Final Homework

Gloria Grace

Supervised Learning

```
head(nasaweather::storms)
```

```
## # A tibble: 6 x 11
##   name     year month   day  hour   lat  long pressure  wind type        seasday
##   <chr>   <int> <int> <int> <int> <dbl> <dbl>    <int> <int> <chr>         <int>
## 1 Allison  1995     6     3     0  17.4 -84.3     1005    30 Tropical D~       3
## 2 Allison  1995     6     3     6  18.3 -84.9     1004    30 Tropical D~       3
## 3 Allison  1995     6     3    12  19.3 -85.7     1003    35 Tropical S~       3
## 4 Allison  1995     6     3    18  20.6 -85.8     1001    40 Tropical S~       3
## 5 Allison  1995     6     4     0  22   -86        997    50 Tropical S~       4
## 6 Allison  1995     6     4     6  23.3 -86.3      995    60 Tropical S~       4
```

```
storms <- nasaweather::storms %>% mutate(type = as_factor(type))
head(storms)
```

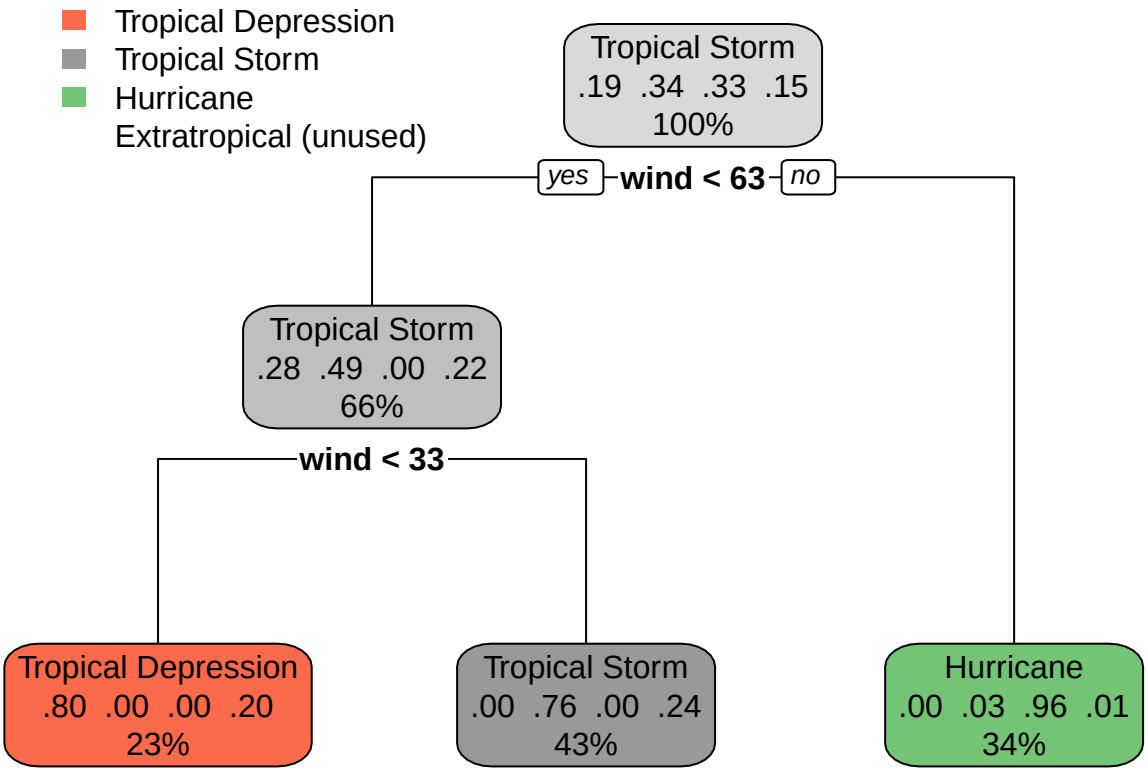
```
## # A tibble: 6 x 11
##   name     year month   day  hour   lat  long pressure  wind type        seasday
##   <chr>   <int> <int> <int> <int> <dbl> <dbl>    <int> <int> <fct>         <int>
## 1 Allison  1995     6     3     0  17.4 -84.3     1005    30 Tropical D~       3
## 2 Allison  1995     6     3     6  18.3 -84.9     1004    30 Tropical D~       3
## 3 Allison  1995     6     3    12  19.3 -85.7     1003    35 Tropical S~       3
## 4 Allison  1995     6     3    18  20.6 -85.8     1001    40 Tropical S~       3
## 5 Allison  1995     6     4     0  22   -86        997    50 Tropical S~       4
## 6 Allison  1995     6     4     6  23.3 -86.3      995    60 Tropical S~       4
```

```
add_predictions <- function(data, model, variable_name = ".pred", model_name = deparse(substitute(model)
  model %>%
    predict(data) %>%
    rename(!!enquo(variable_name) := .pred) %>%
    mutate(model = model_name) %>%
    bind_cols(data)
}
```

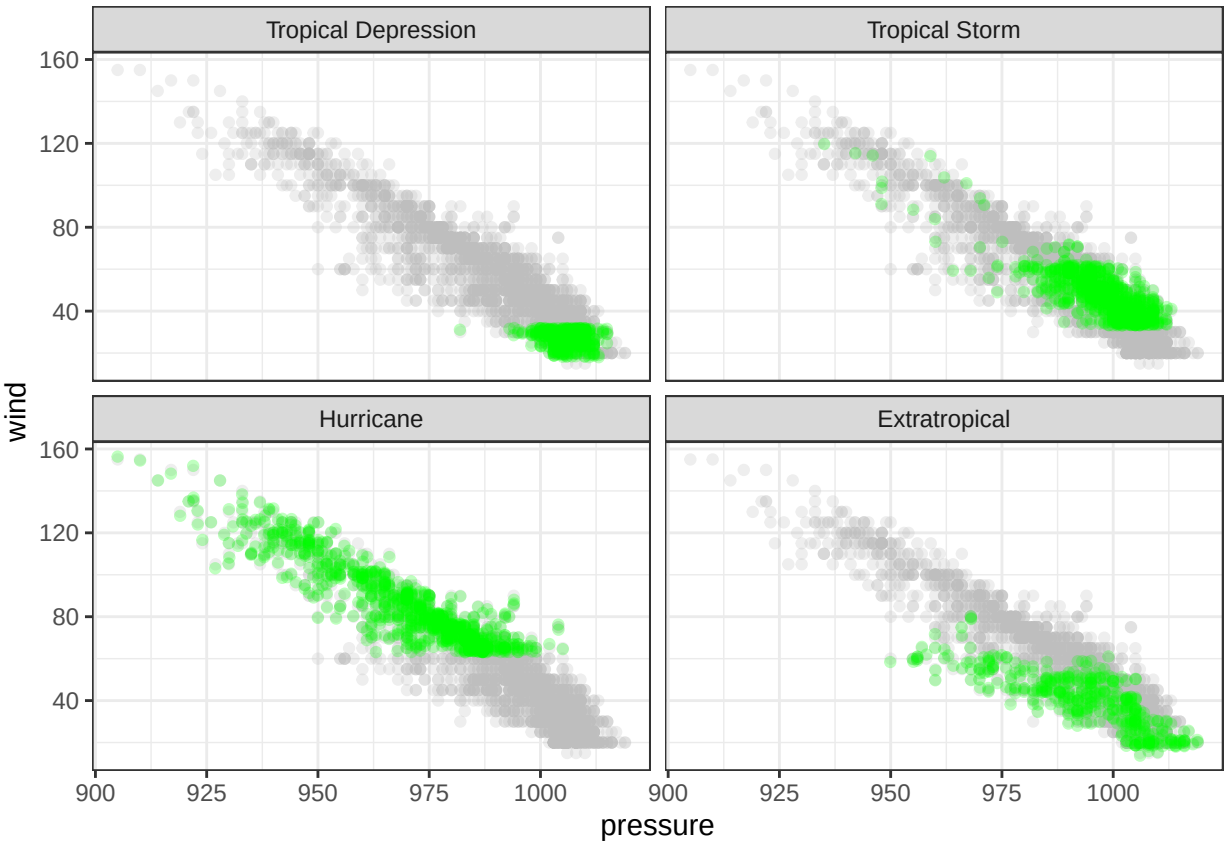
```
set.seed(0)
model <-
  decision_tree(mode = "classification", tree_depth = 2) %>%
  fit(type ~ wind + pressure, data = storms)

model %>%
  extract_fit_engine() %>%
  rpart.plot()
```

```
## Warning: Cannot retrieve the data used to build the model (so cannot determine roundint and is.binary
```

```
## To silence this warning:
##     Call rpart.plot with roundint=FALSE,
##     or rebuild the rpart model with model=TRUE.
```




```
ggplot(data = storms, aes(x = pressure, y = wind)) +
  geom_point(data = storms %>% select(-type), alpha = .25, color = "grey") +
  geom_point(alpha = .25, position = position_jitter(.1), color = "green") +
  facet_wrap(vars(type))
```



Clustering

```
big_cities <- mdsr::world_cities %>%
  arrange(desc(population)) %>%
  slice_head(n = 4000)
```

```
data_for_clustering <- big_cities %>%
  select(latitude, longitude)

set.seed(20211119)
clustering_results <- data_for_clustering %>%
  kmeans(nstart = 10, centers = 2)

cities_with_clusters <- big_cities %>%
  mutate(cluster = as.factor(clustering_results$cluster))
```

```
glance(clustering_results)
```

```
## # A tibble: 1 x 4
##       totss tot.withinss betweenss  iter
##       <dbl>        <dbl>     <dbl> <int>
## 1 24082261.     8637291. 15444970.     1
```

```
tidy(clustering_results)
```

```
## # A tibble: 2 x 5
##   latitude longitude  size withinss cluster
```

```
##      <dbl>     <dbl> <int>    <dbl> <fct>
## 1     27.0      81.4  2406 4359783. 1
## 2     21.3     -45.4  1594 4277507. 2
```

Clustering with k = 2 has a high difference between one and another. But if I added another one to k = 3 the result would be more closer with each other.

Databases

```
HallOfFame <- Lahman::HallOfFame
hr <- Lahman::Batting
people <- Lahman::People
```

```
hr_select <- hr %>%
  select(playerID, H, HR)
```

```
names <- people %>%
  select(nameFirst, nameLast, playerID)
```

```
hr_select %>%
  filter(HR == 500)
```

```
## [1] playerID H        HR
## <0 rows> (or 0-length row.names)
```

```
hr_select %>%
  filter(H == 3000)
```

```
## [1] playerID H        HR
## <0 rows> (or 0-length row.names)
```

There is no more player that have not been inducted into the Baseball Hall of Fame. In the Batting Data, there isn't any players that hit either 500 home runs or 3000 hits.

Text Data

```
macbeth_url <- "http://www.gutenberg.org/cache/epub/1129/pg1129.txt"
#macbeth_raw <- read_file(macbeth_url)
data(Macbeth_raw, package = "mdsr")
```

```
macbeth <- Macbeth_raw %>%
  stringi::stri_split_lines() %>%
  pluck(1)
```

```
pattern <- "\\s + \\s + [A-Z] + \\."
macbeth %>%
  str_detect(pattern) %>%
  sum()
```

```
## [1] 0
```

There is no speaking lines in Macbeth.

```
baby_n <- babynames::babynames %>%
  filter(sex == 'M') %>%
  select(name) %>%
```

```
  stringi::stri_split_lines() %>%
  pluck(1)
```

```
## Warning in stringi::stri_split_lines(.): argument is not an atomic vector;
## coercing
```

```
name <- "George"
name1 <- "Joe"
name2 <- "Charlie"
name3 <-"Jesse"
name4 <- "Diego"
name5 <- "Eugene"
name6 <- "Leo"
name7 <- "Luke"
name8 <- "Joshua"
name9 <- "Dave"
name10 <- "Jake"
```

The most popular name out of the 10 names is Leo and Joe as the second most popular.

```
baby_n %>%
  str_detect(name) %>%
  sum()
```

```
## [1] 231
```

```
baby_n %>%
  str_detect(name1) %>%
  sum()
```

```
## [1] 882
```

```
baby_n %>%
  str_detect(name2) %>%
  sum()
```

```
## [1] 167
```

```
baby_n %>%
  str_detect(name3) %>%
  sum()
```

```
## [1] 381
```

```
baby_n %>%
  str_detect(name4) %>%
  sum()
```

```
## [1] 110
```

```
baby_n %>%
  str_detect(name5) %>%
  sum()
```

```
## [1] 138
```

```
baby_n %>%
  str_detect(name6) %>%
  sum()
```

```
## [1] 1442
```

```
baby_n %>%
  str_detect(name7) %>%
  sum()
```

```
## [1] 168
```

```
baby_n %>%
  str_detect(name8) %>%
  sum()
```

```
## [1] 257
```

```
baby_n %>%
  str_detect(name9) %>%
  sum()
```

```
## [1] 451
```

```
baby_n %>%
  str_detect(name10) %>%
  sum()
```

```
## [1] 315
```

This code didn't work because there are and " that are considered as the end of the name. I don't know how to remove those characters and ended up looking for the names manually.

```
baby_n %>%
  str_detect("[aiueo]$") %>%
  sum()
```

```
## [1] 0
```

‘‘