

An Exploration of Data Mining with Analysis for a Health Care System

Murugananthan Velayutham¹, Mia Torres-Dela Cruz²

¹SEEMIT, Institute Technology Pertama, Negeri Sembilan, Malaysia

²Faculty of Engineering and Technology, Linton University College, Negeri Sembilan, Malaysia

ABSTRACT

There is an undeniable discernment that data mining has a great potential in the improvement of healthcare systems due to its ability to discover patterns and to obtain and isolate important data and information from a large collection of data sets. An in-depth exploration of data mining is essential to understand why this industry is still lagging behind in the implementation of data mining in its technology development. In this paper, datamining techniques has been evaluated with discussion of a broad overview of data mining techniques that point out essentials for health care practitioners and data analytics researchers. Specific examples on health care were shown using evaluation using chosen data mining technique and further discussion were made to show potentials of data mining to the industry.

Keywords: *Clustering, Classification, Health care, Regression, Decision trees.*

1. INTRODUCTION

Data mining is an effective new technology that has a great potential in enabling health systems to solve problems with information they have gathered about their patrons and potential clients. It assists in finding information that would answer questions, improve care, and reduce costs [1]. Nevertheless, because of the complex nature of healthcare and the lag in the acceptance and willingness to embrace technology, the industry is slow in the implementation of effective data mining and analytic strategies. Hence, they are way behind other industries, which successfully use data mining like the retail industry to model customer response. For banks, it helps to

predict and map out customer profitability. The same is true in industries such as telecom, automotive, education, life sciences, manufacturing, and others.

Still at present, data mining in healthcare remains an academic exercise with just a few reasonable success stories. Datamining approaches such as decision trees, clusters, neural networks, and time series are confined to research usage only [4]. However, the healthcare always trailed behind in incorporating the latest research for practical everyday usage.

It is undeniable; however, that datamining is developing into a high potential field to provide insights from a huge amount of data, improving output with lesser costs. It can provide support to systematically utilize data to extract best methods for the healthcare industry and there are great challenges to overcome even when the boundless potential is apparent [4].

This paper explores the potential of datamining for the healthcare industry, delves on the benefits for practitioners and researchers, and discusses the structures, techniques, and specific examples in the health care industry.

2. Data Mining Techniques

Data mining (DM) techniques or methods that use statistical approaches for pattern identification and looks for relations that exist in the given dataset. The DM techniques are classified in the proceeding sections and shown in the following figure (Figure 1):

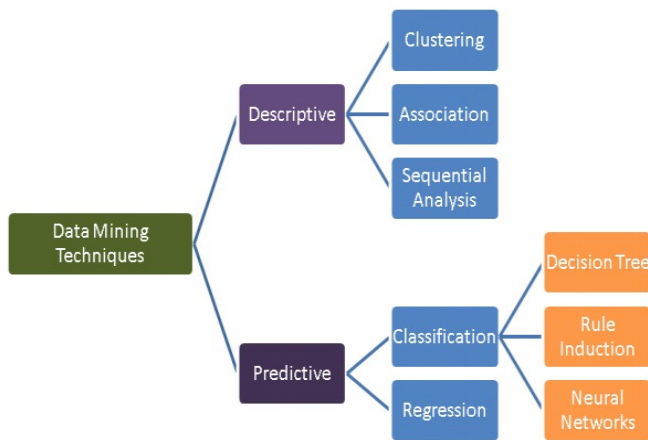


Figure1. Data Mining Techniques

2.1 Descriptive Data Mining Techniques

DM descriptive technique or estimations does unequivocally what the name proposes, they “describe” or gather unrefined data and make it into something that is interpretable by individuals. These are good and sound techniques that explain historical data, which implies motivation behind time that an event has happened, whether it is one-minute back or one year further. Connecting with techniques is useful in a way as they allow picking up from past practices, and perceiving how they may affect future results [6][8]. Further classifications of descriptive data mining are the following:

2.1.1 Clustering

Clustering or bundling is the task of accumulation a game plan of things in a way that challenges in the same social event called a gathering are more tantamount in some sense or another to each other than to those in diverse get-togethers (clusters) [8]. It is an essential undertaking in exploratory of data mining and it grasps the regular assembling or structure in a data set.

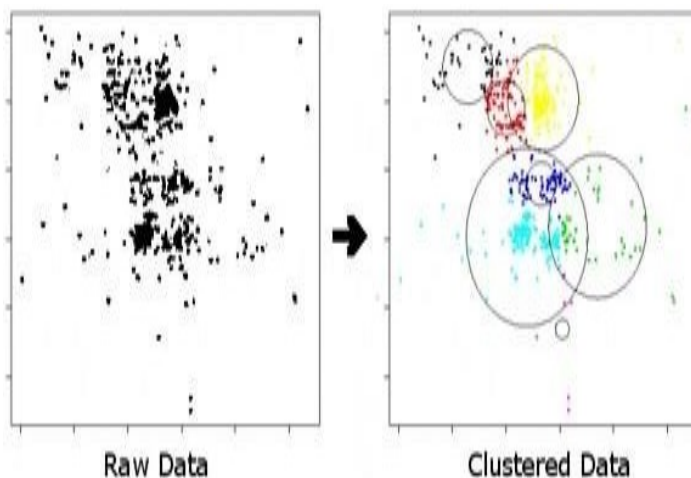


Figure2. Clustering to show raw data to clustered data

2.1.2 Association Rule

Association tenet is a system for finding intriguing associations between items in mining data. It is planned to recognize solid principles found in mining of data utilizing diverse measures of interestingness [7]. An association tenet has two sections, a predecessor (if) and a subsequent (then). A predecessor is a entity found in the information. A subsequent is a item that is found in blend with the forerunner. In information mining, affiliation guidelines are valuable for examining and anticipating client conduct.

2.1.3 Sequential Pattern mining

Sequential pattern mining is a data burrowing procedure for getting customary successive cases in a back to back database. Progressive case of data mining stressed with finding verifiably pertinent samples between data representations where the qualities are passed on in a game plan. It is commonly expected that the qualities are discrete, and thusly time plan mining is solidly related, yet ordinarily considered a substitute activity. Progressive sample mining is a remarkable instance of sorted out data mining [6].

2.2 Predictive Data Mining Techniques

Data mining predictive techniques perceive its roots with the ability to "Anticipate" what may happen. These examinations are about perception about what has to come. Farsighted examination gives associations a significant encounter in perspective of data. Insightful examination give gages about the likelihood of a future result. Keep in mind that no true estimation can "suspect" the future with 100% conviction. Associations use these estimations to figure what may happen later on. This is by virtue of the foundation of insightful examination that relies on probabilities.

Classification: is a data mining strategy used to anticipate bundle enlistment for data events. Gathering is similar to batching where similar areas of customer records are made into unmistakable segments called classes. Yet, not under any condition like gathering, a course of action examination requires that the end- customer/master know early how classes are described. However, we may wish to use portrayal to anticipate whether the atmosphere on a particular day will be "sunny", "stormy" or "cloudy". Renowned request frameworks consolidate decision trees and neural frameworks [7].

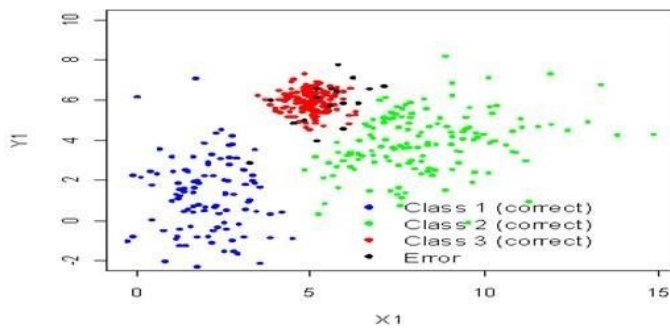


Figure3. Classification for predictive data mining

2.2.1 Decision tree

This is a judicious model, which maps recognitions around a thing to choose about the thing's goal. It is one of the insights that show philosophies as a piece of estimations in data mining. In decision examination, a decision tree can be used to apparently and explicitly address decisions and decision making. In data mining, a decision tree depicts data however not as decisions; rather the resultant gathering tree can be information for decision-making [6].

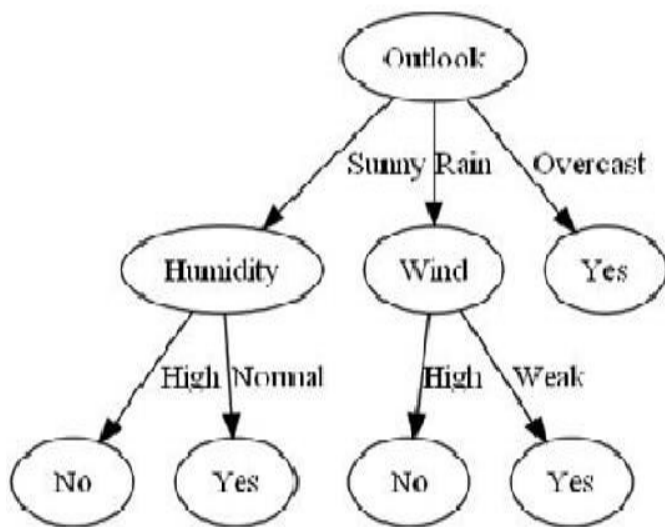


Figure4. An Example of a Decision Tree

2.2.2 Neural Networks

Neural Networks are sensible strategies revealed after the (speculated) methodology of learning in the idiosyncratic structure and the neurological components of the brain and prepared for anticipating new discernments (on specific variables) from distinctive observations (on the same or diverse variables) in the wake of executing a technique of affirmed picking up from existing data [8] [9].

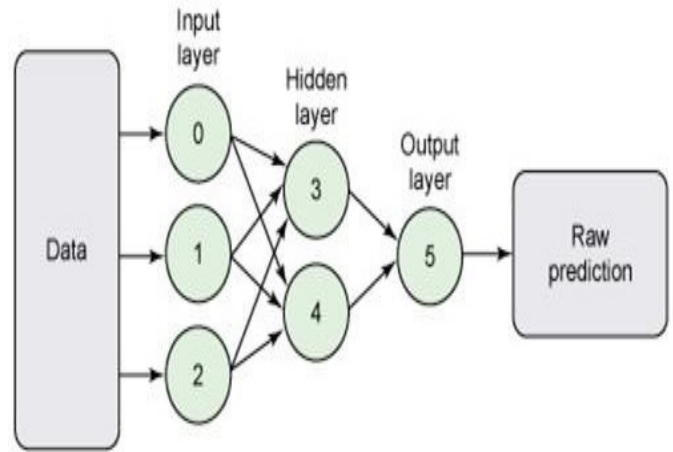


Figure5. An Artificial Neural Networks

2.2.3 Regression

Backslide is a strategy for mining data that used to fit a numerical articulation to a dataset. Backslide is a data mining limit that predicts a number like age, weight, partition, temperature, compensation, which could all be expected using backslide techniques. A backslide undertaking begins with a data set in which the target qualities are known Regression models and are attempted by preparing diverse bits of knowledge that measure the refinement between the foreseen qualities and the ordinary qualities. The least demanding kind of backslide, direct backslide, uses the formula of a straight line ($y = mx + b$) and chooses the suitable qualities for m and b to suspect the estimation of y based upon a given estimation of x [8]. Moved methods, for instance, grant the use of more than one data variable and contemplate the fitting of more perplexing models, for instance, a quadratic correlation.

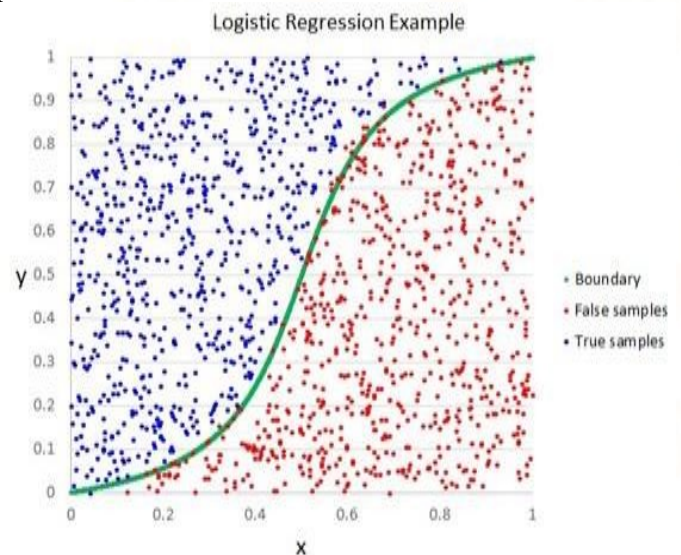


Figure6. An Example of a Logistic Regression

3. SELECTED TECHNIQUES THAT WILL SUIT A MEDICAL CENTER BASED PATIENTS OUTCOME

We have now narrowed our research to the following technique and now we will discuss them based on their effectiveness for our system approach.

- Clustering
- Classification

3.1 Clustering technique

With the usage of collection technique in other to find the therapeutic administrations of patients result, packing system will offer amid the time, some assistance with isolating the course of action of patients data into a plan of noteworthy sub-classes, called bunches [1]. This will offer the wellbeing and help with minding center grasp or choose the patients result in a data set. Batching system can be used either as a stand- alone instrument to get learning into the patient data scattering or as a preprocessing endeavor for distinctive estimations. Grouping strategy will in similar manner give a statistical response for the human administrations of the data [5]. The patients data will be sub-isolated into social occasions (gatherings) such that the data in a bundle are essentially the same (yet not unclear) to one another and though not quite the same as the data in diverse clusters. Gathering is a revelation technique that reveals affiliations, samples, associations, and structures in masses of data.

3.2 Classification Technique

This method will be used to expect human administration center to process the data events of the patients. Plan system is some-how like packing technique in light of present circumstances to aggregate the human administrations patients records into specific areas called classes. In any case, not in the least like grouping system, a portrayal strategy will clear up right on time how classes are described. A count that will realize the course of action procedure, especially in a strong use of the restorative administrations tolerant result, is known as a classifier. The expression "classifier" now and again, implies the numerical limit, completed by a portrayal figuring, which maps data to a grouping.

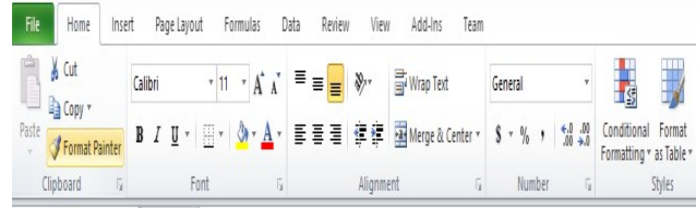
In course of action using this technique, the impression of patients result will be seen as events, the enlightening variables are termed segments (assembled into a component vector), and the possible

groupings of the patient data set to be expected are classes [3].

4. EVALUATION USING THE CHOSEN DATA MINING TECHNIQUES

4.1 Classification

Evaluation of Classification Technique on the health care patients using Weka Data mining tool



	A	B	C	D	E	F	G	H	I	J	K	L
1	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat	Class		
2	40-49	premeno	15-19	0-2	yes		3 right	left_up	no	recurrence-events		
3	50-59	ge40	15-19	0-2	no		1 right	central	no	no-recurrence-events		
4	50-59	ge40	35-39	0-2	no		2 left	left_low	no	recurrence-events		
5	40-49	premeno	35-39	0-2	yes		3 right	left_low	yes	no-recurrence-events		
6	40-49	premeno	30-34		5-Mar yes		2 left	right_up	no	recurrence-events		
7	50-59	premeno	25-29		5-Mar no		2 right	left_up	yes	no-recurrence-events		
8	50-59	ge40	40-44	0-2	no		3 left	left_up	no	no-recurrence-events		
9	40-49	premeno		14-Oct 0-2	no		2 left	left_up	no	no-recurrence-events		
10	40-49	premeno	0-4	0-2	no		2 right	right_low	no	no-recurrence-events		
11	40-49	ge40	40-44	15-17	yes		2 right	left_up	yes	no-recurrence-events		
12	50-59	premeno	25-29	0-2	no		2 left	left_low	no	no-recurrence-events		
13	60-69	ge40	15-19	0-2	no		2 right	left_up	no	no-recurrence-events		
14	50-59	ge40	30-34	0-2	no		1 right	central	no	no-recurrence-events		
15	50-59	ge40	25-29	0-2	no		2 right	left_up	no	no-recurrence-events		
16	40-49	premeno	25-29	0-2	no		2 left	left_low	yes	recurrence-events		
17	30-39	premeno	20-24	0-2	no		3 left	central	no	no-recurrence-events		
18	50-59	premeno		14-Oct	5-Mar no		1 right	left_up	no	no-recurrence-events		
19	60-69	ge40	15-19	0-2	no		2 right	left_up	no	no-recurrence-events		
20	50-59	premeno	40-44	0-2	no		2 left	left_up	no	no-recurrence-events		
21	50-59	ge40	20-24	0-2	no		3 left	left_up	no	no-recurrence-events		
22	50-59	lt40	20-24	0-2	?		1 left	left_low	no	recurrence-events		
23	60-69	ge40	40-44		5-Mar no		2 right	left_up	yes	no-recurrence-events		
24	50-59	ge40	15-19	0-2	no		2 right	left_low	no	no-recurrence-events		
25	40-49	premeno		14-Oct 0-2	no		1 right	left_up	no	no-recurrence-events		
26	30-39	premeno	15-19		8-Jun yes		3 left	left_low	yes	recurrence-events		
27	50-59	ge40	20-24		5-Mar yes		2 right	left_up	no	no-recurrence-events		

Figure7. Dataset for the proposed health care system

The above screenshot (figure 7) is the dataset that have been used to find the desired output for the proposed health care center patients, suffering from breast cancer. The dataset will be used to provide valid output using Weka data mining tool and classification technique [2].

5. Results and discussions

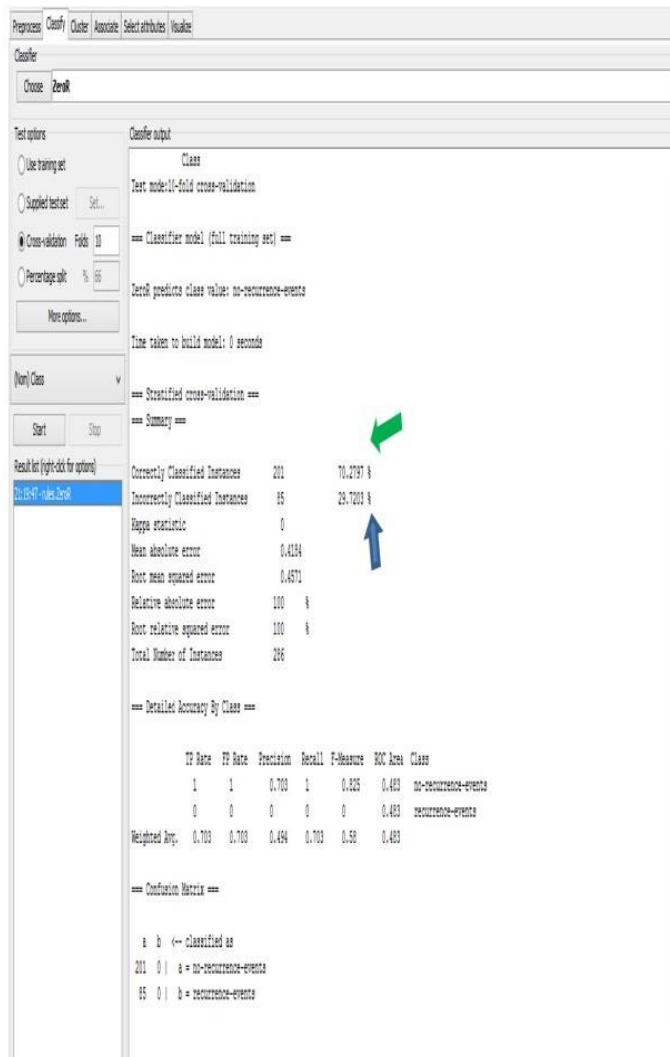


Figure8. Screenshot for the classifier output

The above screenshot (figure 8.) is a classifier output of breast cancer dataset of the medical center patients as indicated by the green arrow above. The output consist of instances, correctly classified instances as indicated by the black arrow above and incorrectly classified instances as indicated by the blue arrow.

Valid Instances: Based on the classification of the dataset instances, 70.2 % instances of the dataset are correctly classified as shown by the green arrow above.

Invalid Instances: Based on the classification of the dataset instances, 29.7 % instances of the dataset are incorrectly classified as shown by the blue arrow above.

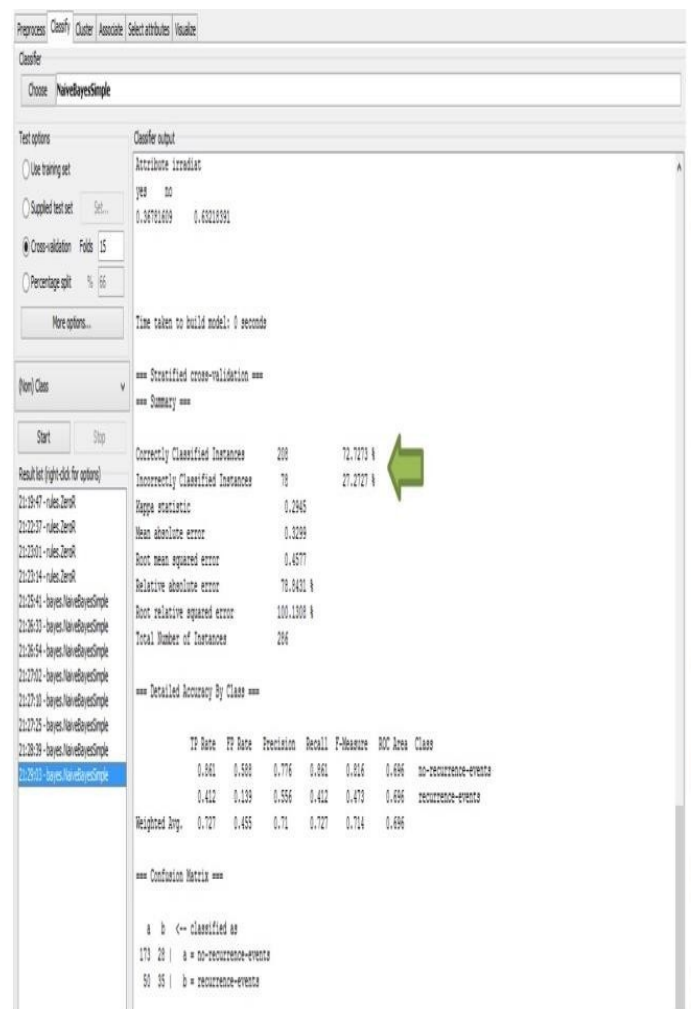


Figure9. Screenshot for the classifier using NavesBayes

The above screenshot is a classifier using NaiveBayes and cross validation output of breast cancer dataset of the medical center patients as indicated above. The output consists of instances, correctly classified instances and incorrectly classified instances as indicated by the green arrow.

Valid Instances: Based on the classification of the dataset instances using Naivebayes and cross validation, 72.7 % instances of the dataset are correctly classified as shown by the above.

Invalid Instances: Based on the classification of the dataset instances using Naivebayes and cross validation, 27.2 % instances of the dataset are incorrectly classified as shown by the above.

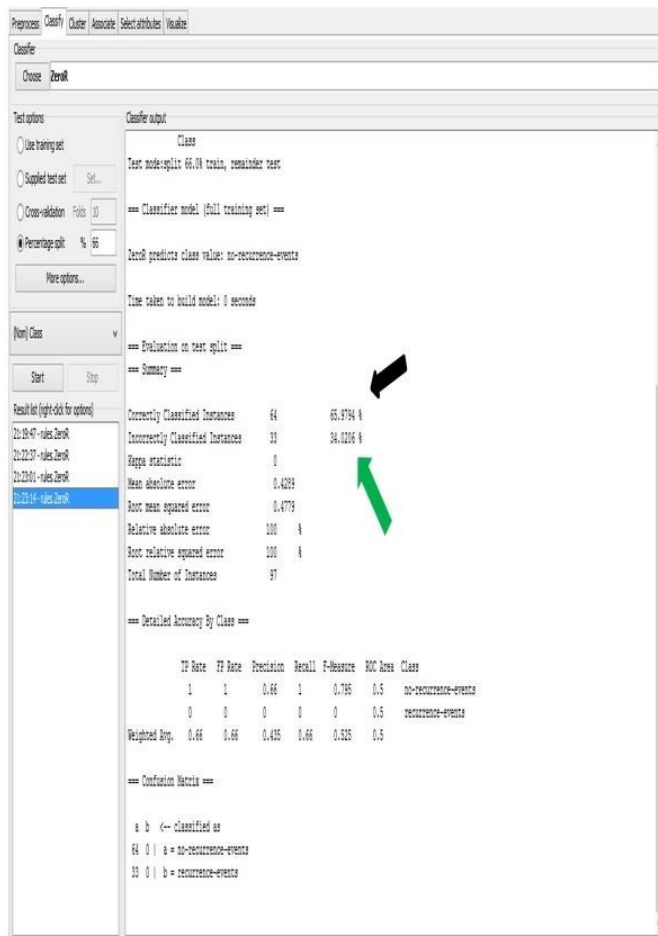


Figure10. Screenshot for the classifier using Rule-ZeroR

The above screenshot (figure 10) is a classifier using percentage split output of breast cancer dataset of the medical center patients as indicated by the blue arrow above. The output consists of instances, correctly classified instances and incorrectly classified instances as indicated by the black and green arrow.

Valid Instances: Based on the classification of the dataset instances using percentage split, 65.9 % instances of the dataset are correctly classified as shown by the black arrow above.

Invalid Instances: Based on the classification of the dataset instances using percentage split, 34.0 % instances of the dataset are incorrectly classified as shown by the green arrow above.



Figure11. Screenshot for the classifier using Bar Chart

The above screenshot (figure 11) is a classifier bar chart output of breast cancer dataset of the medical center patients. The green arrow above indicates age range of the patients while the black arrow indicates the attributes of the datasets. The output shows cancer tumor rate among patients age range.

The screenshot (figure 12.) is a classifier bar chart output of breast cancer dataset for the medical center patients. The screenshot above displays the entire cancer rate among the attributes of the dataset. The output shows cancer rate instances, it also indicates benign cancer rate (low cancer instance) and malignant cancer rate (high cancer instance) in the attributes.

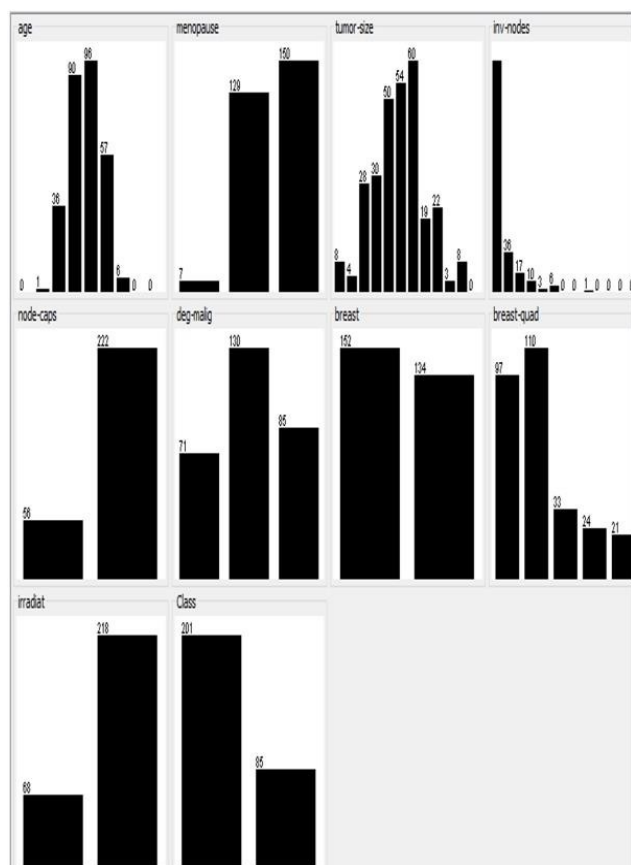


Figure12. Screenshot for the classifier using Bar Chart Type 2

5. Conclusion

It has been observed that industries have implemented data mining in their systems successfully. Healthcare industry, however, is taking a longer time comparatively in embracing DM in its system improvements for technology development. In this paper, it has been pointed out that the potential is great for healthcare providers to utilize DM due to its ability to transform its systems into more effective and efficient information providers. Because of DM's characteristic to extract and identify patterns in clinical and data repositories, decision making in the healthcare organizations will be informative. Continuous improvement of the DM frameworks and methods would make it be more widespread and implementation will show rapid frequencies across the health care industry.

Among the DM techniques, there are specific techniques that would be more suited to medical center based patients, which are the clustering technique and the classification technique. Clustering is the most appropriate for healthcare claims data especially for high cost information, where the cost patterns and sample size is taken into consideration. Classification technique uses the patient results in the

form of events, in segments, and grouping are set into classes.

The transformation of written health records to electronic data has made clinical and healthcare databases more efficient and effective. Added to that, sharing of these information has increased the knowledge demand and distribution across all healthcare sectors all over the world. Through this improvement, the thorough and accurate documentation of patient information has improved healthcare and patient satisfaction. As such, healthcare future may be dependent on data analytics, specifically, data mining in the improvement of overall healthcare services.

6. References

1. Dunsmuir, D., Payne, B. ; Cloete, G. ; Petersen, C. (2014), "Development of mHealth Applications for Pre eclampsia Triage", IEEE Journal of Biomedical and Health Informatics, Vol. PP, No. 99, January , pp. 2168-2194
2. Fayyad, U., Piatetsky- Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.
3. Leventhal, B. (2010). An introduction to data mining and other techniques for advanced analytics. Journal of Direct, Data and Digital Marketing Practice, 12(2), 137-153, doi:10.1057/dddmp. 2010.35.
4. Bennett C, Doub T (2011). Data mining and electronic health records: selecting optimal clinical treatments in practice. CoRR abs/1112: 1668
5. Holzinger, A., Dehmer, M., & Jurisica, I. (2014). Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. BMC bioinformatics, 15(6), I1
6. Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons
7. Dunham, M. (2003). Data Mining: Introductory and Advanced Topics. Upper Saddle River, NJ: Pearson Education.
8. Berson, A., Smith, S., & Thearling, K. (2011). An Overview of Data Mining Techniques.
9. Han, J., Kamber M. (2006). "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers.