R Decision Tree for ELPAC Data

Team 3 - Emma Oo, Luke Awino, Oscar Gil

11/13/2022

```
# R Libraries
library(caret)
library(AppliedPredictiveModeling)
#library(Hmisc)
library(dplyr)
library(tidyverse)
library(ggplot2)
library(corrplot)
library(MASS)
library(ISLR)
library(rpart)
library(partykit)
library(randomForestSRC)
library(earth)
library(MARSS)
library(e1071)
library(summarytools)
library(grid)
library(MLeval)
library(pROC)
```

Load the ELPAC data set from GitHub

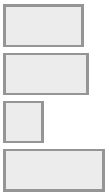
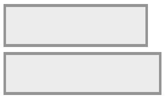
```
df <- read.csv(
  url("https://raw.githubusercontent.com/OscarG-DataSci/ADS-599B/main/Data%20Folder/new_elpac.csv")
                      , header = TRUE)
```

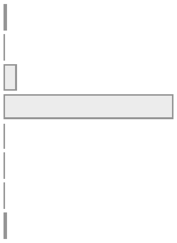
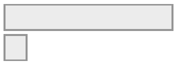
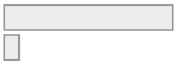
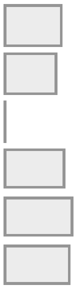
Data Summary

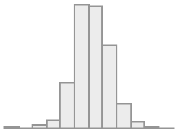
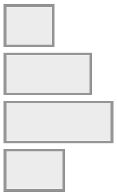
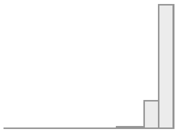
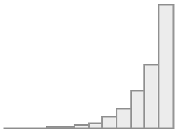
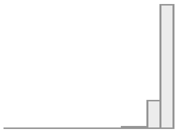
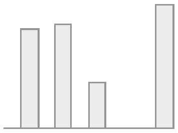
```
# use the view function to view in R Studio
#view(
dfSummary(df,
          plain.ascii  = FALSE,
          style        = "grid",
          graph.magnif = 0.75,
          valid.col    = FALSE,
          tmp.img.dir  = "NA")
```

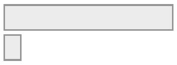
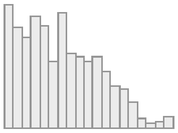
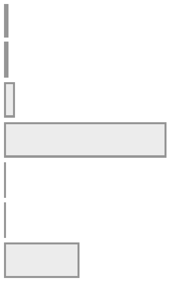
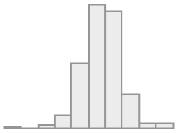
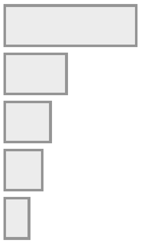
Data Frame Summary

df **Dimensions:** 4314 x 19
Duplicates: 3

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
1	School__deID [integer]	Mean (sd) : 4.2 (2.7) min < med < max: 0 < 4 < 9 IQR (CV) : 4 (0.6)	0 : 422 (9.8%) 1 : 418 (9.7%) 2 : 565 (13.1%) 3 : 477 (11.1%) 4 : 489 (11.3%) 5 : 429 (9.9%) 6 : 460 (10.7%) 7 : 463 (10.7%) 8 : 252 (5.8%) 9 : 339 (7.9%)		0 (0.0%)
2	GradeLevel [integer]	Mean (sd) : 1.9 (1.6) min < med < max: 0 < 1 < 4 IQR (CV) : 4 (0.9)	0 : 1144 (26.5%) 1 : 1201 (27.8%) 2 : 538 (12.5%) 4 : 1431 (33.2%)		0 (0.0%)
3	StudentGender [integer]	Min : 0 Mean : 0.5 Max : 1	0 : 2066 (47.9%) 1 : 2248 (52.1%)		0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
4	StudentEthnicity [integer]	Mean (sd) : 2.9 (0.6) min < med < max: 0 < 3 < 7 IQR (CV) : 0 (0.2)	0 : 55 (1.3%) 1 : 9 (0.2%) 2 : 278 (6.4%) 3 : 3891 (90.2%) 4 : 17 (0.4%) 5 : 22 (0.5%) 6 : 5 (0.1%) 7 : 37 (0.9%)		0 (0.0%)
5	Special_Education [integer]	Min : 0 Mean : 0.1 Max : 1	0 : 3820 (88.5%) 1 : 494 (11.5%)		0 (0.0%)
6	Homeless [integer]	Min : 0 Mean : 0.1 Max : 1	0 : 3944 (91.4%) 1 : 370 (8.6%)		0 (0.0%)
7	SocioEconomically [integer]	Min : 0 Mean : 0.8 Max : 1	0 : 785 (18.2%) 1 : 3529 (81.8%)		0 (0.0%)
8	TestDayName [integer]	Mean (sd) : 2.8 (1.8) min < med < max: 0 < 3 < 5 IQR (CV) : 3 (0.7)	0 : 799 (18.5%) 1 : 744 (17.2%) 2 : 4 (0.1%) 3 : 864 (20.0%) 4 : 967 (22.4%) 5 : 936 (21.7%)		0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
9	OverallScore [integer]	Mean (sd) : 1461.9 (66.1) min < med < max: 1150 < 1462 < 1731 IQR (CV) : 83 (0)	349 distinct values		0 (0.0%)
10	OverallLevel [integer]	Mean (sd) : 2.6 (1) min < med < max: 1 < 3 < 4 IQR (CV) : 1 (0.4)	1 : 712 (16.5%) 2 : 1228 (28.5%) 3 : 1537 (35.6%) 4 : 837 (19.4%)		0 (0.0%)
11	ExpectedAttendanceDays [numeric]	Mean (sd) : 176.9 (7.8) min < med < max: 69 < 180 < 180 IQR (CV) : 0 (0)	63 distinct values		0 (0.0%)
12	DaysAttended [numeric]	Mean (sd) : 163.8 (16.5) min < med < max: 62 < 169 < 180 IQR (CV) : 18 (0.1)	100 distinct values		0 (0.0%)
13	EnrolledPct [numeric]	Mean (sd) : 1 (0) min < med < max: 0.4 < 1 < 1 IQR (CV) : 0 (0)	63 distinct values		0 (0.0%)
14	GradeAttendedPct [numeric]	Mean (sd) : 2.8 (1.6) min < med < max: 0.4 < 2 < 5 IQR (CV) : 3.9 (0.6)	660 distinct values		0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
15	TeacherGender [integer]	Min : 0 Mean : 0.1 Max : 1	0 : 3931 (91.1%) 1 : 383 (8.9%)		0 (0.0%)
16	TeacherTotalYearsOfService [integer]	Mean (sd) : 13.3 (8.6) min < med < max: 1 < 13 < 38 IQR (CV) : 13 (0.6)	37 distinct values		0 (0.0%)
17	TeacherEthnicity [integer]	Mean (sd) : 3.8 (1.5) min < med < max: 0 < 3 < 6 IQR (CV) : 3 (0.4)	0 : 58 (1.3%) 1 : 35 (0.8%) 2 : 145 (3.4%) 3 : 2785 (64.6%) 4 : 6 (0.1%) 5 : 8 (0.2%) 6 : 1277 (29.6%)		0 (0.0%)
18	OverallScoreStd [numeric]	Mean (sd) : 0.6 (0.1) min < med < max: 0 < 0.6 < 1 IQR (CV) : 0.1 (0.2)	730 distinct values		0 (0.0%)
19	TotalAssessments [integer]	Mean (sd) : 2.2 (1.3) min < med < max: 1 < 2 < 5 IQR (CV) : 2 (0.6)	1 : 1887 (43.7%) 2 : 912 (21.1%) 3 : 647 (15.0%) 4 : 532 (12.3%) 5 : 336 (7.8%)		0 (0.0%)

```
#     )
```

Decision Tree

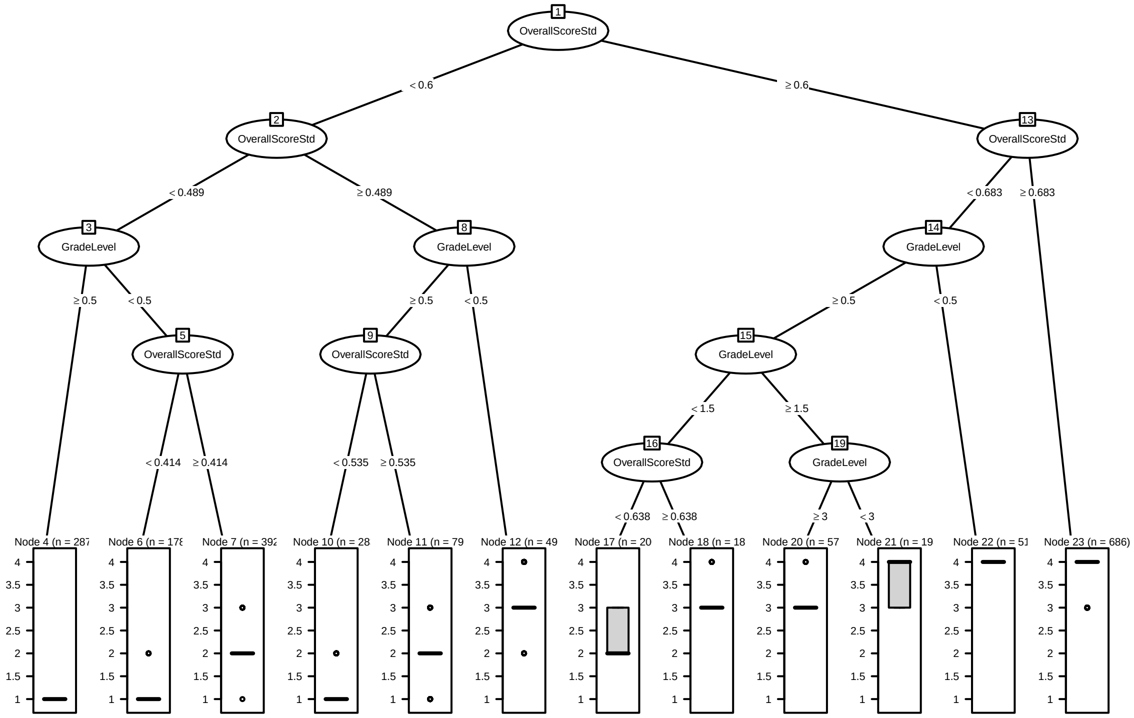
```r
# get column names and their number
colnames(df)
```

```
##  [1] "School_deID"                 "GradeLevel"
##  [3] "StudentGender"               "StudentEthnicity"
##  [5] "Special_Education"           "Homeless"
##  [7] "SocioEconomically"           "TestDayName"
##  [9] "OverallScore"                "OverallLevel"
## [11] "ExpectedAttendanceDays"      "DaysAttended"
## [13] "EnrolledPct"                 "GradeAttendedPct"
## [15] "TeacherGender"               "TeacherTotalYearsOfService"
## [17] "TeacherEthnicity"            "OverallScoreStd"
## [19] "TotalAssessments"
```

```r
#subset, remove unnecessary columns
df2 <- df[-c(9, 11, 13, 14)]

# Begin model...
rPartTree <- rpart(OverallLevel ~ ., data = df2)

rpartTree2 <- as.party(rPartTree)

# Results
rpartTree2
```

```
##
## Model formula:
## OverallLevel ~ School_deID + GradeLevel + StudentGender + StudentEthnicity +
##     Special_Education + Homeless + SocioEconomically + TestDayName +
##     DaysAttended + TeacherGender + TeacherTotalYearsOfService +
##     TeacherEthnicity + OverallScoreStd + TotalAssessments
##
## Fitted party:
## [1] root
## |   [2] OverallScoreStd < 0.59979
## |   |   [3] OverallScoreStd < 0.48895
## |   |   |   [4] GradeLevel >= 0.5: 1.000 (n = 287, err = 0.0)
## |   |   |   [5] GradeLevel < 0.5
## |   |   |   |   [6] OverallScoreStd < 0.41364: 1.051 (n = 178, err = 8.5)
## |   |   |   |   [7] OverallScoreStd >= 0.41364: 2.043 (n = 392, err = 18.3)
## |   |   [8] OverallScoreStd >= 0.48895
## |   |   |   [9] GradeLevel >= 0.5
## |   |   |   |   [10] OverallScoreStd < 0.53494: 1.230 (n = 283, err = 50.1)
## |   |   |   |   [11] OverallScoreStd >= 0.53494: 2.121 (n = 796, err = 158.4)
## |   |   |   [12] GradeLevel < 0.5: 3.132 (n = 491, err = 102.4)
## |   [13] OverallScoreStd >= 0.59979
## |   |   [14] OverallScoreStd < 0.68337
## |   |   |   [15] GradeLevel >= 0.5
## |   |   |   |   [16] GradeLevel < 1.5
## |   |   |   |   |   [17] OverallScoreStd < 0.63836: 2.340 (n = 200, err = 44.9)
## |   |   |   |   |   [18] OverallScoreStd >= 0.63836: 3.171 (n = 181, err = 25.7)
## |   |   |   |   [19] GradeLevel >= 1.5
## |   |   |   |   |   [20] GradeLevel >= 3: 3.016 (n = 571, err = 8.9)
## |   |   |   |   |   [21] GradeLevel < 3: 3.566 (n = 198, err = 48.6)
## |   |   |   [22] GradeLevel < 0.5: 4.000 (n = 51, err = 0.0)
## |   |   [23] OverallScoreStd >= 0.68337: 3.796 (n = 686, err = 111.4)
##
## Number of inner nodes:    11
## Number of terminal nodes: 12
```

```
plot(rpartTree2, gp = gpar(fontsize=4))
```



““