International Journal of Information Sciences and Application.
ISSN 0974-2255 Volume 10, Number 1 (2018), pp. 1-6
© International Research Publication House
http://www.irphouse.com

Machine Learning Algorithm Application in Predicting Children Mortality: A Model Development

Las Johansen B. Caluza

Leyte Normal University, Tacloban City, Philippines

Abstract

World Health Organization (WHO) has shown numerous report regarding child mortality and these are supported by many kinds of literature. However, despite these various reports and studies predicting this phenomenon was not yet taken into consideration, which this research tend to address. This study utilized data mining technique using decision tree called J48 algorithm in classifying child mortality rate, life expectancy at birth, annual population growth, and the gross domestic product. Results revealed that annual population growth is highly correlated in predicting child mortality and generate three distinct rules. Finally, the model generated has of high acceptability with 97.4% ROC curve result of the three classes in predicting child mortality under five years old.

Keywords: J48 Algorithm, model development, decision tree, Child Mortality, Poverty, social science, Philippines.

1.0 INTRODUCTION

Child mortality remains debatable in the world as it questions the government's implementation on poverty alleviation and citizen's welfare. In research findings revealed that Child survival strategies should direct resources toward the leading causes of child mortality, with attention focusing on infectious and neonatal causes (Liu et al., 2012). WHO moreover, UNICEF's Child Health Epidemiology Reference Group (CHERG) has published a series of estimates about the distribution of causes of child death during the past decade, during which time estimation methods and the quality and quantity of input data have improved. Estimation of mortality level in children below the age of five has a profound impact on some demographic parameters (Mekonnen, Ayalew, & Dejene, 2017). Childhood mortality data are also useful in assessing the impact of child survival programs and identifying child populations that are at risk. It is on this premise that the research tends to address by developing a model to predict child mortality as a target variable.

2.0 METHODOLOGY

2.1 Research Design

This research utilized descriptive-correlation design using data mining techniques specifically the decision tree algorithm. Supervised Machine Learning (SML) is one of the most popular applications of machine learning. SML is a machine learning task used to infer labeled datasets (Mohri, Rostamizadeh, & Talwalkar, 2012). An example of SML task is the pattern classification tasks. In this example, predictive modeling is the general concept of constructing a model that is capable of making predictions, in which such a model includes a machine learning algorithm that finds out specific properties from a training dataset to get those predictions (Raschka, 2014).

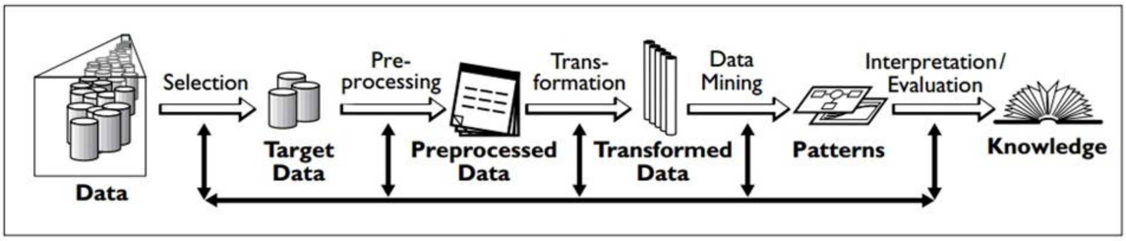

Figure 1: Overview of the steps constituting the KDD process

Source: http://shawndra.pbworks.com/f/The%20KDD%20process%20for%20extracting%20useful%20knowledge%20from%20volumes%20of%20data.pdf

2.2 Research Process

This study in anchor on the Knowledge Discovery in Database (KDD) from the work of Fayyad, Piatetsky-Shapiro, & Smyth (1996). The KDD process is interactive and iterative (with many decisions made by the user), involving numerous steps, summarized as:

1. Learning the application domain: includes relevant prior knowledge and the goals of the application.
2. Creating a target dataset: involves selecting a dataset or focusing on a subset of variables or data samples on which discovery is to be performed.
3. Data cleaning and preprocessing: includes basic operations, such as removing noise or outliers if appropriate, collecting the necessary information to model or account for the noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes, as well as deciding DBMS issues, such as data types, schema, and mapping of missing and unknown values.
4. Data reduction and projection: includes finding useful features to represent the

data, depending on the goal of the task, and using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations of the data

5. Choosing the function of data mining: includes deciding the purpose of the model derive from the data mining algorithm (e.g., summarization, classification, regression, and clustering).
6. Choosing the data mining algorithm(s): includes selecting method(s) to be used for searching for patterns in the data, such as deciding which models and parameters may be appropriate (e.g., models for categorical data are different from models on vectors over real) and matching a particular data mining method with the overall criteria of the KDD process (e.g., the user may be more interested in understanding the model than in its predictive capabilities).
7. Data mining includes searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, clustering, sequence modeling, dependency, and line analysis.
8. Interpretation includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable by users.
9. Using the discovered knowledge: includes incorporating this knowledge into the performance system, taking actions based on the knowledge, or simply documenting it and reporting it to interested parties, as well as checking for and resolving potential conflicts with previously believed (or extracted) knowledge.

Finally, this study utilized a secondary data taken from the World Bank Dataset that include information from 1960 to 2016.

2.3 Variable Definition

a) *Annual population growth rate* for year t is the exponential rate of growth of midyear population from year t-1 to t, expressed as a percentage. The population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship.

b) *Gross Domestic Product (GDP)* at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the goods. It is calculated without making deductions for depreciation of fabricated assets or depletion and degradation of natural resources. Data are in current U.S. dollars. Dollar figures for GDP are converted from domestic currencies using single year official exchange rates. For a few countries where the official exchange rate does not reflect the rate effectively applied to actual foreign exchange transactions, an alternative conversion factor is used.

c) *Life expectancy* at birth used here is the average number of years a newborn is expected to live if mortality patterns at the time of its birth remain constant in the future. It reflects the overall mortality level of a population and summarizes the mortality pattern that prevails across all age groups in a given year. It is calculated in a period life table which provides a snapshot of a population's mortality pattern at a given time.

d) *Mortality Rate Under-five* is the probability per 1,000 that a newborn baby will die before reaching age five, if subject to age-specific mortality rates of the specified year.

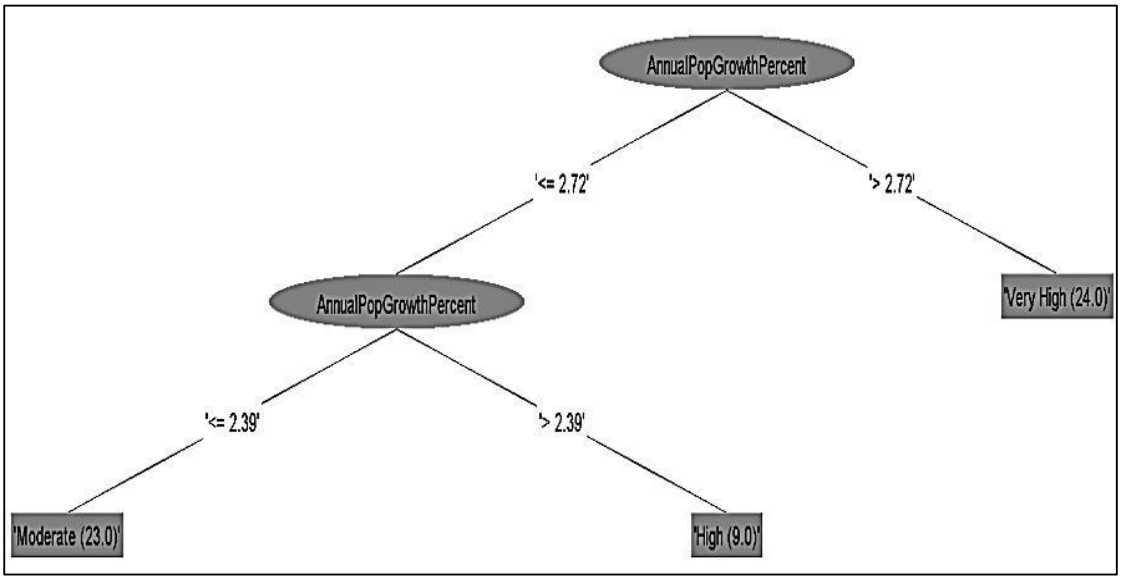

Figure 2: Pruned Tree of Child Mortality Model

3.0 RESEARCH OUTPUT

As shown in the figure above, Annual Population Growth is highly correlated with the Child mortality as it has the highest Information Gain resulting to *Correlation Coefficient* of 96.42% of the model.

3.1 Generated Rule:

Rule1: IF (AnnPopGrowthPercent <= 2.72 AND (AnnPopGrowthPercent <= 2.39 THEN

Mortality Rate is ***Moderate***

Rule 2: IF (AnnPopGrowthPercent > 2.39 THEN Mortality Rate is ***High***

Rule 3: IF (AnnPopGrowthPercent > 2.27 THEN Mortality Rate ***is Very High***

Table 1: Confusion Matrix

A	**B**	**C**	**🡐Classified as**
24	0	0	A=Very High
1	**8**	0	B=High
0	1	**22**	C=Moderate

Table 1 shows the classification of the Child Mortality using 10-fold Cross-Validation. It revealed that:

1. There are 24 correctly classified as *Very High*;
2. There are eight (8) correctly classified as *High,* and one misclassified as *very high*, and finally
3. There are 22 correctly classified as *Moderate,* and one misclassified as *high*.

   To validate the model, the ROC Curve of the generated tree was used.

Table 2: ROC Curve

ROC	**Class**
0.984	Very high
0.934	High
0.978	Moderate
0.974	**Weighted Average**

Table 2 shows that three classes of Child Mortality are very close to 1 or 100%, resulting in an **excellent and acceptable model.**

4.0 CONCLUSION

The annual population growth shows a very significant role in achieving a model in this study. Also, the generated model revealed very high acceptability of the model with a weighted average of 97.4% value of all classes of mortality rate under five years old. Finally, the generated model is an excellent baseline in predicting child mortality under five years old in the Philippines. However, further studies may be conducted to enhance this model.

REFERENCES

[1] 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.

[2] Mekonnen, Y., Ayalew, T., & Dejene, A. (2017). Estimation of child mortality in Addis Ababa. *The Ethiopian Journal of Health Development (EJHD), 9*(3).

[3] Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning. MIT Press.*

[4] Liu, L., Johnson, H. L., Cousens, S., Perin, J., Scott, S., Lawn, J. E., ... & Mathers, C. (2012). Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *The Lancet, 379*(9832), 2151-2161.

[5] Raschka, S. (2014). *Predictive modeling, Supervised machine learning, and pattern classification.* Machine Learning. Retrieved from http://sebastianraschka.com/Articles/2014_intro_supervised_learning.html on 5/21/2017

[6] WHO. *WHO mortality database*: tables. http://www.who.int/

[7] healthinfo/morttables/en/ retrieved 11/12/2016.