Prediction of Type2 Diabetes Patients using Rule Based K-Means Algorithm

Krishnamoorthy.P, R.Gobinath

Abstract: Diabetes Mellitus is an endless glycolysis issue, where the inappropriate administration of this illness can prompt to cardiovascular sickness, kidney malady, eye infection, nerve ailment, pregnancy difficulty and Dental intricacy. The datasets, so for gathered and preprocessed, involve certain qualities which are extremely satisfactory for diabetes mellitus conclusion. The utilization of this credits needs to upgrade the preparation and test order of patients to whether the patient to endure for tablet or insulin. Data Classification could be a prime undertaking in Data mining handling. Accuracy in information grouping undertaking can help the bunching of huge dataset fittingly. In this paper we have tested and proposed a Rule Based K-Means calculation as one of the critical strategy in idealistic field for sorting diabetic patients into two classes for accomplishing better outcomes.

Keywords- Rule Based K-Means, SVM, Decision tree, Naïve Bayes MATLAB.

I. INTRODUCTION

Diabetic Mellitus is endless ailment that is portrayed by high blood glucose level. About portion of the considerable number of diabetics have family heredity factors, which is one of the highlights in diabetic mellitus. Disappointment of pancreas to deliver enough insulin and the body's wasteful utilization of insulin are both pathologic foundations for diabetic mellitus. There are primarily four sorts of Diabetes Mellitus. They are Type1, Type2, Gestational diabetes and innate diabetes. Type 1 likewise called as "Insulin subordinate Diabetes Mellitus" or "Adolescent Onset Diabetes Mellitus" happens when the human body disappointments to create insulin. . They are described by the loss of insulin creating beta cells. Type 2 is likewise called as "Non-Insulin Dependent Diabetes Mellitus" or "Grown-up beginning diabetes "is described by the insulin obstruction. Gestational Diabetes Mellitus (GDM) are impermanent diabetes that looks like Type 2 diabetes in a few similar to the condition in which the pregnant ladies, recently analyzed diabetes show an expanded glucose in the blood. GDM is treatable under cautious supervision and resolves totally once the infant. The Congenital diabetes is caused because of hereditary de insulin discharge, cystic fibrosis-related diabetes, diabetes initiated by high portions of glucocorticoids.

As per the World Health Organization (WHO)[22], 37 cores of people suffer from diabetes around the world before the year 2030, and around 48 lakhs of people with in the year 2012. Most of them belong to lower and middle class families. Common manifestations of diabetes are characterized by insufficient insulin production by pancreas, ineffective use of the insulin produced by the pancreas or hyperglycemia. Causes like fat, hypertension, elevated cholesterol level, high fat diet and sedentary lifestyle are the common factors that contribute to the prevalence of diabetes. Development of kidney diseases, blindness and coronary artery disease are types of the severe damage which are resulted by improper management and late diagnosis of diabetes. Even though there's no established cure for polygenic disorder, indeed, the glucose level of diabetic patients may be controlled by well-established treatments, precise nutrition and regular exercise.

This research work focus on already affected diabetic patient to prevent complexity of diseases.

In this work Rule Based K-Means calculation, DT, Support Vector Machine and NB calculations are performed in MAT LAB to assess the clinical dataset to anticipate the kinds of gatherings who are altogether influenced from the diabetic. Investigational introductions of the four calculations are identified with a few measures and accomplished upstanding accuracy results.

The residual investigation discourse is sorted out as pursues Part-II Literature work for many classification procedures for estimate process for the groups. Part-III Methodology utilized and diverse procedure of dataset. Part-IV states assessed results. Part-V Conclusion of my investigation work.

1.1. Diabetes in Type2 (NIDDM)

It is a perpetual condition that influences the manner in which your body uses sugar (glucose) - a significant wellspring of fuel for your body. It is acclimated be alluded to as grown-up beginning diabetes, yet today more youngsters are being determined to have the confusion, most likely because of the ascent in youth weight. There's no solution for type2 diabetes, however getting in shape, eating great and practicing can help deal with the infection. On the off chance that diet and exercise are insufficient to deal with your blood glucose well, you may likewise

Revised Manuscript Received on July 05, 2019..

Krishnamoorthy.P [1]Ph.D Research Scholar, Vels Institute of Science, Technology and Advanced Studies (VISTAS),Chennai, India.
Dr.R.Gobinath, [2]Associate Professor Vels Institute of Science, Technology and Advanced Studies (VISTAS),Chennai, India.
krishnamoorthy.cp@gmail.com ,iamgobinathmca@gmail.com

require diabetes prescriptions or insulin treatment. Signs and manifestations of sort to type 2 diabetes issue typically grow gradually. Truth be told, the type2 diabetes was not known for a considerable length of time. Searching for: Increased thirst, visit pee, expanded yearning, unintended weight reduction, Fatigue, Blurred vision, Slow-recuperating bruises and incessant diseases.

Components that may expand your danger of type2 diabetes include: Weight, Fat dissemination, Inactivity Family history, Race, Age, Pre-diabetes, Gestational diabetes, Polycystic ovarian disorder, Areas of obscured skin, ordinarily in the armpits and neck.

1.2. Existing Algorithm in Diabetes Type2

Numerous analysts are leading trials for diagnosing the diseases utilizing different order calculations of machine learning approaches like J48, Support Vector Machine, NB, DT, Decision Table and so on as examines and demonstrated that the machine-learning calculations works better in diagnosing various ailments.

1.2.1. Support Vector Machine

It is one of the regular arrangements of directed machine learning standard utilized in characterization. Given a two-class preparing test the point of a help it is to pinpoint the best most noteworthy edge detachment hyperplane among the dual classes. For better conjecture, hyperplane ought not to lie nearer to the information focuses have a place with the different class. Hyper plane ought to be chosen which is a long way from the information exertions from every class. The focuses that untruth nearby to the pinpoint of the classifier are the help vectors.

1.2.2. Naive Bayes

It is a grouping method with a thought which characterizes all highlights to be autonomous and disconnected to one another. It characterizes that position of a particular component in a class not ensure influence the position of alternative element. Since it depends on restrictive likelihood it is considered as a ground-breaking calculation utilized for grouping reason. It functions admirably for the information with unbalancing issues and misplaced qualities. It is a machine learning classifier which utilizes the Bayes Theorem.

1.2.3. Decision Tree

It is a managed machine learning calculation used to tackle characterization issues. The fundamental goal of utilizing DT in this exploration effort is the hope of objective class utilizing choice guideline taken from earlier information. DT utilizes hubs and internodes for the forecast and order. Root hubs order the cases with various highlights. Root hubs can have at least two branches while the leaf hubs speak to characterization. In each stage, It picks every hub by assessing the most elevated data gain among every one of the characteristics.

II. LITERATURE REVIEW

2.1. Decision tree

Sajida et al. in [7] discusses the role of Adaboost and Bagging ensemble machine learning methods [8] using J48 decision tree as the basis for classifying the Diabetes Mellitus and patients as diabetic or non-diabetic, based on diabetes risk factors. Results achieved after the experiment proves that, Adaboost machine learning ensemble technique outperforms well comparatively bagging as well as a J48 decision tree.

The research [5] experimented GA-SVM model for predicting several real world datasets and the results showed that this model provided significant improvement in the performance of classification in comparison with Grid search.

The research paper [4] developed a method using combined dataset of Diabetes disease. Here select (accuracy- 63.54%, specificity- 43.00%, and sensitivity- 99.80%), wrapper (accuracy- 70.69%, specificity- 38.36% and Sensitivity- 89.95) and Ranker (accuracy- 72.61%, specificity- 41.04%, and sensitivity- 90.76%) methods are used for feature selection and LIBSVM for classification feature.

III. METHODLOGY

The following section describes the working principles of the proposed framework and algorithms with an illustration. The illustration and the usage of existing algorithms and proposed algorithm are summarized in the below figure1- the model flow plan. The figure shows the stream of study conducting in fabricating the archetypal.

3.1.1. Datasets are collected from various hospital and patients includes the patient name, age, gender, FBS, PPBS, HbA1C, No of years affected, and the taking of Tablet or Insulin.

3.1.2. The Preprocessing technique is implemented and some of the fields are considered irrelevant and removed e.g. Gender, No of years affected and identity like name is replaced by person1, 2... n.

3.1.3 The classifier algorithms Rule-Based K-Means, Naïve Bayes, SVM, and Decision Tree are applied for the dataset to process the various measures.

3.1.4. The Accuracy, Sensitivity, Specificity and F-measures and ROC are produced by the classifier and cluster algorithms

3.1.5. The measures are compared with proposed algorithm and the results are enlisted or charted.

3.2. Existing algorithms

The predictive analysis algorithm used in Hadoop/Map reduce environment and classified the types of diseases. The Support Vector Machine is used for learning and classification association rules from data in diabetes, healthcare among young and old patients. The SVM solves the problem of interest indirectly, without solving the more difficult problem. The Predictive alarm ranking algorithm measures the blood glucose level to predict the risk of NH with insulin treated diabetes. In this ranking problem we wish not just to accurately predict the pairwise ordering but also preserve the magnitude of the scores or the difference between the ratings. Therefore, regularized solution for regression problem differs from ranking problem solution.

3.3. Proposed algorithm

3.3.1. Rule Based K-Means Algorithm

The process of Rule Based K-Means Algorithm for prediction of hypoglycemia uses available self-monitoring data which was collected from various labs and patients and also association rules generated can be incorporated in self-monitoring devices or applications. It concludes that algorithm offers assessment of the risk of hypoglycemia and can be used for improvement in glycemic control. It is also worth to note that the proposed method indicates that among the variables observed, the presence of the HbA1C, FBS and PPBS was the most strongly associated with type2 diabetes.

In general, Diabetes mellitus is common and serious autoimmune disease characterized by the in ability of the pancreas to produce insulin and an essential hormone needed to convert food into energy, and thus to regulate blood glucose concentration. Several algorithms have been discussed in the literature to predict the method of treatment for the patient. However the proposal is to reduce the complexity of previous methods such as ranking and collaborative filtering method.

In this work, the approach of partitioning of the data space into clusters divides the data into different disjoint groups, each group containing any number dataset. The groups are formed by using Rule Based K Means Algorithm with a marked difference [12][11]. In Rule Based K Means Algorithm, the initial center node is selected arbitrarily; the numbers of groups are not selected and not known in advance [13]. The number of groups is identified based on the way of treatment that is going to suggest the patients. The number of groups in a dataset is two accordingly the center nodes that are identified by selecting one for insulin patient and one for tablet patient. Then the algorithm separates the dataset by measuring not only the squared Euclidean distance but also using the association rules that are framed based on the parameters such as HbA1C, FBS, PPBS, AGE.

3.3.2 Rule Based K-Means Algorithm Steps

Step 1. Identify the two dependent variable from the dataset.

Step 2. Form the scatter diagram based the dependent variables x and y.

Step 3. Label all the data points in the scatter diagram. Assign the number of groups=2.

Step 4.Identify two Initial Center Value (ICV) values randomly based on the mode of treatment (Insulin or Tablet).

Step 5. Determine all the nearest nodes of the ICV using distance formula.

Step 6. By suppression and iteration, select the number of data points in a group based on the lowest distance between the data point and ICV. The data points can be selected for the group until the highest calculated value/2.

Step 7.When all the nodes have been assigned either of the groups, recalculated the passion of ICV values.

Step 8. Derive new ICV.

Step 9. Repeat step 5 to 7 until the ICV reaches fixed position. This produces a separation of the dataset into groups, from which the metric to be minimized can be calculated and goto step 10.

Step 10. The following conditions are used to frame the optimized group from the previous steps.

Association Rules

Rule 1: X (Suggestion=”Insulin”) = (HbA1C>=7) && ((FBS>125) ||(PPBS>105)) &&(AGE>50)

Rule 2: X (Suggestion=”Tablet with WALKING”) = (HbA1C>7) && ((FBS>125) ||(PPBS>105)) ||(AGE<50).

Rule3: X (Suggestion=”Insulin or Tablet with WALKING and Diet”) = (HbA1C<7) && ((FBS<125) ||(PPBS<105)) &&(AGE>50)

Rule4: X (Suggestion=”Tablet with WALKING”) = (HbA1C<7) && ((FBS<125) || (PPBS<105)) && (AGE<50)

Rule 5: X (Suggestion =” Insulin or tablet with WALKING”) = (HbA1C>7) && ((FBS<125) || (PPBS<105)) && (AGE>50)

Description for Association Rules

Rule 1: For age above 50 and HbA1c is greater than or equal to 7 and FBS or PPBS is greater than 125 or 105 respectively then the Suggestion is Insulin.

Rule 2: For age below 50 or FBS is greater than 125 or PPBS is greater than 105 and HbA1c is greater than 7 then the Suggestion is Tablet with Walking.

Rule3: For age above 50 and HbA1c is less than 7 and FBS or PPBS is less than 125 or 105 respectively then the Suggestion is Insulin or Tablet with Walking and Diet.

Rule4: For age below 50 and HbA1c is less than 7 and FBS or PPBS is less than 125 or 105 respectively then the Suggestion is Tablet with Walking.

Rule 5: For age above 50 and HbA1c is above 7 and FBS or PPBS is less than 125 or 105 respectively then the Suggestion is Insulin or Tablet with Walking.

3.3.3. Naive Bayes

NB is a characterization procedure with a thought which characterizes all highlights is independent and random to one another. It characterizes that status of a particular component in a class not ensure influence the status of another element. Since it depends on contingent likelihood it is considered as an amazing calculation utilized for arrangement reason. It functions admirably for the information with unbalancing issues and missing qualities. It is a machine learning classifier which utilizes the Bayes Theorem. Utilizing Bayes hypothesis back likelihood P(C/X) can be determined from P(C), P(X) and P (CX).

3.3.4. Support Vector Machine (SVM)

It is a systematic arrangement of structured machine learning typical is utilized in gathering. Given a two-class getting ready test the purpose of assistance it is to find the best most dumbfounding edge disengaging hyperplane among the two classes. For better theory hyperplane should not to lies closer to the data centers to have a spot with the diverse class. Hyper plane should be picked which is far from the data centers from each grouping. The centers that misrepresentations nearby to the edge of the classifier are the assistance vectors.

3.3.5. Decision Tree (DT)

It is a regulated machine learning calculation used to take care of characterization issues. The primary goal of utilizing DT in this examination effort is the expectation of objective class utilizing choice standard taken from earlier information. It utilizes hubs and internodes for the forecast and order. Root hubs characterize the occasions with various highlights. Root hubs can have at least two branches while the leaf hubs speak to characterization. In each stage, Decision tree picks every hub by assessing the most elevated data gain among every one of the qualities.

3.4. **Accuracy Measures**

Rule Based K-Means Algorithm, Decision Tree, SVM and Naive Bayes algorithms are used in this research work. Investigates are achieved by internal cross-validation 10-folds. Accuracy, Sensitivity, Specificity F-Measure, and ROC processes are regained for the classification of this work. Below Table-3 defines accuracy measures.

3.4.1 Attributes used

The following attributes are used to find the accuracy measures as follows Person, HbA1c, FBS, PPBS, age and Mode of treatment (Insulin/Tablet)

Classifiers formula for Accuracy, Sensitivity, F-Measure, Specificity, and ROC values are listed in Table-2, the expansion of TP is True Positive, TN is True Negative, FP is false positive and FN is False Negative. The comparable classifiers show on the basis of Accuracy, F-measure, Sensitivity, Specificity and ROC values are listed in Table-3

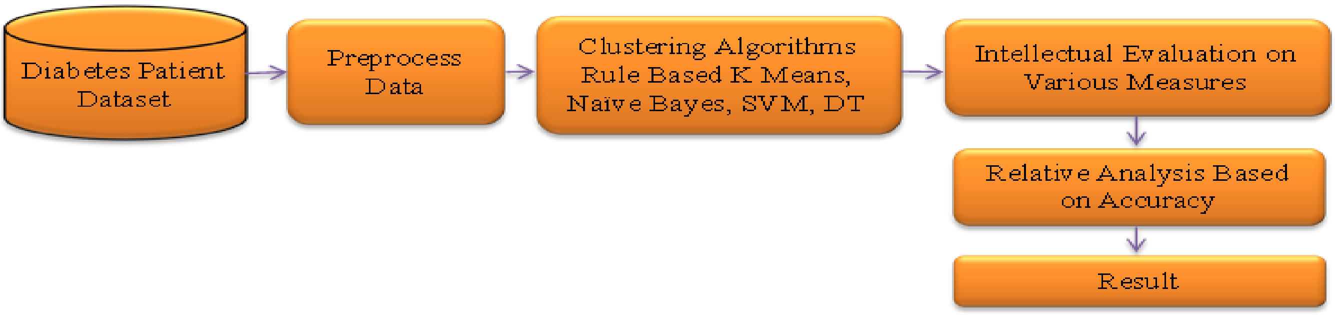
IV.RESULT

Table-3 represents to various execution estimations of all grouping calculations determined on different measures. From Table-3 and 1, it is dissected that Naive Bayes, SVM, and Decision Tree demonstrate the base exactness than the Rule Based K-Means Algorithm. So the Rule Based K-Means Algorithm is conjecturing the conceivable outcomes of diabetes with best accuracy as paralleled to different classifiers for this dataset. Representations of all classifiers dependent on various procedures are structured by means of a diagram in Figure-2 and Figure-3.

4.1. Classifier Performance comparison on Various Measures

In the below figure2, Accuracy measures comparison for four algorithms are Rule Based K-Means algorithm (96%), SVM (93.00%), Naïve Bayes (92.00%) and DT(88.00%) and RB K-Means is better than the others. The Sensitivity measures comparison for four algorithms are Rule Based K-Means algorithm (0.500), Naïve Bayes (0.500%), SVM (0.500%) and DT(0.25%) and first three shares the better Sensitivity than the DT. And the F-measure comparison for four algorithms are Rule Based K-Means algorithm (0.500%), Naïve Bayes (0.200%), SVM (0.39%) and DT(0.250%) and RB K-Means is better than the other three . Also the Specificity measures comparison for four algorithms are Rule Based K-Means algorithm (0.500%), Naïve Bayes (0.125%), SVM (0.390%) and DT(0.125%) and Rule Based K-Means is better than the other three. In the below figure3, ROC area of four algorithms are Rule Based K-Means algorithm (1.0%), SVM (1.0%), Naïve Bayes (0.96%) and DT(0.89%). The minimal is DT which is better than other three.

Figure-1 Process flow Diagram


Diabetes Patient Dataset
Preprocess Data
Clustering Algorithms Rule Based K Means, Naive Bayes, SVM, DT
Intellectual Evaluation on Various Measures
Relative Analysis Based on Accuracy
Result

DOI: 10.35940/ijrte.B1195.0782S419

Table1. Confusion Matrix of several Classification algorithms

	Rule Based K-Means		Naïve Bayes		SVM		Decision Tree	
Tablet	TN =94	FP=2	TN =91	FP=1	TN=91	FP=3	TN=88	FP=3
Insulin	FN=2	TP=2	FN=7	TP=1	FN=4	TP=2	FN=8	TP=1

Table2. Accuracy Procedures

Events	**Descriptions**	**Method**
Accuracy(ACC)	Exactness decides the accuracy of the calculation in foreseeing examples.	ACC=(TP+TN)/(TP+TN+FP+FN)
Sensitivity(SN)	Classifiers rightness/exactness is estimated by Sensitivity.	SN=TP/(TP+FP)
Specificity(SP)	To quantify the classifiers culmination or sensitivity, Recall is utilized.	SP=TP/TP+FN
F-Measure	F-Measure is the weighted normal of precision and recall.	F=2*(P*R)/(P+R)
ROC	ROC (Receiver Operator Curve) arches are utilized to compare the helpfulness of tests.	

Table3. Relative Performance of Classification System on Innumerable Methods

Classifiers Structures	**Sensitivity(SN)**	**Specificity(SP)**	**F-Measure**	**Accuracy(ACC)**	**ROC**
Rule Based K-Means Algorithm	0.500	0.500	0.500	96.70	1.00
SVM	0.500	0.333	0.391	93.00	1.00
Naive Bayes	0.500	0.125	0.200	92.00	0.96
Decision Tree	0.250	0.250	0.250	88.00	0.89

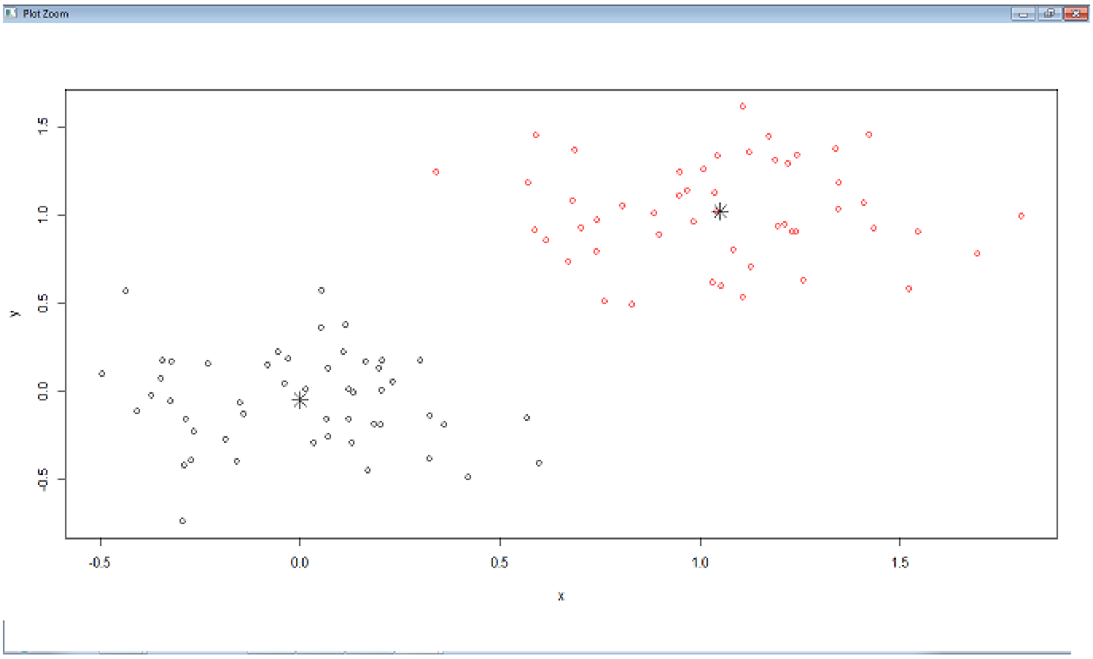


Figure-2 Scatter plot diagram for Rule Based K-means

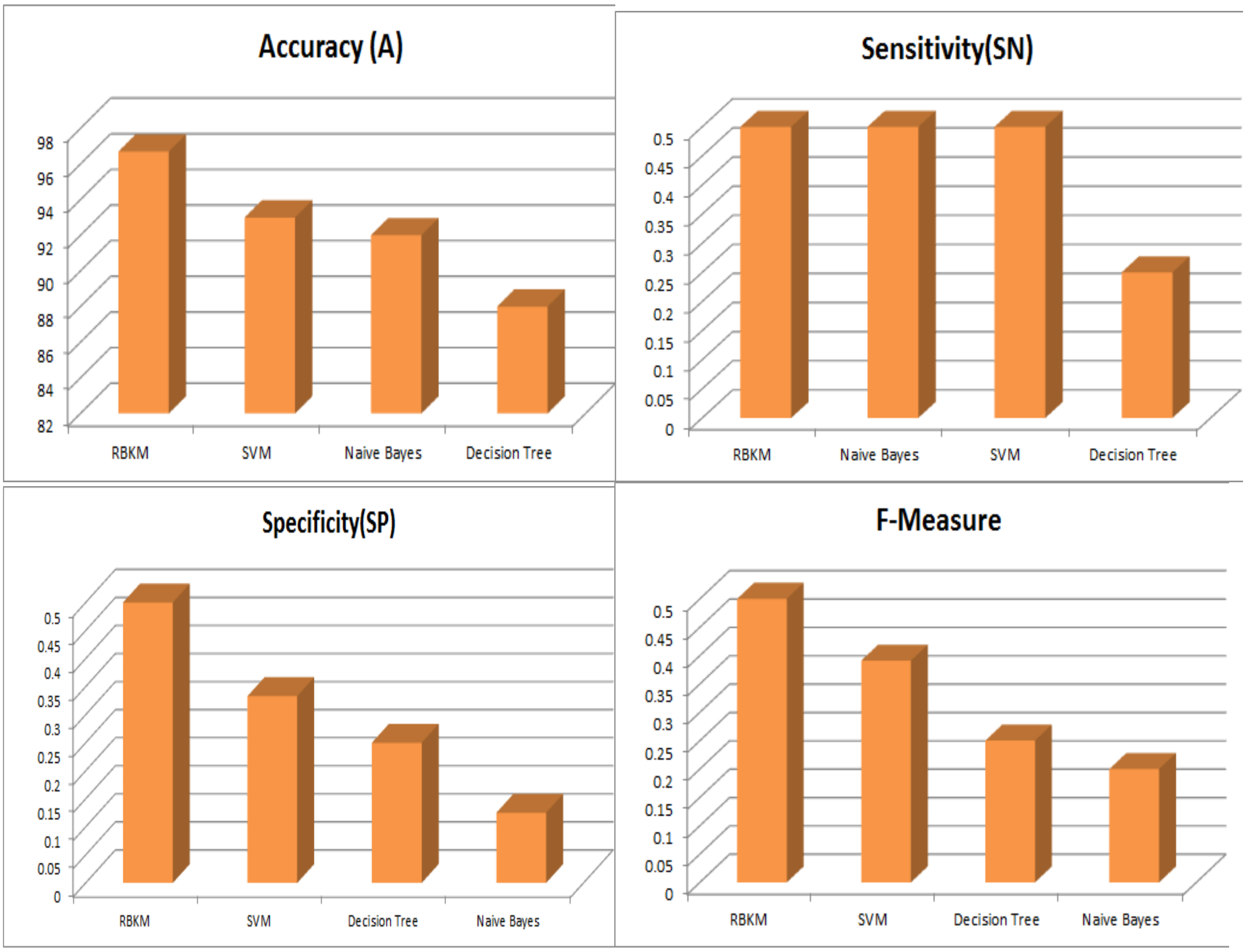


Figure-3. Classifier Performance valuation on many events

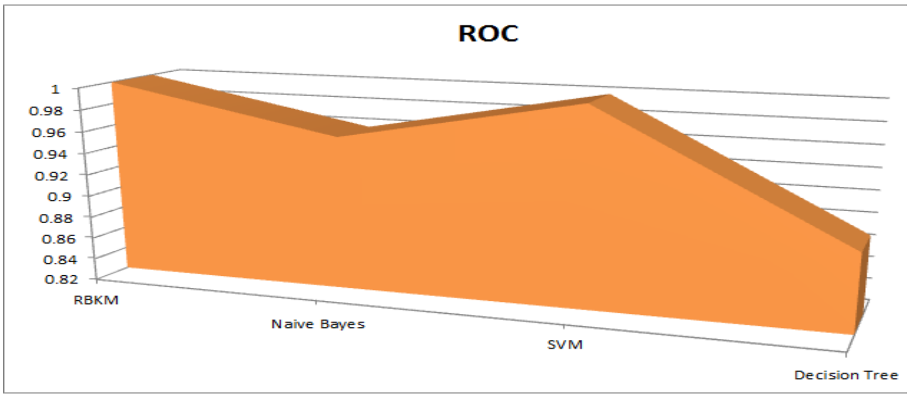


Figure-4. Receiver Operator Curve Area of all several Algorithms results.

V.CONCULSION

One of the normal and significant genuine medicinal issues is the conclusion of diabetes method of treatment. In this investigation, orderly endeavors are made in planning a framework which results in the forecast method of treatment like Insulin or Tablet. During this work, our proposed models are portrayed and contrasted and three machine learning classification algorithms calculations with different measures. Analyses are performed on clinical Diabetes Database. Exploratory outcomes decide the adequacy of the structured framework with an accomplished exactness of 96.0 % utilizing the Rule Based K-Means algorithm calculation which yields best execution than the Naïve Bayes, SVM, and Decision Tree calculation. In future the structured framework will be improved and further better than any similar classification algorithms considered.

VI. REFERENCES:

[1]. Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification

algorithms." Procedia computer science 132 (2018): 1578-1585.

[2]. Krishnamoorthy.P, R.Gobinath, “Preprocessing And Feature Extraction In Clinical Decision Support System For Diabetic Patient”, International Journal of Mechanical and Production Engineering Research and Development (IJMPERD) ISSN (P): 2249-6890; ISSN (E): 2249-8001 Vol. 8, Special Issue 3, Dec 2018, 180-191.

[3]. Negi, Anjli, and Varun Jaiswal. "A first attempt to develop a diabetes prediction method based on different global datasets." 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE, 2016.

[4] Phan, Anh Viet, Minh Le Nguyen, and Lam Thu Bui. "Feature weighting and SVM parameters optimization based on genetic algorithms for classification problems." Applied Intelligence46.2 (2017): 455-469.

[5]. Orabi, Karim M., Yasser M. Kamal, and Thanaa M. Rabah. "Early predictive system for diabetes mellitus disease." Industrial Conference on Data Mining. Springer, Cham, 2016.

[6]. Perveen, Sajida, et al. "Performance analysis of data mining classification techniques to predict diabetes." Procedia Computer Science 82 (2016): 115-121.

[7]. Nai-Arun, Nongyao, and Punnee Sittidech. "Ensemble Learning Model for Diabetes Classification." Advanced Materials Research. Vol. 931. Trans Tech Publications, 2014.

[8]. Iyer, Aiswarya, S. Jeyalatha, and Ronak Sumbaly. "Diagnosis of diabetes using classification mining techniques." arXiv preprint arXiv:1502.03774 (2015).

[9] Han, Jianchao, Juan C. Rodriguez, and Mohsen Beheshti. "Discovering decision tree based diabetes prediction model." International Conference on Advanced Software Engineering and Its Applications. Springer, Berlin, Heidelberg, 2008.

[10].S.Umarani, Dr.PavaiMadheswari S, Dr. NagarajanN,“A New Recovery Scheme for Single and Multiple Link Failures in Crossbar Networks”,International Journal of Soft Computing 8(6),416-423,2013.

[11]. Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011

[12] Bishnu, Partha S., and Vandana Bhattacherjee. "Software fault prediction using quad tree-based k-means clustering algorithm." IEEE Transactions on knowledge and data engineering 24.6 (2011): 1146-1150.

[13]. Rish, Irina. "An empirical study of the naive Bayes classifier." IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. 2001.

[14]. Sisodia, Deepti, Shailendra Kumar Shrivastava, and R. C. Jain. "ISVM for face recognition." 2010 International Conference on Computational Intelligence and Communication Networks. IEEE, 2010.

[15]. Sisodia, Deepti, Lokesh Singh, and Sheetal Sisodia. "Fast and Accurate Face Recognition Using SVM and DCT." Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012. Springer, New Delhi, 2014.

[16]. Pradhan, M. A., et al. "A genetic programming approach for detection of diabetes." Int J Comput Eng Res (ijceronline. com) 2.6 (2012): 91.

[17]. Rashid, Tarik A., Saman M. Abdullah, and Rezhna Mirza Abdullah. "An intelligent approach for diabetes classification, prediction and description." Innovations in Bio-Inspired Computing and Applications. Springer, Cham, 2016. 323-335.

[18]. Nai-arun, Nongyao, and Rungruttikarn Moungmai. "Comparison of classifiers for the risk of diabetes prediction." Procedia Computer Science 69 (2015): 132-142.

[19]. Kumar, D. Ashok, and R. Govindasamy. "Performance and evaluation of classification data mining techniques in diabetes." International Journal of Computer Science and Information Technologies 6.2 (2015): 1312-1319.

[20]. Vijayan, V. Veena, and C. Anjali. "Prediction and diagnosis of diabetes mellitus—A machine learning approach." 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS). IEEE, 2015.

[21]. Kavakiotis, Ioannis, et al. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal 15 (2017): 104-116.

[22]. world health organization in https://www.who.int/