Comprehension Check: Practice with Machine Learning, Part 1

The following questions all ask you to work with the dataset described below.

The **reported_heights** and **heights** datasets were collected from three classes taught in the Departments of Computer Science and Biostatistics, as well as remotely through the Extension School. The Biostatistics class was taught in 2016 along with an online version offered by the Extension School. On 2016-01-25 at 8:15 AM, during one of the lectures, the instructors asked student to fill in the sex and height questionnaire that populated the **reported_heights** dataset. The online students filled out the survey during the next few days, after the lecture was posted online. We can use this insight to define a variable which we will call **type**, to denote the type of student, **inclass** or **online**.

The code below sets up the dataset for you to analyze in the following exercises:

```
library(dslabs)
library(dplyr)
library(lubridate)

data("reported_heights")

dat <- mutate(reported_heights, date_time = ymd_hms(time_stamp)) %>%
  filter(date_time >= make_date(2016, 01, 25) & date_time < make_date(2016, 02,
1)) %>%
  mutate(type = ifelse(day(date_time) == 25 & hour(date_time) == 8 &
between(minute(date_time), 15, 30), "inclass","online")) %>%
  select(sex, type)

y <- factor(dat$sex, c("Female", "Male"))
x <- dat$type
```

Q1

2/2 points (graded)

The `type` column of `dat` indicates whether students took classes in person ("inclass") or online ("online"). What proportion of the inclass group is female? What proportion of the online group is female?

In class

0.667 ✔ **Answer:** 0.667

0.667

Online

0.378 ✔ **Answer:** 0.378

0.378

Explanation
One way to find these values is by using the following code:
`dat %>% group_by(type) %>% summarize(prop_female = mean(sex == "Female"))`.

Submit You have used 1 of 5 attempts

Answers are displayed within the problem

Q2

1/1 point (graded)
In the course videos, height cutoffs were used to predict sex. Instead of using height, use the `type` variable. Use what you learned about Q1 to make an informed guess about sex based on the most prevalent sex for each type. Report the accuracy of your prediction of sex based on type. You do not need to split the data into training and test sets.

Enter your accuracy as a percentage or decimal (eg "50%" or "0.50") to at least the hundredths place.

0.63 ✔ **Answer:** 0.63

0.63

Explanation
This prediction could be done using the following code:

```
y_hat <- ifelse(x == "online", "Male", "Female") %>%
      factor(levels = levels(y))
mean(y_hat==y)
```

Submit You have used 1 of 5 attempts

Answers are displayed within the problem

Q3

1/1 point (graded)

Write a line of code using the `table` function to show the confusion matrix between `y_hat` and `y`. Use the exact format `function(a, b)` for your answer and do not name the columns and rows.

table(y_hat, y)

✔ **Answer:** table(y_hat, y) **or** table(y, y_hat)

Explanation

`table(y_hat, y)` will show the confusion matrix.

Submit

You have used 1 of 3 attempts

Answers are displayed within the problem

Q4

1/1 point (graded)

What is the sensitivity of this prediction? You can use the `sensitivity` function from the **caret** package. Enter your answer as a percentage or decimal (eg "50%" or "0.50") to at least the hundredths place.

0.3823529

✔ **Answer:** 0.382

0.3823529

Explanation

The following example code will give the sensitivity:

```
library(caret)
sensitivity(y_hat, y)
```

Submit

You have used 1 of 5 attempts

Answers are displayed within the problem

Q5

1/1 point (graded)

What is the specificity of this prediction? You can use the `specificity` function from the **caret** package. Enter your answer as a percentage or decimal (eg "50%" or "0.50") to at least the hundredths place.

0.8414634 ✔ **Answer:** 0.841

0.8414634

Explanation

The following example code will give the specificity:

```
library(caret)
specificity(y_hat, y)
```

Submit You have used 1 of 5 attempts

Answers are displayed within the problem

Q6

1/1 point (graded)

What is the prevalence (% of females) in the `dat` dataset defined above? Enter your answer as a percentage or decimal (eg "50%" or "0.50") to at least the hundredths place.

0.4533333 ✔ **Answer:** 0.453

0.4533333

Explanation

`mean(y == "Female")` will give the prevalence of females in the dataset.

Submit You have used 1 of 5 attempts

Answers are displayed within the problem