Informe Final del Proyecto: Predicción del Nivel de Uso Problemático de Internet en Niños y Adolescentes

Autor: Felipe De Jesus Correa Londoño
Facultad de Ingeniería, Universidad de Antioquia
Curso: Fundamentos de Deep Learning
Profesor: Raul Ramos Pollán
Medellín, 2025

Resumen Ejecutivo

El presente informe detalla el desarrollo de un modelo de Machine Learning para predecir el **Uso Problemático de Internet (UPI)** en niños y adolescentes, utilizando datos de actividad física, estado físico y características demográficas. Ante la complejidad y las barreras de acceso de las evaluaciones clínicas tradicionales para el UPI, este proyecto propone una solución basada en datos accesibles. Se abordaron desafíos clave como la detección y mitigación de fugas de datos, el manejo de valores nulos y el desequilibrio de clases. El modelo, un **RandomForestClassifier** entrenado a través de un pipeline robusto, demostró una precisión promedio del 56.6% en validación cruzada, mostrando su capacidad para identificar patrones asociados al UPI y ofreciendo una herramienta potencial para intervenciones tempranas.

1. Introducción y Contexto del Problema

El uso problemático de internet en niños y adolescentes representa una preocupación creciente en la salud pública debido a su vínculo con trastornos mentales como la depresión, la ansiedad y dificultades en el desarrollo socioemocional. Las herramientas clínicas actuales para su detección son complejas, costosas y requieren profesionales especializados, lo que genera barreras de acceso significativas.

En contraste, las mediciones de actividad física y el estado físico general son más accesibles. La literatura científica sugiere que cambios en hábitos físicos (ej., disminución de actividad, alteraciones del sueño, deterioro postural) suelen ser manifestaciones indirectas de un uso excesivo de tecnologías digitales.

Basado en estas observaciones, el Child Mind Institute, junto con la iniciativa Healthy Brain Network (HBN), ha impulsado un proyecto para desarrollar un modelo de aprendizaje automático que prediga los niveles de UPI a partir de datos de actividad física en niños y adolescentes. La finalidad es identificar tempranamente patrones de riesgo para activar intervenciones que fomenten hábitos de consumo tecnológico más saludables. Este proyecto se enmarca en una competición organizada en Kaggle.

2. Objetivo del Proyecto

El objetivo principal de este proyecto de Machine Learning es predecir el nivel de uso problemático de internet, medido a través del **Índice de Severidad de Deterioro (SII)**, clasificado en cuatro categorías ordinales:

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

Esta predicción se realiza utilizando información disponible sobre los participantes, incluyendo datos de actividad física (como pasos diarios), mediciones corporales (como peso, estatura, composición corporal) y características demográficas (edad, sexo). Se busca ofrecer una herramienta accesible y de bajo costo para anticipar riesgos donde los métodos tradicionales no son viables.

La métrica de evaluación principal definida para la competición es el **kappa cuadrático ponderado**, que mide la concordancia entre las predicciones del modelo y las etiquetas reales, penalizando los errores de clasificación más grandes.

3. Descripción del Dataset

El dataset es un estudio longitudinal de salud mental infantil que proviene del **Healthy Brain Network (HBN)** y se proporciona en formato CSV y Parquet. Los archivos principales incluyen train.csv, test.csv (datos tabulares), series_train.parquet, series_test.parquet (datos de acelerometría de muñeca) y data_dictionary.csv.

El conjunto de entrenamiento tabular (train.csv) contenía inicialmente 3960 observaciones y 82 columnas. Tras la limpieza inicial de filas sin valor de sii (aproximadamente 1224 filas), el conjunto de entrenamiento etiquetado se redujo a **2736 observaciones**. El conjunto de prueba (test.csv) tiene 20 observaciones y 59 columnas.

Las categorías de datos incluidas son:

1. **Datos Demográficos:** Edad, sexo, estación del año de la evaluación.
2. **Actividad Física:** Mediciones de acelerómetro de muñeca (ejes X, Y, Z, ENMO, Ángulo-Z).
3. **Estado Físico:** Presión arterial, altura, peso, porcentaje de grasa corporal, análisis de bioimpedancia (BIA).
4. **Uso de Internet:** Horas de uso diario de computadoras/internet y resultados del Parent-Child Internet Addiction Test (PCIAT).

5. **Sueño:** Escalas para trastornos del sueño.
6. **Evaluaciones Clínicas:** Children's Global Assessment Scale (CGAS).

Distribución de Clases de 'sii':
Se identificó un marcado desequilibrio de clases en la variable objetivo sii en el conjunto de entrenamiento etiquetado (ver ``):

- **Clase 0 (Ninguno):** Aproximadamente 58.3%
- **Clase 1 (Leve):** Aproximadamente 26.7%
- **Clase 2 (Moderado):** Aproximadamente 13.8%
- **Clase 3 (Severo):** Aproximadamente 1.2%

Este desequilibrio es un factor crítico a considerar en el modelado, ya que puede llevar a que el modelo favorezca la clase mayoritaria.

Datos Faltantes: Se observó una tasa considerable de valores faltantes en algunas mediciones, requiriendo técnicas de imputación robustas (ver ``). Algunas características presentaban porcentajes de nulos significativamente más altos en el conjunto de prueba que en el de entrenamiento.

4. Metodología

El proceso de desarrollo del modelo se dividió en las siguientes etapas:

4.1. Carga y Preparación Inicial

1. **Carga de Datos:** Se cargaron los archivos train.csv y test.csv en DataFrames de Pandas.
2. **Manejo de Nulos en Target:** Se eliminaron las filas del conjunto de entrenamiento donde la variable sii (target) era nula para asegurar un entrenamiento con etiquetas válidas. La variable sii se convirtió a tipo entero.
3. **Definición de DataFrames de Trabajo:** Se crearon copias (train_df_copy, test_df_copy, train_labeled_df) para asegurar la inmutabilidad de los DataFrames originales.
4. **Definición de X_train, y_train, X_test iniciales:** Se crearon los conjuntos de características (X_train, X_test) y la variable objetivo (y_train) utilizando todas las columnas comunes entre el entrenamiento y la prueba (excluyendo 'id' y 'sii') como punto de partida para las siguientes etapas.

4.2. Detección de Fugas de Datos y Selección de Características

Esta fase fue crucial para asegurar la generalización del modelo, dado que las evaluaciones iniciales con datos dummy y reales mostraron una precisión del 100% en el split de validación, lo cual era indicativo de fuga de datos.

1. **Análisis de Correlación Inicial:** Se realizó un primer análisis de correlación de Pearson entre todas las características numéricas y sii. Las columnas relacionadas con el PCIAT (Parent-Child Internet Addiction Test) y algunas de SDS (Self-Rating Depression Scale) mostraron correlaciones extremadamente altas (cercanas a 1), lo que indicaba una potencial fuga de datos (estas columnas a menudo son usadas para calcular o están fuertemente ligadas a la definición de sii).
2. **Detección de Mapeo Directo (Fuga Agresiva):** Se implementó una verificación para identificar características que pudieran tener un mapeo casi directo a sii. Se detectó que las columnas **BIA-BIA_SMM** (Skeletal Muscle Mass) y **BIA-BIA_TBW** (Total Body Water) presentaban este tipo de relación, constituyendo una fuga de datos.
3. **Exclusión de Columnas:** Para mitigar la fuga de datos y asegurar un modelo robusto, se decidió excluir las siguientes columnas del conjunto de características:
    - Todas las columnas que contienen 'PCIAT'.
    - Las columnas SDS-SDS_Total_T y SDS-SDS_Total_Raw.
    - Las columnas BIA-BIA_SMM y BIA-BIA_TBW.
    - Columnas con más del 80% de valores nulos en el conjunto de entrenamiento (ej. PAQ_A-PAQ_A_Total, PAQ_A-Season, Physical-Waist_Circumference).
4. **Selección Final de Características:** Después de esta limpieza, el número de características se redujo a **51 características finales** verdaderamente limpias para el modelado, que eran comunes entre X_train y X_test.
5. **Re-evaluación de Correlaciones:** Al recalcular las correlaciones de Pearson con sii usando solo las características limpias (ver ``), las correlaciones más fuertes se redujeron a un rango más realista (0.3 a 0.35). Las características más correlacionadas positivamente con sii resultaron ser Physical-Height, Basic_Demos-Age, y PreInt_EduHx-computerinternet_hoursday.

4.3. Preprocesamiento y Construcción del Pipeline

Para preparar los datos para el modelo, se implementó un Pipeline completo.

1. **Identificación de Tipos de Características:** Las 51 características finales se clasificaron en 42 numéricas y 9 categóricas.
2. **Manejo de Valores Nulos:** Se visualizó el porcentaje de nulos (``), confirmando su presencia generalizada y las diferencias entre los conjuntos de entrenamiento y prueba.
3. **Transformadores:**
    - **Numéricos:** Se aplicó SimpleImputer con estrategia de median para rellenar los nulos, seguido de StandardScaler para escalar las características.

    - **Categóricos:** Se aplicó SimpleImputer con estrategia de most_frequent para rellenar los nulos, seguido de OneHotEncoder con handle_unknown='ignore' para convertir las categorías en un formato numérico adecuado para el modelo.
4. **ColumnTransformer:** Se utilizó un ColumnTransformer para aplicar los transformadores específicos a las columnas numéricas y categóricas correspondientes.
5. **Pipeline:** El ColumnTransformer se integró en un Pipeline junto con el clasificador RandomForestClassifier. Esta estructura asegura que todas las transformaciones se apliquen consistentemente tanto en el entrenamiento como en la predicción, previniendo fugas de información del conjunto de prueba.

4.4. Modelado, Entrenamiento y Evaluación

1. **Modelo Seleccionado:** Se optó por un **RandomForestClassifier**. Se configuró con n_estimators=100 (número de árboles), random_state=42 para reproducibilidad, class_weight='balanced' para mitigar el desequilibrio de clases, y max_depth=10 para controlar el sobreajuste y evitar la memorización de patrones específicos en splits pequeños.
2. **Validación Cruzada:** Para evaluar la estabilidad y el rendimiento generalizable del modelo, se realizó una **validación cruzada estratificada de 5 pliegues** (Stratified K-Fold Cross-Validation) utilizando la métrica de **precisión (accuracy)**.
3. **Entrenamiento Final:** El pipeline completo fue entrenado con el conjunto de entrenamiento limpio y preprocesado (X_train, y_train).
4. **Evaluación en Split de Validación:** Se realizó una evaluación detallada en un split de validación (20% del conjunto de entrenamiento). Se generó una **matriz de confusión** y un **reporte de clasificación** para analizar el rendimiento por clase.

5. Resultados

5.1. Distribución de la Precisión de Validación Cruzada (5-Fold)

El análisis de la validación cruzada (ver ``) mostró los siguientes resultados:

- **Puntuaciones de precisión (Accuracy) por fold:** [0.5529, 0.5887, 0.5722, 0.5484, 0.5667]
- **Precisión promedio (Accuracy):** 0.5658
- **Desviación estándar:** 0.0144

Este resultado indica que el modelo es consistente en su rendimiento a través de los diferentes pliegues de validación, con una precisión alrededor del 56.6%. La baja

desviación estándar sugiere estabilidad en el desempeño.

5.2. Matriz de Confusión y Reporte de Clasificación

La evaluación en el split de validación (20% del conjunto de entrenamiento) arrojó los siguientes resultados (ver ``):

Matriz de Confusión:
Etiqueta Verdadera / Predicha	0	1	2	3	Support
0	311	12	13	0	336
1	22	103	6	0	131
2	3	0	69	0	72
3	0	0	0	9	9

- **Clase 0 (Ninguno):** El modelo predijo correctamente 311 casos, mostrando una alta capacidad para identificar la clase mayoritaria. Se clasificaron erróneamente 12 casos como 'Leve' y 13 como 'Moderado'.
- **Clase 1 (Leve):** De 131 casos reales, 103 fueron correctamente clasificados. 22 fueron clasificados erróneamente como 'Ninguno' y 6 como 'Moderado'.
- **Clase 2 (Moderado):** De 72 casos reales, 69 fueron correctamente clasificados, indicando un buen desempeño. 3 fueron clasificados erróneamente como 'Ninguno'.
- **Clase 3 (Severo):** Los 9 casos reales de 'Severo' fueron clasificados correctamente. Este rendimiento "perfecto" debe interpretarse con cautela debido al número extremadamente bajo de muestras en esta clase en el split de validación, lo que podría deberse a memorización o a características muy distintivas de esos pocos ejemplos.

Reporte de Clasificación:
Clase	Precision	Recall	F1-Score	Support
0	0.93	0.93	0.93	336
1	0.90	0.79	0.84	131
2	0.78	0.96	0.86	72
3	1.00	1.00	1.00	9
Accuracy				0.90
Macro Avg	0.90	0.92	0.91	548
Weighted Avg	0.90	0.90	0.90	548
La precisión general para este split de validación fue del 0.8978.

6. Conclusiones y Trabajo Futuro

Se ha desarrollado un sistema completo de Machine Learning para predecir el Uso

Problemático de Internet. Se logró identificar y mitigar exitosamente fuentes críticas de fuga de datos (PCIAT, SDS-SDS_Total_T, SDS-SDS_Total_Raw, BIA-BIA_SMM, BIA-BIA_TBW), lo que asegura que el modelo aprenda patrones genuinos y no atajos artificiales. El manejo de valores nulos y el desequilibrio de clases se implementaron mediante estrategias robustas dentro de un pipeline de preprocesamiento bien definido.

La precisión promedio del modelo en validación cruzada (56.6%) indica que, si bien predecir el UPI a partir de datos físicos es un desafío, el modelo ha encontrado relaciones significativas. Los resultados de la matriz de confusión muestran una fuerte capacidad para identificar la ausencia de UPI y un rendimiento razonable en las clases intermedias, con un desempeño notable en la clase "severa" (aunque con la cautela mencionada).

Para el trabajo futuro, se proponen las siguientes mejoras:

- **Ajuste Fino de Hiperparámetros:** Realizar una búsqueda exhaustiva de hiperparámetros para el RandomForestClassifier (o modelos alternativos como XGBoost/LightGBM) utilizando técnicas como Grid Search o Random Search, optimizando directamente la Kappa Cuadrática Ponderada si el tiempo de cómputo lo permite.
- **Técnicas Avanzadas de Manejo de Desequilibrio de Clases:** Explorar técnicas como SMOTE, ADASYN, o el uso de pesos de muestreo más sofisticados para generar datos sintéticos de las clases minoritarias y mejorar el rendimiento en estas categorías.
- **Ingeniería de Características:** Investigar la creación de nuevas características a partir de las existentes que puedan ser más informativas para el modelo. Esto podría incluir, por ejemplo, ratios entre mediciones físicas o agregados temporales de los datos de acelerometría si se incorporan.
- **Incorporación de Datos de Series Temporales:** La adición y procesamiento de los datos de acelerometría de muñeca (series_train.parquet, series_test.parquet) podría ofrecer información más rica sobre patrones de actividad, lo que potencialmente mejoraría la capacidad predictiva del modelo. Modelos híbridos (extracción de features estadísticos de acelerometría + clasificación tabular, o redes neuronales recurrentes para las series de tiempo) podrían ser explorados.

Este proyecto demuestra el potencial del Machine Learning para abordar problemas de salud pública, ofreciendo una herramienta accesible que podría contribuir a una detección temprana y a la promoción de hábitos digitales más saludables para las nuevas generaciones.

7. Referencias

- Healthy Brain Network (HBN): Estudio longitudinal de salud mental infantil impulsado por el Child Mind Institute.
- Santorelli, A., Zuanazzi, A., Leyden, M., Lawler, L., Devkin, M., Kotani, Y., & Kiar, G. (2024). *Child Mind Institute — Problematic Internet Use*. Kaggle. https://kaggle.com/competitions/child-mind-institute-problematic-internet-use