

# Dataset Description

## Imports

```
library(fairmodels)
library(tidymodels)
```

```
## -- Attaching packages ----- tidymodels 0.2.0 --
```

```
## v broom      0.8.0    v recipes      0.2.0
## v dials      0.1.1    v rsample      0.1.1
## v dplyr      1.0.9    v tibble       3.1.7
## v ggplot2    3.3.6    v tidyr        1.2.0
## v infer      1.0.0    v tune         0.2.0
## v modeldata  0.1.1    v workflows    0.2.6
## v parsnip    0.2.1    v workflowsets 0.2.1
## v purrr      0.3.4    v yardstick    0.0.9
```

```
## -- Conflicts ----- tidymodels_conflicts() --
```

```
## x purrr::discard() masks scales::discard()
## x dplyr::filter()   masks stats::filter()
## x dplyr::lag()      masks stats::lag()
## x recipes::step()   masks stats::step()
## * Use suppressPackageStartupMessages() to eliminate package startup messages
```

```
library(rpart)
```

```
##
```

```
## Attaching package: 'rpart'
```

```
## The following object is masked from 'package:dials':
```

```
##
```

```
##      prune
```

```
library(discrim)
```

```
##
```

```
## Attaching package: 'discrim'
```

```
## The following object is masked from 'package:dials':
```

```
##
```

```
##      smoothness
```

```
source("../scripts/metrics_on_dataset.R")
```

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

```
## v readr 2.1.2 v forcats 0.5.1
## v stringr 1.4.0
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x readr::col_factor() masks scales::col_factor()
## x purrr::discard() masks scales::discard()
## x dplyr::filter() masks stats::filter()
## x stringr::fixed() masks recipes::fixed()
## x dplyr::lag() masks stats::lag()
## x readr::spec() masks yardstick::spec()
```

## The Dataset

```
df <- read_rds("../data/selection.rds")
head(df, 10)
```

```
## nationality gender test_score english_cert extracurricular rating accepted
## 1 Non_Dutch Female 6.7 FALSE TRUE 6.7 FALSE
## 2 Non_Dutch Female 8.8 TRUE TRUE 8.8 TRUE
## 3 Non_Dutch Female 5.1 FALSE FALSE 5.1 FALSE
## 4 Dutch Female 4.8 FALSE FALSE 5.6 FALSE
## 5 Dutch Male 6.4 FALSE FALSE 7.2 FALSE
## 6 Non_Dutch Male 7.1 TRUE FALSE 7.1 FALSE
## 7 Non_Dutch Male 4.9 FALSE FALSE 4.9 FALSE
## 8 Dutch Male 4.8 FALSE FALSE 5.6 FALSE
## 9 Dutch Female 5.1 FALSE FALSE 5.9 FALSE
## 10 Non_Dutch Male 7.6 FALSE TRUE 7.6 FALSE
```

## Characteristics

```
summary(df)
```

```
## nationality gender test_score english_cert extracurricular
## Dutch :497 Female:514 Min. : 1.800 Mode :logical Mode :logical
## Non_Dutch:503 Male :486 1st Qu.: 5.475 FALSE:666 FALSE:600
## Median : 6.500 TRUE :334 TRUE :400
## Mean : 6.511
## 3rd Qu.: 7.525
## Max. :10.000
## rating accepted
## Min. : 2.400 Mode :logical
## 1st Qu.: 5.900 FALSE:837
## Median : 6.950 TRUE :163
```

```
## Mean    : 6.903
## 3rd Qu.: 7.900
## Max.    :10.000
```

```
summary(filter(df, accepted))
```

```
##      nationality      gender      test_score      english_cert      extracurricular
## Dutch      :107  Female:77  Min.      : 7.700  Mode :logical  Mode :logical
## Non_Dutch: 56   Male  :86   1st Qu.: 8.200  FALSE:117  FALSE:102
##                                     Median : 8.600  TRUE :46   TRUE :61
##                                     Mean    : 8.684
##                                     3rd Qu.: 9.050
##                                     Max.    :10.000
##      rating      accepted
## Min.      : 8.500  Mode:logical
## 1st Qu.: 8.800  TRUE:163
## Median : 9.100
## Mean    : 9.175
## 3rd Qu.: 9.600
## Max.    :10.000
```

```
summary(filter(df, !accepted))
```

```
##      nationality      gender      test_score      english_cert      extracurricular
## Dutch      :390  Female:437  Min.      :1.800  Mode :logical  Mode :logical
## Non_Dutch:447  Male  :400   1st Qu.:5.200  FALSE:549  FALSE:498
##                                     Median :6.200  TRUE :288  TRUE :339
##                                     Mean    :6.088
##                                     3rd Qu.:7.000
##                                     Max.    :8.400
##      rating      accepted
## Min.      :2.40  Mode :logical
## 1st Qu.:5.60  FALSE:837
## Median :6.60
## Mean    :6.46
## 3rd Qu.:7.40
## Max.    :8.40
```

## Metrics on the Dataset

```
model <- decision_tree(mode = "classification")
results <- all_metrics(df, model)
```

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

```
print_all_metrics("Unmodified dataset", results)
```

```
## [1] "Unmodified dataset"
## [1] "Group fairness"
## # A tibble: 4 x 4
##   nationality accepted total  perc
##   <fct>         <lgl>    <int> <dbl>
## 1 Dutch        FALSE    390  78.5
## 2 Dutch        TRUE     107  21.5
## 3 Non_Dutch    FALSE    447  88.9
## 4 Non_Dutch    TRUE      56  11.1
## [1] "Causal discrimination"
## [1] 0.123
## [1] "Unawareness"
## # A tibble: 4 x 4
##   nationality predicted_accepted total  perc
##   <fct>         <fct>          <int> <dbl>
## 1 Dutch        FALSE          402  80.9
## 2 Dutch        TRUE           95  19.1
## 3 Non_Dutch    FALSE          435  86.5
## 4 Non_Dutch    TRUE           68  13.5
```