Las relaciones de un todo

Ordaz Guillén Fabián

Junio 2024

1. Introducción

A diferencia de las llamadas "ciencias duras" aquellos ámbitos donde la humanidad es el protagonista han presentado una misma dificultad; la variabilidad del ser humano y su amplio espectro de posibles respuestas aún a situaciones ya conocidas. Por ese motivo se han diseñado varias herramientas a lo largo de la historia que buscan generar grupos que permitan describir, para "bien o para mal" [Dattner(2004)], mediante lo que es y no es el ser humano, responder la pregunta ¿cual es la mejor manera de resumir a un individuo?.

Algunas de estas herramientas son test como el; *Inventario Multifásico de Personalidad de Minnesota (MMPI), 16 Factores de Personalidad (16PF), Cuestionario de Personalidad de Eysenck (EPQ),Test de Rorschach* o bien aquel en el que profundizaremos el ***Inventario de Personalidad NEO (NEO PI-R)***.

El NEO PI-R (1980, Paul T. Costa Jr. y Robert R. McCrae) tiene sus fundamentos en la teoría de los cinco grandes (*The Big Five*), que busca medir los cinco principales dominios de la personalidad (Neuroticismo, Extraversión, Apertura a la Experiencia, Amabilidad y Estabilidad emocional) y sus 30 facetas específicas. Es un modelo muy robusto y respetado dentro de la comunidad tanto en fines clínicos preventivos o reactivos, académicos y laborales, valorado por su capacidad para capturar una amplia gama de comportamientos y características de personalidad con un alto grado de validez. Dando una serie de reactivos en los que en una escala del 1 (completamente en desacuerdo) al 5 (muy de acuerdo) se presentan (habitualmente) 48 situaciones, el usuario debe elegir generando una serie de notas que se someten a alphas de evaluación ya preestablecidos para definir una çonclusión", de recordar que no hay una respuesta correcta o esperada, simplemente es la convicción y naturalidad quien indica que se elige en cada área de la personalidad, así en cada una de ellas.

Ahora bien hemos mencionado ya la duda sobre el modo "más óptimo" de como podemos resumir a un individuo, mediante de una serie de incisos que si bien transparentes, bajo una construcción que nos arrojará información confiable y fundamentados siguen siendo individuales, nos referimos a que tenemos 48 incisos de 5 campos mutuamente excluyentes, es decir, la prueba y sus bases teóricas no están construidas para devolver o tener alguna relación entre áreas y aunque existen cualidades tales como un alto índice de neuroticismo que podría

tener tendencias que también afectan su extraversión o amabilidad, estas últimas se calificarán únicamente con los puntajes de sus escalas así que ¿pudieran estas individualidades esconder un patrón de grupo? mejor dicho ¿un alto A siempre venir causado por un bajo B que consecuente a un C moderado?

2. Descripción de los datos

Los datos evaluados fueron reunidos durante 2 años (2016-2018) mediante un sitio en línea de personalidad. Es importante mencionar que sus participantes fueron avisados que las respuestas podrían ser grabadas y usadas, así como se les pidió confirmaran su aprobación para el uso ético de los mismos.

La respuesta por cada área de la personalidad se agrupo en columnas abreviadas (EXT = extraversión, EST = estabilidad emocional , AGR = amabilidad , CSN = neuroticismo, OPN = apertura a nuevas experiencias) y enumero con un consecutivo después de cada grupo de siglas, Una de las debilidades mas fuertes de este modelo de interpretación es su larga duración y la facilidad de perder a los participantes, así que buscando que se eligiera la mayor cantidad de sentencias en cada área se les mostró a los usuarios de una por cada área a la vez. También generó una marca en milisegundos del tiempo que paso cada usuario en cada reactivo, las dimensiones del dispositivo donde se tomó la prueba (en pixeles) así como una aproximación de la ciudad, latitud y longitud de origen de información. Quizá para otro tipo de enfoques o resolver interrogantes mas adelante podamos usar todo el conjunto de datos pero para este primer acercamiento solo ahondamos en las 50 preguntas con base en los 5 grandes.

Como ya hemos hablado se trata de un conjunto de datos discretos de 50 columnas por 1,015,341 filas, tomando valores del 1 al 5. Para un correcto uso de la información trataremos la misma en búsqueda y eliminación de valores atípicos o vació que nos impidan seguir con el análisis, tomaremos una muestra de 400 datos y buscando una similitud mas exacta con el modelo NEO PI-R, remplazaremos el 2 y 4 por 1 y 5 dejando únicamente 1, 3 y 5 como datos.

Dominio	**Preguntas (total)**	**Respuestas (c/u)**	**Nulos**	**No nulos**
EXT	10	874,434	0	874,434
EST	10	874,434	0	874,434
AGR	10	874,434	0	874,434
CSN	10	874,434	0	874,434
OPN	10	874,434	0	874,434

Cuadro 1: Resumen de datos tratados a analizar

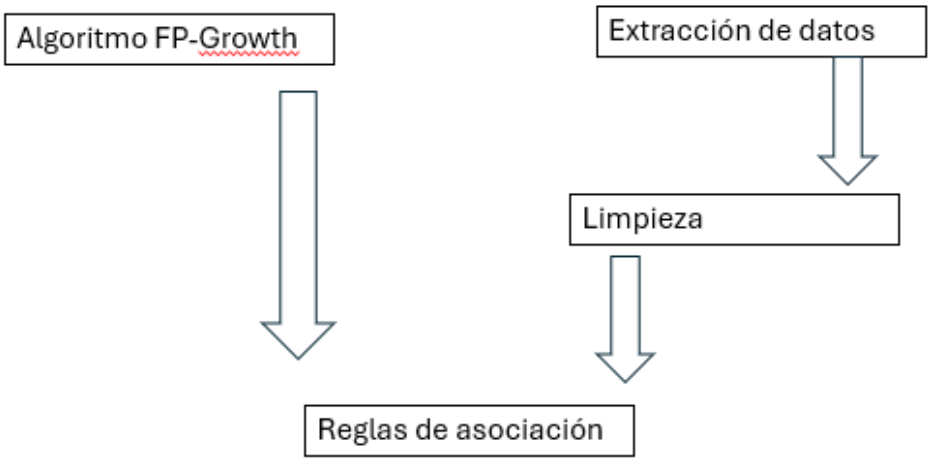

Figura 1: Mapa de calor: Dominios de la personalidad

3. Metodología

3.1. FP-Growth

Como antecedente, de la necesidad de la minería de datos dio inicio a las reglas de asociación que muestra relaciones aparentemente no están "unidas" dentro de los datos. Divididos teóricamente en en dos partes, las reglas de asociación vienen de los conceptos de antecedente (if) y el consecuente (then), usando el criterio de soporte y confianza. El soporte indica que tan frecuente aparece un dato y la confianza es el número de "si-entonces"que fueron encontrados.

El algoritmo empieza generando una serie de árboles de datos de patrones frecuentes, luego, esta misma base de datos se divide en datos consecuentes con patrones de frecuencia único [Sidhu et al.(2014)Sidhu, Meena, Nawani, Gupta, and Thakur]

La figura 1 muestra el proceso de implementación del algoritmo

3.2. Bosques aleatorios

Los bosques aleatorios combinan muchos árboles de decisión y agrega predicciones por promedio, esto muestra mejor desempeño que la singularidad de los árboles aún para grandes números de observaciones y nos otorga medidas de importancia. Al final, según los datos de entrada cada árbol nos dará un resultado y al integrar múltiples resultados nos brindará el resultado del bosque

En la creación de bosques aleatorios, los árboles de decisión siguen tres individualidades

1.- Se usan métodos (baggin) para seleccionar datos únicos en cada árbol

2.- Las características que tendrán dichos árboles también se seleccionan con técnicas de muestreo. Solo se usarán una cantidad de características M, tal que N es la totalidad de ellas y M es menor que N

3.- Todos los árboles crecerán de manera libre conforme sea necesario

Al final el resultado será interpretado usando técnicas de mayoría para los problemas de clasificación y con métodos de promedio matemático para problemas de regresión [Liu(2014)]

4. Aprendizaje no supervisado

4.1. Planteamiento del problema

Siguiendo un enfoque de estudio distinto a lo habitual, no se buscaron respuestas dentro de estándares psicológicos, aunque si nos documentamos al respecto para una mejor interpretación, contraste y debate de resultados [Costa and McCrae(1992)], sino una causalidad de A entonces B dentro de una misma área así como entre ellas y sus posibles combinaciones

4.2. Metodología

Dentro de los modos que tenemos hoy para analizar patrones dentro de los modelos de aprendizaje no supervisados tenemos el algoritmo **FP-Growth**(FPG en adelante), este utilizado comúnmente para revisar y encontrar patrones de asociación de eventos, por ejemplo en una lista de supermercado para cuantificar la cantidad de veces que un alimento esta en la misma compra que otro.

Funciona escaneando dos veces la base de datos, la primera nos entrega el número de veces que ocurre un evento en todas las transacciones, entonces, marcamos un mínimo significante y aquellos que no lleguen a tal frecuencia se descartan y ordenamos de mayor a menor los datos conservados.

Supongamos que tenemos las siguientes transacciones en orden original:

1.- [A, B, C]
2.- [A, C]
3.- [A, B, D]
4.- [B, C]
5.- [A, C, D]

Y pensando que el umbral mínimo de soporte es 2 tenemos
Frecuencias: A: 4 B: 3 C: 4 D: 2
Todos los artículos cumplen con el umbral mínimo de soporte así que los ordenamos
Orden: A (4), C (4), B (3), D (2)

Transacciones ordenadas por frecuencia

1.- [A, C, B]
2.- [A, C]
3.- [A, B, D]
4.- [C, B]
5.- [A, C, D]

Para la segunda pasada se empieza la construcción de un árbol, pudiera verse como uno genealógico, con cada evento se crea una rama, si el evento ya existe se agrega un contador, si es la primera vez se agrega un nodo para una nueva rama
Para la primera transacción [A, C, B]:

$$(A:1)-(C:1)-(B:1)$$

Para la segunda transacción [A, C]:

$$(A:2)-(C:2)-(B:1)$$

Para la tercera transacción [A, B, D]:

$$\begin{array}{l}(A:3)-(C:2)-(B:1)\\ \qquad\quad\ \backslash \\ \qquad\qquad\qquad\quad (B:2)-(D:1)\end{array}$$

Para la cuarta transacción [C, B]:

$$\begin{array}{l}(A:3)-(C:2)-(B:1)\\ \qquad\quad\ \backslash \\ \qquad\qquad\qquad\quad (B:2)-(D:1)\\ (C:1)-(B:1)\end{array}$$

Para la quinta transacción [A, C, D]:

$$\begin{array}{l}(A:4)-(C:3)-(B:1)\\ \qquad\quad\ \backslash \\ \qquad\qquad\qquad\quad (B:2)-(D:1)\\ \qquad\qquad\qquad\quad\ \backslash \\ \qquad\qquad\qquad\qquad\quad (D:1)\\ (C:1)-(B:1)\end{array}$$

Ahora para el análisis necesitamos contemplar los caminos (secuencias de datos desde la raíz hasta cualquier nodo en el árbol FPG) y el soporte (cantidad de transacciones o eventos que contienen esa secuencia exacta de datos). Por ejemplo veamos el evento [A,C], tenemos 3 eventos que tienen dicha secuencia (de los eventos ordenados por frecuencia tendríamos el 1, 2 y 5) y el soporte de [A,C] se define como las 3 transacciones sobre el número total de eventos 3/5 interpretado como el evento [A,C] aparece en el 60 % de todas las transacciones

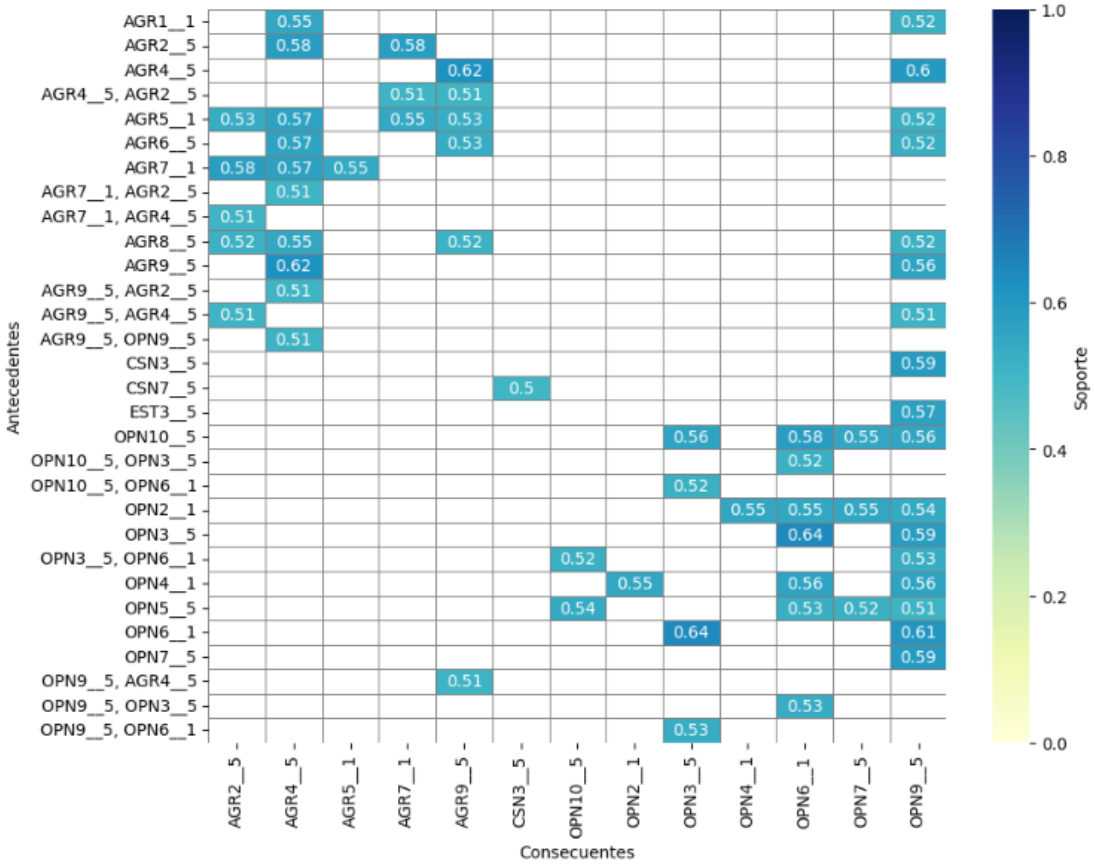

Figura 2: Mapa de calor: Dominios de la personalidad

De este modo, podemos obtener la frecuencia (soporte) con el que los eventos causan o se relacionan con la aparición de otro.

4.3. Resultados

Luego de aplicar FPG se obtuvo una relación del número de pregunta de cada dominio, presentando los resultados en la Figura 2

No solo podemos ver la relación entre los dominios globales de personalidad, sino también, la frecuencia con la que preguntas y el nivel de agrado que se eligió anteceden a otra pregunta y su propio nivel de agrado, podríamos de este modo y con un apoyo entendido en psicología generar un nuevo diseño de NEO PI-R mas eficiente, con menos preguntas que ayuden a los participantes con una prueba mas corta que evite una perdida de atención o foco.

5. Aprendizaje supervisado

5.1. Planteamiento del problema

La variedad dentro de la personalidad es el resultado de multiples factores [Tintaya Condori(2019)] generalmente fuera del control del individuo, si bien los gustos personales y el como reacciona cada quien a las circunstancias que afrontamos pueden ser manipulables o interfieren en menor medida, hay algunos en el entorno macrosocial que están tan normalizados que no distinguimos como nos influyen. Dicho de otro modo, pudiéramos encontrar en otras sociedades constructos que vistos desde nuestra normalidad parecieran ajenos o no entendamos.

Por eso vale la pena detenernos a pensar cuanto de lo que somos respecto a personalidad se desarrolló o se adoptó, a que grado lo compartimos, habremos desarrollado una tendencia personal o existe una tendencia que nos predispone a una respuesta más habitual que en otras zonas

5.2. Metodología

Compuesto por varios arboles de decisión, un **bosque aleatorio** (también conocido como random forest) es un algoritmo de aprendizaje automático que combina la salida de múltiples árboles de decisión para llegar a un resultado, manejando términos de clasificación y regresión. Empezando por la singularidad de un solo árbol que empieza respondiendo preguntas simples, preguntas de respuesta si-no, estas respuestas constituyen los nodos de decisión del árbol como un modo para dividir los datos.

Como inconveniente de los árboles individuales, suelen sobreajustarse (overfitting), ya que al ser entrenados con un set de datos específicos capturan el ruido y las particularidades de los datos, puede ser tan complejo en términos de nodos y ramas que clasifica con casi perfecto rendimiento los datos de entrenamiento pero para datos nuevos no suele tener un desempeño útil.

Para el desempeño de un bosque de decisión se seleccionan varias muestras aleatorias para crear subconjuntos de datos de entrenamiento. Esta selección del mejor split se basa en la reducción de impurezas donde

$$\text{Impureza del nodo padre - } \sum_j \frac{n_j}{n} \text{ Impureza del nodo hijo}$$

De este modo cada árbol se construye de modo independiente a los otros, en lugar de considerar todas las características se seleccionan de manera aleatoria un subconjunto de ellas

$$\text{Importancia de la característica} = \frac{1}{T} \sum_{t=1}^{T} \text{ Importancia de la característica en el árbol t}$$

Donde T es el número de árboles en el bosque

La metodología de los bosques aleatorios se basa en la aleatorización y el ensamblado para crear un modelo que sea más preciso y robusto que un solo árbol de decisión. Al combinar múltiples árboles entrenados en diferentes subconjuntos de datos y características, los bosques aleatorios pueden capturar patrones más generalizables y reducir el riesgo de sobreajuste.

5.3. Resultados

Después de aplicar el bosque aleatorio en la muestra de usuarios que aplicaron el test se obtuvieron los resultados del Cuadro 2

Precisión	**Sensibilidad**	**Puntuación f-1**	**Soporte**	**Continente**
0.8	0	0.01	1,578	África
0	0	0	1	Antártica
0.59	0.43	0.5	15,730	Asia
0.57	0.92	0.7	32,850	Europa
0.39	0.01	0.03	11,053	Oceanía
0.75	0	0	3,464	América

Cuadro 2: Resultados del bosque aleatorio

Debido a las diferencias de tamaño entre la cantidad de usuarios de cada continente se ve la misma diferencia de desempeño (puntuación f-1). El desempeño del continente con mayor representación es del 70 % por lo que podemos concluir que de cada 10 usuarios que, de acuerdo a sus características (respuestas tomadas), correspondían a alguien residente del continente Europeo 7 si o eran, clasificándolos de modo correcto.

6. Métricas de desempeño

Las métricas de desempeño son medidas que se utilizan para evaluar la calidad y eficacia de un modelo, estas dependerán del modelo en cuestión para elegir la más adecuada

En el caso de un modelo de clasificación, el **F1-score** combina la precisión y sensibilidad del modelo en una sola métrica, sirve para clasificar adémas aquellos valores que se creyeron positivos pero no lo fueron (falsos positivos) y los que se creyeron negativos pero no lo fueron (falsos negativos)

Usa los conceptos de:

Precisión: Proporción de verdaderos positivos (TP) entre el total de predicciones positivas (TP + FP).

$$\text{Precisión} = \frac{TP}{TP + FP}$$

Sensibilidad (Recall): Mide la proporción de verdaderos positivos entre todas

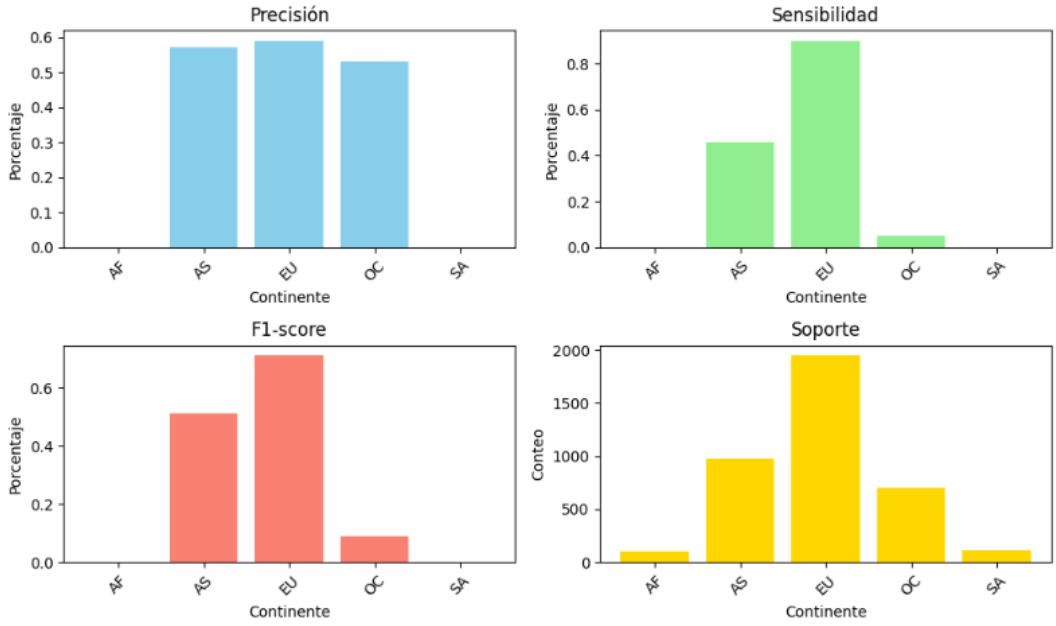

Figura 3: Gráficos de la métrica F1-score

las instancias realmente positivas. Penaliza los falsos negativos.

$$\text{Sensibilidad} = \frac{TP}{TP + FN}$$

F1-score:Combina precisión y recall en una sola métrica mediante la media armónica, proporcionando un balance entre ambos.

$$\text{F1} = 2\frac{(Precisión)(Sensibilidad)}{Precisión + Sensibilidad}$$

En la Figura 3, vemos los resultados del F1-score para el modelo de bosque aleatorio previamente explicado ahora de manera gráfica. En este punto solo añadimos y enfatizamos la idea de que el grupo con muestra mas grande como es Europa, es el que tiene mejor rendimiento en sus métricas con lo que podemos aseverar que el modelo cumple su función

Referencias

[Dattner(2004)] Ben Dattner. El uso y mal uso de los tests de personalidad. *Suplemento Selección de Personal. New York*, 2004.

[Sidhu et al.(2014)Sidhu, Meena, Nawani, Gupta, and Thakur] Shivam Sidhu, U Kumar Meena, Aditya Nawani, Himanshu Gupta, and Narina Thakur.

Fp growth algorithm implementation. *International Journal of Computer Applications*, 93(8):6–10, 2014.

[Liu(2014)] Yingchun Liu. Random forest algorithm in big data environment. *Computer modelling & new technologies*, 18(12A):147–151, 2014.

[Costa and McCrae(1992)] Paul T Costa and Robert R McCrae. *Neo personality inventory-revised (NEO PI-R)*. Psychological Assessment Resources Odessa, FL, 1992.

[Tintaya Condori(2019)] Porfidio Tintaya Condori. Psicología y personalidad. *Revista de investigación psicológica*, (21):115–134, 2019.