Data Science II Final Project Analysis

Camille Okonkwo

Contents

```r
library(tidymodels)
library(splines)
library(caret)
library(glmnet)
library(table1)
library(kableExtra)
library(summarytools)
library(corrplot)
library(cowplot)
```

Background

A research study aims to identify key factors that predict the severity of COVID-19 illness. This study collects demographic information, clinical variables, and disease severity among participants infected with COVID-19 between 2021 and 2023. The goal is to develop a robust prediction model that can accurately predict COVID-19 severity and understand how predictors impact the risk of severe infection.

Data

The training data in "severity_training.RData" includes data from 800 participants.

The test data in "severity_test.RData" includes data from another set of 200 participants.

Here is a description of each variable:

- ID (`id`): Participant ID
- Age (`age`): Age
- Gender (`gender`): 1 = Male, 0 = Female
- Race/ethnicity (`race`): 1 = White, 2 = Asian, 3 = Black, 4 = Hispanic
- Smoking (`smoking`): Smoking status; 0 = Never smoked, 1 = Former smoker, 2 = Current smoker
- Height (`height`): Height (in centimeters)
- Weight (`weight`): Weight (in kilograms)
- BMI (`bmi`): Body Mass Index; BMI = weight (in kilograms) / height (in meters) squared
- Hypertension (`hypertension`): 0 = No, 1 = Yes
- Diabetes (`diabetes`): 0 = No, 1 = Yes
- Systolic blood pressure (`SBP`): Systolic blood pressure (in mm/Hg)
- LDL cholesterol (`LDL`): LDL (low-density lipoprotein) cholesterol (in mg/dL)
- Vaccination status at the time of infection (`vaccine`): 0 = Not vaccinated, 1 = Vaccinated
- Depression score (`depression`): Higher scores indicate higher risk for depression
- Severity of COVID-19 infection (`severity`): **Response variable**; 0 = Not severe, 1 = Severe

Data Preparation

```
# loading training data
load("data/severity_training.RData") # is depression discrete?

# making discrete variables factors
training_data = training_data |>
  janitor::clean_names() |>
  select(-id) |>
  mutate(gender = as.factor(gender),
         race = as.factor(race),
         smoking = as.factor(smoking),
         hypertension = as.factor(hypertension),
         diabetes = as.factor(diabetes),
         vaccine = as.factor(vaccine),
         severity = as.factor(severity))

# matrix of predictors & vector of response for data set exploration
x_train = model.matrix(severity ~., training_data)[, -1]
y_train = training_data$severity

# loading testing data
load("data/severity_test.RData")
```

```r
# making discrete variables factors
test_data = test_data |>
  janitor::clean_names() |>
  select(-id) |>
  mutate(gender = as.factor(gender),
         race = as.factor(race),
         smoking = as.factor(smoking),
         hypertension = as.factor(hypertension),
         diabetes = as.factor(diabetes),
         vaccine = as.factor(vaccine),
         severity = as.factor(severity))

# matrix of predictors and vector of response
x_test = model.matrix(severity ~., test_data)[, -1]
y_test = test_data$severity
```

Exploratory analysis and data visualization

Descriptive Statistics

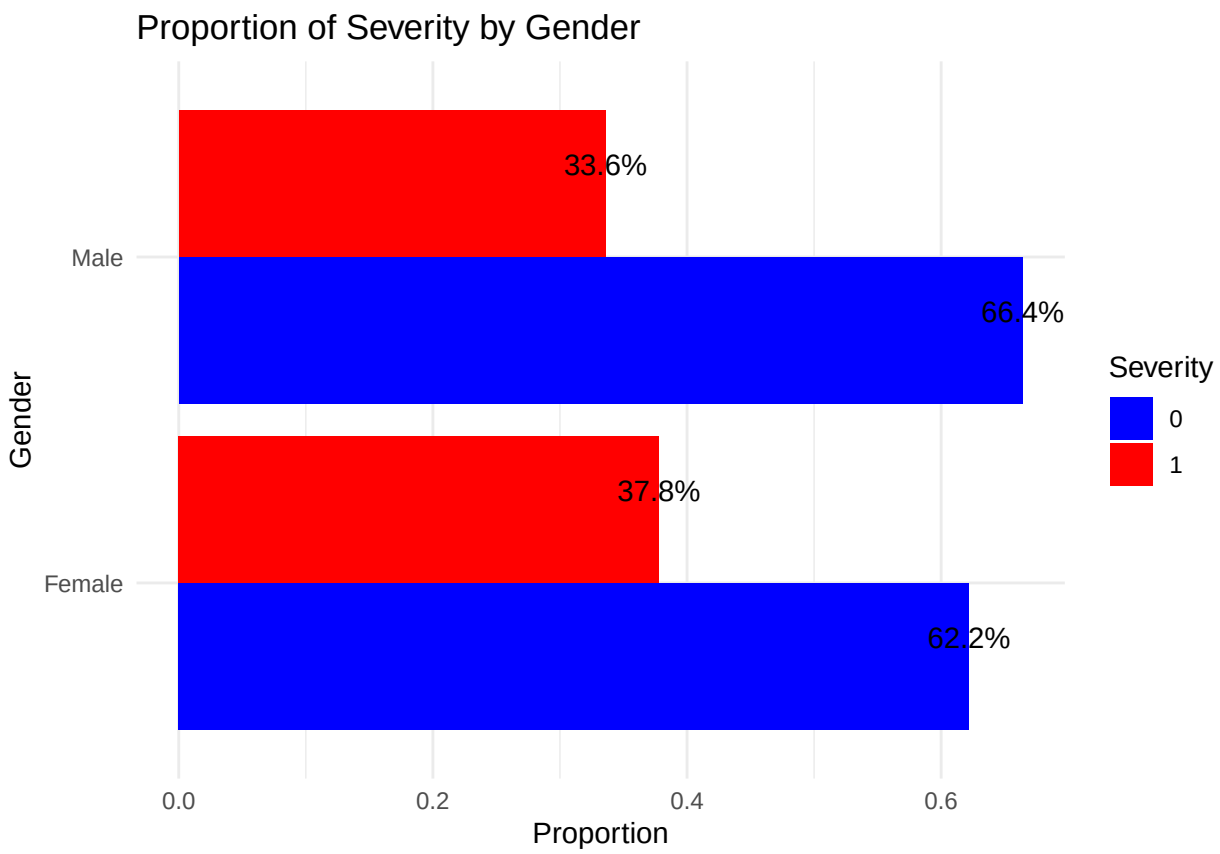
```
descriptive_table = table1(~ age + gender + race + smoking + height + weight + bmi + hypertension + dial
                           data = training_data,
                           overall = "Total",
                           caption = "Descriptive Characteristics of Participants, Stratified by Severi

ds = t1kable(descriptive_table)
ds
```

Discrete Variable Visualization

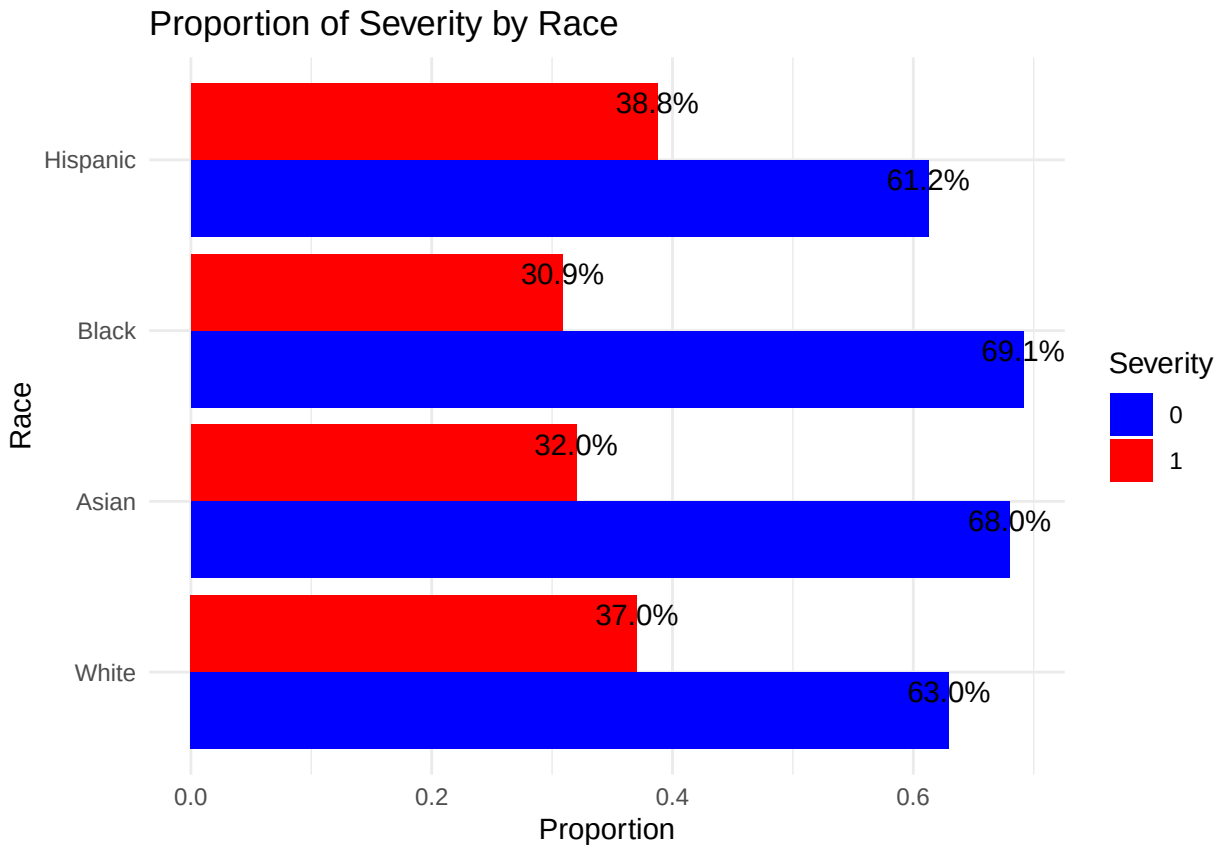
```
# gender x severity
prop_gender <- training_data %>%
  group_by(gender, severity) %>%
  summarise(count = n()) %>%
  mutate(prop = count / sum(count))

gender_labels = c("1" = "Male", "0" = "Female")

ggplot(prop_gender,
       aes(y = factor(gender),
           x = prop,
           fill = factor(severity), label = scales::percent(prop))) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(position = position_dodge(width = 0.9),  vjust = -0.4) +
  labs(x = "Proportion", y = "Gender", fill = "Severity") +
  ggtitle("Proportion of Severity by Gender") +
  scale_fill_manual(values = c("0" = "blue", "1" = "red")) +
  scale_y_discrete(labels = gender_labels) +
  theme_minimal()
```

Table 1: Descriptive Characteristics of Participants, Stratified by Severity of COVID-19 Infection

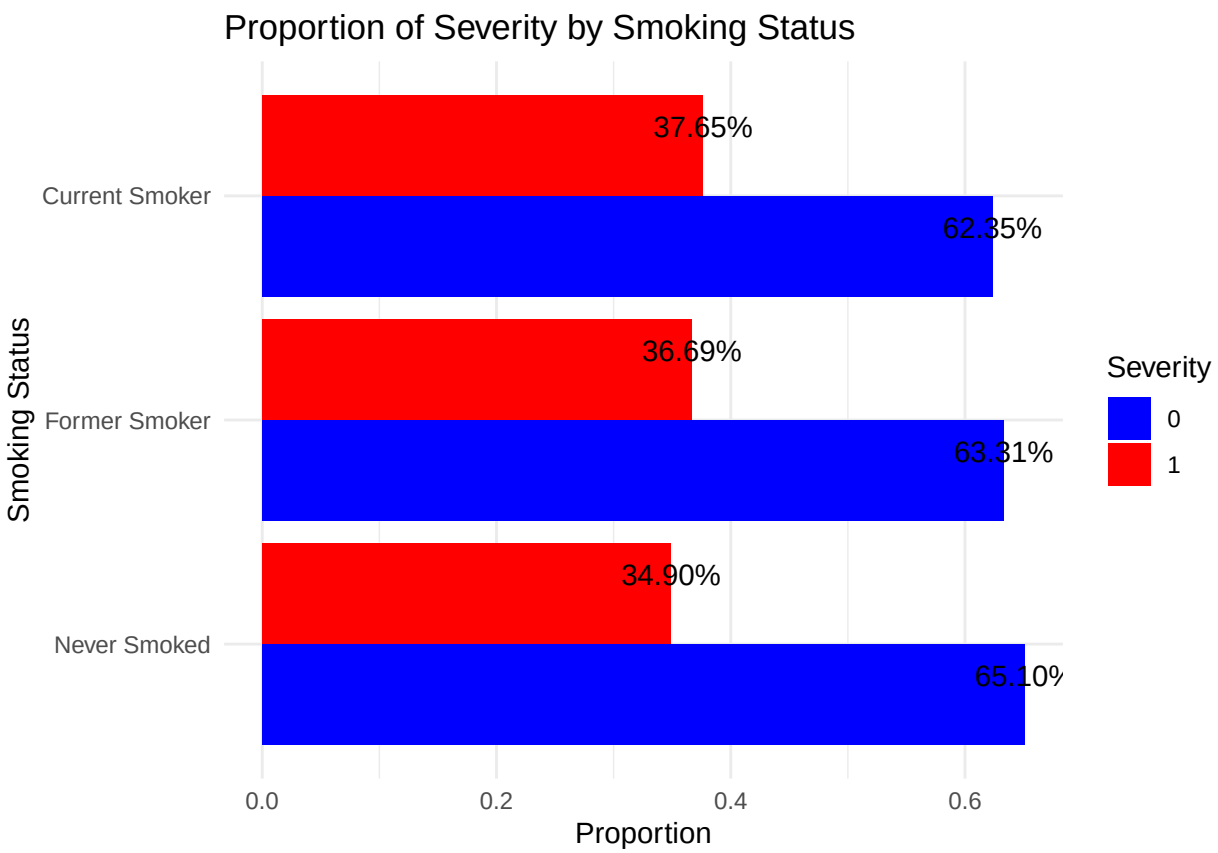
	0	1	Total
	(N=514)	(N=286)	(N=800)
age			
Mean (SD)	59.5 (4.29)	61.0 (4.12)	60.0 (4.30)
Median [Min, Max]	59.0 [46.0, 71.0]	61.0 [48.0, 72.0]	60.0 [46.0, 72.0]
gender			
0	255 (49.6%)	155 (54.2%)	410 (51.3%)
1	259 (50.4%)	131 (45.8%)	390 (48.8%)
race			
1	328 (63.8%)	193 (67.5%)	521 (65.1%)
2	34 (6.6%)	16 (5.6%)	50 (6.3%)
3	103 (20.0%)	46 (16.1%)	149 (18.6%)
4	49 (9.5%)	31 (10.8%)	80 (10.0%)
smoking			
0	304 (59.1%)	163 (57.0%)	467 (58.4%)
1	157 (30.5%)	91 (31.8%)	248 (31.0%)
2	53 (10.3%)	32 (11.2%)	85 (10.6%)
height			
Mean (SD)	170 (6.24)	170 (5.83)	170 (6.09)
Median [Min, Max]	170 [150, 187]	170 [152, 190]	170 [150, 190]
weight			
Mean (SD)	79.0 (7.33)	80.1 (7.09)	79.4 (7.26)
Median [Min, Max]	79.2 [56.6, 105]	79.9 [59.0, 104]	79.3 [56.6, 105]
bmi			
Mean (SD)	27.4 (2.70)	27.9 (2.78)	27.5 (2.74)
Median [Min, Max]	27.4 [19.6, 37.4]	27.9 [19.9, 36.9]	27.6 [19.6, 37.4]
hypertension			
0	332 (64.6%)	100 (35.0%)	432 (54.0%)
1	182 (35.4%)	186 (65.0%)	368 (46.0%)
diabetes			
0	437 (85.0%)	242 (84.6%)	679 (84.9%)
1	77 (15.0%)	44 (15.4%)	121 (15.1%)
sbp			
Mean (SD)	128 (7.58)	133 (7.62)	130 (7.97)
Median [Min, Max]	128 [109, 154]	134 [111, 153]	130 [109, 154]
ldl			
Mean (SD)	108 (20.5)	113 (18.8)	110 (20.1)
Median [Min, Max]	110 [41.0, 167]	112 [54.0, 174]	111 [41.0, 174]
vaccine			
0	96 (18.7%)	240 (83.9%)	336 (42.0%)
1	418 (81.3%)	46 (16.1%)	464 (58.0%)
depression			
Mean (SD)	6.91 (2.13)	6.90 (2.09)	6.91 (2.12)
Median [Min, Max]	7.00 [0, 13.0]	7.00 [1.00, 13.0]	7.00 [0, 13.0]



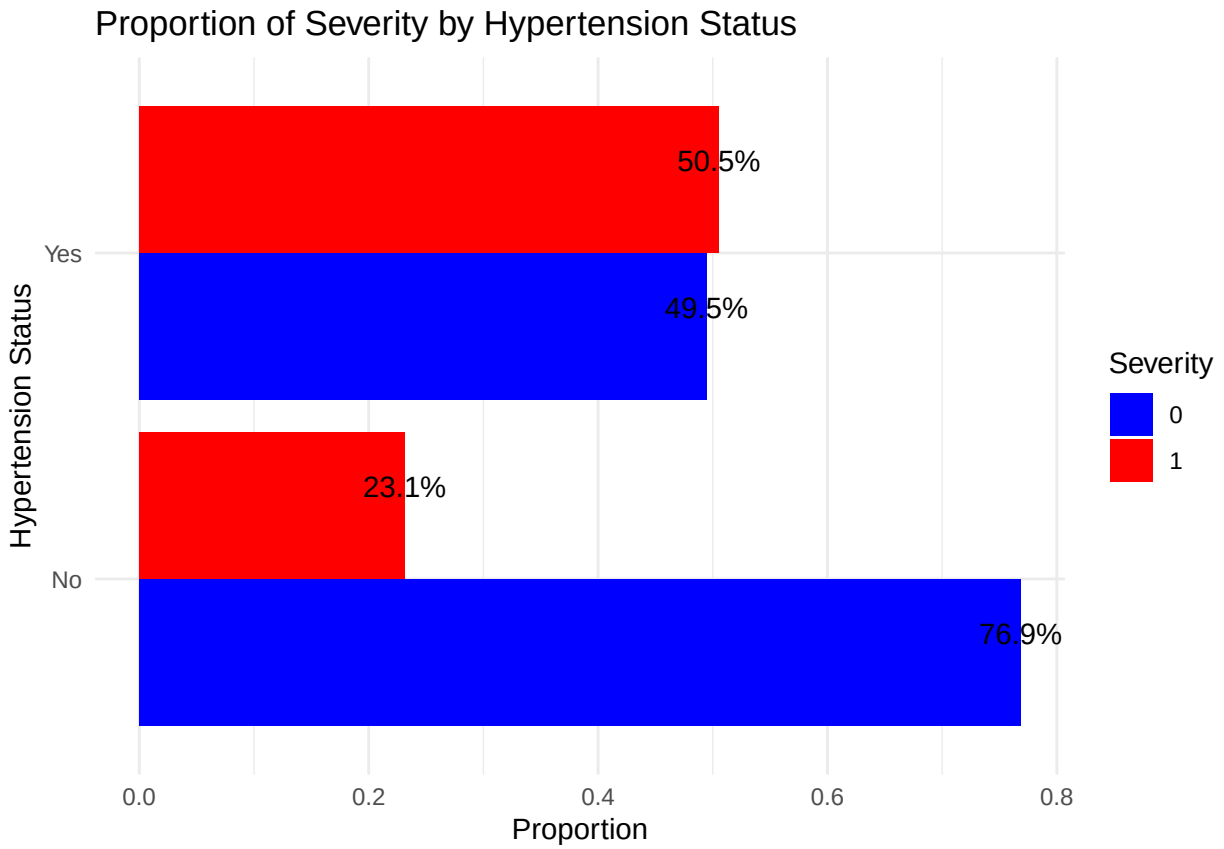

```
# race x severity
prop_race <- training_data %>%
  group_by(race, severity) %>%
  summarise(count = n()) %>%
  mutate(prop = count / sum(count))

race_labels = c("1" = "White", "2" = "Asian", "3" = "Black", "4" = "Hispanic")

ggplot(prop_race,
       aes(y = factor(race),
           x = prop,
           fill = factor(severity), label = scales::percent(prop))) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(position = position_dodge(width = 0.9),  vjust = -0.4) +
  labs(x = "Proportion", y = "Race", fill = "Severity") +
  ggtitle("Proportion of Severity by Race") +
  scale_fill_manual(values = c("0" = "blue", "1" = "red")) +
  scale_y_discrete(labels = race_labels) +
  theme_minimal()
```



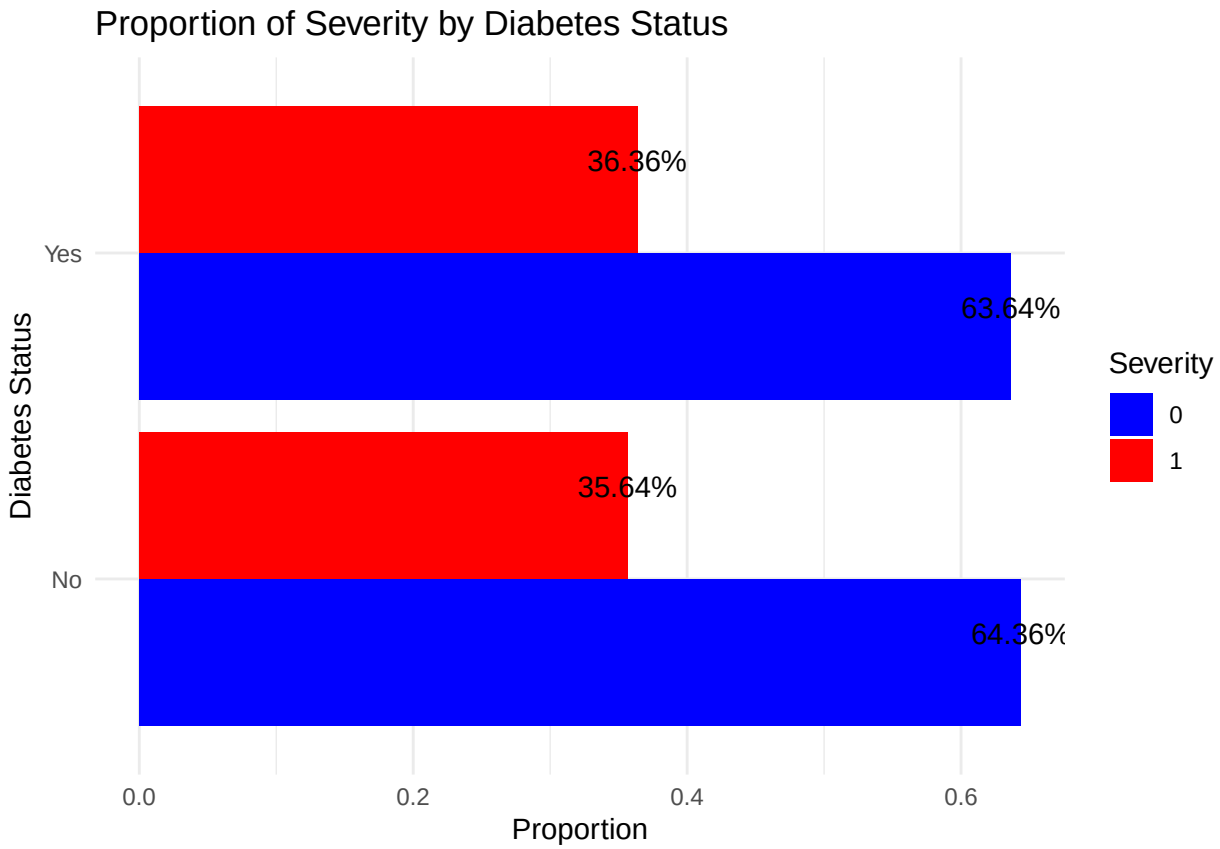

```
# smoking status x severity
prop_smoking <- training_data %>%
  group_by(smoking, severity) %>%
  summarise(count = n()) %>%
  mutate(prop = count / sum(count))

smoking_labels = c("1" = "Former Smoker", "2" = "Current Smoker", "0" = "Never Smoked")

ggplot(prop_smoking,
       aes(y = factor(smoking),
           x = prop,
           fill = factor(severity), label = scales::percent(prop))) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(position = position_dodge(width = 0.9),  vjust = -0.4) +
  labs(x = "Proportion", y = "Smoking Status", fill = "Severity") +
  ggtitle("Proportion of Severity by Smoking Status") +
  scale_fill_manual(values = c("0" = "blue", "1" = "red")) +
  scale_y_discrete(labels = smoking_labels) +
  theme_minimal()
```



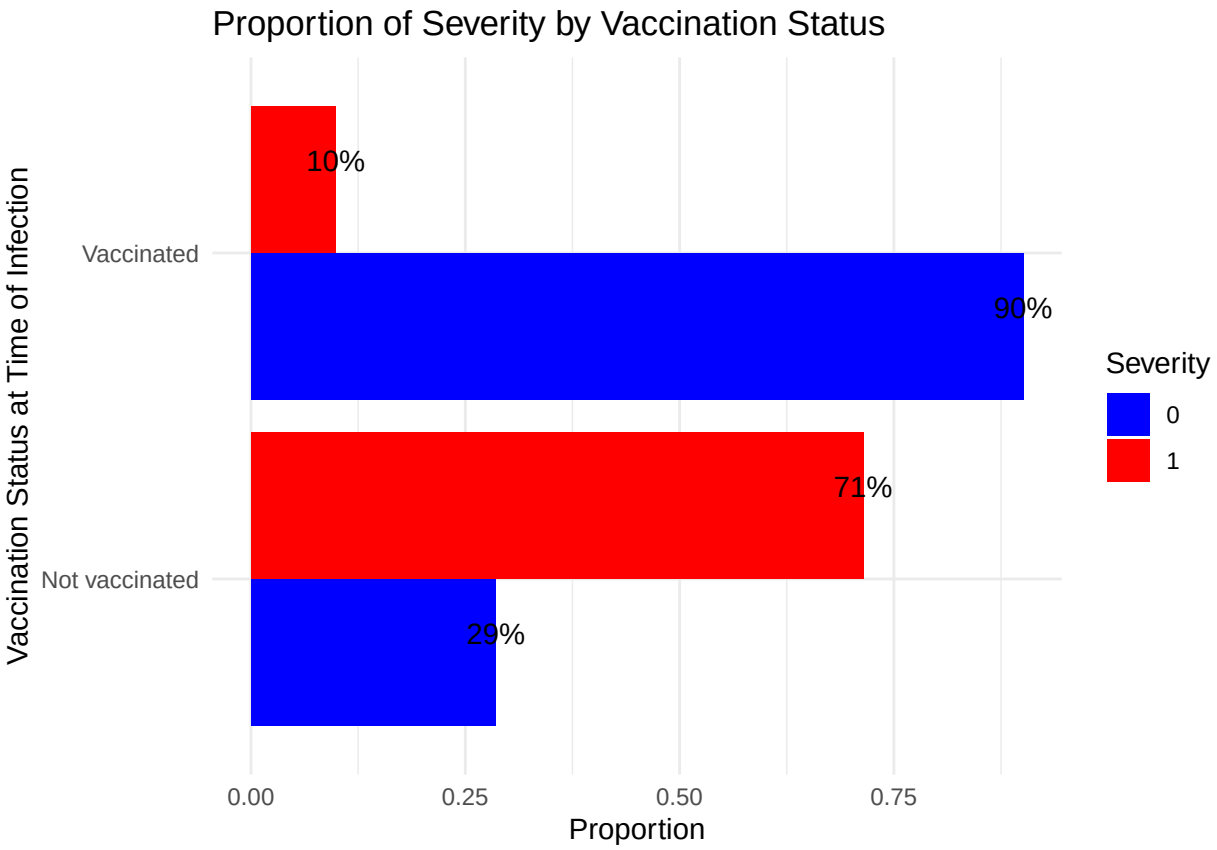

```
# hypertension status x severity
prop_hyp <- training_data %>%
  group_by(hypertension, severity) %>%
  summarise(count = n()) %>%
  mutate(prop = count / sum(count))

hyp_labels = c("0" = "No", "1" = "Yes")

ggplot(prop_hyp,
       aes(y = factor(hypertension),
           x = prop,
           fill = factor(severity), label = scales::percent(prop))) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(position = position_dodge(width = 0.9),  vjust = -0.4) +
  labs(x = "Proportion", y = "Hypertension Status", fill = "Severity") +
  ggtitle("Proportion of Severity by Hypertension Status") +
  scale_fill_manual(values = c("0" = "blue", "1" = "red")) +
  scale_y_discrete(labels = hyp_labels) +
  theme_minimal()
```




```r
# diabetes x severity
prop_dia <- training_data %>%
  group_by(diabetes, severity) %>%
  summarise(count = n()) %>%
  mutate(prop = count / sum(count))

dia_labels = c("0" = "No", "1" = "Yes")

ggplot(prop_dia,
       aes(y = factor(diabetes),
           x = prop,
           fill = factor(severity), label = scales::percent(prop))) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(position = position_dodge(width = 0.9),  vjust = -0.4) +
  labs(x = "Proportion", y = "Diabetes Status", fill = "Severity") +
  ggtitle("Proportion of Severity by Diabetes Status") +
  scale_fill_manual(values = c("0" = "blue", "1" = "red")) +
  scale_y_discrete(labels = dia_labels) +
  theme_minimal()
```

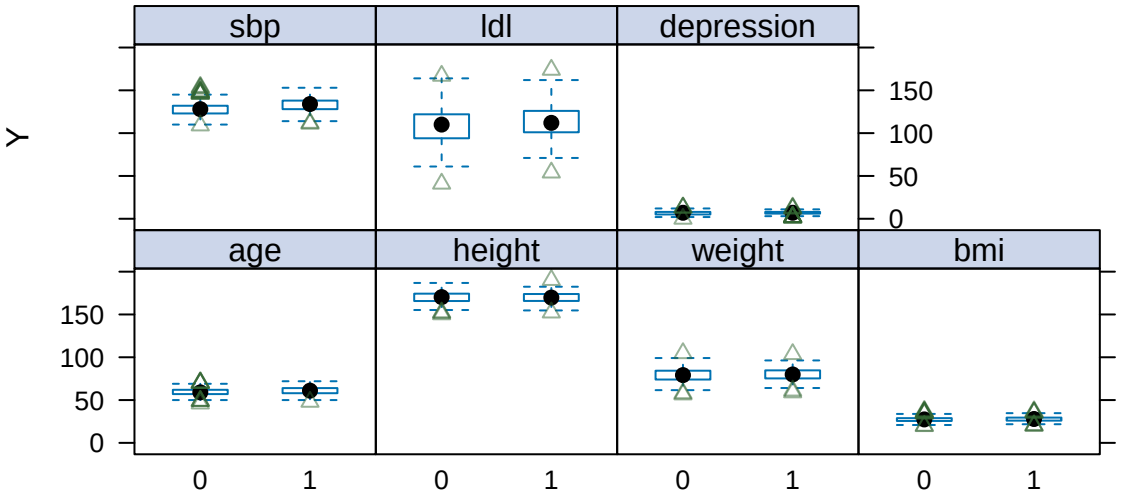



```
# vaccine x severity
prop_vaccine <- training_data %>%
  group_by(vaccine, severity) %>%
  summarise(count = n()) %>%
  mutate(prop = count / sum(count))

vax_labels = c("0" = "Not vaccinated", "1" = "Vaccinated")

ggplot(prop_vaccine,
       aes(y = factor(vaccine),
           x = prop,
           fill = factor(severity), label = scales::percent(prop))) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(position = position_dodge(width = 0.9),  vjust = -0.4) +
  labs(x = "Proportion", y = "Vaccination Status at Time of Infection", fill = "Severity") +
  ggtitle("Proportion of Severity by Vaccination Status") +
  scale_fill_manual(values = c("0" = "blue", "1" = "red")) +
  scale_y_discrete(labels = vax_labels) +
  theme_minimal()
```



Continuous Variable Visualization

```
theme1 <- trellis.par.get()
theme1$plot.symbol$col <- rgb(.2, .4, .2, .5)
theme1$plot.symbol$pch <- 2
theme1$plot.line$col <- rgb(.8, .1, .1, 1)
theme1$plot.line$lwd <- 2
theme1$strip.background$col <- rgb(.0, .2, .6, .2)
trellis.par.set(theme1)

featurePlot(
  x_train[, -c(2, 3, 4, 5, 6, 7, 11, 12, 15)],
  y_train,
  plot = "box",
  labels = c("", "Y"),
  type = c("p", "smooth"),
  layout = c(4, 3))
```


sbp
ldl
depression
age
height
weight
bmi
Y
150
100
50
0
0
1

Model training

Results

Conclusion