Concurso MásMAT

Equipo de Estadística Fundación Bunge y Born

Noviembre 2021

Informe Final Fase de Competencia Concurso MásMAT

Resumen ejecutivo

En el marco del Concurso MásMAT, este informe se centra en la Fase de Competencia en la cual los 3 equipos finalistas recolectaron los datos de los usuarios que descargaron sus aplicaciones en Google Play Store. Este informe está estructurado en las siguientes secciones:

- Experimento: Una introducción al análisis experimental propuesto para evaluar los resultados del concurso.
- Diseño del juego, Muestreo: Una explicación de los criterios de evaluación embebidos en las 3 aplicaciones que posibilitaron la evaluación simultánea de las mismas.
- Proceso Fase de Competencia: Una explicación de los desarrollos acontecidos en la mencionada fase, las características de la competencia y como fue recolectada la información para la evaluación
- Resultados: Esta sección se divide en dos dimensiones: en la primera se mencionan algunos indicadores de performance de las aplicaciones; en el segundo se profundiza en el análisis estadístico de los datos mediante el cual este informe evaluó cuál de las 3 aplicaciones performó mejor en cuanto a propiciar avances en las habilidades matemáticas de los usuarios.
- Conclusiones: En esta sección se presentan las conclusiones surgidas del análisis de los datos mencionado en la sección anterior.

Introducción

A lo largo de cuatro meses (del 28 de junio al 28 de octubre) se llevó a cabo la fase de competencia del Concurso MásMAT. Durante esta etapa, las 3 aplicaciones finalistas estuvieron recolectando los datos de sus usuarios en una base de datos de propiedad de la Fundación Bunge y Born (securizada por la empresa ITRock), para que el equipo estadístico pueda evaluar el avance en el conocimiento matemático de quienes jugaron estos 3 videojuegos.

Para no alterar la evaluación estadística, ninguno de los equipos pudo alterar el código durante los 4 meses de competencia. Además, se siguió una política estricta en cuanto a difusión de

las apps en la cual los equipos no podían comunicar más allá de ciertas excepciones, ya que la pauta publicitaria fue desarrollada de manera equitativa entre las 3 aplicaciones por el equipo de comunicación de la Fundación Bunge y Born. Por parte de la Fundación, la evaluación de los datos fue realizada de manera anonimizada por el equipo estadístico. En la base de datos OxTravel tomaba el nombre de "App 1," Matenautas "App 2," y Mora II "App 3," por este motivo en el apartado estadístico aparecen ambos nombres como referencia para cada aplicación.

El objetivo del concurso fue evaluar si el paso por las aplicaciones (Matenautas, Mora II y OxTravel) generaba un progreso en habilidades matemáticas en niños de 9 a 12 años. Los 3 juegos tuvieron que embeber en su infraestructura 8 preguntas iniciales, 6 niveles de preguntas secuenciales y 8 preguntas finales en caso de que algún usuario llegara a esa instancia. Los dos modelos estadísticos utilizados (explicados en las siguientes secciones del informe) dieron como resultado un claro ganador en los términos de nuestra evaluación de conocimiento matemático.

Objetivos

Los objetivos de la Fase de Competencia del Concurso estuvieron dados por evaluar cuál de las 3 aplicaciones finalistas produjo un mayor avance en las habilidades matemáticas de los usuarios. Se realizó un análisis experimental basado en los datos recopilados por las aplicaciones para poder determinar cual de los 3 juegos performó mejor, es decir, cuál de las 3 aplicaciones generó un mayor avance en las habilidades matemáticas de los usuarios a partir del uso de la misma.

Experimento

El diseño consistió en un análisis experimental para la evaluación del desempeño de las aplicaciones. A cada participante se le asignó aleatoriamente una aplicación y pasó luego a medir una línea de base. Posteriormente se desarrolló un juego común, basado en preguntas de evaluación (iniciales, secuenciales y finales), embebido en las 3 aplicaciones, y se monitoreó el progreso de cada usuario respecto a su propia línea de base (las respuestas a las 8 preguntas

iniciales similares para cada aplicación).

Diseño del Juego

Reglas Comunes

Para poder evaluar el desempeño de las aplicaciones, se crearon reglas comunes a todas las aplicaciones. Cada una de ellas debió implementar un sistema de evaluación y seguimiento sobre cada uno de los participantes. Las mismas consistieron en:

- 6 niveles de aprendizaje.
- A su vez, cada nivel de aprendizaje se subdividió en 4 áreas de aprendizajes $(\mathrm{A}-\mathrm{B}-\mathrm{C}-\mathrm{D})$
- Al inicar el nivel, el jugador contó con un estado:

$$\{\mathrm{A}:0_{pts},\mathrm{B}:0_{pts},\mathrm{C}:0_{pts},\mathrm{D}:0_{pts}\}$$

y un contador de errores

$$\{\mathrm{E}=0\}$$

- A cada jugador se le administró de forma secuencial testeos en forma de preguntas, la cuales estaban referenciadas a un nivel y área.
- Dichas preguntas se sortearon de forma estratificada según área, del pool de preguntas posibles para el nivel, y se le presentaron al jugador en el siguiente orden:

$$\mathrm{A}-\mathrm{B}-\mathrm{C}-\mathrm{D}-\mathrm{A}-\mathrm{B}-\mathrm{C}-\ldots.$$

- Por cada pregunta respondida de forma correcta, acumuló 1_{pto} en el área correspondiente.
- Por cada pregunta respondida de forma incorrecta, acumuló 1_{pts} en su contador de errores.

- En caso de que el jugador hubiera logrado obtener 2_{pts} en un área a lo largo del nivel, esta se consideró como aprobada y dejó de ser evaluada hasta el próximo nivel. Con tal fin, se excluyó dicha área del sorteo de posibles preguntas.
- Si un jugador obtuviese 2_{pts} en cada una de las áreas, entonces pasa de nivel. En otras palabras:

$$\text{Si } \{\text{A} : 2_{pts}, \text{B} : 2_{pts}, \text{C} : 2_{pts}, \text{D} : 2_{pts}\} \longrightarrow \text{ Sube de Nivel}$$

- Si un jugador obtuviera un valor en su contador de errores igual a K, entonces baja de nivel. En otras palabras:

$$\text{Si } \{\text{E} = K\} \longrightarrow \text{ Baja de Nivel}$$

- Si un jugador por haber cometido demasiados errores le correspondiere bajar de nivel y este estuviera en el nivel 1, al ser el nivel 1 el más bajo de todos, se mantendrá en el. En otras palabras:

$$\text{Nivel} = \text{máx}(1, \text{Nivel})$$

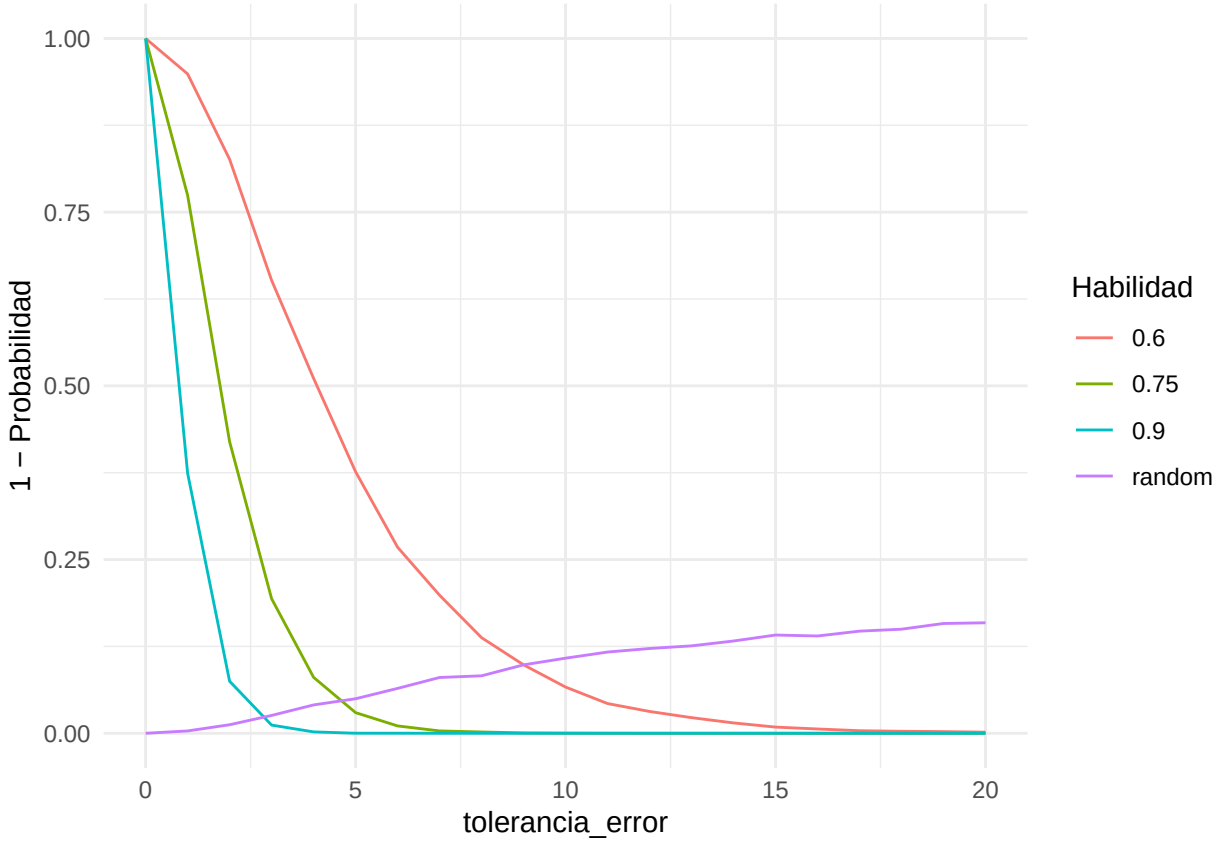
- Si un jugador lograra pasar el nivel 6, entonces gana el juego.

Probabilidad de Pasar de Nivel

Se eligió un K tal que penalizara al jugador que eligiese al azar y al mismo tiempo, que no desmotivase al jugador que cometió un error sincero. Para ello, a partir de una simulación suponiendo un 24 % de preguntas dicotómicas y un 76 % de preguntas trinarias, se obtuvo el comportamiento promedio de las reglas anteriormente descritas.

Si se grafica la probabilidad de pasar de nivel de quien elige al azar (línea violeta) y la probabilidad de bajar de nivel de quien elige a conciencia bajo diferentes niveles de habilidad (línea roja=.6 verde=.75 y azul=.9) se obtiene que se deberiera elegir un $\text{K} < 9$. Así, la probabilidad de pasar de nivel de forma aleatoria sería menor al 10 %. En particular, se eligió

un K = 5, tal que aquel que quien tuviera un nivel de juego por encima del azar (0.6 vs 0.5) le hubiera de tomará 3 veces más completar el juego que aquel que respondiese perfecto.



Línea de Base

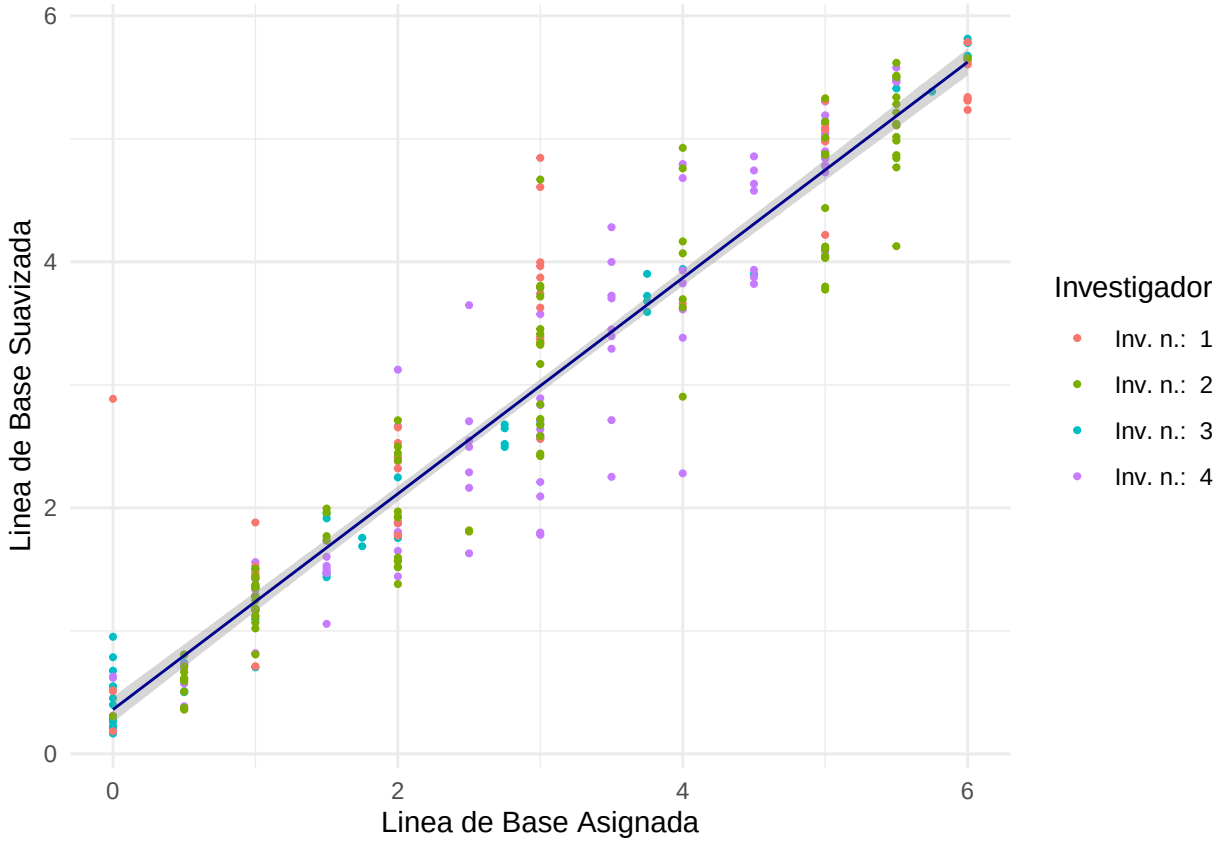
Para la línea de base se utilizaron 8 preguntas:

1. El resultado de 15x100 es:
2. Martín compró 2 facturas y gastó $44
3. ¿Qué otro producto tiene el mismo resultado que 8x5?
4. Lucio coloca empanadas en una bandeja en 6 filas de 8 empanadas cada una
5. ¿Cuántas veces entra el 20 en 3080?
6. ¿Cuál de estos números puede ser el resultado de una multiplicación por 5?
7. Sol preparó ¼ litro de leche chocolatada para cada uno de sus 6 sobrinos. ¿Cuántos

litros de leche chocolatada preparó en total?

8. En una papelera tienen una oferta de invitaciones para casamiento. Se puede elegir entre 3 colores de fondo

Ahora bien, el resultado de estas preguntas arroja un total de 256 combinaciones posibles. Con el fin de generar un score como línea de base, los investigadores miembros puntuaron un conjunto de combinaciones factibles y luego se entrenó un modelo no lineal para asignar el índice. De esta manera, la línea de base resultante fue la predicción promedio para cada una de las puntuaciones.



Captación de Usuarios

Para la estrategia de captación de usuarios se invirtieron ARS$1.705.934 en la plataforma de Google ADS, divididos de manera equitativa entre los 3 equipos, buscando que el público objetivo (niños de 9 a 12 años) descarguen las apps a partir de anuncios en el Google Play Store.

Si bien los anuncios se realizaron de manera similar para los 3 juegos, las diferencias estéticas, el recorrido y la notoriedad previa de los integrantes de los equipos y la conversión en materia de clics y descargas implicó que contemos con diferentes números de descargas, de los cuales no podemos saber cuántas correspondieron efectivamente a población de nuestro grupo objetivo.

Muestreo

Para el muestreo, se aparearon los departamentos para conformar clusters de tamaño 3, equivalentes en sus características, sobre las siguientes variables:

- total de Personas,
- total de Hogares,
- total de Viviendas,
- total de Viviendas Habitadas,
- total de Personas con nivel educativo menor igual a primario completo o incompleto
- total de Personas con nivel educativo mayor igual a universitario completo o incompleto
- total de hogares con al menos un indicador de NBI
- total de viviendas en zonas rurales

Así, cada aplicación capturó participantes de forma equivalente a lo largo del territorio argentino.

Con este procedimiento, se eliminó el riesgo de que hubiera algún sesgo en la captura del dato para cada una de las aplicaciones.

Proceso Fase de Competencia

Durante 4 meses se invirtió en pauta publicitaria para las 3 aplicaciones con el objetivo de captar usuarios para las mismas. Esto fue llevado a cabo a través de un diseño publicitario similar para los 3 juegos, estratificado en cuanto a horarios (para obtener descargas del grupo

etario de 9 a 12 años) y compartimentado según la distribución geográfica en departamentos con características equivalentes como se mencionó en la sección anterior.

A partir de esto, durante todo el período de competencia recolectamos los datos de cada niño que jugó en cada una de las apps en nuestra base de datos. Solo recibimos la información referente a las preguntas de nuestra evaluación, es decir que no les pedimos a los equipos que nos otorgaran acceso al resto de los mini juegos o preguntas propias de cada aplicación.

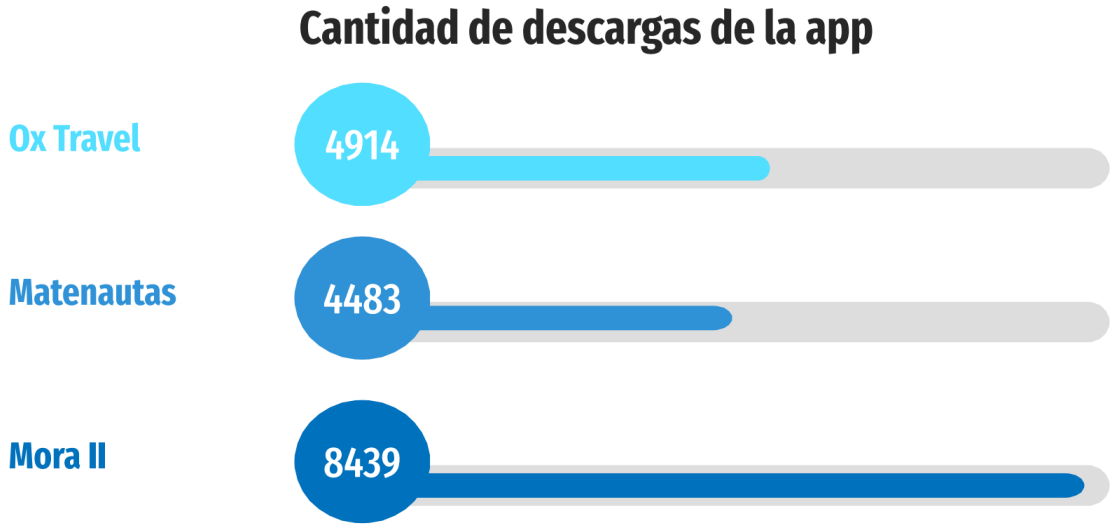
De esta manera, la evaluación fue realizada sobre una base de criterios comunes con el objetivo de comparar el rendimiento de cada aplicación a partir de una herramienta exactamente igual para cada uno de los equipos. El objetivo fue que cada aplicación obtuviera resultados a partir de diferentes mecanismos de llegar al aprendizaje matemático, propio del expertise y las herramientas particulares de cada equipo. Una evaluación similar para las 3 aplicaciones, embebida en la interfaz de cada aplicación, nos permitió comparar las diferentes experiencias de los usuarios en cada aplicación bajo un mismo criterio.

Luego de terminar la Fase de Competencia recopilamos todos los datos impactados en nuestra base para realizar el análisis que sigue a continuación, y así determinar de la manera más rigurosa posible cuál fue la aplicación que presentó una mejor performance respecto a la mejora de habilidades matemáticas en sus usuarios.

Resultados

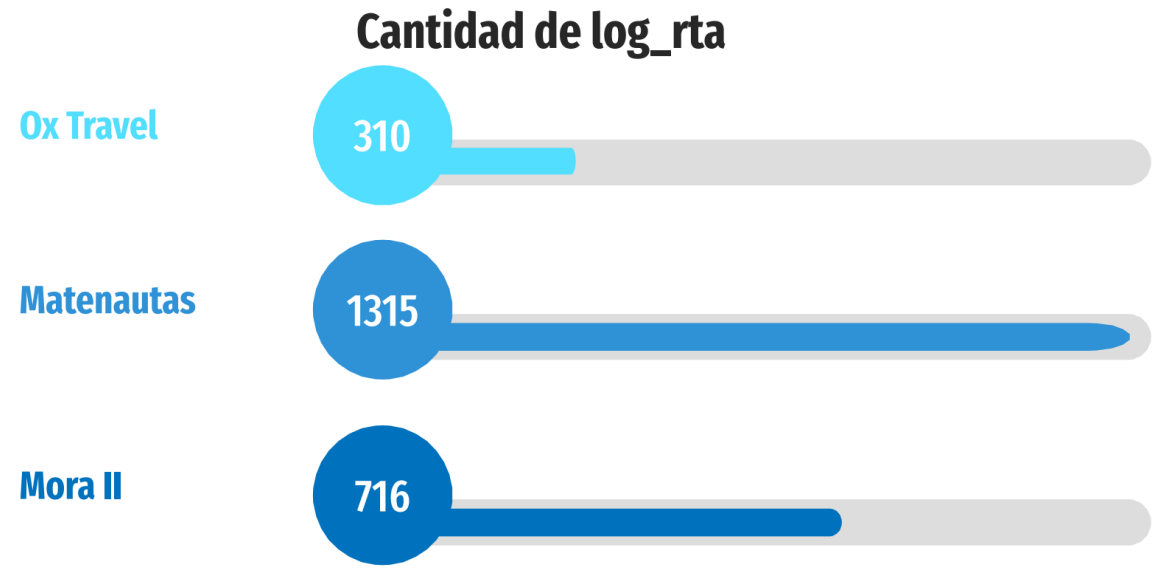
Indicadores de performance de las aplicaciones

Además de los resultados de la evaluación, compartimos también algunos datos clave vinculados a la experiencia de usuario de quienes descargaron y jugaron con las aplicaciones.

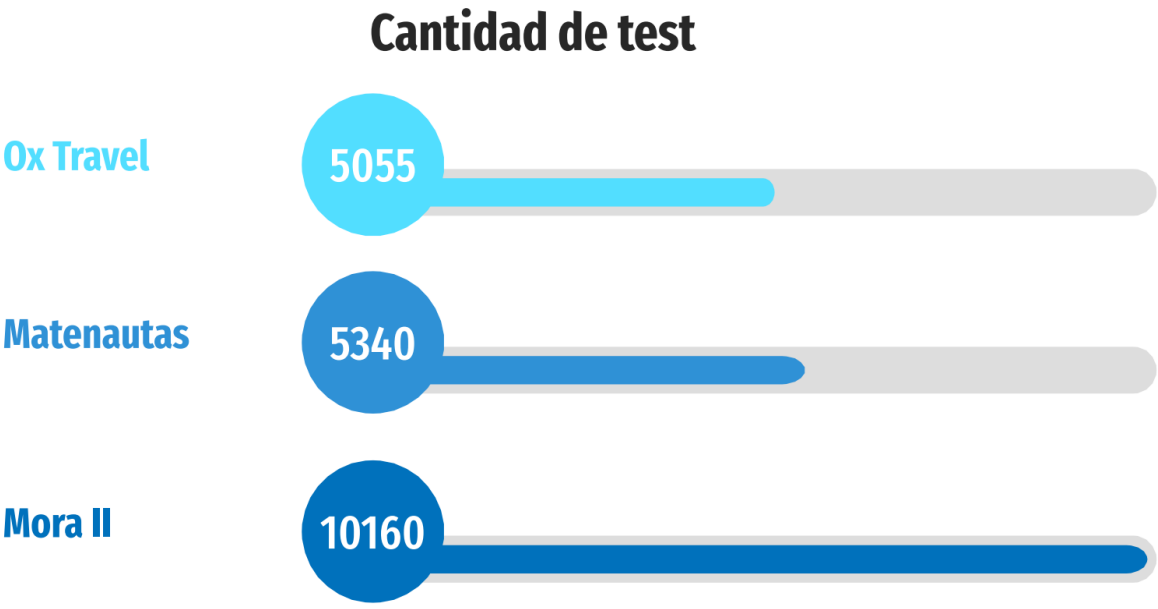


En términos de descargas, Mora II (App 3) fue la aplicación que contó con un mayor número crudo de instalaciones totales.

Cada usuario de OxTravel (App 1) jugó en promedio 3,71 sesiones, los de Matenautas (App 2) jugaron un promedio de 2,44 sesiones, mientras que los usuarios de Mora II (App 3) presentaron un promedio de 1,85.



El indicador de Log Rta fue el principal indicador en términos de evaluación, ya que implicó la cantidad de respuestas impactadas en los niveles de nuestra evaluación luego de las preguntas iniciales. En este apartado Matenautas (App 2) recopiló la mayor cantidad con 1315, seguido de Mora II (App 3) con 716 y OxTravel (App 1) con 310.



El indicador de cantidad de Test implica la cantidad de registros de preguntas iniciales que tuvo cada aplicación. En este indicador Mora II (App 3) produjo 10160 casos, Matenautas (APP 2) 5340 y OxTravel (App 1) 5055.

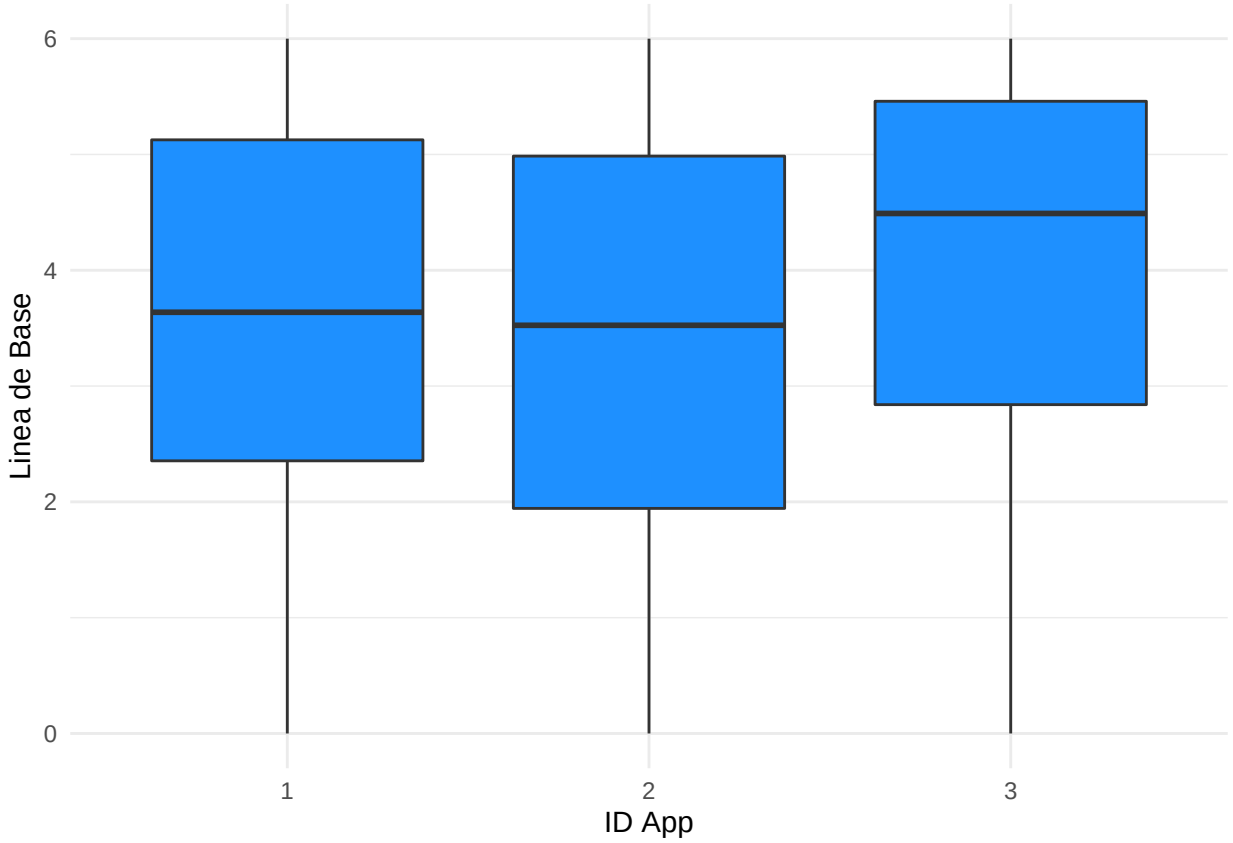
Para la evaluación estadística, tomamos como usuarios válidos solo a aquellos participantes que terminaron las 8 preguntas iniciales, ya que al no contar con la totalidad de las respuestas, los usuarios no podían ser evaluados. En ese sentido Mora II (App 3) contó con 1266 usuarios válidos, Matenautas (App 2) con 608 usuarios y OxTravel (App 1) con 386 usuarios. Del total de usuarios que jugaron al menos el Nivel Uno de los niveles secuenciales, Matenautas (App 2) registró 276 usuarios, Mora II (App 3) 42 y OxTravel (App 1) 7.

Análisis de los datos

Para analizar los datos, se utilizó un modelo del tipo ANCOVA para remover el efecto de la línea de base. La elección de este modelo sigue los lineamientos de McKenzie [4] y de Schneider et al. [5].

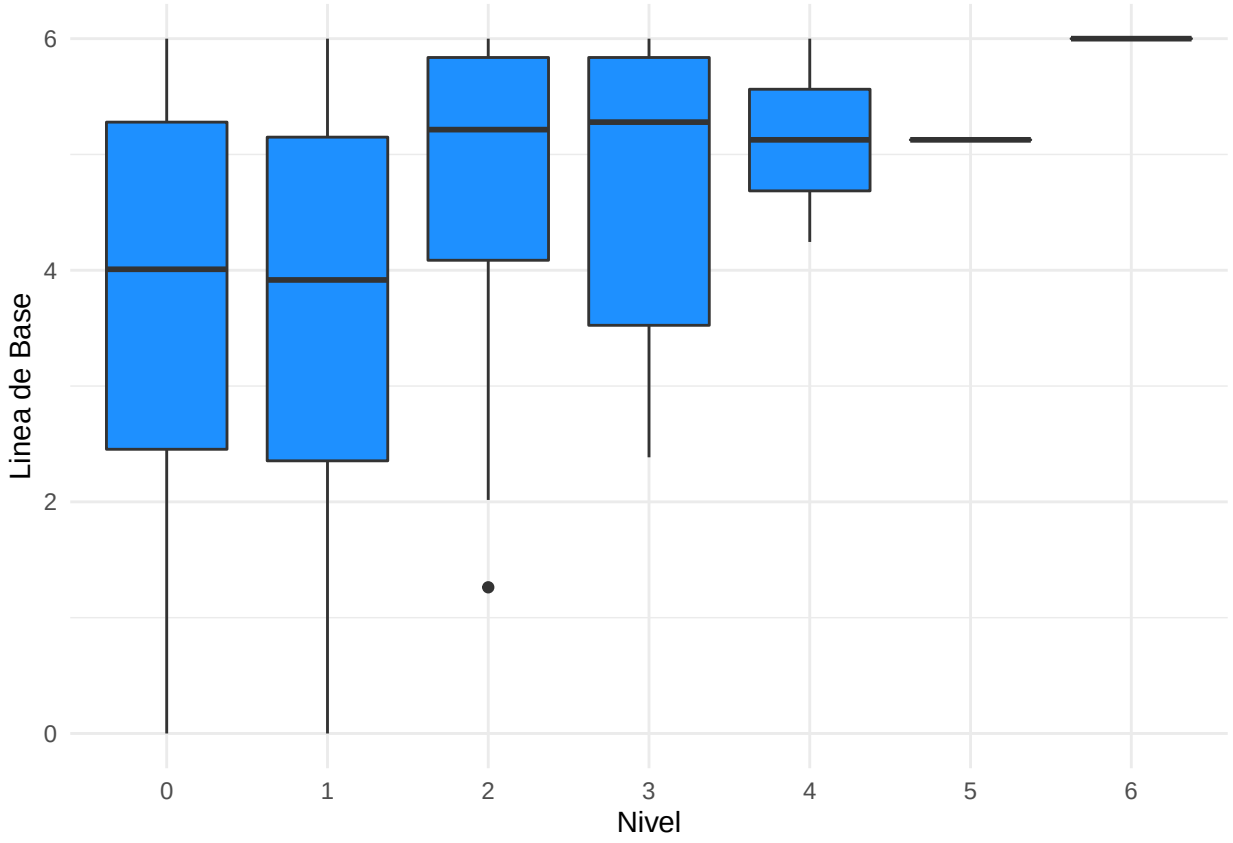
Tal como se mencionó en la introducción, este análisis se realizó de forma anonimizada a partir de IDs que funcionaron como seudónimos de los nombres de las apps para que el equipo estadístico no tuviera conocimiento de los nombres de las aplicaciones que estaban evaluando. Así es que en esta sección App 1 refiere a OxTravel, App 2 a Matenautas y App 3 a Mora II.

Análisis Exploratorio En miras al análisis propuesto, se analizó el comportamiento de la variable de base respecto a cada una de las aplicaciones:



Como puede observarse, la App 3 posee una desviación de posición respecto de las otras aplicaciones. En tanto que la asignación de la publicidad fue aleatoria, no se encuentra explicación estadística para dicha desviación. Cabe aclarar que la App 3 reportó un notable incremento en sus descargas y que muchas de ellas no provinieron de la pauta publicitaria realizada, si no de links externos. Por tanto para el análisis se ha de considerar que existen outliers en los datos y se procedió a utilizar un modelo robusto.

Respecto al efecto de la línea de base sobre los resultados observados, se observa en principio una posible relación monótona (lo esperable) entre dicha línea de base y la variable de respuesta. Tal relación será explorada con mayor detalle utilizando un modelo aditivo generalizado.



Modelo 1 El primer modelo analiza mediante un Modelo Lineal Generalizado (glm) para una variable de respuesta quasi-binomial con 6 categorías (los niveles). Para los tratamientos, se utilizaron contrastes de suma y la línea de base fue centrada y escalada.

Se observa en el Cuadro 1 que los coeficientes de los tratamientos son significativos, siendo la App 2 la que mayor impacto tiene (recuérdese que se utilizó contrastes de suma). A su vez se observa que la interacción entre la línea de base y los tratamientos no resulta significativa, por lo cual podríamos hablar de un efecto de cambio de posición, siendo común el efecto de la línea de base a todas las aplicaciones.

Con esto en mente, se re-estima el modelo y se analiza cuál fue el efecto real de cada aplicación.

Nuevamente, se observa en el Cuadro 2 que la App 2 posee un rendimiento estadísticamente significativo mayor que sus competidores.

Modelo 2 El segundo modelo consistió en un modelo robusto binomial [1] [3] [2] [6]. Los factores experimentales entraron al modelo bajo la parametrización de control de suma y

Cuadro 1: GLM Model

Family: *quasi-binomial* Link function: *logit*				
	Formula: $y \sim scale(LineaBase) * ID\_App$			
	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	-4.1735	0.1408	-29.647	< 2e-16***
scale(LineaBase)	0.5035	0.1471	3.422	0.000632***
ID App1	-0.8439	0.2405	-3.509	0.000458***
ID App2	1.7698	0.1482	11.945	< 2e-16***
scale(LineaBase):ID App1	-0.2620	0.2544	-1.030	0.303194
scale(LineaBase):ID App2	-0.1564	0.1548	-1.011	0.312257
R-sq.(adj) = 0.168		Deviance explained = 28 %		
GCV = 0.5558		Scale est. = 1.6948		
n = 2260				
Signif. codes:	$p_{0.001}$: *** ; $p_{0.01}$: ** ; $p_{0.05}$: * ; $p_{0.1}$: . ; $p_{>0.1}$:			

Cuadro 2: Comparación de los Tratamientos

Estimated Means

ID App	emmean	SE	df	lower.CL	upper.CL
1	-5.04	0.3745	2256	-5.77	-4.30
2	-2.41	0.0891	2256	-2.58	-2.23
3	-4.77	0.1741	2256	-5.11	-4.43

Estimated Contrast

ID App Contrast	estimate	SE	df	t.ratio	p.value
1 - 2	-2.632	0.385	2256	-6.840	<.0001
1 - 3	-0.267	0.411	2256	-0.650	0.7926
2 - 3	2.365	0.194	2256	12.162	<.0001

la línea de base fue modelada como predictora. El objetivo del modelo es poder revisar los resultados a la luz de una posible contaminación de los datos en la App 3

Cuadro 3: Robust GLM Model

Family: *binomial* Link function: *logit* Method: *glmrobMqle(tcc = 1.2)*				
	Formula: $y \sim scale(LineaBase) * ID\_App$			
	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	-4.6433	0.1746	-26.597	< 2e-16***
scale(LineaBase)	0.4692	0.1919	2.445	0.01448*
ID App1	-1.1320	0.3172	-3.568	0.00036***
ID App 2	2.2882	0.1799	12.719	< 2e-16***
scale(LineaBase):ID App1	-0.3380	0.3308	-1.022	0.30700
scale(LineaBase):ID App2	-0.1742	0.1957	-0.890	0.37337
n = 2260				
2172 weights are ∼= 1. The remaining 88 ones are summarized as				
Min.	1st Qu.	Median	3rd Qu.	Max.
0.03904	0.13290	0.31250	0.68350	0.96930
Signif. codes:	$p_{0.001}$: *** ; $p_{0.01}$: ** ; $p_{0.05}$: * ; $p_{0.1}$: . ; $p_{>0.1}$:			

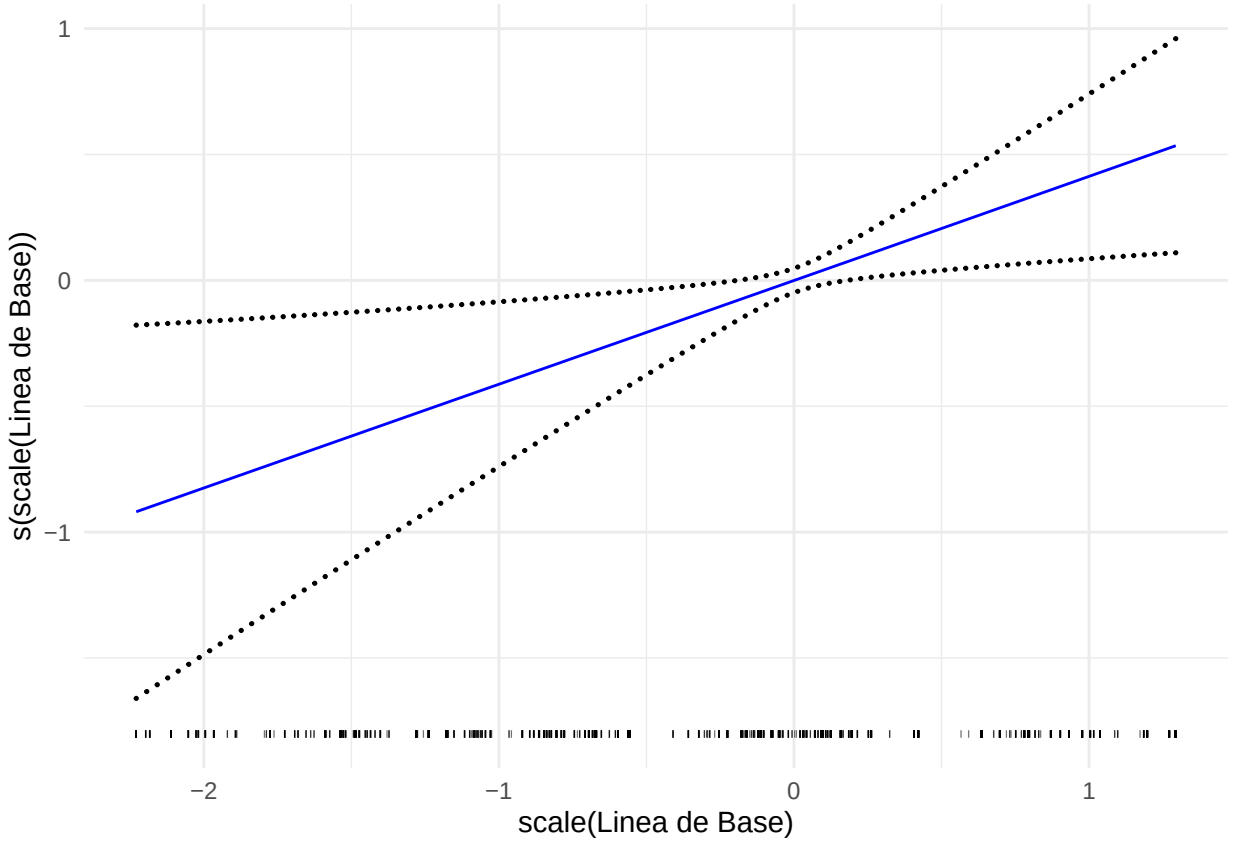
Nuevamente, se observa en el Cuadro 3 que la App 2 es aquella que estadísticamente posee el mejor rendimiento. A su vez, se confirma el hecho de que no son significativas las interacciones entre las aplicaciones y la línea de base. Finalmente, se puede afirmar que la contaminación observada en la App 3 no genera cambios en la interpretación de los resultados, tanto en los valores de los coeficientes, como en su signo y en su significancia.

Modelo 3 Finalmente, mediante un modelo aditivo generalizado [8] [7] de respuesta quasibinomial se analizó la pertinencia del supuesto de monotonicidad de la relación entre la línea de base y la variable de respuesta. Los factores experimentales entraron al modelo bajo la parametrización de control de suma y la línea de base fue modelada vía un spline para favorecer la flexibilidad del modelo.

Cuadro 4: GAM Model

Family: *quasi-binomial* Link function: *logit*				
	Formula: $y \sim s(scale(LineaBase)) * ID\_APP$			
Parametric coefficients:	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	-4.0706	0.1417	-28.720	< 2e-16 ***
ID App1	-0.9666	0.2578	-3.749	0.000182 ***
ID App2	1.6646	0.1501	11.087	< 2e-16 ***
Smooth Terms:	edf	Ref.df	F	p-value
s(scale(LineaBase))	1.008	9	2.727	6.85e-07 ***
R-sq.(adj) = 0.166		Deviance explained = 27.4 %		
GCV = 0.55944		Scale est. = 2.0898		
n = 2260				
Signif. codes:	$p_{0.001}$: *** ; $p_{0.01}$: ** ; $p_{0.05}$: * ; $p_{0.1}$: . ; $p_{>0.1}$:			

Como puede observarse en el Cuadro 4, la App 2 (Matenautas) es claramente superior que el resto de los tratamientos. En particular la línea de base en forma de spline resulta monótona y lineal para con la variable de respuesta.



Conclusiones

En base al análisis de los modelos presentados previamente, la evaluación de los resultados indica que la aplicación con mejor performance en el aprendizaje matemático e interés por parte de sus usuarios es Matenautas (App 2).

Teniendo en cuenta cada uno de los modelos y tomando como referencia el modelo robusto, podemos concluir que el avance / interés en los niveles bajo contrastes de suma y en escala logit por la aplicación Matenautas (App 2) es de 2.2882, mientras que el de Mora II (App 3) es de -1.1562 y el de OxTravel (App 1) es de -1.1320. El signo y valor de los coeficientes, sumado a su altísimo valor de significatividad, hace que inequívocamente pueda afirmarse que Matenautas posee un mejor rendimiento en cuanto al avance de los usuarios en el juego según los criterios de nuestra evaluación.

Entendiendo que el objetivo de la competencia era evaluar el desempeño matemático de los usuarios a partir de la experiencia de juego con cada aplicación, la evaluación estadística

señala que la aplicación ganadora es Matenautas (App 2).

Como complemento al análisis estadístico principal, los indicadores de performance de las apps señalan que la aplicación con mayor cantidad de descargas fue Mora II (App 3), la aplicación con mayor cantidad de usuarios jugando activamente en los niveles secuenciales fue Matenautas (App 2) y que la aplicación con mayor promedio de cantidad de sesiones por jugador fue OxTravel (App 1).

Referencias

[1] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

[2] Martin Maechler, Peter Rousseeuw, Christophe Croux, Valentin Todorov, Andreas Ruckstuhl, Matias Salibian-Barrera, Tobias Verbeke, Manuel Koller, Eduardo L. T. Conceicao, and Maria Anna di Palma. *robustbase: Basic Robust Statistics*, 2021. R package version 0.93-9.

[3] Ricardo A Maronna, R Douglas Martin, Victor J Yohai, and Matías Salibián-Barrera. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

[4] David McKenzie. Beyond baseline and follow-up: The case for more t in experiments. *Journal of development Economics*, 99(2):210–221, 2012.

[5] Bruce A Schneider, Meital Avivi-Reich, and Mindaugas Mozuraitis. A cautionary note on the use of the analysis of covariance (ancova) in classification designs with and without within-subject factors. *Frontiers in psychology*, 6:474, 2015.

[6] Valentin Todorov and Peter Filzmoser. An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47, 2009.

[7] S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36, 2011.

[8] S.N Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition, 2017.