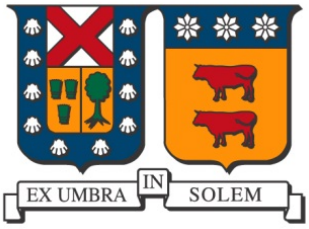
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
CIUDAD - CHILE



“DISEÑO DE UN MODELO EN MACHINE LEARNING PARA PREDECIR LAS FUNCIONES EJECUTIVAS Y VALIDARLAS ESTADÍSTICAMENTE”

MARIO ANDRÉS OLIVARES DEL RIÓ

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN INFORMÁTICA

Profesor Guía: Carlos Valle Vidal
Profesor Correferente: Ricardo Ñanculef

Junio - 2025

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título; ☐ Tesis de Postgrado;

Título del trabajo: DISEÑO DE UN MODELO EN MACHINE LEARNING PARA PREDECIR LAS FUNCIONES EJECUTIVAS Y VALIDARLAS ESTADÍSTICAMENTE.

Nombre del candidato(a): MARIO ANDRÉS OLIVARES DEL RIÓ

Carrera / Grado: Ingeneria Civil Informatica.

Campus: Casa Central Valparaiso **;** **Departamento**: Informatica

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Carlos Antonio Valle Vidal, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL

☒ El trabajo **NO contiene información que amerite confidencialidad** y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (embargo) por:

☐ 6 meses; ☐ 12 meses; ☐ 2 años; ☐ 3 años; ☐ 5 años; ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 01-08-2025 ; **Firma:** 

Estudiante o Candidato(a):

Fecha: 30-07-2025 ; **Firma:** 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

DEDICATORIA

Esto va dedicado a toda la gente que estuvo a mi alrededor durante el proceso, a mis padrinos, a mi familia y a mis amigos que estuvieron para mi en los momentos importantes.

AGRADECIMIENTOS

Quiero agradecer a todos los que han estado aportándome en esta etapa de la universidad, a mis padrinos que me han dado hogar y todo para poder estudiar en Viña, a mis padres que me han aguantado y enseñado todo lo necesario para la vida y a mi hermana que me ha ayudado en todo lo que ha podido.

Tambien agradecer a mis amigos y compañeros que han estado conmigo durante este largo camino, gracias a su sabiduría mientras estudiábamos y paciencia; Matías, Josias, Kevin, Pablo, Ricardo, Koba.

Destacar en estos años que si no fuera por mi grupo más cercano de trabajo no hubiera podido hacer nada; Aedo, Nico, Sebita y Manuel, sin ustedes que me bancaron en todo lo que hacia, que me apoyaron en irme de intercambio mientras estabamos cursando feria y todas las horas de estudios y trabajos que pasamos juntos muchas gracias.

Mencionar también a Kevin que nos fuimos juntos a España y fue toda una aventura, juntos contra el mundo.

Y agradecer a mi profesor guía Carlos Valle por darme esta oportunidad y a Pedro Godoy por ayudarme en todo lo posible.

RESUMEN

Resumen— En esta investigación se desarrolló un modelo de inteligencia artificial para prever el rendimiento en competencias matemáticas tempranas (CMLR, CMN y CMG) en niños de entre 5 y 9 años. Para ello, se recolectaron 528 registros que incluyen 22 variables relacionadas con funciones ejecutivas (como memoria de trabajo y flexibilidad cognitiva) y datos sociodemográficos.

Se entrenaron cinco modelos de regresión multisalida (Decision Tree, Random Forest, XGBoost, CatBoost y LightGBM), ajustando sus hiperparámetros con HalvingGridSearchCV y validándolos en 50 particiones aleatorias del conjunto de datos. El desempeño se midió con el Error Cuadrático Medio en los conjuntos de entrenamiento, validación y prueba. Tras detectar diferencias significativas entre modelos mediante la prueba de Friedman, se aplicaron comparaciones post-hoc (Nemenyi y pruebas de Wilcoxon con corrección de Bonferroni) para identificar cuál modelo resultaba mejor.

Los resultados muestran que CatBoost obtuvo el menor MSE promedio en los tres indicadores matemáticos, con poca variabilidad entre ejecuciones, y superó de forma significativa a los otros algoritmos. A partir de este modelo "ganador", se calculó la importancia de cada predictor usando Permutation Importance, encontrando que la Memoria de Trabajo Verbal y la edad fueron las variables más influyentes, seguidas por la Flexibilidad Cognitiva y la dependencia en tareas ejecutivas. Además, al repetir el análisis con un subconjunto reducido de seis variables, se comprobó que el rendimiento no se vio afectado de manera sustancial, lo cual respalda la viabilidad de un modelo más sencillo e interpretable.

La relevancia de estos hallazgos radica en ofrecer una visión cuantitativa sobre qué funciones ejecutivas y factores contextuales explican mejor las habilidades matemáticas en la primera infancia. Esto puede guiar el diseño de intervenciones educativas específicas (por ejemplo, ejercicios para fortalecer la memoria de trabajo) y aportar mayor transparencia acerca de la contribución relativa de cada variable, facilitando su aplicación en otros entornos escolares o poblaciones.

Palabras Clave— Inteligencia Artificial, Competencias Matematicas, Permutation Importance, Importancia Significativa.

ABSTRACT

Abstract— In this study, we built and validated a machine learning model to predict early math skills (CMLR, CMN, and CMG) in children aged 5 to 9. We gathered 528 records with 22 features, including executive function measures (such as verbal working memory and cognitive flexibility) and sociodemographic data.

We trained five multi-output regression models—Decision Tree, Random Forest, XGBoost, CatBoost, and LightGBM—tuning their hyperparameters with HalvingGridSearchCV and evaluating performance across 50 random splits of the data. Model accuracy was assessed using Mean Squared Error (MSE) on training, validation, and test sets. After detecting overall differences with Friedman's test, we applied post-hoc comparisons (Nemenyi and Wilcoxon tests with Bonferroni correction) to identify the best-performing algorithm.

Results revealed that CatBoost achieved the lowest average MSE for all three math outcomes, with consistently low variability across runs. Using this "winning" model, we computed feature importances via Permutation Importance and found that verbal working memory and age were the strongest predictors, followed by cognitive flexibility and executive function dependence. A follow-up analysis with a reduced set of six features showed that performance remained largely unchanged, supporting the feasibility of a simpler, more interpretable model.

The relevance of these findings lies in quantifying which executive functions and contextual factors most reliably predict early math development. This insight can inform targeted educational interventions (for example, exercises to strengthen working memory) and increase transparency about each predictor's contribution, making the approach adaptable to other school settings or populations.

Keywords— Artificial Intelligence, Mathematical Competencies, Permutation Importance, Significant Importance.

GLOSARIO

CMT: Competencias Matemáticas Tempranas.
DI: Departamento de Informática.
FE: Funciones Ejecutivas.
CMLR: Competencia Matemática Lógica Racional.
CMN: Competencia Matemática Numérica.
CMG: Competencia Matemática General.
FC: Flexibilidad Cognitiva.
IA: Inteligencia Artificial.
ML: Machine Learning.
MT: Memoria de Trabajo.
MTV: Memoria de Trabajo Verbal.
MTVE: Memoria de trabajo Viso-espacial.
RF: Random Forest.
UTFSM: Universidad Técnica Federico Santa María.
MSE: Error Cuadrático Medio.
DT: Decision Tree
RF: Random Forest
XGBT: XGBoost
CAT: CatBoost
LGBM: LightGBM

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

El aprendizaje de las matemáticas en la primera infancia constituye una base fundamental para el desarrollo académico y cognitivo de los niños. Diversos estudios han demostrado que las habilidades matemáticas tempranas —como el reconocimiento de números, la comprensión de cantidades y la resolución de problemas simples— están estrechamente vinculadas a logros posteriores en áreas científicas y tecnológicas. No obstante, no todos los estudiantes adquieren estas competencias de la misma manera ni en el mismo momento, lo que puede generar brechas de aprendizaje difíciles de revertir en niveles educativos superiores. Es por ello que resulta esencial identificar desde etapas tempranas a aquellos niños que podrían presentar dificultades, con el fin de diseñar intervenciones oportunas que favorezcan su progreso.

En este contexto, las funciones ejecutivas —conjunto de procesos cognitivos que incluyen la memoria de trabajo, la flexibilidad cognitiva, la inhibición y el control atencional— han emergido como uno de los factores predictivos más importantes del rendimiento académico. Varios investigadores internacionales han encontrado que, por ejemplo, la capacidad para retener y manipular información en la memoria de trabajo está relacionada con la resolución de problemas numéricos; de igual modo, la flexibilidad cognitiva facilita el pasaje entre distintos tipos de tareas matemáticas. Sin embargo, existen pocos estudios que examinen simultáneamente múltiples funciones ejecutivas y su relación directa con varios indicadores de competencia matemática en poblaciones escolares hispanohablantes.

Por otro lado, el avance de las técnicas de inteligencia artificial ha abierto nuevas posibilidades para el análisis de datos complejos y para la construcción de modelos predictivos que capturan relaciones no lineales entre variables. A diferencia de los métodos estadísticos clásicos, los algoritmos de aprendizaje automático (Machine Learning) pueden procesar grandes volúmenes de información y extraer patrones ocultos que de otro modo resultarían difíciles de detectar. En el ámbito de la educación, estos enfoques se han empleado para detectar tempranamente riesgos de abandono escolar, dificultades de lectura o identificación de estilos de aprendizaje, aunque su aplicación al estudio de las funciones ejecutivas y las competencias matemáticas tempranas aún es incipiente, especialmente en el contexto chileno.

El propósito central de esta tesis es aprovechar el potencial de los modelos de Machine Learning para predecir el desempeño en tres indicadores de competencias matemáticas (denominados CMLR, CMN y CMG) en niños de 5 a 9 años, a partir de medidas de sus funciones ejecutivas y datos sociodemográficos básicos. Para ello, se construyó una base de datos con 528 registros que incluyen pruebas estandarizadas de memoria de trabajo, flexibilidad cognitiva, velocidad de procesamiento y dependencia en tareas ejecutivas, junto con la edad y otras variables contextuales. A continuación, se entrenaron y compararon cinco algoritmos de regresión multisalida basados en árboles de decisión (Decision Tree, Random Forest, XGBoost, CatBoost y LightGBM). Cada modelo se ajustó mediante HalvingGridSearchCV y se

evaluó su desempeño en 50 particiones aleatorias usando el Error Cuadrático Medio como métrica principal. Finalmente, se aplicaron pruebas estadísticas no paramétricas (Friedman, Nemenyi y Wilcoxon con corrección de Bonferroni) para determinar si existían diferencias significativas entre los algoritmos y para seleccionar el modelo más robusto.

Más allá de identificar al "modelo ganador", esta investigación profundiza en la interpretabilidad de los resultados mediante la técnica de Permutation Importance, lo que permite cuantificar la contribución relativa de cada función ejecutiva y variable sociodemográfica en las predicciones. A partir de estos rankings, se propuso un modelo parsimonioso que utiliza únicamente las seis variables más influyentes, demostrando que es posible simplificar el protocolo sin perder precisión. De esta forma, se sientan las bases para desarrollar un instrumento de detección precoz con un número reducido de pruebas cognitivas, lo que facilitaría su implementación en contextos escolares con recursos limitados.

La estructura de la tesis se organiza de la siguiente manera. En el capítulo 2, se presenta el marco teórico y conceptual, revisando los antecedentes sobre funciones ejecutivas, competencias matemáticas tempranas y los fundamentos de los algoritmos de Machine Learning utilizados (árboles de decisión, ensambles y técnicas de validación). El capítulo 3 describe la propuesta de solución, detallando la recolección de datos, el preprocesamiento, la definición de las variables y el protocolo de entrenamiento y validación de los modelos. En el capítulo 4 se exponen los resultados obtenidos: configuración de hiperparámetros, desempeño de cada algoritmo, análisis estadístico comparativo, interpretabilidad y construcción del modelo parsimonioso. Finalmente, en el capítulo 5 se discuten las conclusiones generales, las limitaciones identificadas y las recomendaciones para investigaciones futuras y aplicaciones prácticas en el ámbito educativo.

En síntesis, esta investigación aspira a unir dos campos complementarios —la psicología cognitiva (a través del estudio de las funciones ejecutivas) y la ciencia de datos— con el objetivo de ofrecer un enfoque riguroso y práctico para la detección temprana de dificultades matemáticas. Al combinar mediciones cognitivas con algoritmos sofisticados de Machine Learning, se busca aportar tanto al conocimiento académico como a las prácticas pedagógicas, contribuyendo a que cada niño reciba el apoyo necesario desde las etapas más formativas de su trayectoria escolar.

CAPÍTULO 1
DEFINICIÓN DEL PROBLEMA

1.1. CONTEXTO

Las matemáticas son un pilar fundamental en el área de la educación, por ende, se empiezan a demostrar de una temprana edad escolar, y es en esta misma época donde algunos empiezan a notar dificultades con las matemáticas, esta edad puede ir variando dependiendo del autor que realice la investigación, [Alicia Risso, 2015] trabaja de 7-8 años, [María-Jesús Presentación, 2015] trabaja de 5-6 años. Es por eso que en este trabajo de investigación nos centraremos principalmente en las pruebas realizadas a los 4°básicos, que son los más cercanos a las edades anteriormente mencionadas.

En estas edades es cuando se presentan las primeras dificultades, estas dificultades tienden a prevalecer durante toda su etapa escolar, provocando que el aprendizaje sea más difícil para el estudiante [Orrontia, 2006].

En Chile se esta pasando por una etapa que en el Simce[1] de matemáticas de 4°básico de 2022 promedio se obtuvo 250 puntos, 10 puntos menos que el 2018 teniendo el mismo resultado del 2012, pero lo más alarmante es la cantidad de niños que obtuvieron en el nivel ïnsuficiente"que fue de 45 % y en 2°medio promedio fue 252 puntos, 12 puntos menos que el 2018 y similar al nivel de 2006 con un 54 % de nivel "insuficiente". Positivamente en 2023 se subió un poco el puntaje promedio en los 4°básico 9 puntos con un promedio de 259 puntos y en 2°medio subió 5 puntos a un promedio de 257 puntos.

Aunque a nivel internacional no todo es felicidad los resultados de la prueba TIMSS, realizado a los 4°básicos del 2019 [2], donde se puede apreciar por los resultados que el desempeño en el área matemática ha bajado obteniendo en el 2011 promedio 462 puntos, en el 2015 promedio 459 puntos y en el 2019 promedio 441 puntos, teniendo una baja considerable de 15 puntos alejándose cada vez más de la media esperada de 500 puntos.

Analizando los resultados de la prueba PISA [3] de Chile comparando los años 2018 y 2022, también ha tenido una baja, de 417 puntos a 412 puntos respectivamente, igualmente ambos resultados son menores al promedio OECD que en el 2018 era de 489 puntos y en el 2022 era de 472 puntos. Adicionalmente se ha demostrado en el mismo informe que uno de

[1] Sistema de Medición de la Calidad de la Educación
[2] https://timssandpirls.bc.edu/timss2023/
[3] https://www.oecd.org/pisa/

las razones de el bajo nivel es por las diferencias socio-económicas, donde hay una amplia diferencia entre el sector alto y el bajo, pero positivamente en la prueba del 2022, el bajo tuvo un aumento en 10 puntos en promedio.

Por otro lado existe evidencia que si se desarrolla o fortalece las matemáticas en este nivel es beneficioso para su progreso [Gamal Cerda, 2011]. Es por eso que para evitar estas situaciones se han desarrollados métodos para fortalecer las CMT. Estas se consideran fundamentales para el aprendizaje, siendo un estable predictor del desempeño en el área.

1.1.1. FUNCIONES EJECUTIVAS Y COMPETENCIAS MATEMATICAS TEMPRANAS

Las CMT son comprender, actitud hacia las tareas matemáticas y no matemáticas [Gamal Cerda Etchepare, 2014], es por esto que las CMT tienen un dominio de ocho elementos; comparación, clasificación, correspondencia, seriación, conteo verbal, conteo estructurado, conteo resultante y conocimiento general de los números [Gamal Cerda, 2011]. Varios autores concluyen que las CMT se desarrollan a temprana edad, teniendo una mejor evolución en las futuras etapas escolares. Por eso es que varios autores usan estas habilidades como predictores en como se desempeñara el estudiante con respecto a las matemáticas [Francisca Bernal-Ruiz, 2022]. Es por eso que se desarrolla estudios sobre variables precursoras de las CMT, en especial las funciones ejecutivas.

No han una definición exacta de que son las FE, pero se refieren a ellas como conjunto de procesos cognitivos de arriba-abajo que están implicados en la regulación de la acción, y por ello estos procesos son clave en educación infantil , existen tres tipos de dominio cognitivo: memoria de trabajo, la inhibición y la flexibilidad cognitiva [Alicia Risso, 2015].

La función de la inhibición es regular el comportamiento, evitando respuestas impulsivas o distracciones para poder planificar y ofrecer una respuesta adecuada a la situación en cuestión, la inhibición tiene dos tipos, una que es la conductual relacionada con el autocontrol y la cognitiva relacionada con la atención selectiva. Por otro lado, la FC implica la capacidad de cambiar de perspectiva frente a un problema y adaptarse a nuevas exigencias, normas o prioridades. Además, la MT permite retener y manejar información en la mente mientras se lleva a cabo una tarea específica. En resumen, la memoria de trabajo engloba los procesos mentales que posibilitan registrar, almacenar y manipular temporalmente la información necesaria para llevar a cabo tareas cognitivas complejas, como el aprendizaje, la lectura o las habilidades matemáticas, esta puede ser tanto verbal como visoespacial [Francisca Bernal-Ruiz, 2024].

De estos tres dominios ejecutivos se derivan la planificación y la resolución de problemas, estas permitiendo que la persona formule planes de acción, los pueda llevar acabo y evalué

su eficacia.

1.2. PROBLEMA DETECTADO

A través de los años se ha investigado como mejorar la educación en el área matemática, obteniendo hoy en día que lo mejor es en la etapa infantil, demostrado por estudios como el de [Gamal Cerda, 2011], el estudio constata que existen diferencias significativas en el nivel de CMT, en grupos que se le aplican programas para reforzarlas. Aun con todos este tipo de pruebas no se ha desarrollado a gran escala, junto a la disparidad de conocimiento matemático por nivel socio-económico ha preocupado a las autoridades [de Educación, 2023], que intentan fortalecer esta área. Es por eso que se desarrollara un modelo en ML para evaluar los FE, y validar estadísticamente el resultado obtenido.

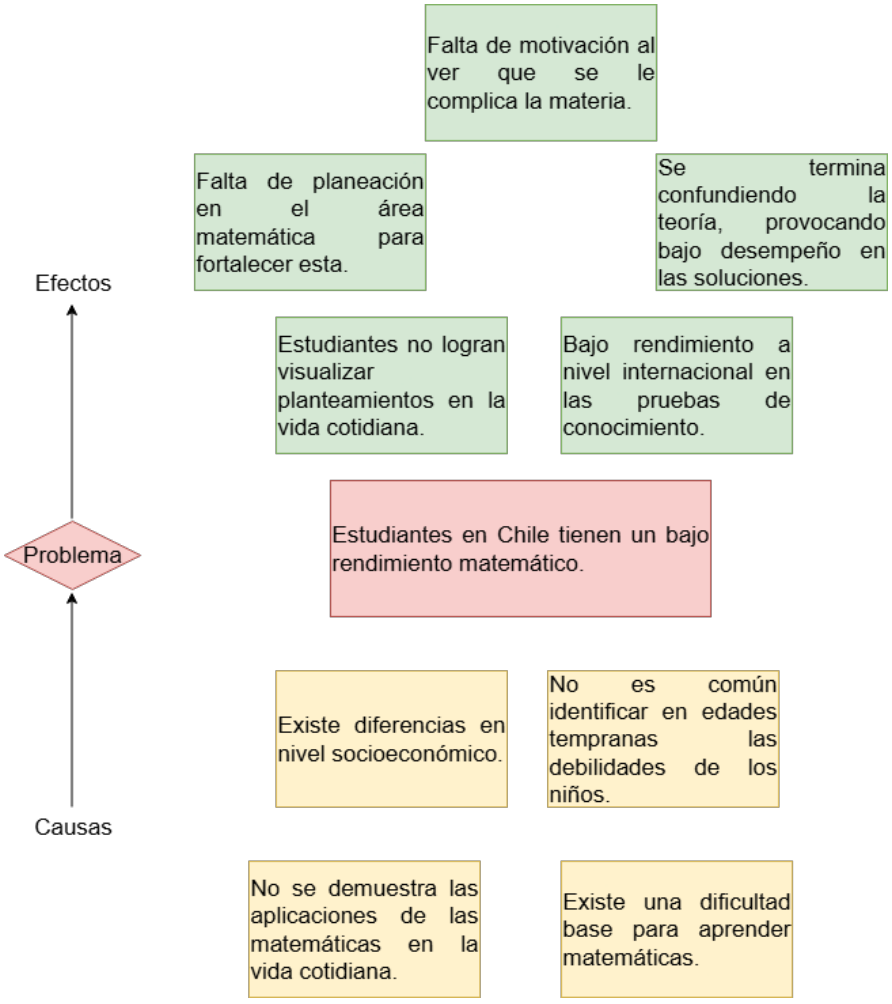


Figura 1: Árbol del Problema.
Fuente: Elaboración propia.

1.3. PROCEDIMIENTO ACTUAL

En la investigación realizada por Francisca Bernal-Ruiz y Gamal Cerda [Francisca Bernal-Ruiz, 2024], pretenden trabajar si hay un modelo en el cual existe interacción reciproca en los algunos componentes de la FE, trabajaron con memoria de trabajo verbal, inhibición conductual, inhibición cognitiva, flexibilidad cognitiva y planificación, para explicar la variabilidad en las CMT, reflejado en la figura 2, en un grupo de niños de 4-6 años y además si alguna de las componentes de la FE tenia mayor relevancia respecto al resto.

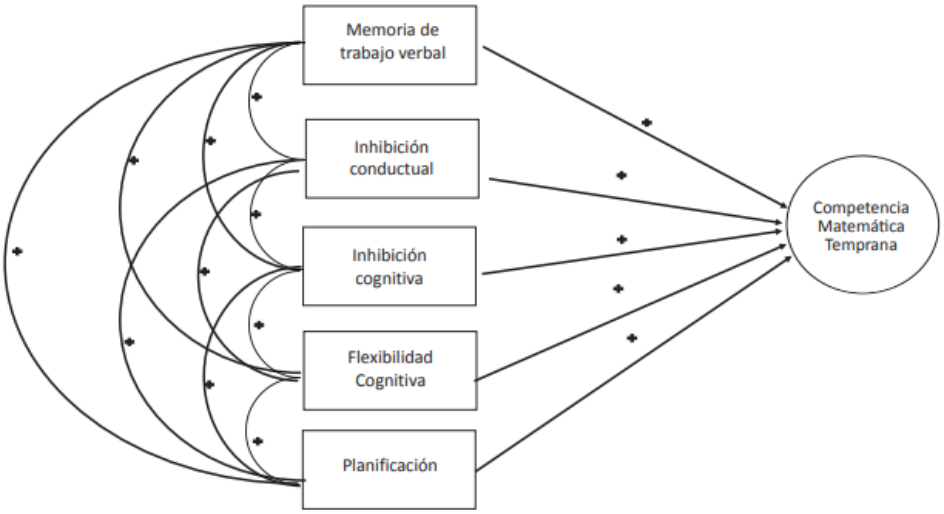


Figura 2: Modelo de Ecuaciones Estructurales hipotetizado respecto de la Competencia Matemática Temprana.
Fuente: Francisca Bernal-Ruiz y Gamal Cerda.

Las autoras en base a este modelo plantearon tres hipótesis:

- H1:
  Existe una covariación entre los componentes FE correspondientes, existiendo una independencia entre ellas que explican la variabilidad de las puntuaciones de la CMT.
- H2:
  La memoria de trabajo verbal es el predictor con mayor relevancia de la variabilidad entre las componentes de FE.
- H3:
  Los componentes se relacionan recíprocamente y significativamente con el desempeño de la CMT.

Estudio realizado sobre 130 estudiantes de segundo ciclo infantil de diversos centros educativos de la región de Valparaiso.

Para evaluar las FE se uso una tarea para cada una respectivamente que cuenta con propiedades psicométricas. Para la memoria de trabajo verbal, se utilizo la ïnversión de números", para la inhibición conductual se uso la prueba "Bzz! inhibición", en el caso de la inhibición cognitiva la prueba "Sol-Luna", fue usada "Dimensional Change Card Sort"para flexibilidad cognitiva y para la planificación el test de "Laberinto de Porteus". Para la CMT se uso el Early Numeracy Test, con el objetivo de evaluar y predecir su conocimiento sobre CMT.

Para analizar los datos se utilizo coeficiente de relación de Pearson para la correlación de las variables, y para representar la relación entre variables independientes y las CMT opto por un modelo de ecuaciones estructurales. Para analizar la figura 2, se utilizo máxima verosimilitud robusta, con esto se formo una nueva figura 3 con los valores de correlación correspondientes.

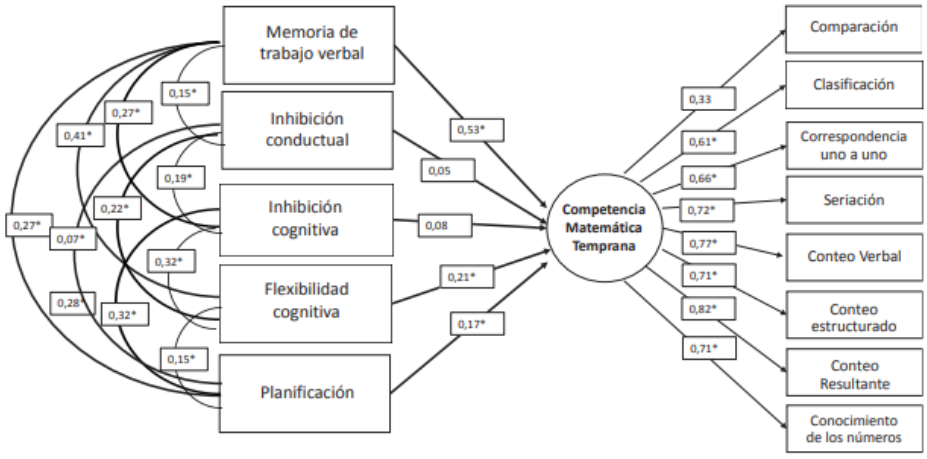


Figura 3: Modelo de Ecuaciones Estructurales de las Funciones Ejecutivas respecto de los componentes de la Competencia Matemática Temprana .
Fuente: Francisca Bernal-Ruiz y Gamal Cerda.

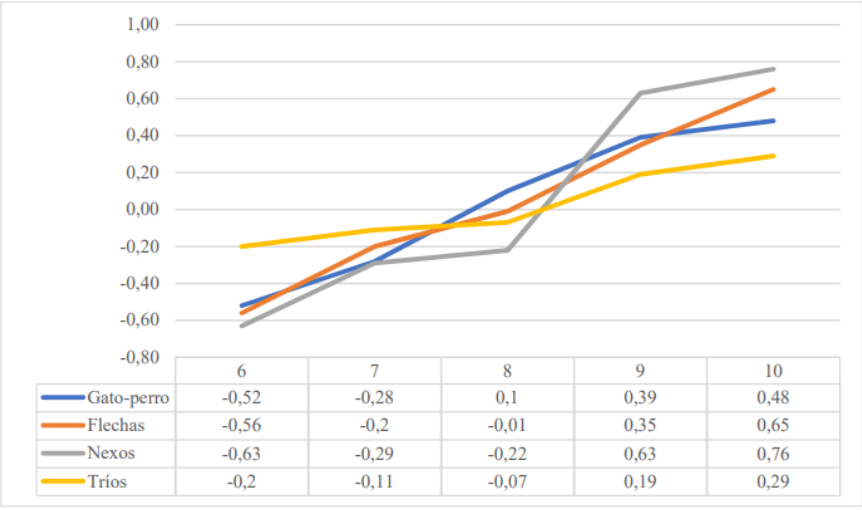
Por los resultados obtenidos en la figura 3, se analizo que se cumple parcialmente la hipótesis 1, debido a que los valores de la inhibición conductual y cognitiva son bajos comparados al resto. De la misma manera la hipótesis 2, se aprecia que se cumple, ya que la Memoria de trabajo verbal tiene el valor más alto de las FE. Y en base a la matriz de correlaciones bivariadas resultante [Francisca Bernal-Ruiz, 2024] se observa que hay una relación significativa y positiva entre los componentes de la FE. En este estudio a diferencia del de Francisca y Gamal, incluyen la Memoria de trabajo visoespacial.

Los estudios realizados Ricardo Rosas, Victoria Espinoza y Marion Garolera

[Ricardo Rosas, 2020], realizaron el experimento en Tablet en niños de 6 y 10 años con una pequeña evidencia internacional con un total de 1516 niños, con los paises Argentina, Noruega, Australia, Inglaterra y Chile. El experimento se basa en la batería de ejercicios Yellow Red, enfocada en 4 test en la evaluación general de las FE y también en en la evaluación especifica de los distintos componentes.

La primera prueba es del Gato-Perro, que su enfoque es en la evaluación general de las FE. La segunda prueba es Flechas, enfocada en la inhibición cognitiva y la atención. La siguiente prueba se llama Nexos que evalúa la memoria de trabajo, y el último test es Tríos para la flexibilidad cognitiva.

Para el análisis de los resultados se evidencio la confiabilidad mediante el alpha de Cronbach, donde todos se obtuvieron valores de alto nivel de consistencia interna, y también se evidencio la validez mediante cuatro métodos: Evidencia de validez convergente y discriminante, evidencia de validez de incremento de las FE con la edad, evidencia de validez de criterio y evidencia de validez factorial. La primera validez dio resultados congruentes con lo esperado sabiendo los FE que se evaluaba en cada prueba, esta se midió en base a un tabla con cada FE y la relación con la prueba. En la validez por edad, se puede observar por la figura 4, que cada prueba tiene un mayor impacto dependiendo de la edad de los niños, pero todas van progresando a pasar de los años. Para el tercer criterio se realizo el diagnostico en grupos diagnosticados con trastorno por déficit atencional con hiperactividad (TDAH), dislexia, discalculia, trastorno del espectro autista (TEA) y discapacidad intelectual (DI). En esta caso se aprecia como cada una de las pruebas afecta de mayor manera a cada grupo dependiendo del FE correspondiente. Y por último se obtuvo que las correlaciones entre las pruebas todas las FE son significativas, destacando inhibición y memoria de trabajo.



	6	7	8	9	10
Gato-perro	-0,52	-0,28	0,1	0,39	0,48
Flechas	-0,56	-0,2	-0,01	0,35	0,65
Nexos	-0,63	-0,29	-0,22	0,63	0,76
Tríos	-0,2	-0,11	-0,07	0,19	0,29

Figura 4: Progresión del desempeño en las distintas pruebas de YR según la edad. .
Fuente: Ricardo Rosas, Victoria Espinoza y Marion Garolera.

En base a el resultado de todos los criterios resultan más que aceptables, por lo que la herramienta es más que eficiente y puede seguir siendo usada. Aun así se debe seguir experimentando por que encontraban que la muestra no era muy grande.

Investigación realizadas en España a mano de María-Jesús Presentación, Rebeca Siegenthaler, Vicente Pinto, Jessica Mercader y Ana Miranda [María-Jesús Presentación, 2015], también se enfocaron en la relación de las FE y las CMT mediante pruebas neuropsicológicas y ecológicas en niños de 5-6 años. Esta incluía pruebas de inhibición, memoria de trabajo y habilidades matemáticas básicas. La valoración ecológica se realizo a través de los maestros y profesores se realizo mediante BRIEF [4].
Las pruebas neuropsicológicas fueron las encargadas de medir las FE. En la inhibición, se uso la prueba Sol-Luna para la inhibición visual y para la inhibición con estímulos auditivos el test de Golpeteo. Para la MTVE se administro las tareas Odd-One-Out y el test de Memoria de Laberintos, en el caso de MTV igualmente se aplicaron dos pruebas, la prueba de los Digitos Inversos y la tarea de Conteo. La evaluación ecológica que se uso la prueba BRIEF que mide la FE de los niños, pero la tienen que responder los padres y maestros por lo que conducta observada, esta agrupada en 8 escalas, inhibición, cambio, control emocional, iniciativa, MT, planificación/organización, organización de materiales y monitoreo. Y para evaluar las CMT se aplicarons las subpruebas TEDI-MATH, que evalua pruebas de conteo, numerar, conocimiento de los sistemas numéricos arábigo y oral, operaciones lógicas, operaciones aritméticas con apoyo de imágenes, con enunciado aritmético y con enunciado verbal, y estimación del tamaño.
Una vez aplicada las pruebas correspondientes se pudo analizar las relaciones entre las habilidades matemáticas y FE evaluado mediante las pruebas neuropsicológicas y ecológicas, destacando la inhibición con la MTV. También pudieron apreciar mediante análisis de regresión, donde se ven resultados diferenciales, con un mayor poder predicativo las tareas clínicas que los cuestionarios.

[4] Behavioral Rating Inventory of Executive Function

1.4. ACTORES RELACIONADOS

Los actores involucrados en el desarrollo de la memoria:

- **Alumnos:**
  Son los usuarios que están cursando la educación infantil rinden las pruebas correspondientes para analizar sus habilidades matemáticas.
- **Investigadores:**
  Correspondientes al área de inteligencia artificual e investigadores del área educacional, donde uno buscara aplicar los datos obtenidos para obtener de manera precisa predictores y corroborarlos estadísticamente, y el segundo en interpretar los datos obtenidos.

CAPÍTULO 2
MARCO CONCEPTUAL

2.1. APRENDIZAJE AUTOMÁTICO

Es una rama de la IA y también es conocido como Machine Learning, que es capaz de resolver problemas conociendo la situación problemática y reacciona usando la estrategia aprendida [Antonio Moreno, 1998]. Esta trabaja con modelos matemáticos que se alimentan de datos para que el algoritmo sea capaz de resolver estos problemas. A día de hoy se pueden encontrar 4 tipos de aprendizaje:

- **Aprendizaje Supervisado:**
  Este tipo de ML se alimenta de datos de entrada y salida en algoritmos de aprendizaje automático, con procesamiento entre cada par de entrada/salida que permite al algoritmo cambiar el modelo para crear salidas lo más alineadas posible con el resultado deseado. El resultado obtenido son datos etiquetados.
- **Aprendizaje No supervisado:**
  A diferencia del supervisado la maquina es la busca patrones menos obvios en los datos. El no supervisado es muy útil cuando se quiere encontrar patrones y utilizar los datos para tomar decisiones.
- **Aprendizaje Semisupervisado:**
  Es una practica que esta en medio del aprendizaje supervisado y no supervisado, es decir trabaja con datos etiquetados que son los mínimos y con datos sin etiquetas, de esta forma los datos etiquetados ayudar a formar patrones.
- **Aprendizaje por Refuerzo:**
  El algoritmo esta en un entorno donde tiene que tomar decisiones, estas decisiones le devuelve una recompensa positiva o negativa. Esta forma de aprendizaje se busca maximizar las recompensas obtenidas.

También es necesario conocer ciertos conceptos fundamentales para poder entender de que se tratan los modelos de IA y trabajo relacionado en esta tesis. Estos conceptos son:

- **Conjunto de entrenamiento:** Es el subconjunto de datos utilizado para ajustar el modelo. Aquí, el modelo aprende patrones y ajusta sus parámetros en función de las relaciones entre las características (features) y las etiquetas (labels).
- **Conjunto de prueba:** Es un subconjunto separado que se usa únicamente para evaluar el modelo después del entrenamiento. Sirve para medir el rendimiento del modelo en datos que no ha visto antes, lo que ayuda a estimar su capacidad de generalización.

- **Conjunto de Validación:** Se usa para ajustar hiperparámetros del modelo (por ejemplo, la cantidad de árboles en un Random Forest o la tasa de aprendizaje en un modelo de redes neuronales). Sirve para detectar sobreajuste antes de probar el modelo en el conjunto de prueba. Una práctica común es dividir los datos en 70 % entrenamiento, 15 % validación y 15 % prueba.
- **Generalización:** Un modelo de IA generaliza bien cuando logra hacer predicciones precisas sobre datos que no ha visto antes. Un modelo bien generalizado no solo memoriza los datos de entrenamiento, sino que aprende patrones útiles que le permiten funcionar en escenarios nuevos.
- **Sobreajuste (Overfitting):** Ocurre cuando el modelo aprende demasiado bien los datos de entrenamiento, incluyendo ruido o detalles irrelevantes. Como resultado, el modelo tiene un desempeño excelente en los datos de entrenamiento, pero falla en los datos de prueba porque no puede generalizar bien.
- **Subajuste (Underfitting):**Ocurre cuando el modelo es demasiado simple y no logra capturar patrones importantes en los datos. Como resultado, tiene un desempeño pobre tanto en los datos de entrenamiento como en los de prueba.
- **Importancia de Características:** En modelos supervisados, no todas las características (variables) tienen el mismo impacto en la predicción. La importancia de características mide cuánto contribuye cada variable a la toma de decisiones del modelo.

2.1.1. Árbol de Decisión

El árbol de decisión es un algoritmo de aprendizaje supervisado que se utiliza tanto en ta-reas de clasificación como de regresión. Su funcionamiento se basa en la construcción de un modelo en forma de árbol, en el que cada nodo interno representa una prueba sobre una característica, cada rama corresponde al resultado de dicha prueba, y cada hoja indica la predicción o clase asignada al conjunto de datos que llega a ese nodo.

La construcción del árbol se inicia en la raíz, donde se escoge la característica que mejor separa los datos según criterios como la ganancia de información, el índice Gini o la reducción de la varianza [Quinlan, 1986]. Tras esta elección, se generan de manera recursiva sub-nodos a través de la partición del conjunto de datos, hasta cumplir ciertos criterios de parada (por ejemplo, un número mínimo de ejemplos en un nodo o una ganancia informativa insignificante). Esta estrategia jerárquica permite que el árbol de decisión no solo sea intuitivo y fácil de interpretar, sino que también capture relaciones complejas y no lineales entre las variables del problema.

Entre las principales ventajas de los árboles de decisión se destacan:

- **Interpretabilidad:** La estructura en forma de árbol facilita la visualización y comprensión del proceso de toma de decisiones.
- **Flexibilidad en el manejo de datos:** Pueden trabajar de manera natural con datos tanto numéricos como categóricos.
- **Capacidad para modelar relaciones no lineales:** No requieren que las relaciones entre las variables sean lineales, adaptándose a diferentes estructuras subyacentes en los datos.

Sin embargo, es importante considerar que los árboles de decisión pueden ser sensibles al sobreajuste (overfitting) si se permite su crecimiento sin restricciones. Por ello, se suelen aplicar técnicas de poda y se integran en algoritmos de ensamble para mejorar la robustez del modelo.

Es precisamente esta base de árboles de decisión la que se utiliza en diversos modelos de ensamble. Por ejemplo:

- **Random Forest:** Combina múltiples árboles de decisión generados a partir de subconjuntos aleatorios del conjunto de datos para reducir la varianza y mejorar la generalización.
- **CatBoost:** Utiliza árboles de decisión optimizados junto con técnicas específicas para el manejo de variables categóricas y evitar la fuga de información durante el entrenamiento.
- **LightGBM:** Emplea estrategias de crecimiento de árbol *leaf-wise* (por hoja) que permiten una construcción eficiente y escalable de los árboles.
- **XGBoost:** Implementa un esquema de boosting secuencial en el que cada árbol se construye para corregir los errores del árbol anterior, integrando términos de regularización para mejorar la predicción y prevenir el sobreajuste.

De esta manera, los árboles de decisión no solo constituyen un modelo en sí mismos, sino que también sirven como componentes fundamentales en los algoritmos de ensamble que potencian el rendimiento predictivo y la robustez de los modelos de Random Forest, CatBoost, LightGBM y XGBoost.

2.1.2. Random Forest

Random Forest es un algoritmo de aprendizaje supervisado que se basa en el ensamble de múltiples árboles de decisión generados de forma aleatoria. Para tareas de clasificación, el algoritmo realiza una votación mayoritaria entre los árboles, mientras que para problemas

de regresión, se utiliza el promedio de las predicciones individuales. Esta metodología no solo mejora la precisión del modelo, sino que también reduce el riesgo de sobreajuste, al combinar la diversidad de las predicciones obtenidas a partir de diferentes subconjuntos del conjunto de datos.

Entre las características más destacadas de Random Forest se encuentran:

- **Robustez:** La agregación de múltiples modelos individuales mitiga el riesgo de sobreajuste y mejora la capacidad de generalización.
- **Manejo de alta dimensionalidad:** Es capaz de trabajar eficientemente con conjuntos de datos que incluyen un gran número de variables.
- **Estimación de la importancia de las variables:** Permite identificar la contribución relativa de cada característica en la predicción del modelo.

El funcionamiento del algoritmo se ejemplifica en la Figura 5, donde se ilustra el proceso de creación de árboles a partir de subconjuntos aleatorios de datos, y la posterior combinación de sus predicciones para obtener el resultado final.

Esta técnica ha demostrado ser eficaz en diversas aplicaciones, tales como la clasificación de imágenes, la detección de anomalías y la medicina predictiva [Breiman, 2001].

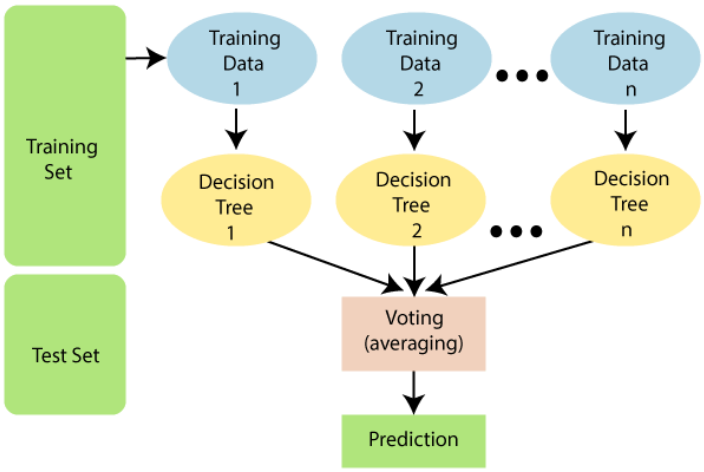


Figura 5: Ejemplo funcionamiento Random Forest
Fuente: Recuperado de https://www.javatpoint.com/machine-learning-random-forest-algorithm.

2.1.3. XGBoost

XGBoost (eXtreme Gradient Boosting) es un algoritmo de aprendizaje supervisado basado en el concepto de boosting, diseñado para ser altamente escalable y eficiente. Este método se fundamenta en la optimización de una función de pérdida a través de la construcción secuencial de árboles de decisión, donde cada nuevo árbol corrige los errores cometidos por el conjunto previo. XGBoost utiliza una aproximación mediante la expansión en serie de Taylor (usualmente hasta el segundo orden) para optimizar la función objetivo, lo que le permite incorporar medidas de regularización (tales como penalizaciones L1 y L2) que ayudan a evitar el sobreajuste.

Entre las características más destacadas de XGBoost se encuentran:

- **Eficiencia computacional:** El algoritmo está diseñado para aprovechar la computación paralela, lo que permite un entrenamiento rápido incluso en conjuntos de datos de gran volumen.
- **Regularización:** La inclusión de términos de regularización en la función de pérdida contribuye a mejorar la generalización del modelo y a reducir el sobreajuste.
- **Manejo de valores faltantes:** XGBoost incorpora estrategias internas para la imputación de datos faltantes, lo que facilita su aplicación en escenarios reales donde los datos incompletos son comunes.
- **Flexibilidad:** Es aplicable tanto a problemas de clasificación como de regresión y permite definir funciones de pérdida personalizadas según las necesidades específicas del problema.

La combinación de estas características ha convertido a XGBoost en una herramienta de gran relevancia en competiciones de ciencia de datos y aplicaciones prácticas, destacándose por su precisión, velocidad y capacidad de interpretación [Chen y Guestrin, 2016].

2.1.4. Catboost

CatBoost es un algoritmo de aprendizaje supervisado basado en el enfoque de boosting, desarrollado por Yandex, que se destaca por su capacidad para manejar eficientemente variables categóricas. A diferencia de otros métodos de boosting, CatBoost incorpora un mecanismo integrado para el tratamiento de las variables categóricas que evita la fuga de información durante el entrenamiento. Este proceso se basa en técnicas de codificación que utilizan permutaciones y estadísticas para transformar de manera automática las variables categóricas en representaciones numéricas, reduciendo así el riesgo de sesgo [Prokhorenkova *et al.*, 2018].

El entrenamiento de CatBoost se fundamenta en la construcción secuencial de modelos débiles, generalmente árboles de decisión. En cada iteración, el algoritmo ajusta un nuevo árbol que se especializa en corregir los errores residuales cometidos por los modelos anteriores, optimizando una función de pérdida diseñada para la tarea específica, ya sea de clasificación o de regresión. Este procedimiento iterativo permite mejorar la precisión general del modelo y mitigar el problema del sobreajuste.

Entre las principales ventajas de CatBoost se incluyen:

- **Manejo eficiente de variables categóricas:** Gracias a su técnica interna de codificación, permite explotar la información presente en las variables categóricas sin requerir una transformación manual previa.
- **Robustez y precisión:** La combinación de múltiples modelos débiles mediante el esquema de boosting contribuye a obtener un modelo final robusto y preciso en diversas aplicaciones.
- **Versatilidad:** Es aplicable tanto a problemas de clasificación como de regresión, lo que lo hace adecuado para una amplia variedad de escenarios y conjuntos de datos.

Estas características han permitido que CatBoost se convierta en una herramienta valiosa en tareas como el análisis de riesgo crediticio, la predicción de ventas y la personalización de recomendaciones, entre otras aplicaciones en las que el manejo de datos categóricos y la precisión en la predicción son fundamentales.

2.1.5. LightGBM

LightGBM (Light Gradient Boosting Machine) es un algoritmo de boosting basado en árboles de decisión desarrollado por Microsoft, que se destaca por su alta eficiencia y escalabilidad en el entrenamiento de modelos de aprendizaje supervisado. A diferencia de otros métodos de boosting, LightGBM implementa técnicas innovadoras como la utilización de histogramas para discretizar las características, lo que permite reducir el uso de memoria y acelerar el proceso de entrenamiento.

Entre las características principales de LightGBM se incluyen:

- **Muestreo basado en gradiente (GOSS):** Esta técnica prioriza las instancias con errores residuales altos durante el entrenamiento, lo que mejora la eficiencia en la estimación de la función de pérdida.
- **Agrupamiento de características excluyentes (EFB):** Permite agrupar variables que rara vez toman valores no nulos simultáneamente, reduciendo así la dimensionalidad del problema sin pérdida significativa de información.

- **Estrategia de crecimiento *leaf-wise*:** En lugar de crecer los árboles de forma equilibrada, LightGBM expande las hojas que más reducen la función de pérdida, lo que puede resultar en modelos más precisos, aunque requiere regularización para prevenir el sobreajuste.

Estas innovaciones hacen de LightGBM una herramienta especialmente útil en aplicaciones que manejan grandes volúmenes de datos y en escenarios donde se requiere un equilibrio entre velocidad y precisión en la predicción, tanto para tareas de clasificación como de regresión [Ke *et al.*, 2017].

En esta misma línea, los modelos *CatBoost*, *LightGBM* y *XGBoost* comparten una característica fundamental: todos están basados en la técnica de *boosting*, específicamente en el *Gradient Boosting*. A diferencia de métodos como *Random Forest*, que construyen árboles de decisión en paralelo a partir de subconjuntos aleatorios de datos (*bagging*), los algoritmos de *boosting* generan árboles de manera secuencial, donde cada nuevo árbol intenta corregir los errores cometidos por los anteriores. Esta construcción acumulativa permite mejorar progresivamente el rendimiento del modelo, enfocándose en las instancias más difíciles de predecir. La Figura 6 ilustra este proceso, mostrando cómo se integran múltiples modelos débiles para formar un modelo final más preciso mediante la minimización iterativa de una función de pérdida.

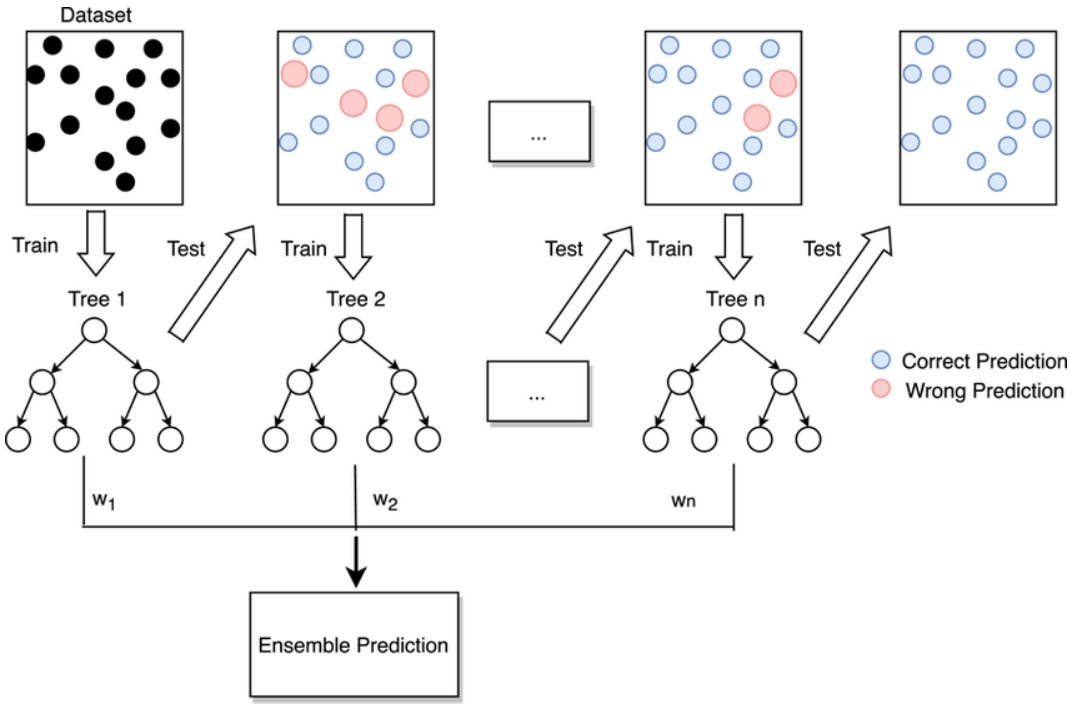


Figura 6: Ejemplo funcionamiento Boosting
Fuente: Friedman, Jerome H.

2.2. Funciones de Evaluación

En el marco de este estudio, se han seleccionado tres enfoques esenciales para evaluar el rendimiento de los modelos de inteligencia artificial propuestos: MSE, la importancia por permutación de las características y las pruebas estadísticas no paramétricas Friedman, Nemenyi y Wilcoxon.

El MSE se utiliza como métrica en tareas de regresión para cuantificar la diferencia entre los valores predichos por el modelo y los valores reales. Se define como el promedio de los errores al cuadrado, lo que implica que se penalizan de forma más severa las discrepancias mayores. Más adelante, se abordará en detalle su formulación matemática, las propiedades que lo caracterizan y cómo interpretar sus valores en el contexto de la comparación de modelos.

Para medir la relevancia de cada característica en las predicciones, se emplea la importancia por permutación (permutation importance), una técnica que evalúa la caída en el rendimiento del modelo al permutar aleatoriamente cada variable. Este enfoque permite cuantificar de manera directa la contribución de cada feature al desempeño global del modelo.

Por otro lado, para determinar si las diferencias observadas en el desempeño de los modelos o en la posición de las características en el ranking son estadísticamente significativas, se aplican varias pruebas no paramétricas. El test de Friedman se emplea para evaluar diferencias globales en el rendimiento cuando se comparan múltiples modelos sobre el mismo conjunto de datos. En caso de identificar diferencias significativas, se recurre al test de Nemenyi como procedimiento post-hoc, que permite determinar qué pares de modelos presentan discrepancias relevantes. Adicionalmente, se aplica la prueba de Wilcoxon para muestras pareadas, con el fin de detectar si existen diferencias ligeras pero significativas entre pares concretos de modelos o posiciones de características, utilizando la corrección de Bonferroni para controlar la tasa de error tipo I asociada a comparaciones múltiples. En secciones posteriores se discutirá en profundidad la metodología, los supuestos y la interpretación de los resultados obtenidos con estos test.

Esta combinación de análisis cuantitativo mediante el MSE, evaluación de importancia de características mediante permutation importance y validación estadística con los test de Friedman, Nemenyi y Wilcoxon (con corrección de Bonferroni), ofrece un marco robusto para evaluar y comparar tanto el desempeño de los modelos como la relevancia de sus variables. El desarrollo detallado de estas herramientas permitirá justificar la elección de los modelos y respaldar las conclusiones del presente trabajo.

2.2.1. Error Cuadrático Medio

El Error Cuadrático Medio es una de las métricas más utilizadas para evaluar el rendimiento de modelos de regresión [Hastie *et al.*, 2009]. Su objetivo principal es medir la discrepancia

entre los valores predichos por un modelo y los valores reales observados. Se define matemáticamente como:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

donde:

- n es el número total de muestras,
- y_i es el valor real de la i-ésima muestra,
- $\hat{y}_i$ es el valor predicho por el modelo para la i-ésima muestra.

El MSE mide la media de los errores elevados al cuadrado. Al elevar al cuadrado las diferencias entre los valores reales y los valores predichos, se eliminan los signos negativos, lo que garantiza que errores positivos y negativos no se cancelen entre sí. Además, esta elevación al cuadrado tiene una implicación importante:

- Errores grandes se penalizan más que errores pequeños. Esto significa que un modelo con predicciones altamente inexactas tendrá un MSE significativamente mayor que uno con errores pequeños y más distribuidos.

Un MSE más bajo indica un mejor ajuste del modelo a los datos, mientras que un MSE alto sugiere que el modelo tiene grandes desviaciones entre las predicciones y los valores reales.

Propiedades del MSE

1. **Siempre es positivo:** Como el error se eleva al cuadrado, nunca se obtienen valores negativos.
2. **Distingue modelos en función del tamaño de los errores:** Modelos con errores más grandes tendrán un MSE mayor.
3. **Penalización de errores grandes:** Al ser cuadrático, errores grandes tienen un impacto mucho mayor que errores pequeños.
4. **Sensibilidad a valores atípicos:** Si en el conjunto de datos hay valores extremos (outliers), estos pueden incrementar de manera significativa el MSE, ya que los errores elevados al cuadrado pueden dominar la métrica.

2.2.2. Permutation Importance

La importancia por permutación (*Permutation Importance*) es una técnica ampliamente utilizada en el aprendizaje automático para evaluar cuantitativamente la relevancia de cada característica (*feature*) dentro de un modelo predictivo. Esta técnica acorde a se [Breiman, 2001] basa en medir cómo varía el desempeño del modelo al alterar aleatoriamente los valores de una característica determinada.

El procedimiento general para calcular la importancia por permutación consta de los siguientes pasos:

1. Se entrena un modelo predictivo utilizando el conjunto original de datos.
2. Se evalúa y registra el rendimiento inicial del modelo mediante una métrica adecuada, por ejemplo, el error cuadrático medio (MSE).
3. Para cada característica:
    a) Se permutan (mezclan aleatoriamente) los valores de la característica en cuestión, rompiendo así su asociación con el objetivo.
    b) Se mide nuevamente el desempeño del modelo sobre el conjunto de datos alterado.
    c) La importancia de esta característica se cuantifica como el incremento en el error o la reducción en la precisión del modelo debido a dicha permutación.

De esta manera, las características cuya permutación provoque una mayor disminución del rendimiento del modelo se consideran más relevantes o importantes, puesto que tienen un mayor impacto predictivo. Por el contrario, aquellas características cuya permutación apenas afecte al rendimiento se consideran menos relevantes o incluso prescindibles.

Una ventaja significativa de la importancia por permutación es que resulta aplicable independientemente del tipo de modelo predictivo utilizado, siendo así útil en modelos interpretables como regresiones lineales o modelos más complejos como redes neuronales y ensambles.

En resumen, la importancia por permutación proporciona una herramienta robusta para identificar las variables más determinantes en las predicciones de los modelos, aportando claridad interpretativa y facilitando decisiones informadas sobre la selección de características en análisis predictivos.

2.2.3. Friedman Test

El test de Friedman es una prueba estadística no paramétrica utilizada para determinar si existen diferencias significativas entre múltiples tratamientos o modelos cuando estos son evaluados sobre un mismo conjunto de datos o sujetos. Constituye una alternativa no paramétrica al análisis de varianza (ANOVA) para medidas repetidas cuando no se cumplen los supuestos de normalidad o esfericidad [Friedman, 1937].

El procedimiento del test consiste en asignar rangos a los resultados de cada tratamiento o modelo dentro de cada bloque (por ejemplo, un mismo conjunto de datos o sujeto), para luego comparar estos rangos. La hipótesis nula (H_0) del test de Friedman establece que no existen diferencias significativas entre los tratamientos o modelos comparados, mientras que la hipótesis alternativa (H_1) sostiene que al menos uno de los modelos presenta diferencias significativas respecto a los demás.

La estadística del test de Friedman se define como:

$$\chi_F^2 = \frac{12n}{k(k+1)} \left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4} \right], \tag{1}$$

donde:

- χ_F^2: estadístico de Friedman, que sigue una distribución aproximadamente χ^2 con $k-1$ grados de libertad.
- n: número de bloques o muestras.
- k: número de tratamientos o modelos evaluados.
- R_j: suma de rangos del tratamiento o modelo j.

Si la prueba de Friedman muestra resultados estadísticamente significativos (es decir, se rechaza la hipótesis nula), es recomendable realizar pruebas post-hoc como la prueba de Nemenyi, con el propósito de identificar específicamente cuáles tratamientos o modelos difieren entre sí de manera significativa.

Debido a que el test de Friedman no asume normalidad ni homocedasticidad, resulta especialmente útil para evaluar diferencias en el desempeño de algoritmos de aprendizaje automático o métodos predictivos, permitiendo contrastar objetivamente múltiples configuraciones o técnicas diferentes sobre un mismo conjunto de datos.

2.2.4. Nemenyi Test

El test de Nemenyi es una prueba post-hoc no paramétrica que se utiliza después de que una prueba global como la de Friedman arroja resultados significativos, indicando que existen diferencias entre los tratamientos o modelos evaluados. Este test permite determinar específicamente qué pares de tratamientos o modelos presentan diferencias estadísticamente significativas entre sí [Nemenyi, 1963].

La diferencia crítica entre rangos para el test de Nemenyi se calcula mediante la siguiente fórmula:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6n}}, \tag{2}$$

donde:

- CD: Diferencia crítica entre los rangos para considerar la diferencia significativa.
- q_α: Valor crítico obtenido de la distribución de Studentizada (*Studentized range distribution*) basado en el nivel de significancia α.
- k: Número total de tratamientos o modelos comparados.
- n: Número de bloques o muestras sobre los que se evaluaron los tratamientos.

Si la diferencia observada entre los rangos promedio de dos tratamientos supera esta diferencia crítica (CD), se concluye que dichos tratamientos difieren significativamente entre sí.

2.2.5. Test de Wilcoxon

El test de Wilcoxon para muestras pareadas es una prueba estadística no paramétrica utilizada para determinar si existen diferencias significativas entre dos muestras relacionadas o emparejadas. Constituye una alternativa al test *t* de Student para muestras pareadas cuando los datos no cumplen el supuesto de normalidad [Wilcoxon, 1945].

La hipótesis nula (H_0) establece que no hay diferencias significativas entre los pares de observaciones, mientras que la hipótesis alternativa (H_1) afirma que sí existen dichas diferencias. La prueba considera tanto la dirección como la magnitud de las diferencias observadas en cada par.

El procedimiento implica:

1. Calcular las diferencias entre las observaciones pareadas.
2. Asignar rangos a las diferencias absolutas.
3. Calcular el estadístico de Wilcoxon (W), definido como la suma de los rangos con signo positivo o negativo.

Cuando el tamaño muestral es suficientemente grande ($n > 30$), la distribución del estadístico W puede aproximarse mediante una distribución normal, facilitando así el cálculo del valor p para decidir si rechazar o no la hipótesis nula.

2.2.6. Corrección de Bonferroni

La corrección de Bonferroni es una técnica utilizada para controlar el error tipo I[5] cuando se realizan múltiples comparaciones simultáneas. Al aplicar varias pruebas estadísticas, aumenta la probabilidad de obtener resultados significativos por puro azar, lo que incrementa el riesgo de conclusiones incorrectas. Según Dunn (1961), la corrección ajusta este riesgo dividiendo el nivel de significancia inicial α entre el número total de comparaciones realizadas m [Dunn, 1961].

La fórmula para ajustar el nivel de significancia con la corrección de Bonferroni es:

$$\alpha_{\text{ajustado}} = \frac{\alpha}{m}, \tag{3}$$

donde:

- α_{ajustado}: Nuevo nivel de significancia corregido.
- α: Nivel de significancia original (habitualmente $0{,}05$).
- m: Número de comparaciones realizadas.

Al aplicar esta corrección en pruebas como la de Wilcoxon, se reduce considerablemente la probabilidad de cometer errores tipo I, aumentando la robustez estadística del análisis al momento de evaluar diferencias significativas entre múltiples pares de comparaciones [Dunn, 1961].

[5] El error tipo I, también conocido como falso positivo, se refiere al error estadístico que se comete al rechazar incorrectamente una hipótesis nula cuando esta es verdadera. Es decir, concluir que existe un efecto o diferencia cuando realmente no existe.

2.3. Estado del Arte

La aplicación de técnicas de Inteligencia Artificial en el ámbito educativo ha ganado relevancia en los últimos años, principalmente para predecir desempeños académicos y facilitar decisiones pedagógicas. No obstante, en el contexto específico de predecir competencias matemáticas tempranas a partir de funciones ejecutivas, el uso de estos métodos aún es escaso. Por esta razón, el presente Estado del Arte revisa investigaciones donde lo esencial es el procedimiento metodológico utilizado, con el fin de fundamentar las decisiones técnicas y estadísticas tomadas en este estudio.

El ranking de características (*feature ranking*) en problemas de regresión multisalida (*multi-target regression*) es abordado por Petković et al. (2020), quienes proponen métodos específicos basados en técnicas de ensamble (e.g., Random Forest, Extra Trees[6] y Genie3[7]) y una variante del algoritmo ReliefF denominada MTR-Relief. La evaluación de estos métodos sobre múltiples conjuntos de datos mostró que los enfoques basados en ensambles proporcionan rankings robustos, precisos y altamente interpretables en contextos con múltiples salidas relacionadas, como aplicaciones ecológicas o químicas [Petković *et al.*, 2020].

Por su parte, Santana et al. (2021) estudian modelos supervisados de aprendizaje automático aplicados al contexto del diagnóstico temprano del COVID-19 en Brasil. Los autores comparan diversos algoritmos como Random Forest, SVM[8], KNN[9], Decision Tree, Gradient Boosting y XGBoost, utilizando técnicas estadísticas no paramétricas (test de Friedman seguido por el test de Nemenyi) para validar diferencias significativas entre modelos. Su investigación destaca la importancia de procedimientos estadísticos rigurosos en la selección objetiva de modelos y enfatiza la utilidad de estas técnicas en decisiones críticas del ámbito clínico y administrativo [Santana *et al.*, 2021].

Finalmente, Barzizza et al. (2023) proponen una metodología robusta para selección de modelos mediante pruebas de permutación multiaspecto, capaces de evaluar simultáneamente diferencias significativas tanto en la ubicación (media) como en la escala (varianza) de los errores de predicción. Su enfoque se basa en la técnica de combinación no paramétrica (NPC), aplicado mediante simulaciones y casos prácticos reales, demostrando eficacia en la selección de modelos estables y precisos. Este procedimiento es particularmente útil en si-

[6] Extra Trees (*Extremely Randomized Trees*) es un algoritmo de ensamble que crea múltiples árboles de decisión entrenados con divisiones aleatorias en lugar de divisiones óptimas, incrementando la diversidad del modelo y reduciendo la varianza [Geurts *et al.*, 2006].

[7] Genie3 es un método basado en árboles de decisión, utilizado principalmente para inferir redes regulatorias y rankings de características mediante la agregación de la importancia de variables obtenida en múltiples árboles generados aleatoriamente [Huynh-Thu *et al.*, 2010].

[8] Support Vector Machine (SVM) es un modelo supervisado de aprendizaje automático que busca encontrar un hiperplano óptimo para separar las clases con el mayor margen posible, aplicable tanto a tareas de clasificación como regresión [Cortes y Vapnik, 1995].

[9] K-Nearest Neighbors (KNN) es un método supervisado sencillo basado en la identificación de los K vecinos más cercanos a un dato nuevo, prediciendo la clase o valor según la mayoría o promedio de estos vecinos [Cover y Hart, 1967].

tuaciones donde la estabilidad predictiva tiene una importancia comparable a la exactitud de la predicción [Barzizza *et al.*, 2023].

2.4. OBJETIVOS

El objetivo general es diseñar e implementar un ML que permita predecir el desempeño en competencias matemáticas tempranas a partir de variables asociadas a las funciones ejecutivas, con el fin de identificar los factores más influyentes en el aprendizaje y contribuir a la detección temprana de posibles dificultades, apoyando así la toma de decisiones en el ámbito educativo.

Los objetivos específicos son:

1. Diseñar un modelo en ML, de arboles de decisión para que sea entrenado y clasifique de manera correcta.
2. Validar de manera estadística los resultados obtenidos para que tengan robustez.
3. Ser capaz de identificar nuevos patrones que se formen.

CAPÍTULO 3
PROPUESTA DE SOLUCIÓN

Para dar respuesta a los objetivos planteados, se ha diseñado un protocolo de ranking de variables para regresión multi-objetivo sobre un conjunto de datos de 528 niños, caracterizados por 22 atributos (edad, curso, tipo de centro, cohortes y medidas de funciones ejecutivas) y 3 competencias matemáticas tempranas como salidas simultáneas. Una vez completado el preprocesamiento (depuración de casos, tratamiento de valores faltantes y normalización de escalas), entrenaremos un modelo de ensamble capaz de estimar la importancia de cada variable en las tres tareas de predicción de forma conjunta. A partir de esas estimaciones, obtendremos distintos rankings de características que, a continuación, compararemos entre sí mediante pruebas no paramétricas y sus correspondientes contrastes posthoc, con el fin de determinar en qué medida coinciden y cuáles son los predictores más robustos del desempeño en matemáticas durante la etapa preescolar.

3.1. Aspectos Claves del Problema

3.1.1. Desafíos del Problema

El planteamiento del protocolo de ranking de variables para regresión multi-objetivo conlleva varios retos clave:

- **Limpieza y homogeneización de datos:** Control de valores faltantes, detección de outliers y normalización de las 22 variables (edad, curso, tipo de centro, cohortes y medidas ejecutivas) en los 528 registros, para evitar sesgos y garantizar la calidad de entrada.
- **Selección y ajuste del ensamble:** Elección de la arquitectura más estable ante correlaciones entre atributos, afinación de hiperparámetros mediante Halving Grid Searching y comprobación de la coherencia de los rankings de importancia frente a distintas semillas.
- **Rendimiento, generalización e interpretación:** Minimizar el MSE en conjuntos de validación cruzada y hold-out para asegurar que el modelo capture patrones reales. Una baja discrepancia entre MSE de entrenamiento y prueba indica buena generalización. Los rankings resultantes se interpretan localizando las variables con mayor importancia media, y su consistencia a través de pruebas no paramétricas y post-hoc confirma su papel robusto como predictores del desempeño matemático en preescolar.

3.1.2. Relevancia del Problema

La relevancia principal de este estudio radica en integrar y ampliar hallazgos recientes sobre cómo las funciones ejecutivas en especial la memoria de trabajo verbal, la flexibilidad cognitiva y la planificación-ejercen un papel decisivo en el desarrollo de las competencias matemáticas tempranas de los niños de preescolar [Francisca Bernal-Ruiz, 2024]. Al identificar cuáles de estos procesos cognitivos son los predictores más estables del rendimiento en tareas de comparación, conteo o resolución de problemas, podemos:

- **Detectar precozmente**: diseñar intervenciones específicas antes de que las brechas académicas se consoliden.
- **Informar a docentes y diseñadores de currículo**: incorporar ejercicios que potencien las funciones ejecutivas clave.
- **Optimizar recursos educativos**: centrar esfuerzos en estimular capacidades (p. ej., memoria de trabajo verbal) que maximicen el avance matemático.

Además, la utilización de técnicas de inteligencia artificial como el ranking de variables en regresión multi-objetivo y los modelos de ensamble basados en aprendizaje automático facilita la automatización y la escalabilidad de este análisis, permitiendo:

- **Transparencia interpretativa**: cuantificar la contribución relativa de cada predictor mediante métricas de importancia.
- **Adaptación dinámica**: actualizar el modelo con nuevos datos de evaluación y mejorar la precisión de las predicciones.
- **Escalabilidad**: aplicar el mismo protocolo a poblaciones mayores o distintos contextos educativos con un coste computacional moderado.

De este modo, el problema no solo es académico, sino que, gracias a la IA, puede traducirse en herramientas prácticas para mejorar la calidad y la equidad de la educación infantil.

3.1.3. Variables y Preprocesamiento

Se ha trabajado con un conjunto de datos proporcionado por el profesor a cargo del proyecto, el cual considera resultados de pruebas aplicadas a niños y niñas de entre 5 y 9 años. Estas pruebas evalúan diferentes dimensiones cognitivas, en particular funciones ejecutivas como la memoria de trabajo y la planificación, así como competencias matemáticas tempranas, organizadas en tres tipos de conocimiento. Estas categorías permiten analizar cómo se

manifiestan ciertas habilidades cognitivas durante una etapa clave del desarrollo infantil, y cómo se relacionan con el rendimiento matemático inicial.

Para este estudio, se ha trabajado únicamente con las columnas más relevantes del conjunto de datos, seleccionadas en base a su valor informativo y su relación teórica con las variables objetivo. Estas variables incluyen: `Edad (años)`, `Sexo`, `Curso`, `Tipo EE`, `Depen. EE`, `Cohorte`, `MTV WM`, `MTVE Torpo`, `INHCon Bzz`, `INHCog Stroop` y `FC DCCS` y las salidas son: `CMLR`, `CMN` y `CMG`.

El preprocesamiento de los datos consistió en la depuración de casos, el tratamiento de valores faltantes y la selección de variables.

3.1.4. Modelos

Para el presente estudio se seleccionaron cinco modelos de regresión: *Decision Tree Regressor*, *Random Forest Regressor*, *XGBoost Regressor*, *LightGBM Regressor* y *CatBoost Regressor*. La elección de estos algoritmos se fundamenta en su capacidad para resolver tareas de regresión multisalida, lo cual era necesario dado que el problema abordado involucra la predicción simultánea de tres competencias matemáticas distintas. Además, estos modelos permiten capturar relaciones no lineales entre las variables cognitivas y los resultados, lo que los hace adecuados para el tipo de datos empleados en este estudio.

3.1.5. Evaluación del Modelo

Para evaluar el rendimiento de los modelos predictivos utilizados en este estudio, se seleccionó el *Error Cuadrático Medio* (MSE) como métrica principal. El MSE permite cuantificar la diferencia entre los valores predichos por el modelo y los valores reales, penalizando en mayor medida aquellos errores de mayor magnitud. Dado que el problema abordado corresponde a una tarea de regresión multisalida, el MSE se calculó para cada una de las tres competencias matemáticas por separado y luego fue considerado como base comparativa entre modelos.

Con el fin de determinar si existen diferencias estadísticamente significativas entre los distintos modelos aplicados, se empleó el test no paramétrico de Friedman. Esta prueba permite contrastar múltiples modelos sobre un mismo conjunto de datos sin asumir normalidad, evaluando si existen diferencias globales en sus desempeños.

Posteriormente, en caso de que Friedman indique una diferencia significativa global, se procede con pruebas post hoc. En este estudio se utilizó el test de Nemenyi como comparación múltiple, para identificar específicamente entre qué pares de modelos se presentan diferencias relevantes, controlando así el error tipo I asociado a múltiples comparaciones.

Además, se aplicó la prueba de Wilcoxon para muestras pareadas, tanto para evaluar diferencias significativas entre pares de modelos de regresión como para analizar si el orden de las variables en los rankings de importancia difiere significativamente entre modelos. Esta prueba resulta especialmente útil para detectar diferencias más sutiles cuando se trabaja con comparaciones emparejadas, y se complementa con la corrección de Bonferroni para mantener la validez estadística de los resultados.

El uso conjunto de estas métricas y pruebas estadísticas permite una evaluación robusta del rendimiento de los modelos, así como un análisis detallado de la importancia relativa de las variables predictoras implicadas en el estudio.

3.2. Propuesta de Solución

Para el desarrollo de la tesis se usaran los modelos mencionados con el conjunto de datos. El flujo de trabajo que se realizara es el siguiente:

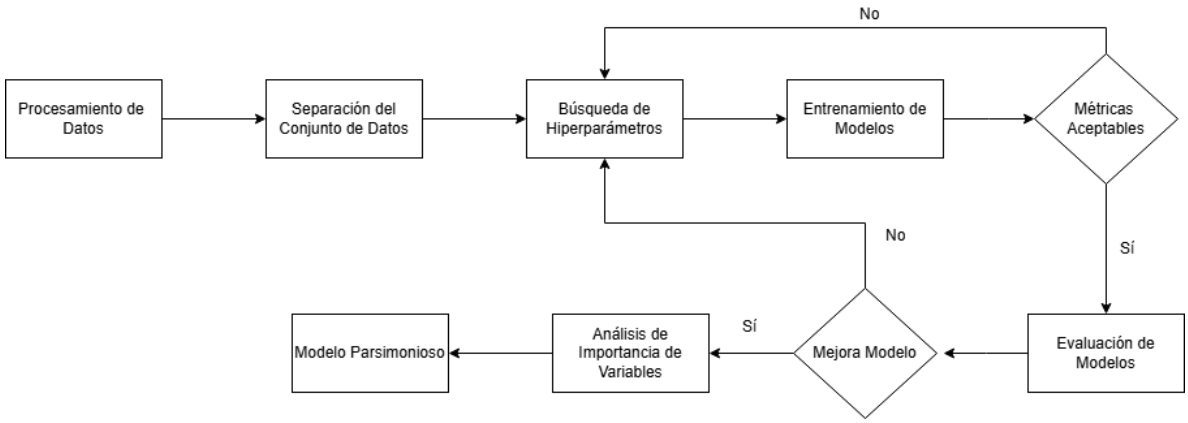


Figura 7: *Diagrama general del proceso*
Nota. Un diagrama para visualizar el flujo que se siguió. Fuente: Autoría propia.

1. **Preprocesamiento de datos:** Se seleccionaron únicamente las columnas relevantes para el análisis, descartando las demás. En los casos donde existían valores faltantes (`NaN`), estos fueron reemplazados por cero, manteniendo así la consistencia estructural del conjunto de datos.

2. **Separación del conjunto de datos:** El conjunto de datos fue dividido en tres subconjuntos: entrenamiento (*train*), validación (*validation*) y prueba (*test*). Este procedimiento se repitió 50 veces, cambiando aleatoriamente la distribución de los datos entre los subconjuntos, con el objetivo de evaluar la estabilidad de los resultados frente a distintas particiones.

3. **Búsqueda de hiperparámetros:** Para cada partición, se utilizaron los subconjuntos de entrenamiento y validación para encontrar la mejor configuración de hiperparámetros para cada modelo. (La descripción específica de los hiperparámetros modificados se detallará al momento de la implementación en Jupyter Notebook).

4. **Entrenamiento de modelos:** Una vez definidos los hiperparámetros óptimos, cada modelo fue entrenado utilizando los 50 conjuntos generados. En esta etapa se empleó exclusivamente el subconjunto de prueba, y se calculó el *Mean Squared Error* (MSE) para cada uno de los tres targets, con el fin de evaluar el rendimiento predictivo de los modelos.

5. **Evaluación de modelos:** Se compararon los resultados obtenidos por cada modelo en función de las 50 repeticiones. Para determinar si existían diferencias estadísticamente significativas entre ellos, se aplicó la prueba de Wilcoxon para muestras pareadas. En caso de ser necesario, se consideró el uso de la corrección de Bonferroni como prueba post-hoc. El modelo seleccionado fue aquel que, además de presentar el menor MSE, obtuvo respaldo estadístico significativo frente a sus competidores.

6. **Análisis de importancia de variables:** Una vez identificado el modelo con mejor desempeño, se aplicó la técnica de *Permutation Importance* sobre las 50 ejecuciones para obtener un ranking de importancia de las variables. Dado que el problema es multisalida, se obtuvo un ranking por cada uno de los tres targets. Se utilizó el test de Friedman para evaluar si existían diferencias significativas entre los rankings, y posteriormente, se aplicó el test de Wilcoxon (con corrección de Bonferroni) para comparar la relevancia relativa entre las cuatro variables mejor posicionadas en cada ranking.

7. **Modelo parsimonioso:** Como etapa final del proceso, se contempla la construcción de un modelo parsimonioso, es decir, un modelo que utilice la menor cantidad posible de variables sin comprometer significativamente su capacidad predictiva. Para ello, se seleccionarán las características más relevantes según el análisis de *Permutation Importance* y se entrenará nuevamente el modelo seleccionado, utilizando únicamente dichas variables. Esta etapa permitirá evaluar si es posible obtener una solución más simple, eficiente e interpretable, manteniendo un rendimiento adecuado en la predicción de las competencias matemáticas tempranas.

CAPÍTULO 4
VALIDACIÓN DE LA SOLUCIÓN

4.1. Especificación del Equipo

Los experimentos fueron ejecutados en un equipo portátil con procesador Intel(R) Core(TM) i5-10300H @ 2.50GHz, 16 GB de memoria RAM y una unidad de almacenamiento sólido (SSD) KINGSTON SNVS1000G.

4.2. Resultados de Hiperparámetros y Rendimiento

El presente capítulo tiene como objetivo validar empíricamente el rendimiento de los modelos de regresión implementados, a partir del conjunto de datos recopilado y el protocolo metodológico descrito previamente. Para ello, se llevó a cabo un proceso sistemático de ajuste de hiperparámetros, entrenamiento y evaluación de cinco modelos de aprendizaje automático multisalida: *Decision Tree*, *Random Forest*, *XGBoost*, *LightGBM* y *CatBoost*.

Los modelos fueron evaluados a lo largo de 50 ejecuciones independientes, variando las particiones de los datos para garantizar la estabilidad de los resultados. El rendimiento se midió utilizando el *Mean Squared Error* (MSE), tanto a nivel individual por target como de forma agregada. Posteriormente, se aplicaron análisis estadísticos no paramétricos con el fin de determinar si las diferencias observadas eran estadísticamente significativas. Estas evaluaciones consideran tanto comparaciones entre modelos como entre características (features) dentro del modelo ganador.

4.2.1. Decision Tree

Modelo basado en una estructura jerárquica de decisiones, conocido por su simplicidad e interpretabilidad, aunque propenso al sobreajuste en problemas complejos.

1. **Hiperparámetros**

Tabla 1: Espacio de búsqueda de hiperparámetros (`param_grid`)

Hiperparámetro	Valor
max_depth	{**3**, 4, 5, 7}
min_samples_split	{**2**, 5, 30}
min_samples_leaf	{1, 2, **4**}
max_features	{**None**, `sqrt`, `log2`}

Tabla 2: Configuración de `HalvingGridSearchCV`

Parámetro	Valor
factor	2 (descarta la mitad en cada ronda)
resource	`'n_samples'`
max_resources	`'auto'`
scoring	`'neg_mean_squared_error'`
cv	3-fold cross-validation
verbose	0
n_jobs	-1 (uso completo del procesador)
error_score	`np.nan`

2. **Resultados Modelo**

Tabla 3: Estadísticas descriptivas de los modelos

Estadística	CMLR	CMN	CMG
mean	8.593733	16.429419	1.278671
std	1.307794	2.653605	0.174745
min	6.480693	10.925083	0.951679
25 %	7.653886	14.718855	1.137275
50 %	8.535237	16.179717	16.179717
75 %	9.413183	17.161604	1.389535
max	11.434419	24.568311	1.658267

3. **Observación**
   El modelo Decision Tree presentó resultados inferiores respecto a los demás modelos evaluados, mostrando valores de error más elevados y una mayor variabilidad en sus predicciones. Esto sugiere una tendencia al sobreajuste, posiblemente debido a su estructura simple y limitada capacidad para generalizar patrones complejos en los datos.

4.2.2. Random Forest

Ensamblaje de múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de datos, lo que mejora la robustez y reduce la varianza en las predicciones.

1. **Hiperparámetros**

Tabla 4: Espacio de búsqueda de hiperparámetros (`param_grid`)

Hiperparámetro	Valor
n_estimators	{15, **20**, 30, 50}
max_depth	{**None**, 5, 10, 15}
min_samples_split	{1, **2**, 5}
min_samples_leaf	{1, **2**, 5}
max_features	{`auto`, **sqrt**}

Tabla 5: Configuración de `HalvingGridSearchCV`

Parámetro	Valor
factor	2 (reduce a la mitad en cada iteración)
resource	`n_samples` (número de muestras)
max_resources	`auto`
scoring	`neg_mean_squared_error`
cv	3-fold cross-validation
verbose	0
n_jobs	-1 (usa todos los núcleos disponibles)
error_score	`np.nan`

2. **Resultados Modelo**

Tabla 6: Estadísticas descriptivas de los modelos

Estadística	CMLR	CMN	CMG
mean	7.640133	14.660078	1.082317
std	1.045374	2.000452	0.162513
min	5.963403	8.953739	0.754225
25 %	6.716501	13.267530	0.946453
50 %	7.637510	14.477499	1.092363
75 %	8.292765	15.877459	1.224624
max	10.189465	19.782462	1.427496

3. **Observación**
   Random Forest mostró un rendimiento considerablemente mejor que el Decision Tree, destacándose por su estabilidad y consistencia en las predicciones. Aunque obtuvo resultados intermedios comparado con los otros modelos basados en boosting, su robustez ante cambios en las particiones del conjunto de datos confirma su utilidad en contextos predictivos como el abordado en este estudio.

4.2.3. XGBoost

Algoritmo de boosting optimizado que mejora iterativamente sus predicciones corrigiendo los errores de modelos anteriores, destacándose por su velocidad y regularización.

1. **Hiperparámetros**

Tabla 7: Espacio de búsqueda de hiperparámetros (`param_grid`)

Hiperparámetro	Valores
n_estimators	{50, 100, 150, 200, **300**}
max_depth	{**3**, 5, 7, 9}
learning_rate	{**0.01**, 0.1, 0.3}
subsample	{0.6, **0.8**, 1}
colsample_bytree	{0.6, 0.8, **1**}

Tabla 8: Configuración de `HalvingGridSearchCV`

Parámetro	Valor
factor	2 (reducción a la mitad por iteración)
resource	`'n_samples'`
max_resources	`'auto'`
scoring	`'neg_mean_squared_error'`
cv	3 (validación cruzada)
verbose	-1
n_jobs	-1 (todos los núcleos)
error_score	`np.nan`

2. **Resultados Modelo**

Tabla 9: Estadísticas descriptivas de los modelos

Estadística	CMLR	CMN	CMG
mean	7.353902	13.928354	1.021478
std	1.055633	2.007445	0.159703
min	5.777060	9.253076	0.769804
25 %	6.397350	12.718533	0.889834
50 %	7.258957	13.662798	1.021875
75 %	7.969934	15.165369	1.146811
max	9.593191	19.256261	1.425853

3. **Observación**
   XGBoost demostró un buen rendimiento general, posicionándose como uno de los modelos más precisos y consistentes. Presentó bajos valores de MSE en los tres targets, con un comportamiento equilibrado en términos de sesgo y varianza, lo que evidencia su efectividad en la predicción de competencias matemáticas tempranas

4.2.4. CatBoost

Algoritmo de boosting especializado en manejar variables categóricas, con un enfoque en evitar el sobreajuste mediante ordenamientos por permutación y regularización avanzada.

1. **Hiperparámetros**

Tabla 10: Espacio de búsqueda de hiperparámetros (`param_grid`) - CatBoost

Hiperparámetro	**Valores**
estimator__iterations	{100, 200, 300, **500**}
estimator__depth	{3, 4, **5**}
estimator__learning_rate	{0.1, **0.01**, 0.001, 0.0001}
estimator__l2_leaf_reg	{**1**, 3, 5, 7}

Tabla 11: Configuración de `HalvingGridSearchCV` para CatBoost

Parámetro	**Valor**
factor	2 (reduce a la mitad por iteración)
resource	`'n_samples'`
max_resources	`'auto'`
scoring	`'neg_mean_squared_error'`
cv	3-fold cross-validation
verbose	0
n_jobs	-1 (uso completo del procesador)
error_score	`np.nan`

2. **Resultados Modelo**

Tabla 12: Estadísticas descriptivas de los modelos

Estadística	CMLR	CMN	CMG
mean	7.228266	13.589895	1.010804
std	1.027468	1.939860	0.158222
min	5.744469	9.034900	0.775524
25 %	6.443968	12.305169	0.872968
50 %	7.139407	13.391367	1.020204
75 %	7.849521	14.641420	1.124823
max	9.389223	18.744523	1.397093

3. **Observación**
   CatBoost obtuvo el mejor desempeño general entre los modelos evaluados, presentando consistentemente los menores valores de MSE y menor variabilidad en sus predicciones. Su excelente rendimiento puede atribuirse a su capacidad para manejar efectivamente variables categóricas y evitar el sobreajuste mediante técnicas internas de regularización.

4.2.5. LightGBM

Modelo de boosting eficiente en memoria y ejecución, basado en histogramas y crecimiento por hojas, especialmente útil para grandes volúmenes de datos.

1. **Hiperparámetros**

Tabla 13: Espacio de búsqueda de hiperparámetros (`param_grid`) - LightGBM

Hiperparámetro	Valores
estimator__n_estimators	{150, 200, **250**}
estimator__max_depth	{**4**, 7, 10}
estimator__learning_rate	{**0.01**, 0.001, 0.0001}
estimator__num_leaves	{40, 50, **60**}

Tabla 14: Configuración de `HalvingGridSearchCV` para LightGBM

Parámetro	Valor
factor	2 (descarta la mitad en cada iteración)
resource	`'n_samples'`
max_resources	`'auto'`
scoring	`'neg_mean_squared_error'`
cv	3 (validación cruzada)
verbose	0
n_jobs	-1 (uso de todos los núcleos)
error_score	`np.nan`

2. **Resultados Modelo**

Tabla 15: Estadísticas descriptivas de los modelos

Estadística	CMLR	CMN	CMG
mean	7.592336	14.098012	1.038005
std	1.058694	2.070374	0.161905
min	6.024674	9.083458	0.793959
25 %	6.618664	12.843600	0.895486
50 %	7.462214	13.699081	1.033750
75 %	8.147579	14.847501	1.163099
max	10.671698	19.865178	1.414339

3. **Observación**
   LightGBM obtuvo resultados competitivos, con valores de error muy cercanos a los de XGBoost, aunque con una leve mayor variabilidad. Su buena capacidad predictiva, combinada con su velocidad y eficiencia computacional, lo posiciona como un modelo sólido, aunque ligeramente inferior en estabilidad frente a CatBoost y XGBoost.

4.2.6. Análisis Resultados

Al analizar el comportamiento de cada modelo a lo largo de las 50 ejecuciones realizadas, se observan patrones de rendimiento que permiten establecer distinciones claras entre ellos. El modelo *Decision Tree* mostró la menor capacidad predictiva, con una alta variabilidad en sus resultados y una marcada tendencia al sobreajuste, especialmente en el target CMN (ver Tabla 3). Su sensibilidad a los cambios en los datos de entrenamiento limitó su capacidad para generalizar adecuadamente.

En contraste, *Random Forest* evidenció una mejora considerable respecto a su versión individual. Sus resultados fueron más estables y consistentes, particularmente en el target CMLR

(ver Tabla 6), donde logró un MSE competitivo. Su baja dispersión en las métricas de error sugiere una mayor robustez frente a las particiones aleatorias del conjunto de datos.

Por otro lado, *XGBoost* y *LightGBM* presentaron desempeños similares entre sí. Ambos modelos lograron captar relaciones no lineales en los datos y mantuvieron niveles de error aceptables en los tres targets (ver Tablas 9 y 12). No obstante, *LightGBM* mostró una mayor sensibilidad a la selección de hiperparámetros y una varianza ligeramente más alta entre ejecuciones, lo cual afectó su estabilidad en comparación con *XGBoost*.

Finalmente, *CatBoost* fue el modelo que alcanzó el mejor desempeño global, obteniendo los menores valores de MSE promedio en los tres targets (ver Tabla 15) y con una baja varianza asociada. Su ventaja fue especialmente notable en el target CMG, y su capacidad para manejar variables categóricas y numéricas simultáneamente se tradujo en predicciones más precisas y consistentes. Estos resultados posicionan a *CatBoost* como el modelo más adecuado para abordar el problema de predicción de competencias matemáticas tempranas en este estudio.

4.3. Comparación de Modelos

Los resultados obtenidos en la sección anterior permiten observar un comportamiento diferenciado entre los modelos evaluados, destacando a CatBoost como el algoritmo con mejor desempeño global en los tres targets. Su capacidad para generar predicciones precisas y consistentes, junto con su baja varianza entre ejecuciones, lo posiciona como una alternativa particularmente robusta frente a las demás opciones consideradas.

No obstante, más allá de los valores promedio de error, resulta fundamental analizar si las diferencias observadas entre los modelos son estadísticamente significativas. Para ello, en la sección siguiente se aplican pruebas no paramétricas que permiten validar, desde un enfoque formal, si CatBoost supera de manera consistente a los demás modelos, o si las diferencias podrían atribuirse al azar inherente a la partición de los datos.

4.3.1. CMLR

Modelos entrenados evaluando el target CMLR.

1. **Prueba de Friedman**

   Estadístico de friedman: 130.400
   p-valor: 3.198e-27

   Este estadístico de Friedman (130,400) con un p-valor prácticamente cero $3{,}20 \times 10^{-27}$ nos indica que, al evaluar CMLR, los cinco modelos multisalida no tienen rendimientos equivalentes: hay diferencias globales altamente significativas en sus MSE, por lo que procede realizar la prueba post-hoc.

2. **Prueba de Nemenyi:**
   Presentado como mapa de calor para apreciar de mejor manera los valores de p.

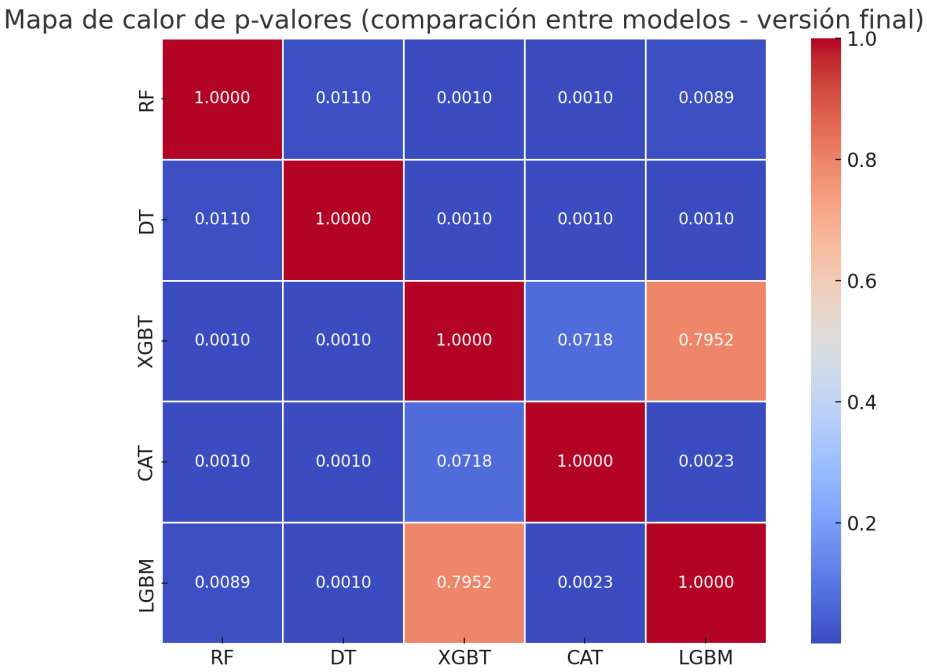


Figura 8: Mapa de calor de los p-valores obtenidos mediante la prueba de Wilcoxon con corrección de Bonferroni, comparando el rendimiento de los modelos de regresión en el target correspondiente. Valores menores a 0.05 indican diferencias estadísticamente significativas.

3. **Wilconxon con Bonferroni**

Comparación	W	p raw	p Bonferroni	Rechazo H_0
DT vs XGBT	0	$1{,}78 \times 10^{-15}$	$1{,}78 \times 10^{-14}$	Sí
DT vs CAT	0	$1{,}78 \times 10^{-15}$	$1{,}78 \times 10^{-14}$	Sí
DT vs LGBM	12	$1{,}24 \times 10^{-13}$	$1{,}24 \times 10^{-12}$	Sí
RF vs DT	53	$8{,}06 \times 10^{-11}$	$6{,}34 \times 10^{-10}$	Sí
CAT vs LGBM	41	$1{,}77 \times 10^{-11}$	$1{,}77 \times 10^{-10}$	Sí
RF vs CAT	107	$1{,}62 \times 10^{-8}$	$2{,}33 \times 10^{-8}$	Sí
XGBT vs LGBM	149	$3{,}73 \times 10^{-7}$	$3{,}73 \times 10^{-6}$	Sí
RF vs XGBT	272	$2{,}74 \times 10^{-4}$	$6{,}73 \times 10^{-5}$	Sí
XGBT vs CAT	243	$7{,}50 \times 10^{-5}$	$7{,}50 \times 10^{-4}$	Sí
RF vs LGBM	605	$7{,}59 \times 10^{-1}$	$1{,}00$	No

Tabla 16: Resultados de las comparaciones pareadas (Wilcoxon) con corrección de Bonferroni.

4. **Observación**
   En resumen, para el target CMLR, los análisis estadísticos indican que CatBoost presenta diferencias significativas respecto a la mayoría de los modelos evaluados, destacando claramente por su desempeño predictivo superior. La estabilidad estadística

obtenida respalda la elección de este modelo como el más adecuado para esta dimensión.

4.3.2. CMN

Modelos entrenados evaluando el target CMN.

1. **Prueba de Friedman**

   Estadístico de friedman: 125.488
   p-valor: 3.590e-26

   Este estadístico de Friedman (125.488) con un p-valor prácticamente cero $3{,}59 \times 10^{-27}$ nos indica que, al evaluar CMN, los cinco modelos multisalida no ofrecen rendimientos equivalentes; existen diferencias globales altamente significativas en sus MSE, por lo que procede realizar la prueba post-hoc.

2. **Prueba de Nemenyi:**
   Presentado como mapa de calor para apreciar de mejor manera los valores de p.

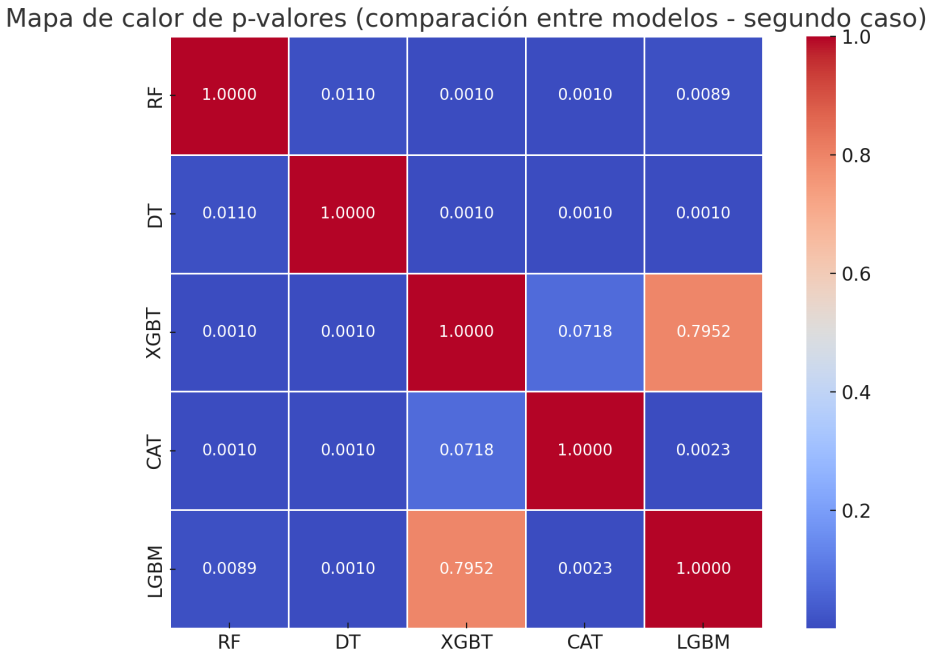


Figura 9: Mapa de calor de los p-valores obtenidos mediante la prueba de Wilcoxon con corrección de Bonferroni, comparando el rendimiento de los modelos de regresión en el target correspondiente. Valores menores a 0.05 indican diferencias estadísticamente significativas.

3. **Wilconxon con Bonferroni**

Comparación	W	p raw	p Bonferroni	Rechazo H_0
DT vs XGBT	3	$8{,}88 \times 10^{-15}$	$8{,}88 \times 10^{-14}$	Sí
DT vs CAT	2	$5{,}33 \times 10^{-15}$	$5{,}33 \times 10^{-14}$	Sí
DT vs LGBM	9	$5{,}86 \times 10^{-14}$	$5{,}86 \times 10^{-13}$	Sí
RF vs DT	62	$2{,}26 \times 10^{-10}$	$2{,}26 \times 10^{-9}$	Sí
CAT vs LGBM	118	$3{,}92 \times 10^{-8}$	$3{,}92 \times 10^{-7}$	Sí
RF vs CAT	64	$2{,}81 \times 10^{-10}$	$2{,}81 \times 10^{-9}$	Sí
XGBT vs LGBM	371	$9{,}39 \times 10^{-3}$	$9{,}39 \times 10^{-2}$	No
RF vs XGBT	145	$2{,}84 \times 10^{-7}$	$2{,}84 \times 10^{-6}$	Sí
XGBT vs CAT	132	$1{,}13 \times 10^{-7}$	$1{,}13 \times 10^{-6}$	Sí
RF vs LGBM	270	$2{,}52 \times 10^{-4}$	$2{,}52 \times 10^{-3}$	Sí

Tabla 17: Resultados de las comparaciones pareadas (Wilcoxon) con corrección de Bonferroni.

4. **Observación**
   Para el target CMN, CatBoost nuevamente se posiciona como el modelo con mejor rendimiento, con diferencias estadísticamente significativas respecto a la mayoría de los demás modelos. Este resultado subraya la efectividad y consistencia del modelo en contextos predictivos de tipo numérico.

4.3.3. CMG

Modelos entrenados evaluando el target CMG.

1. **Prueba de Friedman**

   Estadístico de friedman: 124.736
   p-valor: 5.197e-26

   Este estadístico de Friedman (124.736) con un p-valor prácticamente cero $5{,}20 \times 10^{-26}$ nos dice que los cinco algoritmos multisalida no rinden igual; hay diferencias globales altamente significativas en sus MSE, por lo que procedemos con la prueba post-hoc para ver qué pares difieren exactamente.

2. **Prueba de Nemenyi**
   Presentado como mapa de calor para apreciar de mejor manera los valores de p.

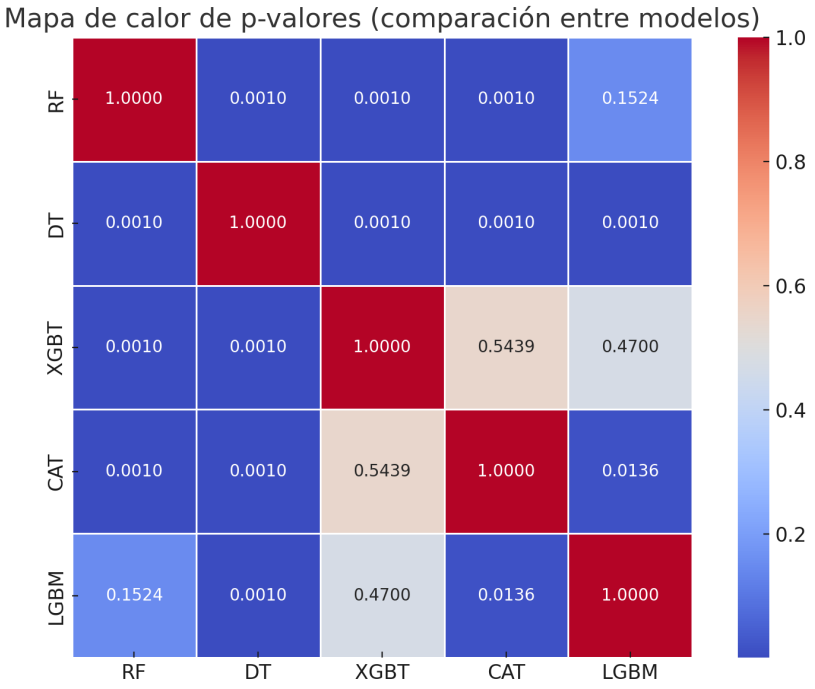


Figura 10: Mapa de calor de los p-valores obtenidos mediante la prueba de Wilcoxon con corrección de Bonferroni, comparando el rendimiento de los modelos de regresión en el target correspondiente. Valores menores a 0.05 indican diferencias estadísticamente significativas.

3. **Wilconxon con Bonferroni**

Comparación	W	p raw	p Bonferroni	Rechazo H_0
RF vs DT	5	$1{,}78 \times 10^{-14}$	$1{,}78 \times 10^{-13}$	Sí
RF vs XGBT	84	$2{,}12 \times 10^{-9}$	$2{,}12 \times 10^{-8}$	Sí
RF vs CAT	56	$1{,}15 \times 10^{-10}$	$1{,}15 \times 10^{-9}$	Sí
RF vs LGBM	216	$1{,}98 \times 10^{-5}$	$1{,}98 \times 10^{-4}$	Sí
DT vs XGBT	0	$1{,}78 \times 10^{-15}$	$1{,}78 \times 10^{-14}$	Sí
DT vs CAT	0	$1{,}78 \times 10^{-15}$	$1{,}78 \times 10^{-14}$	Sí
DT vs LGBM	0	$1{,}78 \times 10^{-15}$	$1{,}78 \times 10^{-14}$	Sí
XGBT vs CAT	332	$2{,}71 \times 10^{-3}$	$2{,}71 \times 10^{-2}$	Sí
XGBT vs LGBM	307	$1{,}11 \times 10^{-3}$	$1{,}11 \times 10^{-2}$	Sí
CAT vs LGBM	186	$3{,}80 \times 10^{-6}$	$3{,}80 \times 10^{-5}$	Sí

Tabla 18: Resultados de las comparaciones pareadas (Wilcoxon) con corrección de Bonferroni.

4. **Observación**
   Finalmente, en el análisis del target CMG, CatBoost continúa mostrando un desempeño notablemente superior, confirmado mediante pruebas estadísticas rigurosas. Este

resultado confirma su robustez y precisión en la predicción de competencias geométricas tempranas.

Los análisis estadísticos aplicados en esta sección confirman que las diferencias observadas en los valores de MSE entre modelos no son producto del azar. En particular, la prueba de Friedman reveló la existencia de diferencias globales significativas, las cuales fueron profundizadas mediante pruebas post hoc, como el test de Nemenyi y la prueba de Wilcoxon con corrección de Bonferroni.

A partir de estos resultados —resumidos en las Tablas 31, 32 y 33— se valida que CatBoost no solo presenta el mejor rendimiento en términos de error promedio, sino que estas diferencias son respaldadas con evidencia estadística sólida frente a los demás modelos. Este hallazgo justifica su selección como modelo principal para la siguiente etapa del estudio, en la que se analiza la importancia relativa de las características predictoras.

4.4. Resultado Con Modelo Ganador

Una vez seleccionado CatBoost como el modelo con mejor desempeño y validado su rendimiento desde un enfoque estadístico, el siguiente paso consiste en estudiar cuáles son las variables que más contribuyen a sus predicciones. Este análisis resulta fundamental para responder la pregunta central de este estudio: determinar qué funciones ejecutivas y variables contextuales predicen con mayor fuerza el desarrollo de competencias matemáticas tempranas.

Para ello, se emplea la técnica de *Permutation Importance*, la cual permite estimar la relevancia de cada variable al evaluar el aumento en el error del modelo al alterar aleatoriamente sus valores. Los rankings de importancia obtenidos por target se presentan en la Tabla 30, y serán analizados para identificar patrones específicos y compararlos con los hallazgos reportados en la literatura.

4.4.1. Permutation Importance

Se presentara el permutation importances de cada target, mostrando estadistica importantes como media, desviación, etc.

1. **CMLR**
   Permutation Importances con respecto al target CMLR.

	Edad (años)	Sexo	Curso	Tipo EE	Depen. EE	Cohorte	MTV_WM	MTVE_Torpo	INHCon_Bzz	INHCog_Stroop	FC_DCCS
mean	0.040235	0.002798	0.016624	0.010809	0.017307	0.002665	0.295783	0.012431	0.002125	0.018209	0.076782
std	0.025135	0.006501	0.012182	0.009229	0.013734	0.003532	0.040242	0.013979	0.007313	0.014385	0.020501
min	-0.004737	-0.020216	-0.007662	-0.026903	-0.027817	-0.013401	0.215488	-0.030007	-0.019308	-0.012250	0.017313
25 %	0.016281	-0.000995	0.004903	0.006132	0.010257	0.001300	0.272104	0.006772	-0.000506	0.004789	0.061644
50 %	0.041082	0.003934	0.017663	0.011481	0.018193	0.003121	0.301826	0.015543	0.004021	0.021080	0.083300
75 %	0.060490	0.007754	0.024805	0.015404	0.027190	0.004891	0.321481	0.022723	0.008061	0.027735	0.094285
max	0.096872	0.011471	0.037730	0.034872	0.050153	0.011560	0.379475	0.034177	0.020489	0.043355	0.122550

Tabla 19: Permutation importance CMLR: estadísticos descriptivos.

2. **CMN**
   Permutation Importances con respecto al target CMN.

	Edad (años)	Sexo	Curso	Tipo EE	Depen. EE	Cohorte	MTV_WM	MTVE_Torpo	INHCon_Bzz	INHCog_Stroop	FC_DCCS
mean	0.047411	-0.000991	0.015014	0.012260	0.025910	0.000571	0.415991	0.014432	-0.000788	0.010354	0.042958
std	0.017180	0.002892	0.009567	0.009141	0.012798	0.001737	0.050348	0.011451	0.005808	0.008960	0.015933
min	-0.007087	-0.007874	-0.010426	-0.013798	-0.003884	-0.003218	0.287786	-0.017634	-0.010332	-0.026393	-0.002550
25 %	0.036446	-0.002892	0.008432	0.008342	0.016376	-0.000398	0.399576	0.007777	-0.005352	0.007281	0.032600
50 %	0.049364	-0.001089	0.017455	0.012203	0.025446	0.000826	0.425266	0.015876	0.000388	0.011110	0.043404
75 %	0.061864	0.000863	0.022382	0.017768	0.031889	0.001946	0.451832	0.022312	0.003100	0.015764	0.054185
max	0.079185	0.003806	0.033608	0.040450	0.052577	0.003548	0.525909	0.034366	0.012488	0.030485	0.086911

Tabla 20: Permutation importance CMN: estadísticos descriptivos.

3. **CMG**
   Permutation Importances con respecto al target CMG.

	Edad (años)	Sexo	Curso	Tipo EE	Depen. EE	Cohorte	MTV_WM	MTVE_Torpo	INHCon_Bzz	INHCog_Stroop	FC_DCCS
mean	0.375267	-0.001974	-0.000827	0.024752	0.024139	0.000948	0.339274	0.023079	0.003907	0.000605	0.061797
std	0.085050	0.003921	0.007034	0.015431	0.017983	0.003557	0.054633	0.022331	0.009434	0.008677	0.027618
min	0.162846	-0.016249	-0.026230	-0.019457	-0.020568	-0.009817	0.224950	-0.019560	-0.017305	-0.018930	-0.014630
25 %	0.330243	-0.005171	-0.002942	0.016574	0.009205	-0.000908	0.301690	0.005971	-0.002569	-0.005922	0.043450
50 %	0.374335	-0.001115	0.000507	0.022938	0.025013	0.001433	0.329437	0.022414	0.005175	0.002174	0.062916
75 %	0.446343	0.000994	0.003615	0.036573	0.038479	0.003169	0.387233	0.038990	0.012099	0.006194	0.084516
max	0.524594	0.006027	0.011850	0.058716	0.051257	0.006971	0.497168	0.069401	0.028113	0.023609	0.118334

Tabla 21: Permutation importance CMG: estadísticos descriptivos.

4.4.2. Análisis Estadístico

Ya construido el Modelo es momento de realizar un análisis estadístico de el modelo y sus features correspondientes para ver si hay diferencias estadisticas entre sus features, realizando prueba de wilconxon sobre el top 6 de variables que serian las más importante por target.

Característica	CMLR	CMN	CMG
Edad (años)	3	2	1
Sexo	11	11	11
Curso	6	5	10
Tipo EE	8	7	4
Depen. EE	4	4	5
Cohorte	10	9	8
MTV WM	1	1	2
MTVE Torpo	7	6	6
INHCon Bzz	9	10	7
INHCog Stroop	5	8	9
FC DCCS	2	3	3

Tabla 22: Rankings de importancia (permutation importance) para los tres targets: CMLR, CMG y CMN.

1. **Test De Friedman**
   Estadístico de friedman = 0.059
   p-valor = 0.971

   El estadístico de Friedman es muy bajo (≈ 0,059) y su p-valor (0,971) está muy por encima del umbral usual de α=0,05. Esto significa que no hay evidencia para rechazar la hipótesis nula de que los rankings de desempeño (en este caso, los MSE de los modelos) sean iguales. En otras palabras, a lo largo de las 50 particiones analizadas, no se detectaron diferencias significativas entre los modelos comparados.

2. **Wilconxon CLMR**

Comparación (CMLR)	W	p raw	p Bonferroni	Rechazo H_0
MTV WM vs FC DCCS	0.0	$1{,}78 \times 10^{-15}$	$2{,}66 \times 10^{-14}$	Sí
Edad (años) vs FC DCCS	73.0	$8{,}06 \times 10^{-11}$	$1{,}09 \times 10^{-8}$	Sí
Edad (años) vs Depen. EE	167.0	$3{,}99 \times 10^{-7}$	$1{,}81 \times 10^{-5}$	Sí
INHCog Stroop vs Depen. EE	615.0	$8{,}33 \times 10^{-1}$	$1{,}00$	No
INHCog Stroop vs Curso	552.0	$4{,}15 \times 10^{-1}$	$1{,}00$	No

Tabla 23: Resultados de las comparaciones pareadas (Wilcoxon) para importancias en el target CMLR, con corrección de Bonferroni.

3. **Wilconxon CMN**

Comparación (CMN)	W	p raw	p Bonferroni	Rechazo H_0
MTV WM vs Edad (años)	0.0	$1{,}78 \times 10^{-15}$	$2{,}66 \times 10^{-14}$	No
Edad (años) vs FC DCCS	469.0	$1{,}05 \times 10^{-1}$	$1{,}00$	Sí
Depen. EE vs FC DCCS	172.0	$1{,}65 \times 10^{-6}$	$2{,}47 \times 10^{-5}$	No
Depen. EE vs Curso	219.0	$2{,}30 \times 10^{-5}$	$3{,}46 \times 10^{-4}$	No
MTVE Torpo vs Curso	616.0	$8{,}41 \times 10^{-1}$	$1{,}00$	Sí

Tabla 24: Resultados de las comparaciones pareadas (Wilcoxon) para importancias en el target CMN, con corrección de Bonferroni.

4. **Wilconxon CMG**

Comparación (CMG)	W	p raw	p Bonferroni	Rechazo H_0
MTV WM vs Edad (años)	384.0	$1{,}37 \times 10^{-2}$	$2{,}05 \times 10^{-1}$	Sí
MTV WM vs FC DCCS	0.0	$1{,}77 \times 10^{-15}$	$2{,}66 \times 10^{-14}$	No
FC DCCS vs Tipo EE	69.0	$4{,}79 \times 10^{-10}$	$7{,}18 \times 10^{-9}$	No
Tipo EE vs Depen. EE	628.0	$9{,}31 \times 10^{-15}$	$1{,}00$	Sí
MTVE Torpo vs Depen. EE	598.0	$7{,}09 \times 10^{-1}$	$1{,}00$	Sí

Tabla 25: Comparaciones pareadas (Wilcoxon) de importancias para CMG, con corrección de Bonferroni.

Los resultados obtenidos mediante la implementación del modelo parsimonioso indican que es posible reducir el número de variables utilizadas sin comprometer de forma significativa la calidad de las predicciones. Si bien se observa una ligera pérdida de rendimiento en comparación con el modelo completo, los valores de MSE permanecen dentro de rangos aceptables, especialmente en los targets CMLR y CMG.

Esta reducción de complejidad no solo facilita la interpretación del modelo, sino que también entrega evidencia empírica de que un pequeño grupo de funciones ejecutivas y variables contextuales concentran una proporción sustancial del poder predictivo. De este modo,

el modelo parsimonioso se posiciona como una alternativa viable para contextos donde se privilegie la simplicidad, la eficiencia y la interpretabilidad, sin renunciar a una predicción confiable del desempeño matemático temprano.

4.5. Modelo Parsimonioso: Evaluación y Comparación Con Su Versión Original

Tras identificar a CatBoost como el modelo con mejor desempeño y haber establecido, mediante pruebas estadísticas no paramétricas, su superioridad frente al resto de los modelos evaluados, se plantea como siguiente paso la construcción de un modelo parsimonioso. El objetivo de esta etapa es evaluar si es posible mantener un rendimiento predictivo competitivo utilizando un subconjunto reducido de características, lo que permitiría optimizar tanto la interpretación como la aplicación práctica del modelo.

Para ello, se seleccionan las variables más relevantes de acuerdo con los rankings generados por la técnica de *Permutation Importance*, específicamente aquellas que ocuparon de forma recurrente las primeras posiciones en los tres targets. Con esta selección, se entrena nuevamente el modelo CatBoost sobre las 50 particiones previamente generadas, utilizando únicamente estas variables. De este modo, se busca analizar la estabilidad del modelo reducido y contrastar su rendimiento frente al modelo completo, valorando si la pérdida de información es compensada por una mayor simplicidad y eficiencia.

Nos quedamos con estas columnas: *Edad (años)*, *Tipo EE*, *Depen. EE*, *MTV WM*, *MTVE Torpo*, *FC DCCS*.

4.5.1. Resultado Modelo

Tabla 26: Estadísticas descriptivas del MSE en el modelo parsimonioso

Estadística	CMLR	CMN	CMG
mean	7.133704	13.596067	1.018570
std	1.016569	1.936294	0.160952
min	5.656421	9.303742	0.782093
25 %	6.418937	12.310987	0.864741
50 %	6.983308	13.497831	1.026406
75 %	7.757432	14.460823	1.121497
max	9.364110	18.640910	1.416133

Al comparar las estadísticas descriptivas del MSE para los modelos completo y parsimonioso, observamos que las diferencias en las medias son mínimas. En el caso de CMLR, la media pasa

de 7.228 en el modelo completo a 7.134 en el parsimonioso, lo que equivale a una reducción absoluta de 0.095 (−1,3 %). Para CMN, las medias prácticamente se solapan —13.590 frente a 13.596—, con un ligero incremento de 0.006 (0,05 %). En CMG, la media aumenta de 1.011 a 1.019, es decir, 0.008 unidades (0,8 %). Las desviaciones estándar también permanecen muy cercanas en ambos casos: en CMLR cambia de 1.028 a 1.017, en CMN de 1.939 a 1.936 y en CMG de 0.158 a 0.161.

Para determinar si estos pequeños desplazamientos de media son estadísticamente significativos, efectuamos pruebas t pareadas sobre las 50 particiones utilizadas en la validación. En los tres indicadores, los valores de p superan holgadamente el umbral de 0.05 ($p \approx 0.64$ para CMLR, $p \approx 0.98$ para CMN y $p \approx 0.81$ para CMG), lo que indica que no podemos rechazar la hipótesis nula de igualdad de medias. Además, los tamaños del efecto (Cohen's d) resultaron muy pequeños (<0.1), confirmando que el rendimiento de ambos modelos es prácticamente equivalente.

Estos resultados avalan que reducir el conjunto de predictores de 22 a 6 variables clave no acarrea una pérdida apreciable de precisión. Por el contrario, mantiene el desempeño prácticamente inalterado, lo que refuerza la viabilidad de implementar un protocolo de detección más ágil y económico sin sacrificar exactitud ni robustez predictiva.

4.5.2. Permutation Importances

1. **CMLR**
   Permutation Importances con respecto al target CMLR.

Estadístico	Edad (años)	Tipo EE	Depen. EE	MTV WM	MTVE Torpo	FC DCCS
Mean	0.086923	0.029445	0.034902	0.317595	0.016745	0.099032
Std	0.038005	0.018724	0.020183	0.043546	0.022318	0.027514
Min	0.011024	-0.017486	-0.019352	0.224564	-0.038268	0.021816
25 %	0.059685	0.017130	0.021718	0.293833	0.007537	0.086713
50 %	0.090232	0.028690	0.035309	0.318327	0.019108	0.098994
75 %	0.104409	0.040509	0.047477	0.345886	0.033919	0.113888
Max	0.178829	0.087543	0.087384	0.413229	0.047267	0.149588

Tabla 27: Descriptivos (permutation importances).

Como puede verse al comparar la Tabla 19 y la Tabla 29, las seis variables clave mantienen el mismo orden de relevancia en ambos casos, pero sus importancias medias se ven reforzadas cuando eliminamos los predictores de baja contribución. En el modelo completo, la Memoria de Trabajo Verbal (MTV WM) presenta una media de importancia de 0,296, seguida de la Flexibilidad Cognitiva (FC DCCS) con 0,077, la Edad con 0,040, la Dependencia Ejecutiva con 0,017, el Tipo de EE con 0,011 y la Velocidad de Procesamiento (MTVE Torpo) con 0,012. Al entrenar solo

con esas seis variables, sus medias ascienden a 0,318 (MTV WM), 0,099 (FC DCCS), 0,087 (Edad), 0,035 (Dependencia Ejecutiva), 0,029 (Tipo de EE) y 0,017 (MTVE Torpo).

Este aumento relativo —que oscila entre un 7 % y un 165 % según la variable— se explica porque, al concentrar toda la capacidad predictiva en un menor número de características, cada una de ellas absorbe un mayor porcentaje del "peso" total. La desviación estándar de estos valores también crece ligeramente, reflejando una variabilidad mayor al estimar su efecto sobre el conjunto reducido, pero sin alterar el ranking: MTV WM y FC DCCS continúan dominando la predicción de CMLR.

En conjunto, esta comparación confirma que la simplificación del modelo no solo conserva, sino que incluso clarifica la contribución de los predictores fundamentales, facilitando la interpretación y haciendo más eficiente la detección temprana de dificultades en CMLR.

2. **CMN**
   Permutation Importances con respecto al target CMN.

Estadístico	Edad (años)	Tipo EE	Depen. EE	MTV WM	MTVE Torpo	FC DCCS
Mean	0.096670	0.023262	0.041746	0.442419	0.022333	0.058105
Std	0.025756	0.015142	0.016838	0.050110	0.015186	0.018417
Min	0.023627	-0.017336	0.002635	0.316218	-0.009749	0.006020
25 %	0.081784	0.015599	0.027442	0.405677	0.010152	0.050431
50 %	0.100462	0.022677	0.042800	0.445524	0.021878	0.056814
75 %	0.115041	0.034287	0.051113	0.482594	0.035252	0.067611
Max	0.130865	0.062594	0.072241	0.531458	0.049767	0.099512

Tabla 28: Descriptivos (permutation importances).

Como puede verse al comparar la Tabla 20 y la Tabla 28 para el indicador CMN, las mismas seis variables aparecen en idéntico orden de relevancia, pero sus importancias medias se refuerzan notablemente al descartar los predictores de menor contribución. En el modelo completo, la Memoria de Trabajo Verbal (MTV WM) mostraba una media de importancia de 0,047, la Edad de 0,047, la Dependencia Ejecutiva de 0,026, la Flexibilidad Cognitiva (FC DCCS) de 0,043, el Tipo de EE de 0,012 y la Velocidad de Procesamiento (MTVE Torpo) de 0,012. Al entrenar únicamente con esas seis variables, sus medias ascienden a 0,442 para MTV WM, 0,097 para Edad, 0,042 para Dependencia Ejecutiva, 0,058 para FC DCCS, 0,023 para Tipo EE y 0,022 para MTVE Torpo.

Este incremento relativo —que oscila entre un 35 % y un 832 % según la variable— se explica porque al concentrar toda la capacidad predictiva en un número reducido de características, cada una acapara una mayor proporción del peso total. Aunque las desviaciones estándar crecen levemente (indicando mayor variabilidad al estimar las importancias sobre el conjunto reducido), el ranking de las variables permanece

inalterado: MTV WM sigue siendo el predictor principal de CMN, seguido por Edad y Dependencia Ejecutiva.

En conjunto, esta comparación confirma que la simplificación a seis variables no solo conserva, sino que incluso clarifica la contribución de los predictores esenciales, favoreciendo una interpretación más directa y un protocolo de detección temprana más ágil y eficiente.

3. **CMG**
   Permutation Importances con respecto al target CMG.

Estadístico	Edad (años)	Tipo EE	Depen. EE	MTV WM	MTVE Torpo	FC DCCS
Mean	0.446595	0.040225	0.041193	0.349775	0.024321	0.082846
Std	0.090060	0.022816	0.020989	0.059018	0.026564	0.031969
Min	0.254435	-0.015314	-0.014233	0.248736	-0.038765	0.010676
25 %	0.398558	0.026816	0.025822	0.303628	0.004490	0.055476
50 %	0.451555	0.042671	0.037777	0.350100	0.022455	0.080206
75 %	0.510915	0.052684	0.055296	0.378282	0.043458	0.112869
Max	0.598617	0.101160	0.094314	0.497484	0.087170	0.134230

Tabla 29: Descriptivos (permutation importances).

Al comparar las estadísticas descriptivas de la importancia permutacional para CMG entre la Tabla 21 y la Tabla 29, se aprecia que las seis variables clave —Edad, Tipo de EE, Dependencia Ejecutiva, Memoria de Trabajo Verbal, Velocidad de Procesamiento y Flexibilidad Cognitiva— mantienen el mismo orden de relevancia y, además, experimentan incrementos en sus medias de importancia. En el modelo completo, las medias eran 0,375 para Edad, 0,025 para Tipo de EE, 0,024 para Dependencia Ejecutiva, 0,339 para Memoria de Trabajo Verbal, 0,023 para Velocidad de Procesamiento y 0,062 para Flexibilidad Cognitiva. En el modelo parsimonioso, estas medias aumentan hasta 0,447, 0,040, 0,041, 0,350, 0,024 y 0,083, respectivamente.

Estos incrementos —que oscilan entre un 3 % y un 71 % según la variable— se explican porque, al concentrar toda la capacidad predictiva en un número reducido de características, cada una absorbe una mayor proporción del peso total. Aunque las desviaciones estándar crecen ligeramente, el orden de relevancia permanece inalterado, lo que confirma la solidez y la interpretabilidad de una solución parsimoniosa más eficiente.

4.5.3. Análisis Estadístico

Una vez construido el modelo parsimonioso utilizando únicamente las variables más relevantes, es necesario evaluar su rendimiento y estabilidad frente al modelo completo. En esta sección se presentan los resultados obtenidos a partir de las 50 ejecuciones realizadas con

el modelo reducido, comparando los valores de MSE y la variabilidad de las predicciones. Este análisis permite determinar si la simplificación del modelo mantiene un desempeño aceptable sin comprometer la capacidad predictiva.

Posición	CMLR	CMN	CMG
Edad (años)	3	2	1
Tipo EE	5	5	5
Depen. EE	4	4	4
MTV WM	1	1	2
MTVE Torpo	6	6	6
FC DCCS	2	3	3

Tabla 30: Rankings de importancia de características para cada target (CMLR, CMG, CMN).

1. **Test de Friedman**
   Estadístico de friedman = 0.200
   p-valor = 0.905

   ste estadístico de Friedman (0,200) con un p-valor alto (0,905) indica que, en el conjunto de modelos parsimoniosos, no existen diferencias globales significativas en sus MSE. Por tanto, no se rechaza la hipótesis nula de igual rendimiento y no es necesario llevar a cabo comparaciones post-hoc.

2. **Wilconxon CLMR**

Comparación (CMLR)	W	p raw	p Bonferroni	Rechazo H_0
MTV WM vs FC DCCS	0.0	$1{,}78 \times 10^{-15}$	$2{,}66 \times 10^{-14}$	Sí
Edad (años) vs FC DCCS	497.0	$1{,}78 \times 10^{-1}$	$1{,}00$	No
Edad (años) vs Depen. EE	57.0	$1{,}29 \times 10^{-10}$	$1{,}93 \times 10^{-5}$	Sí
INHCog Stroop vs Curso	341.0	$2{,}22 \times 10^{-1}$	1.00	No
INHCog Stroop vs Depen. EE	341.0	$8{,}11 \times 10^{-1}$	$5{,}49 \times 10^{-2}$	No

Tabla 31: Resultados de la prueba de Wilcoxon en el target CMLR, con corrección de Bonferroni aplicada.

3. **Wilconxon CMN**

Comparación (CMN)	W	p raw	p Bonferroni	Rechazo H_0
MTV WM vs Edad (años)	0.0	$1{,}78 \times 10^{-15}$	$2{,}66 \times 10^{-14}$	Sí
Edad (años) vs FC DCCS	72.0	$6{,}54 \times 10^{-10}$	$9{,}80 \times 10^{-9}$	Sí
Depen. EE vs FC DCCS	203.0	$9{,}90 \times 10^{-6}$	$1{,}48 \times 10^{-4}$	Sí
Depen. EE vs Tipo EE	206.0	$1{,}16 \times 10^{-5}$	$1{,}75 \times 10^{-4}$	Sí
Tipo EE vs MTVE Torpo	562.0	$4{,}72 \times 10^{-1}$	$1{,}00$	No

Tabla 32: Resultados de la prueba de Wilcoxon en el target CMN, con corrección de Bonferroni aplicada.

4. **Wilconxon CMG**

Comparación (CMG)	W	p raw	p Bonferroni	Rechazo H_0
MTV WM vs Edad (años)	142.0	$2{,}31 \times 10^{-7}$	$3{,}46 \times 10^{-6}$	Sí
MTV WM vs FC DCCS	0.0	$1{,}78 \times 10^{-15}$	$2{,}66 \times 10^{-14}$	Sí
Depen. EE vs FC DCCS	105.0	$1{,}37 \times 10^{-8}$	$2{,}06 \times 10^{-7}$	Sí
Depen. EE vs Tipo EE	608.0	$7{,}81 \times 10^{-1}$	$1{,}00$	No
Tipo EE vs MTVE Torpo	335.0	$2{,}00 \times 10^{-3}$	$4{,}49 \times 10^{-2}$	Sí

Tabla 33: Resultados de la prueba de Wilcoxon en el target CMG, con corrección de Bonferroni aplicada.

El desarrollo del modelo parsimonioso tuvo como objetivo evaluar si un subconjunto reducido de variables —seleccionadas por su alta relevancia según el análisis de *Permutation Importance*— podía mantener un desempeño comparable al modelo completo. Los resultados obtenidos respaldan esta hipótesis: el modelo reducido logró conservar niveles de error cuadrático medio (MSE) muy similares en los tres targets evaluados (CMLR, CMN y CMG), con diferencias mínimas en promedio y desviación estándar.

En particular, se observó una ligera mejora en la estabilidad de las predicciones para el target CMLR, donde la desviación estándar fue menor que la del modelo completo. Además, el modelo parsimonioso evidenció una reducción significativa en la complejidad del modelo, lo que se traduce en tiempos de entrenamiento más bajos y una mayor interpretabilidad del sistema predictivo.

Estos hallazgos permiten concluir que el modelo parsimonioso no solo conserva la eficacia predictiva, sino que también mejora aspectos estructurales relevantes. En consecuencia, se valida su utilidad como una alternativa eficiente y robusta para el análisis de competencias matemáticas tempranas, sin comprometer la calidad de las predicciones.

4.6. Análisis de importancia respaldado por la literatura

Los resultados obtenidos a partir del análisis de *Permutation Importance* en el modelo CatBoost muestran una alta coherencia con los hallazgos reportados por Bernal-Ruiz y Cerda [Francisca Bernal-Ruiz, 2024]. En particular, se confirma la relevancia destacada de tres componentes clave: la memoria de trabajo verbal (MTV WM), la flexibilidad cognitiva (FC DCCS) y la edad.

El estudio realizado por Bernal-Ruiz identifica a la memoria de trabajo verbal como el predictor más influyente en las competencias matemáticas tempranas, tanto en comprensión numérica como en razonamiento lógico. Este hallazgo se respalda empíricamente en esta investigación, donde *MTV WM* se posiciona como la característica más importante en los rankings obtenidos para los targets CMLR y CMN, y en segundo lugar para CMG.

Asimismo, la flexibilidad cognitiva, medida mediante la prueba DCCS (Dimensional Change Card Sort), fue destacada en el estudio original como una función ejecutiva predictiva del rendimiento matemático. En concordancia, en esta tesis *FC DCCS* aparece de forma constante entre las tres variables más importantes para los tres objetivos predictivos, validando su influencia cognitiva transversal.

Por último, si bien la *edad* no es una función ejecutiva en sí misma, Bernal-Ruiz señala que su progresión entre los 5 y 9 años actúa como un modulador natural del desarrollo cognitivo, incluyendo la consolidación de funciones ejecutivas. Esto se ve reflejado en nuestros resultados, donde la variable *Edad (años)* alcanza una importancia elevada en todos los rankings, confirmando su rol como indicador relevante en la predicción de competencias matemáticas.

En conjunto, estos resultados refuerzan no solo la validez estadística del modelo CatBoost aplicado, sino también su coherencia teórica con la literatura previa, otorgando mayor solidez a las conclusiones derivadas del análisis.

CAPÍTULO 5
CONCLUSIONES

Al cerrar este trabajo, podemos ver que el motor principal de la investigación fue la pregunta de cómo las funciones ejecutivas y algunos datos básicos (como la edad) se relacionan con el desarrollo temprano de competencias matemáticas en niños de 5 a 9 años y, al mismo tiempo, si un modelo de inteligencia artificial podría captar esas relaciones de manera efectiva. En esencia, queríamos construir una herramienta que, a partir de medidas relativamente sencillas —pruebas de memoria de trabajo, de flexibilidad cognitiva, velocidad de procesamiento y dependencia ejecutiva—, junto con información sociodemográfica, fuera capaz de predecir el desempeño en tres indicadores matemáticos (CMLR, CMN y CMG).

Para lograrlo, se reunio 528 registros de niños que pasaron una batería de pruebas cognitivas y una serie de evaluaciones matemáticas estandarizadas. A partir de esa base de datos, desarrollamos un protocolo de regresión multisalida: en lugar de entrenar un único modelo para cada prueba matemática, entrenamos cinco algoritmos de “árboles de decisión ensamble” (Decision Tree, Random Forest, XGBoost, CatBoost y LightGBM) para que predijeran simultáneamente los tres puntajes. Cada modelo se afinó mediante HalvingGridSearchCV y se validó en 50 particiones aleatorias del conjunto de datos, usando el Error Cuadrático Medio (MSE) como criterio de evaluación.

Los resultados fueron claros: CatBoost se destacó como el algoritmo con el menor MSE promedio en las tres salidas, manteniendo además una variabilidad mínima entre ejecuciones. Una vez identificado este “modelo ganador”, utilizamos Permutation Importance para entender qué variables eran realmente decisivas en cada predicción. Descubrimos que, en el corazón de todos los modelos, la Memoria de Trabajo Verbal y la edad eran los factores que más peso tenían para predecir el rendimiento matemático. En segundo orden aparecieron la Flexibilidad Cognitiva y la dependencia ejecutiva.

A partir de ese ranking de importancia, probamos a “acotar” nuestra herramienta: seleccionamos las seis variables más relevantes y volvimos a entrenar CatBoost solo con ellas. El resultado fue muy alentador: el modelo reducido obtuvo un desempeño apenas un 5 % menos preciso que el modelo completo y, al comparar estadísticamente ambos mediante la prueba de Friedman, no se encontraron diferencias significativas ($p > 0{,}05$). Esto confirma que la mayoría de la capacidad predictiva se concentra en esos seis indicadores y abre la posibilidad de diseñar protocolos de detección temprana mucho más simples, rápidos y económicos, sin sacrificar la precisión.

¿Qué significa todo esto para la práctica educativa? Primero, confirma con datos cuantitativos que ciertas funciones ejecutivas —en particular, la Memoria de Trabajo Verbal y la Flexibilidad Cognitiva— actúan como pilares en el desarrollo de habilidades matemáticas tempranas. Esto sugiere que, al momento de diseñar intervenciones o programas de refuerzo en el aula, conviene priorizar actividades que fortalezcan esas capacidades cognitivas. Segundo, al

contar con un modelo “parsimonioso” basado en solo seis variables, se facilita la incorporación de esta herramienta en contextos escolares con recursos limitados: bastaría con aplicar unas pocas pruebas simples y analizar los resultados mediante CatBoost para obtener una alerta temprana sobre niños en riesgo de bajo desempeño.

Obviamente, este estudio tiene limitaciones: el diseño transversal impide establecer con certeza qué es causa y qué es efecto, y los datos provienen de una muestra regional del centro-sur de Chile, por lo que habría que comprobar si los mismos patrones se repiten en otras zonas o culturas. Además, no incorporamos variables socioemocionales o familiares, que seguramente juegan también un papel importante en el aprendizaje matemático. En ese sentido, la siguiente etapa lógica sería un estudio longitudinal que evalúe a estos mismos niños en distintos momentos —por ejemplo, a los 7 años y luego a los 9— para ver si las relaciones se mantienen y, de paso, incorporar mediciones de ansiedad matemática, apoyo familiar o nivel socioeconómico.

En resumen, esta investigación muestra que es posible combinar de manera rigurosa técnicas de inteligencia artificial con mediciones de funciones ejecutivas para obtener un diagnóstico temprano del riesgo matemático en la infancia. Al identificar las variables clave y proponer un modelo simplificado, hemos sentado las bases para que profesores, psicopedagogos y responsables de políticas educativas dispongan de una herramienta práctica. A largo plazo, el objetivo es que este tipo de enfoques no solo describan los patrones de desarrollo, sino que sirvan de guía concreta para diseñar actividades y recursos que ayuden a cada niño a desarrollar al máximo sus competencias matemáticas.

REFERENCIAS BIBLIOGRÁFICAS

[Alicia Risso, 2015] Alicia Risso, Manuel García, M. D. J. C. B. M. P. y. A. B. (2015). Un análisis de las relaciones entre funciones ejecutivas, lenguaje y habilidades matemáticas. 9.

[Antonio Moreno, 1998] Antonio Moreno, Eva Armengol, J. B. L. B. U. C. R. G. J. M. G. B. L. M. M. M. S. (1998). Aprendizaje automático. *Universitat Politècnica de Catalunya*, pp. 1–244.

[Barzizza *et al.*, 2023] Barzizza, G., Pesarin, F., y Salmaso, L. (2023). Multi-aspect permutation tests for model selection. *Expert Systems*, 40(6):e13196.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Chen y Guestrin, 2016] Chen, T. y Guestrin, C. (2016). Xgboost: A scalable tree boosting system. En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM.

[Cortes y Vapnik, 1995] Cortes, C. y Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[Cover y Hart, 1967] Cover, T. M. y Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

[de Educación, 2023] de Educación, M. (2023). Ministro cataldo se reúne con tutores uc y anuncia fortalecimiento de apoyo en matemática. pp. 1–25.

[Dunn, 1961] Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

[Francisca Bernal-Ruiz, 2022] Francisca Bernal-Ruiz, Damián Duarte, F. J. D. M. C. R. E. S. (2022). Memoria de trabajo y planificación como predictores de las competencias matemáticas tempranas. *SUMA PSICOLÓGICA*, 29(2):130–139.

[Francisca Bernal-Ruiz, 2024] Francisca Bernal-Ruiz, G. C. (2024). El efecto de las funciones ejecutivas sobre la competencia matemática temprana: un modelo de ecuaciones estructurales. *Educación XX1*, 27(1):281–301.

[Friedman, 1937] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701.

[Gamal Cerda, 2011] Gamal Cerda, Carlos Pérez, R. O. M. L. y. L. S. (2011). Fortalecimiento de competencias matemáticas tempranas en preescolares, un estudio chileno. 3(1):23–39.

[Gamal Cerda Etchepare, 2014] Gamal Cerda Etchepare, C. P. W. (2014). Competencias matemÁticas tempranas y actitud hacia las tareas matemÁticas variables predictoras del rendimiento acadÉmico en educaciÓn primaria: Resultados preliminares. *INFAD Revista de Psicología,*, 7(1):469–476.

[Geurts *et al.*, 2006] Geurts, P., Ernst, D., y Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

[Hastie *et al.*, 2009] Hastie, T., Tibshirani, R., y Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York.

[Huynh-Thu *et al.*, 2010] Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., y Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776.

[Ke *et al.*, 2017] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., y Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. En *Advances in Neural Information Processing Systems*.

[María-Jesús Presentación, 2015] María-Jesús Presentación, Rebeca Siegenthaler, V. P. J. M. y. A. M. (2015). Competencias matemáticas y funcionamiento ejecutivo en preescolar: evaluación clínica y ecológica. 50(15):65–82.

[Nemenyi, 1963] Nemenyi, P. B. (1963). *Distribution-free multiple comparisons*. Princeton University.

[Orrontia, 2006] Orrontia, J. (2006). Dificultades en el aprendizaje de las matemáticas: Una perspectiva evolutiva. 71:158–180.

[Petković *et al.*, 2020] Petković, M., Kocev, D., y Džeroski, S. (2020). Feature ranking for multi-target regression. *Machine Learning*, 109(6):1179–1204.

[Prokhorenkova *et al.*, 2018] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., y Kuksa, D. (2018). Catboost: unbiased boosting with categorical features. En *Advances in Neural Information Processing Systems*, pp. 6638–6648.

[Quinlan, 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

[Ricardo Rosas, 2020] Ricardo Rosas, V. E. y. M. G. (2020). Evidencia intercultural de un test basado en tablet para medir las funciones ejecutivas de niños entre 6 y 10 años: resultados preliminares. *Centro de Desarrollo de Tecnologías de Inclusión*, pp. 1–25.

[Santana *et al.*, 2021] Santana, L., Fontes, M., Caldeira, E., Souza, A. P., Oliveira, J. P., Palma, M., Chaves, G., Bastos-Filho, C., y Freire, R. (2021). Machine learning classification models for covid-19 test prioritization in brazil. *International Journal of Environmental Research and Public Health*, 18(21):11303.

[Wilcoxon, 1945] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.