

PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

ESCUELA DE SISTEMAS

DISERTACIÓN PREVIA A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS Y COMPUTACIÓN

ANÁLISIS DEL COSTO BENEFICIO DE HERRAMIENTAS GRATUITAS Y LICENCIADAS PARA MINERIA DE DATOS

ENVER CAZARES

DIRECTOR: MTR. ESPINOSA VITERI LUIS OSWALDO

QUITO, 2022

Agradecimientos

Quiero agradecer a mis padres Willman y Anadela, quienes me han dado su cariño, apoyo y amor en todo este trayecto de mi vida universitaria. Sé que a pesar de todos los problemas ellos me seguirán guiando y brindando su apoyo como siempre lo han hecho, les agradezco por todo.

A mis tíos Alex García y José Luis Cazarez por estar siempre pendientes de mí, ellos a más de ser unas personas increíbles han sido unos segundos padres y unos buenos amigos, ya que siempre que necesitaba algo podía contar con ellos, muchas gracias tíos.

A todos esos increíbles profesores, quienes con su gran profesionalidad me brindaron su conocimiento el cual me ayudaron a aclarar mis dudas y resolver problemas, aquellos maestros quienes, a pesar de ser increíbles personas, me dieron sus consejos y experiencia para convertirme en un gran profesional.

Muchas Gracias a todos.

Dedicatoria

Esta tesis principalmente quiero dedicarla a mis padres quienes me han apoyado desde el primer día, hasta el último, día tras día, por ser mis consejeros y mi fortaleza para no rendirme y seguir siendo el joven que ellos criaron con mucho amor y humildad, les agradezco de todo corazón por el gran esfuerzo que han hecho por mí.

Además, quiero dedicar esta tesis a mi abuelita por parte de mi padre Teresa Valdivieso, que en paz descanse, lamentablemente ya no se encuentra con nosotros, pero sé que desde el cielo sigue cuidándonos, me hubiera encantado que esté presente en la ceremonia de graduación, lamentablemente la pandemia del coronavirus nos ha despojado de personas queridas.

Para finalizar, quiero dedicar esta tesis a mi abuelita por parte de mi madre, Berta Garcia, quien, con su inmenso amor, me ha cuidado desde que tengo memoria, “Gracias mami Berty por todo su amor y por su apoyo incondicional”.

Resumen

En los últimos cuatro años en el planeta se ha generado el 90% de la información disponible, de la cual apenas el 20% ha sido explorada. La era de la Big Data implica que cada vez es más barato almacenar información, adicional considerando que las redes de comunicación permiten un mayor acceso a la data disponible.

En este contexto la minería de datos se podría definir como el conjunto de técnicas y herramientas que permiten a las organizaciones encontrar patrones de comportamiento y predecir futuros escenarios con el objetivo de obtener una ventaja competitiva, a través de la personalización del mercadeo y el diseño de nuevos productos que encajen con las expectativas de los clientes.

El estado actual del conocimiento prevé que en promedio cada dos años se duplicará la información disponible, convirtiendo al científico de datos en un perfil altamente valorado y con grandes posibilidades de empleo pleno.

El presente documento describe los fundamentos de la minería de datos, dando un vistazo a cada uno de los pasos a realizar desde el KDD hasta la interpretación de datos. Así mismo, se hace una revisión de las herramientas disponibles, tanto licenciadas como gratuitas, no solo como un enfoque descriptivo, sino analítico, proveyendo información suficiente para el momento de decantarse por una u otra solución.

Abstract

In the last four years, 90% of the available information has been generated on the planet, of which only 20% has been explored. The Big Data era implies that it is becoming cheaper and cheaper to store information, in addition to the fact that communication networks allow greater access to the available data.

In this context, data mining could be defined as the set of techniques and tools that allow organizations to find behavioral patterns and predict future scenarios to gain a competitive advantage through the personalization of marketing and the design of new products that match customer expectations.

The current state of knowledge foresees that, on average, every two years, the available information will double, making the data scientist a highly valued profile with great possibilities for total employment.

This paper describes the fundamentals of data mining, taking a look at each step from KDD to data interpretation. It also reviews the available tools, both licensed and accessible, not only as a descriptive approach but also as an analytical one, providing enough information to choose one solution.

Índice de contenidos

Índice de contenidos I

Índice de tablas III

Índice de Figuras IV

Capítulo I: Introducción 1

1.1. Justificación 1

1.2. El Problema 4

1.2.1 Planteamiento del Problema 4

1.3. Objetivos 6

1.3.1. Objetivo General 6

1.3.2. Objetivos específicos 6

1.4. Alcance 7

1.5. Antecedentes y Marco Conceptual 8

1.5.1. Antecedentes 8

1.5.2. Metodología 9

Capítulo II: Marco Teórico 11

2.1. Historia de la minería de datos 12

2.1.1. Minería de datos VS Análisis estadístico 14

2.2. Fundamentos de la Minería de Datos 14

2.2.1. Proceso KDD 14

2.2.2. Dominio del problema 16

2.2.3. Selección 16

2.2.4. Limpieza 18

2.2.5. Transformación 18

2.2.6. Minería de Datos 19

2.2.7. Interpretación 20

Capítulo III: Herramientas para hacer minería de datos 21
3.1. Herramientas licenciadas 22
3.1.1. Rapid miner 22
3.1.2. MicroStrategy 25
3.1.3. SAS 27
3.1.4. Microsoft Azure 29
3.2. Herramientas gratuitas 32
3.2.1. R 32
3.2.2. Weka 35
3.2.3. Orange 37
Capítulo IV: Minería de Datos en las Empresas 39
4.1. Impacto de la Minería de Datos en las Empresas 40
4.2. Costo Beneficio de la Minería de Datos 42
4.3. Comparativa entre herramientas licenciadas y gratuitas 44
Capítulo V: Conclusiones y Recomendaciones 47
5.1. Conclusiones 47
5.2. Recomendaciones 49
Bibliografía 50

Índice de tablas

Tabla 1 Evolución de las tecnologías relacionadas con Data Mining 13
Tabla 2 Algoritmos de Data Mining 20
Tabla 3 Comparativa de herramientas de minería de datos 45

Índice de Figuras

Ilustración 1 Etapas del proceso KDD....15
Ilustración 2 Características del KDD....16
Ilustración 3 Logo Rapid Miner....22
Ilustración 4 Rapid Miner Data Sciencia Life Cycle....23
Ilustración 5 Funcionalidades de Rapid Miner....24
Ilustración 6 Logo MicroStrategy....25
Ilustración 7 Prestaciones MicroStrategy....25
Ilustración 8 Funcionalidades MicroStrategy....26
Ilustración 9 Logo SAS....27
Ilustración 10 SAS Dashboard....27
Ilustración 11 Funcionalidades SAS....28
Ilustración 12 Logo Azure....29
Ilustración 13 Azure vista previa....30
Ilustración 14 Funcionalidades Azure....31
Ilustración 15 Logo R....32
Ilustración 16 Algunas gráficas con R....33
Ilustración 17 Logo Weka....35
Ilustración 18 Vista previa Weka....35
Ilustración 19 Logo Orange Data Mining....37
Ilustración 20 Gráficas con Orange....37
Ilustración 21 Tipos de Sistemas Empresariales....39

Capítulo I: Introducción

En el primer capítulo, se presenta la introducción al trabajo de disertación, en la que detalla un problema identificado en el caso de estudio. Se detallan los objetivos a cumplir a lo largo de este trabajo, además de un corto resumen de lo que se realizó en cada capítulo. Esta información nos ayudará a entender de manera general el tema a tratar y sus diferentes puntos.

1.1. Justificación

La minería de datos implica técnicas y herramientas que se utilizan para extraer información útil de grandes cantidades de datos, también puede comprender mejor el negocio y realizar acciones relacionadas para guiar a la organización hacia el éxito. La minería de datos es una de las principales herramientas que se utilizan para apoyar la toma de decisiones en los planes de gestión de una empresa. Adicional, es un mecanismo de búsqueda que se basa en la exploración de datos para la obtención de información valiosa, la minería de datos usa la bóveda de datos de alguna empresa con el fin de obtener información para poder tomar una decisión.

En la actualidad la tendencia es que las grandes empresas renueven sus herramientas relacionadas con el manejo de datos, es por esto, que en la actualidad muchos de los procesos de negocios se basan en datos y es ahí donde entra la minería de datos, debido a este proceso las empresas tienen la capacidad de reconocer tendencias y patrones que ayuden a mejorar los procesos de negocio con el fin de poder tener más éxito. En la actualidad es bien conocido que son pocas las empresas que implementan herramientas de minería de datos en función del costo y beneficio, esto puede deberse a que, desde el punto de vista productivo, la minería de datos es un proceso nuevo, que de a poco se está estableciendo en las grandes y pequeñas empresas. Dado que el auge de esta nueva tecnología se ve inevitable, se presenta la oportunidad de

realizar el actual proyecto, el mismo que permitirá ofrecer una base de información para que las empresas puedan elegir la herramienta de minería de datos más adaptable a sus condiciones en función del costo beneficio que tendrá dicha herramienta.

Empresas grandes como Coca-Cola son ejemplos de que la minería de datos es importante y muy útil para el mercado, debido a que ayuda a tomar decisiones que pueden ayudar a tener un mayor éxito. Coca-Cola lo usa en redes sociales, con el fin de ayudar a decidir qué anuncio debería ser publicado, con el fin de tener el control de visitas o clics. Con la ayuda de la inteligencia artificial y la minería de datos Coca-Cola mide la presencia de sus productos y de su competencia en los anuncios, fotografías, tips, que se publican en redes sociales, con esto esta empresa tiene una posibilidad de tener 4 veces más probabilidades de que sus publicaciones causen un mayor impacto que la de la competencia. Con la ayuda de la gran cantidad de datos que proporcionan sus seguidores en redes sociales, la compañía realizo una investigación y se descubrió que cada 2 minutos se menciona a sus productos. Coca-Cola con la ayuda de una inteligencia artificial basada en el reconocimiento de imágenes logro alcanzar superioridad con respecto a su competencia, esto se debe a que, gracias a la minería de datos se tomó la decisión adecuada para complacer a los clientes y obtener mayores vistas y clics en sus redes.

Netflix es otra de las grandes empresa que usan minería de datos, esto le ayuda a predecir con exactitud qué es lo que sus usuarios quieren ver, ya que al usar los datos como, las búsquedas que realiza, los dispositivos en los que lo usa, que días son en los que más lo usa, cuanto tiempo lo usa, los comentarios, criticas, son de gran ayuda para que sus analistas logren predecir que películas son las más vistas y los géneros más vistos, Netflix ha recopilado información durante 20 años y con minería de datos pueden decidir qué películas y series pueden incluir en su plataforma, con el fin de que dichas películas y series causen un gran

impacto en los usuarios. Gracias a la minería de datos la firma estadounidense puede tener éxito en cualquier país en el que se implante ya que al usar los datos recogidos de esa región puede ofrecer una programación apta para los usuarios.

R: es considerada como la mejor herramienta gratuita para realizar minería de datos, esto se debe a que tiene código abierto y es fácil de usar sin importar que no tengas ninguna experiencia en programación. Es muy útil para el análisis estadístico, transformación y manipulación de los datos. Según la última encuesta realizada por K. Gnuggets, R se considera la herramienta más popular superando a 93 herramientas diferentes.

RapidMine: La posibilidad de analizar datos encadenando operadores. Esta herramienta que brinda una interfaz gráfica intuitiva y fácil de utilizar, se puede realizar analítica de datos únicamente arrastrando y soltando elementos dentro de su cuadro de mando.

MicroStrategy: es el software de análisis de datos e inteligencia empresarial que complementa a todos los modelos de minería de datos. Con toda una serie de puertas de enlace y controladores nativos, la plataforma puede conectarse a cualquier recurso de la empresa y analizar sus datos. MicroStrategy sobresale transformando datos complejos en visualizaciones accesibles que se distribuirán por el conjunto de una organización.

Azure Machine Learning Studio: orientada para emplear en el desarrollo, prueba y proyección de modelos analíticos con predicción. Una ventaja diferenciadora es que no es necesario tener altos conocimientos en programación para su uso.

1.2. El Problema

1.2.1 Planteamiento del Problema

En el Ecuador el uso de técnica y herramientas para la aplicación de minería de datos no está totalmente afianzada, los casos de aplicación se encuentran básicamente en empresas relacionados con las telecomunicaciones y proveedores de medicina. Se podría inclusive aseverar que las empresas, tanto medianas y pequeñas, no valoran en toda su plenitud, el activo que es la información, perdiendo la oportunidad para generar un mayor impacto para su oferta en marketing, por eso es importante contar con herramientas que puedan ofrecer el análisis de grandes volúmenes de datos, considerando que existen múltiples herramientas de conocidas por las empresas que resultan convenientes de analizar previo al momento de contratar un entorno.

A nivel mundial las grandes empresas se ven obligadas a manejar minería de datos, debido a la gran cantidad de datos generados por sus usuarios diariamente, es por esta razón que se pretende analizar los factores que intervienen dentro de esta problemática:

- Económico.
- Usabilidad.
- Funcionalidad.
- Costo beneficio.
- Eficiencia.
- Portabilidad.

Con esta investigación se pretende aclarar los beneficios de las herramientas, de tal manera que el momento que se desee implementar en las empresas un entorno de minería de daros, esta corresponda a un análisis de las necesidades y recursos que estas tengan.

En el mercado de manejo de datos, se analizará tan solo la rama de marketing de tal manera que se obtendrá las mejores herramientas para este ámbito, considerando el costo beneficio que obtendrá la empresa.

En función de esta problemática se plantea la siguiente pregunta principal de investigación:

- ¿Cuál es el análisis del costo beneficio de utilizar herramientas para hacer minería de datos?

Y las preguntas secundarias:

- ¿Qué beneficio se tiene al utilizar una herramienta gratuita para minería de datos?
- ¿Qué beneficio se tiene al utilizar herramienta propietaria o con licencia para minería de datos?

¿Cuál es el costo beneficio de utilizar herramientas para minería de datos en nuevas empresas?

1.3.Objetivos

1.3.1. Objetivo General

- Analizar el costo beneficio de utilizar herramientas para hacer minería de datos, brindando a las empresas una base de consulta previo a la implementación, mediante la comparativa de las diferentes opciones disponibles en el mercado.

1.3.2. Objetivos específicos

- Analizar las ventajas y desventajas que tienen las herramientas para minería de datos gratuitas y licenciadas.
- Medir que herramienta brinda el mejor costo beneficio para minería de datos.
- Determinar la herramienta para utilizar en minería de datos, en función del marketing.

1.4. Alcance

El trabajo en desarrollo tiene como alcance el **Análisis del costo beneficio de herramientas gratuitas y licenciadas para hacer minería de datos,** debido a que las empresas medianas y pequeñas se están quedando atrás con respecto a las grandes, y esto se debe a que las empresas grandes cuentan con una ventaja muy grande: que ya han implementado en sus entornos de trabajo estrategias de marketing mediante la aplicación de la minería de datos; esto les ha ayudado a colocarse en la cúspide de los mercados, ya que con la minería de datos han logrado descubrir los patrones para ofrecer a los cliente exactamente lo que necesitan. Se plantea realizar este trabajo de tipo exploratorio para conocer el costo beneficio que pueden alcanzar tanto herramientas licenciadas como gratuitas.

Como resultado de esta investigación se entregará un informe, el cual contenga un análisis del costo beneficio entre herramientas licenciadas y gratuitas, con el fin de dar a las empresas que aún no han implementado minería de datos una base de conocimientos para que las empresas puedan tomar la decisión más favorable al momento de contratar una herramienta de software diseñada para hacer minería de datos.

1.5. Antecedentes y Marco Conceptual

1.5.1. Antecedentes

En el presente proyecto de investigación sobre el análisis del costo beneficio entre herramientas licenciadas y gratuitas para realizar minería de datos, el presente trabajo de titulación propuesto se basará en revisiones bibliografías y aportes que han realizado otros autores.

En este contexto se considera que el uso de las herramientas diseñadas para realizar minería de datos en las empresas es importante para que la organización pueda mejorar y evaluar la calidad de sus procesos.

Por lo tanto, se tomará en cuenta los siguientes trabajos de investigación:

Minería de datos. Técnicas y herramientas en la cual (*LOPEZ, GONZALEZ*, 2017) consideran una necesidad mandante la implementación de herramientas que permita a las empresas hacer minería de datos con el fin de mejorar sus procesos de negocio y tener un mayor éxito.

En este proyecto se da a conocer que es la minería de datos y la importancia que tiene en el ámbito empresarial, así como las técnicas que existen para poder realizarlo junto con las herramientas que tienen acceso a este proceso.

Data Mining: Concepts and Techniques en esta investigación, realizada por (Riquelme et al., 2018) da pautas para conocer la forma correcta para procesar datos luego de su recopilación, cuyos resultados pueden ser utilizados en otros procesos dentro de la organización.

En este proyecto se explica la minería de datos y las herramientas utilizadas para descubrir conocimientos a partir de los datos recopilados. Se centra en la viabilidad, utilidad, eficacia y escalabilidad de las técnicas de grandes conjuntos de datos.

Minería de datos en gestión del conocimiento de pymes de Colombia en esta investigación, realizada por (*Minería de Datos En Gestión Del Conocimiento de Pymes de Colombia*, 2017) los datos cobran vital importancia para establecer posibles patrones y tendencias, permitiendo gestionar el conocimiento de las organizaciones.

En este proyecto se explica como la minería de datos ayuda a descubrir patrones, y como se ha utilizado técnicas de minería de datos utilizando una herramienta propietaria, para alcanzar un avance en competencias, herramientas y TIC de las GC de 322 PYMES en las ciudades de Bogotá, Medellín, Cali, Barranquilla, Bucaramanga, Manizales, Pereira y Armenia.

1.5.2. Metodología

El presente trabajo inicia con la investigación del costo beneficio que tienen las herramientas licenciadas y gratuitas, con posibilidad de brindar el servicio de minería de datos a las empresas que deseen implementarlo, de este modo se obtendrá una base teórica en la cual se visualiza que herramienta es la más adecuada para la empresa, con la finalidad de ayudar a la toma de una decisión optima.

Posteriormente se realizará una enumeración de las herramientas tanto licenciadas como gratuitas, con sus respectivas características, precios, ventajas y desventajas, además de mencionar los sistemas operativos con los cuales es compatible.

Después de haber realizado la respectiva numeración, se dará a conocer los beneficios que les ha traído a empresas que ya han implementado minería de datos, y como se han

impuesto en la sima de los mercados, se redactara como las grandes empresas hacían el análisis de los datos antes de implementar minería de datos.

Finalmente obtendremos un informe, con una base teórica en la cual se explique como la minería de datos, puede ayudar a mejorar la empresa, además de los beneficios y costos que se puede obtener con la implementación de herramientas específicas.

Capítulo II: Marco Teórico

En la sociedad actual mucha de la información, la gran mayoría, es digital; es común que los usuarios de tecnología creen perfiles en sistema informáticos, por ejemplo, en las redes sociales. Cada vez que nos registramos en alguna red social, le estamos diciendo a internet cuales son nuestros gustos, por donde nos movemos, entre otra información.

Esta información es almacenada en bases de datos, estructuras organizadas que permiten encontrar patrones mediante consultas, cuyo objetivo es determinar cuál es el comportamiento individual y grupal en cada lugar.

La minería de datos o data mining, se encarga precisamente de eso, encontrar patrones entre las diferentes bases de datos disponibles, es principalmente utilizada en la estadística, así como, en las ciencias computacionales. Para efectivizar la búsqueda de patrones es necesario previamente contar con una base debidamente normalizada, es decir, libre de cualquier tipo de dato irrelevante.

Las aplicaciones para este tipo de tecnología son extensas, considerando que se basa en algoritmos de árbol de decisión o machine learning, es fácil pensar, por ejemplo, en sistemas que sean capaces de predecir la efectividad de algún medicamente, o talvez anticiparnos al comportamiento de nuestros clientes ante un nuevo producto, con el fin de mejorar el posicionamiento y obtener una mejor entrada.

2.1.Historia de la minería de datos

El concepto de minería de datos, contrario a lo que se podría pensar, no es nuevo; ya que, era común entre los estadísticos utilizar datos como Data Fishing o Data Archeology. A principios de los años ochenta, Rakesh Agrawal, GioWiederhold, Robert Blum y Gregory Piatetsky-Shapiro entre otros, empezaron a consolidar los términos de Minería de Datos y KDD. (Cutro, 2010)

La evolución de sus herramientas en el transcurso del tiempo puede dividirse en cuatro etapas principales:

- Colección de Datos (1960)
- Acceso de Datos (1980)
- Almacén de Datos y Apoyo a las Decisiones (principios de la década de 1990)
- Minería de Datos Inteligente (finales de la década de 1990)

Según lo detalla Beatriz Beltrán Martínez en su libro Notas de Minería de Datos, *"Las raíces de la DM se remontan a los años 50. Los departamentos de informática preparaban resúmenes de la información, principalmente de tipo comercial, que se encontraba en los ficheros del ordenador central, con el propósito de facilitar la labor directiva"* De esta manera se dieron origen a los primeros esbozos, para ese entonces los sistemas de información para la toma de decisiones empezaban a nacer, sin embargo utilizaban altos volúmenes de almacenamiento aparte de ser poco flexibles y aparentemente pensados para ser leídos solo por informáticos. Así mismo, la autora menciona que *"En los 60 nacen los sistemas gestores de base de datos que aún se mostraban rígidos y carecían de flexibilidad para realizar consultas"* (Beltrán Martínez, 2022)

Fue posterior a este aparecimiento que se vio la necesidad de crear motores de bases de datos relacionales, esto resolvía parte del problema, sin embargo, los reportes continuaban siendo muy difíciles de elaborar.

Adicional a los problemas detallados, se tuvo que lidiar con la integración de bases de datos, era un trabajo laborioso preparar y depurar información extraída desde diferentes fuentes.

Tabla 1 Evolución de las tecnologías relacionadas con Data Mining

Etapa	**Cuestión planteada**	**Tecnologías**	**Características**
Recolección de datos {Años 60}	'Dime mis beneficios totales en los últimos 4 años.'	Ordenadores, cintas, discos.	Retrospectivo, datos estáticos.
Acceso a los datos. {Años 80}	Ventas en Cataluña durante las últimas Navidades'	Bases de Datos Relacionales (SQL) ODBC	Retrospectivo, datos dinámicos a nivel de registro.
Data Warehouse y soporte a la toma de decisiones. {Años 90}	'Ventas en Andalucía detalle por delegación y descender a nivel tienda.	{OLAP}, bases de datos multidimensionales, data warehouse	Retrospectivo, obtención dinámica de datos a múltiples niveles.
Data Mining	Justifica la tendencia de venta en Castilla para el próximo año	Algoritmos avanzados, ordenadores, multiprocesadores, bases de datos masivas.	Prospectivo, obtención proactiva de información.

Fuente: (Beltrán Martínez, 2022)

2.1.1. Minería de datos VS Análisis estadístico

Según detalla Patricia Blanco Pisado en su artículo acerca del XX Simposio Internacional de Métodos Matemáticos Aplicados a las Ciencias (SIMAAC), organizado por la Universidad de Costa Rica *"... a diferencia de la Estadística, en la minería de datos no se trabaja con muestras de la información, sino que se buscan patrones ocultos en los datos mediante métodos algorítmicos y utilizando un sistema automatizado"* (Blanco Picado, 2016)

En este sentido, si bien es cierto, tanto la estadística como la minería de datos tienen su base en los modelos matemáticos; por una parte, se puede decir que el análisis matemático obtiene información que brinda insumos que posibilitan explicar por qué sucede algo, mientras que, la minería de datos basa su importancia en descubrir patrones y compartimientos que pueden predecir cual sería el comportamiento de algo o alguien bajo ciertas condiciones.

2.2.Fundamentos de la Minería de Datos

Adicional, resulta importante detallar al menos en sus generalidades cuáles son los pasos que la minería de datos utiliza, para lo cual en primera instancia hablaremos de los pasos previos resumidos en el proceso KDD para con este entendimiento poder explicar de mejor manera que es la minería de datos.

2.2.1. Proceso KDD

El proceso de descubrimiento de información en bases de datos o KDD por sus siglas en inglés (Knowledge Discovery in Databases), es proceso novedoso que permite descubrir patrones en grandes volúmenes de datos, con el objetivo de, a partir de estos, generar nuevo conocimiento.

Este proceso cuenta con cinco etapas, cada una dependiente de la anterior. En la siguiente ilustración se detallan:



Ilustración 1 Etapas del proceso KDD

(Párraga Alava, 2021)

Todas estas etapas parten de un paso previo que se conoce que es el dominio del problema.

El proceso KDD se realiza de forma iterativa e interactiva:

- **Iterativo:** Esto debido a que, al finalizar cada etapa se puede continuar con la etapa siguiente, o en su defecto, volver a una etapa anterior. Esto con el objetivo de perfeccionar su funcionamiento.

- **Interactivo:** Si bien es cierto cada uno de los procesos deben ser realizados por un profesional experto en minería de datos, es imprescindible contar con el apoyo de personas que pertenecen al área donde se está resolviendo el problema. De esta menar se cuenta con un criterio proactivo hacia el personal técnico encargado del KDD

De esta manera se garantiza una progresión expedita entre cada una de las etapas.



Ilustración 2 Características del KDD

(Párraga Alava, 2021)

A continuación, se describen cada una de las etapas del KDD.

2.2.2. Dominio del problema

En este paso preliminar se desarrolla una comprensión clara acerca del problema que se enfrenta. Establece preguntas para entender lo que se debería hacer con los datos, además de delimitar claramente los objetivos del proceso a realizar.

Adicional realiza la integración y recolección de datos reconociendo las fuentes de información más importantes con las que se trabajara a futuro.

2.2.3. Selección

Beatriz Beltrán destaca esta sección, más allá de la simple selección, la importancia de realizar este proceso de manera técnica menciona que *"el análisis posterior será mucho más*

sencillo si la fuente es unificada, accesible (interna) y desconectada del trabajo transaccional." (Beltrán Martínez, 2022) De tal manera que las siguientes etapas:

- ✓ **Depende mucho de la fuente:**
  - OLAP u OLTP.
  - Datawarehouse o copia con el esquema original.
  - ROLAP o MOLAP.
- ✓ **Depende también del tipo de usuario:**
  - Picapedreros (o granjeros): se dedican fundamentalmente a realizar informes periódicos, ver la evolución de determinados parámetros, controlar valores anómalos, etc.
  - Exploradores: encargados de encontrar nuevos patrones significativos utilizando técnicas de minería de datos
- ✓ **Recolección de Información Externa.** Aparte de información interna de la organización, los almacenes de datos pueden recoger informaciones externas:
  - Demografía (censo), páginas amarillas, psico geografías, gráficos web, información de otras organizaciones.
  - Datos compartidos en una industria o área de negocio, organizaciones y colegios profesionales, catálogos, etc.
  - Datos resumidos de áreas geográficas, distribución de la competencia, evolución de la economía, información de calendarios y climatológicas, programaciones televisivas, deportivas, catástrofes.

- Bases de datos externas compradas a otras compañías.

2.2.4. Limpieza

Se podría decir que en esta etapa se realiza el preprocesado de los datos; lo cual, entre otros, comprende:

- Eliminación de ruido,
- Estrategias para manejar valores ausentes,
- Normalización de los datos

En esta fase también conocida como data cleaning se deben eliminar el mayor número posible de datos erróneos o inconsistentes (limpieza) e irrelevantes (criba).

Para realizar este trabajo, contamos con varios métodos de índole estadístico, entre otros los siguientes:

- Histogramas (detección de datos anómalos).
- Selección de datos (muestreo, ya sea verticalmente, eliminando atributos u horizontalmente, eliminando tuplas).
- Redefinición de atributos (agrupación o separación).

2.2.5. Transformación

En esta etapa se mejoran los datos que se han logrado procesar hasta aquí, tratando de encontrar patrones útiles para la correcta representación de los datos. También se debe considerar información inherente a los datos, a lo que podríamos definir como el meta proceso.

La información es clasificada en dimensiones, para reducir las variables a las correctas que logren responder las preguntas planteadas en el dominio del problema. La otra clasificación son los atributos, se podría generar rangos para que la interpretación se base exclusivamente en parámetros que se ajusten al objetivo.

Se puede destacar la importancia de esta etapa en que *"... este es un paso crítico dentro del proceso global, que requiere un buen conocimiento del problema y una buena intuición, y que, con frecuencia, marca la diferencia entre el éxito o fracaso de la minería de datos"* (Beltrán Martínez, 2022)

2.2.6. Minería de Datos

Hasta aquí se ha hecho un repaso por los pasos previos, por lo cual, es menester contar con un concepto claro, considerando los aportes de Gutiérrez y Molina (2014), se define la minería de datos como:

> Una tecnología y estrategia de modelado matemático que intenta ayudar a comprender el contenido de una base de datos de forma general, los datos son la materia prima bruta, en el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. (Martínez Moncaleano & Palencia Fajardo, 2021)

El objetivo de esta etapa es la búsqueda de patrones insospechados y de interés. Es aquí donde se selecciona la tarea de minería de datos que, entre otras, pueden ser:

- Clasificación
- Agrupamiento (clustering)
- Patrones secuenciales

- Asociaciones

Adicional a la tarea, se debe seleccionar el algoritmo, a continuación, algunos ejemplos de algoritmos que se pueden utilizar:

Tabla 2 Algoritmos de Data Mining

Algoritmo	**Descripción**
J48	El algoritmo C4.5 (J48) es uno de los más utilizados en data mining. Se basa en la entropía de la información, la cual permite medir el desorden que existe en un conjunto de datos. Los datos del entrenamiento tienen un sistema S = s1, s2, … de muestras ya clasificadas (Quinlan, 1993).
BayesNet	Es un algoritmo basado en el teorema de Bayes. Representa gráficamente la probabilidad condicionada entre un conjunto de variables.
Hoedffging Tree	Es un algoritmo de árbol de decisión aplicado a grandes volúmenes de datos. Este algoritmo asume que la distribución de datos no cambia en el tiempo. Se basa en los modelos matemáticos de la desigualdad de Hoeffding.
OneR	Este algoritmo tiene como particularidad la selección de un atributo o regla que revele mejor la clase de salida. Es un método caracterizado por ser rápido y eficiente, en comparación con otros algoritmos de mayor complejidad.

Fuente: *(Martínez Moncaleano & Palencia Fajardo, 2021)*

2.2.7. Interpretación

Esta es la etapa en la parte más superior del proceso de minería de datos, es aquí donde se interpretan los resultados o patrones descubiertos, posiblemente se deba regresar a las fases anteriores para optimizar los resultados obtenidos. Los patrones encontrados pueden ya ser interpretados por personal no técnico que tenga mayor conocimiento acerca del negocio.

Capítulo III: Herramientas para hacer minería de datos

La ciencia de datos es en los actuales momentos de las pocas profesiones con empleo pleno a nivel mundial, en las empresas y organizaciones, hoy más que nunca, se valora a la información como el activo más importante. Grandes empresas a nivel mundial basan su modelo de negocio en el manejo de datos, por ejemplo, Netflix ha reportado que el 75% de sus clientes deciden mirar las recomendaciones que la plataforma les acerca. (Vivas, 2020)

Esto último, resulta relevante considerando que las recomendaciones en las plataformas de streaming se basan en el aprendizaje que se obtiene del comportamiento de sus propios clientes.

Entonces, si ahora resulta conocido que las grandes empresas a nivel mundial utilizan la ciencia de datos para aumentar sus ganancias, adicional que, considerando que la cantidad de datos generados en los últimos 700 días es tres veces superior a lo que se generó en el resto de la existencia de la humanidad (Loef, 2015), resulta inevitable pensar que exista en mercado una gran cantidad de herramientas de software dedicadas a la Ciencia de Datos.

Con este antecedente, resultaría complejo poder abarcar todas las herramientas de software que puedan ser aplicables a la Minería de Datos, tampoco es ese el objetivo de la presente investigación. Entonces se procederá a continuación a realizar una revisión breve de algunas de las herramientas más populares en la actualidad, desde el enfoque de su comercialización, divididas en herramientas licenciadas y herramientas gratuitas.

3.1. Herramientas licenciadas

Un dato interesante de analizar al momento de hablar de herramientas licenciadas sin duda es el costo de la implementación, lamentablemente durante la investigación resulto difícil acceder a un registro transparente de los costos de implementación de cada una de las herramientas analizadas a continuación.

Las estrategias de venta prácticamente obligan a buscar un acercamiento personal y justificación del proyecto a implementar, en todo caso, se hace referencia a otros aspectos no menos importantes como: cartera de clientes, funcionalidades y fortalezas.

3.1.1. Rapid miner

Según GetApp RapidMiner *"ofrece un paquete de productos que permiten a los analistas de datos crear nuevos procesos de extracción de datos, configurar análisis predictivos y más."* (GetApp, 2022) En definitiva esta herramienta es una Suite de diferentes recursos que acompañan en todo el proceso de análisis de datos.



Ilustración 3 Logo Rapid Miner (Rapid Miner, 2022)

Rapid miner permite montar una infraestructura propia con un server, en el que se puede programar y ejecutar análisis de datos. El paradigma consiste en una interfaz gráfica en la que se puede arrastrar cada uno de los componentes, dependiendo de la etapa del proceso en la que se encuentre. La lista de productos incluye:

- RapidMiner Studio,
- RapidMiner Server,
- RapidMiner Radoop y
- RapidMiner Streams.

De estos productos, quizá el más relevante sea Radoop, según la página oficial del producto está en capacidad de *"proporcionar una plataforma para el procesamiento de Big Data, incluidos análisis y predicciones. Radoop ofrece una interfaz visual para ETL de Big Data, analíticas, creación de informes ad hoc, modelado predictivo y visualización."* (Rapid Miner, 2022)

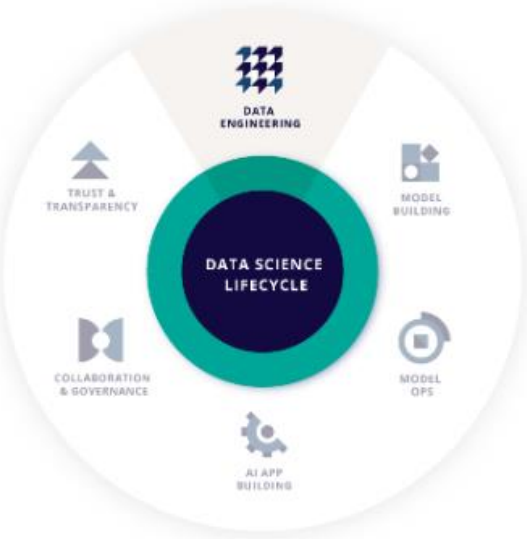


Ilustración 4 Rapid Miner Data Sciencia Life Cycle

(Rapid Miner, 2022)

Esta empresa presume de contar con varias de las empresas más grandes del mundo entre sus clientes, entre algunas se puede mencionar: Visa, Dominos Pizza, Canon, BMW, Hewllet Packard, SONY, entre otras.

Finalmente, en este breve análisis a continuación, se detallan las funcionalidades de esta primera herramienta:



Ilustración 5 Funcionalidades de Rapid Miner

(Capterra, 2022)

3.1.2. MicroStrategy

Al momento de definir que es MicroStrategy Capterra menciona que *"MicroStrategy es un líder mundial en software de analíticas y movilidad empresariales. Pionera en el espacio de inteligencia comercial y analíticas, MicroStrategy ofrece software innovador que permite a las personas tomar mejores decisiones y transformar la forma en que hacen negocios"* (Capterra, 2022)

Ilustración 6 Logo MicroStrategy

(MicroStartegy, 2022)

Lo interesante de esta opción es que la empresa no se enfoca únicamente en vender su producto, sino que además su modelo de negocio gira alrededor del servicio. Es su página se enfocan en brindar diferentes tutoriales acerca del uso de cada uno de sus productos.



Ilustración 7 Prestaciones MicroStrategy

(MicroStartegy, 2022)

A diferencia de Rapid miner que cuenta con clientes de áreas técnicas, MicroStrategy apunta a clientes de negocios, por ejemplo, banco Allianz, Ministerio de Finanzas de Honduras, Banco Fravega o BBVA de Argentina. (MicroStartegy, 2022)

Entre las funcionalidades que se destacan del producto, podemos enumerar las siguientes:



Ilustración 8 Funcionalidades MicroStrategy

(Capterra, 2022)

3.1.3. SAS

Software Analítica y Soluciones SAS destaca como su principal ventaja competitiva el uso de Inteligencia Artificial al momento del procesamiento de analítica de datos.

Ilustración 9 Logo SAS

(SAS, 2022)

En su caso su plataforma lleva el nombre de SAS Viya, y según su página oficial su *"plataforma de IA, analítica y gestión de datos ofrece una IA probada en la realidad, resistencia en medio de la incertidumbre, resultados de confianza y potenciación para todos."* (SAS, 2022)



Ilustración 10 SAS Dashboard

(SAS, 2022)

Honda, Nestle, Levi Satraus o Discover son solo algunos de sus clientes. Quizá una fortaleza diferencial con las anteriores soluciones detalladas es su adaptabilidad con Microsoft Azure, lo que permite combinar dos perspectivas diferentes y complementarias en una sola integración

Si hablamos de funcionalidades, SAS Viya dispone de un conjunto de licenciamientos, cada uno con diferentes características que permiten adaptar el producto a las necesidades del cliente:

Capabilities	SAS Visual Analytics	SAS Visual Statistics	SAS Visual Machine Learning	SAS Visual Data Science	SAS Data Science Programming*	SAS Visual Data Science Decisioning	SAS Visual Forecasting	SAS Visual Text Analytics	SAS Intelligent Decisioning
Data access, preparation & quality	✓	✓	✓	✓	✓	✓	✓	✓	✓
Visualization & reporting	✓	✓	✓	✓	✓	✓	✓	✓	✓
Statistics		✓	✓	✓	✓	✓			
Machine learning & deep learning			✓	✓	✓	✓			
Model deployment & management	^	^	✓	✓	✓	✓			
Forecasting	^	^	^	✓	✓	✓	✓		
Text analytics	^	^	^	✓	✓	✓		✓	
Optimization	^	^	^	✓	✓	✓			
Econometrics	^	^	^	✓	✓	✓			
Digital decisioning	^	^	^	^	^	✓			✓
Event streaming analytics	^	^	^	^	^	✓			

Ilustración 11 Funcionalidades SAS

(SAS, 2022)

3.1.4. Microsoft Azure

Según su web oficial Microsoft “Azure es un servicio de computación en la nube creado por Microsoft para construir, probar, desplegar y administrar aplicaciones y servicios mediante el uso de sus centros de datos” (Azure Microsoft , 2022)



Ilustración 12 Logo Azure

(Azure Microsoft , 2022)

Desde lo teórico estamos frente a la herramienta más robusta para el análisis de datos, ya de entrada desde su presentación hace gala del uso de Machine Learning dentro de sus productos. Y el hecho que sea una herramienta de Micorosoft brinda una mayor integración con el sistema operativo más usado en el sector empresarial.

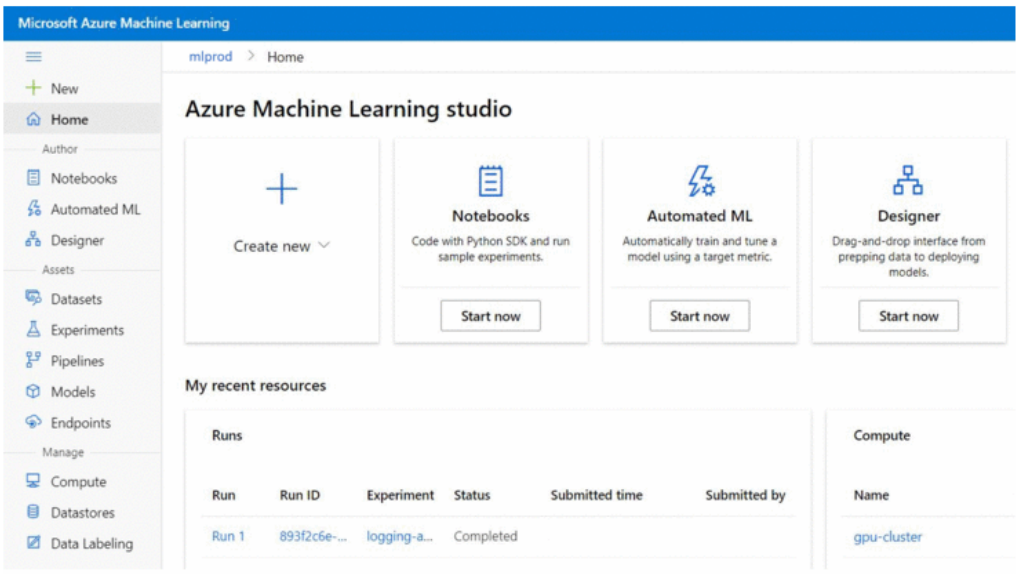


Ilustración 13 Azure vista previa

Precisamente el uso de Machine Learning provoca que cada vez sean menos las líneas de código que se deben ingresar para programar una consulta o ejecutar un proceso. Su página oficial garantiza hasta un 70% menos de pasos para el entrenamiento de modelos y hasta un 90% menos de líneas de código.

Para el efecto, divide el ciclo de vida del aprendizaje en 4 pasos:

- Preparar Datos
- Compilar y entrenar modelos
- Evaluar e implementar
- Administrar y supervisar

Para finalizar, llama la atención las funcionalidades más elaboradas, que aprovechan tecnologías ya comprobadas como Git y cuadernos de colaboración:



Ilustración 14 Funcionalidades Azure

3.2. Herramientas gratuitas

3.2.1. R

R es un entorno y lenguaje de programación con un enfoque al análisis estadístico. En los últimos años ha cobrado mayor relevancia en entornos educativos. Muchas Universidades usan este lenguaje como entorno de aprendizaje para el manejo de elementos estadísticos.



Ilustración 15 Logo R

(R Porject, 2022)

Su licencia es GNU con lo cual su distribución y uso es totalmente gratuito. Su potencialidad de uso es su principal ventaja, sin embargo, la ausencia de una interface gráfica hace que la curva de aprendizaje se sienta muy empinada y se requiera emplear tiempo para dominar el lenguaje.

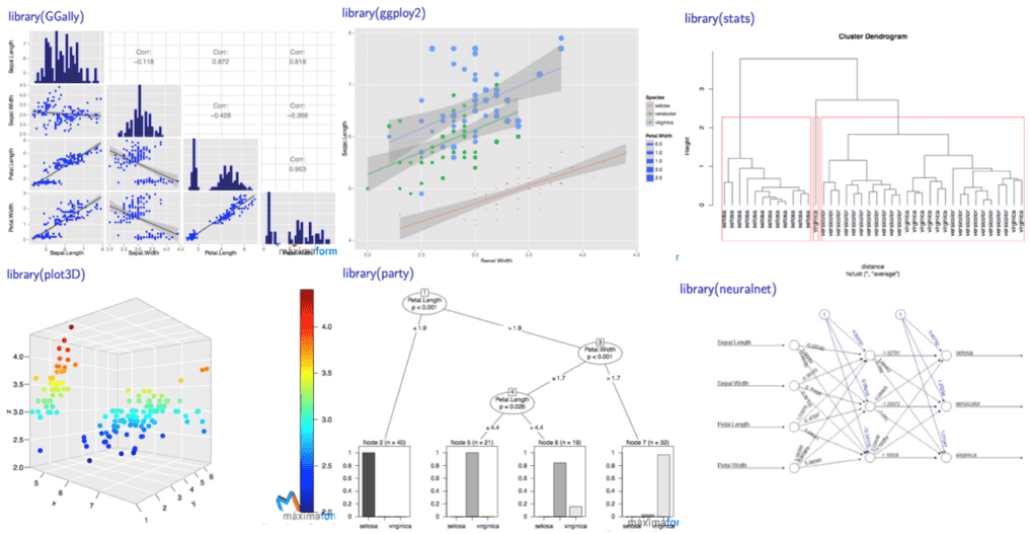


Ilustración 16 Algunas gráficas con R

Sin duda es una herramienta poderosa, si se sabe utilizar. Según el sitio web Máxima Formación (Máxima Formación, 2018), las principales ventajas de usar R son:

- Es una herramienta muy poderosa para todo tipo de procesamiento y manipulación de datos.
- Algunas técnicas avanzadas y robustas solo pueden realizarse con este software.
- Ambiente de trabajo muy flexible y extensible.
- Permite crear gráficos de alta calidad exportables en diversos formatos: PostScript, pdf, bitmap, pictex, png, jpeg, etc.
- Gran cantidad de información sobre sus funciones y paquetes de funciones.
- Consume pocos recursos informáticos.
- Está disponible para todos los sistemas operativos (Windows, Macintosh y sistemas Unix -como Linux-).

- R trabaja con otros lenguajes y permite leer datos de otros softwares como SPSS, SAS, Excel, etc.

- Puedes crear aplicaciones web interactivas (apps) con la herramienta Shiny.

- Puedes crear un flujo de trabajo para escribir informes reproductibles y dinámicos y hacerlo en varios formatos (pdf, word, html).

3.2.2. Weka

Weka (Waikato Environment for Knowledge Analysis) es una plataforma de software destinada al aprendizaje automático y la minería de datos (Data Mining) escrita en lenguaje Java y desarrollada en la Universidad de Waikato (Nueva Zelanda) (Equipo de redacción de B12 España: Marketing, Big Data, Inteligencia Artificial y Ventas., 2020)



Ilustración 17 Logo Weka

(Waikatpo, 2022)

Según el sitio oficial, Weka *"es una colección de algoritmos de aprendizaje automático para tareas de minería de datos. Contiene herramientas para la preparación, clasificación, regresión, agrupación, minería de reglas de asociación y visualización de datos."* (Waikatpo, 2022)

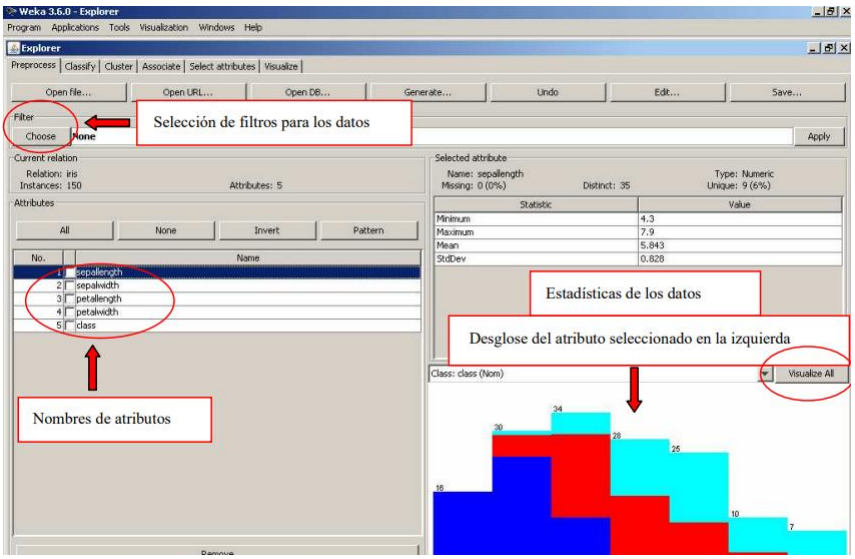


Ilustración 18 Vista previa Weka

A diferencia de R, Weka cuenta con una interface gráfica en formato de menús, divididas en secciones entre las cuales se puede parametrizar las consultas dependiendo de las fuentes de datos a utilizar.

Weka también utiliza técnica de Machine Learning *"para llevarlas a cabo, se sirve de técnicas estadísticas, algoritmos matemáticos y algoritmos de aprendizaje automático que ayudan a mejorar el rendimiento en base a la experiencia."* (Cardenas, Theran Maldonado, & Muñoz Castañeda, 2020)

Weka se denomina a sí mismo un conjunto de Librerías para tareas de minería de datos. Las librerías pueden ser llamadas desde la interface de weka o desde tus propias clases Java. Weka contiene herramientas para diferentes tareas básicas:

- Reprocesó: Multitud de herramientas para el preprocesamiento de los datos (como por ejemplo discretización de variables).
- Clasificar: Algoritmos de clasificación, distribuidos por paquetes, como por ejemplo ID3 o C4.5
- Racimos: Diferentes algoritmos de segmentación como el simple k-means.
- Asociado: Algoritmos para encontrar relaciones de asociación entre variables (Apriori entre otros).
- Seleccionar atributos: Aquí, una vez cargados los datos, Weka es capaz de buscar por nosotros las mejores variables del modelo.
- Visualizar: Herramienta de visualización de datos en los ejes cartesianos, con muchas posibilidades.

3.2.3. Orange

Mediante el uso de flujos de trabajo y una interfaz gráfica, Orange propone que realizar análisis de datos puede ser divertido.

Ilustración 19 Logo Orange Data Mining

(Orange Data Mining, 2022)

De las opciones correspondientes a las herramientas de software libre, Orange se diferencia al presentar un entorno de trabajo dinámico e intuitivo. Si bien es cierto sus funcionalidades, en comparación a R, por ejemplo, son más limitadas, existe la posibilidad que utilizar complementos disponibles.

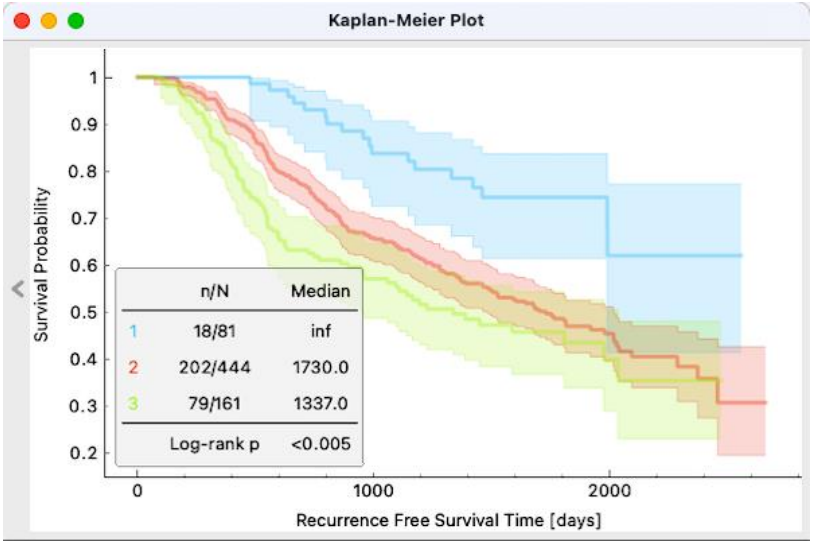


Ilustración 20 Gráficas con Orange

Algunas de las funcionalidades de Orange son:

- Análisis visual
- Búsqueda/filtro
- Creación de informes/análisis
- Descubrimiento visual
- Panel de control personalizable
- Visualización de datos

Capítulo IV: Minería de Datos en las Empresas

Las nuevas tendencias de administración, tanto en empresas públicas como privadas, orienta los esfuerzos al cliente, para lo cual deben estar preparadas para el cambio continuo y la adaptabilidad. Entonces, se podría decir que "una empresa flexible es la que se orienta hacia los clientes, posee tecnología nueva y presenta acuerdos laterales de organización e innovación" (Marcano Aular & Talavera Pereira, 2007).

Si se quiere analizar como los sistemas informáticos proveen información a la empresa, bien podríamos dividir a los mismo en dos grandes grupos: por un lado, los sistemas transaccionales y adiciona los sistemas de indicadores.

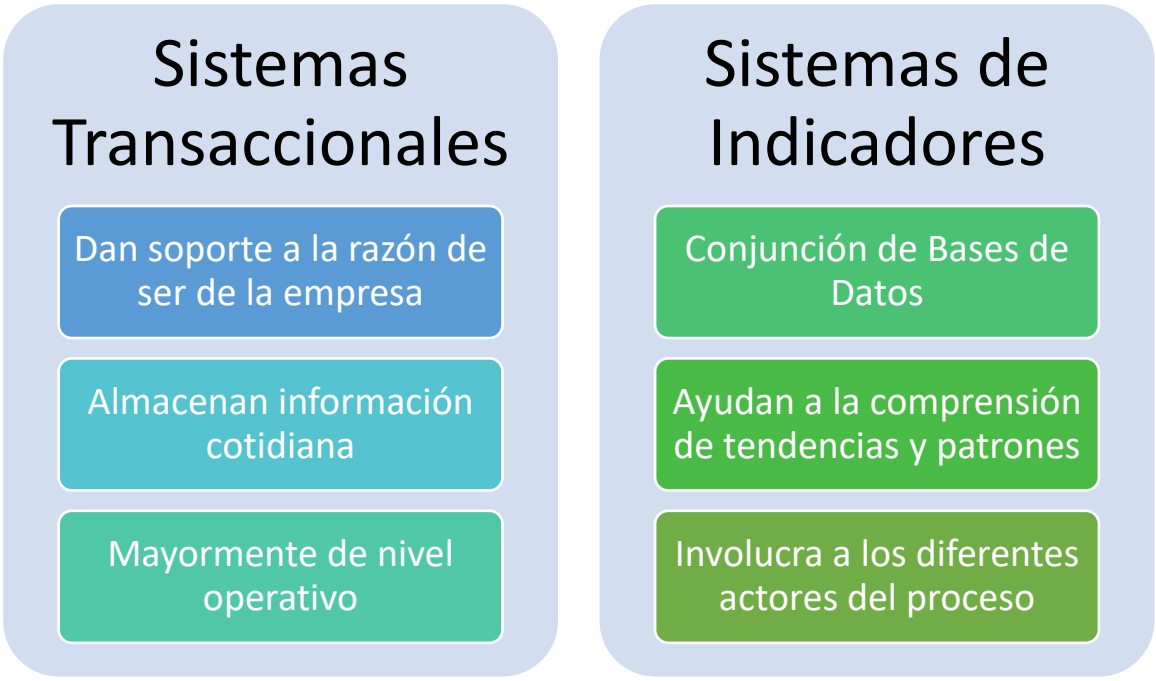


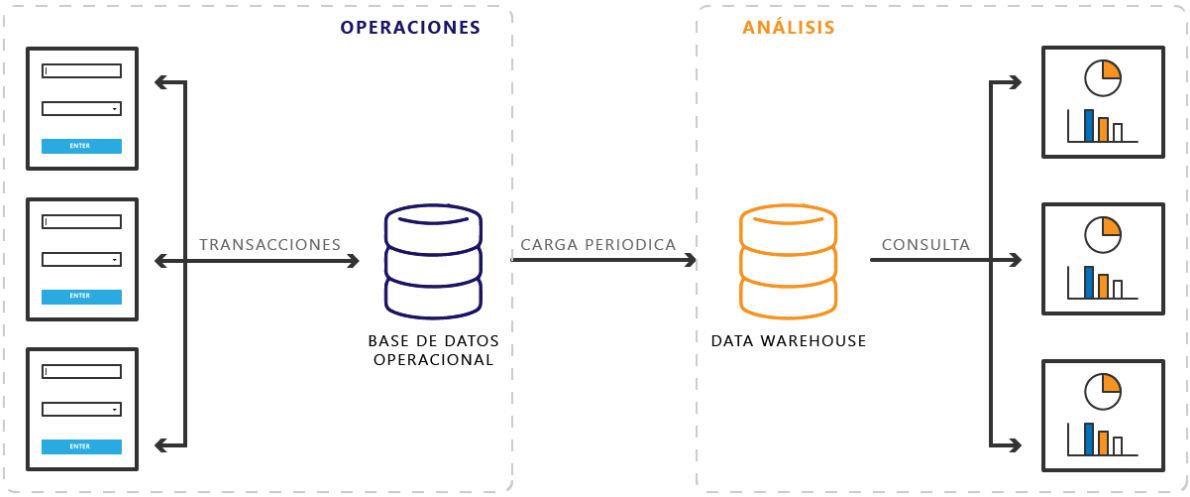
Ilustración 21 Tipos de Sistemas Empresariales

Los datos, si se mantienen en su forma original en bases de datos estructuradas, no suelen entregar beneficios efectivos; la importancia de su valor consiste en la información que se pueda derivar de estos datos, es decir, información que nos facilite la toma de decisiones.

En este sentido es pertinente analizar aun con mayor profundidad de que manera influye el uso de la Minería de Datos en entornos empresariales.

4.1.Impacto de la Minería de Datos en las Empresas

Como se mencionó anteriormente la gestión en las empresas, en las condiciones actuales basa muchas de sus decisiones en los sistemas de información, es común que cada uno de los procesos agregadores de valor cuente con software especializado, brindando acceso a un modelo estructurador de almacenamiento de datos.



Sin embargo, esta información en si misma no contiene un valor diferencial para los tomadores de decisiones. Es por eso por lo que, la tendencia apunta a la big data, a almacenar cada vez más información, para en lo posterior poder extraer tendencias y patrones que logren describir en nuevos niveles el comportamiento de los clientes, proveedores e incluso el personal dentro de las empresas.

La Minería de Datos, en este contexto brinda varios beneficios. En el blog de Intelligent (Intelligent, 2016) se describe al menos 10 ventajas acerca de la implementación:

- La minería de datos descubre información que no se esperaba obtener. Como muchos modelos diferentes son usados, algunos resultados inesperados tienden a aparecer. Las combinaciones de distintas técnicas otorgan efectos inesperados que se transforma en un valor añadido a la empresa.

- Enormes bases de datos pueden ser analizadas mediante la tecnología de data mining.

- Los resultados son fáciles de entender: personas sin un conocimiento previo en ingeniería informática pueden interpretar los resultados con sus propias ideas

- Contribuye a la toma de decisiones tácticas y estratégicas para detectar la información clave

- Te permite encontrar, atraer y retener a los clientes. Reduce el riesgo de perder clientes: ofrecer promociones especificas o productos especiales para retenerlos.

- Mejora la relación con el cliente: la empresa puede mejorar la atención al cliente a partir de la información obtenida.

- Permite ofrecer a tus clientes los productos o servicios que necesitan.

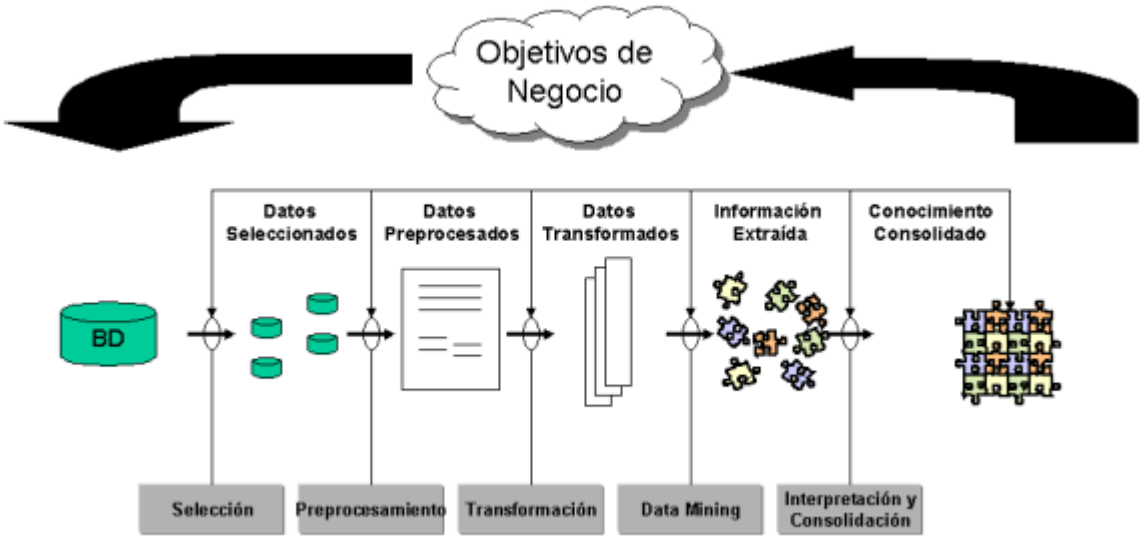
- Los modelos son confiables. Los modelos son probados y comprobados usando técnicas estadísticas antes de ser usado, para que las predicciones que se obtienen sean confiables y válidas.

- En su mayoría, los modelos se generan y construyen de manera rápida. El modelado a veces se torna más fácil puesto que muchos algoritmos han sido probados previamente.

- Abre nuevas oportunidades de negocios y ahorra costes a la empresa.

4.2.Costo Beneficio de la Minería de Datos

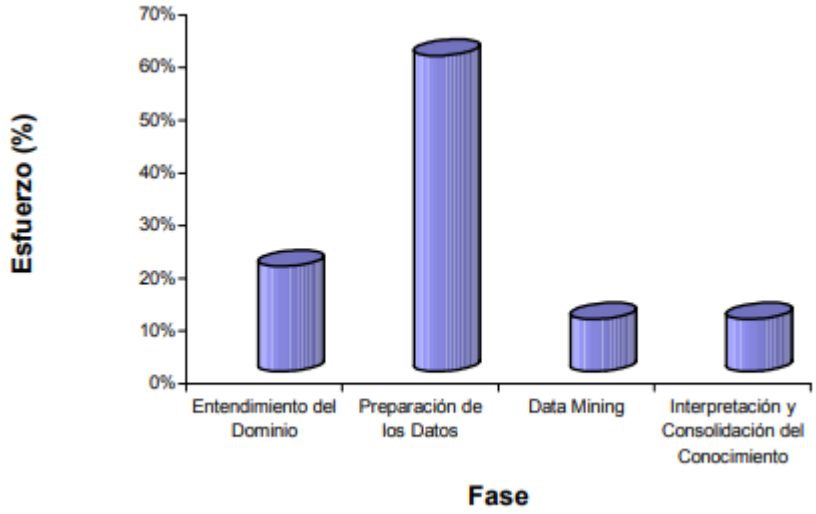
Como se vio en capítulos anteriores, la minería de datos aprovecha el esfuerzo realizado por todos los niveles de la organización. El proceso de KDD se inicia con la identificación de los datos. Para ello hay que imaginar qué datos se necesitan, dónde se pueden encontrar y cómo conseguirlos.



Entonces, la minería de datos se convierte en un proceso que se realiza posterior a la recolección de datos. Para lo cual se debe contar con sistemas de información que recaben datos en cada una de las fases que constituyen el proceso de negocio.

Esto quiere decir que, no es viable realizar un análisis correcto de información de carácter multidimensional, sin antes haber realizado una inversión para automatizar a la organización.

Adicional, al ser la minería de datos un proceso altamente técnico requiere de contar con personal cualificado y un esfuerzo en cada una de las fases del proceso.

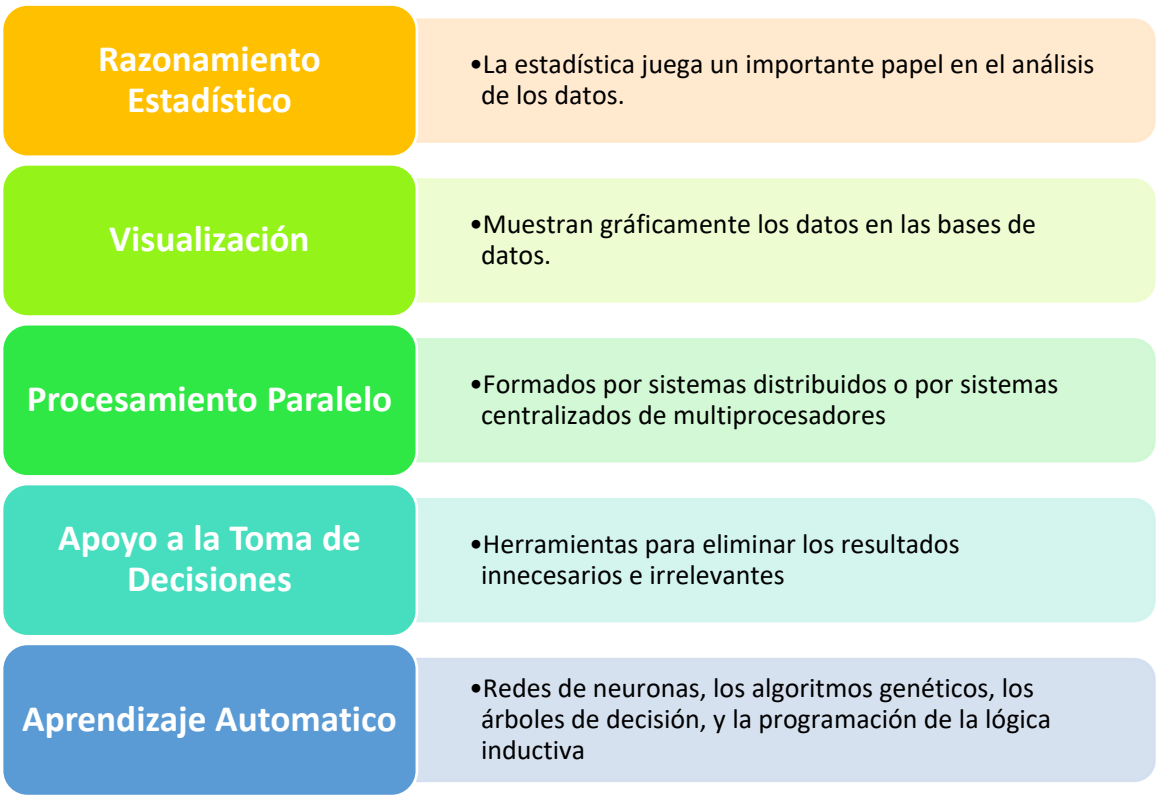


Adicional al esfuerzo e implementaciones previas de sistemas transaccionales, hay que sumarle la complejidad de decisión con respecto a que herramienta utilizar. Por una parte, existen herramientas licenciadas que brindan soporte, sin embargo, el costo es bastante elevado para la gran mayoría de empresas. Las alternativas gratuitas implican un mayor conocimiento para igualar las prestaciones un software más elaborado.

Un último aspecto para considerar es la posible necesidad de las denominadas Tecnologías de Apoyo.

El Data Warehouse es quizás la tecnología que mayor soporte le da a la minería de datos, ya que puede organizar la data de manera accesible. Adicional la posibilidad de tomar varias fuentes de datos y establecer un proceso de limpieza y eliminación de redundancia,

permitiendo enfocar los mayores esfuerzos en la visualización de resultados. Pero, adicional al Data Warehouse debemos describir alguna Tecnologías de Apoyo adicionales:



4.3.Comparativa entre herramientas licenciadas y gratuitas

En el Capítulo III se había realizado una descripción breve de diferentes herramientas para hacer minería de datos desde el enfoque del Costo, en ese sentido se había dividido a las mismas en:

Ahora, con el objetivo de brindar un enfoque más analítico, se describe a continuación la tabla resumen de las opciones que se considera más populares en la actualidad:

Tabla 3 Comparativa de herramientas de minería de datos

HERRAMIENTA	Características	Lenguaje de Programación	Sistema Operativo	Precio / Licencia
RapidMiner	Apto para todos los procesos. Destaca en el análisis predictivo	Java	Windows, MacOS, Linux	Freeware, diferentes formas de pago
SAS	Software potente para grandes empresas. Mayor inversión económica	Lenguaje SAS	Windows, MacOS, Linux	Freeware limitado a instituciones públicas. Costo se establece posterior a la solicitud
MicroStrategy	Permite a las organizaciones convertir los datos en potentes aplicaciones de BI, AI, ML y Analítica	API de Plataforma de MicroStrategy Cloud	Windows, MacOS, Linux	Freeware con funciones limitadas. Costo se establece posterior a la solicitud
Azure	Plataforma en la nube diseñada para simplificar el proceso de creación de aplicaciones modernas	C#, Java, F#	Windows, MacOS, Linux	Freeware con funciones limitadas. Costo se establece posterior a la solicitud
R	Altamente utilizado en la academia. Potente evaluador estadístico	Lenguaje de programación R	Windows, MacOS, Linux	Software libre GPL

Weka	Pueden definirse muchos métodos de clasificación	Java	Windows, MacOS, Linux	Software libre GPL
Orange	Crea vistas atractivas para visualizar datos. No requiere muchos conocimientos previos	C++ y Python	Windows, MacOS, Linux	Software libre GPL

Capítulo V: Conclusiones y Recomendaciones

5.1.Conclusiones

- Al momento de planificar la implementación de minería de datos dentro de una organización, sin duda, las herramientas licenciadas brindan entornos más enriquecidos, que brindan ayuda desde el proceso KDD hasta la interpretación. Sin embargo, son diseñadas casi exclusivamente para grandes empresas. Resulta complicado tener acceso a una cotización clara y casi todas tienen un esquema de suscripción, con lo cual, se debe contar con recursos para una inversión a largo plazo.

- Otro aspecto para considerar, si se desea implementar una herramienta licenciada, es el lenguaje de programación que soporta la plataforma a implementar. En este sentido, debido a: 1)la gran cantidad de documentación existente, 2) la posibilidad de acceder a una versión gratuita para fines académicos, 3) la opción de prueba gratis previo a la suscripción, y 4) la posibilidad de programación en Java; la opción más adecuada para nuestro país sería RapidMiner.

- El proceso de minería de datos es una labor netamente técnica que implica un alto conocimiento en bases de datos y gestión estratégica de la información. Durante este proceso es indispensable identificar los actores claves que realicen el acompañamiento, esto con el objetivo de obtener los resultados deseados.

- Previo a implementar minería de datos dentro de cualquier organización, es necesario la implementación de software transaccional que genere información de manera permanente para tener la data principal a ser analizada.

- Las opciones gratuitas para minería de datos, a diferencia de las licenciadas, son prácticamente escasas. Sin embargo, tienen aspectos muy importantes que no pueden dejarse de considerar, por ejemplo, los lenguajes de programación que soportan (Java, C++ o Python), una comunidad mundial que brinda soporte y comparte experiencias, y la proyección de futuro que siempre brinda el software libre.

- De entre las opciones de herramientas gratuitas estudiadas, llama poderosamente la atención R por los siguientes aspectos: 1) el poco uso de recursos computacionales, 2) un lenguaje nativo de programación débilmente tipado, que lo hace versátil y adaptativo, 3) la alta documentación existente, 4) la alta importancia que ha cobrado en esferas académicas, y 5) una importante cantidad de librerías disponibles para integrar en cualquier solución.

5.2.Recomendaciones

- Si se desea implementar minería de datos por primera vez se debería consultar con personal con experiencia previa y realizar un inventario de fuentes de datos disponibles, sean producto de software transaccional o fuentes de datos de terceros.

- Previo a la implementación de minería de datos, conviene establecer de manera clara los indicadores que se desean obtener e identificar a los actores claves que realizarán el acompañamiento al personal técnico.

- Al seleccionar una herramienta de minería de datos, considerar una opción licenciada, sólo si cuenta con las siguientes características: 1) un proyecto a largo plazo, 2) recursos económicos estables y permanentes, 3) alta capacidad de procesamiento, y 4) personal altamente calificado y con experiencia en gestión de proyectos.

- En caso de que la organización se incline por una herramienta gratuita de minería de datos, se debe considerar los siguientes aspectos: 1) personal altamente calificado en programación de aplicaciones y gestión de bases de datos, 2) contemplar un sistema de gestión de riesgos de la información, 3) implementar sistemas de versionamiento, y 4) la vinculación de la organización con la academia.

Bibliografía

Azure Microsoft . (1 de 06 de 2022). *azure.microsoft.com*. Obtenido de https://azure.microsoft.com/es-es/services/machine-learning/#product-overview

Belinchón Monjas, Y. (31 de Marzo de 2022). *Ingeniería de Telecomunicación.* Obtenido de https://www.it.uc3m.es/jvillena/irc/practicas/10-11/15mem.pdf

Beltrán Martínez, B. (25 de Marzo de 2022). *Benemérita Universidad Autónoma de.* Obtenido de http://bbeltran.cs.buap.mx/NotasMD.pdf

Blanco Picado, P. (25 de Febrero de 2016). *¿Cuál es la relación entre Big Data, minería de datos y Estadística?* Obtenido de Universidad de Costa Rica: https://www.ucr.ac.cr/noticias/2016/02/25/cual-es-la-relacion-entre-big-data-mineria-de-datos-y-estadistica.html#:~:text=A%20diferencia%20de%20la%20Estad%C3%ADstica,de%20datos%20confluyen%20varias%20ciencias.

Capterra. (7 de Junio de 2022). *capterra.ec*. Obtenido de https://www.capterra.ec/software/120/microstrategy-9

Cardenas, I. A., Theran Maldonado, J. D., & Muñoz Castañeda, D. L. (20 de Octubre de 2020). *¿Qué es Weka?* Obtenido de https://medium.com/@isaiaranda15/que-es-weka-926c05050d44

Cutro, A. (26 de Febrero de 2010). *dataprix.com.* Obtenido de https://www.dataprix.com/es/mineria-datos-aplicada-encuesta-permanente-hogares/171-evolucion-historia-mineria-datos

Equipo de redacción de B12 España: Marketing, Big Data, Inteligencia Artificial y Ventas. (7 de Julio de 2020). *agenciab12.com*. Obtenido de https://agenciab12.com/noticia/que-es-weka-que-tiene-que-ver-big-data

GetApp. (6 de Junio de 2022). *getapp.es*. Obtenido de https://www.getapp.es/software/100877/rapidminer

Hastie, T., Tibshirani, R., & Friedman, J. (2020). *Data Mining, Inference, and Prediction.* Springer.

Intelligent. (4 de Agosto de 2016). *Intelligent*. Obtenido de https://itelligent.es/es/10-ventajas-la-mineria-web/

Loef, N. (21 de Julio de 2015). *youtube.com: TEDx Talks*. Obtenido de ¿Querés predecir el futuro? Usá datos: https://www.youtube.com/watch?v=1iqh1B1OZAg

Marcano Aular, Y. J., & Talavera Pereira, R. (11 de Marzo de 2007). *Scielo*. Obtenido de Minería de Datos como soporte a la toma de decisiones empresariales: http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1012-15872007000100008#:~:text=La%20Miner%C3%ADa%20de%20Datos%2C%20bien,identifica%20y%20analiza%20informaci%C3%B3n%20importante.

Martínez Luna, G. L. (Julio de 2011). *revistaciencia.amc.edu.mx.* Obtenido de https://www.revistaciencia.amc.edu.mx/images/revista/62_3/PDF/mineria_aguja.pdf

Martínez Moncaleano, C. J., & Palencia Fajardo, O. (Junio de 2021). *Modelo de minería de datos para el análisis de la productividad y crecimiento personal en las mujeres emprendedoras: el caso de la Asociación las Rosas.* Obtenido de http://www.scielo.org.co/pdf/sdn/v12n26/2027-5692-sdn-12-26-23.pdf

Máxima Formación. (1 de Agosto de 2018). *maximaformacion.es*. Obtenido de https://www.maximaformacion.es/blog-dat/que-es-r-software/

Mena Roa, M. (11 de Agosto de 2021). *statista.com*. Obtenido de https://es.statista.com/grafico/17539/datos-creados-online-en-un-minuto/

MicroStartegy. (25 de Mayo de 2022). *microstrategy.com/es*. Obtenido de https://www.microstrategy.com/es

Orange Data Mining. (15 de 06 de 2022). *orangedatamining.com*. Obtenido de https://orangedatamining.com/

Párraga Alava, J. (1 de Mayo de 2021). *Minería de Datos (Clase 1: Introducción y proceso KDD)*. Obtenido de https://www.youtube.com/watch?v=DmNqizRa1YQ&list=PLBedOvLVWMxwIUMIgh_kVn4Wgp1QYqH2X

Pérez Marqués, M. (2015). *MINERÍA DE DATOS A TRAVES DE EJEMPLOS.* México: Alfaomega Grupo Editor, S.A. de C.V., México.

R Porject. (06 de Junio de 2022). *r-project.org*. Obtenido de https://www.r-project.org/

Rapid Miner. (10 de Junio de 2022). *rapidminer.com*. Obtenido de https://rapidminer.com/

SAS. (19 de Mayo de 2022). *sas.com*. Obtenido de https://www.sas.com/es_mx/home.html

Vivas, F. (17 de Enero de 2020). *youtube.com: TEDx Talks*. Obtenido de Ciencia de Datos: ¿La Profesión Más Sexy del Siglo 21?: https://www.youtube.com/watch?v=AaoM5XhdnG0

Waikatpo. (10 de 06 de 2022). *Weka 3: software de aprendizaje automático en Java*. Obtenido de https://www.cs.waikato.ac.nz/ml/weka/