PREDICTION ANALYSIS FOR DIABETIC PATIENTS USING CLUSTERED BASED CLASSIFICATION

[1]Randheer Singh, [2]Dr. Reema Ajmera, [3]Deevesh Choudhary

Abstract- The inexhaustible measure of information covered up in clinical databases that can be proficiently use to finding of patient's infections. This shrouded data exhibit in clinical databases can be utilized as rules after long procedure of lab examine, master check and positive outcome with understanding infections. This research work implement a hybrid clustering with classification model comprising of enhanced k-means and Logistic Regression classifier. The proposed technique comprises of clustering with classification i.e. clustered data is given to the classification for the evaluating the mining patterns. This research work also analyse the existing computational intelligence techniques for predicting diabetes. Experimental results demonstrate that the proposed technique outperforms the existing techniques.

Keywords – diabetes, clinical databases, clustering, classification, k-means, logistic regression, data mining, prediction.

I. INTRODUCTION

Diabetes is one of the most normal non-transmittable infections around the globe. It is evaluated to be 4th or 5th purpose behind death in the majority high-salary nations. Diabetes is designated by IDF as a standout amongst most difficult medical problems of 21st century. Diabetes is malady since antiquated circumstances. Diabetes is an ailment, basically a metabolic issue, individual has elevated hints of blood glucose occurred by deficient generation on the grounds that the cells in the body don't react the way they need to really react to insulin [1]. In the event that the follow stage of glucose increments then it is indicated by the different manifestations, for example, overwhelming thirst, visit pee, unexplained weight reduction and so on. [1].

Diabetes is a chronic illness that requires ceaseless medicinal care alongside the patient self-care to anticipate intense intricacies and to diminish the danger of long haul entanglements.

Diabetics populace is expanding far and wide inside all age gatherings. The IDF expresses that at regular intervals, two individuals are determined to have diabetes. Their expanding pattern prompts an expansion from their 30 million populace in 1985 to 150 million individuals in 2000 and is anticipated to rise further to 380 million by 2025.

Distinctive areas around the globe are influenced by diabetes pervasiveness in noteworthy diverse degrees. Among 138 million 20-79 years of age diabetics in 2013, Western pacific has the vast majority with diabetics and Africa has the slightest. In view of insights from Kela (Social Insurance Institution of Finland), number of individuals with diabetes in Finland was 195500 people in 2000, 220000 of every 2003 and is anticipated to ascend to a large portion of a million by 2030. Respects to various kinds of diabetes dispersion, Finland is known as a unique case in type 1 with the most astounding rate of rate on the planet.

Computerization in healthcare in general, and in the Operation Theater and intensive care unit in particular, is on the rise. Over last two decades there is rapid growth in the storage of large amount of laboratory examination data in databases. Medical information is characteristic of multi attribution, incompletion and closely related with time, hence medical data mining differs from that of others. Clinical data pertaining to the diagnosis and prognosis carried for various treatments have useful information hidden in them. Data mining in medicine is particular from that in different fields, in light of the fact that the information are heterogeneous; unique moral, legitimate, and social imperatives apply to medicinal data. Historical patient data can be mined to estimate the resource required for each disease category. Patient symptoms over a temporal interval can be mined to decide on better course of actions. Clinical Pathways that assist the physician in identifying the outcome of a treatment course could be generated from the patient historical data. The clinical pathway would assist the physician in identifying the alternate course of treatment that could be administered for a patient.

II. BACKGROUND

Umatejaswi et al. [2] planned model to deal with issues in existing structure in applying information mining systems to be specific grouping and orders that are connected to analyze the sort of diabetes & its seriousness level for each patient from information gathered. They attempted to analyze diabetes in light of the 650 patient's information with which creator investigated and distinguished seriousness of diabetes. As a major aspect of method Simple k-implies calculation is utilized for grouping the whole dataset into 3 bunches i.e., bunch 0 - for gestational diabetes, bunch 1 for type-1 diabetes, bunch 2 for type-2 diabetes. This grouped dataset was given as contribution to the order show which additionally orders every patient's hazard levels of diabetes as gentle, direct and extreme. Execution investigation of various calculations has been done on this information to analyze diabetes.

Kavakiotis et al. [3] aimed to lead an orderly audit of the utilizations of machine learning, information mining strategies and instruments in the field of diabetes with the main classification giving off an impression of being the most well known. An extensive variety of machine learning calculations were utilized. When all is said in done, 85% of those utilized were described

by regulated learning approaches & 15% by unsupervised ones, & all the more particularly, affiliation rules. SVM emerge as the best & generally utilized calculation. Concerning kind of information, clinical datasets were chiefly utilized. Title applications in chose articles venture the convenience of separating profitable information prompting new theories focusing on more profound understanding and further examination in DM.

Selvakumar et al. [4] predicted the persons regardless of whether diabetic or not. In this paper grouping systems are arranged for diabetes information and characterization exactness were looked at for ordering information. Diabetes mellitus is an endless illness and a noteworthy general wellbeing challenge around the world. Utilizing information mining techniques to help individuals to foresee diabetes has increase real ubiquity. This work focused the implementation of Binary Logistic Regression, Multilayer Perceptron and kNN for the diabetes information. From the investigation, it is analyzed that the development of groupings will be distinctive for arrangement strategies. From the histogram, it is seen that the Binary Logistic Regression accuracy is 0.69, Multilayer Perceptron accuracy is 0.71 and KNN gives the accuracy of 0.80.k- Nearest Neighbor is higher than the accuracy of Binary Logistic Regression and Multilayer Perceptron.

Komi et al. [5] explored the early prediction of diabetes via five different data mining methods including: GMM, SVM, Logistic regression, ELM, ANN. The diabetes forecast framework is created utilizing five information mining grouping displaying systems. These models are prepared and approved against a test dataset. Each of the five models can separate examples in light of the anticipated states. The best model to foresee understanding with diabetes give off an impression of being ANN taken after by ELM and GMM. Although not the most effective model, the Logistic regression result is easier to read and interpret, what is more, the training over Logistic regression is very efficient. Although the ANN do outperform other data mining methods, the relationship between attributes and the [mal result is more difficult to understand.

Gauri et al. [6] implemented machine learning calculation in Hadoop MapReduce condition for Pima Indian diabetes informational index to discover missing esteems in it and to find designs from it. The calculations can attribute missing esteems and to perceive designs from the informational collection. Prescient investigation is a strategy that coordinates different information mining procedures, machine learning calculations and insights that utilization present and past informational collections to find learning from it and by utilizing it foresee future events.

Dwivedi et al. [7] uses six computational intelligence techniques for diabetes mellitus prediction namely classification tree, support vector machine, logistic regression, naïve Bayes, and artificial neural network. Diabetes as a chronic disease is becoming a foremost community health concern worldwide. In developing countries, the diabetic patients are increasing rapidly due to lack of sentience and bad eating habits. So, there is a need of a framework that can effectively diagnose thousands of patients using clinical specifics. The performance of these techniques was evaluated on eight different classification performance measurements. Moreover, these techniques were appraised on a receiver operative characteristic curve. Classification accuracy of 77 and 78%was achieved by artificial neural network and logistic regression, respectively, with F 1 measure of 0.83 and 0.84.

Srikanth et al. [8] evaluated Classification Algorithms for the Classification of some Diabetes Disease Patient Datasets. This paper anticipated Diabetes Disease in view of Data Mining Techniques of Classification Algorithms. Order Algorithm and devices may lessen substantial work on Doctors. Characterization Algorithm analyzes the Decision Tree Algorithm, Byes Algorithm and Rule based Algorithm. These calculations assess Error Rates and recognize patients in light of development capacity to quantify the precise outcomes.

Ekta et al. [9] considered holistic approach to analyze and classify the diabetes dataset for data preprocessing. In order to do so, diabetes.arff dataset is used for data preprocessing and prediction of diabetes. From this research work one can easily analyzed that WEKA tool is quite useful for analyzing the given dataset. Finally, it is analyzed that persons who are suffered from diabetes have age more than 40 and mass more than 35. On the other hand, the persons who are not infected from diabetes have age less than 30 and mass less than 35.

Devi et al. [10] explored the early forecast of diabetes utilizing different information mining systems. Diabetes Mellitus is a perpetual sickness to influence different organs of the human body. Early expectation can spare human life and can take control over the infections. The dataset has taken 768 cases from PIMA Indian Dataset to decide exactness of information mining methods in forecast. Examination demonstrates that Modified J48 Classifier give the most astounding exactness than different systems. In restorative field exactness in expectation of infections is most critical factor instead of execution time. In investigation of information mining procedures & instruments Modified J48 Classifier gives 99.87% of most lifted precision using WEKA and MATLAB device. Since diabetes is an unremitting infection it must be stayed away from before it impacts people.

III.PROPOSED TECHNIQUE

This section presents methodology of the proposed technique.

1. **Dataset collection:** The dataset used in this research work is collected from National Institute of Diabetes and Digestive and Kidney Diseases and is based on Pima Indian Diabetic Set from University of California, Irvine (UCI) Repository of machine learning databases. The Pima Indian diabetes database is a collection of medical diagnostic reports of 768 patients.
2. **Pre-processing:** The collected raw data is then pre-processed and formatted. If some missing values are there, it will handle all the missing values by either replacing those values by the mean/average of all the values or by removing.
3. **Proposed Technique:** The proposed technique comprises of clustering with classification i.e. clustered data is given to the classification for the evaluating the mining patterns. For the clustering of data, most commonly used algorithm is K-means but K-means algorithm have many limitations like:
   - K-means algorithm takes initial clusters randomly which is not necessarily true in real-world applications.
   - The K-means algorithm is sensitive to initial centres selection.

The enhanced K-means algorithm is applied for dimensionality reduction to remove outliers and noisy data. This optimized dataset is given as an input to Logistic Regression classifier to find the useful patterns. According to the existing techniques surveyed, when K-means is combined Logistic Regression it will give better results but the issue is the problem of selecting the initial centroids in K-means; this problem is mainly focused in the proposed technique.

1. Data is partitioned into k equal parts. Then the arithmetic mean of each part is taken as the centroid point.
2. K-means is applied on the input dataset by finding the Euclidean distance of each data point from the centroid and clusters are defined. If the distance of centroid of the present nearest cluster is less than or equal to the previous distance, then the data point remains in that cluster and there is no need to find its distance from other cluster centroids.
3. Apply clustering on the dataset for dimensionality reduction and then classify that reduced dataset using Logistic Regression classifier.
   The initial centroids are randomly selected in case of simple K-means algorithm but it is not so in proposed algorithm. The proposed work is to select the initial centroids by partitioning the data into k equal parts and then the arithmetic mean of each part is taken as the centroid point. The efficiency and accuracy of enhanced K-means algorithm is more than simple K-means.
   Logistic regression
   Logistic regression uses an equation as the representation, very much like linear regression. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modelled is a binary values (0 or 1) rather than a numeric value.
   Below is an example logistic regression equation:
   $y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$
   where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data. The actual representation of the model that you would store in memory or in a file are the coefficients in the equation (the beta value or b's).
4. Evaluate the performance of the proposed technique on the basis of accuracy, precision, recall, Root mean squared error.

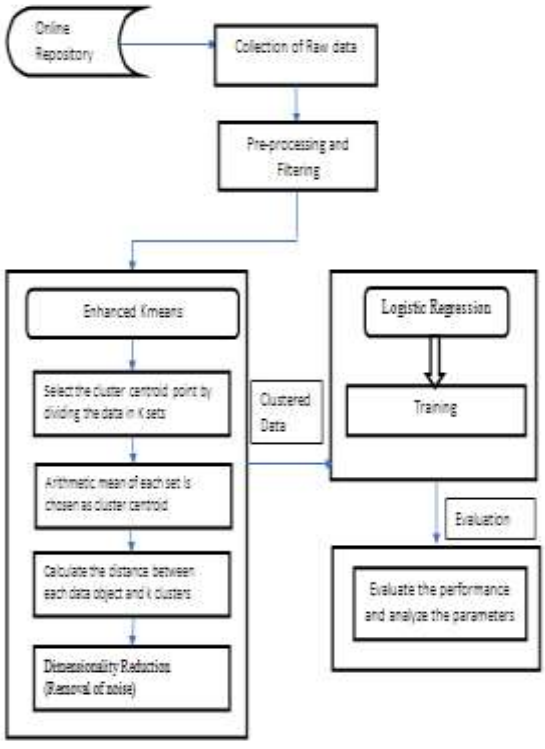

Figure 1: Flowchart of Proposed Technique

IV. EXPERIMENTAL RESULTS

This section presents results of the proposed technique. The implementation of proposed technique is done in Java NetBeans.

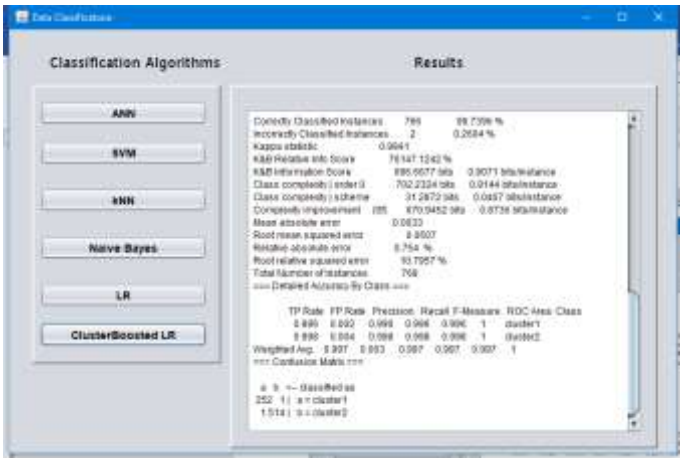

Figure 2: Showing the results of proposed technique

The figure above shows the classification results of proposed algorithm. The results show the accuracy of 99.7396% i.e. 766 instances are correctly classified out of 768 instances. Class details parameters are also shown like precision which is 0.997, recall 0.997, F Measure 0.997, TP Rate 0.997 and FP rate 0.003.

Precision and Recall: Precision and recall are mostly used to evaluate performance in text mining, and in text analysis areas like to retrieve information. These are utilized to measure exactness and completeness respectively.

$$\text{Precision}=\frac{\text{True Positive}}{\text{True Positive}+\text{False Positive}}$$

$$\text{Recall}=\frac{\text{True Positive}}{\text{True Positive}+\text{False Negative}}$$

F-measure: F-Measure is the harmonic mean of precision and recall. The esteem computed utilizing F-measure is stability among precision and recall.

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

Accuracy: Accuracy is the normal measure for arrangement execution. Precision can be estimated as effectively characterized examples to the aggregate number of occurrences, while mistake rate utilizes inaccurately grouped cases rather than accurately ordered cases.

$$\text{Accuracy}=\frac{\text{True Positive}+\text{True Negative}}{\text{True Positive}+\text{False Positive}+\text{True Negative}+\text{False Negative}}$$

The results are formulated using Diabetes dataset on proposed algorithm in contrast with the existing algorithms including ANN, SVM, KNN, naïve bayes, and logistic regression. Results shows that the proposed algorithm classify the diabetes patients more efficiently by showing highest accuracy. The results are illustrated and compared by defining the following tables.

Table 1: Accuracy comparison of Existing and Proposed Classification on Diabetes Dataset

Algorithm	Accuracy
ANN	75.651
SVM	65.1042
KNN	70.0521
Naïve Bayes	76.5625
Logistic Regression	77.6042

Proposed Algorithm	99.7396

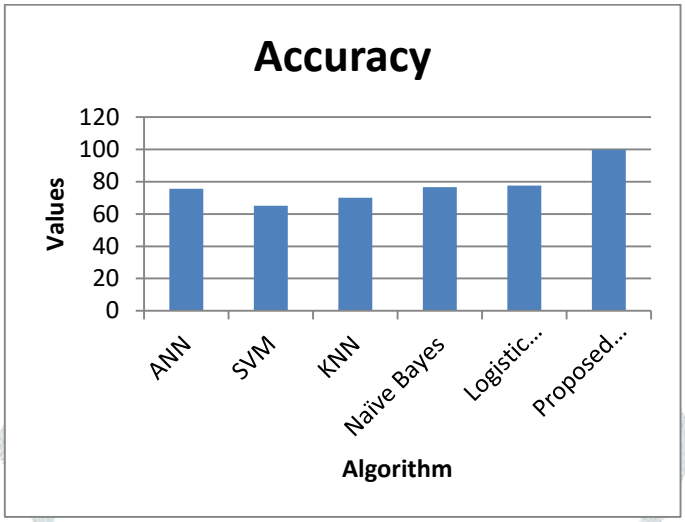

Figure 3: Showing the accuracy comparison of existing with the proposed classification algorithm

The figure above shows the accuracy comparison of the existing algorithms including ANN, SVM, KNN, naïve bayes, logistic regression, and proposed algorithm. The graph clearly shows that the proposed algorithm performs better as its accuracy is 99,7396% i.e. it has more efficiently classifies the patients.

Table 2: Class details parameters comparison of Existing and Proposed Classification on Diabetes Dataset

Para-meters	ANN	SVM	KNN	Naïve Bayes	Logi-stic Regre-ssion	Pro-posed Algo-rithm
Precisi on	0.752	0.424	0.694	0.761	0.771	0.997
Recall	0.757	0.651	0.701	0.766	0.776	0.997
F-Measur e	0.753	0.513	0.697	0.763	0.769	0.997

The table above shows the class parameters comparison between the algorithms. The algorithm with more precision, recall and Fmeasure is the better among all the above. The table clearly shows that the proposed algorithm is better than the existing classification algorithms.

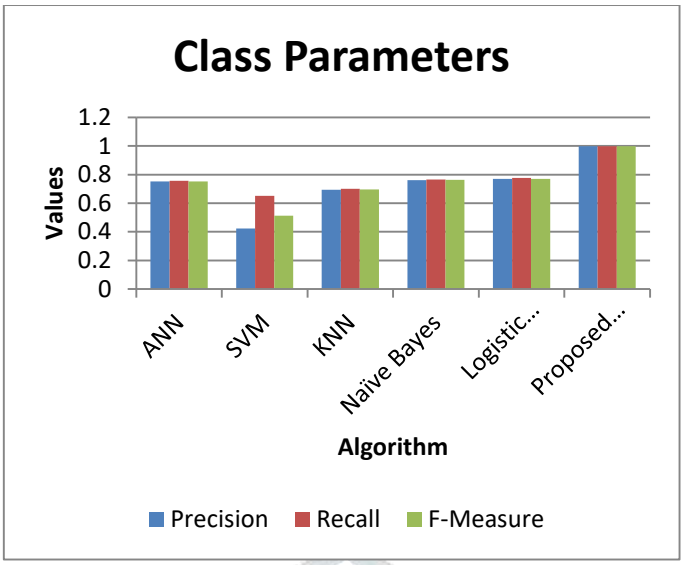

Figure 4: Showing the class parameters comparison of existing with the proposed classification algorithm

The figure above shows the class parameters comparison of the existing algorithms including ANN, SVM, KNN, naïve bayes, logistic regression, and proposed algorithm. The class parameters include precision, recall and Fmeasure. The graph clearly shows that the proposed algorithm performs better as its precision, recall and Fmeasure all are more than the existing base classifiers.

Table 3: Error rate comparison of Existing and Proposed Classification on Diabetes Dataset

Error Rate	ANN	SVM	KNN	Naïve Bayes	Logistic Regre-ssion	Proppsed Algo-rithm
Mean Absolute Error	0.2955	0.349	0.3001	0.2853	0.3097	0.0033
Root Mean Square Error	0.4213	0.5907	0.5465	0.418	0.3956	0.0507

The table above shows the error rate comparison between the algorithms. The algorithm with less error rate i.e. mean absolute error and Root mean square error is the better among all the above. The table clearly shows that the proposed algorithm is better than the existing classification algorithms.

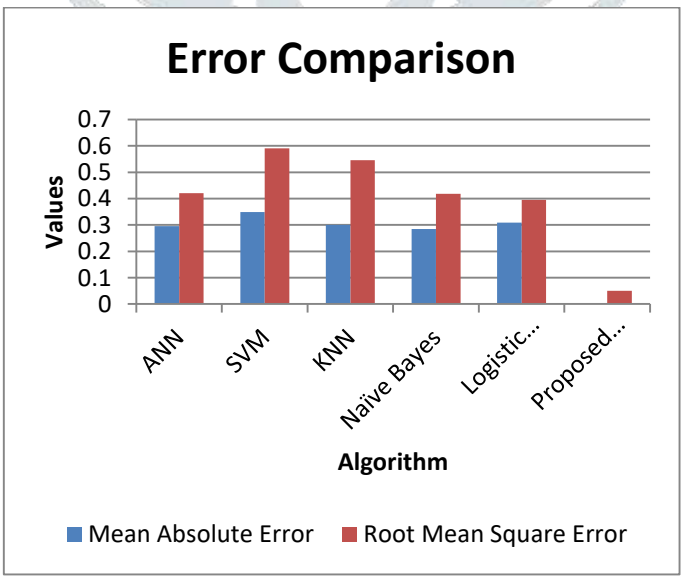

Figure 5: Showing the error rate comparison of existing with the proposed classification algorithm

The figure above shows the error rate comparison of the existing algorithms including ANN, SVM, KNN, naïve bayes, logistic regression and proposed algorithm. The class parameters include precision, recall and Fmeasure. The graph clearly shows that the proposed algorithm performs better as its error rate is less than the existing base classifiers.

V. CONCLUSION

This research work implement a hybrid clustering with classification model comprising of enhanced k-means and Logistic Regression classifier (cluster-boosted Logistic Regression). The proposed technique comprises of clustering with classification i.e. clustered data is given to the classification for the evaluating the mining patterns. This research work also analyse the existing computational intelligence techniques for predicting diabetes. The dataset used in this research work is collected from National Institute of Diabetes and Digestive and Kidney Diseases and is based on Pima Indian Diabetic Set from University of California, Irvine (UCI) Repository of machine learning databases. The Pima Indian diabetes database is a collection of medical diagnostic reports of 768 patients. Results are evaluated on the basis of accuracy, precision, recall, root means squared error and to compare the performance of the proposed technique with the existing techniques. Proposed technique gives accuracy of 99.7396. Proposed technique is compared with ANN, SVM, KNN, naïve bayes and logistic regression algorithm. Proposed technique gives more accuracy, precision, recall, and f-measure and less error as compared to existing algorithms and hence performs better.

REFERENCES

[1] B. Senthil Kumar, Dr. R. Gunavathi. 2016. A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis. International Journal of Advanced Research in Computer and Communication Engineering. 5(12):463-467.
[2] P. Suresh Kumar and V. Umatejaswi. 2017. Diagnosing Diabetes using Data Mining Techniques. International Journal of Scientific and Research Publications. 7(6):705-709.
[3] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda. 2017. Machine Learning and Data Mining Methods in Diabetes Research. Elsevier Computational and Structural Biotechnology. 104–116.
[4] S.Selvakumar, K.Senthamarai Kannan and S.GothaiNachiyar. 2017. Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques. International Journal of Statistics and Systems. 12(2):183-188.
[5] Messan Komi, J un Li, Y ongxin Zhai, Xianguo Zhang. 2017. Application of Data Mining Methods in Diabetes Prediction. IEEE 2nd International Conference on Image, Vision and Computing. 1006-1010.
[6] Gauri D. Kalyankar, Shivananda R. Poojara, Nagaraj V. Dharwadkar. Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop. IEEE International conference on I-SMAC. 619-624.
[7] Ashok Kumar Dwivedi. 2017. Analysis of computational intelligence techniques for diabetes mellitus prediction. Springer.
[8] Panigrahi Srikanth, Dharmaiah Deverapal. 2016. A Critical Study of Classification Algorithms Using Diabetes Diagnosis. IEEE 6th International Advanced Computing Conference. 245-249.
[9] Ekta, Sanjeev Dhawan. 2016. Classification of Data Mining and Analysis for Predicting Diabetes Subtypes using WEKA. International Journal of Scientific & Engineering Research. 7(12):100-103.
[10] Dr. M. Renuka Devi, J. Maria Shyla. 2016. Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus. International Journal of Applied Engineering Research. 11(1):727-730.