**Final Project: Analysis Report**

Jasmin Martinez | P8106: Data Science II | May 12, 2025

**Exploratory cheAnalysis**

To better understand the variables in the severe\_flu dataset and explore potential associations between the outcome of interest and its predictors, an exploratory analysis was conducted. A bar plot was generated to visualize the distribution of severe flu cases in the sample, revealing that non-severe cases were more common than severe ones. The mean age of the sample was 60 years, with an average height of 169.7 cm and an average weight of 80.0 kg. The mean systolic blood pressure (SBP) was 129.9 mmHg, and the mean LDL cholesterol level was 110.5 mg/dL.

The distribution of categorical variables was examined using proportions stratified by flu severity. Among individuals with severe flu, 51.7% were women and 48.2% were men. White individuals had the highest proportion of severe flu cases at 64.8%. Among those who never smoked, 57.3% experienced severe flu. Additionally, 79.4% of individuals without diabetes had severe flu.

Boxplots were used to examine relationships between flu severity and continuous variables. On average, height was slightly greater among those without severe flu (170.2 cm vs. 167.9 cm). Weight, BMI, and LDL cholesterol levels were higher among individuals with severe flu compared to those without (84.0 kg vs. 79.4 kg; 29.6 kg/m² vs. 27.4 kg/m²; 115.3 mg/dL vs. 106.2 mg/dL, respectively).

**Model Training**

To evaluate predictive performance, we split the dataset into an 80% training set and a 20% test set using initial\_split(). All model fitting and cross-validation were performed on the training set, and final predictive performance was assessed on the test set. The outcome was a binary indicator for severe flu, and predictors included age, gender, race, smoking, height, weight, BMI, diabetes, hypertension, SBP, and LDL cholesterol.

A standard logistic regression model was first fit using the glm() function. It achieved modest accuracy on the test set and served as a baseline. To improve performance and prevent overfitting, we fit penalized logistic regression models (Ridge and Lasso) using the glmnet package with 10-fold cross-validation over 100 log-scaled lambda values. Ridge retained all predictors with small coefficients, while Lasso excluded some predictors entirely. Both models were tuned using train() with internal standardization.

Additional models included linear discriminant analysis (LDA), partial least squares (PLS), multivariate adaptive regression splines (MARS), support vector machines (SVM), and gradient boosting machines (GBM). Each was trained using the caret framework with repeated 10-fold cross-validation and grid-based tuning. MARS and GBM were included to capture nonlinear effects and interactions; GBM was tuned over tree depth, shrinkage, and number of trees. Radial SVM models were tuned using tune.svm() across exponential grids of cost and gamma values.

To ensure a fair comparison, all models were evaluated using consistent resampling and the ROC metric via trainControl() with summaryFunction = twoClassSummary. Ridge regression achieved the highest cross-validated ROC and was selected as the final model. It was then refit to the full dataset with the optimal lambda and used to generate participant-level predicted probabilities of severe flu.

Tuning plots and ROC comparisons for all models are shown in Figure X. Results confirm Ridge regression as the most robust model for estimating individual-level risk.

**Results**