Lab02

Amanda Mentesana

2025-02-07

Firstly, set up libraries and read dataset.

```
knitr::opts_chunk$set(echo = FALSE)

#install libraries
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.4.2
```

```
library(EnvStats)
```

```
## Warning: package 'EnvStats' was built under R version 4.4.2
```

```
##
## Attaching package: 'EnvStats'
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.2
```

```
library(ggfortify)
```

```
## Warning: package 'ggfortify' was built under R version 4.4.3
```

```
library(class)
```

```
## Warning: package 'class' was built under R version 4.4.2
```

```
#read the wine data set
wine <- read_csv("C:/Users/amanda/Downloads/wine/wine.data")
```

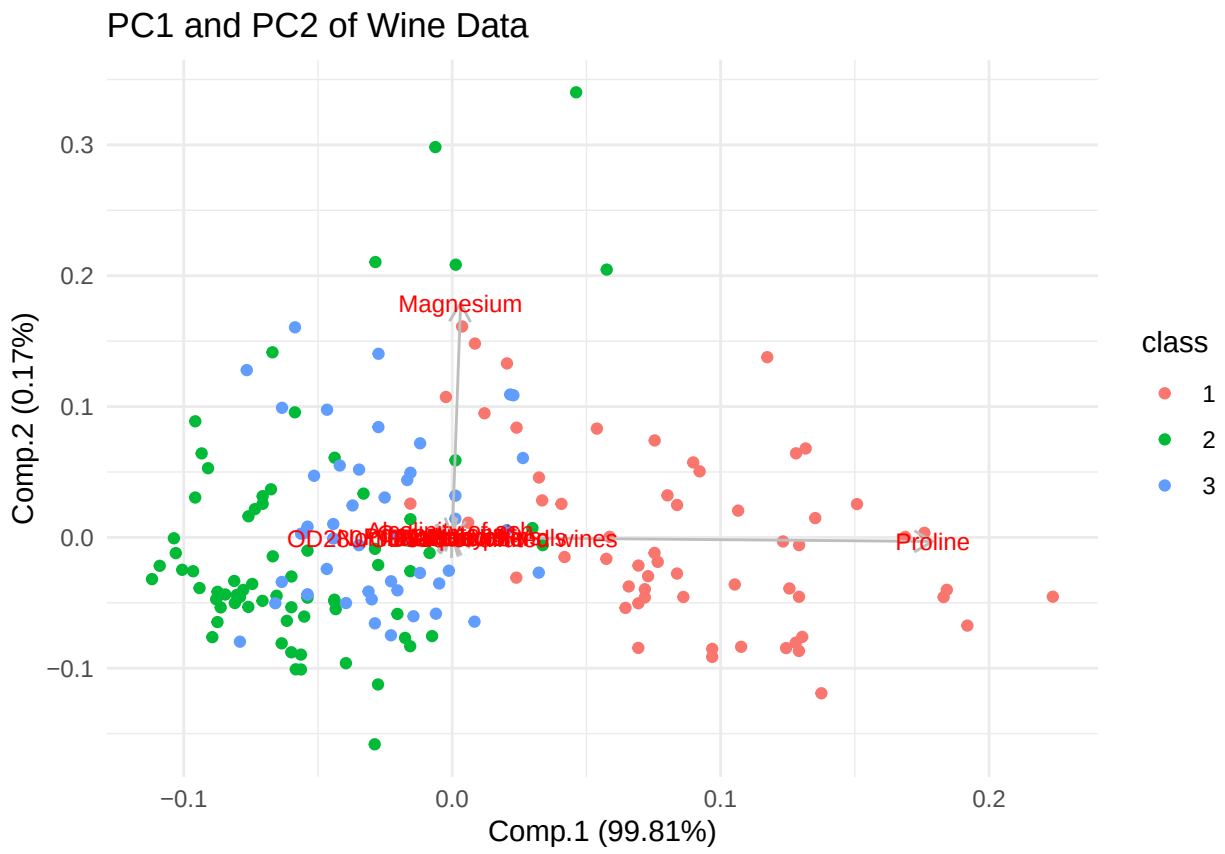
```
## Rows: 177 Columns: 14
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (14): 1, 14.23, 1.71, 2.43, 15.6, 127, 2.8, 3.06, .28, 2.29, 5.64, 1.04,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
colnames(wine) <- c("class","Alcohol","Malic acid","Ash","Alcalinity of ash","Magnesium","Total phenols
```

1. Compute the PCs and plot the dataset using the 1st and 2nd PC.

```
##      class           Alcohol        Malic acid         Ash
##  Min.   :1.000   Min.   :11.03   Min.   :0.74   Min.   :1.360
##  1st Qu.:1.000   1st Qu.:12.36   1st Qu.:1.60   1st Qu.:2.210
##  Median :2.000   Median :13.05   Median :1.87   Median :2.360
##  Mean   :1.944   Mean   :12.99   Mean   :2.34   Mean   :2.366
##  3rd Qu.:3.000   3rd Qu.:13.67   3rd Qu.:3.10   3rd Qu.:2.560
##  Max.   :3.000   Max.   :14.83   Max.   :5.80   Max.   :3.230
##  Alcalinity of ash   Magnesium      Total phenols     Flavanoids
##  Min.   :10.60     Min.   : 70.00   Min.   :0.980   Min.   :0.340
##  1st Qu.:17.20     1st Qu.: 88.00   1st Qu.:1.740   1st Qu.:1.200
##  Median :19.50     Median : 98.00   Median :2.350   Median :2.130
##  Mean   :19.52     Mean   : 99.59   Mean   :2.292   Mean   :2.023
##  3rd Qu.:21.50     3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.860
##  Max.   :30.00     Max.   :162.00   Max.   :3.880   Max.   :5.080
##  Nonflavanoid phenols Proanthocyanins Color intensity       Hue
##  Min.   :0.1300       Min.   :0.410   Min.   : 1.280   Min.   :0.480
##  1st Qu.:0.2700       1st Qu.:1.250   1st Qu.: 3.210   1st Qu.:0.780
##  Median :0.3400       Median :1.550   Median : 4.680   Median :0.960
##  Mean   :0.3623       Mean   :1.587   Mean   : 5.055   Mean   :0.957
##  3rd Qu.:0.4400       3rd Qu.:1.950   3rd Qu.: 6.200   3rd Qu.:1.120
##  Max.   :0.6600       Max.   :3.580   Max.   :13.000   Max.   :1.710
##  OD280/OD315 of diluted wines    Proline
##  Min.   :1.270                Min.   : 278.0
##  1st Qu.:1.930                1st Qu.: 500.0
##  Median :2.780                Median : 672.0
##  Mean   :2.604                Mean   : 745.1
##  3rd Qu.:3.170                3rd Qu.: 985.0
##  Max.   :4.000                Max.   :1680.0
```



2. Identify the variables that contribute the most to the 1st PC.

```
## Importance of components:
##                             Comp.1       Comp.2       Comp.3       Comp.4
## Standard deviation     314.0465241 13.034437573 3.062882e+00 2.234012e+00
## Proportion of Variance   0.9981074  0.001719388 9.494015e-05 5.050804e-05
## Cumulative Proportion    0.9981074  0.999826814 9.999218e-01 9.999723e-01
##                              Comp.5       Comp.6       Comp.7       Comp.8
## Standard deviation     1.107336e+00 9.160683e-01 5.260813e-01 3.887933e-01
## Proportion of Variance 1.240932e-05 8.492685e-06 2.800883e-06 1.529773e-06
## Cumulative Proportion  9.999847e-01 9.999932e-01 9.999960e-01 9.999975e-01
##                              Comp.9      Comp.10      Comp.11      Comp.12
## Standard deviation     3.303978e-01 2.676655e-01 1.937198e-01 1.451319e-01
## Proportion of Variance 1.104749e-06 7.250605e-07 3.797847e-07 2.131645e-07
## Cumulative Proportion  9.999986e-01 9.999993e-01 9.999997e-01 9.999999e-01
##                             Comp.13
## Standard deviation     9.035657e-02
## Proportion of Variance 8.262448e-08
## Cumulative Proportion  1.000000e+00

##                      Alcohol                   Malic acid
##                0.0016464031                -0.0006735032
##                          Ash            Alcalinity of ash
##                0.0001948773                -0.0046271444
```

```
##                   Magnesium                Total phenols
##                0.0174715429                0.0009863499
##                  Flavanoids         Nonflavanoid phenols
##                0.0015575348                -0.0001223031
##              Proanthocyanins              Color intensity
##                0.0005912858                0.0023300597
##                         Hue OD280/OD315 of diluted wines
##                0.0001708674                0.0006850453
##                     Proline
##                0.9998302063
```

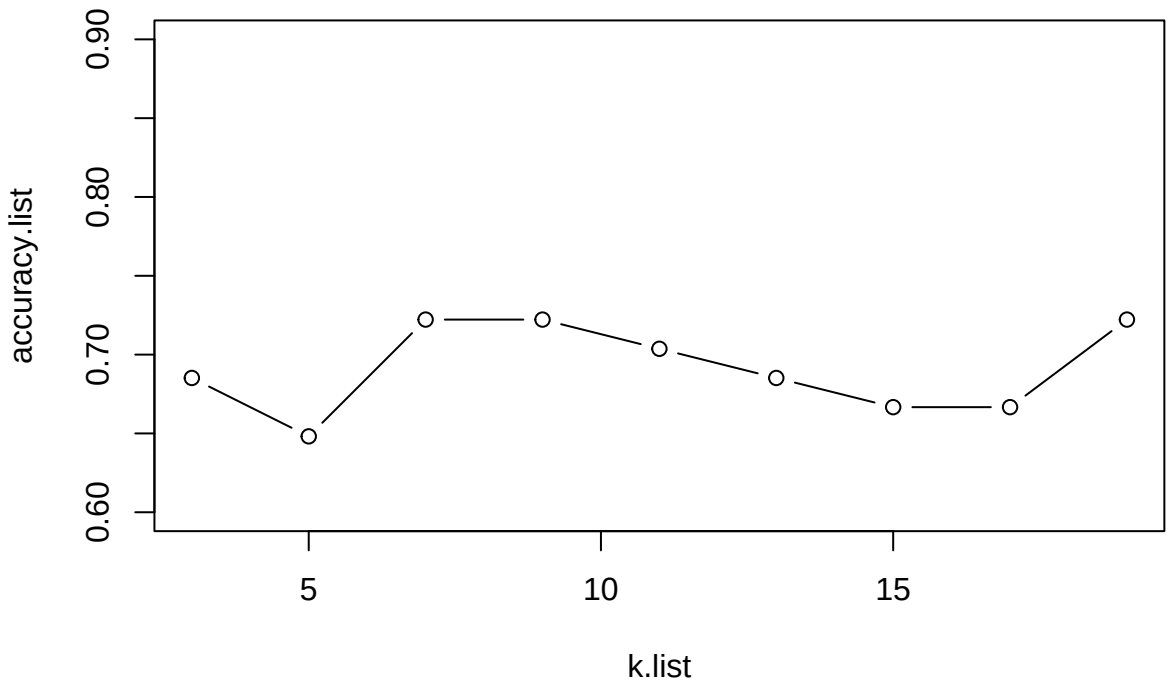
```
##                     Proline                   Magnesium
##                0.9998302063                0.0174715429
##            Alcalinity of ash              Color intensity
##                0.0046271444                0.0023300597
##                     Alcohol                  Flavanoids
##                0.0016464031                0.0015575348
##               Total phenols OD280/OD315 of diluted wines
##                0.0009863499                0.0006850453
##                  Malic acid              Proanthocyanins
##                0.0006735032                0.0005912858
##                          Ash                         Hue
##                0.0001948773                0.0001708674
##         Nonflavanoid phenols
##                0.0001223031
```

3. Drop the variables least contributing to the 1st PC and rerun PCA.

```
##         Nonflavanoid phenols                         Hue
##                0.0001223031                0.0001708674
##                          Ash              Proanthocyanins
##                0.0001948773                0.0005912858
##                  Malic acid OD280/OD315 of diluted wines
##                0.0006735032                0.0006850453
##               Total phenols                  Flavanoids
##                0.0009863499                0.0015575348
##                     Alcohol              Color intensity
##                0.0016464031                0.0023300597
##            Alcalinity of ash                   Magnesium
##                0.0046271444                0.0174715429
##                     Proline
##                0.9998302063
```

4. Train a classifier model (e.g. kNN) to predict wine type using the original dataset.

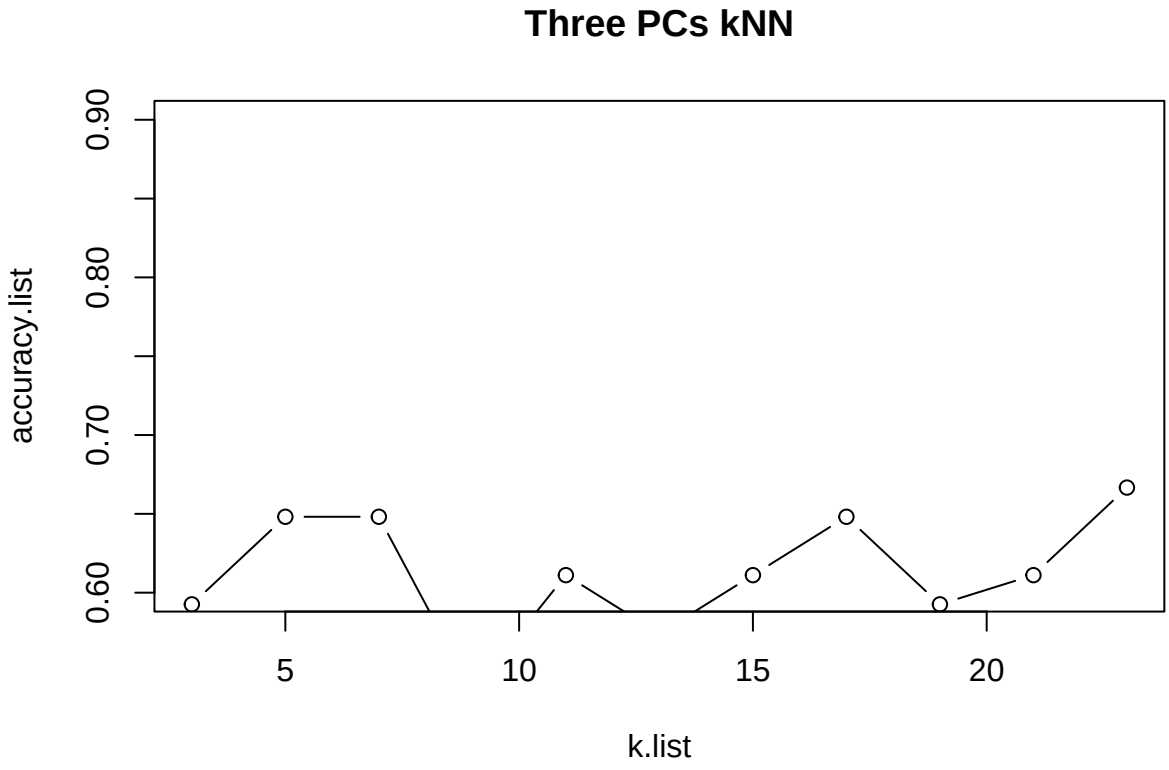
```
## [1] 12
```




```
## [1] 0.6851852 0.6481481 0.7222222 0.7222222 0.7037037 0.6851852 0.6666667
## [8] 0.6666667 0.7222222

## k is maximum at  7

##           actual
## predicted  1  2  3
##         1 16  1  2
##         2  0 15 13
##         3  0  1  6

## [1] 0.6851852

##       Predicted
## Actual  1  2  3
##      1 16  0  0
##      2  1 15  1
##      3  2 13  6

## [1] 0.6851852

##   wine.recall wine.precision   wine.f1
## 1   1.0000000      0.8421053 0.9142857
## 2   0.8823529      0.5357143 0.6666667
## 3   0.2857143      0.8571429 0.4285714
```

5. Train a classifier model to predict wine type using the data projected into the first 3 PCs (scores), from PCA model where lowest PCs are dropped.

```
## [1] 12
```




```
##  [1] 0.5925926 0.6481481 0.6481481 0.5370370 0.6111111 0.5740741 0.6111111
##  [8] 0.6481481 0.5925926 0.6111111 0.6666667

## k is maximum at  23

##           actual
## predicted  1  2  3
##         1 16  1  2
##         2  1 12  8
##         3  0  8  6

## [1] 0.6296296

##       Predicted
## Actual  1  2  3
##      1 16  1  0
##      2  1 12  8
##      3  2  8  6
```

```
## [1] 0.6296296
```

```
##   three.recall three.precision  three.f1
## 1    0.9411765       0.8421053 0.8888889
## 2    0.5714286       0.5714286 0.5714286
## 3    0.3750000       0.4285714 0.4000000
```

6. Compare the 2 classification models using contingency tables and prevision/recall/f1 metrics

```
##           actual
## predicted  1  2  3
##         1 16  1  2
##         2  0 15 13
##         3  0  1  6
```

```
## [1] 0.6851852
```

```
##           actual
## predicted  1  2  3
##         1 16  1  2
##         2  1 12  8
##         3  0  8  6
```

```
## [1] 0.6296296
```

```
##   wine.recall three.recall wine.precision three.precision   wine.f1  three.f1
## 1   1.0000000    0.9411765      0.8421053       0.8421053 0.9142857 0.8888889
## 2   0.8823529    0.5714286      0.5357143       0.5714286 0.6666667 0.5714286
## 3   0.2857143    0.3750000      0.8571429       0.4285714 0.4285714 0.4000000
```