Cheat sheet for Classification

Importing the dataset

The data and metadata need to be imported separately

```
dataset = read_delim('ST000007.csv',
                     delim = ",",
                     col_names = TRUE,
                     trim_ws = TRUE)
metadata = read_delim('metadata.csv',
                     delim = ",",
                     col_names = TRUE,
                     trim_ws = TRUE)
```

Data preprocessing

The samples can be gathered from columns to rows

```
dataset <- gather(dataset, key = … , value = … , …:…)
```

And the features can be spread from rows to columns

```
dataset <- spread(dataset, … , … )
```

The metadata also need preprocessing, R does not automatically understand that the sample id's are text

```
metadata <- metadata %>%
  mutate(local_sample_id = as.character(local_sample_id))
```

The dataset and metadata can be joined together with `left_join`

```
dataset <- dataset %>%
  left_join(metadata)
```

Remove unnecessary columns and rename if needed!

GLM need that the groups are numeric.

```
dataset <- dataset %>%
  mutate('Xoo infection' = case_when(
#discussion: which group should get which number?
    dataset$'Xoo infection' == "XXX" ~ 1,
    …,
  ))
```

Encoding the target feature as factor

```
dataset <- dataset %>%
  mutate('Xoo infection' = factor(dataset$`Xoo infection`))%>%
  select('Xoo infection', everything())
```

Data cleaning: remove variables that do not change from sample to sample and the ones that are strongly correlated

Feature Scaling:

```
… = scale(…)
```

Splitting the dataset into the training and test set

Fitting classification model

Select the suitable classification models with help from `caret`