La presente obra reune una serie de ensayos elaborados por los estudiantes del doctorado de Ciencias de la Administración del Centro Universitario de Ciencias Económico Administrativas de la Universidad de Guadalajara, basados en lo aprendido en la asignatura de Investigación Cuantitativa I. Estos ensayos se orientan, en principio, a realizar un ejercicio de disertación que contribuya a la elaboración de su tesis o se configure como una contribución a la materia, resaltando la pertinencia de su redacción, conceptualizando y proponiendo los modelos revisados como desarrollo de su disertación. Lo anterior sirve como base para realizar la discusión que permite aclarar la contribución esperada, para finalmente concluir en puntos esenciales que sirvan al lector y al expositor, para estudios posteriores.

Es deseo de la coordinación del presente trabajo, que este contribuya al ánimo del lector por conocer los proyectos que se desarrollan e informar de las oportunidades que se muestran, con el fin de dar seguimiento a la evolución de los mismos en su estancia en el posgrado.

ISBN: 978-607-98782-3-8

Centro Universitario de Ciencias
Económico Administrativas

Juan Mejía Trejo
Coordinador

ANÁLISIS MULTIVARIANTE CON ENFOQUE DEPENDIENTE EN LAS CIENCIAS DE LA ADMINISTRACIÓN COMO BASE PARA LA INNOVACIÓN

ENSAYOS 2018

ENSAYOS 2018
ANÁLISIS MULTIVARIANTE CON ENFOQUE DEPENDIENTE EN LAS CIENCIAS DE LA ADMINISTRACIÓN COMO BASE PARA LA INNOVACIÓN

Juan Mejía Trejo
Coordinador

UNIVERSIDAD DE GUADALAJARA

ENSAYOS 2018

ANÁLISIS MULTIVARIANTE CON ENFOQUE DEPENDIENTE EN LAS CIENCIAS DE LA ADMINISTRACIÓN COMO BASE PARA LA INNOVACIÓN

ENSAYOS 2018

Análisis Multivariante con Enfoque Dependiente en las Ciencias de la Administración como Base para la Innovación

Juan Mejía Trejo
Coordinador

Universidad de Guadalajara
2019

Esta obra fue sometido a un proceso de dictamen por pares de acuerdo con las normas establecidas por el comité editorial del Centro Universitario de Ciencias Económico Administrativas de la Universidad de Guadalajara.

Primera edición 2019

D.R. © Universidad de Guadalajara
Centro Universitario de Ciencias Económico Administrativas
Periférico Nte. 799, núcleo universitario Los Belenes
45100, Zapopan, Jalisco

ISBN: 978-607-98782-3-8

Impreso y hecho en México
Printed and made in Mexico

Contenido

CAPÍTULO 1

Herramientas de análisis multivariante predictivo y minería de datos con SPSS Modeler y Statistics: Estudio comparativo

Pascuala Josefina Cárdenas Salazar

DIRECTOR DE TESIS
Juan Mejía Trejo

Palabras clave: IBM SPSS Modeler, IBM SPSS Statistics, análisis multivariante, minería de datos.

Introducción

Una de las grandes limitaciones a la que los investigadores se enfrentan en la búsqueda de modelos predictores es el empleo del software adecuado para el análisis de datos. De acuerdo con León, Castellanos y Vargas (2006) uno de los problemas en la producción científica en su tarea de soportar decisiones empresariales que les permitan anticiparse a cambios, es que no se pone atención al software empleado y se terminan haciendo modelos que no predicen los fenómenos. El empleo del software posibilita a los investigadores de herramientas para la resolución de problemas organizacionales. No obstante, el tipo de programa empleado en la búsqueda de modelos predictivos puede relacionarse con la naturaleza del problema a tratar, por lo que aquí se analizan dos.

Uno de los planteamientos de la búsqueda de modelos predictivos es a partir de modelos estadísticos que previo a un análisis, se establecen

relaciones, es decir, cuando se parte de abstracciones de la realidad representada en modelos que representan los factores y sus relaciones para poder cuantificarlos. Dicho de otra manera, cuando la investigación parte de una hipótesis y se requiere de herramientas de análisis multivariante para cumplir con una serie de pruebas para la aceptación de los supuestos, la consolidación de la predicción y la posibilidad de la generalización a otras poblaciones a través del estudio en la muestra.

Otra propuesta para lograr modelos predictivos se enfoca partir de los datos existentes o históricos de las organizaciones. Por ejemplo, en varias partes del mundo se generan bases de datos pero no se hace uso óptimo de herramientas necesarias para presentar resultados que den solución a las problemáticas de los diferentes sectores de la sociedad (Castañeda, Cabrera, Navarro y de Bries, 2010). Por ello, es necesario extraer conocimiento a través de datos existentes, de manera específica se requiere de una minería de datos que permita conocer patrones antes ocultos en los datos (e.g. Aranda y Sotongo, 2013; Lobaina & Suárez, 2018). Por lo que se requiere de herramientas que permitan el análisis de diversas fuentes de datos y optimización de la información existente.

Con base en lo expuesto, es que se considera necesario revisar estas dos formas de construir o consolidar modelos de predicción así como analizar los softwares enfocados a los modelos de predicción. Para ello se toman como base los programas de la empresa IBM, se trata de SPSS Modeler y SPSS Statistics para descubrir si ¿Es posible el análisis multivariante predictivo a partir de un modelo estadístico con Modeler? o ¿Minería de datos a partir bases de datos preexistentes en Statistics?

De acuerdo a lo anterior es que el presente ensayo tiene como objetivo analizar las características de valor que ofrecen los programas IBM SPSS Modeler y Statistics para atender las incógnitas de investigación y ofrecer alternativas o soluciones óptimas en el establecimiento de modelos de predicción. Así también, conocer en qué circunstancias puede utilizarse uno para el estudio multivariante a partir de un modelo estadístico y, cuál otro, apto para la minería de datos. Para ello el documento está seccionado de la siguiente forma:

En la primera parte se presenta el antecedente de la búsqueda de conocimiento a partir de bases de datos existentes, minería de datos y metodología CRISP-DM, en la segunda, se trata el tema de modelo estadístico y análisis multivariante. En la tercera sección se presentan las características del software IBM SPSS Modeler y en la cuarta, lo relacionado al de Statistics. En la quinta parte se presenta un análisis comparativo de

las funciones que ofrecen los programas en comento. Por último, se presentan la discusión y las ideas concluyentes del presente ensayo.

1. El conocimiento en bases de datos

El descubrimiento del conocimiento de las bases de datos se refiere a un serie de etapas que consiste en el descubrimiento de bases de datos, integración de los mismos, determinar los objetivos, preparación de datos (selección y transformación), minería de datos, evaluación, interpretación y toma de decisiones, siendo la minería de datos una de las partes más importantes dentro del proceso (Aranda y Sotolongo, 2013; Cabena, Hadjinian, Stadler, Verhees y Zanasi, 1998). La minería de datos es uno de los más necesarios en para la optimización de las bases de datos (Lobaina y Suárez, 2018). Este proceso iintegra análisis de datos y extracción de modelos (Fayyad, 1996). El proceso de la extracción de conocimiento en bases de datos se observa a continuación.

Proceso del descubrimiento de conocimiento en bases de datos.

1. Integración y recopilación de los datos. Fase de extracción para decidir de donde se van a sacar los datos almacén de datos, data warehose, (Hernández, Ramírez y Ferri, 2004).
2. Determinación de objetivos. Fase de extracción para determinar el problema a resolver y el objetivo a perseguir (Cabena, *et al*., 1998).
3. Preparación de los datos. (Aranda y Sotolongo, 2013, Cabena, *et al*., 1998).
   Preparación de datos. Selección de fuentes de datos.
   Pre-procesamiento respecto a calidad y determinación de los datos.
   Transformación de datos: conversión en modelo analítico.
4. Minería de datos.
   Técnicas de predicción,
   Técnicas de clasificación (árboles de decisión y reglas de inducción),
   Técnicas de asociación (correlación),
   Técnicas de agrupamiento, (Aranda y Sotolongo, 2013).
5. Evaluación e interpretación de datos. Análisis y evaluación de los resultados (Cabena, *et al*., 1998).
6. Toma de decisiones o difusión y uso del conocimiento (Hernández, Ramírez y Ferri, 2004. Asimilación y aplicación del conocimiento (Cabena, *et al*., 1998).

Minería de datos

Como se mencionó párrafos arriba, la minería de datos es uno de los procesos más importantes en la extracción de conocimiento a partir de las bases de datos existentes. De acuerdo a Aranda y Sotongo (2013) es un proceso que permite establecer modelos predictivos con base a los datos existentes, mismos que pueden aplicarse en las diversas ramas del conocimiento (eg. Rama biológica, aplicaciones educacionales y financieras, procesos industriales, policiales y políticos. Se trata de actividades mediante las cuales se identifican patrones ocultos (Fayyad, 1996; Lobaina & Suárez, 2018, Cabena, et al., 1998). Además de lo anterior, es una fase exploratoria de datos para identificar relaciones sin un objetivo particular (IBM, 2012). Por tal razón, la minería es el proceso que se toma como base para este análisis y no la extracción de conocimiento en general.

Además, dicho proceso reúne todas las características de la extracción de conocimiento a partir de bases datos. Dado que se trata de extraer conocimiento no trivial de bases de datos disponibles de diversas fuentes y mediante diversas técnicas de extracción de conocimiento que resulta de las relaciones y dependencias entres los elementos constitutivos de los factores que se analizan ante un problema dado (e.g. Rodríguez *et. al* 2010). De ahí su importancia para la aplicación óptima de la información existente. Para ello se emplea una metodología o proceso estándar de seis fases para desarrollar e implementar análisis predictivos, Cross-Industry Standard Process for Data Mining, CRISP-DM por sus siglas en inglés y consiste en comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implantación, (Galán-Cortina, 2016). La metodología se aprecia en la tabla 1.

Como se puede observar, la minería de datos es un proceso mediante el cual se trata un problema organizacional a partir de los datos con que se cuenta en un momento dado, se debe tener conocimiento experto en el contexto organizacional así como de la organización misma para entender los datos con que se está trabajando. Es preciso destacar también que el análisis no se desprende de una hipótesis a probar ni de modelos preestablecidos, como se aprecia, parte de un planteamiento y análisis del contexto así como de la elección de las bases útiles para la identificación de conocimiento nuevo. Así también, se requiere de conocimiento en el tipo de datos que se tiene o debería tener para acceder a ellos, sí también para identificar si se trata de formatos diversos y variables distintas y verificar si es posible obtener nuevos datos con los ya existentes. Asimismo, en la modelación de los datos, se requiere de conocimiento de las técnicas

Tabla 1. Minería de Datos a través de la metodología CRISP-DM

Fase	Concepto	Actividades
Comprensión del negocio	Se trata de entender de forma profunda el contexto del problema.	Determinación de objetivos. Valoración de la situación. Determinación de los objetivos. Producción de un plan de proyecto
Comprensión de los datos	Se trata de entender a fondo los datos y su origen.	Recopilación de datos iniciales. Descripción de los datos Exploración de los datos Verificación de los datos.
Preparación de los datos.	Se unifica en un almacén de datos la diversidad de fuentes de datos internos (texto, html, bases de datos) y externos que se requieren para un análisis.	Selección de los datos. Limpieza de datos. Construcción de nuevos datos. Integración de datos. Formato de datos.
Extracción de modelos o modelado.	Se emplean métodos de análisis de datos para la extracción de información.	Selección de técnicas de modelado. Generación de un diseño de comprobación. Generación de modelos. Evaluación del modelo
Evaluación	Se evalúa el modelo de acuerdo al problema planteado.	Evaluación de los resultados. Proceso de revisión. Determinación de pasos siguientes.
Implementación	Se consideran los resultados de la evaluación y se define una estrategia. Técnicas de decisión (e.g. IBM Analytical Decision Management)	Planificación de la implementación. Planificación del control y mantenimiento. Creación de un informa final. Revisión final del proyecto. os resultados obtenidos, las predicciones (no son recomendaciones) se analiza ¿Cuál es la mejor decisión?

Fuente: Elaboración propia con base en IBM, 2016 y CRISP-DM, 2012.

de predicción, clasificación, segmentación y asociación para entender el proceso que se realiza. Por último, para la evaluación e implementación se precisa regresar al problema con que da origen el análisis para verificar si el modelo que se generó es el adecuado para entender el fenómeno organizacional y determinar las acciones a seguir así como seguimiento a su aplicación.

2. Conocimiento a partir de un modelo estadístico

En la sección anterior, se analizó el caso mediante el cual el análisis de datos parte de bases de datos existentes, no se inicia con hipótesis de investigación ni se tiene un modelo en que se establecen las relaciones entre las variables. En esta sección se hace un análisis de datos que tiene su origen en la modelación previa, es decir, en el establecimiento de relaciones y se considera bases de datos que tienen su origen en ello y que fueron recopilados para el estudio específico para la que fue creada. Para explicar entonces este tipo de análisis se parte primero de qué es un modelo estadístico y cuál es el análisis multivariante que debería tener.

Modelo estadístico

El modelo estadístico es un tipo de modelo matemático que representa la realidad y sus aspectos; refiere la contribución de cada factor a un fenómeno determinado para poderlo cuantificar mediante estadística, esta última con el fin de inferir información de los resultados de una muestra al resto de la población. El modelado entonces es el proceso de desarrollar modelos estadísticos (IBM, 2018). En el modelado estadístico se especifica de forma previa el modelo en el que se representa una hipótesis, es decir, se especifican relaciones.

Análisis multivariante

Las técnicas de análisis multivariante son cada vez más aceptadas en la investigación científica pues se requiere de analizar la compleja realidad a partir de relaciones que incluyen más de tres variables, además los software incluyen en sus desarrollados procesamientos las herramientas que permiten realizar estos análisis (Mejía, 2018). De acuerdo con el autor, un

análisis multivariante es el análisis simultáneo de más de dos variables en las que éstas se relacionan y sus efectos no pueden interpretarse de forma independiente, su propósito es medir, explicar y predecir el grado de relación, por lo que aspectos como test de significación, escalas de medida y valor teórico deben estar considerados para su correcta aplicación.

El análisis multivariante es una herramienta para lograr entender la relación significativa y predictiva entre variables. De acuerdo a Mejía (2018) debe haber previo a éste, un constructo teórico conceptual muy sólido, por ello recomienda determinar la significación práctica y estadística, considerar un tamaño muestral adecuado, conocer los datos, verificar de forma constante la parsimonia del modelo, atender a los errores, validar los resultados y dar validez al modelo. Y aunque no exista una metodología como tal, existen procedimientos o fases que deberían considerarse en el diseño de modelos predictivos que facilitan la modelización y permitan identificar la significación de los resultados y modelos que describan mejor a la población en su conjunto.

Dichas fases se centran en un plan de investigación bien definido teniendo como base un modelo conceptual que establezca las relaciones a examinar, seguido de la investigación de campo o empírica, selección de la técnica multivariante y su ejecución. Luego la obtención de los resultados significativos y la interpretación, que se convierte en un asunto central, enfocándose al valor teórico. Por último, las medidas de diagnóstico del modelo posibilitan que el modelo no sea válido sólo para la muestra, sino, que sea generalizable. Las fases son las siguientes, se muestran en la tabla 2.

Como puede observarse y haciendo un comparativo de ambos planteamientos para encontrar modelos predictivos se tiene lo siguiente: en ambos casos es necesario revisar el contexto, su problemática y el objetivo. No obstante, en la minería de datos no se parte de una hipótesis ni de un modelo que establece las relaciones a probar. Así también, mientras que en el establecimiento del diseño (análisis multivariante) el investigador define cuales son las variables a incluir, su operacionalización, el tamaño óptimo de la muestra donde se recaban los datos, así como identifica qué tipos de datos se utilizarán en el análisis inferencial y descriptivo, mientras que en el entendimiento de los datos (minería de datos) se identifican los datos que participarán en el análisis de acuerdo al planteamiento del problema.

Así también, en la evaluación de los supuestos en el análisis multivariante, para cada técnica se debe revisar que los datos cumplan con ciertos requisitos de linealidad, normalidad, homocedasticidad, homogenei-

Tabla 2
Fases para la modelación de un problema multivariante

Fase	Concepto	Actividades
1. Problema, objetivo y técnica multivariante.	Una vez planteado el problema, se definen los objetivos en términos conceptuales. Se determina el modelo conceptual con sus relaciones y por último se identifican las características de medida de las variables.	Definir el problema de investigación y objetivos analíticos. Modelos conceptuales: relación de dependencia: especificar conceptos dependientes y conceptos independientes. Relación de interdependencia: determinar dimensiones de la estructura. Identificar técnica (medida a utilizar) útil para examinar la relación. 1. Análisis dependiente. a) Una dependiente y escala métrica: regresión lineal múltiple y análisis de conjunto. b) Una dependiente y escala no métrica: análisis discriminante múltiple, análisis de regresión logística. c) Varias dependientes y escala métrica: análisis de correlación canónica. d) Varias dependientes y escala no métrica: análisis de la varianza multivariante (MANOVA). e) Múltiples dependientes e independientes: modelo de ecuaciones estructurales. 2. Análisis interdependiente. a) Estructura en variables: análisis factorial. b) Estructura en casos: análisis cluster. Estructura en objetos métrica: análisis multidimensional. d) Estructura objetos no métrica: análisis de correspondencias.

2. Diseño: desarrollo del proyecto de análisis.	Desarrollar un plan de análisis que dirijan el conjunto de supuestos que subyace a la técnica para finalizar la formulación del modelo y las especificaciones de la recolección de datos.	Tamaño de muestra óptimo. Tipos de datos métricos y no métricos, métodos de estimación del modelo.
3. Evaluación de los supuestos de la técnica multivariante.	Evaluar los supuestos subyacentes estadísticos como conceptuales para revisar su capacidad de representar relaciones multivariantes.	Estadística inferencial: Normalidad. Linealidad, independencia en los términos de error e igualdad de varianzas en una relación dependiente: Homocedasticidad. Homogeneidad de la muestra Nexos conceptuales. Cada técnica tiene sus propios supuestos conceptuales: formulación de modelos y representaciones.
4. Ejecución	Estimación efectiva del modelo. Valoración del ajuste del modelo o corroborar que se obtuvieron niveles aceptables sobre los criterios estadísticos para posibilitar la inferencia.	Estimación. Elección de características específicas de los datos: e.g. MANOVA. Maximizar ajuste de los datos: e.g. rotación de factores o funciones discriminantes. Valoración. E.g. nivel de significación.
5. Interpretación	Interpretación del valor teórico (combinaciones múltiples de variables) de los resultados multivariantes para conocer la naturaleza de las relaciones.	Examinar los valores de los coeficientes estimados (ponderaciones) para cada variable. En la interpretación se puede volver a formular el modelo re-especificando las variables.

Fuente: Elaboración propia con base en Mejía, 2017 y 2018.

dad mientras que en la preparación de los datos en la minería de datos se preparan los mismos para realizar el modelado, pero no se trata de cumplir con supuestos, se corre el riesgo de que el modelo predictivo no sea tan preciso (IBM, 2016). Ahora bien, en la ejecución del modelo y modelado, en ambos análisis es parte crucial para la predicción del fenómeno. Asimismo, en la interpretación y evaluación se requiere de una revisión analítica de los resultados por parte del investigador. Por último en la

parte de validación del análisis multivariante se quiere de garantizar el grado de generalización del modelo y en ambos, para la implementación, es preciso combinar los resultados con la experiencia del investigador en la toma de decisiones.

3. Características de IMB SPSS modeler versión 18

El uso de software especializado cada vez es más recurrente en el análisis de grandes bases de datos. Las herramientas de análisis de software son necesarias para analizar grandes cantidades de bases de datos (Aranda y Sotolongo, 2013). Pues cuando se analizan demasiados datos el proceso se convierte en algo engorroso (Soto-Jaramillo, 2009). No obstante, los programas tienen la cualidad de ser de difícil acceso pues son muy costosos (Aranda y Sotolongo, 2013). Al tener poco acceso a dichos programas reduce la posibilidad de realizar una extracción de conocimiento nuevo y de patrones o modelos predictivos

Es así que empresas como Microsoft, Oracle así como IBM crean las herramientas que permiten agilizar el tiempo de respuesta para procesar grandes bases y puedan ser transformados (Galán-Cortina, 2016). El Programa *SPSS Modeler de IBM* ha sido reconocido por su papel en brindar a sus clientes un enfoque integrado y holístico en la gestión empresarial y análisis de grandes bases de datos. Es por ello que en este apartado se analizan las principales características que posee el programa *IBM SPSS Modeler* en las herrami*entas que brinda en sus bases de datos.*

*I*BM SPSS Modeler es una herramienta de software para la minería de datos desarrollada por IBM Corp., (IBM, 2016a). Originalmente se llamaba *SPSS Clementine* y en 2009 se llamó *PASW Modeler* (Galpan-Cortina, 2015). Se trata de un programa orientado las empresas con el objetivo de una mejor comprensión de los datos y obtener los mejores resultados. Esto, mediante una interfaz visual que conlleva a modelos de predicción (pronósticos, clasificaciones, segmentación y asociaciones) más eficaces y en menos tiempo.

En su diseño considera la metodología CRISP-DM, ésta como la minería e datos se muestran en los previos párrafos (IBM, 2012). Este software funciona con interfaz visual y las funciones basadas en nodos –íconos- que se van tomando y forman una ruta o stream que pueden archivarse (de manera individual o por proyectos, mismos a los que se puede acceder para reestructurarse y las bases de datos quedan independientes (IBM, 2016a). La ruta es llamada ruta de datos, consiste en una secuencia de

operaciones en donde los datos fluyen de registro en registro, se manipula un llegan a un destino, esto es, I*MB SPSS Modeler*, mediante el lienzo de rutas lee los datos, los ejecuta y por último los envía a un destino dibujando diagramas de operaciones (con nodos=íconos y rutas=flujo). Como resultado final visual se obtiene un nugget que contiene todas operaciones realizadas. En conjunto con Modeler Server ofrece un rendimiento mayor en el caso que se trabaje con grandes cantidades de datos. Las características del software se muestran en la siguiente tabla 3.

Tabla 3.
Caracteristicas generales de *SPSS Modeler*

Caracteristicas generales del software	
Sistemas operativos	Windows, Linux, Sun Solaris, HP-UX o IBM-AIX, Mac OsX.
Lenguaje	Código abierto, Python (Jython), R. Pago: SPSS
Interfaz	Java
Licencia	Usuarios autorizados y concurrentes limitado en cantidad (de personas), plazo, mantenimiento, alquiler.
Competencia	Oracle Data Mining, WEKA, RapidMiner, Statistic Data Miner, IBM DB2 InfoSphere Warehouse, Microsoft Analysis Services.
Compatible con	Oracle Data Miner, IBM DB2 InfoSphere Warehouse y Microsoft Analysis Services, IBM Netezza Analytics
Compatibilidad	SQL (Structured Query Language), lenguaje de consulta estructurada
Estadísticos	Confirma relaciones sospechosas entre las variables. Los estadísticos de IBM SPSS Statistics pueden emplearse en IBM SPSS Modeler. Se puede acceder y ejecutar determinadas rutinas de IBM SPSS Statistics en IBM SPSS Modeler para generar y puntuar modelos.
Manipulación de datos	Construye nuevas bases de datos a partir de los existentes.
Exploración y visualización	En nodo auditoría de datos desarrolla auditoría inicial de los datos incluyendo gráficos y estadísticos.
Comprobación de hipótesis	Construye modelos que muestran la forma en que se comportan los datos, y verifica estos modelos.

Productos	IBM SPSS Modeler, IBM SPSS Modeler Server, IBM SPSS Modeler Administration Console, IBM SPSS Modeler Batch, IBM SPSS Modeler Solution Publisher, IBM SPSS Modeler Server adaptadores para IBM SPSS Collaboration and Deployment Services
Ediciones	IBM SPSS Modeler Profesional e IBM SPSS Modeler Premiun conformado por IBM SPSS Modeler Entity Analytics, IBM SPSS Modeler Social Network Analysis e IBM SPSS Modeler Text Analytics.
Nodos	Nodos de origen en la paleta de orígenes: lector de datos en forma de círculo Nodos de proceso en la paleta de operadores con registros y con campos: transformador de datos, en forma de hexágono. Nodos de resultados en la paleta de gráficos, informe, exportar: generador de documentos, en forma de triángulo (gráfico) y en forma de cuadrado (informe, exportar) Nodos de modelado en la paleta de modelado: generador de modelos en forma de pentágono.
Preparación de datos	ADP nodo de preparación de datos automática Nodo auditoría de datos valores perdidos, atípicos, sesgos.
Métodos de modelado	Clasificación. Los modelos usan el valor de uno o más campos de entrada para predecir el valor de uno o más resultados o campos de destino, incluyen: aprendizaje automático de las máquinas, inducción de reglas, identificación de subgrupos, métodos estadísticos y generación de varios modelos. Nodo clasificador, nodo auto-numérico, nodo de árbol de clasificación y regresión (C&R), nodo QUEST, nodo CHAID, nodo C5.0, Nodo lista de decisiones, modelos de regresión lineal, nodo PCA/Factorial, Nodo Selección, análisis discriminante, regresión logística, modelo lineal generalizado, nodo de regresión Cox, nod máquina de vectores (SVM), nodo red bayesiana, nodo modelado de respuesta, nodo serial temporal, nodo K (KNN), nodo predicción espacio-temporal (STP).

	Asociación. Modelos que identifican patrones en los datos en los que una o más entidades, se asocian con una o más entidades, útil cuando se desea predecir varios resultados. Nodo a priori, modelo CARMA, nodo secuencia, nodo reglas de asociación.
	Segmentación. Son modelos de agrupación en clústeres que dividen los datos en segmentos o clústeres de registros que tienen patrones similares de campos de entrada, cuando se desconoce el resultado especifico. Nodo agrupación, nodo K-medias, nodo Kohonen, nodo Bietápico, nodo detección de anomalías.
Evaluación	Se pueden comprobar supuestos verificando el comportamiento de los datos y de los modelos.
Implementación	Los resultados obtenidos no son sugerencias, Modeler ofrece un componente central de Predictive Analytics de IBM en Linux para sistema Z que optimiza la decisión de acuerdo al planteamiento.
Productos IBM SPSS Modeler	IBM SPSS Modeler, IBM SPSS Modeler Server, IBM SPSS Modeler Administration Console, IBM SPSS Modeler Batch, IBM SPSS Modeler Solution Publisher, Adaptadores de IBM SPSS Modeler Server para IBM SPSS Collaboration and Deployment Services

Elaboración propia con base en IBM, 2016 y CRISP-DM, 2012.

Como puede observarse *Modeler* cuenta con una serie de algoritmos que proporcionan las herramientas para el análisis de datos cuando se pretende hacer uso óptimo la información con que se cuenta en el planteamiento de determinados problemas. Su interfaz es sencilla de emplear pues cuenta con una serie de íconos –nodos– que el investigador puede ir empleando en el desarrollo de sus modelos de acuerdo con la metodología CRISP-DM. Se pueden comprobar supuestos verificando el comportamiento de los datos y de los modelos.

4. Estudios analizados con spss modeler

Se realizó una búsqueda de información respecto al uso de s*oftware Modeler SPSS* que permita conocer la trascendencia que tiene en la minería de datos. Los estudios revelan que el programa se emplea en el análisis de datos que permite identificar aspectos que previamente, con otras técni-

cas, no ha sido posible. En el caso de Lotfnezhad, Ahmadi, Roudbari y Sadoughi (2015) se extrajeron y descubrieron patrones ocultos en las bases de datos recopilados y se analizaron diversos factores que generan el cáncer de mama; el modelo identificó diez variables predictoras. Asimismo, Shao, Liancheng y Han (2016) concluyeron que el método de predicción funciona de forma efectiva y precisa en la selección en la identificación de la falla de línea single-line-to-ground (SLG), analizando datos de cantidades eléctricas y no eléctricas.

Otros casos como el de Wang, Wen, Lu, Yao y Zhao (2016) desarrollaron modelos de predicción para predecir eventos relacionados con el esqueleto SER en pacientes con cáncer. Y el respecto al análisis de Di, Yang, Fu, Lin y Jiang (2013) desarrollaron un modelo de elección de puntos de acupuntura en la cefalea de tipo tensional, en la que ésta se materializa principalmente en la selección de puntos de acupuntura ubicados en la cabeza combinados con un meridiano distintivo. En éstos y los casos previos, se observa el uso del software en el análisis de datos y en la modelación de predicciones precisa y eficiente. Tal como IBM (2016) establece dentro de sus propósitos.

En los párrafos anteriores se reconoce la función del software I*BM SPSS Modeler* en la minería de datos, es decir en el análisis de datos con el propósito de encontrar nuevos patrones o patrones ocultos y las relaciones que tienen los datos, esto es, cuando el análisis no inicia con un modelo de relaciones de variables. Así también, no se inicia con una hipótesis a probar y los datos con los que se cuenta fueron recopilados para fines diversos en las diferentes áreas y no con un objetivo en particular, además se hizo uso de múltiples bases y diversos formatos y generalmente se reutilizaron. Dicho software ha sido útil en las decisiones que requieren de una modelación rápida con las diversas bases existentes. Se encuentra en todos ellos la eficiencia de los modelos en modelación predictiva, la que ya se encuentra en la fase de implementación.

5. Características de IBM SPSS Statistics versión 24

SPSS corresponde al nombre de Statistical Package for the Social Sciences, software estadístico que ha sido empleado por las ciencias exactas, sociales y aplicadas, incluso las de investigación de mercado (Mejía, 2018). De acuerdo a este autor la primera versión data de 1968 y la última, la versión 24 en 2016. El software *IBM SPSS Statistics* emplea bases casi de

cualquier tipo para convertirlos en informes a manera de tablas, gráficos, diagramas de distribuciones, así también empleado para el análisis de datos estadísticos descriptivos y complejos (IBM, 2011).

Statistics emplea un lenguaje en comandos, menús y cuadros de diálogo. Entre sus principales opciones, este programa cuenta con Statistics base que proporciona una gama de procedimientos estadísticos como tablas de contingencia, estadísticas descriptivas, así como variedad de operaciones como reducción de dimensiones, clasificación, segmentación (análisis factorial, análisis de conglomerados, discriminante). Y sobre todo, ofrece una serie de algoritmos para comparar medias y técnicas predictivas, t, ANOVA, Regresión lineal y regresión ordinal. Otra opción que ofrece es estadísticas avanzadas como análisis de supervivencia Kaplan-Meyer, otra como lo es Bootstrapping (método para obtener estimaciones más precisas, intervalos de confianza para estimar por ejemplo coeficientes de correlación y regresión; otras como conjoint, tablas personalizadas, preparación de datos, árboles de decisión, predicciones (ajustes de curvas), valores perdidos. En la tabla 4 se observan algunas características.

Tabla 4.
Características generales del software, IBM SPSS STATISTICS 24

Caracteristicas generales del software	
Sistemas operativos	Windows, Linux, Mac OSX
Lenguaje o sintaxis	Código abierto, Python (2,7), R Pago: SPSS
Interfaz	Gráfica
Licencia y costo	Usuarios autorizados y concurrentes limitado en cantidad (de personas), plazo, mantenimiento, alquiler
Competencia	SAS, MATLAB, Statistica, Stata, R

Nuevas funciones de la versión 24

Nuevas funciones	Acceso a extensiones, IBM SPSS Extensión Hub, Custom Dialog Builder y Módulo tablas personalizadas
Acceso	A más de 100 extensiones lo que permite acceso a bibliotecas gratuitas (Ver lenguaje)
IBM SPSS Extensión Hub	Explorar, descargar, actualizar eliminar y administrar extensiones

Custom Dialog Builder	Actualización para construir e instalar extensiones.
Módulo tablas personalizadas	Para importar y exportar datos a SPSS
Ediciones	SPSS Statistics Estándar, SPSS Statistic Profesional, SPSS Statistics Premiun.
SPSS Statistics Estándar,	Procedimientos estadísticos para modelos estadísticos lineales, no lineales, de simulación predictiva.
SPSS Statistics Profesional	Preparación de datos, valores perdidos, validez de los datos, árboles de decisión y pronóstico.
SPSS Statistics Premium	Permite modelos de ecuaciones estrcturales y pruebas de maestro. Paquetes de marketing directo, tablas y gráficos de alta gama.

Elaboración propia con base en IBM, 2016.

6. Análisis comparativo de IBM modeler y statistics

En la siguiente tabla se clasificaron las funciones de los dos programas con el objetivo de identificar si se puede hacer minería de datos con *Statistics* o se puede realizar un análisis multivariante a partir de un modelo estadístico con *Statistics*.

Como puede observarse y haciendo un comparativo del software en el análisis de datos bajo los planteamientos de análisis multivariante y minería de datos en el establecimiento de modelos predictivos se tiene lo siguiente: en las primeras fases de establecer objetivo y entender el negocio-entendimiento de los datos en ambos casos es preciso que el investigador revise el contexto, su problemática y el objetivo, independientemente del software. No obstante, *Modeler* cuenta con un espacio en su plataforma que permite dar un seguimiento a las fases de la minería de datos, empezando con entendimiento del negocio y entendimiento de los datos, espacio en que se pueden guardar archivos o datos que el investigador precise para el entendimiento de los datos como del contexto. En el caso del diseño en el análisis multivariante, también se realiza de forma independiente al software que se emplee para su análisis, ya que el investigador define cuales variables elegir, cómo medirlas, así como determinar el tamaño muestra y conocer los datos (métricos o no métricos).

Así también, en la evaluación de los supuestos de normalidad, homocedasticidad, homogeneidad en el análisis multivariante, pueden ejecutarse con comandos a través de cuadros de diálogo en el programa SPSS Statistics, asimismo en *Modeler* los supuestos y preparación de datos ta-

Tabla 6. Análisis comparativo de las herramientas IBM Modeler y Statistics

Fase	Concepto	Actividades	IBM Modeler	IBM Statistics
Objetivo AM /entendi-miento del negocio MD	Se trata de entender de forma profunda el contexto del problema	Analizar el contexto Seleccionar objetivos	En su plataforma cuenta con un espacio para que el investigador guarde, archive y visualice la información útil para el entendimiento del negocio y de los datos.	El investigador revisa de forma independiente al software el planteamiento del problema, así como el objetivo.
Diseño AM/Entendi-miento de los datos MD	**AM** Se trata de establecer el modelo conceptual solido en el cual se establezcan las relaciones a medir, las hipótesis a probar y se cuente con datos representativos de la población	Definir variables. Establecer cómo medir variables. Determinar tamaño muestral. identificar datos métricos y no métricos		De forma independiente el investigador diseña el modelo en que establece las relaciones con las variables a incluir, la operacionalización, el tamaño de la muestra donde se recaban los datos y revisa los datos (métricos y no métricos) para elegir la técnica multivariante.
	MD Se trata de entender a fondo los datos y su origen	Recopilación de datos iniciales. Descripción de los datos. Exploración de los datos. Verificación de los datos	En su plataforma cuenta con un espacio para que el investigador guarde, archive y visualice la información útil para el entendimiento del negocio y de los datos. El investigador puede formular hipótesis si explora los datos, esto último útil en dar forma a la transformación de los datos.	

Supuestos AM/Prepara-ción de datos MD	**AM** Evaluar los supuestos subyacentes estadísticos como conceptuales para revisar su capacidad de representar relaciones multivariantes.	Estadística inferencial: Normalidad. Linealidad, independencia en los términos de error e igualdad de varianzas en una relación dependiente: Homocedasticidad. Homogeneidad de la muestra Nexos conceptuales. Cada técnica tiene sus propios supuestos conceptuales: formulación de modelos y representaciones			Acceso a Excel y archivos SPSS. Rápida instantánea visual de los datos. Permite aplicar reglas de validación que identifiquen valores de los datos no válidos, fuera de rango, perdidos o en blanco. Homocedasticidad: e.g. analizar-comparar medias-ANOVA de un factor-seleccionar en lista de dependientes variable métrica-selección de variable Factor-opciones-prueba de homogeneidad de las varianzas.
	MD Se unifica en un almacén de datos la diversidad de fuentes de dato s internos (texto, html, bases de datos) y externos que se requieren para un análisis.	Selección de los datos. Limpieza de datos. Construcción de nuevos datos. Integración de datos. Formato de datos.	Acceso de datos	Fuentes de datos ODBC, tablas Excel, archivos planos ASCII y archivos SPSS.	
			Pre-procesado de datos	pick & mix, muestreo, particiones, reordenación de campos, nuevas estrategias para la fusión de tablas, nuevas técnicas para recodificar intervalos numéricos, etc Auditoría de datos(perdidos, sesgos, atípicos) Linealidad (eg. Nodo Tipo o GLE)	

Fase	Concepto	Actividades	IBM Modeler		IBM Statistic	
Ejecución AM Extracción de modelos o modelado CRISP -DM.	AM Estimación efectiva del modelo. Valoración del ajuste del modelo o corroborar que se obtuvieron niveles aceptables sobre los criterios estadísticos para posibilitar la inferencia	Estimación. Elección de características específicas de los datos: e.g. MANOVA. Maximizar ajuste de los datos: e.g. rotación de factores o funciones discriminantes. Valoración. E.g. nivel de significación			Árboles de decisión, redes neuronales, agrupamiento, asociación, regresión lineal, combinación de modelos. Técnicas de asociación. Análisis de conjunto	
	MD Se emplean métodos de análisis de datos para la extracción de información.	Selección de técnicas de modelado. Generación de un diseño de comprobación. Generación de modelos. Evaluación del modelo	Técnicas de aprendizaje. Modelado predictivo; de asociación y secuencia y; de seg-mentación	e.g. nodos de á rboles de decisión, redes neuronales, agrupamiento, reglas de asociación, regresión lineal y logística, combinación de modelos.	Statistic permite el modelado a partir de cuadros de diálogo y comandos	e.g. analizar - regresión - lineal, selección de variable dependiente y de variable independiente.
Interpretación AM/Evaluación CRISP - DM.	**AM** Interpretación del valor teórico (combinaciones múltiples de variables) de los resultados multivariantes para conocer la naturaleza de las relaciones.	Examinar los valores de los coeficientes estimados (ponderaciones) para cada variable. En la interpretación se puede volver a formular el modelo re-especificando las variables.		Los resultados pueden desplegarse fácilmente en bases de datos como IBM Statistics y otras aplicaciones. Los modelos generados incluyen los niveles de significación, pruebas t, F	Proceso analítico en la interpretación de resultados mutivariantes. Revisión del valor teórico (ponderaciones de las variables) Significación práctica	Los modelos que se generan pueden probarse los niveles de significación, pruebas t, F

	MD Se evalúa el modelo de acuerdo al problema planteado.	Evaluación de los resultados. Proceso de revisión. Determinación de pasos siguientes.	Técnicas para la evaluación	Modelos guiadas por las condiciones especificadas por el experto		
Implementa-ción MD	Se consideran los resultados de la evaluación y se define una estrategia.	Planificación de la implementación. Planificación del control y mantenimiento. Creación de un informa final. Revisión final del proyecto. Los resultados obtenidos, las predicciones (no son recomendaciones) se analiza ¿Cuál es la mejor decisión?	Visualización de resultados	Ofrece un potente soporte gráfico que permite al usuario tener una visión global de todo el proceso, que comprende desde el análisis del problema hasta la imagen del modelo aprendido.	Para la implementación los resultados no son reco-mendaciones, Modeler cuenta con un componente central PREdictive Analytics de IBM que coadyuva a identificar la mejor opción para el problema planteado.	
			Informes	en HTML y texto, volcar los resultados de la minería de datos obtenidos en bases de datos y exportar los modelos a distintos lenguajes como C, SPSS, HTML, PMML, SQL, etc.		
Validez del modelo AM	Del modelo para garantizar cierto grado de generalidad del moddo obtenido.	Demostrar la generalidad del modelo obtenido de los datos de una muestra al resto de la población. bootstrapping	Boostrap Boosting Para generar modelos más precisos, estables	En cada modelo generado, Modeler obtiene el nivel de significación p ara las pruebas de hipótesis, incluyendo pruebas de bondad de ajuste, F.	Para garantizar cierto grado de generalidad. Validación cruzada Jacknite Bootstrap	En cada modelo, Statistic permte revisar el nivel de sigificación, incluyendo pruebas F, t, ajustes de bondad.

Otras características adicionales a la metodología o fases de AM y MD		
Diseño	Minería de datos en la construcción de modelos predictivos y análisis de datos	Paquete de productos para consolidación de modelos predictivos y análisis de datos profesionales.
Usuarios	Poca habilidad en programación por el diseño de la interfaz. Mucha habilidad en técnicas.	Mucha habilidad por tratarse de comandos, códigos o cuadros de diálogo. Mucha habilidad en técnicas.
Usuarios	Poca habilidad en programación por el diseño de la interfaz. Mucha habilidad en técnicas.	Mucha habilidad por tratarse de comandos, códigos o cuadros de diálogo. Mucha habilidad en técnicas.
Objetivo del modelo	Fuerte para decisiones operativas, débil para investigación y comprensión.	Fuerte para investigación y comprensión, apropiado para procedimientos multivariados clásicos.
Precisión de la predicción	Construye, compara, prueba y combina rápidamente modelos, pero no se comprende la significación, ni se demuestra variación.	Compre nde variación desde el punto de vista estadístico y significación.
Origen de los datos	Datos existentes que se recopilaron	Una base de datos recopilada para un cierto análisis, con un proceso analítico
Fuentes de datos	Diversas fuentes y múltiples formatos. Consolida la diversidad de datos y formatos. Reutilizan datos	Una sola fuente, archivos un solo formato. Base nueva, se analiza cuando aparece.
Informes analíticos	Detección de patrones ocultos	Capacidad de tabulación de datos, creación de gráficos.
Planteamiento del problema	Minería de datos. Para encontrar nuevos patrones y relaciones entre los datos, resultado modelo de predicción. No se especifican relaciones Recorrido exploratorio Sin un objetivo particular No requiere hipótesis Debe cumplir con metodología CRISP-DM	Modelo estadístico. El modelo se especifica de antemano. Se especifican relaciones Se busca una respuesta a una pregunta y un objetivo Requiere de una hipótesis a comprobar Método científico

Entify Analitics	Detección de duplicación para identificar los que tienen una gran probabilidad de pertenecer a la misma entidad.	No tiene la herramienta.
Pronóstico	Identifica el método de series de tiempo que mejor se ajusta a los datos históricos	El paquete SPSS Statistics Premiun permite entre otras cosas el pronóstico.
Datos	Big data, ágil con grandes volúmenes de datos. Se puede tener acceso a ellos sin sacarlos del repositario.	2 millones de registros y 250,000 variables Debe extraerse la base de datos al software para su análisis.
Expresión del modelo de predicción.	Es un objeto físico y portátil que se puede conver tir a PMML y adjuntarlo a nuevos nodos.	Se expresa en forma de coeficientes y estadísticos. El analista analiza estos resultados, le da sentido
Optimización de la predicción	La predicción no es una recomendación. Modeler ofrece IBM Analytical Decision Management para que con los resultados predictivos obtenidos, se tome la mejor decisión. Más allá de la predicción, ofrece escenarios futuros para la toma de decisiones.	La predicción no es una recomendación. El analista recomienda de acuerdo a los resultados, estudios previos y teoría con que se analiza el fenómeno.
Y=mx + c	LINEAL, versión mejorada de regresión lineal	Regresión lineal
Conversión a otros lenguajes.	Los algoritmos de inducción de reglas/árboles de decisión (clasificación) pueden convertirse en SQL y se pueden comparar con otros métodos.	No tiene la herramienta
Exportación de modelos a otros le nguajes	SAS,SPSS	
Exportación de datos	XLS	
Lenguaje	Éxito en analíticas de código abierto. Interfaz directo a R y puede combinarse con otros procedimientos.	La versión 24 ofrece acceso a más de 100 extensiones para aprovechar bibliotecas gratuitas en s intaxis R, Python y SPSS
Y=mx + c	LINEAL, versión mejorada de regresión lineal	Regresión lineal
Conversión a otros lenguajes.	Los algoritmos de inducción de reglas/árboles de decisión (clasificación) pueden convertirse en SQL y se pueden comparar con otr os métodos.	No tiene la herramienta

Exportación de modelos a otros lenguajes	SAS,SPSS	
Exportación de datos	XLS	
Lenguaje	Éxito en analíticas de código abierto. Interfaz directo a R y puede combinarse con otros procedimientos.	La versión 24 ofrece acceso a más de 100 extensiones para aprovechar bibliotecas gratuitas en sintaxis R, Python y SPSS
Análisis de Texto	Text Analytics de Modeler Premiun permite transformar datos no estructurados a datos estructurados.	No tiene la herramienta de procesamiento de texto.
Interfaz	Visual, basada en procesos, flujos de datos (streams), íconos, fáciles de usar. Programación visual que ilustra.	Códigos, comandos o cuadros de diálogo.
Indicadores	Índice de Victoria Huwitz. Modeler permi te el análisis holístico para la administración y análisis de grandes bases de datos.	

Elaboración propia con base a IBM modeler 2016 e IBM statistics 2016.

les como atípicos, sesgo y perdidos (e.g. auditoría de datos) y de linealidad en nodo Tipo o nodo GLE en el que se determina condiciones del modelo.

Ahora bien, en la ejecución del modelo y modelado, en ambos análisis es parte crucial para la predicción del fenómeno, para el análisis multivariante, una vez seleccionada la técnica indicada para el tipo de problema y variables a emplear, así como los datos recolectados en una muestra representativa, se ejecuta en el programa SPSS Statistics mediante comandos y cuadros de diálogo, (e.g. regresión lineal simple: analizar, regresión, lineal, selección de variable dependiente, selección de variables). Mientras que el modelado en SPSS se da mediante diversos nodos, sin tener que buscar los cuadros de diálogo o ejecutar comandos, sino en forma automática empleando los nodos que cuenta con una gran variedad que facilita el trabajo al investigador y optimiza las bases de datos resultando en cada uno de los modelos el más adecuado de acuerdo al objetivo. En ambos se puede hace uso de datos métricos como no paramétricos (e.g. nodo C&R, CHAID en *modeler*).

Asimismo, en la interpretación y evaluación para ambos casos la revisión analítica del investigador es necesaria y ningún programa la sustituye. Para ambos casos la modelación permite analizar los valores de las ponderaciones, en las cuales se revisa el nivel de significancia en la prueba de hipótesis, las de ajuste de bondad, F y t. Así también, en la parte de validación del análisis multivariante se quiere de garantizar el grado de generalización del modelo con algunas técnicas como validación cruzada, *jacknite* y *bootstrap*, mismas que el programa SPSS *Statistics* ofrece mediante comandos y cuadros de diálogo. Así también, *Modeler* ofrece las herramientas de *boosting* y *bootstrap* para generar modelos más precisos más estables y más consistentes.

En ambos, para la implementación, es preciso combinar los resultados con la experiencia del investigador en la toma de decisiones, no obstante, y considerando que los resultados no son recomendaciones, *Modeler* ofrece una alternativa para evaluar los resultados y someterlos a otro proceso para delinear las técnicas de decisión; *IBM Analytical decision management*, herramienta que apoya en proporcionar la mejor alternativa posible para el problema de que se trate.

Discusión

Atendiendo a la pregunta y tratando de contestar si ¿Se puede hacer minería de datos con *Statistics*? y con base en todo lo anteriormente expues-

tos podría decir que no, que la minería de datos es adecuada con software *Modeler* ya que dicho proceso consiste en un análisis de datos con el objetivo de encontrar nuevos patrones o patrones ocultos y las relaciones que tienen los datos, esto es, cuando el análisis no inicia con un modelo de relaciones de variables, ni con un modelo estadístico previo, sino con una metodología estructurada CRISP-DM.

Así también, no se inicia necesariamente con una hipótesis a probar y los datos con los que se cuenta fueron recopilados para fines diversos a lo largo del tiempo y no con un objetivo preciso, además son múltiples bases y diversos formatos y generalmente se reutilizan. *Modeler* puede ser óptimo en las decisiones operativas en la modelación rápida. Asimismo, las características del software permiten llevar a cabo el proceso de describir conocimiento y crear modelos predictivos ofreciendo una serie de facilitadores a los investigadores que aunque no requieren gran experiencia de programación (e.g. IMB, 2016a), sí requiere de conocer las técnicas y la plataforma que se emplea, ya que aunque el sistema ofrece automatización en sus operaciones, debe entenderse todo lo relacionado a los datos y sus resultados. Este programa tiene la opción de tener en la interfaz, la herramienta de *Statistics*, para que, poder realizar algunas pruebas en *Modeler*, por ejemplo modelado.

En en caso del cuestionamiento de si ¿Se puede realizar un análisis multivariante a partir de un modelo estadístico con *Modeler*? No necesariamente, si se requiere de analizar una base de datos para probar una hipótesis, es el software *Statistics* el apropiado cuando se pretende realizar un análisis de datos con el objetivo de encontrar relaciones con un modelo estadístico previo. Así también, el análisis parte con un objetivo claro y una hipótesis de investigación a probar. Cabe destacar que los datos con los que se cuenta fueron recopilados para ese objetivo, además corresponden a una sola base, un solo formato y se trata de una base nueva. Dicho software puede ser óptimo en la investigación y comprensión de un fenómeno en la modelación precisa, pues contiene las herramientas para comprender la variación de los datos estadísticamente así como su significación, aspecto necesario para la demostración de la probabilidad del evento y su precisión. Por lo tanto, para realizar análisis multivariantes a partir de hipótesis y con pruebas de significación el software *IBM Statistics* es el adecuado para hacerlo.

Las herramientas de SPSS Statistics permiten el análisis multivariante de técnicas dependientes de la forma convencional, en la que la investigación encuentra las bases para explorar la variación de los datos, es decir el error de la probabilidad en que se puede dar el evento estudiado,

así como la significación que es de vital importancia en la demostración del potencial explicativo de un modelo estadístico. No obstante, modeler también ofrece el resultado de las pruebas de significancia que requiere un modelo para poder probar su nivel de predicción, su precisión y estabilidad, permitiendo por otro lado, una interfaz más sencilla, pero más completa en el establecimiento de modelos predictivos.

Conclusiones

El propósito de este ensayo fue esclarecer las características de *Modeler* y *Statistics* en la búsqueda de modelos predictivos a partir de dos enfoques, la extracción de conocimiento a través de la minería de datos de bases existentes por un lado y por otro, el análisis multivariante a partir de modelos estadísticos. Y a través de su desarrollo se expusieron las diferentes herramientas que los softwares ofrecen para ambos enfoques.

Respecto a la minería de datos, se apreció que se trata de una de las principales fases del proceso de extracción de datos en la que se identifican datos así como patrones antes ocultos, se modelan relaciones, así como su nivel de predicción a partir de bases existentes. Por lo que el software *Modeler* es más adecuado cuando se pretende llegar a modelos predictivos a partir ésta, pues entre otras cosas puede ser óptimo en las decisiones operativas en la modelación rápida identificando patrones a través de bases de múltiples fuentes y diversos formatos, cuando no se tiene un objetivo claro. Asimismo, las características del software permiten llevar a cabo el proceso de describir conocimiento, y además posibilita las investigaciones científicas que se proponen responder a variaciones y test de significancia de los datos.

Así también, en lo que respecta a los modelos estadísticos, se observó que representan de forma previa los aspectos y relaciones de un evento el cual señala supuestos (hipótesis) para posibilitar su cuantificación mediante estadística y poder inferir los resultados al resto de la población. Así también las técnicas de análisis multivariante permiten realizar estas mediciones de las relaciones entre variables (tres o más) y que requieren de pasar las pruebas de significación, escalas de medida y valor teórico para su correcta aplicación.

Por lo que el software *Statistics* puede ser óptimo en la investigación y comprensión de un fenómeno en la modelación precisa a partir de una hipótesis y una base de datos, pues contiene las herramientas para comprender la variación de los datos estadísticamente así como su significa-

ción, aspecto necesario para la demostración de la probabilidad del evento y su precisión. Por lo tanto, para realizar análisis multivariantes a partir de hipótesis y con pruebas de significación el software *IBM Statistics* es el adecuado para hacerlo. Sus herramientas permiten el análisis multivariante de técnicas dependientes de la forma convencional, en la que la investigación encuentra las bases para explorar la variación de los datos, es decir el error de la probabilidad en que se puede dar el evento estudiado, así como la significación que es de vital importancia en la demostración del potencial explicativo de un modelo estadístico.

Así también, los modelos predictivos que surgen del análisis en Modeler, ofrecen precisión, estabilidad y consistencia. Además de ofrecer el plus de la automatización, ya que esta plataforma hace más sencillo el trabajo complejo de investigadores, académicos y empresarios. En cambio *Statistics* no ofrece la interfaz tan gráfica como la primera, pero sus comandos, cuadros de diálogo y el paquete de programas estadísticos permite al investigador experto obtener el análisis estadístico de la los datos. Por tanto, el análisis multivariante tiene lugar en ambos paquetes de IMB, pues los datos pueden explotarse y emplearse para la identificación de patrones desconocidos y conocimiento que no se considera en los modelos previos. No obstante, la minería de datos, puede hacerse sólo con *Modeler* o programas afines, ya que tiene las herramientas para tratar con varias bases de datos y sobre todo de optimizar los datos, sin una hipótesis u objetivo precisos, pero sí con un planteamiento que guíe el análisis.

Por último, se puede concluir que los softwares analizados proporcionan la oportunidad a investigadores de diversas áreas del conocimiento de explorar ambas plataformas en las que se emplean diversas herramientas en la búsqueda de modelos de predicción. Las herramientas podrían ser complementarias, ya que las investigaciones que inician con una base de datos creada a través de indicadores que tienen su origen en la operacionalización de las variables que se considera tienen dependencia en su relación podrían considerarse para analizarlas a través de *Modeler* en una nueva búsqueda de conocimiento, en conjunto con otras bases de datos relacionadas con el contexto que se analiza.

Referencias

Aranda, Y. R., & Sotolongo, A. R. (2013). Integración de los algoritmos de minería de datos 1R, Prism E ID3 a PostgreSQL. *Journal of Informa-*

tion Systems and Technology Management: JISTEM, 10(2), 389-406.
Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. Zanasi, A. (1998). Discovering Data Mining: From Concept to Implementation.
Castañeda, M. B., Cabrera A., Navarro, Y. & de Bries, W. (2010). *Procesamiento de datos y análisis estadísticos utilizando SPSS: Un libro práctico para investigadores y administradores educativos.* Brasil: Edipucrs.
Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0* Step-by-step data mining guide.
Di, Z., Yang, Y., Fu, Q., Lin, X., Jiang, S. (2013). Exploiting machine learning for predicting skeletal-related events in cancer patients with bone metastases. Proceedings - 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2013 6732633, pp. 31-35
Galán-Cortina V. (2016). Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno Universitario. (Bachelor's thesis) Universidad Carlos III de Madrid Escuela Politécnica Superior Ingeniería en informática.
Hernández J, Ramírez M.J. y Ferri, C. (2004). *Introducción a la Minería de Datos.* España: Ed. Pearson Educación, S.A.
IBM. (2012). *Manual CRISP-DM de IBM SPSS Modeler*. Recuperado de ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf
IBM. (2013). IBM SPSS Statistics 22 Core System guía del Usuario. Recuperado de ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/22.0/es/client/Manuals/IBM_SPSS_Statistics_Core_System_User_Guide.pdf
IBM. (2013). Manual del usuario del sistema básico de IBM SPSS Statistics 20. Recuperado de ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Core_System_Users_Guide.pdf
IBM. (2016a). Guía del usuario de IBM SPSS Modeler 18.0. Recuperado de ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/es/ModelerUsersGuide.pdf
IBM. (2016b). Guía del usuario de IBM SPSS Statistics 24 Core System. Recuperado de ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/24.0/es/client/Manuals/IBM_SPSS_Statistics_Core_System_User_Guide.pdf

Joshig. (2015). El poder que tienen los datos. *Portafolio.*
León, A., & Castellanos, O., & Vargas, F. (2006). Valoración, selección y pertinencia de herramientas de software utilizadas en vigilancia tecnológica. *Ingeniería e Investigación*, 26 (1), 92-102.
Lobaina, E. M. R., & Suárez, C.P.R. (2018). Resultados obtenidos en un proceso de minería de datos aplicado a una base de datos que contiene información bibliográfica referida a cuatro segmentos de la ciencia. *Journal of Information Systems and Technology Management : JISTEM,* 15, 1-11.
Lotfnezhad, H., Ahmadi, M., Roudbari, M., Sadoughi, F. (2015). Prediction of breast cancer survival through knowledge discovery in databases. Global journal of health science 7(4), 392-398.
Mejía, J. (2017). Las ciencias de la administración y el análisis multivariante. Proyectos de investigación, análisis y discusión de resultados. Tomo II. Las técnicas interdependientes. (1ra. Ed.). México: Universidad de Guadalajara.
Mejía, J. (2018). Análisis estadístico multivariante con SPSS para las Ciencias Económico Administrativas. *Teoría y Práctica de las Técnicas Dependientes*. México: D.R. Cloudbook.
Molina López, J. M., & García Herrero, J. (2006). *Técnicas de análisis de datos.* Universidad Carlos: Madrid.
Rodríguez Suárez, Y., Díaz Amador, A. (2011). Herramientas de minería de datos. *Revista Cubana de Ciencias Informáticas*, 3 (3-4)
Rodríguez, D., Pollo-Cattaneo, M. F., Britos, P. V., & García-Martínez, R. (2010). Estimación Empírica de Carga de Trabajo en Proyectos de Explotación de Información. *In XVI Congreso Argentino de Ciencias de la Computación*.
Shao, Z., Liancheng, W., Han, Z. (2016). A fault line selection method for small current grounding system based on big data. *Asia-Pacific Power and Energy Engineering Conference*, pp. 2470-2474.
Soto Jaramillo, C. M. (2009). *Incorporación de técnicas multivariantes en un sistema gestor de bases de datos*. Universidad Nacional de Colombia.
U Fayyad, G. P.-S. (1996). *Data mining and knowledge discovery in databases: an overview, communications of acm*
Wang, Z., Wen, X., Lu, Y., Yao, Y., Zhao, H. (2016). Exploiting machine learning for predicting skeletal-related events in cancer patients with bone metastases. *Oncotarget* 7(11) 12612-12622.