Bulletin of Electrical Engineering and Informatics
Vol. 9, No. 3, June 2020, pp. 877~885
ISSN: 2302-9285, DOI: 10.11591/eei.v9i3.1744

An efficient feature selection algorithm for health care data analysis

Mythily R.[1], Aisha Banu W.[2], Dinesh Mavaluru[3]
[1]Department of Information Technology, B. S. Abdur Rahman Crescent Institute of Science and Technology, India
[2]Department of Computer Science and Engineering, B. S. Abdur Rahman Crescent Institute of Science and Technology, India
[3]Department of Information Technology, Saudi Electronic University, Saudi Arabia

Article Info

Article history:

Received Aug 24, 2019
Revised Nov 12, 2019
Accepted Feb 27, 2020

Keywords:

Advanced clustering
Chi-square
Clustering
Fuzzy
M5

ABSTRACT

Diabete is a silent killer, which will slowly kill the person if it goes undetected. The existing system which uses F-score method and K-means clustering of checking whether a person has diabetes or not are 100% accurate, and anything which isn't a 100% is not acceptable in the medical field, as it could cost the lives of many people. Our proposed system aims at using some of the best features of the existing algorithms to predict diabetes, and combine these and based on these features; This research work turns them into a novel algorithm, which will be 100% accurate in its prediction. With the surge in technological advancements, we can use data mining to predict when a person would be diagnosed with diabetes. Specifically, we analyze the best features of chi-square algorithm and advanced clustering algorithm (ACA). This research work is done using the Pima Indian Diabetes dataset provided by National Institutes of Diabetes and Digestive and Kidney Diseases. Using classification theorems and methods we can consider different factors like age, BMI, blood pressure and the importance given to these attributes overall, and singles these attributes out, and use them for the prediction of diabetes.

Corresponding Author:

Mythily R.,
Department of Information Technology,
B. S. Abdur Rahman Crescent Institute of Science and Technology,
Chennai, India.
Email: mythily@crescent.education

1. INTRODUCTION

As a big nation in Asia, Indonesia keeps developing various fields to keep up with the world's more onset of diseases. Healthcare scenarios have massive data sets and contain values that are not always needed to get the result we desire, to select the features that are required, we use feature selection to get the values that will affect the output of the prediction most and work with those to provide the maximum accuracy possible. Our project will specifically focus on diabetes [1-7].

Data mining is an approach toward dealing with extensive data sets to make out designs and make up connections to think about issues in the course of data examination. Data mining allow undertakings to anticipate potential patterns. In mining, membership rules are through by investigating data for a visit if/at that point designs, at that position utilizing the help and sureness criterion to discover the most critical connections inside the data. Support is the way by which as often as possible the things confirm up in the database, while certainty is the circumstances if/at that point articulations are exact. A classification technique searches for new examples and may carry about a modification in the way the data is composed.

Bunching parameters find and outwardly archive gatherings of actualities that were beforehand obscure. Bunching bunches an arrangement of items and totals them because they are so like each other. There are diverse ways a client can actualize the group, which separate between each grouping model. Cultivating parameters inside data mining can find designs in data that can prompt sensible forecasts about the future, otherwise called prescient examination [8, 9].

Data pre-processing is a method used in data mining to format and organize the data in a way that can be used by the algorithm, often there are a lot of anomalies in raw data, and these can cause incorrect data problems and hence, we clear up such anomalies, like incomplete data or error values. This process is called pre-processing, and the result can use for additional applications. Several steps during pre-processing are:

- Data cleaning: Data is cleaned, as in the missing values and such are filled with their mean values and sometimes the aggregate or neighboring values.
- Data integration: Data is formatted to use with the algorithm.
- Data transformation: Here generally data is aggregated, generalized and normalized.
- Data reduction: Usually data warehouses have large amounts of data which are not needed, so the data reduced to the ones that required by the algorithm.
- Data discretization: It involves the decrease of some values of a continuous attribute by separating the range of attribute intervals.

Feature selection alludes to the way toward diminishing the contributions for handling and examination, or of finding the most significant information sources. A related term, feature building (or feature extraction), alludes to the way toward extricating valuable data or features from existing data. Feature selection is constantly performed before the model prepared. With a few algorithms, feature selection systems are "inherent" so insignificant sections are prohibited and the best features consequently found. Every algorithm has its particular arrangement of default procedures for astutely applying feature decrease. Be that as it may, you can likewise physically set parameters to impact feature selection conduct. Amid programmed feature selection, a score is ascertained for each characteristic and just the qualities that have the best scores chosen for the model. You can likewise alter the edge for the best scores. SQL server data mining gives various strategies to ascertaining these scores, and the correct technique that connected in any model relies upon these variables:

- The algorithm that has integrated into your model
- The data type of the attributes used in the model
- Any parameters set in your model

The advanced clustering algorithm (ACA) technique abstains from registering the separation of every datum protest the group recursively and spares the execution time. ACA requires a straightforward data structure to store data in every cycle, which is to utilize as a part of the following emphasis. The test comes about to demonstrate that the ACA strategy can adequately enhance the speed of clustering and precision, decreasing the many-sided computational quality of the conventional algorithm Kohonen SOM. This paper incorporates ACA, and its reproduced trial comes about with various datasets. Clustering is the way toward sorting out data objects into an arrangement of different classes called clusters. Grouping is an unsupervised method of classification. In the unsupervised procedure, the right answers not known ahead of time. Compromise is a system that doles out data articles to an agreement of classes. A few clustering algorithms had proposed to date. However, every one of them utilized for some particular prerequisite. There does not have a single algorithm that can viably deal with a wide range of necessity. This makes a massive test for the client to choose one among the accessible algorithm for particular purposes. To manage this issues, another algorithm has been proposed in this examination that named as "advanced clustering algorithm" [10-15].

To predict diabetes, by using an efficient feature selection algorithm, this is more accurate than existing algorithms. This project will be of use in the medical field, to predict diabetes more accurately. Most algorithms in use now, are only about 65-70% accurate, at max, which is unacceptable in the medical field. This paper aims at replacing it, by an algorithm which will increase that accuracy.

2. LITERATURE SURVEY

In [16], a comparison of Naïve Bayes, Logistic Regression, J48, and Random Forest algorithms. Naïve Bayes classifier being the simplest classifier has performed well with an accuracy of 76.52% while having relative absolute error 59.56%. Only four algorithms have been compared, which is very minimal. Here, inspection of the execution of learning algorithms namely Naïve Bayes, J48, Random forest, and Logistic Regression to predict the population who are most likely to develop diabetes on Pima Indian diabetes data. The performance depth is verified regarding MAE and NRMSE obtain from the test set and the

results are marginally better quantitatively regarding accuracy and prediction capability. Additional, we plan to recreate the learning of Classification models by introducing the intelligent machine learning algorithms useful to an extensive collection of actual life data set.

In this investigation work [17], discussed that the as frequently as conceivable used portrayal procedures J48, CART, SVMs, and kNN are inspected, on the restorative dataset to find the perfect response for diabetes. The execution pointers precision, specificity, affectability, exactness, bumble rate are registered for the given dataset. Assertion beside with a real data pre-getting ready system can hint at change the precision of the classifier. The limit of data institutionalization had a discernible impact on arrangement execution and astonishingly redesigned the execution of J48. The implementation of the kNN algorithm has minimum exactness. In perspective of the parameters taken for examination, the presentations of the four algorithms are researched. The results exhibit that the execution of the J48 framework is on a fundamental level superior to the next three techniques for the gathering of diabetes data. To upgrade the general accuracy, it is imperative to use a more enlightening list with an immense number of attributes and use the best part assurance system in the future. Future works may in like manner consolidate blend course of action models by joining a bit of the data mining methodologies.

In [18], worked with diabetes prediction by using decision tree and a Naive Bayes model. In this paper, they proposed a more efficient technique for diagnosing the disease compared to the existing system. It is only up to 95% accurate, in prediction. Most of the females are more affected by diabetes with over 246 million individuals.Based on WHO report; by 2025 this number is expected to rise to over 380 million. There is no proper cure for diabetes disease in the world. Through technology; the healthcare sector is improving, as well as symptoms for diabetes are well documented.

The most noteworthy aspect of this study by [19], is to concern data mining knowledge for predicting diabetes. We performed a pre-processing step to compact with a dataset like a feature selection method, normalization and worked with a machine-learning technique such as SVM. The primary endeavor of feature selection is to shrink the dimensions by picking the majority features based on some statistical score. They proposed F-score and achieved improved performance classification compared to other feature selection methods like relief and relief filtering methods. Then the performance of SVM classifier is evaluated in the expression of accuracy, sensitivity, specificity, and AUC. They improved feature selection and normalization of the data performance of SVM classifier. We applied different functions of SVM on various datasets having the same number of features, but with different values then we get the difference in accuracy in diabetes classification.

In this investigation work by [20], the examination of SOM and ACA algorithm in perspective of different points. This can be used as a piece of occasions of significant data sets, which are for the most part exceedingly dimensional educational accumulations. There are only two algorithms which have been examined, which decrease the variety of analysis.SOM algorithm is a standard clustering algorithm, and it is widely used for clustering massive courses of action of data. In this paper reveal ACA and examinations the insufficiencies of the SOM algorithm. Since the versatile computational nature of the SOM algorithm is unpleasantly high inferable from the need to reassign the data centers different conditions in the midst of each cycle, which impacts its adequacy. This paper displays a direct and compelling technique for doling out data centers to packs. The proposed strategy ACA. This paper ensures the whole method of clustering in O (nk) time without giving up the precision of gatherings. The exploratory outcome shows the ACA can get better execution time, nature of SOM algorithm and capacities splendidly on high dimensional instructive list. So the proposed system is conceivable.

3. METHODOLOGY

M5 algorithm is the proposed feature selection algorithm, which aims at improving the chi-square algorithm, using the techniques of clustering, to increase the accuracy of predicting Diabetes, using a specific number of attributes [21-25]. The problem with the existing system is that the accuracy is not up to the mark, and is not always reliable when it comes to the medical field, where every decision is critical. Another problem is that the existing system does not make use of all the data available to it, and therefore there is room for error, due to ignoring specific data, whereas in the proposed system, using clustering, all the collected data is included and taken into consideration as shown in Figure 1.

This system has five modules: preprocessing, clustering, application of chi-square, classification using SVM, and prediction. Preprocessing clears the unwanted and noisy data, while clustering group similar data, and creates clusters which will appear as the output of related data. The application of chi-square is then made to retrieve essential variables, and the classification is done, using which the process of prediction takes place. The existing system which uses F-score method and K-means clustering of checking whether a person has diabetes or not are 100% accurate, and anything which isn't a 100% is not acceptable in the medical field,

as it could cost the lives of many people. We applied special functions of SVM on a variety of datasets having the same no. of features, but with different values then we get variation in accuracy in diabetes classification. Disadvantage: In the existing system, we have performed well with an accuracy of only an average value (around 60%), which is not worth taking a risk on, in the field of medicine.

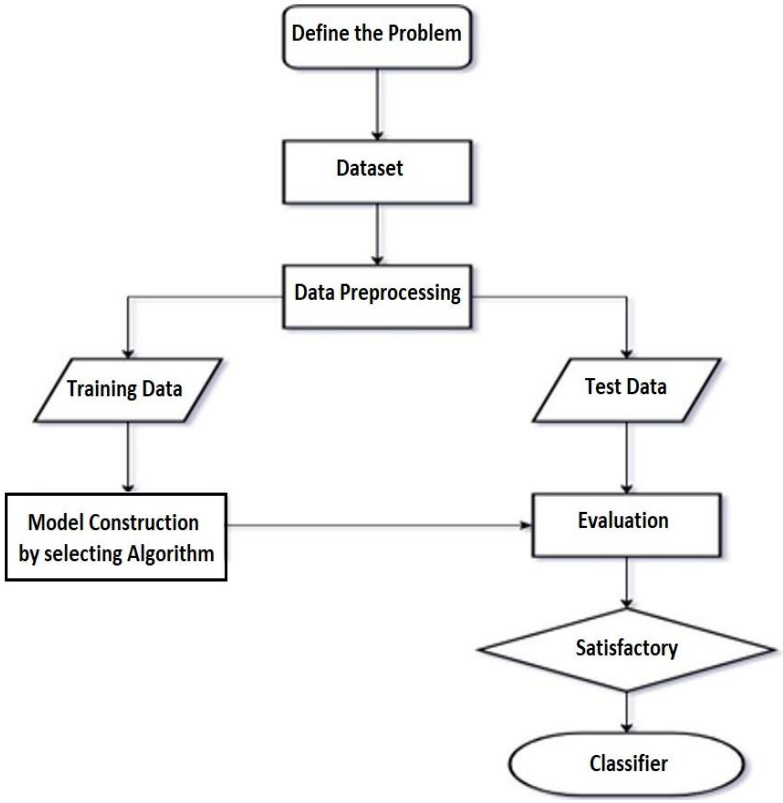

Figure 1. Flowchart of method

3.1. Proposed methodology

We analyze the existing algorithms and use some of the best features of the current algorithms and based on these features; we aim to turn them into a novel algorithm, which will be 100% accurate in its prediction. Our proposed system aims at using some of the best features of the existing algorithms to predict diabetes, and combine these and based on these features; we seek to turn them into a novel algorithm, which will be 100% accurate in its prediction. With the surge in technological advancements, we can use data mining to predict when a person would be diagnosed with diabetes. Specifically, we analyze the best features of chi-square algorithm and ACA. Advantage: the proposed system performs clustering before selecting a feature so that all the items are grouped in advance, in turn improving the overall accuracy. The details of the internal working of the algorithms like architecture diagram and the architecture of the proposed algorithm, "M5 Algorithm". It also has the modules used and how the modules work with the data. The dataset is processed to remove any anomalies in data like incomplete data, clear data. Then the data goes through chi Sq., fuzzy C means, and M5 algorithm. The data obtained from these modules will be used to form an SVM Classified tree. The test dataset is used with the tree to get a prediction. The accuracy and the time taken will be the final results of the forecast. Figure 2 shown algorithm that explains the procedure.

a. Import libraries.
   - Lib for chi.
   - Lib for the dataset.
   - Lib for SVM.
b. Import data. Display Data.
c. Initialize the fuzzy set function with parameters. (Function wrote below)
d. Apply Chi Sq. The algorithm to the dataset.
e. Retrieve the first five values and store it in 'weight

f. Print weights.
g. Use the result to train dataset and store in ‘train.’
h. Initialize the test dataset into ‘test.’
i. Apply the Chi Sq to ‘train.’
j. Use SVM to plot the model for the trained dataset.
k. Predict from the results of SVM tree using Test model.
l. Print.

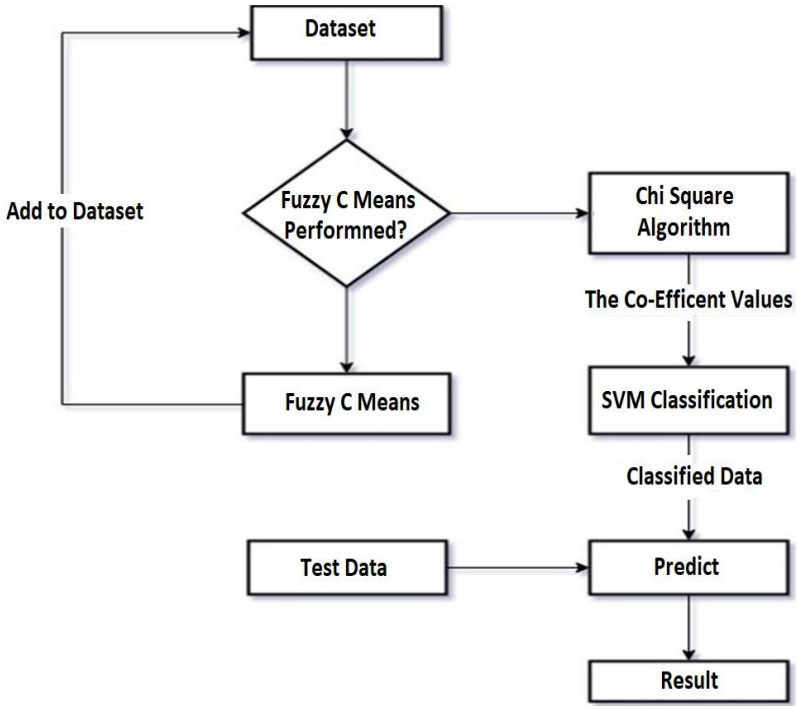

Figure 2. Flow chart of classification

3.2. Preprocessin

The cleaning of the bulk data by using different methods:

- Input: Dataset with incomplete, noisy, inconsistent data.
- Output: The output of preprocessing is adequately cleaned data.

3.3. Clustering

In Fuzzy c-means (FCM) clustering method dataset is assembled into n number of clusters with each datum spot in the data having a place with each group to a certain degree. A data point that is close to the center of a cluster having a high level of having a position or participation with that cluster and another data point that lies far from the center of a group will have a low level of having a place or enrollment with that cluster. C-Means algorithm is a clustering algorithm where each thing may have a home with in excess of one gathering (hence the word `fuzzy'), where the level of enrollment for each item is given by a likelihood circulation over the clusters. What makes FCM diverse is that it doesn't decide the supreme participation of a data point to a given group; instead it calculates the probability (the level of enrollment) that a data point will have a place with that cluster.

Input: Preprocessed data
Output: Cluster values
Procedure: Enter the number of clusters and number of data points.
Create a membership matrix, which consists of a combination of clusters and data points.
The condition specified, i.e., whether all the datasets clustered or not. Will be checked, and if it does satisfy the state, it stops there.
If the condition is not satisfied, then the centroid for each of the clusters is computed.
Finally, the membership coefficient calculated for each point, and then the condition is rechecked.

3.4. Chi-square algorithm

The chi-squared test refers to a class of algebraic tests in which the distribution is a chi-square distribution. When used without further qualification, the term frequently refers to Pearson's chi-squared test, which is used to check whether an observed value might have arisen from an expected value (under some assumption), or whether that assumption is supposed to be wrong. Habitually, the chi-squared test is used to test for independence between two data sets. For instance, in a survey conducted in which the ages of participants are recorded, a chi-squared test can be used to determine if age affects the survey responses, or if the two are independent (since in this case, one would expect the answers to be roughly equivalent across all age groups). The test is also commonly used to test if a population follows a specific distribution; for instance, the test can be used to determine whether a die is fair, or whether a city has an equal number of men and women.

$$x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

where: O is the observed value and E is the expected value.

3.5. SVM classifier

Support vector machine is a supervised machine learning algorithm which can be used for classification. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Here, the sample data is first encoded into the parameter of SVM, and the training dataset is used to calculate the fitness function. If it satisfies the convergent condition, then the SVM model is created. If not, then again, the process of selection, crossover, and mutation is done and then the new fitness function is calculated. After this, once again the condition is checked, and if so, the SVM model is created.

4. EXPERIMENT AND RESULT

Isolating data into preparing and testing sets is an essential piece of assessing data mining models. Typically, when you separate a data set into a preparation set and testing set, the vast majority of the information is utilized for preparing, and a little bit of the information is being used for testing. Investigation Services haphazardly tests the data to help guarantee that the testing and preparing sets are comparative. By using comparative data for preparing and testing, you can limit the effects of data discrepancies and better comprehend the characteristics of the model. After a model has been processed by utilizing the preparation set, you test the model by influencing predictions against the test to set. Because the data in the testing set as of now contains known qualities for ascribe that you need to predict, it is anything but difficult to decide if the model's estimates are correct. According to Table 1, there are 9 attributes (pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes predigree function, age and outcome) which all of them are numerical except outcome.

Table 1. Dataset table

Serial Number	Attributes	Type
1	Pregnancies	Numerical
2	Glucose	Numerical
3	Blood Pressure	Numerical
4	Skin Thickness	Numerical
5	Insulin	Numerical
6	BMI	Numerical
7	Diabetes Pedigree Function	Numerical
8	Age	Numerical
9	Outcome	Categorial

As it can be seen in Table 2, the accuracy of the chi square, Glmnet and M5 algorithms are 72, 73, and 77 respectively. True positives (TP)-values have predicted as correct, and the estimation is categorized as yes, which is a true stamen as the actual value is also a yes. True negatives (TN)-Here values estimated and predicted in the no category and correctly correspond with the actual class, which was also a no. False positives and negatives, qualities happen once actual class contradicts with the predicted level. False positives (FP)-these values that are predicted wrong, for instance, the concrete class is no, but the anticipated group is yes. False negatives (FN)-the real data value yes but the predicted group is no, such

values are false negatives. Accuracy-Accuracy is the percent at which the prediction matches the actual values, the more is, the better as the forecast is going in the right direction. For our model, we have 0.803 that represents approximately 80% accurate.

Accuracy=TP+TN/TP+FP+FN+TN

Precision value-it is a proportion of the predicted values with the actual values and how often it corresponds correctly. High precision identifies with the low false positive rate. We achieved a precision accuracy of 0.78 which is relatively high. Precision=TP/TP+FP Recall (Sensitivity)-Proportion of correctly predicted values with the positive values in the class, as the sensitivity of the model lies with the positive values.

Recall=TP/TP+FN

Table 2. Comparison table

Algorithms	Precision	Accuracy	Recall
Chi Square	52	72	71
Glmnet	46	73	67
M5	50	77	66

The Figure 3 is the comparison of accuracy with different algorithms for the dataset. It shows that the accuracy value of M5 is much higher than the other two algorithms. This increase in efficiency is due to the cluster values obtained by using fuzzy c means algorithm. The above graph shows the number of attributes selected for each algorithm. The chi-square and glmnet select five attributes among the available nine attributes. While the M5 algorithm selects six attributes where the sixth attribute is the cluster values of each observation.

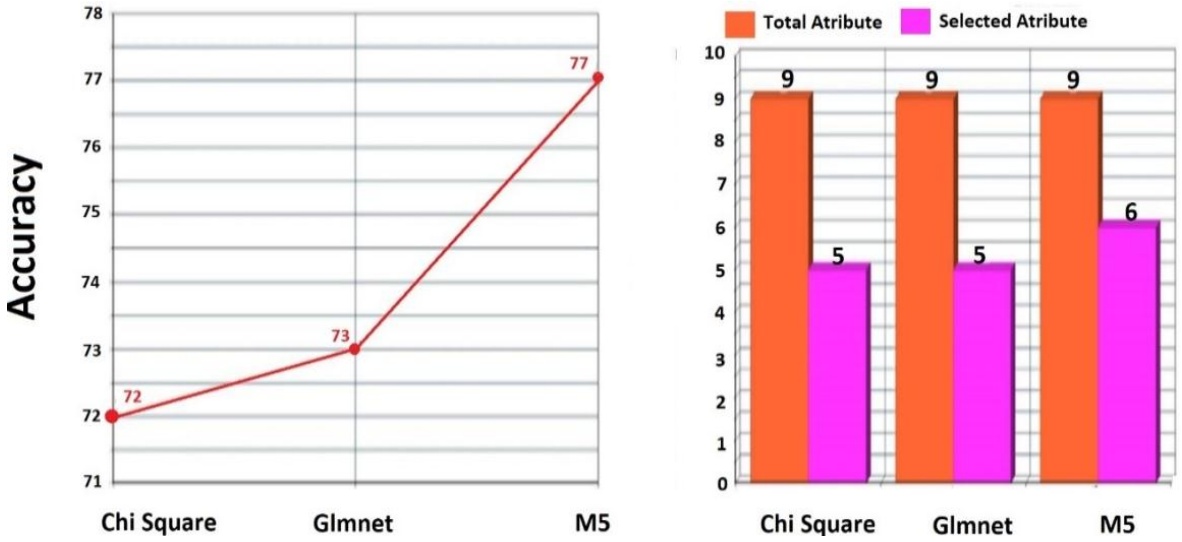

Figure 3. Comparison of accuracy with different algorithms for the dataset

The Figure 4 shows SVM has classified the positive and negative regions for diabetes based on the combined values of glucose and mass from the dataset. The graph shows how SVM has classified the positive and negative regions for diabetes based on the combined values of diabetes and age from the dataset. The graph shows how SVM has classified the positive and negative regions for diabetes based on the combined values of diabetes and mass from the dataset.

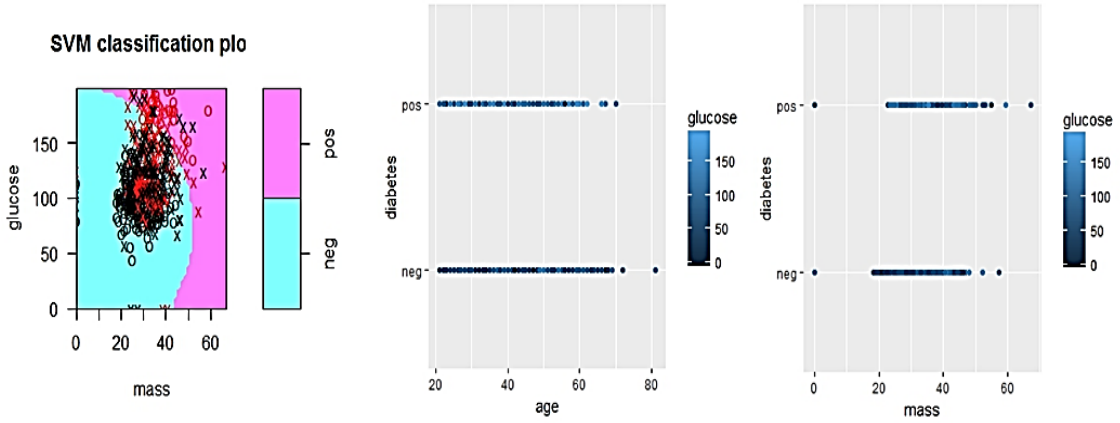

Figure 4. Existing system

5. CONCLUSION

The overall prediction accuracy has been improved, when compared to the previous systems. This process was done by testing five existing algorithms, and using which the M5 algorithm was created based on clustering techniques. Data mining allows the medical professionals with this proper information to intervene earlier, and hence prevent unnecessary deaths. The concept of data aggregation will surely help in the improvement of accuracy. This algorithm can be implemented in healthcare analysis systems, especially in the analysis of diabetes. Future Scope for this research work will be the accuracy of prediction will be improved furthermore. A dedicated user interface will be developed to make sure that the user has a better understanding of how it works. Even the number of datasets will be consumed, and also of a larger size for each of these datasets. Intense testing of each of the modules will be done. The framework of this system will be looked to be used in other areas of application.

REFERENCES

[1] G. Sonam Kumari and N. Vaishali, "Role of competency mapping in indian companies," *International Research Journal of Human Resource and Social Science*, vol. 2, no. 10, pp. 1-10, 2005.
[2] M. Hitt, *et al.*, "Strategic management: Competitiveness and globalization," *Cengage Southwestern Publishing Co.*, 2013.
[3] W. G. Zikmud, et al., "Business research methods," 8th Editions. *South-Western Cengage Learning*, 2010.
[4] S. Atkinson and P. W. Wilson, "Comparing mean efficiency and productivity scores from small samples: a bootstrap methodology," *J. Productivity Anal*, vol. 6, no. 2, pp. 137-152, 2012.
[5] S. Mulyani and S. Fettry, "The influence of Audit committee composition, authority, resource and diligence toward financial reporting quality," *International Business Management*, vol. 10, no. 9, pp. 1756-1767, 2016.
[6] M. Sri, and E. Kasim, "The Influence of business strategy and top management support on the effective of management accounting information system and its impact on corporate performance: Evidence from Indonesia," *IBIMA Confence*, 2017.
[7] M. Ding, et al., "Innovation and marketing in the pharmaceutical industry," *Springer New York Heidelberg Dordrecht London*, 2014.
[8] A. A. Ahmadi, et al., "Information technology; a facilitator for improving dynamic capabilities through knowledge management utilization," *UCT Journal of Management and Accounting Studies*, vol. 2, no.2, pp. 27-36, 2014.
[9] A. S. Jeyalatha and R. Sumbaly, "Diagnosis of diabetes using classification mining techniques," *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol. 5, no. 1, pp. 1-14, 2015.
[10] M. Nabi, et al., "Performance analysis of classification algorithms in predicting diabetes," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 3, 2017.
[11] K. K. Gandhi and P. N. Prajapati, "Diabetes prediction using feature selection and classification," *International Journal of Advanced Engineering and Research Development (IJAERD)*, vol. 1, 2014.
[12] K. Saravananathan and T. Velmurugan, "Analyzing diabetic data using classification algorithms in data mining," *Indian Journal of Science and Technology*, vol. 9, no. 49, pp. 2-6, 2016.
[13] M. Rizwan, et al., "Ideology and politics of Jamiat Ulema-i-Islam (1947-1973)," *Global Social Sciences Review*, vol. 3, no.1, pp. 45-56, 2018.
[14] S. Rauf, *et al.*, "Impact of electronic media on Pakistan's security," *Global Soc. Sci. Rev*, vol. 3, no. 1, pp. 434-446, 2018.
[15] Y. Pourasad, et al., "Design of an optimal active stabilizer mechanism for enhancing vehicle rolling resistance," *Journal of Central South University*, vol. 23, no. 5, pp.1142-1151, 2016.
[16] N. Shah, et al., "Failure in the english subject in government high schools for boys in District Mardan, Khyber Pakhtunkhwa Pakistan," *Global Social Sciences Review*, vol. 3, no. 2, pp. 146-158, 2018.

[17] M. Saleem, et al., "Wh-movement pattern in the spoken discourse of teachers a syntactic analysis," *Global Social Sciences Review*, vol. 3, no. 2, pp. 400-420, 2018.
[18] S. Panigrahi and A. Thakur, "Modeling and simulation of three phases cascaded H-bridge grid-tied PV inverter," *Bulletin of Electrical Engineering and Informatics*, vol 8, no. 1, pp 1-9, 2019.
[19] M. Nawir, et al., "Effective and efficient network anomaly detection system using machine learning algorithm," *Bulletin of Electrical Engineering and Informatics*, vol 8, no. 1, pp 46-51, 2019.
[20] E. Zineb, et al., "The impact of SCRM strategies on supply chain resilience: A quantitative study in the Moroccan manufacturing industry," *International Journal of Supply Chain management*, vol. 6, no. 4, pp. 70-76, 2017.
[21] S. R. Tajuddin, et al., "Analysis and design of directive antenna using frequency selective surface superstrate," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 2, pp. 529-536, 2019.
[22] N. Abdul Malik, et al., "Investigation of lower limb's muscles activity during performance of salat between two age groups," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 2, pp. 608-617, 2019.
[23] M.S. M. Gismalla and M.F. L. Abdullah, "Performance evaluation of optical attocells configuration in an indoor visible light communication," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 2, pp. 668-676, 2019.
[24] H. Wang, et al., "The impact of HVDC links on transmission system collapse," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 16, no. 1, pp. 21-31, 2018.
[25] G. Sinha, et al., "A comparative strategy using PI & fuzzy controller for optimization of power quality control," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 16, no. 1, pp. 118-124, 2018.