Alejandro Naranjo

Assignment 7 Report

For this report there will be two different sections; the first section will be addressing the data analytics for both datasets, and the second section will address the different models that were built. At the end of the report there will be a conclusion that goes over the results and what on these models.

For the two different data sets, I will be using the dataset which contains obesity data based on eating and physical conditions (UCI Machine Learning Repository), and the second dataset that was chosen is the dataset with a decades worth of data on Diabetes (UCI). The reason for picking these two datasets is because I am interested in looking at health data and how we can use it to make predictions about someone’s health. It’s important to track this information so we can make early detections of illnesses or improve the type of health treatment people can receive by analyzing the data that has been collected.

# Data Analysis

## Diabetes Dataset

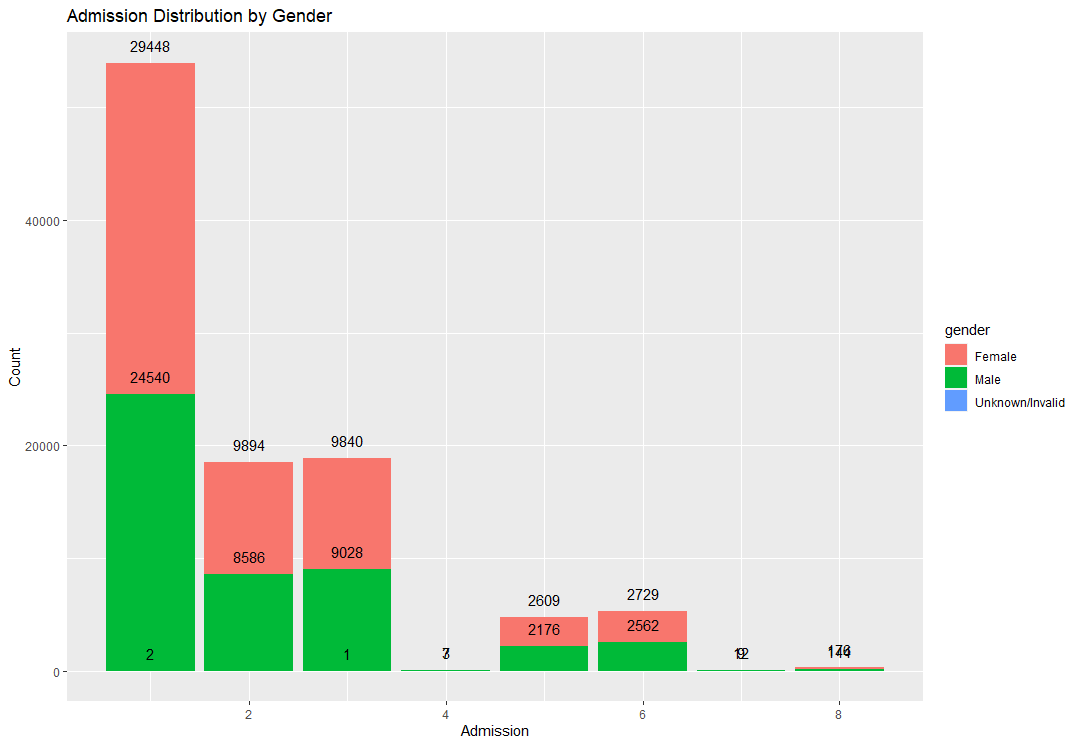


Figure Distribution of Admission Id By Gender

Chart, bar chart

Description automatically generated

Figure Distribution of A1CResult With Readmission Results

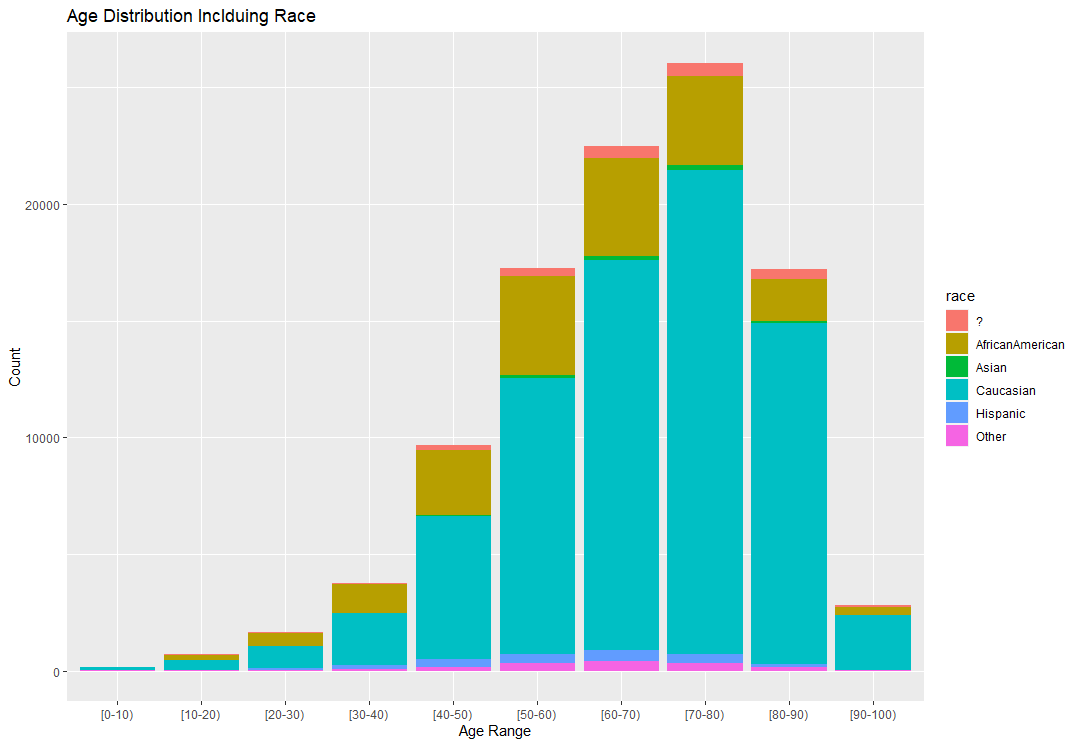


Figure Age Distribution Including Race

Chart, waterfall chart

Description automatically generated

Figure Correlation Matrix of Numeric Data Points

For the Diabetes data set there was a large data set which had over 50 different feature values. For these values we didn’t need all of them and we could get rid of many. The first part of the process was to visualize and see what the data looked like. Looking at Figure 1, the purpose of this was took at what type of data entry made up the majority of the data set. The Admission Id for a patient was mapped to a numerical value and the following is what each represents; 1 is an emergency, 2 is urgent, 3 is elective, 4 is newborn, 5 is Not available, 6 is null, and 7 is trauma center. The data set was largely made up of emergency visits and the good part was neither sex group dominated in the visits for each ID group. The second graph, Figure 2, we are looking at the distribution from an *A1CResult* . What this feature represents is if a test was conducted to understand a person’s glucose levels. Look at the distribution of the results we have a large number of “*None”* indicating the test was never conducted. This might play into effect when we are creating the models.

## Obesity Dataset