Data Mining al servicio de la educación

Un modelo para la detección temprana del mal uso de la tecnología en los niños

Fonzo, Natalia; Kaucher, Martín y Molin, Gino

Big Data
Universidad de San Andrés
Primavera 2018

13 de diciembre de 2018

1 Introducción

En las últimas décadas el rol de la tecnología se ha ido cobrando relevancia en todos los ámbitos de la vida. En la educación, tanto educadores como *policy makers* han comenzado a preguntarse cómo la tecnología puede intervenir en la educación y ayudar a mejorarla.

Tradicionalmente, la introducción de la tecnología en la educación ha tenido dos objetivos. Por un lado, mejorar la pedagogía y facilitar el aprendizaje a través de incorporar los avances tecnológicos a las técnicas educativas en pos de la didáctica, lo que se conoce como "instrucción asistida por computadora"[1]. Por otro lado, la tecnología también adquirió relevancia propia al poder ser utilizada no sólo como una herramienta o medio de llegada al alumno, sino como un fin de enseñanza en sí. A este segundo objetivo se lo conoce en la literatura como "entrenamiento en informática".[2]

Sin embargo, hoy en día la tecnología no sólo entra al aula a través de dichos medios, sino que, además, se encuentra presente en la cotidianidad de los propios alumnos, en sus comunicaciones e interacciones y en el acceso a cualquier información disponible en Internet. Justamente, el rol de estas nuevas tecnologías, definidas como "tecnologías de la información y de la comunicación" (TICs), han sido últimamente el foco de atención en el campo de la educación por la oportunidad de mejorar la calidad educativa que representan.

Las TICs pueden apoyar y mejorar los procesos educativos volviendo a las escuelas promotoras del uso de estas tecnologías que amplíen el repertorio de prácticas, tanto en el contexto escolar como fuera de éste (Tófalo, 2017). Contando con estos dispositivos y acceso a internet, los alumnos pueden adquirir conocimientos más allá de lo que se enseñe en clase e integrar distintos recursos que antes estaban normalmente separados (libros, escritura, grabaciones, bases de datos, etc.), y así, extender los tiempos y lugares donde el aprendizaje ocurre (OECD, 2015).

El impacto de las TICs en la mejora de la calidad educativa da cuenta de la complejidad de la propia problemática. Es difícil hablar del impacto de las TICs, ya que este concepto involucra un muchas tecnologías distintas, con características y potenciales beneficios específicos que, en tanto instrumentos, pueden ser usados de muchas maneras (Sletten, 2010). De hecho, en la literatura disponible, de la pregunta sobre el impacto de las TICs en los aprendizajes se desprenden tres preguntas específicas: sobre los tipos de uso de las TICs y su impacto en los aprendizajes; sobre las condiciones de uso y su impacto sobre el aprendizaje; y sobre quién usa las TICs y su impacto sobre el aprendizaje (Claro, 2010: 6).

Estas nuevas herramientas, por disruptivas y útiles, han vuelto compleja su asimilación en el tradicional ambiente educativo y representan un desafío para quienes se encargan de diseñar la política educativa de un país o ciudad. En vistas de ayudar de alguna manera en este proceso, nos propusimos utilizar técnicas de Machine Learning para tratar de brindarle a los educadores y *policy makers* una herramienta que sirva, en el marco de políticas integrales de inclusión de las TICs al aula, para la detección temprana de aquellos alumnos que presenten dificultades a la hora de integrar las TICs a su educación de manera autónoma y que creamos necesiten una capacitación.

2 Datos

Como ya anticipamos, para este trabajo utilizamos los resultados de las pruebas APRENDER 2016 para alumnos de sexto grado de primaria [3]. Estas consisten en una encuesta de datos personales y dos evaluaciones -de Lengua y Matemática- que, en el marco del llamado "Sistema de Evaluación Nacional de la Calidad y Equidad Educativa", permite al Gobierno de la Nación Argentina monitorear los servicios educativos de las instituciones públicas y privadas del país.

La primera encuesta tiene una serie de noventa y seis preguntas que, además de solicitar el género y la edad, indagan sobre el hogar, la familia y el nivel socioeconómico del alumno, sobre su

[1] Angrist y Lavy (2002)
[2] Dynarski et al (2007)
[3] Base de datos disponible en: https://www.argentina.gob.ar/educacion/aprender2016/bases-de-microdatos

desempeño personal y académico en la escuela, sobre su tiempo libre y, también, sobre su consumo de tecnología. Esta última sección contiene cuarenta preguntas acerca del acceso y uso de la tecnología: si en su casa tienen WiFi y acceso qué dispositivos, a qué edad empezó a usarlos, si tiene celular propio, para qué usa en su tiempo libre mayormente la computadora y qué uso le dan en la escuela, entre otras.

Previo a cualquier análisis, tuvimos que ocuparnos de este último *set* de variables. Dado que nuestro objetivo es predecir si un alumno utiliza responsablemente o no la tecnología, y esta clasificación no existe como tal en la base de datos, tuvimos que construirla. Para ello, comenzamos por definir qué entendemos por responsabilidad a la hora de usar una computadora. Al respecto, decidimos considerar como no responsable cualquier uso que no tuviera como propósito directo servir a la educación del alumno.

Luego, identificamos aquellas preguntas a las que un alumno que usa "bien" la tecnología -bajo el mencionado criterio- respondería afirmativamente. Por ejemplo, preguntas tales como "*¿usás con frecuencia tu computadora o celular para leer artículos o libros digitales?*" fueron seleccionadas y otras tales como "*¿usás con frecuencia tu computadora o celular para seguir a personas conocidas?*", no. A continuación, corroboramos que la respuesta afirmativa a dichas preguntas coincidiera con un 1 en la base y la respuesta negativa, con un 0. En el caso de que algún alumno no hubiera contestado a una determinada pregunta, se le asignó un 0 a esa observación: ante la duda, presumimos que el alumno no usa bien la tecnología y que, potencialmente, necesita ayuda. En otras ocasiones tuvimos que crear esta variable binaria: por ejemplo, cuando preguntas como "*cuando usás la computadora en la escuela, ¿sentís que aprendés más?*" tenían respuestas múltiples "siempre" / "a veces" / "nunca", creamos la variable binaria tomando -según el caso- "siempre" y "a veces" como un 1 y "nunca" como un 0.

Una vez hecho esto, obtuvimos un *subset* de veinte variables que toman valor 1 si el alumno utiliza responsablemente la tecnología y 0, si no. Por último, creamos Z, la variable a predecir. Z toma valor 0 si la suma de las veinte variables preseleccionadas es mayor a 10 -es decir, si el alumno respondió afirmativamente a más de la mitad de las preguntas que indicarían un buen uso de su computadora- y 1 si la suma es menor o igual a 10 -y consideramos que el alumno debiera tener una capacitación-.

Es importante notar que, bajo el criterio empleado, responder sistemáticamente de manera afirmativa a preguntas como "*¿usás con frecuencia tu computadora o celular para ver videos?*" o "*¿usás con frecuencia tu computadora o celular para mandar mensajes a tus amigos?*" no influye de manera alguna en que el estudiante sea considerado irresponsable en el uso de la tecnología. Por el contrario, responder sistemáticamente que no a preguntas tales como "*¿usás con frecuencia tu computadora o celular para buscar información para la escuela?*", sí influye.

Finalmente, se eliminaron de la base las veinte variables que se emplearon en la construcción de Z, siendo noventa y uno las variables restantes, entre las que se incluyen tanto las remanentes de la encuesta personal, como los resultados de las dos evaluaciones y otros datos completados por la organización -no el alumno- tales como la provincia, si la escuela es de gestión pública o privada, si el ámbito es rural o urbano, y otros índices. Además, se habían eliminado previamente las observaciones que no hubieran respondido la encuesta personal, quedándonos así 474.104 observaciones en lugar de las 561.950 originales. De las observaciones conservadas, hay 339.739 observaciones tales que $Z_i = 1$ y 134.365 tales que $Z_i = 0$.

3 Metodología

Una vez definido nuestro objetivo y creada la variable a predecir, planteamos el modelo para nuestro predictor: se trata de un Lasso-Logit. Logit es un modelo de regresión que permite estimar la probabilidad condicional a X de que una observación i pertenezca a una determinada clase de la categoría binaria $Z \in \{0,1\}$:

$$p = P(Z_i = 1/X_i = x_0). \tag{1}$$

Utilizando la estimación del Logit para p y el clasificador de Bayes de penalización simétrica, que penaliza el error tipo I tanto como el error tipo II y que le asigna a cada predicción $j \in \{0,1\}$ un riesgo esperado de la forma

$$R(j) = (1-p)\mathbb{1}[j = Z_i] + (p)\mathbb{1}[j \neq Z_i], \tag{2}$$

es posible minimizar el Error Cuadrático Medio del predictor fuera de la muestra de entrenamiento a partir de minimizar el riesgo de Bayes. Puede demostrarse fácilmente que esto equivale a clasificar las observaciones bajo la siguiente regla:

$$j = \hat{Z}_i = \begin{cases} 1 & \text{si } p \geq 0.5 \\ 0 & \text{if } p < 0.5 \end{cases} \tag{3}$$

En particular, la estimación de p hecha por Logit viene dada por:

$$\hat{p} = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}} \tag{4}$$

donde X_i es el vector de predictores correspondientes a la observación i y β es el vector de coeficientes a estimar por Máxima Verosimilitud. Ahora bien, la cantidad de predictores k en el vector X_i que utilizará el modelo quedará determinada por LASSO. LASSO es un método de selección de variables que tiene por objetivo mejorar la precisión de modelos predictivos. Para lograrlo, este método determina los coeficientes del predictor en cuestión, β en nuestro caso, a partir del siguiente problema de optimización:

$$\min_{\beta} R_l(\beta) = \sum_{i=1}^{n} [Z_i - \hat{Z}_i(\beta)]^2 + \lambda \sum_{s=2}^{k} |\beta_s| \tag{5}$$

donde n es la cantidad de observaciones, k la cantidad de variables en el vector X_i, Z_i la variable (categoría) a predecir, $\hat{Z}_i(\beta)$ la predicción hecha por Logit y λ un número arbitrario no negativo. Resolver este problema equivale a minimizar el Error Cuadrático Medio dentro de la muestra de entrenamiento del modelo Logit -con el primer término- pero penalizando la cantidad de variables incluidas -con el segundo término- para así evitar *overfit*. Es importante destacar que, dado que λ -peso relativo que tiene el problema del ajuste externo en relación al de ajuste interno- será elegido por Cross Validation, en definitiva, minimizar $R_i(\beta)$ equivale a minimizar el Error Cuadrático Medio computado -por Cross Validation- sobre la muestra de test.

El vector de coeficientes β arrojado por LASSO asignará peso nulo a algunas variables del vector X_i y peso positivo a otras. De esta manera, son elegidas las variables necesarias para la estimación $\hat{p}$ y, consecuentemente, para la predicción $j = \hat{Z}_i$. Así, queda determinado el predictor que buscábamos.

4 Resultados

En esta sección realizaremos una exposición de los resultados obtenidos.[4] Procederemos, primero, a una descripción el modelo resultante y, luego, a analizar su performance.

4.1 Modelo seleccionado

Una vez seleccionado el *lambda* (λ=0,000344) que minimiza el error cuadrático medio computado por *Cross Validation* de nuestro poredictor (ver Figura 1), utilizamos la base de prueba para computar el error cuadrático medio que servirá para contrastar la precisión. Este es: 0,3.

[4]El código para correr el modelo en R está disponible en: https://codeshare.io/5vmY3W

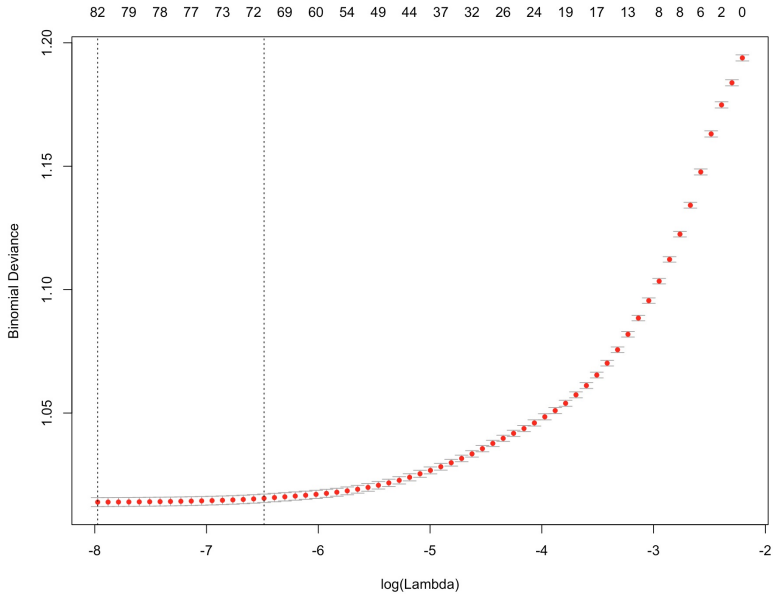

Figure 1: Error Cuadrático Medio computado por *Cross Validation* para los distintos modelos asociados a cada penalidad *lambda*. La línea punteada vertical señala el *log(Lambda)* asociado a nuestro predictor.

Ahora bien, como explicamos previamente en Metodología, asociado a dicho λ LASSO devuelve un vector de coeficientes β. Este tiene noventa y uno coeficientes correspondientes a las noventa y una variables consideradas posibles predictores, de los cuales sólo treinta y tres son distintos de cero. Entre las treinta y tres variables seleccionadas por la predicción, se encuentran variables de índole personal, otras vinculadas al *mal* uso de dispositivos tecnólgicos -según fue definido más arriba-, disponibilidad de dispositivos tecnológicos en el hogar, el desempeño escolar y la conformidad del alumno en el colegio.

A continuación, vemos la evolución de los coeficientes para distintos *lambda*.

4.2 Performance

Veremos ahora distintas medidas del desempeño de nuestro modelo como predictor.

Primeramente, como ya mencionamos, el Error Cuadrático Medio computado sobre la base de test es 0,3.

En segundo lugar, computamos la matriz de confusión con la base de test, que nos permite comparar la predicción hecha por nuestro modelo (columnas) con los verdaderos valores que toma la variable $\hat{Z}_i$ (filas).

Table 1: Matriz de confusión

	0	1	Total
0	10.968	55.996	66.964
1	7.851	162.237	170.088
Total	18.819	218.233	237.052

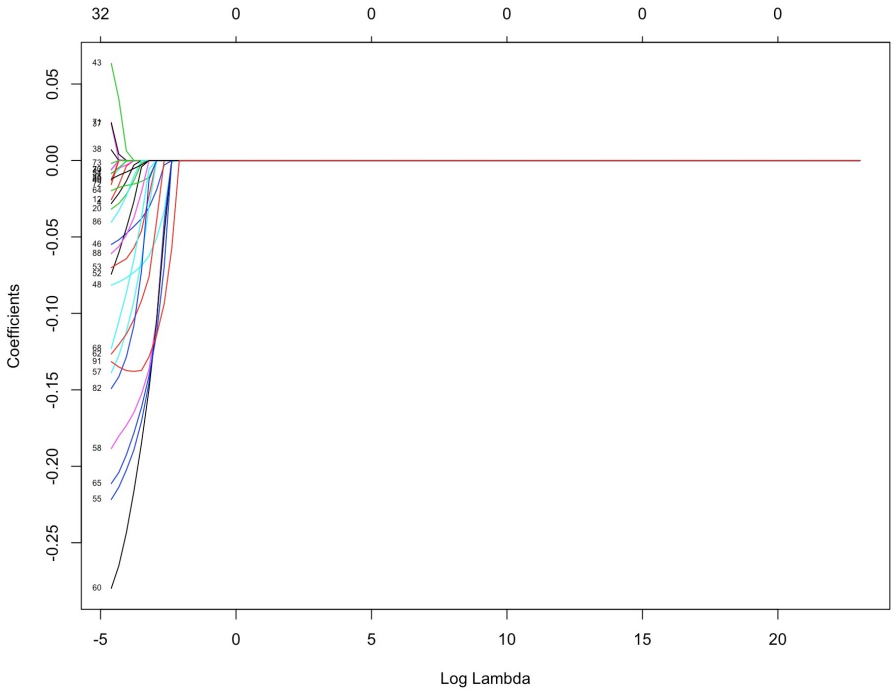

Figure 2: Coeficientes seleccionados por LASSO para distintas penalidades *lambda*.

En la Tabla 1 vemos, por ejemplo, que nuestro modelo predijo correctamente 162.237 de 218.233 observaciones que no utilizan responsablemente la tecnología y 10.968 de 18.819 observaciones que sí utilizan responsablemente la tecnología.

Con estos resultados, calculamos los indicadores de la tabla a continuación.

Table 2: Medidas de desempeño

Precisión	73%
Sensibilidad	95%
Especificidad	16%

La precisión nos dice que nuestro modelo clasifica correctamente cualquier observación un 74% de las veces -valor consistente con el Error Cuadrático Medio de 0,3 previamente hallado. Por otro lado, la sensibilidad -la proporción de verdaderos positivos en el total de positivos- nos dice que predecimos correctamente a los alumnos que usan mal la tecnología un 95%. Sin embargo, la especificidad, es decir, el porcentaje de verdaderos negativos sobre el total de negativos, indica que los casos en que los alumnos utilizan bien la tecnología son predichos correctamente un 16% de las veces.

Considerando que la regla de clasificación utilizada -clasificador simétrico de Bayes- minimiza el Error Cuadrático Medio en la muestra de test y que, a su vez, el modelo resultante tiene tun bajo error tipo II = 5% -que es nuestro error a controlar, pues queremos asegurarnos de detectar lo mejor posible a los alumnos que necesitan ayuda-, entonces podemos pensar que nuestro predictor es lo suficientemente potente para nuestro propósito y que no es necesario modificar el *threshold* de la clasificación y empeorar la precisión. Tenemos una combinación óptima -en todo sentido- de error tipo I y error tipo II.

Más aún, la curva ROC de la Figura 3, que ilustra las posibles combinaciones error tipo I y error tipo II para distintos *thresholds* -o ponderación de los dos errores, que en nuestro caso es simétrica-, indica que el modelo obtenido es relativamente potente. El área bajo la curva es $0,75$: nuestro predictor tiene una potencia intermedia entre un modelo que predice correctamente la totalidad de las observaciones y uno que las clasifica aleatoriamente.

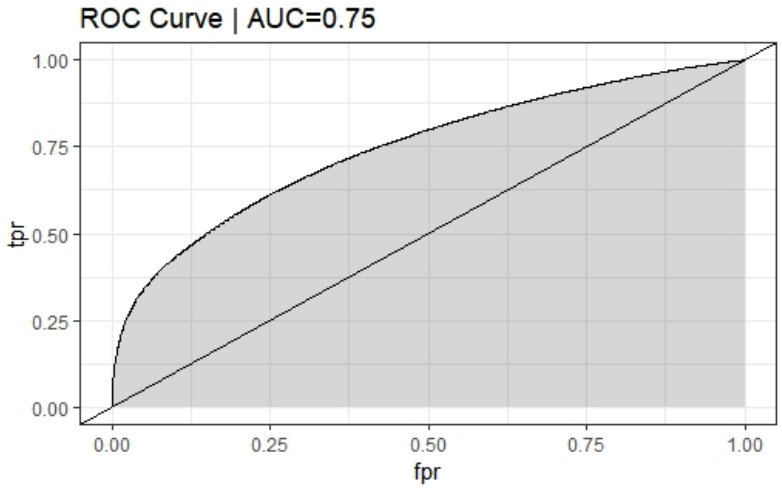

Figure 3: Para distintos *thresholds*, la combinación de $1 - ErrorTipoII$ (eje y) y $ErrorTipoI$ (eje x) de nuestro predictor.

5 Conclusión

Como mencionamos previamente, la aparición de las TICs en la vida cotidiana de las personas y, especialmente, de los alumnos representa un desafío para quienes se encargan de diseñar la política educativa de un país. Se trata de herramientas que, por disruptivas y útiles, han vuelto compleja su asimilación en el tradicional ambiente educativo. En vistas de poder ayudar de alguna manera en este proceso, nos propusimos desarrollar un modelo predictivo que, gracias a que identifica a aquellos alumnos que presentan dificultades a la hora de integrar las TICs a su educación de manera autónoma, permita a las autoridades de las instituciones educativas y, consecuentemente, a los *policy makers* diseñar políticas educativas en torno a las nuevas tecnologías de comunicación e información.

Hemos mostrado que logramos crear una herramienta que cumple con nuestro propósito de manera efectiva: el predictor logra clasificar correctamente a los alumnos el 73% de las veces y es particularmente bueno para identificar a aquellos que necesitan apoyo en el uso de la tecnología.

Sin embargo, cabe destacar que nuestro modelo presenta dos limitaciones dadas por la propia construcción del mismo que y que proponemos como futuros puntos de desarrollo para este tipo de herramientas. En primer lugar, la variable a predecir Z_i fue construida a partir de una selección no arbitraria de variables de la encuesta. En este sentido, la clasificación pierde sentido si el criterio que utilizamos en la construcción de Z_i no se alineara con el propósito de la herramienta en cuestión o la política educativa en la que esta se inserte. Se podría, por ejemplo, utilizar una técnica como Componentes Principales para detectar dónde está la mayor variabilidad en las variables que dan cuenta de cómo usan la tecnología los alumnos y que se emplean en la construcción de Z_i o bien definir un nuevo criterio.

En segundo lugar, nuestro modelo utiliza como predictores datos de la encuesta APRENDER 2016. Proponemos como punto a mejorar en nuestro trabajo, el desarrollo de un modelo que utilice, como

predictores, datos que se puedan encontrar por fuera de ésta encuesta, por ejemplo, datos sobre los alumnos disponibles por las autoridades de las instituciones educativas.

6 Bibliografía

Angrist, Joshua; Lavy, Victor. New evidence on classroom computers and pupil learning. Economic Journal, London, v. 112, n. 482, p. 735-765, 2002.

Baker, R. S., Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. JEDM— Journal of Educational Data Mining, 1(1), 3-17.

Carrillo, P. E., Onofa, M., Ponce, J. (2011). Information technology and student achievement: Evidence from a randomized experiment in Ecuador.

Dynarski et al (2007). Effectiveness of reading and mathematics software products: Findings from the first student cohort.

Ravalli, M. J., Paoloni, P. (2016). Global Kids Online Argentina: Research study on the perceptions and habits of children and adolescents on the use of technologies, the internet and social media.

Rivas, A. (2018). Un sistema educativo digital para la Argentina. CIPPEC

Román, M., Murillo, F. J. (2014). Disponibilidad y uso de TIC en escuelas latinoamericanas: incidencia en el rendimiento escolar. Educação e Pesquisa, 40(4), 879-895.

Sletten, J. (2010). A systematic review of ICT interventions in learning. Master thesis ICT and learning Stord/Haugesund University College

Tófalo, A. (2017). Aprender 2016 - Acceso y uso de TIC. Serie de informes temáticos. Buenos Aires: Secretaría de Evaluación Educativa del Ministerio de Educación de la Nación

7 Anexo I

Predictor	Coeficiente
Intercepto	2,45
¿Cuántas habitaciones tiene la casa en la que vivís, sin contar la cocina y el baño?	-0,03
Pensando en lo que hiciste la semana pasada ¿Cultivaste, cosechaste en la huerta, trabajaste la tierra, o cuidaste animales de granja para utilizar como consumo en tu casa?	-0,03
En la escuela te sentís: Aburrido	-0,03
En la clase de lengua: Entiendo rápido	-0,01
La clase de lengua: Es muy entretenida	-0,01
¿Cómo escribís? (1=muy bien, 4=mal)	0,02
¿Cómo resolvés problemas o cuentas de matemática?	0,01
Cuando están trabajando en el aula. ¿Las maestras y maestros se enojan con ustedes?	-0,01
¿Te llevás bien con tus compañeros? (1= con todos, 4=ninguno)	0,06
¿Entre tus compañeros hay burlas o peleas? (1=muchas veces)	-0,05
Frases sobre tu escuela. Marcá tu grado de acuerdo...Mi escuela es un lugar donde me siento solo (1=de acuerdo)	-0,08
En el último mes, ¿hiciste alguna de estas actividades en tu tiempo libre, fuera del horario escolar? Leíste un libro	-0,01
En el último mes, ¿hiciste alguna de estas actividades en tu tiempo libre, fuera del horario escolar? Te reuniste con amigos	-0,08
En el último mes, ¿hiciste alguna de estas actividades en tu tiempo libre, fuera del horario escolar? Estudiaste algún idioma	-0,07
En el último mes, ¿hiciste alguna de estas actividades en tu tiempo libre, fuera del horario escolar? Fuiste a algún espectáculo o exposición (cine, recital, teatro, museo...)	-0,01
¿En tu casa tenés: Tablet	-0,22
¿En tu casa tenés: Smartphone	-0,15
¿En tu casa tenés: Consola de juegos (PlayStation, Wii o X-Box o similar)	-0,18
¿Para qué actividades usás con mayor frecuencia tu computadora o celular?: Mandar mensajes a tus amigos	-0,29
¿Para qué actividades usás con mayor frecuencia tu computadora o celular?: Seguir a personas conocidas	-0,13
¿Para qué actividades usás con mayor frecuencia tu computadora o celular?: Navegar por redes sociales (Facebook, Google+, Twitter u otras)	-0,03
¿Para qué actividades usás con mayor frecuencia tu computadora o celular?: Comunicarte con familiares	-0,21
¿Para qué actividades usás con mayor frecuencia tu computadora o celular?: Comunicarte con personas que no conozcas	-0,12
¿Qué tipo de actividades te proponen hacer con la computadora cuando estás en el horario de clases?: Jugar con videojuegos educativos	-0,01
¿Qué sentís cuando usás la computadora en la escuela?: La escuela me resulta más entretenida	0,03
Resolver las pruebas de APRENDER, ¿te fue fácil o difícil?: Prueba de Lengua (creciente en dificultad)	-0,01
Resolver las pruebas de APRENDER, ¿te fue fácil o difícil?: Prueba de Matemática	
Sector de Gestión (1=Estatal, 2=Privada, 3=Sin Datos)	-0,16
Cuartil de alumnos según el porcentaje de hogares en el radio de la escuela en estrato socioeconómico muy bajo Total País	-0,04
Índice socioeconómico del alumno	-0,06
Índice socioeconómico del alumno ponderador lengua	-0,01
Índice de Clima escolar	-0,14