**DS6372 – Project 2**

# **Introduction**

The Health and Aging Brain among Latino Elders (HABLE) study is focused on studying the combined impact of depression and inflammation on memory functioning among Mexican-American adults and elders [1]. The authors of this paper have elected to leverage data from this study to determine if it is possible to predict cognitive impairment in the population studied within the HABLE data set. The authors have been able to obtain the HABLE dataset, as described below, for their analysis.

The purpose of this paper is to provide a statistical analysis of the HABLE dataset focused on being able to predict cognitive impairment. The analysis has two main objectives with the dataset. The first objective is to build and interpret a logistic regression model. The second objective is to build and compare additional models to improve on prediction performance metrics.

## **Data Description**

The data leveraged for this analysis is provided by one of the authors as part of analysis being performed in the authors role. The data set contains medical related data for patients in the Health and Aging Brain among Latino Elders (HABLE) study. The data contains information for 381 participants in this study. The total data set contains 741 observations with 161 features or variables. These 161 features contain information relating to various patient attributes such as: age, gender, race, blood pressure, education, income, height, weight, BMI, markers around cholesterol and mental cognition.

The dataset contains a mix of categorial and continuous features. The categorical features indicate information such as cancer status, education level, race consideration, and gender. The continuous features indicate information such as age, blood pressure, BMI, height, weight and cholesterol readings. The features were captured by both an interview process as well as through standard medical result capturing. Thus, some of the values are self-reported by participants in the study

## **Exploratory Data Analysis (EDA)**

The first set in the EDA was to determine which features, if any, could be removed from the data set due to not being relevant/valuable for the analysis. An initial statistical summary analysis (Table 1) was used to look for data quality issues including missing values. Upon review it was initially determined that the following factors or factor groups would be removed from the analysis:

* MedID – removed as it is a record identifier
* Race Identifiers – categorical factors that are summarized in another way
* CM\_notes – removed as it is a notes field
* IMH\_age variables – removed as they contained many missing values
* APOE variables – removed for initial model
* eGRF variabels – removed as they contained many missing values
* cdx\_mci -
* Income – removed for initial model

After removing the factors above the data set contains 117 factors across 400 observations. The next analysis was to determine correlation across factors to look for further simplification. Correlation plots of the data are provided in Figure 1. From this plot it was determined that some variables are clearly highly correlated. The variables can be further simplified as factors such as height, weight and BMI show high correlation.

In order to predict cognitive impairment using the dataset, the authors also elected to collapse the cdx\_cog feature from a 9 factor feature to a binary feature. The various factors were easily grouped into a binary configuration of cognitive impairment or no cognitive impairment. This imputation of variable allowed for additional flexibility in analysis.

## **Analysis Question 1**

## Problem Statement

The goal of analysis question 1 is to perform a logistic regression analysis on the dataset. The interpretation of the regression coefficients including hypothesis testing and confidence intervals are provided for the reader to provide model interpretation

## Model Selection

## Parameter Interpretation

## Conclusion

## **Analysis Question 2**

## Problem Statement

The goal of analysis question 2 is to compare predictive performance of models. The authors will leverage the model developed in analysis question 1 as a baseline for predictive performance. The baseline model will be built upon and predictive performance metrics across the various models compared to determine if predictive performance can be improved.

## Main Analysis

## Conclusion

## **Appendix**

## List of Tables

Table 1 – Summary Statistics

## List of Figures

Figure 1 – Correlation Plots