Chandnee and Jada STAT632 Project details

Jada Smith

4/5/2022

Vino, wine at it's finest

About the Data

Soruce

https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

```
# Libraries we anticipate to use in the project.
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

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

```
library(performance)
library(see)
library(patchwork)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:patchwork':
##
##     area
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(rpart)
```

File

```
vino <- read.csv("wine-quality-white-and-red.csv")
head(vino)
```

```
##    type fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 white           7.0             0.27        0.36           20.7     0.045
```

```
## 2 white           6.3             0.30        0.34            1.6     0.049
## 3 white           8.1             0.28        0.40            6.9     0.050
## 4 white           7.2             0.23        0.32            8.5     0.058
## 5 white           7.2             0.23        0.32            8.5     0.058
## 6 white           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   quality
## 1       6
## 2       6
## 3       6
## 4       6
## 5       6
## 6       6
```

This data set is related to red variants of the Portuguese "Vinho Verde" wine.The dataset describes the amount of various chemicals present in wine and their effect on it's quality. The data set can be viewed as classification or regression tasks.

Our goal, Objective

With this data set, we are looking to understand the data set & cleanup where required. We are also looking to build classification models to predict the wine quality. We will look at only white wines to better narrow our data set.

```
dim(vino)
```

```
## [1] 6497   13
```

```
table(vino$type)
```

```
##
##   red white
##  1599  4898
```

```
# vino_white <- vino %>%
#  filter(type == "white") %>%
#  drop_na()

# dim(vino_white)
# View(vino_white)
```

This data frame contains the following:

Input variables(11) - Predictors (based on physicochemical tests): 1 - fixed acidity 2 - volatile acidity 3 - citric acid 4 - residual sugar 5 - chlorides 6 - free sulfur dioxide 7 - total sulfur dioxide 8 - density 9 - pH 10 - sulphates 11 - alcohol

Dummy variable(1) - Type 0 or 1 - Red or White

Output variable(1) - Response (based on sensory data): 13 - quality (score between 0 and 10)