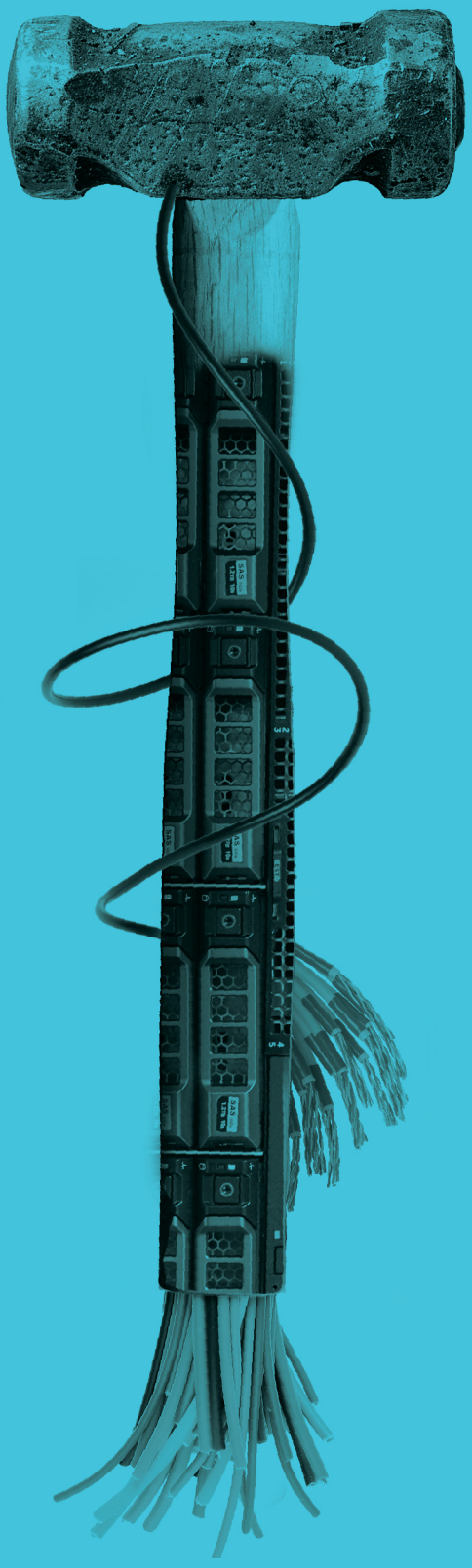
CIENCIAS NATURALES E INGENIERÍA

Datos a gran escala

un enfoque desde la minería de datos



Pedro Pablo González Pérez

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

MATEMÁTICAS APLICADAS Y SISTEMAS

COLECCIÓN LIBRO DE TEXTO

UNIVERSIDAD AUTÓNOMA
METROPOLITANA
Dr. José Antonio De los Reyes
Heredia
Rector General

Dra. Norma Rondero López
Secretaria General

Mtro. Octavio Mercado González
Rector de la Unidad Cuajimalpa

Dr. Gerardo Francisco Kloss
Fernández del Castillo
Secretario de la Unidad

Dra. Ma. Dayanira I. García Toledo
Coordinadora de Cultura

Lic. Gabriela E. Lara Torres
Jefa de Proyecto Editorial de Unidad

Esta obra es una de las ganadoras de la “Convocatoria para libros de texto como apoyo en la impartición de los programas de estudios”, 2024. Fue evaluada para su publicación por el Consejo Editorial de la UAM Unidad Cuajimalpa, con base en los dictámenes solicitados a pares académicos mediante un esquema que preserva el anonimato mutuo. Estos dictámenes resultaron favorables.

Diseño de portada: Aldo Juárez Herrera
Corrección de estilo: Ariana Cuadros Pedral

Universidad Autónoma Metropolitana
Unidad Cuajimalpa
Av. Vasco de Quiroga 4871, col. Santa Fe
Alcaldía Cuajimalpa de Morelos
C.P. 05348, Ciudad de México

www.cua.uam.mx

ISBN: 978-607-28-3298-5
ISBN: 978-607-28-3020-2 (Colección)

Contenido

I. INTRODUCCIÓN

En el contexto actual de la tecnología de la información, el término “datos a gran escala” o “datos masivos” (*big data* en inglés) se ha difundido de manera vertiginosa, ganando una enorme popularidad en muy pocos años (Aguilar, 2013; Marr, 2016; Marz y Warren, 2015; Mayer-Schönberger y Cukier, 2017). Es casi imposible encontrar en el mundo digital algún espacio en el cual este término no tenga cabida. Cuando nos referimos a “datos a gran escala” o “datos masivos”, estamos indicando los enormes y complejos conjuntos de datos que día a día son producidos por múltiples fuentes digitales.

En las últimas décadas, los avances tecnológicos y la creciente digitalización de una gran parte de nuestras actividades cotidianas han impulsado la producción de datos a gran escala, provenientes de una amplia gama de fuentes en diversos dominios. Algunas de las principales fuentes generadoras son: 1) comercio electrónico (*e-commerce*), 2) Internet, motores de búsqueda y redes sociales, 3) mercado financiero, 4) sensores, 5) simulaciones computacionales, 6) sistemas computarizados para el cuidado y monitoreo de la salud, 7) dispositivos móviles y aplicaciones (*apps*), y 8) plataformas digitales de entretenimiento (Aguilar, 2013; Marr, 2016; Mayer-Schönberger y Cukier, 2017).

La importancia de los datos a gran escala se pone de manifiesto en la valiosa ayuda que proporcionan para la toma de decisiones estratégicas de medianas y grandes corporaciones, empresas y negocios. Esto incluye datos estructurados y no estructurados, que se diferencian por su volumen, velocidad, variedad y valor. Los grandes volúmenes de datos se pueden procesar y analizar utilizando métodos, tecnologías y herramientas provenientes de la minería de datos, lo que nos permite generar información que contenga conocimiento crucial para la toma de decisiones en el área en cuestión.

La finalidad del presente material es compartir con el lector aspectos claves sobre las fuentes de producción, las características, la importancia, la preparación y el análisis de los datos a gran escala, así como su papel en la generación de valiosa información y conocimiento que apoyen la toma de decisiones estratégicas en múltiples dominios de la vida cotidiana: desde niveles corporativos y gerenciales del mundo del comercio electrónico, los negocios y las finanzas, hasta en el día a día en el hogar y en el control y monitoreo de la salud personal.

Este material va dirigido, principal —aunque no exclusivamente—, a los alumnos de las licenciaturas en Ingeniería en Computación y en Matemáticas Aplicadas, quienes, desde dos perspectivas diferentes, poseen los conocimientos y

habilidades necesarios para la comprensión, preparación, modelado y análisis de los datos a gran escala.

Por su parte, los alumnos de Ingeniería en Computación poseen conocimientos de métodos, técnicas y herramientas orientados a: 1) gestión de bases de datos, 2) modelos de datos, 3) técnicas de aprendizaje automatizado, tales como redes neuronales artificiales y árboles de decisión, 4) uso de librerías y componentes de aprendizaje automatizado disponibles en sitios web, 5) implementación de programas que permitan complementar actividades relacionadas con la preparación y análisis de los datos, entre otros.

En tanto, los alumnos de la licenciatura en Matemáticas Aplicadas tienen conocimientos y habilidades relacionados con: 1) métodos de regresión, 2) clasificación de datos, 3) técnicas para la selección de características relevantes, tales como el análisis de componentes principales, 4) uso de gráficos para la exploración de datos tabulares, 5) algoritmos de interpolación, entre otros.

A lo largo de cada uno de los capítulos y apartados de este material, se hace énfasis en los métodos, técnicas y herramientas comúnmente utilizados para contender con la preparación y análisis de estos grandes y complejos volúmenes de datos. De forma particular, este libro se centra en uno de los enfoques metodológicos de minería de datos más difundidos y utilizados en los últimos años: el CRISP-DM, por las siglas de Cross-Industry Standard Process for Data Mining (Shearer, 2000).

El capítulo II ofrece un amplio panorama introductorio a los datos a gran escala, enfatizando las propiedades claves que caracterizan, tales como volumen, velocidad, variedad, veracidad y valor. Este apartado se complementa con importantes aspectos relacionados con los datos a gran escala, tales como su importancia en la generación de nueva información y conocimiento, así como sus principales fuentes generadoras.

Por otra parte, en el capítulo III se discute el papel de los datos a gran escala en la inteligencia de negocios. Es decir, la transformación de los datos a información, la cual, a su vez, se convierte en conocimiento que se usa en la toma inteligente de decisiones. Como escenario particular, se describe y se discute ampliamente el comercio electrónico (*e-commerce*) como una de las principales fuentes generadoras de datos a gran escala, que basa gran parte de su toma de decisiones en el conocimiento derivado de estos grandes volúmenes de datos.

En tanto, el capítulo IV introduce los métodos, técnicas y herramientas de la minería de datos necesarios para contender con la preparación, modelado y

análisis de grandes volúmenes de datos. De forma particular, se centra en el CRISP-DM, uno de los enfoques metodológicos más difundidos y utilizados en los últimos años (Shearer, 2000).

En el capítulo V se describen y se discuten a detalle las fases que caracterizan la metodología de minería de datos CRISP-DM: comprensión del dominio del problema, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Cada una de ellas comprende un conjunto de actividades, las cuales son descritas, discutidas y, en algunos casos, ejemplificadas.

Asimismo, el capítulo VI presenta, describe y ejemplifica el uso de la herramienta de minería de datos IDA-WEB TOOL (González Pérez *et al*., 2023), desarrollada a nivel prototipo por alumnos de la Licenciatura en Ingeniería en Computación, durante sus Proyectos terminales I, II y III y a través del proyecto de servicio social. Esta herramienta constituye un valioso apoyo en la fase de preparación de los datos (de acuerdo con el enfoque metodológico CRISP-DM), al dar soporte en cada una de las actividades involucradas en ella. Por otra parte, también es un apoyo durante la fase de modelado, gracias a los modelos de aprendizaje automatizado y los métodos de regresión que pone a disposición del usuario.

Finalmente, el capítulo VII presenta dos interesantes casos de estudio que permiten ilustrar tanto el uso de la metodología de minería de datos CRISP-DM como el de la herramienta de minería de datos IBM SPSS MODELER. El primero está relacionado con la predicción del incremento en las ventas de bienes de consumo, a partir de una promoción aplicada; el segundo, con la clasificación de clientes de crédito como clientes activos o clientes con desgaste.

i. RELACIÓN DEL CONTENIDO CON EL PROGRAMA DE ESTUDIO DE LA LICENCIATURA EN INGENIERÍA EN COMPUTACIÓN

Como se describe en la tabla i.1, el contenido de este material aborda temas clave que se ofrecen en las unidades de enseñanza aprendizaje (UEA) de las licenciaturas en Ingeniería en Computación y en Matemáticas Aplicadas que se mencionan a continuación, y cuyo contenido sintético se adjunta más adelante:

- UEA 4605006 Datos a gran escala
- UEA 4605007 Minería de datos

Tabla i.1. Relación del contenido del material con el programa de estudio de la Licenciatura en Ingeniería en Computación

Temática abordada en el material	UEA que favorece
Introducción a los datos a gran escala	• UEA 4605006 Datos a gran escala
Los datos a gran escala como la base de la inteligencia de negocios: el comercio electrónico	• UEA 4605006 Datos a gran escala
Las técnicas de minería de datos para la preparación y análisis de los datos a gran escala	• UEA 4605006 Datos a gran escala • UEA 4605007 Minería de datos
Enfoques metodológicos de la minería de datos	• UEA 4605006 Datos a gran escala • UEA 4605007 Minería de datos
La herramienta de minería de datos IDA-WEB TOOL	• UEA 4605006 Datos a gran escala • UEA 4605007 Minería de datos

CONTENIDO SINTÉTICO DE LA UEA 4605006 DATOS A GRAN ESCALA

1. Introducción a los datos a gran escala (*big data*)
   1.1. Aproximaciones al concepto de datos a gran escala
   1.2. Características de los datos a gran escala: volumen, velocidad, variedad y veracidad
   1.3. Fuentes de producción de los datos a gran escala
   1.4. Principales aplicaciones de los datos a gran escala
2. Principales componentes tecnológicos de un sistema de datos a gran escala
   2.1. Tecnologías para la captura de datos a gran escala
   2.2. Tecnologías para el almacenamiento de datos a gran escala
   2.3. Tecnologías para el acceso a los datos a gran escala

2.4. Tecnologías para la visualización de los datos a gran escala
2.5. Tecnologías para el análisis de los datos a gran escala

3. Métodos y tecnologías para el análisis de los datos a gran escala
   3.1. Métodos estadísticos
   3.2. Consulta de bases de datos
   3.3. *Data warehouse*
   3.4. Minería de datos
   3.5. Aprendizaje automático
4. Casos de estudio
   4.1. Problemas de clasificación
   4.2. Problemas de predicción

CONTENIDO SINTÉTICO DE LA UEA 4605007 MINERÍA DE DATOS

1. Introducción a la minería de datos
   1.1. El proceso de la minería de datos
   1.2. Problemas que resuelve la minería de datos
   1.3. Representación del conocimiento
2. Métodos de regresión
   2.1. Problemas que se resuelven mediante regresión
   2.2. Regresión lineal
   2.3. Regresión lineal ponderada
   2.4. Regresión no lineal
3. Clasificación de datos
   3.1. Problemas que se resuelven mediante clasificación
   3.2. Regresión logística
   3.3. Redes de función base radial
   3.4. Máquinas de vectores de soporte
   3.5. Redes neuronales
   3.6. Clasificadores basados en reglas y árboles de decisión
4. Agrupamiento de datos (aprendizaje no supervisado)
   4.1. Aplicaciones de agrupamiento de datos
   4.2. Agrupamiento basado en representantes
   4.3. Agrupamiento jerárquico
   4.4. Ensambles
5. Preprocesamiento de datos
   5.1. Selección de características
   5.2. Extracción de características

ii. DESCRIPCIÓN DE LA IMPORTANCIA DE LOS CONOCIMIENTOS A ADQUIRIR, ASÍ COMO DE LAS HABILIDADES Y ACTITUDES A DESARROLLAR

El presente material pretende contribuir tanto a la adquisición de conocimientos como al desarrollo de habilidades de los alumnos de las licenciaturas en Ingeniería en Computación y en Matemáticas Aplicadas durante su recorrido por los planes de estudio homónimos, y, de forma muy particular, cuando cursan alguna de las siguientes UEA:

- 4605006 Datos a gran escala
- 4605007 Minería de datos

En la tabla ii.1, se describen los temas que aborda este libro y su relación con los conocimientos a adquirir, así como las habilidades a desarrollar para cada una de las UEA antes mencionadas. Cabe mencionar que, aunque ambas UEA forman parte del plan de estudios de la Licenciatura en Ingeniería en Computación, también se ofrecen como UEA optativas para los alumnos de la Licenciatura en Matemáticas Aplicadas.

Tabla ii.1. Temas que aborda el presente material y su relación con las UEA del plan de estudios de la Licenciatura en Ingeniería en Computación, los conocimientos a adquirir y las habilidades a desarrollar

Tema abordado en el material	UEA que favorece	Conocimientos a adquirir	Habilidades a desarrollar
Introducción a los datos a gran escala	• 4605006 Datos a gran escala	✓ El término “datos a gran escala”: importancia, características, fuentes generadoras, aplicaciones	✓ Conocer la importancia, las características, fuentes generadoras y principales aplicaciones de los datos a gran escala
Los datos a gran escala como la base de la inteligencia de negocios: el comercio electrónico	• 4605006 Datos a gran escala	✓ El papel de los datos a gran escala como la base de la inteligencia de negocios	✓ Conocer y argumentar los principales dominios donde los datos a gran escala podrían jugar un papel crucial en la toma inteligente de decisiones
Las técnicas de minería de datos para la preparación y análisis de los datos a gran escala	• 4605006 Datos a gran escala • 4605007 Minería de datos	✓ Técnicas de la minería de datos para la preparación y análisis de los datos a gran	✓ Conocer y aplicar las técnicas de la minería de datos para llevar a cabo las fases de preparación de los datos y

			escala		modelado de grandes volúmenes de datos
Enfoques metodológicos de la minería de datos	• 4605006 Datos a gran escala • 4605007 Minería de datos	✓	El enfoque metodológico de minería de datos CRISP-DM	✓	Conocer y aplicar las fases del enfoque metodológico de minería de datos CRISP-DM durante la preparación y modelado de grandes volúmenes de datos
La herramienta de minería de datos IDA-WEB TOOL	• 4605006 Datos a gran escala • 4605007 Minería de datos	✓	Herramientas que soportan las actividades de la minería de datos	✓	Conocer y utilizar herramientas de cómputo que proporcionan un soporte automatizado para la preparación y modelado de grandes volúmenes de datos

II. INTRODUCCIÓN A LOS DATOS A GRAN ESCALA (*BIG DATA*)

2.1. El término "datos a gran escala"

El término "datos a gran escala" (del inglés *big data*) ha ganado gran popularidad en el contexto actual de la tecnología de la información para referirse a los enormes y complejos conjuntos de datos producidas por múltiples fuentes digitales (Aguilar, 2013; Marr, 2016; Mayer-Schönberger y Cukier, 2013, 2017). Resulta difícil encontrar una definición ampliamente aceptada del término, ya que se ha utilizado de manera ubicua, y, en la gran mayoría de los casos, las aproximaciones dependen del área en la que se ha utilizado esta tecnología. Sin embargo, cuando nos referimos a "datos a gran escala" seguramente estamos describiendo problemas que surgen con relación a (Mayer-Schönberger y Cukier, 2017; Tolk, 2015):

- Datos cuyo volumen y complejidad requieren métodos más sofisticados de almacenamiento, recuperación, interacción, preparación, análisis-inferencia y presentación.
- Sistemas de *software* cuya funcionalidad resulta inadecuada para lidiar con el enorme volumen y la gran complejidad de los datos que deben procesar.
- Grandes volúmenes de datos, estructurados y no estructurados, lo que hace que su tratamiento sea mucho más complejo.
- La aplicación de un potente procesamiento computacional a conjuntos de datos altamente masivos y complejos.

El término "datos a gran escala" se pone de manifiesto siempre que los datos obtenidos resultan demasiado voluminosos para ser procesados por una aplicación o un sistema de cómputo. Por otra parte, este término también se revela cuando los sistemas de gestión o servidores de bases de datos no son capaces de proporcionar en tiempo razonable los datos requeridos, debido a problemas con la carga, búsqueda, selección, etcétera.

2.2. Principales características de los datos a gran escala

En los últimos años, la definición de datos a gran escala ha abarcado de tres a cinco dimensiones clave, a las que comúnmente se hace referencia como las "tres V" o las "cinco V" (ver figura 2.1), puesto que la definición original se caracteriza por volumen, velocidad y variedad, mientras que la definición ampliada incluye veracidad y valor (Mayer-Schönberger y Cukier, 2013; Tolk, 2015; Ward y Barker, 2013). A continuación, se define cada una de ellas:

- **Volumen**: Se refiere a la gran cantidad de datos que se generan, recopilan y analizan constantemente; esta variable suele medirse en *gigabytes*, *terabytes* y *petabytes*. El volumen es precisamente la característica que más se asocia a los datos masivos; es imposible no pensar en él cuando nos referimos a *big data*.

- **Velocidad**: Se refiere a la rapidez con la cual los datos son generados, recopilados y procesados. En una gran variedad de dominios, el tiempo de respuesta se convierte en una variable esencial para su uso; este es el caso de los sistemas de cómputo que deben ofrecer respuesta en tiempo real. Pensemos también en la velocidad con la que se generan datos en aplicaciones como los motores de búsqueda, el mercado de valores, las plataformas *e-commerce* y las redes sociales, por mencionar algunos ejemplos.

- **Variedad**: Se refiere a la no homogeneidad o diversidad de los datos, ya que provienen de fuentes diversas. Esto implica que los datos sean de diferente naturaleza, clasificándose comúnmente en dos grandes categorías: estructurados y no estructurados. Los datos estructurados se refieren a datos tabulares, los cuales incluyen valores numéricos, booleanos, categóricos y nominales, y texto estructurado. Por otra parte, como su nombre lo indica, los datos no estructurados se refieren a aquellos que carecen de una estructura, tales como textos no estructurados, presentaciones en visualizadores particulares, fotos, imágenes, videos, y archivos de texto como *e-mails*, PDF, blogs, sitios web, entre otros.

- **Veracidad**: Se refiere a la necesidad de enfrentarse a la incertidumbre contenida en los datos, derivada, principalmente, de la gran variedad que generan las diferentes fuentes. El problema de la veracidad en los datos a gran escala se presenta comúnmente en los datos de texto no estructurado —generados por redes sociales, foros de discusión, correo electrónico, etcétera—, debido a la libertad que caracteriza su creación.

- **Valor**: Este término se refiere a la importancia y el significado que los datos a gran escala pueden proporcionar a empresas, compañías e instituciones en la toma de decisiones a partir de los modelos de predicción y clasificación (principalmente algoritmos de regresión y *machine learning*) basados en los datos existentes, lo que las conduce a ser mucho más rentables y exitosas.

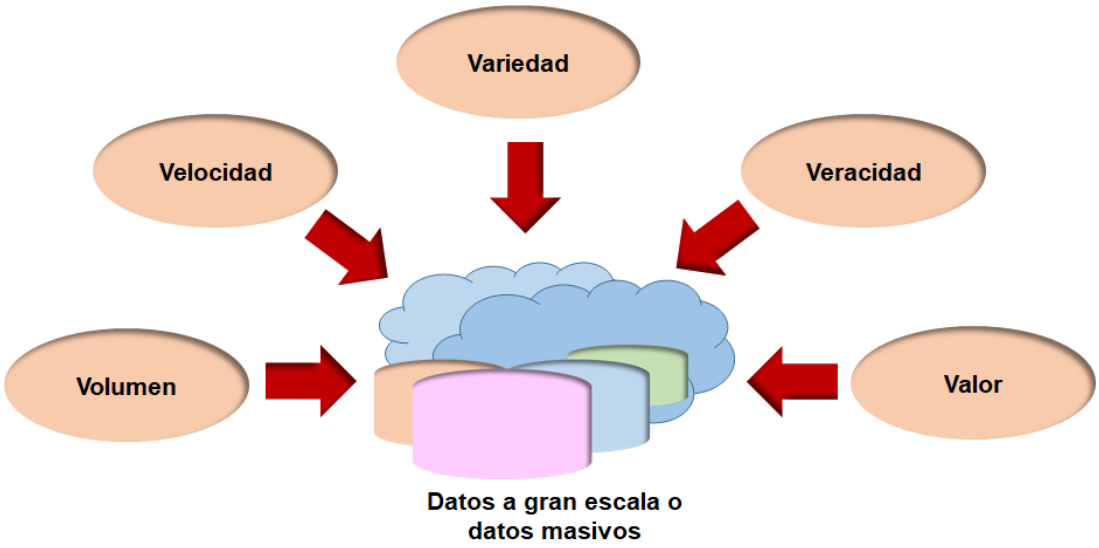


Figura 2.1. Principales características de los datos a gran escala.

2.3. Importancia de los datos a gran escala

La importancia de los datos a gran escala radica en su potencial para proporcionar enormes volúmenes de información, a partir de la cual se pueden tomar valiosas decisiones estratégicas a nivel corporativo, empresarial, académico, o cualquier otro ámbito (Marr, 2016; Mayer-Schönberger y Cukier, 2017). Los métodos, las tecnologías y herramientas para preparar y analizar los datos a gran escala permiten procesar datos estructurados y no estructurados —caracterizados por su volumen, velocidad, variedad y valor—, y producir nuevos datos derivados, los cuales contienen información valiosa para la toma de decisiones.

Con la aplicación de herramientas avanzadas de análisis de datos —tales como técnicas de aprendizaje automatizado y modelos estadísticos— las organizaciones pueden identificar clases, categorías, patrones, tendencias, asociaciones y correlaciones, que, de otro modo, no podrían ser observadas.

La toma de decisiones basada en el análisis inteligente de datos resulta de gran relevancia en dominios como el comercio electrónico, el mercado bursátil, la gestión bancaria, financiera y crediticia, la optimización de procesos, entre otros. En resumen, la importancia de los datos a gran escala se observa en su capacidad para transformar datos brutos en información, que, a su vez se convierte en conocimiento valioso para fundamentar la toma inteligente de decisiones.

2.4. Fuentes generadoras de datos a gran escala

Los avances tecnológicos y la creciente digitalización de una gran parte de

nuestras actividades, en las últimas décadas, han sido el motor que ha impulsado la producción de datos a gran escala, provenientes de una vasta gama de fuentes en diversos dominios (ver figura 2.2). Entre las principales fuentes generadoras se encuentran las siguientes:

- Comercio electrónico (*e-commerce*)
- Internet, motores de búsqueda y redes sociales
- Mercado financiero
- Datos producidos por sensores
- Simuladores computacionales
- Sistemas computarizados para el cuidado y monitoreo de la salud
- Dispositivos móviles y aplicaciones (*apps*)
- Plataformas digitales de entretenimiento
- Gobierno y sector público

Figura 2.2. Principales fuentes generadoras de datos a gran escala.

2.4.1. El comercio electrónico (*e-commerce*)

Las plataformas *e-commerce* son una extensa gama de aplicaciones de comercio electrónico disponibles en Internet, que funcionan como punto de venta de productos y servicios en línea, y a las cuales el usuario comúnmente accede a

través de una URL o desde una *app* en el dispositivo celular. Como se puede apreciar en la figura 2.3, entre las plataformas *e-commerce* más usadas en México se encuentran: Amazon, Mercado Libre, AliExpress, Walmart, Coppel, Liverpool, Sam's Club, The Home Depot, entre otras.

En el próximo capítulo se profundizará en el papel del *e-commerce* como fuente generadora de datos a gran escala, así como en el tipo de datos que se producen a partir de la navegación, interacción, intereses y compras efectuadas por los usuarios, y el gran valor que posee dicha información, una vez procesada y analizada, en la toma de decisiones de la empresa o compañía.

Figura 2.3. Plataformas *e-commerce* más populares en México.

2.4.2. Internet, los motores de búsqueda y las redes sociales

Sin lugar a dudas, una de las principales fuentes generadoras de datos a gran escala proviene de la proliferación que el uso del Internet ha tenido desde mediados de los años noventa, así como los motores de búsqueda y las redes sociales; esto ha provocado la generación de enormes volúmenes de datos producidos a una velocidad inimaginable. Como se ilustra en la figura 2.4, al navegar por la web, visitar sitios específicos, utilizar los motores de búsqueda, interactuar con contenidos en línea, participar en foros de discusión, publicar textos, fotos, audios o videos, entre muchas otras actividades digitales, los usuarios generan enormes volúmenes de datos de forma continua.

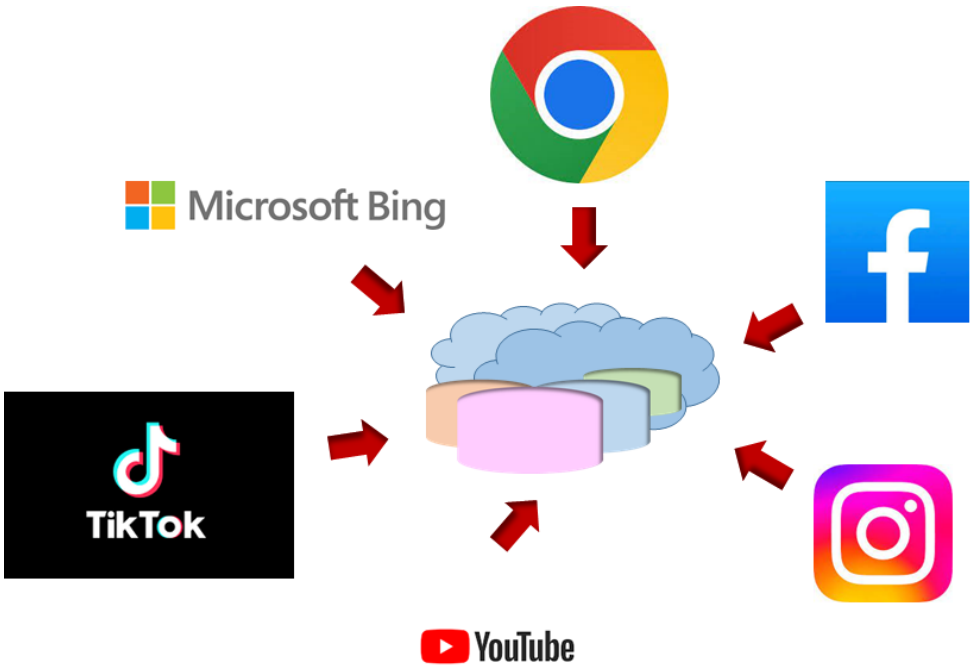


Figura 2.4. Internet, los motores de búsqueda y las redes sociales como principales fuentes de generación continua de enormes volúmenes de datos.

2.4.3. El mercado financiero

El inmenso volumen de datos que, de forma continua y a velocidad vertiginosa, proporciona el mercado bursátil (ver figura 2.5), principalmente a través de las bolsas de valores, es un insumo muy apreciado para mejorar la toma de decisiones de los inversionistas. Estos grandes volúmenes de datos son preprocesados y analizados a través de técnicas y modelos de minería de datos, para producir inferencias, predicciones y estimaciones. De esta forma, la información que genera el mercado bursátil juega un papel relevante en la era de los datos a gran escala; los volúmenes de datos masivos que produce de forma continua es tan importante que entre el 15% y el 50% de los ingresos de las bolsas de valores en el mundo provienen de la venta de esta valiosa información. Para ilustrar lo anterior, la tabla 2.1 muestra un fragmento de un conjunto de datos relacionados con el comportamiento de un grupo de acciones del "Índice Dow Jones", el cual incluye: acción (*stock*), precio de apertura, precio de cierre, precio mínimo, precio máximo, volumen de acciones transferidas, entre otros. El conjunto de datos, nombrado "Dow_Jones_Index_Dataset" se encuentra disponible en el UC Irvine Machine Learning Repository (Dua y Graff, 2019).

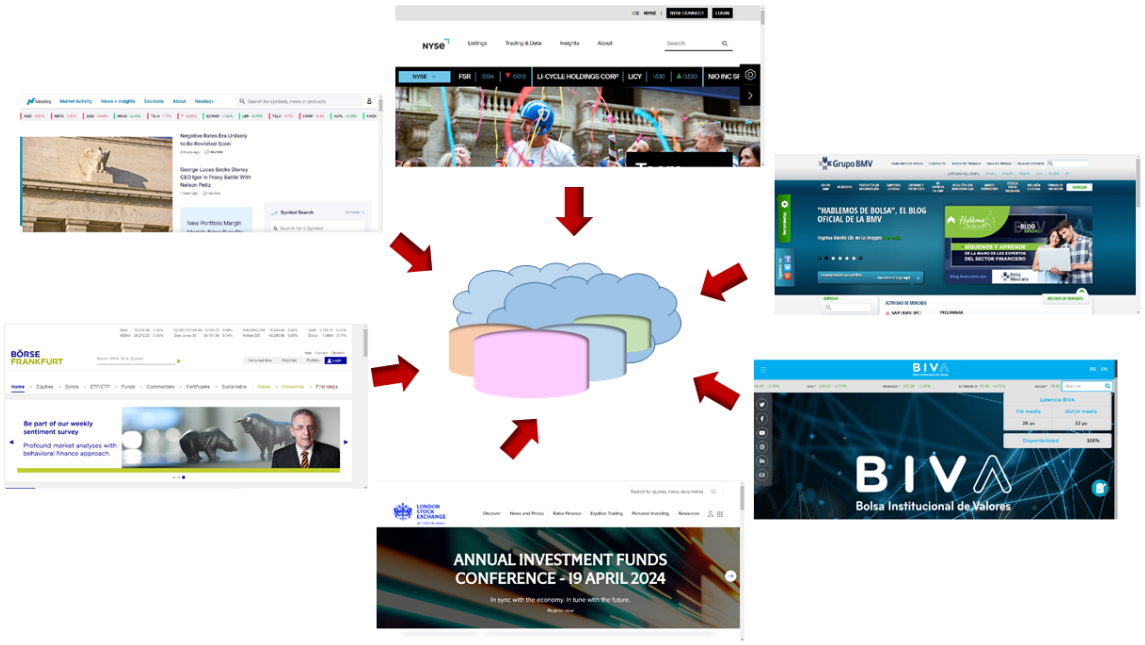


Figura 2.5. El mercado financiero —las principales bolsas de valores en el mundo— constituye uno de los principales generadores de enormes volúmenes de datos en fracciones de segundo y de forma continua.

Tabla 2.1. Conjunto de datos que describen el comportamiento de un conjunto de acciones del "Índice Dow Jones"

	stock	date	open	high	low	close	volume	percent_change_price	next_weeks_open	next_weeks_close	percent_change_next_weeks_price	days_to_next_dividend	percent_retur
2	AA	07/01/2011	15.82	16.72	15.78	16.42	239655616	3.793	16.71	15.97	-4.428	26	
3	AA	14/01/2011	16.71	16.71	15.64	15.97	242963398	-4.428	16.19	15.79	-2.471	19	
4	AA	21/01/2011	16.19	16.38	15.6	15.79	138428495	-2.471	15.87	16.13	1.638	12	
5	AA	28/01/2011	15.87	16.63	15.82	16.13	151379173	1.638	16.18	17.14	5.933	5	
6	AA	04/02/2011	16.18	17.39	16.18	17.14	154387761	5.933	17.33	17.37	0.231	97	
7	AA	11/02/2011	17.33	17.48	16.97	17.37	114691279	0.231	17.39	17.28	-0.633	90	
8	AA	18/02/2011	17.39	17.68	17.28	17.28	80023895	-0.633	16.98	16.68	-1.767	83	
9	AA	25/02/2011	16.98	17.15	15.96	16.68	132981863	-1.767	16.81	16.58	-1.368	76	
10	AA	04/03/2011	16.81	16.94	16.13	16.58	109493077	-1.368	16.58	16.03	-3.317	69	
11	AA	11/03/2011	16.58	16.75	15.42	16.03	114332562	-3.317	15.95	16.11	1.003	62	
12	AA	18/03/2011	15.95	16.33	15.43	16.11	130374108	1.003	16.38	17.09	4.335	55	
13	AA	25/03/2011	16.38	17.24	16.26	17.09	95550392	4.335	17.13	17.47	1.985	48	
14	AXP	07/01/2011	43.3	45.6	43.11	44.36	45102042	2.448	44.2	46.25	4.638	89	
15	AXP	14/01/2011	44.2	46.25	44.01	46.25	25913713	4.638	46.03	46	-0.065	82	
16	AXP	21/01/2011	46.03	46.71	44.71	46	38824728	-0.065	46.05	43.86	-4.756	75	
17	AXP	28/01/2011	46.05	46.27	43.42	43.86	51427274	-4.756	44.13	43.82	-0.702	68	
18	AXP	04/02/2011	44.13	44.23	43.15	43.82	39501680	-0.702	43.96	46.75	6.347	61	
19	AXP	11/02/2011	43.96	46.79	43.88	46.75	43746998	6.347	46.42	45.53	-1.917	54	
20	AXP	18/02/2011	46.42	46.93	45.53	45.53	28564910	-1.917	44.94	43.53	-3.138	47	
21	AXP	25/02/2011	44.94	45.12	43.01	43.53	39654146	-3.138	43.73	43.72	-0.023	40	
22	AXP	04/03/2011	43.73	44.68	42.75	43.72	38985037	-0.023	43.86	44.28	0.958	33	

Hoja1

2.4.4. Datos producidos por sensores

La proliferación de dispositivos físicos que reciben y transfieren datos a través de redes inalámbricas (Internet de las cosas; IoT, por sus siglas en inglés) en diversos entornos —incluidos hogares, fábricas, vehículos e infraestructuras— constituye, sin lugar a dudas, otra de las principales fuentes generadoras de datos a gran escala. Como se pueda apreciar en la figura 2.6, estos dispositivos físicos, comúnmente sensores, recopilan continuamente datos sobre:

- Temperatura
- Humedad
- Presión
- Movimiento
- Ubicación
- Otros parámetros ambientales

Lo anterior conlleva a la generación de grandes volúmenes de datos en tiempo real y, como consecuencia, a la disponibilidad de los mismos para su procesamiento, análisis y toma de decisiones.

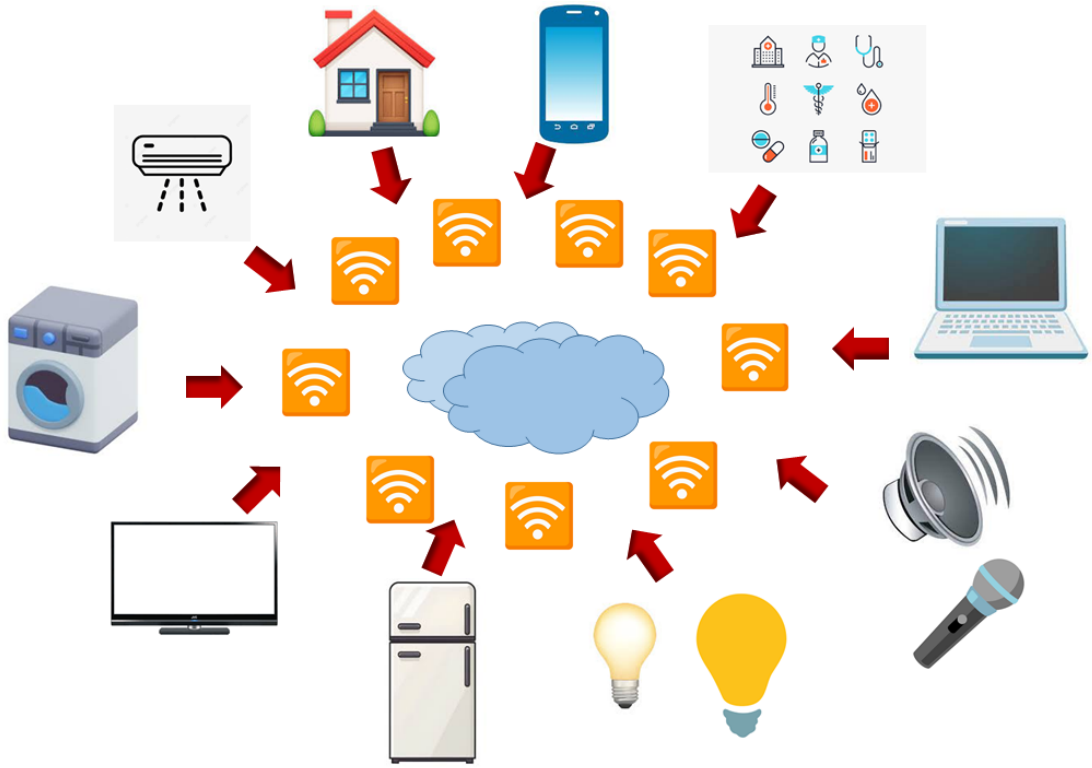


Figura 2.6. La gran cantidad de dispositivos conectados a través del Internet de las cosas (*Internet of Things*) constituyen una importante fuente de generación de grandes volúmenes de datos.

La tabla 2.2 muestra un fragmento de un conjunto de datos relacionados con la contaminación atmosférica; está integrado por 15 características y 9358 registros. Los datos fueron recopilados por cinco sensores químicos de óxidos de metal, integrados en un dispositivo multisensor químico de calidad del aire, que se ubicó en un área significativamente contaminada, al nivel de la carretera, de la periferia

de una ciudad italiana. Los datos se recopilaron durante un año, desde marzo de 2004 hasta febrero de 2005.

Tabla 2.2. Conjunto de datos relacionados con la contaminación atmosférica en una región de Italia

	Date	Time	CO(GT	PT08.S1(C	NMHC(GT)	C6H6(GT)	PT08.S2(NMH	NOx(GT	PT08.S3(NO	NO2(GT)	PT08.S4(NO	PT08.S5(O		R
2	10/03/2004	18:00:00	2.6	1360	150	11.9	1046	166	1056	113	1692	1268	13.6	48.9
3	10/03/2004	19:00:00	2	1292	112	9.4	955	103	1174	92	1559	972	13.3	47.7
4	10/03/2004	20:00:00	2.2	1402	88	9.0	939	131	1140	114	1555	1074	11.9	54.0
5	10/03/2004	21:00:00	2.2	1376	80	9.2	948	172	1092	122	1584	1203	11.0	60.0
6	10/03/2004	22:00:00	1.6	1272	51	6.5	836	131	1205	116	1490	1110	11.2	59.6
7	10/03/2004	23:00:00	1.2	1197	38	4.7	750	89	1337	96	1393	949	11.2	59.2
8	11/03/2004	00:00:00	1.2	1185	31	3.6	690	62	1462	77	1333	733	11.3	56.8
9	11/03/2004	01:00:00	1	1136	31	3.3	672	62	1453	76	1333	730	10.7	60.0
10	11/03/2004	02:00:00	0.9	1094	24	2.3	609	45	1579	60	1276	620	10.7	59.7
11	11/03/2004	03:00:00	0.6	1010	19	1.7	561	-200	1705	-200	1235	501	10.3	60.2
12	11/03/2004	04:00:00	-200	1011	14	1.3	527	21	1818	34	1197	445	10.1	60.5
13	11/03/2004	05:00:00	0.7	1066	8	1.1	512	16	1918	28	1182	422	11.0	56.2
14	11/03/2004	06:00:00	0.7	1052	16	1.6	553	34	1738	48	1221	472	10.5	58.1
15	11/03/2004	07:00:00	1.1	1144	29	3.2	667	98	1490	82	1339	730	10.2	59.6
16	11/03/2004	08:00:00	2	1333	64	8.0	900	174	1136	112	1517	1102	10.8	57.4
17	11/03/2004	09:00:00	2.2	1351	87	9.5	960	129	1079	101	1583	1028	10.5	60.6
18	11/03/2004	10:00:00	1.7	1233	77	6.3	827	112	1218	98	1446	860	10.8	58.4
19	11/03/2004	11:00:00	1.5	1179	43	5.0	762	95	1328	92	1362	671	10.5	57.9
20	11/03/2004	12:00:00	1.6	1236	61	5.2	774	104	1301	95	1401	664	9.5	66.8
21	11/03/2004	13:00:00	1.9	1286	63	7.3	869	146	1162	112	1537	799	8.3	76.4
22	11/03/2004	14:00:00	2.9	1371	164	11.5	1034	207	983	128	1730	1037	8.0	81.1
23	11/03/2004	15:00:00	2.2	1310	79	8.8	933	184	1082	126	1647	946	8.3	79.8
24	11/03/2004	16:00:00	2.2	1292	95	8.3	912	193	1103	131	1591	957	9.7	71.2
25	11/03/2004	17:00:00	2.9	1383	150	11.2	1020	243	1008	135	1719	1104	9.8	67.6
26	11/03/2004	18:00:00	4.8	1581	307	20.8	1319	281	799	151	2083	1409	10.3	64.2

AirQualityUCI

2.4.5. Simulaciones computacionales

Otra importante fuente que genera grandes volúmenes de datos, a través de las simulaciones computacionales y experimentación *in silico*, es la investigación científica. En las últimas tres décadas, estas prácticas se han extendido a la gran mayoría de las disciplinas científico-tecnológicas, destacando entre ellas las ciencias de la vida, la biología celular y molecular, la química, la física, las ciencias ambientales y varias áreas de la ingeniería. Los grandes volúmenes de datos producidos por la simulación computacional constituyen un valioso insumo que fomenta el avance y la innovación en las áreas mencionadas.

Las figuras 2.7 y 2.8 ilustran dos herramientas bioinformáticas de simulación computacional, las cuales generan grandes conjuntos de datos que permiten complementar la investigación teórica y guiar la investigación básica. La figura 2.7, en particular, ilustra la principal interfaz gráfica de la herramienta I-Foldameric, cuyo objetivo es el diseño, simulación y exploración del plegamiento de cadenas de aminoácidos (González Pérez *et al*., 2023).

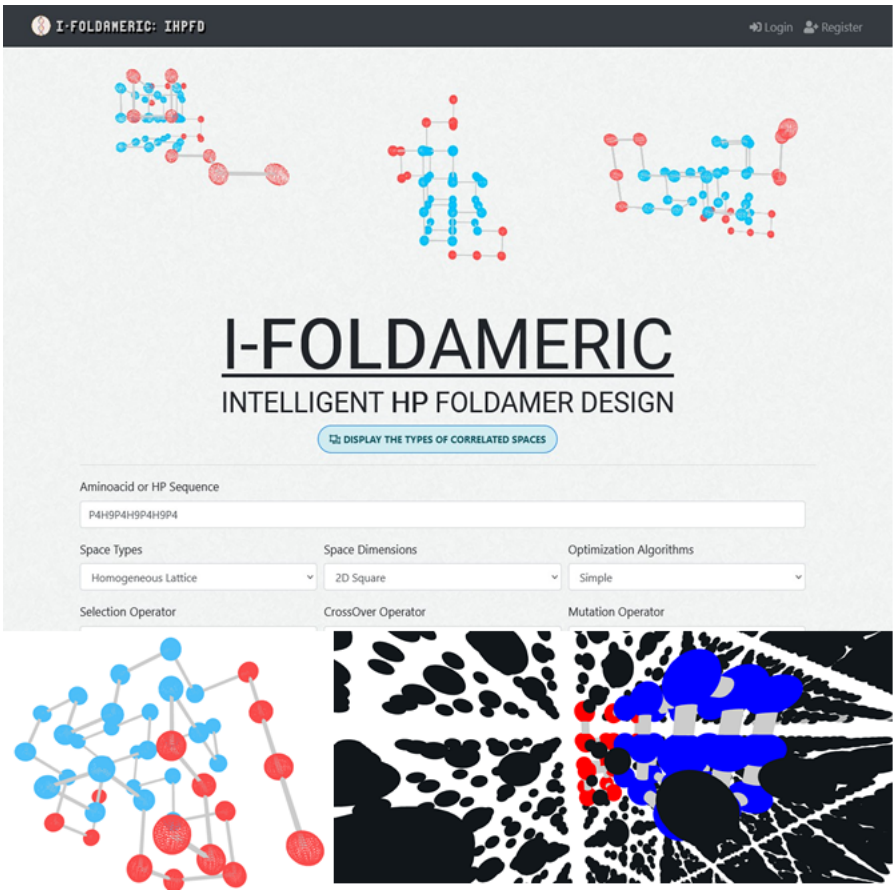


Figura 2.7. La simulación computacional como fuente generadora de grandes volúmenes de datos. Herramienta bioinformática I-Foldameric.

Por otra parte, en la figura 2.8 se puede apreciar la principal interfaz gráfica de la herramienta bioinformática Big Data–Cellulat, dedicada a la simulación y experimentación en vías de señalización celular (Cárdenas-García *et al*., 2016; Cárdenas-García y González-Pérez, 2018; González-Pérez y Cárdenas-García, 2018). Como su nombre lo indica, una de las principales características y valores de esta herramienta es la producción de grandes volúmenes de datos obtenidos a través de la experimentación *in silico*.

En tanto, la tabla 2.3 ilustra un fragmento del conjunto de datos generado por la herramienta de simulación computacional Big Data–Cellulat, durante la simulación del comportamiento de una red de señalización intracelular en células cancerosas (Cárdenas-García *et al*., 2016; Cárdenas-García y González-Pérez, 2018; González-Pérez y Cárdenas-García, 2018).

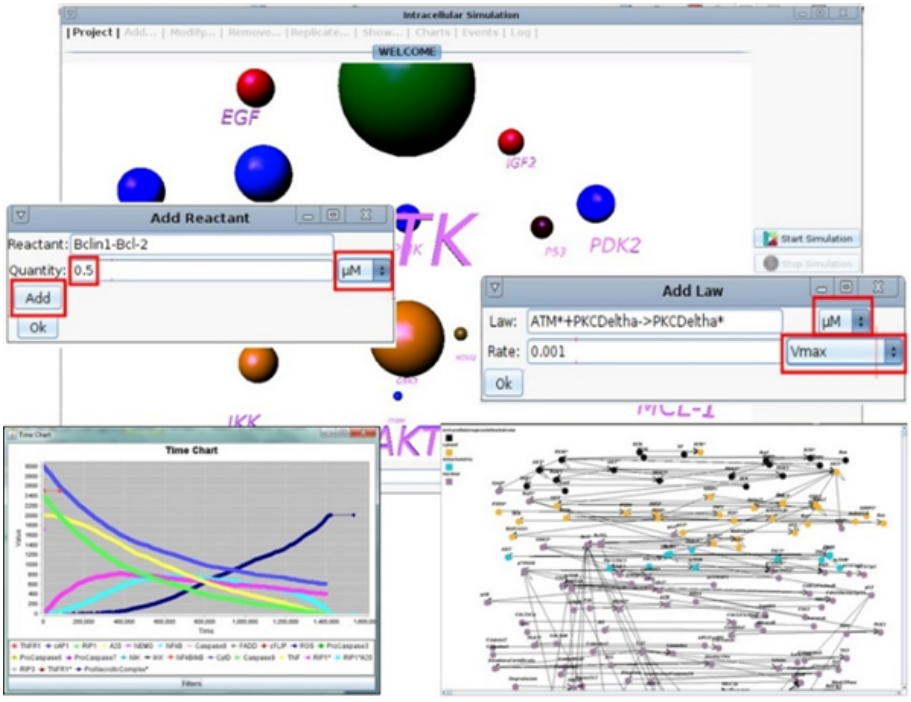


Figura 2.8. La simulación computacional como fuente generadora de grandes volúmenes de datos. Herramienta bioinformática Big-Data Cellulat.

Tabla 2.3. Conjunto de datos generado por la herramienta de simulación computacional Big Data–Cellulat, durante la simulación del comportamiento de una red de señalización intracelular en células cancerosas

CRAF*	RAS*	RAS	mTOR_RAPTOR	FKHR*	mTOR1*	FKHR_FOXO	PROLIFERATION	P21	SHC	P27*	PI3K*	RHEB*	PI3K	P27	TSC2*	AKT*	Cyt	PDK1	E1F4E	GAB1	TSC2	TSC1	GAB2	AUTOPHAGY
0	0	0.8	0.2	0.4	0	0	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.999	0.199	0.1	1	0.078	0.7	1	1	1	0
0	0	0.8	0.2	0.4	0	0	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.998	0.198	0.1	1	0.078	0.7	1	1	1	0
0	0	0.8	0.2	0.4	0	0	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.996	0.196	0.1	1	0.078	0.7	1	1	1	0
0	0	0.8	0.2	0.4	0	0	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.995	0.195	0.1	1	0.078	0.7	1	1	1	0
0	0	0.8	0.2	0.4	0	0	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.994	0.194	0.1	1	0.078	0.7	1	1	1	0
0	0	0.8	0.2	0.399	0	1	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.993	0.192	0.1	1	0.078	0.7	1	1	1	0
0	0	0.8	0.2	0.398	0	1	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.993	0.191	0.1	1	0.078	0.7	1	1	1	0
0	0	0.8	0.2	0.396	0	1	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.993	0.189	0.1	1	0.078	0.7	1	1	1	0
0	0	0.8	0.2	0.395	0	1	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.993	0.188	0.1	1	0.078	0.7	1	1	1	0
0	0	0.8	0.2	0.395	0	1	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.992	0.187	0.1	1	0.078	0.7	1	1	1	0
0	0	0.8	0.2	0.394	0	1	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.991	0.185	0.1	1	0.078	0.7	1	1	1	0
1	0	0.8	0.2	0.394	0	2	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.991	0.184	0.1	1	0.078	0.7	1	1	1	0
1	0	0.8	0.2	0.394	0	2	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.991	0.183	0.1	1	0.078	0.7	1	1	1	0
1	0	0.8	0.2	0.392	0	2	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.991	0.181	0.1	1	0.078	0.7	1	1	1	0
1	0	0.8	0.2	0.392	0	2	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.991	0.18	0.1	1	0.078	0.7	1	1	1	0
1	0	0.8	0.2	0.392	0	2	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.99	0.179	0.1	1	0.078	0.7	1	1	1	0
1	0	0.8	0.2	0.391	0	2	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.989	0.177	0.1	1	0.078	0.7	1	1	1	0
1	0	0.8	0.2	0.391	0	2	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.988	0.176	0.1	1	0.078	0.7	1	1	1	0
1	0	0.8	0.2	0.391	0	2	0	0.27	0.76	0.27	0	0.8	0.9	0.27	0.987	0.175	0.1	1	0.078	0.7	1	1	1	0
1	0	0.8	0.2	0.39	0	2	0	0.27	0.759	0.27	0	0.8	0.9	0.27	0.987	0.174	0.1	1	0.078	0.7	1	1	1	0
1	0	0.8	0.2	0.39	0	2	0	0.27	0.758	0.27	0	0.8	0.9	0.27	0.987	0.174	0.1	1	0.078	0.7	1	1	1	0

Hoja1

2.4.6. Sistemas computarizados para el cuidado y monitoreo de la salud

El bienestar de las personas y el cuidado y monitoreo de la salud son otras de las áreas que en los últimos años se han caracterizado por un crecimiento exponencial en la generación de datos masivos, a través de:

- Monitoreo y transmisión en tiempo real de datos personales generados durante la actividad deportiva, así como aquellos relacionados con el estado de salud y bienestar. En este caso, los sensores vienen en forma de bandas, pulseras y relojes.
- Monitoreo de pacientes a través de aplicaciones basadas en sensores instalados en el teléfono celular.

Para ejemplificar lo anterior, basta mencionar la amplia gama de dispositivos portátiles (principalmente del tipo *smart band*) que las personas pueden llevar consigo en forma de bandas, pulseras, relojes, etcétera, para el monitoreo de aspectos como: nivel de glucosa, ritmo cardíaco, presión arterial, número de pasos caminados en un intervalo de tiempo, calorías quemadas, calidad del sueño durante la noche, entre otras características asociadas a la salud y al bienestar.

2.4.7. Dispositivos móviles y aplicaciones

Otra importante fuente de producción de datos a gran escala proviene del continuo auge del uso de dispositivos móviles y las aplicaciones que hospedan (ver figura 2.9). Los vastos volúmenes de datos recopilados por estas tecnologías son producto de:

- Las interacciones de los usuarios
- El seguimiento de la ubicación
- El uso de aplicaciones
- Las preferencias de los usuarios
- Los diferentes tipos de sensores ya disponibles en estos dispositivos

La información recopilada contiene un gran valor para los operadores de telefonía móvil, los desarrolladores de aplicaciones y las plataformas publicitarias, quienes la utilizan para elaborar perfiles de usuarios, optimizar aplicaciones y realizar *marketing* dirigido.

Figura 2.9. Los dispositivos móviles y aplicaciones (*apps*) como fuente generadora de grandes volúmenes de datos a gran escala.

2.4.8. Plataformas digitales de entretenimiento

Cada día, grandes volúmenes de datos se generan a través del uso de plataformas digitales de entretenimiento, incluyendo videos en *streaming* —como es el caso de Netflix, Prime, y Claro Video (ver figura 2.10)—, videojuegos, música, libros electrónicos, entre otras. Al usar este tipo de plataformas, los usuarios generan información relacionada con sus preferencias, tiempo que les dedican, interacciones con los contenidos, frecuencia de uso, entre otros aspectos. La información recopilada es utilizada por las mismas plataformas para maximizar las sugerencias de contenido y personalizar las experiencias de los usuarios.

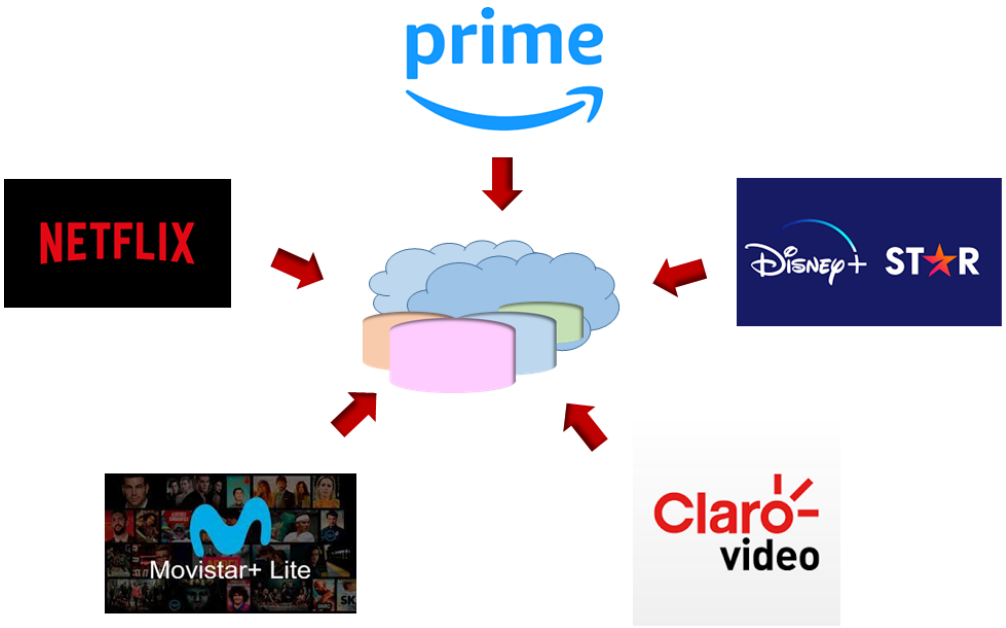


Figura 2.10. Las plataformas digitales de entretenimiento generan grandes volúmenes de datos a partir de los diferentes tipos de interacciones del usuario, tales como número de visitas, preferencias, tiempo que les dedican, etcétera.

2.4.9. Gobierno y sector público

El Gobierno y el sector público son otras importantes fuentes de producción de datos masivos, producidos a través de registros y estadísticas de salud pública, como en el caso de la pandemia por COVID-19, las encuestas políticas, los censos de población y viviendas, los sistemas de información geográfica del INEGI, las imágenes producidas por cámaras de seguridad, los sistemas de transporte público, entre otros (ver figura 2.11). El Gobierno y las entidades públicas, en sus diferentes niveles, utilizan estos grandes volúmenes de información para obtener información y conocimiento como guía en la ejecución de tareas y acciones de su competencia.

Para ejemplificar la generación de datos por parte del Gobierno y el sector público, la tabla 2.4 muestra un fragmento del enorme volumen de información generada por la Secretaría de Salud en México, durante la pandemia por COVID-19, conformado por 67,383 registros y 46 atributos.

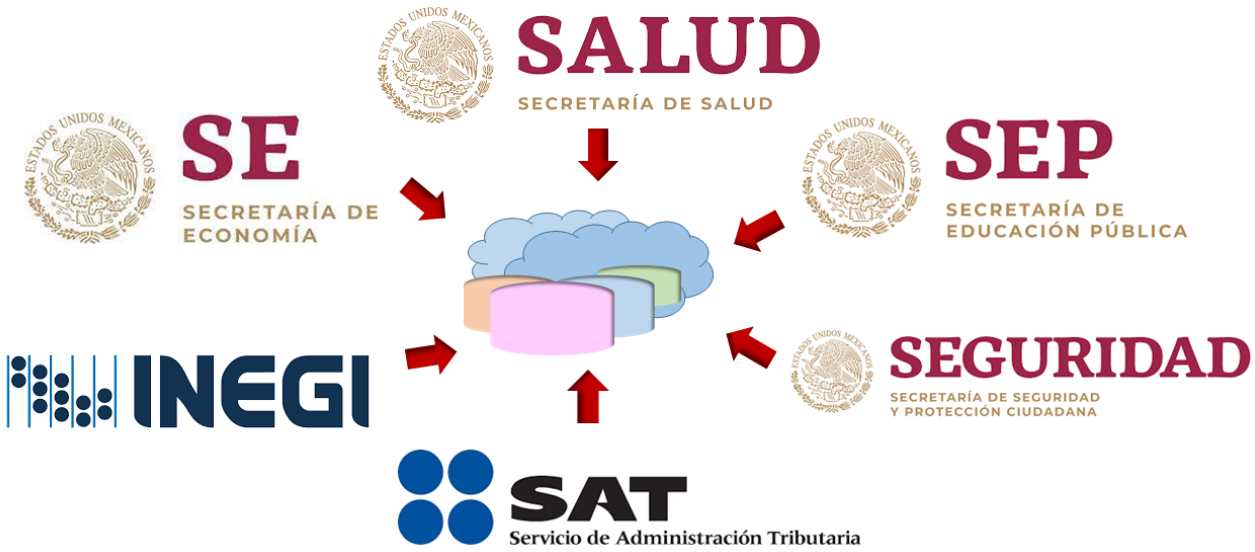


Figura 2.11. El Gobierno y las dependencias públicas fungen también como fuentes generadoras de datos, gran parte de los cuales se encuentran disponibles en línea para su consulta y procesamiento.

Tabla 2.4. Conjunto de datos generado por la Secretaría de Salud en México durante la pandemia por COVID-19

	FECHA INGRESO	FECHA SINTOMAS	FECHA DEFUNCION	INTUBADO	NEUMONIA	EDAD	NACIONALID	EMBARAZO	HABLA LENG	DIABETES	EPOC	ASMA	INMUNOSUP	HIPERTENSI
2	16/06/2020	02/06/2020	19/06/2020	SI	NO	68	MEXICANA	NO	NO	NO	NO	NO	NO	SI
3	06/04/2020	01/04/2020	NO APLICA	NO APLICA	NO	43	MEXICANA	NO	NO	NO	NO	NO	NO	SI
4	03/06/2020	28/05/2020	NO APLICA	NO	SI	56	MEXICANA	NO	NO	SI	NO	NO	NO	NO
5	23/06/2020	11/06/2020	NO APLICA	NO APLICA	NO	47	MEXICANA	NO	NO	NO	NO	NO	NO	NO
6	06/04/2020	02/04/2020	NO APLICA	NO APLICA	NO	24	MEXICANA	NO	NO	NO	NO	NO	NO	NO
7	21/05/2020	14/05/2020	NO APLICA	NO APLICA	SI	42	MEXICANA	NO APLICA	NO	NO	NO	NO	NO	NO
8	02/05/2020	30/04/2020	NO APLICA	SI	SI	41	MEXICANA	NO	NO	NO	NO	NO	NO	NO
9	12/04/2020	05/04/2020	NO APLICA	NO APLICA	NO	22	MEXICANA	NO	NO	NO	NO	NO	NO	NO
10	11/05/2020	06/05/2020	NO APLICA	NO APLICA	NO	29	MEXICANA	NO	NO ESPECIFI	NO	NO	NO	NO	NO
11	20/04/2020	10/04/2020	NO APLICA	NO APLICA	NO	29	MEXICANA	NO	NO	NO	NO	NO	NO	NO
12	25/05/2020	25/05/2020	09/06/2020	NO	NO	48	MEXICANA	NO	NO	NO	NO	NO	NO	NO
13	27/04/2020	17/04/2020	NO APLICA	NO	NO	24	MEXICANA	NO APLICA	NO ESPECIFI	NO	NO	NO	NO	NO
14	02/07/2020	28/06/2020	NO APLICA	NO APLICA	NO	31	MEXICANA	NO APLICA	NO	NO	NO	NO	NO	NO
15	28/04/2020	18/04/2020	NO APLICA	NO	SI	41	MEXICANA	NO APLICA	NO	NO	NO	NO	NO	NO
16	25/04/2020	25/04/2020	29/04/2020	NO	NO	72	MEXICANA	NO	NO	NO	SI	NO	NO	SI
17	28/05/2020	25/05/2020	NO APLICA	NO APLICA	NO	37	MEXICANA	NO APLICA	NO	NO	NO	NO	NO	NO
18	16/05/2020	16/05/2020	NO APLICA	NO APLICA	NO	22	MEXICANA	NO APLICA	NO	SI	SI	NO	SI	SI
19	22/04/2020	20/04/2020	NO APLICA	NO APLICA	NO	30	MEXICANA	NO APLICA	NO	NO	NO	NO	NO	NO
20	22/04/2020	16/04/2020	NO APLICA	NO	SI	40	MEXICANA	NO APLICA	NO	SI	NO	NO	NO	NO
21	01/06/2020	01/06/2020	NO APLICA	NO APLICA	NO	29	MEXICANA	NO	NO	NO	NO	NO	NO	NO
22	13/04/2020	04/04/2020	NO APLICA	NO	NO	66	MEXICANA	NO APLICA	NO	NO	NO	NO	NO	SI

dataset20k

III. LOS DATOS A GRAN ESCALA COMO LA BASE DE LA INTELIGENCIA DE NEGOCIOS: EL COMERCIO ELECTRÓNICO

3.1. La toma inteligente de decisiones basada en el uso de los datos a gran escala

En los últimos años, el uso de métodos y técnicas para la recopilación, preparación y modelado de datos a gran escala ha cambiado la naturaleza del comercio y de los negocios. Esto se debe principalmente a la velocidad con la cual los datos generados por los propios usuarios de plataformas *e-commerce* y otros sitios web pueden ser recolectados, procesados y analizados, permitiendo que la toma de decisiones de las compañías, empresas e instituciones sea mucho más rentable y exitosa.

En la manufactura, agricultura o producción industrial, una vez que un producto ha sido elaborado o cultivado, es necesario proceder a la publicidad, colocación, venta y distribución del mismo. Es aquí donde los datos masivos recopilados por plataformas *e-commerce* y otros sitios web resultan de gran utilidad, al indicar al comercio y a los negocios quiénes podrían ser los clientes (compradores) potenciales de estos productos. Estos datos son comúnmente empleados para:

- Pronosticar las preferencias de los clientes
- Efectuar promociones, ofertas o descuentos
- Efectuar ventas cruzadas
- Determinar dónde exhibir un determinado producto
- Identificar cuáles serán los mejores puntos de venta de un producto en particular
- Rastrear los movimientos de los clientes, tanto en tiendas físicas como en línea

Hoy en día, la industria, la manufactura, la agricultura, el comercio y los negocios intentan aprovechar al máximo los grandes beneficios que proporciona la tecnología de datos masivos. Por tal motivo, estos sectores se han dado a la tarea de innovar en los métodos de recopilación, preparación y análisis de grandes volúmenes de datos, de forma que buscan avanzar en la predicción certera y anticipada de las necesidades, gustos y preferencias de los clientes. Es decir, aplican la inteligencia de negocios, la cual se basa en el principio de transformar los datos en información y la información en conocimiento, con la finalidad de optimizar la toma de decisiones (ver figura 3.1).

Por otra parte, el enfoque de datos a gran escala no sólo ha sido adoptado por estos sectores para incrementar sus ventas y beneficios, sino que también está siendo utilizado para la búsqueda de recursos humanos que posean la experiencia, los conocimientos y las habilidades que requieren los diferentes cargos de líderes,

ejecutivos y operativos en las empresas.

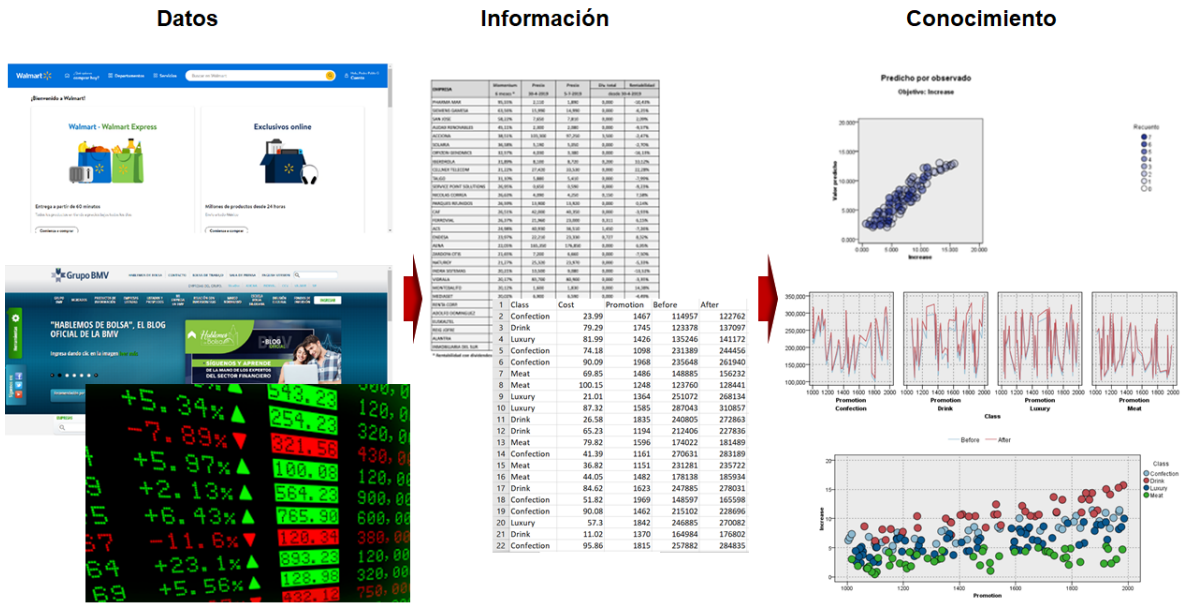


Figura 3.1. Inteligencia de negocios: transformar los datos en información y la información en conocimiento, con la finalidad de optimizar la toma de decisiones.

Para convertir un negocio en una empresa inteligente (*smart business*) y lograr mayor rentabilidad, éxito y posicionamiento en el mercado, no sólo es necesario tener acceso a los grandes volúmenes de datos creados por los visitantes, usuarios o clientes de la empresa, también es esencial utilizar una metodología, modelo o enfoque de minería de datos que oriente, de manera organizada y estructurada, la preparación, análisis y presentación de la información.

3.2. Las plataformas de comercio electrónico (*e-commerce*)

Como se ha mencionado, el término *e-commerce* se refiere a una amplia variedad de herramientas de comercio electrónico disponibles en Internet, a las cuales el usuario comúnmente accede mediante una URL o una aplicación en su dispositivo celular. Un punto de venta en línea destinado al comercio de bienes y servicios es conocido como plataforma *e-commerce*.

Generalmente, un sistema de comercio electrónico engloba componentes de *hardware*, *software*, comunicaciones (comúnmente el Internet), y un conjunto de herramientas en línea que le permiten al negocio:

- Publicar sus productos o servicios
- Proporcionar al consumidor la navegación, búsqueda y selección de

productos o servicios
- Proporcionar al consumidor la adquisición y pago de los productos o servicios

Entre las principales ventajas que trae consigo el uso del *e-commerce*, podemos resaltar las siguientes (Martínez Reyes, 2021; Mendoza Becerril, 2021):

- La utilización de la web permite que las empresas y los clientes interactúen directamente, eliminando intermediarios y garantizando una entrega instantánea, lo que mejora la distribución de productos o servicios. Gracias a que el Internet es la principal vía de comunicación, este medio llega a puntos donde antes no era posible, abarcando así un mercado más amplio en términos de territorio.
- Reduce considerablemente los costos de la transacción, eliminando formularios para pedidos, cotizaciones y otros procedimientos. Al no ser necesario procesar todo este tipo de información, la operación de compraventa a nivel del *e-commerce* reduce significativamente los costos.
- Garantiza tanto al negocio como al consumidor el fácil acceso a la información, ya que se puede ingresar a una gama de bases de datos que permite identificar productos, encontrar y colocar ofertas, crear mercados, acceder a mercados nuevos, etcétera.
- Mejora las relaciones entre el negocio y los consumidores, ya que se pueden establecer diferentes canales de comunicación en línea, a través de los cuales el consumidor puede plantear sus inquietudes, efectuar cualquier tipo de reclamo, obtener más información sobre los productos o servicios; mientras que, por parte del negocio, esta variedad de canales de comunicación permite atender las inquietudes de los consumidores, conocer sus preferencias y llegar a acuerdos que beneficien a ambas partes.
- Mejora el *marketing*, ya que el negocio informa de forma continua a sus clientes sobre los productos o servicios que ofrece, permitiendo que los clientes puedan acceder a éstos las 24 horas del día, los 365 días del año.

3.3. Las plataformas de *e-commerce* en México

En los últimos años el mercado electrónico en México ha mostrado un notable crecimiento, contando con varios actores importantes que dominan el sector. Entre las principales plataformas *e-commerce* dedicadas a la venta de productos al consumidor destacan:

- Amazon México (https://www.amazon.com.mx/)
- Mercado Libre (https://www.mercadolibre.com.mx/)

- Walmart (https://www.walmart.com.mx)
- Coppel (https://www.coppel.com/)
- Liverpool (https://www.liverpool.com.mx/)
- Palacio de Hierro (https://www.elpalaciodehierro.com/)
- Sam's Club (https://www.sams.com.mx/)
- The Home Depot México (https://www.homedepot.com.mx/)
- Office Depot México (https://www.officedepot.com.mx/)

Estos sitios de comercio electrónico en México atienden una amplia variedad de necesidades, demandas y preferencias de los consumidores, a la vez que recolectan de forma continua enormes volúmenes de datos generados por las interacciones, búsquedas, selecciones y adquisiciones de los usuarios. A medida que aumenta el comercio electrónico y la competencia en el mercado mexicano, la popularidad y el uso de estas plataformas continúan evolucionando para adaptarse a las exigencias de los consumidores.

3.4. Tipos de *e-commerce*

Dependiendo de quién proporciona el servicio y a quién va dirigido, se pueden identificar cuatro tipos de comercio electrónico (Martínez Reyes, 2021; Mendoza Becerril, 2021):

- *E-commerce* negocio a negocio
- *E-commerce* negocio a cliente
- *E-commerce* cliente a negocio
- *E-commerce* cliente a cliente

3.4.1. *E-commerce* negocio a negocio

Cuando una empresa intercambia bienes o servicios con otra, se produce el comercio electrónico negocio a negocio (B2B, del inglés *Business to Business*). Muchas compañías o empresas poseen sus propios sitios web para ofrecer este tipo de comercio electrónico (ver figura 3.2), en el que los proveedores pueden colaborar con sus clientes presentándoles su inventario, donde incluyen precios exclusivos para cada empresa con la que trabajan, lo que facilita la toma de decisiones de compra. Las transacciones comerciales negocio a negocio son similares a las que suelen ocurrir entre un fabricante y un distribuidor o entre un distribuidor y un comercio minorista.

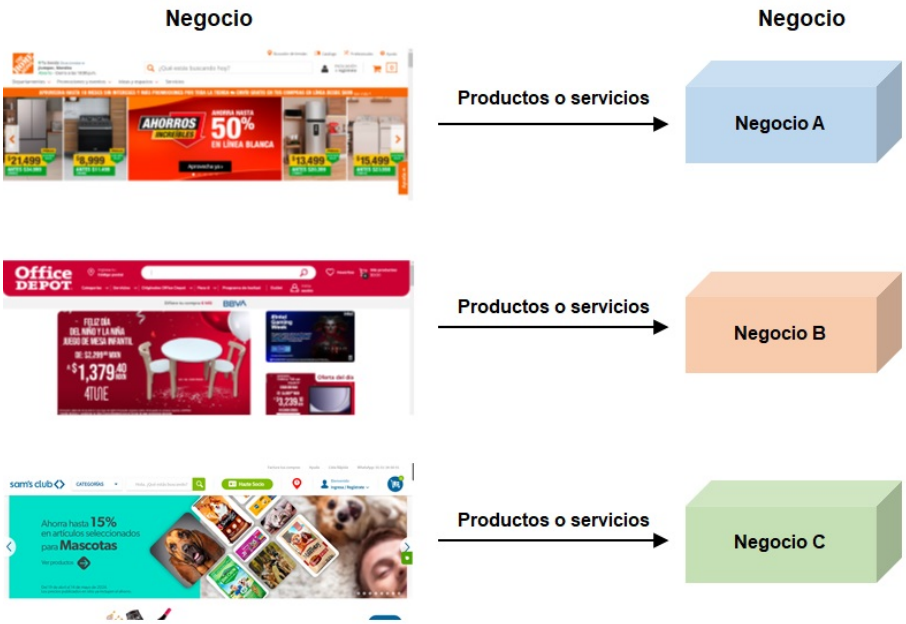


Figura 3.2. Comercio electrónico tipo negocio a negocio (C2C).

3.4.2. *E-commerce* negocio a cliente

El comercio electrónico negocio a cliente (B2C, del inglés *Business to Customer*) es el tipo de *e-commerce* más difundido y utilizado en la web. Se refiere al enfoque que emplean las empresas para llegar directamente a los clientes o consumidores finales (ver figura 3.3). Este comercio es el tipo de transacción que lleva a cabo una empresa cuando posee un vasto portafolio de clientes. Con él, las pequeñas y grandes empresas pueden mostrar sus catálogos de productos a sus clientes finales y vender directamente, por lo tanto, muchos distribuidores importantes utilizan su portal para comercializar a través de Internet. A este tipo de *e-commerce* pertenecen empresas líderes en México como: Walmart, Coppel, Liverpool, The Home Depot, Palacio de Hierro, Amazon, entre otras.

Figura 3.3. Comercio electrónico tipo negocio a cliente (B2C).

3.4.3. *E-commerce* cliente a negocio

Se refiere a la relación que se da entre el cliente y la empresa, caracterizada porque el cliente es quien inicia la operación de compraventa (ver figura 3.4). En este tipo de *e-commerce* (C2B, del inglés *Customer to Business*), el cliente o un grupo de clientes hacen una oferta a la empresa, generalmente a través de Internet, mostrando las características de sus productos o servicios, los precios asociados y alguna otra información de interés. De esta forma, los clientes ofertan bienes y servicios y las empresas pagan por ellos. Este modelo empresarial es una inversión del modelo tradicional (B2C), en el que las empresas ofrecen bienes y servicios a los consumidores.

Figura 3.4. Comercio electrónico tipo cliente a negocio (C2B).

3.4.4. *E-commerce* cliente a cliente

El tipo de *e-commerce* cliente a cliente (C2C, del inglés *Customer to Customer*) es un vínculo entre dos clientes, de los cuales cada uno es un consumidor final. Es una especie de oferta clasificada en línea, cuyo objetivo es facilitar la comercialización de productos y servicios entre particulares (ver figura 3.5). Por lo tanto, C2C es un modelo de negocio en la red que busca conectar a los usuarios con fines comerciales, donde la empresa sirve sólo como intermediaria y cobra por sus servicios, como sucede con eBay o Mercado Libre.

Figura 3.5. Comercio electrónico tipo cliente a cliente (C2C).

3.5. La recopilación de datos a través de las plataformas de comercio electrónico

Las plataformas de comercio electrónico tratan de recopilar la mayor cantidad posible de información sobre los datos personales, financieros, sociales, culturales, de sus clientes y usuarios, para satisfacer sus necesidades de cambio (mayor rentabilidad, mejor posicionamiento en el mercado, etcétera), de una forma mucho más eficaz.

Las plataformas de *e-commerce* producen volúmenes masivos de datos debido a la interacción que el usuario realiza a través de la navegación, las búsquedas, consultas y compras de los productos o servicios ofrecidos. Estos datos son un valioso insumo que, además de contener información personal y financiera de los usuarios, también reflejan sus preferencias por determinados productos, su capacidad y frecuencia de consumo, su nivel socioeconómico, etcétera.

El comercio electrónico produce una vasta gama de tipos de datos, entre los que destacan: datos del consumidor, datos de navegación, datos de transacciones efectuadas, datos de los producto*s* ofertados y datos relacionados con el propio sitio web. Comprender las diferentes categorías y los datos que genera a gran velocidad y volumen el *e-commerce* permitirá su uso adecuado en la toma de decisiones y optimización de procesos. En la tabla 3.1 se relacionan las principales categorías y tipos de datos producidos por el comercio electrónico.

Tabla 3.1. Categorías y tipos de datos producidos por el comercio electrónico

Categoría	Tipo de dato
Datos del consumidor	❖ Información personal: datos demográficos, información de contacto, de sus cuentas de acceso, preferencias, etcétera. ❖ Historial de compras: información sobre cada compra efectuada por el consumidor, incluyendo categorías de productos, compras repetidas, ventas cruzadas, patrones de compra, etcétera.
Datos de transacciones	❖ Datos relacionados con las órdenes de compra del consumidor. ❖ Datos relacionados con las facturas generadas por el consumidor. ❖ Datos relacionados con reembolsos y devoluciones solicitados por el consumidor.
Datos del producto	❖ Catálogo de productos: información acerca del producto, incluyendo identificador, nombre, descripción, categoría, precio, características, etcétera. ❖ Datos de inventario: niveles de inventario, tipos de productos, disponibilidad de existencias, avisos de

	reposición de existencias, pedidos pendientes, etcétera. ❖ Reseñas y calificaciones de productos: opiniones, comentarios, calificaciones, que genera el usuario acerca de un producto o servicio particular.
Datos de navegación en el sitio web generados por el usuario	❖ Visitas al sitio web del negocio: número de visitas efectuadas y tiempo de duración de cada una de ellas, patrones de navegación en el sitio web. ❖ Patrones de búsqueda: selección de productos y visualización de sus características sin efectuar compras, interés mostrado por el cliente acerca de un producto particular.
Datos sobre promociones, ofertas y descuentos	❖ Historial del negocio sobre promociones, ofertas y descuentos efectuados, así como el comportamiento de las ventas en dicho período.

3.6. El procesamiento de los datos recolectados

Los grandes volúmenes de datos que producen de forma continua las plataformas *e-commerce* constituyen un valioso insumo para los modelos de minería de datos que analizan dicha información, y que, como resultado, producen inferencias, predicciones y estimaciones de gran utilidad para la toma de decisiones en la venta de productos en línea.

Para procesar y analizar estos grandes volúmenes de datos, y obtener a partir de ellos información y conocimiento importante, se requiere aplicar métodos, técnicas y herramientas de la minería de datos; es decir, un enfoque metodológico que, a través de cada una de sus fases, permita transformar los datos en información y la información en conocimiento útil para la toma de decisiones que mejore el posicionamiento y la rentabilidad del negocio (ver figura 3.6).

Los siguientes capítulos de este material están dedicados a presentar, discutir y ejemplificar el enfoque de la minería de datos para la preparación, modelado, análisis e interpretación de grandes volúmenes de datos.

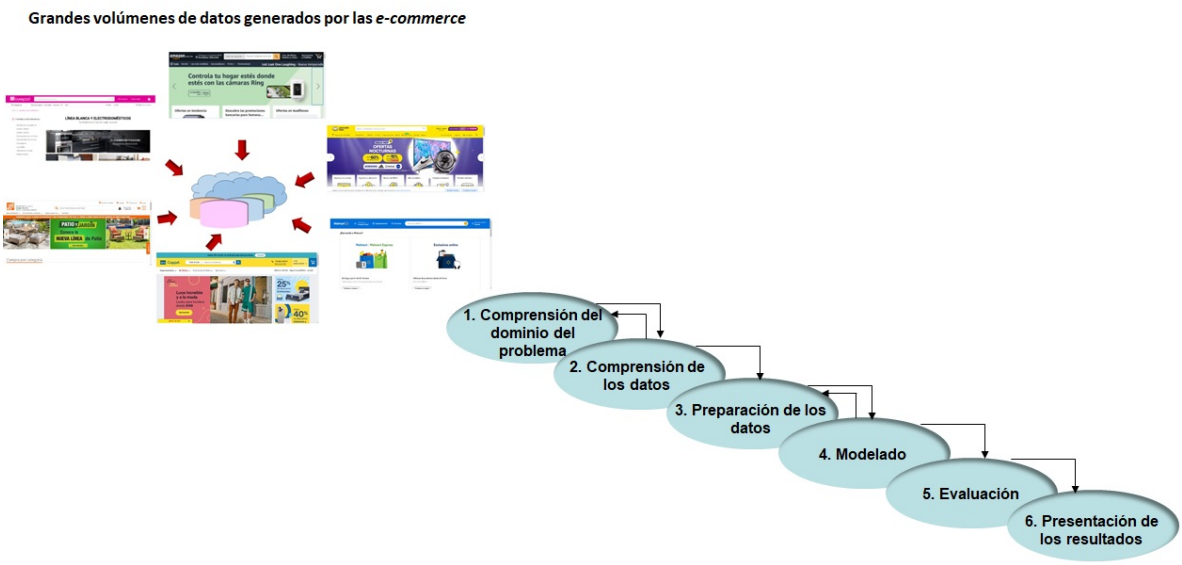


Figura 3.6. Aplicación del enfoque de minería de datos a los grandes volúmenes de datos generados por el *e-commerce*.

IV. EL ENFOQUE DE LA MINERÍA DE DATOS PARA LA PREPARACIÓN Y ANÁLISIS DE LOS DATOS A GRAN ESCALA

Para que un negocio sea inteligente (del inglés *smart business*), y, por lo tanto, logre una mayor rentabilidad, éxito y posicionamiento en el mercado, no basta con poseer el acceso a los grandes volúmenes de datos generados por los visitantes, usuarios o clientes del negocio, también es imprescindible basarse en una metodología, modelo o enfoque de minería de datos que guíe de forma organizada y estructurada la preparación, análisis y presentación de esta información.

4.1. La minería de datos

La minería de datos es un área que hereda y usa elementos de disciplinas como ciencias de la computación, inteligencia artificial, matemáticas y estadística (Aggarwal, 2015; Hernández Orallo, 2010; Tan *et al.*, 2018). Su objetivo principal es el descubrimiento de patrones, relaciones y correlaciones entre grandes volúmenes de datos. Es decir, los métodos y técnicas propios de esta área permiten procesar y analizar grandes volúmenes de datos, y, como resultado, producir nuevos datos derivados, los cuales contienen información y conocimiento de gran valor para la toma de decisiones basadas en los resultados de los modelos aplicados, comúnmente de correlación, predicción, clasificación, agrupamiento, reconocimiento de patrones, entre otros.

Al descubrir patrones, relaciones, correlaciones, inferencias y conocimientos ocultos en conjuntos de datos grandes y complejos, la minería de datos juega un papel importante para su preparación y análisis. Como ya se indicó, los datos a gran escala se caracterizan principalmente por su volumen, velocidad y variedad, al abarcar diferentes formatos y fuentes. De aquí que las técnicas de minería de datos resulten de gran utilidad para extraer información significativa y transformarla en conocimiento valioso para la toma de decisiones.

4.2. Fases de la minería de datos

La minería de datos consiste en la extracción de patrones, conceptos y conocimientos importantes de grandes conjuntos de datos. Aunque el flujo de trabajo puede variar según el dominio del problema y los objetivos del análisis que se llevará acabo, las principales fases o etapas involucradas en esta metodología son las siguientes (ver figura 4.1) (Aggarwal, 2015; Hernández Orallo, 2010; Shearer, 2000):

1. **Comprensión del dominio del problema**: Independientemente de la metodología o enfoque de minería de datos que se haya elegido, la comprensión del dominio del problema o negocio es crucial y tiene un gran impacto en las fases posteriores, debido a que es en esta fase donde se define claramente el problema que se intenta resolver. Se centra en comprender los objetivos y las metas del trabajo a desarrollar y se proporciona una perspectiva adecuada para comprender qué datos deben analizarse.
2. **Comprensión de los datos**: Consiste en la descripción y análisis inicial de los datos, con la finalidad de identificar si hay problemas de calidad (omisión, incompletitud, redundancia, falta de veracidad, etcétera) presentes, para así descubrir o proponer ideas iniciales acerca de los datos y establecer hipótesis de la información que describen.
3. **Análisis exploratorio de los datos (EDA, por sus siglas en inglés):** Implica la exploración y visualización de los datos para descubrir sus características, distribución, relaciones y posibles patrones contenidos entre ellos. Mediante el uso de técnicas EDA, tales como estadísticas descriptivas, visualización de datos, análisis de correlación y reducción de dimensionalidad, es posible descubrir patrones o tendencias preliminares en los datos y comprender su estructura básica.
4. **Preparación de los datos**: Como su nombre lo indica, en esta fase se preparan o se da forma a los datos, para construir el conjunto final que servirá como insumo para la construcción del modelo. De existir problemas de calidad en los datos, tales como omisión, errores de medición, errores de codificación, incompletitud, redundancia, falta de veracidad, etcétera, deben solucionarse en esta fase.
5. **Modelado**: Implica seleccionar los algoritmos o modelos de minería de datos apropiados para lograr el objetivo; estos pueden ser de tipo clasificación, regresión o agrupación. Entre los principales modelos y algoritmos encuentran las redes neuronales artificiales, los árboles de decisión y los métodos de regresión. Usualmente, cada modelo se ejecuta, al inicio, con los propios parámetros propuestos “por defecto”; posteriormente, estos parámetros pueden ser ajustados para ver si es posible mejorar los resultados que produce el modelo (bondad, error, coeficiente de correlación, etcétera).
6. **Evaluación de los modelos**: Una vez que los modelos han sido construidos (es decir, la etapa de entrenamiento del modelo ha concluido), necesitan ser validados y evaluados para conocer su rendimiento y capacidad de generalización sobre nuevos datos. Esta evaluación se realiza al final de la fase de entrenamiento del modelo, con el apoyo de importantes métricas como la matriz de confusión (para objetivos de

minería de datos del tipo clasificación), coeficiente de correlación (para objetivos de minería de datos del tipo regresión), exactitud, precisión, sensibilidad y F1 Score.

7. **Análisis de los resultados**: Los conocimientos y descubrimientos de la minería de datos deben interpretarse y comunicarse de manera efectiva a las partes interesadas una vez que los modelos fueron construidos, validados y evaluados. Esto implica explicar la importancia de los patrones descubiertos, las relaciones y correlaciones identificadas entre los datos, las predicciones o clasificaciones resultantes, así como las implicaciones para el problema de estudio y recomendaciones prácticas basadas en los resultados del análisis.

Estas fases se discutirán a detalle más adelante, en el capítulo V, dedicado al enfoque metodológico CRISP-DM (Shearer, 2000).

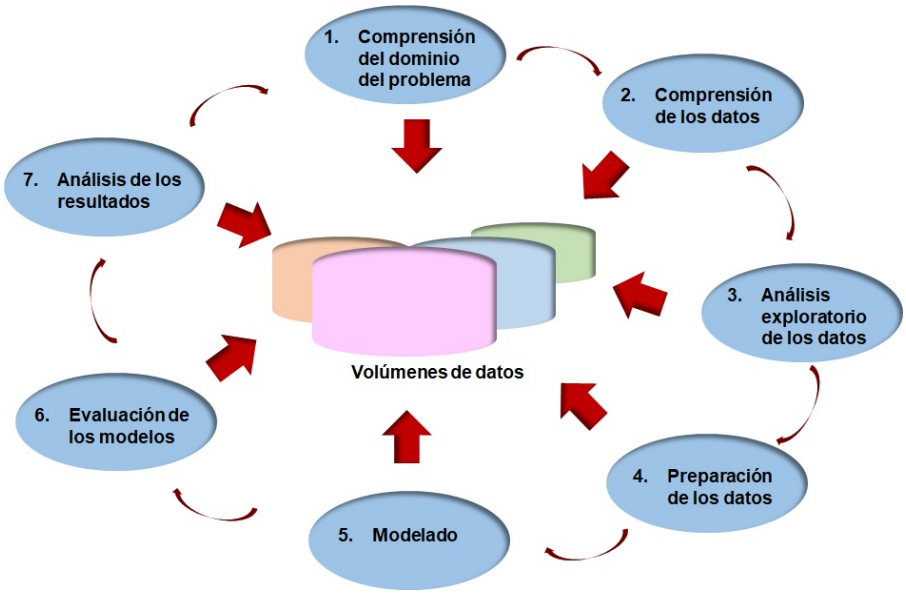


Figura 4.1. Principales fases o etapas involucradas en la minería de datos.

4.3. Métodos de la minería de datos

De manera común, los términos “método de minería de datos” y “técnica de minería de datos” son usados indistintamente, ya sea para hacer referencia a métodos de la minería de datos para extraer información valiosa (clasificación, regresión, o agrupamiento) o para referir las técnicas que permiten la implementación de estos métodos (redes neuronales artificiales supervisadas y no supervisadas, árboles de decisión, métodos de regresión, entre otros).

Tomando esto en consideración, y para evitar ambigüedades, en este material utilizaremos el término “método de minería de datos” para referirnos a los métodos de extracción de información, mientras que reservaremos el término “técnica de minería de datos” para hacer referencia a técnicas de inteligencia artificial, aprendizaje automatizado, estadísticas o matemáticas, que permiten la implementación de dichos métodos.

Entre los principales métodos utilizados en la minería de datos para extraer información de grandes conjuntos, podemos identificar los siguientes, los cuales serán descritos en los apartados contiguos:

- Clasificación
- Regresión
- Agrupamiento (*clustering*)
- Minería de reglas de asociación
- Minería de secuencias
- Minería de textos

4.3.1. Clasificación

En la minería de datos, la clasificación es un método utilizado para catalogar datos en clases o categorías predefinidas. Su aplicación requiere que para cada registro (fila) del conjunto de datos exista un atributo indicando la clase o categoría (etiqueta) a la cual pertenece. Este método también es conocido como clasificación supervisada, ya que el algoritmo o técnica que le corresponde aprende de los datos de entrenamiento etiquetados.

La clasificación supervisada se caracteriza comúnmente por dos fases de trabajo: la fase de entrenamiento o aprendizaje y la fase de generalización (Cerulli, 2023; Suthaharan, 2015). Durante la fase de entrenamiento o aprendizaje, el algoritmo se adiestra utilizando los datos etiquetados (ver figura 4.2). Posteriormente, durante la fase de generalización, el modelo construido es utilizado para clasificar datos nuevos o desconocidos en las categorías apropiadas (ver figura 4.3).

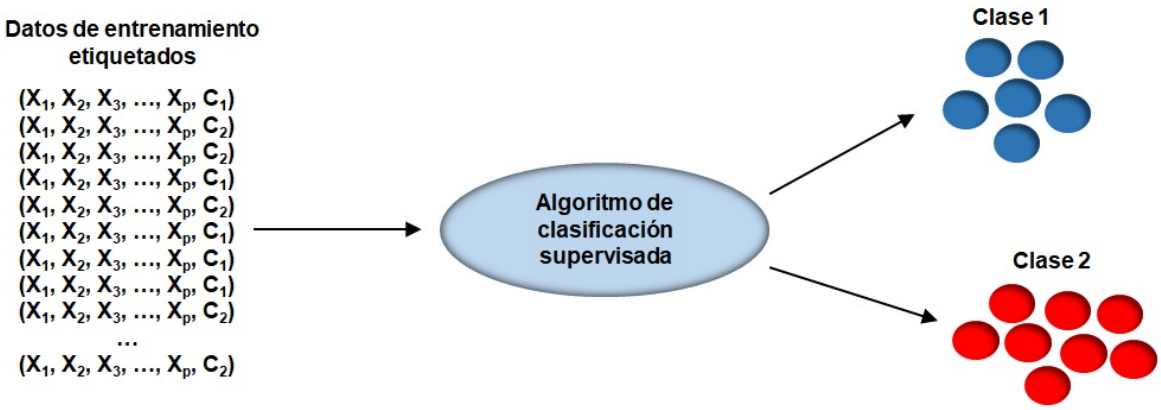


Figura 4.2. Fase de entrenamiento o aprendizaje de la clasificación supervisada.

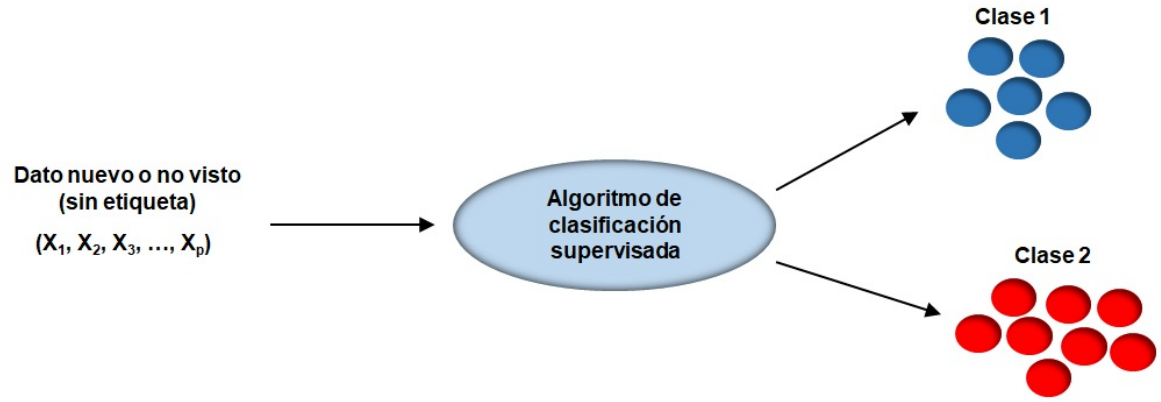


Figura 4.3. Fase de generalización de la clasificación supervisada.

Asimismo, la clasificación supervisada puede ser binaria o multi-clase:

- **Clasificación binaria**: Como su nombre lo indica, la clasificación binaria se refiere a que un dato sólo puede ser incluido en una de dos clases o categorías definidas (ver figura 4.4). Un ejemplo de esta clasificación es el diagnóstico médico basado en casos positivos y negativos, donde cada registro de un individuo o paciente tiene asociado un atributo que indica si presenta (caso positivo) o no (caso negativo) la enfermedad.
- **Clasificación multi-clase**: La clasificación supervisada multi-clase tiene lugar cuando un dato puede ser clasificado en dos o más clases o categorías (ver figura 4.5). Un ejemplo es la categorización de clientes de instituciones crediticias de acuerdo con su historial de pagos, donde pueden ser clasificados como: cliente al corriente en sus pagos en todas sus cuentas (clase 1), cliente con atraso de 1 a 89 días en al menos una de sus cuentas (clase 2), cliente con atraso superior a 90 días en al menos una de sus cuentas (clase 3), y cliente con al menos una cuenta sin recuperar (clase 4).

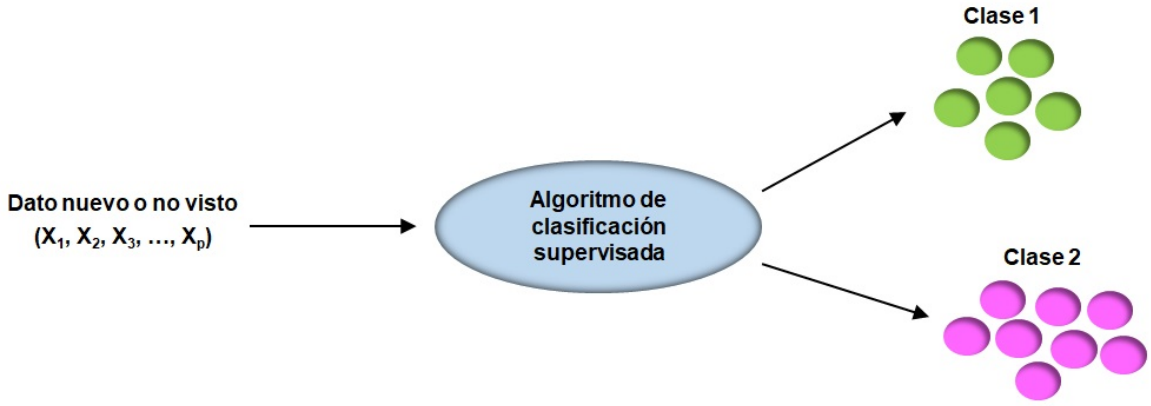


Figura 4.4. Clasificación supervisada binaria.

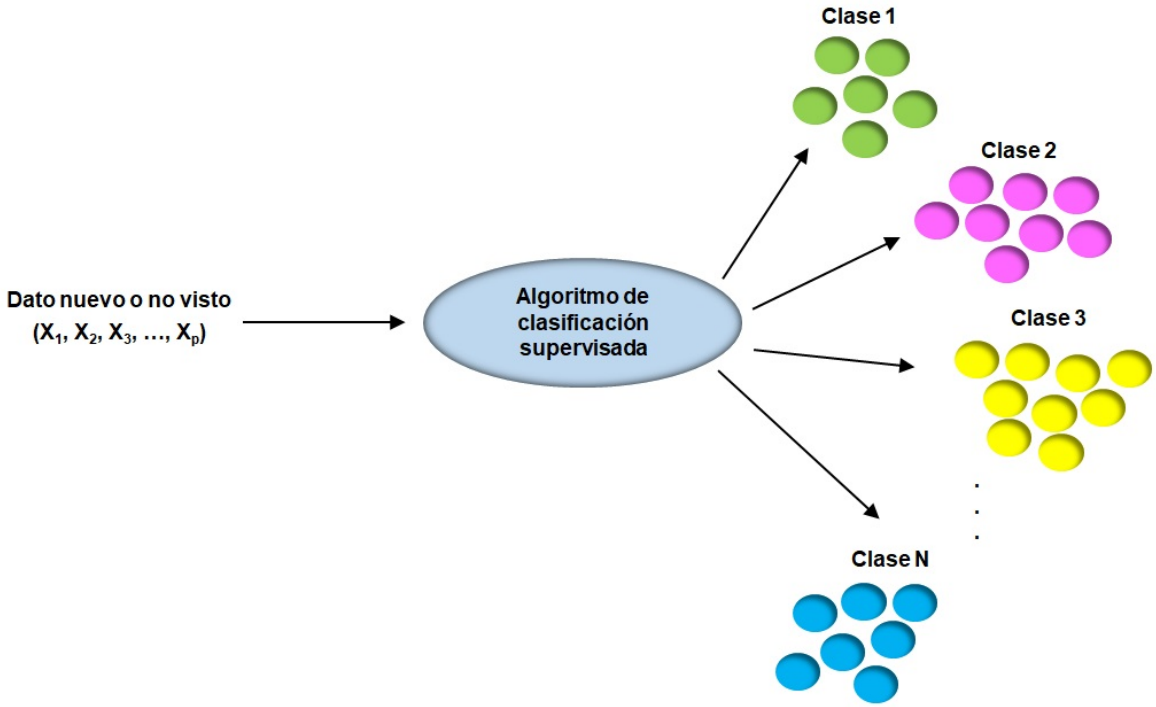


Figura 4.5. Clasificación supervisada multi-clase.

4.3.2. Regresión

La regresión es un método estadístico utilizado en la minería de datos para modelar y evaluar la relación entre una o más variables independientes, conocidas como predictores o características, y una variable dependiente, comúnmente llamada variable objetivo o variable a predecir (Chatterjee y Simonoff, 2013; Young, 2017). El análisis de regresión tiene como objetivo determinar cómo se asocian los cambios en las variables independientes con los cambios en la variable dependiente, y hacer

predicciones basadas en esta relación. Algunos ejemplos de problemas de predicción en minería de datos son los siguientes:

a) Predicción del incremento semanal del precio de las acciones en el mercado bursátil.
b) Predicción del incremento de las ventas en un *e-commerce*, a partir de la aplicación de una promoción.
c) Predicción de la variación de la inflación de los precios al consumidor.

De acuerdo con la forma en que interactúan las variables entre sí, se pueden identificar los siguientes tipos de modelos o técnicas de regresión (ver figura 4.6):

- **Regresión lineal**: Se emplea cuando una ecuación lineal puede describir aproximadamente la relación entre las variables independientes y la variable dependiente. La regresión lineal trata de encontrar la línea de mejor ajuste que minimice las variaciones entre los valores predichos y los observados (ver figura 4.6a).
- **Regresión no lineal**: Cuando las relaciones que se establecen entre las variables independientes y la variable dependiente no pueden ser descritas mediante una ecuación lineal, se pueden utilizar funciones no lineales, las cuales permiten modelar de forma mucho más flexible (ver figura 4.6b).
- **Regresión polinomial**: Al igual que en la regresión lineal y en la no lineal, en la polinomial el problema consiste en describir la relación entre las variables independientes y la variable dependiente; en este caso, se ajusta una curva polinómica a los datos, lo que permite capturar relaciones más complejas entre las variables (ver figura 4.6c).
- **Regresión logística**: A diferencia de los modelos de regresión antes descritos, en los cuales el objetivo es predecir el valor de la variable dependiente (variable continua) a partir de las variables independientes, la regresión logística no está asociada a problemas de predicción sino a problemas de clasificación. Es decir, la variable dependiente es una variable categórica que comúnmente toma valores binarios. La regresión logística modela la probabilidad de que la variable dependiente pertenezca a una de dos categorías (ver figura 4.6d).

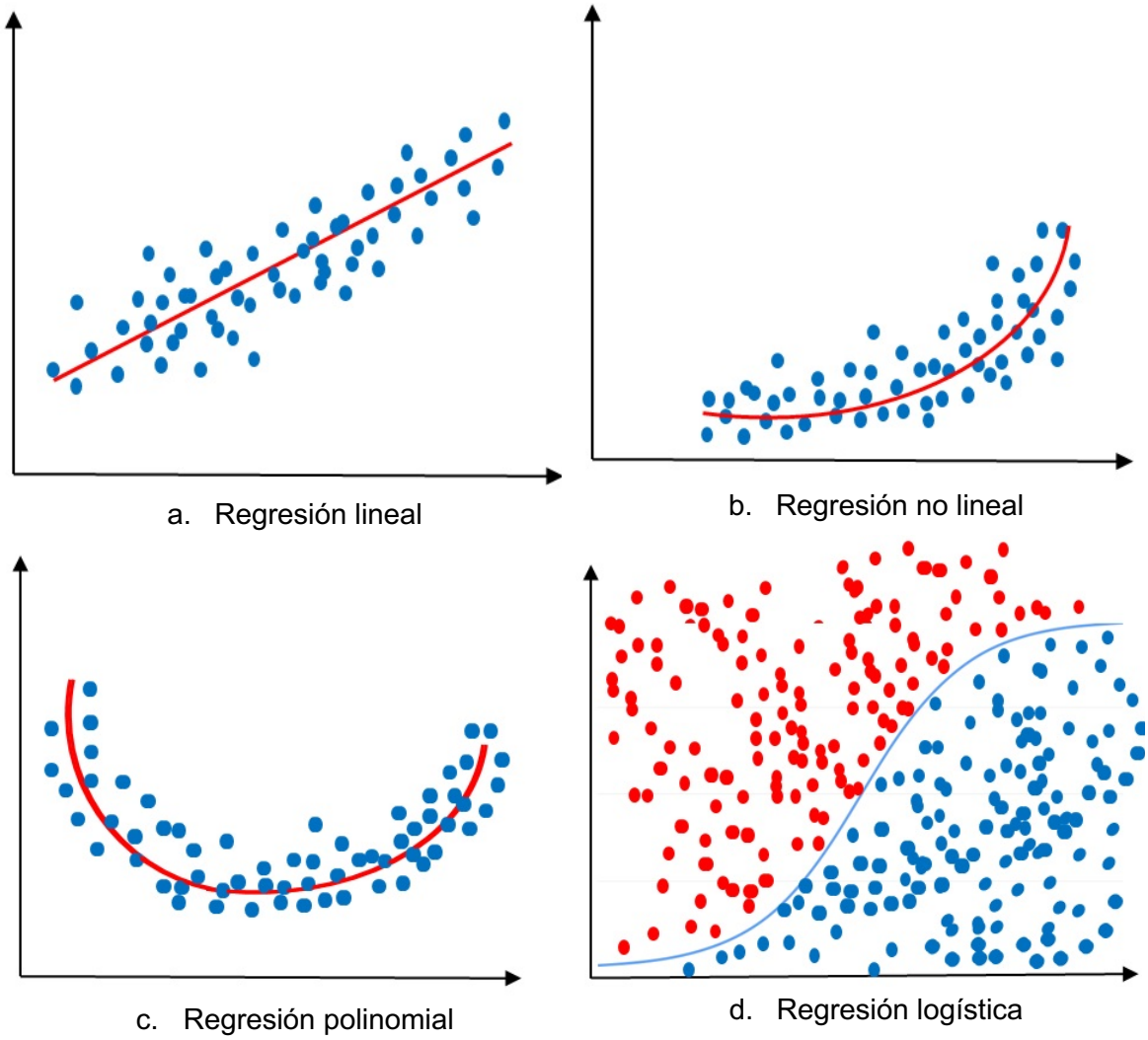


Figura 4.6. Principales tipos de modelos de regresión en la minería de datos.

4.3.3. Agrupamiento

El agrupamiento es la organización de un conjunto de datos en grupos establecidos de acuerdo con la similitud entre ellos. A diferencia de la clasificación, que es un problema de aprendizaje supervisado (cada registro en el conjunto de datos posee un atributo indicando la clase o categoría a la cual pertenece), el agrupamiento (del inglés *clustering*) es un problema de aprendizaje no supervisado, esto significa que la clase o categoría a la cual pertenece cada registro o dato de un conjunto no se conoce *a priori*, por lo que este modelo es el encargado de encontrar dichas clases (Aggarwal y Reddy, 2013; Everitt *et al.*, 2011).

El agrupamiento da como resultado la formación de clases o categorías (*clusters*), de las cuales, cada una debe ser más o menos homogénea —dada la similitud entre

los datos que agrupa— y diferentes de las demás (ver figura 4.7). En un grupo, clase o categoría, cada dato está representado por un vector de características, o como un punto en un espacio multidimensional.

La mayoría de las técnicas de agrupamiento se basa en algún tipo de métrica o medida para determinar qué tan similares son los objetos del grupo. Entre las más utilizadas se encuentran: distancia Euclidiana, distancia de Minkowski, distancia de Manhattan, distancia de Canberra, distancia de Karl Pearson y distancia de Mahalanobis.

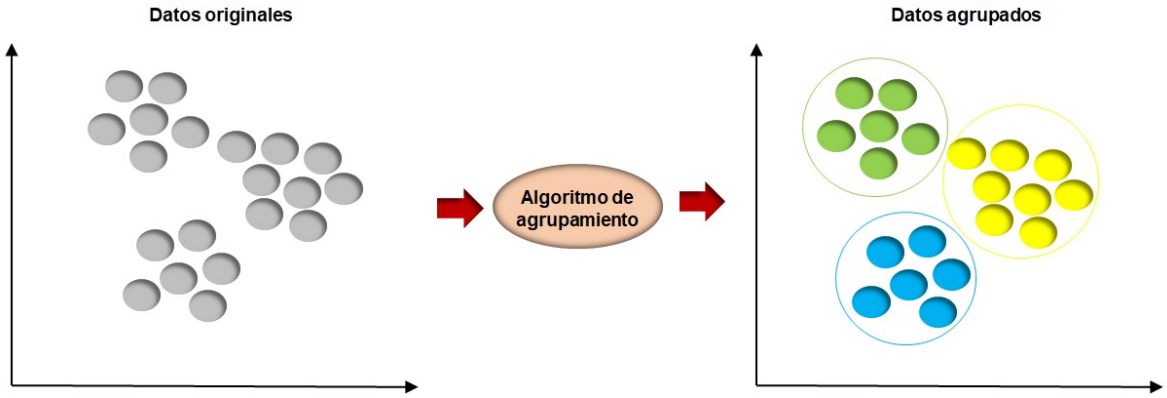


Figura 4.7. Algoritmo de agrupamiento.

Por su parte, los algoritmos de agrupamiento se clasifican en dos categorías principales:

- Agrupamiento jerárquico
- Agrupamiento partitivo

Aquí sólo nos referiremos al agrupamiento jerárquico, ya que es el algoritmo más utilizado en las técnicas de aprendizaje automático. En él, los grupos están anidados y organizados como nodos de un árbol jerárquico. Dependiendo de la técnica utilizada para la formación de grupos (*clusters*), el *clustering* jerárquico se subdivide en:

- **Aglomerativo (de abajo hacia arriba)**: La agrupación jerárquica aglomerativa comienza con un conjunto de N grupos, donde cada uno contiene un solo objeto. En cada paso del algoritmo, los dos objetos más similares se fusionan en un mismo grupo, y el proceso continúa hasta llegar a un único grupo que contiene todos los objetos iniciales.

- **Divisivo (de arriba hacia abajo)**: La agrupación jerárquica divisiva comienza con un único grupo que contiene N objetos que se van a agrupar y, a través de iteraciones sucesivas, separa los objetos en grupos más delgados o específicos.

El agrupamiento jerárquico suele esquematizarse mediante un diagrama bidimensional denominado dendograma (ver figura 4.8), que representa las fusiones o divisiones que se llevan a cabo sucesivamente, a medida que pasamos de los nodos particulares al nodo raíz o viceversa.

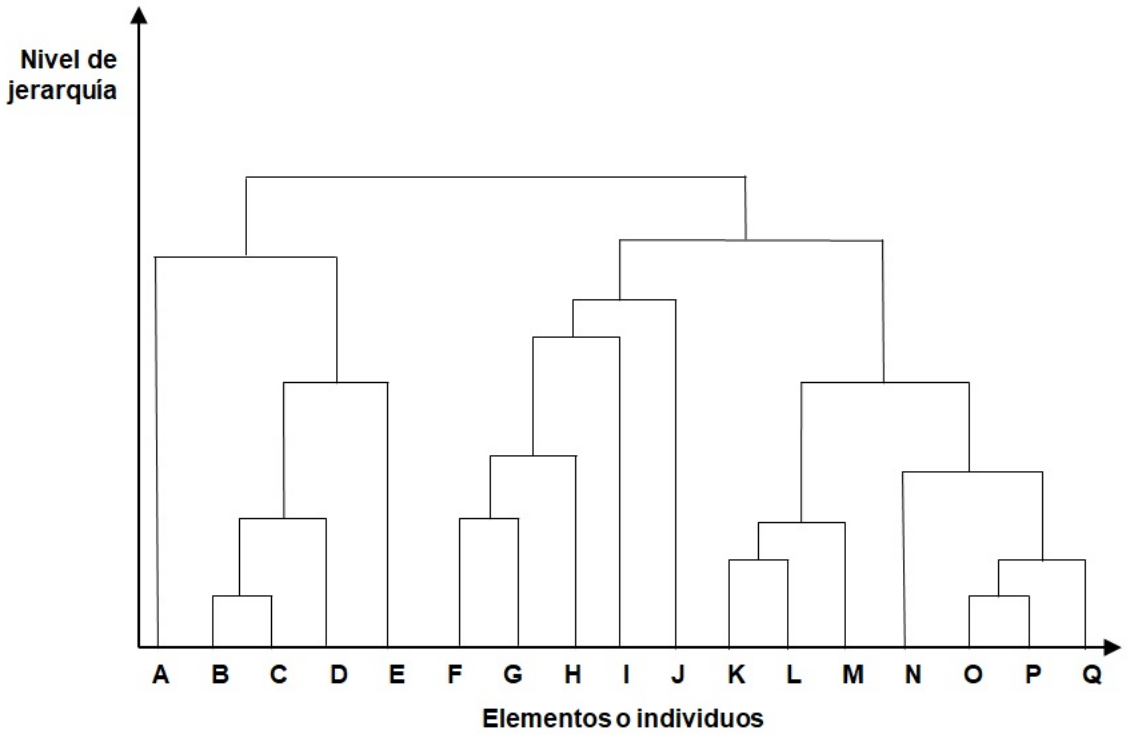


Figura 4.8. Representación gráfica del agrupamiento jerárquico.

4.3.4. Minería de reglas de asociación

A diferencia de los métodos de clasificación o regresión, los cuales predicen a qué clase o categoría pertenecerá un elemento particular o el valor que tomará una variable dependiente como resultado de los valores de las variables independientes, el objetivo de la minería de reglas de asociación es descubrir correlaciones o asociaciones importantes en forma de implicaciones lógicas (si <premisa> entonces <conclusión>), entre los valores de las características (variables) en grandes conjuntos de datos (Zhang y Zhang, 2002).

Uno de los dominios donde la minería de reglas de asociación resulta de gran valor

es el análisis de la canasta de compra del comercio electrónico, donde la aplicación de este método consiste en el análisis de datos transaccionales para identificar relaciones entre artículos que se compraron juntos. De esta forma, la empresa logra sugerir y vender productos relacionados con un cliente, aumentando así el valor total de la compra.

Otros dominios de aplicación de la minería de reglas de asociación son los siguientes:

- Diagnóstico médico y tratamiento
- Análisis del comportamiento de clientes del comercio electrónico
- Detección de fraude en transacciones con tarjetas bancaria
- Identificación de patrones en la interacción de usuarios en páginas web
- Minería de textos

4.3.5. Minería de patrones secuenciales

Si bien son métodos diferentes, la minería de patrones secuenciales se asemeja más a la minería de reglas de asociación que a los métodos de clasificación y regresión que ya se han revisado anteriormente (Wang y Yang, 2005). El objetivo de la minería de patrones secuenciales es encontrar secuencias de eventos que ocurren a lo largo del tiempo en los datos. Estos patrones muestran correlaciones temporales o secuenciales que pueden revelar información importante sobre los comportamientos, procesos y tendencias subyacentes en los datos.

En muchas situaciones del mundo real, los conjuntos de datos no corresponden a una colección de registros o instancias diferentes, sino que se presentan como una serie de eventos o transacciones que ocurren a lo largo del tiempo. Cualquier conjunto de datos que corresponda a series de tiempo son en sí datos secuenciales. Ejemplos de ellos son:

- Comportamiento de las acciones en el mercado bursátil, a través del tiempo
- Registro de transacciones efectuadas por clientes de *e-commerce*
- Datos producidos por simulaciones computacionales de sistemas biológicos
- Datos recolectados por un conjunto de sensores medioambientales a través del tiempo

4.3.6. Minería de textos

Cuando los datos de interés no son datos estructurados, es decir, no poseen una organización como la que exhiben los datos tabulares, requerida para la aplicación

de los métodos de clasificación, regresión y agrupamiento, es necesario recurrir a otros métodos como la minería de texto (Aggarwal, 2022; Ignatow y Mihalcea, 2016). También conocido como análisis de texto o procesamiento del lenguaje natural (*NLP,* por sus siglas en inglés), este método se enfoca en extraer información, patrones y conocimientos valiosos a partir de datos de texto no estructurados.

El auge que han tenido en los últimos años los motores de búsqueda, la publicación de contenidos en Internet, los foros de discusión, el correo electrónico, la publicación digital de literatura científica y tecnológica, así como las redes sociales, ha conllevado a una producción continua de grandes volúmenes de datos no estructurados. Es precisamente aquí donde los métodos de minería de texto resultan de gran utilidad, al permitir analizar estos grandes volúmenes y extraer información significativa de los mismos.

La minería de textos abarca varias tareas y técnicas. Entre ellas destacan:

- Reconocimiento de textos
- Clasificación de textos
- Agrupamiento de textos (*clustering*)
- Minería de opiniones
- Modelado de tópicos
- Resumir textos
- Reconocimiento de entidades nombradas (*NER,* por sus siglas en inglés)

4.4. Técnicas de minería de datos

Como ya se mencionó previamente, en este material se utilizará el término “técnica de minería de datos” para referir aquellos procedimientos a base de inteligencias artificiales, aprendizaje automatizado, estadísticas o matemáticas, que permiten la implementación de los métodos de la minería de datos, tales como limpieza de datos, selección de características, predicción, clasificación, agrupamiento, entre otros. De forma particular, este texto se centrará en técnicas de la minería de datos que permiten implementar los métodos de las siguientes fases:

- Preparación de los datos
- Modelado

No son objetivos de este material presentar y discutir de forma detallada las diferentes técnicas que se implementan en estas dos fases de la minería de datos, ya que la extensión y nivel de detalle de estos contenidos conllevaría a perder la idea principal que se intenta transmitir aquí: la preparación y análisis de grandes

volúmenes de datos. Sin embargo, si el lector está interesado en estos temas, puede consultar los textos especializados de aprendizaje automatizado (Cerulli, 2023; Suthaharan, 2015), redes neuronales artificiales (Aggarwal, 2019; Chowdhary, 2020), árboles de decisión (Hartshorn, s. f.; Sheppard, 2017) o modelos de regresión (Chatterjee y Simonoff, 2013; Young, 2017) referenciados en la bibliografía.

4.4.1. Técnicas de minería de datos utilizadas en la fase de preparación de los datos

A continuación, la tabla 4.1 relaciona las principales técnicas utilizadas para los métodos que caracterizan la fase de preparación de datos:

Tabla 4.1. Técnicas comúnmente utilizadas en la fase de preparación de los datos de la minería de datos

Método de minería de datos	Técnica utilizada
Selección de datos	❖ Filtros para la selección de campos o selección de registros
Limpieza de datos	❖ Relleno de datos nulos o perdidos con la media, moda, o mediana ❖ Eliminación de valores atípicos
Construcción de nuevos datos	❖ Derivación de nuevos campos a partir de campos existentes, utilizando expresiones que contienen operadores aritméticos, relacionales o lógicos
Integración de datos	❖ Fusión de archivos.
Balanceo de clases	❖ Algoritmo SMOTE (del inglés *Synthetic Minority Oversampling Technique*) ❖ Algoritmo GAN (del inglés *Generative Adversarial Networtk*)
Selección de características	❖ Análisis de componentes principales (PCA, por sus siglas en inglés) ❖ Métricas de similitud, tales como Chi-Cuadrada, Coeficiente de correlación de Pearson, *Low variance*, *Recursive feature elimination*, *Least Absolute Shrinkage and Selection Operator*, entre otras

4.4.2. Técnicas de minería de datos utilizadas en la fase de modelado

La tabla 4.2 relaciona las principales técnicas que permiten implementar los métodos de clasificación, regresión y agrupamiento característicos de esta fase de

minería de datos (ver figura 4.9).

Tabla 4.2. Técnicas comúnmente utilizadas en la implementación de los métodos de la fase de modelado de la minería de datos

Método de minería de datos	Técnica utilizada
Predicción	❖ Regresión lineal ❖ Regresión no lineal ❖ Regresión polinomial
Clasificación	❖ Redes neuronales supervisadas: ○ Perceptrón multicapa (MLP) ○ Máquina de vectores de soporte (SVM) ○ Máquina de vectores de soporte Lineal (LSVM) ○ Máquina de Boltzmann restringida (RBM) ❖ Árboles de decisión ❖ Algoritmo de los K vecinos más cercanos ❖ Regresión logística
Agrupamiento	❖ Redes neuronales no supervisadas ❖ Algoritmos de agrupamiento

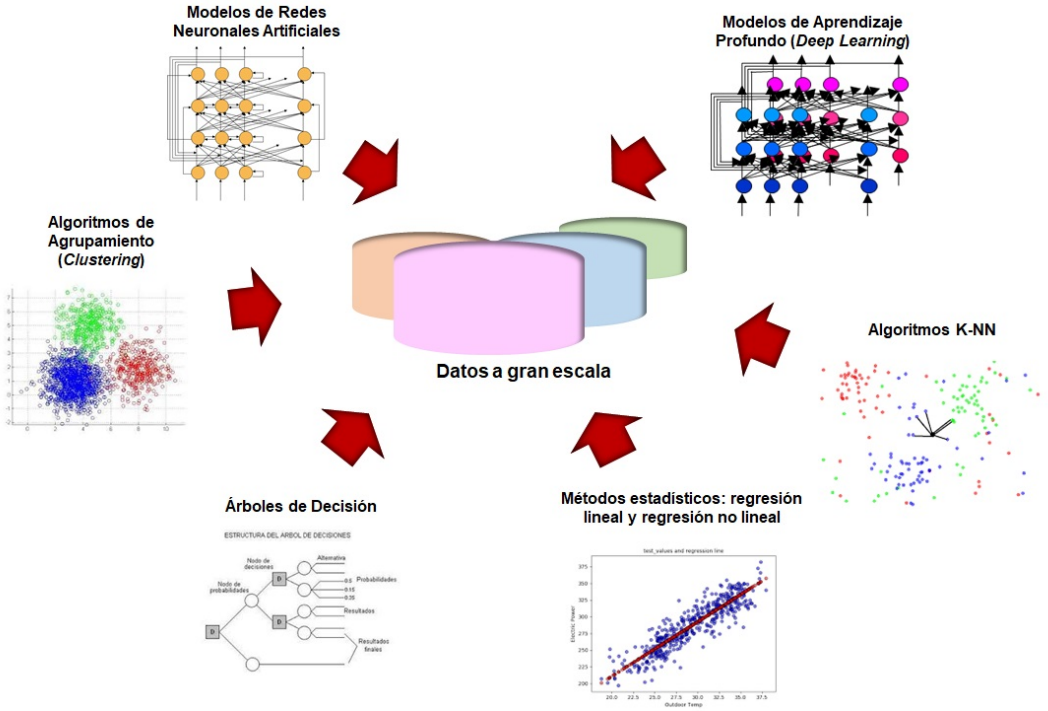


Figura 4.9. Técnicas en las que comúnmente se apoya la fase de modelado de la minería de datos.

4.5. Herramientas de minería de datos

En las últimas dos décadas se han desarrollado y liberado el acceso a una amplia gama de herramientas de minería de datos, las cuales comúnmente soportan una variedad de funciones para la preparación de datos, la exploración, el modelado y el análisis. Resulta difícil identificar entre estas herramientas de minería de datos la ideal para el tipo de tarea a realizar: algunas se especializan mucho más en la exploración y preparación de los datos, mientras que otras hacen mayor énfasis en el modelado; algunas más se enfocan en otros aspectos como la curva de aprendizaje asociada al uso y dominio de la herramienta, las características del *hardware* requerido para su instalación y ejecución, el volumen máximo de datos que la herramienta puede procesar o el tipo de acceso a la misma, es decir, si es un *software* gratuito o de paga. No obstante, entre las herramientas de minería de datos de mayor poder disponibles, ya sean de uso gratuito o de paga, podemos identificar las siguientes:

- IBM SPSS Modeler (https://www.ibm.com/mx-es/products/spss-modeler)
- WEKA (Waikato Environment for Knowledge Analysis) (https://www.weka.io/)
- RapidMiner Studio (https://docs.rapidminer.com/latest/studio/)
- KNIME (Konstanz Information Miner) (https://www.knime.com/)
- Librerías de Python: Scikit-learn (https://scikit-learn.org/stable/), Pandas (https://pandas.pydata.org/) y NumPy (https://numpy.org/)
- Oracle DM (Oracle Data Miner) (https://www.oracle.com/big-data/technologies/dataminer/)
- SSAS (Microsoft SQL Server Analysis Services) (https://learn.microsoft.com/en-us/analysis-services/ssas-overview?view=asallproducts-allversions)

La tabla 4.3 indica las fases, métodos y actividades de minería de datos que soporta cada una de las herramientas antes mencionadas. En este material haremos uso de dos en particular:

- IBM SPSS Modeler (licencia para uso académico)
- IDA-WEB TOOL (herramienta desarrollada a nivel prototipo para uso académico, de uso gratuito, disponible en: http://bioinformatics.cua.uam.mx/ida-tool/)

La herramienta IBM SPSS Modeler (https://www.ibm.com/mx-es/products/spss-

modeler) proporciona un valioso soporte a cada una de las fases de la minería de datos (ver figura 4.1), brindando al usuario una extensa paleta de herramientas para automatizar las tareas que engloba cada una de éstas.

Por otra parte, la herramienta de minería de datos IDA-WEB TOOL será presentada de forma íntegra en el capítulo VI.

Tabla 4.3. Herramientas de minería de datos con las fases y métodos en las que brindan soporte

Herramienta	Fases y métodos de la minería de datos que soporta
IBM SPSS Modeler	Proporciona soporte en todas las fases de minería de datos, siguiendo la metodología CRISP-DM: ❖ Comprensión del negocio ❖ Comprensión de los datos ❖ Preparación de los datos ❖ Modelado: clasificación, regresión y agrupamiento ❖ Evaluación ❖ Despliegue
WEKA	❖ Preparación de los datos ❖ Modelado: clasificación, regresión y agrupamiento ❖ Minería de reglas de asociación ❖ Selección de características
RapidMiner	❖ Preparación de los datos ❖ Modelado ❖ Despliegue
KNIME	❖ Preparación de los datos ❖ Modelado ❖ Minería de textos
Librerías de Python	❖ Preparación de los datos ❖ Modelado: clasificación, regresión y agrupamiento ❖ Reducción de dimensionalidad ❖ Balanceo de clases en conjuntos de datos para tareas de clasificación ❖ Evaluación de los modelos
Oracle DM	❖ Preparación de los datos ❖ Modelado: clasificación, regresión y agrupamiento ❖ Minería de reglas de asociación

SSAS	❖ Preparación de los datos ❖ Modelado: clasificación, regresión y agrupamiento ❖ Minería de reglas de asociación

4.6. Enfoques metodológicos de minería de datos

En los últimos años, varios enfoques metodológicos para guiar el proceso de minería de datos han sido propuestos en la literatura (Aggarwal, 2015; Hernádez Orallo, 2010; Tan *et al.*, 2018; Witten *et al.*, 2016). Sin embargo, entre todos estos, SMART Model (Marr, 2015) y CRISP-DM (Cross Industry Standard Process for Data Mining) han sido las dos metodologías de minería de datos más difundidas y utilizadas en las últimas décadas (Perez, 2021; Shearer, 2000).

La figura 4.10 ilustra las fases que integran el enfoque metodológico de minería de datos SMART Model, las cuales son:

- (S) Iniciar con una estrategia
- (M) Medir métricas y datos
- (A) Analizar los datos
- (R) Reportar los resultados
- (T) Transformar el negocio y la toma de decisiones

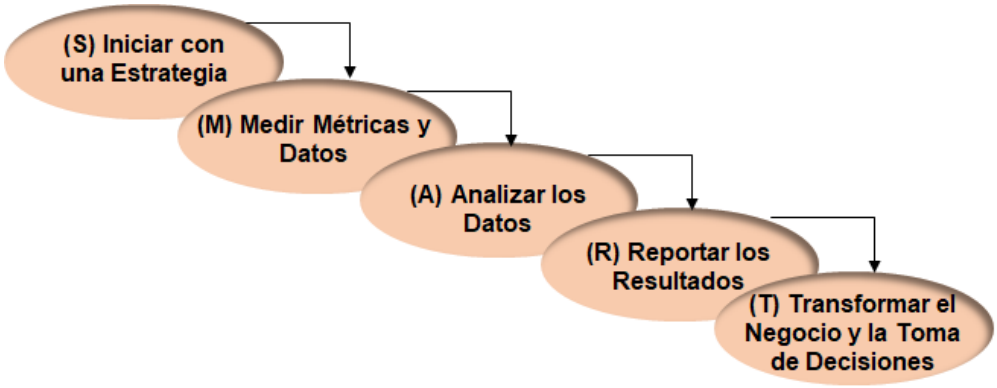


Figura 4.10. El enfoque metodológico de minería de datos SMART Model.

Por otra parte, la metodología CRISP-DM será tratada de forma íntegra en el próximo capítulo, ya que es precisamente ésta la que guiará el trabajo de preparación y análisis de grandes volúmenes de datos en este material.

V. EL ENFOQUE METODOLÓGICO DE MINERÍA DE DATOS CRISP-DM

A pesar de no constituir un enfoque metodológico reciente, Cross Industry Standard Process for Data Mining (CRISP-DM) sigue siendo relevante como guía en la minería de datos, por lo que hoy en día aún es ampliamente usada, difundida, extendida y adaptada a proyectos de datos a gran escala (Perez, 2021; Shearer, 2000). Lo anterior justifica que se trate dicho enfoque en este material.

Como se puede apreciar en la figura 5.1, el enfoque metodológico de minería de datos CRISP-DM abarca las siguientes fases:

- Comprensión del problema
- Comprensión de los datos
- Preparación de los datos
- Modelado
- Evaluación
- Presentación de los datos

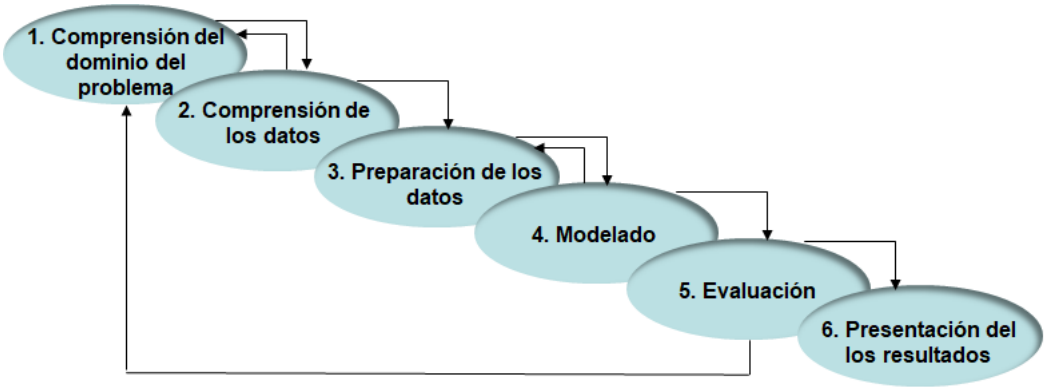


Figura 5.1. El enfoque metodológico de minería de datos CRISP-DM.

CRISP-DM es una metodología de minería de datos muy flexible, la cual se puede adaptar fácilmente a necesidades concretas o a un problema específico, relacionados con el volumen de datos a procesar y analizar. En ocasiones, dependiendo de dichos aspectos, las fases de comprensión del problema y preparación de los datos pueden llegar a ser las más relevantes, mientras que en otros casos las fases de modelado y evaluación serán las más significativas para la toma de decisiones.

A continuación, se describirá el alcance de cada una de las fases de la

metodología de minería de datos CRISP-DM (ver figura 5.1), haciendo énfasis en las actividades que comprende cada una.

Para un mejor entendimiento del enfoque metodológico CRISP-DM, se ha tratado de incluir fragmentos de ejemplos ilustrativos de todas fases. Cada uno de ellos se encuentran insertados en cuadros de texto de fondo azul claro, de forma que el lector los puede identificar con facilidad. Por otra parte, de forma complementaria, en los capítulos VI y VII se ilustra la aplicación de las fases del enfoque metodológico CRISP-DM en tres conjuntos de datos muy difundidos y ampliamente utilizados en tareas de clasificación y predicción con aprendizaje automatizado:

1. El primer caso describe el comportamiento semanal de un conjunto de acciones en el mercado bursátil, a partir de importantes atributos como: identificador de la acción, precio de apertura, precio de cierre, precio máximo, precio mínimo, volumen de acciones transferido, entre otros. Este conjunto de datos ha sido utilizado para demostrar la operación de la herramienta de minería de datos IDA-WEB TOOL, presentada en el capítulo VI. Asimismo, es de carácter público y se encuentra en el UC Irvine Machine Learning Repository de la Universidad de California, Estados Unidos (https://archive.ics.uci.edu/dataset/312/dow+jones+index).

2. El segundo conjunto de datos está dedicado al mercado de bienes de consumo. De forma particular, al incremento en las ventas de cuatro tipos de bienes de consumo: bebidas y licores, artículos confeccionados, productos cárnicos y artículos de lujo. La información se tomó de los datos de prueba que provee la herramienta IBM SPSS MODELER en su versión para la comunidad estudiantil (https://www.ibm.com/mx-es/products/spss-modeler), cuyo uso requiere licencia. El tamaño del conjunto original era de 200 registros o instancias, y fue aumentado a 2500 para mejorar el desempeño de los modelos de aprendizaje automatizado.

3. El tercer conjunto de datos está relacionado con la tarea de predicción del abandono de los servicios de tarjetas de crédito, a partir de características sociodemográficas y financieras de un grupo de 10,000 clientes. El conjunto de datos es de acceso público y se encuentra disponible en: https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers.

5.1. Fase de comprensión del problema de CRISP-DM

La fase de comprensión del dominio del problema o negocio (ver figura 5.2) resulta

de gran importancia e impacto en las fases sucesivas, independientemente de la metodología o enfoque de minería de datos que se haya seleccionado. Esto se debe a que en esta fase se define de forma clara el problema que se intenta resolver, se focaliza en la comprensión de las metas u objetivos a lograr y proporciona una perspectiva de minería de datos para comprender qué datos deben ser analizados.

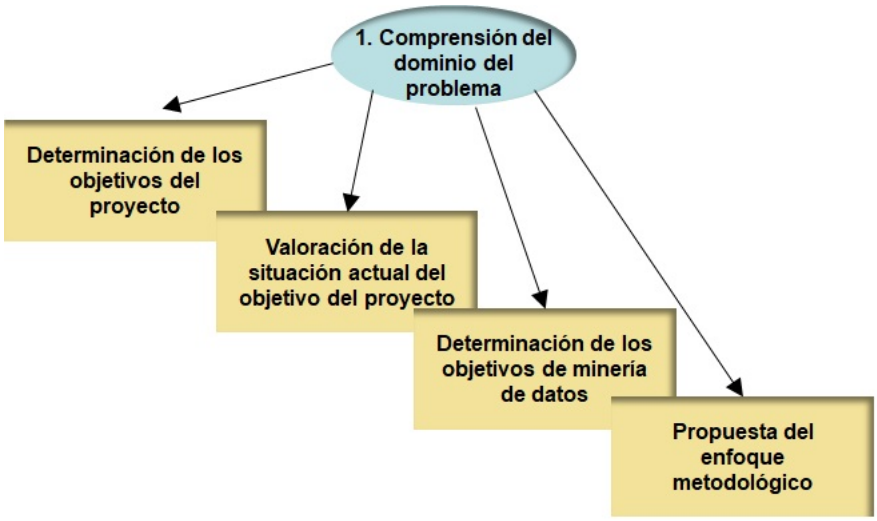


Figura 5.2. Actividades de la fase de comprensión del dominio del problema, del enfoque metodológico de minería de datos CRISP-DM.

Como se puede apreciar en la figura 5.2, cuando se sigue el enfoque CRIPS-DM, la comprensión del dominio del problema o negocio puede llevarse a cabo ejecutando las siguientes actividades (Perez, 2021; Shearer, 2000):

1. Determinación de los objetivos del proyecto
2. Valoración de la situación actual del objetivo del proyecto
3. Determinación de los objetivos de minería de datos
4. Propuesta del enfoque metodológico para desarrollar el proyecto

5.1.1. Determinación de los objetivos del proyecto

En la fase de comprensión del dominio del problema es imprescindible conocer de forma clara las razones que justifican llevar a cabo el proyecto o estudio. De aquí que un aspecto clave de esta fase sea precisamente la determinación de los objetivos. Esto se refiere identificar: ¿qué se pretende hacer?, ¿cuál es el problema que se necesita resolver?, ¿cuál es la pregunta a responder?

Los objetivos del proyecto dependerán completamente del dominio del problema o negocio en el cual se enmarca (ver figuras 5.3 y 5.4).

Ejemplos de dominio del problema o negocio son los siguientes:

- Ventas y mercadotecnia (marketing)
- Finanzas y mercado bursátil
- Gestión bancaria y crediticia
- Medicina
- Lotería y juegos de apuesta
- Clima y contaminación atmosférica
- Investigación científica
- Otras ramas de las ciencias e ingenierías

Figura 5.3. Ejemplos de dominio del problema o negocio.

Ejemplos de objetivos de proyecto son los siguientes:

- En el dominio de ventas y mercadotecnia, el objetivo podría ser la predicción del incremento de las ventas y la retención de clientes, a partir de acciones tales como ofertas y descuentos.
- En el dominio de otorgamiento de créditos, el objetivo podría ser la clasificación de los solicitantes de crédito bancario en "aceptados" o "rechazados", según su historial crediticio, capacidad de pago, etc.
- En el dominio de medicina, el objetivo del proyecto podría ser el diagnóstico de enfermedades, a partir de información heredo-familiar, signos y síntomas, resultados de exámenes de laboratorio, entre otro tipo de información.
- En el dominio de clima y contaminación atmosférica, el objetivo podría ser la predicción de la calidad del aire en determinadas zonas metropolitanas.

Figura 5.4. Ejemplos de objetivos de proyectos.

5.1.2. Valoración de la situación actual del objetivo del proyecto

Una vez que se ha definido el objetivo del proyecto es necesario valorar la situación actual del mismo, considerando las siguientes cuestiones (Shearer, 2000):

- ¿Se comprende de forma clara el problema que se intenta abordar?
- ¿Existen datos disponibles para efectuar el análisis?
- De contar con datos disponibles, ¿cuál es su fuente y de qué tipo son?

- ¿Se dispone de los recursos humanos y tecnológicos para desarrollar el proyecto?
- ¿Se han identificado factores de riesgo que afecten el desarrollo del proyecto?

5.1.3. Determinación de los objetivos de minería de datos

En esta fase es imprescindible que, una vez que se determinen los objetivos del proyecto, se traduzcan a objetivos de minería de datos. Los objetivos de minería de datos se refieren a qué tipo de información se desea explorar y encontrar en los datos, por ejemplo:

- Datos que satisfagan ciertas condiciones
- Relaciones existentes entre ciertas características o campos de los datos
- Clasificación de los datos
- Agrupamiento de los datos
- Predicción a partir de los propios datos
- Identificación de patrones en el conjunto de datos

La figura 5.5 ilustra varios ejemplos de objetivos de minería de datos.

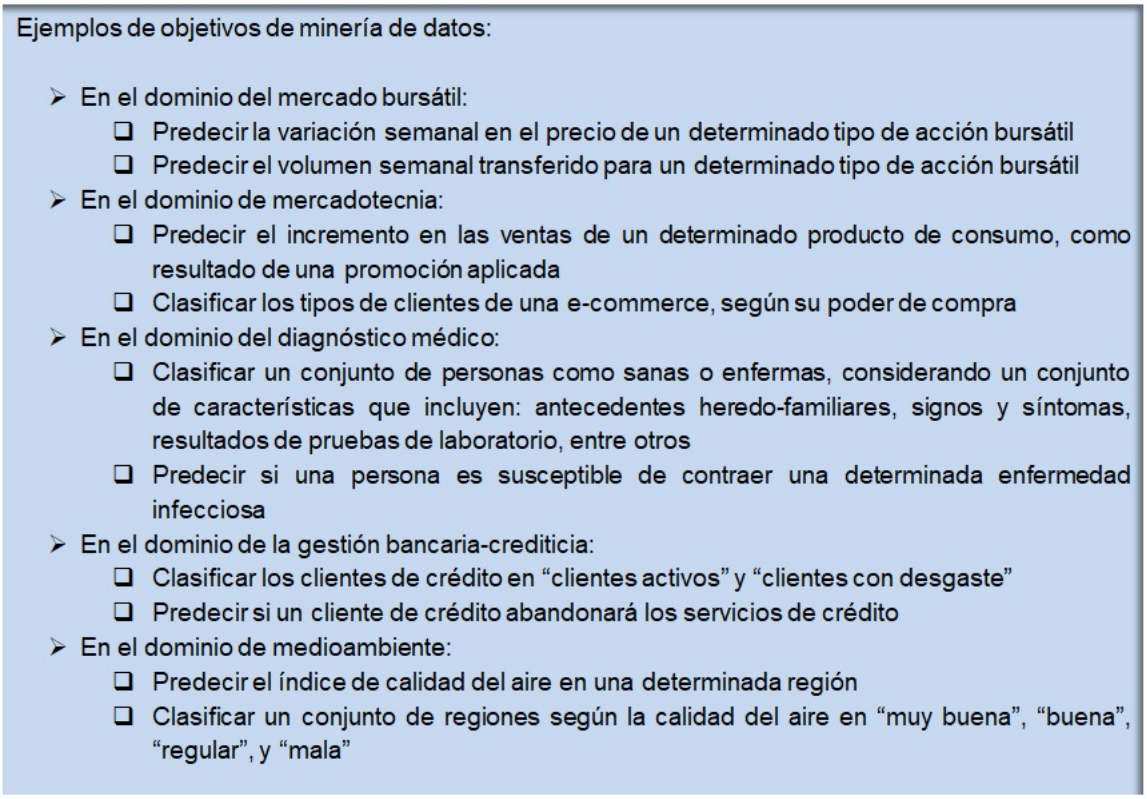


Figura 5.5. Ejemplos de objetivos de minería de datos.

5.1.4. Propuesta del enfoque metodológico para desarrollar el proyecto

El enfoque metodológico o plan de proyecto de minería de datos se propone a partir de todos los aspectos que se han identificado, hasta el momento, como parte de la fase de comprensión del dominio del problema o negocio. Dentro de estos aspectos se encuentran el objetivo del proyecto, la valoración de la situación actual del objetivo del proyecto y los objetivos de minería de datos.

El enfoque metodológico o plan de proyecto de minería de datos puede ser visto como un cronograma de trabajo, comúnmente presentado en forma de tabla (ver figura 5.6), el cual debe especificar, entre otros aspectos:

- Nombre de la fase de la metodología CRISP-DM
- Tiempo en semanas dedicado a dicha fase
- Recursos humanos y tecnológicos que requiere
- Posibles riesgos a mitigar en cada fase

Fase	Tiempo a dedicar	Recursos humanos y tecnológicos	Riesgos atribuibles
Comprensión del dominio del problema			
Comprensión de los datos			
Preparación de los datos			
Modelado			
Evaluación			
Presentación			

Figura 5.6. Enfoque metodológico o plan de proyecto de minería de datos, según la metodología CRIP-DM.

5.2. Fase de comprensión de los datos de CRISP-DM

La fase de comprensión de los datos consiste en la descripción, exploración y análisis inicial, con la finalidad de:

- Conocer las características de los datos: número de registros, número de características o campos, significado de cada característica o campo, tipo de cada característica o campo, relaciones entre las características o campos.
- Identificar problemas de calidad presentes en los datos (omisión, incompletitud, redundancia, falta de veracidad, etcétera).
- Descubrir y proponer ideas iniciales acerca de los datos.
- Establecer hipótesis acerca de la información que describen dichos datos.

La fase de comprensión de los datos implica estudiarlos con mayor detenimiento, tratando de entender las posibles relaciones que se manifiesten entre ellos. Esta fase también conlleva a la exploración de los datos con el apoyo de tablas, gráficos, resúmenes y otras herramientas estadísticas que proporcionen diferentes perspectivas de los mismos. Nótese que esta fase no significa la preparación ni limpieza de los datos, sino que se trata sólo de una exploración inicial para comprenderlos y descubrir relaciones entre ellos.

Como se puede apreciar en la figura 5.7, la fase de comprensión de los datos incluye las siguientes actividades (Shearer, 2000):

- Recopilación de los datos iniciales
- Descripción de los datos
- Exploración de los datos

- Verificación de la calidad de los datos

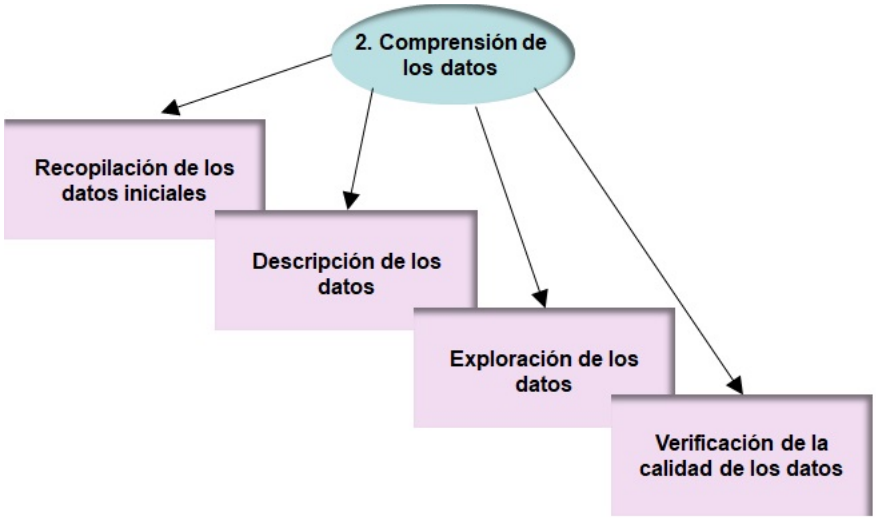


Figura 5.7. Actividades de la fase de comprensión de los datos del enfoque metodológico de minería de datos CRISP-DM.

5.2.1. Recopilación de los datos iniciales

La recopilación de los datos iniciales implica identificar las fuentes de las cuales provienen. Se pueden identificar los siguientes tipos de datos:

- **Datos existentes**: Son con los que se cuenta inicialmente para el proyecto de minería de datos. Comúnmente, son obtenidos por la propia compañía, empresa, institución, grupo de investigación, etcétera.
- **Datos adquiridos**: Se refiere a aquellos adquiridos por un tercero y que pueden complementar los datos existentes, si se cuenta con ellos, o constituir los datos iniciales para el proyecto de minería de datos.
- **Datos adicionales**: De ser el caso, se deberá recurrir a otros datos que complementen las fuentes ya mencionadas.

Como parte de la recopilación de los datos iniciales, es necesario efectuar un análisis preliminar de la información, centrando la atención en sus atributos o campos (columnas), mediante preguntas como:

- ¿Cuáles son los atributos más prometedores?
- ¿Cuáles son los atributos menos relevantes?, ¿podrían excluirse del conjunto de datos?
- ¿Se cuenta con datos suficientes?

- ¿Son suficientes los atributos?

La figura 5.8 muestra algunos ejemplos de datos existentes, adquiridos y adicionales.

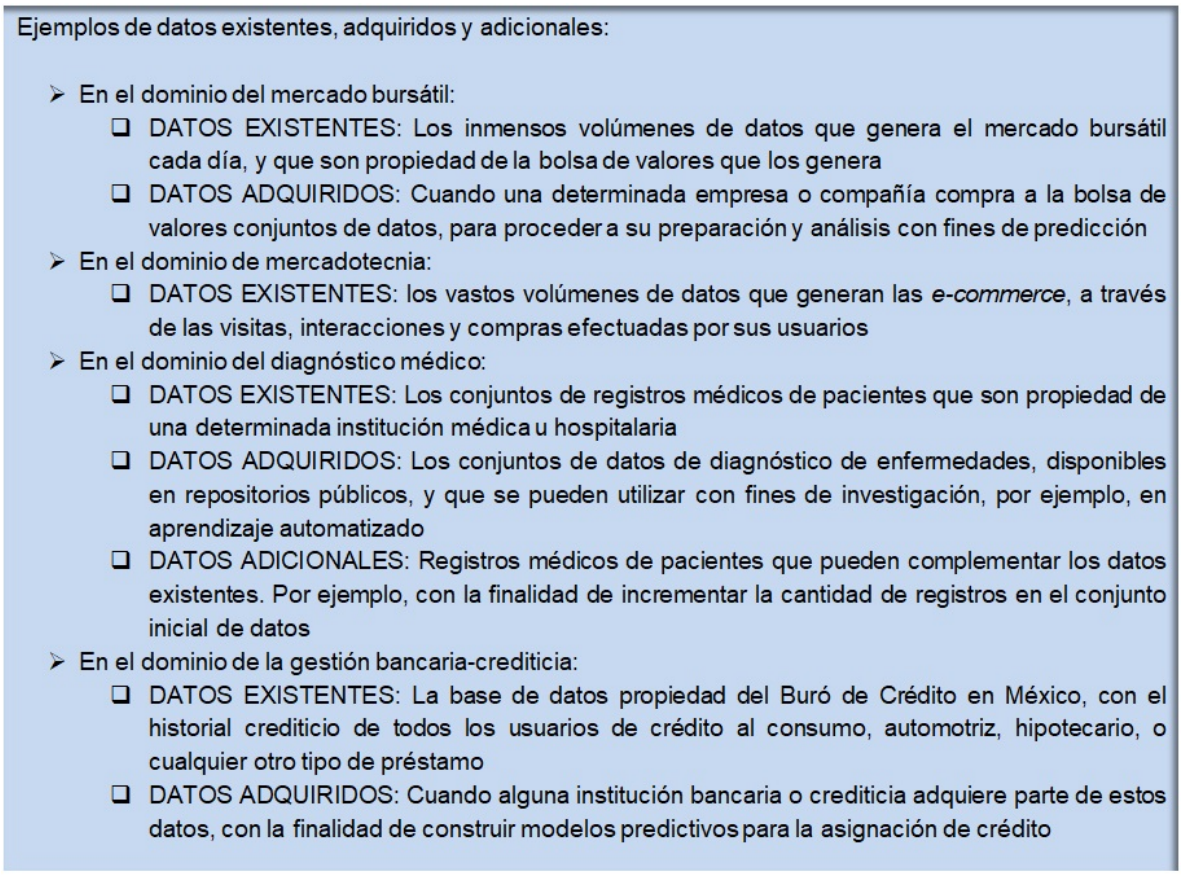


Figura 5.8. Ejemplos de datos existentes, adquiridos y adicionales.

5.2.2. Descripción de los datos

Esta actividad se refiere a la descripción del conjunto de datos, tomando en consideración los siguientes aspectos por cada conjunto disponible:

- Cantidad de registros que conforman el conjunto de datos.
- Cantidad de atributos o características.
- Tipos de valores (numéricos, caracteres, booleanos): identificar los tipos de datos simbólicos (fecha, hora, acceso a páginas web, etcétera) y proponer su conversión a datos numéricos.
- En el caso de conjuntos de datos para tareas de clasificación, información sobre posibles clases o categorías que agrupan diferentes subconjuntos y cantidad de datos por clase.

Ejemplo de la caracterización de un conjunto de datos correspondiente al comportamiento de clientes de crédito al consumo, hipotecario, automotriz, y otros tipos de préstamos:

- Cantidad total de registros: 2,485,890
- Cantidad total de atributos: 24
  - Número de atributos predictores: 23
  - Número de atributos representando clases: 1
- Número de clases y significado de las mismas:
  - Número de clases: 4
    - Clase 1: Cliente al corriente en sus pagos en todas sus cuentas
    - Clase 2: Cliente con atraso de 1 a 89 días en al menos una de sus cuentas
    - Clase 3: Cliente con atraso superior a 90 días en al menos una de sus cuentas
    - Clase 4: Cliente con al menos una cuenta sin recuperar
- Cantidad de atributos numéricos: 11
- Cantidad de atributos nominales: 2
- Cantidad de atributos ordinales: 10

Figura 5.9. Ejemplo de la caracterización de un conjunto de datos.

5.2.3. Exploración de los datos

La exploración de los datos constituye un primer análisis efectuado sobre el conjunto para lograr lo siguiente (Shearer, 2000):

- Comprender mejor los datos.
- Corroborar o hacer más explícitos los objetivos de minería de datos.
- Formular hipótesis sobre los datos.
- Identificar tareas requeridas en la fase de preparación de los datos, tales como limpieza, eliminación de datos redundantes, completar datos faltantes, etcétera.

La exploración de los datos se apoya en diferentes tipos de gráficos, tablas, herramientas de visualización y resúmenes estadísticos que proporcionan una mejor comprensión de los mismos, y, como resultado, permiten identificar atributos claves y eliminar atributos irrelevantes. Las figuras 5.10 y 5.11 muestran las paletas de gráficos que proporcionan las herramientas de minería de datos IDA-WEB TOOL e IBM SPSS MODELER, respectivamente, como soporte a la exploración de datos.

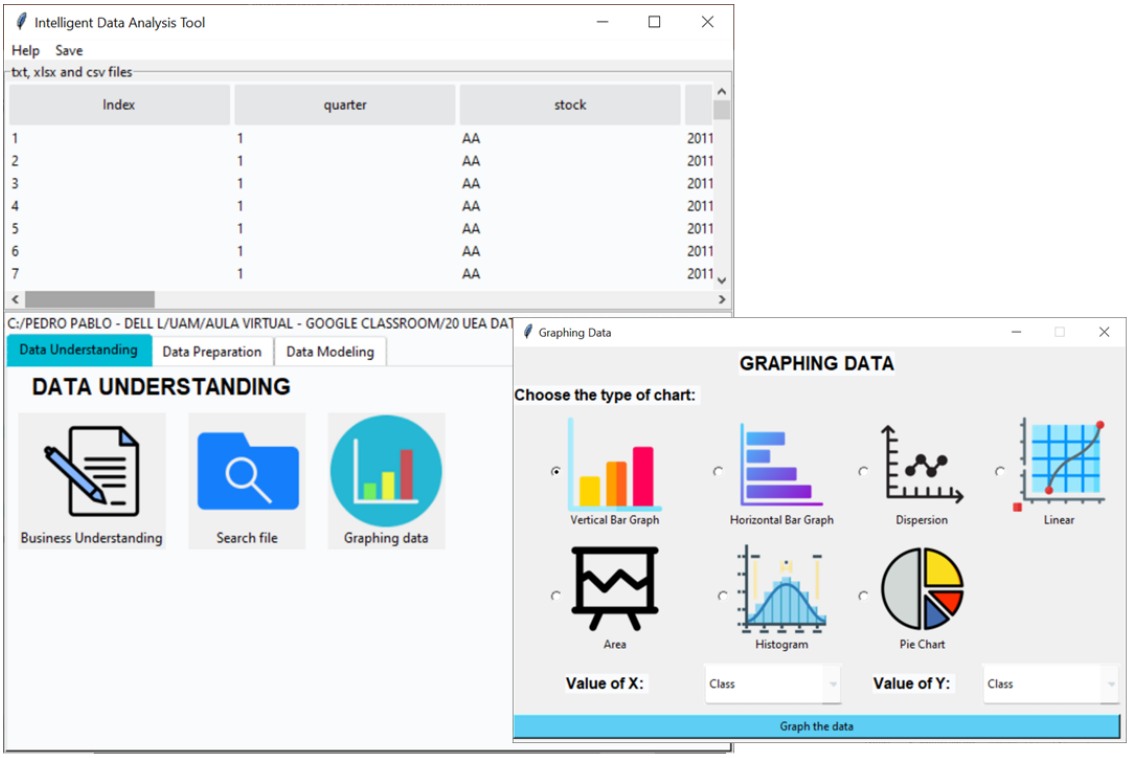


Figura 5.10. Paleta de gráficos que proporciona la herramienta de minería de datos IDA-WEB TOOL para la exploración de los datos.

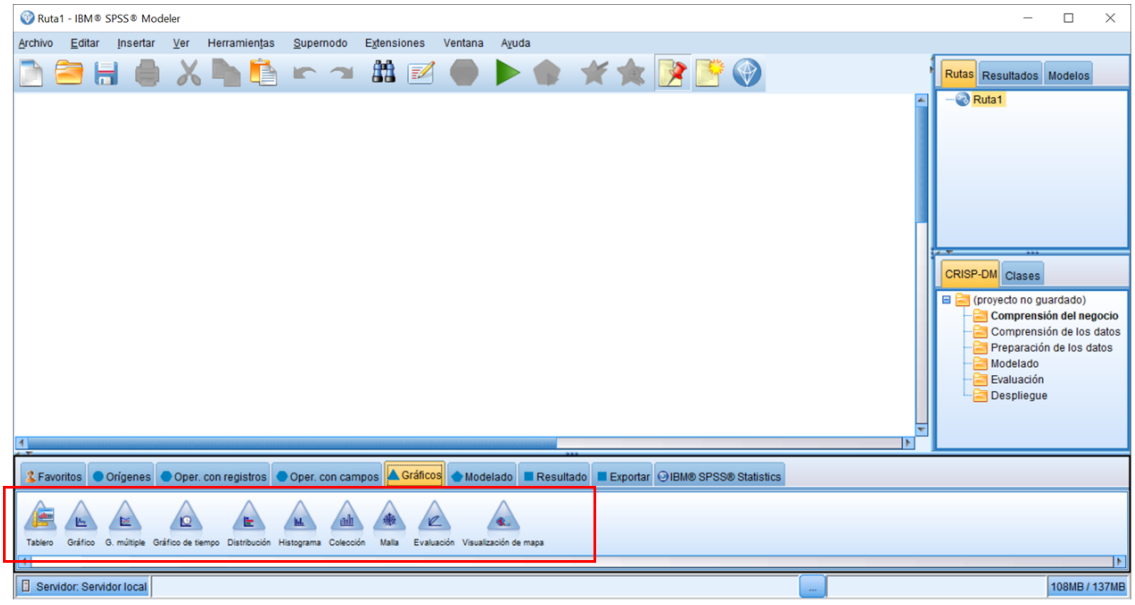


Figura 5.11. Paleta de gráficos que proporciona la herramienta de minería de datos IBM SPSS MODELER para la exploración de los datos.

5.2.4. Verificación de la calidad de los datos

La verificación de la calidad de los datos se refiere a identificar errores, inconsistencias o incompletitud en el conjunto de datos. Esta actividad se centra en distinguir los siguientes tipos de problemas:

- Datos perdidos o vacíos
- Errores en los datos
- Errores de medición
- Errores de codificación
- Atributos redundantes o de escasa utilidad

Todos los problemas identificados durante la verificación de los datos deberán solucionarse en la próxima fase (preparación de los datos), antes de que ser proporcionados como insumo al modelo que permitirá su análisis final.

5.3. Fase de preparación de los datos de CRISP-DM

Como su nombre lo indica, la fase de preparación de los datos prepara o da forma a los datos sin procesar para elaborar el conjunto final que servirá como insumo para la construcción del modelo (Perez, 2021; Shearer, 2000). De existir problemas de calidad en los datos, tales como omisión, errores en los datos, errores de medición o de codificación, incompletitud, redundancia, falta de veracidad, etcétera, es en esta fase donde deben ser solucionados.

Esta fase es la que, comúnmente, consume más tiempo, no sólo en la metodología CRISP-DM, sino en cualquier enfoque de minería de datos. En esta fase los datos en bruto deben ser preprocesados y convertidos en un conjunto sobre el cual se puedan aplicar los objetivos de la minería de datos.

Como se ilustra en la figura 5.12, la fase de preparación de los datos implica las siguientes actividades (Shearer, 2000):

- Selección de datos
- Limpieza de datos
- Construcción de nuevos datos
- Integración de datos
- Formato de datos

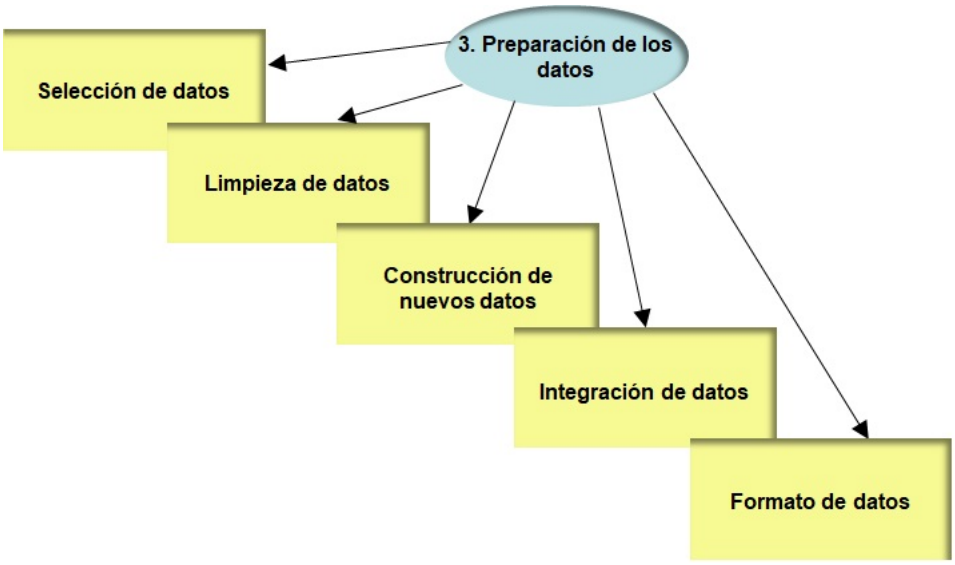


Figura 5.12. Actividades de la fase de preparación de los datos del enfoque metodológico de minería de datos CRISP-DM.

5.3.1. Selección de datos

Una vez realizada la recolección de datos iniciales (fase de comprensión de los datos), es posible iniciar la selección de aquellos que satisfagan los objetivos propuestos. Dicha selección abarca dos criterios fundamentales:

- **Selección de registros (filas)**: Se selecciona el subconjunto o subconjuntos de datos a utilizar. No suele requerirse el volumen completo de datos para el análisis y toma de decisiones, puesto que mientras mayor sea el conjunto, mayor será el tiempo requerido para las fases de preparación de los datos y modelado.
- **Selección de atributos o características (columnas)**: Se seleccionan los atributos o características más relevantes para el análisis y toma de decisiones. No todos los atributos o características resultarán relevantes para dicho fin.

La mayoría de las herramientas que soportan parcial o completamente la minería de datos ofrece la opción de “selección de datos”, tanto para la selección de registros (filas) como para la de atributos o características (columnas). Las figuras 5.13 y 5.14 muestran las funciones de soporte para la selección de datos que proporcionan las herramientas IDA-WEB TOOL e IBM SPSS MODELER, respectivamente.

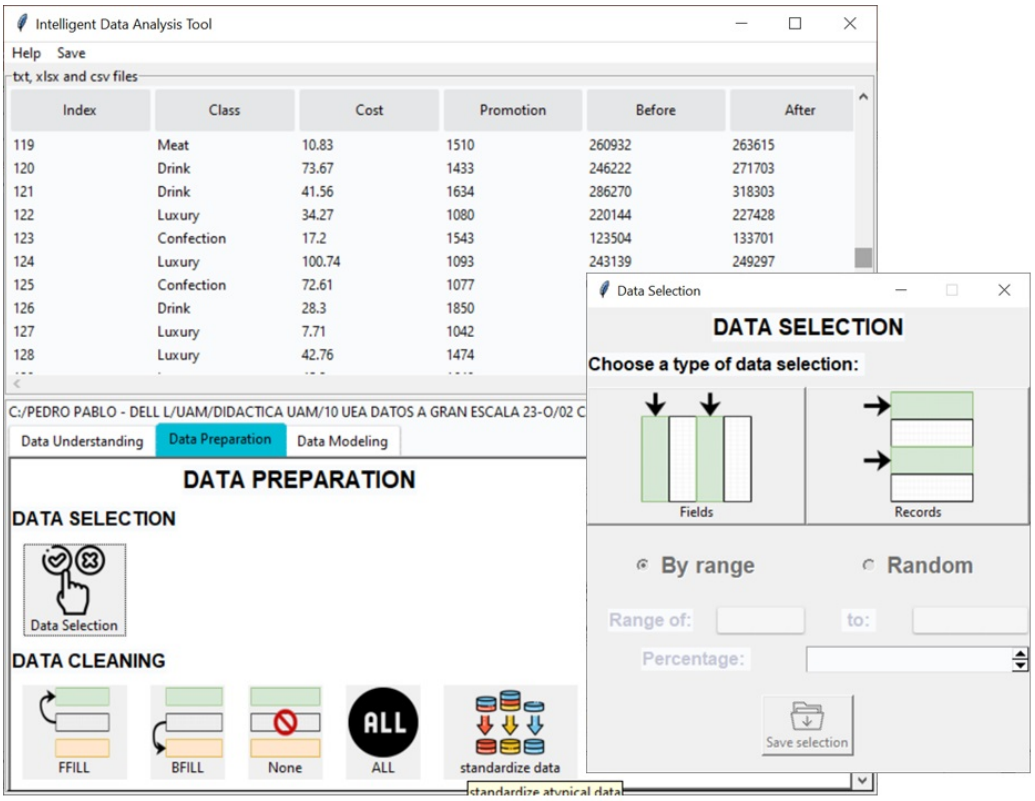


Figura 5.13. Funciones que proporciona la herramienta de minería de datos IDA-WEB TOOL para la selección de datos.

Figura 5.14. Funciones que proporciona la herramienta de minería de datos IBM SPSS MODELER para la selección de datos.

5.3.2. Limpieza de datos

La limpieza de datos se refiere a la corrección de problemas detectados durante la fase de comprensión de los datos, tales como:

- Datos perdidos (datos nulos o datos vacíos)
- Datos atípicos
- Errores en los datos
- Errores en la codificación de los datos
- Errores en la unidad de medida de los datos
- Errores en los metadatos

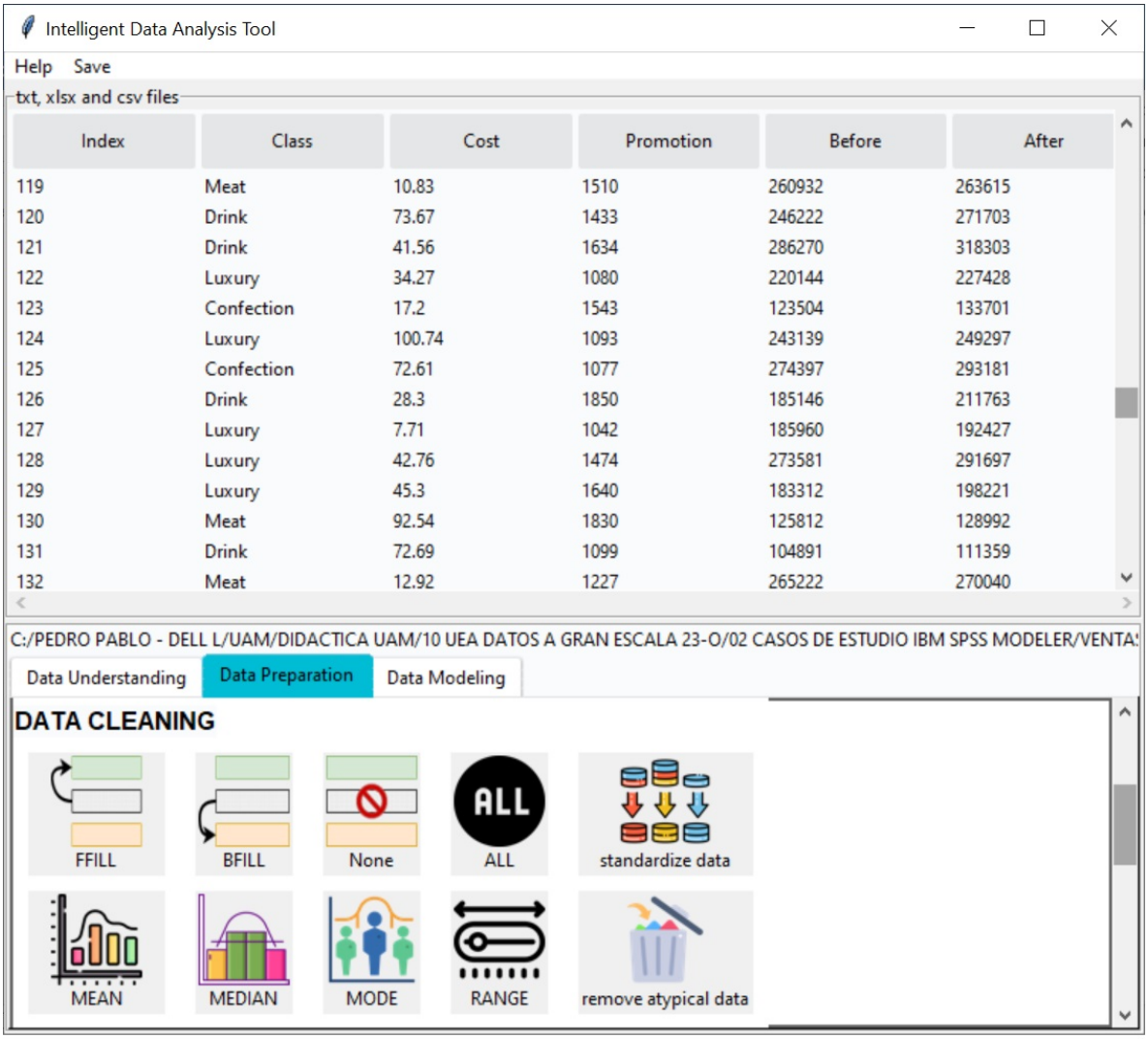


Figura 5.15. Funciones que proporciona la herramienta de minería de datos IDA-WEB TOOL para la limpieza de datos.

En esta actividad se requiere identificar qué tipo de ruido contienen los datos y qué métodos utilizar para eliminarlo. La mayoría de las herramientas que soportan parcial o completamente la minería de datos ofrece la opción de “limpieza de datos”, en la cual se incluye el reemplazo de datos perdidos, para la que se toma en cuenta el valor de la media, moda o mediana del atributo, según sea continuo, nominal u ordinal, respectivamente. Las figuras 5.15 y 5.16 muestran las funciones de soporte para la limpieza de datos que proporcionan las herramientas IDA-WEB TOOL e IBM SPSS MODELER, respectivamente.

Figura 5.16. Funciones que proporciona la herramienta de minería de datos IBM SPSS MODELER para la limpieza de datos.

5.3.3. Construcción de nuevos datos

Comúnmente, la construcción de nuevos datos se traduce en la derivación de nuevos atributos a partir de los ya existentes. Por ejemplo, valores porcentuales calculados a partir de dos atributos originales.

La mayoría de las herramientas que soportan parcial o completamente la minería de datos ofrece las opciones de “derivación de nuevos datos”, lo que permite la construcción de expresiones a partir de las cuales se construirán los nuevos datos, basándose en el uso de operadores matemáticos, lógicos, y relacionales. Las figuras 5.17 y 5.18 ilustran las funciones de soporte para la derivación de nuevos datos que proporcionan las herramientas IDA-WEB TOOL e IBM SPSS MODELER, respectivamente.

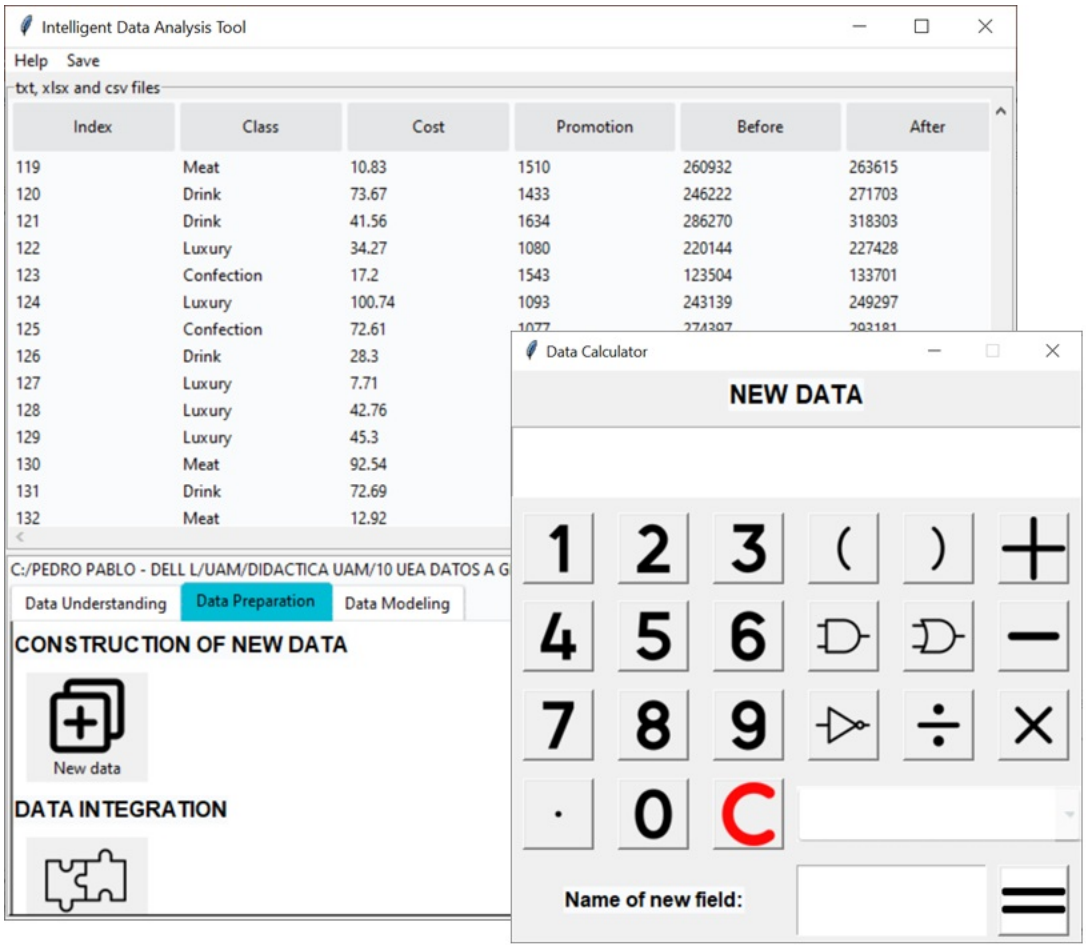


Figura 5.17. Funciones que proporciona la herramienta de minería de datos IDA-WEB TOOL para la construcción o derivación de nuevos datos.

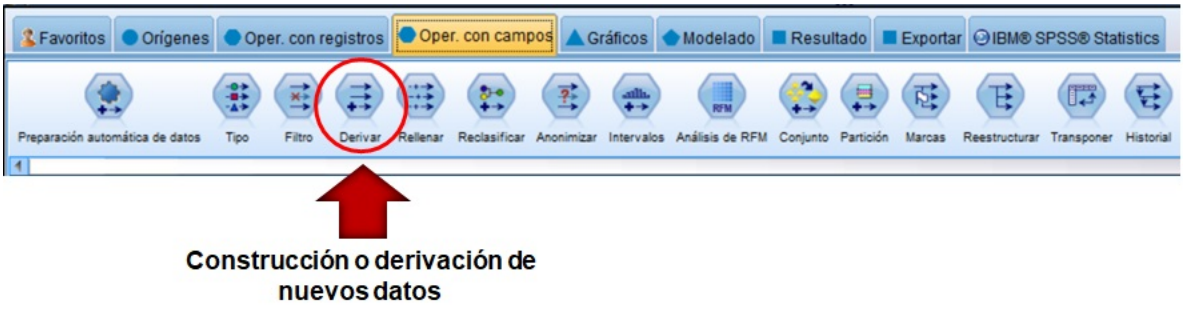


Figura 5.18. Funciones que proporciona la herramienta de minería de datos IBM SPSS MODELER para la construcción o derivación de nuevos datos.

5.3.4. Integración de datos

Se refiere al incremento del conjunto de datos, a partir de aquellos provenientes de otras fuentes (por ejemplo, tablas pertenecientes a la misma base de datos, tablas procedentes de otras fuentes, etcétera). Esta integración puede darse incorporando nuevos atributos (columnas) o nuevos registros (filas).

- **Incorporación de nuevos atributos**: Se refiere a la unión de dos conjuntos de datos con registros similares y atributos diferentes, lo cual resulta en un incremento de los atributos (o columnas) del conjunto de datos.
- **Incorporación de nuevos registros**: Se refiere a la unión de dos conjuntos de datos con atributos similares y registros diferentes, lo cual resulta en un incremento de los registros (o filas) del conjunto de datos.

Muchas de las herramientas que soportan parcial o completamente la minería de datos ofrecen opciones de integración de datos, permitiendo la incorporación de nuevos atributos o de nuevos registros. Las figuras 5.19 y 5.20 ilustran las funciones de soporte para la integración de datos que proporcionan las herramientas IDA-WEB TOOL e IBM SPSS MODELER, respectivamente.

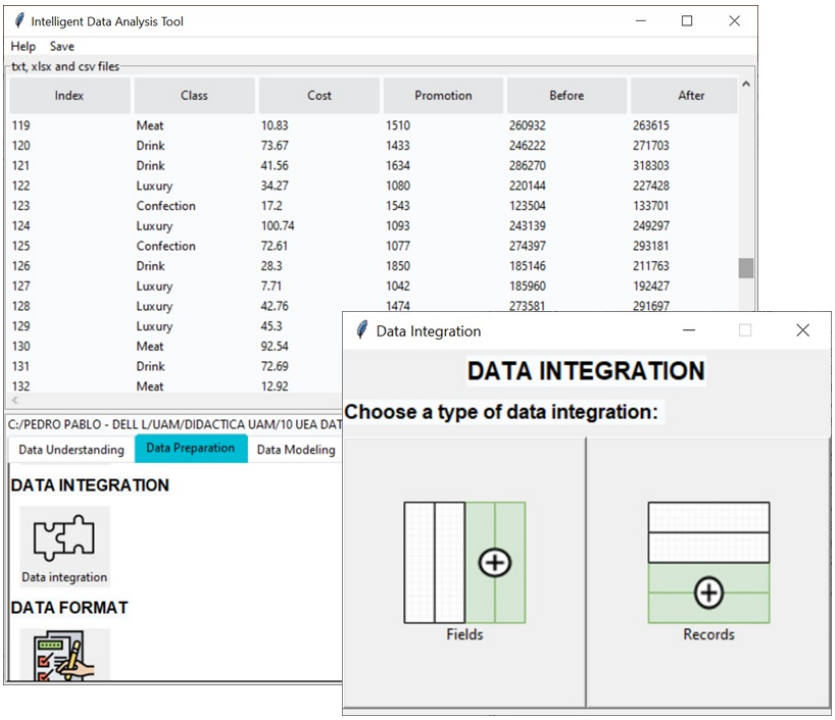


Figura 5.19. Funciones que proporciona la herramienta de minería de datos IDA-WEB TOOL para la integración de datos.

Figura 5.20. Funciones que proporciona la herramienta de minería de datos IBM SPSS MODELER para la integración de datos.

5.3.5. Formato de datos

El formato de datos se considera si el modelo, que posteriormente se construirá en la fase de modelado, requiere de algún orden o clasificación particular de los registros o los atributos, o de un tipo de datos particular para los atributos. La mayoría de las herramientas para el procesamiento de datos permiten otorgar el formato solicitado (ya sea del orden o del tipo de dato). Por ejemplo, al ordenar los datos para un modelo de clasificación supervisada, sería conveniente que los atributos que representan clases se ubiquen en las últimas columnas del conjunto de datos. Por otra parte, al ordenar los datos para un modelo de clasificación supervisada, es necesario que los atributos que representan clases sean de tipo categórico.

La mayoría de las herramientas para el procesamiento de datos permiten dar algún formato particular al conjunto de datos (ya sea orden o tipo de dato). Las figuras 5.21 y 5.22 ilustran las funciones de soporte para el formato de datos que proporcionan las herramientas IDA-WEB TOOL e IBM SPSS MODELER, respectivamente.

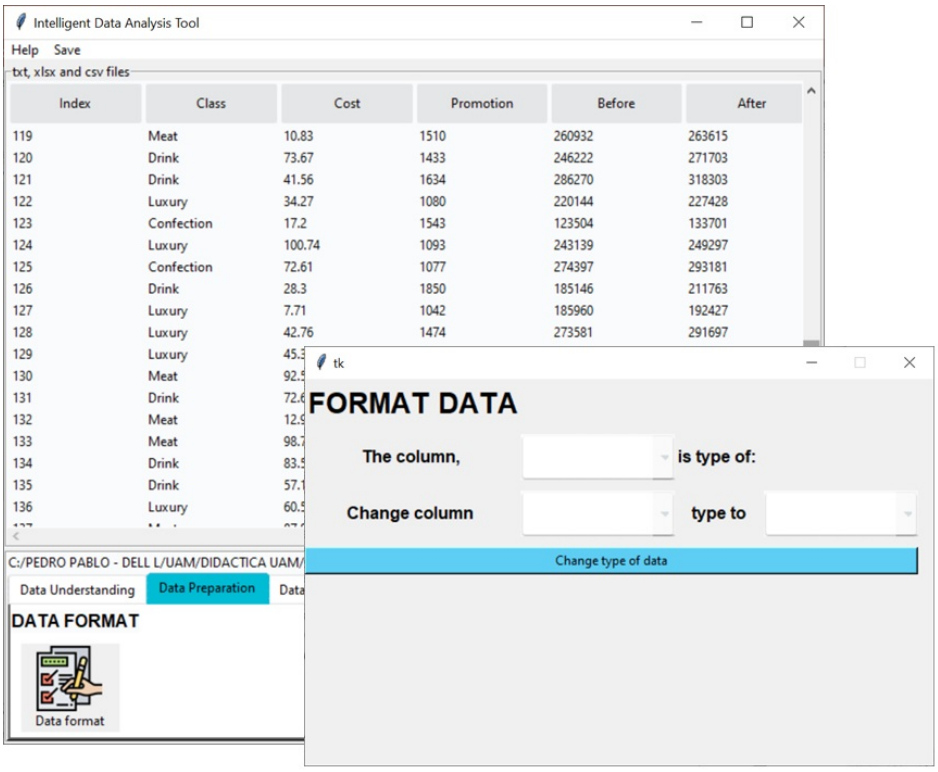


Figura 5.21. Funciones que proporciona la herramienta de minería de datos IDA-WEB TOOL para el formato de datos.

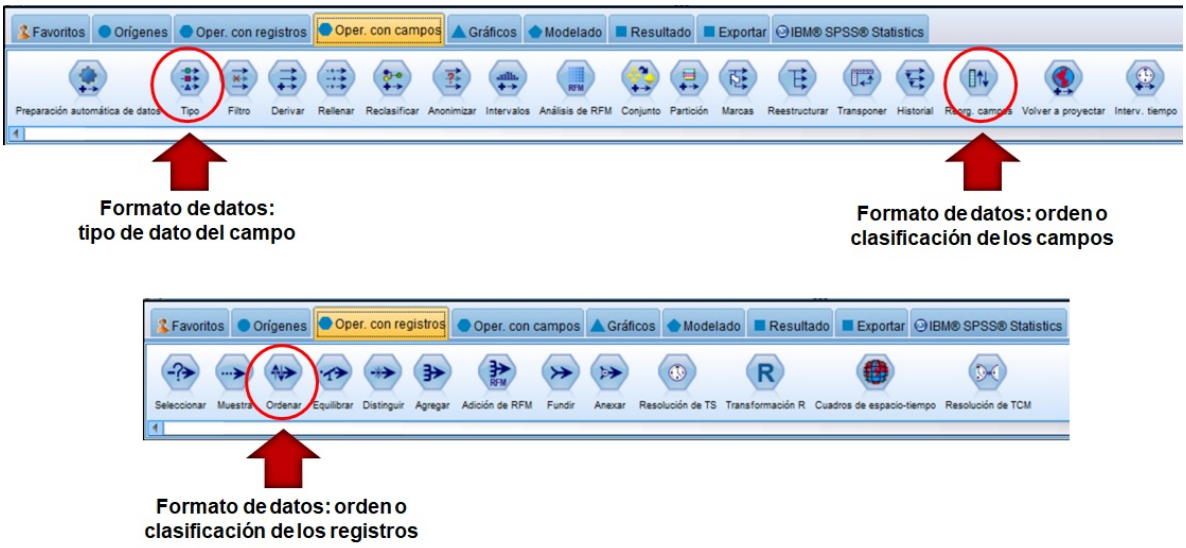


Figura 5.22. Funciones que proporciona la herramienta de minería de datos IBM SPSS MODELER para el formato de datos.

5.4. Fase de modelado de CRISP-DM

El modelado se refiere a la construcción del modelo de aprendizaje automatizado o algoritmo de regresión, a partir del conjunto de datos (Perez, 2021; Shearer, 2000). Esta fase no debe iniciarse si alguna de las siguientes cuestiones no ha sido manejada o considerada previamente:

- ¿Se ha garantizado un fácil acceso a los datos desde las herramientas de modelado a utilizar?
- ¿A través de la comprensión de los datos y de su exploración se ha podido identificar el subconjunto de datos más prometedores?
- ¿Se ha efectuado una adecuada limpieza de los datos?
- ¿Se conocen las herramientas de modelado?
- ¿Los datos se encuentran en el formato requerido por el modelo?
- En caso de haber efectuado la integración de datos, ¿existe algún problema en la unión o fusión de los mismos?

Comúnmente, cada uno de los modelos seleccionados se ejecuta inicialmente con los parámetros propuestos por defecto por el modelo. Posteriormente, pueden ser ajustados para ver si es posible mejorar los resultados que produce el modelo (bondad, error, coeficiente de correlación, etcétera).

Como se ilustra en la figura 5.23, la fase de modelado incluye las siguientes actividades:

- Selección de técnicas de modelado
- Métodos de comprobación
- Generación de los modelos
- Evaluación de los modelos

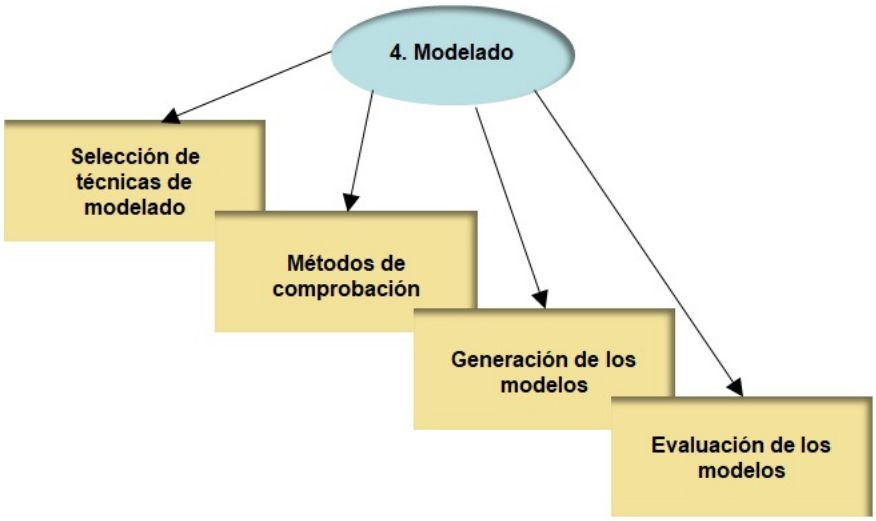


Figura 5.23. Actividades de la fase de modelado del enfoque metodológico de minería de datos CRISP-DM.

5.4.1. Selección de técnicas de modelado

La fase de modelado comúnmente se ejecuta utilizando varios modelos, de tal forma que se pueden obtener diferentes resultados a partir de un mismo conjunto de datos y, de esta forma, es posible comparar cuál fue el modelo que mejor se adecúa al objetivo del proyecto.

La elección de los modelos a utilizar depende estrechamente del objetivo de minería de datos propuesto. Por ejemplo:

- Si se trata de un problema de predicción, será necesario utilizar técnicas o métodos basado en la regresión, tales como:
  - Regresión lineal
  - Regresión no lineal
  - Modelos de aprendizaje supervisado

- Si se trata de un problema de clasificación, entonces lo adecuado sería utilizar modelos de aprendizaje supervisado, por ejemplo:
  - Perceptrón multicapas (MLP, de la sigla en inglés)
  - Máquina de vectores de soporte (VSM, de la sigla en inglés)
  - Máquina de vectores de soporte lineal (LVSM)
  - Algoritmo de los K vecinos más cercanos (K-NN)
  - Árboles de decisión

Para la selección del modelo o de los modelos adecuados se deben considerar los

siguientes aspectos:

- Los tipos de datos con los que se cuenta para la minería de datos. Es decir, considerar si son continuos, categóricos, ordinales, nominales, etcétera. Es común que cada modelo imponga restricciones sobre el tipo de datos, ya sean predictores u objetivos.
- Los objetivos de minería de datos. Por ejemplo, si es problema de predicción, problema de clasificación, problema de agrupamiento, etcétera.
- Requerimientos específicos del modelado. Por ejemplo: ¿qué tipos de resultados necesita que proporcione el modelo y cómo requiere que se presenten?

La tabla 5.1 ejemplifica la selección de modelos para un conjunto de objetivos de minería de datos pertenecientes a diferentes dominios.

Tabla 5.1. Ejemplos de la actividad de selección de técnicas de modelado según el tipo de objetivo de minería de datos

Dominio	Objetivo de minería de datos	Modelo
Mercado bursátil	❑ Predecir la variación semanal en el precio de un determinado tipo de acción bursátil.	❖ Método de regresión lineal, MLP, SVM, LSVM
	❑ Predecir el volumen semanal transferido para un determinado tipo de acción bursátil.	❖ Método de regresión lineal, MLP, SVM, LSVM
Mercadotecnia	❑ Predecir el incremento en las ventas de un determinado producto de consumo, como resultado de una promoción aplicada.	❖ Método de regresión lineal, MLP, SVM, LSVM
	❑ Clasificar los tipos de clientes de un *e-commerce*, según su poder de compra.	❖ MLP, K-NN, SVM, LSVM, regresión logística, árboles de decisión
Diagnóstico médico	❑ Clasificar un conjunto de personas como sanas o enfermas, considerando características que incluyen: antecedentes heredofamiliares, signos y síntomas, resultados de pruebas de laboratorio, entre otros.	❖ MLP, K-NN, SVM, LSVM, regresión logística, árboles de decisión
	❑ Predecir si una persona es susceptible de contraer una enfermedad infecciosa.	❖ Método de regresión lineal, MLP, SVM, LSVM
Gestión bancaria-crediticia	❑ Clasificar los clientes de crédito en “activos” y “con desgaste”.	❖ MLP, K-NN, SVM, LSVM, regresión logística, árboles de decisión
	❑ Predecir si un cliente de crédito	❖ Método de regresión

		abandonará los servicios de crédito.		lineal, MLP, SVM, LSVM
Medioambiente	❑	Predecir el índice de la calidad del aire en una determinada región.	❖	MLP, K-NN, SVM, LSVM, regresión logística, árboles de decisión
	❑	Clasificar un conjunto de regiones según la calidad del aire en “muy buena”, “buena”, “regular” y “mala”.	❖	MLP, K-NN, SVM, LSVM, regresión logística, árboles de decisión

5.4.2. Métodos de comprobación

Los métodos de comprobación se refieren a la forma en que se verificarán los resultados producidos por el modelo. Estos métodos deben incluir:

- **El criterio de bondad del modelo**: La bondad se refiere a una medición del desempeño del modelo. Por ejemplo, para modelos supervisados, el criterio de bondad se conoce por la tasa de error del modelo, usando métricas como exactitud, precisión, sensibilidad y F1 Score; mientras que, para modelos no supervisados, la bondad puede referirse a la facilidad de interpretación, tiempo de procesamiento, entre otros aspectos.
- **Nuevos datos para comprobar el criterio de bondad**: Para comprobar el criterio de bondad del modelo es necesario contar con nuevos datos, es decir, aquellos que no hayan sido suministrados al modelo como insumo, y para los cuales comúnmente se conoce o sospecha el resultado esperado, en términos de clase, grupo o resultado predictivo.

Los métodos de comprobación permiten constatar los resultados producidos por cada uno de los modelos seleccionados, antes de decidir cuál o cuáles se utilizarán. En la actualidad, muchas herramientas de minería de datos incluyen, entre sus prestaciones, la partición de los datos en dos conjuntos: uno para el entrenamiento del modelo y el otro para la prueba o verificación del mismo.

Como ya se mencionó, todos los modelos supervisados (tales como redes neuronales con aprendizaje supervisado, árboles de decisión supervisados, algoritmos de predicción supervisados, entre otros) requieren de un conjunto de datos de prueba para comprobar el desempeño del modelo.

5.4.3. Generación de los modelos

La generación se refiere a la ejecución de los modelos seleccionados para analizar los datos. Afortunadamente, existe una extensa gama de herramientas para realizar dicho análisis, que incluyen una amplia variedad de modelos inteligentes y de minería de datos implementados y disponibles para su uso. Por ejemplo: árboles de

búsqueda, modelos de predicción, algoritmos de agrupamiento, redes neuronales artificiales, entre otras técnicas.

La generación de un modelo comúnmente incluye tareas como:

- Selección y preparación de los datos para el modelo particular
- Configuración de los parámetros del modelo
- Ejecución del modelo
- Exploración de los resultados del modelo

Las figuras 5.24 y 5.25 ilustran las técnicas de modelado que proporcionan las herramientas de minería de datos IDA-WEB TOOL e IBM SPSS MODELER, respectivamente.

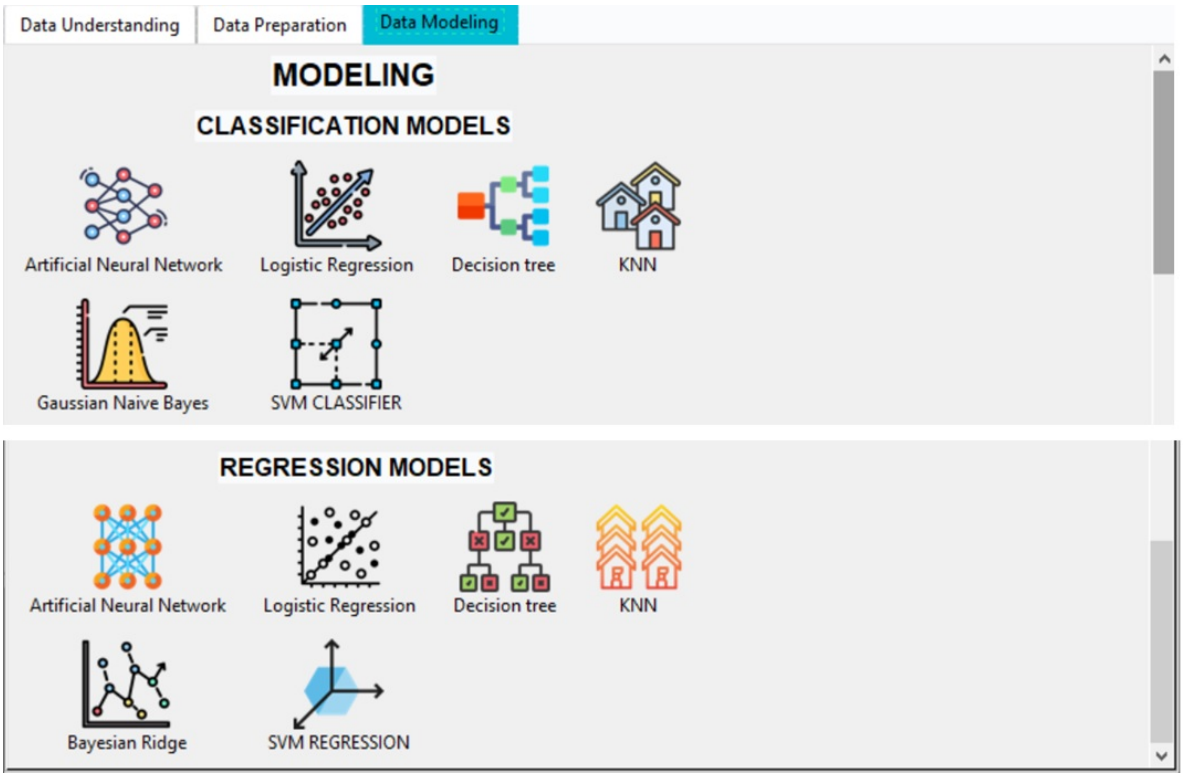


Figura 5.24. Técnicas de modelado que proporciona la herramienta de minería de datos IDA-WEB TOOL.

Figura 5.25. Técnicas de modelado que proporciona la herramienta de minería de datos IBM SPSS MODELER.

5.4.4. Evaluación de los modelos

Una vez seleccionados los modelos a utilizar y ya que han sido ejecutados sobre el conjunto de datos a analizar, se observan, mediante la evaluación, los resultados producidos por cada uno de ellos, para así decidir qué modelo resultó el más preciso. En el caso de modelos supervisados, se busca aquel con el menor error producido.

Un método comúnmente utilizado para la evaluación de los modelos es la exploración de los resultados producidos, a través de gráficos, tablas y métricas de desempeño. Una vez evaluados los modelos, pueden ser clasificados a partir de criterios como efectividad o precisión, tiempo de ejecución, facilidad para la interpretación de los resultados, etcétera.

5.4.4.1. La matriz de confusión y métricas de desempeño para la evaluación de los modelos de clasificación supervisada

En el aprendizaje supervisado, particularmente en la clasificación supervisada, la matriz de confusión es una herramienta que permite visualizar el desempeño del algoritmo. Como se puede apreciar en la figura 5.26, cada fila de la matriz representa el número de registros o instancias en la clase real, mientras que cada columna representa el número de predicciones efectuadas por el algoritmo de clasificación para cada clase. Esta herramienta permite visualizar qué tan bien está clasificando el algoritmo, o, si bien, está confundiendo las clases.

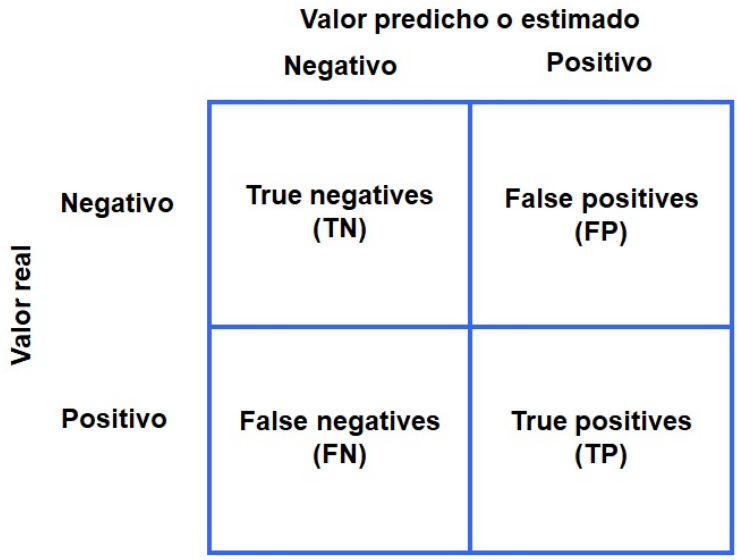


Figura 5.26. Matriz de confusión para tareas de clasificación binaria.

Nótese que en la figura 5.26 se identifican:

- **True negatives (TN)**: El algoritmo de clasificación predijo falso (o negativo), siendo esa la respuesta correcta.
- **True positives (TP)**: El algoritmo de clasificación predijo verdadero (o positivo), siendo esa la respuesta correcta.
- **False positives (FP)**: El algoritmo de clasificación predijo verdadero (o positivo), pero esa no es la respuesta correcta.
- **False negatives (FN)**: El algoritmo de clasificación predijo falso (o negativo), pero esa no es la respuesta correcta.

A partir de los valores de TN, TP, FP y FN de la matriz de confusión es posible calcular las siguientes métricas de desempeño del algoritmo de clasificación:

- **Exactitud (*accuracy*):** Porcentaje de predicciones verdaderas.

$$\boldsymbol{Exactitud} = (\boldsymbol{TN} + \boldsymbol{TP})/(\boldsymbol{TN} + \boldsymbol{TP} - \boldsymbol{FP} + \boldsymbol{FN})$$

- **Precisión (*precision*):** Porcentaje de predicciones positivas verdaderas.

$$\boldsymbol{Precisión} = \boldsymbol{TP}/(\boldsymbol{TP} + \boldsymbol{FP})$$

- **Sensibilidad o exhaustividad (*recall*):** Porcentaje de casos positivos detectados.

$$Sensibilidad = TP/(TP + FN)$$

- **Especificidad (*specify*):** Porcentaje de casos negativos detectados.

$$Especificidad = TN/(TN + FP)$$

- **F1 Score:** Es la media armónica entre sensibilidad y precisión. Esta métrica balancea el equilibrio entre falsos positivos y falsos negativos.

$$F1\ Score = 2\ (* Precisión * Sensibilidad)\ /\ ((Precisión + Sensibilidad)\ )$$

5.5. Fase de evaluación de CRISP-DM

En esta fase, los resultados producidos por el modelo de análisis son evaluados para asegurar que los objetivos del proyecto o de la organización sean alcanzados. Se refiere a la evaluación e interpretación de los resultados y descubrimientos producidos por el modelo, no a la evaluación del modelo en sí, ya que esto se realizó al final de la fase de modelado, durante la actividad de evaluación de los modelos. Para ejecutar esta fase, se deben comprender claramente los objetivos comerciales de la organización o, en su caso, los objetivos de proyecto en cuestión (ver ejemplo en figura 5.27).

Por ejemplo, para los objetivos comerciales relacionados con el posicionamiento, rentabilidad y retención de clientes de una plataforma e-commerce, los resultados globales podrían ser:

1. Recomendaciones sobre mejoras en los productos ofrecidos a los clientes
2. Recomendaciones sobre mejoras en el mecanismo de ventas cruzadas, de forma tal que el cliente no perciba un comportamiento voraz por parte del negocio
3. Recomendaciones sobre mejoras en el sitio web, de forma tal que ofrezca una mejor navegación e interacción al usuario.

Por ejemplo, para los objetivos comerciales de un estudio de mercadotecnia, relacionado con el incremento de las ventas de bienes de consumo a partir de la ejecución de determinadas promociones, los resultados globales podrían ser:

1. Recomendaciones de mejoras para que la promoción resulte mucho más atractiva, en aquellos tipos de bienes de consumo donde no se obtuvieron los resultados previstos
2. Recomendaciones sobre cuáles variaciones efectuar en el monto de la promoción, en dependencia de los resultados obtenidos para cada tipo de bien de consumo

Figura 5.27. Ejemplos de objetivos comerciales de la organización.

Los siguientes aspectos pueden proporcionar un gran soporte para la evaluación de los resultados producidos por el modelo (Shearer, 2000):

- Claridad con la que se expresan los resultados del modelo.
- Facilidad y claridad para la presentación de los resultados del modelo.
- Descubrimientos relevantes efectuados a partir de los resultados producidos por el modelo.
- Correspondencia de los resultados producidos por el modelo a los objetivos del proyecto (por ejemplo, objetivos comerciales de la compañía).
- Otras conclusiones adicionales generadas en los resultados producidos por el modelo.

5.6. Fase de presentación de los resultados de CRISP-DM

La fase de despliegue se refiere a la presentación final de los resultados, así como al uso de los nuevos conocimientos encontrados para responder al objetivo del proyecto: por ejemplo, en el caso de objetivos comerciales, podría efectuar mejoras en la organización o compañía, tales como un posicionamiento más alto, incremento de la rentabilidad, retención de clientes, etcétera.

En esta fase se decide si los nuevos patrones descubiertos en los resultados de los modelos son lo suficientemente fuertes como para conllevar a grandes cambios en las estrategias de la organización, cuando se trata de objetivos comerciales (ver ejemplo en figura 5.28). Por lo tanto, la fase de despliegue considera:

- Cómo utilizar los resultados producidos por la minería de datos para efectuar modificaciones y mejoras en la organización, cuando se trata de objetivos comerciales.
- Cómo utilizar los resultados producidos por la minería de datos para guiar las fases claves de un proyecto de investigación, cuando se trata de objetivos de investigación científica. Por ejemplo, la evaluación experimental, tomando como guía la predicción efectuada por el modelo.

Por ejemplo, para los objetivos comerciales relacionados con el posicionamiento, rentabilidad y retención de clientes de una plataforma e-commerce, acciones a tomar durante la fase "Despliegue" podrían ser:

1. Efectuar mejoras en los productos ofrecidos a los clientes
2. Efectuar mejoras en el mecanismo de ventas cruzadas, de forma tal que el cliente no perciba un comportamiento voraz por parte del negocio
3. Efectuar mejoras en el sitio Web, de forma tal que ofrezca una mejor navegación e interacción al usuario

Por ejemplo, para los objetivos comerciales de un estudio de mercadotecnia, relacionado con el incremento de las ventas de bienes de consumo a partir de la ejecución de determinadas promociones, acciones a tomar durante la fase "Despliegue" podrían ser:

1. Efectuar mejoras para que la promoción resulte mucho más atractiva, en aquellos tipos de bienes de consumo donde no se obtuvieron los resultados previstos
2. Efectuar variaciones en el monto de la promoción, en dependencia de los resultados obtenidos para cada tipo de bien de consumo

Figura 5.28. Ejemplos de la fase de despliegue en objetivos comerciales.

Comúnmente, la fase de despliegue conlleva a la elaboración de un informe final que incluye los siguientes aspectos:

- Descripción detallada del problema original.
- Resumen del procedimiento utilizado para llevar a cabo la minería de datos.
- Resumen de los resultados del proyecto de minería de datos, incluyendo modelos, resultados, nuevos conocimientos, etcétera.
- Resumen del plan propuesto para el despliegue.
- Recomendaciones para futuros proyectos de análisis inteligente de datos a gran escala.

El desarrollo de la fase de despliegue requiere de mayor argumentación, discusión y documentación. Sin embargo, de ser necesario, se pueden utilizar las herramientas, aplicaciones y paquetes de cómputo relacionados en la fase de modelado.

VI. LA HERRAMIENTA DE MINERÍA DE DATOS IDA-WEB TOOL

6.1. Alcance de la herramienta IDA-WEB TOOL

Intelligent Data Analysis Tool (IDA-WEB TOOL) es una herramienta computacional de soporte para las principales actividades de la minería de datos, tales como comprensión del dominio del problema, comprensión de los datos, exploración de los datos, preparación de los datos y modelado (González Pérez *et al.*, 2023). De forma particular, IDA-WEB TOOL brinda al usuario un apoyo automatizado e integrado para ejecutar las principales fases de metodologías de minería de datos comúnmente conocidas, como las ya mencionadas CRISP-DM y SMART MODEL.

La herramienta de minería de datos IDA-WEB TOOL, incluyendo los programas fuentes y documentos técnicos que contiene, ha sido liberada en: http://bioinformatics.cua.uam.mx/ida-tool/, y se encuentra disponible para su uso por la comunidad académica, es decir, quienes se dedican al análisis inteligente de datos y, principalmente, los alumnos de las carreras en Ingeniería en Computación, Ingeniería de *Software*, Matemáticas Aplicadas, o afines, enfocadas a tareas de minería de datos.

Entre las principales actividades de minería de datos que IDA-WEB TOOL proporciona al usuario se encuentran las siguientes:

- Comprensión del dominio del problema
  - Plan de proyecto de minería de datos
- Comprensión de los datos
  - Gestión de archivos en formato .xlsx, .csv y .txt
    - Abrir archivo
    - Guardar archivo
    - Guardar como…
    - Conversión de archivos
  - Graficación de datos
    - Graficación de datos con gráfico de barras horizontal
    - Graficación de datos con gráfico de barras vertical
    - Graficación de datos con gráfico de dispersión
    - Graficación de datos con gráfico lineal
    - Graficación de datos con gráfico de área
    - Graficación de datos con gráfico histograma
    - Graficación de datos con gráfico de pastel

- Preparación de los datos
  - Selección de datos
  - Limpieza de datos
    - Relleno de datos perdidos con el método Forward Fill (FFILL)
    - Relleno de datos perdidos con el método Backward Fill (BFILL)
    - Relleno de datos perdidos utilizando la media
    - Relleno de datos perdidos utilizando la moda
    - Relleno de datos perdidos utilizando la mediana
    - Estandarización/normalización de datos
    - Eliminar datos atípicos
  - Construcción de nuevos datos
  - Integración de datos
  - Formato de datos
- Modelado para problemas de clasificación
  - Redes neuronales artificiales supervisadas, tales como Multi-Layer Perceptron (MLP) y Support Vector Machine (SVM)
  - Árboles de decisión
  - Algoritmo de los K vecinos más cercanos (K-NN)
  - Algoritmo de clasificación probabilística (Gaussian Naive Bayes)
  - Algoritmos de regresión logística
- Modelado para problemas de predicción
  - Redes neuronales artificiales supervisadas, tales como Multi-Layer Perceptron (MLP) y Support Vector Machine (SVM)
  - Árboles de decisión
  - Algoritmo de los K vecinos más cercanos (K-NN)
  - Algoritmos de regresión
  - Regresión lineal bayesiana

6.2. Manual de usuario

6.2.1. Requerimientos para la instalación de IDA-WEB TOOL

Los requerimientos para la instalación de la herramienta de minería de datos IDA-WEB TOOL son los siguientes:

1. Contar con una versión de Python igual o superior a la 3.10. La descarga de Python se puede efectuar utilizando el siguiente enlace: https://www.python.org/downloads/
2. Garantizar que durante la instalación de Python también se instale PIP, el sistema de gestión de paquetes de Python requerido para la instalación y

administración de *software* escrito en Python; así como *tkinter*, un componente de la biblioteca gráfica Tcl/Tk de Python.

6.2.2. Instalación de IDA-WEB TOOL

La herramienta IDA-WEB TOOL se puede descargar directamente desde el sitio web http://bioinformatics.cua.uam.mx/ida-tool/, tal como se muestra en la figura 6.1.

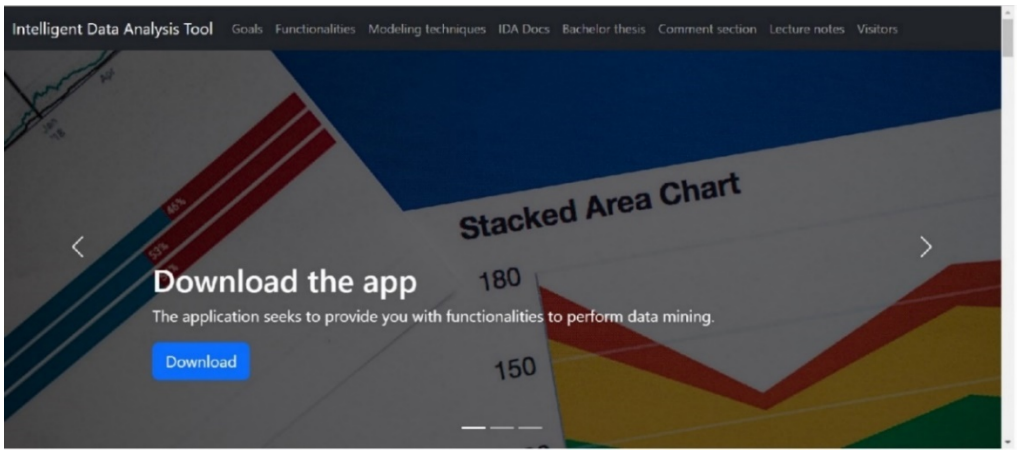


Figura 6.1. Interfaz gráfica de usuario principal del sitio web **http://bioinformatics.cua.uam.mx/ida-tool/** desde el cual es posible efectuar la descarga de la herramienta IDA-WEB TOOL.

Una vez presionado el botón “Download”, se desplegará la interfaz gráfica que se ilustra en la figura 6.2, desde la cual es posible iniciar la descarga de la herramienta IDA-WEB TOOL.

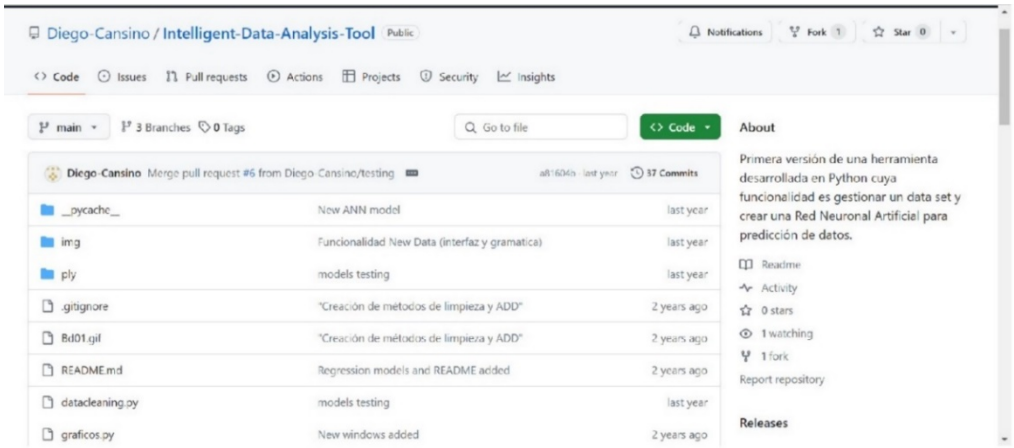


Figura 6.2. Interfaz gráfica desde la cual se inicia la descarga de la herramienta IDA-WEB TOOL.

Al presionar el botón “Code” (ver figura 6.2) se abrirá una ventana emergente (ver figura 6.3), desde la cual se puede seleccionar la opción “Download ZIP” para proceder a la descarga del archivo comprimido que contiene el código fuente y los archivos requeridos para la correcta instalación y ejecución de IDA-WEB TOOL.

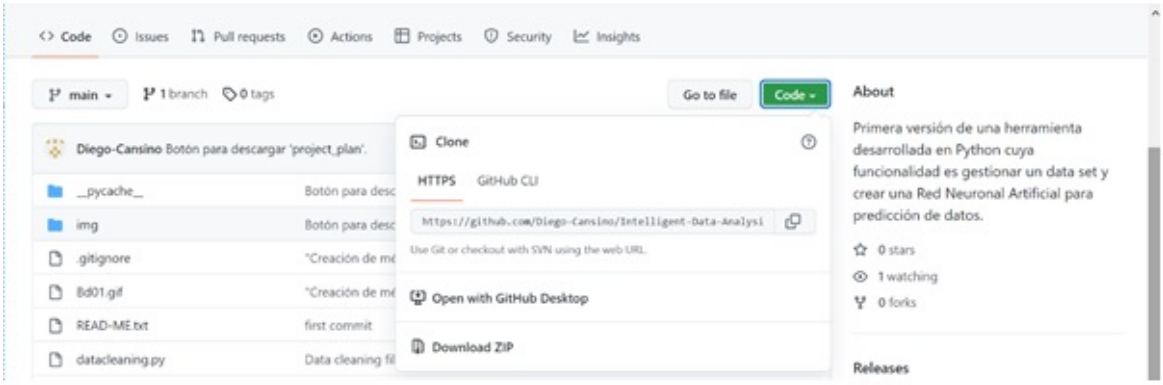


Figura 6.3. Interfaz gráfica desde la cual se selecciona la opción para iniciar la descarga de la herramienta IDA-WEB TOOL.

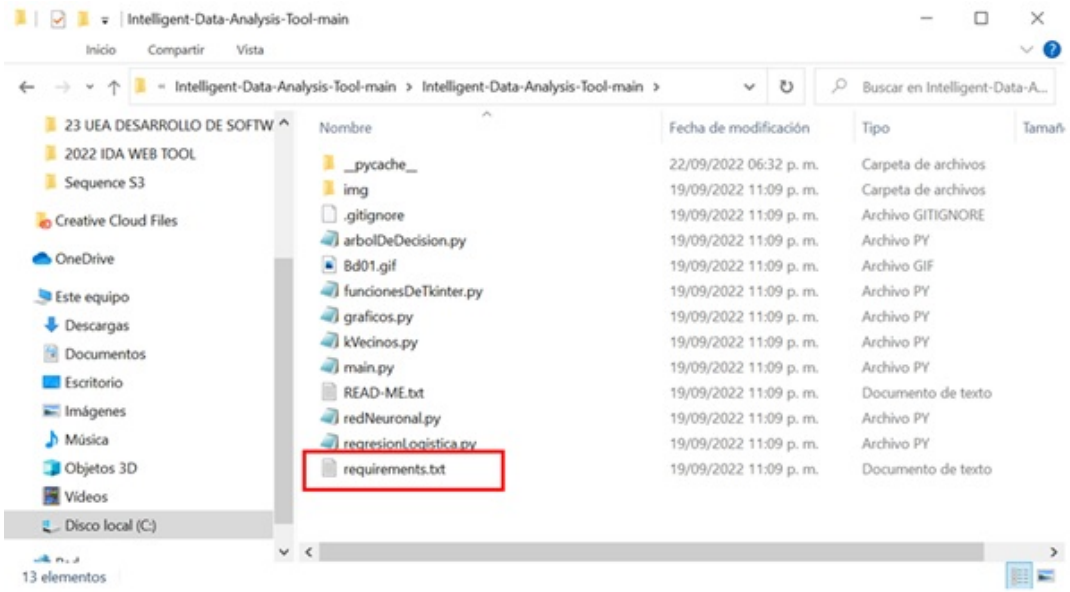


Figura 6.4. Carpeta “Intelligent-Data-Analysis-Tool-main” descomprimida, al interior de la cual se encuentra el archivo de texto “requirements.txt”, que contiene todas las dependencias que se utilizaron en el desarrollo de la herramienta IDA-WEB TOOL.

Una vez concluida la descarga, se debe proceder a descomprimir el archivo “Intelligent-Data-Analysis-Tool-main.zip”; el contenido de la carpeta se ilustra en la figura 6.4. Como se puede observar, al interior de la carpeta descomprimida “Intelligent-Data-Analysis-Tool-main” se localiza el archivo de texto “requirements.txt”, el cual contiene todas las dependencias utilizadas en el desarrollo de IDA-WEB TOOL, y que deben instalarse para ejecutar dicha herramienta.

Como se ilustra en la figura 6.5, para instalar estas dependencias, se debe ejecutar el siguiente comando por terminal: “pip install -r requirements.txt”.

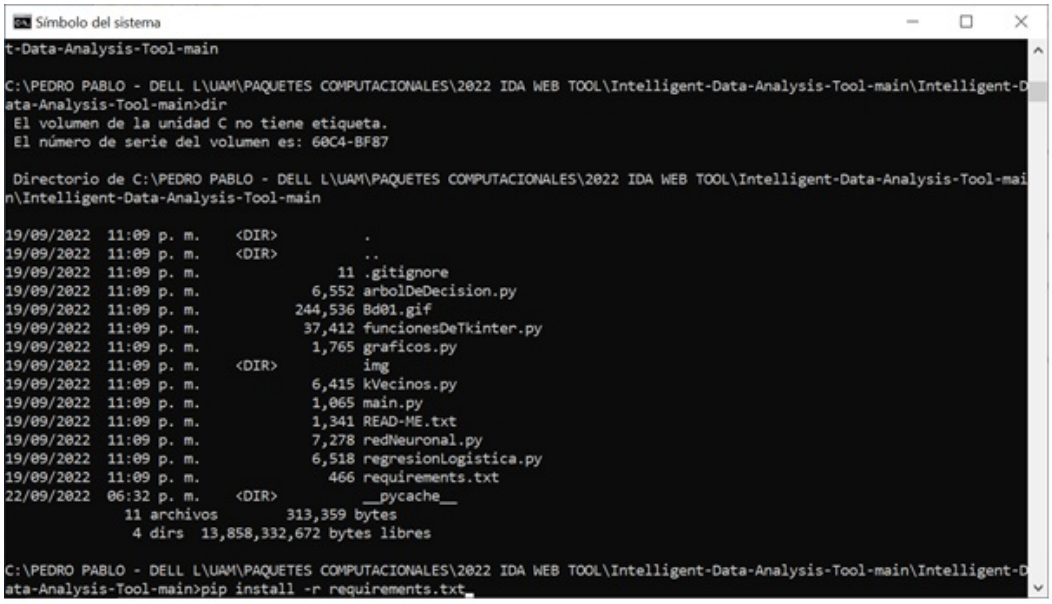


Figura 6.5. Ejecución en la terminal del sistema del comando “pip install -r requirements.txt”, lo cual permite la instalación de todas las dependencias que se utilizaron en el desarrollo de la herramienta IDA-WEB TOOL.

Una vez concluida la instalación de todas las dependencias, la herramienta IDA-WEB TOOL estará lista para su lanzamiento en ejecución, la cual se puede llevar a cabo a través de dos modalidades:

1. Utilizando el entorno de desarrollo integrado de Python (IDLE, por sus siglas en inglés), tal como se ilustra en las figuras de la 6.6 a la 6.10. Como se puede apreciar en la figura 6.9, la ejecución de la herramienta IDA-WEB TOOL se lanza en el IDLE, con el programa “main.py”.
2. Directamente en la terminal del sistema, utilizando el comando “$> python main.py”, tal como se muestra en la figura 6.11.

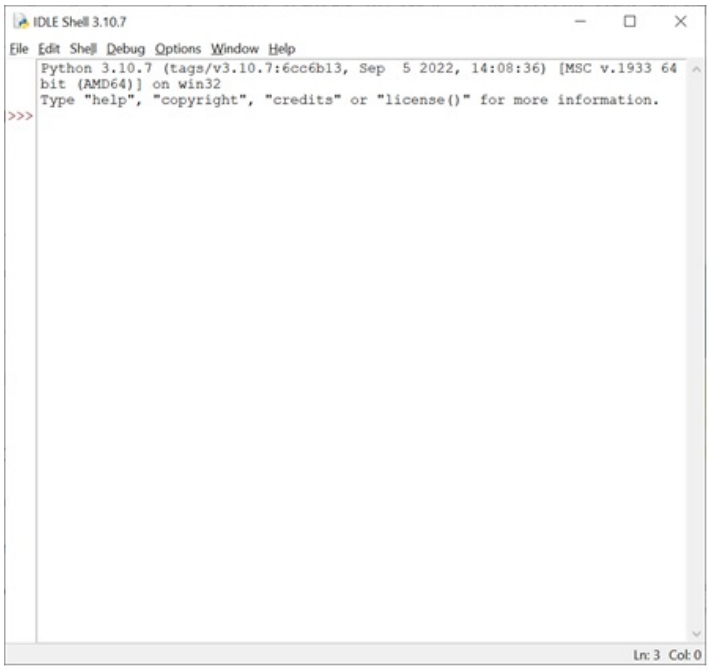


Figura 6.6. El entorno de desarrollo integrado de Python, a través del cual se lanza en ejecución la herramienta IDA-WEB TOOL.

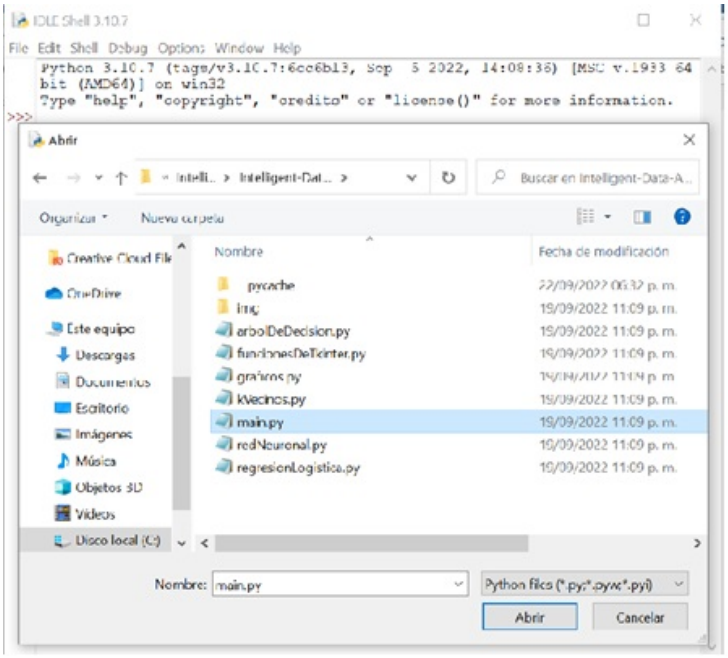


Figura 6.7. Carga del programa “main.py” en el IDLE de Python, para ejecutar la herramienta IDA-WEB TOOL.

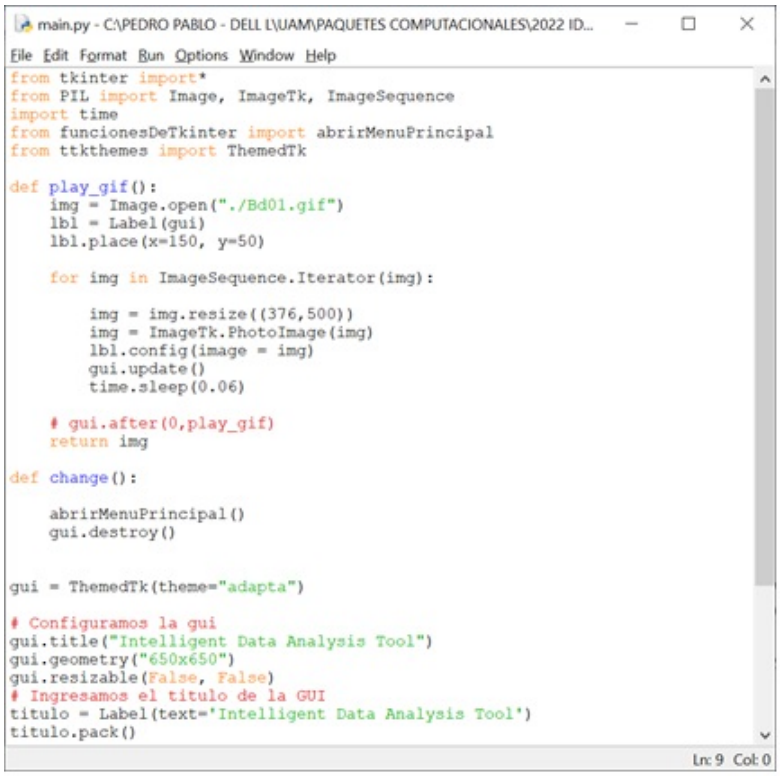


Figura 6.8. Despliegue del programa “main.py” en el IDLE de Python, para lanzar en ejecución la herramienta IDA-WEB TOOL.

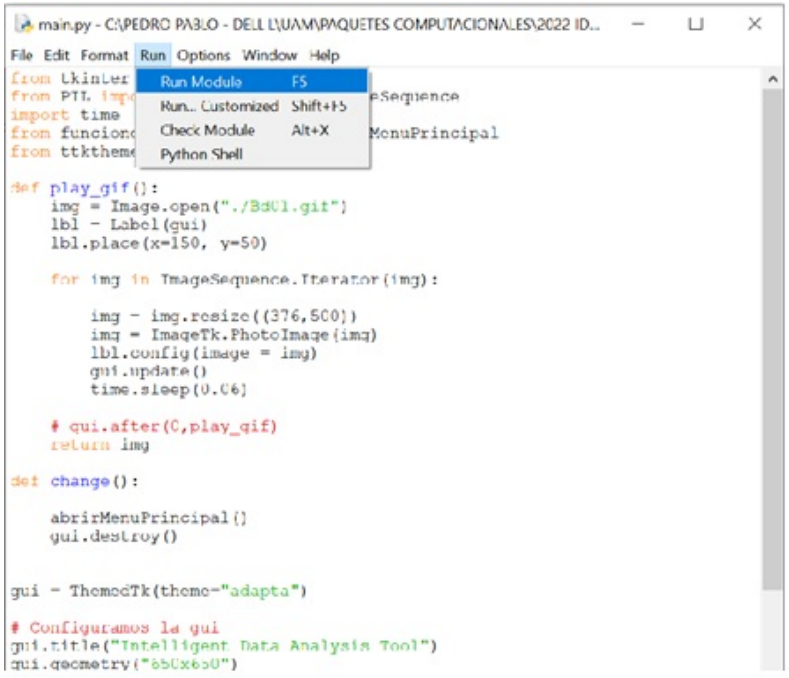


Figura 6.9. Con la opción “Run Module” del submenú “Run” se ejecuta la herramienta IDA-WEB TOOL.

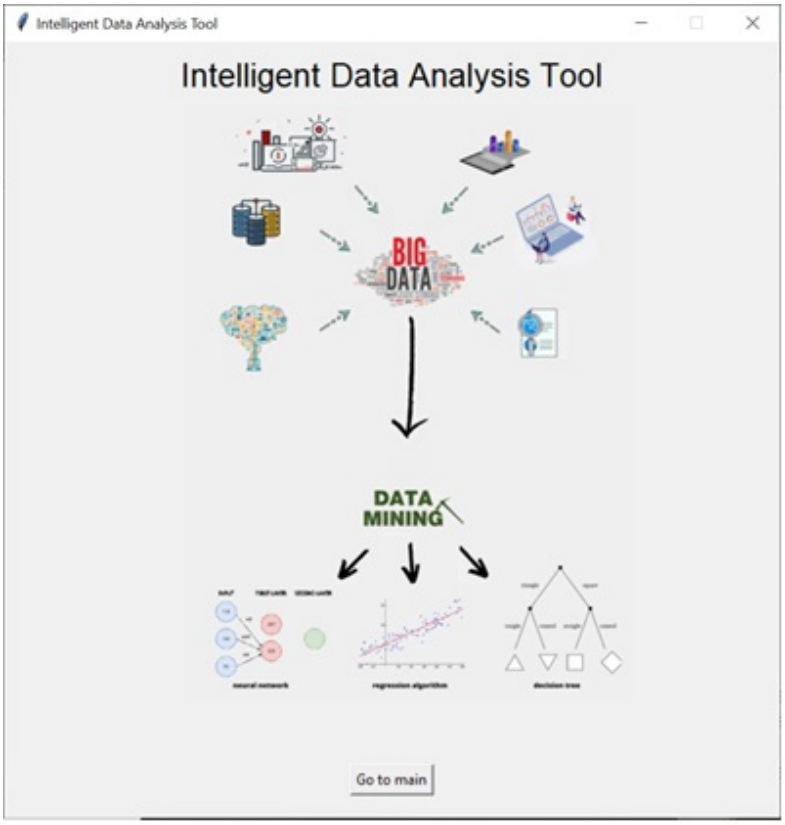


Figura 6.10. La herramienta IDA-WEB TOOL en ejecución. Interfaz gráfica principal.

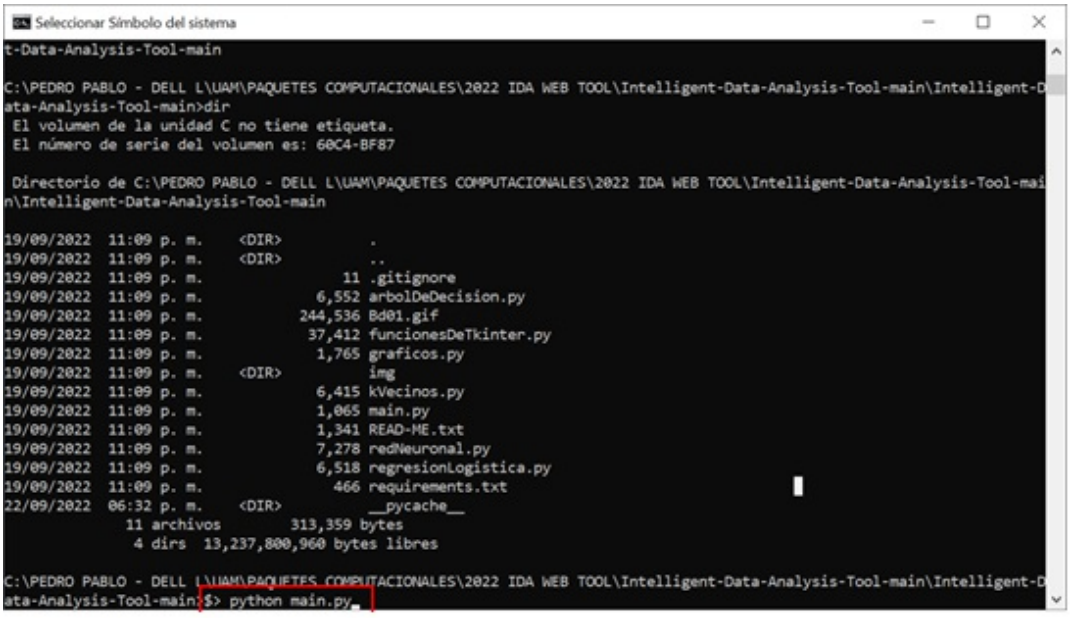


Figura 6.11. Ejecución de la herramienta IDA-WEB TOOL desde la terminal del sistema, utilizando el comando “$> python main.py”.

6.3. Demostración de operación

Para ilustrar la operación de la herramienta IDA-WEB TOOL, se ejecutarán las principales actividades que conlleva la minería de datos, según el enfoque metodológico CRISP-DM, sobre un conjunto de datos que describe el comportamiento semanal de un grupo de acciones que cotizan en el mercado bursátil de Nueva York, de forma particular, aquellas correspondientes al "Índice Dow Jones". El conjunto de datos, nombrado "Dow_Jones_Index_Dataset", se encuentra disponible en el UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/) (Dua y Graff, 2019).

6.3.1. Comprensión del dominio del problema

Como se puede apreciar en las figuras 6.12 y 6.13, al presionar el botón "Go to main" (GUI, por sus siglas en inglés) en la interfaz gráfica principal de la herramienta IDA-WEB TOOL (figura 6.12), se despliega la GUI que permite acceder y ejecutar las diferentes actividades involucradas en la minería de datos, tales como: comprensión de los datos, preparación de los datos y modelado (figura 6.13).

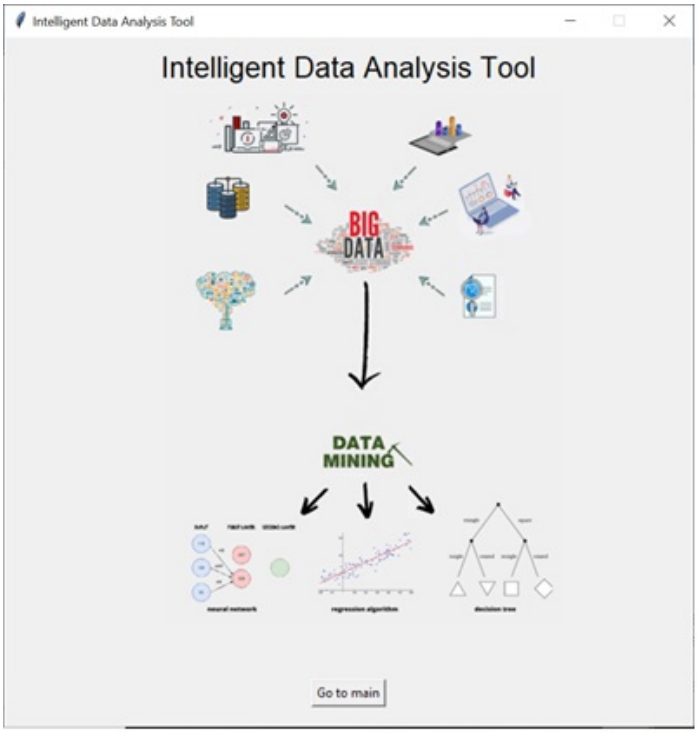


Figura 6.12. GUI principal de la herramienta IDA-WEB TOOL, desplegada al presionar el botón "Go to main".

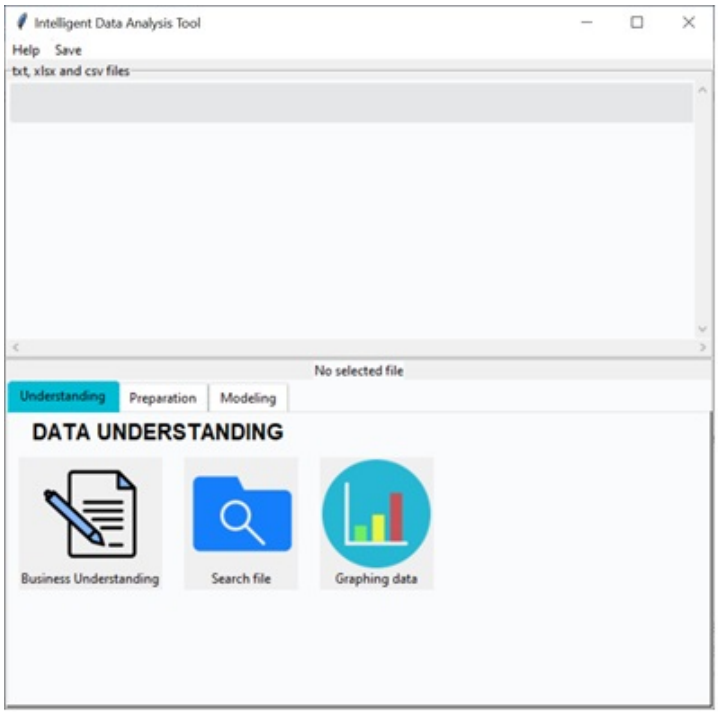


Figura 6.13. GUI de la herramienta IDA-WEB TOOL que permite acceder y ejecutar las diferentes actividades de minería de datos.

Como se ilustra en la figura 6.14, al seleccionar en la barra de herramientas de minería de datos la opción “Business Understanding”, se proporciona al usuario una tabla para que registre el plan del proyecto de minería de datos, como parte de la fase de comprensión del dominio del problema.

Dicha fase resulta de gran importancia y ejerce un gran impacto en las fases sucesivas, independientemente de la metodología o enfoque de minería de datos que se haya seleccionado. Esto se debe a que es en ésta donde se define de forma clara el problema que se intenta resolver, se focaliza en la comprensión de las metas u objetivos del trabajo a desarrollar y se proporciona una perspectiva de minería de datos que permite comprender qué datos deben ser analizados.

Como se puede apreciar en la figura 6.14, la comprensión del dominio del problema o negocio puede llevarse a cabo ejecutando las siguientes actividades:

- Determinación de los objetivos del proyecto
- Valoración de la situación actual del objetivo del proyecto
- Determinación de los objetivos de minería de datos
- Propuesta del enfoque metodológico para desarrollar el proyecto

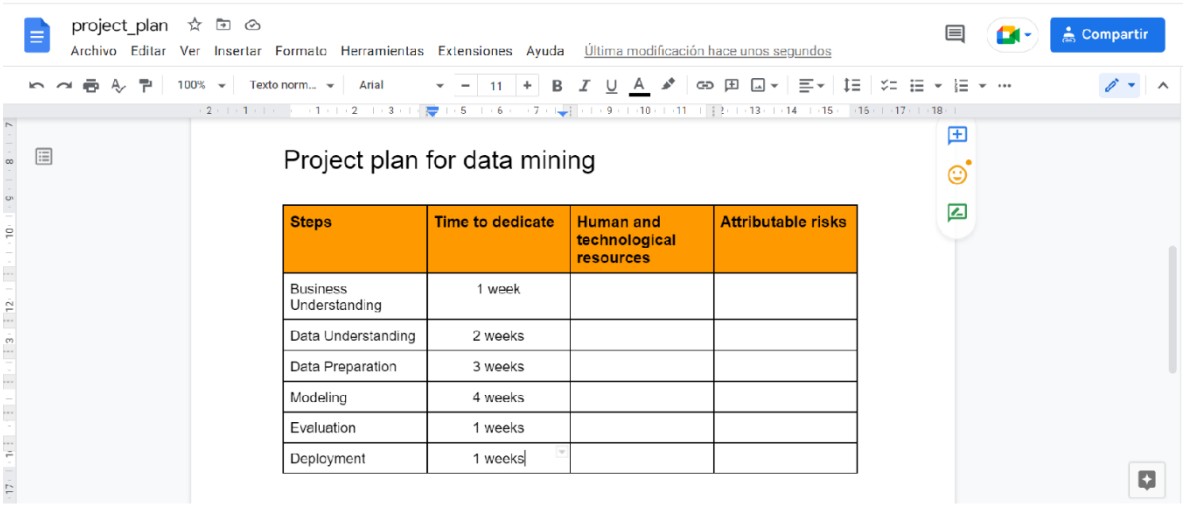


Figura 6.14. Registro del plan del proyecto de minería de datos en la herramienta IDA-WEB TOOL, como parte de la fase de comprensión del dominio.

6.3.2. Comprensión de los datos

Las figuras 6.15 y 6.16 muestran la carga y despliegue, respectivamente, del archivo .xlsx que contiene el conjunto de datos del "Índice Dow Jones", el cual será utilizado como caso de estudio para ilustrar las diferentes actividades de minería de datos que proporciona la herramienta IDA-WEB TOOL. Como se puede apreciar en la figura 6.15, para cargar o abrir un archivo de datos es necesario seleccionar en la barra de herramientas la opción "Data understanding" y posteriormente el submenú "Search file".

Una vez que el conjunto de datos ha sido cargado y desplegado, se puede proceder a la exploración del mismo, utilizando el ícono "Graphing data" de la opción "Data understanding" en la barra de herramientas de minería de datos.

La figura 6.17 muestra la gama de gráficos que proporciona IDA-WEB TOOL al usuario. Por otra parte, las figuras 6.18 a la 6.21 ilustran varios gráficos elaborados para el conjunto de datos del "Índice Dow Jones".

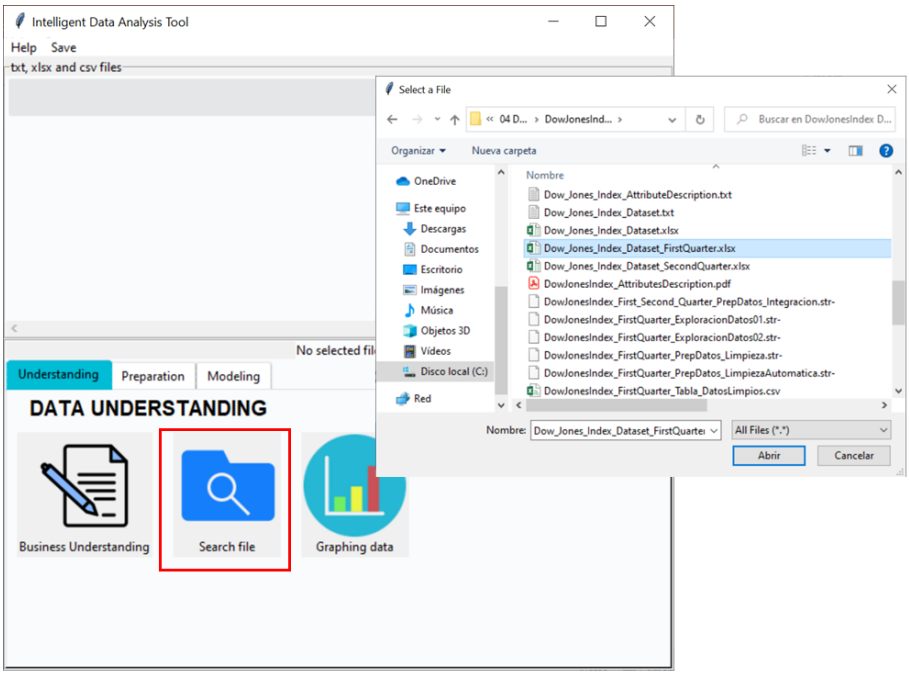


Figura 6.15. Carga del archivo .xlsx a través del menú “Understanding” y el submenú “Search file” en la herramienta IDA-WEB TOOL.

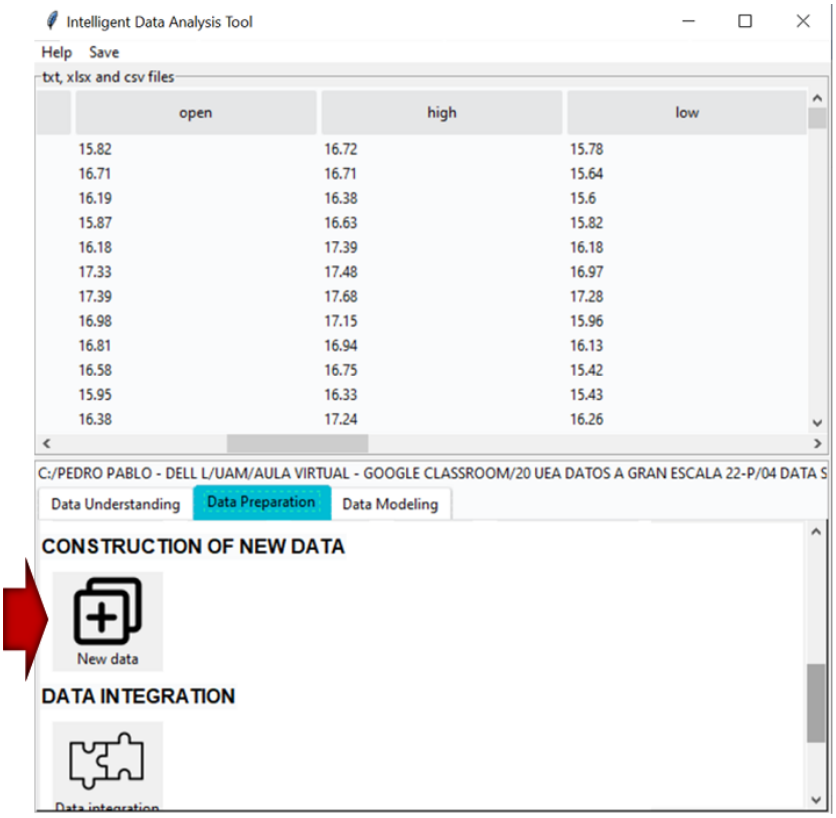


Figura 6.16. Despliegue del archivo .xlsx a través del menú “Understanding” y el submenú “Search file” en la herramienta IDA-WEB TOOL.

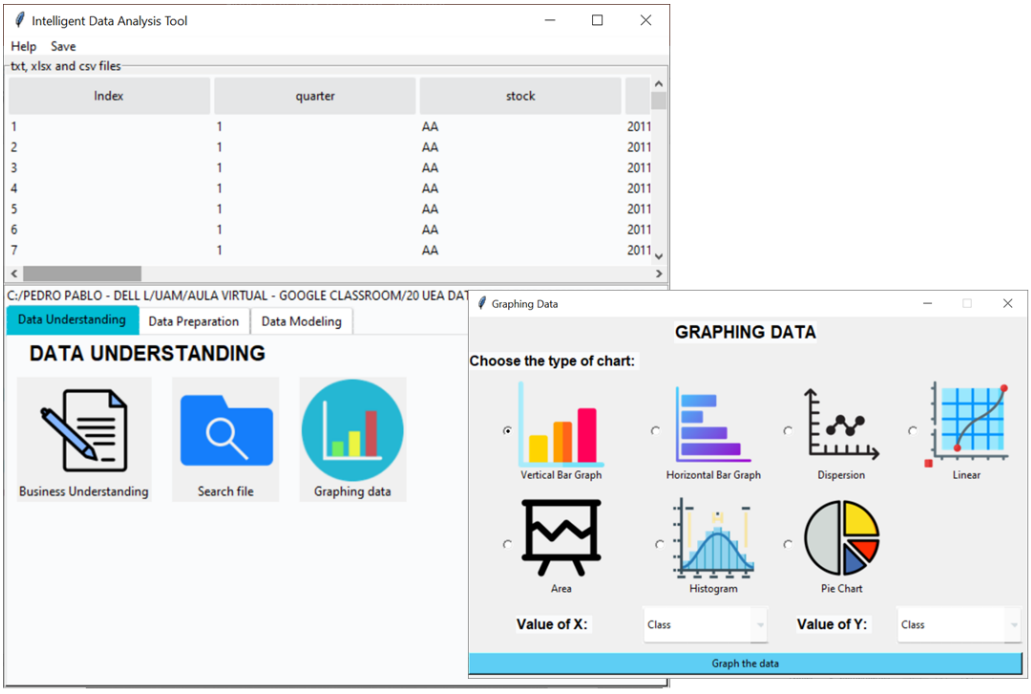


Figura 6.17. Gama de gráficos que proporciona la herramienta IDA-WEB TOOL para la exploración de datos.

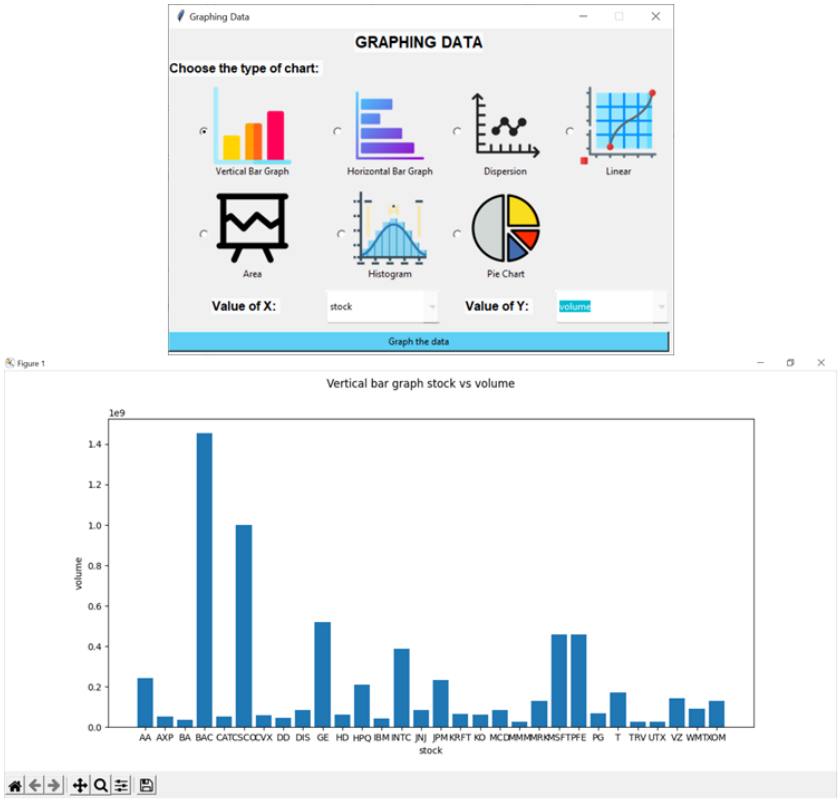


Figura 6.18. Exploración de datos utilizando el gráfico “Vertical Bar Graph” en la herramienta IDA-WEB TOOL.

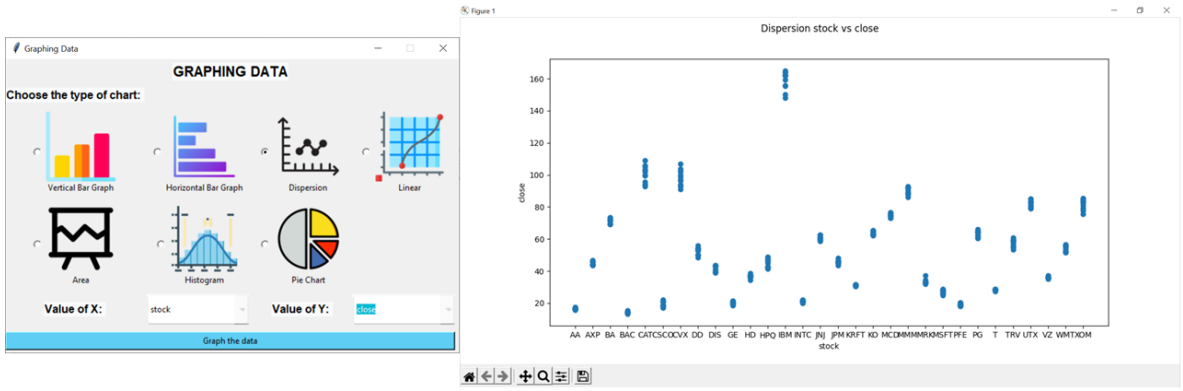


Figura 6.19. Exploración de datos utilizando el gráfico “Dispersion” en la herramienta IDA-WEB TOOL.

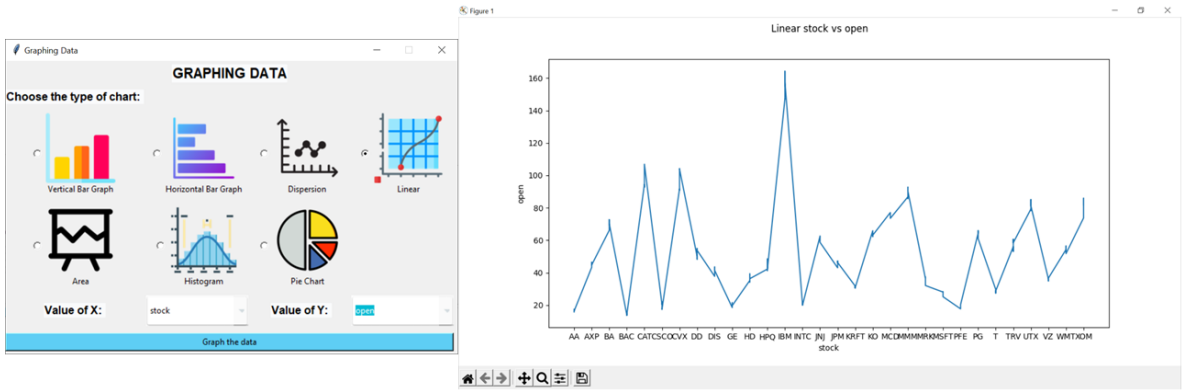


Figura 6.20. Exploración de datos utilizando el gráfico “Linear” en la herramienta IDA-WEB TOOL.

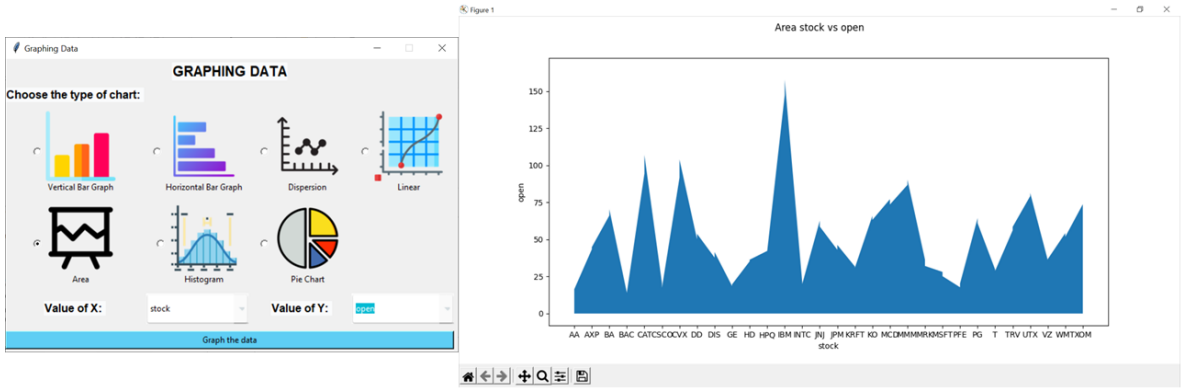


Figura 6.21. Exploración de datos utilizando el gráfico “Area” en la herramienta IDA-WEB TOOL.

6.3.3. Preparación de los datos

Como se puede apreciar en la figura 6.22, al seleccionar la opción “Data Preparation” de la barra de herramientas de minería de datos, IDA-WEB TOOL pone a disposición del usuario las siguientes actividades, que comúnmente caracterizan la fase de preparación de los datos:

- “Data Selection” (Selección de datos)
- “Data Cleaning” (Limpieza de datos)
- “Construction of New Data” (Derivación de nuevos datos)
- “Data Integration” (Integración de datos)
- “Data Format” (Formato de datos)

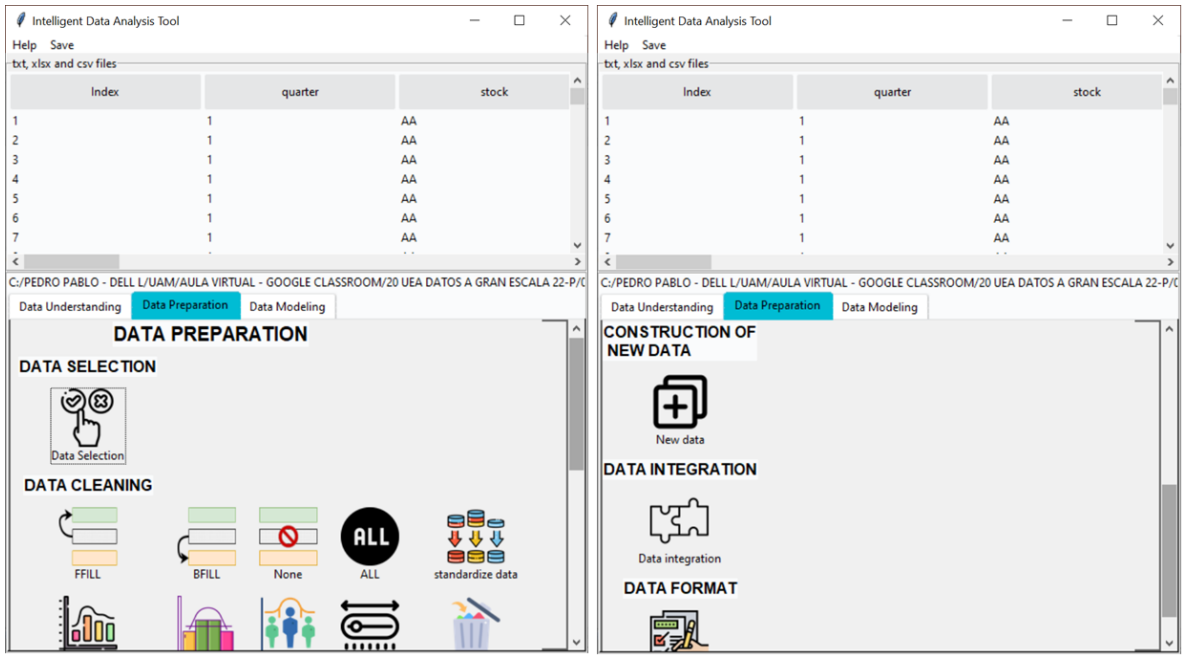


Figura 6.22. La opción “Data Preparation” en la barra de herramientas de minería de datos de IDA-WEB TOOL.

Como se puede apreciar en la figura 6.23, al pasar el cursor encima del ícono que representa cada método de limpieza de datos, se despliega una ventana emergente con una breve explicación acerca del mismo.

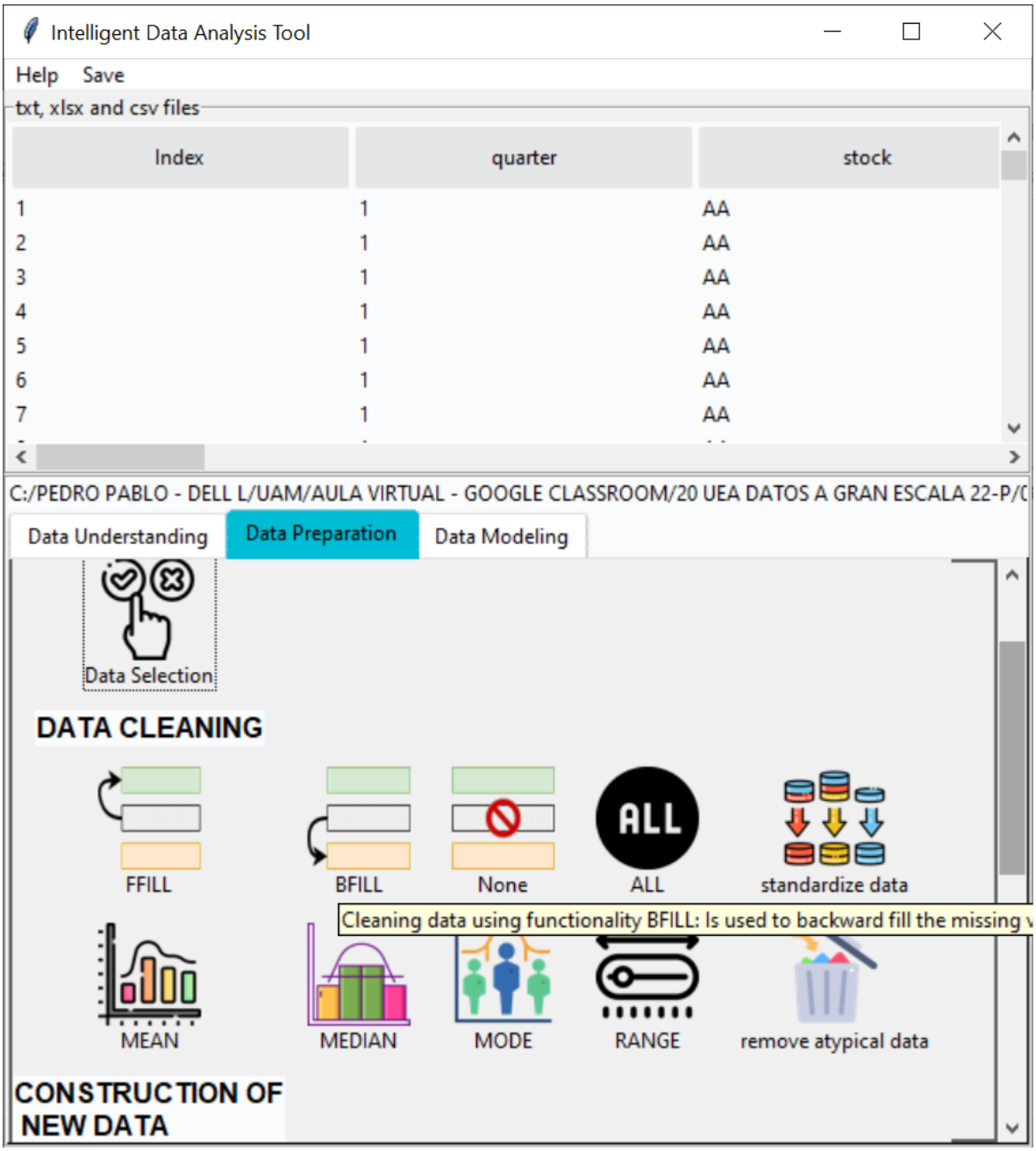


Figura 6.23. Los métodos de limpieza de datos que ofrece la herramienta IDA-WEB TOOL.

6.3.3.1. Selección de datos ("Data Selection")

La actividad "Data Selection" permite al usuario dos tipos de selecciones:

- Selección de campos
- Selección de registros

Tanto la selección de campos como la de registros producen como resultado un nuevo archivo que contendrá la selección efectuada. Las figuras de la 6.24 a la 6.29 muestran la actividad “Data Selection”, correspondiente al menú “Data Preparation”, sobre el conjunto de datos “Índice Dow Jones”.

En la figura 6.24 se puede apreciar el ícono que permite iniciar con la selección de datos. Como se muestra en la figura 6.25, una vez presionado el ícono “Data Selection” se debe elegir el tipo de selección que se desea ejecutar, ya sea de campos o de registros.

En tanto, las figuras 6.26 y 6.27 ilustran la tarea de selección de campos (*fields*). En la figura 6.26 se indican para selección los campos *quarter*, *stock*, *open*, *close* y *volume*. El resultado de esta acción se ilustra en la figura 6.27.

Por otra parte, las figuras 6.28 y 6.29 ilustran la tarea de selección de registros (*records*). En la figura 6.28 se indica la selección de los registros del 1 al 180 del *dataset*. El resultado de esta acción se puede apreciar en la figura 6.29.

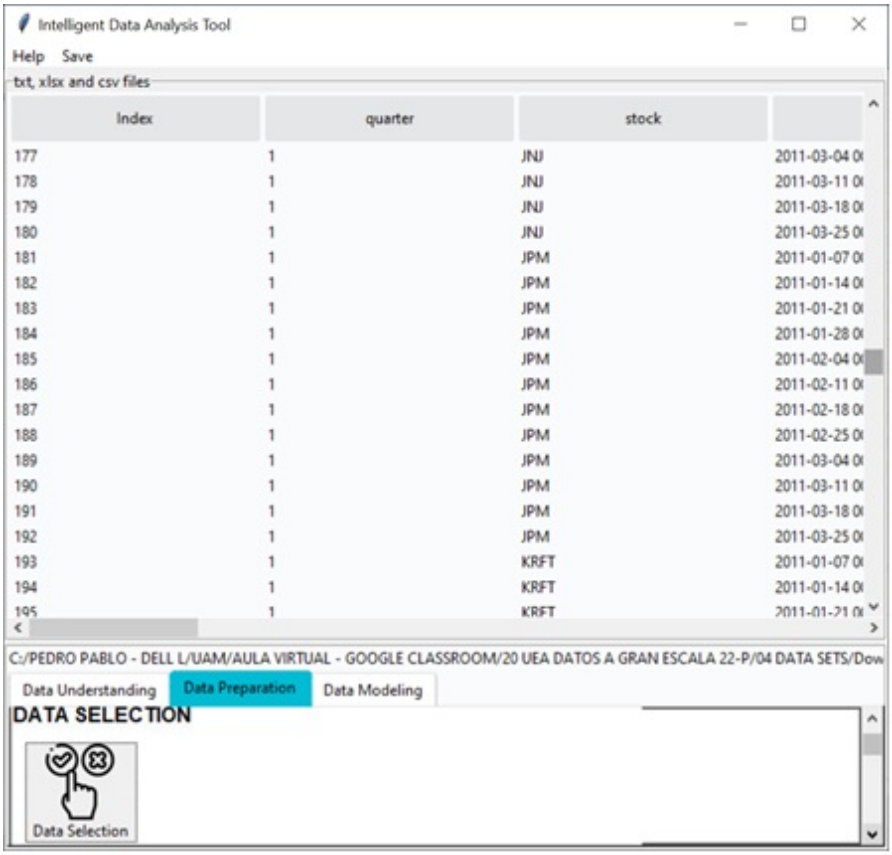


Figura 6.24. La opción “Data Selection” del menú “Data Preparation” de la herramienta IDA-WEB TOOL.

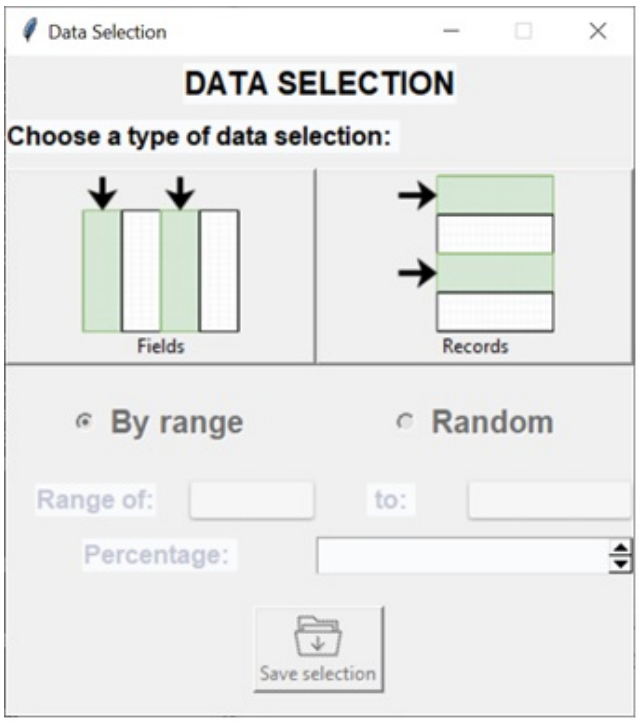


Figura 6.25. La selección de campos y la selección de registros como parte de la opción "Data Selection" en la herramienta IDA-WEB TOOL.

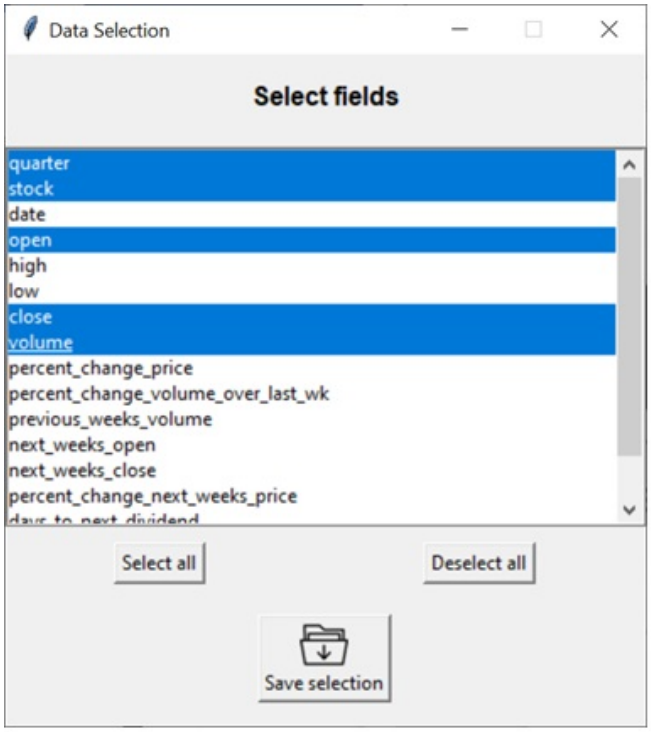


Figura 6.26. La selección de campos en el *dataset* del "Índice Dow Jones" en la herramienta IDA-WEB TOOL.

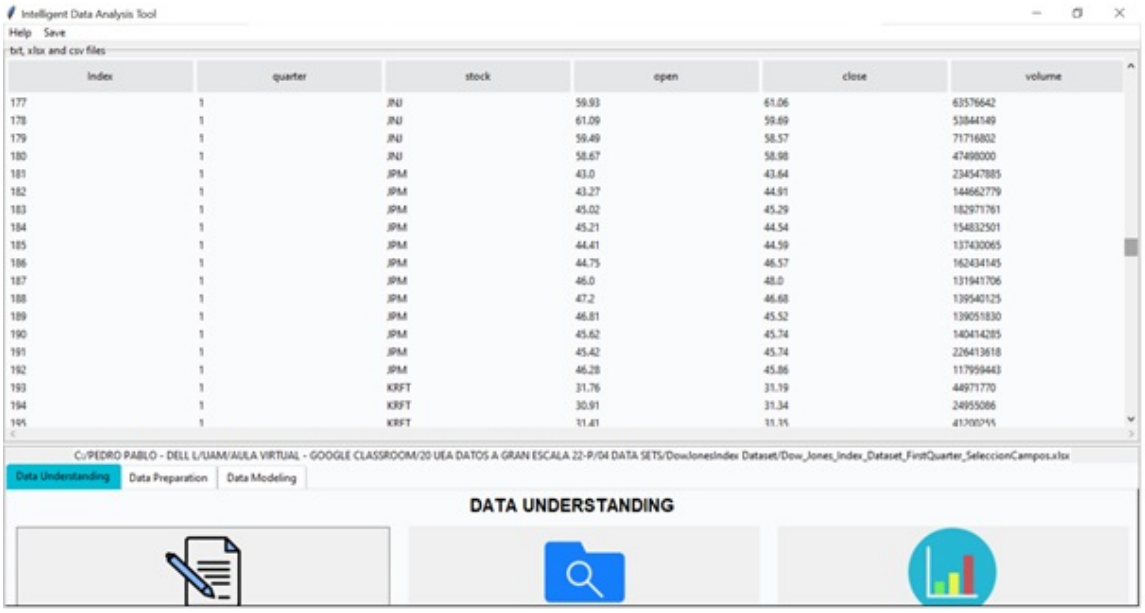


Figura 6.27. El *dataset* resultante de la selección de campos en el *dataset* original "Índice Dow Jones" en la herramienta IDA-WEB TOOL.

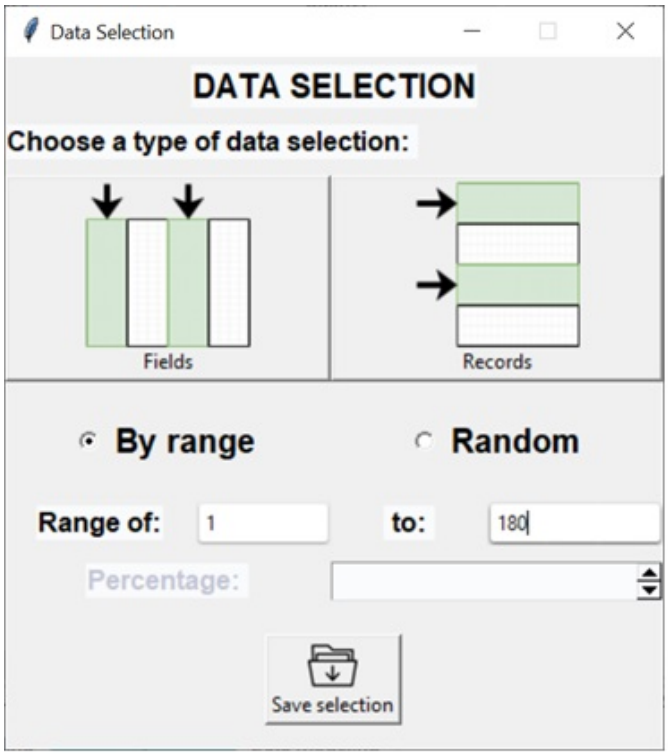


Figura 6.28. La selección de registros en el *dataset* "Índice Dow Jones" en la herramienta IDA-WEB TOOL.

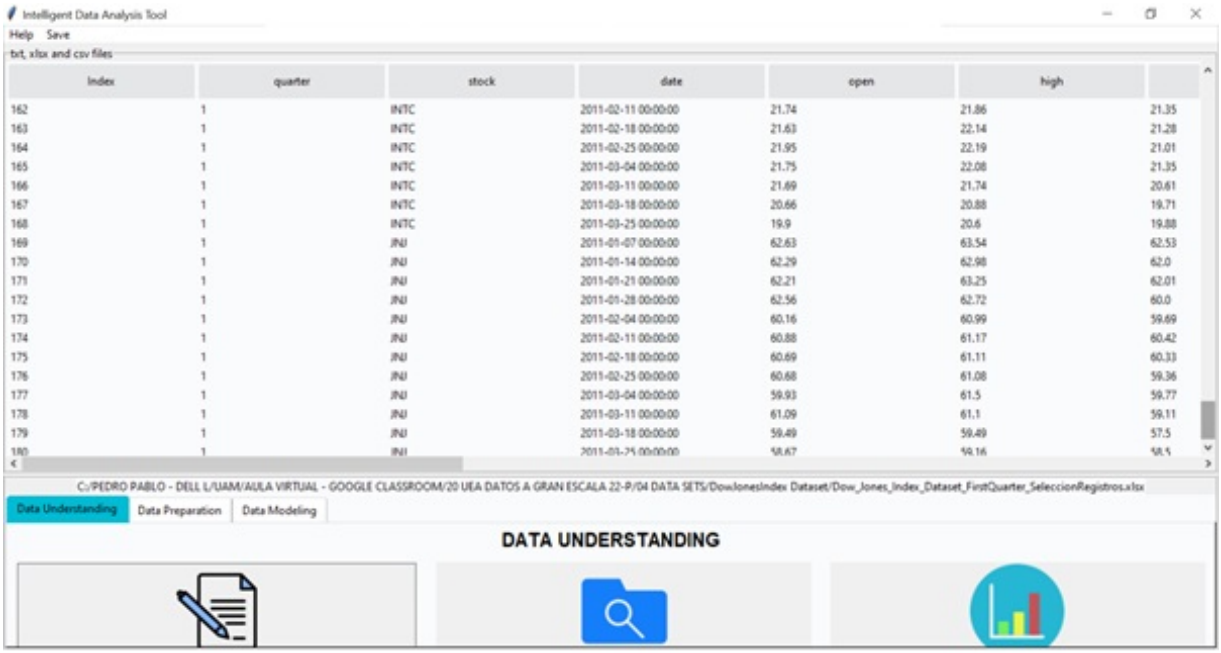


Figura 6.29. El *dataset* resultante de la selección de registros en el *dataset* original del "Índice Dow Jones" en la herramienta IDA-WEB TOOL.

6.3.3.2. Limpieza de datos ("Data Cleaning")

La actividad "Data Cleaning" permite al usuario efectuar la limpieza de datos, utilizando uno o más de los siguientes métodos:

- Limpieza de datos con el método Forward Fill (FFILL)
- Limpieza de datos con el método Backward Fill (BFILL)
- Limpieza de datos con el método NONE
- Limpieza de datos con el método ALL
- Limpieza de datos utilizando la media (MEAN)
- Limpieza de datos utilizando la moda (MODE)
- Limpieza de datos utilizando la mediana (MEDIAN)
- Limpieza de datos con el método RANGE
- Estandarización/normalización de datos
- Eliminar datos atípicos

Las figuras de la 6.30 a la 6.34 ilustran la actividad "Data Cleaning" en el conjunto de datos "Índice Dow Jones", utilizando los métodos FFILL y MEAN. La figura 6.30 identifica parte de los datos nulos o perdidos en este conjunto. Posteriormente, como se puede apreciar en la figura 6.31, se seleccionan todos los campos del "Índice Dow Jones" para proceder a la limpieza mediante el método FFILL. La figura 6.32 ilustra el próximo paso a ejecutar, el cual consiste en el envío de los campos

seleccionados en la figura 6.31 para realizar la limpieza. Finalmente, la figura 6.33 ilustra el resultado de emplear el método FFILL. Nótese que, como este método utilizó los datos del registro previo para proceder a la limpieza en el registro sucesivo, no es posible rellenar datos perdidos del primer registro del *dataset*.

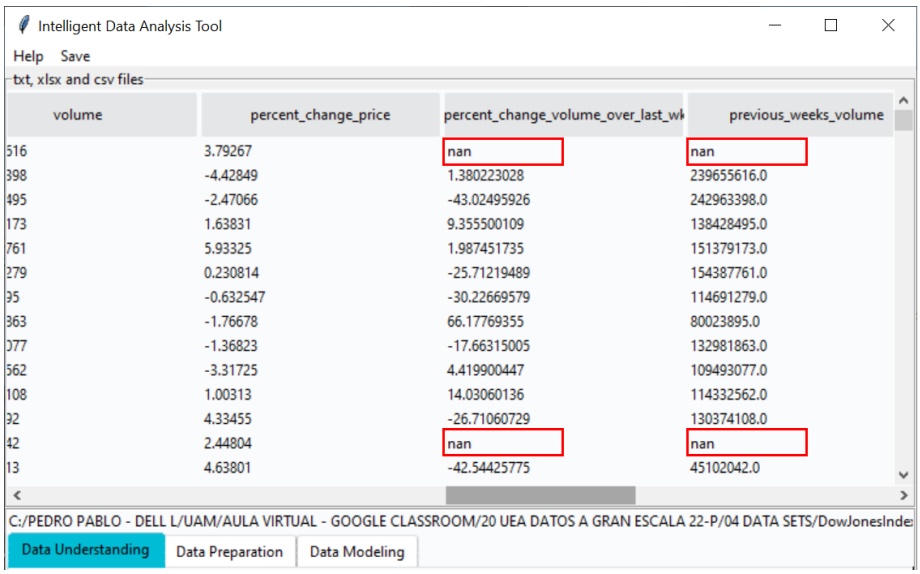


Figura 6.30. Identificación de datos nulos o perdidos en el conjunto de datos del "Índice Dow Jones" en la herramienta IDA-WEB TOOL.

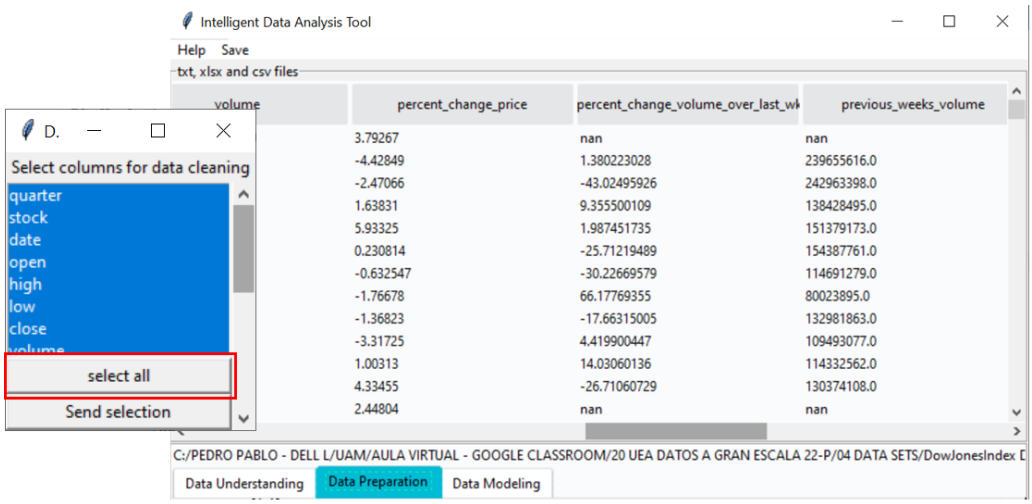


Figura 6.31. Selección de todos los campos en el conjunto de datos "Índice Dow Jones" para proceder a la limpieza utilizando el método FFILL en la herramienta IDA-WEB TOOL.

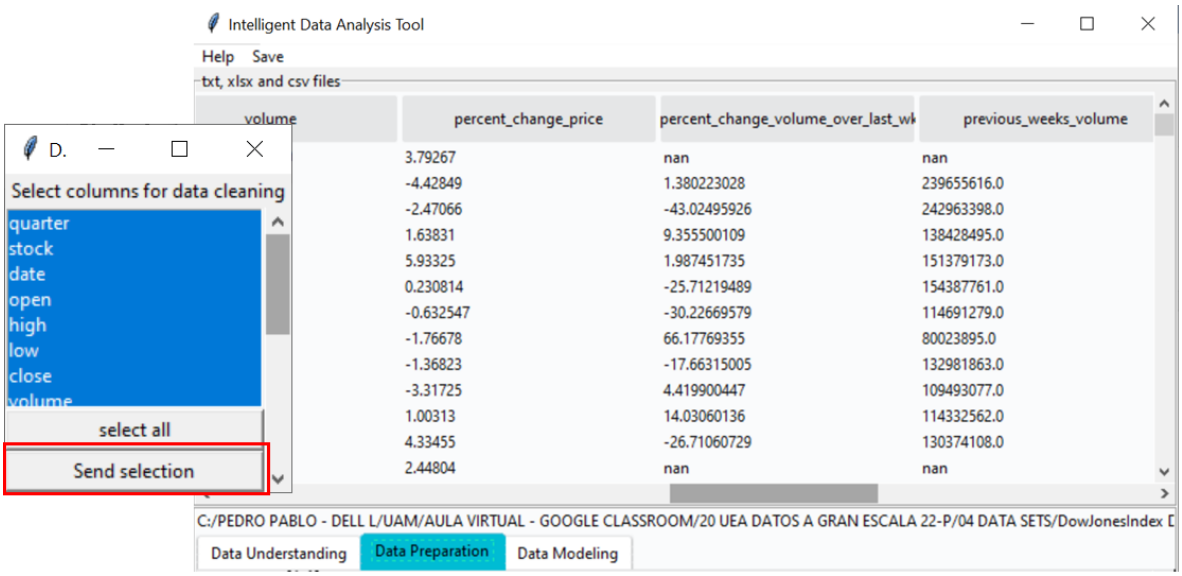


Figura 6.32. Envío de los campos seleccionados en el conjunto de datos "Índice Dow Jones" para proceder a la limpieza utilizando el método FFILL en la herramienta IDA-WEB TOOL.

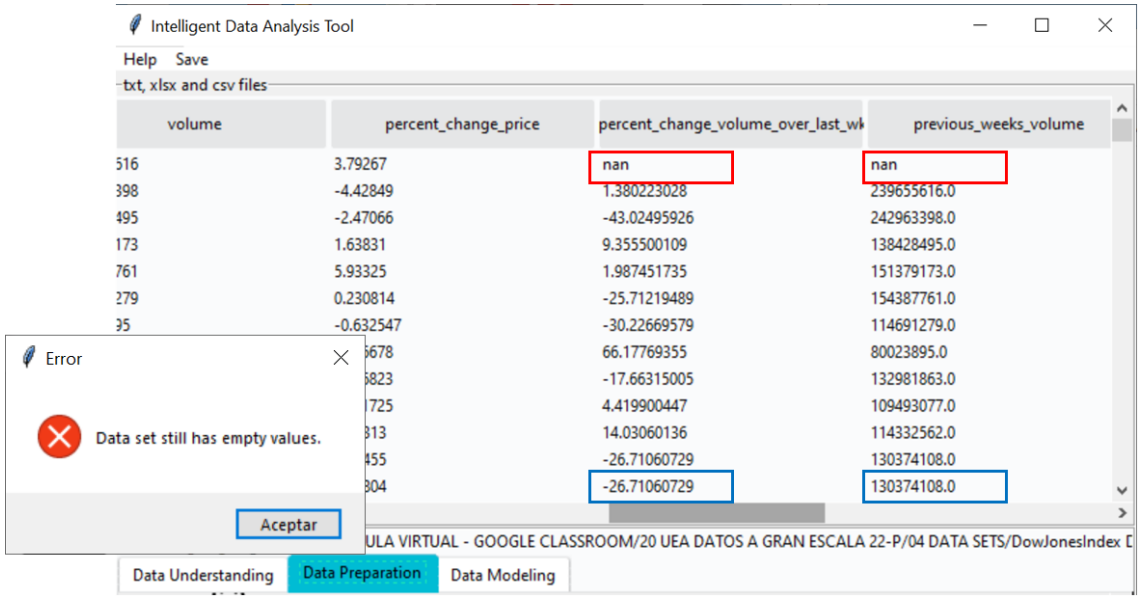


Figura 6.33. Conclusión de la limpieza de los datos del conjunto de datos "Índice Dow Jones" utilizando el método FFILL en la herramienta IDA-WEB TOOL.

Siguiendo la misma dinámica de las figuras 6.31 y 6.32, a continuación, se ilustra la limpieza de datos utilizando el método MEAN. La figura 6.34 muestra el resultado obtenido. Nótese que cuando se utiliza este segundo método se logran rellenar todos los datos perdidos del *dataset*, a diferencia de lo que ocurre con el método

FFILL.

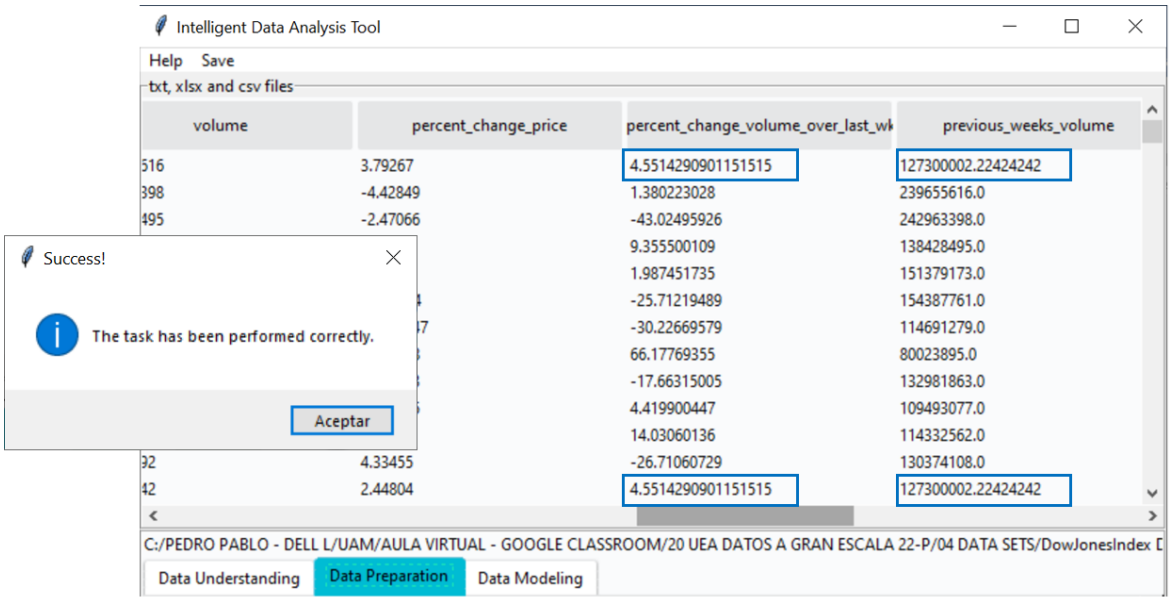


Figura 6.34. Conclusión de la limpieza de los datos del conjunto de datos "Índice Dow Jones" utilizando el método MEAN en la herramienta IDA-WEB TOOL.

6.3.3.3. Construcción o derivación de nuevos datos ("Construction of New Data")

La actividad "Construction of New Data" consiste en derivar nuevos campos a partir de operaciones que se realizan entre los campos ya existentes en el conjunto de datos a trabajar. Las figuras de la 6.35 a la 6.45 ilustran esta actividad sobre el conjunto de datos "Índice Dow Jones". Como se puede apreciar en la figura 6.35, dicha acción forma parte del menú "Data Preparation". En tanto, en la figura 6.36, se observa que al accionar el ícono "New Data" se desplegará la interfaz gráfica que permite la selección de los campos y operaciones requeridos para derivar el nuevo campo deseado.

Para ilustrar esta actividad en el conjunto de datos "Índice Dow Jones", se derivó el nuevo dato *Percent_High_Low*, el cual se desea incorporar como un nuevo campo. Los pasos a seguir son ilustrados en las figuras de la 6.37 a la 6.45.

Cabe mencionar que la expresión simple para el cálculo del campo derivado es la siguiente:

$$Percent\_High\_Low = \frac{(high - low)}{low} * 100$$

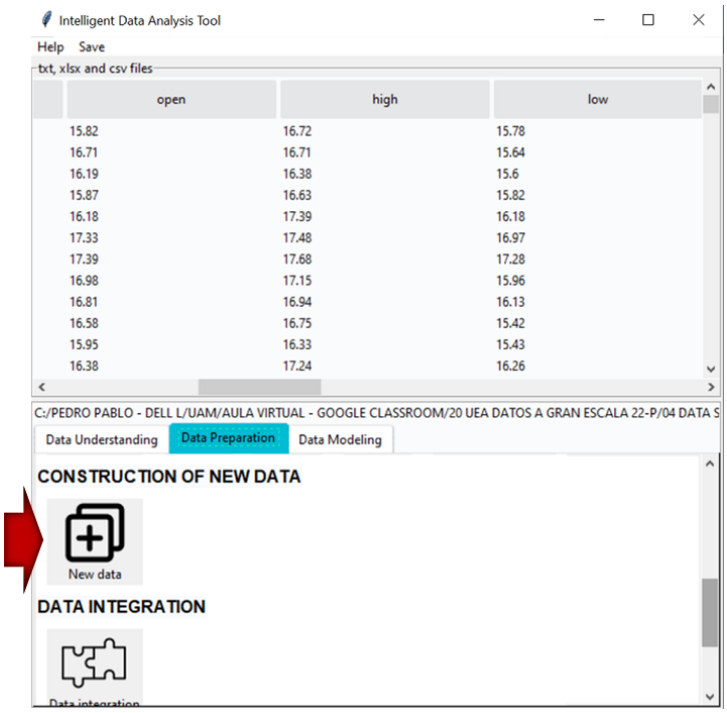


Figura 6.35. La actividad “Construction of New Data” en el menú “Data Preparation” de la herramienta IDA-WEB TOOL.

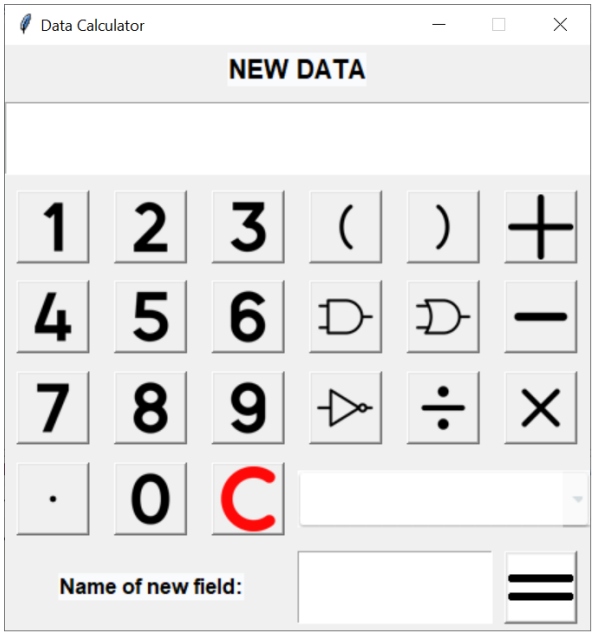


Figura 6.36. La interfaz gráfica para la construcción de nuevos datos, a través de la cual es posible la selección de los campos y operaciones requeridos para derivar el nuevo campo deseado en la herramienta IDA-WEB TOOL.

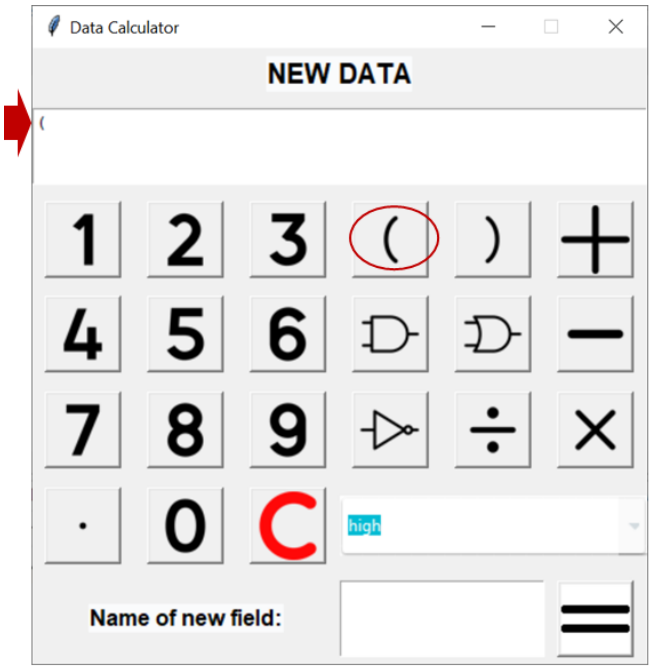


Figura 6.37. Derivación del campo *Percent_High_Low* a través de la actividad "Construction of New Data" en la herramienta IDA-WEB TOOL (paso 1).

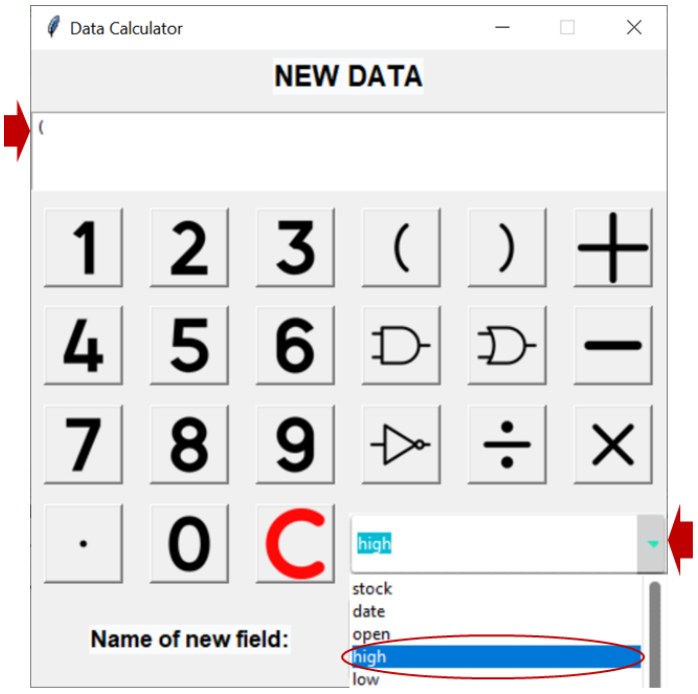


Figura 6.38. Derivación del campo *Percent_High_Low* a través de la actividad "Construction of New Data" en la herramienta IDA-WEB TOOL (paso 2).

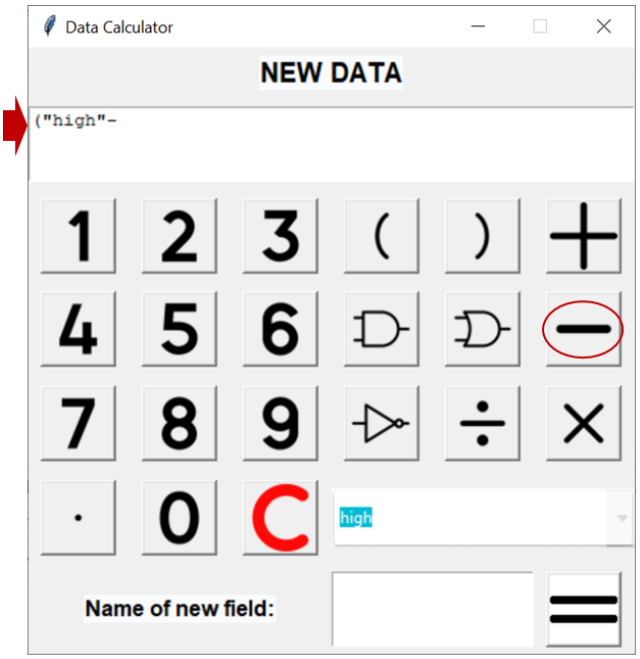


Figura 6.39. Derivación del campo *Percent_High_Low* a través de la actividad "Construction of New Data" en la herramienta IDA-WEB TOOL (paso 3).

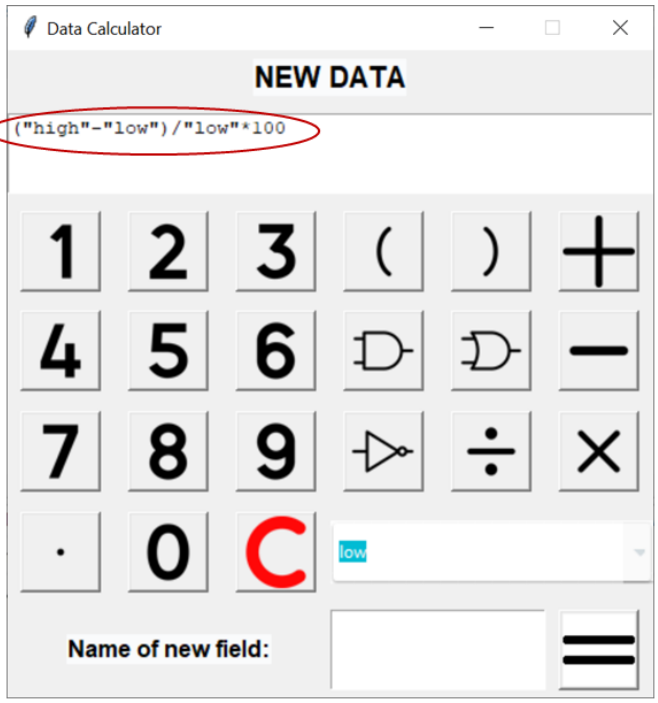


Figura 6.40. Derivación del campo *Percent_High_Low* a través de la actividad "Construction of New Data" en la herramienta IDA-WEB TOOL (pasos restantes).

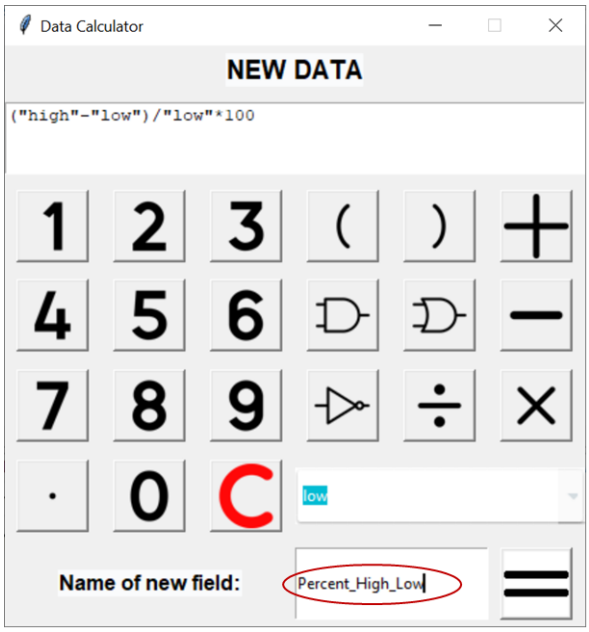


Figura 6.41. Se introduce el nombre del campo derivado *Percent_High_Low* en la herramienta IDA-WEB TOOL.

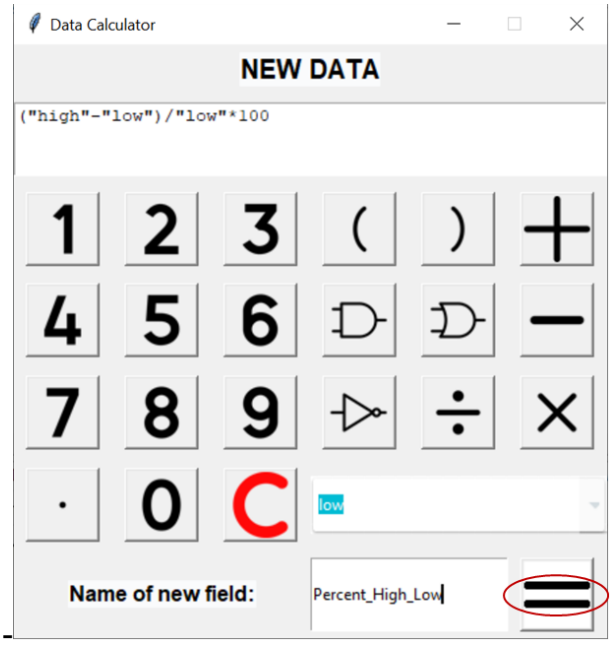


Figura 6.42. Al presionar el símbolo “=”, en la herramienta IDA-WEB TOOL, se despliega la ventana emergente para proceder a salvar el nuevo archivo que contendrá el campo derivado *Percent_High_Low*.

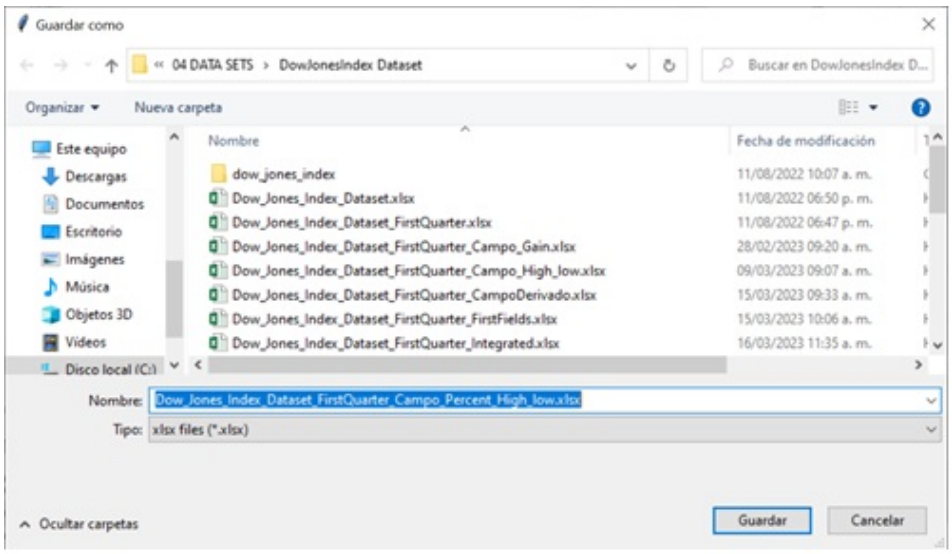


Figura 6.43. Selección del directorio y nombre del archivo que contendrá el nuevo campo derivado *Percent_High_Low* en la herramienta IDA-WEB TOOL.

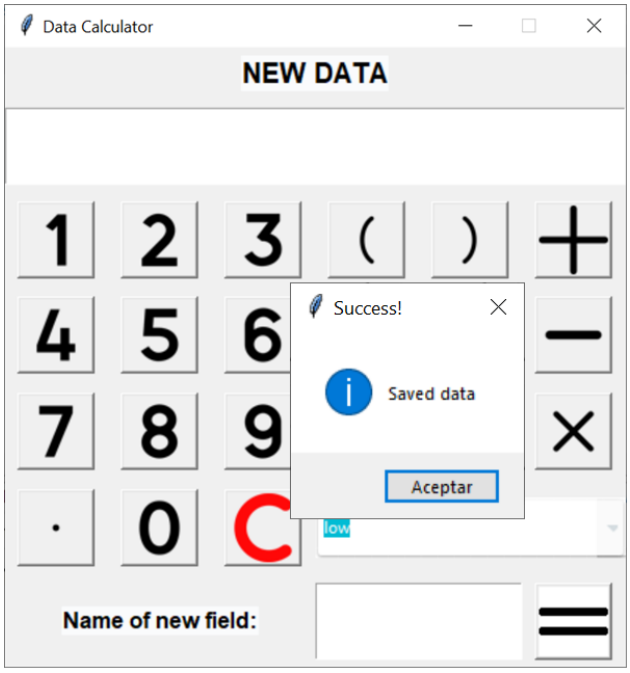


Figura 6.44. Notificación del almacenamiento correcto del archivo con el nuevo campo derivado *Percent_High_Low* en la herramienta IDA-WEB TOOL.

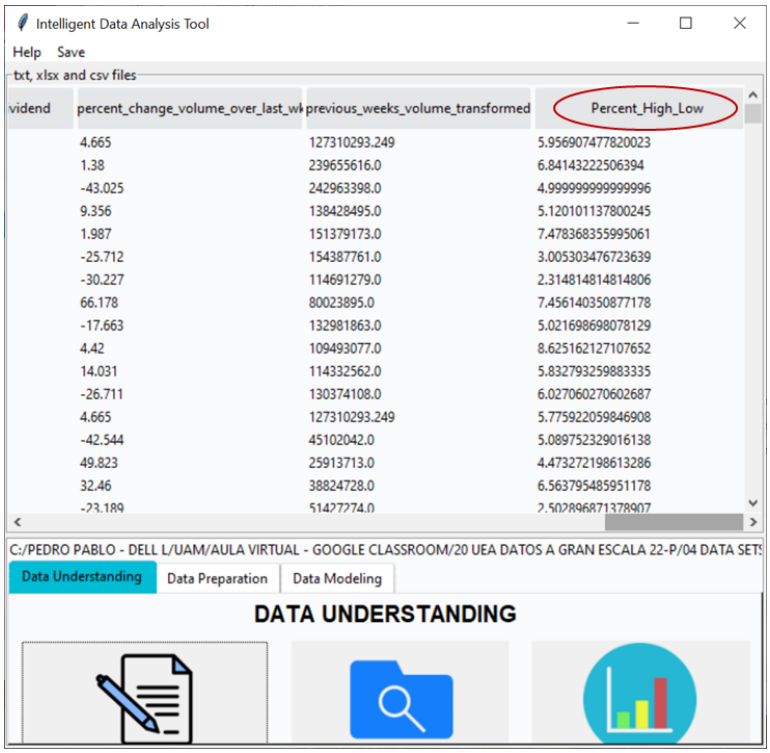


Figura 6.45. Visualización del archivo resultante con el nuevo campo derivado *Percent_High_Low* en la herramienta IDA-WEB TOOL.

6.3.3.4. Integración de datos (“Data Integration”)

La integración de datos (“Data Integration”) permite al usuario fusionar dos archivos compatibles atendiendo a alguno de los siguientes criterios:

- Integración de campos
- Integración de registros

La compatibilidad de los archivos a fusionar se refiere a que ambos presenten el mismo tipo de campos o registros a integrar.

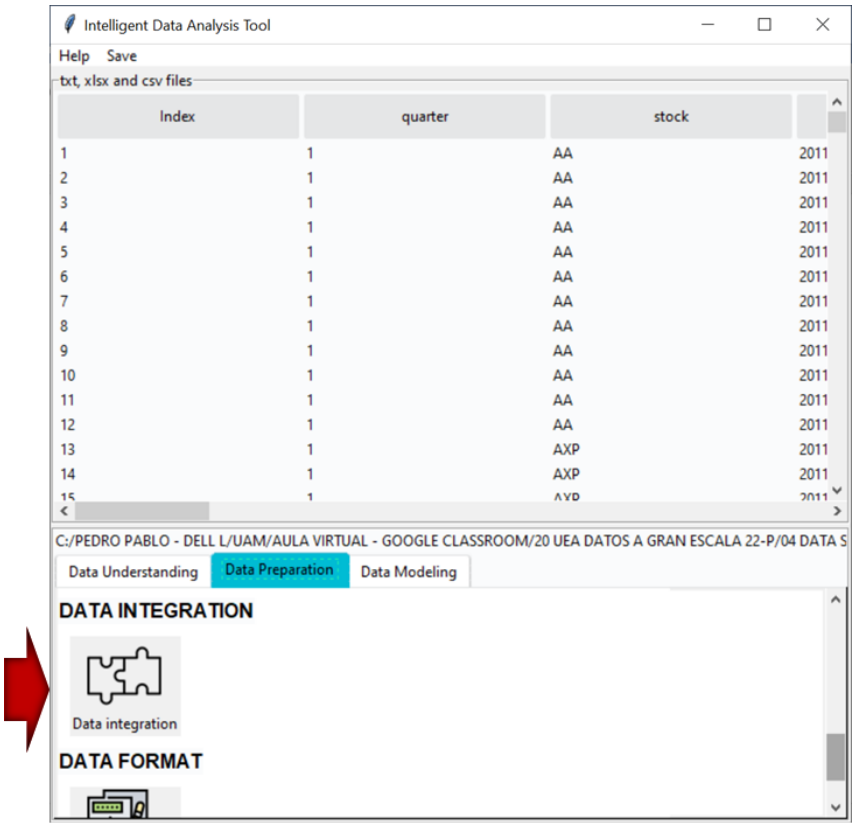


Figura 6.46. La actividad “Data Integration” en el menú “Data Preparation” de la herramienta IDA-WEB TOOL.

Como se puede apreciar en la figura 6.46, la actividad “Data Integration” forma parte del menú “Data Preparation”. Las figuras de la 6.46 a la 6.52 ilustran esta actividad sobre dos conjuntos de datos compatibles del “Índice Dow Jones”. Para ello, cada uno de los conjuntos mencionados representa un grupo diferente al conjunto global de datos del “Índice Dow Jones” y poseen el mismo número de registros. Los dos conjuntos cuyos campos se desean integrar son:

Archivo número 1: Dow_Jones_Index_Dataset_FirstQuarter_FirstFields.xlsx
Archivo número 2: Dow_Jones_Index_Dataset_FirstQuarter_SecondFields.xlsx

La figura 6.47 muestra la carga del archivo número 1, el cual contiene el primer grupo de campos del conjunto global de datos del “Índice Dow Jones” (desde el campo *quarter* hasta el campo *volume*). Como se ilustra en la figura 6.48, al seleccionar el ícono “Data Integration” (ver figura 6.46) se desplegará la interfaz gráfica que permitirá seleccionar el tipo de integración requerida: de campos o de registros. Las figuras de la 6.49 a la 6.52 ilustran los pasos restantes de esta actividad.

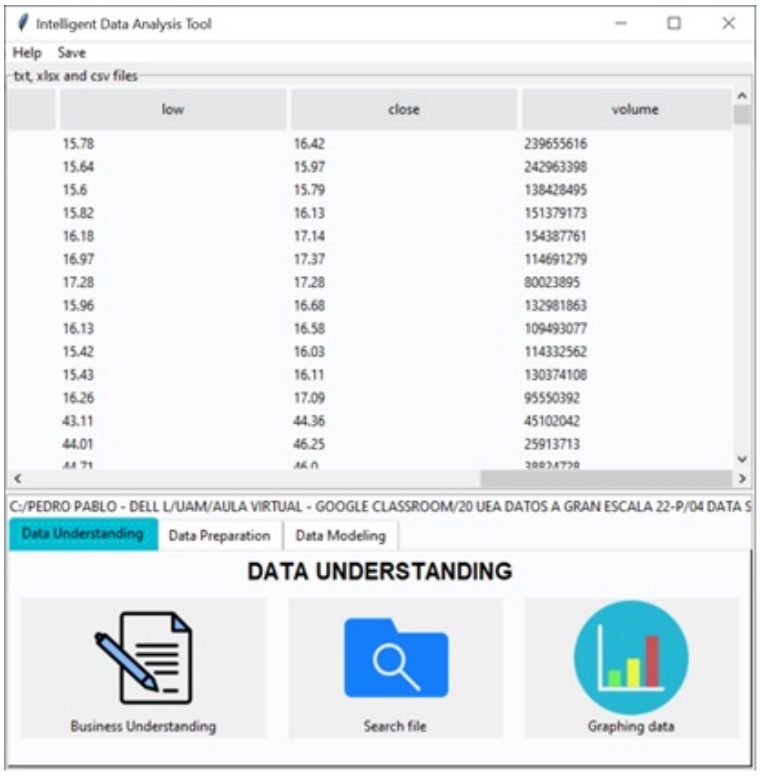


Figura 6.47. Carga del archivo número 1 a fusionar en la actividad fusión de campos en la herramienta IDA-WEB TOOL.

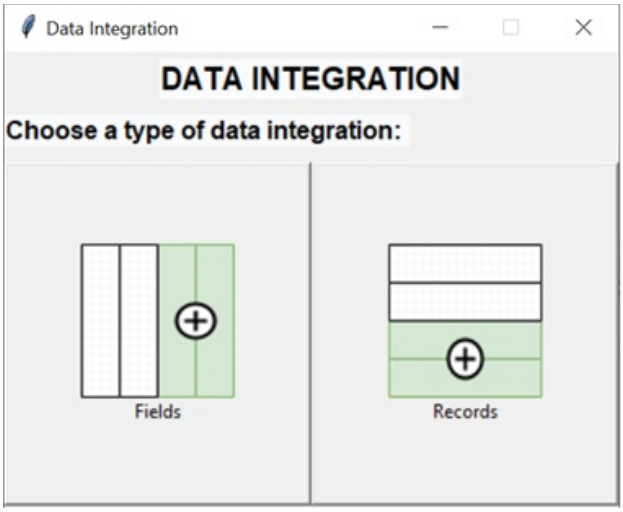


Figura 6.48. Interfaz gráfica de la herramienta IDA-WEB TOOL que permite seleccionar el tipo de integración de datos deseada: de campos o de registros.

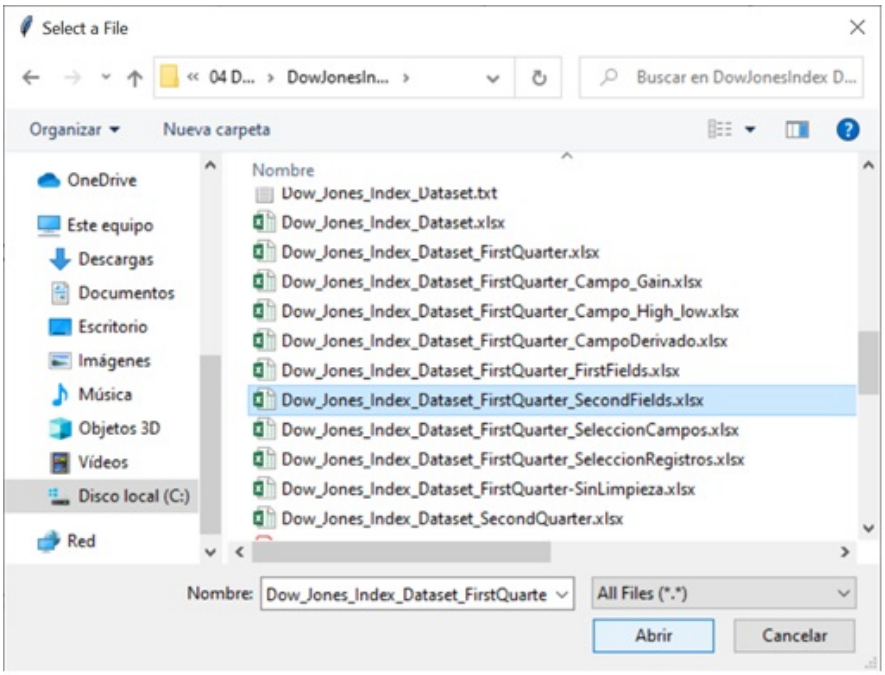


Figura 6.49. Selección del segundo conjunto de datos (Dow_Jones_Index_Dataset_FirstQuarter_SecondFields.xlsx), compatible con el primero, para proceder a su integración en la herramienta IDA-WEB TOOL.

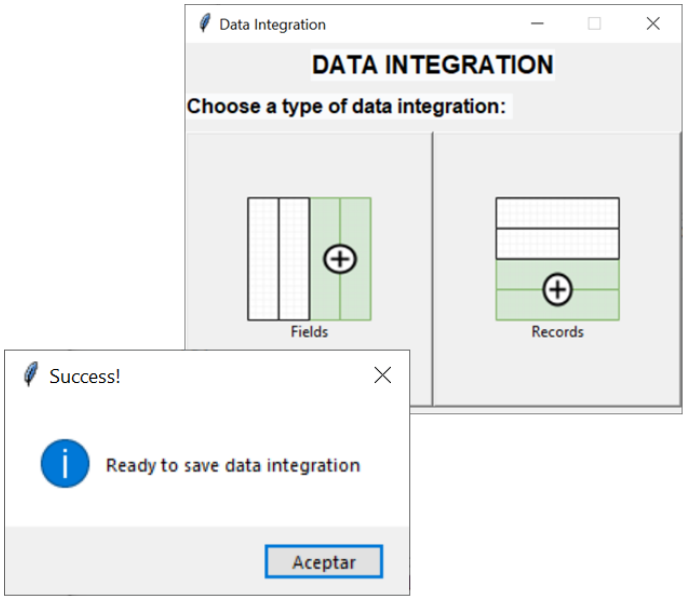


Figura 6.50. La selección del segundo conjunto de datos (Dow_Jones_Index_Dataset_FirstQuarter_SecondFields.xlsx) en la herramienta IDA-WEB TOOL se efectuó de forma correcta y se puede proceder a la integración de ambos conjuntos.

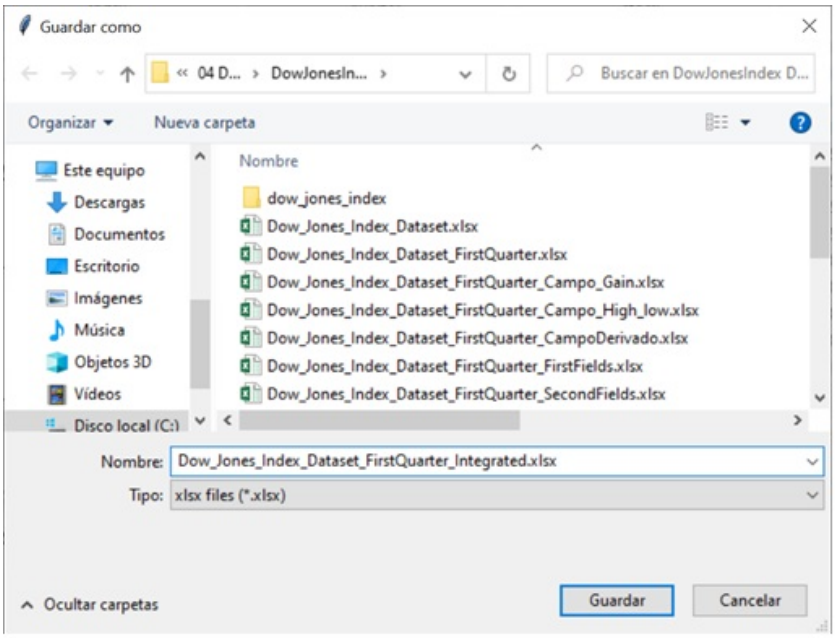


Figura 6.51. Elección de la ubicación y el nombre del nuevo archivo que contendrá la integración de datos en la herramienta IDA-WEB TOOL.

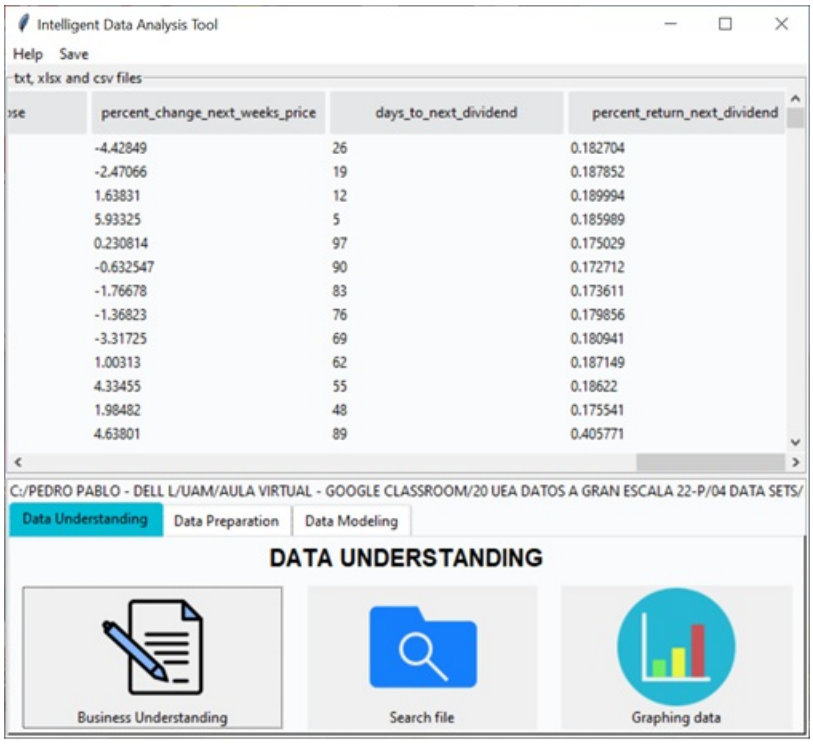


Figura 6.52. Archivo resultante de la integración de datos en la herramienta IDA-WEB TOOL, el cual contiene los campos del primer y el segundo archivo.

6.3.3.5. Formato de datos (“Data Format”)

La actividad de formato de datos (“Data Format”) permite al usuario visualizar y modificar el tipo de dato de un campo seleccionado. Esta actividad es de gran importancia en la preparación de los datos, ya que, de acuerdo con el tipo de modelado a aplicar, tanto los datos predictores como los predictivos requerirán de un formato de datos particular. Por ejemplo, si la tarea a resolver a través de la fase de modelado es del tipo clasificación, entonces el campo o campos objetivos (predictivos) deberán poseer un formato de tipo categórico, nominal u ordinal. Por otra parte, si la tarea a resolver es del tipo predicción, entonces el campo objetivo (predictivo) deberá ser de tipo numérico.

Las figuras de la 6.53 a la 6.57 ilustran el formato de datos sobre el conjunto de datos del “Índice Dow Jones”. Como se puede apreciar en la figura 6.53, esta actividad es parte del menú “Data Preparation”; al desplegarlo y seleccionar el ícono “Data Format” aparecerá la interfaz gráfica que se muestra en la figura 6.54, a través de la cual se puede visualizar el tipo de dato de cada uno de los campos, así como cambiar el tipo de dato de los campos requeridos. Las figuras de la 6.55 a la 6.57 ilustran cómo puede ser utilizada la función “Data Format”.

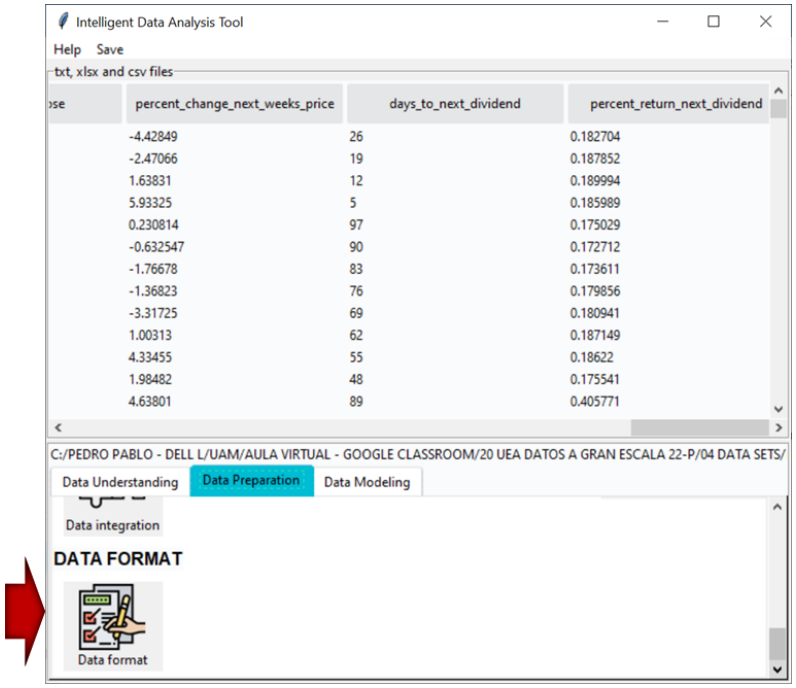


Figura 6.53. La actividad “Data Format” en el menú “Data Preparation” de la herramienta IDA-WEB TOOL.

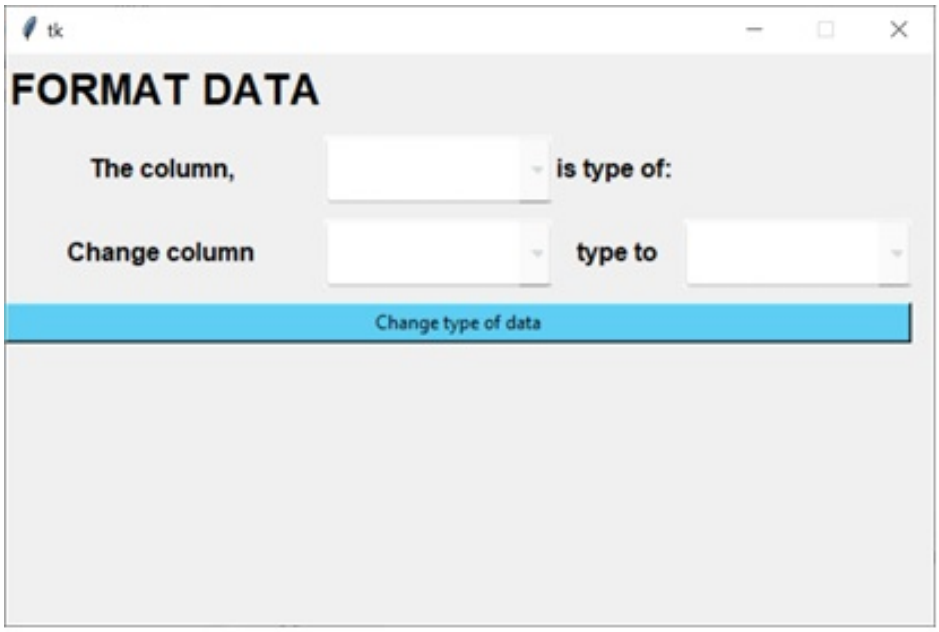


Figura 6.54. Interfaz gráfica de la actividad “Data Format” en el menú “Data Preparation” de la herramienta IDA-WEB TOOL.

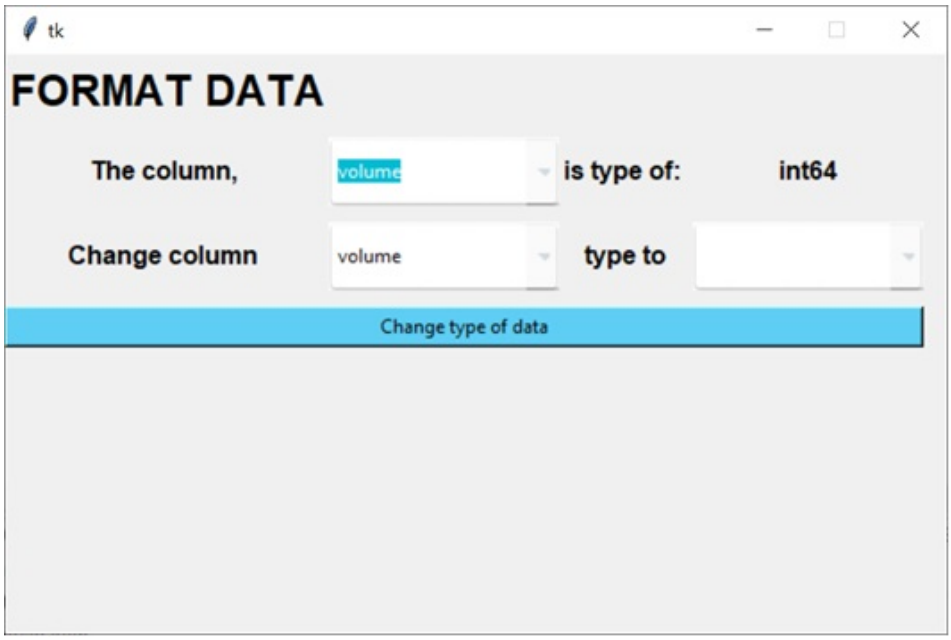


Figura 6.55. Consulta del formato del campo *volume* en la interfaz gráfica de la actividad “Data Format” de la herramienta IDA-WEB TOOL.

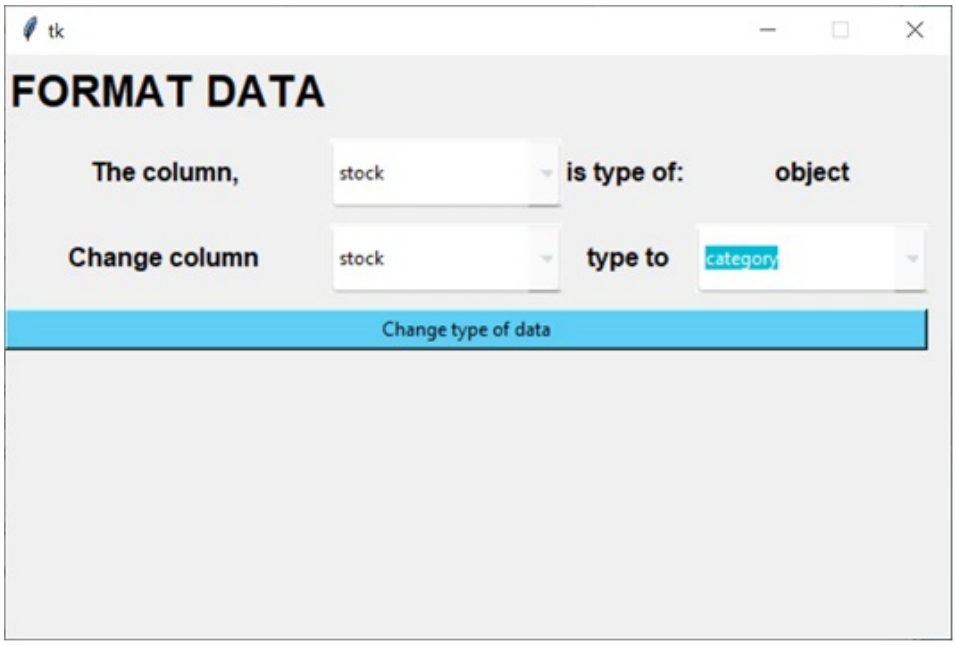


Figura 6.56. Cambio de formato del campo *volume* de tipo *object* a tipo *category* en la interfaz gráfica de la actividad “Data Format” de la herramienta IDA-WEB TOOL.

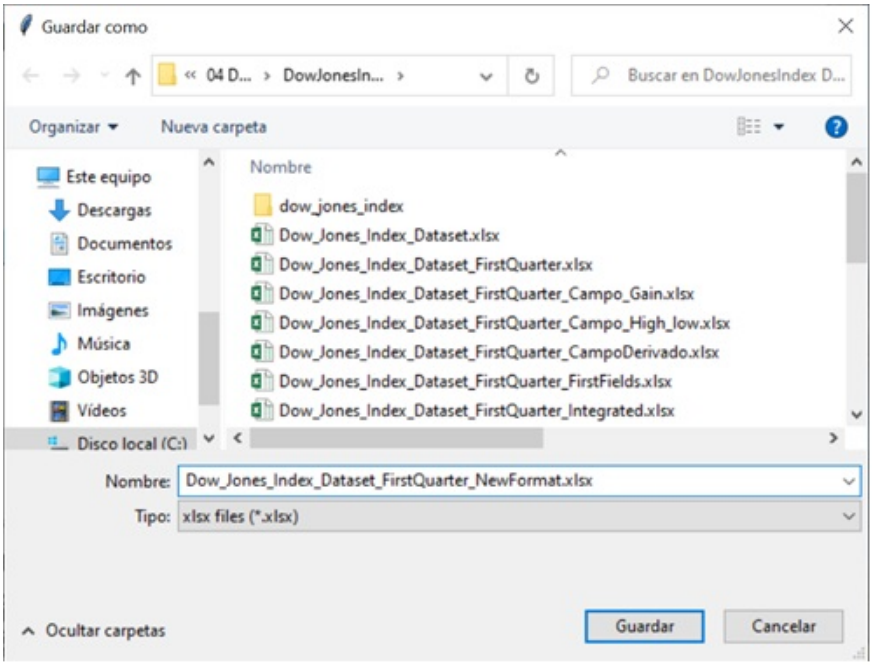


Figura 6.57. Almacenamiento del nuevo archivo con el cambio de formato del campo *volume* de tipo *object* a tipo *category*.

6.3.4. Modelado

La herramienta IDA-WEB TOOL está enfocada a dos tipos de tareas fundamentales

de la minería de datos: clasificación y predicción. Para soportar estas tareas, pone a disposición del usuario una extensa gama de modelos de redes neuronales artificiales, árboles de decisión, algoritmos de clasificación supervisada y algoritmos de regresión. De forma particular, los modelos de *machine learning* y algoritmos estadísticos que integra IDA-WEB TOOL son los siguientes:

- Clasificación:
  - Redes neuronales artificiales supervisadas, tales como Multi-Layer Perceptron (MLP) y Support Vector Machine (SVM)
  - Árboles de decisión
  - Algoritmo de los K vecinos más cercanos (K-NN)
  - Algoritmo de clasificación probabilística (Gaussian Naive Bayes)
  - Algoritmos de regresión logística

- Predicción:
  - Redes neuronales artificiales supervisadas, tales como, Multi-Layer Perceptron (MLP) y Support Vector Machine (SVM)
  - Árboles de decisión
  - Algoritmo de los K vecinos más cercanos (K-NN)
  - Algoritmos de regresión
  - Regresión lineal bayesiana

La figura 6.58 muestra el despliegue del menú “Data Modeling” de la herramienta IDA-WEB TOOL, donde se aprecian todos los modelos disponibles para realizar las tareas de clasificación y regresión para esta importante fase de la minería de datos.

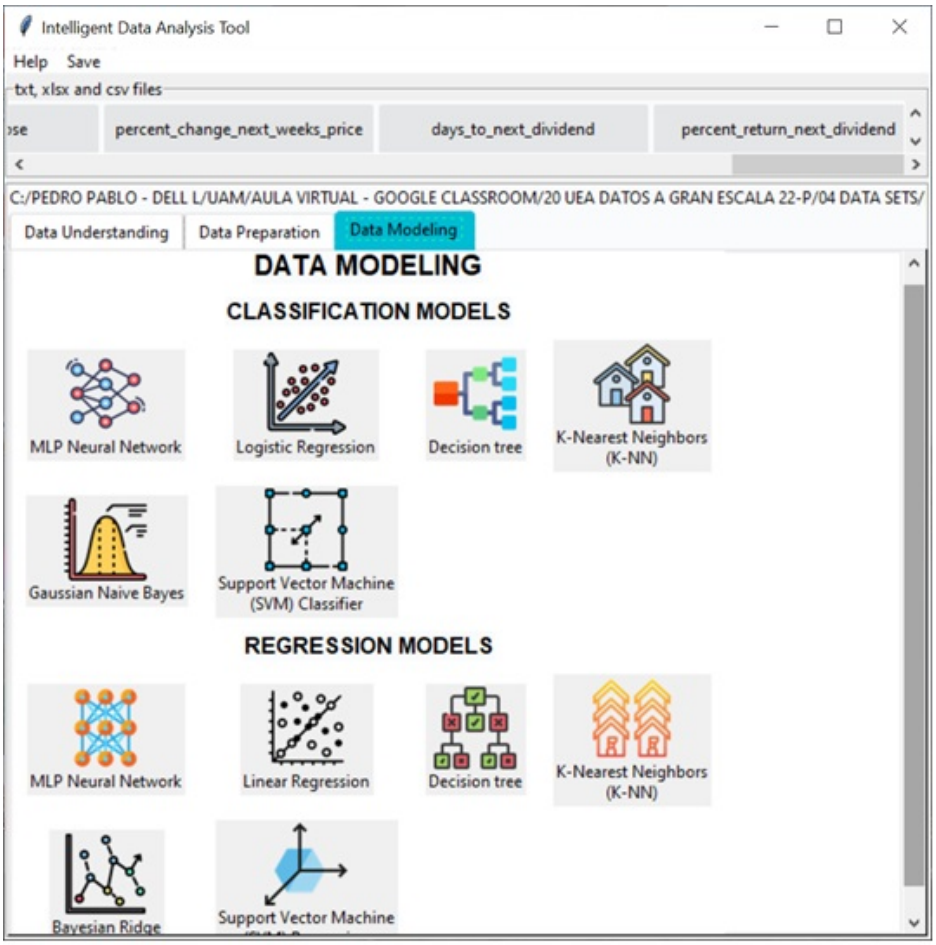


Figura 6.58. El menú “Data Modeling” de la herramienta IDA-WEB TOOL.

6.3.4.1. Construcción y evaluación de un modelo de clasificación en IDA-WEB TOOL

Las figuras 6.59 a la 6.68 ilustran la construcción y evaluación de un modelo de clasificación en IDA-WEB TOOL. Para ello, se utilizó el conjunto de datos del “Índice Dow Jones”, considerando sólo aquellos campos que fungen como buenos predictores y el total de registros.

El campo a clasificar es *stock,* mientras que los campos predictores son:

- *open*
- *high*
- *low*
- *close*
- *volumen*
- *next_weeks_open*
- *next_weeks_close*
- *percent_return_next_dividend*

En las figuras 6.59 y 6.60 se muestra el proceso de carga del archivo “Índice Dow Jones” para la construcción del modelo de clasificación. En tanto, como se puede apreciar en la figura 6.61, una vez cargado el archivo con los datos, el primer paso consiste en la selección del modelo de clasificación a utilizar, que en este caso se trata de una red neuronal perceptrón multi-estrato (MLP, por sus siglas en inglés).

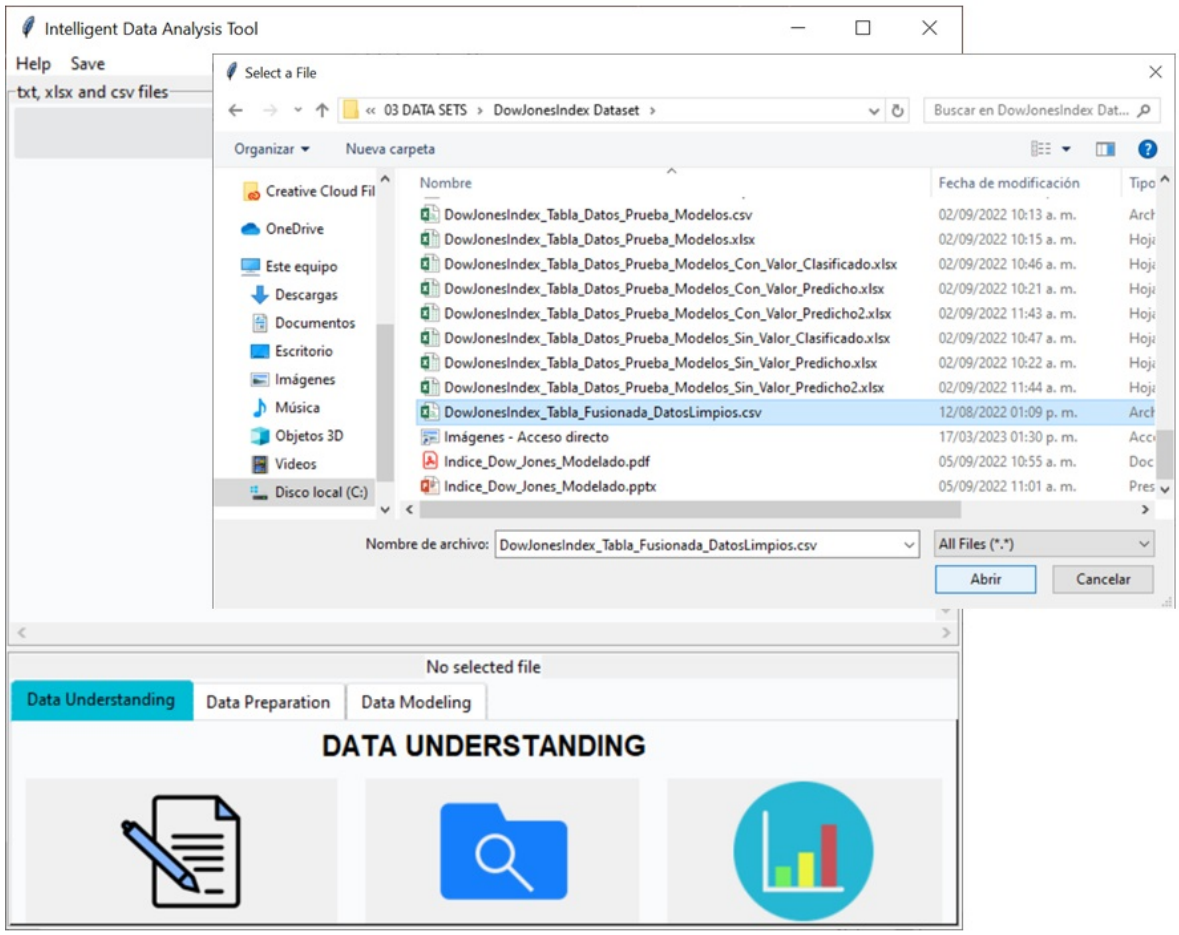


Figura 6.59. Paso 1 de 2 de la carga del archivo “Índice Dow Jones” para la tarea de clasificación.

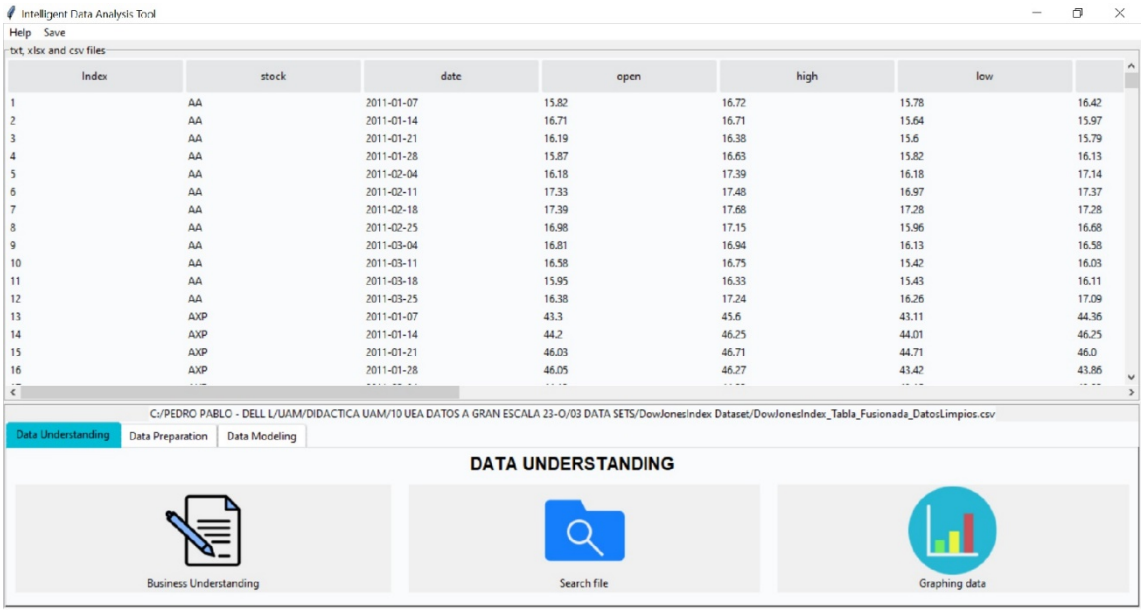


Figura 6.60. Paso 2 de 2 de la carga del archivo “Índice Dow Jones” para la tarea de clasificación.

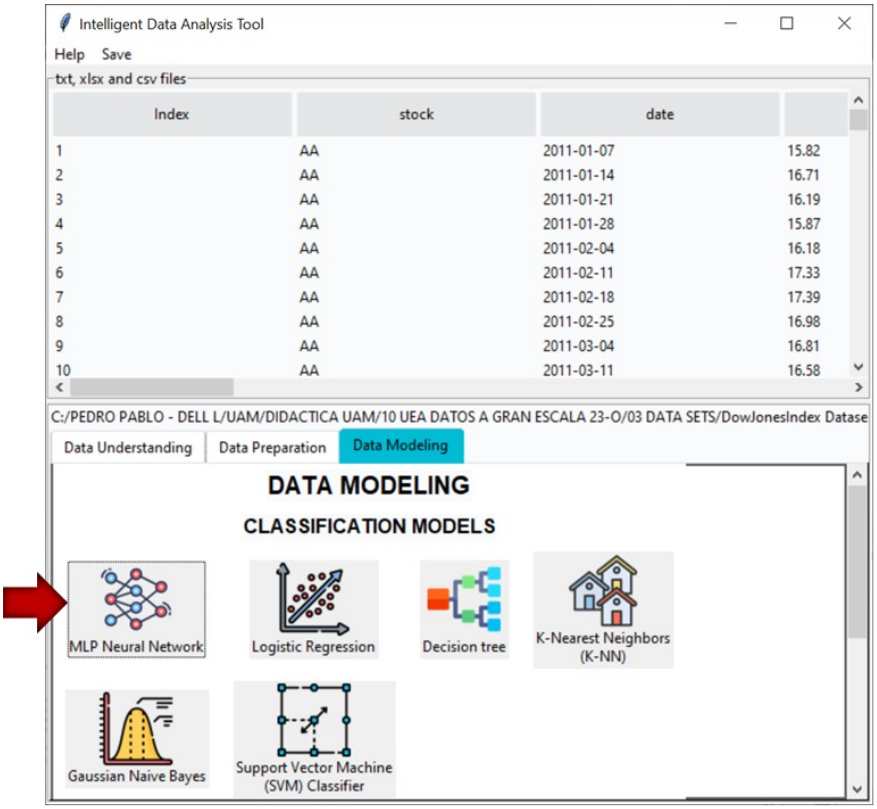


Figura 6.61. Selección del modelo MLP Neural Network, para la tarea de clasificación, desde el menú “Data Modeling” de la herramienta IDA-WEB TOOL.

Una vez elegido el modelo para la tarea de clasificación, se desplegará en pantalla una nueva interfaz gráfica para seleccionar los campos predictores (entradas al modelo MLP) y el campo de la clase o categoría (salida que debe producir el modelo MLP) (ver figura 6.62). Los campos predictores y objetivo (predictivo) seleccionados se ilustran en la figura 6.63. Posteriormente, se desplegará en pantalla la interfaz que permitirá seleccionar los íconos para iniciar la construcción del modelo de predicción. Antes de esto, es necesario cargar el conjunto adicional de datos que contiene los datos de prueba, asegurándose de que no estén contenidos en el conjunto inicial (ver figura 6.64). Una vez que el archivo con los datos de prueba se haya cargado, es posible proceder a la generación del modelo, seleccionando el ícono “Make prediction”, tal como se ilustra en la figura 6.65.

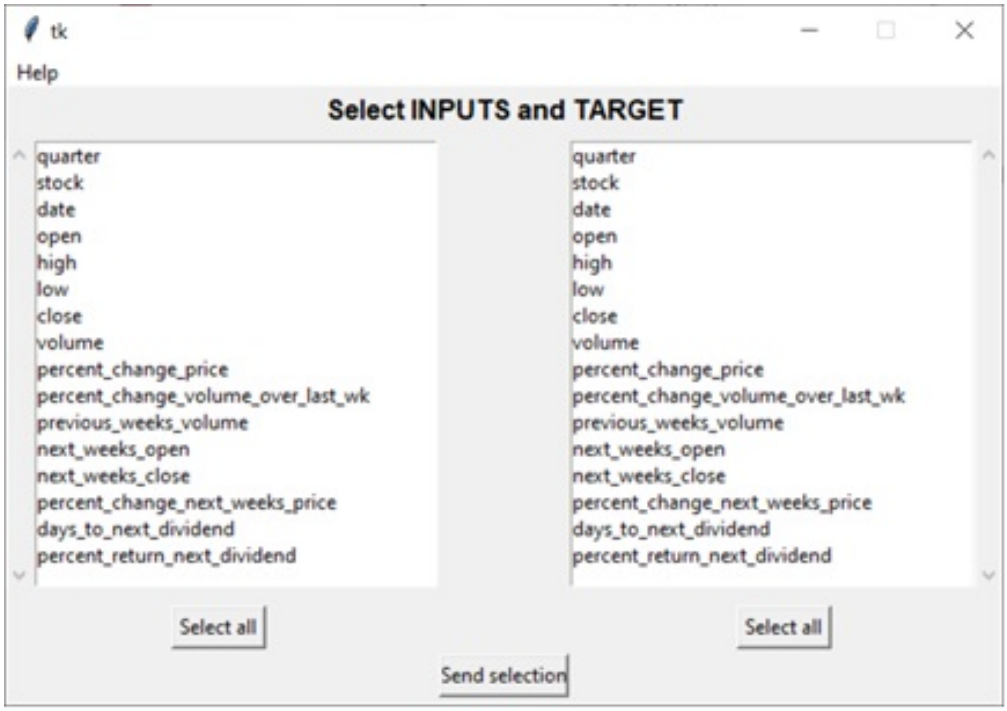


Figura 6.62. Interfaz gráfica para la selección de los campos predictores y el campo predictivo.

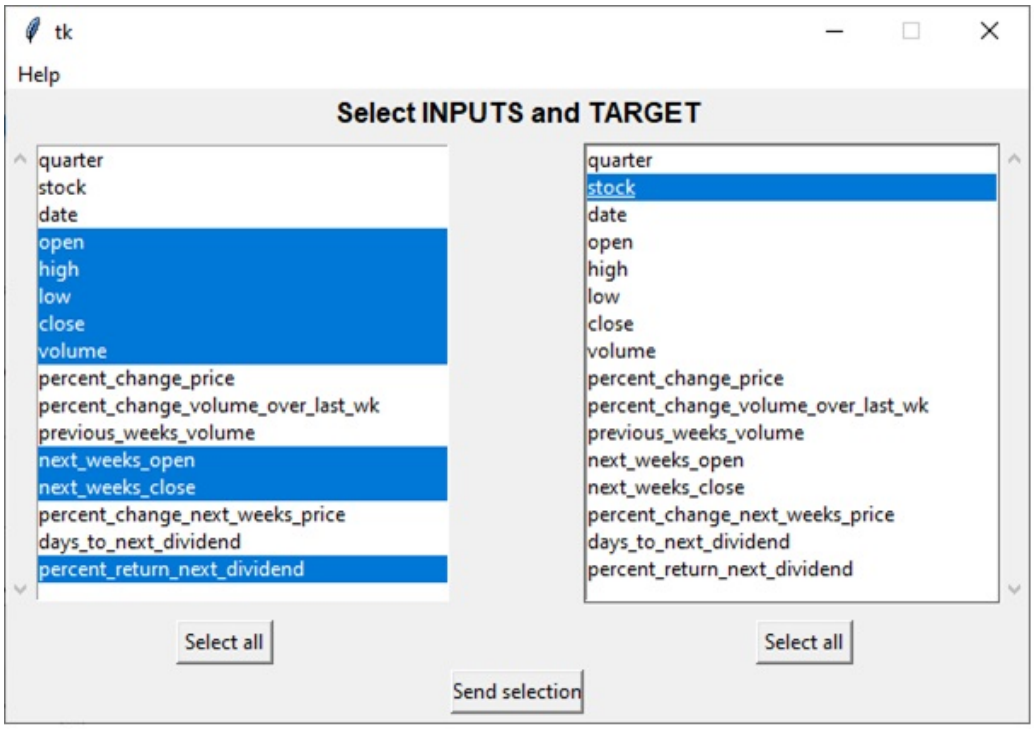


Figura 6.63. Selección de los campos predictores (entradas al modelo MLP) y el campo de la clase (salida que debe producir el modelo MLP).

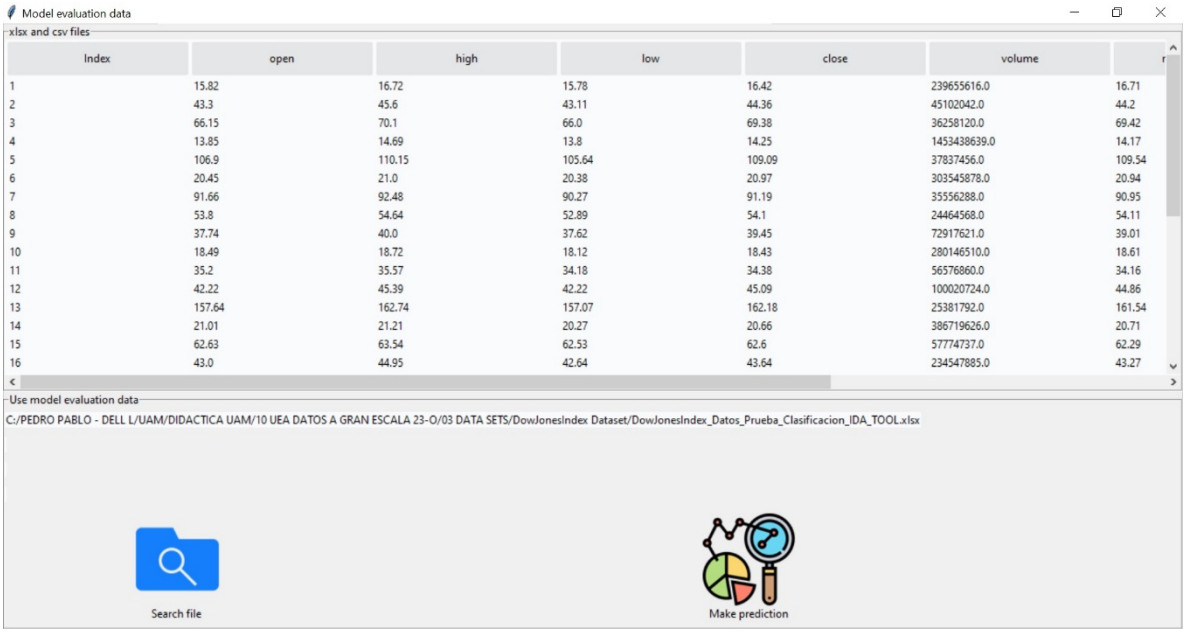


Figura 6.64. Carga del archivo con los datos de prueba, paso previo a la generación del modelo de predicción.

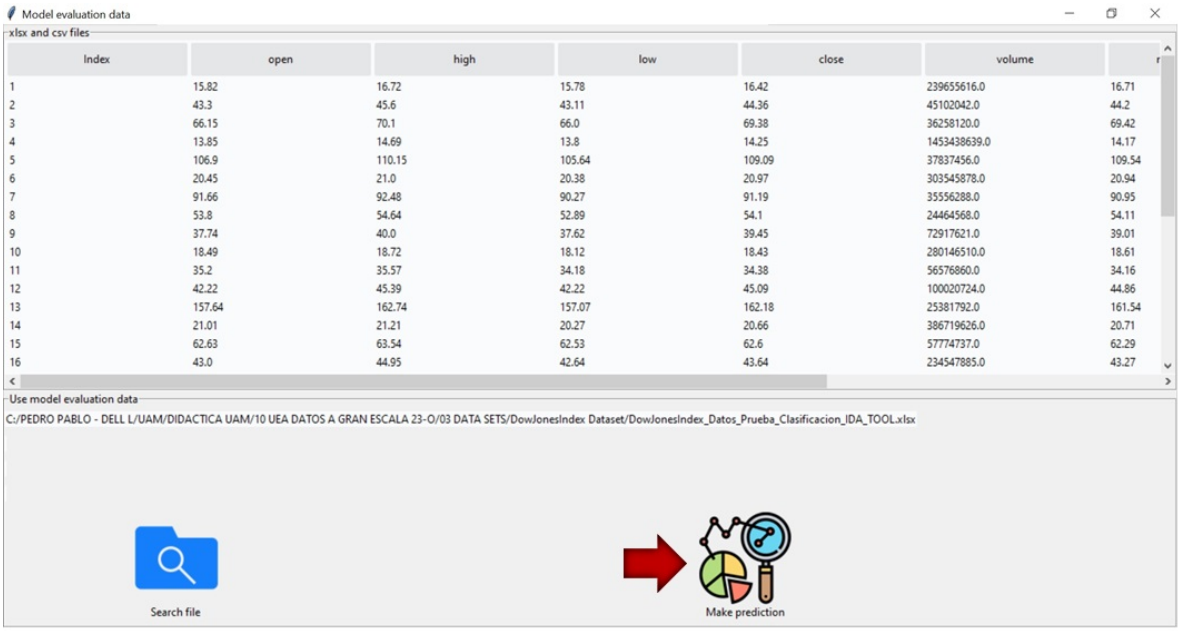


Figura 6.65. Construcción del modelo de clasificación, una vez que haya sido cargado el archivo con los datos de prueba.

Tras presionar el botón “Make prediction”, se iniciará la creación del modelo. Al concluir este proceso, se desplegará una ventana emergente donde deberá confirmar la acción pulsando “Aceptar” (ver figura 6.66).

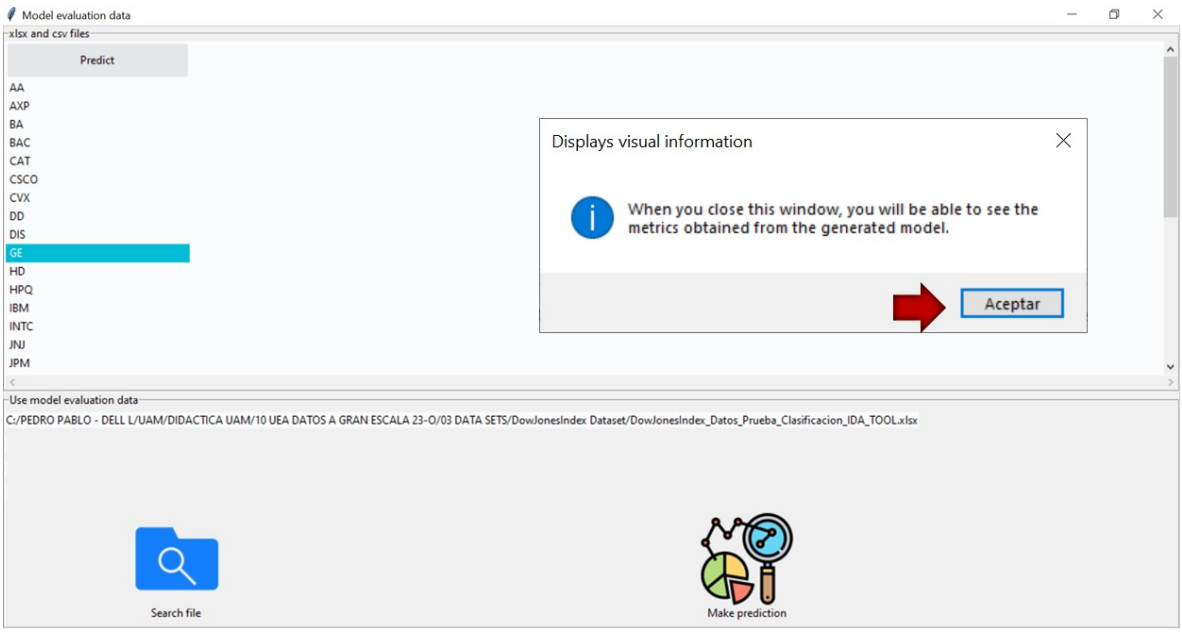


Figura 6.66. Conclusión de la construcción del modelo de clasificación y ventana emergente que permite visualizar los resultados.

Una vez concluido el proceso de creación del modelo de clasificación, y después de haber presionado “Aceptar” en el ícono de la ventana emergente que indica que la construcción del modelo finalizó (ver figura 6.66), se mostrarán las métricas de desempeño del modelo de clasificación, incluyendo la matriz de confusión, lo cual permite proceder a la evaluación del mismo (véanse figuras 6.67 y 6.68).

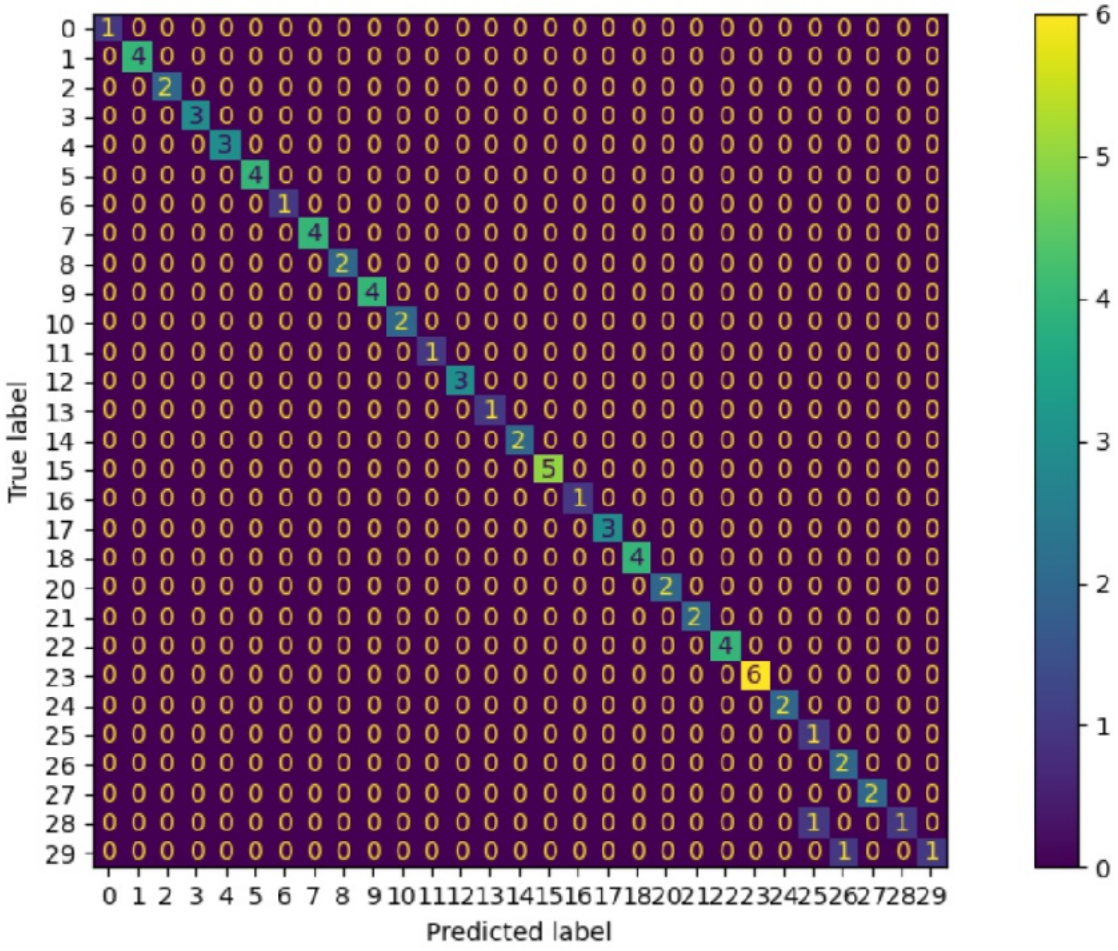


Figura 6.67. Matriz de confusión para las 30 clases que agrupan los registros del conjunto de datos “Índice Dow Jones” para el modelo de clasificación construido.

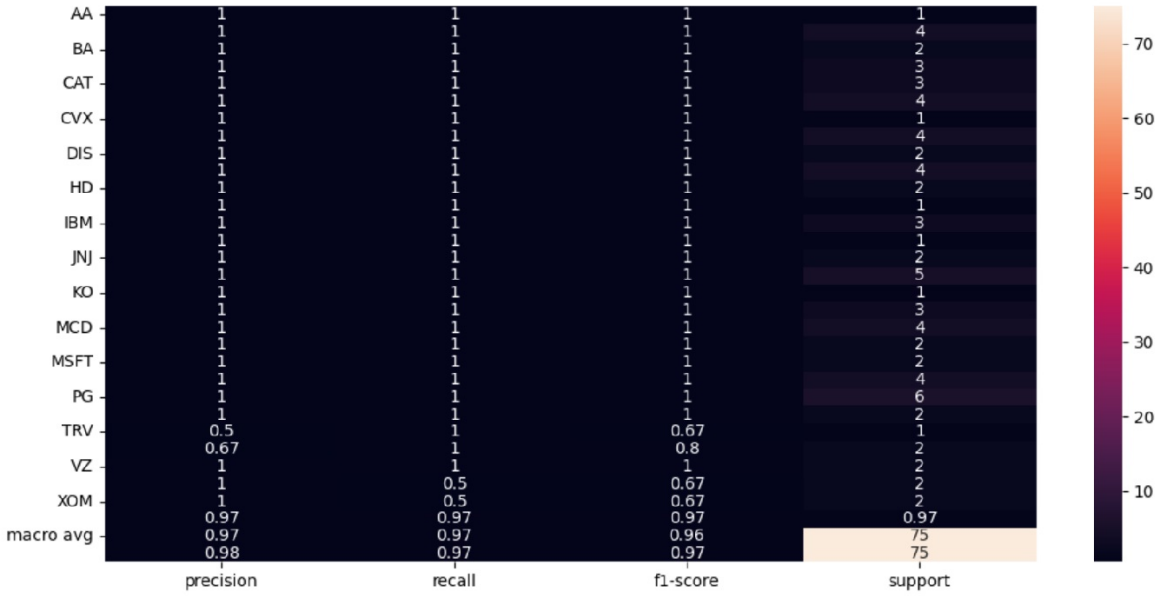


Figura 6.68. Métricas de desempeño para cada una de las 15 clases que integran el conjunto de datos de prueba.

6.3.4.2. Construcción y evaluación de un modelo de regresión en IDA-WEB TOOL

El proceso de construcción y evaluación de un modelo de regresión en IDA-WEB TOOL es muy similar al que caracteriza la construcción de un modelo de clasificación, tal como se ilustra en las figuras de la 6.59 a la 6.66. La principal diferencia radica en que, en este caso, el modelo debe seleccionarse del panel "Regression Models". Las figuras de la 6.69 a la 6.76 ilustran los principales pasos para la construcción y evaluación de un modelo para predicción. Para esta tarea, se utilizó nuevamente el conjunto de datos del "Índice Dow Jones", considerando sólo aquellos campos que fungen como buenos predictores y el total de registros.

El campo a predecir es: *percent_change_next_weeks_price*, mientras que los campos predictores son:

- *stock*
- *open*
- *high*
- *low*
- *close*
- *volumen*
- *next_weeks_open*
- *next_weeks_close*
- *percent_return_next_dividend*

Como se puede apreciar en la figura 6.69, una vez cargado el archivo con los datos, el siguiente paso consiste en la selección del modelo de predicción a utilizar, que en este caso se trata de una red neuronal perceptrón multi-estrato (MLP).

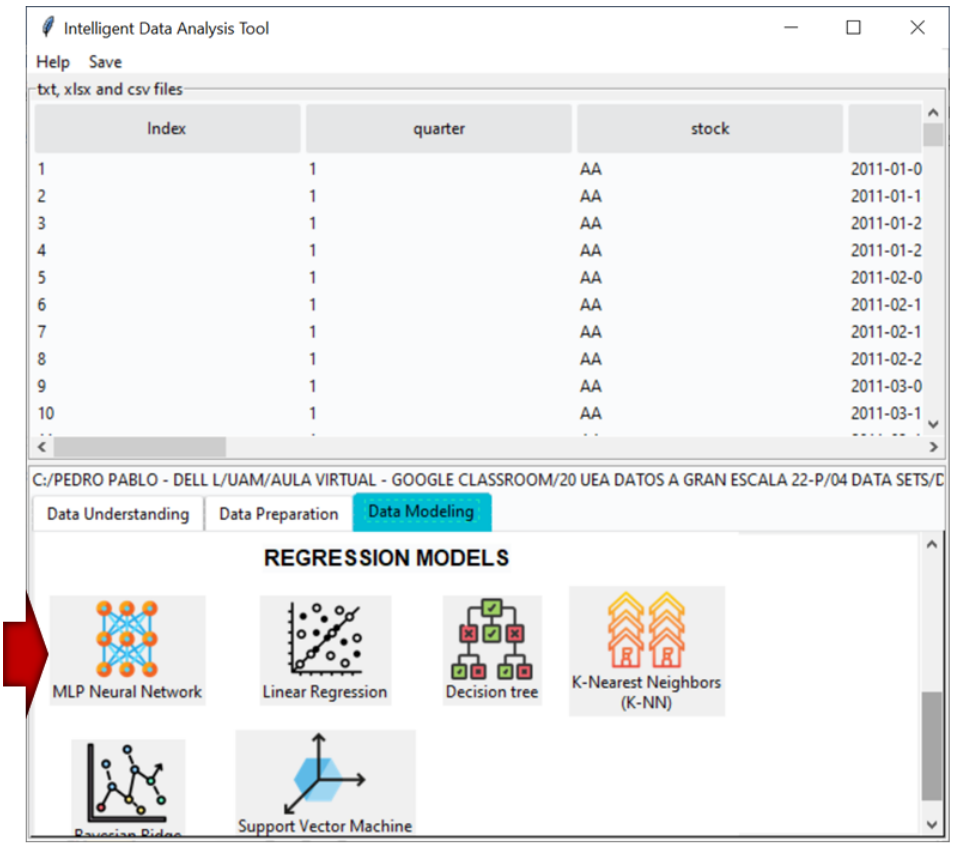


Figura 6.69. Selección del modelo MLP Neural Network, para la tarea de predicción, desde el menú “Data Modeling” de la herramienta IDA-TOOL.

Una vez que se ha elegido el tipo de modelo para la tarea de predicción, aparecerá en pantalla una interfaz gráfica para proceder a la selección de los campos predictores (entradas al modelo MLP) y el campo objetivo (predictivo) (ver figura 6.70); ambos campos se ilustran en la figura 6.71. Posteriormente, se desplegará en pantalla una nueva interfaz que permitirá seleccionar los íconos para iniciar con la construcción del modelo de predicción basado en el MLP. Antes de ello, es necesario cargar el conjunto de datos adicional que contiene los datos de prueba, asegurándose de que no estén contenidos en el conjunto inicial (ver figuras 6.72 y 6.73). Una vez hecho esto, es posible pasar a la generación del modelo, seleccionando el ícono “Make prediction”, tal como se ilustra en la figura 6.74.

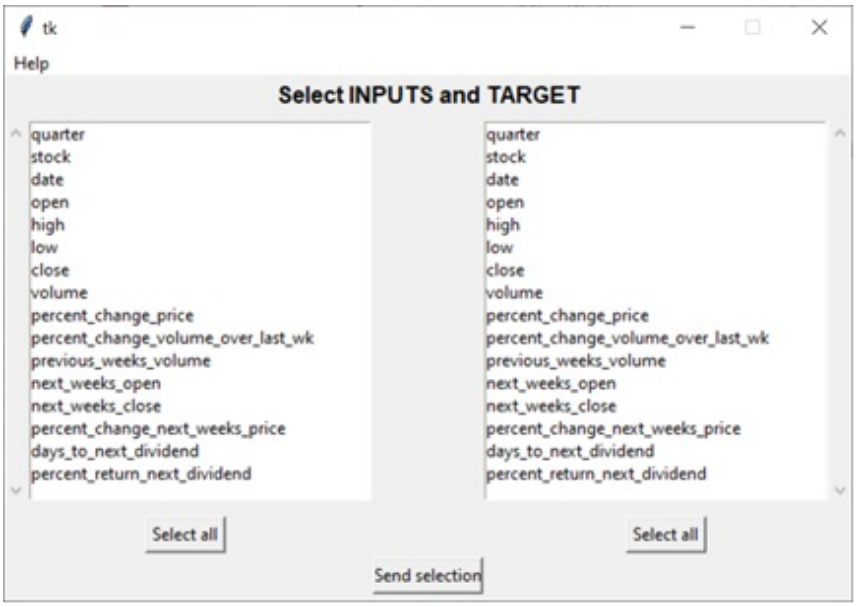


Figura 6.70. Interfaz gráfica para la selección de los campos predictores y el campo predictivo.

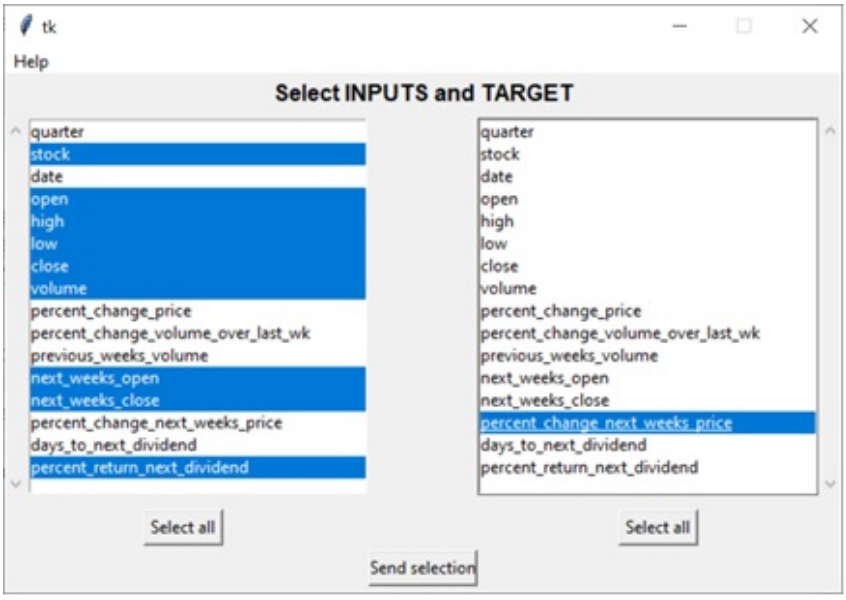


Figura 6.71. Selección de los campos predictores (entradas al modelo MLP) y el campo predictivo (salida que debe producir el modelo MLP).

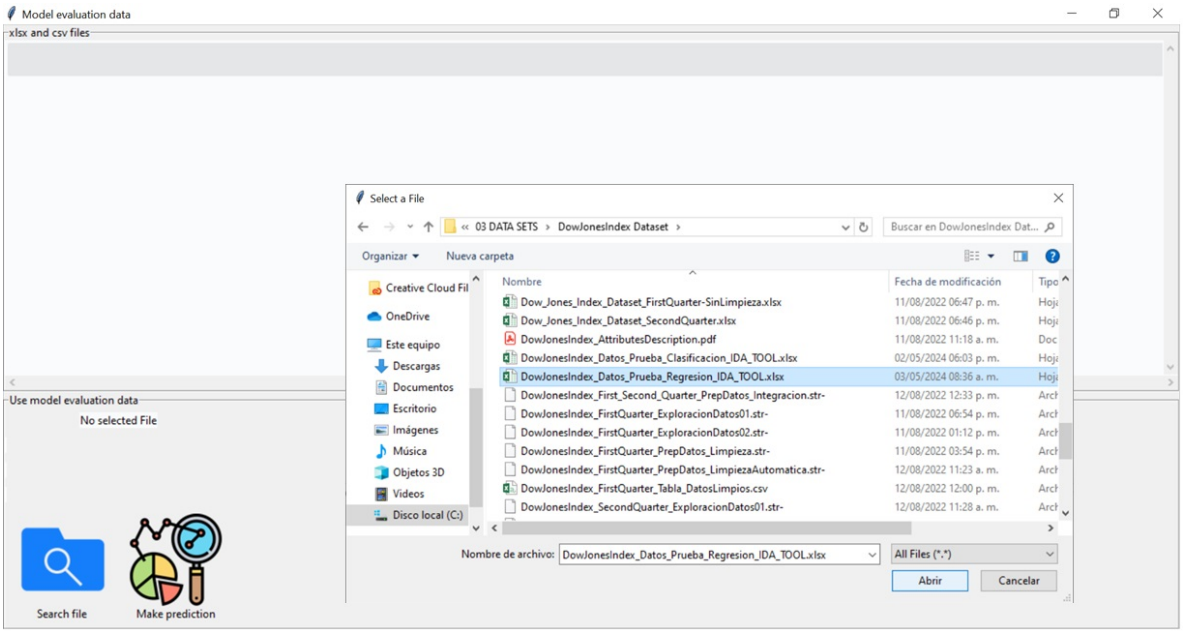
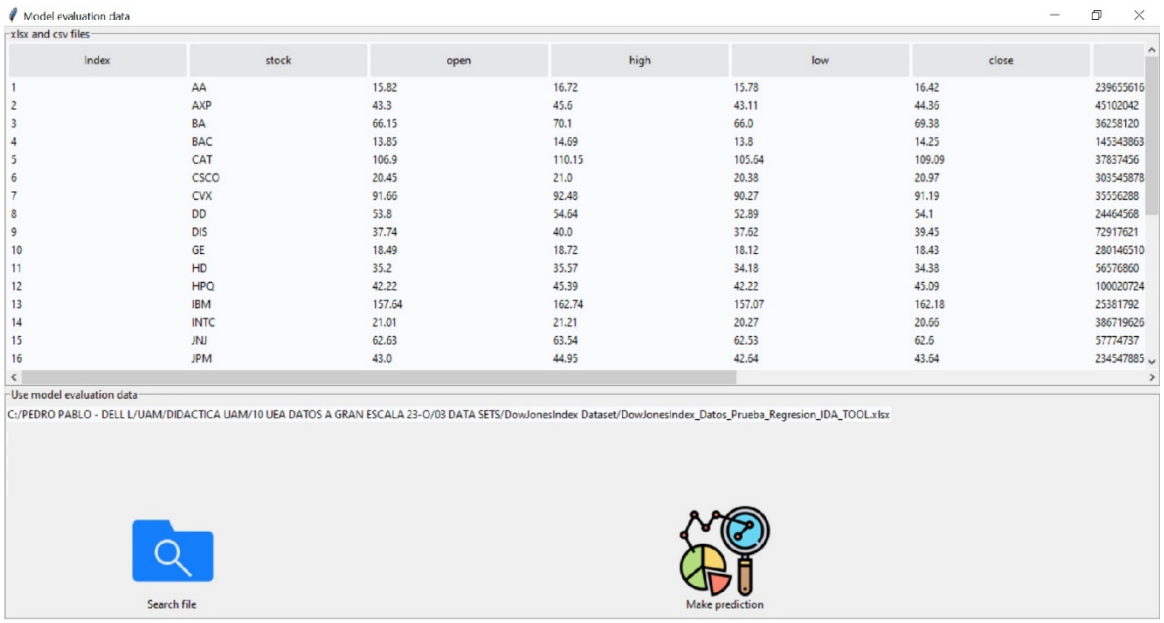


Figura 6.72. Paso 1 de la carga del archivo con los datos de prueba, previo a la generación del modelo de predicción.



Index	stock	open	high	low	close	
1	AA	15.82	16.72	15.78	16.42	239655616
2	AXP	43.3	45.6	43.11	44.36	45102042
3	BA	66.15	70.1	66.0	69.38	36258120
4	BAC	13.85	14.69	13.8	14.25	145343863
5	CAT	106.9	110.15	105.64	109.09	37837456
6	CSCO	20.45	21.0	20.38	20.97	303545878
7	CVX	91.66	92.48	90.27	91.19	35556288
8	DD	53.8	54.64	52.89	54.1	24464568
9	DIS	37.74	40.0	37.62	39.45	72917621
10	GE	18.49	18.72	18.12	18.43	280146510
11	HD	35.2	35.57	34.18	34.38	56576860
12	HPQ	42.22	45.39	42.22	45.09	100020724
13	IBM	157.64	162.74	157.07	162.18	25381792
14	INTC	21.01	21.21	20.27	20.66	386719626
15	JNJ	62.63	63.54	62.53	62.6	57774737
16	JPM	43.0	44.95	42.64	43.54	234547885

Figura 6.73. Paso 2 de la carga del archivo con los datos de prueba, previo a la generación del modelo de predicción.

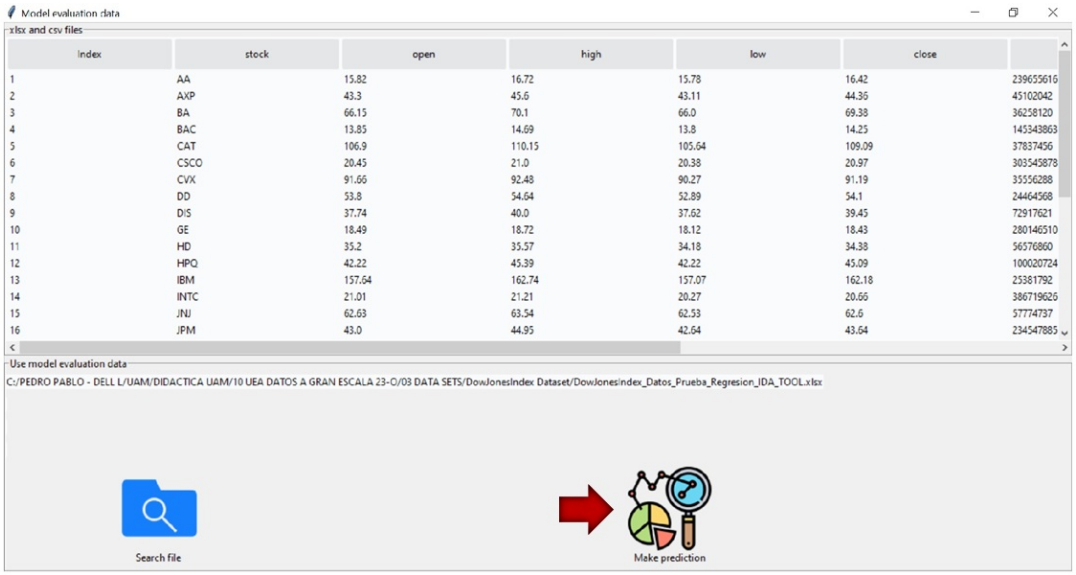


Figura 6.74. Selección del ícono “Make prediction” para iniciar la construcción del modelo de predicción con la técnica elegida (MPL model).

Tras presionar el botón “Make prediction”, se iniciará la creación del modelo de regresión. Una vez concluido este proceso, se desplegará una ventana emergente donde se deberá confirmar la acción de continuar pulsando “Aceptar” (ver figura 6.75). Posteriormente, se mostrarán las métricas de desempeño del modelo de regresión (ver figura 6.76).

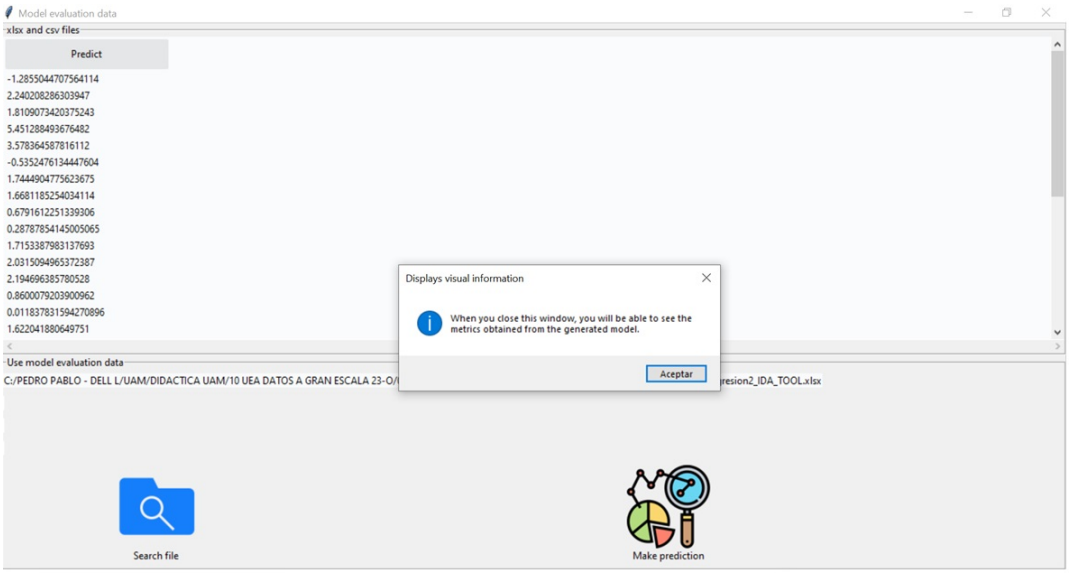


Figura 6.75. Conclusión de la construcción del modelo de regresión y ventana emergente que permite continuar a la visualización de los resultados.

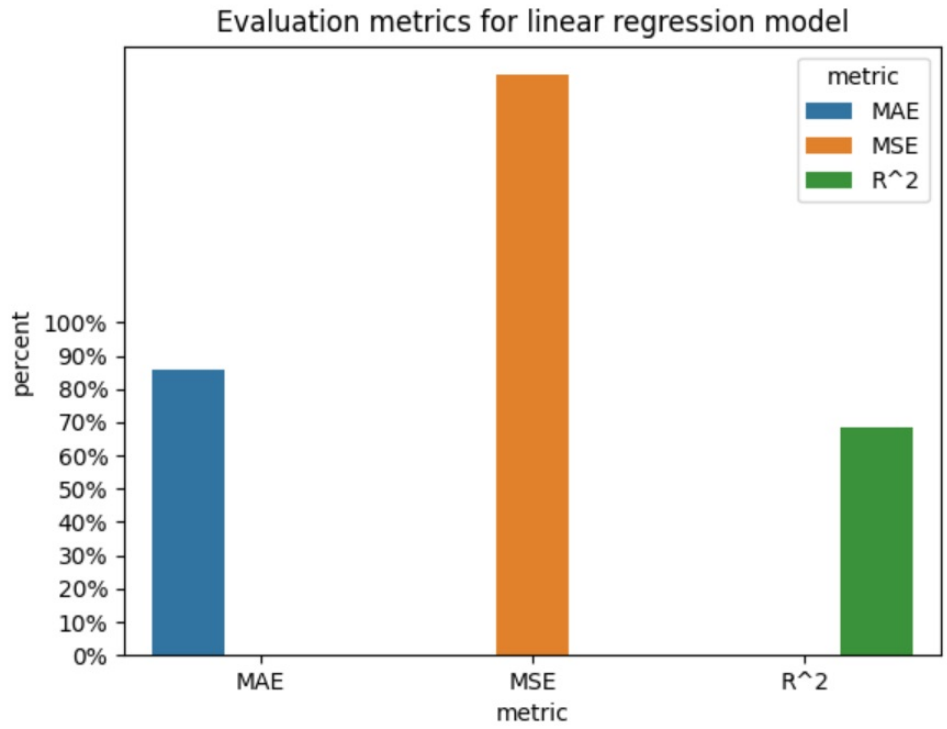


Figura 6.76. Métricas de evaluación del desempeño del modelo de regresión.

6.4. Consideraciones finales

A través del caso de estudio del "Índice Dow Jones" se han ilustrado las principales funciones que IDA-WEB TOOL proporciona al usuario. Como se ha indicado previamente, esta no es una herramienta de minería de datos profesional, como IBM SPSS Modeler, sino que está construida a nivel prototipo, por lo que su funcionalidad se debe continuar probando, refinando e incrementando. Sin embargo, hemos querido presentarla en este material por el gran esfuerzo que representa su desarrollo, al ser el resultado del trabajo realizado durante los cursos de Proyectos terminales I, II, y III, y como proyecto de servicio social de tres alumnos de la licenciatura en Ingeniería en Computación de la Universidad Autónoma Metropolitana, Unidad Cuajimalpa.

Como parte de nuestro trabajo actual y futuro, las siguientes mejoras están siendo incorporadas a la herramienta IDA-WEB TOOL:

1. Una mayor interactividad en los modelos de regresión y clasificación que permita el ajuste de parámetros y proporcione más resultados útiles al usuario, tales como otras métricas de evaluación del desempeño del modelo, la importancia de los predictores, entre otros.

2. Un componente de selección de características que proporcione varias técnicas al usuario, entre ellas, el análisis de componentes principales.
3. Un componente basado en el algoritmo SMOTE para equilibrar la cantidad de registros en cada clase o categoría, cuando se trate de un objetivo de minería de datos de clasificación. Esto permitirá afrontar el problema del desbalance de las clases.

VII. CASOS DE ESTUDIO

Tomando en cuenta que la herramienta de minería de datos IDA-WEB TOOL se encuentra aún en su fase de evaluación y pruebas, la actividad de exploración de los datos y las fases de preparación de los datos y modelado de los dos casos de estudio presentados a continuación se llevarán a cabo utilizando principalmente la herramienta de minería de datos IBM SPSS MODELER.

7.1. Caso de estudio 1: Mercado de bienes de consumo

El conjunto de datos original en el que se basa el presente caso de estudio se tomó de los conjuntos de datos que provee la herramienta de minería de datos IBM SPSS MODELER, en su versión para la comunidad estudiantil (https://www.ibm.com/mx-es/products/spss-modeler), para la cual poseemos licencia académica. El tamaño del conjunto original era de 200 registros o instancias, y se aumentó a 2500, mediante la técnica Synthetic Minority Over-sampling Technique (SMOTE), para mejorar el desempeño de los modelos de aprendizaje automatizado.

7.1.1. Comprensión del dominio del problema

7.1.1.1. Determinación de los objetivos del proyecto

Determinar los objetivos del proyecto implica:

- Comprender los factores que influyen positivamente en el incremento de las ventas de cuatro categorías de bienes de consumo: artículos de lujo (*luxury*), bebidas y licores (*drink*), productos cárnicos (*meat*) y productos confeccionados (*confection*).
- Comparar el comportamiento del incremento de las ventas entre las cuatro categorías de bienes de consumo, con base en la promoción aplicada.
- Predecir el incremento en las ventas de bienes de consumo, con base en las categorías predispuestas y la promoción aplicada.

7.1.1.2. Valoración de la situación actual del objetivo del proyecto

- ¿Se comprende de forma clara el problema a abordar?
    - Un bien de consumo es el producto final de un proceso de producción, el cual va dirigido a satisfacer las necesidades de las personas. Por lo tanto, el mercado de bienes de consumo de refiere a aquel donde los productos comercializados están destinados a satisfacer las

necesidades del cliente final.

- Los bienes de consumo pueden ser de consumo inmediato (no perdurables), por ejemplo, los alimentos y las bebidas, o de consumo duradero, como la vivienda, los muebles, equipos electrodomésticos, electrónica, etcétera.
- A diferencia de otros bienes, como los de capital, los bienes de consumo no crean otros productos, sino que benefician directamente al cliente final.
- Aunque existen diferentes categorías para clasificar los bienes de consumo, una de las principales es según su tiempo de uso o duración:
    - Bienes de consumo duraderos
    - Bienes de consumo no duraderos
- De forma paralela, existen varias formas de clasificar los bienes de consumo según el tipo de productos que representa, por ejemplo:
    - Electrodomésticos
    - Electrónica
    - Bebidas y licores
    - Alimentos confeccionados
    - Carnes
    - Perfumería, entre otros.
- Dos características claves de los bienes de consumo son su clasificación y precio. Por ejemplo:
    - Bienes de consumo duradero:
        - Electrodomésticos:
            - Refrigerador marca AX200, precio: $18,600.00
            - Lavadora marca BZ1000, precio: $14,500.00
        - Muebles:
            - Sala de cuatro piezas modelo FT500, precio: $24,000.00
            - Recámara modelo ITA300, precio: $18,400.00
        - Electrónica:
            - TV de 60 pulgadas 4K marca IO2000, precio: $18,900.00.
            - Laptop marca Spin600, precio: $14,800.00
        - Automóviles:
            - Auto sedán marca AT150, modelo MS200, precio: $360,000.00
    - Bienes de consumo no duradero:
        - Carnes:
            - Milanesa de res, precio: $180.00/kg

        - Pechuga de pollo, precio: $120.00/kg
    - Alimentos procesados:
        - Galletas surtidas, precio: $80.00 la caja de 400 g
        - Atún en lata, marca AT30, precio: $24.00 la lata de 380 g
        - Pastel de chocolate, precio: $120.00/kg
    - Bebidas y licores:
        - Vino tinto, marca VT10, precio: $265.00
        - Whisky, marca Wh80, precio: $320.00
        - Cerveza, marca CR40, precio: $45.00

- La gran mayoría de los bienes de consumo se encuentran sujetos a ofertas y promociones sobre su precio de mercado, en determinados períodos del año, con la finalidad de incentivar el consumo.
- Para lograr que las ofertas y promociones logren realmente incentivar el consumo, y por lo tanto reflejen un incremento real en las ventas, es necesario considerar varios factores, entre ellos:
    1) Período del año en que se lanza la oferta o promoción
    2) Tipo de bien de consumo (duradero o no duradero)
    3) Categoría a la que pertenece el bien de consumo (electrodoméstico, electrónica, automóviles, alimentos elaborados, bebidas y licores, perfumería, etcétera)
    4) Precio comercial del bien de consumo
    5) Porcentaje de descuento respecto al precio comercial
    6) Características de la población de las áreas o regiones consideradas para la oferta o promoción

❖ ¿Existen datos disponibles para efectuar el análisis?

- Sí. Se cuenta con un conjunto de datos con 200 registros aumentado a 2500, que describe los volúmenes de ventas antes y después de aplicar promociones, para cuatro clases de productos de consumo:
    - ❑ Bebidas y licores
    - ❑ Carnes
    - ❑ Alimentos elaborados
    - ❑ Artículos de lujo

- El registro de estos datos se realizo en el siguiente enlace: https://www.ibm.com/mx-es/products/spss-modeler (ver figura 7.1).

1	Class	Cost	Promotion	Before	After
2	Confection	23.99	1467	114957	122762
3	Drink	79.29	1745	123378	137097
4	Luxury	81.99	1426	135246	141172
5	Confection	74.18	1098	231389	244456
6	Confection	90.09	1968	235648	261940
7	Meat	69.85	1486	148885	156232
8	Meat	100.15	1248	123760	128441
9	Luxury	21.01	1364	251072	268134
10	Luxury	87.32	1585	287043	310857
11	Drink	26.58	1835	240805	272863
12	Drink	65.23	1194	212406	227836
13	Meat	79.82	1596	174022	181489
14	Confection	41.39	1161	270631	283189
15	Meat	36.82	1151	231281	235722
16	Meat	44.05	1482	178138	185934
17	Drink	84.62	1623	247885	278031
18	Confection	51.82	1969	148597	165598
19	Confection	90.08	1462	215102	228696
20	Luxury	57.3	1842	246885	270082
21	Drink	11.02	1370	164984	176802
22	Confection	95.86	1815	257882	284835

Figura 7.1. Fragmento del conjunto de datos “Bienes de consumo”.

7.1.1.3. Determinación de los objetivos de minería de datos

- **Problema de predicción**: Esto implica construir varios modelos predictivos supervisados que permitan predecir el incremento en las ventas para cada clase de bien de consumo, así como seleccionar el modelo predictivo con mayor precisión (menor tasa de error).

7.1.1.4. Propuesta del enfoque metodológico

La propuesta metodológica para el desarrollo del proyecto “Bienes de consumo” se presenta en la figura 7.2.

Fase	Tiempo a dedicar	Recursos humanos y tecnológicos	Riesgos atribuibles
1. Comprensión del dominio del problema	1 semana	Experto en mercadotecnia, experto en minería de datos.	No se han identificado riesgos.
2. Comprensión de los datos	2 semana	Experto en el dominio del problema, experto en minería de datos. Tablas, gráficos y resúmenes estadísticos que faciliten la comprensión de los datos.	No se trata de un gran volumen de datos, lo cual podría influir en la precisión de los resultados, y por lo tanto en el poder de la herramienta predictiva.
3. Preparación de los datos	2 semanas	Experto en mercadotecnia, experto en minería de datos. Herramientas para el análisis exploratorio de datos. Herramienta *IDA WEB TOOL* Paquete *IBM SPSS Modeler*.	
4. Modelado	2 semanas	Experto en minería de datos y experto en técnicas de *machine learning*. Herramientas para la implementación de modelos de *machine learning*. Herramienta *IDA WEB TOOL* Paquete *IBM SPSS Modeler*.	No se han identificado riesgos.
5. Evaluación	1 semana	Experto en mercadotecnia, experto en minería de datos. Herramienta *IDA WEB TOOL* Paquete *IBM SPSS Modeler*.	No se han identificado riesgos.
6. Presentación	1 semana	Experto en mercadotecnia, experto en minería de datos, directiva de la organización. Paquete *IBM SPSS Modeler*.	No se han identificado riesgos.

Figura 7.2. Propuesta metodológica para el desarrollo del proyecto “Bienes de consumo”.

7.1.2. Comprensión de los datos

7.1.2.1. Recopilación de los datos iniciales

La fuente del conjunto de datos “Bienes de consumo” es el propio repositorio de datos de la herramienta de minería de datos IBM SPSS MODELER, en su versión para la comunidad estudiantil (https://www.ibm.com/mx-es/products/spss-modeler). Como se ha mencionado, el tamaño del conjunto de datos original era de 200 registros o instancias, y fue aumentado a 2500 para mejorar el desempeño de los modelos de aprendizaje automatizado. La figura 7.3 presenta un fragmento de dicho conjunto, utilizando el visor de archivos de IDA-WEB TOOL.

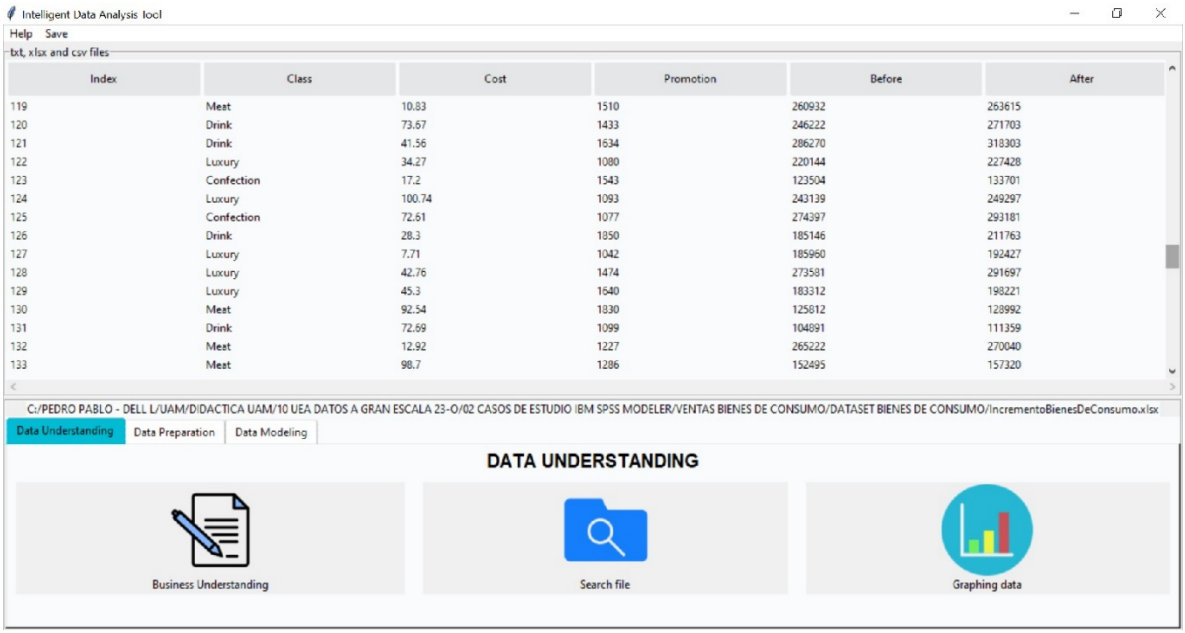


Figura 7.3. Fragmento ilustrativo del conjunto de datos “Bienes de consumo”, utilizando el visor de la herramienta de minería de datos IDA-WEB TOOL.

7.1.2.2. Descripción de los datos

Como se puede apreciar en la tabla 7.1, la estructura inicial del conjunto de datos “Bienes de consumo” comprende 5 atributos. Como ya se había mencionado, el tamaño del conjunto de datos original era de 200 registros, y fue incrementado a 2500 para incrementar el desempeño de los modelos de aprendizaje automatizado.

Tabla 7.1. Atributos y tipos de datos del conjunto “Bienes de consumo”

Nombre del atributo	Descripción	Tipo
Class	Clase o tipo del bien de consumo. En el conjunto de datos se identifican cuatro clases: ❖ Artículo confeccionado ❖ Bebidas y licores ❖ Productos cárnicos ❖ Productos de lujo	Categórico
Cost	Costo unitario del producto	Continuo
Promotion	Monto de la promoción	Continuo
Before	Monto de las ventas antes de aplicar la promoción	Continuo
After	Monto de las ventas después de aplicar la promoción	Continuo

7.1.2.3. Exploración de los datos

Las figuras de la 7.4 a la 7.7 ilustran parte de la actividad de exploración de los datos, a través de gráficos que permiten analizar la importancia de los campos predictores.

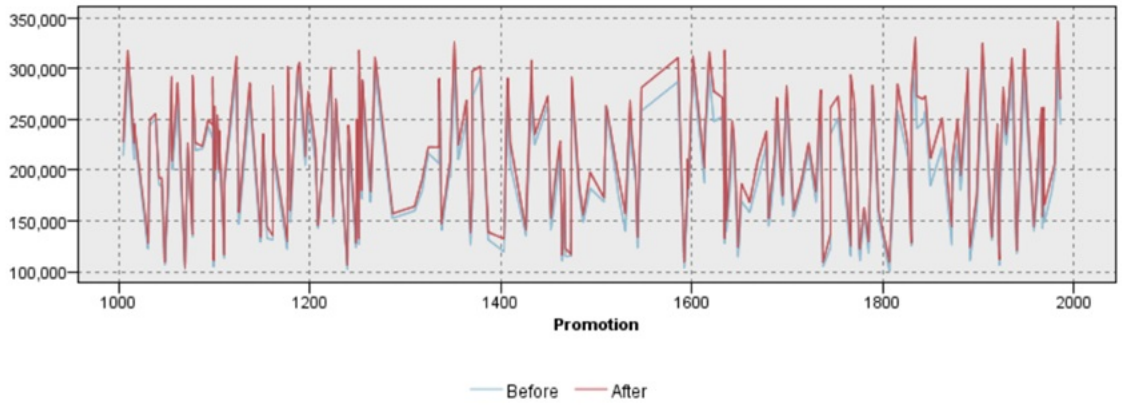


Figura 7.4. Exploración de los datos. Gráfico múltiple.

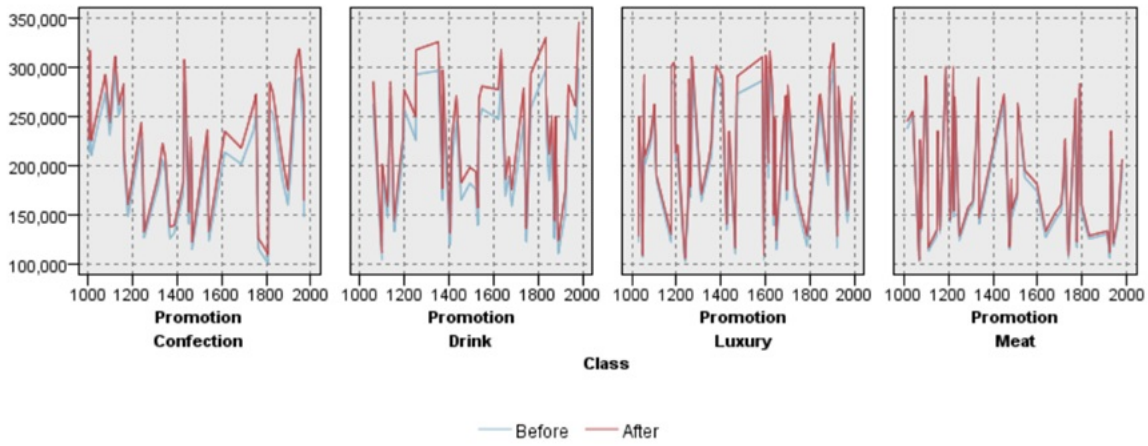


Figura 7.5. Exploración de los datos. Gráfico múltiple por clase del bien de consumo. Comportamiento de las ventas antes y después de aplicar la promoción respecto al monto de la promoción.

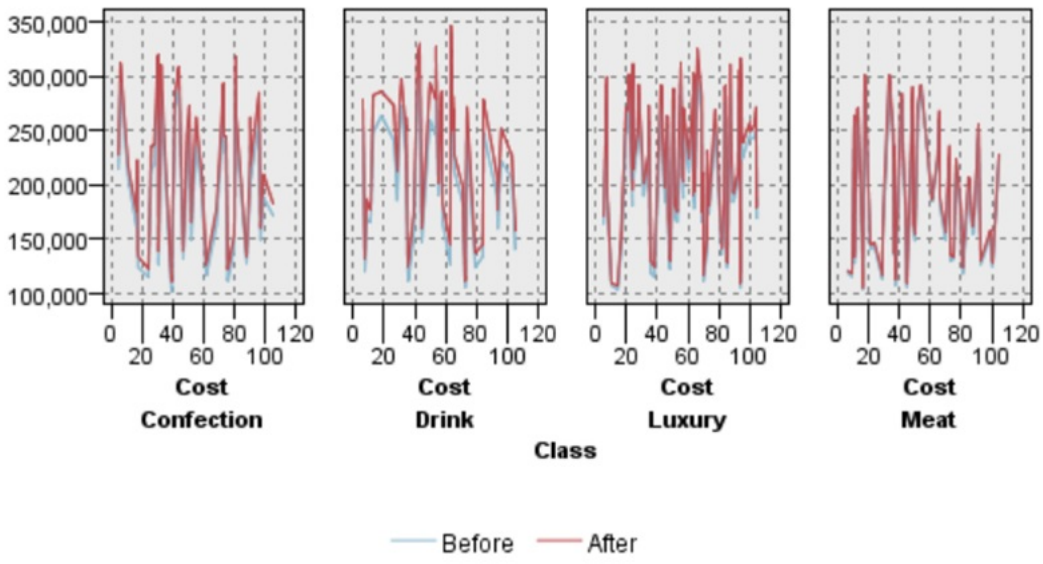


Figura 7.6. Exploración de los datos. Gráfico múltiple por clase del bien de consumo. Comportamiento de las ventas antes y después de aplicar la promoción respecto al costo unitario del producto.

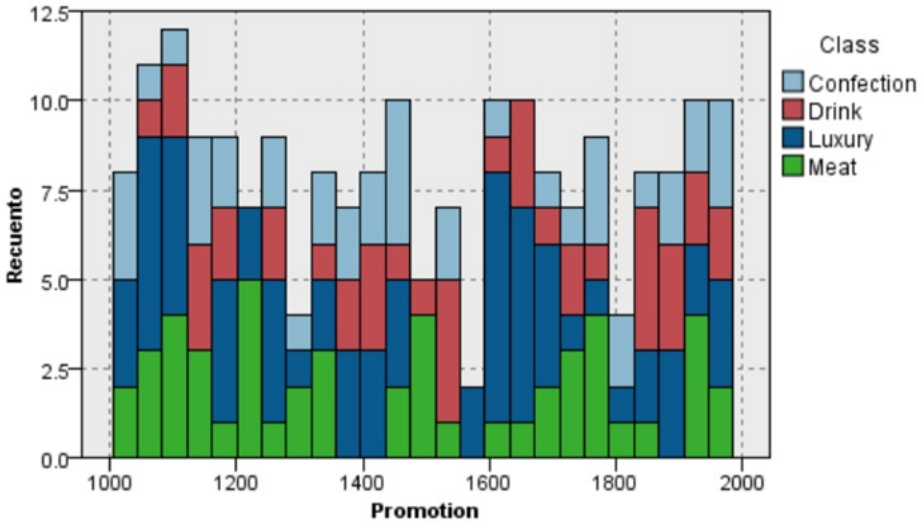


Figura 7.7. Exploración de los datos. Gráfico de recuento por clase del bien de consumo.

7.1.2.4. Verificación de la calidad de los datos

- Datos perdidos o vacíos: No existen.
- Errores en los datos: Aparentemente no existen.
- Errores de medición: Aparentemente no existen.
- Errores de codificación: Aparentemente no existen.
- Atributos redundantes o de escasa utilidad: No se han identificado.

7.1.3. Preparación de los datos

7.1.3.1. Selección de los datos

- Inicialmente, se mantendrán los 6 atributos y 200 registros que componen el conjunto de datos. Más adelante se considerará la derivación de un nuevo atributo, lo que podría ocasionar la exclusión de uno de los existentes, al dejar de ser relevante
- Como ya se indicó en la fase de comprensión del dominio del problema o negocio, el número de registros podría resultar insuficiente para garantizar precisión en los objetivos de minería de datos. Por tal motivo, en una fase posterior de este caso de estudio, el conjunto de datos fue aumentado de 200 a 2,500 registros utilizando la técnica SMOTE.

7.1.3.2. Limpieza de los datos

Como ya se indicó en la fase de comprensión de los datos, no será necesario corregir problemas relacionados con datos perdidos o vacíos, errores en los datos o errores de codificación. En este caso de estudio, por lo tanto:

- Datos perdidos o vacíos: No existen.
- Errores en los datos: Aparentemente no existen.
- Errores de medición: Aparentemente no existen.
- Errores de codificación: Aparentemente no existen.
- Atributos redundantes o de escasa utilidad: No se han identificado.

7.1.3.3. Construcción de nuevos datos

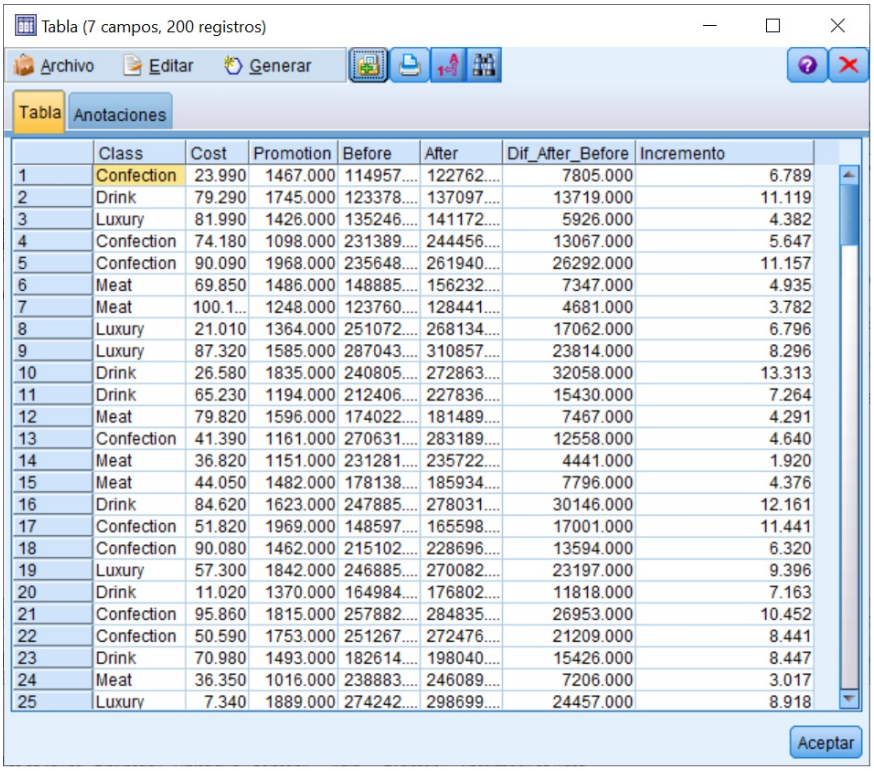
Como se analizó previamente, el conjunto de datos está integrado por los siguientes campos o atributos:

- *Class* (clase o categoría)
- *Cost* (costo)
- *Promotion* (promoción)
- *Before* (monto de ventas antes de la promoción)
- *After* (monto de ventas posterior a la promoción)

Sin embargo, se requiere derivar el atributo *increase* (incremento), a partir de los atributos iniciales *before* y *after*, puesto que representa el incremento de las ventas con relación a la promoción aplicada. De esta forma, *increase* será el atributo objetivo, es decir, aquel cuyo valor se desea predecir. La figura 7.8 ilustra la construcción de este nuevo atributo utilizando la herramienta IBM SPSS MODELER.

En tanto, la figura 7.9 muestra un fragmento del conjunto de datos que contiene el atributo *increase*.

Figura 7.8. Construcción del campo *increase*.



	Class	Cost	Promotion	Before	After	Dif_After_Before	Incremento
1	Confection	23.990	1467.000	114957....	122762....	7805.000	6.789
2	Drink	79.290	1745.000	123378....	137097....	13719.000	11.119
3	Luxury	81.990	1426.000	135246....	141172....	5926.000	4.382
4	Confection	74.180	1098.000	231389....	244456....	13067.000	5.647
5	Confection	90.090	1968.000	235648....	261940....	26292.000	11.157
6	Meat	69.850	1486.000	148885....	156232....	7347.000	4.935
7	Meat	100.1...	1248.000	123760....	128441....	4681.000	3.782
8	Luxury	21.010	1364.000	251072....	268134....	17062.000	6.796
9	Luxury	87.320	1585.000	287043....	310857....	23814.000	8.296
10	Drink	26.580	1835.000	240805....	272863....	32058.000	13.313
11	Drink	65.230	1194.000	212406....	227836....	15430.000	7.264
12	Meat	79.820	1596.000	174022....	181489....	7467.000	4.291
13	Confection	41.390	1161.000	270631....	283189....	12558.000	4.640
14	Meat	36.820	1151.000	231281....	235722....	4441.000	1.920
15	Meat	44.050	1482.000	178138....	185934....	7796.000	4.376
16	Drink	84.620	1623.000	247885....	278031....	30146.000	12.161
17	Confection	51.820	1969.000	148597....	165598....	17001.000	11.441
18	Confection	90.080	1462.000	215102....	228696....	13594.000	6.320
19	Luxury	57.300	1842.000	246885....	270082....	23197.000	9.396
20	Drink	11.020	1370.000	164984....	176802....	11818.000	7.163
21	Confection	95.860	1815.000	257882....	284835....	26953.000	10.452
22	Confection	50.590	1753.000	251267....	272476....	21209.000	8.441
23	Drink	70.980	1493.000	182614....	198040....	15426.000	8.447
24	Meat	36.350	1016.000	238883....	246089....	7206.000	3.017
25	Luxury	7.340	1889.000	274242....	298699....	24457.000	8.918

Figura 7.9. Fragmento del conjunto de datos “Bienes de consumo” con el nuevo atributo derivado *increase*.

7.1.3.4. Integración de datos

No se cuenta con otras fuentes de datos a integrar al conjunto inicial. Por lo tanto, éste será analizado en la fase de modelado.

7.1.3.5. Formato de datos

Como se puede apreciar en la figura 7.10, los datos se encuentran en el formato requerido para la aplicación de los modelos predictivos, por lo tanto, en relación al tipo de dato, sólo será necesario verificar que los atributos *cost*, *before*, *promotion* y *after* sean de tipo continuo, mientras que el atributo *class* debe ser de tipo nominal. En cuanto al rol, todos los atributos predictores deben ser de tipo entrada, mientras que el nuevo atributo derivado *increase* deberá ser de tipo destino. Dado que ahora se cuenta con este nuevo atributo, *after* deja de ser relevante, por lo que no será considerado como atributo de entrada para los modelos predictivos.

Campo	Medida	Valores	No se enc...	Comprobar	Rol
Class	Nominal	Confection,...		Ninguno	Entrada
Cost	Continuo	[5.08,104.98]		Ninguno	Entrada
Promotion	Continuo	[1004.0,19...		Ninguno	Entrada
Before	Continuo	[100751.0,...		Ninguno	Entrada
After	Continuo	[104393.0,...		Ninguno	Ninguna
Increase	Continuo	[0.4365444...		Ninguno	Destino

Figura 7.10. Formato de datos.

7.1.3.6. Nueva exploración de los datos

Las figuras de la 7.11 a la 7.13 muestran la nueva exploración de los datos, considerando el atributo derivado *increase*. Nótese en estas figuras la importancia de este nuevo atributo.

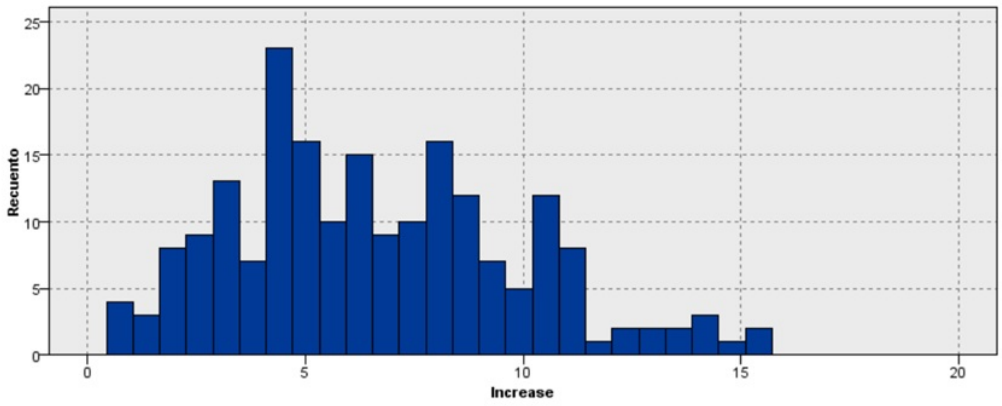


Figura 7.11. Nueva exploración de los datos. Gráfico de recuento para la característica *increase*.

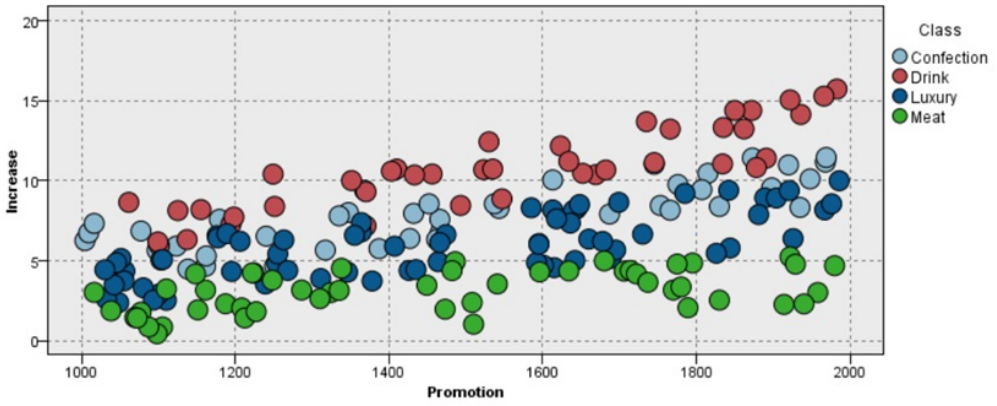


Figura 7.12. Nueva exploración de los datos. Gráfico *increase vs. promotion* por clase de bien de consumo.

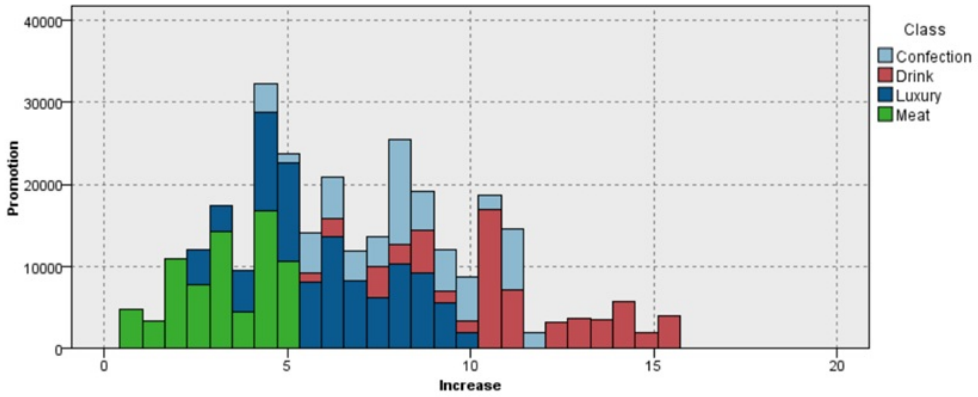


Figura 7.13. Nueva exploración de los datos. Gráfico *promotion vs. increase* por clase de bien de consumo.

7.1.4. Modelado

7.1.4.1. Selección de técnicas de modelado

Se escogieron cuatro modelos que responden al objetivo de minería de datos que se trata en este material, el cual es la predicción del incremento en las ventas según la clase del bien de consumo. Los modelos seleccionados fueron los siguientes:

- Regresión lineal
- Árbol de decisión C&R
- Red neuronal MLP
- Máquina de vectores de soporte (SVM)
- Algoritmo KNN

7.1.4.2. Métodos de comprobación

Considerando que los modelos propuestos son técnicas de aprendizaje supervisado, el método de comprobación es seleccionado por el criterio de bondad del modelo. En este caso es el coeficiente de correlación, dado que se trata de un problema de regresión. Con relación a los datos necesarios para comprobar el criterio de bondad, los cuatro modelos propuestos incluyen entre sus cualidades la partición de los datos en dos conjuntos, uno para la fase de entrenamiento y el otro para la fase de prueba o generalización.

7.1.4.3. Generación de los modelos

Como se puede apreciar en la figura 7.14, los siguientes modelos fueron generados utilizando el paquete IBM SPSS MODELER, e inicialmente se emplearon los parámetros propuestos por defecto para cada modelo.

- Regresión lineal
- Árbol de decisión C&R
- Red neuronal MLP
- Máquina de vectores de soporte (SVM)
- Algoritmo KNN

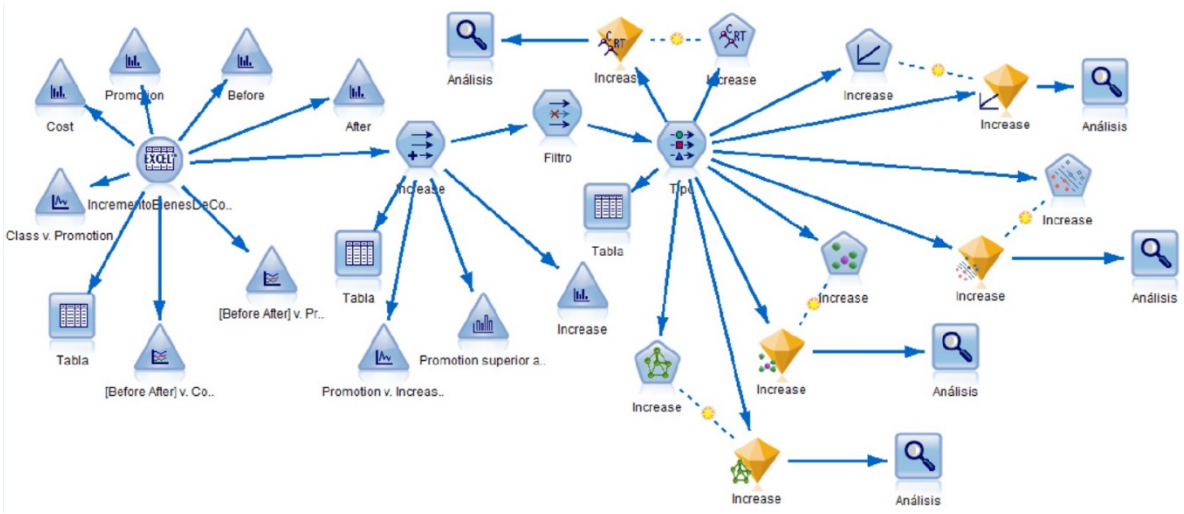


Figura 7.14. Creación de los cinco modelos de predicción utilizando la herramienta IBM SPSS MODELER.

Las figuras de la 7.15 a la 7.17 muestran la creación del modelo de regresión lineal. En la figura 7.15 se puede apreciar el modelo de regresión generado; en la 7.16 se muestra la importancia de las características predictoras, y finalmente, la 7.17 indica las métricas de evaluación del desempeño del modelo generado.

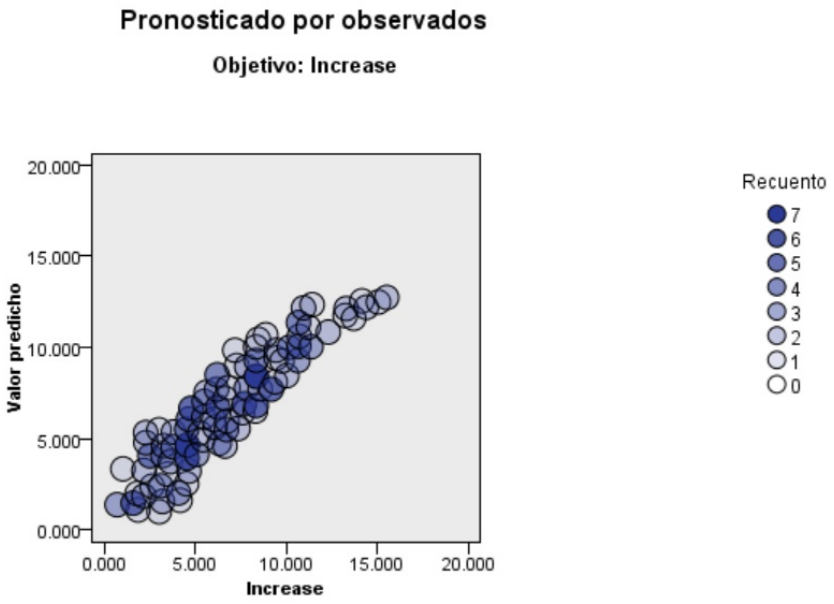


Figura 7.15. Creación del modelo de regresión lineal.

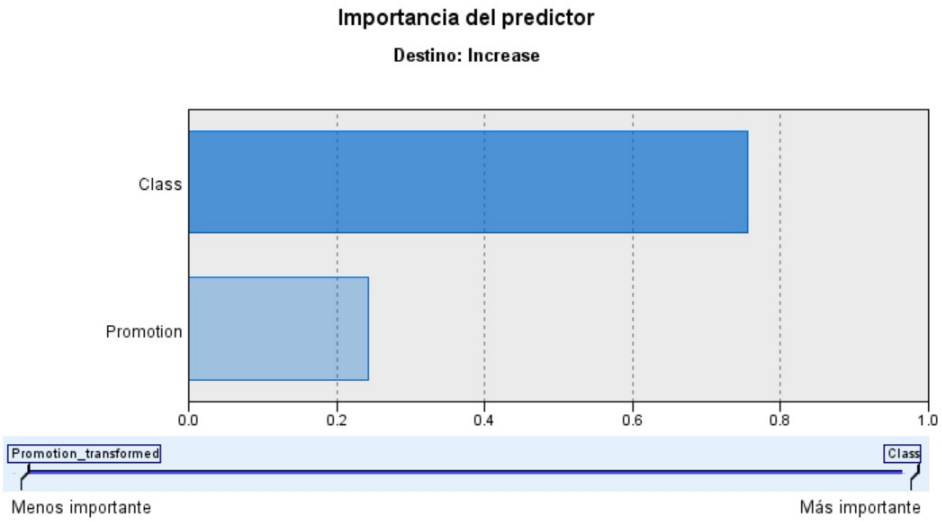


Figura 7.16. Importancia del predictor en la creación del modelo de regresión lineal.

Resultados para el campo de resultado Increase

Comparando $L-Increase con Increase

Error mínimo	-3.107
Error máximo	2.932
Error promedio	-0.0
Error absoluto promedio	1.118
Desviación estándar	1.34
Correlación lineal	0.913
Ocurrencias	200

Figura 7.17. Métricas de evaluación del desempeño del modelo de regresión lineal.

La creación del modelo de árbol de decisión C&R se ilustra en las figuras de la 7.18 a la 7.20. En la figura 7.18 se aprecia la estructura del árbol de decisión generado; en la 7.19 se muestra la importancia de las características predictoras, y, finalmente, la 7.20 indica las métricas de evaluación del desempeño del modelo generado.

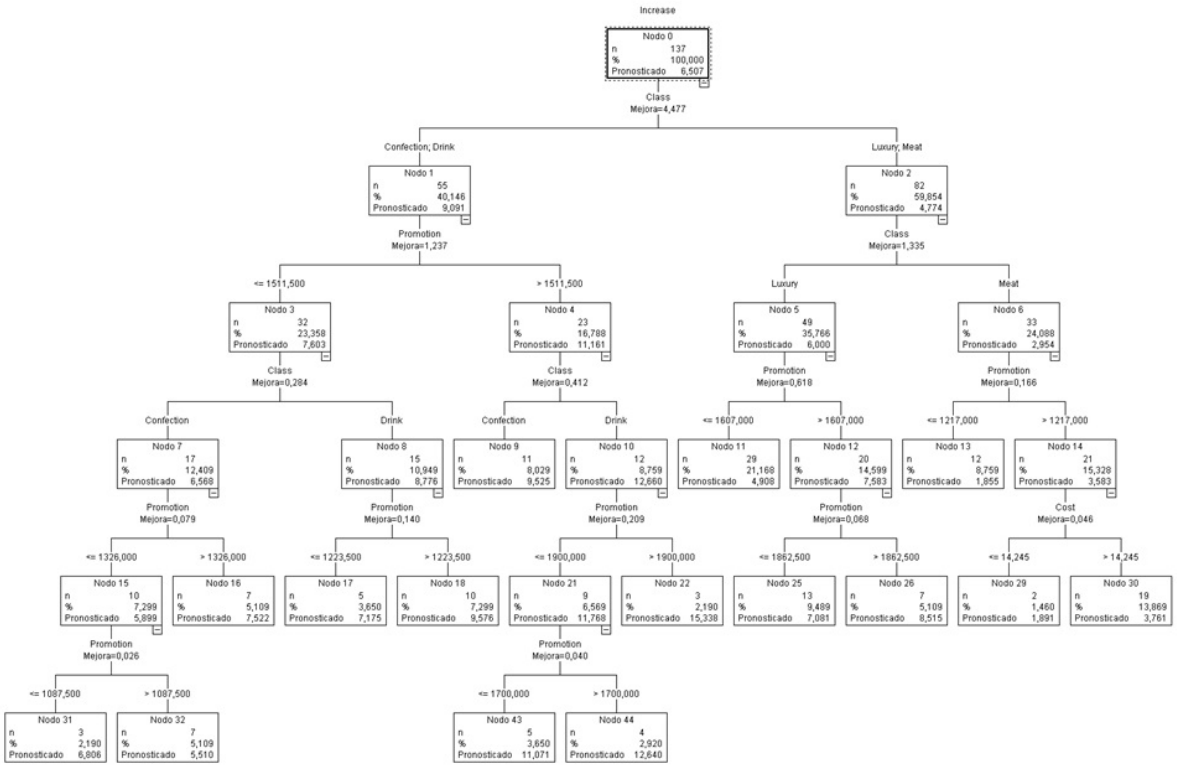


Figura 7.18. Creación del modelo de árbol de decisión C&R.

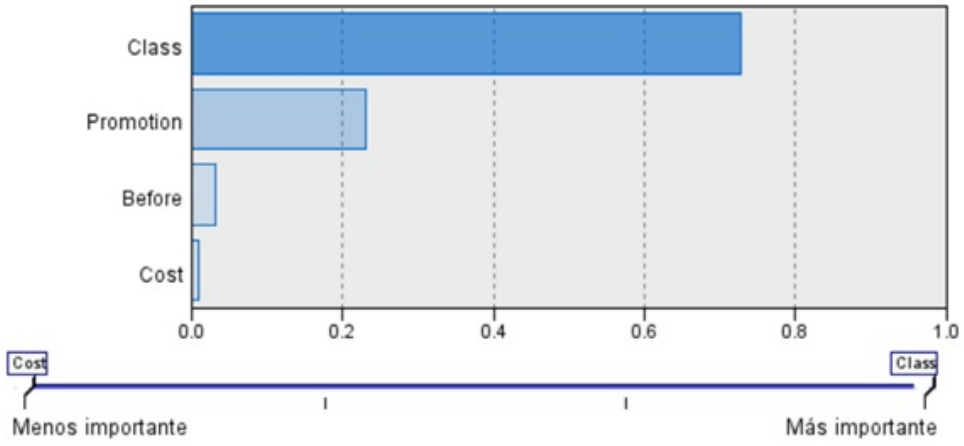


Figura 7.19. Importancia del predictor en la creación del modelo de árbol de decisión C&R.

Resultados para el campo de resultado Increase

Comparando $R-Increase con Increase

Error mínimo	-2.535
Error máximo	3.388
Error promedio	0.024
Error absoluto promedio	0.967
Desviación estándar	1.176
Correlación lineal	0.934
Ocurrencias	200

Figura 7.20. Métricas de evaluación del desempeño del modelo de árbol de decisión C&R.

Las figuras de la 7.21 a la 7.23 muestran la creación del modelo de red neuronal MLP. En la figura 7.21 se puede apreciar el modelo de red neuronal MLP generado; en la 7.22 se muestra la importancia de las características predictoras, y finalmente, la 7.23 indica las métricas de evaluación del desempeño del modelo generado.

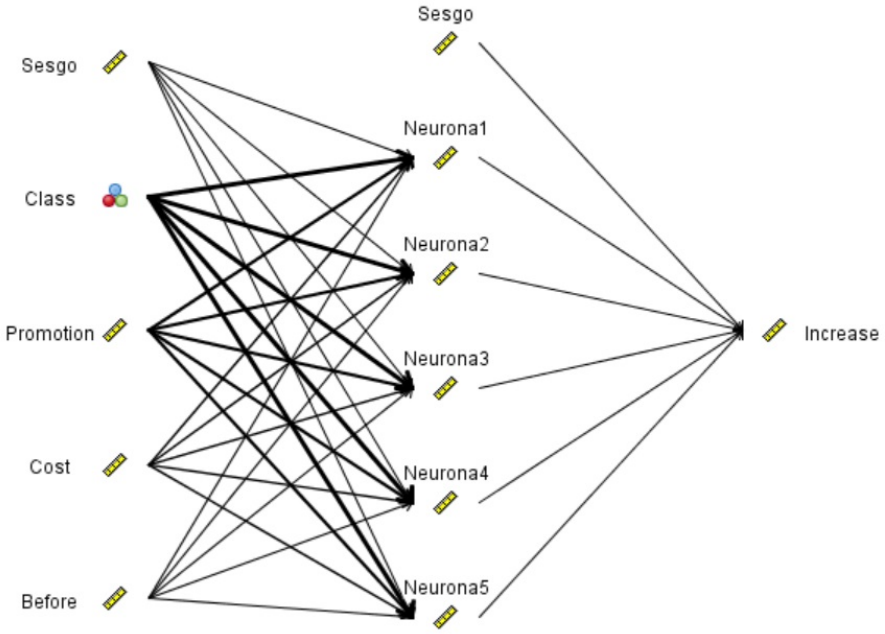


Figura 7.21. Creación del modelo de red neuronal MLP.

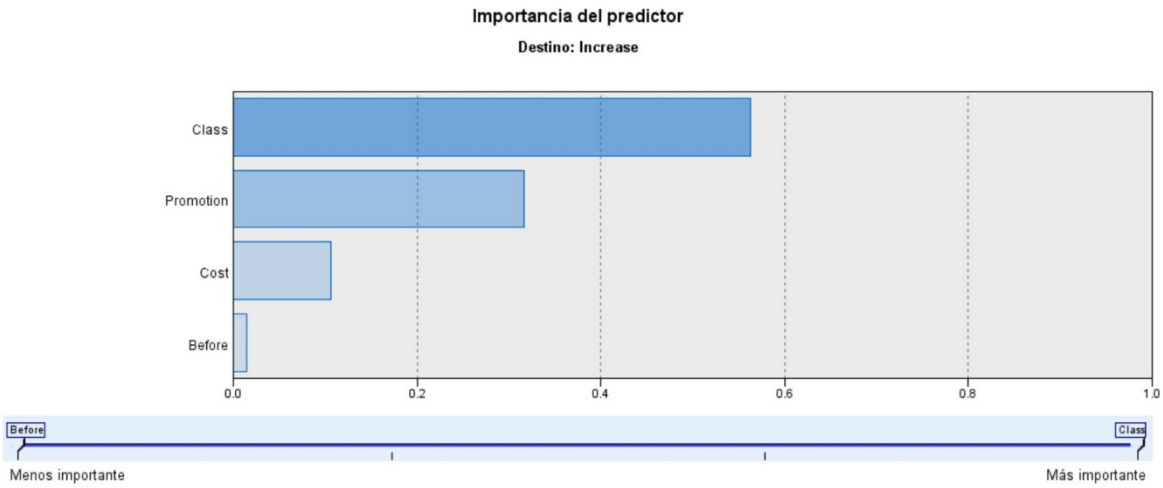


Figura 7.22. Importancia del predictor en la creación del modelo de red neuronal MLP.

Resultados para el campo de resultado Increase

Comparando $N-Increase con Increase

Error mínimo	-2.686
Error máximo	2.808
Error promedio	-0.015
Error absoluto promedio	1.087
Desviación estándar	1.299
Correlación lineal	0.919
Ocurrencias	200

Figura 7.23. Métricas de evaluación del desempeño del modelo de red neuronal MLP.

La creación del modelo de máquina de vectores de soporte (SVM, por sus siglas en inglés) se ilustra en las figuras 7.24 y 7.25. En la 7.24 se muestra la importancia de las características predictoras, mientras que la 7.25 indica las métricas de evaluación del desempeño del modelo generado.

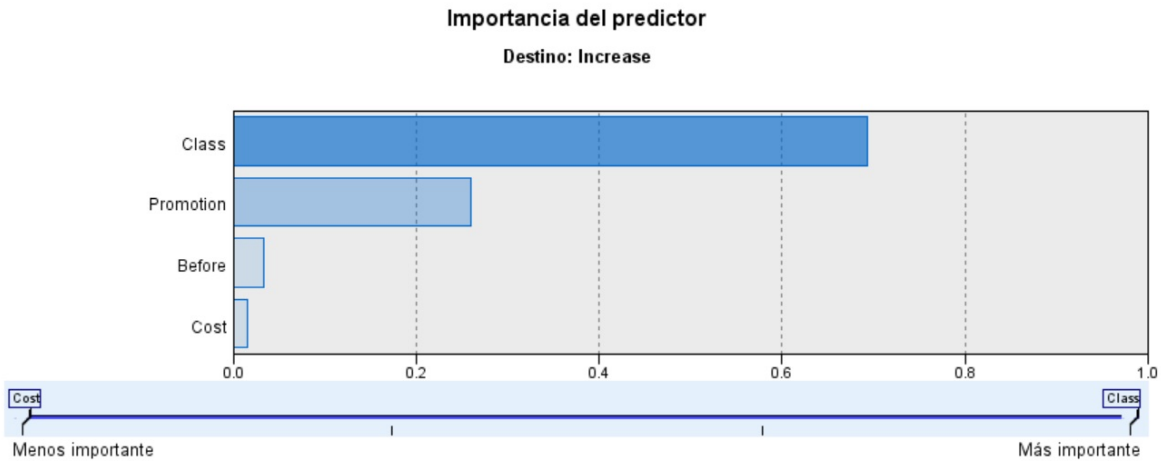


Figura 7.24. Importancia del predictor en la creación del modelo de máquina de vectores de soporte (SVM).

Resultados para el campo de resultado Increase

Comparando $S-Increase con Increase

Error mínimo	-2.875
Error máximo	2.625
Error promedio	-0.114
Error absoluto promedio	1.04
Desviación estándar	1.273
Correlación lineal	0.923
Ocurrencias	200

Figura 7.25. Métricas de evaluación del desempeño del modelo de máquina de vectores de soporte (SVM).

El último modelo generado se basa en el algoritmo de los K vecinos más cercanos (K-NN), cuya generación se ilustra en las figuras 7.26 y 7.27. En la 7.26 se aprecia el modelo K-NN generado, y la 7.27 indica las métricas de evaluación del desempeño del modelo.

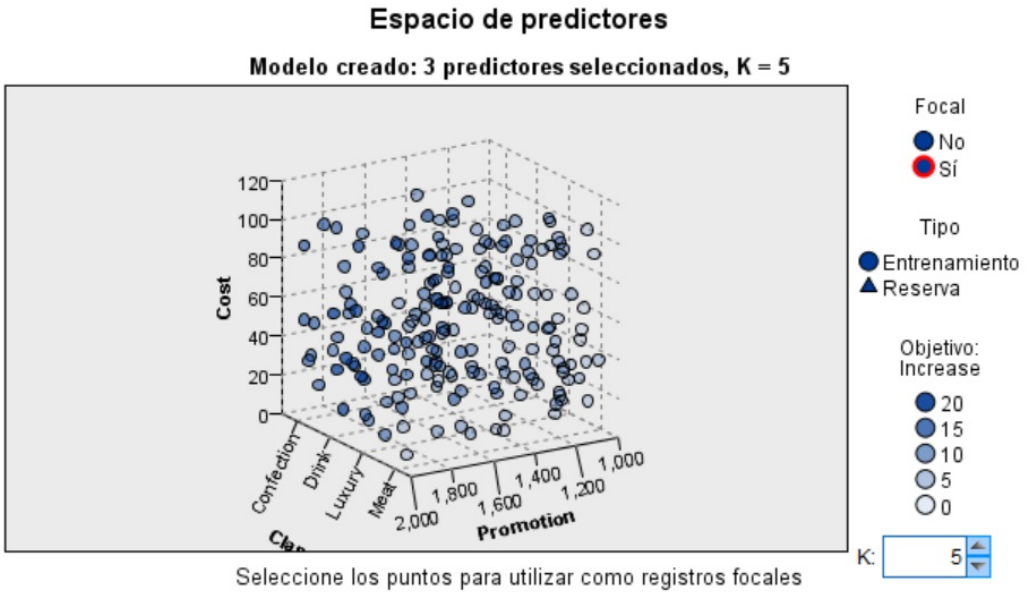


Figura 7.26. Creación del modelo basado en el algoritmo de los K vecinos más cercanos.

Resultados para el campo de resultado Increase

Comparando $KNN-Increase con Increase

Error mínimo	-2.794
Error máximo	2.928
Error promedio	0.055
Error absoluto promedio	1.024
Desviación estándar	1.199
Correlación lineal	0.931
Ocurrencias	200

Figura 7.27. Métricas de evaluación del desempeño del modelo basado en el algoritmo de los K vecinos más cercanos.

7.1.4.4. Evaluación de los modelos

En la tabla 7.2 se relacionan las métricas de evaluación del desempeño de los cinco modelos generados:

- Regresión lineal (RL)
- Árbol de decisión C&R
- Red neuronal MLP
- Máquina de vectores de soporte (SVM)
- Algoritmo KNN

Tabla 7.2. Métricas del desempeño de los modelos de regresión generados

	RL	C&R	MLP	SVM	K-NN
Error mínimo	-3.107	-2.535	-2.686	-2.875	-2.794
Error máximo	2.932	3.388	2.808	2.625	2.928
Error promedio	-0.0	0.024	-0.015	-0.114	0.055
Error absoluto promedio	1.118	0.967	1.087	1.04	1.024
Desviación estándar	1.34	1.176	1.299	1.273	1.199
Correlación lineal	0.913	0.934	0.919	0.923	0.931

Para la evaluación de los cinco modelos de predicción generados no sólo se tomó en consideración las métricas de desempeño relacionadas en la tabla 7.1, sino también otros aspectos importantes, entre los que destacan:

- Facilidad de interpretación de los resultados producidos por el modelo
- Identificación de los campos predictores más importantes

Ante los resultados obtenidos, los modelos de árbol de decisión C&R y de red neuronal MLP fueron seleccionados.

7.1.5. Evaluación

Como su nombre lo indica, el objetivo de esta fase es la evaluación de los resultados producidos tanto por el análisis exploratorio de los datos como por los modelos generados, para identificar mejoras que permitan, en el caso de estudio propuesto, que la promoción aplicada resulte mucho más atractiva en los tipos de bienes de consumo donde no se obtuvieron los resultados previstos.

A partir de los resultados producidos por los gráficos y modelos seleccionados (árbol de decisión C&R y red neuronal MLP) se efectuaron las siguientes recomendaciones, las cuales fueron discutidas y aceptadas por la directiva de la organización:

- Diseñar promociones mucho más atractivas para el consumidor en aquellos bienes de consumo donde no se obtuvieron los resultados previstos, como es el caso de artículos de lujo y productos cárnicos.
- Utilizar los modelos seleccionados para predecir el incremento que se observará en las ventas de un tipo particular de bien de consumo, a partir de los datos de *class*, *cost*, *promotion* y *before*.

7.1.6. Despliegue

Considerando los objetivos comerciales del proyecto y los resultados

proporcionados por el proyecto de minería de datos, la directiva de la compañía tomó las siguientes decisiones:

- Efectuar mejoras para que la promoción resulte mucho más atractiva en aquellos tipos de bienes de consumo donde no se obtuvieron los resultados previstos.
- Efectuar variaciones en el monto de la promoción, de acuerdo con los resultados obtenidos para cada tipo de bien de consumo.
- Utilizar los modelos seleccionados para la predicción del incremento en las ventas de bienes de consumo, con base en la categoría de bien de consumo y la promoción aplicada.

A continuación, se ejemplifica el uso, por parte de la directiva de la compañía, de los modelos generados, para la predicción del incremento en las ventas de determinados bienes de consumo (ver tabla 7.3), con base en la categoría del bien y la promoción aplicada. Las figuras de la 7.28 a la 7.32 ilustran el uso de los dos modelos generados en la predicción del incremento en las ventas para nuevos datos.

Tabla 7.3. Nuevos datos para predecir el incremento en las ventas

Class	***Cost***	***Promotion***	***Before***
Confection	24.5	1467	117400
Drink	85	1745	132262
Luxury	90	1426	148458
Meat	75	1486	159862

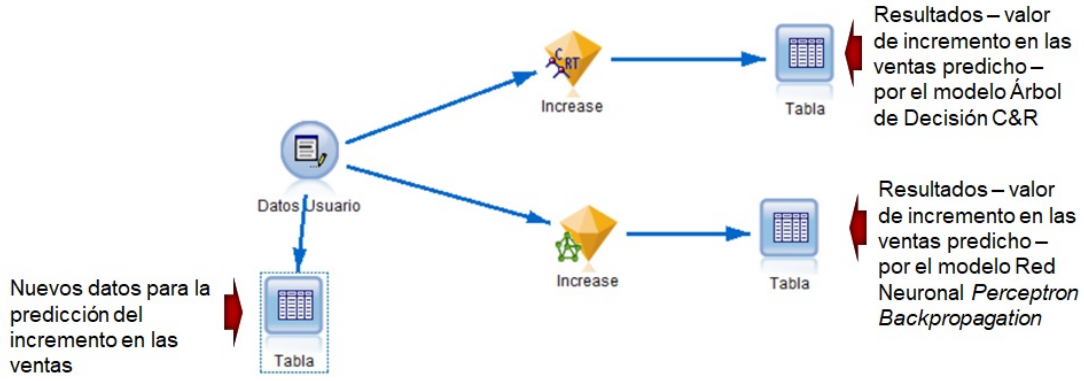


Figura 7.28. Uso de los modelos generados para la predicción del incremento en las ventas de nuevos datos.

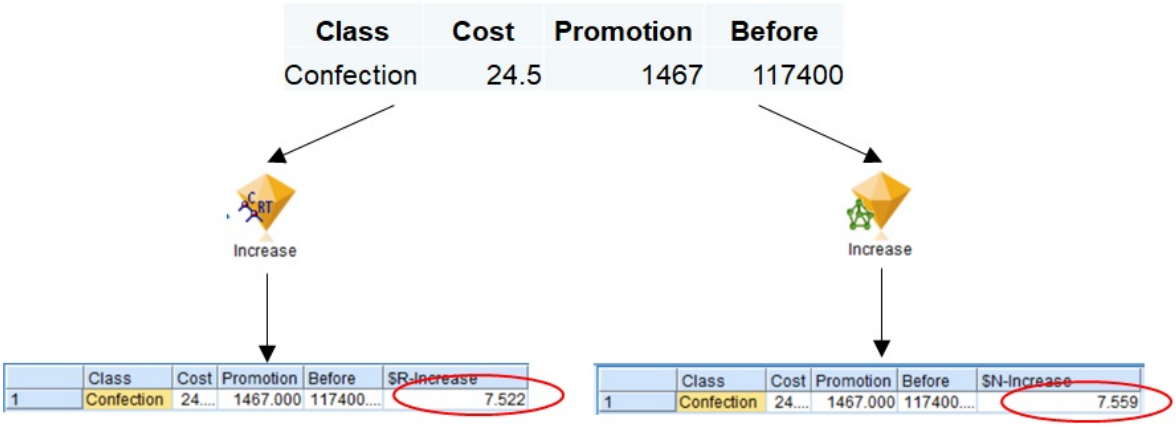


Figura 7.29. Predicción del incremento en las ventas de un nuevo bien de consumo de tipo *confection*.

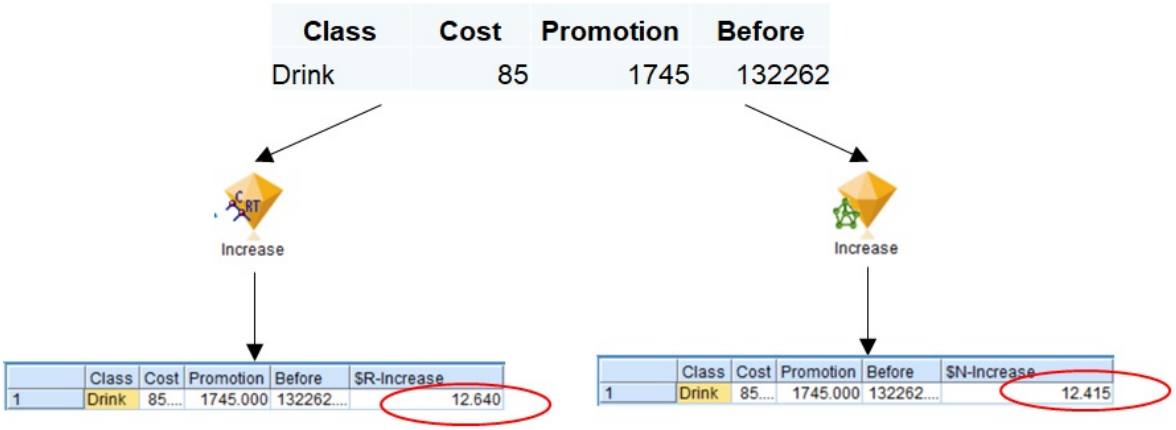


Figura 7.30. Predicción del incremento en las ventas de un nuevo bien de consumo de tipo *drink*.

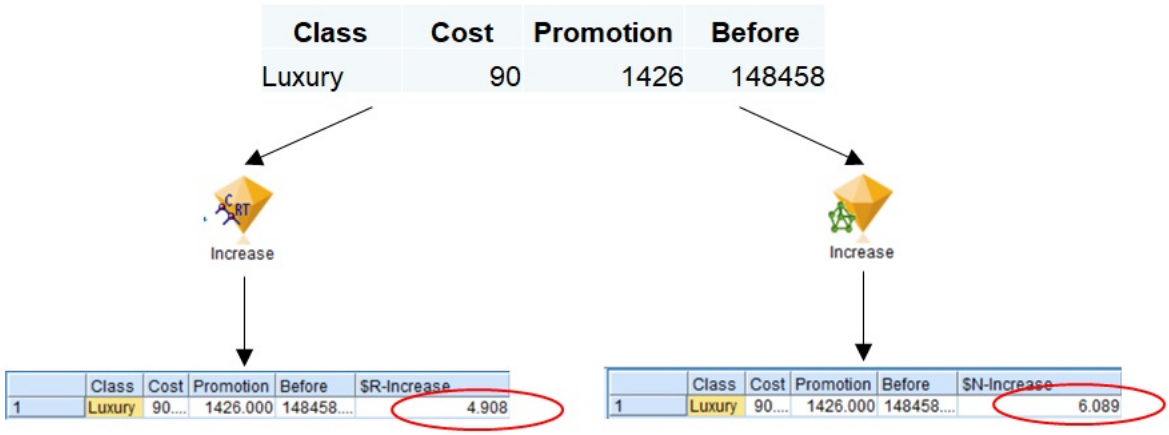


Figura 7.31. Predicción del incremento en las ventas de un nuevo bien de consumo de tipo *luxury*.

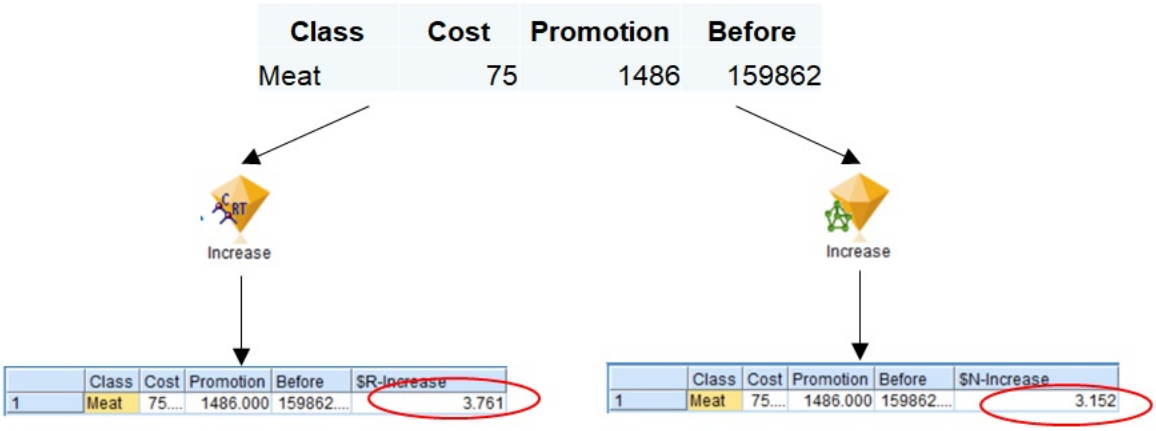


Figura 7.32. Predicción del incremento en las ventas de un nuevo bien de consumo de tipo *meat*.

7.2. Caso de estudio 2: Desgaste del cliente de tarjetas de crédito

7.2.1. Comprensión del dominio del problema

7.2.1.1. Determinación de los objetivos del proyecto

Los bancos, al ofrecer tarjetas de crédito, comúnmente se enfrentan al problema del desgaste del cliente, el cual se refiere a cómo la relación entre un cliente y su tarjeta de crédito se deteriora gradualmente, hasta que deja de utilizarla. Esto se manifiesta de diferentes formas y puede originarse por diversas causas.

El objetivo del presente proyecto consiste en identificar las principales causas que conllevan al desgaste del cliente de tarjetas de crédito, a partir de la exploración, preparación y modelado de un conjunto de datos que los clasifica como clientes activos y con desgaste. Lo anterior, con la finalidad de que el banco pueda, de forma proactiva, identificar o predecir a aquellos clientes con riesgo de desgaste o abandono, y ofrecerles un plan de crédito atractivo y mejorado, revirtiendo así la decisión de abandonar el uso de su tarjeta.

7.2.1.2. Valoración de la situación actual del objetivo del proyecto

Entre los tipos de créditos más comunes que ofrecen las instituciones dedicadas a esta labor se encuentran:

- Crédito hipotecario
- Crédito automotriz
- Crédito al consumo

El crédito al consumo, también conocido como crédito personal, se otorga para financiar una variedad de gastos, como la compra de electrodomésticos, muebles, viajes, pago de colegiatura de estudios, entre otros. La forma de ejercerlo, generalmente, es a través del otorgamiento y uso de la tarjeta de crédito.

Entre los principales factores que influyen en el desgaste del cliente de tarjetas de crédito se identifican los siguientes:

- Tasas de interés elevadas
- Cargos adicionales que se reflejan en el estado de cuenta
- Cambios en las condiciones personales y laborales del cliente
- Cambios en las condiciones financieras del cliente
- Incremento de los gastos fijos del cliente (hipoteca, auto, etcétera)

- Experiencias negativas por parte del servicio al cliente
- Ofertas mucho más atractivas provenientes de otras instituciones bancarias y de crédito
- Problemas de seguridad relacionados con los fraudes, extorsiones o robos de información financiera

Las tarjetas de crédito se caracterizan por pertenecer a un determinado nivel, el cual ofrece ciertos beneficios adicionales, ventajas o recompensas al titular, de acuerdo con sus características, como son el nivel de gastos y el de ingresos o solvencia económica. Los niveles de tarjetas que generalmente otorgan las instituciones bancarias y de créditos son (figura 7.33):

- Tarjeta de crédito estándar
- Tarjeta de crédito oro
- Tarjeta de crédito platino
- Tarjeta de crédito *black*

Figura 7.33. Niveles de tarjetas de crédito que comúnmente ofrecen las instituciones bancarias y crediticias.

Los beneficios, ventajas y recompensas que ofrecen los diferentes niveles de tarjeta de crédito varían de forma creciente: van del estándar —que es el nivel básico de tarjetas de crédito y ofrece muy pocos beneficios— hasta el *black* —el nivel máximo, reservado para clientes con elevada solvencia económica e impecable historial crediticio—.

El uso de las tarjetas de crédito no significa necesariamente un incremento en la capacidad de gastos del cliente, ya que su uso siempre conlleva a diferentes riesgos que los clientes deben considerar. Entre ellos se encuentran los siguientes:

- Pago de tasas de interés crecientes
- Pago de comisiones adicionales, no previstas o no informadas
- Cargos moratorios por impuntualidad o falta de pagos
- Endeudamiento excesivo
- Dependencia financiera, al utilizar la tarjeta de crédito para gastos básicos
- Problemas de seguridad, tales como fraudes o robo de identidad, al utilizar la tarjeta de crédito en sitios web o establecimientos pocos seguros
- Impacto negativo en el historial crediticio

Para reducir estos riesgos, los clientes de tarjetas de crédito deben usarla de manera responsable, realizar sus pagos a tiempo y evitar endeudarse con más de lo que puedan pagar. Por otra parte, para contrarrestar el problema del desgaste del cliente de tarjetas de crédito, las instituciones bancarias y de crédito deben esforzarse por mejorar la transparencia en sus políticas, brindar un servicio al cliente eficiente, mantener tasas de interés competitivas, ofrecer un servicio de atención al cliente de calidad y adaptarse a las necesidades cambiantes de sus usuarios.

Para efectuar el análisis correspondiente de los objetivos del proyecto se cuenta con un voluminoso conjunto de datos de acceso público, cuyas características serán descritas en la siguiente fase de la metodología CRISP-DM: comprensión de los datos.

7.2.1.3. Determinación de los objetivos de minería de datos

En este proyecto se identifican dos objetivos de minería de datos:

- **Predicción a través de clasificación**: Predecir si un cliente abandonará los servicios de la tarjeta de crédito, de forma que la institución bancaria o crediticia pueda actuar anticipadamente y tratar de revertir la decisión del usuario.
- **Clasificación**: Asignar o promocionar un nivel superior de tarjeta de crédito, a partir de los datos demográficos y financieros del cliente.

7.2.1.4. Propuesta del enfoque metodológico

El enfoque metodológico que guiará este proyecto de minería de datos, relacionado con el desgaste del cliente de crédito, se muestra en la figura 7.34.

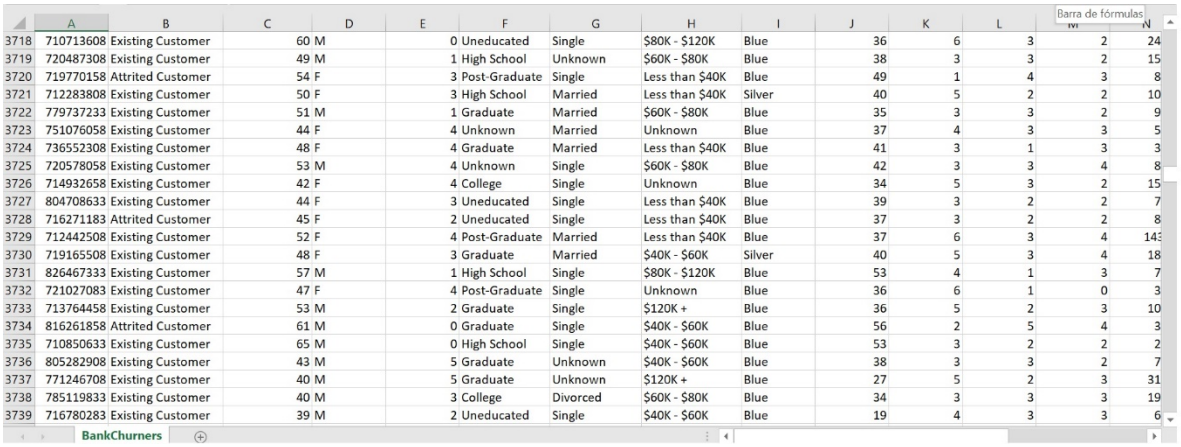
Fase	Tiempo a dedicar	Recursos humanos y tecnológicos	Riesgos atribuibles
1. Comprensión del dominio del problema	2 semana	Experto en crédito al consumo, experto en minería de datos	Curva de aprendizaje en el dominio del problema
2. Comprensión de los datos	3 semana	Experto en el dominio del problema, experto en minería de datos. Tablas, gráficos y resúmenes estadísticos que faciliten la comprensión de los datos	Curva de aprendizaje en el dominio del problema
3. Preparación de los datos	4 semanas	Experto en mercadotecnia, experto en minería de datos Herramientas para el análisis exploratorio de datos Paquete *IBM SPSS Modeler* Herramienta *IDA-WEB TOOL*	No se han identificado riesgos
4. Modelado	3 semanas	Experto en minería de datos y experto en técnicas de *machine learning* Herramientas para la implementación de modelos de *machine learning* Paquete *IBM SPSS Modeler* Herramienta *IDA-WEB TOOL*	No se han identificado riesgos
5. Evaluación	1 semana	Experto en crédito al consumo, experto en minería de datos Paquete *IBM SPSS Modeler* Herramienta *IDA-WEB TOOL*	No se han identificado riesgos
6. Presentación	1 semana	Experto en crédito al consumo, experto en minería de datos, directiva de la organización Paquete *IBM SPSS Modeler* Herramienta *IDA-WEB TOOL*	No se han identificado riesgos

Figura 7.34. Enfoque metodológico.

7.2.2. Comprensión de los datos

7.2.2.1. Recopilación de los datos iniciales

El conjunto de datos inicial de este proyecto es de acceso público y se encuentra disponible en: https://www.kaggle.com/sakshigoyal7/credit-card-customers. El conjunto “Desgaste del cliente de crédito” está compuesto por 21 atributos y 10,000 registros. La figura 7.35 muestra un fragmento del mismo.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
3718	710713608	Existing Customer	60	M	0	Uneducated	Single	$80K - $120K	Blue	36	6	3	2	24
3719	720487308	Existing Customer	49	M	1	High School	Unknown	$60K - $80K	Blue	38	3	3	2	15
3720	719770158	Attrited Customer	54	F	3	Post-Graduate	Single	Less than $40K	Blue	49	1	4	3	8
3721	712283808	Existing Customer	50	F	3	High School	Married	Less than $40K	Silver	40	5	2	2	10
3722	779737233	Existing Customer	51	M	1	Graduate	Married	$60K - $80K	Blue	35	3	3	2	9
3723	751076058	Existing Customer	44	F	4	Unknown	Married	Unknown	Blue	37	4	3	3	5
3724	736552308	Existing Customer	48	F	4	Graduate	Married	Less than $40K	Blue	41	3	1	3	3
3725	720578058	Existing Customer	53	M	4	Unknown	Single	$60K - $80K	Blue	42	3	3	4	8
3726	714932658	Existing Customer	42	F	4	College	Single	Unknown	Blue	34	5	3	2	15
3727	804708633	Existing Customer	44	F	3	Uneducated	Single	Less than $40K	Blue	39	3	2	2	7
3728	716271183	Attrited Customer	45	F	2	Uneducated	Single	Less than $40K	Blue	37	3	2	2	8
3729	712442508	Existing Customer	52	F	4	Post-Graduate	Married	Less than $40K	Blue	37	6	3	4	143
3730	719165508	Existing Customer	48	F	3	Graduate	Married	$40K - $60K	Silver	40	5	3	4	18
3731	826467333	Existing Customer	57	M	1	High School	Single	$80K - $120K	Blue	53	4	1	3	7
3732	721027083	Existing Customer	47	F	4	Post-Graduate	Single	Unknown	Blue	36	6	1	0	3
3733	713764458	Existing Customer	53	M	2	Graduate	Single	$120K +	Blue	36	5	2	3	10
3734	816261858	Attrited Customer	61	M	0	Graduate	Single	$40K - $60K	Blue	56	2	5	4	3
3735	710850633	Existing Customer	65	M	0	High School	Single	$40K - $60K	Blue	53	3	2	2	2
3736	805282908	Existing Customer	43	M	5	Graduate	Unknown	$40K - $60K	Blue	38	3	3	2	7
3737	771246708	Existing Customer	40	M	5	Graduate	Unknown	$120K +	Blue	27	5	2	3	31
3738	785119833	Existing Customer	40	M	3	College	Divorced	$60K - $80K	Blue	34	3	3	3	19
3739	716780283	Existing Customer	39	M	2	Uneducated	Single	$40K - $60K	Blue	19	4	3	3	6

Figura 7.35. Fragmento ilustrativo del conjunto de datos relacionados con el desgaste del cliente de crédito.

7.2.2.2. Descripción de los datos

La tabla 7.4 muestra la relación de los 21 atributos del conjunto de datos “Desgaste del cliente de crédito”, indicando su nombre y descripción. De forma complementaria, en la tabla 7.5 se indica el tipo de dato de cada atributo (continuo, categórico, ordinal, nominal, etcétera) y si el atributo es predictor (entrada), objetivo (salida) o ambos (entrada/salida).

Tabla 7.4. Nombre y descripción de los atributos que conforman el conjunto de datos “Desgaste del cliente de crédito”

Atributo	Descripción
ID cliente	Número de identificación de cada cliente
Estado del cliente	Estado de actividad del cliente: “existente”, si se encuentra activo, o “desgaste”, si ha abandonado el uso de la tarjeta de crédito
Edad	Edad en años del cliente
Género	Sexo del cliente
Estado marital	Estado civil del cliente: soltero, casado, divorciado, viudo o desconocido
Ingresos	Monto de ingresos del cliente expresado en rangos o en miles de dólares
Límite de tarjeta de crédito	Límite máximo del crédito, expresado en miles de dólares
Categoría de tarjeta de crédito	Nivel de la tarjeta de crédito: *blue*, *silver*, *gold*, o *platinum*
Período de relación con el banco	Tiempo de relación del cliente con el banco, expresado en meses
Número de dependientes	Cantidad de personas que dependen económicamente del titular de la tarjeta de crédito
Nivel de educación	Nivel de escolaridad del cliente, expresado como atributo de tipo ordinal
Total de productos en poder del cliente	Total de productos de crédito que posee el cliente, aquí se consideran todos los tipos de créditos o préstamos

Número de meses inactivo en los últimos 12 meses	Tiempo de inactividad del cliente en su cuenta durante los últimos 12 meses
Número de contactos en los últimos 12 meses	Número de veces que el cliente se puso en contacto directo con el banco
Saldo rotativo total en la tarjeta de crédito	Capital total adeudado en la tarjeta de crédito
Línea de crédito abierta para compras (promedio de los últimos 12 meses)	Monto total de dinero disponible para compras
Monto total de las transacciones	Monto total de dinero involucrado en las transacciones del cliente
Cambio en el monto de la transacción	Atributo derivado por el banco que representa un cambio en el monto total de transacciones
Recuento total de transacciones (últimos 12 meses)	Cantidad de transacciones efectuadas por el cliente en los últimos 12 meses
Cambio en el recuento de transacciones	Atributo derivado por el banco que representa un cambio en el recuento total de transacciones
Índice de utilización promedio de la tarjeta	Cantidad de crédito renovable que se está utilizando dividida entre el crédito total disponible

Tabla 7.5. Tipo de valor y papel (predictor/objetivo) de los atributos que conforman el conjunto de datos “Desgaste del cliente de crédito”

Atributo	Tipo de valor	Papel
ID cliente	Numérico	Predictor
Estado del cliente	Nominal, expresado como caracteres	Objetivo/predictor
Edad	Numérico	Predictor
Género	Nominal, expresados como caracteres	Predictor
Número de dependientes	Numérico	Predictor
Nivel de educación		Predictor
Estado marital	Nominal, expresado como caracteres	Predictor
Categoría de ingresos	Ordinal, expresado en rangos de ingresos	Predictor
Categoría de tarjeta de crédito	Ordinal, expresado como caracteres	Objetivo/predictor
Período de la relación con el banco	Numérico	Predictor
Total de productos de crédito en poder del cliente	Numérico	Predictor
Número de meses inactivo en los últimos 12 meses	Numérico	Predictor
Número de contactos en los últimos 12 meses	Numérico	Predictor
Límite de tarjeta de crédito	Numérico	Predictor
Saldo rotativo total en la tarjeta de crédito	Numérico	Predictor
Línea de crédito abierta para compras (promedio de los últimos 12 meses)	Numérico	Predictor
Cambio en el monto de la transacción	Numérico	Predictor
Monto total de las transacciones	Numérico	Predictor

Recuento total de transacciones (últimos 12 meses)	Numérico	Predictor
Cambio en el recuento de transacciones	Numérico	Predictor
Índice de utilización promedio de la tarjeta	Numérico	Predictor

7.2.2.3. Exploración de los datos

Las figuras de la 7.36 a la 7.47 muestran aspectos clave de la actividad de exploración de los datos, a través de gráficos que permiten analizar la importancia de los campos predictores. Nótese en las figuras 7.36 y 7.37 el nivel de desbalance del conjunto de datos respecto a la categoría "estado del cliente" (*Attrition_Flag*).

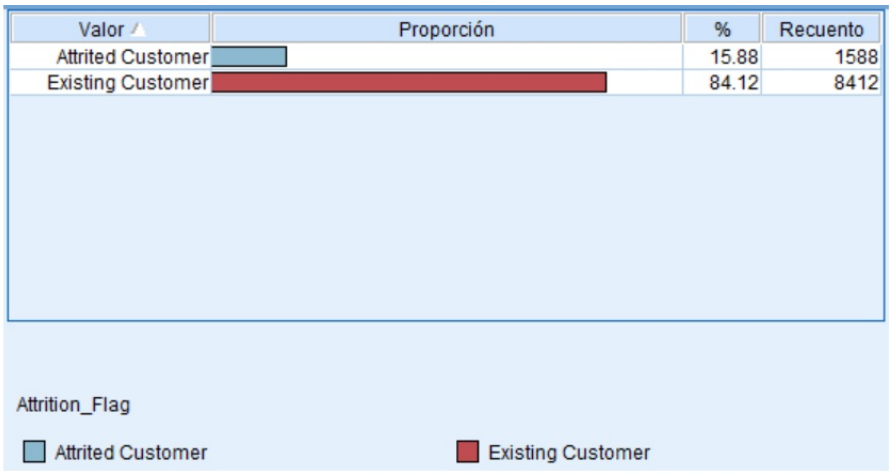


Figura 7.36. Exploración de los datos. Tabla de recuento para la categoría "estado del cliente" (*Attrition_Flag*).

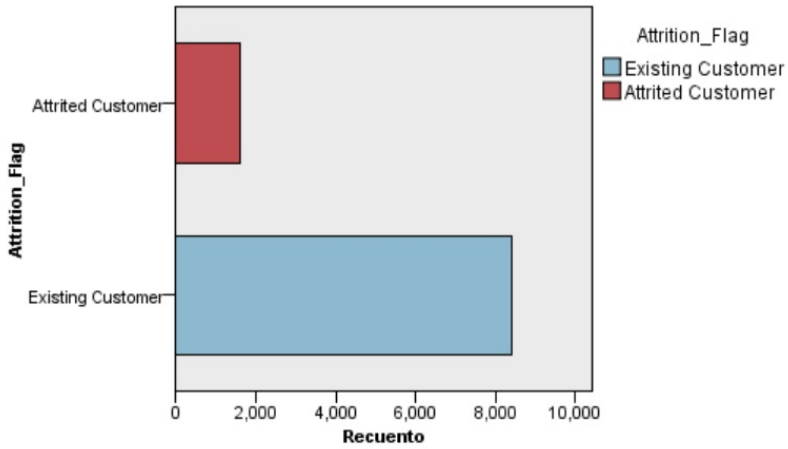


Figura 7.37. Exploración de los datos. Gráfico de recuento para la categoría "estado del cliente" (*Attrition_Flag*).

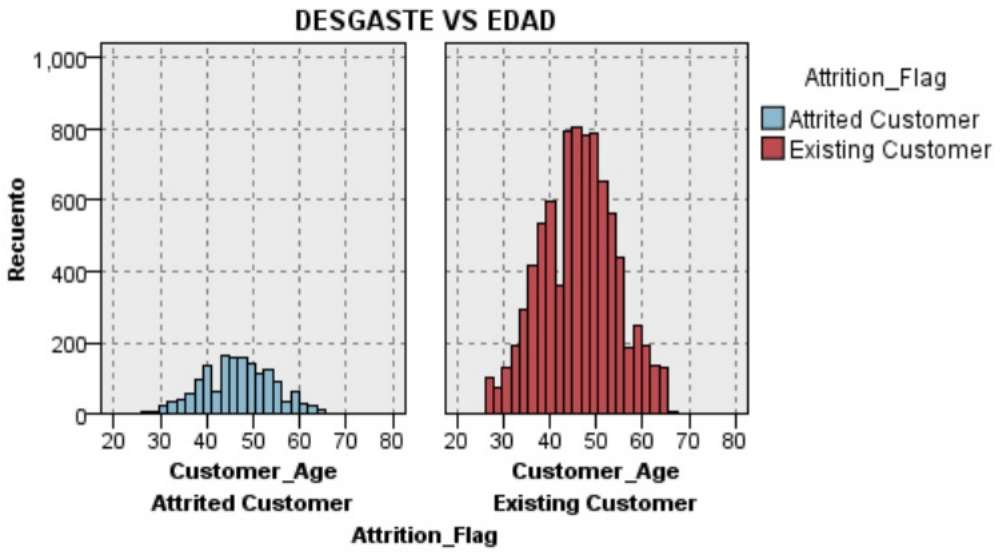


Figura 7.38. Exploración de los datos. Gráfico de recuento: desgaste del cliente *vs*. edad.

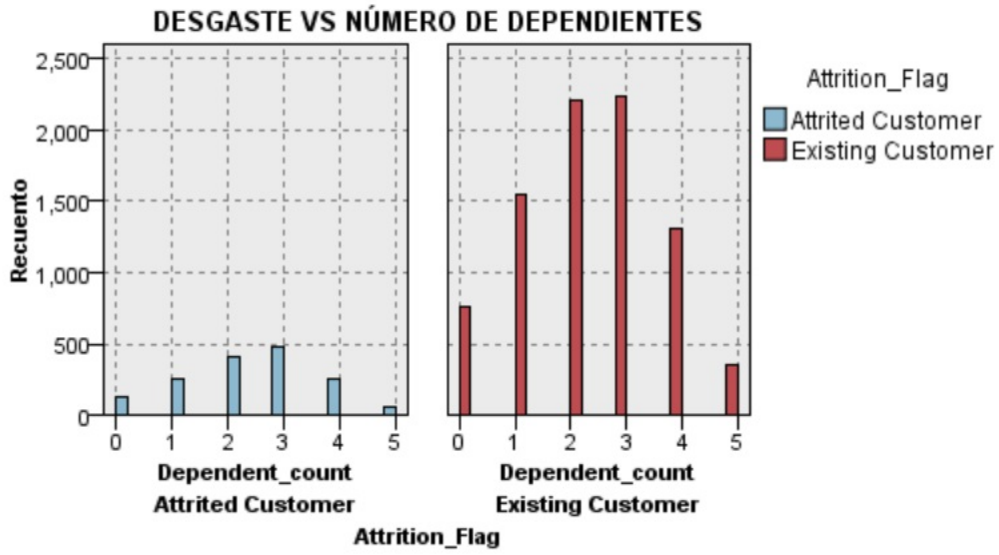


Figura 7.39. Exploración de los datos. Gráfico de recuento: desgaste del cliente *vs.* número de dependientes.

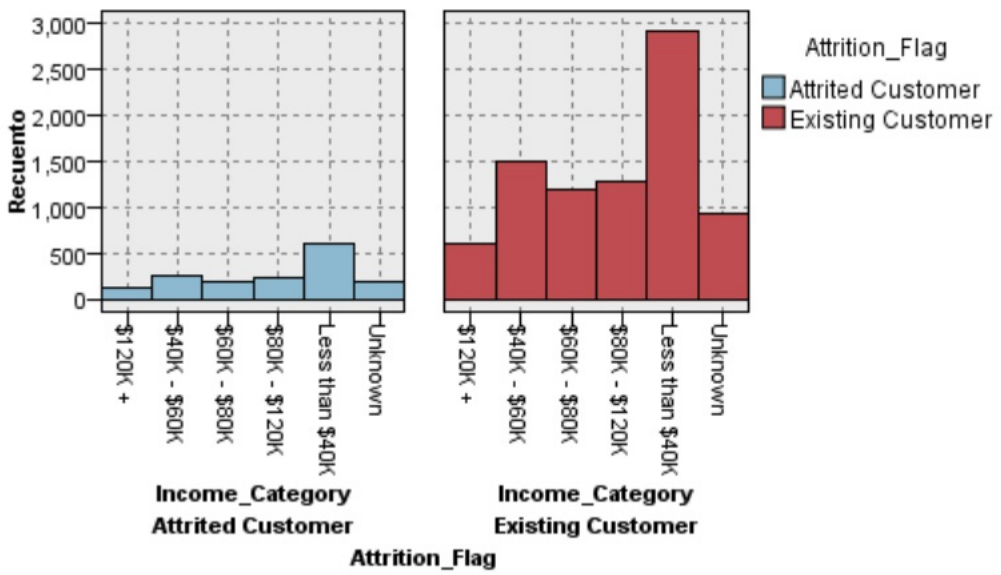


Figura 7.40. Exploración de los datos. Desgaste del cliente *vs*. categoría de ingresos.

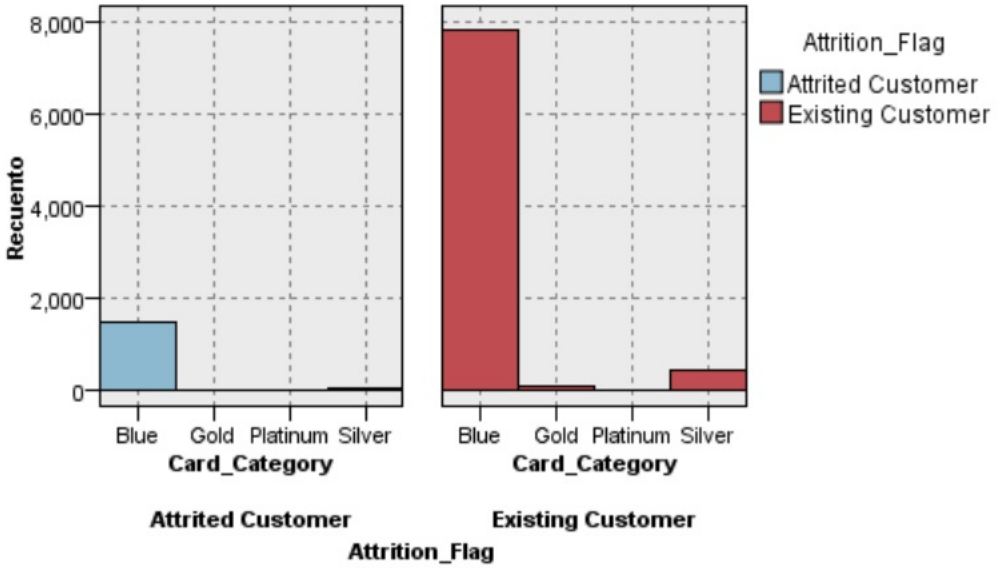


Figura 7.41. Exploración de los datos. Desgaste del cliente *vs*. nivel de la tarjeta de crédito.

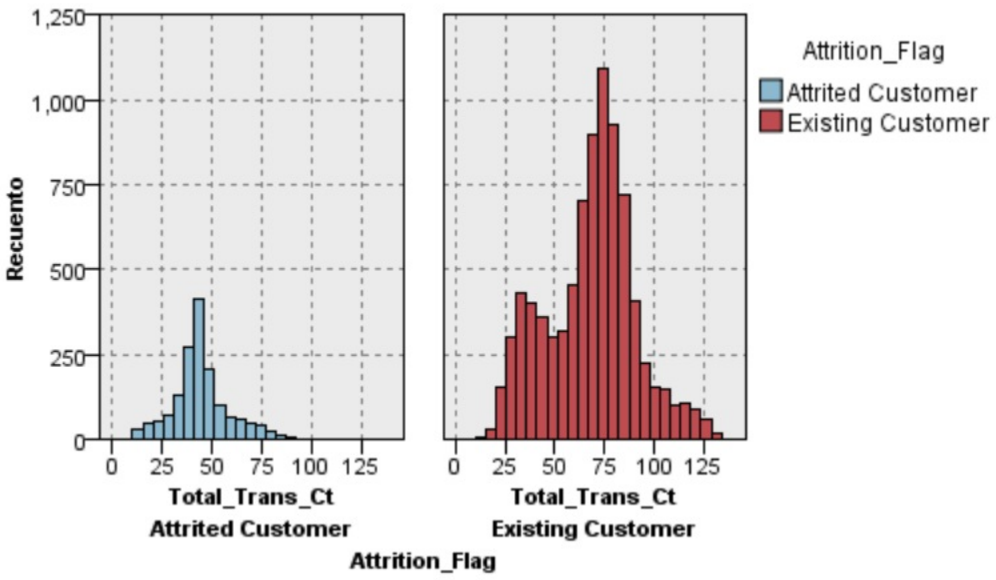


Figura 7.42. Exploración de los datos. Desgaste del cliente *vs*. número total de transacciones.

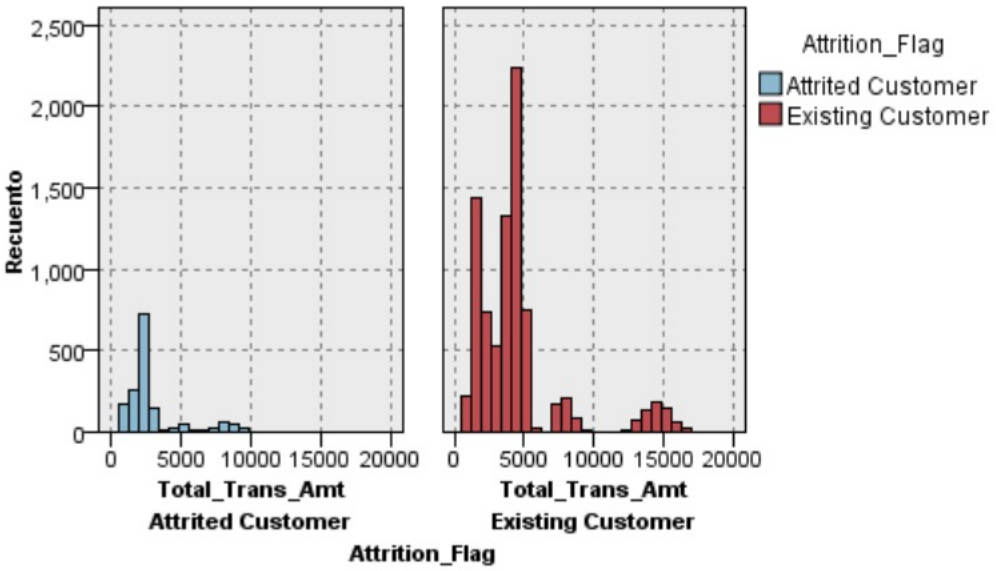


Figura 7.43. Exploración de los datos. Desgaste del cliente *vs*. monto total de transacciones.

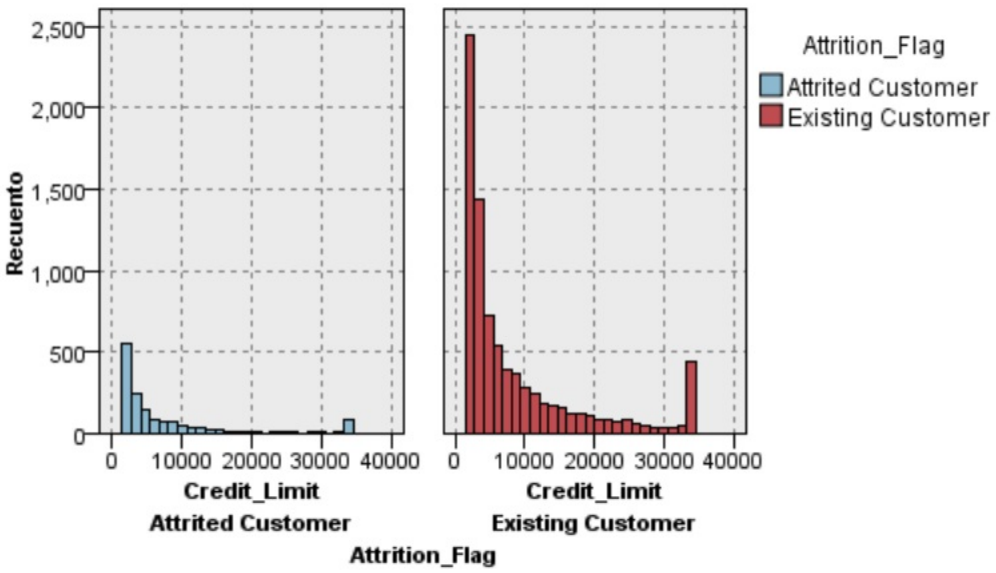


Figura 7.44. Exploración de los datos. Desgaste del cliente *vs*. límite de crédito.

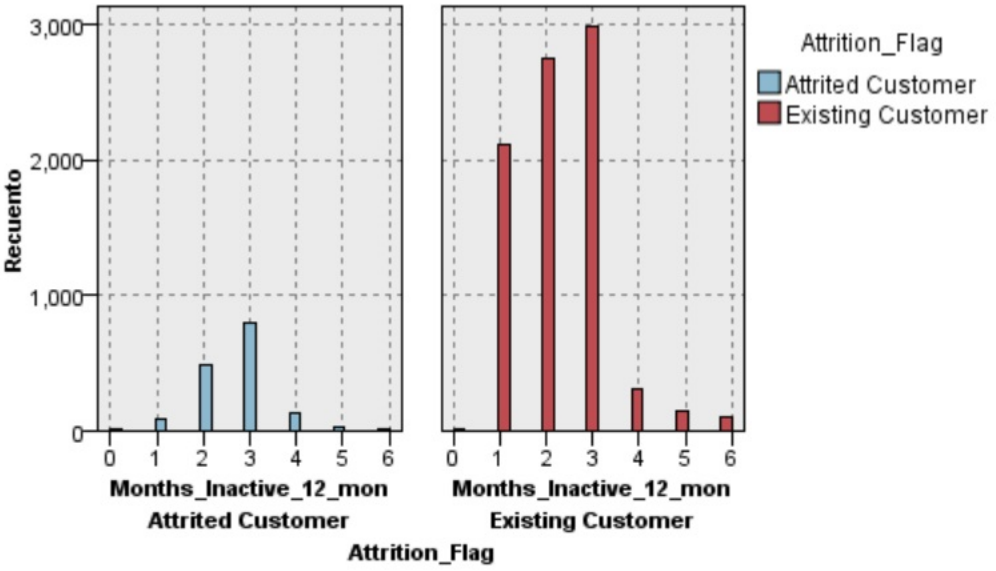


Figura 7.45. Exploración de los datos. Desgaste del cliente *vs*. número de meses inactivo en los últimos 12 meses.

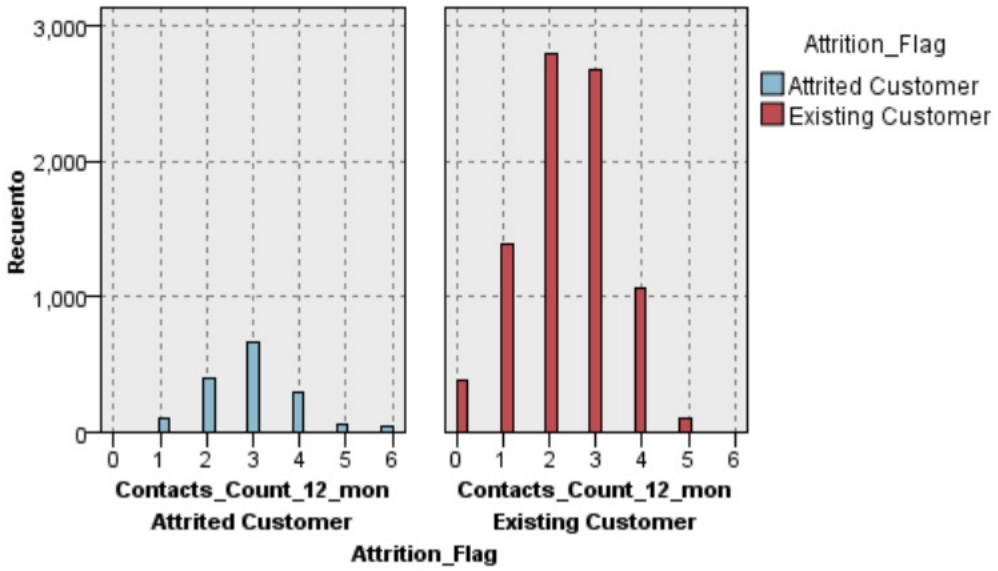


Figura 7.46. Exploración de los datos. Desgaste del cliente *vs*. número de contactos con el banco en los últimos 12 meses.

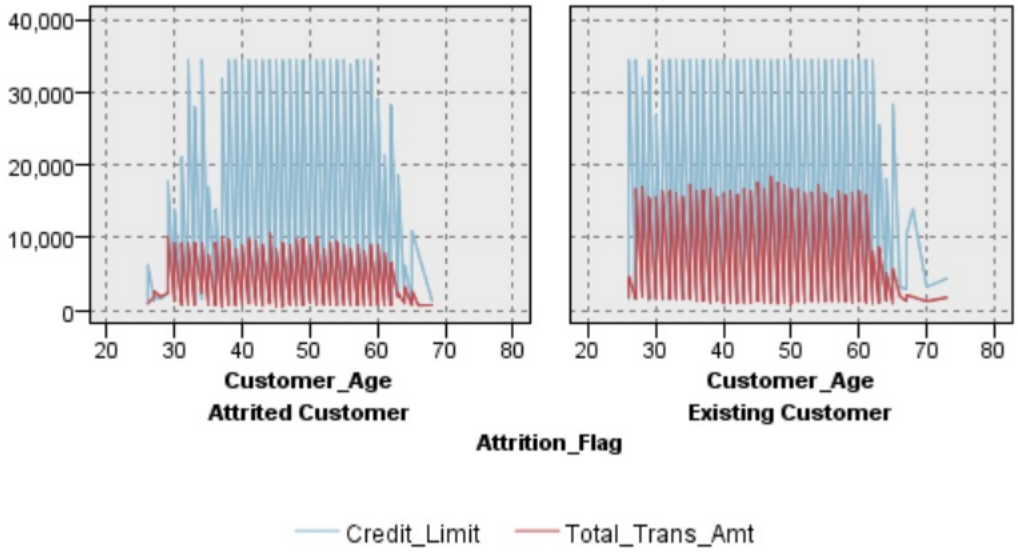


Figura 7.47. Exploración de los datos. Límite de crédito y monto total de las transferencias efectuadas por edad, para clientes con desgaste (panel de la izquierda) y para clientes activos (panel de la derecha).

7.2.2.4. Verificación de la calidad de los datos

De manera previa a la fase de preparación de los datos, se consideraron y se analizaron los siguientes aspectos:

- Datos perdidos o vacíos: No existen.

- Errores en los datos: Aparentemente no existen.
- Errores de medición: Aparentemente no existen.
- Errores de codificación: Es necesario efectuar la codificación de los siguientes atributos que aparecen como cadena de caracteres a tipo nominal u ordinal:
    - Estado del cliente
    - Género
    - Estado marital
    - Ingresos del cliente
    - Nivel de educación
    - Categoría de la tarjeta
- Atributos redundantes o de escasa utilidad: No se han identificado.
- Errores en la identificación de las posibles categorías o clases: No se han identificado.

La figura 7.48 muestra la verificación preliminar de la calidad de los datos, mientras que en la figura 7.49 se observa la verificación preliminar de la calidad de los atributos, a partir del análisis de componentes principales.

Campo	Gráfico de muest...	Medida	Mín.	Máx.	Media	Desv. est.	Sesgo	Exclusivo	Válido
CLIENTNUM		Continuo	708082083...	828343083.000	739273048.898	36957585.731	0.990	--	10000
Attrition_Flag		Categórico	--	--	--	--	--	2	10000
Customer_Age		Continuo	26.000	73.000	46.337	8.012	-0.030	--	10000
Gender		Categórico	--	--	--	--	--	2	10000
Dependent_count		Continuo	0.000	5.000	2.346	1.298	-0.024	--	10000
Education_Level		Categórico	--	--	--	--	--	7	10000
Marital_Status		Categórico	--	--	--	--	--	4	10000
Income_Category		Categórico	--	--	--	--	--	6	10000

Figura 7.48. Verificación de la calidad de los datos.

Comunalidades

	Inicial	Extracción
CLIENTNUM	1.000	.052
Customer_Age	1.000	.840
Dependent_count	1.000	.114
Months_on_book	1.000	.872
Total_Relationship_Count	1.000	.390
Months_Inactive_12_mon	1.000	.103
Contacts_Count_12_mon	1.000	.162
Credit_Limit	1.000	.990
Total_Revolving_Bal	1.000	.897
Avg_Open_To_Buy	1.000	.987
Total_Amt_Chng_Q4_Q1	1.000	.616
Total_Trans_Amt	1.000	.849
Total_Trans_Ct	1.000	.804
Total_Ct_Chng_Q4_Q1	1.000	.632
Avg_Utilization_Ratio	1.000	.879

Método de extracción: análisis de componentes principales.

Figura 7.49. Verificación de la calidad de los atributos. Análisis de componentes principales.

7.2.3. Preparación de los datos

7.2.3.1. Selección de datos

Como ya habíamos indicado, los objetivos de minería de datos de este proyecto son los siguientes:

- **Número 1. Predicción a través de clasificación:** Predecir si un cliente abandonará los servicios de tarjeta de crédito, de forma que la institución bancaria o crediticia pueda actuar anticipadamente y trate de revertir la previsible decisión del usuario.
- **Número 2. Clasificación:** Asignar o promocionar un nivel superior de tarjeta de crédito, a partir de los datos demográficos y financieros del cliente.

Para ambos objetivos se considerarán los 10,000 registros del conjunto de datos y los 21 atributos iniciales. Posteriormente se procederá a una selección de características o atributos tomando como base:

1. Los resultados del análisis de componentes principales (ver figura 7.49)

2. El orden de los predictores más importantes propuestos por el modelo de red neuronal MLP

7.2.3.2. Limpieza de datos

Considerando que aparentemente no existen datos perdidos o vacíos, errores en los datos, errores de medición, atributos redundantes o de escasa utilidad, o errores en la asignación de las clases o categorías, la limpieza de datos se enfocará en corregir los errores de codificación, previamente señalados, para lo cual será necesario codificar los siguientes atributos, mismos que aparecen como cadena de caracteres a tipo nominal u ordinal:

- Estado del cliente
- Género
- Estado marital
- Ingresos del cliente
- Nivel de educación
- Categoría de la tarjeta

La recodificación de los campos indicados se llevará a cabo utilizando el nodo “Rellenar” del menú “Operaciones con campos”, tal como se ilustra en la figura 7.50. En tanto, las figuras 7.51 y 7.52 muestran fragmentos del conjunto de datos “Desgaste del cliente de crédito” antes y después de la recodificación de campos.

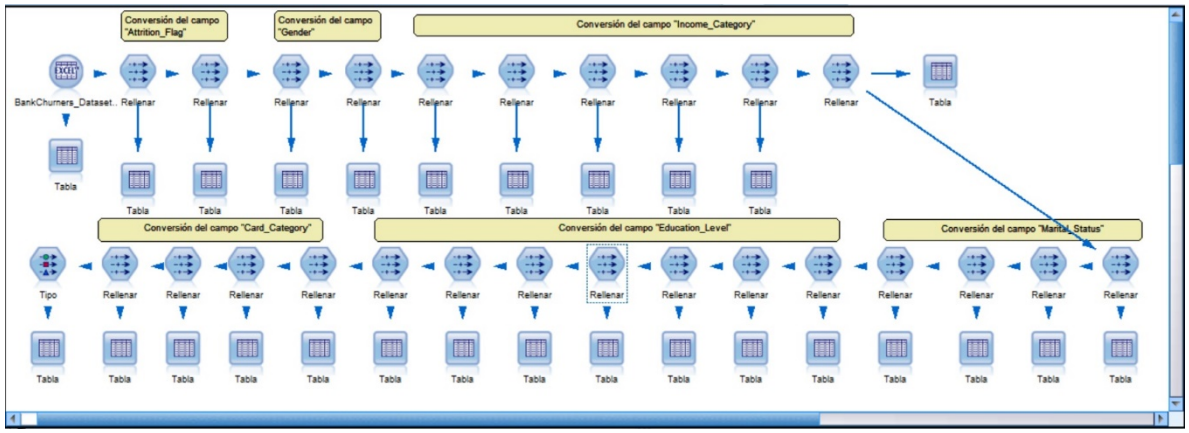


Figura 7.50. Conversión de campos utilizando el nodo “Rellenar” del menú “Operaciones con campos” de la herramienta de minería de datos IBM SPSS MODELER.

	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book	Total_Relationship_Count	Months_Inactive_12
2150	718777008.0...	Existing Customer	59.000	M	0.000	High School	Married	$80K - $120K	Blue	47.000	5.000	
2151	713352183.0...	Existing Customer	30.000	M	1.000	Post-Graduate	Married	Less than $40K	Blue	17.000	5.000	
2152	716128083.0...	Existing Customer	56.000	M	2.000	High School	Married	$40K - $60K	Blue	36.000	5.000	
2153	721286158.0...	Existing Customer	33.000	M	2.000	Unknown	Single	$120K +	Blue	15.000	4.000	
2154	757881258.0...	Existing Customer	31.000	M	0.000	Graduate	Divorced	$60K - $80K	Blue	24.000	6.000	
2155	715843533.0...	Existing Customer	26.000	M	1.000	College	Single	$40K - $60K	Blue	36.000	3.000	
2156	717234183.0...	Existing Customer	29.000	M	1.000	Graduate	Married	Less than $40K	Blue	18.000	4.000	
2157	717260058.0...	Existing Customer	53.000	F	2.000	Uneducated	Married	Less than $40K	Blue	42.000	6.000	
2158	709710783.0...	Existing Customer	35.000	F	0.000	High School	Single	Less than $40K	Blue	36.000	4.000	
2159	787284858.0...	Existing Customer	35.000	M	2.000	Uneducated	Unknown	$120K +	Blue	29.000	5.000	
2160	815845758.0...	Existing Customer	56.000	F	4.000	High School	Married	Less than $40K	Blue	51.000	4.000	
2161	720639783.0...	Existing Customer	35.000	F	1.000	Uneducated	Single	Unknown	Silver	27.000	4.000	
2162	713331408.0...	Existing Customer	35.000	M	3.000	Uneducated	Married	$40K - $60K	Blue	36.000	6.000	
2163	778693458.0...	Attrited Customer	32.000	M	1.000	High School	Divorced	$60K - $80K	Blue	13.000	3.000	
2164	788610258.0...	Existing Customer	49.000	M	4.000	College	Unknown	Less than $40K	Blue	43.000	5.000	
2165	711173883.0...	Existing Customer	36.000	M	4.000	Graduate	Divorced	$120K +	Blue	29.000	3.000	
2166	717628608.0...	Existing Customer	37.000	M	3.000	Unknown	Married	$60K - $80K	Blue	29.000	5.000	
2167	789261558.0...	Existing Customer	58.000	M	1.000	High School	Married	$60K - $80K	Blue	47.000	3.000	
2168	770647233.0...	Existing Customer	35.000	F	2.000	High School	Married	Less than $40K	Blue	21.000	5.000	
2169	712826208.0...	Existing Customer	39.000	F	2.000	Graduate	Married	$40K - $60K	Blue	26.000	5.000	
2170	718181433.0...	Existing Customer	65.000	F	0.000	Uneducated	Single	Less than $40K	Blue	56.000	6.000	
2171	714460683.0...	Existing Customer	65.000	M	0.000	High School	Married	Less than $40K	Blue	56.000	3.000	
2172	716460933.0...	Existing Customer	51.000	M	2.000	Unknown	Married	$120K +	Blue	42.000	4.000	
2173	710787708.0...	Existing Customer	56.000	F	1.000	Uneducated	Married	$40K - $60K	Blue	40.000	4.000	
2174	713800158.0...	Existing Customer	58.000	M	0.000	High School	Married	$80K - $120K	Blue	36.000	6.000	
2175	712551258.0...	Attrited Customer	39.000	M	1.000	Uneducated	Married	Less than $40K	Blue	36.000	3.000	
2176	715379133.0...	Existing Customer	51.000	F	1.000	College	Single	Less than $40K	Blue	36.000	4.000	
2177	712091433.0...	Existing Customer	58.000	M	4.000	Uneducated	Single	$40K - $60K	Silver	50.000	5.000	
2178	771756333.0...	Existing Customer	34.000	F	3.000	Graduate	Married	Less than $40K	Blue	23.000	5.000	
2179	779895108.0...	Existing Customer	39.000	M	3.000	Unknown	Single	$80K - $120K	Blue	23.000	6.000	
2180	771976683.0...	Existing Customer	54.000	F	1.000	College	Married	Unknown	Blue	43.000	3.000	
2181	716800908.0...	Existing Customer	26.000	F	2.000	High School	Single	Unknown	Blue	13.000	5.000	

Figura 7.51. Fragmento del conjunto de datos “Desgaste del cliente de crédito” antes de efectuar la recodificación de campos.

	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mor
2150	718777008.0...	0	59.000	1	0.000	1	2	4	1	47.000	5.000	1.00
2151	713352183.0...	0	30.000	1	1.000	5	2	1	1	17.000	5.000	1.00
2152	716128083.0...	0	56.000	1	2.000	1	2	2	1	36.000	5.000	2.00
2153	721286158.0...	0	33.000	1	2.000	0	1	5	1	15.000	4.000	2.00
2154	757881258.0...	0	31.000	1	0.000	4	3	3	1	24.000	6.000	2.00
2155	715843533.0...	0	26.000	1	1.000	3	1	2	1	36.000	3.000	4.00
2156	717234183.0...	0	29.000	1	1.000	4	2	1	1	18.000	4.000	1.00
2157	717260058.0...	0	53.000	0	2.000	1	2	1	1	42.000	6.000	1.00
2158	709710783.0...	0	35.000	0	0.000	1	1	1	1	36.000	4.000	3.00
2159	787284858.0...	0	35.000	1	2.000	1	0	5	1	29.000	5.000	1.00
2160	815845758.0...	0	56.000	0	4.000	1	2	1	1	51.000	4.000	2.00
2161	720639783.0...	0	35.000	0	1.000	1	1	0	2	27.000	4.000	3.00
2162	713331408.0...	0	35.000	1	3.000	1	2	2	1	36.000	6.000	2.00
2163	778693458.0...	1	32.000	1	1.000	1	3	3	1	13.000	3.000	3.00
2164	788610258.0...	0	49.000	1	4.000	3	0	1	1	43.000	5.000	2.00
2165	711173883.0...	0	36.000	1	4.000	4	3	5	1	29.000	3.000	2.00
2166	717628608.0...	0	37.000	1	3.000	0	2	3	1	29.000	5.000	1.00
2167	789261558.0...	0	58.000	1	1.000	1	2	3	1	47.000	3.000	1.00
2168	770647233.0...	0	35.000	0	2.000	1	2	1	1	21.000	5.000	6.00
2169	712826208.0...	0	39.000	0	2.000	4	2	2	1	26.000	5.000	1.00
2170	718181433.0...	0	65.000	0	0.000	1	1	1	1	56.000	6.000	1.00
2171	714460683.0...	0	65.000	1	0.000	1	2	1	1	56.000	3.000	3.00
2172	716460933.0...	0	51.000	1	2.000	0	2	5	1	42.000	4.000	1.00
2173	710787708.0...	0	56.000	0	1.000	1	2	2	1	40.000	4.000	2.00
2174	713800158.0...	0	58.000	1	0.000	1	2	4	1	36.000	6.000	3.00
2175	712551258.0...	1	39.000	1	1.000	1	2	1	1	36.000	3.000	3.00
2176	715379133.0...	0	51.000	0	1.000	3	1	1	1	36.000	4.000	4.00
2177	712091433.0...	0	58.000	1	4.000	1	1	2	2	50.000	5.000	2.00
2178	771756333.0...	0	34.000	0	3.000	4	2	1	1	23.000	5.000	3.00
2179	779895108.0...	0	39.000	1	3.000	0	1	4	1	23.000	6.000	1.00
2180	771976683.0...	0	54.000	0	1.000	3	2	0	1	43.000	3.000	3.00
2181	716800908.0...	0	26.000	0	2.000	1	1	0	1	13.000	5.000	0.00

Figura 7.52. Fragmento del conjunto de datos “Desgaste del cliente de crédito” después de efectuar la recodificación de campos.

7.2.3.3. Construcción de nuevos datos

Considerando que los 21 atributos que integran el conjunto de datos “Desgaste del cliente de crédito” son suficientes, no será necesaria la construcción o derivación de nuevos atributos.

7.2.3.4. Integración de datos

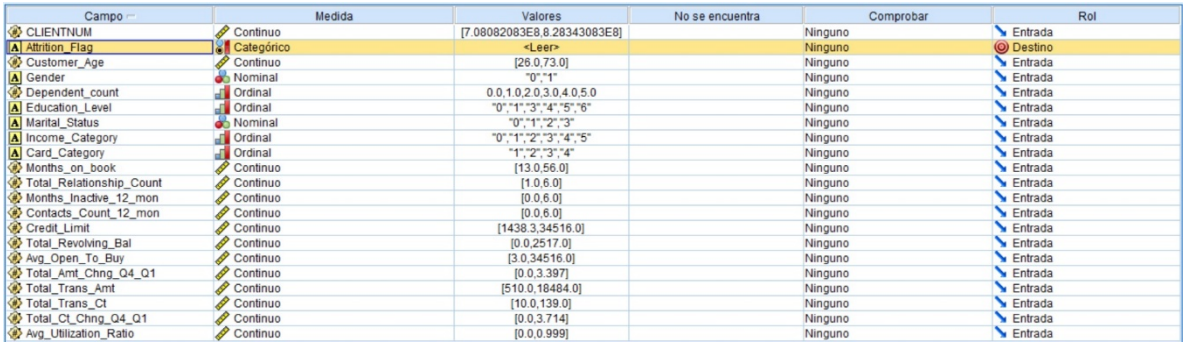
Asimismo, considerando suficientes los 10,000 campos que integran el conjunto de

datos “Desgaste del cliente de crédito”, no será necesaria la integración de registros.

7.2.3.5. Formato de datos

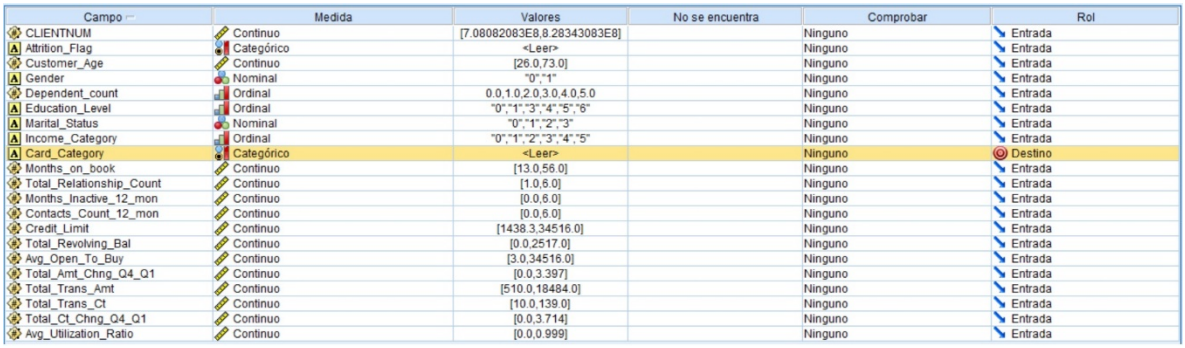
Los únicos cambios requeridos en el formato de datos, en cuanto a tipo y papel, son los siguientes:

- Para el objetivo de minería de datos número 1, el campo “estado del cliente” (ver tabla 7.5) debe ser de tipo “categórico” y su papel debe ser “destino” (objetivo). Los 20 restantes atributos mantienen su tipo, mientras que su papel debe ser “entrada” (ver figura 7.53).
- Para el objetivo de minería de datos número 2, el campo “categoría de la tarjeta de crédito” (ver tabla 7.5) debe ser de tipo “categórico” y su papel debe ser “destino” (objetivo). Los 20 atributos restantes mantienen su tipo, mientras que su papel debe ser “entrada” (ver figura 7.54)



Campo	Medida	Valores	No se encuentra	Comprobar	Rol
CLIENTNUM	Continuo	[7.08082083E8,8.28343083E8]		Ninguno	Entrada
Attrition_Flag	Categórico	<Leer>		Ninguno	Destino
Customer_Age	Continuo	[26.0,73.0]		Ninguno	Entrada
Gender	Nominal	"0","1"		Ninguno	Entrada
Dependent_count	Ordinal	0.0,1.0,2.0,3.0,4.0,5.0		Ninguno	Entrada
Education_Level	Ordinal	"0","1","3","4","5","6"		Ninguno	Entrada
Marital_Status	Nominal	"0","1","2","3"		Ninguno	Entrada
Income_Category	Ordinal	"0","1","2","3","4","5"		Ninguno	Entrada
Card_Category	Ordinal	"1","2","3","4"		Ninguno	Entrada
Months_on_book	Continuo	[13.0,56.0]		Ninguno	Entrada
Total_Relationship_Count	Continuo	[1.0,6.0]		Ninguno	Entrada
Months_Inactive_12_mon	Continuo	[0.0,6.0]		Ninguno	Entrada
Contacts_Count_12_mon	Continuo	[0.0,6.0]		Ninguno	Entrada
Credit_Limit	Continuo	[1438.3,34516.0]		Ninguno	Entrada
Total_Revolving_Bal	Continuo	[0.0,2517.0]		Ninguno	Entrada
Avg_Open_To_Buy	Continuo	[3.0,34516.0]		Ninguno	Entrada
Total_Amt_Chng_Q4_Q1	Continuo	[0.0,3.397]		Ninguno	Entrada
Total_Trans_Amt	Continuo	[510.0,18484.0]		Ninguno	Entrada
Total_Trans_Ct	Continuo	[10.0,139.0]		Ninguno	Entrada
Total_Ct_Chng_Q4_Q1	Continuo	[0.0,3.714]		Ninguno	Entrada
Avg_Utilization_Ratio	Continuo	[0.0,0.999]		Ninguno	Entrada

Figura 7.53. Formato de datos para el objetivo de minería de datos número 1. Nótese que el campo “estado del cliente” (*Attrition_Flag*) es de tipo “categórico” y su papel es “destino” (objetivo).



Campo	Medida	Valores	No se encuentra	Comprobar	Rol
CLIENTNUM	Continuo	[7.08082083E8,8.28343083E8]		Ninguno	Entrada
Attrition_Flag	Categórico	<Leer>		Ninguno	Entrada
Customer_Age	Continuo	[26.0,73.0]		Ninguno	Entrada
Gender	Nominal	"0","1"		Ninguno	Entrada
Dependent_count	Ordinal	0.0,1.0,2.0,3.0,4.0,5.0		Ninguno	Entrada
Education_Level	Ordinal	"0","1","3","4","5","6"		Ninguno	Entrada
Marital_Status	Nominal	"0","1","2","3"		Ninguno	Entrada
Income_Category	Ordinal	"0","1","2","3","4","5"		Ninguno	Entrada
Card_Category	Categórico	<Leer>		Ninguno	Destino
Months_on_book	Continuo	[13.0,56.0]		Ninguno	Entrada
Total_Relationship_Count	Continuo	[1.0,6.0]		Ninguno	Entrada
Months_Inactive_12_mon	Continuo	[0.0,6.0]		Ninguno	Entrada
Contacts_Count_12_mon	Continuo	[0.0,6.0]		Ninguno	Entrada
Credit_Limit	Continuo	[1438.3,34516.0]		Ninguno	Entrada
Total_Revolving_Bal	Continuo	[0.0,2517.0]		Ninguno	Entrada
Avg_Open_To_Buy	Continuo	[3.0,34516.0]		Ninguno	Entrada
Total_Amt_Chng_Q4_Q1	Continuo	[0.0,3.397]		Ninguno	Entrada
Total_Trans_Amt	Continuo	[510.0,18484.0]		Ninguno	Entrada
Total_Trans_Ct	Continuo	[10.0,139.0]		Ninguno	Entrada
Total_Ct_Chng_Q4_Q1	Continuo	[0.0,3.714]		Ninguno	Entrada
Avg_Utilization_Ratio	Continuo	[0.0,0.999]		Ninguno	Entrada

Figura 7.54. Formato de datos para el objetivo de minería de datos número 1. Nótese que el campo “categoría de la tarjeta de crédito” (*Card_Category*) es de tipo “categórico” y su papel es “destino” (objetivo).

7.2.4. Modelado

Para ilustrar la ejecución de las actividades que integran la fase de modelado, se considerará sólo el objetivo de minería de datos número 1:

- **Objetivo de minería de datos número 1**: Predecir, a través de clasificación, si un cliente abandonará los servicios de tarjeta de crédito, de forma que la institución bancaria o crediticia pueda actuar anticipadamente y trate de revertir la previsible decisión del usuario.

7.2.4.1. Selección de técnicas de modelado

Inicialmente, fueron seleccionados 10 modelos que responden al objetivo de minería de datos mencionado. En otras palabras, este objetivo busca construir un modelo de clasificación supervisada, cuya tarea sea ubicar a un cliente de crédito en las categorías de activo o con desgaste. Con base en la herramienta de minería de datos IBM SPSS MODELER, los modelos seleccionados fueron los siguientes:

- Red neuronal MLP
- Máquina de soporte vectorial (SVM)
- Máquina de soporte vectorial lineal (LSVM)
- Red bayesiana
- Algoritmo de los K vecinos más cercanos (K-NN)
- Regresión logística
- Árbol de decisión *QUEST*
- Árbol de decisión CHAID
- Árbol de decisión C&R
- Árbol de decisión C 5.0

7.2.4.2. Métodos de comprobación

Considerando que los modelos propuestos son técnicas de clasificación supervisada, debido al tipo de objetivo de minería de datos establecido, los métodos de comprobación a utilizar serán la matriz de confusión y métricas de desempeño para las tareas de clasificación supervisada, los cuales se presentan en el apartado 5.4.4.1. Las métricas de evaluación del desempeño a considerar serán las siguientes:

- Exactitud
- Precisión
- Sensibilidad

- F1 Score

En relación a los datos para comprobar el criterio de bondad del modelo, todos los modelos de aprendizaje supervisado propuestos incluyen entre sus funciones la partición de los datos en dos conjuntos, uno para la fase de entrenamiento del modelo y el otro para la de prueba o generalización del mismo.

7.2.4.3. Generación de los modelos

Como se puede apreciar en la figura 7.55, utilizando el paquete IBM SPSS MODELER, fueron generados los siguientes 10 modelos:

- Red neuronal MLP
- Máquina de soporte vectorial (SVM)
- Máquina de soporte vectorial lineal (LSVM)
- Red bayesiana
- Algoritmo de los K vecinos más cercanos (K-NN)
- Regresión logística
- Árbol de decisión QUEST
- Árbol de decisión CHAID
- Árbol de decisión C&R
- Árbol de decisión C 5.0

Inicialmente se utilizaron los parámetros propuestos por defecto por cada modelo. Como se indicó, los 10 modelos pertenecen a la categoría de aprendizaje supervisado, para resolver un problema de clasificación, por lo que como criterio de bondad del modelo se consideraron las métricas de desempeño obtenidas a partir de la matriz de confusión y relacionadas en el apartado anterior.

Aquí también es necesario hacer notar que, para la generación de estos primeros 10 modelos de clasificación, no se consideró el conjunto inicial de los 20 predictores (ver tablas 7.4 y 7.5), sino que, tomando en cuenta la selección de características resultante del análisis de componentes principales y la importancia de los predictores indicada por la generación del modelo de red neuronal MLP, se llevó a cabo un proceso incremental de exclusión de características. De esta forma, los resultados de los modelos generados que se describen a continuación corresponden a una primera selección de características, en la cual los siguientes predictores no fueron considerados, dada su escasa utilidad:

- ID cliente
- Edad

- Género

Nótese que los resultados de los modelos de clasificación que a continuación se presentan podrían mejorarse al excluir otras características irrelevantes en el conjunto de datos.

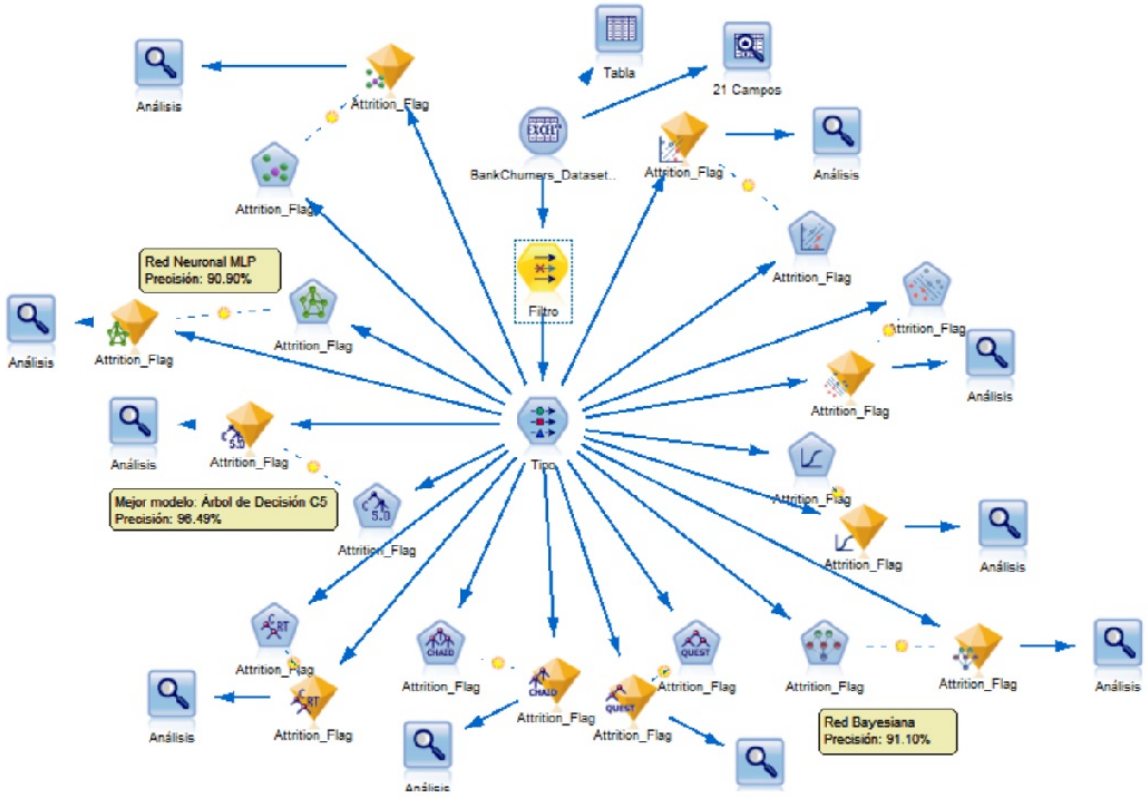


Figura 7.55. Generación de los modelos de clasificación supervisada para el objetivo número 1 de minería de datos, utilizando la herramienta IBM SPSS MODELER.

Las figuras de la 7.56 a la 7.58 muestran los resultados relacionados con el valor de exactitud y matriz de confusión de los tres modelos de clasificación generados que exhibieron un mejor desempeño. Como se puede apreciar en la figura 7.55, los mejores modelos generados fueron:

- Árbol de decisión 5.0
- Red bayesiana
- Red neuronal MLP

Resultados para el campo de resultado Attrition_Flag

Comparando $C-Attrition_Flag con Attrition_Flag

Correctos	3,958	96.49%
Erróneos	144	3.51%
Total	4,102	

Matriz de coincidencias para $C-Attrition_Flag (las filas muestran las reales)

	0.000000	**1.000000**
0.000000	2,457	57
1.000000	87	1,501

Evaluación del rendimiento

0.000000	0.455
1.000000	0.912

Figura 7.56. Valor de exactitud y matriz de confusión para el modelo de árbol de decisión 5.0 generado.

Resultados para el campo de resultado Attrition_Flag

Comparando $B-Attrition_Flag con Attrition_Flag

Correctos	3,737	91.1%
Erróneos	365	8.9%
Total	4,102	

Matriz de coincidencias para $B-Attrition_Flag (las filas muestran las reales)

	0.000000	**1.000000**
0.000000	2,326	188
1.000000	177	1,411

Evaluación del rendimiento

0.000000	0.416
1.000000	0.824

Figura 7.57. Valor de exactitud y matriz de confusión para el modelo de red bayesiana generado.

Resultados para el campo de resultado Attrition_Flag

Comparando $N-Attrition_Flag con Attrition_Flag

Correctos	3,729	90.91%
Erróneos	373	9.09%
Total	4,102	

Matriz de coincidencias para $N-Attrition_Flag (las filas muestran las reales)

	0.000000	**1.000000**
0.000000	2,324	190
1.000000	183	1,405

Evaluación del rendimiento

0.000000	0.414
1.000000	0.822

Figura 7.58. Valor de exactitud y matriz de confusión para el modelo de red neuronal MLP generado.

7.2.4.4. Evaluación de los modelos

En la tabla 7.6 se relacionan las métricas de evaluación del desempeño de cada uno de los 10 modelos de predicción generados. Nótese que la evaluación no sólo consideró las métricas de desempeño relacionadas en la tabla 7.6, sino también otros aspectos importantes, entre los que destacan:

- Facilidad de interpretación de los resultados producidos por el modelo
- Identificación de los campos predictores más importantes

Como resultado, los siguientes modelos fueron seleccionados:

- Árbol de decisión 5.0
- Red bayesiana
- Red neuronal MLP

Tabla 7.6. Métricas de evaluación del desempeño de los 10 modelos de clasificación generados para el objetivo de minería de datos número 1

Modelo	Exactitud (bondad)	Precisión	Sensibilidad (*recall*)	Especificidad
Red neuronal MLP	0.9091 (90.91%)	0.88	0.88	0.92
Red neuronal SVM	0.8944 (89.44%)	0.87	0.84	0.92
Red neuronal LSVM	0.8564 (85.64%)	0.81	0.80	0.88
Red neuronal Bayesiana	0.9110 (91.10%)	0.88	0.88	0.92
Algoritmo KNN	0.8718 (87.18%)	0.85	0.80	0.91
Árbol de decisión C5.0	0.9649 (96.49%)	0.96	0.94	0.97
Árbol de decisión CRT	0.8927 (89.27%)	0.84	0.88	0.89
Árbol de decisión CHAID	0.8727 (87.27%)	0.83	0.84	0.89
Árbol de decisión QUEST	0.8567 (85.67%)	0.85	0.76	0.91
Regresión logística	0.8564 (85.64%)	0.82	0.80	0.88

A diferencia del primer caso de estudio, en este segundo caso no se discutirán las fases de evaluación y despliegue, ya que el proceso de selección de características (predictores) debe continuar de forma incremental, hasta que ya no sea posible mejorar la exactitud (bondad) de los modelos de clasificación generados.

VIII. CONCLUSIONES

El término datos a gran escala —también referido como datos masivos o *big data*— significa un cambio fundamental en la forma en que se crean, mantienen y examinan los grandes volúmenes de datos que se generan día a día en la era digital en la cual estamos inmersos. Las organizaciones, el sector empresarial, la industria y la academia pueden impulsar la innovación, obtener conocimientos importantes y generar valor sustancial a través del análisis de los datos a gran escala. Sin embargo, para utilizar plenamente este concepto, se necesitan una infraestructura sólida, análisis potentes y una estrategia de minería de datos que garantice la privacidad y la correcta manipulación de la información recolectada.

Los datos a gran escala son cruciales para las medianas y grandes corporaciones, empresas y negocios, puesto que las ayudan a tomar decisiones estratégicas. Los métodos, tecnologías y herramientas derivados de la minería de datos se pueden utilizar para procesar y analizar estos grandes volúmenes de datos, lo que permite generar datos derivados que contienen información y conocimiento esenciales para la toma de decisiones en el dominio en cuestión.

El objetivo del presente material fue proporcionar al lector información básica, ejemplos y aplicaciones de las fuentes de producción de grandes volúmenes de datos, así como sus características, importancia, preparación, análisis y su papel en la generación de información y conocimiento valioso para la toma de decisiones estratégicas en una vasta variedad de dominios de la vida cotidiana: desde el hogar y la vida cotidiana hasta los niveles corporativos y gerenciales del mundo del comercio. Asimismo, se puede ver su utilidad en una serie de actividades que ejecutamos día a día como, por ejemplo, el deporte, la actividad física, el control y monitoreo del bienestar y la salud personal.

Se hizo especial énfasis en las fases, métodos, actividades y herramientas de la minería de datos, así como en su valioso papel en la comprensión, preparación, modelado y análisis de grandes volúmenes de datos. De forma particular, este material se centró en uno de los enfoques metodológicos de minería de datos más difundidos y utilizados en las últimas dos décadas: la metodología CRISP-DM. Para automatizar las fases y actividades de esta metodología, se presentaron dos herramientas de cómputo:

- IBM SPSS MODELER
- IDA-WEB TOOL

Como se exploró en este material, IBM-SPSS MODELER es una potente herramienta que proporciona un enorme soporte a todas las fases de la metodología de minería de datos CRISP-DM, y cuya gran ventaja es su accesibilidad para la comunidad académica. Por otra parte, IDA-WEB TOOL, aun siendo una herramienta de minería de datos en fase de prototipo y evaluación, posee para los autores de este texto un gran valor, ya que fue desarrollada por alumnos de la licenciatura en Ingeniería en Computación, de la Universidad Autónoma Metropolitana, Unidad Cuajimalpa, a través de sus Proyectos terminales I, II, y III, y de sus proyectos de servicio social. Aquí es necesario hacer notar nuevamente que, IDA-WEB TOOL reutiliza un gran número de componentes de *software* pertenecientes a la librería Pandas de Python y a otras librerías de aprendizaje automatizado disponibles.

Finalmente, en este material se presentaron tres casos de estudio con el objetivo de ilustrar el uso de las fases, actividades y herramientas de la minería de datos en la comprensión, preparación y análisis de los grandes volúmenes de datos que se producen de forma continua en la era digital. Es decir, los datos se convierten en información que, a su vez, se convierte en valioso conocimiento para la toma de decisiones estratégica. Cabe señalar que muy pocas veces es necesario trabajar con el volumen completo de datos, puesto que basta con una muestra representativa del mismo. Es por eso que los conjuntos de datos de los tres casos no representan el volumen total, sino sólo una muestra que permite su modelado y análisis.

REFERENCIAS

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.

Aggarwal, C. C. (2019). *Neural Networks and Deep Learning: A Textbook*. Springer.

Aggarwal, C. C. (2022). *Machine Learning for Text*. Springer.

Aggarwal, C. C. y Reddy, C. K. (Eds.). (2013). *Data Clustering: Algorithms and Applications*. Chapman & Hall/CRC.

Aguilar, L. J. (2013). *Big data: Análisis de grandes volúmenes de datos en organizaciones*. Alfaomega.

Cárdenas-García, M., González-Pérez, P. P., Montagna, S., Cortés, O. S. y Caballero, E. H. (2016). Modeling Intercellular Communication as a Survival Strategy of Cancer Cells: An In Silico Approach on a Flexible Bioinformatics Framework. *Bioinformatics and Biology Insights*, *10*, 5-18. https://doi.org/10.4137/BBI.S38075

Cardenas-Garcia, M. y Pérez, P. (2018). *An in Silico Approach for Understanding the Complex Intercellular Interaction Patterns in Cancer Cells* (p. 195). https://doi.org/10.5220/0006722601880195

Cerulli, G. (2023). *Fundamentals of Supervised Machine Learning: With Applications in Python, R, and Stata*. Springer.

Chatterjee, S. y Simonoff, J. S. (2013). *Handbook of Regression Analysis*. Wiley.

Chowdhary, K. R. (2020). *Fundamentals of Artificial Intelligence*. Springer.

Dua, D. y Graff, C. (2019). *UCI Machine Learning Repository* [dataset]. University of California, School of Information and Computer Science, Irvine, CA. http://archive.ics.uci.edu/ml

Everitt, B. S., Landau, S., Leese, M. y Stahl, D. (2011). *Cluster Analysis*. Wiley.

González Pérez, P. P., Cansino Malpica, D. J., López Vázquez, A. y García Torres, D. A. (2023). *IDA WEB TOOL: Una Herramienta de Minería de Datos* [Python]. Universidad Autónoma Metropolitana, Unidad Cuajimalpa. https://garciad955.github.io/IW-IDAT/

González Pérez, P. P., Ponce Rodríguez, M., García Martínez, J. A. y Pérez Pérez, A. (2023). *I-FOLDAMERIC: INTELLIGENT HP FOLDAMER DESIGN* (Versión 2.0) [Javascript - PHP]. Universidad Autónoma Metropolitana, Unidad Cuajimalpa. http://bioinformatics.cua.uam.mx/i-foldameric/#/

González-Pérez, P. P. y Cárdenas-García, M. (2018). Inspecting the Role of PI3K/AKT Signaling Pathway in Cancer Development Using an In Silico Modeling and Simulation Approach. En I. Rojas y F. Ortuño (Eds.), *Bioinformatics and Biomedical Engineering*, pp. 83-95. Springer International Publishing. https://doi.org/10.1007/978-3-319-78723-7_7

Hartshorn, S. (s. f.). *Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners*. Edición Kindle.

Hernández Orallo, J. (2010). *Introducción a la minería de datos*. Prentice Hall/Pearson.

Ignatow, G. y Mihalcea, R. F. (2016). *Text Mining: A Guidebook for the Social Sciences*.

Marr, B. (2015). *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*. Wiley.

Marr, B. (2016). *Big Data in Practice: How 45 Successful Companies Used Big Data*

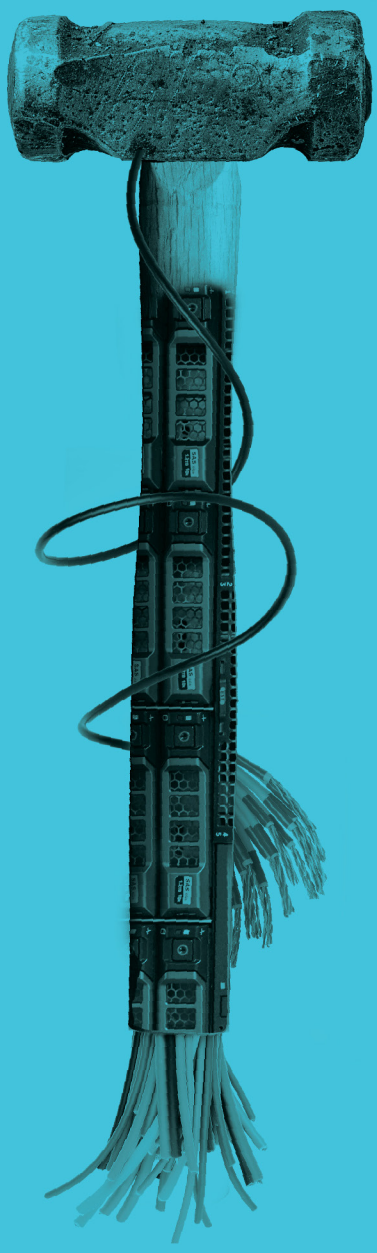
Analytics to Deliver Extraordinary Results. Wiley.
Martínez Reyes, J. C. (2021). *Desarrollo de un Prototipo para la Aplicación de Marketing Personalizado* [Tesina de Proyecto Terminal]. Universidad Autónoma Metropolitana, Unidad Cuajimalpa.
Marz, N. y Warren, J. (2015). *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications.
Mayer-Schönberger, V. y Cukier, K. (2013). *Big data: La revolución de los datos masivos*. Turner Noema.
Mayer-Schönberger, V. y Cukier, K. (2017). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. John Murray.
Mendoza Becerril, J. de J. (2021). *Análisis Inteligente de Datos en E-Commerce: Un Enfoque desde la Ingeniería de Software* [Tesina de Proyecto Terminal]. Universidad Autónoma Metropolitana, Unidad Cuajimalpa.
Perez López, C. (2021). *Data mining. The CRISP-DM methodology. The clem language and IBM SPSS modeler*. PDF Books.
Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of data warehousing*, *5*, 13-22.
Sheppard, C. (2017). *Tree-based Machine Learning Algorithms: Decision Trees, Random Forests, and Boosting*. Edición Kindle.
Suthaharan, S. (2015). *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*. Springer.
Tan, P.-N., Steinbach, M. y Kumar, V. (2018). *Introduction to Data Mining*. University of Minnesota.
Tolk, A. (2015). The next generation of modeling & simulation: Integrating big data and deep learning. *Proceedings of the Conference on Summer Computer Simulation*, 1-8.
Wang, W. y Yang, J. (2005). *Mining Sequential Patterns from Large Data Sets*. Springer.
Ward, J. S. y Barker, A. (2013, septiembre 20). *Undefined By Data: A Survey of Big Data Definitions*. arXiv.Org. https://arxiv.org/abs/1309.5821v1
Witten, I. H., Frank, E., Hall, M. A. y Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. MK.
Young, D. S. (2017). *Handbook of Regression Methods*. CRC Press.
Zhang, C. y Zhang, S. (2002). *Association Rule Mining: Models and Algorithms*. Springer.

GLOSARIO

Término	Descripción
CRISP-DM	Metodología de minería de datos ampliamente difundida y utilizada en las dos últimas décadas. Las fases que la integran son: comprensión del dominio del problema, comprensión de los datos, preparación de los datos, modelado, evaluación, y despliegue.
Datos a gran escala o datos masivos (del inglés *big data*)	Enormes y complejos conjuntos de datos producidos por múltiples fuentes digitales; se caracterizan por su volumen, velocidad, variedad, veracidad y valor.
e-commerce	Se refiere a las plataformas de comercio electrónico disponibles en sitios web, tales como Amazon, Mercado Libre, Walmart, Liverpool, entre otras. El comercio electrónico constituye una de las principales fuentes generadoras de datos a gran escala.
IBM SPSS MODELER	Plataforma profesional de minería de datos desarrollada por IBM, la cual permite ejecutar todas las fases de la metodología de minería de datos CRISP-DM.
IDA-WEB TOOL	Herramienta de minería de datos desarrollada a nivel prototipo por alumnos de la licenciatura en Ingeniería en Computación, de la Universidad Autónoma Metropolitana, Unidad Cuajimalpa. La herramienta reutiliza un gran número de componentes de la librería Pandas de Python y de otras librerías de aprendizaje automatizado.
IoT	Internet de las cosas (IoT por sus siglas en inglés). Sistemas de dispositivos físicos —comúnmente sensores y actuadores— que reciben y transfieren datos a través de redes inalámbricas. Esta tecnología permite conectar dispositivos domésticos a Internet, constituyendo así una fuente generadora de datos a gran escala.
K-NN	K-NN (del inglés *K-nearest neighbors*) es un algoritmo de clasificación de aprendizaje supervisado. Utiliza la proximidad entre los puntos en el espacio como criterio de clasificación de nuevos puntos. La cercanía entre los puntos es calculada mediante alguna métrica, comúnmente, la distancia Euclidiana.
LSVM	Máquina de vectores de soporte lineal.
Minería de datos	El objetivo principal de la minería de datos es el descubrimiento de patrones, relaciones, correlaciones, etcétera, en grandes volúmenes de datos. Sus métodos y técnicas permiten procesar y analizar grandes volúmenes de datos, y, como resultado, producen nuevos datos derivados, los cuales contienen información y conocimiento de gran valor para la toma de decisiones en el dominio en cuestión.
MLP	Red neuronal perceptrón multiestrato.
SVM	Máquina de vectores de soporte.
SMOTE	Del inglés Synthetic Minority Over-sampling Technique. Algoritmo utilizado para tratar el problema del desbalance de clases en conjuntos de datos en tareas de clasificación supervisada. La técnica consiste en aumentar el número de ejemplos en la clase minoritaria. SMOTE utiliza un algoritmo de interpolación y se basa en la técnica de los K vecinos más cercanos (K-NN).

Datos a gran escala. Un enfoque desde la minería de datos, de Pedro Pablo González Pérez, es una obra que se puede encontrar en la página web del repositorio de la UAM Cuajimalpa Concéntric@.

La asistencia editorial estuvo a cargo de Denise Ocaranza.



Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA