¿CÓMO SE DESARROLLA UN PROYECTO DE DATA SCIENCE?

1.1 Implementación de un modelo mínimo viable.

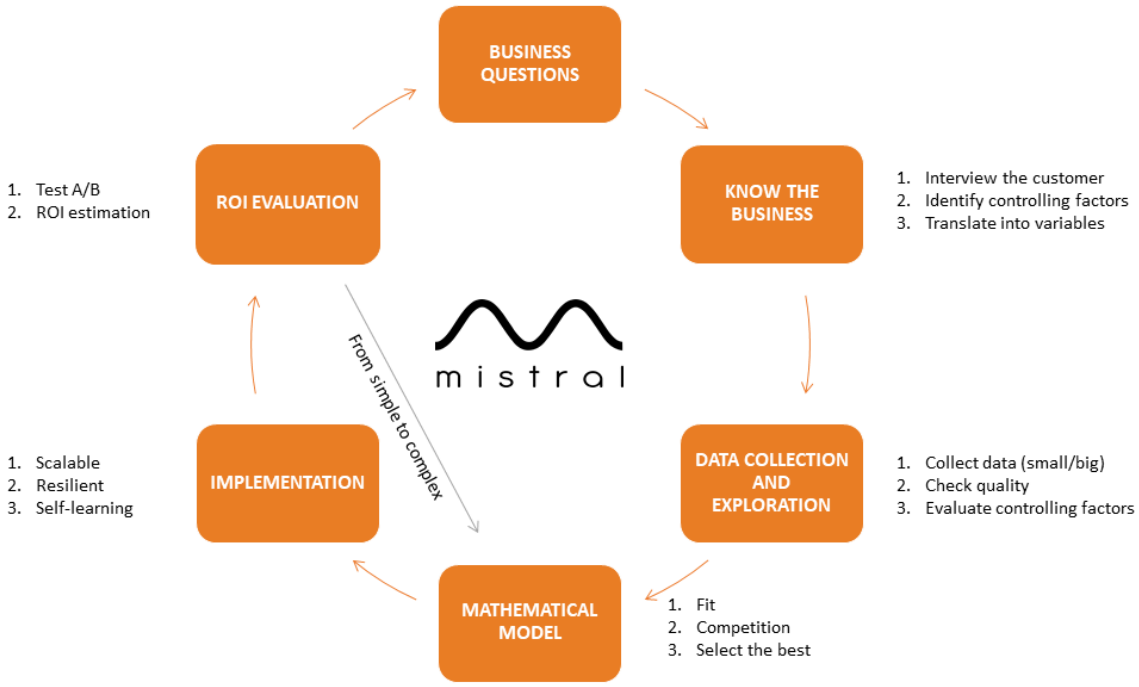

Fig 1. Ciclo de vida de un proyecto de Data Science

Business Question

El ciclo de vida de un proyecto de Data Science (véase figura 1) empieza por la **Business Question** (BQ) mediante la cual **el cliente se plantea una necesidad**, ya sea específica de su propia empresa o más general, común a las empresas de un mismo sector.

Por ejemplo, en sectores industriales predominan preguntas del tipo cuántas ventas tendré el año que viene, al objeto de poder optimizar la compra de materiales (reducir costes) y la gestión del stock (evitar roturas de stock), lo que se resuelve mediante modelos de predicción de la demanda. Otro tipo de pregunta relacionada con la cadena de producción es cuándo fallarán las máquinas. Su respuesta permitiría predecir el mantenimiento de las mismas y evitar paradas no deseadas (algoritmos de mantenimiento predictivo). En este sector también se preguntan, cómo puedo reducir la merma generada para minimizar costes, lo que requiere un estudio de los parámetros recogidos por los sensores distribuidos y relacionarlos con la calidad del producto en sus diferentes etapas de la cadena de producción.

Sin embargo, en las empresas relacionadas con el e-commerce o el retail, además de querer predecir la demanda, este sector se pregunta qué tipos de clientes tiene, lo que se responde mediante técnicas de

segmentación de la clientela con las que se caracterizan diferentes perfiles. Para aumentar las ventas, en este sector es recurrente preguntarse qué productos se pueden recomendar de forma acertada para cada cliente, fomentando el Up-Selling y Cross-Selling. En este último caso, es fundamental poder relacionar los datos históricos de compra disponibles y mejor aún, con las valoraciones hechas por los propios clientes.

Know the business

En el siguiente paso del ciclo de Data Science, **Know the business**, **Mistral recopila** cuanta **información** se pueda disponer que permita responder a estas BQ y, la resume en esquemas o diagramas de flujo para poder identificar sus factores de control y cuellos de botella potenciales. Por ejemplo, en el caso del mantenimiento predictivo se representarían las máquinas de producción junto su función y la información que emiten sus sensores (véase figura 2).

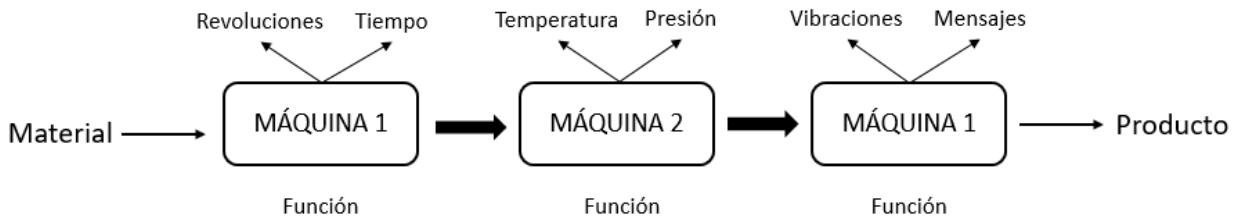

Fig 2. Esquema de la información disponible en una cadena de producción

Data Collection and Exploration

Conocidos todos los procesos e identificados los factores de control, traducimos esta información en **variables y datos**, momento en el que se produce la recolección de esta información (small/big Data) y su posterior **análisis exploratorio** para **evaluar la calidad** de estos y cuantificar el efecto de esos factores de control (**"Data Collection and Exploration"**).

Por un lado, evaluamos el tamaño muestral para ver si hay suficientes datos históricos con los que desarrollar los modelos matemáticos, así como también la coherencia de los datos (descriptivos estadísticos básicos, valores extremos, "outliers", tipo de distribución y visualización). Por otro lado, se cuantificaría la importancia de los factores de control (por ejemplo, la varianza explicada) y su efecto en el ajuste de los modelos matemáticos.

The mathematical model

Durante la cuarta fase y partiendo siempre **de lo más simple a lo más complejo**, Mistral aplica diferentes metodologías de **machine learning** para desarrollar varios modelos matemáticos que compiten entre sí

para ver cuál es el mejor enfoque científico que explique y prediga los datos. Este **procedimiento** es **cíclico y reiterativo**, hasta que se obtienen los parámetros más adecuados.

Empezamos con modelos más sencillos para después incrementar su complejidad, ofreciendo de esta manera una respuesta más rápida a las necesidades del cliente y que éste pueda obtener beneficios desde los momentos más tempranos del desarrollo del proyecto (véase figura 1).

Implementation

Cuando se dispone de un modelo mínimo viable, se inicia la fase de **Implementación**, durante la cual se introduce éste en el sistema de control correspondiente de la empresa. El modelo desarrollado tiene que ser dinámico, **aprendiendo de forma continua**, con una implementación escalable y tolerable a los cambios del modelo: es decir que sea capaz de auto entrenarse a medida que se tienen más datos.

ROI Evaluation

La última fase del ciclo se alcanza con la **medición del ROI** (Return Of Investment) para evaluar el retorno económico de la inversión realizada, puesto que el mejor modelo matemático no tiene por qué ser siempre el que mayor beneficios aporte.

1.2 Monitorización del modelo.

Implantado el modelo mínimo viable, el ciclo continúa con el desarrollo de modelos matemáticos más complejos que reemplazarán al previamente instalado si se considera que pudiesen reducir los costes o aumentar los beneficios. Además, a medida que se obtengan más datos o nuevas variables, se irán incluyendo, volviéndose a evaluar la bondad de ajuste y ROI.

En esta fase Mistral monitoriza la efectividad real de los algoritmos desarrollados mediante la implementación de cuadros de mando (integrados o no en el sistema de control), con los que visualizar de forma interactiva y en tiempo real la evolución de los parámetros más significativos (*e.g.* temperatura), para que el cliente pueda adoptar en cualquier momento las decisiones más oportunas con la máxima información posible, de calidad y reciente.