ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

Diabetes Prediction for the Patient Based on JRIP Technique

Er. Veena Rani, Er. Sandeep Singh Dhaliwal

P.G Student, Department of Computer Engineering, Bhai Maha Singh College of Engineering and Technology, Shri Muktsar Sahib, Punjab, India

Assistant Professor, Department of Computer Engineering, Bhai Maha Singh College of Engineering and Technology, Shri Muktsar Sahib, Punjab, India

ABSTRACT: Information Mining is the present stream where organizations are giving more weight on. Different types of applications are there in today world which are generating large amount of data. Different types of classifiers like J-Rip, Decision table, naives bayes and logistic regression has been performed. Based on the classifiers the accuracy for the different classification process has been performed. J-Rip is the best classification technique with the accuracy of 96.617%. in comparison to the J-Rip the accuracy for Hybrid prediction model of improved k-means and logistic regression, J48(pruned) and naïve bayes accuracy is 95.41%, 89.3%, 74.92% respectively. Total instances for the dataset items are 768. F-Measure value for the J-Rip is 0.985 and the F-Measure value for the Hybrid prediction model of improved k-means and logistic regression 0.965. J-Rip is better classification technique as compared to the existing models for the correctly classifying the instances based on different attributes.

KEYWORDS: JRip, logistic regression, Decision Table, Diabetes mellitus

I. INTRODUCTION

Various search engines and social networking sites collect huge amounts of data by using different data mining techniques; they try to find the hidden patterns within the data. Data mining is widely used in diverse areas such as banking, e-commerce, health and medicine, genetics, education, stock exchange and various other fields' .The data is available in structured, semi-structured and unstructured format which is initially processed and then analyzed using different techniques. These techniques are broadly classified into two main categories, namely, predictive and descriptive. The predictive data mining techniques focus on understanding the future, making predictions using the available datasets whereas the descriptive techniques summarize and analyze the past data and properties to make it useful in predicting new ones.

Various fields like marketing, education, banking sector, bio-informatics, and healthcare make use of these techniques. But nowadays, a lot of focus is given to the healthcare sector to improve the process of decision making and provide better facilities to the patients. The data for analyzing the health data can be collected from various sources like paper records, x-rays, etc and can be stored in electronic health records. The accumulated data can then be processed and analyzed. The predictive analysis considers the various symptoms, their effects on health and can help in early detection and prevention of these diseases. Different data mining techniques have been implemented to obtain the results regarding breast cancer, heart disease and diabetes to analyze the current status and make future predictions regarding the occurrence of disease, early detection and preventable patient deaths.

These days, there has been an extensive increase in the number of diabetic patients. According to WHO report, about 69 million people in India were diagnosed with diabetes in 2015. It is estimated that this count will rise to 98 million

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

by the end of 2030. Diabetes is caused due to high increase in blood sugar levels. The most common types of diabetes are Type 1, Type 2 and gestational diabetes. Type 1 diabetes is caused when the body is unable to generate the required amount of insulin. In type 2 diabetes, the body is unable to make proper use of produced insulin. Gestational diabetes develops in pregnant women and in most cases goes away after the baby is born but they are highly prone to type 2 diabetes later on in life. In order to identify the high risk patients, these advanced technologies can be used. They play a pivotal role in addressing these challenges and Data mining can prove to be an appropriate field.

The success of data mining is obvious in highly visible business fields such as marketing, retail, and e-commerce. This has caught the attention of other fields such as the healthcare sector where researchers are using data mining approaches such as expert system, statistics, data visualization, multidimensional databases in prognosis and diagnosis of diseases. In healthcare, other applications of data mining are in the judgment of treatment effectiveness, healthcare management, CRM, and detection of fraud and abuse. A survey conducted in 2011 by Pew's Research Center revealed that up to 80% of Internet users have gone online to find information regarding to healthcare. Hype in medical costs and greater internet connectivity has made online healthcare attractive. An online diagnosis would greatly benefit in not only in the reduction of the medical expenses, but in providing a more accurate detection and being available all the time and accessible from any place with Internet connectivity.

A. Feature Extraction Approaches

Data preprocessing is an important step in data mining as real world data are generally noisy (containing errors and outliers), inconsistent (containing discrepancies such as gender = male and pregnant = yes) and incomplete(missing attributes). The major tasks in preprocessing are data cleaning, data integration, data transformation and data reduction.

Data Cleaning consists of filling in missing values by ignoring the tuple or by filling in the missing value with the attribute mean or by projecting the missing value using a machine learning algorithm, identifying outlier and smoothing noisy data which can be done by binning, clustering and regression, correct inconsistent data by using domain knowledge or expert decisions.

Data Integration involves associating data from multiple sources into single data. This includes resolving inconsistencies among different data sets.

Data Transformation consists of normalization where the attributes are spanning to fall within a specified range, aggregation (moving up in the concept hierarchy on numeric attributes), generalization (moving up in the concept hierarchy on nominal attributes) and attribute construction (replacing or adding new attributes inferred by existing attributes). Information is changed into a frame that is suitable for data mining.

Data Reduction involves reducing the number of attributes, lessen the number of attribute values and minimizes the number of tuples. Discretization is a type of information decrease.

B. Measure of algorithm output

A confusion matrix carries data about the actual and predicted classifications done by a classification system. Using the information in a confusion matrix, performance of the systems can be evaluated.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

TABLE 1
Confusion Matrix

Actual \ Predicted	**Positive**	**Negative**
Positive	TP	FN
Negative	FP	TN

The meaning of the elements in the confusion matrix is as follows

- TP is the number of right predictions that an instance is positive.
- FP is the number of false predictions that an instance is positive.
- FN is the number of false predictions that an instance is negative.
- TN is the number of right predictions that an instance is positive.

Using the above given parameters, various evaluation metrics are used as:
Accuracy is the percentage of correctly predicted instances or the proportion of the total number of predictions that are correct. Accuracy = (TP+TN)/ (TP+TN+FP+FN)

Sensitivity or Recall or true positive rate is the proportion of positive instances that are correctly predicted as positive. Recall= TP/ (TP+FN)

Specificity or the true negative rate is the proportion of negative instances that are correctly predicted as negative. Specificity= TN/ (TN+FP)

Precision is the proportion of predicted positive instances that are correct. P = TP/ (TP+FP)

F-measure is the harmonic mean of precision and recall. It is given by F-measure = 2* precision*recall/ (precision+ recall).

II. LITERATURE SURVEY

Han Wu et al. [4] have proposed a model that implemented k-means algorithm and logistic regression to predict type-2 diabetes. The main aim was to improve the accuracy and implement the model using various datasets. This model produced satisfying results and a lesser amount of time consumed by it. Accuracy of their proposed model was 3.04% more than the existing ones. The performance of the model was evaluated using two other datasets as well.

Isaac et al. [5] have compared the accuracy obtained by implementing k-means algorithm and decision tree for diagnosis of breast cancer. They have trained the model using 15 attributes. Results showed that both the techniques resulted in high accuracy but statistical results show that k-means algorithm has higher performance than decision tree. On the other hand, Chen et al [6] have proposed hybrid model for prediction of type-2 diabetes. This model

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

implemented k-means for data reduction and decision tree for classification which resulted in better accuracy and promising results.

Kanika et al. [7] have implemented an optimized hierarchical clustering method which aimed at reducing the computation cost. They have used Genetic algorithm and SVM for selection and classification respectively. They have used PIMA Indians Diabetes dataset was used and the accuracy was improved by 1.351%. Another hybrid approach used k-means clustering using feature similarity and SVM for classification was implemented using PIMA Indian Diabetes dataset [17]. The results showed that this combination has high prediction accuracy.

Nahato et al. [9] have proposed combination of fuzzy sets and extreme learning machine for classification of three different datasets regarding heart disease and diabetes. In this hybrid approach features of the dataset were divided into fuzzy sets and then extreme learning machine was used to classify them. Three different datasets were used to implement the proposed algorithm, namely, Cleveland heart disease dataset (CHD), Statlog heart disease dataset (SHD) and Pima Indian diabetes dataset (PID). The algorithm was run by varying the number of hidden layer neurons and the ones that produced the most efficient results were selected. This model performed better in terms of accuracy and training time. The accuracy of the classifier for CHD, SHD and diabetes data came out to be 73.77%, 94.44% and 92.54%, respectively.

III. ALGORITHM

This approach is an extension of k-means clustering where the weights are assigned to individual features based on their importance. An initial set of centroids is fed to k-means and observations are clustered to the nearest centroid based on the distance obtained. This is the underlying bunching. Further, based on these clusters, new centroids are calculated. Then, based on the current clustering, weights are calculated for each variable within the cluster. The weights define the importance of each variable. Using these weights and the centroids obtained, observations are again clustered. This typically reduces the distance for important variables.

The objective function for weighted k-means is given using.

$$\sum_{i=1}^{k} \sum_{x \in C_i} w(x) ||x - m_i||^2 \qquad (13)$$

where k is the number of clusters, C_i is the i^{th} cluster and m_i is the mean of C_i.

Further, new centroids are calculated and observations are again clustered using the weighted distance measure. This continues until the final clusters are obtained. Then, the correctly clustered observations are used to train the classifier.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

IV. FLOWCHART

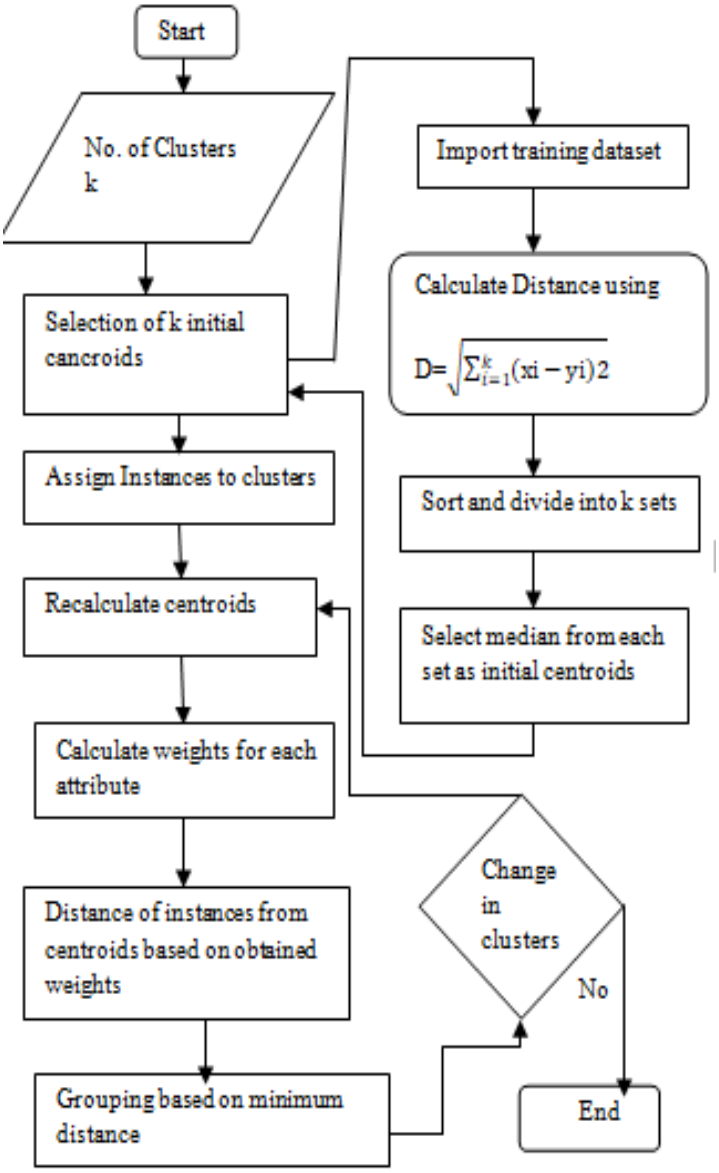

Fig.1 Flowchart of proposed model

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

V. RESULTS AND DISCUSSIONS

A. Collected the data

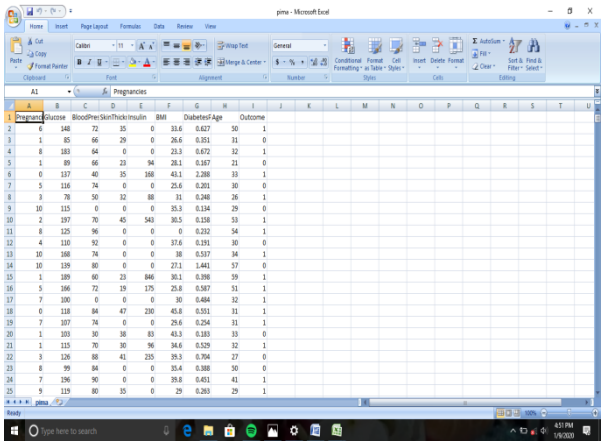

Fig. 2 PIMA Dataset

B. Experimental results

TABLE 2
Performance Analysis of Existing and Proposed results

Parameters	Existing Results	Proposed Results
Correctly classified instances	95.41	96.6
Incorrectly classified instances	4.5	3.39
Kappa Statistics	0.89	0.93
Mean absolute error	0.094	0.0932
Root mean square error	0.2093	0.1087
Relative absolute error	20.86	20
Root relative squared error	43.93	37.4721
True Positive	377	550
True Negative	185	20
False Positive	20	11
False Negative	7	8

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

Table 2 displays the results of existing technique and proposed technique. This table shows that proposed technique have result improvement in true positive, true negative etc. The relative absolute error and root relative squared errors are decreased.

C. Accuracy Comparison

TABLE 3
Comparison of accuracy

Sr. No.	Method	Accuracy
1.	Our proposed model	96.60%
2.	Base Model	95.42%
3.	HPM	92.38%
4.	J48(pruned)	89.3%
5.	Logistic	78.2%
6.	Naïve Bayes	74.9%
7.	KNN	67.6%

Table 3 displays the comparison of proposed model's accuracy with other experiments. This table shows that our model's accuracy of prediction has achieved a certain level of improvement. We compared our result with other researchers' experiments using the same dataset. The accuracy of our proposed model is 96.60%.

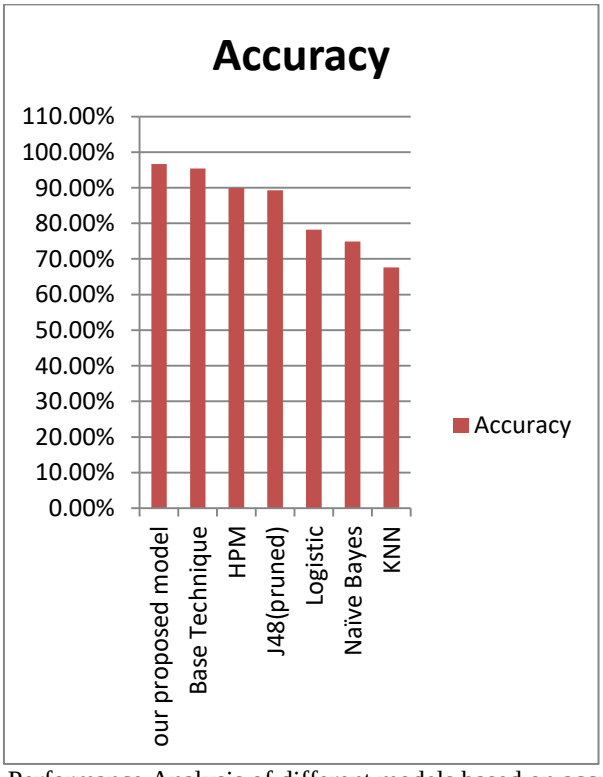

Fig.3 Performance Analysis of different models based on accuracy

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

TABLE 4
Performance Analysis using various parameters

Parameter	Existing Model	Proposed Model
Predicted Accuracy	0.954	0.966
Precision	0.954	0.98
Recall	0.954	0.985
Specificity	0.92	0.64
MCC	0.899	0.661
Kappa Statistics	0.897	0.924
F-Measure	0.965	0.985

Table 4 shows the values of base and proposed models for various parameters. Our proposed model shows improvement in accuracy, precision, recall, kappa statistics and F-Measure etc which proves that the proposed model is reliable and effective.

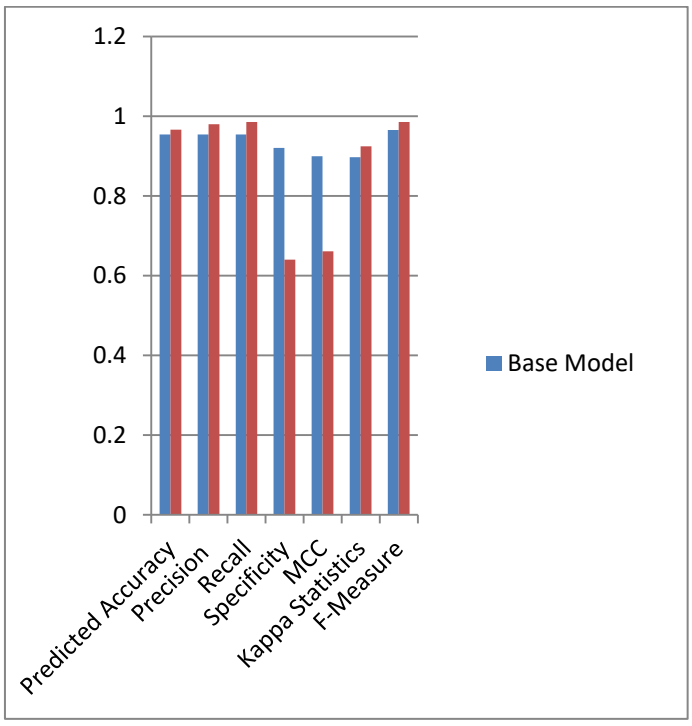

Fig. 4 Proposed model performances

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

5.4 Analysis and Interpretation of the Result

The data collected as patient data. This dataset is also known as PIMA dataset. Data has nine attributes, which are responsible for the diabetes to the patients. These attributes are Number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure, Triceps skin fold thickness, 2-Hour serum insulin, Body mass index, Diabetes pedigree function, Age (years). JRIP is the best available algorithm for the classification with higher level of accuracy and lowest level of error rate.

VI. CONCLUSION

From the above analysis for the dataset belongs to the PIMA dataset it is clear that the pre prediction for the patient based on different parameters will remain best way to make precaution for the patient. PIMA dataset includes 9 different parameters. These parameters are responsible for the diabetes to the patient. Our proposed model used both cluster and classification method. our proposed model consisted k-means for clustering and JRip for classification. Our proposed model implies that JRip the best algorithm withhigher level of accuracy in determining the true positive and true negative, false positive and false negatives. In medical applications, the sensitivity or true positive rates are extremely important as diagnosing a diseased person as non-diseased could be a matter of life and death. A high false positive is tolerable but a high specificity or true positive is desired.

VII. FUTURE WORK

The current research focused on hybridization of k-means and JRip for diabetes prediction based on eight different parameters. In the future, focus can be laid on using several other combinations of clustering and classification approaches for disease prediction on a larger dataset having more number of attributes. Different feature weighing approaches can also be used to enhance the process of clustering.

REFERENCES

[1] Athmaja, S., Hanumanthappa, M, and Kavitha, V (2017) "A survey of machine learning algorithms for big data analytics," in Innovations in Information, Embedded and Communication Systems (ICIIECS), *2017* International Conference on, pp. 1-4, IEEE.
[2] Kunjir, A., Sawant, H. and Shaikh, N. F. (2017) "Data mining and visualization for prediction of multiple diseases in healthcare," in Big Data Analytics and Computational Intelligence (ICBDAC), 2017 International Conference on, pp. 329-334, IEEE,
[3] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I. (2017) "Machine learning and data mining methods in diabetes research," Computational and structural biotechnology journal, vol. 15, pp. 104-116.
[4] Wu, H., Yang, S., Huang, Z., He, J., and Wang X. (2018) "Type 2 diabetes mellitus prediction model based on data mining," Informatics in Medicine Unlocked, vol. 10, pp. 100-107, 2018.
[5] Isaac , L. D. and Sureshkumar, C. (2018) "Diagnosis prognosis and prevention of breast cancer based on present scenario of human life," in 2018 International Conference on Communication information and Computing Technology (ICCICT), pp. 1-7, IEEE,
[6] Chen, W., Chen, S., Zhang, H., and Wu, T. (2017) "A hybrid prediction model for type diabetes using k-means and decision tree," in Software Engineering and Service Science (ICSESS), 2017 8th IEEE International Conference on, pp. 386-390, IEEE.
[7] Bhatia, K. and Syal, R. (2017) "Predictive analysis using hybrid clustering in diabetes diagnosis," in Control, Automation & Power Engineering (RDCAPE), Recent Developments in, pp. 447-452, IEEE.
[8] Osman, A.H. and Aljahdali, H.M. (2017) "Diabetes disease diagnosis method based on feature extraction using K-SVM," Int J Adv Computer Sci Appl, vol. 8, no. 1.
[9] Nahato, K. B., Nehemiah,K. H., and Kannan A. (2016) "Hybrid approach using fuzzy sets and extreme learning machine for classifying clinical datasets," Informatics in Medicine Unlocked, vol. 2, pp. 1-11.
[10] Fatemidokht, H. and Rafsanjani ,M. K. (2018) "Development of a hybrid neuro-fuzzy system as a diagnostic tool for type 2 diabetes mellitus," in Fuzzy and Intelligent Systems (CFIS), 2018 6th Iranian Joint Congress on, pp. 54-56, IEEE.
[11] Melin, P., Ramirez,E., and Arechiga, G. Prado (2018) "A new variant of fuzzy k-nearest neighbor using interval type-2 fuzzy logic," IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1-7, IEEE.
[12] Vaishali, R., Sasikala, R., Ramasubbareddy, S., Remya S., and Nalluri S. (2017) "Genetic algorithm based feature selection and moe fuzzy classification algorithm on pima indians diabetes dataset," in Computing Networking and Informatics (ICCNI), International Conference on, pp. 1-5, IEEE.
[13] Shinde, Sachin, and Bharat Tidke. (2014) "Improved K-means Algorithm for searching Research papers." International Journal of Computer Science & Communication Networks 4.6 (2014): 197-202.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

[14]Khalil, R.M. and Al-Jumaily, A., (2017) "Machine learning based prediction of depression among type 2 diabetic patients", In 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE) (pp. 1-5). IEEE.
[15] Lekha, S. and Suchetha, M., (2017),"Real-time non-invasive detection and classification of diabetes using modified convolution neural network", IEEE journal of biomedical and health informatics, 22(5), pp.1630-1636.
[16] Liu, L., (2018) , "Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning", In 2018 International Conference on Robots & Intelligent System (ICRIS) (pp. 157-160). IEEE.
[17] Kaggle.com. (2016) "Pima Indians Diabetes Database", [online] Available at: https://www.kaggle.com/uciml/pima-indians-diabetes-database [Accessed 12 Nov. 2018].