Death by Chocolate? No, Death by County. Modelling County-Level Mortality Rates

Introduction

Motivation/Purpose

Data Sources

The first data source we used was a compilation of county-level data from a variety of sources compiled by Github user Deleetdk. Most of the data sources included are those from a paper published by Emil O. W. Kirkegaard in the journal *Open Quantitative Sociology and Political Science* in 2016, "Inequality across US counties: an S factor analysis". This data source was, very luckily for us, an RData file on GitHub that was very easy to download.

The demographic data (such as racial composition, gender, and age) come from the United States Census Bureau's annual American Community Survey through the American FactFinder. This is also the source for some economic indicators (educational attainment, sector composition of employment). Another data source is the American Human Development Index of the Measure of America, which contains health, education, and income indicators. Many health indicators come from the Robert Wood Johnson Foundation's County Health Rankings & Roadmaps. Finally, the county-level electoral data comes from the failing New York Times.

The first step of data wrangling with this dataset was selecting the variables we wanted from it. We then removed the state-level rows, which all have FIPS codes of X000, so that the data was only counties.

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

The key dataset for this analysis is a set of estimates of "annual mortality rates by US county for 21 mutually exclusive causes of death from 1980 through 2014" constructed by Laura Dwyer-Lindgren, Amelia Bertozzi-Villa, Rebecca Stubbs, et al. from the National Center for Health Statistics' National Vital Statistics System and published in the Journal of the American Medical Association. We downloaded the dataset itself from Kaggle, which it was provided to by the Institute for Health Metrics and Evaluation.

The causes of death (and our short names for them) are:

HIV: HIV-AIDS and Tuberculosis INF: Diarrhea, lower respiratory, and other common infectious diseases TROP: Neglected tropical diseases and malaria MAT: Maternal disorders NEON: Neonatal disorders NUT: Nutritional deficiencies OTHC: Other communicable, maternal, neonatal, and nutritional diseases NEOP: Neoplasms CHRON: Chronic respiratory diseases CIRR: Cirrhosis and other chronic liver diseases DIG: Digestive diseases NEUR: Neurological disorders MENSUB: Mental and substance use disorders DIAB: Diabetes, urogenital, blood, and endocrine diseases MUSC: Musculoskeletal disorders OTHN: Other non-communicable diseases TRAN: Transport injuries UNIN: Unintentional injuries SELF: Self-harm and interpersonal violence WAR: Forces of nature, war, and legal intervention

This dataset was an Excel file with estimates with estimated mortality rates for each county for each of 20 causes of mortality for each of many years (1980, 1985, 1990, 1995, 2000, 2005, 2010, and 2014). The first thing we had to do was change each estimate from a midpoint and a confidence interval to just the midpoint.

We also renamed the sheets and columns to make them easier to work with. Then, because each cause of death was its own sheet, we had to make a function to read in each sheet.

```
library(readxl)
read_excel_allsheets <- function(filename) {
    sheets <- readxl::excel_sheets(filename)
    x <-    lapply(sheets, function(X) readxl::read_excel(filename, sheet = X))
    names(x) <- sheets
    x
}
mortalityrates <- read_excel_allsheets("MortalityRatesClean.xlsx")
```

```
## Warning in strptime(x, format, tz = tz): unknown timezone 'default/America/
## Los_Angeles'
```

```
mortalityrates <- as.data.frame(mortalityrates)
```

Because each cause of death was its own sheet, when we imported the data there were 20 columns of the FIPS codes and 20 columns of location (the county names). As a result, we selected one of the location and FIPS and all of the mortality rate estimates to keep.

Similar to with the other dataset, there were state rows as well as county rows. However, in this dataset the states had low FIPS codes (below 100) instead of FIPS codes of multiples of 1000. Additionally, the mortality rates dataset had counties in Puerto Rico (FIPS codes in the 72000s), but we removed these because they did not have data in the Deleetdk data.

```
mortalityrates <- mortalityrates %>% filter(fips>100) %>% filter(fips<70000) %>% arrange(fips)
```

At this point, we combined the two datasets.

```
require(dplyr)
countyrates <- inner_join(countydata,mortalityrates,by=c("fips"))
```

After combining the datasets, we noticed there were several observations with missing data for some variables. In particular, all of Alaska was missing most of the Deleetdk data and a few stray observations were missing the education data and a number of other variables. We removed those observations.

```
countyratesnogaps <- countyrates %>% filter(fips>100) %>% filter(fips<70000) %>% filter(!is.na(rep16_fra
```

We also chose to remove some of the variables that had a lot of missing data by selecting the others.

Finally, we formally created the dataframe we would use of all the county observations with tidy data for all of our variables of interest.

```
countyratesfullcases <- countyratesnogaps[complete.cases(countyratesnogaps),]
```

Model Creation/Statistical Computation

We created two models for each of the twenty causes of death - a regression tree and a k nearest neighbors model.

```
require(caret)
```

```
## Loading required package: caret
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
require(rpart)
```

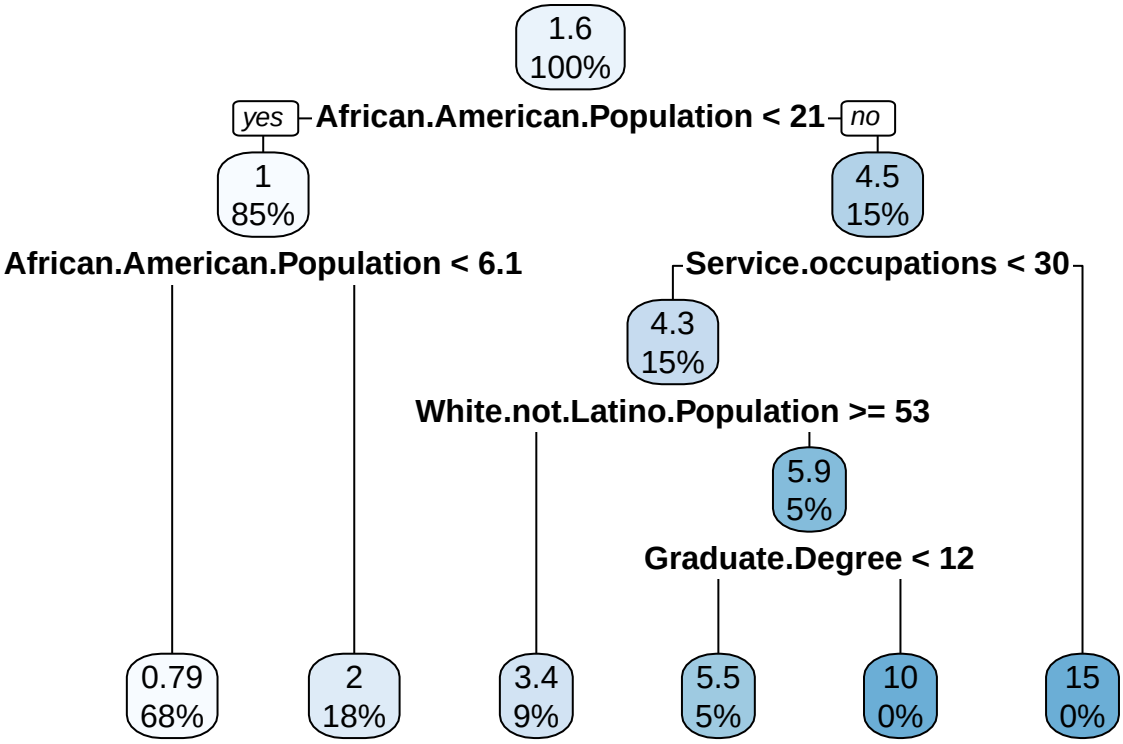
```
## Loading required package: rpart
```

```
require(rpart.plot)
```

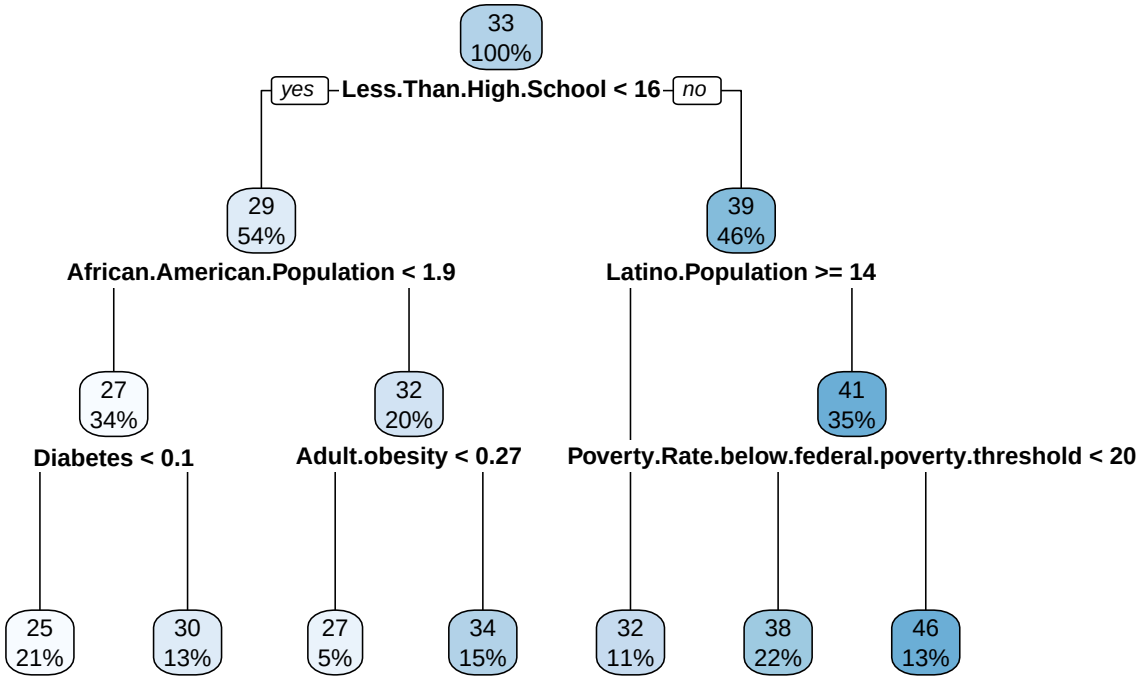
```
## Loading required package: rpart.plot
```

```
library(tree)
```

HIV: HIV-AIDS and Tuberculosis

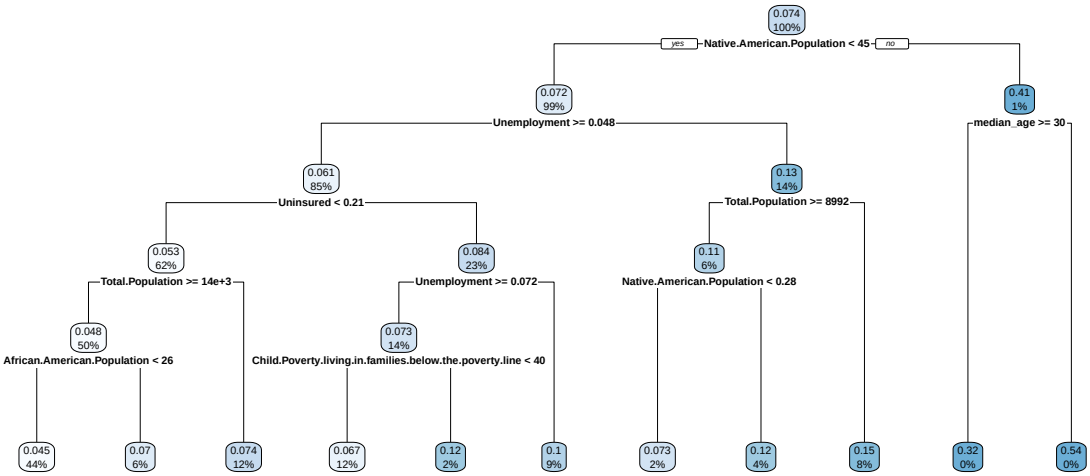


INF: Diarrhea, lower respiratory, and other common infectious diseases

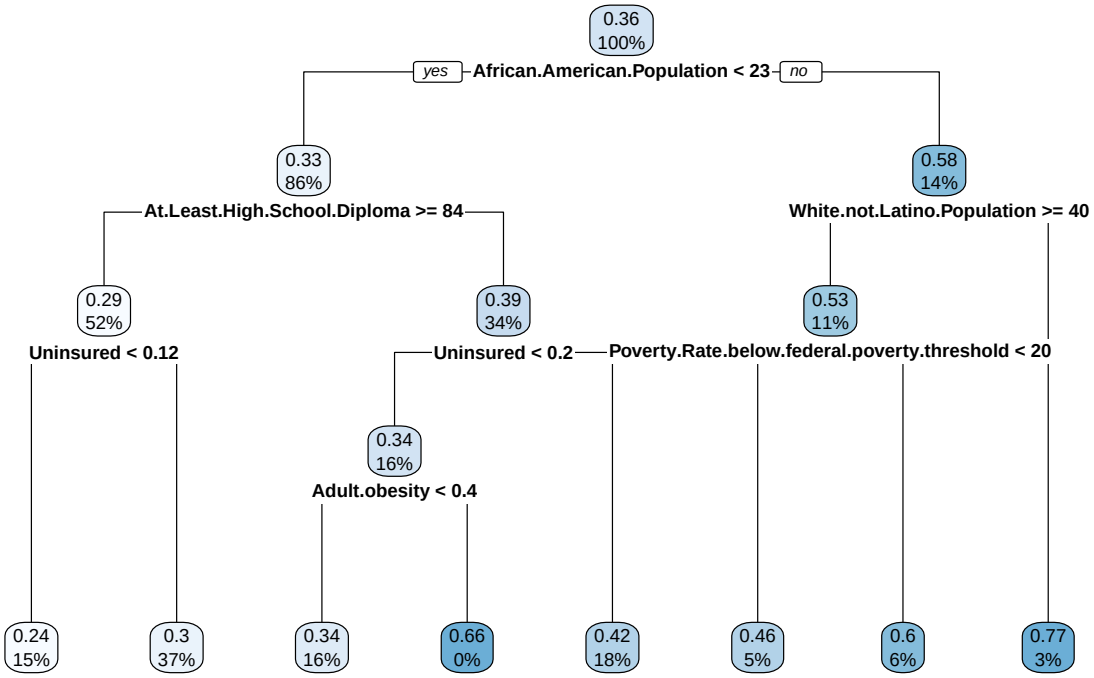


TROP: Neglected tropical diseases and malaria

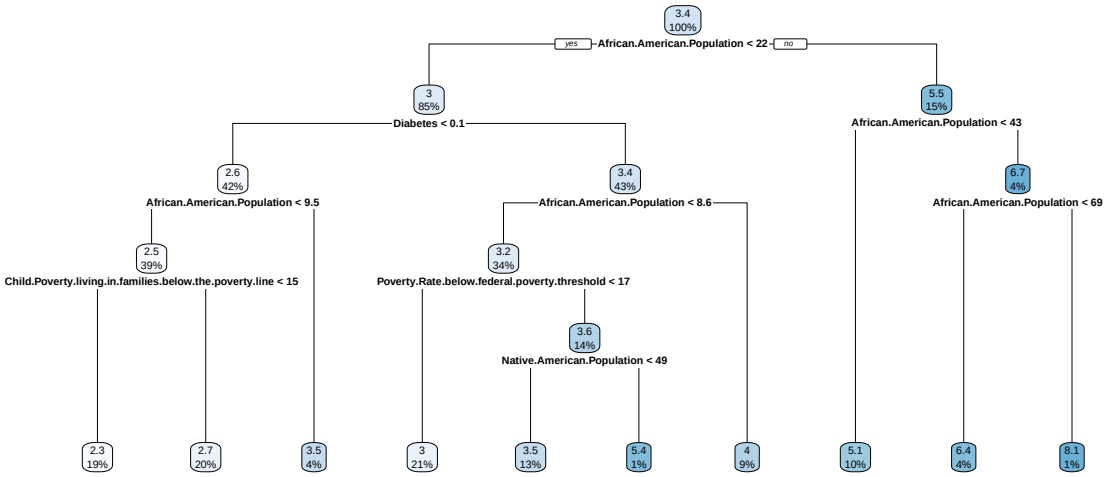
```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =
## trainInfo, : There were missing values in resampled performance measures.
```



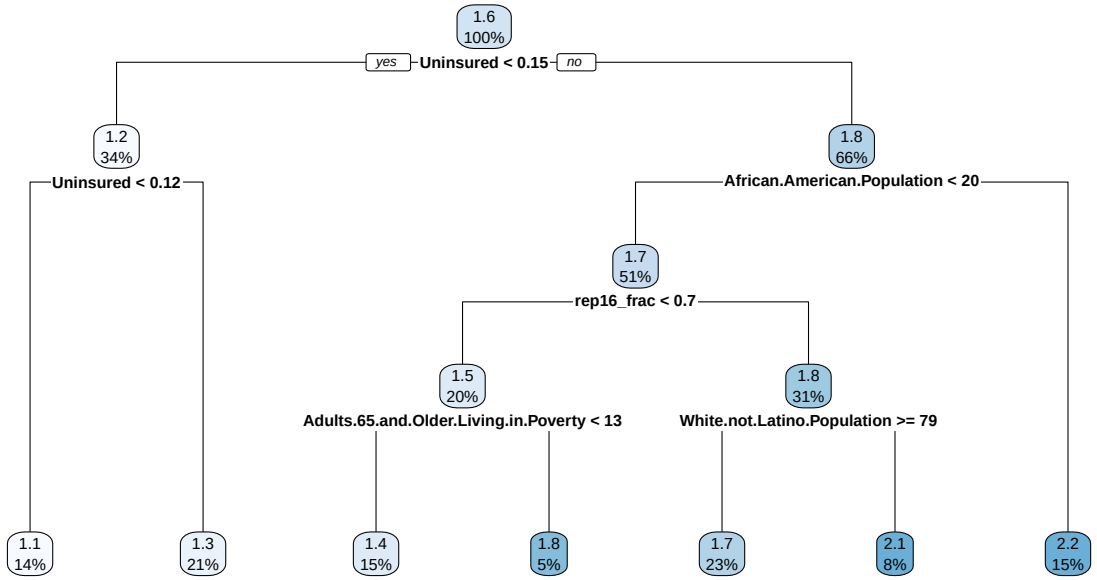
MAT: Maternal disorders



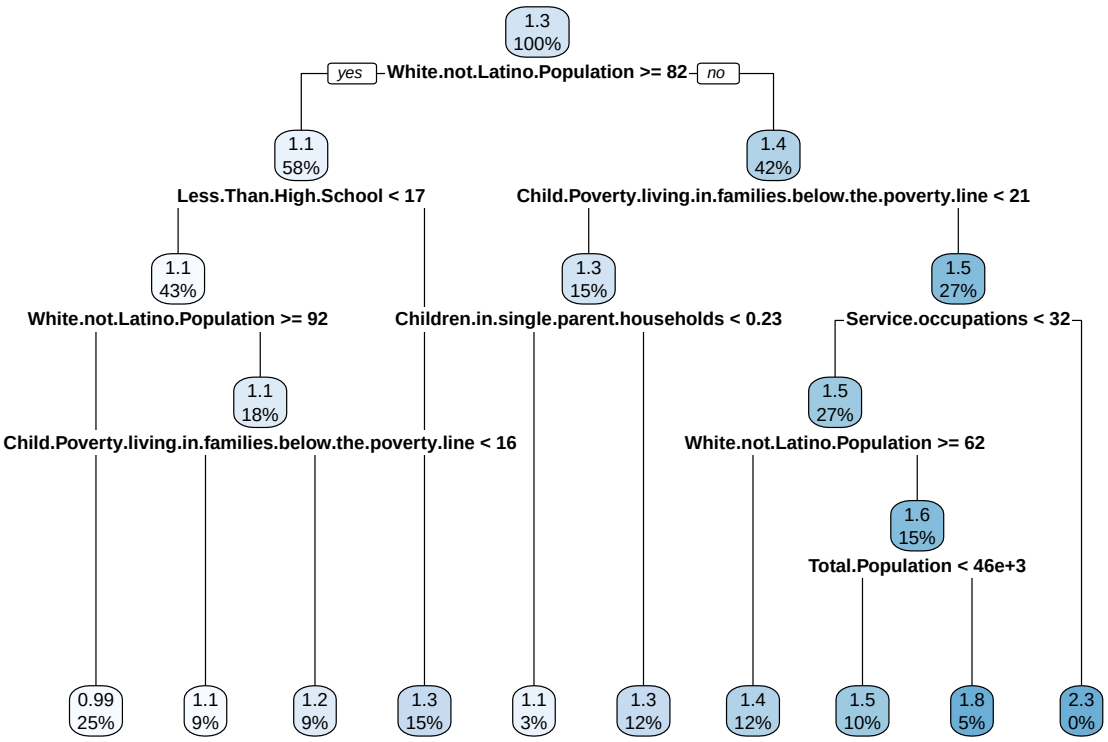
NEON: Neonatal disorders



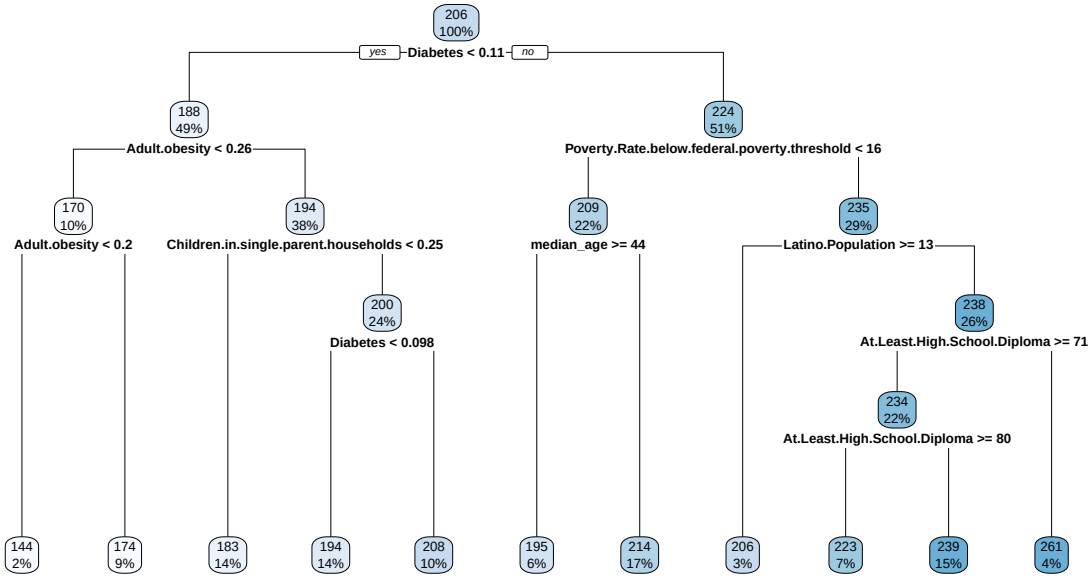
NUT: Nutritional deficiencies



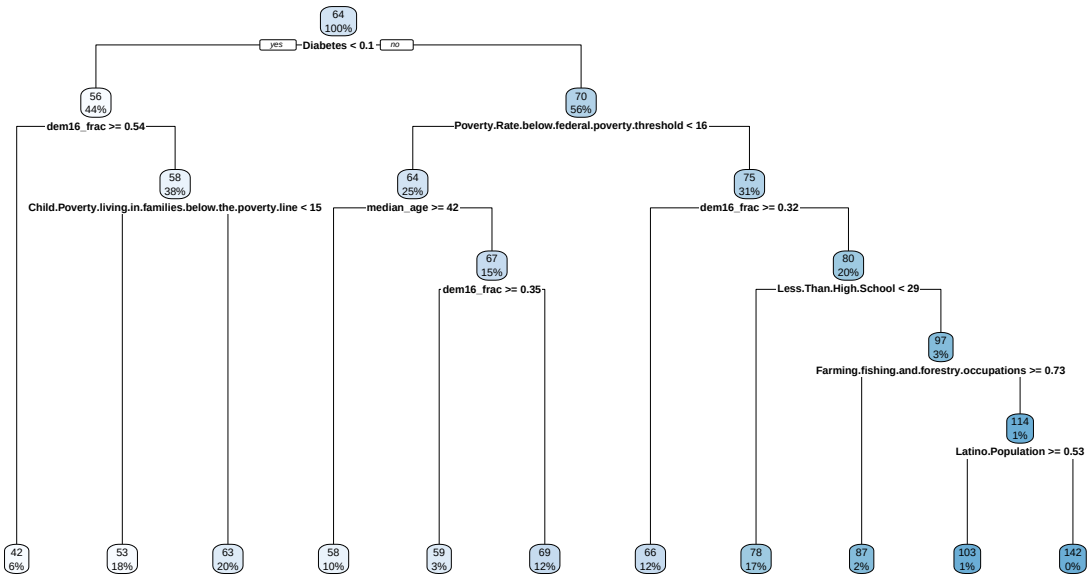
OTHC: Other communicable, maternal, neonatal, and nutritional diseases



NEOP: Neoplasms

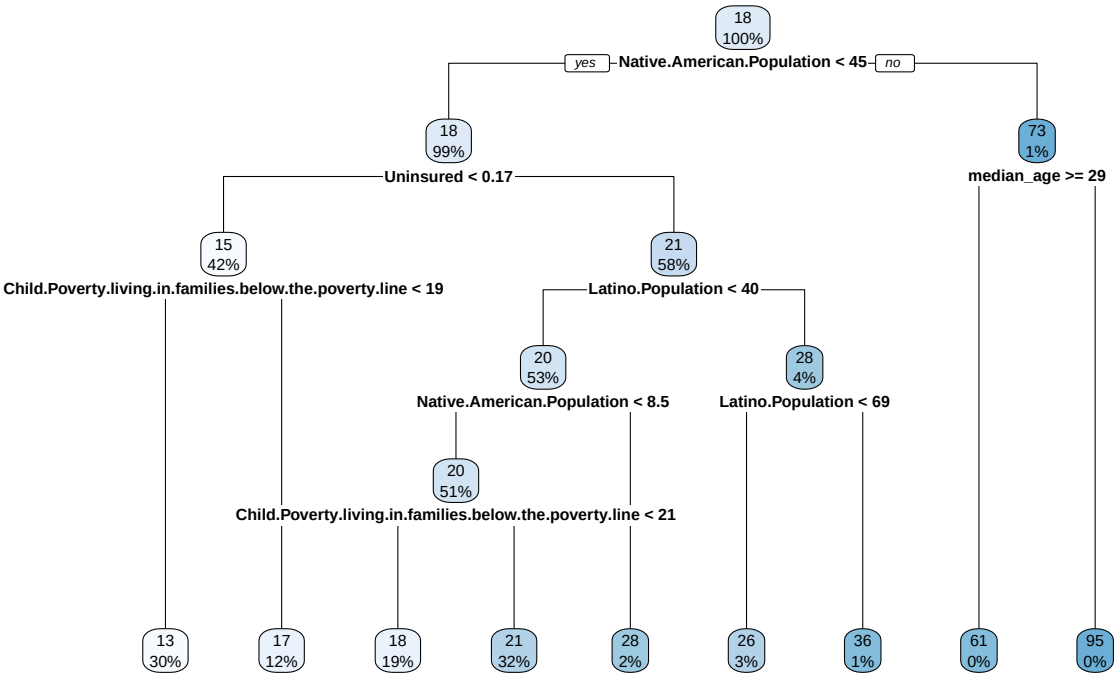


CHRON: Chronic respiratory diseases

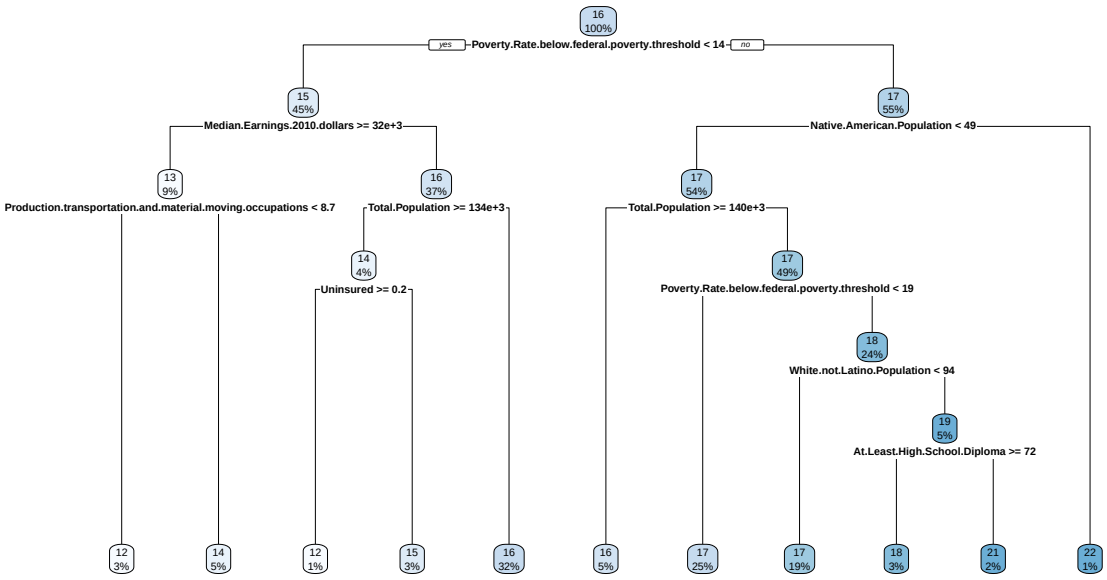


CIRR: Cirrhosis and other chronic liver diseases

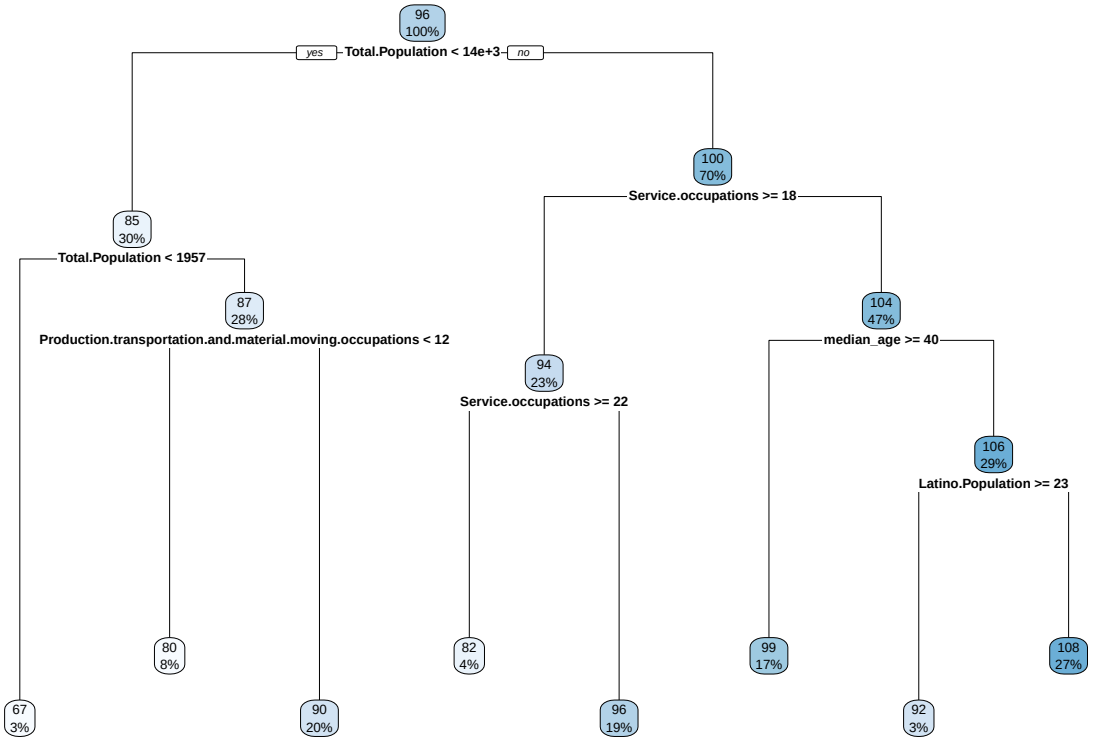
```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =
## trainInfo, : There were missing values in resampled performance measures.
```



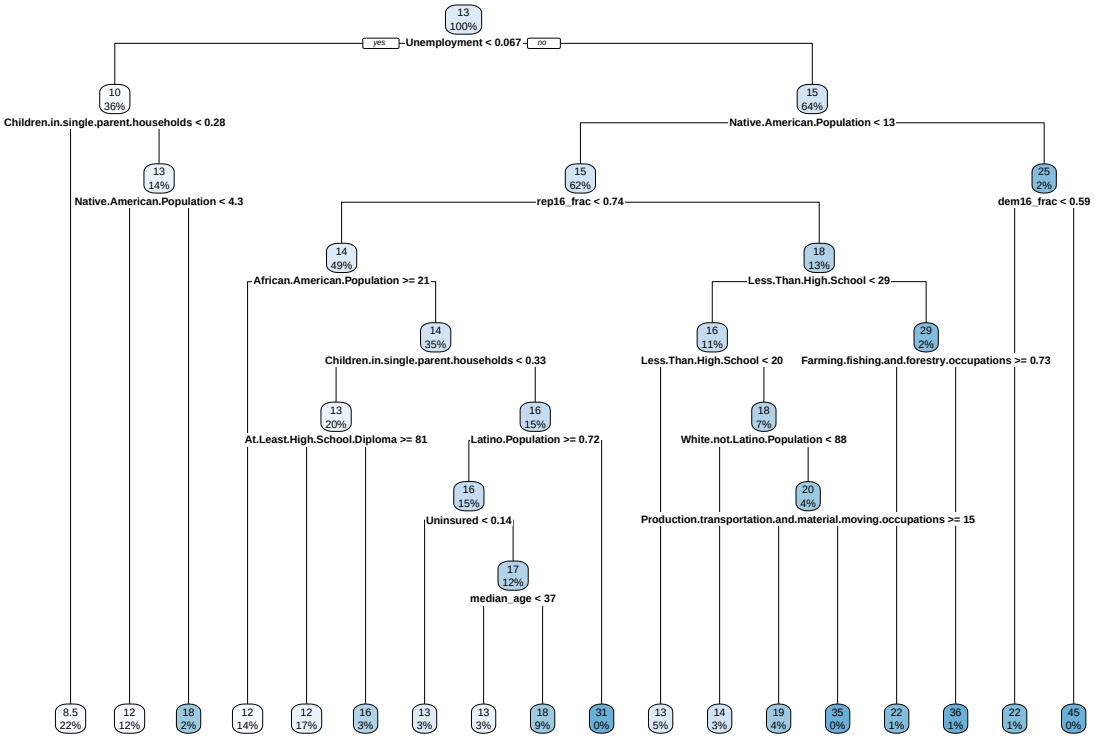
DIG: Digestive diseases



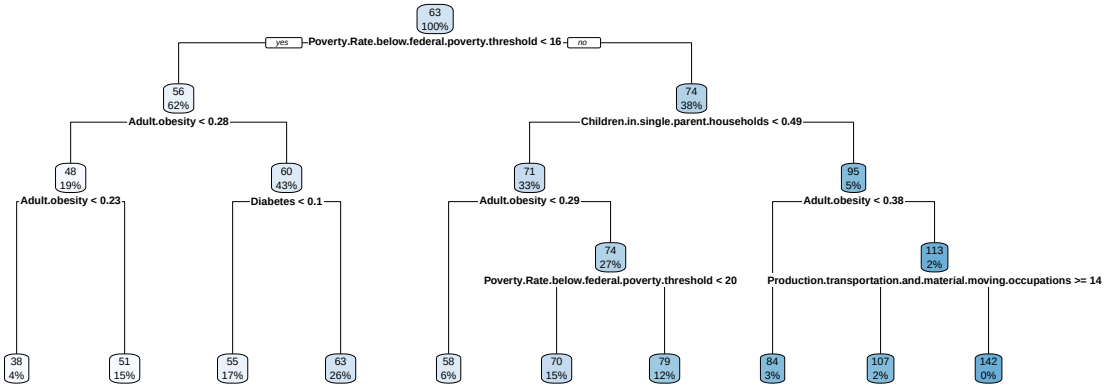
NEUR: Neurological disorders



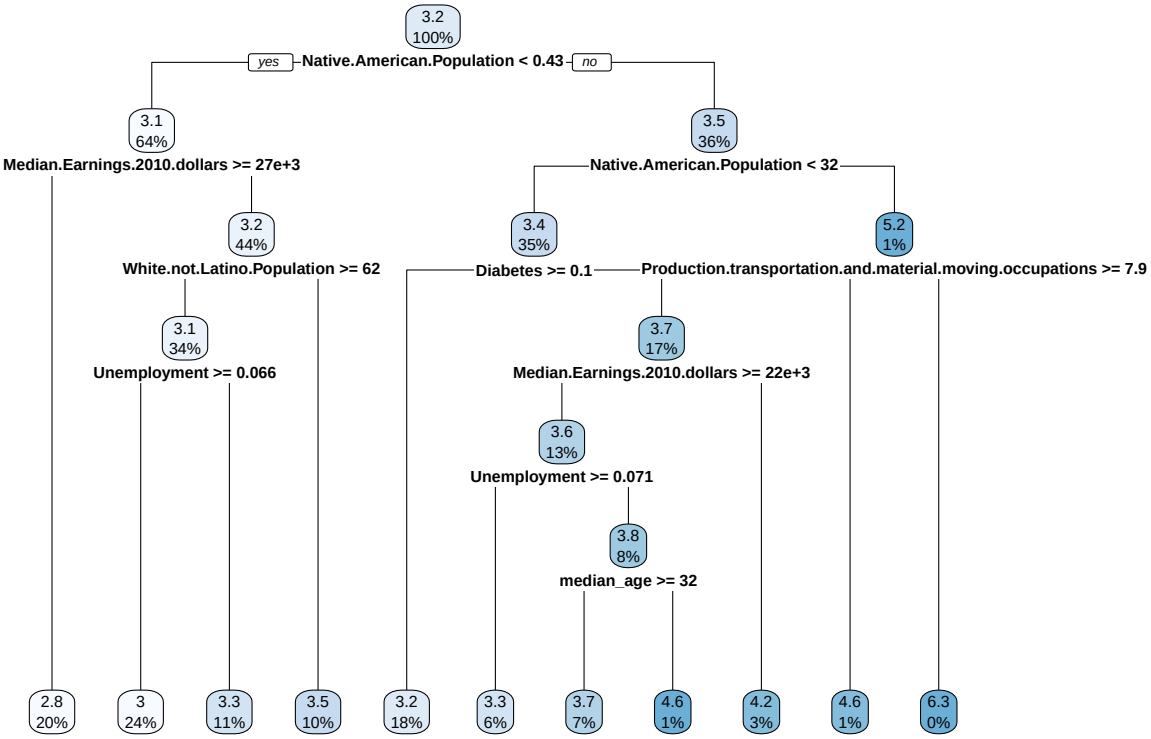
MENSUB: Mental and substance use disorders



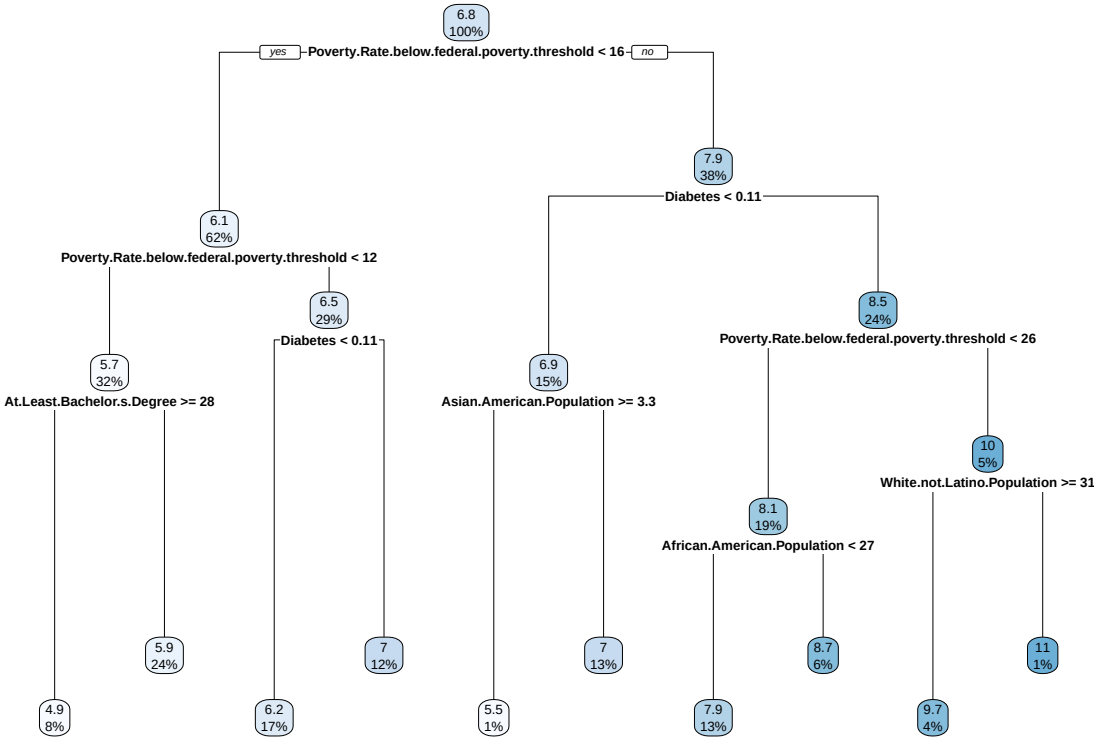
DIAB: Diabetes, urogenital, blood, and endocrine diseases



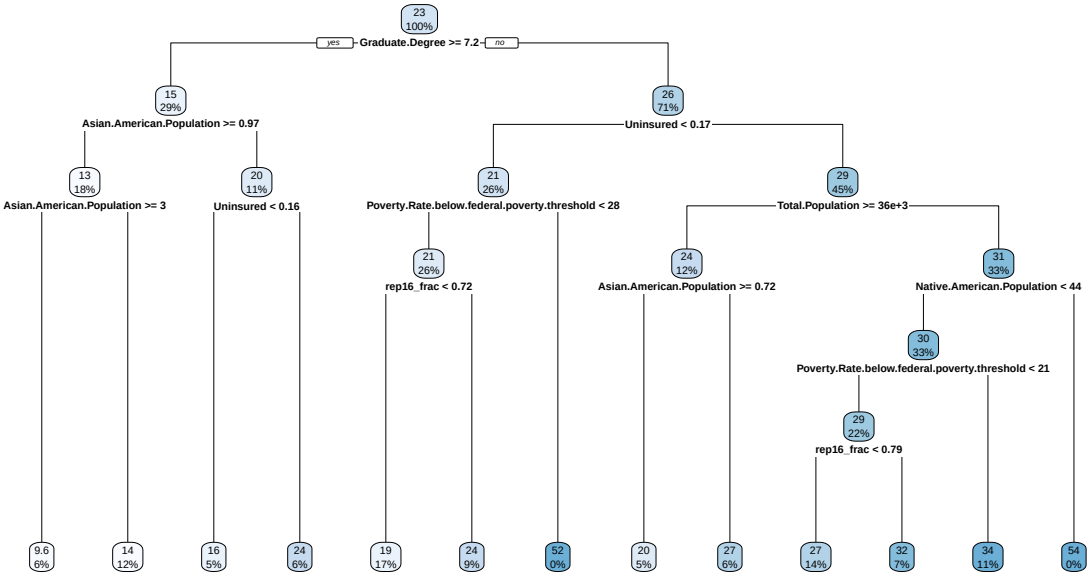
MUSC: Musculoskeletal disorders



OTHN: Other non-communicable diseases

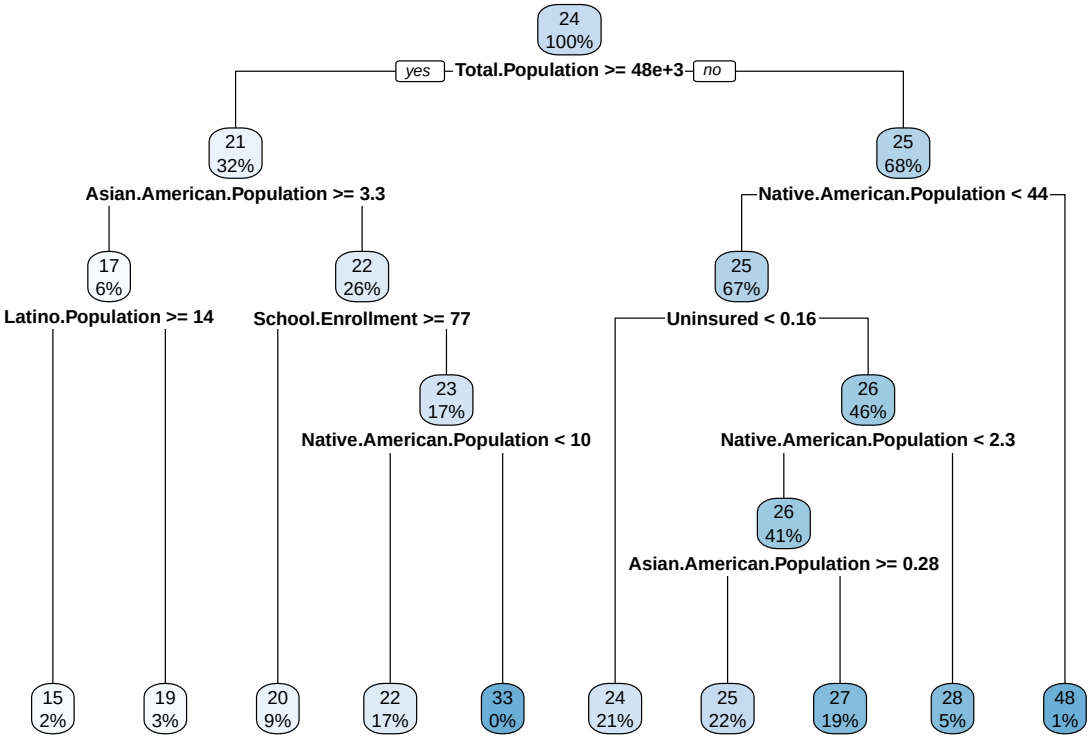


TRAN: Transport injuries

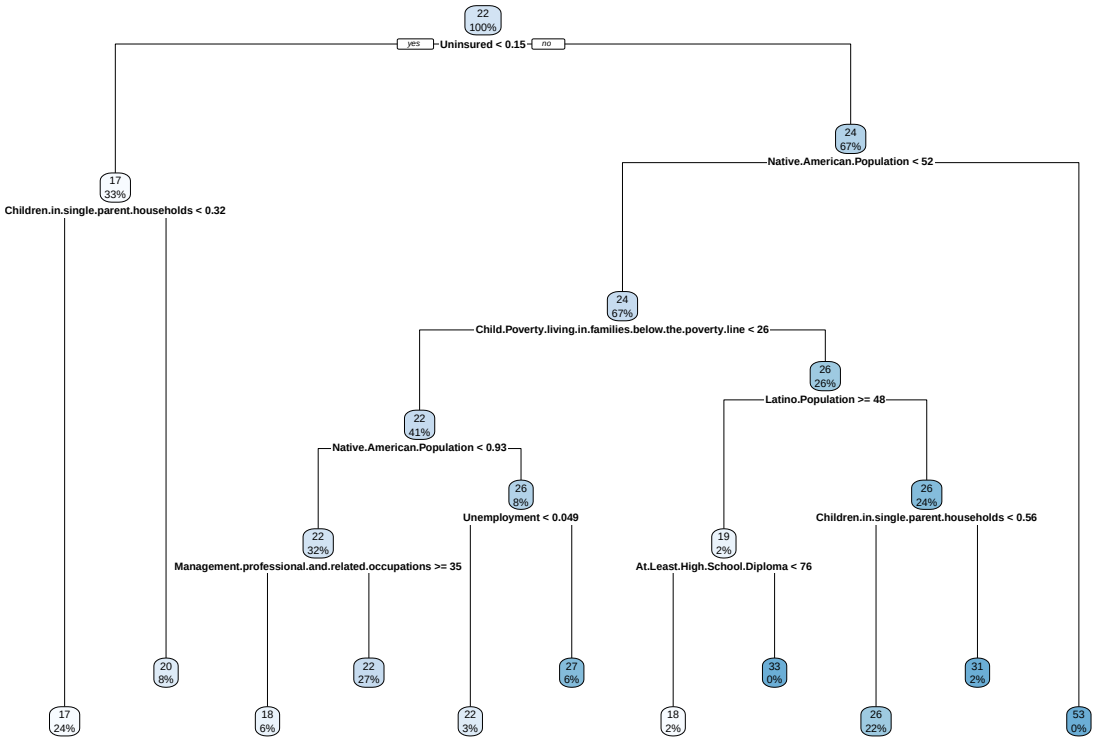


UNIN: Unintentional injuries

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =
## trainInfo, : There were missing values in resampled performance measures.
```



SELF: Self-harm and interpersonal violence



WAR: Forces of nature, war, and legal intervention