Universitat Oberta de Catalunya

uoc.edu

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

Javier Pérez Esteras
Máster Universitario de Ciencia de Datos
Área de minería de datos y aprendizaje automático

Raúl Parada Medina *(Profesor colaborador)*
Jordi Casas Roma *(Profesor responsable)*

Junio de 2020

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Uso de técnicas de minería de datos para la detección de anomalías y predicción de la producción en una turbina eólica*
Nombre del autor:	*Javier Pérez Esteras*
Nombre del consultor/a:	*Raúl Parada Medina*
Nombre del PRA:	*Jordi Casas Roma*
Fecha de entrega:	*06/2020*
Titulación:	*Máster Universitario de Ciencia de Datos*
Área del Trabajo Final:	*Minería de datos y machine learning*
Idioma del trabajo:	*Español*
Palabras clave	*Wind Power, Machine Learning, Preventive Maintenance, Energy prediction*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

En las últimas décadas se ha experimentado un importante incremento en la generación de electricidad procedente de energías renovables, especialmente la eólica, lo que convierte este sector en uno de especial interés.

Este trabajo valida el uso de algoritmos de minería de datos para la detección de anomalías en tareas de mantenimiento preventivo de turbinas, usando solamente las variables de control, sin necesidad de sistemas CMS específicos.

Se han aplicado y comparado algoritmos que determinan las relaciones entre parámetros durante el funcionamiento normal, como árboles aleatorios, potenciación de gradiente, máquinas de soporte vectorial para regresión y redes neuronales multicapa. Para ello hemos usado un histórico de datos propietario.

Se han simulado dos de los fallos más comunes, la pérdida de potencia y la desalineación. Se han aplicado técnicas de detección de anomalías basadas en métodos estadísticos y autoencoders. Los métodos estadísticos son los que mejor resultado han dado. Con una producción por encima del 20% nominal, son capaces de detectar una pérdida de potencia del 10% con una precisión del 100% y una sensibilidad del 97%.

También se han validado y comparado métodos de predicción de la potencia usando solamente los datos del pasado, como modelos auto regresivos, LSTM, y ARIMA, siendo estos últimos los que mejor resultado han dado.

Como conclusión, este estudio demuestra la viabilidad de usar técnicas de minería de datos, para la detección de anomalías y la anticipación al fallo en sistemas que no disponen de CMS específico, así como la predicción de producción usando datos del pasado.

Abstract (in English, 250 words or less):

In recent years, there has been a significant increase in the generation of electricity from renewable energies, especially wind power. This raises special interest on this sector.

The present work validates the use of data mining algorithms for anomaly detection and preventive maintenance tasks on turbines. It uses only SCADA control variables, it does not require specific CMS system.

The project has validated and compared algorithms determining relationships between parameters during regular operation. These algorithms included random trees, gradient boosting, support vector machines for regression, and multilayer neural networks. We have made use of a proprietary data log for these purposes.

Power loss and misalignment are common faults on wind turbines. We have simulated them while applying anomaly detection techniques. These were based on auto encoders and statistical methods, with the latter achieving better results. These techniques were able to detect a 10% power loss with energy production levels above 20% of the nominal value, performing with 100% precision level and showing 97% of sensitivity.

Power prediction based on historic series has also been validated and compared using different methods such as auto regression, LSTM, and ARIMA, giving this one the best result.

This work has proved and validated the feasibility in the use of data mining techniques for maintenance tasks in wind turbines. We anticipated failures by detecting anomalies in systems without specific CMS, and predicted production levels using historic data.

Índice

1 Introducción ... 1

1.1 Contexto y justificación del trabajo ... 1
1.2 Objetivos del Trabajo ... 4
1.3 Enfoque y método seguido ... 5
1.4 Planificación del Trabajo ... 6
1.5 Breve sumario de productos obtenidos ... 9
1.6 Breve descripción de los otros capítulos de la memoria ... 9

2 Estado del arte ... 10

2.1 La energía eólica ... 10
2.1.1 Funcionamiento general ... 10
2.1.2 Anatomía ... 11
2.1.3 Estrategia de control ... 14
2.1.4 Variables aportadas por el SCADA ... 15
2.1.5 Predicción de energía ... 16
2.1.6 El mantenimiento preventivo ... 17
2.1.7 Sistemas de monitorización ... 18
2.2 El aprendizaje automático ... 19
2.2.1 Detección de anomalías ... 20
2.2.2 Aplicación en aerogeneradores ... 25
2.2.3 Predicción de la producción ... 26
2.3 Trabajos realizados ... 27
2.3.1 Sector eólico ... 27
2.3.2 Mantenimiento preventivo y detección anomalías ... 28
2.3.3 Aplicación de modelos ... 29

3 Diseño e implementación ... 31

3.1 Selección de herramientas ... 31
3.1.1 Librerías usadas ... 31
3.2 Limpieza y análisis exploratorio de datos ... 32
3.2.1 Determinación de los datos de normalidad ... 33
3.2.2 Selección de características ... 34
3.3 Obtención de modelos ... 36
3.3.1 Métricas y sistema de evaluación ... 37
3.3.2 Normalización ... 38
3.3.3 Algoritmos de minería de datos ... 39
3.3.4 Búsqueda de hiperparámetros ... 40
3.3.5 Predicción de la potencia ... 40
3.3.6 Predicción del pitch ... 42
3.3.7 Tiempo de computación ... 44
3.4 Detección de anomalías ... 45
3.4.1 Simulación del fallo ... 46
3.4.2 Método a aplicar ... 46
3.4.3 Pérdidas en el tren de potencia ... 48
3.4.4 Desalineación del rotor ... 51
3.4.5 Uso de *autoencoder* para la detección ... 54
3.5 Predicción de energía ... 56

3.5.1 Aplicación de modelos para la previsión a corto plazo 56
3.5.2 Aplicación de modelos para la previsión a medio plazo..... 60

4 Conclusiones..... 62

5 Mejoras y trabajos futuros 64

6 Glosario..... 66

7 Bibliografía 67

8 Anexos 69

8.1 Anexo I. Variables de SCADA..... 69
8.1.1 Descripción gráfica por estado de máquina..... 69
8.1.2 Curva de potencia antes y después del filtrado 74
8.1.3 Correlación entre variables principales 75
8.2 Anexo II. Resultado de los modelos. 77
8.2.1 Predicción de la potencia..... 77
8.2.2 Predicción del pitch..... 83
8.3 Anexo III. Código fuente..... 88

Lista de figuras

Figura 1. Fotografía de parque eólico 1
Figura 2. Estructura de generación de la energía 1
Figura 3. Evolución de potencia eólica instalada..... 2
Figura 4. Página web mostrando la predicción de consumo. 3
Figura 5. Obtención de modelos de normalidad..... 4
Figura 6. Uso de los modelos para varios objetivos. 5
Figura 7. Cronograma de los trabajos 8
Figura 8. Incremento del área de rotor desde 1980 10
Figura 9. Elementos en la nacelle 12
Figura 10. Ángulo de desalineación 13
Figura 11. Curva ideal de potencia..... 14
Figura 12. Curva de potencia 15
Figura 13. Costes de mantenimiento..... 17
Figura 14. Métodos de predicción de fallo..... 18
Figura 15. Métodos de aprendizaje automático y algoritmos más usados 20
Figura 16. Distribución del error de predicción 24
Figura 17. Proceso de detección de anomalías 25
Figura 18. Distribución de las variables en condiciones de trabajo 34
Figura 19. Descripción del viento en función del estado 34
Figura 20. Correlación entre variables 35
Figura 21. Relaciones entre variables 36
Figura 22. Diagrama de partición de conjuntos 38
Figura 23. Resultados de la búsqueda de hiperparámetros de RF 40
Figura 24. Variables candidatas para predicción de potencia 41
Figura 25. Mejores resultados en la predicción de potencia 42
Figura 26. Ajuste y residuos en la predicción de potencia 42
Figura 27. Variables candidatas para la predicción de Pitch..... 43
Figura 28. Mejores resultados en la predicción de Pitch 44
Figura 29. Ajuste y residuos en la predicción de Pitch 44
Figura 30. Tiempos de computación de los modelos 45
Figura 31. Varianza en la relación viento-potencia..... 46
Figura 32. Potencia real y modificada 48
Figura 33. Residuos en la predicción de la potencia 49
Figura 34. Residuos en función de la potencia..... 49
Figura 35. Curvas de sensibilidad y precisión 50
Figura 36. Sensibilidad y precisión en la detección de perdida de potencia 50
Figura 37. Curva de potencia con conjunto normal y modificado 51
Figura 38. Residuos es la predicción de potencia 52
Figura 39. Residuos en función del viento y la potencia 52
Figura 40. Residuos promediados y umbral seleccionado 53
Figura 41. Sensibilidad y precisión en la detección de desalineación 53
Figura 42. Esquema de autoencoder para la detección de anomalías..... 54
Figura 43. Error de reconstrucción de los días normales y modificados 55
Figura 44. Reconstrucción de la dirección y la potencia 55
Figura 45. Sensibilidad y precisión obtenidos por autoencoder 55
Figura 46. Modo de operación de una auto regresión..... 57

Figura 47. Predicción de la potencia como serie temporal 57
Figura 48. Construcción de secuencias para la red LSTM 58
Figura 49. Predicción de la potencia con LSTM 58
Figura 50. Potencia normalizada y estacionaria 59
Figura 51. Comparativa de modelos ordenados por MAE y RMSE 60

Lista de tablas

Tabla 1. Paquetes de trabajo 6
Tabla 2. Variables habituales de los SCADA 15
Tabla 3. Modelos de predicción habituales 16
Tabla 4. Técnicas empleadas 30
Tabla 5. Variables del conjunto de datos 32
Tabla 6. Principales estadísticos 33
Tabla 7. Modelos para la predicción de potencia 41
Tabla 8. Ajuste de los modelos de potencia 41
Tabla 9. Modelos para la predicción del Pitch 43
Tabla 10. Ajuste de los modelos de pitch 43
Tabla 11. Sensibilidad y precisión en la detección de perdida de potencia 51
Tabla 12. Sensibilidad y precisión en la detección de desalineación 54
Tabla 13. Resultados de los modelos de auto regresión 57
Tabla 14. Resultados de los modelos LSTM 58
Tabla 15. Resultados de los modelos ARIMA 59

1 Introducción

1.1 Contexto y justificación del trabajo

Una de las energías más limpia e inagotable, es la eólica, que obtiene electricidad gracias a la fuerza del viento. Entre las ventajas de la energía eólica destaca que no contamina, es inagotable y está disponible prácticamente en la todo del planeta. Además, contribuye a un modelo de desarrollo sostenible.

Según la Asociación Empresarial Eólica [1] actualmente en España hay más de 1.100 parques eólicos como el que se observa en la Figura 1, que suman casi 24.000 MW de potencia eléctrica instalada.

Figura 1. Fotografía de parque eólico Autor: J. Pérez

En 2018, según el informe de Red Eléctrica [2], con fecha 26 de junio de 2019, la potencia eléctrica instalada procedente de renovables alcanzó el 46.7%, correspondiendo a la eólica un 22.6%. En cuanto a la producción, la energía eólica supuso el 19% del total. La estructura de generación de energía en 2018 puede observarse en la Figura 2.

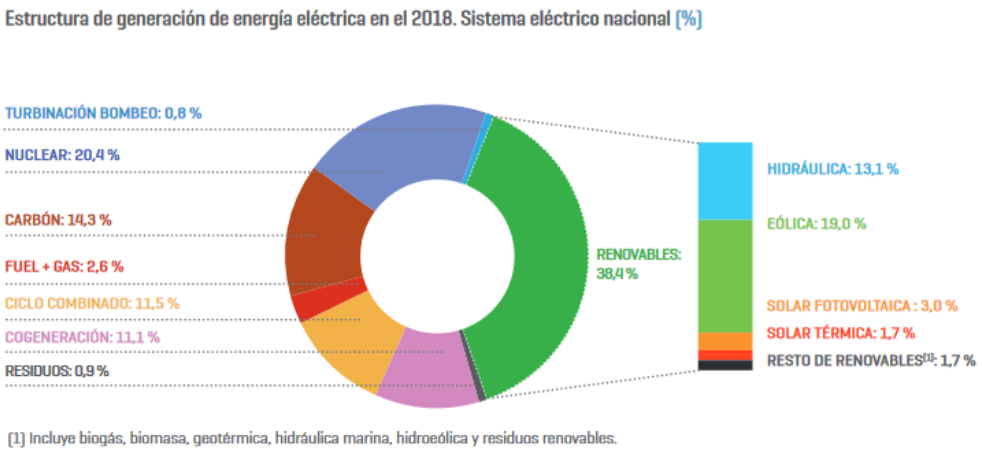


Figura 2. Estructura de generación de la energía Fuente: Red Eléctrica [2]

En cuanto a la evolución a lo largo de los años, este tipo de energía experimentó un importante auge en la década del 2.000 como se observa en la Figura 3, hasta estabilizarse en 2012, debido mayormente a la crisis económica y los cambios regulatorios, que supusieron la supresión de ayudas y primas a la producción, conocidos como “impuesto al sol” y “moratoria verde” [3]

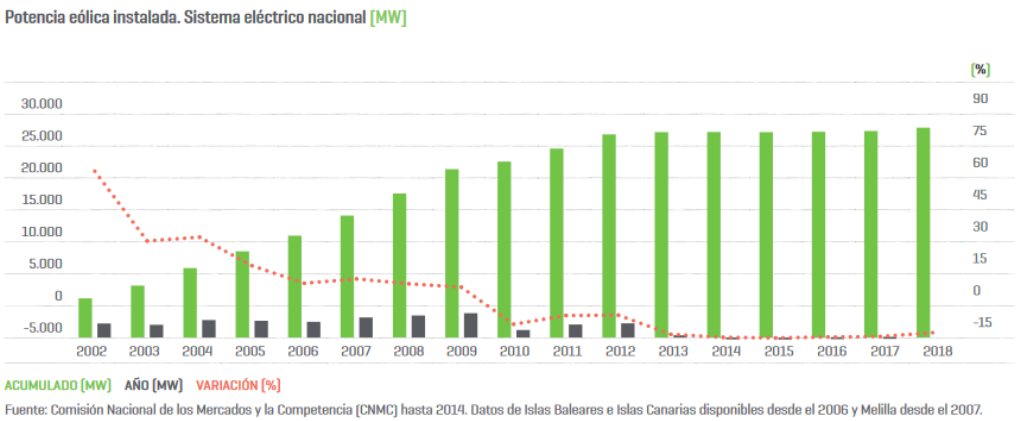


Figura 3. Evolución de potencia eólica instalada.
Fuente: Red Eléctrica de España [2]

Las cifras ponen de manifiesto la importancia de este sector, que contribuye a la transición hacia un modelo sostenible y respetuoso con el medio ambiente que ponga fin a la dependencia de combustibles fósiles. Además, resulta clave para cumplir los objetivos europeos (Directiva 2009/28, que establece que al menos un 20% del consumo final bruto de energía proceda de fuentes renovables en 2020. [4]

Según informa Siemens-Gamesa [5], uno de los principales fabricantes de turbinas a nivel mundial, se espera que la demanda de la electricidad siga aumentando a nivel global, y la eólica *offshore* (en el mar) se posiciona como la respuesta a la búsqueda de nuevas fuentes de energía. Ésta, tiene como ventaja la gran cantidad de espacio disponible y el aumento uniforme de los recursos eólicos, pero por contrapartida tiene unos mayores costes de mantenimiento e intervención, que gracias al mantenimiento preventivo y a las técnicas de *big data*, pueden ser reducidos.

Un aerogenerador es sin duda una increíble pieza de ingeniería que contribuye a la reducción de gases de efecto invernadero y preserva el medio ambiente, además de contar con un futuro prometedor. Esto junto con la posibilidad de trabajar con datos reales obtenidos directamente de máquinas, han despertado un interés por estudiar la aplicación de algoritmos de minería de datos en este sector.

Actualmente algunos fabricantes como Siemens-Gamesa [5] ya incorporan CMS (*Condition Monitoring System*) o sistemas de monitorización que recopilan la información de cientos de sensores específicos que contribuyen a supervisar el estado y funcionamiento de la turbina con el objetivo de ampliar la vida útil y optimizar la producción. Por otro lado, hay numerosos estudios que modelizan la curva de potencia, (relación viento-potencia) mediante modelos de regresión, aunque en ocasiones se recurre a juegos de datos sintéticos, ya que el sector eólico no es partidario de compartir datos reales de forma pública.

En este trabajo tendremos la oportunidad de trabajar con datos reales y procedentes de turbinas sin sensores específicos, es decir solo con los parámetros mínimos que la máquina supervisa para su funcionamiento, con el objetivo de encontrar modelos que sean de aplicación para mantenimiento preventivo, principalmente la detección de anomalías.

La utilización de datos únicamente del SCADA o PLC, con variables únicamente de control, hace más interesante la propuesta que si se utilizan datos de sensores específicos, ya que abre la posibilidad de realizar un mantenimiento preventivo a máquinas que no estaban preparadas para ello con hardware adicional.

Un segundo uso de estos modelos puede ser la predicción de la producción de energía a corto y medio plazo: El mayor problema de la energía eléctrica es que no se puede almacenar, de modo de la producción y la demanda debe ir coordinada. Existe un complejo sistema a través del cual los productores y las comercializadores compran y venden energía en diferentes mercados y precios. De ahí la necesidad de contar con un buen sistema de predicción. La Figura 4 ilustra un visualizador online que Red Eléctrica de España [2] tiene a disposición en su página web (www.ree.es/es/datos/aldia) y que permite observar la demanda prevista, programada y real.

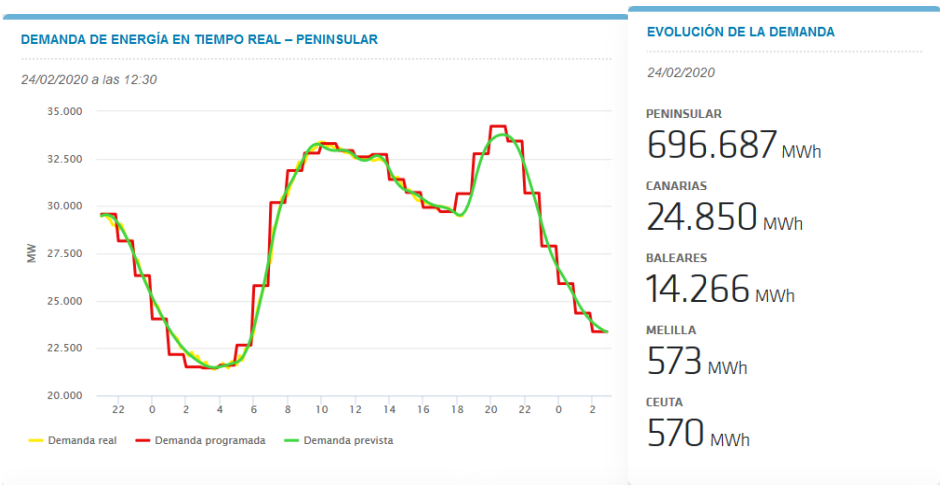


Figura 4. Página web mostrando la predicción de consumo.
Fuente: Página web de Red Eléctrica [2]

1.2 Objetivos del Trabajo

A partir de un histórico de datos recibido de un SCADA se plantean dos objetivos principales.

- Validar el uso de técnicas de *machine learning* para un conjunto de datos específico, que permita la detección de anomalías y la anticipación al fallo, pudiendo así incrementar la disponibilidad y reducir de los tiempos de parada.
- Validar la capacidad predictiva de la producción utilizando los datos de viento e históricos para una gestión más eficiente de los recursos de la red.

Para ello será necesaria la consecución de objetivos secundarios como:

- Aprender el funcionamiento y modos de trabajo de los aerogeneradores.
- Aprender las técnicas *machine learning* usadas para mantenimiento preventivo.
- Descubrir los modelos de minería de datos más idóneos para nuestros objetivos.
- La aplicación de varios modelos con diferentes técnicas de regresión que describan el funcionamiento correcto de una turbina, y que permitan efectuar predicciones con datos futuros.
- La aplicación de los modelos a los datos del SCADA para detectar desviaciones sobre la normalidad.
- La adquisición de la predicción meteorológica desde proveedores de internet.

Las Figura 5 y Figura 6 muestran el proceso de obtención de los modelos de normalidad y cómo posteriormente, a partir de ellos es posible la detección de anomalías y la predicción de la producción.

Figura 5. Obtención de modelos de normalidad.

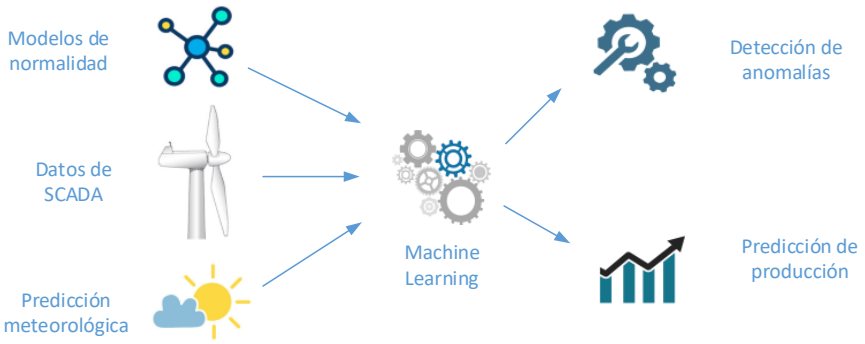


Figura 6. Uso de los modelos para varios objetivos.

1.3 Enfoque y método seguido

La metodología para llevar a cabo el trabajo se llevará a cabo siguiendo las siguientes fases o etapas:

- Como punto de partida se localizará y se estudiará el funcionamiento general de una turbina. Este punto es fundamental para poder realizar una correcta interpretación del histórico de datos con el que contaremos en las siguientes fases.
- Se realizará un estado del arte de las técnicas de modelado actuales y su uso para mantenimiento preventivo y predicción.
- Se realizará la selección de herramientas, entornos de desarrollo y librerías a utilizar.
- Se buscarán proveedores meteorológicos a través de los cuales poder acceder a una predicción de vientos en una ubicación específica y se analizará la técnica necesaria para su recogida.
- Posteriormente se realizará una exploración de los datos recibidos, donde se analizarán, limpiarán, graficarán y se identificarán las características más significativas.
- Se estudiarán diferentes modelos de regresión y sus características y se aplicarán a los datos con el objetivo de modelar el comportamiento en estado funcional. Lo modelos que lo permitan serán entrenados con técnicas de búsqueda óptima de hiperparámetros y se evaluará la bondad de ajuste a través de varias métricas, realizando una comparativa entres modelos. Para evitar la sobre especialización se utilizará validación cruzada.

Una vez se disponga de los modelos se intentará aplicarlos con diferentes objetivos:

- Detección de anomalías, lo que supone anticiparse a la aparición de un problema. Para esto probablemente será necesaria la simulación del fallo, es decir, la alteración de un conjunto de datos de test de forma controlada que permita analizar el desajuste de los modelos de normalidad frente a estos datos alterados.

- Predicción de energía a corto plazo, lo que supone prever la producción a minutos vista, teniendo como datos de entrada la producción obtenida a tiempo pasado.
- Predicción de energía a medio plazo, lo que supone prever la producción a días vista, teniendo como única entrada la predicción de vientos de servicios meteorológicos externos.

1.4 Planificación del Trabajo

Para llevar a cabo con éxito la consecución de los objetivos anteriores se expone una propuesta de planificación que engloba los paquetes de trabajo de la Tabla 1.

Paquete de trabajo	Días	F. Inicio	F. Fin
WP1. Definición y planificación del trabajo	10	19/02/20	01/03/20
WP2. Búsqueda de información y estado del arte	20	02/03/20	22/03/20
WP3. Diseño e implementación	60	23/03/20	23/05/20
WP4. Redacción de la memoria	15	24/05/20	10/06/20
WP5. Elaboración de la presentación	6	11/06/20	16/06/20

Tabla 1. Paquetes de trabajo

Entrando más en detalle, establecemos tareas en cada uno de los paquetes que finalizarán con el hito de la entrega de las PEC.

WP1. Definición y planificación del trabajo	
Objetivos	Definir la temática del trabajo justificando el interés. Definir la planificación del trabajo
Requerimientos	Acordar el alcance con el tutor del trabajo.
Tareas	Búsqueda de información Contexto y justificación Establecimiento de objetivos Enfoque y método Planificación Abstract
Resultado	Un entregable la especificación del trabajo a realizar (PEC1) El conocimiento necesario para emprender la búsqueda de información y estado del arte.

WP2. Búsqueda de información y estado del arte	
Objetivos	Documentarse sobre los estudios actuales en la temática elegida. Adquirir el conocimiento necesario para abordar el trabajo
Requerimientos	Haber definido unos objetivos
Tareas	Funcionamiento de aerogenerador Mantenimiento preventivo Técnicas de regresión Técnicas de predicción de series temporales Servicios meteorológicos Trabajos existentes
Resultado	Un entregable con los resultados de la investigación (PEC2) El refinamiento de los objetivos planteados en la fase anterior.

WP3. Diseño e implementación	
Objetivos	La realización de la tareas definidas en las fases anteriores
Requerimientos	Contar con los conocimiento adquiridos del punto anterior. Contar con el juego de datos necesario
Tareas	Selección de herramientas Limpieza y análisis exploratorio de datos Obtención y comparativa de modelos Síntesis del fallo Aplicación de los modelos para la detección de la anomalías Obtención de la previsión meteorológica Aplicación de modelos para la previsión a corto plazo Aplicación de modelos para la previsión a medio plazo
Resultado	Una o varias entregas parciales con los avances en la implementación. (PEC3)

WP4. Redacción de la memoria	
Objetivos	Documentar y justificar el desarrollo y conclusiones del trabajo. Integrar toda la documentación en un único documento.
Requerimientos	Finalizar el trabajo y haber llegado a unas conclusiones
Tareas	Redacción de los resultados del trabajo Integración de toda la documentación en formato oficial.
Resultado	La memoria del trabajo

WP5. Elaboración de la presentación	
Objetivos	Preparar una exposición pública y oral del trabajo
Requerimientos	Haber finalizado la memoria
Tareas	Elaboración de la presentación Trabajo de la exposición oral Grabación del video-presentación
Resultado	Transmitir la relevancia y resultados del trabajo.

En cuanto a recursos materiales se precisa:

- Histórico de datos del SCADA de al menos una turbina eólica.
- Equipo informático con capacidad suficiente para cálculo de modelos acorde al conjunto de datos.
- Software y librerías que serán determinadas más adelante.

En la Figura 7 se ilustra el cronograma o secuenciación en tiempo de las tareas descritas.

Id	Nombre de tarea	Duración	Comienzo	Fin
1	**WP1. Definición y planificación del trabajo**	**11 días**	**mié 19/02/20**	**dom 01/03/20**
2	Búsqueda de información	3 días	mié 19/02/20	sáb 22/02/20
3	Contexto y justificación	3 días	sáb 22/02/20	mar 25/02/20
4	Establecimiento de objetivos	3 días	mar 25/02/20	vie 28/02/20
5	Enfoque y método	2 días	mar 25/02/20	jue 27/02/20
6	Planificación	1 día	vie 28/02/20	sáb 29/02/20
7	Abstract	1 día	sáb 29/02/20	dom 01/03/20
8	Entrega PEC1	0 días	dom 01/03/20	dom 01/03/20
9	**WP2. Búsqueda de información y estado del arte**	**20 días**	**lun 02/03/20**	**dom 22/03/20**
10	Funcionamiento de aerogenerador	4 días	lun 02/03/20	vie 06/03/20
11	Mantenimiento preventivo	7 días	mié 04/03/20	mié 11/03/20
12	Servicios meteorológicos	3 días	vie 06/03/20	lun 09/03/20
13	Técnicas de regresión	7 días	dom 08/03/20	dom 15/03/20
14	Técnicas de predicción de series temporales	7 días	mié 11/03/20	mié 18/03/20
15	Trabajos existentes	7 días	dom 15/03/20	dom 22/03/20
16	Entrega PEC2	0 días	dom 22/03/20	dom 22/03/20
17	**WP3. Diseño e implementación**	**1,88 días**	**lun 23/03/20**	**sáb 23/05/20**
18	Selección de herramientas	2 días	lun 23/03/20	mié 25/03/20
19	Limpieza y análisis exploratorio de datos	7 días	mié 25/03/20	mié 01/04/20
20	Obtención y comparativa de modelos	13 días	mié 01/04/20	mar 14/04/20
21	Síntesis del fallo	5 días	mar 14/04/20	dom 19/04/20
22	Aplicación de los modelos para la detección de la anomalía	15 días	dom 19/04/20	lun 04/05/20
23	Obtención de la previsión meteorológica	7 días	dom 03/05/20	dom 10/05/20
24	Aplicación de modelos para la previsión a corto plazo	7 días	dom 10/05/20	dom 17/05/20
25	Aplicación de modelos para la previsión a medio plazo	6 días	lun 18/05/20	sáb 23/05/20
26	Entrega PEC3	0 días	sáb 23/05/20	sáb 23/05/20
27	**WP4. Redacción de la memoria**	**17 días**	**lun 25/05/20**	**mié 10/06/20**
28	Redacción de los resultados del trabajo	11 días	lun 25/05/20	jue 04/06/20
29	Integración de toda la documentación en formato oficial.	6 días	vie 05/06/20	mié 10/06/20
30	Entrega PEC4	0 días	mié 10/06/20	mié 10/06/20
31	**WP5. Elaboración de la presentación**	**6 días**	**jue 11/06/20**	**mar 16/06/20**
32	Elaboración de la presentación	3 días	jue 11/06/20	sáb 13/06/20
33	Trabajo de la exposición oral	3 días	dom 14/06/20	mar 16/06/20
34	Grabación del video-presentación	3 días	dom 14/06/20	mar 16/06/20
35	Entrega PEC5	0 días	mar 16/06/20	mar 16/06/20
36	Defensa Proyecto	0 días	mié 24/06/20	mié 24/06/20

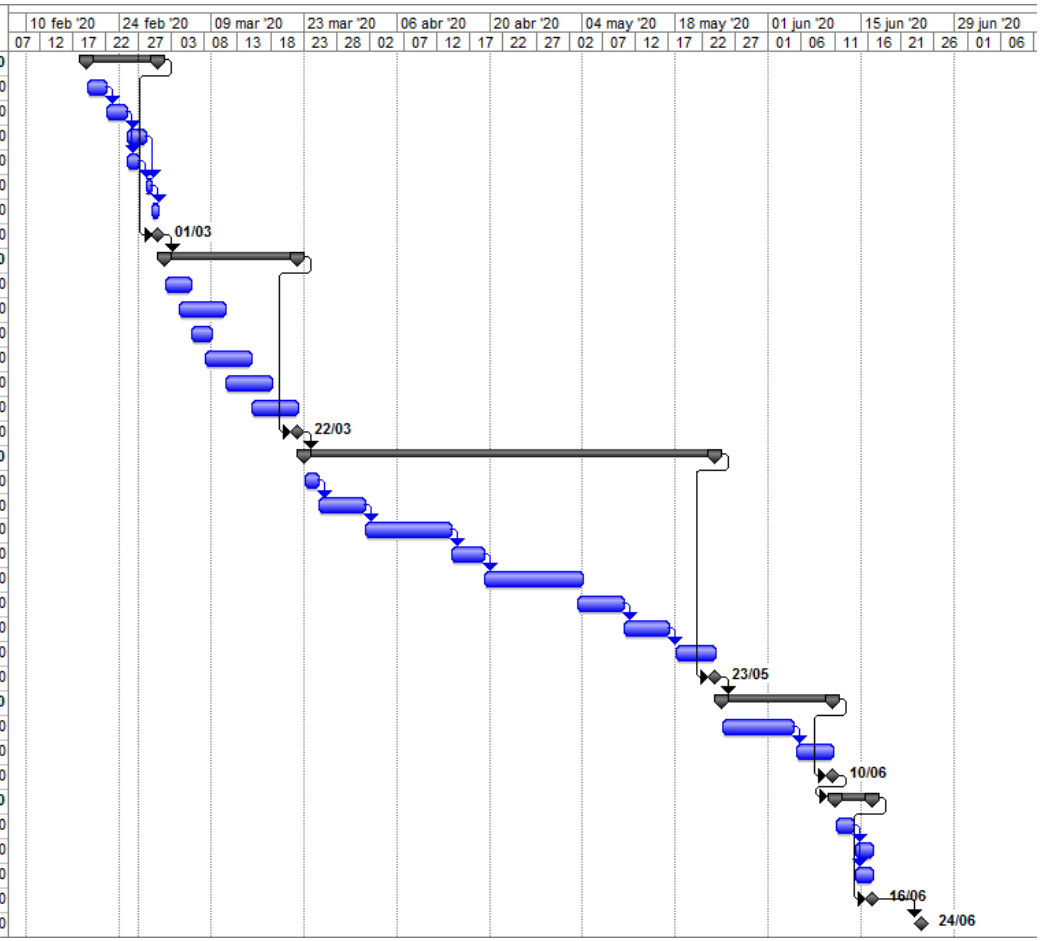


Figura 7. Cronograma de los trabajos

1.5 Breve sumario de productos obtenidos

Como resultado del trabajo se obtendrán los siguientes productos:

- Una serie de modelos que describan el funcionamiento normal de una turbina y una comparación de sus ajustes.
- La construcción de un sistema clasificador, utilizando varias técnicas de detección de anomalías, de modo que sea capaz de diferenciar un día normal de un día anómalo.
- La validación y comparativa de diferentes técnicas de predicción de energía a corto y medio plazo.
- Una valoración sobre el posible uso de estas técnicas.

1.6 Breve descripción de los otros capítulos de la memoria

El capítulo 2 contiene un estado del arte, fruto de la búsqueda y recopilación de información, donde se puede diferenciar los apartados:

- Energía eólica. Se exponen los conceptos básicos sobre los tipos de aerogeneradores, las partes que los componen, y sus estrategias de control. También se introduce la importancia del mantenimiento preventivo y los sistemas de monitorización
- Aprendizaje automático. Se explican las diferentes técnicas que existen, se define el concepto de anomalía y los tipos que existen, así como los métodos más utilizados para su detección.
- Trabajos existentes. Se expone un resumen de la información recopilada y cuáles son las líneas generales de investigación, objetivos planteados, algoritmos, técnicas empleadas, resultados, etc.

El capítulo 3 contiene el desarrollo y la implementación, tras comenzar con la selección de herramientas se desglosa en los siguientes apartados:

- Análisis exploratorio, donde se analizan y limpian los datos originales de entrada. Se localizan las características relevantes y los periodos donde la maquina permanece en régimen de producción.
- Obtención de modelos. Se buscan características que junto con diferentes algoritmos sean capaces de predecir a otras, usando técnicas como selección de hiperparametros y validación cruzada.
- Detección de anomalías. Se utilizan métodos estadísticos y redes encoder-decoder para la identificación de días con diferentes fallos simulados.
- Predicción de energía. Se calcula la precisión de algoritmos como LSTM, ARIMA y métodos auto regresivos para la estimación de la potencia en función de los datos del pasado.

Finalmente, los capítulos 4 y 5, contienen las conclusiones y los trabajos futuros, respectivamente.

2 Estado del arte

2.1 La energía eólica

La energía eólica consiste en convertir la energía del viento en electricidad a través de un generador que puede ser impulsado de diferentes formas, como puede ser una hélice. Existen turbinas o aerogeneradores domésticos, con una potencia de varios kilovatios, y grandes construcciones, de varios megavatios, que suelen agruparse en forma de parque eólico. En la última década hemos visto como las dimensiones de los aerogeneradores ha incrementado considerablemente a la par que su capacidad para la producción de energía, según lo detalla la Agencia Danesa de Energía [6] en la Figura 8.

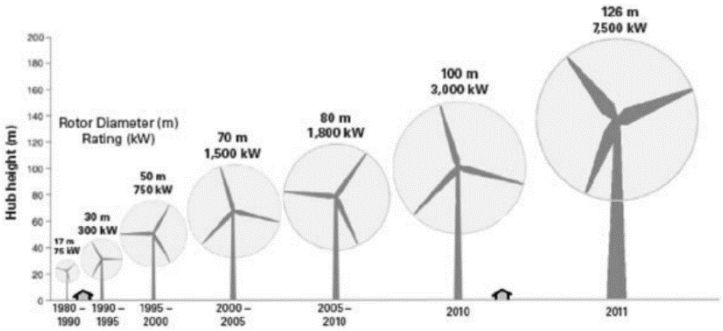


Figura 8. Incremento del área de rotor desde 1980 Fuente: [6]

Entender en funcionamiento de estas turbinas es importante para interpretar de forma correcta las variables recibidas del SCADA, sobre todo de cara a selección de características en la construcción de modelos. En los siguientes puntos se aporta un resumen sobre los resultados de la investigación al respecto.

2.1.1 Funcionamiento general

Para la elaboración de este trabajo nos centraremos en una turbina de eje horizontal, con una potencia nominal de 2 MW, uno de los modelos más habituales, y que consta de un rotor ubicado en una góndola (*nacelle*) en lo alto de una torre con dos o tres palas.

Con el objetivo de capturar la mayor energía posible del viento, el aerogenerador se orienta de forma automática gracias a un mecanismo de giro denominado *yaw,* [7] que se comanda en base a las lecturas de la veleta y el anemómetro que se encuentran en la parte superior y que son piezas claves como veremos más adelante en la anatomía.

Cuando el viento sopla con una fuerza superior a un umbral que se sitúa en torno a los 4 m/s, las palas giran y aerogenerador entra en situación de poder generar energía, no obstante, también hay un límite máximo de

viento, a partir del cual no son capaces de funcionar debido a su diseño aerodinámico, optimizado para los vientos más frecuentes. A partir de 25 m/s los aerogeneradores entran en modo bandera, es decir se detienen por seguridad y solo se orientan evitando ofrecer resistencia al viento. Estas situaciones no son frecuentes y se trata de evitar daños mayores como la ruptura de partes o la caída al suelo.

En la parte superior, la *nacelle,* es donde se encuentra la mayor parte de la maquinaria, como el tren de potencia, donde se ubica la multiplicadora (una especie de caja de cambios que eleva las revoluciones del eje de baja, o rotor, que se encuentra conectado a las palas) para incrementarlas hasta alcanzar las revoluciones necesarias para accionar el generador.

También se ubica como elemento final el transformador, encargado de elevar la tensión de salida del generador (típicamente 690v) a la adecuada para evacuar la energía hasta la subestación más cercana, donde aún es elevada más e inyectada a la red de distribución.

2.1.2 Anatomía

Las principales partes de las que consta un aerogenerador de eje horizontal, el más extendido en la geografía española son:

- Cimentación. El punto de anclaje al terreno. Sus dimensiones son acordes al tamaño del aerogenerador y en ocasiones puede estar monitorizada para comprobar su estabilidad. En las versiones *offshore*, es sustituida por diferentes sistemas, como estructuras flotantes o pilotes.
- Torre. Soporta la *nacelle*, que es la parte donde se encuentra la mayoría de elementos. Su altura está condicionada por el terreno y por los estudios de vientos. Normalmente a mayor altura, mayor intensidad de viento. En los grandes aerogeneradores las torres tubulares pueden ser de acero, de celosía o de hormigón.
- Nacelle. En lo más alto de torre, se trata de la estructura que tiene capacidad de girar sobre el eje Z para adaptarse a la dirección del viento. Alberga el rotor con las palas, tren de potencia, y transformador.

En de la *nacelle*, se encuentran ubicados los sistemas que pueden observase en la Figura 9.

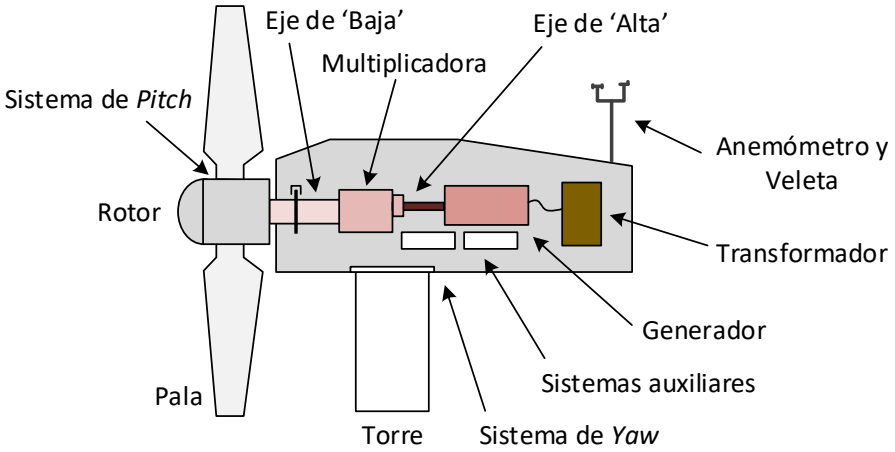


Figura 9. Elementos en la nacelle Elaboración propia.

- Veleta y anemómetros. Su cometido es capturar la dirección y la intensidad del viento. Son medidas claves que suelen ser redundantes, es decir, son adquiridas por duplicado y con sensores de diferentes tecnologías, como pueden ser los sistemas mecánicos tradicionales (cazoletas y veleta) y ultrasónicos, que detectan dirección y velocidad sin piezas móviles. Algunos fabricantes comienzan a emplear tecnologías más novedosas, como la medición por un haz láser, que tiene la ventaja de poder medir el viento por delante del rotor, evitando las interferencias que causa éste.
- Rotor: En la parte externa de la *nacelle* hace de interfaz entre las palas y el eje de baja. Dispone de un mecanismo de giro de las palas denominado *pitch*, normalmente hidráulico, que permite modificar el ángulo de ataque con respecto al viento. La potencia nominal de un aerogenerador aumenta con el área de barrido del rotor que es proporcional a la longitud de las palas. Un aerogenerador de 2MW suele tener un rotor de unos 80 metros de diámetro.
- Eje de baja. Denominado así por ser el eje que gira a pocas revoluciones y conecta el rotor con la multiplicadora. En la mayoría de aerogeneradores gira en torno a las 15 revoluciones por minuto.
- Multiplicadora: Es una caja de cambios de conecta el eje de baja con el eje de alta. Su cometido es elevar las revoluciones multiplicándolas por un factor. Es una pieza sometida a grandes esfuerzos y desgaste.
- Eje de alta. Conecta la multiplicadora con el generador. Suele alcanzar unas 15.000 revoluciones por minuto, necesarias para excitar el generador.
- Freno. Bloquea el tren de potencia. Se trata de un mecanismo hidráulico que se acciona cuando la maquina se encuentra fuera de servicio o bien ante una situación de parada de emergencia.
- Generador. Normalmente un generador asíncrono, trifásico de 690V.
- Transformador. Eleva la tensión de salida del generador para transportarla de forma más eficiente hasta la subestación más cercana, normalmente entre 10.000 y 30.000 V.

- Mecanismo de giro. Denominado *yaw*. Permite orientar la nacelle en la dirección del viento para maximizar la captura de energía. Suele estar comandado por motores eléctricos.
- Grupo hidráulico. Sistema de apoyo para al accionamiento de algunos mecanismos como el *pitch* y el freno.

2.1.2.1 Sistema de pitch

El *pitch* es un mecanismo de giro de las palas, que permite modificar el ángulo de ataque con respecto al viento consiguiendo ajustar la resistencia aerodinámica para aprovechar al máximo diferentes velocidades de viento. De este modo es posible generar la máxima potencia de energía tanto a altas como a bajas velocidades sin poner en riesgo la integridad estructural.

El ángulo de *pitch* se mide en grados, y existen unos valores para los cuales, el empuje de viento es máximo, nulo o variable en función de los grados. Este es un valor fundamental de control como veremos más adelante.

2.1.2.2 Sistema de yaw

Se denomina así al mecanismo de orientación que hace girar la *nacelle* para mantener el rotor enfrentado al viento, normalmente mediante motores eléctricos y reductoras.

Cuando el rotor no está perpendicular al viento, se puede considerar un error de orientación, o desalineación, teniendo varias consecuencias, como que la cantidad de energía del viento que pasa a través del área rotor es inferior, y que las palas están sometidas a mayores cargas de fatiga que si se encuentra completamente perpendicular al viento. Por consiguiente, los sistemas de control buscaran mantener siempre el rotor enfrentado al viento para incrementar la generación de energía.
Este ángulo también es importante en los sistemas de control y se detalla en la Figura 10.

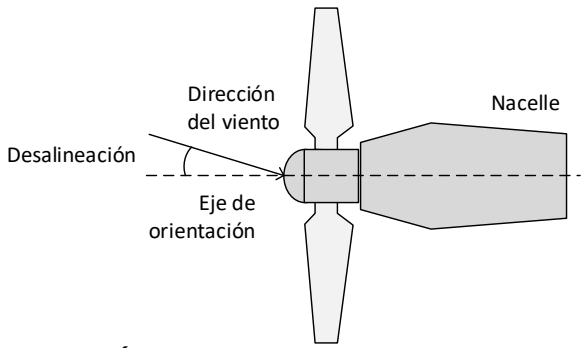


Figura 10. Ángulo de desalineación Elaboración propia.

2.1.3 Estrategia de control

El objetivo de la estrategia es maximizar la generación de energía de acuerdo a la Figura 11, que ilustra la curva ideal de potencia según la describe [8].

Existen 4 zonas de vientos, delimitadas por tres valores clave:

- Valor de corte inferior. Es el valor de viento mínimo necesario para comenzar la producción de energía. Típicamente 3 o 4 metros por segundo. (Vmin)
- Valor de corte superior. Es el valor de viento máximo a partir de la cual no es posible generar energía debido a la aparición de riesgos estructurales. Alrededor de 25 metros por segundo. (Vmax)
- Valor de viento nominal. Separa las zonas de trabajo de carga parcial y carga total. Por encima de este valor la producción de energía es constante e independiente de la velocidad de viento. Por debajo de este valor, la generación se incrementa progresivamente con el viento hasta alcanzar la producción máxima. Se encuentra entorno a los 12 metros por segundo. (Vnom)

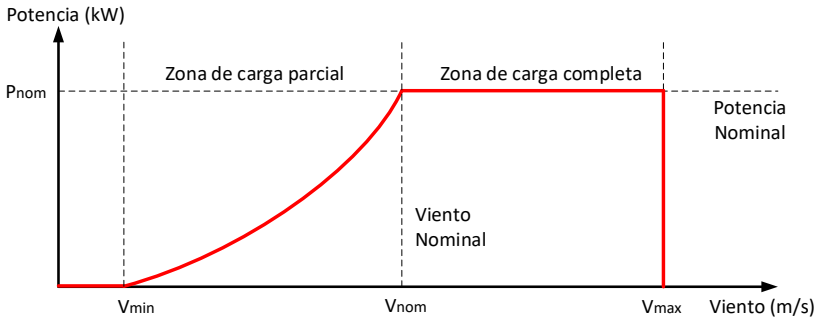


Figura 11. Curva ideal de potencia Elaboración propia.

En cada una de las zonas se plantea un objetivo distinto para el control:

- Maximizar la generación de energía cuando el viento está por debajo del valor nominal, mientras que cuando se encuentra por encima es minimizar la carga mecánica al mismo tiempo que se mantiene la generación de energía máxima.
- A lo largo de zona de carga completa el sistema de control busca las revoluciones óptimas del generador, regulando el ángulo de *pitch* y manteniendo constante el par del generador.
- En la zona de trabajo de carga parcial, el sistema utiliza el *pitch* optimo con máximo empuje y regula el par del generador para conseguir la relación optima entre viento y revoluciones.

El sistema de control alterna entre los modos trabajo de forma automática usando como parámetros principales de regulación el *pitch* y el par del generador. El par afecta a carga del tren de potencia mientras que el *pitch* afecta a la carga estructural.

Otros trabajos [9] hacen referencia a sistemas de control más sofisticados que incluyen limitaciones de potencia, para evitar la emisión de ruido ambiente o ajustarse a la demanda de la red, en cuyo casos la regulación persigue el objetivo de ajustarse a una determinada producción.

2.1.4 Variables aportadas por el SCADA

En la Tabla 2 podemos observar las variables más importantes aportadas por los SCADA, según describe Meik Schlechtingen [10] y pertenecientes a un aerogenerador con multiplicadora de 2MW de potencia nominal. También se pueden observar los valores habituales de funcionamiento en estado sin defectos.

Variable	Rango normal
Angulo de pitch	-5º a 90º
Velocidad de viento	0 a 35 m/s
Revoluciones del rotor	0 a 16 rpm
Revoluciones del generador	0 a 1700 rpm
Potencia de salida del generador	−50 to 2100 kW
Potencia reactiva del generador	−25 to 300 kWAr
Temperatura de la nacelle	5 a 50 ºC
Temperatura aceite de la multiplicadora	30 a 35 ºC
Temperatura aceite del grupo hidráulico	30 a 80 ºC
Presión del grupo hidráulico	170 a 190 hPa

Tabla 2. Variables habituales de los SCADA

Lo más significativo son las relaciones entre estas variables, sobre todo la curva de potencia o relación viento-potencia, ilustrada en la Figura 12, que es muy usada para el análisis de rendimiento y detección de anomalías.

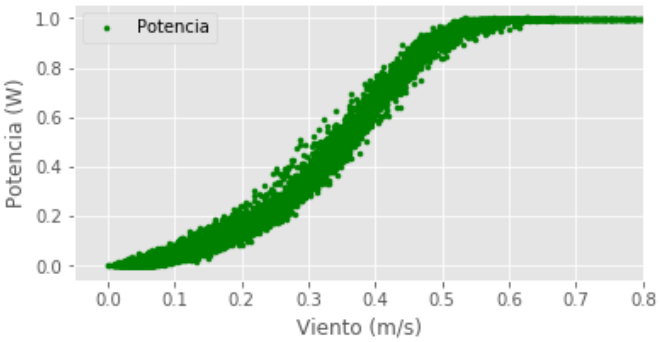


Figura 12. Curva de potencia Elaboración propia.

Otras relaciones interesantes son descritas por [9] donde se detallan la características utilizadas para predecir otras.

Variable a predecir	Potencia nominal	Potencia limitada
Potencia total (t)	Velocidad de viento (t) Angulo de *yaw* (t) Angulo de *pitch* (t)	Velocidad de viento (t) Angulo de *yaw* (t) Angulo de *pitch* (t)
Revoluciones del rotor (t)	Velocidad de viento Angulo de *yaw* Angulo de *pitch*	Velocidad de viento (t) Angulo de *yaw* (t) Angulo de *pitch* (t)
Temp. de componente (t)	Velocidad de viento Potencia total (t) Temp. ambiente (t) Temperatura de componente (t-1)	Potencia total (t) Temp. ambiente (t) Temperatura de componente (t-1)

Tabla 3. Modelos de predicción habituales

2.1.5 Predicción de energía

Regular la producción y vertido de energía a la red eléctrica de un país es un proceso sumamente complejo. No hemos encontrado referencias españolas, pero la agencia danesa de energía [6] explica el proceso que comienza 28 días antes de la producción, con una serie de aproximaciones sobre la demanda y la generación. Estas previsiones se van haciendo más rigurosas y con una granularidad menor hasta llegar a los 5 minutos un dia antes de la generación. El objetivo es regular en tiempo real la oferta y la demanda. con la mayor eficiencia y menor coste, primando el uso de energías baratas, frente al uso de las de reserva, más caras.

Para esto es necesario saber de antemano con qué energía se puede contar de cada tipo de fuente, y para ello es fundamental disponer de un robusto sistema de previsión de la producción. Con esto, se podrá hacer un balance seguro y eficiente.

Actualmente dispone dos métodos de predicción:

- *Offline*: Se realiza combinado diferentes modelos. Por un lado, modelos meteorológicos de predicción de vientos, y por otro una estimación de la producción en función del viento, obtenida desde los históricos. El error de estos modelos incrementa en función de los días a los que se realiza la predicción.

- *Online*: Puede ser mejorada la predicción anterior a corto plazo, si se combina con datos de los SCADA en tiempo real, ya que esto permite observar las pequeñas variaciones de la predicción frente a lo obtenido, como puede ser por ejemplo que un frente llegue antes de lo previsto. De este modo se mejora significativamente la predicción a intervalos de 5 minutos vista.

2.1.6 El mantenimiento preventivo

El mantenimiento preventivo es un modo de racionalizar costes, proteger equipos, mantener constante el ritmo de producción y garantizar el normal funcionamiento de una instalación o una maquinaria.

A medida que las perspectivas de crecimiento del sector eólico se afianzan, sobre todo debido a su expansión hacia el mar, es importante tener en mente la minimización de los costes de operación y mantenimiento que pueden ser muy elevados en este sector, e incrementar la fiabilidad y disponibilidad con objetivo de maximizar la producción.

Tradicionalmente el mantenimiento se ha clasificado en tres grupos:

- Mantenimiento correctivo: Es aquel que se ejecuta después del fallo para corregir el mismo. Tiene unos costes elevados ya que supone una parada normalmente no programada.

- Mantenimiento preventivo: Es aquel que tiene como objetivo evitar la aparición del fallo mediante intervenciones periódicas y anticipadas sobre unas instalaciones funcionales, como pueden ser inspecciones de rutina, ajustes, limpiezas y reemplazo de piezas desgastadas, etc.

- Mantenimiento predictivo: Es aquel que se efectúa tras observar ciertos indicios que pueden augurar algún tipo de fallo, y con el objetivo de adelantarse a la avería.

El coste de mantenimiento de una instalación de forma general depende de la estrategia elegida según se observa en la Figura 13.

Con un mantenimiento preventivo intensivo se reduce considerablemente el número de fallos, pero se eleva el coste, lo mismo que con un mantenimiento reactivo, que resulta más económico, pero produce sobrecostes en las reparaciones. La combinación equilibrada de ambos métodos es lo que ofrece el punto óptimo. [11]

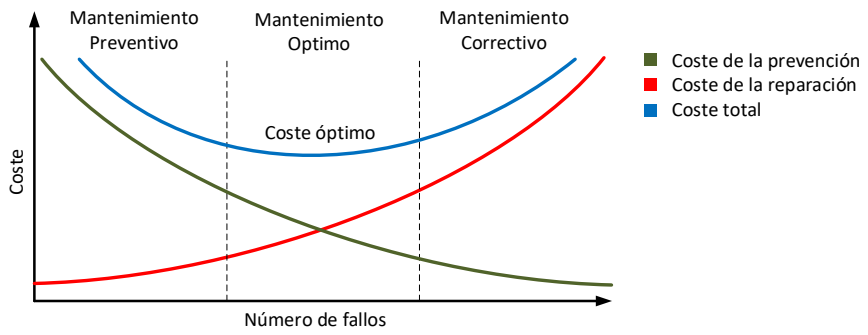


Figura 13. Costes de mantenimiento Elaboración propia.

Las técnicas de mantenimiento preventivo tradicionales se basan en el análisis de medidas de temperatura, vibración, termografía, ultrasonidos, etc. [12]. Recientemente con el auge del *big-data* se unieron las técnicas basadas en datos o *data-driven*. Estas técnicas amplían la capacidad de detección de anomalías de una forma más eficiente basándose en la detección de patrones de conducta. [13]

2.1.7 Sistemas de monitorización

El sistema de monitorización y supervisión es una herramienta utilizada para el mantenimiento de las máquinas. Proporciona información sobre cuáles son las condiciones de los componentes y puede permitir la predicción de un fallo antes de que ocurra.

Existen diferentes métodos destinados a la predicción del fallo [13] como se ilustra en Figura 14 .

- Basados en la experiencia y en histórico de incidencias. Es un método guiado por la intuición y su capacidad predictiva es cada vez más baja a medida que los sistemas aumentan en complejidad.
- Basado en datos: Es el método en el que nos centraremos en este trabajo. Se basa en la determinación del fallo basándose en los datos recientemente observados. No es necesario un conocimiento exhaustivo de modelos matemáticos ni de los sistemas monitorizados.
- Basado en modelos: Este método se basa en la confección de complejos modelos matemáticos que describen el comportamiento a través de principios físicos. Estos modelos son complicados de construir, se necesitan costosos experimentos y análisis numéricos que no siempre son viables.

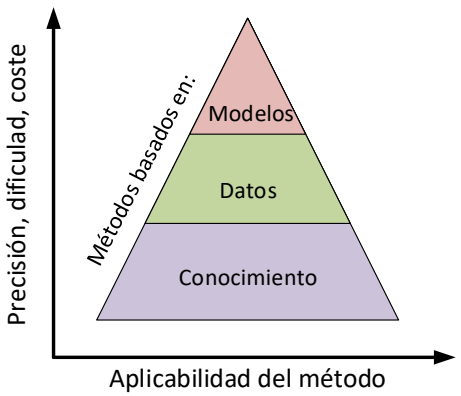


Figura 14. Métodos de predicción de fallo. Elaboración propia.

Los modernos aerogeneradores están siendo instalados cada ven en lugares más inaccesibles y en condiciones ambientales más agresivas,

sometidos continuamente a un estrés mecánico. Esto crea la necesidad de incluir sistemas de monitorización y supervisión de serie, no obstante equipos de cierta antigüedad sólo disponen de sistemas básicos. Es por esto por lo cual explorar los datos de SCADA puede aportar un valor añadido para las compañías energéticas de cara para prolongar la vida útil y reducir costes de mantenimiento en los aerogeneradores más antiguos.

Los elementos monitorizados suelen ser el generador, los convertidores de potencia, ejes de transmisión, multiplicadora, rodamientos y sistemas de control. Una extensa lista puede consultarse en [11] .

Por otro lado Pierre Tchakoua [14] ofrece una guía de conceptos y métodos tradicionales aplicados a sistemas de monitorización para aerogeneradores a la que añade el análisis de las variables de SCADA.

2.2 El aprendizaje automático

El aprendizaje automático engloba las técnicas y métodos que permiten aprender de los datos, generalizar comportamientos y realizar inferencias o predicciones para un conjunto más amplio y desconocido. Principalmente estos métodos se clasifican como supervisados y no supervisados como ilustra la Figura 15, aunque también existen los semi-supervisados y de una sola clase.

- Los métodos supervisados son aquellos que aprenden de conjuntos cuyas instancian están previamente etiquetadas o clasificadas dentro un numero finito de conjuntos. Los algoritmos tienen la tarea de aprender de estos datos para asignar posteriormente la etiqueta correcta a un dato nuevo que no ha visto previamente.

- En contra los métodos no supervisados no disponen de estas etiquetas, y su labor es encontrar similitudes o características comunes que clasifiquen o agrupen las instancias en un número de conjuntos desconocido.

- Los métodos semi-supervisados combinan los dos anteriores en los que se parte de un conjunto de datos en el cual se sabe a qué clase pertenecen ciertas instancias, pero también hay otras de las cuales se desconoce su etiqueta. En el caso de detección de anomalías, suelen utilizarse algoritmos de este tipo.

- Los métodos de una clase o *one-class*, son métodos de clasificación que parten de un conjunto de datos donde todas las etiquetas son de una clase. Esto hace que los algoritmos supervisados no puedan reconocer etiquetas no vistas anteriormente, no obstante, se sabe que existen y pueden ser consideradas anomalías. El funcionamiento de estos algoritmos se basa en crear fronteras de decisión que agrupen

los datos iniciales para que, en caso de encontrar una anomalía, esta quede fuera de este espacio.

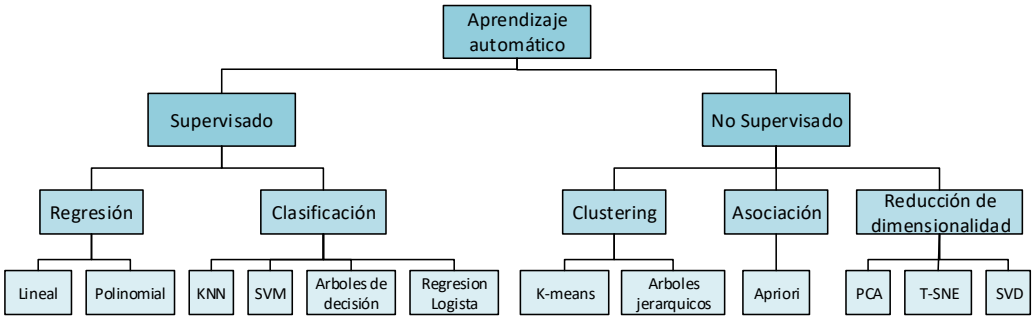


Figura 15. Métodos de aprendizaje automático y algoritmos más usados

2.2.1 Detección de anomalías

La detección de anomalías tiene una gran variedad de aplicaciones entre las que destacan la detección de fraude bancario, aplicaciones de seguros, sistemas anti intrusión, seguridad de la red, sistemas críticos, etc. En general cualquier sistema que genera datos que puedan ser almacenados y analizados mediante técnicas *big data* puede beneficiarse de esta tecnología.

Lo primero es definir lo que se considera una anomalía: La detección de anomalías se refiere a la tarea de entrar patrones en los datos que no encajan con lo previsto. Estos patrones son considerados anomalías o *outliers*, y una vez detectados pueden considerarse un síntoma de fallo si superan, por ejemplo, un umbral establecido.

Desde este punto de vista, se presentan varios desafíos que hacen que no sea una tarea fácil. [15]

- La definición de comportamiento normal no es trivial y una vez establecida puede no ser representativa en un futuro.
- No se puede asumir que las anomalías sean siempre fortuitas, y pueden ser enmascaradas intencionadamente.
- Muchas de las técnicas son válidas solo para ciertos ámbitos y la generalización es complicada.
- La disponibilidad de datos anómalos no siempre está disponible y la mayoría de las veces solo se dispone de datos de normalidad.
- Los datos suelen contener ruido que en ocasiones puede confundirse con las anomalías.

V. Chandola [15] define extensamente una serie de algoritmos usados para la detección de anomalías, que pueden ser agrupados en métodos de clasificación, métodos de detección de vecindad y métodos

estadísticos. La naturaleza de los atributos de los datos determinará la capacidad de aplicar un método u otro, dependiendo por ejemplo si se tratan atributos continuos o categóricos.

Las anomalías pueden clasificarse según su tipo:

- Individuales: Se puede considera una simple muestra anómala, como un evento individual e independiente del resto. Por ejemplo, una transacción bancaria de una cantidad fuera de lo habitual.

- Anomalías contextualizadas: Donde el factor que determina si se trata de una anomalía o no, depende de un contexto como por el ejemplo el momento en el que ocurra. Por ejemplo, una temperatura habitual en ciertos momentos puede ser anómala en otros.

- Anomalías colectivas: Donde la anomalía en sí, no es valor de una muestra sino el valor sucesivo de varias muestras, por ejemplo, una desviación progresiva de una temperatura.

También puede considerarse tres maneras diferentes de detectar anomalías en función de si se trabaja con datos etiquetados o no. Las etiquetas en un conjunto de datos harían referencia a aun estado correcto o anómalo establecido de antemano. En muchas ocasiones esta información no estará disponible.

- Detección de anomalías supervisada: Engloba las técnicas aplicadas a conjuntos de datos etiquetados como normales y anómalos. Este método tiene varios problemas, como el desbalanceo del número de muestras de cada clase y la asignación previa de la clase a cada instancia que debe realizarse en base al conocimiento. Esta detección suele basarse en métodos de clasificación.

- Detección de anomalías semi-supervisada: En este método se dispone de un conjunto de datos que representa la normalidad, y se sabe que puede haber anomalías que van a diferir en aspecto con respecto a la normalidad, pero no se han observado todavía. Las técnicas usadas se basarán en medir el grado de desviación de las muestras de prueba con respecto a la normalidad.

- Detección de anomalías no supervisada: En este método se dispone de datos válidos y anómalos de forma conjunta, pero sin saber cual corresponde a cada clase, de modo que se tendrán que aplicar algoritmos para tratar de diferenciarlos.

La consideración de anomalía o no se realizará dependiendo de los algoritmos utilizados. Aquellos que se basen en métodos de clasificación

asignará una categoría o etiqueta que marcará la anomalía, mientras que otros métodos tendrán como salida una métrica o *score* sobre la que se establecerá un umbral que separe la normalidad de la anomalía.

La detección de anomalías en ocasiones se confunde con la clasificación de sucesos raros, que corresponde a un método supervisado de clasificación donde el suceso raro es una clase desbalanceada, es decir, tiene muy pocas ocurrencias dentro del conjunto de datos.

2.2.1.1 Técnicas más relevantes

Algunos de los métodos tradicionales utilizados para la detección de anomalías son basados en técnicas *one-class,* es decir, algoritmos que son entrenados conociendo una sola clase y cuya labor es descubrir otro tipo de clase desconocida.

Algunos algoritmos destacables. [16]

- *Isolation forest*: Este algoritmo de clasificación, es usado para la detección de anomalías no supervisada, es decir, se parte de un conjunto de datos donde coexisten datos normales con anómalos, pero no se sabe de antemano a que clase pertenecen.
  Está basado en árboles de decisión e intenta separar todas las muestras mediante divisiones aleatorias hasta completar un árbol. En principio es más fácil aislar las nuestras anómalas con un número menor de separaciones, así pues, establece una puntuación de anomalía en función del número de separaciones que necesita para aislar cada muestra.

- *One class SVM*: Se basa en máquinas de soporte vectorial, pero con solo una clase y se suele usar en la detección de anomalías semi-supervisada. El algoritmo utiliza funciones *kernel* para trasladar los datos a una dimensionalidad mayor y poder utilizar hiperplanos para delimitar las instancias con las que ha sido entrenado, asumiendo que son todas normales. Cuando el algoritmo intente clasificar instancias nuevas que no encajan con la normalidad es de esperar que queden fuera de las fronteras de decisión trazadas por los hiperplanos.

- *K-Nearest neighbor*: El uso de este algoritmo permite la separación de las muestras en dos grandes clases, datos normales y datos anómalos, en función de las distancias entre las muestras y sus centroides. De modo que pueden ser usados en la detección de anomalías no supervisada. Otros algoritmos de agrupación como *DBScan* pueden ser usados de la misma manera.

Algunos otros métodos se basan en redes neuronales:

- *Autoencoders*: se denomina así a un tipo de red neuronal que se compone de dos partes. Un *encoder*, o etapa en la que se reduce la dimensionalidad desde un número elevado de entradas a un número inferior de salidas, y el *decoder*, o la etapa que reconstruye la señal de alta dimensionalidad a partir de la salida de la anterior etapa de baja dimensionalidad.
  Este tipo de redes son útiles para funciones de compresión de la información, clasificación y detección de anomalías. Para éste último caso, el funcionamiento se basa en el que la etapa de *encoder* y *decoder* se entrenan de forma conjunta bajo un dataset considerado normal. La etapa *decoder* será capaz de reconstruir la información de encoder con un determinado grado de error. En caso de utilizar la red para una predicción donde la entrada presente anomias o diferencias con respecto a los datos entrenados, se traducirá en una ineficiencia de la relación compresión-descompresión que se manifestará en un mayor grado de error a la salida de red. La cuantificación de este error puede determinar la existencia o no de datos anómalos.

- Máquinas de *Boltzman*: Se trata de un método no supervisado basado en una red neural compuesta solo por las neuronas de entrada y las de una capa oculta, y que además todas las neuronas están conectadas con todas. Están diseñadas para aprender la distribución de probabilidad del conjunto de datos con el que se entrena. Existen variantes como la *restricted* donde las neuronas de la misma capa no están conectadas entres si, y son usadas para la detección de anomalías.

- *Long short-term memory models*. Es un tipo de red neuronal que se usa para modelar el conjunto de datos acorde a una secuencia temporal. Se realizado enviando la salida de la red en el momento T, como una entrada del momento T+1. La idea es que la red pueda olvidar datos pasados irrelevantes mientras que pueda recordar datos que son relevantes para el futuro. Se suele usar en series temporales del mismo modo que otros métodos autorregresivos como ARIMA, comparando lo observado con la predicción a futuro realizada con los datos del pasado.

2.2.1.2 Métodos de detección estadísticos

Los métodos estadísticos son muy usados para la detección anomalías semi-supervisada, se basan en la asunción de que los datos sin anomalías se ubican en las regiones de probabilidad más altas del modelo estocástico, mientras que los datos con fallo o anomalías se ubican en las regiones de probabilidad más baja. [10]

Así pues, el principio de funcionamiento para la detección de anomalías se basa en buscar modelos que sean capaces de predecir unas características a partir de otras. Para esto son de aplicación numerosos

algoritmos de regresión, como árboles aleatorios, SVR o redes neuronales. Se evalua la bondad del ajuste con diferentes métricas como error absoluto medio o error cuadrático medio de cara a determinar los mejores de ellos.
Una vez encontrados los modelos más efectivos, se realiza una predicción a partir de un modelo entrenado con datos que describen la normalidad contra datos del SCADA obtenidos a posteriori. El error obtenido en las estimaciones habitualmente sigue una distribución normal con media entorno a cero.

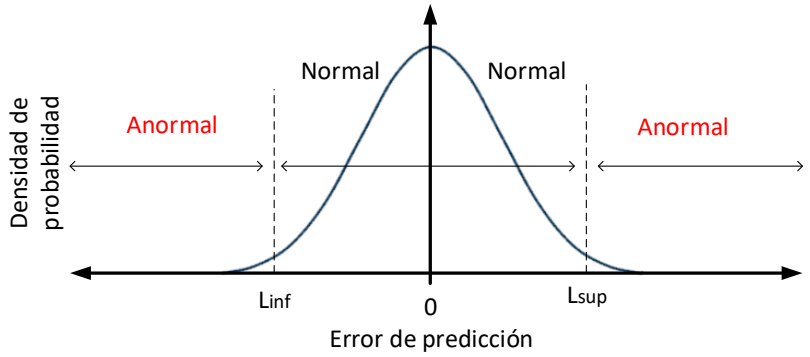


Figura 16. Distribución del error de predicción Elaboración propia.

Durante el proceso de entrenamiento se determinan los parámetros que engloban la distribución de los errores en la normalidad, que luego durante la fase de predicción servirán para discriminar las anomalías, identificándolas por aquellas mediciones cuyo error de estimación se encuentra fuera de los límites. La Figura 16 ilustra los rangos de discriminación del error.

Los limites deben ser ajustados cuidadosamente para evitar falsos positivos, que pueden provocar la pérdida de confianza en el modelo, pero al mismo tiempo tienen que ser lo suficientemente ajustados como para mantener la capacidad de detectar desviaciones que revelen la anomalía. M. Schlechtingen [10] ilustra el proceso y plantea por ejemplo los siguientes límites:

- Los errores cometidos por el modelo que tiene una probabilidad de aparecer superior al 0.01% con considerador como normales.
- Los errores cometidos por el modelo que tienen una probabilidad inferior al 0.01% son considerados anomalías.

La Figura 17 ilustra el proceso completo de la detección de anomalías mediante éste método, uno de los más usados en la bibliografía consultada.

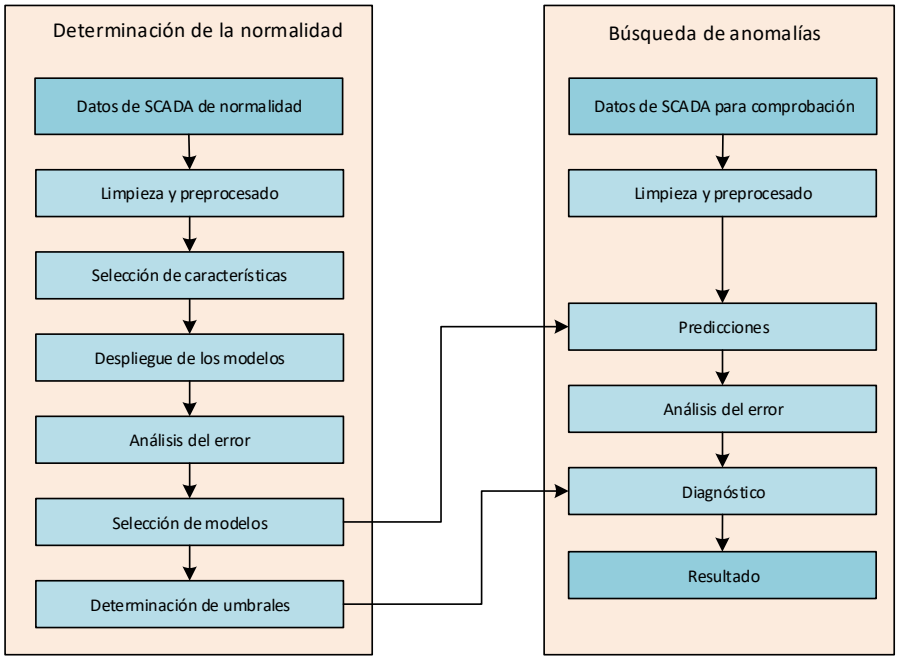


Figura 17. Proceso de detección de anomalías

2.2.2 Aplicación en aerogeneradores

Los datos que vamos a tratar en este trabajo pertenecen al ámbito industrial, donde hay que diferenciar dos tipos de sistemas [13]

- Continuos: Donde el sistema trabaja la mayoría del tiempo en un régimen óptimo y constante desde del inicio, como por ejemplo motores y maquinas rotativas. La mayoría de los estudios se desarrollan en este entorno.

- Discretos: Donde los sistemas tienen diferentes estados y modos de operación que alternan en función de una lógica de control, haciendo que funcionen en condiciones diferentes durante periodos de tiempo variables.

Un aerogenerador estaría incluido en un sistema discreto, que en general son más complejos de tratar. Las máquinas alternan estados de funcionamiento y parámetros de control que puede hacer que la definición de normalidad sea muy compleja [17].

El tipo de anomalía habitual en estos sistemas es la anomalía individual o colectiva, como una secuencia de valores anómalos.

En el ámbito industrial, la detección de anomalías puede aplicarse fundamentalmente para el pronóstico de fallos en sistema mecánicos, donde los datos suelen tener una componente temporal como por ejemplo los procedentes de sensores y lecturas de sistemas de adquisición, pero también como pronóstico más avanzado que pueda detectar

funcionamientos erróneos de, por ejemplo: un sistema de control o regulación que no consiga alcanzar el comportamiento programado.

En este entorno la tenencia de datos de normalidad es relativamente fácil y son de aplicación los métodos semi-supervisados incluyendo técnicas de modelado estadístico y redes neuronales.

Alguno de los problemas que pueden plantearse y que se abordaran más adelante pueden ser:

- Limitaciones del conjunto de datos. No todos SCADA aportan la misma cantidad de señales y hemos visto que algunas son de antemano mucho más significativas que otras.
- La calidad de los datos.
- La selección de características para construir modelos, para lo cual será necesario el conocimiento del sector o la aplicación de algoritmos para su determinación.
- La ausencia de anomalías, para lo cual deberían sintetizarse.

2.2.3 Predicción de la producción

El uso del aprendizaje automático participa en los modelos de estimación de la de producción como se vio en el punto 2.1.5. tanto en los modelos de predicción *offline* como *online*.

En la predicción *offline* son de aplicación algoritmos no lineales como redes neuronales, máquinas de soporte vectorial o procesos gaussianos, que correlacionan principalmente viento con potencia y permite hacer una predicción utilizando como entrada una estimación del viento que puede proceder de modelos numéricos propios o de servicios meteorológicos externos.

Existen numerosos portales web que ofrecen la predicción de vientos a días vista y con diferentes granularidades que van desde dos predicciones al día hasta 24. Algunos de estos servicios son accesibles a todo el mundo con tan solicitar una clave API que permite su consulta a través de una interfaz de llamadas REST. Por ejemplo, *Wheather Underground* (www.wunderground.com) ofrece predicciones a nivel mundial y la Agencia Estatal de Meteorología (www.aemet.es) hace lo mismo, pero cubriendo solo el territorio nacional.

Para la predicción *online* intervienen métodos de predicción de series temporales, en especial los denominados métodos auto regresivos, es de decir, que no usan como características variables independientes, sino que usan los propios datos pasados para predecir los futuros, por ejemplo, el método ARIMA o basados en redes neuronales como LSTM.

2.3 Trabajos realizados

Se han consultado numerosas fuentes de documentación en búsqueda de trabajos relacionados, obteniendo los mejores resultados en conocidos portales que facilitan el acceso a artículos de revistas de carácter científico-técnico, libros y publicaciones de muy alto nivel relacionadas con la ingeniería y los datos, como son:

- 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.

También se han consultado diferentes repositorios de universidades sin haber tenido éxito en este sector. En cambio, sí que se han encontrado trabajos relacionados con la detección de anomalías en otras áreas, como tráfico de red o fraude.

La información consultada nos ha sido el punto de partida para la elaboración de este estado del arte, donde se ha resumido y organizado los conceptos y técnicas encontradas. La hemos agrupado en tres apartados.

2.3.1 Sector eólico

El conocimiento del área de aplicación es importante para un despliegue exitoso de las herramientas de minería de datos, si bien, tampoco hace falta ser un experto, son varios los trabajos que hacen referencia a este conocimiento sobre todo a la hora de la selección de características.
En ese sentido hemos recopilado información extensa sobre el funcionamiento, anatomía y sistemas de control de un aerogenerador.

En internet se pueden encontrar gran cantidad de recursos que sirven como introducción, no obstante existen trabajos mucho más rigurosos como por el ejemplo el de Tomas Esbensen [8] donde se establece un modelo dinámico de control de aerogeneradores basado en modelos matemáticos que formulan el comportamiento íntegro del sistema, teniendo como objetivo la propuesta de un sistema robusto de control. Todo está formulado matemáticamente: viento, sistema aerodinámico, tren de fuerza, generador, torre, *pitch*, etc. para permitir simulaciones íntegras con software específico Simulink.

Un trabajo similar es aportado por Farsoni Saverio [18] en su tesis doctoral.

En estos trabajos se analiza en profundidad la partes que componen el sistema y cómo son gobernadas por los sistemas de control interno de modo que se puede obtener una lógica de funcionamiento del mismo, y serán de referencia para el análisis del conjunto de datos en los siguientes capítulos.

Un importante recurso es también el informe de la agencia energética danesa: "*Integration of Wind Energy in Power Systems"* [6] de 2017, donde refleja la importancia del sector y la integración en el sistema energético nacional de las energías procedente de las renovables.

2.3.2 Mantenimiento preventivo y detección anomalías

Son numerosas las referencias que hemos consultado sobre mantenimiento preventivo en aerogeneradores, casi todos hacen hincapié en su importancia debido a los altos costes de reparación e intervención es este sector.

Mientras que un gran número de trabajos se centran en el mantenimiento preventivo tradicional, muchos trabajos analizan el uso de técnicas de minería de datos con registros de los SCADA, ya no solo para mantenimiento preventivo, sino también para la predicción de la producción, o sistemas de control más robustos que sean tolerantes a fallos y adapten su funcionamiento a un autodiagnóstico.

Estos trabajos dejan de manifiesto que el uso de datos SCADA junto con técnicas de minería de datos ya es de aplicación. Estos son algunos trabajos destacados, aunque han sido muchos más los consultados, con temáticas similares:

- Kai Zhong [13] expone diversos métodos de pronóstico de fallo en entornos industriales, como los basados en la experiencia, en modelos y en datos.

- Z. Hameed [19] realiza un planteamiento completo de un sistema de monitorización en aerogeneradores desglosándolo por subsistemas e incluyendo las técnicas tradicionales: análisis de vibraciones, análisis de aceite, termografía, salud estructural, ultrasonidos, etc. También realiza una supervisión de parámetros de control.
  Pierre Tchakoua [14] tiene un trabajo similar, y dan pie a muchos otros donde ya se aplican técnicas de minería de datos.

- Papatheou [20] aplica redes neuronales y procesos gaussianos con Matlab para realizar un estudio de la salud estructural de un parque eólico. El objetivo que plantea es determinar si un aerogenerador presenta alguna anomalía con respecto al resto a partir de los datos

de los SCADA del parque utilizando la curva de potencia.

- Peng Sun [9] utiliza una red neuronal para realizar la predicción de algunos parámetros del SCADA y determinar anomalías con el análisis de los residuos. En este caso dispone de datos reales con históricos de fallo. En otra publicación suya [17] utiliza algoritmos genéticos para la selección de características cuando se dispone de datos de muy alta dimensionalidad.

- Peng Qian [21] propone una red neuronal de una sola capa oculta (SLFN) para estimar el grado de salubridad de una turbina de cara a usarlo como entrada a los sistemas de control, y de ese modo ajustar la producción de energía de un modo inteligente y protector.

- Wenna Zhang [22] utiliza un método de reducción de dimensionalidad llamada Análisis de Factores Paralelos (PARAFAC) usando un gran número de variables de entrada en vez de modelos más específicos, para la detección de anomalías. En este caso dispone de datos completos con fallos de un parque operacional y demuestra su eficacia junto al algoritmo k-means en la clasificación de las máquinas acorde a tres estados de salubridad: normal, aviso y alarma.
  Mian Du [23] también utiliza un método de reducción de dimensionalidad a la hora de seleccionar las características, esta vez basado en mapas auto organizados (SOM)

- Meik Schlechtinger [10] propone un método de monitorización de variables a través de modelos de normalidad obtenidos con redes neuronales ANFIS y lógica borrosa desde los cuales puede establecer de forma automática estados operacionales intuitivos que representen la salud de las máquinas.

2.3.3 Aplicación de modelos

También hemos encontrado muchos trabajos que se centran solo en diferentes técnicas para modelar la curva de potencia, como la relación más significativa del rendimiento de un aerogenerador.

Son trabajos más orientados a la aplicación de modelos paramétricos como: *Polynomial regression, Locally Weighted Polynomial Regression, Four-parameter logistic function (4-PL), Five-parameter logistic function (5-PL), Spline Regression, Penalized Spline Regression, Modified hyperbolic tangent (MHTan),* etc. donde el objetivo es ajustar unos parámetros que describan la curva a través de una fórmula. [24] [25] y otros.

También existen trabajos que hacen lo mismo con modelos no paramétricos como: *K-nearest neighbors regression (KNN), Random forest regression (RF), Extremely randomized trees (ERTs), Stochastic gradient boosted regression trees (SGBRT), Gaussian Process (GP), Artificial Neural Network (ANN),* etc. [24] [26] [27] y otros.

La mayoría de ellos tienen como objetivo evaluar la bondad de ajuste de diferentes métodos o validar su uso en el sector. Muchos trabajos recurren a conjuntos de datos sintéticos [27] mientras que otros disponen de datos reales para la validación de los modelos.

En casi todos se concluye con la validez de los modelos, no obstante, observamos que los resultados son muy dependientes del conjunto de datos y el tipo de máquina de la cual proceden.

Hemos notado que las diferencias entre máquinas son muy evidentes y sustanciales, a la vez que las diferencias entre las curvas teóricas ofrecidas por el fabricante y las reales obtenidas en campo.

Por lo cual puede ser que un modelo paramétrico, basado en una regresión polinómica, se ajuste bien a una curva de potencia de una máquina de 2 MW pero no lo haga para un de 650 W, ya que la distribución es muy diferente debido a que las maquinas aprovechan la fuerza del viento de formas mecánicas diferentes.

Esto es lo que nos obligará seguramente a realizar nuestra propia selección de modelos.

Como resumen, la Tabla 4 incluye las técnicas empleadas en las referencias consultadas, quedando patente su utilización actual con diferentes objetivos como reducción de la dimensionalidad, selección de características, predicción y detección de anomalías. No obstante, los resultados son muy dependientes de los datos y las máquinas.

Técnicas	Referencias
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]

Tabla 4. Técnicas empleadas

3 Diseño e implementación

3.1 Selección de herramientas

Llegado el momento de abordar los datos, lo primero es determinar las herramientas que vamos a utilizar para realizar el trabajo. Hay muchas soluciones disponibles, pero la que destaca por su potencia, sencillez y sobre todo, porque es libre, es *Python* en unión con una serie de librerías relacionadas con la manipulación de datos y algoritmos de *machine learning*. También destaca R, un conocido software estadístico cuyas capacidades son similares.

Nos decantamos por *Python*, por tener una mayor afinidad con él. Además, el uso *Python* ha tenido en los últimos años un auge sustancial con respecto a otros lenguajes de programación.

Anaconda y *WinPython*, son las distribuciones más conocidas que integran un gran número de librerías y herramientas listas para trabajar, están orientados a un entorno científico e incluyen *Spyder* con entorno de desarrollo.

3.1.1 Librerías usadas

Python solo es un intérprete y la potencia le llega a través de un sinfín de librerías y paquetes de código abierto que le aportan funcionalidades específicas para cualquier entorno de aplicación.

En nuestro caso las librerías más conocidas que usaremos serán:

- *Numpy*: Es una librería que añade soporte para trabajar con vectores y matrices de datos a través de instrucciones de alto nivel
- *Pandas*: Es una extensión de *Numpy* y ofrece estructuras de datos y operaciones para la manipulación de tablas y series temporales.
- *Matplotlib*: facilita la generación de potentes gráficos a partir de datos contenidos en estructuras de la librería *Numpy* y *Pandas*.
- *Scikit-learn*: Implementa una gran cantidad de algoritmos de aprendizaje automático, métricas y funciones para el tratamiento de datos.
- *Keras* y *Tensorflow*. Es una biblioteca que hace de interfaz y facilita la construcción de redes neuronales independientemente del motor utilizado, que puede ser *TensorFlow*, *Microsoft Cognitive Toolkit* o *Theano*. Nosotros usaremos *Tensorflow* de *Google*.

3.2 Limpieza y análisis exploratorio de datos

La recepción de los datos se realiza en un fichero CSV con 23 variables procedentes del SCADA de un aerogenerador de 2 MW de una instalación privada cuyos datos no son públicos. Contiene 44.640 adquisiciones correspondientes a medidas diezminutales. Este método de presentación, en medias de diez minutos sigue un estándar en eólica para el análisis de datos que tiene como objetivo evitar la influencia excesiva de la turbulencia y reducir la varianza.

En total supone un tiempo aproximado de 300 días de datos.

En las labores de este análisis exploratorio se incluye la descripción de las variables, estudio de la existencia de datos anómalos o perdidos y cálculo de los principales estadísticos.

Las variables que trabajaremos, una vez retiradas del conjunto original las que no aportan información de la máquina como la carga del procesador o memoria utilizada por el autómata son las indicadas en Tabla 5 y sus principales estadísticos se reflejan en la Tabla 6.

El significado y las unidades son inferidas, ya que el juego se encuentra sin documentación adicional.

Variable	Descripción	Unidades
EstadoOp	Estado operacional	--
ParGen	Par del generador	N/m
PitchA	Angulo de pitch pala 1	Grados
PitchB	Angulo de pitch pala 2	Grados
PitchC	Angulo de pitch pala 3	Grados
PitchMin	Angulo de pitch mínimo	Grados
PitchRef	Angulo de pitch de referencia	Grados
PotEstatorRef	Potencia en el estator referencia	W
PotReactiva	Potencia reactiva	W
PotStator	Potencia en el Estator	W
PotTotal	Potencia Total	W
PresGH	Presión del grupo hidráulico	hPa
RpmGen	Revoluciones del generador	Rpm
RpmGenFil	Revoluciones del generador filtrado	Rpm
RpmRotor	Revoluciones del rotor	Rpm
VGrid	Tensión de la red de salida	V
WindCazo	Velocidad viento	m/s
WindDirection	Dirección del viento	Grados
WindSonico	Velocidad viento	m/s

Tabla 5. Variables del conjunto de datos

En el anexo I se incluye un graficado completo de todas las señales y una descripción estadística más completa.

Variable	mean	std	min	25%	50%	75%	max
EstadoOp	54,00	5,46	11,00	55,00	55,00	55,00	55,00
ParGen	211,51	300,05	0,00	12,34	100,08	241,94	1174,19
PitchA	5,96	15,04	-2,88	-0,07	0,77	1,64	90,08
PitchB	6,06	15,21	-0,64	-0,05	0,81	1,65	90,04
PitchC	6,06	15,20	-0,63	-0,04	0,81	1,65	90,01
PitchMin	6,08	15,88	-0,68	-0,09	0,79	1,63	83,00
PitchRef	5,97	15,01	-0,68	-0,09	0,77	1,63	90,00
PotEstatorRef	331,70	471,54	0,00	18,41	155,55	380,00	1844,43
PotReactiva	6,74	30,45	-60,00	-0,17	0,00	1,00	127,00
PotStator	331,67	471,54	-0,84	18,36	155,61	380,05	1843,87
PotTotal	317,30	526,26	-19,60	-1,86	86,76	327,62	2000,53
PresGH	188,61	9,83	0,30	186,51	189,23	191,55	196,39
RpmGen	1052,05	465,33	-110,39	999,74	1000,24	1349,49	1692,04
RpmGenFil	1052,04	465,32	-110,43	999,73	1000,22	1349,28	1692,10
RpmRotor	10,41	4,61	0,00	9,89	9,90	13,35	16,75
VGrid	692,78	5,25	674,12	689,11	692,65	696,27	708,79
WindCazo	5,08	3,62	0,00	3,00	4,43	6,47	25,26
WindDirection	0,13	25,30	-177,97	-5,58	-0,26	5,59	179,26
WindSonico	5,31	3,35	0,14	3,00	4,48	6,55	25,47

Tabla 6. Principales estadísticos

3.2.1 Determinación de los datos de normalidad

Los datos que usaremos para realizar los modelos han de incluir a la máquina en estado operativo, es decir, en funcionamiento de carga parcial (sin llegar a la producción máxima) o carga completa (producción máxima

Tenemos que descartar los periodos de inactividad donde la maquina se encuentra en parada, o en condiciones donde no es posible la producción de energía.

Examinando los modos de operación de las máquinas, podemos establecer el estado funcional de acuerdo a los siguientes criterios que aseguran que maquina entra en la zona de carga parcial.

- Potencia de salida mayor de 0 W
- Velocidad de viento superior a 2 m/s
- Revoluciones de generador superiores a 960 rpm
- Angulo de pitch inferior a 40º

Aplicado las anteriores restricciones, el 74% de las muestras se encuentran en condiciones de trabajo y serán las que utilizaremos. La Figura 18 muestra la distribución de sus valores.

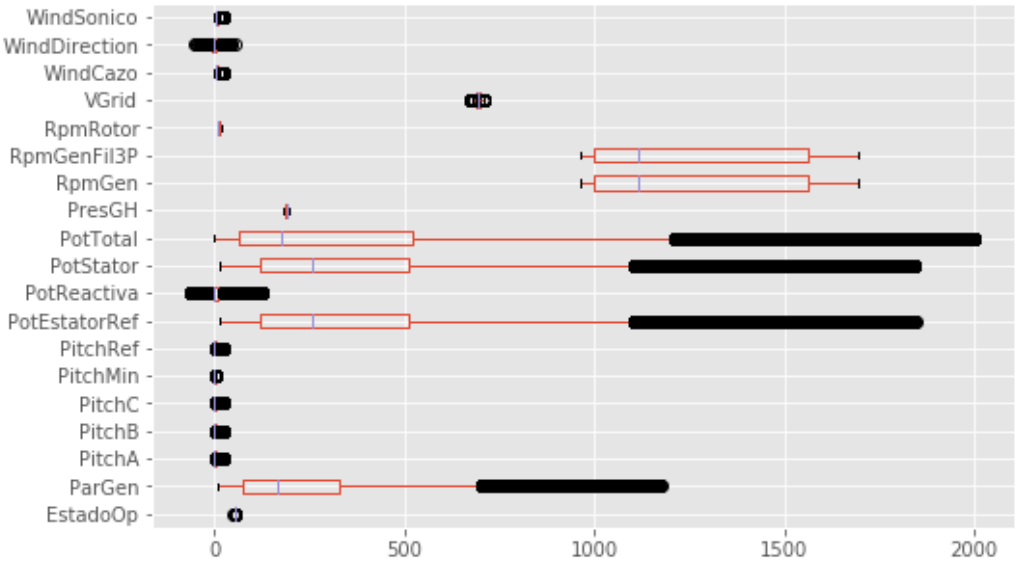


Figura 18. Distribución de las variables en condiciones de trabajo

En el anexo I se pueden observar la graficas más relevantes diferenciando los estados operacionales. Como ejemplo se muestra la serie temporal del viento, y su distribución en función del estado de funcionamiento de la máquina en la Figura 19

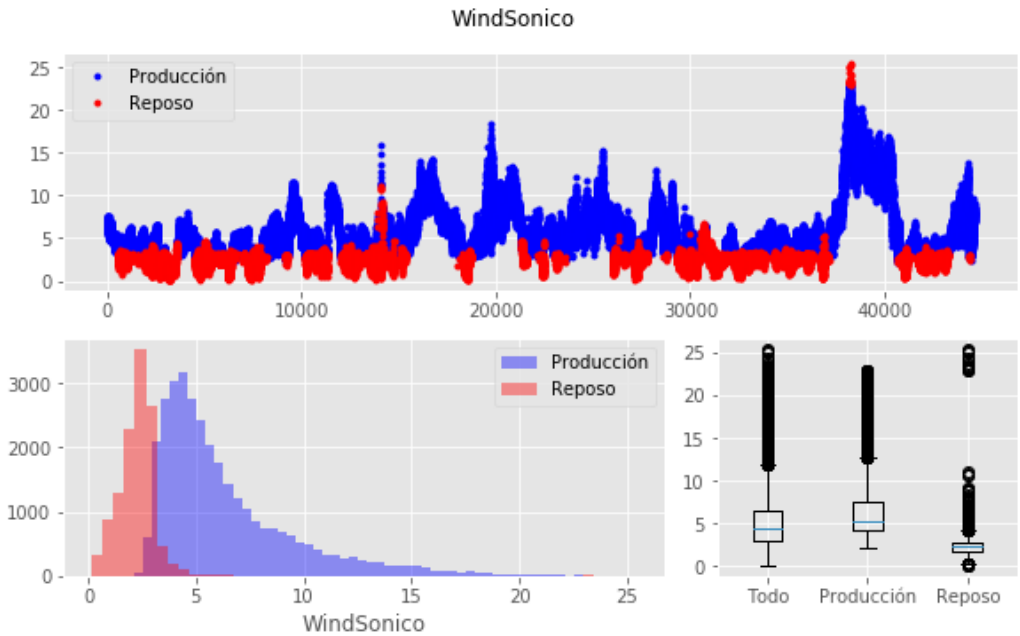


Figura 19. Descripción del viento en función del estado

3.2.2 Selección de características

Algunas de las variables que hemos recibido han sido descartadas por no aportar información de la máquina, y otras pueden ser descartables por otros motivos, como estar altamente correlacionas a nivel mecánico y aportar por lo tanto la misma información. Calculamos el índice de

correlación de *Pearson* y lo visualizamos en el mapa de calor de la Figura 20. También incluimos una gráfica *pairplot* para descubrir relaciones entre ellas de forma visual en la Figura 21

Crearemos una variable nueva (día), que agrupa las muestras pertenecientes al mismo día, ya que más adelante la usaremos para separar conjuntos de entrenamiento y test.

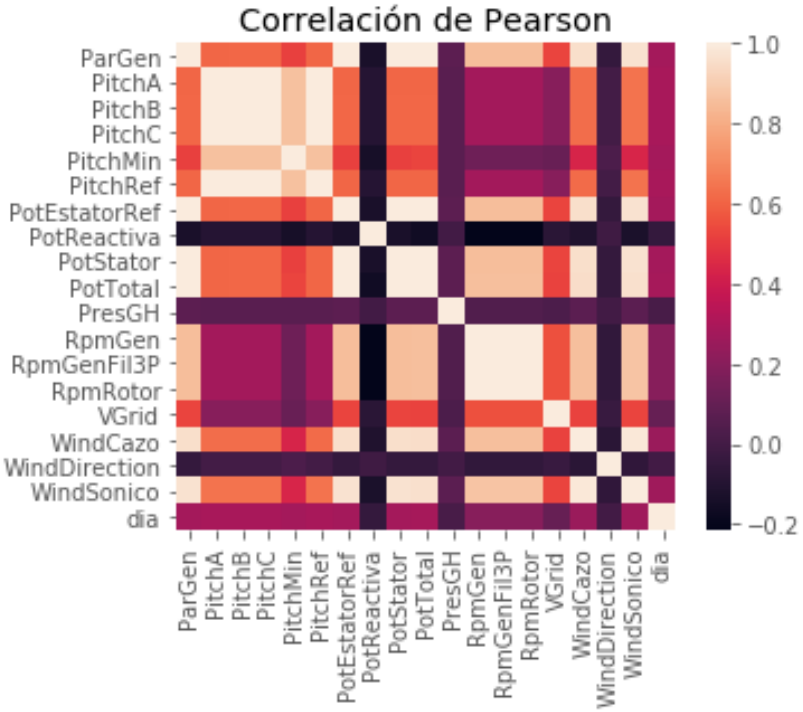


Figura 20. Correlación entre variables

Todas las variables de *pitch* están altamente correlacionadas entre ellas con un índice prácticamente de 1. El *pitch* de referencia es el valor de *pitch* establecido por el control como óptimo para el funcionamiento de la máquina en condiciones actuales, los valores de *pitch* de las palas 1,2,3, son el valor instantáneo medido, y el mínimo hace referencia a algún tipo de estadística. Como los datos están promediados apenas hay diferencias entre ellos, consideramos que el más relevante es el de referencia, denominado *PitchRef*.

En el generador podemos encontrar diferentes medidas también muy correlacionadas, como el par y la potencia total, así como diferentes medidas de potencia dentro del generador, cuya suma es la potencia total, siendo esta última la más significativa.

En cuanto a vientos tenemos dos medidas, que hace alusión a diferentes métodos de medida. Hemos observado bastante más estable, sobre todo a bajas velocidades de viento la medición sónica frente al sistema de cazoletas, que al contener elementos móviles es más impreciso.

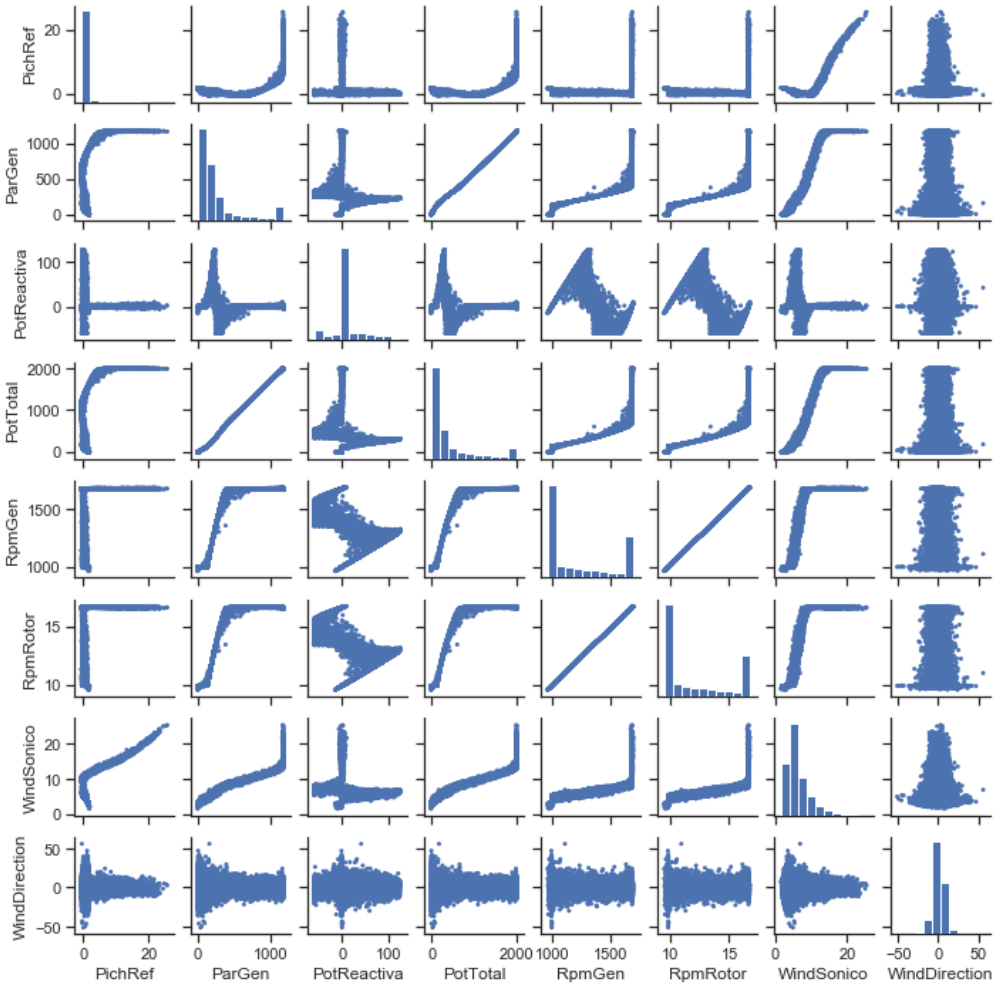


Figura 21. Relaciones entre variables

3.3 Obtención de modelos

En este apartado, buscaremos características que puedan ser inferidas a partir de otras dentro de los datos que describen la normalidad en régimen de producción de energía. Analizaremos la correlación entre variables y la importancia de éstas de cara a la construcción de modelos que serán evaluados por diferentes métricas para poder ser comparados.

Como primer paso, para cada variable a predecir, buscaremos entre el resto de variables, aquellas cuyo índice de correlación de *Pearson* sea elevado. Estas variables las usaremos luego para construir un modelo preliminar *Random Forest Regressor.* Este algoritmo ofrece estadísticas sobre cuál de las características empleadas es más significativa en el modelo, de modo que es de ayuda para determinar las más influyentes y descartar aquellas que no aporten información.

Pero antes vamos a describir la métricas y algoritmos que usaremos.

3.3.1 Métricas y sistema de evaluación

Para esta valoración utilizaremos diferentes métricas que indican cómo de bueno es el ajuste del método. La diferencia obtenida entre las predicciones efectuadas y la realidad se denomina error o residuo, y a partir de estos valores hay numerosas formas de obtener indicativos de calidad del ajuste. Algunas de que usaremos serán: MAE, MSE, RMSE, MAPE y R2 que se detallan a continuación.

- Error absoluto medio (MAE): Es la métrica quizás más fácil de entender ya que esta expresada en las mismas unidades que la característica predicha. Se trata simplemente de la media de los valores de los residuos, dando una idea de cuál es el error medio. Un valor de 0 indica un ajuste perfecto.

$$MAE = \frac{1}{N}\sum|y - \hat{y}|$$

- Error medio cuadrático (MSE): Se calcula igual que error absoluto, solo que la diferencia entre predicción y valor real se eleva al cuadrado. Esto hace que puntos con mayor error contribuyan de mayor manera. El problema de esta métrica es que no está expresada en las mismas unidades y los *outliers* pueden interferir demasiado.

$$MSE = \frac{1}{N}\sum(y - \hat{y})^2$$

- Raíz del error medio cuadrático: (RMSE): Básicamente es la raíz cuadrada del MSE, de este modo se expresa en unidades de la variable predicha haciendo más fácil la interpretación.

$$RMSE = \sqrt{\frac{\sum(y - \hat{y})^2}{N}}$$

- Error porcentual absoluto medio (MAPE). Es el equivalente al MAE en porcentaje. Se calcula igual, pero escalando cada residuo por el valor real. De este modo ofrece la misma idea que MAE, pero en porcentaje. Tiene el problema de que cuando el valor predicho es cero, queda indeterminado.

$$MAPE = \frac{100\%}{N}\sum\frac{|y - \hat{y}|}{y}$$

- Coeficiente de determinación R2: Es una medida estadística que describe la bondad del ajuste de un modelo, indicando cual es la proporción de la variación total en la variable dependiente que es explicada por el modelo. Está en el rango 0-1, indicando un ajuste perfecto con el valor de 1.

$$R^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

Por otro lado, a la hora de validar un modelo hay que asegurarse de que es capaz de generalizar y ser efectivo para datos nuevos que nunca ha visto. Para ello se utilizan dos técnicas: la partición del conjunto de datos en dos subconjuntos, uno destinado al entrenamiento y otro a la validación o *test*.

El objetivo es evitar el sobre entrenamiento, es decir, que el modelo se ajuste en exceso a los datos con los que ha sido entrenado. Se suelen comparar las precisiones obtenidas entre los datos de entrenamiento y de validación, que deben ser similares.

Existe otra técnica, denominada validación cruzada o *k-fold CV,* que permite examinar el comportamiento más exhaustivamente dividendo el conjunto en varios subconjuntos, y evaluando la predicción de cada uno de esos conjuntos usando como datos de entrenamiento el resto de conjuntos. Esta técnica puede usarse conjuntamente a la anterior de entrenamiento y validación. Su funcionamiento se ilustra en la Figura 22.

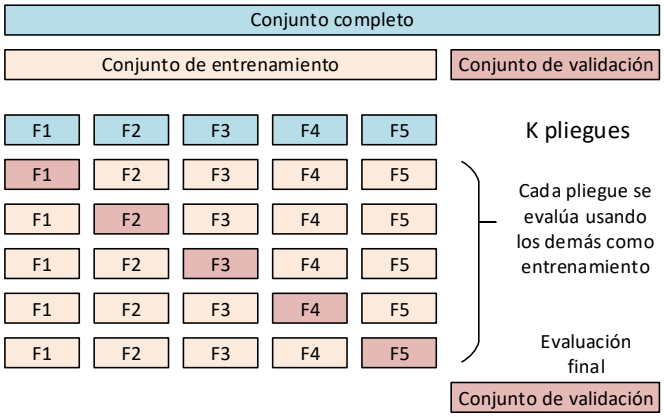


Figura 22. Diagrama de partición de conjuntos

3.3.2 Normalización

Hay diferentes modos de normalizar y según el método luego se podrá aplicar con mayor o menor éxito ciertos algoritmos. Normalizar significa, adaptar los valores de la variable para que estén en un rango definido. En general se obtienen mejores resultados si todas las variables están escaladas, ya que todas tendrán pesos similares a la hora de computar distancias y otros estadísticos.

- Normalización entre el mínimo y el máximo. Este proceso va a escalar la señal entre dos valores, normalmente 0 y 1 en función del valor mínimo y máximo de la señal. De este modo nos desprendemos de la magnitud, pero tiene cierta peligrosidad, como que, si la señal es muy estable, puede amplificar el ruido y parecer que no lo es. La forma de calcularla es restar a cada muestra, el valor mínimo y dividirlo entre el rango observado, es decir, el máximo menos el mínimo.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- Normalización por la media y desviación estándar. En este método a cada muestra se le resta la media de la variable y se le divide por la desviación típica. Este método concentra la información entorno a cero, haciendo que la media sea cero.

$$X_{norm} = \frac{X - X_{mean}}{X_{stddev}}$$

Es importante ser capaz de realizar la transformación inversa, ya que en muchas ocasiones los resultados los deberemos expresar en las unidades originales previas al escalado.

Usaremos la normalización por mínimo y máximo ya que utilizaremos redes neuronales que trabajan mejor con los datos escalados entre cero y uno.

3.3.3 Algoritmos de minería de datos

Para la realización de predicciones usaremos los siguientes métodos de regresión.

- *Random Forest Regressor:* O “bosques aleatorios” es uno de los métodos de *bagging* o combinación paralela de clasificadores más conocidos. Se trata de una combinación de árboles predictores donde cada uno se especializa en un subconjunto de los datos. La valoración final del algoritmo se conforma a partir de los votos individuales de cada árbol. Se pueden utilizar tanto para clasificación como para regresión, es rápido y capaz de lidiar con una gran cantidad de datos y de alta dimensionalidad.
- *Gradient Booting Regression:* Se trata de un método de *boosting* que se basa también en la utilización de numeroso predictores débiles, al igual que *Random Forest,* pero a diferencia de éste, cada árbol dispone de la información sobre los errores que ha cometido el anterior de modo que ayuda a afinar el resultado.
- *Support Vector Machine.* Las máquinas de soporte vectorial, se basan en la búsqueda de hiperplanos que separen los conjuntos trasladándolos a una dimensionalidad superior. Existe una versión que permite aplicarse a la regresión.
- *K Nearest Neighbor.* Originalmente es un método de clasificación que asocia una etiqueta basándose en la similitud de la muestra con los vecinos más cercanos. También sirve como método de predicción.
- *Neural Network*. Las redes neuronales son algoritmos que son capaces de aprender de la experiencia, extraer las características principales de los datos y generalizar posteriormente a conjuntos nuevos.

3.3.4 Búsqueda de hiperparámetros

Hay algoritmos con parámetros de ejecución que influyen en el ajuste que es capaz de conseguir. En algunos casos pueden ser sencillos de establecer, pero en otros casos, no es fácil saber a priori cuales son los que obtendrán mejores resultados.

Una forma de encontrarlos es mediante una técnica llamada, “búsqueda de hiperparámetros” basada en establecer un *grid*, o matriz con una combinación de estos parámetros. Se entrenarán todos ellos y se dará con el mejor de ellos según la métrica especificada. El método es por fuerza bruta, es decir, se prueban todos y se selecciona el mejor.

Hemos creado un algoritmo para buscar estos parámetros representando el resultado en un mapa de calor y un diagrama de cajas con la distribución de los resultados obtenidos en cada prueba. La Figura 23, muestra como ejemplo los resultados de buscar los parámetros óptimos de un algoritmo *Random Forest* en cuanto al número de árboles y la profundidad máxima.

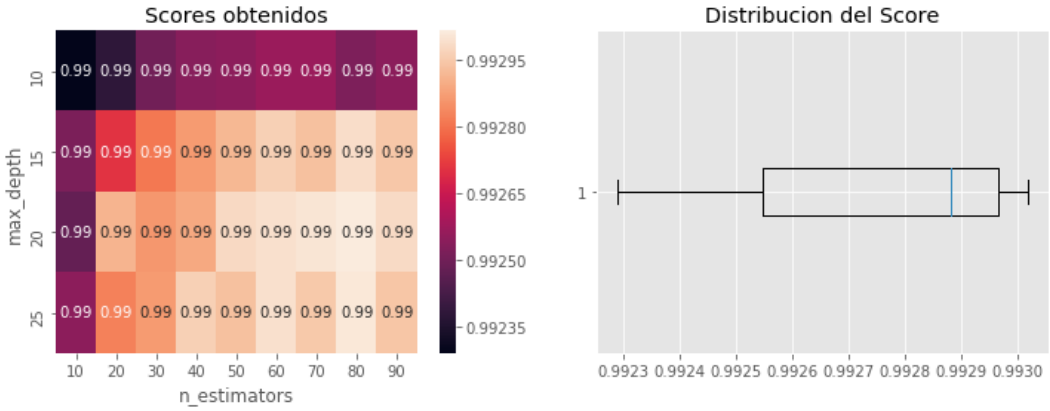


Figura 23. Resultados de la búsqueda de hiperparámetros de RF

La búsqueda se realiza para todos los algoritmos empleados, de esta forma nos aseguramos de que el algoritmo se aplica de forma óptima. Los gráficos y métricas obtenidas para cada algoritmo en la predicción de la potencia y *pitch*, se incluyen en el anexo II mientras que los resultados finales se exponen en los puntos 3.3.5 y 3.3.6 respectivamente.

3.3.5 Predicción de la potencia

Para la selección de variables candidatas a la predicción de potencia, preseleccionamos aquellas cuyo índice de correlación de *Pearson* es superior a 0.5, aunque esto no es una regla estricta.

El algoritmo *ramdom forest*, es capaz de ofrecernos un índice de qué características son más relevantes a la hora realizar una predicción. Para esta prueba, se suele introducir como variables independientes las características candidatas junto con una aleatoria como control.

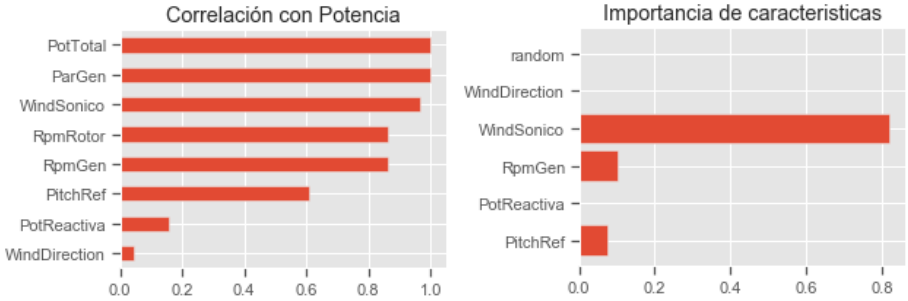


Figura 24. Variables candidatas para predicción de potencia

Se observa que la dirección de viendo no es relevante, aunque introduciéndola en el modelo mejora unas décimas como se verá más adelante. Aplicando cierta lógica desde el conocimiento del sistema, la lista de modelos para predicción de la potencia queda según la Tabla 7.

Nombre	Variables independientes	Variable dependiente
W->Pot	WindSonico	PotTotal
WRP->Pot	WindSonico + RpmGen + PitchRef	PotTotal
WRPD->Pot	WindSonico + RpmGen + PitchRef +WindDirection	PotTotal

Tabla 7. Modelos para la predicción de potencia

Computando más de 100 modelos y después de varias horas, se arrojan los resultados reflejados en la Tabla 8. El resultado gráfico de la selección de hiperparámetros, las distribuciones de residuos y las predicciones de los principales modelos, pueden encontrarse en el anexo II.

En los nombres de los modelos usamos la siguiente nomenclatura: A_CCC->P siendo A el algoritmo, C las iniciales de los predictores usados y P la variable a predecir.

Modelo	CV R2	MSE	RMSE	MAE	MAPE	TST R2
RF_W->Pot	0,99443	0,00045	0,02118	0,01321	62,38791	0,99463
KNN_W->Pot	0,99484	0,00042	0,02049	0,01290	61,39661	0,99497
SVR_W->Pot	0,99482	0,00042	0,02055	0,01253	43,94289	0,99494
GB_W->Pot	0,99472	0,00043	0,02073	0,01308	64,83367	0,99485
LR_W->Pot	0,93027	0,00557	0,07465	0,05486	811,09717	0,93324
NN_W->Pot	0,99478	0,00043	0,02069	0,01370	75,08820	0,99487
DNN_W->Pot	0,99447	0,00045	0,02111	0,01325	37,15605	0,99466
RF_WRP->Pot	0,99849	0,00011	0,01065	0,00544	50,67733	0,99864
KNN_WRP->Pot	0,99826	0,00013	0,01146	0,00580	43,71671	0,99843
SVR_WRP->Pot	0,99796	0,00016	0,01254	0,00643	40,48510	0,99812
GB_WRP->Pot	0,99850	0,00011	0,01058	0,00542	53,54754	0,99866
LR_WRP->Pot	0,93229	0,00544	0,07376	0,05308	672,93430	0,93481
NN_WRP->Pot	0,99765	0,00017	0,01296	0,00773	61,28640	0,99799
DNN_WRP->Pot	0,99724	0,00020	0,01431	0,00836	60,09850	0,99755
RF_WRPD->Pot	0,99859	0,00010	0,01017	0,00528	51,41877	0,99876
KNN_WRPD->Pot	0,99805	0,00015	0,01221	0,00623	50,10767	0,99822
SVR_WRPD->Pot	0,99803	0,00015	0,01235	0,00637	39,81602	0,99817
GB_WRPD->Pot	0,99864	0,00010	0,00999	0,00523	48,85665	0,99881
LR_WRPD->Pot	0,93239	0,00543	0,07369	0,05300	665,17593	0,93495
NN_WRPD->Pot	0,99782	0,00014	0,01183	0,00685	38,76251	0,99832
DNN_WRPD->Pot	0,99784	0,00014	0,01175	0,00646	63,45316	0,99835

Tabla 8. Ajuste de los modelos de potencia

Sorprendentemente *Gradient Boosting* (GB_WRPD->Pot) y *Random Forest* (RF_WRPD->Pot) han conseguido los mejores resultados utilizando cuatro predictores (viento, revoluciones, *pitch* y dirección), por encima incluso de las redes neuronales, aunque también hay que decir que la diferencia ha sido minina ya que todos los regresores a excepción de la regresión lineal han conseguido unos resultados muy buenos. El resultado de la comparación se muestra en la Figura 25 ordenados por MAE y RMSE.

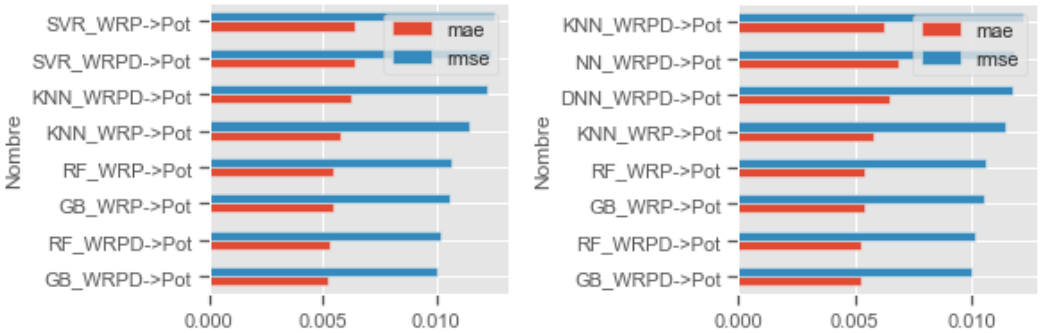


Figura 25. Mejores resultados en la predicción de potencia

Un MAE de 0.005 en la predicción de la potencia equivale a un error medio de 10 W, que, en un fondo de escala de 2000, es una buena aproximación.

Como ejemplo de la aproximación y residuos resultantes de un modelo de soporte vectorial, se incluye la Figura 26.

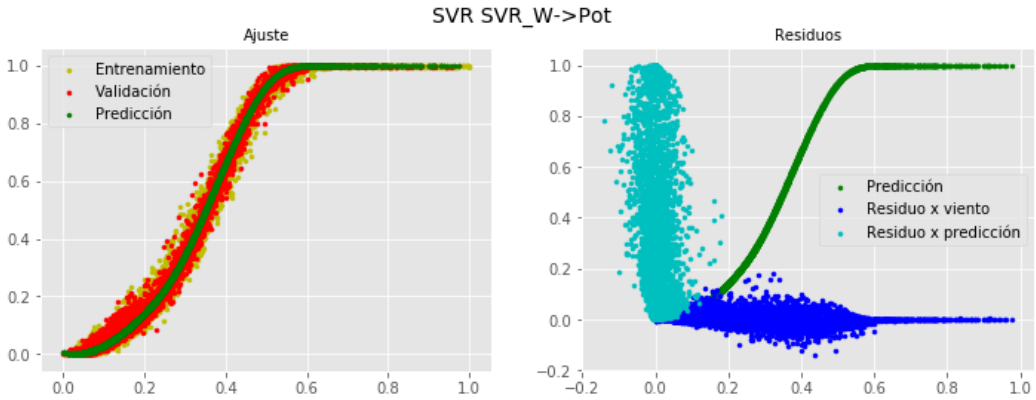


Figura 26. Ajuste y residuos en la predicción de potencia

3.3.6 Predicción del pitch

Siguiendo el mismo procedimiento para la predicción de la variable de *pitch*, que hemos usado para la predicción de potencia, las características seleccionadas se indican en la Figura 27

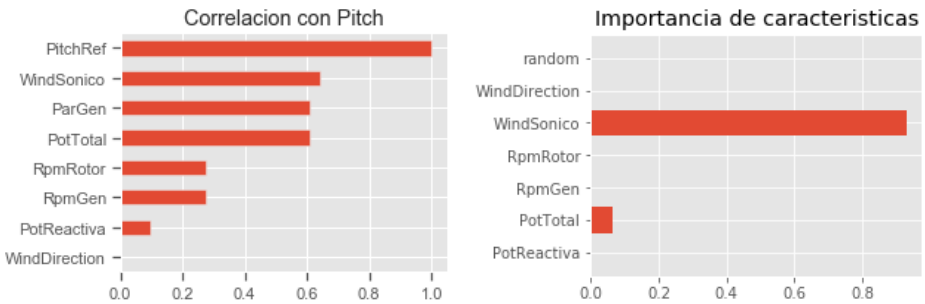


Figura 27. Variables candidatas para la predicción de Pitch

La lista de modelos para la predicción del *pitch* se muestra en la Tabla 9

Nombre	Variables independientes	Variable dependiente
W->Pitch	WindSonico	PitchRef
WR->Pitch	WindSonico + RpmGen	PitchRef
WRP->Pitch	WindSonico + RpmGen + PotTotal	PitchRef

Tabla 9. Modelos para la predicción del Pitch

En general, las predicciones son bastante buenas, excepto para la regresión lineal, como se observa en la Tabla 10 .

Modelo	CV R2	MSE	RMSE	MAE	MAPE	TST R2
RF_W->Pitch	0,98983	0,00012	0,01116	0,00575	17,22150	0,98986
KNN_W->Pitch	0,99105	0,00011	0,01051	0,00560	17,11895	0,99101
SVR_W->Pitch	0,99112	0,00011	0,01049	0,00548	15,25328	0,99104
GB_W->Pitch	0,99101	0,00011	0,01061	0,00569	18,04693	0,99084
LR_W->Pitch	0,40931	0,00723	0,08500	0,06334	253,55549	0,41176
NN_W->Pitch	0,99064	0,00012	0,01087	0,00578	17,11422	0,99039
DNN_W->Pitch	0,98977	0,00011	0,01058	0,00611	18,08869	0,99088
RF_WR->Pitch	0,99301	0,00009	0,00942	0,00366	9,05029	0,99278
KNN_WR->Pitch	0,99353	0,00008	0,00868	0,00351	8,92294	0,99387
SVR_WR->Pitch	0,99337	0,00008	0,00878	0,00372	11,90463	0,99372
GB_WR->Pitch	0,99306	0,00008	0,00904	0,00365	9,88956	0,99335
LR_WR->Pitch	0,75004	0,00313	0,05592	0,04632	156,08659	0,74545
NN_WR->Pitch	0,98920	0,00009	0,00949	0,00434	13,59275	0,99267
DNN_WR->Pitch	0,99231	0,00014	0,01175	0,00452	8,85289	0,98876
RF_WRP->Pitch	0,99543	0,00006	0,00745	0,00280	5,76981	0,99548
KNN_WRP->Pitch	0,99539	0,00005	0,00742	0,00292	6,38489	0,99552
SVR_WRP->Pitch	0,99536	0,00005	0,00716	0,00296	6,46704	0,99583
GB_WRP->Pitch	0,99516	0,00005	0,00741	0,00295	6,98923	0,99553
LR_WRP->Pitch	0,75201	0,00309	0,05563	0,04601	154,80872	0,74804
NN_WRP->Pitch	0,99365	0,00008	0,00871	0,00416	10,38556	0,99382
DNN_WRP->Pitch	0,99277	0,00008	0,00877	0,00547	16,13558	0,99374

Tabla 10. Ajuste de los modelos de pitch

El resultado comparativo de los mejores modelos está en la Figura 28 ordenados por MAE y RMSE.

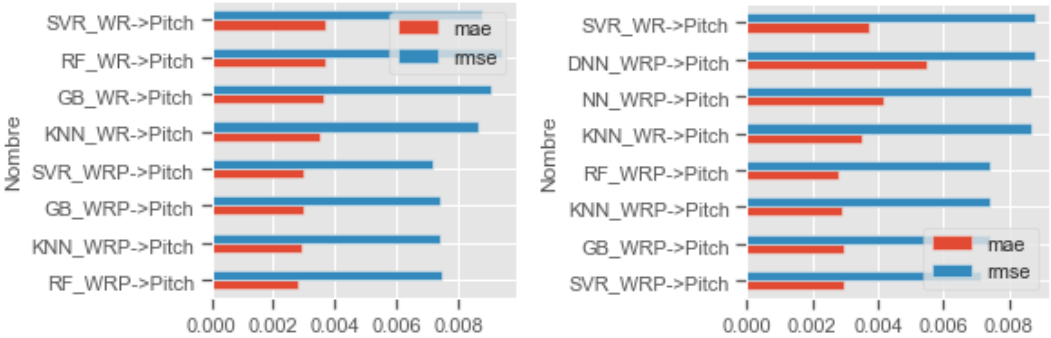


Figura 28. Mejores resultados en la predicción de Pitch

Como ejemplo, la aproximación y residuos resultantes de un modelo de red neuronal con dos capas ocultas reflejan en la Figura 29.

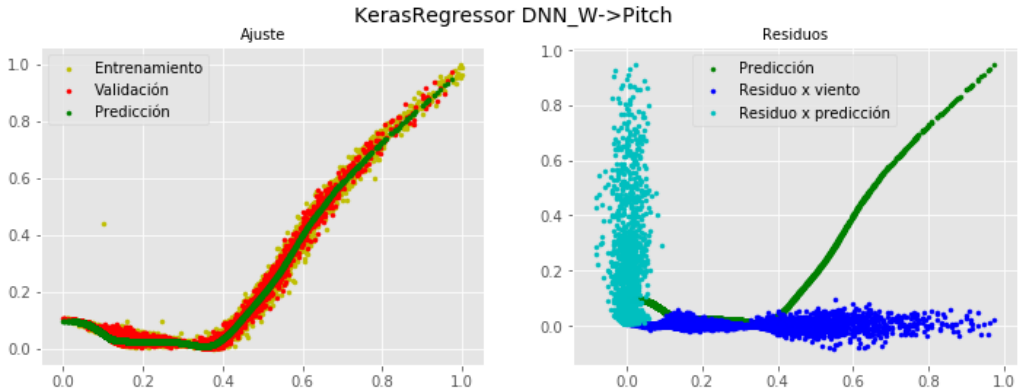


Figura 29. Ajuste y residuos en la predicción de Pitch

3.3.7 Tiempo de computación

Un factor que a veces se pasa por alto es el tiempo empleado por los modelos en las operaciones de entrenamiento y predicción. Si bien en determinados entornos de desarrollo y análisis no es crítico debido a que se dispone de una infraestructura con generosas capacidades de computación, hay que tenerlo en cuenta cuando la ejecución puede planearse para llevarse a cabo en arquitecturas más modestas o hechas a medida.

Un ejemplo es el uso de estos algoritmos en hardware especifico o dispositivos de bajo coste orientados a la industria 4.0 e internet de las cosas. En este entorno las predicciones pueden realizarse en campo y de forma desatendida, en forma de sensores integrados e inteligentes.

Aunque el coste de computación no es tan trivial como medir el tiempo de ejecución, ya que intervienen factores como la paralelizacion, en la Figura 30 se incluyen los tiempos empleados para el entrenamiento y para la predicción de los modelos WRP->Pitch en un PC de escritorio con procesador i5 y 8GB de RAM, con la mejor configuración obtenida en el proceso de selección de hiperparámetros.

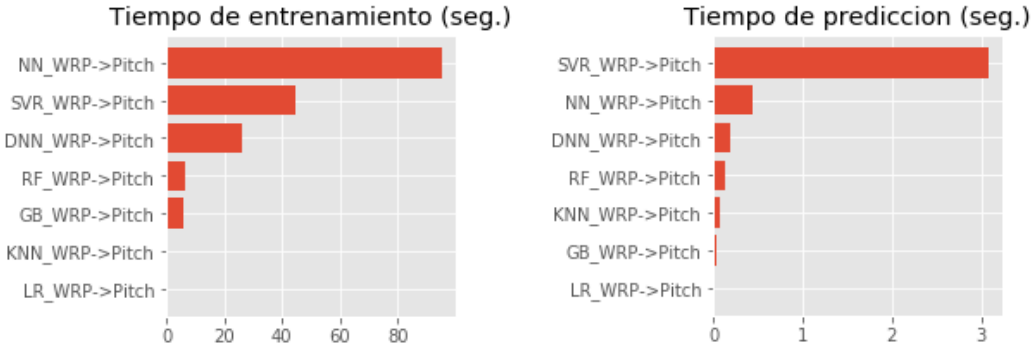


Figura 30. Tiempos de computación de los modelos

Como puede observarse, el entrenamiento es muy lento en las redes neurales y máquinas de soporte vectorial comparado con los métodos basados en árboles, pasando de más de un minuto, a segundos.

En cuanto a la predicción, los tiempos son muy inferiores, de apenas unos segundos en todos los modelos, siendo la máquina de soporte vectorial la que más tiempo consume.

En cuanto a la velocidad de KNN es engañosa, ya que es un método perezoso que no tiene entrenamiento, y los vecinos se calculan en tiempo de predicción. Además, este método precisa de los datos de entrenamiento en el tiempo de realizar las predicciones por lo cual su uso queda descartado para muchas aplicaciones.

3.4 Detección de anomalías

Para determinar en qué grado los algoritmos de detección son efectivos, vamos a simular diferentes fallos en diferentes grados, y aplicar estos algoritmos de modo que podamos ver su eficacia en función del grado de anomalía.

Las turbinas modernas disponen de sistemas CMS con sensores específicos que ayudan a una predicción más precisa, pero en nuestro caso utilizaremos solo variables de control, siendo interesante por su posible utilización en máquinas que no cuenten con CMS y que corresponden con modelos antiguos que ahora son objeto de programas para prolongar su vida útil.

Uno de los mayores problemas que vamos a encontrar es la alta varianza observada en la zona de vientos media. Como puede observarse en la Figura 31, a ciertas velocidades de viento puede registrarse un gran abanico de potencias, entendiendo que esto es debido a que las máquinas cuentan con inercia y periodos de adaptación a los vientos y modos de funcionamiento que no quedan registrados en las variables de control de las que disponemos, y por lo tanto, la predicción cuenta con un peor ajuste en esas zonas que se traduce en una distribución de residuos no homogénea.

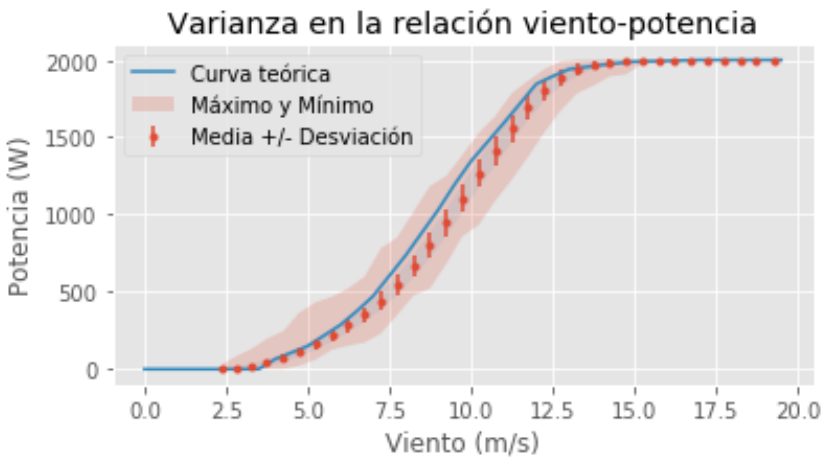


Figura 31. Varianza en la relación viento-potencia

Por otro lado, la anomalía que vamos a trabajar es la anomalía contextualizada, es decir, la observación de ciertos valores que pueden ser validos o habituales en ciertas situaciones, pero no en otras. Los aerogeneradores trabajan 24x7 y son máquinas pesadas que alternan modos de funcionamiento y donde podemos observar una alta varianza a corto plazo, de modo que la detección de una anomalía la haremos a nivel de día, es decir, trataremos de averiguar si el comportamiento de un día completo ha sido normal o anómalo.

3.4.1 Simulación del fallo

El conjunto de datos del que disponemos asumimos que se encuentra en la normalidad y no ha registrado ningún tipo de anomalía.

En este caso, al no disponer de datos de fallo, tenemos que simularlos. Una vez tenemos un modelo que describe la normalidad, alteraremos el conjunto de test de forma controlada, provocando un desajuste con arreglo al funcionamiento de la máquina, para ver en qué grado los mecanismos de detección de anomalías se activan.

3.4.2 Método a aplicar

Para la detección de anomalías usaremos el método estadístico descrito en el punto 2.2.1.2 e ilustrado en la Figura 17.

En el entorno que nos encontramos, las anomalías pueden ser principalmente de dos tipos: Sobrevenidas, es decir fallos drásticos que ocurren de forma instantánea y que supone la parada de emergencia por los sistemas de seguridad, o degradaciones del rendimiento progresivas que se dan en ciertas condiciones. Este es el caso en el que trabajaremos.

El objetivo es conseguir un sistema clasificador para que, a partir de los datos adquiridos de un día completo, se le pueda aplicar una etiqueta como día normal o día anómalo. Una vez tengamos el clasificador lo

someteremos a evaluación para analizar su precisión. Las principales métricas que tenemos a disposición son las propias de una matriz de confusión:

- Verdaderos positivos (TP): Los días anómalos que son identificados como anómalos.
- Falsos negativos (FN): Los días anómalos que se identifican como normales.
- Falsos positivos (FP): Los días normales que se identifican como anómalos.
- Verdaderos negativos (FN): Los días normales que se identifican como normales.

A partir de esos cuatro valores se obtienen otras dos métricas interesantes cuyas formulas se indican a continuación:

- Sensibilidad: Es la capacidad para encontrar los días anómalos. Esta métrica oscila entre 0-1 siendo 1, la capacidad plena de encontrar las anomalías.
- Precisión: Es la capacidad para no clasificar mal un día que es normal. También oscila entre 0-1, y 1 indica que ningún día normal se ha etiquetado como anómalo.

$$Sensibilidad = \frac{TP}{TP + FN} \qquad Precisión = \frac{TP}{TP + FP}$$

De la curva precisión-sensibilidad se puede obtener la valoración del modelo, lo idóneo es una sensibilidad alta y una precisión alta lo más cercano ambas a 1.

Partiremos de un modelo de predicción de potencia, pero en este caso debe ser entrenado de diferente manera para garantizar que en el proceso de validación se enfrenta a datos completamente nuevos. Usaremos los días pares para entrenamiento y los días impares para validación. De este modo los modelos disponen de datos distribuidos para entrenar que están repartidos a lo largo de todo el dataset en condiciones de normalidad.

Para la validación, crearemos un conjunto compuesto por:

- Los días reservados para validar los modelos en estado de normalidad, la etiqueta asignada es la de “normal”
- Los mismos días anteriores pero modificados con una simulación de fallo, la etiqueta asignada es la de “fallo”

De este modo someteremos a prueba el clasificador con un conjunto balanceado con 50% de días normales y 50% días anómalos, con el que podremos probar si clasifica bien los días con fallo, pero también si se equivoca con los días normales. Ningún dato de estos ha sido visto previamente por los modelos.

3.4.3 Pérdidas en el tren de potencia

El fenómeno de perdida de potencia es el caso en el que se genera menos energía que la habitual con motivos de fallos en el generador o en la multiplicadora.

Partiendo del conjunto sano, modificaremos la potencia total generada multiplicanda por un factor de amortiguación, en primera instancia, igual en todo el rango de la potencia.

Una pérdida de potencia afecta la variable de potencia total, pero también a otras relacionadas mecánicamente o eléctricamente, como el par, la potencia del estator y otras. Pero no afecta a las usadas en los modelos desarrollados con anterioridad. Es decir, viento, revoluciones de rotor, *pitch* y la orientación de la maquina permanecen independientes ya que son operados por el control en busca de la máxima potencia, que en el dataset modificado no llegará a alcanzar.

Estudiaremos si una pérdida del 5% o del 10% en la producción es detectable por algoritmos de minera de datos. Hay que tener en cuenta que un 5% en un fondo de escala de 2.000W solo supone 100W en el punto de máxima potencia, algo que puede ser fácil de detectar en máxima producción, pero puede quedar enmascarado sobre todo en la zona de carga parcial.

Comparamos la potencia en el dataset modificado con una potencia al 90% frente a la potencia original para los días asignados a validación en la Figura 32:

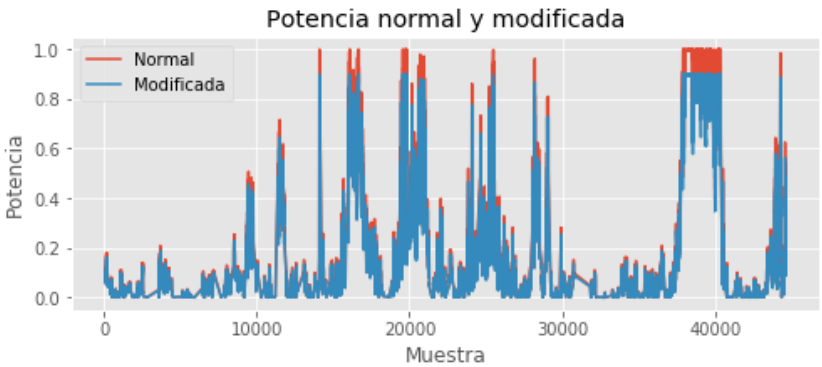


Figura 32. Potencia real y modificada

Aplicando la técnica descrita en el punto 2.2.1.2, partimos de un modelo de predicción de potencia usando las variables WRPD entrenado con datos de normalidad. Realizamos una predicción sobre el conjunto de validación que contiene tanto días normales como con fallo simulado. Observamos los residuos en la Figura 33.

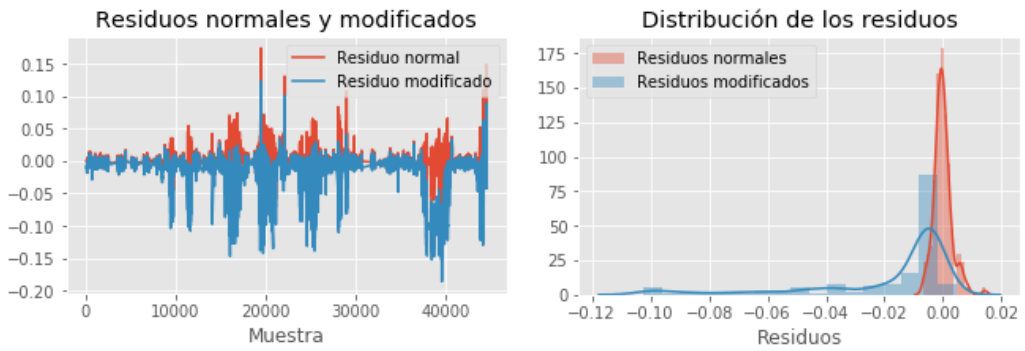


Figura 33. Residuos en la predicción de la potencia

Como puede observarse hay una gran varianza en los residuos y están distribuidos de forma desigual en función de la potencia predicha. Esto es debido a que sobre todo en la zona de vientos medios, puede alcanzarse un amplio abanico de potencias, debido posiblemente a periodos transitorios de la máquina, por eso el ajuste de los modelos predictivos es peor en esta zona.

Como era de esperar, a mayor potencia, mayor es la diferencia entre los residuos de la normalidad y la anomalía, como se muestra en la Figura 34.

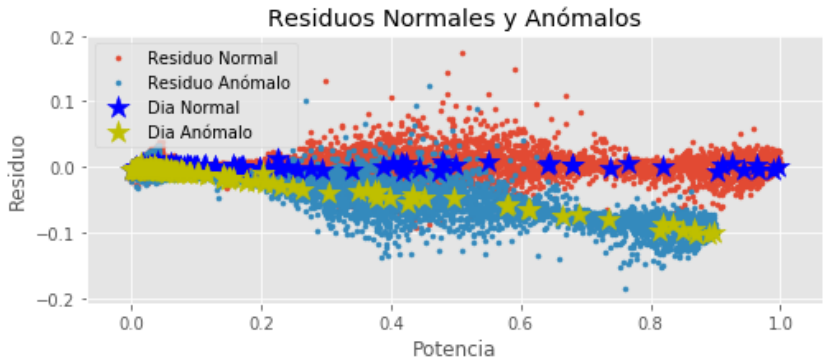


Figura 34. Residuos en función de la potencia

Tenemos que encontrar la forma de diferenciar los residuos procedentes de los días normales con respecto a los días anómalos. Normalmente se calcula un umbral en base a la distribución de los errores en el conjunto de entrenamiento.

Para intentar evitar el solapamiento de los residuos, reduciremos la varianza promediando a valores diarios, de esta forma resultara más sencillo encontrar un umbral que separe las muestras con efectividad, así pues, para clasificar un día como normal o anómalo, simplemente se comparara la media de los residuos diarios con respecto al umbral: si el residuo es superior se clasifica como día anómalo.

Para ilustrar la efectividad a diferentes valores de umbral hemos creado un algoritmo que calcula la sensibilidad y la precisión para determinar el punto óptimo según se observa en la Figura 35.

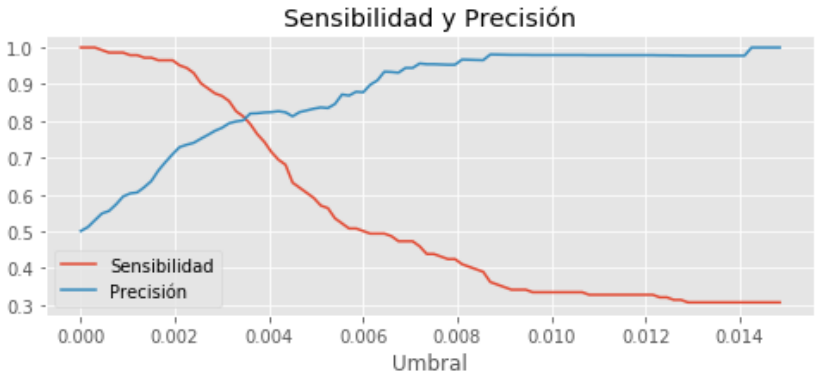


Figura 35. Curvas de sensibilidad y precisión

Vemos que el resultado no es muy bueno. La sensibilidad es sumamente baja para una precisión aceptable, eso se traducirá en que habría que elegir entre un clasificador que pase por alto las anomalías o que identificara de forma errónea muchos días correctos.

El origen de esta imprecisión puede ser, según se observa en los residuos, a que, a valores muy pequeños de potencia, la diferencia es mínima y poco significativa. Es donde más errores se producen.

Para solucionar esto aplicaremos dos umbrales adicionales:

- Umbral de potencia a partir del cual se evaluarán los días. Es decir, para evaluar una desviación de potencia, necesitaremos un día que disponga de una potencia mínima.
- Umbral de soporte. Para evaluar ese día deberá contar con un número mínimo de muestras. Como trabajamos con medias diarias, esto descartaría aquellos días en los que por ejemplo solo ha estado en producción unos minutos, en cuyo caso el promedio puede falsear la realidad.

Aplicando un soporte de 36 muestras, es decir, que tengamos al menos 6 horas de datos en el día, y diferentes valores para el umbral de potencia conseguimos las curvas de precisión y sensibilidad de la Figura 36.

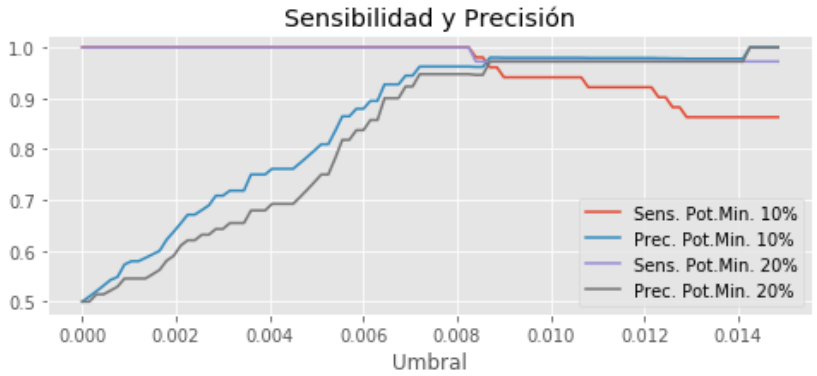


Figura 36. Sensibilidad y precisión en la detección de perdida de potencia

Como vemos para un umbral de potencia de 10%, es decir, aplicando el clasificador solo a los días que supere una producción media de 200W, se obtienen unos resultados muy buenos.

Y si se aplica un mínimo del 20% se consigue una precisión del 100%, es decir, ningún día normal se clasifica como anómalo, mientras que se consigue una sensibilidad del 97%, es decir solo un 3% de los días anómalos se clasifican como normales. Los datos exactos se muestran en la Tabla 11. Puede observarse como al aumentar la potencia mínima necesaria para someter el día a clasificación, quedan fuera aquellos en los que más errores se comete.

Potencia mínima	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

Tabla 11. Sensibilidad y precisión en la detección de perdida de potencia

3.4.4 Desalineación del rotor

El efecto de una desalineación es que para producir la potencia se necesita más viento del habitual, se produce por diversas causas y el resultado es que el rotor no se encuentra correctamente alineado al viento.

Para simularla incrementaremos la variable de viento en un factor, pero recortando los valores que superen el máximo observado durante el entrenamiento al propio valor máximo, mientras que el resto de variables permanecen intactas.

Simularemos un incremento del 10% de viento para mantener la producción, el dataset original y modificado lo observamos en la Figura 37 y graficamos el viento frente a la potencia, para apreciar mejor como la curva del dataset modificado se mueve hacia la derecha reflejando que necesita más viento para producir la misma cantidad de energía.

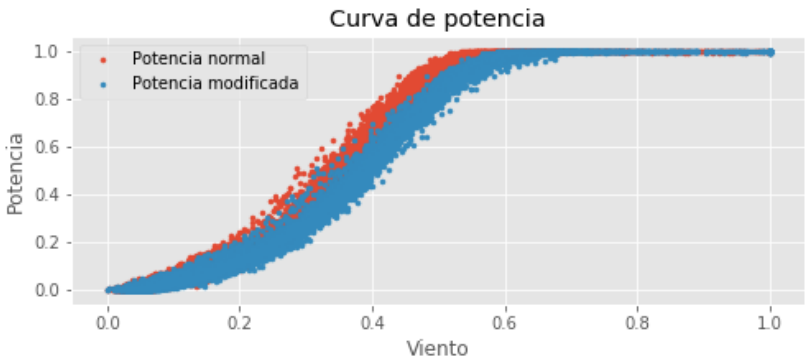


Figura 37. Curva de potencia con conjunto normal y modificado

Hay que tener en cuenta, que para niveles bajos de viento va a ser difícil diferenciar las predicciones, y a niveles de viento muy altos de viento, también, ya que una vez superado el valor nominal de viento, la producción es constante al nivel máximo, como se observa en la Figura 37.

El procedimiento para detectar anomalías, de forma general es el mismo que el aplicado en el punto 3.4.3 simulando una pérdida de potencia, solo que en este caso la variable modificada es una de las variables independientes.

Comenzaremos con un modelo básico RF_W->Pot con solo la variable de viento para predecir la potencia. Comparamos la distribución y forma de los residuos en la Figura 38.

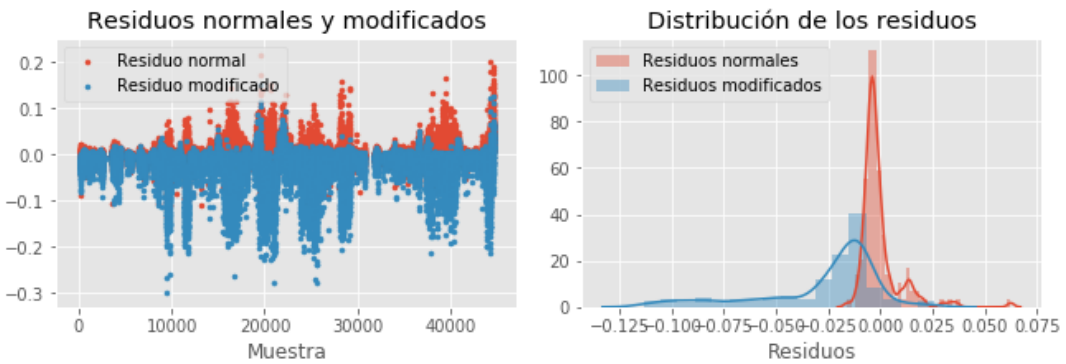


Figura 38. Residuos es la predicción de potencia

La vista de los residuos en función del viento y la potencia corresponde la Figura 39.

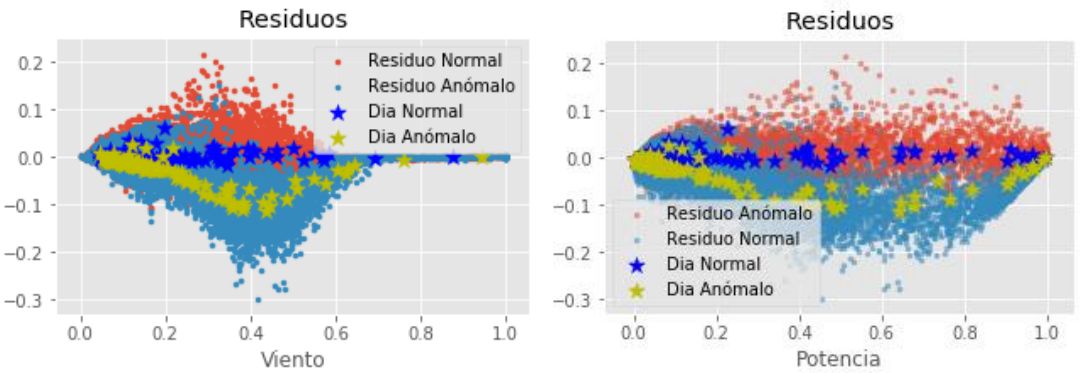


Figura 39. Residuos en función del viento y la potencia

Aplicaremos una media diaria para reducir la varianza y volveremos a calcular un umbral para diferenciar los residuos anómalos de los normales.

En este caso, la diferenciación puede ser más complicada, ya que la anomalía solo puede ser detectada en los valores de viento intermedios. Para ello el mismo umbral que en el caso de pérdida de potencia se aplicó para discriminar los días con una potencia baja, aquí lo aplicaremos también para discriminar los días con potencia alta, en igual medida.

En la Figura 40 se muestra un umbral de 0.015. Vemos que hay días anómalos que entran en la zona de clasificación normal, no obstante, no computan al encontrarse fuera del umbral mínimo de potencia o de soporte.

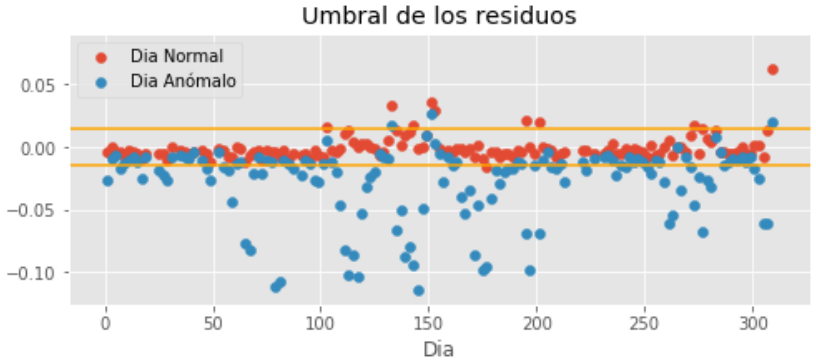


Figura 40. Residuos promediados y umbral seleccionado

De este modo las curvas de sensibilidad y precisión quedan como en la Figura 41 para distintos valores de umbral.

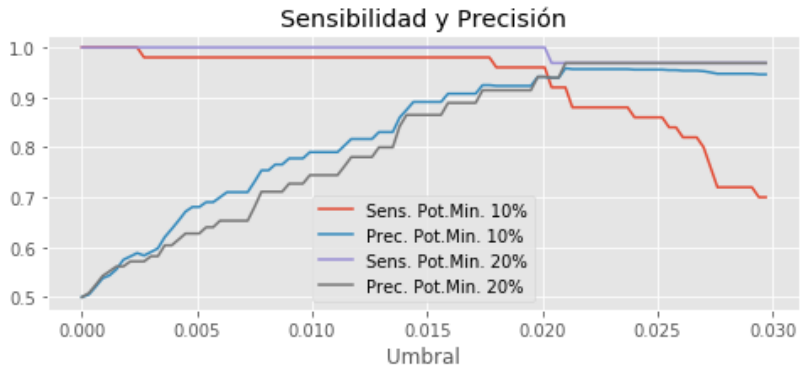


Figura 41. Sensibilidad y precisión en la detección de desalineación

Los resultados numéricos se muestran en la Tabla 12. Y como puede observarse también son unos resultados buenos, obteniendo umbrales que resultan en un 97% de precisión y 97% de sensibilidad.

Potencia mínima	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

Tabla 12. Sensibilidad y precisión en la detección de desalineación

3.4.5 Uso de *autoencoder* para la detección

Vamos a explorar el posible uso de una red neuronal denominada autoencoder. Según se describe en el punto 2.2.1.1 este tipo de redes reducen la dimensionalidad de entrada a una inferior para luego reconstruir la señal de nuevo a la dimensión original. La red descarta la información no relevante ya que se entrena teniendo como función a minimizar, el error de construcción.

Crearemos una red acorde a la Figura 42. La idea es introducir las mismas variables empleadas en el punto 3.4.3 para la predicción de potencia, junto con la potencia, como variables de entrada y salida.

En una primera etapa se reducirá la dimensionalidad de 5 características a cuatro, y posteriormente a tres, para reconstruirla de nuevo a cuatro y finalmente a cinco. Usaremos funciones de transferencia sigmoide. Se entrenará la red con los datos de los días pares que representan la normalidad, y se observará el error de reconstrucción cometido sobre los conjuntos de validación formados por los días impares, y que incluirán el conjunto normal y el conjunto con error simulado.

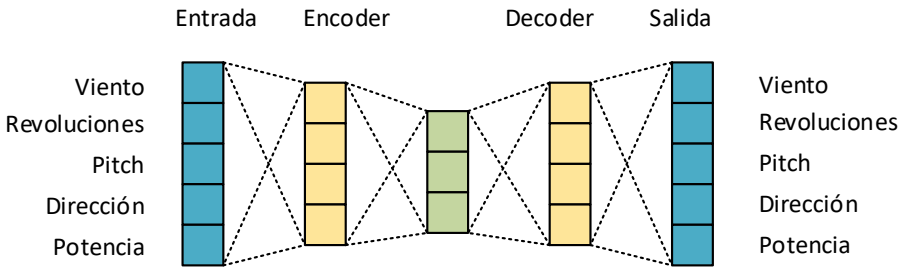


Figura 42. Esquema de autoencoder para la detección de anomalías

En la Figura 43 podemos ver el error de reconstrucción de ambos conjuntos. Los datos se presentan como promedio de los días completos. Se puede observar las diferencias son mínimas e insuficientes para poder diferenciar con éxito los días anómalos.

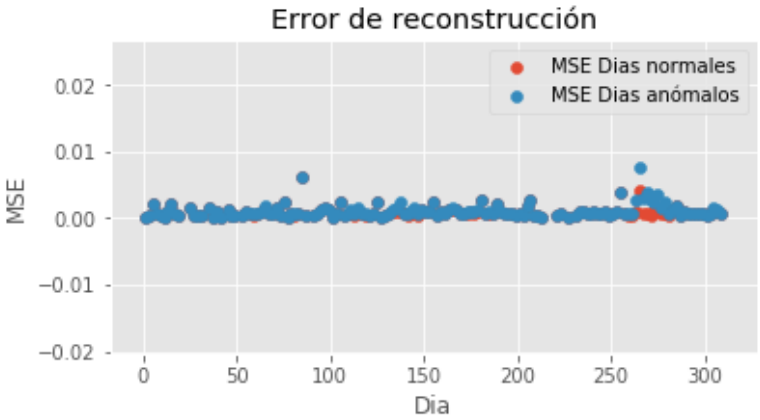


Figura 43. Error de reconstrucción de los días normales y modificados

No obstante, si observamos con detalle las señales reconstruidas en la Figura 44 en el caso de alimentar la red con los datos modificados, vemos que la dirección de viento ha quedado prácticamente anulada, posiblemente por ser la menos significativa, mientras que la señal de potencia es reconstruida acorde a lo aprendido a los datos de entrenamiento. Es decir, la red se comporta como un modelo de predicción.

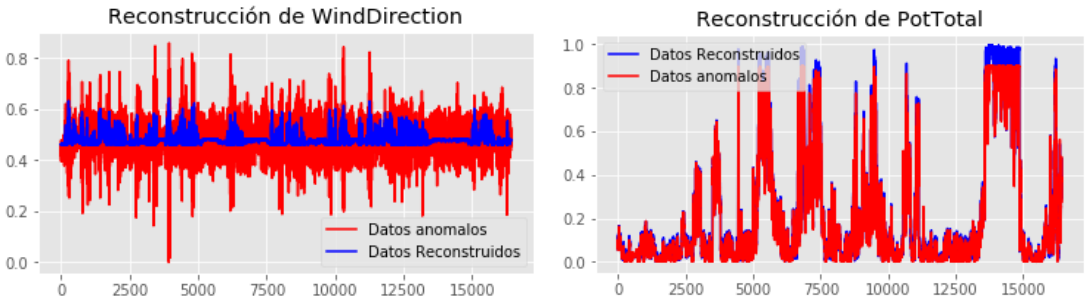


Figura 44. Reconstrucción de la dirección y la potencia

En este caso, si analizamos la diferencias entre la entrada y la salida de la potencia, podríamos aplicar el mismo método del umbral utilizado en el punto 3.4.3 para intentar discriminar los días. La Figura 45 muestra los resultados de la clasificación, para diferentes niveles de potencia mínima.

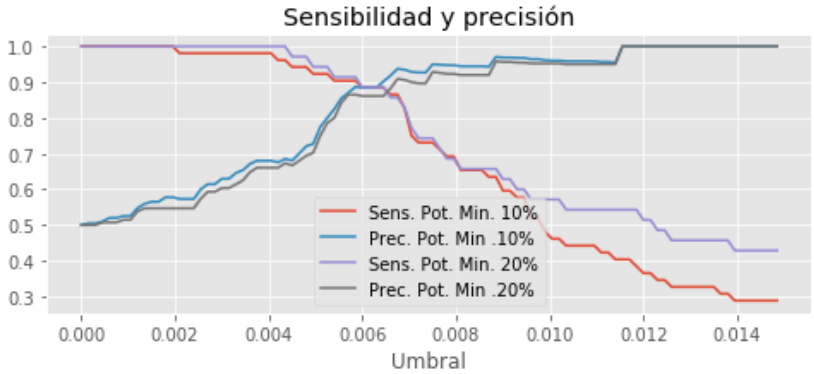


Figura 45. Sensibilidad y precisión obtenidos por autoencoder

Como se puede observar el resultado es peor que el obtenido mediante el metodo estadístico donde se podia obtener un 100% de precisíon y 97% de sensibilidad.

En este caso, si queremos una sensibilidad o precisión realmente alta, vemos comprometida la otra. Aun así, es un buen resultado si se adopta como solucion la eleccion del umbral donde sensibilidad y precision adquieren un compromiso y se igualan entorno al 90%, según se aprecia en la Figura 45.

3.5 Predicción de energía

3.5.1 Aplicación de modelos para la previsión a corto plazo

En este apartado aplicaremos varios métodos con el fin de obtener una predicción de la producción de energía basada en los datos recientes.

Dado que el conjunto de datos está formado por muestras diezminutales, usaremos la información de los últimos 30 o 60 minutos para predecir la potencia generada en los próximos 10 minutos, que es lo mismo que utilizar las 3 o 6 últimas muestras para predecir la siguiente.

Usaremos el conjunto completo, sin seleccionar solo las partes de producción de energía como hicimos para la detección de anomalías.

En este caso la producción de energía se considera una serie temporal que no depende de ninguna otra, y tiene que ser continua en el tiempo, sin saltos ni cortes que la interrumpan, ya que algunos de los modelos que usaremos utilizarán secuencias dependientes del tiempo. Por ello en este caso, los conjuntos de entrenamiento y validación estarán formados, por el 50% de las muestras iniciales, para el conjunto de entrenamiento, y el 50% restantes para validación, de este modo se hace un reparto equitativo y sin cortes, de la potencia observada.

Los métodos que exploraremos serán auto regresivos formados por diferentes algoritmos de regresión, redes neuronales LSTM y el método ARIMA.

3.5.1.1 Modelo auto regresivo

Estos modelos utilizan los datos pasados para predecir el futuro según se ilustra en la Figura 46, creando una serie de características con datos desfasados *n* muestras. El concepto de ventana es el número de muestras pasadas que se utiliza para la creación de las características, y la distancia de predicción es la muestra futura que se predice.

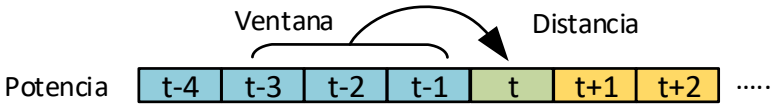


Figura 46. Modo de operación de una auto regresión

Con estas nuevas características, se pueden aplicar diferentes algoritmos de regresión vistos anteriormente incluyendo redes neuronales.

La Tabla 13 muestra los resultados de la validación para los algoritmos de árboles aleatorios, regresión lineal y máquina de soporte vectorial, incluyendo un método peculiar de predicción que es considerar la potencia constante y proponer como futuro, el ultimo valor visto.

Modelo	MSE	RMSE	MAE	R2
RF_6	0.001603	0.040040	0.019926	0.982582
LR_6	0.001461	0.038222	0.019014	0.984129
SVR_6	0.001524	0.039045	0.019419	0.983438
RF_3	0.001663	0.040778	0.020250	0.981934
LR_3	0.001492	0.038621	0.019253	0.983795
SVR_3	0.001533	0.039155	0.019549	0.983344
Previo Valor	0.001621	0.040262	0.018849	0.982389

Tabla 13. Resultados de los modelos de auto regresión

Como podemos ver los resultados son muy similares, y el hecho de utilizar más tiempo pasado para la predicción no mejora sustancialmente el resultado. La Figura 47 muestra un detalle ampliado con la predicción usando los tres últimos datos con los algoritmos de regresión lineal, bosques aleatorios y máquina de soporte vectorial para regresión.

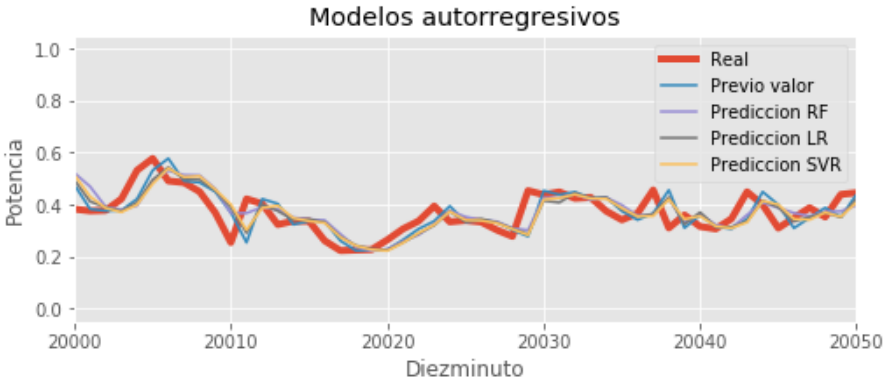


Figura 47. Predicción de la potencia como serie temporal

3.5.1.2 Modelo LSTM

Las redes de memoria a corto plazo o redes LSTM son un tipo de red neuronal recurrente utilizada en el aprendizaje profundo. Está diseñada para manejar la dependencia de secuencias detectando patrones de comportamiento. Aunque en principio la potencia producida no presenta

ningún tipo de patrón de forma visible, probaremos este tipo de redes por si existe algún tipo de tendencia oculta que ayude a la predicción.

La potencia se ha de desglosar en secuencias como los de la Figura 48, formadas por el valor a predecir y los *n* valores previos. Utilizaremos secuencias de 3, 6, y 10 valores. El ejemplo ilustra una secuencia de cuatro valores.

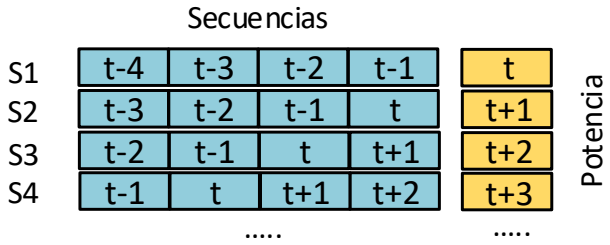


Figura 48. Construcción de secuencias para la red LSTM

La Tabla 14 contempla las precisiones obtenidas sobre el conjunto de validación para diferentes valores de longitud de secuencia y numero de celdas de memoria, y la Figura 49 un detalle ampliado de predicción.

Modelo	MSE	RMSE	MAE	R2
LSTM-3-50-50	0.001549	0.039358	0.020472	0.983171
LSTM-6-50-50	0.001499	0.038721	0.020483	0.983713
LSTM-10-50-50	0.001588	0.039852	0.021259	0.982750
LSTM-3-50-150	0.001595	0.039932	0.021836	0.982676
LSTM-6-50-150	0.001554	0.039427	0.021885	0.983114
LSTM-10-50-150	0.001473	0.038375	0.019040	0.984005

Tabla 14. Resultados de los modelos LSTM

Los resultados también son similares entre sí independientemente del tamaño de la red y la longitud de la secuencia, y en general son algo peores que los modelos auto regresivos, posiblemente debido a que las secuencias no delatan ningún tipo de patrón o conducta que ayuden a la predicción, según habíamos intuido.

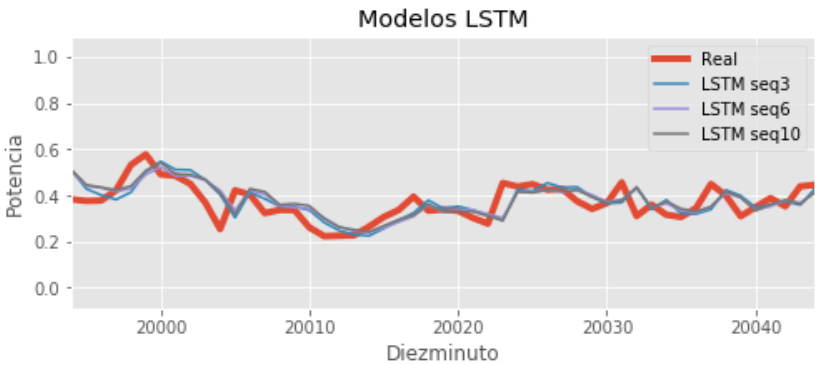


Figura 49. Predicción de la potencia con LSTM

3.5.1.3 Modelo ARIMA

El modelo ARIMA es un modelo auto regresivo integrado de media móvil. Se trata de un modelo estadístico donde la predicción a futuro se realiza en base a los datos del pasado en vez de en variables independientes.

En realidad, es un método que está compuesto por tres componentes: auto regresivo, integrado y de media móvil. Cuando se especifica un modelo ARIMA se utilizan los parámetro *(p,d,q)* que indican el orden con el que se aplican estos componentes, pudiendo desactivar alguno de ellos si le establecemos un valor de cero.

- p: Numero de términos auto regresivos, el número de muestras atrás usadas para la predicción.
- d: Orden en de diferenciación. Permite eliminar la tendencia y la variabilidad, restando a cada muestra el valor anterior.
- q: Número de términos de media móvil. Representa el error del modelo como combinación de términos de error anteriores.

Probaremos con distintos órdenes de auto regresión y aplicando una diferenciación de valor uno, lo que supone convertir la señal en estacionaria restando a cada valor, el valor anterior, con el resultado que muestra la Figura 50.

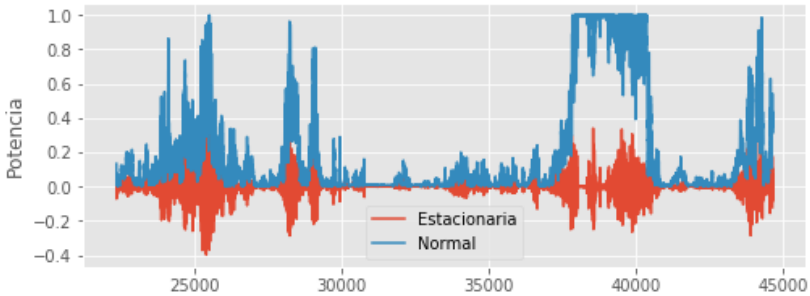


Figura 50. Potencia normalizada y estacionaria

Los resultados de los modelos se incluyen la Tabla 15.

Modelo	MSE	RMSE	MAE	R2
ARIMA-1-1-0	0.001542	0.039272	0.018728	0.983243
ARIMA-3-1-0	0.001464	0.038265	0.018440	0.984092
ARIMA-6-1-0	0.001449	0.038063	0.018414	0.984259
ARIMA-1-0-0	0.001614	0.040180	0.019383	0.982459
ARIMA-3-0-0	0.001484	0.038519	0.018796	0.983879
ARIMA-6-0-0	0.001452	0.038109	0.018653	0.984220

Tabla 15. Resultados de los modelos ARIMA

Lo errores obtenidos son muy similares no pudiendo decir que haya un parámetro determinante que mejore los resultados.

3.5.1.4 Comparativa

Como normal general no ha habido ningún método que haya destacado por encima de los demás de forma clara. Aunque sí que es evidente que los métodos ARIMA encabezan los resultados tanto en los valores MAE como en los RMSE según se ve en la Figura 51.

Los LSTM son los que peor resultado han obtenido debido a que estos algoritmos están pensados para señales con patrones repetitivos que en nuestro caso de estudio no se suelen dar, al menos en periodos de tiempo tan cortos. Hay que tener en cuenta que la producción depende principalmente del viento, y esta es una variable que puede mostrar ciertos patrones o estacionalidad, pero en periodos muy superiores, como por ejemplo un año.

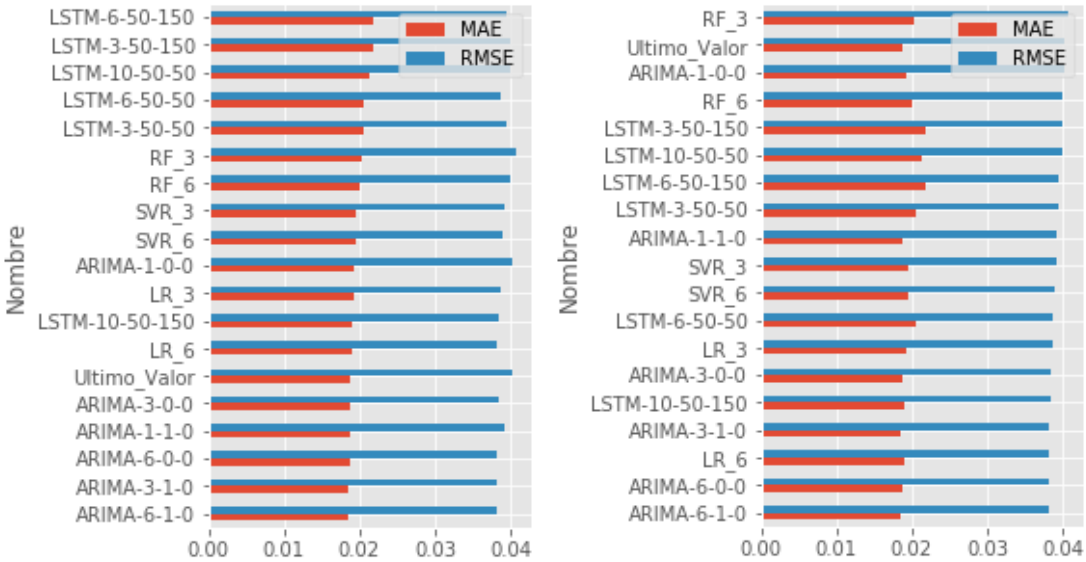


Figura 51. Comparativa de modelos ordenados por MAE y RMSE

En general los mejores modelos son capaces de predecir la potencia con un MAE, o error absoluto medio de entorno al 0.018, lo que supone 36W en un fondo de escala de 2000W. En cuanto al RMSE el indicador es del 0.038, lo que supone 76W.

Estos resultados demuestran que la predicción de potencia del siguiente instante es viable con un error relativamente pequeño, utilizando los datos recientes.

3.5.2 Aplicación de modelos para la previsión a medio plazo

Mientras que la predicción a corto plazo se basa en predecir la potencia en los siguientes minutos, la predicción a medio plazo incluiría periodos más largos como días o semanas, siendo útil para objetivos como:

- Estimar la producción en términos de energía en megavatios hora, integrando la potencia frente al tiempo.
- Anticipar periodos de inactividad para destinarlos a mantenimiento preventivo.
- Anticipar condiciones meteorológicas adversas que faciliten una intervención protectora del parque.

Inicialmente la idea era simular estas predicciones recogiendo la información meteorológica de vientos de proveedores de internet para implementar una secuencia temporal de vientos a partir de la cual poder determinar la potencia generada en cada momento con uno de los modelos creados en el punto 3.3.5

Pero nos hemos encontrado con dos problemas: la precisión y la validación:

- El primero es la precisión de la predicción. Los proveedores gratuitos facilitan la información de vientos de forma muy poco precisa, con intervalos de tiempo de unas seis horas. Algunas de las fuentes consultadas son *Weather Underground* y la Agencia Estatal de Meteorología. Con esta resolución, no podemos realizar una predicción de energía precisa aunque usáramos alguna técnica de interpolación.
  Para estimar la energía se necesita integrar la potencia junto a una unidad de tiempo cuanto más pequeña mejor.
- El otro problema, es la validación de los resultados. Los proveedores gratuitos no ofrecen la estimación de tiempo pasado, lo cual sería necesario para evaluar el grado de ajuste comparando las predicciones efectuadas con los datos reales que disponemos.
  No es muy útil tener una previsión si no es posible saber su precisión.

La solución de este problema de predicción de vientos podría ser acudir a servicios profesionales con mayor resolución y combinar con técnicas de interpolación hasta llegar a la resolución adecuada, o bien implementar un sistema propio de predicción de vientos a nivel de parque basado en modelos meteorológicos, para lo cual haría falta seguramente recursos e inversión en equipamientos.

No hay solución en nuestro caso al problema de la validación. Sería necesaria lo obtención de nuevos datos de SCADA en un tiempo del cual tuviéramos efectuada una predicción de vientos previa.

Por estos motivos, este punto no se simulará.

4 Conclusiones

El sector de la energía eólica es un sector clave debido al auge que ha experimentado en las últimas décadas, su tendencia al alta, y sobre todo, por su inminente expansión al mar. Además, jugará un papel importante en el objetivo de cumplir con las directivas europeas en cuanto al consumo de energía verde [4].

Las máquinas modernas disponen de complejos sistemas de monitorización destinados al mantenimiento preventivo para evitar pérdidas en la producción y sobrecostes. No obstante, hay un importante número de instalaciones antiguas que no disponen de ellos, por lo cual detectar posibles anomalías a partir de solo los datos de control, puede ser de especial interés, ya que permitiría utilizar estas técnicas de reducción de costes en máquinas que están llegando al final de su vida útil y sufren por lo tanto un mayor número de averías.

El uso de técnicas de minería de datos en el sector eólico esta constatado en la numerosa documentación consultada, tanto para conjuntos de datos sintéticos como reales, no obstante, se ha observado que los resultados de los algoritmos son muy dependientes del modelo y fabricante, ya que cada máquina tiene unas características peculiares.

En este trabajo se ha validado el uso de estas técnicas para un conjunto de datos real y propietario procedente de un SCADA de un aerogenerador de 2 MW sin sensores específicos de mantenimiento, es decir, sólo con las variables que al autómata utiliza como control.

El dataset contiene 23 variables, de las cuales algunas no son de aplicación y otras han sido descartadas por tener una alta correlación mecánica o proceder de sistemas redundantes. En ese sentido el dataset podría haber sido mejor, no obstante, aun así, se ha conseguido realizar una selección de características que nos permiten predecir con precisión tanto la potencia como el *pitch*.

Los algoritmos entrenados han sido sometidos a una búsqueda de hiperparámetros para una configuración óptima, y se ha usado validación cruzada para la comprobación del ajuste. Algunos como *Random Forest* y *Gradient Boosting* han conseguido resultados excelentes, al igual que *Support Vector Regression*, quedando por encima incluso de las redes neuronales y dejando patente su validez pese a no ser los más referenciados en la documentación consultada en el estado del arte.

En nuestro caso es un resultado importante ya que estos algoritmos son fáciles de ajustar y entrenan rápido. *Gradient Boosting* por ejemplo consigue predecir la potencia con un error absoluto medio de apenas 10W frente a un fondo de escala de 2,000W.

En cuanto a la detección de anomalías, conseguir datos de fallos reales en máquinas es algo difícil, de modo que se han simulado dos de las anomalias más frecuentes en los aerogeneradores: la perdida de potencia y la desalineación.

La primera tiene el efecto de que no se puede alcanzar la producción nominal por problemas en el generador o en la multiplicadora, mientras que la segunda tiene como efecto que se necesita más viento del habitual para alcanzar la producción nominal, es decir no se aprovecha la fuerza del viento en su plenitud.

La simulación del fallo comienza con la determinación de los periodos de tiempo donde la máquina permanece en régimen de producción de energía, y que se considera la normalidad, para posteriormente sintetizar un dataset anómalo después de modificar de forma controlada el valor de algunas variables, como su reducción en un factor.

En la detección de anomalías, se han utilizado los días pares para entrenamiento y los impares para validación, de este modo los algoritmos tienen la opción de aprender de los datos distribuidos a lo largo de todas las condiciones de viento del conjunto.

Para la creación del conjunto de validación se han usado los días impares como días normales, y esos mismos días, tras simular el fallo, como días anómalos, de modo que se obtiene un conjunto equilibrado con un 50% de cada tipo, para someter a prueba el clasificador.

El mayor problema que se ha encontrado es la alta varianza que se observa en la zona de vientos media. A ciertas velocidades el abanico de potencias registradas es muy amplio, y eso se traduce en que el ajuste de los modelos es peor en esta zona, provocando una distribucion no uniforme de los residuos.

Se han aplicado con éxito técnicas basadas en métodos estadísticos y basadas en una red neuronal *encoder-decoder*.

Los metodos estadísticos han demostrado ser más precisos, siendo capaces de no fallar en un dia normal, e identificar con un 97% de certeza un día anómalo con un 10% de perdidas, siempre que éste supere la potencia media del 20% nominal, lo que significa una precisión del 100% y una sensibilidad del 97%.

La técnica de *encoder-decoder* ha demostrado tener un comportamiento algo peor, pero también aceptable, donde nos obliga a tener elegir entre sensibilidad y precisión, existiendo un punto de equilibrio con un 85% en ambos estadísticos.

En cuanto a la predicción de energía, se han probado diferentes algoritmos para la predicción a corto plazo, como modelos auto regresivos en los que se ha usado la regresión lineal, máquinas de soporte vectorial y arboles aleatorios. También se han probado redes neuronales LSTM y modelos combinados como ARIMA, demostrado que son efectivos y con resultados muy similares, quedando ligeramente por encima los métodos

ARIMA. En general son capaces de predecir la potencia del siguiente instante con un error absoluto medio de 36W.

En cuanto a la previsión a medio plazo, no se llevado a cabo finalmente la simulación debido a la imposibilidad de contar con una predicción meteorológica con suficiente resolución y la incapacidad de poder evaluar la precisión que se puede obtener.

Como conclusión podemos decir que los resultados obtenidos hacen viable el uso de las técnicas de minería de datos para nuestro dataset en concreto con uno resultados aceptables, siendo capaz de identificar con bastante precisión las anomalías simuladas y predecir la potencia a corto plazo basándose en datos del pasado.

5 Mejoras y trabajos futuros

Una vez obtenidos los objetivos de este trabajo, se plantean una serie de mejoras.

Algunas pasan con conseguir un conjunto de datos más completo que incluyan anomalías reales.

En el actual juego, además hemos encontrado alguna limitación, como la falta de características, en concreto las de temperatura de los componentes principales, que habrían permitido casi con seguridad realizar una predicción más precisa y con una menor dispersión de los residuos, lo que podría traducirse en una mejora de la efectividad de los clasificadores.

También habría sido conveniente que hubiera cubierto mayor tiempo de adquisición, de al menos un año para haber observado de forma completa el ciclo anual de vientos.

Un estudio interesante seria la determinación de los datos de entrenamiento mínimos para obtener una clasificación precisa. El conjunto que hemos tratado incluye información de unos 300 días de funcionamiento, donde hemos usado la mitad para entrenar y la otra mitad para validar la detección de anomalías. Sería aconsejable disponer de un periodo superior y determinar mediante la partición de conjuntos, como influye la cantidad de datos de entrenamiento, en la calidad de la predicción. Esto tiene como interés el conocer a partir de cuantos días de datos se puede empezar a contar con un sistema efectivo.

La detección anomalías podría ser mejorable mediante el diseño de un sistema de *scoring.* donde el algoritmo clasificador de anomalías obtenga como resultado un valor numérico que represente el porcentaje de normalidad observada. De este modo no solo se obtiene una etiqueta de

“día normal” o “día anómalo” sino un valor que refuerza la pertenencia a cada una de esas clases.

Para ello se propone el estudio de varios métodos, como promediar los datos del día a datos horarios y devolver el porcentaje de horas dentro y fuera del umbral, o bien de un sistema de puntuación donde cada dato horario se le asigne un valor en función de la distancia al umbral, de modo que una hora muy por encima del umbral cuente más que otra más cercana.

Otros trabajos interesantes podrían estar en el estudio a nivel de parque. Estudiar si el modelo de clasificación es efectivo de forma individualizada en cada una de las máquinas o si existe algún método de generalización por el cual se pueda extender a todas.

Finalmente puede ser interesante estudiar la posible integración del método de detección de anomalías a un sistema embarcado para la construcción de un demostrador. Esto permitiría que tras diseño de un pequeño hardware pueda ser instalado en los propios aerogeneradores o subestaciones y trabaje de forma autónoma, recogiendo datos diarios, evaluando el estado, y lanzando señales de alarma.

Sería de aplicación en instalación antiguas donde la recogida de datos de SCADA se hace complicada más allá de la subestación o incluso de la máquina.

6 Glosario

ANFIS: *Adaptive Neuro-Fuzzy Inference System*. Método que integra los principios de las redes neuronales y de la lógica borrosa.

API: *Application Programming Interface*. Interfaz de programación de aplicaciones usada para intercambio de información entre proveedores y clientes.

ARIMA: *Autoregressive Integrated Moving Average.* Método auto regresivo utilizado para la predicción de series temporales.

CMS: *Condition Monitoring System*. Sistema de registro de datos comúnmente usado para supervisión de máquinas.

IEEE: *Institute of Electrical and Electronics Engineers*. Asociación mundial de ingenieros dedicada a la normalización y el desarrollo en áreas técnicas.

LSTM: *Long short-term memory*. Red neuronal recurrente utilizada en el aprendizaje profundo utilizada en series temporales.

MDPI: *Multidisciplinary Digital Publishing Institute*. Portal de acceso abierto a cientos de revistas digitales tecnológicas.

MW: Megavatio. Unidad de potencia.

PARAFAC: *Parallel Factor Analysis*. Método de descomposición de arrays.

PLC: *Programmable logic controller.* Controlador lógico programable que en la industria se utiliza para el control de sistemas.

REST: *REpresentational State Transfer*. Arquitectura de servicios web para sistemas distribuidos.

SCADA: *Supervisory Control And Data Acquisition*. Sistema de supervisión y control de procesos industriales.

SLFN: *Single Layer Feedforward neural network*. Tipo de red neural no recurrente con una sola capa oculta.

SOM: *Self organized map*. Método de reducción de dimensionalidad basado en redes neuronales.

SVM*: Support Vector Machine.* Técnica de *machine learning* basada en la construcción de hiperplanos.

SVR: *Support Vector Regression.* Técnica de *machine learning* basada en la construcción de hiperplanos para regresión.

7 Bibliografía

[1] Asociación Empresarial Eólica, «La eólica en España,» [En línea]. Available: https://www.aeeolica.org/sobre-la-eolica/la-eolica-espana. [Último acceso: Febrero 2020].

[2] Red Eléctrica de España, «Informe de las energías renovables,» [En línea]. Available: http://www.ree.es. [Último acceso: Febrero 2020].

[3] El boletín, «Auge y caída de las energías renovables,» [En línea]. Available: https://www.elboletin.com/economia/142322/auge-caida-energias-renovables.html. [Último acceso: Febrero 2020].

[4] Diario Oficial de la Unión Europea, «Directiva 2009/28/Ce Del Parlamento Europeo Y Del Consejo,» [En línea]. Available: https://www.boe.es/doue/2009/140/L00016-00062.pdf. [Último acceso: Febrero 2020].

[5] Siemens Gamesa, «Servicios de diagnóstico: El poder del big-data,» [En línea]. Available: https://www.siemensgamesa.com/es-es/products-and-services/service-wind/servicios-dediagnostico. [Último acceso: Febrero 2020].

[6] Danish Energy Agency, «Ea Energy Analyses. Integration of wind energy in power systems,» [En línea]. Available: https://en.energinet.dk. [Último acceso: Marzo 2020].

[7] M. Entezami, «Wind Turbine Condition Monitoring System,» Tesis doctoral. School of electronic, electrical and computer engineering, University of Birmingham, 2010.

[8] T. Esbensen, «Joint Power and Speed Control of Wind Turbines,» Department of Electronic Systems, Denmark, 2008.

[9] P. Sun, «A generalized model for wind turbine anomaly identification based on SCADA data,» *Applied Energy 168 (2016) 550–567,* 2016.

[10] M. Schlechtingen, «Wind turbine condition monitoring based on SCADA data using normal behavior models,» *Applied Soft Computing 13 259–270,* 2013.

[11] P. Tchakoua, «New trends and future challenges for wind turbines condition monitoring,» Second International Conference on Control, Automation and Information Sciences, Nha Trang City, Vietnam, 2013.

[12] R. K. Mobley, An introduction to predictive maintenance. Second Edition. ISBN 0-7506-7531-4, Butterworth-Heinemann, 2002.

[13] M. H. Kai Zhong, «Data-Driven Based Fault Prognosis for Industrial Systems: A Concise Overview,» *IEEE/CAA Journal of automatica sinica, vol. 7, no. 2,* March 2020.

[14] P. Tchakoua, «A Review of Concepts and Methods for Wind Turbines Condition Monitoring,» School of Engineering, Université du Québec à Chicoutimi, Canada, 2013.

[15] V. Chandola, «Anomaly Detection : A Survey,» Technical Report Department of Computer Science and Engineering, University of Minnesota, 2007.

[16] S. Alla, Beginning Anomaly Detection Using Python-Based Deep Learning, New Jersey, NJ, USA ISBN-13 (pbk): 978-1-4842-5176-8: Apress., 2019.

[17] P. Sun, «Wind Turbine Anomaly Detection Using Normal Behavior Models based on SCADA Data,» Department of High Voltage and Insulation Engineering, School of Electrical Engineering, Chongqing University.

[18] F. Saverio, «Data-Driven Fault Diagnosis and Fault Tolerant Control of Wind Turbines,» Tesis doctoral., Università degli Studi di Ferrara, 2016.

[19] Z. Hameed, «Condition monitoring and fault detection of wind turbinesand related algorithms: A review,» *Renewable and Sustainable Energy Reviews 13 (2009) 1–39.*

[20] E. Papatheou, «Wind Turbine Structural Health Monitoring: A Short Investigation Based on SCADA Data,» *EWSHM - 7th European Workshop on Structural Health Monitoring, IFFSTTAR, Inria, Université de Nantes, Jul 2014, Nantes, France,* 2014.

[21] P. Qian, «Data-driven condition monitoring approaches to improving power output of wind turbines».

[22] W. Zhang, «Simultaneous Fault Detection and Sensor Selection for Condition Monitoring of Wind Turbines,» *Energies. 9. 280. 10.3390/en9040280. ,* 2016.

[23] M. Du, «A SCADA Data based Anomaly Detection Method for Wind Turbines,» *China International Conference on Electricity Distribution (CICED 2016) Xi'an, 10-13 Aug, 2016,* p. Paper No. CP0429, 2016.

[24] S. Shokrzadeh, «Wind Turbine Power Curve Modeling Using Advanced Parametric and Nonparametric Methods,» *IEEE transactions on sustainable energy, vol. 5, no. 4, october 2014. ,* pp. p1262-1269, 2014.

[25] E. Taslimi-Renani, «Development of an enhanced parametric model for wind turbine power curve,» *Applied Energy 177 (2016) 544–552,* 2016.

[26] R. K. Pandit, «SCADA based nonparametric models for condition monitoring of a wind turbine,» The 7th International Conference on Renewable Power Generation. doi: 10.1049/joe.2018.9284, 2018.

[27] O. Janssens, «Data-driven multivariate power curve modeling of offshore wind turbines,» *Engineering Applications of Artificial Intelligence 55 (2016) 331–338,* 2016.

8 Anexos

En este apartado se incluyen los anexos I y II, que contienen información de referencia de los procesos de exploración de las variables y la creación de modelos.

También se incluye el anexo III con la referencia al código fuente del trabajo.

8.1 Anexo I. Variables de SCADA

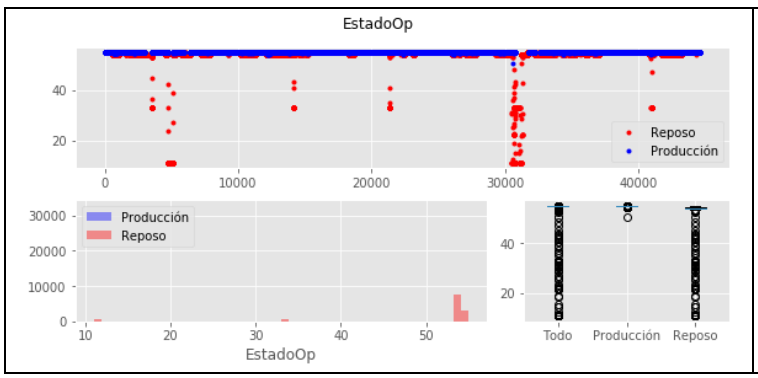
En el punto 3.2 se lleva a cabo un análisis exploratorio de los datos donde se analizan las variables recibidas del SCADA y en el punto 3.2.1 se determina los periodos de tiempo en los que la maquina permanece en régimen de producción de energía.

Adjuntamos como referencia:

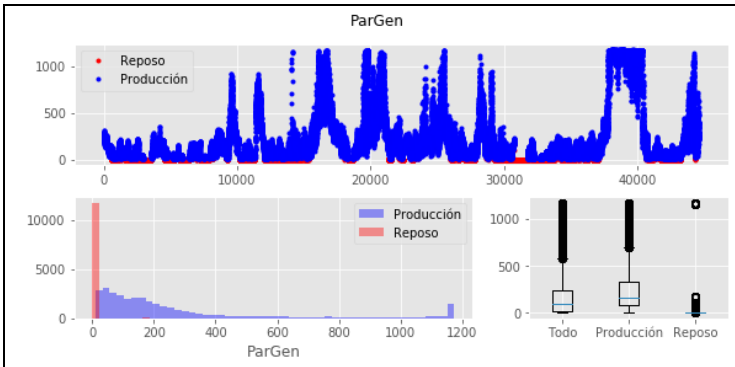
- El graficado temporal de las señales más relevantes y algunos estadísticos.
- La observación de la curva de potencia antes y después de retirar los periodos donde la maquina no produce energía
- Mapa de calor y los valores numéricos de los índices de correlación de Pearson entre las variables.

8.1.1 Descripción gráfica por estado de máquina

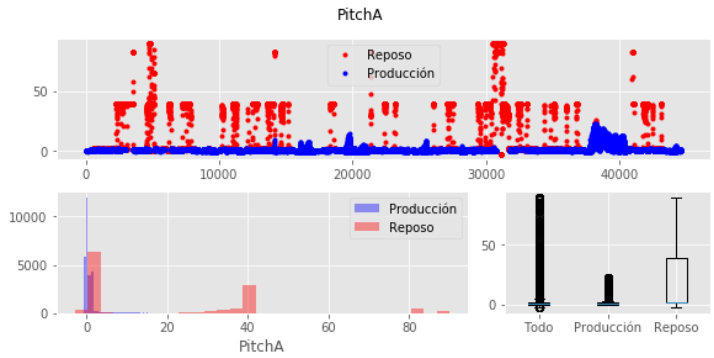
Para cada señal se grafica la línea temporal, y la distribución de los valores de forma íntegra y diferenciando los estados operacionales de producción de energía o reposo.



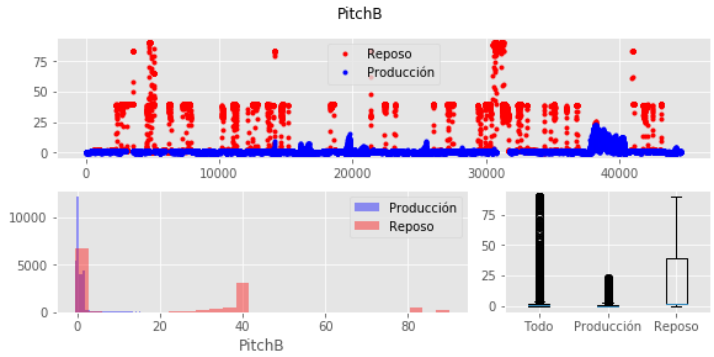

```
mean     53.999711
std       5.460979
min      11.000000
25%      55.000000
50%      55.000000
75%      55.000000
max      55.000000
```



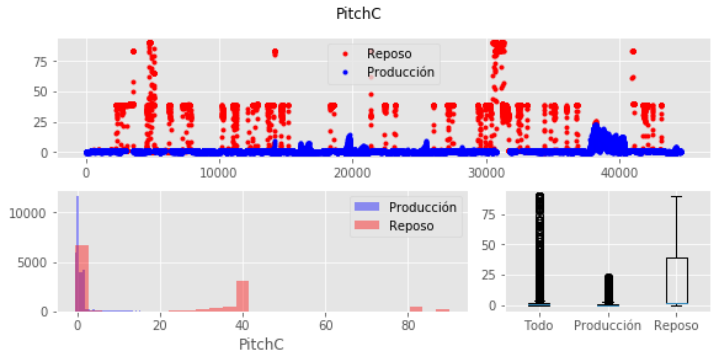
mean	211.511012
std	300.051626
min	0.000000
25%	12.341167
50%	100.076918
75%	241.940129
max	1174.186521



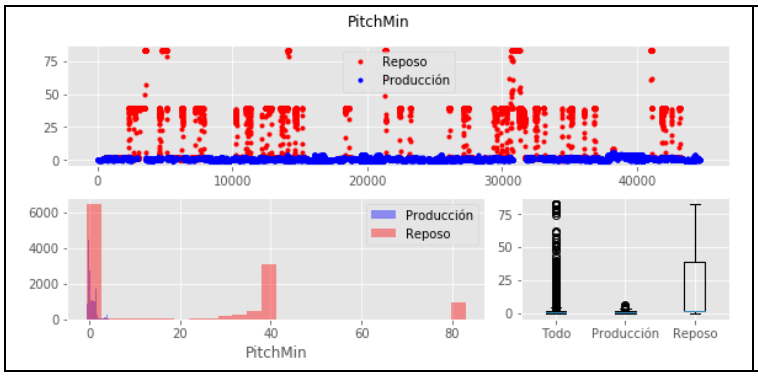
mean	5.964586
std	15.036366
min	-2.876067
25%	-0.074817
50%	0.774566
75%	1.636187
max	90.079861



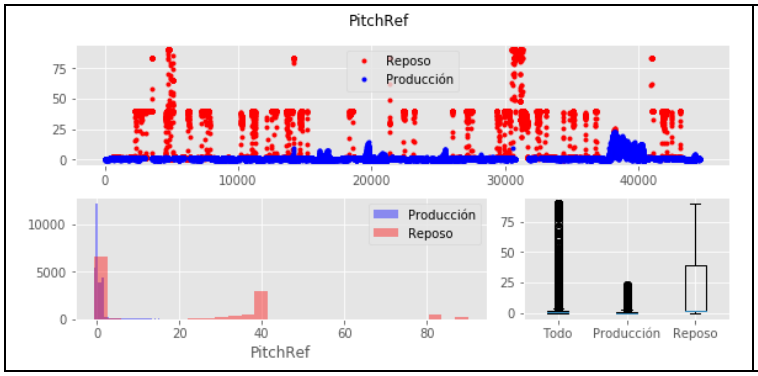
mean	6.055010
std	15.206042
min	-0.644800
25%	-0.049067
50%	0.810234
75%	1.648646
max	90.036264



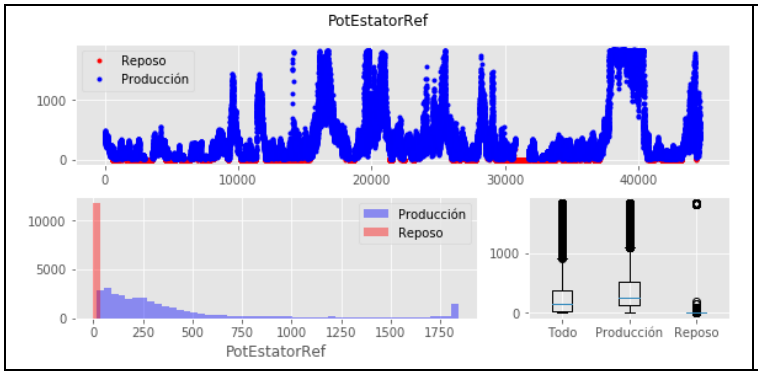
mean	6.055865
std	15.204390
min	-0.626133
25%	-0.044392
50%	0.813850
75%	1.648929
max	90.006533



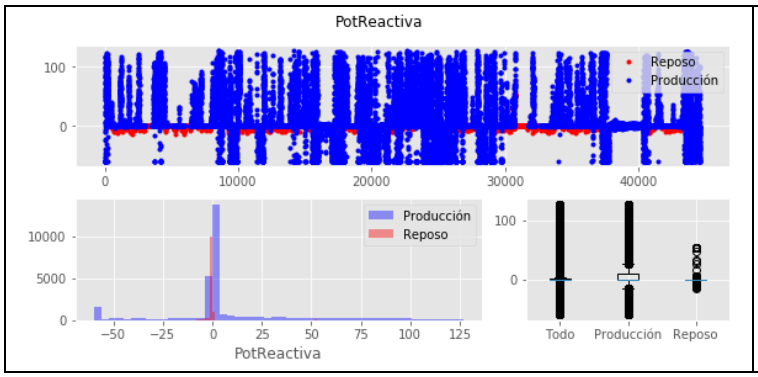
mean	6.080520
std	15.882357
min	-0.676117
25%	-0.086800
50%	0.791275
75%	1.632008
max	83.000000



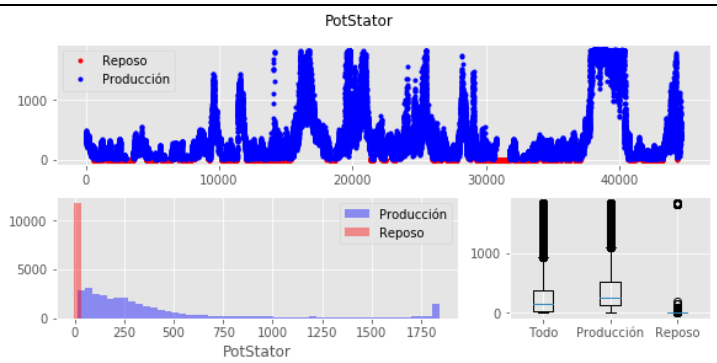
mean	5.969520
std	15.009974
min	-0.676183
25%	-0.085904
50%	0.772558
75%	1.625167
max	90.000000



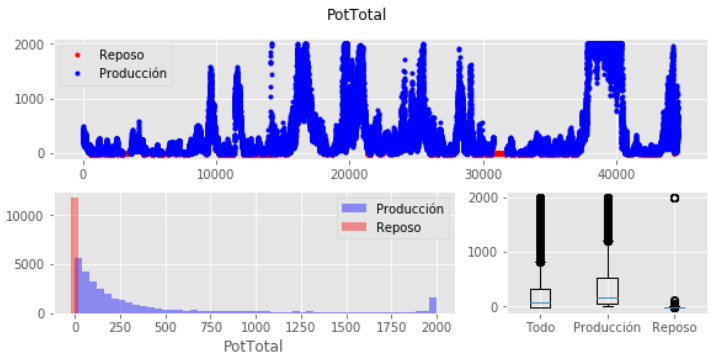
mean	331.704013
std	471.541790
min	0.000000
25%	18.407292
50%	155.549253
75%	380.004590
max	1844.427191



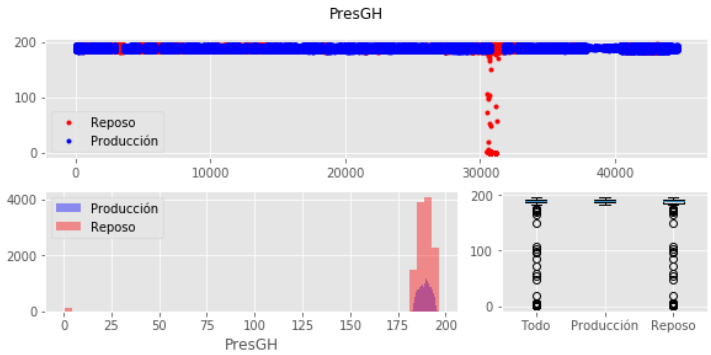
mean	6.735466
std	30.445788
min	-60.000000
25%	-0.166667
50%	0.000000
75%	0.997084
max	126.995000



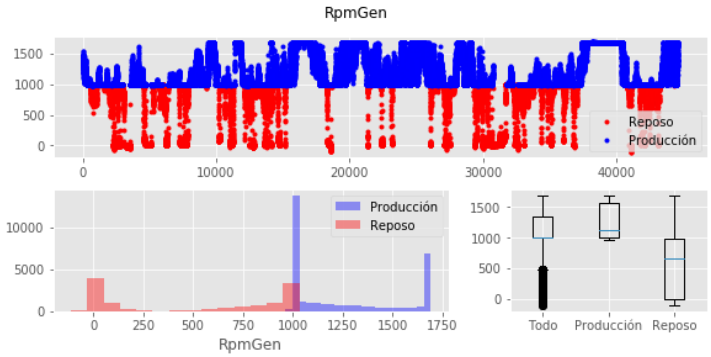
mean	331.668701
std	471.535500
min	-0.842833
25%	18.360667
50%	155.613419
75%	380.045089
max	1843.872857



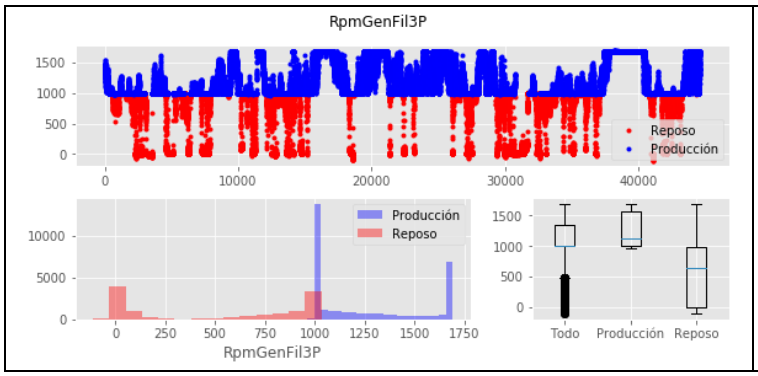
mean	317.295751
std	526.259970
min	-19.597667
25%	-1.861875
50%	86.756251
75%	327.620797
max	2000.526356



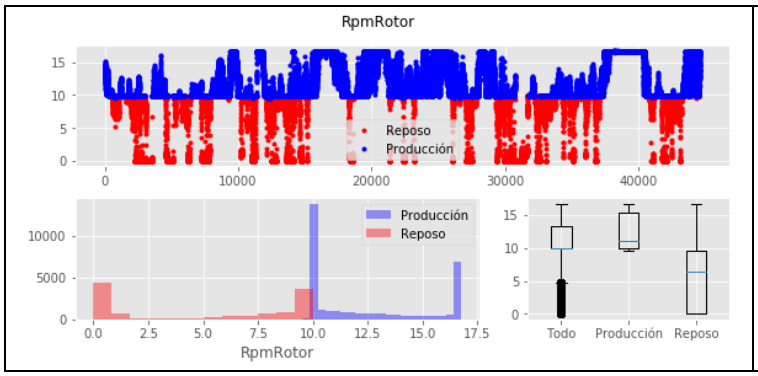
mean	188.608135
std	9.829763
min	0.297833
25%	186.513003
50%	189.233837
75%	191.554836
max	196.393669



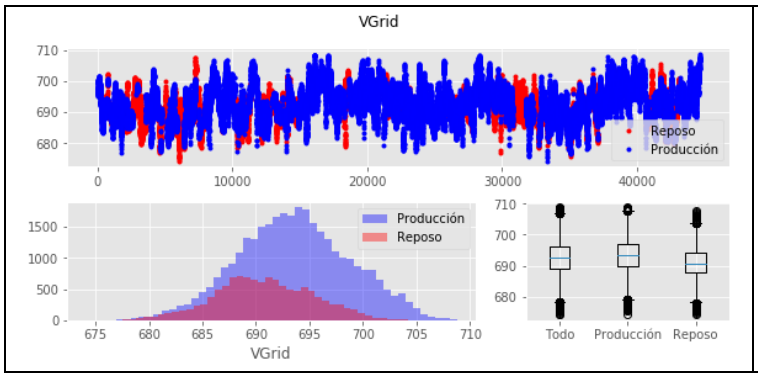
mean	1052.050228
std	465.331451
min	-110.387502
25%	999.739845
50%	1000.236180
75%	1349.491481
max	1692.043357



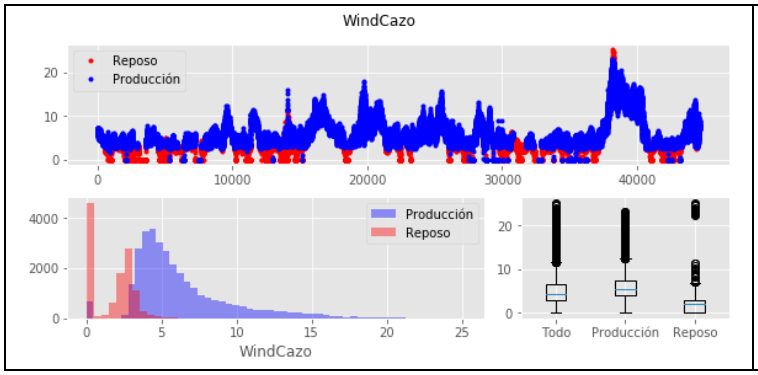
mean	1052.036132
std	465.324113
min	-110.430335
25%	999.727013
50%	1000.223679
75%	1349.278273
max	1692.099024



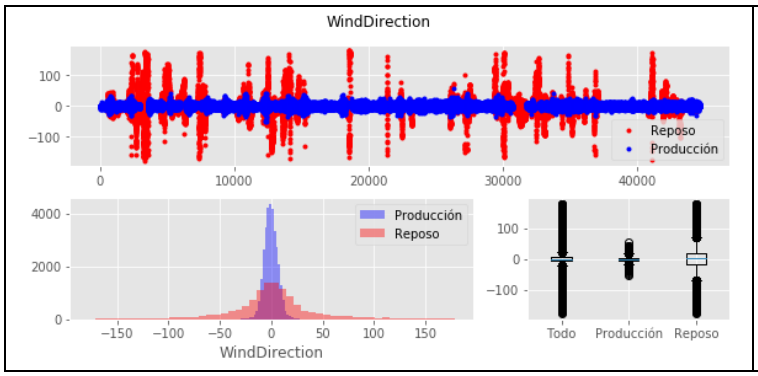
mean	10.407155
std	4.606552
min	0.000000
25%	9.890517
50%	9.896050
75%	13.348400
max	16.745633



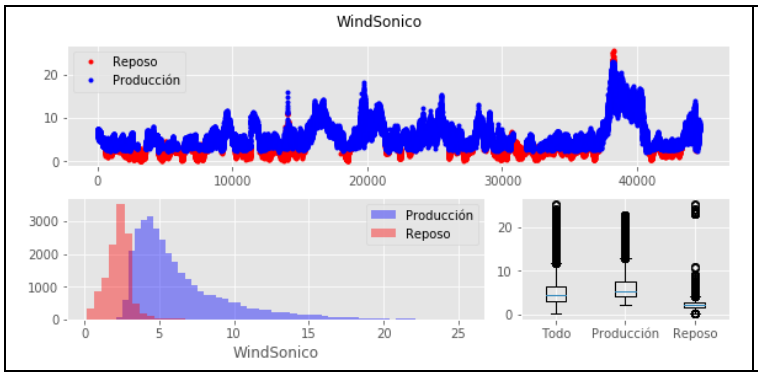
mean	692.780493
std	5.245891
min	674.118679
25%	689.113720
50%	692.645013
75%	696.265679
max	708.791511



mean	5.083187
std	3.624386
min	0.000000
25%	2.996063
50%	4.432092
75%	6.471596
max	25.260816



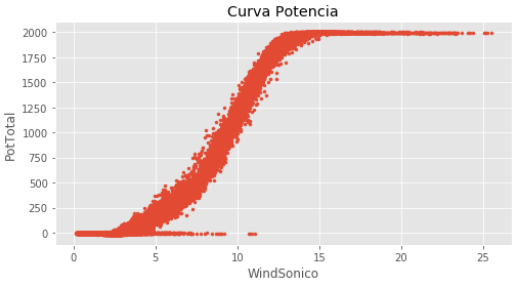
mean	0.132255
std	25.297682
min	-177.965171
25%	-5.580875
50%	-0.255500
75%	5.586250
max	179.259336

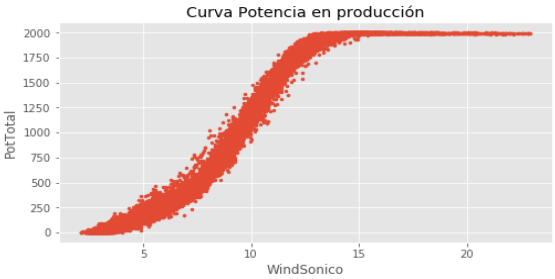


mean	5.311494
std	3.353631
min	0.139467
25%	3.003367
50%	4.477884
75%	6.546425
max	25.473749

8.1.2 Curva de potencia antes y después del filtrado

Las siguientes graficas ilustran la curva de potencia con el conjunto completo y con solo los datos en régimen de producción de energía.

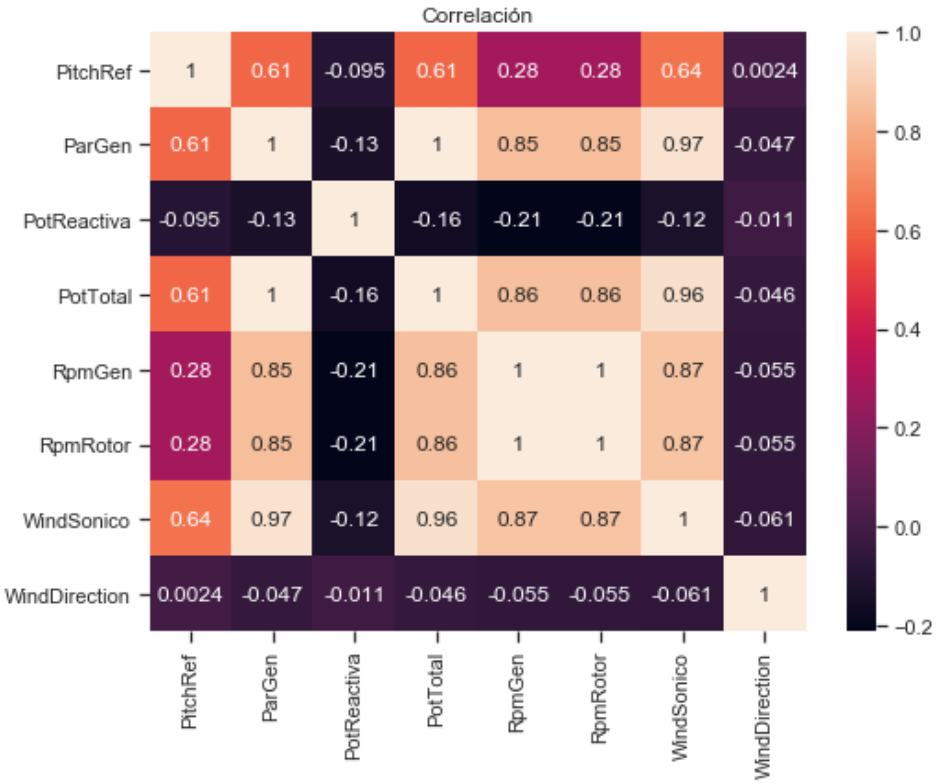




Como puede observarse en la curva en modo de producción, los valores de potencia entorno a cero con vientos superiores a 5 m/s han sido suprimidos, ya que obviamente la maquina se encuentra fuera de servicio.

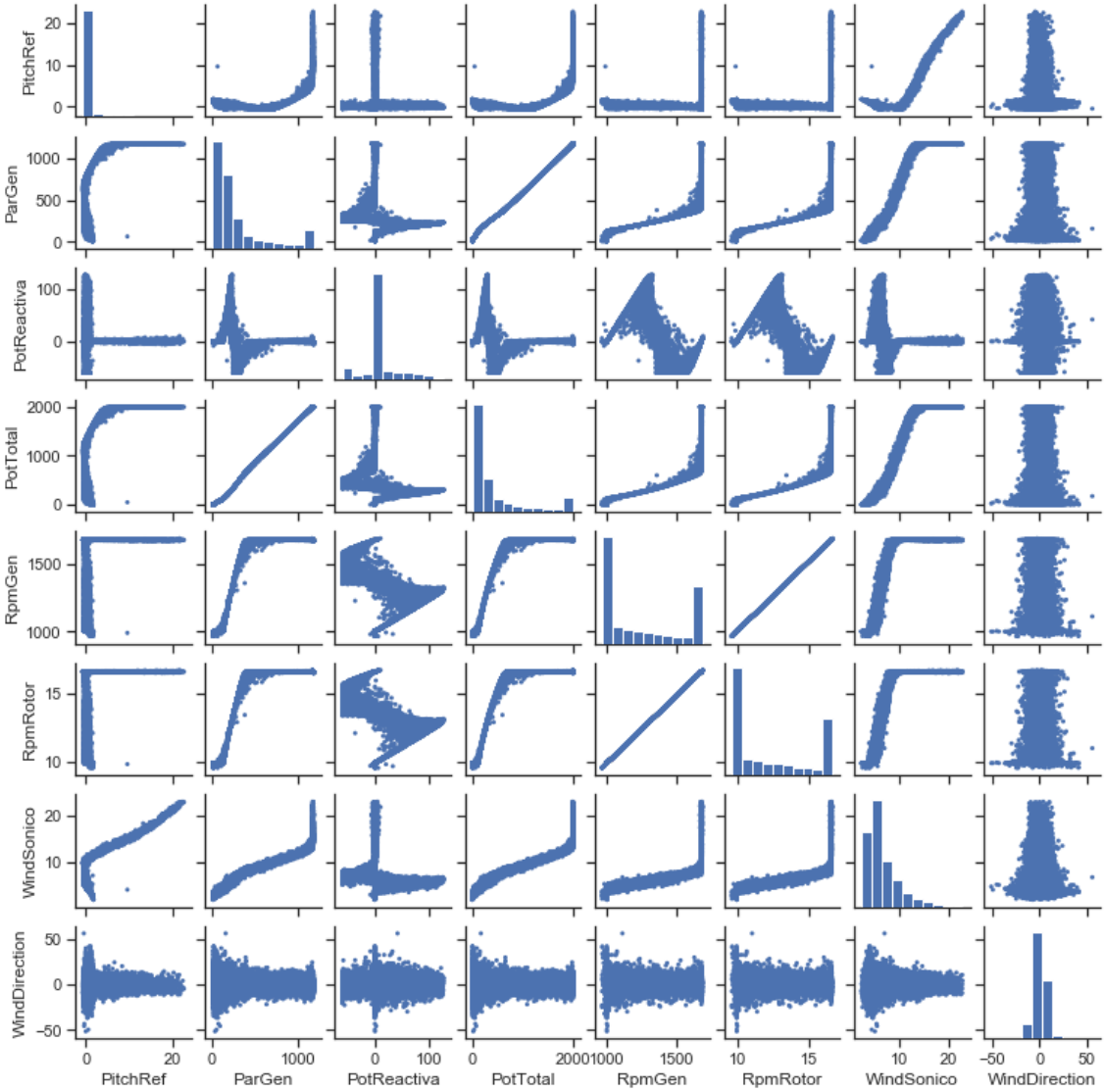
8.1.3 Correlación entre variables principales

Se adjunta el mapa de calor con los índices de correlación de Pearson obtenidos y la tabla con los valores numéricos.



	PitchRef	ParGen	PotReactiv	PotTotal	RpmGen	RpmRotor	WindSonico	WindDirec
PitchRef	1,00	0,61	-0,09	0,61	0,28	0,28	0,64	0,00
ParGen	0,61	1,00	-0,13	1,00	0,85	0,85	0,97	-0,05
PotReactiva	-0,09	-0,13	1,00	-0,16	-0,21	-0,21	-0,12	-0,01
PotTotal	0,61	1,00	-0,16	1,00	0,86	0,86	0,96	-0,05
RpmGen	0,28	0,85	-0,21	0,86	1,00	1,00	0,87	-0,05
RpmRotor	0,28	0,85	-0,21	0,86	1,00	1,00	0,87	-0,05
WindSonico	0,64	0,97	-0,12	0,96	0,87	0,87	1,00	-0,06
WindDirection	0,00	-0,05	-0,01	-0,05	-0,05	-0,05	-0,06	1,00

Grafica de pares de las variables principales


PitchRef
ParGen
PotReactiva
PotTotal
RpmGen
RpmRotor
WindSonico
WindDirection

8.2 Anexo II. Resultado de los modelos.

En el proceso de creación de modelos, tras seleccionar las variables candidatas, se ejecutan los diferentes algoritmos en una secuencia que determina la mejor combinación de sus parámetros.

En este proceso se genera una serie de graficas que representan:

- Un mapa de calor con el *score* R2 obtenido en la validación cruzada en cada una de las combinaciones de parámetros.
- Un diagrama de cajas con la distribución de todos los *scores* obtenidos.

También se muestran los siguientes estadísticos:

- Los parámetros que han obtenido mejor resultado.
- El valor R2 obtenido del proceso de validación cruzada y de la validación del conjunto de test.
- Otros estadísticos como MAE, MSE, RMSE y MAPE de la mejor opción
- La media y desviación estándar de los residuos

Y finalmente se aporta unas graficas que representa el ajuste:

- Se grafican los conjuntos de entrenamiento, validación y las predicciones sobre el conjunto de validación.
- Los residuos frente al viento y a la variable predicha.
- Los valores observados, la predicción y los residuos en serie temporal.
- La distribución de los residuos.

8.2.1 Predicción de la potencia

Se incluyen los resultados de la evaluación del ajuste de todos los modelos.

Los nombres modelos se corresponden con la nomenclatura AA_VVV->P, siendo A el algoritmo, V las iniciales de los predictores empleados y P la variable a predecir.

Modelo	CV R2	MSE	RMSE	MAE	MAPE	TST R2
RF_W->Pot	0,99443	0,00045	0,02118	0,01321	62,38791	0,99463
KNN_W->Pot	0,99484	0,00042	0,02049	0,01290	61,39661	0,99497
SVR_W->Pot	0,99482	0,00042	0,02055	0,01253	43,94289	0,99494
GB_W->Pot	0,99472	0,00043	0,02073	0,01308	64,83367	0,99485
LR_W->Pot	0,93027	0,00557	0,07465	0,05486	811,09717	0,93324
NN_W->Pot	0,99478	0,00043	0,02069	0,01370	75,08820	0,99487
DNN_W->Pot	0,99447	0,00045	0,02111	0,01325	37,15605	0,99466
RF_WRP->Pot	0,99849	0,00011	0,01065	0,00544	50,67733	0,99864
KNN_WRP->Pot	0,99826	0,00013	0,01146	0,00580	43,71671	0,99843
SVR_WRP->Pot	0,99796	0,00016	0,01254	0,00643	40,48510	0,99812
GB_WRP->Pot	0,99850	0,00011	0,01058	0,00542	53,54754	0,99866
LR_WRP->Pot	0,93229	0,00544	0,07376	0,05308	672,93430	0,93481

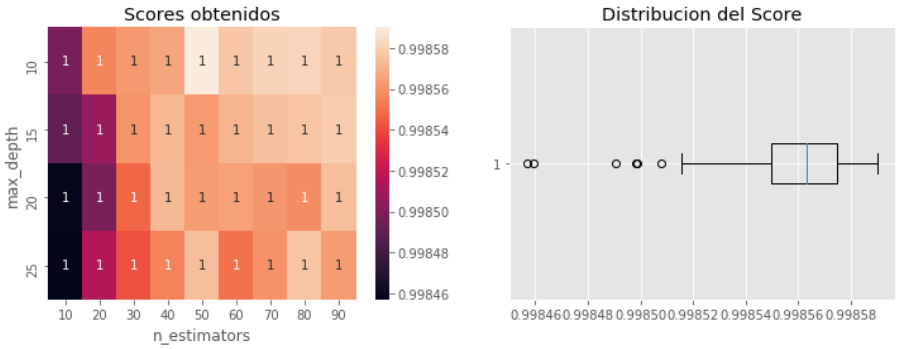
NN_WRP->Pot	0,99765	0,00017	0,01296	0,00773	61,28640	0,99799
DNN_WRP->Pot	0,99724	0,00020	0,01431	0,00836	60,09850	0,99755
RF_WRPD->Pot	0,99859	0,00010	0,01017	0,00528	51,41877	0,99876
KNN_WRPD->Pot	0,99805	0,00015	0,01221	0,00623	50,10767	0,99822
SVR_WRPD->Pot	0,99803	0,00015	0,01235	0,00637	39,81602	0,99817
GB_WRPD->Pot	0,99864	0,00010	0,00999	0,00523	48,85665	0,99881
LR_WRPD->Pot	0,93239	0,00543	0,07369	0,05300	665,17593	0,93495
NN_WRPD->Pot	0,99782	0,00014	0,01183	0,00685	38,76251	0,99832
DNN_WRPD->Pot	0,99784	0,00014	0,01175	0,00646	63,45316	0,99835

En la predicción de la potencia los modelos que mejor resultado han obtenido son lo que usan como predictores: el viento, las revoluciones, el pitch y la dirección de viento.

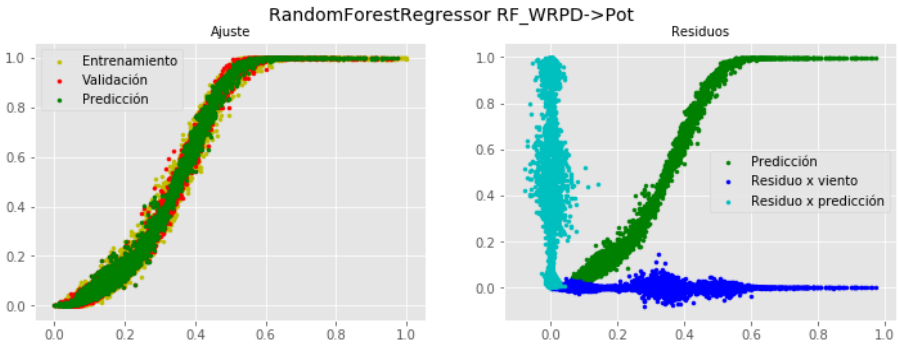
Se incluyen a continuación los resultados de la búsqueda de hiperparámetros.

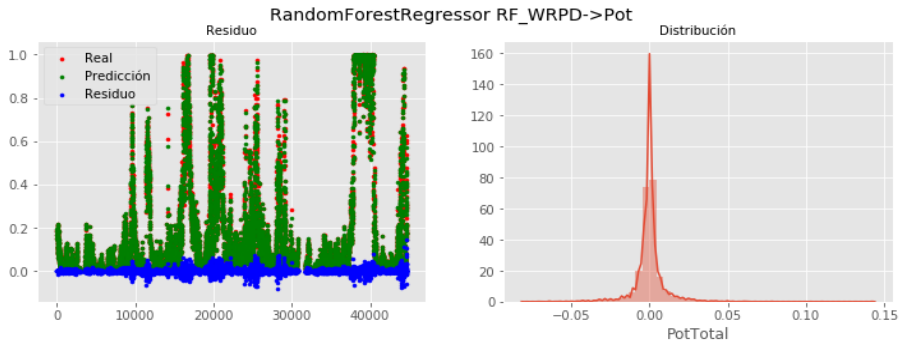
8.2.1.1 Random Forest Regresor (WRPD->Pot)

```
Modelos computados: 36
Precisiones obtenidas:
```



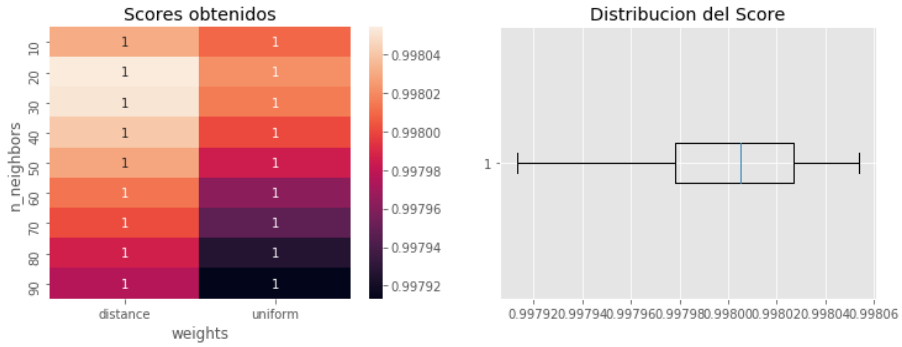

```
Mejores parametros: {'max_depth': 10, 'n_estimators': 90}
Mejor accuracy: 0.9985883715108643
CV R2 Score: 0.99859 (+/- 0.00017)
MSE: 0.000103417 RMSE: 0.010169414 MAE: 0.005281574 MAPE:
51.418768496 R2: 0.998760943
Residual: Mean: -0.000040810 Std: 0.010169332
```



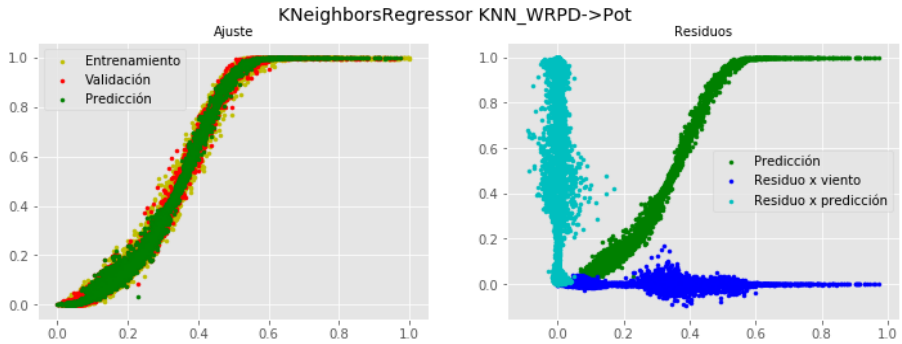


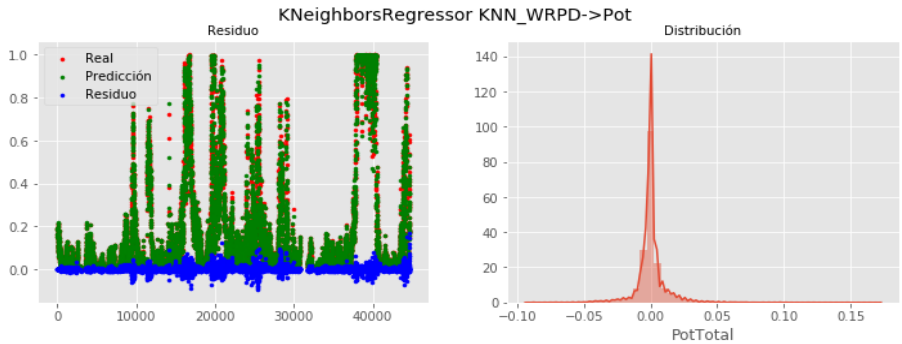
8.2.1.2 K-NN (WRPD->Pot)

```
Modelos computados: 18
Precisiones obtenidas:
```



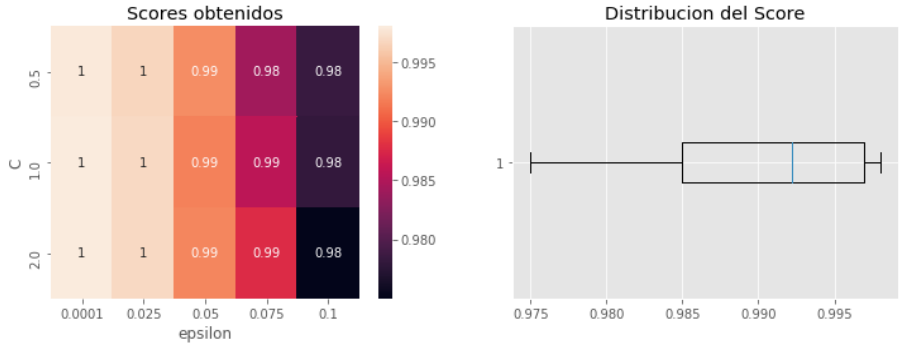

```
Mejores parametros: {'n_neighbors': 20, 'weights': 'distance'}
Mejor accuracy: 0.9980538945470278
CV R2 Score: 0.99805 (+/- 0.00021)
MSE: 0.000148978 RMSE: 0.012205641 MAE: 0.006234704 MAPE:
50.107666436 R2: 0.998215072
Residual: Mean: 0.000418454 Std: 0.012198466
```



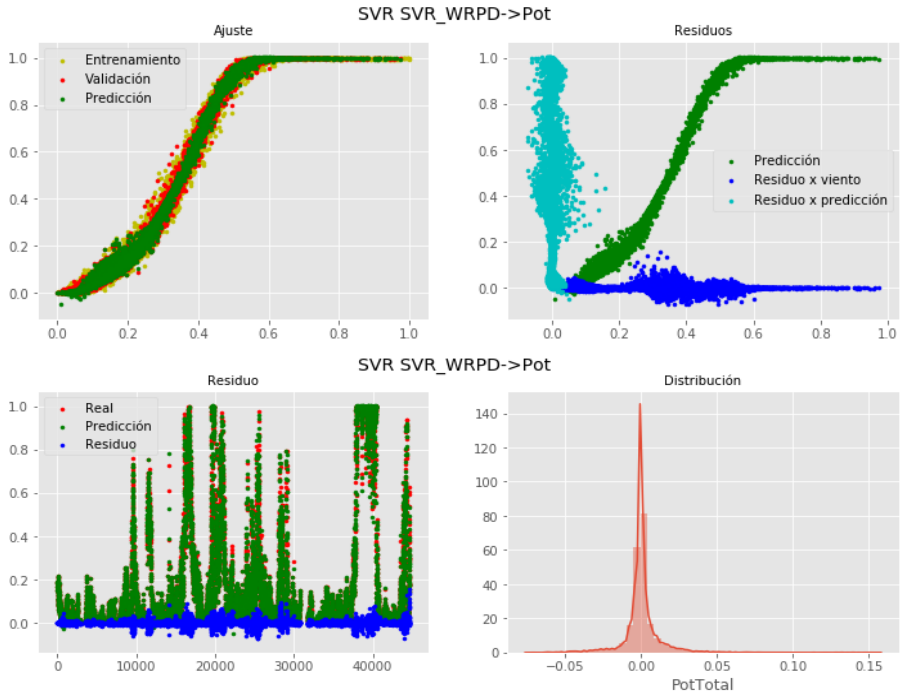


8.2.1.3 SVR (WRPD->Pot)

```
Modelos computados: 15
Precisiones obtenidas:
```

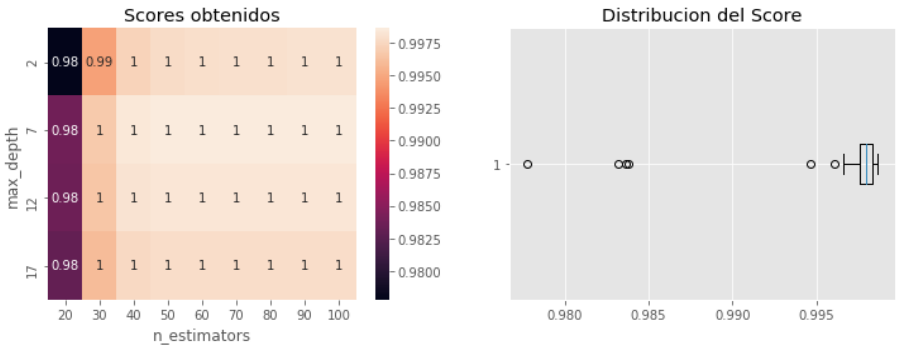



```
Mejores parametros: {'C': 2, 'epsilon': 0.0001}
Mejor accuracy: 0.9980300127451983
CV R2 Score: 0.99803 (+/- 0.00019)
MSE: 0.000152436 RMSE: 0.012346514 MAE: 0.006373485 MAPE:
39.816022212 R2: 0.998173632
Residual: Mean: 0.000527632 Std: 0.012335235
```

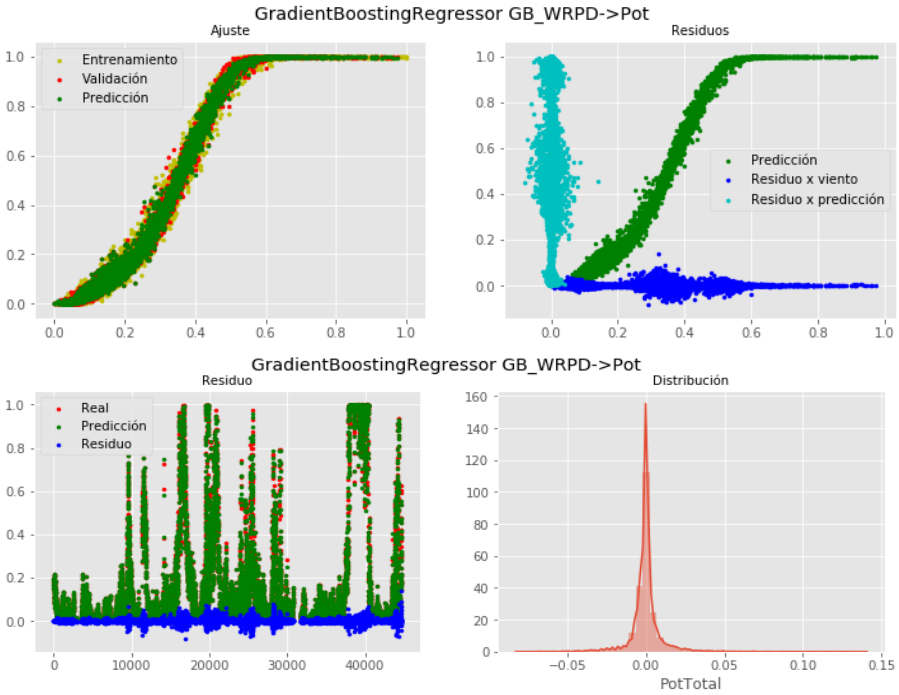


8.2.1.4 Gradient Boosting (WRPD->Pot)

```
Modelos computados: 36
Precisiones obtenidas:
```

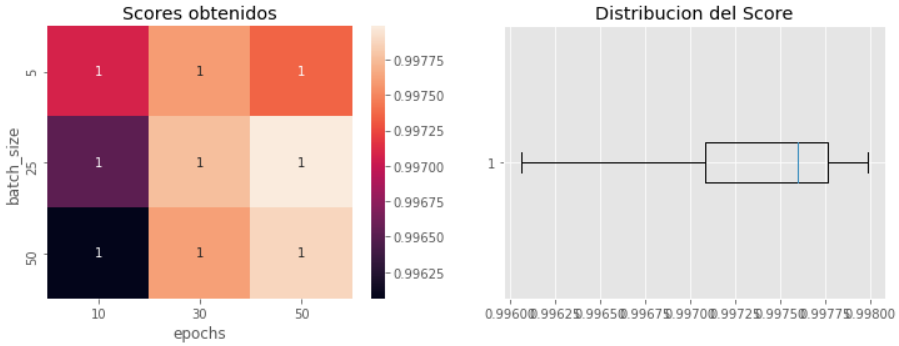



```
Mejores parametros: {'max_depth': 7, 'n_estimators': 70}
Mejor accuracy: 0.9986350450791779
CV R2 Score: 0.99864 (+/- 0.00017)
MSE: 0.000099733 RMSE: 0.009986640 MAE: 0.005229809 MAPE:
48.856652226 R2: 0.998805081
Residual: Mean: -0.000041618 Std: 0.009986554
```



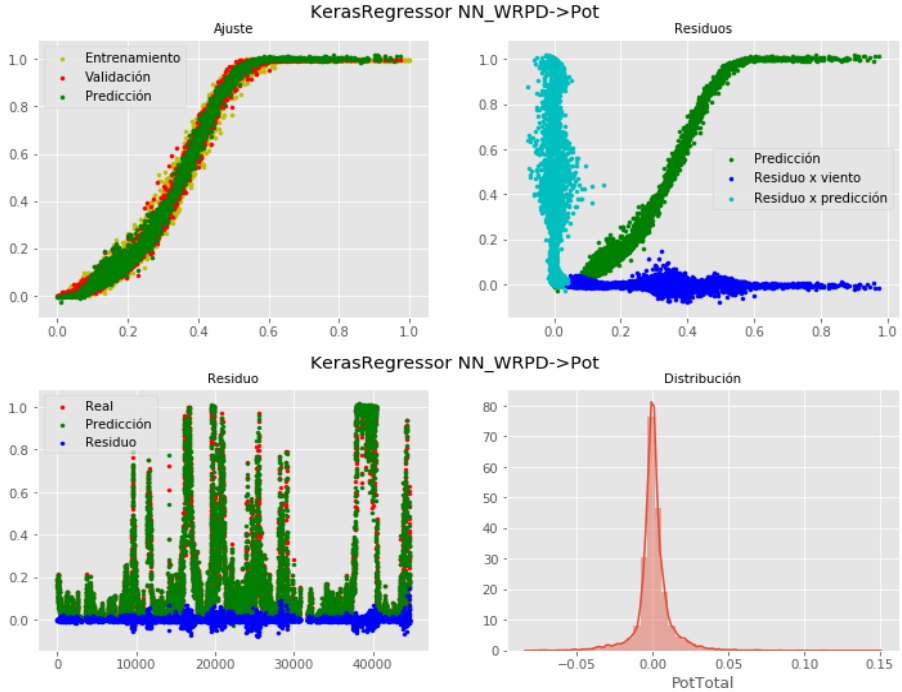
8.2.1.5 Neural Network (WRPD->Pot)

```
Modelos computados: 9
Precisiones obtenidas:
```



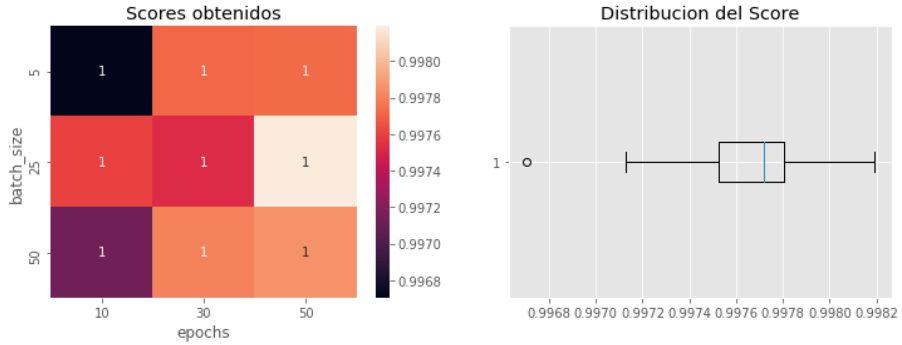

```
Mejores parametros: {'batch_size': 25, 'epochs': 50}
Mejor accuracy: 0.997984557110954
CV R2 Score: 0.99782 (+/- 0.00052)
MSE: 0.000139964 RMSE: 0.011830621 MAE: 0.006848284 MAPE:
38.762512938 R2: 0.998323071
```

```
Residual: Mean: 0.000545086 Std: 0.011818057
```

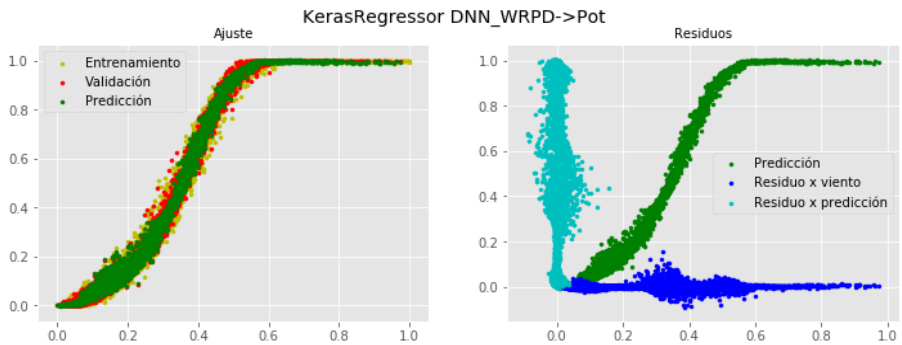


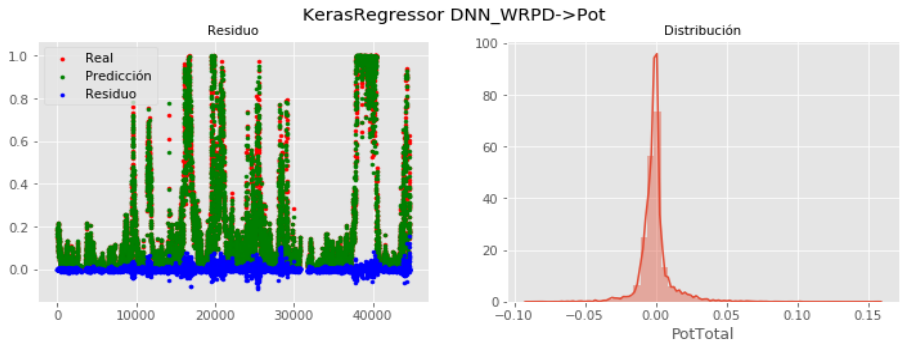
8.2.1.6 Deep Neural Network (WRPD->Pot)

```
Modelos computados: 9
Precisiones obtenidas:
```




```
Mejores parametros: {'batch_size': 25, 'epochs': 50}
Mejor accuracy: 0.9981905066508312
CV R2 Score: 0.99784 (+/- 0.00046)
MSE: 0.000137959 RMSE: 0.011745583 MAE: 0.006460156 MAPE:
63.453162054 R2: 0.998347092
Residual: Mean: -0.000376902 Std: 0.011739534
```





8.2.2 Predicción del pitch

Se incluyen los resultados de la evaluación del ajuste de todos los modelos.

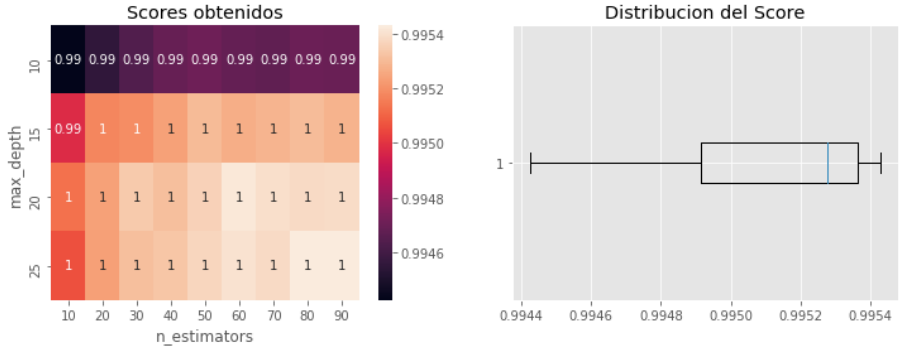
Los nombres modelos se corresponden con la nomenclatura AA_VVV->P, siendo A el algoritmo, V las iniciales de los predictores empleados y P la variable a predecir.

Modelo	CV R2	MSE	RMSE	MAE	MAPE	TST R2
RF_W->Pitch	0,98983	0,00012	0,01116	0,00575	17,22150	0,98986
KNN_W->Pitch	0,99105	0,00011	0,01051	0,00560	17,11895	0,99101
SVR_W->Pitch	0,99112	0,00011	0,01049	0,00548	15,25328	0,99104
GB_W->Pitch	0,99101	0,00011	0,01061	0,00569	18,04693	0,99084
LR_W->Pitch	0,40931	0,00723	0,08500	0,06334	253,55549	0,41176
NN_W->Pitch	0,99064	0,00012	0,01087	0,00578	17,11422	0,99039
DNN_W->Pitch	0,98977	0,00011	0,01058	0,00611	18,08869	0,99088
RF_WR->Pitch	0,99301	0,00009	0,00942	0,00366	9,05029	0,99278
KNN_WR->Pitch	0,99353	0,00008	0,00868	0,00351	8,92294	0,99387
SVR_WR->Pitch	0,99337	0,00008	0,00878	0,00372	11,90463	0,99372
GB_WR->Pitch	0,99306	0,00008	0,00904	0,00365	9,88956	0,99335
LR_WR->Pitch	0,75004	0,00313	0,05592	0,04632	156,08659	0,74545
NN_WR->Pitch	0,98920	0,00009	0,00949	0,00434	13,59275	0,99267
DNN_WR->Pitch	0,99231	0,00014	0,01175	0,00452	8,85289	0,98876
RF_WRP->Pitch	0,99543	0,00006	0,00745	0,00280	5,76981	0,99548
KNN_WRP->Pitc	0,99539	0,00005	0,00742	0,00292	6,38489	0,99552
SVR_WRP->Pitc	0,99536	0,00005	0,00716	0,00296	6,46704	0,99583
GB_WRP->Pitch	0,99516	0,00005	0,00741	0,00295	6,98923	0,99553
LR_WRP->Pitch	0,75201	0,00309	0,05563	0,04601	154,80872	0,74804
NN_WRP->Pitch	0,99365	0,00008	0,00871	0,00416	10,38556	0,99382
DNN_WRP->Pitc	0,99277	0,00008	0,00877	0,00547	16,13558	0,99374

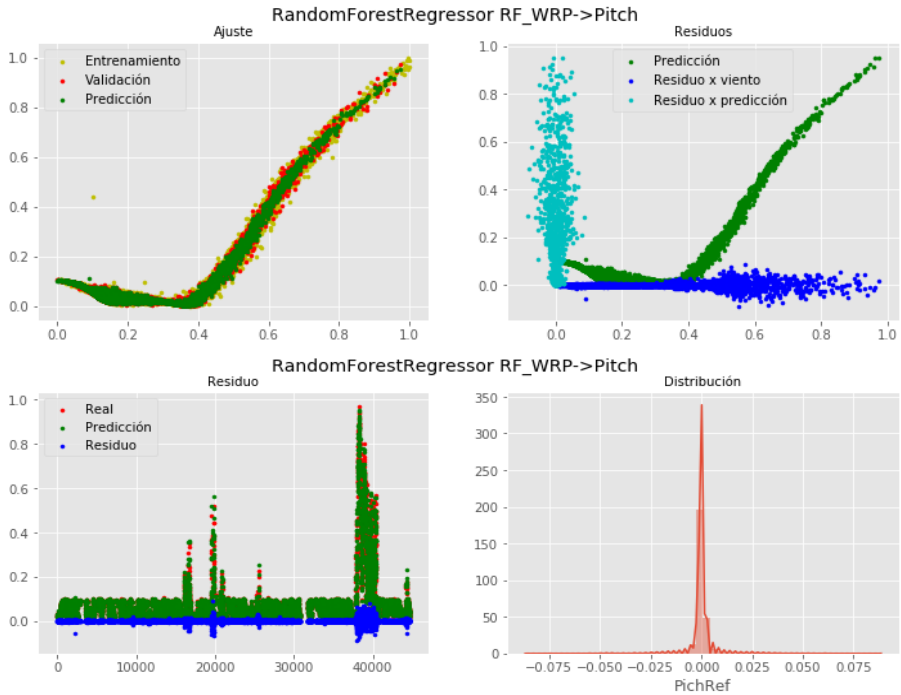
En la predicción de la pich los modelos que mejor resultado han obtenido son lo que usan como predictores: el viento, las revoluciones, y la potencia.

8.2.2.1 Random Forest Regresor (WRP->Pitch)

```
Modelos computados: 36
Precisiones obtenidas:
```

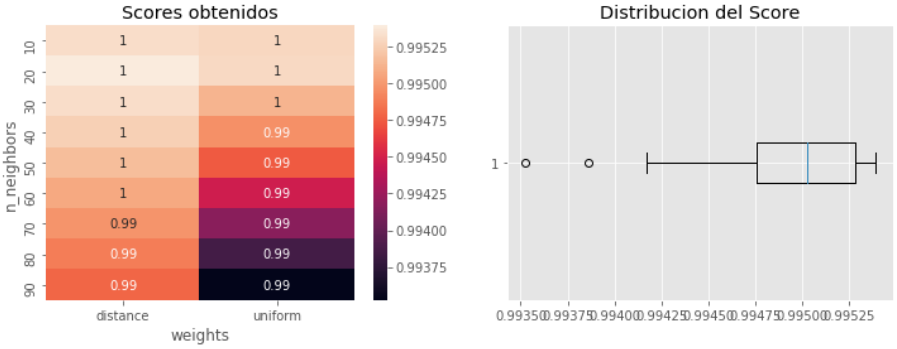



```
Mejores parametros: {'max_depth': 20, 'n_estimators': 90}
Mejor accuracy: 0.9954162759442504
CV R2 Score: 0.99543 (+/- 0.00203)
MSE: 0.000055468 RMSE: 0.007447683 MAE: 0.002795194 MAPE:
5.769806874 R2: 0.995484265
Residual: Mean: 0.000020044 Std: 0.007447656
```

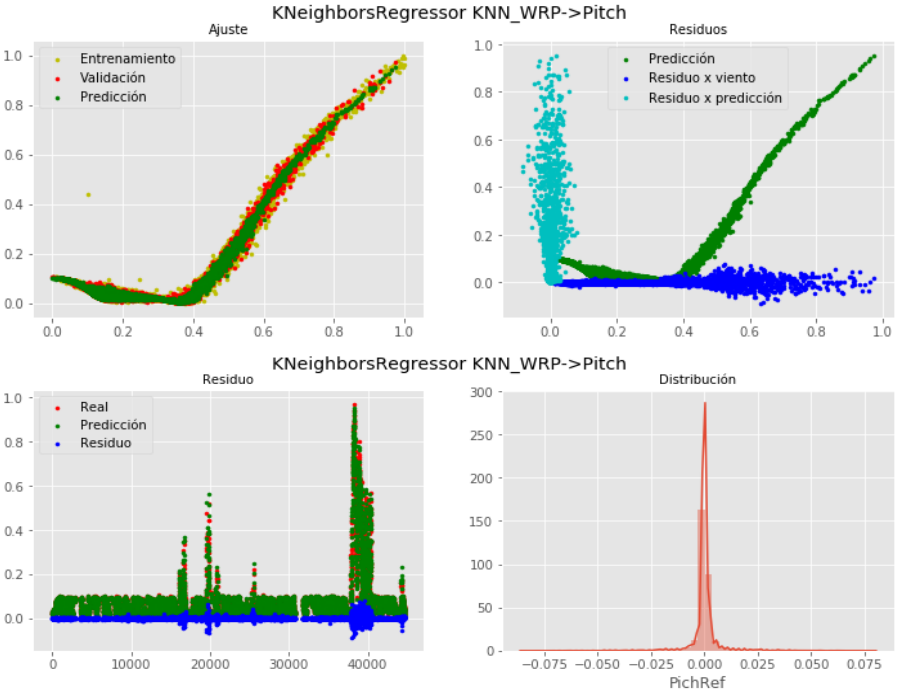


8.2.2.2 K-NN (WRP->Pitch)

```
Modelos computados: 18
Precisiones obtenidas:
```

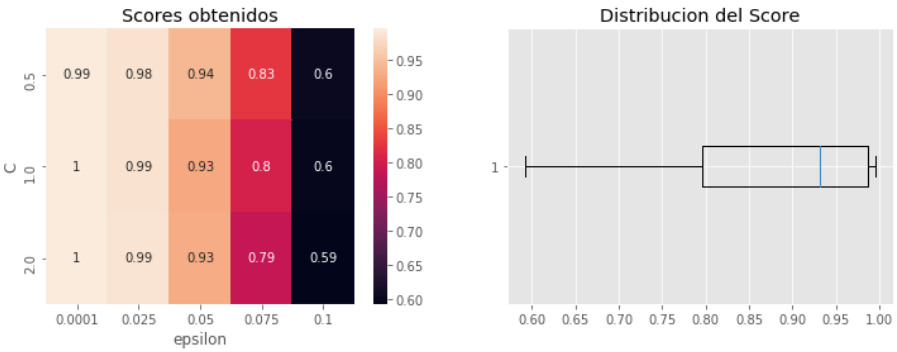



```
Mejores parametros: {'n_neighbors': 20, 'weights': 'distance'}
Mejor accuracy: 0.9953917061941399
CV R2 Score: 0.99539 (+/- 0.00195)
MSE: 0.000054983 RMSE: 0.007415040 MAE: 0.002918080 MAPE:
6.384887926 R2: 0.995523763
Residual: Mean: 0.000177593 Std: 0.007412913
```



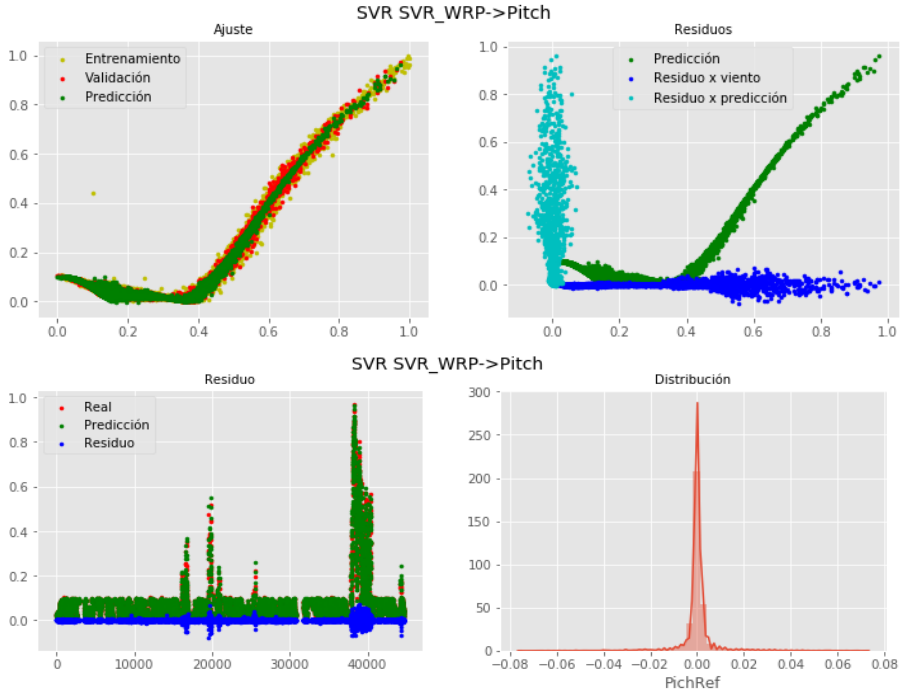
8.2.2.3 SVR (WRP->Pitch)

```
Modelos computados: 15
Precisiones obtenidas:
```



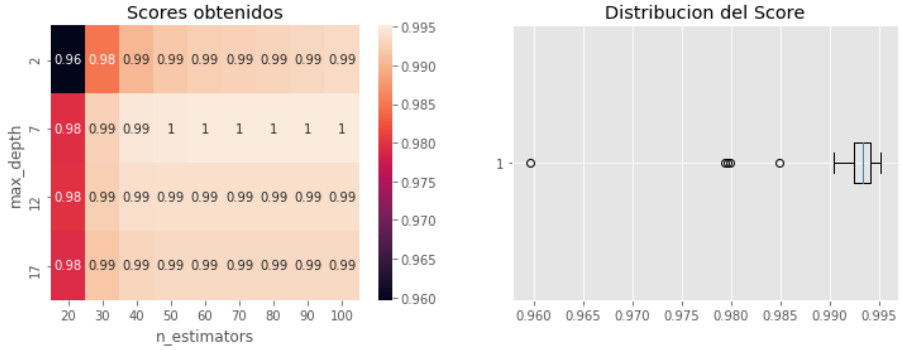

```
Mejores parametros: {'C': 2, 'epsilon': 0.0001}
Mejor accuracy: 0.9953625367651147
CV R2 Score: 0.99536 (+/- 0.00164)
```

```
MSE: 0.000051247 RMSE: 0.007158733 MAE: 0.002956925 MAPE:
6.467041693 R2: 0.995827864
Residual: Mean: 0.000118785 Std: 0.007157748
```

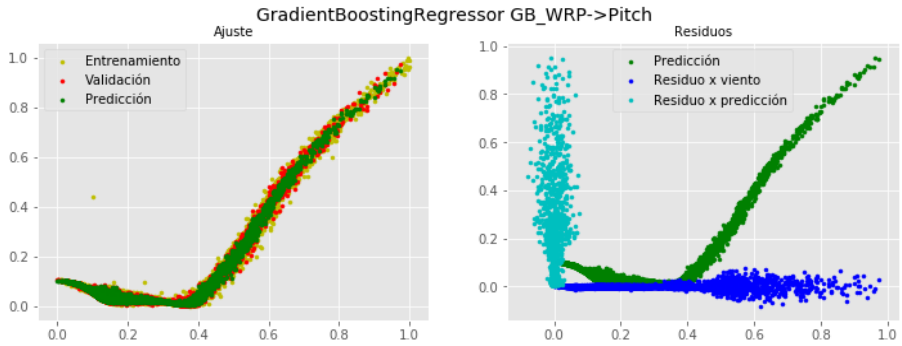


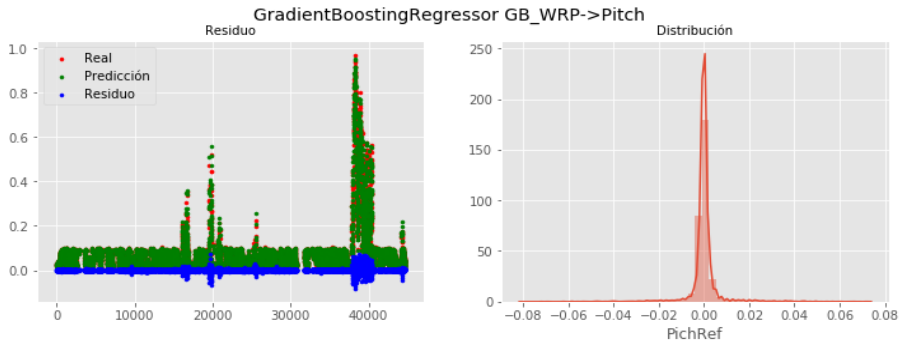
8.2.2.4 Evaluando Gradient Boosting (WRP->Pitch)

```
Modelos computados: 36
Precisiones obtenidas:
```



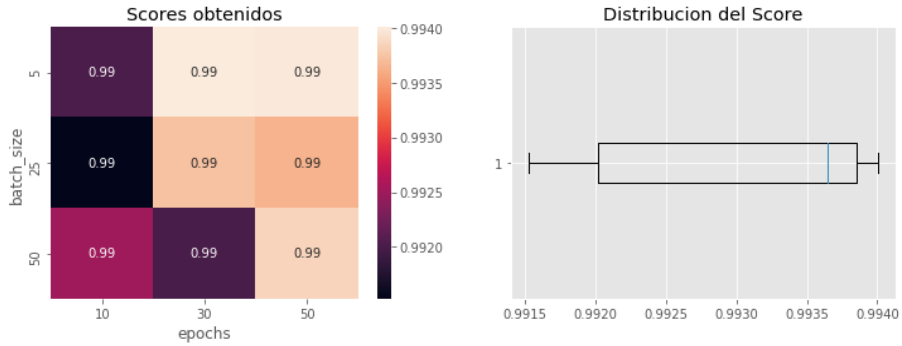

```
Mejores parametros: {'max_depth': 7, 'n_estimators': 90}
Mejor accuracy: 0.9951774574089313
CV R2 Score: 0.99516 (+/- 0.00213)
MSE: 0.000054959 RMSE: 0.007413457 MAE: 0.002953179 MAPE:
6.989231880 R2: 0.995525673
Residual: Mean: 0.000033177 Std: 0.007413383
```



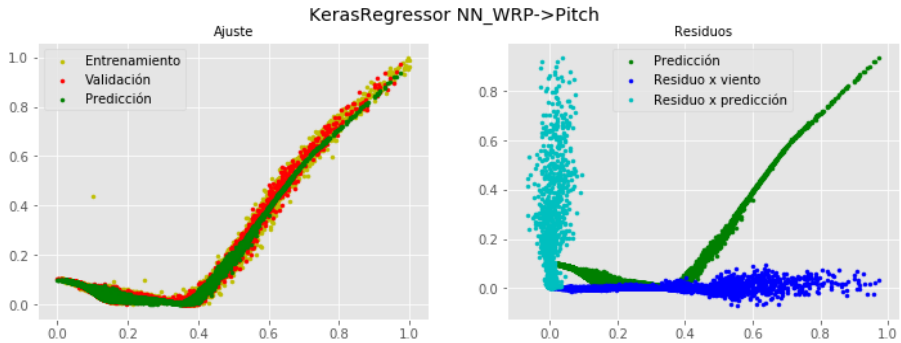


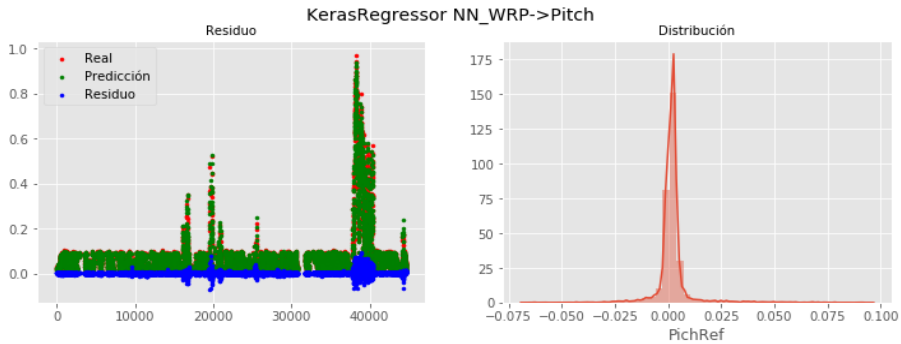
8.2.2.5 Neural Network (WRP->Pitch)

```
Modelos computados: 9
Precisiones obtenidas:
```




```
Mejores parametros: {'batch_size': 5, 'epochs': 30}
Mejor accuracy: 0.9940049363727446
CV R2 Score: 0.99365 (+/- 0.00257)
MSE: 0.000075928 RMSE: 0.008713648 MAE: 0.004162293 MAPE:
10.385561893 R2: 0.993818610
Residual: Mean: 0.002041864 Std: 0.008471036
```

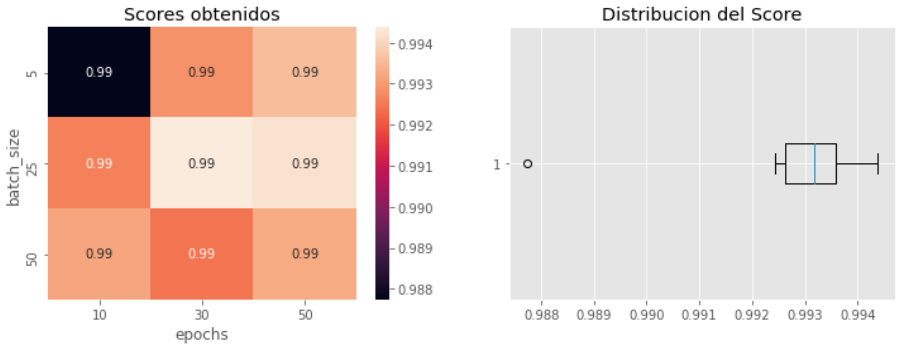




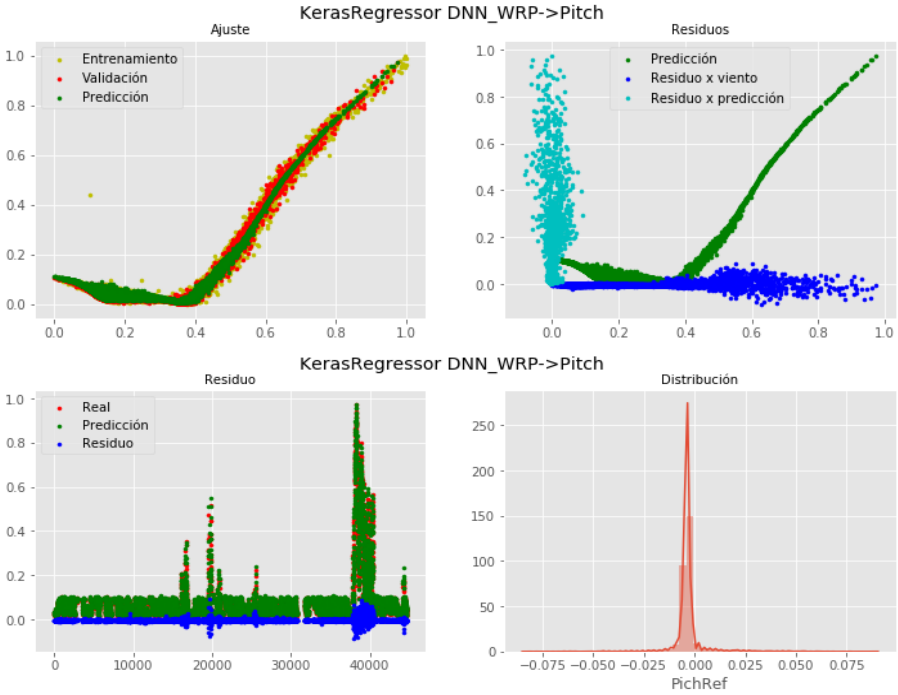
8.2.2.6 Evaluando Deep Neural Network (WRP->Pitch)

```
Modelos computados: 9
```

```
Precisiones obtenidas:
```




```
Mejores parametros: {'batch_size': 25, 'epochs': 30}
Mejor accuracy: 0.9943767481049697
CV R2 Score: 0.99277 (+/- 0.00363)
MSE: 0.000076918 RMSE: 0.008770314 MAE: 0.005473598 MAPE:
16.135575103 R2: 0.993737952
Residual: Mean: -0.003198327 Std: 0.008166340
```



8.3 Anexo III. Código fuente

Debido a la extensión de trabajo, no incluimos en la memoria los scripts y recursos utilizados.

De modo que se podrán encontrar, junto con los resultados en el siguiente repositorio público de internet: https://github.com/copazohm/mdstfm.