Understanding the Risk Factors of Long COVID Using the Behavioral Risk Factor Surveillance System

Team: Devin Gee, Juan P. Selame Fernandez

Overview: The CDC’s Behavioral Risk Factor Surveillance System (BRFSS) is an annual phone survey that collects information centered on health and behavior among American adults. The 2022 survey was the first year that included information relating to COVID, providing new opportunities to understand the behaviors, demographics, and health conditions linked to long COVID.

Dataset: Modeled after a Kaggle dataset that used the 2015 data

- **Source**: CDC’s Population Health Surveillance Branch
- **Features**: 22 in total including but not limited to: demographic (age, sex, race, income), lifestyle (sleep amount, exercise, diet,smoking), health (BMI, chronic illnesses, cholesterol, blood pressure)
- **Samples**: 121,379 responses (pre-processing)
- **Goal**: Predict whether or not the respondent has/had long COVID and identify key predictive features across models.

Work Done So Far: Previous project have attempted to predict Long COVID largely through the use of symptom focused datasets. None have used more sociobehavioral datasets to understand the backgrounds and contexts of those with the disease.

Methodology:

- **Preprocessing**: Clean the 2022 CDC data for analysis selecting features of interest and removing invalid samples; partition into training and testing sets. Strong focus on feature selection such as filter methods (Pearson correlation) and embedded methods (Lasso and Ridge regression)
- **Model Experimentation**: Test machine learning models including KNN,Logistic Regression, Random Forest, Naive Bayes and Neural Networks.
    - K-Fold Cross Validation to improve models
- **Optimization & Evaluation:** Standardize data, and tune hyperparameters with random search. Evaluate using AUROC and AUPRC due to dataset imbalance