ISSN 2007-9737

Metodología multitareas automatizada para estudiar indicadores organizativos usando ciencia de datos

Jorge Pérez Rave[1,2], Gloria Jaramillo Álvarez[3], Favián González Echavarría[4]

[1] Grupo de investigación IDINNOV, IDINNOV S.A.S, Medellín, Colombia

[2] Universidad Nacional de Colombia, Programa de Formación Doctoral en Ingeniería de Sistemas, Facultad de Minas, Medellín, Colombia

[3] Universidad Nacional de Colombia, Departamento de Ciencias de la Computación y de la Decisión, Facultad de Minas, Medellín, Colombia

[4] Universidad de Antioquia, Departamento de Ingeniería Industrial, Facultad de Ingeniería, Medellín, Colombia

investigacion@idinnov.com, gpjarami@unal.edu.co, favian.gonzalez@udea.edu.co

Resumen. El objetivo es proveer una metodología multitareas para estudiar indicadores organizativos de forma automática, usando Ciencia de Datos. Consta de 7 etapas: preparación de datos, análisis univariado, análisis bivariado, patrones de agrupación, reducción de dimensiones, entrenamiento de modelos y validación. Se usa *R, RStudio* (procesar) y *Rmarkdown* (visualizar). Se aplica en cuatro casos (dos de manufactura y dos de servicios) y aporta información de utilidad a nivel organizativo y de enseñanza-aprendizaje. La metodología distingue entre indicadores de medios y de respuesta, integra más de 10 métodos de análisis, abarca los tres alcances estadísticos (univariado, bivariado y multivariado) y responde seis preguntas de interés para los analistas.

Keywords. Ciencia de datos, indicadores de negocios, programación en R, analítica, análisis de datos, metodología de análisis.

Automated Multitasking Methodology for Studying Business Indicators using Data Science

Abstract. The objective is to provide a multitasking methodology to study business indicators automatically using Data Science. This consists of 7 stages: data preparation, univariate analysis, bivariate analysis, grouping patterns, reduction of dimensions, supervised model training and validation. R, R-Studio (processing) and Rmarkdown (visualization) are used. The methodology is applied on four study cases (two from manufacturing sector and two from services) and provide useful information for firms and teaching-learning processes. The methodology distinguishes between media and response indicators, comprises more than 10 analysis methods, includes the three statistical scopes (univariate, bivariate and multivariate) and automatically answers six questions of interest for analysts.

Palabras clave. Data science, business indicators, R-programming, analytics, data analysis, analysis methodology.

1. Introducción

La Ciencia de Datos es un área multidisciplinaria enfocada en el estudio de patrones subyacentes en los datos (no necesariamente masivos, aunque allí viene teniendo mayor auge), combinando métodos estadísticos, matemáticos, computacionales y de los negocios [1-3]. En el ámbito organizativo esta disciplina tiene mucho por aportar, ya que prevalecen premisas como *lo que no se mide no se controla y lo que no se controla no se puede mejorar*. Según [4-5] el conocimiento generado a partir de la Ciencia de Datos puede apoyar la toma de decisiones en la organización y favorecer su

desempeño. De hecho, se argumenta que los estrategas son cada vez más dependientes de información derivada del análisis de datos [6].

A pesar de la importancia de esta disciplina y de su utilidad para las organizaciones, su uso en este contexto es declarado en infancia. Por ejemplo, [4, 6] advierten que son pocas las organizaciones que han incursionado en la Ciencia de Datos y que, además, las pocas que lo han hecho presentan un alcance limitado, por motivos como: ausencia de personal entrenado y poca calidad de las fuentes de información. En experiencias de consultoría de empresas, esto se refleja, por ejemplo, en alcances limitados al mero uso desarticulado de técnicas o paquetes computacionales sin un propósito global, o sin explotación de tareas previas, como la captura y la preparación de datos (limpieza, estructuración, creación de nuevas variables, etc.). Asimismo, el analista tiende a depender de instrucciones manuales cada vez que se enfrenta a un nuevo conjunto de datos, debido al uso de software basado en botones o de instrucciones de código básicas o no generalizables. Esto abre oportunidades para la automatización de las tareas, a través de algoritmos integrales que reduzcan tiempo y recursos, y estimulen la oportunidad en la publicación de información y la toma de decisiones.

Adicionalmente, [7] expresa que los algoritmos creados para automatizar tareas (ej.: análisis – visualización) suelen minar patrones en los datos sin considerar el interés y las necesidades del analista. Al respecto, es común delimitar el estudio de indicadores a uno o dos de los siguientes alcances: descriptivo, comparativo, correlacional, de agrupación, o de predicción.

Sin embargo, desde una perspectiva pragmática, el analista requiere de todos ellos, articulados de forma lógica y automática, donde los resultados de una etapa sean, a su vez, insumos de la siguiente. La razón de esta necesidad es simple, el analista de datos busca ayudar a comprender a los estrategas el estado actual del sistema, los patrones de agrupación, las relaciones entre los procesos, los efectos de las acciones tomadas y los predictores del desempeño. Y, para ello, es fundamental desde la mera descripción univariada hasta la inducción de patrones y estudios de. Dicha postura cobra aún más importancia, en la medida en que algunos indicadores empresariales son redundantes o no representan razonablemente los objetivos a los que se deben [8]. De ahí la necesidad de descubrir y abordar indicadores latentes (inducidos desde conjuntos de indicadores observables), que faciliten enfocarse en objetivos globales y no solo en métricas individuales, las cuales pueden llevar a direcciones contradictorias, malentendidos y esfuerzos ineficaces [9]. Sumado a esto, [5] precisa que las metodologías existentes no suelen detallar las actividades en cada paso, lo cual dificulta la comprensión, el uso y la reproducibilidad por parte del analista.

El objetivo, entonces, es proponer una metodología multitareas automatizada para estudiar indicadores organizativos usando Ciencia de Datos, que abarque los alcances univariado, bivariado y multivariado, y brinde respuestas a seis interrogantes pragmáticos: ¿Cómo preparar la base de datos?, ¿cuál es el *status* de cada uno de los indicadores?, ¿cuál es el grado de asociación entre pares de indicadores?, ¿qué patrones de agrupación subyacen en los indicadores y cuáles son sus características?, ¿qué estructuras subyacen en los grupos de indicadores y qué nuevas medidas las resumen?

Y ¿qué factores latentes y relación funcional permiten predecir el desempeño del sistema? La principal contribución de este trabajo se resume en que la metodología propuesta contempla los tres alcances estadísticos (univariado, bivariado y multivariado) y articula más de 10 métodos de análisis en una misma sistemática automatizada (estadísticos descriptivos; correlaciones Pearson, Kendall, tablas de contingencia con estadísticos Chi-2, Phi/V-de-Cramer; Análisis Clúster con enlaces simple, promedio y completo; Análisis de Componentes Principales, Regresión Logística, Árboles de Clasificación y estadísticos de la matriz de confusión) para brindar respuestas a seis preguntas de interés para el analista de datos.

Dichos métodos son reconocidos en el campo estadístico y están debidamente documentados en libros de texto, pero al interpretarlos desde la teoría de recursos y capacidades [10] se sabe que el mero uso de estos no garantiza ventajas competitivas.

ISSN 2007-9737

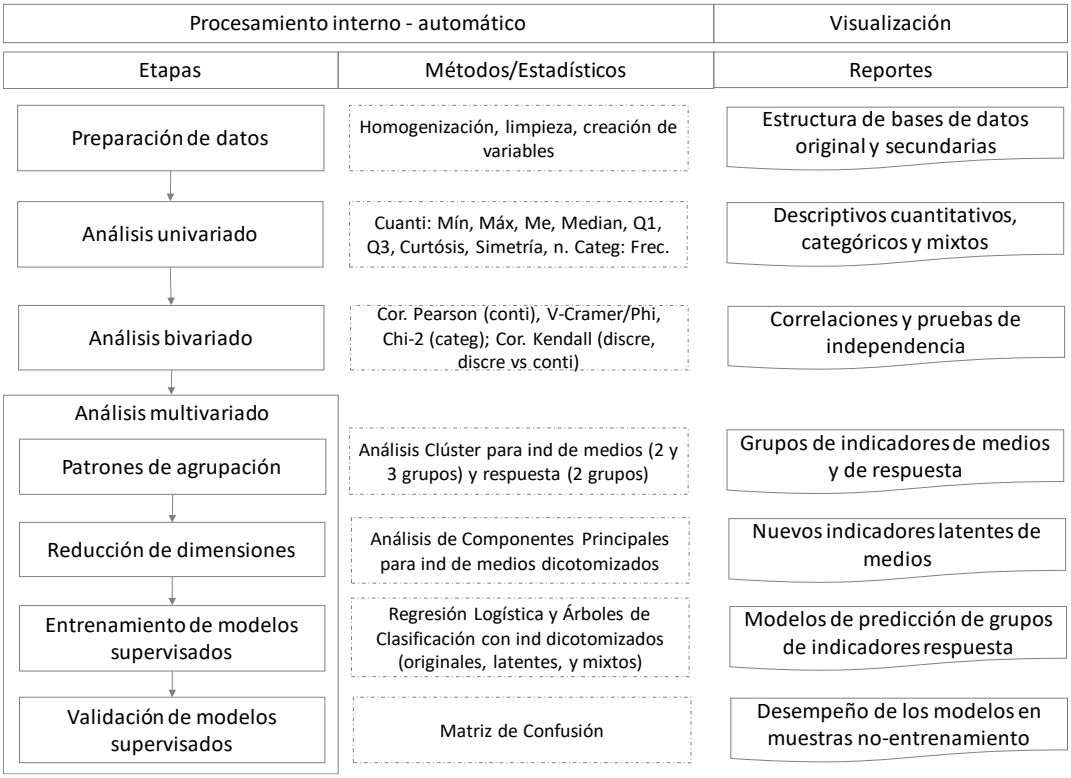


Fig. 1. Metodología propuesta: etapas, métodos/estadísticos y reportes automáticos

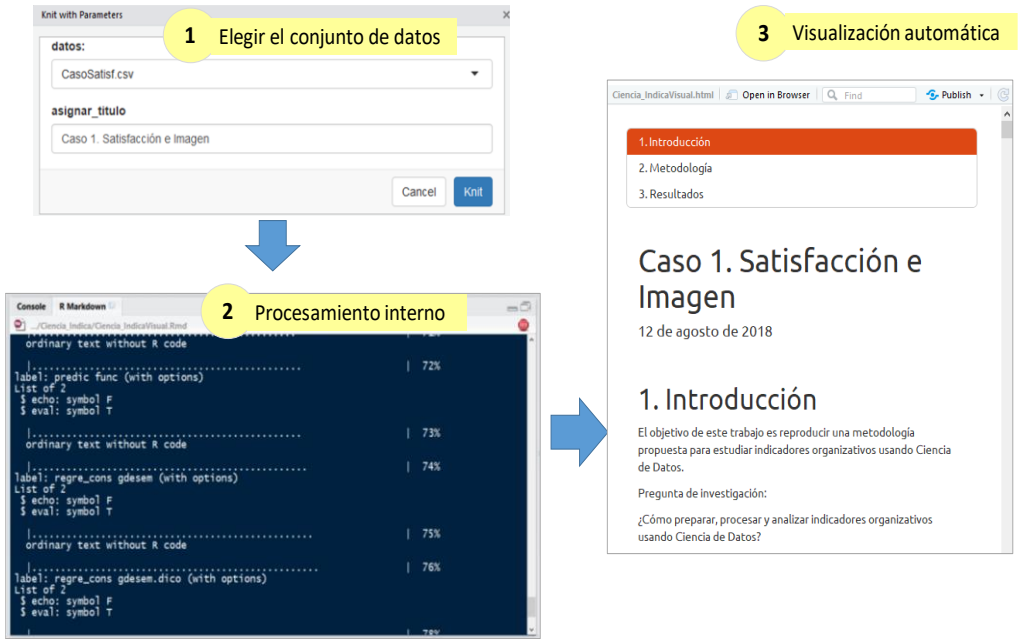


Fig. 2. Ilustración de los tres momentos requeridos para ejecutar la metodología propuesta, usando *Rmarkdown*

ISSN 2007-9737

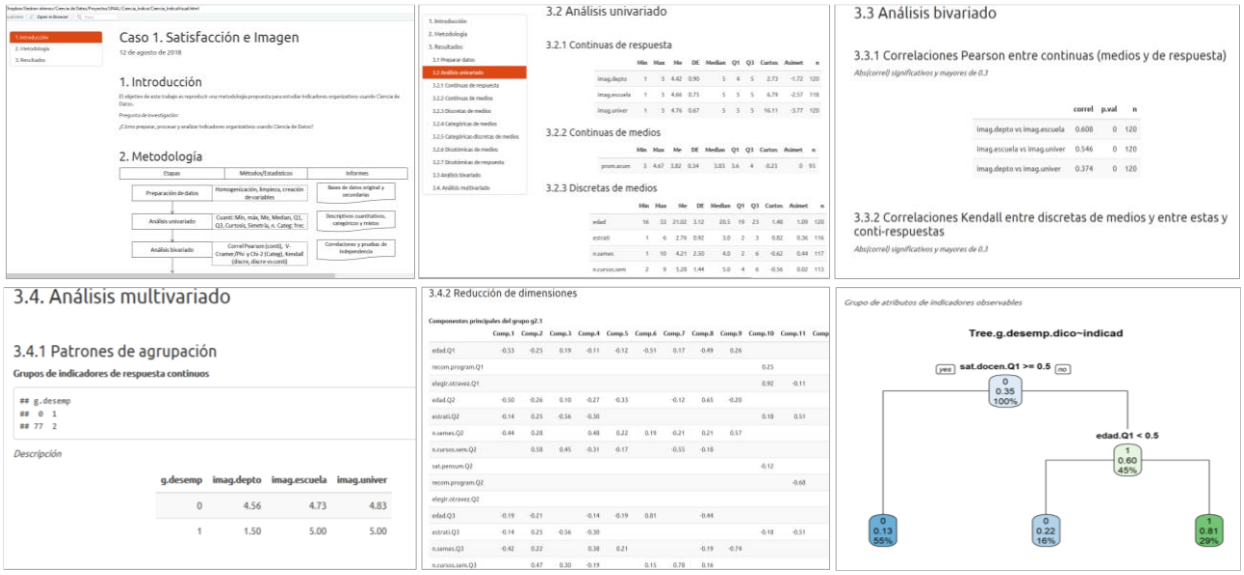


Fig. 3. Ilustración de los reportes automáticos en *HTML*, generados por la ejecución de la metodología, usando *Rmarkdown*

No obstante, la forma de utilizarlos, relacionarlos, combinarlos, automatizarlos, etc., (procesos de orquestación de los recursos: [11]) representa todo un desafío y puede derivar en nuevas capacidades dinámicas–analíticas que distingan la organización de sus competidores, renueven las operaciones y generen valor [12].

Esta posición es consistente con [13, p.30], al referirse a tecnologías como la minería de datos "En el mercado, existe hasta cierto punto la expectativa de que la minería de datos es una tecnología consistente en pulsar un botón. Sin embargo, esto no es cierto…", "…depende [su éxito] de la combinación adecuada de buenas herramientas y analistas expertos.

Además, requiere una metodología sólida y una gestión efectiva del proyecto". El artículo está organizado así: la sección 2 describe la metodología propuesta. La sección 3 resume los resultados y análisis de aplicación en cuatro casos de estudio. La sección 4 plasma las conclusiones y los trabajos futuros.

2. Metodología propuesta

La metodología consta de siete etapas, las cuales se exponen en la Figura 1. Dichas etapas fueron automatizadas por medio de lenguaje *R* [14] y el entorno *RStudio* [15]. En la columna central de la Figura 1 se nombran los métodos y los estadísticos usados en cada etapa y, en la columna derecha, los reporte generados de forma automática por medio de *Rmarkdown* [16].

Para la ejecución de la metodología, basta con que se desplieguen tres momentos, expuestos en la Figura 2. Un breve extracto de diferentes visualizaciones que provee la metodología es presentado en la Figura 3. A continuación, se describe brevemente cada una de las etapas que conforman la metodología.

2.1. Preparación de datos

El objetivo de esta etapa es facilitar la homogenización y limpieza de datos, y generar nuevos conjuntos de datos, a partir de los indicadores originales. Para lo primero, automáticamente todos los datos tipo "carácter" se unifican en mayúsculas y sin tilde.

Luego, se genera un reporte del estado de la base de datos original, en términos de siete campos: número de observaciones (*observ*), variables (*variab*), cantidad de valores faltantes (*NAs*) en la variable con el máximo número de estos (*var.na.max*), números mínimo y máximo de

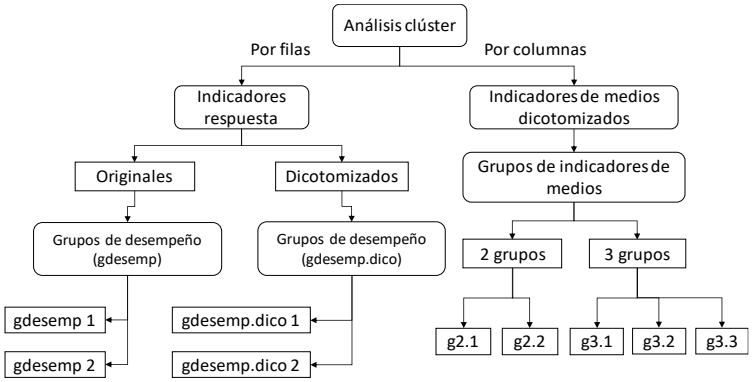


Fig. 4. Inducción de grupos de desempeño y de indicadores de medios

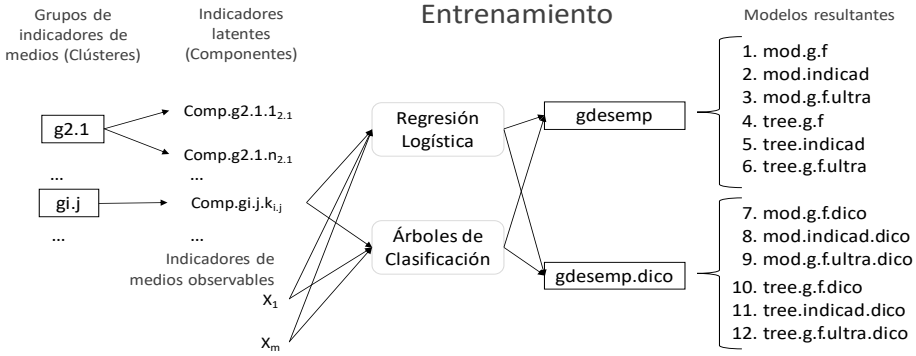


Fig. 5. Inducción de grupos de desempeño y de grupos de indicadores de medios

niveles de respuesta entre las variables categóricas (*var.level.min*, *var.level.max*), cantidad de variables con valores negativos (*var.neg*), cantidad de valores cero (0) en la variable con el máximo número de estos (*var.cero.max*) y nombres de las variables que se sugiere revisar, con base en los campos previos (*alerta*).

Con este reporte el usuario tendrá insumos para revisar su conjunto de datos original, realizar ajustes y continuar visualizando los demás reportes, o bien, ajustar el conjunto de datos y repetir la ejecución de la metodología propuesta.

El diseño del conjunto de datos que se acostumbra para la alimentación y el monitoreo de indicadores organizativos sigue el formato filas (observaciones) y columnas (indicadores). A su vez, de acuerdo con diversos modelos de gestión (Ej: EFQM), una organización debe distinguir entre medios y resultados.

Ello significa que unos pocos indicadores son los que representan los resultados principales en el proceso u organización en general, en tanto que otros corresponden a medios para impactarlos. Los primeros se denominarán *indicadores de respuesta* y, los segundos, *indicadores de medios*. Para la ejecución automática de la metodología basta que el analista provea el conjunto de datos y especifique qué indicadores son de respuesta.

Luego, esta misma etapa ejecuta la creación automática de seis bases de datos secundarias, clasificadas según tipologías de indicadores: categóricos, continuos, discretos, dicotómicos, respuesta-continuos y respuesta-dicotómicos. La dicotomización se realiza para indicadores categóricos (politómicos) y cuantitativos (continuos o discretos), tanto de medios como de respuesta. La partición de los indicadores categóricos se realiza con base en la moda (*mo*), en tanto que, para las demás, adicional a la moda

también considera los cuartiles (*Q1*, *Q2* y *Q3*). Esto significa que una determinada variable cuantitativa "X", del conjunto de datos original, genera cuatro variables secundarias: *X.mo*, *X.Q1*, *X.Q2* y *X.Q3*, dando lugar a una gran cantidad de nuevas variables para estudio.

2.2. Análisis de indicadores individuales (univariado)

El objetivo es posibilitar el reconocimiento individual de cada indicador, tomando en consideración tendencia central, dispersión, localización y forma distribucional. En esa vía, una vez creadas las bases de datos secundarias, también internamente se ejecuta el análisis univariado. Para los indicadores cuantitativos se recurre a nueve estadísticos resumen. Para los indicadores categóricos se automatiza la obtención de frecuencias (absolutas). Para los indicadores dicotomizados previamente, se reportan, entre otros, las frecuencias absolutas y relativas de cada categoría ("1", "0").

2.3. Análisis de pares de indicadores (bivariado)

Esta etapa tiene como fin explorar y retratar posibles asociaciones estadísticamente significativas entre pares de indicadores. Internamente, la metodología realiza cuatro análisis. El primero para indicadores continuos (correlaciones Pearson); el segundo para discretos o una mezcla de estos con continuos (correlaciones Kendall), el tercero para indicadores categóricos puros, y el cuatro para los indicadores dicotomizados. En estos dos últimos se usa la prueba de independencia Chi-2 y la V-de Cramer (en dicotómicos equivale al test Phi). Esto ejecuta una gran cantidad de comparaciones, según las dimensiones de los conjuntos de datos. Por lo mismo, el reporte exhibe solo las asociaciones significativas al 0.05 y, además, cuyas medidas de correlación (o asociación), en valor absoluto, supere un umbral determinado |p| (entre 0 y 1). Un resumen de los métodos/estadísticos de esta etapa puede verse en la anterior Figura 1.

Tabla 1. Salidas automáticas del proceso de preparación de datos del caso 1

Métricas	Datos originales	Datos secundarios	
observ	120	**Variables**	**Cant**
variab	15	Categóricas	4
var.na.max	27	Continuas	1
var.level.min	2	Discretas	7
var.level.max	3	Dicotómicas	28
var.neg	0	Respuesta-continuas	3
var.cero.max	0	Respuesta-dicotómicas	9
alerta	prom.acum		

2.4. Análisis de grupos de indicadores (multivariado)

Este análisis multivariado automatizado considera inicialmente métodos no supervisados para inducir patrones desde los datos.

Luego, utiliza métodos supervisados, guiados por grupos de indicadores respuesta. A continuación, se describen las sub-etapas de cada uno de ellos.

– **Patrones de agrupación:**

El objetivo es inducir posibles conglomerados a través de Análisis Clúster. Uno para las filas (registros) y otro para las columnas. Para los registros (filas) se ha fijado como parámetro explorar el mínimo número de grupos que pueden distinguir el desempeño del sistema (2 grupos). Esto tiene varias razones: 1) la exclusión de puntos intermedios en las clasificaciones de desempeño facilita discriminación de los registros (filas) en dos extremos (alto y bajo desempeño). 2) Se reduce el consumo computacional y 3) se facilita la interpretación por parte del usuario, en comparación con 3 o más categorías de desempeño.

Esto genera internamente una nueva variable respuesta (dicotómica: 1, 0), alusiva a cada grupo de desempeño. El Análisis Clúster amerita inicialmente el cálculo de la matriz de distancias entre los elementos.

Tabla 2. Salidas automáticas de la descripción univariada para el caso 1

Indicadores de respuesta (continuos o discretos)										
Nombres	Min	Max	Me	DE	Median	Q1	Q3	Curtos	Asimet	n
imag.depto	1	5	4.42	0.9	5	4	5	2.73	-1.72	120
imag.escuela	1	5	4.66	0.8	5	5	5	6.79	-2.57	118
imag.univer	1	5	4.76	0.7	5	5	5	16.11	-3.77	120
Indicadores de medios (continuos o discretos)										
Nombres	Min	Max	Me	DE	Median	Q1	Q3	Curtos	Asimet	n
prom.acum	3	4.7	3.82	0.3	3.83	4	4	-0.23	0	93
edad	16	33	21	3.1	20.5	19	23	1.48	1.09	120
estrati	1	6	2.76	0.9	3	2	3	0.82	0.36	116
n.semes	1	10	4.21	2.5	4	2	6	-0.62	0.44	117
n.cursos.sem	2	9	5.28	1.4	5	4	6	-0.56	0.02	113
sat.metod	1	5	4.17	1	4	4	5	1.49	-1.34	120
sat.docen	1	5	4.26	1.1	5	4	5	2.29	-1.7	120
sat.pensum	1	5	4.17	1.2	5	4	5	-0.01	-1.15	120
Indicadores categóricos										
género		semes.categ			cursos.categ			trabaja		
f: 64		adaptados: 69			hasta 4: 33			no: 75		
m: 55		avanzados: 14			mas de 4: 80			si: 42		
NA's: 1		nuevos: 34			NA's: 7			NA's: 3		
		NA's: 3								
Indicadores de medios – Dicotomizados (Extracto)										
Nombres			Min	Max	Frec.1	Prop.1	Frec.0	Prop.0	n	
1. cursos.categ.mo			0	1	80	0.71	33	0.29	113	
2. edad.Q1			0	1	81	0.68	39	0.32	120	
3. edad.Q2			0	1	60	0.5	60	0.5	120	
28. sat.pensum.Q3			0	0	0	0	120	1	120	

Esta se calcula para los indicadores respuesta en su forma natural (Ej: continuas, que es lo más común) usando la distancia euclidiana. Y también se calcula para dichos indicadores dicotomizados (según sección 2.1, con base en cuartiles), a través del coeficiente de similaridad de Jaccard.

Para los indicadores de medios (columnas) se estudia la conformación de 2 y 3 posibles conglomerados (nuevos indicadores latentes); cada uno de ellos servirá en posteriores etapas para explorar su papel predictor del desempeño. Internamente, la conformación de grupos se ejecuta con base en los enlaces simple (“vecino más cercano”), completo (“vecino más lejano”) y “promedio”, y se elige aquel que arroja la clasificación con mayor diversidad, representado por la menor desviación estándar. Por ejemplo, una clasificación que arroje un grupo de 99 elementos y otro de 1 elemento infiere baja diversidad, pues la mayoría pertenece a una sola categoría (desviación: 69.3). En cambio, una clasificación de 51 (grupo “1”) y 49 elementos (grupo “2”) tiene alta diversidad (desviación: 1.4). Esta etapa arroja dos grupos de indicadores de desempeño y cinco grupos de indicadores de medios. En la Figura 4 se ilustra la conformación de los grupos.

– **Reducción de dimensiones:**

Se busca generar combinaciones lineales de los grupos de indicadores de medios, arrojados por la etapa anterior, de modo que se cuente con indicadores latentes que los representen. Para ello, de forma automática se ejecuta el Análisis de Componentes Principales (ACP) y los *scores* de dichas combinaciones se almacenan en nuevas variables (una por componente).

– **Entrenamiento de modelos supervisados:**

Son dos los tipos de modelos, uno paramétrico (regresión logística) y otro no paramétrico (árboles de clasificación). En la Figura 5 se ilustra la generación de estos modelos.

En total son 12 los modelos que se obtienen, seis de cada tipología (regresión, árbol). El entrenamiento se realiza con el 70% de las observaciones (filas), seleccionadas de forma aleatoria, y el 30% restante se separa para la validación en la siguiente etapa. Dicho entrenamiento se realiza en cuatro fases. La primera toma como predictores los indicadores de cada grupo por separado (componentes). La segunda combina los mejores predictores de dichos modelos y ejecuta nuevamente el entrenamiento, dando lugar a los modelos resultantes basados en indicadores latentes (componentes), los cuales se denominan para regresiones: *mod.g.f* y *mod.g.f.dico*, y para árboles: *tree.g.f* y *tree.g.f.dico*. La tercera ejecuta la modelación solo con indicadores observables como predictores, dando lugar a *mod.indicad*, *mod.indicad.dico*, *tree.indicad*, *tree.indicad.dico*. En la cuarta fase se entrenan nuevos modelos usando los predictores de los modelos previos (latentes y observables), llevando a: *mod.g.f.ultra* y *mod.g.f.ultra.dico*, y *tree.g.f.ultra* y *tree.g.f.ultra.dico*.

– **Validación de modelos supervisados:**

El objetivo de esta etapa es explorar la capacidad predictiva de los modelos entrenados, a través de la ejecución automática de cada uno de ellos en la muestra no-entrenamiento, que corresponde al 30% del total de registros. El desempeño se estima por medio de tres métricas de la matriz de confusión: Tasa de Aciertos (TA), Tasa de Aciertos Positivos (TAP) y Tasa de Aciertos Negativos (TAN).

3. Resultados

El despliegue de la metodología se esboza para cuatro casos. Vale anotar que el analista solo debe ingresar el conjunto de datos y especificar los indicadores de respuesta; todo lo demás se genera de forma automática y se obtienen reportes en *HTML* por medio de *Rmarkdown* (previas Figuras 2-3).

Tabla 3. Salidas automáticas de medidas de asociación de pares de indicadores cuantitativos y categóricos del caso 1 (bajo significancia de 0.05 y $|p|\geq 0.3$)

Correlaciones Pearson (Continuas)	
Pares de indicadores	Correl
imag.depto vs imag.escuela	0.608
imag.escuela vs imag.univer	0.546
imag.depto vs imag.univer	0.374
Cor. Kendall (Discretas, cont) (Extracto)	
Pares de indicadores	Correl
imag.depto vs imag.escuela	0.579
sat.metod vs sat.docen	0.547
imag.escuela vs imag.univer	0.482
edad vs n.semes	0.443
sat.metod vs imag.depto	0.324

Cor. V-Cramer/Phi (dicotómicas, politóm) (Extracto)		
Pares de indicadores	Coef.Cram	Chi2
edad..n.semes.Q1Q2	0.527	31.15
sat.metod..sat.pensum.Q1Q1	0.453	22.83
sat.metod..sat.docen.Q1Q1	0.448	22.27
sat.docen..imag.depto.Q1Q1	0.41	18.51
edad..trabaja.Q1mo	0.362	14.15

3.1. Caso 1: satisfacción e imagen del programa

Esta base de datos se deriva de una encuesta a estudiantes acerca de la satisfacción e imagen con aspectos de una institución de educación superior.

3.1.1. Preparación de datos

En la Tabla 1 se presenta los reportes automáticos de preparación de datos. En la parte izquierda para la base de datos original y en la parte derecha para las bases de datos secundarias.

Dicho conjunto consta de 120 registros (filas) y 15 (indicadores). Cada registro hace referencia a las respuestas dadas por cada estudiante, ante la consulta sobre diversos aspectos de la institución

Tabla 4. Salidas automáticas para los patrones de agrupación bajo Análisis Clúster

Descripción de grupos de indicadores respuesta continuos					
g.desemp	Cant	imag.depto	imag.escuela	imag.univer	Significado
0	77	4.56	4.73	4.83	Sin distinción
1	2	1.5	5	5	
Descripción de grupos de indicadores respuesta dicotomizados					
g.desemp.dico	Cant	**imag.depto**	**imag.escuela**	**imag.univer**	**Significado**
0	48	5	4.96	4.92	Mejor desempeño
1	31	3.68	4.39	4.71	Peor desempeño
Conformación de grupos de indicadores de medios					
Cant. Grupos		**1**	**2**		**3**
2 (g2)		14	14		NA
3 (g3)		5	13		10

Tabla 5. Salidas automáticas de estructuras de indicadores latentes bajo ACP

Variables	Comp.1	Comp.2	Comp.5
edad.Q1	-0.55	-0.38	-0.23
estrati.Q1	-0.16	0.55	
edad.Q2	-0.50	-0.48	0.16
n.semes.Q2	-0.46	0.43	-0.6
n.semes.Q3	-0.44	0.37	0.75

de educación superior. Estos aspectos corresponden a las columnas, que pueden dividirse en indicadores de medios y de resultados (respuesta). Los de medios comprenden: satisfacción con docente, métodos de enseñanza y pensum; e indicadores de segmentación de quien responde (edad, género, etc.). Los de respuesta corresponden a la imagen del departamento, la escuela y la institución en general. Estos últimos, amparados en que ante una imagen favorable de la institución es más probable desencadenar intenciones conductuales positivas (Ej: recomendación, etc.).

La escala de los indicadores de satisfacción e imagen fue de 1 (peor percepción) a 5 (mejor percepción). Este proceso de preparación también permitió informar sobre la calidad de los datos originales. Por ejemplo, se alertó sobre el indicador “promedio acumulado” del estudiante, pues fue el que más valores faltantes presentó (*var.na.max*: 27). En el resto no se encontraron razones de alerta.

A partir de los 15 indicadores originales se crearon nuevas métricas dicotomizadas. Se pasó a 37 indicadores (28 de medios y 9 de respuesta, ambos dicotomizados), lo que equivale a un aumento del 146%. La división del dominio de valores se hizo con base en la moda (para categóricos) y los cuartiles.

3.1.2. Análisis univariado

En la tabla 2 se resume cada indicador del caso 1, según tipología (cuantitativos, categóricos, dicotómicos), fruto de las salidas automáticas que arroja el despliegue de la metodología.

El análisis univariado posibilitó obtener una mirada descriptiva de cada indicador, considerando tendencia central, dispersión, localización y forma distribucional. Esto es importante en el medio organizativo, ya que periódicamente el analista debe exhibir ante los decisores el estado actual de los procesos, a partir de los indicadores que los representen. Este

reporte ayuda a detectar posibles datos extraños y saber si se está dentro o fuera de las metas establecidas. Ilustrando los indicadores de respuesta (imagen institucional) de este caso, se tiende a una percepción favorable entre los estudiantes, con medias, medianas y cuartiles de mínimo 4 puntos de calificación (Tabla 2). En cuanto a los indicadores categóricos, prevalece: género "femenino" (53%), estudiantes "adaptados" a la institución en cuanto al tiempo de permanencia (57.5%), inscritos en más de 4 cursos en el semestre (66.7%), y que no están laborando (62.5%).

3.1.3. Análisis bivariado

En la Tabla 3 se resumen las salidas automáticas para el análisis de asociación entre pares de indicadores.

La descripción bivariada permite explorar posibles asociaciones entre pares de indicadores. El sistema reporta solo aquellas estadísticamente significativas (α: 0.05) y con medidas de asociación (en valor absoluto) superiores a 0.3 (p; este parámetro puede modificarse). Desde el punto de vista práctico, resulta útil comprender cómo se relacionan los indicadores. Nótese la asociación significativa entre los indicadores de respuesta (imagen, Tabla 3), con correlaciones *Pearson* desde 0.374 hasta 0.608, todas significativas, lo cual da una idea de lo consistente que puede ser este constructo subyacente. En esa misma tabla, sección de *Cor. Kendall*, se observan otros pares con correlaciones significativas, entre ellas: indicadores de satisfacción e imagen (0.324), lo cual es consistente con el fenómeno y lo que dicta la teoría (a mas satisfacción, mejor imagen) [17-18]. También se incluye *V-Cramer/Phi* para categóricos (incluyendo dicotómicos), con el fin de cubrir toda la región de combinaciones.

3.1.4. Análisis multivariado

Patrones de agrupación:

La tabla 4 presenta los reportes del Análisis Clúster para grupos de desempeño (ejecutado sobre registros: filas) y para grupos de indicadores de medios (columnas).

Pasando a los patrones de agrupación, la Tabla 4 ofrece resultados para la conformación de posibles grupos de desempeño (Análisis Clúster por filas), tomando como insumos tanto indicadores continuos (*g.desemp*) como dicotomizados (*g.desemp.dico*). Para el escenario de *g.desemp*, la agrupación no arrojó una diversidad razonable, pues la mayoría de filas (respuestas de los estudiantes) se ubicó en el grupo "1" (77 de 79 registros completos: 97.4%). Contrario sucede en los clúster de indicadores ya dicotomizados (*g.desemp.dico*), en los que la partición binaria es más diversa.

Tabla 6. Salidas automáticas de regresiones de *g.desemp.dico* usando grupos de indicadores de medios

	g2.1	g2.2	g3.1	g3.2	g3.3
(Interc)	−0.70	−0.81*	−0.72*	−0.74*	−0.64*
	(0.31)	(0.33)	(0.31)	(0.34)	(0.28)
Comp.1	−0.89*		−0.95*	1.49***	
	(0.36)		(0.39)	(0.44)	
Comp.13		−315.6*			
		(124.2)			
Aldrich-N. R2	0.10	0.14	0.10	0.22	
Likelih.Ratio	6.09	8.93	6.37	15.36	
p-val	0.01	0.003	0.012	0	
Log-likelih.	−32.4	−31.0	−32.3	−27.8	−35.5
Deviance	64.8	62.0	64.5	55.5	70.9
AIC	64.8	66.0	68.5	59.5	72.9
N	55	55	55	55	55

*p-val<0.05; **p-val<0.01; ***p-val<0.001; (Err.Est)

El grupo de etiqueta "0" reúne 48 registros (60.7%), en tanto que el de etiqueta "1" presenta 31 registros (39.3%). Además, al explorar las medias de los grupos, sobresalen dos patrones de estudiantes. Los del grupo "0" tienden a calificar más alto la imagen en los tres indicadores de respuesta (depto: 5, escuela: 4.96 y universidad: 4.92), en comparación con los del grupo "1" (depto: 3.68, escuela: 4.39 y universidad: 4.71). Esto resulta útil para inducir patrones de desempeño y explorar sus predictores.

Adicionalmente, el Análisis Clúster es ejecutado para conformar posibles grupos de indicadores de medios (columnas), probando con 2 y 3 conglomerados. Los grupos obtenidos (Tabla 4) para el presente caso de estudio, se muestran diversos; fijando 2 grupos: 14 indicadores en cada grupo; y fijando 3 grupos: 5, 13 y 10 indicadores.

ISSN 2007-9737

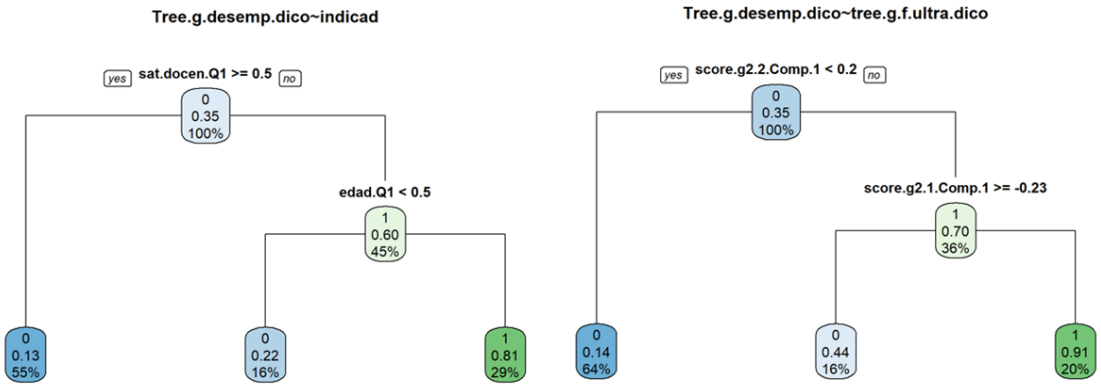


Fig. 6. Salidas automáticas. Caso 1. Árboles para *g.desemp.dico*. A la izquierda solo con indicadores observables (*tree.indicad*) y, a la derecha, el resultante combina observables y latentes (*tree.g.f.ultra.dico*)

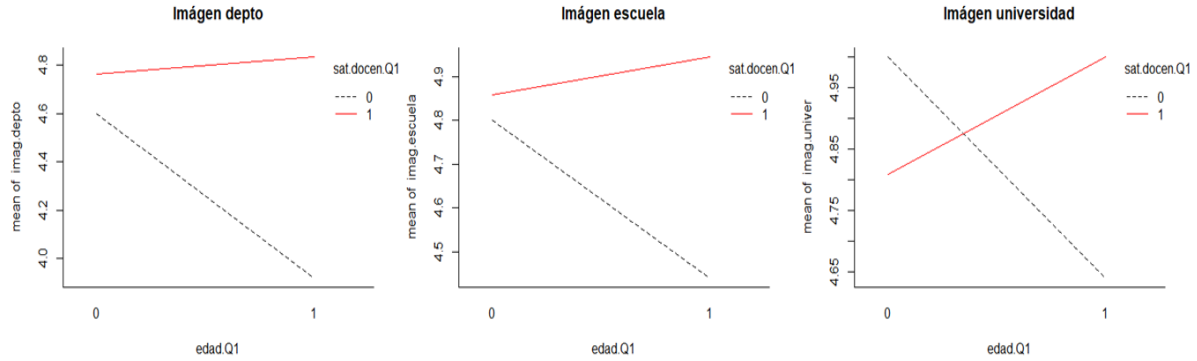


Fig. 7. Gráfico de interacción entre los niveles de edad y satisfacción con docente, caso 1, con respecto a imagen del departamento

Tabla 7. Salidas automáticas de validación de modelos para el caso 1

modelos	TA	TAN	TAP
1. mod.g.f			
2. mod.indicad			
3. mod.g.f.ultra			
4. mod.g.f.dico	**0.58**	**0.5**	**0.67**
5. mod.indicad.dico	0.54	0.5	0.58
6. mod.g.f.ultra.dico	0.54	0.5	0.58
10. tree.g.f.dico	0.58	0.75	0.42
11. tree.indicad.dico	**0.63**	**0.75**	**0.5**
12. tree.g.f.ultra.dico	0.58	0.75	0.42

TA: Tasa de aciertos; TAN: Tasa de Aciertos Negativos; Tasa de Aciertos Positivos

Respecto a la reducción de dimensiones, en la Tabla 5 se presenta las componentes del grupo g3.1, las cuales ocupan menor extensión y son suficientes para ilustrar el reporte de la metodología.

La reducción de dimensiones bajo ACP permite generar indicadores latentes, los cuales consisten en combinaciones lineales de los indicadores observables en cada grupo de medios. Por ejemplo, al considerar las componentes del grupo g3.1 (3 conglomerados, primer grupo de medios, que está compuesto por 5 indicadores observables; Tablas 4-5), este incorpora *edad.Q1*, *estrati.Q1*, *edad.Q2* y *n.semes.Q3*. Los signos de estos indicadores en la primera componente principal son todos los mismos (negativos; Tabla 5). En la medida en que los *scores* de este indicador latente tomen valores más negativos (más bajos), tiende hacia un estudiante más avanzado en edad, en semestres cursados y en condiciones económicas; contrario a cuando este indicador latente toma valores más positivos. El uso de componentes principales, además de reducir dimensiones para efectos del entrenamiento de modelos supervisados (siguiente etapa), es útil para la práctica organizacional, con el fin de tener una mejor comprensión de complejidades e información presente en conjuntos de indicadores. Esto le facilita al analista proponer nuevas métricas de resultados globales para equipos de trabajo, áreas o procesos.

Entrenamiento de modelos supervisados:

La metodología automatizada arroja 12 alternativas de modelización (6 de regresión logística y 6 de árbol), fruto de varios análisis internos que fueron ilustrados por medio de la Figura 5. A modo de interpretación, la Tabla 6 ofrece los modelos de regresión para *g.esemp.dico*, en función de las componentes de los grupos de indicadores de medios. El primer modelo es obtenido usando las componentes principales de los indicadores del grupo g2.1 (véase Figura 4 para recordar la conformación de grupos). Luego de ejecutar rutinas internas con base en regresión por pasos (*Backward*) y los criterios *AIC* y el *valor-p*, se obtuvo la componente principal 1 como único predictor significativo (en el grupo g2.2, además de esta, también fue significativa la componente 13). La interpretación del coeficiente de los *scores* de la componente 1 (-0.89) es: a medida que los *scores* aumentan (se hacen más positivos) tiende a ser mayor la probabilidad de calificar en el grupo “0” (mejor desempeño). Esta componente comprende indicadores de edad, número de semestres, estrato (con signos negativos) y promedio académico en el semestre (signo positivo).

En otras palabras, aquellos estudiantes más jóvenes (en edad y semestres cursados) y de condiciones económicas menores, pero con mejores promedios académicos, tienden a ubicarse en el grupo “0” (perciben más favorable la imagen de la institución).

En cambio, los estudiantes más avanzados en edad y semestres cursados y con menores promedios académicos tienden a calificar menor la imagen del servicio (en comparación con los primeros).

Todo esto ofrece información fundamental, subyacente, que puede alimentar la toma de decisiones estratégica y táctica. Así, el usuario puede ir a los reportes de entrenamiento de modelos y de patrones de agrupación que arroja el sistema y proceder con la interpretación y obtención de nueva información de interés.

Otro elemento de utilidad en la sección de entrenamiento de modelos es los árboles de clasificación, los cuales son más simples de interpretar que los resultados de regresión. Por ejemplo, en la Figura 6, árbol de la izquierda, se está modelando *g.desemp.dico*, en función de solamente indicadores observables.

De todos los indicadores observables dicotomizados, solo dos resultaron ser predictores importantes: la satisfacción con el docente (dividida con base en cuartil 1) y la edad (también binaria usando cuartil 1).

La interpretación es la siguiente: aquellos estudiantes que están satisfechos con el docente tienden a ubicarse en el grupo “0” (mejor desempeño), así como los que tienen menores puntuaciones en la satisfacción con el docente (inferiores a Q1), pero que a su vez son de edades menores (inferior a Q1).

Pero si estos presentan edades mayores, tienden a ubicarse en el grupo “1” (puntuaciones menores en imagen).

ISSN 2007-9737

— **Validación de modelos supervisados en muestras no-entrenamiento:**

La validación de los modelos supervisados se realizó con el 30% de los registros no-entrenamiento. La Tabla 7 ofrece los resultados comparativos entre los modelos. Esta tabla refleja dos modelos como los más prometedores: *mod.g.f.dico* (regresión) y *tree.indicad.dico* (árbol). Este último presenta mejores tasas de acierto en comparación con el primero, excepto para la TAP.

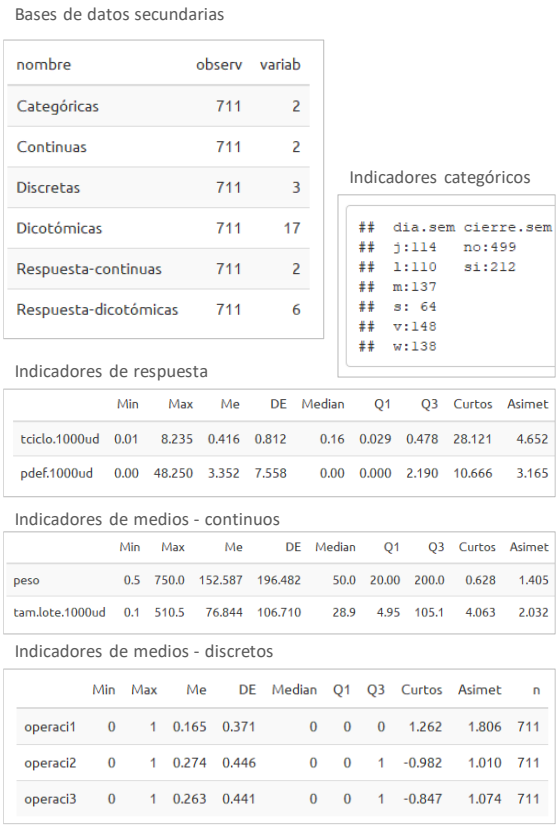
A modo de ejemplo se tomará el modelo basado en árboles de clasificación, dada la parsimonia que presenta y las mayores tasas de acierto en TA y TAN. Dicho árbol fue expuesto en la Figura 6, lado izquierdo, y consta solo de *sat.docen.Q1* y *edad.Q1*. Para nutrir la discusión, se ha realizado un análisis de segmentación usando un gráfico de interacción, el cual se plasma en la Figura 7.

Nótese, en la Figura 7, que los estudiantes que reflejan más satisfacción con respecto al docente, tienden a percibir la imagen de la universidad y la escuela relativamente similar (muy favorable); en cambio, cuando reflejan menor satisfacción docente, estas difieren según si el estudiante es joven o de más edad. Nótese que las menores puntuaciones medias en imagen de universidad y escuela fueron generadas por estudiantes de mayor edad y con menor puntuación en satisfacción docente.

Esto es consistente con lo arrojado por el árbol de clasificación y el modelo de regresión (este último también incluyó: promedio académico y estrato). Con relación a la imagen de la universidad (Figura 7, gráfico derecho), el patrón es similar (bajo test analítico de contraste de medias no hay diferencias significativas en el nivel "0"). Todos estos hallazgos, que son un ejemplo de lo que posibilita ser realizado con la información de los reportes automáticos, evidencia utilidad de la metodología propuesta.

3.2. Caso 2: calidad y eficiencia manufacturera

Este conjunto de datos consta de 711 registros de producción diaria de una empresa manufactura y 9 variables. Los indicadores de respuesta objeto de estudio son: número de defectos (*pdef.1000ud*) y tiempo de ciclo en horas (*tciclo.100ud*); ambos


Bases de datos secundarias

nombre	observ	variab
Categóricas	711	2
Continuas	711	2
Discretas	711	3
Dicotómicas	711	17
Respuesta-continuas	711	2
Respuesta-dicotómicas	711	6

Indicadores categóricos

```
##  dia.sem cierre.sem
##  j:114   no:499
##  l:110   si:212
##  m:137
##  s: 64
##  v:148
##  w:138
```

Indicadores de respuesta

	Min	Max	Me	DE	Median	Q1	Q3	Curtos	Asimet
tciclo.1000ud	0.01	8.235	0.416	0.812	0.16	0.029	0.478	28.121	4.652
pdef.1000ud	0.00	48.250	3.352	7.558	0.00	0.000	2.190	10.666	3.165

Indicadores de medios - continuos

	Min	Max	Me	DE	Median	Q1	Q3	Curtos	Asimet
peso	0.5	750.0	152.587	196.482	50.0	20.00	200.0	0.628	1.405
tam.lote.1000ud	0.1	510.5	76.844	106.710	28.9	4.95	105.1	4.063	2.032

Indicadores de medios - discretos

	Min	Max	Me	DE	Median	Q1	Q3	Curtos	Asimet	n
operaci1	0	1	0.165	0.371	0	0	0	1.262	1.806	711
operaci2	0	1	0.274	0.446	0	0	1	-0.982	1.010	711
operaci3	0	1	0.263	0.441	0	0	1	-0.847	1.074	711

Fig. 8. Salidas automáticas de reporte *html* bajo *Rmarkdown* para descripción univariada del caso 2

por cada 1000 unidades. Los siete indicadores de medios, son: peso del producto (en mg); día de la semana en que se produce (*dia.sem*: L – D); evento de si se produce al cierre de semana o no (*cierre.sem;* "si": viernes a domingo); tamaño de lote en miles de unidades (*tam.lote.1000ud*) y operación que se realiza (*operaci1*, …, *operaci3*; cada una de tipo binario, mutuamente excluyentes; cuando estas toman valor cero, se hace referencia a la operación 4).

En la Figura 8 se aportan extractos del reporte de descripción univariada, obtenido de forma automática. Nótese que, a partir de las nueve variables originales, se han creado 17 variables dicotomizadas de medios y seis de respuesta, fruto de la preparación de datos. A modo de ilustración, véase que, respecto a indicadores categóricos, los reportes de producción son relativamente estables en la muestra para los días en semana (L – V), oscilando entre 110 y 148

registros; para el día sábado (S) la cantidad de registros es menor (64).

Con relación a indicadores de respuesta, enfatizando, por ejemplo, sobre la cantidad de defectos por cada mil unidades elaboradas (pdef.1000ud), en el 50% de los casos la calidad es totalmente favorable (Q2=0) y en el 25% de los casos ocurren 2.19 defectos por cada mil unidades (Q3). Si bien a primera vista este último valor se muestra bajo, dista considerablemente del referente pragmático de 3.4 partes defectuosas por millón producidas, que promulga el nivel Seis Sigma; por lo que existen amplias necesidades de mejora para aspirar a una calidad de clase mundial.

Respecto a los indicadores de medios-continuos, vale ilustrar el tamaño de lote, el cual fluctúa entre 0.1 unidades de mil (100ud) y 510 mil, con un coeficiente de variación de 138% (*100 x DE/Me*). En alusión a indicadores de medios – discretos (en este caso todos binarios), el 16.5% de los registros de producción se debe a la operación 1, el 53.7% a las operaciones 2 o 3, y el restante 29.8% a la operación 4 (ocurre cuando las operaciones 1-3 toman valor cero).

Como se describió en el caso 1, el analista podrá encontrar en el reporte información útil sobre el análisis bivariado. Por efectos de extensión se pasa directamente a las salidas multivariadas. La Figura 9 muestra pantallazos de resultados sobre el proceso de patrones de agrupación.

En la Figura 9 se muestra que, la formación de los dos grupos de desempeño (análisis clúster) empleando las variables de respuesta originales (continuas), previamente estandarizadas, llevó configurar el grupo “0” con 656 registros de producción y el grupo “1” con 55 registros (*g.desemp*). No obstante, la ejecución para las variables respuesta ya dicotomizadas, proporcionó una configuración más balanceada que la anterior (*g.desemp.dico*), con 298 registros en el grupo “0” y 413 en el grupo “1”. La cuestión ahora es ¿cuál de los grupos (“0” o “1”) infiere mejor desempeño que el otro?

Al observar las medias de los indicadores de respuesta originales (continuos) según grupos, en la Figura 9 se reflejan dos situaciones en conflicto, dado que, dentro de cada grupo, un menor tiempo

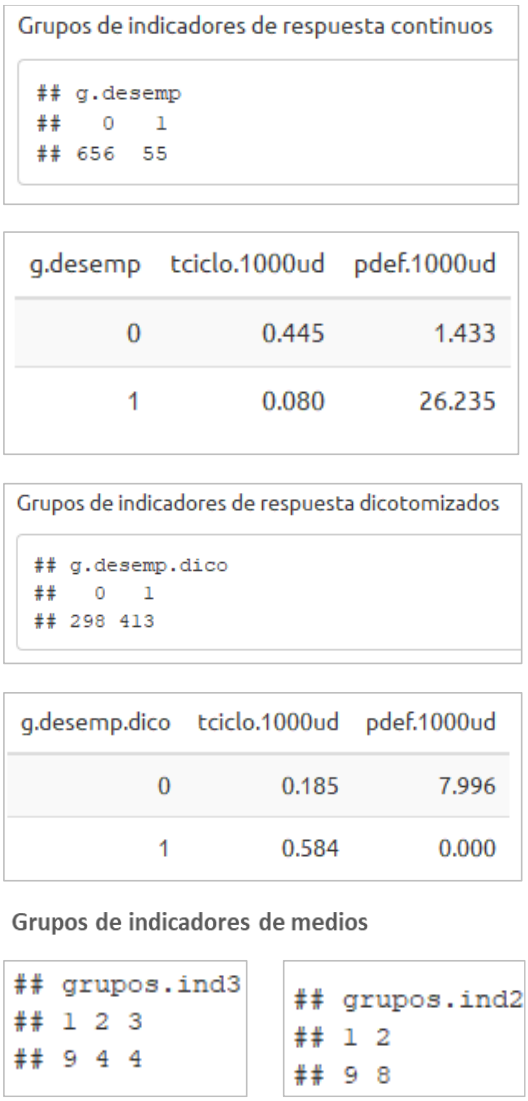


Fig. 9. Salidas automáticas de reporte *html* bajo *Rmarkdown* para descripción multivariada – patrones de agrupación, caso 2

de ciclo (*tciclo.1000ud*) está acompañado por una mayor tasa de defectos (*pdef.1000ud*) y viceversa.

En otras palabras, estas métricas se muestran antagónicas dentro de cada grupo, lo cual ocurre en situaciones donde los esfuerzos por eficiencia desmedida conducen al detrimento de la calidad. Entonces, este comportamiento contrario, visto tanto para las agrupaciones basadas en *g.desemp* como en *g.desemp.dico*, en lugar de retratar alto o bajo desempeño puede interpretarse como el perfil

de producción que se adopte; en uno de ellos prima la eficiencia (grupo "0"; *g.desemp.dico*) y en el otro la calidad (grupo "1"). Sin embargo, la eficiencia a la que se hace alusión es una eficiencia miope, ya que, en este caso, la velocidad de producción (menor *tciclo.1000ud*) está en asocio con más defectos. Dicho de otro modo, la estimación de un *tciclo.1000ud* efectivo resulta mayor al monitoreado en la organización, al requerirse horas adicionales para producir unidades conformes.

Al enfatizar sobre patrones de agrupación de los 17 indicadores de medios-dicotomizados, en la Figura 9 también se detalla la conformación de indicadores para dos y tres conglomerados. Por ejemplo, el reporte para dos grupos (*grupos.ind2*, parte inferior derecha de la Figura 9), arrojó nueve indicadores en el grupo 1 y ocho en el grupo 2.

El analista puede profundizar en las diferentes salidas arrojadas por el despliegue de la metodología, con el fin de facilitar los procesos de exploración de conocimiento. Por ejemplo, considerar resultados de la reducción de dimensiones y de entrenamiento de modelos de regresión/árbol (tareas desplegadas en la sección 2.4 y para las que se mostró aplicación en el primer caso de estudio).

Considérese ahora, en la Figura 10, los resultados de la validación de los modelos supervisados (regresión y árbol) en la muestra no-entrenamiento (30% de los registros).

Véase, en la Figura 10, diferentes alterativas de desempeño de los modelos en prueba. Respecto a regresión logística, los resultados son prometedores y similares desde una mirada general, con leves diferencias en la matriz de confusión. Por ejemplo, al considerar la tasa de acierto global (TA), el mejor resultado lo obtuvo la regresión logística basada netamente en indicadores (sin componentes principales) (*mod.indicad;* TA: 86.4%), el cual resultó ser el mismo *mod.g.f.ultra.*

En referencia a modelos basados en árboles, no hubo diferencias en la tasa de aciertos global (TA: 78.5% de los casos); sin embargo, sobresale el modelo *tri.g.f.dico* (coincidente con *tri.g.f.ultra.dico)*, al presentar valores más estables en las demás métricas (TAN: 70.8% y TAP: 84.0%).

modelo	umbral.1	TA	TAN	TAP
mod.g.f	0.077	0.860	0.872	0.722
mod.indicad	0.077	0.864	0.883	0.667
mod.g.f.ultra	0.077	0.864	0.883	0.667
mod.g.f.dico	0.581	0.785	0.708	0.840
mod.indicad.dico	0.581	0.785	0.685	0.856
mod.g.f.ultra.dico	0.581	0.785	0.730	0.824
tree.g.f				
tree.indicad				
tree.g.f.ultra				
tree.g.f.dico		0.785	0.708	0.840
tree.indicad.dico		0.785	0.607	0.912
tree.g.f.ultra.dico		0.785	0.708	0.840

Fig. 10. Salidas automáticas de validación de modelos para el caso 2

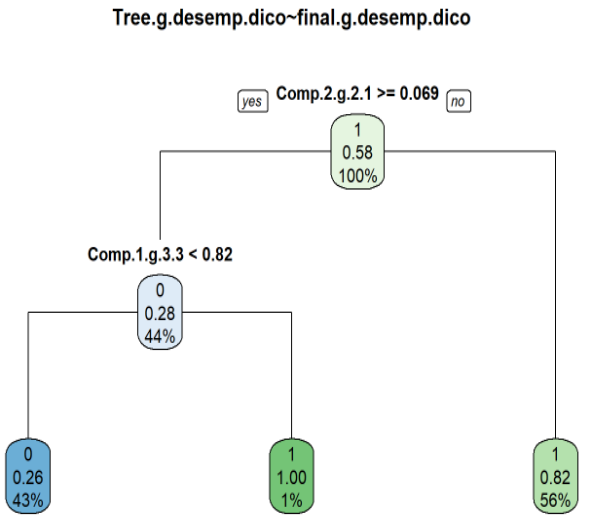


Fig. 11. Salidas automáticas para el entrenamiento de árbol de clasificación para *g.desemp.dico* (*tree.g.f.dico*). Caso 2

ISSN 2007-9737

El analista-usuario de la metodología puede profundizar en cada uno de los modelos obtenidos, todos con tasas razonables de acierto, con miras a proceder con tareas de interpretación, discusión y exploración de conocimiento. En esta oportunidad, por cuestiones de parsimonia de los modelos, se discutirá sobre el modelo *tri.g.f.dico*, cuyo árbol se presenta en la Figura 11.

El árbol presentado en la Figura 11 consta de dos predictores. El principal es el *score* de la segunda componente (*Comp.2*) del grupo 1 de indicadores de medios-dicotomizados, cuando se fijaron dos grupos en el Análisis Clúster (véase la composición del grupo en Figura 9, parte inferior derecha). El predictor segundario es el *score* de la primera componente principal (Comp.1) del grupo g3.3. Se destaca que con la primera partición (*score de Comp.2.g.2.1>= 0.069*), el 56% de la muestra de entrenamiento fue clasificado en la categoría "1" (enfoque hacia la calidad). Allí, el 82% de las observaciones coincide con esta categoría. Respecto al 44% restante, la segunda partición (*score de Comp.1.g.3.3<0.82*) culmina la segmentación. Para profundizar en el contenido de los *scores* mencionados, tome en cuenta la Figura 12.

La componente 2 (*Comp.2*), señalada en la parte superior de la Figura 12, muestra los indicadores de medios que la conforman (operaciones 1 y 3). En vista de que estas variables son originalmente dicotómicas (0,1) la partición con base en los cuartiles 1 y 2 coincide.

En otras palabras, lo que induce el *score* de *Comp.2* es la presencia de la operación 1 o 3, vs la de las operaciones 2 o 4, lo que puede asociarse con la intensidad de la operación responsable. Este indicador latente (*Comp.2*) tenderá a aumentar, en la medida en que la producción de unidades pase por las operaciones 1 o 3, y tenderá a disminuir en caso contrario (2 o 4).

Respecto a *Comp.1.g.3.3*, sus *scores* están en función del peso-dicotomizado del producto, considerando los cuartiles 1 y 2 (20mg y 50mg, respectivamente; véase estadísticos descriptivos en Figura 8). En la medida en que el peso vaya superando dichos cuartiles de segmentación, el *score* de *Comp.1.g.3.3* aumentará, por lo que este puede entenderse como una métrica latente de la masa del producto. Para una mejor comprensión los *scores* de interés y de su relación con los

Comp.princ. grupo g2.1

	Comp.1	Comp.2
operaci1.Q1	0.43	0.39
operaci3.Q1	-0.47	0.52
operaci1.Q2	0.43	0.39
operaci3.Q2	-0.47	0.52
operaci1.Q3	0.43	0.39
operaci2.Q3		
operaci3.Q3		
dia.sem.mo		
peso.contiQ3	0.12	

Comp.princ. grupo g3.3

	Comp.1
cierre.sem.mo	
peso.contiQ1	-0.66
tam.lote.1000ud.contiQ1	
peso.contiQ2	-0.74

Fig. 12. Extractos de salidas automáticas de reducción de dimensiones

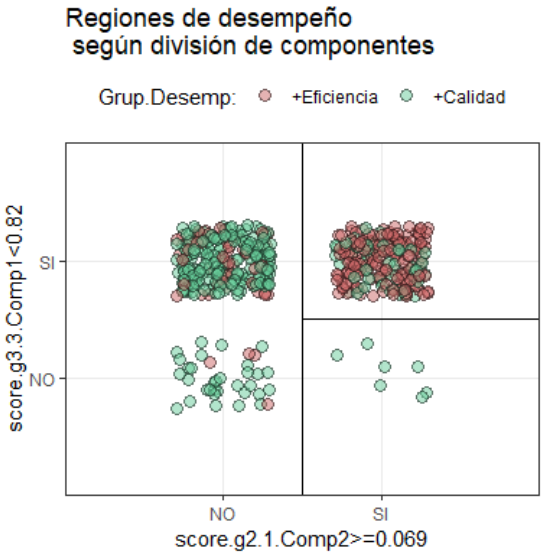


Fig. 13. Regiones delimitadas por la segmentación de los *scores* de las componentes predictoras, caso 2

ISSN 2007-9737

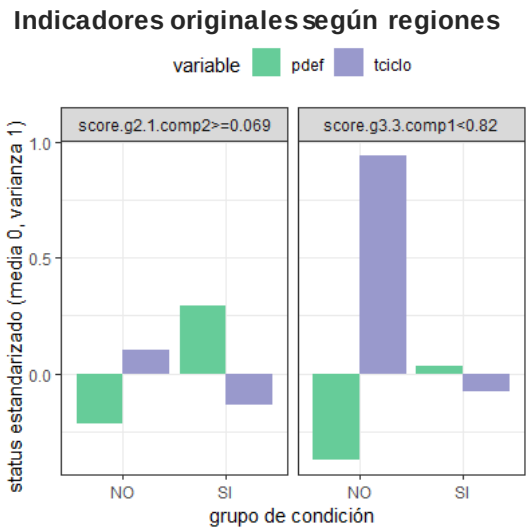


Fig. 14. Valores medios de indicadores de respuesta originales (estandarizados), según regiones delimitadas por el árbol entrenado

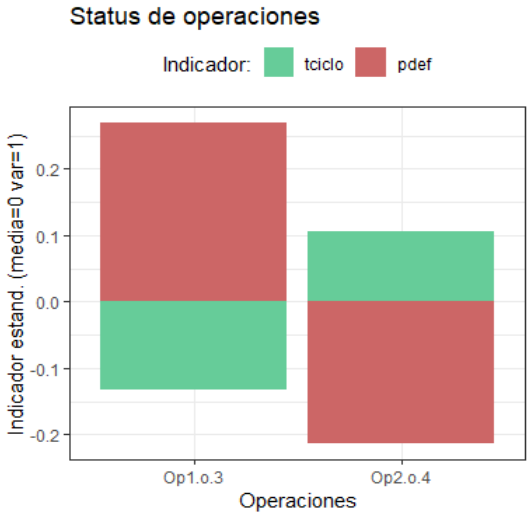


Fig. 15. Valores medios de los indicadores de respuesta originales (estandarizados), según tipo de operación responsable

enfoques de producción ("0": eficiencia, "1": calidad), analícese la Figura 13.

La Figura 13 refleja lo razonable de la clasificación dada por el árbol en la muestra de entrenamiento. Se destaca que cuando el *score* de la componente 2 (intensidad de la operación responsable) supera 0.069, incluso de manera visual se logra distinguir los enfoques de producción hacia la calidad (región izquierda, Figura 13) y hacia la eficiencia (región derecha). Además, en la región derecha, que en su mayoría reúne registros donde ha primado la eficiencia, el modelo ejecuta una segunda división, según la intensidad de la masa del producto (indicador latente: scores *Comp.1.g.3.3<0.82*).

Mediante esta subdivisión, se logra distinguir una situación en la que ocurren unos pocos casos (siete) donde prima la calidad en vez de la eficiencia (cuadrante inferior derecho, Figura 13). Estas regiones motivan a analizar el comportamiento de los indicadores de respuesta originales (continuos), según cada partición, lo cual se plasma en la Figura 14.

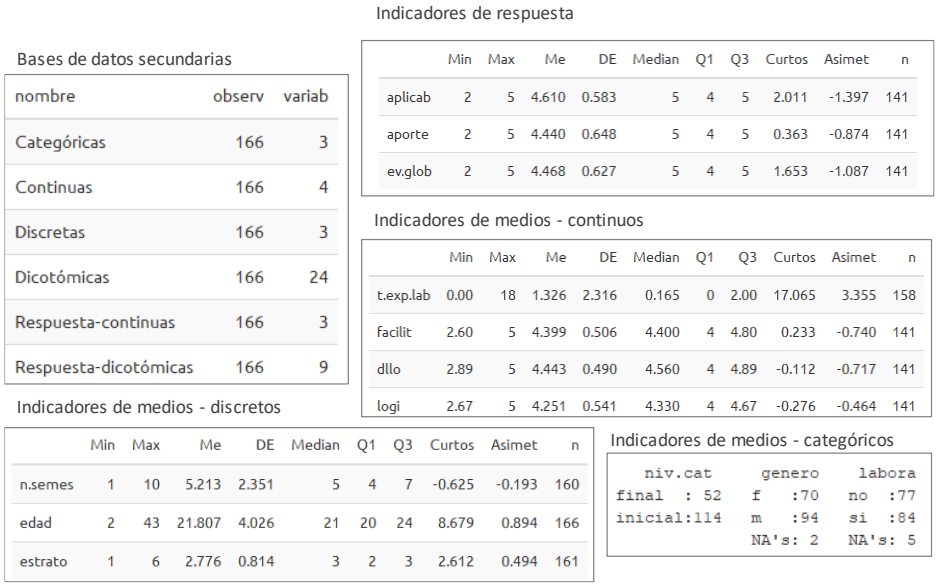
En cuando a la primera partición (*score.g2.1.comp.2*>=0.069; gráfico izquierdo, Figura 14), sobresale el antagonismo que se ha venido interpretando para los enfoques de producción, al encontrarse valores estandarizados (media cero, varianza uno) contrarios para los indicadores de respuesta continuos (cuando *pdef.1000ud* es positivo, *tciclo.1000ud* es negativo, y viceversa).

Adicionalmente, al detallar en el *score* de *comp.2.g2.1* (intensidad de la operación responsable), cuando este no satisface la condición inducida por el árbol, se nota un enfoque de primacía de calidad (*pdef.1000ud_estand.*<0) en vez de eficiencia (*tciclo.1000ud_estand.*>0); en cambio, cuando sí se satisface la condición, se invierte el enfoque que prima (*pdef.1000ud_estand.* > 0; *tciclo.1000ud_estand.* < 0). Lo mismo puede verse en la componente 1 del grupo g.3.3 (gráfico derecho, Figura 14).

Véase, ahora, en la Figura 15, los valores medios de los indicadores de respuesta originales (continuos; previa estandarización bajo media cero y varianza 1), pero esta vez según el tipo de operación.

La Figura 15 es consistente con lo que se ha inducido sobre los enfoques de producción, pues en ella sobresale también los objetivos en conflicto entre disminuir defectos (énfasis en calidad) o disminuir tiempos de ciclo de forma desmedida (eficiencia – "ciega"). Cabe destacar que ante el escenario de producción bajo las operaciones 1 o 3, tiende a primar el enfoque hacia la eficiencia, en vez de hacia la calidad.

ISSN 2007-9737



Bases de datos secundarias

nombre	observ	variab
Categóricas	166	3
Continuas	166	4
Discretas	166	3
Dicotómicas	166	24
Respuesta-continuas	166	3
Respuesta-dicotómicas	166	9

Indicadores de respuesta

	Min	Max	Me	DE	Median	Q1	Q3	Curtos	Asimet	n
aplicab	2	5	4.610	0.583	5	4	5	2.011	-1.397	141
aporte	2	5	4.440	0.648	5	4	5	0.363	-0.874	141
ev.glob	2	5	4.468	0.627	5	4	5	1.653	-1.087	141

Indicadores de medios - continuos

	Min	Max	Me	DE	Median	Q1	Q3	Curtos	Asimet	n
t.exp.lab	0.00	18	1.326	2.316	0.165	0	2.00	17.065	3.355	158
facilit	2.60	5	4.399	0.506	4.400	4	4.80	0.233	-0.740	141
dllo	2.89	5	4.443	0.490	4.560	4	4.89	-0.112	-0.717	141
logi	2.67	5	4.251	0.541	4.330	4	4.67	-0.276	-0.464	141

Indicadores de medios - discretos

	Min	Max	Me	DE	Median	Q1	Q3	Curtos	Asimet	n
n.semes	1	10	5.213	2.351	5	4	7	-0.625	-0.193	160
edad	2	43	21.807	4.026	21	20	24	8.679	0.894	166
estrato	1	6	2.776	0.814	3	2	3	2.612	0.494	161

Indicadores de medios - categóricos

```
   niv.cat      genero      labora
final  : 52   f    :70   no  :77
inicial:114   m    :94   si  :84
              NA's: 2    NA's: 5
```

Fig. 16. Extractos del reporte HTML bajo *Rmarkdown* para la descripción univariada, caso

modelo	umbral.1	TA	TAN	TAP
mod.g.f	0.045	0.75	0.789	0.000
mod.indicad				
mod.g.f.ultra	0.045	0.75	0.789	0.000
mod.g.f.dico	0.519	0.70	0.571	0.842
mod.indicad.dico	0.519	0.60	0.429	0.789
mod.g.f.ultra.dico	0.519	0.70	0.571	0.842
tree.g.f				
tree.indicad				
tree.g.f.ultra				
tree.g.f.dico		0.50	0.286	0.737
tree.indicad.dico		0.50	0.238	0.789
tree.g.f.ultra.dico		0.55	0.524	0.579

Fig. 17. Salidas automáticas de validación de modelos para el caso 3

En cambio, en las operaciones 2 o 4 el fenómeno es contrario. Estos hallazgos son solamente unos de tantas posibilidades de interpretación, comprensión del proceso y exploración de conocimiento, a partir de los reportes automáticos que se derivan del despliegue de la metodología propuesta. Sin embargo, reflejan varias oportunidades de mejora del proceso en cuestión, que van desde la necesidad de balancear los tiempos de operaciones; identificar las causas de que a las operaciones 1 o 3 se les atribuya alta tasa de

Grupos de desempeño

g.desemp.dico	aplicab	aporte	ev.glob
0	4.312	3.844	4.031
1	4.884	5.000	4.855

Modelo de regresión logística elegido

	mod.g.f.dico
	g.desemp.dico
(Intercept)	0.536
	(0.399)
Comp.1.g.2.2	2.804***
	(0.625)
Comp.7.g.2.2	4.948**
	(1.709)
Comp.2.g.3.1	−1.472*
	(0.634)
Comp.7.g.3.2	5.010**
	(1.786)
Aldrich-Nelson R-sq.	0.441
McFadden R-sq.	0.572
Likelihood-ratio	73.499
p	0.000
Log-likelihood	−27.449
Deviance	54.899
AIC	64.899
N	93

Fig. 18. Grupos de desempeño y modelo de clasificación elegidos

Ind.medios g.2.2

	Comp.1	Comp.7
niv.cat.mo	0.19	
facilit.contiQ1	0.40	-0.72
dllo.contiQ1	0.34	0.68
logi.contiQ1	0.43	
facilit.contiQ2	0.45	
dllo.contiQ2	0.42	
logi.contiQ2	0.35	0.10

Ind.medios g.3.1

	Comp.2
n.semes.Q1	0.44
edad.Q1	0.26
estrato.Q1	
n.semes.Q2	0.52
edad.Q2	0.25
genero.mo	0.22
labora.mo	-0.35
t.exp.lab.contiQ1	-0.35
t.exp.lab.contiQ2	-0.32

Ind.medios g.3.2

	Comp.7
estrato.Q2	0.15
n.semes.Q3	-0.20
edad.Q3	
estrato.Q3	0.15
t.exp.lab.contiQ3	
facilit.contiQ3	-0.66
dllo.contiQ3	
logi.contiQ3	0.69

Fig. 19. Contenido de las componentes principales predictoras de *g.desemp.dico*

defectos, así como de que las operaciones 2 o 4 presenten altos tiempos de procesamiento; cuantificar el impacto de los dos enfoques de producción inducidos; unificar el tipo de enfoque que debe primar en la organización; profundizar en las implicaciones del peso del producto; entre muchas otras cuestiones que motiva el reporte completo.

3.3. Caso 3: satisfacción con la capacitación

Se trata de indicadores de una encuesta a estudiantes universitarios, que busca evaluar el grado de satisfacción con una capacitación impartida. La base de datos consta de trece indicadores, seis de los cuales son perceptuales, medidos entre 1 (puntuación más desfavorable) y 5 (puntuación más favorable). Los indicadores de respuesta (perceptuales) son: aplicabilidad de los conceptos empleados (*aplicab*), aporte de los contenidos a la vida personal (*aporte*) y evaluación global de la sesión (*ev.glob*). Los indicadores de medios son: número de semestres cursados por el estudiante (*n.semes*); nivel categorizado (*niv.cat*; "inicial": 1 a 6 semestres, "final": 7 a 10 semestres); *edad* (en años); *estrato* (socioeconómico); *género*; evento de si está laborando o no (*labora*); años de experiencia laboral (*t.exp.lab*); puntuaciones promedio de facilitador (*facilit*); de desarrollo de la sesión (*dllo*) y de logística de la sesión (*logi*).

En la Figura 16 se aportan algunos de los descriptivos univariados para los indicadores descritos. Por criterios de extensión, se pasa

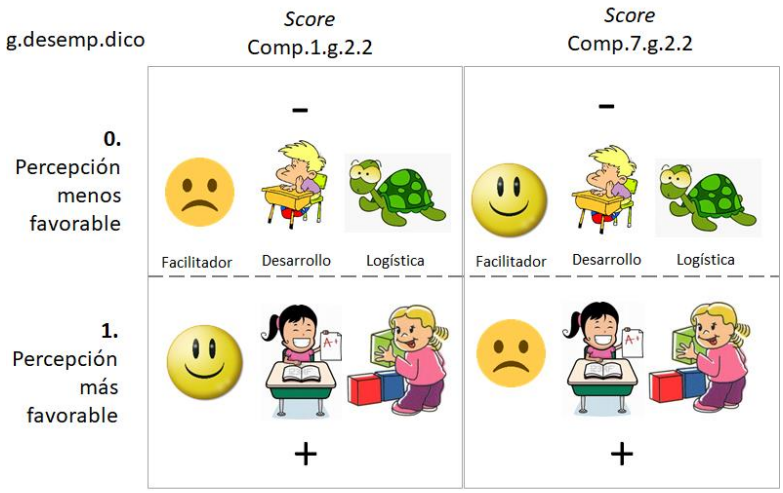


Fig. 20. Representación gráfica-pragmática de los *scores* de las componentes predictoras en asocio con *g.desemp.dico*. Imágenes obtenidas por medio "imágenes en línea" de Microsoft Office – Word (2016) ®, *Bing* ®, etiquetadas *creative commons*

directamente a los resultados de validación de los modelos, expuestos en la Figura 17.

En la Figura 17 se observa que el modelo más prometedor (*mod.g.f.dico*, 70% tasa de acierto global: TA; 57.1% TAN y 84.2% TAP) no pertenece a la familia de árboles sino a la de regresión logística. De hecho, el modelo que se construye con los predictores significativos en modelos previos (*mod.g.f.ultra.dico*) no encontró una mejor alternativa, por lo que se limitó a reproducirlo.

Al observar la conformación de grupos de desempeño, la agrupación con base en las variables respuestas dicotomizadas (g*.desemp.dico*) condujo a grupos prácticamente balanceados (grupo "0": 64 obs.; grupo "1": 69 obs.). A su vez, en la Figura 18, parte superior, puede verse que el grupo "1" de observaciones induce un mejor desempeño que el grupo "0", dado que en el primero se presentan puntuaciones medias más elevadas.

Al enfatizar en el modelo de clasificación que busca predecir dichos grupos de desempeño, sobresalen cuatro *scores* regresores. Por ejemplo, *Comp.1.g.2.2* es uno de ellos, y denota el *score* de la primera componente principal, obtenida para el segundo grupo de indicadores de medios (previo análisis de componentes principales; véase aspectos metodológicos en sección 2.4, *reducción de dimensiones*). La Figura 19 facilita profundizar en el contenido de los predictores de interés.

La Figura 19 muestra los indicadores que definen cada predictor significativo en el modelo de clasificación.

Los *scores* de las componentes 1 (de g.2.2) y 7 (de g.2.2 y g.3.2) inducen tipos de escenarios de capacitación; mientras que los *scores* de la componente 2 (g.3.1), pueden interpretarse como perfiles de estudiantes.

Por ejemplo, en la medida en que el *score* de *Comp.1.g.2.2* aumente, denota la presencia de una capacitación favorable en calidad del facilitador (*facilit.conti*), desarrollo de la sesión (*dllo*), de las tareas logísticas al respecto (*logi*) y que está dirigida a estudiantes de semestres iniciales (1 a 6) (niv.cat.mo; moda es "inicial"). Cuando tal indicador latente (*score*) se hace menor, tiende a representar lo contrario. Con relación a *Comp.7.g.2.2* (parte superior, Figura 19), se encuentra otro tipo de escenario de capacitación, cuyo aumento en el *score* puede asociarse con un docente percibido favorablemente por el estudiante, en cuanto a conocimientos, dominio, entusiasmo,…, pero el despliegue de los contenidos (*dllo*) tiende a ser desfavorable, por falencias en orden de los temas, participación, logro de los objetivos,…, así como en la logística implicada (*logi*).

Para este escenario aplican frases como: "se nota que el profesor sabe mucho, pero no se le entiende", o "buen profesor, pero en este entorno

difícilmente se aprende". Por otro lado, en la medida en que el *score* de esta componente (*Comp.7.g.2.2*) disminuya, induce un escenario de capacitación con desarrollo de contenidos y logística favorables, pero con un docente que refleja inseguridad, poco entusiasmo, dominio, etc. Un resumen gráfico-pragmático de las interpretaciones expresas, se plasma en la Figura 20.

Respecto a los *scores* de *Comp.7.g.2.2,* mientras más positivos sean, se induce un tipo de estudiante hombre (*genero.mo;* moda: género masculino), de más edad y avance en la carrera (*n.semes*), que no labora (signo negativo; moda: "si") o que tiene poca experiencia laboral (signo negativo; *t.ex.lab*). Mientras más bajo se torne tal indicador latente (*scores*), representará lo contrario. Pasando a los *scores* de *Comp.7.g.3.2*, emerge otro tipo de escenario de capacitación, que contempla aspectos del estudiante (*estrato* y *n.semes*), del facilitador (*facilit*) y de la logística del evento (*logi*).

3.4. Caso 4: desempeño operativo

Este conjunto de datos consta de 951 registros de producción diaria de una empresa manufactura. Los indicadores de respuesta son: proporción de generación de defectos (*p.defec*), de no generación de valor (*p.anv*) y eficiencia *(productos/unidad de tiempo*).

Para este caso 4 se discute únicamente los reportes de los modelos finales, pues en casos previos se ha ejemplificado los demás procesos. En esa vía, los 21 indicadores originales de medios, luego son transformados automáticamente, dando lugar a 53 indicadores dicotomizados (ver, en el anexo 1, un extracto de los reportes automáticos sobre este caso 4).

Yendo a los reportes finales de la metodología para este caso 4, la Tabla 8 ofrece los resultados comparativos entre los modelos. En ella sobresalen dos modelos como los más prometedores: *mod.g.f.ultra.dico* (regresión) y *tree.indicad.dico* (árbol), con resultados idénticos en la matriz de confusión (TA: 72.4%, TAN: 71.6% y TAP: 73.3%).

Considerando el principio de la parsimonia, se elegirá el modelo basado en árbol, que consta solo del indicador de *5S* (dicotomizado con base en cuartil 2). Las *5S* son una filosofía y herramienta de mejora empleada con frecuencia en entornos

Tabla 8. Salidas automáticas de validación de modelos para el caso 4

Modelos	TA	TAN	TAP
1. mod.g.f	0.69	0.68	0.75
2. mod.indicad	0.64	0.62	0.90
3. mod.g.f.ultra	0.69	0.68	0.75
4. mod.g.f.dico	0.66	0.62	0.71
5. mod.indicad.dico	0.68	0.63	0.74
6. mod.g.f.ultra.dico	**0.72**	**0.72**	**0.73**
10. tree.g.f.dico	0.68	0.66	0.71
11. tree.indicad.dico	**0.72**	**0.72**	**0.73**
12. tree.g.f.ultra.dico	0.67	0.63	0.72

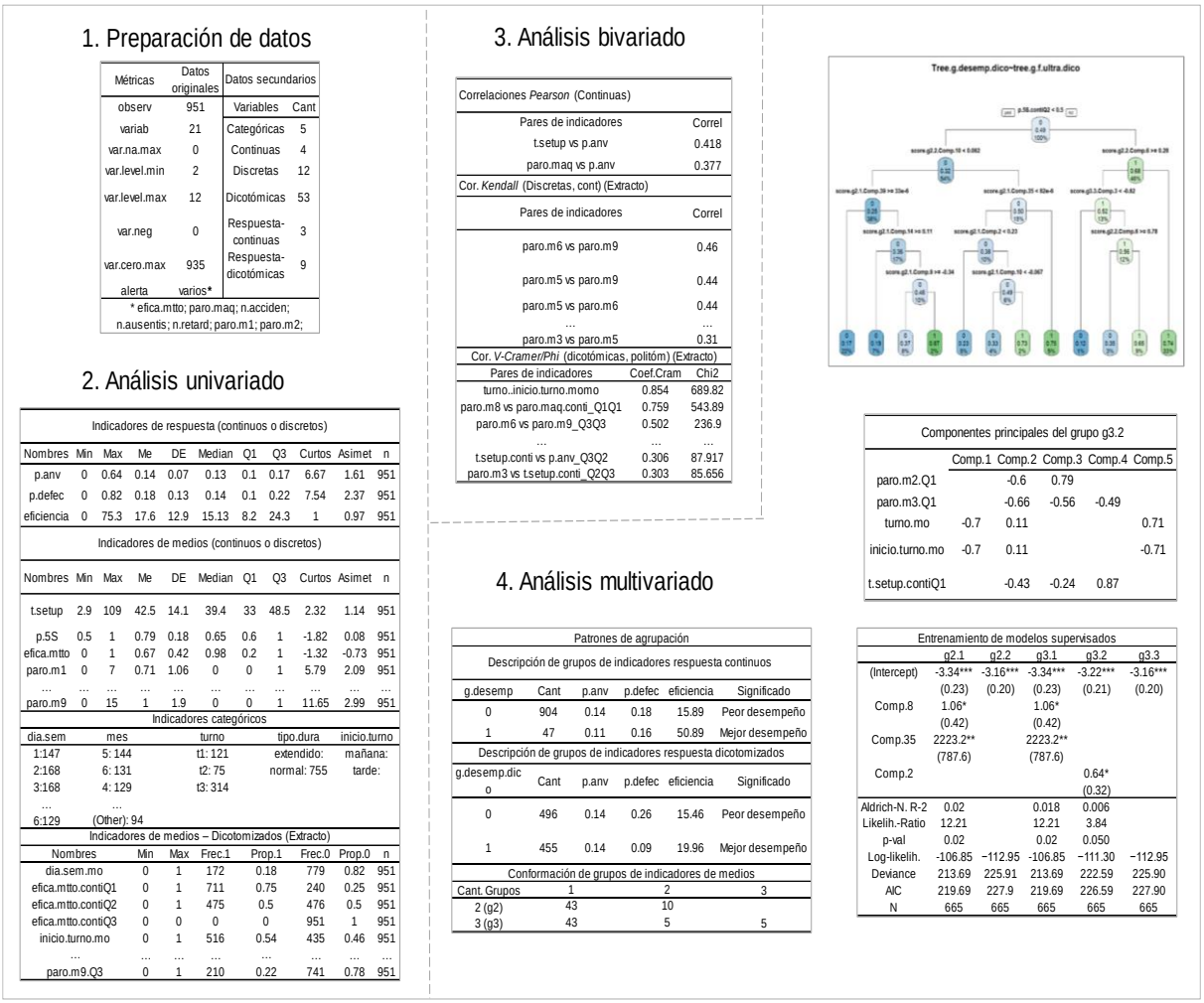
TA: Tasa de aciertos; TAN: Tasa de Aciertos Negativos; Tasa de Aciertos Positivos

Tabla 9. Contraste de medias para niveles del predictor del *desempeño.dico* (derivado de árbol), caso 4, respecto a indicadores de respuesta

p.5S.contiQ2	0. Bajo desempeño	1. Alto desempeño
Medias:		
p.anv	0.15	**0.13**
p.defec	0.21	**0.14**
eficiencia	14.75	**20.83**
Manova:		
Pillai	0.12071	
Pr(>F)	**< 2.20E-16*** **	

Tabla 10. Descripción del consumo computacional

Casos	Obs	Variab origin	Variab secund	Durac. (min)
1. Satisfac. e imagen	120	15	52	0.34
2. Eficiencia y calidad	711	9	32	0.38
3. Satisfac. capacitación	166	13	46	0.34
4. Desemp. operativo	951	24	86	2.53



1. Preparación de datos

Métricas	Datos originales	Datos secundarios	
observ	951	Variables	Cant
variab	21	Categóricas	5
var.na.max	0	Continuas	4
var.level.min	2	Discretas	12
var.level.max	12	Dicotómicas	53
var.neg	0	Respuesta-continuas	3
var.cero.max	935	Respuesta-dicotómicas	9
alerta	varios*		

* efica.mtto; paro.maq; n.acciden; n.ausentis; n.retard; paro.m1; paro.m2;

2. Análisis univariado

Indicadores de respuesta (continuos o discretos)

Nombres	Min	Max	Me	DE	Median	Q1	Q3	Curtos	Asimet	n
p.anv	0	0.64	0.14	0.07	0.13	0.1	0.17	6.67	1.61	951
p.defec	0	0.82	0.18	0.13	0.14	0.1	0.22	7.54	2.37	951
eficiencia	0	75.3	17.6	12.9	15.13	8.2	24.3	1	0.97	951

Indicadores de medios (continuos o discretos)

Nombres	Min	Max	Me	DE	Median	Q1	Q3	Curtos	Asimet	n
t.setup	2.9	109	42.5	14.1	39.4	33	48.5	2.32	1.14	951
p.5S	0.5	1	0.79	0.18	0.65	0.6	1	-1.82	0.08	951
efica.mtto	0	1	0.67	0.42	0.98	0.2	1	-1.32	-0.73	951
paro.m1	0	7	0.71	1.06	0	0	1	5.79	2.09	951
…	…	…	…	…	…	…	…	…	…	…
paro.m9	0	15	1	1.9	0	0	1	11.65	2.99	951

Indicadores categóricos

dia.sem	mes	turno	tipo.dura	inicio.turno
1:147	5: 144	t1: 121	extendido:	mañana:
2:168	6: 131	t2: 75	normal: 755	tarde:
3:168	4: 129	t3: 314		
…	…			
6:129	(Other): 94			

Indicadores de medios – Dicotomizados (Extracto)

Nombres	Min	Max	Frec.1	Prop.1	Frec.0	Prop.0	n
dia.sem.mo	0	1	172	0.18	779	0.82	951
efica.mtto.contiQ1	0	1	711	0.75	240	0.25	951
efica.mtto.contiQ2	0	1	475	0.5	476	0.5	951
efica.mtto.contiQ3	0	0	0	0	951	1	951
inicio.turno.mo	0	1	516	0.54	435	0.46	951
…	…	…	…	…	…	…	…
paro.m9.Q3	0	1	210	0.22	741	0.78	951

3. Análisis bivariado

Correlaciones *Pearson* (Continuas)

Pares de indicadores	Correl
t.setup vs p.anv	0.418
paro.maq vs p.anv	0.377

Cor. *Kendall* (Discretas, cont) (Extracto)

Pares de indicadores	Correl
paro.m6 vs paro.m9	0.46
paro.m5 vs paro.m9	0.44
paro.m5 vs paro.m6	0.44
…	…
paro.m3 vs paro.m5	0.31

Cor. *V-Cramer/Phi* (dicotómicas, politóm) (Extracto)

Pares de indicadores	Coef.Cram	Chi2
turno..inicio.turno.momo	0.854	689.82
paro.m8 vs paro.maq.conti_Q1Q1	0.759	543.89
paro.m6 vs paro.m9_Q3Q3	0.502	236.9
…	…	…
t.setup.conti vs p.anv_Q3Q2	0.306	87.917
paro.m3 vs t.setup.conti_Q2Q3	0.303	85.656

Componentes principales del grupo g3.2

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
paro.m2.Q1		-0.6	0.79		
paro.m3.Q1		-0.66	-0.56	-0.49	
turno.mo	-0.7	0.11			0.71
inicio.turno.mo	-0.7	0.11			-0.71
t.setup.contiQ1		-0.43	-0.24	0.87	

4. Análisis multivariado

Patrones de agrupación

Descripción de grupos de indicadores respuesta continuos

g.desemp	Cant	p.anv	p.defec	eficiencia	Significado
0	904	0.14	0.18	15.89	Peor desempeño
1	47	0.11	0.16	50.89	Mejor desempeño

Descripción de grupos de indicadores respuesta dicotomizados

g.desemp.dico	Cant	p.anv	p.defec	eficiencia	Significado
0	496	0.14	0.26	15.46	Peor desempeño
1	455	0.14	0.09	19.96	Mejor desempeño

Conformación de grupos de indicadores de medios

Cant. Grupos	1	2	3
2 (g2)	43	10	
3 (g3)	43	5	5

Entrenamiento de modelos supervisados

	g2.1	g2.2	g3.1	g3.2	g3.3
(Intercept)	-3.34***	-3.16***	-3.34***	-3.22***	-3.16***
	(0.23)	(0.20)	(0.23)	(0.21)	(0.20)
Comp.8	1.06*		1.06*		
	(0.42)		(0.42)		
Comp.35	2223.2**		2223.2**		
	(787.6)		(787.6)		
Comp.2				0.64*	
				(0.32)	
Aldrich-N. R-2	0.02		0.018	0.006	
Likelih.-Ratio	12.21		12.21	3.84	
p-val	0.02		0.02	0.050	
Log-likelih.	-106.85	−112.95	-106.85	−111.30	−112.95
Deviance	213.69	225.91	213.69	222.59	225.90
AIC	219.69	227.9	219.69	226.59	227.90
N	665	665	665	665	665

Fig. 21. Algunos resultados del uso de la metodología en el caso 4

de producción, para sembrar el hábito de mantener el lugar de trabajo con lo necesario, ordenado, limpio, estandarizado y seguro, bajo principios de mejora continua. Mucho se habla de las bondades de las *5S*, las cuales son relativamente simples de instaurar, pero complejas de conservar en el tiempo. Una de las estrategias es mediante la auditoría de 5S, que en este caso se puntúa en proporción de cumplimiento.

En la empresa en estudio, por ejemplo, el porcentaje de *5S* fue de 79% (media), con una mediana de 65% (Q2). En esta oportunidad, los resultados arrojados por la metodología destacan el valor de las *5S* como predictor potencial del desempeño. Para profundizar en la relación entre grado de complimiento de las *5S* (dicotomizada según Q2) y los indicadores de respuesta: *p.anv* (proporción de tiempo en que no se agrega valor), *p.defec* (porcentaje de defectos) y eficiencia (productos/hora), véase la Tabla 9.

Cuando el porcentaje de cumplimiento de las *5S* fue mayor al cuartil 2 (grupo “1”), también se encontró un mejor desempeño del sistema: *p.def* y *p.anv* fueron menores en comparación a cuando tal cumplimiento fue menor de Q2 (grupo “0”). Asimismo, la eficiencia cambió de 14.75 u/h a 20.8 u/h al pasar de bajo (grupo “0”) a alto (grupo “1”) cumplimiento de *5S*. Al explorar la veracidad de

tales diferencias usando métodos adicionales, como *Manova*, se halló razones estadísticamente significativas (valor-p casi nulo; Tabla 9). De nuevo, estos hallazgos son una muestra de la utilidad de la metodología.

3.5. Duración del despliegue de la metodología

En la Tabla 10 se presenta la duración media de ejecutar, en cada caso, tres veces la metodología propuesta, usando un computador portátil *Intel® Core™ i5-5200 CPU de 2.20Ghz y 8GB*.

De la Tabla 10 vale mencionar las siguientes estimaciones para la región específica de ensayo: la duración por observación osciló entre 0.03 segundos y 0.17 segundos (*Durac/Obs*); por variable original estuvo entre 1.36 seg y 6.33 seg (*Durac/Variab_origin*); y por variable secundaria entre 0.39 seg y 1.77 seg (*Durac/Variab_secund*). En resumen, en segundos o en pocos minutos, el analista de datos puede obtener automáticamente un reporte en *HTML* de uso articulado de más de 10 métodos de análisis, con alcances univariado, bivariado y multivariado. Esto favorece la diminución de costos y de tiempo operativo, lo cual representa mayor posibilidad de dedicación a tareas intelectuales de interpretación, discusión y transferencia de hallazgos, así como de toma de decisiones. Igualmente, profesionales sin dominio en la ejecución de los métodos empleados, pero sí en la interpretación del reporte automático generado, también pueden encontrar utilidad en la metodología propuesta.

4. Conclusiones

Se aportó una metodología que de forma automática ayuda a preparar, procesar, analizar y visualizar indicadores organizativos usando Ciencia de Datos. Su principal valor es la sistemática multitareas automática que abarca los tres alcances estadísticos (univariado, bivariado y multivariado), articula más de 10 métodos de análisis, brinda respuestas a seis preguntas de interés para el analista y usa software libre. Es decir, contribuye con procesos de orquestación de recursos (analíticos), lo cual es un desafío pragmático, en la medida en que el uso de métodos (recursos) *per se* no asegura ventajas competitivas [11-12].

El procedimiento metodológico se automatizó en *R* (a través de *R-Studio*), con reportes *HTML* por medio de *RMarkdow*. Se probó en cuatro casos, dos de servicios y dos de manufactura, obteniendo resultados útiles para una mejor comprensión de la situación actual del sistema, de sus patrones y de la predicción de posibles grupos de desempeño.

La concepción de desempeño no es restringida en esta metodología, sino que da flexibilidad al analista para que indique lo que ha de considerar como variables respuesta y, a partir de ellas, se configuran los grupos de desempeño. En otras palabras, se toma en cuenta necesidades específicas del analista, lo cual ayuda a superar oportunidades descritas en [7].

Otro elemento que vale nombrar es el tiempo computacional, el cual es reducido en comparación con lo que representaría ejecutar manualmente todas las tareas de la metodología. En pocos minutos o segundos, el analista puede obtener de forma automática un reporte de la ejecución articulada de más de 10 métodos de análisis, cuyos resultados en los casos de prueba, consumieron cerca de 20 páginas por caso.

Con este artículo se espera contribuir a una mayor motivación de las organizaciones para prepararse y aprovechar las potencialidades de la Ciencia de Datos, en la búsqueda incesante del mejoramiento de procesos. Asimismo, estimular el trabajo conjunto empresa – academia para propiciar nuevas soluciones de Ciencia de Datos, guiadas por el dominio de problemas reales en la empresa.

Este estudio también motiva preguntas emergentes para trabajos en pregrado y posgrado: P.1 ¿Cómo favorecer la calidad de los de datos, antes de ingresarlos al sistema? Futuros estudios podrían apoyarse en [6], quienes proponen un método inspirado en TQM y control estadístico de procesos. P.2 ¿Qué efectos se produce al adicionar más de dos niveles de desempeño? Para ello, merecería incluirse también la regresión logística multinomial. P.3 ¿Cómo se desempeña la metodología en otros casos (sectores, mayores tamaños de muestra, etc.)? P.4 ¿Cómo puede la metodología apalancar los procesos de enseñanza-aprendizaje basados en Ciencia de

Datos? Es decir, se gira la mirada hacia el proceso formativo de futuros analistas, encargados del tratamiento de indicadores en las organizaciones. P.5 ¿Cuál es el impacto, en calidad de la información para la toma de decisiones, eficiencia, satisfacción y costos, del uso de la metodología por parte de organizaciones, en comparación con sus tareas convencionales de análisis de datos?

Referencias

1. **Pop, F. (2014).** High performance numerical computing for high energy physics: a new challenge for big data science*. Advances in High Energy Physics*, pp. 1–13.
2. **Davis, J. & Burgoon, L. (2015).** Can data science inform environmental justice and community risk screening for type 2 diabetes? *Plos one*, Vol 10, No. 4, pp. 1–14. DOI:10.1371/journal.pone.0121855.
3. **Ames, D., Quinn, N., & Rizzoli, A. (2014).** Predicting impact of natural calamities in era of big data and data science. international environmental modelling and software society (iEMSs). *7th Intl. Congress on Env. Modelling and Software.*
4. **Schoenherr, T., & Speier-Pero, C. (2015).** Data science, predictive analytics, and big data in supply chain management: current state and future potential. *Journal of Business Logistics*, Vol. 36, No. 1, pp. 120–132. DOI:10.1111/jbl.12082.
5. **Pacheco, F., Rangel, C., Aguilar, J., Cerrada, M., & Altamiranda, J. (2014).** Methodological framework for data processing based on the Data Science paradigm. In Computing Conference (CLEI), XL Latin American, pp. 1–12. DOI: 10.1109/CLEI.2014. 6965184.
6. **Hazen, B., Boone, C., Ezell, J., & Jones-Farmer, L. (2014).** Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International *Journal of Production Economics*, Vol. 154, pp. 72–80. DOI:10.1016/j.ijpe.2014.04.018.
7. **Leung, C. & Jiang, F. (2014).** A data science solution for mining interesting patterns from uncertain big data. *Big data and cloud computing (BdCloud), IEEE Fourth International Conference*, pp. 235–242. DOI: 10.1109/BDCloud.2014.136.
8. **Wu, X., Zhu, X., Wu, G., & Ding, W. (2014).** Data mining with big data. Knowledge and Data Engineering, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 1, pp. 97–107. DOI: 10.1109/TKDE.2013.109.
9. **Peral, J., Maté, A., & Marco, M. (2017).** Application of Data Mining techniques to identify relevant Key Performance Indicators. *Computer Standards & Interfaces*, Vol. 50, pp. 55–64. DOI:10.1016/ j.csi.2016.09.009
10. **Barney, J. (1991).** Firm resources and sustained competitive advantage. *Journal of Management*, Vol. 17, pp. 99–120. DOI: 10.1177/014920639 101700108.
11. **Teece, D., Pisano, G., & Shuen, A. (1997).** Dynamic capabilities and strategic management. *Strategic Management Journal*, Vol. 18, No. 7 pp. 509–533.
12. **Popovič, A., Hackney, R., Tassabehji, R., & Castelli, M. (2016).** The impact of big data analytics on firms' high value business performance. *Information Systems Frontiers*, Vol. 20, pp. 1–14. DOI: 10.1007/s10796-016-9720-4.
13. **Wirth, R. & Hipp, J. (2000).** CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pp. 29–39. DOI:10.1.1.198.5133.
14. **R Project. (2008).** *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Austria.* https://www.R-project.org
15. **R Studio Team (2015).** *RStudio: Integrated Development for R*. RStudio. https://www. rstudio.com/.
16. **Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., Atkins, A., Hyndman, R., & Arslan, R. (2017).** rmarkdown: Dynamic Documents for R. R package version 1.6. https://CRAN.R-project.org/package=rmarkdown
17. **Hu, H., Kandampully, & J., Juwaheer, D. (2009).** Relationships and impacts of service quality, perceived value, customer satisfaction, and image: An empirical study. *Service industries Journal,* Vol. 29, No. 2, pp. 111–125. DOI:10.1080/026 42060802292932.
18. **Chen, C. (2008).** Investigating structural relationships between service quality, perceived value, satisfaction, and behavioral intentions for air passengers: evidence from Taiwan. *Transportation Research Part A: Policy and Practice*, Vol. 42, No. 4, pp. 709–717. DOI:10.1016/j.tra.2008.01.007.

Article received on 30/09/2018; accepted on 16/01/2020. Corresponding author is Jorge Pérez Rave.