# An Exposome Data Analysis Pipeline in R

Jeff Sorbo
*Department of Computer Science*
*Texas Tech University*
*Lubbock, Texas 79409-3104*
*Email: jeffrey.s.sorbo@ttu.edu*

***Abstract*—Brief description of the work described in this paper.**

## 1. Introduction

The Exposome data are described; the goals of the analytic pipeline are defined; data loading, cleaning, and merging are detailed; feature selection and data modeling are detailed; model evaluation and results are discussed; and areas of future work are listed.

## 2. Data

The Exposome data were provided by the Exposome research group and consisted of 3,125 data points representing county and parish units across the United States. The data were provided in two files: the independent variables file and the dependent variables file.

The independent variables file contained data in 63 attributes: 3 unique identifiers including a string attribute consisting of the county and state name; and 60 numeric attributes consisting of various data aggregated at the county level, such as population, bank offices, housing unit values, per capita income, and average daily precipitation.

The dependent variables file contained data in 9 attributes: the unique identifier consisting of county and state name, 7 numeric attributes related to cardiovascular disease (CVD) death; and an attribute containing the quintile of the age-adjusted CVD death rate.

## 3. Methodology

A pipeline was developed to load, clean, merge, and preprocess the data and to train an ensemble learning model to predict the CVD rate. Based on [1], the ensemble learning model combined clustering, decision trees, and association mining. The pipeline was written in the R language to provide for potential reuse and adaptation by members of the Exposome research group.

### 3.1. Data Loading and Conversions

The Exposome data were loaded from the independent and dependent attributes files in comma-separated values (CSV) format.

Many of the attribute names in the files were based on codes in the original data sources, *e.g.*, "AGE030200D," "HEA010200D," "HSG680200D;" such attributes were given friendly names based on a data dictionary provided by the Exposome group.

Data points with missing values were removed, and the independent data were merged with the dependent data based on the county and state names.

All numeric attributes were grouped in quintiles.

The CVD attribute in the merged file was converted to a binary type: the highest quintile (*i.e.*, the highest rate of CVD) was set to 1, and the lower 4 quintiles were set to 0.

### 3.2. Feature Selection

The Exposome data were grouped into subsets for model training and evaluation. The first subset, hereafter referred to as "data set 1," consisted of 10 attributes identified as a paraclique by members of the Exposome group. The second subset, hereafter referred to as "data set 2," included all 23 statistical attributes from the independent attributes file.

Some feature selection techniques were applied to data set 2: the $\chi^2$ test, symmetrical uncertainty, and gain ratio. The data sets resulting from these methods are hereafter referred to "data set 2a," "data set 2b," and "data set 2c" respectively.

### 3.3. Data Modeling

K-Means clustering was applied to the paraclique data. 3 clusters were found, and each data point was labeled with the cluster id 1-3.

Decision trees using recursive partitioning were trained against each cluster. The first tree is shown in Figure 1.

The apriori association mining algorithm was run against the data.

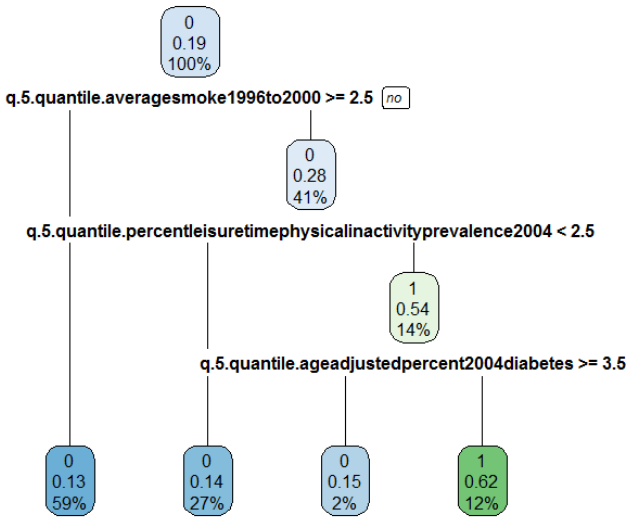


Figure 1. Decision tree based on paraclique features, no clustering

# 4. Results

# 5. Remaining Work

# 6. Conclusion

# References

[1] S. Datta, *A Multi-Stage Decision Algorithm for Rule Generation for Minority Class*. PhD thesis, Texas Tech University, Lubbock, TX, 8 2014.