Analysing Citizen Profiles with Data Mining

W. A. Mohotti

139172C

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of Degree of Master of Science in Information Technology.

April 2016

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student

W. A Mohotti

Signature of Student

.....

Date: 10 - 04 - 2016

Supervised by

Name of Supervisor

S. C. Premaratne

Signature of Supervisor

.....

Date: 10/04/2016

Dedication

We dedicate the output of this research work and thesis to government policy setters who are trying to uplift lifestyles of Sri Lankans by addressing the citizens' problems. Also we specially dedicate this system to all those who generously contributed their valuable time, advising and helping in doing this research, specially to my supervisor Mr. S.C. Premaratne. In Sri Lanka, area of analyzing citizen profiles is not effectively done with appropriate techniques. It is with this thought in mind that we have done this research. I hope the research and the findings described below will provide a useful insight for analyzing lifestyle data to provide solutions to issues attach with citizens' life patterns.

Acknowledgement

First of all I would like to thank my project supervisor Mr. S.C. Premaratne who spent his valuable time for guiding this research to make it a success. Furthermore, my next big thank goes to Prof. Asoka Karunandha who taught us Research Methodology and Literature Review and thesis writing subjects which were the basis for this research.

Not only that my thanks should go to all the lecturers in M.Sc in Information Technology degree program of Faculty of IT, who gave their hands to sharpen our knowledge and ideas throughout these two years as they were the illumination which lit up our path ways to success.

Apart from the people who were directly involved, many more helped to make this project a success. Department of Census and Statistics contribute to this research by giving their HIES 2012/2013 data. So thank you all for your great support. Finally, I would like to thank all the batch mates of the M.Sc. in IT degree program who gave their valuable feedbacks to improve the results of the research.

Abstract

There is an exponential growth in issues attached with lifestyles of Sri Lankans over the past few decades. These may contribute to low down the life quality within citizens. In Sri Lanka, there are no adequate researches in the field of analyzing lifestyle data. Though there are few researches which have analyzed the causes for the socio-economic problems, such approaches are not capable of handling big data effectively and not efficient in predicting or describing the issues attach with lifestyle.

Hence, the research has been conducted to analyze citizen profiles in effective way to explore different lifestyle issues. It is hypothesized that analyzing citizen profiles can be done through data mining according to the output want to achieve through predictive or descriptive techniques. The solution takes HIES data set as the input and predict the factors attach with a particular lifestyle issue or describe specific lifestyle issue with its associative causes. Having received the input, this approach preprocessed the dataset to remove the anomalies. Then build data models to represent the lifestyle issue by extracting attributes from HIES data set. Then proceed with pattern recognition for the issues. The important patterns recognized through this approach will be useful for government and policy makers to set up appropriate government policies to uplift the life quality of citizen. The overall design of the research consists of two research question, one question used predictive mining based solution and other one is based on descriptive mining. Classification in data mining was used in finding the factors and their relationships that associated with no schooling and dropouts as those were predictive mining tasks. Clustering is used to explore the relationship between chronic diseases and family.

The overall research is designed using WEKA data mining tool and SPSS statistical tool. Finally, the data models build for citizen profile analysis using data mining techniques are evaluated for their performance using measurements such as value for accuracy, error rate, training time, TP rate, FP rate and ROC measurement.

Contents

List of Figures

School Dropouts

List of Tables