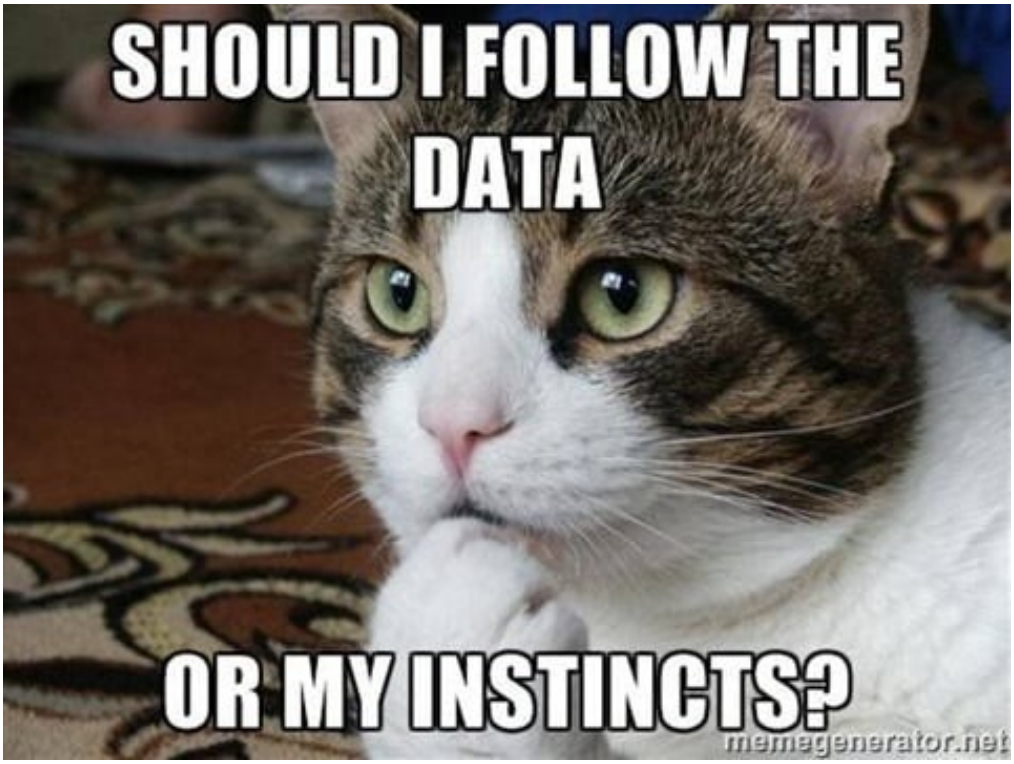
Análisis CSVs



Índice

Explicación:

En este apartado se va a explicar los principales datos de los csvs **train.csv** y *test.csv*.
No se abordará el **dictionary.csv** ya que el objetivo de esto es tener un entendimiento claro de los campos principales, lo cual era el objetivo de dicho csv.
Tampoco se abordará el tema de los archivos formato **parquet**, ya que tienen que ver con otra cosa.

Los principales columnas se dividen en los llamados **instruments**, sin contar la columna **id** que será por supuesto el id del individuo a analizar. Son 11 categorías de instruments, y cada categoría puede albergar más de una columna asociada a ella, haciendo un total de **81** columnas en el train.csv, y **58** columnas en el test.csv. Si os preguntáis por qué una tiene más columnas que otra, es porque las columnas que faltan son las columnas objetivo a calcular con el entrenamiento del modelo. No os preocupéis porque dichas columnas son de un mismo tipo de instrument que se abordará más tarde.

Aunque se ha explicado que se tienen que calcular esas nuevas columnas, lo cierto es que eso solo es para hacer una suma entre ellas, con el objetivo de calcular la verdadera columna final, y la meta definitiva para este trabajo: la columna llamada **sii**.

La columna sii (Severity Impairment Index), es un estándar para medir el uso problemático de Internet, y según el valor que tenga, puede caer en la siguientes numéricas: (Más adelante se explicará la columna **PCIAT_PCIAT_Total**)

- 0 : Ninguna = (PCIAT_PCIAT_Total entre 0 y 30)
- 1 : Leve = (PCIAT_PCIAT_Total entre 31 y 49)
- 2 : Moderado = (PCIAT_PCIAT_Total entre 50 y 79)
- 3 : Severo = (PCIAT_PCIAT_Total entre 80 o más)

Antes de explicar los instrumets, voy a explicar unos términos que voy a utilizar a lo largo de este PDF, y creo que es necesario explicarlos:

- **PIU**: Es el acrónimo en inglés de *Problematic Internet User*. No es como tal una columna en ninguno de los csv, pero puede sustituir a las columnas "**PreInt_EduHx-computerinternet_hoursday**", ya que se refiere a la cantidad de horas en Internet. Es un término real que podéis buscar en internet libremente.
- **EDA**: Es el acrónimo en inglés de *Exploratory Data Analysis*. Simplemente es como se le llama algunas veces a cuando exploramos el análisis de datos en inglés.

Ahora entraré a la clasificación de los instruments. Me he tomado la libertad de agruparlos en 4 grupos según su tipo:

1. **Objetivo**: Estos son los instrumentos que queremos calcular al entrenar nuestro modelo. Está formada por un solo tipo de instrumento, y es el utilizado para calcular el sii.
2. **Salud física y fitness**: Aquí se encuentran los instruments con datos más objetivos, como lo podrían ser la altura del individuo o su resultado en alguna prueba.
   Estos a su vez se subdividirán en 4 partes:
   1. *Escala de Evaluación Global Infantil*: Destinada al análisis de niños menores de 18 años.
   2. *Medidas de salud física*: Su objetivo es medir la salud en el ámbito físico del objetivo.
   3. *Medidas de actividad física*: Miden la actividad física del objetivo.
   4. *Sueño*: Análisis del desorden del sueño del individuo.

   En total, suman 7 instruments.
3. **Demografía**: Se trata de análisis según el sexo y la edad. Está formada por un solo instrument.
4. **Conducta**: Se tratan de cuestionarios que rellena el usuario. Estas se subdividen en dos categorías:
   1. Internet al día: Trata de las horas al día que invierte en internet.
   2. Puntuación actividad física: Cuestionario sobre la actividad física del individuo.

   En total suman 2 instruments.

En adición, pondré un apartado especial al final de cada una de las 4 clasificaciones principales para incluir brevemente su relación con el **PIU**. (El Conducta no tendrá porque a parte de subjetivo, de ahí saco los datos necesarios para apollarme en crear el PIU).

Objetivos:

Parent-Child Internet Addiction Test (PCIAT)

Este instrumento trata sobre la puntuación que dan los padres a sus hijos mediante una serie de preguntas hechas en un cuestionario.
Son un total unas 22 columnas, y estas **NO** se encuentran en el test.csv, sino que son exclusivas del train.csv.
La columnas que la conforman son las siguientes:

- **PCIAT-PCIAT_N (categorical int)**: Siendo N un número del 1 al 20, esta columna representa la puntuación que los padres han dado a los cuestionarios, siendo cada cuestionario un número, (es decir, hay 20 cuestionarios distintos). La puntuación va desde 0 a 5, siendo 0 nunca, y 5 con extremada frecuencia.
- **PCIAT-Season (string)**: Es la época del año en la que se participó en los cuestionarios. Esta se divide en 4: Spring, Summer, Fall(Otoño) y Winter. Solo es para tener más información, y no influye directamente en el cálculo de sii. A partir de ahora, si vemos más adelante alguna columnaa con “Season” en el nombre, se aplicará la misma lógica que esta, teniendo los mismo valores y objetivo, el cual solo es informar.
- **PCIAT-PCIAT_Total**: La suma total de todas las columnas PCIAT-PCIAT_N. Dependiendo de la puntuación de esta columna, el sii será un número u otro. Por ejemplo, si estaascolumnas tienen un 33, el sii será de 1, mientras que si es 100, será de 3.

NOTAS IMPORTANTES:

- Es posible que en algunas filas, algún cuestionario esté sin contestar, es decir, vacío (Null), por lo que se tendría que discutir como rellenarlo. Sin embargo hay dos casos muy particulares: El primero es que hay filas que da igual si pones un 0 o 5 en sus PCIAT vacíos, la puntuación del sii no varía debido a que no pasa de intervalo, incluso con todos los huecos rellenados con 5; el segundo es que dependiendo de los valores que se pongan en los PCIAT vacíos, el sii puede variar porque el total cambia de intervalo, siendo esto bastante sensible. (Creo que se daban 17 filas en este caso).
- Hay casos en las que todos los cuestionarios PCIAT están a 0 o directamente están a nulos. Esto es muy sospechoso, ya que se pueden tomar como algún tipo de hojas en blanco, equivocación o directamente datos no verídicos.
- El sii parece aumentar con la edad, aunque los adolescentes son los que tienen más PCIAT-Total. De todos modos, da la sensación de hay muchos datos de adultos perdidos.
- Debido a que los cuestionarios son más subjetivos, es muy posible que las puntuaciones se vean influidas por las emociones u situaciones de los que contestan.

Relación PIU:

- El PIU normalmente es más alto en adolescentes según los análisis externos.
- Cuanto más alto el sii, más alto el PIU, aunque hay casos en los que no es así, teniendo un sii bajo, y un PIU alto.
- Si os preguntáis de dónde saco el PIU, es de las columnas ***PreInt_EduHx-computerinternet_hoursday*** y ***PreInt_EduHx-computerinternet_hoursday*** que se verán más adelante.

Demografía:

Demographics (Basic_Demos):

Siendo el único instrument de este apartado, Demographics es la información de del sexo y la edad del participante. Aquí sus columnas:

- **Basic_Demos-Enroll_Season(string)**: Época del año de la participación.
- **Basic_Demos-Age(float)**: Edad del participante.
- **Basic_Demos-Sex(categorical int)**: Sexo del participante, siendo 0 Male, y 1 Female.

NOTAS IMPORTANTES:

- Están equitativamente repartidas según las estación del año, aunque la participación es ligeramente más activa en primavera, y menor en otoño(Fall).
- Los hombres superan a las mujeres en cantidad en las edades más jóvenes, aunque se va igualando conforme pasa el tiempo.

Relación PIU:

- No hay mucho o nada destacable que contar sobre el PIU en este apartado, ya que sabemos que va ligado sobretodo a los más jóvenes. Aunque se tendría que ver si afecta el sexo también en este caso.

Salud Física y Fitness:

Pondré el número de su respectiva subclasificación para poder distinguirlos mejor a la derecha de su título.

Children Global Assesment Scale(CGAS)(1):

Este instrument tiene como objetivo analizar de forma numérica mediante doctores especializados en la salud mental, el desempeño de la mentalidad del individuo. Las columnas son las siguientes:

- **CGAS-Season(string)**: Época del año de la participación.
- **CGAS-CGAS_Score(int)**: Puntuación dada por el medico al individuo. Esta puntuación está comprendida entre 1 y 100, siendo 1 la peor y 100 la mejor.

NOTAS IMPORTANTES:

- La mayoría de individuos oscilan entre el 51 y el 80, y hay un caso de puntuación 999, aunque obviamente este último es un error.
- Es difícil describir la relación entre CGAS y sii, y si cogemos muestras pequeñas hay veces en las que CGAS nos confunde aún más, por lo que hay que tener cuidado con este instrument.

Physical Measures(Physical)(2):

Es una colección de presión sanguínea, latidos del corazón, peso, altura, y tamaño de cintura y cadera. Aquí sus columnas:

- **Physical-Season(str)**: Época del año de la participación.
- **Physical-BMI(float)**: Masa corporal calculada con la fórmula (kg/m^2), es decir, peso en kilogramos partido altura en metros cuadrados.
- **Physical-Height(float)**: Altura del individuo. Está en pulgadas (In).
- **Physical-Weight(float)**: Peso del individuo. Está en libras (lbr).
- **Physical-Waist_Circumference(int)**: Medida dela cintura del individuo. Está en pulgadas (In).
- **Physical-Diastolic_BP(int)**: Es la mínima presión sanguínea registrada. Se mide en milímetros de mercurio (mmHg).
- **Physical-HeartRate(int)**: Es la cadencia de latidos media. Está en latidos por minuto (late/min).

- **Physical-Systolic_BP(int)**:Es la máxima presión sanguínea registrada. Se mide en milímetros de mercurio (mmHg).

NOTAS IMPORTANTES:

- Hay valores demasiado valores grandes en el peso o altura como para ser normales, pero claro, esto podría ser, o que son erróneos, o que se trata de algún caso excepcional o de gigantismo, cosa que no podemos saber con certeza.
- Hay valores directamente erróneos en la presión sanguínea (Diastolic y Systolic), como por ejemplo casos en el que Systolic es menor que Diastolic, o que el Diastolic es tenga como registro 0, ya que significaría que el individuo no tiene pulso. También se puede notar que los valores son bajos, dando a entender que cuando se midieron, se hizo en un estado de reposo.
- Aunque la columna BMI se relaciona con la presión sanguínea, no parece haber una correlación fuerte entre estas dos características, aunque si que lo hay entre Diastolic y Systolic.
- Si se analizan los valores de BMI y la presión sanguínea, parece que muchos valores se salen de la media establecida (mirarlo externamente), además de haber un tamaño grande en general de cadera. Esto quiere decir que o están muy gordos o tiene algún tipo de malformación respecto a la cadera.
- Hay correlación entre la altura, peso, presión sanguínea y tamaño de la cadera, por lo que da a entender que personas con más altura y peso tienen más sii. Sabemos que estos valores aumentan con la edad, pero también sabemos que los adolescentes son por lo general los que más sii tienen.

Bio-electric Impedance Analysis(BIA)(2):

Es una recopilación de las medidas de los componentes esenciales del cuerpo, como el BMI (Índice de masa corporal), grasa, músculo y contenido de agua. Sus columnas son las siguientes:

- **BIA-Season(string)**: Época del año de la participación.
- **BIA-BIA_Activity_Level_num(categorical int)**: Nivel de actividad del individuo. Esta puntuación va desde 1 a 5, siendo 1 muy ligera, y 5 excepcional.
- **BIA-BIA_Frame_Num(categorical int)**: Es el tipo/tamaño de esqueleto del individuo. Esta puntuación va desde 1 a 3, siendo 1 pequeño, 2 mediano, y 3 grande. Este aspecto se calcula según la altura y el tamaño de la circunferencia de la muñeca del individuo.
- **BIA-BIA_{Acrónimos}(float)**: Son las medidas de distintos aspectos y componentes del cuerpo del individuo. La parte {Acrónimos} refleja qué parte del cuerpo o que característica se está midiendo.

NOTAS IMPORTANTES:

- No se se sabe con certeza si los datos están hechos mediante una fórmula que sigue el BIA o son raw. Parece ser que es los primero, ya que tiende a sobrestimar la masa muscular según el sexo, altura, peso y otros factores.
- Las mediciones de BIA son un poco inexactas, ya que contiene valores negativos o absurdamente grandes.
- Se podría compara la columna **Physical-BMI** con la columnas **BIA-BIA_BMI**, sin embargo, parece haber valores 0 incluso después de haber sido comparados y entrenados. Eso es bastante confuso la verdad.

FitnessGram Vitals and Treadmill(Fitness_Endurance)(2):

Medidas de las aptitudes cardiovasculares captadas usando el protocolo NHANES, (un test de ejercicio físico basado en género, edad, masa corporal, y ejercicio físico auto reportado). Sus columnas son las siguientes:

- **Fitness_Endurance-Season(string)**: Época del año de la participación.
- **Fitness_Endurance-Max_Stage (int)**: Máxima fase a la que han llegado los participantes.
- **Fitness_Endurance-Time_Mins(int)**: Tiempo en el que se completó la prueba en minutos (participante exhausto).
- **Fitness_Endurance-Time_Sec(int)**: Tiempo en el que se completó la prueba en segundos(participante exhausto).

NOTAS IMPORTANTES:

- Algunas veces podemos encontrarnos los minutos o segundos con valores nulos. Se podrían tomar como 0, y aunque algunas veces podría no afectar, (por ejemplo, si tenemos los minutos y no los segundos), hay veces que es demasiado sensible, como por ejemplo en el que caso de que falten ambos campos o falten los minutos.
- La media de los participantes está en el stage 5.
- Algunos participantes fallaron al completar el stage 1 (0 minutos). Esto podría ser un error o directamente son inútiles.
- Hay algunos individuos de 7 y 8 años con una resistencia demasiado grande.
- Particularmente en este instrument hay bastantes datos faltantes.

FitnessGram Child(FCG)(3):

Evaluación de aptitudes físicas relacionadas con la salud midiendo 5 parámetros en concreto: capacidad aeróbica, fuerza muscular, resistencia muscular, flexibilidad, y composición del cuerpo. Aquí están sus columnas:

- **FGC-Season(string)**: Época del año de la participación.
- **FGC-FGC_CU(int)**: Número de abdominales hechos.
- **FGC-FGC_CU_Zone(categorical int)**: Puntuación para determinar si está lo suficientemente sano en temas de realizar curl. 0 es que necesita mejorar, y 1 es que está en una zona sana.
- **FGC-FGC_GSND(float)**: Puntuación de la fuerza de agarre con su mano NO dominante.
- **FGC-FGC_GSND_Zone(categorial int)**: Clasificación de la fuerza de agarre de su mano NO dominante. Va desde 1 a 3, siendo 1 débil, 2 moderada, y 3 fuerte.
- **FGC-FGC_GSD(float)**: Puntuación de la fuerza de agarre con su mano dominante.
- **FGC-FGC_GSD_Zone(categorial int)**: Clasificación de la fuerza de agarre de su mano dominante. Va desde 1 a 3, siendo 1 débil, 2 moderada, y 3 fuerte.
- **FGC-FGC_PU(int)**: Número total de flexiones.
- **FGC-FGC_PU_Zone(categorical int)**: Clasificación del estado del individuo haciendo flexiones. 0 es que necesita mejorar, y 1 es que está sano.
- **FGC-FGC_SRL(float)**: Puntuación de estirarse sentado hacia el lado izquierdo.
- **FGC-FGC_SRL_Zone(categorial int)**: Clasificación del estado del individuo estirándose sentado hacia su izquierda. 0 es que necesita mejorar, y 1 es que está sano.
- **FGC-FGC_SRR(float)**: Puntuación de estirarse sentado hacia el lado derecho.
- **FGC-FGC_SRL_Zone(categorial int)**: Clasificación del estado del individuo estirándose sentado hacia su derecha. 0 es que necesita mejorar, y 1 es que está sano.
- **FGC-FGC_TL(int)**: Puntuación de realizar el ejercicio trunk lift.
- **FGC-FGC_TL_ Zone(categorical int)**: Clasificación del estado del individuo realizando el ejercicio trunk lift. 0 es que necesita mejorar, y 1 es que está sano.

NOTAS IMPORTANTES:

- Aunque el total de los resultados tienden a ser bajos, la mayoría de los participantes están sanos en el ejercicio trunk lift. (Aunque es un ejercicio relativamente fácil así que no sé si contar eso).

- Los valores para distintos ejercicios se solapan entre sí, esto puede deberse que con con la edad, unos ejercicios pueden aumentar de puntuación mientras que otros bajan.
- Aun con lo anterior, parece que incluso en las mismas edades, hay diferentes valores para los ejercicios, posiblemente debido al sexo u otros factores como condiciones especiales.
- Hay correlación entre la fuerza de agarra de las dos manos y los estiramientos en ambas direcciones. Tienen sentido, ya que se espera que si una aumenta la otra también.
- Aunque se tenga buena puntuación en las pruebas, eso no significa que refleje su nivel actual, ya que no sabemos a ciencia cierta cuando se realizó a excepción de la época del año. Tampoco sabemos si fueron las mismas pruebas para todos o personalizadas.

Actigrahy(---)(3):

Análisis de la actividad física mediante un biotracker. Esto se corresponde con los archivos parquet, (aunque se encuentre como instrument), ya que ellos son los que contienen los datos. Sus columnas están definidas en su respectivo apartado en la página de Kaggle, por lo que no se tratará aquí.

Sleep Disturbance Scale(SDS)(4):

Escala utilizada para la categorizar el desorden del sueño del individuo. Aquí sus columnas:

- **SDS-Season(string)**: Época de participación del individuo.
- **SDS-SDS_Total_Raw(int)**: Puntuación total “raw” del sueño del individuo.
- **SDS-SDS_Total_T(int)**: Puntuación total del sueño del individuo.

NOTAS IMPORTANTES:

- Tanto la columna raw como T, tienen valores moderadamente variables, indicando efectivamente un desorden del sueño en varios participantes.
- Aunque T por lo general suelen tener valores más altos que Raw, no entiendo muy bien su diferencia, aunque podría ser que Raw es el tomado por los médicos, y T tienen alguna especie de fórmula o método aplicada.
- El sexo y la edad puede también influir en el sueño, además de haber valores que no están definidos (nulos).

Relación PIU:

- La columnas "**CGAS-CGAS_SCORE**" podría indicar el rango de efecto del PIU sobre el individuo. No obstante, no hay participantes con el sii máximo (3), que tengan buen CGAS score, sugiriendo que las respuestas de los padres puede reflejar PIU en temas de salud o funcionalidad.
- Enlazando con lo anteriormente dicho, hay casos de individuos con un PIU alto pero ningún problema de salud, aunque claro, no son la mayoría.
- El **BMI** podría aumentar también según el PIU del usuario, debido a la inactividad física.
- Coincidiendo con lo de arriba, el PIU también podría ser un reflejo en el apartado del instrumento **BIA**.
- En el instrumento del sueño **SDS**, el PIU podría ser una de las razones del desorden del sueño de los individuos.
- Al parecer, los problemas vasculares, que no salud, no tienen mucho que ver con el PIU, aunque se necesitaría una confirmación más contundente.
- Hay una parte bastante contraintuitiva, y es que al parecer en los test del instrument **FGC**, tienen una correlación positiva al PIU, indicando que cuanto más sano esté en las pruebas como la de abdominales o trunk up, más PIU tendrá el individuo… curioso cuanto menos.
- También se sabe que la actividad física aumenta con la edad, por lo que la edad indirectamente podría estar relacionada en el aumento del PIU.

Conducta:

Physical Activity Questionnaire(PAQ):

Información sobre la participación del individuo en un periodo de 7 días.
Se divide en dos partes según sus objetivos, siendo estos niños (Children,C), y adolescentes (Adolescents,A). Estas son sus columnas:

- **PAQ_A-Season(string)**: Época del año de la participación.
- **PAQ_A-PAQ_A-Total(float)**: Total de la suma de los puntos del cuestionario para adolescentes
- **PAQ_C-Season(string)**: Época del año de la participación.
- **PAQ_C-PAQ_C-Total(float)**: Total de la suma de los puntos del cuestionario para niños.

NOTAS IMPORTANTES:

- Parece que la división entre adolescentes y niños está mal, ya que en los niños va de 7 a 17 años, mientras que en los adolescentes van de 13 a 18 años.
- Parece que la actividad física de los niños y adolescentes es bastante estable.. ¿ estarán mintiendo en los test ?
- Hay bastantes valores que faltan en este instrumento.

Internet Use(PreInt):

Número de horas al día que consume al uso del internet el individuo. Es te es el instrumento en concreto que he usado para poder referirme al PIU, ya que el PIU es básicamente las horas que alguien le hecha al internet. Estas son sus columnas:

- **PreInt_EduHx-Season(string)**: Época de participación del individuo.
- **PreInt_EduHx-computerinternet_hoursday(categorical int)**: Horas que el individuo le echa al ordenador. Está dividido según puntuación, siendo 0 menos de 1 hora al día, 1 alrededor de una hora, 2 alrededor de 2 horas, y 3 más de 3 horas al día.

NOTAS IMPORTANTES:

- Este fue el instrument usado para el PIU en este análisis. El PIU de todos modos se puede buscar de manera general en internet, pero está bien saber las horas que le dedican al día los individuos, ya que pienso que es de lo más importante.
- Habría que fijar un número del cual a partir de ahí el uso de Internet sea preocupante.