



Universidad
del Caribe

2000

CANCUN, QUINTANA ROO, MÉXICO

CONOCIMIENTO Y CULTURA PARA EL DESARROLLO HUMANO

Cálculo de entropía con repositorios.

ASIGNATURA:

Teoría de la Información

Realizado por:

ABDIEL GABRIEL HAU TUN

Matricula: 200300588

ANGEL JESUS JIMENEZ BE

Matricula: 200300603

Docente:

Ismael Domínguez Jiménez

1.- Obtener un repositorio abierto de datos

Repositorio usado: [Predict People Personality Types](#)

El repositorio abierto de datos en Kaggle, titulado Predict People Personality Types (Predecir Tipos de Personalidad de las Personas) es un conjunto de datos que está alojado por Stealth Technologies y está diseñado para proyectos de análisis de datos y aprendizaje automático (Machine Learning) orientados a la predicción de tipos de personalidad.

Descripción del contenido del dataset:

Variables incluidas: El dataset contiene diversas características relacionadas con rasgos de personalidad como la apertura a nuevas experiencias, la extroversión, la responsabilidad, la amabilidad y la estabilidad emocional, junto con variables demográficas y de comportamiento.

Uso esperado: La finalidad es proporcionar a los usuarios un recurso para crear modelos que puedan clasificar o predecir los tipos de personalidad, con aplicaciones potenciales en recursos humanos, marketing personalizado, orientación académica, entre otros.

Formato: Los datos se encuentran disponibles en formatos tabulares (generalmente CSV), lo que facilita su uso para algoritmos de Machine Learning en plataformas como Python o R.

2.- Preparar los datos

```
# Importar las bibliotecas necesarias
import pandas as pd
import numpy as np

# Paso 1: Preparar los datos
# Leer los datos del archivo CSV
data = pd.read_csv("data.csv")
print(data.head(5))
```

Para la preparación de los datos, se importan las bibliotecas Pandas y NumPy para el análisis de datos. Se lee un archivo CSV llamado "data.csv" (el cual corresponde a repositorio descargado) y se almacena en un data frame para trabajar con los datos de manera eficiente.

Se imprimen las primeras 5 filas para ir conociendo la estructura del dataframe.

```
a informacion/entropia.py"
  Age  Gender  Education  Introversion Score  Sensing Score  Thinking Score  Judging Score  Interest Personality
0  19.0   Male         0         9.47080      7.141434      6.03696      4.360278      Unknown      ENFP
1  27.0  Female         0         5.85392      6.160195      0.80552      4.221421      Sports      ESFP
2  21.0  Female         0         7.08615      3.388433      2.66188      5.127320      Unknown      ENFP
3  28.0   Male         0         2.01892      4.823624      7.30625      5.986550      Others      INTP
4  36.0  Female         1         9.91703      4.755080      5.31469      4.677213      Technology  ENFP
```

De manera adicional, se imprime si existen valores nulos en cada columna para verificar la calidad de los datos y asegurarnos de que no haya valores faltantes que puedan afectar el análisis.

```
Valores faltantes por columna:
Age          0
Gender       0
Education    0
Introversion Score  0
Sensing Score  0
Thinking Score  0
Judging Score  0
Interest     0
Personality  0
dtype: int64
```

3.- Determinar de los datos obtenidos que se quiere conocer (de que se desea obtener la probabilidad)

	Age	Gender	Education	Introversion Score	Sensing Score	Thinking Score	Judging Score	Interest	Personality
0	19.0	Male	0	9.47080	7.141434	6.03696	4.360278	Unknown	ENFP
1	27.0	Female	0	5.85392	6.160195	0.80552	4.221421	Sports	ESFP
2	21.0	Female	0	7.08615	3.388433	2.66188	5.127320	Unknown	ENFP
3	28.0	Male	0	2.01892	4.823624	7.30625	5.986550	Others	INTP
4	36.0	Female	1	9.91703	4.755080	5.31469	4.677213	Technology	ENFP

A partir del DataFrame obtenido, elegimos calcular la probabilidad de las diferentes clases que componen las categorías de Personality e Interest:

- **Personality** : Queremos conocer la distribución de las distintas personalidades (e.g., ENFP, ESFP, INTP) y calcular la probabilidad de que una persona seleccionada al azar pertenezca a cada tipo de personalidad en el conjunto de datos.
- **Interest**: También deseamos analizar la distribución de los intereses y calcular la probabilidad de que una persona seleccionada al azar tenga un interés específico.

4.- Identificar las clases

Para la identificación de las clases, contamos la frecuencia de cada columna en las categorías de Personalidad e Interés para entender cómo se distribuyen los datos. Esto nos ayuda a identificar cuántas veces aparece cada tipo de personalidad e interés en el conjunto de datos.

Para ello, utilizamos el método `value_counts()` en las columnas relevantes del DataFrame. Esto generará dos series que muestran las diferentes clases y la cantidad de ocurrencias de cada una.

Clases - Personality:

```

Personality
ENFP    34404
INTP    24718
INFP    24711
INTP    17132
ESFP     4832
ENFJ     3883
ISFP     3456
ESTP     3334
INFJ     2919
ENTJ     2783
ISTP     2390
INTJ     1920
ESFJ      554
ESTJ      392
ISFJ      371
ISTJ      262
Name: count, dtype: int64

```

Clases - Interest:

```

Interest
Unknown    48835
Arts       25489
Others     21733
Technology  19103
Sports     12901
Name: count, dtype: int64

```

Esta información es esencial para el análisis posterior, ya que nos permitirá calcular las probabilidades de cada clase y la entropía del sistema, brindando así una mejor comprensión de la distribución de las variables en el conjunto de datos.

5.- Calcular las probabilidades de las clases

Para calcular la probabilidades de cada clase en las categorías de Personalidad e Intereses, utilizamos la siguiente fórmula:

$$Probabilidad = \frac{Frecuencia\ de\ la\ clase}{Total\ de\ datos}$$

Aplicamos esta fórmula a las frecuencias obtenidas previamente:

- **Personality:** Se calcula dividiendo la frecuencia de cada tipo de personalidad entre el total de registros en la columna Personality, en código queda:
“frecuencias_personality / len(data_personality)”
- **Interest:** Se calcula dividiendo la frecuencia de cada interés entre el total de registros en la columna Interest, en código queda “frecuencias_interest / len(data_interest)”

```
# Paso 4: Calcular las probabilidades de las clases
# Calcular la probabilidad de cada clase dividiendo su frecuencia por el total
probabilidades_personality = frecuencias_personality / len(data_personality)
probabilidades_interest = frecuencias_interest / len(data_interest)
```

6.- Calcular la entropía del sistema

Para calcular la entropía de las clases en las categorías de Personalidad e Intereses, utilizamos la siguiente fórmula:

$$H = -\sum_{i=1}^N P(i) \log P(i)$$

Donde:

- H es la entropía del sistema.
- P_i es la probabilidad de cada clase
- n es el número total de clases.

La entropía proporciona una medida de la incertidumbre o el desorden en un conjunto de datos. Una entropía alta indica una mayor diversidad entre las clases, lo que sugiere que las clases están más uniformemente distribuidas. Por el contrario, una entropía baja implica que algunas clases son mucho más frecuentes que otras, lo que indica una concentración en ciertos tipos de personalidad o intereses.

Para el cálculo de la entropía, se creó una función llamada "calcular_entropia", que realiza los siguientes pasos:

1. **Cálculo del logaritmo:** La función toma las probabilidades de las clases y calcula su logaritmo en base 2.
2. **Producto de probabilidades y logaritmos:** Multiplica cada probabilidad por su logaritmo correspondiente, lo que permite determinar la contribución de cada clase a la entropía total.
3. **Suma de productos:** Suma todos los productos obtenidos en el paso anterior y multiplica por -1 para obtener el valor final de la entropía.
4. **Resultados:** Finalmente, la función devuelve tanto los productos como el valor total de la entropía, que se utilizará para analizar la distribución de las clases en el conjunto de datos.

```
# Paso 5: Calcular la entropía del sistema
# Definir una función para calcular la entropía basada en las probabilidades
def calcular_entropia(probabilidades):
    log_probabilidades = np.log2(probabilidades)
    productos = probabilidades * log_probabilidades
    entropia = -productos.sum()
    return productos, entropia
```

7.- Resultados

a) Las primeras 5 filas

	Age	Gender	Education	Introversion Score	Sensing Score	Thinking Score	Judging Score	Interest	Personality
0	19.0	Male	0	9.47080	7.141434	6.03696	4.360278	Unknown	ENFP
1	27.0	Female	0	5.85392	6.160195	0.80552	4.221421	Sports	ESFP
2	21.0	Female	0	7.08615	3.388433	2.66188	5.127320	Unknown	ENFP
3	28.0	Male	0	2.01892	4.823624	7.30625	5.986550	Others	INTP
4	36.0	Female	1	9.91703	4.755080	5.31469	4.677213	Technology	ENFP

El conjunto de datos muestra información detallada sobre diversos individuos, combinando atributos demográficos (como edad y género), niveles educativos, puntuaciones en dimensiones psicológicas, y categorías de interés y personalidad.

Intereses y Personalidades:

Los intereses se clasifican en categorías como Sports (Deportes), Technology (Tecnología) y Others. Sin embargo, algunos individuos tienen el interés clasificado como Unknown.

Las personalidades indican el tipo de personalidad del individuo, según una clasificación como ENFP, INTP, o ESFP. Estos tipos corresponden a la teoría de

Myers-Briggs, que clasifica las personalidades en combinaciones de indicadores (e.g., ENFP: Extrovertido, Intuitivo, Emocional, Perceptivo).

b) Valores faltantes

```
4 30.0 Female 1
Valores faltantes por columna:
Age          0
Gender       0
Education    0
Introversion Score  0
Sensing Score  0
Thinking Score  0
Judging Score  0
Interest     0
Personality  0
dtype: int64
Classes - Personality:
```

Significado de los valores Para cada columna, se muestra el número de valores faltantes. En este caso, todas las columnas tienen 0 valores faltantes, lo que significa que no hay datos ausentes en ninguna de las columnas del DataFrame.

Calidad de los Datos: La ausencia de valores faltantes en todas las columnas es una señal de que los datos están completos y no requieren imputación o limpieza en este aspecto.

Preparación para Análisis: Este resultado sugiere que los datos están en buen estado y listos para realizar su análisis, sin necesidad de preocuparse por problemas derivados de datos incompletos.

c) Clases en Personality

```
31
Classes - Personality:
Personality
ENFP    34404
ENTP    24718
INFP    24711
INTP    17132
ESFP     4832
ENFJ     3883
ISFP     3456
ESTP     3334
INFJ     2919
ENTJ     2783
ISTP     2390
INTJ     1920
```


Muestra un conteo de las clases en la categoría "Personality" del DataFrame. Cada fila representa un tipo de personalidad basado en la teoría de Myers-Briggs (MBTI), junto con la cantidad de veces que cada tipo de personalidad aparece en los datos.

Desigualdad en la Frecuencia: Se observa una notable variación en la frecuencia de los tipos de personalidad. Las personalidades ENFP, ENTP e INFP son las más comunes, mientras que ISTJ y ISFJ son las menos frecuentes.

Prevalencia de Perceptivos: Los tipos con el rasgo Perceptivo (P) como ENFP e INFP son más comunes que los tipos con rasgos de Juicio (J) como ESTJ e ESFJ.

d) Intereses

```
Clases - Interest:
Interest
Unknown      48835
Arts          25489
Others        21733
Technology    19103
Sports        12901
Name: count, dtype: int64
```

Se muestra el conteo de las clases en la categoría "Interest" del DataFrame. Cada fila representa una categoría de interés, junto con la cantidad de individuos que pertenecen a cada una de ellas.

Prevalencia del interés "Unknown" Con 48,835 individuos, esta categoría es, con diferencia, la más grande. Esto podría indicar falta de datos o individuos sin un interés claramente definido.

Distribución en los demás intereses: Las Artes son el segundo interés más frecuente, seguido de Others y Technology. Deportes es el interés menos común entre las categorías identificadas.

Posible sesgo en los datos: La gran cantidad de individuos en la categoría "Unknown" podría afectar la capacidad de interpretar patrones claros sobre los intereses. Esto sugiere que podría ser necesario investigar por qué tantos datos aparecen como desconocidos.

e) Probabilidades en personalidad

Probabilidades - Personality:	
Personality	
ENFP	0.268653
ENTP	0.193017
INFP	0.192963
INTP	0.133780
ESFP	0.037732
ENFJ	0.030321
ISFP	0.026987
ESTP	0.026034
INFJ	0.022794
ENTJ	0.021732
ISTP	0.018663
INTJ	0.014993
ESFJ	0.004326
ESTJ	0.003061
ISFJ	0.002897
ISTJ	0.002046
Name: count, dtype: float64	

Se muestra las probabilidades de las clases en la categoría "Personality". Cada fila representa un tipo de personalidad del modelo MBTI (Myers-Briggs Type Indicator),

con su probabilidad relativa calculada. Las probabilidades se obtienen al dividir la frecuencia de cada clase entre el total de observaciones en esta categoría.

Personalidades Dominantes: ENFP, ENTP e INFP son las personalidades más comunes, representando una gran parte del total. Estas personalidades tienden a ser más abiertas, intuitivas y extrovertidas.

Personalidades Menos Comunes: Las personalidades como ISTJ e ISFJ tienen probabilidades muy bajas, indicando que son raras en la muestra.

Diversidad en la Distribución: Aunque algunas personalidades son más comunes que otras, la muestra incluye una amplia gama de tipos. Esto sugiere que el conjunto de datos es diverso, pero con cierta tendencia hacia personalidades e

ENFP: 0.2687 (26.87%) Es la personalidad más común en la muestra, presente en más de una cuarta parte de los datos. extrovertidas y perceptivas. Mientras que ESFP a ISTJ son las personalidades que tienen probabilidades más bajas, como: ISFJ: 0.0029 (0.29%) y ISTJ: 0.0020 (0.20%), la personalidad menos común en la muestra.

f) Probabilidades en Interés

Probabilidades - Interest:	
Interest	
Unknown	0.381342
Arts	0.199038
Others	0.169708
Technology	0.149171
Sports	0.100741
Name: count, dtype: float64	

Se muestran las probabilidades de las clases en la categoría "Interest". Cada fila representa una categoría de interés, con su probabilidad relativa. Estas probabilidades se calculan dividiendo la frecuencia de cada categoría entre el total de registros en esta columna.

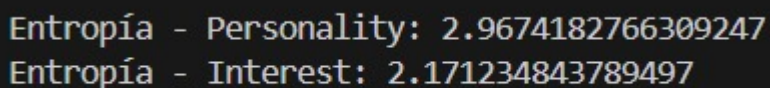
Prevalencia de "Unknown": Casi el 40% de los registros tienen su interés marcado como Unknown. Esto podría reflejar falta de información o un desafío en la recolección de datos sobre los intereses de los individuos.

Distribución de Intereses Identificados: Entre los intereses conocidos, Arts (19.90%) es el más frecuente, seguido de Others (16.97%) y Technology (14.92%) Sports (10.07%) es el interés menos frecuente en esta muestra.

Unknown: 0.3813 (38.13%) Esta categoría representa a los individuos cuyo interés es desconocido o no identificado, siendo la clase más grande en la muestra. Mientras que **Sports: 0.1007 (10.07%)** muestra la menor proporción de la muestra se interesa en deportes.

Tendencias de la muestra: Las artes y la tecnología parecen ser áreas populares de interés, lo que puede reflejar la orientación de las personas en la muestra hacia actividades creativas y tecnológicas. La baja proporción de interés en deportes sugiere que podría ser menos prioritario para esta población.

g) Entropía



```
Entropía - Personality: 2.9674182766309247
Entropía - Interest: 2.171234843789497
```

Entropía - Personality: 2.9674

El valor de entropía para Personality indica una alta diversidad en las clases, lo que sugiere que las personalidades están relativamente equilibradas en la muestra, dificultando la predicción del tipo de personalidad de un individuo al azar. La amplia distribución, con varias categorías significativas como ENFP, INFP y ENTP, contribuye a este elevado valor de entropía.

Entropía - Interest: 2.1712

El valor más bajo de entropía para Interest indica una menor diversidad en comparación con Personality, lo que refleja que la distribución de los intereses está más concentrada en algunas categorías, especialmente en Unknown (38.13%). Esta menor entropía sugiere que es más fácil predecir el interés de un individuo, ya que los datos están dominados por un número reducido de categorías principales.

Comparación e Interpretación:

Mayor Entropía en Personality: La entropía más alta en "Personality" indica que las personalidades están distribuidas de forma más equitativa entre muchas categorías, lo que sugiere una mayor incertidumbre al intentar predecir una personalidad específica.

Menor Entropía en Interest: La entropía más baja en "Interest" muestra que los intereses están más concentrados, especialmente en categorías como Unknown. Esto reduce la incertidumbre, ya que algunas categorías dominan la muestra.

h) Cálculo de Entropía

Cálculo de Entropía - Personality:				
	Clase	Probabilidad	$\log_2(\text{Probabilidad})$	Producto
0	ENFP	0.268653	-1.896183	-0.509416
1	ENTP	0.193017	-2.373197	-0.458068
2	INFP	0.192963	-2.373606	-0.458017
3	INTP	0.133780	-2.902066	-0.388238
4	ESFP	0.037732	-4.728067	-0.178400
5	ENFJ	0.030321	-5.043516	-0.152927
6	ISFP	0.026987	-5.211584	-0.140646
7	ESTP	0.026034	-5.263433	-0.137031
8	INFJ	0.022794	-5.455213	-0.124345
9	ENTJ	0.021732	-5.524046	-0.120048
10	ISTP	0.018663	-5.743677	-0.107194
11	INTJ	0.014993	-6.059581	-0.090850
12	ESFJ	0.004326	-7.852729	-0.033971
13	ESTJ	0.003061	-8.351762	-0.025565
14	ISFJ	0.002897	-8.431196	-0.024426
15	ISTJ	0.002046	-8.933049	-0.018276

Cálculo de Entropía - Interest:				
	Clase	Probabilidad	$\log_2(\text{Probabilidad})$	Producto
0	Unknown	0.381342	-1.390844	-0.530387
1	Arts	0.199038	-2.328885	-0.463536
2	Others	0.169708	-2.558872	-0.434262
3	Technology	0.149171	-2.744960	-0.409469
4	Sports	0.100741	-3.311276	-0.333581

Distribución en Personality: La entropía refleja una amplia diversidad de personalidades, con varias categorías importantes como ENFP, INFP y ENTP. Esta amplia distribución resulta en una entropía relativamente alta.

Distribución en Interest: La entropía es más baja para los intereses, ya que la categoría Unknown domina significativamente (38.13%), lo que indica una concentración de datos en pocas categorías principales.

Personality tiene mayor entropía que Interest, lo que sugiere que la distribución de personalidades es más diversa. Mientras que en Interest, los datos están más concentrados en unas pocas categorías, especialmente en "Unknown".

Sección 1: Cálculo de Entropía - Personality

ENFP tiene la mayor probabilidad (0.2687) y, por lo tanto, aporta más al valor total de la entropía. Las personalidades menos frecuentes, como ISTJ (0.0020), contribuyen menos a la entropía.

Sección 2: Cálculo de Entropía - Interest

La categoría Unknown tiene la mayor probabilidad (0.3813) y aporta significativamente al valor total de la entropía. Sports tiene la probabilidad más baja (0.1007) y, en consecuencia, su contribución a la entropía es menor.

Referencias

Serrano, L. (2018, October 26). Shannon Entropy, Information Gain, and Picking Balls from Buckets. Medium.
<https://medium.com/udacity/shannon-entropy-information-gain-and-picking-balls-from-buckets-5810d35d54b4>

Predict people personality types. (2024, September 14).
<https://www.kaggle.com/datasets/stealthtechnologies/predict-people-personality-types>

Palacios, I. N. (2022, March 30). Entendiendo la entropía en la información - I. N. Palacios - Medium.
<https://kaldt-slange.medium.com/entendiendo-la-entrop%C3%ADa-en-la-informaci%C3%B3n-6bb434cdc377>

Calderón, J. C. H. (2024, March 24). La entropía de Shannon como medida de la incertidumbre y la información potencial. Parte I. Medium.
<https://medium.com/@JuanEnredado/la-entrop%C3%ADa-de-shannon-como-medida-de-la-incertidumbre-parte-i-6a12c4d5d36>

Westreicher, G. (2022, 24 noviembre). Cálculo de probabilidades Qué es, definición y concepto. Economipedia.
<https://economipedia.com/definiciones/calculo-de-probabilidades.html>