Uso de técnicas de minería de datos para la detección de anomalías y predicción de la producción en una turbina eólica.

Máster Universitario de Ciencia de Datos

Trabajo fin de máster

Javier Pérez Esteras (Autor)
Raúl Parada Medina (Profesor colaborador)
Jordi Casas Roma (Profesor responsable)

Junio de 2020

Uso de técnicas de minería de datos para la detección de anomalías y predicción de la producción en una turbina eólica.

1. Justificación.
2. Objetivos principales y secundarios.
3. Metodología y productos obtenidos.
4. Investigación y estado del arte.
5. Desarrollo e implementación.
6. Conclusiones y trabajos futuros.

1. Justificación del trabajo

¿Por que el sector eólico y la detección de anomalías?

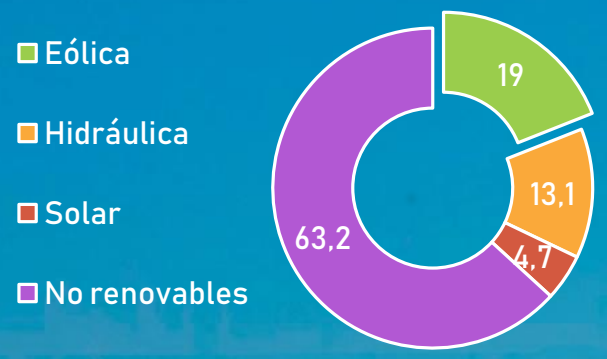
- Energía Limpia e Inagotable
- Sector en auge
- 19% de la estructura nacional
- 24.000 MW instalados
- Clave Para objetivos europeos

¿ Que beneficios puede aportar ?

- Reducción de costes
- Incremento de la producción
- Extensión de la vida útil
- Mejor gestión de los recursos

Datos de 2018 según Red Eléctrica de España

Estructura de generación (%)



Evolución de la potencia instalada (GW)

2. Objetivos

2.1 Objetivo Principal

Validar técnicas de minería de datos en un conjunto propietario para:

- Detección de anomalías
- Predicción de la producción

2.2 Objetivos secundarios

- Conocer el funcionamiento de un aerogenerador.
- Aprender las técnicas *machine learning* empleadas para detección de anomalías y predicción de series temporales.
- Aplicación de los modelos a los datos del SCADA para detectar desviaciones sobre la normalidad.
- Adquisición de la predicción meteorológica desde proveedores de internet.

Fases para realización del trabajo

1. **Documentación**
   - Funcionamiento de un aerogenerador y del sector eólico.
   - Algoritmos y técnicas de minería de datos.
   - Trabajos realizados.

2. **Selección de herramientas**
   - Python, scikit-learn, keras y tensorflow.

3. **Desarrollo e implementación**
   - Análisis exploratorio de los datos.
   - Determinación de la normalidad.
   - Selección de características y creación de modelos.
   - Simulación de los fallos.
   - Evaluación de técnicas de detección de anomalías.
   - Adquisición de la previsión meteorológica.
   - Evaluación de algoritmos de predicción de potencia.

4. **Productos obtenidos**

Modelos | Clasificador | Validación | Valoración

scikit learn

Keras

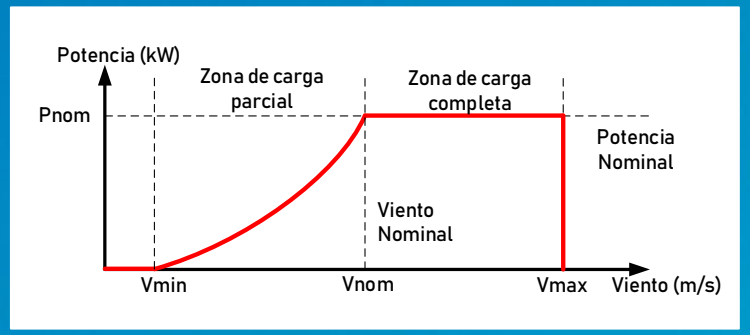
TensorFlow

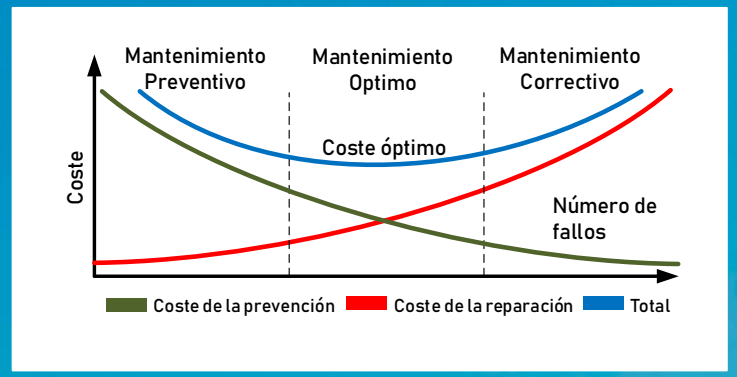
python™

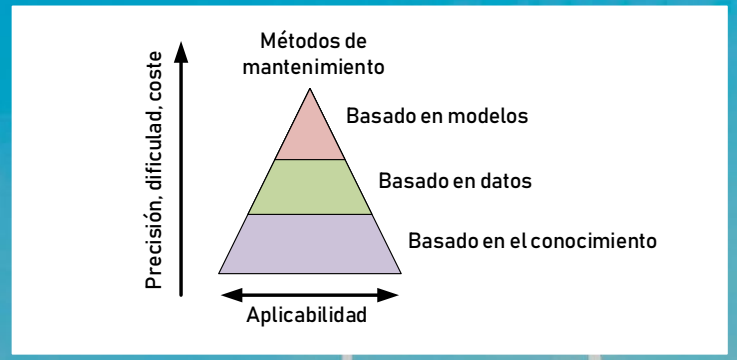
SPYDER

4.1 Sector eólico

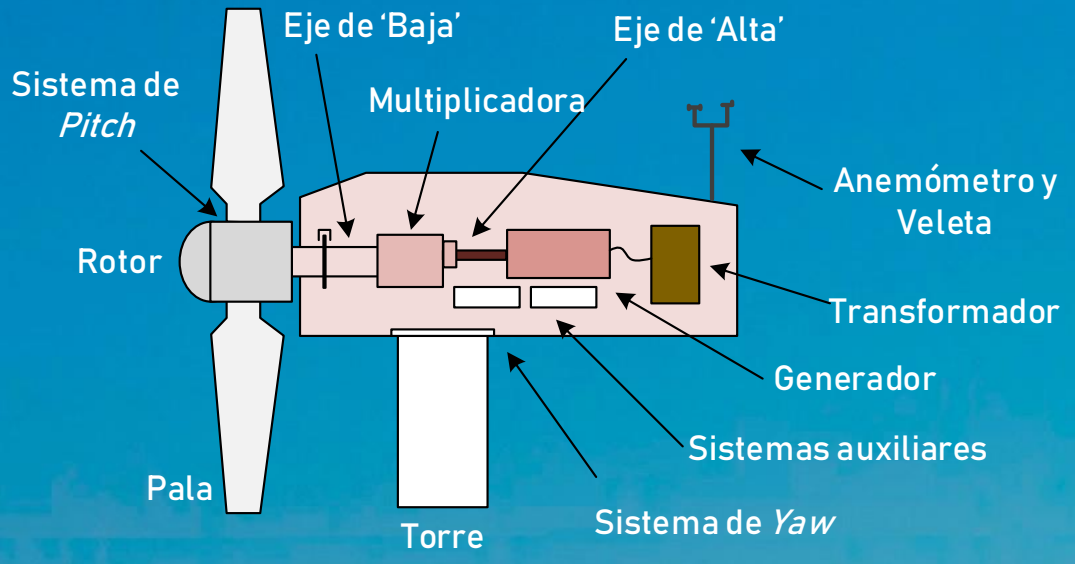
- Principales sistemas: *Yaw* y *Pitch.*
- Estrategia de control: maximización de la producción.
- La importancia del mantenimiento preventivo.
- Complejos sistemas de monitorización en auge.
- La integración de la energía en parque con previsiones *Online* y *Offline.*

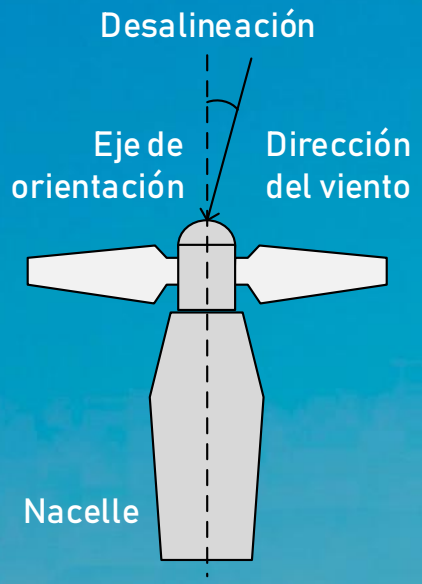






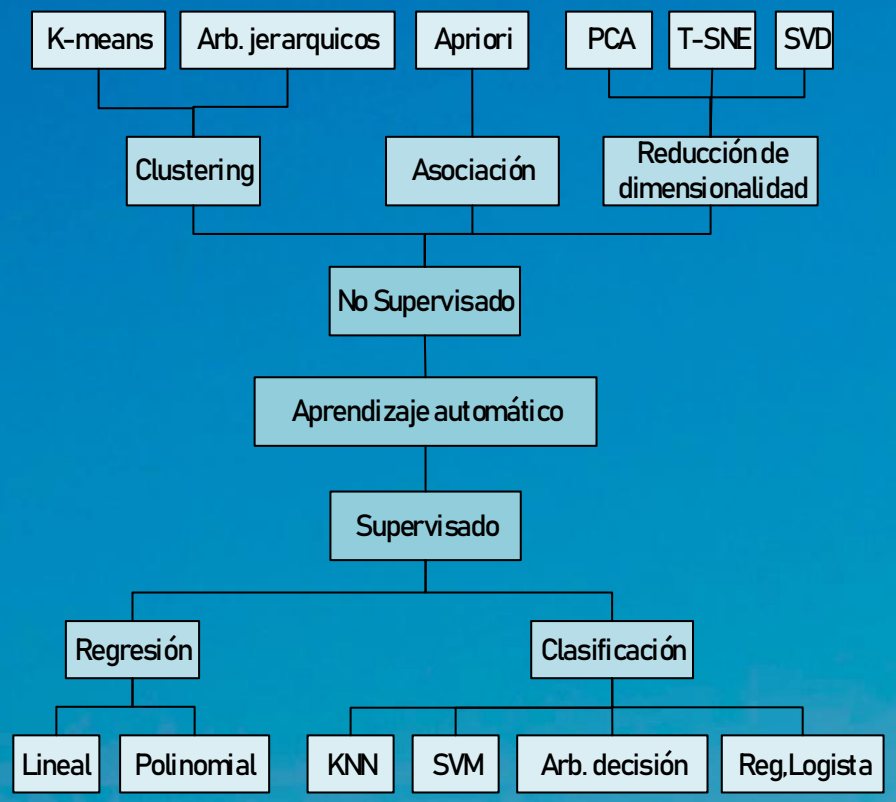
Anatomía





4.2 Aprendizaje automático

Engloba la técnicas y métodos que permiten aprender de los datos, generalizar comportamientos y realizar inferencias o predicciones para un conjunto más amplio y desconocido.



4.2.1 Predicción de la producción

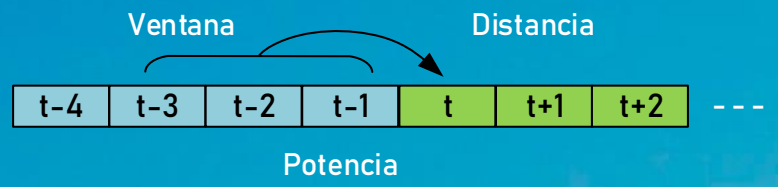
A medio plazo. En el próximo día. (*Offline*)

- Basado en modelos.

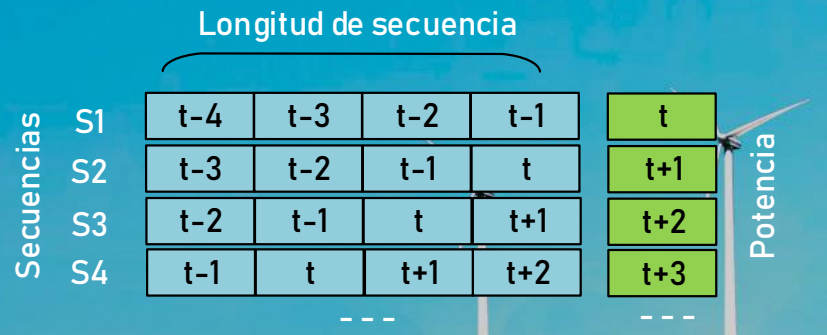
A corto plazo. En los próximos 10 minutos. (*Online*)

- Métodos auto regresivos: RF, LR, SVR.
- Basado en redes neurotales. LSTM .
- Combinados: ARIMA.
- Ultimo valor.

Autoregresión



Secuencias para LSTM



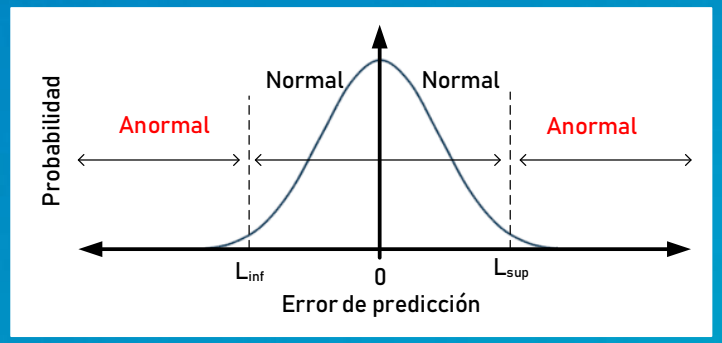
4.2.2 Detección de anomalías

- Tipos de datos: continuos y discretos.
- Tipos de anomalía: Individual, contextual y colectiva.
- Tipos de detección: supervisada, semisupervisada y no supervisada.

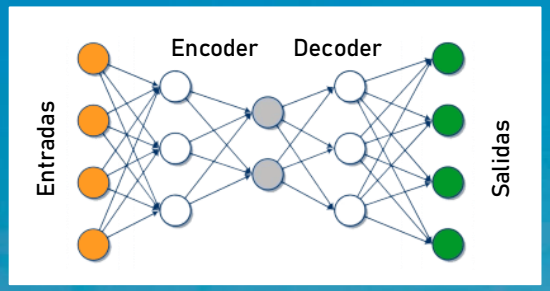
Detección basada en métodos estadísticos:
Se basan en la asunción de que los datos sin anomalías se ubican en las regiones de probabilidad más alta, mientras que los datos con fallo se ubican en las regiones de probabilidad más baja.

Detección basada en autoencoders
Basada en la eficacia de reconstrucción de una red neuronal que aprende a comprimir y descomprimir los datos.

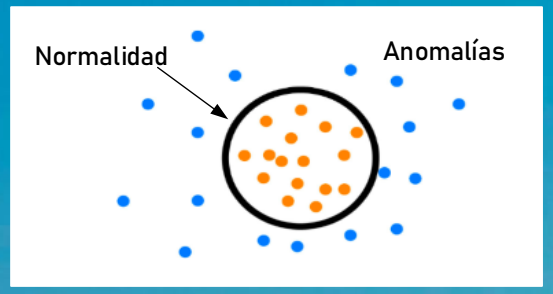
Basada en métodos de una clase
Donde se asume me un porcentaje de las muestras son diferentes a las demás.
Ej. Isolation Forest, One Class SVM, Etc.



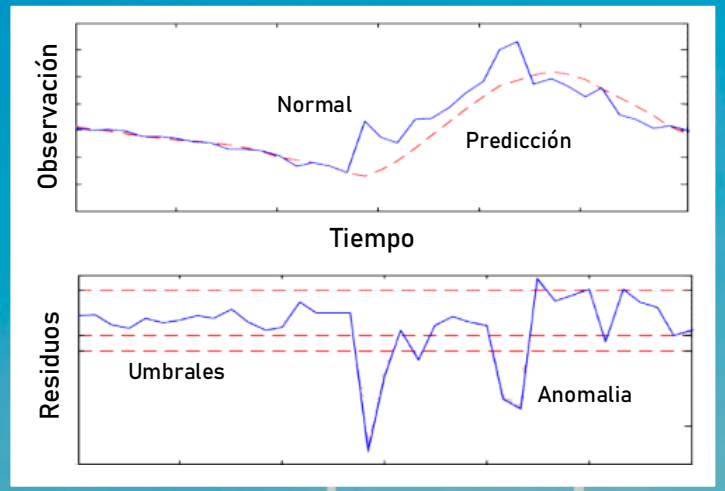
Detección basada en métodos estadísticos



Autoencoder



Métodos de una sola clase



Detección de anomalías basada en umbrales

4.3 Trabajos realizados

Principales portales consultados:

- IEEE Explore: Institute of Electrical and Electronics Engineers. ieeexplore.ieee.org.
- MDPI: Multidisciplinary Digital Publishing Institute. www.mdpi.com.
- ScienceDirect: Base de datos y plataforma digital de Elsevier. www.sciencedirect.com.

Información sobre el sector eólico.

- Tomas Esbensen [8] y Farsoni Saverio [18], Modelos matemáticos y modos de funcionamiento.
- Agencia energética danesa: Integración de la energía del viento en red.

Información sobre el uso de técnicas de minería de datos.

- Meik Schlechtinger [10], Peng Qian [21], Peng Sun [9], Papatheou [20] y otros. Numerosos trabajos con técnicas de todo tipo aplicadas a los aerogeneradores.
- Otros autores aportan números modelos paramétricos y no paramétricos de las principales variables.

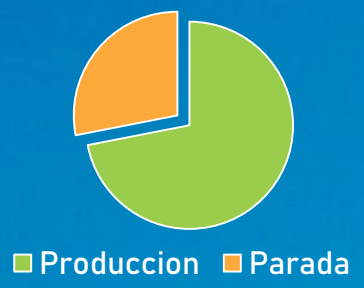
Técnicas observadas:

Técnicas	Referencias en la memoria
Redes Neuronales multicapa. (MLP ANN)	[20] [9] [17] [21] [23] [10]
Sistema adaptativo de inferencia neuro-difusa (ANFIS)	[9] [10]
Máquinas de soporte vectorial (SVM)	[9] [23]
K vecinos más cercanos (KNN)	[27]
Procesos gaussianos (GP)	[26]
Arboles aleatorios (RT, RF, ERT)	[26] [27]
Técnicas de reducción de dimensionalidad (PCA,SOM,PARAFAC)	[13] [17] [22] [23]

5.1 Análisis exploratorio y determinación de los periodos de producción

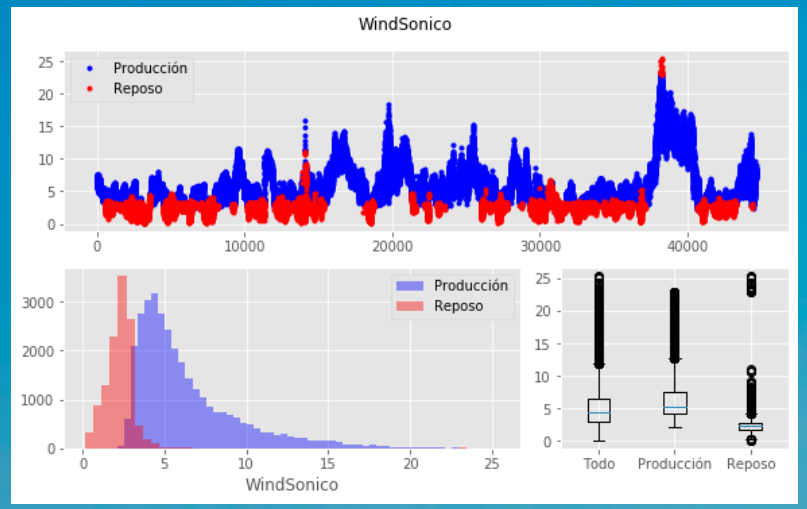
Determinación del régimen de trabajo:

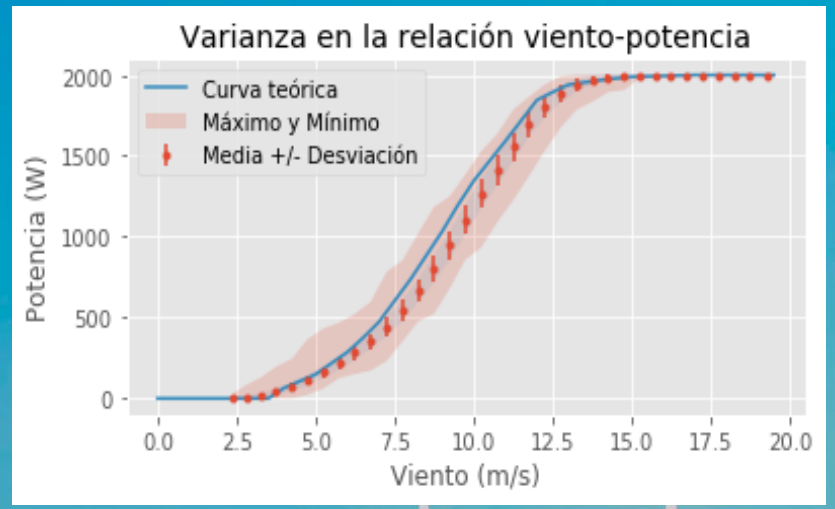
0 W	2 m/s	960 rpm	40 º
Potencia mínima	Viento mínimo	Rev. mínimas	Pitch máximo



Variables destacadas

Descripción	Unidades
Par del generador	N/m
Angulo de pich mínimo	Grados
Angulo de pich de referencia	Grados
Potencia en el estator referencia	W
Potencia reactiva	W
Potencia en el Estator	W
Potencia Total	W
Presión del grupo hidráulico	hPa
Revoluciones del generador	Rpm
Rev. del generador filtrado	Rpm
Revoluciones del rotor	Rpm
Velocidad viento	m/s
Dirección del viento	Grados
Velocidad viento	m/s





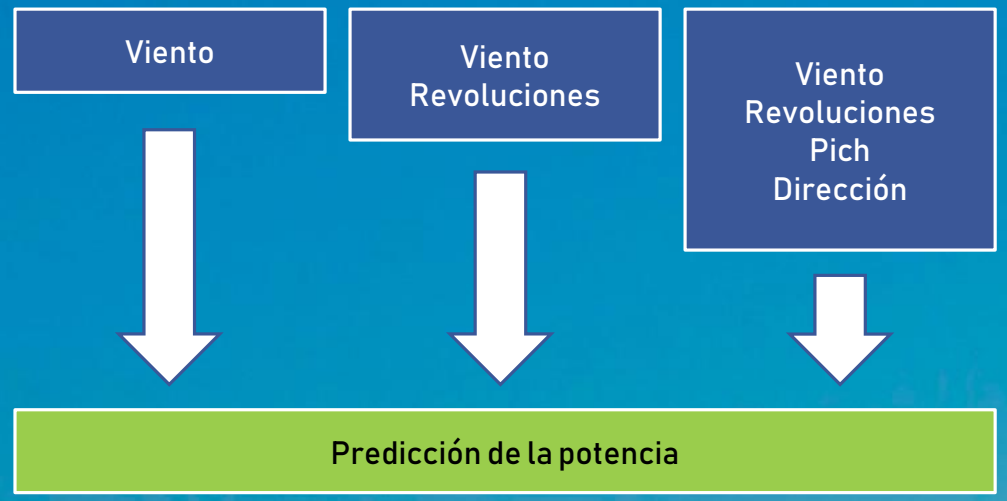
5.2 Selección de características

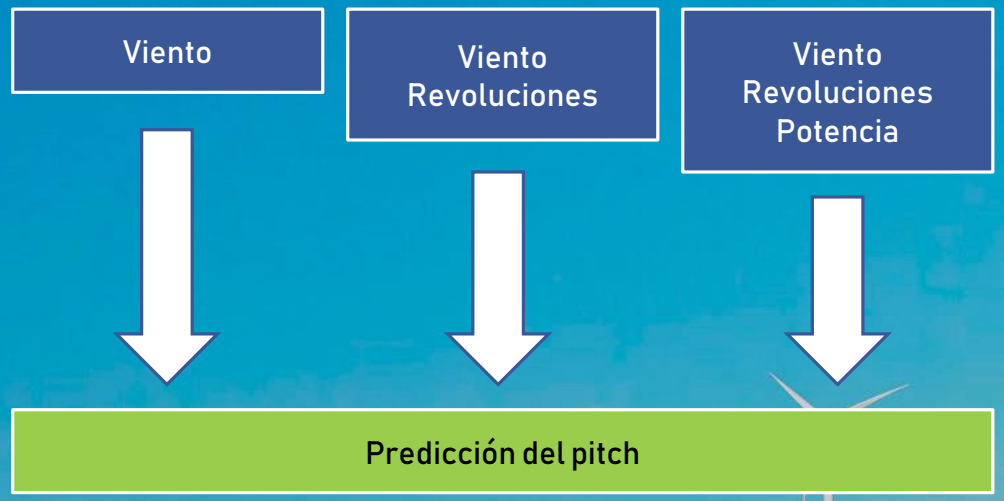
Determinación de las variables que pueden predecir otras variables gracias a :

- Conocimiento de campo de aplicación
- Índice de correlación de *Pearson*.
- Grafica de pares.
- Importancia de características de *Random Forest.*

5.3 Algoritmos usados

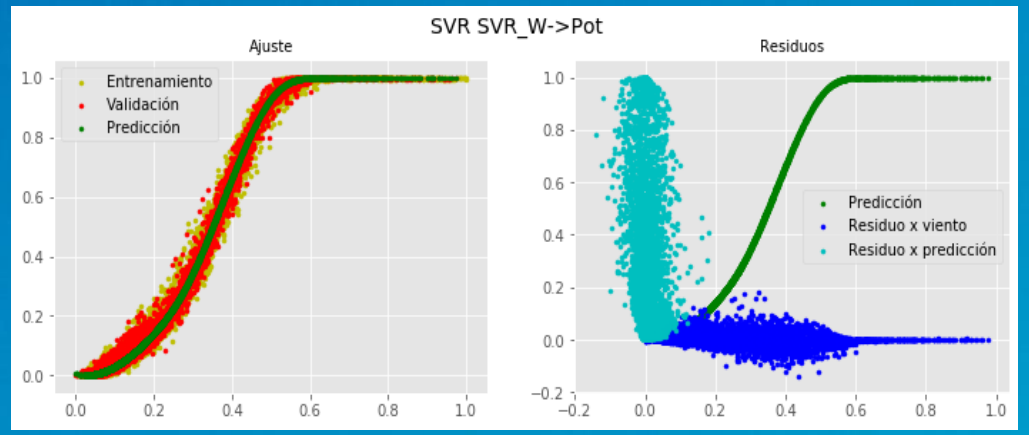
- Arboles aleatorios
- Potenciación del gradiente.
- Maquina de soporte vectorial para regresión.
- Red neural de una capa oculta.
- Red neural de dos capas ocultas.
- K-NN.
- Regresión lineal.





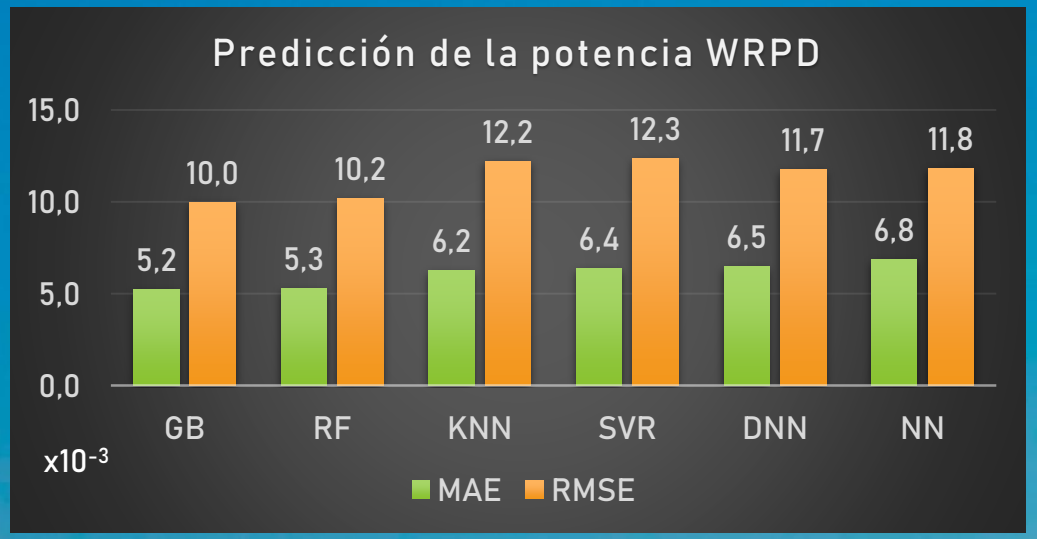
5.3 Creación de modelos

- Normalización por mínimo y máximo.
- Conjuntos de entrenamiento del 75% y validación del 25%.
- Búsqueda de hiperparámetros.
- Validación cruzada.
- Métricas de ajuste: R2, MSE, RMSE, MAE y MAPE.
- Los ajustes son buenos en general excepto para LR.
- Los residuos no siguen la distribución normal.



Ajuste de la predicción de potencia por SVR usando solo el viento

Resultados



5.4 Detección de anomalías

- Buscamos anomalías contextualizadas.
- Ata varianza en la zona media de vientos.
- Simulación de dos de los principales fallos.
- Construcción de un sistema clasificador.
- Utilizaremos métodos estadísticos y autoencoders.
- Evaluación mediante matriz de confusión y otros estadísticos.
- Entrenamiento con días pares y validación con días impares.

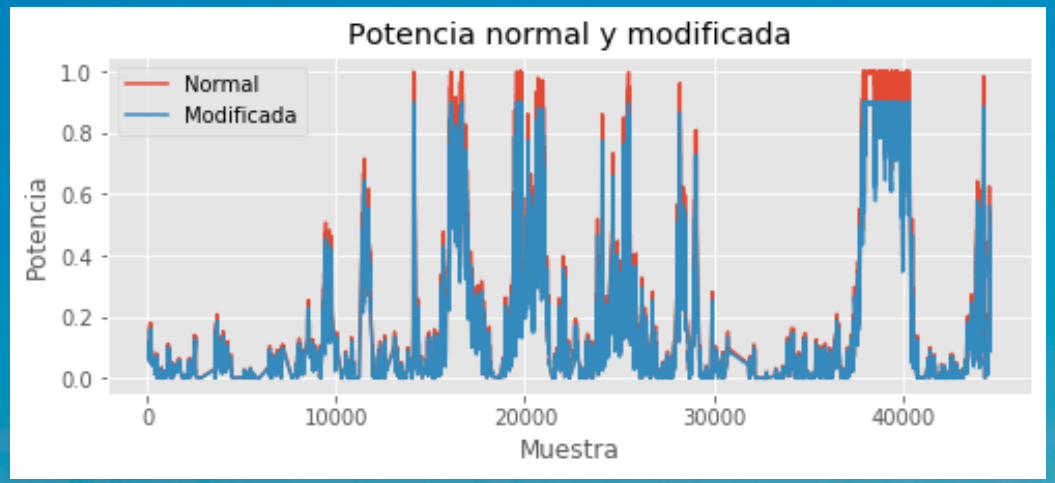
Evaluación

- Conjunto balanceado de test
- Matriz de confusión
- Sensibilidad
- Precisión
- Curvas Sensibilidad-Precisión

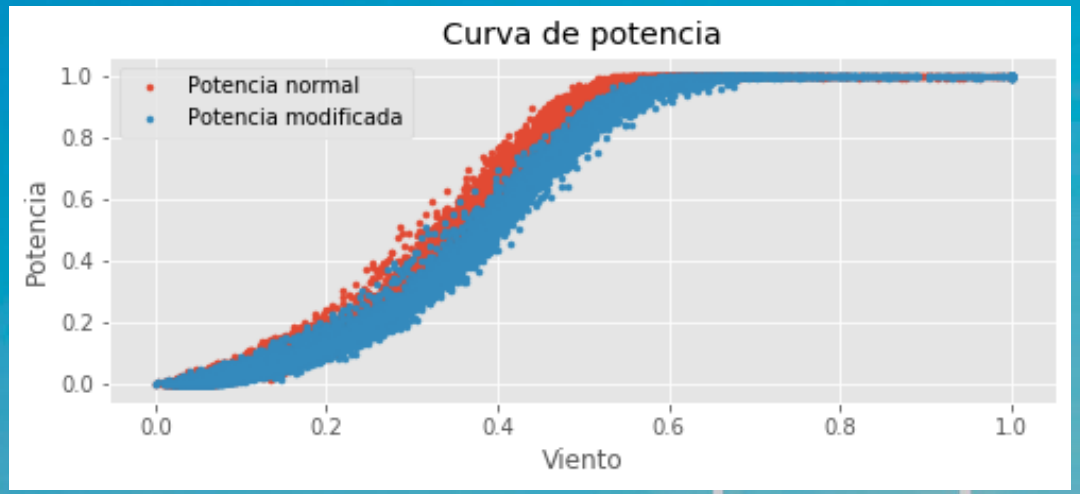
		Predicción	
		P	N
Observación	P	VP	FP
	N	FN	VN

Matriz de confusión

Pérdida de potencia



Desalineación



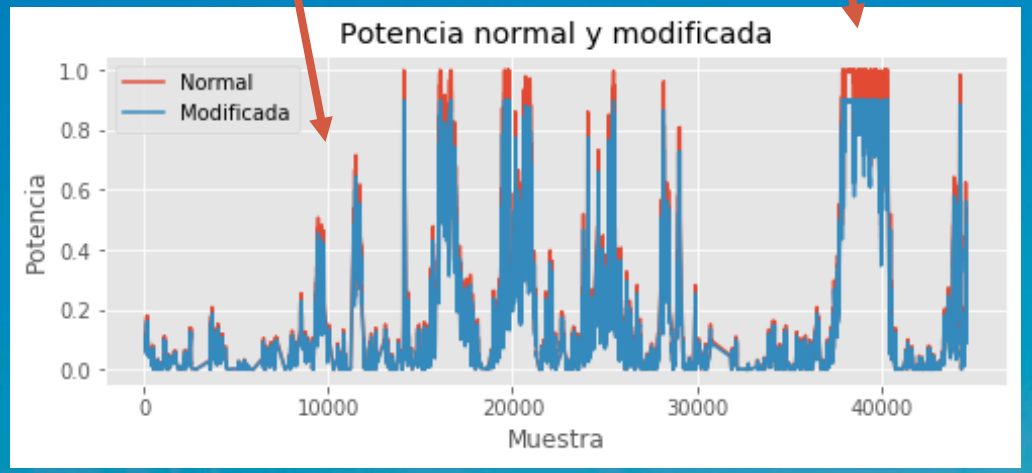
5.5 Procedimiento de métodos estadísticos

5.6 Pérdida de potencia

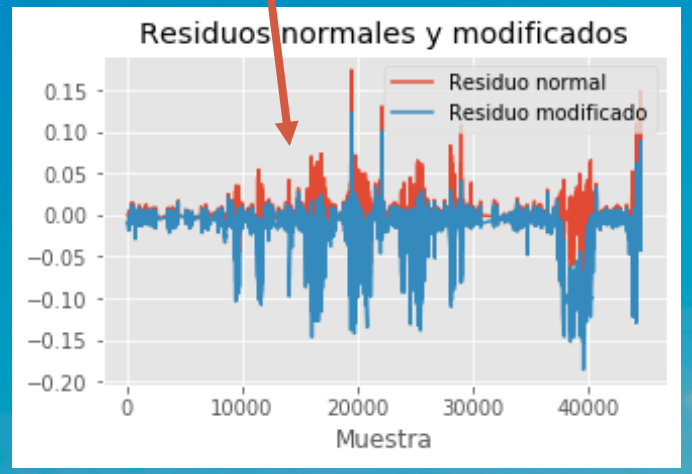
- Simulamos el fallo con una reducción de la potencia generada en un 10%.
- Conjunto de validación de 50% días normales y 50% días con fallo simulado.
- Clasificación de “día normal” o “día anómalo”.

La perdida de potencia queda enmascarada en vientos medios

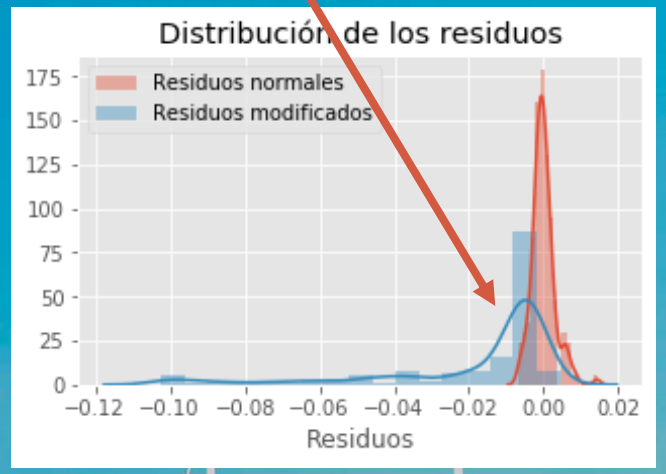
La perdida de potencia es evidente cuando el viento supera el valor nominal



Hay mayor error en las muestras anómalas



La distribución de los residuos es diferenciadora. Hay que localizar un umbral con el que podamos separar los normales de los anómalos



5.6 Detección de la pérdida de potencia

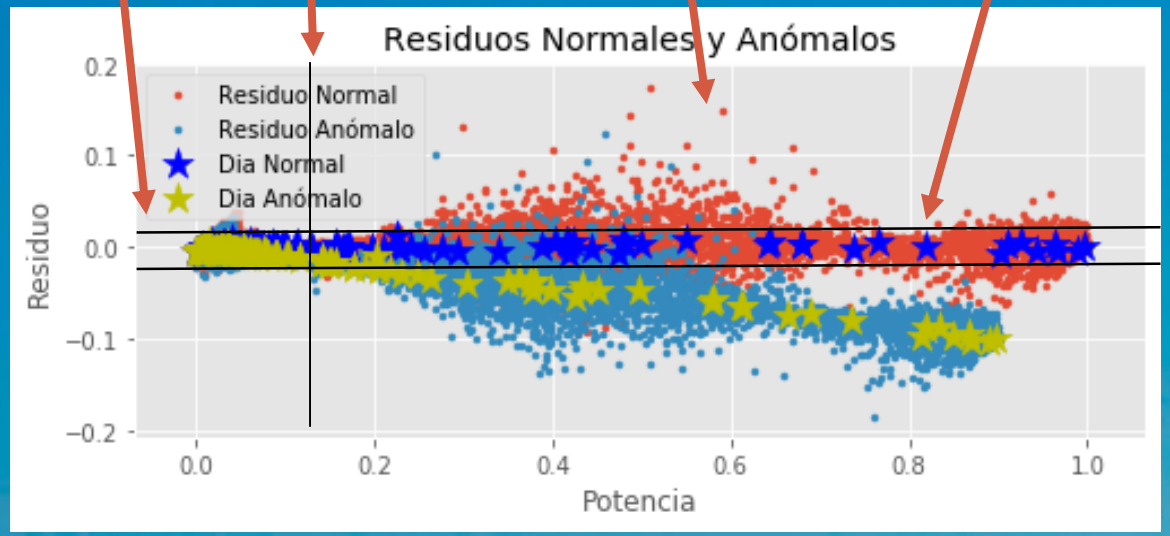
- Predecimos la potencia usando WRPD->Pot.
- Se reduce la varianza promediando a valores diarios.
- El establecimiento de los umbrales determina la precisión (Soporte, potencia mínima y umbral de residuo)

Umbral de los residuos

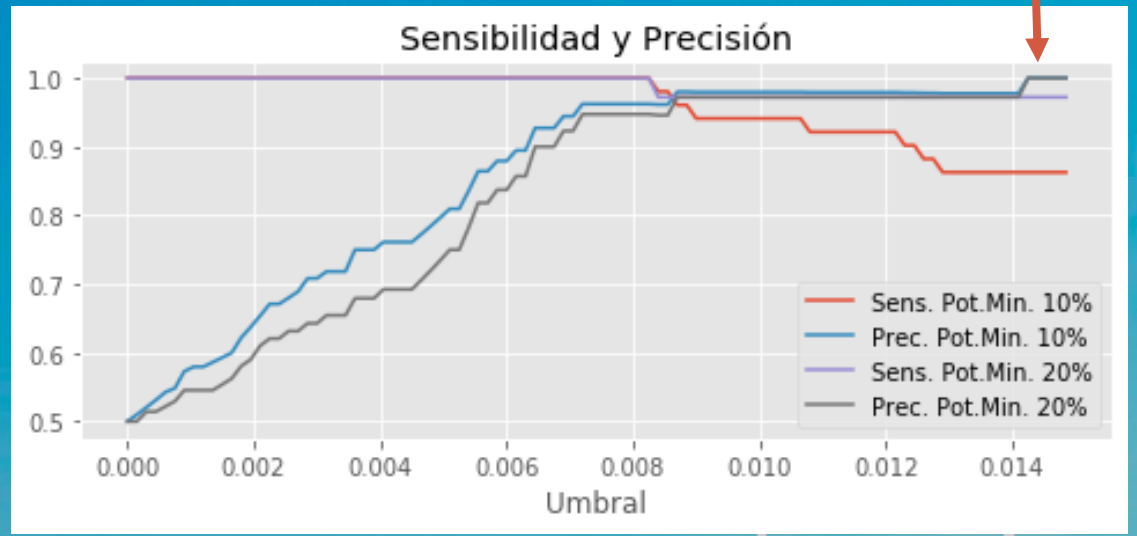
Umbral de potencia mínima

Los residuos individuales están muy dispersos

Los residuos promediados por días son diferenciadores



Pot. Min.	Umbral	TP	FN	FP	TN	Sensibilidad	Precisión
10 %	0,0030	51	0	21	30	1,00	0,71
	0,0068	51	0	4	47	1,00	0,93
	0,0105	48	3	1	50	0,94	0,98
	0,0143	44	7	0	51	0,86	1,00
20 %	0,0030	36	0	20	16	1,00	0,64
	0,0068	36	0	4	32	1,00	0,90
	0,0105	35	1	1	35	0,97	0,97
	0,0143	35	1	0	36	0,97	1,00



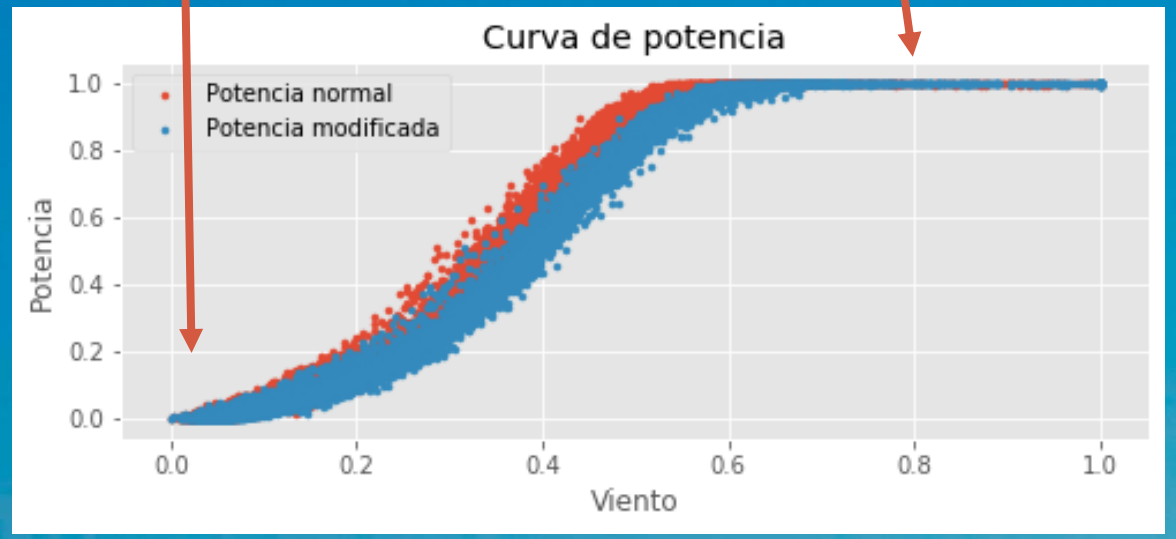
5.7 Desalineación del rotor

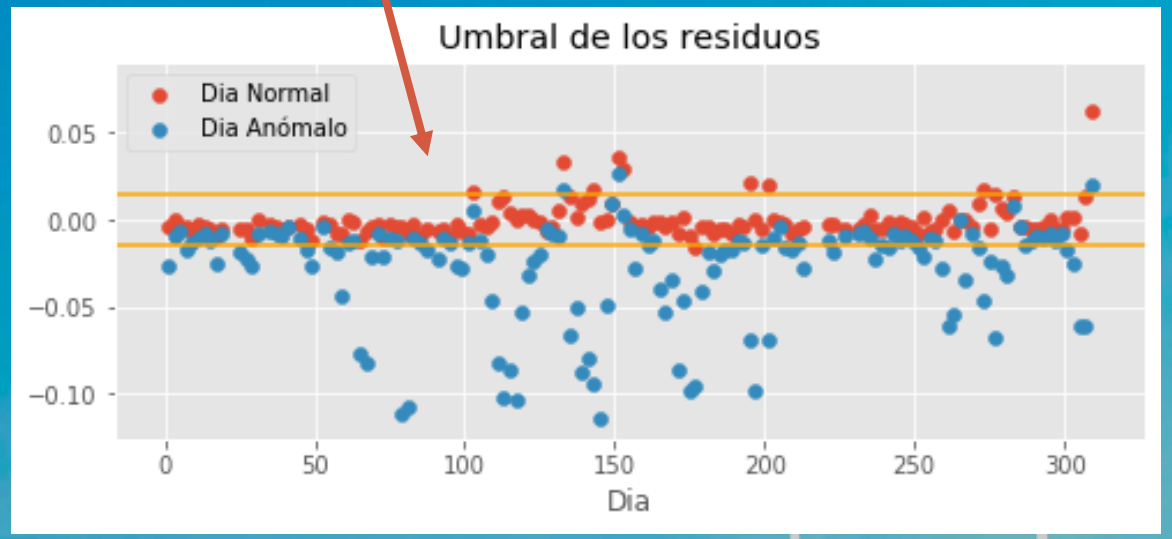
- Simulamos la desalineación incrementando el viento en un 10% sin sobrepasar limites observados.
- Para producir la potencia nominal se necesita más viento del habitual, pero se alcanza igualmente la producción máxima.

Si el viento es muy bajo, la detección va a ser difícil

Si el viento alcanza la potencia máxima, también va a ser difícil.,

Umbral de los residuos promediados para la diferenciación de días normales y anómalos

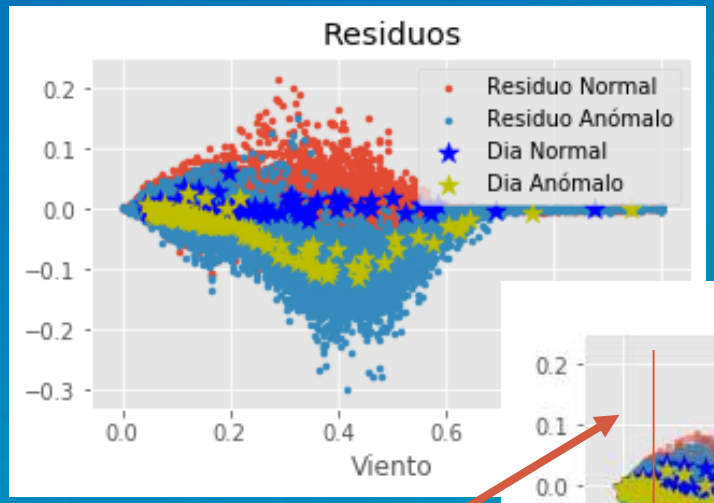




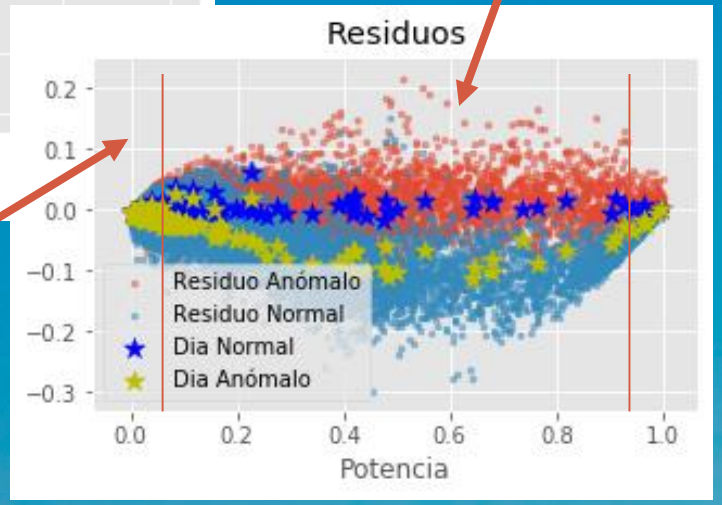
5.8 Detección de la desalineación del rotor

- Usamos modelo de predicción Viento->Potencia.
- El umbral de potencia ahora es de potencia mínima y máxima.

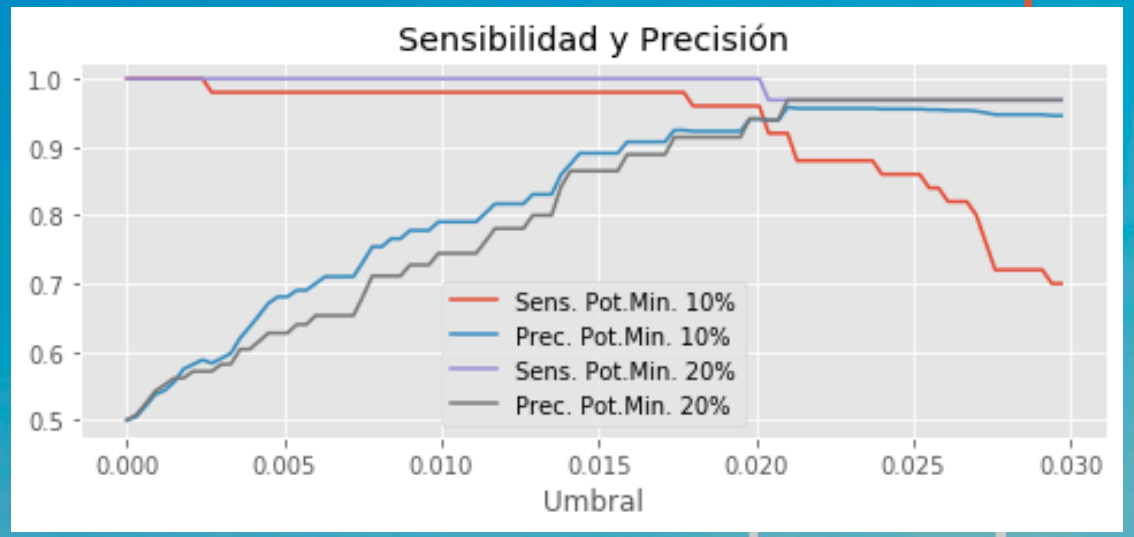
Pot. Min.	Umbral	TP	FN	FP	TN	Sensibilidad	Precisión
10 %	0,0060	49	1	21	29	0,98	0,70
	0,0135	49	1	10	40	0,98	0,83
	0,0210	46	4	2	48	0,92	0,96
	0,0285	36	14	2	48	0,72	0,95
20 %	0,0060	32	0	17	15	1,00	0,65
	0,0135	32	0	8	24	1,00	0,80
	0,0210	31	1	1	31	0,97	0,97
	0,0285	31	1	1	31	0,97	0,97



Los residuos promediados por días son diferenciadores

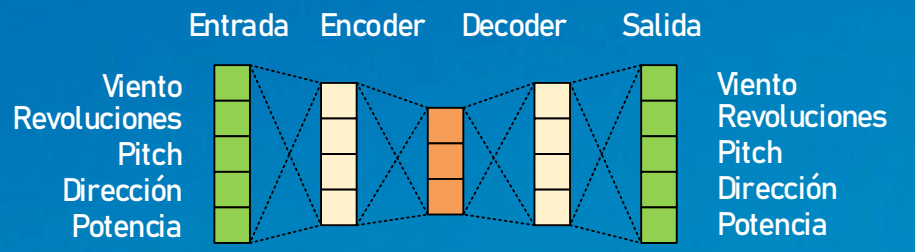


Umbral de potencia mínima



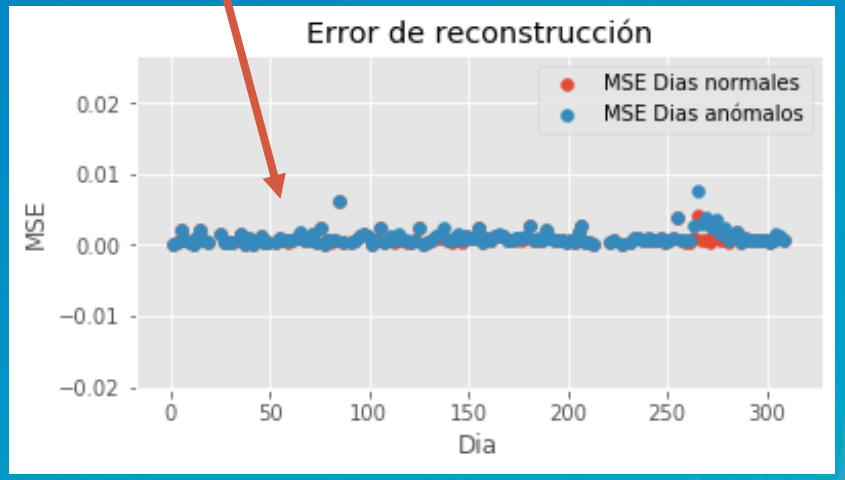
5.9 Procedimiento de autoencoder

- El método usa como indicador de normalidad el error de reconstrucción.
- Como no es diferenciador, usamos la red como método de predicción.
- Los residuos entre la señal reconstruida y la original, son relevantes.

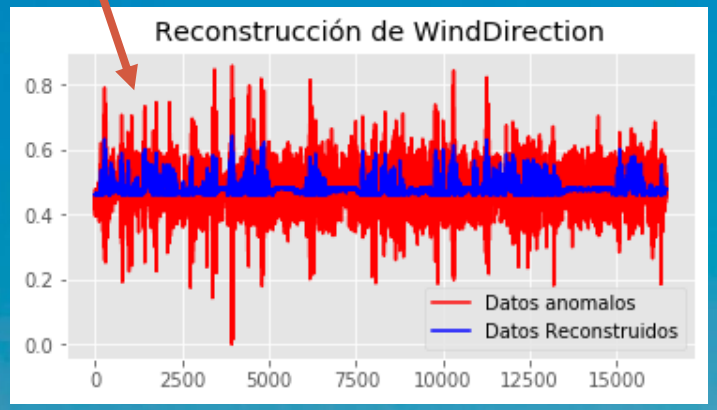


Autoencoder utilizado

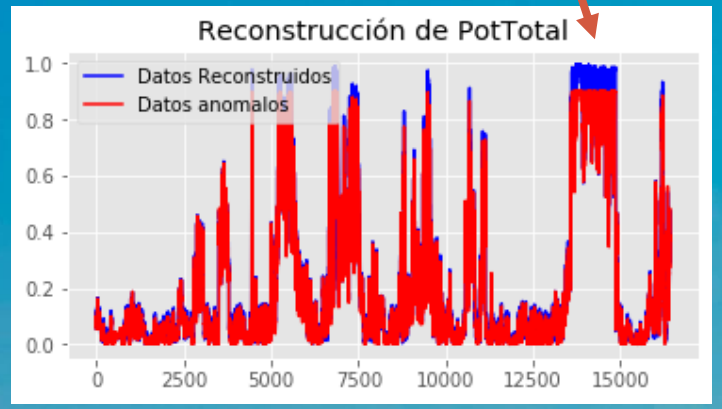
El error global de reconstrucción no es diferenciador

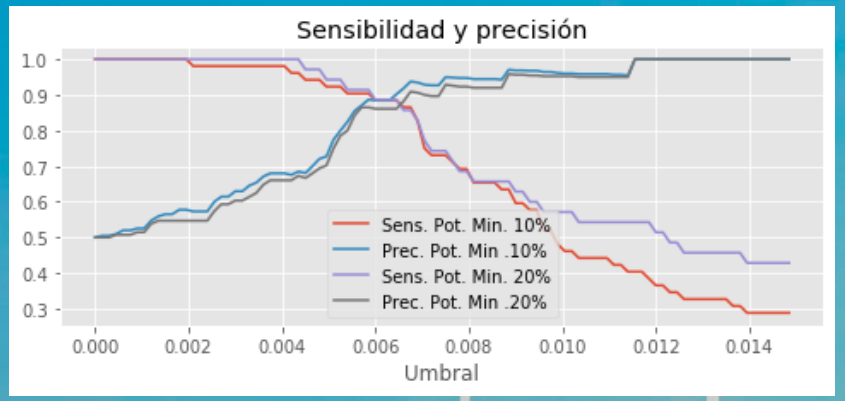


La red ha descartado los datos menos relevantes como la dirección del viento



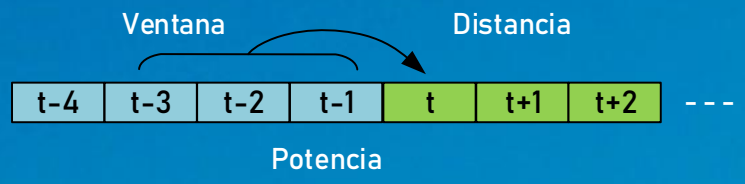
Los residuos de la potencia reconstruida si son diferenciadores.



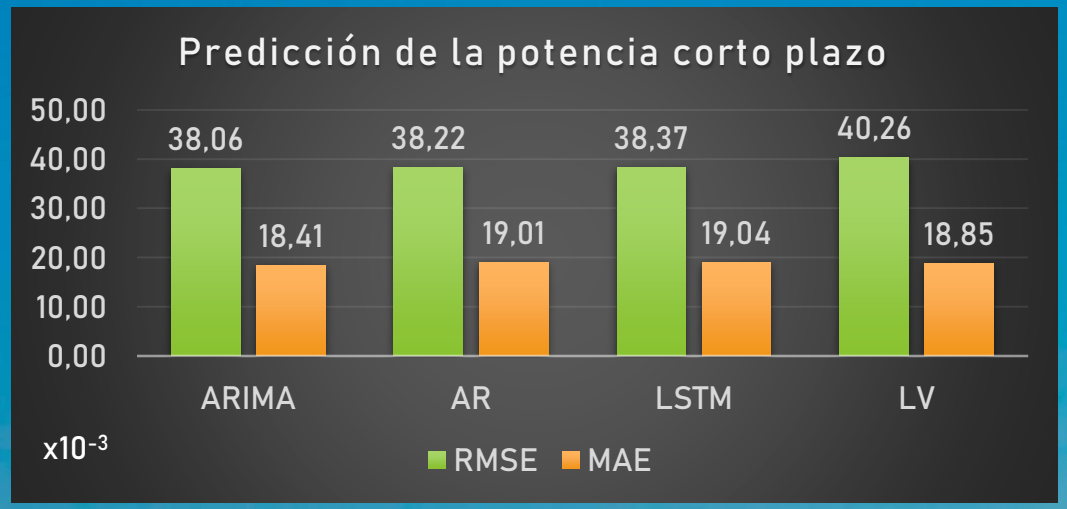


5.10 Predicción de energía a corto plazo

- Utilizamos solo los datos de potencia del pasado.
- Técnicas LSTM, ARIMA y autorregresión con RF, LR y SVR.
- Resultados muy similares usando los 30 y 60 últimos minutos para predecir los 10 siguientes.



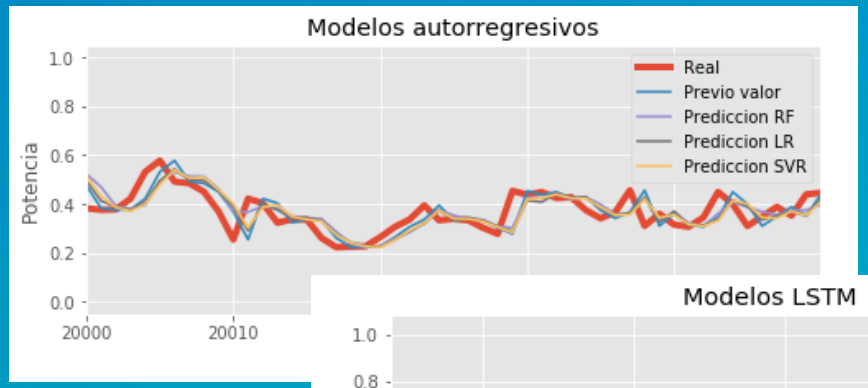
Resultados

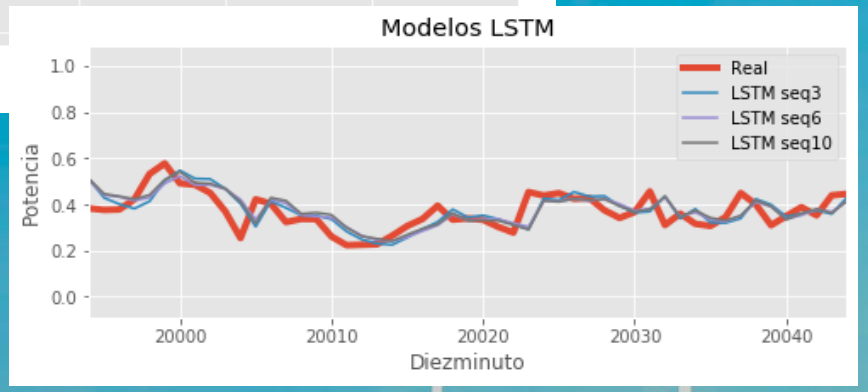


5.11 Predicción de medio plazo

Las simulación no se ha podido llevar a cabo:

- Los proveedores gratuito no ofrecen suficiente resolución.
- No podemos tener las predicciones del pasado por lo cual no podemos evaluar la precisión.





Predicciones a corto plazo

6. Conclusiones y trabajos futuros

6.1 Resultados

Se ha logrado con éxito:

- Adquirir el conocimiento necesario para la ejecución del trabajo.
- Analizar los datos del SCADA.
- Crear modelos efectivos de predicción de potencia y *pitch*.
- Simular y detectar anomalías con buena precisión.
- Predecir la potencia a corto plazo.

No se ha conseguido sin embargo:

- La predicción de energía a medio plazo.

Como conclusión:

- El uso de técnicas de minería de datos hace viable la detección de anomalías y por lo tanto la anticipación al fallo en nuestro conjunto de datos propietario.
- Es viable la predicción de la producción a corto plazo.

Detección de anomalías

100% Precisión
97% Sensibilidad

Perdida de potencia del 10% con una potencia media superior al 20%

Predicción de la producción

36W MAE
76W RMSE

Predicción de la potencia en el siguiente instante

6.2 Mejoras y trabajos futuros

- Obtención de mas y mejores datos de SCADA
- Validación con de datos de fallos reales
- Determinación de los tiempos de entrenamiento mínimos
- Crear un clasificador con sistema de *score*.
- Realizar estudios a nivel de parque
- Construcción de demostrador de CMS integrado para su uso en campo.

IGEP *Board*. Basado en arquitectura ARM

Uso de técnicas de minería de datos para la detección de anomalías y predicción de la producción en una turbina eólica.

Gracias por su atención