Facultad de Ciencias de la Computación

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD CIENCIAS DE LA COMPUTACIÓN

TIME-WASTERS ON SOCIAL MEDIA

EMILIANO MASTRANZO JUÁREZ

PROFESOR: JAIME ALEJANDRO ROMERO SIERRA

FECHA: 02/12/2024

1. INTRODUCCIÓN

- **Objetivo**
  Encontrar el nivel de impacto que tienen el uso de celular en las redes sociales y como afecta en el tiempo y productividad perdida de los usuarios.

- **Justificación**
  Encontrar los resultados esperados nos ayudará a implementar acciones para evitar que los usuarios que tienen un mayor impacto en su productividad durante el día contraigan una adicción más fuerte y dependiente hacia el uso del celular; de esta forma podremos hacer que los usuarios se enfoquen y sean más productivos en cualquier actividad que hagan, beneficiando a todos los sectores involucrados en estas mismas. De otra forma puede ser usado para encontrar ciertos patrones de consumo que tienen los usuarios y cómo el marketing podría ser redirigido a ciertas personas.

- **Fuentes de datos**

  El conjunto de datos proporciona una visión completa de las interacciones y el compromiso de los usuarios con varias plataformas de redes sociales. Este conjunto de datos abarca una amplia gama de atributos que facilitan un análisis exhaustivo de cómo las redes sociales afectan la gestión del tiempo y la productividad de los usuarios. Sirve como un recurso esencial para investigadores, especialistas en marketing y científicos sociales que buscan profundizar en las complejidades de los patrones de consumo de las redes sociales.

2. METODOLOGÍA

- ## Proceso de limpieza de datos

  Empezamos convirtiendo nuestras variables de tipo objeto a tipo flotante que tienen datos numéricos; seguido a esto, para no borrar las filas que tienen NaN, usamos el llenado por promedio de cada una de nuestras variables, de esta forma seguimos conservando toda la fila y tenemos una respuesta del dato que nos faltaba.

  Después borramos todos los NaN de nuestras variables categóricas, en este caso se borra toda la fila de nuestros NaN porque no podemos darles valores por nosotros mismos o rellenarlos por medio de un algoritmo.

  Borramos nuestras columnas 'Frequency', 'Income', 'Debt', 'Owns Property', 'Video ID', 'Importance Score', 'Scroll Rate', 'Watch Reason', 'ConnectionType' debido que en nuestro estudio no son tan esenciales y si queremos borrar los NaN que tienen, estaríamos perdiendo demasiadas filas quedándonos con una diminuta base de datos.

  Comprobamos que nuestra base de datos no tenga valores inválidos y una vez hecho esto, procedemos a resetear el índice para poder apreciar mejor el orden de nuestra base.

- ## Análisis Exploratorio de Datos (EDA)

a) Descripción General de los Datos

- **Visión general:** La base de daros cuenta con 485 filas y 22 columnas.

- **Tipos de variables:**

UserID	object
Age	float64
Gender	object
Location	object
Profession	object
Demographics	object
Platform	object
Total Time Spent	float64
Number of Sessions	float64
Video Category	object
Video Length	float64
Engagement	float64
Time Spent On Video	float64
Number of Videos Watched	float64
ProductivityLoss	float64
Satisfaction	float64
DeviceType	object
OS	object
Watch Time	object
Self Control	float64
Addiction Level	float64
CurrentActivity	object

- **Resumen estadístico:**

	Age	Total Time Spent	Number of Sessions	Video Length	Engagement	Time Spent On Video	Number of Videos Watched	ProductivityLoss	Satisfaction	Self Control	Addiction Level
count	475.000000	475.000000	475.000000	441.000000	475.000000	475.000000	475.000000	475.000000	475.000000	475.000000	475.000000
mean	40.510210	154.278943	9.844345	15.321995	4969.381802	15.080000	25.240093	5.094737	4.911923	7.048421	3.004080
std	13.358442	81.820103	5.208445	8.326611	2818.151687	8.103362	13.762885	2.176174	2.110233	2.110795	2.022086
min	18.000000	10.000000	1.000000	1.000000	48.000000	1.000000	1.000000	1.000000	1.000000	3.000000	0.000000
25%	29.000000	87.000000	6.000000	8.000000	2448.000000	8.000000	14.000000	3.000000	4.000000	5.000000	2.000000
50%	40.889571	151.326844	9.930857	15.000000	4991.535677	16.000000	25.036008	5.000000	4.883745	7.000000	2.944269
75%	52.000000	223.000000	14.000000	22.000000	7221.000000	22.000000	36.000000	6.500000	7.000000	8.000000	5.000000
max	64.000000	298.000000	19.000000	29.000000	9982.000000	29.000000	49.000000	9.000000	9.000000	10.000000	7.000000

Frecuencia de categorías para la columna 'UserID':

```
UserID
invalid    11
205.0       2
186.0       2
119.0       2
811.0       2
           ..
397.0       1
395.0       1
394.0       1
393.0       1
488.0       1
Name: count, Length: 464, dtype: int64
```

Frecuencia de categorías para la columna 'Gender':

```
Gender
Male       242
Female     154
Other       79
invalid     10
Name: count, dtype: int64
```

Frecuencia de categorías para la columna 'OS':

```
OS
Android    234
iOS        126
Windows     61
MacOS       53
invalid     11
Name: count, dtype: int64
```

Frecuencia de categorías para la columna 'Watch Time':

```
Watch Time
2:00 PM     71
9:00 PM     59
5:00 PM     43
3:55 PM     40
4:25 PM     36
10:15 PM    34
11:30 PM    33
8:30 PM     29
7:25 PM     23
3:45 PM     22
5:45 PM     20
6:05 PM     18
8:00 AM     16
7:45 AM     14
invalid     14
9:15 AM      7
9:55 AM      6
Name: count, dtype: int64
```

Frecuencia de categorías para la columna 'Location':

```
Location
India            117
United States     81
Indonesia         39
Vietnam           39
Barzil            38
Mexico            35
Pakistan          35
Philippines       34
Japan             30
Germany           29
invalid            8
Name: count, dtype: int64
```

Frecuencia de categorías para la columna 'Profession':

```
Profession
Students         115
Waiting staff     98
Labor/Worker      92
driver            46
Cashier           35
Engineer          28
Teacher           23
Manager           21
Artist            19
invalid            8
Name: count, dtype: int64
```

Frecuencia de categorías para la columna 'Demographics':

```
Demographics
Rural      365
Urban      110
invalid     10
Name: count, dtype: int64
```

Frecuencia de categorías para la columna 'Platform':

```
Platform
Instagram    132
TikTok       131
YouTube      116
Facebook      94
invalid       12
Name: count, dtype: int64
```

Frecuencia de categorías para la columna 'Video Category':

```
Video Category
Life Hacks       86
Jokes/Memes      83
Gaming           61
Vlogs            52
Trends           49
Entertainment    48
Pranks           41
ASMR             37
Comedy           17
invalid          11
Name: count, dtype: int64
```

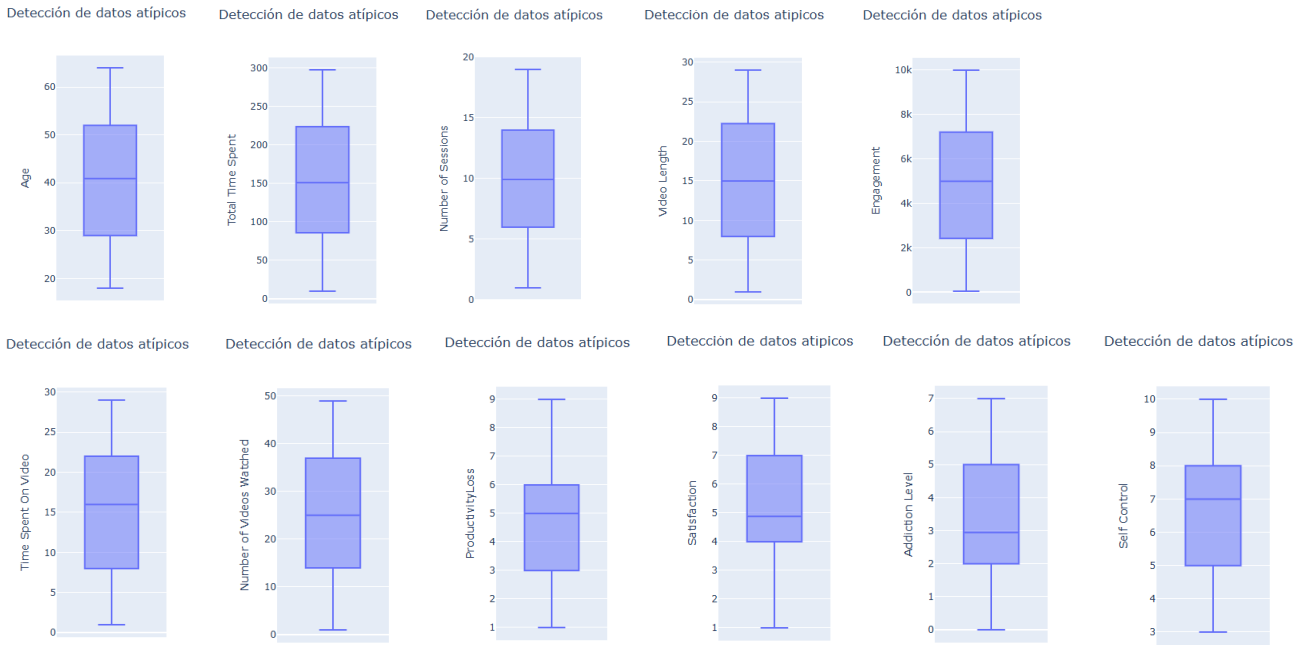
Frecuencia de categorías para la columna 'DeviceType':

```
DeviceType
Smartphone    258
Tablet        146
Computer       66
invalid        15
Name: count, dtype: int64
```

Frecuencia de categorías para la columna 'CurrentActivity':

```
CurrentActivity
At home       174
At school     117
At work       116
Commuting      56
invalid        12
Name: count, dtype: int64
```

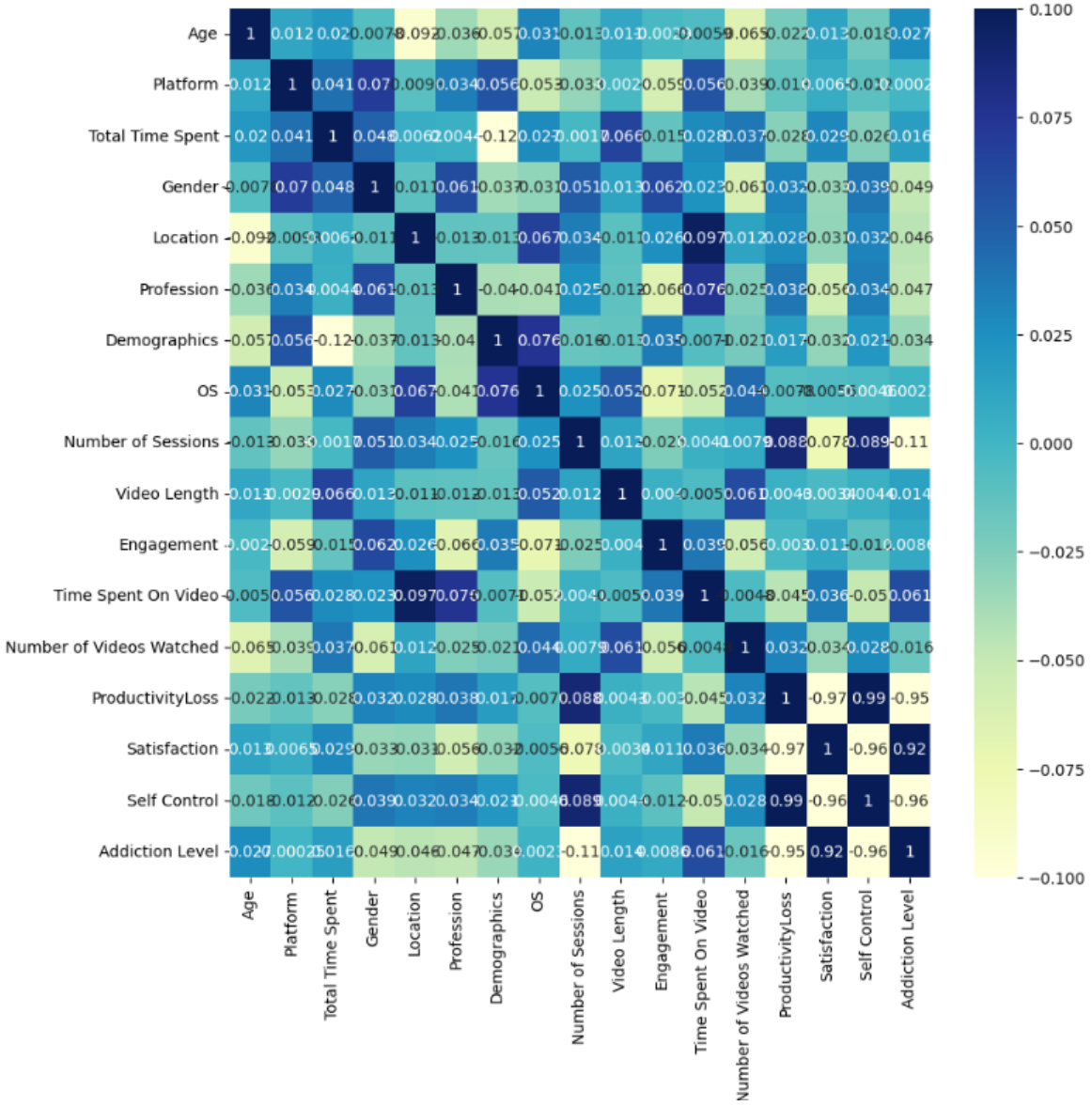
b) Visualización y Distribución de Variables Individuales



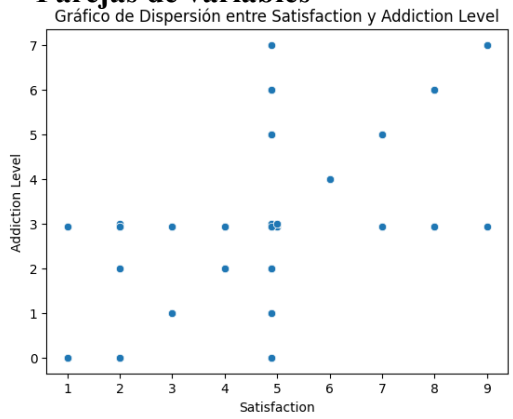
c) Correlación entre variables

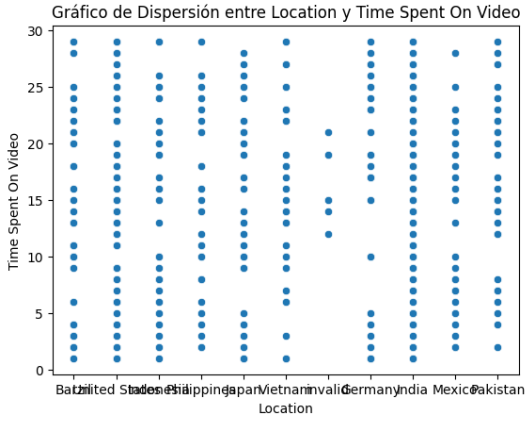
▪ Matriz de correlación

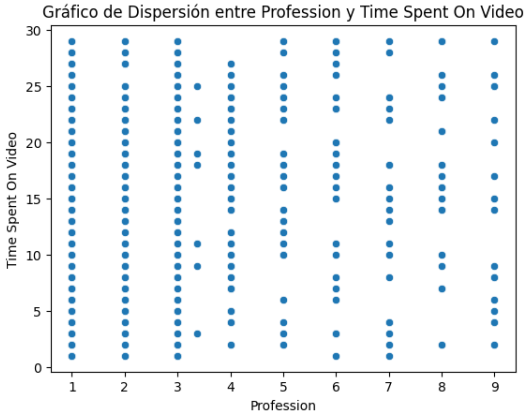
	Age	Platform	Total Time Spent	Gender	Location	Profession	Demographics	OS	Number of Sessions	Video Length	Engagement	Time Spent On Video	Number of Videos Watched	ProductivityLoss	Satisfaction	Self Control	Addiction Level
Age	1.000000	0.011587	0.019722	-0.007839	-0.092354	-0.035595	-0.056964	0.030757	-0.012833	0.010803	-0.002886	-0.005937	-0.064961	-0.022180	0.012531	-0.018061	0.027236
Platform	0.011587	1.000000	0.041041	0.069584	-0.009826	0.034368	0.055972	-0.053006	-0.032747	-0.002913	-0.058556	0.056057	-0.038899	-0.013404	0.006546	-0.011583	-0.000251
Total Time Spent	0.019722	0.041041	1.000000	0.047943	0.006201	0.004351	-0.124779	0.027362	-0.001714	0.065690	-0.014843	0.027608	0.037153	-0.028034	0.029278	-0.026131	0.015888
Gender	-0.007839	0.069584	0.047943	1.000000	-0.011259	0.060796	-0.036822	-0.031093	0.051439	0.012702	0.062232	0.023485	-0.060779	0.031561	-0.033281	0.038667	-0.048622
Location	-0.092354	-0.009826	0.006201	-0.011259	1.000000	-0.012613	-0.013357	0.066619	0.034401	-0.010793	0.025616	0.096786	0.011789	0.027719	-0.031457	0.031730	-0.045687
Profession	-0.035595	0.034368	0.004351	0.060796	-0.012613	1.000000	-0.040476	-0.041155	0.024826	-0.012088	-0.066387	0.076373	-0.025275	0.037834	-0.056068	0.033760	-0.046898
Demographics	-0.056964	0.055972	-0.124779	-0.036822	-0.013357	-0.040476	1.000000	0.075729	-0.015625	-0.013380	0.035384	-0.007085	-0.021074	0.017423	-0.031894	0.020637	-0.034004
OS	0.030757	-0.053006	0.027362	-0.031093	0.066619	-0.041155	0.075729	1.000000	0.024665	0.052020	-0.070829	-0.052423	0.044255	-0.007839	-0.005569	-0.004640	0.002303
Number of Sessions	-0.012833	-0.032747	-0.001714	0.051439	0.034401	0.024826	-0.015625	0.024665	1.000000	0.012480	-0.024783	0.004070	0.007920	0.088253	-0.078083	0.089410	-0.112081
Video Length	0.010803	-0.002913	0.065690	0.012702	-0.010793	-0.012088	-0.013380	0.052020	0.012480	1.000000	0.003981	-0.005251	0.061341	0.004289	-0.003363	0.004395	0.014405
Engagement	-0.002886	-0.058556	-0.014843	0.062232	0.025616	-0.066387	0.035384	-0.070829	-0.024783	0.003981	1.000000	0.038975	-0.055931	-0.002981	0.010953	-0.011536	0.008616
Time Spent On Video	-0.005937	0.056057	0.027608	0.023485	0.096786	0.076373	-0.007085	-0.052423	0.004070	-0.005251	0.038975	1.000000	-0.004780	-0.045175	0.035890	-0.049564	0.060709
Number of Videos Watched	-0.064961	-0.038899	0.037153	-0.060779	0.011789	-0.025275	-0.021074	0.044255	0.007920	0.061341	-0.055931	-0.004780	1.000000	0.031576	-0.033772	0.028064	-0.016153
ProductivityLoss	-0.022180	-0.013404	-0.028034	0.031561	0.027719	0.037834	0.017423	-0.007839	0.088253	0.004289	-0.002981	-0.045175	0.031576	1.000000	-0.969740	0.993810	-0.953950
Satisfaction	0.012531	0.006546	0.029278	-0.033281	-0.031457	-0.056068	-0.031894	-0.005569	-0.078083	-0.003363	0.010953	0.035890	-0.033772	-0.969740	1.000000	-0.963221	0.922063
Self Control	-0.018061	-0.011583	-0.026131	0.038667	0.031730	0.033760	0.020637	-0.004640	0.089410	0.004395	-0.011536	-0.049564	0.028064	0.993810	-0.963221	1.000000	-0.958711
Addiction Level	0.027236	-0.000251	0.015888	-0.048622	-0.045687	-0.046898	-0.034004	0.002303	-0.112081	0.014405	0.008616	0.060709	-0.016153	-0.953950	0.922063	-0.958711	1.000000

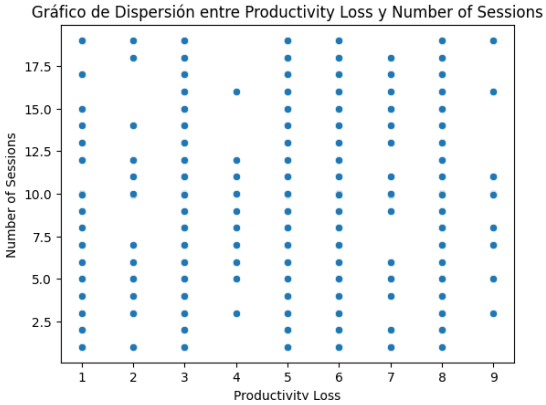

Age
Platform
Total Time Spent
Gender
Location
Profession
Demographics
OS
Number of Sessions
Video Length
Engagement
Time Spent On Video
Number of Videos Watched
ProductivityLoss
Satisfaction
Self Control
Addiction Level
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100

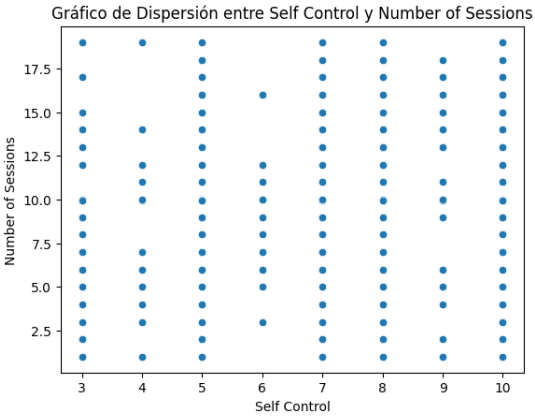
- **Parejas de variables**

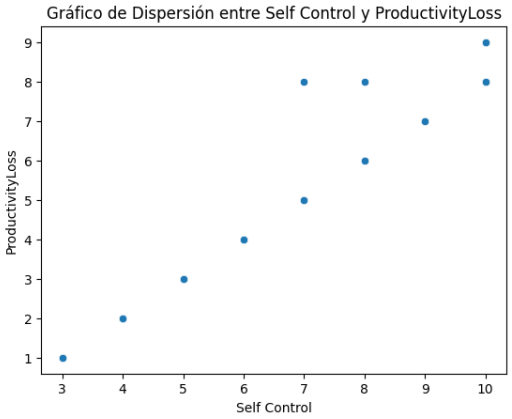








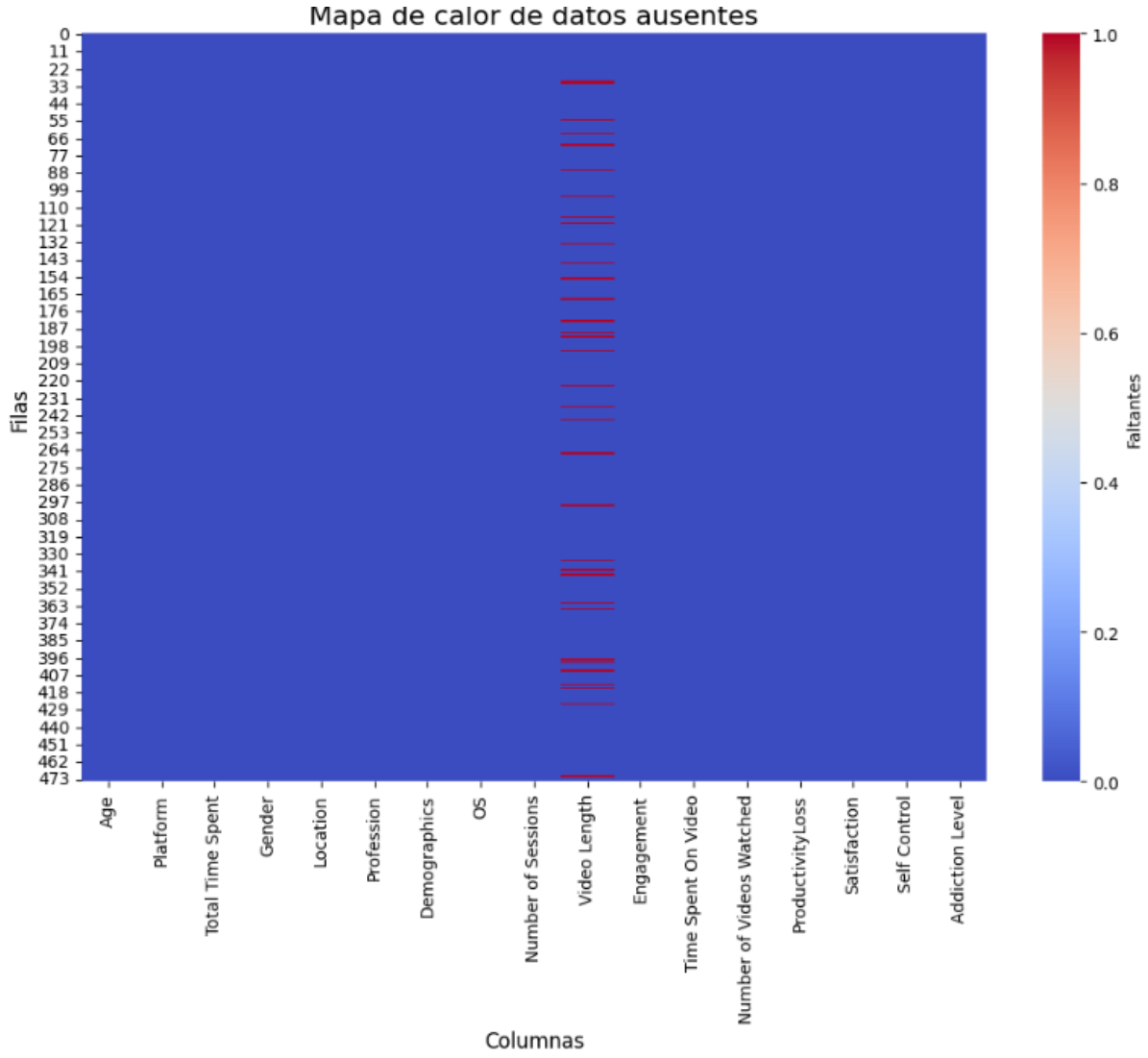




d) Análisis de Valores Atípicos (Outliers)

Sin Outliers.

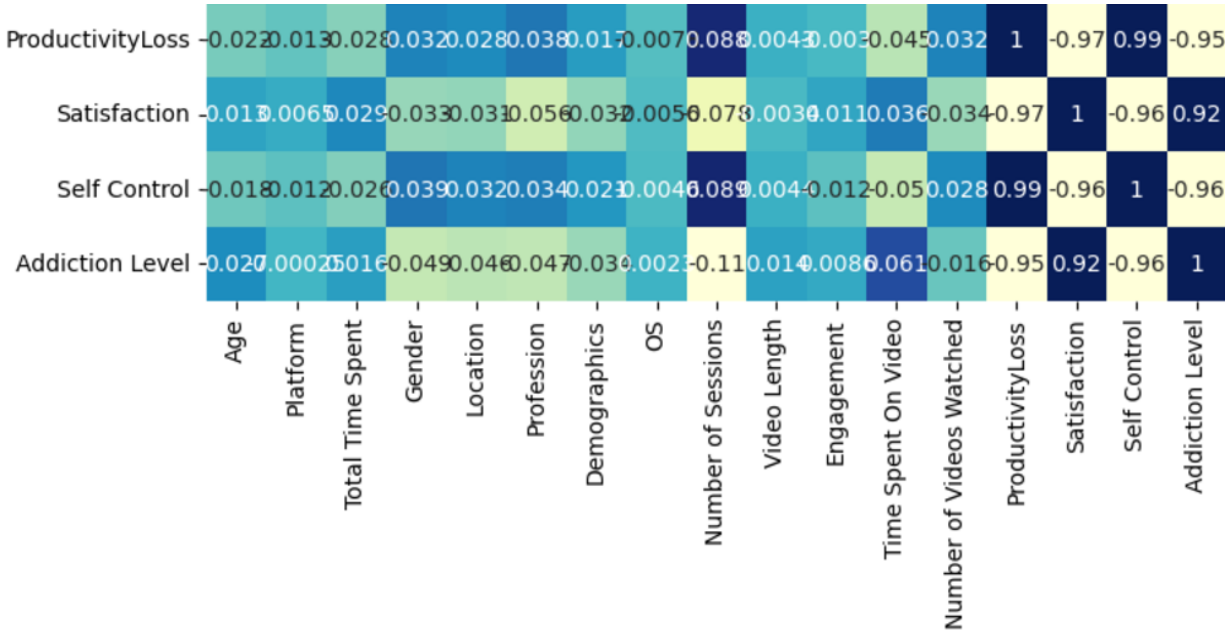
e) Análisis de Valores Faltantes



A los datos faltantes se les hizo un llenado en base a su media, siendo este el método más práctico, ya que, si eliminamos las filas con los valores faltantes, se perderían datos importantes en nuestro estudio de la productividad.

f) Observaciones y Hallazgos Importantes

Se quiere encontrar la variable de “ProductivityLoss” y las variables que están muy fuertemente correlacionadas con esta misma son “Self Control”, “Addiction Level” y “Satisfaction”.



- **Hallazgos Clave**
  Podemos observar en el mapa de calor completo que el tiempo que pasan viendo videos está muy correlacionada con el país en el que residen los usuarios, siendo un factor que se podría estudiar aparte, teniendo mente un objetivo distinto al que nosotros estamos estudiando.
  La correlación entre la satisfacción y el nivel de adicción, así como el autocontrol con la productividad perdida tienen una correlación lineal, si baja uno el otro es casi seguro que baje, de igual manera si uno sube, el otro tendrá que subir.
  Uno llegaría pensar de primera instancia que columnas como la edad, el tiempo total que emplean al uso de las redes o el tiempo que pasan viendo videos estarían muy de la mano con el tiempo de productividad perdida; pero en esta base de datos nos encontramos que incluso son los que menos se relacionan con la productividad perdida.

- **Implicaciones para el Modelo**
  Sabiendo que hay un conjunto de variables que pueden modificar el resultado dependiendo su valor, podríamos darnos una idea de qué modelo se acopla mejor a nuestras necesidades.

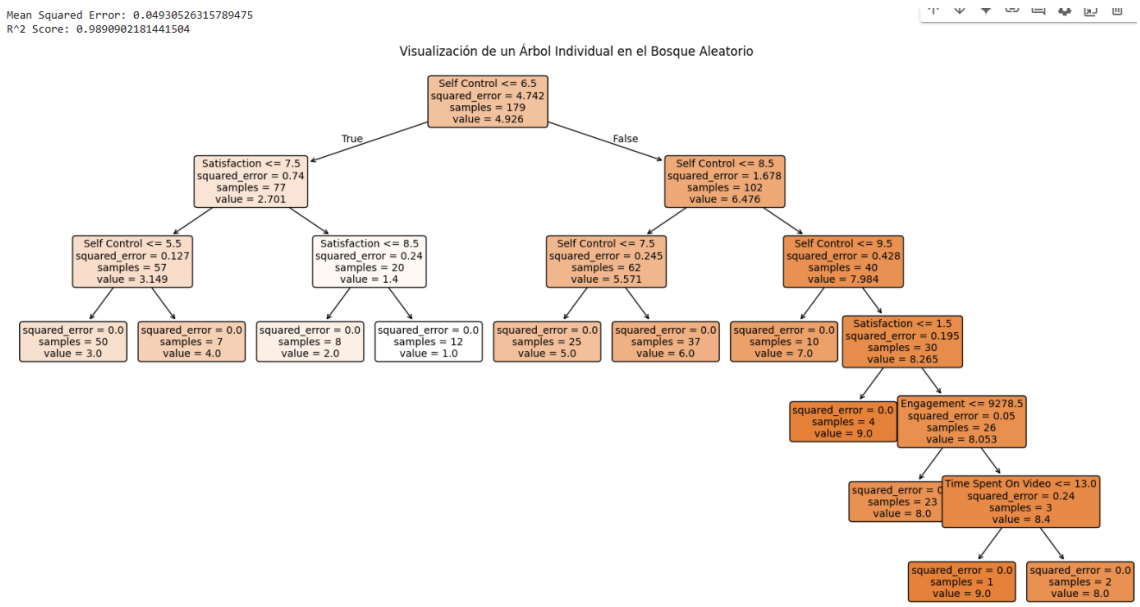
3. MODELO DE MACHINE LEARNING

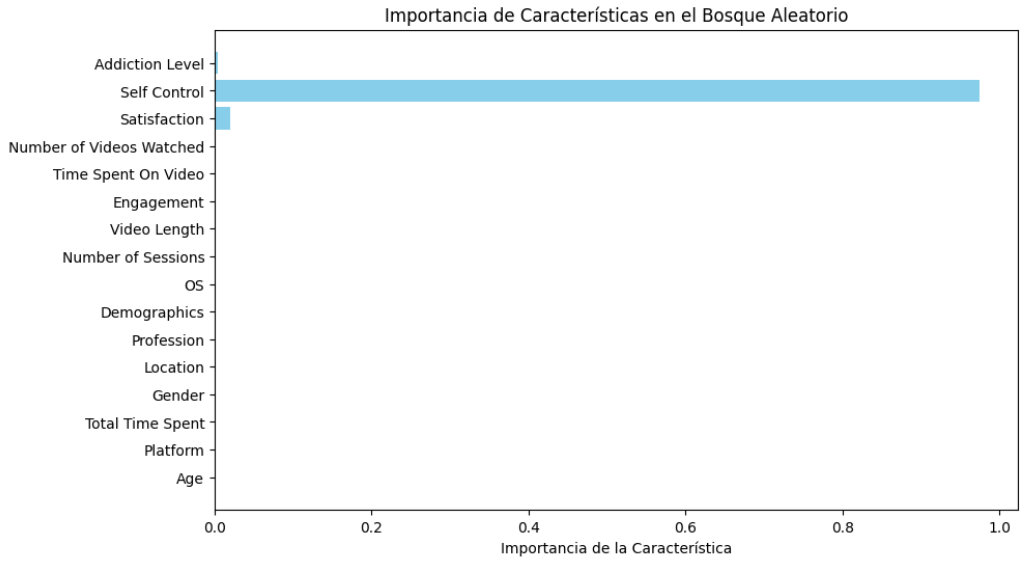
- **Descripción del modelo**
  Random Forest, un modelo de aprendizaje automático supervisado.

- **Justificación**
  Al darnos cuenta de que distintas variables son esenciales para saber el impacto en la pérdida de productividad que tiene la gente, podemos hacer uso de un Random Forest como proceso de machine learning, de esta forma podríamos encontrar el nivel de productividad que pierde la gente con distintos valores de nuestras variables.

- **Implementación y Entrenamiento**
  - ❖ Dividimos nuestras características de nuestra variable objetivo.
  - ❖ En la parte de prueba vamos a poner que agarre el 40% de los datos para que evalúe más robustamente el rendimiento del modelo y lo sobrante para el entrenamiento, que sería el 60% de nuestros datos.
  - ❖ Creamos nuestro modelo de Random Forest y ajustamos el modelo a los datos del entrenamiento, que busca los patrones que relacionan nuestras características con la variable objetivo.
  - ❖ El modelo predice los valores de "y" para el conjunto de prueba.
  - ❖ Calcula el error cuadrático medio (MSE).
  - ❖ R2_score mide que tan bien se ajustan las predicciones a los datos reales; mientras más cerca al 1, el modelo tiene un buen ajuste.
  - ❖ Extraemos el primer árbol de nuestro bosque y nos genera un gráfico de cómo el primer árbol tomó las decisiones, etiquetando las ramas del árbol con los nombres de las características.
  - ❖ Como última, por medio de un gráfico de barras indicamos la importancia de las características al hacer nuestras predicciones.

- **Resultados**





EL MSE nos indica el porcentaje de error que tiene nuestro árbol de decisiones y por el contrario nuestro r2_score nos indica el porcentaje de validación.

En la gráfica de barras podemos notar nuestras variables que tienen influencia y cuanta al querer encontrar la productividad perdida.

```
Scores de validación cruzada: [0.9999634  0.9999128  0.99886483 0.99951539 0.9682334 ]
Promedio de validación cruzada: 0.9932979641543559
```

De igual forma usamos una validación cruzada para ver si nuestros parámetros del entrenamiento son cercanos a nuestra validación cruzada.

```
usuario = {
    'Addiction Level': 2.94,
    'Satisfaction': 4,
    'Self Control': 8
    }

# Obtener las columnas del conjunto de datos de entrenamiento
columnas_entrenamiento = X_train.columns

# Convertir el diccionario a un DataFrame con el orden correcto de columnas
usuario_df = pd.DataFrame([usuario], columns=columnas_entrenamiento)

# Predecir la calificación final utilizando el modelo entrenado
prediccion_calificacion = rf_model.predict(usuario_df)

# Mostrar la predicción
print(f"El impacto de productividad perdida final predicha para el usuario es: {prediccion_calificacion[0]:.2f}")
```

```
El impacto de productividad perdida final predicha para el usuario es: 6.00
```

Como último paso, ponemos a prueba nuestro Random Forest para que empiece a hacer predicciones con datos que nosotros le demos de nuestras 3 variables; en este caso nosotros le pusimos una 2.94 para “Addiction Level”, 4 para “Satisfaction” y 8 en “Self Control”, arrojándonos que el **impacto en la productividad perdida será de 6.**

Comprobando con nuestro dataset que esto es correcto.

ProductivityLoss	Satisfaction	DeviceType	OS	Watch Time	Self Control	Addiction Level
3.0	7.000000	Tablet	2.0	9:00 PM	5.0	5.000000
6.0	4.000000	Smartphone	1.0	2:00 PM	8.0	2.944269

4. CONCLUSIONES

Después del exhaustivo análisis podemos observar que el objetivo planteado en un principio, dió el resultado esperado, ayudando a encontrar personas con una fuerte dependencia y por ende, ayudar a que la productividad en su día a día vaya mejorando.
No es el único estudio que podemos hacer; como ya lo había mencionado antes, puede ser estudiado para cuestiones de marketing encontrando patrones que tiene la gente al hacer uso de las redes sociales.

5. REFERENCIAS

Zeesolver. (n.d.). *Dark Web* [Dataset]. Kaggle. Retrieved December 2, 2024, from https://www.kaggle.com/datasets/zeesolver/dark-web