Lab 4: KNN Classification, Cross-validation, Decision boundary plots

Amelia and Remy

Oct 7

Example: Forensic Glass Classification

This data set is provided as part of the MASS package for R, and the description below borrows from the data documentation provided in that package.

We have measurements of different attributes of fragments of glass, as well as the type of glass. If we could develop a reliable mechanism for classifying glass, glass shards found at crime scenes could potentially be used as evidence in criminal trials. There are 7 glass types in the originally published data, but only 6 occur in this data set: "window float glass (WinF: 70), window non-float glass (WinNF: 76), vehicle window glass (Veh: 17), containers (Con: 13), tableware (Tabl: 9) and vehicle headlamps (Head: 29)".

For each glass fragment, we have measurements of the refractive index of the glass, as well as the percentage by weight of 8 different elements in the glass. These are our explanatory/predictive variables.

Read in data, train/test split, validation splits

I've written this code for you. No need to update.

```
library(readr)
library(dplyr)
library(ggplot2)
library(GGally)
library(gridExtra)
library(caret)

set.seed(9215)

# Read in data, tqke a first look
glass <- read_csv("http://www.evanlray.com/data/mass/fgl.csv") %>%
  mutate(type = factor(type))

head(glass)
```

```
## # A tibble: 6 x 10
##       RI    Na    Mg    Al    Si     K    Ca    Ba    Fe type
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct>
## 1  3.01   13.6  4.49  1.1   71.8 0.06   8.75     0  0    WinF
## 2 -0.39   13.9  3.6   1.36  72.7 0.48   7.83     0  0    WinF
## 3 -1.82   13.5  3.55  1.54  73.0 0.39   7.78     0  0    WinF
## 4 -0.340  13.2  3.69  1.29  72.6 0.570  8.22     0  0    WinF
## 5 -0.580  13.3  3.62  1.24  73.1 0.55   8.07     0  0    WinF
## 6 -2.04   12.8  3.61  1.62  73.0 0.64   8.07     0  0.26 WinF
```

```
dim(glass)
```

```
## [1] 214  10
```

```
# Train/test split
tt_inds <- caret::createDataPartition(glass$type, p = 0.7)

train_glass <- glass %>% slice(tt_inds[[1]])
test_glass <- glass %>% slice(-tt_inds[[1]])
```

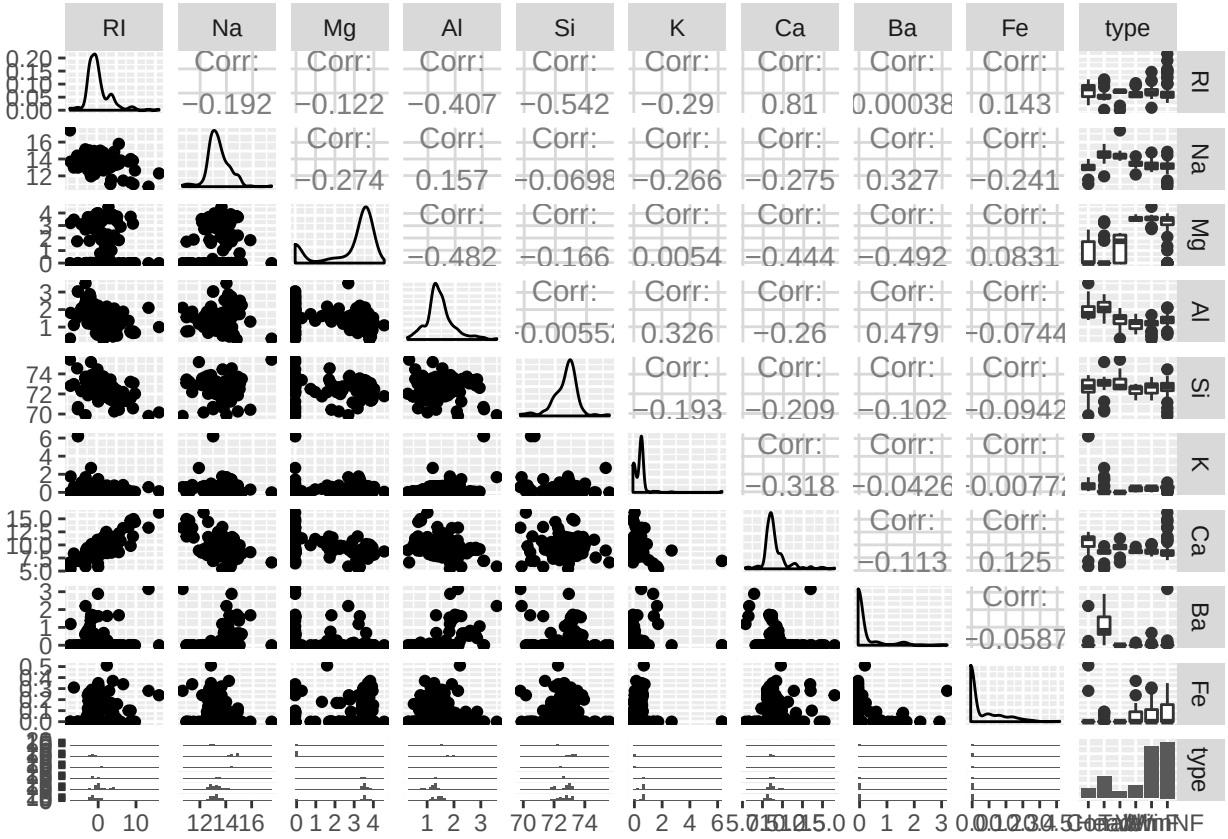
```
# Cross-validation splits
crossval_val_fold_inds <- caret::createFolds(
  y = train_glass$type, # response variable as a vector
  k = 10 # number of folds for cross-validation
)
```

1. Make some exploratory plots of the training data.

Your goal is to understand the data well enough that you can identify two features (two of the measured attributes) that would be useful for classifying glass in part 2. Ideally, each feature you choose will individually have something to say about glass type (for example, maybe the distribution of values for the feature is different for the different glass types), and there will not be a strong relationship between the two features you choose (so they carry different information about glass type and you don't just have basically the same information from your two features). Don't spend a huge amount of time on this though. A pairs plot could be good enough.

```
library(GGally)
ggpairs(glass)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



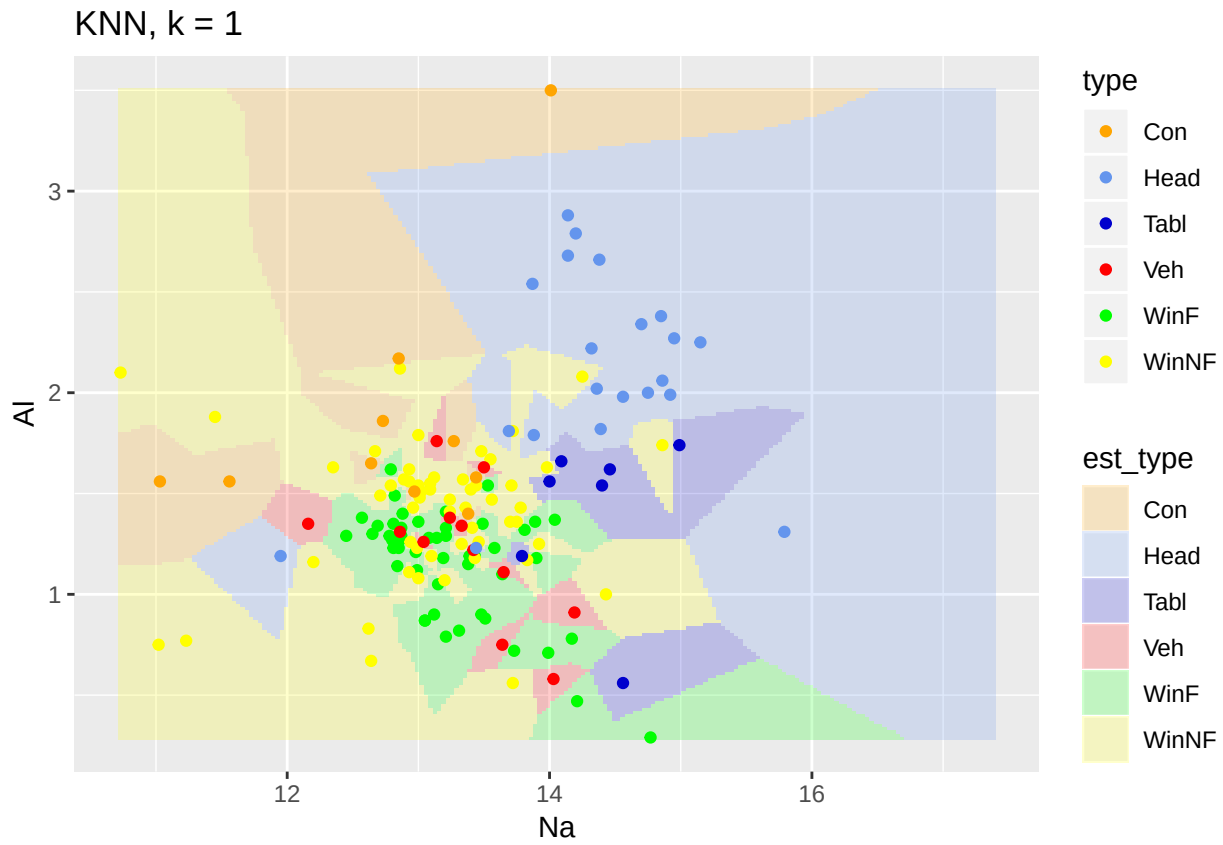
Based on the plots, Na and Al are features whose distribution of values is different for different glass types and there is not a strong relationship between Na and Al.

I will use Na and Al.

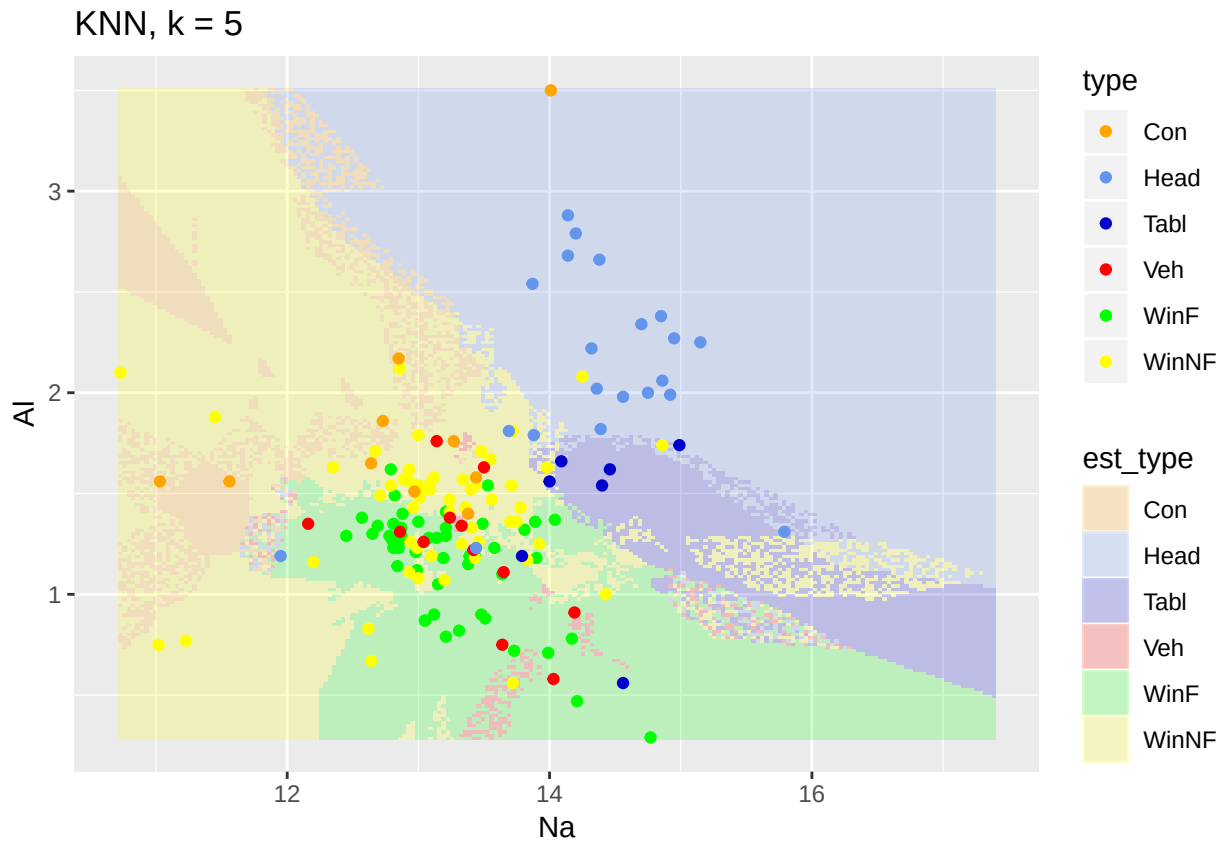
2. Using your two selected features, fit a KNN model and make a plot showing the decision boundaries in the two-dimensional feature space. Optional, if time: make a few different versions of the plot with different values of k to see how the decision boundaries change with the number of neighbors used.

To save you some time retyping all my code from the handout, I'm pasting in the code I used for the party affiliation example below. You will need to make the appropriate changes to tailor the code to the data set you're working with in this lab.

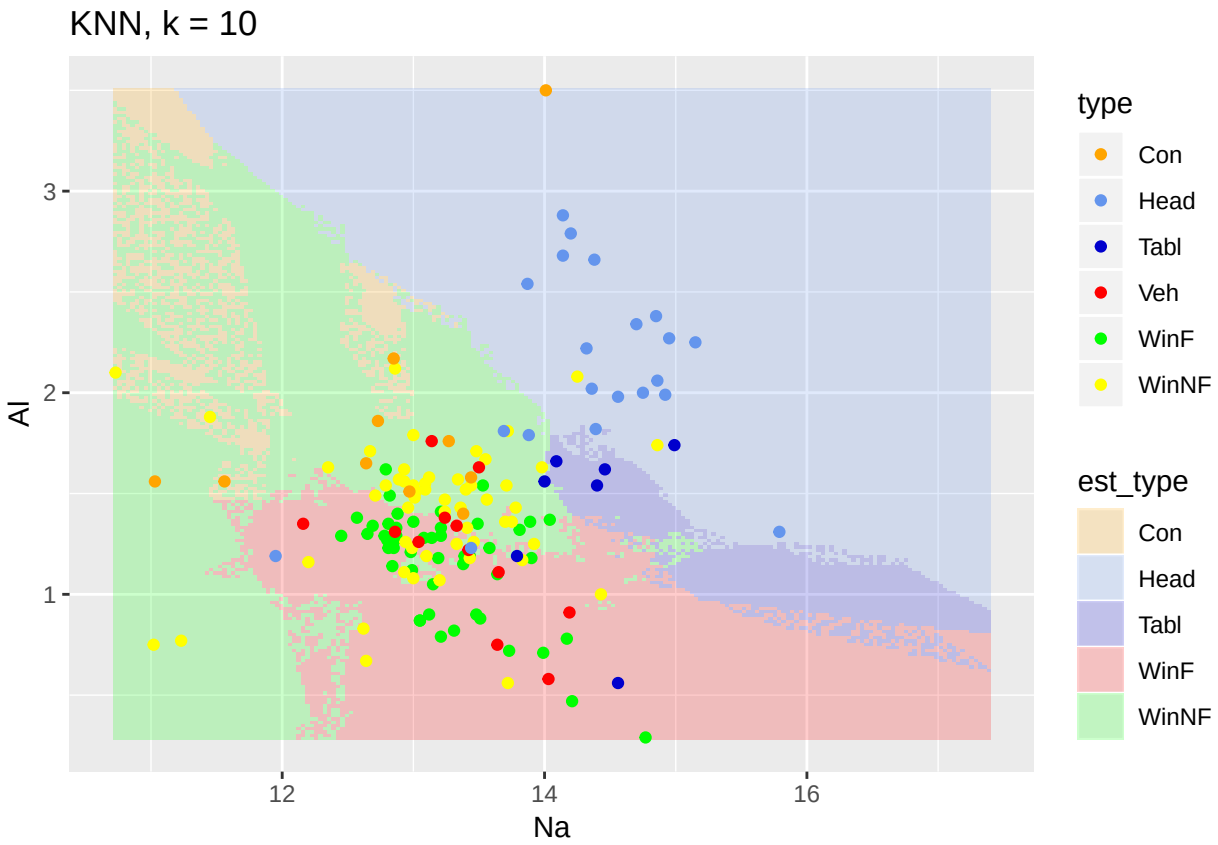
```
k <- c(1, 5, 10, 15, 50, 100, 150)
background <- vector("list", length(k))

for (i in 1:length(k)) {
  # "train" the KNN model
  knn_fit <- caret::train(
    form = type ~ Na + Al,
    data = train_glass,
    method = "knn",
    preProcess = "scale",
    trControl = trainControl(method = "none"),
    tuneGrid = data.frame(k = k[i])
  )

  # a grid of values for age and popul at which to get the estimated class.
  # it's not a test data set in the sense that we don't have observations of party to go with these points,
  # but we will treat it as a "test set" in the sense that we will obtain predictions at these points
  test_grid <- expand.grid(
    Na = seq(from = 10.73, to = 17.38, length = 201),
    Al = seq(from = 0.29, to = 3.50, length = 201)
  )
  head(test_grid)

  # use predict to find the estimated most likely class at each point in our grid
  y_hats <- predict(knn_fit, newdata = test_grid, type = "raw")

  # add the estimated types into the test_grid data frame
  background_knn <- test_grid %>%
    mutate(
      est_type = y_hats
    )
  background[[i]] <- background_knn
}

# make the plot.  geom_raster does the shading in the background, alpha = 0.2 makes it transparent
title <- paste0("KNN, k = ", 1)
ggplot() +
  geom_raster(data = background[[1]],
    mapping = aes(x = Na, y = Al, fill = est_type), alpha = 0.2) +
  geom_point(data = train_glass, mapping = aes(x = Na, y = Al, color = type)) +
  scale_color_manual("type", values = c("orange", "cornflowerblue", "mediumblue", "red", "green", "yellow")) +
  scale_fill_manual(values = c("orange", "cornflowerblue", "mediumblue", "red", "green", "yellow")) +
  ggtitle(title)
```



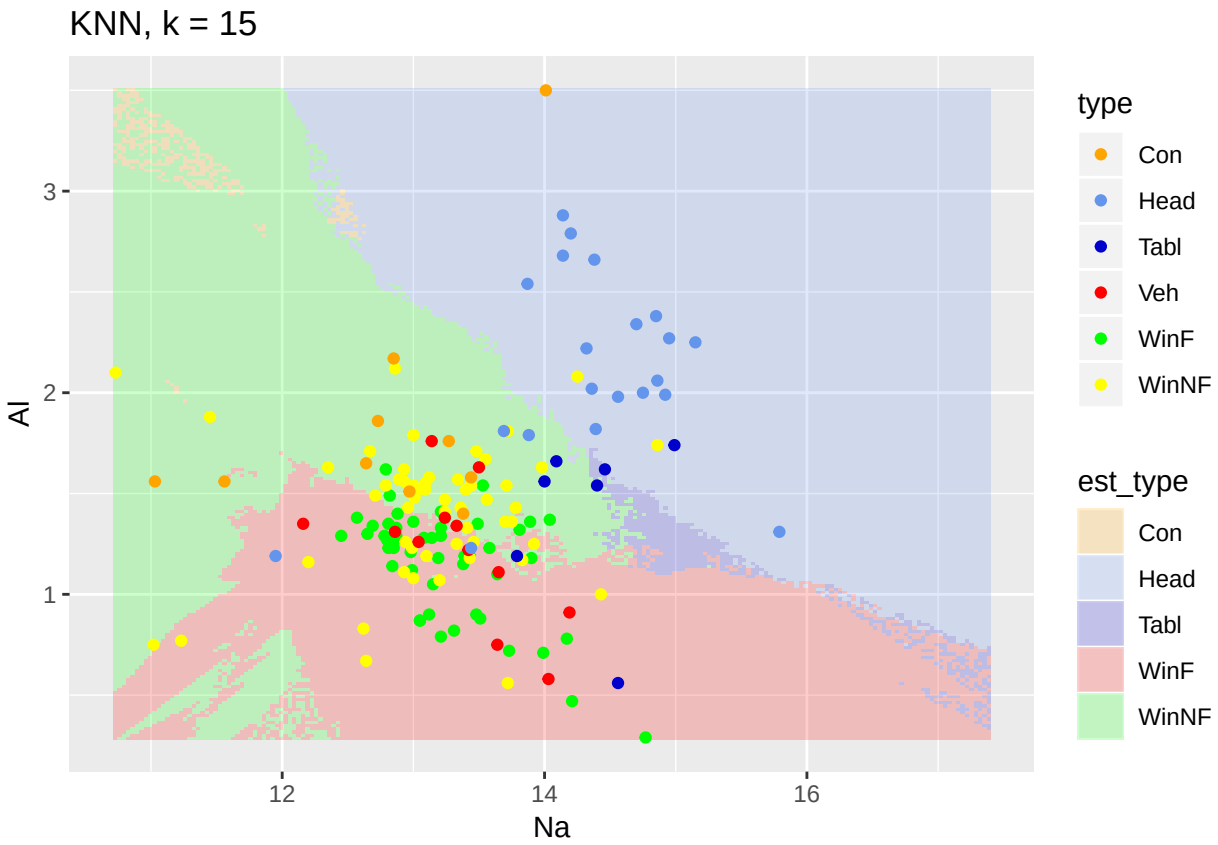

```
title <- paste0("KNN, k = ", 5)
ggplot() +
  geom_raster(data = background[[2]],
    mapping = aes(x = Na, y = Al, fill = est_type), alpha = 0.2) +
  geom_point(data = train_glass, mapping = aes(x = Na, y = Al, color = type)) +
  scale_color_manual("type", values = c("orange", "cornflowerblue", "mediumblue", "red", "green", "yellow")) +
  scale_fill_manual(values = c("orange", "cornflowerblue", "mediumblue", "red", "green", "yellow")) +
  ggtitle(title)
```



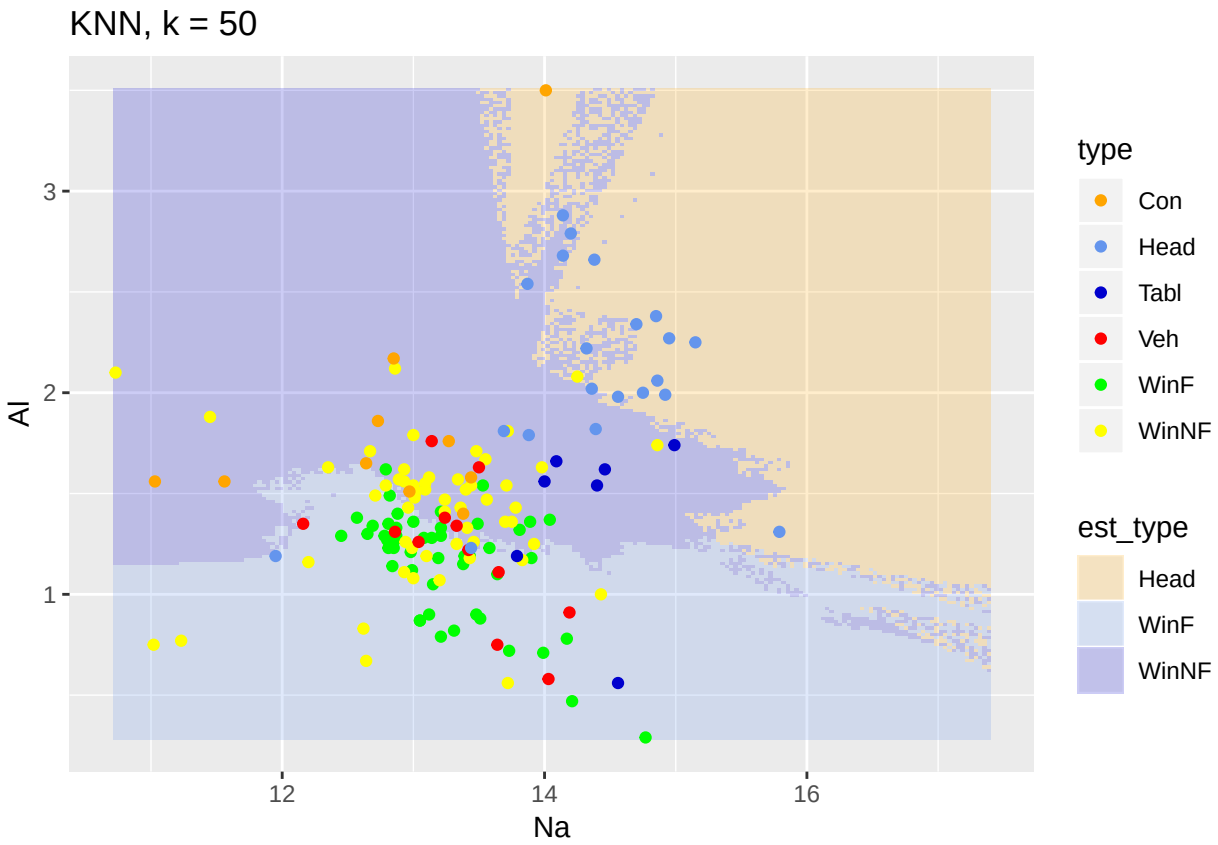

```
title <- paste0("KNN, k = ", 10)
ggplot() +
  geom_raster(data = background[[3]],
    mapping = aes(x = Na, y = Al, fill = est_type), alpha = 0.2) +
  geom_point(data = train_glass, mapping = aes(x = Na, y = Al, color = type)) +
  scale_color_manual("type", values = c("orange", "cornflowerblue", "mediumblue", "red", "green", "yellow")) +
  scale_fill_manual(values = c("orange", "cornflowerblue", "mediumblue", "red", "green", "yellow")) +
  ggtitle(title)
```



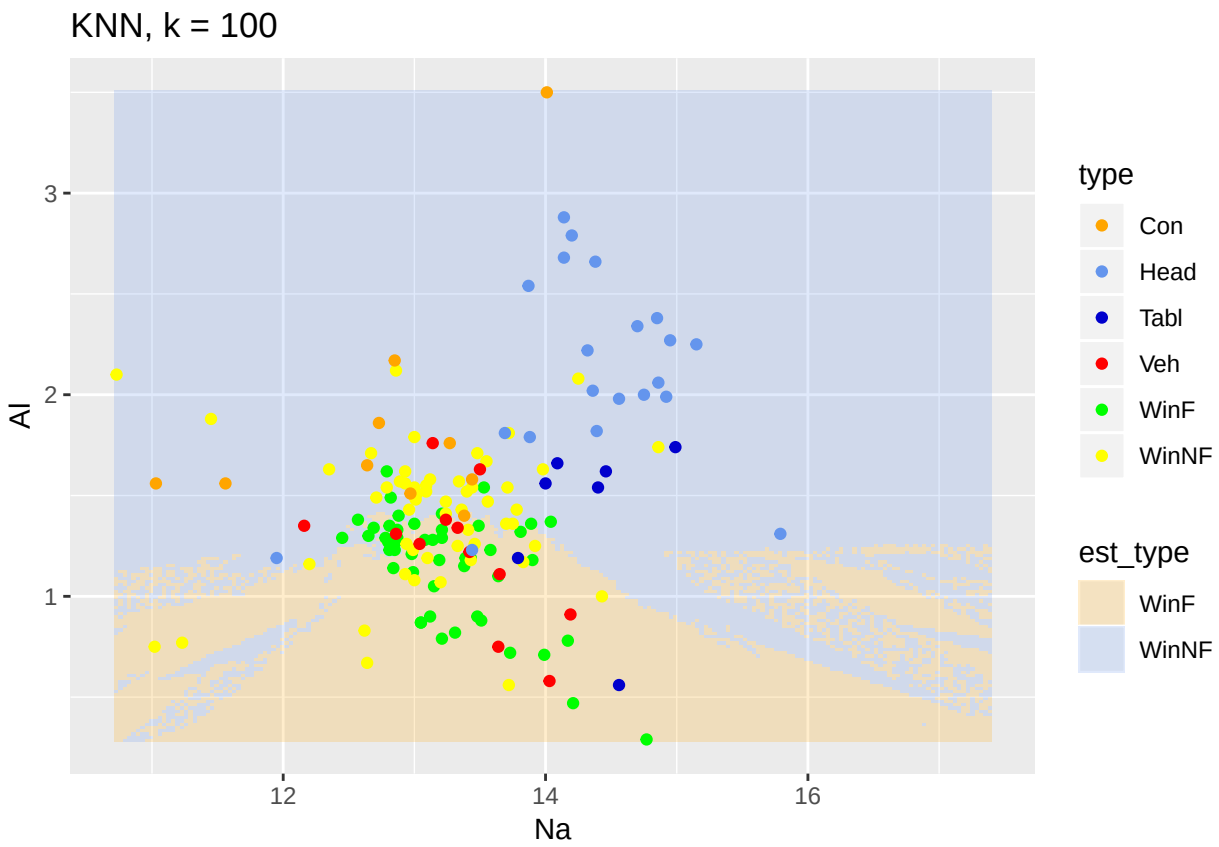

```
title <- paste0("KNN, k = ", 15)
ggplot() +
  geom_raster(data = background[[4]],
    mapping = aes(x = Na, y = Al, fill = est_type), alpha = 0.2) +
  geom_point(data = train_glass, mapping = aes(x = Na, y = Al, color = type)) +
  scale_color_manual("type", values = c("orange", "cornflowerblue", "mediumblue", "red", "green", "yellow")) +
  scale_fill_manual(values = c("orange", "cornflowerblue", "mediumblue", "red", "green", "yellow")) +
  ggtitle(title)
```



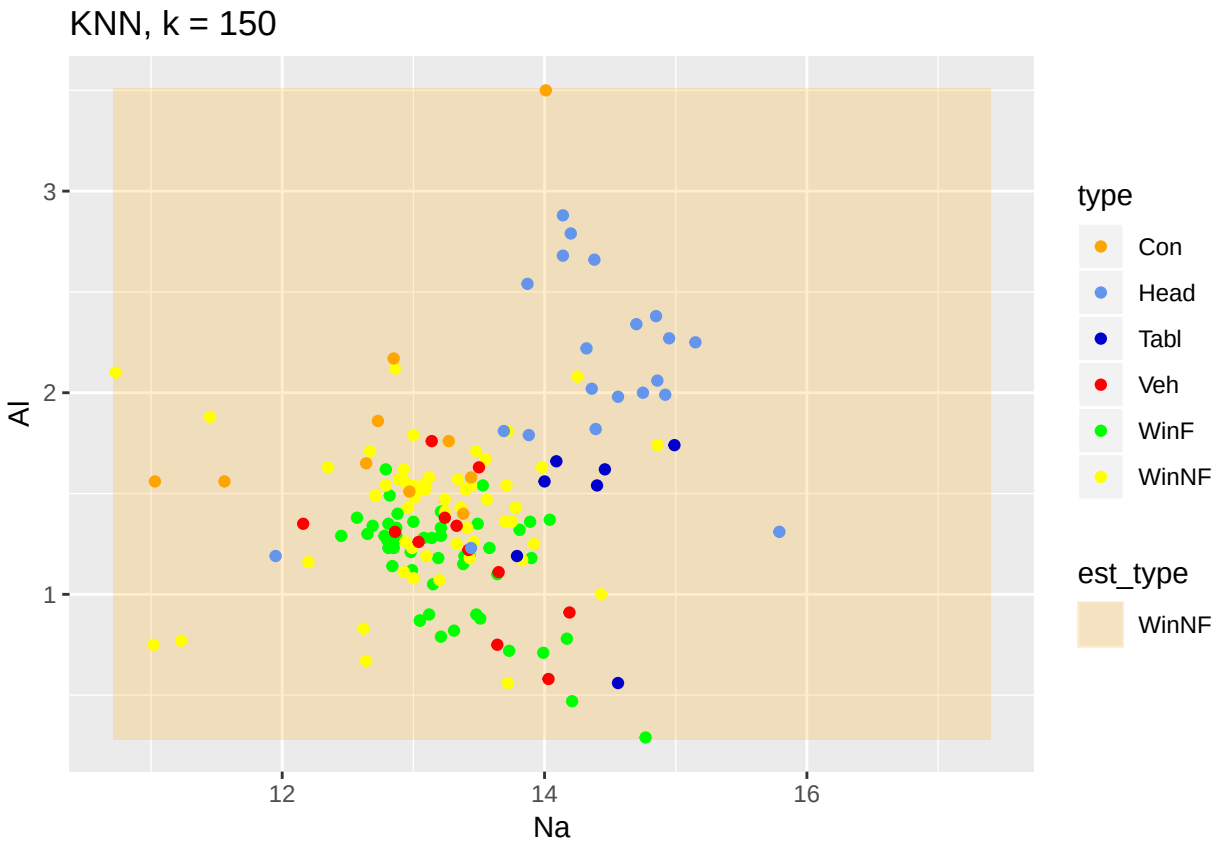

```
title <- paste0("KNN, k = ", 50)
ggplot() +
  geom_raster(data = background[[5]],
    mapping = aes(x = Na, y = Al, fill = est_type), alpha = 0.2) +
  geom_point(data = train_glass, mapping = aes(x = Na, y = Al, color = type)) +
  scale_color_manual("type", values = c("orange", "cornflowerblue", "mediumblue", "red", "green", "yellow")) +
  scale_fill_manual(values = c("orange", "cornflowerblue", "mediumblue", "red", "green", "yellow")) +
  ggtitle(title)
```




```
title <- paste0("KNN, k = ", 100)
ggplot() +
  geom_raster(data = background[[6]],
    mapping = aes(x = Na, y = Al, fill = est_type), alpha = 0.2) +
  geom_point(data = train_glass, mapping = aes(x = Na, y = Al, color = type)) +
  scale_color_manual("type", values = c("orange", "cornflowerblue", "mediumblue", "red", "green", "yellow")) +
  scale_fill_manual(values = c("orange", "cornflowerblue", "mediumblue", "red", "green", "yellow")) +
  ggtitle(title)
```




```
title <- paste0("KNN, k = ", 150)
ggplot() +
  geom_raster(data = background[[7]],
    mapping = aes(x = Na, y = Al, fill = est_type), alpha = 0.2) +
  geom_point(data = train_glass, mapping = aes(x = Na, y = Al, color = type)) +
  scale_color_manual("type", values = c("orange", "cornflowerblue", "mediumblue", "red", "green", "yellow")) +
  scale_fill_manual(values = c("orange", "cornflowerblue", "mediumblue", "red", "green", "yellow")) +
  ggtitle(title)
```



3. Use cross-validation to evaluate the classification performance of a KNN classifier using all 9 features for a few values of k.

Your code should do the following things:

1. Allocate space to store your cross-validation results. At a minimum, this will be a vector of values long enough to store the validation set classification error rate for each validation fold
2. For i = 1, . . . , number of validation folds
   a. Assemble training set data and validation set data specific to the index i
   b. Fit your model to the training set data
   c. Generate predictions for the validation set data
   d. Calculate validation set classification error rate and save it in the space you allocated in step 1 above.
3. Calculate the mean classification error rate across all 10 validation folds.

You will need to do this for each value of k you investigate. For today, you can organize this in any way that makes sense to you. Options include using two nested for loops where one iterates over the values of k you're looking at and the other iterates over cross-validation folds; manually doing steps b, c, and d for each model within a single for loop; or repeating your whole block of code multiple times, once for each value of k.

```
# mean classification error
MCE <- data.frame(
  val_fold = 1:10,
  MCE = NA
)

# different values for k neighbors
k_vector <- c(1, 5, 10, 15, 50, 100, 150)

# result of cross validation for each k
results <- data.frame(
  k_neighbor = k_vector,
  MCE_diff_k = NA
)
```

```
for (i in 1:length(k_vector)) {
  k <- k_vector[i]
  # cross - validation for each k
  for (j in 1:10) {
    crossval_val_glass <- train_glass %>% slice(crossval_val_fold_inds[[j]])
    crossval_train_glass <- train_glass %>% slice(-crossval_val_fold_inds[[j]])

    # "train" the KNN model
    knn_fit <- train(
      form = type ~ RI + Na + Mg + Al + Si + K + Ca + Ba + Fe,
      data = crossval_train_glass,
      method = "knn",
      preProcess = "scale",
      trControl = trainControl(method = "none"),
      tuneGrid = data.frame(k = 10)
    )

    y_hats <- predict(knn_fit, newdata = crossval_val_glass, type = "raw")
    head(y_hats)

    MCE$MCE[j] <- mean(y_hats != crossval_val_glass$type)
  }
  results$MCE_diff_k[i] <- mean(MCE$MCE)
}
results
```

```
##   k_neighbor MCE_diff_k
## 1          1  0.3473810
## 2          5  0.3473810
## 3         10  0.3541071
## 4         15  0.3282143
## 5         50  0.3353571
## 6        100  0.3348810
## 7        150  0.3219643
```

4. Using your selected value of k from cross-validation, refit your KNN model to the full training set and get the test set error rate. How does your performance compare to what you'd get if you just predicted the most common class in the training set?

- The value 15 of k neighbors gives the smallest misspecified classification error from cross-validation.

```
# "train" the KNN model
knn_fit <- train(
  form = type ~ RI + Na + Mg + Al + Si + K + Ca + Ba + Fe,
  data = train_glass,
  method = "knn",
  preProcess = "scale",
  trControl = trainControl(method = "none"),
  tuneGrid = data.frame(k = 15)
)

# to get estimate type probabilities, specify type = "prob" in the predict function
f_hats <- predict(knn_fit, newdata = test_glass, type = "prob")
head(f_hats)
```

```
##   Con Head Tabl        Veh      WinF      WinNF
## 1   0    0    0 0.20000000 0.4666667 0.33333333
## 2   0    0    0 0.13333333 0.7333333 0.13333333
## 3   0    0    0 0.13333333 0.7333333 0.13333333
```

```
## 4   0    0    0 0.06666667 0.6000000 0.33333333
## 5   0    0    0 0.06666667 0.7333333 0.20000000
## 6   0    0    0 0.13333333 0.8000000 0.06666667
```

```
# to get the most likely type
y_hats <- predict(knn_fit, newdata = test_glass, type = "raw")
head(y_hats)
```

```
## [1] WinF WinF WinF WinF WinF WinF
## Levels: Con Head Tabl Veh WinF WinNF
```

```
# classficiation error rate
mean(y_hats != test_glass$type)
```

```
## [1] 0.3442623
```

```
# compare to predicting the most common type in the training set
train_glass %>% count(type)
```

```
## # A tibble: 6 x 2
##   type      n
##   <fct> <int>
## 1 Con      10
## 2 Head     21
## 3 Tabl      7
## 4 Veh      12
## 5 WinF     49
## 6 WinNF    54
```

```
mean("WinNF" != test_glass$type)
```

```
## [1] 0.6393443
```

- The KNN performance with the classification error rate 0.344 is so much better what I'd get if just predicted the most common type (WinNF) in the training set with the classification error rate 0.639.