Titanic

```
library(titanic)    # loads titanic_train data frame
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.2     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
```

```
library(rpart)

# 3 significant digits
options(digits = 3)

# Clean the data
titanic_clean <- titanic_train %>%
  mutate(Survived = factor(Survived),
         Embarked = factor(Embarked),
         Age = ifelse(is.na(Age), median(Age, na.rm = TRUE), Age), # NA age to median age
         FamilySize = SibSp + Parch + 1) %>%    # count family members
  select(Survived,  Sex, Pclass, Age, Fare, SibSp, Parch, FamilySize, Embarked)
```

1. Training and Test sets: Use the caret package to create a 20% data partition based on the Survived column. Assign the 20% partition to test_set and the remaining 80% partition to train_set.

```
test_index <- createDataPartition(titanic_clean$Survived, times = 1, p = 0.2, list = FALSE)
test <- titanic_clean[test_index,]
train <- titanic_clean[-test_index,]
```

2. Baseline Prediction by Guessing the Outcome: The simplest prediction method is randomly guessing the outcome without using additional predictors. These methods will help determine whether the machine learning algorithm performs better than chance.

```r
guess_pred <- sample(c(0, 1), size = nrow(test), replace = TRUE)
#accuracy of this guessing method
mean(guess_pred == test$Survived)
```

```
## [1] 0.508
```

3. Predicting Survival by Sex

```r
library(broom)

#a. Use the training set to determine whether members of a given sex were more likely to survive or die
train %>%
  group_by(Sex) %>%
  summarize(Survived = mean(Survived == 1))
```

```
## # A tibble: 2 x 2
##   Sex    Survived
##   <chr>     <dbl>
## 1 female    0.733
## 2 male      0.198
```

```r
#b. Predict survival using sex on the test set: if the survival rate for a sex is over 0.5, predict sur
sex_model <- train %>%
  group_by(Sex) %>%
  summarize(Survived_predict = ifelse(mean(Survived == 1) > 0.5, 1, 0))

test_set1 <- test %>%
  inner_join(sex_model, by = 'Sex')

cm1 <- confusionMatrix(data = factor(test_set1$Survived_predict), reference = factor(test_set1$Survived
cm1 %>% tidy() %>% filter(term == "accuracy")
```

```
## # A tibble: 1 x 6
##   term     class estimate conf.low conf.high       p.value
##   <chr>    <chr>    <dbl>    <dbl>     <dbl>         <dbl>
## 1 accuracy <NA>     0.821    0.757     0.874 0.00000000172
```

4. Predicting survival by Passenger Class

```r
pclass_model <- train %>%
  group_by(Pclass) %>%
  summarize(Survived_predict = ifelse(mean(Survived == 1) > 0.5, 1, 0))

test_set2 <- test %>%
  inner_join(pclass_model, by = 'Pclass')

cm2 <- confusionMatrix(data = factor(test_set2$Survived_predict), reference = factor(test_set2$Survived
cm2 %>% tidy() %>%
  filter(term == 'accuracy')
```

```
## # A tibble: 1 x 6
##   term     class estimate conf.low conf.high p.value
##   <chr>    <chr>    <dbl>    <dbl>     <dbl>   <dbl>
## 1 accuracy <NA>     0.670    0.596     0.739  0.0712
```

Use the training set to group passengers by both sex and passenger class. Which sex and class combinations were more likely to survive than die (i.e. >50% survival)?

```
combined_model <- train %>%
  group_by(Pclass, Sex) %>%
  summarize(Survived_predict = ifelse(mean(Survived == 1) > 0.5, 1, 0))
```

```
## `summarise()` has grouped output by 'Pclass'. You can override using the `.groups` argument.
```

```
test_set3 <- test %>%
  inner_join(combined_model, by = c('Pclass', 'Sex'))

cm3 <- confusionMatrix(data = factor(test_set3$Survived_predict), reference = factor(test_set3$Survived
cm3 %>% tidy() %>%
  filter(term == 'accuracy')
```

```
## # A tibble: 1 x 6
##   term     class estimate conf.low conf.high      p.value
##   <chr>    <chr>    <dbl>    <dbl>     <dbl>        <dbl>
## 1 accuracy <NA>     0.804    0.739     0.860 0.0000000358
```

5. Confusion Matrix: create confusion matrices for the combined sex and class model and inspect sensitivity, specificity and balanced accuracy.

```
cm3 %>% tidy() %>%
  filter(term == 'sensitivity' | term == 'specificity' | term == 'balanced_accuracy')
```

```
## # A tibble: 3 x 6
##   term              class estimate conf.low conf.high p.value
##   <chr>             <chr>    <dbl>    <dbl>     <dbl>   <dbl>
## 1 sensitivity       0        0.982       NA        NA      NA
## 2 specificity       0        0.522       NA        NA      NA
## 3 balanced_accuracy 0        0.752       NA        NA      NA
```

6. Calculate scores for the sex model, class model, and combined sex and class model.

```
F_meas(data = factor(test_set1$Survived_predict), reference = factor(test_set1$Survived))
```

```
## [1] 0.856
```

```
F_meas(data = factor(test_set2$Survived_predict), reference = factor(test_set2$Survived))
```

```
## [1] 0.763
```

```r
F_meas(data = factor(test_set3$Survived_predict), reference = factor(test_set3$Survived))
```

```
## [1] 0.861
```

7. Survival by Fare - LDA and QDA

```r
#Train a model using linear discriminant analysis (LDA)
fit_lda <- train(Survived ~ Fare, data=train, method = 'lda')
survived_hat <- predict(fit_lda, test)
#accuracy
mean(survived_hat == test$Survived)
```

```
## [1] 0.654
```

```r
#qda model
fit_qda <- train(Survived ~ Fare, data=train, method = 'qda')
survived_hat <- predict(fit_qda, test)
#accuracy
mean(survived_hat == test$Survived)
```

```
## [1] 0.642
```

8. Logistic Regression Models

```r
#Train a logistic regression model using age as the only predictor
fit_sex <- glm(Survived ~ Sex, data=train, family = binomial)
survived_hat <- predict(fit_sex, test)
survived_hat <- ifelse(survived_hat >= 0, 1, 0)
mean(survived_hat == test$Survived)
```

```
## [1] 0.821
```

```r
#Train a logistic regression model using all predictors
fit_all <- glm(Survived ~ ., data=train, family = binomial)
survived_hat <- predict(fit_all, test, type = "response")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```r
survived_hat <- ifelse(survived_hat >= 0.5, 1, 0)
mean(survived_hat == test$Survived)
```

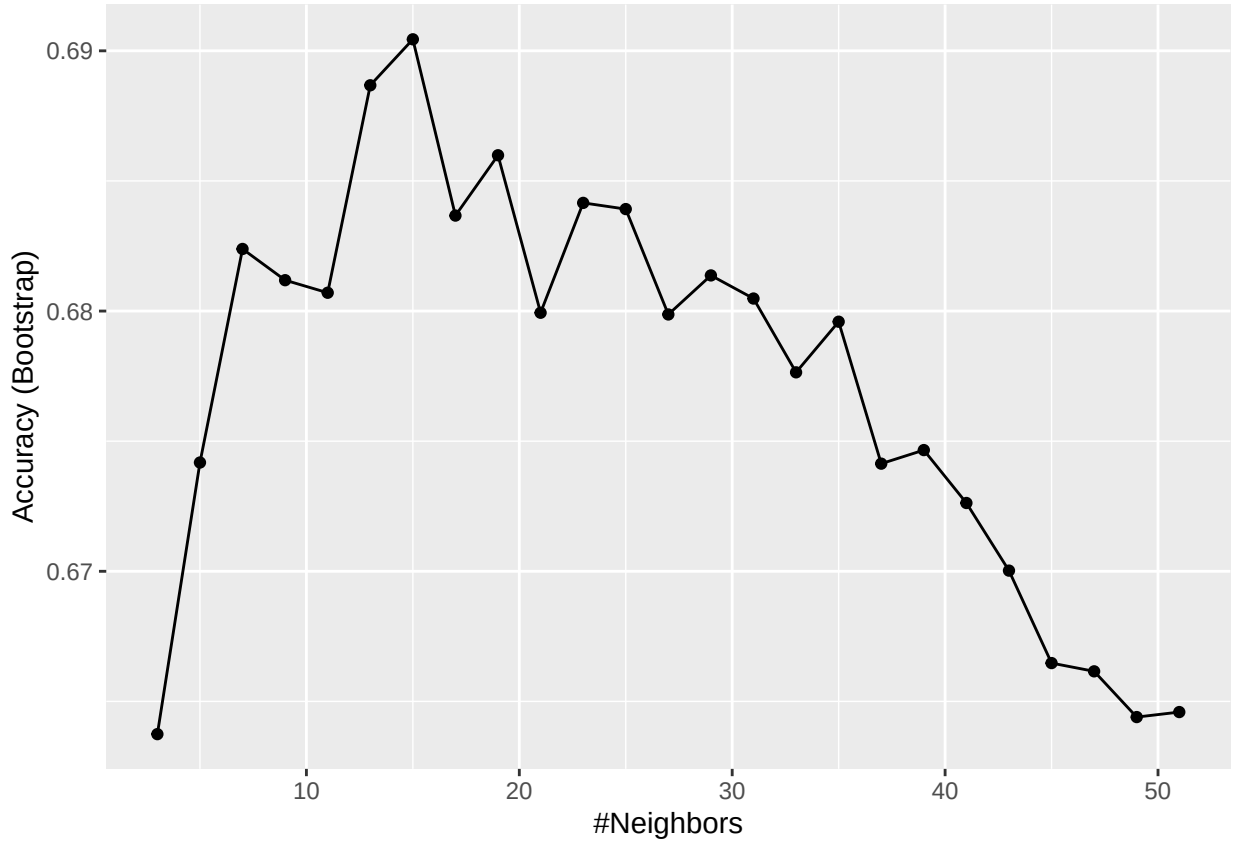
```
## [1] 0.816
```

9.kNN model

```r
#tuning model with k = seq(3, 51, 2)
k <- seq(3, 51, 2)
fit_knn <- train(Survived ~ ., data = train, method = "knn", tuneGrid = data.frame(k))
fit_knn$bestTune
```

```
##    k
## 7 15
```

```r
#plotting k values
ggplot(fit_knn)
```




```r
#of 7, 11, 17 and 21, which yields the highest accuracy
fit_knn$results %>% filter(k %in% c(7, 11, 17, 21)) %>% pull(Accuracy)
```

```
## [1] 0.682 0.681 0.684 0.680
```

```r
#accuracy
survived_hat <- predict(fit_knn, test)
mean(survived_hat == test$Survived)
```

```
## [1] 0.693
```

10. Cross-Validation: Instead of the default training control, use 10-fold cross-validation where each partition consists of 10% of the total.

```r
#tuning model with k = seq(3, 51, 2)
control <- trainControl(method = "cv", number = 10, p = .9)
train_knn_cv <- train(Survived ~ ., method = "knn",
                      data = train,
                      tuneGrid = data.frame(k = seq(3, 51, 2)),
                      trControl = control)

#optimal value of k
train_knn_cv$bestTune
```

```
##    k
## 8 17
```

```r
#accuracy of test set
survived_hat <- predict(train_knn_cv, test)
mean(survived_hat == test$Survived)
```

```
## [1] 0.693
```

11. Classification Tree Model

a) Tune the complexity parameter with cp = seq(0, 0.05, 0.002)

```r
fit_tree <- train(Survived ~ ., method = "rpart",
                  data = train,
                  tuneGrid = data.frame(cp = seq(0, 0.05, 0.002)))
fit_tree$bestTune
```

```
##      cp
## 8 0.014
```

```r
fit_tree$results
```

```
##       cp Accuracy Kappa AccuracySD KappaSD
## 1  0.000    0.788 0.545     0.0216  0.0435
## 2  0.002    0.790 0.548     0.0221  0.0444
## 3  0.004    0.796 0.559     0.0230  0.0479
## 4  0.006    0.799 0.562     0.0211  0.0457
## 5  0.008    0.802 0.567     0.0199  0.0472
## 6  0.010    0.803 0.568     0.0179  0.0427
## 7  0.012    0.807 0.573     0.0194  0.0444
## 8  0.014    0.807 0.572     0.0186  0.0452
## 9  0.016    0.802 0.560     0.0211  0.0478
## 10 0.018    0.800 0.557     0.0210  0.0472
## 11 0.020    0.800 0.553     0.0179  0.0442
## 12 0.022    0.800 0.552     0.0182  0.0438
## 13 0.024    0.799 0.550     0.0166  0.0421
## 14 0.026    0.798 0.548     0.0165  0.0424
## 15 0.028    0.798 0.548     0.0165  0.0424
## 16 0.030    0.796 0.542     0.0166  0.0418
```

```
## 17 0.032    0.793 0.538     0.0193  0.0447
## 18 0.034    0.791 0.534     0.0207  0.0491
## 19 0.036    0.791 0.533     0.0208  0.0497
## 20 0.038    0.789 0.532     0.0211  0.0491
## 21 0.040    0.787 0.527     0.0206  0.0475
## 22 0.042    0.785 0.524     0.0187  0.0436
## 23 0.044    0.781 0.516     0.0188  0.0431
## 24 0.046    0.781 0.515     0.0188  0.0424
## 25 0.048    0.780 0.513     0.0181  0.0417
## 26 0.050    0.780 0.513     0.0181  0.0417
```

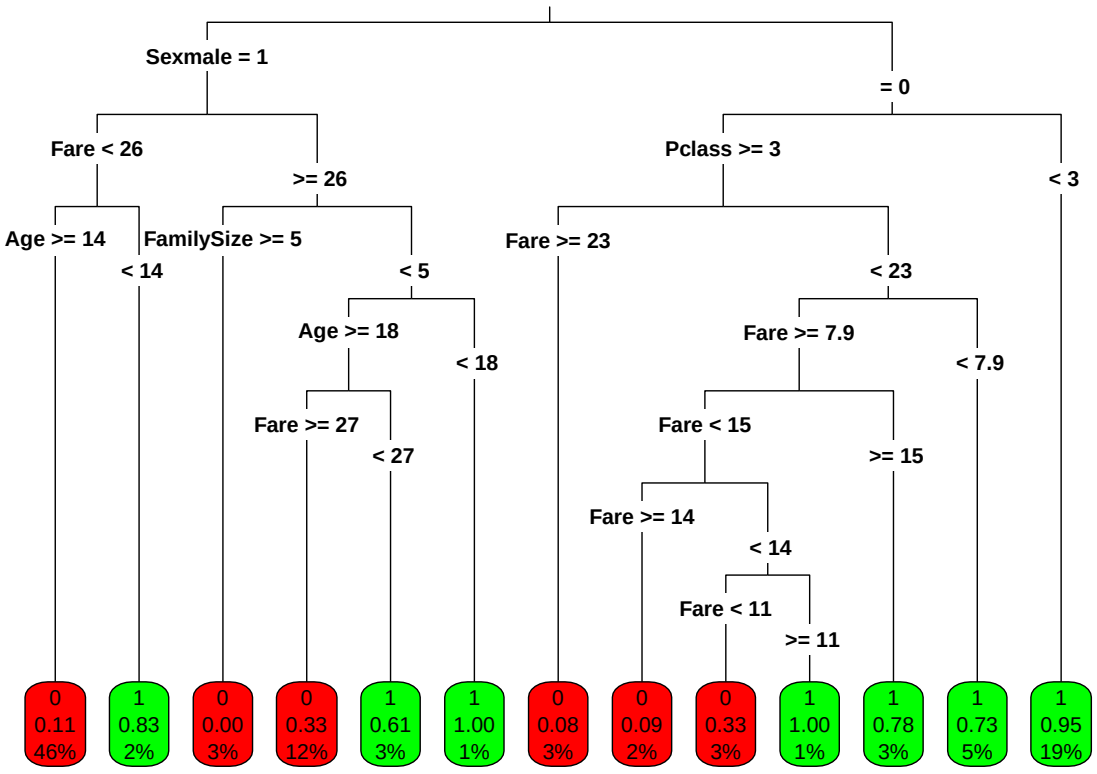
```r
#accuracy with test set
survived_hat <- predict(fit_tree, test)
mean(survived_hat == test$Survived)
```

```
## [1] 0.838
```

b. Inspect the final model and plot the decision tree.

```r
library(rpart.plot)

rpart.plot(fit_tree$finalModel, type = 3, box.palette = c("red", "green"), fallen.leaves = TRUE)
```



12. Random Forest Model

```r
#Test values of mtry = seq(1:7)
#Set ntree to 100
fit_rf <- train(Survived ~ ., data = train, method = "rf",
                tuneGrid = data.frame(mtry = seq(1:7)), ntree = 100)
fit_rf$bestTune
```

```
##   mtry
## 4    4
```

```r
#accuracy
survived_hat <- predict(fit_rf, test)
confusionMatrix(data = factor(survived_hat), reference = factor(test$Survived))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 97 14
##          1 13 55
##
##                Accuracy : 0.849
##                  95% CI : (0.788, 0.898)
##     No Information Rate : 0.615
##     P-Value [Acc > NIR] : 5.17e-12
##
##                   Kappa : 0.681
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.882
##             Specificity : 0.797
##          Pos Pred Value : 0.874
##          Neg Pred Value : 0.809
##              Prevalence : 0.615
##          Detection Rate : 0.542
##    Detection Prevalence : 0.620
##       Balanced Accuracy : 0.839
##
##        'Positive' Class : 0
##
```

```r
mean(survived_hat == test$Survived)
```

```
## [1] 0.849
```

```r
#determine the importance of various predictors to the random forest model
varImp(fit_rf$finalModel)
```

```
##            Overall
## Sexmale      83.87
## Pclass       23.34
```

```
## Age          57.72
## Fare         71.34
## SibSp        11.39
## Parch         7.36
## FamilySize   16.44
## EmbarkedC     5.59
## EmbarkedQ     1.53
## EmbarkedS     4.18
```