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

Bruno Tancredi

Instituto de Estadística - Facultad de Ciencias Económicas y de Administración, Universidad de la República, Uruguay

Fondo Sectorial de Educación Inclusión digital 2020 - Modalidad financiado por ANIII. Proyecto "Monitor y evaluación de uso de las plataformas educativas"

Resumen

En este trabajo se utilizan datos de uso de plataformas educativas del programa Ceibal en Inglés, en particular Little Bridge y se vincula el uso de dicha plataforma con el rendimiento en las pruebas adaptativas de Inglés. En los modelos predictivos se combina información sociodemográfica de los estudiantes y el centro, en conjunto con indicadores relacionados a la actividad del estudiante, para predecir el desempeño educativo. La implementación de los modelos predictivos se hizo utilizando intensivamente la biblioteca `tidymodels`.

Descripción del problema y datos a utilizar

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 públicas. En 2021, comenzaron a utilizar la plataforma Little Bridge (LB), donde los estudiantes realizan actividades asignadas por docentes a distancia. Un objetivo del trabajo es predecir el desempeño de los estudiantes en la prueba final de diciembre utilizando datos recopilados hasta julio.

Rplot

Figura 1: Plataforma Little Bridge

- Se utilizaron datos proporcionados por Ceibal, que incluyen tareas en LB, resultados de pruebas adaptativas, actividades asignadas por profesores, actividades en otras plataformas Ceibal, y registros de mensajes en la plataforma. Para trabajar con estos volúmenes de datos, se utilizo la librería `data_table`.
- El total de alumnos es de 71222, la prueba adaptativa la dieron 12449. De estos 12449 solo para 8663 estudiantes hay registro de las actividades realizadas y asignadas.
- Se construyeron indicadores relacionados a las actividades realizadas por los alumnos, tanto tareas como mensajes enviado, además de indicadores relacionados a las actividades todos estos agrupados por mes.

Modelos predictivos

Se utilizó el framework `tidymodels` para la preparación y el ajuste de distintos modelos.

- Se ajustaron distintos modelos. Random Forest(RF), Bayesian Additive Regression Trees Additive Trees (BART), XGBoost (XGB) y Support Vector Regression (SVR) con el kernel polinomial.
- Se usó cross-validation con 10 particiones para encontrar los mejores hiperparámetros para cada modelo. Se creó una cuadrícula de búsqueda utilizando el método de hipercubos latinos, con 20 puntos para cada modelo. La selección de la mejor combinación de hiperparámetros se basó en la Raíz del Error Cuadrático Medio (RMSE).
- Se comparó los distintos modelos utilizando el RMSE de vuelto en la validación cruzada para la mejor combinación de hiperparámetros.

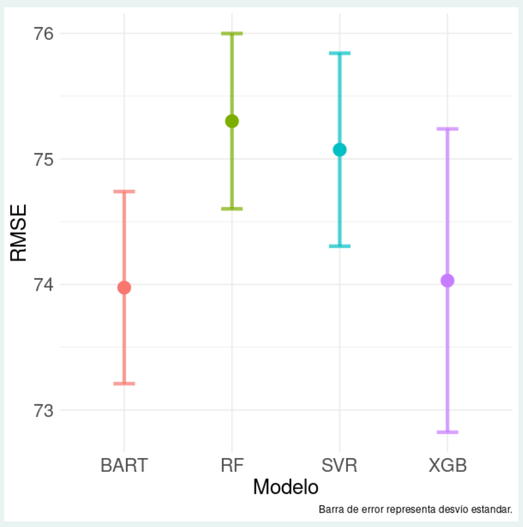


Figura 2: Comparación de distintos modelos predictivos

Debido a que el modelo BART fue el modelo que presento más bajo RMSE, se decidió utilizarlo como modelo final.

Resultado de modelo predictivo (I)

Se reportó un RMSE de 74.3. Se observa que el modelo tiene un sesgo para predecir resultados relacionados al nivel más bajo de inglés y más alto, en conjunto se percibe que el modelo falla a la hora de predecir el nivel de inglés del alumno.

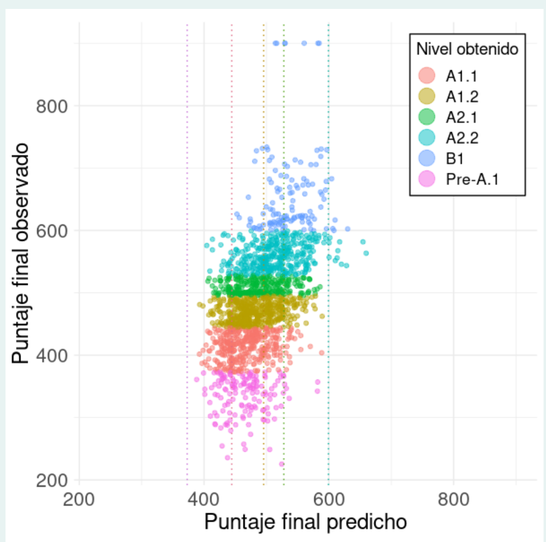


Figura 3: Puntaje final predicho contra puntaje final observado

Resultado de modelo predictivo (II)

Para calcular la importancia de las variables, se implemento un indicador que calcula la suma de veces que la variable aparece en los árboles de las 1000 muestras posteriores, ponderando por la cantidad de observaciones que existen en el nodo de la variable.

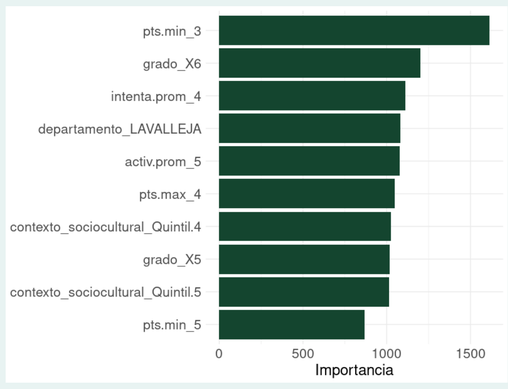
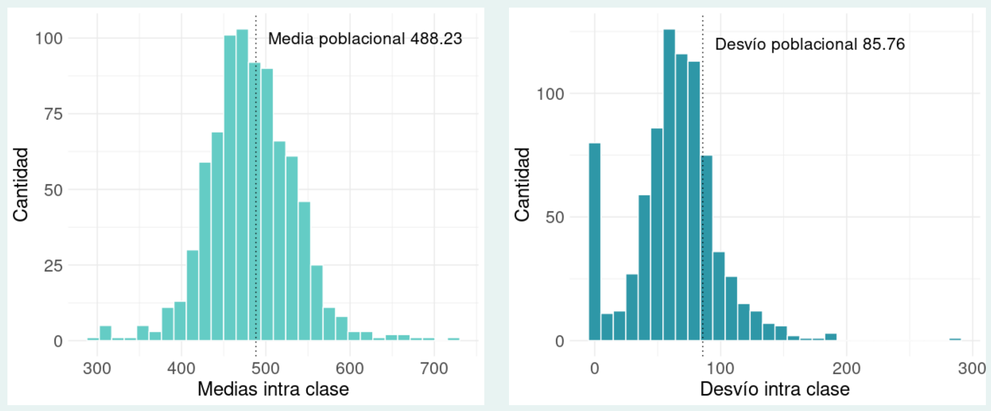


Figura 4: Importancia de variables para modelo predictivo.

Descripción de clases

Debido a que dentro de las variables más importantes figuran variables de la clase (Grado, departamento, contexto sociocultural) se decidió crear un modelo cuyos individuos sean las clases y la variable a predecir el promedio del puntaje final, utilizando datos hasta julio.



(a) (b)

Figura 5: Media (a) y Desvío Estándar (b) de los resultados dentro de la clase

- Existen 852 clases en la que al menos un alumno hizo la prueba y le fueron asignadas actividades, no todos los alumnos realizaron la prueba.
- 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 construyeron variables para indicar la proporción de alumnos por clase. Las clases varían de tamaño durante el año ya que un alumno puede cambiar de clase.

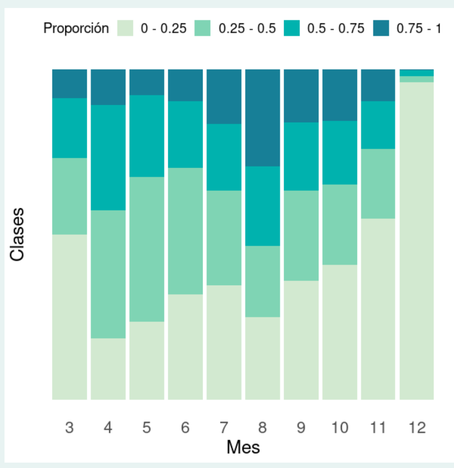


Figura 6: Proporción de alumnos que realizaron actividades

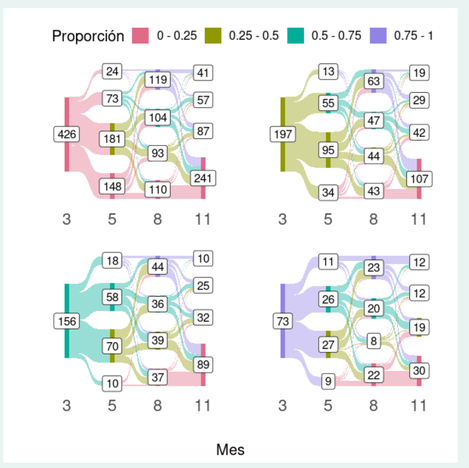


Figura 7: Evolución de la proporción de clases para meses seleccionados.

Modelo predictivo de clases

Se utilizó Random Forest para la predicción del puntaje promedio final dentro la clase, se reportó un RMSE de 42.7. Se observó la importancia que tiene que la clase pertenezca a sexto grado frente a los otros dos grados.

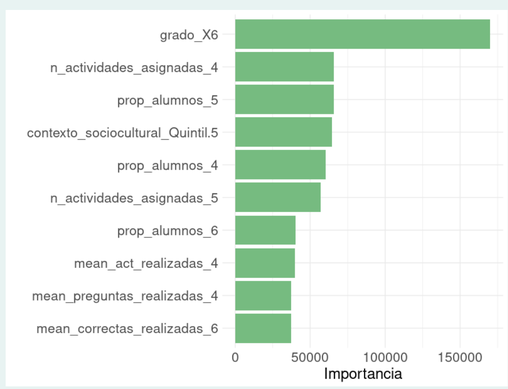


Figura 8: Importancia de variables para modelo de clase.

Conclusiones y futuro trabajo

- Debido a que el puntaje es cercano entre niveles el modelo falla al predecir el nivel final.
- La baja proporción de alumnos que realizan actividades complejiza la predicción individual de resultados.
- Estudiar el efecto de la clase, por ejemplo, utilizando BART para estudiar modelos mixtos.