SC1_Proj

Alessio

13 January 2020

Description of the dataset and problem

TODO: describe

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

Dataset Preprocessing Visualisation and Exploration

```
colSums(is.na(data))
```

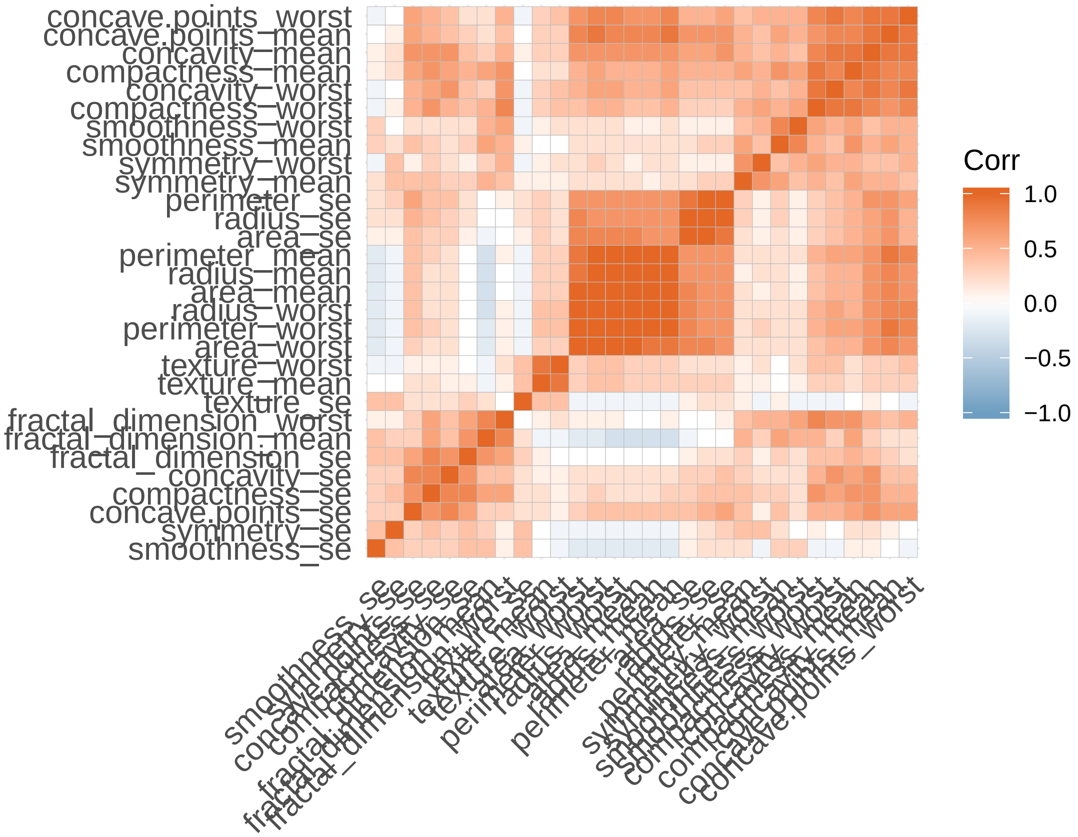
```
##                      id               diagnosis             radius_mean
##                       0                       0                       0
##            texture_mean          perimeter_mean               area_mean
##                       0                       0                     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

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

```
sum(is.na(data))
```

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

```
corr <- data[,-1] %>%
          cor() %>%
           round(1)
ggcorrplot(corr,
           hc.order = TRUE,
           colors = c("#6D9EC1", "white", "#E46726"),
           ggtheme = ggplot2::theme_minimal)
```



Dimensionality Reduction and Feature Selection

- Correlation Feature Selection
- PCA
- tSNE
- LDA

Classification

- Naive Bayes
- SVM
- Logistic Regression
- Lasso

Conclusion

- Evaluation of results
- Discuss outliers