P8106: Severe Flu Prediction Project

Naomi Simon-Kumar ns3782

05/10/2025

Contents

1. Introduction

1.1. Background and Study Objectives

This study aims to build prediction models for identifying factors associated with the incidence of severe flu within 6 months post-vaccination. The analysis focuses on a dataset of 1,000 participants to understand risk factors for severe flu in a population administered the flu vaccine.

The primary objectives of this project are to: (1) evaluate whether advanced predictive modeling techniques like boosting and support vector machines (SVM) provide superior predictive performance compared to simpler models, (2) develop a predictive risk score that quantifies the probability of experiencing severe flu based on individual characteristics, and (3) identify key demographic and clinical factors that predict severe flu risk and assess how these factors influence this risk.

1.2. Data Source and Description

The `severe_flu` dataset used for analysis contains demographic and clinical information from 1,000 participants. Variables include age (in years), gender (1 = Male, 0 = Female), race/ethnicity (1 = White, 2 = Asian, 3 = Black, 4 = Hispanic), smoking status (0 = Never smoked, 1 = Former smoker, 2 = Current smoker), height (in centimeters), weight (in kilograms), BMI (Body Mass Index, calculated as weight in kg divided by height in meters squared), diabetes status (0 = No, 1 = Yes), hypertension status (0 = No, 1 = Yes), systolic blood pressure (SBP, in mmHg), and LDL cholesterol (in mg/dL). The outcome of interest is a binary variable severe_flu indicating whether a participant experienced severe flu within 6 months post-vaccination (0 = No, 1 = Yes).

2. Exploratory Analysis

Exploratory data analysis was undertaken to examine the structure of the `severe_flu` dataset, including assessment of data distributions, and relationships between variables. The dataset contains 1,000 observations with 12 variables after removing the ID column. Categorical variables (gender, race, smoking status, diabetes, hypertension, and severe flu outcome) were converted to factors with appropriate labels. The dataset was verified to be complete with no missing observations.

Summary statistics were calculated for all demographic and clinical variables. Visualizations were produced to compare characteristics between participants who did and did not experience severe flu.

2.1. Summary Statistics

Table 1. Summary statistics of demographic and clinical characteristics in the study population

	Overall
	(N=1000)
Gender	
Female	522 (52.2%)
Male	478 (47.8%)
Race/Ethnicity	
White	656 (65.6%)
Asian	64 (6.4%)
Black	184 (18.4%)
Hispanic	96 (9.6%)
Smoking Status	
Never	584 (58.4%)
Former	313 (31.3%)
Current	103 (10.3%)
Height (cm)	
Mean (SD)	170 (6.11)
Median [Min, Max]	170 [152, 192]
Weight (kg)	

	Overall
Mean (SD)	80.0 (7.12)
Median [Min, Max]	80.1 [59.1, 104]
Body Mass Index (weight/ (height)^2)	
Mean (SD)	27.9 (2.76)
Median [Min, Max]	27.7 [20.1, 36.7]
Diabetes	
No	855 (85.5%)
Yes	145 (14.5%)
Hypertension	
No	536 (53.6%)
Yes	464 (46.4%)
Systolic Blood Pressure (mmHg)	
Mean (SD)	130 (7.88)
Median [Min, Max]	130 [108, 154]
LDL cholesterol (mg/dL)	
Mean (SD)	110 (19.7)
Median [Min, Max]	111 [41.0, 174]

Table 1 summarises the demographic and clinical characteristics of the study population (N = 1000). The sample was balanced by gender (52.2% female, 47.8% male), with predominantly White participants (65.6%). Most participants were never smokers (58.4%), with 14.5% reporting diabetes and 46.4% identified as hypertensive. Notably, the mean BMI was 27.9 kg/m² (SD=2.76), indicating the average subject would be considered clinically overweight. Mean systolic blood pressure was 130 mmHg (SD=7.88), and mean LDL cholesterol was 110 mg/dL (SD=19.7).

2.2. Exploratory Plots

2.2.1. Histograms

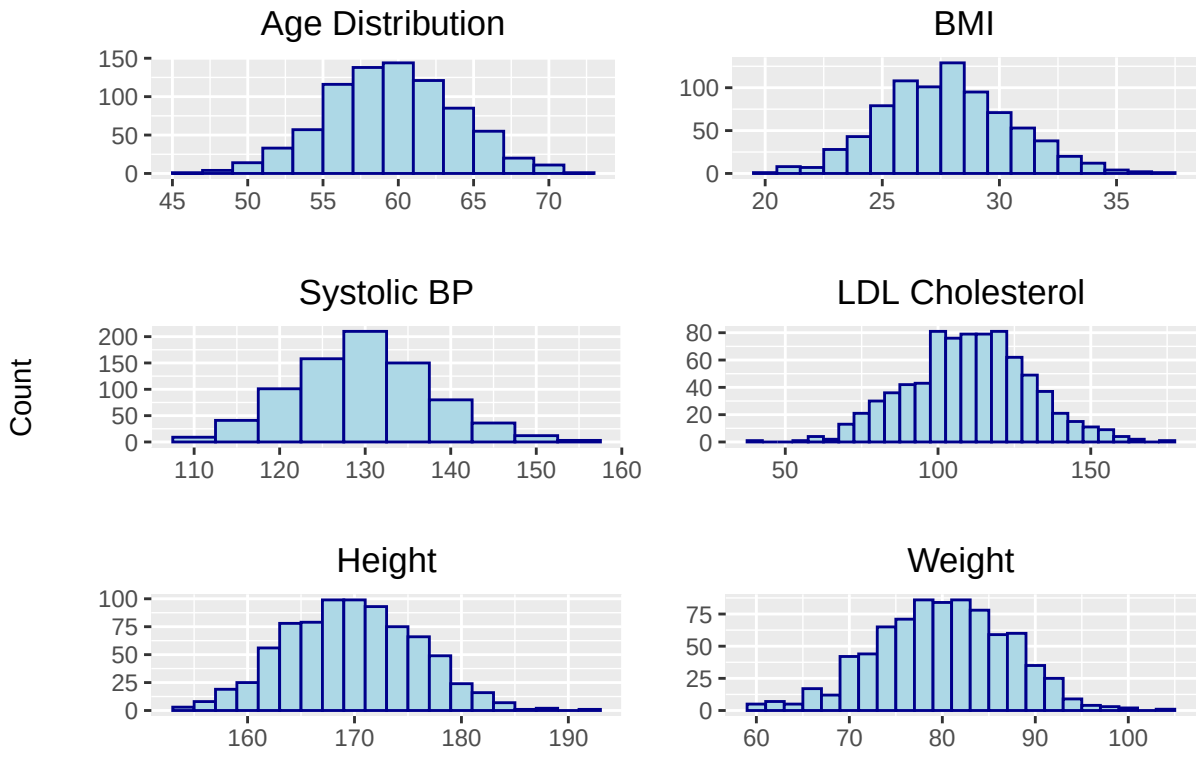

Figure 1. Distribution of continuous predictors in the study dataset

Figure 1 presents histogram plots for numeric predictors in the study dataset. Age, BMI, SBP, LDL, height and weight are approximately normally distributed, with minimal right skew for LDL.

2.2.2. Boxplots

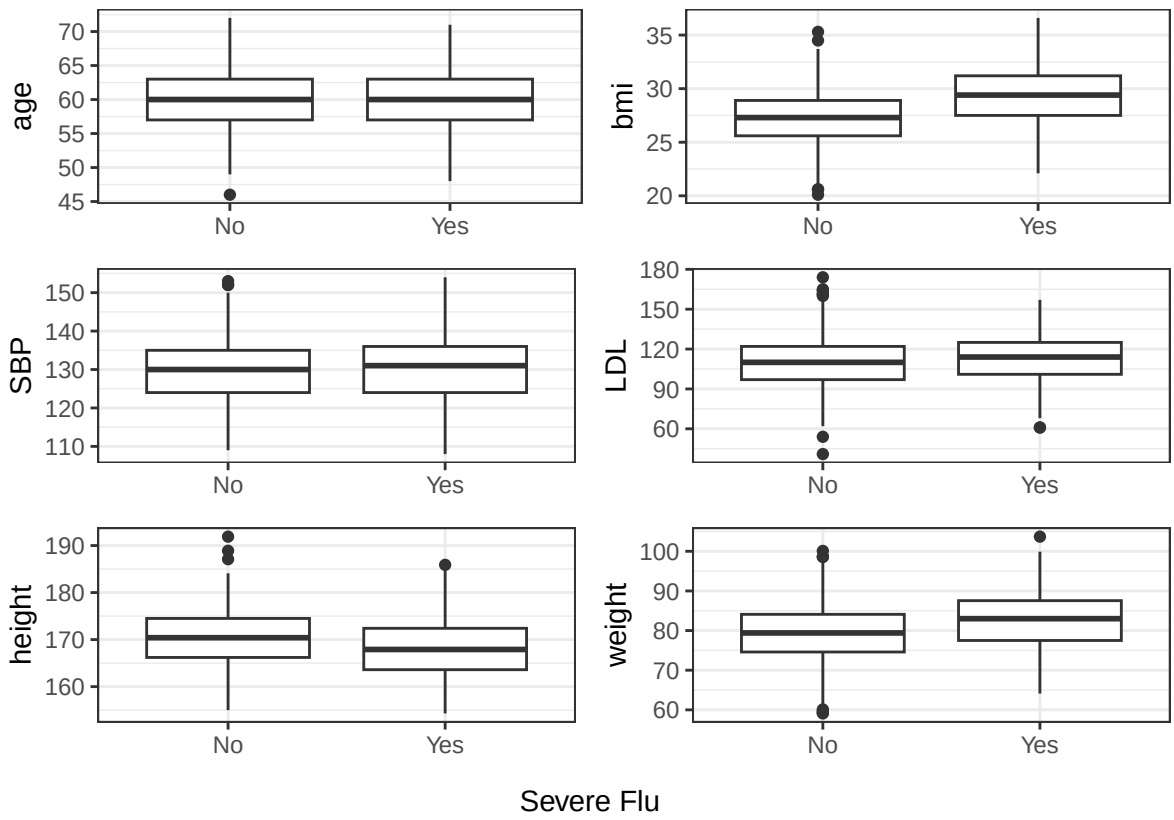

Figure 2. Distribution of continuous predictors by severe flu status in the study dataset

From the boxplots of continuous predictors, age, SBP, LDL, and height distributions appear similar between severe flu and non-severe flu groups, with considerable middle quartile overlap (Figure 2). BMI and weight show slightly higher median values in the severe flu group, however, the substantial overlap in their distributions suggest these differences may not be statistically significant.

2.2.2. Barplots

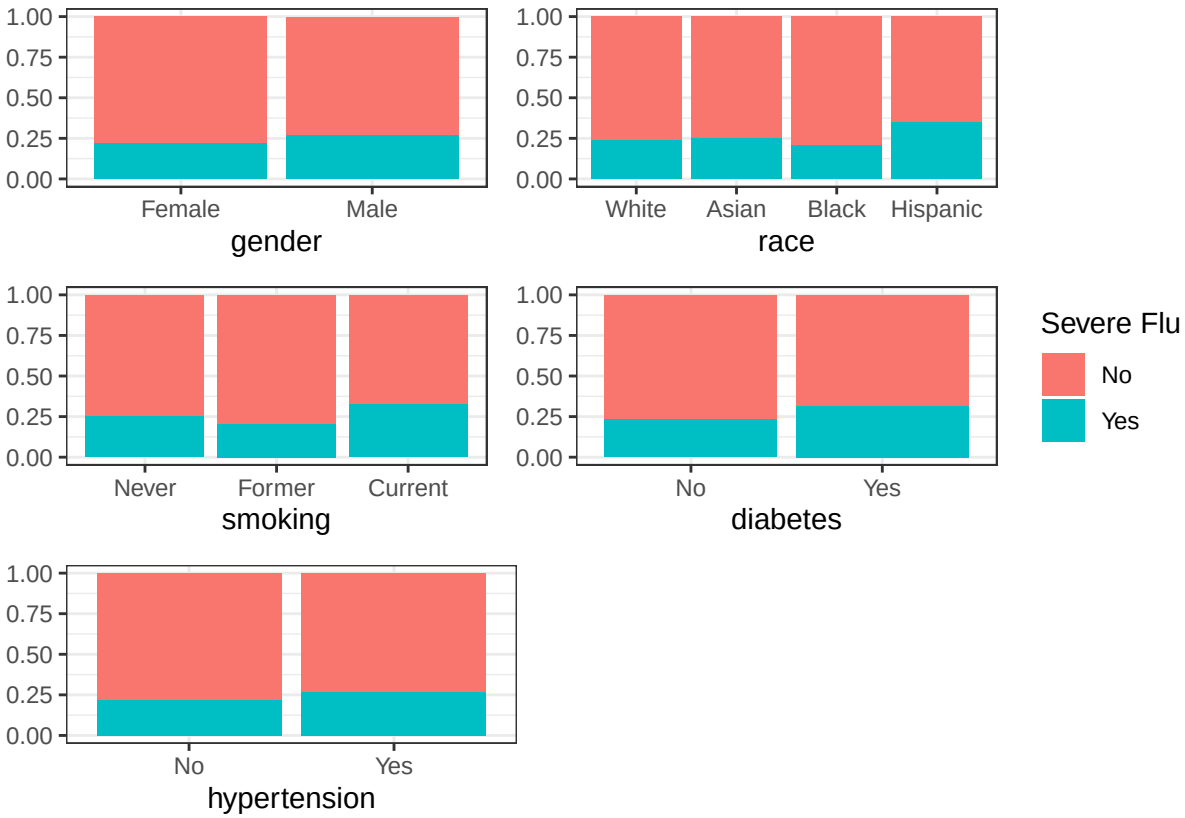

Figure 3. Proportion of severe flu outcomes across categorical variables in the study dataset

Among categorical predictors, there appears to be differences in severe flu rates (Figure 3). Hispanic subjects show a higher proportion of severe flu cases compared to other racial groups. Diabetic patients also appear to have a slightly higher severe flu rate than non-diabetic individuals. Similarly current smokers have a slightly higher proportion of severe flu cases compared to never and former smokers. Gender and hypertension status show minimal differences in severe flu proportions across categories.

2.3. Correlation Analysis

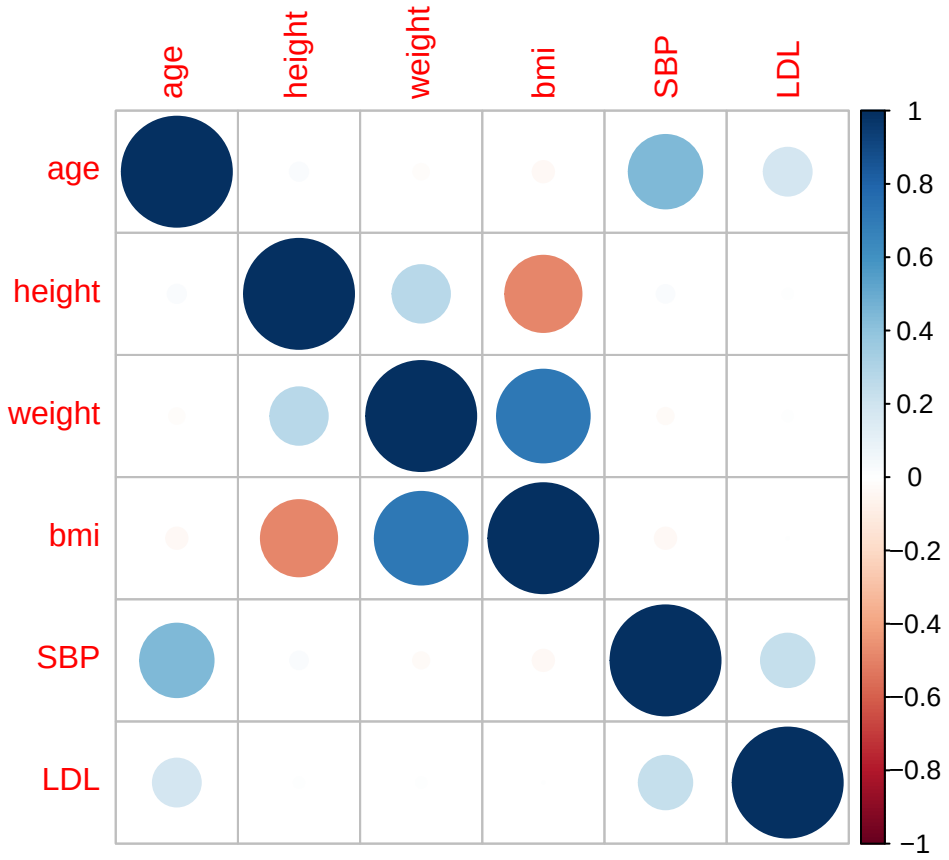

Figure 4. Correlation matrix between continuous predictors in the study dataset

The correlation plot shows several relationships among continuous predictors. Weight and BMI have a strong positive correlation ($r \approx 0.8$), which may indicate collinearity. This is expected as BMI is calculated directly from weight and height. Therefore, interpretation of model results, including identifying important predictors of severe flu, should consider that these variables overlap in what they measure. Age and SBP have a moderate positive correlation ($r \approx 0.4$). Height and weight are weakly positively correlated ($r \approx 0.2$). By comparison, height has a moderate negative correlation with BMI ($r \approx -0.4$).

3. Model Training

3.1. Data preprocessing

The data was examined to identify missing values and undertake data cleaning. No missing observations were identified. Categorical variables were converted into appropriate factor variables: gender (Female/Male), race/ethnicity (White, Asian, Black, Hispanic), smoking status (Never, Former, Current), diabetes (No/Yes), hypertension (No/Yes), and the response variable severe flu (No/Yes). The ID variable was removed as it was not a meaningful predictor for flu severity. The preprocessed dataset was then split into training (80%) and testing (20%) sets for model development and evaluation.

3.2. Model Evaluation metrics

Area Under the Receiver Operating Characteristic curve (AUC) was selected as the metric for evaluating classification performance across all candidate models. AUC measures a model's ability to discriminate between severe and non-severe flu cases across all possible classification thresholds, with values ranging from 0.5 (no better than random chance) to 1.0 (perfect discrimination) (Hosmer et al., 2013). AUC was computed during cross-validation (10-fold with 5 repeats) on the training dataset to select optimal model parameters. Final model performance was then evaluated on the held-out test dataset. This approach allowed for fair comparison between simple models (logistic regression, LDA) and more complex techniques (SVM, boosting) while being robust to class imbalance in the dataset (Kuhn & Johnson, 2013). Confusion matrices were generated for the final selected model to report classification accuracy, sensitivity, and specificity at the 0.5 probability threshold, following the Bayes decision rule, which assigns an observation to the most likely class. This was undertaken to assess predictive performance after model selection based on cross-validated AUC.

3.3. Models

References

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). John Wiley & Sons.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.RetryClaude can make mistakes. Please double-check responses.