USO DE LA ENTROPÍA CONDICIONAL COMO MÉTODO PARA LA REDUCCIÓN DE DIMENSIONALIDAD.
UNA APLICACIÓN EN GESTIÓN DE LA INNOVACIÓN

MARÍA DEL C. ROMERO - MARÍA I. CAMIO - MARÍA B. ÁLVAREZ[9]
Facultad de Ciencias Económicas (CEA). UNICEN
romero@econ.unicen.edu.ar - camio@econ.unicen.edu.ar - maria.alvarez@econ.unicen.edu.ar

Fecha Recepción: Julio 2014 - Fecha Aceptación: Abril 2015

RESUMEN

En una gran cantidad de contextos de investigación se presentan situaciones de alta dimensionalidad dada por la gran cantidad de observaciones, de variables, o por mayor cantidad de variables que de observaciones. La clasificación supervisada suele ser una técnica estadística muy utilizada para detectar variables que sean relevantes en la distinción entre grupos. No obstante, la alta dimensionalidad dada por una mayor cantidad de variables que de observaciones, hace que las técnicas convencionales sean inestables y poco confiables. Los métodos filtro resultan una buena estrategia para realizar un ordenamiento de las variables dada su importancia en la distinción entre grupos y reducir dimensionalidad. En este trabajo, se aplica un filtro que trabaja con la entropía condicional como medida de evaluación, en datos referidos al área de gestión de la innovación caracterizados por la presencia de variables cualitativas ordinales. Una vez ordenadas las variables, el investigador especialista en la temática decidirá las acciones a tomar. Por un lado, puede seleccionar las variables más relevantes para analizar su comportamiento con mayor detalle. Por otro, puede "descartar" las menos relevantes para, de esta manera, reducir la dimensionalidad y poder aplicar técnicas convencionales a las variables remanentes.

PALABRAS CLAVE: Entropía **–** Alta dimensionalidad **–** Filtros **–** *Software* y servicios informáticos.

ABSTRACT

High dimensional databases caused by large number of observations, attributes or by a greater number of attributes than observations are present in a large amount of research contexts. Supervised classification is often a statistical technique used to discover the attributes that are relevant to the distinction between groups. However, in contexts of high dimensionality with higher amount of attributes than observations, the application of conventional

[9] Becaria CIC, Beca de Perfeccionamiento, Lugar de trabajo: CEA. UNICEN.

classification techniques generally do not work quite right. The filter methods are a good strategy to order the attributes given its importance in the distinction between groups and to reduce dimensionality. In this paper a filter that works with the conditional entropy as a measure of evaluation is applied to innovation management data characterized by the presence of ordinal attributes. Given a ranking of the attributes, the specialist researcher will decide the actions to take. On one hand, the most relevant attributes can be selected to analyze their behavior in more detail. On the other, the less relevant ones can be discarded in order to reduce the dimensionality and to apply conventional techniques to the remaining attributes.

KEYWORDS: Entropy – High dimensionality – Filters – Software and computer services.

1. INTRODUCCIÓN

Este trabajo se enmarca en el Proyecto de Investigación "Innovación y Modalidades de Gestión" del Centro de Estudios en Administración (CEA) de la Facultad de Ciencias Económicas de la Universidad Nacional del Centro de la Provincia de Buenos Aires (UNICEN), Argentina. El objetivo general de este proyecto se centra en el estudio de la innovación y su medición a nivel empresarial y en la forma en que aquélla se ve influenciada por diferentes modalidades de gestión, en empresas del sector de *software* y servicios informáticos.

En una primera etapa, se trabajó sobre la operacionalización del concepto "nivel de innovación", descifrando sus dimensiones inherentes y seleccionando las variables pertinentes para medir cada una de ellas (Camio, Romero y Álvarez, 2012). De esta manera, se construyó un Modelo de Medición de la Innovación con tres dimensiones centrales: Capacidades, Resultados e Impactos. Cada una de ellas, compuesta a su vez, por una gran cantidad de variables.

En Romero, Camio y Álvarez (2013) se presentó una metodología de construcción de índices para traducir un concepto en una variable "medible" y resumirlo en un único valor. Esto permite "sintetizar" las variables componentes de cada una de las dimensiones propuestas (Capacidades, Resultados e Impactos) en una nueva variable ordinal con las categorías: Alto, Medio y Bajo. De esta manera, a partir de la combinación de las variables que componen, por ejemplo, la dimensión Resultados, puede concluirse que una determinada empresa tiene un nivel de Resultados Alto.

Además de la necesidad de "resumir" en un único valor la información referida a cada una de las dimensiones, se plantea el objetivo de detectar la presencia de asociaciones entre las variables componentes. Por ejemplo, ¿alguna de las variables de Capacidades está asociada con el nivel obtenido para los Resultados?, o lo que es lo mismo, agrupando las empresas según el nivel obtenido para los Resultados (Alto, Medio y Bajo), ¿alguna de las variables

de las Capacidades podría ser importante en la distinción entre estos grupos de Resultados?

En contextos convencionales, esta problemática podría resolverse trabajando con métodos estadísticos clásicos multivariados de clasificación tales como análisis de conglomerados o análisis discriminante. Sin embargo, la alta dimensionalidad con mayor cantidad de variables que de observaciones, dificulta considerablemente la aplicación de estos métodos.

Las bases de datos con las cuales se trabaja en este proyecto se caracterizan por contener una gran cantidad de variables principalmente cualitativas y una baja cantidad de observaciones. Las observaciones las constituyen cada una de las empresas pequeñas del sector del software y servicios informáticos que respondieron el cuestionario[10], siendo la tasa de respuesta muy baja. Es entonces necesario trabajar con métodos de selección de variables para reducir dimensionalidad y detectar aquellas variables que distinguen entre grupos. En Romero, Camio y Álvarez (2014) se presentó el uso de la entropía condicional como medida de evaluación en un método filtro para construir un ordenamiento de las variables, considerando su importancia (relevancia) en la determinación de los grupos.

En este trabajo se presenta una aplicación de dicho método en un conjunto de datos referidos a la gestión de la innovación que se caracterizan por tener variables cualitativas y una alta dimensionalidad dada por una mayor cantidad de variables que de observaciones. Se plantean, además, los alcances y potencialidades del método aplicado.

2. DESARROLLO

2.1. Modelo de medición de la innovación en empresas de *software* y servicios informáticos

La innovación empresarial es concebida como un proceso de cambio, tanto incremental como sustancial, en productos, procesos, organización y mercadotecnia, que impregna a toda la empresa y no se restringe a un área específica e involucra la interacción con el entorno (Cotec, 2006). En el contexto empresarial actual, altamente competitivo y con productos y tecnología con ciclos de vida cortos, es fundamental para la industria del *software* innovar continuamente. Resulta necesario avanzar en iniciativas de medición que evalúen la capacidad de innovación, la producción y el rendimiento (bin Ali y Edison, 2010).

En OCDE y EUROSTAT (2005) se introdujeron algunas modificaciones menores en las definiciones de innovación de producto y de proceso para así reflejar mejor las actividades de innovación en el sector de servicios. En relación a las particularidades de este sector adquiere relevancia la definición de

[10] Para la obtención de los datos se construyó un cuestionario estructurado el cual fue enviado vía web mediante el *software Lime Survey* a los gerentes de las empresas (pertenecientes a una base de datos construida *ad-hoc* a los fines del presente proyecto).

Investigación y Desarrollo (I+D). El Manual de Frascati (OCDE, 2002) resulta la referencia directa para esta definición ya que remiten a él las diferentes legislaciones para el fomento de la industria del *software*.

En la actualidad, la innovación es considerada una capacidad dinámica. Este concepto hace referencia a "la capacidad de la organización de crear, extender o modificar su base de recursos intencionalmente" (Helfalt, Finkelstein, Mitchell, Peteraf, Singh, Teece y Winter, 2007) por la adición de nuevo conocimiento en los nuevos productos, servicios, procesos, tecnologías o métodos de gestión.

En Camio *et al.* (2012) se presentó un Modelo de Medición de la Innovación - FIGURA 1 - que fue construido tomando como punto de partida los componentes del Nivel de Innovación para empresas en general, el análisis de los datos recolectados en estudios previos, el relevamiento bibliográfico de los elementos distintivos de la medición de la innovación para el sector de *software* y servicios informáticos y especialmente la importancia del concepto de I+D para el sector. En su nivel superior el modelo comprende los tres elementos que se identificaron como distintivos en la medición de la innovación: Capacidades, Resultados e Impactos/*Performance*.

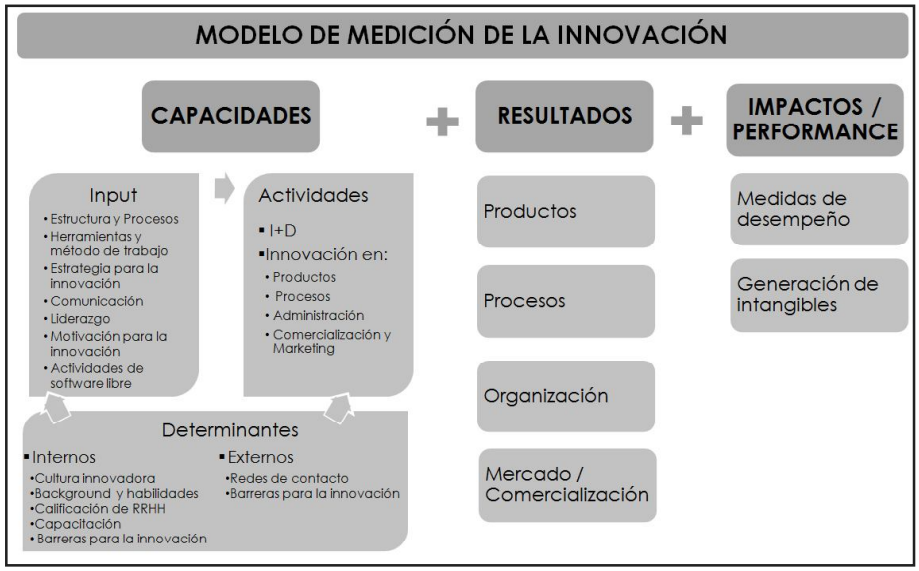

FIGURA 1: Modelo de medición de la innovación.
Fuente: Elaboración propia.

1 Capacidades

1.1 *Input* (Gestión de la Innovación)

1.1.1 Estructura y Procesos: Existencia de áreas o departamentos especiales (por ej.: I+D), Número de personas asignadas a cada área, Nivel identificado para diferentes características de la estructura, Participación y trabajo en redes, Nivel de procesos.

1.1.2 Herramientas y metodología de trabajo: Nivel de herramientas de *software*, Espacios de documentación y de discusión de metodologías

1.1.3 Estrategia para la innovación: Prioridades estratégicas, Grado de orientación al

cliente, Indicadores para medir el nivel de satisfacción de los clientes, Explicitación de objetivos, estrategias, programas o indicadores para la innovación, Decisiones estratégicas para la innovación. Cartera de productos y servicios

1.1.4 Comunicación: Circulación y frecuencia de las comunicaciones, Transmisión de las decisiones tomadas en materia de innovación

1.1.5 Liderazgo: Rasgos de los líderes (con orden de importancia)

1.1.6 Motivación para la generación de innovación: Generación de ideas vinculadas a innovaciones, Sistema de reconocimiento para las personas que presentan ideas innovadores

1.1.7 Actividades de *software* libre: Conocimiento de *software* libre, Motivos de uso de *software* libre

1.2 Actividades de Innovación: Porcentaje de personas, de tiempo por persona y de presupuesto asignado a las actividades de innovación, Nivel de productos y servicios

1.3 Determinantes

1.3.1 Internos

1.3.1.1 Cultura: Valores que defiende y promueve la cultura organizacional

1.3.1.2 *Background* y habilidades: Nivel de relevancia de la experiencia laboral previa, Diversificación de las habilidades de los empleados

1.3.1.3 Calificación de los recursos humanos: Porcentaje de personas según su máximo nivel de educación alcanzado

1.3.1.4. Capacitación: Formas de capacitación

1.3.1.5 Barreras internas: Nivel de importancia de barreras internas y de las barreras económico / financieras para la innovación

1.3.2 Externos

1.3.2.1 Relación con actores externos: Nivel de contacto con actores externos

1.3.2.2 Barreras Externas: Nivel de importancia de barreras externas para la innovación

2 Resultados

2.1 Productos: Introducción de productos nuevos o mejorados, Mercado para el cual son nuevos, Innovación que afecta las características principales

2.2 Procesos: Introducción de procesos nuevos o de mejoras en los procesos existentes

2.3 Organización: Innovaciones organizacionales y su importancia

2.4 Comercialización: Innovaciones en comercialización y su importancia

3 Impactos / Performance

3.1 Medidas de desempeño: Porcentaje de participación de las ventas de productos innovados sobre las ventas totales para los últimos 3 años, Impacto por la introducción de innovaciones en productos, procesos, organización, comercialización

3.2 Generación de intangibles: Solicitud y obtención de patentes, Licencia de tecnología, Uso de licencias de *software* libre / *open source,* Certificación de normas de calidad

El modelo presentado se construyó considerando un enfoque de tipo integral (Edison, bin Ali y Torkar, 2013). Este abordaje considera las capacidades (medidas de *inputs* de innovación, los determinantes y los procesos), los resultados (productos, procesos, de comercialización y organizacionales) y medidas del impacto de la innovación. Además, se tomaron como base otros estudios específicos en empresas pequeñas de los sectores de electrónica y *software* (Romijn y Albaladejo, 2002) que identifican como determinantes de la capacidad de innovación tanto recursos internos (*background* de fundadores o gerentes, habilidades de la fuerza de trabajo, entre otros), como recursos externos (intensidad de las redes de contacto, ventajas de proximidad relativas a las redes de contacto, apoyo institucional).

Otros modelos (Miranda y Figueiredo, 2010) presentan una metodología de medición de las capacidades innovativas en empresas de *software*. Se

distinguen niveles avanzados, de innovación intermedia y de innovación básica. Se identifica en cada uno de ellos el tipo de herramientas de ingeniería de *software* utilizadas, el tipo de productos o servicios y la relación de los procesos con medidas de calidad, certificaciones y la actualización continua de los procesos a partir de nuevos métodos y tecnologías.

2.2. Descubrimiento de conocimiento en bases de datos de alta dimensionalidad

En las últimas décadas las bases de datos crecieron no sólo en cantidad de observaciones sino también de variables. Esto se debe, principalmente, a que las nuevas tecnologías posibilitan la adquisición, almacenamiento y administración de los datos. Estas bases de datos se encuentran en variadas y diversas actividades, tales como bases de datos genómicos, bibliográficos, de documentos de texto, de datos de clientes, de imágenes satelitales y de transacciones comerciales (Fayyad, Piatetsky-Shapiro y Smith, 1996; Cano de Amo, 2004; Ruiz Sánchez, 2006).

Dado que los datos recolectados sobre el entorno representan, en su gran mayoría, la evidencia básica que se usa para construir teorías y modelos sobre el universo en el cual se vive, la obtención de información-conocimiento a partir de los datos obtenidos es de fundamental importancia, ya sea para tomar decisiones, explorar o comprender los fenómenos que los originaron.

La alta dimensionalidad cambia el paradigma de abordaje del proceso analítico y propone nuevos desafíos metodológicos asentados en la teoría estadística y en la capacidad computacional.

El tratamiento estadístico de las grandes bases de datos engloba un conjunto de técnicas que constituyen, junto a otras, la estrategia de análisis conocida como KDD: Descubrimiento de conocimiento en bases de datos (*Knowledge Discovery in Databases* (Fayyad *et al.,* 1996)). En un nivel abstracto, KDD se refiere al desarrollo de métodos y técnicas para hacer que los datos tengan sentido y tiene como objetivo principal la extracción de conocimiento de alto nivel a partir de grandes conjuntos de datos de bajo nivel. Consiste en el proceso completo de descubrir conocimiento útil a partir de los datos. Es un proceso interactivo e iterativo en el cual se usa el conocimiento del dominio y que comprende varios pasos que van desde la obtención de los datos hasta la aplicación del conocimiento adquirido. Lo esencial en este proceso es la minería de datos (*data mining*) que consiste en la aplicación de métodos o algoritmos específicos para el descubrimiento y la extracción de patrones (modelos) a partir de los datos. “Extraer un patrón” se refiere a ajustar un modelo a los datos, encontrar una estructura a partir de los mismos o, en general, realizar cualquier descripción de alto nivel de los datos.

La estadística constituye un componente fundamental en la minería de datos, no sólo porque la extracción de conocimiento se realiza generalmente mediante la aplicación de técnicas estadísticas (por ejemplo, métodos de clasificación supervisada, clasificación no supervisada, medidas de resumen, modelización de dependencias), sino porque además, el conocimiento extraído juega el rol de conocimiento inferido. Las técnicas estadísticas comúnmente

usadas incluyen la clasificación supervisada (clasificación de observaciones pertenecientes a una cantidad conocida de grupos y "detección" de un subconjunto de variables que permita clasificar nuevas observaciones) y no supervisada (clasificación cuando no se conoce previamente la cantidad de grupos).

En alta dimensionalidad, es imprescindible trabajar con algoritmos eficientes ya que la manipulación efectiva de los datos es de fundamental importancia. La gran cantidad de observaciones y de variables genera problemas en términos del incremento del tamaño del espacio de búsqueda para la inducción del modelo de una manera combinatoriamente explosiva, ya que se incrementan las posibilidades de que un algoritmo de minería de datos pueda encontrar patrones espurios que no sean válidos en general (Fayyad *et al*., 1996).

No sólo las bases de datos con gran cantidad de observaciones y de variables generan alta dimensionalidad, sino también las situaciones, cada vez más frecuentes, en las cuales la cantidad de variables supera a la cantidad de observaciones. Esta situación suele estar asociada con mayor ruido[11], lo que provoca que los clasificadores convencionales (análisis de conglomerados, análisis discriminante y técnicas clásicas de inferencia) tiendan a producir resultados inestables por efecto de la parametrización y del sobreajuste (caso en el cual no sólo se modelan los patrones generales en los datos sino también cualquier ruido específico a dicho conjunto de datos). Lo anterior es, sin duda, una consecuencia directa de una gran cantidad de variables que no sólo incluyen información sino también ruido.

2.3. Métodos de selección de variables

En contextos de grandes bases de datos con la particularidad de tener una menor cantidad de observaciones que de variables, es necesario reducir la dimensionalidad. En los últimos años surgieron propuestas de reducción para tratar problemas de clasificación, las cuales proponen la disminución de la cantidad de variables como paso previo a la aplicación de técnicas convencionales. La reducción de la cantidad inicial de variables y la eliminación de variables irrelevantes, mejoraría el rendimiento de las técnicas estadísticas aplicadas posteriormente. En particular, al hablar de clasificación supervisada, la **selección de variables** tiene como objetivo central "descubrir" aquellas variables relevantes en la distinción entre grupos y que puedan ser usadas en la clasificación.

Estos métodos suelen clasificarse considerando la dependencia que tienen con el algoritmo inductivo (clasificador) que finalmente usará el subconjunto seleccionado para realizar la clasificación de observaciones. Puede hablarse de tres enfoques centrales: filtros, *wrappers* (Blum y Langley, 1997; Kohavi y John, 1997) y embebidos.

Los métodos filtro consideran el problema de la selección de variables

[11] Puede definirse como una perturbación que perjudica la transmisión y que impide que la información llegue con claridad.

independientemente del diseño de un clasificador particular. Las variables son seleccionadas considerando las características generales de los datos y utilizando medidas que describen la habilidad de las mismas para distinguir entre los distintos grupos desde una perspectiva genérica. Trabajan, en general, con medidas de distancia, información, dependencia y consistencia. Los *wrappers* (Kohavi y John, 1997), en cambio, realizan la selección y el diseño del clasificador conjuntamente, la utilidad de una variable se evalúa considerando la precisión estimada del clasificador resultante (tasa de error de mala clasificación) que, posteriormente los utilizará para clasificar nuevas observaciones o para predecir el grupo de pertenencia de las observaciones. Los métodos embebidos (*embedded*) - Breiman (1996) - realizan la selección de variables y el diseño del clasificador conjuntamente y combinan las predicciones de múltiples clasificadores para producir uno solo, el cual se denomina clasificador resultante o ensamblador.

2.3.1. Métodos filtro

Debido a que los métodos filtro se ocupan sólo de la selección del subconjunto de variables sin considerar el diseño del clasificador que se utilizará posteriormente, pueden ser usados en conjunción con otros objetivos. Se consideran, esencialmente, métodos de preprocesamiento o de filtrado de datos que ayudan a reducir la dimensionalidad y los efectos no deseados de la sobreparametrización.

El resultado de la aplicación de un algoritmo de filtrado puede ser (Ruiz Sánchez, 2006) un subconjunto de variables o un ordenamiento de todas ellas. En este último caso, para cada variable se genera un puntaje (asociado con la relevancia que tiene cada una de ellas en la distinción entre grupos) que permite ordenarlas por dicha importancia. Resultan una alternativa muy tentadora si el objetivo principal es el ordenamiento de variables considerando su relevancia en la distinción de grupos, lo que permite además, distinguir entre variables relevantes y no relevantes. Estos métodos presentan buen éxito empírico, son muy eficientes computacional y algorítmicamente, son robustos al sobreajuste y muy útiles en contextos con una cantidad de variables superior a la de observaciones.

En el proceso de generación del ordenamiento de variables es necesario contar con una función (medida) de evaluación, la cual resulte una medida de la relevancia que tiene la variable en la distinción entre grupos. Se define a la “relevancia” como la medida de la capacidad que tiene una variable para diferenciar los grupos, mayor relevancia implica mayor capacidad para distinguirlos. Existe una gran cantidad de trabajos en los cuales se clasificaron las diferentes funciones de evaluación utilizadas en la selección de variables. Dash y Liu (1997), las dividieron en cinco categorías: distancia, información, dependencia, consistencia y tasa de error del clasificador (*misclassification*) y Ruiz Sánchez (2006) extendió la clasificación presentada agregando nuevas funciones y reclasificando algunos métodos. Más allá del extenso y variado trabajo realizado respecto a evaluaciones y comparaciones de los métodos de selección de variables, no es posible determinar en todos los casos, la

existencia de un método que resulte más adecuado que el resto.

2.3.2 Entropía condicional como medida de relevancia

La selección de los filtros a considerar se ve fuertemente influenciada por las características del área de aplicación, en este caso, datos de tipo cualitativo. Se trabajó entonces con filtros que generen un ordenamiento de variables utilizando a la entropía condicional como medida de evaluación (Romero *et al.*, 2014) de la importancia (capacidad discriminatoria) que tiene cada variable en la distinción entre grupos.

La entropía (presentada por Shannon en 1948) representa una medida de la incertidumbre media de una variable aleatoria:

$$H(X) = -\sum_{i} p(x_i).log_2\, p(x_i)$$

donde X representa la variable aleatoria, $p(x_i)$ la probabilidad de ocurrencia de cada uno de sus valores y el logaritmo en base 2 considera que la información se representará mediante código binario (*bits*).

El mínimo valor es 0, indicando que no hay incertidumbre, y el máximo se obtiene en casos en los cuales la variable aleatoria pueda tomar diferentes valores equiprobables. Si por ejemplo, X es una variable binaria, esto es, toma sólo dos posibles valores y estos tienen una probabilidad 0,5: $H(X) = -0{,}5.log_2(0{,}5) - 0{,}5.log_2(0{,}5) = -0{,}5.(-1) - 0{,}5.(-1) = 1$

La entropía condicional referencia la entropía que tiene una determinada variable Y conociendo la información que aporta otra variable X:

$$H(Y/X) = -\sum_{x} p(x) \sum_{y} p(y/x)\,.log_2 p(y/x)$$

El mínimo valor que puede tomar la entropía condicional es 0, lo que indica que, conociendo la variable X, no se tiene incertidumbre sobre el valor que tomará la variable Y. Suponiendo que la variable Y representa grupos (dados por los valores que puede tomar), en caso de entropía condicional 0, dado el valor de X no se tiene incertidumbre sobre el grupo (dado por el valor de Y) al cual pertenecerá la observación. Esto puede interpretarse como que la variable X está altamente asociada al grupo y que, dado el valor de la misma, puede determinarse con certeza el grupo al que pertenecerá.

El máximo valor está dado por la cantidad de valores que pueden tomar las variables. Cuanto más cerca del valor máximo se encuentre la entropía condicional, menos distingue la variable entre grupos. Altos valores de la entropía condicional indicarían que la variable no toma valores diferenciales entre grupos. Un caso extremo está dado por una constante en la cual, claramente, no toma valores diferentes dependiendo del grupo.

Las situaciones planteadas pueden ejemplificarse considerando dos variables binarias.

$$H(Y/X) =$$
$$-p(X=0).(p(Y=0/X=0).log_2 p(Y=0/X=0) + p(Y=1/X=0).log_2 p(Y=1/X=0))$$
$$-p(X=1).(p(Y=0/X=1).log_2 p(Y=0/X=1) + p(Y=1/X=1).log_2 p(Y=1/X=1))$$

Situación 1			Situación 2		
X	Y		X	Y	$H(Y/X)$
0	0	$H(Y/X) =$	0	0	$= -0{,}5\,.\,(0{,}5\,.\,log_2 0, +0{,}5.log_2 0{,}5)$
0	0	$-0{,}5\,.\,(1\,.\,log_2 1 + 0.log_2 0)$	0	1	$-\,0{,}5\,.\,(0{,}5.log_2 0{,}5 + 0{,}5\,.\,log_2 0{,}5)$
1	1	$-\,0{,}5\,.\,(0.log_2 0 + 1\,.\,log_2 1) = 0$	1	0	$=-0{,}5\,.\big(0{,}5\,.\,(-1) + 0{,}5.(-1)\big) -$
1	1		1	1	$0{,}5\,.\big(0{,}5.(-1) + 0{,}5\,.\,(-1)\big) = 1$

En la Situación 1 se obtiene el mínimo valor de entropía condicional ya que, como puede observarse, al conocer el valor de la variable X no se tiene incertidumbre sobre el valor de la variable Y. En cambio, en la Situación 2 se obtiene el mayor valor. Conocer la variable X no modifica la incertidumbre que se tiene sobre la variable Y.

El algoritmo genera entonces, un ordenamiento de variables según la importancia que tengan en la determinación de los grupos (basado en la entropía condicional).

2.4. Especificación del problema

La tarea realizada a nivel teórico para "desagregar" el concepto de nivel de innovación en sus dimensiones componentes y encontrar las variables contextuales adecuadas (descripto en el punto 2.1), fue acompañada por un proceso en el cual a partir de los valores de dichas variables se pudiera dar respuesta a la variable de nivel superior. En Romero *et al.* (2013) se presentó una metodología de construcción de índices que permite dar respuesta a cada una de las dimensiones (Capacidades, Resultados e Impactos) mediante una variable cualitativa ordinal que puede tomar los valores: Alto, Medio o Bajo.

Este trabajo contribuye a dar respuesta a la problemática de descubrir la existencia o no de asociación entre las variables componentes de las Capacidades y el nivel obtenido para los Resultados. Corresponde a un tema de clasificación supervisada, en la cual, dados diferentes grupos, se pretenden encontrar las variables que se comportan diferencialmente: dadas las empresas clasificadas según su nivel de Resultados (Alto, Medio o Bajo), el interés reside en detectar las variables de las Capacidades que tienen valores diferentes para cada grupo. Las variables de Capacidades son:

Nombre	Descripción y Preguntas del Anexo
V1_DEPTOS	Existencia de áreas o departamentos especiales y cantidad de personas asignadas a cada uno de ellos (Preguntas 1 y 2)
V2_DECIS	Nivel (Bajo / Alto) de centralización de las decisiones y el uso de sistemas de planificación y control (Pregunta 3)
V3_RED	Trabajo en red (dentro y con otras empresas) (Pregunta 4)
V4_PROCESOS	Nivel de los procesos que posee la empresa (Pregunta 5)
V5_HERRAM	Nivel de herramientas de *software* utilizadas (Pregunta 6)
V6_DOCUM_METOD	Espacios de documentación y de discusión de metodologías (Pregunta 7)
V7_PRIORID_ESTRAT	Prioridades estratégicas de la empresa (Pregunta 8)
V8_ORIENTACION_ CLIENTE	Orientación al cliente y existencia de indicadores que permitan medir el nivel de satisfacción de los mismos (Preguntas 9 y 10)

V9_INDIC_INNOV	Explicitación de objetivos, estrategias, programas o indicadores para la innovación (Pregunta 11)
V10_DECIS_ESTRAT	Forma en la que se toman las decisiones estratégicas para la innovación en su empresa (Pregunta 12)
V11_CARTERA_ PRODYSERV	Actividades que realiza la empresa (desarrollo de productos, prestación de servicios, etc.) (Pregunta 13)
V12_COMUNIC_CIRC	Circulación de la comunicación en la empresa (Pregunta 14)
V13_COMUNIC_FREC	Frecuencia de comunicaciones dentro y entre áreas (Pregunta 15)
V14_COMUNIC_ TRANSMISION	Comunicación. Transmisión de las decisiones tomadas en materia de innovación (Pregunta 16)
V15_RASGOS_ LIDERES	Rasgos de los líderes que resultan más importantes para la empresa (Pregunta 17)
V16_MOTIVACION	Motivación para la generación de ideas vinculadas a innovaciones y existencia de un sistema de reconocimiento (Preguntas 18 y 19)
V17_SOFT_LIBRE	Actividades de *software* libre. Conocimiento del concepto, utilización y motivos de uso del mismo (Preguntas 20 y 21)
V18_ACTIV_RRHH	Cantidad de personas y porcentaje del tiempo que le dedican a actividades de innovación (Preguntas 22 y 23)
V19_ACTIV_PRESUP	Presupuesto asignado a las actividades de innovación (Pregunta 24)
V20_NIVEL_ PRODYSERV	Nivel de productos y servicios que más representa la situación de la empresa (Pregunta 25)
V21_CULTURA	Valores que defiende y promueve la cultura organizacional (Preguntas 26 y 27)
V22_BACK_HABIL	*Background* y habilidades. Relevancia de la experiencia laboral previa (de los gerentes) y diversificación de las habilidades (técnicas - analíticas) de los empleados (Preguntas 28 y 29)
V23_BARR_INT	Nivel de importancia de barreras internas y de barreras económico / financieras para la innovación (Preguntas 30 y 31)
V24_CALIF_RRHH	Nivel de educación de los recursos humanos (Pregunta 32)
V25_CAPACITACION	Formas de capacitación utilizadas (Pregunta 33)
V26_ACTORES_EXT	Relación y nivel de contacto con actores externos (Pregunta 34)
V27_BARR_EXT	Identificación de barreras externas para la innovación (Pregunta 35)

Tanto las variables componentes de Capacidades como la variable resumen para los Resultados son ordinales con las categorías Bajo, Medio y Alto. En el caso de las variables de Capacidades, la asignación del valor fue realizada por especialistas en la temática a partir de las preguntas enunciadas en el ANEXO y en función de su contribución a la innovación. La variable resumen para los Resultados fue construida a partir de la metodología planteada en Romero *et al.* (2013) considerando innovaciones realizadas en: Productos: (Anexo – Preguntas 36, 37 y 38), Procesos: (Anexo – Preguntas 39 y 40), Organización (Anexo – Preguntas 41 y 42) y Comercialización (Anexo – Preguntas 43 y 44).

En el contexto de aplicación, se tienen 20 observaciones que corresponden a empresas pequeñas y 28 variables cualitativas ordinales (una de ellas corresponde al grupo de los Resultados y las 27 restantes a los aspectos de las Capacidades).

La matriz de datos puede representarse de la siguiente manera:

		Capacidades (innovación)					Resultados Grupo
		Cap_1	Cap_2	Cap_3	…	Cap_p	
Empresa	1	y_{11}	y_{12}	y_{13}	…	y_{1p}	Bajo
	2	…	…	…	…	…	…
	3	y_{31}	y_{32}	y_{33}	…	y_{3p}	Bajo
	4	y_{41}	y_{42}	y_{43}	…	y_{4p}	Medio
	…	…	…	…	…	…	…
	…	…	…	…	…	…	Alto
	n	y_{n1}	y_{n2}	y_{n3}	…	y_{np}	Alto

p: cantidad de variables (de las Capacidades), p = 27.
n: cantidad de observaciones (empresas), n = 20.
y_{ij}: valor de la variable j-ésima en la empresa i-ésima; i = 1, …, n; j = 1, …, p
Grupo: variable de clasificación para las empresas según los Resultados de la innovación.

Dada la particularidad de una cantidad de observaciones menor que de variables (n<p), el problema puede enunciarse como un problema de clasificación supervisada en alta dimensionalidad con una cantidad de variables mayor que de observaciones en el cual cobra sentido el filtrado de datos para la reducción de dimensionalidad.

Si la entropía (incertidumbre) de los Resultados de la innovación disminuye por el conocimiento de alguna variable de las Capacidades de la innovación, dicho aspecto es relevante para la distinción entre los grupos de Resultados. La formulación particular de la entropía condicional para este contexto donde tanto la variable X como Y son cualitativas ordinales con los valores: Bajo, Medio y Alto sería:

$$H(Y/X) = -\sum_{x} p(x) \sum_{y} p(y/x) \,.\, log_2 p(y/x)$$

H (Y / X) = - p (X=Bajo) . (p (Y=Bajo / X=Bajo). $\log_2$ (Y=Bajo / X=Bajo) +
p (Y=Medio / X=Bajo). $\log_2$ (Y=Medio / X=Bajo) +
p (Y=Alto / X=Bajo). $\log_2$ (Y=Alto / X=Bajo))
- p (X=Medio) . (p (Y=Bajo / X=Medio). $\log_2$ (Y=Bajo / X=Medio) +
p (Y=Medio / X=Medio). $\log_2$ (Y=Medio / X=Medio) +
p (Y=Alto / X=Medio). $\log_2$ (Y=Alto / X=Medio))
- p (X=Alto) . (p (Y=Bajo / X=Alto). $\log_2$ (Y=Bajo / X=Alto) +
p (Y=Medio / X=Alto). $\log_2$ (Y=Medio / X=Alto) +
p (Y=Alto / X=Alto). $\log_2$ (Y=Alto / X=Alto))

donde X representa a cada una de las variables de Capacidades e Y determina los grupos de Resultados.

Para el cálculo, se desarrolló un programa en R[12] y se utilizó el *package entropy* de R (Hausser y Strimmer, 2013). Se trabajó con la función “entropy” la cual estima la entropía Shannon de una variable aleatoria a partir de frecuencias observadas. La salida del programa resulta una lista ordenada de forma decreciente con cada una de las variables y su relevancia

[12] **R** es un lenguaje y entorno de programación para análisis estadístico desarrollado inicialmente por Robert Gentlemen y Ross Ihaka del Departamento de Estadística de la Univ. Auckland (1993). Su desarrollo actual es responsabilidad del *R Development Core Team. http://www.r-project.org/*

correspondiente en la distinción entre grupos (menor entropía condicional implica mayor relevancia en la distinción entre grupos, por lo tanto, se esperaría que las variables que estén en el tope de la lista, tengan valores de entropía prácticamente nulos).

3. RESULTADOS

En general, se encuentra que para ninguna de las variables se obtiene una entropía condicional cercana al mínimo – valor igual a 0 – (TABLA 1). Esto indicaría, en principio, que ninguna de las variables consideradas en Capacidades es "determinante" en los Resultados.

TABLA 1: Ordenamiento de las variables considerando la entropía condicional

Orden	Variable	Entr. cond.	Orden	Variable	Entr. cond.
1	V24_CALIF_RRHH	1,0477	15	V25_CAPACITACION	1,3711
2	V2_DECIS	1,0639	16	V6_DOCUM_METOD	1,3755
3	V17_SOFT_LIBRE	1,2297	17	V7_PRIORID_ESTRAT	1,3868
4	V26_ACTORES_EXT	1,2419	18	V10_DECIS_ESTRAT	1,3908
5	V16_MOTIVACION	1,2876	19	V13_COMUNIC_FREC	1,3924
6	V4_PROCESOS	1,3069	20	V5_HERRAM	1,4082
7	V27_BARR_EXT	1,3132	21	V22_BACK_HABIL	1,4087
8	V3_RED	1,3142	22	V12_COMUNIC_CIRC	1,4333
9	V15_RASGOS_LIDERES	1,3337	23	V11_CARTERA_PRODYSERV	1,4348
10	V20_NIVEL_PRODYSERV	1,3427	24	V8_ORIENTACION_CLIENTE	1,4460
11	V18_ACTIV_RRHH	1,3479	25	V14_COMUNIC_TRANSMISION	1,4460
12	V19_ACTIV_PRESUP	1,3498	26	V9_INDIC_INNOV	1,5000
13	V1_DEPTOS	1,3629	27	V21_CULTURA	1,5000
14	V23_BARR_INT	1,3692			

Las variables con mayor relevancia en la distinción entre empresas con distintos niveles de Resultados son: la calificación de los recursos humanos en cuanto a su nivel de educación (V24_CALIF_RRHH) en primer lugar, el nivel de centralización de las decisiones y el uso de sistemas de planificación y control (V2_DECIS) en segundo lugar. Les siguen las actividades de *software* libre (V17_SOFT_LIBRE), la relación y nivel de contacto con actores externos (V26_ACTORES_EXT) y la motivación para la generación de innovación (V16_MOTIVACION). Los valores que defiende y promueve la cultura organizacional (V21_CULTURA) y la explicitación de objetivos, estrategias, programas o indicadores para la innovación (V9_INDIC_INNOV) resultaron ser

las variables menos relevantes en la distinción entre los diferentes grupos de Resultados.

El ordenamiento de variables puede utilizarse con dos fines: detectar las variables más relevantes y/o detectar las menos relevantes. El especialista en la temática es el encargado de decidir qué uso darle a dicho ordenamiento. Puede detectar las variables más relevantes (se ubican entre los primeros lugares de la lista), descartar aquéllas que sean irrelevantes en la distinción entre grupos (las que ocupan los últimos lugares) o ambas cosas. Una variable sería irrelevante si la probabilidad de asignarle un grupo determinado a una observación no se ve modificada por la presencia o ausencia de tal variable.

4. CONCLUSIONES

Son cada vez más frecuentes los contextos de alta dimensionalidad determinados no sólo por grandes volúmenes de datos sino también por mayor cantidad de variables que de observaciones. En las ciencias sociales y en áreas como la administración y la gestión organizacional, estas situaciones resultan habituales, agregándose además, la presencia de variables principalmente cualitativas.

En este trabajo se presenta una aplicación concreta en el área de la gestión de la innovación donde se trabaja con un método filtro que usa la entropía condicional como medida de evaluación para detectar las variables relevantes (irrelevantes) en la distinción entre grupos. Esta medida resulta ser muy útil ya que puede aplicarse a variables pertenecientes a cualquier escala de medición y no tiene restricciones sobre la cantidad de grupos en los cuales se clasifica a las observaciones.

Dados tres grupos de empresas pequeñas de *software* y servicios informáticos determinados por el nivel de Resultados asociados a la innovación (Alto, Medio y Bajo), se aplica la medida de entropía condicional como método filtro para detectar qué variables de las Capacidades se comportan diferencialmente entre los diferentes grupos de Resultados. Se observa que la calificación de los recursos humanos y que el nivel de centralización de las decisiones y el uso de sistemas de planificación y control son los aspectos de las Capacidades que tienen mayor relevancia en la distinción de empresas con diferentes niveles de Resultados. Esto apoya los resultados de otros estudios en la temática y subraya la importancia determinante de estos aspectos en la gestión de la innovación (tal como se expone en Camio *et al.*, 2012).

El método propuesto permite ordenar variables considerando la importancia que tienen en la distinción entre grupos. A partir de dicho ordenamiento, el investigador puede seleccionar las variables más relevantes (para analizar su comportamiento con mayor detalle), y “descartar” las menos relevantes para reducir la dimensionalidad y poder aplicar técnicas convencionales a las variables restantes. En ambos casos es importante considerar la cercanía o lejanía de las relevancias respecto de los valores mínimo y máximo de la entropía, que una variable esté al tope del ordenamiento no significa que sea un “buen” clasificador, sino que es mejor clasificador que el

resto de las variables de la lista. Debe entonces determinarse, además, la cantidad de variables a considerar, esto es, "dónde" cortar la lista fijando el método que resulte más conveniente en el contexto en el cual se esté trabajando.

La identificación de las variables clave permite hacer más eficientes estudios futuros, salvando los problemas que trae consigo la alta dimensionalidad, especialmente en casos con mayor cantidad de variables que de observaciones.

Es importante tener en cuenta que el método propuesto es de tipo descriptivo y cuyo objetivo es esclarecer, en la medida de lo posible, situaciones de clasificación supervisada en alta dimensionalidad en las cuales las técnicas convencionales de agrupamiento no son pertinentes de aplicar. Resulta una primera aproximación para "ordenar" variables considerando su importancia en la distinción entre grupos en contextos de alta dimensionalidad con variables cualitativas. El próximo paso consiste en el desarrollo de un algoritmo que genere el subconjunto de variables que resulten diferenciales.

5. ANEXO

Pregunta 1. Áreas o departamentos especiales posee la firma: a) I+D, b) Gestión de la Calidad /Certificación (Procesos), c) Gestión de la Calidad /Certificación (Productos), d) Otras

Pregunta 2. Número de personas asignadas a cada área: a) I+D, b) Gestión de la Calidad /Certificación (Procesos), c) Gestión de la Calidad /Certificación (Productos), d) Otras

Pregunta 3. Nivel identificado (Alto/Bajo) para las siguientes características de la estructura: a) Nivel de centralización de las decisiones, b) Uso de sistemas de planificación y control

Pregunta 4. Indique Sí o No a las siguientes afirmaciones: a) Su empresa participa en redes de colaboración con otras empresas, b) Dentro de su empresa los integrantes trabajan en red

Pregunta 5. Nivel de los procesos que más representa la situación de la empresa: a) Procesos operacionales no formalizados, b) Estandarización básica de los procesos, c) Normalización del proceso de ingeniería de *software*, d) Capacitación en metodologías de gestión de procesos, e) Gestión estratégica de la calidad, obtención de certificaciones, f) Procesos controlados con medidas de calidad, g) Mejora continua de los procesos

Pregunta 6. Nivel de herramientas de *software* que más representa la situación de la empresa: a) Copias de seguridad, reutilización de código fuente, prácticas de ingeniería de *software ad hoc*, b) Estandarización y documentación de las prácticas de ingeniería de *software*, c) Normalización de las prácticas de pruebas y la inspección del código, creación y control de versiones automatizadas, creación de componentes de la biblioteca, d) Herramientas de integración continua, e) Integración con herramientas de otras áreas específicas, herramientas de generación de código y manejo de equipos geográficamente dispersos

Pregunta 7. Documentación y metodologías (Sí/No): a) Documentación: Su empresa tiene como política general documentar las diferentes alternativas consideradas acerca de las herramientas y tecnologías utilizadas en la implementación de los productos, b) Metodologías: Periódicamente se producen en la empresa espacios de discusión acerca de las metodologías usadas en el ciclo de desarrollo de un producto de *software* (análisis, diseño y validación)

Pregunta 8. 3 prioridades estratégicas de su empresa: a) Incrementar rentabilidad (utilidad / capital invertido), b) Reducción de costos / precios competitivos, c) Incrementar el margen por diferenciación de producto / servicio, d) Generación de nuevos productos / servicios, e) Foco en principales competencias, g) Desarrollar nuevos mercados geográficos, h) Desarrollar nuevas áreas de negocio, i) Posicionarse como empresa líder en innovación

Pregunta 9. 3 afirmaciones que mejor representan la relación con sus clientes: a) No se cuenta con la información suficiente respecto de las necesidades o problemas actuales de los clientes, b) Existen dificultades para mantener una comunicación fluida con los clientes, c) Dedica tiempo a la atención de cada cliente para satisfacer las demandas que se le plantean, d) Comprende que el cliente es fundamental para su empresa y se actúa en consecuencia, e) Es una prioridad estratégica el mantenimiento de las relaciones a largo plazo con el cliente y se planifican las acciones de los equipos de trabajos en base a esto, f) Es un referente interno y externo cuando se busca aportar soluciones o satisfacer las demandas

Pregunta 10. ¿Cuenta con indicadores para medir el nivel de satisfacción de los clientes? Sí / No

Pregunta 11. Indique si la empresa tiene explicitados objetivos, estrategias, programas o indicadores para la innovación: a) Objetivos de innovación, b) Estrategias de innovación, c) Programa de innovación, d) Indicadores para medir el nivel de innovación, e) Ninguno, f) Otro

Pregunta 12. Afirmación que mejor represente la forma en la que se toman las decisiones estratégicas para la innovación: a) Las toman los socios con la información con que cuenten en ese momento, b) Las toman los socios luego de consultar a miembros del equipo de trabajo, c) Los directivos comparten el problema con el grupo de trabajo y luego ellos toman la decisión

Pregunta 13. Actividades que realiza la empresa (porcentaje de realización de cada una): a) Desarrollo de productos, b) Prestación de servicios, c) *Outsourcing* (tercerización, subcontratación), d) Búsqueda de soluciones, e) Otras

Pregunta 14. Circulación de la comunicación: a) Los canales de comunicación son mayormente ascendentes y en menor medida descendentes con comunicaciones horizontales, b) Los canales de comunicación en su empresa son mayormente descendentes y en menor medida ascendentes con algunas comunicaciones horizontales, c) Los canales de comunicación en su empresa son descendentes, con escasas relaciones horizontales

Pregunta 15. Frecuencia de las comunicaciones (muy frecuentemente, frecuentemente, poco frecuentemente, no se dan): a) Dentro del mismo área, b) Entre áreas diferentes

Pregunta 16. Los resultados de las decisiones tomadas en materia de innovación: a) Se transmiten en forma completa a todos los niveles, b) Se transmiten en forma completa a los niveles intermedios y en forma reducida a los niveles inferiores, c) Se trasmiten en forma parcial a todos los demás niveles, d) No se transmite la información a los demás niveles

Pregunta 17. 5 rasgos de los líderes más importantes: a) Creativo, b) Realista, c) Explorador, d) Conservador, e) Motivador, f) Fuertemente enfocado en las tareas, g) Toma riesgos, h) Promotor de la participación, i) "Jefe", j) Con ideas insuperables

Pregunta 18. Las ideas vinculadas a mejoras o innovaciones a realizar en la empresa: a) Son generadas en su gran mayoría por los directivos, b) Se alienta y se analizan las propuestas innovadoras de cualquier miembro del personal, c) Se escuchan y analizan las propuestas innovadoras de cualquier miembro del personal

Pregunta 19. ¿Se aplica un sistema de reconocimiento a aquellas personas que presenten ideas innovadoras o de mejora?: a) Siempre, b) Algunas veces, c) No se realiza

Pregunta 20. *Software* libre (Sí / No): a) ¿Conoce el concepto?, b) ¿Su empresa lo utiliza?

Pregunta 21. Motivos por los que utiliza *software* libre: a) Para evitar el pago de licencias, b) Porque a largo plazo considera que va a generar mayores beneficios, c) Considera que no es correcto utilizar *software* ilegal, d) Entiende que puede ser un modelo de negocios que potencia la innovación tecnológica, e) Otro

Pregunta 22. Cantidad de personas asignadas a las actividades innovativas: a) Proyectos de I+D; b) Innovaciones en comercialización; c) Innovaciones en la gestión organizacional; d) Capacitación como soporte para la innovación

Pregunta 23. Porcentaje de tiempo de trabajo por persona (si se tiene más de una persona por actividad y presentan tiempos de dedicación diferentes, indicar el porcentaje promedio de tiempo por persona) - de las personas asignadas a las actividades innovativas - a) Proyectos de I+D; b) Innovaciones en comercialización; c) Innovaciones en la gestión organizacional; d) Capacitación como soporte para la innovación

Pregunta 24. Porcentaje fijado para cada actividad (del presupuesto total asignado): a) Proyectos de I+D, b) Innovaciones en comercialización, c) Innovación en la gestión organizacional, d) Adquisición de *hardware* que potencie la innovación, e) Adquisición de tecnología no incorporada al capital físico, f) Capacitación como soporte para la innovación

Pregunta 25. Nivel de productos y servicios que más representa la situación de la empresa: a) Replicación de especificaciones técnicas y funcionales determinadas por los clientes, pequeñas soluciones en partes de proyectos, mantenimiento de soluciones existentes, b) Realiza las especificaciones funcionales del cliente (especificación técnica); hace reingeniería de productos existentes en el mercado, c) Realiza el análisis, definición y especificación de los requisitos para el cliente, la implantación de *software* corporativo y la reingeniería de productos agregando funcionalidades, d) Soluciones desarrolladas con conocimiento de los negocios del cliente, configuración y personalización de *software* empresarial, evolución continua de los productos y nuevos productos utilizando conocimientos adquiridos en productos anteriores, e) Soluciones de alto valor agregado y alta complejidad a partir de conocimientos técnicos y de negocio, soluciones completas con integración y personalización de *software* corporativo, uso de tecnologías avanzadas, f) I+D con la aplicación de tecnología de punta, por ejemplo, *grid computing*.

Pregunta 26. 5 valores que defiende y promueve la cultura organizacional: a) Libertad y apoyo para explotar las habilidades personales, b) Preferencia del trabajo en equipo al individual, c) Se prefieren empleados que acaten las normas y órdenes de sus superiores, d) Compromiso de la dirección con la innovación, e) Se prefieren empleados dispuestos a tolerar la incertidumbre y a asumir riesgos, f) Todo el personal conoce claramente los objetivos de la organización, g) Existen controles estrictos en las tareas que desarrollan los empleados, h) Preferencia del trabajo individual al trabajo en equipo, i) Sólo los directivos conocen claramente los objetivos de la organización, j) Se prefieren empleados conservadores y la empresa elige proyectos de bajo riesgo, k) Se entiende el error como una posibilidad de aprendizaje

Pregunta 27. Jerarquización de los 5 valores que defiende y promueve la cultura organizacional (seleccionados en la pregunta anterior), siendo 1 el más relevante y 5 el menos relevante.

Pregunta 28. Nivel de relevancia de la experiencia laboral previa (Muy relevante / Relevante / Poco relevante): a) Pequeñas y medianas empresas, b) Grandes empresas, c) Instituciones científicas, d) Organizaciones / Entidades públicas, e) Otra

Pregunta 29. Teniendo en cuenta las habilidades técnico–analíticas de los empleados, evalúe el nivel de los siguientes aspectos: (Alto / Medio / Bajo): a) Solidez, b) Profundidad, c) Variedad

Pregunta 30. Nivel de importancia de las barreras internas (Alto/Medio/Bajo/No es una barrera): a) Falta de personal cualificado, b) Falta de información del mercado, c) Falta de información tecnológica, d) La presión del día a día, que absorbe todo el tiempo de los directivos, e) Una excesiva orientación interna sin establecer ni gestionar relaciones con agentes externos en los proyectos de innovación (con clientes, centros tecnológicos, etc.), f) Una cultura que no favorece la creatividad y no permite los errores, g) La definición de las estrategias sin contar con las ideas y el

talento del resto de la organización, h) Fallos a lo largo de los proyectos de innovación (por ejemplo: no involucrar a los agentes adecuados, tiempos de desarrollo demasiado largos, etc.), i) La falta de indicadores que midan los avances y resultados de la innovación

Pregunta 31. Nivel de importancia de las barreras económico/financieras (Alto / Medio / Bajo / No es una barrera): a) Costos para la innovación muy altos, b) Dificultad de financiamiento, c) Período de recupero muy extenso, d) Alto riesgo de la inversión, e) La presión por los resultados a corto plazo que dificulta realizar inversiones en proyectos de innovación, f) Una excesiva focalización hacia la reducción de costos

Pregunta 32. Número de personas según su máximo nivel de educación alcanzado: a) Hasta secundaria incompleta, b) Hasta secundaria completa, c) Hasta universitaria incompleta, d) Hasta universitaria completa, e) Hasta postgrado incompleto, f) Hasta postgrado completo

Pregunta 33. Indique si la empresa usa alguna de las siguientes formas de capacitación y la frecuencia promedio por persona. a) Formación *in company* (a medida), b) Participación en capacitaciones abiertas, c) Formación por integrantes de la propia empresa, d) Otra

Pregunta 34. Nivel de contacto con los actores externos. (Alto / Bajo / Medio / No hay relación): a) Clientes, b) Proveedores, c) Competidores, d) Universidad, centro de investigación o desarrollo tecnológico, e) Empresas de industrias relacionadas, f) Empresas de consultoría, g) Gobierno

Pregunta 35. Nivel de importancia de las siguientes barreras externas (Alto / Medio / Bajo / No es barrera): a) Dificultad de cooperación con instituciones de tecnología, b) Normativa legal, c) Falta de información tecnológica, d) Rechazo de productos por parte del mercado, e) Facilidad de copia

Pregunta 36. ¿Introdujo al mercado productos nuevos o significativamente mejorados en los últimos 3 años? Sí / No

Pregunta 37. Productos nuevos para el mercado: a) Internacional, b) Nacional, c) Su empresa

Pregunta 38. ¿La innovación afecta las características principales del producto en la mayoría de los casos?: Sí / No

Pregunta 39. ¿Introdujo la empresa procesos nuevos en los últimos 3 años? Sí / No

Pregunta 40. ¿Introdujo la empresa mejoras significativas en los procesos existentes en los últimos 3 años? Sí / No

Pregunta 41. ¿Realizó la empresa innovaciones organizacionales en los últimos 3 años? Sí / No

Pregunta 42. Innovaciones organizacionales realizadas en los últimos 3 años: a) Desverticalización de las relaciones, b) Reducción de áreas funcionales de la organización, c) Mayor participación en la toma de decisiones, d) Mayor nivel de delegación de tareas, e) Interacción entre departamentos, f) Introducción de nuevos métodos de gestión del negocio, g) Introducción de nuevos métodos de gestión de RRHH, h) Introducción de nuevos métodos de distribución de responsabilidades, i) Mejoras en la gestión de procesos por la introducción de estándares, descripción de procedimientos, generación de indicadores cuali-cuantitativos, j) Desarrollo de relaciones externas, k) Otras

Pregunta 43. ¿Realizó innovaciones en comercialización en los últimos 3 años?: Sí / No

Pregunta 44. Innovaciones en comercialización realizadas en los últimos 3 años: a) Introducción de productos/servicios en nuevos mercados, b) Apertura de mercados no existentes, c) Desarrollo de nuevos canales de distribución/comercialización, d) Desarrollo de nuevas estrategias de distribución/comercialización, e) Desarrollo de nuevas estrategias de fijación de precios, f) Desarrollo de nuevas estrategias de comunicación/publicidad, g) Otras

6. REFERENCIAS BIBLIOGRÁFICAS

BIN ALI, N.; EDISON, H. (2010): "TOWARDS INNOVATION MEASUREMENT IN SOFTWARE INDUSTRY". School of Computing at Blekinge Institute of Technology.

BLUM, A.; LANGLEY, P. (1997): "SELECTION OF RELEVANT FEATURES AND EXAMPLES IN MACHINE LEARNING". *Artificial Intelligence*, 97(1-2): 245-271.

BREIMAN, L. (1996): "BAGGING PREDICTORS*". Machine Learning*, 24(2): 123-140.

CAMIO, M. I., ROMERO, M. del C.; ÁLVAREZ, M. B. (2012): "MEDICIÓN DEL NIVEL DE INNOVACIÓN EN EMPRESAS DEL SECTOR DE SOFTWARE". *XVII Reunión Anual de la RedPyMEs Mercosur.* Escola Politécnica da Universidade de Sao Paulo - São Paulo, SP – Brasil.

CANO DE AMO, J. R. (2004): "REDUCCIÓN DE DATOS BASADA EN SELECCIÓN EVOLUTIVA DE INSTANCIAS PARA MINERÍA DE DATOS". Tesis doctoral. Departamento de Cs. de la Computación e Inteligencia Artificial. Univ. de Granada. España.

COTEC (2006): "FUNDACIÓN PARA LA INNOVACIÓN TECNOLÓGICA. MARCO DE REFERENCIA DE INNOVACIÓN". Madrid: Editorial Club de excelencia en Gestión.

DASH, M.; LIU, H. (1997): "FEATURE SELECTION FOR CLASSIFICATION". *Intelligent Data Analysis*, 1: 131–156.

EDISON, H.; BIN ALI, N.; TORKAR, R. (2013): "TOWARDS INNOVATION MEASUREMENT IN THE SOFTWARE INDUSTRY". *Journal of Systems and Software,* 86(5): 1390-1407.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMITH, P. (1996): "FROM DATA MINING TO KNOWLEDGE DISCOVERY IN DATABASES". *Artificial Intelligence Magazine*, 17(3): 37-54.

HAUSER, J.; STRIMMER, K. (2013): "PACKAGE ENTROPY: ESTIMATION OF ENTROPY, MUTUAL INFORMATION AND RELATED QUANTITIES". R package, versión 1.2.0. Versión obtenida en abril de 2013. http://strimmerlab.org/software/entropy

HELFAT, E.; FINKELSTEIN, S.; MITCHELL, W.; PETERAF, M.; SINGH, H.; TEECE, D.; WINTER, S. (2007): "DYNAMIC CAPABILITIES: UNDERSTANDING STRATEGIC CHANGE IN ORGANIZATIONS". Malden, London. Blackwell publishing.

KOHAVI, R.; JOHN, G. H. (1997): "WRAPPERS FOR FEATURE SUBSET SELECTION". *Artificial Intelligence* Journal, 97(1-2): 273-324.

MIRANDA, E.; FIGUEIREDO, P. (2010): "DINÂMICA DA ACUMULAÇÃO DE CAPACIDAD ES INOVADORAS: EVIDÊNCIAS DE EMPRESAS DE SOFTWARE NO RIO DE JANEIRO E EM SÃO PAULO". *RAE: Revista de Administração de Empresas*, 50:75-93.

OCDE (2002): "MANUAL DE FRASCATI. ORGANIZACIÓN PARA LA COOPERACIÓN Y DESARROLLO ECONÓMICOS". OECD, París.

OCDE y Eurostat (2005): "MANUAL DE OSLO: GUÍA PARA LA RECOGIDA E INTERPRETACIÓN DE DATOS SOBRE INNOVACIÓN". OECF/ European Communities.

ROMERO, M. del C.; CAMIO, M. I., ÁLVAREZ, M. B. (2013): "CONSTRUCCIÓN DE ÍNDICES. UNA APLICACIÓN PARA MEDIR EL NIVEL DE INNOVACIÓN EN EMPRESAS DE SOFTWARE Y SERVICIOS INFORMÁTICOS". *XXVI Encuentro Nacional de Docentes en Investigación Operativa y XXIV Escuela de Perfeccionamiento en Investigación Operativa*. Córdoba, mayo de 2013.

ROMERO, M. del C.; CAMIO, M. I.; ÁLVAREZ, M. B. (2014): "CLASIFICACIÓN SUPERVISADA EN ALTA DIMENSIONALIDAD. UNA APLICACIÓN EN GESTIÓN ORGANIZACIONAL". *XXVII Encuentro Nacional de Docentes en Investigación Operativa y XXIV Escuela de Perfeccionamiento en Investigación Operativa*. San Nicolás, 21 al 23 de mayo de 2014.

ROMIJN, H.; ALBALADEJO, M. (2002): "DETERMINANTS OF INNOVATION CAPABILITY IN SMALL ELECTRONICS AND SOFTWARE FIRMS IN SOUTHEAST ENGLAND". *Research Policy*, 31(7):1053-1067.

RUIZ SÁNCHEZ, R. (2006): "HEURÍSTICAS DE SELECCIÓN DE ATRIBUTOS PARA DATOS DE GRAN DIMENSIONALIDAD". Memoria de Tesis Doctoral para optar al grado de Doctor en Informática por la Universidad de Sevilla. Sevilla, España.

SHANNON, C. E. (1948): "A MATHEMATICAL THEORY OF COMMUNICATION". *Bell System Technical Journal*, 7:379-423 y 623-656.