DSC::Energy Analytics

Proyecto de Machine Learning en Mantenimiento

Cómo enfocarlo para el éxito

Antonio Marín Écija

Sevilla, 11 de octubre de 2019

Email: dsc@dscenergy.com Web: www.dscenergy.ai

DSC::Energy Analytics

DSC Energy Analytics es una empresa de consultoría de **analítica avanzada** y modelos predictivos con técnicas de "**Machine Learning**", que da servicios principalmente en el sector de la energía renovable.

Motivos fundacionales:

DSC Energy se crea partiendo de la experiencia acumulada del equipo, en los sectores que ha trabajado durante 25 años, con empresas tanto pequeñas como grandes multinacionales, en más de 20 países, y aprovecharlo para usar las **nuevas tecnologías** que se presentan con la transformación digital que están teniendo las empresas, especialmente en el mundo de la **inteligencia artificial y el big data.**

Recursos de computación:

Para abordar proyectos con alto volumen de datos es necesaria capacidad de computación elevada.
DSC Energy cuenta con estaciones de trabajo con las siguientes capacidades:

- RAM: 4 x 128 Gb = 512 Gb
- CPUs: 34 núcleos en total (68 hilos)
- GPUs (Deep Learning): 6 GPUs en total con 113 Tflops y 53 Gb de memoria

Email: dsc@dscenergy.com Web: www.dscenergy.ai

DSC::Energy
Analytics

ÍNDICE

DSC::Energy
Analytics

1.1. Definiciones

Data Science = Ciencia de Datos

es un término unifica **estadística, análisis de datos, aprendizaje automático y sus métodos relacionados** con el fin de "comprender y analizar fenómenos reales" con datos. Emplea técnicas y teorías extraídas de muchos campos dentro del contexto de las matemáticas, las estadísticas, la informática y las ciencias de la información.

Machine Learning = Aprendizaje Automático

es un subcampo de las **ciencias de la computación** y una rama de la **inteligencia artificial**, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan, usando **algoritmos y modelos estadísticos.** Los algoritmos de aprendizaje automático crean un **modelo matemático basado en datos**, para hacer predicciones o decisiones sin ser programado explícitamente para realizar la tarea.

DSC::Energy
Analytics

1.2. Tipos

Artificial Intelligence

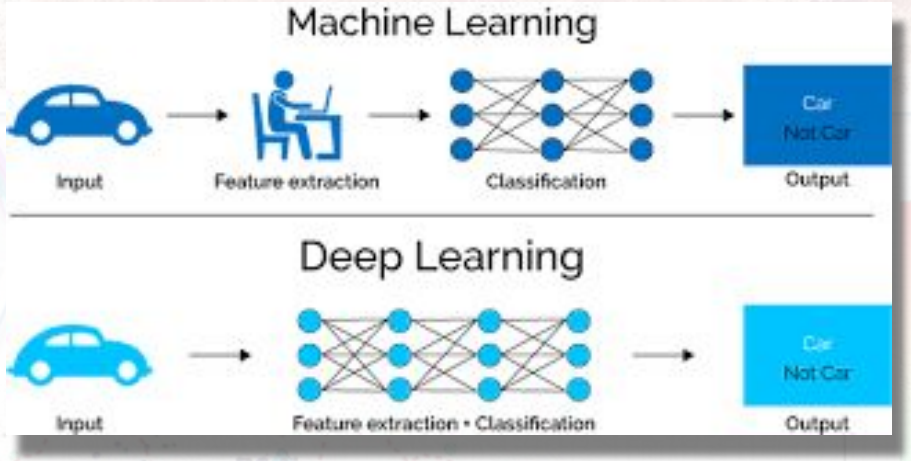
Any technique that enables computers to mimic human intelligence. It includes *machine learning*

Machine Learning

A subset of AI that includes techniques that enable machines to improve at tasks with experience. It includes *deep learning*

Deep Learning

A subset of machine learning based on neural networks that permit a machine to train itself to perform a task.



DSC::Energy
Analytics

1.3. ¿Por qué ahora?

Democratización de la IA - software abierto

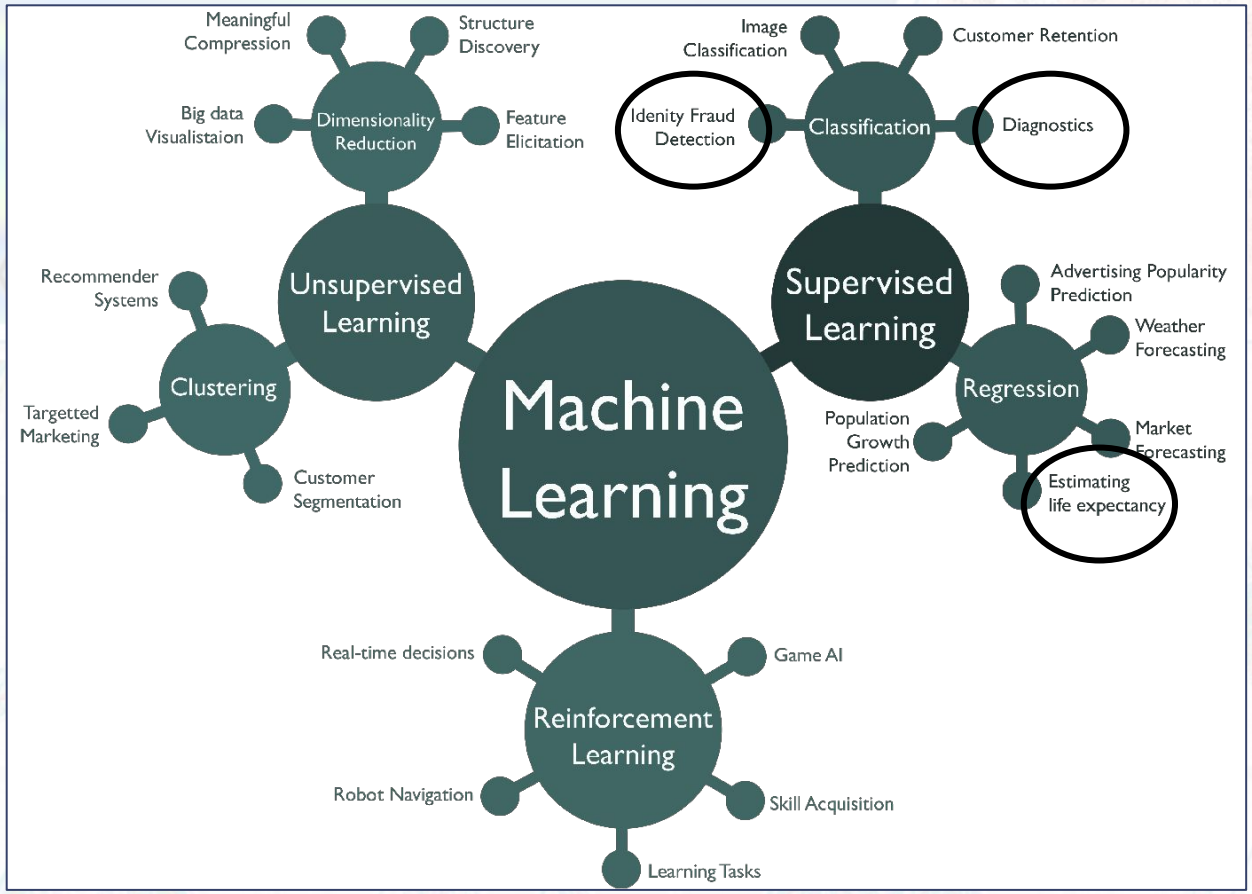
- Lenguajes especializados abiertos: Python, R
- Algoritmos avanzados abiertos: LightGBM (Microsoft), XGBoost, Catboost (Yandex)
- Redes neuronales: Tensorflow (Google), Torch (usado por Facebook AI Research Group, IBM, Yandex)
- Infraestructuras big data: Hadoop, Spark
- Cloud y sus herramientas: Microsoft Azure, Amazon Web Services, Google Cloud

Revolución computación

- CPUs cada vez más avanzados (incluso opción de alquilar en la nube). En los últimos 25 años la computación ha aumentado su velocidad en 1 millón de veces ! Y sigue creciendo
- Evolución espectacular de las GPUs para las redes neuronales, impulsadas por la industria del videojuego y el minado de bitcoin. Hoy por algo más de 1000 euros se puede tener una GPU con 30 TFlops, algo impensable hace pocos años (1997: 30000 USD/GFlops -> 2019: 0.03 USD/GFlops)

DSC::Energy Analytics

1.4. Lo más avanzado en ML



March, 2014
XGBoost initially started as research project by Tianqi Chen but it actually became famous in 2016

Jan, 2017
Microsoft released first stable version of LightGBM

April, 2017
Yandex, one of Russia's leading tech companies open sources CatBoost

Microsoft LightGBM

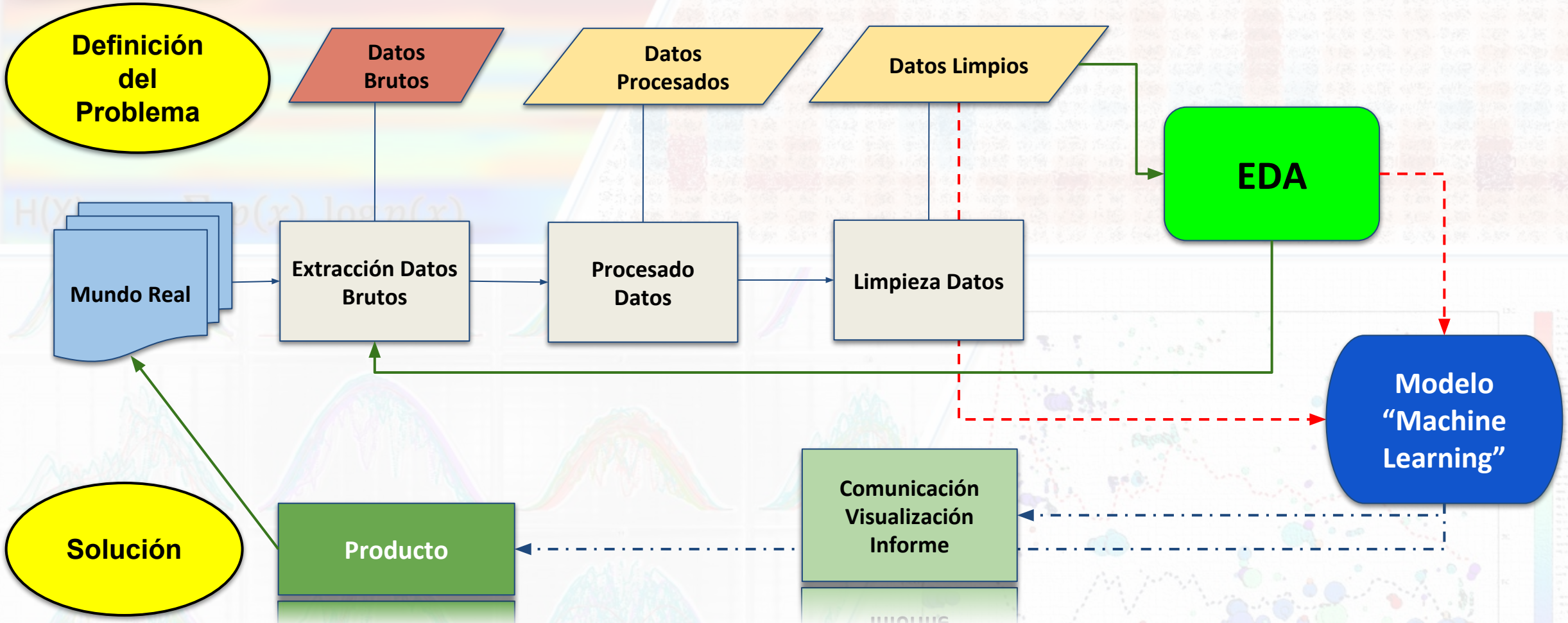
dmlc XGBoost

Yandex CatBoost

DSC::Energy Analytics

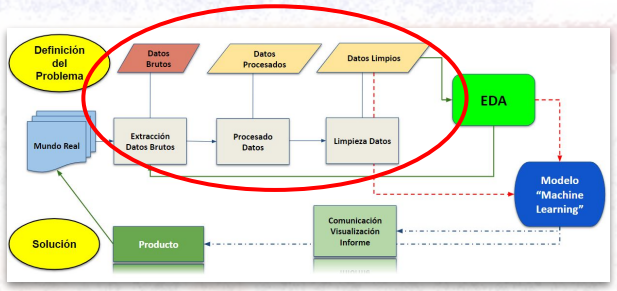
2. Proceso Proyecto Machine Learning

2.1. Esquema



DSC::Energy
Analytics

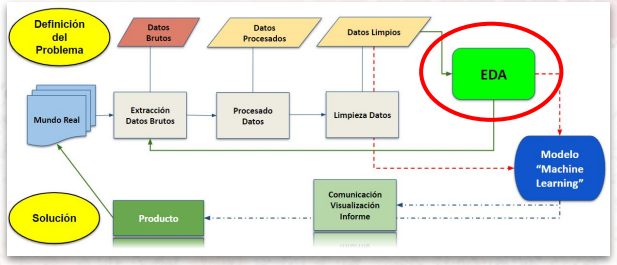
2.2. Los Datos



- El algoritmo aprenderá a partir de los datos, por lo que son muy importantes. De hecho, se estima que en un proyecto de Machine Learning, el **80% del tiempo es para la preparación de los datos**.
- Puede haber **datos estructurados y no estructurados**, pero al final los algoritmos sólo entienden de datos estructurados, con un orden generalmente tabular
- **Etapas** de procesado de datos:
  - Adquisición y almacenamiento datos brutos
  - Limpieza de datos erróneos
  - Imputación de datos faltantes ("missing")
  - Incorporación de nueva información procesada (ingeniería de variables o "Feature Engineering")

DSC::Energy
Analytics

2.3. El EDA[1] (Análisis Exploratorio de Datos - "Exploratory Data Analysis")

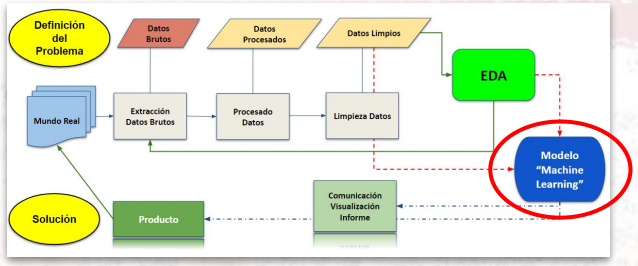


- Es una etapa inicial imprescindible, y que en muchas ocasiones aporta mucho valor, para conocer lo que está pasando en el problema que se está afrontando
- Se trata de analizar los datos siguiendo unos pasos mínimos:
  - Estadísticas principales de las variables independientes **continuas**: min, max, media, mediana, desviación estándar, valores nulos, valores faltantes (missing), valores fuera de rango
  - Estadísticas de las variables independientes **categóricas**: tipos, valores únicos, cardinalidad
  - Estadística de la **variable objetivo**: min, max, media, desviación estándar
  - Visualización de **relaciones entre variables** independientes y variable objetivo
  - **Visualización** de variables: boxplots (caja y bigotes), distribuciones, scatters (nube de puntos), heatmaps, ….

[1] El término EDA fue utilizado por primera vez por John W. Tukey, un reconocido estadístico estadounidense

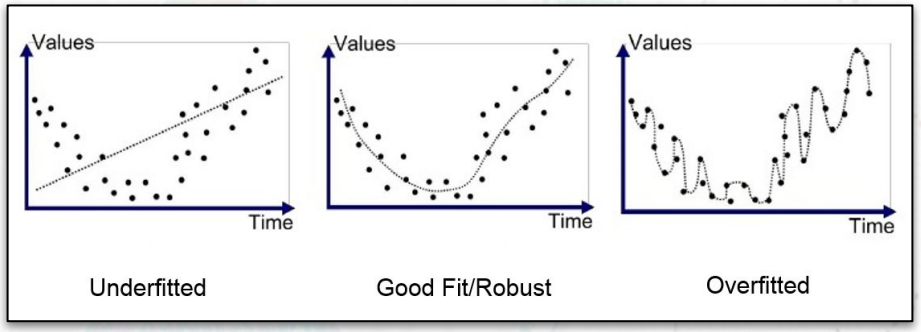
DSC::Energy Analytics

2.4. Entrenamiento del modelo



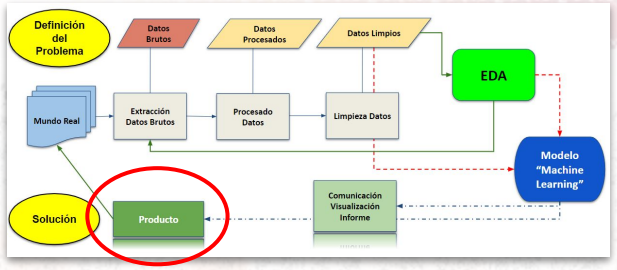
- Selección del **método de validación**: validación cruzada ("cross validation"). Partición de datos en entrenamiento, validación y test.
- Selección de la **métrica**: RMSE, MAE, AUC, ...
- Selección del **algoritmo**
- Selección de los **parámetros** del algoritmo ("hyperparameter tuning")
- Evaluación del modelo, según la métrica y con el sistema de validación

- Objetivo principal: el modelo debe **generalizar** y mantener su rendimiento para datos no vistos, evitar el **sobreajuste** ("**overfitting**")



DSC::Energy
Analytics

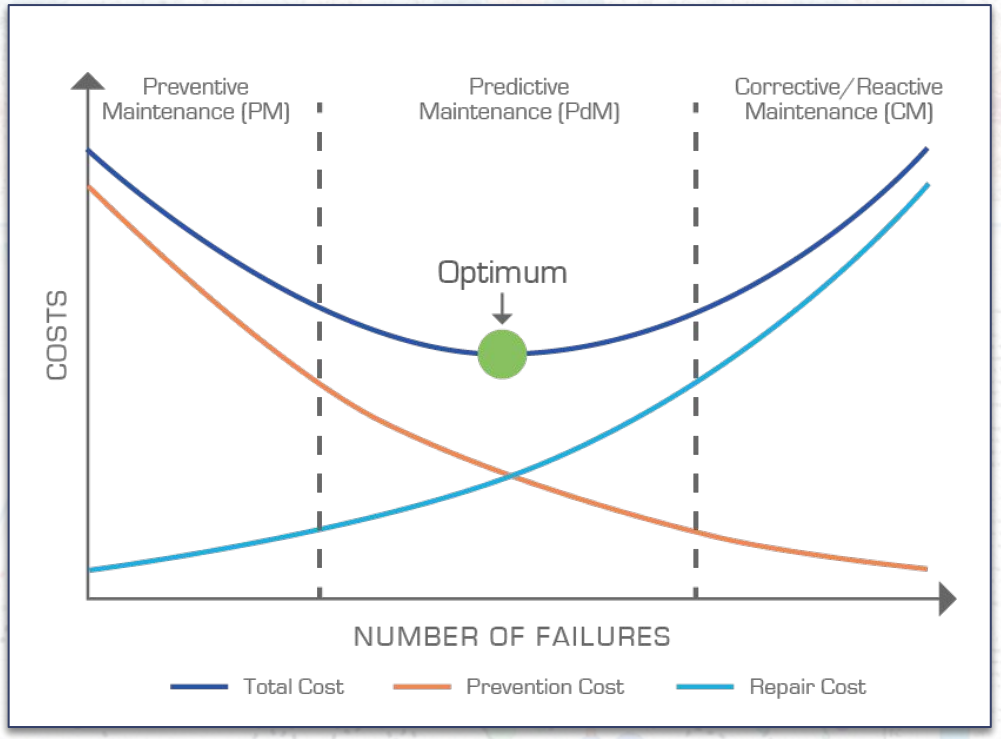
2.5. Implantación (“deployment”)



- Puede ser en equipos locales o en la nube (“cloud”)
- En equipos locales: requiere inversión, dependiendo del volumen de datos, la necesidad de memoria RAM y la rapidez de ejecución de los algoritmos
- En la nube: **Azure** (Microsoft), Amazon Web Services **AWS** (Amazon), **Google Cloud** (Google)
- ¿Quiénes serán los usuarios del modelo? ¿Cómo lo consumirán?
- No olvidar la **gestión y mantenimiento del modelo**. Revisión de rendimiento periódico, por si hay que volver a entrenar

3.1. Clasificación mantenimientos

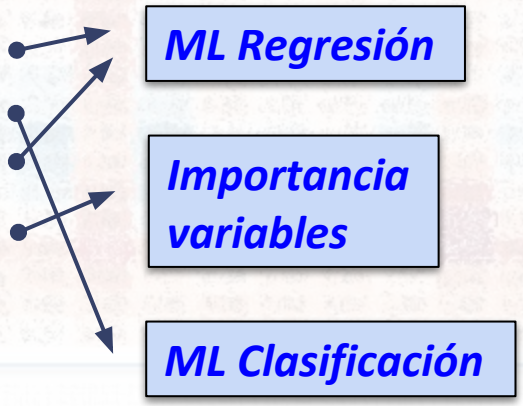
- Mantenimiento **correctivo**: las piezas se reemplazan una vez presentan errores
  - Se garantiza que las piezas se usan completamente
  - Pero para el negocio implica costos en tiempo de inactividad
- Mantenimiento **preventivo**:
  - Se estima tiempo de vida útil y se reemplaza antes de que ocurra algún error
  - Alto costo del tiempo de inactividad programado e infrautilización de los componentes
- Mantenimiento **predictivo**:
  - Se optimiza equilibrio entre mantenimiento correctivo y preventivo, mediante reemplazo justo a tiempo de los distintos componentes



DSC::Energy Analytics

3.2. Problemas a resolver

- Detectar **anomalías** en el **rendimiento** o la **funcionalidad** del sistema o los equipos
- **Predecir** si un componente puede presentar **un error en un futuro próximo**
- Calcular la **vida útil** restante de un recurso
- Identificar las **principales causas** del error de un recurso
- Identificar **qué acciones de mantenimiento** se deben llevar a cabo en un recurso y en qué plazo



Condiciones a tener en cuenta para seleccionar los problemas a abordar:

- La **naturaleza** del problema debe ser **predictiva**
- El problema debe tener un **registro del historial** de operaciones, que contenga tanto resultados buenos como malos. Deben estar registrados los informes de errores, los registros de mantenimiento, registro de reemplazos y reparaciones realizadas
- Los datos deben tener la **suficiente calidad**
- Es imprescindible la **colaboración de expertos del dominio**, que tengan una idea clara del problema

DSC::Energy Analytics

3.3. Casos de uso[1]

Sector energético:

- **Problema**: *errores en turbinas eólicas*, debido a fallo en los componentes principales: multiplicadora, generador, transformador

- **Ventajas**: la modelización para detección de anomalías puede informar de componentes con probabilidad de fallo, y permita programar actividades de mantenimiento

- **Problema**: *errores en disyuntores* de redes de distribución

- **Ventajas**: la predicción de errores ayuda a reducir los costes de reparación, y ayudan a mejorar la calidad de la red energética mediante la reducción de interrupciones del servicio y errores inesperados

[1] Casos de uso extraídos de la "Guía de Azure AI para soluciones de mantenimiento predictivo"

DSC::Energy Analytics

3.3. Casos de uso

Sector aeronáutico:

- **Problema**: *retraso y cancelaciones de vuelos* por problemas mecánicos

- **Ventajas**: ML puede predecir la probabilidad de que un avión se retrase o se cancele

Sector financiero:

- **Problema**: *errores en cajeros automáticos* que provocan interrupción de transacciones

- **Ventajas**: en lugar de que la máquina presente errores en medio de una transacción, la alternativa deseable es programar la máquina para denegar el servicio en función de la predicción

DSC::Energy Analytics

3.3. Casos de uso

Otros sectores:

• **Problema**: *errores en puertas de ascensores* La principal preocupación de los clientes son la seguridad, la confiabilidad y el tiempo de actividad	• **Ventajas**: el mantenimiento predictivo puede predecir las posibles causas de los errores en las puertas, brindando una ventaja competitiva a la hora de ofrecer los servicios en períodos prolongados de vida de los ascensores
• **Problema**: *errores en ruedas de trenes*, pues son causa de la mitad de todos los descarrilamientos	• **Ventajas**: predecir los errores en las ruedas ayudará a reemplazarlas justa a tiempo
• **Problema**: *errores en puertas de metro*	• **Ventajas**: predecir los errores en las puertas ayudará a optimizar las programaciones de mantenimiento

DSC::Energy
Analytics

3.4. Los datos (requisitos)

El éxito de cualquier aprendizaje depende de la calidad de lo que se enseña (**datos**) y la capacidad del aprendiz (**algoritmo**). Los modelos predictivos aprenden patrones de los datos y predicen resultados futuros con cierta probabilidad. **Los datos deben cumplir unos requisitos.**

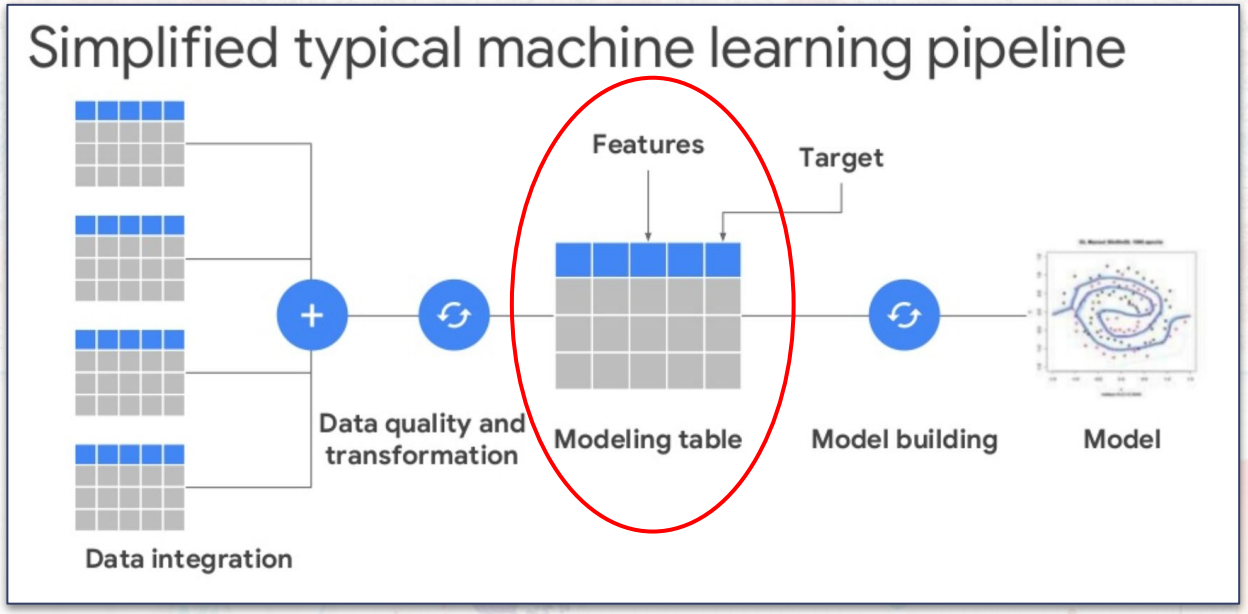
1. Que **existan** y que sean **accesibles**
2. **Relevantes** para el problema. El experto del dominio puede aportar mucho
3. **Suficientes** para el entrenamiento del modelo de Machine Learning
4. **De calidad**, para que el modelo aprenda adecuadamente

DSC::Energy
Analytics

3.4. Los datos (origen, tipos y procesamiento)

Origen:

- Historial de errores
- Historial de mantenimiento o reparaciones
- Condiciones de funcionamiento de la máquina
- Características estáticas



Tipos de datos:

- Datos temporales
- Datos estáticos

Procesamiento:

- Organizar los datos en una **tabla de registros**, en los que cada fila representa una instancia de aprendizaje, y las columnas representan las características (variables o “features”)
- Incorporar la variable destino (“target”) con el valor de cada instancia, que es lo que se pretende aprender

DSC::Energy
Analytics

3.5. Procesamiento de los datos (1 / 2)

Preparación de la tabla global de datos:

- **Datos temporales:** cada registro debe pertenecer a una unidad de tiempo y debe ofrecer información distinta. Si la frecuencia es alta y los datos apenas cambian, hay que agrupar en otras unidades
- **Datos estáticos:**
  - **Registros de mantenimiento**. Deben tener identificador del recurso o máquina, así como marca de tiempo, para poder unir a la tabla de datos temporales
  - **Registros de errores**. Deben registrar los tipos de errores y el recurso, así como también la marca de tiempo, para también unir a la tabla de datos temporales
  - **Metadatos de máquinas y operador**: las propiedades de cada recurso o máquina, con su identificador, se integrarán mediante una unión a la tabla principal

DSC::Energy Analytics

3.5. Procesamiento de los datos (2 / 2)

Preprocesado:

- Limpieza
- Rellenado o eliminación de valores que faltan (“missing”)
- Dependiendo del algoritmo, puede requerir normalización

Ingeniería de variables (“feature engineering”):

- Una vez tenemos la tabla de datos unificada podemos incorporar nuevas variables que aporten información no contenida en la tabla, como por ejemplo:
  - Agregados acumulados (Ej. media o desviación acumulada de la última semana, mes, día, ..)
  - Agrupación de variables
  - Transformación de variables: logaritmos, ratios entre variables
  - Codificación de variables categóricas: “one-hot encoding”, “label encoding”, ...

DSC::Energy
Analytics

3.6. Modelos

Clasificación binaria:

- Predicción de probabilidad de que una pieza de un equipo presente un error en un período futuro

Regresión (predicción de variable continua):

- Predicción de la vida útil restante de un recurso
- Predicción de variables para detección de anomalías

Clasificación multiclase:

- Predicción de probabilidad de error según varios tipos posibles (clases)
- Predicción del intervalo de error según un número discreto de períodos (ej: fallo en 1, 2, 3, 4, 5 meses)

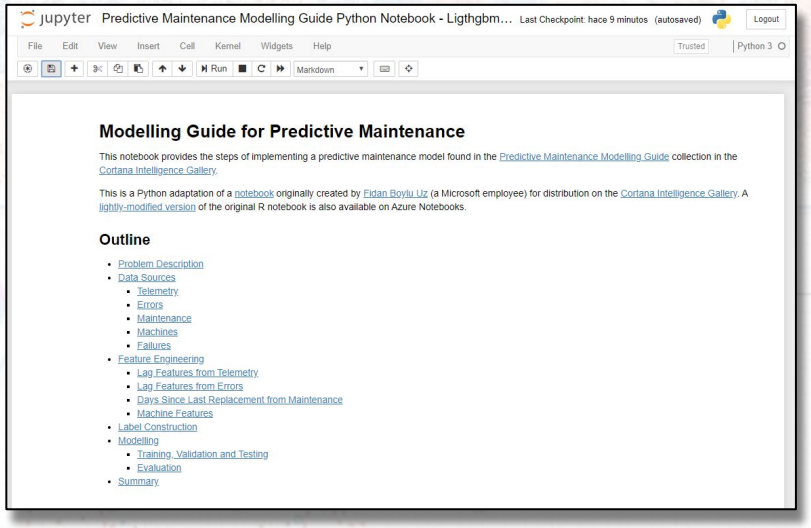
DSC::Energy
Analytics

4. Ejemplo con datos (en Python)

Problema: Predecir qué componente va a fallar próximamente
Sector: Fabricación
Tipo modelo: Clasificación multiclase (5 clases)
Datos:

- ❏ Telemetría: 876100 registros, 6 columnas (horario)
- ❏ Errores: 3919 registros, 3 columnas
- ❏ Mantenimiento (4 componentes): 3286 registros, 3 columnas
- ❏ Máquinas (4 modelos): 100 registros, 3 columnas
- ❏ Fallos: 761 registros, 3 columnas

Algoritmo: Light GBM (Gradient Boosting Trees)
Lenguaje programación: Python
Entorno: Jupyter notebook



DSC::Energy
Analytics

5. Conclusiones

1. Definir claramente el **problema** a resolver y lo que se busca
2. Empezar por un **proyecto piloto**
3. Definir bien el **equipo de trabajo**
4. Recopilar los **datos**
5. Seguir el esquema de proyecto, **muy importante el EDA**
6. Tras implantación, **analizar valor aportado** y vender internamente

DSC::Energy
Analytics

Gracias