AN IMPROVED PREDICTION MODEL FOR TYPE 2 DIABETES MELLITUS DISEASE USING CLUSTERING AND CLASSIFICATION ALGORITHMS

Mrs. P. Laura Juliet[1], T. Bhavadharani[2]

[1]Assistant Professor, Dept. of Computer Applications, Vellalar College for Women, Erode, Tamilnadu, India
[2]Research Scholar, Department of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India

---***---

***Abstract -** Due to its continuously increasing occurrence, more and more families are influenced by diabetes mellitus. In this study, we have proposed a novel model based on data mining techniques for predicting Type 2 diabetes mellitus (T2DM). The main problems that we are trying to solve are to improve the accuracy of the prediction model, and to make the model adaptive to more than one dataset. In order to get the experimental result, we used the Pima Indians Diabetes Dataset from UCI Machine Learning Repository. Pre-process the dataset using K-means clustering algorithm. CFS Subset Evaluation method is used for feature selection. The aim of this paper is to select the correlated features. In the proposed model, K-means is used for data reduction with classifiers for classification. The conclusion shows that, this model attained a higher accuracy when compared with other models. Moreover, this model ensures that the dataset quality is sufficient. To further evaluate the performance of this model, we applied it to two other datasets. Both experiments results show good performance. The result shows that the proposed model has reached better accuracy compared to other previous studies. On the basis of the result, it can be proven that the proposed model would be helpful in Type 2 diabetes diagnosis.*

Key Words: K-Means, Clustering, Feature Selection, CFS Subset Evaluation.

I. INTRODUCTION

Data mining, also known as knowledge discovery in databases (KDD) could meet this need by providing tools to determine knowledge from data. Data mining is the process of determining interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information sources, or data that are streamed into the system enthusiastically.

DIABETES DISEASE

The population of India is nowadays more than 1000 million. The estimation of the actual number of diabetics in India is around 40 million. This means that India truly has the highest number of diabetics of any one country in the entire world. IGT (Impaired Glucose Tolerance) is also a rising problem in India. The occurrence of IGT is thought to be around 8.7 per cent in urban areas and 7.9 per cent in rural areas, although this estimate may be too high. It is thought that about 35 per cent of IGT sufferers go on to develop Type 2 diabetes, so India is openly facing a healthcare crisis. In India, the type of diabetes varies considerably from that in the Western world. Type I is considerably rarer and only about 1/3 of Type 2 diabetics are overweight or obese. Data mining plays a major role in the field of medicine. The prediction model also helps the doctors in the process of diagnosis. Preventing the diabetes disease is an ongoing area of interest to the health care community.

The goal of the data mining methodology is to extract data from a data set and change it into a reasonable structure for further use. Our examination concentrates on this part of Medical conclusion learning design through the gathered data of diabetes. The primary goal of this examination is to assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes disease utilizing diabetes patient's database.

II. LITERATURE REVIEW

Juntao Wang and Xiaolong Su, et.al [2010] has proposed **"An improved k-means clustering algorithm"** and it is used widely in cluster analysis for that the K-means algorithm has higher efficiency and scalability and converges fast when dealing with large data sets. The K-Means clustering algorithm is a partition-based cluster analysis technique. According to the algorithm we firstly select k objects as initial cluster centers, then calculate the distance between each object and each cluster center and assign it to the nearest cluster, update the averages of all clusters, repeat this method until the criterion function converged [1].

Rojalina Priyadarshini et.al [2014] has proposed, **"A novel approach to predict diabetes mellitus using modified extreme learning machine".** Diabetes mellitus otherwise known as a slow poison by the medical experts is a major, alarming and gradually becoming a global problem. This paper experimented and used the concept of modified extreme learning machine to identify the patients of being diabetic or non-diabetic basing on some previously given data which in turn helps the medical people to identify whether someone is affected by diabetes or not.. The main aim of classification task is to classify objects into a number of categories or classes. [4].

Veena Vijayan V et.al [2015] has proposed **"Decision support systems for predicting diabetes mellitus –a review".** Diabetes mellitus is caused due to the increased level of sugar content in the blood. This can cause series complications like kidney failure, stroke, cancer, heart disease and blindness. The main objective of this study is to review the benefits of different pre-processing techniques for decision support systems for predicting diabetes which are based on Support Vector Machine (SVM), Naive Bayes classifier and Decision Tree. Accuracy of the decision tree is increased from 75.1 % to 79.01% and Naive Bayes is improved from 75.82 %to 79.01%, when the PCA and discretize is done before classification. But the accuracy of the SVM is decreased from 76.6 to 75.69 [6].

Md Abul Basar, Hassan Nomani Alvi, Gazi, et.al [2015] has proposed **"A review on diabetes patient lifestyle management using mobile application".** The objective of this review is to analyse the articles and the existing mobile apps that are currently available to support diabetes patient and based on the analysis to describe the opportunities for developing mobile apps giving special concern to the people of the developing and the least developed countries integrating most of the common features like blood glucose monitoring, data entry for various parameter, data storage and analysis, short message service (SMS) based service, communication with healthcare team, diet plan, nutrition information, exercise plan, medication suggestion etc. Our primary aim for the app user is to control blood glucose and the secondary aim is lifestyle management [2].

Phattharat Songthung et.al [2016], has proposed **"Improving type 2 diabetes mellitus risk prediction using classification".** Diabetes is a chronic disease that contributes to a significant portion of the healthcare expenditure for a nation as individuals with diabetes need continuous medical care. In order to prevent or delay the onset of type 2 diabetes, it is necessary to identify high risk populations and introduce behavior modifications as early as possible. There are six important attributes used to compute the score: age (years), gender, BMI, waist circumference (cm), presence of hypertension, and family history of diabetes in parents or siblings. The presence of each attribute is given a score based on the severity of the attribute, and scores are summed up into a total risk score ranging from 0-17. If the total risk score is six or higher, the individual is recommended to get a follow-up lab test for fasting blood glucose and undergo behavior modification [3].

III. SYSTEM METHODOLOGY

Step 1: Select the diabetes dataset from UCI Machine Learning Repository. The dataset consists of 768 patient information.
Step 2: Pre-process the dataset using missing value imputation approach (K-means clustering algorithm) and predict misclassified data and replace imputation value.
Step 3: Feature Selection has been done based on CFS Subset evaluation method using logistic regression algorithm in order to extract the best features.
Step 4: Cluster the data using K-means clustering algorithm to remove the incorrectly clustered data.
Step 5: Classify the clustered data using classification algorithms such as Naïve Bayes, Decision Tree, K Star, Logistic Regression and SVM.
Step 6: Compare and display the result
Step 7: Accuracy will be analysed and displayed.
Step 8: Finally evaluation metrics are calculated using the parameters.

DATA PRE PROCESSING

First, we have analysed each attribute's medical implication and its correlation to DM. We determined that the number of pregnancies has little connection with DM. Therefore, we transformed this numeric attribute into a nominal attribute. The value 0 indicates non-pregnant and 1 indicates pregnant. The complexity of the dataset was reduced by this process.

Second, there are some missing and incorrect values in the dataset due to errors or deregulation. Most of the inaccurate experimental results were caused by these meaningless values. For example, in the original dataset, the values of diastolic blood pressure and body mass index could not be 0, which indicates that the real value was missing. To reduce the influence of meaningless values, we used the means from the training data to replace all missing values.

After the above steps were applied, the unsupervised normalize filter for attribute was used to normalize all the data into the section [0, 1] by using (1), where x' is the mean or average value for the variable and s is the standard deviation for the variable. Value is the new normalized value. This avoids the complexity of calculation and accelerates the speed of the operation.

$$\text{Value} = \frac{value - X'}{S} \qquad (1)$$

FEATURE SELECTION

CFS Subset Evaluation

The aim of this paper is to select the correlated features or attributes of medical dataset so that patient need not to go for several tests and in future it is used for preparing the clinical decision support system which is helpful for decision making of disease prediction in a cheaper way. It evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.

CLASSIFICATION

The model consists of double-level algorithms. In the first level, K-means algorithm was used to remove incorrectly clustered data. The enhanced dataset was used as input for next level. Then, Classification algorithms were used to classify the remaining data.

K-means cluster algorithm

The K-means is one of the most standard cluster algorithms. It is a typical distance-based cluster algorithm, and the distance is used as a measure of similarity, i.e., the smaller distance between objects shows the greater similarity. Fig. 4.2 shows a graphic procedure of the K-means algorithm, and the procedures of the K-means Cluster algorithm are as follows:

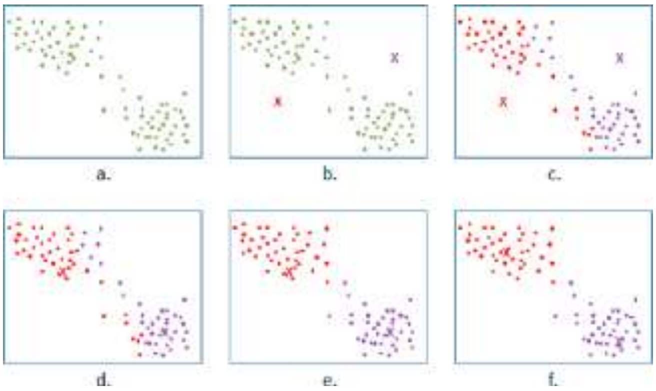

Fig. 1 Procedures of the K-means algorithm.

1) Show all objects (step a). Select K from provided N as the number of initial cluster centre (step b). In Fig. 3b, the value of K is 2, and we use the ‘ ✕ ’ to present the categories.

2) Calculate distance between each object and cluster centre. Cluster every object to the nearest cluster according to the distance using (2) extracted (step c).

$$S_i^{(t)} = \left\{\forall j, 1AjAk\, X_p : \| x_p - m_i^{(t)} \|^2 \leq \| x_p - m_j^{(t)} \|^2\ \forall j, 1 \leq j \leq k\right\} \forall j, 1AjAk \qquad (2)$$

3) Recalculate every cluster centre to verify whether they are changed using (3) extracted (step d).

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \varepsilon S_i^{(t)}} X_j \qquad (3)$$

Table 1 K-means cluster
Result of the 2-means cluster of the initial dataset.

No.	Label	Count
1	Cluster0	458
2	Cluster1	310

4) Circulate step 2 and step 3 until the new cluster centre is the same as the original one.

Logistic regression algorithm

The main purpose our experiment is to predict whether one person is diabetic or not, which is a typical binary-classification problem. Besides, the logistic regression algorithm is always used in data mining, disease automatic diagnosis and economic prediction, especially predicting and classifying of medical and health problem. In conclusion, we decided to use the logistic regression as one part of our proposed model. The logistic regression algorithm is based on the linear regression model expressed as (4).

$$P = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \qquad (4)$$

The predictive value of the classification problem can only be 0 or 1, so we may set a critical point. The output is 1 if the value is greater than the threshold, otherwise the output is 0. The output variable range of logistic regression is always between 0 and 1. Logistic regression is a regression model that reduces the prediction range and limits the prediction value to [0, 1]. Based on linear regression, the logistic regression enhances a layer of sigmoid function (non-linearity). The features are first summed linearly and then predicted using the sigmoid function. The main formulas of the logistic regression algorithm are shown in (5), (6), and (7).

Pr(Y=+1|X)$^{\sim}\beta$.XandPr(Y=-1|X)=1-Pr(Y=+1|X) (5)

$$\downarrow \sigma(x) := \frac{1}{1+e^{-x}} \epsilon [0,1] \qquad (6)$$

Pr(Y=+1|X)$^{\sim}\sigma(\beta.X)$ and Pr(Y=-1|X) = 1- Pr(Y = +|X) (7)

Naive Bayes

Naive Bayes is a classification technique with a notion which defines all features are independent and unrelated to each other. It defines that status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability it is considered as a powerful algorithm employed for classification purpose. It works well for the data with imbalancing problems and missing values. Naive Bayes is a machine learning classifier which employs the Bayes Theorem. Using Bayes theorem posterior probability $P\ (C|X)$ can be calculated from P(C), P(X) and $P\ (X|C)$ [5].

Decision Tree

Decision Tree is a supervised machine learning algorithm used to solve classification problems. The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data. It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, Decision tree chooses each node by evaluating the highest information gain among all the attributes [26].

K Star

K Star is an instance-based classifier that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. For the implementation of an instance-based classifier which uses the entropic distance measure. The intuition is that the distance between instances be defined as the complexity of transforming one instance into another. The calculation of the complexity is done in two steps. First a finite set of transformations which map instances to instances is defined. A "program" to transform one instance to another is a finite sequence of transformations starting at one instance and terminating at other [5].

SVM

SVM is one of the standard set of supervised machine learning model employed in classification. Given a two-class training sample the aim of a support vector machine is to find the best highest-margin separating hyperplane between the two classes. For better generalization hyperplane should not lies closer to the data points belong to the other class. Hyperplane should be selected which is far from the data points from each category. The points that lie nearest of the classifier are the support vectors. The SVM finds the optimal separating hyperplane by maximizing the distance between the two decision boundaries [5].

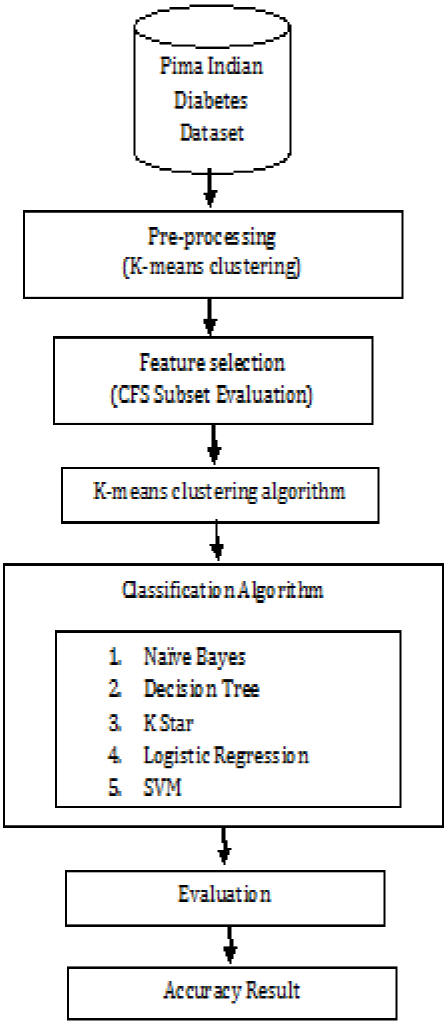
Pima Indian Diabetes Dataset
Pre-processing (K-means clustering)
Feature selection (CFS Subset Evaluation)
K-means clustering algorithm
Classification Algorithm
1. Naïve Bayes
2. Decision Tree
3. K Star
4. Logistic Regression
5. SVM
Evaluation
Accuracy Result

Fig 2 System Flow Diagram

IV. RESULT AND DISCUSSIONS

PERFORMANCE METRICS ANALYSIS

Precision Value: Precision is calculated as the number of true positive predictions divided by the total number of positive predictions.

$$Precision\ value = \frac{True\ positive}{(True\ positive + False\ positive)}$$

Recall value: Recall is calculated as the number of right positive predictions divided by the total number of positives.

$$Recall\ value = \frac{True\ positive}{(False\ positive + False\ Negative)}$$

F measure: The F measure is the harmonic mean of precision and recall

$$F\ measure = 2 * \frac{precision * recall}{precision + recall}$$

Table 2 Precision, Recall

Classification Algorithms	Precision	Recall
Naïve Bayes	0.77	0.775
Decision Tree	0.742	0.749
K Star	0.691	0.699
Logistic Regression	0.772	0.777
SVM	0.767	0.771

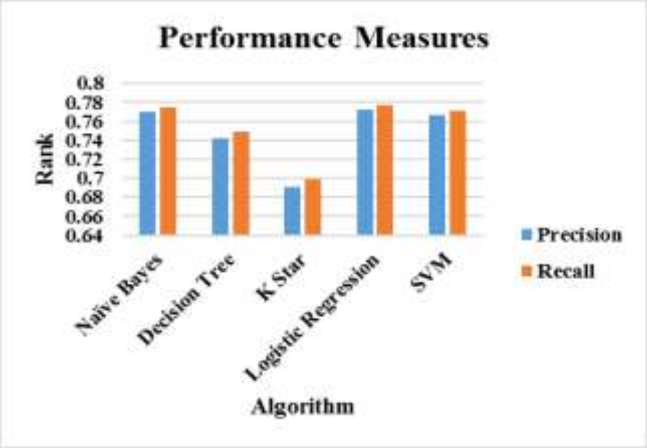

Fig 3 Precision, Recall

Table 3 F-Measure

Classification Algorithms	F-Measure
Naïve Bayes	0.769
Decision Tree	0.743
K Star	0.693
Logistic Regression	0.77
SVM	0.76

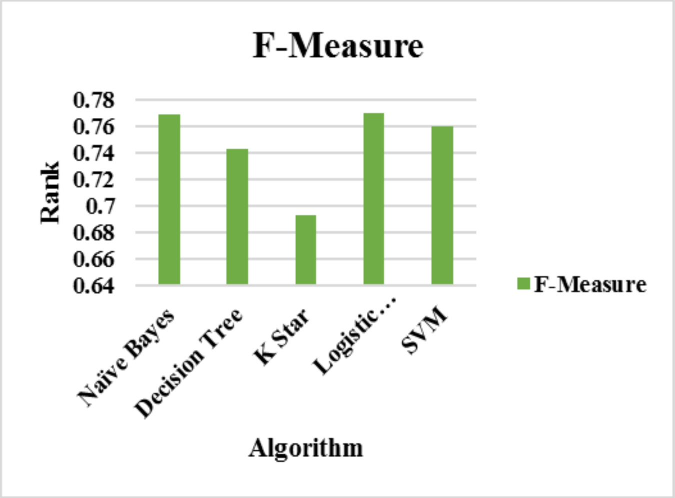

Fig 4 F-Measure

Accuracy

Accuracy is calculated as the number of totally correct predictions divided by the total number of the dataset.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

Table 4 Accuracy

Classification Algorithms	Existing Accuracy (Without Cluster)	Proposed Accuracy (With Cluster)
Naïve Bayes	75.30	77.474
Decision Tree	73.82	74.87
K Star	67.76	69.922
Logistic Regression	76.44	77.734
SVM	75.66	77.083

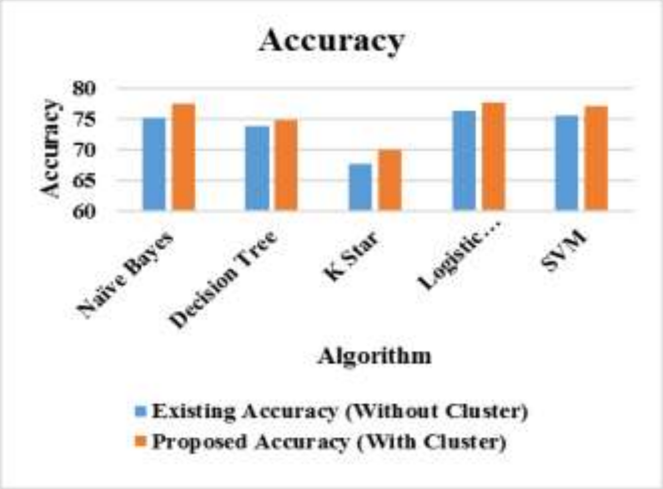

Fig 5 Accuracy

The efficient classification algorithms namely Naive Bayes, Decision Tree, K Star, Logistic Regression and SVM are used to develop the model. These algorithms are compared, and accuracy is evaluated. From the above table 5.2.8, it is observed that after applying K-means clustering algorithm all the classification algorithms gives better result and Logistic Regression had the best predictive power with high accuracy 77.73% as compared to Naive Bayes, Decision Tree, K Star, and SVM.

V. CONCLUSION AND FUTURE WORK

The proposed model that consisted of both cluster and class method ensured the enhancement of prediction accuracy. The proposed model has proven to be appropriate for predicting T2DM and selecting best features. One of the proposed model's benefits is that it avoids deleting overmuch original data. It ensures the high quality of experimental data. The other benefit is that, this model can apply in the Pima Indian Diabetes Dataset as well as other various datasets. While the limitation is that it consumes more time during the part of pre-processing.

The main problems that solved are improving accuracy of prediction model and making the model to adapt to different datasets. In this paper, after applying K-means clustering algorithm, the proposed model gives better result. The accuracy of Naive Bayes, Decision Tree, K Star, Logistic Regression and SVM was increased. Also Logistic Regression gives better accuracy with 77.73 % when compared to Naive Bayes, Decision Tree, K Star, and SVM. And the proposed K-means algorithm contributed a lot to the prediction model.

For future work, it is necessary to bring in hospital's real and latest patients' data for continuous training and optimization of this proposed model. The quantity of the dataset should be large enough for training and predicting. Some advanced algorithms and models should be applied in the study of DM. Grading forecasting standards are also necessary for potential diabetes patients. Developing a series of rules and standards is a valid method to prevent people from developing DM. Based on that, a more effective model for predicting DM and grading potential patients is presented. This will help to lower the growth rate of diabetes and eventually decrease the risk of developing DM.

REFERENCES

JOURNAL REFERENCES

1. Juntao Wang and Xiaolong Su, An improved K-Means clustering algorithm, 2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN).
2. Md Abul Basar, Hassan Nomani Alvi, Gazi, A Review on Diabetes Patient Lifestyle Management Using Mobile Application, 18th International Conference on Computer and Information Technology (ICCIT), 21-23 December, 2015.
3. Phattharat Songthung and Kunwadee Sripanidkulchai, Improving Type 2 Diabetes Mellitus Risk Prediction Using Classification, 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE).
4. Rojalina Priyadarshini, Nilamadhab Dash and Rachita Mishra, A Novel approach to Predict Diabetes Mellitus using Modified Extreme Learning Machine.
5. Tamilvanan.B, Dr.V. Murali Bhaskaran, An Experimental Study of Diabetes Disease Prediction System Using Classification Techniques, IOSR-JCE e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 19, Issue 1, Ver. IV (Jan.-Feb. 2017), PP 39-44.
6. Veena Vijayan V. and Anjali C., Decision support systems for predicting diabetes mellitus –a review. Proceedings of 2015 global conference on communication technologies (GCCT 2015).