

Breast Cancer Prediction

Andrea Becsek

2 January 2020

```
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.2.1 --
## v ggplot2 3.2.1      v purrr  0.3.2
## v tibble  2.1.3      v dplyr  0.8.0.1
## v tidyr   0.8.3      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.4.0

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(ggplot2)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##   set_names

## The following object is masked from 'package:tidyr':
##
##   extract
```

```
library(dplyr)
```

Import data

```
data <- read.csv("../data/data.csv")
dim(data)
```

```
## [1] 569  33
```

```
print(colnames(data))
```

```
## [1] "id"                "diagnosis"
## [3] "radius_mean"       "texture_mean"
## [5] "perimeter_mean"    "area_mean"
## [7] "smoothness_mean"   "compactness_mean"
## [9] "concavity_mean"    "concave.points_mean"
## [11] "symmetry_mean"     "fractal_dimension_mean"
## [13] "radius_se"         "texture_se"
## [15] "perimeter_se"      "area_se"
## [17] "smoothness_se"     "compactness_se"
## [19] "concavity_se"      "concave.points_se"
```

```
## [21] "symmetry_se"          "fractal_dimension_se"
## [23] "radius_worst"        "texture_worst"
## [25] "perimeter_worst"     "area_worst"
## [27] "smoothness_worst"    "compactness_worst"
## [29] "concavity_worst"     "concave.points_worst"
## [31] "symmetry_worst"      "fractal_dimension_worst"
## [33] "X"
```

Check for NAs

```
data %>%
  summarise_all(funs(sum(is.na(.))))
```

```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
```

```
## please use list() instead
```

```
##
```

```
## # Before:
```

```
## funs(name = f(.))
```

```
##
```

```
## # After:
```

```
## list(name = ~f(.))
```

```
## This warning is displayed once per session.
```

```
##   id diagnosis radius_mean texture_mean perimeter_mean area_mean
## 1  0          0          0          0          0          0
##   smoothness_mean compactness_mean concavity_mean concave.points_mean
## 1          0          0          0          0
##   symmetry_mean fractal_dimension_mean radius_se texture_se perimeter_se
## 1          0          0          0          0          0
##   area_se smoothness_se compactness_se concavity_se concave.points_se
## 1          0          0          0          0          0
##   symmetry_se fractal_dimension_se radius_worst texture_worst perimeter_worst
## 1          0          0          0          0          0
##   area_worst smoothness_worst compactness_worst concavity_worst
## 1          0          0          0          0
##   concave.points_worst symmetry_worst fractal_dimension_worst   X
## 1          0          0          0          0 569
```

Remove *id* and *X*

```
data %<>%
  dplyr::select(-c(id, X))
```