



La adicción a la redes sociales
Proyecto de ML – Bootcamp Data Science
The Bridge

Hélène van Caloen
Julio 2025

🎯 Objetivo

Predecir la adicción en los jóvenes, porque nos puede ayudar a tomar medidas preventivas y evitar esta alteración, reorientandoles hacia una vida más sana y un uso responsable de las tecnologías.

📁 Dataset utilizado (Kaggle)

Dataset limpio, 13 columnas y 705 instancias.

Nos encontramos con un problema supervisado de clasificación multclases, no balanceado.

Realizamos un análisis de nuestras variables para entenderlas, categorizarlas y tener una primera idea de priorización:

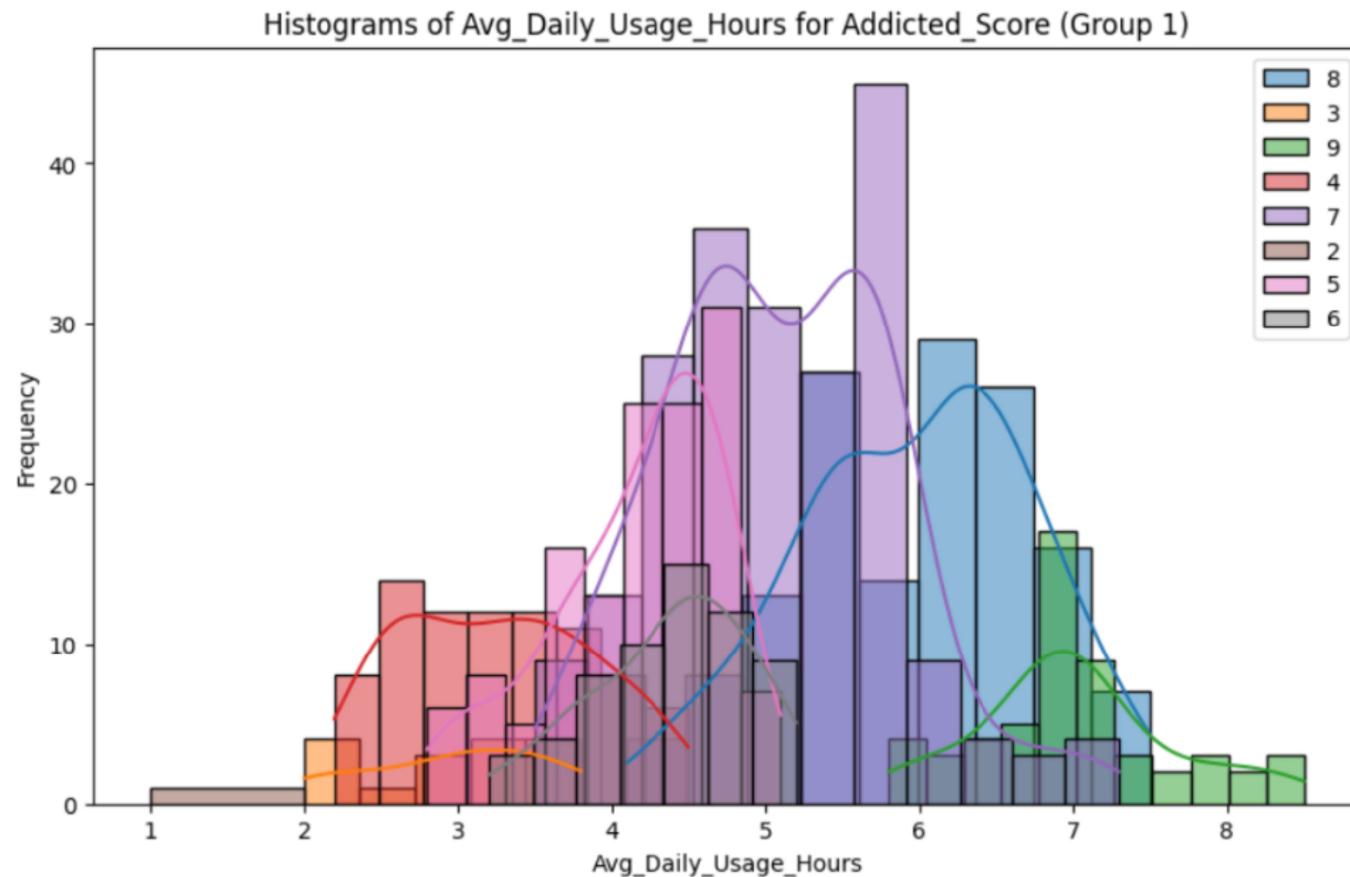
Columna/Variable	Descripción	Tipo de Categoría	Priorizacion
Age	Edad del estudiante	Numerica discreta	Baja
Gender	Genero del estudiante	Binaria	Media
Academic_Level	Nivel academicico (3 tipologias)	Categorica	Alta
Country	Pays del estudiante	Categorica	Alta
Avg_Daily_Usage_Hours	Horas de uso de redes sociales al dia	Numerica continua	Alta
Most_Used_Platform	Plataformas de conección	Categorica	Media
Affects_Academic_Performance	Si afecta resultados academicos	Binaria	Alta
Sleep_Hours_Per_Night	Horas de sueño	Numerica continua	Alta
Mental_Health_Score	Evaluacion de 0 a 10	Categorica	Alta
Relationship_Status	estado actual de relaciones (3 tipologias)	Categorica	Media
Conflicts_Over_Social_Media	numero de conflicto con familiares sobre el uso de redes sociales	Categorica	Alta
Addicted_Score	Clasificación evaluación adicción de 0 a 10	Categorica	Target



Pasos seguidos

- Análisis univariante, con visualizaciones varias (barplot, histogramas, boxplot, etc.), cardinalidad y nunique, etc.
- Análisis bivariantes, otras visualizaciones y análisis de correlación numéricas y de Cramers para categóricas
- Tratamiento de features sobre todo:
 - enteras que son en realidad categóricas,
 - Bins,
 - “One Hot Encoding”
- Target → “Not addicted”, “Moderate”, “Addicted”
- Nuevo análisis bivariantes sobre features creadas (visualizaciones: histplot pairplot, heatmap, y analisis de correlacion con Cramers para estas categoricas nuevas)
- Selección de features importantes (acabamos con 5 de 24)

Ejemplo de impacto de features visual





Modelo Base

1. Decidimos instanciar un Random Forest como baseline
2. Añadimos un SMOTE → evitar predicciones principalmente de clases mayoritarias
3. Aplicamos nuestra selección de features previa
4. Métrica: “Balanced Accuracy” – limita sesgos de una clase desbalanceada

Vesion Modelo	Metrica	Resultado	Numero de columnas
Random Forest clf	Balanced accuracy	0.9404	24
Random Forest clf + SMOTE	Balanced accuracy	0.9896	24
Random Forest clf + SMOTE	Balanced accuracy	0.9323	5

No esta mal... y, ¿si intentamos mejorar?



Selección de features

- Aplicamos varias metodologías de selección de features para intentar mejorar nuestro modelo:
 - Selección visual
 - “Feature importance” → Random Forest
 - “Mutual info classifier”
 - “Recursive Feature Elimination” (RFE)
 - “Recursive Feature Elimination” con validación cruzada (RFECV)
- Descartamos la PCA porque tampoco tenemos tantas features, y sobre todo queremos mantener la “explicabilidad” de nuestro modelo y entender los indicadores mas relevantes que lo influencian.



Probamos varios modelos...

Probamos los modelos, con las listas de features y optamos por ...



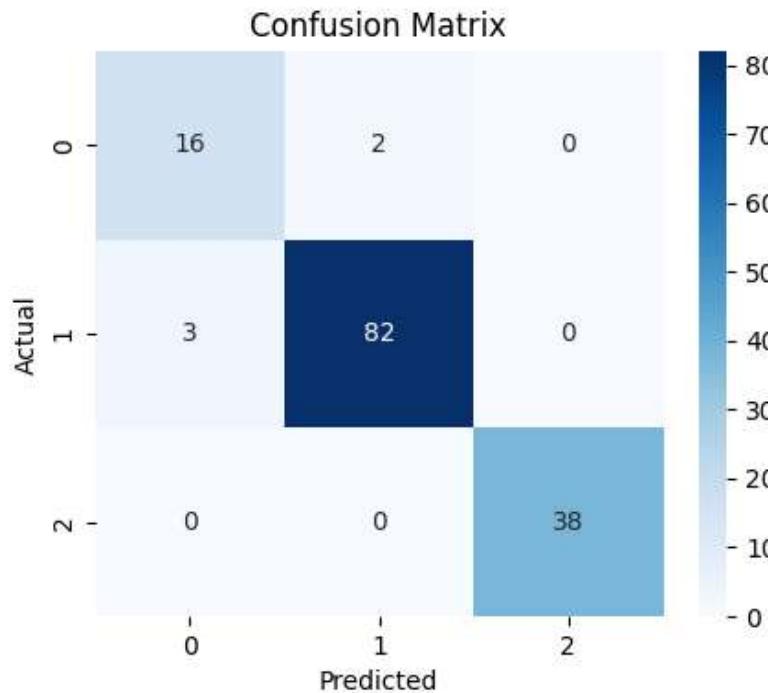
La linea 5: con un numero bastante reducido de features y siendo un modelo rápido y estable, y seguimos teniendo un resultado de casi 97.8 %! Nos parece bastante satisfactorio → **OPTIMIZACIÓN de Hiper-parametros**

Resultados y conclusiones

De 38 predicciones, de jóvenes adictos, todas son correctas

De 84 predicciones, de jóvenes como moderados, solo 2 estarían No adictos.

De 17 predicciones, de jóvenes NO adictos, 3 lo son a nivel moderado.



CONCLUSIONES:

Este modelo responde a la pregunta que nos hacíamos ya que vemos que por las personas mas adictas, acierta muy bien, lo que nos permitirá aplicar medidas preventivas deseadas.