William Paterson University

WELCOME

Will. Power.

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

Ana K. Ocampo and Cyril S. Ku
Department of Computer Science

Computer Science Advisory Board Meeting
May 5, 2017

Overview

- Project Description
- Research Goal
- YRBSS (CDC)
- YRBS Data Mart (WPU)
- Client/Server Architecture
- Preliminary Studies
- Future Research Plan
- Acknowledgement
- Bibliography

Project Description

- An inter-disciplinary research project:
  - ➢ Department of Public Health (WPU) – Dr. Corey H. Basch and Dr. Alex Kecojevic
  - ➢ Department of Computer Science (WPU) – Dr. Cyril S. Ku and Ms. Ana K. Ocampo (Research Assistant)
  - ➢ Department of Health and Behavior Studies (Columbia University) – Dr. Charles E. Basch
- Data Warehouse:
  - ➢ YRBSS (Youth Risk Behavior Surveillance System) from CDC (Centers for Disease Control and Prevention)
- Data Analytics Environment at WPU:
  - ➢ MySQL Server (YRBS Data Mart)
  - ➢ MySQL Workbench
  - ➢ R Studio (R Console/RGui)
  - ➢ WEKA (Waikato Environment for Knowledge Analysis)

Research Goal

- 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
    - ➢ Collaborate with the Public Health Department (William Paterson) and Health and Behavior Studies (Columbia) on their behavioral research in terms of data collection, analysis, and prediction
    - ➢ Correlate the results from statistics and the results from machine learning

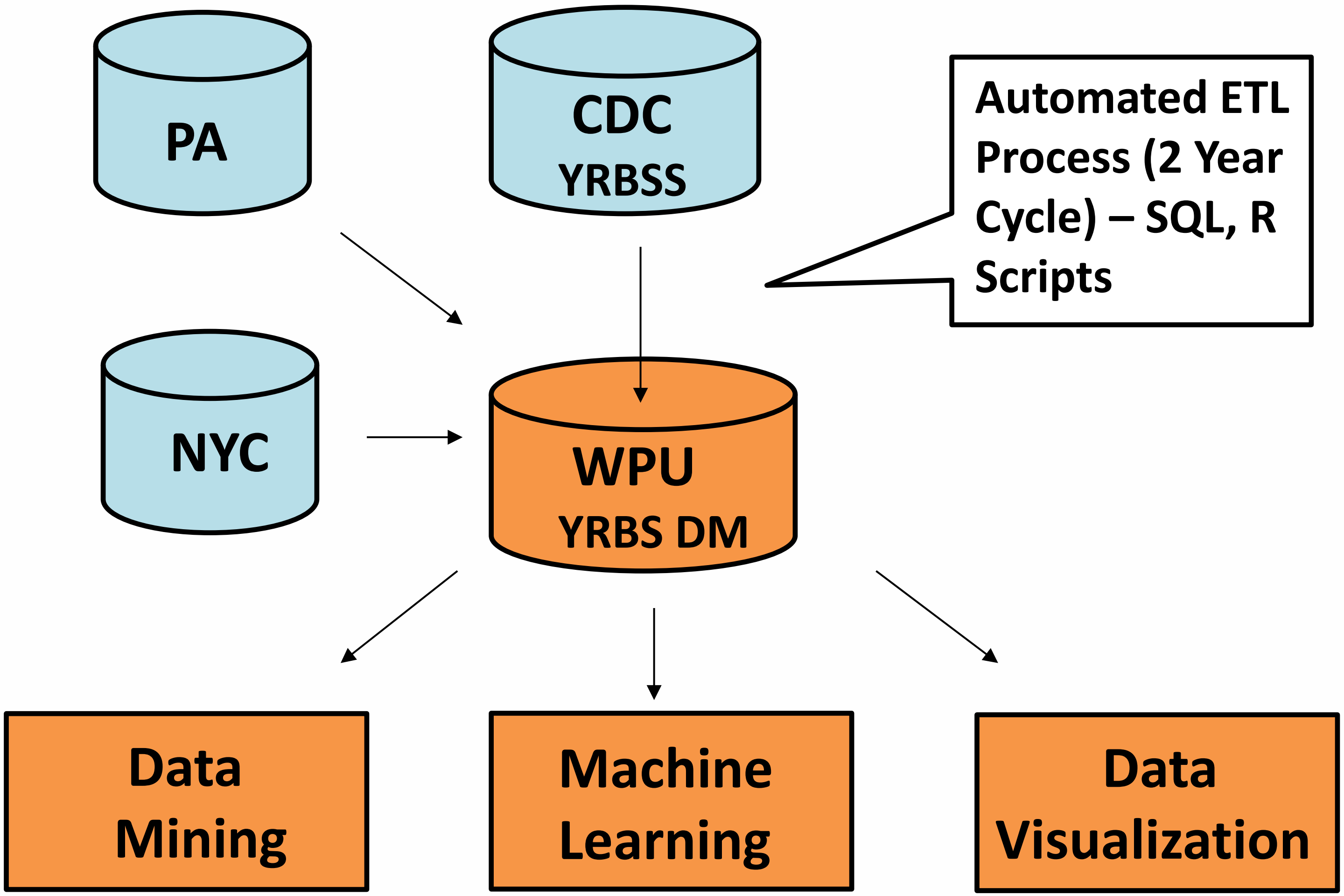
YRBSS (CDC)

- 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)

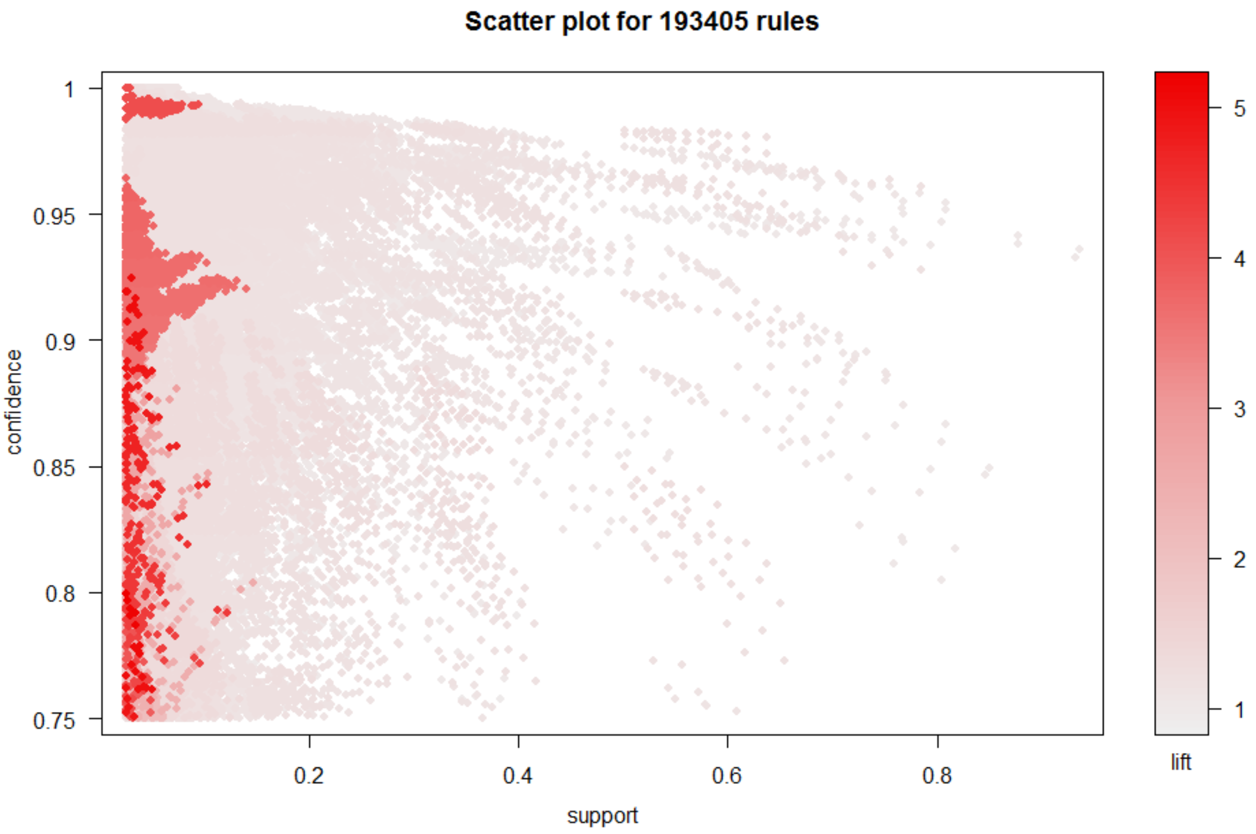
- Aggregated subsets of YRBSS from CDC and survey data from New York City and Pennsylvania
  - ➢ National (2011, 2013, 2015)
  - ➢ New Jersey (2011, 2013)
  - ➢ New York (2011, 2013, 2015)
  - ➢ New York City (2011, 2013, 2015)
  - ➢ Pennsylvania (2015)
  - ➢ Philadelphia (Pending)

Client/Server Architecture



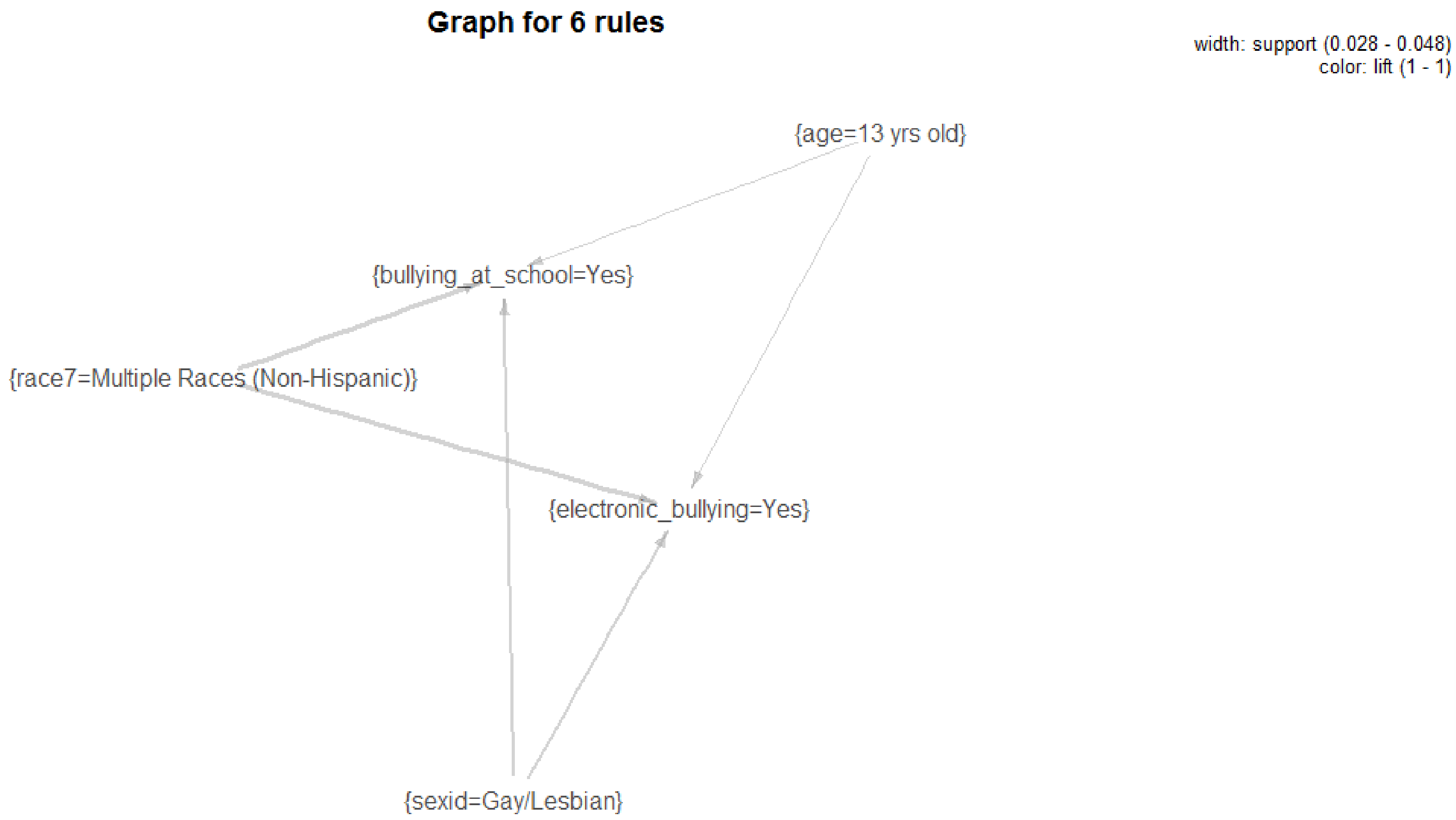
Preliminary Studies

- 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 slide shows the association scatter plot generated after running the Apriori Algorithm, showing only association rules with confidence > 0.75



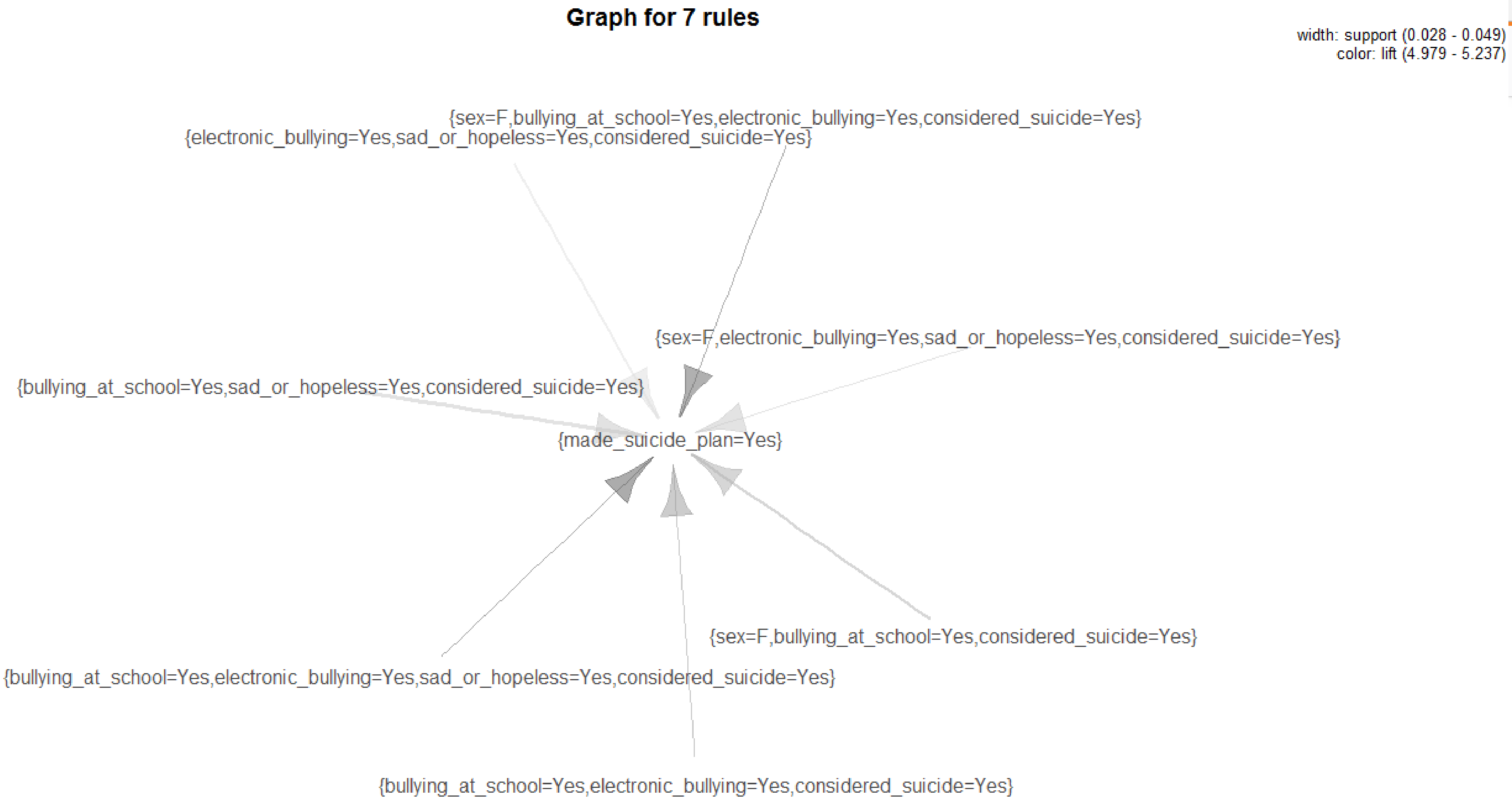
Metrics for Association Rule

- **Support** – how frequently the items in the rule occur together
- **Confidence** – probability of both the antecedent and the consequent appearing together *(the conditional probability of the consequent given the antecedent)*
- **Lift** – strength of a rule over the random co-occurrence of the antecedent and the consequent, given their individual support



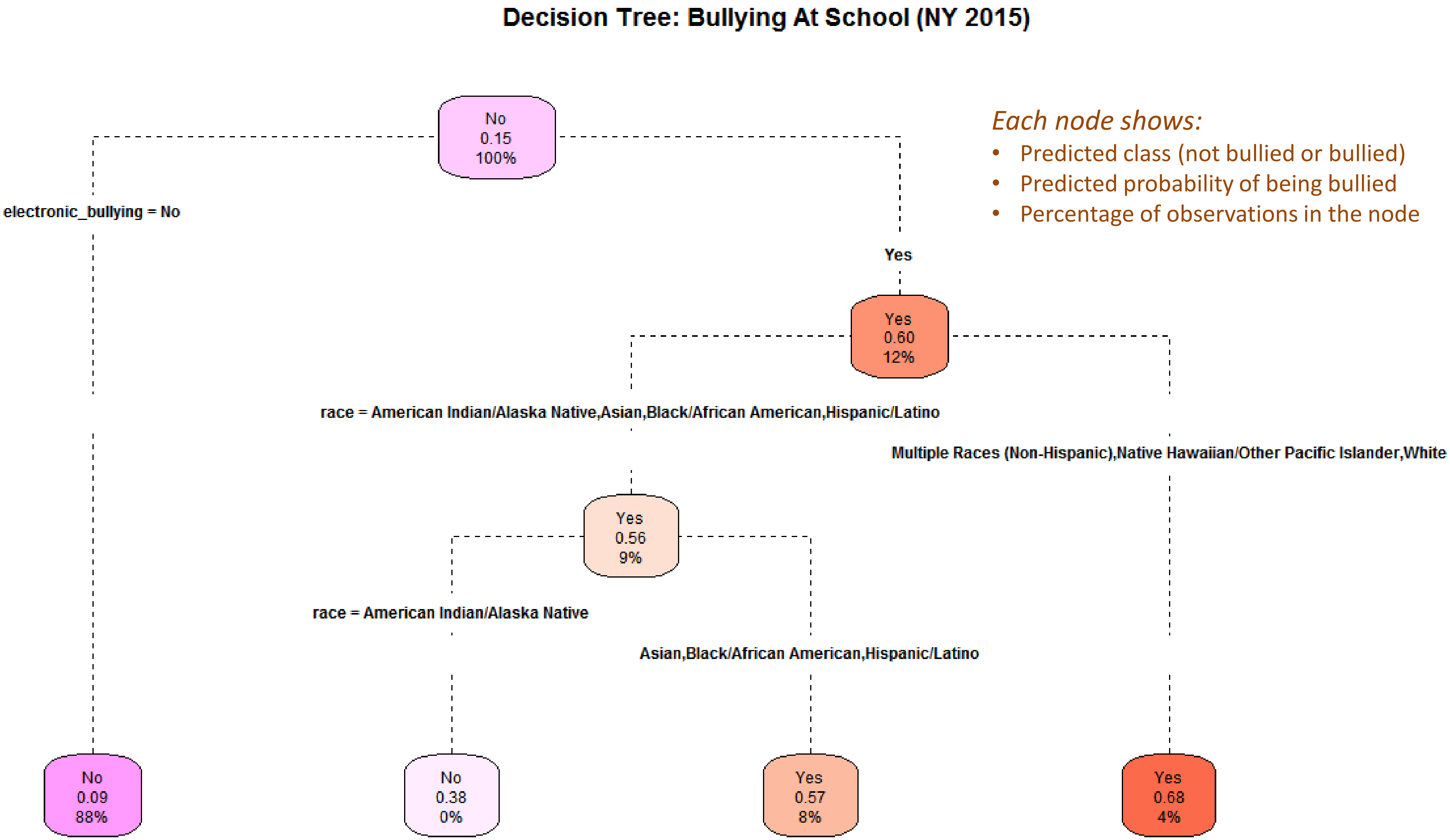
antecedent	consequent	support	confidence	lift
{age=13 yrs old}	{bullying_at_school=Yes}	0.028011204	1	1
{race7=Multiple Races (Non-Hispanic)}	{bullying_at_school=Yes}	0.047619048	1	1
{sexid=Gay/Lesbian}	{bullying_at_school=Yes}	0.040616246	1	1
{age=13 yrs old}	{electronic_bullying=Yes}	0.028011204	1	1
{race7=Multiple Races (Non-Hispanic)}	{electronic_bullying=Yes}	0.047619048	1	1
{sexid=Gay/Lesbian}	{electronic_bullying=Yes}	0.040616246	1	1

Example of strong association rules


Graph for 7 rules
width: support (0.028 - 0.049)
color: lift (4.979 - 5.237)
{sex=F,bullying_at_school=Yes,electronic_bullying=Yes,considered_suicide=Yes}
{electronic_bullying=Yes,sad_or_hopeless=Yes,considered_suicide=Yes}
{sex=F,electronic_bullying=Yes,sad_or_hopeless=Yes,considered_suicide=Yes}
{bullying_at_school=Yes,sad_or_hopeless=Yes,considered_suicide=Yes}
{made_suicide_plan=Yes}
{sex=F,bullying_at_school=Yes,considered_suicide=Yes}
{bullying_at_school=Yes,electronic_bullying=Yes,sad_or_hopeless=Yes,considered_suicide=Yes}
{bullying_at_school=Yes,electronic_bullying=Yes,considered_suicide=Yes}

Preliminary Studies

- 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


Decision Tree: Bullying At School (NY 2015)
No
0.15
100%
electronic_bullying = No
Yes
Yes
0.60
12%
race = American Indian/Alaska Native,Asian,Black/African American,Hispanic/Latino
Multiple Races (Non-Hispanic),Native Hawaiian/Other Pacific Islander,White
Yes
0.56
9%
race = American Indian/Alaska Native
Asian,Black/African American,Hispanic/Latino
No
0.09
88%
No
0.38
0%
Yes
0.57
8%
Yes
0.68
4%

Each node shows:

- Predicted class (not bullied or bullied)
- Predicted probability of being bullied
- Percentage of observations in the node

Future Research Plan

- 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
  - ➢ Use various data mining and machine learning (neural network algorithms) techniques of classification, association, and clustering analyses on the YRBS data
  - ➢ Summer 2017: extending tanning trending to 2015 using current statistical method; using machine learning algorithm (e.g., decision tree) for prior years to predict and correlate 2015 results; establish criteria to predict future tanning trending

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

WILLIAM
PATERSON
UNIVERSITY