Exploración de factores asociados al aprendizaje

Juan F. Mancilla-Caceres, Luis Furlán y Lynette García
Centro de Estudios en Informática Aplicada, Instituto de Investigaciones, Universidad del Valle de Guatemala
jfmancilla@uvg.edu.gt

RESUMEN: En este artículo, se analizaron los resultados del desempeño de 1.2 millones de estudiantes de último año de diversificado en las evaluaciones que el Ministerio de Educación (MINEDUC) administró durante los años 2006 al 2016. Los datos constan de las notas obtenidas en las evaluaciones de Matemática y Lectura, así como información acerca de factores asociados al aprendizaje, tales como lenguaje materno y estrato socio-económico, entre otros. El análisis consistió en la exploración de las relaciones entre los promedios obtenidos a nivel municipal y de las zonas dentro de la Ciudad de Guatemala y las variables asociadas entre sí. En este estudio, también mostramos un modelo de regresión que muestra qué factores son los más importantes para determinar el éxito en las evaluaciones de los siguientes años. Los resultados muestran una alta correlación entre estrato socioeconómico promedio del municipio y entre la educación de los padres con los resultados de los exámenes. Esto indica que se debe hacer un mayor esfuerzo para ayudar a los jóvenes que se consideran "primera generación" respecto a ir al colegio. Algunas de las otras conclusiones incluyen la necesidad de encontrar formas alternativas de medir el desempeño en Matemáticas debido a la alta correlación entre la habilidad de Lectura y dicho desempeño.

PALABRAS CLAVE: Factores Asociados, Desempeño en Educación, Predicción de Notas.

Exploring Associated Factors on Education

ABSTRACT: In this article, we analyze the performance of 1.2 million students on standardize tests from the Ministry of Education of Guatemala (MINEDUC for its name in Spanish). These tests were administered on senior students before graduation during 2006 and 2016. Data consists on the grades obtained on Math and Reading tests, as well as factors that may affect learning such as the language spoken at home and socio-economical status. The analysis consisted in exploring the correlations between the municipal averages and the demographical data. We also present a regression model used to predict the grades based on previous years and the chosen factors that most influence the performance. The model shows a high impact of the socio-economical status and the parents education on the grades, which suggests that policy makers should increase support for first-generation students.Another conclusion suggests that educators should find alternative ways to evaluate Math performance given the high correlation with the Reading tests.

KEYWORDS: Associated Factors, Performance in Education, Predicting Grades.

Introducción

A partir del año 2006, el Ministerio de Educación (MINEDUC) ha administrado pruebas de Matemáticas y Lectura a todos los graduados de 5o año de diversificado a nivel nacional. Estas

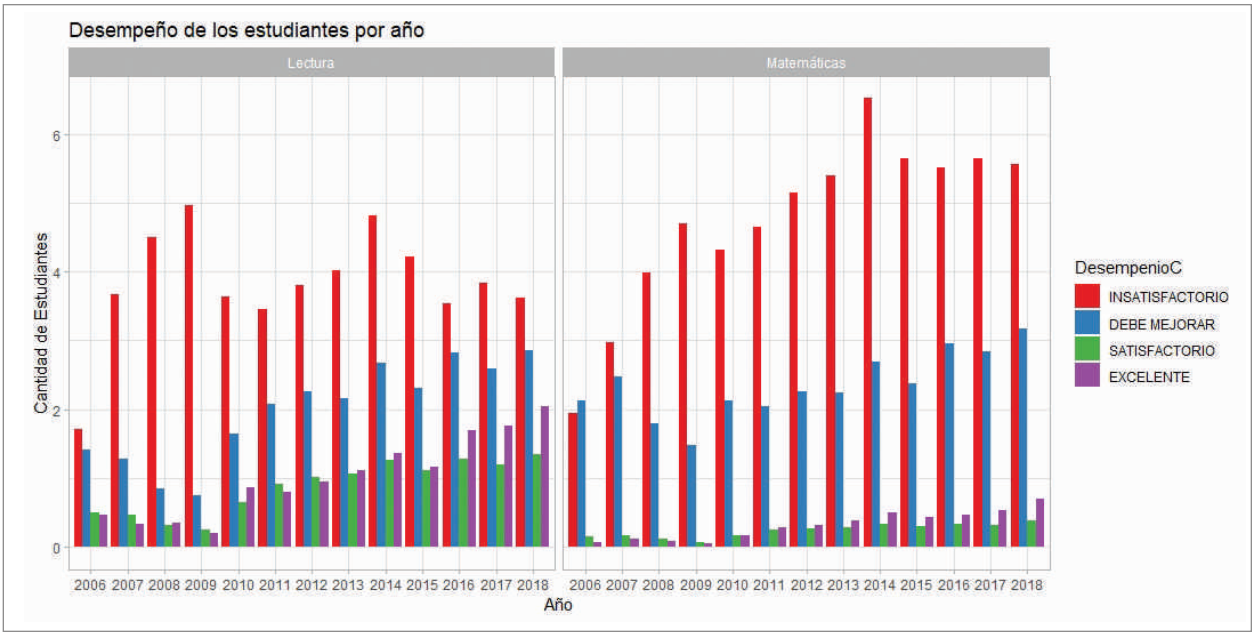

Figura 1. Desempeño de los estudiantes por año. 52.4 % de los estudiantes tiene un desempeño insatisfactorio mientras un 25.06 % debe mejorar.

pruebas constituyen un censo ya que todos los graduados, tanto de colegios privados como instituciones públicas, deben someterse a ellas. Además de las pruebas de desempeño, los estudiantes deben llenar una encuesta de factores asociados, que incluyen preguntas al respecto de su familia (por ejemplo, el idioma materno y paterno, su identificación étnica, etc.) así como factores socio-económicos (por ejemplo, que electrodomésticos poseen, tipo de piso, paredes y techo de su casa, etc.), y conductuales (por ejemplo, cuantos libros leen al año, periódicos, etc.).

En este estudio, se presenta la exploración de las relaciones entre los factores asociados y el rendimiento en lectura y matemática a nivel municipal. También se presenta un modelo de regresión para predecir el desempeño en lectura de cada municipio basado en los factores asociados y el desempeño del año anterior. El objetivo de este análisis es incrementar el entendimiento de los factores que afectan el desempeño de los estudiantes y así poder planificar intervenciones que mejoren el desempeño académico de los mismos.

Una de las mayores limitaciones de los datos es que no se poseen las mismas variables en todos los años, por ejemplo, el tipo de agua que se usa para beber y el nivel educativo de los padres. Además, estos datos son obtenidos por auto reporte de los estudiantes que no necesariamente reflejan la realidad en su totalidad. Debido a estas razones, en el presente estudio, se utilizaron los promedios a nivel municipal, que permite evitar sesgos individuales de los estudiantes, así como analizar patrones más generales. Sin embargo, esto a su vez introduce nuevos retos como el que no todos los municipios tienen la misma cantidad de estudiantes e instituciones.

Descripción de los datos

Los datos consisten en los resultados de 10 años de pruebas administradas por el MINEDUC, que incluyen evaluaciones de lectura y matemática, así como un cuestionario de factores asociados. En total, se cuenta con 3,390 observaciones a nivel municipal (entre los años del 2006 al 2016). Se identificaron 68 variables que eran comunes en todos los años de recolección. Estas variables incluyen información acerca del desempeño en las áreas de matemática y lectura, uso de computadora, número de períodos por materia, etnia del estudiante, idioma(s) hablado(s) en casa, información laboral del estudiante, nivel educativo de los padres, características de la institución educativa (por ejemplo, tipo de jornada, sector oficial o privado), hábitos de lectura, así como factores económicos como las características del hogar (e.g., tipo de piso, paredes, techo, número de cuartos), disponibilidad de agua, electrodomésticos y automóviles.

Análisis correlativo

Como parte del análisis exploratorio, en esta sección se muestran los resultados de una serie de correlaciones encontradas entre distintas medidas de desempeño y los factores presentes en las encuestas. El objetivo de este análisis preliminar es mostrar algunos de los retos y observaciones más comunes en los datos. En la Figura 1 se muestran los promedios de todos los estudiantes que tomaron las evaluaciones a nivel nacional. A pesar de que se puede observar que la proporción de alumnos que tiene un desempeño excelente aumenta, por lo general se ve que más de la mitad de estudiantes (el 52.4 %) tiene un desempeño insatisfactorio en ambas áreas y un 25.06 % un desempeño que debe mejorar.

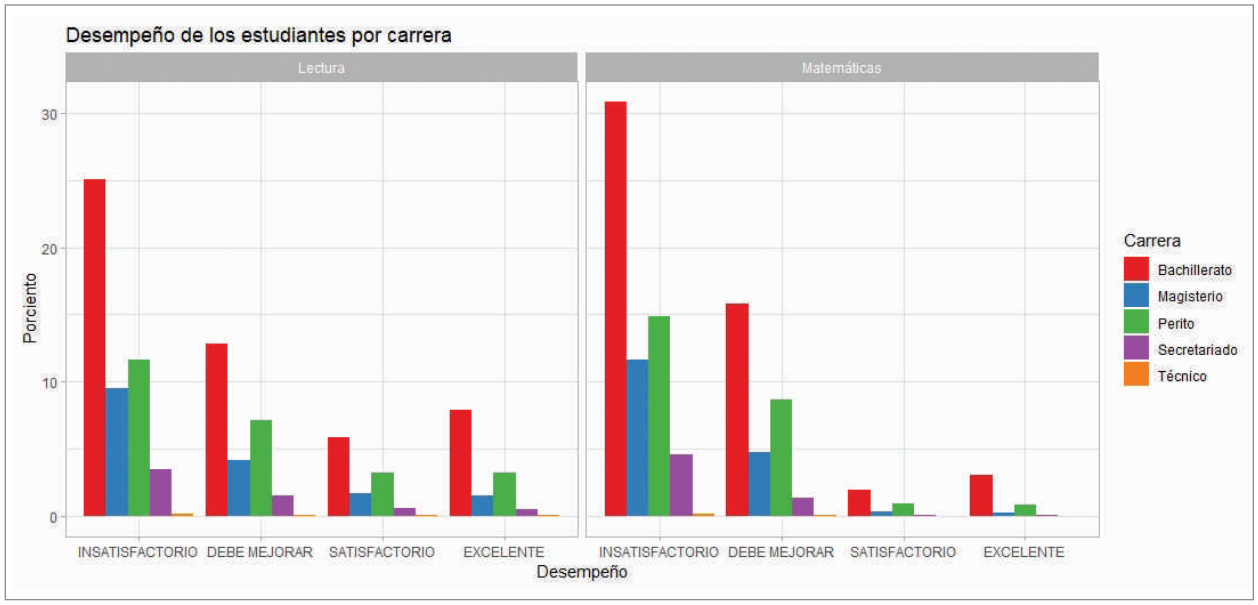

Figura 2. Desempeño de los estudiantes por rama. Los estudiantes de bachillerato muestran mayor proporción de desempeño insatisfactorio.

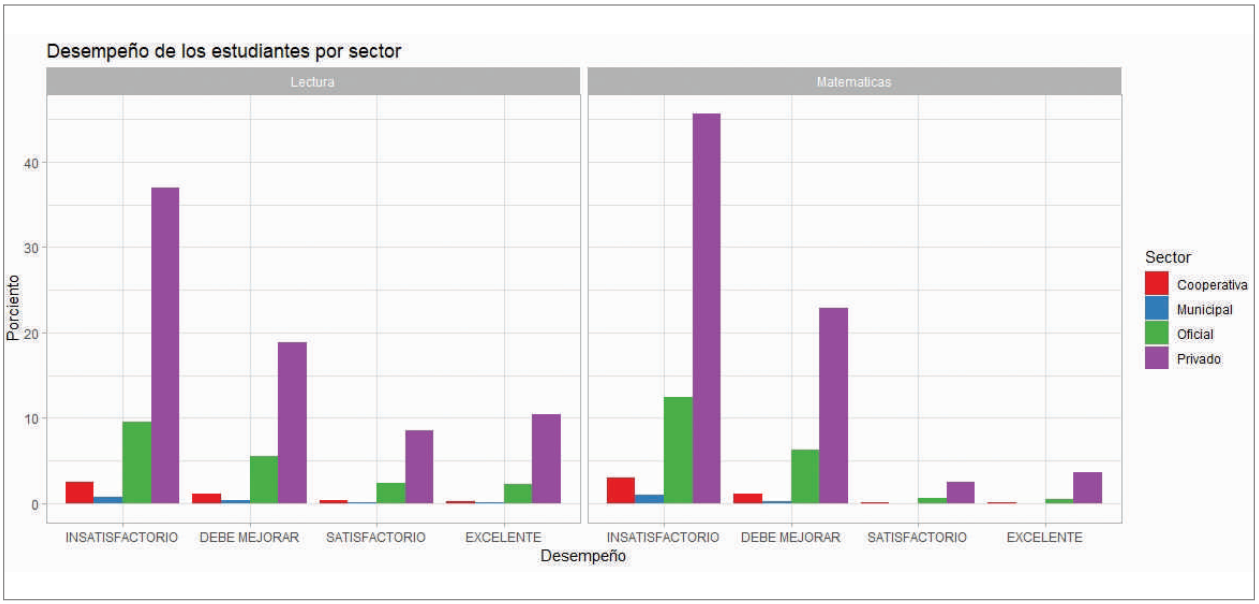

Figura 3. Desempeño por sector de la institución. Los estudiantes que atienden a instituciones privadas tienden a mostrar el mayor grado de insatisfactorio.

Debido a la gran proporción de estudiantes con desempeño insatisfactorio, en el resto de esta sección se enfoca, únicamente, en mostrar la distribución de estudiantes con desempeño insatisfactorio en distintas dimensiones. En la Figura 2 se ilustra dicha distribución en función de la rama de estudio de los estudiantes. El 49.6 % de los estudiantes estudiaron bachillerato, 6.3 % secretariado, un 19.1 % magisterio, 24.5 % perito, y 0.3 % un técnico. Como se puede ver en la figura, la rama con mayor proporción de desempeño insatisfactorio es el bachillerato, reportando un 25 % en el caso de lectura y un 30 % en el caso de matemáticas.

En el caso de desempeño por sector de la institución, se encontró que el sector privado (que representa el 74.9 % de todos los estudiantes) es donde más estudiantes tienen un desempeño insatisfactorio en ambas áreas de lectura y matemática. Ver Figura 3.

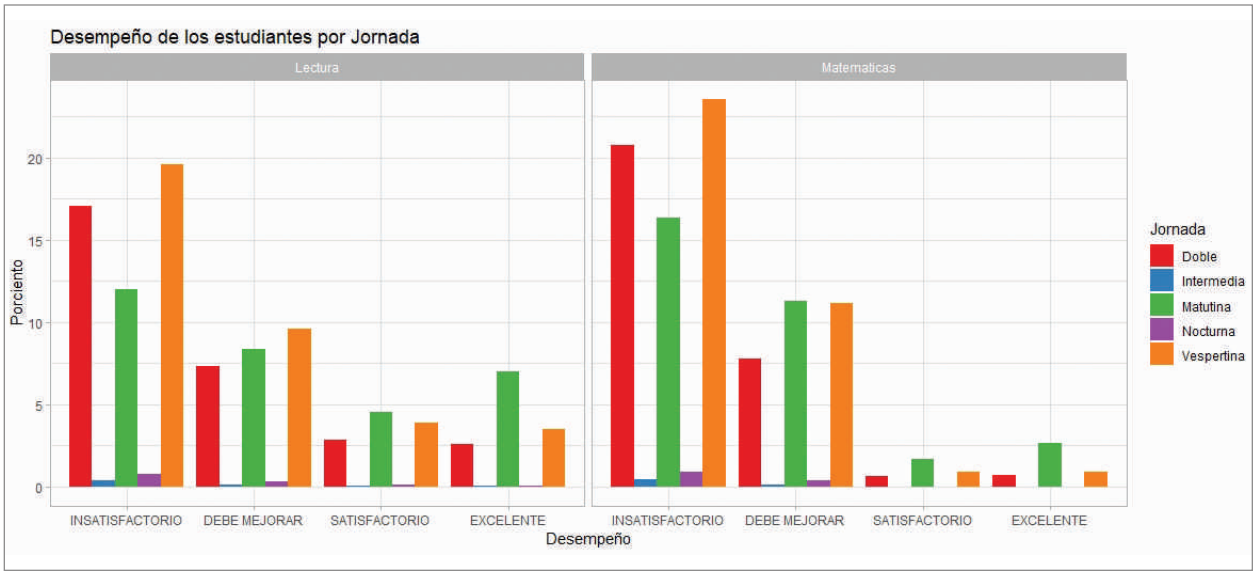

Figura 4. Desempeño por jornada. Los estudiantes de las jornadas vespertina y doble muestran la mayor proporción de desempeño insatisfactorio.

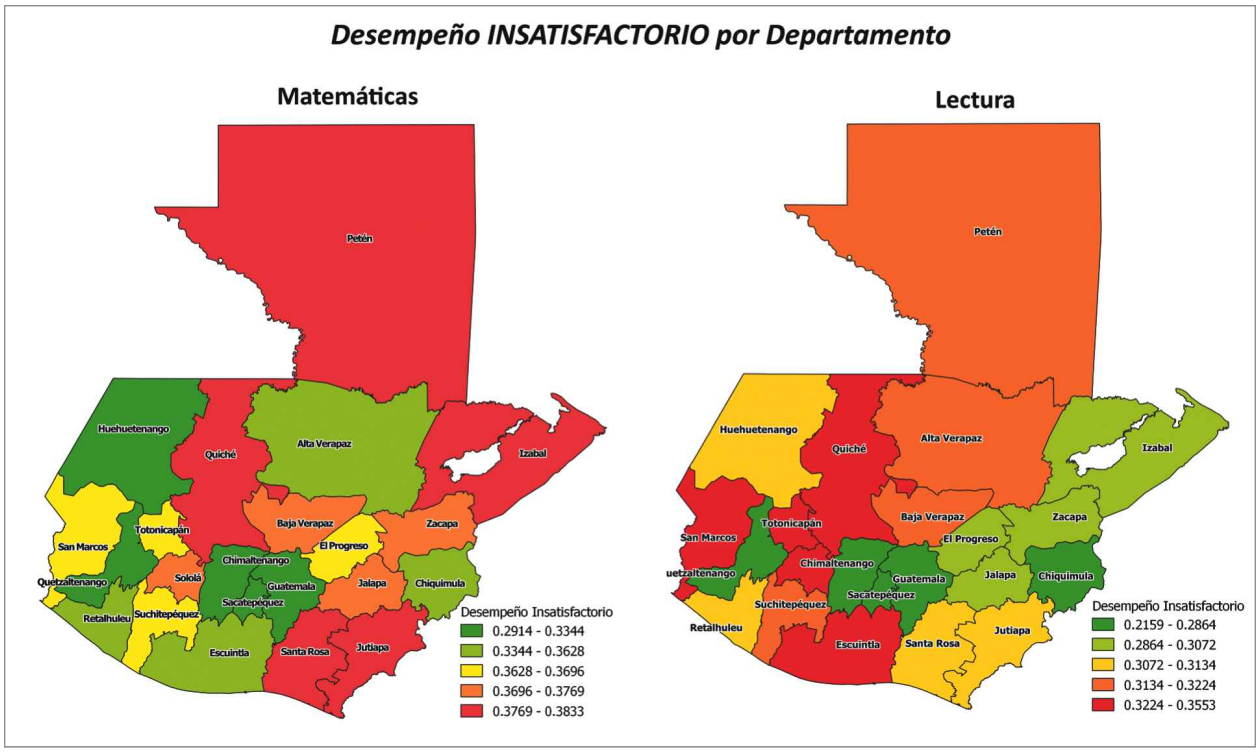

Figura 5. Distribución de estudiantes con desempeño insatisfactorio a nivel departamental.

Cuando se analizó el desempeño por la jornada en que los estudiantes atienden a las instituciones, se observó que aquellos que van a la jornada vespertina y doble tienden a mostrar mayor nivel de desempeño insatisfactorio tanto en lectura como en matemáticas. En total, el 65.9 % de los estudiantes atienden a la jornada doble o a la vespertina. La Figura 4 muestra la distribución de desempeño por jornada.

Para finalizar, la Figura 5 muestra la distribución de estudiantes con desempeño insatisfactorio a nivel departamental. El color en la figura, en este caso, representa la proporción de estudiantes que reportan un desempeño insatisfactorio dentro de cada departamento, siendo así que un color rojo representa el peor de los casos, mientras que el color verde representa el menos insatisfactorio de los casos. En la figura también se puede

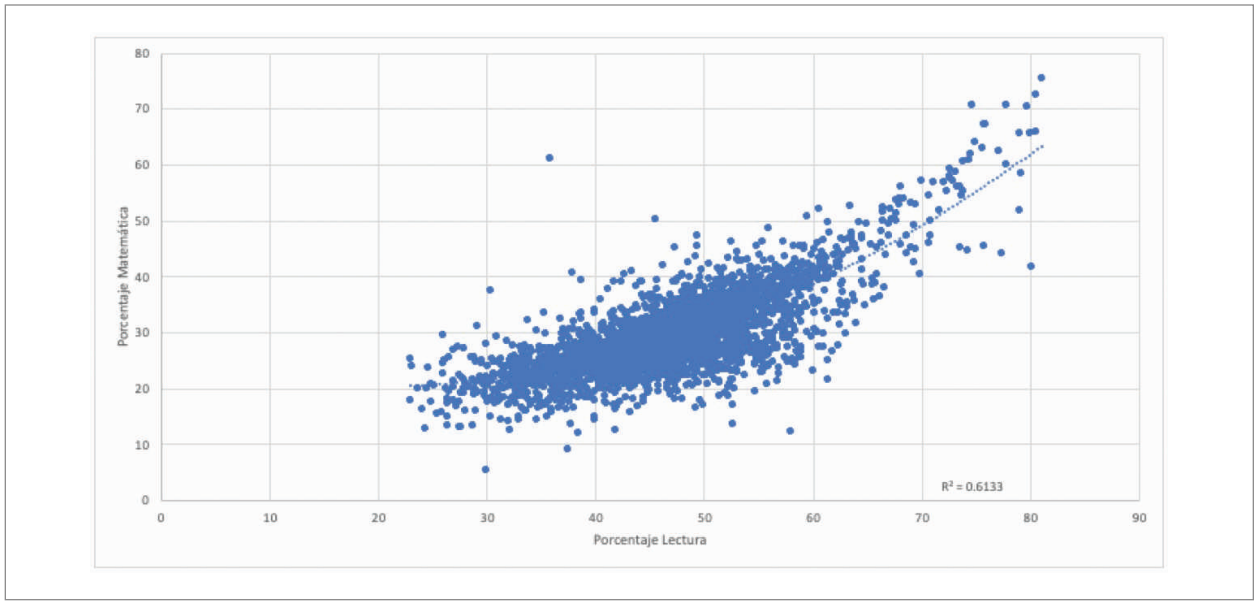

Figura 6. Correlación entre notas de matemática y lectura. Se observa un coeficiente de correlación R2 = 0.61.

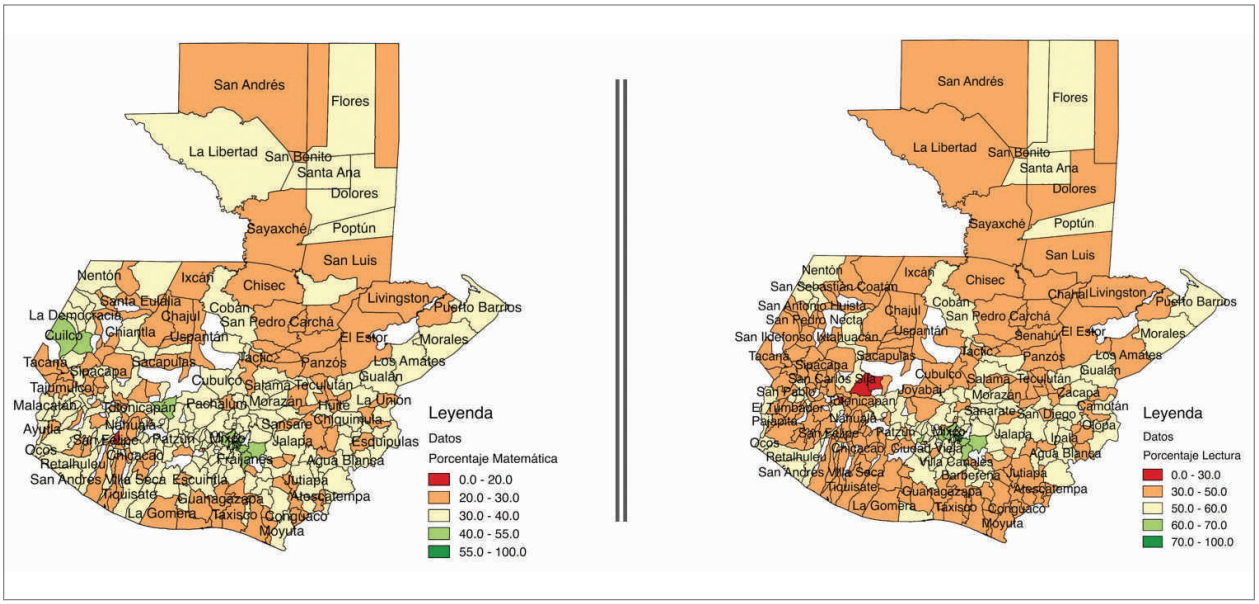

Figura 7. Matemática y lectura a nivel municipal.

observar que, aunque existe una correlación entre departamentos que tienen grave desempeño insatisfactorio en ambos lectura y matemática, algunos parecen mostrar menor problema en matemática que en lectura, como es el caso de Alta Verapaz, Escuintla y San Marcos.

Modelo predictivo

En la presente sección mostramos la creación de un modelo predictivo con el objetivo de predecir el desempeño promedio a nivel municipal en la prueba de lectura. Decidimos construir este modelo para así poder identificar las variables que más poder predictivo tienen para el desempeño en lectura. Escogimos únicamente predecir la nota de lectura debido a que encontramos una alta correlación entre la nota de la prueba de matemática y de lectura (ver Figura 6). Este resultado es esperado ya que, para poder obtener una buena nota en matemáticas, primero es necesario poder entender el enunciado de los problemas. Esto apunta a una de nuestras primeras recomendaciones para futura investigación: explorar métodos distintos de evaluación matemática para así no tener resultados altamente correlacionados.

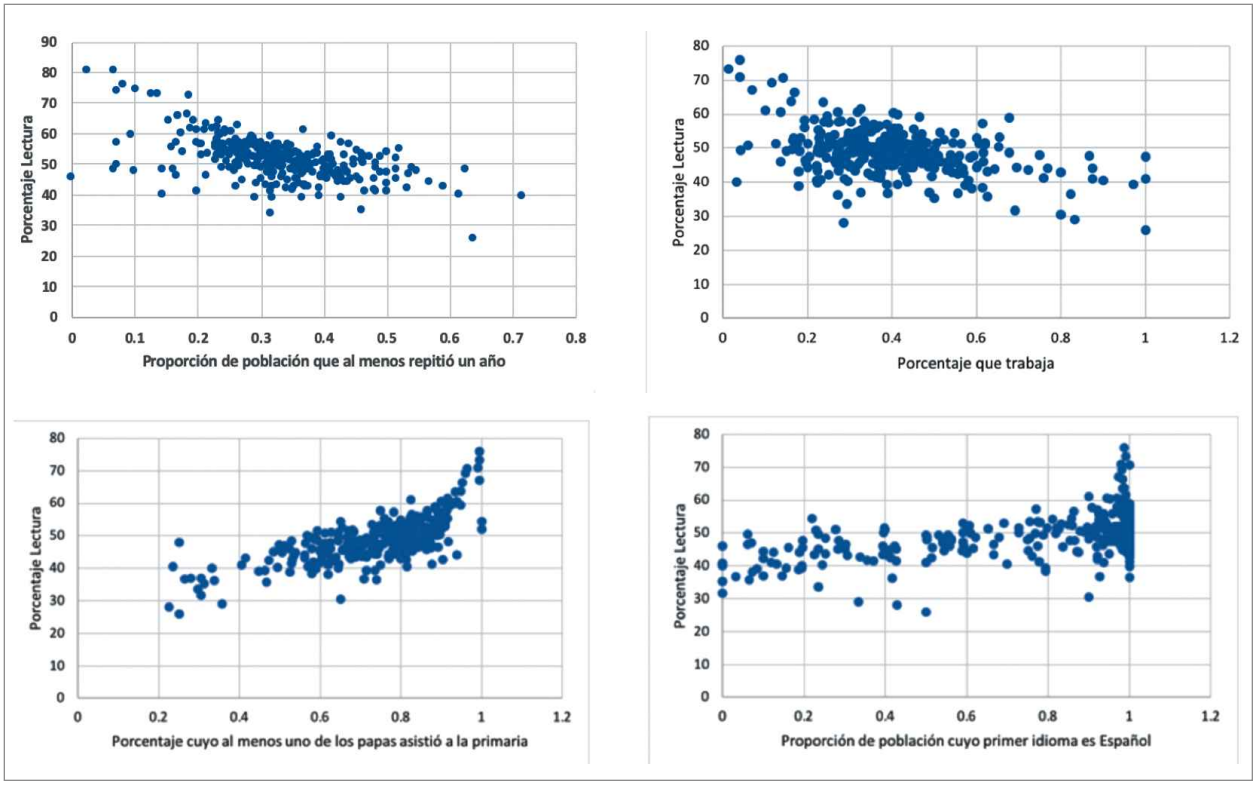

Figura 8. Correlaciones entre nota en lectura y proporción de estudiantes que han repetido un año, que trabaja, cuyos padres asistieron a la primaria, y cuyo primer idioma es español

En la Figura 7, se muestra la distribución de notas para ambas áreas de lectura y matemáticas. La correlación que se mencionó previamente se puede observar claramente. Cabe notar que las mejores notas se encuentran en el área central del país (cercano al municipio de Guatemala) con excepción de Cuilco, Huehuetenango donde hay un alto desempeño en el área de las matemáticas.

Para la construcción del modelo predictivo, se exploró la relación de cada variable con la nota obtenida en lectura. De esa forma se puede ver la influencia de algunas de esas variables en los resultados. Algunas de estas variables exploradas fueron: el porcentaje de estudiantes en el municipio que reportan trabajar mientras estudian, la proporción de estudiantes que reportan tener computadora, el porcentaje de estudiantes que reportan haber repetido al menos un año durante sus estudios, la proporción de estudiantes que atendió a pre-primaria, los que reportan idioma español como primer idioma, la jornada en la que atienden y el nivel educativo de los padres.

En la Figura 8 se pueden observar algunas de estas correlaciones. Cabe resaltar que no todas las correlaciones son lineales y se puede observar que solo aquellos municipios en los cuales al menos el 80 % de padres asistieron a la primaria tienen oportunidad de tener un desempeño promedio satisfactorio. De la misma forma, solo aquellos donde la mayoría reporta hablar español como primer idioma pueden llegar a tener un promedio en lectura de más de 60 puntos. Esto muestra otra de nuestras conclusiones: se deben explorar distintas formas de evaluación de lenguaje para capturar y representar de mejor manera el rendimiento de las comunidades donde el español no es el primer idioma.

Otra observación importante de la Figura 8 es la correlación positiva de la nota de lectura y el poseer computadora. Esta señal es importante porque más que referirse a la posible influencia de las computadoras en la habilidad de lectura, es posible que se refiera a la correlación entre el desempeño académico y el estrato socio económico. Para medir esto, se diseñó una variable socio- económica a partir de otras de las variables en el cuestionario de factores asociados. En particular, se utilizó información acerca del tipo de piso, pared y techo del hogar en que los estudiantes viven, el agua que usan para lavar, el número de automóviles que poseen, el número total de electrodomésticos, el número de servicios contratados que pagan, el tipo de combustible que usan para cocinar y el tamaño del hogar. Estas variables fueron ponderadas según su representatividad de estatus socioeconómico, por ejemplo, entre más caro sea el material usado para piso, más alto el valor en la variable socio económica.

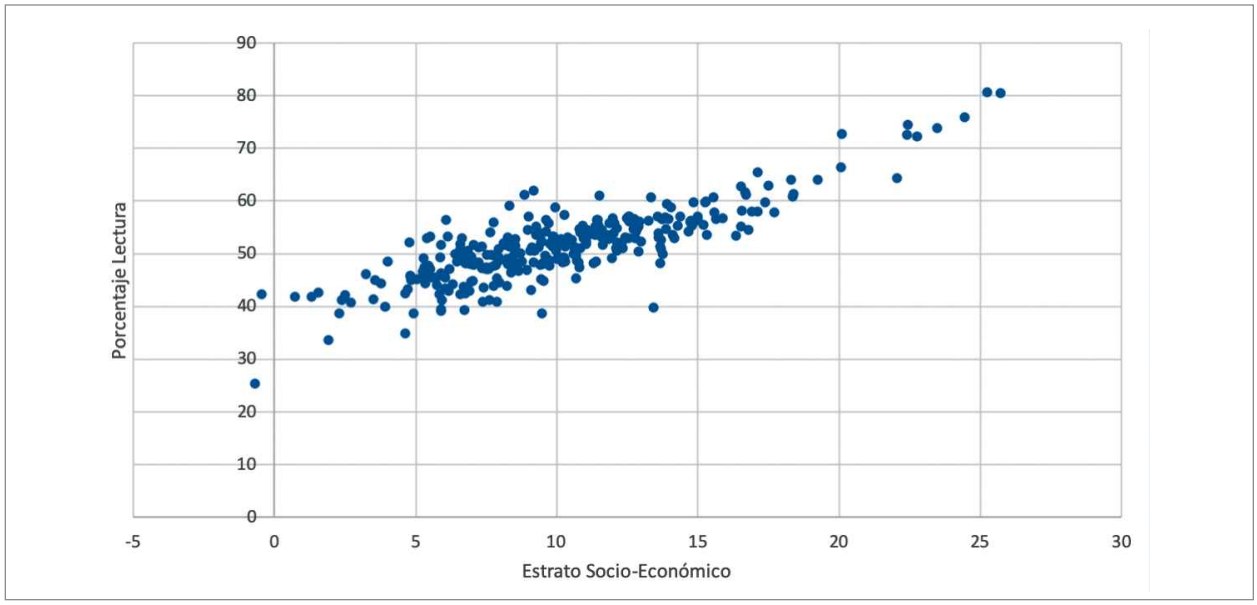

Figura 9. Correlación entre la nota de lectura y la variable socio económica. El valor en el eje horizontal de esta variable no tiene una interpretación directa y debe de ser considerado solo como un valor ordinal, es decir, entre más a la derecha se encuentra un municipio, en promedio sus habitantes tienen mayor estatus socio-económico.

La Figura 9 muestra la correlación entre la variable creada y los resultados de lectura mostrando una alta correlación entre estatus socioeconómico y desempeño. Esto es consistente con teorías de factores asociados ya que generalmente este estatus está correlacionado con mejor alimentación y recursos para estudios (Van der Berg, 2008).

Regresión usando redes neuronales: funciones de base radial (RBF regression)

Utilizando las variables descritas en la sección anterior, se entrenó un modelo lineal basado en redes neuronales (con bases Gaussianas radiales) para predecir el desempeño de cada municipio en función de los factores asociados. Estas redes neuronales son usualmente utilizadas para la clasificación de instancias desconocidas. Sin embargo, es posible también utilizarlas para regresión, lo que las hace especialmente útiles para aplicaciones en finanzas, física, biología y en este caso, educación.

Las redes con funciones de base radiales (RBFnets por sus siglas en inglés) son usadas para aproximar una función o hacer regresión. La idea es utilizar varias funciones Gaussianas para aproximar la función deseada, haciendo que una RBFnet sea prácticamente una red neuronal de 2 capas donde las entradas entran a una capa totalmente conectada y la salida es la suma ponderada (o combinación lineal) de la capa oculta (ver Figura 10)

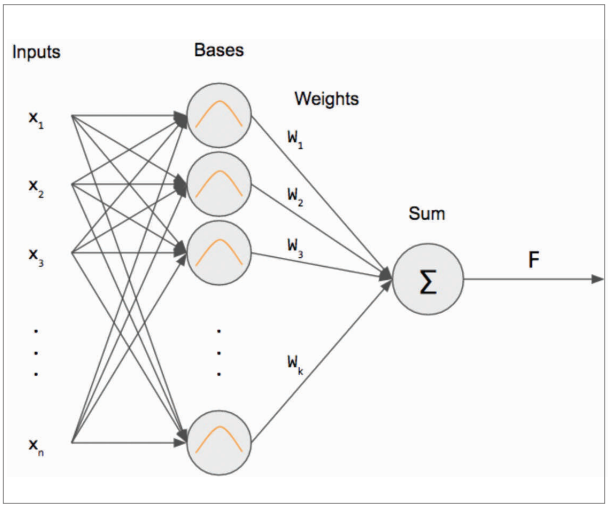

Figura 10. Modelo conceptual de una regresión utilizando bases radiales gaussianas.

Se entrenó el modelo RBFNet con los datos del año 2006 al 2015 para ajustar los parámetros de la regresión y predecir, para cada año, el año siguiente. Se usó una validación cruzada de 10 iteraciones (10-fold cross validation), el modelo se entrenó con un total de 2,674 instancias. Esto generó una correlación de 0.83 con un error absoluto de 3.21. Para terminar de evaluar el modelo, se realizó una segunda prueba utilizando como datos de prueba solamente los del año 2016 e intentar predecirlos con el modelo ajustado anteriormente. Los resultados se muestran

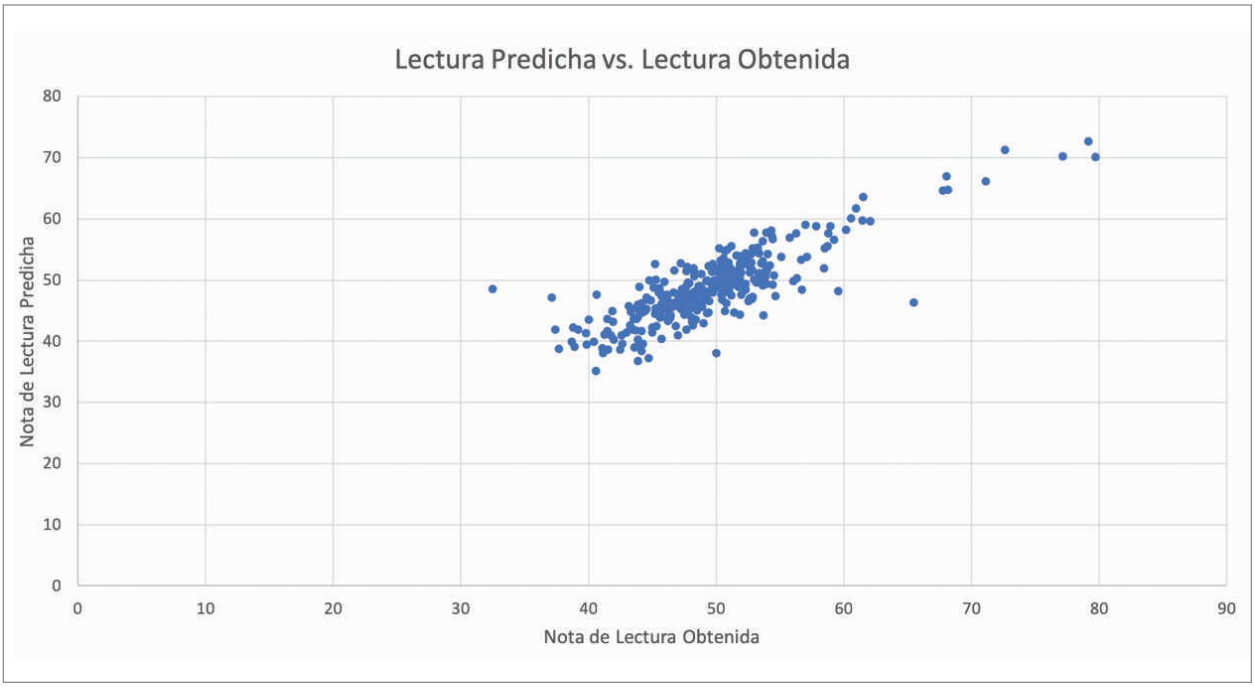

Figura 11. Correlación durante el entrenamiento utilizando como prueba los datos del 2016.

en la Figura 11. En este caso, se utilizaron 308 instancias que generaron un coeficiente de correlación de 0.84 con un error absoluto de 2.53.

Variables con mayor respuesta

Explorando los parámetros del modelo presentado anteriormente, se encontró que las variables que más ayudan a predecir el rendimiento son las siguientes: uso de computadora, tiempo para ir de la casa al establecimiento, repetición de grado, si se asistió a preprimaria o no, grado más alto estudiado por la madre, grado más alto estudiado por el padre, si alguno de los padres asistió a la escuela o no, si el idioma principal en el hogar es el español, cuantos libros ha leído, la frecuencia de lecturas de periódicos, la suma total de periódicos que se lee, el número de periodos semanales de computación y la variable socio-económica.

Valores atípicos

Después de observar los resultados expuestos anteriormente, se observaron al menos 2 valores atípicos. Estos valores corresponden a la nota promedio obtenida por los municipios de San Francisco La Unión, Quetzaltenango y Ocós, San Marcos. Estos dos municipios presentan un caso de estudio particularmente interesante, no solo por ser valores atípicos de modelo, es decir, que el modelo no predijo correctamente la nota, sino por sus diferencias y similitudes. En particular, ambos municipios reportan que el 75 % de sus estudiantes reciben remesas, todos sus estudiantes asistieron a preprimaria, ninguno trabaja actualmente, casi ninguno de sus estudiantes ha repetido grado, todos los estudiantes asisten a una institución del sector oficial y todos cursan el bachillerato. Además, ambos municipios muestran un nivel socio económico similar, medido por la variable explicada anteriormente. En el caso de San Francisco la Unión, el valor de su estatus socio-económico es 9.25 mientras que Ocós reporta un valor de 9. La mayor diferencia entre estos dos municipios es que San Francisco La Unión reporta que todos sus estudiantes son de etnia Maya y ninguno habla español como primer idioma. Mientras tanto, Ocós reporta el 100 % de estudiantes ladinos de los cuales el 100 % habla español como idioma materno. A pesar de las similitudes expuestas anteriormente y de la diferencia principal (que tiende a estar relacionada con un menor rendimiento para el caso de San Francisco La Unión), Ocós tiene un promedio de lectura de 32.5 (que lo coloca en el percentil 0 de todo el país), mientras que San Francisco La Unión tiene una nota promedio de 65.5, colocándolo en el percentil 97 del país.

Lo anterior indica que, a pesar de que las circunstancias socio-económicas de los municipios son iguales, y que la distribución étnica sugiriera mayor dificultad para la evaluación de lectura en español, existe alguna estrategia o circunstancia no capturada por los datos que causa que San Francisco La Unión sea uno de los mejores municipios en términos de lectura del país. Este

resultado, aunque ayuda a disminuir el rendimiento del modelo predictivo, indica que debe realizarse una mayor investigación para entender qué puede explicar este resultado. Se puede especular que en las instituciones de San Francisco La Unión se han implementado estrategias de enseñanza que son más efectivas para el contexto de ellos, que podrían ser utilizadas en otros municipios similares.

Conclusiones y recomendaciones

En este estudio se muestran las relaciones entre factores asociados y el desempeño escolar en las áreas de matemáticas y lectura. El modelo que se presenta enfatiza la importancia de las variables relacionadas al estrato socio-económico, la educación de los padres, y hábitos de lectura relacionados con libros y periódicos. Es importante enfatizar que, a pesar de las limitaciones del análisis presente, se han encontrado señales importantes de diferencias en los desempeños promedio de algunos municipios del país, tal como los casos de Ocós, San Marcos y San Francisco La Unión, Quetzaltenango. Estos a pesar de ser muy similares en cantidad de estudiantes, instituciones y rasgos socio-económicos, rompen con lo esperado tanto positiva como negativamente con respecto a su desempeño. Mas investigación es necesaria para poder explicar estar diferencias y así crear planes de apoyo para las comunidades que más ayuda necesitan.

Las variables que más se relacionan con el rendimiento académico son las que miden el estrato-socioeconómico, la educación de los padres, hábitos de lectura (como lectura de periódicos y libros), el idioma que se habla en el hogar, así como el uso de computadora y el tiempo requerido para llegar al establecimiento. Otra conclusión importante de este estudio es la posible reducción del cuestionario de factores asociados. Actualmente, el cuestionario cuenta con una alta cantidad de preguntas que podría causar fatiga a los estudiantes (Porter et. al. 2004)

Como se discutió anteriormente, una de las mayores limitaciones de este estudio es que el análisis se realizó a nivel municipal, utilizando solamente las variables que se encuentran presentes en todos los años. Subsiguientes investigaciones se concentrarán a nivel de instituciones (en lugar de municipios) para explicar las diferencias a nivel institucional y evitar así los sesgos mostrados en este estudio.

Se recomienda también, que se investiguen métodos alternativos de evaluación para así poder evaluar, de forma independiente la habilidad matemática y la habilidad de lectura, así como a estudiantes cuyo primer idioma no sea el español.

Agradecimientos

Este trabajo fue posible gracias a la cooperación del Centro de Investigaciones Educativas, Jorge Andrés Galvez-Sobral, M.A. y a todo el equipo.

Bibliografía

Porter, S.R., Whitcomb, M.E., Weitzer, W.H. (2004) *Multiple surveys of students and survey fatigue* New Directions for Institutional Research, 121: 63-73. doi:10.1002/ir.101

Stephens, N.M., Fryberg, S.A., Markus, H.R., Johnson, C S., Covarrubias, R. (2012) *Unseen Disadvantage: How American Universities' Focus on Independence Undermines the Academic Performance of First-Generation College Students* Journal of Personality and Social Psychology. 102 (6) 1178-1197. doi: 10.1037/a0027143

R Core Team (2016) *R: A language and environment for statistical computing R Foundation for Statistical Computing,* Vienna, Austria. URL https://www.R-project.org/

Van der Berg, S. (2008) *How effective are poor schools? Poverty and educational outcomes in South Africa.* Studies in Educational Evaluation, 34 (3): 145-154.

Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis* Springer-Verlag New York