Universitat Oberta de Catalunya (UOC)

Máster Universitario en Ciencia de Datos (*Data Science*)

TRABAJO FINAL DE MÁSTER

Área: 004

xxx Mantenimiento Predictivo xxx

xxx Predicción del fallo de máquinas de producción a partir de datos de un sistema de sensores. xxx

---

Autor: Oscar Rojo Martín

Tutor: Lorena Polo Navarro

Profesor: Antonio Lozano Bagén

Donostia-San Sebastián, 8 de enero de 2023

Índice general

Capítulo 1

Introducción

1.1. Introducción

Como hemos visto y vivido durante el último siglo y medio, las máquinas - en general - durante su proceso de vida tienen periodos en que requieren de reparaciones o de sustitución de piezas por el desgaste de estas. Con la evolución de la tecnología inalámbrica y conectada 4G y 5G cada vez existe una mayor oferta de sensores inteligente, que permiten monitorizar las máquinas y en base a los comportamientos, predecir estos últimos. El TFM que presento va destinado a generar una WEB-APP donde se mostrarán los datos registrados por los sensores recopilados de diferentes maquinarias y se realizarán evaluaciones de machine learning con el fin de predecir anomalías o comportamientos repetitivos en dichas maquinarias.

1.2. Contexto y justificación del Trabajo

Este trabajo trata concebir una herramienta / aplicación web con un enfoque de aprendizaje automático para la detección de anomalías en sistemas de control industrial basados en datos de medición.

El contexto de la elección del presente trabajo se basa en:

- Resido en Donostia - Gipuzkoa.
- Trabajo como Analista de Datos en una pequeña empresa y por ahora el nicho de mercado es la creación de Paneles de Control de carácter económico financiero para las empresas de servicios. Esta web-app puede ser una buena herramienta para poder ofrecer otros servicios a las empresas, en particular, PYMEs radicadas en la provincia que se dedican al sector industrial y no tengan suficiente capacidad económica para contratar estos servicios a grandes empresas de consultoría.

- Como decía antes, la gran cantidad de pequeña empresa que tiene maquinas fresadoras, de corte, semiautomáticas, etc.. y que se beneficiarían de esta tecnología para poder reducir costes y evitar parones no previstos o innecesarios.

1.3. Objetivos del Trabajo

Una web-app que muestre tanto los datos provenientes de los sensores de las máquinas como las predicciones en el caso en que las variables cambien o en caso de no cambiar, predecir las fechas en que se puede producir la anomalía.

1.4. Impacto en sostenibilidad, ético-social y de diversidad

"Actuar de manera honesta, ética, sostenible, socialmente responsable y respetuosa con los derechos humanos y la diversidad, tanto en la práctica académica como en la profesional, y diseñar soluciones para mejorar estas prácticas."

1.5. Enfoque y método seguido

El método elegido es el siguiente:

- Analizar el origen de los datos.
- Recopilar información.
- Almacenamiento.
- Extracción, Transformación y Carga.
- Analizar los datos.
- Visualización / representación gráfica.
- Generación de modelo predictivo que permita detectar las anomalías, fallos, etc de la maquinaria de donde se recolecta la información.
- Integrar todo ello en una aplicación web para desplegarla en internet y con acceso restringido de usuario y contraseña.

La estrategia es ir avanzando con rapidez ante posibles cambios y la de realizar entregas periódicas del trabajo.

Capítulo 2

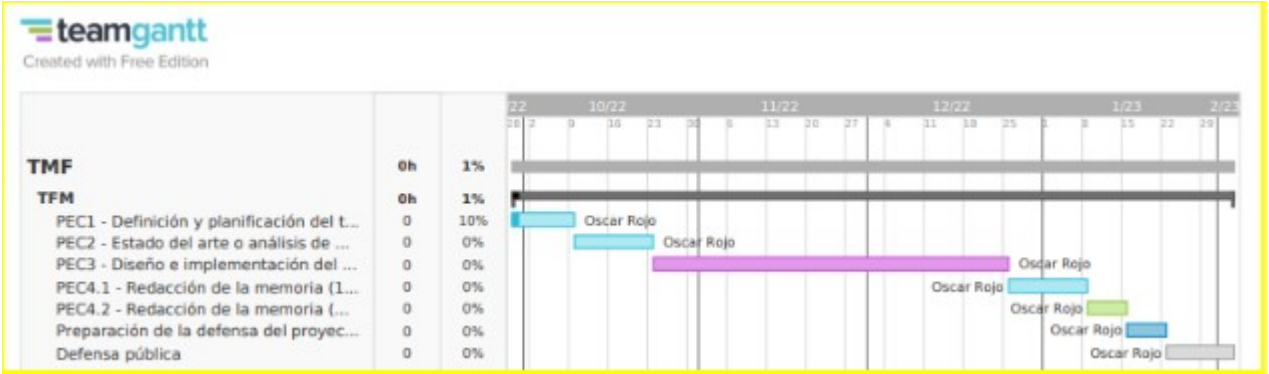
Planificación del Trabajo

2.1. Planificación del trabajo

Este trabajo se realizará bajo la planificación realizada mediante un diagrama de Gantt a traves de la aplicación https://app.teamgantt.com/projects/gantt?ids=3246752

2.1.1. Detalle planificación

Cuadro 2.1: Detalle planificación TFM.



2.2. Planificación del trabajo

Cuadro 2.2: Planificación

Etapa	Título	Descripción	Inicio	Fin
PEC1	Definición y planificación del trabajo final	Fase que tiene como objetivo el de definir cuál es la temática del TFM, el objetivo del mismo, justificar su relevancia, definir los objetivos principales y la planificación temporal del proyecto.	28/09/2022	09/10/2022
PEC2	Estado del arte o análisis del mercado del proyecto	Recopilación de información. Entrevistar a empresarios de la industria del metal en Gipuzkoa.	10/10/2022	23/10/2022
PEC3	Diseño e implementación del trabajo	Programar y desarrollar la web-app utilizando Python como lenguaje de programación y la librería principal Django uno de los frameworks de servidores web más populares, para configurar un entorno de desarrollo y crear aplicaciones web. Para llegar al final a ponerlo en producción en AWS	24/10/2022	25/12/2022
PEC4	Redacción de la memoria.	Subdividido en 2 partes, Aquí se deberá mostrar tanto todo el trabajo realizado en anteriores bloques como la información recopilada.	26/12/2022	08/01/2023
Preparación de la defensa del proyecto.	Preparación de la defensa del proyecto.	Mediante la creación de un vídeo de presentación del Trabajo de Fin de Máster.	09/01/2023	15/01/2023
Defensa del proyecto.	Defensa del proyecto.	Delante de un tribunal evaluador, se realizará la defensa del TFM. Esta defensa síncrona constará de una presentación-resumen y de la respuestas a las consultas realizadas por el tribunal evaluador	16/01/2023	22/01/2023

Capítulo 3

Estado del Arte

3.1. Introducción

El mantenimiento predictivo (MP) es una técnica que emplea el análisis basado en datos para detectar condiciones de trabajo anómalas y predecir los riesgos de fallos futuros de los activos. A pesar de su amplia aplicación en las industrias manufacturera y del petróleo y el gas, la aplicación del Mantenimiento predictivo en instalaciones de infraestructura, en el sector de PYMEs industriales en Gipuzkoa, es escasa de acuerdo a las consultas realizadas a diferentes empresarios del sector industrial en el territorio. En este sentido destaco:

1. Utilizan los sensores instalados en la maquinaria para controlar el output, el producto (calidad, resistencia, etc)

2. Varios empresarios no les interesa esta tecnología de MP, ya que creen que les supone un esfuerzo muy grande en personal, en tiempo y en coste (no inversión).

3. Directamente quitan o desconectan todos los sensores conectados a las máquinas.

La llagad de las TICs y de la IA ofrece una gran oportunidad para mejorar la práctica del mantenimiento de infraestructuras mediante la integración de técnicas de MP.

3.2. Un poco de historia

La Revolución industrial constituyo uno de los periodos más decisivos de la trayectoria de la Humanidad. Las principales áreas a las que se dirigió la revolución industrial fueron la tecnológica, la socioeconómica y la cultural, cambiando el mundo tal y como es hoy, nuestra forma de vivir y la organización social, política y económica actual. Destacamos los siguientes cambios:

1. Utilización de nuevos materiales, principalmente el hierro y el acero;
2. Uso de nuevas fuentes de energía, como el carbón, la máquina de vapor, la electricidad, el petróleo y los motores de combustión interna motores de combustión interna;
3. La invención de nuevas máquinas para la industria textil (por ejemplo, la rueda de torsión);
4. La organización de la producción en el sistema tipo fábrica que conocemos hoy y la división del del trabajo;
5. Desarrollo de los transportes y medios de comunicación como: locomotoras de vapor, barcos de vapor barcos de vapor, aviones, telégrafo, radio;
6. Aplicación de los principios científicos en la industrialización.

Todos los cambios tecnológicos han contribuido en gran medida a la utilización de los recursos naturales y han permitido la producción en masa de bienes. Hoy en día, el mantenimiento es una cuestión de suma importancia en el entorno industrial y no hace más que mejorar el entorno de trabajo los rendimientos y la rentabilidad.

3.3. Qué es el mantenimiento

El objetivo o elemento principal de las empresas son los ***productos***. Por ello es necesario un plan de gestión de productos que garantiza el buen desarrollo y la organización de todas las etapas, recursos y medios implicados en el proceso de producción El mantenimiento presente desde la fase de diseño de un producto, debe garantizar el buen funcionamiento del producto en función de su vida útil. dentro de las cinco fases en el ciclo de vida de un producto: idea, definición, realización, uso y reciclaje. Fases comprendidas en 3 categorías:

1. La fase incipiente: idea, definición, realización;
2. La fase intermedia: uso;
3. Fase final: reciclaje.

La fase intermedia es la que más tiempo dura, ya que el producto está en fase de uso, por lo que es necesario mantenerlo para que funcione en parámetros óptimos. El concepto de mantenimiento está definido por la norma UNE-EN 13306:2018: es la **Combinación de todas las acciones técnicas, administrativas y de gestión realizadas durante el ciclo de vida de un**

Cuadro 3.1: Análisis comparativo de las ventajas e inconvenientes del mantenimiento reactivo y proactivo.

Table 1. Reactive and proactive maintenance advantages and disadvantages.

	Advantages	Disadvantages
Reactive maintenance	Low initial costs Easy to implement	Unscheduled stops High associated costs (work over schedule, delivery urgency, manufacturing urgency) Poorly optimized resources
Proactive maintenance	Increased system reliability Minimizing logistical stops Reduction of non-scheduled stops Lowers costs (optimizing parts, optimizing work) Maintenance planning Optimization of logistic support	High initial set-up costs The cost reduction is not immediate Not reliable for all equipment

elemento, destinadas a conservarlo o a devolverlo a un estado en el que pueda desempeñar la función requerida.

El mantenimiento de los equipos ha evolucionado a lo largo del tiempo, desde el mantenimiento que se realizaba sólo cuando los elementos sufrían daños hasta los métodos modernos, el mantenimiento proactivo y el predictivo, siendo este último el más popular en la actualidad. El mantenimiento realizado sólo en caso de daños o fallos, ha demostrado que es un método ineficiente por fallos en el sistema. Esto derivó en el mantenimiento programado (revisión total si hay problemas) y su consecuente problema de sustituir elementos sin haber llegado a su vida útil. Por ejemplo: **La Correa de distribución del coche, esta tiene programada su sustitución con 100.000 km, pero el nº de km recorridos sin que se rompa puede ser superior a los 135.000 km. el problema es que se rompa... pero eso es otra historia**

El siguiente paso fue el mantenimiento predictivo. Según **A state of the art of predictive maintenance techniques** las tareas de mantenimiento se pueden clasificar en:

1. Mantenimiento reactivo: el sistema se utiliza hasta que se producen averías. Cuando se produce una avería, los Cuando se produce una avería, se sustituyen los elementos dañados para restablecer la capacidad de funcionamiento del sistema.

2. Mantenimiento proactivo - las acciones de mantenimiento se planifican o tienen lugar como resultado de los indicadores de seguimiento. La planificación puede realizarse mediante sensores que anuncian las primeras fases de una avería. Este tipo de mantenimiento tiene como objetivo mantener la funcionalidad del sistema.

Independientemente del tipo de mantenimiento que se aplique, el objetivo de estas actividades es reducir al máximo las paradas causadas por las averías. Para poder minimizar las paradas, es necesario entender los modos y mecanismos de los fallos.

Debemos de tener en cuenta que la los procesos industriales se llevan a cabo con sistemas eléctricos y mecánicos.

3.4. El mantenimiento predictivo

El mantenimiento predictivo es un concepto enmarcado como un enfoque de mantenimiento proactivo.

Una de las técnicas más utilizadas es el **mantenimiento predictivo condicional**, el cual se realiza observando ciertos parámetros o ciertos componentes del sistema. En él, el estado del sistema se presenta en tiempo real, en función de los parámetros seguidos, siendo algunos de los más importantes:

- Análisis de vibraciones: es el método más eficiente para detectar problemas en los equipos que realizan movimientos rotativos.
- Análisis acústico: puede detectar o monitorizar grietas en tuberías o conductos;
- Análisis de aceites de lubricación: se analizan las partículas encontradas en los aceites utilizados para determinar el grado de desgaste de los componentes se analizan.
- Análisis de partículas en el entorno de trabajo: método generalmente utilizado en equipos que trabajan en un entorno fluido.
- Análisis de corrosión: e realizan mediciones de ultrasonido para determinar la corrosión en diferentes estructuras.
- Análisis térmico: se utiliza especialmente en el caso de sistemas mecánicos y eléctricos para detectar sobrecalentamiento en general.
- Análisis de rendimiento: una técnica eficaz para determinar los problemas de funcionamiento del sistema.

Utilizando alguno o algunos de los métodos indicados, se desarrollan soluciones de mantenimiento para equipos y sistemas sistemas. Suponiendo que 30 % de los costes de mantenimiento se deben a una mala planificación del mantenimiento, con costes añadidos en el proceso de producción, el mantenimiento predictivo es un método interesante en la industria, donde las paradas pueden causar grandes pérdidas.

3.4.1. El mantenimiento predictivo en la Industria 4.0

En estos momentos en que la industria se enfrenta a dificultades debido al desarrollo de la tecnología, la disminución de los recursos naturales, la incidencia de guerras y desastres naturales y al mismo tiempo cuestiones sociales como la globalización y el aumento de la edad de jubilación de la mano de obra, que producen efectos económicos. Además los consumidores de hoy en día quieren una variedad y una alta calidad de los productos, así como una calidad servicios de calidad durante su uso.

La visión de la Industria 4.0 viene con un impacto masivo en la forma actual de trabajar en la industria caracterizada por un nuevo modelo de interacción socio-tecnológica: **Nuevas fábricas inteligentes**.

La industria 4.0 la define el artículo **The concept Industry 4.0** como que *representa básicamente una sinergia entre el Internet de las Cosas, el Internet de los Servicios y por por supuesto, el proceso industrial.* y estos se pueden aprovechar para desarrollar una estrategia de mantenimiento predictivo utilizando las tecnologías mencionadas, siendo esto alcanzable con reducidos recursos materiales y recursos humanos. La Industria 4.0 está dando forma a los procesos de fabricación y mantenimiento y a los que se producen de la interacción máquina-hombre.

3.4.2. Técnicas

Las podemos dividir en tres técnicas básicas:

- Mantenimiento usando los sensores ya instalados.
- Mantenimiento basado en sensores de prueba.
- Mantenimiento basado en la técnica de señales de prueba.

Todas tienen como objetivo proporcionar una mejora en el análisis de fallos, aunque representan enfoques diferentes. Las dos primeras técnicas pueden considerarse **pasivas** y la tercera **activa**, cuya aplicación tiene en cuenta las respuestas del bucle y las pruebas en tiempo real.

3.5. Bombas de Agua

"Las bombas de agua son componentes esenciales y costosos en los sistemas de distribución de agua. Su papel es clave en la transmisión y distribución del líquido, conectando todas las etapas del proceso. Estos equipos son propensos a fallos debido a su complejo funcionamiento en diferentes condiciones, tales como altas temperaturas, sobre-esfuerzos y uso prolongado al

aire libre. En consecuencia, las bombas de agua representan un riesgo para la seguridad del sistema de distribución de agua potable en las ciudades."

3.5.1. Fallos en las Bombas de Agua

La falla de esta pieza clave del sistema de distribución de agua puede tener graves consecuencias, como costos de reparación o reemplazo y pérdidas de ingresos. Además, la frecuencia de fallos puede afectar la confiabilidad del sistema a largo plazo. Por ello, es fundamental conocer los componentes de las bombas de agua y entender cómo pueden fallar. Algunos de estos componentes incluyen el platillo o polea, la carcasa, los rodamientos, el cierre, la turbina, los circuitos eléctricos y de aceite, y los controles de presión. Un problema en cualquiera de estos elementos puede poner en riesgo la operación de la bomba y llevar a su retiro del sistema de distribución de agua en las ciudades.

3.5.1.1. Curvas de fallo en equipos industriales

En la actualidad existen seis curvas que se consideran modelos de fallos para equipos industriales. Estas curvas se pueden dividirse en dos grupos. El primer grupo es el mantenimiento basado en el tiempo y el segundo grupo es el mantenimiento basado en la condición.

Los mantenimientos basados en el tiempo, también conocidos como mantenimiento preventivo, se realizan de forma regular a intervalos predeterminados de tiempo. Estos mantenimientos tienen como objetivo evitar fallos y prolongar la vida útil de la bomba. Algunas tareas comunes de mantenimiento preventivo incluyen:

- Inspecciones visuales: consisten en examinar la bomba y sus componentes para detectar signos de desgaste o daño.
- Limpieza: se realiza para eliminar suciedad, polvo y otros contaminantes que pueden afectar el funcionamiento de la bomba.
- Lubricación: se aplica para reducir la fricción y el desgaste en las piezas móviles de la bomba.
- Cambio de filtros: se realiza para mantener el flujo de agua limpio y evitar obstrucciones o daños a la bomba.

El intervalo de tiempo entre los mantenimientos preventivos puede variar según el tipo de bomba y el entorno en el que se utiliza. Es importante seguir el plan de mantenimiento recomendado por el fabricante para garantizar el correcto funcionamiento y prolongar la vida útil de la bomba.

Existen tres tipos comunes de curvas utilizadas en el mantenimiento basado en el tiempo:

- Curva de bañera: muestra el tiempo de vida útil de un componente según la frecuencia de uso. Los componentes que se usan con mayor frecuencia tienen una vida útil más corta y requieren mantenimiento más frecuente.
- Curva de desgaste: muestra el tiempo de vida útil de un componente según el desgaste o el deslizamiento. Los componentes que experimentan más desgaste tienen una vida útil más corta y requieren mantenimiento más frecuente.
- Curva de basura: muestra el tiempo de vida útil de un componente según la acumulación de basura o contaminantes. Los componentes que se contaminan más rápidamente tienen una vida útil más corta y requieren mantenimiento más frecuente.

Estas curvas se utilizan para establecer planes de mantenimiento preventivo y determinar cuándo deben realizarse las tareas de mantenimiento.

Los mantenimientos basados en la condición son aquellos que se realizan en función de la condición actual de la bomba, en lugar de siguiendo una frecuencia predeterminada. Algunos ejemplos de mantenimientos basados en la condición son:

- Mantenimiento predictivo: se basa en el monitoreo de la bomba para detectar posibles fallos antes de que ocurran. Esto puede incluir pruebas de rendimiento o mediciones de parámetros como la vibración o el consumo de energía.
- Mantenimiento basado en el tiempo de funcionamiento: se realiza en función de la cantidad de tiempo que ha estado en funcionamiento la bomba, en lugar de seguir una frecuencia fija.
- Mantenimiento basado en el uso: se realiza en función de la cantidad de uso que ha tenido la bomba, en lugar de siguiendo una frecuencia fija.

Los mantenimientos basados en la condición pueden ser más eficientes que otros tipos de mantenimiento, ya que se realizan cuando es necesario y no siguiendo una frecuencia predeterminada. Sin embargo, es importante contar con un sistema de monitoreo adecuado para poder llevar a cabo este tipo de mantenimiento de manera efectiva.

Las curvas de mantenimiento basado en la condición son gráficos que se utilizan para determinar cuándo se deben realizar mantenimientos en función de la condición actual de la bomba. Algunos ejemplos de curvas que se pueden utilizar en el mantenimiento basado en la condición son:

- Curva de Degradación: muestra cómo la condición de la bomba se va deteriorando con el tiempo. Se pueden establecer umbrales en la curva para indicar cuándo se debe realizar un mantenimiento preventivo o correctivo.

- Curva de Cumplimiento: muestra el porcentaje de tiempo en el que la bomba cumple con ciertos parámetros de rendimiento o calidad. Se pueden establecer umbrales en la curva para indicar cuándo se debe realizar un mantenimiento preventivo o correctivo.
- Curva de Carga: muestra el porcentaje de carga de la bomba en función del tiempo. Se pueden establecer umbrales en la curva para indicar cuándo se debe realizar un mantenimiento preventivo o correctivo.

  Estas curvas pueden ser útiles para planificar el mantenimiento de las bombas de agua y tomar decisiones sobre cuándo realizar ciertas tareas de mantenimiento.

3.5.2. Métodos de mantenimiento de las bombas de agua

El mantenimiento es considerado una actividad estratégica que garantiza la credibilidad de la operación de equipos y procesos industriales. El mantenimiento debe buscar la intervención en los equipos reduciendo el tiempo de intervención, dejando el sistema inaccesible por el menor tiempo posible . Existen varios métodos de mantenimiento para las bombas de agua, algunos de ellos son:

- Mantenimiento preventivo: se realiza de forma regular para evitar fallos y prolongar la vida útil de la bomba. Puede incluir tareas como inspecciones visuales, limpieza y lubricación de piezas, y cambio de filtros.
- Mantenimiento correctivo: se realiza cuando se detecta un problema o fallo en la bomba. Puede incluir reparaciones o reemplazo de piezas dañadas.
- Mantenimiento predictivo: se basa en el monitorización de la bomba para detectar posibles fallos antes de que ocurran. Esto puede incluir pruebas de rendimiento o mediciones de parámetros como la vibración o el consumo de energía.
- Mantenimiento adaptativo: se realiza para adaptar la bomba a nuevas condiciones de trabajo o cambios en el sistema en el que está instalada.

Es importante seguir un plan de mantenimiento adecuado para garantizar el correcto funcionamiento y prolongar la vida útil de la bomba de agua.

3.5.2.1. Flujo de trabajo del sistema de mantenimiento predictivo

Es fundamental conocer los posibles fallos que pueden presentar los equipos en el **mantenimiento predictivo (PdM)**, ya que este sistema se basa en un conjunto de indicadores que reflejan el estado del equipo y son utilizados para tomar decisiones. Por lo tanto, se llevan a cabo

diferentes tipos de análisis para detectar estos fallos, como el análisis físico-químico, el análisis de la firma eléctrica, el análisis de partículas, la inspección termográfica, el método de emisión acústica y el análisis de gases disueltos, entre otros. Para hacer predicciones sobre el estado de los equipos, una buena opción es utilizar algoritmos de aprendizaje automático. Por lo general, el proceso de **PdM** se divide en cuatro fases. La primera fase es la adquisición de datos, en la que se recogen los datos de los sensores y los análisis mencionados anteriormente. Luego, se lleva a cabo una fase de preprocesamiento para limpiar los datos adquiridos. A continuación, los datos procesados se pasan a un algoritmo de aprendizaje automático para entrenar un modelo. Finalmente, se utiliza el modelo para predecir anomalías en un nuevo conjunto de datos. Si se detecta alguna anomalía, se informa al usuario para que decida si es necesario intervenir.

Cuadro 3.2: Textos analizados

Titulo	**Autores**	**Año**
A new paradigm of cloud-based predictive maintenance for intelligent manufacturing	Wang, J., Zhang, L., Duan, L.	2015
Continuous time predictive maintenance scheduling for a deteriorating system	L. Dieulle, C. Berenguer, A. Grall and M. Roussignol	2001
Maintenance 4.0: Intelligent and Predictive Maintenance System Architecture	Cachada, Ana and Barbosa, Jose and Leitño, Paulo and Gcraldcs, Carla A.S. and Deusdado, Leonel and Costa, Jacinta and Teixeira, Carlos and Teixeira, João and Moreira, António H.J. and Moreira, Pedro Miguel and Romero, Luís	2018
A predictive maintenance policy with imperfect monitoring	Giuseppe Curcurù, Giacomo Galante, AlbertoLombardo	2010
Predictive maintenance techniques	Bin Lu; David B. Durocher; Peter Stemper	2009
Predictive Maintenance of Power Substation Equipment by Infrared Thermography Using a Machine-Learning Approach	Ullah I, Yang F, Khan R, Liu L, Yang H, Gao B, Sun K.	2017

3.6. Líneas de investigación

3.6.1. Resumen

Cuadro 3.3: Tabla resumen

Titulo	A new paradigm of cloud-based predictive maintenance for intelligent manufacturing	Continuous time predictive maintenance scheduling for a deteriorating system	Maintenance 4.0: Intelligent and Predictive Maintenance System Architecture	A predictive maintenance policy with imperfect monitoring	Predictive maintenance techniques	Predictive Maintenance of Power Substation Equipment by Infrared Thermography Using a Machine-Learning Approach
Autores	Wang, J., Zhang, L., Duan, L.	L. Dieulle, C. Berenguer, A. Grall and M. Roussignol	Cachada, Ana and Barbosa, Jose and Leitño, Paulo and Gcraldcs, Carla A.S. and Deusdado, Leonel and Costa, Jacinta and Teixeira, Carlos and Teixeira, João and Moreira, António H.J. and Moreira, Pedro Miguel and Romero, Luís	GiuseppeCurcurúGiacomoGalanteAlbertoLombardo	Bin Lu; David B. Durocher; Peter Stemper	Ullah I, Yang F, Khan R, Liu L, Yang H, Gao B, Sun K.
Año	2015	2001	2018	2010	2009	2017
Resumen	Se genera un nodo de detección y computación en la nube de bajo coste, un middleware de agente móvil y bibliotecas numéricas de código abierto. El intercambio de información y la interacción se logran mediante un agente móvil que distribuye los algoritmos de análisis al nodo de detección y computación en la nube para procesar localmente los datos y compartir los resultados del análisis. El "agente móvil" mejora la flexibilidad y adaptabilidad del sistema, reduce la transmisión de datos en bruto y responde instantáneamente a los cambios dinámicos de las operaciones y tareas.	Se desarrolla un modelo matemático para el coste del sistema mantenido utilizando la teoría de los procesos de renovación. Los experimentos numéricos muestran que la tasa de coste de mantenimiento en un horizonte infinito puede minimizarse mediante una optimización conjunta del umbral de reposición y los tiempos de inspección aperiódica, y que la estructura de mantenimiento propuesta rinde más que las políticas clásicas de mantenimiento preventivo, que pueden tratarse como casos particulares.	El artículo describe la arquitectura de un sistema de mantenimiento inteligente y predictivo, alineado con los principios de la Industria 4.0, que considera el análisis avanzado y en línea de los datos recogidos para la detección temprana de la ocurrencia de posibles fallos de la máquina, y apoya a los técnicos durante las intervenciones de mantenimiento proporcionando un soporte de decisión inteligente guiado.	Adopta un modelo estocástico para el proceso de degradación e hipotetizando el uso de un sistema de monitorización imperfecto, el procedimiento actualiza, mediante un enfoque bayesiano, la información a-priori, utilizando los datos procedentes del sistema de monitorización.	Aqui se analiza la importancia del Mantenimiento Predictivo para las aplicaciones de procesos industriales e investiga teconologías emergentes, como la evaluación de la eficiencia energética en línea y la supervisión continua del estado. También se discuten dos métodos para la detección de fallos en los rodamientos y la estimación de la eficiencia energética.	Se aborda un mecanismo de prevención subestaciones eléctricas utilizando un enfoque de visión por ordenador aprovechando las imágenes térmicas infrarrojas, para prevenir problemas de contacto, cargas irregulares, grietas en el aislamiento, relés defectuosos, uniones de terminales y otros problemas similares, aumentan la temperatura interna de los instrumentos eléctricos.
Algoritmo	Se seleccionan cinco características estadísticas, tres de frecuencia y los coeficientes del modelo autorregresivo de la envolvente de la corriente del motor para diagnosticar fallos motor. SVM(Support vector machine)	Modelo matemático para el coste del sistema mantenido utilizando la teoría de los procesos de renovación	Neural Networks, Bayesian Belief Network, Fuzzy Logic Prediction	Markovian degradation process	Algoritmo matemático. No especifica	Perceptrón multicapa (MLP) para clasificar las condiciones térmicas de los componentes de las subestaciones eléctricas en clases "defectuosas" y "no defectuosas". Las redes neuronales artificiales (ANN) se utilizan para la clasificación de un defecto en diferentes materiales.
Variables, Parámetros	Se extraen 600 conjuntos de vectores de características correspondientes a seis motores diferentes. Cada vector consta de 23 características: 5 estadísticas, 3 dominio de la frecuencia y 15 coeficientes AR.	Variables de decisión de mantenimiento: el umbral de sustitución preventiva y el programa de inspección basado en el estado del sistema.	Tipo de defecto, Dia y hora, ID, Presión, Temperatura, Humedad, Vigraciones, Ruido Operacional		Velocidad, Ruido, Vibración, Desajustes	150 imágenes térmicas de diferentes equipos eléctricos.
Caso de Aplicación	Se utilizan seis motores de inducción con diferentes modos de fallo en un sistema de prueba de motores para imitar los procesos de fabricación distribuidos.		Máquina de prensado		Industria Papelera. La planta de envasado de cartón de Weyerhaeuser en Manitowoc (Wisconsin, EE.UU.) supervisa tres motores de inducción críticos: un motor de soplado de 75 CV, un motor de bomba hidráulica de 50 CV y un motor de compresor de 200 CV.	Subestaciones eléctricas

3.6.2. A predictive maintenance policy with imperfect monitoring

El método de mantenimiento predictivo que se ejecuta con un sistema de monitorización sujeto a errores. Para la generación de datos, se considera un sistema sujeto a degradación continua que se modela mediante cadenas de Markov, que permite realizar una simulación eficaz para seguir la evolución del proyecto principal: Correlacionar la eficacia de los equipos de análisis y adquisición de datos con la reducción de costes en el mantenimiento predictivo en comparación con otros métodos de mantenimiento.

Su valoración depende de la calidad de los sensores y del coste de reparación de los daños. Cuando se modifican los parámetros de simulación y disminuye la calidad de los sensores utilizados para la adquisición, la eficacia del mantenimiento predictivo disminuye considerablemente, situándose por debajo del mantenimiento preventivo.

Los resultados de esta simulación refuerzan la idea de adoptar sistemas de mantenimiento predictivo, que mejoran los costes de mantenimiento, incluso cuando se toma el coste de los equipos de análisis (siempre que los sensores y los sistemas de adquisición proporcionen datos cualitativos).

3.6.3. A new paradigm of cloud-based predictive maintenance for intelligent manufacturing

Aquí se proponen un sistema de mantenimiento predictivo basado en un nuevo sistema de control, en una nueva arquitectura en la nube que utiliza un cliente móvil en lugar de la clásica arquitectura servidor-cliente, con el que se pretende incrementar la flexibilidad/adaptabilidad del sistema, reducir el volumen de datos brutos que se transmiten y mejorar el tiempo de respuesta a los cambios dinámicos que se producen en los sistemas monitorizados.

Inconvenientes: la planificación automática del mantenimiento, la gestión de la energía, la gestión de análisis y transmisión de datos, almacenamiento de datos, implementación de sistemas a gran escala. Para abordar algunas de estos problemas, como se menciona desde el principio, esta solución propone el uso de agentes móviles para distribuir tareas en la nube.

El **agente móvil** es un método derivado del estudio IA que pretende dividir las tareas y su ejecución en paralelo. Los agentes están representados por programas de software que migran dentro de la red. Las principales características de un agente móvil son la autonomía, la inteligencia y la adaptabilidad, siendo adecuado para sistemas Cloud. Esto reduce la cantidad de datos transmitidos y permite la modificación y adaptación de los algoritmos que se ejecutan en los equipos, aumentando así su versatilidad.

Para ello se utilizaron seis motores de corriente alterna, cinco de ellos con diferentes defectos. Con el uso de agentes móviles, se envían programas de software para que se ejecuten y adquieran

diferentes parámetros, según la metodología de mantenimiento predictivo. Luego, los datos se procesan con algoritmos específicos, consiguiendo una característica del equipo monitorizado. Los resultados confirmaron la eficacia del sistema y proporcionaron datos relevantes sobre el estado de los motores analizados en comparación con el motor estándar.

3.6.4. Predictive maintenance techniques

Consideran como principal el rendimiento de los motores eléctricos y los fallos en los rodamientos, siendo éstos los problemas más comunes. Estudio basado en dos casos prácticos realizados en un motor de soplador de 75 hp y en un motor de compresor de aire de 200 hp en una planta activa. De ambos experimentos se obtuvieron resultados y conclusiones que confirman una desalineación en el eje del motor del soplador y una pérdida de rendimiento y una degradación para el motor del compresor debido a pequeños periodos de sobrecarga.

3.6.5. Continuous time predictive maintenance scheduling for a deteriorating system

Los investigadores proponen un método utilizando sistemas que están sometidos a un deterioro continuo, para tratar:

- La creación de una estructura para aplicar el mantenimiento predictivo condicional, y la determinación de un modelo matemático de análisis de costes a partir de la aparición de los fallos.
- Demostrar que los costes de mantenimiento a largo plazo pueden reducirse sustancialmente, utilizando dicho enfoque.

Utiliza un modelo aleatorio para simular el deterioro continuo del sistema. Puede simular un envejecimiento en el caso de los modelos mecánicos, la evolución de los productos defectuosos en el caso de una línea de producción, el nivel de corrosión / erosión en el caso de las estructuras. El modelo elegido ofrece una rentabilidad teniendo en cuenta el tiempo necesario para realizar el mantenimiento durante un periodo indefinido. El objetivo es realizar una estimación mínima de los costes teniendo en cuenta el nivel de desgaste y el momento adecuado para detener la máquina. Como el mantenimiento basado en la monitorización continua de los parámetros tiene un mejor rendimiento que otros métodos de mantenimiento, el modelo matemático presentado debe ser capaz de prevenir aleatorios. La solución se perfila en torno a dos ideas principales:

- Determinar el método de mantenimiento del sistema en cuestión, en función de la evolución de los daños.

- Determinar y establecer los datos para la siguiente inspección. Las simulaciones realizadas ofrecen una amplia gama de soluciones para determinar el modo de mantenimiento o la utilización del equipo hasta un fallo inminente para obtener el resultado deseado: la eficiencia de los costes de mantenimiento.

3.6.6. Predictive Maintenance of Power Substation Equipment by Infrared Thermography Using a Machine-Learning Approach

Los autores proponen una solución de mantenimiento predictivo mediante la **termografía** utilizando una cámara termográfica de infrarrojos y utilizando técnicas de ML (Machine Learning). Este tipo de anomalía se produce principalmente en los equipos de energía, donde los contactos imperfectos, la corrosión, el aislamiento imperfecto, etc. provocan el sobrecalentamiento del equipo.

Es método permite monitorizar el equipo en funcionamiento. El análisis se realiza sobre un total de 150 fotos, en las que se determinan unos 300 puntos de interés. Mediante Machine Learning los puntos de interés se clasifican como defectos o funcionamiento en parámetros. Utilizando la cámara de infrarojos, como en el deterioro de los equipos eléctricos el principal efecto del envejecimiento de los cables está representado por un aumento de la resistencia, el análisis termográfico puede ofrecer información relevante para la implementación de las prácticas de mantenimiento predictivo.

3.6.7. Maintenance 4.0: Intelligent and Predictive Maintenance System Architecture

Aquí se proponen una innovadora estructura de mantenimiento predictivo mantenimiento predictivo basado en la arquitectura OSA-CBM (Open System Architecture for Condition Based Monitoring) donde describe la arquitectura de un sistema de mantenimiento inteligente y predictivo, alineado con los principios de la Industria 4.0, que considera el análisis avanzado y en línea de los datos recogidos para la detección más temprana de la ocurrencia de posibles fallos de la máquina y apoya a los técnicos durante las intervenciones de mantenimiento proporcionando un soporte de decisión inteligentemente guiado.

Esta estructura consta de seis pasos:

1. Adquisición de datos: proporciona el acceso a los datos digitalizados del sensor o transductores digitalizados y registra estos datos.

2. Manipulación de datos: puede realizar transformaciones de señales de uno o varios canales y aplicar algoritmos especializados de extracción de características a los datos recopilados.

3. Detección de estados: realiza la supervisión de las condiciones comparando las características con los valores esperados o los límites operativos y devolviendo indicadores de condiciones y/o alarmas. límites operativos y devolviendo indicadores de estado y/o alarmas.

4. Evaluación de la salud: determina si la salud del sistema está está sufriendo una degradación teniendo en cuenta las tendencias en el historial de salud, el estado operativo y el historial de mantenimiento. de salud, el estado operativo y el historial de mantenimiento.

5. Evaluación de pronóstico: proyecta el estado de salud actual estado de salud actual del activo teniendo en cuenta una estimación de los perfiles de uso futuros.

6. Generación de consejos: proporciona recomendaciones relacionadas con de mantenimiento y modificación de la configuración del activo, teniendo en cuenta configuración del activo, teniendo en cuenta el historial operativo, los perfiles de perfiles de misión actuales y futuros y las limitaciones de recursos.

3.7. La gestión de los datos

Con el avance de la Industria 4.0, la implantación del mantenimiento predictivo cambia en comparación con el clásico, sobre todo en el campo de la gestión de los datos, debido al exponencial incremento en el número de datos que se generan.

Los procesos de gestión de datos tienen como objetivo proporcionar un procedimiento estándar para el análisis de datos con el fin de evaluar y proponer una solución basada en los datos analizados.

Destacamos algunos de los procesos de datos disponibles más importantes: Por supuesto, hay muchos procesos de datos disponibles, cada uno de los cuales se adapta a una necesidad de gestión de datos diferente; algunos de los más importantes son:

- El modelo de proceso Cross Industry Standard Process for data Mining (CRISP-DM).
- El modelo de proceso Sample, Explore, Modify, Model and Access (SEMMA).
- La metodología Team Data Science Process.

Todos estos métodos representan un enfoque novedoso en términos de técnicas de mantenimiento predictivo, convirtiéndose en el estándar en las técnicas de mantenimiento debido a la multitud de beneficios.

Con los nuevos avances tecnológicos y el movimiento de la Industria 4.0, las fábricas , las maquinarias empezaron a ser más avanzadas y más conectadas. Esto abrió el camino a la **monitorización de parámetros en tiempo real**, que es el pilar fundacional en el mantenimiento predictivo.

La rapidez y facilidad con la que ahora se adquieren y envían los datos es algo que hay que aprovechar. Con las nuevas innovaciones, como el Internet de las cosas y el Internet de los servicios, aparecen nuevos retos, siendo la gestión y el procesamiento de datos uno de ellos.

En los últimos años, la ciencia de datos ha acaparado mucha atención y ha dedicado grandes esfuerzos a desarrollar análisis sofisticados, mejorar los modelos de datos y cultivar nuevos algoritmos. Sin embargo, estos proyectos pueden enfrentarse a retos organizativos y sociotécnicos a medida que avanzan, como la falta de visión, estrategia y objetivos claros, un énfasis sesgado en cuestiones técnicas, la falta de reproducibilidad y la ambigüedad de roles, por nombrar algunos. Estos retos contribuyen a un bajo nivel de madurez en los proyectos de ciencia de datos que se gestionan al azar. Dicho esto, las metodologías de gestión de proyectos y procesos son beneficiosas para los proyectos de ciencia de datos. Metodologías como éstas pueden ayudar a tener éxito y superar algunos de los problemas mencionados anteriormente. Por otro lado, a los equipos de ciencia de datos puede resultarles difícil ceñirse a una metodología de proyecto. Los proyectos de ciencia de datos pueden planificarse utilizando muchas metodologías. Una encuesta realizada por KDnuggets en 2014 muestra que la metodología más utilizada es CRISP-DM, con un 43. % de las respuestas.

3.7.1. CRISP-DM

CRISP-DM son las siglas de Cross-Industry Standard Process for Data Mining, y fue creado a mediados de los años 90 por SPSS y Teradata. Describe enfoques comunes utilizados por los expertos en minería de datos. Se desglosa en 6 fases del ciclo de vida de un proyecto de minería de datos, como se puede ver en la Figura: comprensión del negocio, comprensión de datos, preparación de datos, modelado, evaluación y despliegue.

- Entendimiento del negocio - Esta es la primera etapa, y es aquí donde se definen los objetivos y necesidades del proyecto desde el punto de vista empresarial. Aquí es también donde se proyecta el plan del proyecto.

- Comprensión de los datos - Es fundamental analizar y familiarizarse con los datos una vez recopilados. El principal objetivo de esta fase es estudiar los datos y garantizar su calidad. Esta fase suele estar vinculada a la fase de comprensión del negocio, ya que es fundamental tener en cuenta la calidad y las características de los datos para garantizar que los objetivos del proyecto estén claramente definidos.

- Preparación de datos - El objetivo principal de esta fase es manejar los datos y crear el conjunto de datos final que se introducirá en el modelado. En este proceso se lleva a cabo la selección de características, la transformación de datos y la limpieza.
- Modelización - En esta fase se eligen y aplican las técnicas de modelización. Se divide el conjunto de datos, se generan los conjuntos de datos de prueba y de entrenamiento, y se construyen y aplican los modelos.Este paso también puede estar relacionado con la fase de preparación de los datos si es necesario realizar alguna transformación o selección para obtener mejores resultados.
- Evaluación: es la fase en la que se analizan y revisan los resultados del modelo para comprobar si se han alcanzado o no los objetivos definidos en la fase de comprensión del negocio.
- Despliegue: en esta fase del ciclo de vida, el modelo se organiza y se entrega al cliente.

El uso del aprendizaje automático en este tipo de problemas es una práctica muy extendida hoy en día.

Capítulo 4

Modelos

En este capítulo analizaremos algunos algoritmos de aprendizaje automático.

4.0.0.1. Aprendizaje automático aplicado a la PdM

El auge de los sistemas de mantenimiento predictivo (**PdM**) se debe en gran medida a la creciente utilización de enfoques basados en datos, como el aprendizaje automático. Este tipo de tecnología ha revolucionado el mantenimiento, permitiendo predecir fallos en los equipos y otros sucesos relevantes durante su ciclo de vida. Los enfoques de aprendizaje automático utilizados en los sistemas de **PdM** pueden variar dependiendo de los datos disponibles. Muchos artículos de investigación han abordado este tema. En el contexto del **PdM**, las metodologías de apoyo a las máquinas más importantes son la agrupación, la clasificación, la regresión y la detección de anomalías.

4.1. Aprendizaje automático

La cantidad de datos que se recopilan ha aumentado enormemente en los últimos años, lo que ha llevado a un auge en el uso de técnicas de aprendizaje automático. Este subcampo de la ciencia de datos es muy valioso porque permite obtener información valiosa de los datos de una manera que sería imposible para una persona promedio debido al tamaño de la información.

El aprendizaje automático puede dividirse en cuatro subcampos principales:

- Aprendizaje supervisado.
- Aprendizaje no supervisado.
- Aprendizaje semisupervisado.
- Aprendizaje por refuerzo.

4.2. Aprendizaje supervisado

En este subcampo del aprendizaje automático, el objetivo es predecir el resultado de una tarea basándose en las características de los datos. El modelo recibe un conjunto de características y una variable objetivo. Con esto, el modelo aprende la función que relaciona las características con la variable objetivo. El aprendizaje supervisado puede dividirse en dos tipos principales:

- Regresión: Encuentra un modelo de relación (función matemática) que relaciona un conjunto de variables numéricas o categóricas con una variable numérica.
- Clasificación: Encontrar un modelo de relación (función matemática) que relacione un conjunto de variables numéricas o categóricas con una variable categórica.

4.2.1. Modelos de aprendizaje supervisado

Muchos algoritmos de aprendizaje supervisado pueden utilizarse para resolver problemas de clasificación. Los más relevantes para este trabajo son las redes bayesianas, Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest, K-Nearest Neighbors (KNN), Artificial Neural Network (ANN). A continuación se describe cada algoritmo en su propia sección.

4.2.1.1. Redes bayesianas

Una red bayesiana es un modelo gráfico de relaciones de probabilidad entre un conjunto de variables. La red debe representarse. A continuación, se determinan los parámetros, lo que dificulta su aplicación sin la opinión de un experto. sin la opinión de un experto. BN tampoco tiene éxito con grandes conjuntos de datos porque las redes grandes no son viables en términos de tiempo y espacio.

4.2.1.2. Naive Bayes

Naive Bayes es un clasificador probabilístico basado en el teorema de Bayes. Se basa en la suposición ingenua de que cada par de características es independiente, lo que significa que cada característica es independiente de las demás para una clase determinada. Este método es sencillo, intuitivo y puede funcionar con gran eficacia.

4.2.1.3. Support vector machine

Las máquinas de vectores soporte son algoritmos de aprendizaje supervisado que utilizan una función kernel para transformar las características de los datos en un espacio de alta dimensión.

El objetivo es encontrar un hiperplano que separe los datos en dos clases. El algoritmo se basa en la idea de que la distancia entre los puntos de datos y el hiperplano es el factor más crítico para determinar la clase del punto de datos.

4.2.1.4. Decision trees and random forest

Los árboles de decisión son un algoritmo muy utilizado para problemas de clasificación. Son fáciles de entender y explicar una vez que muestra la decisión de dividir los datos. Este algoritmo es robusto frente al ruido de los datos. También proporciona un alto rendimiento para un tiempo de cálculo relativamente rápido. Un problema de este algoritmo es que le resulta difícil manejar conjuntos de datos de alta dimensión. Otro problema es que sin el uso adecuado de la poda puede llevar fácilmente a un sobreajuste. Para resolver este problema, se utilizan los bosques aleatorios. Un bosque aleatorio es una colección de árboles de decisión entrenados utilizando una muestra bootstrap de los datos originales. El objetivo es encontrar un conjunto de árboles de decisión que sean más robustos al ruido de los datos.

4.2.1.5. Vecinos más próximos (K-Nearest Neighbors)

K-Nearest Neighbors (KNN) es un método no paramétrico de clasificación. Es un algoritmo sencillo que puede utilizarse para encontrar los vecinos más cercanos de un punto dado en un conjunto de datos. El objetivo es encontrar la clase de los vecinos más cercanos. Se trata de un algoritmo simple de aprendizaje perezoso cuya eficacia depende en gran medida del valor del parámetro k y del tamaño de los datos.

4.2.2. Red neuronal artificial

Las redes neuronales artificiales (RNA) son una interconexión entre modelos computacionales y una estructura en capas. Esta red consta de nodos (neuronas artificiales), conexiones ponderadas y funcionalidad. La idea de las RNA es parametrizar una estructura de forma repetitiva para ajustar los parámetros durante el entrenamiento. Las neuronas están dispuestas en capas, y cada una de ellas está asociada a 2 variables, un conjunto de pesos y un sesgo. Podemos ver esta estructura en la Figura 2.2. Las RNA pueden dividirse en 3 fases: entrada, procesamiento y salida. Cuanto mayor es la fase de procesamiento, más profunda es la red neuronal.

4.2.3. Problemas de datos desequilibrados

Los conjuntos de datos desequilibrados son un problema común en el aprendizaje automático. En esta sección se discutirán algunos problemas comunes que surgen al tratar con conjuntos de datos desequilibrados. Es frecuente encontrar problemas con conjuntos de datos desequilibrados en el mundo real. Por ejemplo, en la detección de fallos hay una gran diferencia entre el número de puntos de datos con fallos y sin fallos. Esto puede suponer un problema a la hora de entrenar un modelo de detección de fallos, ya que los algoritmos no funcionan bien porque suelen estar diseñados para manejar conjuntos de datos equilibrados. Una de las soluciones más comunes para tratar este tipo de problema es utilizar técnicas de remuestreo.

4.2.4. Overfitting y Underfitting

El overfitting y el underfitting son dos problemas frecuentes en los algoritmos de aprendizaje automático. El sobreajuste(overfitting) se produce cuando el algoritmo se ajusta a los datos de entrenamiento y memoriza su ruido. Esto conduce a un deterioro de la generalización de las propiedades del modelo, lo que se traduce en un bajo rendimiento cuando se aplica a los datos de prueba. Los conjuntos de datos con tamaños pequeños son más propensos al sobreajuste que los conjuntos de datos con tamaños grandes, aunque puede ocurrir en conjuntos de datos grandes debido a la complejidad de los datos. Por otro lado, el infraajuste se produce cuando el algoritmo no puede detectar la variabilidad de los datos. Esto conduce a un rendimiento deficiente tanto en los datos de entrenamiento como en los de prueba. Para solucionar el overfitting y el underfitting, se pueden utilizar varias técnicas. Algunas de ellas son:

- Validación cruzada: Esta técnica divide los datos en uno o varios conjuntos de entrenamiento y de prueba. El objetivo es entrenar el modelo sólo con el conjunto de entrenamiento original.
- Regularización: es una técnica que penaliza el modelo para evitar el sobreajuste. También puede ser un hiperparámetro que puede ajustarse en función del algoritmo utilizado.
- Selección de características: es una técnica que selecciona las características más relevantes de los datos.
- Detención temprana: es una técnica que detiene el proceso de entrenamiento cuando el modelo no mejora.
- Entrenar con más datos: entrenar con más datos puede ser útil para que los algoritmos detecten mejor la señal. El problema es que si los datos añadidos añaden ruido, esta técnica no es útil.

Las técnicas relevantes para este trabajo son la selección de características y el entrenamiento con más datos.

4.2.5. Métricas más comunes para la clasificación

Es crucial evaluar el rendimiento del modelo para saber si tiene un buen rendimiento y validarlo. Se pueden utilizar muchas métricas para evaluar el rendimiento del modelo. A continuación se describirán las métricas más comunes, cada una en su sección.

4.2.5.1. Precisión

La precisión es el porcentaje de predicciones correctas. Es la métrica más utilizada en el aprendizaje automático. Se calcula dividiendo el número de predicciones correctas por el número total de predicciones. La precisión es una buena métrica si los datos están equilibrados. De lo contrario, puede dar una falsa sensación de alcanzar un alto rendimiento. Donde:

- TP: verdadero positivo
- TN: verdadero negativo
- FP: falso positivo
- FN: falso negativo

4.2.5.2. Precisión

La precisión se refiere a la capacidad de un modelo para predecir correctamente los patrones convencionalmente considerados positivos. Se calcula dividiendo el número de patrones positivos predichos correctamente por el número total de patrones positivos predichos. En otras palabras, se divide el número de verdaderos positivos por la suma de verdaderos positivos y falsos positivos.

$$Precision = \frac{TP}{TP + FP} \tag{4.1}$$

4.2.5.3. Recuperación

La recuperación mide cuántos patrones positivos se clasifican correctamente. Se puede calcular dividiendo el número de verdaderos positivos entre la suma de verdaderos positivos y verdaderos negativos.

$$Recall = \frac{TP}{TP + FN} \tag{4.2}$$

4.2.5.4. ROC AUC

La área bajo la curva (AUC) es una medida muy utilizada para evaluar el rendimiento de un clasificador binario. Esta métrica proporciona una visión general del rendimiento del clasificador. La AUC ROC es una curva de probabilidad que muestra la tasa de verdaderos positivos frente a la tasa de falsos positivos.

$$TFP = \frac{FP}{TN + FP} \tag{4.3}$$

4.2.5.5. F1 Score

La F1 Score es una medida que combina la precisión y la recuperación en una sola métrica. Se calcula como la media armónica de estas dos medidas. Esta métrica puede ser útil para evaluar el desempeño y la fiabilidad del modelo, ya que busca encontrar un equilibrio entre la precisión y la recuperación.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4.4}$$

La matriz de confusión es una herramienta comúnmente utilizada para evaluar el rendimiento de un modelo en problemas de clasificación binaria. Una de las métricas derivadas de la matriz de confusión que se suele utilizar es la precisión. Sin embargo, esta métrica puede no ser apropiada cuando los datos están desequilibrados, es decir, cuando una de las clases es mucho más frecuente que la otra. En este caso, la clase mayoritaria puede tener un mayor impacto en el rendimiento del modelo, lo que puede dar lugar a una alta precisión pero un mal rendimiento con la clase minoritaria.

Cuando se trata de conjuntos de datos desequilibrados, se pueden utilizar algunas métricas para evaluar el rendimiento del modelo. Algunas de las métricas que pueden ser útiles son la F1 Score, que es una medida promedio ponderada de precisión y recuperación, y el AUC ROC, que es el área bajo la curva ROC. Estas dos medidas no están sesgadas ni hacia las clases mayoritarias ni hacia las minoritarias.

4.3. Aprendizaje no supervisado

El aprendizaje no supervisado se centra en buscar patrones y relaciones entre las características de los datos de entrada, por lo que se basa principalmente en el uso de datos no etiquetados. Estos métodos pueden ser útiles para descubrir patrones previamente desconocidos en un conjunto de datos.

Los usos más comunes del aprendizaje no supervisado son:

- Clustering: un enfoque para encontrar patrones similares y relaciones entre características en el conjunto de datos, agrupando los puntos de datos con características comunes.
- Reglas de asociación: como su nombre indica, son simples reglas de asociación que ayudan a descubrir relaciones entre conjuntos de datos aparentemente independientes. Este enfoque se utiliza con grandes conjuntos de datos transaccionales. Más concretamente, en estudios de cestas de mercado para analizar los hábitos de compra de los clientes.
- Reducción de la dimensionalidad: es el proceso de proyectar conjuntos de datos de alta dimensión a un espacio de menor dimensión. Uno de los métodos más utilizados es el Análisis de Componentes Principales (ACP). Para esta tesis, la agrupación es la metodología explorada. La siguiente sección presenta una visión general de los algoritmos de clustering existentes.

4.3.1. Algoritmos de clustering

Debido al aumento de los datos y de la potencia de cálculo, se desarrollan constantemente nuevos algoritmos para responder a las necesidades de las empresas. Con este aumento, es esencial estructurar una taxonomía para catalogar los diferentes algoritmos. El enfoque más utilizado en la literatura es distinguir entre algoritmos de clustering basados en partición, jerárquicos, basados en densidad, basados en malla y basados en modelos. Los algoritmos basados en la partición dividen los datos en un conjunto de k conglomerados y, a continuación, asignan cada punto de datos al conglomerado más cercano.

Algunos ejemplos de algoritmos basados en particiones son k-means, k-medoids, fuzzy c-means

Los algoritmos jerárquicos organizan los datos en forma de árbol. La raíz del árbol es el primer clúster, y las hojas son los clústeres más cercanos a la raíz. Algunos ejemplos de algoritmos jerárquicos son BIRCH, CURE, ROCK.

Los algoritmos basados en la densidad se utilizan para encontrar clusters en los datos cercanos entre sí en regiones de alta densidad. Algunos ejemplos de algoritmos basados en la densidad son DBSCAN, OPTICS y Mean-Shift.

En los algoritmos basados en rejilla, los datos se dividen en una estructura de rejilla definida. Cuantifica las áreas de los objetos en un número finito de celdas que forman una estructura de rejilla en la que se implementan todas las operaciones para la agrupación. La ventaja del método es su rápido tiempo de procesamiento, que suele ser independiente del número de objetos de datos, y sigue dependiendo únicamente de las múltiples celdas de cada dimensión en el espacio cuantificado. Algunos ejemplos de algoritmos basados en cuadrículas son WaveCluster, STING, CLIQUE. Por último, los algoritmos basados en modelos pretenden optimizar el ajuste

entre la base de datos dada y un modelo concreto para cada conglomerado. Algunos ejemplos de algoritmos basados en modelos son EM, COBWEB, SOM. Para esta tesis se utilizaron algoritmos basados en la partición y algoritmos basados en la densidad, en concreto, DBSCAN y k-means.

4.3.2. Clases de algoritmos

4.3.2.1. K-Means

K-means es un algoritmo basado en la partición que divide los datos en k clusters diferentes. Inicializando k centroides diferentes, el algoritmo asigna cada punto de datos al cluster más cercano. A continuación, el algoritmo vuelve a calcular los centroides de cada clúster y repite el proceso hasta que los centroides no cambian. El algoritmo es rápido y puede utilizarse para encontrar patrones en los datos. es posible ver un ejemplo de K-Means con k=2 y k=3.

4.3.2.2. DBSCAN

DBSCAN es un algoritmo basado en la densidad que encuentra clusters en los datos que están muy juntos y marca los valores atípicos que no están cerca de ninguna región. El algoritmo es sencillo y rápido y puede utilizarse para encontrar patrones en los datos.

4.3.3. Métricas para el aprendizaje no supervisado

Después de utilizar los algoritmos de clustering, es esencial evaluar el rendimiento del algoritmo. Varios aspectos son cruciales a considerar cuando se evalúa el rendimiento del clustering. Estos aspectos son:

- Las tendencias de clustering en los datos
- El número correcto de conglomerados
- La calidad de los clusters sin información externa
- La comparación de los resultados con información externa

Por lo tanto, existen dos tipos de validaciones para los algoritmos de clustering, la validación interna y la validación externa.

La validación interna es la evaluación del rendimiento del algoritmo sin información externa, utilizando únicamente la información proporcionada por los datos de entrada. Existen dos tipos de métricas: medidas de cohesión y de separación. La cohesión evalúa la proximidad entre los

elementos de un mismo cluster, mientras que las medidas de separación cuantifican el nivel de separación entre clusters.

Algunos ejemplos de métricas que miden la separación y la cohesión al mismo tiempo:

- El coeficiente Calisnki-Harabasz: también conocido como criterio de relación de varianza, se basa en la dispersión interna y la dispersión entre clusters.
- La puntuación Xie-Beni: esta métrica se diseñó para la agrupación difusa. Sin embargo, puede aplicarse al clustering duro. Es un cociente que divide el nivel de compactación de los datos dentro de un mismo cluster y la separación de los datos de clusters diferentes.
- El índice Ball-Hall: El índice de Ball-Hall: se basa en las distancias cuadráticas entre los puntos del cluster y el centroide del cluster.
- Coeficiente de silueta: es la medida más común que combina cohesión y separación.

Esta medida se define en el intervalo [-1, 1] para cada punto de datos. En caso de valor positivo, se experimenta una gran separación entre clusters. En cambio, si es negativo, los clusters están mezclados. Por último, si es cero, indica que el conjunto de datos está uniformemente distribuido en el espacio euclídeo. La validación externa es la evaluación del rendimiento del algoritmo con información externa. Los métodos de validación externa se dividen en 3 grandes grupos, que son: conjuntos coincidentes, correlación entre pares y teoría de la información. Los métodos de conjuntos coincidentes comparan los clusters detectados con la correspondencia natural. Algunos ejemplos de métodos de conjuntos coincidentes son:

- Precisión: mide los verdaderos positivos, es decir, el número de puntos de datos clasificados adecuadamente dentro del mismo cluster.
- Recall: mide el porcentaje de elementos que se incluyen adecuadamente en el mismo clúster.
- F1 Score: combinación de precisión y recuperación.
- Pureza: mide si cada clúster contiene sólo muestras de la misma clase.

La correlación entre pares mide la similitud entre dos particiones en condiciones similares, como un proceso de agrupación para el mismo conjunto. Se supone que los ejemplos de un mismo cluster deben pertenecer a la misma clase y viceversa. Algunos ejemplos de métodos de correlación entre iguales son:

- Coeficiente de Jaccard: evalúa la similitud de un cluster detectado con una partición proporcionada.

- Coeficiente de Rand: es el equivalente a la precisión en un enfoque de aprendizaje supervisado.

- Coeficiente de Folkes y Mallows: calcula la similitud entre los clusters encontrados por el algoritmo con respecto a los marcadores independientes.

Por último, los métodos de teoría de la información se basan en conceptos de la Teoría de la Información, como la incertidumbre actual en la predicción de las clases naturales proporcionadas por otras particiones. Esta familia incluye medidas básicas como la entropía y la información mutua y sus respectivas variantes normalizadas. Algunos ejemplos de métodos de la Teoría de la Información son

- Entropía: una medida recíproca de pureza que mide el grado de desorden en la agrupación.

- Información mutua: una métrica que mide la reducción de la incertidumbre sobre los resultados de la agrupación los resultados de la agrupación a partir de un conocimiento previo.

4.4. Aprendizaje semisupervisado

El aprendizaje semisupervisado es una rama del aprendizaje automático que combina técnicas de aprendizaje supervisado y no supervisado. Esta metodología utiliza tanto datos etiquetados como no etiquetados para entrenar el modelo. Normalmente, se utiliza para ayudar a mejorar el rendimiento en uno de esos tipos de tareas de aprendizaje mediante el uso de métodos que pertenecen a la otra tarea de aprendizaje. La mayor parte del uso del aprendizaje semisupervisado se centra en la clasificación. Para situaciones con datos etiquetados limitados, los enfoques de clasificación semisupervisada son útiles porque las técnicas de aprendizaje supervisado son insuficientes para resolver el problema y no son fiables. Esto es posible porque los datos etiquetados pueden ser caros o difíciles de recopilar. Si los datos no etiquetados son suficientes en esas circunstancias, pueden mejorar el rendimiento del modelo. El aprendizaje semisupervisado se basa en tres supuestos principales que son la base de la mayoría de los algoritmos de aprendizaje semisupervisado.

Estos supuestos son:

- Suposición de suavidad: Según este supuesto, si dos puntos de entrada en el espacio de entrada están próximos entre sí, lo más probable es que pertenezcan a la misma clase. Este supuesto también es común en el aprendizaje supervisado, pero la ventaja del aprendizaje semisupervisado es que puede manejar datos no etiquetados.

- Suposición de baja densidad: Según este supuesto, la frontera de decisión debe atravesar zonas de baja densidad en lugar de zonas de alta densidad. La hipótesis de suavidad está estrechamente relacionada con esta hipótesis.

- Hipótesis múltiple: La hipótesis de los colectores afirma que el espacio de entrada está formado por colectores de dimensiones inferiores en los que se encuentran todos los puntos de datos y que los puntos de datos de un mismo colector pertenecen a la misma clase.

4.4.1. Métodos inductivos

os métodos inductivos son aquellos que buscan construir un modelo que pueda hacer predicciones para cualquier tipo de datos disponibles en el espacio de entrada. Estos métodos son una ampliación de los métodos de aprendizaje supervisado que incluyen datos tanto etiquetados como no etiquetados.

4.4.2. Métodos envolventes

Los métodos envolventes utilizan técnicas de aprendizaje supervisado para entrenar clasificadores en datos previamente etiquetados, y luego producen más datos etiquetados utilizando las predicciones del modelo. Una vez que se han generado estos datos etiquetados adicionales, se vuelve a entrenar al clasificador utilizando tanto los datos previamente etiquetados como los nuevos datos etiquetados. De esta manera, el enfoque envolvente convierte los datos no etiquetados en datos etiquetados, y luego se utilizan para construir el modelo final mediante un algoritmo de aprendizaje puramente supervisado.

Los enfoques envolventes incluyen los siguientes, que son los más conocidos:

- El autoentrenamiento es el método de pseudoetiquetado más sencillo. Estos métodos utilizan un clasificador supervisado para entrenar iterativamente sobre datos etiquetados y pseudo-etiquetados hasta que no se clasifican más datos.

- Co-entrenamiento: una extensión del autoentrenamiento en la que se utilizan dos o más clasificadores supervisados en lugar de uno solo. Para que esto funcione, los clasificadores supervisados no deben estar demasiado conectados en sus predicciones. En este caso, la cantidad de datos nuevos generados será limitada.

- Refuerzo: Un conjunto de clasificadores se construye creando sucesivamente clasificadores individuales, de forma similar a los algoritmos de refuerzo estándar.

4.4.3. Preprocesamiento no supervisado

El preprocesamiento no supervisado utiliza los datos etiquetados y no etiquetados en dos etapas diferentes, a diferencia de los métodos envolventes. Normalmente, los datos no etiquetados se procesan primero en la etapa no supervisada para extraer características, agrupar los datos para etiquetarlos después o preentrenar un aprendiz basado e inicializarlo con los pesos adecuados. El siguiente proceso es la aplicación de un algoritmo de aprendizaje supervisado.

Las técnicas utilizadas en esta etapa son:

- Extracción de características: Se trata de una técnica ventajosa que ha desempeñado un papel esencial en la construcción de clasificadores. Esta técnica encuentra transformaciones en los datos de entrada que pueden mejorar el rendimiento o la eficiencia del clasificador.

- Cluster-then-label: Como su nombre indica, este enfoque une los procesos de agrupación y clasificación. En primer lugar, se agrupan todos los datos disponibles y, a continuación, los grupos resultantes se utilizan para guiar el proceso de clasificación.

- Pre-entrenamiento: El preentrenamiento es una técnica muy utilizada actualmente en el aprendizaje profundo. Los datos no etiquetados sugieren el límite hacia regiones potencialmente interesantes antes de aplicar el algoritmo supervisado.

4.4.4. Métodos intrínsecamente semisupervisados

Los métodos intrínsecamente semisupervisados son aquellos que pueden funcionar tanto con datos etiquetados (supervisados) como con datos sin etiquetar (no supervisados). Estos métodos son capaces de aprender de los datos sin necesidad de una etiqueta o clase predefinida para cada ejemplo de datos. Por lo general, estos métodos se basan en uno de los supuestos de aprendizaje semisupervisado comentados anteriormente.

Los métodos utilizados en esta etapa son:

- Máximo-margen: Es un enfoque que busca maximizar la distancia entre dos clases de datos al mismo tiempo que se minimiza el error de clasificación. Es el método más directo e intenta maximizar la distancia entre los puntos de datos de entrada y la frontera de decisión. Esta técnica corresponde a la hipótesis de baja densidad semisupervisada.

- Métodos basados en la perturbación: consiste en añadir ruido o perturbaciones a los datos no etiquetados para generar múltiples conjuntos de datos modificados que se pueden utilizar para entrenar un modelo. Esta técnica se basa en el supuesto de suavidad. El modelo de predicción debe ser robusto a las perturbaciones locales en el espacio de entrada,

lo que significa que cuando un punto de datos es perturbado por una pequeña cantidad de ruido, la predicción para los datos ruidosos y los datos limpios debe ser similar. Estos métodos suelen aplicarse con redes neuronales.

- Métodos mainfold: es un enfoque que busca preservar la estructura de los datos al proyectarlos a un espacio de menor dimensión. Como se ha visto antes con los métodos basados en perturbaciones, añadir pequeñas perturbaciones al espacio de entrada funciona bien bajo el supuesto de suavidad. Sin embargo, en algunos conjuntos de datos de baja dimensión, las perturbaciones pueden diferir de los datos de entrada. Por este motivo, aquí se utiliza el supuesto de la multiplicidad.

- Modelos generativos: son modelos que intentan generar nuevos ejemplos de datos a partir de una distribución de datos conocida. Todos los métodos anteriores son discriminativos. Su único objetivo es suponer una función que pueda clasificar los puntos de datos. En cambio, el objetivo de los modelos generativos es modelar la distribución p(x, y), de la que se pueden extraer muestras(x, y).

4.4.5. Métodos transductivos

A diferencia de los métodos inductivos, que producen un modelo que puede clasificar cualquier punto de datos en el espacio de entrada, las técnicas transductivas no tienen en cuenta la distinción entre el conjunto de entrenamiento y el de prueba. Se proporcionan tanto datos etiquetados como no etiquetados como entrada, y la salida consiste exclusivamente en predicciones para los datos no etiquetados.

Estos métodos a menudo se definen mediante un grafo que conecta todos los puntos de datos y codifica las relaciones entre ellos a través de aristas que pueden estar ponderadas. Se define y optimiza una función objetivo que asegura que las etiquetas predichas coincidan con las etiquetas reales de los datos etiquetados, y que los puntos de datos similares, según el grafo de similitud, tengan las mismas predicciones de etiquetas. Hay una cierta similitud entre los métodos transductivos e inductivos, ya que ambos construyen un grafo sobre los puntos de datos y utilizan relaciones entre pares para aproximar estructuras más complejas.

La principal diferencia es que los métodos inductivos tratan de crear un modelo capaz de clasificar cualquier punto de datos en el espacio de entrada, mientras que los métodos transductivos solo dan predicciones para un conjunto de datos específico sin etiquetar.

Capítulo 5

Software y Hardware

5.1. Requerimientos técnicos

En este trabajo final de máster, se han utilizado una serie de herramientas de software y hardware para llevar a cabo el análisis y la visualización de los datos. En lo que respecta al software, se han utilizado lenguajes de programación como Python y R, así como librerías y paquetes especializados en el análisis de datos y la ciencia de datos, como Pandas, NumPy y Matplotlib. Además, se han utilizado diversas herramientas de visualización de datos, como Plotly, para crear gráficos y visualizaciones atractivas y fáciles de interpretar. En cuanto al hardware, se ha utilizado una computadora equipada con un procesador de alta potencia y suficiente memoria RAM para manejar el procesamiento de grandes cantidades de datos. También se han utilizado discos duros externos para almacenar y respaldar los datos utilizados en el análisis.

Las herramientas tanto de software como de hardware que utilizaré en este proyecto son las siguientes:

- Sistema Operativo: Linux. Versión Ubuntu 22.04.
- IDE: Visual Studio Code.
- Lenguaje de programación: Python, Html, CSS.
- Control de versiones: Github.
- Alojamiento: AWS.
- Navegador web: Brave.
- Procesador de textos Latex: Overleaf.

Detallamos.

5.2. Sistema Operativo: Linux

Linux es un sistema operativo basado en el kernel del sistema operativo Unix. Es un sistema operativo de código abierto, lo que significa que el código fuente está disponible para cualquier persona para examinar, modificar y mejorar. Linux es conocido por ser un sistema operativo seguro, estable y altamente configurable.

Las razones de elegir Linux para el análisis de datos y la ciencia de datos son:

- Libre: Linux es un sistema operativo libre y de código abierto, .
- Estabilidad: Linux es conocido por ser un sistema operativo estable y fiable.
- Herramientas de análisis de datos: Linux viene con una gran cantidad de herramientas de análisis de datos y ciencia de datos preinstaladas, como R, Python y Julia.
- Comunidad: Linux tiene una gran comunidad de usuarios y desarrolladores que contribuyen al sistema operativo y crean paquetes de software adicionales.

5.2.1. Ubuntu

Ubuntu es una distribución de Linux, lo que significa que es una versión específica del sistema operativo Linux que viene con un conjunto particular de programas y herramientas. Ubuntu fue creado con el objetivo de ser fácil de usar y está diseñado para ser intuitivo para usuarios que no son expertos en informática.

5.3. IDE

Un IDE (Integrated Development Environment, o Entorno de Desarrollo Integrado en español) es un software que proporciona a los desarrolladores un entorno de trabajo completo para crear aplicaciones de software. Un IDE normalmente incluye un editor de código, un depurador y herramientas de construcción. Son muy útiles para los desarrolladores ya que facilitan y agilizan el proceso de creación de software. Al tener todas las herramientas necesarias en un mismo entorno, los desarrolladores pueden trabajar de manera más eficiente y tener un mayor control sobre el código que están creando. Características:

- Sintaxis de destacado: resaltado de diferentes elementos del código para hacerlo más legible.
- Autocompletado: sugerencia de código mientras se escribe.

- Depuración: herramientas para encontrar y solucionar errores en el código.
- Consola: una consola integrada para ejecutar código y ver los resultados.
- Control de versiones: herramientas para trabajar con repositorios de código y controlar los cambios realizados.

5.3.1. VSCode

Visual Studio Code (también conocido como Visual Studio Code o simplemente VS Code) es un editor de código fuente y un entorno de desarrollo integrado (IDE) desarrollado por Microsoft para Windows, Linux y macOS. Es gratuito y de código abierto, y está diseñado para ser una herramienta ligera y fácil de usar para desarrolladores de todos los niveles.

VS Code incluye una amplia gama de características para facilitar el desarrollo de software, como:

- Sintaxis de destacado: resaltado de diferentes elementos del código para hacerlo más legible.
- Autocompletado: sugerencia de código mientras se escribe.
- Depuración: herramientas para encontrar y solucionar errores en el código.
- Integración con control de versiones: soporte para Git y otras herramientas de control de versiones.
- Extensiones: miles de extensiones disponibles para añadir funcionalidades adicionales al editor.

5.4. Lenguajes de Programación

Un lenguaje de programación es un lenguaje que se utiliza para escribir y ejecutar programas de computadora. Un programa de computadora es un conjunto de instrucciones que le dice a la computadora qué hacer. Los lenguajes de programación se utilizan para crear todo tipo de aplicaciones, desde juegos hasta programas de análisis de datos.

5.4.1. PYTHON, HTML, CSS

Los lenguajes utilizados son:

- Python es un lenguaje de programación de alto nivel y orientado a objetos que se utiliza ampliamente en el análisis de datos, la ciencia de datos, el aprendizaje automático y el desarrollo de aplicaciones web. Es conocido por su sintaxis clara y fácil de leer, lo que lo hace ideal para principiantes y programadores experimentados.
- HTML (HyperText Markup Language) es un lenguaje de marcado utilizado para estructurar y dar formato a documentos web. Es el lenguaje principal utilizado para crear páginas web y se utiliza junto con CSS (Cascading Style Sheets) para dar estilo y diseño a las páginas.
- CSS es un lenguaje de diseño utilizado para dar estilo y diseño a las páginas web. Se utiliza junto con HTML para definir cómo se deben mostrar los elementos en una página web, como los colores, fuentes, márgenes y posicionamiento.

5.5. Control de versiones

Un control de versiones es un sistema que registra los cambios realizados en un archivo o conjunto de archivos a lo largo del tiempo, de manera que pueda recuperar versiones específicas más adelante. Los sistemas de control de versiones son muy útiles para el desarrollo de software, ya que permiten a los desarrolladores revertir fácilmente los cambios si algo sale mal o comparar versiones del código para ver qué ha cambiado. Algunos ejemplos populares incluyen Git, Subversion y Mercurial.

5.5.1. Github

GitHub es una plataforma en línea que ofrece alojamiento de repositorios de código y herramientas para el control de versiones. Se basa en el sistema de control de versiones Git, que permite a los desarrolladores controlar y gestionar los cambios en el código de un proyecto de software.

5.6. Cloud

Un sistema cloud para ciencia de datos es un servicio en línea que proporciona a los usuarios acceso a una plataforma de análisis de datos en la nube. Los usuarios pueden subir sus datos al sistema cloud, procesarlos y analizarlos utilizando diversas herramientas y lenguajes de programación, y luego guardar los resultados en la nube.

Ejemplos de sistemas cloud para ciencia de datos incluyen AWS (Amazon Web Services), Azure (de Microsoft) y Google Cloud Platform.

5.6.1. AWS

Amazon Web Services (AWS) es una plataforma en la nube que ofrece una amplia gama de servicios para el desarrollo de aplicaciones, incluyendo servidores virtuales. Una de las opciones disponibles para el despliegue de este proyecto Django en AWS es utilizar EC2 (Amazon Elastic Compute Cloud), que es un servicio de Amazon que proporciona servidores virtuales a través de la nube.

5.7. Navegador web

Un navegador web es una aplicación que se utiliza para acceder a Internet y ver páginas web.

5.7.1. Brave

Brave es el navegado rutilizado y enfocado en la privacidad y la seguridad de los usuarios y bloquea automáticamente los anuncios y el seguimiento de terceros. Brave también tiene una función de privacidad integrada que permite a los usuarios navegar de manera anónima y proteger su privacidad en línea. Está basado en Chromium y es compatible con todas las páginas web estándar y admite extensiones de Chrome y Firefox.

5.8. Latex

LaTeX es un lenguaje de marcado de documentos utilizado principalmente en el campo de la ciencia y la tecnología. Se utiliza para crear documentos profesionales de alta calidad, como artículos científicos, libros y presentaciones. Es utilizado ampliamente en la comunidad científica y académica debido a su soporte para la notación matemática y formato de alta calidad.

Algunas de las ventajas de LaTeX incluyen:

- Produce documentos de alta calidad: LaTeX utiliza una tipografía de alta calidad y permite la creación de documentos con un diseño profesional.
- Facilita la colaboración: Los archivos LaTeX son fáciles de compartir y colaborar en equipo.
- Permite la creación de documentos largos y complejos: LaTeX es muy adecuado para la creación de documentos largos y complejos, como tesis y artículos científicos.

- Formatea automáticamente los elementos del documento: LaTeX se encarga del formateo automático de cosas como tablas, índices y bibliografías, lo que ahorra tiempo y esfuerzo al usuario.
- Mejora la legibilidad del código: LaTeX utiliza un lenguaje de marcado fácil de leer y entender, lo que hace que sea más fácil de mantener y modificar el código a lo largo del tiempo.

Algunos de los inconvenientes de LaTeX son:

- Aprendizaje curva: LaTeX tiene una sintaxis especial y puede llevar un tiempo aprender a utilizarlo de manera efectiva.
- No es tan intuitivo como otros procesadores de texto: Algunas personas pueden encontrar LaTeX menos intuitivo que otros procesadores de texto como Microsoft Word.
- No es tan versátil: LaTeX está diseñado especialmente para la creación de documentos científicos y técnicos y puede ser menos adecuado para otros tipos de documentos.
- Puede ser más lento que otros procesadores de texto: Generar el documento final a partir del archivo LaTeX puede tomar más tiempo que simplemente escribir el documento en un procesador de texto como Word.

5.8.1. Overleaf

Overleaf es una plataforma en línea que permite a los usuarios crear y colaborar en documentos LaTeX en tiempo real. La plataforma ofrece un editor LaTeX en línea y una interfaz de usuario amigable que hace que sea fácil de usar para personas sin experiencia previa en LaTeX.

5.9. Hardware

Para realizar el trabajo final (TFM), se utilizaron diversas herramientas de hardware para almacenar y procesar los datos. Esto incluyó servidores en la nube, computadoras de escritorio y dispositivos móviles. Los servidores en la nube se utilizaron para almacenar y procesar grandes cantidades de datos, mientras que las computadoras de escritorio y dispositivos móviles se utilizaron para acceder a los datos y realizar análisis y visualizaciones. Además, se utilizaron diversas herramientas de hardware para la captura y el procesamiento de datos, como sensores y dispositivos de medición. Estas herramientas de hardware fueron esenciales para poder realizar el TFM de manera eficiente y obtener resultados precisos y confiables.

Cuadro 5.1: Especificaciones Hardware I/II

Clase	**Descripción**
system	ROG Strix G713IC_G713IC
bus	G713IC
memory	64KiB BIOS
memory	16GiB Memoria de sistema
memory	8GiB SODIMM DDR4 Síncrono Unbuffered 3200 MHz (0,3 ns)
memory	8GiB SODIMM DDR4 Síncrono Unbuffered 3200 MHz (0,3 ns)
memory	512KiB L1 caché
memory	4MiB L2 caché
memory	8MiB L3 caché
processor	AMD Ryzen 7 4800H with Radeon Graphics
bridge	Renoir/Cezanne Root Complex
generic	Renoir/Cezanne IOMMU
bridge	Renoir PCIe GPP Bridge
display	GA107M [GeForce RTX 3050 Mobile]
multimedia	NVIDIA Corporation
bridge	Renoir/Cezanne PCIe GPP Bridge
network	RTL8111/8168/8411 PCI Express Gigabit Ethernet Controller
bridge	Renoir/Cezanne PCIe GPP Bridge
network	MT7921 802.11ax PCI Express Wireless Network Adapter
bridge	Renoir/Cezanne PCIe GPP Bridge
storage	SAMSUNG MZVLQ1T0HBLB-00B00
disk	NVMe disk
disk	NVMe disk
disk	1024GB NVMe disk
volume	99MiB Windows FAT volumen
volume	15MiB reserved partition
volume	482GiB Windows NTFS volumen
volume	616MiB Windows NTFS volumen
volume	470GiB partición EXT4

Cuadro 5.2: Especificaciones Hardware II/II

Clase	**Descripción**
bridge	Renoir Internal PCIe GPP Bridge to Bus
display	Renoir
multimedia	Renoir Radeon High Definition Audio Controller
input	HD-Audio Generic HDMI/DP,pcm=3
generic	Family 17h (Models 10h-1fh) Platform Security Processor
bus	Renoir/Cezanne USB 3.1
bus	xHCI Host Controller
input	Asus Keyboard
bus	xHCI Host Controller
bus	Renoir/Cezanne USB 3.1
bus	xHCI Host Controller
input	USB OPTICAL MOUSE Keyboard
communication	Wireless_Device
bus	xHCI Host Controller
multimedia	Raven/Raven2/FireFlight/Renoir Audio Processor
multimedia	Family 17h (Models 10h-1fh) HD Audio Controller
input	HD-Audio Generic Headphone
bus	FCH SMBus Controller
bridge	FCH LPC Bridge
system	PnP device PNP0c01
system	PnP device PNP0b00
generic	PnP device ATK3001
system	PnP device PNP0c02
input	Lid Switch
input	Power Button
input	Asus Wireless Radio Control
input	Asus WMI hotkeys
input	Sleep Button
input	ASUE120A:00 04F3:319B Mouse
input	ASUE120A:00 04F3:319B Touchpad
input	Video Bus

Capítulo 6

Dataset

El conjunto de datos fue creado por el Laboratorio Nacional de Energías Renovables (NREL) como parte de un proyecto de investigación sobre sistemas de bombeo de agua.

El National Renewable Energy Laboratory (NREL) es un laboratorio de investigación financiado por el Departamento de Energía de Estados Unidos. El NREL se creó en 1974 con el nombre de Instituto de Investigación de la Energía Solar, y pasó a denominarse Laboratorio Nacional de Energías Renovables en 1977. Está situado en Golden, Colorado, y se centra en la investigación y el desarrollo de energías renovables y tecnologías de eficiencia energética. Como parte de sus actividades de investigación, el NREL ha realizado varios estudios sobre sistemas de bombeo de agua, incluida la recogida de datos de 52 sensores en un sistema de bombeo de agua.

El NREL es:

- Un laboratorio de investigación líder en el campo de las energías renovables y la eficiencia energética, y se dedica a impulsar el uso de tecnologías energéticas limpias y sostenibles. Investiga una amplia gama de tecnologías de energías renovables, como la solar, la eólica, la geotérmica y la bioenergía.
- Cuenta con varias instalaciones de investigación, así como varios bancos de pruebas y proyectos piloto.
- Colabora estrechamente con la industria, el mundo académico y otras organizaciones de investigación para impulsar el desarrollo y la implantación de tecnologías de energías renovables.

Estos datos se recogieron como parte de un proyecto de investigación sobre sistemas de bombeo de agua, y el objetivo era recopilar datos que pudieran utilizarse para optimizar el diseño y el funcionamiento de los sistemas de bombeo de agua. Los datos recopilados por

el NREL se han puesto a disposición a través del conjunto de datos de Kaggle denominado "Dataset from Water Pump with 52 sensors". Este conjunto de datos puede ser utilizado por investigadores e ingenieros que trabajen en el desarrollo y la optimización de sistemas de bombas de agua, así como para el mantenimiento predictivo, el diagnóstico de fallos y la optimización de sistemas de bombas de agua.

Este conjunto de datos puede utilizarse para diversos fines, como el desarrollo y la optimización de sistemas de bombas de agua, el mantenimiento predictivo, el diagnóstico de fallos y la optimización de sistemas de bombas de agua.

Los datos se recogieron de un sistema de bombeo de agua situado en un entorno de laboratorio, y los sensores se colocaron en varias partes del sistema, como la bomba, el motor y el panel de control.

El conjunto de datos incluye datos sin procesar y datos procesados, con los datos procesados transformados y limpiados para facilitar el análisis. Se proporciona en formato CSV.

Capítulo 7

Proyecto Principal

En esta sección, se describirá detalladamente el proceso de desarrollo del proyecto. En primer lugar, se realizará un análisis exhaustivo del conjunto de datos que se ha obtenido para el proyecto. Durante este análisis, se examinarán las diferentes características de los datos y se discutirá su relevancia y importancia para el desarrollo del modelo que se está construyendo.

Una vez que hayamos superado este obstáculo y hayamos desarrollado una forma eficiente de clasificar los datos, se presentarán los resultados de la clasificación obtenidos. Se detallará cómo se han dividido los datos en diferentes categorías y se analizará la precisión y eficacia del modelo utilizado. Además, se discutirán las posibles mejoras y optimizaciones que se podrían realizar en el futuro para aumentar la precisión de la clasificación.

Comenzamos:

7.1. Contexto

Los datos se refieren a una bomba de agua de una zona pequeña que sufrió 7 fallos en el sistema durante el año 2018. Al analizar los datos, el equipo no ha podido identificar ningún patrón en los momentos en que el sistema falló, por lo que no saben con certeza dónde deben enfocar sus esfuerzos para evitar futuros problemas. Un pequeño equipo se encargo de las bombas de agua en una pequeña zona alejada de una gran ciudad. El año anterior, el sistema de bombeo falló 7 veces. Estos fallos causan enormes problemas a muchas personas, empresas y también, en algunos casos, lesiones graves.

El objetivo es predecir futuros fallos de las bombas de agua utilizando los datos históricos de los sensores, lo que permitirá al equipo reaccionar de forma proactiva y planificar la solución de mantenimiento más adecuada.

Detallamos los sensores:

Cuadro 7.1: Sensores I

SENSOR_00 – Motor Casing Vibration
SENSOR_01 – Motor Frequency A
SENSOR_02 – Motor Frequency B
SENSOR_03 – Motor Frequency C
SENSOR_04 – Motor Speed
SENSOR_05 – Motor Current
SENSOR_06 – Motor Active Power
SENSOR_07 – Motor Apparent Power
SENSOR_08 – Motor Reactive Power
SENSOR_09 – Motor Shaft Power
SENSOR_10 – Motor Phase Current A
SENSOR_11 – Motor Phase Current B
SENSOR_12 – Motor Phase Current C
SENSOR_13 – Motor Coupling Vibration
SENSOR_14 – Motor Phase Voltage AB
SENSOR_16 – Motor Phase Voltage BC
SENSOR_17 – Motor Phase Voltage CA
SENSOR_18 – Pump Casing Vibration
SENSOR_19 – Pump Stage 1 Impeller Speed
SENSOR_20 – Pump Stage 1 Impeller Speed
SENSOR_21 – Pump Stage 1 Impeller Speed
SENSOR_22 – Pump Stage 1 Impeller Speed
SENSOR_23 – Pump Stage 1 Impeller Speed
SENSOR_24 – Pump Stage 1 Impeller Speed
SENSOR_25 – Pump Stage 2 Impeller Speed
SENSOR_26 – Pump Stage 2 Impeller Speed
SENSOR_27 – Pump Stage 2 Impeller Speed
SENSOR_28 – Pump Stage 2 Impeller Speed
SENSOR_29 – Pump Stage 2 Impeller Speed
SENSOR_30 – Pump Stage 2 Impeller Speed
SENSOR_31 – Pump Stage 2 Impeller Speed
SENSOR_32 – Pump Stage 2 Impeller Speed
SENSOR_33 – Pump Stage 2 Impeller Speed
SENSOR_34 – Pump Inlet Flow
SENSOR_35 – Pump Discharge Flow
SENSOR_36 – Pump UNKNOWN
SENSOR_37 – Pump Lube Oil Overhead Reservoir Level
SENSOR_38 – Pump Lube Oil Return Temp
SENSOR_39 – Pump Lube Oil Supply Temp
SENSOR_40 – Pump Thrust Bearing Active Temp
SENSOR_41 – Motor Non Drive End Radial Bearing Temp 1

Cuadro 7.2: Sensores II

SENSOR_42 – Motor Non Drive End Radial Bearing Temp 2
SENSOR_43 – Pump Thrust Bearing Inactive Temp
SENSOR_44 – Pump Drive End Radial Bearing Temp 1
SENSOR_45 – Pump non Drive End Radial Bearing Temp 1
SENSOR_46 – Pump Non Drive End Radial Bearing Temp 2
SENSOR_47 – Pump Drive End Radial Bearing Temp 2
SENSOR_48 – Pump Inlet Pressure
SENSOR_49 – Pump Temp Unknown
SENSOR_50 – Pump Discharge Pressure 1
SENSOR_51 – Pump Discharge Pressure 2
Pump Status

7.2. Formulación de ML

El problema de las bombas de agua es un problema de clasificación en el que se utiliza el aprendizaje automático para predecir si una bomba de agua está funcionando correctamente o no en base a los datos proporcionados por 52 sensores. Los sensores proporcionan datos sobre diversos aspectos del rendimiento de la bomba, como la presión del agua, la temperatura y el caudal. El objetivo del aprendizaje automático es analizar estos datos y predecir si la bomba está funcionando correctamente o no.

7.3. Métricas de rendimiento

En este estudio de caso, es importante que el modelo de aprendizaje automático sea capaz de clasificar adecuadamente tanto las bombas de agua que están funcionando correctamente como las que no lo están. Si el modelo predice incorrectamente la clase de una bomba, esto puede tener consecuencias graves, como averías, costes de reparación y problemas de seguridad. Por ejemplo, si el modelo predice que una bomba que en realidad está funcionando correctamente no lo está, se podría llevar a cabo un mantenimiento innecesario en la bomba, lo que podría ocasionar costes adicionales y pérdida de tiempo. Por otro lado, si el modelo predice que una bomba que en realidad está fallando está funcionando correctamente, esto podría tener consecuencias aún más graves, como averías y problemas de seguridad.

- Para evaluar el rendimiento del modelo en este problema de clasificación, se utilizan dos métricas: la puntuación F1-Score y la matriz de confusión. La macro F1-Score se calcula tomando la F1-Score de cada clase (es decir, la precisión y el recall de cada clase) y calculando su media. Esta métrica nos permite dar la misma importancia a ambas clases y evaluar el rendimiento del modelo en términos de precisión y recall promedio.

- La matriz de confusión, por otro lado, nos permite visualizar el rendimiento del modelo y ver cuántos puntos ha clasificado correctamente y cuántos ha clasificado incorrectamente para cada clase. Esta información es útil para identificar las fortalezas y debilidades del modelo y para tomar medidas para mejorar su rendimiento.

7.4. Resumen de datos

Los datos del sensor de la bomba se recogen de kaggle:

https://www.kaggle.com/datasets/nphantawee/pump-sensor-data

Los datos contienen lecturas de 52 sensores y su correspondiente marca de tiempo. Se proporcionan los datos de los sensores de 5 meses, de abril a agosto. Cada fila contiene lecturas de sensores después de un minuto y no se especifica qué parámetro mide cada sensor. El conjunto de datos contiene un total de 220320 puntos de datos.

Vemos las primeras y últimas filas del dataset:

```
                timestamp  sensor_01  sensor_02  sensor_03   sensor_04  \
0       2018-04-01 00:00:00   47.09201  53.211800  46.310760  634.375000
1       2018-04-01 00:01:00   47.09201  53.211800  46.310760  634.375000
2       2018-04-01 00:02:00   47.35243  53.211800  46.397570  638.888900
3       2018-04-01 00:03:00   47.09201  53.168400  46.397568  628.125000
4       2018-04-01 00:04:00   47.13541  53.211800  46.397568  636.458300
...                     ...        ...        ...        ...         ...
220315  2018-08-31 23:55:00   47.69965  50.520830  43.142361  634.722229
220316  2018-08-31 23:56:00   47.69965  50.564240  43.142361  630.902771
220317  2018-08-31 23:57:00   47.69965  50.520830  43.142361  625.925903
220318  2018-08-31 23:58:00   47.69965  50.520832  43.142361  635.648100
220319  2018-08-31 23:59:00   47.69965  50.520832  43.142361  639.814800

        sensor_05  sensor_06  sensor_07  sensor_08  sensor_09  ...  sensor_41  \
0        76.45975   13.41146   16.13136   15.56713   15.05353  ...  30.989580
1        76.45975   13.41146   16.13136   15.56713   15.05353  ...  30.989580
2        73.54598   13.32465   16.03733   15.61777   15.01013  ...  30.468750
3        76.98898   13.31742   16.24711   15.69734   15.08247  ...  30.468750
4        76.58897   13.35359   16.21094   15.69734   15.08247  ...  30.989580
...           ...        ...        ...        ...        ...  ...        ...
```

```
220315   64.59095   15.11863   16.65220   15.65393   15.16204  ...  30.468750
220316   65.83363   15.15480   16.70284   15.65393   15.11863  ...  30.208332
220317   67.29445   15.08970   16.70284   15.69734   15.11863  ...  29.947920
220318   65.09175   15.11863   16.56539   15.74074   15.11863  ...  29.947916
220319   65.45634   15.11863   16.65220   15.65393   15.01013  ...  29.947916

        sensor_42  sensor_43  sensor_44  sensor_45  sensor_46  sensor_47  \
0       31.770832   41.92708  39.641200   65.68287   50.92593  38.194440
1       31.770832   41.92708  39.641200   65.68287   50.92593  38.194440
2       31.770830   41.66666  39.351852   65.39352   51.21528  38.194443
3       31.510420   40.88541  39.062500   64.81481   51.21528  38.194440
4       31.510420   41.40625  38.773150   65.10416   51.79398  38.773150
...           ...        ...        ...        ...        ...        ...
220315  30.208330   38.28125  68.287030   52.37268   48.32176  41.087960
220316  29.947920   38.28125  66.840280   50.63657   48.03241  40.798610
220317  30.208330   39.06250  65.393520   48.90046   48.03241  40.798610
220318  30.208332   40.62500  64.236110   47.74306   48.32176  40.509258
220319  30.208332   41.40625  62.789350   46.29630   48.90046  40.219910

        sensor_48  sensor_49  machine_status
0        157.9861   67.70834          NORMAL
1        157.9861   67.70834          NORMAL
2        155.9606   67.12963          NORMAL
3        155.9606   66.84028          NORMAL
4        158.2755   66.55093          NORMAL
...           ...        ...             ...
220315   212.3843  153.64580          NORMAL
220316   213.8310  156.25000          NORMAL
220317   217.3032  155.38190          NORMAL
220318   222.5116  153.93520          NORMAL
220319   227.4306  150.46300          NORMAL

[220320 rows x 50 columns]
```

Descripción del dataset: Utilizamos el método describe de pandas para calcular resúmenes estadísticos para cada columna del conjunto de datos. Los resúmenes incluyen estadísticas como la media, la mediana, el mínimo, el máximo y los cuartiles. Luego, se utiliza el método T para transponer la tabla, de modo que las columnas se conviertan en filas y las filas en columnas. Finalmente, se utiliza el método reset_index para restablecer el índice de la tabla y convertirlo en una columna normal.

Cuadro 7.3: Resumen dataset

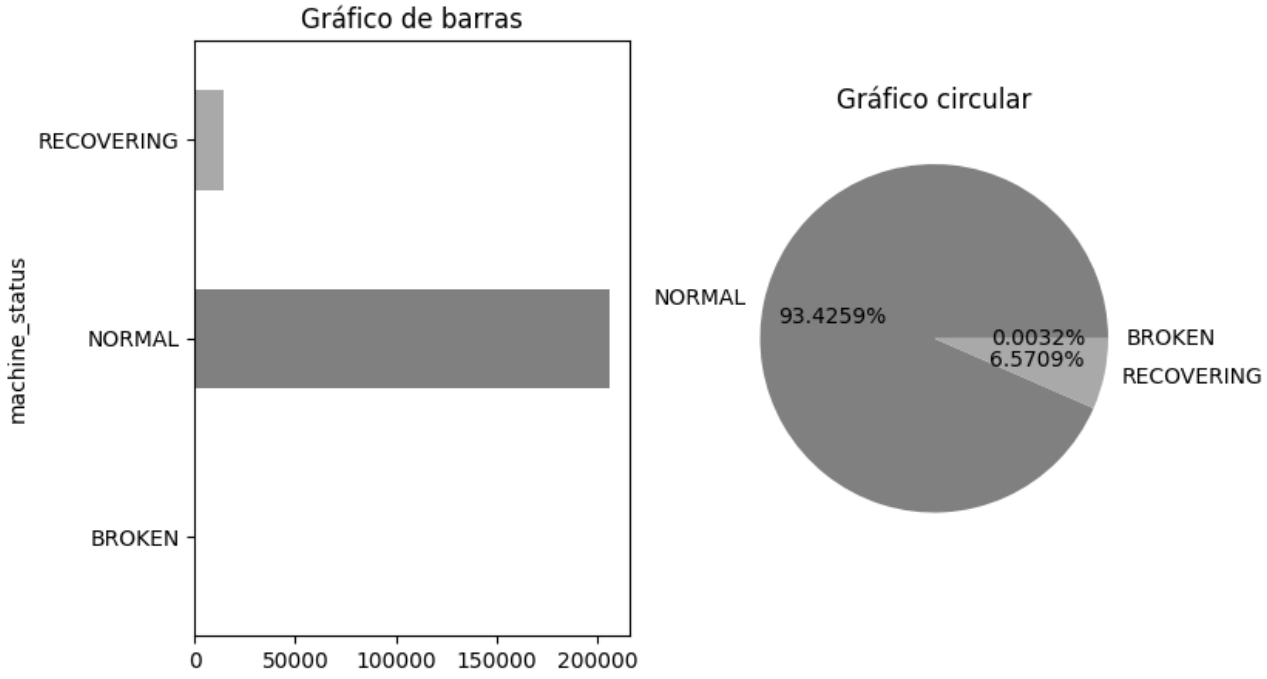
index	count	mean	std	min	0.25	0.50	0.75	max
sensor_01	219951	47.59	3.30	0.00	46.31	48.13	49.48	56.73
sensor_02	220301	50.87	3.67	33.16	50.39	51.65	52.78	56.03
sensor_03	220301	43.75	2.42	31.64	42.84	44.23	45.31	48.22
sensor_04	220301	590.67	144.02	2.80	626.62	632.64	637.62	800.00
sensor_05	220301	73.40	17.30	0.00	69.98	75.58	80.91	100.00
sensor_06	215522	13.50	2.16	0.01	13.35	13.64	14.54	22.25
sensor_07	214869	15.84	2.20	0.00	15.91	16.17	16.43	23.60
sensor_08	215213	15.20	2.04	0.03	15.18	15.49	15.70	24.35
sensor_09	215725	14.80	2.09	0.00	15.05	15.08	15.12	25.00
sensor_10	220301	41.47	12.09	0.00	40.71	44.29	47.46	76.11
sensor_11	220301	41.92	13.06	0.00	38.86	45.36	49.66	60.00
sensor_12	220301	29.14	10.11	0.00	28.69	32.52	34.94	45.00
sensor_13	220301	7.08	6.90	0.00	1.54	2.93	12.86	31.19
sensor_14	220299	376.86	113.21	32.41	418.10	420.11	421.00	500.00
sensor_16	220289	416.47	126.07	0.00	459.45	462.86	464.30	739.74
sensor_17	220274	421.13	129.16	0.00	454.14	462.02	466.86	600.00
sensor_18	220274	2.30	0.77	0.00	2.45	2.53	2.59	4.87
sensor_19	220304	590.83	199.35	0.00	662.77	665.67	667.15	878.92
sensor_20	220304	360.81	101.97	0.00	398.02	399.37	400.09	448.91
sensor_21	220304	796.23	226.68	95.53	875.46	879.70	882.13	1107.53
sensor_22	220279	459.79	154.53	0.00	478.96	531.86	534.25	594.06
sensor_23	220304	922.61	291.84	0.00	950.92	981.93	1090.81	1227.56
sensor_24	220304	556.24	182.30	0.00	601.15	625.87	628.61	1000.00
sensor_25	220284	649.14	220.87	0.00	693.96	740.20	750.36	839.58
sensor_26	220300	786.41	246.66	43.15	790.49	861.87	919.10	1214.42
sensor_27	220304	501.51	169.82	0.00	448.30	494.47	536.27	2000.00
sensor_28	220304	851.69	313.07	4.32	782.68	967.28	1043.98	1841.15
sensor_29	220248	576.20	225.76	0.64	518.95	564.87	744.02	1466.28
sensor_30	220059	614.60	195.73	0.00	627.78	668.98	697.22	1600.00
sensor_31	220304	863.32	283.54	23.96	839.06	917.71	981.25	1800.00
sensor_32	220252	804.28	260.60	0.24	760.61	878.85	943.88	1839.21
sensor_33	220304	486.41	150.75	6.46	489.76	512.27	555.16	1578.60
sensor_34	220304	234.97	88.38	54.88	172.49	226.36	316.84	425.55
sensor_35	220304	427.13	141.77	0.00	353.18	473.35	528.89	694.48
sensor_36	220304	593.03	289.39	2.26	288.55	709.67	837.33	984.06
sensor_37	220304	60.79	37.60	0.00	28.80	64.30	90.82	174.90
sensor_38	220293	49.66	10.54	24.48	45.57	49.48	53.65	417.71
sensor_39	220293	36.61	15.61	19.27	32.55	35.42	39.06	547.92
sensor_40	220293	68.84	21.37	23.44	57.81	66.41	77.86	512.76
sensor_41	220293	35.37	7.90	20.83	32.55	34.90	37.76	420.31
sensor_42	220293	35.45	10.26	22.14	32.81	35.16	36.98	374.22
sensor_43	220293	43.88	11.04	24.48	39.58	42.97	46.61	408.59
sensor_44	220293	42.66	11.58	25.75	36.75	40.51	45.14	1000.00
sensor_45	220293	43.09	12.84	26.33	36.75	40.22	44.85	320.31
sensor_46	220293	48.02	15.64	26.33	40.51	44.85	51.22	370.37
sensor_47	220293	44.34	10.44	27.20	39.06	42.53	46.59	303.53
sensor_48	220293	150.89	82.24	26.33	83.91	138.02	208.33	561.63
sensor_49	220293	57.12	19.14	26.62	47.74	52.66	60.76	464.41

7.5. EDA - Análisis Exploratorio de Datos

7.5.1. Distribución de las clases

Los datos contienen 3 clases: Normal, Recuperación y Roto. 205836 puntos de datos pertenecen a la clase Normal, 14477 puntos de datos pertenecen a la clase Recuperación y 7 puntos de datos pertenecen a la clase Roto.

Figura 7.1: Distribución de clases



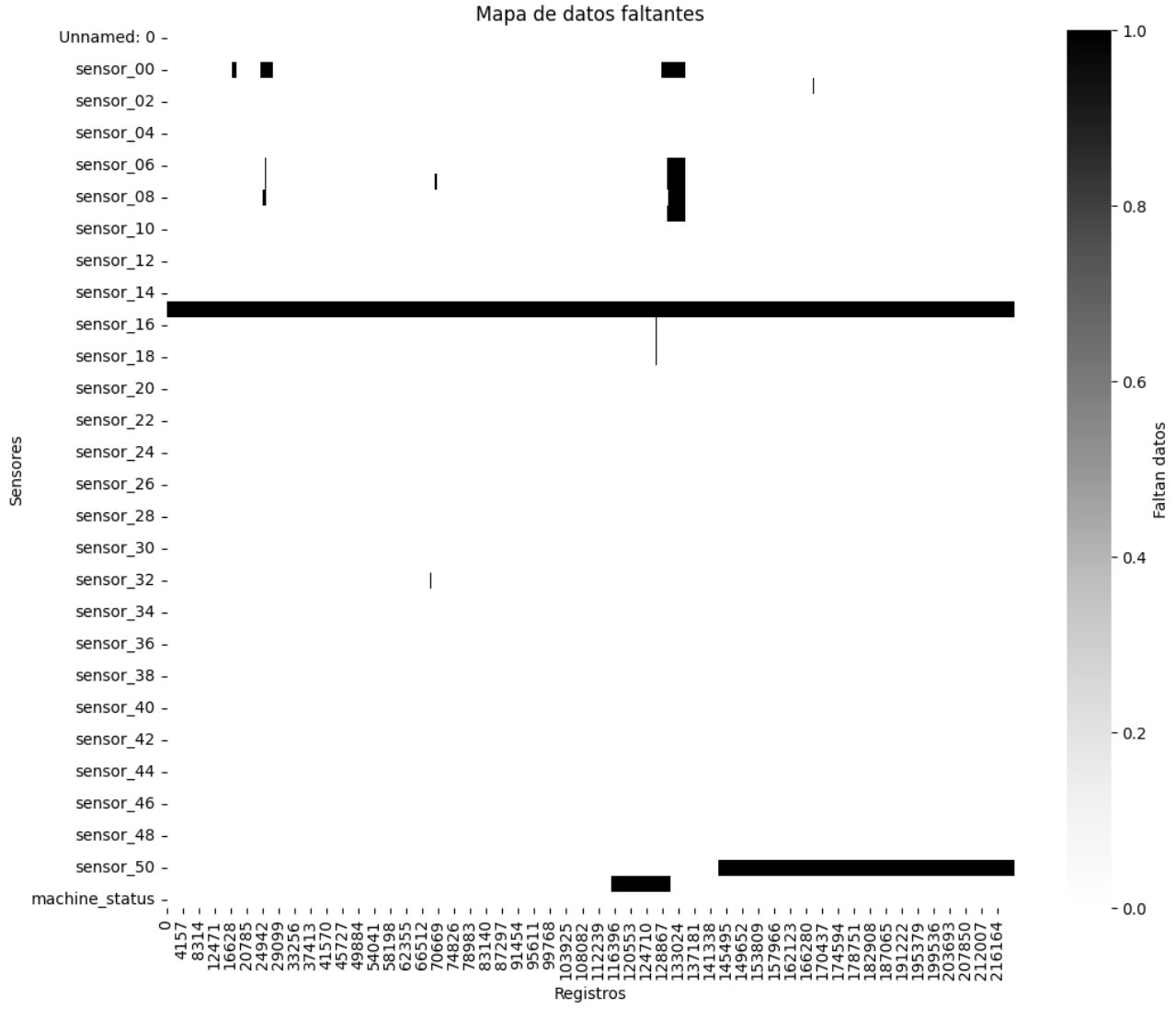
7.5.2. Missing Values

Para crear el gráfico, primero se utiliza el método isna de pandas para seleccionar solo los valores faltantes del conjunto de datos. Luego, se utiliza el método transpose para transponer el conjunto de datos, de modo que las columnas se conviertan en filas y las filas en columnas.

A continuación, se utiliza la función heatmap de seaborn para crear un mapa de calor del conjunto de datos. El mapa de calor muestra una tabla de valores en forma de mapa de colores, donde los colores más oscuros indican valores más altos y los colores más claros indican valores más bajos. En este caso, el mapa de calor muestra la cantidad de valores faltantes en cada columna.

Es importante tener en cuenta que esta función solo muestra el gráfico y no devuelve ningún valor.

Figura 7.2: Mapa de Calor de Missing Values



7.5.3. Visualización de sensores

En esta sección, mostramos gráficas con los datos de los sensores tras eliminar los sensores con mayor número de **Missing Values**. Estas gráficas nos permiten visualizar de manera clara y precisa la información obtenida de cada sensor y analizar cómo varían los valores a lo largo del tiempo. Además, nos ayudan a detectar patrones y tendencias que pueden ser útiles para el análisis y la clasificación de los datos. Por tanto, la visualización de los datos es una herramienta esencial para comprender y explotar al máximo el potencial de los datos que se han recogido.

Aunque no tenemos mucho conocimiento sobre las bombas de agua y sus sensores, es importante examinar los datos y buscar patrones que puedan ser útiles para predecir fallos o anomalías. Al observar las gráficas, podemos ver que hay grandes cambios en los valores de muchos de los sensores en determinados momentos. En estos momentos, algunos sensores caen a valores muy bajos, mientras que otros aumentan y otros, que normalmente son más estables, fluctúan mucho.

Podemos inferir que en estos momentos la máquina ha fallado y se está recuperando al menos una vez. Esto podría ser un indicador importante para predecir futuros fallos y es algo que deberemos tener en cuenta a la hora de desarrollar nuestro modelo de clasificación. Otra tarea que deberemos abordar es el tratamiento de los datos que faltan, ya que esto puede afectar a la precisión de nuestros resultados.

Figura 7.3: sensor_01

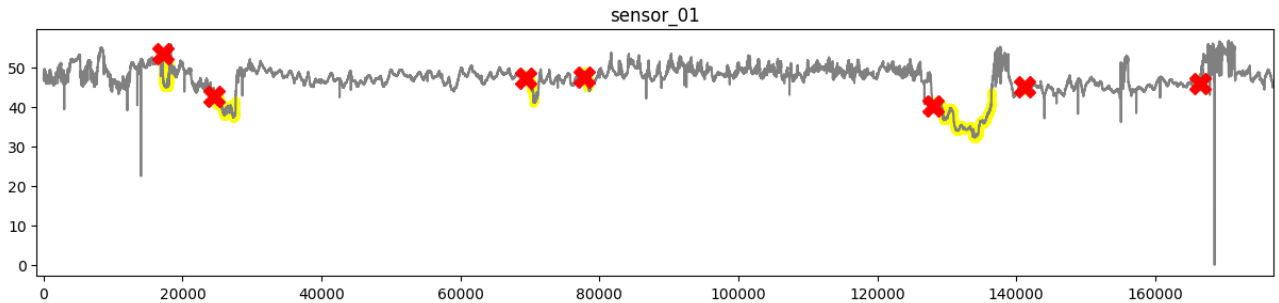


Figura 7.4: sensor_02

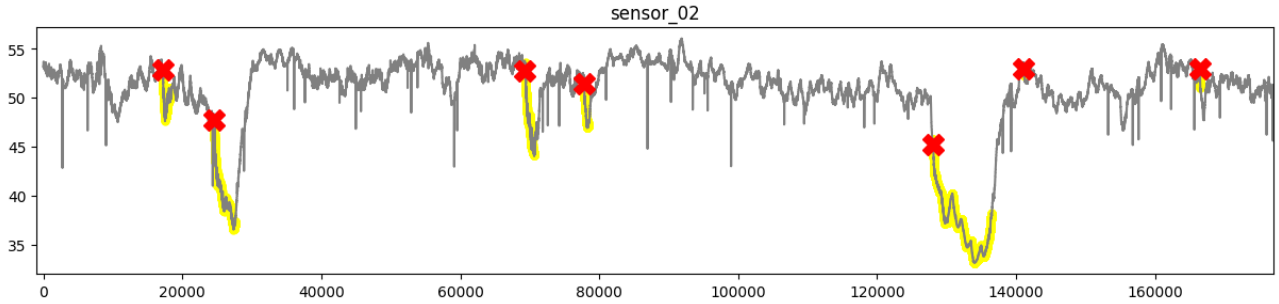


Figura 7.5: sensor_03

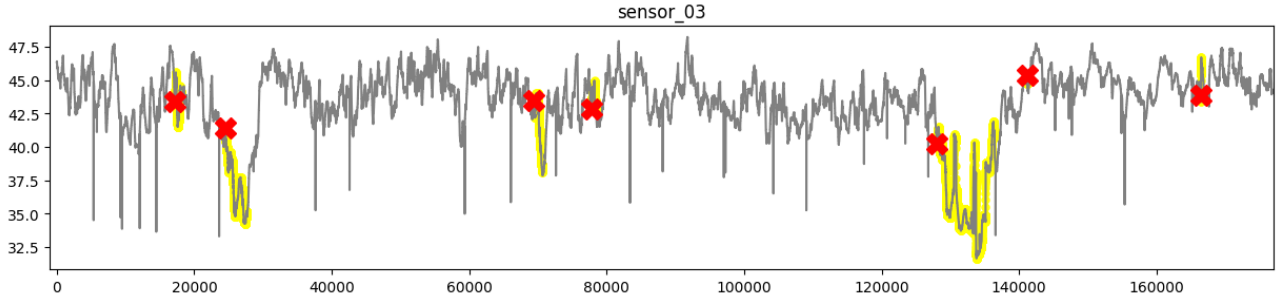


Figura 7.6: sensor_04

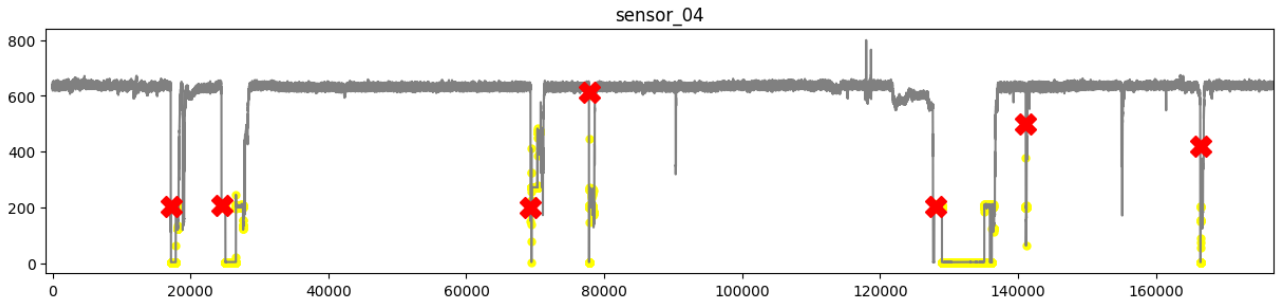


Figura 7.7: sensor_05

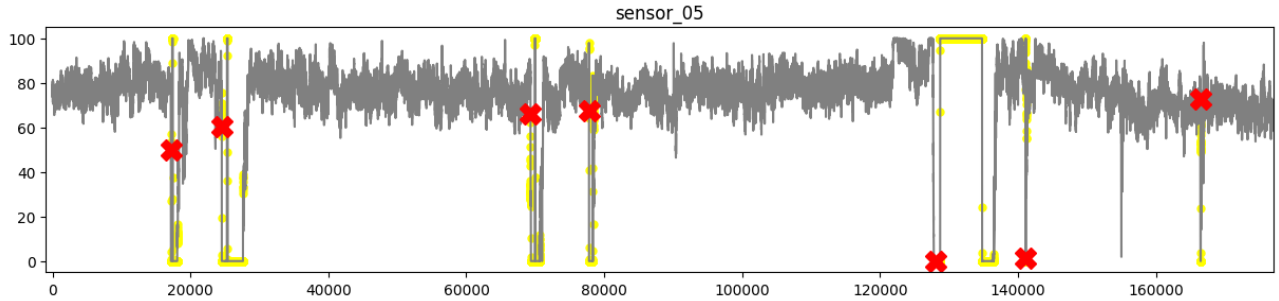


Figura 7.8: sensor_06

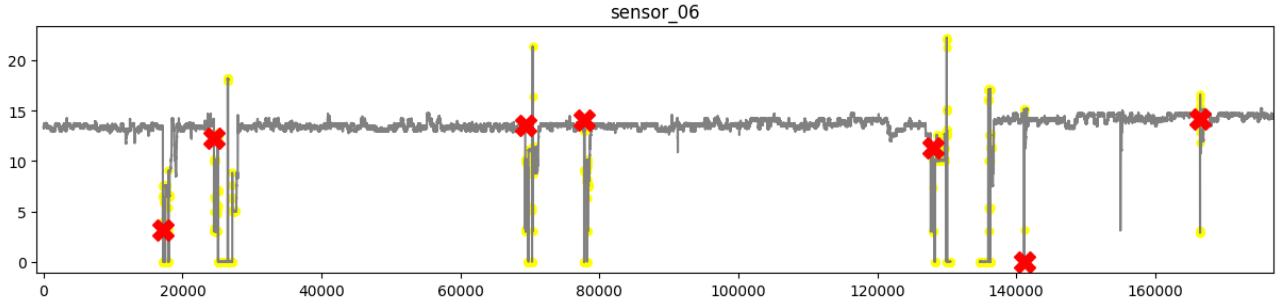


Figura 7.9: sensor_07

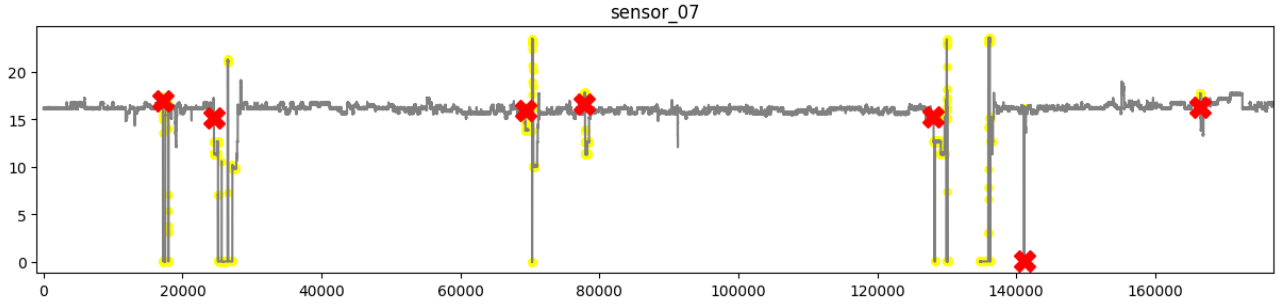


Figura 7.10: sensor_08

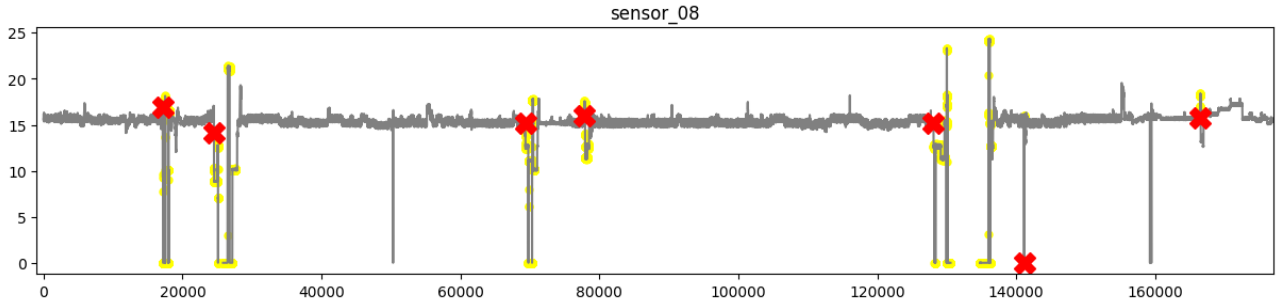


Figura 7.11: sensor_09

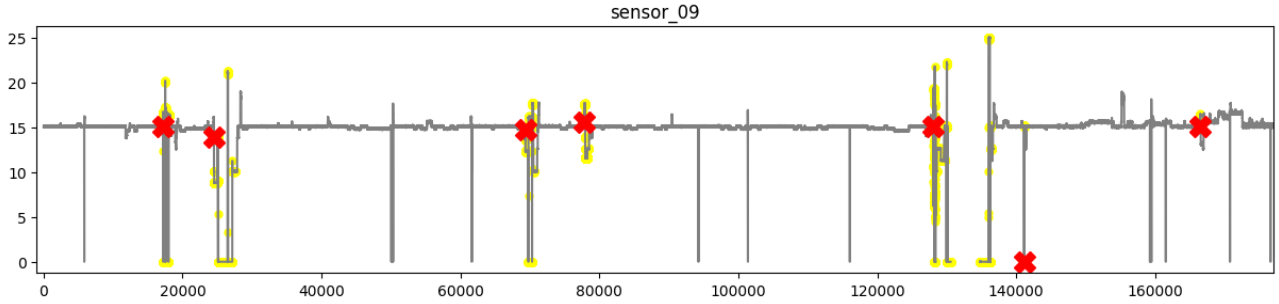


Figura 7.12: sensor_10

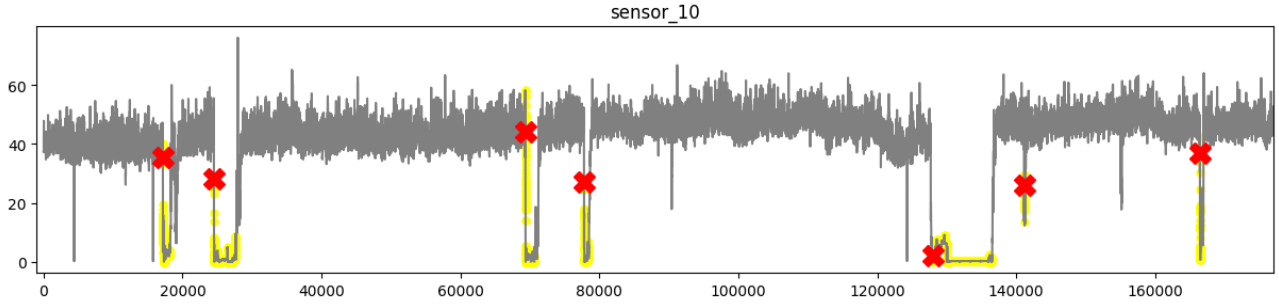


Figura 7.13: sensor_11

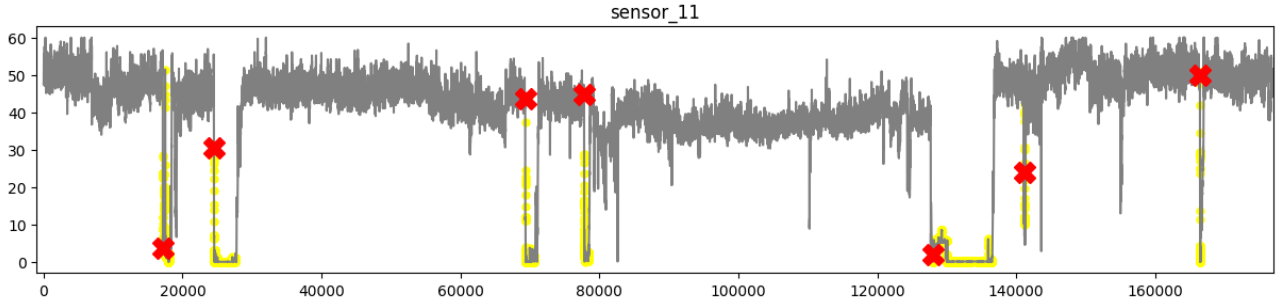


Figura 7.14: sensor_12

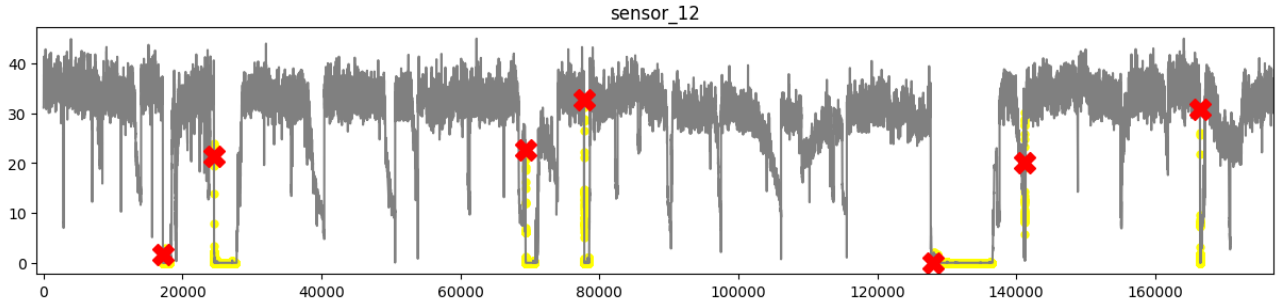


Figura 7.15: sensor_13

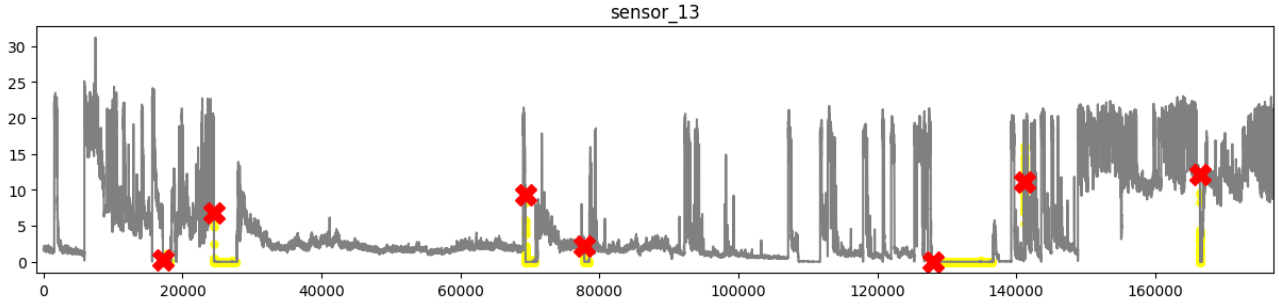


Figura 7.16: sensor_14

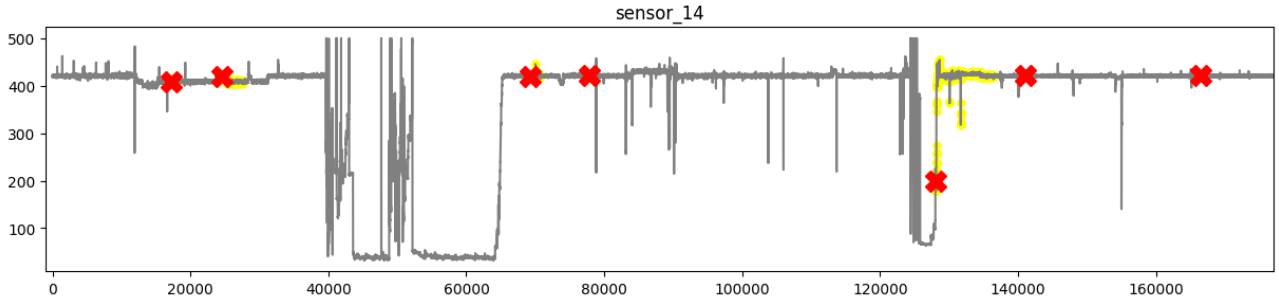


Figura 7.17: sensor_16

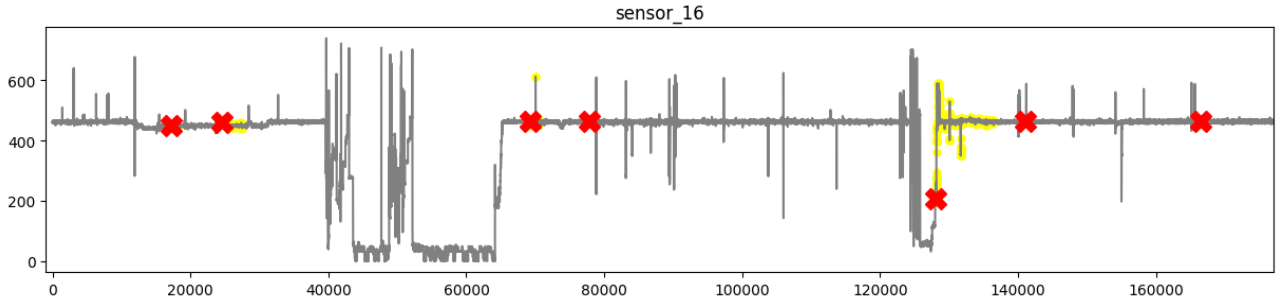


Figura 7.18: sensor_17

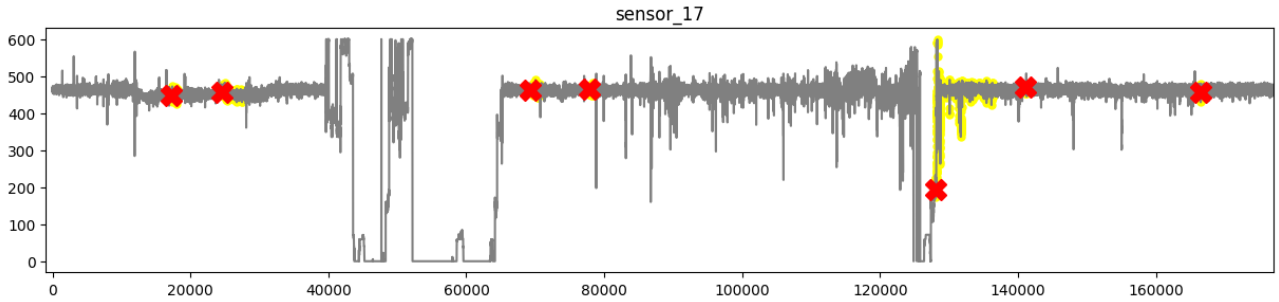


Figura 7.19: sensor_18

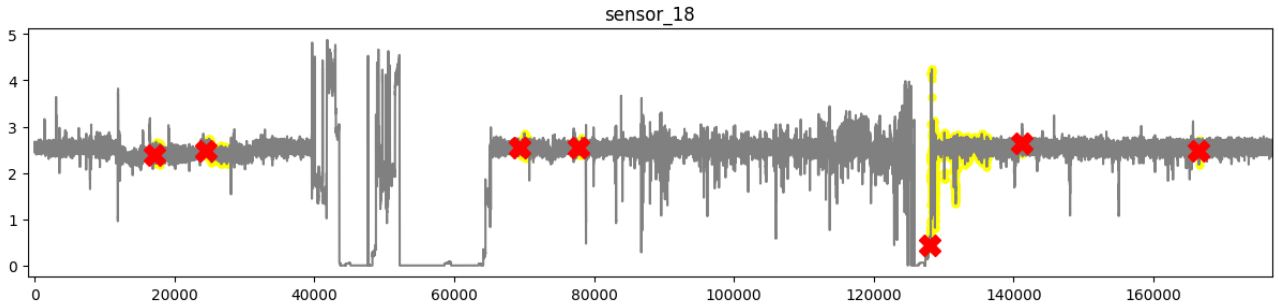


Figura 7.20: sensor_19

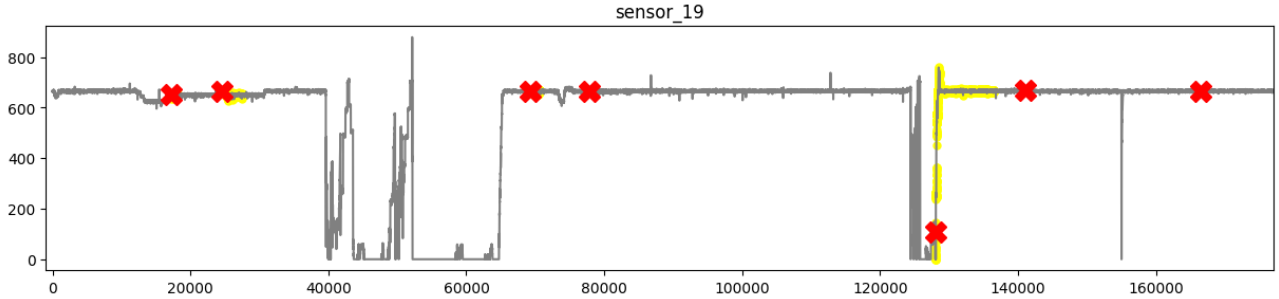


Figura 7.21: sensor_20

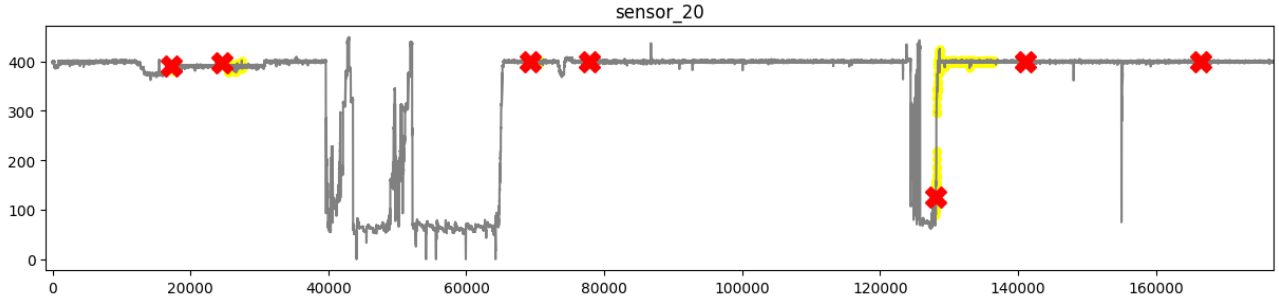


Figura 7.22: sensor_21

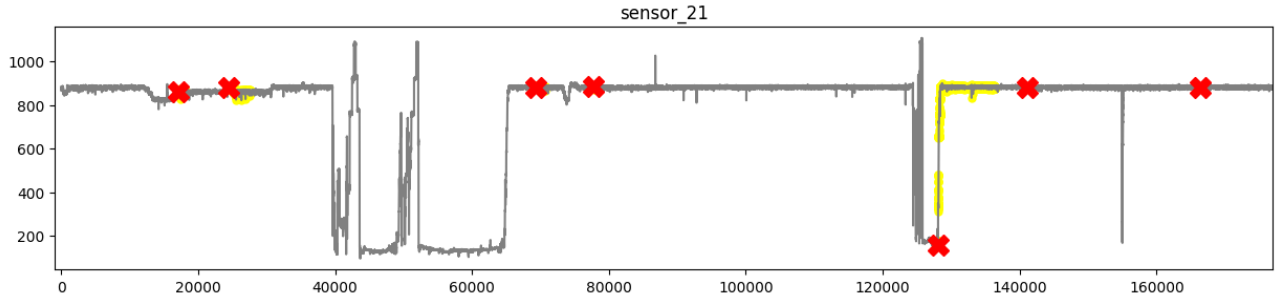


Figura 7.23: sensor_22

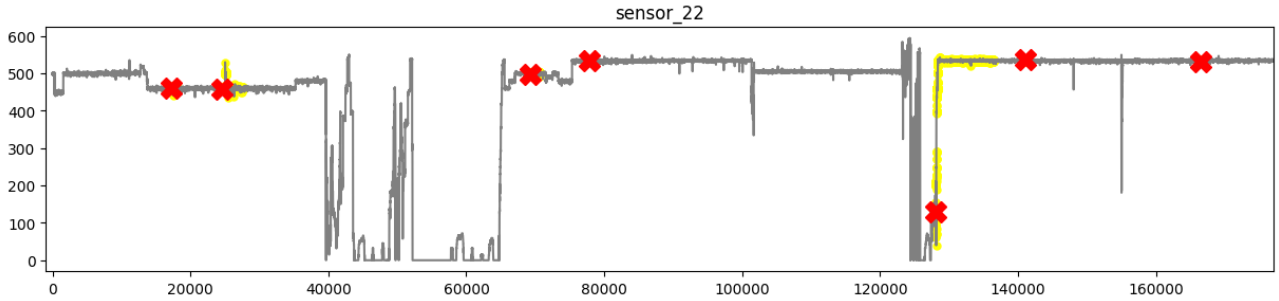


Figura 7.24: sensor_23

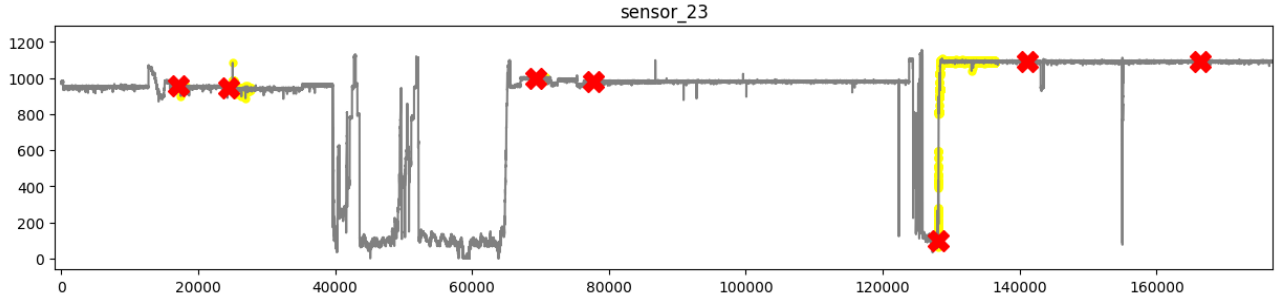


Figura 7.25: sensor_24

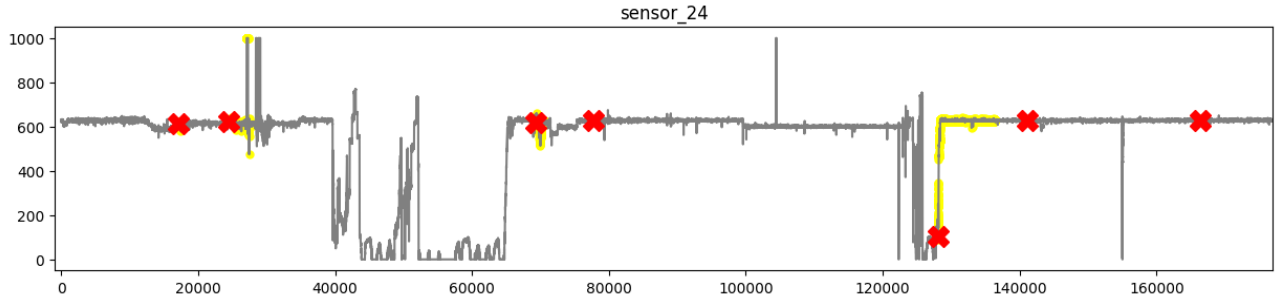


Figura 7.26: sensor_25

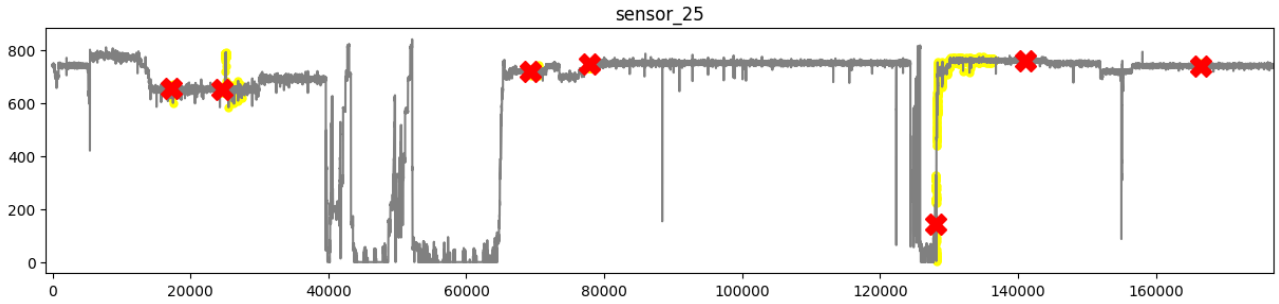


Figura 7.27: sensor_26

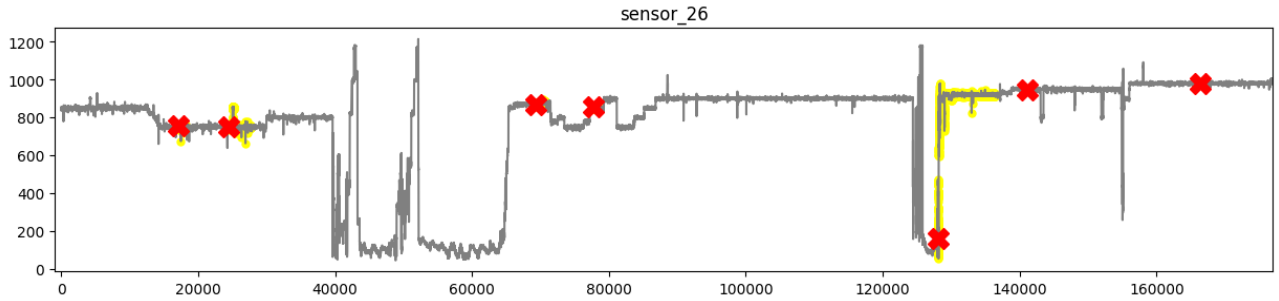


Figura 7.28: sensor_27

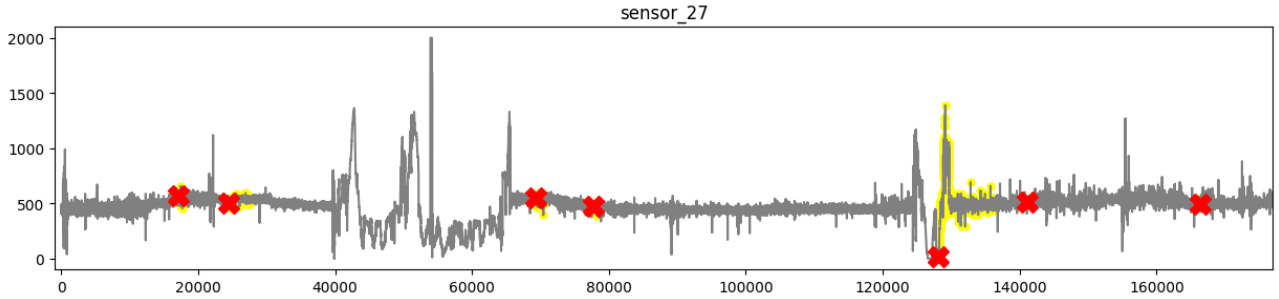


Figura 7.29: sensor_28

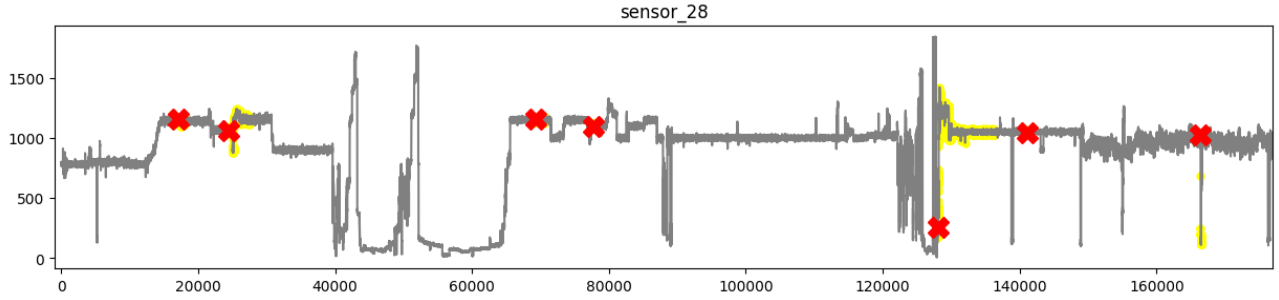


Figura 7.30: sensor_29

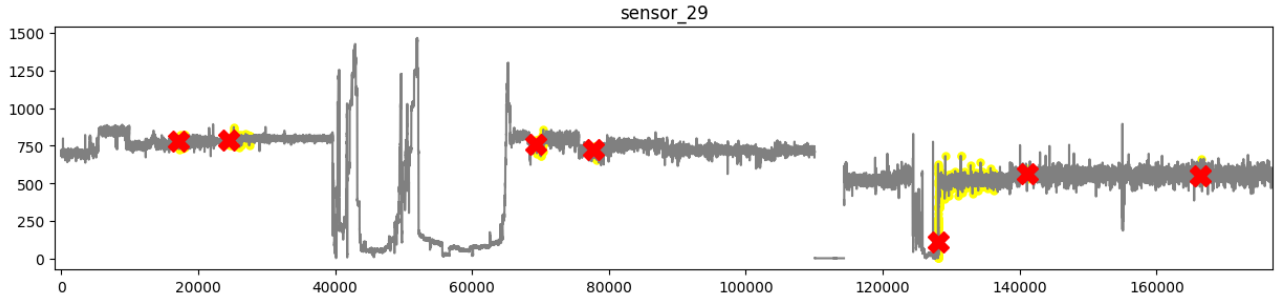


Figura 7.31: sensor_30

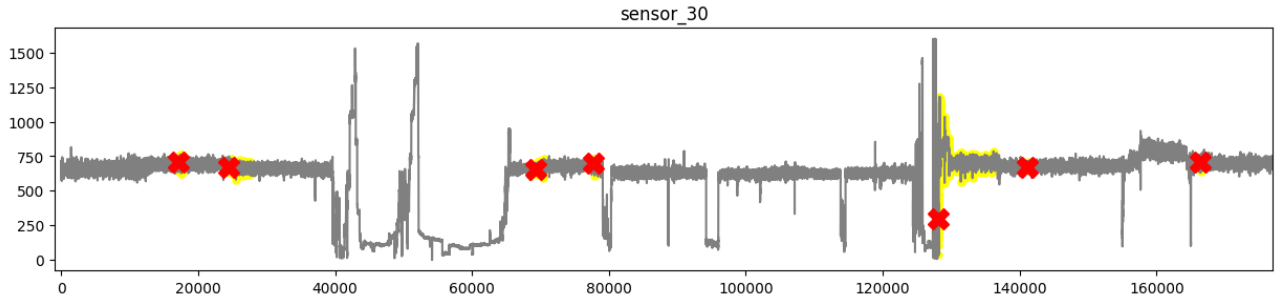


Figura 7.32: sensor_31

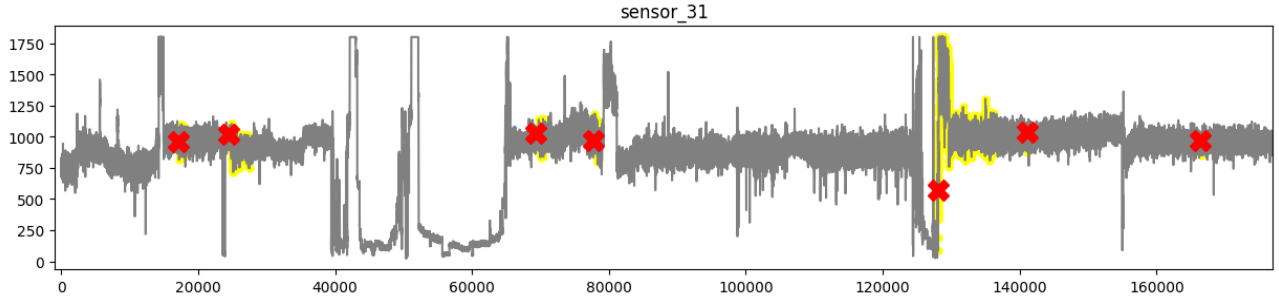


Figura 7.33: sensor_32

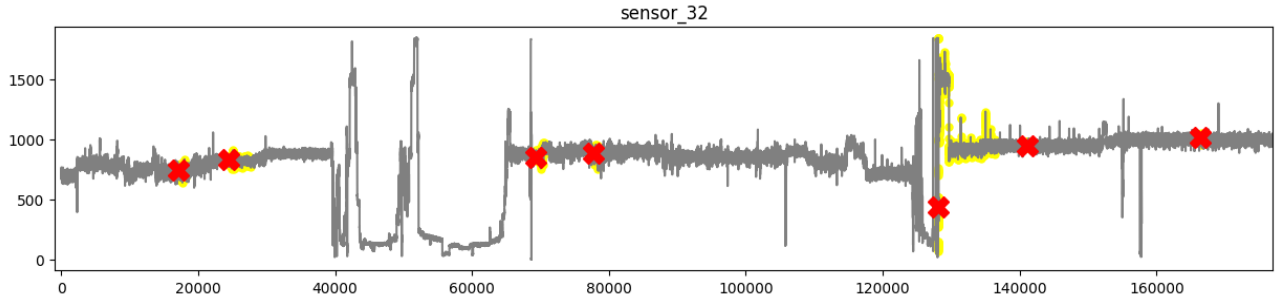


Figura 7.34: sensor_33

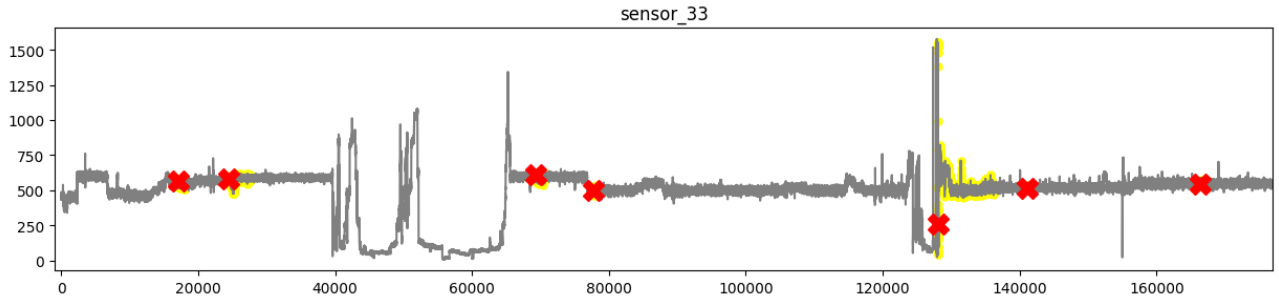


Figura 7.35: sensor_34

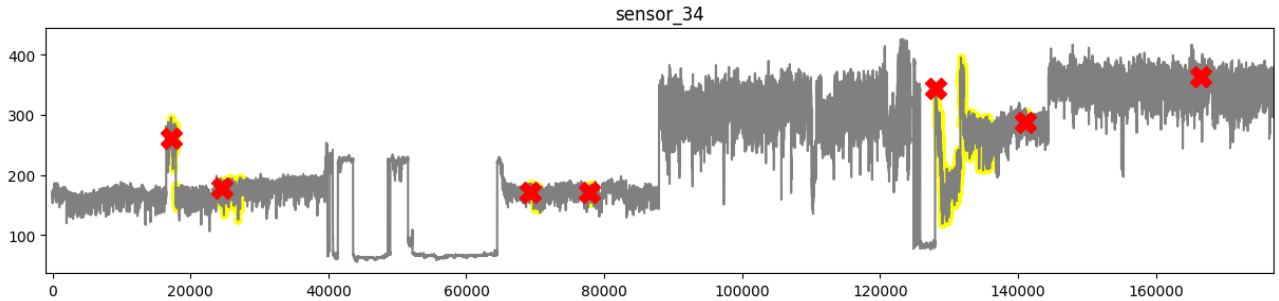


Figura 7.36: sensor_35

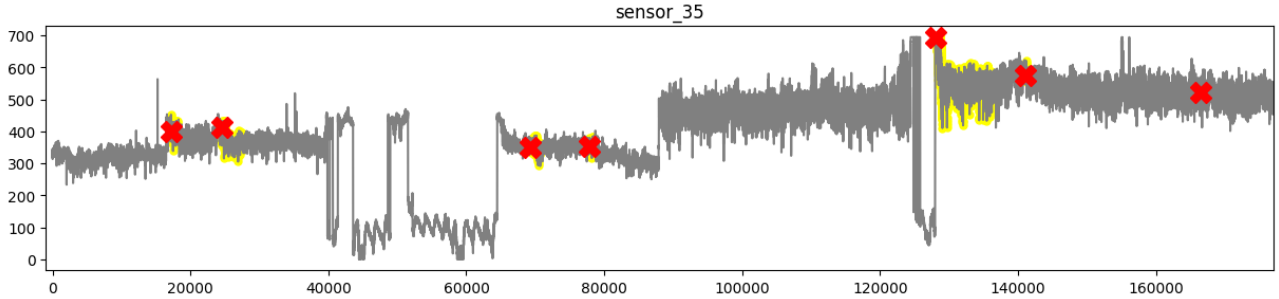


Figura 7.37: sensor_36

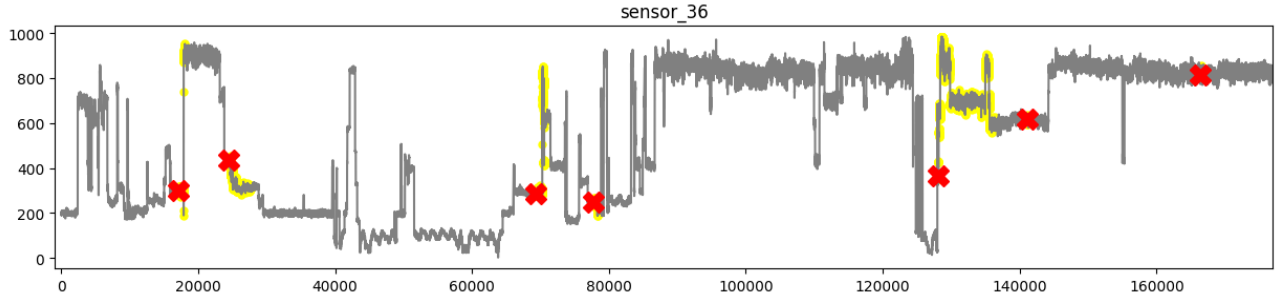


Figura 7.38: sensor_37

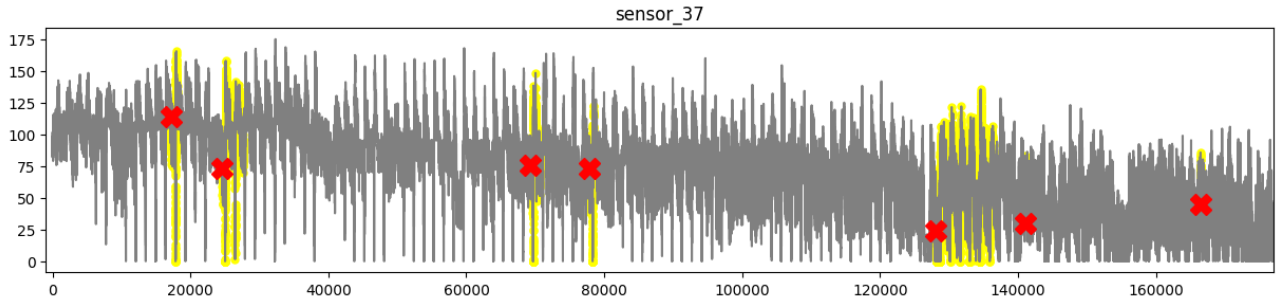


Figura 7.39: sensor_38

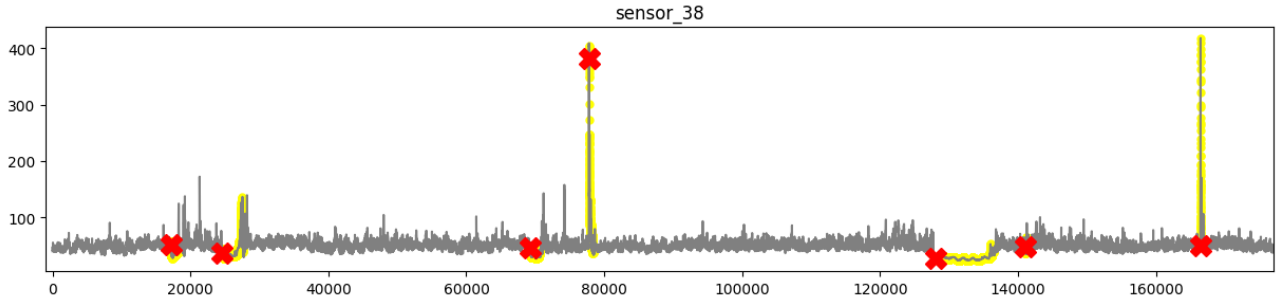


Figura 7.40: sensor_39

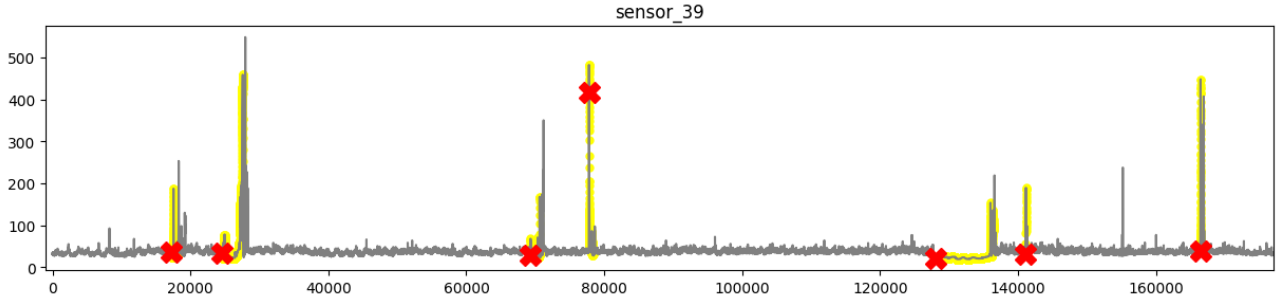


Figura 7.41: sensor_40

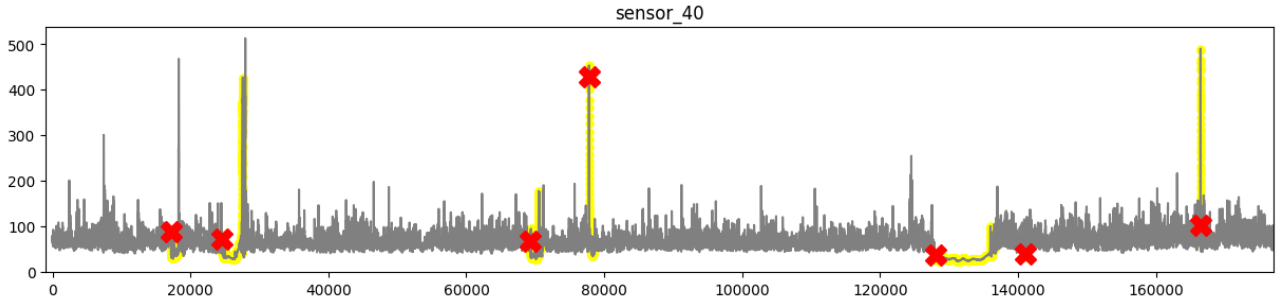


Figura 7.42: sensor_41

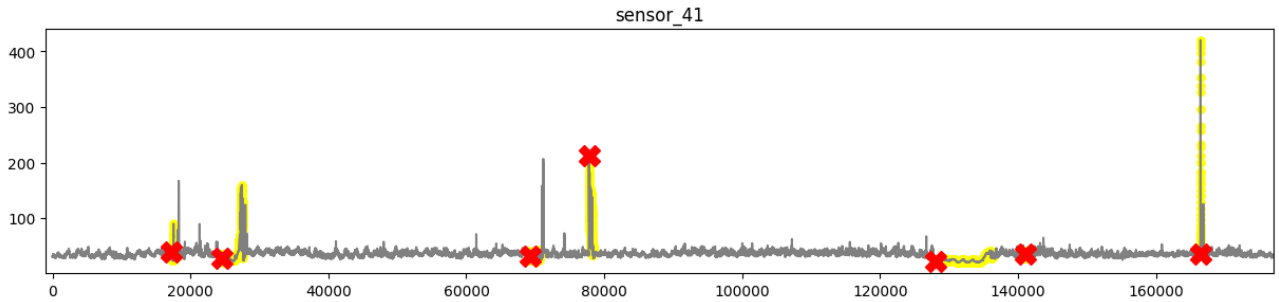


Figura 7.43: sensor_42

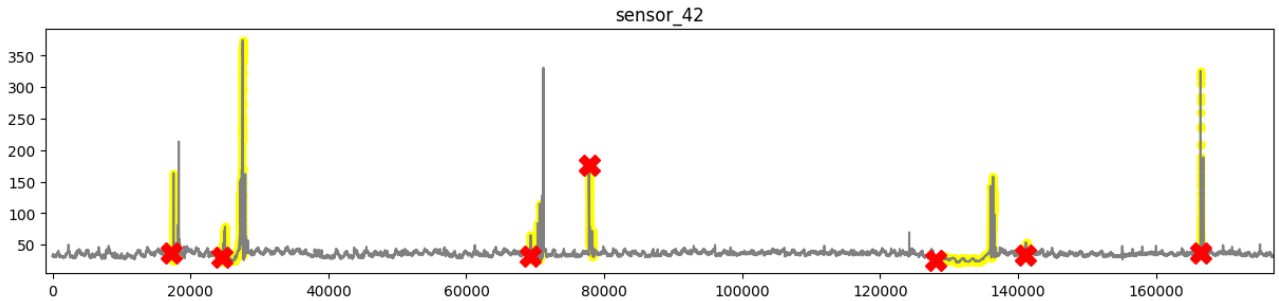


Figura 7.44: sensor_43

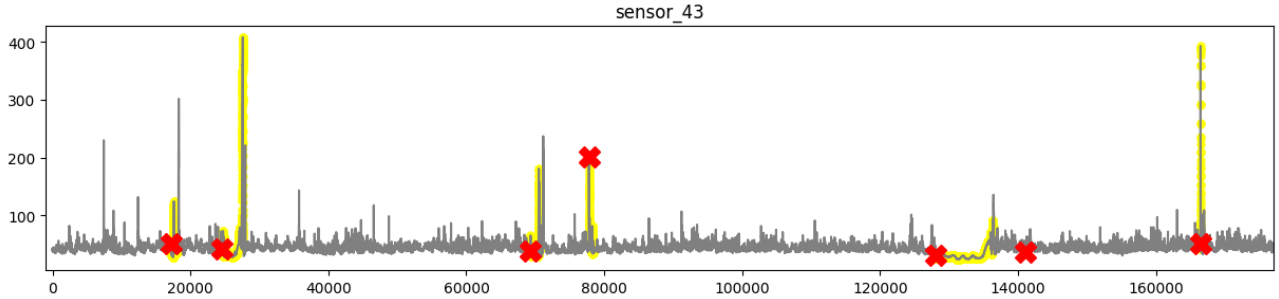


Figura 7.45: sensor_44

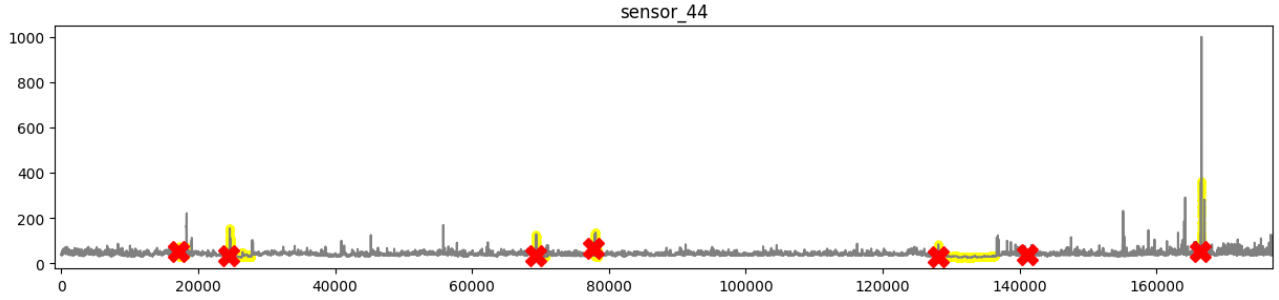


Figura 7.46: sensor_45

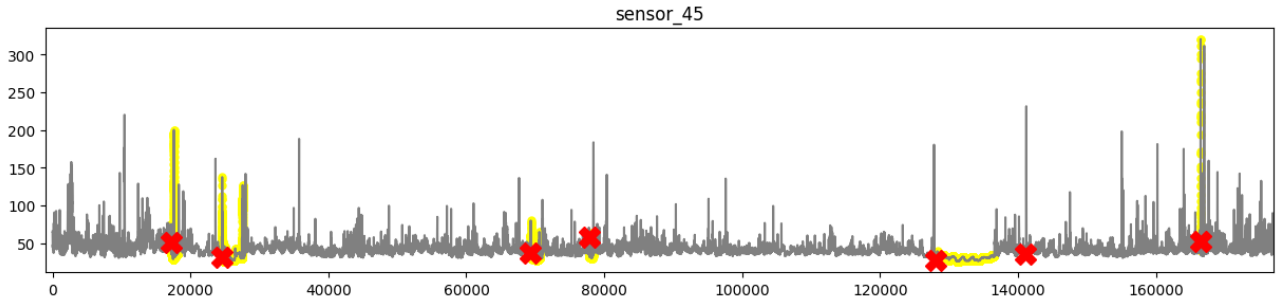


Figura 7.47: sensor_46

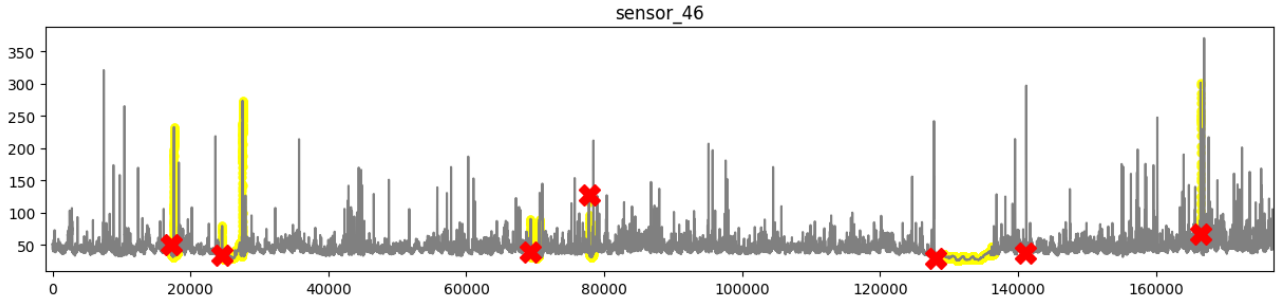


Figura 7.48: sensor_47

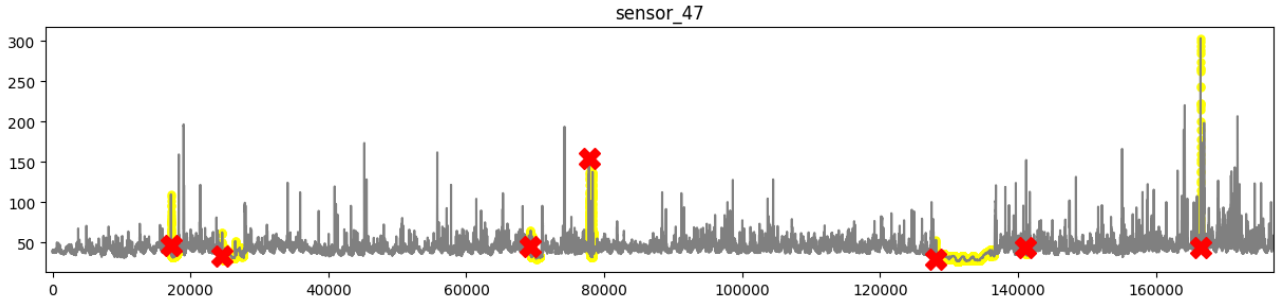


Figura 7.49: sensor_48

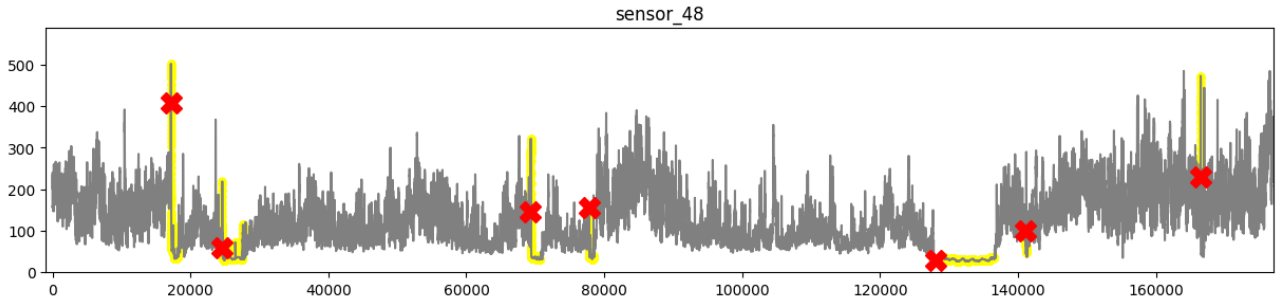
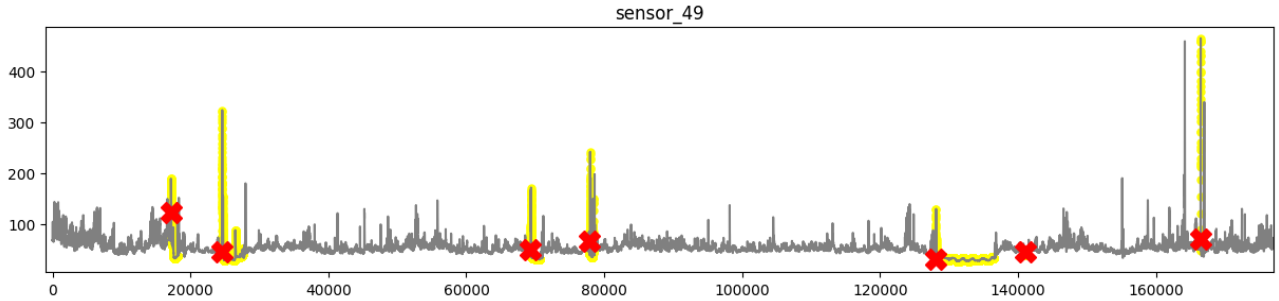


Figura 7.50: sensor_49



Para las 2 últimas gráficas hemos generado 2 nuevos valores:

- Primero, se define un diccionario m que asigna valores numéricos a los distintos estados de la máquina: 'NORMAL' se asigna el valor 2, 'RECOVERING' se asigna el valor 1 y 'BROKEN' se asigna el valor 0.

- Luego, se utiliza el método map de pandas para reemplazar los valores de la columna 'machine_status' con los valores numéricos del diccionario m. El resultado se asigna a la nueva columna 'stat'.

- A continuación, se utiliza el método rolling de pandas para calcular el promedio móvil de la columna 'stat'. El promedio móvil es el promedio de una ventana de datos móviles a lo largo del tiempo. En este caso, se utiliza una ventana de dos elementos, lo que significa que se calcula el promedio de cada par de elementos consecutivos. El resultado se asigna a la nueva columna 'rol'.

Figura 7.51: stat

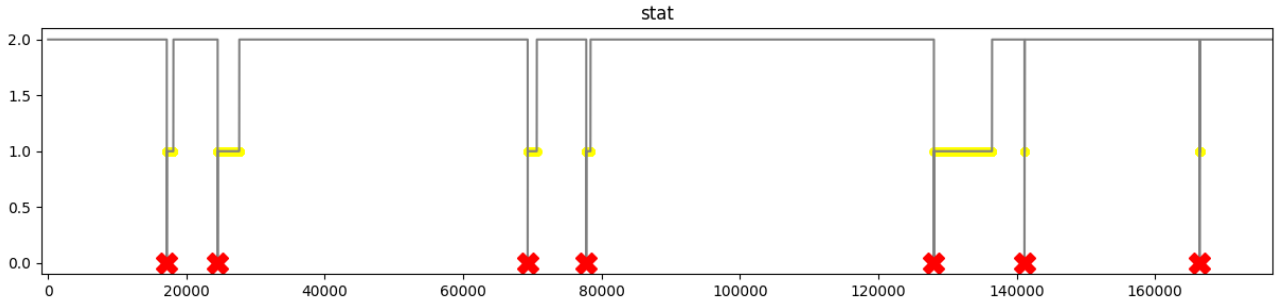
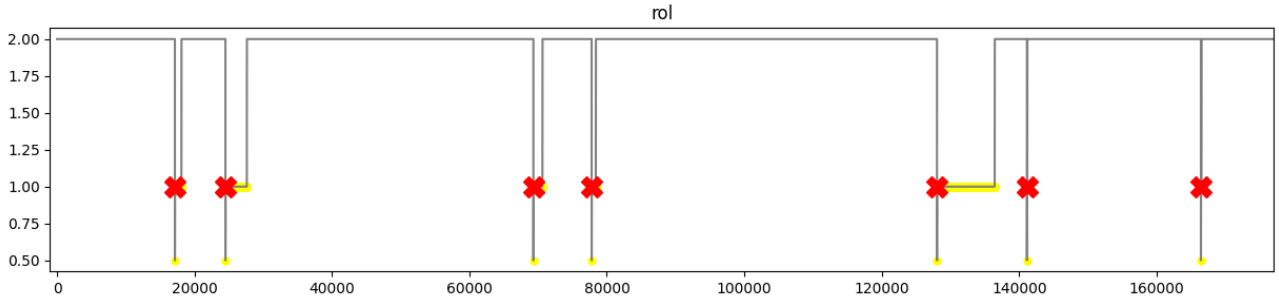


Figura 7.52: rol



7.5.4. Matriz de correlaciones

La función corr_heat_map devuelve un gráfico de mapa de calor que muestra la correlación entre las columnas numéricas del conjunto de datos.

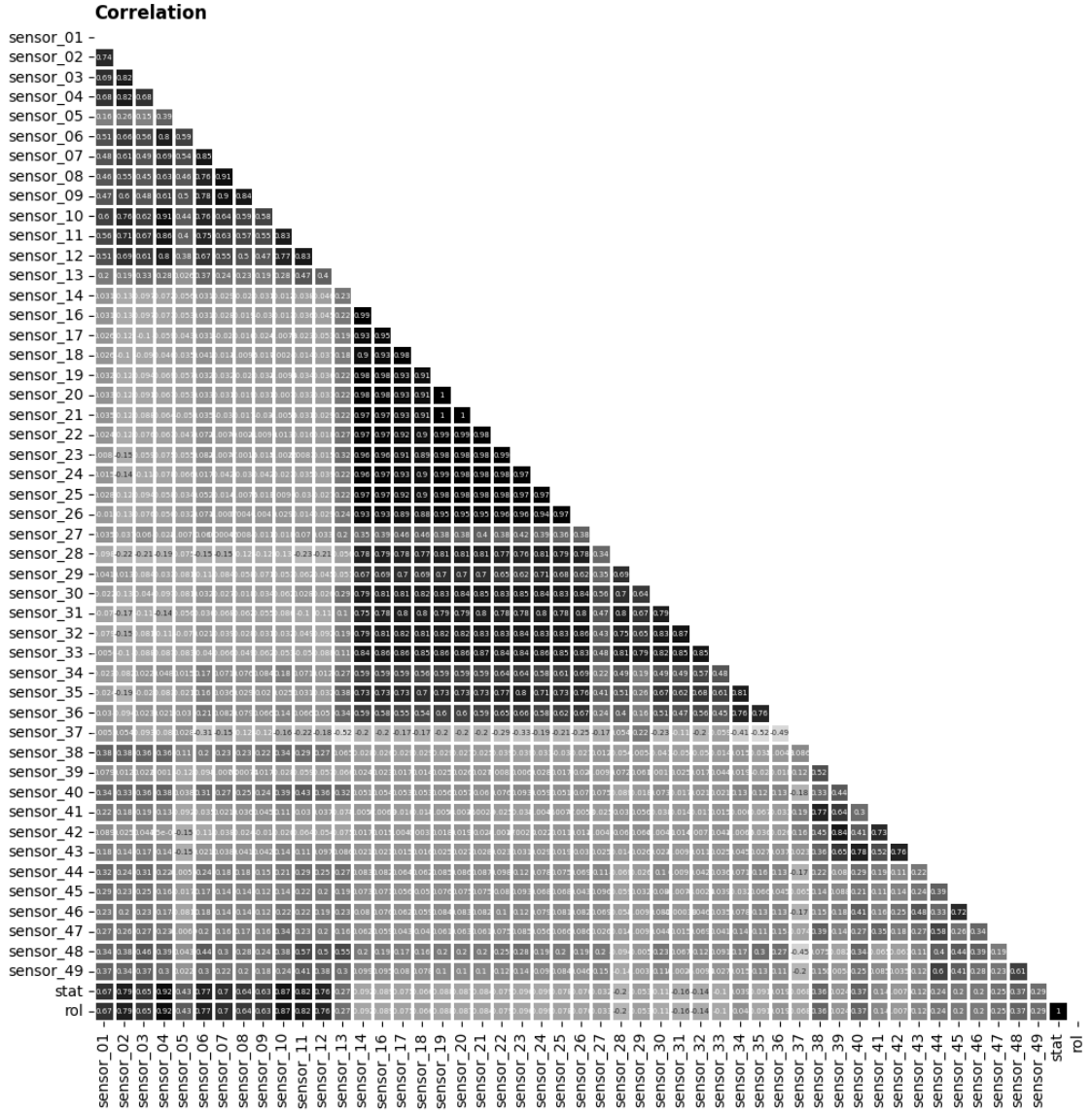
La correlación es una medida de la relación lineal entre dos variables. Una correlación alta indica que las variables están estrechamente relacionadas y una correlación baja indica que las variables no están relacionadas.

Para calcular la correlación, se utiliza el método corr de pandas, que calcula la correlación entre todas las parejas de columnas del conjunto de datos. El resultado se asigna a la variable data_clean_corr.

A continuación, se utiliza la función heatmap de seaborn para crear un mapa de calor de la matriz de correlación. El mapa de calor muestra una tabla de valores en forma de mapa de colores, donde los colores más oscuros indican valores más altos y los colores más claros indican valores más bajos. En este caso, el mapa de calor muestra la correlación entre las columnas numéricas del conjunto de datos.

Para la selección de las variables seleccionamos los sensores con una correlación superior al 75 % con la variable Stat.

Figura 7.53: Matriz de correlaciones



7.5.4.1. Resumen de EDA:

- El conjunto de datos contiene 3 clases: Normal, Recuperación y Roto. 205836 puntos de datos pertenecen a la clase Normal, 14477 puntos de datos pertenecen a la clase Recuperación y 7 puntos de datos pertenecen a la clase Roto.
- El gráfico de la variación del estado de la máquina con el tiempo muestra que el estado de recuperación sigue al estado de avería y que en el estado de recuperación la bomba intenta recuperarse del estado de avería, por lo que el estado de recuperación se considera estado de avería.
- Hemos reasignado 'NORMAL' con el valor 2, 'RECOVERING' conel valor 1 y 'BROKEN' con el valor 0.
- Al comprobar los valores que faltan, encontramos que todos los sensores tienen valores que faltan, todas las lecturas del sensor_15 faltan, el 34 % de las lecturas del sensor_50 faltan, el 6 % de las lecturas del sensor_51 faltan, el 4 % de las lecturas del sensor_00 faltan y todos los demás sensores tienen menos de un 2 % de valores que faltan.
- Hemos observado en el gráfico de valores perdidos que las lecturas de los sensores faltan tanto en el estado Normal como en el Roto, y que la mayoría de los datos faltan en el estado Roto.
- El sensor_15 se elimina del conjunto de datos, ya que faltan todos sus valores, y los valores que faltan de todos los demás sensores se rellenan con un valor fuera de distribución, -1.
- Se seleccionan los sensores 02, 04, 06, 10, 11 y 12 como características finales por tener una correlación superior al 75 % con la variable stat.

7.6. Modelado

Comenzamos el modelado, que es el proceso de entrenar un modelo a partir de datos de entrada para que pueda hacer predicciones precisas sobre datos futuros. El modelo se construye utilizando un algoritmo de aprendizaje automático y se entrena con un conjunto de datos de entrenamiento. Una vez que el modelo ha sido entrenado, se puede utilizar para hacer predicciones sobre nuevos datos.

Primeramente seleccionamos las variables X e Y

- X son los sensores previamente seleccionados
- Y es el output, en este caso stat.

7.6.1. Normalización de datos

La normalización de los datos es un proceso que se utiliza para ajustar los datos a una escala

$$X\_norm = \frac{X - X.mean()}{X.std()} \tag{7.1}$$

común. En este caso, se están restando la media de cada columna a cada valor de la columna y luego dividiendo el resultado por la desviación estándar de la columna. Esto asegura que todas las columnas tengan una media de cero y una desviación estándar de uno.

La normalización de los datos es a veces útil cuando se comparan diferentes series de datos que pueden estar en escalas muy diferentes. También puede ser útil para algunos algoritmos de aprendizaje automático que se benefician de tener todas las características en la misma escala.

7.6.2. Seleccionar el tamaño de la muestra de validación

Con la función la función 'train_test_split' del módulo 'model_selection' de scikit-learn para dividir el conjunto de datos en dos conjuntos: un conjunto de entrenamiento y un conjunto de validación.

Los conjuntos de entrenamiento y validación se utilizan para evaluar y ajustar el modelo de aprendizaje automático. El conjunto de entrenamiento se utiliza para entrenar el modelo y el conjunto de validación se utiliza para evaluar cómo bien el modelo generaliza a datos que no ha visto antes. Esta división es importante porque queremos evaluar el rendimiento del modelo en datos que no ha visto durante el entrenamiento, ya que esto es más representativo de cómo el modelo se desempeñará en el mundo real.

En este caso, la función 'train_test_split' está dividiendo los datos de entrada 'X_norm' y las etiquetas 'y' en dos conjuntos: 'X_train' y 'X_validation' para los datos de entrada y 'Y_train' y 'Y_validation' para las etiquetas. La proporción de datos en el conjunto de entrenamiento y el conjunto de validación se controla con el parámetro 'test_size'. El parámetro 'random_state' asegura que los datos se dividan de manera consistente cada vez que se ejecuta el código.

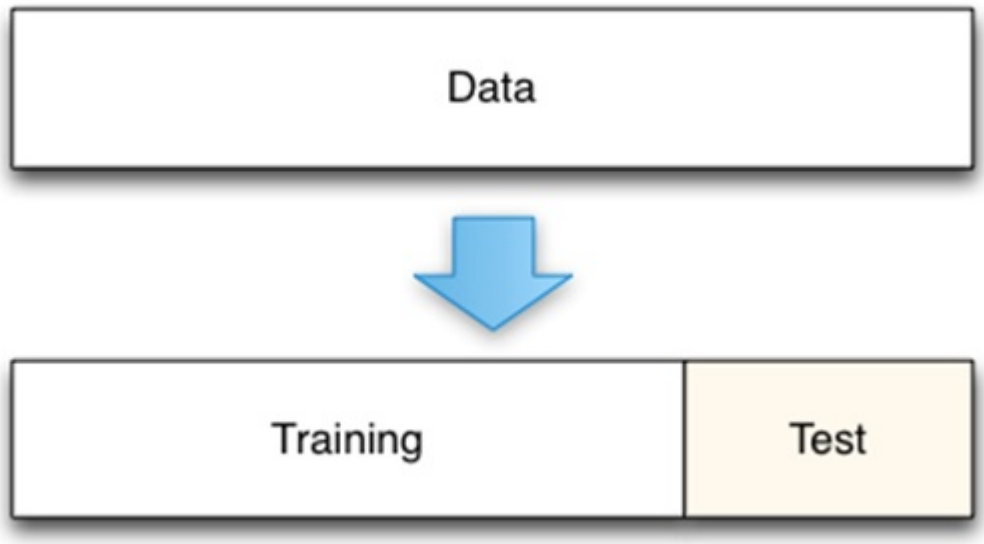
$$validation\_size = 0{,}20 \tag{7.2}$$

$$seed = 7 \tag{7.3}$$

7.6.3. Modelos utilizados

Hemos entrenado seis modelos diferentes:

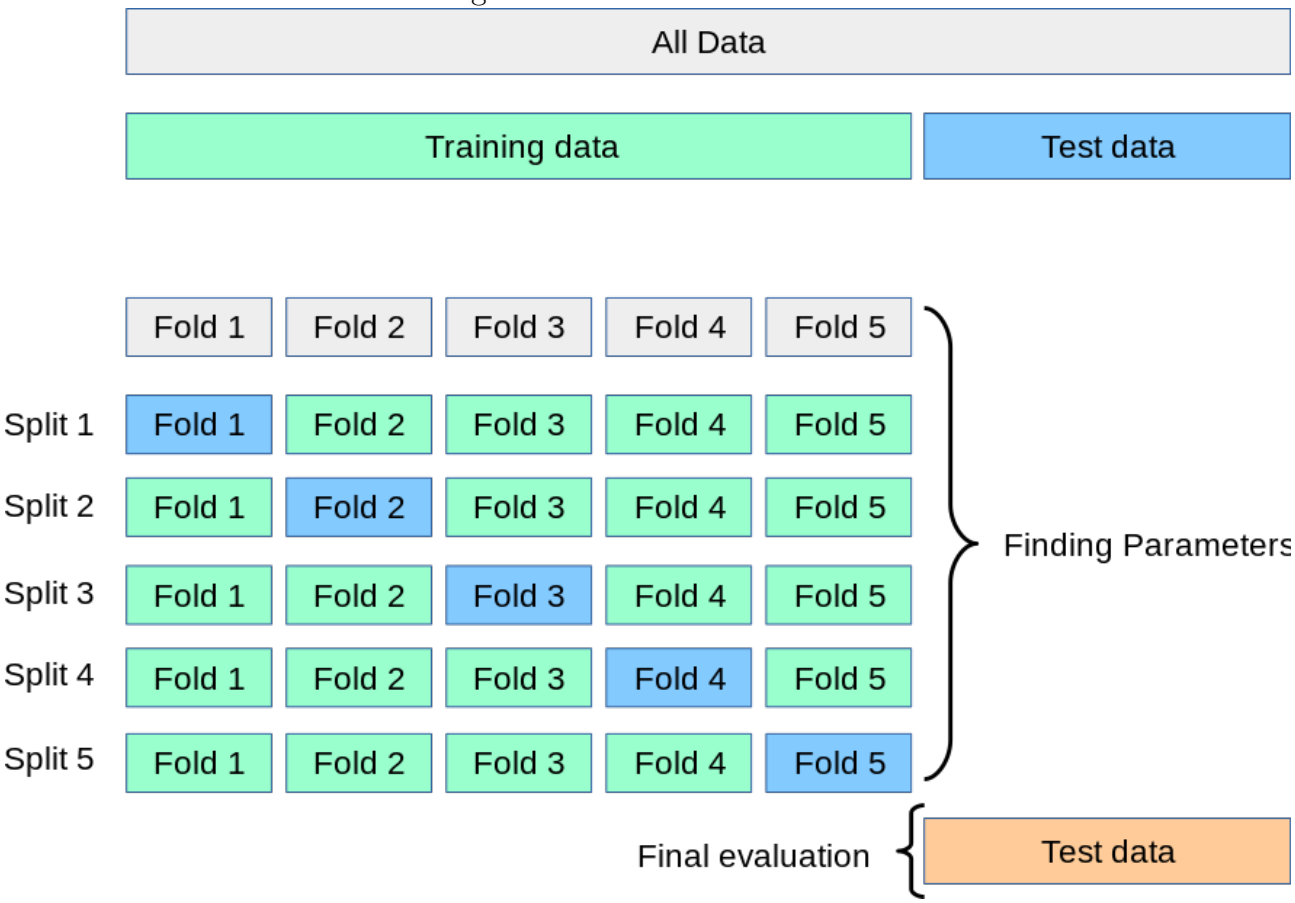
Figura 7.54: División de datos



- Logistic Regression (LR)
- Linear Discriminant Analysis (LDA)
- K-Nearest Neighbors (KNN).
- Classification and Regression Trees (CART).
- Gaussian Naive Bayes (NB).
- Support Vector Machines (SVM).

Para ello hemos utilizado la validación cruzada, que es una técnica utilizada para evaluar el rendimiento de un modelo de aprendizaje automático. Consiste en dividir el conjunto de datos en 'k' pliegues (o "folds"), donde 'k' es un número especificado por el usuario. Luego, se entrena el modelo 'k' veces, cada vez utilizando un pliegue diferente como el conjunto de validación y los 'k-1' pliegues restantes como el conjunto de entrenamiento. Finalmente, se calcula la puntuación media y la desviación estándar para cada modelo. Esto proporciona una medida más precisa del rendimiento del modelo que si simplemente entrenáramos y evaluáramos el modelo una vez con todos los datos.

Figura 7.55: División de datos



7.6.4. Comparación de modelos

Los resultados son:

- LR: 0.994117 (0.000545) La media de los resultados es: 0.9941165216117712 y la desviación estándar es: 0.0005447067033496082
- LDA: 0.987059 (0.000673) La media de los resultados es: 0.9870585989588235 y la desviación estándar es: 0.000672909331079805
- KNN: 0.999677 (0.000144) La media de los resultados es: 0.9996766073329775 y la desviación estándar es: 0.00014363935391012618
- CART: 0.999523 (0.000132) La media de los resultados es: 0.9995234216377644 y la desviación estándar es: 0.0001323239445151695
- NB: 0.982508 (0.001177) La media de los resultados es: 0.9825084033660783 y la desviación estándar es: 0.0011767107835311446
- SVM: 0.998162 (0.000331) La media de los resultados es: 0.9981617648976006 y la desviación estándar es: 0.0003310197121410058

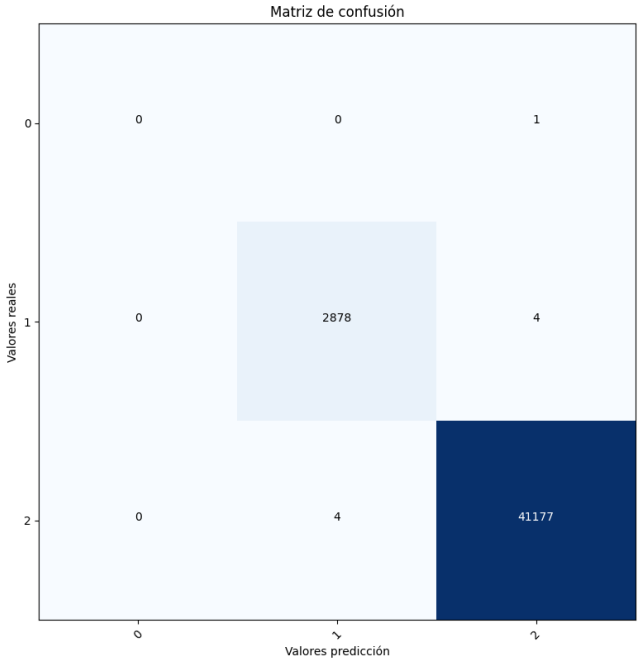
Como vemos, el mejor modelo es KNN.

Con dicho modelo obtenemos la matriz de confusión, una herramienta comúnmente utilizada en el análisis de resultados de clasificación. Muestra la cantidad de predicciones correctas y incorrectas que ha realizado un modelo de clasificación. Se suele presentar en forma de tabla con dos filas y dos columnas.

La fila superior de la matriz de confusión se refiere a las predicciones positivas y la fila inferior se refiere a las predicciones negativas. La columna izquierda de la matriz se refiere a las etiquetas verdaderas negativas y la columna derecha se refiere a las etiquetas verdaderas positivas.

Hemos preprocesado nuestros datos de nuevo para asegurar una pizarra limpia en el nuevo modelo. Ahora usaremos GridSearchCV con KNN para encontrar los mejores parámetros para usar con este modelo.

Figura 7.56: Matriz de confusión para KNN



7.6.5. Ajuste de hiperparámetros

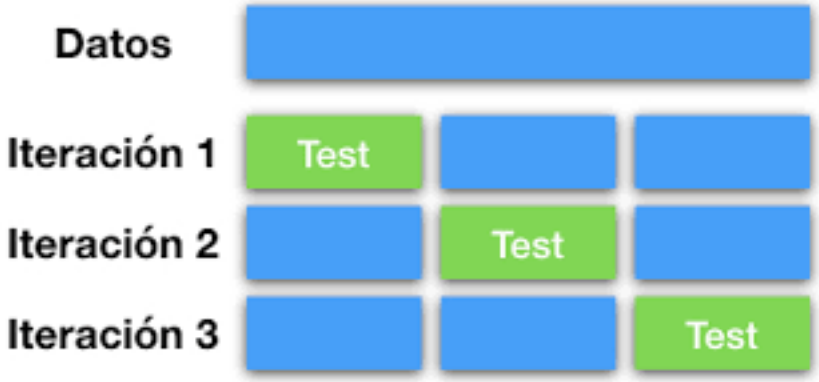
Para obtener el mejor rendimiento de los modelos de aprendizaje automático, ajustaremos los hiperparámetros mediante la técnica de validación cruzada. Debido a la naturaleza temporal de los datos, se debe tener cuidado al dividir los datos para evitar la fuga de datos. Por lo tanto, se implementa la validación cruzada TRIPLE basada en el tiempo y los datos de entrenamiento se dividen como se muestra a continuación.

scikit-learn dispone de varias clases que implementan la metodología de la validación cruzada. En el caso de que se desee utilizar para seleccionar los parámetros de entrenamiento de un modelo una de las opciones es GridSearchCV. Siendo uno de los más simples y fáciles de utilizar. El constructor de esta clase se ha de llamar indicándole la instancia de un modelo, los valores a probar y el número de conjuntos en el que se dividen los datos. Esto se realiza mediante los siguientes parámetros:

- estimator: el modelo que se ha de evaluar
- param_grid: un diccionario en que se indicar los parámetros a evaluar como clave y el conjunto elementos como valor
- cv: el número de conjuntos en los que se divide los datos para la validación cruzada.

En la késima división, los k primeros pliegues se utilizan como datos de entrenamiento y el (k+1)º pliegue se utiliza como datos de prueba. Los hiperparámetros n_estimators (número de

Figura 7.57: CrossValidation



estimadores) y max_depth (profundidad máxima del árbol) se ajustan mediante Grid Search CV. Los parámetros utilizados son:

- 'n_neighbors': [3, 5, 10]
- 'weights': ['uniform', 'distance']
- 'algorithm': ['ball_tree', 'kd_tree']
- 'p': [1, 2]

Y los mejores parámetros son: Best parameters: 'algorithm': 'ball_tree', 'n_neighbors': 5, 'p': 1, 'weights': 'distance'

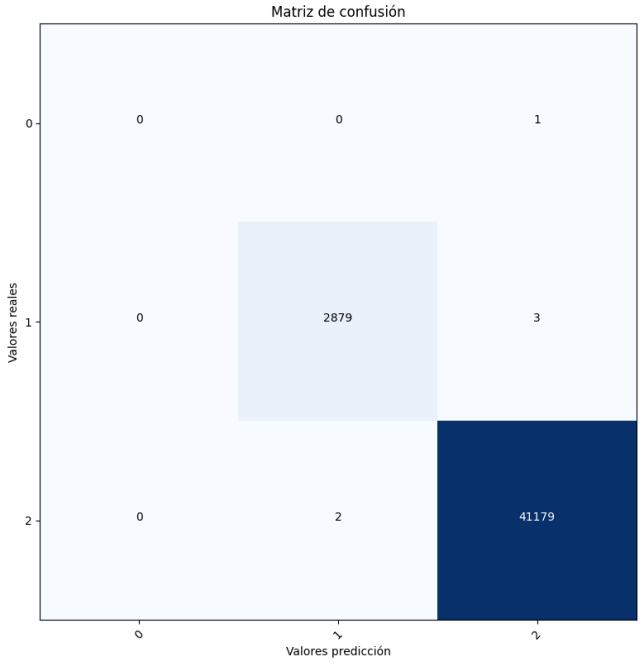
Capítulo 8

Conclusiones y Lineas de trabajo futuras

8.1. Conclusiones

Con dichos parámetros la precisión del modelo es de 0.9998638344226579
Y la matriz de confusión es:

Figura 8.1: KNN Best



Como se ve es más precisa que el anterior modelo.

8.2. Lineas de trabajo futuras

Se desarrolla una aplicación web utilizando HTML y el framework web Django que es capaz de predecir entradas de un único punto y entradas múltiples. Para obtener la predicción de múltiples puntos de datos se debe cargar un archivo csv. La predicción de clase y los valores de probabilidad correspondientes pueden descargarse como un archivo csv.

Para ello guardaremos el modelo con la función "to_picke"de pandas. La función 'to_pickle' de pandas se utiliza para serializar un objeto de Python y guardarlo en un archivo con formato 'pickle'. En este caso, se está serializando el modelo de aprendizaje automático KNN y guardándolo en un archivo llamado 'KNN_model.pickle' en la carpeta 'media'.

La serialización es el proceso de convertir un objeto en una secuencia de bytes, que se puede guardar en un archivo o transmitir a través de una red. El formato 'pickle' es uno de los formatos más comunes para serializar objetos de Python. Al serializar el modelo y guardarlo en un archivo, podemos guardar el estado del modelo entrenado y luego cargarlo más tarde y utilizarlo sin tener que volver a entrenar el modelo.

El producto a obtener con este TFM es una web-app destinada a mostrar los datos recogidos por diferentes sensores instalados en maquinaria para conocer su comportamiento. También poder generar un algoritmo de Machine-Learning para la predicción de anomalías y comportamientos de dichas máquinas.

8.2.1. Desarrollo web

Utilizando Django, un framework de desarrollo de aplicaciones web de código abierto escrito en Python, hemos generado esta página web.

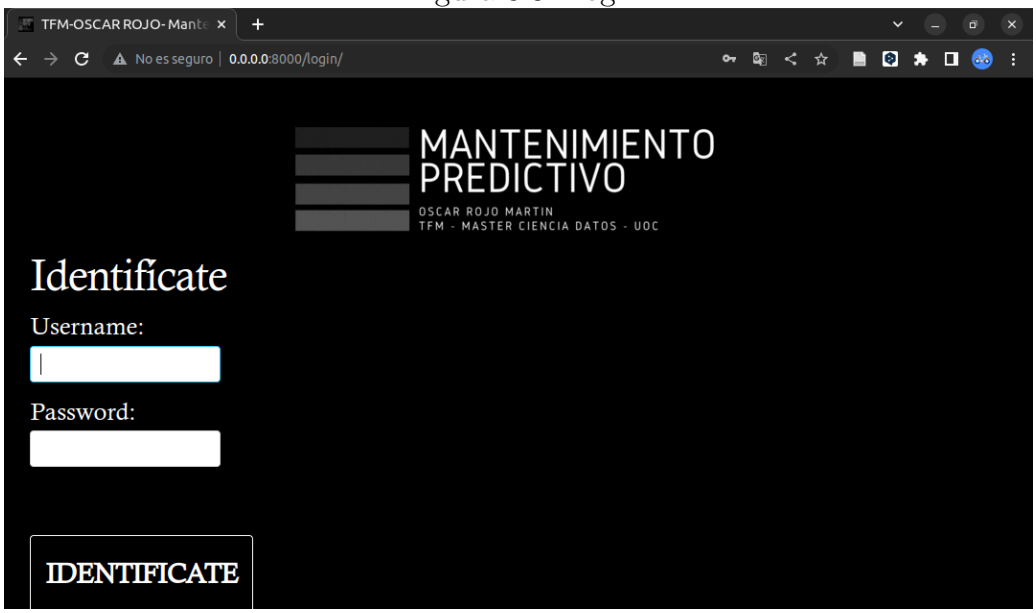
Figura 8.2: Pantalla inicial.

Django,fue diseñado para ser rápido, seguro y escalable y proporciona un conjunto de herra-

mientas para el desarrollo de aplicaciones web, como un sistema de gestión de bases de datos, un motor de plantillas y una capa de abstracción de URLs.

Se previsto que el acceso sea privado, por lo que se ha habilitado una opción para logearse.

Figura 8.3: Login



Para el almacenamiento de los datos, se utiliza una Base de Datos Relacional SQLite.

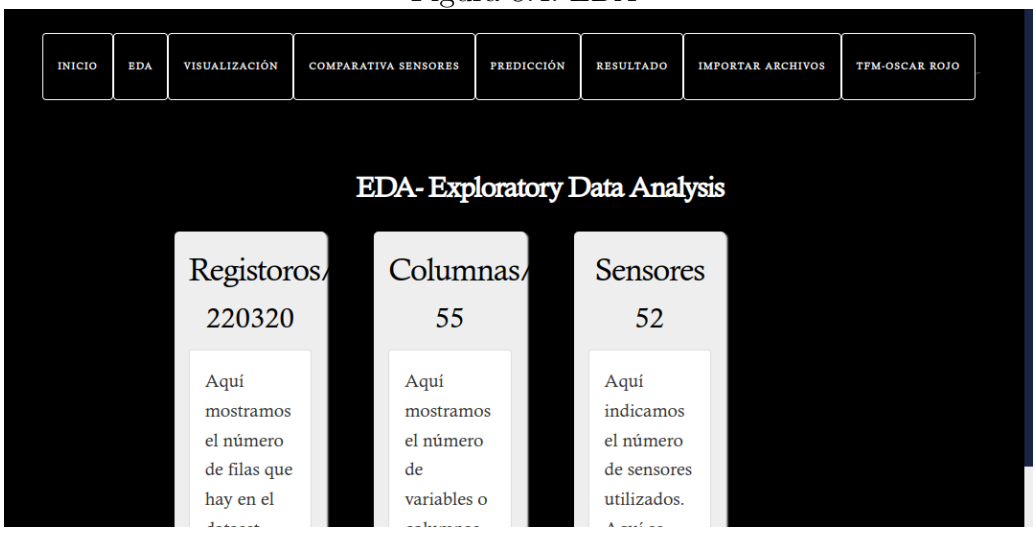
8.2.2. Apartados de la Aplicación-Web

Pasamos a describir brevemente los diferentes apartados de la aplicación web.

8.2.2.1. EDA - Exploratory Data Analysis

En este apartado tenemos un pequeño panel de control donde nos indica el número de registros, parámetros, Missing value, etc, del dataset así como las gráficas expuestas anteriormente : Distribucion de clases (Figura 7.1) en página 58, Mapa de Calor de Missing Values (Figura 7.2) en página 59 y Matriz de correlaciones (Figura 7.53) en página 75. Tambien se muestra en tabla html el Resumen dataset (Figura 7.3) en página 57.

Figura 8.4: EDA



8.2.2.2. Visualización

En este enlace mostramos las gráficas de los sensores mostrados en el apartado 7.5.3 Visualización de sensores

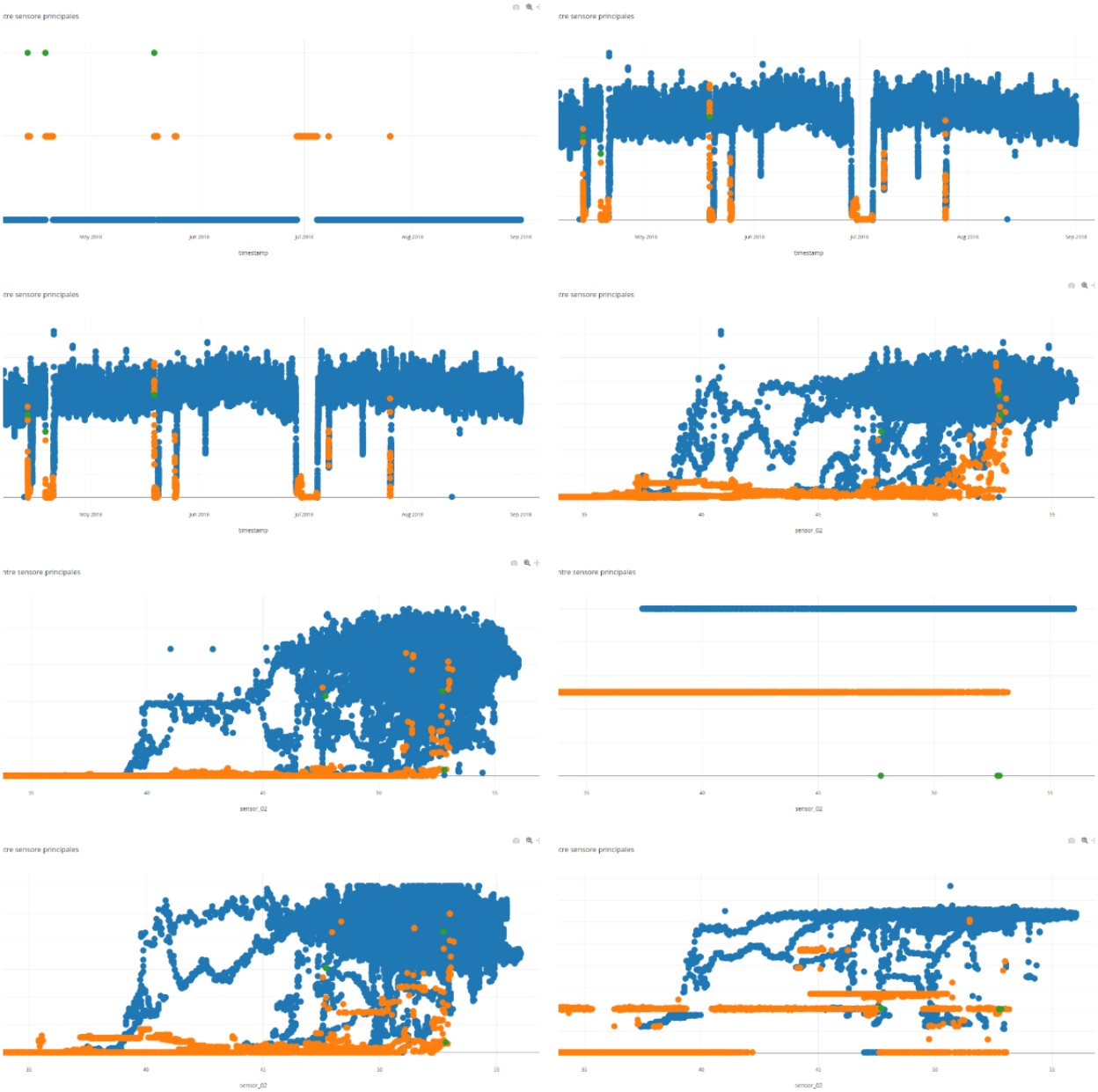
Figura 8.5: Visualización

Se muestran las gráficas sensor_01, sensor_02, sensor_03, sensor_06, sensor_07, etc.

8.2.2.3. Comparativa de Sensores

Para este apartado utilizamos la biblioteca plotly, que nos permite realizar gráficos interactivo. El tipo de gráfico realizado es Scatter, con el fin de comparar dos registros simultaneamente. En este caso por los tiempos de carga y la gran necesidad de recursos que necesita el ordenador y AWS-EC2, hemos seleccionado los sensores destinados a realizar el modelo de predicción: sensor_02, sensor_04, sensor_05, sensor_10, sensor_11 y sensor_12 junto con stat, el output "machine_status y el las fechas de registros.

Figura 8.6: ScatterPlot by Plotly



8.2.2.4. Predicción

Para este apartado hemos generado unos cuadros de imput de los valores de cada sensor e informamos en dichos input de la media, el valor mínimo y máximo de los sensores con el fin de aportar un valor adicional al input que inserten

Figura 8.7: Input predicción

8.2.2.5. Resultado

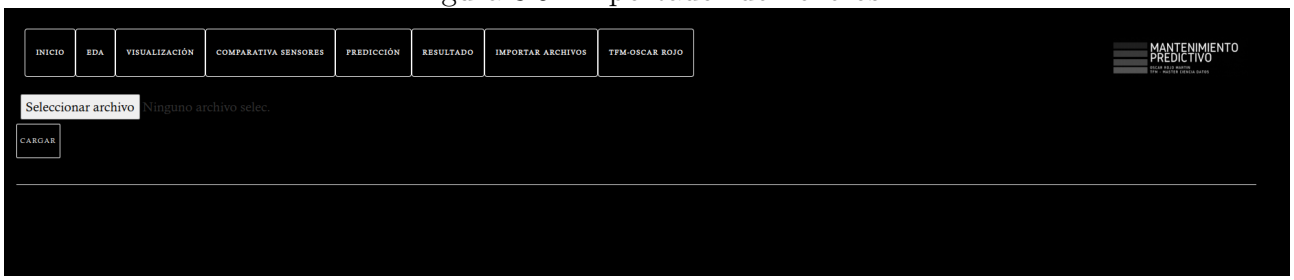
Este último aportado, es el resultado del calculo de las predicción según el modelo utilizado. Mostramos los datos de los input introducidos y el resultado correspondiente en una tabla... (pendiente de terminar)

Figura 8.8: Tabla de Resultados

8.2.2.6. Importar archivos

Añadimos un apartado para la importación de fichero en caso en que se actualicen los datos recogidos

Figura 8.9: Importador de ficheros



8.2.2.7. TFM-Oscar Rojo

Y por último habilitamos un link en el que se descarga el TFM en formato pdf en una pestaña nueva del navegador.

8.2.3. Puesta en Producción de la aplicación

Ver el README del repositorio:

https://github.com/zumaiaUOC/TFM_dashboard

Allí se encontrará toda la información para darse de alta en AWS, clonar el repositorio, Generar y configurar la instancia de la MV. Así como información para utilizar varios termanales-shell mediante ScreenRC

Capítulo 9

Bibliografía

262588213843476. (n.d.). Hosting a Django application on AWS EC2 (running AMI Linux) - steps involved. Gist. Retrieved from https://gist.github.com/tanuka72/79ae58b16be4ac3fafe0

An Introduction to Predictive Maintenance. (2002). doi:10.1016/B978-0-7506-7531-4.X5000-3

Bartodziej, C. J. (2017). The concept Industry 4.0. In C. J. Bartodziej (Ed.), The Concept Industry 4.0 : An Empirical Analysis of Technologies and Applications in Production Logistics (pp. 27–50). doi:10.1007/978-3-658-16502-4_3

Berry, M. W., Mohamed, A., & Yap, B. W. (2019). Supervised and Unsupervised Learning for Data Science (1st ed.). Springer Publishing Company, Incorporated.

Busche, R. (2022, July). Thesis Template. Retrieved from https://github.com/JarnoRFB/thesis-template

Cachada, A., Barbosa, J., Leitño, P., Gcraldcs, C. A. S., Deusdado, L., Costa, J., ... Romero, L. (2018, September). Maintenance 4.0: Intelligent and Predictive Maintenance System Architecture. 2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA), 1, 139–146. doi:10.1109/ETFA.2018.8502489

ChatGPT: Optimizing Language Models for Dialogue. (2022, November). OpenAI. Retrieved from https://openai.com/blog/chatgpt/

Coandă, P., Avram, M., & Constantin, V. (2020). A state of the art of predictive maintenance techniques. IOP Conference Series: Materials Science and Engineering, 997(1), 012039. doi:10.1088/1757-899X/997/1/012039

Curcurù, G., Galante, G., & Lombardo, A. (2010). A predictive maintenance policy with imperfect monitoring. Reliability Engineering & System Safety, 95(9), 989–997. doi:10.1016/j.ress.2010.04.010

Dieulle, L., Berenguer, C., Grall, A., & Roussignol, M. (2001, January). Continuous time predictive maintenance scheduling for a deteriorating system. Annual Reliability and Maintainability Symposium. 2001 Proceedings. International Symposium on Product Quality and Integrity (Cat. No.01CH37179), 150–155. doi:10.1109/RAMS.2001.902458

Eckerson, W. W., Hanlon, N., & Barquin, R. (2000). DIRECTOR OF EDUCATION AND RESEARCH. 5(4).

Ghavami, P. K. (2019). Big data analytics methods: analytics techniques in data mining, deep learning and natural language processing. De Gruyter.

Hashemian, H. M., & Bean, W. C. (2011). State-of-the-Art Predictive Maintenance Techniques*. IEEE Transactions on Instrumentation and Measurement, 60(10), 3480–3492. doi:10.1109/TIM.2009.2036347

Ji, C., Li, Y. U., Qiu, D., Jin, Y., Xu, Y., Awada, U., ... Qu, W. (2013). Big data processing: Big challenges. Journal of Interconnection Networks, 13. doi:10.1142/S0219265912500090

Kotsiantis, S. B. (n.d.). Supervised Machine Learning: A Review of Classification Techniques.

Las tecnologías de la Industria 4.0 mejoran la economía circular. (n.d.). UPV/EHU. Retrieved from https://www.ehu.eus/es/-/industria-4-0-economia-circular

Li, H., Parikh, D., He, Q., Qian, B., Li, Z., Fang, D., & Hampapur, A. (2014). Improving rail network velocity: A machine learning approach to predictive maintenance. Transportation Research Part C: Emerging Technologies, 45, 17–26. doi:10.1016/j.trc.2014.04.013

Marr, B. (n.d.). 27 Incredible Examples Of AI And Machine Learning In Practice. Forbes. Retrieved from https://www.forbes.com/sites/bernardmarr/2018/04/30/27-incredible-examples-of-ai-and-machine-learning-in-practice/

Martinez, I., Viles, E., & Olaizola, I. G. (2021). Data Science Methodologies: Current Challenges and Future Approaches. doi:10.48550/ARXIV.2106.07287

Mokhtari, S., Abbaspour, A., Yen, K. K., & Sargolzaei, A. (2021). A Machine Learning Approach for Anomaly Detection in Industrial Control Systems Based on Measurement Data. Electronics, 10(4), 407. doi:10.3390/electronics10040407

Mongkolchaichana, S., Phruksaphanrat, B., & Panjavongroj, S. (2021). Prioritization of Sustainable Supply Chain Management Practices in an Automotive Elastomer Manufacturer in Thailand Sustainable supply chain management Multiple attribute decision making Logarithmic fuzzy preference programming Prioritization A case study. Advances in Science Technology and Engineering Systems Journal, 6, 1079–1090.

Morais, J., Pires, Y., Cardoso, C., & Klautau, A. (2009). An Overview of Data Mining Techniques Applied to Power Systems. doi:10.5772/6463

Nantasenamat, C. (2021, June). How to Build a Machine Learning Model. Medium. Retrieved from https://towardsdatascience.com/how-to-build-a-machine-learning-model-439ab8fb3fb1

Nasteski, V. (2017). An overview of the supervised machine learning methods. HORIZONS. B, 4, 51–62. doi:10.20544/HORIZONS.B.04.1.17.P05

Olson, D., & Delen, D. (2008). Advanced Data Mining Techniques. doi:10.1007/978-3-540-76917-0

Palacio-Niño, J.-O., & Berzal, F. (2019, May). Evaluation Metrics for Unsupervised Learning Algorithms. doi:10.48550/arXiv.1905.05667

Park, C., Moon, D. H., Do, N., & Bae, S. (2016). A predictive maintenance approach based on real-time internal parameter monitoring. The International Journal of Advanced Manufacturing Technology, 85. doi:10.1007/s00170-015-7981-6

Petrasch, R., & Hentschke, R. (2016). Process Modeling for Industry 4.0 Applications Towards an Industry 4.0 Process Modeling Language and Method. doi:10.1109/JCSSE.2016.7748885

Practical Machinery Vibration Analysis and Predictive Maintenance - 1st Edition. (n.d.). Retrieved from https://www.elsevier.com/books/practical-machinery-vibration-analysis-and-predictive-maintenance/scheffer/978-0-7506-6275-8

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., ... Lin, C.-T. (2017). A review of clustering techniques and developments. Neurocomputing, 267, 664–681. doi:10.1016/j.neucom.2017.06.053

Seel, N. M. (Ed.). (2012). Encyclopedia of the sciences of learning. New York: Springer.

Sehra, S., Flores, D., & Montanez, G. (2021). Undecidability of Underfitting in Learning Algorithms.

Sunilkumar. (2022, January). Predictive Maintenance of Pumps. Medium. Retrieved from https://medium.com/@suniaidvpr/predictive-maintenance-of-pumps-7c358f0efe68

Susto, G. A., Wan, J., Pampuri, S., Zanon, M., Johnston, A. B., O'Hara, P. G., & McLoone, S. (2014, August). An adaptive machine learning decision system for flexible predictive maintenance. 2014 IEEE International Conference on Automation Science and Engineering (CASE), 806–811. doi:10.1109/CoASE.2014.6899418

Theissler, A., Pérez-Velázquez, J., Kettelgerdes, M., & Elger, G. (2021). Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. Reliability Engineering & System Safety, 215, 107864. doi:10.1016/j.ress.2021.107864

thesis.pdf. (n.d.). Google Docs. Retrieved from https://drive.google.com/file/d/1JCwNjQNW-FDqSqdSn5DvGlCzF_8CkJTg/edit?usp=embed_facebook

Tunkelang, D. (2017, September). 10 Things Everyone Should Know About Machine Learning. Medium. Retrieved from https://dtunkelang.medium.com/10-things-everyone-should-know-about-machine-learning-15279c27ce96

Ullah, I., Yang, F., Khan, R., Liu, L., Yang, H., Gao, B., & Sun, K. (2017). Predictive Maintenance of Power Substation Equipment by Infrared Thermography Using a Machine-Learning Approach. Energies, 10(12), 1987. doi:10.3390/en10121987

UNE-EN 13306:2018 Mantenimiento. Terminología del mantenimiento. (n.d.). Retrieved from https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0060338

Vij, A., Vijendra, S., Jain, A., Bajaj, S., Bassi, A., & Sharma, A. (2020). IoT and Machine Learning Approaches for Automation of Farm Irrigation System. Procedia Computer Science, 167, 1250–1257. doi:10.1016/j.procs.2020.03.440

Water Pump Dataset Exploration – Ozan ŞAHİN. (n.d.). Retrieved from http://sahinozan.com/inc pump-dataset-exploration/

Zhu, X. (2008). Semi-Supervised Learning Literature Survey. Comput Sci, University of Wisconsin-Madison, 2.

Créditos/Copyright

Copyright © 2022 Oscar Rojo Martín.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

Se concede permiso para copiar, distribuir y/o modificar este documento bajo los términos de la Licencia de Documentación Libre de GNU, Versión 1.3 o cualquier versión posterior publicada por la Fundación para el Software Libre; sin secciones invariables, sin textos de portada y sin textos de contraportada. Se incluye una copia de la licencia en la sección titulada "Licencia de Documentación Libre GNU".

GNU Free Documentation License.

FICHA DEL TRABAJO FINAL

Título del trabajo:	Mantenimiento Predictivo
Nombre del autor:	Oscar Rojo Martín
Nombre del colaborador/a docente:	Lorena Polo Navarro
Nombre del PRA:	Antonio Lozano Bagén
Fecha de entrega (mm/aaaa):	01/2023
Titulación o programa:	Máster Ciencia de Datos
Área del Trabajo Final:	Área 4 - Data Science en el ámbito industrial
Idioma del trabajo:	Español
Palabras clave	Industria, Sensores, Predicción

Dedicatoria/Cita

Este trabajo se lo dedico a todos los empresarios que desde principios del siglo pasado han contribuido a este territorio a ser uno de los más importantes de España en materia de riqueza, educación y bienestar.

"**La educación es el arma más poderosa que puedes usar para cambiar el mundo.**"
- Nelsón Mandela (1918-2013)

Agradecimientos

Te elijo para ser mi esposa, para hacer la vida de la mano, codo con codo. Te elijo para amar, con todo mi ser, incondicionalmente. Te elijo al principio y al final de cada día. Y te elegiría en cien vidas, en cien mundos, en cualquier versión de la realidad que te encontrara, y te elegiría.

¡Gracias **Lei**!

Y como no, mis otras 3 chicas: "**Arritxu, Emma y Maya**"

Abstract

The importance of industrial activity is one of the main characteristics of the economy of Gipuzkoa, a densely populated territory with a high level of economic and social development and a high level of income per capita. Industrial activity is highly specialized in sectors that work with metal, such as iron and steel, parts and components, metal products, capital goods, machine tools, and transport material. All these sectors have a wide range of machinery. This work aims at processing, visualizing, and generating a predictive model from the available multiple sensors. Machine Learning models will be applied and the results will be displayed in a Dashboard for consultation.

Keywords: Industry, Sensors, Prediction, Master Final Project

Resumen

La importancia de la actividad industrial es una de las características principales de la economía de Gipuzkoa, un territorio densamente poblado, con un alto nivel de desarrollo económico y social y un nivel de renta per capital. La actividad industrial se encuentra muy especializada en sectores que trabajan el metal, como son siderometalurgia, piezas y componentes, productos metálicos, bienes de equipo, máquina herramienta y material de transporte. Todos estos sectores cuentan con una amplia gama de maquinaria. Este trabajo está destinado a tratar, visualizar y generar un modelo predictivo provenientes de los múltiples sensores con los que cuentan actualmente. Se intentarán aplicar modelos de Machine Learning y los resultados se mostrarán en un Panel de Mandos para su consulta.

Palabras clave: Industria, Sensores, Predicción, Trabajo de Final de Máster

Índice de figuras

Índice de cuadros