Data Analytics for YRBS (Youth Risk Behavior Survey) Data using Machine Learning and Data Mining Techniques

Ana K. Ocampo and Cyril S. Ku (Computer Science), Corey H. Basch and Aleksandar Kecojevic (Public Health); William Paterson University

Outline

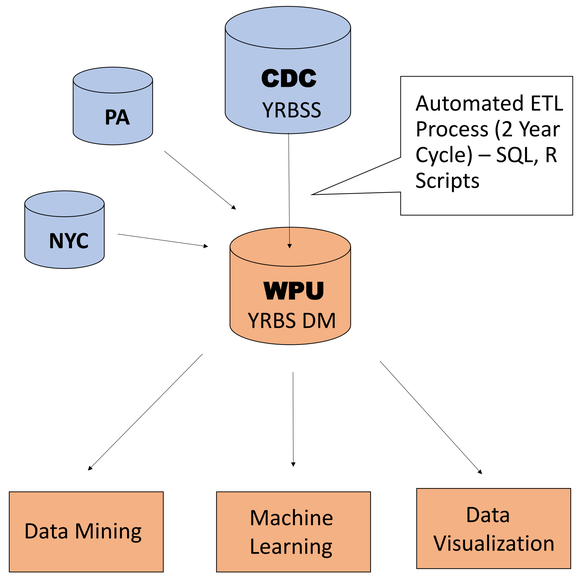
Abstract

A data analytics environment has been created in the Department of Computer Science at William Paterson University. This environment consists of a data mart of YRBS (Youth Risk Behavior Survey) data from CDC (Centers for Disease Control and Prevention) and several machine learning and data mining tools. This presentation outlines our inter-disciplinary research collaboration with the Department of Public Health on the YRBS data. The objectives are to find interesting data patterns, discover hidden relationships, and perform predictive analyses on the YRBS data. The architecture of the data analytics environment as well as the preliminary results of using the R programming language will be presented.

Project Description

- An inter-disciplinary research project:
  - Department of Public Health – Dr. Corey H. Basch and Dr. Alex Kecojevic
  - Department of Computer Science – Dr. Cyril S. Ku and Ms. Ana K. Ocampo (Research Assistant)
- Data Warehouse
  - YRBSS (Youth Risk Behavior Surveillance System) from CDC (Centers for Disease Control and Prevention)
- Data Analytics Environment at William Paterson University
  - MySQL server (YRBS Data Mart)
  - MySQL Workbench
  - R Studio (R Console/RGui)
  - Other software will be considered in the future such as WEKA (Waikato Environment for Knowledge Analysis)

Client/Server Architecture



YRBSS (CDC)

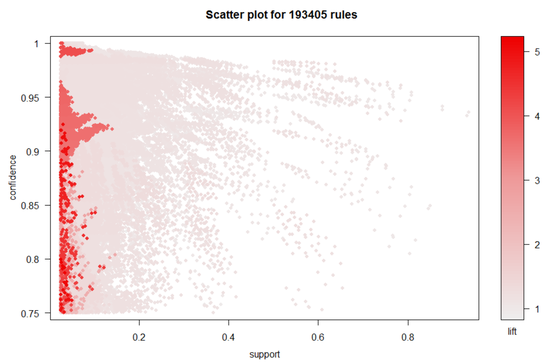
- Developed in 1990 to monitor priority health risk behaviors that contribute to the leading causes of death, disability, and social problems among youth and young adults in the U.S.
  - Behaviors that contribute to unintentional injuries and violence
  - Sexual behaviors related to unintended pregnancy and sexually transmitted infections, including HIV infection
  - Alcohol and other drug use
  - Tobacco use
  - Unhealthy dietary behaviors
  - Inadequate physical activity
  - Monitors the prevalence of obesity and asthma and other health-related behaviors plus sexual identity and sex of sexual contacts
- From 1991 through 2015, the YRBSS has collected data from more than 3.8 million high school students in more than 1,700 separate surveys

YRBS Data Mart (WPU)

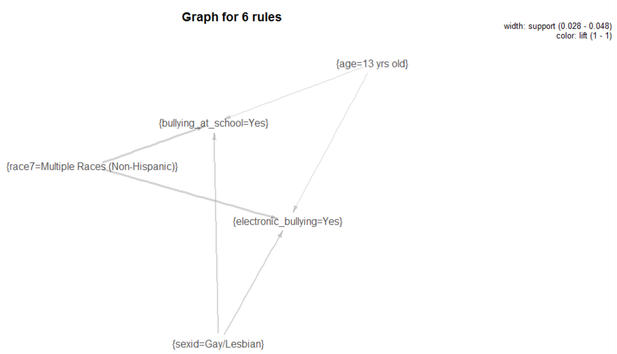
- Aggregated subsets of YRBSS from CDC and survey data from New York City and Pennsylvania
  - National (2011, 2013, 2015)
  - New Jersey (2011, 2013)
  - New York (2011, 2013, 2015)
  - New York City (2011, 2013, 2015)
  - Pennsylvania (2015)

Preliminary Study

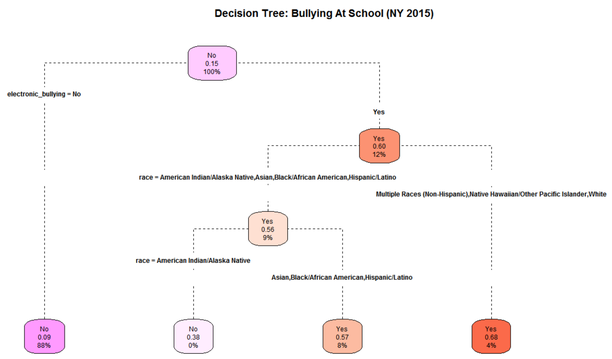
- Used a subset of the New York dataset (2015)
- Focused on machine learning algorithms to explore relationships and patterns between variables in the dataset
- Performed association mining rule to discover frequent co-occurring associations among variables (focused on two variables: bullied at school, electronic bullying)
- The following diagram shows the association scatter plot generated after running the Apriori Algorithm, showing only association rules with confidence > 0.75



- The following association graph shows the top 6 association rules by confidence for those who answered "Yes" to both bullying questions



- Used classification to indicate if a student is bullied at school based on their race, and their answer from the question of being electronically bullied (yes/no)
- The following decision tree was generated to show the results



Future Research Plan

- Use various data mining and machine learning (neural network algorithms) techniques of classification, association, and clustering analyses on the YRBS data
- The goal of the research is to use knowledge discovery approach instead of the traditional statistics-based approach to find interesting or hidden relationships, including anomaly detection and data prediction

Acknowledgement

- This research was supported in part by the ART (Assigned Released Time for Research) program, Office of the Provost; and in part by the Student Research Funds of the College of Science and Health, William Paterson University.

Bibliography

- Brener, N. D., Kann, L., Shanklin, S., Kinchen, S., Eaton, D. K., Hawkins, J., and Flint, K. H., "Methodology of the Youth Risk Behavior Surveillance System – 2013," *CDC MMWR Recommendations and Reports,* Vol. 62, No. 1, March 1, 2013
- http://www.cdc.gov/healthyyouth/data/yrbs/index.htm
- Torgo, L., *Data Mining with R: Learning with Case Studies,* 2nd Edition, CRC Press, 2017
- Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J., *Data Mining: Practical Machine Learning Tools and Techniques,* 4th Edition, Morgan Kaufmann, 2017