Academic Analytics: Predictions around Argentine “Aprender” National Evaluation

EPPS 6323 Knowledge Mining

Dr. Karl Ho

Student: Federico Ferrero

Objective

To conduct both an **Exploratory Data Analysis** and a **Predictive Analysis** that uses Machine Learning techniques with the purpose of finding the most accurate and adequate predictive hypotheses for the Argentine "Aprender" National Evaluation.

National Evaluation Operation "Apren

- Argentine **Ministry of Education**, Culture, Science and Technology.
- **Annual** administration.
- Population: all **high-school seniors** in Argentina.
- Evaluation **purpose**: "to generate timely and quality information to better understand the achievements and pending challenges around students' learning" (Aprender, 2019).
- Traditionally, **predominant use of descriptive** techniques.
- Collects data on knowledge of **Mathematics**, **Language**, and **contextual information** of the respondent students.

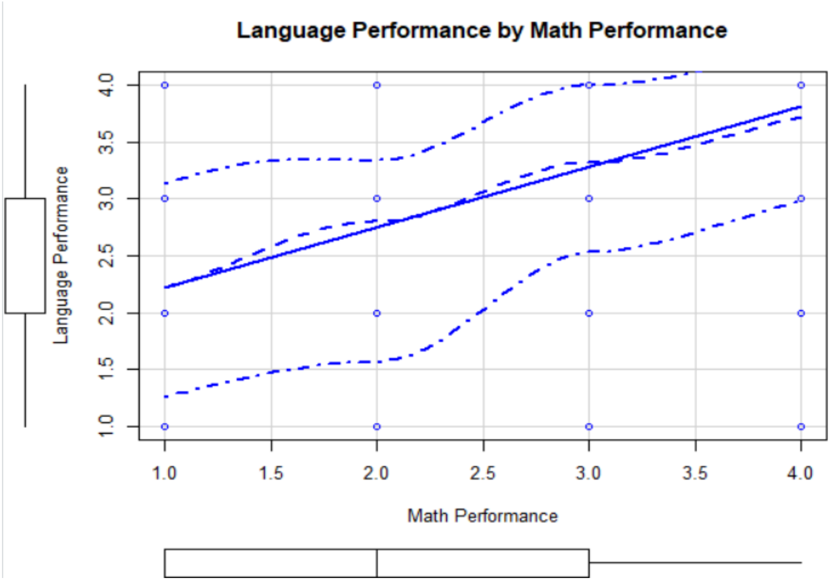
Data

- 2019 edition.
- **N=34,191** high-school seniors (Cordoba Province).
- **Dependent Variables** (2):
  - **Language Performance** (ldesemp) and **Math Performance** (mdesemp):
    - 4 categories: below basic level, basic, satisfactory, and advanced.
- **Independent Variables** (246):
  - gender, sector (public or private), ambit (rural or urban), student socio-economic situation, student cultural consumption, school climate, student self-perception, educational practices and use of technology, migration status, etcetera.

Analysis strategies

1. **Exploratory Data Analysis** using **visualizations:**
   - Traditional variables: Sector, Ambit, Gender, Repetition, Student Employment, Student Socioeconomic Level.
2. **Finding the Best Model** using *regsubsets* with leaps package:
   - Math Performance: Forward selection.
   - Language Performance: Backward selection.
3. **Supervised Learning** techniques:
   - **Simple regression**.
   - **Tree-Based-methods**.

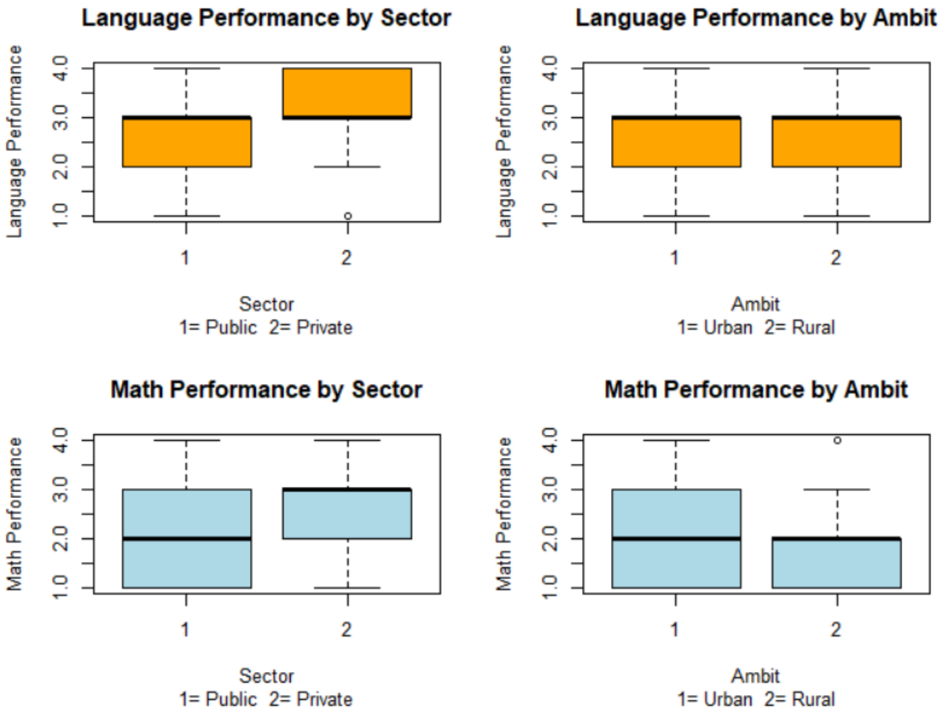
Exploratory Data Analysis



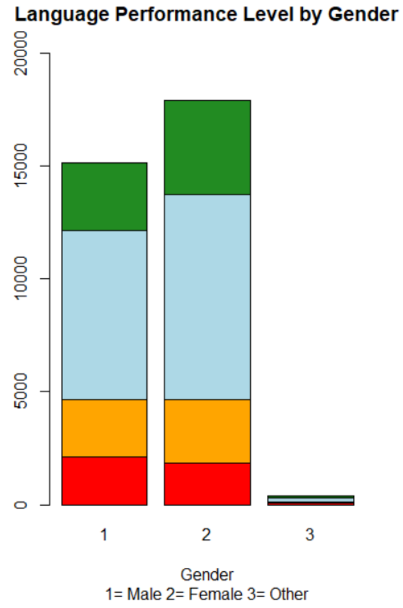
Positive and linear association between Language Performance and Math Performance.

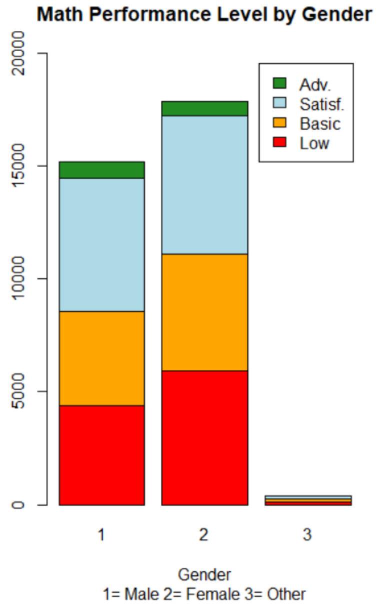
Students' performance and Sector by Ambit

- Sector:
  - 1= Public
  - 2= Private
- Ambit:
  - 1= Urban
  - 2= Rural
- **Language Scores**: *better performance at private schools and no differences between ambits*. 50% of cases in private schools are between satisfactory and advanced. In the public system, 50% between basic and satisfactory.
- **Math Scores**: *better performance in private and urban schools*. 50 % of cases in private schools are between satisfactory and advanced. In public system, 50% between below basic and satisfactory. Performance in rural is worse than in urban schools.



Students' performance by Gender





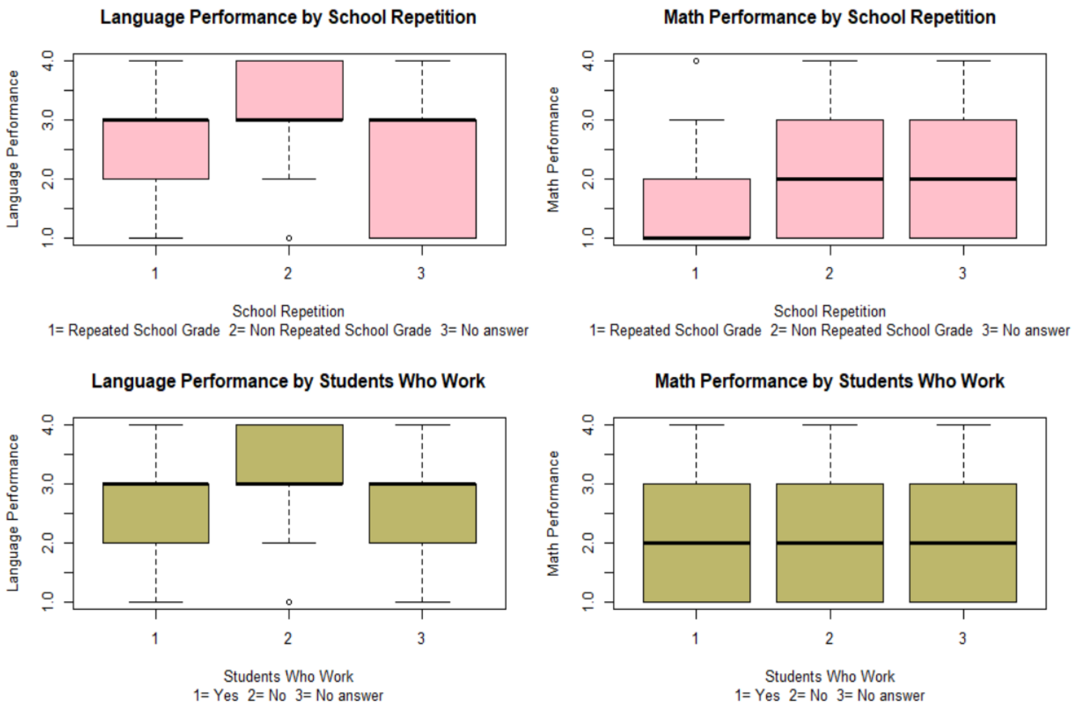
- **Language Scores**: *better performance in women* (39% at least obtain the satisfactory level while 31% of men achieve this level). Both groups have 13% of students in basic and low level.

- **Math Scores**: *worse performance in women* (32% obtain low and basic levels against 25% in the case of male students). Both groups have 19% of students in satisfactory and advanced levels.

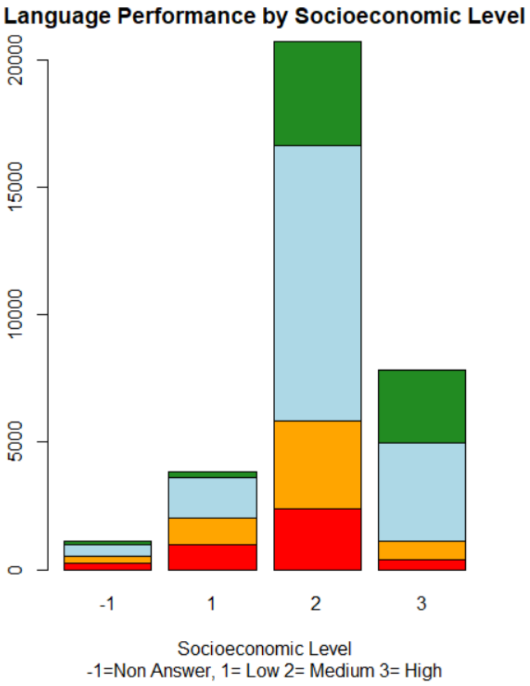
Students' performance by Repetition and Work Experience

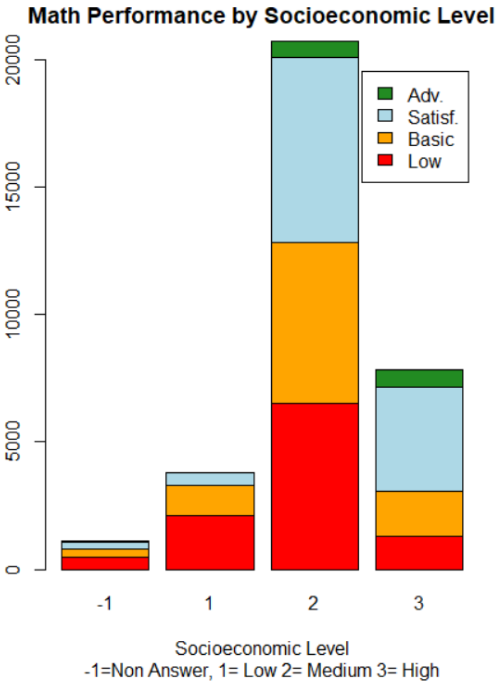
- **Language Scores**: *clear better performance in non-repitent students and students who don't work.*

- **Math Scores**: *better performance in non-repitent students. No apparent differences when considered worker students.* A considerable number of missing values.



Students’ Performance by Socioeconomic Level



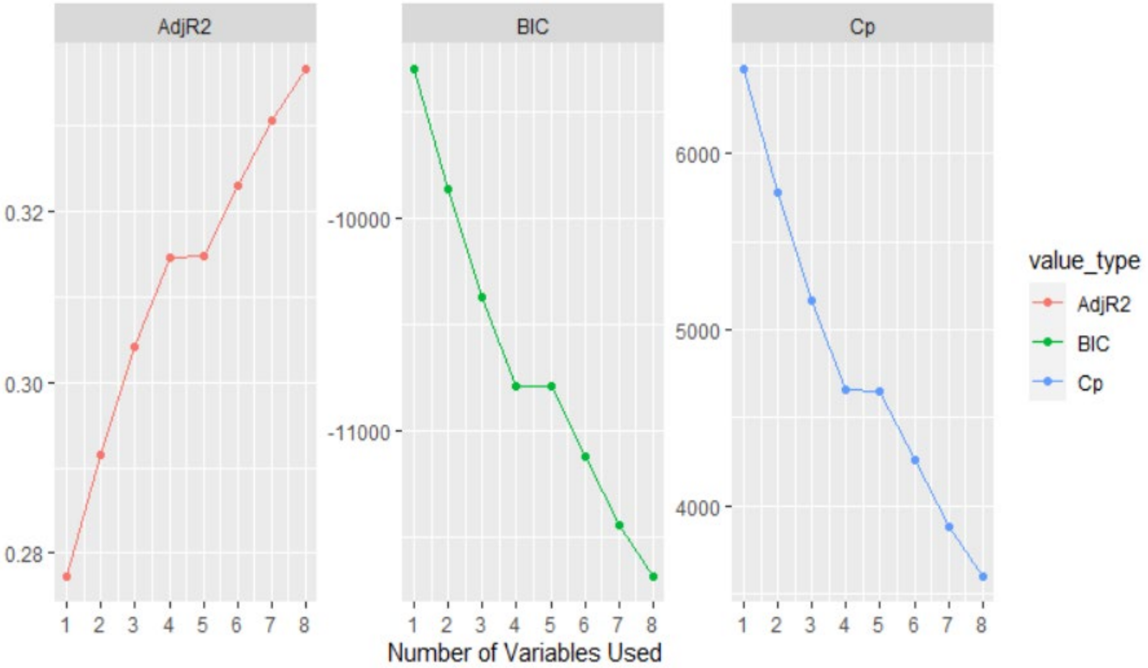


- **Language Scores:** *predominantly middle socioeconomic level students who obtain at least a satisfactory level* (44%).

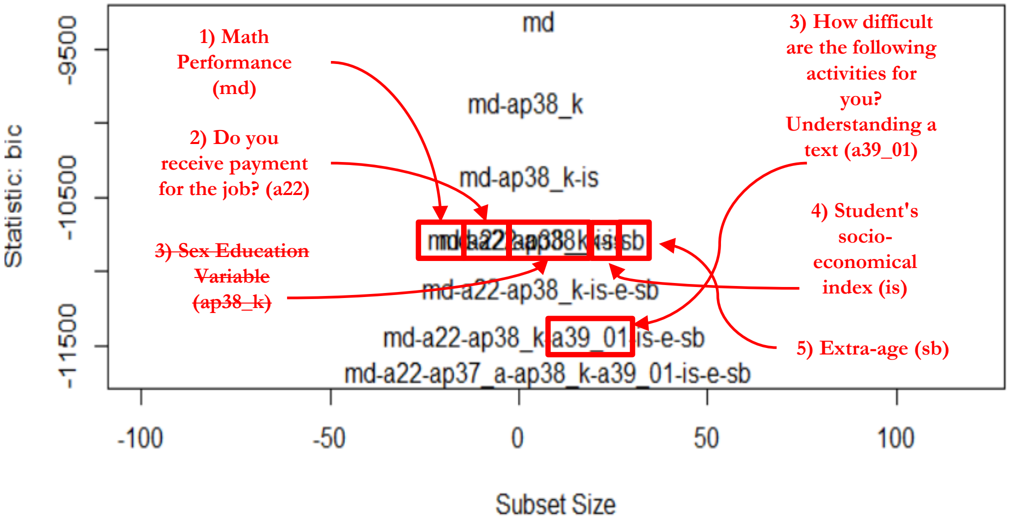
- **Math Scores**: *predominantly middle socioeconomic level students who obtain low and basic level* (37%).

How Many are the Optimal Number When Predicting Language Performance?

- **5** seems to be the **better number of predictors** for the model when predicting **Language Performance**: high AdjR2 and low BIC and Cp.



What are the Best Predictors of Language Performance?

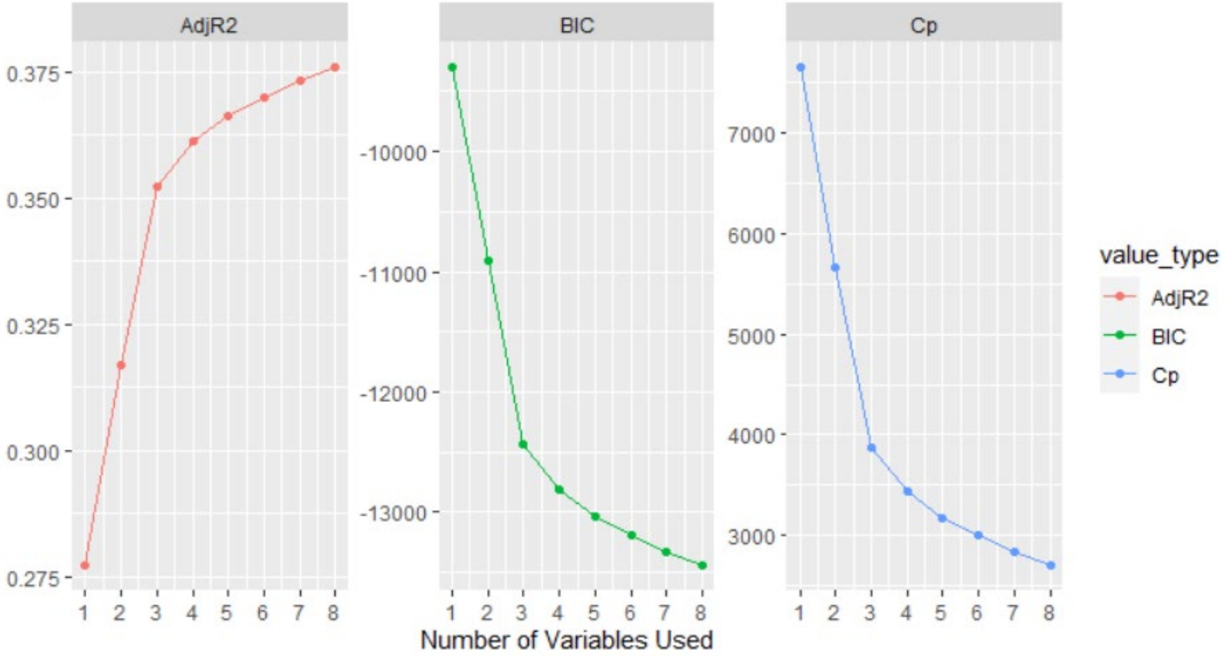


5 Best Predictors of Language Performance

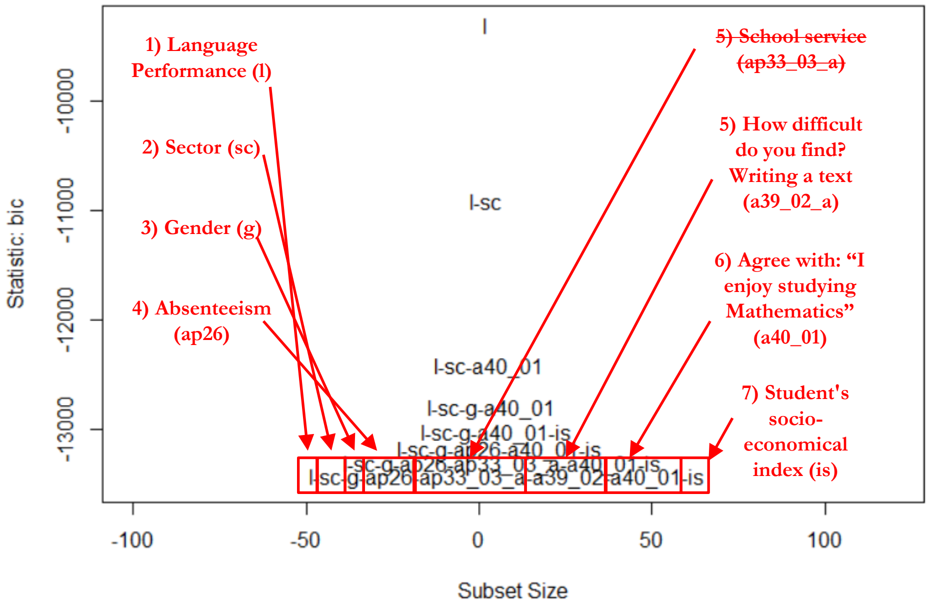
1. **Math Performance** (mdesemp)
2. Do you receive **payment for the job** you do outside your home? (ap22)
3. How difficult are the following activities for you? **Understanding a text** (a39_01)
4. Student's **socio-economical index** (isocioa)
5. **Extra-age** (sobreedad)

How Many are the Optimal Number When Predicting Math Performance?

- **8** seems to be the **better number of predictors** for the model when predicting **Math Performance**: high AdjR2 and low BIC and Cp.



What are the Best Predictors of Math Performance?



7 Best Predictors of Math Performance

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)

Linear Regression Outputs

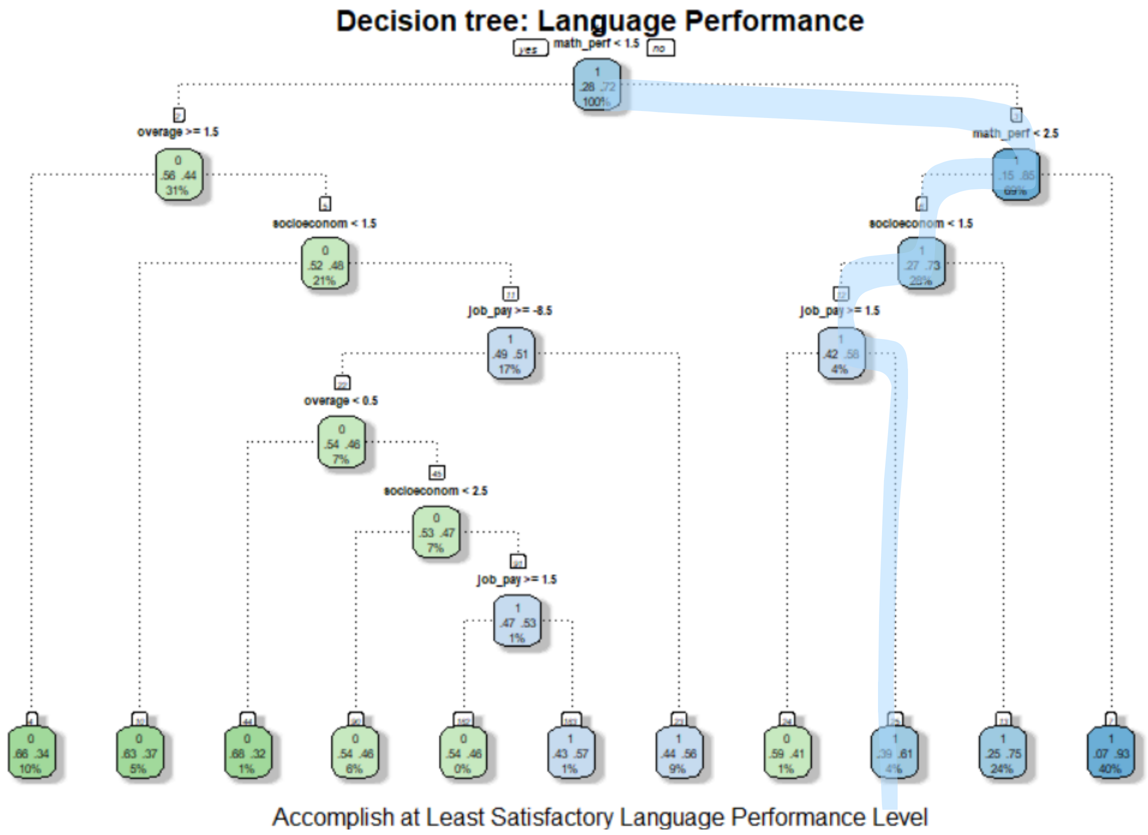
Comparing Regression Models Outputs

	Dependent variables	
	Language Performance (1)	Math Performance (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

- **Language Performance:**
  - *Positive association:*
    - Math performance, Difficulty in understanding a text, Student' medium and high socioeconomic level ($p<0.01$).
  - *Negative association:*
    - Payment for a job ($p<0.01$), Student's low socioeconomic level ($p<0.01$), over-age ($p<0.05$).

- **Math Performance**:
  - *Positive association:*
    - Language performance, Private sector, Enjoying Maths, Student' medium and high socioeconomic level ($p<0.01$).
  - *Negative association:*
    - Absenteeism, Female, Difficulty in writing a text, Student's low socioeconomic level ($p<0.01$).

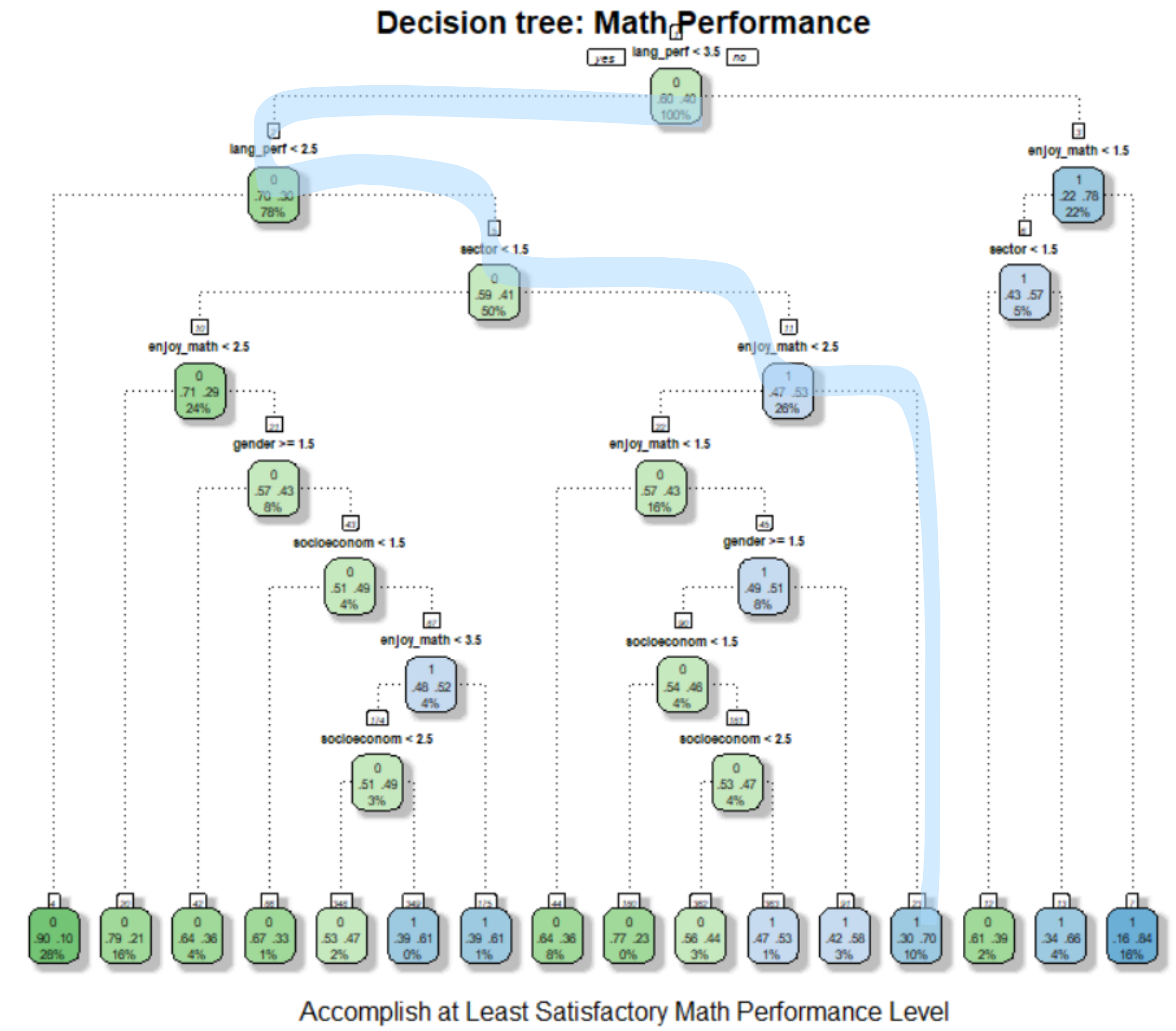
Tree-Based Methods: Language Performa



Let's try a case: a student who has…

- **Math Performance** higher than 1.5 but less than 2.5 (basic level)
- **Socioeconomic level** lower than 1.5 (low index)
- And a **job payment value** higher or equal to 1.5 (which means that the student doesn't work)

Has **4% of chances** of accomplishing at least satisfactory Language Performance Level.



Tree-Based Method Math Performance

Let's try another case: a student who…

- **Language Performance** lower than 3.5 but higher than 2.5 (satisfactory level)
- Attends **to a public school** (Sector value lower than 1.5)
- And the **enjoy math value** is lower than 2.5 (which means that does not agree with the sentence)

Has **10% of chances** of accomplishing at least satisfactory Math Performance Level.

Tree-Based Methods: Which predictive method is more

		Language Performance	Math Performance
Decision Tree	*Accuracy*	0.7679725	0.7649435
	Predicted Accomplish rate	0.8090665	0.7415507
Conditional Inference Tree	*Accuracy*	0.7653473	0.7610057
	Predicted Accomplish rate	0.7986811	0.7513612

Code and Outputs

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

Final ideas

- Most **accurate predictive hypotheses** for the Argentine "Aprender" National Evaluation can be identified using Machine Learning techniques.
- **Traditional/classic pedagogical variables** are not always the ones that best predict performance according to the Machine Learning techniques used here.
- An analysis of this type can help to adequately identify the dimensions to promote in projects for the **design of educational public policies**.
- The prediction used to **identify students at risk** and then to make interventions aimed at strengthening desired performances can be an interesting pedagogical strategy.

References

Acharya, M. S., Armaan, A., & Antony, A. S. (2019, February). A comparison of regression models for prediction of graduate admissions. In *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)* (pp. 1-5). IEEE.

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

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

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.

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

Thank you!