TP AID 2021 cuat. 1
Análisis de asociación entre personalidad y consumo de drogas

Alumno: G. Sebastián Pedersen (sebasped (at) gmail (dot) com)
Docente: Débora Chan

Índice

1. Introducción

Los problemas de detectar o evaluar el riesgo de consumo y abuso de drogas, en especial ilegales, son de considerable importancia actual. En el presente trabajo se aborda el análisis de la asociación entre personalidad y consumo de drogas, haciendo énfasis en ilegales, y usando datos recolectados mediante encuestas anónimas, las cuales registraron datos de personas sobre la personalidad (cinco scores del NEO-FFI-R, una variante del conocido Five Factor Model; la impulsividad mediante el BIS-11 y la búsqueda de sensaciones mediante el ImpSS) y el consumo de diversas drogas legales e ilegales, además de rango etario, sexo, nivel de educación, nacionalidad y etnia.

2. Datos y prepocesado

2.1. Datos crudos

Para el análisis se utilizaron datos provenientes de encuestas anónimas. La base de datos está disponible online (ver [1]) y consta de 1885 observaciones. En cada encuesta, es decir cada persona u observación, se registraron las siguientes variables, divididas simplemente por mayor claridad en tres grupos:

I) Variables demográficas:

- Edad: 18-24; 25-34; 35-44; 45-54; 55-64 y 65 o más.
- Sexo: femenino; masculino.
- Educación:
  1) Dejó la escuela antes de los 16 años.
  2) Dejó la escuela a los 16 años.
  3) Dejó la escuela a los 17 años.
  4) Dejó la escuela a los 18 años.
  5) Terminó la escuela o cursó algo de terciario o universitario, pero no tiene algún título.
  6) Tiene título profesional.
  7) Tiene título universitario.
  8) Tiene título de maestría.
  9) Tiene título de doctorado.
- País de residencia.
- Etnia.

II) Variables sobre la personalidad (NEO-FFI-R ver [2], BIS-11 ver [3], ImpSS ver [4]):

- Nscore: score sobre neurotismo, medido por NEO-FFI-R.

- Escore: score sobre extraversión, medido por NEO-FFI-R.
- Oscore: score la apertura a experimentar, medido por NEO-FFI-R.
- Ascore: score sobre la complacencia, medido por NEO-FFI-R.
- Cscore: score sobre la conciencia, medido por NEO-FFI-R.
- Impulsive: score sobre la impulsividad, medido por BIS-11.
- SS: score sobre la búsqueda de sensaciones, medido por ImpSS.

III) Variables sobre el consumo de drogas, medidas con las siguientes categorías:

1) CL0: nunca consumió.
2) CL1: consumió hace más de una década.
3) CL2: consumió en la última década.
4) CL3: consumió en el último año.
5) CL4: consumió en el último mes.
6) CL5: consumió en la última semana.
7) CL6: consumió en el último día.

Las variables en sí mismas son las siguientes:

- Alcohol.
- Amphet: consumo de amfetaminas
- Amyl: consumo de amyl nitritos.
- Benzos: consumo de benzodiazepina.
- Caff: consumo de cafeína.
- Cannabis.
- Choc: consumo de chocolate.
- Coke: consumo de cocaína.
- Crack.
- Ecstasy: consumo de éxtasis.
- Heroin: consumo de heroina.
- Ketamine: consumo de ketamina.
- Legalh: consumo de *legal highs* (drogas psicoactivas contenidas en diversos químicos).
- LSD.
- Meth: consumo de metadona.
- Mushrooms: consumo de hongos mágicos.
- Nicotine.
- Semer: droga ficticia.
- VSA: consumo de sustancias volátiles.

2.2. Preprocesado

Se identificó que tanto la variable país de residencia como la variable etnia, estaban muy concentradas en pocas categorías. País de residencia, además de estar concentrada en unos pocos países (6 países y una categoría otros), el casi 85 % de los datos corresponden a UK o USA. Por su parte etnia está concentrada en algo más del 91 % en la raza blanca. Por lo tanto se decidió excluir estas variables del análisis, pues no hay una suficiente diversidad y adecuada distribución como para que tenga sentido considerar su influencia sobre la clasificación. También se decidió excluir del análisis la variable sexo por considerarse no adecuada su influencia en la clasificación.

Para decidir el criterio a tener en cuenta a la hora de considerar el consumo de drogas, tanto cuáles drogas como qué umbral de consumo, se realizó un análisis de correspondencia entre todas las drogas y tomando como variable *sí consume la droga* la correspondiente a los niveles CL3, 4, 5 y 6 (es decir consume significa que consumió en el último año o menos).

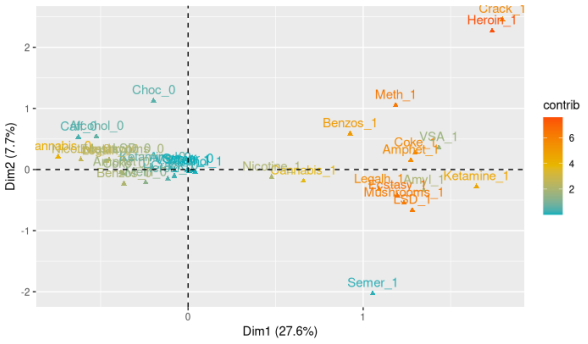

Figura 1. Análisis de correspondencia entre las drogas. La categoría 1 corresponde a haber consumido en el último año o menos.

En la figura 1 se observa que el consumo de Semer (*Semer_1*) está aislado, y como esta droga

es ficticia y fue introducida a modo de control, se decidió eliminar las observaciones que indicaron consumo de dicha droga. También se observa que los consumos de nicotina y cannabis están fuertemente asociados, y dado que en los últimos años algunos países han legalizado el consumo de cannabis, se decidió no tenerlas en cuenta (recordemos que el análisis hace énfasis en el consumo de drogas ilegales). Se observa otro grupo que incluye consumo de Meth, Benzos, VSA, Coke, Amphet, Ketamine, Legalh, Amyl, Ecstasy, Mushrooms y LSD. Tomando en cuenta la contribución de cada una de estas drogas y su ilegalidad, se decidió conservar Coke, Ecstasy, LSD y Meth. Se distingue otro grupo, que incluye Crack y Heroin, y se decidió conservar ambas tomando en cuenta ilegalidad y contribución. Finalmente se observa un último grupo, concentrado alrededor de lo más usual, que incluye a todos los no consumos y a los consumos de drogas no ilegales.

Cabe mencionar que se realizaron también análisis de correspondencia tanto exluyendo la droga ficticia Semer, como endureciendo el criterio de *sí consume* a si consumió en el último mes o menos, arrojando resultados similares a los de la figura 1, aunque endureciendo el criterio de consumo conllevaba a un mayor desbalanceo de clases. No se consideró como adecuado tanto endurecer aún más el criterio de consumo, como relajarlo más allá del último año.

Finalmente considerando que las variables edad y nivel educativo no iban a ser utilizadas de forma directa en el análisis, pues este se enfoca en la asociación de la personalidad con el consumo de drogas (con énfasis en ilegales), se decidió agrupar aún más estas variables para brindar por un lado mayor interpretabilidad y por el otro una mejor distribución entre sus respectivas categorías (las distribuciones originales se pueden ver de [1]). Las nuevas variables se detallan a continuación:

- Variable edad nueva: 18-24 (34.11 % de las obs.); 25-34 (25.52 %); 35 o más (40.37 %).
- Variable nivel educativo nueva:
  1) Dejó la escuela a los 18 años o antes (13.64 % de las obs.).
  2) Terminó la escuela o cursó algo de terciario o universitario, pero no tiene algún título (26.84 %).
  3) Es un profesional no universitario (14.32 %).
  4) Tiene título universitario o mayor (45.19 %).

Observemos que el nivel educativo está claramente sesgado hacia el nivel mayor. Deberemos tener en cuenta esto a la hora de realizar el análisis y derivar conclusiones.

Resumiendo, y tomando en cuenta los preprocesos antes mencionados, la cantidad de observaciones disminuyó a 1877, quedando balanceadas en 42 % y 58 % entre consume y no consume drogas respectivamente, y las variables finales utilizadas fueron las siguientes:

- Edad nueva: 18-24; 25-34 y 35 o más.
- Eduación nueva:
  1) Dejó la escuela a los 18 años o antes.
  2) Terminó la escuela o cursó algo de terciario o universitario, pero no tiene algún título.
  3) Es un profesional no universitario.
  4) Tiene título universitario o mayor.
- Nscore: score sobre neurotismo.
- Escore: score sobre extraversión.
- Oscore: score la apertura a experimentar.
- Ascore: score sobre la complacencia.
- Cscore: score sobre la conciencia.
- Impulsive: score sobre la impulsividad.
- SS: score sobre la búsqueda de sensaciones.
- Consume: variable que indica si consume o no drogas, siendo que sí consume si es que consumió alguna de las drogas cocaína, crack, éxtasis, heroína, LSD o metadona, en el último año o menos.

3. Análisis exploratorio

Como primera exploración de los datos, de los boxplots de cada score (N, E, O, A y C), de Impulsive y de SS, se observan distribuciones univariadas simétricas y con algunos outliers no severos, considerando tanto cada variable entera como su separación por la variable Consume. En un plano más formal y técnico se comprueban la normalidad univariada de las mencionadas variables tanto mirando los qqplots como realizando el test de Anderson-Darling (con p-valores menores al 2 %).

Ya que el análisis es sobre la asociación entre la personalidad y el consumo de drogas, una interesante primera medida es contrastar la media del grupo de observaciones que sí consumen contra la de los que no consumen, para analizar si existen diferencias e intentar discernir las variables que mejor pueden clasificar los mencionados grupos.

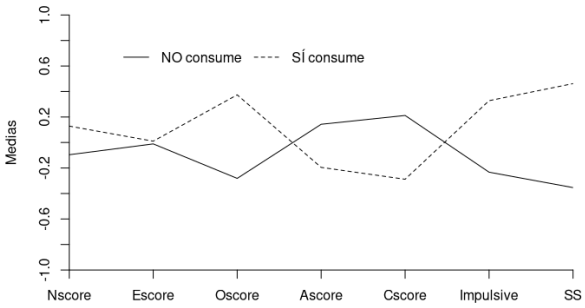

Figura 2. Comparación de los perfiles medios multivariados.

La figura 2 muestra los perfiles medios de ambos grupos (cabe aclarar que la estimación robusta no presentó diferencias significativas). A simple viste se observa que el grupo que consume tiene, en promedio, más alto el Oscore (apertura a experimentar), la impulsividad y el SS (búsqueda de sensaciones), y más bajo el Ascore (complacencia) y el Cscore (conciencia); observándose lo contrario en el grupo que no consume. Los Nscore (neurotismo) y Escore (extraversión) parecen, al menos en promedio, no influir.

Un resultado análogo se pudo apreciar de los boxplots de las mencionadas variables categorizadas por consume y no consume, arrojando más evidencia sobre la posibilidad de asociar la personalidad al consumo de drogas.

Lamentablemente el test de Shapiro arroja rechazo de la normalidad multivariada, y el M-test de Box rechaza la homocedasticidad entre los dos grupos. Más allá de esto, si de todos modos se aplica el test de Hotelling de comparación de medias multivariadas entre dos grupos (que tiene asidero en su validez asintótica), el resultado es de contundente rechazo de igualdad de medias entre ambos grupos, en la misma línea con lo que se observa en la figura 2.

El análisis de componentes principales (ACP) también arroja resultados cualitativamente en la misma línea que los perfiles medios. La figura 3 muestra el biplot de ACP, cabiendo recalcar que las primeras dos componentes explican algo más del 58 % de la variabilidad (información) total, y que no se observan outliers multivariados.

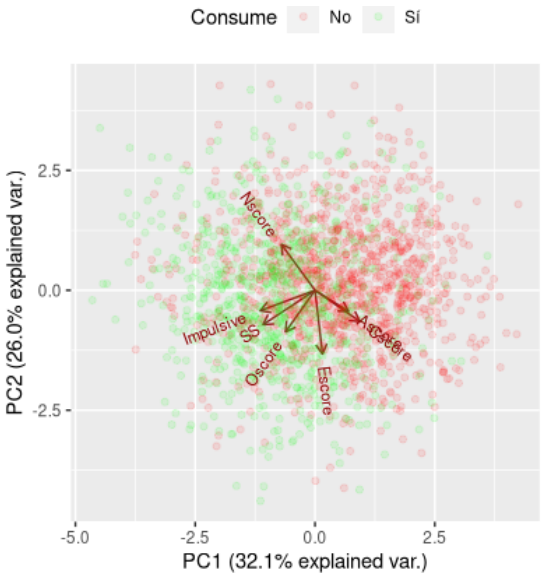

Figura 3. Biplot de componentes principales.

Al igual que para los perfiles medios, la figura 3 de ACP muestra la misma tendencia: para el grupo que consume altas las variables Oscore (apertura a experimentar), la impulsividad y el SS (búsqueda de sensaciones); y para el grupo que no consume altas las variables Ascore (complacencia) y el Cscore (conciencia); no observándose una influencia clara para las variables Nscore (neurotismo) y Escore (extraversión).

Si se realiza el mismo biplot de la figura 3 pero discriminando por rango de edad, se distin-

gue una tendencia a no consumir drogas para los mayores de 35 y a consumir en las personas pertenecientes al rango de edad 18-24. Algo similar se observa en el nivel de educación observándose una tendencia a asociar menor consumo con mayor nivel educativo, aunque todavía menos clara que en el caso de la edad y, dada la asimétrica distribución del nivel de educación en los datos, con menor asidero.

Correlaciones entre variables: mirando la figura 3 se puede malinterpretar la correlación entre las variables. Desde el dato duro arrojado por el cálculo de las correlaciones, destacamos un 0.62 entre Impulsive y SS; alrededor de un 0.4 entre Nscore y Escore, entre Nscore y Cscore y entre Oscore y SS; con el resto de las correlaciones en el orden de 0.3 o menor.

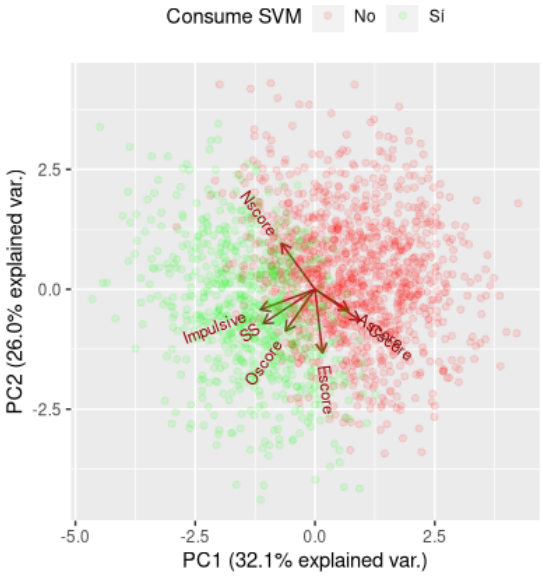

Figura 4. Biplot de la clasificación por SVM.

4. Clasificación supervisada

Para estas técnicas de clasificación se separaron los datos en entrenamiento y testeo, en 70 % y 30 % respectivamente, utilizando los datos de entrenamiento para construir el clasificador, y los datos de testeo para testearlo. Además se escalaron los datos de entrenamiento, y se utilizaron estos parámetros de escalado para escalar los datos de testeo. La métrica a maximizar fue siempre el accuracy, es decir el porcentaje de clasificación correcta, tanto positiva como negativa (clasificación positiva es que sí consuma drogas).

Teniendo en cuenta tanto la métrica a maximizar como que no se satisfacen los supuestos de normalidad multivariada ni homocedasticidad, como ya mencionamos en la sección 3, la técnica que mejores resultados arrojó fue la de support vector machine (SVM), clasificando correctamente (accuracy) el 72.5 % de las observaciones para testeo, consiguiendo además una precisión del 76.6 % (detección de falsos positivos) y un recall del 75 % (detección de falsos negativos). La figura 4 muestra el biplot de clasificación del SVM, es decir el biplot construido por ACP pero con la clasificación del SVM (léase aplicando el clasificador a la totalidad de los datos y no solamente a los de testeo). Si se compara con el biplot de clasificación real de la figura 3 se observa el buen comportamiento de este clasificador.

Resulta también interesante analizar el clasificador del tipo regresión logística. Además de presentar un accuracy similar a SVM, con una precisión algo menor y un recall algo mayor, lo interesante de este clasificador radica en analizar sus coeficientes, pues los mayores positivos están en las variables SS y Oscore, los mayores negativos en las variables Cscore y Ascore, y los relativamente más cercanos a cero en las variables Nscore, Escore y Impulsive. Salvo Impulsive, el resto está en directa correlación con lo observado en la comparación de los perfiles medios de la figura 2 y en lo observado en el ACP de la figura 3. Que Impulsive tenga un coeficiente cercano a cero se explica por su alta correlación lineal con SS (0.62 de coef. de corr.).

Dado que SVM y regresión logística arrojaron resultados similares en cuanto a métricas de evaluación (accuracy, precisión y recall) se procedió a analizar las observaciones (todas, como en la figura 4) clasificadas de manera distinta entre ambos clasificadores. En cantidad total son pocas, siendo poco más del 5 % (46 mal clasificadas por SVM pero bien por regresión logística, y 53 en el caso contrario, de un total de 1877). Analizando las observaciones donde hubieron diferencias, en ambos casos se detectaron tres grupos predominantes de observaciones: cercanas al promedio,

asociadas al Nscore, y asociadas al Escore. Es interesante relacionar esto con la tendencia observada en la figura 3 la cual resulta razonable interpretar indicando que el Nscore y el Escore no influyen a la hora de asociar la personalidad con el consumo de drogas. Es decir las observaciones donde los clasificadores difieren o bien son cercanas al promedio, o bien están asociadas a variables de la personalidad que parecen no influir respecto al consumo de drogas.

A pesar de no satisfacerse las hipótesis de normalidad multivariada y homocedasticidad, igualmente se entrenó un clasificador del tipo discriminante lineal (LDA), clasificando correctamente (accuracy) un 72 % de los datos de testeo, con una precisión del 79 % y un recall de 73.6 % (todos valores similares al SVM). Algo interesante de este clasificador LDA radica en analizar sus coeficientes, presentando resultados similares a la regresión logística, y por lo tanto avalando una similar interpretación de las variables en cuanto al consumo o no de drogas.

El clasificador del tipo discriminante cuadrático (QDA), que en principio tendría más sentido que el LDA al no necesitar la hipótesis de homocedasticidad, pero entrenado con la misma salvedad que para LDA de no cumplirse normalidad, arrojó una clasificación (accuracy) similar al LDA, pero con una precisión más baja (63.4 %) y un recall más alto (82.7 %). Este clasificador puede ser de utilidad si se desea focalizar la clasificación en evitar los falsos negativos pero manteniendo el accuracy.

Al no cumplirse el supuesto de normalidad, tampoco es más adecuado el clasificador del tipo discriminante robusto (RDA), que de todos modos fue entrenado, arrojando resultados similares al LDA (accuracy del 71.8 %, precisión del 77.6 % y recall del 73.7 %).

Cabe mencionar que en todos los casos los clasificadores arrojaron un resultado similar al mostrado para SVM en la figura 4, y por lo tanto avalando con mayor evidencia la relación entre variables y consumo o no de drogas antes mencionado para regresión logística y LDA, y también exhibido gráficamente por los perfiles medios de la figura 2 y por el biplot de ACP de la figura 3.

4.1. Conclusiones

La clasificación arroja evidencia en favor de asociar el consumo de drogas a las personalidades con alto Oscore (apertura a experimentar), SS (búsqueda de sensaciones) e Impulsive (impulsividad); y a las personalidades con alto Ascore (complacencia) y Cscore (conciencia) con el no consumo de drogas. Es decir hay evidencia para pensar que un test de personalidad contiene variables que arrojan información importante a la hora de intentar detectar a una persona con riesgo de comenzar a o que está consumiendo drogas.

Además esta clasificación arroja evidencia de que el Nscore (neurotismo) y Escore (extraversión) no tienen una clara influencia en el consumo o no de drogas.

5. Clasificación no supervisada

Para estas técnicas de clasificación se utilizaron todos los datos prescindiendo de su clasificación original (es decir si consume o no), y también fueron escalados. Recordemos que estas técnicas buscan estructuras o agrupamientos internos de los datos sin conocer ningún tipo de clasificación sobre los mismos. Cabe aclarar que el análisis se realizó asumiendo que no se posee ningún tipo de clasificación de los datos, e intentando discernir tres (potencialmente dos) grupos de variables que sinteticen: tendencia a experimentar, tendencia a ser cautos, y neutrales respecto de estas dos (potencialmente vacío).

La técnica que mejores resultados arrojó fue la de k-means, pues fue la que mejor agrupó a las observaciones respecto a la interpretación individual de las variables. La figura 5 muestra el biplot de agrupamiento arrojado por k-means, es decir el biplot construido por ACP pero con el agrupamiento dado por k-means, en donde se observa un grupo de observaciones con alto Ascore (complacencia) y Cscore (conciencia); y otro grupo con alto Impulsive (impulsividad), SS (búsqueda de sensaciones), Oscore (apertura a experimentar), y en menor medida Nscore (neurotismo). El Escore (extraversión) parece ser neutro. Es decir hay un grupo con una fuerte tendencia a experimentar de manera impulsiva, y otro grupo mucho más

complaciente y concienzudo.

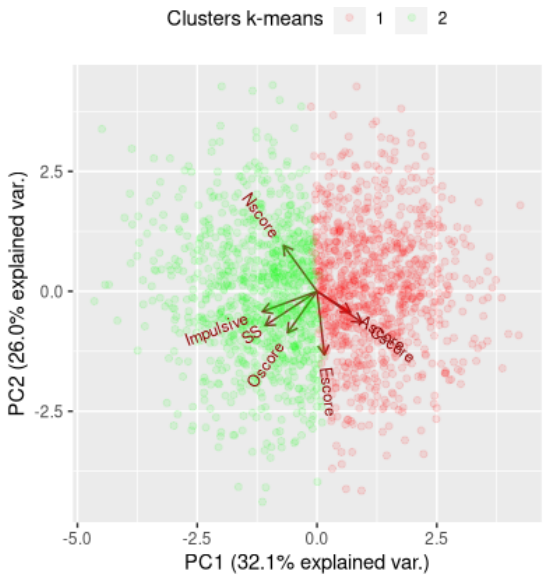

Figura 5. Biplot del agrupamiento por k-means.

Si computamos los perfiles medios de cada cluster, mostrados en la figura 6, podemos observar un comportamiento coherente con la figura 5, donde de nuevo se evidencian los mencionados dos grupos provenientes de interpretar los grupos de variables resultantes.

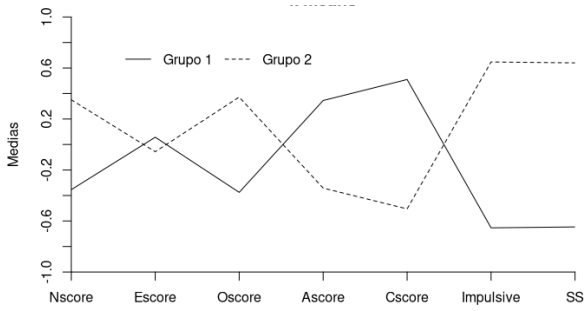

Figura 6. Comparación de los perfiles medios multivariados de los grupos de k-means.

La técnica de k-means, al no utilizar la clasificación real de los datos, no conoce a priori la cantidad de grupos reales. Por lo tanto es razonable evaluar internamente cuán bien se están separando los datos respecto a la cantidad de agrupamientos. En este sentido se utilizó el coeficiente de Silhouette, el cual arrojó resultados enteramente consistentes con que el mejor agrupamiento es en dos grupos, pues su valor bajó sistemáticamente desde dos hasta quince grupos.

También se probaron técnicas de agrupamiento jerárquico, pero en ningún caso se obtuvieron clasificaciones en grupos razonables de ser interpretados teniendo en cuenta los significados de las variables.

Por último también es interesante mostrar el resultado de k-means agrupando por variables (y no por observaciones como al principio). Tomando en cuenta el objetivo del análisis en este tipo de clasificación, tiene sentido agrupar a las variables en dos o en tres grupos. Para dos grupos el resultado fue:

- Grupo 1: Nscore, Oscore, Impusilve, SS.
- Grupo 2: Escore, Ascore, Cscore.

Si analizamos este agrupamiento respecto a las variables, es razonable interpretar al grupo 1 como las variables que sí influyen a la tendencia a experimentar de manera impulsiva, y al grupo 2 como las variables que no. Para tres grupos el resultado fue:

- Grupo 1: Oscore, Impusilve, SS.
- Grupo 2: Escore, Ascore, Cscore.
- Grupo 3: Nscore.

Si nuevamente analizamos este agrupamiento respecto a las variables, es razonable interpretar al grupo 1 como las variables que sí influyen a la tendencia a experimentar de manera impulsiva, al grupo 2 como las que no, y al grupo 3 como las neutras. Si comparamos este agrupamiento con las interpretaciones anteriormente extraídas del agrupamiento por observaciones (figuras 5 y 6), como única diferencia se observa que hay un intercambio entre Escore y Nscore respecto a ser neutra o a influir a la tendencia de experimentar.

5.1. Conclusiones

La clasificación arroja evidencia en favor de agrupar por un lado a las personalidades con alto Oscore (apertura a experimentar), SS (búsqueda de sensaciones) e Impulsive (impulsividad), y

agrupar por otro lado a las personalidades con alto Ascore (complacencia) y Cscore (conciencia). Es decir hay evidencia para pensar que un test de personalidad contiene variables que arrojan información importante a la hora de discernir personas con tendencia a experimentar de manera impulsiva, de personas más cautas y concienzudas.

Adicionalmente esta clasificación arroja evidencia de que el Nscore (neurotismo) y Escore (extraversión) no tienen una clara influencia a la hora de discernir la clasificación antes mencionada.

Referencias

[1] Drug consumption (quantified) Data Set
`https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29#`

[2] McCrae RR, Costa PT; *A contemplated revision of the NEO Five-Factor Inventory*;
Personality and Individual Differences. 2004; 36(3):587–596.

[3] Stanford MS, Mathias CW, Dougherty DM, Lake SL, Anderson NE, Patton JH; *Fifty years of the Barratt Impulsiveness Scale: An update and review*;
Personality and Individual Differences. 2009; 47(5):385–395.

[4] Zuckerman M; *Behavioral expressions and biosocial bases of sensation seeking*;
New York: Cambridge University Press; 1994.