Efficiency Analysis of Health Care Centers Using Data Envelopment Analysis

Hassan Najadat*, Qutaibah Althebyan*, Abedallah Khamaisheh†, Mohammad Al-Saad† and Ahmad Al Rawashdeh†

*Jordan University of Science and Technology
Irbid, Jordan, 22110
Email: {najadat, qaalthebyan}@just.edu.jo
†Jordan University of Science and Technology
Irbid, Jordan, 22110
Email: {amkhamaiseh14, mzalsaad15, amalrwashdeh15}@cit.just.edu.jo

Abstract—Efforts to produce more accountable health care organizations are an emerging global health trend. High performing organizations hold health care providers accountable for both efficiency and the quality of care provided. Increasing demands for greater accountability will require managers who focus more time and attention on "best practices" performance measurement. For starters, managers need an approach to measure and assess, as well as benchmark their organizations’ performance against that of similar organizations. The main key idea of this paper is to propose an approach to evaluate the technical efficiency of health care centers located in Jordan. The proposed approach is based on two techniques: Data Envelopment Analysis (DEA) and Data Mining Techniques. The results showed that out of 21 Jordanian regions, 13 were efficient and 8 were not efficient. Several classifiers (J48, SMO, Nave Bayes) were used in the evaluation of the results. The obtained results showed that the J48 classifier, which represents a decision tree classifier, produced the highest accuracy in almost all experiments studied.

I. INTRODUCTION

Recently, health care has become a major concern all over the world. In fact, most high developed countries besides other countries pay more attention to the health care services as well as health care providers. These attempts aim at improving the kind of services offered to the people. This in fact resulted in improving the health of people in these countries. It also resulted in improving peoples’ practices and behaviors in terms of health. Although this has an impact on the quality of health services provided, however, it has substantial impact on the health care expenditures. Although such expenditures will not have bad impacts on the economies of rich countries, it will have bad impacts on the economies of the developed countries due to several and non-endless reasons. Hence, the importance of evaluating the efficiency of health care services as well as health care institutions and providers has risen and entitled [1]. In fact, less efficiency means a waste of resources that can be used as needed resources for other issues. For a long time discussing the efficiency of the health care services in the developed countries was not a major issue. One reason behind that is that such countries used not to provide the minimal health care services. Moreover, the most and major concern behind that was the financial support that might be an impediment to these countries’ economies. An attempt to evaluate the efficiency of health care services is a very important and vital process that needs to be considered and adopted. This in fact helps in the knowledge of the levels and the determinants of health care services provided in any country. Such knowledge can be used by health care professionals and policy-makers to avoid wasted resources and hence, redirect these wasted resources to the places where it should be. Several methods had been proposed in the literature that handled measuring the efficiency of any underlying methodologies, namely, frontier analysis, Data Envelopment Analysis. In fact the Data Envelopment Analysis has become a popular benchmark used for evaluating the efficiency of decision making units where any system can be decomposed into several components that can be thought of as decision making units. Data Envelopment Analysis (DEA) is a theory based mathematical approach introduced to health care policy makers and managers to measure and evaluate the relative performance of health care organizations. DEA is a powerful performance evaluation methodology capable of identifying top performers in relation to less effective performers. The methodology can handle multiple, non-commensurate inputs and outputs, including qualitative factors such as patient satisfaction. DEA can be used to process multiple inputs and outputs. Most importantly, DEA estimates a single, summary measure of relative performance, without requiring prior weights.

II. BACKGROUND

In this Section, we will briefly talk about data mining techniques and Data Envelopment Analysis approach in evaluating the performances of organizational units. In addition, we will talk briefly about DEAP which is data envelopment analysis software and an open source data mining tool called WEKA. All the aforementioned concepts will be used in our proposed approach for evaluating the efficiency of health care services and centers.

A. Data Mining Techniques

Data mining [6] refers to the process of extracting interesting patterns or useful knowledge from large amounts of data. Several steps are involved in a data mining process in
terms of the concept of knowledge discovery. Data mining is widely used and has many applications in telecommunication industry, retail industry, financial data analysis, biological data analysis, intrusion detection, and many scientific applications. Several techniques and methods are used in data mining but not limited to: classification, clustering, and association rule mining. Classification [6] is a two-step process of finding a model that best describes groups of objects that may have common properties, to distinguish data objects and predict categorical labels or the class for objects when the class label is unknown. This model is derived from the analysis of data called training data where the trained data represents data objects that have unknown class label. However, to predict numerical data values, another technique called prediction is used, where the prediction process finds models for continuous-valued functions to predict missing or unavailable numeric values.

Clustering: [6] is the process of analyzing data objects without previous knowledge of class labels. Objects are grouped together by maximizing the intra-class similarity and minimizing the interclass similarity. In clustering, each group represents a class of objects, and objects in the same cluster or group have high similarity when compared to another object in the same cluster, but very dissimilar with other objects in the other clutters.

Association Rule Mining: [6] is a two-step process that finds all frequent item sets then generates a strong association rules from that frequent item sets. These item sets are determined as frequent based on a minimum support threshold. The resulting rules must satisfy both minimum support and minimum confidence thresholds.

B. Data Envelopment Analysis (DEA)

Data Envelopment analysis (DEA) [2], [3] is a nonparametric linear programming approach for assessing, measuring and evaluating the efficiency, productivity and the performance of a set of peer entities called Decision Making Units (DMUs) which produce multiple outputs from multiple inputs. DEA was first introduced in 1978 as a managerial and performance measurement tool. A Decision Making Unit (DMU) definition is generic and flexible, in general it refers to any entity to be evaluated in terms of its efficiency in converting inputs into outputs. DEA approach has many features and becomes widely used in evaluating many different kinds of entities in many different contexts. One key feature is that it requires just a few assumptions, which makes this approach appropriated in evaluating the performance in many cases where the relations between inputs and outputs involved in DMUs are very complex or unknown. DEA approach has many applications in evaluating the performance of entities of various forms of DMUs including: schools, hospitals, banks, airports and others.

DEA can handle multiple inputs and multiple outputs at the same time without any assumptions on data distribution. As a result, DEA models can be divided in terms of proportional change in inputs or outputs into two main models: the input-oriented model, where the concentration is on minimizing inputs while at least preserving the given output levels, and the output-oriented model, where the concentration is on maximizing outputs without the need of maximizing the given inputs.

DEA models also can be divided in terms of returns to scale into two main efficiency measurement models by adding weight constraints, the first model is constant return to scale (CRS) and the second model is variable return to scale (VRS). CRS efficiency measurement model was originally proposed in [4] where all DMUs are assumed to be operated at their optimal scales. VRS model was firstly introduced in [5] as a non-constant DEA efficiency measurement model, where DMUs can operate in a variable scales, by allowing the breakdown of efficiency into technical and scale efficiencies in DEA.

III. RELATED WORK

DEA has been extensively used for analyzing and evaluating health care services and providers. In fact the DEA has been used widely in the assessment of health care organizations all over the world [11]–[14]. Although there are several approaches used in health care efficiency evaluation, however, the DEA is considered as the most widely used and adopted approach.

The authors of [11], [15] estimated the technical efficiency of a primary health care centers in Chile. The authors used data collected from 259 Chilean municipalities; 82 in the urban group and 177 in the rural group. Their data spanned the era from 2006 and 2008. The use of the DEA approaches gave good and very helpful results in classifying the areas where some areas are found to be efficient while others are found to be non-efficient; hence, subsequent steps and remedies need to be done by health care officials.

In another study, Satyanarayana [8] measured the technical efficiency for the health care centers in Indian rural areas. The datas were collected from several resources. Again, the use of the DEA approaches gave good and very helpful results in classifying the areas where some areas are found to be efficient while others are found to be non-efficient; hence, subsequent steps and remedies need to be done by health care officials.

The authors of [9] used the Data Envelopment Analysis to calculate the technical efficiency of Basic Health Units (BHUs) in Sargodha with the choice of inputs and outputs being specific to BHUs operation. In their work, they used the data of 116 basic BHUs in district Sargodha for the year 2010. Data used to be handled by the Statistical department of the health care center. Four outputs and inputs were used to evaluate the efficiency of the BHUs. Input variables are Number of Medical staff, Number of Para-medical staff, Number of Lady Health Workers, Number of other staff while, Number of outdoor patients, Number of Child immunized, Number of Family Planning Visits and Number of First Antenatal care visits are output variables. DEA model results revealed that 76% BHUs were inefficient and destructing the infrastructure. The overall findings were consistent with the expectation.
that public health services delivery mechanism in developing countries is inefficient. Hence, governments should allocate more resources on the health sector to improve the devastated infrastructure. Akazili [10] conducted a study to estimate the technical efficiency of 89 randomly selected health care centers in Ghana using the DEA method. The aim was to determine the degree of efficiency for health centers and consequently, draw recommendations for performance targets for inefficient facilities. The findings showed that 65% of health centers were technically inefficient and using resources that they did not actually need.

IV. DATASET INFORMATION

Our dataset consists of health data collected from health centers in 21 Jordanian regions each with three input variables and two output variables. The dataset has been collected from ministry of health website (http://www.moh.gov.jo/Pages/viewpage.aspx?pageID=169/). Table I shows the variables with their descriptions.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>Input</td>
<td>Total number of healthcare centers</td>
</tr>
<tr>
<td>X2</td>
<td>Input</td>
<td>Number of available clinics</td>
</tr>
<tr>
<td>X3</td>
<td>Input</td>
<td>Number of patients</td>
</tr>
<tr>
<td>Y1</td>
<td>Output</td>
<td>Number of dental patients treatment</td>
</tr>
<tr>
<td>Y2</td>
<td>Output</td>
<td>Patients who are not hospitalized for 24 hours</td>
</tr>
</tbody>
</table>

Table II shows several statistics for the input and output variables of the dataset, the statistics were computed using the WEKA tool.

<table>
<thead>
<tr>
<th>Name</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>StdDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>7</td>
<td>62</td>
<td>32.238</td>
<td>355.84</td>
</tr>
<tr>
<td>X2</td>
<td>305</td>
<td>426</td>
<td>148.857</td>
<td>82.528</td>
</tr>
<tr>
<td>X3</td>
<td>115756</td>
<td>2685048</td>
<td>619096.81</td>
<td>612248.257</td>
</tr>
<tr>
<td>Y1</td>
<td>5027</td>
<td>272323</td>
<td>72514</td>
<td>65487.945</td>
</tr>
<tr>
<td>Y2</td>
<td>103674</td>
<td>2282470</td>
<td>529750.476</td>
<td>529542.975</td>
</tr>
</tbody>
</table>

Table I: Health Centers Dataset Variables

Table II: Health Centers Dataset Statistics

The collected dataset contains only regions of health centers information lacking the efficiency scores. We used DEAP software to compute the efficiency of the regions of health centers to produce the final complete data set which contains the health centers data along with the efficiency scores.

V. METHODOLOGY AND RESULTS

Our methodology combines the results of DEA with the data mining techniques which are the clustering and classification techniques. Several algorithms for each of the techniques were incorporated using WEKA [7]. The results of DEA are efficiency scores for regions of health centers. These efficiency scores were calculated using the input oriented model.

A. Applying Clustering

The problem found in the dataset was that the regions differ in population size. To make a fair decision, clustering algorithms were used to group data based on similarities among their population sizes. When the DEAP software is used for the whole dataset, only four regions which represent the largest population sizes were classified as efficient regions. However, when the datasets were grouped using Simple K Means algorithm; the efficient regions are increased to 13 regions. The number of clusters has been determined using elbow method to specify the optimal number of clusters. In our study the best number of clusters was found to be 3. After that, the dataset became ready to apply the classification techniques.

The result of applying clustering on the original dataset shows that the Simple K Means clustering algorithm is able to group the dataset into 3 clusters with sum of squared errors value 19.56. The number of regions assigned to each cluster differs from one cluster into another one where the first cluster has 9 regions, the second one has 4 regions, and the third cluster has 8 regions.

B. Applying Classification

Classification divides the data into set of classes based on the value of class label. The class label that is regarding to the DEA represents the efficiency score; where 1.00 represents efficient classes, and any value less than 1.00 represents inefficient classes. To be able to use the results of DEA with different classification algorithms, we converted the efficiency score with 1.00 values to "1" and other values to "0". By using the classification process we can predict the efficiency state of any new region of health centers given only its inputs and outputs. For the given data set we have 8 inefficient regions and 13 efficient regions. We apply set of classification algorithms using the WEKA tool [7] to the dataset and for each classifier.

1) Decision Tree Classifier: The decision tree classifier breaks down a complex decision making process into smaller processes [8], and apply this process on multistage to produce decisions. At each stage the classifier choses the best attribute that can be used to assess the decision. We used J48 classification algorithm in WEKA for the decision tree. Results of the decision tree classifier can be a set of rules that provide a factor that identify efficient and inefficient health centers.

2) Nave Bayes Classifier: The Nave Bayes classifier is based on the assumption that all attributes are independent from the class labels. Bayesian analysis is based on known prior probability of the class label. This means that we want to know how accurate we can classify new unknown classes regions to be efficient given a set of inputs and outputs. This classifier will give an indication of the amount of inputs and outputs needed for the new health centers in order to be efficient.

3) Support Vector Machine Classifier: The support vector machine classifier analyzes and recognizes patterns in data that are used for classification. The support vector machine maps data into a higher dimensional input space and finds
the optimal separating hyper plan in this space. Based on
the separation hyper plan the support vectors are identified
for each class and are used for classifying new unknown
class label instances. We used SMO classification algorithm
to support vector machine.

In our approach, the above mentioned classifiers will be
applied to the data sets several times so that it is easy to decide
whether to use input-output in classification or use input only.
Cross validation with 10 folds test option has been used for
all classifiers.

C. Using Classification algorithms with Input Variables Only

In this section we present the results of applying classification
algorithm using input variables only. This can be used in
predicting efficiency of new investment in the Health care
sector by predicting how efficient the health centers will be
based on the available inputs. Table III shows the results
of running the above classifiers on the data set using input
variables. Over the three classifiers; the best accuracy results
appeared when using the J48 decision tree classifier with
accuracy value 76.1905%. In the decision tree classifier; the
major splitting point in the tree is the X3 input variable, which
will be used to determine efficiency of any upcoming health
center.

Table III
RESULTS OF RUNNING CLASSIFIERS USING INPUT VARIABLES

<table>
<thead>
<tr>
<th>Name</th>
<th>Accuracy</th>
<th>F-Measure</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>76.1905%</td>
<td>0.747</td>
<td>0.762</td>
<td></td>
</tr>
<tr>
<td>Nave Bayes</td>
<td>52.381%</td>
<td>0.517</td>
<td>0.605</td>
<td>0.524</td>
</tr>
<tr>
<td>SMO</td>
<td>52.381%</td>
<td>0.426</td>
<td>0.358</td>
<td>0.524</td>
</tr>
</tbody>
</table>

Since our dataset is comparatively small to datasets in other
studies, we need to enlarge our dataset using the “SMOTE”
filter in WEKA tool to balance the classes. We applied this
filter several times to get a semi-balanced dataset with 84
instances; 52 of them now efficient and 32 inefficient. Then,
we repeated the same classification process used in the original
dataset with the classifiers; J48, Nave Bayes, and SMO on the
enlarged dataset.

Table IV shows the results of the classifiers on enlarged
dataset. As in the original dataset the best accuracy is resulted
from the J48 decision tree classifier with value 88.0952%.
The enlarged dataset tends to produce better results than the
original dataset.

Table IV
RESULTS OF RUNNING THE CLASSIFIER WITH ENLARGED DATASET

<table>
<thead>
<tr>
<th>Name</th>
<th>Accuracy</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>88.0952%</td>
<td>0.875</td>
<td>0.9</td>
<td>0.881</td>
</tr>
<tr>
<td>Nave Bayes</td>
<td>86.9048%</td>
<td>0.87</td>
<td>0.878</td>
<td>0.869</td>
</tr>
<tr>
<td>SMO</td>
<td>87.4711%</td>
<td>0.835</td>
<td>0.837</td>
<td>0.835</td>
</tr>
</tbody>
</table>

D. Using Classification algorithms with input and output variables

In this section we present the results of applying classification
algorithms using input and output variables. This
can be used in predicting efficiency classes of new unknown
health centers based on the set of available inputs and outputs.
Instead of applying the DEA models to calculate efficiency for
new health centers, the classifiers can predict the class labels.
Table V shows the result of running the above classifiers on
the data set using input and output variables. Over the three
classifiers the best accuracy results appear when using the J48
decision tree classifier with accuracy value 71.4286

Table V
RESULTS OF RUNNING THE CLASSIFIERS USING ALL VARIABLES

<table>
<thead>
<tr>
<th>Name</th>
<th>Accuracy</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>91.4286%</td>
<td>0.714</td>
<td>0.714</td>
<td>0.714</td>
</tr>
<tr>
<td>Nave Bayes</td>
<td>52.381%</td>
<td>0.517</td>
<td>0.524</td>
<td>0.605</td>
</tr>
<tr>
<td>SMO</td>
<td>61.9048%</td>
<td>0.581</td>
<td>0.591</td>
<td>0.619</td>
</tr>
</tbody>
</table>

The decision tree that represents the classification results
affected by the two output attributes (Y1, Y2) and one input
attribute (X1) is shown in Figure 1.

![Visualization Tree of J48 classifier](image)

Fig. 1. Visualization Tree of J48 classifier

Table VI shows the results of the classifiers after data
enlargement process. The best accuracy results from the SMO
classifier with 91.6667% accuracy value.

Table VI
RESULTS OF RUNNING CLASSIFIERS WITH ENLARGED DATASET

<table>
<thead>
<tr>
<th>Name</th>
<th>Accuracy</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>89.2857%</td>
<td>0.888</td>
<td>0.909</td>
<td>0.893</td>
</tr>
<tr>
<td>Nave Bayes</td>
<td>69.0476%</td>
<td>0.692</td>
<td>0.76</td>
<td>0.69</td>
</tr>
<tr>
<td>SMO</td>
<td>91.6667%</td>
<td>0.893</td>
<td>0.927</td>
<td>0.917</td>
</tr>
</tbody>
</table>

The decision tree that represents the classification results
affected by one output attribute (Y2) is shown in Figure 2.

From the above discussions and results of our proposed
model, we can conclude that any attempt to evaluate the
efficiency of healthcare agencies should follow our proposed
approach while making sure to use the J48 classifier as it gives
the best results in almost all experiments. Moreover, attempts
should also look directly to the decision tree that is resulted
from the J48 classifier which makes the evaluation process an
easy evaluation process.
VI. CONCLUSION

This paper applies data mining techniques; Classification, and clustering on the data that resulted from the DEA. This study shows that by using classification it is feasible to predict the efficiency of new investment in the health care sector. We are able to form rules that lead to efficient investment. Classification can be used to predict operating health care centers. To use classification with DEA results we used the data enlargement method to overcome the small amounts of data. In this paper we three classifiers have been used and studied. These classifiers are the support vector machine classifier (SMO), the Nave Bayes classifier, and the Decision tree classifier (J48). It is found that the J48 classifier achieved the best accuracies as shown in the above tables. Nevertheless; SMO achieved the best accuracy only in the last experiment. The worst accuracies resulted from Nave Bayes classifier in all experiments. Given that the J48 classifier achieved the best results, recommendations for using this classifier can be highlighted and given to any attempt to evaluate health care centers. Moreover, future evaluations for health care centers can rely on the decision tree obtained by the J48 classifier to build their decisions and conclusions.

REFERENCES
