Proyecto de semestre. Entrega 1

Autor:

Felipe De Jesus Correa Londoño

Facultad de Ingeniería Universidad de Antioquia

Fundamentos de Deep Learning

Raul Ramos Pollán

Medellín, 2025

Contexto de aplicación.

En la era digital actual, el uso problemático de internet en niños y adolescentes representa una preocupación creciente en materia de salud pública, dado su vínculo con trastornos mentales como la depresión, la ansiedad y las dificultades en el desarrollo social y emocional. A pesar de la importancia de detectar este tipo de conductas de manera temprana, las herramientas clínicas disponibles actualmente son complejas, costosas y requieren la intervención de profesionales altamente capacitados. Esta situación genera barreras de acceso importantes, especialmente en comunidades donde existen limitaciones culturales, económicas o lingüísticas, afectando la posibilidad de realizar diagnósticos oportunos.

Por otro lado, las mediciones relacionadas con la actividad física y el estado general de forma física resultan ser alternativas mucho más accesibles y fáciles de obtener, ya que no requieren intervención clínica especializada. La literatura científica señala que ciertos cambios en los hábitos físicos —como la disminución en los niveles de actividad, alteraciones en los patrones de sueño, irregularidades en la alimentación y deterioro postural— suelen ser manifestaciones indirectas en personas que hacen un uso excesivo de tecnologías digitales.

Con base en estas observaciones, el Child Mind Institute, en conjunto con la iniciativa Healthy Brain Network (HBN), ha impulsado un proyecto que busca desarrollar un modelo de aprendizaje automático capaz de predecir los niveles de uso problemático de internet a partir de datos de actividad física recogidos en niños y adolescentes. La finalidad de esta propuesta es identificar de manera temprana los patrones de riesgo, permitiendo con ello activar intervenciones oportunas que fomenten hábitos de consumo tecnológico más saludables y que contribuyan a la prevención de trastornos asociados.

La participación en este proyecto se da en el marco de una competencia organizada a través de la plataforma Kaggle, con el patrocinio de Dell Technologies y NVIDIA, y con apoyo financiero del Departamento de Servicios de Salud de California (DHCS) mediante la iniciativa Children and Youth Behavioral Health Initiative (CYBHI). Dentro de la competencia, el desafío consiste en construir un modelo predictivo eficiente que analice registros de actividad física y sea capaz de clasificar el nivel de uso problemático de internet, utilizando como métrica de evaluación el kappa cuadrático ponderado, que mide la concordancia entre las predicciones del modelo y las etiquetas reales.

Desde una perspectiva académica, la correcta detección de señales físicas asociadas al uso excesivo de internet no solo puede mejorar el bienestar de las nuevas generaciones, sino también fortalecer el conocimiento científico acerca de la relación entre el comportamiento físico y el entorno digital. Además, abre la posibilidad de

implementar soluciones de bajo costo en contextos donde las evaluaciones clínicas tradicionales no son factibles. A largo plazo, considero que este tipo de iniciativas son fundamentales para avanzar hacia un futuro más saludable, equitativo e inclusivo, donde los niños y adolescentes puedan desenvolverse de manera equilibrada en un mundo digital cada vez más complejo.

Objetivo de machine learning

Queremos predecir el nivel de uso problemático de internet en niños y adolescentes, medido mediante el índice "Severity Impairment Index" (SII), que clasifica el nivel de severidad en cuatro categorías ordinales:

0: Ninguno,

1: Leve,

2: Moderado,

3: Severo.

Esta predicción se realizará dada la información disponible sobre los participantes, que incluye datos de actividad física (como pasos diarios, intensidad de la actividad, patrones de movimiento), mediciones corporales (como peso, estatura y composición corporal) y características demográficas complementarias (como edad y sexo).

El objetivo principal es construir un modelo predictivo que, a partir de estas señales físicas y demográficas, sea capaz de identificar tempranamente patrones asociados al uso excesivo de internet y tecnología. De este modo, se busca ofrecer una herramienta accesible y de bajo costo para anticipar riesgos en contextos donde los métodos tradicionales de evaluación clínica no son viables.

Queremos encontrar el valor de SII (0, 1, 2, 3) dado un conjunto de variables de actividad física, condición física y datos demográficos, optimizando la clasificación mediante la métrica de kappa cuadrático ponderado, que mide la concordancia entre las predicciones y los valores reales.

Dataset: tipo de datos, tamaño (número de datos y tamaño en disco), distribución de las clases

Origen: Estudio Healthy Brain Network (HBN).

Formato: Archivos CSV y Parquet.

Tamaño:

- **Datos Tabulares:** Archivos en formato CSV para entrenamiento y prueba.

- **Datos de Series Temporales:** Acelerometría de muñeca en formato Parquet, por participante.
- **Cantidad de Archivos:** Aproximadamente 1002 archivos.
- **Tamaño Total:** Aproximadamente 6.73 GB.

Descripción de los Datos:

1. **Datos Demográficos:** Información sobre la edad, sexo y estación del año de la evaluación de cada participante.
2. **Actividad Física:** Datos obtenidos de un acelerómetro de muñeca, incluyendo las mediciones en los ejes X, Y, Z, ENMO (Euclidean Norm Minus One), y el Ángulo-Z, para capturar la actividad física de los participantes.
3. **Estado Físico:** Información relacionada con la salud, incluyendo la presión arterial, altura, peso, porcentaje de grasa corporal, y datos derivados de un análisis de bioimpedancia, entre otros indicadores de salud física.
4. **Uso de Internet:** Información sobre las horas de uso diario de computadoras o internet, junto con los resultados del *Parent-Child Internet Addiction Test* (PCIAT), una evaluación de la adicción a Internet.
5. **Sueño:** Escalas para categorizar los trastornos del sueño en los participantes.
6. **Evaluaciones Clínicas:** Incluye medidas del *Children's Global Assessment Scale* (CGAS), una evaluación que mide el funcionamiento general de los niños.

Distribución de Clases: El objetivo de este conjunto de datos es predecir el Índice de Severidad de Impairment (SII), una medida del uso problemático de internet. Los valores del SII están categorizados como sigue:

- **0:** Ninguno
- **1:** Leve
- **2:** Moderado
- **3:** Severo

Aunque la distribución exacta no ha sido publicada, se espera que haya una sobrerrepresentación de casos leves a moderados, ya que los participantes provienen de una muestra clínica.

Observaciones Importantes:

- **Datos Faltantes:** Hay una tasa considerable de valores faltantes en algunas mediciones, lo que puede requerir técnicas de manejo de datos ausentes.
- **Series Temporales:** Cada participante tiene registros continuos de actividad diaria obtenidos con un acelerómetro, con intervalos de tiempo de 5 segundos.

Estos datos incluyen el **timestamp**, las medidas de aceleración (ejes X, Y, Z), y otros indicadores como la **ENMO**, **Ángulo-Z**, y **flag de no uso** del dispositivo.

Archivos Disponibles:

- **train.csv y test.csv:** Archivos con datos tabulares para entrenamiento y prueba.
- **series_train.parquet y series_test.parquet:** Archivos con los datos de acelerometría en formato Parquet para cada participante.
- **data_dictionary.csv:** Diccionario de los campos y variables en el dataset.
- **sample_submission.csv:** Formato de ejemplo para las entregas de la competencia.

Propósito: El objetivo del concurso es predecir el Índice de Severidad de Impairment (SII) a partir de los datos, lo que ayuda a evaluar el impacto del uso de internet en la salud mental y el comportamiento de los participantes.

Métricas de desempeño (de machine learning y negocio)

Las presentaciones se puntúan utilizando el **kappa ponderado cuadrático**, una métrica que mide el grado de acuerdo entre dos resultados. Esta métrica típicamente varía de 0 (acuerdo aleatorio) a 1 (acuerdo total). En caso de que haya menos acuerdo del esperado por azar, la métrica puede caer por debajo de 0.

Para calcular el **kappa ponderado cuadrático**, se construyen tres matrices: **O**, **W** y **E**, donde **N** es el número de etiquetas distintas.

- La **matriz O** es una matriz de histograma de **N × N**, en la cual **O(i,j)** corresponde al número de instancias que tienen un valor real **i** y un valor predicho **j**.
- La **matriz W** es una matriz de pesos de **N × N**, calculada con base en la diferencia al cuadrado entre los valores reales y predichos:

$$W(i,j) = ((i - j)^2) / (N - 1)^2$$

- La **matriz E** es una matriz de histograma de **N × N** de resultados esperados, calculada bajo la suposición de que no existe correlación entre los valores. Se calcula como el producto externo entre el vector de histograma de los resultados reales y el vector de histograma de los resultados predichos, normalizado de tal manera que las sumas de **E** y **O** sean iguales.

A partir de estas tres matrices, el **kappa ponderado cuadrático** se calcula con la siguiente fórmula:

$$\kappa = 1 - \Sigma(i,j)\,[W(i,j) * O(i,j)] \,/\, \Sigma(i,j)\,[W(i,j) * E(i,j)]$$

Referencias y resultados previos

- **Healthy Brain Network (HBN):** Estudio longitudinal de salud mental infantil impulsado por el Child Mind Institute.
- **Investigaciones recientes:** Estudios previos muestran correlaciones entre menor actividad física y mayor uso problemático de dispositivos electrónicos.
- **Resultados esperados:** Modelos anteriores en problemas similares han logrado $\kappa \approx 0.6$ usando solamente características físicas básicas; al incluir series de acelerometría, se espera superar $\kappa > 0.7$.
- **Métodos exitosos previos:**
    - Modelos basados en **XGBoost** y **LightGBM** para datos tabulares.
    - **Redes neuronales recurrentes (RNN, LSTM)** para series de acelerometría.
    - **Modelos híbridos:** extracción de features estadísticos de acelerometría + clasificación tabular.

Metadata del dataset

Instrument

Parent-Child Internet Addiction Test	27%
Bio-electric Impedance Analysis	21%
Other (42)	52%

Valid	81	100%
Mismatched	0	0%
Missing	0	0%
Unique	12	
Most Common	Parent-Chil...	27%

Field

81
unique values

Valid	81	100%
Mismatched	0	0%
Missing	0	0%
Unique	81	
Most Common	id	1%

Description

Season of participation	12%
Participant's ID	1%
Other (70)	86%

Valid	81	100%
Mismatched	0	0%
Missing	0	0%
Unique	72	
Most Common	Season of ...	12%

Type

categorical int	38%
float	30%
Other (26)	32%

Valid	81	100%
Mismatched	0	0%
Missing	0	0%
Unique	4	
Most Common	categorical...	38%

Values

[null]	48%
0,1,2,3,4,5	25%
Other (22)	27%

Valid	42	52%
Mismatched	0	0%
Missing	39	48%
Unique	6	
Most Common	0,1,2,3,4,5	25%

Value Labels

[null]	60%
0=Does Not Apply, 1=Rarely, 2=Occa...	25%
Other (12)	15%

Valid	32	40%
Mismatched	0	0%
Missing	49	60%
Unique	8	
Most Common	0=Does No...	25%