Care Cost Prediction Model for Orphanage Organizations in Saudi Arabia

Huda N Alhazmi[1], Alshymaa Alghamdi, Fatimah Alajlani, Samah Abuayied, and Fahd M Aldosari

College of Computer and Information Systems, Umm Al-Qura University, Saudi Arabia

Summary

Care services are a significant asset in human life. Care in its overall nature focuses on human needs and covers several aspects such as health care, homes, personal care, and education. In fact, care deals with many dimensions: physical, psychological, and social interconnections. Very little information is available on estimating the cost of care services that provided to orphans and abandoned children. Prediction of the cost of the care system delivered by governmental or non-governmental organizations to support orphans and abandoned children is increasingly needed. The purpose of this study is to analyze the care cost for orphanage organizations in Saudi Arabia to forecast the cost as well as explore the most influence factor on the cost. By using business analytic process that applied statistical and machine learning techniques, we proposed a model includes simple linear regression, Naive Bayes classifier, and Random Forest algorithms. The finding of our predictive model shows that Naive Bayes has addressed the highest accuracy equals to 87% in predicting the total care cost. Our model offers predictive approach in the perspective of business analytics.

Key words:

Care cost, Orphanage organizations, Machine learning, Business analytic, Naive Bayes, Simple liner regression, Random Forest

1. Introduction

Today, business analytics is a major engine of competitiveness and development across sectors. Therefore, business analytics helps gain insights into making data-driven decisions [1] Lately, data has become a topic of particular interest because of the high but sometimes hidden, potential that it possesses [2]. In the recent past, the importance of data has been proven in almost all aspects of life. As a result, it provides managers with key information to predict the future of the organization's sales or costs while developing the present.

The purpose of this research is using business analytics approaches to build a cost prediction model for orphanage organizations. That, it focuses on studying the cost of residential care provided by orphanage institutes in Saudi Arabia. In Saudi Arabia, residential care is a service provided to abandoned children, which refers to those who born of unknown parents and they are called by orphans. Residential care for orphans and abandoned children in Saudi Arabia is provided by governmental institutions through Ministry of Social Affairs (MOSA), also by non-governmental organizations. In 1960, formal state welfare for orphans and abandoned children began. In 1962, MOSA issued the policy of social security in which orphans and abandoned children receive financial assistants [3].

A review of the studies on the cost estimation of orphans and abandoned children programs provide little information to document the costs and structure of costs [4]. A small number of studies have applied methods to evaluate the costs of these programs [5] [6], but these methods are not usually integrated into the routine practice of the organization's programs. To inform the program operation, costing analysis can be performed as a routine practice on the organization rather than reviewing activities that applied to get information and take decisions.

To fill this void, this study involves a new perspective to analyze the cost using machine learning algorithms to predict the care cost for orphanage organizations in Saudi Arabia, as well as determine the most influence factor of care on the cost. To achieve this goal, we built a model includes a simple linear regression model, Naive Bayes classifier, and Random Forest. To the best of our knowledge, this is the first study applies the cost analysis on residential care organization in Saudi Arabia using machine learning techniques. Findings from our analysis reveal a significant performance by Naive Bayes to predict minimum, maximum, and average cost.

The paper is structured as follows. Section 2 includes related studies. Section 3 presents the data description and preprocessing. The method is explained in Section 4. Section 5 provides the results and discussion. The last section concludes the work.

2. Related Work

This section presents a comparative review of the current studies in the area of orphan's care costs related to orphanage organizations in how they can estimate the total care cost. Moreover, this section also presents the new investigations of business analytics and machine learning in the area of total care cost prediction in order to establish a clear perspective of total care cost prediction for orphanage organizations. Table 1 shows a summary for

Manuscript received April 5, 2021
Manuscript revised April 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.4.13

this review. An orphanage organization refers to a residential center, home, or institution dedicated to the care of children whose parents have died and to children who are disconnected from their biological families for various family challenges [7]. Some of the underlying conditions that can contribute to children being put in orphanages in Saudi Arabia include deceased parents, abusive families who don't think of the repercussions of such abusiveness to the children, different forms of illness for example mental illness [8]. Another conditions where parents are not willing to take care of their children at all [9].

To date, there is not much known about how much it really costs to implement orphanage care services like provision of shelter, education, food and nutrition, psychological support, clothing and special needs, child protection programs, and case management. Authors in [10]. based on studying the orphanage organizations in Botswana, they found out that most care costs offered by the government institutions. The research revealed varying results within and across the service focal points in the organizations. According to [3], there are about 15 voluntary and 20 government-oriented institutions offering care for abandoned children and orphans in Saudi Arabia. From author's point of view, the care cost estimation is based on the duration of the orphan's stay in the orphanage. The length of stay is related to the emotional stability that orphans go through, which is very costly to revive and find solutions [9]. According to many studies conducted to investigate the care cost, the results varied from one service to the other. For example, [10] found out that the highest care cost for orphanage organizations in Botswana is based on education with 29%, then psychological support 24%, food and nutrition 23%, accommodation 8%, bedding and clothing 6%, health care 5%, special needs 1% and child protection takes the least percentage among other service. On the other hand, a study carried by [11] in Eritrea and Benin showed that the average economic annual cost per one orphan in an orphanage organization vary between the two countries. Neglecting the fundamental needs of children in orphanages today has high and expensive futuristic risks and repercussion not only to the orphanage organizations but also to the family and government at large [12]. According to the study of [13], very few orphanage organizations provided by the country with the best orphans' education grants, school fees, and other basic needs. Study [14] showed that in terms of review of the costs by resource, the highest orphanage care costs are linked to labor and program material, then transport, and finally the cost bestowed upon furniture and equipment. In terms of financial costs, most orphanage organizations receive donations from various donors in the name of clothes, food, medical care, etc. As the authors note in [15] the care cost for orphanage organization always hikes when the children experience emotional anguish which results in depression in most abandoned children. The study, for example, found that children who lived in the orphanage organization for more than two years had a lower IQ and retarded brain development which is very costly to fix. It highlighted high costs paid to therapists and counselors who tried to help children get back on track in the desired way. Rosenthal [16] revealed that the negative effects that orphanage organizations show on the children cost them in fact more than expected. Furthermore, [17] and [18] revealed that informal care given to orphans in home-based humanitarian where abandoned children experience family love reduces care cost for orphanage organizations. In such a setup, members were likely to act in the abandoned children's best interest, hence reducing the cost involved in taking care of children in orphanages. According to [19] and [15], the increased cost of health care is a significant global economic and public health problem.

Over time, statistical and machine learning techniques have been used for cost prediction in many domains. For example, many research have examined the estimation of healthcare costs by treating the issue either as a regression problem or as a classification problem [21], [22]. The study [17] estimated the increases in health care expenses for patients in the next year and recognized variables that make a major contribution to the projection. In particular, the researchers concentrated on the function of therapeutic care and other medical characteristics, including hospitalization and hospital visits. They viewed this issue as a binary classification challenge and projected that the net cost for patients would rise in 2015 based on its 2014 functionality. They also developed large-scale features and implemented innovative industrial features such as drug management to capture various trends in the results. The analysis of cost assessments of healthcare is also aimed to providing the most reliable estimation of the mean costs of treating the condition or determining the features of the patients/structure that driving the costs and obtaining an estimate of projected costs. This study [24] investigated the use of three state-of-the-art machine learning algorithms including regression tree, M5 model tree, and Random Forest to estimate the healthcare costs of patients based on the data of their past medical and expense experience. The results show that the Random Forest model was the best performing model to predict healthcare cost. Authors of [25] have also carried out their analysis from 2009 to 2013. The estimated cost of health insurance in 2013 were forecasted from 2009 to 2012 for comorbidity metrics.

Orphanage programs cost prediction is often a difficult issue from the perspective of data mining costs. To our knowledge, no prior studies have examined machine learning algorithms to predict or classify cost care for orphanages organizations. To fill this literature gap, our study offers a model includes machine learning algorithms

to predict the cost care for orphanages organizations in Saudi Arabia.

Table 1: Summary of related work

Paper	Author	Year	Model
[4]	Ahmed Albar		Shaping behavior model
[9]	Alghamdi H.	2016	Phenomenological qualitative model
[8]	Al-Jobair	2020	Analytical cross-sectional study
[14]	Bettmann	2013	Psychomotor model
[12]	King, Dewey	2015	Qualitative and quantitative model
[11]	Loren, L.P	2015	Open-plan model
[6]	Diane	2012	Study analysis
[17]	Munaaba	2004	Cross-sectional unmatched case control model
[15]	Eric Rosenthal	2017	Research analysis model
[22]	Stacia	2019	Measure evaluation
[16]	Subbarao et al.	2001	Case-study analysis
[7]	Taneja	2002	Prospective developmental assessment
[13]	Whetten et al.	2011	Correlation study report model
[23]	Sushmita et al.	2015	Predicting healthcare cost using machine learning
[10]	Kan et al.	2010	The potential of penalized linear regression models

3. Data Generating and Preprocessing

The study was started with data generating, cleaning, and preprocessing processes, which we explain in the following.

3.1 Data Description

The data used in this research is synthetic data, the synthesized data sets are created using generatedata website [26]. It's a free open-source platform written in JavaScript, PHP, and MySQL that can be used for fast production of large amounts of customized data across a broad range of software testing formats. We built the synthetic data which mimic the real data based on the information collected from different orphanage organizations in Saudi Arabia. Access to the data in these organizations is highly restricted due to the confidentiality. The dataset was created on November 1, 2020 in total of 2218 records. Table 2 presents the main dataset attributes. The description of the generated data is shown in Fig.1 and Fig. 2. In the figures, the x-axis represents the financial cost for service of orphanage care system, while the y-axis represents the number of orphans. All the cost range expressed in Saudi SAR.

Table 2: Dataset attributes description

Attribute Name	Description	Data Type
The cost based on the child's stay	Estimate cost for orphan's stay in the orphanage organization	String
Placement cost range form	The cost of recruiting orphans in the orphanage	Integer
Supervision replacement cost range form	The cost of monitoring orphans	Integer
Kids costs	Costs of orphan children	Integer
Salary of staff	The salaries of workers in the orphanage	Integer
Cost of education-foods-dresses	The cost of food, drink and clothing for each orphan	Integer
Cost of medical care	The cost of healthcare for each orphan	Integer

3.2 Data Preprocessing

The data preprocessing is essential step to address the data quality issues such as missing values, duplicate data, and outliers. The quality can be enhanced by providing additional preprocessing step. We started by changing the dataset columns names in the Excel file, in which each column was referred to a combination of 3 to 4 words. We shorten the names to simplify the file reading step in Python code as shown in Table 3. In the next step, we examined the existing of null values that conducted on Google coLab using Python programming language. Furthermore, we eliminated the outliers in some attributes using the Interquartile Range (IQR) role in Python. IQR is a measure of variability that divided a data set into quarterlies, then the quarterlies divide a rank-ordered data set into four equal parts. The values that separate the parts are called the first, second, and third quarterlies; and they are denoted by Q1, Q2, and Q3, respectively. Then the duplicate values are removed using python, as a result the total number of the records are decreased. Moreover, data reduction is one of the most important steps in any data preprocessing, in which it assists in reducing the data dimensionality. That lead to increasing the model performance, identifying the irrelevant data, simplifying the data visualizing, and enhancing the prediction results. Feature extraction is one of the well-known methods in

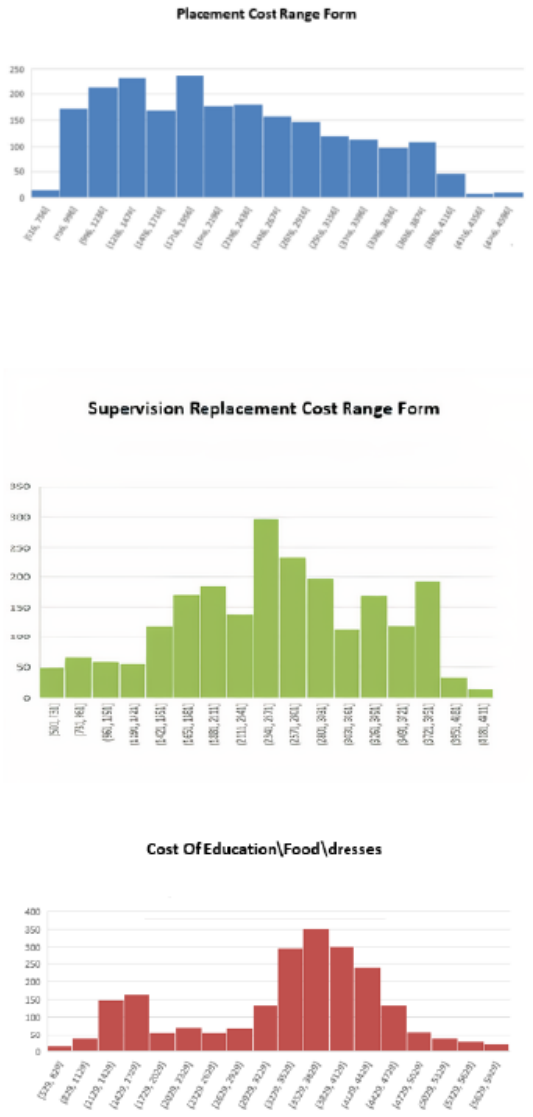

Fig. 1 Dataset description 1

data reduction. The extraction has been achieved using python script in Google coLab, we extracted six attributes from the dataset: CBCS, PCRF, SRCRF, KC, CEFD, CMC, which are used to obtain the total care cost (TCC).

Table 3: Dataset acronyms

Abbreviation	Full name of the shortcut
CBCS	The cost based on the child's stay
PCRF	Placement cost range form
SRCRF	Supervision replacement cost range form
KC	Kids costs
SOS	Salary of staff
CEFD	Cost of education, foods, dresses
CMC	Cost of medical care

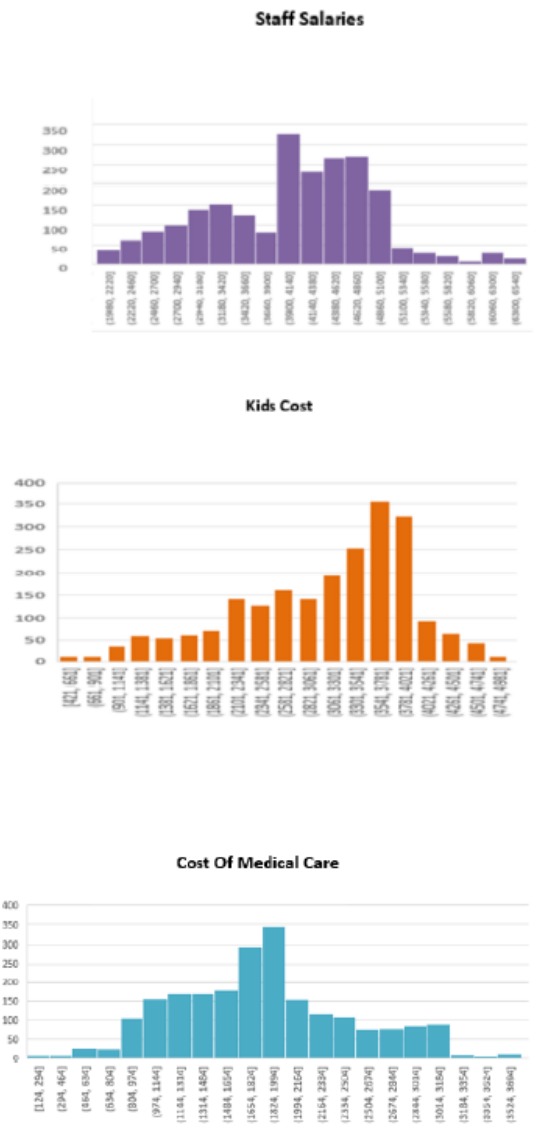

Fig. 2 Dataset description 2

4. Methods

This section introduces the research methodology and the proposed model for predicting the care cost. The Google coLab has been used as the main platform for Python 3 programming language, and Rapid Miner 9.8 for model implantation. Rapid Miner is a data science and machine learning software, that can deal with noisy data in a robotic manner [27]. In our approach, we first applied features selection, and then we used the simple linear regression to study the relations between different attributes in the dataset to obtain the most contributed

attribute on the cost. Finally, to predict the total care cost, multiple machine learning classification techniques are applied on the extracted attributes.

4.1 Simple Linear Regression

Linear regression is one of the most common supervised machine learning statistical analysis techniques; it is typically used to find the linear correlation between two or more response and predictor variables. This technique is classified into two types depending on the number of variables in the model: simple linear regression and multiple linear regression. In simple linear regression, one response variable corresponds to one predictor variable. Whereas in multiple linear regression, more than two response variables correspond to predictor variable. In our work, we used simple linear regression to study the correlation between the care total cost and other attributes in the datasets to obtain the most affected attributes on care total cost. The dataset was divided into 70%t rain and 30% test. Then, we calculated the Pearson's correlation coefficient (PCC) for each simple linear regression model. PCC determines the covariance and strength of a linear regression relation between two factors and is calculated as shown in Eq.(1)

$$PCC = covariance(X,Y)/(stdv(X) * stdv(Y)) \quad (1)$$

where **X** and **Y** are the response and predictor factors in the simple linear regression and **std** refers to the standard deviation of the two variables.

To evaluate the performance of the simple linear regression, we used the most used metrics such as mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). Eq.(2), Eq.(3), and Eq.(4) represent the three metrics, respectively.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - y_j^*| \quad (2)$$

$$MSE = \frac{1}{n}\sum_{j=1\to n}(y_j - y_j^*)^2 \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1\to n}(y_j - y_j^*)^2} \quad (4)$$

These regression metrics are the standard metrics for the determination of continuous variables and model accuracy.

4.2 Machine Learning Classification Techniques

Classification is one of the data mining techniques that aim to determine dataset records classes depending on the value of the predefined target attribute from the dataset. There are many classification methods that proved their benefits in prediction and identified various life problems. Ira Ekanda Putri [28] used two different classification methods: Naive Bayes classifier and support vector machine (SVM) to predict heart disease, in which SVM was addressed the best prediction performance. On the other hand, Dedy Hartama [29] used the C4.5 decision tree to predict patterns of interest of high school graduates. They record a good performance of C4.5 algorithm in classifying students into their interesting study program according to the given attributes in the dataset. In our work, we applied two classification methods, naming, Naive Bayes classifier and Random Forest to predict the total care cost. The dataset has been divided into 70%train and 30% test. In the following, we will discuss the two methods.

4.2.1 Naive Bayes Classifier

Naive Bayes depends on probability theory for finding the best classification class for the target attribute. The main advantages of Naive Bayes, that it is a robust method for isolating noise points in the dataset, handling the missing values by ignoring the missing values during calculating the probability. Also, it has the ability for dealing with the irrelevant attributes. In this work, we applied Naive Bayes on all attributes using Rapid Miner 9.8. The data has been split into 70% training and 30% testing and the important feature in the model is KC. The model predicts the maximum, minimum, and average cost. On the other hand, we also applied Naive Bayes on the three important attributes: TCC, CMC, CEFD, that have been selected from the simple linear regression model.

4.2.2 Random Forest

A Random Forest is a combination of predictors trees, in which the value of each tree depends on a random vector sampled separately and with the same distribution for all trees in the Random Forest. The algorithm was discovered by Leo Breiman [30]. Moreover, a study by Ramón Díaz-Uriarte [31], has used a Random Forest algorithm to select the relevant gene expression for sample classification, such as distinguish between cancer and non-cancer patients. The author argued that Random Forest recorded an optimal performance in predicting and classification gene expression, and it can deal with noisy data such as the microarray data. For our model, we have applied the Random Forest on our dataset using Rapid Miner 9.8. We selected three attributes TCC, CMC, CEFD

to apply the model, and we choose about 100 trees with maximum depth =2 with error rate = 39.4% as shown in figure Fig. 5.

4.2.3 Classification Models Evaluation

To evaluate the performance of the classification models, we used four performance metrics, which are accuracy, recall, precision, and F1 score

- **Accuracy**: Defined as the ratio of the correct predicted observation of the model to the total of observations.
- **Recall**: Defined as the ratio of the correct predicted of the positive observations to the total observations in the actual class.
- **Precision**: Defined as the ratio of the correct predicted of the positive observations to the total predicted positive observations.
- **F1-Score**: Presented the average of Precision, Recall, and described by Eq.(5)

$$F1-Score = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (5)$$

5. Results and Discussion

In this section, we will illustrate some experimental results of our study. We summaries the findings of the simple linear regression and the machine learning classification techniques that used to predict the care cost and find out the most influence factor on the cost.

5.1 Simple Linear Regression

Fig. 3 shows five different simple linear regression model that represent the relation between the dataset attributes and TCC. The result of the calculated Pearson's correlation coefficient (PCC) for simple linear regression model is shown in Table 4. According to the numbers illustrated in the table, we note that the results lead to the existing of positive relations between TCC and the other attributes. This relationship becomes strong between TCC, CMC and CEFD, which is a good indication to use these two attributes in building the machine learning classification models.

Table 5 shows the results of the three metrics on our simple linear regression models. Moreover, the errors rate differs from one simple linear regression model to another. We can see that CMC has addressed the highest values for the three metrics as illustrated in Fig. 3.

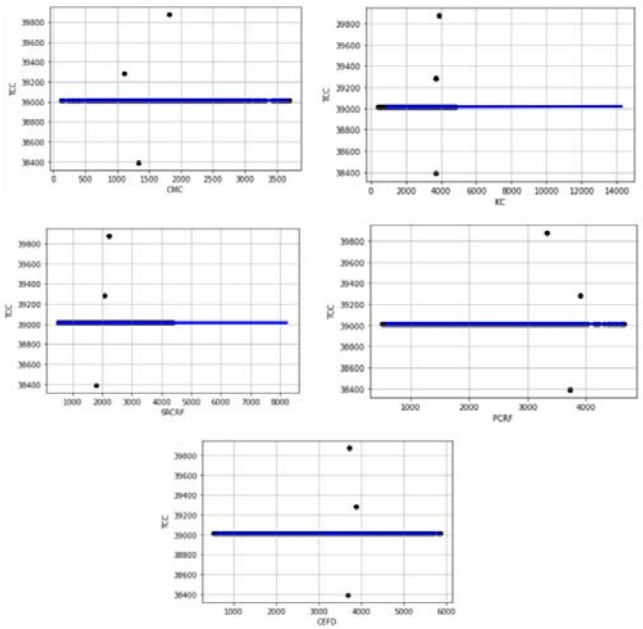

Fig. 3 Simple linear regression correlation result among the dataset attributes and care total cost

5.2 Naive Bayes Classifier

Table 6 shows the result of the Naive Bayes classifier for predicting the total care cost. BA registered 87% accuracy when used with the all attributes. We also applied BA on the three important attributes: TCC, CMC, CEFD, that have been selected from the simple linear regression model. The model predicted the minimum, the maximum, and the average cost as shown in Fig. 4. The classifier shows less accuracy about 65%.

Table 4: Simple linear regression PCC results

Response	Predictor	PCC	Direction	Strength
TCC	CMC	0.8	Positive	Strong
TCC	CEFD	0.9	Positive	Strong
TCC	KC	0.4	Positive	Weak
TCC	SRCRF	0.4	Positive	Weak
TCC	PCRF	0.1	Positive	Weak

Table 5: Regression metrics values.

Response	Predictor	MAE	MSE	RMSE
TCC	CMC	159	543	233
TCC	CEFD	1.44	583	24.15
TCC	KC	1.49	583	24.16
TCC	SRCRF	1.43	583	24.15
TCC	PCRF	1.48	584	24.2

Table 6 Naïve Bayes cost result for all attribute

Features	Minimum	Maximum	Average
KC	421	14267	3127
CEFD	529	5872	3324
CMC	750	1910	1500
PCRF	516	4652	2177
SOS	300	6503	4041
SRCRF	501	8237	2539

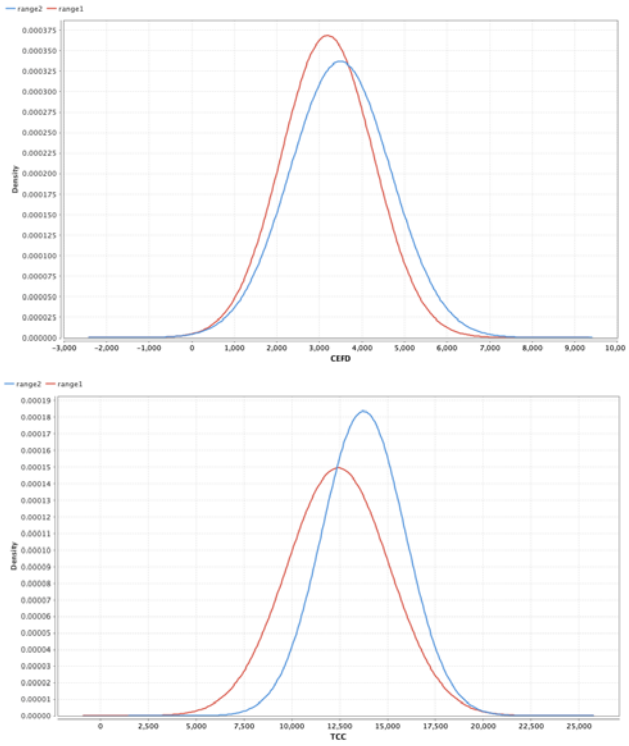

Fig. 4 Naive bayes result for CEFD and TCC

5.3 Random Forest

Figure Fig. 5 illustrates the result of Random Forest algorithm. The feature that affected the most the predicting process is TCC = 0.189 then CEFD =0.185. The result shows that random classifier has 63% accuracy. The minimum, maximum and average cost are shown in Table 7.

Table 7: Random forest cost result

Features	Minimum	Maximum	Average
TCC	4947	24948	13001
CEFD	529	5872	3324
CMC	750	1910	1500

5.4 Performance Result

The results of the performance of the prediction using the performance metrics: accuracy, F1 score, recall, and precision are presented in Table 8. Naive Bayes for all attributes in the dataset exhibits the highest accuracy rate equals to 87%. Followed by the performance of the same model with the three attributes 65%, then Random Forest recorded accuracy rate equals to 63%. Remarkably, Naive Bayes exhibits the best performance when applied on all attributes. However, the three models show good performance for prediction the care cost and provided the estimated cost as maximum, minimum, and average value.

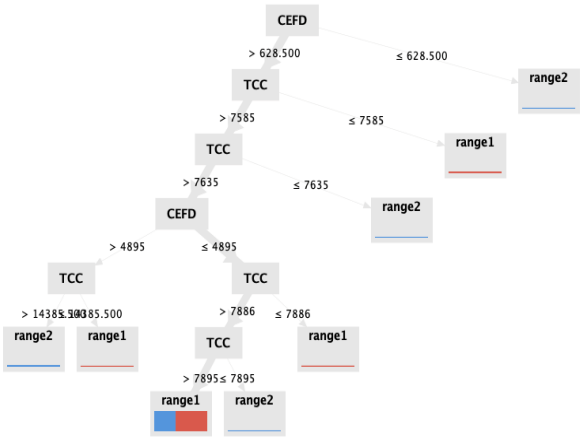

Fig. 5 Random Forest result.

Table 8: Classification models' evaluation metrics

Model	Accuracy	F1 score	Recall	Precision
Naive Bayes (all)	87%	91%	92%	90%
Naive Bayes (3)	65%	70%	70%	70%
Random Forest	63%	76.5%	99.5%	62%

5.5 Limitations

This study limits its scope to the costs per child. The study doesn't consider the quantity and the quality of services to the children, for instance, the number of years of staying or schooling. Also, the benefits of the services to the children, which vary in terms of health and educational outcomes. However, regardless of these challenges, the study measures the average annual costs per child which is the best measure of overall economic cost.

6. Conclusion

A successful use of data through the process of business analytics help organization to gain insights and operate efficiently. Care cost predicting is one of the most important topics in the care services area, in which it can affect the care services process and goals. In this research, we have focused on predicting the care cost of orphanage organizations in Saudi Arabia. However, during the research, we have faced many limitations in gathering real data from orphanage organizations due to confidentiality and privacy. Thus, the data that used in this research is synthetic data gathered from "generatedata". Furthermore, we proposed a predictive model includes a simple linear regression and two machine learning algorithm, naming, Naive Bayes classifier, and Random Forest. The result shows that Naive Bayes exhibits the significant performance when it applied on the all attributes of the dataset. Our model can be applied on real data which help orphanage organizations to gain insights about their care system that are seldom considered from the perspective of business analytics.

References

[1] G. Cao, Y. Duan and G. Li, "Linking business analytics to decision making effectiveness: A path model analysis," IEEE Transactions on Engineering Management, vol. 62, no. 3, pp. 384–395, 2015.

[2] Z. A. Al-Sai and L. M. Abualigah, "Big data and e-government: A review," in 2017 8th international conference on information technology (ICIT). IEEE, pp. 580–587, 2017.

[3] A. A. Albar, "Residential child and youth care in saudi arabia: a case study of abandoned children and young people," GLOBAL PERSPECTIVES, p. 156, 2016.

[4] B.A. Larson, N. Wambua, "How to calculate the annual costs of ngo-implemented programmes to support orphans and vulnerable children: a six-step approach," Journal of the International AIDS Society, vol. 14, 14–59, 2011.

[5] C. Desmond, J. Gow, H. Loening-Voysey, T. Wilson B, Stirling, "Approaches to caring, essential elements for a quality service and cost-effectiveness in south africa," Evaluation and Program Planning, vol. 25, PP. 447–458, 2002.

[6] P. Hutchinson and T.R. Thurman, "Analyzing the cost effectiveness of interventions to benefit orphans and vulnerable children: evidence from Kenya and Tanzania," 2009.

[7] V. Taneja, S. Sriram, R. Beri, V. Sreenivas, R. Aggarwal, and R. Kaur, "'not by bread alone': impact of a structured 90-minute play session on development of children in an orphanage," Child: Care, Health and Development, vol. 28, no. 1, pp. 95–100, 2002.

[8] A. M. Al-Jobair, S. A. Al-Sadhan, A. A. Al-Faifi, R. I. Andijani, and S. K. Al-Motlag, "Medical and dental health status of orphan children in central saudi arabia," Saudi Med J, vol. 34, no. 5, pp. 531–536, 2013.

[9] H. O. Alghamdi, "Relationship between behavioural disorders and social cognition among orphans in saudi arabia." International Education Studies, vol. 13, no. 6, pp. 85–95, 2020.

[10] C. Formson and S. Forsythe, "A costing analysis of selected orphan and vulnerable children (ovc) programs in botswana," Health Policy Initiative, Task Order, vol. 1, 2010.

[11] M. Prywes, D. Coury, G. Fesseha, G. Hounsounou, A. Kielland et al., Costs of projects for orphans and other vulnerable children: Case studies in Eritrea and Benin. World Bank, Washington, DC, 2004.

[12] L. P. Loren, "Abandoning the orphans: An open access approach to hostage works," Berkeley Tech. LJ, vol. 27, p. 1431, 2012.

[13] N. King, C. Dewey, and D. Borish, "Determinants of primary school non-enrollment and absenteeism: results from a retrospective, convergent mixed methods, cohort study in rural western kenya," PLoS One, vol. 10, no. 9, p. e0138362, 2015.

[14] R. Whetten, L. Messer, J. Ostermann, K. Whetten, B. W. Pence, M. Buckner, N. Thielman, K. O'Donnell, P. O. for Orphans (POFO) Research Team et al., "Child work and labour among orphaned and abandoned children in five low and middle income countries," BMC International Health and Human Rights, vol. 11, no. 1, p. 1, 2011.

[15] J. E. Bettmann, J. M. Mortensen, and K. O. Akuoko, "Orphanage caregivers' perceptions of children's emotional needs," Children and Youth Services Review, vol. 49, pp. 71–79, 2015.

[16] E. Rosenthal, "A mandate to end placement of children in institutions and orphanages: The duty of governments and donors to prevent segregation and torture," Protecting Children Against Torture in Detention: Global Solutions for a Global Problem (2017), 2017.

[17] K. Subbarao, A. Mattimore, and K. Plangemann, Social protection of Africa's orphans and other vulnerable children: Issues and good practice program options. World Bank, Africa Region, 2001.

[18] F. N. Munaaba, J. Owor, P. Baguma, S. Musisi, F. Mugisha, D. Muhangi, V. I. Matovu, J. R. Owor, E. Ezati, J. Okumu et al., "Comparative studies on orphans and non-orphans in uganda," Center for International Health and Development, Tech. Rep., 2004.

[19] D. Pritchard, A. Petrilla, S. Hallinan, D. H. Taylor Jr, V. F. Schabert, and R. W. Dubois, "What contributes most to high health care costs? health care spending in high resource patients," Journal of Managed Care & Specialty Pharmacy, vol. 22, no. 2, pp. 102–109, 2016.

[20] Z. Hu, S. Hao, B. Jin, A. Y. Shin, C. Zhu, M. Huang, Y. Wang, L. Zheng, D. Dai, D. S. Culver et al., "Online prediction of health care utilization in the next six months based on electronic health record information: a cohort and validation study," Journal of medical Internet research, vol. 17, no. 9, p. e219, 2015.

[21] D. Bertsimas, M. V. Bjarnad´ottir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G. Wang, "Algorithmic prediction of health-care costs," Operations Research, vol. 56, no. 6, pp. 1382–1392, 2008.

[22] S. Tamang, A. Milstein, H. T. Sørensen, L. Pedersen, L. Mackey, J.-R. Betterton, L. Janson, and N. Shah,

"Predicting patient 'cost blooms' in denmark: a longitudinal population-based study," BMJ open, vol. 7, no. 1, p. e011580, 2017.

[23] A. M. J¨odicke, U. Zellweger, I. T. Tomka, T. Neuer, I. Curkovic, M. Roos, G. A. Kullak-Ublick, H. Sargsyan, and M. Egbring, "Prediction of health care expenditure increase: how does pharmacotherapy contribute?" BMC health services research, vol. 19, no. 1, p. 953, 2019.

[24] S. Sushmita, S. Newman, J. Marquardt, P. Ram, V. Prasad, M. D. Cock, and A. Teredesai, "Population cost prediction on public healthcare datasets," in Proceedings of the 5th International Conference on Digital Health 2015, 2015, pp. 87–94.

[25] H. J. Kan, H. Kharrazi, H.-Y. Chang, D. Bodycombe, K. Lemke, and J. P. Weiner, "Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults," PloS one, vol. 14, no. 3, p. e0213258, 2019.

[26] Generatedata.com, 2021, URL: https://www.generatedata.com/.

[27] Rapidminer,2021, URL: https://rapidminer.com/.

[28] T. Putri, S. Sriadhi, R. Sari, R. Rahmadani, and H. Hutahaean, "A comparison of classification algorithms for hate speech detection," in IOP Conference Series: Materials Science and Engineering, vol. 830, no. 3. IOP Publishing, 2020, p. 032006.

[29] D. Hartama, A. P. Windarto, and A. Wanto, "The application of data mining in determining patterns of interest of high school graduates," in Journal of Physics: Conference Series, vol. 1339, no. 1. IOP Publishing, 2019, p. 012042.

[30] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001

[31] R. D´ıaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," BMC bioinformatics, vol. 7, no. 1, p. 3, 2006.