Cuadernos de Psicología del Deporte, vol. 15, 1, 31-40
Recibido: 18/07/2014
Aceptado: 24/11/2014

Murcia (España)
ISSN edición impresa: 1578-8423
ISSN edición web (*http://revistas.um.es/cpd*): 1989-5879

Algoritmos de clasificación y redes neuronales en la observación automatizada de registros

Classification algorithms and neural networks in automated observation records

Algoritmos de classificação e redes neurais em registros de observação automatizados

S.L. González-Ruiz*, I. Gómez-Gallego, J.L. Pastrana-Brincones y A. Hernández-Mendo

Universidad de Málaga

Resumen: El objetivo del presente estudio es analizar los datos obtenidos a través de una plataforma *on-line*, mediante diferentes técnicas de clasificación y aprendizaje orientadas al descubrimiento del conocimiento. Se aplican técnicas de minería de datos para obtener relaciones de fiabilidad que informen del interés de los usuarios por cumplimentar de manera rigurosa el cuestionario *on-line* atendiendo al modo de realizar el mismo. Aunque existen técnicas que nos permiten observar el comportamiento de los usuarios mientras realizan el cuestionario, en este caso se emplean Redes Neuronales Artificiales para predecir el comportamiento de aquellos, atendiendo a variables obtenidas al realizar el cuestionario. La muestra consta de 1.636 participantes de diferentes zonas geográficas y rangos de edad, obtenida al contestar de manera anónima o identificada al cuestionario Inventario Psicológico para el Seguimiento de Talentos Deportivos (IPSETA). Los resultados obtenidos mediante las diferentes técnicas de análisis informan que el género femenino prefiere realizar el registro en la plataforma para cumplimentar el cuestionario, alcanzando un alto porcentaje de fiabilidad (70%).
Palabras clave: minería de datos, Redes Neuronales Artificiales, WEKA reglas de asociación, análisis de grupos.
Abstract: The aim of this study is to analyse a set of data got through an *on-line* platform, using some ranking and knowledge oriented discovery rules techniques. Data mining techniques are applied to obtain a reliable relationship which can show the interest of the users in order to fill rigorously the *on-line* questionnaire attending to the way they do. Although there are programming techniques which allows us to observe the behaviour of users while filling the survey, current work uses artificial neural networks to predict their behaviour, based on variables obtained from the own survey. The sample is made up of 1,636 participants from different geographical areas and age ranges, obtained anonymously by answering the IPSETA questionnaire which is used for a psychological monitoring of sport talents. The results obtained using the analysis techniques show that females prefer to register on the platform to fill the survey, getting a high reliability (70%).
Key words: data mining, artificial neural networks, WEKA, association rules, cluster analysis.
Resumo: O objetivo deste estudo é analisar um conjunto de dados através de uma plataforma *on-line* utilizando diferentes técnicas ou regras de descoberta de conhecimento orientado. Técnicas de mineração de dados são aplicados para obter uma relação de confiança é relatado o interesse dos usuários de forma confiável para o preenchimento do questionário *on-line* modo de endereçamento de fazer o mesmo. Embora existam técnicas de programação que nos permite observar o comportamento dos usuários durante a realização da pesquisa, graças às novas ferramentas podem prever o comportamento do mesmo, com base em variáveis obtidas através da realização de questionário. A amostra foi composta por 1.636 participantes de diferentes áreas geográficas e faixas etárias obtidas anonimamente responder o questionário e identificar formas IPSETA Os resultados obtidos pelas diferentes técnicas é relatado que o sexo feminino que você preferir registrar na plataforma para o questionário
Palabras-chave: mineração de dados, redes neurais artificiais, WEKA, regras de associação, análise de cluster.

1. Introducción

El objetivo de las técnicas de minería de datos (*Data Mining*) consiste en procesar y analizar la información para encontrar patrones repetitivos, tendencias, o reglas que expliquen el comportamiento de los datos en un determinado contexto (López, 2007; Vieira, Ortíz y Ramírez, 2009). Básicamente, las técnicas de descubrimiento de conocimiento en bases de datos (*Knowledge Discovery in Database o* KDD) surgen como una forma de comprender su contenido. Para ello utilizan procedimientos y técnicas que abarcan desde la estadística hasta técnicas de Inteligencia Artificial y de Redes Neuronales Artificiales, en toda su extensión. Según Hand, Mannila y Smyth (2001), la información a procesar y analizar, por lo general, está formada por grandes conjuntos de datos, y un análisis pormenorizado de este conjunto permitirá extraer conocimiento explicativo del mismo.

Los campos de investigación de las técnicas KDD son muy dispares, y engloban desde la supercomputación, hasta

Dirección para correspondencia [Correspondence address]: Sergio Luis González Ruiz. Departamento de Psicología Social. Universidad de Málaga (España). E-mail: sergioluisgr@gmail.com

la estadística, bases de datos, reconocimiento de patrones, y centran su atención principalmente en el proceso de extraer, almacenar y acceder a conocimiento en grandes volúmenes de datos. Ya existen estudios similares realizados con esta técnica en el ámbito de la Psicología, (Gervilla, Cajal, Jiménez y Palmer, 2010; Gervilla, Jiménez, Montaño, Sesé, Cajal y Palmer, 2009; Montaño, Gervilla, Cajal y Palmer, 2014; Sancesario, 2012; Romeo, Codina, Yepes, Pestana y Guardia, 2013). Otros estudios sobre la aplicación de esta técnica en otros ámbitos lo podemos encontrar en Palma, Palma y Pérez (2009).

En este sentido, uno de los conceptos más importantes relacionados con las técnicas KDD y la minería de datos es el concepto de aprendizaje automático o *machine learning*, un área cuyo objetivo es desarrollar modelos computacionales capaces de inducir conocimiento a partir de datos. Entre los principales modelos o algoritmos de aprendizaje automático están aquellos denominados de caja negra, como pueden ser las Redes Neuronales Artificiales, y métodos orientados hacia el conocimiento, como los Arboles de Decisión o las Reglas de Asociación.

En este trabajo, y como una aplicación práctica de estos algoritmos, presentaremos los resultados obtenidos mediante su aplicación a un conjunto de datos procedentes de un formulario, utilizando un algoritmo de cada clase, una Red Neuronal Artificial, concretamente una red de retro-propagación (Galushkin, 2007), y algoritmos para la obtención de Reglas de Asociación y agrupamiento de instancias (Witten y Frank, 2005).

2. Método

2.1. Participantes

La muestra del estudio está compuesta por 1.636 participantes de diferentes zonas geográficas (España, Colombia, Argentina, México, Chile, etc.) de los cuales 400 son hombres (24,45%), 1.178 mujeres (72%) y 58 no contestaron la pregunta sobre el género (3,55%), con edades comprendidas entre los 10 y 56 años, (media = 23,9; DE=7,8).

Cabe destacar que de esa muestra, 308 participantes realizaron el cuestionario de manera anónima y 1.328 lo cumplimentaron registrándose en la plataforma de evaluación. Las características de la población aparecen en la tabla 1.

Tabla 1. Características socio-demográficas de los participantes

Muestra		**n**	**%**
Edad *(años)*	10-20	400	24,46
	20-30	1006	61,52
	30-40	69	4,20
	40-50	82	5,00
	50-60	18	1,11
	No responde	61	3,71
Género	Masculino	400	24,45
	Femenino	1178	72,00
	No responde	58	3,55
Estudios	Estudios Superiores	1257	76,84
	Estudios Primarios	11	0,68
	Estudios medios	263	16,06
	Sin Estudios	4	0,25
	No responde	101	6,18
Nacionalidad	Argentina	27	1,65
	Chile	5	0,30
	Colombia	1268	77,50
	España	184	11,24
	México	53	3,24
	otros	35	2,13
	No responde	64	3,91
Estado Civil	Viudo/a	2	0,12
	Casado/a	70	4,28
	Divorciado/a	63	3,85
	Soltero/a	1428	87,29
	No responde	73	4,46

2.2. Instrumentos

Se ha utilizado el cuestionario Inventario Psicológico para el Seguimiento de Talentos Deportivos IP-SETA (Yubelly-García, 2005) compuesto por 19 ítems, que tiene como objetivo la detección y seguimiento de talentos deportivos, evaluando tres variables: motivación intrínseca, motivación de logro y autoeficacia. Para la obtención de las reglas de asociación y agrupamiento se ha usado la herramienta de minería de datos WEKA versión 3.6.8 (Witten, Frank, Trigg, Hall, Holmes y Cunningham, 1999; Witten y Frank, 2005). Para el análisis de datos se utilizó el programa Microsoft Excel en su versión 2010. La recogida de la información se realizó a través de la plataforma de Evaluación Psicosocial *on-line* MenPas (www.menpas.com)(González-Ruiz, Hernández-Mendo y Pastrana-Brincones, 2010).

2.3. Procedimiento

Es habitual encontrar investigaciones donde la recolección de la muestra se realiza a través de internet, (Chung, Des Roches, Meunier y Eavey, 2005; Nielsen, Stenstrom y Levin, 2006; Gosling, Vazire, Srivastava y John, 2004) con muestras de 9.600, 23.900 y 361.000 participantes respectivamente. Según diversos autores (Carlbring, et al, 2007; Holländare, et al., 2010; Hedman, et al., 2010) no existen diferencias significativas entre realizar el cuestionario en papel o cumplimentarlo *on-line*.

La muestra se recogió en el periodo comprendido entre el 13/12/2012 y 24/02/2014, a través de la plataforma www.menpas.com. Los usuarios (registrados y no registrados) han accedido a la plataforma y cumplimentado una de las dos implementaciones posibles del cuestionario IP-SETA (Yubelly-García, 2005): anónima y nominal. Entre los datos obtenidos por la aplicación se encuentra el tipo de usuario (REGISTRADO, NO REGISTRADO), el número de horas semanales dedicadas al deporte, el género (MASCULINO, FEMENINO, NO RESPONDE), el deporte practicado, estado civil, profesión, tiempo empleado en responder a cada ítem del cuestionario y tiempo total en responder el cuestionario, entre otros.

El interés de este trabajo se centra principalmente en estimar la relación entre la fiabilidad en completar el cuestionario y el perfil de usuario. Para ello se tendrá en cuenta fundamentalmente la información relacionada con el modo mediante el cual el usuario realiza el cuestionario, sin tener en cuenta las respuestas a cada uno de los ítems del mismo.

Esta fiabilidad se define a partir del tiempo dedicado a completar los ítems del cuestionario, en su totalidad. Esta variable tiempo se mide en segundos y es obtenida gracias a técnicas implementadas en el cuestionario que ayudan a controlar el comportamiento de los usuarios mientras están realizando el cuestionario (Stieger y Reips, 2010).

3. Técnicas de análisis de datos

3.1. Reglas de asociación

Las Reglas de Asociación intentan identificar o descubrir relaciones consistentes en bases de datos usando diferentes medidas de interés. Agrawal, Imielinski y Swami (1993) introdujeron el concepto de "reglas fuertes" para descubrir regularidades entre productos en transiciones comerciales a gran escala registrados en puntos de venta de supermercados. Así por ejemplo, la regla {*onions,potatoes*}=>{*burguer*} encontrada en el conjunto de datos de venta del supermercado indicaría que si el cliente compra cebollas y patatas juntos, es probable que además compre carne para hamburguesas. Tal información puede resultar importante como base para la toma de decisión relativa a actividades de marketing, tales como precios promocionales en determinados productos y localización de productos en la superficie.

Las reglas de asociación también pueden ser usadas en otros contextos, incluyendo análisis Web, detección de intrusos, producción continua y bioinformática. Un problema asociado a esta técnica de descubrimiento del conocimiento es que, en todas las aplicaciones, el resultado será abultado; es decir, encontraremos un grandísimo conjunto de reglas de asociación. Sin embargo, es importante realizar un análisis minucioso con el fin de realizar un proceso de eliminación de aquellas reglas obvias para el conocimiento, y mantener aquellas que son de interés para el estudio.

Para seleccionar las reglas interesantes de todo el conjunto generado por la aplicación, existen algunas restricciones en varias medidas de significancia e interés que pueden ser usadas. Las restricciones más conocidas son umbrales mínimos de soporte y confianza. El **Soporte** Supp (X) de un itemset X (itemset es un conjunto de atributos) es definido como la proporción de transacciones en el conjunto de datos que contienen al itemset X. La **confianza** de una regla es definida como:

$$Conf(X=>Y)=Supp(XUY)/Supp(X)$$

y puede ser interpretada como una estimación de la probabilidad condicional P(Y/X). La confianza puede ser interpretada como la probabilidad condicionada de un evento X, dado otro evento Y (P(Y/X)).

Existen además otras medidas como son el **Lift** y la **Convicción**, que si bien no suelen ser usadas en la mayoría de las ocasiones, sí que pueden ser de ayuda como medidas complementarias a la confianza.

3.2. Análisis de Grupos

El análisis de grupo pertenece, al igual que otras tipologías y que el análisis discriminante, al conjunto de técnicas que tiene por objetivo la clasificación de los individuos en grupos maximizando la homogeneidad intra-grupo y la mayor diferencia inter-grupos. Existen numerosos algoritmos de agrupamiento, clasificados comúnmente en jerárquicos y no-jerárquicos.

En este estudio, se aplicará un algoritmo no jerárquico denominado *Simple EM* (*Expectation maximisation*), es uno de los más simples, rápidos y precisos en los resultados, siendo estos similares a los obtenidos por otros métodos disponibles en la aplicación. El algoritmo se ha configurado para un máximo de iteraciones de 300, valor recomendado por la *Suite* WEKA.

EM asigna una distribución de probabilidad a cada instancia o ejemplo que indica la probabilidad de pertenecer a cada uno de los grupos. EM decide cuantos grupos crear median-

te validación cruzada, aunque se puede especificar cuantos grupos se quieren generar. La validación cruzada ejecutada determina el número de grupos mediante los siguientes pasos:

1. El número de grupos se inicializa a 1.
2. El conjunto de entrenamiento se divide en 10 segmentos.
3. El algoritmo se ejecuta 10 veces usando los 10 segmentos.
4. El *log- likelihood* (medida de similitud estadística) es promediado sobre los 10 resultados.
5. Si el *log-likelihood* ha incrementado el número de grupos, el algoritmo continúa en el paso 2.

Las técnicas utilizadas para la obtención de reglas de asociación y agrupamiento son parte del módulo de aprendizaje de la herramienta de minería de datos WEKA v. 3.6.8 (*Waikato Environment for Knowledge Analysis*). Esta herramienta es un entorno para experimentación de análisis de datos que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario. WEKA se distribuye como software de libre distribución, y desarrollado en Java. Está constituido por una serie de paquetes de código abierto con diferentes técnicas de pre-procesado, clasificación, agrupamiento, asociación y visualización, así como facilidades para su aplicación y análisis de prestaciones cuando son aplicadas a los datos de entrada seleccionados. Estos paquetes pueden ser integrados en cualquier proyecto de análisis de datos, e incluso pueden extenderse con contribuciones de los usuarios que desarrollen nuevos algoritmos. Con objeto de facilitar su uso por un mayor número de usuarios, WEKA además incluye una interfaz gráfica de usuario para acceder y configurar las diferentes herramientas integradas.

4. Resultados

4.1. Reglas de Asociación

Se centra la atención en el conjunto de variables que serán indicativos del grado de compromiso del encuestado en la realización del cuestionario, atendiendo al tiempo empleado en completar el mismo. De esta forma, el atributo tiempo total empleado en realizar el test ha sido discretizado en 3 partes iguales, considerando el segundo intervalo (aquel que está en medio), como el tiempo empleado normalmente por un usuario con algún interés en las preguntas del cuestionario. Por eliminación, se consideran no interesados aquellos usuarios que emplearon tiempos demasiado bajos o demasiado altos. También se ha discretizado el atributo número de horas dedicadas al deporte.

En relación a la obtención de las Reglas de Asociación, WEKA mantiene como herramienta de análisis el algoritmo *a priori*. Hay que señalar que cualquier algoritmo o procedimiento aplicado para la obtención de Reglas de Asociación puede generar una cantidad importante de éstas. Es útil prestar atención sólo a aquellas que arrojan unos valores altos del parámetro Soporte. En este caso, y aplicando esto último, aparecen las siguientes reglas:

1. Fem Pt2 -> Reg H1 *Sop (0,99)*
2. Fem Pt2 H1 -> Reg *Sop (0,98)*
3. Fem Pt2 -> Reg*Sop (0,99)*

siendo:

Fem: género femenino,
Pti: Tercil en el tiempo dedicado a rellenar el cuestionario,
Hi: Tercil del número de horas dedicadas al deporte,
Reg: valor que indica el registro de los datos personales del usuario.

Así por ejemplo, la lectura de la primera regla sería: con una confianza del 99%, los usuarios de género femenino que respondieron en el tercil 2 de tiempo estaban registrados y practicaron deporte un número de horas semanales correspondientes al tercil 1. La segunda regla obtenida es una redundancia de la primera. Mediante esta técnica la mayoría de las reglas obtenidas son redundancias de otras con significado más fuerte. La tercera regla sí es más concluyente, ya que se obtiene una confianza del 99%, y su lectura es: los usuarios del género femenino emplean un tiempo correspondiente al segundo tercil y además se registran en la plataforma.

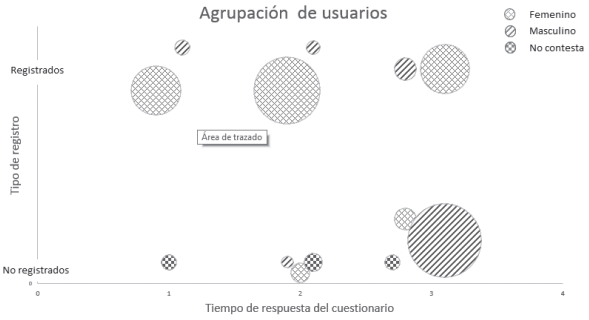
El análisis de las reglas obtenidas mediante el algoritmo, después de eliminar aquellas reglas realmente obvias o irrelevantes, muestran que, con un alto grado de confianza, las personas del género femenino prefieren realizar el cuestionario en la plataforma de manera registrada (es decir, no hacer el cuestionario de manera anónima). Además, existe una evidencia que se contrastará después con el análisis de grupos, y es que siempre que encontremos en las reglas el valor de Pt2 (interés en completar el cuestionario), son los usuarios de género femenino los que acompañan en la regla.

4.2. Análisis de grupos

En el eje X de la figura 1 se representa el tercil de respuesta al cuestionario (1= tiempo excesivamente corto, 2= tiempo normal, 3= tiempo excesivamente largo), y, en el eje Y, usuarios registrados y no registrados. En el eje X se han diferenciado los tipos de usuario según su género (masculino, femenino, y no contesta). Claramente se puede apreciar que la gran mayoría de los individuos de género femenino suelen inscribirse, es decir, el género masculino prefiere rellenar el cuestionario de forma anónima. También se puede apreciar que una gran fracción de los encuestados que caen en el tercil 2, es decir, el más fiable, son de género femenino.

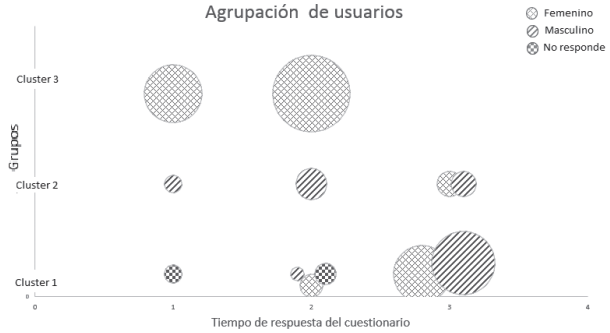
Se utiliza la facilidad de la función *Jitter*. Esta opción permite introducir un desplazamiento aleatorio (ruido) en las instancias, con objeto de poder visualizar todas aquellas que comparten un par de valores de atributos simbólicos, de manera que se puedan observar la agrupación de instancias que aparece en cada región.

Figura 1. Agrupamiento obtenido por EM



En la figura 2, se ha representado el número de grupos obtenido por el algoritmo *Simple EM* en función del tercil de tiempo dedicado a completar el cuestionario. Uno de los aspectos principales que se observan al ejecutar este tipo de algoritmo sobre los datos disponibles es que el atributo principal que el algoritmo ha tomado para construir los grupos ha sido el género de los encuestados. Además, claramente se aprecia como el género femenino, apilado mayormente en el grupo 3 (cluster 3 en la figura), parece más interesado por completar la entrevista, mientras que el género masculino se encuentra algo más disperso, aunque algún porcentaje sí se decante por una realización interesada del cuestionario. De hecho, el género masculino está disperso en dos grupos, lo que indica que no siguen una misma línea de actuación, como sí parece que lo hace el género femenino.

Figura 2. Grupos obtenidos por *Simple EM*



5. Redes Neuronales Artificiales

Determinados algoritmos y métodos del aprendizaje computacional permiten obtener buenos resultados de predicción una vez entrenado el modelo con un número conocido de datos, aplicando alguna técnica de presentación de los mismos, esperando que el modelo produzca una respuesta, sino idéntica, sí cercana a la correcta. De entre estos modelos predictivos, las *Redes Neuronales Artificiales* (RNA) han constituido en los últimos tiempos un foco de investigación importante y con una actividad intensa, siendo un paradigma de aprendizaje computacional muy extendido en la resolución de problemas de diversas áreas de la Ingeniería y la Ciencia, como el problema del viajante (Hilera y Martinez*,* 1995). Debido a sus excelentes capacidades de ajuste, las RNA se aplican de manera exitosa en distintos ámbitos científicos, sociales y tecnológicos: manufacturación, biología, finanzas, previsión del tiempo, análisis de tendencias y patrones, etc. Entre las propiedades más destacables, la capacidad de generalización confiere a estos modelos una amplia aplicabilidad en tareas de clasificación y aproximación, entendiendo capacidad de generalización como la propiedad de la RNA para computar correctamente ejemplos de un conjunto de datos que no le han sido mostrados previamente, después de una fase de entrenamiento con ejemplos del mismo conjunto de datos.

Sin embargo, existen algunas propiedades importantes de los datos que deben ser tenidas en cuenta cuando deseamos aplicar algunos de estos algoritmos para predicción, y que influyen notablemente en la capacidad de generalización del modelo. De entre ellas, dos propiedades importantes son la calidad de los datos de entrenamiento (es decir, que éste se realice con una porción significativa de datos del problema), y su complejidad. La complejidad de un conjunto de datos se puede cuantificar de muchas maneras, y fundamentalmente dará una idea del grado de facilidad con el que un conjunto de datos puede ser aprendido y, en caso de las RNA, de la arquitectura y topología de la misma.

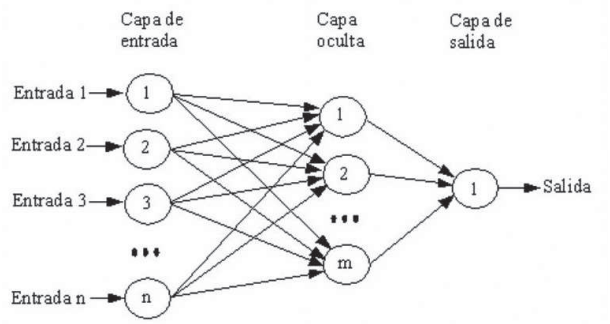
5.1. Aplicación de una RNA a la predicción del género del usuario

Como aplicación de las RNA como modelos computacionales de aprendizaje, se ha seleccionado un método de presentación de datos denominado *validación cruzada* (Flórez y Fernández 2008). Mediante este método, el conjunto total de ejemplos se divide en un número determinado de partes iguales denominadas segmentos, habitualmente de 10. El entrenamiento del modelo consistirá en presentar 9 segmentos para entrenamiento y uno para clasificación, rotando este último ente los 10 segmentos. Por tanto, el procedimiento se repite 10 veces, una vez por cada segmento para la clasificación.

Otra característica importante de las RNA es el algoritmo

de aprendizaje usado. Existe una gran variedad, diferenciándolos fundamentalmente en la función de coste a minimizar. En este caso, se ha utilizado el algoritmo de *Back-Propagation*, o retroalimentación.

Figura 3. Topología de una RNA



En la figura 3 se presenta una RNA completamente conectada. Para esta configuración, hemos seleccionado como atributo a aproximar el género del usuario que cumplimenta el cuestionario, en función del tiempo requerido para ello, del tipo de registro y del número de horas semanales empleadas en realizar algún tipo de deporte.

La construcción del modelo tomó un tiempo de 11,84 segundos, y después de la etapa de aprendizaje, se obtuvo en el segmento de test que el porcentaje de aciertos fue de 80%.

La matriz de confusión es la siguiente:

Figura 4. Matriz relacionada

```
  a    b    c   <-- classified as
184  216    0 |    a = Masculino
 65 1114    0 |    b = Femenino
 43   14    0 |    c = No_responde
```

Esta matriz relaciona el número de instancias de una determinada clase que han sido clasificadas correctamente (diagonal principal), y aquellas que han sido clasificadas como pertenecientes a otra clase. A modo de ejemplo, en el caso de la matriz de confusión de la figura 4, se puede observar que 65 instancias de la clase *a* se han clasificado erróneamente como clase *b*.

Como se puede apreciar en este caso, el género femenino lo clasifica en un alto porcentaje de los casos, mientras que el genero de No_responde falla de manera absoluta.

Como se indicó anteriormente, el comportamiento de una RNA para predecir datos nuevos depende en gran medida de la calidad de los datos de entrenamiento, y de su complejidad. Esta característica se pone de manifiesto en este caso, cuando se pretende predecir el grado de fiabilidad que podemos esperar de un usuario conociendo los demás atributos (género, registro, etc), ya que al ejecutar este modelo con los datos disponibles, el porcentaje de acierto está alrededor del 70%, porcentaje que se puede considerar relativamente bueno, lo que además indica cierta complejidad en los datos a analizar.

La matriz de confusión, en este caso, es la siguiente:

Figura 5. Matriz de confusión

```
  a   b   c   <-- classified as
420 107  16 |   a = 1
313 126  99 |   b = 2
166  58 331 |   c = 3
```

Se toman datos de carácter descriptivo y Alfa de Cronbach, proporcionados por la plataforma MenPas para contrastar los obtenidos por el procedimiento de redes neuronales.

Tabla 2. Estadísticos y Alfa de Cronbach suministrados por la plataforma MenPas para los usuarios identificados.

Usuarios	Factores	n	Media	Moda	Varianza	Sx	Alfa Cronbach	Alfa Cronbach Global
Totales	M_Intrinseca	1328	15,014	17	4,911	2,216	0,688	0,809
	M_Logro	1328	15,652	16	2,902	1,704	0,493	
	Autoeficacia	1328	17,995	18	5,831	2,415	0,666	
Femeninos	M_Intrinseca	1112	15,035	17	4,784	2,187	0,684	0,800
	M_Logro	1112	15,65	16	2,742	1,656	0,468	
	Autoeficacia	1112	18,078	18	5,566	2,359	0,656	
Masculinos	M_Intrínseca	216	14,907	17	5,513	2,348	0,708	0.843
	M_Logro	216	15,662	17	3,734	1,932	0,593	
	Autoeficacia	216	17,569	18	6,918	2,63	0,703	

Tabla 3. Estadísticos y Alfa de Cronbach suministrados por la plataforma MenPas sobre los usuarios anónimos.

Usuarios	Factores	n	Media	Moda	Varianza	Sx	Alfa Cronbach	Alfa Cronbach Global
Total	M_Intrínseca	308	14,515	17	8,93	2,988	0,830	0,900
	M_Logro	308	15,319	16	6,276	2,505	0,766	
	Autoeficacia	308	17,264	19	11,127	3,336	0,805	
Femeninos	M_Intrínseca	66	13,788	17	10,346	3,217	0,869	0,921
	M_Logro	66	14,773	16	5,592	2,365	0,731	
	Autoeficacia	66	16,182	20	15,325	3,915	0,847	
Masculinos	M_Intrínseca	184	14,587	17	8,904	2,984	0,821	0,891
	M_Logro	184	15,565	16	5,85	2,419	0,755	
	Autoeficacia	184	17,571	19	9,657	3,108	0,793	
No contestan	M_Intrínseca	58	15,123	17	6,418	2,533	0,780	0,887
	M_Logro	58	15,158	18	7,849	2,802	0,822	
	Autoeficacia	58	17,526	21	9,278	3,046	0,748	

Los datos suministrados por la plataforma MenPas se distinguen entre usuarios identificados y anónimos. En la tabla 2 se recogen los valores estadísticos y de fiabilidad para los identificados y en la tabla 3 para los usuarios que han cumplimentado el cuestionario de forma anónima. Si se observa la columna Alfa de Cronbach global de las dos tablas:

Los datos son similares a los obtenidos mediante análisis de Redes Neuronales (el Alfa de Cronbach es ligeramente superior). El Alfa de Cronbach global para usuarios de género femenino está en la misma línea de los resultados estimados por el análisis de RNA. En esta misma línea se sitúan los factores Motivación extrínseca y Autoeficacia. Sin embargo, los resultados alcanzados por los usuarios masculinos son mejores que los femeninos en los usuarios identificados y es mayor en los usuarios anónimos.

Los valores obtenidos en la tabla 3 (usuarios anónimos) son ligeramente mayores que los pertenecientes a los usuarios identificados (tabla 2), situación que se produce también para cada uno de los factores.

Respecto al valor más repetido (moda) y desviación típica, no se aprecian diferencias a destacar entre las escalas de las dos tablas. No pasa lo mismo con la varianza, donde los valores de las dos tablas son dispares.

6. Discusión

Los métodos tradicionales, para obtener a partir de una serie de datos predicciones, pueden ser lentos y costosos. Se ha usado esta novedosa técnica, para predecir la fiabilidad de los usuarios, al rellenar el cuestionario, dependiendo de variables como el género, tiempo total de realización, deporte practicado y tipo de identificación. Tras aplicar las técnicas expuestas, los resultados obtenidos ayudan a comprender mejor el interés de los usuarios a la hora de realizar pruebas *on-line*.

Se ha contrastado los datos obtenidos por el análisis de RNA con datos procedentes de la plataforma MenPas y se comprueba que son coherentes, aunque se aprecia que dependiendo del anonimato se incrementa la fiabilidad. Este dato señala que el anonimato permite disminuir la deseabilidad social (Muhlenfeld, 2005). Se puede comprobar las diferencias entre los índices Alfa globales y por factores entre usuarios anónimos e identificados (ver tablas 2 y 3).

Esta cuestión de anonimato y el incremento de la fiabilidad asociada, plantea otra cuestión en la línea del estudio de la estabilidad de la medida (fiabilidad). Tradicionalmente se ha estudiado esta cuestión a través de indicadores –que podríamos denominar estáticos-. La incorporación de los estudios *on-line* presentan nuevas posibilidades. Una de las mejores posibilidades es poder obtener muestras de grandes dimensiones (Gosling, Vazire, Srivastava y John, 2004) que permitan nuevas posibilidades de análisis como la utilización de las redes neuronales. Otras posibilidades están relacionadas con la utilización de nuevas variables, como es el estudio de los tiempos de respuesta a los ítems o al total del cuestionario, el número de modificaciones de cada ítem o el estudio del orden de respuesta. Esta última cuestión permite comprobar, por ejemplo, si ha contestado el cuestionario de forma creciente, decreciente, aleatoria, si ha modificado el valor de un ítem, etc.

Tabla 4. Tiempos de respuesta a los items proporcionados por MenPas.

tn1	tn2	tn3	tn4	tn5	tn6	tn7	tn8	tn9	tn10	tn..	tn19
9.9	6,5	5,8	4,5	4,6	2,8	3,2	3,4	3,1	6,1	..	3,2

Figura 6. Orden de respuesta a un ítem proporcionado por MenPas.

Nota: Se observa que primero realiza el ítem 1 (RBL1) con valor 3 (V3) en un tiempo de 57,40 segundos, luego realiza el ítem 2 (RBL2) con valor 3 (V3) en un tiempo de 12,50 segundos, etc. El separador entre ítems es el carácter @

Esta situación de obtención de datos automatizados a través de investigaciones *on-line* (p.e. MenPas) que permite la obtención de grandes muestras permite la utilización de procedimientos análiticos que hasta ahora no habían sido utilizados con profusión en este área y además pone un reto, en la utilización de nuevas variables para el estudio de las propiedades metodológicas psicométricas de las herramientas y procesos de medida, tanto desde el punto de vista cualitativo, cuantitativo o a través de *Mixed Methods* (Anguera, Camerino, Castañer y Sánchez-Algarra, 2014).

Consideramos que el uso de redes neuronales para el estudio de datos obtenidos por procedimientos semi y automatizados es óptimo y adecuado. Y además en el caso de estudios *on-line*, la utilización de estrategias y datos complementarios puede ayudar a reducir los sesgos producidos por la deseabilidad social y la percepción de los usuarios.

Según Muhlenfeld (2005), el anonimato en los cuestionario, anima a los participantes a dar respuestas más honestas, es posible que los valores obtenidos en tabla 3, ligeramente superiores, sean consecuencia de este factor. La fiabilidad de las evaluaciones se puede ver afectada por la distorsión del participante que responde el cuestionario. Según Pérez, Labiano y Brusasca (2010) a mayor edad, disminuye la deseabilidad social.

Como trabajo futuro se pretenden realizar estudios mediante técnicas similares en bases de datos con muestras de tamaño superior.

Aplicaciones Prácticas

Los resultado obtenidos, como se ha podido comprobar, pueden ayudar a los investigadores, y en especial a los psicólogos, a desarrollar investigaciones sobre distintos tipo de comportamiento con datos semi y automatizados, comprobando la utilización de nuevas variables en la optimización de los procesos de fiabilidad y validez.

Becas, ayudas o soporte financiero: Este artículo se ha realizado con el apoyo y financiación del Proyecto I+D+I *Observación de la interacción en deporte y actividad física: Avances técnicos y metodológicos en registros automatizados cualitativos-cuantitativos*. Secretaria de Estado de Investigación, Desarrollo e Innovación del Ministerio de Economía y Competitividad [Referencia: DEP2012-32124].

Referencias

1. Agrawal, R., Imielinski, T. y Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of the 1993 AcmSigmod International Conference on Management of Data,* 207-216.
2. Anguera, M.T., Camerino, O., Castañer, M. y Sánchez-Algarra, P. (2014). Mixed methods en la investigación de la actividad física y el deporte. Mixed methods en la investigación de la actividad física y el deporte *Revista de Psicología del Deporte*, *23*(1), 123-130.
3. Carlbring, P., Brunt, S., Bohman, S., Austin, D., Richards, J., Ost, L. y Andersson, G. (2007). Internet Vs paper and pencil administration of questionnaires commonly used in panic/agoraphobia research. *Computers in Human Behavior, 23*, 1421–1434.
4. Chung J.H., Des Roches C.M., Meunier J. y Eavey, R.D. (2005). Evaluation of noise-induced hearing loss in young people using a web-based survey technique. *Pediatrics*, *115*, 861-867.
5. Flórez R. y Fernández J.M. (2008). *Las Redes Neuronales Artificiales, fundamentos teóricos y aplicaciones prácticas*. Serie Metodología y análisis de datos en Ciencias Sociales. Madrid: Netbiblo S.L.
6. Galushkin, A.I. (2007). *Neural networks theory*. New York: Springer.
7. Gervilla, E., Cajal, B., Jiménez, R. y Palmer, A. (2010). Estudio de los factores asociados al uso de sustancias en la adolescencia mediante reglas de asociación. *Adicciones, 22*(4), 293-300.
8. Gervilla, E., Jiménez, R., Montaño, J.J., Sesé, A., Cajal, B. y Palmer, A. (2009*).* La metodología del Data Mining. Una aplicación al consumo de alcohol en adolescentes. *Adicciones, 21*(1), 65-80.
9. González-Ruiz, S. L., Hernández-Mendo, A., Pastrana-Brincones, J.L. (2010). Herramienta software para la evaluación psicosocial de deportistas y entornos deportivos. *Lecturas: EF y Deportes. Revista Digital,* 15(144).
10. Gosling, S.D., Vazire, S., Srivastava, S. y John, O.P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions. *American Psychologist, 59*, 93-104.
11. Hand, D., Mannila, H. y Smyth, P. (2001). *Principles of Data Mining*. Cambridge, MA: The MIT Press.
12. Holländare, F., Andersson, G. y Engström, I. (2010). A Comparison of psychometric properties between internet and paper versions of two depression instruments (BDI-II and MADRS-S) administered to clinic patients. *Journal of Medical Internet Research, 12, e49.*
13. Hedman, E., Ljótsson, B., Rück, C., Furmark, T., Carlbring, P., Lindefors, N. y Andersson, G. (2010). Internet administration of self-report measures commonly used in research on social anxiety disorder: a psychometric evaluation. *Computers in Human Behavior,* 26, 736–740
14. Hilera, J.R. y Martínez, V.J. (1995). *Redes Neuronales Artificiales*. Madrid: RA-MA
15. López, C.P. (2007). *Minería de datos: técnicas y herramientas.* Madrid: Paraninfo.
16. Montaño, J.J., Gervilla, E., Cajal, B. y Palmer, A. (2014). Data mining classification techniques: an application to tobacco consumption in teenagers. *Anales de Psicología, 30*(2), 633-641.
17. Muhlenfeld, H. U. (2005). *Differences between "talking about" and "admitting" sensitive behaviour in anonymous and non-anonymous web-based interviews*. Computers in Human Behavior, 21, 993-1003.
18. Nielsen, T.A., Stenstrom, P. y Levin R. (2006). Nightmare Frequency as a Function of Age, Gender, and September 11. *Findings From an Internet Questionnaire*, *16*(3), 145-158.
19. Palma, C., Palma, W. y Pérez., R. (2009). Data Mining.El arte de

anticipar.10 casosreales. Santiago de Chile: RIL Editores.
20. Pérez, M., Labiano, M. y Brusasca, C. (2010). Escala de deseabilidad social: análisis psicométrico enmuestra argentina. *Evaluar, 10*, 53-67.
21. Rodriguez, J.E. (2008). Minería de datos para la determinación del grado de exclusión social. *Colombia Vínculos*, 5(1), 23 - 31.
22. Romeo, M., Codina, N., Yepes-Baldó, M., Pestana, J.V. y Guardia, J. (2013). Data mining and mall users profile. *Universitas Psychologica, 12*(1), 195-207.
23. Sancesario, L.A. (2012). La minería de datos como herramienta novedosa para el estudio de contrarios en el deporte. *Lecturas: EF y Deportes. Revista digital, 16*(165).
24. Stieger, S. y Reips, U. (2010).What are participants doing while filling in an *on-line* questionnaire: A paradata collection tool and an empirical study. *Computers in HumanBehavior*; 26 (6), 1488-1495.
25. Vieira, L.P., Ortíz, L.I. y Ramírez, S.S. (2009). *Introducción a la Minería de Datos.* Rio de Janeiro: E-papers.
26. Witten, I.H. y Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd. ed.). San Francisco: Morgan Kaufmann.
27. Witten, I.H., Frank, E., Trigg, L., Hall, M., Holmes, G. y Cunningham, S.J. (1999). WEKA: Practical machine learning tools and techniques with Java implementations (pp. 192-196). In N. Kasabov and K. Ko (Ed.), *Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems*. New Zealand: Dunedin.
28. Yubelly-García, S. (2005).*InventarioPsicologico para el Seguimiento de Talentos Deportivos IPSETA*. Bogota: Rendimiento Óptimo.