HarvardX PH125.9x Data Science: Capstone Project II
Applying Machine Learning Techniques to the SexLab Dataset

Luke Holmes Ph.D., University of Essex

2022-11-06

Introduction

Background

Experimental Psychology has been formally exploring the topic of sexual orientation since the Kinsey Institute was founded in 1947. Often, the approach to conducting these experiments is that participants are asked to self-report their sexual orientation, and then various measures are taken from them - for example, how masculine or feminine they consider themselves compared to others of the same sex, or how they respond to videos of attractive people.

When analysing this data, the approach often taken is to use the self-reported sexual orientation as the predictor, and examine its correlation with the other measures. However, we could also approach this from differently, by attempting to predict sexual orientation using the other variables. Approaching the problem "backwards" in this manner allows us to use multiple predictors in our predictive models, and thus fine-tune our predictions to explain as much of the variance in sexual orientation as possible.

Dataset

The current dataset is real data gathered in the Human Sexuality Laboratory in the Department of Psychology at the University of Essex over the course of several years. It contains anonymised data from 279 male participants who visited the laboratory in person during this time. Each participant self-reported their sexual orientation, then took part in various tasks, during which data was gathered. Specific information on the variables is given in the Method section.

Project Goal

The current project is the second capstone coursework project for the HarvardX PH125.9x Data Science course. Our goal is to predict sexual orientation using several different kinds of models: Firstly, Linear Regression, which was the model used to analyse this data when it was first gathered. This is not a "true" Machine Learning algorithm, but it is included for the same of completeness and as a learning exercise for R. Secondly we will use K-Nearest Neighbours, a supervised learning algorithm which makes predictions based on "distance" between data points. Finally, we will use a Classification Tree, another kind of model which repeatedly partitions data in order to make predictions about the outcome variable based on the other variables.

Key Steps

- **Introduction**: Background, dataset and project goals will be explained.
- **Methods**: All variables will be explained, and the dataset will be explored using summary statistics and visualisations.
- **Modelling**: A series of Machine Learning Models will be trained and tested on the data: Linear Regression, K-Nearest Neighbours and Classification Trees will be used.
- **Conclusions**: The models will be compared with one another, and the conclusions which can be drawn from the models will be explained.

Methods

Obtaining the Dataset

The dataset must first be downloaded, some column types changed, and the columns renamed appropriately.

```r
# Downloads and loads the sexlab dataset from my github repository
sexlabdata <- read.csv(
  "https://raw.githubusercontent.com/LukeH91/SexLabProject/main/sexlab-r-datasheet.csv",
                      head=TRUE, sep=",",encoding="utf-8")


# changes the column names to more R-friendly ones
colnames(sexlabdata) <- c("id", "sexual_orientation", "pupil_dilation",
                          "subjective_ratings", "self_reported_gnc", "observer_rated_gnc")

# convert two columns from character to numeric for analyses
sexlabdata$pupil_dilation <- as.numeric(sexlabdata$pupil_dilation)
sexlabdata$subjective_ratings <- as.numeric(sexlabdata$subjective_ratings)
```

Data Exploration

We will first look at the variables available in our data set.

```
## Column Names:  id sexual_orientation pupil_dilation
```

```
## [1] "subjective_ratings" "self_reported_gnc"  "observer_rated_gnc"
```

In order, we have:

- **ID**: A unique ID given to each participant. This is only used to verify that no-one ends up in both the training and test sets.
- **Sexual Orientation**: A participant's sexual orientation, measured on a 7-point Kinsey scale from 0 (exclusively straight), through 3 (bisexual with no preference) to 6 (exclusively gay). For verification purposes, participants were also asked about their sexual attraction on a similar scale, and the two scores were averaged to give this composite score. Hence, some participants have a sexual orientation score ending in .5.

- **Pupil Dilation (PD)**: Participants were shown a series of videos featuring attractive men and women, and their pupil dilation was measured as they watched. In theory, pupils dilate when an individual is watching a video of someone they find attractive. The average dilation value while watching women was then deducted from the dilation value when watching men. Thus, a participant with a negative pupil dilation score responded more strongly to women, and a participant with a positive pupil dilation score responded more strongly to men.
- **Subjective Ratings (SR)**: Participants were also asked to give ratings of how attractive they found the people in the videos on a scale from 0-6. Again, scores for videos featuring women were deducted from scores for videos featuring men. Thus, a participant with a negative score preferred women, and a participant with a positive score preferred men.

- **Self-Reported Gender Nonconformity (SRG)**: Participants were asked how feminine they feel compared to other men of their same age, on a scale from 0-6. Lower scores indicated that participants felt less feminine, and higher scores indicated that they felt more feminine.
- **Observer-Rated Gender Nonconformity (ORG)**: Participants were briefly interviewed, and the video clips of these interviews were rated by neutral observers for how masculine or feminine the participant appeared in their speech, dress, and mannerisms on a scale from 0-6. Lower scores indicated that participants appeared more masculine, and higher scores indicated that they appeared more feminine.

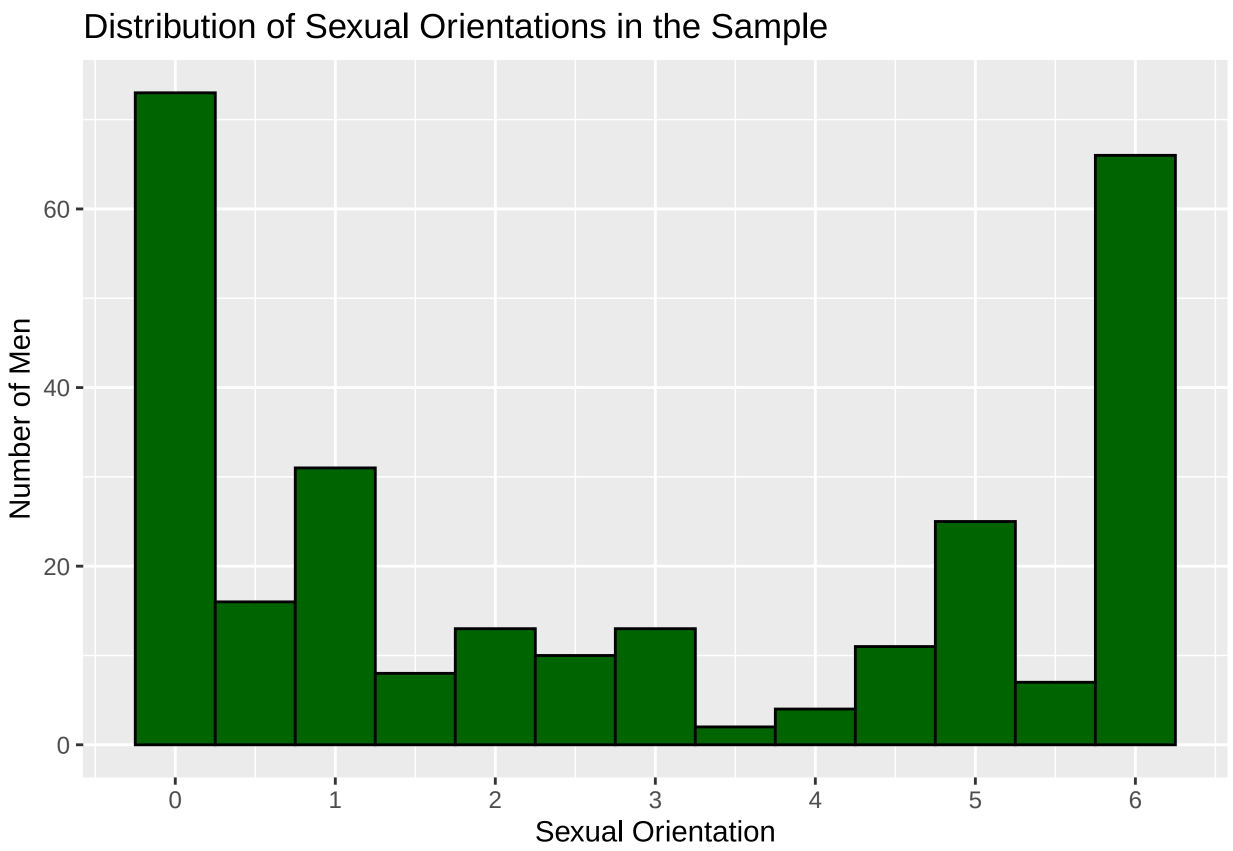
All of these variables can be theorised to have a relationship with sexual orientation - in general, research indicates that gay men are more feminine (on average) in their behaviours and mannerisms than straight men, and we can reasonably expect that they would respond more strongly to videos featuring men than videos featuring women. Thus, we should be able to predict a participant's sexual orientation from these variables with acceptable accuracy.

Data Visualisation

We will first verify that our dataset has the 279 participants that we expect it to have:

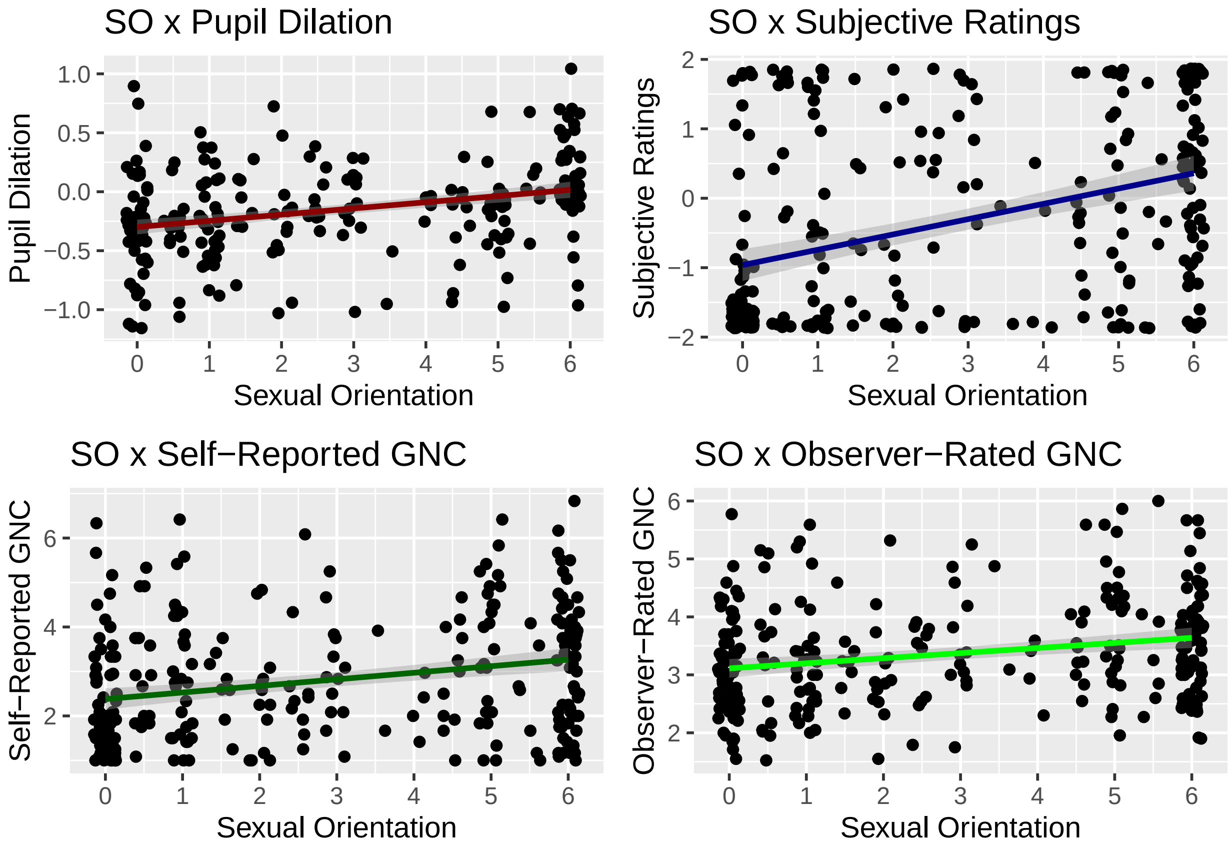
```
## Number of Participants: 279
```

We can now take a look at the distribution of their sexual orientations. As stated previously, since the sexual orientation variable is the average of two similar questions, we expect to see some participants who have orientations ending in a decimal:



From this, we can see that there are a significantly higher number of exclusively straight and exclusively gay participants. This is partly due to these specific groups being the focus of several of the research projects which resulted in this data. This will be less of a problem later, since we will group participants categorically into bins based on their sexual orientation, and bisexuals have a wider bin than either straight or gay men.

We will now examine the relationship between sexual orientation and the four predictive variables: Pupil Dilation, Subjective Ratings, Self-Reported Gender Nonconformity, and Observer-Rated Gender Nonconformity. In all cases, these graphs can be interpreted in the same way - sexual orientation is on the x axis, with straight men (0-1) on the left, bisexual men (2-4) in the middle, and gay men (5-6) on the right.

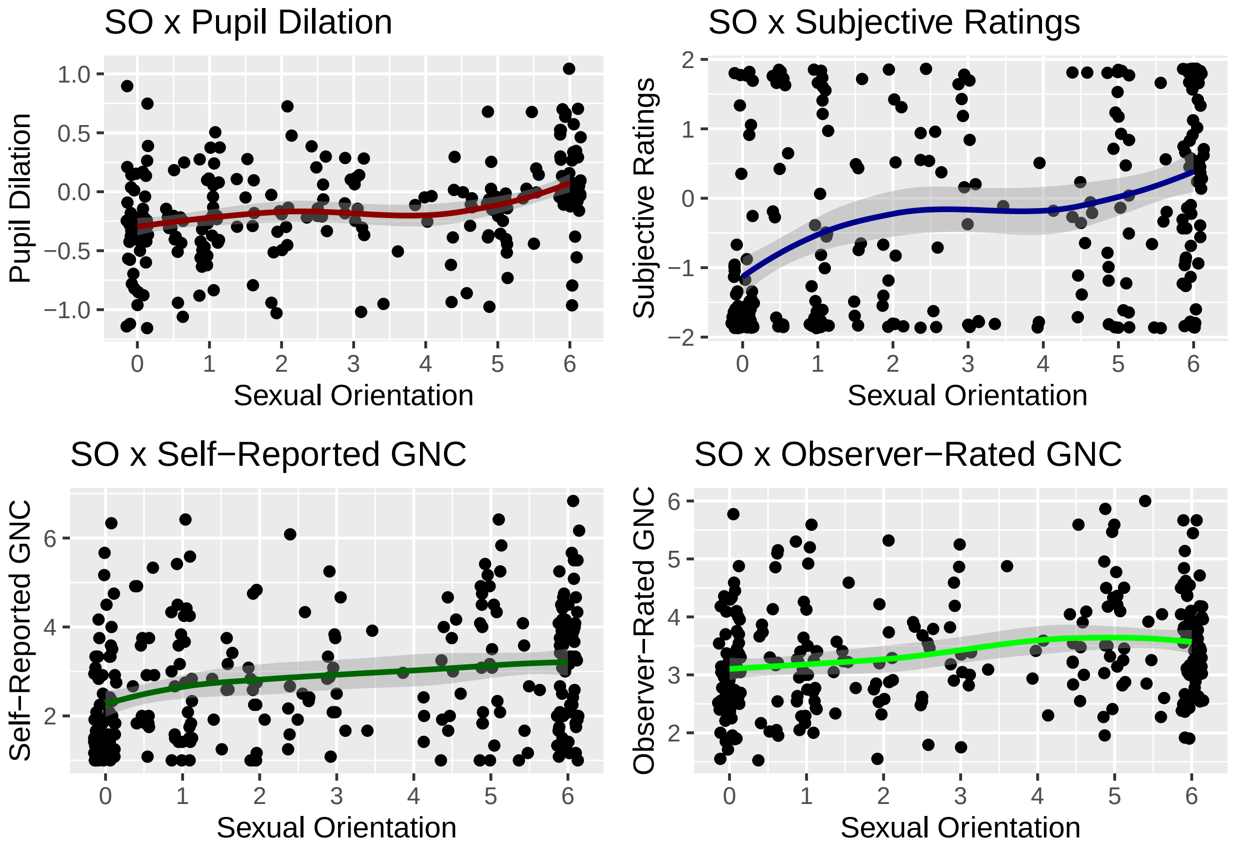

SO x Pupil Dilation

SO x Subjective Ratings

SO x Self−Reported GNC

SO x Observer−Rated GNC

As we can see here, all four of the variables have a positive relationship with sexual orientation as predicted. In other words, compared to straight men, gay men tend to respond more strongly to videos featuring men (both in pupil dilation and in their subjective ratings), and gay men tend to be more feminine than straight men, both in self-reports and observer ratings. Yet, these linear fit lines may not be the best way to depict the relationships between the variables - we will also try non-linear fits.



As we can see here, in some cases, the relationship between the predictive variables and sexual orientation appears to be better-represented using a non-linear model. These effects seem to be primarily driven by bisexual men, who are closer to straight men on some measures (e.g. pupil dilation) and closer to gay men on others (e.g. subjective ratings).

To be sure of the exact nature of the relationships between sexual orientation and the predictors, we will calculate the Pearson correlations (r) and corresponding significance values (p) for each pairwise comparison.

```
##      Pupil Dilation Subjective Ratings Self-Reported GNC Observer-Rated GNC
## 1 r           0.427              0.407             0.287              0.231
## 2 p           0.000              0.000             0.000              0.000
```

The above table gives us two interesting insights: Firstly, every predictive variable is correlated with sexual orientation. All four p-values were below .05, indicating that these are statistically significant relationships. Secondly, although the variables are all correlated (and in the same direction, positively) with sexual orientation, the magnitude of these correlations differs, and the two gender nonconformity variables have a weaker relationship with sexual orientation than pupil dilation or subjective ratings.

Insights and Model Building

Exploring and visualising the data has given us several valuable insights which we must take into account when building our models. In summary, they are:

1. There are more exclusively gay and exclusively straight men in our sample than bisexual men. This is (somewhat) acceptable for linear regression, but would be problematic for other kinds of continuous modelling. We will thus need to group participants by sexual orientation and use classification modelling techniques for our other models.

2. The relationship between sexual orientation and the other variables may be non-linear. Thus, it makes sense to explore options other than linear regression when trying to make predictions based on these variables.
3. All four predictive variables were significantly correlated with sexual orientation. However, their magnitudes differed - the relationship between sexual orientation, pupil dilation and subjective ratings was stronger than the relationship between sexual orientation and either of the gender nonconformity variables. This may mean that the latter two variables are less suitable for use in predicting sexual orientation.

Modelling

With the information we have gathered from examining the data set, we will now build a series of models, with the aim of predicting sexual orientation. The first set of 4 models will be simple Linear Regression models, and this will be followed by a set of 4 K-Nearest Neighbour models and 4 Classification Tree models, for a total of 12 models.

For all three sets of models, we will add predictors stage-by-stage in descending order of their correlation with sexual orientation; thus, in each set of models, the order will be:

1. Sexual Orientation predicted by **Pupil Dilation**
2. Sexual Orientation predicted by **Pupil Dilation and Subjective Ratings**
3. Sexual Orientation predicted by **Pupil Dilation, Subjective Ratings, and Self-Reported Gender Nonconformity**
4. Sexual Orientation predicted by **Pupil Dilation, Subjective Ratings, Self-Reported Gender Nonconformity and Observer-Rated Gender Nonconformity**

At the end of each section, we will add the relevant figures for measuring the effectiveness of the models (R-Squared in the case of Linear Regression and Accuracy in the case of KNN and CT) to a table for easy comparison.

The dataset must first be split into training and testing sets. The code for doing this is below. All models will first be trained on the training set, and then will be used to predict the sexual orientation of participants in the testing set. Owing to the size of the data set (which is very large for this kind of Psychology research, but relatively small by Machine Learning standards), we decided on a train/test split of 80/20, to avoid either set being too small to function properly.

```r
# sets seed
set.seed(1, sample.kind = "Rounding")
# split the data into training and test sets
test_index <- createDataPartition(y = sexlabdata$sexual_orientation, times = 1,
                                  p = 0.2, list = FALSE)
train_set <- sexlabdata[-test_index,]
test_set <- sexlabdata[test_index,]
# verify that the same id does not appear in both sets
train_set$id %in% test_set$id
```

Model Set 1: Linear Regression

We will first attempt to predict the sexual orientation of the test set using a series of Linear Regression models.

```
# variables will be added to the model in descending order of predictive power
# LR model 1: PD
# model is fitted using the training set
fit1 <- lm(sexual_orientation ~ pupil_dilation, data = train_set)
# predictions are generated on the test set
y_hat1 <- predict(fit1,newdata=test_set)
# summary of model is shown
summary(fit1)
```

```
##
## Call:
## lm(formula = sexual_orientation ~ pupil_dilation, data = train_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3803 -2.1436 -0.3396  2.1700  5.1957
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.1753     0.1683  18.872  < 2e-16 ***
## pupil_dilation   2.4616     0.4072   6.045 6.37e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.271 on 219 degrees of freedom
## Multiple R-squared:  0.143,  Adjusted R-squared:  0.1391
## F-statistic: 36.54 on 1 and 219 DF,  p-value: 6.366e-09
```

```
# LR model 2: PD/SR
# model is fitted using the training set
fit2 <- lm(sexual_orientation ~ pupil_dilation + subjective_ratings, data = train_set)
# predictions are generated on the test set
y_hat2 <- predict(fit2,newdata=test_set)
# summary of model is shown
summary(fit2)
```

```
##
## Call:
## lm(formula = sexual_orientation ~ pupil_dilation + subjective_ratings,
##     data = train_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8418 -1.8049 -0.2946  1.9521  5.4668
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.3346     0.1611  20.693  < 2e-16 ***
## pupil_dilation       2.2733     0.3850   5.905 1.34e-08 ***
## subjective_ratings   0.5531     0.1026   5.392 1.81e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 2.138 on 218 degrees of freedom
## Multiple R-squared:  0.2438, Adjusted R-squared:  0.2369
## F-statistic: 35.15 on 2 and 218 DF,  p-value: 5.876e-14
```

```
# LR model 3: PD/SR/SRG
# model is fitted using the training set
fit3 <- lm(sexual_orientation ~ pupil_dilation + subjective_ratings +
             self_reported_gnc, data = train_set)
# predictions are generated on the test set
y_hat3 <- predict(fit3,newdata=test_set)
# summary of model is shown
summary(fit3)
```

```
##
## Call:
## lm(formula = sexual_orientation ~ pupil_dilation + subjective_ratings +
##     self_reported_gnc, data = train_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8503 -1.7611 -0.3288  1.9357  5.5424
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.12555    0.41342   7.560 1.12e-12 ***
## pupil_dilation       2.30450    0.38975   5.913 1.29e-08 ***
## subjective_ratings   0.51179    0.12737   4.018 8.09e-05 ***
## self_reported_gnc    0.07198    0.13108   0.549    0.583
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.141 on 217 degrees of freedom
## Multiple R-squared:  0.2449, Adjusted R-squared:  0.2344
## F-statistic: 23.46 on 3 and 217 DF,  p-value: 3.439e-13
```

```
# LR model 4: PD/SR/SRG/ORG
# model is fitted using the training set
fit4 <- lm(sexual_orientation ~ pupil_dilation + subjective_ratings +
             self_reported_gnc + observer_rated_gnc, data = train_set)
# predictions are generated on the test set
y_hat4 <- predict(fit4,newdata=test_set)
# summary of model is shown
summary(fit4)
```

```
##
## Call:
## lm(formula = sexual_orientation ~ pupil_dilation + subjective_ratings +
##     self_reported_gnc + observer_rated_gnc, data = train_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6682 -1.7793 -0.4235  1.9147  5.5583
```

```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.5043     0.6766   5.180 5.09e-07 ***
## pupil_dilation        2.3260     0.3914   5.943 1.11e-08 ***
## subjective_ratings    0.5325     0.1308   4.070 6.59e-05 ***
## self_reported_gnc     0.1403     0.1629   0.861     0.39
## observer_rated_gnc   -0.1680     0.2374  -0.708     0.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.144 on 216 degrees of freedom
## Multiple R-squared:  0.2466, Adjusted R-squared:  0.2327
## F-statistic: 17.68 on 4 and 216 DF,  p-value: 1.439e-12
```

Our results table so far looks like this:

```
##           Model LR (R^2) KNN (Acc) Class (Acc)
## 1            PD   0.1391
## 2         PD/SR   0.2369
## 3     PD/SR/SRG   0.2344
## 4 PD/SR/SRG/ORG   0.2327
```

As we can see, the Adjusted R-Squared value - a measure of how much of the total variance in sexual orientation is explained by the model - improves significantly in model 2, when subjective ratings are added as a predictor. However, following that, the addition of either of the gender nonconformity variables does not improve the predictive power of the model further; rather, it causes it to decrease slightly.

This is possibly due to colinearity between the variables, or because their predictive power over sexual orientation is weaker than the first two variables to begin with. The results of the models themselves back this up - in model 4, we can clearly see that although PD and SR are significantly related to sexual orientation, SRG and ORG are not. In other words, the variance in sexual orientation that they account for is better explained by PD and SR, and the relationship between SRG/ORG and sexual orientation vanishes when PD and SR are controlled for in a model like this.

Model Set 2: K-Nearest Neighbours

As mentioned previously, the distribution of sexual orientations in our sample is not even - there are more exclusively gay and straight men in the sample than bisexual men. This is largely due to the priorities of the research projects which make up this dataset. This presents a problem for models which treat sexual orientation as a continuous variable.

As such, we will take a different approach - our second and third set of models will be based on classification, and we will convert sexual orientation into a factor to accommodate this. Specifically, we will categorise participants from 0.0 to 1.0 as "straight", 1.5 to 4.5 as "bi", and 5.0 to 6.0 as "gay". We must also split this new dataset into train and test sets.

```
# create categorical sexual orientation variable through binning
knndata <- sexlabdata %>%
  mutate(so_category=ifelse(sexual_orientation %in% c(0.0, 0.5, 1.0), "straight",
                     ifelse(sexual_orientation %in% c(1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5),
                                                      "bi", "gay")))
# turn this new categorical variable into a factor
```

```
knndata$so_category <- as.factor(knndata$so_category)
# drop the old sexual orientation variable to prevent it being used for predictions
knndata <- select(knndata,-sexual_orientation)
# drop id variable for the same reason
knndata <- select(knndata,-id)
# sets seed
set.seed(1, sample.kind = "Rounding")
# split the new dataset into training and test sets
test_index <- createDataPartition(y = knndata$so_category, times = 1,
                                  p = 0.2, list = FALSE)
knn_train_set <- knndata[-test_index,]
knn_test_set <- knndata[test_index,]
# sets parameters for cross-validation of knn models
trctrl <- trainControl(method = "repeatedcv", number = 20, repeats = 5)
```

The slightly larger bin size given to bisexual participants helps correct the problem of uneven sample distribution, although it is still not balanced:

```
##       bi      gay straight
##       61       98      120
```

Although not ideal, this is adequate for training classification models. We will again compute four models - this time K-nearest Neighbour models - and each time we will add a new predictor and measure the effectiveness of the model. Since this is a classification model, we will now be measuring the effectiveness of its predictions using Accuracy. At each stage, the optimal K will be found through cross-validation with 20 folds repeated 5 times. All variables were centered and scaled (standardised) before predictions were made.

```
# KNN model 1: PD
# knn model is fitted
knn_fit1 <- train(so_category ~pupil_dilation,
                  data = knn_train_set, method = "knn", trControl=trctrl,
                  preProcess = c("center", "scale"), tuneLength = 10)
# shows accuracy of knn model in cross-validation
knn_fit1
```

```
## k-Nearest Neighbors
##
## 222 samples
##   1 predictor
##   3 classes: ’bi’, ’gay’, ’straight’
##
## Pre-processing: centered (1), scaled (1)
## Resampling: Cross-Validated (20 fold, repeated 5 times)
## Summary of sample sizes: 211, 211, 212, 210, 211, 212, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    5  0.5503081  0.2866888
##    7  0.5741919  0.3203654
##    9  0.5801919  0.3288764
##   11  0.5771465  0.3195170
##   13  0.5865152  0.3211736
##    15  0.5996061  0.3381318
##    17  0.6102273  0.3532944
##    19  0.6113939  0.3549640
##    21  0.6086515  0.3502433
##    23  0.6114697  0.3557006
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 23.
```

```
# use trained model to predict test set
test_pred1 <- predict(knn_fit1, newdata = knn_test_set)
# confusion matrix showing the results
knncf1 <- confusionMatrix(test_pred1, knn_test_set$so_category)

# KNN model 2: PD/SR
# knn model is fitted
knn_fit2 <- train(so_category ~pupil_dilation + subjective_ratings,
                 data = knn_train_set, method = "knn", trControl=trctrl,
                 preProcess = c("center", "scale"), tuneLength = 10)
# shows accuracy of knn model in cross-validation
knn_fit2
```

```
## k-Nearest Neighbors
##
## 222 samples
##   2 predictor
##   3 classes: 'bi', 'gay', 'straight'
##
## Pre-processing: centered (2), scaled (2)
## Resampling: Cross-Validated (20 fold, repeated 5 times)
## Summary of sample sizes: 212, 211, 211, 210, 211, 210, ...
## Resampling results across tuning parameters:
##
##    k   Accuracy   Kappa
##     5  0.5925354  0.3494402
##     7  0.5826263  0.3287452
##     9  0.5928283  0.3393733
##    11  0.6127727  0.3683130
##    13  0.6160455  0.3685676
##    15  0.6128182  0.3650879
##    17  0.6146212  0.3690276
##    19  0.6257273  0.3861567
##    21  0.6115455  0.3639607
##    23  0.6061667  0.3517629
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 19.
```

```
# use trained model to predict test set
test_pred2 <- predict(knn_fit2, newdata = knn_test_set)
# confusion matrix showing the results
knncf2 <- confusionMatrix(test_pred2, knn_test_set$so_category)
```

```
# KNN model 3: PD/SR/SRG
# knn model is fitted
knn_fit3 <- train(so_category ~pupil_dilation + subjective_ratings + self_reported_gnc,
                  data = knn_train_set, method = "knn", trControl=trctrl,
                  preProcess = c("center", "scale"), tuneLength = 10)
# shows accuracy of knn model in cross-validation
knn_fit3
```

```
## k-Nearest Neighbors
##
## 222 samples
##   3 predictor
##   3 classes: ’bi’, ’gay’, ’straight’
##
## Pre-processing: centered (3), scaled (3)
## Resampling: Cross-Validated (20 fold, repeated 5 times)
## Summary of sample sizes: 211, 211, 211, 212, 211, 211, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    5  0.5840253  0.3305381
##    7  0.5629646  0.2962688
##    9  0.5846818  0.3298334
##   11  0.5738182  0.3070262
##   13  0.5698939  0.2987091
##   15  0.5619394  0.2857351
##   17  0.5527424  0.2711740
##   19  0.5527727  0.2703950
##   21  0.5580909  0.2765989
##   23  0.5630758  0.2835668
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 9.
```

```
# use trained model to predict test set
test_pred3 <- predict(knn_fit3, newdata = knn_test_set)
# confusion matrix showing the results
knncf3 <- confusionMatrix(test_pred3, knn_test_set$so_category)

# knn model 4: PD/SR/SRG/ORG
# knn model is fitted
knn_fit4 <- train(so_category ~., data = knn_train_set, method = "knn",
                  trControl=trctrl,
                  preProcess = c("center", "scale"),
                  tuneLength = 10)
# shows accuracy of knn model in cross-validation
knn_fit4
```

```
## k-Nearest Neighbors
##
## 222 samples
##   4 predictor
##   3 classes: ’bi’, ’gay’, ’straight’
##
## Pre-processing: centered (4), scaled (4)
## Resampling: Cross-Validated (20 fold, repeated 5 times)
## Summary of sample sizes: 211, 212, 210, 211, 211, 210, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    5  0.5732323  0.3181956
##    7  0.5485960  0.2742224
##    9  0.5535202  0.2794071
##   11  0.5537929  0.2774075
##   13  0.5552626  0.2809798
##   15  0.5493232  0.2692709
##   17  0.5548687  0.2764590
##   19  0.5607475  0.2841818
##   21  0.5651566  0.2939243
##   23  0.5596717  0.2843418
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 5.
```

```
# use trained model to predict test set
test_pred4 <- predict(knn_fit4, newdata = knn_test_set)
# confusion matrix showing the results
knncf4 <- confusionMatrix(test_pred4, knn_test_set$so_category)
```

The results of testing these models on the testing set are shown below:

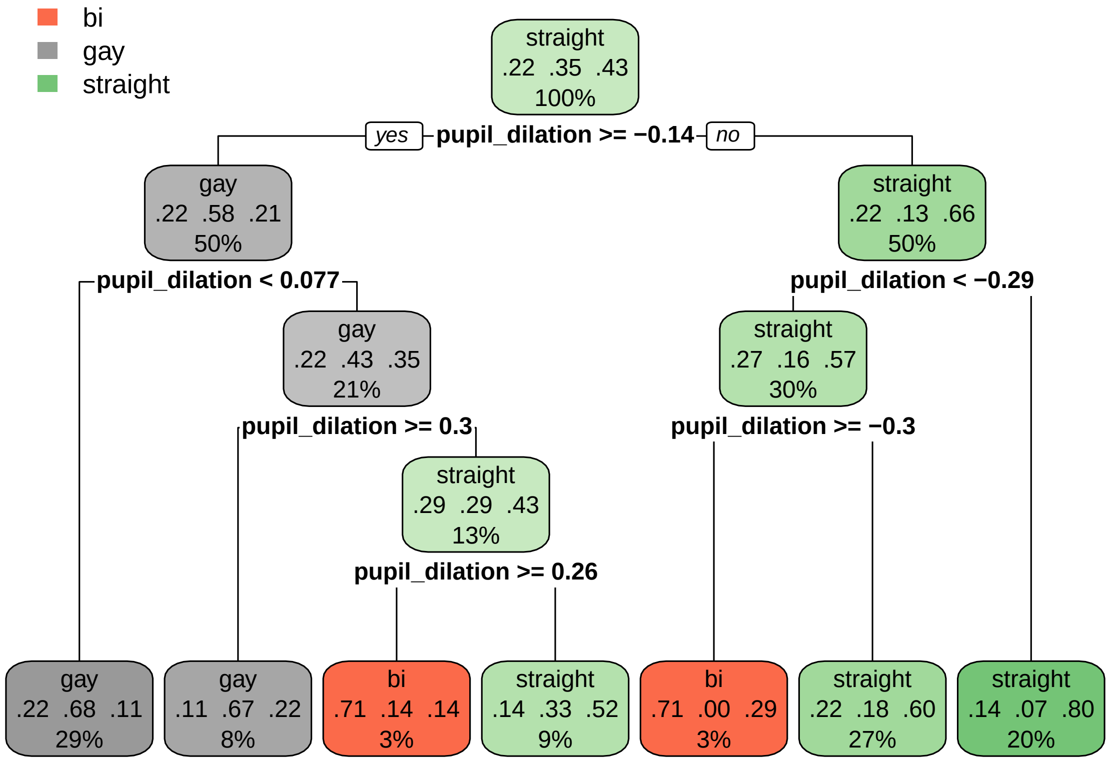
```
##           Model LR (R^2) KNN (Acc) Class (Acc)
## 1            PD   0.1391 0.6140351
## 2         PD/SR   0.2369 0.6666667
## 3     PD/SR/SRG   0.2344 0.6140351
## 4 PD/SR/SRG/ORG   0.2327 0.5614035
```

Here we see a broadly similar pattern to the one we found using Linear Regression: It is not the case that simply adding more variables makes the model more accurate. Instead, the most accurate model is #2 (PD & SR), and adding the gender nonconformity variables actually reduces the overall accuracy of the model.

Model Set 3: Classification Trees

Finally, we will repeat the same set of four models, but this time using a Classification Tree method. Again, sexual orientation will be treated as a factor rather than a continuous variable. The models will be trained on the training set, and we will use their Accuracy in predicting the sexual orietation of the test set as a measure of the success of the model.

```
# Classification model 1: PD
# classification model is fitted and plotted
rpartfit1 <- rpart(so_category~pupil_dilation, data = knn_train_set, method = 'class')
rpart.plot(rpartfit1)
```

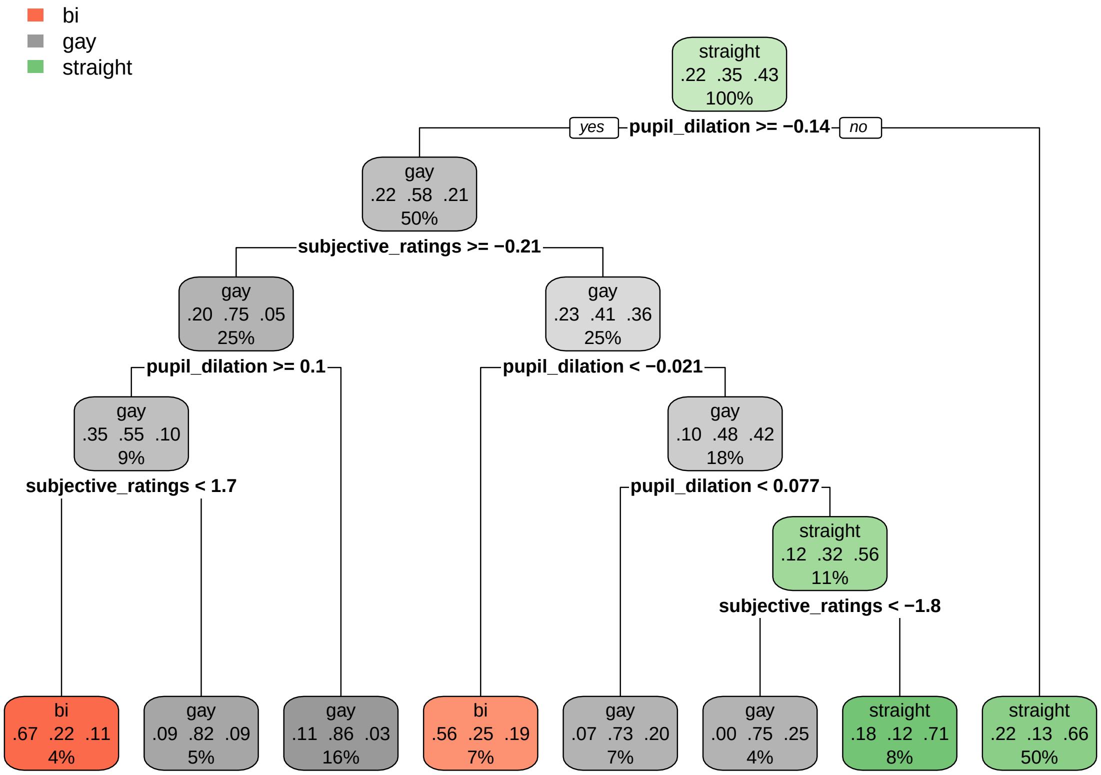


```
# fitted model used to predict test set
classpredict1 <- predict(rpartfit1,newdata=knn_test_set, type ="class")
# confusion matrix created and printed
classcf1 <- confusionMatrix(classpredict1, knn_test_set$so_category)
classcf1
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction bi gay straight
##   bi        0   3        0
##   gay       4  14        2
##   straight  9   3       22
##
## Overall Statistics
##
##                Accuracy : 0.6316
##                  95% CI : (0.4934, 0.7555)
##     No Information Rate : 0.4211
##     P-Value [Acc > NIR] : 0.001093
##
##                   Kappa : 0.3997
##
##  Mcnemar's Test P-Value : 0.025063
##
## Statistics by Class:
```

```
##
##                      Class: bi Class: gay Class: straight
## Sensitivity            0.00000     0.7000          0.9167
## Specificity            0.93182     0.8378          0.6364
## Pos Pred Value         0.00000     0.7000          0.6471
## Neg Pred Value         0.75926     0.8378          0.9130
## Prevalence             0.22807     0.3509          0.4211
## Detection Rate         0.00000     0.2456          0.3860
## Detection Prevalence   0.05263     0.3509          0.5965
## Balanced Accuracy      0.46591     0.7689          0.7765
```

```
# Classification model 2: PD/SR
# classification model is fitted and plotted
rpartfit2 <- rpart(so_category~pupil_dilation+subjective_ratings,
                   data = knn_train_set, method = 'class')
rpart.plot(rpartfit2)
```

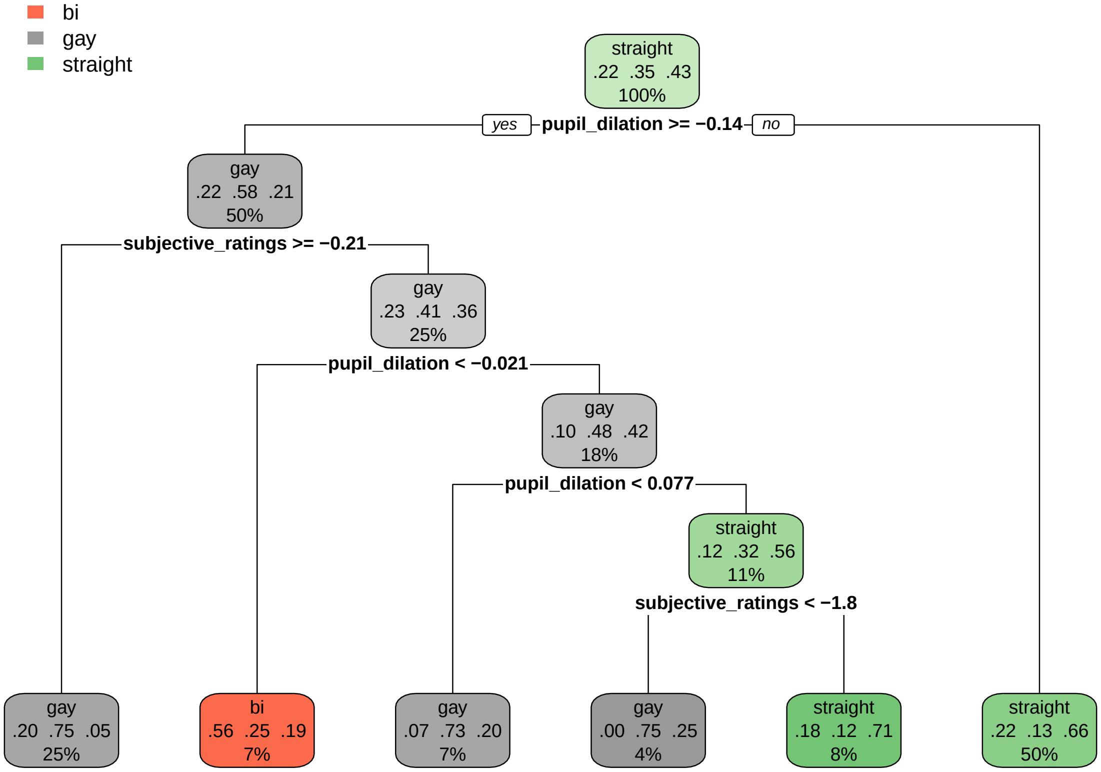


```
# fitted model used to predict test set
classpredict2 <- predict(rpartfit2,newdata=knn_test_set, type ="class")
# confusion matrix created and printed
classcf2 <- confusionMatrix(classpredict2, knn_test_set$so_category)
classcf2
```

```
## Confusion Matrix and Statistics
##
```

```
##           Reference
## Prediction bi gay straight
##   bi        1   3        0
##   gay       5  12        2
##   straight  7   5       22
##
## Overall Statistics
##
##                Accuracy : 0.614
##                  95% CI : (0.4757, 0.74)
##     No Information Rate : 0.4211
##     P-Value [Acc > NIR] : 0.002581
##
##                   Kappa : 0.3733
##
##  Mcnemar's Test P-Value : 0.032280
##
## Statistics by Class:
##
##                      Class: bi Class: gay Class: straight
## Sensitivity            0.07692     0.6000          0.9167
## Specificity            0.93182     0.8108          0.6364
## Pos Pred Value         0.25000     0.6316          0.6471
## Neg Pred Value         0.77358     0.7895          0.9130
## Prevalence             0.22807     0.3509          0.4211
## Detection Rate         0.01754     0.2105          0.3860
## Detection Prevalence   0.07018     0.3333          0.5965
## Balanced Accuracy      0.50437     0.7054          0.7765
```

```
# Classification model 3: PD/SR/SRG
# classification model is fitted
rpartfit3 <- rpart(so_category~pupil_dilation+subjective_ratings+
                     self_reported_gnc, data = knn_train_set, method = 'class')
rpart.plot(rpartfit3)
```

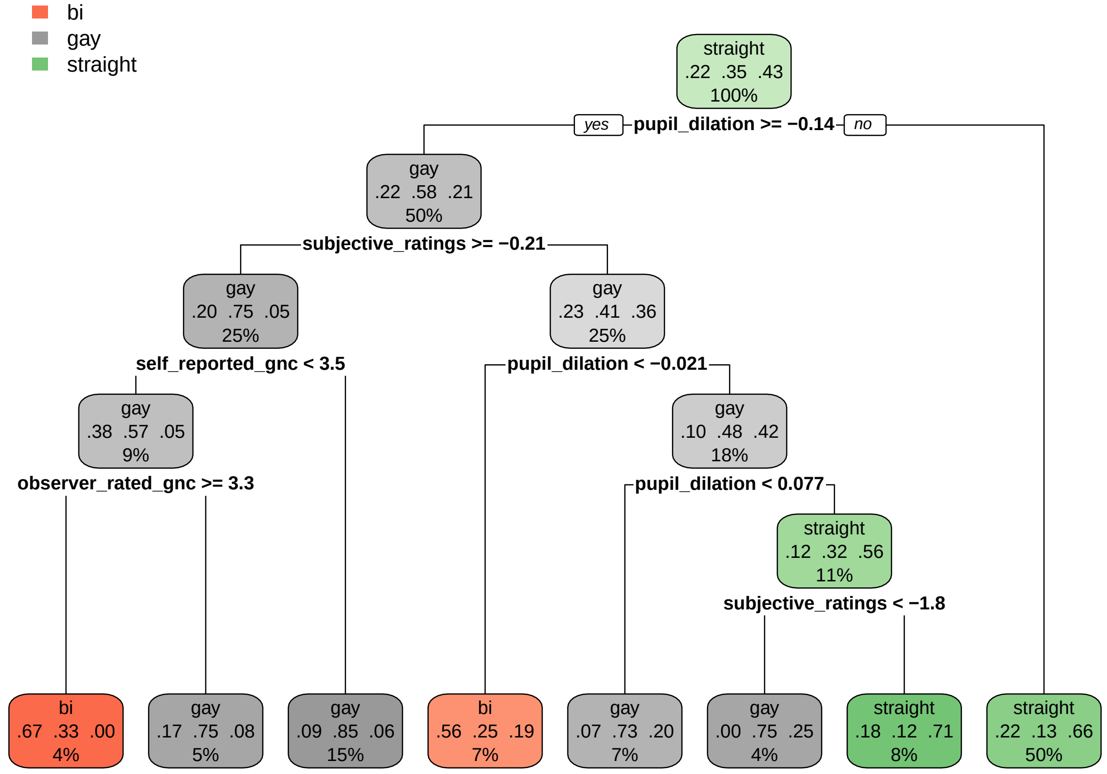


```
# fitted model used to predict test set
classpredict3 <- predict(rpartfit3,newdata=knn_test_set, type ="class")
# confusion matrix created and printed
classcf3 <- confusionMatrix(classpredict3, knn_test_set$so_category)
classcf3
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction bi gay straight
##   bi        0   1        0
##   gay       6  14        2
##   straight  7   5       22
##
## Overall Statistics
##
##                Accuracy : 0.6316
##                  95% CI : (0.4934, 0.7555)
##     No Information Rate : 0.4211
##     P-Value [Acc > NIR] : 0.001093
##
##                   Kappa : 0.3955
##
##  Mcnemar's Test P-Value : 0.007889
##
## Statistics by Class:
```

```
##
##                      Class: bi Class: gay Class: straight
## Sensitivity            0.00000     0.7000          0.9167
## Specificity            0.97727     0.7838          0.6364
## Pos Pred Value         0.00000     0.6364          0.6471
## Neg Pred Value         0.76786     0.8286          0.9130
## Prevalence             0.22807     0.3509          0.4211
## Detection Rate         0.00000     0.2456          0.3860
## Detection Prevalence   0.01754     0.3860          0.5965
## Balanced Accuracy      0.48864     0.7419          0.7765
```

```
# Classification model 4: PD/SR/SRG/ORG
# classification model is fitted
rpartfit4 <- rpart(so_category~., data = knn_train_set, method = 'class')
rpart.plot(rpartfit4)
```



```
# fitted model used to predict test set
classpredict4 <- predict(rpartfit4,newdata=knn_test_set, type ="class")
classcf4 <- confusionMatrix(classpredict4, knn_test_set$so_category)
# confusion matrix created and printed
classcf4
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction bi gay straight
##   bi        1   5        0
##   gay       5  10        2
##   straight  7   5       22
##
## Overall Statistics
##
##                Accuracy : 0.5789
##                  95% CI : (0.4408, 0.7086)
##     No Information Rate : 0.4211
##     P-Value [Acc > NIR] : 0.01176
##
##                   Kappa : 0.3211
##
##  Mcnemar's Test P-Value : 0.04046
##
## Statistics by Class:
##
##                      Class: bi Class: gay Class: straight
## Sentivity              0.07692     0.5000          0.9167
## Specificity            0.88636     0.8108          0.6364
## Pos Pred Value         0.16667     0.5882          0.6471
## Neg Pred Value         0.76471     0.7500          0.9130
## Prevalence             0.22807     0.3509          0.4211
## Detection Rate         0.01754     0.1754          0.3860
## Detection Prevalence   0.10526     0.2982          0.5965
## Balanced Accuracy      0.48164     0.6554          0.7765
```

The results of testing these models on the testing set are shown below

```
##             Model LR (R^2) KNN (Acc) Class (Acc)
## 1              PD   0.1391 0.6140351   0.6315789
## 2           PD/SR   0.2369 0.6666667   0.6140351
## 3       PD/SR/SRG   0.2344 0.6140351   0.5789474
## 4 PD/SR/SRG/ORG   0.2327 0.5614035   0.5789474
```

We see a similar pattern as before: Adding more predictors does not necessarily increase the accuracy of the model. Rather, the most accurate model is the first one, with only pupil dilation as a predictor, and the accuracy decreases each time a new predictor is added.

Conclusions

In summary, we found that adding more variables does not necessarily improve the fit of a model. This applied to both the Linear Regression models and also the factor-based classification models. Broadly, the models followed the same pattern: Pupil dilation, which had the strongest correlation with sexual orientation, was always a very strong predictor, and subjective ratings, with the second-strongest correlation, usually improved the model by being added. However, the two gender nonconformity variables tended to reduce the predictive power of a model when taken into consideration. It is possible that this is because they are correlated with one another, or correlate with the other predictors, and thus are unable to explain any additional variance when added to a model in stages 3 and 4. Another potential explanation is that their relationship with sexual orientation, although statistically significant, was simply weaker than the relationship between sexual orientation and the other two predictors, and thus they were less predictive.

```
## [1] "Final Results Table:"
```

```
##           Model LR (R^2) KNN (Acc) Class (Acc)
## 1            PD   0.1391 0.6140351   0.6315789
## 2         PD/SR   0.2369 0.6666667   0.6140351
## 3     PD/SR/SRG   0.2344 0.6140351   0.5789474
## 4 PD/SR/SRG/ORG   0.2327 0.5614035   0.5789474
```

One clear limitation of this project is the sample size. Gathering data such as this is expensive and labour-intensive, and although the sample is large by the standards of research done in this area, it is still small by the standards of most Machine Learning models. Although our best efforts were made to mitigate this factor - for example, through the use of carefully-selected train/test splits and binning, we cannot rule out that issues of sample size or distribution may have affected the findings.

Future research could approach this problem with an even wider range of different predictive models. If a more evenly-distributed sample was recruited, these could include continuous Machine Learning models, and this may shed more light on the issue of why certain predictors did not seem to benefit the models. Additionally, it would be worth focusing not just on Accuracy, but also on Sensitivity and Specificity. Exploring these parameters in addition to the ones already present was outside of the scope of this paper, but it is possible that a clearer picture of what is going on in these models may be gleaned from examining these alongside Accuracy.