# Herramientas de analítica para la explotación de datos

**09.07.2018**

**Versión 1.0**

GOBIERNO DE COLOMBIA

MINTIC

GOBIERNO DIGITAL

BIOS

**Ministerio de Tecnologías de la Información y las Comunicaciones**
**Viceministerio de Economía Digital**
**Dirección de Gobierno Digital**

## Equipo de trabajo

Carlos Eduardo Rozo Bolaños – Director de Gobierno Digital
Jaime Enrique Cusba – Subdirector de Gobierno en línea
Luisa Fernanda Medina Martínez – Líder de Datos y Soluciones Abiertas
Carlos Julio León Caicedo – Líder Técnico de Datos y Soluciones Abiertas

## Centro de Bioinformática y Biología Computacional – BIOS

## Equipo de trabajo

Dany Leon Molina Orrego – Director ejecutivo BIOS
Eduardo Gómez Restrepo – Coordinador de Bionegocios BIOS
Yohan Ricardo Céspedes Villar – Asesor BIOS
Juan Salvador Estrada – Asesor BIOS

| Versión | Observaciones |
|---|---|
| Versión 1<br>Octubre 2018 | **Inventario de herramientas libres de analítica**<br>Dirigido a estudiantes, emprendedores, ciudadanos, entidades del Estado y demás actores del ecosistema de datos |

Comentarios, sugerencias o correcciones pueden ser enviadas al correo electrónico: gobiernodigital@mintic.gov.co

Herramientas de analítica para la explotación de datos

# Tabla de contenido

# 1. Introducción

Nuevas tecnologías y modelos de servicio están transformando las expectativas de los usuarios frente al modo de interacción con las empresas de diversos sectores, creando nuevos y diferentes modelos organizacionales y de negocio, además de crear un panorama competitivo cambiante en las diferentes industrias. Esto hace que grandes compañías se adapten para entrar en sectores no tradicionales para ellas y/o realizar expansiones geográficas con modelos de negocio diferentes. Esto se ve reflejado en el uso de datos abiertos que hacen diferentes organizaciones para realizar la expansión hacia otros mercados geográficos. Esta revolución tecnológica está afectando la forma en que los usuarios tienen que evolucionar, para interactuar de manera diferente y cubrir sus expectativas crecientes, creando consumidores de datos abiertos más empoderados.

Cuatro tendencias tecnológicas son las que están entrando con fuerza y están cambiando el comportamiento de los consumidores de estos datos:

1. La irrupción y fortalecimiento de los **canales digitales** permite que los consumidores requieran servicios cada vez más complejos y busquen interactuar de una manera más ágil y eficiente con menores tiempos de respuesta.

2. El **crecimiento de las redes sociales** como nueva forma de comunicación y de relación permite a los consumidores generar opiniones de forma viral y acceder a nuevos canales de servicio y consumo. A su vez para las organizaciones es una **fuente de información** muy valiosa, casi inexplorada para interactuar con el cliente con un conocimiento profundo de la situación y necesidades de estos. Estudios indican que el 90% de la información que tenemos actualmente ha sido generada en los dos últimos años y muchos a través de datos abiertos disponibles a los usuarios (OpenMind, 2018).

3. El desarrollo de nuevos servicios en la nube permite tener servicios más agiles, escalables y fáciles de evolucionar e innovar, produciendo una carrera en la búsqueda de nuevas formas de interactuar con el cliente.

4. Adicionalmente las **nuevas tecnologías de computación cognitiva** están revolucionando la forma como se llega al cliente. **International Data Corporation indica que para el 2020 el 50% de los consumidores interactuarán regularmente con servicios cognitivos**. Las organizaciones pondrán a disposición de usuarios y clientes, las capacidades de lenguaje natural, procesamiento de información estructurada y no estructurada, con entendimiento de contexto, con análisis sentimental de la información, para dar una mejor asesoría al cliente. La computación cognitiva será el centro neurálgico de las organizaciones y modo de interacción de los clientes en los canales digitales, así como asesor de los canales físicos. (Pat Research, 2018)

Esta **nueva norma en la experiencia del cliente,** producida por el advenimiento de nuevas tecnologías, hace que las entidades del estado necesiten como pilar fundamental poner a disposición no sólo datos abiertos sino también información no estructurada, esto en conjunto con la aplicación de **capacidades analíticas y cognitivas** que permitan impactar los objetivos estratégicos, el cumplimiento de la misión y la mejora en la satisfacción de los servicios prestados a la ciudadanía.

En consonancia con el **Conpes 3920 de 2018** de la Política Nacional de Explotación de Datos (Big Data), cuyo objetivo es el de aumentar el aprovechamiento de datos en Colombia, mediante el desarrollo de condiciones para que sean gestionados como activos y generar valor social y económico; a través de líneas de acción como por ejemplo, la línea de acción número 13 del Conpes, la cual tiene por objetivo, incentivar el uso de los bienes y servicios basados en datos, así como permitir la toma de decisiones basadas en estos por parte de ciudadanos, poniendo al alcance de los mismos el concepto la analítica y fomentar el uso de la explotación de datos, así como la apropiación de los productos que ésta genera, orientado a dinamizar el ecosistema de datos en el país.

Por lo anterior, en cumplimiento con el **Conpes de explotación de datos** de incentivar la apropiación digital del concepto de analítica y poner a disposición de estudiantes, emprendedores, actores del ecosistema de datos, y a ciudadanos en general, un conjunto de herramientas libres utilizadas en la construcción de analítica, el Ministerio TIC ha realizado un diagnóstico e inventario de las herramientas más utilizadas en el mundo, con base en los estudios realizados por Gartner una empresa consultora y de investigación de las tecnologías de la información, y lo ha consolidado en este documento, en donde se pretende además de introducir al lector en el concepto de analítica, herramientas y tecnologías, proporcionar la información suficiente para la elección y apropiación de alguna de las plataformas aquí presentadas.

# 2. Marco conceptual

## 2.1. Analítica y Big Data

Los datos que almacenan las entidades del estado se generan rápidamente y son complejos, sin embargo, dentro de éstos está contenida una gran cantidad de información sumamente útil para la toma de decisiones; es por eso por lo que resalta la importancia de analizar e interpretar estos datos y convertirlos en información que las entidades del estado pueden utilizar a su favor para mejorar su funcionamiento y ser más eficientes.

Al hablar del análisis de datos, se hace referencia a la revisión de los datos con el fin de extraer conclusiones de la información. Esta actividad permite mejorar la toma de decisiones de las empresas, entidades e instituciones en general.

El análisis de datos es esencial para el cumplimiento de los objetivos de todos los organismos gubernamentales, ya que permite obtener resultados más efectivos y de una manera más eficiente.

El creciente volumen de datos y su naturaleza cada vez más diversa permiten contar con más información enriquecedora para la toma de decisiones, sin embargo, pueden también paralizar la organización, ya que la cantidad tan grande de información ha provocado que los medios tradicionales de almacenamiento y procesamiento sean insuficientes, y es aquí donde entra el procesamiento de Big Data, ya que permite el almacenamiento y la examinación de grandes volúmenes de datos a gran velocidad, con el objetivo de extraer patrones y relaciones desconocidas y útiles para la comprensión del funcionamiento y mejora de la institución; estas mejoras se traducen en la agilización de la toma de decisiones para la generación de valor social y económico, ya sean productos, servicios o procesos.

Es importante mencionar que, para poder llegar a los resultados deseados, es necesario pasar por distintas etapas (CONPES, 2018), mismas que se mencionan a continuación:

- Generación y recolección: Consta de la creación de datos digitales por parte de las personas como resultado de su interacción con sistemas, herramientas y servicios digitales, o por máquinas con software y dispositivos que capturen fenómenos.
- Compartición y agregación: Consiste en la disponibilidad de los datos para que sean conocidos y usados mediante procesos y plataformas.
- Explotación: Es aquí donde entra el análisis de los datos mediante técnicas científicas y herramientas automatizadas, mismas que permiten generar información que a su vez deriva en conocimiento respecto de los fenómenos de la realidad, ya que antes de pasar por esta etapa la información permanecía oculta debido a su complejidad.
- Innovación: Corresponde a la generación de bienes, servicios y procesos que aporten a la diversificación y sofisticación de la economía y a la generación de valor social como una fuente de crecimiento.

El gobierno puede hacer uso de las tecnologías de Big Data y de analítica para obtener predicciones y prevención de crímenes, poder anticipar tasas de desempleo, alerta temprana sobre enfermedades, entre otros. Los beneficios del uso correcto de los datos son extremadamente amplios y cualquier institución puede hacer uso de ellos para poder crear análisis estratégicos para la toma de decisiones.

## 2.2. Universo de datos

Los datos corresponden a la representación primaria de variables cualitativas y cuantitativas que son almacenables, transferibles y que pueden ser visualizadas, controladas y entendidas (CONPES, 2018). Las fuentes de datos que utiliza el Big Data van desde la información producida por medios de comunicación, pasando por los datos generados por sensores, hasta los registros de la navegación de los clientes y la información generada en redes sociales.

Dentro de las fuentes de datos utilizadas están los datos estructurados, mismos que se encuentran en un formato bien definido; en estos se encuentran en las bases de datos relacionales y hojas de cálculo.

Por otro lado, se encuentran los datos no estructurados, dentro de los cuales no existen campos definidos como en los datos estructurados. Dentro de éstos se encuentran las imágenes, videos, audios, formatos de texto, mensajes, artículos, correos electrónicos, entre otros.

## 2.3. Perfil del científico de datos

En los últimos años, debido a la inercia que ha tenido el uso creciente de los datos para la toma de decisiones, se han creado distintos cargos con capacidades tales como: el manejo de herramientas de Big Data para el procesamiento de información, conocimientos de programación, dominio en el uso de bases de datos (tanto relacionales como no relacionales), manejo de herramientas de visualización, capacidades analíticas y el uso de herramientas de machine learning; algunos de los más representativos (CAOBA et. al., 2017) son:

- Chief Data Officer (CDO): Miembro de la dirección ejecutiva encargado de liderar la gestión de datos y analítica asociada con el negocio, es responsable de los diferentes equipos especializados en datos en la empresa.
- Data Scientist (científico de los datos): Interpreta y transforma los grandes volúmenes de datos en información útil para la empresa. Se caracteriza por tener habilidades en matemáticas, estadística, programación, es creativo y cuenta con habilidades comunicativas para explicar los resultados de su trabajo.
- Citizen Data Scientist: En palabras de Gartner" El citizen Data Scientist es una persona que crea o genera modelos que aprovecha el análisis predictivo o

prescriptivo, pero cuya principal función de trabajo se encuentra fuera del ámbito de la estadística y análisis".

- Data Engineer: Es el responsable de proporcionar los datos requeridos por el científico de datos. Tiene gran conocimiento en bases de datos, arquitecturas de clúster, lenguajes de programación y sistemas de procesamiento de datos.
- Data Steward (administrador de datos): Se encarga de mantener la calidad, disponibilidad y seguridad de los datos. Posee conocimientos en los procesos del negocio e identifica cómo son usados dentro de la empresa.
- Business Data Analyst (analista de datos): Participa en el análisis de los datos con el fin de recolectar las necesidades del cliente para sustentarlas de manera clara al científico de datos.
- Data Artist: Es un experto en Business Analytics, encargado de mostrar de manera sencilla (gráficos, infografías y herramientas visuales), los resultados del análisis de los datos para comprender grandes volúmenes de información.
- Estadístico: Sus funciones se centran en obtener, analizar e interpretar datos cualitativos y cuantitativos usando los métodos estadísticos.
- Administrador de bases de datos: Tiene conocimientos fuertes en el manejo de bases de datos, típicamente relacionales (CAOBA et. al. , 2017).

El perfil de Big Data y análisis de datos, cuenta con dominios en distintas áreas, mismos que permiten expresar un perfil completo; estos dominios se presentan a continuación:
“

- **Negocio:** Definición y comprensión de la temática específica del proyecto, que está definida por las necesidades y reglas del negocio.
- **TI:** Administración de la tecnología de TI, despliegue de la solución y manejo del ciclo de vida de los datos, al igual que temas transversales de los mismos.
- **Analítica:** Selección y análisis de los datos de forma apropiada, al igual que selección y construcción de los modelos adecuados para la solución.
- **Administración de proyectos:** Gestión del proyecto, al igual que la gestión del equipo de trabajo colaborativo.
- **Habilidades transversales:** Conjunto de habilidades que apoyan el desarrollo efectivo de un proyecto de BD&DA

” (CAOBA et. al. , 2017)

La interacción de personas con distintos perfiles dentro de un equipo de trabajo permite el correcto funcionamiento e integración de las habilidades individuales para un fin común. “Existen 4 perfiles distintos: el perfil núcleo (P0), el de negocio (P1), el de tecnologías de la información (P2), el de analítica (P3) y el de administración de proyectos (P4). Para comprender la definición de cada uno de los perfiles, es necesario comprender la taxonomía de Bloom” (CAOBA et. al. , 2017), ya que permite identificar el grado de conocimiento dentro de cada área. Su estructura es la siguiente:


*Figura 1: Pirámide de conocimiento, taxonomía del Bloom*

### 2.3.1. Perfil Núcleo – P0

El perfil núcleo es el que facilita la comunicación interna del equipo ya que maneja un lenguaje común, conoce y comprende los conceptos de los perfiles especializados, lo que facilita la toma de decisiones dentro de cada proyecto.

"El tener conocimientos y habilidades en los 2 primeros niveles de Bloom permite al profesional con el perfil núcleo estar en la capacidad de conocer la idea del negocio, las posibles herramientas en TI a usar, los modelos matemáticos existentes y los procedimientos para la administración de proyectos que junto con habilidades (transversales) como trabajo en equipo, comunicación, responsabilidad, liderazgo, disciplina, y pasión entre otras que le permiten comprender las decisiones y las consecuencias de las decisiones que se toman durante el proyecto (desde la concepción hasta el cierre)" (CAOBA et. al. , 2017).

### 2.3.2. Perfil de Negocio – P1

Por su parte, el perfil de negocio (P1) se caracteriza por su capacidad de actuación debido a su conocimiento del negocio y de la organización, lo que le permite tomar decisiones cruciales en el desarrollo de los proyectos. Sus conocimientos le permiten realizar análisis costo-beneficio, así como de las herramientas que pueden ayudar dentro del proyecto.

### 2.3.3. Perfil en TI – P2

En cuanto al perfil en TI (P2) es el que ejecuta las acciones y toma las decisiones respecto al proyecto, con base en las tecnologías de información. Es un experto en tecnologías de la información que soportan el análisis de datos y Big Data, por lo que es el encargado de la administración de la infraestructura, así como del despliegue de la solución (CAOBA et. al.,

2017). Algunas de sus habilidades están encaminadas a la manipulación de datos, en términos de su ciclo de vida y calidad. Cuenta con liderazgo en temas de lenguajes, herramientas y metodologías en temas relacionado con TI.

### 2.3.4. Perfil en Analítica – P3

El perfil en analítica (P3) cuenta con habilidades de manejo de los datos fuente, los explora, selecciona, prepara y visualiza en modelos analíticos. También tiene la capacidad de comunicar los resultados de la solución construida, así como habilidades para la selección de herramientas y metodologías apropiadas para cada problema.

### 2.3.5. Perfil en Administración de proyectos – P4

Finalmente, el perfil en administración de proyectos (P4) "tiene mayores habilidades en el conocimiento de gestión de proyectos en cuanto a la planeación, tiempos, manejo de recursos y estructuración de las acciones a realizar en pro de cumplir los acuerdos realizados con los distintos actores involucrados en el proyecto. Específicamente, conoce y comprende de forma detallada cómo gestionar un proyecto, los recursos y el presupuesto disponible, y además es capaz de organizar con el uso de herramientas y conocimiento en metodologías la planeación, ejecución, monitoreo y ajuste de tareas antes, durante y después del desarrollo del proyecto. Incentiva y monitorea adicionalmente el trabajo en equipo" (CAOBA et. al., 2017).

## 2.4. Ciclo de vida del proceso analítico

Las soluciones de analítica presentadas en este documento responden a una visión de los procesos que deben internalizar las entidades para desarrollar de modo continuo una capacidad o inteligencia que le permita tomar decisiones e implementar acciones en base a las ideas principales que surgen del análisis de los datos.

A continuación, se presentan los niveles de madurez de la Analítica en las organizaciones:


*Figura 2: Niveles de la Analítica (SAS – 2018)*

La analítica se puede conceptualizar cómo una serie de actividades o fases de trabajo iterativas que en conjunto representan un ciclo analítico continuamente retroalimentado de nuevas preguntas y desafíos de negocios, cómo de los resultados y experiencias de las acciones en curso. Gráficamente el ciclo analítico puede pensarse como una secuencia:

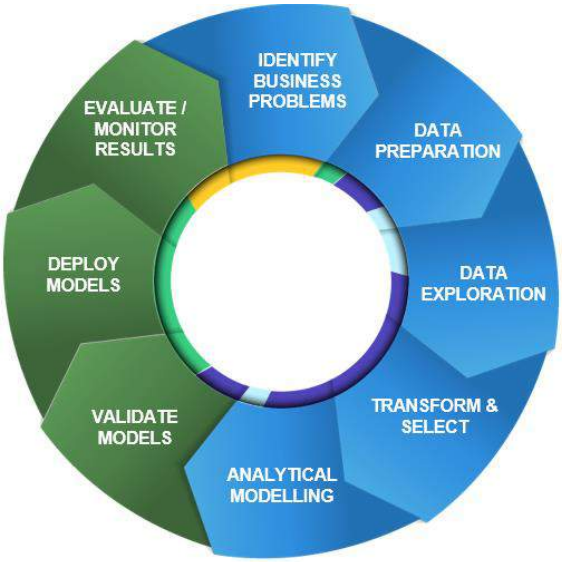


*Figura 3: El ciclo Analítico (SAS – 2018)*

Dentro de esta serie de fases o actividades, hay dos grandes agrupamientos:

- Las fases de exploración y de aprendizaje (color azul):

- Identificación y formulación de problemas (Identify business needs).
- Preparación y selección de datos para el análisis del problema (Data Preparation).
- Exploración y análisis preliminares (Data Preparation).
- Revisión, ajuste y adecuación de los datos (Transform & Select).
- Desarrollo de modelos analíticos (Analytical Modelling).

- Las fases de implementación y de acción (color verde):

    - Validación de modelos (Validate Models).
    - Implementación de modelos (Deploy Models).
    - Ejecución, evaluación, monitoreo y análisis de resultados (Evaluate, Monitor & Results).

Cómo se mencionaba anteriormente, hay una idea implícita en este ciclo, y es la del aprendizaje continuo. El ciclo analítico comienza con un problema, una idea o inquietud, que deriva en un modelo de la realidad o la generalización de una regla, que será utilizada para tomar decisiones y actuar sobre la realidad. De la implementación de estos modelos y reglas, surgen los resultados que generan nuevas preguntas e inquietudes que se traducirán en nuevos problemas originando así nuevos ciclos de aprendizaje.

Otro aspecto para tener en cuenta es que esta abstracción del ciclo analítico no es un caso aislado dentro de la organización, sino que en paralelo son múltiples los ciclos que ocurren en simultáneo para responder a necesidades de negocio diferentes.

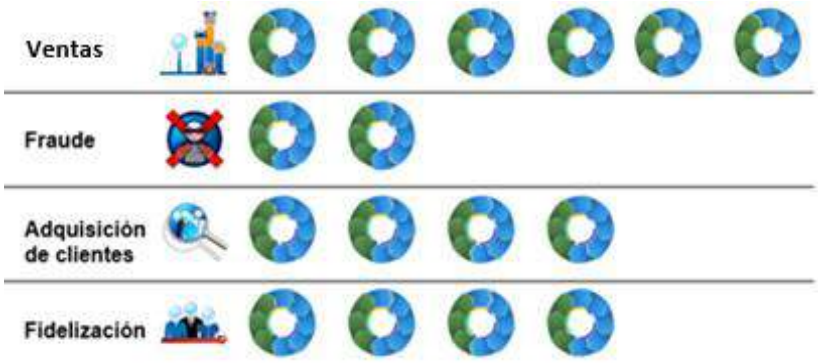


*Figura 4: Múltiples ciclos analíticos dentro de la organización (SAS – 2018)*

Esto lleva a poner relevancia la necesidad de:

- Contar con soluciones de analítica robustas para soportar el desarrollo descentralizado de modelos.
- Disponer de soluciones que permitan administrar y gestionar estos ciclos en simultáneo, para un mejor control y gobierno de los procesos analíticos.

## 2.5. CONPES 3920 de 2018

Hoy en día la economía mundial se caracteriza por reconocer el valor, relevancia y vigencia que los datos tienen en prácticamente todas las actividades humanas. Su análisis permite descubrir la cura a diferentes enfermedades, impulsar avances científicos, identificar riesgo sociales y ambientales, reconocer tendencias de mercado, comportamiento del consumidor, segmentar la población, optimizar y automatizar procesos, tomar decisiones empresariales y estatales.

Reconociendo esta realidad y la oportunidad para producir beneficios sociales y económicos, Colombia se une a las naciones más desarrolladas del planeta, teniendo en cuenta que, en los últimos seis años, ocho de las principales economías del mundo avanzaron en una política pública nacional de aprovechamiento de datos, para adecuar las condiciones de cada país con el objetivo de maximizar los beneficios sociales y económicos y minimizar los riesgos que se asocian a la explotación de datos[1].

Es así como el Conpes (Consejo Nacional de Política Económica y Social) en su documento 3920 de 2018 de política nacional de explotación de datos, propone una serie de elementos, políticas, acciones y metas para regular e incentivar el uso y explotación de datos, acciones para promulgar una cultura de datos y cerrar la brecha de habilidades analíticas. Dentro de estas habilidades analíticas, una de las oportunidades de mejora que se ha detectado, es la falta de claridad, disponibilidad y ordenamiento de las diferentes herramientas analíticas que sean de acceso libre o bajo licenciamiento y que permitan y faciliten la materialización de beneficios por el uso de estas.

## 2.6. Tipos de licenciamiento de herramientas libres

Antes de hablar de tipos de licenciamiento, es importante mencionar que existen dos tipos de software: propietario y libre. El primero corresponde al programa cuyo uso, redistribución o alteración está prohibida, o que en su defecto se requiere un permiso por parte del titular del software.

Por otro lado, el software libre es un programa en el que el usuario tiene ciertos privilegios sobre el mismo, permitiéndole hacer uso, copiarlo, modificarlo y distribuirlo a su gusto. Este tipo de software es mucho más económico que el propietario, y otorga una gran libertad para el crecimiento y mejora del mismo gracias a la colaboración de múltiples usuarios o comunidades de desarrolladores.

Algo que es importante resaltar, es que a pesar de que el software libre permite modificaciones y distribución de este, estos programas pueden tener ciertas limitantes

[1] Estos son: Estados Unidos, Unión Europea, Australia, Corea del Sur, Reino Unido, Francia, Japón, y China, Tomado del Conpes 3920 de 2018 – Política Nacional de Explotación de Datos

estipuladas en el tipo de licenciamiento que tiene. A continuación, se mencionan algunos de los principales tipos de licencias para software libre existentes en el mercado.

- **Licencia GPL/GNU GPL:** Es una de las más utilizadas y se caracteriza porque el desarrollador conserva sus derechos de autor, y permite su uso, distribución y modificación, con la única condición de que cuando el software se modifique, la herramienta resultado quede bajo la misma licencia. Es un software totalmente gratuito y es posible pagar únicamente por gastos de copiado y distribución.
- **Licencia AGPL:** Además de las cláusulas incluidas en una GNU GPL, considera que, si se quiere usar como parte de un software nuevo, éste está obligado a su distribución libre.
- **Licencia BSD:** Esta licencia es la menos restrictiva, permite la venta de nuevas versiones y pueden ser distribuidas bajo otras licencias.
- **Licencia Apache:** Permite la distribución, modificación y la distribución de versiones modificadas conservando el copyright y la renuncia de responsabilidad, de tal manera que los desarrolladores deben incluir dos archivos en el directorio principal de los paquetes de software: la copia de la licencia, y un documento de texto que incluya los avisos obligatorios del software presente en la distribución (BBVA, 2014).
- **Licencia Creative Common:** Es una combinación de cuatro condiciones: atribución (siempre se debe reconocer al autor), no comercial (no permite su uso comercial), no derivadas (no se puede modificar) y compartir igual (considera la creación de obras derivadas siempre que mantengan la licencia original).
- **Licencia Gratuita**: El usuario puede descargarla sin ningún tipo de costo.

Por otro lado, las formas de distribución del software son:

- **Shareware**: En esta forma de distribución el usuario puede evaluar de forma gratuita el producto, con ciertas limitaciones en tiempo o formas de uso, o en las capacidades finales. Para desbloquear el uso completo del software es necesario realizar un pago.
- **Freeware:** Es distribuido libre de pagos, comúnmente sin fechas de expiración y con funcionalidad completa. El autor mantiene el copyright del programa.
- **Donationware:** Es un Freeware que se distribuye con un recordatorio regular una solicitud para que los usuarios hagan una donación al autor o a un tercero como una organización benéfica.
- **Crippleware:** Son versiones básicas y limitadas de un programa: que tienen menos funcionalidades, consumen menos memoria, ocupan menos espacio en disco duro..., etc., que el programa completo original, siendo generalmente gratuitas
- **Free Software** (software libre): Da al usuario la libertad de ejecutar, copiar, distribuir, cambiar y mejorar el software. Cualquier versión redistribuida debe distribuirse bajo los términos originales de uso, modificación y distribución.

Es importante mencionar la diferencia entre el Free Software y el Open Source, que, aunque son conceptos muy cercanos, no son idénticos. El software Open Source, a diferencia del

Free Software, permite a los usuarios acceder al código fuente, pero bajo un copyright, principalmente con el objetivo de que la comunidad ayude a eliminar errores y proporcionar un producto más útil.

A continuación, se presenta una tabla con las principales diferencias de los softwares presentados.

| Tipo | Tiene costo | Permite modificación y redistribución | Funcionalidad completa | Expira |
|---|---|---|---|---|
| **Shareware** | No inicialmente, pero puede haber un cargo posterior | No aplica | | Si |
| **Freeware** | No, aunque los Donationware mandan recordatorios para hacer alguna donación | No | Usualmente si, los que no, son denominados Crippleware | No |
| **Free Software** | Algunas pueden tener costo | Si | Si | No |

# 3. Herramientas de analítica

En la actualidad, gracias al crecimiento del internet y los avances en TI, se han generado una gran cantidad de datos en el mundo, desde los generados por las empresas hasta los generados a través de la navegación y el comportamiento de los usuarios en internet. Estos datos cubren un amplio espectro de información que puede ser útil para la toma de decisiones, y es aquí donde entra en acción la analítica avanzada.

Existen diversas técnicas de analítica avanzada para cumplir con distintos objetivos que van desde la reducción de dimensión, hasta la clasificación y el pronóstico de comportamientos, mismas que han crecido junto con el uso del internet y el almacenamiento masivo de datos.

Actualmente son muchas las herramientas y softwares que permiten realizar este tipo de análisis, así como la creación de modelos, de los cuales se mencionarán algunos de los más relevantes en las siguientes secciones.

## 3.1. Análisis de tendencias de uso de herramientas de analítica

Gartner es una empresa consultora y de investigación de las tecnologías de la información con sede en Estados Unidos, proporciona análisis de investigación y consejos para profesionales de las tecnologías de la información y comunicación y para empresas de tecnología. Una de las maneras de presentar sus análisis es el cuadrante mágico de Gartner, que es el resultado de la evaluación de diversas plataformas, en las que se muestran las principales herramientas o software dentro de una categoría; en este caso, las plataformas para ciencia de datos y machine learning.

Este gráfico es una herramienta que permite identificar las plataformas con mayor visión de negocio y una mejor adaptación para utilizar las distintas herramientas que posee cada software; en pocas palabras son herramientas que lideran el mercado dentro de su categoría.

Por ejemplo, en su versión de 2018, éste cuadrante muestra que las mejores plataformas para la **ciencia de datos** y machine learning pertenecen a KNIME, Alteryx, H2O.ai, SAS, RapidMiner, TIBCO Software, IBM, Microsoft, Domino, MathWorks, Databricks, SAP, Angoss y Anaconda, plataformas que se presentarán más adelante.


*Figura 5: Cuadrante mágico de Gartner para plataformas de ciencia de datos y machine learning, 2018*

La manera de leer este gráfico es la siguiente: en el eje horizontal se tiene la visión de la empresa, y en el eje vertical la habilidad de adaptación de sus herramientas, de tal manera que las empresas que se encuentran en el cuadrante superior derecho son aquellas con mayor visión y las que han sabido ejecutar mejor su negocio en esta área (ciencia de datos y machine learning) y por lo tanto son consideradas las empresas líderes. Por su parte, las del cuadrante superior izquierdo son consideradas retadoras ya que se han desempeñado bien, sin embargo, no tienen tanta visión como las líderes. Las visionarias se encuentran en la parte inferior derecha y las empresas jugadoras de nicho en el lado izquierdo.

Además de Gartner, existen otras empresas que realizan investigaciones similares, una de las cuales es Forrester Research, que es una empresa independiente de investigación de mercados que brinda asesoramiento sobre el impacto existente y potencial de la tecnología a sus clientes y al público en general, uno de sus informes es la Forrester Wave que es una metodología objetiva de software, hardware o servicios, que evalúa la oferta actual, estrategia y presencia en el mercado de los proveedores frente a criterios de evaluación predefinidos. Forrester Wave expone informes y hojas de cálculo donde aparecen los criterios utilizados para calificar las ofertas de los proveedores recogidos en el mercado.

El proceso de investigación de Forrester Wave se basa en la participación de cuatro actores clave:

- **Analista**: Es el experto en contenido de Forrester Wave. El analista determina los criterios de inclusión y evaluación, y el marco de puntuación basado en la metodología de Forrester Wave y añade su conocimiento del mercado y de las necesidades de los clientes.
- **Investigador asociado**: Gestiona el proceso de Forrester Wave. El investigador impulsa y mantiene el cronograma del proyecto Forrester Wave y se comunica con los vendedores durante todo el proceso.
- **Equipo del proveedor**: Proporciona información detallada de los productos y del servicio. El equipo del proveedor comprende los contactos clave de los vendedores participantes con los que se trabaja.
- **Referencias de clientes**: Comparten sus experiencias con los productos de forma anónima. Forrester solicita un máximo de tres referencias de clientes de los participantes para las evaluaciones de software y hardware, y diez referencias de clientes para las evaluaciones de servicios. Cada una de las referencias es contactada por el equipo de investigación para programar una breve llamada o completar una encuesta. Forrester realiza preguntas a los clientes en función de su experiencia con el servicio o producto que se evalúa. Las referencias de clientes no están listadas ni son nombradas en los materiales publicados a menos que se acuerde específicamente antes de su publicación.

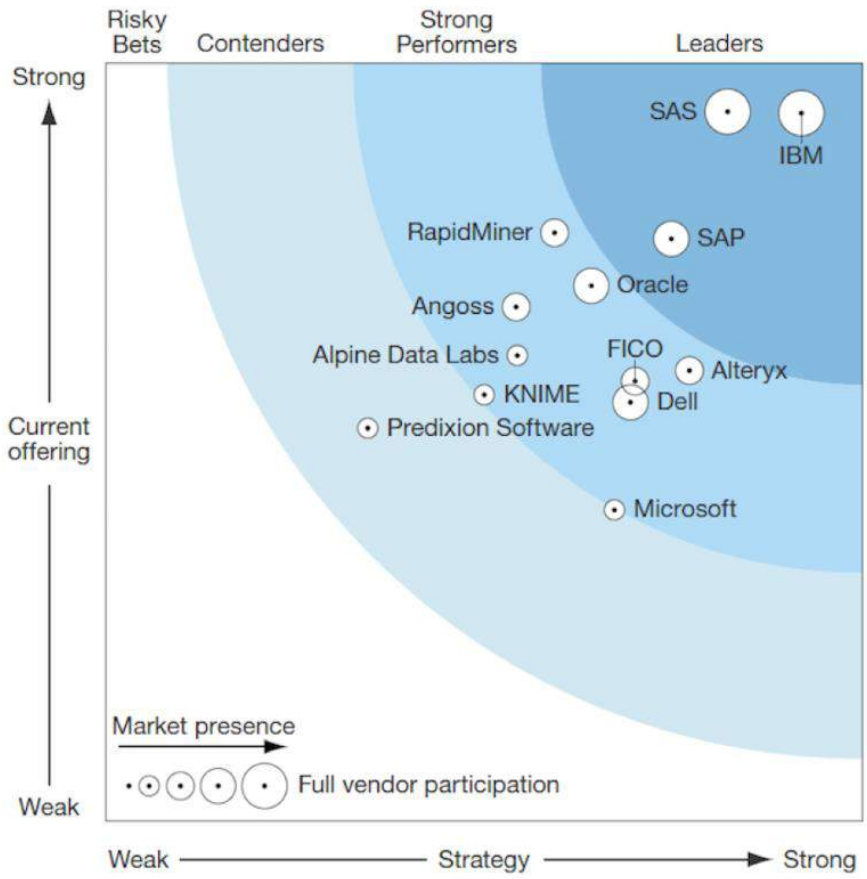


*Figura 6: Tendencias de Herramientas de Analítica por estudios de Forrester*

De manera similar al cuadrante mágico de Gartner, Forrester Wave muestra en el eje horizontal la estrategia de las empresas y en el vertical su oferta actual. La segmentación que hace es respecto a "ondas", donde a partir de la esquina superior derecha se encuentran las empresas líderes, seguidas de empresas fuertes, contendientes y finalmente apuestas riesgosas. Además, el tamaño del círculo indicador de cada empresa muestra la presencia de mercado de ésta.

Por otro lado, dentro de las herramientas en software libre más utilizados para la analítica están R y Python y existe una fuerte discusión entre la comunidad de cuál es el mejor. Mientras que R es principalmente utilizado por académicos e investigadores, Python es usado por programadores que quieren ahondar en análisis de datos o aplicar técnicas estadísticas, o desarrolladores que se convierten en científicos de datos.

A continuación, se muestran algunos gráficos respecto a la popularidad que han tenido estas dos plataformas:

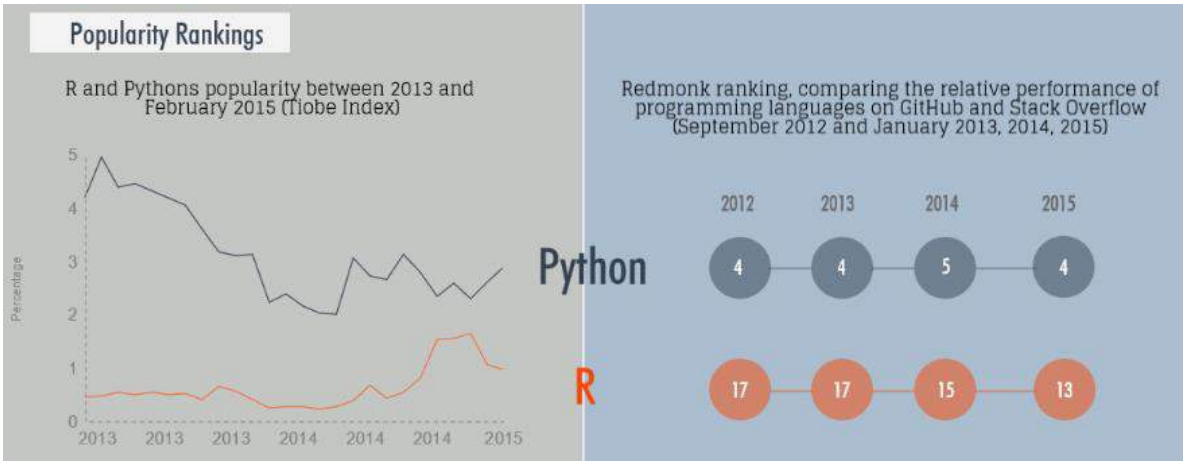


*Figura 7: Popularidad de las plataformas R y Python*

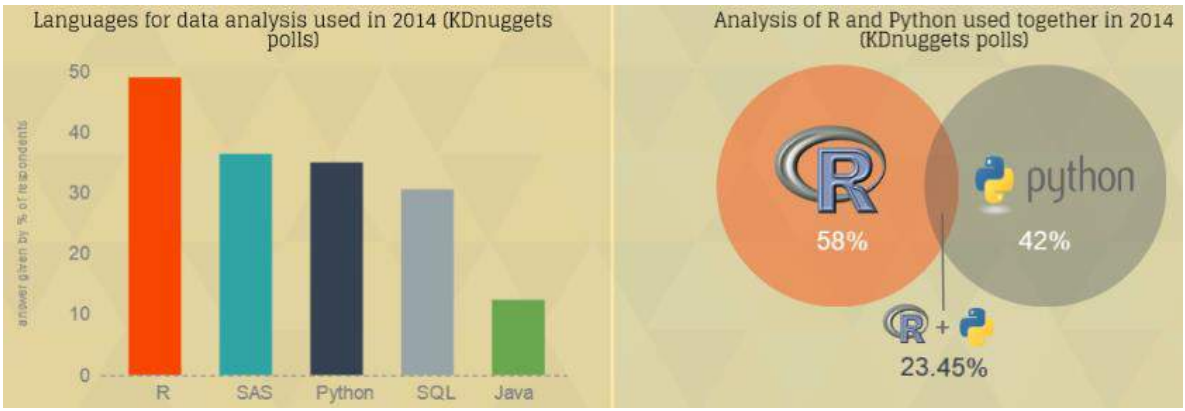


*Figura 8: Uso de lenguajes para análisis de datos en 2014*

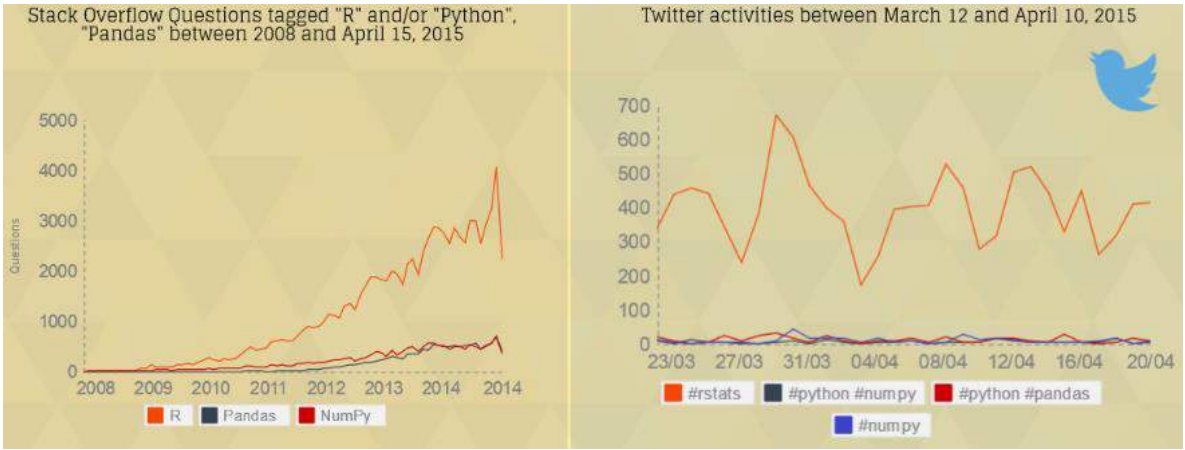


*Figura 9: Dinamismo de la comunidad de R y Python*

R se ocupa principalmente cuando se requieren cálculos solitarios o análisis en servidores individuales, mientras que Python se utiliza más cuando las tareas requieren ser integradas en aplicaciones web o si el código requiere ser insertado en las bases de datos de producción.

Ambas plataformas cuentan con pros y contras. Por ejemplo, Python resalta por su código sencillo, donde cualquier funcionalidad se escribe siempre de la misma forma, situación que no ocurre con R. Por su parte, en el caso de R, una vez aprendidas las bases es fácil desarrollar aplicaciones avanzadas.

Las ventajas de R se encuentran en sus capacidades gráficas y su ecosistema. Fue desarrollado por y para estadísticos, por lo que tiene un gran uso entre ingenieros, estadísticos y científicos. Por otro lado, una de sus desventajas es que es lento ya que le exige mucha capacidad a la computadora.

Las ventajas de Python son su IPython Notebook, contar con programas ya escritos, texto y ecuaciones para documentación en un ambiente. La lectura del lenguaje Python es muy simple y su curva de aprendizaje es muy ligera; Así mismo es una herramienta que se integra en cualquier parte del flujo de trabajo. Por otro lado, las visualizaciones en este lenguaje son complejas, y el resultado no necesariamente es el deseado, así mismo no existen paquetes en esta plataforma que reproduzcan lo que hacen paquetes relevantes en R.

A continuación, se presentan el listado de herramientas libres y propietarias identificadas y recomendadas en los procesos de analítica básica y avanzada en las organizaciones.

## 3.2. Herramientas libres

### 3.2.1. Programación R

R es uno de los lenguajes y entornos de programación para análisis estadístico y gráfico más utilizado en el campo de la minería de datos, que puede aplicarse a gran variedad de disciplinas.

Es un proyecto colaborativo y abierto por lo que los desarrolladores pueden descargar el código de forma gratuita y modificarlo para incluir mejoras. Proporciona una amplia gama de herramientas estadísticas que incluyen análisis de datos y generación de gráficos, además de ofrecer la posibilidad de cargar bibliotecas y paquetes con diversas funcionalidades, lo que permite a los usuarios extender su configuración básica.

En la siguiente imagen vamos a poder observar el panel de “R” que se divide en cuatro cuadrantes. El primero es donde se puede ver el historial del código, después viene la parte de espacio de trabajo, el tercero donde se introduce el código y el cuarto donde podemos visualizar gráficos.

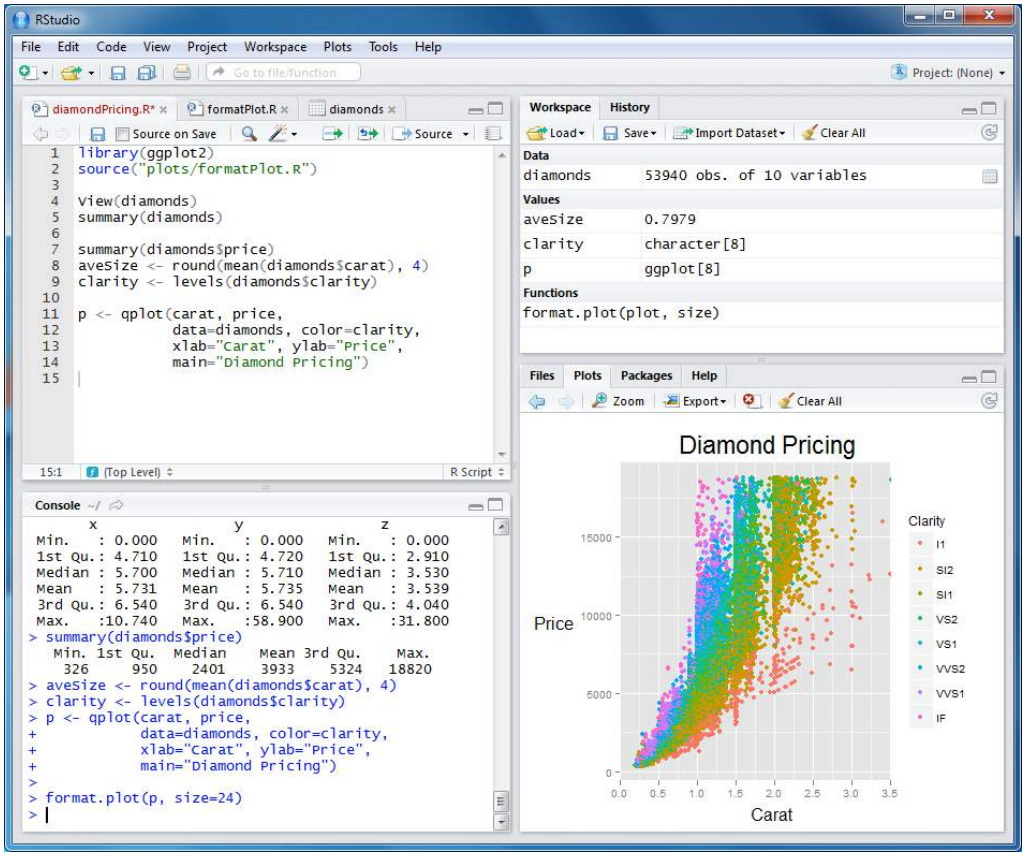


*Figura 10: Visualización de pantalla de RStudio*

Adicionalmente, R puede integrarse con distintas bases de datos. Existen bibliotecas que facilitan su utilización desde lenguajes de programación interpretados como Perl, Python y Ruby. Y por supuesto existen proyectos que permiten utilizar R desde Java o .Net.

Junto con R se incluyen ocho bibliotecas o paquetes (llamadas bibliotecas estándar), sin embargo, existen muchos paquetes más disponibles a través de Internet en (http://www.r-project.org). Actualmente se encuentran disponibles 2337 librerías desarrolladas en R, que cubren multitud de campos desde aplicaciones Bayesianas, financieras, gráfico de mapas, análisis de datos espaciales, etc. Esto es lo que define a R como un entorno vivo, que se actualiza con frecuencia y que está abierto a la mejora continua.

Funcionalidades:

- Software libre y gratuito.
- Multiplataforma.
- Código abierto.
- Actualizaciones constantes.
- Buena interfaz visual sin necesidad de programar.

Licenciamiento: GNU GPL.

Principales paquetes:

- e1071: Funciones para el análisis de clase latente, transformada de Fourier a corto plazo, clustering difuso, máquinas de soporte vectorial (SVM), computación de ruta más corta, clústeres empaquetados, etc.
- Rpart: Árboles recursivos de Partición y Regresión.
- Igraph: Una colección de herramientas de análisis de red.
- Nnet: Para redes neuronales y modelos lineales multinomiales (con más de dos categorías).
- Randomforest: Bosques aleatorios para clasificación y regresión.
- Caret: Conjunto de funciones que intentan simplificar el proceso para crear modelos predictivos.
- Kernlab: machine learning basado en núcleos.
- Glmnet: Modelos lineales generalizados de malla elástica.
- ROCR: Visualización del rendimiento de los clasificadores de puntuación.
- Gbm: Modelos de regresión potenciados generalizados.

Para más información consultar los siguientes enlaces:

- Página oficial: https://www.r-project.org/
- Documentación: https://www.rdocumentation.org/
- Comunidad: https://community.rstudio.com/

### 3.2.2. Programación Python

Python es un lenguaje de programación interpretado, orientado a objetos y de alto nivel con semántica dinámica. Sus estructuras de datos integradas de alto nivel, combinadas con el tipado dinámico y el enlace dinámico, lo hacen muy atractivo para el desarrollo rápido de aplicaciones, así como también para usarlo como scripting o lenguaje de pegado para conectar componentes existentes. La sintaxis simple y fácil de aprender de Python enfatiza la legibilidad y, por lo tanto, reduce el costo del mantenimiento del programa. Python admite módulos y paquetes, lo que fomenta la modularidad del programa y la reutilización de código. El intérprete de Python y la extensa biblioteca estándar están disponibles en formato fuente o binario sin cargo para todas las plataformas principales, y se pueden distribuir libremente.

Es relativamente fácil de aprender y portátil, lo que significa que sus declaraciones se pueden interpretar en una serie de sistemas operativos, incluidos los sistemas basados en UNIX, Mac OS, MS-DOS, OS / 2 y varias versiones de Microsoft Windows 98. El código fuente está disponible y abierto para modificación y reutilización.

Una característica notable de Python es su sangría de las declaraciones fuente para hacer que el código sea más fácil de leer. Python ofrece tipos de datos dinámicos, clases preparadas e interfaces para muchas llamadas y bibliotecas del sistema.

Funcionalidades:

- Fácil de aprender cuando se tienen bases de programación y se viene de un entorno Web.
- Lenguaje Generalista.
- Gratuito.
- Software abierto.
- Rápido de desarrollar.
- Sencillo y veloz.
- Multiplataforma.
- Fácil gestión de errores mediante las excepciones.

Licenciamiento: Licencia Python Software Fundation (licencia de software libre al estilo de la licencia BSD y compatible con la licencia GNU GPL).

Para más información consultar los siguientes enlaces:

- Página oficial: https://www.python.org/
- Documentación: https://www.python.org/doc/
- Comunidad: https://www.python.org/community/
- Tutoriales: https://docs.python.org/2/tutorial/

Principales librerías:

- Keras: Diseñada específicamente para hacer experimentos con redes neuronales. Permite crear prototipos rápidamente y de manera fácil, pues está pensada para que sea fácil de usar
- Pytorch: Es un framework de Python que permite el crecimiento rápido del Deep Learning, con una fuerte aceleración de la GPU. La característica principal de Pytorch es que utiliza gráficos computacionales dinámicos.
- TensorFlow: Cuando una red neuronal se vuelve más compleja, llena de capas y parámetros, el proceso se vuelve mucho más complejo. Es aquí donde entra TensorFlow, una librería de computación numérica que computa gradientes automáticamente.
- Scikit-learn: Otra librería para diseñar y entrenar redes neuronales. Emplea algoritmos de clasificación (determina a qué categoría pertenece un objeto), regresión (asocia atributos de valor continuo a objetos) y agrupamiento (agrupa objetos similares en conjuntos); opera de manera simultánea con librerías como NumPy y SciPy.

- NumPy es una extensión de Python que agrega mayor soporte para vectores y matrices a las ya existentes, constituyendo así una biblioteca de funciones matemáticas de alto nivel.
- SciPy: Es una biblioteca de herramientas y algoritmos matemáticos para Python que contiene módulos para optimización, álgebra lineal, integración, interpolación, funciones especiales, procesamiento de señales y de imagen entre otras tareas para la ciencia e ingeniería.
- Pandas: Es una librería para el análisis de datos que cuenta con las estructuras requeridas para limpiar los datos en bruto y que sean aptos para el análisis. Dado que Pandas lleva a cabo tareas como alinear datos para su comparación, fusionar conjuntos de datos, gestión de datos perdidos, etc., se ha convertido en una librería muy importante para procesos estadísticos en Python.
- Matplotlib: Es una biblioteca para la generación de gráficos a partir de datos contenidos en listas o arreglos.
- Theano: Librería para definir, evaluar y optimizar funciones matemáticas. Tiene un funcionamiento fluido y se puede integrar con NumPy.

### 3.2.3. KNIME

Es una plataforma de código abierto para analíticas, reporte e integración de información. Considera varios componentes para machine learning y minería de datos a través de su concepto de flujo de datos modular, y permite una interfaz de usuario gráfica que permite el ensamble de nodos para procesamiento de información, modelación y análisis y visualización de datos (PAT Research, 2018).

Provee más de 1000 rutinas de análisis de datos, tanto nativas como a través de R o Weka, algunas de las cuales se muestran en la siguiente imagen.

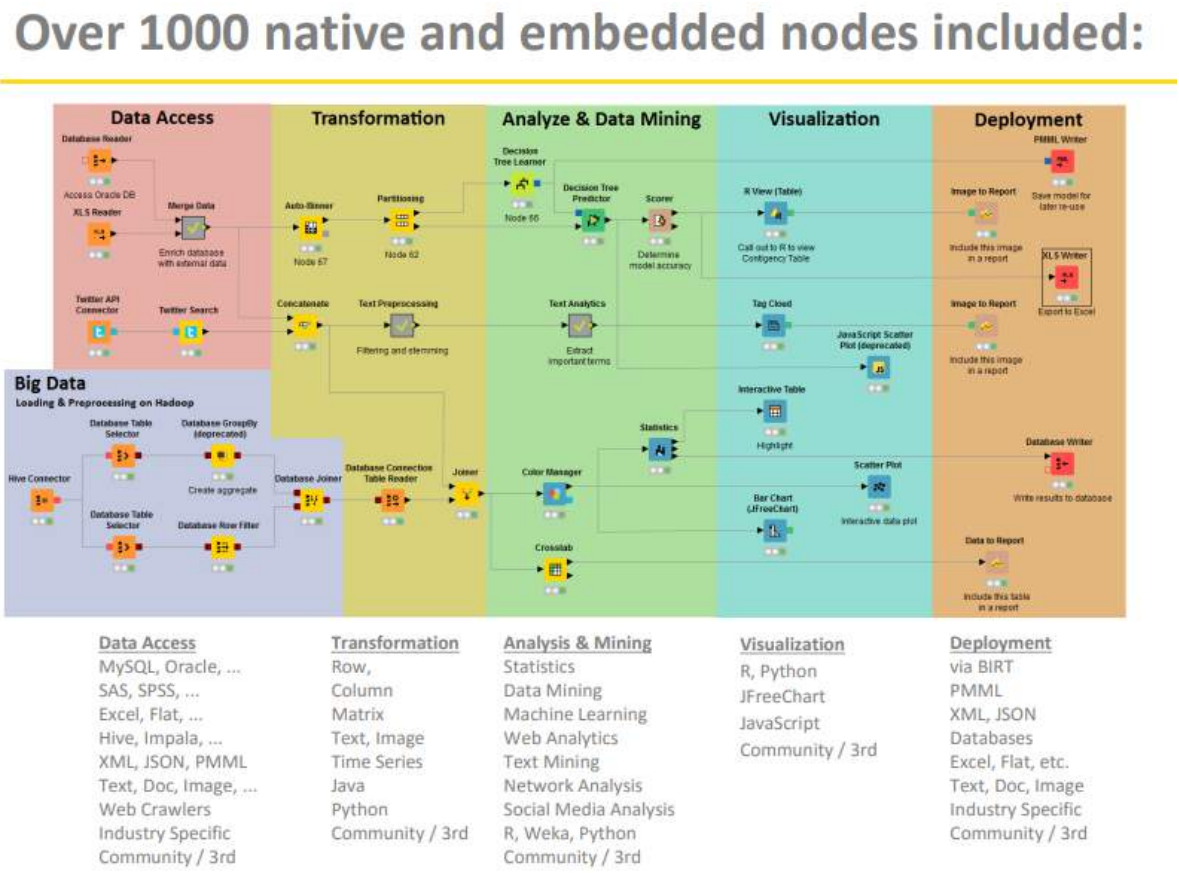


*Figura 11: Tipos de nodos incluidos en KNIME*

Los flujos analíticos de KNIME pueden ser ejecutados a través de una interfaz de usuario interactiva y en modo de ejecución por bloques, habilitando el proceso de análisis de datos para ser fácilmente integrado en la gestión local de trabajos y ejecutado en una base periódica.

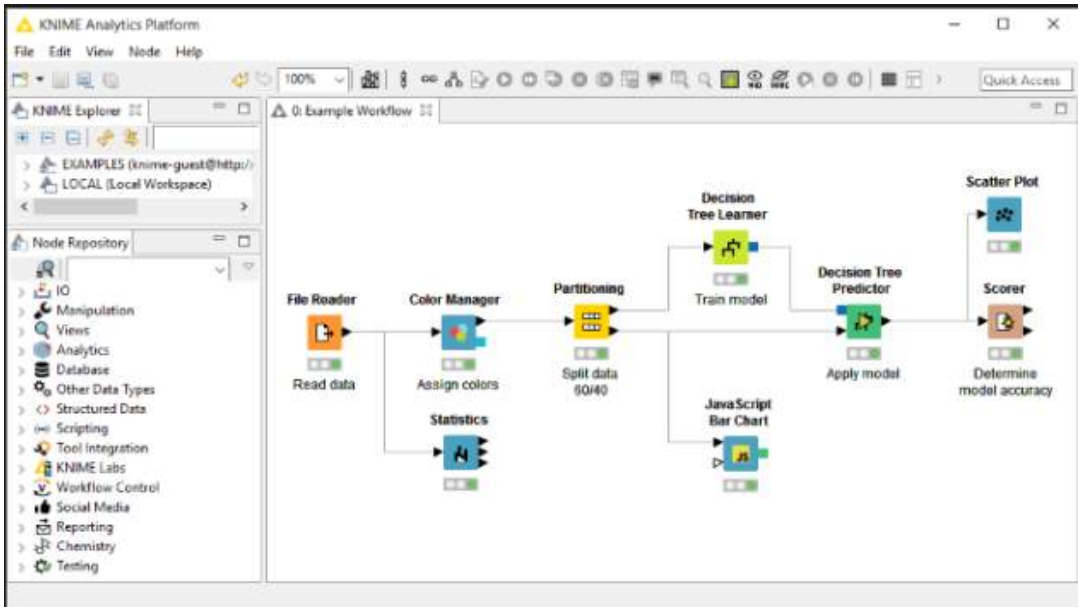


*Figura 12: Visualización de pantalla de KNIME*

Está escrito en Java y basado en Eclipse y hace uso de sus mecanismos de extensión para añadir complementos proveyendo funcionalidades adicionales.

La versión central incluye cientos de módulos para la integración, y los nodos de bases de datos soportan todos los sistemas gestores de bases de datos comunes. Usando la extensión gratuita Report Designer los flujos de trabajo pueden ser usados como conjuntos de datos para crear plantillas que pueden ser exportadas a documentos en formato .doc, .ppt, .xls, .pdf, entre otros. También existen complementos que permiten la integración de métodos de minería de texto y de imágenes, así como análisis de series de tiempo (PAT Research, 2018).

Funcionalidades:

- Software libre bajo licencia GNU.
- Combinación de datos y herramientas.
- Analíticas poderosas.
- Más de 1000 módulos y creciendo.
- Conectores para todos los formatos de archivos y bases de datos más utilizadas.
- Soporte para una gran variedad de tipos de datos.
- Combinación y transformación de datos nativa y en la base de datos.
- Funciones matemáticas y estadísticas.
- Algoritmos de predicción avanzados y de machine learning.
- Control de flujo.
- Herramienta de unión para R, Python, SQL, Java, Weka, etc.
- Vistas de datos y reportes interactivos.

Licenciamiento: GNU GPL 3

Para más información consultar los siguientes enlaces:

- Página oficial: https://www.knime.com/
- Documentación: https://www.knime.com/documentation
- Comunidad: https://www.knime.com/knime-community
- Tutoriales: https://www.knime.com/resources

### 3.2.4. Rapid Miner

Es una solución que facilita el análisis predictivo permitiendo procesos de analítica avanzada mediante un esquema de programación drag & drop, y opcionalmente la posterior generación de código.

Este modelo de solución acelera la creación de prototipos y la validación de modelos, al mismo tiempo que agiliza la transformación, el desarrollo y la validación de datos.

RapidMiner está escrita en Java y contiene más de 500 operadores con diferentes enfoques para mostrar las conexiones en los datos: hay opciones para minería de datos, minería de texto o minería web, y también para el análisis de sentimientos o minería de opinión.

Proporciona tres herramientas: RapidMiner Studio, RapidMiner Server y RapidMiner Radoop, cada una enfocada a una técnica diferente de la minería de datos.

**RapidMiner Studio:** Es un software para los equipos de ciencia de datos que utiliza machine learning para el desarrollo de modelos predictivos. Provee una amplia librería de algoritmos, preparación de información y funciones de exploración, así como herramientas de validación de modelos. (PAT Research, 2018)

El software permite utilizar código R y Python, así como añadir nuevas funcionalidades a través de un Marketplace de extensiones preconstruidas para temas como estadística univariable y multivariable, minería de datos, series de tiempo, procesamiento de imágenes, analíticas web, texto, etc. Usa un modelo cliente/servidor, que ofrece el servidor como un software, un servicio o en la infraestructura de la nube.

RapidMiner provee una interfaz gráfica de usuario para diseñar y ejecutar flujos analíticos, a los que se refiere como procesos y constan de múltiples operadores. Cada operador ejecuta una tarea, donde su salida es la entrada del siguiente operador. Las funciones individuales pueden ser llamadas desde la línea de comandos.

**RapidMiner Server:** Hace fácil compartir, reutilizar y operacionalizar los modelos predictivos y resultados creados en RapidMiner Studio. El repositorio central y gestión dedicados al poder de cálculo y a las flexibles opciones de despliegue soporta el trabajo en equipo y la rápida puesta en acción; hace el trabajo en equipo fácil y rápido ya que los usuarios de la organización pueden trabajar juntos de manera eficiente, debido a que son capaces de acceder, reutilizar y compartir modelos y procesos en una versión controlada, segura y en un ambiente gestionado centralmente (PAT Research, 2018).

**RapidMiner Radoop:** Simplifica el análisis predictivo con un volumen de datos elevado sobre los clusteres Hadoop, ya que aprovecha la potencia de procesamiento de Hadoop.

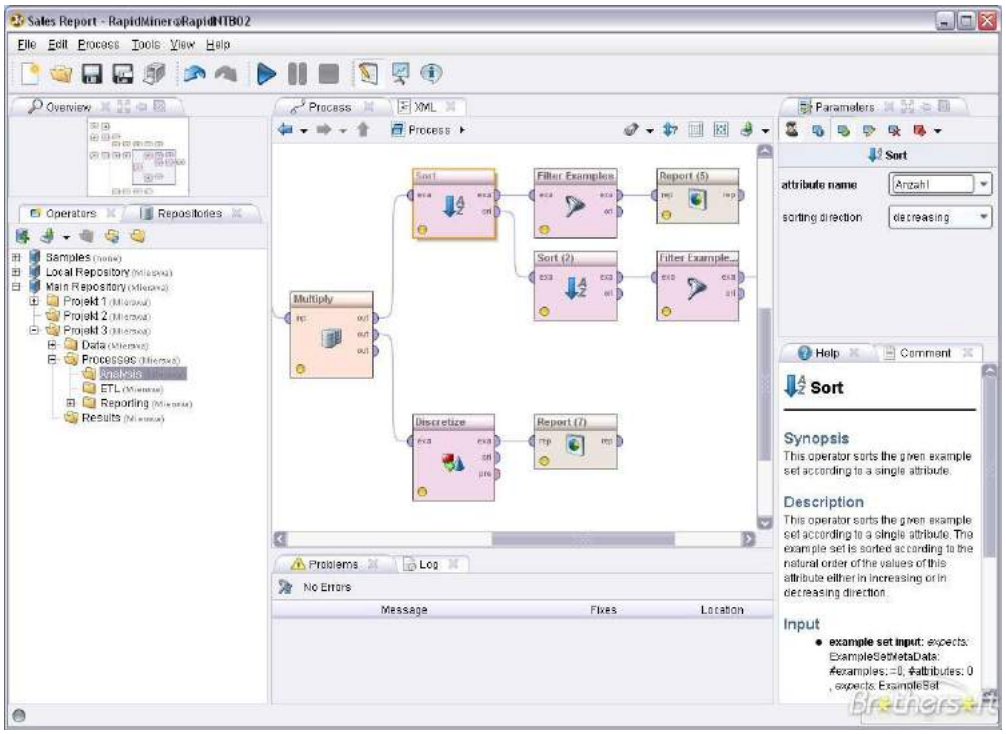


*Figura 13: Interfaz gráfica que provee RapidMiner*

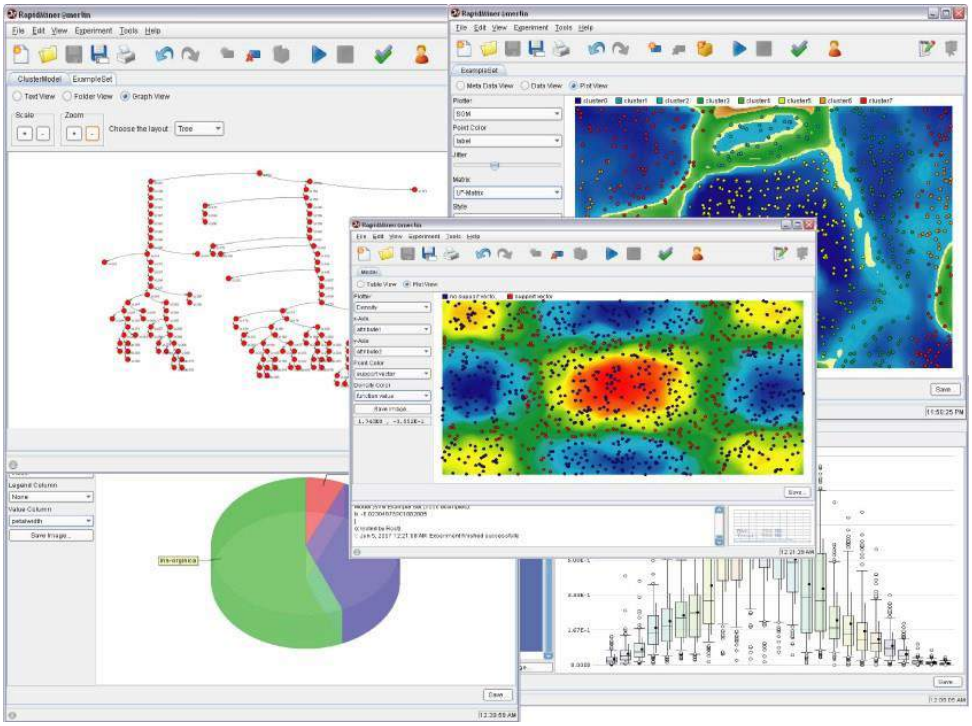

*Figura 14: Visualización de pantalla de RapidMiner*

Funcionalidades:

- Acceso a datos, limpieza, preparación y modelación con RapidMiner Studio.
- Reducción de la complejidad de Hadoop, procesamiento escalable, acceso a grandes volúmenes de datos y optimización para Hadoop con RapidMiner Radoop.
- Plataforma unificada.
- Amplia conectividad.
- Cuenta con una plataforma web que permite publicar reportes, implementar sistemas de scoring, diseño y navegación web de reportes, integración vía servicios web.
- Fácil desarrollo debido a su interfaz gráfica (drag & drop).
- Permite el uso de código en R y Python.

Licenciamiento: AGPL 3.0

Para más información consultar los siguientes enlaces:

- Página oficial: https://rapidminer.com/
- Documentación: https://docs.rapidminer.com/
- Comunidad: https://community.rapidminer.com/
- Tutoriales: https://rapidminer.com/training/videos/

## 3.2.5. H2O.ai

H2O.ai es una API de machine learning de código abierto y de fácil escalabilidad para las aplicaciones más inteligentes. H2O hace posible para todos aplicar analíticas predictivas y machine learning para resolver los problemas de negocio más desafiantes.


*Figura 15: Estructura del motor de predicción H2O*

Dentro de H2O se utiliza un valor para acceder y referenciar datos, modelos, objetos, entre otros, en todos los nodos y máquinas. Los algoritmos se implementan sobre el marco Map / Reduce distribuido de H2O y utilizan el marco Java Fork / Join para multihilo. Los datos se leen en paralelo, se distribuyen por el clúster y se almacenan en la memoria en un formato de columna de forma comprimida. El analizador de datos de H2O tiene una inteligencia incorporada para adivinar el esquema del conjunto de datos entrante y admite la incorporación de datos de múltiples fuentes en varios formatos (H2O.ai, 2018).

Funcionalidades:

- Código abierto.
- Interfaz gráfica de usuario basada en la web fácil de usar.
- Carga, exploración, depuración y selección de información.
- Creación y evaluación de modelos.
- La capacidad de respuesta del procesamiento en memoria y la capacidad de ejecutar una socialización rápida entre nodos y clústeres se combinan para que pueda soportar las solicitudes de grandes conjuntos de datos.
- La API REST de H2O permite el acceso a todas las capacidades de H2O desde un script externo vía JSON a través de HTTP.
- Alta velocidad, calidad, facilidad de uso y despliegue del modelo.
- Involucra lenguajes como R, Python, Java, Scala, JSON y a través de APIs poderosas.
- Incluye conexiones a fuentes de datos desde HDFS, S3, SQL y No SQL.
- Análisis de Big Data escalable de manera masiva.
- Scoring de datos en tiempo real.

Licenciamiento: Licencia de código abierto Apache 2.0

Para más información consultar los siguientes enlaces:

- Página oficial: https://www.h2o.ai/
- Documentación y tutoriales: http://docs.h2o.ai/
- Comunidad: https://www.h2o.ai/community/

### 3.2.6. Microsoft R

La familia de productos de Microsoft R incluye Microsoft R Server, Microsoft R Client, Microsoft R Open y SQL Server R Services. Microsoft R es la plataforma de análisis de nivel empresarial de despliegue más amplio para R. Soporta una gran variedad de estadísticas de Big Data, modelación predictiva y capacidades de machine learning. Así mismo, soporta el rango completo de analíticas de exploración, análisis, visualización y modelación basadas en R.

Microsoft R Client es una herramienta de ciencia de datos para análisis de alto rendimiento. Está construido sobre Microsoft R Open, por lo que se puede usar cualquier paquete de R para construir analíticas propias. Adicionalmente, Microsoft R Client integra el poder de la tecnología ScaleR y sus funciones propietarias para beneficiar con la paralelización e informática remota (PAT Research, 2018).

Microsoft R Open es una distribución mejorada de R de Microsoft Corporation; es una plataforma completa de código abierto para análisis estadísticos y ciencia de datos. Al ser

desarrollada en la máquina de código abierto R, Microsoft R Open es completamente compatible con paquetes, scripts y aplicaciones que trabajan con esa versión de R. Permite una mejora en el desempeño, en comparación con la distribución de R estándar, ya que se apalanca del alto rendimiento y librerías matemáticas multihilo (PAT Research, 2018).

Por su parte SQL Server R Services provee una plataforma de desarrollo de aplicaciones inteligentes. Se puede usar el poder de R y de sus paquetes para crear modelos y generar predicciones usando datos de SQL Server.

Funcionalidades:

- Realiza el análisis en los datos.
- Construcción de aplicaciones con capacidad para inteligencia artificial.
- Experiencia mejorada y velocidad de despliegue.
- Adaptable a las necesidades futuras.
- Escala las analíticas en R para Big Data.
- Soporte.

Licenciamiento: GNU GPL 3.0

Para más información consultar los siguientes enlaces:

- Página oficial: https://mran.microsoft.com/
- Documentación y tutoriales: https://mran.microsoft.com/documents/getting-started
- Comunidad:
  - Foro:https://social.msdn.microsoft.com/Forums/en-S/home?forum=ropen
  - Blog: http://blog.revolutionanalytics.com/

### 3.2.7. Anaconda

Anaconda es una plataforma abierta de ciencia de datos alimentada por Python. La versión de código abierto es una distribución de Python y R de alto rendimiento, e incluye más de 100 de los paquetes de ciencia de datos más populares de estos lenguajes. También cuenta con acceso a más de 720 paquetes que pueden ser fácilmente instalados con Anaconda (el paquete, dependencia y gestor de ambiente que es incluido en Anaconda) (PAT Research, 2018).

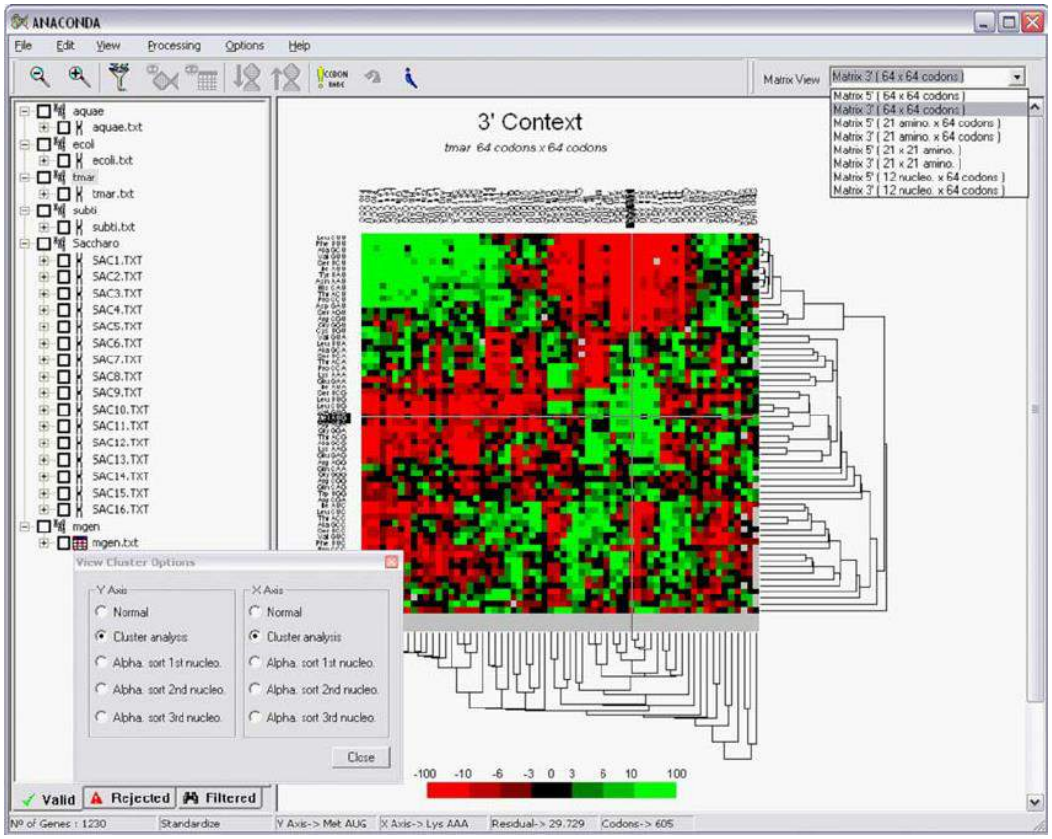


*Figura 16: Visualización de pantalla de Anaconda*

El repositorio de Anaconda otorga a los equipos de ciencia de datos una manera fácil de reproducir, desarrollar y publicar activos de ciencia de datos. Los paquetes, blocs de notas y ambientes pueden ser compartidos para mejorar la productividad del equipo y asegurar la consistencia.

Funcionalidades:

- Análisis de flujos de trabajo.
- Análisis de interacción.
- Distribución de alto rendimiento.
- Analíticas avanzadas.
- Escalamiento de alto rendimiento.
- Reproducibilidad.
- Despliegue analítico.

Licenciamiento: BSD.

Para más información consultar los siguientes enlaces:

- Página oficial: https://anaconda.org/

- Documentación: https://docs.anaconda.com/
- Comunidad: https://anaconda.org/community

## 3.2.8. GNU Octave

GNU Octave es un lenguaje de alto nivel, destinado principalmente a cálculos numéricos. Proporciona una interfaz de línea de comando conveniente para resolver problemas lineales y no lineales numéricamente, y para realizar otros experimentos numéricos utilizando un lenguaje que es principalmente compatible con Matlab.

Es fácilmente extensible y personalizable a través de funciones definidas por el usuario escritas en el propio lenguaje de Octave, o utilizando módulos cargados dinámicamente escritos en C ++, C, Fortran u otros lenguajes.

Es un software libre que corre sobre GNU/Linux, MacOS, BSD y Windows.

El objetivo principal de este software es ayudar a los usuarios a resolver problemas realísticos o usarlos como aplicación de enseñanza, investigación y comercial.

A continuación, se muestra la visualización de pantalla de este software:

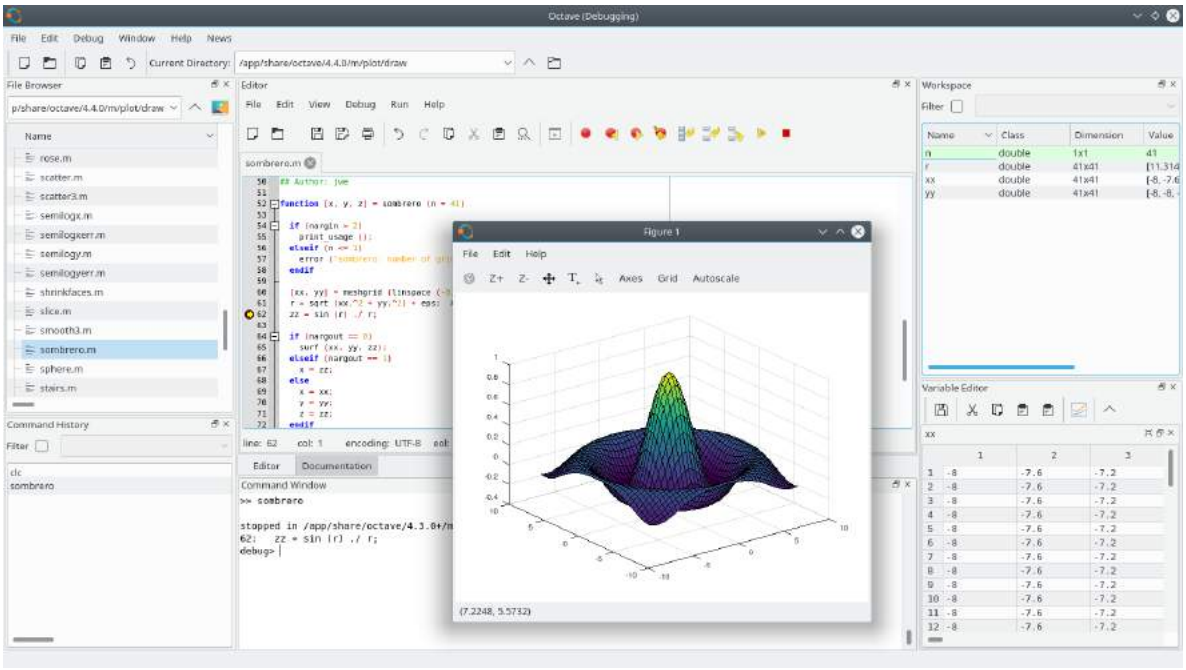


*Figura 17: Visualización de pantalla de GNU Octave*

Funcionalidades:

- Puede correr de distintas formas: en la interfaz de usuario, en la consola, o invocada como parte de un script Shell.
- Sintaxis altamente compatible con Matlab.
- Destinado a cálculos numéricos.

Licenciamiento: GNU GPL.

Para más información consultar los siguientes enlaces:

- Página principal: https://www.gnu.org/software/octave/
- Documentación: https://octave.org/doc/interpreter/

### 3.2.9. Weka Data Mining

Es una colección de algoritmos de machine learning para tareas de minería de datos. Contiene herramientas para la preparación de datos, clasificación, regresión, agrupación, minería de reglas de asociación y visualización.

Es importante mencionar que Weka se basa en la suposición de que los datos están disponibles como un solo archivo plano o de relación, donde cada punto de datos se describe por un número fijo de atributos. Los datos pueden ser importados de formatos como ARFF, CSV, C4.5, binarios, así como de un URL o de una base de datos SQL usando Java Database Connectivity (JDBC).

La interfaz principal de usuario es el explorador, y se puede acceder a las mismas funcionalidades a través de la interfaz de flujo de conocimiento basada en componentes o a través de la línea de comandos. También cuenta con Experimenter, el cual permite la comparación sistemática de la ejecución de los algoritmos predictivos de machine learning de Weka.

La interfaz Explorer presenta varios paneles que proveen acceso a los principales componentes (como muestra la siguiente imagen), como el panel de preprocesamiento que facilita la importación de datos, el panel de clasificación que habilita al usuario para aplicar algoritmos de clasificación y regresión, el panel de clúster que da acceso a técnicas de agrupación, el de selección de atributos que permite seleccionar las variables que mejor predicen en un conjunto de datos, y el panel de visualización que muestra una matriz de diagrama de dispersión.

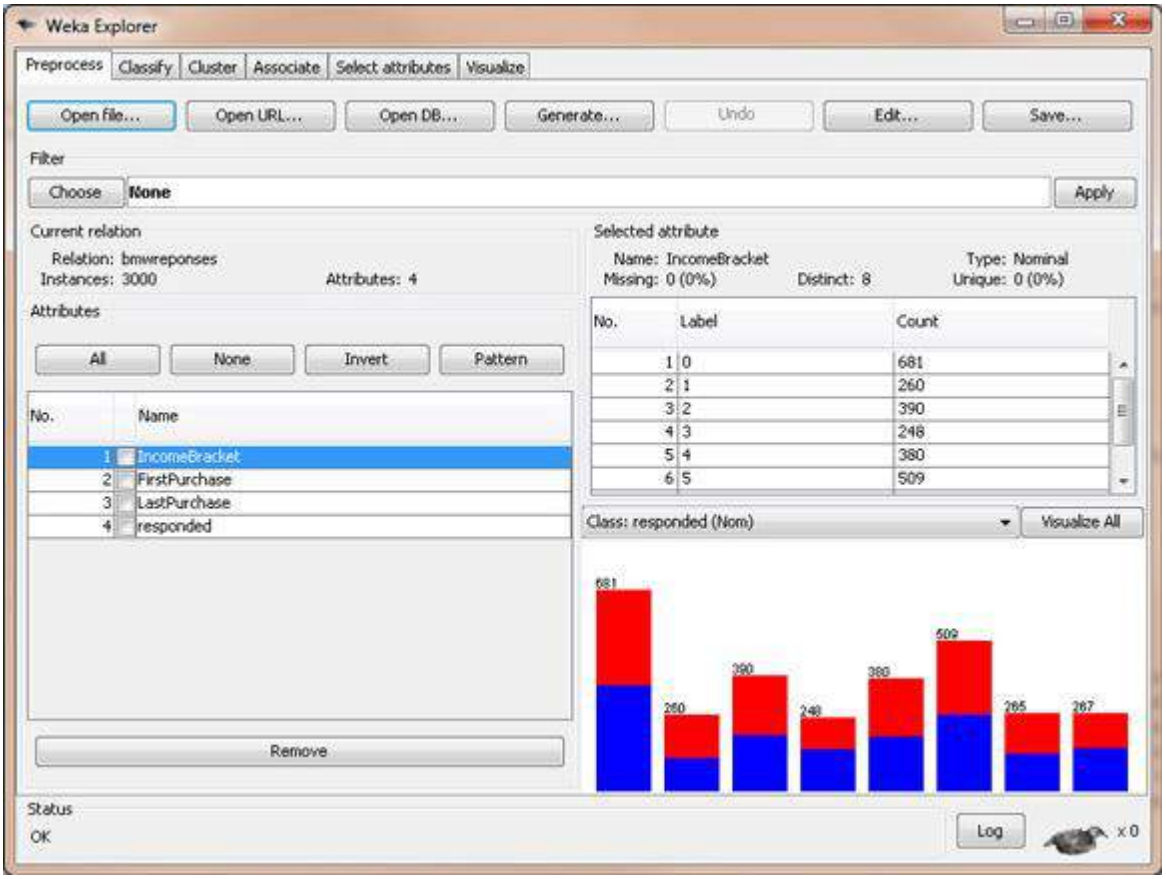


*Figura 18: Visualización de pantalla de Weka Data Mining*

Funcionalidades:

- Acceso a bases de datos SQL usando JDBC.
- Los algoritmos pueden ser aplicados directamente o llamados a partir de código Java.
- Preparación de datos.
- Técnicas de machine learning como: regresión logística, perceptrones multicapa, máquinas de vectores de soporte, árboles de decisión, etc.
- Visualización.

Licenciamiento: GNU GPL.

Para más información consultar los siguientes enlaces:

- Página principal: https://www.cs.waikato.ac.nz/ml/weka/index.html
- Documentación: https://www.cs.waikato.ac.nz/ml/weka/documentation.html
- Tutoriales: https://www.cs.waikato.ac.nz/ml/weka/courses.html
- Videos: https://www.youtube.com/user/WekaMOOC

## 3.2.10. Orange

Es una herramienta de Código abierto para análisis y visualización de datos donde la minería de datos se hace a través de programación visual o por código de Python. Cuenta con componentes para machine learning, adiciones para bioinformática y minería de texto.

Consiste en una interfaz en la que el usuario coloca widgets y crea flujos de trabajo para el análisis de datos. Estos widgets otorgan funcionalidades como: leer datos, mostrar una tabla, selección de variables, comparación de algoritmos, visualización de elementos, etc.

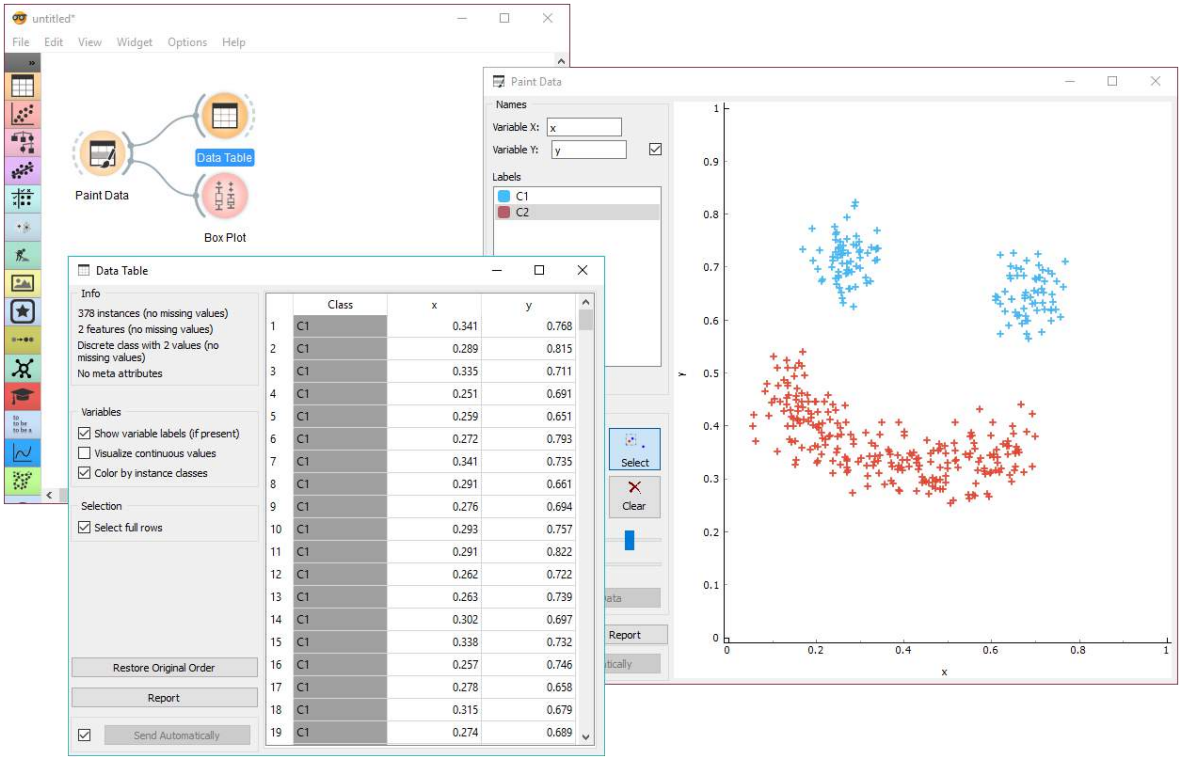


*Figura 19: Visualización de pantalla de Orange*

Se pueden desarrollar widgets, los cuales pueden ser extendidos como complementos, permitiendo que los componentes y códigos puedan ser reutilizados.

Orange corre sobre Windows, Mac OS X y una variedad de sistemas operativos Linux.

Funcionalidades:

- Visualización interactiva de datos.
- Programación visual (interfaz gráfica de usuario).
- Adiciones disponibles para minar datos de fuentes de datos externas, minería de texto, análisis de redes.
- Recuerda las selecciones y sugiere las combinaciones más utilizadas.

- Existen más de 100 widgets que cubren las tareas estandarizadas del análisis de datos.

Licenciamiento: GNU GPL.

Para más información consultar los siguientes enlaces:

- Página oficial: https://orange.biolab.si/
- Documentación: https://orange.biolab.si/docs/
- Comunidad: https://blog.biolab.si/
- Tutoriales: https://orange.biolab.si/training/introduction-to-data-mining/

## 3.2.11. TANAGRA

Tanagra es un software gratuito de minería de datos para fines académicos y de investigación. Propone varios métodos de minería de datos, de análisis exploratorio, aprendizaje estadístico, aprendizaje automático y área de bases de datos.

En Tanagra el usuario puede diseñar visualmente un proceso de minería de datos en un diagrama, donde cada nodo es una técnica estadística o de aprendizaje automático y la conexión entre dos nodos representa la transferencia de datos, de tal manera que los tratamientos están representados en un diagrama de árbol. Los resultados se muestran en formato HTML, por lo que es fácil exportar los resultados para ser visualizados en un navegador. También es posible copiar las tablas de resultados en una hoja de cálculo.

Tanagra hace un buen uso de los enfoques estadísticos (por ejemplo, pruebas estadísticas paramétricas y no paramétricas), los métodos de análisis multivariable (por ejemplo, análisis factorial, análisis de correspondencia, análisis de conglomerados, regresión) y las técnicas de aprendizaje automático (por ejemplo, red neuronal, máquina de soporte vectorial, árboles de decisión).

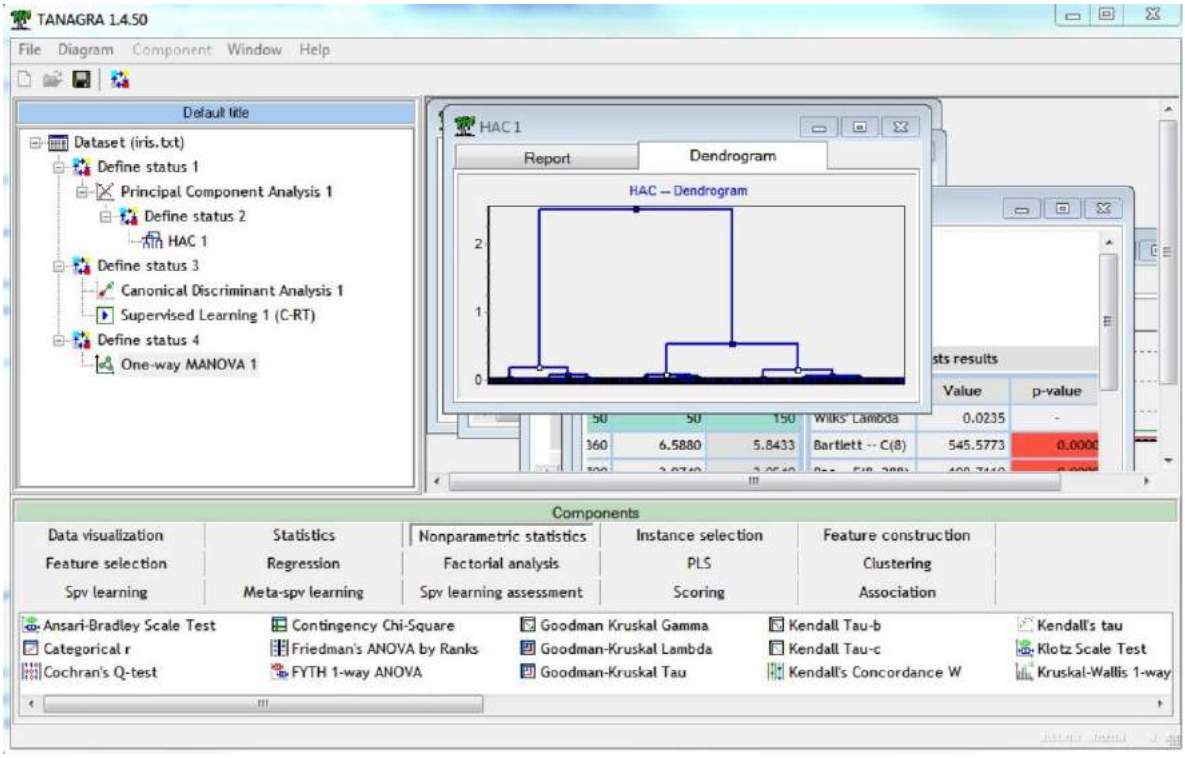


*Figura 20: Visualización de pantalla de Tanagra*

Funcionalidades:

- Análisis exploratorio.
- Minería de datos.
- Aprendizaje automático.
- Interfaz de usuario gráfica.

Licenciamiento: Gratuito.

Para más información consultar los siguientes enlaces:

- Página Principal: http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
- Documentación: http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
- Tutoriales: http://data-mining-tutorials.blogspot.com/
- Videos: https://www.youtube.com/watch?v=5MtsExnb_J0

## 3.2.12. C++

C++ es un lenguaje de programación desarrollado a partir de C, que proporciona funcionalidades tales como administración de memoria con punteros y funcionalidad OPP de alto nivel. Para utilizarlo es suficiente contar con un editor de texto y un compilador (traductor de código fuente a lenguaje máquina), aunque lo más recomendable para usar este lenguaje es usar un entorno de desarrollo integrado (IDE por sus siglas en inglés), que es un software que agrupa herramientas que permiten el desarrollo de un programa informático, incluyendo un compilador.

Dentro de los compiladores más conocidos de este lenguaje se encuentran GCC (GNU Compiler Collection), MinGW (implementación para Windows de GCC), ambos distribuidos bajo licencia GNU GPL. Los distribuidos como Software propietario son: C++Builder y Visual C++ (Forma parte de Visual Studio)

En cuanto a entornos de desarrollo destacan:

**Dev-C++:** Emplea el compilador MinGW y esta creado mediante el Entorno Delphi y el lenguaje Object Pascal. Se trata de un software libre, sencillo, ligero y eficiente, para la plataforma Windows.

Licenciamiento: GNU GPLv2.

- Página oficial:
    - http://www.bloodshed.net/dev/index.html
    - http://orwelldevcpp.blogspot.com/
- Documentación: http://www.bloodshed.net/dev/doc/index.html

**CodeBlocks:** Este es un software libre, multiplataforma. Code Blocks es una alternativa a Dev-C++ que ha sido desarrollada mediante el propio lenguaje C++. Sus capacidades son bastante buenas y es muy popular entre los nuevos programadores. Se puede encontrar separado del compilado o la versión “mingw” que incluye g++ (GCC para C++).

Licenciamiento: GNU GPL.

- Página oficial: http://www.codeblocks.org/
- Comunidad: http://forums.codeblocks.org/

**Eclipse:** Su principal propósito es programar mediante Java, también es libre y multiplataforma.

Licenciamiento: Licencia Pública Eclipse.

- Página oficial: https://www.eclipse.org/
- Documentación: https://help.eclipse.org/photon/index.jsp
- Comunidad:
    - https://blogs.eclipse.org/
    - https://www.eclipse.org/forums/
- Videos: https://www.youtube.com/user/EclipseFdn

**MonoDevelop:** Esta es una alternativa a Visual Studio, pero este IDE es multiplataforma y de software libre. Posee un editor de interfaces gráficas que implementa la biblioteca GTK y es compatible con el Framework .Net de Microsoft.

Licenciamiento: GNU GPL.

- Página oficial: https://www.monodevelop.com/
- Documentación: https://www.monodevelop.com/documentation/
- Comunidad: https://www.monodevelop.com/developers/

**Anjuta:** Este software tiene como propósito principal utilizar herramientas proporcionadas por GTK+ para desarrollar aplicaciones para el escritorio GNOME. Esta opción es propia de los sistemas GNU/Linux y BSD, y se distribuye bajo licencia GNU GPL.

Licenciamiento: GNU GPL.

- Página oficial: http://anjuta.org/

El lenguaje de programación C++ cuenta con librerías que permiten la aplicación de analítica avanzada y machine learning, entre las que destacan:

- Algebra lineal: Eigen y blaze.
- machine learning: Tensorflow, dlib, mlpack.
- Redes neuronales: dnn, tiny-dnn, fast-dnn.
- Optimizaciones no lineales: Ceres.

Para más información referente al código C++ consultar los siguientes enlaces:

- Página oficial:
    - http://www.cplusplus.com/
    - https://isocpp.org/
- Documentación:
    - http://www.cplusplus.com/reference/
    - http://www.cplusplus.com/articles/

- Comunidad: http://www.cplusplus.com/forum/
- Tutoriales:
    - http://www.cplusplus.com/doc/tutorial/

Videos: https://isocpp.org/blog/category/video-on-demand

## 3.3. Herramientas Propietarias

### 3.3.1. SAS

Es el software líder en analíticas de negocio y servicios, y el vendedor independiente más grande en el mercado de Business Intelligence (Inteligencia de Negocios). A través de soluciones innovadoras, SAS ayuda a sus clientes en más de 70.000 sitios, mejorando la ejecución y el valor de entrega, haciendo que la toma de decisiones sea más rápida y eficiente.

Las analíticas avanzadas de SAS están infundidas de algoritmos innovadores y de vanguardia que pueden ayudar a resolver los problemas que se tengan.

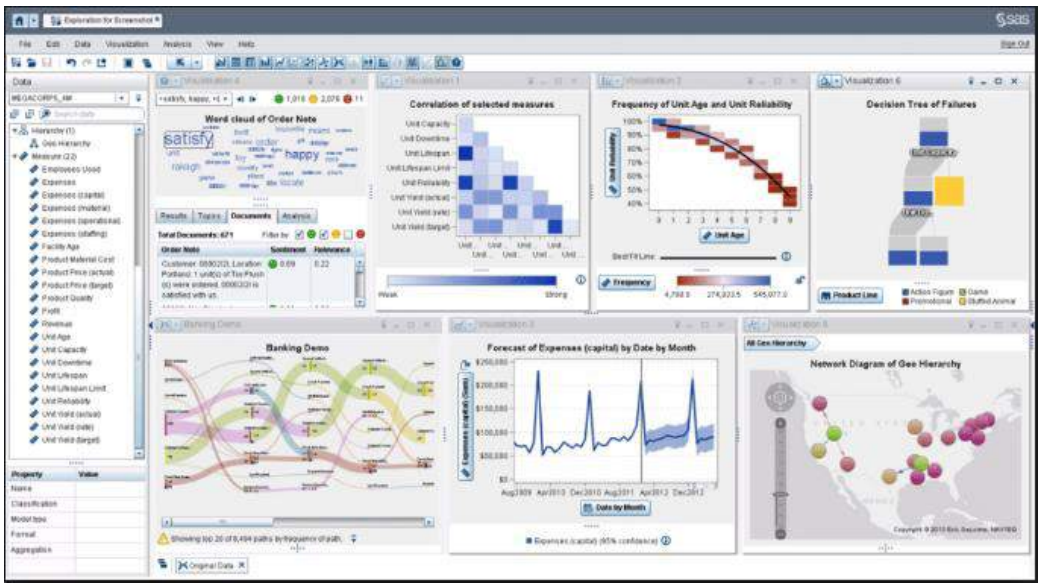


*Figura 21: Visualización de pantalla de SAS*

En su portafolio se incluye herramientas para distintas actividades tales como: SAS Visual Analytics, SAS Enterprise Miner, SAS Model Manager, SAS High Performance Data Mining, SAS Scoring Acelerator, SAS Decision Management.

Funcionalidades:

- Interfaz visual amigable sin necesidad de programar.
- No presenta problemas de compatibilidad de versiones.
- Minería de datos.
- Análisis estadístico.
- Pronóstico y econometría.
- Analíticas de texto.
- Optimización y simulación.

Licenciamiento: Software propietario, y cuenta con una versión Crippleware en la edición SAS Universitaria, con la cual solo se podrá utilizar para fines académicos de aprendizaje, pero no para fines comerciales.

Para más información consultar los siguientes enlaces:

- Página oficial: https://www.sas.com/en_us/home.html
- Documentación: https://support.sas.com/documentation/
- Comunidad: https://communities.sas.com/
- Tutoriales: https://www.sas.com/en_in/training/home/academic-program/sas-tutorials.html

### 3.3.2. Alteryx Analytics

El Diseño de Alteryx permite mezclar datos, construir poderosas aplicaciones de análisis predictivos y espaciales, escalar flujos de trabajo analíticos críticos para cumplir con los requisitos analíticos y de datos, programar múltiples flujos de trabajo, publicar y compartir aplicaciones analíticas de la organización (PAT Research, 2018).

Es el software líder en la unión de datos y analíticas avanzadas. Provee a los analistas con un flujo de trabajo intuitivo para la unión de datos y analíticas avanzadas, que conducen a profundos conocimientos en horas y no en semanas, típico de los enfoques tradicionales. En la siguiente imagen vamos a poder visualizar la pantalla principal de Alteryx Analytics:

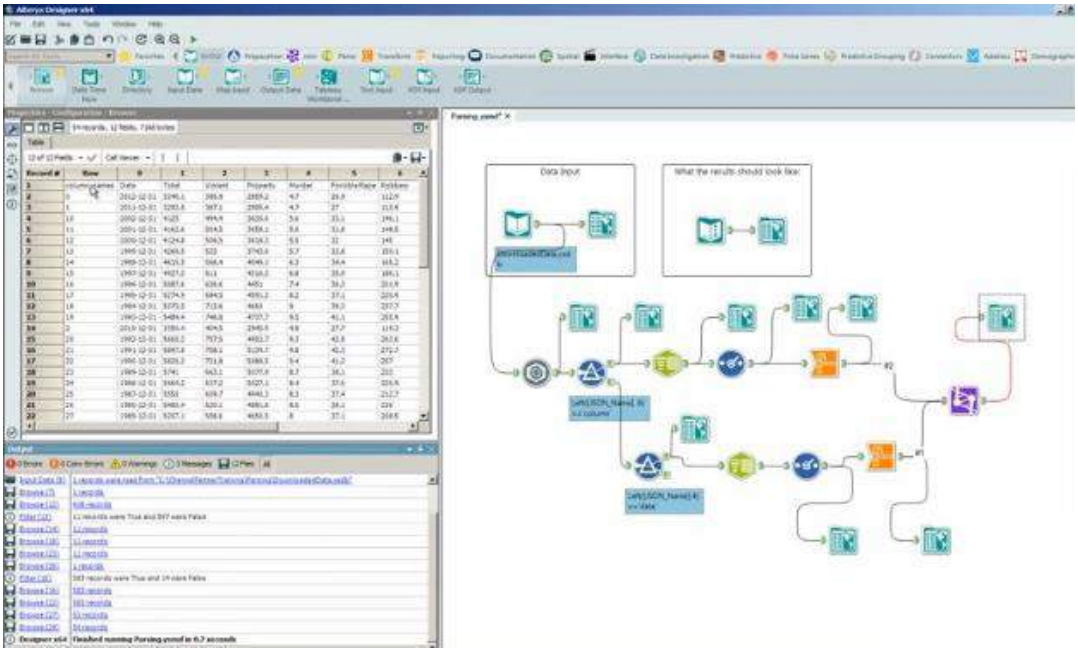

*Figura 22: Visualización de pantalla de Alteryx Analytics*

Su portafolio está compuesto por Alteryx Designer, Alteryx Server y Alteryx Analytics Gallery. El primero posibilita unir datos internos, de terceros, y basados en la nube, construir poderosos modelos analíticos predictivos y espaciales basados en R sin programar y compartir conocimientos profundos de los datos con las personas que toman decisiones. Todas las técnicas de modelación predictivas pueden ser incluidas sin usar programación.

Por su parte, Alteryx Server escala los flujos analíticos críticos para unir los datos y los requerimientos analíticos. Finalmente, Alteryx Analytics Gallery permite consumir, compartir y publicar aplicaciones analíticas y macros.

Funcionalidades:

- Es una solución de tiempo de ejecución basada en la nube.
- Empodera a quien sea que tome decisiones basadas en datos.
- Crea un estudio privado para aplicaciones analíticas.
- Personaliza y ejecuta aplicaciones analíticas desde la nube.
- Basada en el escritorio es una solución de tiempo de diseño.
- Preparación y unión de datos en flujos de datos replicables
- Ejecuta análisis predictivos, espaciales y estadísticos sin código.
- Resultados de los análisis a todos los formatos populares.

Licenciamiento: Software propietario con Shareware de 14 días.

Para más información consultar los siguientes enlaces:

- Página oficial: https://www.alteryx.com/
- Documentación y tutoriales: https://help.alteryx.com/DocPortal/index.htm
- Comunidad: https://community.alteryx.com/

## 3.3.3. Microsoft Azure machine learning

Es un ambiente de desarrollo visual y colaborativo que permite construir, probar y desarrollar soluciones de analítica predictiva que operan en los datos. Ofrece analíticas avanzadas basadas en la nube, diseñadas para simplificar el proceso de machine learning para negocios, provee recursos y memoria flexible, y elimina la preocupación de la preparación e instalación ya que el trabajo es realizado en un navegador web.

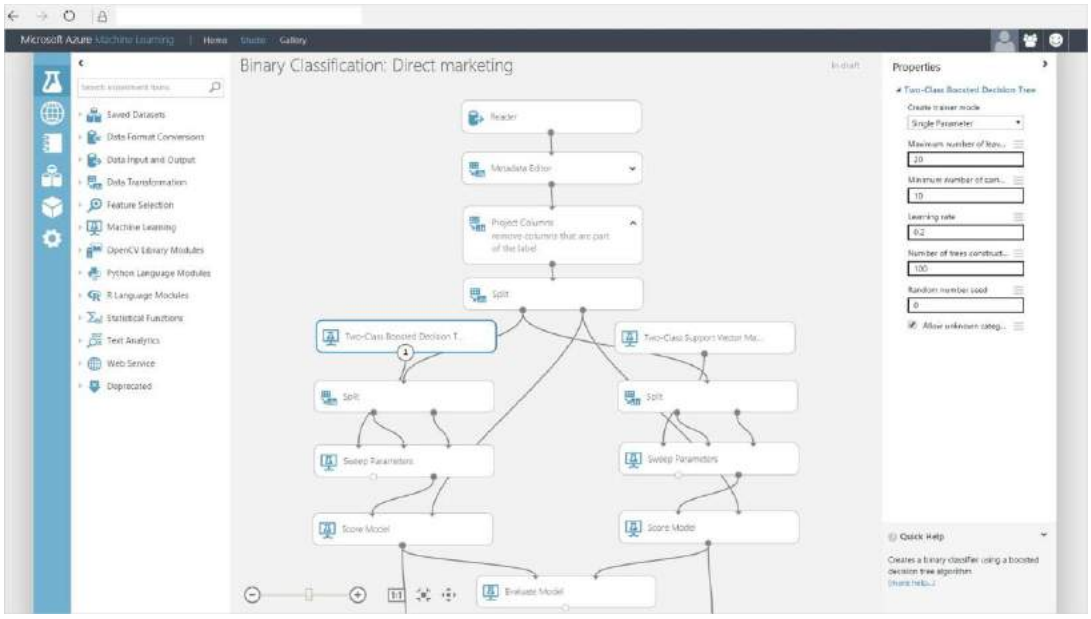


*Figura 23: Visualización de pantalla de Azure machine learning*

Azure soporta cualquier sistema operativo, lenguaje, herramienta y framework; desde Windows a Linux, SQL Server a Oracle, C# a Java. Azure es ambas: una infraestructura como servicio y una plataforma como servicio.

Los usuarios de negocios pueden crear modelos a su manera con lo mejor de los algoritmos de Xbox, Bing, paquetes de R o Python, o profundizando en códigos de R o Python. El modelo final puede ser desarrollado en minutos como un servicio web, el cual puede conectarse a cualquier información, donde sea. También puede publicarse a la comunidad en la galería de productos o en machine learning Marketplace.

Funcionalidades:

- Marketing digital.
- Móviles.
- Comercio electrónico.
- SharePoint en Azure.
- Dynamics en Azure.
- Red Hat en Azure.
- Desarrollo y prueba.
- Monitoreo.
- Inteligencia de Negocio.
- Big Data y analíticas.
- Data Warehouse.
- Aplicaciones de negocio SaaS.
- Migración a la nube.
- Integración híbrida.
- Almacenamiento y respaldo.
- Recuperación de desastres.
- Internet de las cosas.
- Medios digitales.
- Cálculos de alto rendimiento.
- Interfaz bajo el esquema drag & drop.

Licenciamiento: Software propietario en la nube, con Crippleware disponible, que permite el desarrollo e implemente de soluciones de Machine Learning con una cuenta gratuita de Azure durante 12 meses.

Para más información consultar los siguientes enlaces:

- Página oficial: https://azure.microsoft.com/en-us/
- Documentación y tutoriales: https://docs.microsoft.com/en-us/azure/machine-learning/
- Comunidad: https://azure.microsoft.com/en-us/support/community/

## 3.3.4. SAP Predictive Analytics

Es una solución de análisis estadístico, minería de datos y análisis predictivos que trabaja con un ambiente de datos heredados, así como con las fuentes de datos de la plataforma SAP. Permite construir modelos predictivos para descubrir patrones y relaciones en los datos, con el objetivo de realizar predicciones precisas de los eventos futuros.

Ofrece una experiencia intuitiva bajo un esquema de programación drag & drop, de código libre, con suficiente poder para que los científicos de datos conduzcan los análisis más sofisticados usando Big Data, pero lo suficientemente simple para permitir que los analistas de negocios realicen análisis prospectivos usando datos en Excel. Hace de los análisis predictivos un proceso simple, permitiendo incrementar la precisión, incluir todas las variables predictoras potenciales y eliminar errores manuales.

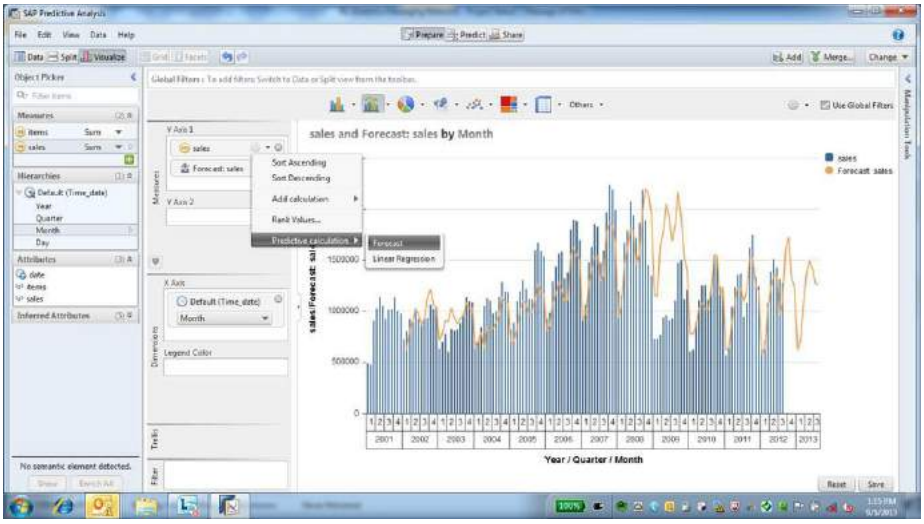


*Figura 24: Visualización de pantalla de SAP Predictive Analytics*

Los usuarios de alta demanda pueden extender y personalizar las funcionalidades agregando sus propios scripts de R.

Funcionalidades:

- Automatización de la preparación de los datos, modelación predictiva y desarrollo, y fácil reentrenamiento.
- Aprovechamiento del scoring predictivo en la base de datos para una amplia variedad de sistemas objetivo.
- Apalancamiento de las capacidades de visualización avanzadas para revelar información rápidamente.
- Integración con R para habilitar un gran número de algoritmos y scripts en R personalizados.

- Despliegue autónomo de analíticas predictivas SAP o con SAP HANA.

Licenciamiento: Software propietario con Shareware de 30 días.

Para más información consultar los siguientes enlaces:

- Página Oficial: https://www.sap.com/latinamerica/products/predictive-analytics.html
- Documentación: https://www.sap.com/latinamerica/products/predictive-analytics/features.html
- Comunidad: https://www.sap.com/latinamerica/community.html
- Tutoriales: https://www.sap.com/developer/tutorial-navigator.html

### 3.3.5. Herramientas de analítica de IBM

IBM ofrece productos y soluciones de analíticas predictivas de fácil uso, que hacen frente a las necesidades específicas de los diferentes usuarios y niveles de habilidad, desde principiantes hasta expertos analíticos.

Ayuda a los negocios a transformar datos en información predictiva para guiar decisiones e interacciones de primera línea, predecir lo que los clientes quieren y lo que harán después para aumentar la rentabilidad y la retención, maximizar la productividad, procesos y bienes, detectar y prevenir fraudes antes de que afecten a la organización, y ejecutar análisis estadísticos.

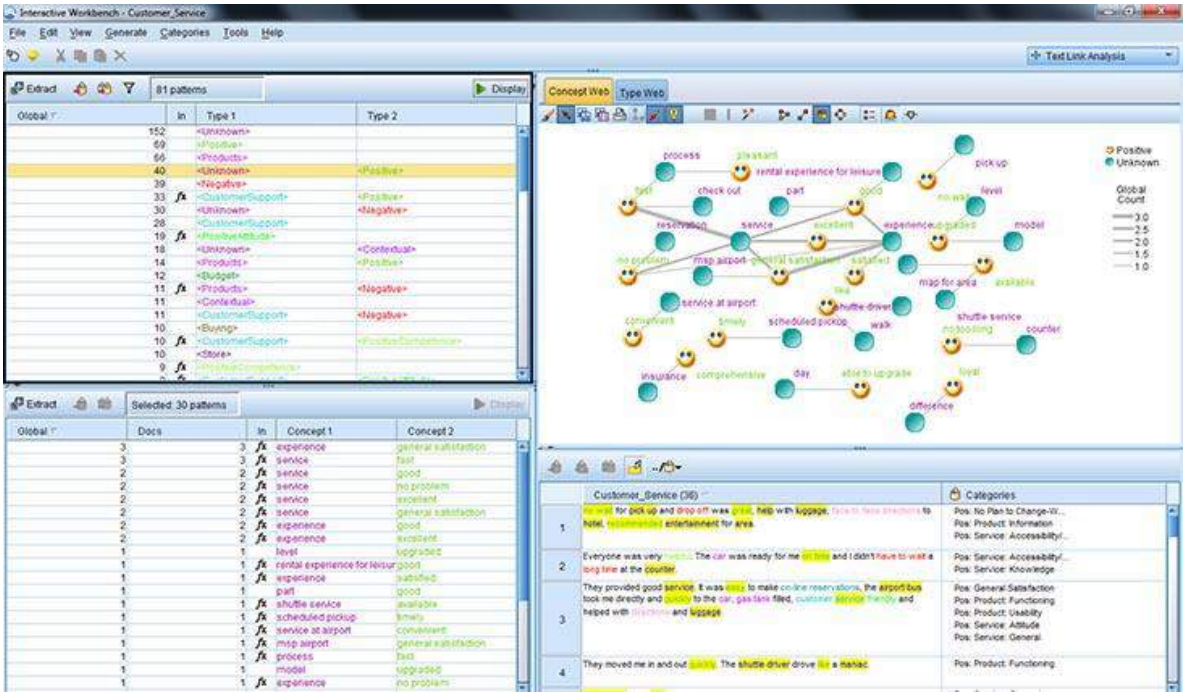

*Figura 25: Visualización de pantallas de herramientas IBM*

## Familia IBM SPSS

La familia de software IBM SPSS ayuda a las organizaciones a predecir eventos futuros y de esta forma poder actuar de manera proactiva para obtener mejores resultados empresariales. Al incorporar analíticas de predicción en sus operaciones diarias las organizaciones se convierten en empresas de predicción con capacidad de direccionar y automatizar las decisiones para cumplir los objetivos de negocio y lograr una ventaja competitiva cuantificable. SPSS proporciona software para análisis de predicción que complementa el IBM Cognos Business Intelligence de consulta e informe de los productos.

| Servicios de colaboración y despliegue de IBM SPSS | Funcionalidades | Tipo de Software |
|---|---|---|
| IBM SPSS Data Collection | Proporciona un conjunto de herramientas que otorgan beneficios para desarrollar exclusivamente enfoques de encuesta. | Software Propietario |
| IBM Analytical Decision Management | Conjunto de herramientas para ayudar a automatizar las decisiones de gran volumen en toda la empresa. | |
| IBM SPSS Modeler | Ayuda a las empresas descubrir conocimientos claves, patrones y tendencias en los datos. | |

| Servicios de colaboración y despliegue de IBM SPSS | Funcionalidades | Tipo de Software |
|---|---|---|
| IBM SPSS Statistics | Estadísticas avanzadas y administración de datos para investigar problemas de negocios. | |

Cada uno de estos cuenta con versión Shareware de 10 o 30 días.

Para más información consultar las siguientes páginas:

- Página oficial: https://www.ibm.com/analytics/spss-statistics-software
- Documentación y tutoriales: https://www.ibm.com/support/knowledgecenter/es/
- Comunidad: https://www.ibm.com/communities/

**Watson Analytics**

Watson Analytics es un servicio inteligente de análisis y visualización de datos que se puede utilizar para descubrir rápidamente a patrones y significado en sus datos. Con el descubrimiento orientado de datos, el análisis predictivo automatizado y capacidades cognitivas, como el diálogo natural del idioma, puede interactuar con datos conversacionalmente para obtener respuestas que usted entienda.

Funcionalidades:

- Diálogo en lenguaje natural.
- Análisis predictivo automatizado.
- Análisis de un clic.
- Descubrimiento inteligente de datos.
- Análisis simplificado.
- Analítica avanzada accesible.
- Paneles de instrumentos de autoservicio.

Licenciamiento: Software propietario con Crippleware disponible de 30 días.

Para más información consultar:

- Página principal: https://www.ibm.com/watson-analytics
- Documentación y tutoriales: https://www.ibm.com/support/knowledgecenter/es/
- Comunidad: https://www.ibm.com/communities/

## 3.3.6. TIBCO Spotfire

Es una plataforma de analíticas de clase empresarial que ayuda a los usuarios a explorar rápidamente para desarrollar conocimiento accionable, sin requerir la intervención de TI. TIBCO Spotfire proporciona tableros interactivos, visualizaciones y analíticas predictivas y conducidas por eventos para facilitar la comprensión inmediata en cualquier dispositivo.

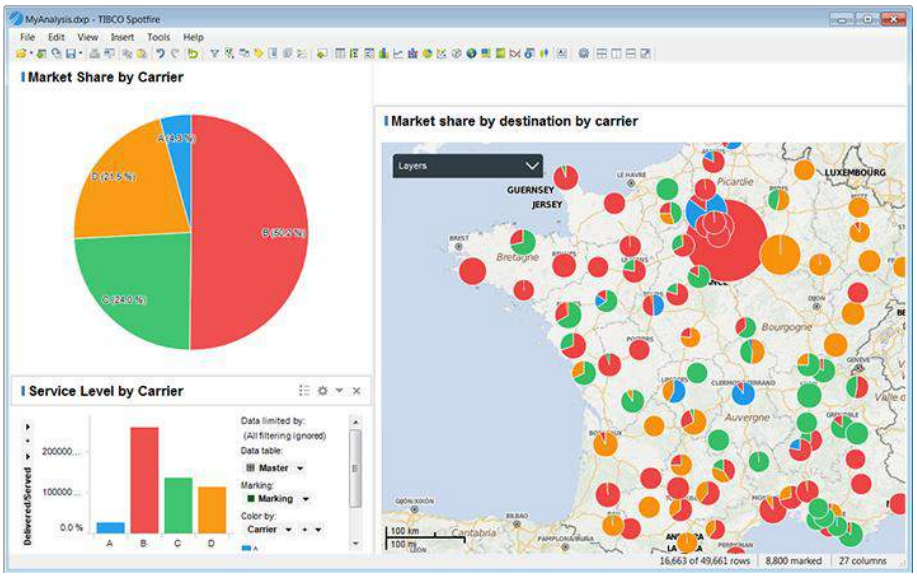


*Figura 26: Visualización de pantalla de TIBCO Spotfire*

Spotfire Analyst hace el ciclo analítico rápido, comprensible y fácil para una gran variedad de usuarios, permitiéndoles generar información útil al momento para realizar análisis adecuados, accediendo a fuentes sin requerir asistencia de TI. Da visibilidad a lo desconocido al publicar instantáneamente aplicaciones analíticas, construir indicadores de negocio y métricas para usuarios móviles y realizar secuencias de comandos en R.

TIBCO Spotfire Statistics Services (TSSS) se apalanca de los análisis predictivos en la base de datos a través de Teradata Aster y permite la integración a R, S+, SAS, MATLAB y otras aplicaciones (PET Research, 2018). Por su parte TIBCO Enterprise Runtime for R (TERR) permite ejecutar scripts de R y paquetes.

Spotfire es una plataforma extensible donde se pueden crear herramientas analíticas avanzadas. Esto es hecho a través de código C# en la interfaz Spotfire.

Funcionalidades:

- Descubrimiento de datos.
- Análisis predictivo.
- Análisis de Big Data.

- Análisis de localización.
- Análisis de nivel empresarial.

Licenciamiento: Software propietario, cuenta con Shareware para educación y sin fines de lucro válido por un año.

Para más información consultar los siguientes enlaces:

- Página oficial: https://www.tibco.com/
- Documentación y tutoriales: https://docs.tibco.com/
- Comunidad: https://community.tibco.com/

### 3.3.7. Databricks Unified Analytics

Esta plataforma fue concebida por los creadores originales de Apache Spark, unifica a la ciencia y la ingeniería de datos a través del ciclo de vida de machine learning. (Databricks, 2018). Es un servicio de clúster Apache Spark basado en la nube.

Databricks permite a los usuarios consumir datos e información a través de la interfaz con la que se sientan más cómodos. Proporciona un espacio de trabajo interactivo que expone las interfaces nativas de R, Scala, Python, y SQL de Spark, una REST API para acceso programático remoto, la capacidad de ejecutar trabajos arbitrarios de Spark desarrollados sin conexión, y soporte sin interrupciones para aplicaciones de terceros como BI y herramientas de dominio específico.

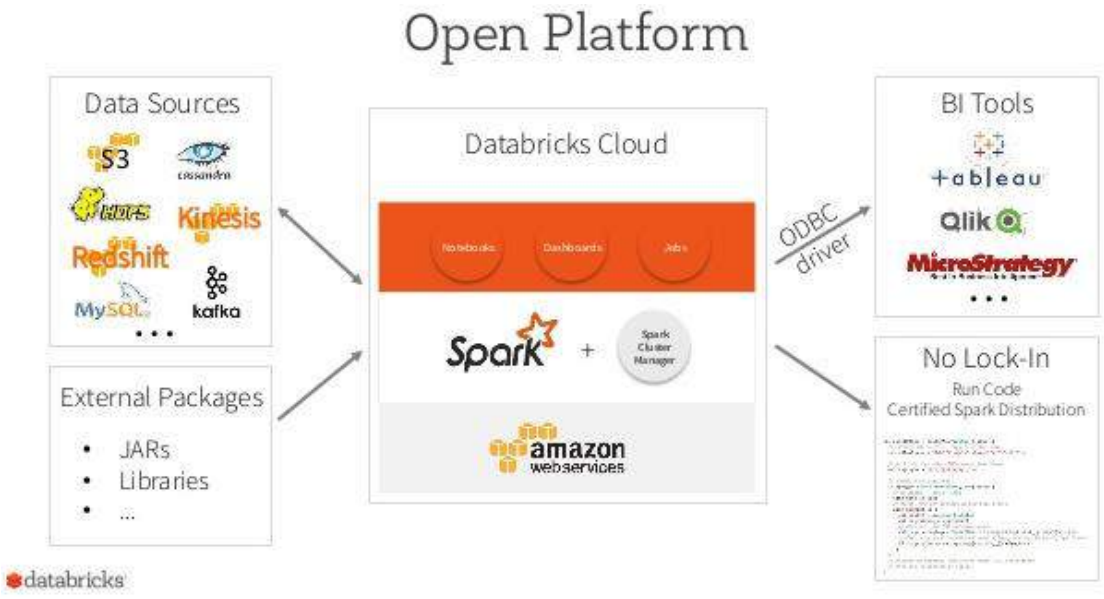


*Figura 27: Integración de Databricks*

Manipula todos los procesos analíticos, desde la extracción y transformación de los datos hasta el entrenamiento y desarrollo de modelos, lenguajes y funcionalidades vía notebooks y a través de la interfaz de programación de aplicaciones.

Databricks proporciona un espectro de características únicas y robustas que permiten a las empresas aprovechar el poder de la inteligencia artificial. Algunas de sus características clave incluyen: optimizaciones para el entorno de la nube, administración de costos, características de exploración integradas, herramientas de producción incorporadas y herramientas de seguridad. Las características forman una columna vertebral sobre la cual las empresas pueden agilizar los procesos y acelerar los procesos para aumentar la productividad.

Funcionalidades:

- Programador flexible.
- Notificaciones.
- Ejecución de notebooks como trabajos.
- Adición de librerías Spark y aplicaciones personalizadas a los trabajos.
- Soporte de clúster flexible.
- Control de acceso basado en roles.
- Espacio de trabajo interactivo.
- Colaboración.
- Editar y compartir en vivo.
- Notebooks como scripts.
- Soporte multilenguaje.

Licenciamiento: Software propietario, cuenta con su versión gratuita Community Edition que es un Crippleware.

Para más información consultar los siguientes enlaces:

- Página oficial: https://databricks.com/
- Documentación: https://docs.databricks.com/index.html
- Comunidad: https://community.cloud.databricks.com/login.html

### 3.3.8. Domino

Domino proporciona una plataforma abierta y unificada para construir, validar, entregar y monitorear modelos a escala. Esto acelera la investigación, genera colaboración, aumenta la velocidad de iteración y elimina la fricción de implementación para ofrecer modelos impactantes.

Permite descubrir, compartir y reutilizar fuentes de datos, incluidas bases de datos en la nube y sistemas distribuidos como Hadoop y Spark, ejecutar cargas de trabajo de desarrollo y producción en contenedores totalmente configurables para crear entornos compartidos, reutilizables y revisados, aprovechar el cómputo escalable en la nube o en las instalaciones y utilizar las últimas técnicas de Deep Learning mediante el acceso con un clic al hardware de la Unidad de Procesamiento Gráfico (GPU); así mismo, permite capturar automáticamente todas las dependencias para cada experimento con Reproductibility Engine, incluidos los datos, el código, el entorno, los parámetros, los resultados y la discusión.

Algunas de sus características clave incluyen: optimizaciones para el entorno de la nube, administración de costos, características de exploración integradas, herramientas de producción incorporadas y herramientas de seguridad.

Licenciamiento: Software propietario con un Crippleware en la nube para educación.

Para más información consultar los siguientes enlaces:

- Página oficial: https://www.dominodatalab.com/
- Documentación y tutoriales: https://support.dominodatalab.com/hc/en-us
- Comunidad: https://blog.dominodatalab.com/

### 3.3.9. Matlab

La plataforma de Matlab está optimizada para resolver problemas de ingeniería y ciencia. El lenguaje basado en matrices es la manera más natural para expresar matemáticas computacionales, lo que hace sencillo capturar las ideas matemáticas detrás de los usuarios, y facilita la escritura, lectura, entendimiento y mantenimiento de los códigos. Las operaciones matemáticas son distribuidas en los múltiples núcleos de las computadoras de los usuarios, y las llamadas a librerías están fuertemente optimizadas.

El lenguaje Matlab también provee características de lenguajes de programación tradicionales, incluyendo flujo de control, manipulación de errores, programación orientada a objetos, pruebas de unidades e integración de fuentes de control.

Proporciona un entorno de escritorio sintonizado para ingeniería iterativa y flujos de trabajo científico. Soporta lenguajes incluyendo C/C++, Java, .NET y Python, así como lenguajes para sistemas integrados. El código Matlab puede ser fácilmente desplegado a los sistemas Hadoop.

Permite aplicar estadísticas, machine learning y herramientas de visualización de datos que no se pueden almacenar en memoria.

Gracias a MATLAB, se han implementado miles de aplicaciones de mantenimiento predictivo, analítica de sensores, finanzas, y electrónica de comunicaciones. Facilita las partes más difíciles de machine learning permitiendo apuntar y hacer clic para entrenar y comparar modelos, cómo hacer técnicas de procesamiento de señales. (MathWorks, 2018).

Funcionalidades:

- Lenguaje de alto nivel aplicado a cálculos por parte de científicos e ingenieros.
- Ambiente de escritorio para la exploración iterativa, diseño y resolución de problemas.
- Gráficos para la visualización de datos y herramientas para gráficas personalizadas.
- Toolbox para ajuste de curvas, clasificación de datos, análisis de señales y otras tareas.
- Herramientas para construir aplicaciones con interfaces de usuario personalizadas.
- Funciones para código C/C++, Java, .Net, Python, SQL, Hadoop, y Excel.

Licenciamiento: Software propietario con Shareware de 30 días.

Para más información consultar los siguientes enlaces:

- Página oficial: https://la.mathworks.com/
- Documentación: https://la.mathworks.com/help/matlab/
- Comunidad: https://la.mathworks.com/matlabcentral/
- Tutoriales: https://la.mathworks.com/support/learn-with-matlab-tutorials.html

## 3.3.10. Lavastorm Analytics Engine

Proporciona un entorno analítico poderoso, visual y versátil para abordar los desafíos analíticos más difíciles y críticos para el negocio en una amplia gama de negocios.

Lavastorm Analytics Engine, ayuda a identificar problemas rápidamente, acelerar la resolución de estos y a implementar controles persistentes.

Posee la capacidad de mejorar el negocio a través de más de 100 bloques analíticos configurables que los usuarios ensamblan, ejecutan y reutilizan, para realizar los análisis que necesitan y controles automáticos persistentes.

Puede escalar a la perfección desde una única instalación de escritorio a una implementación de granja de servidores de alto rendimiento capaz de satisfacer las necesidades analíticas de la empresa más exigente.

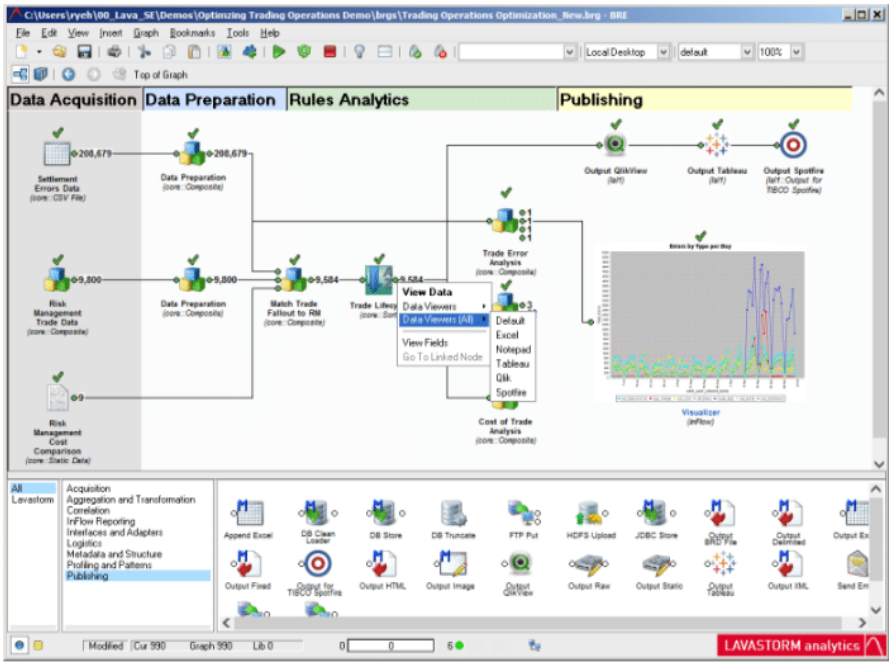


*Figura 28: Visualización de pantalla de Lavastorm Analytics Engine*

Funcionalidades

- Integración de datos / Extracción, Transformación y Carga (ETL): Permite adquirir rápidamente, transformar, combinar y enriquecer datos de prácticamente cualquier

fuente, incluidas las fuentes de Big Data, como MongoDB y Hadoop, aplicaciones basadas en la nube, almacenes, bases de datos, sistemas de empresas, sistemas heredados y archivos independientes.

- Gestión de calidad de datos: Inspección y análisis continuo de capacidades para descubrir y corregir sistemáticamente problemas de datos que afectan la precisión, integridad, coherencia, validez, puntualidad y accesibilidad.
- Análisis y descubrimiento de datos: Permite combinar y analizar datos para responder preguntas ad hoc, explorar ideas, exponer tendencias, descubrir patrones ocultos, descubrir causas raíz e identificar excepciones.

Licenciamiento: Software propietario, con versión Public Edition tipo Crippleware.

Para más información consultar los siguientes enlaces:

- Página Principal: http://www.lavastorm.com
- Comunidad: https://lavastorm.zendesk.com/hc/en-us
- Documentación: https://preview-desktop.lavastorm.com/?__hstc=215508872.ae40e3e1e3c90f1ce90a80290e7ac73e.1535388649037.1535388649037.1535388649037.1&__hssc=215508872.1.1535388649038&__hsfp=3906306934
- Tutoriales: http://www.lavastorm.com/product/lavastorm-academy/

## 3.3.11. Teradata Analytics Platform

La plataforma Teradata Analytics Platform simplifica y mejora la experiencia del usuario al proporcionar análisis a escala, integrando los mejores motores analíticos, admitiendo las herramientas y lenguajes más populares y operacionalizando los análisis a escala empresarial. Teradata Analytics Platform le permite utilizar sus herramientas de analítica favoritas, los idiomas más adecuados y los mejores motores analíticos para acceder a una mayor variedad de datos, incluidas nuevas fuentes. Proporciona un entorno de análisis e información integrado que ofrece las mejores funciones de analítica y motores a escala, además de considerar un soporte para múltiples tipos de datos, formatos y almacenamiento de datos heterogéneos.

Teradata Analytics Platform utiliza Teradata SQL Engine, que ofrece una amplia gama de funciones analíticas integradas, lo que permite realizar análisis descriptivos, predictivos y prescriptivos utilizando datos integrados. Es un potente motor que permite la implementación de análisis complejos y algoritmos analíticos sin exponer su complejidad, a través de un lenguaje SQL simple e intuitivo ampliamente utilizado y compatible con sus herramientas analíticas favoritas. Su arquitectura patentada de procesamiento paralelo masivo (MPP) permite que las cargas de trabajo de analítica complejas se desglosen y

distribuyan en múltiples procesos paralelos que se desempeñen de la manera más eficiente posible.

Teradata Analytics Platform incluye un motor de aprendizaje automático que permite a los científicos de datos utilizar funciones de machine learning. Teradata Analytics Platform también proporciona más de 180 funciones analíticas preconstruidas para transformar, preparar, analizar y visualizar datos multiestructurados, abordando así una gran cantidad de casos de uso comercial.

Teradata Analytics Platform proporciona un motor de procesamiento de gráficos nativo para el análisis de gráficos que identifica y mide las relaciones entre personas, productos y procesos. Además, proporciona funciones analíticas preprogramadas para facilitar la resolución de problemas empresariales complejos, como análisis de redes sociales, influenciadores, detección de fraudes, gestión de la cadena de suministro, análisis de redes, detección de amenazas, y lavado de dinero.

Funcionalidades:

- Permite conexiones con Hadoop.
- Usa JSON para soportar el almacenamiento y procesamiento dentro de la base de datos de Teradata.
- Permite el almacenamiento de datos.
- Es una herramienta escalable.
- Tiene habilidad de modelado para los negocios.

Licenciamiento: Software propietario.

Para más información consultar los siguientes enlaces:

- Página oficial: https://www.teradata.com/
- Documentación: https://www.info.teradata.com/browse.cfm
- Comunidad: https://community.teradata.com/

# 4. Herramientas de Inteligencia de Negocios

En esta sección se hablará acerca de las principales herramientas de inteligencia de negocios. Estas herramientas sirven para poder visualizar tableros de control y facilitar la toma de decisiones. En concreto, este grupo de herramientas se utilizan para hacer una mejor descripción e interpretación de conjuntos de datos de una organización.

Las tendencias en inteligencia de negocios nacen con la finalidad de transformar una serie de datos en información. A la hora de tomar decisiones, una empresa debe estudiar dicha

información mediante unos datos analizados y procesados para ser, posteriormente, utilizados de manera adecuada.

La Inteligencia de negocios es una metodología que tiene por objetivo aumentar el rendimiento de una institución y su efectividad mediante una inteligente organización de sus datos, como pueden ser resultados de operaciones diarias, transacciones, entre otros.

Dentro de este revolucionario sector, los datos no son solo números, sino también cifras que proyectan una gran inversión por parte de importantes empresas de inteligencia de negocios.

Este apartado permite descubrir algunas de las tecnologías de BI más relevantes en el mercado.

## 4.1. Tendencias de herramientas de inteligencia de negocios

En la siguiente imagen se presenta la tendencia de las herramientas de inteligencia de negocios más representativas de acuerdo con el cuadrante mágico de Gartner analizado en febrero de 2018.


*Figura 29: Tendencias de herramientas de BI en el cuadrante mágico de Gartner para plataformas de Business Intelligence, 2018*

## 4.2. Power BI

Power BI es un servicio de análisis de negocio basado en la nube que proporciona una vista única de los datos más críticos del negocio. Permite supervisar el estado de la empresa mediante un panel activo, crear informes interactivos enriquecidos con Power BI Desktop y obtener acceso a los datos sobre la marcha con aplicaciones Power BI Mobile nativas. Es fácil, rápido y gratuito.

Las aplicaciones de Power BI incluyen paneles, informes y conjuntos de datos que proporcionan a todos los usuarios una vista personalizada de sus métricas empresariales más importantes.

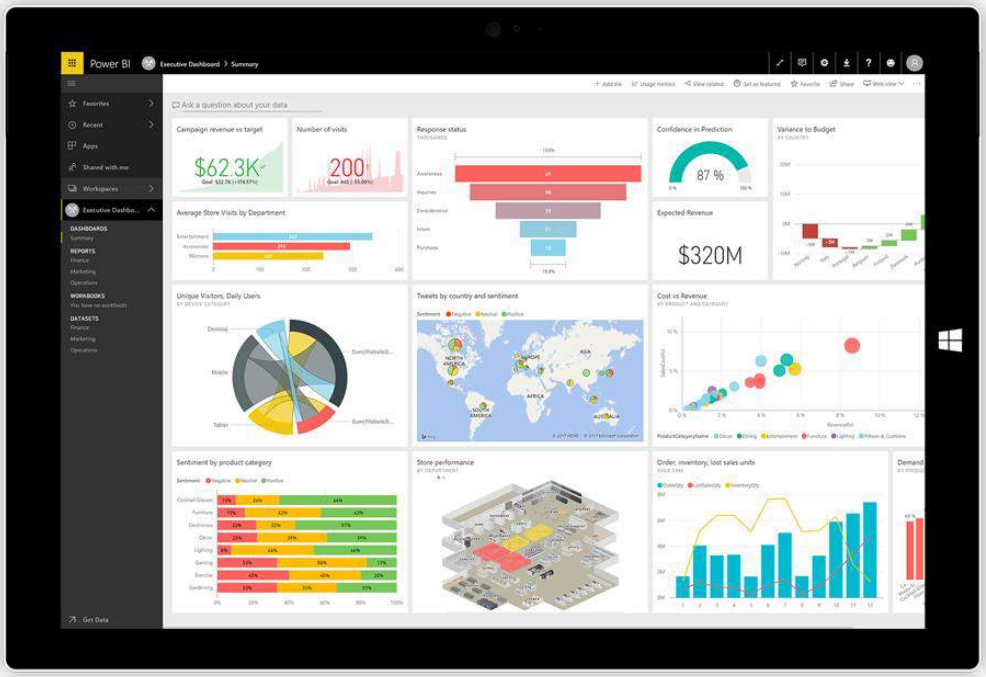


*Figura 30: Visualización de pantalla de Power BI*

Permite usar el API tipo REST abierta y basada en estándares para integrar la aplicación o servicio con Power BI, lo que ayuda a entregar las soluciones con más rapidez. Cuenta también con aplicaciones táctiles nativas para Windows, iOS y Android, además cuenta con la opción de publicar en Power BI web. Otra de sus características es que se pueden formular preguntas en lenguaje natural y obtener como respuesta los gráficos apropiados.

La integración de Power BI permite el acceso a orígenes de datos locales, bases de datos y servicios en la nube, así como la interacción con los gráficos a través de filtros, generar alertas, y consultar datos de Power View.

Algunas de sus ventajas son el incremento de la eficiencia, ligereza en la elaboración de informes, acceso a los datos importantes, así como el aumento de la documentación en español.

Funcionalidades:

- Creación de paneles, es decir tableros en los que de un solo vistazo se pueda evaluar toda la información.
- Creación de informes.
- Vista de edición.
- Vista de lectura.
- Power BI web para compartir informes.
- Más de 65 fuentes de datos: SQL, Excel, xml, JSON, csv, web, servicios de análisis, etc.

Licenciamiento: Cuenta con versión Crippleware.

Para más información consultar los siguientes enlaces:

- Página oficial: https://powerbi.microsoft.com/es-es/
- Documentación y tutoriales: https://docs.microsoft.com/en-us/power-bi/
- Comunidad: http://community.powerbi.com/

## 4.3. Qlik view

Es una plataforma enfocada al análisis visual de datos y aplicaciones interactivas que tiene por objetivo mejorar el proceso de acceso a los datos de cara al usuario, como acceder a ciertas visualizaciones “limpias” y fáciles de comprender, diseños de gráficos llamativos, entre otros.

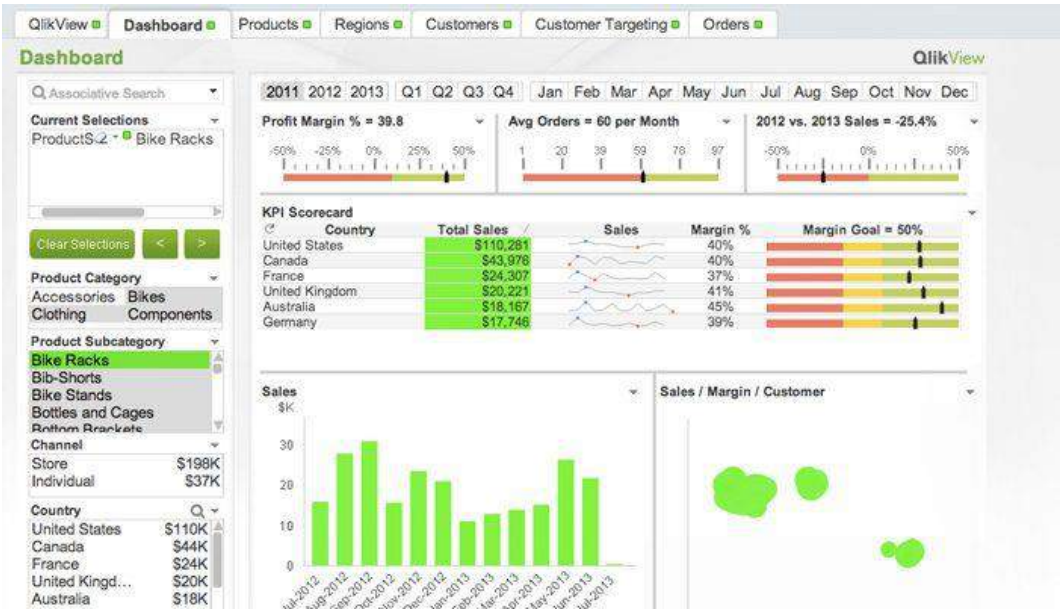


*Figura 31: Visualización de pantalla de Qlik View*

Además, Qlik funciona con el motor Qlik Indexing Engine (QIX), motor de indexación asociativa de datos más potente del mundo, según explican en la web principal de la herramienta. El motor asociativo de Qlik permite combinar cualquier tipo y número de fuentes de datos para explorar libremente todos los datos y cambiar, al instante, la manera de pensar en función de lo que se observa.

Algunas de sus ventajas son el acceso a documentación, tutoriales, análisis colaborativos e interactivos, la creación de informes automatizados y ahorro en tiempo en la creación de informes.

Funcionalidades:

- Integración de datos.
- Mejor gestión en RAM para elaboraciones en memoria.
- Exportación de cuadros de mando en formato xml.
- Guardar datos en formato texto desde script.
- Listas con más de un campo.
- Listas jerárquicas.
- Modelos asociativos.
- Soporta el descubrimiento de datos móviles.

Licenciamiento: Su versión QlikView Personal Edition es un Freeware.

Para más información consultar los siguientes enlaces:

- Página oficial: https://www.qlik.com/us

- Documentación: https://help.qlik.com/en-US/qlikview/November2017/Subsystems/QMC/Content/QMC_Documents.htm
- Comunidad: https://community.qlik.com/welcome
- Tutoriales: https://help.qlik.com/en-US/sense/3.1/Content/Tutorials.htm

## 4.4. Tableau

Esta otra herramienta BI sirve para la visualización interactiva de los datos, con los que los usuarios pueden interactuar de varias maneras: comparando datos, filtrándolos o creando una conexión entre unas variables y otras.


*Figura 32: Visualización de pantalla de Tableau*

Algunas de sus ventajas es que se tiene acceso a grandes bases de datos como MySQL, Oracle, Greenplum y Microsoft. Se puede utilizar la API para la extracción sistemática de los datos. Adicionalmente, acepta formatos Acces, texto y Excel.

Funcionalidades:

- Historial de revisiones.
- Vistas de licenciamiento.
- Suscripciones para otros usuarios para visualización de tableros.
- Administración de dispositivos móviles.
- Conector de datos 2.0
- Actualización de ETL.

Licenciamiento: Software propietario con disponibilidad de Shareware de 14 días.

Para más información consultar los siguientes enlaces:

- Página oficial: https://www.tableau.com/es-es
- Documentación: https://onlinehelp.tableau.com/current/pro/desktop/en-us/help.htm
- Comunidad: https://community.tableau.com/welcome
- Tutoriales: https://www.tableau.com/es-es/learn/training

## 4.5. Sisense

El software de BI de Sisense hace fácil revelar instantáneamente información del negocio de datos complejos, de cualquier recurso de datos y de cualquier tamaño.

La compañía fue nombrada recientemente en la lista Forbes 2017 Cloud 100, la lista definitiva de las 100 mejores empresas privadas de la nube en el mundo.

Sisense está concentrada en redefinir cada aspecto de las analíticas de negocios para hacer fácil para cualquiera descubrir información del negocio. Es fácil limpiar y combinar datos contaminados que vienen de muchas fuentes y las analíticas visuales permiten explorar los datos y obtener información instantáneamente.


*Figura 33: Visualización de pantalla de Sisense*

Funcionalidades:

- Conecta y analiza datos de múltiples fuentes.
- Interfaz intuitiva.
- Sugerencias inteligentes.
- Integra funciones de R para mejores analíticas predictivas.
- Acceso a cientos de visualizaciones clásicas y Open Source.
- Detección de anomalías.
- Alertas automáticas.
- Alertas móviles.

Licenciamiento: Software propietario con versión completa de prueba gratuita de 14 días.

Para más información consultar los siguientes enlaces:

- Página oficial: https://www.sisense.com/
- Documentación: https://documentation.sisense.com/
- Comunidad: https://support.sisense.com/hc/en-us
- Tutoriales: https://www.sisense.com/training/

## 4.6. MicroStrategy

Permite aprovechar las mejores fuentes y sistemas de datos de la industria con facilidad. Con docenas de controladores y puertas de enlace listas, MicroStrategy facilita comenzar a obtener respuestas de los conjuntos de datos más desafiantes del mundo a través de la creación de informes sofisticados y personalizados.

Mejora la productividad en toda su organización mediante la ampliación del acceso a los datos de más equipos comerciales. Ofrece la libertad de explorar los datos por su cuenta a los usuarios de negocio sin sacrificar el gobierno de datos, la escalabilidad o la seguridad.


*Figura 34: Visualización de pantalla para MicroStrategy*

Faculta la automatización de la entrega de informes personalizados, dosieres y cuadros de mando para cada usuario y permite aprovechar la integración con los paquetes estadísticos de terceros y conectarse fácilmente a fuentes de Big Data.

Su herramienta MicroStrategy SDK proporciona la característica de crear aplicaciones web personalizadas de informes que puedan satisfacer cualquier tipo de negocio que necesite.

Funcionalidades:

- Creación de infraestructura orientada a objetos.
- Generación y distribución de aplicaciones de inteligencia de negocios universales.
- Administración de los ciclos de vida de un proyecto.
- Repositorio centralizado de metadatos reutilizables.
- Reutilización de objetos.
- Arquitectura escalable en memoria.
- Conectores de datos optimizados.
- Herramientas de optimización de alto rendimiento.

Licenciamiento: Software propietario con disponibilidad de Shareware de 30 días.

Para más información consultar los siguientes enlaces:

- Página oficial: https://www.microstrategy.com/es
- Documentación: https://microstrategyhelp.atlassian.net/wiki/spaces/MSTRDOCS/overview?mode=global
- Comunidad: https://community.microstrategy.com/s/

## 4.7. Pentaho BI Suite

Es un conjunto de programas libres para generar inteligencia empresarial, que incluye herramientas para generar informes, minería de datos, ETL, aprendizaje de máquina, así como para Big Data.

Dentro de sus herramientas están:

- Pentaho Analysis Services: Es un servidor OLAP (procesamiento analítico en línea) escrito en Java. Es compatible con el MDX (expresiones multidimensionales) y el lenguaje de consulta XML para el Análisis y especificaciones de la interfaz olap4j.
- Pentaho Reporting: Consiste en un motor de presentación, capaz de generar informes programáticos sobre la base de un archivo de definición XML. Un uso notable de esta herramienta es el Generador de informes para OpenOffice.org
- Pentaho Data Mining: Es una envoltura alrededor del proyecto Weka. Es una suite de software que usa estrategias de aprendizaje de máquina, aprendizaje automático y minería de datos. Cuenta con series de clasificación, de regresión, de reglas de asociación, y de algoritmos de agrupación, para así apoyar las tareas de análisis predictivo.
- Pentaho Dashboard: Es una plataforma integrada para proporcionar información sobre sus datos, donde se pueden ver informes, gráficos interactivos y los cubos creados con las herramientas Pentaho Report Designer.
- Pentaho para Apache Hadoop: Es un conector de bajo nivel para facilitar el acceso a grandes volúmenes de daos manejados en el proyecto Apache Hadoop.

Funcionalidades:

- Existe soporte para más de 34 tipos de bases de datos (SQL, MS Access, Teradata, Oracle, Netezza, etc.).
- Soporte para lectura de archivos XML, Excel, texto plano, etc.
- Soporte para lenguajes Javascript y expresiones regulares, además de SQL.
- Permite ejecutar múltiples trabajos secuencialmente y en paralelo.

Licenciamiento: La mayoría del software que comprende Pentaho Business Analytics es de código abierto, con licencia bajo la licencia pública general GNU versión 2. Business Analytics también contiene un gran volumen de bibliotecas de código abierto de terceros que tienen licencia bajo una serie de licencias diferentes. La mayoría de este software se puede redistribuir libremente, con las notables excepciones de los siguientes programas creados por Pentaho:

- Dashboard Designer.
- Analyzer.
- Interactive Reporting.

- Various individual BI Platform JARs.

Para más información consultar los siguientes enlaces:

- Página oficial: https://www.hitachivantara.com/go/pentaho.html?source=pentaho-redirect
- Comunidad: https://community.hitachivantara.com/

# 5. Algoritmos de analítica avanzada

Existen dos grandes grupos que segmentan los algoritmos existentes dentro de la analítica avanzada, estos son el aprendizaje supervisado y no supervisado; todas las herramientas de analítica presentadas en la sección 3 incluyen la mayoría de estos algoritmos.

Por un lado, los algoritmos de aprendizaje supervisado trabajan con datos etiquetados intentando encontrar una función que, dadas las variables de entrada les asigne la etiqueta de salida adecuada. El algoritmo utiliza un conjunto de datos de entrenamiento y un conjunto de prueba. Dentro del aprendizaje supervisado encontramos los problemas de predicción (predicciones meteorológicas, de ventas, de crecimiento, etc.) y clasificación (identificación de dígitos, diagnósticos, detección de fraude, entre otros), para los que se ocupan técnicas como árboles de decisión, regresión, máquinas de soporte vectorial (SVM), métodos de ensamble, algoritmo de clasificación Naïve Bayes, entre otros.

En cuanto al aprendizaje no supervisado, éste tiene lugar cuando no se dispone de datos etiquetados para el entrenamiento. Sólo conocemos los datos de entrada, pero no existen datos de salida que correspondan a una entrada determinada. Por tanto, sólo podemos describir la estructura de los datos, para intentar encontrar algún tipo de organización que simplifique el análisis. Por ello, tienen un carácter exploratorio (Recuero, 2017). Dentro de esta categoría se encuentran los problemas de clustering, el análisis de componentes principales, entre otros.

## 5.1. Aprendizaje supervisado

**Árboles de decisión:** Son modelos precisos, estables y más sencillos de interpretar básicamente porque construyen unas reglas de decisión que se pueden representar como un árbol. A diferencia de los modelos lineales, pueden representar relaciones no lineales para resolver problemas. Al ser más precisos y elaborados, se gana en capacidad predictiva, pero se pierde en rendimiento. Pueden emplearse para predecir la respuesta del público ante el lanzamiento de un nuevo producto o para averiguar la idoneidad de una campaña de marketing (Rayón, 2017).

**Regresión lineal:** Esta técnica trata de encontrar una línea que se ajuste bien a la nube de puntos que se disponen. Este modelo posee el problema de datos no estacionarios, dinámicas altamente variables, por lo que no puede ser modelada con metodologías tan simples, es decir, no todo se puede modelar con una recta. El análisis de regresión permite desde desarrollar pronósticos de futuro, hasta identificar los factores que mayor incidencia tienen en la generación de beneficios de una corporación o determinar cuánto afectará un cambio en las tasas de interés a una cartera de bonos (Logicalis, 2017).

**Redes neuronales:** Las redes están concebidas como nodos conectados simulando las conexiones de las neuronas en el cerebro, pero no es su principal objetivo. Las habilidades cognitivas de razonamiento adquiridas por esta técnica hacen que sea uno de los algoritmos más usados actualmente. El reconocimiento de imágenes o vídeos, por ejemplo, es un mecanismo complejo que sólo una red neuronal es capaz de abordar. Su complicación es que tienen un entrenamiento lento y requieren de mucha capacidad de cómputo (Rayón, 2017).

**Algoritmo de clasificación Naïve Bayes:** Es un modelo basado en el Teorema de Bayes que permite determinar la probabilidad de posibles resultados. Requiere ser alimentado de datos suficientes, aunque, cuando el volumen de información es muy grande, debe recurrirse a la hipótesis de la independencia condicional, que permite simplificar la expresión del Teorema de Bayes, factorizando la probabilidad. Este tipo de algoritmos de machine learning se utilizan para los softwares de reconocimiento facial o para determinar si la emoción contenida en un texto es positiva o negativa, o incluso, marcar un correo electrónico como spam o no spam (Logicalis, 2017).

**Support Vector Machines (SVM):** Las denominadas SVM basan su funcionamiento en un algoritmo de clasificación binario que facilita a los negocios identificar el género de los usuarios de sus webs o escoger el tipo de publicidad que debe aparecer en la pantalla, entre otras aplicaciones (Logicalis, 2017).

## 5.2. Aprendizaje no supervisado

**Clustering:** Este tipo de algoritmos consisten en agrupar un conjunto de objetos de tal manera que los objetos en el mismo grupo (clúster) sean más similares entre sí que a los de otros grupos (Raona Enginyers, 2017). Existen diversas técnicas de clustering, entre las que destacan los dendogramas, k-medias, por estimación de distancias Este grupo de algoritmos son utilizados para procesos de recomendación.

**Principal Components Analysis (PCA)**: PCA es una técnica que se utiliza para describir un conjunto de datos en términos de nuevas variables (componentes) no correlacionadas (Raona Enginyers, 2017). Principalmente se ocupa para la reducción de dimensionalidad que permite simplificación de procesos, así como de visualizaciones.

**Independent Components Analysis (ICA):** Corresponde a una generalización de PCA. Se ocupa para revelar los factores ocultos que subyacen a conjuntos de variables, mediciones o señales aleatorias. En el modelo, se supone que las variables de datos son mezclas lineales de variables latentes desconocidas, cuyas mezclas también son desconocidas. Las variables latentes se asumen no gaussianas y mutuamente independientes (Raona Enginyers, 2017). Independent Components Analysis es heurístico, razón por la que su uso en Big Data conlleva un alto tiempo de cómputo.

# 6. Conclusiones y recomendaciones

El crecimiento masivo de los datos producidos en los últimos años ha derivado en la creación de herramientas que permiten la extracción, manipulación, depuración, análisis e interpretación de los datos, así como en la introducción de técnicas avanzadas de análisis y modelación de datos que facilitan la explotación de la información que puede ser extraída de estos.

Existen una gran cantidad de plataformas que sirven para visualizar y modelar información con el fin de alimentar los conocimientos dentro del contexto que se maneje. Sea un ciudadano, estudiante, emprendedor, o actor del ecosistema de datos, el lector de este documento puede sacar provecho del conocimiento que puede generar a través del uso de las herramientas aquí presentadas.

Cada plataforma tiene características propias como el costo, capacidades e integración con otros servicios, que pueden ser de gran ayuda para distintos objetivos. Las necesidades de cada persona, proyecto y negocio determinan las características de las herramientas que deben ser utilizadas, por lo que hacer una recomendación para la utilización de una sola plataforma sería equivocado.

# 7. Glosario

**Algoritmo:** Conjunto ordenado de instrucciones o reglas sistematizadas que hace posible la ejecución de determinadas actividades y dar solución a un problema en específico.

**Analítica avanzada:** Es una forma de dar uso a los datos de manera exhaustiva y tomar decisiones con base en estos, se basa en herramientas de análisis estadístico y en la creación de modelos.

**Analítica descriptiva:** Surge a partir del análisis de información histórica, visualizándola de manera que ayude a la comprensión de la situación actual y previa del negocio. El uso de métricas clave o KPIs es fundamental en este tipo de análisis, ya que permiten una visualización más comprensible y emitir notificaciones en caso de valores no esperados.

**Analítica predictiva:** El objetivo consiste en encontrar patrones y tendencias, así como relaciones en los datos que permitan comprender las causas de tal forma que se puedan pronosticar los eventos futuros.

**Apache Hadoop:** Es un framework de software que soporta aplicaciones distribuidas bajo una licencia libre, permite a las aplicaciones con miles de nodos y petabytes de datos.

**Apache Spark:** Es un framework de computación en clúster Open Source orientado a la velocidad.

**API:** Es la interfaz de programación de aplicaciones, y corresponde al conjunto de rutinas, funciones y procedimientos que permite hacer uso de un servicio web de un tercero dentro de una aplicación web propia.

**Big Data:** Término que hace referencia a grandes conjuntos de datos ya sean estructurados o no estructurados. Se caracteriza por su elevado volumen de información, la alta velocidad de procesamiento, y la variedad de datos. El análisis de Big Data proporciona información para tomar mejores decisiones y realizar operaciones estratégicas en los negocios.

**Business Intelligence (BI):** Concepto que hace referencia a la inteligencia empresarial, la creación de conocimiento e información relevante para la toma de decisiones, a partir del análisis de datos recopilados.

**Ciencia de datos:** Campo que permite analizar e interpretar grandes fuentes de datos, a través del uso de herramientas matemáticas, estadísticas, y tecnologías de la información. Su fin principal es la obtención de información a partir de los datos.

**Clúster:** Es un grupo o conglomerado de objetos con características similares.

**ETL:** Significa extraer, transformar, y cargar. Es el proceso que permite mover datos desde múltiples fuentes, reformatearlos, limpiarlos, y cargarlos en otro destino.

**GPU:** Unidad de Procesamiento Gráfico.

**Internet de las cosas:** IoT por sus siglas en inglés es un concepto que habla de la interconexión de los objetos cotidianos con el internet.

**Kubernete**: Es un sistema de código libre para la automatización del despliegue, ajuste de escala y manejo de aplicaciones en contenedores.

**machine learning:** Es una rama de la inteligencia artificial que permite a las computadoras identificar y generalizar patrones y tendencias a partir de ejemplos.

**MapReduce:** Modelo de programación para dar soporte a la computación paralela sobre grandes colecciones de datos en grupos de computadoras.

**Minería de datos:** Conjunto de técnicas y herramientas que permiten explorar grandes cantidades de datos para determinar patrones y tendencias. Utiliza métodos de estadística, inteligencia artificial, machine learning, así como conocimientos de bases de datos. Existe también la minería de texto y de imágenes.

**MongoDB:** Es un sistema de base de datos NoSQL orientado a documentos, desarrollado bajo el concepto de código abierto.

**Multihilo:** Dentro de un proceso programático la programación multihilo hace referencia a la ejecución de múltiples tareas dentro de un mismo proceso.

**Pruebas no paramétricas**: Corresponden a técnicas estadísticas que no suponen ningún modelo probabilístico sobre los datos analizados.

**Pruebas paramétricas:** Son técnicas estadísticas que se basan en la suposición de que los datos analizados provienen de un modelo probabilístico.

**Red neuronal:** Técnica computacional que trata de replicar el comportamiento del cerebro para resolver problemas complejos. Las habilidades cognitivas de razonamiento adquiridas por esta técnica hacen que sea uno de los algoritmos más usados actualmente.

**Regresión:** Técnica estadística que trata de encontrar una línea que se ajuste bien a la nube de puntos que se disponen.

**Script:** Es un documento que contiene instrucciones escritas en códigos de programación.

**Software:** Conjunto de programas, instrucciones y reglas informáticas que permiten ejecutar distintas tareas en una computadora.

**Software libre:** Software que otorga libertades a los usuarios y comunidad, tales como libre uso, copiado, modificación y distribución.

**Weka:** Es una plataforma de software escrita en Java para el aprendizaje automático y la minería de datos.

# 8. Bibliografía


Alteryx Inc. (2018). [Página de inicio]. Recuperado de: https://www.alteryx.com/

Anaconda Inc. (2018). *Anaconda Cloud*. Recuperado de: https://anaconda.org/

Angoss Software Corporation. (2018). [Página de inicio]. Recuperado de: http://www.angoss.com/

Autibux Training. (2018). *Herramientas y librerías Python para machine learning.* Recuperado de: http://blog.auriboxtraining.com/big-data/herramientas-librerias-python-machine-learning/

Aziza, Bruno. (2018, Febrero 25). *Gartner Magic Quadrant for Business Intelligence (BI) 2018: The Good, The Bad, The Ugly….* Recuperado de: http://blog.atscale.com/gartner-magic-quadrant-for-business-intelligence-bi-2018-the-good-the-bad-the-ugly?wvideo=h2jeisy6i1

BBVA. (2014). *Las 5 licencias de software libre más importantes que todo desarrollador debe conocer*. Recuperado de: https://bbvaopen4u.com/es/actualidad/las-5-licencias-de-software-libre-mas-importantes-que-todo-desarrollador-debe-conocer

CAOBA et. al. (2017, Julio). *Perfil alianza CAOBA Reporte técnico*. Recuperado de: http://www.colciencias.gov.co/sites/default/files/upload/convocatoria/anexo_no._1_-_perfil_citizen_data_scientist_caoba.pdf

Clarcat. (Sin Fecha). *RapidMiner: Plataforma de análisis predictivo de código abierto*. Recuperado de: https://www.clarcat.com/rapidminer/

CONPES. (2018). *Política Nacional de Explotación de Datos (Big Data).* Recuperado de: https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3920.pdf

Databricks. (2017). *Databricks Features A Primer*. Recuperado de: http://go.databricks.com/hubfs/pdfs/Databricks-Feature-Primer.pdf

Databricks. (2018). *Databricks Unified Analytics*. Recuperado de: https://databricks.com/

Domino Data Lab. (2018). [Página de inicio]. Recuperado de: https://www.dominodatalab.com/

GNU Octave. (2018). [Página de inicio]. Recuperado de: https://www.gnu.org/software/octave/

Grupo Korporate. (Sin Fecha). *Los cinco tipos de fuentes de datos*. Recuperado de: https://grupokorporate.com/los-cinco-tipos-de-fuentes-de-datos/

H2O.ai. (2018). [Página de inicio]. Recuperado de: https://www.h2o.ai/

IBM. (Sin Fecha). [Página de inicio]. Recuperado de: https://www.ibm.com/mx-es/

KNIME. (Sin Fecha). [Página de inicio]. Recuperado de: https://www.knime.com/

Logicalis. (2017). *Learning machine, los usos del aprendizaje supervisado*. Recuperado de: https://blog.es.logicalis.com/analytics/learning-machine-los-usos-del-aprendizaje-supervisado

MathWorks. (2018). *Analice y modelice datos mediante estadísticas y aprendizaje automático*. Recuperado de: https://la.mathworks.com/products/statistics.html

Microsoft. (2018). *Azure machine learning Studio*. Recuperado de: https://azure.microsoft.com/es-mx/services/machine-learning-studio/

Microsoft. (2018). [Página de inicio]. Recuperado de: https://powerbi.microsoft.com/es-es/

Microsoft. (Sin Fecha). [Página de inicio]. Recuperado de: https://mran.microsoft.com/

Microstrategi Iberica S.L.U. (2018). [Página de inicio]. Recuperado de: https://www.microstrategy.com/es

OpenMind. (2018). Reinventando la empresa en la era digital. Recuperado de: https://www.bbvaopenmind.com/wp-content/uploads/2015/01/BBVA-OpenMind-libro-Reinventar-la-Empresa-en-la-Era-Digital-empresa-innovacion1-1.pdf

Pat Research. (2018). Top 24 predictive analytics free software. Recuperado de: https://www.predictiveanalyticstoday.com/top-predictive-analytics-freeware-software/

Pat Research. (2018). *Top 52 predictive analytics & prescriptive analytics software.* Recuperado de: https://www.predictiveanalyticstoday.com/top-predictive-analytics-software/

Peddibhotla, Geethika. (2015). *Top 20 R machine learning and Data Science packages.* Recuperado de: https://www.kdnuggets.com/2015/06/top-20-r-machine-learning-packages.html

Qlick Tech International AB. (2018). [Página de inicio]. Recuperado de: https://www.qlik.com/

Raona Enginyers S.L. (2017). [Página de inicio]. Recuperado de: https://www.raona.com/los-10-algoritmos-esenciales-machine-learning/

RapidMiner. (2018). *Gartner Magic Quadrant For Data Science And machine learning Platforms.* Recuperado de: https://rapidminer.com/resource/gartner-magic-quadrant-data-science-platforms/

RapidMiner. (2018). *Lightning Fast Data Science Platform*. Recuperado de: https://rapidminer.com/

Rayón, Alex. (2017). *Guía para comenzar con algoritmos de machine learning*. Recuperado de: https://blogs.deusto.es/bigdata/guia-para-comenzar-con-algoritmos-de-machine-learning/

Recuero, Paloma. (2017). *Los 2 tipos de aprendizaje en machine learning: supervisado y no supervisado*. Recuperado de: https://data-speaks.luca-d3.com/2017/11/que-algoritmo-elegir-en-ml-aprendizaje.html

SAP. (Sin Fecha). [Página de inicio]. Recuperado de: https://www.sap.com/index.html

SAS Institute Inc. (Sin Fecha). *Analytics Software y Solutions*. Recuperado de: https://www.sas.com/en_us/home.html

SAS. (2018). [Gráficas] Niveles de Analítica y *Ciclo Analítico*. Recuperado de: https://www.sas.com/es_pe/campaigns/analytics/gestionando-el-ciclo-analitico.html

Sisense Inc. (2018). [Página de inicio]. Recuperado de: https://www.sisense.com/

Tanagra. (2008). [Página de inicio]. Recuperado de: http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html

Tableau Software. (2018). [Página de inicio]. Recuperado de: https://www.tableau.com/

TechTarget. (2018). *Principios de la analítica de datos: una guía esencial*. Recuperado de: https://searchdatacenter.techtarget.com/es/guia/Principios-de-la-analitica-de-datos-una-guia-esencial

Teradata. (2018). [Página de inicio]. Recuperado de: https://www.teradata.com/

The Lab MathWorks Inc. (2018). *Matlab para deep learning*. Recuperado de: https://la.mathworks.com/

TIBCO Sporfire Inc. (2018). [Página de inicio]. Recuperado de: https://spotfire.tibco.com/

Universidad de Oviedo. (Sin Fecha). *Sistemas Inteligentes*. Recuperado de: http://www.aic.uniovi.es/ssii/ssii-t12-aprendizajenosupervisado.pdf

University of Ljubljana. (Sin Fecha). [Página de inicio]. Recuperado de: https://orange.biolab.si/

University of Waikato. (Sin Fecha). [Página de inicio]. Recuperado de: https://www.cs.waikato.ac.nz/ml/index.html

GOBIERNO DE COLOMBIA

MINTIC

GOBIERNO DIGITAL

BIOS