COSC 5557: Practical Machine Learning

Exploratory Data Analysis- Primary Tumor Data

Abiodun Awosola

2024-02-15

Note: Not all lines of code are displayed.

Loading the Primary Tumor Data

```
knitr::opts_chunk$set(comment = NA) # removes '##' from outputs

#Imports Data

tumor_dat1 <- read.csv(

  "primary-tumor.data", sep = ",",  check.names = TRUE, header=F,
  col.names=c("class", "age", "sex", "histologic-type", "degree-of-diffe", "bone",
"bone-marrow", "lung", "pleura", "peritoneum", "liver",
"brain", "skin", "neck", "supraclavicular", "axillar", "mediastinum", "abdominal") )
```

Exploring the Data

As a first step, we explore the data and look for simple problems such as constant or duplicated features. This can be done quite efficiently with a package like `DataExplorer` or `skimr` which can be used to create a large number of informative plots.

Below we summarize the most important findings for data cleaning, but we only consider this aspect in a cursory manner:

Data Attributes

```
[1] "C"
```

Table 1: Data summary

Name	tumor_dat1
Number of rows	339
Number of columns	18
_____	
Column type frequency:	
character	5
numeric	13
_____	
Group variables	None

Variable type: character

skim_variable	n_missing
sex	1
histologic.type	67
degree.of.diffe	155
skin	1
axillar	1

Variable type: numeric

skim_variable	n_missing
class	0
age	0
bone	0
bone.marrow	0
lung	0
pleura	0
peritoneum	0
liver	0
brain	0
neck	0
supraclavicular	0
mediastinum	0
abdominal	0

For this data, all the variables are categorical, and so they are expected to be character data type and in levels. They are already in levels. However, not all the variables are character data type, as seen from the output from skimming the data. There are also missing values.

The next thing is to clean up the data by fixing those anomalies.

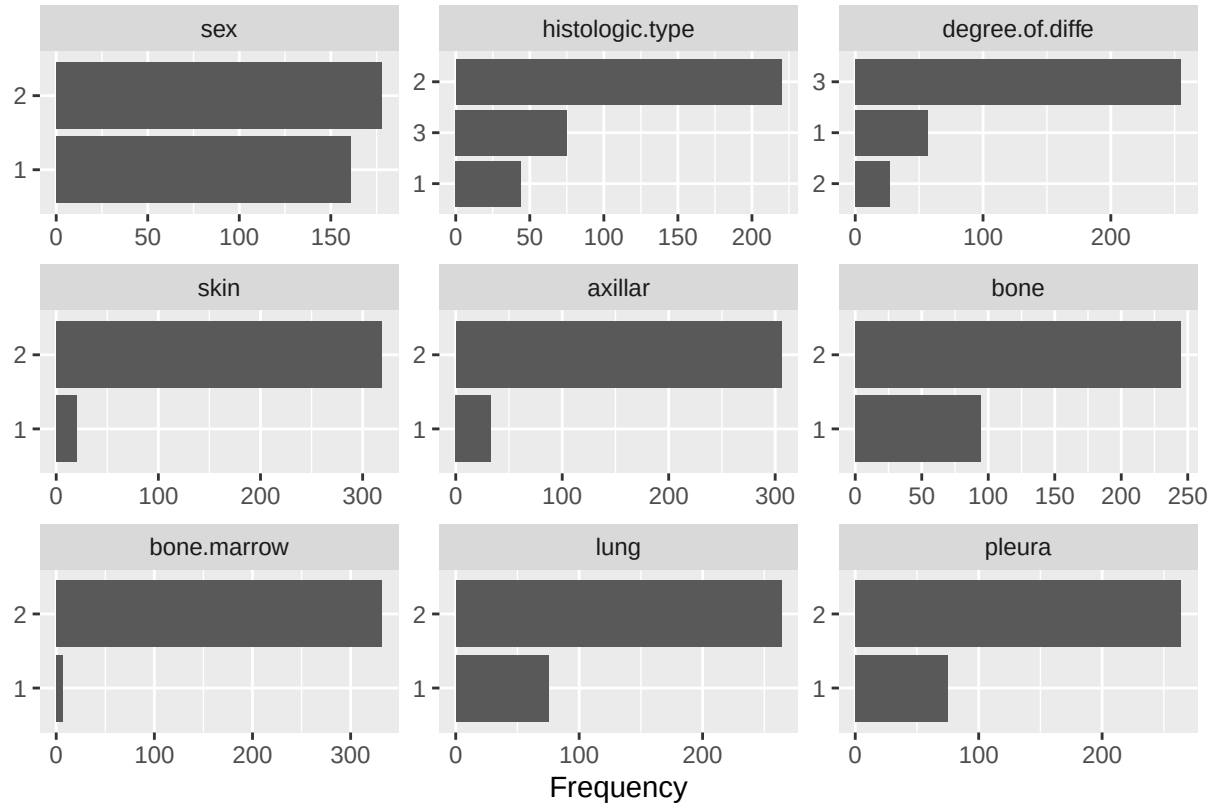
Data Cleaning

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
class	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
age	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
sex	1	1	2	2	2	2	1	1	1	1	1	1	1	1	1
histologic.type	3	3	2	3	3	3	1	1	1	1	1	1	1	1	1
degree.of.diffe	3	3	3	3	3	3	1	1	1	2	3	3	3	3	3
bone	2	2	1	1	1	1	1	1	2	1	1	1	1	1	2
bone.marrow	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
lung	1	2	2	1	1	2	2	2	2	2	2	2	2	2	2
pleura	2	2	2	1	1	2	2	2	2	2	1	2	2	2	2
peritoneum	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
liver	2	1	2	2	2	2	2	2	2	2	2	2	2	2	1
brain	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2
skin	2	2	2	2	2	2	1	2	2	2	2	1	2	2	2
neck	2	2	2	2	2	2	1	2	1	1	2	2	2	2	2

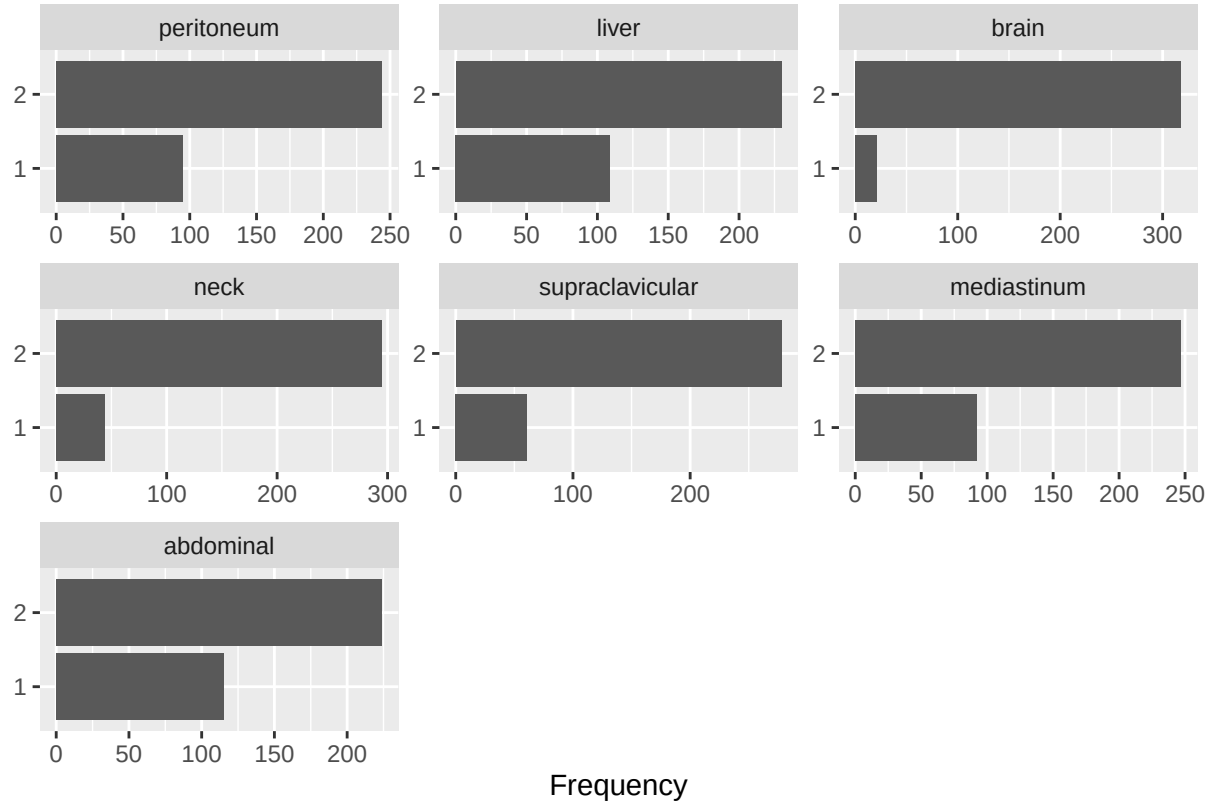
supraclavicular	2	1	2	2	2	1	1	2	2	1	1	2	2	2	2
axillar	2	2	2	2	2	1	2	2	2	2	2	1	2	2	2
mediastinum	2	1	1	1	1	1	2	2	2	2	1	2	2	2	1
abdominal	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

Table 1: First 15 Rows of the Tumor Data

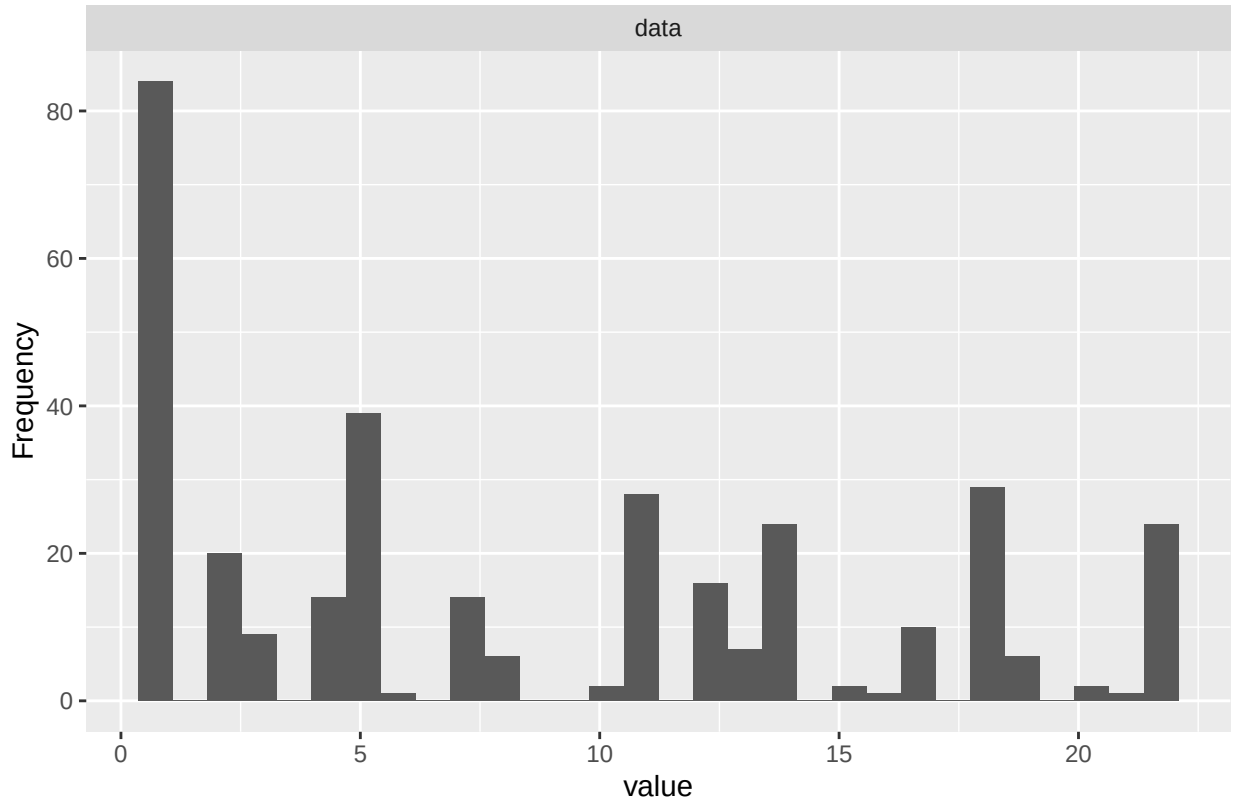
*** Frequency Distribution of the Variable Levels**



Page 1



Page 2



Variable Encoding

These categorical variables can't be use as is in a mathematical equation. They could be have converted or encode to numbers so they could be used an algorithms.The factor function in R could be used to do this.

Splitting the Model

The data set split into two sets which are `training` set and the `test` set. This step is necessary, as in order to evaluate the performance of the machine learning model, a separate data set from the training set is needed.

```
<TaskClassif:encoded_data> (339 x 25)
* Target: class
* Properties: multiclass
* Features (24):
  - int (24): abdominal, age, axillar_1, axillar_2, bone, bone.marrow,
    brain, degree.of.diffe_1, degree.of.diffe_2, degree.of.diffe_3,
    histologic.type_1, histologic.type_2, histologic.type_3, liver,
    lung, mediastinum, neck, peritoneum, pleura, sex_1, sex_2, skin_1,
    skin_2, supraclavicular
```

```
[1] 339  25
```

```
features <- c(Features = tsk_tumor$feature_names,
  Target = tsk_tumor$target_names)

library(knitr)
```

```
kable(features, caption = "Features and Target")
```

Table 5: Features and Target

	x
Features1	abdominal
Features2	age
Features3	axillar_1
Features4	axillar_2
Features5	bone
Features6	bone.marrow
Features7	brain
Features8	degree.of.diffe_1
Features9	degree.of.diffe_2
Features10	degree.of.diffe_3
Features11	histologic.type_1
Features12	histologic.type_2
Features13	histologic.type_3
Features14	liver
Features15	lung
Features16	mediastinum
Features17	neck
Features18	peritoneum
Features19	pleura
Features20	sex_1
Features21	sex_2
Features22	skin_1
Features23	skin_2
Features24	supraclavicular
Target	class

```
tail(tsk_tumor$row_ids)  # last 6 rows ID
```

```
[1] 334 335 336 337 338 339
```

```
library(xtable)

# retrieves all data
ln1 <- xtable(t(head(tsk_tumor$data(), 18)))

kable(ln1, caption = "Preprocessed Data")
```

Table 6: Preprocessed Data

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
class	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
abdominal	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
age	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2
axillar_1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
axillar_2	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1
bone	2	2	1	1	1	1	1	1	2	1	1	1	1	1	2	2	2	2
bone.marrow	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
brain	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2
degree.of.diffe_1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
degree.of.diffe_2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
degree.of.diffe_3	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	0
histologic.type_1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0
histologic.type_2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
histologic.type_3	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
liver	2	1	2	2	2	2	2	2	2	2	2	2	2	2	1	1	2	2
lung	1	2	2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	1
mediastinum	2	1	1	1	1	1	2	2	2	2	1	2	2	2	1	2	1	1
neck	2	2	2	2	2	2	1	2	1	1	2	2	2	2	2	2	2	1
peritoneum	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
pleura	2	2	2	1	1	2	2	2	2	2	1	2	2	2	2	1	2	2
sex_1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
sex_2	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
skin_1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
skin_2	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1
supraclavicular	2	1	2	2	2	1	1	2	2	1	1	2	2	2	2	2	1	1

Table 7: Preprocessed Data Distribution

	V1	V2	V3	V4	V5	V6	V7
class	1 : 84	5 : 39	18 : 29	11 : 28	14 : 24	22 : 24	(Other):111
abdominal	Min.	1st	Median	Mean	3rd	Max.	NA
	:1.000	Qu.:1.000	:2.000	:1.661	Qu.:2.000	:2.000	
age	Min.	1st	Median	Mean	3rd	Max.	NA
	:1.000	Qu.:2.000	:2.000	:2.248	Qu.:3.000	:3.000	
axillar_1	Min.	1st	Median	Mean	3rd	Max.	NA
	:0.00000	Qu.:0.00000	:0.00000	:0.09735	Qu.:0.00000	:1.00000	
axillar_2	Min.	1st	Median	Mean	3rd	Max.	NA
	:0.0000	Qu.:1.0000	:1.0000	:0.9027	Qu.:1.0000	:1.0000	
bone	Min.	1st	Median	Mean	3rd	Max.	NA
	:1.000	Qu.:1.000	:2.000	:1.723	Qu.:2.000	:2.000	
bone.marrow	Min.	1st	Median	Mean	3rd	Max.	NA
	:1.000	Qu.:2.000	:2.000	:1.979	Qu.:2.000	:2.000	
brain	Min.	1st	Median	Mean	3rd	Max.	NA
	:1.000	Qu.:2.000	:2.000	:1.938	Qu.:2.000	:2.000	
degree.of.diffe_1	Min.	1st	Median	Mean	3rd	Max.	NA
	:0.0000	Qu.:0.0000	:0.0000	:0.1681	Qu.:0.0000	:1.0000	
degree.of.diffe_2	Min.	1st	Median	Mean	3rd	Max.	NA
	:0.00000	Qu.:0.00000	:0.00000	:0.07965	Qu.:0.00000	:1.00000	
degree.of.diffe_3	Min.	1st	Median	Mean	3rd	Max.	NA
	:0.0000	Qu.:1.0000	:1.0000	:0.7522	Qu.:1.0000	:1.0000	
histologic.type_1	Min.	1st	Median	Mean	3rd	Max.	NA
	:0.0000	Qu.:0.0000	:0.0000	:0.1298	Qu.:0.0000	:1.0000	
histologic.type_2	Min.	1st	Median	Mean	3rd	Max.	NA
	:0.000	Qu.:0.000	:1.000	:0.649	Qu.:1.000	:1.000	
histologic.type_3	Min.	1st	Median	Mean	3rd	Max.	NA
	:0.0000	Qu.:0.0000	:0.0000	:0.2212	Qu.:0.0000	:1.0000	
liver	Min.	1st	Median	Mean	3rd	Max.	NA
	:1.000	Qu.:1.000	:2.000	:1.678	Qu.:2.000	:2.000	

	V1	V2	V3	V4	V5	V6	V7
lung	Min. :1.000	1st Qu.:2.000	Median :2.000	Mean :1.779	3rd Qu.:2.000	Max. :2.000	NA
mediastinum	Min. :1.000	1st Qu.:1.000	Median :2.000	Mean :1.729	3rd Qu.:2.000	Max. :2.000	NA
neck	Min. :1.00	1st Qu.:2.00	Median :2.00	Mean :1.87	3rd Qu.:2.00	Max. :2.00	NA
peritoneum	Min. :1.00	1st Qu.:1.00	Median :2.00	Mean :1.72	3rd Qu.:2.00	Max. :2.00	NA
pleura	Min. :1.000	1st Qu.:2.000	Median :2.000	Mean :1.779	3rd Qu.:2.000	Max. :2.000	NA
sex_1	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.4749	3rd Qu.:1.0000	Max. :1.0000	NA
sex_2	Min. :0.0000	1st Qu.:0.0000	Median :1.0000	Mean :0.5251	3rd Qu.:1.0000	Max. :1.0000	NA
skin_1	Min. :0.000	1st Qu.:0.000	Median :0.000	Mean :0.059	3rd Qu.:0.000	Max. :1.000	NA
skin_2	Min. :0.000	1st Qu.:1.000	Median :1.000	Mean :0.941	3rd Qu.:1.000	Max. :1.000	NA
supraclavicular	Min. :1.00	1st Qu.:2.00	Median :2.00	Mean :1.82	3rd Qu.:2.00	Max. :2.00	NA

```
#Makes prediction on new data
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.3.2
```

```
library(knitr)

prediction = lrn_rpart$predict(tsk_tumor, row_ids = split$test)

prediction
```

```
<PredictionClassif> for 113 observations:
    row_ids truth response
          1     1        1
         10     1        2
         16     1        1
---
        333    22       22
        336    22       22
        337    22        1
```

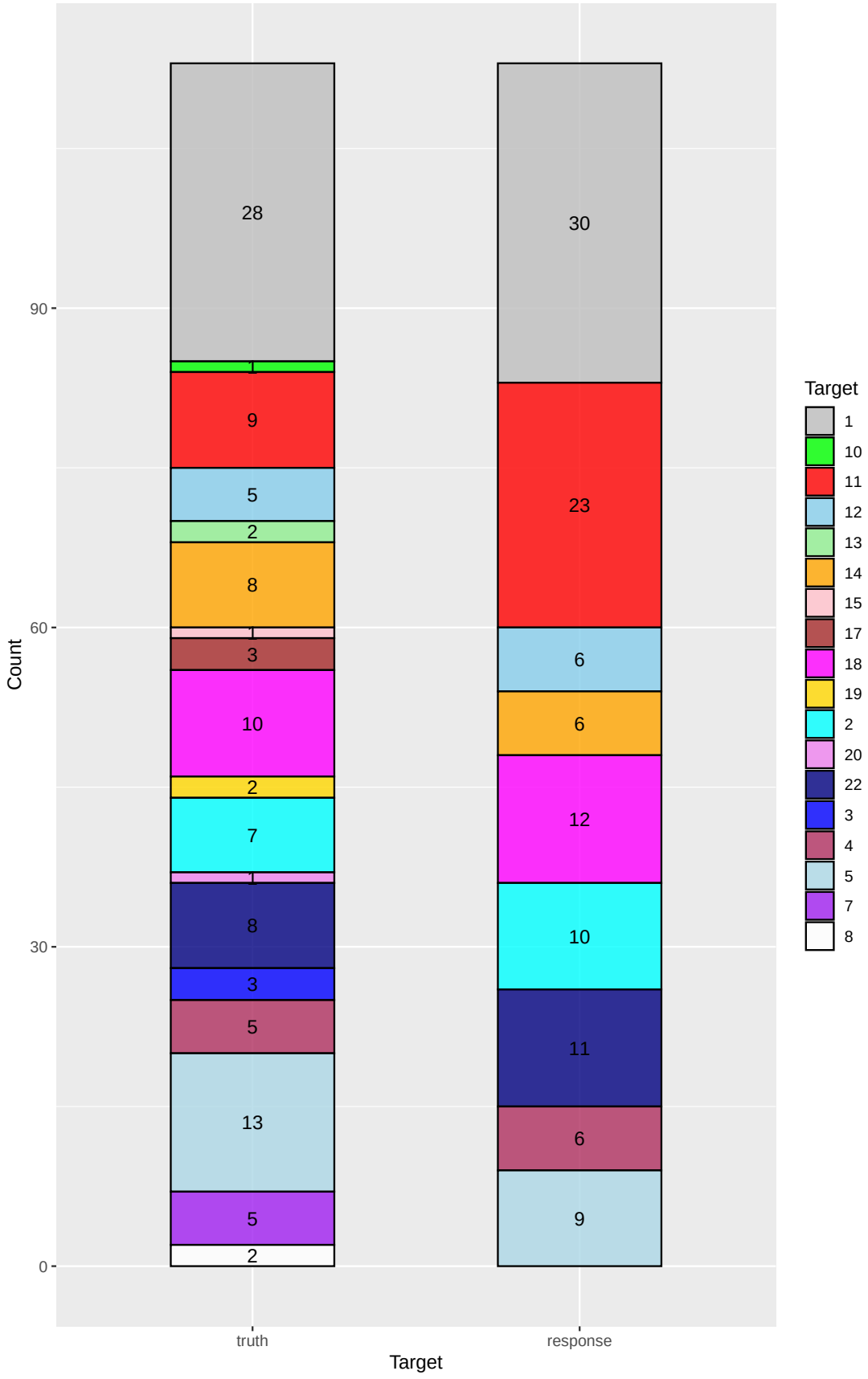
```
prediction$response[1:8]
```

```
[1] 1  2  1  14 22 5  5  22
Levels: 1 2 3 4 5 6 7 8 10 11 12 13 14 15 16 17 18 19 20 21 22
```

```
# 'maxdepth = 13' predicts 100% of feature 6
library(ggplot2)
library(mlr3viz)
prediction = lrn_rpart$predict(tsk_tumor, split$test)
autoplot(prediction, type = "stacked", theme = theme_grey()) +
  scale_fill_manual(values = c("gray", "green", "red", "skyblue", "lightgreen", "orange", "pink", "brow
```

```r
  labs(x = "Target", y = "Count")  # Set x-axis and y-axis labels
```

```
Scale for fill is already present.
Adding another scale for fill, which will replace the existing scale.
```


Count
90
60
30
0
28
1
9
5
2
8
1
3
10
2
7
1
8
3
5
13
5
2
30
23
6
6
12
10
11
6
9
truth
response
Target
Target
1
10
11
12
13
14
15
17
18
19
2
20
22
3
4
5
7
8

Measures

This is to measure the performance of the implemented machine learning code.

```
library(mlr3)
set.seed(544)
# load and partition our task
tsk_tumor = as_task_classif(encoded_data, target = "class")
splits = partition(tsk_tumor)
# load featureless learner
lrn_featureless = lrn("classif.featureless")
# load decision tree and set hyperparameters
lrn_rpart = lrn("classif.rpart", cp = 0.2, maxdepth = 5)
# load accuracy measure

measure = msr("classif.acc")
# train learners
lrn_featureless$train(tsk_tumor, splits$train)
lrn_rpart$train(tsk_tumor, splits$train)
# make and score predictions
lrn_featureless$predict(tsk_tumor, splits$test)$score(measure)
```

```
classif.acc
  0.2477876
```