# A Survey Study Support Vector Machines and K-MEAN Algorithms for Diabetes Dataset

**Nora Ibrahem Alghurair**

Fahad Bin Sultan University-KSA

noraibrahem018@gmail.com

**Advisor: Dr. Mohammad A. Mezher**

Fahad Bin Sultan University-KSA

mmezher@fbsu.edu.sa

## Abstract

Diabetes is actually one of the primary causes of human mortality. Diabetes is a intense disease affecting various parts of the human body. Diabetes can rise long-range complications including, renal failure and cardiac failure.. It is therefore imperative that diabetes be diagnosed in a timely manner people all over the world. In this study, a survey was conducted on data testing on SVM technology using a different kernel as well as data test results were surveyed with a complementary algorithm between SVM and K-mean to diagnose diabetes and compare their results with previous studies. This comparison was made to UCI PIMA Indian Dataset and using Anaconda python.

**Keywords:** K-means, SVM, Diabetes, PIMA Indian, UCI.

## Introduction

Today the worldwide buzzword is "healthcare". Predicting an disease in a community can command an important part, in improving healthcare in society. Diabetes, is a group of metabolic defect that lead to high grade, of blood sugar a long period .Diabetes is classified in two .The first one is the patient's inability to produce insulin in the body. Secondly, the body becomes insulin resistant, so insulin cannot function normally. Therefore diabetes may be said to be a serious incurable condition(Ramesh ; Caytiles ; Iyengar , 2017).

Statistics show that there are about 194 million people living with diabetes worldwide. Statistics also point to an rise to 333 million by 2025(Ateeq ; Ganapathy, 2017). People have long suffered from various diseases that have in some cases been able to treat and provide ways to help them, but often, poor luck, because of undiagnosed symptoms in patients for a long time may even threaten life Patient. Therefore, Recent studies (Sundaram , 2018)& (Mirza ; Mittal; Zaman , 2018) shown that a trending field of diabetes was the prediction.

In the medical part, one of the application of decision support models is the diagnosis of diseases such as diabetes. Delays in diagnosing and predicting diabetes due to uncontrolled blood glucose increases the risk of failure of human organs such as the kidneys, eyes, heart, nerves, etc. (Dadgar ; Kaardaan)

However, there have been many predictions of diabetes in the past, such as: probabilistic neural network Bayes modeling (Sujarani; Kalaiselvi, 2018), SMOTE (Mirza ; Mittal; Zaman, 2018), decision tree algorithm (Kadhm ; Ghindawi ; Mhawi, 2018), genetic algorithm ( Choubey ; Paul, 2017), neural network (El_Jerjawi; Abu-Naser, 2018).

Therefore, this study aims to conduct a survey of the previous techniques and studies and compare their results with the results obtained from data testing using SVM technology with a various kernels as well as with an incorporated technology based on the SVM and K-mean algorithm to diagnose diabetes Support vector machine (SVM) is a supervised learning algorithm exercised to carry out classification as well as regression problems. The main function of this algorithm is to predict the separation membership of the categorical goal by creating hyper planes in a multi-dimensional space separating the cases of denominations of different classes.( Sethi ; Goraya ; Sharma, 2017) The main algorithm is the Support Vector Classification (SVC) and it is about visualizing "margin" both sides of the hyperlink that divides two data categories with the most widen margin.

SVM works on two categories of data, linear SVM for insulate data and linear SVM for non-insulate data. In the case of linear SVM for separable data, it takes only one plane to insulate the data, but in the state of linear SVM for non-insulate data, one of the excessive planes is required and the kernel trick is used. SVM supports the most accurate prediction accuracy that provides outstanding performance in the field of bioinformatics. SVM is also primarily a high-performance method of classification (Stoean ; Stoean ; Preuss ; El-Darzi ; Dumitrescu, 2006).

Clustering is the essence function of producing data and a public technique for analyzing statistical data applied in many field,, including machine learning, pattern recognition, image analysis, and bioinformatics. One from the most commonly applied algorithms in clustering is K-means.

Clustering K-means is a form of unsupervised learning that is applied when you have unnamed data (i.e. data without specific groups).The objective from that algorithm is to search for collection in the data , with the number of groups represented by the variable K. The algorithm works frequently to specify each data point into one of the K groups based on available features. Data points are Clustering established on the similarity of the feature (Sharmila; Vetha Manickam, 2016).

## Literature review

***Many researchers have developed and designed diabetes prediction systems established on various algorithms and style.***

D. Kumar et al. ( Choubey ; Paul, 2017), suggest a system for classifying diabetes through two step; in the first step, (GA) applied as a feature, and in the second step it was used (RBF NN) to classify the selected traits among all traits. In (Dadgar ; Kaardaan),a proposal of the UTA optimization algorithm applied to displace ineffective and destructive features and the GA used to generate approach a optimum value of weight.

In (Ateeq ; Ganapathy, 2017), this study a modern hybrid classification algorithm called MPSO-NN. The suggest technique combined with the algorithm with Multi-Layer Perceptron Network.

In (Mirza ; Mittal; Zaman , 2018) ,the suggest system was used in two step. In the first step, the data defect is removed utilize SMOTE and in the second stage diabetes is diagnosed using Decision Tree (DT) classifier.

In this work (Thoma;Joseph;Johnson;Thomas, 2019).they suggested studying the DT algorithm and evaluating it based on accuracy. Decision trees tend to match quickly and also have poor prediction accuracy for low-volume sample responses.

In (Kadhm ; Ghindawi ; Mhawi, 2018), the proposed model applied K-NN algorithm to extract undesirable data ,therefore decrease processing time. However ,the suggest classification approach established on the DT to assign each data pattern to its appropriate group. through experiments ,the suggest system accomplished a high classification consequence ..

In (Giveki;Salimi;Bahmanyar;Khademian, 2012), the suggest model consists of three step: first, the PCA is utilized to define an ideal subset of features from a set of all features. Second, mutual information is used to construct (Feature Weighted Support Vector Machines) by weighting various features based on their importance. Finally, since parameter selection plays a vital role in SVM rating accuracy, MCS is applied to determine the best parameter values.

In this study (Kose;Guraksin;Deperlioglu, 2016), a diabetes diagnostic system formulated by both Transmission Support Servers (SVM) and Knowledge Development Algorithm (CoDOA). Besides SVM training, CoDOA conducted to determine the sigma coefficient for the kernel RBF function. The suggest approach was able to Determine diabetes.

the research in (Osman;Aljahdali, 2017), propose a system decided on data mining technology for predicting diabetes. The proposed system contains three main steps:step1, preparation and study of the data set. In preprocessing. step2, collect diabetes data using the K. method algorithm. Step 3, Diabetes prediction and diagnosis using K-SVM. Obtaining optimal diagnostic.

However, many researchers utilize different technique in order to obtain the better prediction rate. In ( Sethi ;  Goraya ; Sharma, 2017) ,they proposed five main techniques for categorizing diabetes. Technologies applied, ANN, SVM, KNN, Naive Bayes and Ensemble.

## Model and implementation

In this study, a set of Pima Indians Diabetes data extracted from the UCI website, which aims to diagnose whether a patient has diabetes or not, is obtained. This data set contains 768 samples taken from women with at least 21 years of age. The data set Composed of 9 features and a binary value for a class, If the value of the test is 1, then this means that the patient has diabetes, and if the test value is 0, then this means that the patient does not have diabetes. Table I shows the features of this dataset used.

TABLE I. PIMA INDIAN DIABETES DATASET

| NO | Feature |
|---|---|
| 1 | Amount of pregnancies |
| 2 | Concentration of plasma glucose 2 hours during an oral glucose tolerance test |
| 3 | Blood pressure diastolique (mm Hg) |
| 4 | Triceps skin fold thickness (mm) |
| 5 | Insulin 2-Hour Serum (mu U / ml) |
| 6 | Index of body mass (weight in kg/(hight in m)^2) |
| 7 | A pedigree function of diabetes |
| 8 | Age (years) |
| 9 | Class attribute (0 or 1) |

To implement the proposed method, Anaconda Python is utilized as an illustrative and high-standard programming language that is utilized in many fields, such as the web and desktop programs, This is because it utilize a lot number strong frame and libraries That enabled her to gain wide popularity. Furthermore, this language is quite simple because the code is simple to write and read for this language.

In this survey, it is divided into three phase. in the first phase, sundry tests are performed on a collection of diabetes data by applying SVM and using three kernels .The first is linear, the second is RBF, and the third is sigmoid.

The second step, many data tests are performed by applying the integrative technology between svm and k-mean.

The third step, a comparison is made between the test results obtained from the first step and the second step and comparing their consequence with past studies and define the best techniques.

## Experimental and Results discussion

Survey, the consequence are calculated as follows:

$$\text{Accuracy}= (TN+TP)/(TN+FP)+(TP+FN)x100$$

Where, True Positive (TP):Diabetic counts and the number of patients without diabetes are measured correctly. False Positive (FP):Calculates the number of diabetics without diabetes. True Negative (TN)The number of diabetes sufferers and the number of patients without diabetes was incorrectly calculated. False Negative (FN)This is the number of people who have no diabetes listed as diabetic.

A. **first experiment** At this phase, the diabetes data was divided to test it many times in order to get better accuracy using SVM and three of the kernels which are linear, RBF and sigmoid. The results were as follows:

TABLE II. RESULTS ON THE ORIGINAL PIMA INDIAN DIABETES DATASET USING SVM ALGORITHM

| kernel | Accuracy |
|---|---|
| **linear** | 83% |
| **RBF** | 82% |
| **sigmoid** | 69% |

Receiver Operating Characteristic (ROC) metric to evaluate classifier output fineness using cross-validation Typically ROC curves are true positive on the Y axis and false positive on the X axis. It pointing that the top left angle is the “perfect” dot a false positive rate of zero, and a true positive rate of one. This isn't very actual, , but it aims to have a wide area underneath the curve (AUC) is usually better.

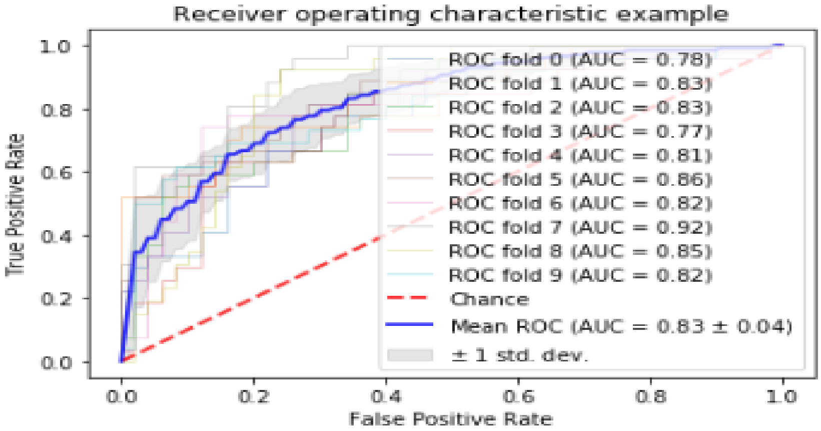

Fig. 1. Accuracy of linear SVM

Use the linear SVM classifier ROC response, generated from K-fold cross-validation (10 folds).

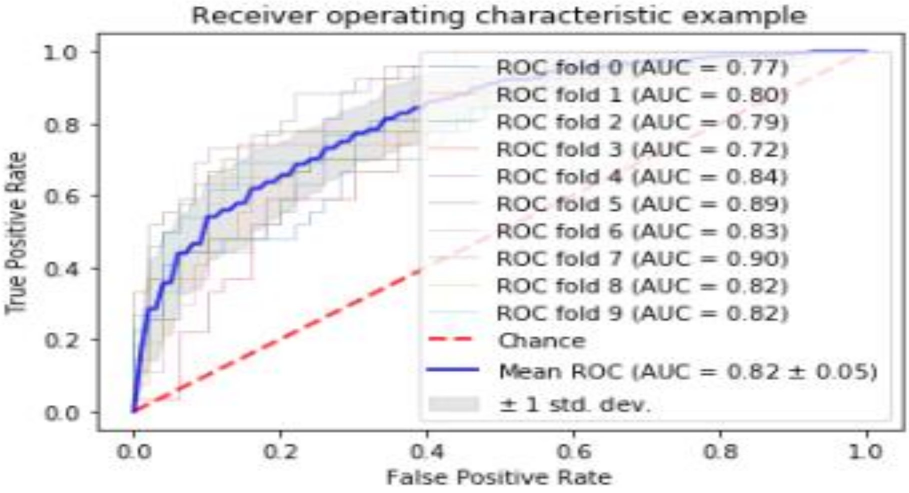

Fig. 2. Accuracy of RBF SVM

Use the RBF SVM classifier ROC response, generated from K-fold cross-validation (10 folds).

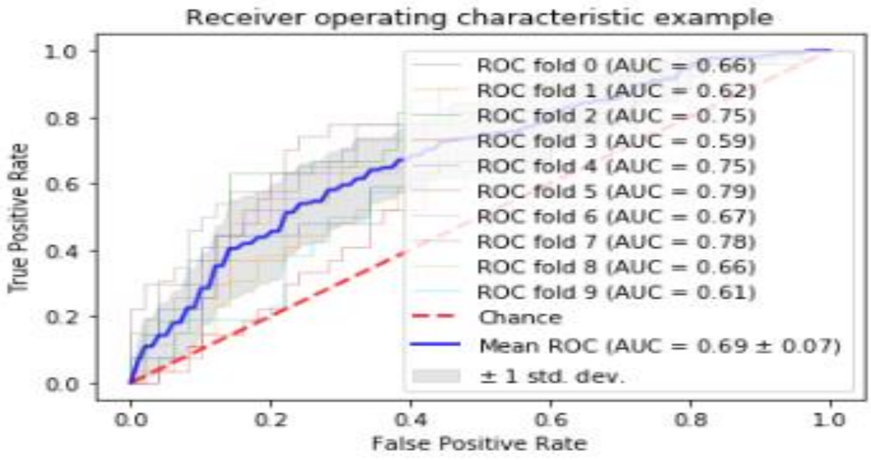

Fig. 3. Accuracy of Sigmoid SVM

ROC response of different datasets, created from K-fold cross-validation (10 folds) using Sigmoid SVM classifier.

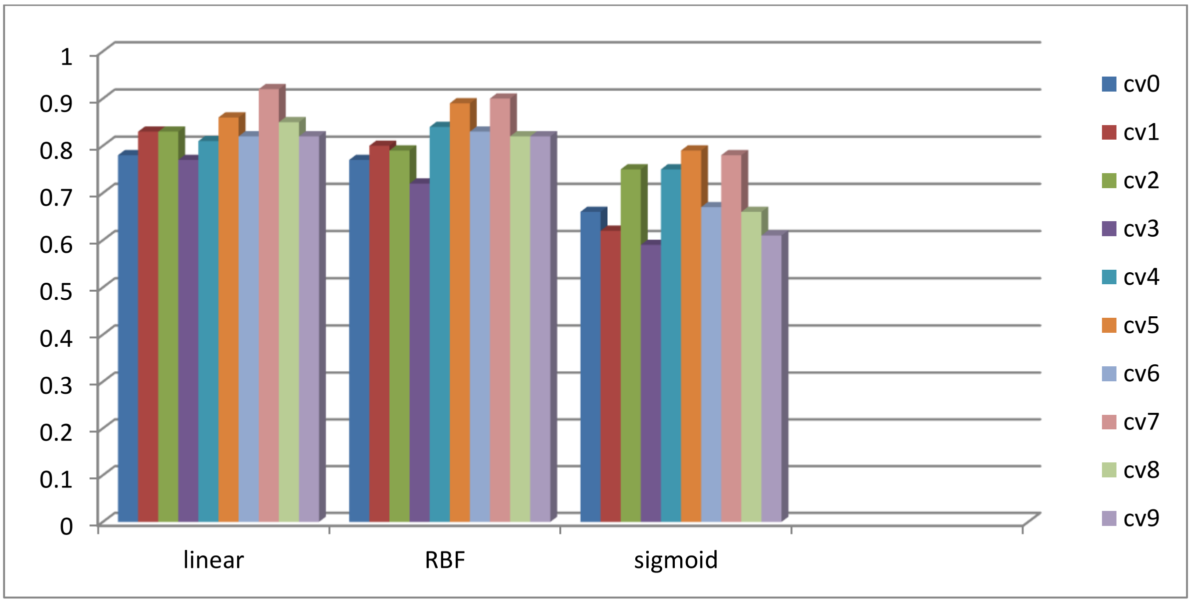

Fig. 4. Comparison of a different kernel

Figures4 show the accuracy of SVM algorithm consequence in with the better category in the test trials .These are the diagrams gained for the SVM output test. The better consequence was the accuracy of the linear kernel test with a accuracy of 83%.

B. **second experiment** At this phase, diabetes data were divided to test it multiple times in order to obtain better accuracy using an complementary algorithm among the k-means algorithm (as a feature extractor) and SVM. The result were as follows:

TABLE III. RESULTS ON THE ORIGINAL PIMA INDIAN DIABETES DATASET USING K-MEAN + SVM ALGORITHM

| Algorithm | **Accuracy** |
|---|---|
| k-means + SVM | 82% |

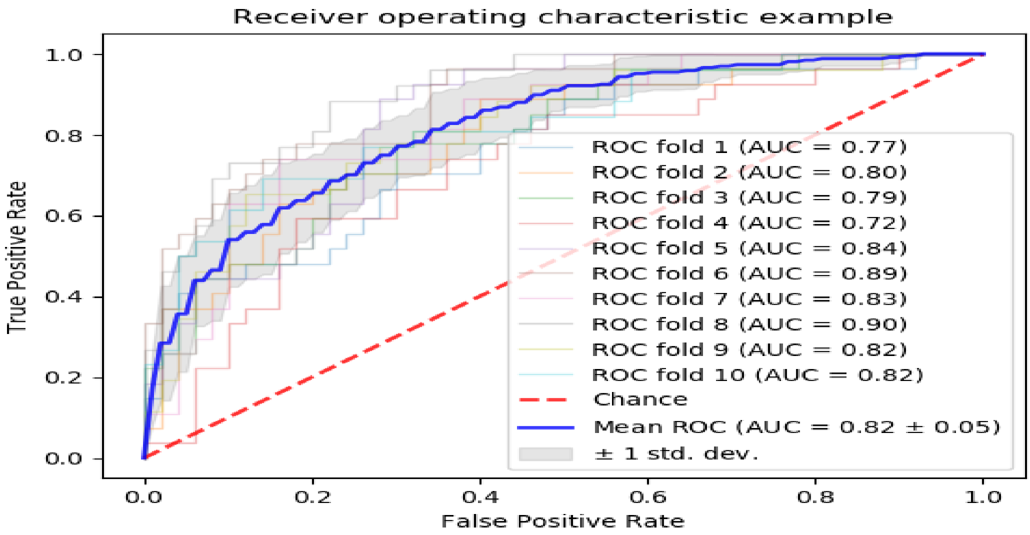

Fig.5. Accuracy of k-means + SVM

C. **third experiment** Table 4 shows a compare of the accuracy consequence obtained in this study and other previous for diabetes prediction. Using the same dataset.

TABLE IV. COMPARISON OF ALGORITHMS

| Method | Accuracy | References |
|---|---|---|
| K-means Clustering & Decision Tree | 98.7% | (Kadhm ; Ghindawi ; Mhawi, 2018) |
| Elman Neural network & soft max | 95.7% | (Sundaram , 2018) |
| PNN model &Bayes classification | 89.56% | (Sujarani; Kalaiselvi, 2018) |
| Levenberg-Marquardt | 82% | ( Zhang;Lin;Kang;Ning;Meng, 2018) |
| SVM technique & K-means | (99.9%) data with 70% data size | (Osman;Aljahdali, 2017) |
| GA_RBF NN | 77.4% | ( Choubey ; Paul, 2017) |
| MPSO-NN algorithm | 81.8% | (Ateeq ; Ganapathy, 2017) |
| UTA&genetic algorithm | 87.46% | (Dadgar ; Kaardaan) |
| (SVM) and (CoDOA) | 87,50% | (Kose;Guraksin;Deperlioglu, 2016) |
| Decision tree | 87 % | (Thoma;Joseph;Johnson;Thomas, 2019) |
| MI-MCS-FWSVM method | 93.58% | (Giveki;Salimi;Bahmanyar;Khademian, 2012) |
| linear SVM | 83% | this study |
| RBF SVM | 82% | this study |
| Sigmoid SVM | 69% | this study |
| k-means + SVM | 82% | this study |

In reference (Osman;Aljahdali, 2017), A pretty result show, and thus for using them they are part of the data set which is only 70%. Thus, the survey concluded that utilize SVM with K - MEAN on the entire data set, its results were poor compared to other algorithms. And if SVM is joint with the decision tree, it will give better data resolution results.

## Conclusion

One of the world's most important modern medical challenges is early diabetes detection. Trials were performed on the Pima Indian patient database using anaconda python. Prediction analysis is how the future user predicts the basis for current situations. During this study, a survey was conducted on the results obtained from data testing using SVM technology with a different kernels as well as a survey on the results of a complementary technique established on the SVM and K-mean algorithm to diagnose and evaluate diabetes based on accuracy.

Our result was poor compared to other previous algorithms, where it obtained the highest accuracy is the linear kernels by up to 83%. Integrated technology obtained 82% accuracy. Therefore, in the future, we suggest combining two SVM technologies and a decision tree in order to obtain the best accuracy results from previous studies.

## References

[1] Ramesh, S., &Caytiles, R. D., &Iyengar, N. C. S. (2017). A Deep Learning Approach to Identify Diabetes. Advanced Science and Technology Letters, 145, 44-49.

[2] Ateeq, K., & Ganapathy, G. (2017). The novel hybrid Modified Particle Swarm Optimization–Neural Network (MPSO-NN) Algorithm for classifying the Diabetes. International Journal of Computational Intelligence Research, 13(4), 595-614.

[3] Dadgar, S. M. H., &Kaardaan, M. A Hybrid Method of Feature Selection and Neural Network with Genetic Algorithm to Predict Diabetes.

[4] Sundaram, N. M. (2018). An Improved Elman Neural Network Classifier for classification of Medical Data for Diagnosis of Diabetes. International Journal of Engineering Science, 16317.

[5] Mirza, S., &Mittal, S., & Zaman, M. (2018). Decision Support Predictive model for prognosis of diabetes using SMOTE and Decision tree. International Journal of Applied Engineering Research, 13(11), 9277-9282.

[6] Kadhm, M. S.,& Ghindawi, I. W., &Mhawi, D. E. (2018). An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach. International Journal of Applied Engineering Research, 13(6), 4038-4041.

[7] Sujarani,P.& Kalaiselvi, K.(2018).Prediction of Diabetes Using Artificial Neural Networks: A Review. Jour of Adv Research in Dynamical & Control Systems.

[8] El_Jerjawi,N.S., & Abu-Naser.S.S.(2018).Diabetes Prediction Using Artificial Neural Network.International Journal of Advanced Science and Technology.

[9] Choubey, D. K., & Paul, S. (2017). GA_RBF NN: a classification system for diabetes. International Journal of Biomedical Engineering and Technology, 23(1), 71-93.

[10] Sethi,H., & Goraya,A.,& Sharma,V.(2017).Artificial Intelligence based Ensemble Model for Diagnosis of Diabetes.International Journal of Advanced Research in Computer Science,0976-5697.

[11] Stoean,R., & Stoean,C.,& Preuss,M.,&El-Darzi,E.,&Dumitrescu,D (2006). Evolutionary Support Vector Machines for Diabetes Mellitus Diagnosis .International IEEE Conference Intelligent Systems, pp. 182-187.

[12] Sharmila,K.,& Vetha Manickam,S.(2016).Diagnosing Diabetic Dataset using Hadoop and K-means Clustering Techniques.Indian Journal of Science and Technology, DOI: 10.17485.

[13] Thoma,J.,& Joseph,A.,&Johnson,I.,&Thomas,J.(2019).Machine Learning Approach For Diabetes Prediction.International Journal of Information Systems and Computer Sciences, 2319 – 7595.

[14] Giveki,D.,&Salimi,H.,&Bahmanyar,G.,&Khademian,Y.(2012).Automatic Detection of Diabetes Diagnosis using Feature Weighted Support Vector Machines based on Mutual Information and Modified Cuckoo Search .ArXiv, 2319 – 7595.

[15] Kose,U.,&Guraksin,G.,&Deperlioglu,O.(2016).Cognitive Development Optimization Algorithm Based Support Vector Machines for Determining Diabetes .Brain Journal of Artificial Intelligence Research, 2067 – 3957.

[16] Osman, A. H., &Aljahdali, H. M. (2017). Diabetes disease diagnosis method based on feature extraction using K-SVM. Int J Adv Comput Sci Appl, 8(1).

Zhang, Y., &Lin, Z., &Kang, Y., &Ning, R., & Meng, Y. (2018).A Feed-Forward Neural Network Model For The Accurate Prediction Of Diabetes Mellitus. INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, 2277-8616.