DSC::Energy Analytics

CLANER ASOCIACIÓN DE ENERGÍAS RENOVABLES DE ANDALUCÍA

Aprendizaje Automático ("Machine Learning") en Energías Renovables

Antonio Marín Écija

Sevilla, 30 de junio de 2020

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

DSC::Energy Analytics

CLANER ASOCIACIÓN DE ENERGÍAS RENOVABLES DE ANDALUCÍA

ÍNDICE

DSC::Energy Analytics

CLANER ASOCIACIÓN DE ENERGÍAS RENOVABLES DE ANDALUCÍA

1. Introducción DSC Energy

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 y la industria, así como a otros sectores.
Entre otras áreas destacan la **Supervisión de Rendimiento**,
los **Modelos Predictivos** y el **Mantenimiento Predictivo Inteligente**.

DSC::Energy Analytics

CLANER ASOCIACIÓN DE ENERGÍAS RENOVABLES DE ANDALUCÍA

Fotovoltaica y Termosolar

Eólica

Biomasa y Agroenergía

Emisiones
Mercado Carbono

Distribución y Smartcities

Construcción y
Explotación

Depuración y Tratamiento
de Aguas Residuales

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 multinacionales como pymes, 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.**

Data Science = Ciencia de Datos

es un concepto que 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 el **estudio científico** de **algoritmos y modelos estadísticos** que utilizan los sistemas informáticos para realizar una tarea específica sin usar instrucciones explícitas, sino que **se basan en patrones e inferencia de los datos**. Es visto como un subconjunto de **inteligencia artificial**.

Los algoritmos de **aprendizaje automático** crean un modelo matemático basado en datos de muestra, conocidos como "datos de entrenamiento", para hacer predicciones o decisiones sin ser programado explícitamente para realizar la tarea.

2.2. Tipos

Artificial Intelligence

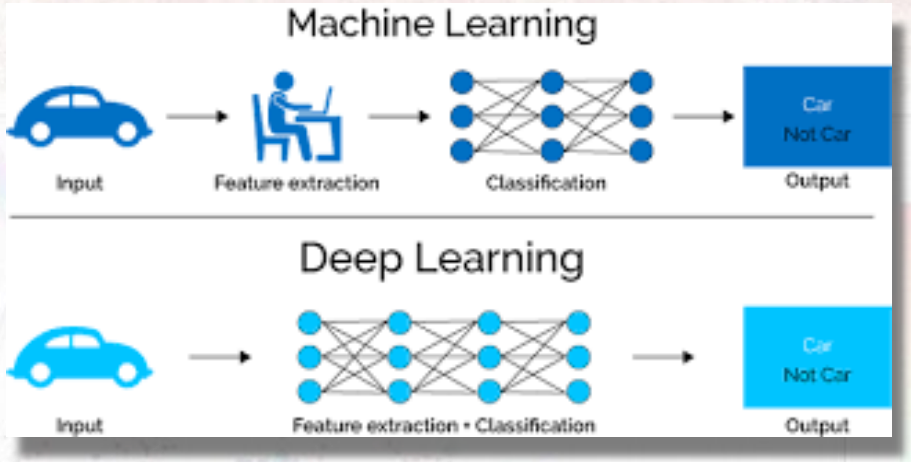
Any technique that enables computers to mimic human intelligence. It includes *machine learning*

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.



2.3. Por qué ahora?

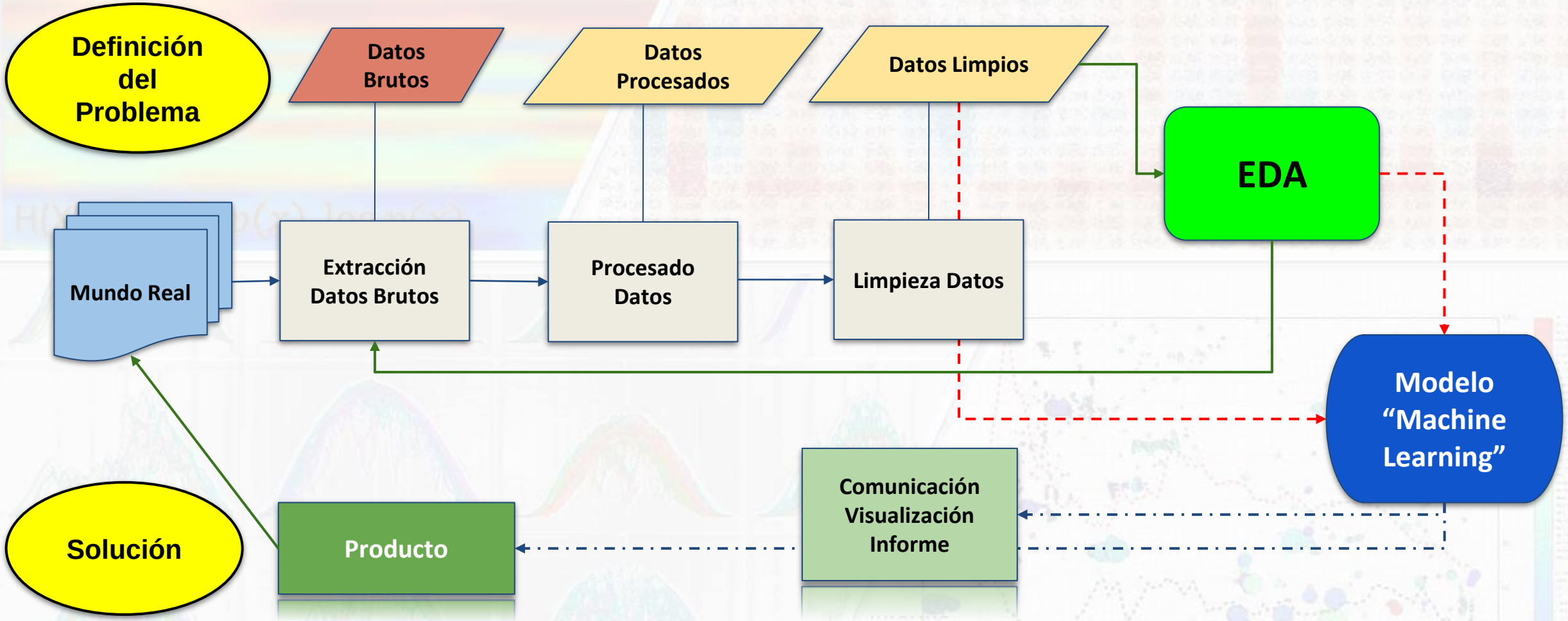
Democratización de la IA - software abierto

- 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

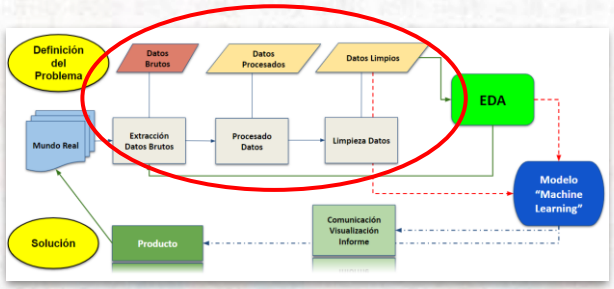
Revolución computación

- 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)

3.1. Proceso Proyecto Machine Learning. Esquema

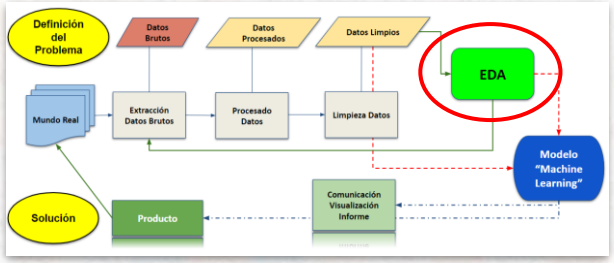


3.2. Los Datos



- 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
- Según se disponga de la solución para aprender o no, se puede hablar de **aprendizaje supervisado o no supervisado**
- **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”)

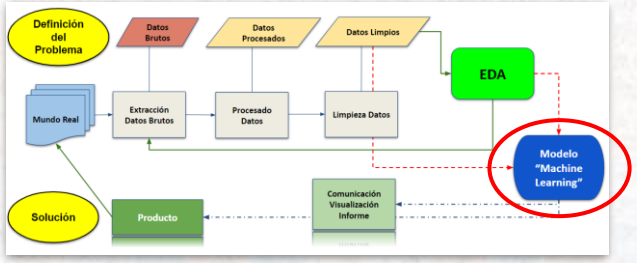
3.3. El EDA[1] (Análisis Exploratorio de Datos - "Exploratory Data Analysis")



- 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 variables: boxplots (caja y bigotes), distribuciones, scatters (nube de puntos), heatmaps, ….
  - Visualización de **relaciones entre variables** independientes y variable objetivo

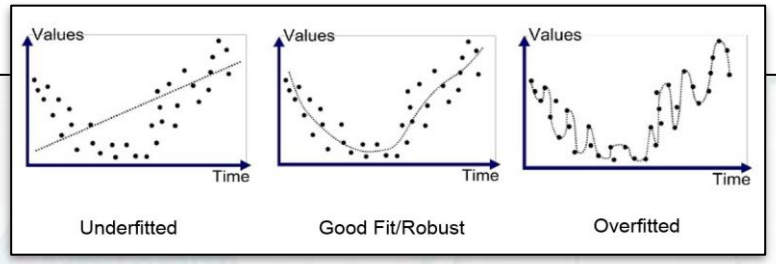
[1] El término EDA fue utilizado por primera vez por John W. Tukey, un reconocido estadístico estadounidense

3.4. Entrenamiento del modelo

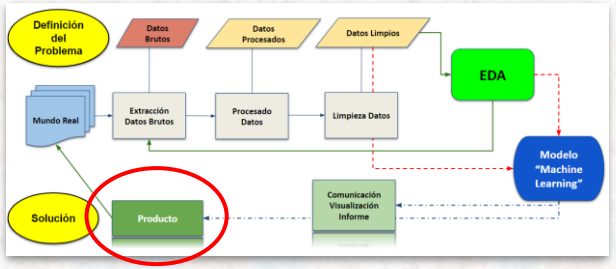


- Selección del **método de validación**: validación cruzada ("cross validation"). Partición de datos en entrenamiento, validación y test.
- Selección de la **métrica**: RMSE, MAE, AUC, ...
- Selección del **algoritmo**
- Selección de los **parámetros** del algoritmo ("hyperparameter tuning")
- Evaluación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



3.5. 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

DSC::Energy Analytics

CLANER ASOCIACIÓN DE ENERGÍAS RENOVABLES DE ANDALUCÍA

3.6. ¿Por dónde empezar?

1. ¿Tenemos **datos** registrados digitalmente?
2. ¿Qué **problemas** pensamos podríamos resolver?
3. ¿Cómo está enfocando mi **organización** la transformación digital y el uso los datos de forma avanzada?
4. Hacer una **lista** de posibles proyectos, definiendo siempre el problema a resolver y los beneficios que podría tener
5. Empezar por un **proyecto piloto inicial**, que pueda ser representativo de más que se pueda afrontar
6. Medir el **impacto económico** del proyecto y promocionarlo
7. Seguir avanzando con más proyectos de la lista, aumentando el despliegue y el alcance

DSC::Energy Analytics

CLANER ASOCIACIÓN DE ENERGÍAS RENOVABLES DE ANDALUCÍA

3.7. A tener en cuenta

1. Herramienta de **mejora de la productividad**
2. Evolución de la **computación**: en 25 años se ha multiplicado por 1 millón !! Y de los precios: 30k USD/GFlop en 1997 a 0.03 USD/GFlop en 2019
3. Se necesitan **recursos**: locales/nube. Pero son cada vez más baratos
4. Sin **datos**, hay poco que hacer
5. Es muy importante la implicación de la **Dirección** si se quiere impulsar a escala
6. Quien conoce los problemas es **Negocio/Operaciones**, pero hay que contar con IT. Es buena la figura impulsora transversal: “Transformación Digital”.
7. Y todo esto tiene sentido si al final vemos que aporta a la **Cuenta de Resultados** de alguna forma o de otra, conviene siempre valorarlo

DSC::Energy Analytics

CLANER ASOCIACIÓN DE ENERGÍAS RENOVABLES DE ANDALUCÍA

4. Tipos de Proyectos

Evaluación de Mejoras

Modelado de energía generada en parques eólicos, clúster de turbinas o turbinas individuales para **evaluar mejoras implantadas**

Detección de Anomalías

Modelado de temperaturas y/o vibraciones de componentes de equipos, para **mantenimiento predictivo**

Modelado de Generación de Energía

Modelado para **seguimiento de rendimiento** de parques eólicos, tanto a nivel parque como a nivel individual por turbina

DSC::Energy Analytics

4.1. Evaluación de Mejoras en Turbinas eólicas

CLANER ASOCIACIÓN DE ENERGÍAS RENOVABLES DE ANDALUCÍA

Problema Planteado:

- Parques en funcionamiento
- Mejoras puntuales con alto coste económico
- Confirmación de dicha mejora
- Elección de próximas turbinas a mejorar

Solución Ejecutada:

- Captura y preparación de datos alta frecuencia
- Análisis, modelado y predicción
- Modelo ML de Turbinas Modificadas a partir de comportamiento de las vecinas

Mejora Conseguida:

- Confirmación y cuantificación de la mejora
- Reducción del tiempo de evaluación al usar datos de alta frecuencia
- Evaluación en todos los sectores

DSC::Energy Analytics

4.2. Detección de Anomalías en temperaturas componentes principales

CLANER ASOCIACIÓN DE ENERGÍAS RENOVABLES DE ANDALUCÍA

Problema Planteado:

- Fallos muy costosos y no previsibles
- Tiempos de reparación largos
- Pérdidas de producción

Solución Ejecutada:

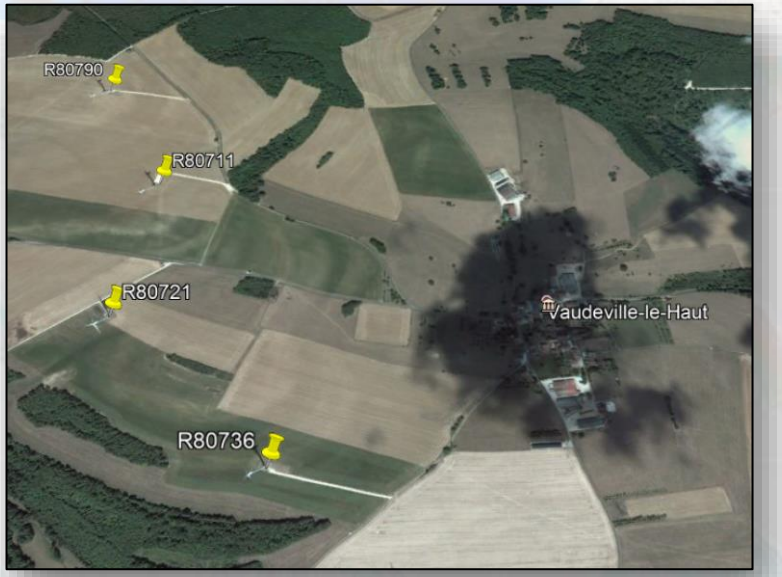
- Captura y preparación de datos
- Identificación momentos de fallos
- Modelo ML de predicción de temperaturas

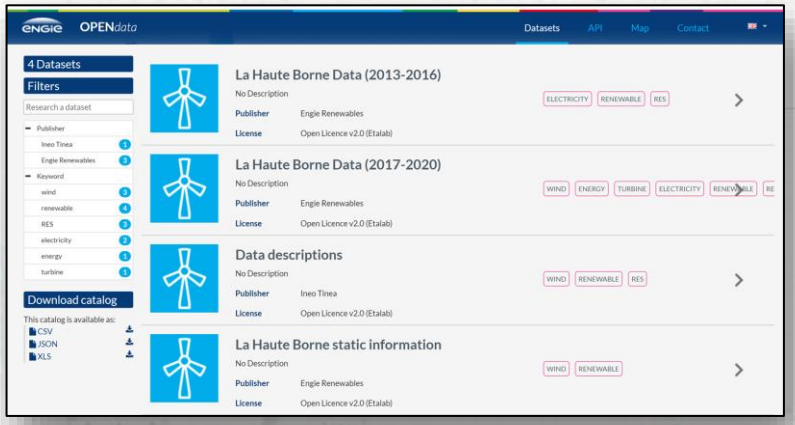
Mejora Conseguida:

- Sistema de alerta de anomalías
- Posibilitar Mantenimiento Predictivo
- Reducción costes mantenimiento

Parque Eólico La Haute Borne (8.2 MW):

- Propietario y operador: Engie
- Turbinas: 4 x Senvion MM82 - 2050 kW – Diámetro rotor 82 m – Altura 80 metros
- Localización: Vaudeville-le-Haut (departamento: Meuse, región: Grand Est) - Francia
- Dataset: datos 10 minutales años 2013 a 2016 (840380 observaciones, 138 variables)
- Variables: potencias, velocidad, pitch, yaw, voltaje, frecuencia, par, revoluciones, temperaturas, …





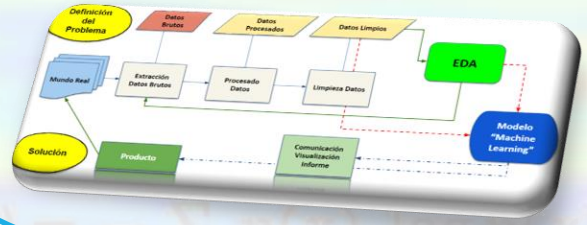
https://opendata-renewables.engie.com/explore/index

DSC::Energy Analytics

Resumen

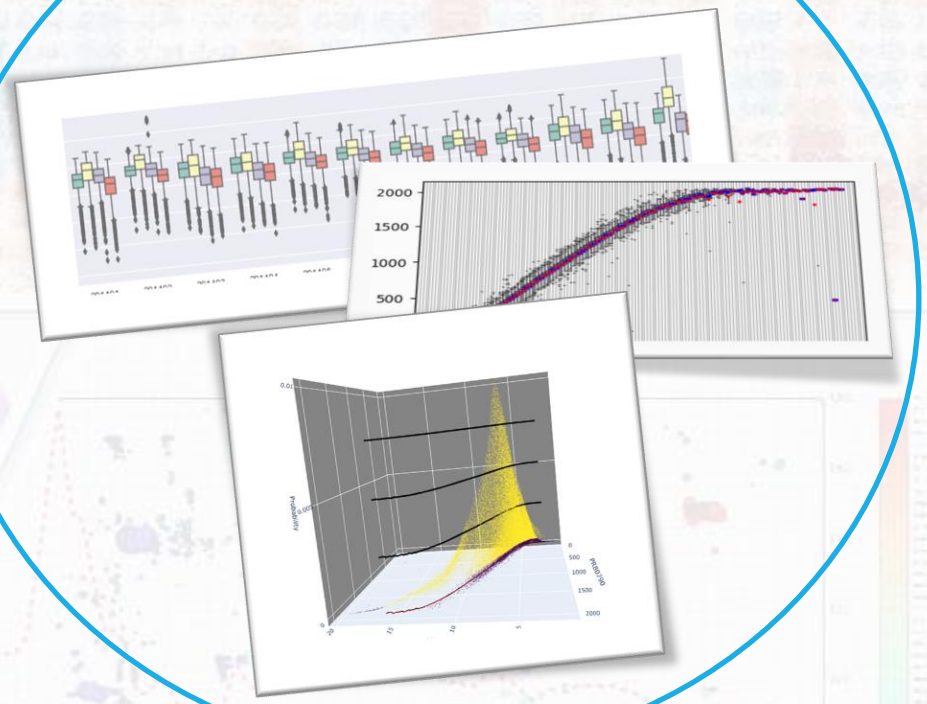
CLANER ASOCIACIÓN DE ENERGÍAS RENOVABLES DE ANDALUCÍA

Procedimiento (primero el "Problema")



Machine Learning

Visualización ("EDA")



Los "Datos"

DSC::Energy Analytics

CLANER ASOCIACIÓN DE ENERGÍAS RENOVABLES DE ANDALUCÍA

PREGUNTAS

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