A Knowledge Discovery Approach to Analyzing Mental Health Problems in NYC

Sara M. Steinel (Department of Computer Science), Nellie K. Cordova (Department of Mathematics),
Cyril S. Ku (Department of Computer Science), and David M. Freestone (Department of Psychology)
William Paterson University

Introduction

- This research analyzes the relationship between artificial tanning and some of the mental health problems in New York City (NYC).
- Previous research has discovered a relationship between frequent use of artificial tanning devices and suboptimal mental health using a logistic regression.
- Our research expands upon previous research, as we introduce a knowledge discovery approach.
- Our goal is to compare traditional hypothesis-driven statistical analyses that test an already existing statistical model (e.g. a logistic regression) and exploratory analyses that are not driven by an experimenter's model, but by hidden patterns in the raw data (e.g. a knowledge discovery approach).

Methods

- **Data collection:** The data collection for this research consists of the Centers for Disease Control and Prevention (CDC) Youth Risk Behavior Survey (YRBS) 2015 NYC data (1889 total responses).
- **Logistic Regression:** The logistic regression uses independent variables (mental health predictor questions and demographic covariates) to fit a model for a binary dependent variable (artificial tanning behaviors).
- **K-Nearest Neighbor (KNN):** KNN is a supervised classification machine learning algorithm. Instances of training data are used to classify new data objects by calculating the minimum distance.
- **Agglomerative and Divisive Clustering:** Agglomerative and Divisive clustering are two unsupervised hierarchical clustering machine learning algorithms. Agglomerative clustering is a bottom-up method, while divisive clustering is a top-down method.

Results

Logistic Model Summary

	OR	2.5 %	97.5 %
Age	0.3149804	0.2553083	0.3862507
Sex	1.0958631	0.9328389	1.2878688
Race	0.4943878	0.3803335	0.6375408
Bullied	0.7644283	0.5485792	1.0564053
Hopeless	0.7469698	0.5892021	0.9444811
Suicide Attempt	1.3830065	0.8297413	2.2782777
Seek Help	0.8097645	0.6090636	1.0715947

Fig. 1

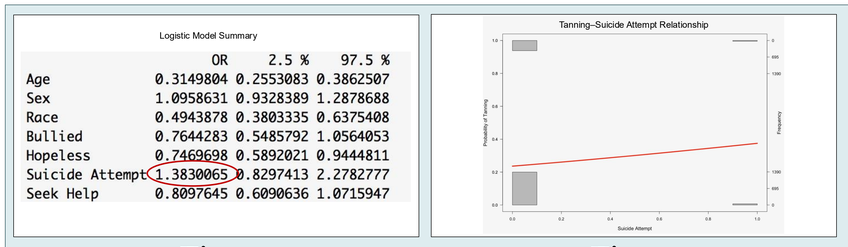

Fig. 2

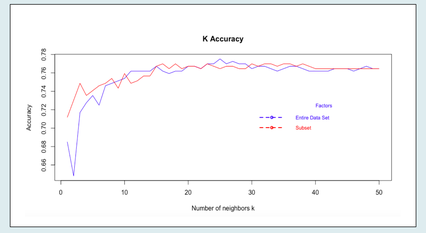

Fig. 3

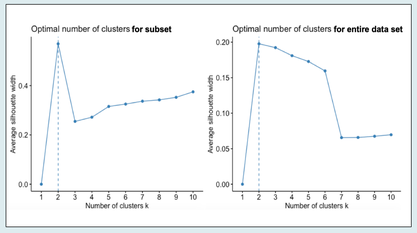

Fig. 4

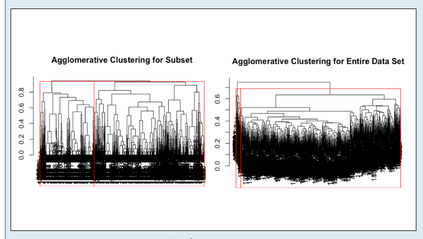

Fig. 5

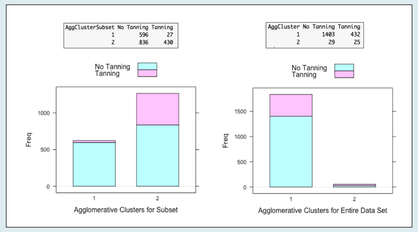

Fig. 6

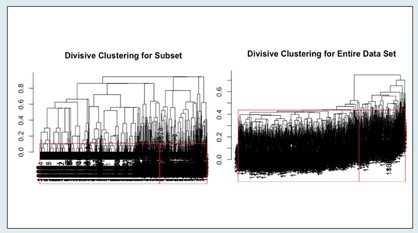

Fig. 7

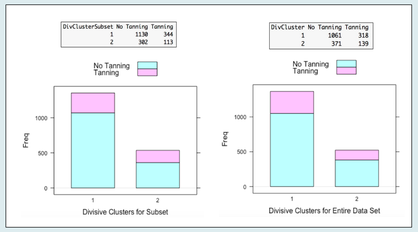

Fig. 8

Discussion

- Fig. 1: Odds ratio (OR) results from binary logistic regression (OR > 1 indicates positive correlation with tanning).
- Fig. 2: Graphical representation of OR showing positive correlation between suicide attempts and tanning.
- Fig. 3: Plot of accuracy of KNN for every K from 1 through 50, where K = number of neighbors.
- Fig. 4: Plot of average silhouette width for every K, where K = number of clusters (K = 2 is optimal).
- Fig. 5: Dendrogram for agglomerative clustering method.
- Fig. 6: Distribution of tanners and non-tanners for the agglomerative clustering method.
- Fig. 7: Dendrogram for divisive clustering method.
- Fig. 8: Distribution of tanners and non-tanners for the divisive clustering method.
- **Conclusion: All algorithms are sufficient for this study, but the subset yields more accurate results.**

References

Brener, N., Kann, L., Shanklin, S., Kinchen, S., Eaton, D., Hawkins, J., & Flint, K. (2013). Methodology of the Youth Risk Behavior Surveillance System—2013. *Morbidity and Mortality Weekly Report: Recommendations and Reports, 62*(1), 1- 20.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Waltham, MA: Morgan Kaufmann Publishers.

Acknowledgements

We would like to acknowledge the contributions of Corey Bash (Department of Public Health at William Paterson University), and Grace Hillyer (Department Epidemiology at Columbia University).