

Minería de reglas de asociación para instrumentos psicométricos

Proyecto de título

**Alumno**: Nicolás Rojas Ulriksen.

**Profesor guía**: John Atkinson.

**Compañía**: eb|Consulting.

**Programa**: Master in Business Analytics.

**Director de programa**: Alexis Montecinos.

Índice

[Resumen Ejecutivo 2](#_Toc92216682)

[Contexto del problema: 3](#_Toc92216683)

[Reglas de asociación 6](#_Toc92216684)

[Desarrollo e implementación 8](#_Toc92216685)

[Datos: 8](#_Toc92216686)

[Conversión de las bases a variables *Dummy*: 9](#_Toc92216687)

[Conversión a variables indicadoras: 10](#_Toc92216688)

[Categorías de cargo: 10](#_Toc92216689)

[Edad: 11](#_Toc92216690)

[Nivel educativo: 11](#_Toc92216691)

[Puntajes de corte: 12](#_Toc92216692)

[Promedios por faceta: 12](#_Toc92216693)

[Preguntas: 12](#_Toc92216694)

[Minería de reglas de asociación: 15](#_Toc92216695)

[Conclusiones 24](#_Toc92216696)

[Bibliografía y documentación relacionada 27](#_Toc92216697)

[Anexos: 29](#_Toc92216698)

# Resumen Ejecutivo

Con más de 10 años de experiencia en consultoría de recursos humanos, eb|Consulting se conforma por un equipo multidisciplinario y altamente especializado en la consultoría organizacional tanto privada como pública. Con el objetivo de entregar a sus clientes soluciones estratégicas, y sustentadas en evidencia, eb|Consulting se apoya en el enfoque *Evidence-Based Management* [[1]](#footnote-1) (2006, Pfeffer y Sutton) y en el *Scientist Practitioner Model* [[2]](#footnote-2).

El estudio de las relaciones inter-faceta de la personalidad, tradicionalmente, consta de distintas metodologías asociadas, principalmente, a relaciones lineales entre las variables. Sin embargo, dada la naturaleza de la personalidad; estas relaciones suelen presentar correlaciones considerablemente bajas, dejando las siguientes incógnitas ¿Cuánta relación entre las facetas es no lineal? Y el siguiente problema: ¿Por dónde empezamos a buscar relaciones?

Con el propósito de sentar una base para futuras investigaciones que respondan la pregunta inicial, el presente informe responde al problema por medio de la minería de reglas de asociación, metodología de aprendizaje automático que busca relaciones de coocurrencia estadística entre las variables de interés, utilizando el algoritmo Apriori [[3]](#footnote-3) a través de RStudio, ocupando las librerías Arules, arulesViz, dplyr, readxl y tidyr. Siendo las primeras dos las encargadas de minar y visualizar las reglas de asociación extraídas, y el resto librerías para el trabajo y la carga de datos.

Ocupando el algoritmo Apriori, se lograron minar más de 35.000 relaciones de coocurrencia (Reglas de asociación) entre las 192 variables involucradas. Cabe destacar que, por efectos de simplicidad de análisis, extensión y representación gráfica, el presente informe se centra únicamente en las 20 reglas de asociación con mayores puntajes en los parámetros *Lift* y *Support*, y que estén relacionadas con la variable BIN\_P\_106 [[4]](#footnote-4). No obstante, miles de las reglas de asociación extraídas cuentan con indicadores atractivos para desarrollar posteriores estudios.

# Contexto del problema:

Los test psicológicos laborales son instrumentos construidos con el propósito de rescatar una muestra del comportamiento de una persona en el ámbito laboral, de una manera estandarizada, para poder hacer comparaciones entre las distintas respuestas, y poder hacer estudios de desempeño sobre las mismas personas en el transcurso del tiempo.

La validación de test psicométricos para la medición de facetas de personalidad requiere de un arduo proceso de investigación que contempla análisis descriptivo e inferencial. En la creación de test psicométricos, como se puede observar en la *Tabla 1*[[5]](#footnote-5), la significancia estadística suele tener valores más bajos que las ciencias aplicadas como la biología molecular o la química.

Valores referenciales para el tamaño del efecto 
de las diferentes pruebas de significación 
estadística.

Tabla 1: Valores referenciales para el tamaño del efecto de las diferentes pruebas de significación estadística.

La correlación es una medida que expresa hasta que punto dos variables están relacionadas linealmente, y como se puede observar en la *Tabla 2*, tanto las correlaciones (*r Linear*), como las relaciones cuadráticas y cúbicas entre las distintas facetas de personalidad medidas es considerablemente baja, presentando una barrera en el estudio del comportamiento humano.

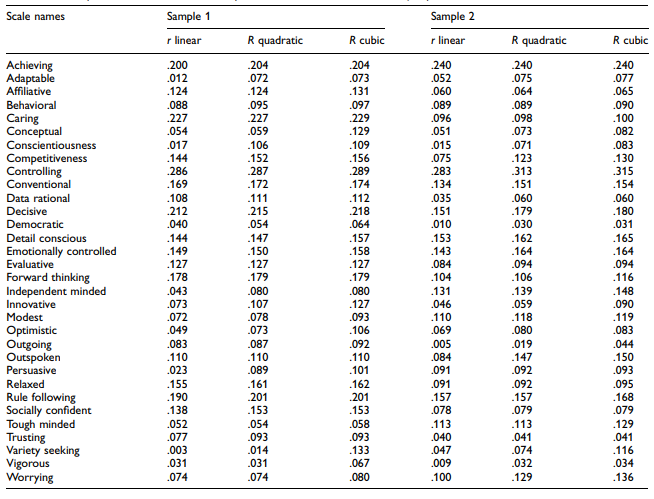


Tabla 2: Relaciones lineales, cuadráticas y cúbicas entre 32 facetas de la personalidad y desempeño laboral.

Dado a que las facetas son bloques fundamentales de la personalidad, que refieren a patrones estables en el comportamiento, sentimientos y pensamientos de los individuos, entender como estas se relacionan entre ellas es sumamente importante para poder:

1. Comprender el comportamiento humano.
2. Predecir el comportamiento y desempeño de las personas.
3. Profundizar en nuestro conocimiento sobre la personalidad.

Ya que hay diversas facetas de personalidad, consolidadas bajo 5 grandes familias – Apertura a las experiencias, Responsabilidad, Amabilidad/Agradabilidad, Extraversión y Estabilidad Emocional – estudiar las relaciones entre cada una de estas requiere un profundo trabajo de análisis sobre las distintas facetas y su comportamiento en conjunto, dificultando el análisis debido a la extensión de dichas facetas.

El problema fundamental que busca resolver este proyecto es reducir la extensión de las mencionadas conexiones inter-faceta, y dirigir futuras investigaciones sobre las relaciones no lineales respecto a las facetas de la personalidad. Debido a la extensa lista de combinaciones que se pueden presentar, pero que pueden llegar a resultados negativos, se introducen las reglas de asociación como un método alternativo para detectar relaciones de coocurrencia entre las distintas preguntas en los test psicométricos, facilitando el posterior estudio de estas.

# Reglas de asociación

Las reglas de asociación son una parte sumamente importante del aprendizaje automático cuyo propósito es buscar relaciones de causalidad (antecedente-consecuencia) entre los items que tienden a ocurrir de forma conjunta. En el mundo de Data Science, las reglas de asociación se ocupan para buscar correlaciones y coocurrencias entre las variables de un data set.

Las reglas de asociación, al buscar relaciones de coocurrencia, se definen como implicancias del tipo “Si X entonces Y”, siendo X el antecedente – A partir de ahora LHS (*Left Hand Side*) – e Y la consecuencia causada por la presencia del antecedente X – A partir de ahora RHS (*Rigth Hand Side*) – Un claro ejemplo de esto es {A, B} => {C}, explicando que cuando A y B ocurren simultáneamente dentro del data set, C puede ocurrir con una probabilidad determinada.

Las reglas de asociación, al medir coocurrencias y buscar relaciones entre los items de la data set, están sujetas a distintos parámetros los cuales permiten rescatar las asociaciones de mayor utilidad. Estos parámetros son:

1. ***Support* (Soporte)**: Mide la frecuencia del ítem en todas las transacciones registradas. Ayuda a identificar las reglas que valen la pena analizar más profundamente.
2. ***Confidence* (Confianza)**: Mide cuan probable es la ocurrencia de un ítem como consecuencia de otro. Entre más cercano a 1, más confianza tenemos de que si hay un ítem aparecerá el otro.
3. ***Lift***: Control para el *Support* (frecuencia) de consecuencias mientras calcula la probabilidad condicional de 𝑌 dado 𝑋. Es cuanto aumenta nuestra probabilidad de tener 𝑌 dado 𝑋. Es el cálculo de la probabilidad de tener 𝑌 dado 𝑋 dividido a las veces que tenemos Y sin saber que tenemos X.

Si tenemos un *Lift* mayor a 1, entonces aumenta la confianza de nuestra regla de asociación.

1. ***Leverage***: Muestra el computo de la diferencia entre la frecuencia observada de cuando 𝑋 e 𝑌 aparecen en conjunto, y la frecuencia que sería esperada si 𝑋 e 𝑌 fuesen independientes. Un valor de 0 indica independencia total.
2. ***Conviction***: Una alta convicción significa que la consecuencia es altamente dependiente del antecedente. En el caso de una confianza de 1, el denominador se convierte en 0 (1-1), por lo que la convicción se define como 𝑖𝑛𝑓. Al igual que el *Lift*, si los ítems son independientes, la convicción es 1.

Teniendo en cuenta lo anterior, la minería de reglas de asociación sobre datos que miden las distintas facetas de personalidad facilita el estudio de las conexiones entre las distintas facetas. De este modo, se abren caminos dirigidos a la investigación sobre las facetas relacionadas extraídas del modelo. Dado a que es un modelo de coocurrencias, el proyecto sienta un precedente para poder definir puntos de origen a nuevas investigaciones sobre la relación y el comportamiento entre dichas facetas. Si bien las reglas de asociación no afirman la relación entre las variables, tenemos factores como el soporte y la confianza que nos permiten rescatar aquellas reglas que pueden servir como precedente para nuevas investigaciones.

# Desarrollo e implementación

## Datos:

La base de datos utilizada para el presente proyecto consta de las respuestas del test psicométrico Personali-T desarrollado por eb|Consulting. El *dataset* nace de la consolidación de las respuestas de los candidatos a distintos procesos de selección realizados por diferentes clientes de eb|Consulting, alcanzando un total de 28,163 observaciones iniciales y 21,393 observaciones post limpieza.

La minería de reglas de asociación exige que las variables estudiadas sean del tipo indicativo (variables “*Dummy*”). Dado a que el objetivo del estudio es encontrar asociaciones entre las facetas de la personalidad, y que la información original está presentada en forma de variables no indicativas, la primera fase del presente proyecto consta del trabajo y conversión de las correspondientes variables a una forma útil para el algoritmo Apriori.

Para la minería de reglas de asociación, se consolidaron las respuestas de la prueba Personali-T [[6]](#footnote-6) de los candidatos a los distintos procesos de selección de los diversos clientes de eb|Consulting. Si bien las pruebas miden las mismas facetas de la personalidad para todos los clientes, se observó tempranamente en el estudio que los ítems evaluados no necesariamente estaban almacenados de la misma forma (i.e.: base de datos 1: variable a = columna 1, base de datos 2: variable a = columna 2), por lo que el primer paso del estudio consistió en la reestructuración de las bases existentes para su posterior consolidación.

Considerando lo anterior, las bases se reordenaron con sus respectivas variables tomando la base de datos con mayor número de observaciones como punto de referencia. Dada la necesidad de corroborar constantemente que las variables fueran iguales, se tomó la decisión de consolidar las distintas bases de datos en Microsoft Excel.

Una vez unificadas las bases, se tomó la decisión de renombrar las variables en Excel, con nombres sencillos y reducidos (Ver ***Anexo 1***), con el propósito de facilitar el posterior manejo de las variables en RStudio. Para esta tarea en específico se decidió utilizar Excel ya que no todas las variables cuentan con los mismos requisitos para el cambio de nombre. Un ejemplo claro de esto son los nombres de las variables correspondientes a cada pregunta, en donde el nombre de cada variable será P\_N siendo N el número de la pregunta, y donde PI\_N correspondiente a aquellas preguntas que cuentan con una escala inversa – Si el candidato marcó 5 en la pregunta, se registra un 1 en la base de datos – siendo I el indicador de esto, y N el número de la pregunta.

## Conversión de las bases a variables *Dummy*:

Para la minería de las reglas de asociación, es sumamente importante que la información sea de carácter booleano, en donde cada variable debe registrar si la observación cumple con el criterio de la variable con un 1 = TRUE, y un 0 = FALSE en el caso contrario.

Dada la naturaleza de la base de datos estudiada, la versión original cuenta con variables del tipo *Character* por ejemplo la variable ‘CARGO’, y otras del tipo *Numeric*, cómo lo serían las variables correspondientes a los puntajes. Tomando esto en consideración, una vez unificadas las bases, el siguiente paso consistió en la conversión de las distintas variables a variables indicadoras (“*Dummy*”). A raíz de esta necesidad, en conjunto con el amplio espectro del tipo de variables, se tomó la decisión de convertir las variables por medio de Microsoft Excel.

A raíz de la extensión de las variables involucradas en el modelo, se explicarán solamente las “familias” de variables y el criterio ocupado para convertirlas en indicadoras, y no el proceso de cada una ya que la base de datos consolidada consta de 229 variables. Las distintas variables fueron agrupadas en las siguientes categorías para el posterior detalle del trabajo realizado sobre ellas:

1. **Cargos**: Para poder convertir en variables indicadoras la información relevante al tipo de cargo al que postula el candidato se ocupó el “Manual de familias de cargo” realizado por eb|Consulting, el cual sentó base para categorizar la alta variedad de cargos en la base de datos en 7 grandes variables indicadoras.
2. **Edad**: Con el propósito de convertir la variable ‘EDAD’ original en variables indicadoras, se segmentó la información bajo 9 categorías en donde se determina el rango etario del candidato.
3. **Nivel educativo**: Variables relacionadas al nivel educativo señalado por el candidato. Se categorizó el nivel educativo de los candidatos bajo variables indicadoras del nivel de estudios alcanzados. Dada la ausencia de respuestas en la variable original de nivel educativo, se tomó como premisa el siguiente supuesto: En aquellas observaciones con nivel educativo indeterminado, el mayor grado educativo alcanzado es educación media.
4. **Puntajes de corte**: Las variables relacionadas con el puntaje de corte de cada pregunta fueron descartadas de posterior análisis ya que la variabilidad de los puntajes es provocada a los distintos requisitos exigidos por los diferentes clientes de eb|Consulting.
5. **Promedios por facetas**: Variables en donde se calcula el promedio de las preguntas relacionadas a cada faceta de la personalidad medida.
6. **Preguntas**: Las variables asociadas a cada pregunta del cuestionario Personali-T, de eb|Consulting. Miden el puntaje asignado por cada candidato a la pregunta correspondiente.
7. **Resultados**: Estas variables (BAJO, MEDIO y ALTO) definen las categorías de la variable original *Resultados*, la cual clasifica, en categorías del mismo nombre de las variables anteriores, los resultados de los candidatos.

### Conversión a variables indicadoras:

A continuación, se especificará el trabajo realizado para la conversión de las distintas “familias” de variables, detallas anteriormente, a variables indicadoras con el propósito de facilitar la comprensión de las reglas de asociación minadas como consecuencia del algoritmo ocupado.

### Categorías de cargo:

La variable original detalla, en forma de *strings*, el cargo al que postuló el candidato a la hora de contestar el test Personali-T. Para poder incluir esta información en la minería de reglas de asociación y ver si existen coocurrencias entre las respuestas del candidato y las preguntas del cuestionario, se ocupó el manual de familias de cargo desarrollado por eb|Consulting, en el cual se detallan los requisitos generales que exige cada familia de cargo. Un ejemplo de esto serían aquellos cargos catalogados bajo la variable DGSG (Directores, Gerentes y Subgerentes), ya que, según el manual, la mayoría de las posiciones abiertas en estos cargos contemplan ciertos requisitos mínimos de capacidad cognitiva y personalidad similares. Bajo este criterio, se definieron variables para los cargos de jefatura, ventas, técnicos y demás; siendo cada una de estas una nueva variable indicadora, en dónde 1 = TRUE indica que el candidato postulo a un cargo dentro de esa familia de cargos, y 0 = FALSE el caso contrario.

Para cerciorar que no hubo pérdida de información en el proceso de convertir la información a variables indicadoras, se creó una variable llamada CHECK, en la cual se suman los valores de cada fila para las variables asociadas a cargos. De esta forma, al correr el comando *summary(CHECK)* podemos notar que tanto el máximo como el mínimo para cada observación es de 1, lo que implica que cada observación está categorizada dentro de una de las familias de cargo.

## Edad:

Para poder traducir la edad de los candidatos en variables indicadoras, se tomó la decisión de segmentar a los candidatos por grupo etario. A diferencia de un método en donde se ocupe la media o la moda para generar la variable indicadora, con este método, si bien tenemos más variables, podemos buscar asociaciones entre las respuestas y el grupo etario de los candidatos, y potencialmente conocer más en detalle el comportamiento de los distintos grupos.

Dada la reducida cantidad de observaciones menores a 18 años y mayores a 100, se eliminaron todas aquellas observaciones fuera de dicho rango. Con esto, las categorías definitivas contemplan menores a 24 y mayores a 60 años, mientras que los rangos intermedios se separan cada 5 años, causando un resultado de 9 variables indicadoras respecto a la edad.

Para evitar la pérdida de información relevante, se definió una variable – ocupada temporalmente – llamada CHECK2, la cual es la suma de las variables indicadoras relacionadas con el tiempo. De esta forma, al ejecutar el código *summary(CHECK2)* podemos observar que tanto el mínimo como el máximo valor es de 1, demostrando que no hay observaciones que se encuentren en más de una categoría, ni observaciones que no se encuentren en ninguna.

## Nivel educativo:

Para poder convertir en variables indicadoras la información correspondiente al nivel educativo alcanzado por cada candidato, se ocuparon tablas dinámicas de Excel para determinar los niveles especificados por estos. Una vez rescatado el rango educativo alcanzado (entre educación básica y doctorado), se crearon variables indicadoras en donde 1 = TRUE indica si el candidato alcanzó el nivel señalado por la variable, y 0 = FALSE en el caso contrario.

En el caso de aquellas observaciones en las que se registró un nivel educativo incompleto estas fueron categorizadas con un 1 = TRUE bajo las variables indicadoras del último nivel completado. Por ejemplo, si un candidato marcó maestría incompleta como último nivel de estudios, la observación fue categorizada bajo la variable indicativa correspondiente a educación universitaria completa.

Para corroborar que no hubo pérdida de información, se creó temporalmente la variable CHECK3, la cual consiste en la suma de todas las variables indicadoras asociadas al nivel educativo. Una vez creada la variable, se ejecutó el comando *summary(CHECK3)* con el cual se observó que el mínimo y el máximo es de 1, indicando que no hubo observaciones categorizadas en más de una variable, ni observaciones sin categoría.

## Resultados:

La última variable temporal definida fue CHECK4, la cual, al igual que las anteriores, está conformada por la suma de las variables BAJO, MEDIO y ALTO – Variables indicadoras de la categoría del puntaje obtenido por el candidato – por lo que, al sumar 1, implica que la observación se encuentra solo en una de las categorías (clasificación correcta), de ser 0, la observación no se encuentra en ninguna de las variables indicadoras (clasificación errónea). Finalmente, el ultimo escenario posible es que la variable CHECK4 tome un valor mayor a 1, significando que la observación ha sido categorizada en múltiples variables indicadoras del tipo **Resultados** y, por lo tanto, ha sido erróneamente categorizada.

### Puntajes de corte:

Las 36 variables asociadas a los puntajes de corte requeridos para la faceta evaluada fueron descartadas del estudio por los siguientes motivos:

1. No tienen relación con las respuestas de los candidatos.
2. Los candidatos no tienen incidencia en estas.
3. La única variabilidad presente en dichas variables corresponde a las exigencias definidas por los clientes de eb|Consulting para cada faceta. Es decir, el puntaje de corte tradicional se define en 3 en una escala del 1 al 5, pero en aquellos procesos de selección donde una empresa considera más importante una faceta, el puntaje de corte se aumentó a 4.

## Promedios por faceta:

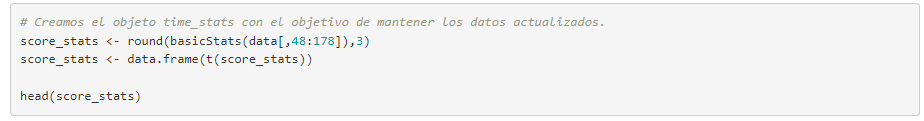
Dado a que estas variables están construidas con el promedio de cada candidato por faceta, y tienen como objetivo determinar si la faceta está por sobre o por debajo del puntaje de corte definido, estas fueron descartadas del análisis. Al igual que las variables relacionadas con el puntaje de corte, los criterios para excluirlas del estudio fueron:

1. No tienen relación con las respuestas de los candidatos.
2. Los candidatos no tienen incidencia en estas.

### Preguntas:

Para las variables relacionadas a las preguntas del cuestionario Personali-T de eb|Consulting, se determinó que las variables indicadoras serían creadas ocupando el promedio de la pregunta más/menos una desviación estándar de la media. Se definió el prefijo “BIN\_” como indicador para todas aquellas variables indicativas relacionadas con las preguntas.

Para crear las variables indicadoras mencionadas, se ocuparon los paquetes de R fBasics y dplyr. Con el primer paquete se ejecutó el código que se puede observar en el ***Código 1***. En donde se almacena, en el objeto score\_stats, la información estadística de las variables asociadas a las preguntas. Esta información es rescatada por medio de la función de fBasics basicStats. Posteriormente, se redondean los valores, se traspone la matriz almacenada en score\_stats y se renombran las variables con el propósito de facilitar la lectura y la comprensión de los resultados.



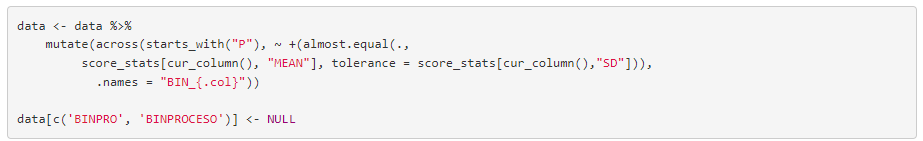
Código 1: Generación de las estadísticas básicas para la posterior creación de las variables indicadoras.

Se ocupó el comando *head()* para ver el comportamiento estadístico de las primeras 6 variables relacionadas con las preguntas del cuestionario. Inicialmente, se tomó en consideración la mediana para generar las variables, sin embargo, como se puede observar en la ***tabla 3***, el comportamiento de la mediana es repetitivo a lo largo de las variables, y dado a que el rango de valores se encuentra entre 1 y 5, se tomó la decisión de ocupar la media y la desviación estándar como el factor determinante para convertir el puntaje de las preguntas en variables indicativas.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | *P\_1* | *P\_2* | *P\_3* | *P­\_4* | *P\_5* | *P\_6* |
| *N.OBS* | 21,393 | 21,393 | 21,393 | 21,393 | 21,393 | 21,393 |
| *NAS* | 11 | 11 | 11 | 11 | 11 | 11 |
| *MIN* | 1 | 1 | 1 | 1 | 1 | 1 |
| *MAX* | 5 | 5 | 5 | 5 | 5 | 5 |
| *X1.QUARTILE* | 4 | 4 | 2 | 4 | 4 | 3 |
| *X3.QUARTILE* | 5 | 4 | 4 | 5 | 5 | 4 |
| *MEAN* | 4.160 | 3.940 | 2.927 | 4.512 | 4.614 | 3.450 |
| *MEDIAN* | 4 | 4 | 3 | 5 | 5 | 4 |
| *SUM* | 88948 | 84244 | 62578 | 96467 | 98658 | 73763 |
| *SE.MEAN* | 0.005 | 0.005 | 0.007 | 0.004 | 0.004 | 0.006 |
| *LCL.MEAN* | 4.150 | 3.930 | 2.912 | 4.504 | 4.606 | 3.437 |
| *UCL.MEAN* | 4.170 | 3.950 | 2.941 | 4.519 | 4.622 | 3.462 |
| *VAR* | 0.554 | 0.588 | 1.118 | 0.332 | 0.343 | 0.881 |
| *SD* | 0.744 | 0.767 | 1.057 | 0.576 | 0.585 | 0.938 |
| *SKEW* | -0.835 | -0.575 | 0.075 | -1.095 | -1.721 | -0.439 |
| *KURT* | 1.304 | 0.532 | -0.503 | 2.955 | 4.826 | -0.048 |

Tabla 3: Primeras 5 observaciones de score\_stats.

Una vez definido el criterio para la creación de las nuevas variables, se ocupó la librería dplyr y la función *mutate()* para generar un comando que automáticamente identificará la media y la desviación estándar de las variables y categorizará las observaciones, con un 1 = TRUE si es que la observación entra en el rango y 0 = FALSE en el caso contrario, en variables nuevas con el mismo nombre de la original, pero con el prefijo ‘BIN\_’. Como se puede observar en el ***Código*** 2, primero se definen las nuevas variables bajo la lógica de que aplique el criterio a aquellas variables que empiecen con la letra **P**, y posteriormente se eliminan las variables BIN\_PRO y BIN\_PROCESO ya que no son de interés con el estudio.



Código 2: Creación de las variables indicativas usando los valores de score\_stats y eliminación de variables erróneas.

Luego de corroborar que todas las variables han sido creadas correctamente, se procedió a eliminar aquellas variables irrelevantes para el caso de estudio. Con el propósito de mantener la base de datos consolidada y convertida a variables indicativas – A partir de este punto se referirá a esta base de datos como BDCP (Base de Datos Consolidada y Preparada) – sin mayores alteraciones, se crea una copia de la base llamada personalidad\_bin, de la cual se eliminan aquellas variables no utilizadas posteriormente con el comando ‘*NULL*’ (Véase ***Código 3***).



Código 3: Eliminación de variables irrelevantes para el estudio y consolidación en personalidad\_bin.

## Minería de reglas de asociación:

A raíz de tener la base consolidada, limpia y representada la información en variables indicadoras (“*Dummy*”), tenemos la base para poder buscar las coocurrencias entre las distintas variables por medio de la minería de reglas de asociación.

Para poder minar las reglas de asociación relevantes se utilizó el algoritmo Apriori, el cual fue planteado inicialmente en 1994 por los *computer scientist* Rakesh Agrawal y Ramakrishnan Srikant en una publicación denominada “*Fast Algorithms for Mining Association Rules*”. El desarrollo e implementación del algoritmo Apriori en el lenguaje R fue inspiración de Michael Hahnsler, Christian Buchta, Bettina Gruen, Kurt Hornik, Ian Johnson y Makhlouf Lembdi, siendo la patente original del código fuente a nombre de Christian Borgelt.

El algoritmo tiene como objetivo inicial encontrar *itemset*s lo suficientemente recurrentes dentro de la base de datos para poder seleccionar aquellos que cumplen con la frecuencia determinada (*Support* o soporte). Para esto, el proceso lógico que lleva el algoritmo consta, a grandes rasgos, de los siguientes pasos observables en la ***Ilustración 1*** [[7]](#footnote-7), en donde C\_k es el ítem siendo evaluado, y L\_K el *itemset* frecuente de orden K en el que se está evaluando.

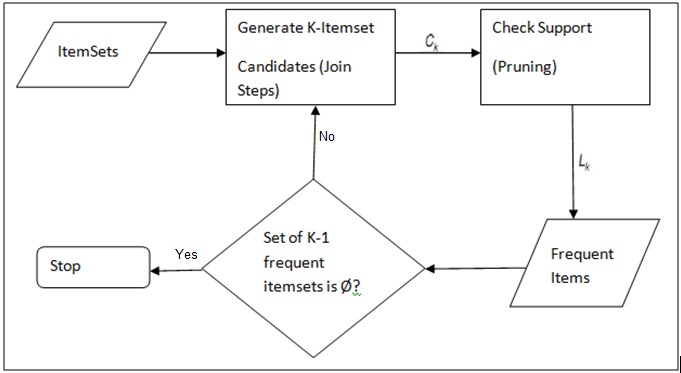


Ilustración 1: Proceso lógico del algoritmo Apriori.

El algoritmo se acerca a la base de datos en búsqueda de los *itemset* frecuentes desde la base más simple, empezando desde *itemset*s de una sola variable, y luego procede a ir agregando variable a los distintos *itemset*s generados. Sin embargo, minar absolutamente todos los *itemset*s frecuentes es computacionalmente muy exigente, y, además, inútil ya que habría que iterar por cada posible *itemset*. Apriori soluciona el problema de los recursos computacionales de una forma notablemente eficiente, en donde el algoritmo dejará de buscar *itemset* frecuentes con todos aquellos *itemset*s que no cumplieron con la frecuencia (*Support* o soporte), y en este caso la confianza, determinadas al comienzo. De esta forma, se optimiza la minería de *itemset*s frecuentes y reglas de asociación por medio del descarte inicial de todos aquellos *itemset* que no cumplen con los criterios definidos. Se puede visualizar el funcionamiento del algoritmo en la ***Ilustración 2*** [[8]](#footnote-8).

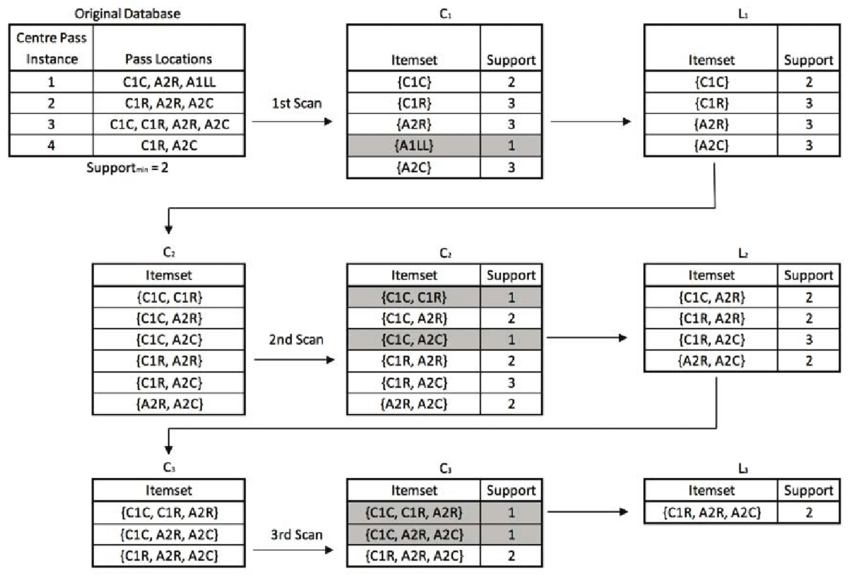


Ilustración 2: Demostración de la selección de itemsets frecuentes.

En el análisis se consideraron tres de los parámetros mencionados anteriormente, Soporte, Confianza y *Lift*. Para la generación inicial de las reglas se definió que el algoritmo Apriori descartara aquellas reglas con un soporte menor al 85% (frecuencia en *itemset*) y una confianza (Probabilidad de la coocurrencia) mayor o igual al 95%. A pesar de la exigencia de los parámetros iniciales, el algoritmo entregó 34,255 reglas de asociación, aunque con mayor capacidad computacional se podrían haber extraído más. Es por esto que, para efectos de reducir la extensión del estudio, y dada la imposibilidad de abordarlas todas, el informe se centrará en mostrar y estudiar únicamente 20 de las reglas minadas, seleccionadas bajo el criterio de tener aquellas con mayores puntajes en soporte, confianza y *Lift*, y relacionadas a la variable BIN\_P\_106 (pregunta relacionada con la faceta de entusiasmo).

Dados los requerimientos del algoritmo Apriori, en donde los valores ingresados deben ser booleanos, el primer acercamiento que se tomó para minar las reglas de asociación fue convertir los valores de la BDCP de valores numéricos (0 y 1) a valores booleanos TRUE y FALSE. Para esto, se eliminó la columna *X* puesto que contiene el indicador de fila, información irrelevante para el proyecto; y, como se puede observar en el ***Código 4***, por medio de la librería dplyr se ocupó la función *mutate\_if*, la cual cumplió la función de mutar condicionalmente todos aquellos valores del tipo *Numeric*, identificados por medio del comando *is.numeric*, al tipo *Logical* por medio del comando *as.logical*.



Código 4: Transformación de los datos del tipo numérico al tipo booleano.

Una vez convertidos los valores al tipo lógico, se procedió a implementar el algoritmo Apriori desde la librería Arules. Para esto, se definió un objeto, llamado *‘rules’*, cuyo objetivo es contener todas las reglas de asociación minadas. La función *apriori()* cuenta con diversos parámetros modificables por el usuario. Dada la flexibilidad de la función, se tomó la decisión de incluir como parámetros de selección un *Support = 0.85* y un *Confidence = 0.95*. Sin embargo, dado el extenso número de variables, el poder computacional requerido seguía siendo mayor al disponible, causando problemas en la minería, e impidiendo la ejecución rápida del código. Para poder controlar y optimizar el funcionamiento del algoritmo Apriori se incluyeron dos parámetros adicionales de control:

1. ***Memopt***: El propósito de este comando lógico de control es determinar si se prefiere minimizar el uso de memoria computacional (*memopt = TRUE*), o alternativamente, maximizar la velocidad del algoritmo (*memopt = FALSE, default*).
2. ***Load***: el propósito de este comando lógico de control es determinar si se deben almacenar (*load = TRUE, default*) o no (*load = FALSE*) las transacciones en la memoria computacional.



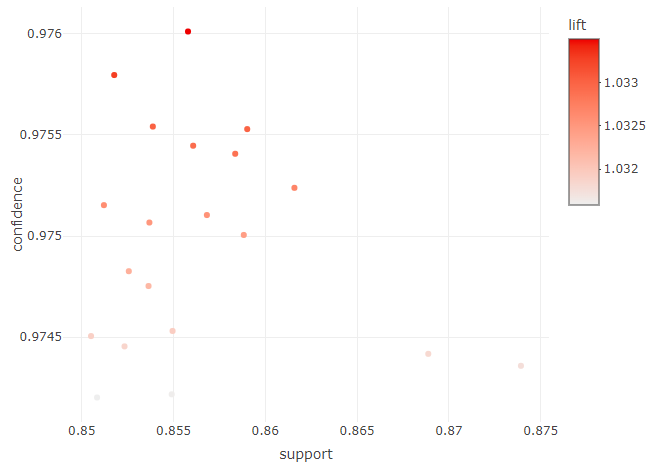
Código 5: Minería de las reglas de asociación.

Como se puede observar en el ***Código 5***, para el proyecto se definieron los parámetros mencionados como *memopt = TRUE* y *load = FALSE*, de este modo se priorizó la protección de la memoria RAM disponible y la minimización del uso de esta. Si bien es contraintuitivo ocupar esta configuración para acelerar el proceso de minado, en este caso particular, la lentitud inicial del algoritmo se debía a los limitados recursos computacionales disponibles, particularmente relacionado a la memoria RAM. Es por esto que, al reconfigurar dichos parámetros de control, se optimizó el uso de los recursos disponibles, logrando que el algoritmo Apriori minara correctamente las reglas de asociación buscadas, pero limitando la cantidad de reglas posibles a minar.

Una vez ejecutado exitosamente el código con el algoritmo Apriori, se procedió a ocupar la biblioteca arulesViz, librería de R dedicada a la visualización de reglas de asociación del paquete arules. Por medio de esta librería se generaron tres visualizaciones útiles, las cuales se centraron en dos objetivos principales:

1. Visualizar el comportamiento de las 20 reglas de asociación con mejores puntajes en las variables *Support*, *Confidence* y *Lift*, y relacionadas con la variable BIN\_P\_106.
2. Comprender la relación entre las distintas variables involucradas en dichas reglas de asociación rescatadas de forma global.

Para comprender mejor el comportamiento de los parámetros se generó un gráfico de dispersión agregando un factor de tonalidad para poder contrastar el *Lift* entre las distintas reglas. Para efectos del informe se puede observar en la ***Ilustración 3*** la dispersión y gradiente de los distintos puntos, no obstante, se generó un gráfico interactivo que detalla el valor de los parámetros y el contenido de cada regla de asociación en el ***Anexo 2***. Cabe destacar, que el gráfico de dispersión de la figura mencionada sólo se muestran reglas las mejores 20 reglas de asociación relacionadas a la variable BIN\_P\_106, sin embargo, más adelante se abordan nuevas asociaciones entre las variables.



2

1

3

Ilustración 3: Visualización estilo Jitter Plot sobre las 20 principales reglas de asociación relacionadas a la variable BIN\_P\_106. Explicación de los recuadros más adelante.

Como se puede observar en la ***Ilustración 3***, las 20 principales reglas de asociación cuentan con una confianza mayor al 97%, un soporte mayor al 85% y un *Lift* mínimo de 1.03. Los recuadros marcados en la figura mencionada muestran las siguientes reglas de asociación:

1. Recuadro 1 ***Ilustración 3***:{BIN\_P\_11, BIN\_P\_16, BIN\_P\_36, BIN\_P\_43} => {BIN\_P\_106}
   1. Con una confianza del 0.976, es la que cuenta con mayor probabilidad de coocurrencia dentro de las 20 graficadas, mientras que la regla con menor, cuenta con un valor de 0.974.
2. Recuadro 2 ***Ilustración 3***:{BIN\_P\_11, BIN\_P\_16, BIN\_P\_43} => {BIN\_P\_106}
   1. Con un soporte del 0.874, es la que mayor frecuencia presenta dentro de las reglas de asociación mostradas en la ***Ilustración 3***, siendo la menor frecuencia alcanzada de 81,5%.
3. Recuadro 3 ***Ilustración 3***:

{BIN\_P\_16, BIN\_P\_18, BIN\_P\_43, BIN\_P\_103} => {BIN\_P\_106}

* 1. Finalmente, este recuadro muestra la regla de asociación con menores valores en todos los criterios presentes, alcanzando un soporte del 81,5%, una confianza del 97,4% y un *Lift* menor a 1,032.

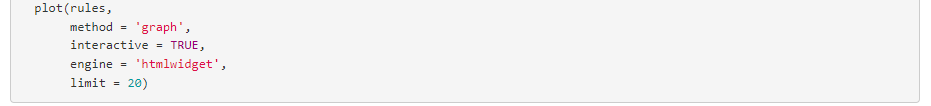
Como se puede observar en el ***Anexo 2***, todas las reglas presentadas en la ***Ilustración 3*** tienen como consecuencia – A partir de este momento RHS, por sus siglas en inglés *Rigth Hand Side* – la variable BIN\_P\_106. El temor inicial del proyecto consistía en que la minería de reglas de asociación encontrara únicamente asociaciones entre preguntas que miden las mismas facetas de la personalidad, sin embargo, podemos notar en el anexo mencionado, y en la ***Tabla 4*** que no todas las variables antecedentes – A partir de ahora LHS, por sus siglas en inglés *Left Hand Side* – tienen relación con la faceta de entusiasmo como RHS.

|  |  |  |  |
| --- | --- | --- | --- |
| Variables involucradas LHS | Faceta que mide | Variable RHS | Faceta que mide |
| BIN\_P\_4 | Curiosidad | BIN\_P\_106 | Entusiasmo |
| BIN\_P\_11 | Bienestar | BIN\_P\_106 | Entusiasmo |
| BIN\_P\_16 | Adaptabilidad | BIN\_P\_106 | Entusiasmo |
| BIN\_P\_17 | Deliberación | BIN\_P\_106 | Entusiasmo |
| BIN\_P\_18 | Orientación al trabajo | BIN\_P\_106 | Entusiasmo |
| BIN\_P\_29 | Orientación al trabajo | BIN\_P\_106 | Entusiasmo |
| BIN\_P\_36 | Obediencia | BIN\_P\_106 | Entusiasmo |
| BIN\_P\_43 | Emociones positivas | BIN\_P\_106 | Entusiasmo |
| BIN\_P\_46 | Orden | BIN\_P\_106 | Entusiasmo |
| BIN\_P\_54 | Tendencia a ser pacífico | BIN\_P\_106 | Entusiasmo |
| BIN\_P\_67 | Obediencia | BIN\_P\_106 | Entusiasmo |
| BIN\_P\_103 | Cooperación | BIN\_P\_106 | Entusiasmo |
| BIN\_P\_115 | Curiosidad | BIN\_P\_106 | Entusiasmo |
| BIN\_PI\_124 | Baja ansiedad | BIN\_P\_106 | Entusiasmo |

Tabla 4: Variables LHS involucradas en las 20 reglas de asociación estudiadas, y sus respectivas facetas que miden, en contraposición con la variable RHS.

Como se puede observar en la ***Tabla 4***, distintas variables involucradas en LHS miden distintas facetas, demostrando que el foco del proyecto es beneficioso para eb|consulting y para la academia puesto que se han encontrado relaciones de coocurrencia estadística entre distintas facetas de la personalidad, facilitando el estudio en profundidad de la personalidad.

Para determinar el comportamiento de las relaciones entre las variables indicativas descritas en la tabla anterior, se realizó un gráfico interactivo por medio del motor *htmlwidget* en RStudio (Véase ***Código 6***). Los resultados generales se pueden observar en la ***Ilustración 4***, sin embargo, se recomienda ver el detalle del mapa interactivo en el ***Anexo 2***.



Código 6: Generación del mapa interactivo sobre la interacción entre variables.

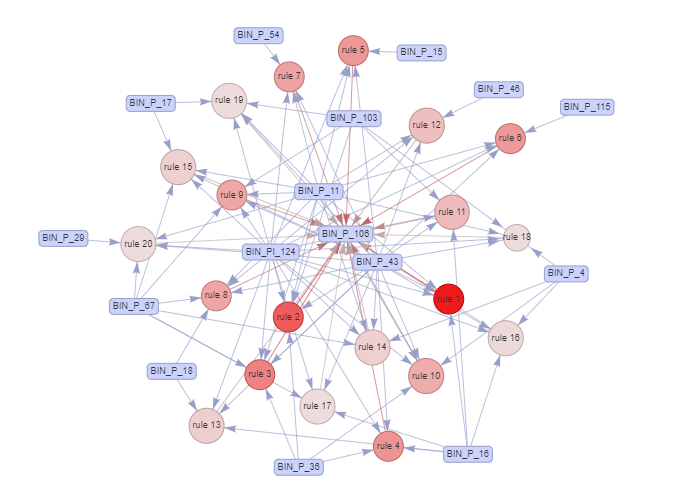


Ilustración 4: Visualización sobre las relaciones entre las variables LHS y RHS estudiadas.

La figura anterior detalla las relaciones entre las variables y la composición de las 20 reglas de asociación más importantes relacionadas con la variable BIN\_P\_106. Dada la complejidad de la visualización, el detalle se presenta en la ***Tabla 5***.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Regla | LHS | RHS | *Support* | *Confidence* | Coverage | *Lift* | Count | Order |
| 1 | {BIN\_P\_11, BIN\_P\_16, BIN\_P\_43, BIN\_PI\_124} | {BIN\_P\_106} | 0,811 | 0,982 | 0,826 | 1 | 17300 | 5 |
| 2 | {BIN\_P\_11, BIN\_P\_36, BIN\_P\_43, BIN\_P\_103, BIN\_PI\_124} | {BIN\_P\_106} | 0.801 | 0.981 | 0.816 | 1.04 | 17100 | 6 |
| 3 | {BIN\_P\_11, BIN\_P\_36, BIN\_P\_43, BIN\_P\_67, BIN\_PI\_124} | {BIN\_P\_106} | 0.803 | 0.981 | 0.819 | 1.04 | 17200 | 6 |
| 4 | {BIN\_P\_16, BIN\_P\_36, BIN\_P\_43, BIN\_PI\_124} | {BIN\_P\_106} | 0.806 | 0.981 | 0.822 | 1.04 | 17200 | 5 |
| 5 | {BIN\_P\_11, BIN\_P\_15, BIN\_P\_43, BIN\_PI\_124} | {BIN\_P\_106} | 0.803 | 0.981 | 0.818 | 1.04 | 17200 | 5 |
| 6 | {BIN\_P\_11, BIN\_P\_43, BIN\_P\_115, BIN\_PI\_124} | {BIN\_P\_106} | 0.807 | 0.981 | 0.823 | 1.04 | 17300 | 5 |
| 7 | {BIN\_P\_11, BIN\_P\_43, BIN\_P\_54, BIN\_PI\_124} | {BIN\_P\_106} | 0.809 | 0.981 | 0.824 | 1.04 | 17300 | 5 |
| 8 | {BIN\_P\_11, BIN\_P\_18, BIN\_P\_43, BIN\_P\_67, BIN\_PI\_124} | {BIN\_P\_106} | 0.8 | 0.981 | 0.816 | 1.04 | 17100 | 6 |
| 9 | {BIN\_P\_11, BIN\_P\_43, BIN\_P\_67, BIN\_P\_103, BIN\_PI\_124} | {BIN\_P\_106} | 0.807 | 0.981 | 0.822 | 1.04 | 17300 | 6 |
| 10 | {BIN\_P\_4, BIN\_P\_11, BIN\_P\_36, BIN\_P\_43, BIN\_PI\_124} | {BIN\_P\_106} | 0.802 | 0.981 | 0.818 | 1.04 | 17200 | 6 |
| 11 | {BIN\_P\_16, BIN\_P\_43, BIN\_P\_103, BIN\_PI\_124} | {BIN\_P\_106} | 0.809 | 0.981 | 0.825 | 1.04 | 17300 | 5 |
| 12 | {BIN\_P\_11, BIN\_P\_43, BIN\_P\_46, BIN\_PI\_124} | {BIN\_P\_106} | 0.81 | 0.981 | 0.825 | 1.04 | 17300 | 5 |
| 13 | {BIN\_P\_16, BIN\_P\_18, BIN\_P\_43, BIN\_PI\_124} | {BIN\_P\_106} | 0.803 | 0.981 | 0.819 | 1.04 | 17200 | 5 |
| 14 | {BIN\_P\_4, BIN\_P\_11, BIN\_P\_43, BIN\_P\_67, BIN\_PI\_124} | {BIN\_P\_106} | 0.807 | 0.981 | 0.823 | 1.04 | 17300 | 6 |
| 15 | {BIN\_P\_11, BIN\_P\_17, BIN\_P\_43, BIN\_P\_67, BIN\_PI\_124} | {BIN\_P\_106} | 0.805 | 0.981 | 0.821 | 1.04 | 17200 | 6 |
| 16 | {BIN\_P\_4, BIN\_P\_16, BIN\_P\_43, BIN\_PI\_124} | {BIN\_P\_106} | 0.809 | 0.981 | 0.825 | 1.04 | 17300 | 5 |
| 17 | {BIN\_P\_16, BIN\_P\_43, BIN\_P\_67, BIN\_PI\_124} | {BIN\_P\_106} | 0.81 | 0.98 | 0.827 | 1.04 | 17300 | 5 |
| 18 | {BIN\_P\_4, BIN\_P\_11, BIN\_P\_43, BIN\_P\_103, BIN\_PI\_124} | {BIN\_P\_106} | 0.805 | 0.98 | 0.822 | 1.04 | 17200 | 6 |
| 19 | {BIN\_P\_11, BIN\_P\_17, BIN\_P\_43, BIN\_P\_103, BIN\_PI\_124} | {BIN\_P\_106} | 0.803 | 0.98 | 0.819 | 1.04 | 17200 | 6 |
| 20 | {BIN\_P\_11, BIN\_P\_29, BIN\_P\_43, BIN\_P\_67, BIN\_PI\_124} | {BIN\_P\_106} | 0.801 | 0.98 | 0.816 | 1.04 | 17100 | 6 |

Tabla 5: 20 principales reglas de asociación estudiadas y visualizadas en la Ilustración 4.

Podemos notar que las reglas de asociación minadas de la BDCP de la ***Tabla 5***, presentan un comportamiento similar, en donde los valores del *Lift* se redondean todos a 1.04, hay 15 variables diferentes LHS involucradas en la coocurrencia de la variable BIN\_P\_106, de las cuales ninguna tiene una relación inicial con la faceta de entusiasmo – faceta medida parcialmente por la variable BIN\_P\_106 – mostrando resultados atractivos e importantes para el proyecto.

Finalmente, se tomó la decisión de generar un gráfico mediante el método *grouped matrix*, con el propósito de determinar que otras reglas de asociación más complejas se obtuvieron del modelo. Para esto, se ocupó la librería arulesViz por medio del código observable en el ***Código 7***. Como se puede observar en la figura, se muestran 20 grupos LHS conformados por diversas variables, y 10 RHS diferentes, indicando el éxito del modelo para determinar coocurrencias estadísticas dentro de los datos de la BDCP.





Código 7: Generación de visualización estilo Grouped Matrix para las principales asociaciones minadas en general.

En la ***Ilustración 5***, podemos observar en detalle que el algoritmo logró rescatar reglas de asociación y coocurrencias entre distintos grupos de variables involucradas en el *dataset*. Entregando a eb|Consulting sustento suficiente para que se inicien investigaciones inter-faceta entre las distintas variables involucradas.

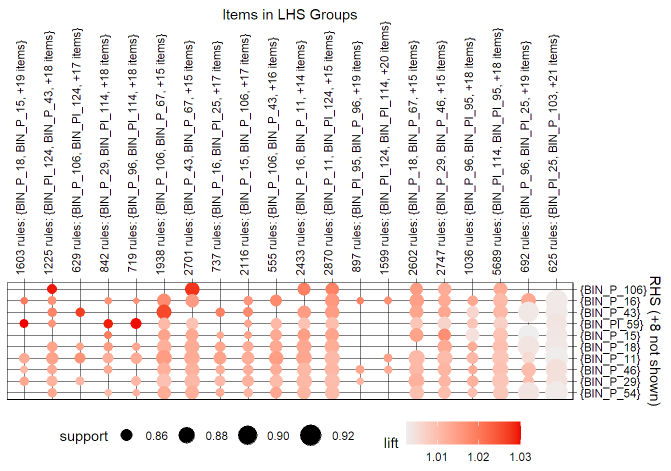


Ilustración 5: Gropued matrix, visualización de las principales asociaciones LHS-RHS producto del algoritmo Apriori.

# Conclusiones

Con el objetivo de determinar la posible existencia de reglas de asociación relevantes entre las distintas preguntas psicométricas, y los distintos rasgos de la personalidad, se ocuparon las respuestas al test Personali-T, información demográfica y datos sobre las familias de cargo de 21,393 postulantes a distintos procesos de selección.

El estudio consistió en tres pasos fundamentales, siendo el primero la preparación y selección de la data correspondiente, en donde se empezó por la consolidación de las bases de cada proceso de selección realizado con el test psicométrico de eb|Consulting. Una vez estandarizadas las bases y adecuadamente unificadas se procedió a la depuración de observaciones inconsistentes (i.e.: Mayores a 100 años), proceso con el que se concluyó con la presencia de 21,393 observaciones útiles. En segundo lugar, y posiblemente el paso fundamental para el estudio, se procedió a convertir las variables originales de la base de datos consolidada en variables indicadoras – De esta forma, la información puede ser ocupada para la minería de reglas de asociación – ocupando diversos criterios estadísticos y lógicos para la correcta categorización y redefinición de la información en variables indicadoras (“*Dummy*”).

Una vez concluida la fase de modificaciones, se procede a aplicar el algoritmo Apriori – Algoritmo desarrollado para la minería de reglas de asociación y el hallazgo de *itemset*s frecuentes – para minar las reglas de asociación que cumplan con un soporte mayor o igual al 85%, y cuenten con una confianza mayor o igual al 95%. A pesar de la exigencia de los parámetros, y de la reducida capacidad computacional, el modelo entrego 34,255 reglas de asociación, en la cuales el estudio se enfocó en las 20 principales relacionadas a la variable BIN\_P\_106, la cual mide parcialmente el puntaje de los candidatos en la faceta de Entusiasmo, logrando determinar la presencia de 15 variables LHS coocurrentes con la variable de estudio.

El éxito del proyecto radica en que la minería de reglas de asociación para evaluaciones psicométricas presentó resultados en donde las principales coocurrencias no presentan variables LHS que midan la misma faceta que las variables RHS, lo cual implica la existencia, por lo menos estadística, de relaciones entre las variables que miden diferentes rasgos de la personalidad, abriendo paso al futuro estudio de conexiones inter-faceta y facilitando la investigación de estas por medio de la determinación de las asociaciones principales presentes en el *dataset*.

Eb|Consulting puede encontrar valor en el proyecto de investigación expuesto ya que, por medio de la minería de reglas de asociación, se han encontrado diversas relaciones de coocurrencia estadística entre las respuestas a las distintas preguntas del cuestionario Personali-T, las cuales sientan una base nueva para validar el instrumento psicométrico de nuevas maneras – Por medio del estudio de dichas relaciones – y facilitan el estudio de la medición de las distintas facetas de la personalidad como predictor del desempeño. Esto es sumamente valioso ya que, recolectando data suficiente sobre el desempeño de los candidatos seleccionados, las reglas de asociación minadas abren camino al estudio en profundidad de las relaciones inter-faceta como predictores del desempeño, permitiéndole a la compañía profundizar en la validación de sus instrumentos y en la ampliación de sus focos de investigación para aumentar el valor agregado hacia los clientes.

Adicionalmente, cabe destacar que, si bien han sido pocos, ya se han hecho estudios sobre las relaciones inter-faceta por medio de la minería de reglas de asociación, encontrando conexiones relevantes entre la irritabilidad y la apertura a nuevas ideas (89.95%) – Extraído y traducido de: Khalid Perwez, Syed et al. "*Association Rule Mining Technique for Psychometric Personality Testing and Behaviour Prediction*". *International Journal of Engineering and Technology*, Vol 5, No 5, 2013. – comprobando la utilidad del método para encontrar relaciones no evidentes entre las distintas facetas medidas.

Si bien los resultados del proyecto sientan precedente futuras investigaciones, es importante destacar que las reglas de asociación no describen el tipo de relación entre las distintas variables, solamente detallan la coocurrencia estadística de estas, por lo que es importante que eb|Consulting profundice en el estudio de la relevancia de dichas conexiones como predictores del desempeño con el propósito de posicionar el test Personali-T como un instrumento aún mejor a la hora de predecir el desempeño organizacional de los candidatos.

# Bibliografía y documentación relacionada

Albert Maydeu-Olivares, J. J. (2005). *Contemporary Psychometrics.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Barros, E. (s.f.). Problemas y desafíos de la selección de personas en Chile. *Administración y economía UC*, 16-21.

Burnett, R. P. (2003). A Personality Trait-Based Interactionist Model of Job Performance. *Journal of Applied Psychology*.

Deborah L. Whetzel, M. A. (2010). *Linearity of Personality–Performance Relationships: A Large Scale Examination.* Tampa: International Journal of Selection and Assessment.

Eduardo Barros, D. D. (s.f.). *La integración de la evidencia científica en los procesos de selección: una propuesta para prácticas en Chile.* Santiago de Chile: UAH/Ediciones.

Frank L. Schimdt, J. E. (1998). *The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings.* Iowa: Psychological Bulletin.

Hans-Georg Wolff, S. K. (2011). *The relationship between networking behaviors and the Big Five personality dimension.* Emeral group.

Jeffrery Pfeffer, R. I. (2006). Evidence-Based Management. *Harvard Business Review*.

Jennifer L. Jones, S. L. (2007). Foundations and Assummptions of the Scientist-Practitioner Model. *Sage Journals*.

Manuel Cárdenas, H. A. (2014). *Potencia Estadística y Cálculo del Tamaño del Efecto en G\*Power: Complementos a las Pruebas de Significación Estadística y su Applicación en Psicología.* Valparaíso, Chile: Universidad de Valparaíso.

Michael Hahsler, S. C. (s.f.). *Visualizing Association Rules: Introduction to the R-extension Package arulesViz.*

Mining Association Rules on Students’ Profiles and Personality Types. (2015). *Proceedings of the International MultiConference of Engineers and Computer Scientists.* Hong Kong.

Price, L. R. (2017). *Psychometric Methods: theory into Practice.* New York: The Guildford Press.

Rakesh Agrawal, R. S. (1994). Fast Algorithms for Mining Association Rules. *IBM Almaden Research Center*.

Syed Khalid Perwez, H. M. (2013). Association Rule Mining Technique for Psychometric Personality Testing and Behaviour Prediction. *International Journal of Engineering and Technology*.

Tan, J. (2012). Different Types of Association Rules Mining Review. *Applied Mechanics and Materials*.

# Anexos:

1. **Anexo 1: Diccionario de variables.xlsx**
   1. Supuestos (Hoja 1): Detalla los supuestos utilizados para la transformación de la data y las modificaciones realizadas a los datos para la conversión a variables indicativas.
   2. Categorización de cargos: Detalla los cargos de los distintos procesos de selección y las distintas familias de cargo utilizadas para la creación de sus respectivas variables indicadoras.
   3. Diccionario personalidad: Explicación en detalle de todas las variables involucradas en la base de datos consolidada y preparada. Cuenta con las variables que se ocuparon y las que fueron descartadas por ser irrelevantes para el estudio.
2. **Anexo 2: Association-Rule-Mining.html**
   1. Documento generado a partir del archivo RMD de RStudio con los códigos utilizados para la minería de reglas de asociación y su posterior visualización.
3. **Anexo 3: PERSONALI-T Data manipulation.html**
   1. Documento generado a partir del archivo RMD de RStudio con los códigos utilizados para la modificación y limpieza de variables de la base de datos consolidada.
4. **Anexo 4: PERSONALI-T\_SCORE-STATS.xlsx**
   1. Archivo Excel conteniendo las estadísticas fundamentales de las preguntas del cuestionario Personali-T. Archivo utilizado con el propósito de determinar el comportamiento de dichas variables y establecer criterios para la creación de las variables indicadoras.
5. **Anexo 5: CONSOLIDADO\_PERSONALI-T\_BINARY\_MODS.xlsx**
   1. Archivo Excel anonimizado con todas las observaciones y fórmulas ocupadas para la redefinición de las variables originales al tipo indicativa (“*Dummy*”).

1. Evidence-based Management: “*The conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients.*” (David Sackett, BHR.) [↑](#footnote-ref-1)
2. Scientist Practitioner Model: “*The scientist-practitioner model is founded on the ideology that trained professional psychologists should be knowledgeable in both research and clinical practice. Emphasis should be placed on the successful integration of science and practice, where the relationship between the two variables is carefully considered.”* (Jones; Mehr, 2007) [↑](#footnote-ref-2)
3. Algoritmo Apriori: Algoritmo desarrollado para el descubrimiento de *itemsets* frecuentes dentro de *datasets* y, posteriormente, para la mineríoa de reglas de asociación. [↑](#footnote-ref-3)
4. BIN\_P\_106: Variable indicadora de la respuesta de los candidatos en la pregunta 106, relacionada con la faceta de entusiasmo. [↑](#footnote-ref-4)
5. Tabla extraída de: Cárdenas Castro, J. M., y H. Arancibia Martini. «Potencia estadística Y cálculo Del tamaño Del Efecto En G\*Power: Complementos a Las Pruebas De significación estadística Y Su aplicación En psicología.». Salud & Sociedad, vol. 5, n.º 2, noviembre de 2016, pp. 210-44, doi:10.22199/S07187475.2014.0002.00006 [↑](#footnote-ref-5)
6. Personali-T: Test psicométrico estandarizado desarrollado por eb|Consulting para la predicción de desempeño de los candidatos a distintos procesos de selección. [↑](#footnote-ref-6)
7. Ilustración 1 extraída de: [https://dwgeek.com/mining-frequent-*itemset*s-apriori-algorithm.html/](https://dwgeek.com/mining-frequent-itemsets-apriori-algorithm.html/) [↑](#footnote-ref-7)
8. Ilustración 2 extraída de: <https://www.researchgate.net/figure/Tabular-example-of-the-Apriori-algorithm-process-Processing-example-of-the-Apriori_fig3_335323554> [↑](#footnote-ref-8)