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

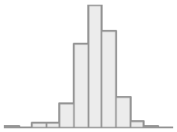
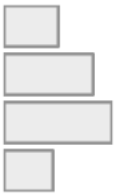
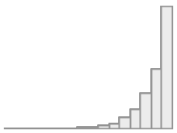
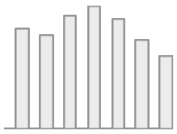
```
# use the view function to view in R Studio
#view(
dfSummary(df,
          plain.ascii  = FALSE,
          style        = "grid",
          graph.magnif = 0.75,
          valid.col    = FALSE,
          tmp.img.dir  = "NA")
```

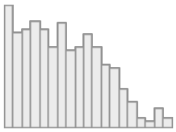
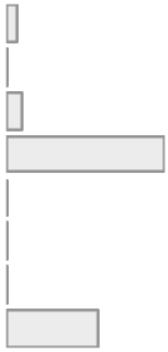
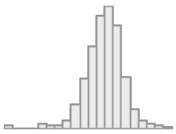
Data Frame Summary

df Dimensions: 9460 x 18
Duplicates: 7

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
1	School_deID [integer]	Mean (sd) : 4.2 (2.8) min < med < max: 0 < 4 < 9 IQR (CV) : 5 (0.7)	0 : 1002 (10.6%) 1 : 975 (10.3%) 2 : 1185 (12.5%) 3 : 1069 (11.3%) 4 : 995 (10.5%) 5 : 868 (9.2%) 6 : 979 (10.3%) 7 : 994 (10.5%) 8 : 573 (6.1%) 9 : 820 (8.7%)		0 (0.0%)
2	GradeLevel [integer]	Mean (sd) : 2.9 (1.9) min < med < max: 0 < 3 < 6 IQR (CV) : 3 (0.7)	0 : 1357 (14.3%) 1 : 1253 (13.2%) 2 : 1532 (16.2%) 3 : 1641 (17.3%) 4 : 1492 (15.8%) 5 : 1208 (12.8%) 6 : 977 (10.3%)		0 (0.0%)
3	StudentGender [integer]	Min : 0 Mean : 0.5 Max : 1	0 : 4502 (47.6%) 1 : 4958 (52.4%)		0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
4	StudentEthnicity [integer]	Mean (sd) : 3.9 (0.6) min < med < max: 0 < 4 < 8 IQR (CV) : 0 (0.2)	0 : 1 (0.0%) 1 : 130 (1.4%) 2 : 25 (0.3%) 3 : 595 (6.3%) 4 : 8544 (90.3%) 5 : 32 (0.3%) 6 : 40 (0.4%) 7 : 13 (0.1%) 8 : 80 (0.8%)		0 (0.0%)
5	Special_Education [integer]	Min : 0 Mean : 0.1 Max : 1	0 : 8234 (87.0%) 1 : 1226 (13.0%)		0 (0.0%)
6	Homeless [integer]	Min : 0 Mean : 0.1 Max : 1	0 : 8669 (91.6%) 1 : 791 (8.4%)		0 (0.0%)
7	SocioEconomically [integer]	Min : 0 Mean : 0.8 Max : 1	0 : 1664 (17.6%) 1 : 7796 (82.4%)		0 (0.0%)
8	TestDayName [integer]	Mean (sd) : 3.4 (2.3) min < med < max: 0 < 4 < 6 IQR (CV) : 4 (0.7)	0 : 1884 (19.9%) 1 : 1564 (16.5%) 2 : 23 (0.2%) 3 : 12 (0.1%) 4 : 1883 (19.9%) 5 : 1936 (20.5%) 6 : 2158 (22.8%)		0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
9	OverallScore [integer]	Mean (sd) : 1476.2 (65.5) min < med < max: 1150 < 1479 < 1731 IQR (CV) : 80 (0)	407 distinct values		0 (0.0%)
10	OverallLevel [integer]	Mean (sd) : 2.5 (1) min < med < max: 1 < 3 < 4 IQR (CV) : 1 (0.4)	1 : 1687 (17.8%) 2 : 2781 (29.4%) 3 : 3431 (36.3%) 4 : 1561 (16.5%)		0 (0.0%)
11	ExpectedAttendanceDays [numeric]	Mean (sd) : 176.7 (7.6) min < med < max: 69 < 180 < 180 IQR (CV) : 3.2 (0)	71 distinct values		0 (0.0%)
12	DaysAttended [numeric]	Mean (sd) : 164.2 (16.5) min < med < max: 20 < 170 < 180 IQR (CV) : 18 (0.1)	111 distinct values		0 (0.0%)
13	EnrolledPct [numeric]	Mean (sd) : 1 (0) min < med < max: 0.4 < 1 < 1 IQR (CV) : 0 (0)	71 distinct values		0 (0.0%)
14	GradeAttendedPct [numeric]	Mean (sd) : 3.8 (1.9) min < med < max: 0.4 < 3.9 < 7 IQR (CV) : 3 (0.5)	1345 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 : 8573 (90.6%) 1 : 887 (9.4%)		0 (0.0%)
16	TeacherTotalYearsOfService [integer]	Mean (sd) : 13.9 (8.8) min < med < max: 1 < 13 < 38 IQR (CV) : 13 (0.6)	38 distinct values		0 (0.0%)
17	TeacherEthnicity [integer]	Mean (sd) : 4.2 (2.1) min < med < max: 0 < 3 < 7 IQR (CV) : 4 (0.5)	0 : 332 (3.5%) 1 : 46 (0.5%) 2 : 516 (5.5%) 3 : 5387 (56.9%) 4 : 18 (0.2%) 5 : 8 (0.1%) 6 : 16 (0.2%) 7 : 3137 (33.2%)		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)	1264 distinct values		0 (0.0%)

```
#    )
```

Decision Tree

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

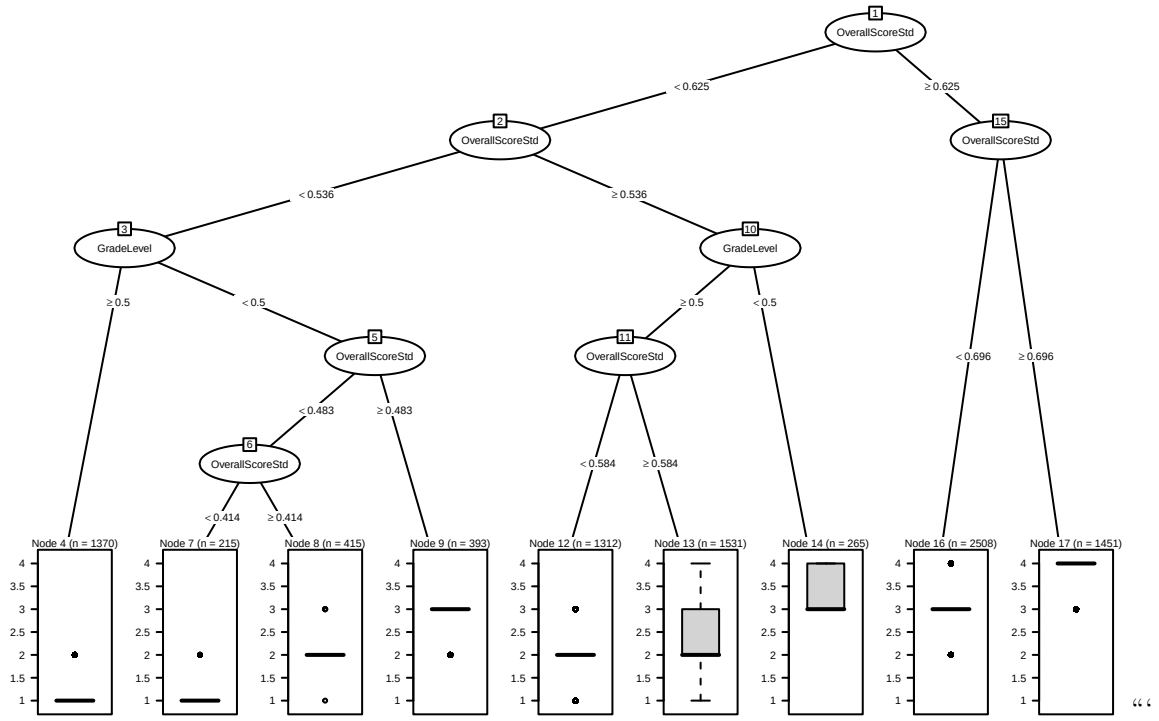
```
##  [1] "School_deID"                "GradeLevel"
##  [3] "StudentGender"              "StudentEthnicity"
##  [5] "Special_Education"          "Homeless"
```

```
##  [7] "SocioEconomically"          "TestDayName"
##  [9] "OverallScore"               "OverallLevel"
## [11] "ExpectedAttendanceDays"     "DaysAttended"
## [13] "EnrolledPct"                "GradeAttendedPct"
## [15] "TeacherGender"              "TeacherTotalYearsOfService"
## [17] "TeacherEthnicity"           "OverallScoreStd"
```

```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
##
## Fitted party:
## [1] root
## |   [2] OverallScoreStd < 0.62476
## |   |   [3] OverallScoreStd < 0.53588
## |   |   |   [4] GradeLevel >= 0.5: 1.135 (n = 1370, err = 160.0)
## |   |   |   [5] GradeLevel < 0.5
## |   |   |   |   [6] OverallScoreStd < 0.48273
## |   |   |   |   |   [7] OverallScoreStd < 0.41364: 1.065 (n = 215, err = 13.1)
## |   |   |   |   |   [8] OverallScoreStd >= 0.41364: 2.010 (n = 415, err = 6.0)
## |   |   |   |   [9] OverallScoreStd >= 0.48273: 2.855 (n = 393, err = 48.7)
## |   |   [10] OverallScoreStd >= 0.53588
## |   |   |   [11] GradeLevel >= 0.5
## |   |   |   |   [12] OverallScoreStd < 0.58356: 1.830 (n = 1312, err = 323.1)
## |   |   |   |   [13] OverallScoreStd >= 0.58356: 2.413 (n = 1531, err = 421.1)
## |   |   |   [14] GradeLevel < 0.5: 3.491 (n = 265, err = 66.2)
## |   [15] OverallScoreStd >= 0.62476
## |   |   [16] OverallScoreStd < 0.6962: 3.018 (n = 2508, err = 574.2)
## |   |   [17] OverallScoreStd >= 0.6962: 3.761 (n = 1451, err = 264.0)
##
## Number of inner nodes:    8
## Number of terminal nodes: 9
```

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


1
OverallScoreStd
< 0.625
≥ 0.625
2
OverallScoreStd
15
OverallScoreStd
< 0.536
≥ 0.536
3
GradeLevel
10
GradeLevel
≥ 0.5
< 0.5
≥ 0.5
< 0.5
5
OverallScoreStd
11
OverallScoreStd
< 0.696
≥ 0.696
< 0.483
≥ 0.483
6
OverallScoreStd
< 0.584
≥ 0.584
< 0.414
≥ 0.414
Node 4 (n = 1370)
Node 7 (n = 215)
Node 8 (n = 415)
Node 9 (n = 393)
Node 12 (n = 1312)
Node 13 (n = 1531)
Node 14 (n = 265)
Node 16 (n = 2508)
Node 17 (n = 1451)
4
3.5
3
2.5
2
1.5
1

“‘