Módulo 3. Aplicaciones y generación de valor

Unidad 1. Aplicación y generación de valor a partir del Big Data

Big Data es un término que se viene utilizando asociado a un sinnúmero de herramientas y tecnologías. En esta unidad buscaremos mostrar en concreto algunas herramientas que se utilizan para su aplicación, explotación y, por ende, generar valor a partir de datos. Algunas de estas herramientas son *open source* (gratuitas), mientras que otras requieren de una licencia paga. Este no es un listado exhaustivo de herramientas, sino más bien un detalle de las más comunes a la hora de explotar los datos y generar valor de ellos. Si bien todas las herramientas que se detallarán son utilizadas para Big Data, pueden ser utilizadas con menor cantidad, variedad o velocidad de datos.

Cabe destacar que, así como Big Data está en constante crecimiento, las herramientas para explotar grandes volúmenes de datos también están en constante avance. No sería extraño que cuando lean esto haya salido alguna herramienta que supere las detalladas. Por lo que aquí va un consejo: no se "enamoren" de ninguna, busquen constantemente nuevas herramientas que puedan ayudarlos a sacar lo mejor de sus datos.

En la segunda parte de esta unidad, buscaremos algunos casos de éxito de la implementación de Big Data. El objetivo es que se puedan llevar ideas concretas de formas de bajar la teoría a la práctica, casos de usos que podrían aplicar en sus empresas o un lineamiento para comenzar una exploración más profunda.

3.1.1 ¿Cómo generar valor a partir de Big Data?

En mi experiencia, trabajando en la explotación y generación de información a través de Big Data, distingo tres tipos de herramientas: **herramientas de explotación o minería de datos, herramientas de visualización y herramientas**

enfocados en *machine learning*. Todas tienen relación entre sí, ya lo veremos más en profundidad en cada caso.

A continuación, detallaré algunas de las herramientas más utilizadas en cada campo. Solo nos enfocaremos en las dedicadas a la generación de valor a través de los datos. Por lo que no haremos foco en la recolección o almacenamiento de los datos, solamente en la explotación.

Como aclaré al comienzo de esta unidad, es un listado orientativo de las herramientas que más se utilizan en cada campo. Es importante poder distinguir el fuerte de cada una y las posibles prestaciones. También les será de mucha utilidad definir si trabajarán con una herramienta *open source* o una que requiera licencia paga. Si bien las herramientas *open source* suelen presentarse como la mejor opción, algunas de las que requieren licencia paga suelen ser más sencillas de utilizar o logran excelentes integraciones con otras herramientas conocidas por los usuarios.

Herramientas de explotación o minería de datos

Estas herramientas permiten bucear en los datos que tenemos para generar hallazgos o *insights* que permitan mejorar la toma de decisiones. A fin de cuentas, si estamos desarrollando una estrategia de Big Data, es para eso, para mejorar la toma de decisiones. Gartner define al *data mining* como

> el proceso de descubrimiento de correlaciones, patrones y tendencias significativas al examinar grandes cantidades de datos almacenados en repositorios de información. La minería de datos [o data mining] emplea tecnologías de reconocimiento de patrones, así como técnicas avanzadas de estadística y matemática. ("Data mining", s.f., https://gtnr.it/2TVtum4).

Este tipo de herramientas permite conectarnos con las diferentes fuentes de datos como, por ejemplo, un *data warehouse on premise* o en la nube, un archivo .csv o .txt, o con cualquier fuente de información estándar que podamos encontrar.

Figura 1: *Data mining*

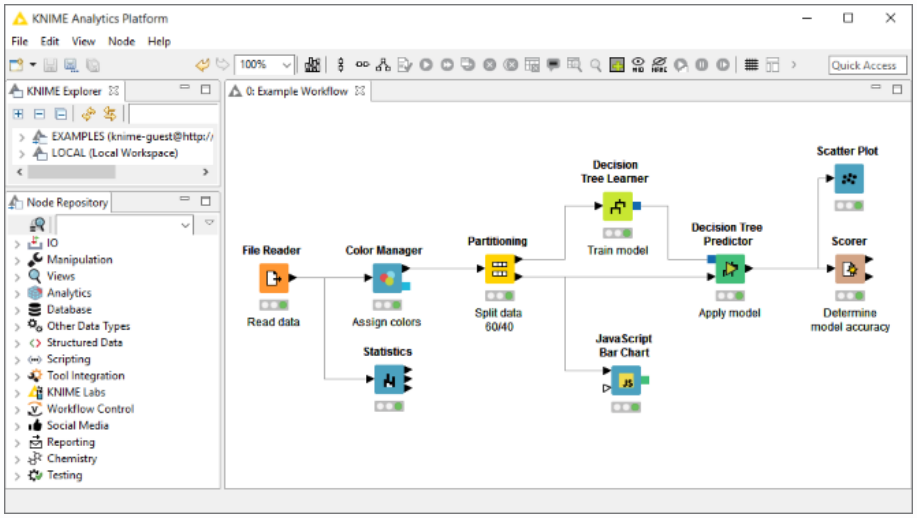
Fuente: [Imagen sin título sobre *data mining*]. (s.f.). Recuperada de http://bit.ly/2UasCJ9

Cabe aclarar que se podría hacer minería de datos en lenguaje SQL, Python o R si fuera necesario. Pero el objetivo de estas herramientas es acercar la minería de datos a otros públicos sin tanto conocimiento de lenguajes de programación.

Aquí van las herramientas más utilizadas en el mundo de la minería de datos:

- **Knime**: es una herramienta *open source*, una de las más utilizadas para la explotación de datos. También puede ser utilizada para procesos robustos de ingesta de datos o para generación de modelos de *machine learning*, aunque su fuerte pasa por la explotación de datos. Como programa *open source* tiene las ventajas de costo cero y cuenta con constantes actualizaciones con nuevas funcionalidades de gran ayuda. Además, al ser tan utilizado, existen un sinfín de tutoriales y ejemplos de cómo aplicar algunas técnicas de minería, que son de gran utilidad a la hora de comenzar desde cero. Funciona como un *drag and drop*. Uno selecciona la función que quiere realizar y la arroja a la hoja de trabajo. Tiene una visualización muy simple, lo que permite crear flujos de datos que son muy fáciles de leer. Una de las principales desventajas que tiene este *soft*, en comparación con sus competidores con licencia paga, es su capacidad de procesamiento. Si bien soporta grandes volúmenes de datos, podemos notar una gran diferencia entre esta
herramienta y otras como SAS o SPPS Modeler en su capacidad de procesarlos. Para más información o descargarse el *soft*, les recomiendo entrar a su página oficial: https://www.knime.com/

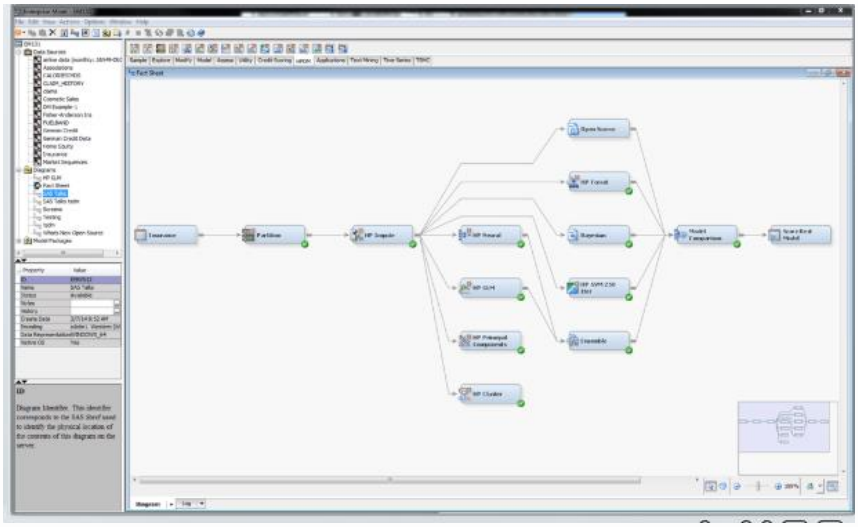
Figura 2: Knime



Fuente: [Imagen sin título sobre Knime]. (s.f.). Recuperada de http://bit.ly/2IPZWQr

- **SAS**: es la principal herramienta de *data mining* a nivel mundial en ambientes corporativos. Es sencilla de utilizar y ofrece, al igual que Knime, una interfaz gráfica que facilita el uso a personas con escasos conocimientos técnicos. Su fuerte es el *data mining*, *text mining* y optimización. Es una de las herramientas predilectas de campañas de *email marketing*, o cualquier tipo de campañas masivas, en donde se busca optimizar los envíos. A diferencia de los *softwares open source*, ofrece soporte técnico, algo que las grandes compañías valoran mucho. Además, su capacidad de procesamiento es muy buena. Este es un atributo muy importante cuando estamos hablando de una gran cantidad de datos, ya que podemos acortar tiempos en el procesamiento y en el análisis. Genera gráficos sencillos de analizar, aunque la interfaz gráfica no es uno de sus fuertes. Con algunos conocimientos, se pueden realizar modelos productivos de manera sencilla.
Para encontrar más información, les dejo el enlace de su página oficial: https://www.sas.com/es_ar/software/enterprise-miner.html

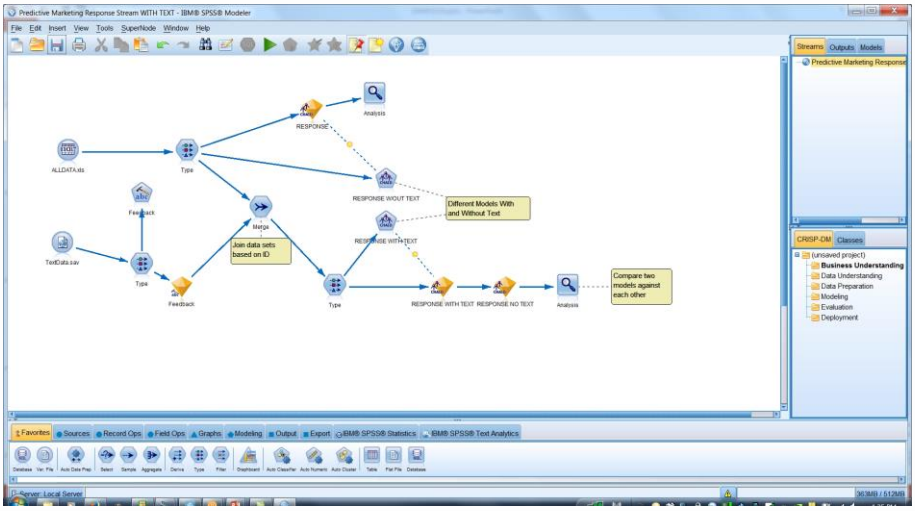
Figura 3: SAS



Fuente: [Imagen sin título sobre SAS]. (s.f.). Recuperada de http://bit.ly/2Qmcgfk

- **IBM SPSS Modeler**: esta es otra de las herramientas de licencia paga más potentes que existen en el mercado. Sin duda, sus mayores atributos son la capacidad de integrar distintas fuentes de información y su capacidad de procesamiento.
  Es un programa muy sencillo de utilizar, con el que los usuarios no expertos técnicamente pueden lograr grandes resultados con conocimientos mínimos de la herramienta.
  Este *software* fue una evolución del programa estadístico SPSS Statistics, uno de los programas más utilizados en el mundo académico, cuya compañía fue luego adquirida por IBM.
  Su principal desventaja, como todo enlatado, es la falta de actualizaciones recurrentes con nuevas funcionalidades y su costosa licencia. Por el costo que tiene esta herramienta son pocas las empresas que pueden pagarlo.
  Para más información, pueden ingresar a su página oficial y descargarse una demo: https://www.ibm.com/ar-es/products/spss-modeler

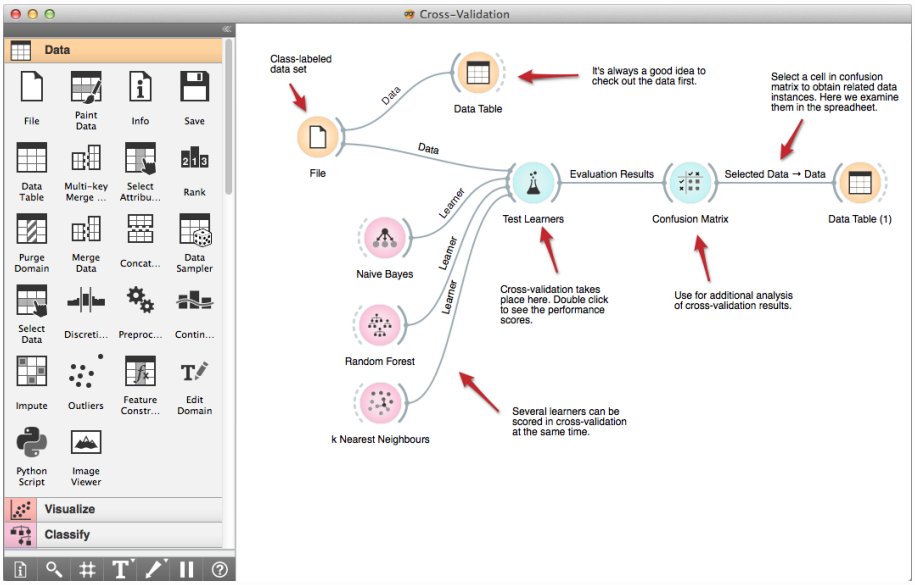
Figura 4: IBM SPSS Modeler



Fuente: [Imagen sin título sobre IBM SPSS Modeler]. (s.f.). Recuperada de http://bit.ly/39VVmfj

- **Orange**: es otra herramienta gratuita. Muy parecida a Knime, pero con muchas integraciones con Python. Así como las otras herramientas, sirve tanto para *data mining* como para *data preparation*, *data visualization* y *machine learning*, aunque su fuerte es el *data mining*.
  Dos de las grandes ventajas de este *soft open source* son la gran cantidad de tutoriales gratuitos que se pueden encontrar y la cantidad de ejemplos de modelos de datos que podemos descargar para aprender.
  Su gran integración con lenguaje Python permite importar librerías de una manera muy sencilla, aunque rara vez es necesario, ya que suelen salir actualizaciones muy frecuentemente con todo lo que un usuario estándar necesita. Este programa se volvió muy popular dentro del ambiente de *data science* una vez que fue incorporado dentro del paquete de programas que viene en Anaconda (un paquete de programas y librerías para *data science*).
  Al igual que Knime, el procesamiento de grandes volúmenes de información no es uno de sus puntos fuertes.
  Para más información, pueden visitar su página oficial y descargarse el programa: https://orange.biolab.si/

Figura 5: Orange



Fuente: [Imagen sin título sobre Orange]. (s.f.). Recuperada de http://bit.ly/2wbNPu7

Estos simplemente son algunos de los programas más utilizados, pero no quiero dejar de nombrar otros como RapidMiner, Teradata, Oracle DataMining o Weka. Estas herramientas también son muy utilizadas y sería muy bueno que puedan profundizar en ellas o conocerlas.

El valor que aportan todos estos programas depende fundamentalmente de dos cosas: la calidad de los datos y las preguntas que queramos resolver.

Respecto al primer punto, la calidad de los datos es un tema *cross* a todas las herramientas que detallaremos y a todo lo referido a Big Data. Una frase muy conocida es *"garbage in, garbage out"*, haciendo referencia a que, si la información que ingresa a un tablero, a un análisis puntual o a un modelo de inteligencia artificial no es confiable o es de mala calidad, los resultados los serán también.

Respecto al segundo punto (las preguntas que queramos resolver) es indispensable trabajar sobre la generación de hipótesis y resolverlas mediante datos. Algunos *insights* son hallados de manera casual, sin buscarlos, pero este es el menor porcentaje de los casos. Lo más habitual es buscar aprobar o refutar una hipótesis que se haya planteado con anterioridad, y en esa exploración pueden surgir, además, nuevas ideas.

Herramientas de visualización

Estas herramientas permiten generar reportes o tableros muy sencillamente. De esta manera se pueden monitorear o reconocer tendencias de una forma visual y simple. El gran valor que estas herramientas agregan a una organización es el de poder contar con información confiable y detallada, de manera simple de leer; también pueden ayudar a monitorear alguna situación crítica o permitir detectar tendencias o despertar preguntas que pueden luego ser resueltas mediante datos.

Gartner define la visualización de datos como

> una forma de representar información gráficamente, resaltando patrones y tendencias en los datos y ayudando al lector a obtener información rápida. También conocida como "exploración visual interactiva", permite la exploración de datos mediante la manipulación de imágenes de gráficos, con el color, brillo, tamaño, forma y movimiento de los objetos visuales que representan aspectos del conjunto de datos que se analiza. Incluye una variedad de opciones de visualización que van más allá de las de los gráficos circulares, de barras y de líneas, incluidos los mapas de calor y de árboles, mapas geográficos, diagramas de dispersión y otras imágenes de propósito especial. Estas herramientas permiten a los usuarios analizar los datos interactuando directamente con una representación visual de estos. ("Data visualization", s.f., https://gtnr.it/2Wi0qqh).

Al igual que con las herramientas de minería de datos, existen herramientas *open source* y otras con licenciamiento pago. La única diferencia, es que, en este caso, casi todos los programas con licenciamiento pago pueden ser utilizados de forma gratuita, pero con algunas limitaciones. Hay que tener en cuenta que a la hora de elegir una de estas herramientas para una empresa, es muy importante analizar la seguridad de la publicación de tableros y la confidencialidad de la información.

Gartner hace un reporte todos los años sobre las principales herramientas informáticas. Pueden leer el reporte de herramientas de visualización **acá**.

Figura 6: Cuadrante mágico de Gartner para plataformas de análisis de inteligencia empresarial

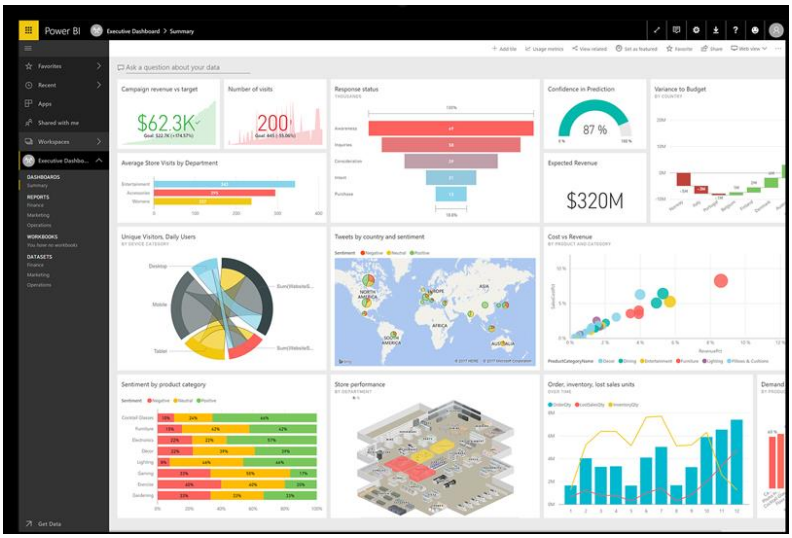
Fuente: [Imagen sin título sobre el cuadrante de Gartner para plataformas de análisis de inteligencia empresarial]. (s.f.). Recuperado de https://gtnr.it/3d3p1VR

Las herramientas de visualización están muy asociadas al *storytelling*, una nueva rama de la visualización que se enfoca en mostrar datos de manera gráfica y con un sentido, como si estuviera contando una historia. Es muy interesante todo lo que se viene hablando y contando al respecto, ya que la forma en la que se muestra la información es clave para que las personas entiendan de manera simple y rápida lo que se está mostrando. El libro *Storytelling with data*, de Cole Nussbaumer Knaflic, detalla muy bien cómo sacar lo mejor de los datos con ejemplos reales.

Como en el caso anterior, estas son algunas de las herramientas más utilizadas en el mundo:

- **Microsoft Power BI**: esta es una herramienta muy simple de aprender. Gracias a su integración con todo el ecosistema Microsoft, logró posicionarse como una de las principales en este rubro. Permite procesar gran cantidad de datos y unir distintas fuentes de información de manera muy sencilla. Al ser una herramienta de tanta difusión, existen un sinnúmero de tutoriales y cursos *online* que ayudan a perfeccionar el arte de hacer visualizaciones. Se puede trabajar tanto de manera *online* como *desktop*. Pueden leer más información en su página oficial: https://powerbi.microsoft.com/es-es/

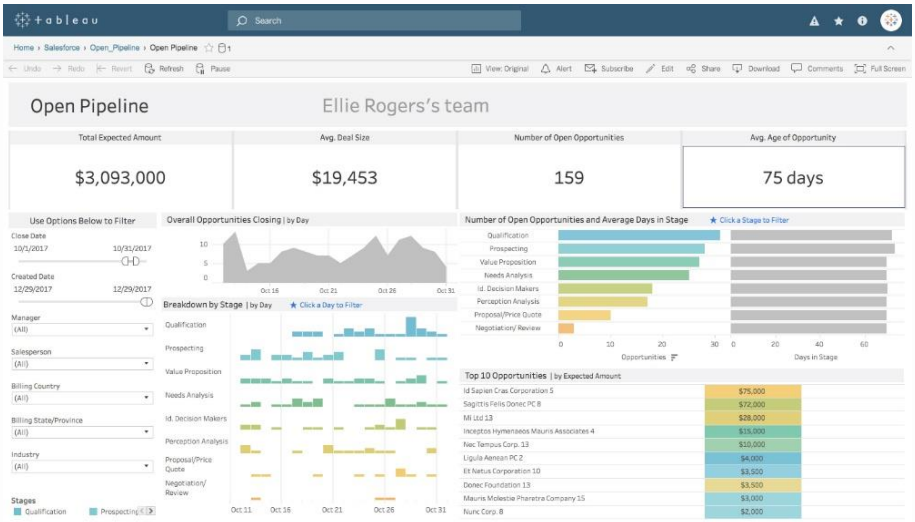
Figura 7: Microsoft Power BI



Fuente: [Imagen sin título sobre Microsoft Power BI]. (s.f.). Recuperado de http://bit.ly/2UfyNf3

- **Tableau**: es una herramienta comercial muy similar a Power BI. Tiene funcionalidades muy parecidas y se puede trabajar tanto en la versión web como en la versión *desktop*. Es una de las herramientas más utilizadas por compañías de todo el mundo, debido a lo fácil que resulta hacer tableros y compartirlos. Al igual que Power BI, tiene gráficos interactivos que permiten hacer un filtro o un análisis más en detalle de alguna variable, lo que lo hace muy intuitivo.
  Para más información, pueden consultar la web oficial: https://www.tableau.com/

Figura 8: Tableau



Fuente: [Imagen sin título sobre Tableau]. (s.f.). Recuperada de https://tabsoft.co/3aZiKbS

Otra de las herramientas que compite y está a la altura de Power BI y Tableau es QLIK. Luego le siguen algunas como MicroStrategy, Salesforce, SAS, entre otras.

Herramientas de *machine learning*

Estas herramientas permiten construir modelos de *machine learning* o inteligencia artificial. Existen de todo tipo, algunas para un uso profesional y para aquellos con experiencia en algún lenguaje de programación, y otros más cercanos a usuarios con poca experiencia en programación. Estas herramientas suelen ser la pata más avanzada de implementación y generación de valor a través de Big Data. Son las que finalmente hacen un uso intensivo y provechoso de la información que tengamos de nuestra organización.

Cabe aclarar que algunas de estas herramientas ya las hemos nombrado, porque sirven para múltiples usos (minería, visualización y *machine learning*). Esto se debe a que cada *software* intenta captar todo el ciclo del procesamiento de datos (procesar, limpiar y transformar los datos).

Herramientas como Knime, Orange y SPSS Modeler permiten, una vez procesados y transformados los datos, entrenar modelos de *machine learning* de manera muy sencilla. Esto acerca muchísimo a usuarios con poca experiencia en la programación al armado de modelos. Más allá de lo fácil que resulta armar un modelo en estos

programas, no deja de ser un requisito muy importante interpretar los resultados y las métricas de cada modelo.

Las herramientas de visualización vienen apuntando cada vez más a este segmento. Ya que permiten que uno pueda hacer, por ejemplo, *forecasting* de alguna variable dado un histograma.

Figura 9: *Forecasting*

Fuente: [Imagen sin título sobre *forecasting*]. (s.f.). Recuperada de http://bit.ly/2x1jFtA

La figura que vemos arriba es un ejemplo de cómo una herramienta de fácil uso, como es Microsoft Power BI, con pasos muy simples, hace una proyección estimada de cómo podría seguir ese histograma (en este caso las ventas mensuales).

Como vemos en el cuadrante mágico de Gartner, existe una gran variedad de herramientas y una gran concentración en su lado derecho. Esta gran variedad que vemos tiene que ver con la madurez de la industria. Es un campo relativamente nuevo (menos de 10 años) y en donde aún no hay un claro ganador.

Figura 10: Cuadrante mágico de Gartner para plataformas de *data science* y *machine learning*

Fuente: [Imagen sin título sobre el cuadrante de Gartner para plataformas de *data science* y machine *learning*]. (s.f.). Recuperado de http://bit.ly/33lenp1

Ya habíamos hablado de algunas como SAS, Knime, IBM Modeler (que está dentro del paquete de Watson Studio) y RapidMiner. Pero aparecen otros tipos de herramientas muy interesantes, como lo son las de Auto Machine Learning (AutoML). Estas son H2O.ai y DataRobot. Son herramientas que permiten probar, entrenar y testear una gran variedad de modelos. Basta con indicarle el set de datos, la variable que se quiere predecir, la métrica para evaluar, y el programa busca cuál es el mejor modelo (o la combinación de varios) que mejor predice. Muestran de una manera gráfica e intuitiva los resultados del entrenamiento y las métricas del test. Luego de esto es muy simple implementar uno de estos modelos en producción. Recomiendo que prueben estas herramientas.

Por otro lado, una de las más recomendadas para Gartner es Anaconda. Esta es la herramienta predilecta de todo *Data Scientist*. Anaconda es un paquete de

programas, como Jupyter Notebooks, Orange, Spyder, etc. Tiene todos los paquetes y librerías necesarios comprimidos en un solo programa. Suele ser la herramienta de inicio de cualquier aficionado al tema. Requiere de mayor nivel de programación y por eso su público es un poco más acotado. Su fuerte es el de ser *open source*.

Google viene creciendo con su paquete de *machine learning*. Recientemente sacó Google AutoML, que acerca mucho más el mundo de *machine learning* a personas que no programan. Aunque aún está lejos de herramientas como DataRobot y H2O.ai, no faltará mucho para que ofrezca soluciones todavía más sencillas que estos dos competidores. Sin dudas que siempre hay que estar atento a las novedades de este gigante de Mountain View.

3.1.2 Mejores prácticas y casos de éxito de Big Data

Se suele decir que el 80 % de los proyectos de Big Data quedan en la nada, ya sea sin implementar o sin ser utilizados. En mi experiencia con la aplicación de Big Data y modelos de inteligencia artificial he aprendido que la clave para que una implementación sea exitosa es contar con un sponsor fuerte del lado del negocio. De nada sirve tener el mejor modelo que prediga el churn de clientes si nadie hace campañas dirigidas a esos clientes, o de nada sirve tener el mejor modelo que prediga el próximo producto para venderle a un cliente (next product to buy) si los asesores comerciales de una sucursal no creen en él y no lo utilizan. Por eso, los proyectos de Big Data no son solo tecnológicos, sino más bien de negocio. Deben tener un objetivo concreto de sumar valor al negocio y contar con un sponsor fuerte de ese lado para poder traccionar un cambio.

Por lo general, se recomienda comenzar con un quick win, un proyecto sencillo, pero que aporte una ganancia asegurada en el corto plazo. Dependiendo del estadio de la organización, esto puede ser muy simple o un poco más complejo. Es indispensable mostrar el diferencial que aportamos con la implementación de alguna de estas técnicas, por eso siempre es recomendable trabajar con grupos de control para mostrar este uplift generado y atribuible al modelo.

Casos de éxito de grandes empresas hay miles. En la actualidad, las técnicas de Big Data están detrás de todo lo que hacemos día a día. Aplicaciones como Google Maps, Spotify, Netflix y tantas otras funcionan constantemente con algoritmos que en tiempo real nos ayudan a ahorrar tiempo, nos sugieren las mejores

recomendaciones o cualquier mejora que facilite nuestras vidas. Por ejemplo, Amazon al tener el historial de transacciones y productos que miraba cada cliente, logró en su tienda online desarrollar un algoritmo de recomendación de productos tan eficiente que se le atribuye el 37 % de su facturación. Esto quiere decir que el 37 % de las ventas del sitio son producto de una recomendación hecha por ese modelo. Y no hace falta ser una tienda online para poder sacarle provecho a esto; las tiendas de retail vienen trabajando en algunas de estas técnicas hace mucho tiempo. Fueron de los primeros casos en utilizar la asociación de productos (el famoso caso de Wal-Mart de principios de los 90 en donde se halló una fuerte correlación entre la venta de pañales y cervezas los viernes por la tarde), el análisis de la canasta y las ofertas personalizadas por cliente.

Estas técnicas son utilizadas en un sinfín de industrias, ya no quedan ramas sin ser tocadas por el Big Data. En la industria de la medicina, por ejemplo, se han logrado grandes avances en la detección temprana de enfermedades mediante el análisis de miles de análisis y diagnósticos. En la industria del deporte, cada vez son más utilizadas, ya sea para transmisiones televisivas como para análisis de rendimiento y entrenamiento de jugadores.

El sector financiero es uno de los que más avances ha hecho en este campo, debido a la gran cantidad de datos que pueden recolectar y la ventaja de poder tomar decisiones a gran velocidad. Desde la posibilidad de ofrecerle a una persona que está navegando por internet la alternativa de ser cliente, acceder a un préstamo o a una tarjeta de crédito hasta realizar y modificar inversiones en tiempo real con base en la lectura de opiniones, tweets y cualquier otro tipo de información que esté disponible en la red.

Pero más allá de la industria, lo que siempre hay que evaluar es el valor que agregamos con este tipo de técnicas. Como comentamos en los dos módulos anteriores, dependiendo el caso, tal vez no haga falta implementar grandes arquitecturas o grandes proyectos, tal vez con small data sea suficiente. Es importante aclararlo, ya que muchas organizaciones con el afán de implementar este tipo de técnicas pierden el foco de por qué hacerlo y termina siendo un fin en sí mismo.

Conceptos clave

- A la hora de extraer valor de los datos existen un sinfín de herramientas, pero que podemos catalogarlas como herramientas de minería de datos, herramientas de visualización y herramientas enfocadas en machine learning. Todas ellas tienen puntos en común.
- La selección de una de ellas depende del estadio de la organización, de la profesionalización del sector que encarará la tarea y del presupuesto del que pueden disponer.
- Es muy importante que el proyecto de Big Data apalanque una necesidad puntual; de nada vale realizar un proyecto de estas características que no agregue valor al negocio.

Referencias

Data mining. (s.f.). *Gartner Glossary*. [Traducción propia]. Recuperado de https://www.gartner.com/en/information-technology/glossary/data-mining

Data visualization. (s.f.). *Gartner Glossary*. [Traducción propia]. Recuperado de https://www.gartner.com/en/marketing/glossary/data-visualization

Imagen sin título sobre el cuadrante de Gartner para plataformas de análisis de inteligencia empresarial. (s.f.). Recuperado de https://www.gartner.com/doc/reprints?id=1-1YAE9AY1&ct=200206&st=sb&signin=7951cf25f058f8ea2703be59e31e6396

Imagen sin título sobre el cuadrante de Gartner para plataformas de *data science* y *machine learning*. (s.f.). Recuperado de https://www.tibco.com/es/node/48081

Imagen sin título sobre *data mining*. (s.f.). Recuperada de https://towardsdatascience.com/data-mining-tools-f701645e0f4c

Imagen sin título sobre *forecasting*. (s.f.). Recuperada de https://powerbi.microsoft.com/fr-fr/blog/introducing-new-forecasting-capabilities-in-power-view-for-office-365/

Imagen sin título sobre IBM SPSS Modeler. (s.f.). Recuperada de https://developpaper.com/summary-of-19-best-used-free-data-mining-tools/

Imagen sin título sobre Knime. (s.f.). Recuperada de https://www.capterra.es/software/158739/knime-analytics-platform

Imagen sin título sobre Microsoft Power BI. (s.f.). Recuperada de https://atx.mx/producto/power-bi-pro/

Imagen sin título sobre Orange. (s.f.). Recuperada de https://www.gameload.xyz/

Imagen sin título sobre SAS. (s.f.). Recuperada de https://www.sas.com/es_ar/software/enterprise-miner.html

Imagen sin título sobre Tableau. (s.f.). Recuperada de https://www.tableau.com/es-es/products/dashboard-starters