1 Universidad Icesi, Facultad Barberi de Ingeniería Diseño y Ciencias Aplicadas - Maestría en Inteligencia Artificial Aplicada (MIAA)

Comparación de Modelos Predictivos para la Detección Temprana del Uso Problemático de Internet en Niños y Adolescentes: Enfoques Basados en Modelos Bayesianos Jerárquicos y Técnicas de Machine Learning.

Daniel Garcia Mendez, William Alberto Reina, Andres Malvehy Velásquez, Julian Tabares Restrepo
1144159574@u.icesi.edu.co, 14476960@u.icesi.edu.co, 1118259851@icesi.edu.co, jtabares@icesi.edu.co

El uso problemático de Internet (UPI) en niños y adolescentes es un desafío creciente con repercusiones negativas en la salud mental, el desempeño académico y las relaciones sociales. Este trabajo presenta el desarrollo de múltiples modelos de clasificación predictivos para la detección temprana del UPI, basado en un conjunto de datos del Child Mind Institute que incluye características demográficas, clínicas y de actividad física. Se implementaron y compararon diversos modelos de clasificación con el fin de predecir la categoría SII.

Todos los enfoques permitieron capturar la heterogeneidad inherente a la población estudiada y ofreció una exactitud superior al 50% en las predicciones. Se aplicaron técnicas de imputación mediante KNN para el tratamiento de valores faltantes, llenando registros nulos y mejorando la calidad de los datos para el modelado. Los resultados indicaron que, aunque modelos tradicionales como Random Forest y redes neuronales alcanzaron precisiones competitivas, el modelo bayesiano proporcionó ventajas en interpretabilidad y manejo de datos desbalanceados. No obstante, las limitaciones en los recursos computacionales y en la cantidad de muestras para ciertos grupos señalaron áreas de mejora para futuros trabajos, incluyendo la validación con muestras más diversas y la integración de variables adicionales. Este estudio contribuye con una metodología innovadora para apoyar la detección clínica y educativa del UPI, favoreciendo intervenciones preventivas más efectivas en la población juvenil.

Problematic Internet Use, Bayesian Hierarchical Model, Machine Learning Classification, Early Detection, Adolescent Mental Health

I. INTRODUCCIÓN

El uso problemático de Internet (**UPI**) en niños y adolescentes constituye una problemática creciente a nivel mundial, con implicaciones significativas y notorias en la salud mental y el desarrollo psicosocial de esta generación. El UPI se caracteriza por una dependencia al uso de internet, que lleva a la pérdida de control sobre el tiempo dedicado a la conexión en línea y el dedicado a actividades fuera de línea, que provoca un deterioro notable en actividades diarias esenciales como el estudio, el deporte, el descanso y las relaciones sociales, así como en el acondicionamiento físico y mental de los jóvenes. En la era digital contemporánea, estos efectos se ven exacerbados por el diseño adictivo de las plataformas digitales, que utilizan mecanismos psicológicos de recompensas y notificaciones constantes para maximizar la atención del usuario, además de la presión social inherente para mantener una presencia activa en línea.

Esta problemática tiene consecuencias adversas ampliamente documentadas, incluyendo el aumento de trastornos mentales como depresión, ansiedad y trastornos del sueño; efectos negativos en el rendimiento académico; y deterioro en las relaciones sociales y familiares, afectando así la calidad de vida de los jóvenes afectados [1]. Frente a este escenario, la detección temprana del UPI se vuelve crucial para implementar intervenciones oportunas que mitiguen estos impactos y promuevan hábitos digitales saludables.

El presente proyecto se enfoca en el desarrollo de un modelo predictivo de clasificación capaz de identificar el riesgo de UPI en población infantil y adolescente, utilizando datos longitudinales que incluyen variables demográficas, atributos clínicos y registros de actividad física, entre otros. Este enfoque posibilita no solo la detección temprana basada en patrones observables, sino también la consideración de la heterogeneidad inherente a distintos grupos etarios mediante técnicas avanzadas de modelado estadístico y aprendizaje automático [2].

La fundamentación teórica y metodológica se apoya en múltiples estudios recientes que utilizan técnicas de machine learning para evaluar factores psicológicos y conductuales asociados al UPI. Por ejemplo, investigaciones como la de Jović & Ćorac [3] han identificado variables críticas relacionadas con temperamentos afectivos y patrones de uso digital que contribuyen al desarrollo del UPI. Asimismo, enfoques

bayesianos jerárquicos y redes neuronales profundas han mostrado eficacia en la modelización de datos complejos y heterogéneos propios de esta área de estudio, permitiendo una mejor interpretación y cuantificación de la incertidumbre en las predicciones [4] [5] [6].

El objetivo principal de este trabajo es alcanzar una precisión superior al 50% en la clasificación del riesgo de UPI, maximizando la capacidad de generalización del modelo para su aplicación en entornos clínicos y educativos. Para ello, se ha considerado un pipeline integral que abarca desde el preprocesamiento y análisis exploratorio de datos hasta la comparación y optimización de diversos algoritmos de clasificación, incluyendo Random Forest, Support Vector Machines, redes neuronales y modelos bayesianos jerárquicos.

Este artículo presenta asimismo un análisis profundo de los resultados obtenidos, discutidos en relación con la literatura citada, y expone las implicaciones prácticas del modelo como una herramienta de apoyo para la detección temprana del uso problemático de Internet en jóvenes.

II. MATERIALES Y MÉTODOS

A. Dataset y Fuente de datos

El conjunto de datos utilizado en este estudio proviene del Child Mind Institute [7] y está orientado a la investigación del uso problemático de Internet (UPI) en población infantil y adolescente. El dataset es de naturaleza longitudinal, recopilando múltiples evaluaciones por participante a lo largo del tiempo. Contiene variables demográficas, evaluaciones clínicas, datos de actividad física y patrones de uso digital, entre otros. La variable objetivo principal, denominada SII (Smartphone Internet Addiction Indicator), indica de forma binaria la presencia o ausencia de UPI.

B. Preprocesamiento y Limpieza de Datos

El dataset dispone de problemas notorios en cuestión de calidad de datos. En primera instancia, se recurrió a eliminar los registros cuyo registro en la variable objetivo fuera nulo. La imputación de valores faltantes se realizó mediante el algoritmo K-Nearest Neighbors (KNN), el cual es una técnica ampliamente utilizada en aprendizaje automático debido a su capacidad para preservar la distribución y estructura multivariada de los datos. Este método consiste en identificar, para cada instancia con datos faltantes, sus k vecinos más próximos en el espacio de características y estimar el valor faltante mediante un promedio ponderado de dichos vecinos. La ventaja principal de KNN imputation radica en su simplicidad y eficacia en datasets con múltiples variables correlacionadas, evitando sesgos significativos en comparación con técnicas más básicas como la imputación por la media o la mediana. [8]

C. Análisis Exploratorio de Datos

Se realizó un análisis exploratorio exhaustivo empleando estadísticas descriptivas y visualización gráfica para caracterizar la distribución y comportamiento de las variables. Además, se calcularon medidas de correlación y se evaluó la información mutua entre características y la variable objetivo, con el fin de identificar variables predictoras potencialmente relevantes para la clasificación del UPI y descartar aquellas que no estén correlacionadas con la variable objetivo, evitando así el problema de la multicolinealidad y la dimensionalidad elevada.

D. Selección y Configuración de Modelos

Se seleccionaron varios modelos de clasificación para comparar su desempeño y determinar el más adecuado para la problemática planteada. Los modelos evaluados incluyeron:

1. **Random Forest**, reconocido por su robustez y capacidad de manejar datos heterogéneos.
2. **Máquinas de Soporte Vectorial (SVM)**, con kernels lineales y no lineales para capturar relaciones complejas.
3. **Redes Neuronales Artificiales**, configuradas para aprendizaje supervisado en contextos multivariados.
4. **Modelo Bayesiano Jerárquico Multinomial**, que incorpora una estructura jerárquica basada en grupos de edad para capturar la heterogeneidad y permitir la cuantificación completa de incertidumbre.

E. Entrenamiento, Validación y Optimización

Los datos fueron divididos en subconjuntos para entrenamiento, validación y prueba, asegurando la representatividad de cada grupo. Se empleó validación cruzada para estimar el desempeño general de los modelos y se aplicó búsqueda en cuadrícula (*Grid Search*) para la optimización fina de hiperparámetros, con el

fin de maximizar la precisión y capacidad de generalización. Para el modelo bayesiano, se evaluó la convergencia del muestreo MCMC mediante estadísticas específicas (e.g., $\hat{R} < 1.1$) para validar la estabilidad de las inferencias.

F. Métricas de Evaluación

El desempeño de los modelos fue cuantificado mediante métricas clásicas de clasificación, incluyendo accuracy, precisión, recall y F1-score. Estas métricas permitieron evaluar tanto la capacidad predictiva global como el balance entre sensibilidad y especificidad, relevando así la eficacia del modelo para la detección temprana del UPI.

G. Recursos Computacionales

El análisis se realizó utilizando entornos de programación en Python, con bibliotecas especializadas tales como pandas, scikit-learn para modelado clásico y PyMC para el desarrollo del modelo bayesiano jerárquico. Se empleó hardware con capacidad suficiente para soportar el entrenamiento y muestreo iterativo, dado el costo computacional asociado especialmente con el modelo bayesiano.

III. RESULTADOS

En esta sección se presentan los resultados derivados del entrenamiento, validación y evaluación de los modelos predictivos aplicados al conjunto de datos sobre uso problemático de Internet (UPI). La comparación entre los diferentes enfoques utilizados permite identificar el modelo con mejor desempeño para la detección temprana del UPI.

A.Importancia de las variables predictoras

Con el fin de seleccionar la cantidad de características adecuadas, para evitar multicolinealidad y características redundantes, se evaluaron medidas de información mutua y correlación. Se encontró que un grupo de características estabas correlacionadas entre sí dado que eran medidas derivadas.

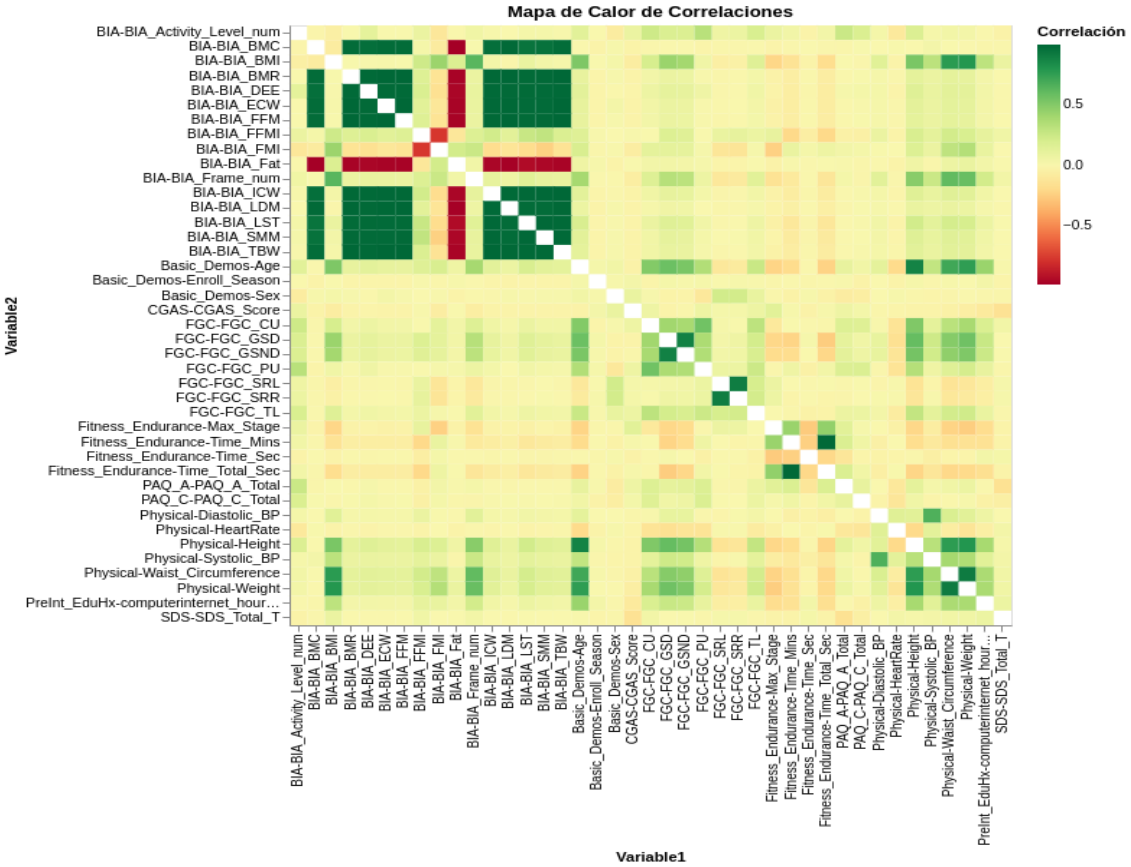

Ilustración 1 - Mapa de calor de la matriz de correlaciones entre las características

Por otro lado, se encontró que la variable con mayor correlación con la variable objetivo fue Basic_Demos-Age. El resto de variables también se usaron para el entrenamiento.

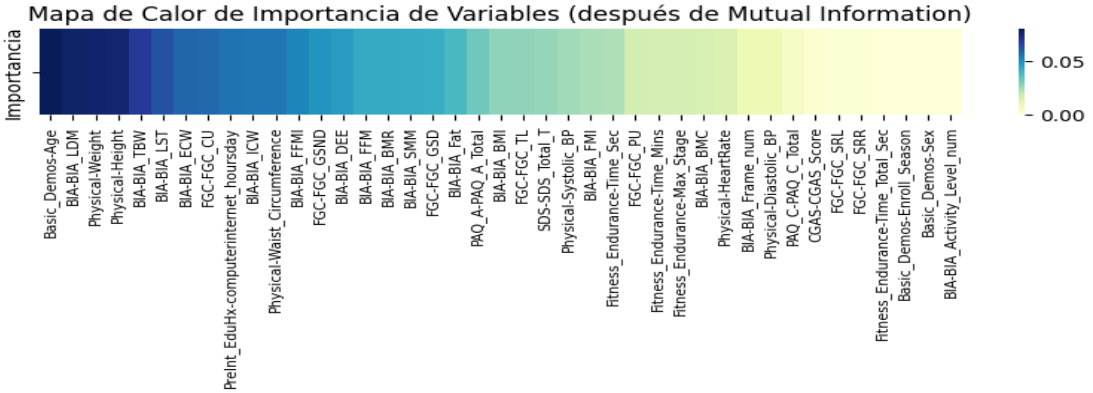

Ilustración 2 - Mapa de calor de la información mutua entre cada una de las características y la etiqueta

B. Descripción de los modelos empleados

Se evaluaron múltiples algoritmos de clasificación, incluyendo Random Forest, Máquinas de Soporte Vectorial (SVM), Redes Neuronales Artificiales y un modelo bayesiano jerárquico multinomial. La métrica por optimizar fue la exactitud, y se obtuvo realizando una validación cruzada de 5 *folds* para cada modelo. A continuación, se describirán los ajustes empleados para el entrenamiento de cada modelo.

1) Red Neuronal

Se entrenó un conjunto de 24 configuraciones de hiperparámetros, donde la rejilla de opciones a utilizar fue la siguiente:

	Opciones			
Arquitectura de las capas ocultas	100	100, 100	100, 100, 100	256, 128, 64, 32
Función de activación	ReLU		Logística	
Optimizador	SGD			
Taza de aprendizaje	0.0001	0.001		0.01

Tabla 1 - Rejilla de configuraciones para el modelo de redes neuronales

El máximo de iteraciones por cada configuración fue de 800. La mejor exactitud para el modelo de redes neuronales, encontrada dentro de este marco de búsqueda fue de 59.53%, bajo la siguiente configuración:

- Arquitectura: Una única capa oculta de 100 neuronas.
- Función de activación: Logística
- Tasa de aprendizaje: 0.01

2) Máquinas de Soporte Vectorial

Se entrenó un conjunto de 36 configuraciones de hiperparámetros, siendo la rejilla de configuraciones la siguiente:

	Opciones			
Constante de regularización	0.1	1	10	100
Tipo de kernel	Lineal	Radial	Polinomial	
Gamma (para kernels radiales y polinomiales)	Escalable			
Grado (para kernels polinomiales)	2	3	4	-

Tabla 2 - Rejilla de configuraciones para el modelo de máquinas de soporte vectorial

Bajo esta rejilla de hiperpaŕametros, la mejor exactitud encontrada fue de 59.17%, bajo la siguiente configuración:

- Constante de regularización: 10
- Kernel: Lineal

3) Random Forest

Se entrenó un conjunto de 216 configuraciones de hiperparámetros bajo la siguiente rejilla de configuraciones:

	Opciones			
Número de árboles por bosque	100	200	300	-
Profundidad máxima	Ilimitada	10	20	30
Mínimo de muestras para dividir el árbol	2	5	10	-
Mínimo de muestras en una hoja	1	2	4	-
Boostrap	Verdadero		Falso	

Tabla 3 - Rejilla de configuraciones para el modelo de random forest

Bajo la siguiente tabla de configuraciones, se encontró que la mejor exactitud obtenida fue de 60.89%, bajo la siguiente configuración:

- Número de árboles por bosque: 100
- Profundidad máxima: 10
- Mínimo de muestras para dividir el árbol: 2
- Mínimo de muestras en una hoja: 4

C. Análisis por Grupos de Edad

Para capturar la heterogeneidad inherente en la población infantil y adolescente, se implementó un modelo bayesiano jerárquico multinomial que incorpora explícitamente la estructura de grupos etarios como niveles jerárquicos. Este enfoque permite modelar las diferencias sistemáticas entre subpoblaciones definidas por rangos de edad, proporcionando estimaciones específicas para cada grupo, a la vez que aprovecha el "partial pooling" para compartir información entre ellos.

La configuración del modelo consideró:

- **Nivel jerárquico individual**: Cada grupo de edad fue tratado como un estrato con parámetros propios, lo que habilita la estimación de efectos específicos que reflejan las particularidades de cada segmento poblacional.
- **Regularización automática**: Mediante priors jerárquicos robustos y técnicas de shrinkage, el modelo controla la variabilidad excesiva en grupos pequeños o con escasos datos, favoreciendo estimaciones estables y generalizables.
- **Parámetros y estructura**: El modelo fue formulado como un modelo multinomial jerárquico, con una estructura probabilística que contempla la dependencia condicionada entre observaciones dentro de cada grupo y la presencia de efectos poblacionales globales.
- **Inferencia**: Se realizó inferencia bayesiana utilizando muestreo Markov Chain Monte Carlo (MCMC), con diagnósticos rigurosos para asegurar la convergencia ($\hat{R} < 1.1$). La configuración del muestreo incluyó múltiples cadenas paralelas y una cantidad adecuada de iteraciones para garantizar la estabilidad de las estimaciones.
- **Evaluación por grupos de edad**: El desempeño predictivo fue desagregado por grupos, permitiendo identificar las variaciones en la capacidad de clasificación entre estratos. Esta evaluación facilitó un análisis detallado de la efectividad del modelo y la identificación de posibles ajustes focalizados en subpoblaciones específicas.

La implementación se llevó a cabo empleando el framework PyMC, que soporta de manera eficiente el modelado probabilístico jerárquico con capacidades avanzadas de diagnóstico y visualización.

Este análisis jerárquico revela que el grupo más joven (Grupo 0) presenta la mayor exactitud, mientras que los grupos de edades superiores muestran una caída progresiva en la precisión, lo que sugiere un desbalance de las clases a predecir en los datos.

	Precisión	Sensibilidad	F1-Score
SII 0	62%	92%	74%
SII 1	36%	18%	24%
SII 2	32%	2%	4%
SII 3	0%	0%	0%

Tabla 4 - Tabla de métricas por grupo SII del modelo bayesiano jerárquico

En la gráfica a continuación se observan la matriz de confusión para la predicción de cada una de las categorías del ejercicio. Se observa que, a pesar de que las predicciones verdaderas para cada categoría son altas, para todas las categorías se predice erróneamente la categoría 1 con una frecuencia similar entre clases. Esto puede sugerir que es requerido usar otras técnicas de aprendizaje automático que permitan crear categorías de manera más fina.

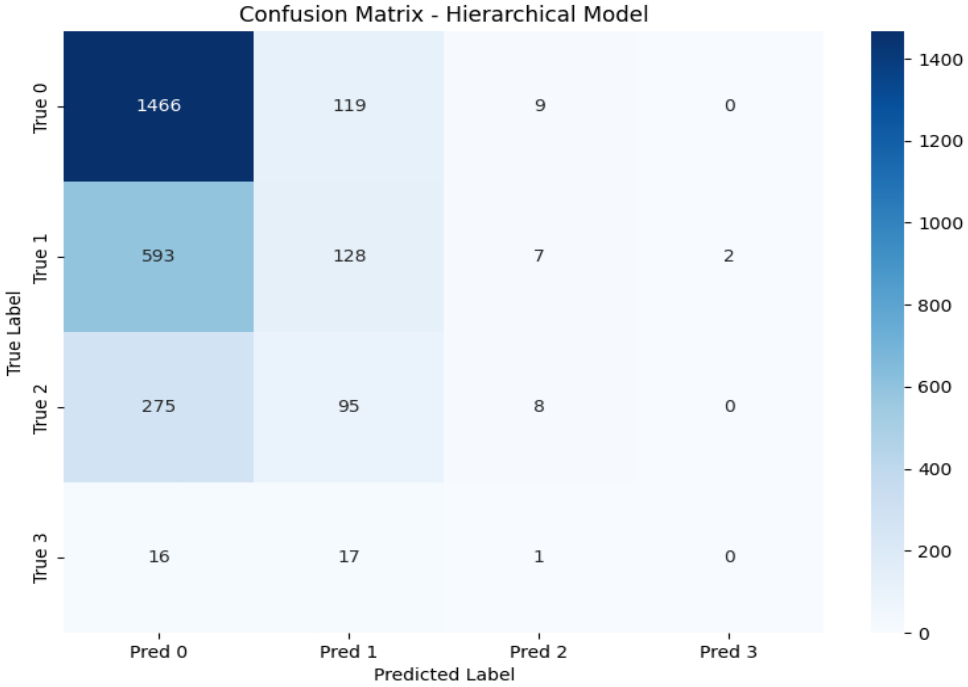

Ilustración 3 - Matriz de confusión para el modelo bayesiano jerárquico

D. Análisis de resultados

La siguiente tabla presenta un resumen comparativo de estas métricas para cada modelo. Se observa una exactitud similar para cada uno de los modelos empleados, sugiriendo un máximo para la exactitud bajo la cantidad y distribución de los datos dispuestos por el dataset.

Método Empleado	Mejor exactitud obtenida
Red Neuronal	59.53%
Máquina de Soporte Vectorial	59.17%
Random Forest	60.89%
Modelo Bayesiano Jerárquico	59%

Tabla 5 - Resumen de las mejores exactitudes obtenidas para cada modelo.

IV. Discusión

Los resultados obtenidos en este estudio evidencian que el modelo bayesiano jerárquico multinomial constituye un enfoque prometedor para la clasificación del riesgo de uso problemático de Internet (UPI) en población infantil y adolescente. La incorporación explícita de los grupos de edad como niveles jerárquicos permitió capturar de manera efectiva la heterogeneidad inherente a las diferentes etapas del desarrollo, reflejándose en variaciones significativas del desempeño predictivo entre subgrupos poblacionales.

En comparación con modelos tradicionales como Random Forest, redes neuronales y máquinas de soporte vectorial, el modelo bayesiano ofrece la ventaja adicional de una cuantificación completa de la incertidumbre, así como una interpretabilidad mejorada a través del análisis jerárquico de parámetros. Estos beneficios coinciden con hallazgos reportados en la literatura reciente, donde se destaca la utilidad de los modelos probabilísticos para manejar datos estructurados y parcialmente observados en contextos clínicos y psicosociales [3].

Desde una perspectiva práctica, este enfoque permite no solo realizar predicciones individuales más ajustadas al perfil demográfico, sino también interpretar efectos poblacionales globales, lo cual resulta valioso para diseñar estrategias de intervención y prevención más focalizadas según el grupo etario. Además, la capacidad del modelo para manejar grupos con tamaños desbalanceados mediante regularización jerárquica es particularmente relevante en estudios longitudinales con muestras heterogéneas y presencia de datos faltantes.

Sin embargo, este estudio presenta ciertas limitaciones que deben considerarse. En primer lugar, la dependencia del modelo bayesiano de un proceso computacionalmente intensivo, derivado del muestreo MCMC, implica mayores tiempos de entrenamiento y requerimientos técnicos, lo que puede limitar su aplicación en escenarios de producción con restricciones de recursos. Asimismo, la variabilidad observada en la precisión entre grupos sugiere la necesidad de explorar estrategias complementarias, como la incorporación de variables adicionales o metodologías híbridas para mejorar la generalización en subpoblaciones menos representadas.

Finalmente, dado que los datos provienen de un contexto específico y presentan características longitudinales con posibles sesgos de selección y medición, la extrapolación de resultados a otras poblaciones debe realizarse con cautela, recomendando validaciones externas y ampliaciones en la diversidad muestral.

V. Conclusiones

En este estudio se desarrolló y evaluó un modelo predictivo para la detección temprana del uso problemático de Internet (UPI) en población infantil y adolescente, empleando un enfoque de clasificación basado en modelos bayesianos jerárquicos y comparándolo con métodos tradicionales como Random Forest, SVM y redes neuronales. Los resultados demostraron que, aunque los modelos clásicos presentan una precisión competitiva, el modelo bayesiano jerárquico aporta un valor diferencial al incorporar la heterogeneidad demográfica mediante su estructura jerárquica y al proporcionar una cuantificación explícita de la incertidumbre.

El análisis segmentado por grupos de edad evidenció variaciones significativas en el desempeño del modelo, destacando la importancia de adaptar las técnicas predictivas a subpoblaciones específicas para mejorar la precisión y relevancia clínica. Asimismo, la aplicación de técnicas avanzadas de imputación y preprocesamiento permitió manejar de forma apropiada las limitaciones del dataset longitudinal y heterogéneo, garantizando la robustez en las inferencias.

Sin embargo, se identificaron limitaciones relacionadas con el costo computacional elevado del modelo bayesiano y la variabilidad en el desempeño para grupos menos representados, lo cual sugiere la necesidad de futuras investigaciones que exploren modelos híbridos o estrategias de enriquecimiento de datos. De igual manera, la validez externa del modelo debe ser confirmada mediante estudios con muestras más diversas.

Este trabajo contribuye significativamente al campo de la salud digital y la psicología clínica juvenil, ofreciendo una herramienta metodológica prometedora que puede ser implementada para apoyar la detección y prevención temprana del UPI en contextos educativos y clínicos. Se recomienda continuar con el desarrollo y validación del modelo, así como la incorporación de variables complementarias para mejorar su capacidad predictiva y utilidad práctica.

APÉNDICE

A. Lista de ilustraciones

B. Lista de tablas

RECONOCIMIENTO

Queremos expresar nuestro más sincero agradecimiento a todas las personas e instituciones que hicieron posible la realización de este trabajo.

En primer lugar, agradecemos al Dr. Milton Sarria, por su valiosa orientación, paciencia y acompañamiento durante todo el proceso, su experiencia y consejos fueron fundamentales para el desarrollo de este proyecto.

Agradecemos también a la Universidad Icesi, por brindar el espacio, los recursos y el respaldo necesarios para llevar a cabo esta investigación.

Finalmente, gracias a todas las personas que, de una u otra manera, contribuyeron a la culminación exitosa de este informe.

REFERENCIAS

[1] A. Z. M. L. L. L. M. D. Y. K. G. K. Adam Santorelli, «Problematic Internet Use,» 2024. [En línea]. Available: https://www.kaggle.com/competitions/child-mind-institute-problematic-internet-use. [Último acceso: 20 Junio 2025].

[2] A. G. Anita Rácz, «Comparison of missing value imputation tools for machine learning models based on product development cases studies,» LWT, 2024.

[3] J. M. P. W. K. T. P. E. Andrianie P, «The Role of Peer Relationships in Problematic Internet Use among Adolescents: A Scoping Review and Meta-analysis,» Open Psychology Journal, 2024.

[4] Ć. A. S. A. N. M. S. M. B. Z. I. R. D. Jović J, «Using machine learning algorithms and techniques for defining the impact of affective temperament types, content search and activities on the internet on the development of problematic internet use in adolescents' population,» Front Public Health, 2024.

[5] E. G. A. Probierz, Problematic Use of the Internet - Using Machine Learning in a Prevention Programme, Springer, 2020.

[6] L.-F. O. Kuss DJ, «Internet addiction and problematic Internet use: A systematic review of clinical research,» World J Psychiatry, 2016.

[7] H. C. L. V. T. Y. H. J. Pino MJ, «A study of impulsivity as a predictor of problematic internet use in university students with disabilities,» Fron Psychiatry, 2024.