Universidad de la República

Facultad de Ciencias Económicas y de Administración

Trabajo final de grado para obtener el título de

Licenciado en Estadística

Aprendizaje estadístico aplicado para potenciar la enseñanza de inglés en primaria: El caso de Ceibal en Inglés en Uruguay

Bruno Tancredi

Tutores:

Ignacio Alvarez-Castro

Natalia da Silva

Montevideo, Uruguay

Junio 2024

El tribunal docente integrado por los abajo firmantes aprueba el trabajo final de grado:

Aprendizaje estadístico aplicado para potenciar la enseñanza de inglés en primaria: El caso de Ceibal en Inglés en Uruguay

Bruno Tancredi

Tutores:

Ignacio Alvarez-Castro

Natalia da Silva

Licenciatura en Estadística

Calificación: 12

Fecha: 6/6/2024

Tribunal:

Profesora Cecilia Marconi _____

Profesor Juan José Goyeneche _____

Profesora Natalia da Silva _____

Resumen ejecutivo

En este trabajo se presentan modelos predictivos que vinculan el uso de los estudiantes de la plataforma educativa del programa Ceibal en Inglés, en particular Little Bridge, con el rendimiento en las pruebas adaptativas de Inglés. Los modelos predicen la performance en las pruebas adaptativas de Inglés en base a distintas medidas de uso que hacen los estudiantes de la plataforma Little Bridge, así como información socio demográfica de los estudiantes y del centro al que pertenecen. La implementación de los modelos predictivos se hizo utilizando intensivamente la biblioteca `tidymodels` que permite un flujo de trabajo estandarizado y coherente para ajustar y comparar varios métodos de aprendizaje estadístico al mismo tiempo simplificando la reproducibilidad y coherencia en la implementación.

Palabras claves: Bayesian Additive Regression Trees, Modelos mixtos, Aprendizaje supervisado, Educational data mining.

Índice

1. Introducción

Este trabajo fue desarrollado en el marco del Fondo Sectorial "Inclusión Digital: Educación con Nuevos Horizontes" - 2020 - Modalidad A (duración 2 años), desarrollado por el Instituto de Estadística de la Facultad de Ciencias Económicas y de Administración siendo Natalia da Silva, responsable académica del mismo. El objetivo general del proyecto es desarrollar herramientas estadísticas para la evaluación y monitoreo de plataformas educativas como CREA que colaboren en la elaboración de políticas educativas y la toma de decisiones basadas en evidencia. Dentro de este objetivo general, uno de los objetivos específicos es vincular el uso de las plataformas educativas con el desempeño de los estudiantes. En este trabajo el vínculo entre uso y desempeño se enfoca en el aprendizaje y enseñanza de Inglés dentro del programa Ceibal en Inglés.

Ceibal es el centro de innovación educativa de Uruguay que busca integrar tecnologías digitales en la educación para mejorar el aprendizaje, fomentar la innovación, inclusión y crecimiento personal. Ceibal en Inglés (CEI) se enfoca en la enseñanza de inglés en educación pública de Primaria y Media (Secundaria y UTU). La enseñanza de inglés es mediante videoconferencias donde los y las estudiantes participan de clases semanales con docentes de lengua extranjera. Para el desarrollo de la clase, docentes de aula y docentes a distancia trabajan colaborativamente a través de la plataforma CREA, pero a partir del 2021 se incorpora una nueva plataforma específica para el aprendizaje de inglés llamada Little Bridge (LB). En LB los y las estudiantes realizan actividades asignadas por docentes a distancia. El presente documento se enfoca en cómo combinar información de uso de la plataforma Little Bridge del programa Ceibal en Inglés en el año 2021 para predecir el desempeño de los alumnos en las pruebas

adaptativas de inglés para primaria (4.º, 5.º y 6.º año).

En este trabajo se exploran diversos métodos de aprendizaje estadístico supervisado, empleando varias herramientas computacionales con el propósito de alcanzar el objetivo antes mencionado. En la Sección 2, se ofrece una descripción detallada de la transformación realizada a los datos originales proporcionados por Ceibal, que son los insumos fundamentales para los distintos modelos, así como un análisis descriptivo de los mismos. La Sección 3 se centra en desarrollar la metodología esencial para llevar a cabo este trabajo. Se hace hincapié en `tidymodels` (Kuhn & Wickham, 2020), dada su amplia utilización a lo largo del proyecto, y en el modelo de árboles de regresión aditivos bayesianos (BART, por sus siglas en inglés), debido a los buenos resultados obtenidos con este modelo. Se realiza una breve reseña del modelo original presentado en (Chipman et al., 2010), junto con algunas extensiones. En la Sección 4 se presentan los principales resultados de los modelos implementados, así como indicadores de rendimiento.

2. Datos de uso de plataformas LB y CREA

Los datos utilizados para este trabajo fueron brindados por Ceibal, los mismos corresponden al año 2021, que es el primer año en que funciona la plataforma LB. Estos datos se entregaron en formato `CSV` y en formato `XLSX`. El Cuadro 1 muestra los archivos de datos recibidos, una descripción general y la cantidad de filas y columnas en cada uno. Estos archivos varían en tamaño, siendo el más grande de ellos el que tiene 4,072,559 filas. Debido al gran volumen de datos se decidió utilizar la librería `data.table` (Dowle & Srinivasan, 2023) en conjunto con `dtplyr` (Wickham et al.,

2023) debido a la sencillez de su sintaxis y eficiencia computacional. La primera etapa consistió en la transformación de los datos, en el que se pasó de tener los archivos separados a un solo conjunto de datos utilizado para modelar. Antes de la etapa de modelado se hizo un análisis descriptivo de este conjunto de datos final.

Cuadro 1: Descripción de los archivos disponibles para 2021

Archivo	**Descripción**	**Tamaño**
act_crea_cei_C_ID_ALEAT	Actividad del alumno en la plataforma CREA.	4072559 filas, 21 columnas.
actividad_diaria_est_lb_2021_fin	Actividad del alumno en la plataforma Little Bridge.	1030280 filas, 19 columnas.
mensajes_diarios_est_lb_2021_fin	Logs de mensajes recibidos y enviados.	585082 filas, 15 columnas.
docente_activiad_lb_2021_EMI fin	Lecciones asignadas por los docentes a las clases.	84449 filas, 7 columnas.
TAI-43032022517 - Evaluaci¢n Adaptativa de Ingles 2021	Pruebas adaptativas finales realizadas por los alumnos.	35032 filas, 13 columnas.
TAI-31302022517 - Evaluaci¢n Adaptativa de Ingles 2021 Reading	Reading final realizado por los alumnos	33309 filas, 13 columnas.
Base_Cruce_Lesson_Activity	Indica el nombre de la actividad y a qué lección pertenece	915 filas, 3 columnas.

2.1. Transformación de los datos

Con el objetivo de facilitar la interpretación de los datos y crear una base para poder construir sobre estos, se optó por representar los datos mediante un modelo entidad-relación. En la Figura 1 se presenta el diagrama correspondiente al modelo entidad-

relación. Un modelo entidad-relación es un modelo conceptual (P. P.-S. Chen, 1976) en el que se trata de representar la realidad a través de entidades y sus relaciones. Para representar al modelo entidad-relación se utiliza un diagrama de entidad-relación en el que las entidades quedan representadas por rectángulos, las relaciones entre entidades quedan representadas por rombos, las entidades débiles quedan representadas con una flecha en la dirección de la entidad fuerte (una entidad débil es una entidad que no existe sin su entidad fuerte, e.g. una instancia de prueba no existe si el alumno que la dio no existe). A su vez, en dicho diagrama se indica la cardinalidad de la relación, donde un N del lado de una entidad **A** indica que la otra entidad **B** se puede relacionar con N entidades **A** y un 1 indica que **B** se relaciona con a lo sumo un único **A**.

Para lograr la representación se llevó a cabo un proceso de limpieza y manipulación de datos en el que se requirió tomar ciertos criterios. Estas decisiones fueron necesarias para poder representar los datos de manera que puedan ser usados para un modelo predictivo. A continuación, se describen los criterios empleados para la representación de cada entidad.

Alumnos

Cada alumno está caracterizado por un identificador, así como por el centro, grado, departamento, zona y contexto sociocultural a los cuales pertenece. Si bien en la realidad entendemos que existe la posibilidad de que estas variables cambien a lo largo del tiempo, en los datos proporcionados no se refleja ningún cambio. Se filtró a los alumnos que no pertenecen a ninguna clase, en conjunto también se filtraron a los alumnos que no realizaron ninguna actividad o que si hizo actividades estas no figuran en el conjunto de datos brindado por Ceibal, ya que el objetivo del trabajo es

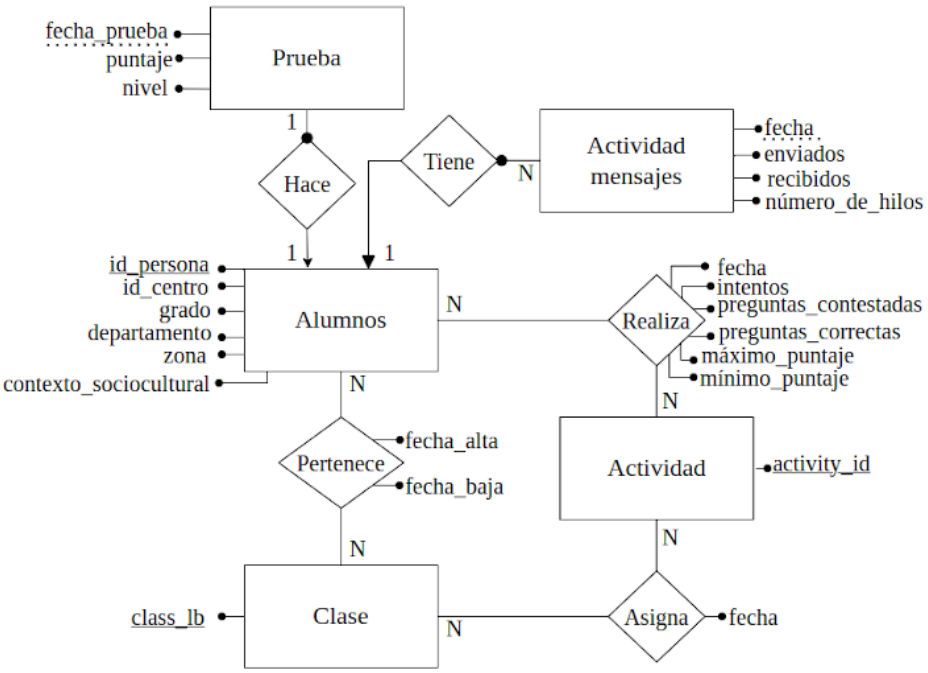


Figura 1: Diagrama entidad-relación. Las entidades se representan con rectángulos, las relaciones con rombos y las entidades débiles con una flecha en la dirección de la entidad fuerte.

vincular el uso de la plataforma con el resultado final obtenido. De los 71,222 alumnos originales se conservan 45,485.

Prueba

Una prueba queda representada por la fecha en la que se tomó, el puntaje obtenido y el nivel. El puntaje máximo registrado fue de 900 y el mínimo de 225.21. El nivel se calcula a partir del puntaje, utilizando unos puntos de cortes ya establecidos. Debido a un cambio en los puntos de corte en el año 2022, la variable nivel fue recalculada para los datos del año 2021. La Tabla 2 muestra los puntajes asociados a cada nivel.

Nivel	Puntos
Pre-A.1	< 372.9
A1.1	$372.9 - 444.4$
A1.2	$444.4 - 495.5$
A2.1	$495.5 - 527.9$
A2.2	$527.9 - 599.4$
B1	> 599.4

Cuadro 2: Puntaje asociado a los niveles.

En el conjunto de datos originales, algunos estudiantes tienen múltiples registros de evaluaciones. Se estableció el criterio de seleccionar únicamente la primera evaluación realizada por cada estudiante. En casos en los que se realizaron dos intentos en el mismo día, se eligió el intento con el puntaje más bajo.

Actividad mensajes

La entidad de actividad de mensajes queda determinada por la fecha, la cantidad de mensajes enviados, recibidos y la cantidad de hilos. Se registran los días que hubo un mensaje recibido o uno enviado. Muy pocas acciones fueron requeridas, únicamente

se llevó a cabo la sustitución de los valores nulos por 0 y se designó el primero de marzo de 2021 como fecha de actividad para estos casos. La actividad de mensajes queda determinada por el alumno del que es la actividad y la fecha de la actividad.

Actividad

La entidad de actividad cuenta únicamente con su identificador. Las actividades son partes de las lecciones, debido a que no contamos con información sobre dificultad de la lección u otra información de valor sobre la misma, se decidió excluirlas del modelo. Los docentes asignan lecciones a la clase. Dado que una actividad puede asociarse con más de una lección, se registra únicamente la primera asignación. Es importante señalar que la asignación por parte del docente se realiza a nivel de clase, no a nivel de alumno. A su vez los alumnos realizan las actividades, para cada actividad que realizan se guarda las veces que lo hizo, las preguntas que contestó, cuantas de estas fueron correctamente respondidas y el puntaje máximo entre los intentos y el mínimo (si el o la estudiante intentaron una sola vez, el puntaje máximo y el mínimo coinciden).

Clase

Una clase queda representada por su identificador. Los alumnos pueden cambiar la clase a lo largo del tiempo. Dado que no se dispone de la fecha exacta en que un alumno fue inscripto en una nueva clase, se consideró como fecha de alta el primer registro de actividad del alumno en dicha clase y como fecha de baja el día anterior. Un estudiante no pertenece simultáneamente a dos clases durante el mismo periodo.

2.2. Conjunto final de datos

Basándonos en el modelo entidad-relación propuesto, generamos tres conjuntos de datos finales, datos de alumnos, datos de alumnos de sexto año y datos de clase. El primero y el segundo tienen a los alumnos como unidad de análisis, mientras el tercer conjunto de datos tiene a la clase como unidad de análisis.

Datos de alumnos

Este conjunto se representa en formato longitudinal, mostrando por mes los distintos indicadores de uso, junto con la clase a la que pertenecen en cada mes y los datos propios de la entidad `Alumno`. Es importante señalar que las variables de la entidad `Alumno` no están vinculadas al mes. Además, los conjuntos cuentan con las variables a predecir, es decir, los puntos y el nivel obtenido en la prueba por el estudiante. Se le agregó un indicador de la actividad de los estudiantes en la plataforma CREA. El indicador queda determinado por $\frac{log(1+sum(actividad\_registrada))}{log(1+dias\_del\_mes)}$, fue ideado con el objetivo de que el indicador sea 1 cuando la cantidad de actividades es igual a la cantidad de días del mes, de esta forma se podría leer que los alumnos hicieron al menos una actividad por día de forma promedio. Cabe aclarar que no tenemos información de dichas actividades para todos los estudiantes. Para la creación de este conjunto se excluyó a los estudiantes que no hicieron la prueba. Debido a que se querían construir indicadores de uso relacionados al compromiso del estudiante con la materia, se descartaron a los alumnos que no tienen actividades asignadas. El conjunto es utilizado en el modelo 4.1.

Datos de alumnos de sexto grado

El conjunto sigue la misma estructura que el primero, exceptuando el indicador de CREA, y tiene únicamente a los alumnos de sexto grado. No se utiliza el indicador de CREA debido a que solo se cuenta con registros en la plataforma de CREA para 6 alumnos de los 2810 existentes. También se usó el mismo criterio de filtrado para los alumnos. En el modelo 4.3 se utilizan estos datos.

Datos de clase

Los datos también se encuentran de forma longitudinal, donde en cada mes se encuentran la cantidad de estudiantes que pertenecen a la clase en conjunto con la suma, el promedio y el desvío de los indicadores de actividad. Existen alumnos de distintos grados y distintos contextos dentro de una clase, se utilizó el valor más frecuente dentro de la clase como representativo. Se descartaron las clases en la que ningún alumno hizo la prueba. Cabe aclarar que dentro de una clase no todos los alumnos realizan la prueba. Como variable objetivo se utiliza el promedio obtenido en la prueba por los estudiantes que la realizaron dentro de la clase, se tomó el criterio de que un estudiante se pondera en el promedio si estuvo en la clase en algún momento del año, a modo de ejemplo, si un estudiante estuvo en la clase A en el mes de julio y la prueba la da en diciembre estando en la clase B, el alumno aporta al promedio tanto de la clase A como en la clase B. Consultar el Apéndice A para ver la estructura de los conjuntos de datos.

2.3. Análisis descriptivo

Con el objetivo de mejorar la legibilidad y mantener coherencia con los conjuntos de datos presentados en la Sección 2.2, se realizará una distinción en el análisis descriptivo entre los alumnos y las clases. Es importante destacar que esta separación es por definición de naturaleza abstracta, y el análisis descriptivo de un conjunto proporciona información relevante sobre el otro.

2.3.1. Análisis descriptivo de los alumnos

La cantidad total de alumnos luego de aplicar los filtros antes mencionados es de 45,485, siendo 12,449 los que participaron en la prueba adaptativa de inglés. De este grupo, únicamente se cuenta con registros de las actividades asignadas y realizadas para 8,663 alumnos.

La Figura 2 ilustra la variación del puntaje final y la cantidad de actividades asignadas en función del grado. Se observa que a los alumnos de cuarto grado se les asignan menos actividades en comparación con los de quinto y sexto grado. Además, se aprecia que a medida que aumenta el grado, comienzan a surgir valores más altos en la prueba final.

Hay una disparidad considerable en los atributos de los alumnos. El departamento con más alumnos dentro del conjunto final es Canelones con 1747 y el menor es Flores con 80 alumnos, Canelones núclea aproximadamente 20 % del total de alumnos. Debido a la naturaleza propia de la variable, los alumnos de zonas rurales son considerablemente menos que los de entornos urbanos (8372 urbanos y 291 rurales). La cantidad de alumnos por contextos no es balanceada, siendo el quintil 5 el que tiene más alumnos (2570) y el quintil 1 el que tiene menos (1098). Las actividades

realizadas tampoco se hacen de forma equitativa entre contextos, siendo los alumnos de quintiles más altos los que hacen más actividades. El total de alumnos por grado está distribuido de forma pareja entre los 3 grados, incluso cuando se agrupa marginalmente por otra variable, la proporción sigue siendo aproximadamente equitativa. En el Apéndice B se deja una figura con los distintos gráficos de caja de puntos por distintas variables del alumno (Grado, Departamento, Zona, Contexto).

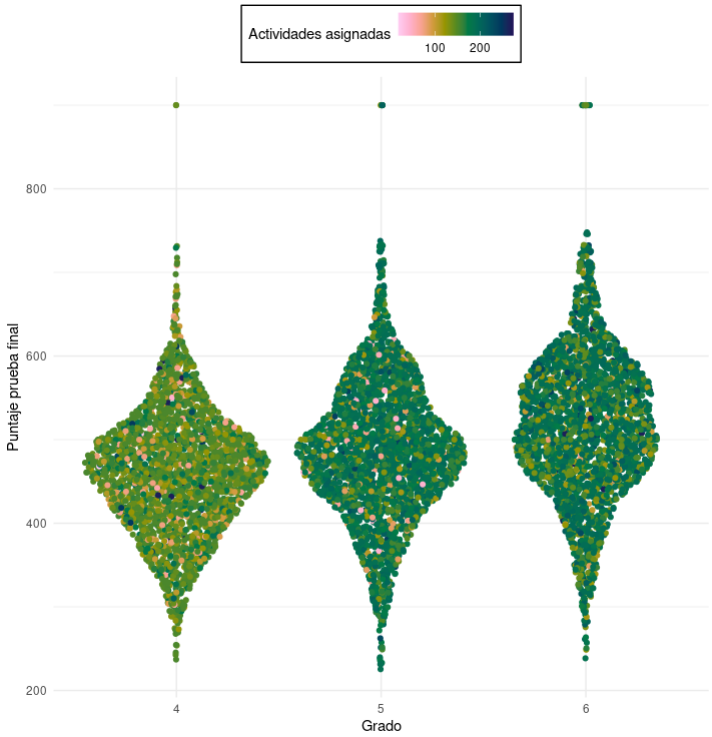


Figura 2: Gráfico sina con las actividades asignadas y puntaje final por grado del estudiante, donde el color representa la cantidad de actividades asignadas en el año.

2.3.2. Análisis descriptivo de las clases

Existen 852 clases en la que al menos un alumno hizo la prueba, realizó y se le asignaron actividades. Dentro de una clase no todos los alumnos realizan actividades,

esta proporción puede ser bastante baja (0 % y 25 %), aunque se debe tener en cuenta que es posible que varios alumnos hagan trabajo en conjunto con una única máquina, lo que disminuye artificialmente la proporción de alumnos con trabajo registrado.

Para estudiar el nivel de uso de las clases a lo largo del año lectivo se construye la variable *Nivel de uso* con cuatro niveles correspondientes a un uso Bajo de la plataforma (menos de 25 % de los estudiantes), uso Medio (entre 25 % y 50 %), uso Alto (50 % y 75 %) y Muy Alto (más de 75 %). Esta variable es calculada en cada mes, por lo que una misma clase puede comenzar con un uso Bajo y moverse a niveles mayores en el correr del año.

Figura 3 muestra barras apiladas al 100 % para visualizar la proporción de clases por categoría de nivel de uso en cada mes del año. Se observa que durante Marzo, aproximadamente la mitad de las clases muestra un nivel de uso Bajo y luego los meses siguientes esta cantidad disminuye bastante, indicando que el uso de LB es dilatado respecto del inicio del año escolar. Un segundo aspecto a destacar es que la proporción de clases con Muy Alto uso es relativamente pequeña en todos los meses del año siendo Agosto el mes donde dicha proporción es mayor.

Una limitación del diagrama de barras mes a mes, es que esconde la evolución de cada clase individual ya que las proporciones de clases en cada categoría en cada mes se calcula de forma independiente del resto. La Figura 4 presenta un gráfico Sankey, el cual está diseñado para mejorar esta limitación, al menos parcialmente. El diagrama Sankey muestra la cantidad de clases en cada nivel de uso en 4 momentos del año (Marzo, Mayo, Agosto y Noviembre) y resalta el cambio de un nivel a otro de cada clase en el tiempo. Para mejorar su lectura, el diagrama se presenta en paneles correspondientes al valor de nivel de uso en el mes de Marzo. Se observa que las clases

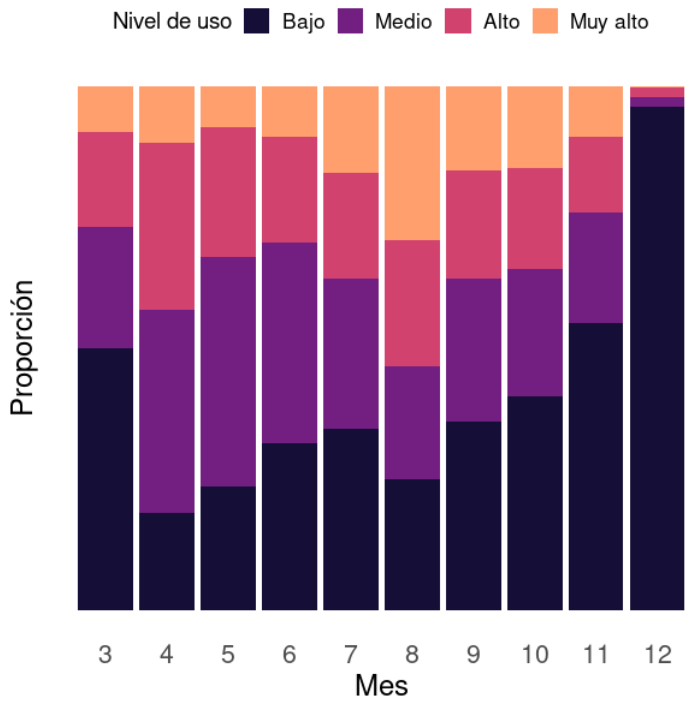


Figura 3: Gráfico de barras apiladas al 100 % del nivel de uso según mes. La altura de cada barra para cada mes, representa la proporción de clases en cada nivel de uso representado con color.

que muestran niveles Bajo o Medio de uso de la plataforma LB al inicio del año se mantienen en esos niveles, mientras que las clases que inician con niveles Muy Altos de uso les cuesta mantenerse en ese nivel de uso.

En relación a los resultados de la prueba adaptativa de inglés, es interesante explorar la homogeneidad de resultados en cada clase en relación a las 852 clases observadas. Con este fin, se calcula el promedio y desvío de los resultados de la prueba para cada clase por separado, esto es, se obtienen 852 medias y desvíos. En la Figura 5 se muestra un histograma de los promedios de resultados por clase y un histograma de los desvíos por clase. En ambos casos se agrega una línea vertical con la media o el desvío general con todos los estudiantes observados. Con respecto al histograma de medias, se observa que las medias por clase muestran una distribución aproximadamente simétrica entorno a la media general (488 puntos), también se observa variabilidad en el nivel promedio de cada clase. Con respecto a los desvíos, se observa que las clases muestran distinto grado de homogeneidad en los resultados de la prueba, con algunas clases de bajo desvío estándar y otras con mucha dispersión en los resultados de la prueba. Cabe aclarar que la gran cantidad de clases con desvío 0 se debe a que en estas clases solo un alumno realizó la prueba dentro de la clase. Los resúmenes de la Figura 5 sugieren que tener efectos individuales por clase puede ser de interés para la etapa de modelado.

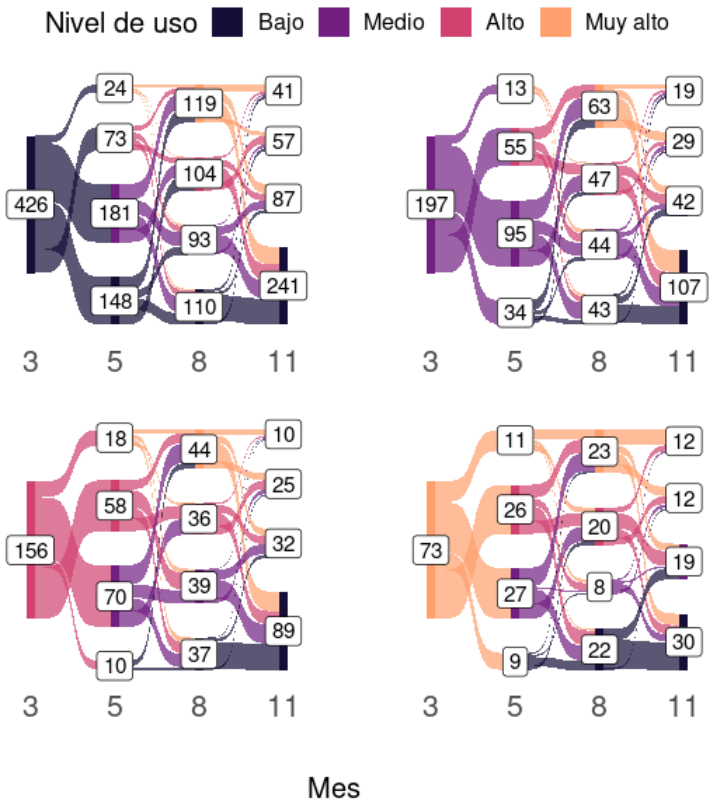
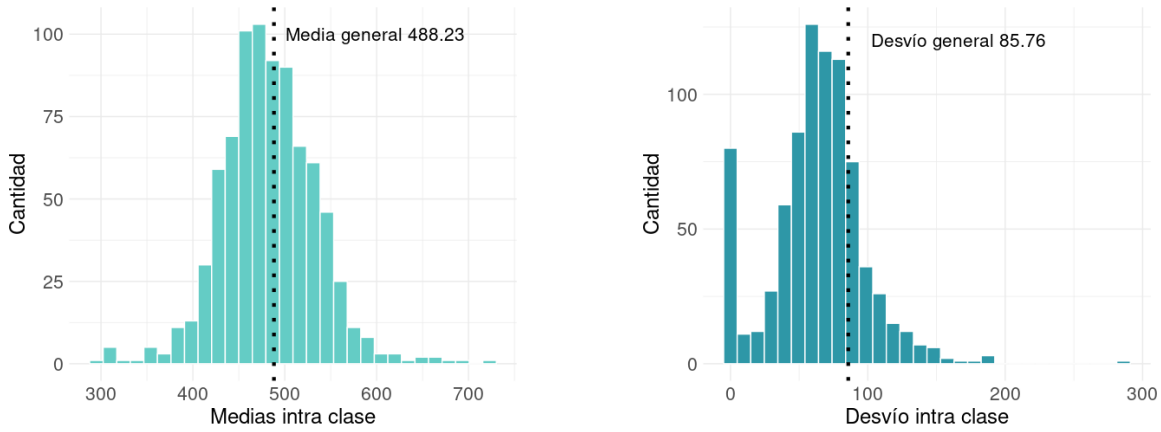


Figura 4: Diagrama Sankey que ilustra la evolución de la cantidad de clases en cada nivel de uso para los meses de Marzo, Mayo, Agosto y Noviembre. En cada panel se separan según el valor de nivel de uso en el mes de Marzo. El ancho de la barra en cada mes representa la frecuencia de clases en ese nivel.



(a) (b)

Figura 5: Histograma con la media (a) y el desvío estándar (b) de los resultados dentro de la clase.

3. Métodos: Árboles Bayesianos

En esta sección se comentarán los métodos estadísticos de aprendizaje supervisado, centrándonos en el modelo de Bayesian Additive Regression Trees (BART) (Chipman et al., 2010). Dado que BART es un método de aprendizaje bayesiano, se proporcionará una breve introducción a los conceptos clave de la inferencia bayesiana. Asimismo, dado que los datos exhiben una estructura de agrupación, se revisarán los modelos mixtos. Además, se describirá el flujo de trabajo esperado en un proyecto de ciencia de datos, junto con su implementación en `tidymodels`.

Un problema fundamental del aprendizaje supervisado es hacer inferencia acerca de una función desconocida f que dada una entrada $x = (x_1, \ldots, x_p)$ predice una respuesta y donde

$$y = f(x) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

En el presente trabajo, la variable de respuesta y representa una medida de desempeño educativo y las variables explicativas x contienen variables de uso de LB, en conjunto con otras variables de contexto.

Existen diversas técnicas disponibles para estimar f. Durante la investigación se utilizaron distintos métodos con el objetivo de comparar y seleccionar el mejor para responder las preguntas de interés de forma más precisa. Los métodos utilizados fueron Random Forest (Breiman, 2001), Support Vector Regression(Cortes & Vapnik, 1995), XGBoost (T. Chen & Guestrin, 2016) y el ya mencionado BART. Se omitió la descripción de los primeros tres métodos debido a que son ampliamente utilizados. En la Subsección 3.3 se detalla el cuarto método.

3.1. Inferencia Bayesiana

La inferencia estadística se ocupa de elaborar métodos para estimar características generales de un fenómeno de interés a partir de una cantidad finita de datos observados. Los denominados *métodos Bayesianos* son procedimientos estadísticos que utilicen la probabilidad como medida de incertidumbre y la regla de Bayes para actualizarla (Hoff, 2009).

Un modelo Bayesiano paramétrico tiene dos componentes básicos. En primer lugar, un *modelo de probabilidad para los datos* $\mathbb{P}(y|\theta)$, que representa la incertidumbre sobre las cantidades observables, y, condicional al valor de cantidades **no** observables o parámetros θ. En segundo lugar, un *modelo de probabilidad para los parámetros*, o previa $\mathbb{P}(\theta)$ que representa la incertidumbre sobre los parámetros del modelo para los datos antes de haberlos observado. Este segundo componente es fundamental en el proceso de inferencia bayesiana.

A partir de estos dos componentes y utilizando la regla de Bayes (debido a esto es que se le llama inferencia bayesiana), se puede obtener la distribución de los parámetros luego de haber observado los datos y. Esta distribución recibe el nombre de **posterior** y queda definida por

$$\mathbb{P}(\theta|y) = \frac{\mathbb{P}(y|\theta)\mathbb{P}(\theta)}{\mathbb{P}(y)} \overset{1}{=} \frac{\mathbb{P}(y|\theta)\mathbb{P}(\theta)}{\int \mathbb{P}(y|\theta)\mathbb{P}(\theta)d\theta}$$

El objetivo principal de los métodos Bayesianos es hallar la distribución posterior, $\mathbb{P}(\theta|y)$. Dicha distribución constituye la base de todas las afirmaciones inferenciales sobre las cantidades no observables θ ya sean, estimaciones puntuales, por intervalos, etc.

Al aplicar inferencia bayesiana en problemas estadísticos podemos incorporar información previa que conozcamos de la realidad, ya sea por conocimiento en el área específica o por experimentación previa. Esto nos permite ajustar el modelo en función del grado de certeza que tengamos en esa información. Además, gracias a la distribución posterior podemos obtener estimaciones puntuales de los parámetros y sus respectivos intervalos de probabilidad.

Debido a la complejidad que puede presentar obtener la distribución posterior es que se utilizan métodos de simulación Montecarlo basados en cadenas de Markov (MCMC) (Andrieu et al., 2003). Los MCMC son métodos de simulación que nos permiten obtener muestras de una distribución de probabilidad donde conocemos una función que es proporcional a esta, para el caso de la distribución posterior tenemos que $\mathbb{P}(\theta|y) \propto \mathbb{P}(y|\theta)\mathbb{P}(\theta)$, siendo $\mathbb{P}(\theta|y)$ la distribución de la que queremos obtener muestras y $\mathbb{P}(y|\theta)\mathbb{P}(\theta)$ una función conocida. Dentro de los métodos existentes, el

[1] Teorema de la probabilidad total

algoritmo Metropolis-Hasting (Hastings, 1970) y el muestreador de Gibbs (Geman & Geman, 1984) tienen una alta importancia, siendo este último utilizado para obtener muestras de la posterior en BART.

3.2. Modelos mixtos

Los modelos mixtos son modelos que utilizan efectos fijos y efectos aleatorios, se entiende como efecto fijo a los términos que no varían. Los efectos aleatorios son útiles cuando se trabaja con datos que muestran cierta estructura de agrupación, como datos longitudinales o en los que existen factores a los que pertenecen los individuos. Su utilidad radica en que nos permiten contemplar la variabilidad de pertenecer a un grupo o a otro cuantificando el efecto de cada grupo en la variable respuesta. A modo de ejemplo de datos agrupados, en esta investigación contamos con la agrupación propia de la realidad en la que los alumnos pertenecen a clases y a centros educativos. Debido a esto el interés de los modelos mixtos radica en la naturaleza del problema de estudio.

Dentro de las categorías de modelos mixtos existentes, están los modelos con intercepto aleatorio, que son un tipo de modelo mixto en el que al modelo solo se le suma un efecto aleatorio relacionado con el factor al que pertenece el individuo y que no interactúa con las otras variables.

Para aclarar estos conceptos se detallará la forma más básica de un modelo lineal mixto, que es el modelo lineal mixto con un intercepto aleatorio.

$$y_{ik} = x_i\beta + \alpha_k + \epsilon_i, \qquad \alpha_k \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

Donde x_i es un vector con las observaciones de los predictores del individuo i sin la constante y β es un vector con los coeficientes, pero sin el coeficiente asociado al intercepto.

Se puede demostrar que la correlación entre dos individuos que pertenecen al mismo grupo es $\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2}$. Por otro lado, esta correlación es nula cuando los individuos pertenecen a grupos distintos, lo que indica que este modelo considera la dependencia entre observaciones del mismo grupo.

En la regresión lineal clásica tenemos dos enfoques para trabajar con datos estructurados en grupos. Por un lado se puede tener parámetros para cada grupo por separado y que su estimación sólo depende de los datos del grupo correspondiente. (no pooling). Esta opción puede verse como el tener un modelo para cada grupo, es decir, el conjunto de parámetros es único para cada grupo. Las observaciones dentro de cada grupo solo sirven para estimar su propio grupo. Cuenta con alta varianza, ya que puede suceder que existan grupos con pocas observaciones. Por otro lado, es posible no incluir la información de pertenencia a cada grupo (complete pooling). En este caso, existe un único modelo, es decir, todas las observaciones comparten los mismos parámetros, por lo que todas van a influir en la estimación de estos. La desventaja es que presenta alto sesgo. Los modelos mixtos nos permiten introducir una tercera variante que es el partial pooling. Es un punto intermedio entre el complete pooling y el no pooling. Hay parámetros que se comparten entre los grupos y otros que son

propios del grupo.

Tenemos que la estimación del efecto aleatorio queda determinada por

$$\hat{\alpha}_j \approx \frac{\frac{n_j}{\sigma_\epsilon^2}}{\frac{n_j}{\sigma_\epsilon^2} + \frac{1}{\sigma_\alpha^2}}(\bar{y}_j - \beta\bar{x}_j) + \frac{\frac{1}{\sigma_\alpha^2}}{\frac{n_j}{\sigma_\epsilon^2} + \frac{1}{\sigma_\alpha^2}}\mu_\alpha$$

Se observa como la estimación es una forma de partial pooling, ya que pondera el complete pooling con μ_α y el no pooling con $(\bar{y}_j - \beta\bar{x}_j)$. Notar que en el límite cuando $\sigma_\alpha^2 \to 0$ estamos ante complete pooling y cuando $\sigma_\alpha^2 \to \infty$ estamos ante no pooling, lo cual es acorde a lo esperado, ya que cuanta más información tengamos del grupo (es decir, más varianza dentro del grupo) es deseable tener menos pooling y viceversa. Omitiremos la demostración de cómo fue derivada esta estimación.

Hemos repasado la forma más básica de modelo lineal mixto que es el modelo con pendiente aleatoria, pero dependiendo de la realidad del problema, se le pueden ir aplicando sucesivas complejizaciones. Un modelo mixto más complejo puede ser agregar predictores al nivel de grupo, es decir $\alpha_k \sim \mathcal{N}(\mu_\alpha + \gamma U_k, \sigma_\alpha^2)$. Para ilustrar este modelo dentro del contexto de este estudio, podemos considerar un modelo donde el efecto aleatorio presentado por el centro educativo tenga como predictor al departamento al que pertenece este. Otro tipo de modelo mixto es el modelo con pendientes aleatorias, que permite que las pendientes varíen según el grupo. El modelo de pendientes aleatorias permite que exista cierta correlación entre los efectos. Así se especifica un modelo mixto con dos pendientes aleatorias.

$$y_{ik} = x_i\beta + w_i(\gamma_1 + \mu_{1k}) + z_i(\gamma_2 + \mu_{2k}) + \epsilon_i$$

$$\text{Con} \quad \begin{pmatrix} \mu_{1k} \\ \mu_{2k} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2_{\mu_1} & \rho\sigma_{\mu_1}\sigma_{\mu_2} \\ \rho\sigma_{\mu_1}\sigma_{\mu_2} & \sigma^2_{\mu_2} \end{bmatrix} \right), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Donde γ_1 y γ_2 son dos efectos fijos y μ_1, μ_2 los efectos aleatorios. $\sigma^2_{\mu_1}$ y $\sigma^2_{\mu_2}$ son la varianza del efecto aleatorio μ_1, μ_2 respectivamente y ρ la correlación entre estos.

Cabe aclarar que estos modelos no son excluyentes y podemos, por ejemplo, trabajar con un modelo mixto con pendiente e intercepto aleatorio.

Para una lectura más completa sobre modelos lineales mixtos se recomienda (Gelman & Hill, 2006).

3.3. Bayesian Additive Regression Trees

A continuación se comentará el modelo Bayesian Additive Regression Trees (BART). Para una lectura más completa se recomienda (Y. V. Tan & Roy, 2019).

3.3.1. BART para regresión

Como se mencionó al inicio de la sección, el objetivo es estimar una función desconocida f 1. Para resolver este problema, BART aproxima a $f(x) = \mathbb{E}(y|x)$ usando

$$f(x) \approx h(x) \equiv \sum_{j=1}^{m} g_j(x) \tag{2}$$

Donde cada g_j es un árbol de regresión y m la cantidad de árboles.

Sea T un árbol completamente binario (i.e. cada padre tiene dos hijos) y $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$ un conjunto de parámetros donde cada μ_k está asociado al conjunto de hojas del árbol T. En general se asume que los μ_k son independientes entre sí. Observar que cada individuo x_i está asociado con un $\mu_k \in M$ por lo que dado T y M podemos definir a $g(x, T, M)$ como la función que asocia a x con μ_k. Por lo que podemos definir a la Ecuación 2 como

$$h(x) = \sum_{j=1}^{m} g(x, T_j, M_j) \Rightarrow y \simeq \sum_{j=1}^{m} g(x, T_j, M_j) \tag{3}$$

Observar que cuando fijamos la cantidad de árboles, el modelo de suma de árboles queda determinado por $(T_1, M_1), \ldots, (T_m, M_m)$ y σ. Estos parámetros son desconocidos y se usa inferencia bayesiana para estimarlos.

La Figura 6 exhibe dos árboles completamente binarios pertenecientes a un modelo de suma de árboles. Se presentan los elementos fundamentales de un árbol: los nodos internos con sus reglas de decisión, que incluyen una variable y un punto de corte, y las hojas de los árboles etiquetadas con μ_{jk}. Aunque no se ilustra de manera explícita, otro aspecto relevante es la profundidad del árbol, que representa la cantidad máxima de niveles que contiene. Ambos árboles de la figura presentan la misma profundidad (2).

Distribuciones previas

La distribución previa queda determinada por $\mathbb{P}\big[(T_1, M_1), \ldots, (T_m, M_m), \sigma\big]$. Especificando que $[(T_1, M_1), \ldots, (T_m, M_m)]$ y σ son independientes y que cada $(T_1, M_1), \ldots, (T_m, M_m)$ es independiente el uno del otro, es decir, (T_1, M_1) es independiente de (T_2, M_2), de (T_3, M_3),..., así hasta m. Con estas especificaciones podemos escribir la

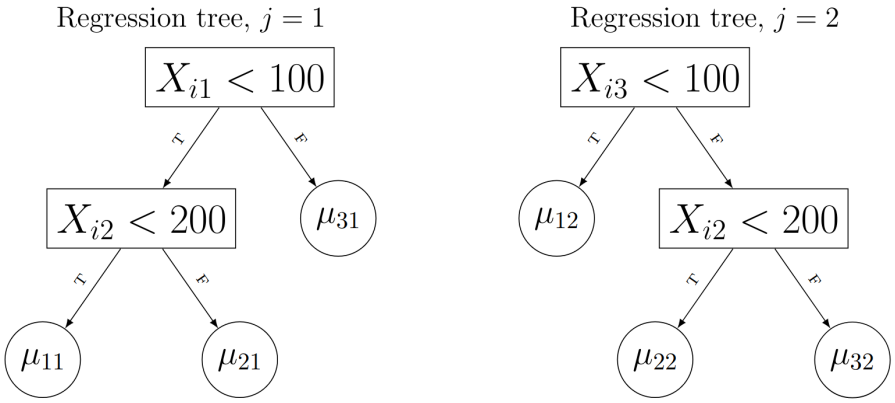


Figura 6: Figura obtenida de *Bayesian additive regression trees and the General BART model* (Y. V. Tan & Roy, 2019), Pág. 9. Representación de árboles binarios en el modelo BART. Los nodos internos están etiquetados con las reglas de partición, mientras que las hojas están etiquetadas con sus respectivos valores μ_{jk}.

previa como

$$\begin{aligned}
\mathbb{P}\big[(T_1, M_1), \ldots, (T_m, M_m), \sigma\big] &= \Big[\prod_{j=1}^{m} \mathbb{P}(T_j, M_j)\Big]\mathbb{P}(\sigma) \\
&= \Big[\prod_{j=1}^{m} \mathbb{P}(M_j|T_j)\mathbb{P}(T_j)\Big]\mathbb{P}(\sigma) \\
&= \Big[\prod_{j=1}^{m} \{\prod_{k=1}^{b_j} \mathbb{P}(\mu_{jk}|T_j)\}\mathbb{P}(T_j)\Big]\mathbb{P}(\sigma)
\end{aligned} \tag{4}$$

De 4 se desprende que hay que definir las distribuciones previas de σ, T_j y $\mu_{jk}|T_j$.

Previa de T_j: Para definir la previa de T_j hay que definir la distribución de los tres componentes que la determinan: La probabilidad de que un nodo sea terminal, qué variable se utiliza para dividir los datos en cada nodo interno y cuál es el punto de corte para la variable seleccionada en cada nodo. Una previa muy utilizada es:

- $\mathbb{P}(\textit{nodo sea terminal}) \propto \frac{\alpha}{(1+d)^{\beta}}$, donde d representa la profundidad del árbol. α y β se asumen conocidos.
- Selección uniforme de la variable para dividir entre todos los predictores.
- Selección uniforme del punto de corte entre los valores observados de la variable seleccionada

Previa de $\mu_{jk}|T_j$: $\mu_{jk}|T_j \sim \mathcal{N}(0, \sigma_\mu^2)$. En general se reescala a la variable a predecir y para que $y_{\text{mín}}$ = -0.5 y $y_{\text{máx}} = 0.5$. Con este reescalamiento definimos a $\sigma_\mu^2 = \frac{0.5}{k\sqrt{m}}$. k es un hiperparámetro que se fija previamente. Se recomienda usar $k = 2$ para que la probabilidad de que $\mathbb{E}(y|x)$ pertenezca al intervalo $(y_{\text{mín}}, y_{\text{máx}})$ sea de 0.95.

Previa de σ: Se utiliza la distribución $IG(\frac{\nu}{2}, \frac{\nu\lambda}{2})$ como previa de σ^2. Es necesario fijar ν y λ, por defecto se usa $\nu = 3$ y λ es el valor tal que $\mathbb{P}(\sigma^2 < s^2; \nu; \lambda) = 0.9$ donde s^2 es la varianza estimada de los residuos de una regresión lineal múltiple en la que y es la variable de respuesta y x las variables explicativas.

3.3.2. BART con respuesta binaria

En el artículo original (Chipman et al., 2010) ya se plantea la extensión para trabajar con respuesta binaria, utilizando el modelo probit

$$\mathbb{P}\big[y_i = 1|x_i, (T_1, M_1), \ldots, (T_m, M_m)\big] = \Phi[\sum_{j=1}^{m} g(x_i, T_j, M_j)] \tag{5}$$

Donde $\Phi[.]$ es la distribución acumulada de la normal estándar. A diferencia de otros métodos de agregación, BART no usa un sistema de votación para la clasificación.

Las distribuciones previas difieren del modelo BART original. La diferencia más notable es que no es necesario una previa para σ. Tomando las mismas suposiciones que para el modelo original, la distribución previa queda determinada por

$$\mathbb{P}((T_1, M_1), \ldots, (T_m, M_m)) = \prod_{j}^{m} [\prod_{k=1}^{b_j} \mathbb{P}(\mu_{jk}|T_j)] \mathbb{P}(T_j) \tag{6}$$

Para $\mathbb{P}(T_j)$ se utiliza exactamente la misma previa que para el modelo original. La previa $(\mu_{jk}|T_j)$ también sigue una distribución normal, pero tomamos al desvío estándar de la distribución cómo $\sigma_\mu = \frac{3}{k\sqrt{m}}$. Esto es debido a que se considera que $(-\Phi[3.0], \Phi[3.0])$ contiene a la mayoría de los valores de $p(x)$.

3.3.3. BART con intercepto aleatorio

Como se evidenció en la Subsección 3.2, es crucial la capacidad de incorporar efectos aleatorios. A pesar de la gran flexibilidad que muchas técnicas de aprendizaje automático poseen en la estimación de la función f, es común que asuman la independencia de los datos. Sin embargo, como ya dijimos anteriormente en los casos de estructuras con grupos, esta suposición no se cumple.

La extensión de BART para trabajar con un intercepto aleatorio fue presentada en (Y. Tan et al., 2016). Se extiende la Ecuación 1 para agregar el intercepto aleatorio α_k, el subíndice k se usa para indicar el factor y el subíndice i para indicar el individuo dentro del factor.

$$y_{ik} = \sum_{j=1}^{m} g(x_{ik}, T_j, M_j) + \alpha_k + \epsilon_{ik}, \qquad \alpha_k \sim \mathcal{N}(0, \tau^2), \quad \epsilon_{ik} \sim \mathcal{N}(0, \sigma^2) \tag{7}$$

Se específica que α_k y ϵ_{ik} sean independientes. Tomando los mismos supuestos respecto a la previa del modelo BART original y asumiendo que σ^2 y τ^2 son independientes, podemos descomponer la previa como:

$$\mathbb{P}[(T_1, M_1), \ldots, (T_m, M_m), \sigma, \tau] = \big[\prod_{j=1}^{m} \big\{ \prod_{k=1}^{b_j} \mathbb{P}(\mu_{jk}|T_j) \big\} \mathbb{P}(T_j) \big] \mathbb{P}(\sigma) \mathbb{P}(\tau) \tag{8}$$

De la Ecuación 8 se desprende que hay que definir las previas para T_j, $\mu_{jk}|T_j$, σ y τ. La distribución previa por defecto para τ^2 es $IG(1, 1)$. Para T_j se utiliza exactamente la misma previa que para el modelo BART original, para $\mu_{jk}|T_j$ y σ se conserva la familia de distribución para la previa, pero varían los hiperparámetros. Respecto a la previa de σ se conserva el hiperparámetro ν, pero cambia la definición del hiperparámetro λ. Para λ primero se estima un modelo MARS (Multivariate Adaptive Regression Splines) (Friedman, 1991) donde se utilizan los mismos predictores y la misma respuesta que para el modelo BART. Luego de estimado el modelo se utiliza como un estimador del intercepto aleatorio α_k a la media de los residuos para cada factor k, obtenemos una estimación de la varianza con $s^2 = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_{ik} - \hat{y}_{ik} - \hat{\alpha}_k)^2}{N - \mathcal{N}(1 - \sqrt{(\frac{RSS}{GCV \times N})})}$. N es la cantidad de observaciones, RSS es la suma de residuos al cuadrado y GCV es el valor de la validación cruzada generalizada. Por lo tanto, definimos a λ como el valor que hace que se cumpla $\mathbb{P}(\sigma^2 < s^2; \nu, \lambda) = 0.9$. Se utiliza el hiperparámetro $\sigma_\mu = \frac{1.96}{k\sqrt{m}}$.

Obtener predicciones para factores no observados

Un problema que se desprende naturalmente es el de obtener predicciones de factores que no hayan sido previamente observados, debido a que vamos a desconocer los efectos aleatorios de estos. A diferencia de métodos más tradicionales en los que a veces se requiere de técnicas particulares para realizar predicciones con factores no observados, para el caso de los modelos mixtos existe una forma inherente al modelo para resolver esta problemática.

En el marco de los modelos lineales mixtos, más concretamente el modelo con intercepto aleatorio que fue con el que trabajamos, el procedimiento para obtener una predicción de una observación perteneciente a un factor que no haya sido visto previamente [1], consta de primero simular $\hat{\alpha}_{j+1} \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$. Con un $\hat{\alpha}_{j+1}$ simulado se obtiene la predicción puntual de $y_{i+1,j+1}$ con $\hat{y}_{i+1,j+1} = x_{i+1}\beta + \hat{\alpha}_{j+1} + \epsilon_{i+1}$. Podemos realizar varias simulaciones de $y_{i+1,j+1} \sim \mathcal{N}(x_{i+1}\beta + \hat{\alpha}_{j+1} + \epsilon_{i+1}, \sigma_\epsilon^2)$ con el objetivo de medir la incertidumbre que tenemos respecto a esta predicción, para cada una de estas observaciones tenemos que previamente realizar una nueva simulación de $\hat{\alpha}_{j+1}$.

Para el caso de BART, la librería `dbarts` utiliza como efecto aleatorio una muestra obtenida de la distribución $\mathcal{N}(0, \tau^2)$, donde τ^2 se obtiene de la muestra posterior $\mathbb{P}(\tau|y)$.

3.3.4. Posterior de BART

Para realizar predicciones en BART, se emplean simulaciones de la distribución posterior para obtener muestras de las componentes de la suma de árboles. Estas muestras se utilizan para generar varias predicciones. En general, se obtienen múltiples mues-

[1] Teniendo a $\mu_\alpha, \sigma_\alpha, \beta, \sigma_\epsilon^2$ previamente estimados

tras, lo que resulta en diversas predicciones, siendo la predicción puntual el promedio de estas. La disponibilidad de varias muestras facilita la derivación de conclusiones sobre la incertidumbre asociada a la predicción.

Las distribuciones previas presentadas en la Subsección 3.3.1 inducen a una distribución posterior

$$\begin{aligned}\mathbb{P}[(T_1, M_1), \ldots, (T_m, M_m), \sigma | y] \propto \mathbb{P}[y | (T_1, M_1), \ldots, (T_m, M_m), \sigma] \\ \times \mathbb{P}[(T_1, M_1), \ldots, (T_m, M_m), \sigma]\end{aligned} \tag{9}$$

En este trabajo no se detallará como obtener muestras de la posterior 9, se recomienda referirse a los artículos recomendados.

Se requieren pequeñas modificaciones para obtener muestras de la posterior en BART con respuesta binaria y BART con intercepto aleatorio. Cuando se trata de BART con respuesta binaria es necesario definir

$$\begin{aligned} z_i &\sim \mathcal{N}_{(-\infty,0)}\Big[\sum_{j=1}^{m} g(x_i, T_j, M_j), 1\Big], \qquad \text{cuando } y_i = 0 \\ z_i &\sim \mathcal{N}_{(0,+\infty)}\Big[\sum_{j=1}^{m} g(x_i, T_j, M_j), 1\Big], \qquad \text{cuando } y_i = 1 \end{aligned} \tag{10}$$

Podemos tratar a z como la respuesta continua de un modelo BART

$$z = \sum_{j=1}^{m} g(x, T_j, M_j) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 1) \tag{11}$$

Para este z podemos usar la misma estimación de la posterior que se utiliza para el

modelo BART original, con σ fijado en 1.

En el caso de BART con intercepto aleatorio primero es necesario obtener muestras de la posterior de σ, τ y α_k. Luego con el α_k actualizado se obtiene $\tilde{y}_{ik} = y_{ik} - \alpha_k$. Con esta transformación podemos obtener muestras como en el modelo original.

3.4. Flujo de trabajo

Cuando trabajamos en proyectos de estadística aplicados a datos reales o simulados es importante y deseable definir flujos de trabajo que faciliten la reproducibilidad de los resultados de dicha investigación. En términos generales podemos decir que una investigación es reproducible si somos capaces de obtener los mismos resultados de una investigación ya realizada. Para ello en general es necesario no solo contar con los datos sino con una descripción detallada de los pasos que se siguieron en dicha investigación, tanto en la manipulación y transformación de datos, definición de los métodos utilizados, heramientas computacionales y versiones de las mismas entre otros (Gentleman & Temple Lang, 2007). En este contexto el flujo de trabajo es el proceso sistemático e iterativo que va desde la generación, transformación y limpieza de datos, exploración de los mismos, modelado y posterior comunicación de resultados. En esta sección nos concentraremos en mencionar el flujo de trabajo enfocado a la etapa de modelado sin embargo en la Sección 2 detallamos parte de la transformacion y exploración de datos que son relevantes y forman parte del flujo de trabajo completo. En lo que sigue se enumerarán los pasos que componen un flujo de trabajo normal con foco en la etapa de modelado, aunque existan una cantidad diversa de variaciones de este flujo. El número de los pasos indica el orden en el que estos son ejecutados. Se presupone que el lector posee conocimientos básicos en

análisis de datos, incluyendo conceptos como validación cruzada para la selección de hiperparámetros y la elección de modelos, por lo que se omitirá la descripción detallada de estos métodos.

1. **Preparación de los datos:** Consiste en transformar el conjunto de datos original a un conjunto de datos que sirva de entrada para el modelo. Entre las prácticas comunes está la estandarización, la codificación de variables categóricas, la imputación de datos faltantes, etc. Dependiendo de que tan preparados vengan los datos de origen y los datos que acepta el modelo, este paso puede llegar a ser extenso, además de que muchas decisiones quedan a criterio del investigador. Usualmente en este paso también se deciden los roles de las variables, es decir, cuál va a ser la variable respuesta y cuáles las variables predictoras.

2. **Exploracion de datos:** Consiste en el proceso iterativo donde nos hacemos preguntas y buscamos respuestas sobre nuestros datos con distintas herramientas estadísticas donde la visualizacion de datos juega un rol muy importante. Esta etapa nos permite encontrar inconsitencia en los datos, identificar posibles variables relevantes para nuestro modelo entre otros.

3. **Definición del modelo:** Se define el modelo a ser utilizado, además si existen distintas implementaciones es necesario determinar cuál de estas es la que se va a utilizar. También se elige qué hiperparámetros quedan determinados y para cuáles es necesario llevar a cabo una búsqueda.

4. **Definición de la grilla de hiperparámetros:** Se establece el espacio de búsqueda de los hiperparámetros. Además, es necesario definir si se va a utilizar una grilla regular (se exploran todas las combinaciones) o una grilla no regular

(de todas las combinaciones posibles se utilizan algunos puntos). Dentro de los tipos de grillas no regulares más conocidas están la grilla aleatoria y la grilla de hipercubos latinos.

5. **Creación de particiones para la validación cruzada:** Se determina que técnica de validación cruzada se va a utilizar. Usualmente, se utiliza validación cruzada con k-particiones. A partir del conjunto de datos de entrenamiento se realizan las particiones necesarias para realizar la validación cruzada, generalmente el número de particiones es 10.

6. **Ajuste del modelo con los distintos hiperparámetros:** Para cada punto de la grilla de hiperparámetros se ajustan y evalúan la cantidad de modelos determinados por la validación cruzada, en general se utilizan técnicas de computación paralela para acortar los tiempos.

7. **Selección del mejor modelo:** Utilizando una métrica de performance se escoge a la mejor combinación de hiperparámetros. La métrica a ser utilizada depende de si se trata de una regresión o una clasificación. Dentro de las métricas más comunes están RMSE y R^2 para la regresión, mientras que la precisión y AUC son las más frecuentes para la clasificación.

8. **Predicción:** Con el modelo final se realiza la predicción del conjunto de datos de test.

9. **Evaluación de los resultados:** A partir de las predicciones y los datos observados se calculan distintas métricas para evaluar la performance del modelo final. Muchas veces las métricas pueden estar acompañadas de herramientas

visuales para facilitar la comprensión de los resultados.

A lo largo del trabajo se utilizó la librería `tidymodels` para la calibración, selección, predicción y comparación de distintos modelos. Se decidió utilizar `tidymodels` debido a la facilidad que brinda para seguir el proceso de modelado de principio (preparación de los datos) hasta el final (evaluación de las predicciones). La facilidad radica en que permite un flujo de trabajo estandarizado y coherente para ajustar y comparar varios métodos de aprendizaje estadístico al mismo tiempo simplificando la reproducibilidad y coherencia en la implementación (Kuhn & Silge, 2022). A continuación se detallan los paquetes principales que conforman a `tidymodels`.

- `parsnip` (Kuhn & Vaughan, 2023): Este paquete proporciona una interfaz consistente para especificar modelos. Permite especificar un modelo mediante una fórmula y luego ajustar ese modelo a los datos.

- `dials` (Kuhn & Frick, 2023): Se utiliza para crear y gestionar parámetros de ajuste para modelos. Ayuda con la optimización de hiperparámetros, permitiendo buscar los mejores valores para el rendimiento del modelo.

- `recipes` (Kuhn, Wickham & Hvitfeldt, 2023): Proporciona una manera de preparar datos para el modelado. Permite especificar una serie de pasos de preprocesamiento de datos, facilitando la creación de flujos de trabajo reproducibles.

- `workflows` (Vaughan & Couch, 2023): Este paquete proporciona una manera de combinar una especificación de modelo, una receta de preparación de datos y un conjunto de parámetros de ajuste del modelo en un solo objeto. Esto facilita la gestión de todo el proceso de modelado.

- `tune` (Kuhn, 2023): Este paquete ayuda en el proceso de ajuste de hiperparámetros. Se utiliza en conjunto con los paquetes `dials` y `workflows` para buscar eficientemente los mejores hiperparámetros para un modelo dado.

- `yardstick` (Kuhn, Vaughan & Hvitfeldt, 2023): Facilita el cálculo de diversas métricas de rendimiento para evaluar qué tan bien se desempeña un modelo. Se puede trabajar con métricas de clasificación y de regresión.

En la Figura 7 se encuentra un diagrama de flujo que ilustra la relación del flujo de trabajo descrito con los paquetes de `tidymodels`.

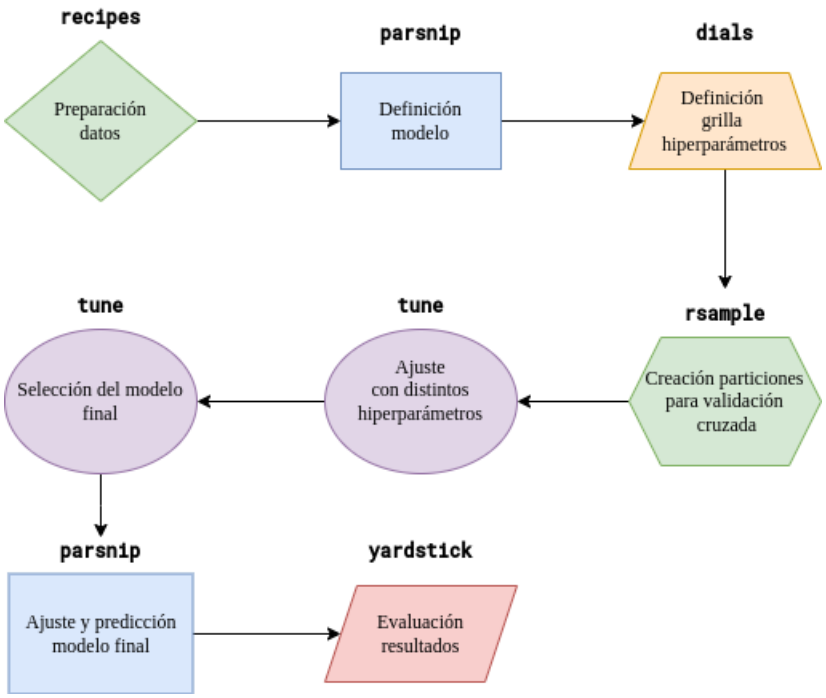


Figura 7: Flujo de trabajo en `tidymodels`, indicando en cada paso que librería se ha usado. La dirección de las flechas indica el progreso del flujo.

4. Resultados: Desempeño académico

En el transcurso del proyecto se realizaron tres modelos predictivos. El primero es un modelo predictivo para el total de los alumnos en el que se desea predecir el puntaje obtenido en la prueba final (4.1), el segundo es un modelo predictivo para clases, en el que el objetivo es predecir el puntaje promedio obtenido por los alumnos que dieron la prueba dentro de la clase (4.2) y el tercero es un modelo predictivo para alumnos de sexto grado en el que se quiere predecir si un alumno alcanzó el nivel mínimo (4.3). En esta sección se comentan los resultados obtenidos para cada uno de los modelos.

4.1. Modelo predictivo para alumnos

El objetivo de este modelo es predecir el resultado de la prueba adaptativa de inglés, en puntos, de los estudiantes con los datos disponibles hasta julio ya que es deseable poder obtener una alerta temprana. Como la variable a predecir es de naturaleza cuantitativa, se escogieron distintos modelos de regresión. Se utilizaron los modelos Random Forest(RF), Bayesian Additive Regression Trees (BART), XGBoost (XGB) y Support Vector Regression (SVR) con el kernel polinomial. Los `engine` utilizados para estos modelos fueron `ranger` (Wright & Ziegler, 2017), `dbarts` (Dorie, 2023), `xgboost` (T. Chen et al., 2023), `kernlab` (Karatzoglou et al., 2023) (Karatzoglou et al., 2004) respectivamente.

El conjunto de datos original de 8,663 estudiantes fue particionado en 6928 estudiantes ($\approx 80\,\%$ del total) para entrenamiento de los modelos y 1735 estudiantes ($\approx 20\,\%$) para un test del modelo final. Se usó validación cruzada con 10 particiones para encontrar los mejores hiperparámetros para cada modelo. En el Anexo C se en-

cuentran los hiperparámetros que se ajustaron por modelo. Se creó una cuadrícula de búsqueda utilizando el método de hipercubos latinos (McKay et al., 1979), con 20 puntos para cada modelo. La selección de la mejor combinación de hiperparámetros se basó en la Raíz del Error Cuadrático Medio (RMSE). Debido al alto coste computacional de encontrar la mejor combinación para cada modelo, fue necesario utilizar la plataforma Vertex AI de Google Cloud, ya que en total era necesario ajustar 1000 modelos (10 particiones por 20 combinaciones para cada uno de los 5 modelos). Se aprovechó de la librería `doParallel` (Daniel et al., 2022) para ejecutar el ajuste en paralelo. Ver Figura 8 para los resultados obtenidos. Debido a su mejor performance en el indicador elegido, se decidió utilizar BART con los hiperparámetros ajustados como el modelo final.

Se encontró que la configuración óptima de hiperparámetros para este modelo consiste en 482 árboles, con valores específicos para los parámetros de las previas: k = 0.795 que permite una alta variación de los nodos terminales, α = 0.702 y β = 2.23, favoreciendo así la elección de árboles de menor tamaño.

Se reportó un RMSE de 74.3 para el conjunto de test. En la Figura 9 se encuentra un gráfico de puntaje final predicho contra el puntaje final observado, en el que se observa que el modelo tiene un sesgo para predecir resultados relacionados con el nivel más bajo de inglés y más alto, tendiendo a predecir valores más cercanos al promedio. A partir del puntaje predicho se obtuvo el nivel utilizando los puntos de corte ya establecidos, se observa que el modelo no pudo predecir a ningún alumno como nivel Pre-A1. Además el modelo presenta problemas al predecir el nivel final ya que en la mayoría de los niveles definidos el rango entre sus puntos de corte es inferior al valor del RMSE.

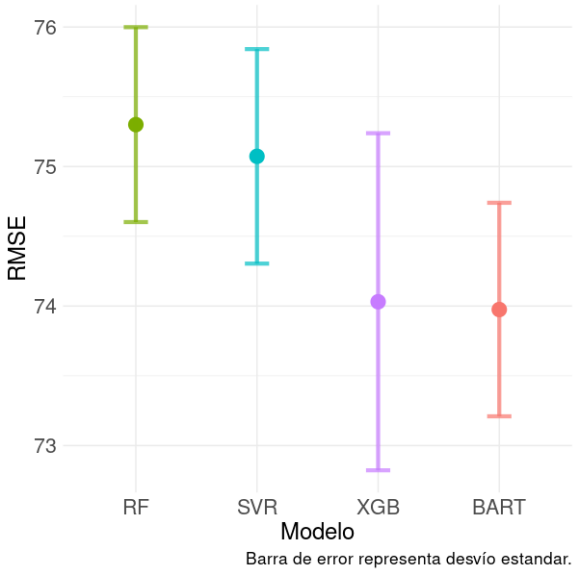


Figura 8: RMSE promedio obtenido en la mejor combinación de hiperparámetros para cada modelo, con su desvío estándar.

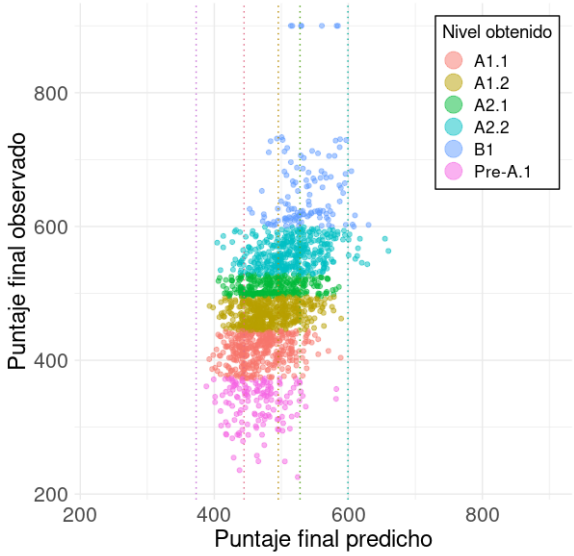


Figura 9: Gráfico de dispersión del puntaje final observado y puntaje final predicho donde el color representa el nivel obtenido en la prueba y las líneas punteadas los distintos puntos de corte.

Para calcular la importancia de las variables, se implementó un indicador que calcula la suma de veces que la variable aparece en los 482 árboles de cada una de las 1000 muestras posteriores, ponderando por la cantidad de observaciones que existen en el nodo de la variable y dividiendo por la cantidad de reglas de decisión que hay en cada muestra (i.e. la cantidad de nodos no terminales que hay en los 482 árboles). Este indicador es una variación del original planteado por (Chipman et al., 2010), la propuesta de esta variación figura en el mismo artículo. El problema del indicador original, ya advertido por los autores, es que variables irrelevantes pueden sumar importancia al aparecer en varios árboles en nodos más profundos. Ver Figura 10 para las 10 variables más importantes. Se destaca que dentro de las variables más importantes figuran variables de la clase (grado, departamento, contexto sociocultural). Las 3 variables más importantes son el puntaje mínimo por actividad promedio obtenido en el mes de julio, el puntaje mínimo por actividad promedio obtenido en el mes de mayo y si el alumno es del departamento de Lavalleja.

4.2. Modelo predictivo para clases

En la Subsección 2.3.2, se resaltó la observación de que el desvío intraclase era inferior al desvío general, indicando la posible presencia de información dentro de cada clase. Esto se complementa con los resultados obtenidos en la Subsección 4.1, donde se identificaron variables relacionadas con las clases como las más importantes. Con base en esta información, se tomó la decisión de desarrollar un modelo en el cual las unidades de análisis sean las clases y la variable a predecir sea el promedio de los puntajes finales de las mismas, utilizando datos recopilados hasta julio. Se utilizó únicamente un modelo Random Forest con 1000 árboles, en el que se ajustó el hiperparámetro

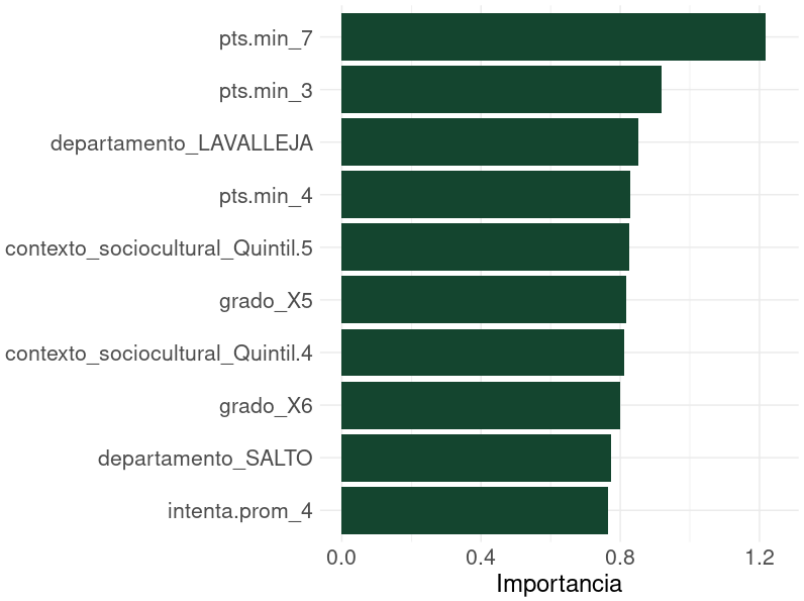


Figura 10: Importancia de las variables para BART 4.1, se seleccionan las primeras 10 más importantes.

utilizado para definir la cantidad de variables que se usan por árbol (`mtry`) y el hiperparámetro que controla la cantidad mínima de observaciones que tiene que tener un nodo para poder dividirse (`min_n`). Nuevamente se utilizo validación cruzada con 10 particiones para encontrar los mejores hiperparámetros. Para el ajuste se creó una grilla de 10x10 equiespaciada para abarcar todo el espacio de `mtry` y para `min_n` que cubra los valores entre 1 y 40, para así identificar una región donde podría encontrarse la combinación óptima (aquella que minimice el RMSE). Esta región óptima queda determinada por valores entre 20 y 30 para `min_n` y entre 30 y 50 para `mtry`. Luego, se llevó a cabo una búsqueda localizada mediante una grilla de 5x5, también equiespaciada, que cubriera específicamente esta región. Se seleccionó la mejor combinación de esta grilla para entrenar el modelo final. Se encontró que `mtry` $= 35$ y `min_n` $= 30$ es la mejor combinación.

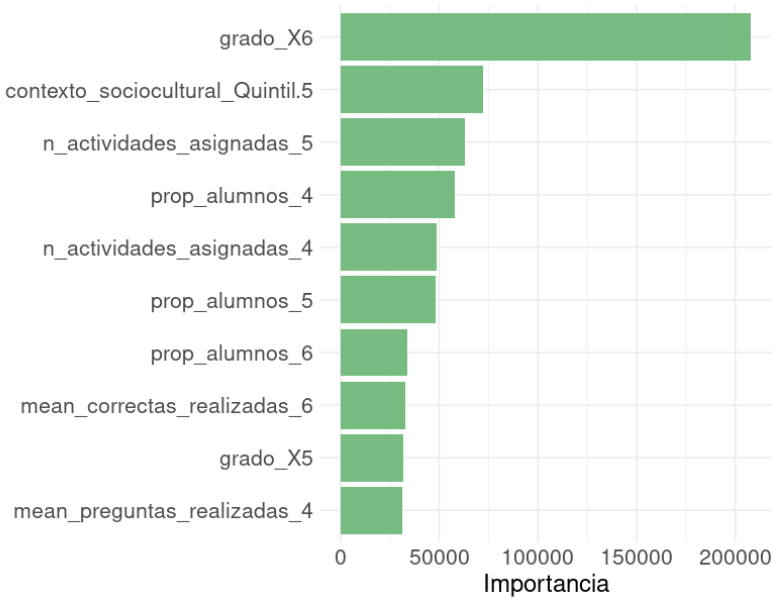


Figura 11: Importancia de las variables para Random forest 4.2, se seleccionan las primeras 10 más importantes.

Se reportó un RMSE de 46.7 para el conjunto de test. La importancia fue calculada con el paquete `vip` (Greenwell & Boehmke, 2020), utilizando el criterio del decrecimiento de la impureza. En la Figura 11 se visualizan las 10 variables más importantes, se destaca la importancia que tiene que la clase pertenezca a sexto grado, las otras dos variables más importantes son si la clase está relacionada con el quintil del contexto sociocultural más alto y la cantidad de actividades asignadas en el mes de mayo.

4.3. Modelo predictivo para alumnos de sexto grado

Interesa un modelo que pueda clasificar si un estudiante de sexto grado llego al nivel mínimo o no, nuevamente con los datos disponibles hasta julio. La meta de egreso establecida por Políticas Lingüisticas de ANEP para Educación Primaria es el nivel

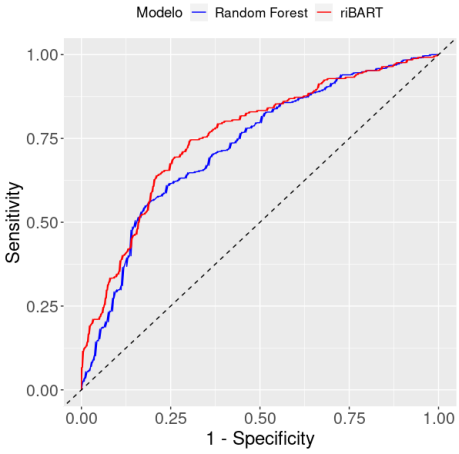
A2, por lo que se toma A2.1 como nivel mínimo. Se recodificó la variable de respuesta, en donde 1 es que el estudiante alcanzó el nivel y 0 es que no. Se utilizó un modelo BART con efecto aleatorio, el efecto aleatorio seleccionado fue el centro al que pertenecen los estudiantes, no se escogió a la clase como efecto aleatorio debido a que existen muchas clases y varias con pocos alumnos. Se tomó un modelo Random Forest como referencia. Debido a que el modelo BART con intercepto aleatorio no está implementado para `tidymodels` se usó directamente la librería `dbarts`. Para el modelo BART no se realizó ningún ajuste de los hiperparámetros y se utilizaron los valores por defecto: $\alpha = 0.95$, $\beta = 2$ y $k = 2$. Se fijó para que se usen 200 árboles, 4 cadenas, que se descarten las primeras 1000 iteraciones y quedándonos con 200 muestras posteriores por cadena. Para el modelo Random Forest se ajustaron los hiperparámetros `mtry` y `min_n`, utilizando validación cruzada con 10 particiones y una grilla de 20 puntos armada con el método de hipercubos latinos, la mejor combinación fue `mtry` $= 24$ y `min_n` $= 29$. En la Figura 12, se presentan los resultados de ambos modelos, calculadas a partir de la predicción en el conjunto de prueba. Se observa que el rendimiento del modelo BART es ligeramente superior.

Se estudió el efecto aleatorio de los distintos centros educativos, para el estudio se tomó el promedio de las muestras de la posterior y se le realizó el desvío, se usó como criterio que el efecto era considerable cuando incluyendo el desvío el efecto no incluía al 0. Ver Figura 13 para los distintos efectos aleatorios. Existen 9 centros de estudio con efecto negativo y 17 con efecto positivo.

Como método de análisis de resultado se crearon tres individuos ficticios relacionados con los deciles de los indicadores de uso, un individuo forma parte del primer decil, otro del quinto decil y el tercero forma parte del último decil. Se repitió el expe-

Indicador	Random Forest	riBART
Accuracy	0.686	0.704
Sensitivity	0.711	0.806
Specificity	0.655	0.575
AUC	0.727	0.759

(a)



(b)

Figura 12: Resultados de ambos modelos. (a) Métricas de la performance (b) Curva ROC

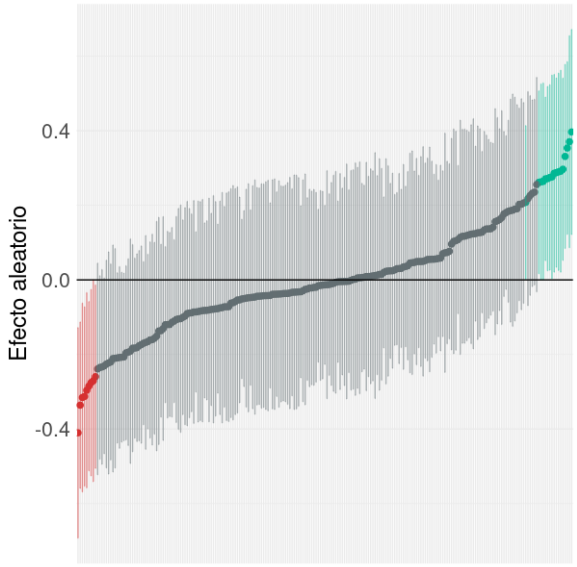
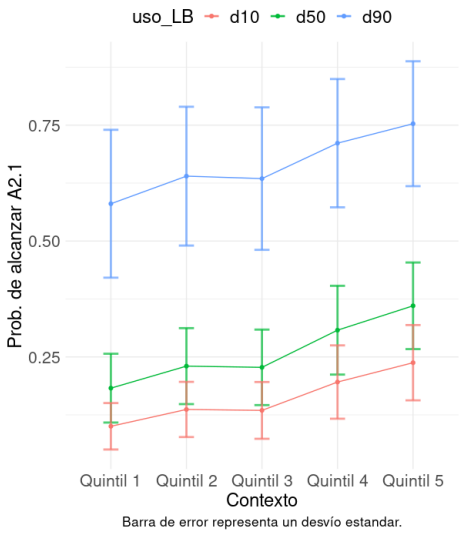


Figura 13: Efecto aleatorio de los distintos centros. Se remarcan los centros con efecto aleatorio negativo y positivo. La barra de error representa el desvío.

rimento variando el contexto sociocultural. Se dejan fijos los otros atributos que no se relacionan con la actividad, en concreto se les asignó que pertenezcan al departamento de Montevideo, que sean de zona urbana y formen parte de la clase 1016, los criterios de asignación de estos atributos no siguieron un criterio particular. En la Figura 14 se muestran los resultados obtenidos. Como es esperable, el modelo le asigna mayor probabilidad de llegar al puntaje mínimo a los de deciles más altos, confirmando la relación que hay entre el uso de la plataforma y el nivel obtenido. Se destaca como la probabilidad de llegar al nivel mínimo aumenta a medida que aumenta el contexto sociocultural, dejando lo demás constante.

Decil	Contexto	Prob	Std.
d90	Quintil 5	0.753	0.135
d90	Quintil 4	0.711	0.138
d90	Quintil 3	0.635	0.154
d90	Quintil 2	0.640	0.150
d90	Quintil 1	0.580	0.159
d50	Quintil 5	0.360	0.093
d50	Quintil 4	0.308	0.096
d50	Quintil 3	0.228	0.081
d50	Quintil 2	0.230	0.082
d50	Quintil 1	0.183	0.074
d10	Quintil 5	0.238	0.081
d10	Quintil 4	0.196	0.079
d10	Quintil 3	0.135	0.061
d10	Quintil 2	0.137	0.060
d10	Quintil 1	0.100	0.050

(a)



(b)

Figura 14: Predicción de los individuos ficticios. (a) Tabla con los resultados ordenados de forma descendente por decil y contexto. (b) Gráfico de los resultados con barra de error de un desvío estándar.

5. Discusión

En el transcurso del trabajo se han presentado distintos modelos para relacionar el uso de la plataforma educativa Little Bridge, hasta el mes de julio, con el rendimiento en las pruebas adaptativas de inglés.

Se ha abordado en detalle la preparación de los datos para poder utilizarlos como fuente de entrada de los distintos modelos, fue presentado un modelo de entidad-relación, con el objetivo de representar los datos de manera comprensible, facilitando así su manipulación. Del análisis descriptivo surgió que la cantidad de alumnos ni por departamento ni por contexto sociocultural es balanceada, siendo Canelones el departamento con más alumnos y el quintil 5 el quintil con más alumnos, destacando que en promedio los alumnos con el quintil más alto son los que hacen más actividades. También del análisis descriptivo se observó que la proporción de alumnos que hacen actividades dentro de una clase puede llegar a ser baja, siendo común observar clases que caen en el primer cuartil, la proporción que realizan actividades varía concentrándose en abril y agosto los meses con más actividad.

Se ha proporcionado un extenso marco teórico sobre el modelo BART, en conjunto con el flujo de trabajo esperado en un proyecto de ciencia de datos y su debida implementación con la librería `tidymodels`. Se destaca que el extenso uso de `tidymodels` ha permitido la reproducibilidad del proyecto.

Se presentaron tres modelos predictivos con distintas entradas. El primer modelo consideró como individuos a todos los estudiantes y tenía como objetivo la predicción de puntajes. Se evaluaron diversos modelos de regresión, y se destacó el rendimiento superior del modelo BART. Aunque este se desempeña bien como modelo de regresión, presenta dificultades al convertirlo en un modelo de clasificación debido a la

proximidad entre niveles. Se identificó que entre las variables más importantes se encuentran aquellas relacionadas con la clase, como grado, departamento y contexto sociocultural.

El segundo modelo tiene como individuos las clases, como resultado destaca la importancia que tiene la variable de si un alumno pertenece a sexto grado o no.

El tercer modelo tiene como individuos a los alumnos de sexto grado y tiene como objetivo predecir si un alumno alcanza el nivel deseado o no. Otra vez destaca la mejor performance del modelo BART en comparación con el modelo Random Forest. Este modelo permite identificar en julio, con buen nivel de precisión, si un alumno va a alcanzar el nivel deseado a fin de año, permitiendo aplicar intervenciones específicas para mejorar el desempeño del estudiante. Debido a la naturaleza de BART, se puede cuantificar la incertidumbre sobre la probabilidad asignada a que el alumno alcance el nivel deseado. El uso de efectos aleatorios permitió identificar centros educativos cuyo efecto difiere significativamente de 0, tanto positiva como negativamente. A través de individuos ficticios se pudo cuantificar como el uso de la plataforma afecta el desempeño en la prueba final.

Como problemática se identificó la falta de control en las actividades hechas por los alumnos, es sabido que en las instancias de las clases presenciales hay estudiantes que comparten el dispositivo, por lo que hay actividades hechas por varios estudiantes y que solo figuran para uno. La falta de dispositivos permite la situación de que un estudiante figure con pocas actividades hechas, pero en realidad haya trabajado todo el año. Es desconocido que tan frecuente se da esta situación. Podría ser útil para palear este efecto, saber cuándo los estudiantes entran a la plataforma y si lo hacen fuera de horario escolar. No tener la dificultad asociada a las tareas que realiza

el alumno complejiza inferir el nivel que tiene el alumno en el transcurso del año. Tampoco se cuenta con el nivel previo al inicio de clases que tiene el alumno o si el alumno tiene clases de forma particular. El inglés es un idioma que está presente culturalmente, lo que significa que algunos estudiantes pueden estar expuestos a este idioma fuera del entorno escolar, debido a su naturaleza, este factor resulta difícil de cuantificar.

A modo de futuro trabajo, se plantea explorar nuevas técnicas estadísticas como la selección de variables con BART propuesta en (Bleich et al., 2014) y en el trabajo sin publicar, (Luo & Daniels, 2021) donde se proponen métodos más refinados para corregir el sesgo a favor de las variables continuas. También se propone utilizar técnicas de predicción conformal (Vovk et al., 2022) para predecir un conjunto formado por niveles de inglés que cubran con cierta confianza el nivel de inglés que el alumno pueda obtener en la prueba final. Por último, se plantea la propuesta de realizar un experimento controlado en el que algunas clases reciban el tratamiento de utilizar la plataforma Little Bridge y otras no. Es importante eliminar todos los factores de confusión, por lo que se tienen que buscar clases con similares características. Además, es necesario medir correctamente las actividades de todos los alumnos pertenecientes a las clases que reciban el tratamiento. Este experimento permitiría estimar el efecto de utilizar la plataforma Little Bridge en la enseñanza de inglés.

Referencias

Andrieu, C., Freitas, N., Doucet, A., & Jordan, M. (2003). An Introduction to MCMC for Machine Learning. *Machine Learning*, *50*, 5-43. https://doi.org/10.1023/A:1020281327116

Bleich, J., Kapelner, A., George, E. I., & Jensen, S. T. (2014). Variable selection for BART: An application to gene regulation. *The Annals of Applied Statistics*, *8*(3), 1750-1781. https://doi.org/10.1214/14-AOAS755

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chen, P. P.-S. (1976). The entity-relationship model—toward a unified view of data. *ACM Trans. Database Syst.*, *1*(1), 9-36. https://doi.org/10.1145/320434.320440

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., & contributors, X. (2023). *xgboost: Extreme Gradient Boosting*. https://github.com/dmlc/xgboost

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266-298. https://doi.org/10.1214/09-AOAS285

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Daniel, F., Weston, S., & Tenenbaum, D. (2022). *doParallel: Foreach Parallel Adaptor for the 'parallel'' Package*. https://github.com/RevolutionAnalytics/doparallel

Dorie, V. (2023). dbarts: Discrete Bayesian Additive Regression Trees Sampler [R package version 0.9-23]. https://CRAN.R-project.org/package=dbarts

Dowle, M., & Srinivasan, A. (2023). *data.table: Extension of 'data.frame'* [https://r-datatable.com, https://Rdatatable.gitlab.io/data.table, https://github.com/Rdatatable/data.table].

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, *19*(1), 1-67. https://doi.org/10.1214/aos/1176347963

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721-741. https://doi.org/10.1109/TPAMI.1984.4767596

Gentleman, R., & Temple Lang, D. (2007). Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, *16*(1), 1-23.

Greenwell, B. M., & Boehmke, B. C. (2020). Variable Importance Plots—An Introduction to the vip Package. *The R Journal*, *12*(1), 343-366. https://doi.org/10.32614/RJ-2020-013

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97-109. https://doi.org/10.1093/biomet/57.1.97

Hoff, P. D. (2009). *A first course in Bayesian statistical methods* (Vol. 580). Springer.

Karatzoglou, A., Smola, A., & Hornik, K. (2023). *kernlab: Kernel-Based Machine Learning Lab* [R package version 0.9-32]. https://CRAN.R-project.org/package=kernlab

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, *11*(9), 1-20. https://doi.org/10.18637/jss.v011.i09

Kuhn, M. (2023). *tune: Tidy Tuning Tools* [R package version 1.1.2]. https://CRAN.R-project.org/package=tune

Kuhn, M., & Frick, H. (2023). *dials: Tools for Creating Tuning Parameter Values* [R package version 1.2.0]. https://CRAN.R-project.org/package=dials

Kuhn, M., & Silge, J. (2022, julio). *Tidy Modeling with R*. O'Reilly Media, Inc.

Kuhn, M., & Vaughan, D. (2023). *parsnip: A Common API to Modeling and Analysis Functions* [R package version 1.1.1]. https://CRAN.R-project.org/package=parsnip

Kuhn, M., Vaughan, D., & Hvitfeldt, E. (2023). *yardstick: Tidy Characterizations of Model Performance* [R package version 1.2.0]. https://CRAN.R-project.org/package=yardstick

Kuhn, M., & Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.* https://www.tidymodels.org

Kuhn, M., Wickham, H., & Hvitfeldt, E. (2023). *recipes: Preprocessing and Feature Engineering Steps for Modeling* [R package version 1.0.8]. https://CRAN.R-project.org/package=recipes

Luo, C., & Daniels, M. (2021, diciembre). *Variable Selection Using Bayesian Additive Regression Trees*. https://doi.org/10.48550/arXiv.2112.13998

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, *21*(2), 239-245. https://doi.org/10.1080/00401706.1979.10489755

Tan, Y., Flannagan, C., & Elliott, M. (2016). Predicting human-driving behavior to help driverless vehicles drive: random intercept Bayesian Additive Regression Trees. *Statistics and Its Interface*, *11*(4), 557-572. https://doi.org/10.4310/SII.2018.v11.n4.a1

Tan, Y. V., & Roy, J. (2019). Bayesian additive regression trees and the General BART model. *Statistics in Medicine*, *38*(25), 5048-5069. https://doi.org/https://doi.org/10.1002/sim.8347

Vaughan, D., & Couch, S. (2023). *workflows: Modeling Workflows* [R package version 1.1.3]. https://CRAN.R-project.org/package=workflows

Vovk, V., Gammerman, A., & Shafer, G. (2022). Conformal Prediction: Classification and General Case. En *Algorithmic Learning in a Random World* (pp. 71-106). Springer International Publishing. https://doi.org/10.1007/978-3-031-06649-8_3

Wickham, H., Girlich, M., Fairbanks, M., & Dickerson, R. (2023). *dtplyr: Data Table Back-End for 'dplyr'* [https://dtplyr.tidyverse.org, https://github.com/tidyverse/dtplyr].

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1), 1-17. https://doi.org/10.18637/jss.v077.i01

Apéndice

A. Estructura de datos

Campo	Descripción
id_persona	Identificador del estudiante.
departamento	Departamento del estudiante.
zona	Zona: Rural o Urbana.
grado	Grado al que pertenece el estudiante (4, 5, 6).
id_centro	Identificador del centro al que pertenece el estudiante
contexto_sociocultural	Contexto sociocultural
actividades_acumuladas_**MES**	Actividades acumuladas hasta el mes determinado.
preguntas_acumuladas_**MES**	Preguntas realizadas acumuladas hasta el mes determinado.
correctas_acumuladas_**MES**	Respuestas correctas acumuladas hasta el mes determinado.
intenta.prom_**MES**	$\frac{intentos_realizados}{actividades_realizada}$ en el mes determinado
activ.prom_**MES**	Actividades promedio realizada por día.
pregu.prom_**MES**	Cantidad de preguntas por actividad promedio en el mes determinado.
pts.min_**MES**	Puntos mínimos por actividad promedio en el mes determinado.
pts.max_**MES**	Puntos máximos por actividad para el mes determinado.
acierto_**MES**	Acierto promedio en el mes determinado.
dias_acumulados_**MES**	Cantidad de días en los que hizo actividades.
intentos_acumulados_**MES**	Intentos acumulados hasta el mes determinado.
actividades_asignadas_acumuladas_**MES**	Actividades asignadas acumuladas hasta el mes determinado.
porc_actividades_hechas_**MES**	$\frac{actividades\_acumuladas\_\boldsymbol{MES}}{actividades\_asignadas\_acumuladas\_\boldsymbol{MES}}$
mensajes_enviados_acumulados_**MES**	Mensajes enviados acumulados hasta el mes determinado.
mensajes_recibidos_acumulados_**MES**	Mensajes recibidos acumulados hasta el mes determinado.
mensajes_hilos_acumulados_**MES**	Cantidad de hilos hasta el mes determinado.
delta_**MES**	Indicador de la actividad del alumno en la plataforma CREA. $\frac{log(1+sum(actividad\_registrada))}{log(1+dias\_del\_mes)}$
class_lb_**MES**	Clase a la que pertenece el alumno en el mes.
puntos	Puntos obtenidos en la prueba final.
nivel	Nivel obtenido en la prueba final.

Descripción de la estructura de datos para el modelo predictivo para alumnos.

Campo	Descripción
class_lb	Identificador de la clase.
departamento	Departamento más frecuente dentro de la clase.
grado	Grado más frecuente dentro de la clase.
contexto_sociocultural	Contexto sociocultural más frecuente dentro de la clase.
zona	Zona más frecuente dentro de la clase.
cant_alumnos_**MES**	Cantidad de alumnos que realizaron actividades en el mes determinado.
n_total_alumnos_**MES**	Número total de alumnos en el mes determinado.
act_realizadas_**MES**	Promedio de actividades realizadas por alumno en el mes determinado.
preguntas_realizadas_**MES**	Promedio de preguntas realizadas por alumno en el mes determinado.
correctas_realizadas_**MES**	Promedio de preguntas realizadas por alumno en el mes determinado.
mean_act_realizadas_**MES**	Promedio de intentos realizado por actividad en el mes determinado.
mean_preguntas_realizadas_**MES**	Promedio de preguntas realizadas por actividad en el mes determinado.
mean_correctas_realizadas_**MES**	Promedio de respuestas correctas por actividad realizadas en el mes determinado.
var_act_realizadas_**MES**	Varianza de intentos realizados en el mes por actividad determinado.
var_preguntas_realizadas_**MES**	Varianza de preguntas realizadas por actividad en el mes determinado.
var_correctas_realizadas_**MES**	Varianza de respuestas correctas realizadas por actividad en el mes determinado.
n_actividades_asignadas_**MES**	Número de actividades asignadas en el mes determinado.
prop_alumnos_**MES**	Proporción de alumnos que realizaron actividades en el mes determinado.
ptos	Puntaje promedio obtenido por los alumnos que realizaron la prueba final.

Descripción de la estructura de datos para el modelo predictivo de clase.

Campo	Descripción
id_persona	Identificador del estudiante.
departamento	Departamento del estudiante.
zona	Zona: Rural o Urbana.
id_centro	Identificador del centro al que pertenece el estudiante
contexto_sociocultural	Contexto sociocultural
actividades_acumuladas_**MES**	Actividades acumuladas hasta el mes determinado.
preguntas_acumuladas_**MES**	Preguntas realizadas acumuladas hasta el mes determinado.
correctas_acumuladas_**MES**	Respuestas correctas acumuladas hasta el mes determinado.
intenta.prom_**MES**	$\frac{intentos_realizados}{actividades_realizada}$ en el mes determinado
activ.prom_**MES**	Actividades promedio realizada por día.
pregu.prom_**MES**	Cantidad de preguntas por actividad promedio en el mes determinado.
pts.min_**MES**	Puntos mínimos por actividad promedio en el mes determinado.
pts.max_**MES**	Puntos máximos por actividad para el mes determinado.
acierto_**MES**	Acierto promedio para el mes determinado.
días_acumulados_**MES**	Cantidad de días en los que hizo actividades.
intentos_acumulados_**MES**	Intentos acumulados hasta el mes determinado.
actividades_asignadas_acumuladas_**MES**	Actividades asignadas acumuladas hasta el mes determinado.
porc_actividades_hechas_**MES**	$\frac{actividades\_acumuladas\_\boldsymbol{MES}}{actividades\_asignadas\_acumuladas\_\boldsymbol{MES}}$
mensajes_enviados_acumulados_**MES**	Mensajes enviados acumulados hasta el mes determinado.
mensajes_recibidos_acumulados_**MES**	Mensajes recibidos acumulados hasta el mes determinado.
mensajes_hilos_acumulados_**MES**	Cantidad de hilos hasta el mes determinado.
class_lb_**MES**	Clase a la que pertenece el alumno en el mes.
puntos	Puntos obtenidos en la prueba final.
nivel	Nivel obtenido en la prueba final.

Descripción de la estructura de datos del modelo predictivo para alumnos de sexto grado.

B. Gráfico de cajas

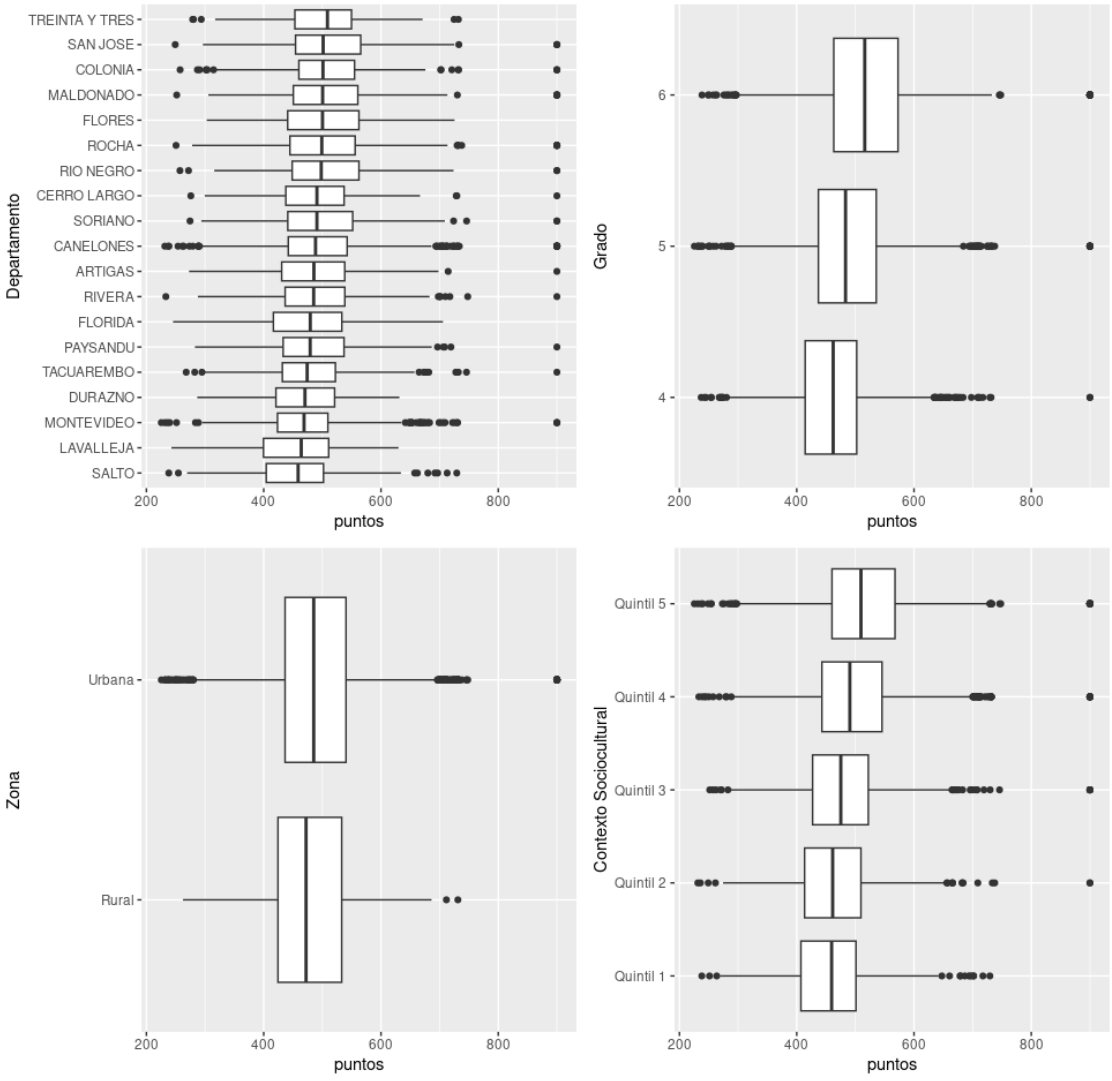


Gráfico de caja de puntos por variables del estudiante.

C. Hiperparámetros por modelo

Modelo	Fijados	Ajustados
Random Forest	trees = 1000	mtry = 85, min_n = 3
BART	Se dejó por defecto el número de cadenas (1), la cantidad de muestras de la posterior (1000) y la cantidad de muestras que se descartan (100)	trees = 482, prior_terminal_node_coef = 0.702, prior_terminal_node_expo = 2.23, prior_outcome_range = 0.795
XGBoost	trees = 1000	tree_depth = 5, min_n = 11, loss_reduction = 0.02, sample_size = 0.892, mtry = 68, learn_rate = 0.0346
Support Vector Regression	degree = 2	cost = 1.27, margin = 0.143

Hiperparámetros por modelo diferenciando los que se fijaron y los que se ajustaron.