Analysis and Prediction Of Pima Indian Diabetes Dataset Using SDKNN Classifier Technique

To cite this article: Radhanath Patra and Bonomali khuntia 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1070** 012059

Analysis and Prediction Of Pima Indian Diabetes Dataset Using SDKNN Classifier Technique

Radhanath Patra
Berhampur University, Berhampur, Ganjam, Odisha-760007
E-mail: 1radhanath.patra@gmail.com

Bonomali khuntia
Berhampur University, Berhampur, Ganjam, Odisha-760007
E-mail: bonomalikhuntia@gmail.com

Abstract.
The newly proposed weighted k nearest neighbour is known as standard deviation K nearest neighbour(SDKNN) classifier technique. It is based on the principle of standard deviation. Standard deviation measures spreading of attribute about mean. Spreading of attribute plays a significant role to improve the classification accuracy of a dataset. Most of our distance calculation method between two points is determined by using euclidean distance process for finding nearest neighbour. Our proposed technique is based on a new distance calculation formula to find nearest neighbour in KNN. We apply here standard deviations of attributes as power for calculating distance between train dataset and test dataset. Distance calculation between two points in k nearest neighbour classifier is modified according to the standard deviation of attribute. In this paper, standard deviation of attributes are used. In first attempt, we have used standard deviation of attributes as power for calculating K Nearest Neighbour to improve classification accuracy and in second attempt, based on mean of standard deviation attributes, distance in K Nearest Neighbour is processed to further improve the classification accuracy. Our concept is implemented on Pima Indian Diabetes Dataset (PIDD). The analysis on Pima Indian Diabetes Dataset (PIDD) is carried out by splitting dataset in to 90% training data and 10% testing data. We have found that, in our proposed technique, average classification accuracy gives result 83.2%, a great improvement as compared to other conventional technique.

Keywords: SDKNN(Standard Deviation K Nearest Neighbour), KNN (K Nearest Neighbour), PIDD (Pima Indian Diabetes Dataset)

1. Introduction

Diabetes or Diabetes Mellitus in medical field is called as silent killer. Now a day's diabetes (DM) is widely spreading in all over the world and the effect of diabetes is showing a major decline of health condition in human beings directly or indirectly. Due to diabetes various human organs are severely affected and ill functioned which may cause heart stroke, blindness, brain dead, kidney failure e.t.c. More than 422 million people are suffering from diabetes as per world health organization index (WHO)[1]. Machine learning now days plays a crucial role for detection and prediction of medical diseases at an early stage of safe human life. Machine

Learning makes diagnosis process easier and deterministic. So in need of early detection and prediction of diabetes. Therefore machine learning methods are real exemplary, in this aspect for accurate predictions of diabetes data which may help the patient in more consolidated way[2]. Diabetes occurs due to the various diet factors is now slowly affecting from youth to older people across the world. Like other diseases, diabetes is a chronic disease that is slowly affecting human beings and will be probably increased in coming future. Increase of sugar level in blood indicates diabetes. Diabetes can't be cured completely but it can be controlled or prevented. It not only affects immune system of patient but also cause of other diseases like heart attack, blindness, kidney diseases etc. Diabetes is such a severe disease, a patient requires to visit the diagnostic center most often for consulting a doctor. Diabetes Mellitus (DM) is due to abnormal insulin secretion. Diabetes Mellitus (DM) is generally classified into two categories. They are Type 1 diabetes (T1D) and Type 2 diabetes (T2D)[3]. Most of the peoples is affected by T2D as compared to T1D. The main causes of T2D includes heredity, eating habit and lack of physical exercise, whereas T1D is thought to be due to autoimmune logical destruction of the Langerhans islets hosting pancreatic- cells[4]. Machine learning plays a crucial role to analyze the medical data. Machine learning application has already brought a lot of revolution in health care sector for diagnosis and prognosis of diseases. ML helps in proper prediction of diseases such that effective treatment of diseases can be possible. Various clustering algorithm, machine learning classifier as well as evolutionary algorithm has already proposed in medical field for the analysis of Pima Indian Diabetic Dataset(PIDD)[5]. ML thus helps to classify diabetes data in to diabetic or non diabetic. Our paper is one such approach to analyze the Pima Indian Diabetic Dataset (PIDD) using weighted knn classifier. Our proposed system as shown in figure 1, focuses on improvement of classification accuracy. We have directed our concept in to two ways. (i)First of all, given dataset is normalized such that any supervised algorithm is used to learn the dataset. With this approach all values of dataset lies within range of 0 to 1. We have adopted here is column normalization.

$$Normalization = \frac{input\ column\ data - minimum\ value\ in\ a\ column}{Maximum\ value\ of\ a\ column - minimum\ value\ in\ a\ column} \quad (1)$$

(ii) We have found the standard deviation from normalized attributes. Standard deviations of attributes are then used as power instead of power 2 in euclidean distance formula to modify distance calculation in KNN. This approach greatly improves the calculation accuracy.

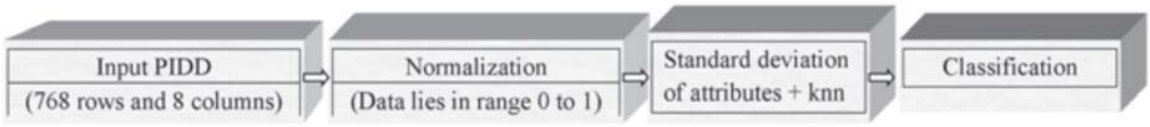

Figure 1. Proposed flow model

2. Related Work:

Many researchers have carried out using with machine learning and data mining tool to predict the medical diseases and the performance clearly reflects the use of machine learning approach on medical field does well. Thus it helps on accurate prediction of diseases. Jabbbar et al. (2013) used K-nearest neighbour algorithm and genetic algorithm (GA) to classify various data sets of UCI machine learning repository with a great accuracy[6]. In their implementation technique they used genetic search to rank the attributes and used the useful attributes for KNN classifier. Kourou et al.(2015) outlined a review of various machine learning approach on several cancer data and concluded that integrated approach of feature selection with classifier will provide a

promising result for analysis of cancer data[7]. Zhang et al. (2017) proposed a correlation matrix KNN (CM-KNN) to learn different values of k for different test data for classification, regression and missing data computation. The proposed method is dependent on correlation matrix of test data with the nearest neighbour to achieve better classification accuracy[8]. Kaur et al. (2018) used linear kernel support vector machine (SVM-Linear), Radial Basis Function (RBF), KNN, ANN Machine Learning algorithm in R data mining tool to analyze the Pima Indian Diabetic Dataset (PIDD)and concluded that SVM and KNN are the best classifier providing a accuracy nearly 89%[10]. Sisodia et al.(2018) used naive bayes, decision tree and support vector machine approach to classify pima indian diabetic dataset (PIDD). They performed the analysis using weka tool and concluded that the naïve Bayes had a better classification of accuracy of 76.30%[14]. Wu et al.(2018) developed a hybrid approach by combining K-means clustering to filter the noisy data from PIDD and used logistic regression to get an accuracy of 95.2% in Weka tool [9] . Sneha et al.(2019) applied naïve Bayes technique on pima indian diabetic dataset (PIDD) and got an accuracy of 82.30%[15]. Swapna G et al. applied long short term memory unit of recurrent neural network with convolutional neural network for feature selection from HRV data and these extracted features were used an input for SVM classifier to increase the accuracy up to 95.7% [16]. Atik Mahabub (2019)Proposed ensemble voting classifier, a Combination of k nearest neighbour, Multilayer perceptron and support vector classifier to classify PIDD with a classification accuracy of 86% [17].

3. Dataset Used:

Pima indian diabetes dataset (PIDD) is originated from national institute of diabetes and digestive and kidney diseases. Pima indian diabetes dataset (PIDD) consists of 9 attributes (8 predictor and 1 class label). It is the representation of 8 characteristics of 768 women having age more than 21 years. Pima indian diabetes dataset (PIDD) comes under supervised binary classification. The various attributes of this data base is described here.

3.1. Attribute Description of PIDD

Table 1:Dataset attribute information

Sl.No	Attribute	Description of attributes	Mean value	STD value
1	Pregnancies	Number of times	0.2262	0.19
2	Glucose	Plasma glucose concentration (mg/dL)	0.6075	0.16
3	Blood pressure	Diastolic blood pressure(mm Hg)	0.5664	0.15
4	Skin thickness	Tricep skin fold thickness(mm)	0.2074	0.16
5	Insulin	2-hour serum insulin(mu U/ml)	0.09	0.13
6	BMI	Body mass index(weight in kg/(height in m).^{2}	0.47	0.11
7	Pedi	Diabetespedigreefunction	0.16	0.14
8	Age	years	0.20	0.19
9	Target	1:diabetic,0:Non diabetic	0.34	0.47

Class Distribution: (class value 1 is interpreted as "tested positive for diabetes" and class value 0 is interpreted as non diabetic.) In Class label Number of instances 500 refers to 0 and 268 refers to 1. Here we have developed a modified knn supervised machine learning to classify in to diabetic and non diabetic of PIDD.

4. Learning

Learning process is an intelligent approach for developing a suitable algorithm or model which learns from its own experience, perform analysis on dataset to achieve a desired output with

error minimization. Our proposed machine learning approach is based on supervised learning. For analysis of Pima indian diabetes dataset (PIDD) various supervised learning approach like j48, SVM, knn, random forest e.t.c. were applied for classification[18]. Different feature selection techniques are proposed and compared with traditional feature selection technique for improvement of classification accuracy[19,20]. Newly adopted deep learning method is also used to classify diabetic and non diabetic patients[16].

4.1. Supervised Machine Learning

Supervised learning approach has two methods for processing data. One is regression and another is classification. Most of the analysis is done by supervised learning for Pima indian diabetes dataset (PIDD) [21]. In our paper we have adopted classification approach in which target is present and data analysis is carried out based on target.

4.2. Classification in Machine Learning

Classification in machine learning is one of prior decision making techniques used for data analysis. Various classifier techniques are too used to classify data samples. The main objective of our paper is to use a novel approach of machine learning for analysis of Pima indian diabetes dataset (PIDD) to achieve good accuracy. Some of popular classifier used in PIDD is described as

4.3. Support Vector Machine

SVM mostly preferred for supervised machine learning because of its simplicity, high accuracy and realiability. It can handle both linearly separable and non linear separable dataset in effective manner. Need of suitable hyper plane shown in below figure 2 is required to distinguish two class. This can be constructed with the help of support vectors to have a good margin for better classification and the technique is well preferred for PIDD[14].

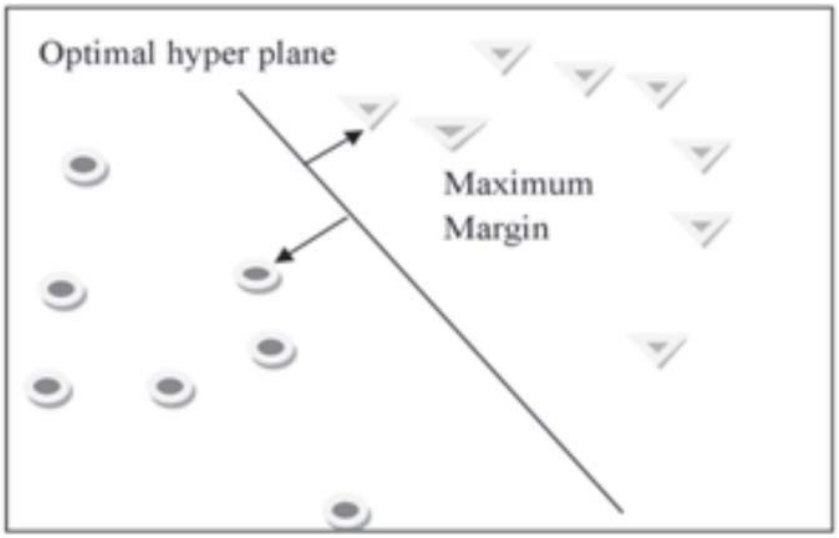

Figure 2. SVM

4.4. Logistic Regression

Generally occurrence of any event is predicted with the probability approach. One such statistical model which is used in medical field to analyze the severity of disease is logistic regression. Logistic regression is simple statistical model for analysis of dependent and independent features of input through sigmoid function. One of the finest approaches of logistic regression is binary logistic regression to analyze the binary input data through binary sigmoid function to classify input data of PIDD [22]. It predicts the probability occurrence of event by fitting to a logistic function.

4.5. Decision Tree classifier

Decision tree classifier is one of the best classifier to classify binary data. Decision tree classifier follows divide and conquer approach in which the dataset is represented in a tree structure [23, 24]. The structure is formed using the concept of information gain and depending on the importance of information the classifier performs the classification which is adopted by many researchers for PIDD [28]. The information gain for a particular attribute X at a node is calculated as

$$InformationGain(N, X) = Entropy(N) - \sum_{(valueatx)} \frac{|N|}{|N_I|} Entropy(N) \tag{2}$$

N: set of instance at that particular node and $|N|$ is its cardinality
Ni: set of attributes in N.
Entropy of N is found as

$$Entropy(N) = \sum_{i=1}^{no.ofclasses} -p_i log_2 p_i \tag{3}$$

p_i *is the portion of instance at* N.

4.6. Naïve Bayes (NB) classifier

Naive Bayes classifier is constructed from a family of machine learning classifier based on Bayes theorem. NB classifier develops a probabilistic model which assumes inter dependency of features to each other for prediction of outcome and also used for PIDD [25]. If y is class variable and x is dependent feature vector then class with maximum probability is defined as

$$y = argmax_y p(y) \prod (p\frac{x_i}{y}) \tag{4}$$

P(y) is class probability and P(xi/y) is conditional probability. Using the above formula we can find the class label of a given predictor.

4.7. Random forest classifier

Random forest(RF) classifier is mostly used to classify dataset having missing values. Through RF best features are extracted to classify dataset. Random forest is a collection of decision tree forming a random forest. For a given dataset random samples are extracted and these are fed to decision tree to obtain the output. Each tree in RF uses its own mechanism to classify the data sample in to a class. The most predicted result having more votes gives final output and used for PIDD [26, 27].

4.8. KNN classifier
K nearest neighbour(KNN) is simple, supervised learning algorithm to classify a given dataset with most accuracy. KNN has two property which are lazy learning and non parametric also. Nearest neighbour is measured by distance function that can be euclidean distance, manhattan distance, hamming distance or minkowski distance. KNN classifier is more easy to implement and versatile as compared to other machine learning technique for which many weighted KNN algorithm is implemented for analysis of PIDD[29]. So In our proposed method we modified the distance approach formula of KNN to find a new technique. Here we have discussed the weighted KNN algorithm and its impact for the analysis of PIDD. A generalized structure of a classifier through neural network is clearly shown in figure 3.

5. Proposed method and Model
After careful analysis and consideration of above result, a modified approach of KNN is purposed. KNN is a useful classifier for binary classification. In the initial process pre-analysis is done in which data cleaning, normalization and randomization process is applied [30]. The k nearest neighbour approach is based on distance matrix [32, 33]. Even though KNN is a non parametric classifier, here the analysis is carried out with consideration of test data value. In each occurrence of processing time the row data is shuffled randomly and distance value is calculated depending on the user selected K value for accuracy prediction. Euclidean distance is one of the mostly widely used methods to find the neighbour distance in KNN. PIDD dataset which is used in our analysis is divided in to training and testing part. In our paper a modified KNN algorithm is proposed to analyze the PIDD dataset with better classification accuracy. Thus instead of square in an euclidean distance, standard deviation(STD) as well as mean of STD of attributes is applied to modify distance equation [31].

5.1. Standard deviation:
Standard deviation is one of the important concepts of statistics used to measure the distribution of data about the mean. It is the root means square value of variance. Higher is the standard deviation, means more is the deviation of data from mean value. Standard deviation is classified in to two types Standard deviation in a whole dataset

$$Population\ Sample\ standard\ deviation\ (PSTD) = \sqrt{\frac{\sum(X-u).^2}{n}} \tag{5}$$

X=individual value of population
u=average of population
n= number of data points in the population
Sample standard deviation STD

$$Sample\ standard\ deviation\ (STD) = \sqrt{\frac{\sum(X-X_i).^2}{n-1}} \tag{6}$$

X=no attributes of dataset
$X_i = meanvalueofattributes$
$n = numberofdatapointsinthesample$

Here the standard deviation of PIDD after normalization is considered for weighted KNN structure. Using Standard deviation nearest neighbour is encountered and the method performs well as compared to simple Euclidean distance approach.

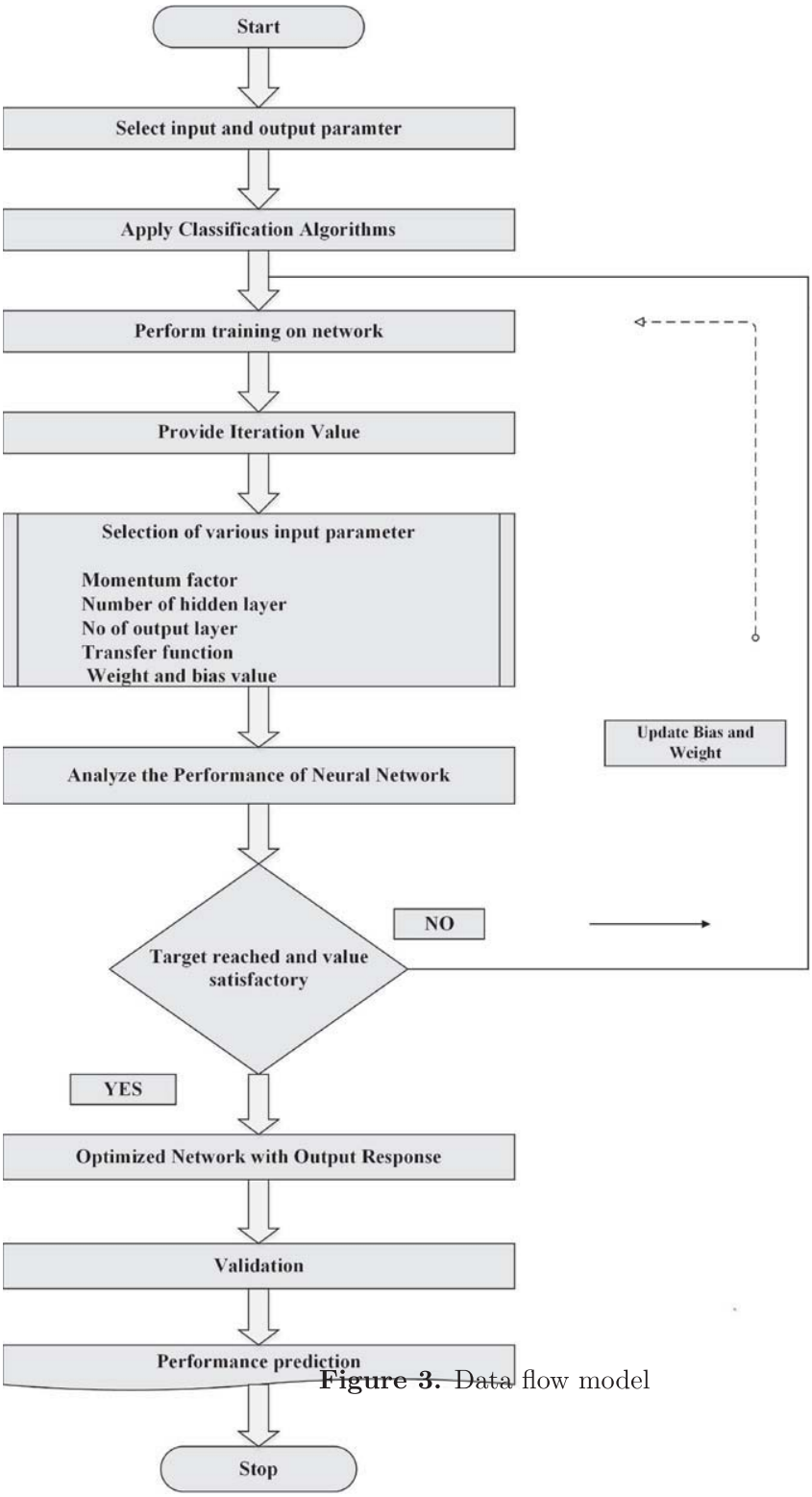

Figure 3. Data flow model

Method 1:

If Input vectors
X=$x_1, x_2, x_3, \ldots\ldots, x_N$
$Y = y_1, y_2, y_3, \ldots\ldots, y_N$

Distance between two input vectors is calculated, where x is input pattern from training part and y from testing dataset.

$$D(X,Y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^p} \tag{7}$$

p: is a positive integer i: 1, 2, 3...N so depending up on the value of P various distance formula is used to analyze the dataset.If P=2 then the approach is confined to Euclidean Distance. In our modified KNN approach instead of P=2 we have replaced with standard deviation (STD) of attribute. Standard deviation is used to measure spreading of data about the mean. In the proposed model standard deviation of the attribute was found out and it is applied as power to modify distance calculation of KNN. The approach is proposed in which the equation of distance is modified. Instead of power p, standard deviation (STD) of attribute will be used in the equation and it has shown that accuracy is improved for different values of k.

$$D(X,Y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^{STD}} \tag{8}$$

$x_i/c_i Dataset$
x_i : $Attributes$
c_i : $Classlabel$

Modified KNN (KNN + STD) Algorithm Pseudo code:

(Load the dataset, perform pre analysis and Normalization, Partition in to train dataset and test dataset).
1. Load training dataset (x_{train}) and test dataset (x_{test}).
2. Find the standard deviations of attributes (STD).
3. Calculate "distance D (x_{train}, x_{test})" where i=1,2..n and d represents the distance expressed in terms of standard deviation.

$$D(X_{\text{train}}, X_{\text{test}}) = \sqrt{\sum_{i=1}^{m}(X_{\text{train}} - X_{\text{test}})^{STD}} \tag{9}$$

If X_{train} = training data set value
X_{test}= test dataset value
4. Arrange the distance in decreasing order.
5. Choose the value of K and take the K distance from the sorted list.
6. Find those k values and depending on majority of class label to point class label is determined.
End

Method 2:

In another method mean of standard deviation was found out. The attributes having larger value than mean and attributes with lesser value than mean values are selected and applied as power for their corresponding attributes in distance calculation of KNN.
If Input vectors
X=x$_1, x_2, x_3, \ldots\ldots, x_N$
$Y = y_1, y_2, y_3, \ldots\ldots, y_N$

Value of distance of attributes greater than mean of standard deviation:

$$D(X_1, Y_1) = \sqrt{\sum_{i=1_i}^{m} (X_{1\mathrm{i}} - Y_{2\mathrm{i}}))^{attributes>mean(STD)}} \tag{10}$$

Value of distance of attributes lesser than mean of standard deviation:

$$D(X_2, Y_2) = \sqrt{\sum_{i=1_i}^{m} (X_{2\mathrm{i}} - Y_{2\mathrm{i}})^{attributes<mean(STD)}} \tag{11}$$

In this case suppose the number of class label is c then c 1, 2, 3....
X_{train} = training data set value and X_{test}= test dataset value

Modified KNN (KNN + mean (STD)) Algorithm Pseudo code:

Load the dataset, perform pre analysis and Normalization, Partition in to train dataset and test dataset

1. Load training dataset (X_{train}) and test dataset (X_{test}).
2. Find the mean of standard deviations of attributes (STD)
3. Find Value of distance of attributes greater than mean of standard deviation

$$D(X_1, Y_1) = \sqrt{\sum_{i=1_i}^{m} (X_{1\mathrm{i}} - Y_{1\mathrm{i}})^{attributes>mean(STD)}} \tag{12}$$

$$D(X_2, Y_2) = \sqrt{\sum_{i=1_i}^{m} (X_{2\mathrm{i}}) - Y_{2\mathrm{i}})^{attributes<mean(STD)}} \tag{13}$$

4. Find Value of distance of attributes lesser than mean of standard deviation:
5. Find the value D as sum of D1 and D2.
4. Arrange the distance in decreasing order.
5. Choose the value of K and take the K distance from the sorted list.
6. Find those k values and depending on majority of class label to point class label is justified.

End.

Our KNN approach is involved in three steps,as shown in figure 4, first step basic KNN is used for classification, second step weighted KNN (KNN with standard deviation) is used and in third step attribute with mean value of standard deviation is used for finding nearest neighbour with a improvement of classification accuracy. So detailed process is described as below

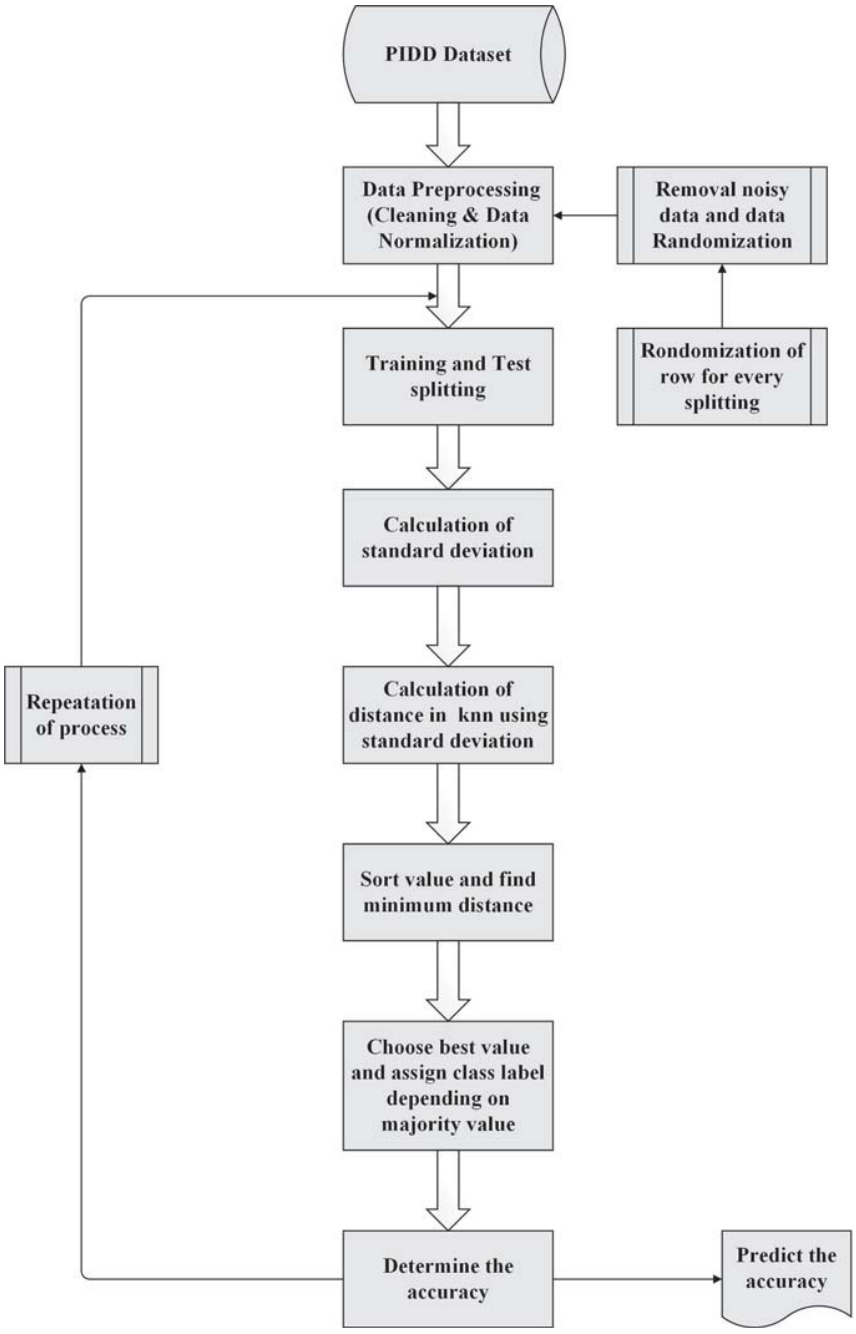

Figure 4. SDKNN Data flow model

6. Performance measure of previous counterpart work with relative classification accuracy

Table 2:Dataset classification comparison

classification accuracy			
Authors	Data size	Methods	Accuracy
Fayssal Belo-ufa and M.A chikh(2016)[13]	768 and 9 at-tributes	Modified artificial bee colony	84.21%
Yoichi Hayashi and shonosuke Yukita(2016)[32]	768 and 9 at-tributes	Modified Sam-pling Re-Rx with j48 graft	83.83%
Md Maniruz-zaman et al.(2017)[11]	768 and 9 at-tributes	Gaussian process based classifica-tion model using RBF kernel-k10	81.97%
Harleen Kaur and vinita ku-mari(2018)[10]	768 and 9 at-tributes	Modified Linear kernel SVM	89%
Deepti Sisodia and dilip singh sisodia(2018)[14]	768 and 9 at-tributes	Modified Naïve Bayes	76.3%
Han Wu (2018)[9]	589 with 9 at-tributes	Improved Means with logistic re-gression	95.4%
Ch.Sanjeeb ku-mar Dash et al.(2019)[31]	768 9 attributes	Modified iTLBO and RBFN (im-proved teaching learning based optimization+ radial Basis Function kernel)	81.77%
Gopi Battineni et al.(2019)[33]	768 and 9 at-tributes	Logistic regres-sion	77%
Atik Ma-habub(2019)[34]	768 and 9 at-tributes	Modified Ensem-ble voting classi-fier	85.71%
Dilp kumar chubeY t et al.(2020)[35]]	768 and 9 at-tributes	Modified PSO Naive Bayes	78.6%
Our Proposed Method	768 and 9 at-tributes	Modified Weighted knn(SDKNN)	83.76%

7. Result Analysis:

As we have discussed the PIDD dataset is normalization before any supervised classification process gets implemented for classification. The standard deviation of all attributes is calculated for our process. Initially we applied kNN .We have divided total dataset in to (9:1) ratio for train data and test data respectively. Various frequencies of operation and analysis are represented

here in tabulation form.

Table 5 :Table for KNN Average accuracy of 81.77

KNN	k=5	k=6	k=7	k=10	k=15	k=20	k=25	k=27
	85.71	79.22	83.11	83.11	83.11	80.51	74.02	83.11
	81.81	80.51	.81	77.92	84.41	85.71	81.81	81.811
	83.11	85.1	87.01	85.71	83.11	77.92	80.51	79.22
	87.01	80.51	84.41	81.81	75.32	79.22	76.62	80.51
	84.41	81.81	79.22	76.62	85.71	81.81	85.71	80.85
Average Accuracy	84.41	81.43	83.11	81.03	82.33	81.03	79.73	81.1

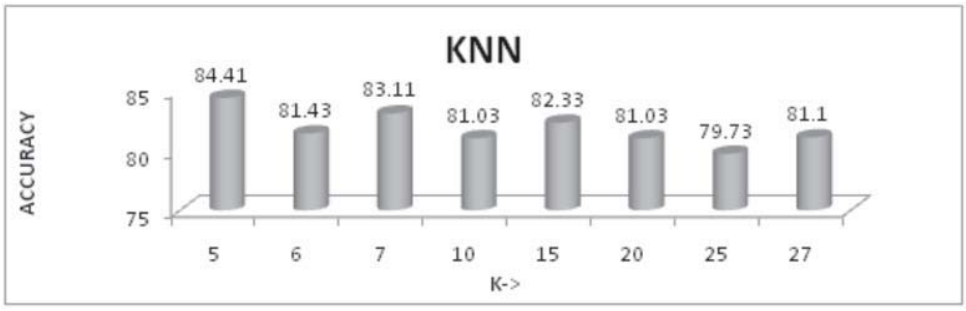

Figure 5. KNN with Euclidean distance approach having Average accuracy of 81.77%

Table 4:KNN with standard deviation approach having Average accuracy of 82.27%

KNN+STD	k=5	k=6	k=7	k=10	k=15	k=20	k=25	k=27
	84.41	90.9	87.01	80.51	81.81	79.22	85	81.81
	83.11	88.31	84.41	83.11	79.22	77.92	79.22	72.27
	87.01	85.56	81.81	81.81	85.71	76.62	85.71	77.92
	83.11	80.51	79.22	84.41	84.41	84.41	764.02	80.51
Average Accuracy	84.41	84.67	82.59	82.07	83.11	80.25	81.67	79.38

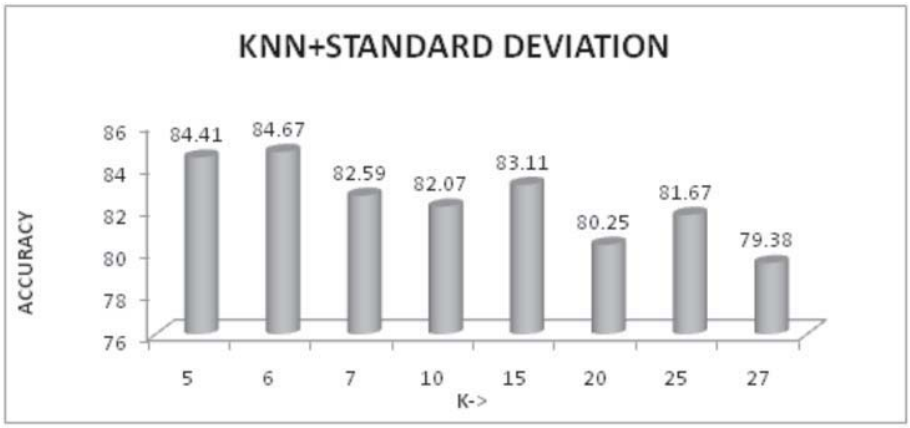

Figure 6. KNN with standard deviation approach having Average accuracy of 82.27%

Table 3:Table for KNN Average accuracy of 83.76

KNN	k=5	k=6	k=7	k=10	k=15	k=20	k=25	k=27
	83.01	83.11	79.22	85.71	83.11	83.11	81.11	79.22
	83.11	81.11	87.01	81.81	87.01	84.41	80.51	81.81
	88.31	85.71	83.11	88.31	85.71	80.51	83.11	80.51
	80.51	84.41	88.31	83.11	84.41	85.71	85.71	84.41
Average Accuracy	85.71	84.41	85.45	84.67	84.41	82.59	82.07	80.77

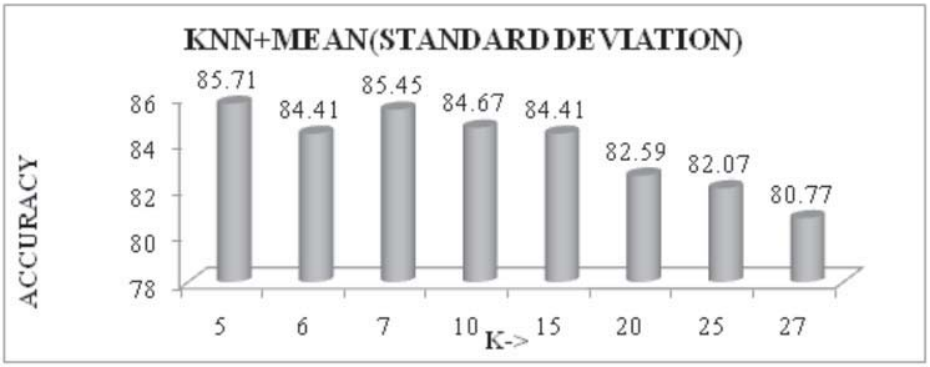

Figure 7. KNN with mean standard deviation(SDKNN) approach having Average accuracy of 83.76%

8. Conclusion:

Using adaptive KNN classifier we have achieved an accuracy of 83% for our PIDD. As we have found the proposed algorithm uses modified Euclidean distance concept in for analysis of PIDD to predict diabetic or non diabetic patient as compared to other classifier. The value of K plays an important role. Deep learning method with proper choice of K value will possibly improve the accuracy performance for more concrete prediction.

References

[1] Global Report on Diabetes - World Health Organization
[2] Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2020, American Diabetes Association, Diabetes Care 2020 Jan; 43(Supplement 1): S14-S31
[3] Kharroubi, A. T., Darwish, H. M. (2015). Diabetes mellitus: The epidemic of the century. World journal of diabetes, 6(6), 850–867. doi:10.4239/wjd.v6.i6.850
[4] Wu, Y., Ding, Y., Tanaka, Y., Zhang, W. (2014). Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. International journal of medical sciences, 11(11), 1185–1200. doi:10.7150/ijms.10001
[5] Larabi-Marie-Sainte, S., Aburahmah, L., Almohaini, R., Saba, T. (2019). Current Techniques for Diabetes Prediction: Review and Case Study. Applied Sciences, 9(21), 4604
[6] Jabbar, M. A., Deekshatulu, B. L., Chandra, P. (2015). Classification of heart disease using k-nearest neighbor and genetic algorithm. arXiv preprint arXiv:1508.02061. biotechnology journal, 13, 8-17.
[7] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.
[8] Zhang, S., Li, X., Zong, M., Zhu, X., Cheng, D. (2017). Learning k for knn classification. ACM Transactions on Intelligent Systems and Technology (TIST), 8(3), 1-19.
[9] Wu, H., Yang, S., Huang, Z., He, J., Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10, 100-107.ISSN 2352-9148,https://doi.org/10.1016/j.imu.2017.12.006
[10] Kaur, H., Kumari, V. (2018). Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics.
[11] Maniruzzaman, M., Kumar, N., Abedin, M. M., Islam, M. S., Suri, H. S., El-Baz, A. S., Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Computer methods and programs in biomedicine, 152, 23-34

[12] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116.,
[13] Beloufa, F., Chikh, M. A. (2013). Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm. Computer methods and programs in biomedicine, 112(1), 92-103
[14] Sisodia, D., Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585,doi.org/10.1016/j.procs.2018.05.122
[15] Sneha, N., Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data, 6(1), 13
[16] Swapna, G., Vinayakumar, R., Soman, K. P. (2018). Diabetes detection using deep learning algorithms. ICT Express, 4(4), 243-246
[17] Mahabub, A. (2019). A robust voting approach for diabetes prediction using traditional machine learning techniques. SN Applied Sciences, 1(12), 1667
[18] Kandhasamy, J. P., Balamurali, S. J. P. C. S. (2015). Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 47, 45-51, https://doi.org/10.1016/j.procs.2015.03.182
[19] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 9, 515
[20] Alehegn, M., Joshi, R., Mulay, P. (2018). Analysis and prediction of diabetes mellitus using machine learning algorithm. International Journal of Pure and Applied Mathematics, 118(9), 871-878.
[21] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116
[22] Rahimloo, P., Jafarian, A. (2016). Prediction of Diabetes by Using Artificial Neural Network, Logistic Regression Statistical Model and Combination of Them. Bulletin de la Société Royale des Sciences de Liège, 85, 1148-1164
[23] Pradeep, K. R., Naveen, N. C. (2016, December). Predictive analysis of diabetes using J48 algorithm of classification techniques. In 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I) (pp. 347-352). IEEE
[24] Orabi, K. M., Kamal, Y. M., Rabah, T. M. (2016, July). Early predictive system for diabetes mellitus disease. In Industrial Conference on Data Mining (pp. 420-427). Springer, Cham.
[25] Pei, D., Gong, Y., Kang, H., Zhang, C., Guo, Q. (2019). Accurate and rapid screening model for potential diabetes mellitus. BMC medical informatics and decision making, 19(1), 41.. https://doi.org/10.1186/s12911-019-0790-3
[26] Mukasheva, A., Saparkhojayev, N., Akanov, Z., Apon, A., Kalra, S. (2019). Forecasting the Prevalence of Diabetes Mellitus Using Econometric Models. Diabetes Therapy, 10(6), 2079-2093
[27] Kumar Das, Sujit and Kumar Mishra, Arnab and Roy, Pinki, Automatic Diabetes Prediction Using Tree Based Ensemble Learners (March 19, 2019). International Journal of Computational Intelligence IoT, Vol. 2, No. 2, 2019. Available at SSRN: https://ssrn.com/abstract=3355532
[28] Sohail, M. N., Jiadong, R., Uba, M. M., Irshad, M., Iqbal, W., Arshad, J., John, A. V. (2019). A hybrid Forecast Cost Benefit Classification of diabetes mellitus prevalence based on epidemiological study on Real-life patient's data. Scientific reports, 9(1), 1-10.
[29] Yan, X., Li, W., Chen, W., Luo, W., Zhang, C., Wu, Q., Liu, H. (2013). Weighted K-nearest neighbor classification algorithm based on Genetic Algorithm. Telkomnika, 11(10), 6173-6178.
[30] Zhang S, Li X, Zong M, Zhu X, Wang R. (2018). Efficient kNN Classification With Different Numbers of Nearest Neighbors. IEEE Transactions on Neural Networks and Learning Systems, 29, 1774-1785. doi: 10.1109/TNNLS.2017.2673241. Epub 2017 Apr 12
[31] Dash, C. S. K., Behera, A. K., Dehuri, S., Cho, S. B. (2019). Building a novel classifier based on teaching learning based optimization and radial basis function neural networks for non-imputed database with irrelevant features. Applied Computing and Informatics
[32] Hayashi, Y., Yukita, S. (2016). Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. Informatics in Medicine Unlocked, 2, 92-104
[33] Battineni, G., Sagaro, G. G., Nalini, C., Amenta, F., Tayebati, S. K. (2019). Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods. Machines, 7(4), 74.
[34] Mahabub, A. (2019). A robust voting approach for diabetes prediction using traditional machine learning techniques. SN Applied Sciences, 1(12), 1667.
[35] Choubey, D. K., Kumar, P., Tripathi, S., Kumar, S. (2020). Performance evaluation of classification methods with PCA and PSO for diabetes. Network Modeling Analysis in Health Informatics and Bioinformatics, 9(1), 5