Stan Ana-Maria Bianca CEN 4.S2B

# Machine learning – project

## Wine quality

**Dataset**: Wine Quality | Kaggle

**Load dataset**:

```
filename <- "wines.csv"
data <- read.csv(filename)
```

**Load libraries**:

```
library(ggplot2)
library(mlbench)
library(caret)
library(lattice)
library(e1071)
library(corrplot)
library(correlation)
library(randomForest)
library(rpart)
library(rpart.plot)
```

**Load data**

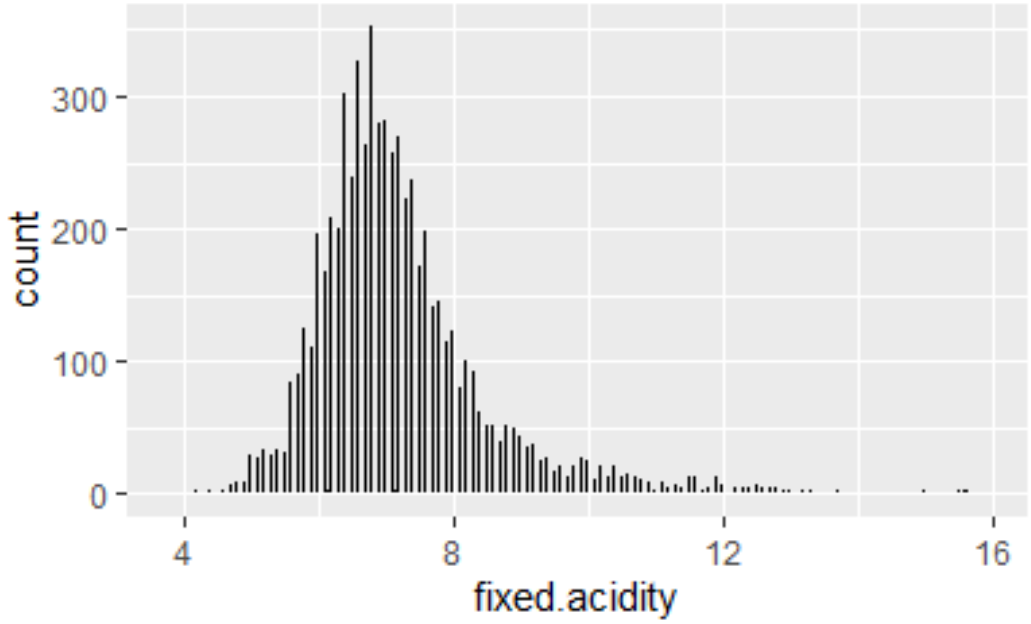
```
filename <- "wines.csv"
data <- read.csv(filename)
```

**Data visualizations**

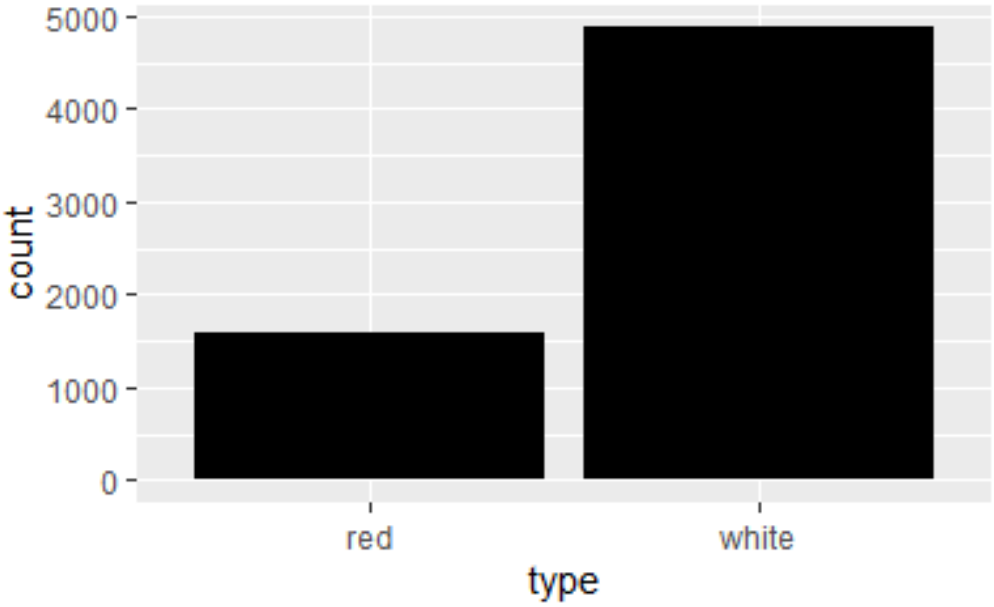
**Univariate Visualization:**

**Bar plot:**

ggplot(data, aes(fixed.acidity)) + geom_bar(fill="black")

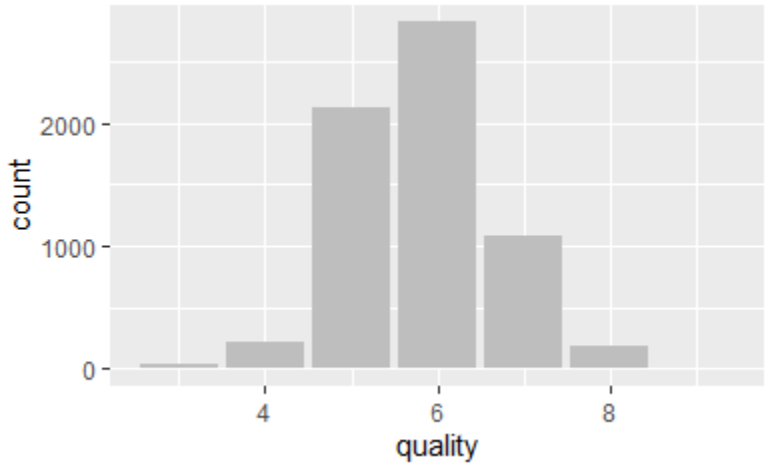



ggplot(data, aes(type)) + geom_bar(fill="black")

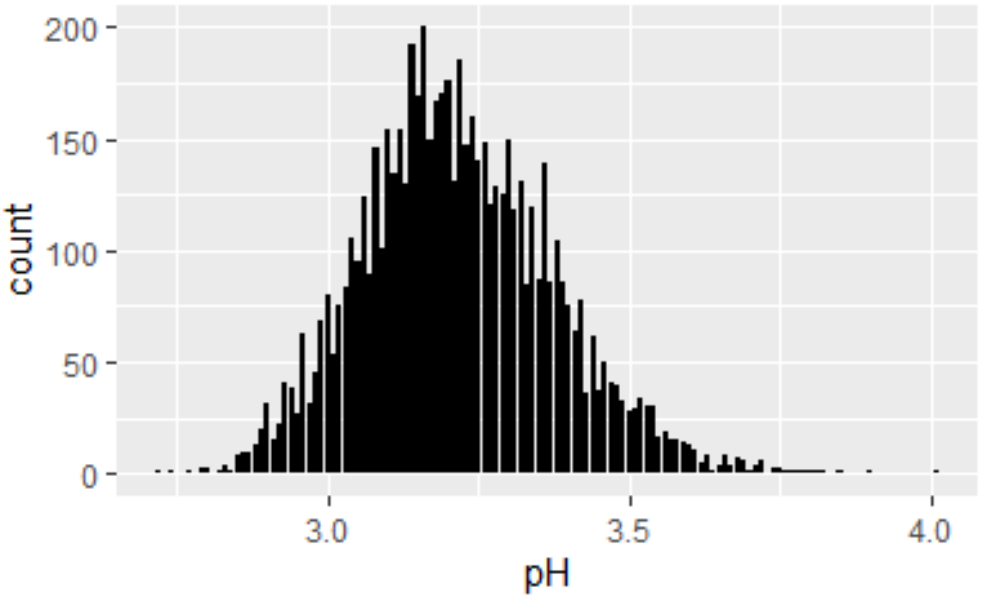



ggplot(data = data) +

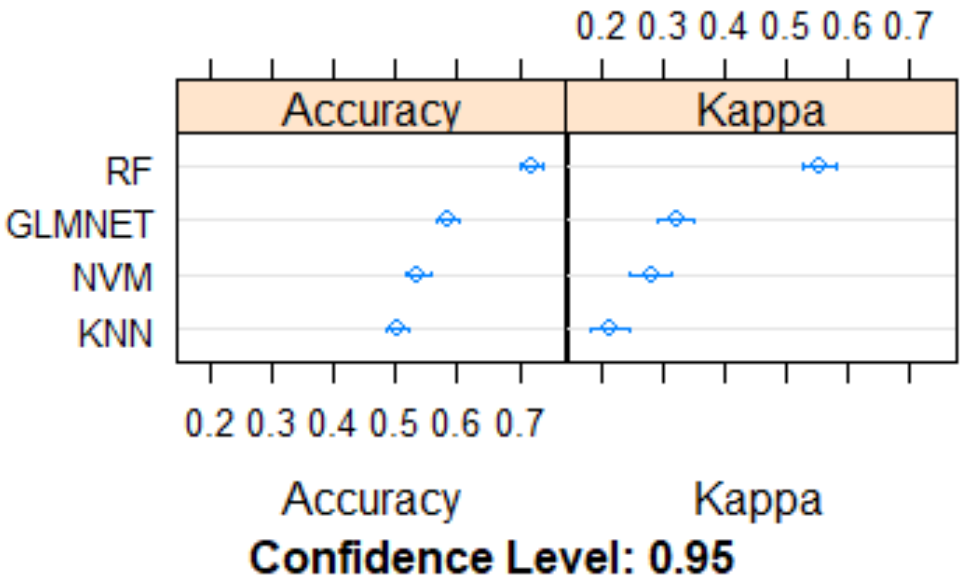
geom_bar(mapping = aes(x = quality), fill="gray")




```
ggplot(data, aes(pH)) + geom_bar(fill="black")
```




**Choose the best algorithm:**




## Predictions. Confusion matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction bad good normal
    bad    347    4    108
    good     8  159     49
    normal 108   98    412

Overall Statistics

               Accuracy : 0.71
                 95% CI : (0.6844, 0.7346)
    No Information Rate : 0.4401
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5388

 Mcnemar's Test P-Value : 0.0005153

Statistics by Class:

                     Class: bad Class: good Class: normal
Sensitivity              0.7495      0.6092        0.7241
Specificity              0.8651      0.9448        0.7155
Pos Pred Value           0.7560      0.7361        0.6667
Neg Pred Value           0.8609      0.9053        0.7674
Prevalence               0.3581      0.2019        0.4401
Detection Rate           0.2684      0.1230        0.3186
Detection Prevalence     0.3550      0.1671        0.4780
Balanced Accuracy        0.8073      0.7770        0.7198
>
```