The University of Texas at Dallas

School of Economic, Political and Policy Sciences

EPPS 6323 Knowledge Mining

Spring 2021

Dr. Karl Ho

Final Paper

Academic Analytics:
Predictions around Argentine "Aprender" National Evaluation

Student: Federico Ferrero

Academic Analytics:
Predictions around Argentine “Aprender” National Evaluation

Federico Ferrero

In Argentina, each year the National Evaluation Operation “Aprender” is carried out. This assessment is aimed at high school seniors and has the purpose to generate timely and quality information to better understand the achievements and pending challenges around students' learning (Aprender, 2019). The evaluation is developed by the Argentine Ministry of Education, Culture, Science and Technology, through the Secretariat for Educational Evaluation and collects data on knowledge of Mathematics, Language, and contextual information of the respondent students (sociodemographic variables, school climate, student self-perception, educational practices and use of technology, among other data).

In general terms, results are used for educational decision-making every year, also considering the construction of time series to conduct analyzes from a longitudinal point of view. These approximations tend to be predominantly descriptive and although the use of algorithms dedicated to the prediction of students' success is increasingly frequent at the global level (Jayaprakash, Moody, Lauría, Regan and Baron, 2014), such Machine Learning techniques are not usually applied to this specific evaluation.

If the use of these techniques is reviewed in other latitudes, it is observed that new evaluative developments have a place not only at the educational systemic level but also at the level of individual learning. Indeed, efforts to develop “Predictive Analytics” (Holmes, Bialik, and Fadel, 2019; Siegel, 2016; Williamson, 2016) are increasing with the objective of, for example, identifying students “at-risk” and tailoring, in this way, pedagogical interventions. In any case, there are long controversies around the accepted degree of granularity of the predictions: if it is convenient to make predictions on each individual or if it should be done on groups or institutions.

In this context, this study proposes not only an Exploratory Data Analysis but also a Predictive Analysis that use Machine Learning techniques with the purpose of finding the most accurate and adequate predictive hypotheses for the Argentine “Aprender” National Evaluation.

In other words, this Academic Analytics exercise contemplates the use of knowledge mining techniques that allow predicting student performance beyond the conventional approach of hypothesis testing inscribed in what Leo Breiman (2001) calls the “Data Modeling Culture”.

Following this author, the challenge is to be able to account for the nature of phenomenon according to the structure of data we have and not to implant a predefined model in the form of a template that is insensitive to the behavior of the data with which we work. In this line, Breiman argues that the data and the problem should guide the solutions and not the *a priori* adoption of techniques that only permit conclusions to be drawn on the model and not on the data. In his own words: “If all a man has is a hammer, then every problem looks like a nail. The trouble is that recently some of the problems have stopped looking like nails” (2001; 204).

Based on these discussions and our approximation to the “Algorithmic Modeling Culture”, it is expected that the results to be obtained in this exercise will offer considerable predictive precision in models that, in short, explain more exquisitely the educational phenomena here studied. In the longer term, this would also imply that teachers and administrators can take advantage of such results to better understand their students’ learning processes and, thus, be able to personalize pedagogical interventions based on the patterns and trends of success identified.

Data and Methodology

“Aprender” National Evaluation data is online available for the 2019 edition, the last year in which the operation was conducted due to the COVID-19 pandemic and school lockdowns. In that opportunity 343,751 high school senior students were assessed but, in this study, we will work with a particular jurisdiction, the province of Cordoba. To take that decision, computational power limitations to carry out the analyzes were considered. In this province, the second-biggest jurisdiction in the country, N = 34,191 students were evaluated and their data were anonymized to preserve the identity of the respondents.

The dependent variables are Language Performance (ldesemp) and Math Performance (mdesemp) and their values are presented in a classification of 4 categories: below basic level, basic level, satisfactory level, and advanced level.

A considerable number of other 246 independent variables include gender, sector (public or private school management), ambit (rural or urban), student socioeconomic situation, student cultural consumption, school climate, student self-perception, educational practices and use of technology, migration status, among others.

The analytical strategies were deployed in three different moments.

First, an Exploratory Data Analysis is carried out selecting classic pedagogical variables historically used to account for student performances. These variables include: sector, ambit, gender, repetition, student employment, and student socioeconomic level. This exploratory instance was supported by visualizations of the data constructed with R.

Second, using *regsubsets* with *leaps* package, the work focused on finding the best possible models. For Language Performance backward selection was used while for Math Performance forward selection was chosen. In this instance, two questions were answered: how many the optimal predictors in each case are and what these predictors are. AdjR2, BIC, and Cp were calculated to determine the number of covariates and then graphs were produced with *car* library to facilitate the interpretation of the outputs of the *regsubsets* functions.

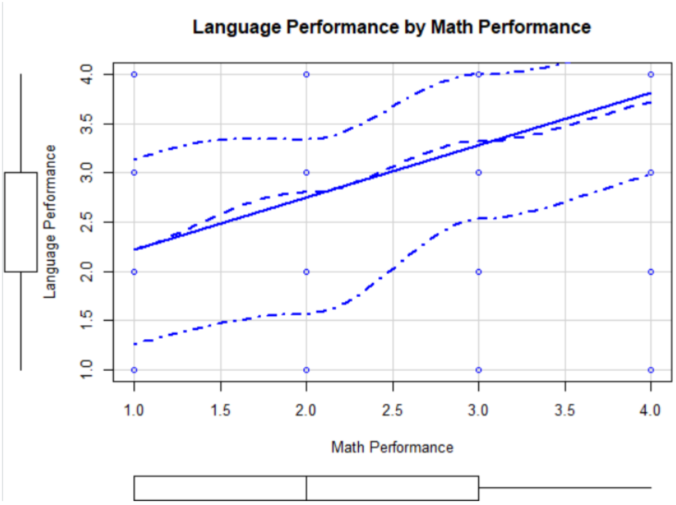
Finally, Supervised Learning techniques were applied. In the first place, multiple regressions were run with the predictors found relevant both in the case of Language and Math Performances. Subsequently, Tree-Based-methods were used with decision trees and conditional decision trees and then cross-validations were conducted. After the interpretation of performance prediction considering specific students' cases, via confusion matrixes and accuracy rates, the techniques were compared to identify the most powerful one.

Exploratory Data Analysis

Based on our initial explorations with traditional variables used to predict performance in educational settings, some findings can be listed.

First, Figure 1 shows that there is a positive and linear association between Language Performance and Math Performance.

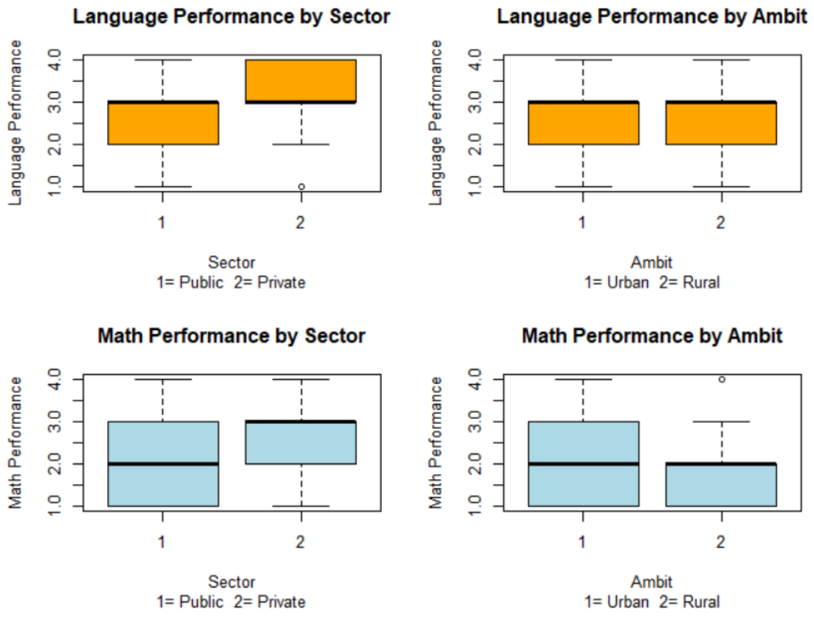
Figure 1



```
library("car")
scatterplot(ldesemp ~ mdesemp, data=mydata,
            xlab="Math Performance", ylab="Language Performance",
            main="Language Performance by Math Performance")
```

Second, when performances are analyzed according to Sector (public or private school management) and Ambit (either urban or rural), it is observed that in the case of Language scores there are better results at private schools and no differences between ambits: 50% of cases in private schools are between satisfactory and advanced while in the public system, 50% of observations are between basic and satisfactory. Also, according to the boxplots, it should be noted that data for Language are more dispersed in the public sector and less in the private type of school management.

Figure 2



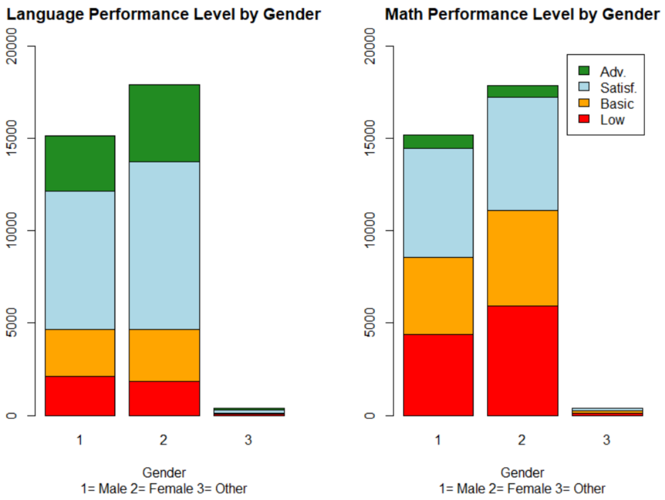
```
par(mfrow=c(2, 2))
boxplot(ldesemp~sector,data=mydata, main="Language Performance by Sector", sub="1= Pub
lic  2= Private",
        xlab="Sector", ylab="Language Performance", col="orange")

boxplot(ldesemp~ambito,data=mydata, main="Language Performance by Ambit", sub="1= Urba
n  2= Rural",
        xlab="Ambit", ylab="Language Performance", col="orange")
```

```
boxplot(mdesemp~sector,data=mydata, main="Math Performance by Sector", sub="1= Public
2= Private",
        xlab="Sector", ylab="Math Performance", col="lightblue")


boxplot(mdesemp~ambito,data=mydata, main="Math Performance by Ambit", sub="1= Urban  2
= Rural",
        xlab="Ambit", ylab="Math Performance", col="lightblue")
```

Regarding Mathematics scores, there is evidence of better performance in private and urban schools: 50% of cases in private schools are between basic and satisfactory while in the public system 50% of students are between below-basic and satisfactory. Likewise, Math performance in rural ambit seems to be significantly worse than in urban schools. The existence of an outlier value should be also noted in the case of the rural ambit for Mathematics: this presence will require further examinations and perhaps treatments to balance the data.

In Figure 3, the stacked bar graphs show absolute frequencies of students according to their performance by gender. On the one hand, in terms of Language scores, it is reported a better performance in women (39% at least obtain the satisfactory level while 31% of men achieve this level). Both groups have 13% of students in basic and low level. On the other hand, Math scores show worse performance in women (32% obtain low and basic levels against 25% in the case of male students). Both last groups have 19% of students in satisfactory and advanced levels.

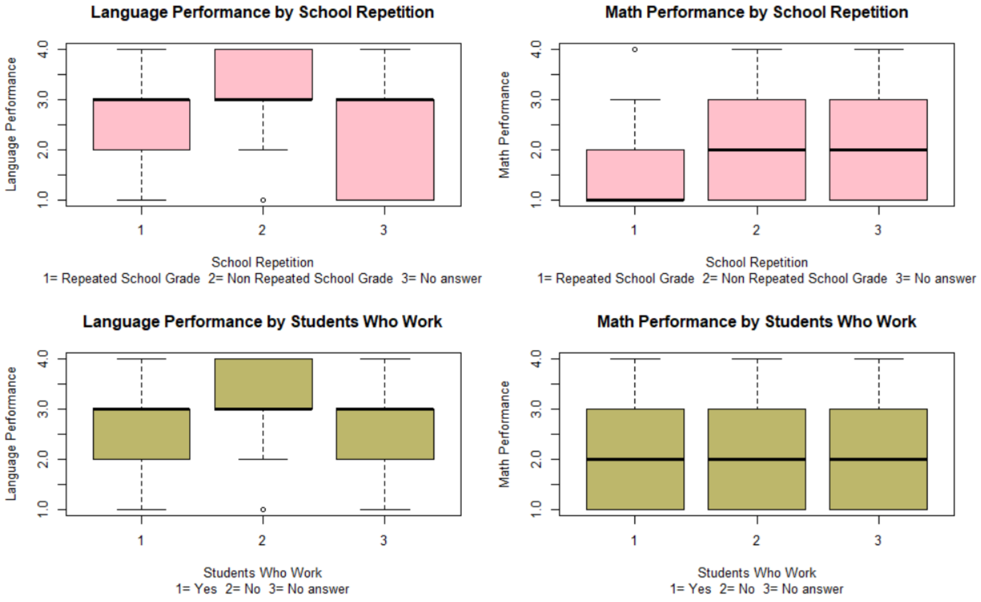
Figure 3



```
par(mfrow=c(1, 2))


counts1 <- table(mydata$ldesemp, mydata$gender)

barplot(counts1, main="Language Performance Level by Gender", sub="1= Male 2= Female 3
= Other",

        xlab="Gender", col=c("red","orange","lightblue","forestgreen"), ylim=c(0,20000
))




counts2 <- table(mydata$mdesemp, mydata$gender)

barplot(counts2, main="Math Performance Level by Gender", sub="1= Male 2= Female 3= Ot
her",

        xlab="Gender", col=c("red","orange","lightblue","forestgreen"),

        legend.text = c("Low", "Basic", "Satisf.", "Adv."), ylim=c(0,20000))
```

Figure 4 focuses on performances by school repetition and students' employment status. In the case of Language, a clear better performance is recorded in non-repeating students and students who do not work. Along the same lines, for Math scores, better performance is also observed in non-repeating students although there are no apparent differences when considered worker students. Anyway, it should be said that a considerable number of missing values in this variable may have skewed these results. As before, outliers values should be noticed and explored in depth.

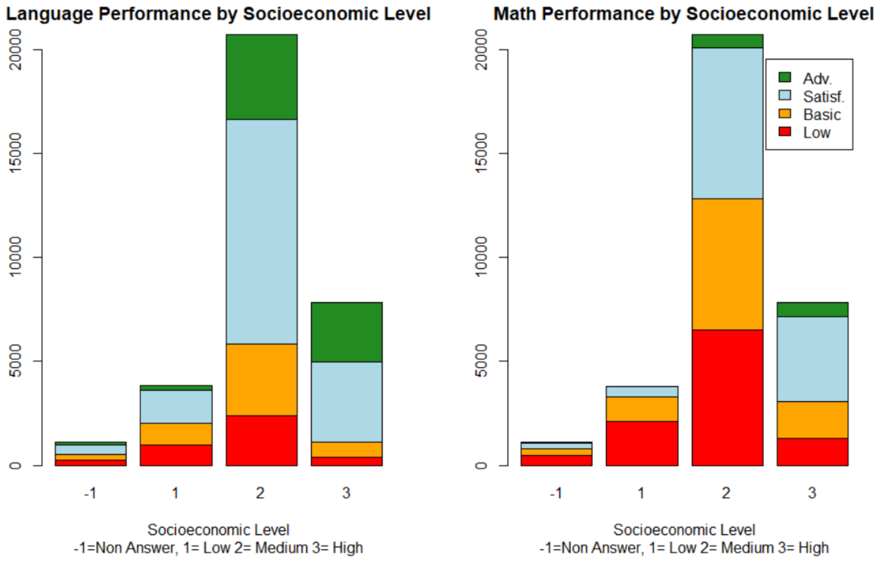
Figure 4



```r
par(mfrow=c(2, 2))

boxplot(ldesemp~repitencia_dicotomica,data=mydata, main="Language Performance by School Repetition", sub="1= Repeated School Grade  2= Non Repeated School Grade  3= No answer",

        xlab="School Repetition", ylab="Language Performance", col="pink")


boxplot(mdesemp~repitencia_dicotomica,data=mydata, main="Math Performance by School Repetition", sub="1= Repeated School Grade  2= Non Repeated School Grade  3= No answer",

        xlab="School Repetition", ylab="Math Performance", col="pink")


boxplot(ldesemp~trabaja_fuera_hogar,data=mydata, main="Language Performance by Students Who Work", sub="1= Yes  2= No  3= No answer",

        xlab="Students Who Work", ylab="Language Performance", col="darkkhaki")


boxplot(mdesemp~trabaja_fuera_hogar,data=mydata, main="Math Performance by Students Who Work", sub="1= Yes  2= No  3= No answer",

        xlab="Students Who Work", ylab="Math Performance", col="darkkhaki")
```

Finally, Figure 4 presents trends related to performances by students' socioeconomic level. When analyzing Language scores, we find predominantly middle socioeconomic level students who obtain at least a satisfactory level (44%). In the case of Math, results are not so favorable since middle socioeconomic level students who obtain low and basic levels in their performances prevail (37%).

Figure 4



```r
par(mfrow=c(1, 2))


counts1 <- table(mydata$ldesemp, mydata$isocioa)

barplot(counts1, main="Language Performance by Socioeconomic Level", sub="-1=Non Answer, 1= Low 2= Medium 3= High",

        xlab="Socioeconomic Level", col=c("red","orange","lightblue","forestgreen"), ylim=c(0,20000))




counts2 <- table(mydata$mdesemp, mydata$isocioa)

barplot(counts2, main="Math Performance by Socioeconomic Level", sub="-1=Non Answer, 1= Low 2= Medium 3= High",

        xlab="Socioeconomic Level", col=c("red","orange","lightblue","forestgreen"),

        legend.text = c("Low", "Basic", "Satisf.", "Adv."), ylim=c(0,20000))
```

Finding the Best Models

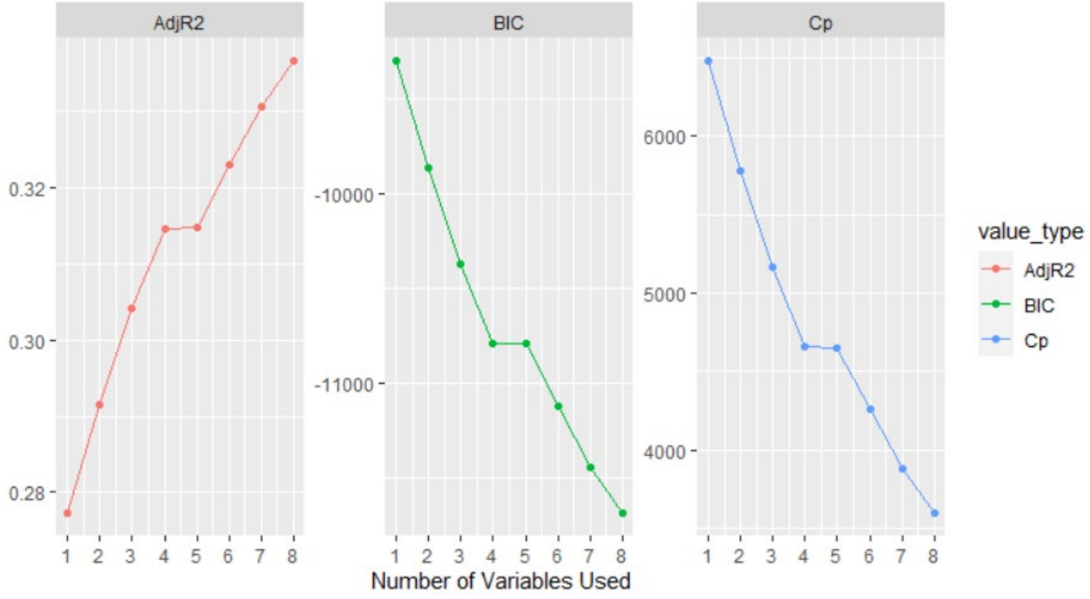
Beyond these preliminary analyzes oriented by the use of classical variables in the field of educational knowledge, the next question that arose was for the most precise hypotheses when predicting performances. In this framework, backward algorithm was applied to identify the best predictors of Language Performance and forward selection in the case of Math Performance.

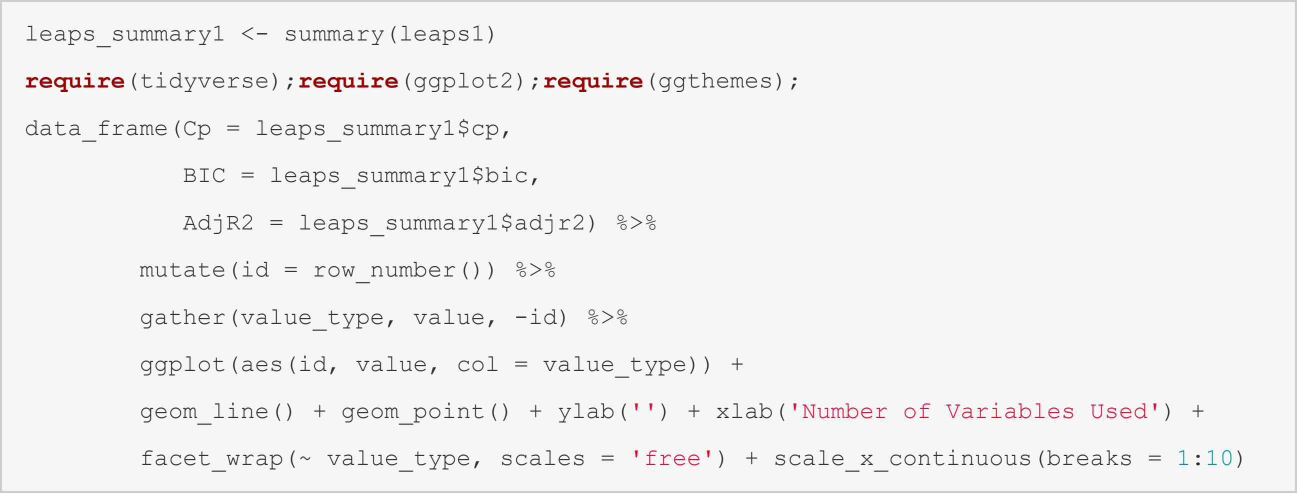
```r
library(leaps)

leaps1<- regsubsets(ldesemp ~., data= mydata, nbest=1, method = "backward")

leaps2<- regsubsets(mdesemp ~., data= mydata, nbest=1, method = "forward")
```

Later, we focus on the identification of the optimal number of predictors. In the case of Language Performance (see Figure 5), 5 seems to be the better number of predictors for the model due to this number of variables shows the more adequate combination: high AdjR2 at the same time that low BIC and Cp.

Figure 5

How Many IVs are the Optimal Number When Predicting Language Performance?



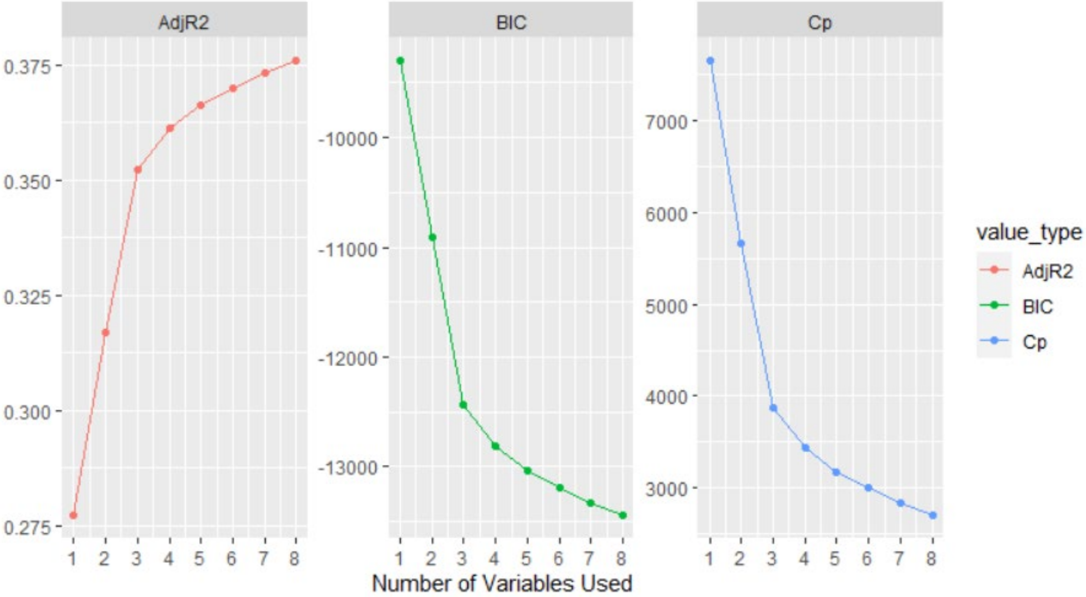


```
leaps_summary1 <- summary(leaps1)
require(tidyverse);require(ggplot2);require(ggthemes);
data_frame(Cp = leaps_summary1$cp,
           BIC = leaps_summary1$bic,
           AdjR2 = leaps_summary1$adjr2) %>%
        mutate(id = row_number()) %>%
        gather(value_type, value, -id) %>%
        ggplot(aes(id, value, col = value_type)) +
        geom_line() + geom_point() + ylab('') + xlab('Number of Variables Used') +
        facet_wrap(~ value_type, scales = 'free') + scale_x_continuous(breaks = 1:10)
```

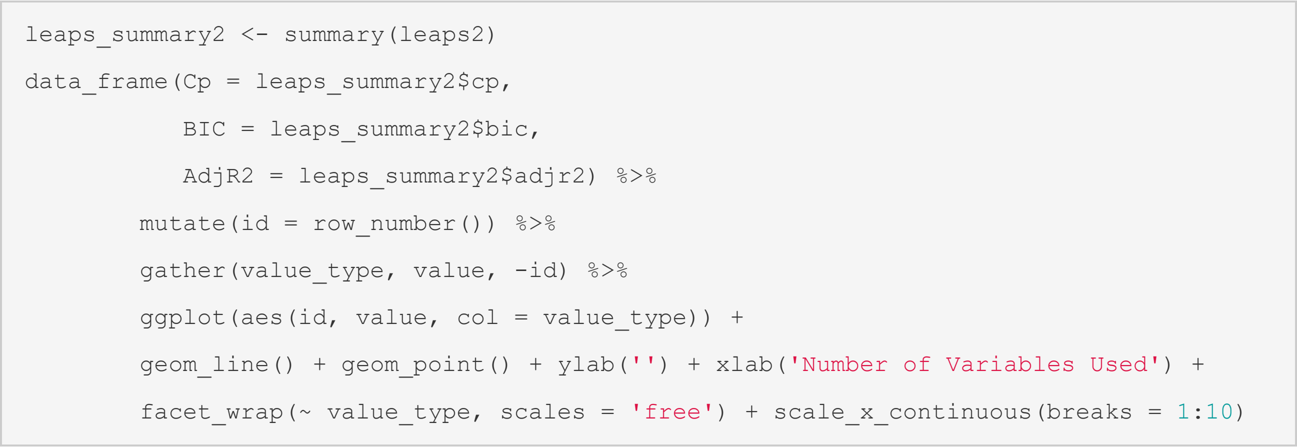
In the case of Math Performance (Figure 6), 8 is apparently the better number of predictors for the model because of the same combination of high AdjR2 and low BIC and Cp.

Figure 6

How Many IVs are the Optimal Number When Predicting Math Performance?



.



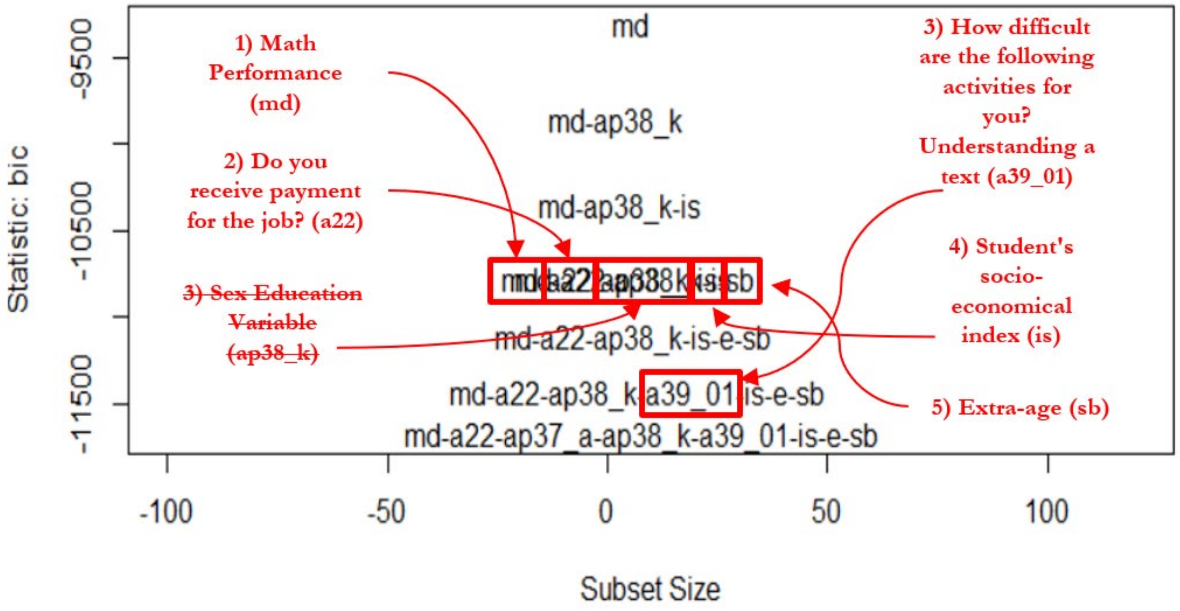
```
leaps_summary2 <- summary(leaps2)
data_frame(Cp = leaps_summary2$cp,
           BIC = leaps_summary2$bic,
           AdjR2 = leaps_summary2$adjr2) %>%
        mutate(id = row_number()) %>%
        gather(value_type, value, -id) %>%
        ggplot(aes(id, value, col = value_type)) +
        geom_line() + geom_point() + ylab('') + xlab('Number of Variables Used') +
        facet_wrap(~ value_type, scales = 'free') + scale_x_continuous(breaks = 1:10)
```

Subsequently, it was necessary to identify these optimal predictors. Two main ways can be adopted to achieve this objective: one is by analyzing the output of the subsets selection, and the other one is through graphics. Because the interpretation of the outputs can generate confusion given the considerable number of covariates, it was decided to work with the graphics alternative.

In Figure 7 it is presented the plot crossing BIC statistic with subset sizes when predicting Language Performance using the function *subsets* from *car* library. This plot yields a list of codes for the predictors and, as can be seen below, these equivalencies are highlighted in red.

Figure 7

What are the Best Predictors of Language Performance?



```r
library(car)
subsets(leaps1, statistic="bic", xlim=c(-100,120), legend = FALSE)
```

Therefore, we can say that according to the structure of our data, the 5 best predictors of Language Performance are:

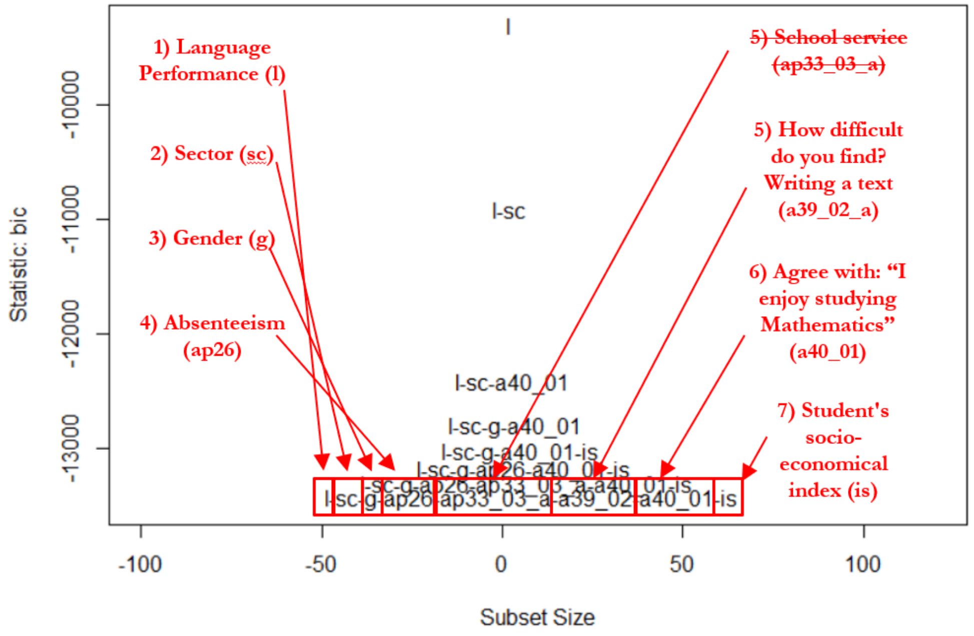
1. Math Performance (mdesemp)
2. Do you receive payment for the job you do outside your home? (ap22)
3. How difficult are the following activities for you? Understanding a text (a39_01)
4. Student's socio-economical index (isocioa)
5. Extra-age (sobreedad)

It is suggestive how most of the variables detected make perfect sense in their conceptual contribution to the prediction of Language Performance. However, it is important to note here that one of the identified independent variables (Sex Education variable) was excluded because it was not considered conceptually relevant for the prediction of our dependent variable.

Now, in Figure 8 are presented the 8 best predictors for Math Performance. In this case, also one of the independent variables was deleted (School service) due to its theoretical irrelevancy when predicting Math scores.

Figure 8

What are the Best Predictors of Math Performance?



```
library(car)
subsets(leaps2, statistic="bic", xlim=c(-100,120), ylim=c(-13500,-9300), legend = FALS
E)
```

As a result of applying the forward algorithm, the 7 best predictors of Math Performance can be listed as follows:

1. Language Performance (ldesemp)
2. Sector (either public or private) (sector)
3. Gender (gender)
4. Absenteeism. So far this year, how many times have you missed school? (ap26)
5. How difficult do you find the following activities? Writing a text (ap39_02)
6. To what extent do you agree with the following statements? I enjoy studying Mathematics (ap40_01)
7. Student's socio-economical index (isocia)

In Table 1, we are now able to present linear regressions outputs to predict Language and Math performances. Focusing on the most important coefficients in terms of absolute values, the improvement in Language Performance is associated with an increase of Math Performance by 46 percentage points (statistically significant at p <0.01). This coefficient is followed by medium and high socioeconomic students' level which increments Language performance by 21% and 36% respectively at p <0.01. There are also positive associations with difficulty to understand a text (p <0.01) as well as negative expectable associations with covariates such as working student (p <0.01), low student socioeconomic level (p <0.01), and over-age (p <0.05).

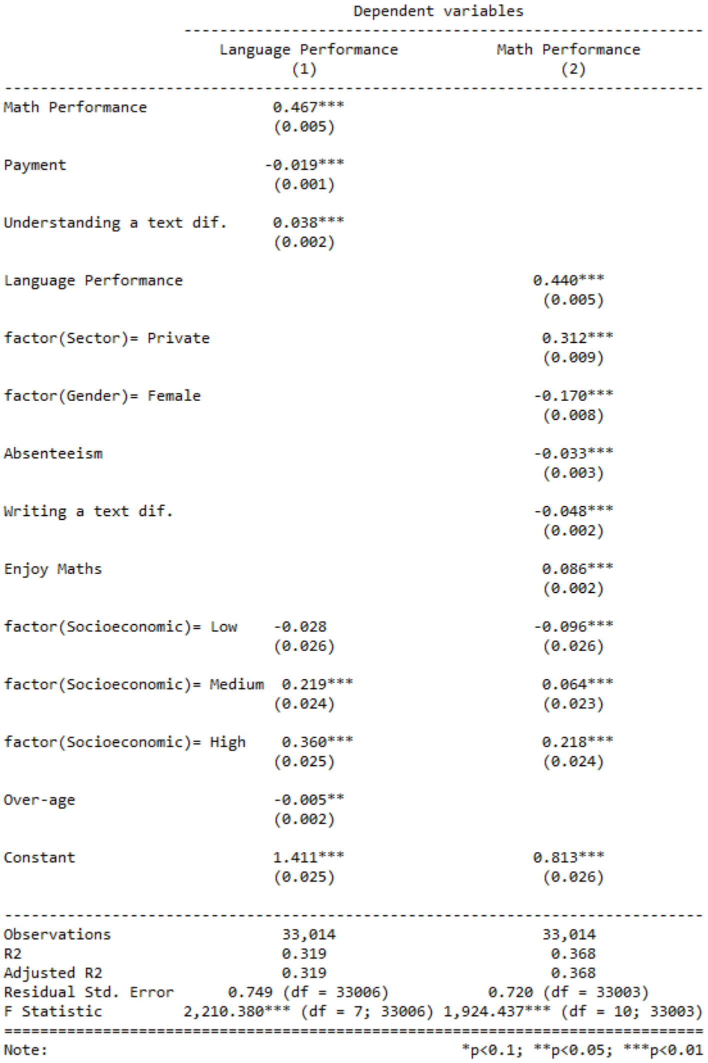
Regarding Math Performance, the coefficients that contribute the most to its increase are Language Performance (by 44%), private sector school management (by 31%), high socioeconomic status of students (by 21%), all of them statistically significant at p <0.01.

Likewise, there are positive associations with the fact of enjoying Maths and with medium socioeconomic level of students. Conversely, the coefficients that negatively impact Math Performance are being a female (by 17%), difficulty in writing a text (by 4.8%), low socioeconomic level (by 0.9%), and student absenteeism (by 0.03%) (all of them statistically significant at p <0.01). Here, it is important to note that explicative hypotheses around gender effects should be considered in terms of gender stereotypes and possible differential training paths intended for women in the Math realm.

Regarding the robustness of models, standard errors are in general low which means precise estimations. Additionally, R^2 shows how well the regression models fit the observed data and, in our cases, both of them indicate values higher than 30% which can be considered acceptable in our field of knowledge.

Table 1

Linear Regressions Outputs



```
                                       Dependent variables
                    ----------------------------------------------------------
                        Language Performance           Math Performance
                                (1)                           (2)
------------------------------------------------------------------------------
Math Performance              0.467***
                              (0.005)

Payment                      -0.019***
                              (0.001)

Understanding a text dif.     0.038***
                              (0.002)

Language Performance                                       0.440***
                                                            (0.005)

factor(Sector)= Private                                     0.312***
                                                            (0.009)

factor(Gender)= Female                                     -0.170***
                                                            (0.008)

Absenteeism                                                -0.033***
                                                            (0.003)

Writing a text dif.                                        -0.048***
                                                            (0.002)

Enjoy Maths                                                 0.086***
                                                            (0.002)

factor(Socioeconomic)= Low    -0.028                       -0.096***
                              (0.026)                       (0.026)

factor(Socioeconomic)= Medium  0.219***                     0.064***
                              (0.024)                       (0.023)

factor(Socioeconomic)= High    0.360***                     0.218***
                              (0.025)                       (0.024)

Over-age                     -0.005**
                              (0.002)

Constant                      1.411***                     0.813***
                              (0.025)                       (0.026)

------------------------------------------------------------------------------
Observations                   33,014                       33,014
R2                             0.319                         0.368
Adjusted R2                    0.319                         0.368
Residual Std. Error      0.749 (df = 33006)           0.720 (df = 33003)
F Statistic         2,210.380*** (df = 7; 33006) 1,924.437*** (df = 10; 33003)
==============================================================================
Note:                                              *p<0.1; **p<0.05; ***p<0.01
```

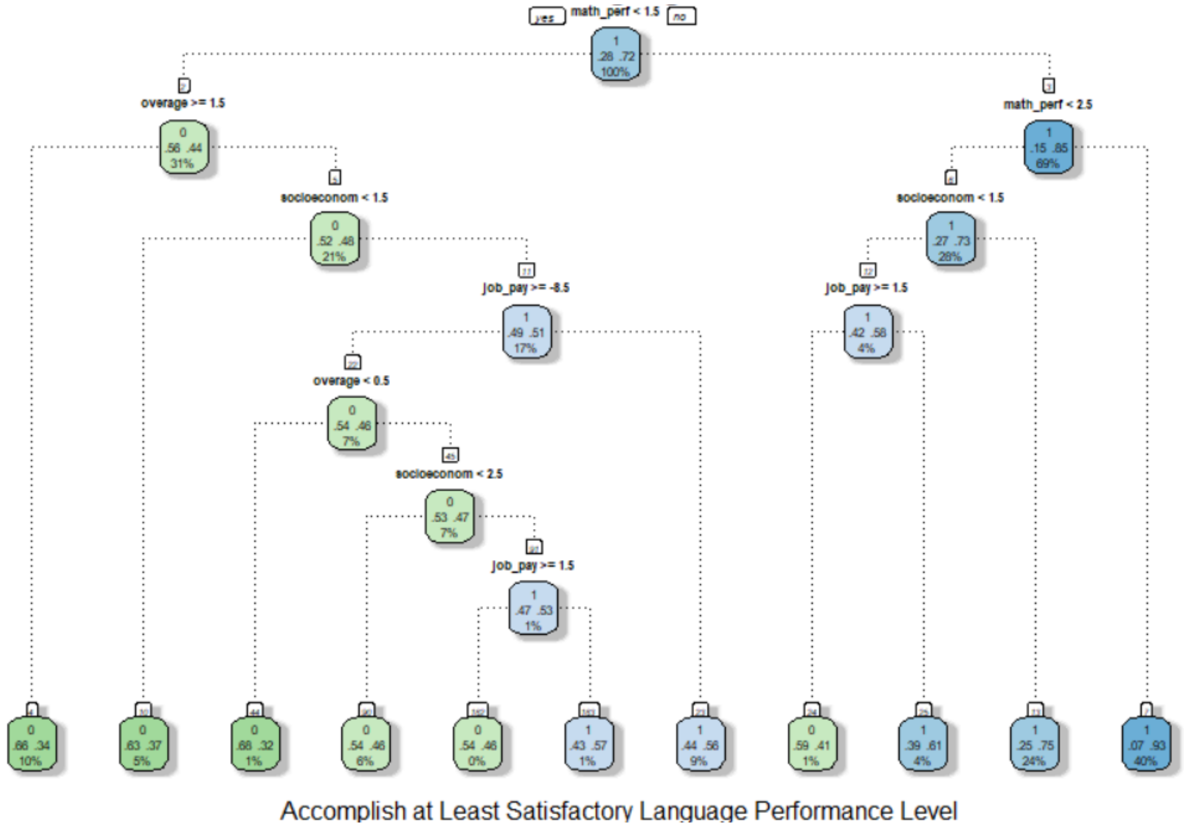
```
rightfit_lang=lm(ldesemp~ mdesemp + ap22 + ap39_01 +factor(isocioa) + sobreedad, data=
mydata)

summary(rightfit_lang)

rightfit_math=lm(mdesemp~ ldesemp + factor(sector)+ factor(gender) +

                      ap26 + ap39_02 + ap40_01 + factor(isocioa),data=mydata)

summary(rightfit_math)
```

Finally, this analytical exercise explored Tree-Based Models and analyzed their predictive accuracy. To do this, the following steps were followed: (1) use *subset* function to group variables selected previously; (2) rename such variables; (3) create new dataset without missing data through simple omissions; (4) recode lang_perf and math_perf as dummy variables (1 for satisfactory and advanced, 0 for basic and below) and then set seed; (5) order data by row number; (6) sample for training data (selection of 70% of data); (6) build training data and testing data; (7) run decision tree models; (8) plot decision trees; and (9) construct confusion matrixes and calculate accuracies.

For the case of the Language Performance decision tree, the graph shown in Figure 9 was obtained. When interpreting the tree, we can take the case of a particular student. Suppose we want to predict the chances that a student accomplishes at least satisfactory Language Performance Level when she/he presents: (1) Math Performance higher than 1.5 but less than 2.5 (basic level), (2) socioeconomic level lower than 1.5 (low level), and (3) a job payment value lower or equal to 1.5 (which means that the student does not work). According to the decision tree, in this particular case, this student has a 4% of chance of accomplishing at least a satisfactory Language Performance level.

Figure 9

Decision Tree Language Performance



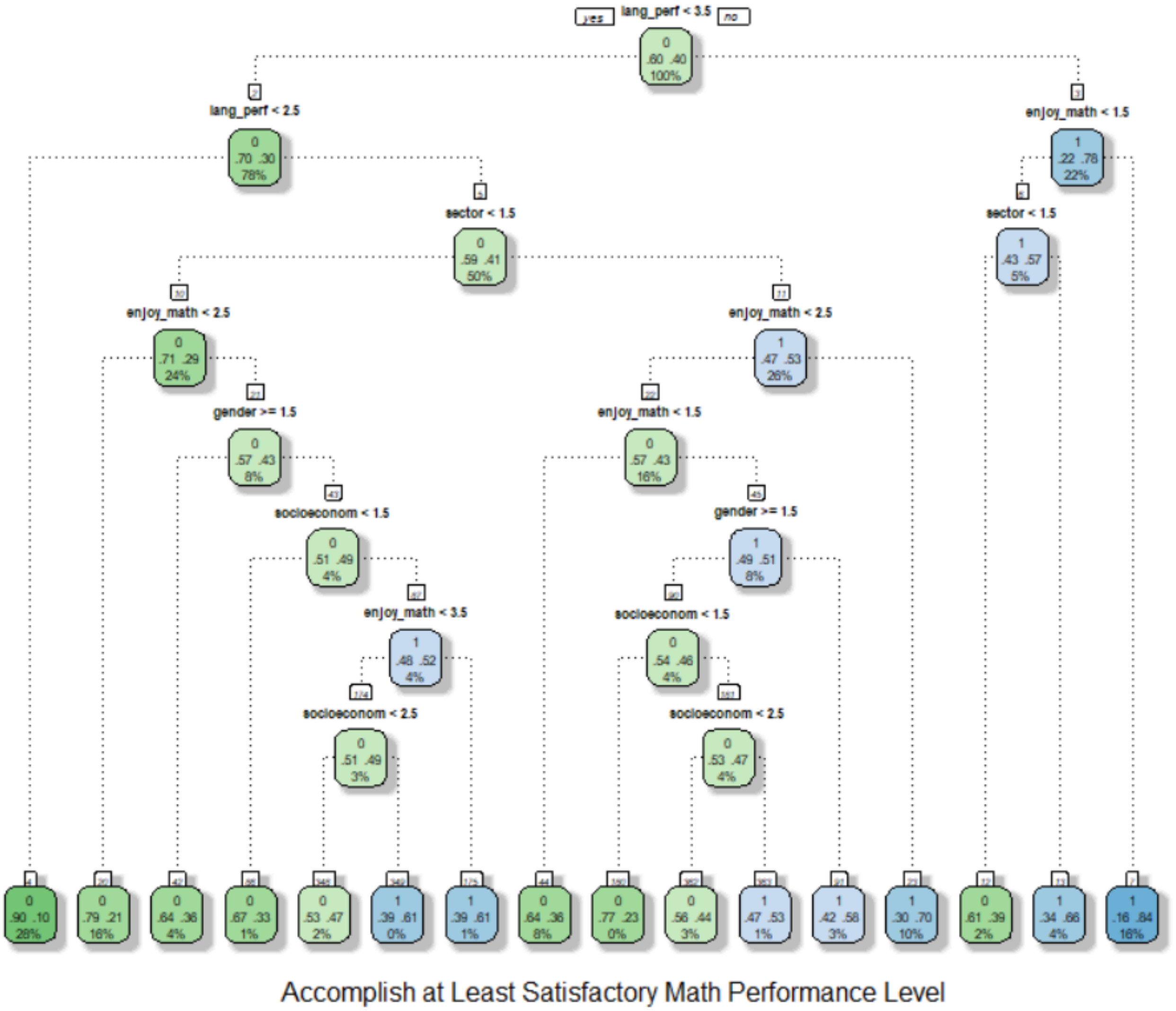
```
library("rpart")

library("rpart.plot")

library("rattle")

library(AER)
```

```
tree_base <- subset(mydata, select = c(ldesemp, mdesemp, ap22, isocioa, sobreedad))
names(tree_base)[1] <- "lang_perf"
names(tree_base)[2] <- "math_perf"
names(tree_base)[3] <- "job_pay"
names(tree_base)[4] <- "socioeconom"
names(tree_base)[5] <- "overage"
tree_base <- na.omit(tree_base)
tree_base$lang_perf <- ifelse(tree_base$lang_perf >= "3", 1, 0);
set.seed(1001)
new_tree_base <- tree_base[sample(nrow(tree_base)),]
t_idx <- sample(seq_len(nrow(tree_base)), size = round(0.70 * nrow(tree_base)))
traindata <- new_tree_base[t_idx,]
testdata <- new_tree_base[ - t_idx,]
dtree_lang <- rpart::rpart(formula = lang_perf ~ ., data = traindata, method = "class"
, control = rpart.control(cp = 0.001)) # complexity parameter
rattle::fancyRpartPlot(dtree_lang, type = 1, main = "Decision tree: Language Performan
ce", caption = "Accomplish at Least Satisfactory Language Performance Level" )
```

In Figure 10 the decision tree for Math Performance is included. In the same way as in the previous one, analysis of different profiles can be conducted. For example, a student with Language Performance lower than 3.5 but higher than 2.5 (satisfactory level), who attends a private school (sector value higher than 1.5) and has a "enjoy Math value" higher than 2.5 (which means that does agree with the sentence) presents a final 10% of chance of accomplishing at least satisfactory Math Performance level.

Figure 10

Decision Tree Math Performance



```
tree_base <- subset(mydata, select = c(mdesemp, ldesemp, sector, gender, ap26,
                                         ap39_02, ap40_01, isocioa))
names(tree_base)[1] <- "math_perf"
names(tree_base)[2] <- "lang_perf"
names(tree_base)[5] <- "absent"
names(tree_base)[6] <- "dif_writing"
names(tree_base)[7] <- "enjoy_math"
names(tree_base)[8] <- "socioeconom"
tree_base <- na.omit(tree_base)
tree_base$math_perf <- ifelse(tree_base$math_perf >= "3", 1, 0);
set.seed(1001)
```

```
new_tree_base <- tree_base[sample(nrow(tree_base)),]

t_idx <- sample(seq_len(nrow(tree_base)), size = round(0.70 * nrow(tree_base)))

traindata <- new_tree_base[t_idx,]

testdata <- new_tree_base[ - t_idx,]

dtree_math <- rpart::rpart(formula = math_perf ~ ., data = traindata, method = "class"
, control = rpart.control(cp = 0.001)) # complexity parameter

rattle::fancyRpartPlot(dtree_math, type = 1, main = "Decision tree: Math Performance",
caption = "Accomplish at Least Satisfactory Math Performance Level" )
```

Additional Conditional Inference Trees were modeled and cross-validation has been completed to compare the accuracy of our trees. Table 2 shows these results in which decision trees for Language Performance and Math Performance indicate greater accuracy compared to Conditional Inference Trees (0.767 and 1.764 respectively). In turn, the predicted accomplish rate is higher in Language Performance decision tree (0.809) in comparison to the decision tree for Math. Inversely the predicted not accomplish rate is higher in Math Performance decision tree (0.777). It could be suggested that the tree-based model for Language Performance does a slightly better job according to the structure of our data.

Table 2
Accuracy Comparison among Decision Trees

		Language Performance	**Math Performance**
Decision Trees	*Accuracy*	0.7679725	0.7649435
	Predicted Accomplish Rate	0.8090665	0.7415507
	Predicted Not Accomplish Rate	0.6147971	0.7778474
Conditional Inference Trees	*Accuracy*	0.7653473	0.7610057
	Predicted Accomplish Rate	0.7986811	0.7513612
	Predicted Not Accomplish Rate	0.6218532	0.7658381

```
# Decision Tree: Language Performance
resultdt <- predict(dtree_lang, newdata = testdata, type = "class")
cm_langdt <- table(testdata$lang_perf, resultdt, dnn = c("Actual", "Predicted"))
cm_langdt
##        Predicted
## Actual    0    1
##       0 1288 1491
##       1  807 6318
cm_langdt[4] / sum(cm_langdt[, 2])
## [1] 0.8090665
cm_langdt[1] / sum(cm_langdt[, 1])
## [1] 0.6147971
accuracydt <- sum(diag(cm_langdt)) / sum(cm_langdt)
accuracydt
## [1] 0.7679725
```

```
# Conditional Decision Tree: Language Performance
cm_langcit = table(testdata$lang_perf, round(predict(cit, newdata = testdata)), dnn =
c("Actual", "Predicted"))
cm_langcit
##        Predicted
## Actual    0    1
##       0 1161 1618
##       1  706 6419
cm_langcit[4] / sum(cm_langcit[, 2])
## [1] 0.7986811
cm_langcit[1] / sum(cm_langcit[, 1])
## [1] 0.6218532
accuracycit <- sum(diag(cm_langcit)) / sum(cm_langcit)
accuracycit
## [1] 0.7653473
```

```
# Decision Tree: Math Performance
resultdt <- predict(dtree_math, newdata = testdata, type = "class")
```

```
cm_mathdt <- table(testdata$math_perf, resultdt, dnn = c("Actual", "Predicted"))

cm_mathdt
##       Predicted
## Actual    0    1
##      0 4965  910
##      1 1418 2611
cm_mathdt[4] / sum(cm_mathdt[, 2])
## [1] 0.7415507
cm_mathdt[1] / sum(cm_mathdt[, 1])
## [1] 0.7778474
accuracydt <- sum(diag(cm_mathdt)) / sum(cm_mathdt)

accuracydt
## [1] 0.7649435
```

```
# Conditional Decision Tree: Math Performance

cm_mathcit = table(testdata$math_perf, round(predict(cit, newdata = testdata)), dnn =
c("Actual", "Predicted"))

cm_mathcit
##       Predicted
## Actual    0    1
##      0 5053  822
##      1 1545 2484
cm_mathcit[4] / sum(cm_mathcit[, 2])
## [1] 0.7513612
cm_mathcit[1] / sum(cm_mathcit[, 1])
## [1] 0.7658381
accuracycit <- sum(diag(cm_mathcit)) / sum(cm_mathcit)

accuracycit
## [1] 0.7610057
```

Code and Outputs

https://federico-jf.github.io/Knowledge-Mining/Final-Project.html

Conclusions and Discussions

This analytical exercise has addressed different predictive strategies that could guide pedagogical decisions in Argentine educational institutions including not just content-delivery strategies but also resources allocation, design of strengthening programs for students' performances, among others.

As observed, the analysis highlights the differences found between predictions made with more traditional hypothesis testing methods and those of Machine Learning aimed not just at achieving more precision but also identifying correct hypotheses (James, Witten, Hastie and Tibshirani, 2013). In fact, the main conclusions that emerge from this exercise indicate:

First, the most accurate predictive hypotheses for the Argentine “Aprender” National Evaluation can be identified using Machine Learning techniques when distinguishing optimal predictors for Language and Math Performances. In return, it should be noticed that traditional/classic pedagogical variables are not always the ones that best predict performance according to the Machine Learning techniques here utilized.

Second, an analysis of this type can help to adequately identify the dimensions to promote in projects for the design of educational public policies. The structure of the data itself “dictates” particular decisions without the need to assume *a priori* models that are often ineffective to understand the phenomenon submitted to analysis (Breiman, 2001).

Third, the prediction used to identify students at risk and then to make interventions aimed at strengthening desired performances can be an interesting pedagogical strategy to develop. In this sense, predictive systems that prevent unwanted results are considered more suitable than predictive systems aimed at the selection of students. Ethical discussions are relevant at this point when addressing such issues in the pedagogical field due to the possible existence of errors that can affect students' life projects (O'neil, 2016).

Finally, some limitations of this study should be noticed. Undoubtedly, data preprocessing could be bolstered with more intensive work on balancing the dataset, also taking advantage of knowledge mining techniques. At the same time, the treatment of missing values could be approached with more sophisticated techniques of Machine Learning such as K-Nearest Neighbor (KNN) algorithm instead of the direct omission of observations that in this opportunity was used. Additionally, other knowledge mining techniques not explored here remain to be applied (such as Random Forests, etcetera) whose accuracy rates could offer better options when predicting the performances studied.

Future analyzes could consider comparisons of the results of the “Aprender” evaluation among Argentine provinces. The Argentine educational reality is clearly heterogeneous and interesting trends and patterns could emerge that allow a detailed look at the phenomena in each of the regions. In return, some findings could undoubtedly be generalized at the national level. Thus, educational decision-making could be nourished from this accurate and regional evidence.

References

Aprender (2019). National Evaluation Operation Aprender. Retrieved May 12, 2021, from https://www.argentina.gob.ar/educacion/aprender2019

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3), 199-231.

Holmes, W., Bialik, M. & Fadel, C. (2019) Artificial intelligence in education: promises and implications for teaching and learning. Boston, MA: The Center for Curriculum Redesign.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open-source analytics initiative. *Journal of Learning Analytics*, 1(1), 6-47.

O'neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Siegel, E. (2016). Predictive Analytics. The power to predict who will click, buy, lie or die. New Jersey: John Wiley and Sons.

Williamson, B. (2016). Digital education governance: data visualization, predictive analytics, and ‘real-time’ policy instruments. *Journal of Education Policy*, 31(2), 123-141.