*Modelos Analíticos para Ciencia de Datos I*

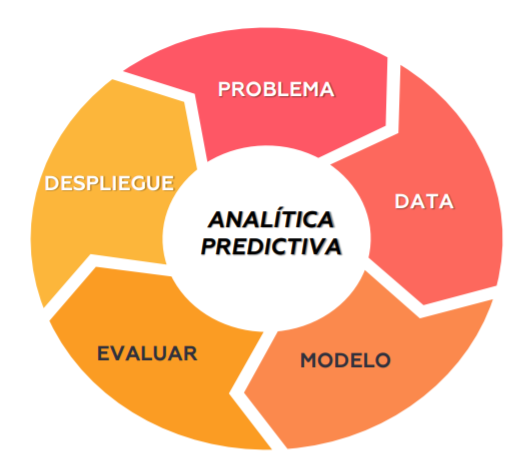
**¿Qué es un Modelo Analítico?**

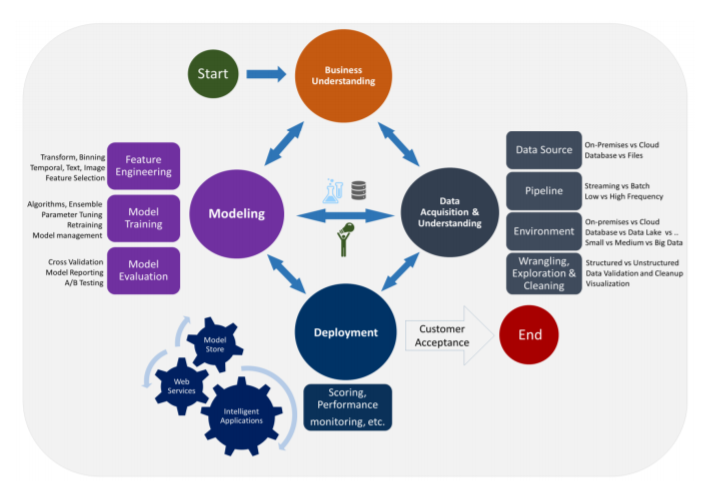
Existen múltiples definiciones al respecto. Sin embargo, para el desarrollo de este curso diremos entonces que: “Un Modelo Analítico de Datos, es un proceso que permite combinar datos heterogéneos de muchas fuentes diferentes. Por lo tanto, facilita que distintos tipos de datos sean fusionados para posteriormente poder ser analizados de manera conjunta”.

**Etapas en la creación de Productos Analíticos con Ciencia de Datos**

****

Este proceso está basado en una metodología estructurada de pasos:





Un aspecto que resulta importante de destacar, es que para crear un modelo analítico, utilizamos generalmente técnicas vinculadas con Data Science y con el Machine Learning, pero entonces, **¿Qué es el Machine Learning?** y **¿Cuáles son sus particularidades y características principales?**

**Machine Learning**

El Machine Learning o ML, es un método de análisis de datos que automatiza la construcción de *“Modelos Analíticos”*.

Es una rama de la Inteligencia Artificial, basada en la idea de que los sistemas pueden aprender de datos, identificar patrones y tomar decisiones con una mínima intervención humana.

Aprender en este contexto quiere decir, identificar patrones complejos en millones de datos.



**Evolución del machine learning**

El ML surge a mediados de los años 50. Se empezó a utilizar, en problemas de predicción complejos donde los modelos estadísticos clásicos no eran eficientes como por ejemplo, la predicción de series temporales no lineales, la predicción de los mercados financieros, el reconocimiento de texto escrito, etc.

La característica principal de este tipo de algoritmos, es que son capaces de reajustarse automáticamente, para mejorar su rendimiento en función del número de aciertos y de fallos producidos en un proceso de entrenamiento previo a su aplicación y durante la ejecución en tiempo real del mismo.

Éstos son algunos ejemplos ampliamente conocidos de aplicaciones de machine learning con los que quizás estemos todos un poco familiarizados:

* ¿El automóvil de conducción autónoma de Google? La esencia del Machine Learning.
* ¿Ofertas de recomendación en línea como las de Amazon y Netflix? Aplicaciones de ML para la vida diaria.
* ¿Saber lo que los clientes dicen acerca de nosotros en Twitter? Machine learning combinado con creación de reglas lingüísticas.
* ¿Detección de fraudes? Uno de los usos más importantes en la actualidad del ML.

**¿Por qué es importante el Machine Learning?**

El resurgimiento del interés en el aprendizaje basado en máquina, se debe a los volúmenes y variedades crecientes de datos disponibles, el procesamiento computacional más económico y poderoso y el almacenaje de datos asequible y mucho menos costoso. Todas estas cosas, hacen posible producir modelos de manera rápida y automática que puedan analizar datos grandes y complejos. Con la construcción de modelos precisos, una organización tiene una mejor oportunidad de identificar oportunidades rentables o de evitar riesgos desconocidos.

**Ventajas del Machine Learning:**

* Mayor conocimiento de las necesidades, gustos y hábitos de compra de los clientes.
* Innovación en productos y soluciones tecnológicas.
* Optimización de la producción y de la productividad.
* Capacidad de realizar acciones preventivas y correctivas.
* Predicción de tendencias y necesidades.



**¿Quién utiliza ML?**

**Servicios Financieros:**

Los bancos y otras empresas de la industria financiera utilizan la tecnología del aprendizaje basado en máquina para dos fines principales: identificar insights importantes en los datos y prevenir el fraude.

**Salud:**

El machine learning es una tendencia en rápido crecimiento en la industria de atención a la salud, gracias a la aparición de dispositivos y sensores que pueden usar datos para evaluar la salud de un paciente en tiempo real.

**Petróleo y Gas:**

* Cómo encontrar nuevas fuentes de energía.
* Análisis de minerales del suelo.
* Predicción de fallos de sensores de refinerías.
* Optimización de la distribución de petróleo para hacerla más eficiente y económica, etc.

**Gobierno:**

Dependencias de gobierno como seguridad pública y los servicios públicos tienen una necesidad particular del Machine Learning porque tienen múltiples fuentes de datos de las que se pueden extraer insights. Por ejemplo, el análisis de datos en las Smart Cities o también conocidas como Ciudades Inteligentes.

**Marketing y Ventas:**

Los sitios Web que recomiendan artículos que podrían llegar a gustarnos con base en nuestras compras anteriores, utilizan el machine learning para analizar nuestro historial de compras y de tal forma promocionar otros artículos que podrían llegar a resultarnos de interés. Esta capacidad de capturar datos, analizarlos y usarlos para personalizar la experiencia del cliente, es una de las grandes tendencias dentro del mundo de Marketing y Ventas.

**Transporte:**

Analizar datos para identificar patrones y tendencias, es clave para la industria del transporte, que se sustenta en hacer las rutas más eficientes y anticipar problemas potenciales para incrementar la rentabilidad. Los aspectos de análisis y modelado de datos del machine learning, son herramientas importantes para las compañías de mensajería, transporte público y otras organizaciones de transporte.

**¿Qué se requiere para crear buenos sistemas de machine learning?**

* Recursos de preparación de datos.
* Algoritmos – básicos y avanzados.
* Automatización y procesos iterativos.
* Escalabilidad del Modelo en producción.

**Fases del Proyecto de Machine Learning:**

**Etapa 1 - Definir el objetivo**:

Es vital entender el problema a resolver, y cuáles son nuestros objetivos dados las características de la empresa, así como de la data que tendremos a disposición.

Las siguientes preguntas son típicas en esta etapa:

* ¿Qué exactamente deseamos hacer?
* ¿Cómo exactamente podremos hacerlo?
* ¿Es posible lo que deseo dada la data que tengo?

**Etapa 2 - Recolección de la data**:

* **Data First Party:** Data propia de la empresa (ERP,CRM,BD, etc).
* **Data Second Party**: Suele ser data que comparte una organización con sus aliados estratégicos.
* **Data Third Party**: Datos de tercero que podemos obtener ya sea de forma gratuita o incurriendo en algún tipo de costo asociado.

**Etapa 3 - Preparar la data**:

Una vez que disponemos de la data, continuamos con el preprocesamiento de la misma, esto normalmente lo conocemos como la limpieza de los datos o el formateo del dato.

El objetivo de esta etapa es manipular y convertir la data en formas que produzcan mejores resultados. Como ejemplos típicos de preparación de datos tenemos: Eliminar o inferir datos perdidos, categorizar los valores de las variables, normalizar los valores numéricos o escalarlos para que puedan ser comparables.

**Etapa 4 - Elección del algoritmo**:

Una vez que ya hemos preprocesado la data, nos corresponde elegir el algoritmo más adecuado en relación al problema que deseamos resolver.

En este punto tenemos que decidir por el **Tipo de Aprendizaje** que vamos a implementar. Resulta importante mencionar, que en esta sesión simplemente comentaremos sobre esta temática. En las próximas clases de: *Modelos Analíticos para Ciencia de Datos II y III* se abordará los Tipos de Aprendizaje en ML con mucho mayor detalle y profundidad.

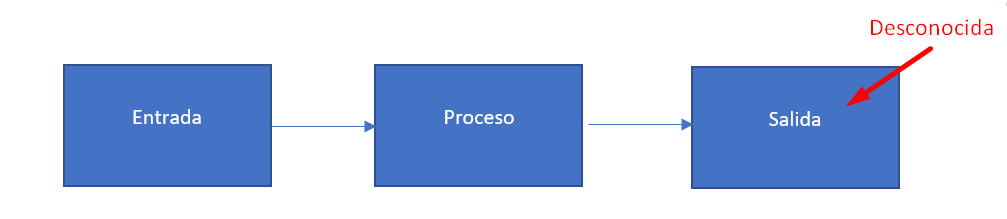


**Aprendizaje Supervisado:** Los algoritmos de aprendizaje supervisado son entrenados utilizando ejemplos etiquetados, como una entrada donde se conoce el resultado deseado. El algoritmo de aprendizaje recibe un conjunto de entradas junto con los resultados correctos correspondientes, y el algoritmo aprende comparando su resultado real con resultados correctos para encontrar errores. Luego modifica el modelo en consecuencia es decir, la salida de este algoritmo es conocida.

Diagrama

Descripción generada automáticamente con confianza mediaEstructura básica de un Algoritmo

**Aprendizaje No Supervisado:** Se utiliza contra datos que no tienen etiquetas históricas. No se da la "respuesta correcta" al sistema. El algoritmo debe descubrir lo que se muestra. El objetivo es explorar los datos y encontrar alguna estructura en su interior. Por ejemplo, puede identificar segmentos de clientes con atributos similares que después puedan ser tratados de manera semejante en campañas de marketing o bien puede encontrar los atributos principales que separan los segmentos de clientes.



Estructura básica de un Algoritmo

**Aprendizaje por Refuerzo:** Se utiliza a menudo para robótica, juegos y navegación. Con el aprendizaje con refuerzo, el algoritmo descubre a través de ensayo y error qué acciones producen las mayores recompensas. Este tipo de aprendizaje tiene tres componentes principales: el agente (el que aprende o toma decisiones), el entorno (todo con lo que interactúa el agente) y acciones (lo que el agente puede hacer).

**Etapa 5 - Entrenar el modelo**:

Este paso tiene una relación directa con conceptos que abordaremos más adelante en el curso (Training y Test). Sin embargo, el proceso de entrenamiento de un modelo de ML, consiste en proporcionarle al modelo datos de entrenamiento de los cuales pueda aprender.

**Etapa 6 - Validación del modelo**:

Dado que ya tenemos el modelo entrenado, lo siguiente es validarlo. Esto lo realizaremos con la data de validación y procederemos a correr el algoritmo y a evaluar los resultados obtenidos.

En el caso de que los resultados no sean satisfactorios, deberemos volver a la etapa 5 hasta que nuestro modelo se ajuste bien a las dos particiones (data de entrenamiento y data de validación).

**Etapa 7: Deployment del modelo**:

Consiste en la implementación en producción de nuestro modelo. Generalmente para este paso, solemos ayudarnos de la nube a través de los 3 vendors más conocidos que existen actualmente: AWS, Azure y GCP.



**Conceptos complementarios de ML:**

* En el aprendizaje basado en máquina, un destino se conoce como etiqueta o target.
* En estadística, un destino se conoce como variable dependiente.
* Una variable en estadística se conoce como característica en el Machine Learning o Feature.
* Una transformación en estadística se conoce como creación de característica en el ML.