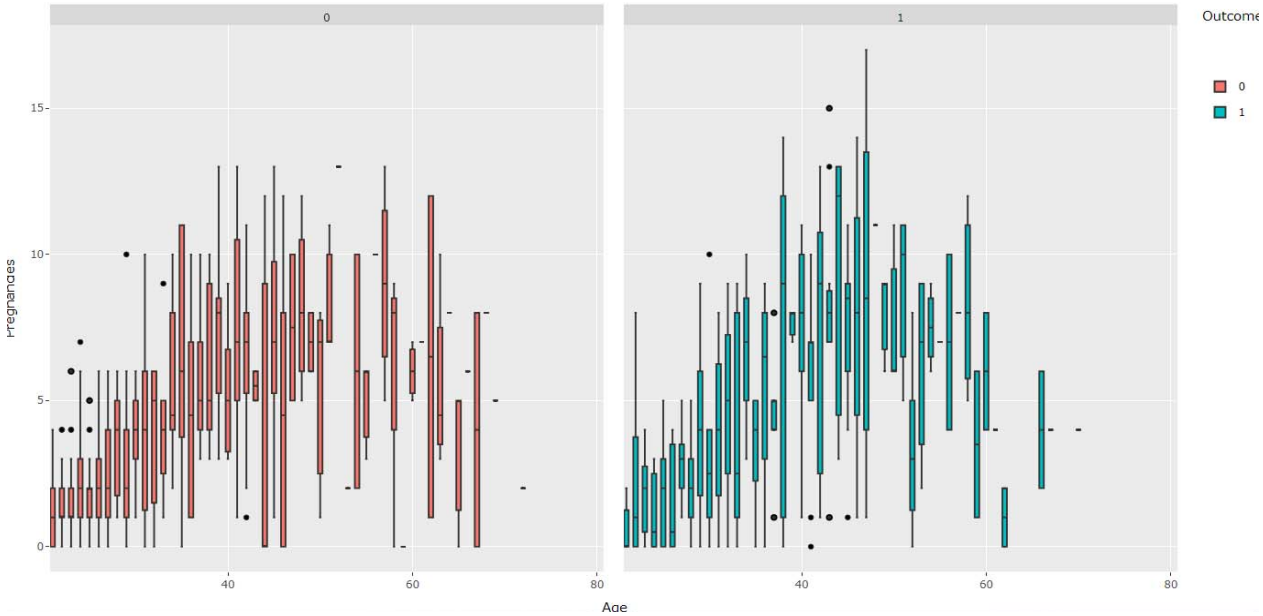

Figure 4: The Boxplot shows the details about pregnancies and its distribution.

3.4.3 *Density Plot*

Researchers can find the distribution of univariate variables in case the pregnancies of the test subjects.

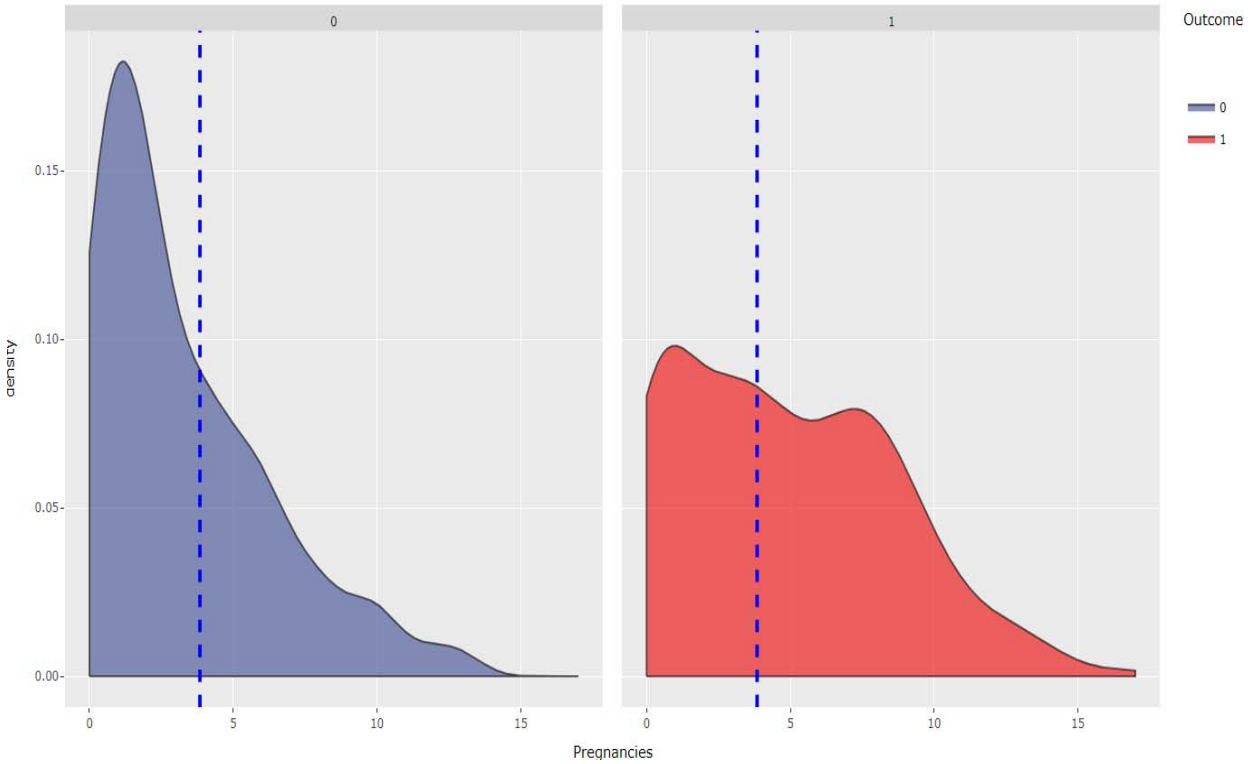

Fig 5: Distribution of univariate variables in case the pregnancies

3.4.4 *Scatter Plot*

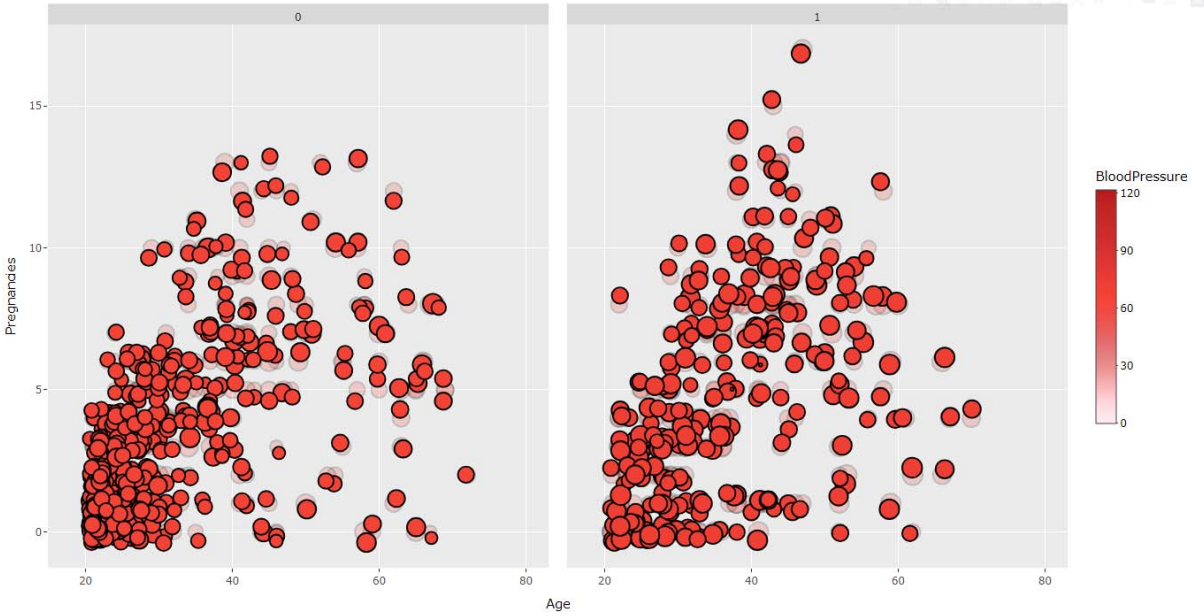

Fig 6: Relation between Glucose, Blood Pressure, Age, Pregnancy

4.Comparison of Data Mining Algorithms

The present study compared five data mining algorithms in terms of accuracy (ACC), sensitivity (Sens), specificity (Spec) and receiver operating characteristics (ROC) with the help of R-programming as shown in Figure7 and Table 5,6, 7, 8 and 9 where ROC Curve (Receiver Operating Characteristic) in which curves have significantly apply in medical decision making.

4.1 *Linear Discriminant Analysis*

Predictions set: 615 rows and 7 columns

Parameter	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
none	0.83099	0.8725	0.58605	0.02862	0.03687	0.01946

Table 5: Result of Linear Discriminant Analysis (LDA):1 row 7 columns

4.2 *k Nearest Neighbour*

Predictions set: 615 rows and 7 columns

k	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
5	0.75677	0.82	0.49767	0.03152	0.0570088	0.0709207
7	0.77151	0.83	0.50698	0.03514	0.0420193	0.1201675

Table6: Result of k Nearest Neighbour (kNN):2 rows 7 columns

4.3Support Vector Machine

Predictions set: 615 rows and 7 columns

Sigma	C	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
0.11786	0.25	0.8331395	0.86	0.5906977	0.00362	0.0205396	0.0746363
0.11786	0.5	0.8363372	0.8725	0.6046512	0.00494	0.03354102	0.0716792

Table7: Result of Support Vector Machine (SVM):2 rows 7 columns

4.4 Random Forest

Predictions set: 615 rows and 7 columns

cp		ROC	Sens	Spec	ROCSD	SensSD	SpecSD
0.06511628		0.6915698	0.8175	0.5534884	0.04369	0.08551316	0.06453
0.23255814		0.6387791	0.845	0.4325581	0.07931	0.11911129	0.24863

Table8: Result of Random Forest:2 rows 7 columns

4.5 Adaboost

Predictions set: 615 rows and 7 columns

nlter	method	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
50	Adaboost M1	0.7763372	0.795	0.5953488	0.0363201	0.02877716	0.0482243
50	Real adaboost	0.6205814	0.8425	0.5069767	0.0582777	0.02270738	0.0601962
100	Adaboost M1	0.7715116	0.7875	0.5860465	0.0041255	0.03186887	0.0601962
100	Real adaboost	0.6042442	0.8425	0.5162791	0.0677549	0.02877716	0.0724298

Table9: Result of Adaboost:4 rows 8 columns

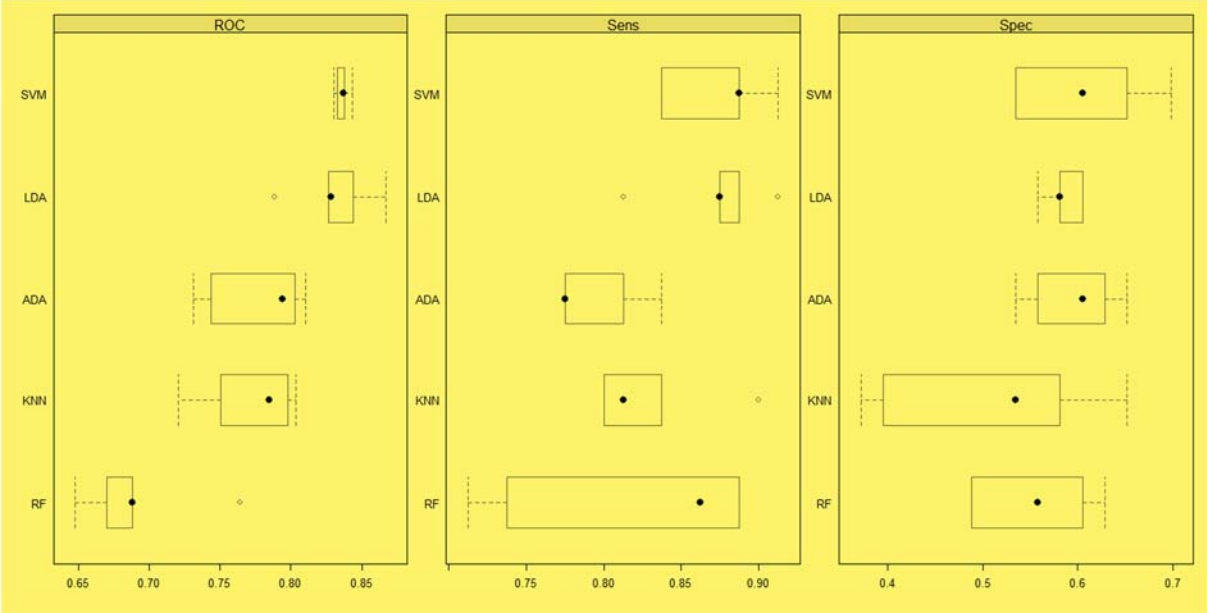

Fig 7: RF, KNN, ADA, LDA & SVM in terms of specificity (Spec), sensitivity (SEN) and receiver operating characteristics (ROC)

Algorithms	Min	1st Qu	Median	Mean	3rd Qu.	Max	NA's
LDA	0.7886628	0.826744	0.82849	0.83099	0.844186	0.8669	0
KNN	0.720494	0.750727	0.78488	0.77151	0.7976744	0.8038	0
SVM	0.8305233	0.832849	0.83721	0.83634	0.8377907	0.8433	0
RF	0.64782	0.670204	0.68794	0.69157	0.687936	0.764	0
ADA	0.7311047	0.743605	0.7939	0.77634	0.802907	0.8102	0

Table 10: Comparison of Data Mining Algorithms (ROC= Receiver Operating Characteristics)

Data Mining Algorithms	Min	1st Qu	Median	Mean	3rd Qu.	Max	NA's
LDA	0.8215	0.875	0.875	0.8725	0.8875	0.9125	0
KNN	0.8	0.8	0.8125	0.83	0.8375	0.9000	0
SVM	0.8375	0.8375	0.8875	0.8875	0.88725	0.9125	0
RF	0.7125	0.7375	0.8625	0.8175	0.8875	0.8875	0
ADA	0.775	0.775	0.775	0.795	0.8125	0.837	0

Table 11: Comparison of Data Mining Algorithms (SEN=Sensitivity)

Algorithms	Min	1st Qu	Median	Mean	3rd Qu.	Max	NA's
LDA	0.5581395	0.5813953	0.5813953	0.5860465	0.6046512	0.6046512	0
KNN	0.3720930	0.3953488	0.5348837	0.5069767	0.5813953	0.6511628	0
SVM	0.5348837	0.5348837	0.6046512	0.6046512	0.6511628	0.6976744	0
RF	0.4883721	0.4883721	0.5581395	0.5534884	0.6046512	0.6279070	0
ADA	0.5348837	0.5581395	0.6046512	0.5953488	0.6279070	0.6511628	0

Table 12: Comparison of Machine Learning Algorithms (SPEC=Specifications)

5. Confusion Matrix and Statistics

This study predicted diabetes from samples with an accuracy of approximately 73%. Table 7: Column 0 represent predictive probability that each observation is a member of non - diabetic group. Column 1 represent predictive probability that each observation is a member of diabetic group

Class0	88	29
Class1	12	14

Table 13: Column 0 represent predictive probability that each observation is a member of non - diabetic group. Column 1 represent predictive probability that each observation is a member of diabetic group

6. Mcnemear'sTest

Mcnemar's Test is applicable to explain where there are differences on a dichotomous dependent variable between two related groups. 95% confidence interval, p-value is 0.01246, accepting the null hypothesis.TheMcNemar test is a non-parametric test for paired nominal data. The kappa is being measured agreement between the outcomes. According to Landis and Koch (1977) kappa ranges 0.21-0.40 means the result of the study is fair.

Accuracy	0.732
95% CI	(0.6545, 0.8003)
No Information Rate	0.6536
P-Value [Acc> NIR]	0.02367
Kappa	0.36
Mcnemar's Test P-Value	0.01246
Sensitivity	0.88
Specificity	0.4528

PosPred Value	0.7521
Neg Pred Value	0.6667
Precision	0.7521
Recall	0.88
F1	0.8111
Prevalence	0.6536
Detection Rate	0.5752
Detection Prevalence	0.7647
Balanced Accuracy	0.6664
'Positive' Class	Class0

Table 14: LDA Algorithm Confusion Matrix

Accuracy	**0.7386**
95% CI	(0.6615, 0.8062)
No Information Rate	0.6536
P-Value [Acc> NIR]	0.01536
Kappa	0.3902
Mcnemar's Test P-Value	0.08199
Sensitivity	0.8600
Specificity	0.5094
PosPred Value	0.7679
Neg Pred Value	0.6585
Precision	0.7679
Recall	0.8600
F1	0.8113
Prevalence	0.6536
Detection Rate	0.5621
Detection Prevalence	0.7320
Balanced Accuracy	0.6847
'Positive' Class	Class0

Table 15: KNN Model Confusion Matrix

Accuracy	**0.7386**
95% CI	(0.6615, 0.8062)
No Information Rate	0.6536
P-Value [Acc> NIR]	0.01536
Kappa	0.4069
Mcnemar's Test P-Value	0.42920
Sensitivity	0.8300
Specificity	0.5660
PosPred Value	0.7830
Neg Pred Value	0.6383
Precision	0.7830
Recall	0.8300
F1	0.8058
Prevalence	0.6536
Detection Rate	0.5425
Detection Prevalence	0.6928
Balanced Accuracy	0.6980
'Positive' Class	Class0

Table16: SVM Model Confusion Matrix

Accuracy	**0.7516**
95% CI	(0.6754, 0.8179)
No Information Rate	0.6536
P-Value [Acc > NIR]	0.005891
Kappa	0.4313
Mcnemar's Test P-Value	0.256145
Sensitivity	0.8500
Specificity	0.5660

Pos Pred Value	0.7870
Neg Pred Value	0.6667
Precision	0.7870
Recall	0.8500
F1	0.8173
Prevalence	0.6536
Detection Rate	0.5556
Detection Prevalence	0.7059
Balanced Accuracy	0.7080
'Positive' Class	Class0

Table 17. Adaboost Model Confusion Matrix

Accuracy	**0.7255**
95% CI	(0.6476, 0.7945)
No Information Rate	0.6536
P-Value [Acc > NIR]	0.03543
Kappa	0.3597
Mcnemar's Test P-Value	0.08963
Sensitivity	0.8500
Specificity	0.4906
Pos Pred Value	0.7589
Neg Pred Value	0.6341
Precision	0.7589
Recall	0.8500
F1	0.8019
Prevalence	0.6536
Detection Rate	0.5556
Detection Prevalence	0.7320
Balanced Accuracy	0.6703

'Positive' Class	Class0

Table18: Random Forest Model Confusion Matrix

6. Results and Discussion

The present study used receiver operating characteristic (ROC) curves to compare sensitivity versus specificity across a range of values for the ability to predict a dichotomous outcome suggests no discrimination (i.e., potential to examine patients with and without the disease), In ROC 0.7 to 0.8 is assumed acceptable, 0.8 to 0.9 is assumed excellent, and more than 0.9 is assumed outstanding as shown in Figure 7.Classifier algorithms such as Linear Discriminant Analysis (LDA), k-Nearest Neighbour (kNN), Support Vector Machine (SVM), Random Forest (RF), and Adaboost were evaluated in terms of their discriminative classification accuracy. The Linear Discriminant Analysis acted superb in Classification Scheme -- I, III and IV, and the k-Nearest Neighbour in acted superb in Classification Scheme II. Evaluation of all five algorithms is based on parameters like accuracy, sensitivity and specificity were presented in five algorithms confusion matrix. The SVM was recorded the best classification accuracy of 0.7386 with sensitivity0.8300and specificity 0.5660 as shown in Table 16. Followed by LDA algorithm with accuracy of 0.732 with sensitivity 0.8800 and specificity 0.4528 as shown in Table 14.

7. Conclusion

The present study compared five data mining classifiers and their applications in the healthcare sector to extract knowledgeand to predict diseases based on their symptoms. There is a rapid shift in the volume of restoration information, data extraction methods are very useful in this area.SVMhas proved its accuracy in predicting diabetes after pregnancy. The datasets consist of several medical predictors/features and one target/response variable named as Outcome i.e 768 Observations of 9 Variables which is taken from 798 Pima Indian Diabetes Data Set (PIDD). The result of the study showed that SVM data mining classifier is integrated into knowledge management and suitable for early prediction of diabetes disease as compared to others classifier algorithms. It is clear that data mining classifier can help healthcare sectors in terms of predicting diseases and knowledge management on a prior basis.

Acknowledgements
No funding sources

References

[1] Anderin, C., Gustafsson, U. O., Heijbel, N., & Thorell, “A. Weight loss before bariatric surgery 388 and postoperative complications: data from the Scandinavian Obesity Registry (SOReg)”, Ann 389 Surg, vol.261, no.5, pp.909-913,2015.
[2] Breault J.L., Goodall C.R. & Fos P.J, Artificial Intelligence in Medicine, vol.26, no.1, pp.37-54,2002.
[3] Cruz, A.” Predicting the Relapse Category in Patients with Tuberculosis: A Chi-Square Automatic Interaction Detector (CHAID) Decision Tree Analysis”, Open Journal of Social Sciences, vol.6, pp.29-36,2018.
[4] Chen, M. S., Han, J., &Yu, P. S., “Data mining: an overview from a database perspective”, IEEE Transactions on Knowledge and Data Engineering, vol.8, no.6, pp.866–883, 1996.
[5] Chen, S.Y. and Liu, X, “Data mining from 1994 to 2004: an application-oriented review”, International Journal of Business Intelligence and Data Mining, vol. 1 no. 1, pp. 4-11, 2005.
[6] Cheng, H., Lu, Y. & Sheu, C, “An ontology-based business intelligence application in a financial knowledge management system”, Expert Systems with Applications, 36, pp.3614–3622, 2009.
[7] Cantú, F.J. & Ceballos, H.G, “A multiagent knowledge and information network approach for managing research asset”, Expert Systems with Applications, vol 37 no.7, pp.5272-5284,2006.
[8] Dalkir, K, “Knowledge Management Theory and Practice”, Elsevier, Burlington.2005.
[9] Drucker, P,” Knowledge worker: new target for management”, Christian Science Monitor,1964.
[10] D. Senthil Kumar, G. Sathyadevi, S. Sivanesh, JCSI International Journal of Computer Science vol.8, no.3, pp.147-153,2011.
[11] El-Sappagh, S. H. and El-Masri, S, “A distributed clinical decision support system architecture”, Journal of King Saud University-Computer and Information Sciences, vol.26, no1, pp.69–78,2014.
[12] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P, “From Data Mining to Knowledge Discovery in Databases”,AI Magazine, vol.17,no.3, pp.37-54,1996.
[13] Fikirte Girma Woldemichael, Sumitra Menaria.” Prediction of Diabetes Using Data Mining Techniques”, 2nd International Conference on Trends in Electronics and Informatics (ICOEI), pp.414 – 418,2018.
[14] Fang Ye et al., “Chi-squared Automatic Interaction Detection Tree Analysis of Risk factors for Infant Anemia in Beijing, China,” Chin Med J, vol. 129, no. 10, pp. 1193-1199, 2016.
[15] Hwang, H.G., Chang, I.C., Chen, F.J. & Wu, S.Y, “Investigation of the application of KMS for diseases classifications: A study in a Taiwanese hospital”, Expert Systems with Applications, vol.34.no1, pp.725-733, 2008.
[16] Huang, Y., P. McCullagh, N. Black and R. Harper, “Feature selection and classification model construction on type 2 diabetic patient’s data”, Artificial Intelligence in Medicine, vol.41, pp. 251-26,2007.
[17] Hashi, E. K., Zaman, M. S. U., & Hasan, M. R. (2017). An expert clinical decision support system to predict disease using classification techniques. Electrical, Computer and Communication Engineering, pp.396-400,2017.
[18] https://www.ibef.org/industry/healthcare-india.aspx.
[19] Inmon, W. H. Building the Data Warehouse. New York, Wiley,2012.

[20] J.Kandhasamy and Balamurali, "Performance Analysis of Classifier Models to Predict Diabetes Mellitus", Procedia Computer Science vol. 47, pp.45-51,2015.
[21] K. Srinivas, B. Kavitha Rani and Dr. A. Govardhan, "Application of Data Mining Techniques in Healthcare and Prediction of Heart Attacks," International Journal on Computer Science and Engineering, vol.2, no.2, pp.250 – 255, 2011.
[22] Khan, N., Gaurav, D., &Kandl, T, "Performance evaluation of Levenberg Marquardt technique in error reduction for diabetes condition classification", Procedia Computer Science, vol.18, pp.2629-2637, 2013.
[23] Kusiak A, Dixon B, Shah S, "Predicting survival time for kidney dialysis patients: a data mining approach." Comput Biol Med, vol.35, no.4, pp.311–27,2005.
[24] K. Gomathi, Dr. D. Shanmuga Priyaa, "Multi Disease Prediction using Data Mining Techniques",2017.
[25] Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I, "Machine learning applications in cancer prognosis and prediction", Comput. Struct. Biotechnol. J, vol.13, pp. 8–17, 2015.
[26] https://kommandotech.com/statistics/big-data-statistics/
[27] Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A, "Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indian's diabetes", International Journal on Soft Computing, vol. 2, no.2, pp.15-23, 2011.
[28] Lavrac, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M. &Kobler, A, "Data mining and visualization for decision support and modeling of public health-care resources", Journal of Biomedical Informatics, vol.40, pp.438-447,2007.
[29] Liao, S.H., Chen, C.M., Wu, C.H, "Mining customer knowledge for product line and brand extension in Retailing", Expert Systems with Applications, vol.34, no.3, pp.1763-1776,2008.
[30] Li, X., Zhu, Z. & Pan, X, "Knowledge cultivating for intelligent decision making in Small & middle Businesses", Procedia Computer Science, vol.1, no.1, pp.2479-2488,2010.
[31] Mahmoudinejad Dezfuli S A, Mahmoudinejad Dezfuli S R, Mahmoudinejad Dezfuli S V, Kiani Y, "Early Diagnosis of Diabetes Mellitus Using Data Mining and Classification Techniques", Jundishapur J Chronic Dis Care, vol.8, no.3, e94173, 2019.
[32] M A. Jabbar, Priti Chandra and B. L. Deekshatulu, "Cluster based association rule mining for heart attack prediction", Journal of Theoretical and Applied Information Technology, vol.32, no.2, pp.197 –201,2011.
[33] Meng X-H, Huang Y-X, Rao D-P, Zhang Q, Liu Q, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors", Kaohsiung J Med Sci, vol.29, no.2, pp.93–99, 2013.
[34] Naveen Kishore G, V.Rajesh, A.Vamsi Akki Reddy, K.Sumedh, T.Rajesh Sai Reddy, "Decision Tree. Prediction Of Diabetes Using Machine Learning Classification Algorithms," International Journal of Scientiic & Technology Research, vol.19, no.1,2020.
[35] Olatubosun Olabode and Bola Titilayo Olabode, "Cerebrovascular Accident Attack Classification Using Multilayer Feed Forward Artificial Neural Network with Back Propagation Error", Journal of Computer Science, vol.8, no., pp.18 – 25, 2012.
[36] [36] Pan, L.; Liu, G.; Lin, F.; Zhong, S.; Xia, H.; Sun, X.; Liang, "Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia", Sci. Rep.vol. 7, pp.7402, 2017.
[37] R.priya and P Aruna, " SVM and Neural Network based Diagnosis of Diabetic Retinopathy", International Journal of Computer Applications, vol. 41, no.1, pp.6-12, 2012.
[38] Sathyadevi, G, "Application of CART algorithm in hepatitis disease diagnosis, Recent Trends in Information Technology (ICRTIT)", International Conference on, IEEE, pp. 1283–1287, 2011.
[39] Srinivas, K., Kavihta Rani, B., & Govrdhan, A., "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Journal on Computer Science and Engineering, pp.250-255, 2010.
[40] Selvakumar, S., Kannan, K.S. and GothaiNachiyar, S., "Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques", International Journal of Statistics and Systems, vol.12, no.2, pp.183-188, 2017.
[41] Sisodia, D., Shrivastava, S.K., Jain, R.C, "ISVM for face recognition. Proceedings", – International Conference on Computational Intelligence and Communication Networks, CICN, pp.554–559, 2010.
[42] Senthilkumar D and Paulraj S, "Prediction of Low-Birth-Weight Infants and Its Risk Factors Using Data Mining Techniques" in Proceedings of the International Conference on Industrial.
[43] Shearer, C, "The CRISP-DM Model: The New Blueprint for Data Mining", Journal of Data Warehousing, vol.5, pp.13-22, 2010.
[44] Sarwar A. and Sharma V, "Comparative analysis of machine learning techniques in prognosis of type II diabetes", Al & Society, vol. 29, no.1, pp.123-129, 2014.
[45] S. Sathya and A. Rajesh, "An Effective Prediction of Diabetics Using ID3 Classification Algorithm. Middle East Journal of Scientific Research" pp. 207-211, 2016
[46] J.Kandhasamy and Balamurali, "Performance Analysis of Classifier Models to Predict Diabetes Mellitus", Procedia Computer Science vol.47, pp.45-51, 2015.
[47] Saravananathan, K., & Velmurugan, T, "Analyzing Diabetic Data using Classification Algorithms in Data Mining", Indian Journal of Science and Technology,vol. 9, no.43, 2016.
[48] Thangaraju, P., Deepa, B., & Karthikeyan, T. , "Comparison of Data mining Techniques for Forecasting Diabetes Mellitus", International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no.8, 2014.
[49] Wang, H. & Wang, S. "A knowledge management approach to data mining process for business intelligence", Industrial Management & Data Systems, vol.108, no.5,pp. 622-634, 2008.
[50] Yu, W., Liu, T., Valdez, R. et al. "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes" BMC Med Inform Decis Mak vol.10, no.16, 2010.

Authors Profile

Dr. Shruti Traymbak is an Associate Professor of HR and Marketing at Jagannath International Management School, New Delhi. She has more than 6 years of Industrial experience and 9 years of teaching and research experience. She has graduated from Miranda House, Delhi University and has been awarded Ph.D. in Human Resource Management in 2019 from the reputed Institute Birla Institute of Technology, Mesra (Ranchi). Her areas of interest are, HRM, Organizational Behaviour, Payroll Management, Marketing and HR Analytics, Training and Development and Industrial Relations and Labour Law. She has many publications of national and International fame to her credit like ABDC, Scopus, Web of Science, UGC care etc.

.

Ms. Neha Issar is a part of the Faculty at Lloyd Business School as Assistant Professor. She is IBM Spark & Scala Certified and 10 years of Corporate and Academic experience. Database & Query Language, Data Science using R, Management Information System and Enterprise Resource Planning. She is a commerce graduate from Dr. B.R. Ambedkar University and holds a post graduate diploma in Finance and Marketing. She has more than nine years of experience in academics and industry. She has worked as a Data Researcher at S&P Global Market Intelligence Pvt. Ltd. She has also worked on Projects at Aditya Birla Group, Times Internet in et.com. She was a Senior Manager Investor Relations at Opera Gratia Pvt Ltd. She is an IBM Spark and Scala Certified Professional.