Volume: 11 Issue: 01 Pages: 4142-4149(2019) ISSN: 0975-0290

Diabetes Disease Detection through Data Mining Techniques

Amira Hassan Abed
Department of Information Systems center
Egyptian Organization for Standardization & Quality, Egypt
mirahassan61286@gmail.com
Mona Nasr
Department of Information Systems, Faculty of Computers & Information
Helwan University, Egypt
m.nasr@helwan.edu.eg

ABSTRACT

Diabetes is a inveterate defect and disturbance resulted from metabolic conk out in carbohydrate metabolism thus it has occupied a globally serious health problem. In general, the detection of diabetes in early stages can greatly has significant impact on the diabetic patients treatment in which lead to drive out its relevant side effects.
Machine learning is an emerging technology that providing high importance prognosis and a deeper understanding for different clustering of diseases such as diabetes. And because there is a lack of effective analysis tools to discover hidden relationships and trends in data, so Health information technology has emerged as a new technology in health care sector in a short period by utilizing Business Intelligence 'BI' which is a data-driven Decision Support System.
In this study, we proposed a high precision diagnostic analysis by using k-means clustering technique. In the first stage, noisy, uncertain and inconsistent data was detected and removed from data set through the preprocessing to prepare date to implement a clustering model. Then, we apply k-means technique on community health diabetes related indicators data set to cluster diabetic patients from healthy one with high accuracy and reliability results.

Keywords – **Business Intelligence, Health Care, Data Mining, Data-Driven Decision Support System**.

Date of Submission: June 19, 2019 Date of Acceptance: July 18, 2019

I. INTRODUCTION

By 2040, researchers and statisticians are expected that about 642 million adults (1 in 10 adults) will have diabetes. moreover, 46.5% of those diabetic adults have not been diagnosed [1]. In order to reduce this high numbers of deaths according to diabetes, it is important for providing many advanced methods and techniques that will help efficiently in diagnosis of diabetes in early stages and be devised, because a large number of deaths between diabetic patients are resulted from the late in diagnosis of diabetes. In order to develop and implement an advanced techniques for the early diagnosis of diabetes, we extensively need to utilize sophisticated information technology solutions, Business Intelligence and data mining is a suitable IT tools for this situations.

Business Intelligence (BI) and Data mining techniques have a critical role in the medical and the healthcare sectors depending on the Patient Electronic Health Record thanks to that the Business intelligence is considered a broad category of methodologies, solutions, and applications for capturing, collecting, maintaining, analyzing, and providing easily data access to help users in making successful and faster decisions. it also include various activities and functions of decision support systems such as querying and reporting, online analytical processing 'OLAP', statistical analysis, forecasting, and data and text mining.

Business intelligence solutions are used in many industries to gain significant insight from different data sources to help business executives in making more informed decisions towards main goal to be achieved in efficient and effective manners. Also, it is shown as a mechanism to emphases on a robust and systematic methods to successful healthcare management with a goal of ensuring its great impact on a quality improvement and cost control. **[2]** BI tools are technology that efficiently supports the business operation by providing an interesting value to the enterprise-wide information and thus the way this information is used. **[3]** Some of the key characteristic and features that has been introduced through many researches for BI tools are its ability to capture and gather data from multiple heterogeneous data sources, its ability to perform advanced analytical processes, and the ability to support multi-users needs and demands. **[4]**

BI is built based on various components such as an Extraction, Transformation and Loading 'ETL' system, data warehouse technology, database query and reporting tools, online analytical processing system, data mining techniques and data visualization tools. **[5]** In our work the data will be extracted from the electronic health individuals records as a source of patient medical records and health status and then moved to the data warehouse repositories then the data mining performed its role in discovering extensive hidden patterns and relationships.

Volume: 11 Issue: 01 Pages: 4142-4149(2019) ISSN: 0975-0290

finally the results introduced through the business intelligence visualization tools, in which these results were used by organizations to support their decision making processes.

On the other hand, the Data mining supports the ability to extract and discover previously unknown, hidden, on other words it discovers interesting patterns from a large database repository. These patterns can aid medical diagnosis and decision-making. Various interesting data mining techniques and algorithms have been designed and developed for many application in different sectors to extract serious knowledge and information in the diagnosis and treatment of disease based on large medical datasets.

Data mining is considered a tool of business intelligence for successful knowledge discovery. The predictive power of data mining is generated from principles of pattern recognition, machine learning, and statistics in which they enables it automatically to extract knowledge and also to determine interest interrelations and patterns from large databases**.[6]** Data mining involves a number of complex and advanced data analysis tools to discover the previously hidden and unknown valid patterns and relationships in large available data sets. These analysis tools can be categorized to mathematical algorithms, statistical methods, and learning algorithms.

For the purpose of our work the Data mining depends on the data from extensive repositories in which the data are represented in a structure of EHR with supporting from many DM tools such as "Waikato Environment for Knowledge Analysis 'WEKA', Konstanz Information Miner 'KNIME', Rapid Miner, Orange, ...etc." to propose an interesting analysis models using the DM techniques such as "classification, clustering and association rule" to get the desired analysis results.

Data mining Techniques; the tasks and activities of data mining that can be modeled in either Predictive or Descriptive fashion. Regarding the Predictive models; they support a prediction about data values using known and identified results captured from available datasets, while the descriptive models discover patterns or relationships in data so they unlike the predictive model. The Predictive data mining models include classification, prediction, regression and time series analysis. Classification is probably the best classic prediction technique in comparison to all the data mining techniques based on the machine learning. Where it classifies each object in set of data into one or more predefined set of classes or groups using one of the classification methods such as Decision trees. **[7,8,9]**

Secondly, the Descriptive model which explore the properties and features of the data examined, not subject to predict any new features and they includes various methods such as Clustering, Association Rules, Summarizations, and Sequence pattern analysis. Descriptive data mining is basically designed to generate frequency and Sequence, cross tabulate and interesting correlation. Descriptive data mining models can be defined to find interesting unknown regularities in the data, to discover hidden patterns and find interesting set of subgroups in the supported bulk of data. From the descriptive models the Clustering techniques is considered the best techniques used to discover groups of objects that are similar to each other in one cluster. The main goal is to group similar objects in one cluster and different objects in another clusters. One of the most well-known clustering technique is K-means that we would depend on using it through our work.

K-means is considered the simplest and well known unsupervised learning clustering algorithm that divides or clusters a data set into a number of groups k. K-means is a well-known partitioning clustering method that tries to obtain a user specified number of clusters represented by their center points (centroids). This method subjects, in other words to generate k clusters based mainly on k centroids (center points of clusters) for each of them. The centroid is defined as the average value of the tuples of a training set.

This algorithm can act better when the centroids are significantly disclose from each other and also there are have high similarity within the corresponding clusters, and low similarity between separate clusters.

After specifying centroids, each object in the data set is assigned to the most closest similar cluster. To this end, a proximity measure is a power for k -means such as Euclidean distance quantifies the notion of the closest. After assigning all objects to their corresponding clusters, each cluster centre is recomputed and position of each centroid is converted to new point. Then, distance between objects and new centroids are computed and assigned to the closest clusters. This process is repeated until all data points are allocated into the relevant clusters **[10, 11].**

II. RELATED WORK

Diabetes is become the most well-known non-transmittable disease in the global. It is assessed to be the seventh leading reason for the death **[12].** It is predicted that the diabetes rate between adults all over the world will become 642 million by 2040 [**1**]. The early diagnosis of diabetes in patients has been a main goal for medical researches and also for healthcare professionals. With the availability of vast technological innovation tools and solutions in computer science, collaborative studies have shown that by applying computer skills. Information technology and advanced algorithms (supported by data mining), efficient, cost effective and rapid methods techniques or sophisticated tools can be designed for the diagnosis of diabetes.

Many researchers have introduced various analysis models using data mining to discover interest and diagnose diabetes. Han et al. in reference **[1]** developed an extensive model that based on the k-means clustering algorithm and the logistic regression algorithm for predicting diabetes. The model obtained about 95.42% accuracy. Patil **[13]** proposed a integrated prediction model that applied k-means clustering partitioning

Volume: 11 Issue: 01 Pages: 4142-4149(2019) ISSN: 0975-0290

algorithm for the original dataset and then using theC4.5d. algorithm for building the classifier model. The classification accuracy findings was about 92.38%. e. Iyer et. al. **[14]** in their study proposed the classification model using the Naïve Bayes algorithm to predict thef. onset of diabetes. The study gave an accuracy result of 79.56%. In Ref. **[15]**, the authors used k-means clustering in determining and removing the outliers data, a geneticg. algorithm and finding interesting correlation based feature selection method for relevant feature extraction, andh. finally used k-nearest neighbor technique for optimally classifying the diabetic patients. in reference **[16]** thei. author followed a systematic methodology based on Principal Component Analysis method to decrease the extracted features in supporting with the Neural Networkj. classifier. The accuracy this analysis results was about 92.2%. k.

III. METHODOLOGY l.

Basically, the given dataset views one of the latest community health diabetes related indicators. This diabetes related indicators data set provides a subset of data (40 indictors) for the diabetes topic. we follow three systematic steps for implementing cluster Modeling based on k-means Algorithm, As shown in figure 1.

The reason why for choosing and using this the k-means clustering Algorithm , as it is one of unsupervised machine learning techniques. It used greatly in the problems where the data is unlabeled (data without specified clusters or groups), in which it identifies numbers of centroids (center objects) and then connects every object to the most closest similar cluster, while maintaining the centroids as small as possible. The Mean in k-means method refers to average of the data that is finding the centroid.

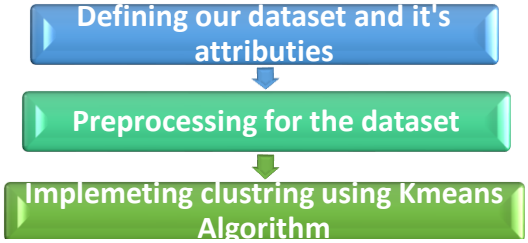


Figure 1: Steps for Clustering Using K-Means Algorithm

1- Defining the dataset:

This community health obesity and diabetes related indicators provides **12 attributes** with their significant values related to **2733** instances/objects:

a. **Country Name:** as shown in Figure 2, there are **72 Country** contributed in this database.
b. **Country Code:** each country has a specific country code, thus there are 72 unique codes.
c. **Region Name:** this attribute identifies 9 Regions values as indicated in Figure 4.

Indicator Number: each Indicator has a specified Indicator Number value.
Indicator : Figure 6 shown the dataset presents 40 obesity and diabetes related indicators.
Denominator Note: presents that the data object Indicator Percentage / Rate based on average annual population.
Measure Unit: this attribute identifies the measurement unit for evaluating the Indicator in each country.
Percentage / Rate: Figure 9 shown that the values have been categorized into 4 groups.
Data Comments: this attributes values confirms that the data have been supported and it applicable for analysis processes.
Date by Years: data have been collected through 3 periods.
Data Sources: data have been collected **6 various sources**, as indicated in Figure 12.
Quartile: data have been categorized into 4 quarters.

2 - Preprocessing and filtering Step

Nowadays the real world databases are highly involved a noisy, missing, and inconsistent data for many reasons such as their typically huge sizes and being origin and captured from multiple and heterogeneous sources **[17].** Data quality in this aspect is occupied an extensive success factor for the data mining analysis process for disease prediction and diagnosis effectively, because low quality data may lead to inaccurate or low prediction result.

In order to make our original dataset more efficient and applicable for clustering and predicting diabetes, we applied several preprocessing techniques using one of various packages offered within the analysis integrated development environment.

Data Preprocessing included the extensive numbers of processes, in which we performed the following processes: cleaning, editing, normalization, transformation, and attribute selection. The product after all this processes, is the final **training set**.

- **Data cleaning:** in this process, the objects that have incomplete, incorrect, inaccurate or irrelevant values were identified. And then the editing or removing actions were considered.
- **Data editing :** in this process the review and adjustment of the dataset has been performed. The purpose is to control the quality of the data that will be used in clustering model process.
- **Normalization:** it is done in order to scale the data values in a specified range.
- **Transformation:** it is taken in order to convert the data in appropriate forms suitable for the mining (clustering) process.
- **Attribute selection:** new attributes were constructed from the given set of attributes to help in the mining process.

Volume: 11 Issue: 01 Pages: 4142-4149(2019) ISSN: 0975-0290

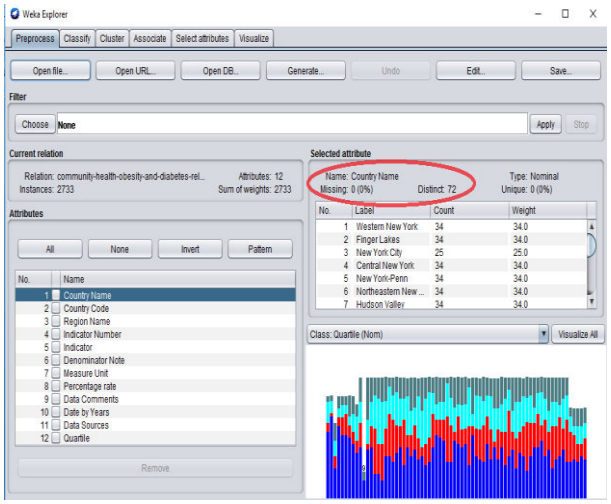


Figure 2: Country Name Attribute Missing 0% of its values

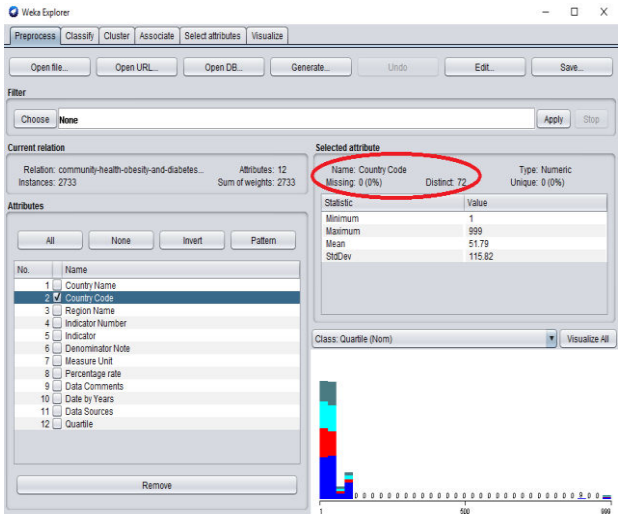


Figure 3: Country Code Attribute Missing 0% of its values

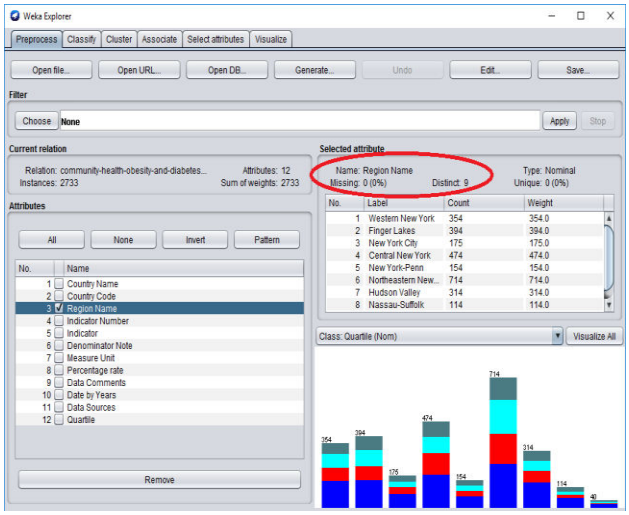


Figure 4: Region Name Attribute Missing 0% of its values

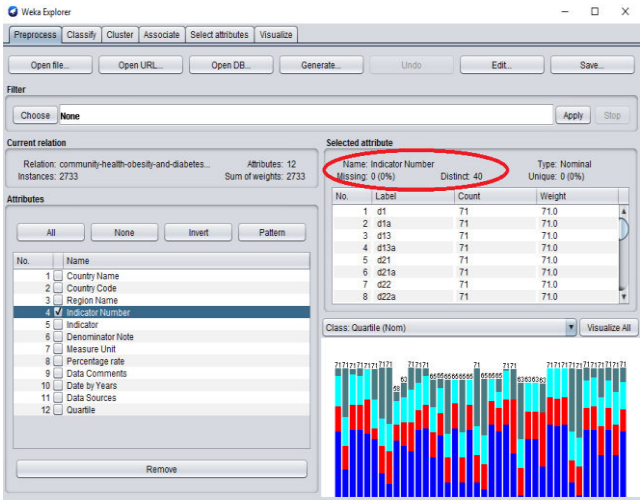


Figure 5: Indicator Number Attribute - Missing 0% of its values

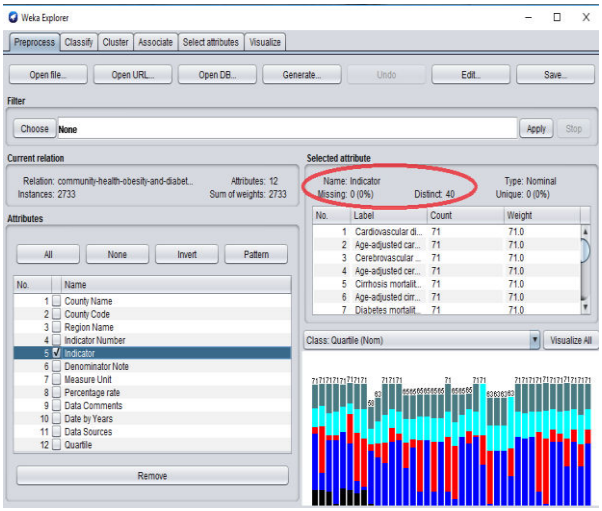


Figure 6: Indicator Attribute - Missing 0% of its values

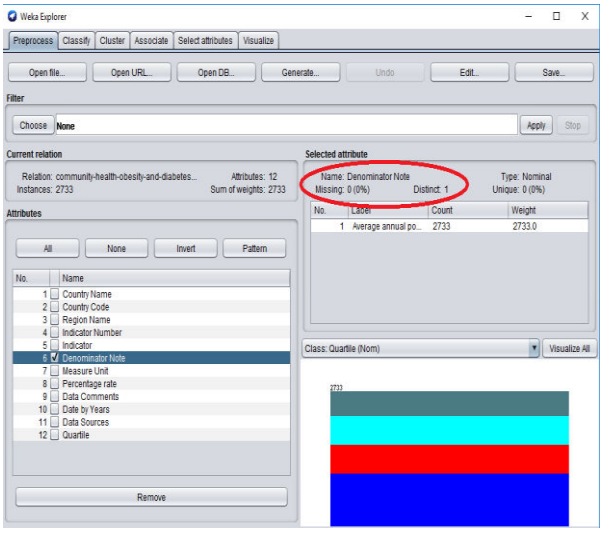


Figure 7: Dominator Note Attribute - Missing 0% of its values

Volume: 11 Issue: 01 Pages: 4142-4149(2019) ISSN: 0975-0290

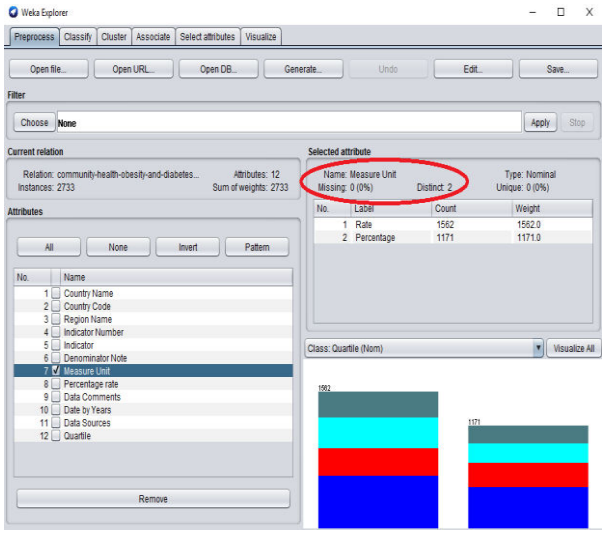


Figure 8: Measure unit Attribute - Missing 0% of its values

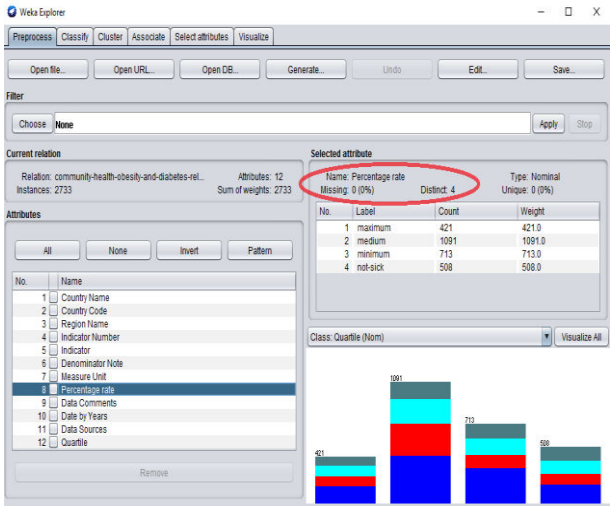


Figure 9: Percentage/ Rate Attribute - Missing 0% of its values

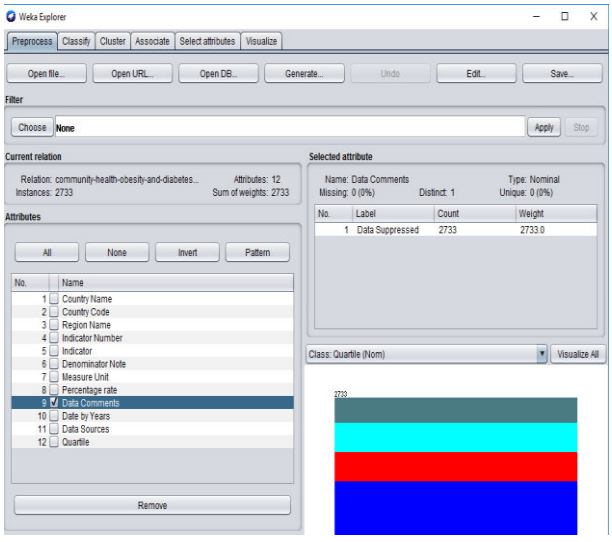


Figure 10: Data comments Attribute - Missing 0% of its values

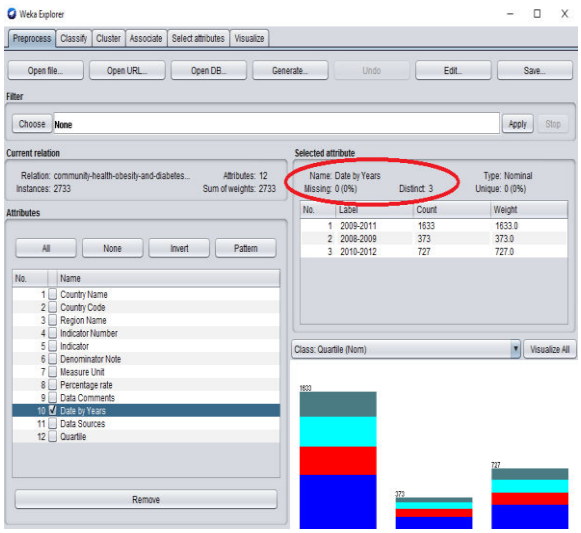


Figure 11: Date by years Attribute - Missing 0% of its values

During the **preprocessing step**, the missing values have been found in two attributes (Data Sources attribute & Quartile attribute), As shown in figures 12 & 13. Figure 12 reveals that the Data Sources attribute missing 15 (1%) values for 15 instances/objects of the given dataset, and Figure 13 finds that the Quartile attribute missing 63 (2%) values for 63 instances. As a result, the **Filtering Capabilities** are used to handle these situations **(look Figure 14)** and prepare the dataset to be applicable for using in the next step (clustering model).

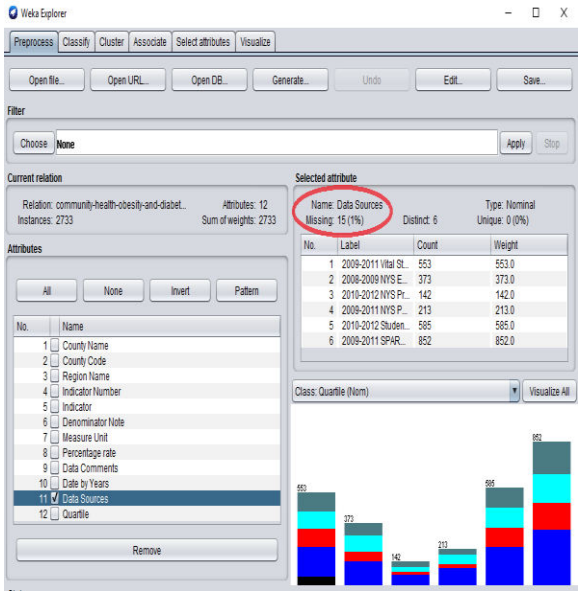


Figure 12: Data Sources Attribute Missing 1% of its values

Volume: 11 Issue: 01 Pages: 4142-4149(2019) ISSN: 0975-0290

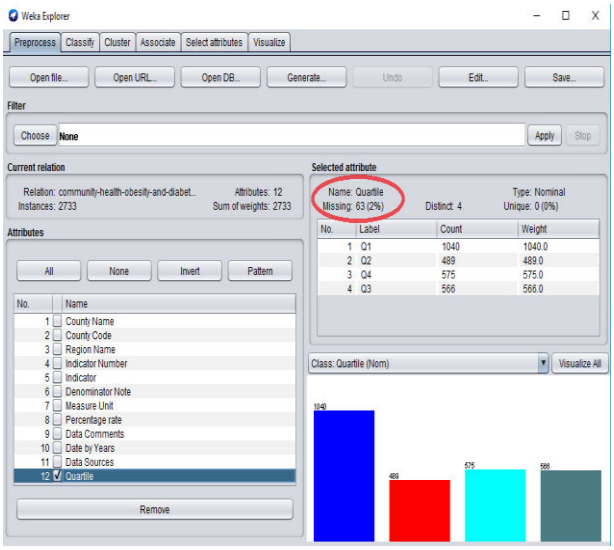


Figure 13: Quartile Attribute Missing 2% of its values

From the Filtering Capabilities, **the missing values capability** is selected to determine the objects that missed values and edit this missed values with a customized value, which I supported to handle this situation.

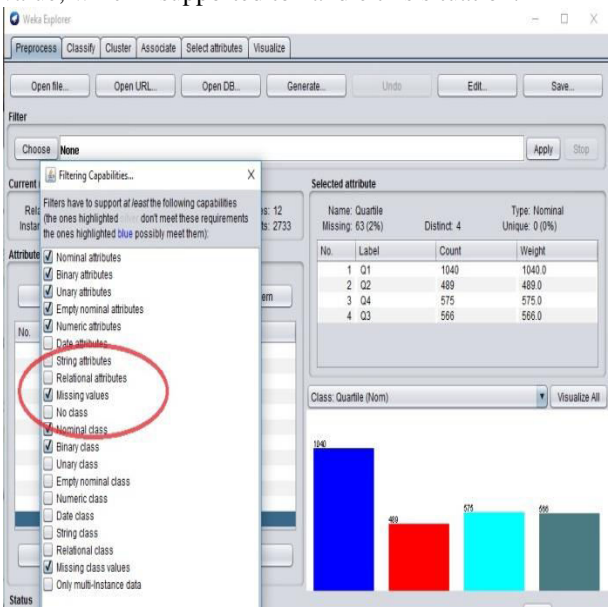


Figure 14: The selection of the missing values capability filtering

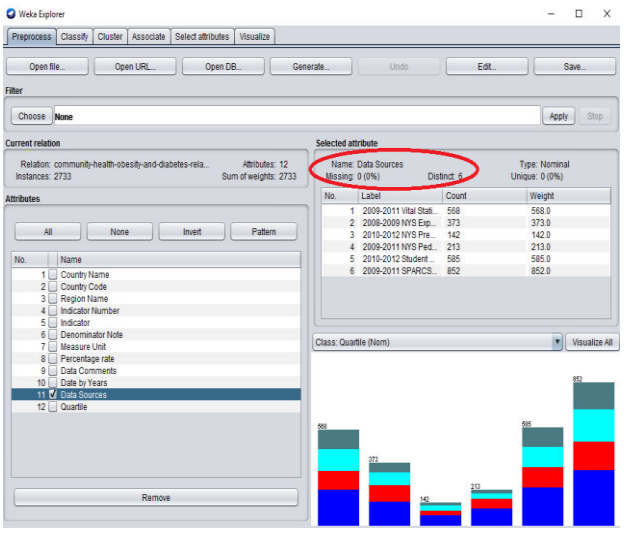


Figure 15: Data Sources Attribute after filtering process– 0% Missing values

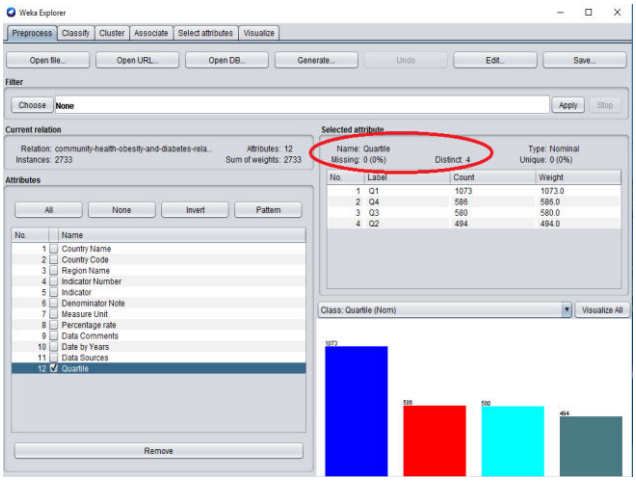


Figure 16: Quartile Attribute after filtering process– 0% Missing values

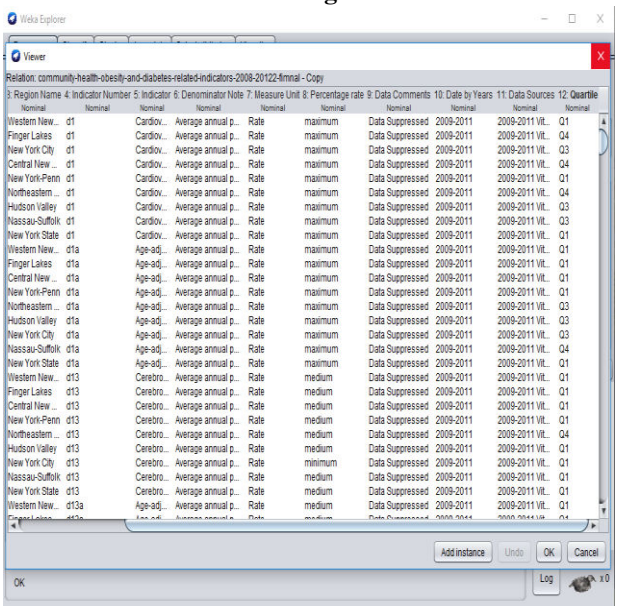


Figure 17: the completeness of dataset after Preprocess & filtering processes

Volume: 11 Issue: 01 Pages: 4142-4149(2019) ISSN: 0975-0290

3- Clustering using K-means

Our task here to grouping together a set of objects in a way that objects/instances in the same cluster are more similar to each other than objects in other clusters. In other words, we try to investigate the structure of the data by grouping the data objects into distinct subgroups (clusters). K-means algorithm is an iterative algorithm tries to partition the dataset into distinct nonoverlapping subgroups where each data object belongs to only one group.

In the following figures, the run for Clustering the given dataset using K-means algorithm is shown in figure 19. The output present 2 clusters and it performed 4 iteration through 0.05 second to build the clustering model. The discovered clusters are:

Cluster 0: Oneida,30,'Central New York',g76,'Age-adjusted percentage of adults who did not participate in leisure time physical activity in last 30 days','Average annual population',Percentage,medium,'Data Suppressed',2008-2009,'2008-2009 NYS Expanded Behavioral Risk Factor Surveillance System Data as of 2010',Q1.

Cluster 1: Nassau-Suffolk,108,Nassau-Suffolk,g62,'Percentage of WIC mothers breastfeeding at least 6 months','Average annual population',Percentage,medium,'Data Suppressed',2009-2011,'2009-2011 NYS Pediatric Nutrition Surveillance System Data as of September, 2013',Q1.

So we can indicate the following results"

- **cluster 0:** discovers that the instances from Oneida in 'Central New York have **Medium Percentage** for adults who did not participate in leisure time physical activity in last 30 days' during 2008-2009.
- **Cluster 1:** reveals that the instances from Nassau-Suffolk,in Nassau-Suffolk , have , **Medium 'Percentage** of WIC mothers breastfeeding at least 6 months' during 2009-2011.

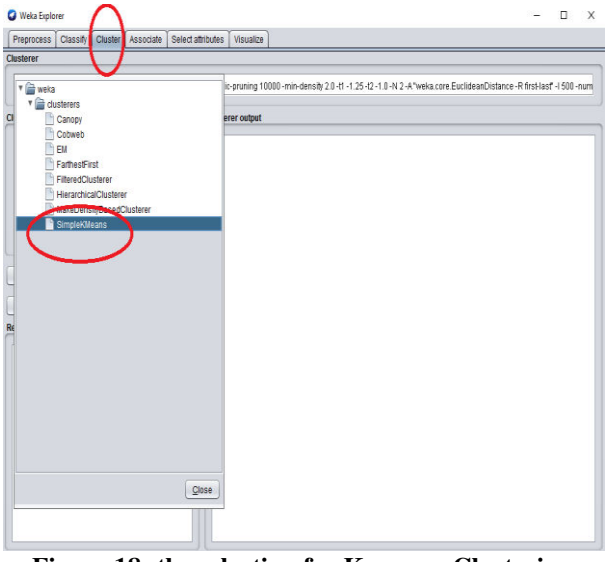


Figure 18: the selection for K-means Clustering algorithm

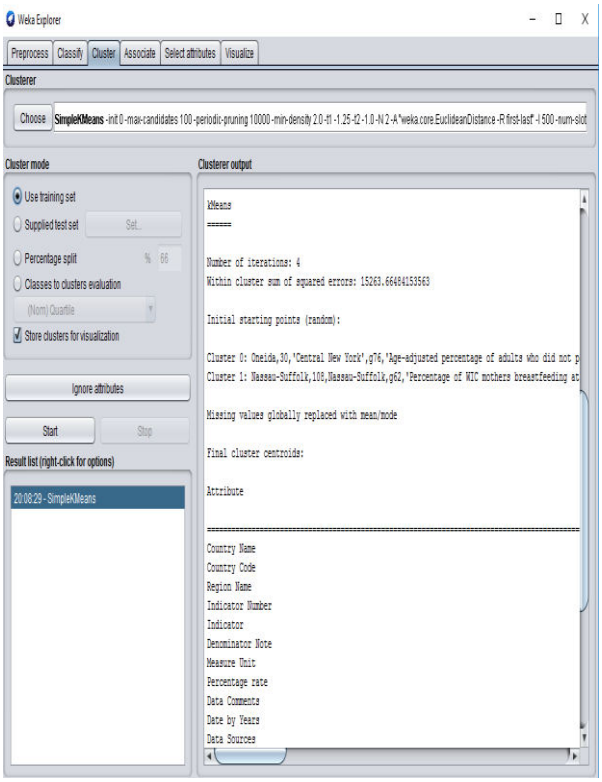
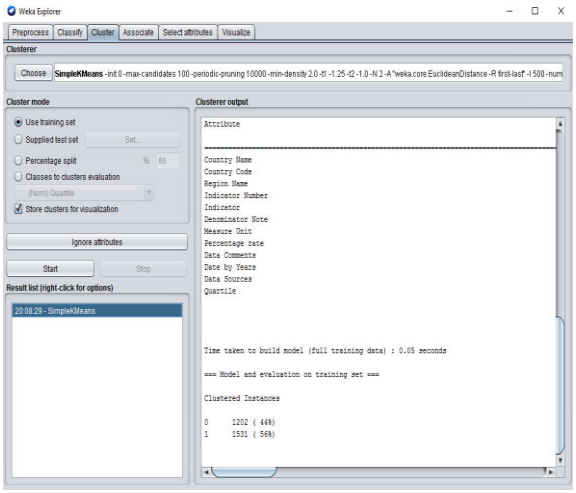


Figure 19: the output of Clustering using K-means



Cont. Figure 19: the output of Clustering using K-means.

IV. CONCLUSION

The aim of this work was to design an efficient model for the discovering an interest knowledge about diabetes based on the latest community health diabetes related indicators dataset. A fast and accurate diabetes clustering analysis is proposed in this paper.

The proposed cluster modeling used 2733 instances within 12 attributes for each one of the dataset. The used data is preprocessed in order to remove the inconsistent or unwanted data, and handle missing values to finally obtain faster processing time. Moreover, the grouping K-means cluster technique for the latest community health

Volume: 11 Issue: 01 Pages: 4142-4149(2019) ISSN: 0975-0290

diabetes related indicators dataset into subsets (K clusters), achieves an optimal clustering results.

In other words the research attempt to employ Data mining techniques based on the utilization of Patient dataset through the Business intelligence application to provide the significant results to make the right decision at the right time.

REFERENCES

1. https://www.diabetesdaily.com/learn-about-diabetes/what-is-diabetes/how-manypeople-have-diabetes/.
2. Ashrafi, N. et al (2014), The impact of Business Intelligence on Healthcare Delivery in the USA, Interdisciplinary *Journal of Information Knowledge and Management*, vol. 11, no. 2, pp117-130.
3. Stackowiak, R. a. (2007), Oracle Data warehouse and Business Intelligence Solutions, *Wiley Publishing*.
4. Tvrdikova, M. (2007), Support of Decision Making by Business Intelligence Tools, 6th International Conference (p. 368), *Computer information system and industrial management application.*
5. Wang, Y. (2010), Business Intelligence and Data Mining in MBS Carbon management.
6. Ahmed, S., Seddawy, A. Nasr, A Proposed Framework for Detecting and Predicting Diseases through Business Intelligence Applications, *International Journal of Advanced Networking and Applications (IJANA)*, Vol. 10, Issue 04, Jan - Feb 2019 issue, pp 3951-3957.
7. Soni, Y. et al (2011), Predictive Data Mining for Medical Diagnosis: An overview of heart disease Prediction, International Journal of Computer Application, pp.43-48.
8. Srinivas, R. (2010), Application of Data Mining techniques in healthcare & Prediction of heart attacks, *International Journal on computer science and engineering*, pp 250-255.
9. Sudhakar, K. et al (2014), Study of Heart Disease Prediction Using Data Mining, *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 14, no. 3, pp.1157-1160.
10. Jain A., Murty M., and Flynn P. (1999), "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323.
11. Abbas O., "Comparisons Between Data Clustering Algorithms," *The International Arab Journal of Information Technology*, vol. 5, no. 3, pp. 320-325, 2008.
12. Khandegar Anjali. Khushbu Pawar diagnosis of diabetes mellitus using PCA, neural Network and cultural algorithm. *International Journal of Digital Application Contemp Res* 2017; vol.5, no.6, pp. 115-125.
13. Patil BM, Joshi RC, Durga Toshniwal. Hybrid prediction model for Type-2 diabetic patients. *Expert Systems Applications* 2010; Vol.37, no.8.pp:8102–8115.
14. Iyer A, Jeyalatha S, Sumbaly R. Diagnosis of diabetes using classification mining techniques. Int J Data Min Knowl Manag Process (IJDKP) 2015; Vol.5, no.1.
15. Gowda Karegowda Asha, Jayaram MA, Manjunath AS. Cascading K-means clustering and K-nearest neighbor classifier for categorization of diabetic patients. *International Journal of Eng Adv Technology* 2012; Vol.1, no.3. ISSN: 2249 – 8958.
16. Novakovic J, Rankov S. Classification performance using principal component analysis and different value of the ratio R. *International Journal of Computer Communication Control* 2011;Vol. VI, no.2, pp. :317–27. ISSN 1841-9836, E-ISSN 1841-9844.
17. Han J, Kamber M, Pei J. Data mining concepts and techniques. 3rd USA: *Morgan Kaufmann Publishers*; 2012.