Universidad de Valladolid

Escuela de Ingeniería Informática

# Desarrollo de modelos predictivos en un entorno de fabricación industrial

Trabajo Fin de Grado de Informática


Autor: **Miguel Martín Mateos**

Tutores: **Álvaro García García (CIDAUT)**
**Valentín Cardeñoso Payo (UVa)**


Universidad de Valladolid

Cidaut

# Agradecimientos

*" En primera instancia, agradecer a Álvaro, tutor de empresa, por brindarme la posibilidad de realizar este TFG en colaboración con fundación CIDAUT. Además, agradecer su esfuerzo semana a semana por interesarse por cómo me encontraba y cómo llevaba el trabajo en estas circunstancias especiales provocadas por el Covid19 en la que la realización de videoconferencias programadas durante estos meses me ha ayudado a poder comentar los trabajos actualizados".*

*" En segunda instancia, agradecer también a Valentín, tutor de la Universidad, por confiar en el proyecto propuesto con Álvaro allá por el mes de Noviembre. Además, destacar su ayuda en llevar el trabajo realizado a un plano más académico y científico así como variedad de consejos sobre cómo mejorar aspectos de la memoria. "*

*" Agradecer a compañeros de carrera, el soporte y la ayuda durante estos meses atípicos en los que nos hemos ayudado entre nosotros en diferentes cuestiones ya sean teóricas o prácticas con la consecución de este trabajo fin de grado".*

*" Por último, agradecer a mi familia y mi pareja la ayuda ofrecida durante estos meses en la que ha habido momentos más complicados y de estrés, en la que sin su ayuda no hubiera sido posible la consecución de este objetivo".*

# Índice general

# Índice de figuras

# Índice de cuadros

# Resumen

Este trabajo fin de grado aborda el estudio y caracterización de condiciones de funcionamiento de una máquina industrial a partir de modos de operación que pueden ser clasificados y predichos de forma automática empleando técnicas de *Machine Learning*.

En particular se va a tratar con uno de los comportamientos más característicos de las máquinas industriales como es el arranque, el cual es clave en la puesta en marcha del proceso de fabricación. Caracterizar y segmentar las fases de operación del proceso de arranque , será importante para analizar el comportamiento de la máquina y poder detectar anomalías o incidencias durante el transcurso del mismo. La no tenencia de información a priori con valores etiquetados para comparar si la máquina se comporta correctamente durante un arranque, llevará a utilizar técnicas de aprendizaje no supervisado. La segmentación a través del clustering permitirá clasificar los distintos arranques en grupos de acuerdo a sus similitudes en cuánto a duración y consumo.

La generación de los patrones más característicos entre los modelos de arranque extraídos del clustering ayudará a determinar un modelo de arranque inicial. Con esta información, se podrá comprobar la existencia de anomalías comparando datos de test (nuevos arranques) con el comportamiento del arranque de base, segundo a segundo.

# Abstract

This Final Degree Project deals with the study and characterization of the operating conditions of an industrial machine based on operating modes that can be automatically classified and predicted using Machine Learning techniques.

In particular, we will deal with one of the most characteristic behaviours of industrial machines, the start-up, which is key in the start of the manufacturing process.Characterizing and segmenting the operating phases of the start-up process will be important to analyze the behavior of the machine and to be able to detect anomalies or incidents during the course of the process. Not having a priori information with labeled values to compare if the machine behaves correctly during a start, it will lead to the use of unsupervised learning techniques. The segmentation through clustering will allow the classification of the different startups into groups according to their similarities in terms of duration and energy consumption.

Generating the most characteristic patterns among the boot models extracted from the clustering will help to determine an initial boot model. With this information, the existence of anomalies can be verified by comparing test data (new starts) with the behaviour of the base start, second by second.

# Capítulo 1

# Introducción

## 1.1 Motivación

La eficiencia en los procesos de fabricación es algo fundamental dentro de las industrias. La cuarta revolución industrial junto a las tecnologías digitales ha supuesto un momento de cambio que favorece la transformación de las industrias y modelos de fabricación existentes a partir de los nuevos principios de la 'Industria 4.0'[37].

La digitalización de las fábricas acompañada de las nuevas tecnologías habilitadoras (de la Industria 4.0) introduce un escenario donde gobiernos, investigadores y expertos en nuevas tecnologías deben trabajar juntos para afrontar este proceso de cambio.

Debido a la necesidad de minimizar o solucionar problemas en sistemas complejos se está generando mucha demanda de profesionales de 'ingeniería de datos'. Este TFG basado en el desarrollo de modelos predictivos en un entorno de fabricación industrial reúne varios conocimientos adquiridos durante mi etapa académica realizando el doble grado Indat. Desde el análisis de series temporales, aplicar modelos de clustering, realizar un análisis descriptivo/exploratorio de los datos con diferentes estacionalidades hasta la aplicación de modelos de machine learning o aplicación de minería de datos. Se corresponde con algunos de los métodos que he utilizado durante el transcurso del TFG y que caracteriza el contenido de éste con los conocimientos adquiridos en la formación universitaria. La unión de ambas carreras universitarias da un abanico de conocimientos que me permiten la realización de este trabajo.

Uno de los aspectos más influyentes que me llevaron a realizar un trabajo conjunto con una empresa, es el poder trabajar con datos reales, lo cual supone sumar experiencia para el futuro laboral como complemento a los conocimientos adquiridos en el proceso académico. Además, es interesante conocer de primera mano como 'la Ciencia de Datos' se encuentra en muchas aplicaciones, en este caso en el sector industrial y como poder actuar para prevenir fallos, realizar mejoras o predecir resultados con la ayuda del análisis de datos.

A nivel práctico es interesante estar trabajando 'diariamente' en un lenguaje específico como Python en este caso y adquirir mayor habilidad a la hora de programar. También para tener mayor conocimiento para la limpieza o el procesado de los datos como complemento a lo aprendido en la formación académica.

Desde un punto de vista ya más personal orientado a un 'analista de datos' procedente de la formación universitaria, mi objetivo se basaría en poder dar una explicación teórico/práctica a un conjunto de datos de partida sobre un problema concreto, analizarlos, caracterizarlos y clasificarlos.

Mas allá del mero objetivo de realizar el TFG, pongo un mayor énfasis en la situación de estar en una fase de aprendizaje, comenzar a investigar, analizar datos, cometer errores, realizar pruebas o afrontar nuevos retos diarios. Tras finalizar los estudios, este trabajo de investigación realizado en colaboración con una empresa, me ha ayudado a saber como actuar, investigar o trabajar con datos reales. En definitiva, ha ayudado a prepararse para el futuro laboral que al final es uno de los objetivos en la formación universitaria.

## 1.2 Objetivos

El objetivo general de este trabajo es el desarrollo de modelos predictivos sobre el comportamiento del arranque de una fresadora CNC( Control Numérico por Computadora), a partir de unos conjuntos de datos extraídos de sensores instalados en la máquina, que se ubica en un centro de Investigación.

La cantidad de datos que se generan actualmente en todos los sectores, y en concreto en el de la industria, está aumentando cada vez más, por lo que la caracterización y la clasificación automática mediante herramientas de aprendizaje, supone una ventaja competitiva que no se puede menospreciar.

En concreto, se trata de crear modelos que predigan si los arranques diarios de una fresadora se adecúan a un patrón de comportamiento habitual extraído de analizar una serie de arranques ó, por el contrario, existen ciertas anomalías en la serie que nos avisan que ha ocurrido una incidencia durante el arranque.

Uno de los objetivos iniciales más importantes es entender que información aportan los conjuntos de datos, o indicadores de proceso obtenidos de los sensores instalados en la máquina fresadora CNC. Con ello, se podrá lograr interpretar los resultados de una manera más efectiva. Para el arranque de la máquina será necesario conocer las variables de fase de corriente así como su tendencia. Con más detalle se describe en el capítulo 4.

Tanto el preprocesado como la limpieza de los datos serán importantes, así como la aplicación de un modelo de aprendizaje, ya que este mismo dependerá en gran medida de las transformaciones realizadas anteriormente. Con ayuda de transformaciones como la normalización nos permitirá detectar patrones en el arranque que en una etapa posterior servirán para la verificación de los modelos de aprendizaje.

Por otra parte, antes de aplicar un modelo de aprendizaje, es necesario comprender como se comporta la serie temporal a utilizar, que en este caso, se corresponde con las distintas fases de operación que caracterizan el arranque de la máquina. El objetivo es determinar la segmentación de dichas fases, es decir, la explicación de cuando empieza y termina una fase para poder después caracterizarlas adecuadamente. Con esta información, se plantea una hipótesis basada en agrupar los distintos arranques en base a características similares, para después, con la aplicación de un

algoritmo de aprendizaje, sea posible validar dicha hipótesis.

Finalmente, para llegar a desarrollar un modelo predictivo es muy útil la aplicación de herramientas de aprendizaje automático, en concreto en este caso, aprendizaje no supervisado. El objetivo consistirá en clasificar los distintos arranques de una fresadora en base a su consumo y su tendencia en el tiempo.

Para ello se utilizará una metodología de aprendizaje en la que no tenemos información a priori sobre como clasificar una observación o, en este caso, el arranque. Gracias a la aplicación de modelos de clustering se podrá clasificar de acuerdo a medidas como las 'distancias' entre observaciones para detectar si un arranque es anómalo o pertenece a una serie de arranques característicos determinados por un modelo concreto.

Con la obtención de modelos de arranque característicos tras la aplicación de modelos de Machine Learning, se pretende comprobar con nuevos arranques de la fresadora, en que medida se adecúan a los patrones de comportamiento asociados a los arranques. Con esto, se pretende simular el arranque segundo a segundo de la fresadora en tiempo real, monitorizando si se comporta de acuerdo al modelo implementado o si hay ligeras desviaciones en cada momento, pudiendo determinar la existencia de anomalías.

## 1.3 Metodología

### 1.3.1 Problemática Existente

En la actualidad, cada vez más industrias tienen como objetivo obtener una reducción de costes y un mantenimiento de los activos más eficiente. Esto tiene su origen en la denominada industria 4.0 [37] de la que hablaremos en el siguiente capítulo con mayor detalle. Dentro del sector industrial, las líneas de producción están formadas por máquinas que realizan diferentes procesos con el objetivo de que el producto final sea satisfactorio. Obviamente, cuánto mayor sea la eficiencia en el mantenimiento de los activos, mayor repercusión tendrá en el resultado y en el consumo. Es una necesidad que cada vez más fábricas industriales demandan.

Debido a esta creciente demanda de obtener los mejores productos, el mejor mantenimiento, a la vez que se minimizan costes, la investigación en técnicas del desarrollo de modelos predictivos está totalmente alineada en este objetivo.
En este caso concreto, como es la detección de anomalías en máquinas industriales, se proporcionan herramientas que ayudan a prevenir dichos fallos, obtener mejores resultados y/o disminuir el consumo.

Haciendo referencia a la Conferencia de Directores y Decanos de Ingenieria Informatica de la Univesidad de Deusto[37], se puede comprobar que con el paso del tiempo se demandan más profesionales en 'ciencia de datos' con conocimientos en las tecnologías digitales(TIC), lo cual es la base de la nueva revolución industrial (Industria 4.0).

### 1.3.2 Implementación

El punto de partida pretende la identificación y modelado de un 'escenario de aprendizaje sobre un sistema industrial' que permita registrar y posteriormente reproducir una serie de condiciones

operativas en el arranque de una máquina fresadora CNC, para la implementación de un modelo virtual o Gemelo Digital. Con este modelo será más sencillo la creación, prueba y validación de los modelos de aprendizaje automático y su interacción con el sistema físico real [16].

Ese escenario de aprendizaje trabajará en la detección de anomalías en máquinas industriales,concretamente de una fresadora con control numérico por computadora de la marca *Nicolás Correa CF-20* ubicada en un taller de mecanizado perteneciente a un Centro de I+D(Fundación Cidaut).

En este caso concreto,entendemos por anomalía a las observaciones que no siguen una tendencia habitual en la serie. El análisis de dichas observaciones en relación a valores de referencia nos indicará si esa anomalía supone un cambio de tendencia positivo o negativo.

Para ello se detectará la tendencia de una observación en relación a la anterior y la posterior con el objetivo de detectar la tendencia de cada dato de la serie temporal. Detectar incrementos y decrementos nos permitirá posteriormente detectar las distintas fases de operación, diferenciadas en cada arranque. Con la diferenciación de las distintas fases de operación, se podrá detectar de manera individual como se comporta cada serie de arranque diaria respecto a cada fase, donde se utilizarán indicadores de referencia basados en el consumo eléctrico medio,máximo o la duración del tiempo de cada fase de operación de la máquina.

El siguiente paso consiste en parametrizar e implementar el framework o entorno de trabajo para la 'Ciencia de Datos' con aquellas variables e indicadores que nos ayuden a validar el modelo virtual del sistema de fabricación industrial, que permita posteriormente realizar mejoras sin necesidad de intervenir en el sistema físico a la hora de aplicar las técnicas de análisis de datos y control predictivo.

Con la obtención de las variables más influyentes, se aplica un modelo de aprendizaje no supervisado, concretamente un modelo de clustering para la extracción de los diferentes arranques característicos o la detección de un arranque no característico debido por anomalías producidas en las fases de operación.

Con el resultado de este modelo base se pretenderá extrapolar estos resultados a otros conjuntos de datos (lineas de trabajo futuras en capítulo 6).

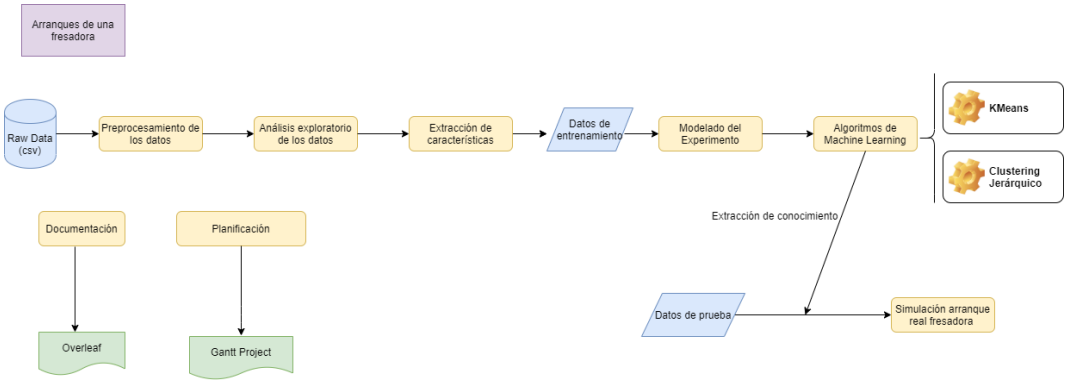



Figura 1.1: Boceto Fases Implementación

A continuación se expone de manera breve las herramientas que se han utlizado para poner en marcha este TFG y que se han expuesto de manera esquemática en la Figura [1.1]:

- Conjuntos de datos de Diciembre de 2019 y Enero,Febrero y Marzo de 2020 sobre los indicadores del proceso CNC de una fresadora 3.1a
- Extracción datos de entrenamiento y de prueba del preprocesamiento realizado al raw Data.
- Para el análisis descriptivo de los datos se ha hecho uso de la herramienta de visualización gráfica **Tableau**.
- Utilización del lenguaje de programación Python en todas las fases del trabajo.
- Utilización del lenguaje de R como complemento a la utilización de Python en la extracción de características.
- Aplicación de modelos de machine learning de aprendizaje no supervisado: KMeans y Clustering jerárquico.
- **Overleaf** para realizar el informe del trabajo en latex.
- Se usará la herramienta **GanttProject** para realizar la planificación del trabajo.

Por último, mencionar que toda la documentación( archivos y conjuntos de datos) se encuentran en el repositorio : https://gitlab.inf.uva.es/mimarti/tfg-desarollo-de-modelos-predictivos-en-un-entorno-de-fabricaci-n-industrial.git.

## 1.4 Estructura del documento

Este TFG trabaja en el desarrollo de un modelo predictivo en un entorno de fabricación industrial. A continuación, de manera breve, se resume de lo que trata cada uno de los capítulos:

*El capítulo 1* introduce la motivación del TFG, así como los objetivos planteados y su puesta en marcha, exponiendo las distintas herramientas utilizadas para llevarlo a cabo. Es importante detallar el motivo por el que se ha llevado a realizar este TFG así como cuales son los resultados que se quieren obtener tras terminar dicha investigación. Además se detallará la planificación realizada anteriormente y posteriormente de la realización del trabajo mediante diagramas de Gantt.

*El capítulo 2* explica de manera concisa el cambio que supone la aplicación de las nuevas tecnologías en el sector industrial. Como en la realización de cualquier trabajo o investigación es necesario familiarizarse en el contexto de trabajo, en este caso el de la Industria. Inicialmente, se introducirá el concepto de Industria 4.0 así como las tecnologías llevadas a cabo en esta nueva revolución industrial. También se detallará los tipos de mantenimiento industrial, enfocándonos especialmente en el mantenimiento predictivo. Es importante destacar como veremos en *la sección 2.3*, la digitalización y el retrofitting, dos conceptos muy importantes en la convergencia de las tecnologías de la información y tecnologías operacionales.
Una de las secciones más importantes y con las que más se ha trabajado en el transcurso del TFG hace referencia a los modelos de aprendizaje automático. Por ello se expone en detalle, valorando como su aplicación a la industria puede ofrecer muchos beneficios,desde reducir costes a la prevención de fallos. De manera breve se detallará qué es un sistema de fabricación por control

numérico , el cual está incorporado en la fresadora de la que obtendremos los datos para analizar. Por último, se describirá el concepto de gemelo digital, encargado de reproducir un sistema virtual que contiene toda la información sobre un sistema físico real.

*En el capítulo 3* se describe la procedencia de los datos a utilizar, así como las características técnicas de la máquina utilizada. Será importante destacar los indicadores clave de rendimiento con los que se obtiene la información para analizar, así como el tipo de datos de cada uno de ellos y a qué representa. Por último se concretará en la *sección 3.2* un resumen sobre la evolución global del trabajo con sus distintas fases del proyecto.

*El capítulo 4*, comienza abordando un proceso de limpieza, procesamiento y descripción de los datos. El objetivo es la transformación de la información de tal forma que se logre obtener un dataset adecuado para su tratamiento. Además, se realiza un análisis descriptivo que nos ayude a entender de una mejor manera los datos que estamos tratando así como qué representan.
*En las secciones 4.3 y 4.4* se comenzará detallando el modo en el que se han detectado las diferentes fases de operación de las que se compone un arranque. Para ello es necesario obtener el comportamiento del consumo en cada segundo para ver los incrementos o decrecimientos de consumo destacables y con ello la segmentación de las distintas fases de operación del arranque. Tras la diferenciación de las distintas fases se realizará una serie de normalizaciones a unas variables representadas por los estadísticos máximo y media del consumo y a la duración de cada fase del arranque. Con este procedimiento y los valores estadísticos comentados se podrá determinar una hipótesis de clasificación en grupos de arranques que se contrastará posteriormente con la aplicación de un algoritmo de aprendizaje no supervisado. *Las secciones 4.5 y 4.6* introducen el concepto del sistema de aprendizaje utilizado, en concreto una metodología no supervisada, en la que no existen etiquetas para describir si una observación pertenece o no a una categoría en cuestión. Por último se describirá la implementación de los algoritmos utilizados así como los resultados obtenidos.

*El capítulo 5* tratará de extraer los patrones más característicos de las fases de los arranques pudiendo caracterizar un nuevo modelo de arranque como agrupaciones comunes a varios de los modelos de arranque, extraídos de los algoritmos de aprendizaje no supervisado. En resumen, caracterizar el modelo de aprendizaje resultante a otro nivel, pudiendo caracterizar un modelo de comportamiento en base a ciertos patrones que ocurren frecuentemente. Por último, simular un arranque de una fresadora en tiempo real con un nuevo conjunto de datos de prueba con objetivo de determinar en qué medida se repite el tipo de arranque más característico y encontrar posibles anomalías en los distintos arranques.

Por último, *en el capítulo 6* se detallan las conclusiones obtenidas tras realizar el modelo predictivo así como la línea a seguir si se tuviera un tiempo mayor para realizar el trabajo.

## 1.5 Planificación

Este proyecto tiene como fecha de inicio el mes de Febrero de 2020, habiéndose realizado la siguiente planificación de tareas y fechas asociadas **a priori**(en la que las semanas incluyen sábados y

domingos). En el diagrama de Gantt se explican subtareas asociadas a las fases que se enumeran a continuación:

1. Fase iniciación*(5 Semanas)*
2. Fase Procesamiento de los datos *(7 Semanas)*
3. Fase Clasificación*(6 Semanas)*
4. Fase Prueba *(2 Semanas)*
5. Fases Revisión y Presentación (*3 Semanas*)

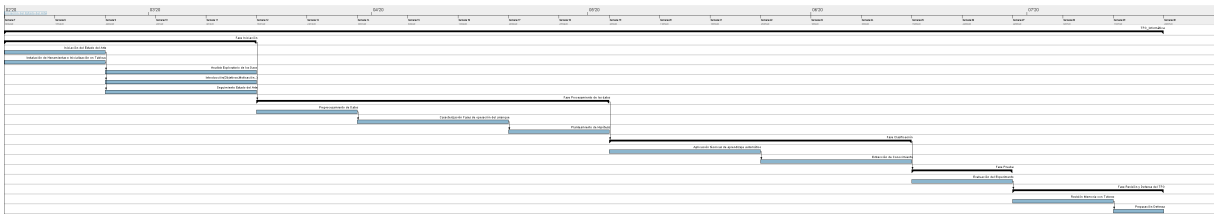


Figura 1.2: Planificación Diagrama Gantt A priori

Sin embargo, tras haber finalizado con el trabajo fin de grado se ha estructurado diferente las distintas fases del trabajo fin de grado. La dependencia del tiempo para estudiar las asignatura cursadas durante el transcurso del cuatrimestre así como la realización a la par del TFG de Estadística ha provocado ciertas discrepancias. Además, antes de la realización del TFG es difícil saber que tiempos puede llevar cada fase del TFG. Además, habrá alguna fase del proyecto que no se añadió en la planificación inicial. El análisis del trabajo realizado fue documentándose en Gitlab, cada una o dos semanas, actualizando el trabajo realizado y con fechas de inicio y de fin que son añadidas en el diagrama de Gantt [1.3].

De manera esquemática se organizó este trabajo fin de grado del siguiente modo, con un inicio y fin similar al planificado a priori. De nuevo, existen subtareas en el Diagrama de Gantt para cada fase de las enumeradas:

1. Fase iniciación (*3 Semanas*)
2. Fase Estado del Arte/Análisis Exploratorio de los datos(*3 Semanas*)
3. Fase Tratamiento de los datos(*3 Semanas*)
4. Fase Normalización de los Datos y Clustering(*6 Semanas*)
5. Fase Extracción de patrones de comportamiento y Simulación(*5 Semanas*)
6. Fases Revisión y Presentación (*3 Semanas*)

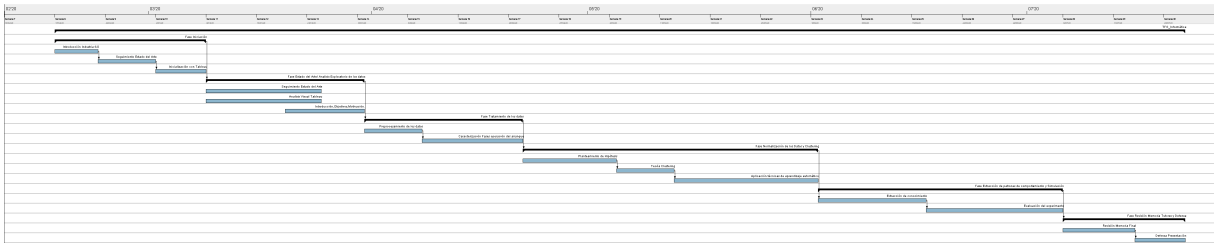


Figura 1.3: Planificación Diagrama Gantt A posteriori

Existe alguna diferencia con el contenido de las diferentes fases, como por ejemplo en la planificación a priori, el planteamiento de Hipótesis se encuentra dentro de la **Fase de Preprocesamiento**, mientras que en la planificación a posteriori, el planteamiento de Hipótesis se ha incluido en la **Fase de Normalización de los Datos y Clustering**.

Sin embargo, los tiempos estimados para cada subfase del proyecto son similares, planificando 1 semana más o menos en cuánto a la planificación inicial. La inclusión de procedimientos como *Teoría Clustering* o haber dedicado un mayor tiempo al aprendizaje con Tableau , ha llevado a diferenciar las planificaciones explicadas. También la primera parte sobre el seguimiento del Estado del Arte, fue más largo que el tiempo planificado inicialmente. A cambio, la fase de preprocesamiento de los datos( anterior a la caracterización Fases de operación) fue 1 semana más corta que el tiempo planificado inicialmente.

# Capítulo 2

# Estado del Arte

## 2.1 Industria 4.0

El origen del concepto 'Industria 4.0' aparece a principios de la década de 2010 al amparo del sector automovilístico y del Gobierno de Alemania. En la feria de Hannover de 2011, fue presentada la Estrategia de Alta Tecnología del ejecutivo alemán Henning Kagermann, en la que se describía una producción industrial, cuyos productos y máquinas estarían interconectados entre sí digitalmente. Este informe recogía por primera vez el concepto de Industria 4.0 para denominar el conjunto de acciones dirigidas a lograr la denominada fábrica inteligente [3].

De acuerdo con la *agencia de desarrollo económico de la República Federal de Alemania* [3]:

> *"La Industria 4.0 conecta tecnologías de producción de sistemas integrados y procesos de producción inteligente para allanar el camino hacia una nueva era tecnológica que transforma radicalmente la industria y las cadenas de valor de producción y los modelos comerciales."*

En definitiva, el concepto de Industria 4.0 puede resumirse como la convergencia entre las tecnologías digitales y las líneas de producción [37]. Esta revolución representa una integración entre el Internet de las Cosas (IoT) y la aparición de nuevas tecnologías como son la robótica o la Inteligencia Artificial(IA),entre otras muchas.

Una de la bases generales de la Industria 4.0 se centra en la obtención y la utilización de datos en tiempo real para proporcionar unos resultados más competitivos. Junto a ello el Internet Industrial de las Cosas(IIoT),que con los datos y los objetos físicos, representa una fuente de valor que permite construir cadenas de suministro más inteligentes o procesos de fabricación mejorados [3].

En esta nueva revolución industrial, los procesos de fabricación se encuentran en una fase de *digitalización* producida por el avance de las Tecnologías de la Información, especialmente por el avance de la Informática [37]. Esto se traduce en una nueva forma de tener mayores beneficios en el sector industrial o reducir costes de producción.

De acuerdo a un artículo de primeros de 2020 publicado por el periódico *el País*[25] en colaboración con *Telefónica*, la aplicación de las últimas tecnologías a los procesos de fabricación provocan ahorros en costes de producción o mantenimiento del 20 % según el instituto Fraunhofer.

Exponen que, ante este nuevo paradigma, provocará que sectores industriales tradicionales deban 'renovarse' si quieren optimizar sus procesos.

### 2.1.1 Tecnologías habilitadoras y sistemas de información

La Industria 4.0 se fundamenta en la aplicación de tecnologías habilitadoras, algunas de las cuales se pueden ver en la Figura [2.1]. A continuación se hace una breve descripción de algunas de ellas.


Figura 2.1: Tecnologías habilitadoras industria 4.0. *Fuente* [38]

Una de las tecnologías más importantes es el **Cloud Computing o almacenamiento en la nube**, la cual permite el flujo de inmensas cantidades de datos y su almacenamiento deslocalizado, provocando una mayor flexibilidad. Además, gracias a esta flexibilidad, las empresas dejan de depender de infraestructuras locales para alojar los datos, que pasan a estar accesibles de forma distribuida. Por ello, muchas empresas hacen uso de estos servicios que permiten un fácil acceso a través de dispositivos [1].

De esta forma, actúa como un habilitador de la innovación en el entorno industrial, provocando una mayor agilidad, ahorro de costes al pagar por lo que se necesita realmente y acceso a las últimas tecnologías de forma mas sencilla y ágil [25].

Figura 2.2: Cloud Computing *Fuente* [7]

Una vez definido de manera general el significado de Cloud Computing, se quiere destacar que para la realización de este TFG se ha utilizado la tecnología Edge Computing, el que se describe a continuación.

Gartner definió el Edge Computing como una parte de una topología de computación distribuida en la que la información procesada es alojada en una ubicación local donde personas pueden producir o consumir esa información [34].

Esta tecnología permite que los datos producidos por los dispositivos del Internet de las Cosas se procesen más cerca de donde se crearon, en lugar de enviarlos a través de largos recorridos para que lleguen a centros de datos y nubes de computación". Esto es muy beneficios ya que permite analizar los datos casi en tiempo real, como es el caso de este TFG gracias a la colaboración con Fundación CIDAUT.

Imad Sousou, vice presidente del Grupo de software y servicios así como director del Centro de Tecnología Open Source de Intel Corporation, destacó en una entrevista cuatro razones por las cuales Edge computing ha sido exitoso [26] :

- Velocidad, que reduce la latencia porque los datos no tienen que viajar sobre la red a un centro de datos remoto o a través de la nube para ser procesados.
- Mejor seguridad de los datos
- Escalabilidad, que reduce las cargas de red y permite mayor crecimiento.
- Costes más bajos debido a la reducción de la cantidad de datos transferidos hacia una ubicación central para almacenarlos.

Otra de las tecnologías mas innovadoras es el **Big Data and Data Analytics**. En el caso de este TFG se enfoca a la definición de Big Data ya que en la sección [2.4] se entra en detalle sobre aspectos como la Inteligencia Artificial o el Machine Learning.Por otro lado, aunque se detallen las características del Big Data de manera resumida, en este trabajo de fin de grado se utiliza más bien Smart Data ya que el concepto de Big Data tiene un alcance más amplio.

Se puede hablar de Big Data cuando existe una combinación de las siguientes características:

- Gran volumen de todo tipo de datos estructurados y no estructurados.

- Procesamiento veloz para obtener la información en tiempo 'real'.
- Tolerancia a fallos para recuperar información sin mucho esfuerzo.
- Almacenamiento específico para dicho volumen en centros de procesamiento de datos (CPD).

En general un proyecto Big Data puede contener mucha variedad de fases, como puede ser almacenamiento( albergar gran cantidad de datos en un centro de procesamiento de datos (CPD)),visualización de datos para ver comportamiento de estos o el análisis mediante el uso de diferentes algoritmos [1].

Otro de los aspectos que se comenta brevemente en la sección anterior y que tiene una gran relevancia es **Internet de las Cosas(IoT)**. Fue en 2009 cuando Kevin Ashton, profesor del MIT, usó la expresión *Internet of Things* de forma pública por primera vez y desde entonces el crecimiento en base a este término ha sido exponencial.

*'Si tuviésemos ordenadores que fuesen capaces de saber todo lo que pudiese saberse de cualquier cosa –usando datos recolectados sin intervención humana– seríamos capaces de hacer seguimiento detallado de todo, y poder reducir de forma importante los costes y malos usos. Sabríamos cuándo las cosas necesitan ser reparadas, cambiadas o recuperadas, incluso si están frescas o pasadas de fecha. El Internet de las Cosas tiene el potencial de cambiar el mundo como ya lo hizo Internet. O incluso más' [39].*

Su argumentación, en efecto, no iba muy desencaminada,pero ¿cómo se puede definir actualmente su uso?
Se puede definir como la representación de objetos que son capaces de recibir instrucciones y emitir datos utilizando una conexión a Internet [1]. El IoT se considera como la tecnología base para la implementación de la Industria 4.0 debido a la optimización y automatización basada en los datos o la transformación digital [9]. Gracias a la conexión de objetos a través del IoT es posible capturar información de cualquier proceso y darle un valor con la Inteligencia Artificial.

Otras tecnologías no tan relacionadas con este trabajo pero también influyentes en la industria son [1]:

- Fabricación aditiva o impresión 3D: Se puede definir como una técnica que se lleva a cabo para crear objetos físicos a partir de su representación en formato CAD,con la superposición de capas impresas utilizando diversos materiales.
- Ciberseguridad: Contempla la protección de la información contenida en un dispositivo a través del tratamiento de las diferentes amenazas que suponen un riesgo. Algunas de ellas pueden ser un virus( que altera el funcionamiento del PC sin consentimiento), fuga de datos(datos confidenciales que van a terceros), software malicioso, contraseñas inseguras o existencia de fallos de seguridad en programas.
- Robótica colaborativa: Describe una nueva generación de robots que trabajan sin barreras de separación con los humanos para favorecer el proceso y aumentar la colaboración. Surge como complemento del operario en orden de conseguir una mayor productividad.

- Realidad Aumentada: Consiste en la unión del contenido digital con contenido físico para construir una realidad dual en tiempo real. Un caso de uso de aplicación en el entorno en el que se desarrolla este TFG se observa en la Figura [2.3]
- Gemelo Digital: Creación de modelos virtuales de cualquier proceso/activo a partir de un modelo físico real que permitirá monitorizar,representar y prevenir posibles problemas o probar funcionalidades sin riesgos. Con más detalle en la sección 2.6.


Figura 2.3: Realidad Aumentada en Mantenimiento Industrial *Fuente CIDAUT*

## 2.2 Tipos de Mantenimiento Industrial

De acuerdo a la Real Academia Española [11], entendemos por **mantenimiento** :

*'Conjunto de operaciones y cuidados necesarios para que instalaciones, edificios, industrias, etc., puedan seguir funcionando adecuadamente.*

Pero ... ¿como lo relacionamos con la industria?

El mantenimiento dentro de la industria ha sufrido una evolución importante llevada en gran medida por el desarrollo tecnológico de los equipos de control y medida. De manera simple se puede medir la evolución del mantenimiento en cuatro etapas:

- Correctivo
- Preventivo
- Predictivo
- Proactivo o prescriptivo

Inicialmente, el mantenimiento estaba asociado a trabajos para resolución de averías, junto a los pertinentes costes de reparación o costes derivados de la producción. Este tipo de mantenimiento se conoce como **correctivo** [10].

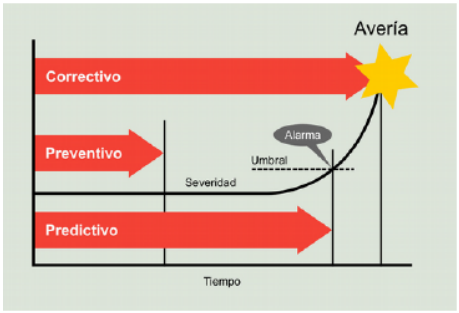
Pero la necesidad de reducir estos costes, llevaron a los técnicos de mantenimiento a programar revisiones periódicas con el objetivo de mantener las máquinas en el mejor estado posible y reducir su probabilidad de fallo. Este tipo de mantenimiento es el **preventivo**. No obstante, hay medidas que no se pueden cuantificar, como la reducción de los periodos de intervención sin que se introduzcan consecuencias perjudiciales para las máquinas [10].

Como consecuencia de estas medidas difíciles de ajustar y con ayuda del desarrollo tecnológico, se planteó un nuevo concepto: **el mantenimiento predictivo**. Gracias a este tipo de mantenimiento, es posible entrar en el terreno de la anticipación a la avería ya se dispone información acerca de como se comporta una máquina y como debería hacerlo en condiciones determinadas, permitiendo preveer que elementos puede fallar y tener una previsión estimada de cuándo. De este modo, es posible evitar costes recurrentes en averías [10].

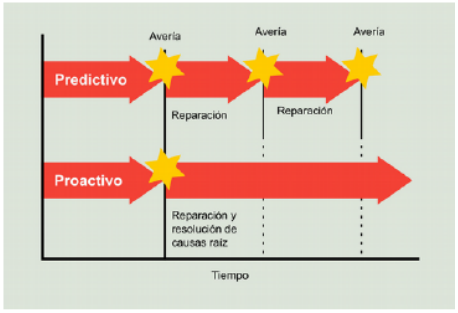
La crisis económica entre los años 2008 y 2015 ha sido un factor determinante para la adopción de nuevas estrategias de mantenimiento predictivo en las empresas. También ha supuesto la especialización en la oferta de mejores servicios por parte de compañías especializadas en mantenimiento predictivo a nivel nacional [12].

Como complemento a la evolución del mantenimiento predictivo se desarrolló el llamado **mantenimiento proactivo**. Este tipo de mantenimiento lleva a cabo el análisis de la causa raiz del suceso. Este analiza las causas raíz de la repetitibidad de la avería ó la causa raiz del fallo, lo cual no es realizado por el mantenimiento predictivo, ya que sólo determina cuando puede fallar algún componente [10].

Para ver de manera gráfica la explicación aportada entre los primeros tres tipos de mantenimiento [2.4a] y la diferencia entre predictivo y proactivo [2.4b] se adjuntan dos imágenes explicativas:




(a) Tipos de Mantenimiento *Fuente: [10]*




(b) Predictivo vs Proactivo *Fuente: [10]*

Figura 2.4: Mantenimiento Industrial

### 2.2.1 Mantenimiento Predictivo

La base del mantenimiento predictivo radica en la monitorización de los equipos, evaluando los parámetros registrados con los indicadores en funcionamiento normal. Por ello, no es necesario realizar una parada para evaluar cada condición, ya que conocemos el estado de la máquina mientras trabaja.

El mantenimiento tanto correctivo como preventivo son los que suponen el mayor gasto de recursos e impacto en la productividad. La utilización de un mantenimiento predictivo en el que

se disponga de un entorno conectado con indicadores sobre una plataforma en tiempo real, para tomar decisiones de forma temprana sobre posibles anomalías puede reducir costes. Entre otras ventajas, proporciona reducción de la mano de obra, mejora de la fiabilidad global,pérdidas de producción por paradas no planificadas y rearranques o aumento de seguridad [30].

No obstante no todas las empresas tienen la capacidad para llevarlo a cabo( por ejemplo por la existencia de máquinas antiguas no digitalizadas que no proporcionan acceso a los datos)[12].

Para Ballesteros,antiguo director de la empresa Preditec (2017)[4], *'monitorizar las máquinas no es algo nuevo, pero en el contexto de la Industria 4.0, el desarrollo de esta actividad permite obtener datos que se generan por los sensores incorporados a la máquina'.*

Por último cabe citar algunas de las principales técnicas de mantenimiento predictivo como son [30] :

- Análisis de vibraciones: principal técnica para supervisar y diagnosticar la maquinaria rotativa e implantar un plan de mantenimiento predictivo.
- Ultrasonidos : Identificación de defectos en rodamientos o fallos eléctricos.
- Análisis de lubricantes son fundamentales para determinar el deterioro del lubricante, la entrada de contaminantes o presencia y de partículas de desgaste.
- Análisis de Aceites: Análisis periódico de muestras de aceite en las que se obtiene información del índice de acidez o de contaminación por particulas.
- Termografías: Identificación periódica de puntos calientes en cuadros eléctricos,cojinetes..
- Análisis de firma eléctrica: Identificación defectos en motores eléctricos como fallos mecánicos.

En un artículo realizado por Ballesteros [5], define el **mantenimiento prescriptivo** como un nuevo concepto que describe una estrategia de mantenimiento similar al predictivo, pero totalmente automatizado en el que se destaca la capacidad de prescribir las soluciones a los problemas detectados o incluso la programación automática de las tareas correctivas. Este tipo de mantenimiento, cita, que se está proponiendo para sistemas futuros donde el análisis predictivo y la programación de mantenimiento se realizan de manera automática.

## 2.3 Convergencia de Tecnologías Operacionales(TO) y Tecnologías de la Información(TI)

### 2.3.1 Digitalización

'El foco de la llamada 'transformación digital' está muy ligado a la 'cuarta revolución industrial', en la que digitalización e Industria 4.0 van de la mano, lo cual supone la aplicación a escala industrial de sistemas automatizados consiguiendo crear redes de producción digitales que permiten acelerarla y utilizar los recursos de manera más eficiente' [18].

La transformación digital no solo está cambiando nuestra economía, sino también aspectos como la mano de obra. La cuarta revolución industrial o la inteligencia artificial cambiaran el mercado laboral [18].

Hay muchas definiciones de digitalización [18], una de ellas podría definirse como 'la adopción masiva de la tecnología digital a través de los servicios y los dispositivos conectados'. Para ello resulta clave algunas áreas tecnológicas que ya hemos comentado como: IoT, cloud computing, impresión 3D, ciberseguridad, big data y data analytics.

La inteligencia que permite un nuevo concepto de fábrica digital es el resultado de la unión entre nuevos ecosistemas de fabricación y mantenimiento industrial [37]. Se introducen entre otros aspectos, indicadores en tiempo real que puedan permitir optimizar consumo o mejorar el control de la calidad de los productos [12].

De acuerdo a un informe elaborado por PwC (PriceWaterhouseCoopers,una de las firmas de consultoría más importante del mundo)[31],en el que realizaron entrevistas a decenas de directivos industriales dentro y fuera de España,llevaron a cabo un análisis de como la digitalización en los procesos industriales empezaba a tener efecto en reducción de costes y de eficiencia.

Una prueba de ello es la imagen adjunta [2.5] :

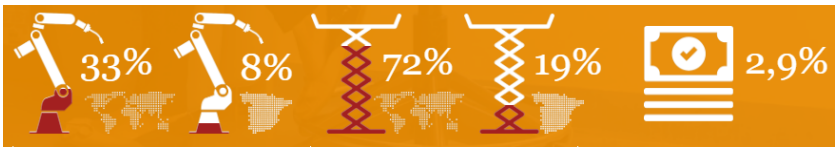



Figura 2.5: Evolución del % de nivel de digitalización en España y a nivel global en 2016. *Fuente [31]*

En el año 2016, cuándo se realizó este estudio[31], se estimaba que el 33 % de las empresas industriales globales tenían un nivel de digitalización avanzado frente al 8 % en España. Además, se estimaba que para el año 2020, habría un incremento de este porcentaje a nivel global hasta el 72 % mientras en España el incremento seguiría siendo insuficiente llegando sólo a un 19 %. Esto se traduciría en un 2.9 % de incremento adicional en la facturación en los siguientes años.

En dicho informe destacan algunas de las ventajas que ofrece la digitalización a las empresas industriales como es un incremento añadido de los ingresos, efectos directos en materia de reducción de costes y eficiencia y por último una amortización de la inversión en un corto periodo de tiempo.

### 2.3.2 Retrofitting(Adaptación tecnológica de equipos antiguos)

Hoy en día, las industrias han evolucionado con tecnologías avanzadas como sensores, robots,identificación automática ... [23]. Es posible que algunos componentes del sistema de producción de la industria sean automáticos, mientras que otros haya que operar manualmente.

Los sistemas de producción automatizados que operan en una industria son generalmente implementados por sistemas informáticos y conectado a sistemas de soporte de producción y sistemas de gestión de información a diferentes niveles de la operación. A diferencia de la industria automatizada, Industria 4.0 tiene un conjunto de tecnologías como IoT, Cloud Computing, fabricación aditiva, robots autónomos, realidad aumentada y seguridad de la información que trabajan en conjunto para mejorar el sistema de producción [23].

Por ello, siempre que sea viable, el método más eficiente,rápido y con menor coste en industria 4.0 es el *retrofitting*, el cual ayuda aumentar la eficiencia del equipo, reducir la producción del coste y aumentar la conectividad de la industria [23].

No obstante, en la actualidad hay obstáculos que impiden dicha evolución tecnológica. Una encuesta ofrecida en la Figura [2.6] realizada por la consultora PwC [31] nos indica algunos de esos obstáculos y en que medida son más 'culpables' de la no evolución digital en España.

En este informe, destacan como elemento principal(76 % de las respuestas de los directivos) la falta de una cultura digital, además de una formación adecuada. Además, creen que existe una clara falta de visión de liderazgo en alta dirección (64 %). Sin embargo, es reseñable que solo el 20 % se achaque a un talento insuficiente, lo que puede indicar que existe talento suficiente pero no se le da la formación adecuada para exprimirlo.

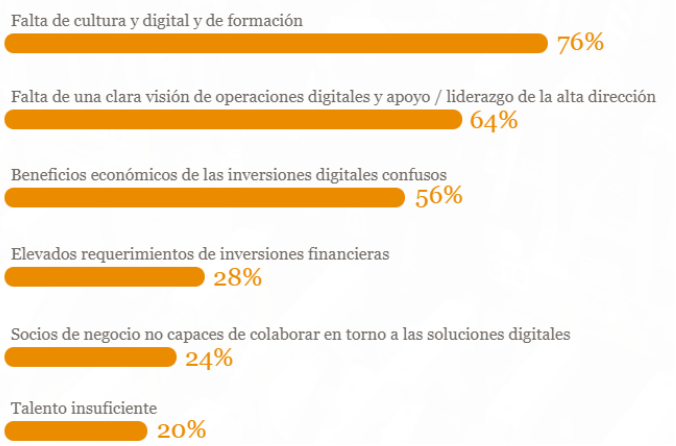



Figura 2.6: Principales obstáculos para la digitalización de industrias españolas. *Fuente [31]*

Por tanto, en este camino hacia la transformación digital de las industrias, el primer objetivo a alcanzar es la convergencia entre el mundo físico y el mundo digital, entre las tecnologías operacionales (TO) y las tecnologías de la información(TI) [12].

## 2.4 Modelos de aprendizaje automático

Hoy en día, gracias a la existencia de grandes volúmenes de datos en diversos sectores como el industrial, ciencias de la salud, ciencias sociales, mundo del deporte.., debida a las interacciones entre sistemas y seres humanos, es posible obtener patrones que nos ayudan a obtener conclusiones o resultados minimizando el error posible. Para lograrlo, una de las técnicas en las que podemos apoyarnos es el aprendizaje automático, como rama de la inteligencia artificial [27]. Antes de explicar qué es el aprendizaje automático o Machine learning(ML), cabe explicar brevemente el origen de la IA.

'El término *inteligencia artificial*(IA) fue acuñado formalmente en 1956 durante la conferencia de Dartmouth, pero para entonces ya se había estado trabajando en ello durante cinco años en los cuales se había propuesto muchas definiciones distintas que en ningún caso habían logrado ser aceptadas totalmente por la comunidad investigadora' [42]. En ese mismo año, John McCarthy,uno de los precursores de la IA, acuñó la expresión 'inteligencia artificial', y la definió como 'la ciencia e ingenio de hacer máquinas inteligentes' [42]. En este trabajo nos vamos a centrar en una de sus ramas que es el aprendizaje automático.

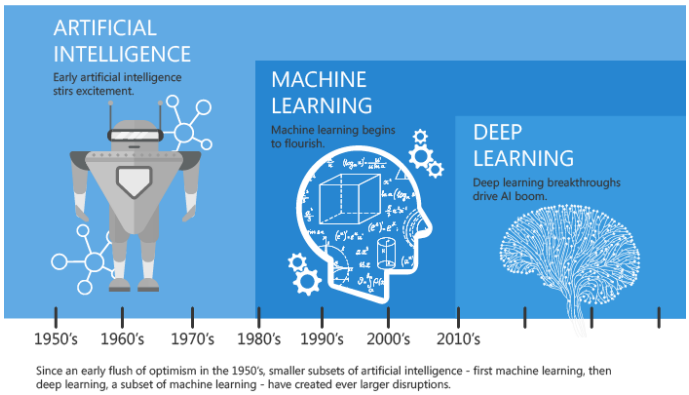



Figura 2.7: Comparativa Inteligencia Artificial, Machine Learning y Deep Learning. *Fuente [21]*

'Debido a nuevas tecnologías de cómputo, hoy día el Machine Learning no es como el del pasado. Nació del reconocimiento de patrones y de la teoría que dice que las computadoras pueden aprender sin ser programadas para realizar tareas específicas' [35]. Una de las ventajas que proporciona el ML es que es iterativo, ya que a medida que los modelos son expuestos a nuevos datos, dichos modelos se pueden adaptar de forma independiente. Además, aprenden en base a la experiencia de realizar cálculos para producir resultados confiables y repetibles. Es una ciencia que no es nueva, pero que ha cobrado un nuevo impulso [35].

No obstante, la utilización de un sistema de Machine Learning tiene muchos parámetros a tener en cuenta como pueden ser [35]:

- Preparación de los datos: Preprocesamiento, Limpieza
- Utilización de algoritmos más simples o complejos
- Automatización y procesos iterativos

- Escalabilidad

La mayoría de las industrias que utilizan grandes cantidades de datos se han hecho eco del valor de que proporciona la tecnología del machine learning. Gracias a su uso y con la obtención de datos en tiempo real, muchas empresas pueden trabajar de manera más eficiente o lograr una ventaja sobre sus competidores. Marketing,transporte, ciencias de la salud o servicios financieros ya hacen uso de esta técnica.

Por último, cabe mencionar un subconjunto de técnicas del Machine Learning como es el *deep learning*, aunque no se utilizan para la realización de este trabajo.

Se puede definir el **'deep learning'** de muchas formas. Según Ahmed Banafa, experto en Inteligencia Artificial(IA), explica como con el tiempo el aprendizaje profundo está adquiriendo mayor relevancia en el campo de la IA. Explica que el aprendizaje profundo trata del uso de redes neuronales, las cuales se utilizan para mejorar aspectos como el procesamiento de lenguaje natural(NLP) o reconocimiento por voz. Las redes neuronales son un sistema de programas y estructuras de datos que se asemejan al funcionamiento del cerebro humano [6].

Cabe resaltar otra de las técnicas derivadas de la Inteligencia Artificial y que va de la mano con el Machine Learning que es la **Mineria de datos**. Se puede definir como el análisis de conjuntos de datos mayormente de gran dimensión, con el objetivo de obtener unos datos más comprensibles y útiles. Tiene relación con otras disciplinas como bases de datos, estadística, visualización de datos o aprendizaje automático [28].

### 2.4.1 Machine Learning en Industria

La industria persigue ofrecer productos de alta calidad con el mínimo coste posible. Gracias a la introducción de las nuevas tecnologías como medio de conseguir Fábricas Inteligentes, el aprendizaje automático se ha convertido en una de las tendencias más influyentes [9].

Unido a la mayor obtención y almacenamiento de datos, esta tecnología tiene un impacto directo en la mejora de la eficiencia de los sistemas productivos, la calidad de los productos y la seguridad de las personas [9].No obstante como hemos explicado en la sección 2.3.2, a veces la existencia de muchas máquinas antiguas no digitalizadas y sin acceso a datos de forma centralizada dificultan esa obtención y almacenamiento de datos con mayor facilidad.

El machine learning puede aportar a la industria numerosas ventajas que abarcan a todo el proceso de producción. Esto se podría traducir en un incremento de la productividad, reducción de costes y aumento de la eficiencia. La principal ventaja del machine learning puede ser construir modelos predictivos usando exclusivamente datos históricos del proceso; esos modelos se pueden usar para predecir distintos parámetros como puede ser el consumo.
No obstante, para ello es necesario contar con herramientas que capturen la información producida en planta. A continuación se exponen algunas de las aplicaciones o utilidades que puede aportar el machine learning [15]:

1. Predecir productividad ,es decir, medir la capacidad de la cadena de montaje de manera virtual, pudiendo obtener información del sistema en tiempo real y actuar en consecuencia.

2. Personalización: Las propias máquinas aprenderán y podrán adaptarse al gusto del cliente con mayor precisión.

3. Automatización: El aprendizaje automático de las máquinas está basado en la información obtenida del mundo real. A partir de los datos obtenidos de las máquinas desde paradas imprevistas hasta la falta de personal, pueden extraerse soluciones para cada incidencia y automatizar dichas respuestas.

4. Control de calidad : Capturar información en tiempo real a través de los sensores, permitiendo estimar la calidad de los productos.

En el caso del impacto del aprendizaje automático en la industria 4.0 puede asociarse entre otros aspectos a mejorar el mantenimiento, a optimización de recursos, detección de anomalías en el arrancado de la máquina así como en las diferentes operaciones u optimización de la capacidad de producción [15].

Esta forma de utilizar el aprendizaje automático se traduce en una disminución de los tiempos de inactividad imprevistos, ya que los fabricantes tienen la posibilidad de pedir piezas de repuesto a un provedor antes de que se produzca la avería, lo que se denomina mantenimiento predictivo.

Según una encuesta realizada por *Deloitte*, utilizar tecnologías de aprendizaje automático en el sector de la fabricación reduce los tiempos de inactividad imprevistos entre un 15 y un 30 % y, con ello, los costes de mantenimiento en un 30 % [2].

## 2.5 Sistema de fabricación por control numérico(CNC)

El CNC tiene sus orígenes en la intención de la industria de elevar la producción. El hombre que empezó a diseñarlo fue John T.Parsons en las fábricas de Detroit y aún hoy en día se está mejorando continuamente. El CNC esta basado en un conjuntos de códigos de letras y números(correspondiente el primero a comando específico y el segundo en valores deseados), que combinados provocan el movimiento de los ejes de la máquina [40].

La aparición del control numérico por computadora parece haber superado en prestaciones a la máquina convencional pero realmente también tiene sus defectos. En contra de la máquina convencional se encuentra una menor repetitividad y un menor rendimiento ante bajas tolerancias a fallos, ya que el personal no puede tener ese rendimiento y exactitud como si lo tiene el control numérico por computadora. No obstante, para utilizar una máquina de control numérico es necesario un proceso de aprendizaje sobre su uso. Además el tiempo de preparación de la máquina es mayor debido a su mayor digitalización [36].

Los elementos principales con los que suele contar un máquina de control numérico suelen ser un dispositivo de entrada, un controlador, la máquina-herramienta( que en el desarrollo de este TFG corresponde a una fresadora propiedad de Fundación Cidaut con marca Nicolás Correa CF-20), sistema de accionamiento, dispositivos de realimentación y un monitor [17].

A continuación nos vamos a centrar en el control numérico por computadora en fresadoras, ya que es la máquina de la que proceden los datos con los que se ha trabajado.

Las fresadoras con control numérico por computadora permiten una automatización programable de la producción. Su principal aplicación se centra en volúmenes de producción medios de piezas sencillas y en volúmenes de producción medios y bajos de piezas complejas. Esto permite realizar mecanizados con alta precisión pudiendo cambiar de pieza de manera más fácil con el uso

de la consola. El equipo de control numérico se controla mediante códigos G(movimientos y ciclos fijos) y códigos M(funciones auxiliares) [41].

## 2.6 Concepto de Gemelo Digital

El concepto de *Gemelo Digital* tiene su origen en una conferencia en la Universidad de Michigan para la industria en 2002 sobre la gestión del ciclo de vida de un producto(PLM: Product Lifecycle Management) [16].

Si bien la terminología ha cambiado con el tiempo, el concepto básico de Gemelo Digital se mantiene estable desde sus inicios en 2001. La idea "raíz" se basa en la construcción de un sistema de información digital sobre un sistema físico, actuando ambos como una única entidad en sí misma. De este modo permanecerá vinculado a él a lo largo de todo su ciclo de vida [16].

La premisa bajo este concepto de Gemelo Digital se basa en dos sistemas, un sistema físico que ha existido siempre y un nuevo sistema virtual que contiene toda la información necesaria sobre el sistema físico, lo que significa que hay dos 'gemelos' de sistemas en los que hay una relación entre el espacio real y el virtual.

La creación de los gemelos digitales requiere 'conexión a los datos representados por la contraparte física', los cuales son actualmente accesibles gracias a sensores e inteligencia que ofrecen los *Sistemas Ciberfísicos*. Esta capacidad de trabajar con los datos y sobre los sistemas sin afectar al proceso productivo real, abre la posibilidad a nuevas aplicaciones como optimización,simulación o monitorizaciones en tiempo real. Todas ellas están enfocadas a nuevos planes de fabricación o con previsiones a tener un mejor nivel de precision y de fiabilidad a los actuales sistemas de simulación en planta [19].

Para llevar a cabo la implementación de un gemelo digital, la empresa Gartner junto con su vicepresidente Marc Halpern[14] exponen 4 consejos:

1. Establecer prácticas bien documentadas para construir y modificar los modelos ya que como decía Halpern: ' los gemelos digitales tienen enormes beneficios potenciales, pero crearlos y mantenerlos puede ser muy difícil.'
2. Garantizar largos ciclos de vida de acceso: aumentar la vida viable de los gemelos digitales estableciendo un objetivo para los arquitectos de TI en vías de planificar la evolución a largo plazo de los formatos y el almacenamiento de datos.
3. Incluir datos de muchas fuentes: definir una arquitectura que permita el acceso y el uso de datos de muchas fuentes.
4. Involucrar a toda la cadena de valor del producto, desde el director de la cadena de suministro hasta el director de la tecnología.

Empresas industriales y tecnológicas como Siemens [24] ya utilizan el gemelo digital de manera muy frecuente. De hecho a finales del año 2019, Siemens ha propuesto ahorros de costes y eficiencias en la producción a todos sus clientes gracias al gemelo digital. La compañia replica en entornos virtuales todas las fases del ciclo de producción. El resultado final de esta estrategia facilita la optimización de las tareas de montaje de fabricación y por ende del ahorro de gastos.

# Capítulo 3

# Fundamento teórico del piloto industrial

## 3.1 Orígenes de los datos

La cantidad de datos que se generan actualmente en los entornos industriales permite abordar nuevas soluciones de caracterización y clasificación automática de la información mediante herramientas de aprendizaje.

Los datos utilizados para esta investigación se enmarcan dentro de un proyecto de I+D realizado por la Fundación Cidaut. Proceden de una fresadora CNC Nicolás Correa CF-20 que es controlada por un operario diariamente. En concreto la fresadora, instalada en el año 1994, está situada en una bancada fija sin ningún tipo de conectividad de datos y fue retroactualizada con un sistema de digitalización no intrusivo y un panel táctil (Figura 3.1b) en 2018. Se extraen datos cada segundo sobre valores de consumo (trifásico) vibraciones en 3 ejes, temperatura y sonido. También se caracteriza la operación de mecanizado y las herramientas utilizadas por el operario.

La caracterización de los datos de producción proporciona la ventaja de poder realizar mayor número de repeticiones y asegurar que la calidad del producto es consistente.

Las dimensiones de esta fresadora(Figura 3.1a) son 5200 mm x 3127 mm x 2450mm.


(a) Fresadora

(b) Sistema digitalización y panel táctil

Figura 3.1: Fresadora y sistema digitalización

En cuánto a características técnicas de la fresadora en cuestión: el recorrido de los ejes X(longitudinal), Y(transversal) y Z(vertical) es de 1800mm,800mm y 800mm respectivamente.

Tiene un cabezal con una potencia del mandrino de 15kW con una gama de velocidades de entre 20 y 2500 revoluciones por minuto. El avance de trabajo varía entre 5 y 5000 milímetros por minuto mientras que el avance rápido va a 12000 milímetros por minuto.

### 3.1.1 Indicadores del proceso CNC

Las plantas de fabricación actuales se caracterizan por una mayor exigencia de productividad, reducción de costes y mejora de la calidad final del producto. Si bien existen múltiples estrategias para avanzar en estos requerimientos, la mejora del control y el diagnóstico de las máquinas que intervienen en el proceso productivo se han convertido en una piedra angular y una de las estrategias más exitosas para alcanzar este objetivo [32].

Pero la mejora del control y el diagnóstico de máquinas choca sistemáticamente contra una barrera: falta de digitalización en infraestructuras antiguas. Para la toma de decisiones en tiempo real, es necesario saber más sobre el funcionamiento, lo que supone abordar el desarrollo e implementación de nuevas soluciones de sensorización y conectividad. Afortunadamente en este estudio tenemos los indicadores del proceso necesarios para obtener la máxima información posible a partir de una retroactualización digital de la maquinaria industrial.

Estos indicadores clave se conocen como KPI (Key Performance Indicator). Su utilidad para medir los procesos no es otra que la de conocer el estado actual de sus actividades y recoger datos históricos para tener un seguimiento a lo largo del tiempo. Con esto se puede conocer la evolución del desempeño del proceso, y se facilita la toma de decisiones y la identificación de resultados anormales o de tendencias positivas o negativas. Además, podemos fijar valores de referencia para saber si nuestras actividades funcionan correctamente, o si debemos hacer cambios.

A continuación se exponen los diferentes indicadores que existen en las bases de datos extraídas de la fresadora gracias a los sensores que tiene incorporados:

| **Indicador Base de Datos** | **Tipo de Dato** | **Observaciones** |
|---|---|---|
| Time_Stamp | yyyy-mm-dd hh:mm:ss | Fecha medición<br>año-mes-dia hora:min:seg |
| Time_Stamp_ms | Entero Positivo | Tiempo en milisegundos en cada medición |
| CNC_Corriente_Fase1 | Real Positivo | Representan información de la corriente medido por<br>3 Sondas de corriente de fase1,fase2 y fase3. |
| CNC_Corriente_Fase2 | Real Positivo | |
| CNC_Corriente_Fase3 | Real Positivo | |
| CNC_Acelerómetrox | Real | Representan información de las vibraciones medidas por<br>los Acelerómetros transversal,longitudinal y axial. |
| CNC_Acelerómetroy | Real | |
| CNC_Acelerómetroz | Real | |
| CNC_Sonido | Real Positivo | Introduce información con<br>un sensor de sonido |
| CNC_Temperatura | Entero Positivo | Introduce información con un sensor de temperatura |
| CNC_Vallado | Categórico(0 ó 1) | Introduce información del sensor de la puerta:<br>Abierta:0<br>Cerrada:1 |
| operación_id | Categórico(0 a 7) | Representa el código de la operación realizada |
| herramienta_id | Categórico(0 a 31) | Representa el código de la herramienta usada en la operación |
| material_id | Categórico(0 a 5) | Representa el código del material utilizado |
| orden_fab | String | Representa identificador de la pieza |

Cuadro 3.1: Tabla Resumen Indicadores del proceso CNC

| Operación | Código |
|---|---|
| Toma de Ceros | 1 |
| Planeado | 2 |
| Contorneado o Copiado | 3 |
| Mandrinado | 4 |
| Fresado | 5 |
| Taladrado | 6 |
| Especiales | 7 |

Cuadro 3.2: Tabla Resumen código de las operaciones

| Material | Código |
|---|---|
| Plástico | 1 |
| Aluminio | 2 |
| Acero | 3 |
| Acero Inox o 316 | 4 |
| Otros | 5 |

Cuadro 3.3: Tabla Resumen código de los materiales

Los tres indicadores principales utilizados en la investigación son CNC_Corriente_Fase1, CNC_Corriente_Fas y CNC_Corriente_Fase3. Representan las tres corrientes alternas de la fresadora.

La alimentación trifásica representa un sistema de producción, distribución y consumo de energía eléctrica formado por tres corrientes alternas de igual frecuencia y amplitud que presentan una diferencia de fase entre ellas de 120 grados eléctricas( se suelen denominar las fases como R,S y T) [43].Ver Figura [3.2]:

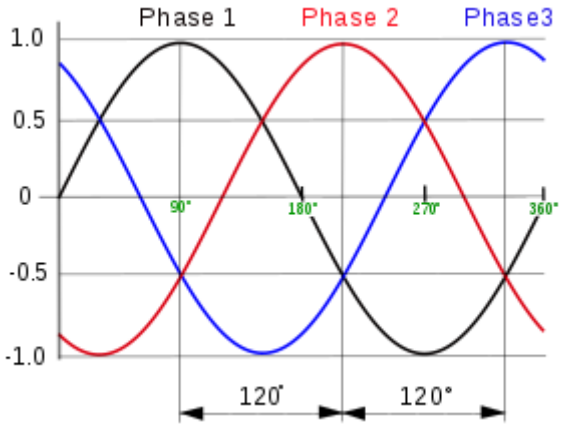



Figura 3.2: Explicación tres corrientes alternas R,S y T. *Fuente [43]*

La corriente alterna de la primera fase representa la producción, la segunda la distribución y la tercera el consumo. Con ayuda de estas variables se determina la detección de consumos anómalos sin tener en cuenta otras variables salvo el tiempo.

Otro de los indicadores que nos ayudará a conocer si la máquina esta sobrecalentada o si una operación excede la temperatura estimada es la variable CNC_Temperatura.

Para modificar las condiciones de corte sin correr riesgo de dañar la máquina se incorporan tres variables que sean capaces de medir en tiempo real las vibraciones sufridas por el cabezal. Estas tres variables son CNC_AcelerómetroX, CNC_Acelerómetroy y CNC_AcelerómetroZ, formando un acelerómetro triaxial incorporado en el cabezal de fresado(mandrino). Estos sensores longitudinal,transversal y vertical fueron especificados en las características técnicas de la fresadora.

Por último se encuentran las variables asociadas a CNC_Sonido,CNC_Vallado, operación_id, herramienta_id y material_id.

## 3.2 Metodología general

Los datos que vamos a tratar proceden del piloto industrial comentado en la *sección 3.1.* Gracias a los sensores y al control numérico por computadora que posee la fresadora, obtenemos los datos reales desde el primer segundo del arranque de la máquina, así como de las operaciones realizadas. A lo largo de esta sección se va a explicar de manera general el desarrollo que se ha seguido en este trabajo. De manera más detallada será explicado en *los capítulos 4 y 5.*

El alcance de ese trabajo se centra en uno de los comportamientos más representativos de las máquinas industriales. Se trata del proceso de arranque desde un estado de parada completa. Cualquier anomalía detectada durante el proceso de arranque es clave y condiciona la puesta en marcha del proceso de fabricación. Una de las variables más características que se puede observar en el arranque es el consumo de corriente( determinada por los indicadores CNC_Corriente_Fase1, CNC_Corriente_Fase2 y CNC_Corriente_Fase3). Existen otros factores como la vibración o la temperatura pero se ha considerado para el alcance de este trabajo el uso exclusivo de las variables de consumo de corriente, las cuales son las más influyentes en el proceso de arranque de la máquina.

Un operario se encarga del proceso de arranque de la fresadora y de la supervisión de la puesta en marcha. Esta puesta en marcha en frío permite disponer de un estado repetitivo característico, por lo que estudiar los valores de consumo durante el arranque puede ayudar a establecer patrones de condición o de estado muy útiles para el mantenimiento predictivo.

Los datos en crudo utilizados en este trabajo tienen el formato que se observa en la tabla [3.3], en la que se observan algunos de los indicadores de proceso detallados en la tabla [3.1]. En concreto, los datos empleados se corresponden con fechas desde Diciembre de 2019 hasta Marzo de 2020.

| | Time_Stamp | Time_Stamp_ms | CNC_Temperatura | CNC_Acelerometrox | CNC_Acelerometroy | CNC_Acelerometroz | CNC_CorrienteFase1 |
|---|---|---|---|---|---|---|---|
| 0 | 2019-12-01 00:00:01.0000000 | 230 | 16 | 0,047 | -1,053 | -0,262 | 0 |
| 1 | 2019-12-01 00:00:02.0000000 | 229 | 16 | 0,047 | -1,053 | -0,262 | 0 |
| 2 | 2019-12-01 00:00:03.0000000 | 230 | 16 | 0,047 | -1,053 | -0,262 | 0 |
| 3 | 2019-12-01 00:00:04.0000000 | 229 | 16 | 0,047 | -1,053 | -0,262 | 0 |
| 4 | 2019-12-01 00:00:05.0000000 | 230 | 16 | 0,047 | -1,053 | -0,262 | 0 |
| 5 | 2019-12-01 00:00:06.0000000 | 232 | 16 | 0,047 | -1,053 | -0,262 | 0 |
| 6 | 2019-12-01 00:00:07.0000000 | 231 | 16 | 0,047 | -1,053 | -0,262 | 0 |

Figura 3.3: Ejemplo Datos en bruto

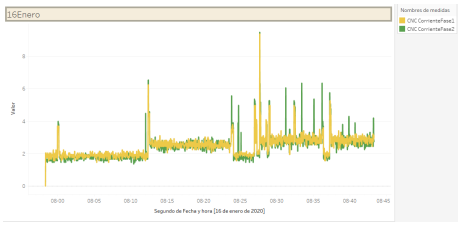
Para la realización de este trabajo se utilizarán las variables de consumo, además de la variable Time_Stamp, que determina la fecha y hora de cada consumo de la máquina para cada segundo. Además, se usan como apoyo dos variables 'Fecha' y 'Hora' a partir de la variable Time_Stamp, la cual fue renombrada por 'Fecha y Hora'. Por tanto, el formato del conjunto de datos resultante fue reducido a 6 indicadores tal y como se observa en la siguiente tabla [3.4]:

| | Fecha y hora | Fecha | CNC_CorrienteFase1 | CNC_CorrienteFase2 | CNC_CorrienteFase3 | Hora |
|---|---|---|---|---|---|---|
| 0 | 2019-12-01 00:00:01 | 2019-12-01 | 0.0 | 0.0 | 0.0 | 00:00:01 |
| 1 | 2019-12-01 00:00:02 | 2019-12-01 | 0.0 | 0.0 | 0.0 | 00:00:02 |
| 2 | 2019-12-01 00:00:03 | 2019-12-01 | 0.0 | 0.0 | 0.0 | 00:00:03 |
| 3 | 2019-12-01 00:00:04 | 2019-12-01 | 0.0 | 0.0 | 0.0 | 00:00:04 |
| 4 | 2019-12-01 00:00:05 | 2019-12-01 | 0.0 | 0.0 | 0.0 | 00:00:05 |

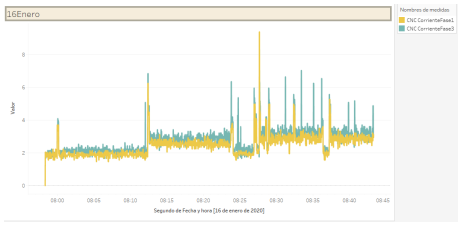
Figura 3.4: Datos filtrados del dataset en crudo

Con el conjunto de datos filtrado se realiza un preprocesamiento de los datos con el fin de determinar el inicio y el fin de cada arranque,tomando en consideración cuándo un arranque se considera válido para usarse o no. Se consideró válido el arranque en el que existieran datos con consumo nulo desde el inicio del día (00:00h) hasta antes del arranque de la máquina( cambio de consumo nulo a no nulo).

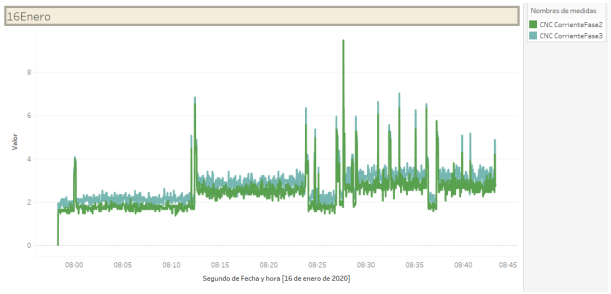
Observando de manera gráfica el comportamiento de los arranques para las tres fases de corriente(CNC_Corriente_FaseX), se determinó utilizar solo una de las tres variables debido a su similitud en cuánto a la tendencia y el comportamiento de las mismas. La elección de una de ellas se realizó con el objetivo de trabajar de un modo más flexible y rápido.A continuación se observa de manera visual la tendencia de las tres fases de corriente durante un arranque diario de la máquina, comparadas dos a dos:




(a) CNC_Corriente_Fase1 vs CNC_Corriente_Fase2




(b) CNC_Corriente_Fase1 vs CNC_Corriente_Fase3




(c) CNC_Corriente_Fase2 vs CNC_Corriente_Fase3

Figura 3.5: Comparación fases de corriente durante un arranque diario

Se observa en las gráficas anteriores como la duración del 'arranque' debe ser analizado para determinar cuando se considera ya efectuado. Por ello, inicialmente para comparar las tres fases de corriente, se escogió un tiempo grande para poder acotar a partir de un número de minutos similar para todos los arranques. Sin embargo, como se verá más adelante en el capítulo 4, la duración del arranque está íntimamente relacionada con un segmento/fase del arranque que varía entre 95 segundos hasta más de 2000 segundos( ejemplo de duración de 800 segundos en la Figura[3.6]. Durante este tiempo, el consumo oscila en torno a unos valores medios, por lo que la información proporcionada a partir de los 95 segundos se estimó como tiempo de espera en la que no se proporciona información adicional influyente. Es por ello que, finalmente, el arranque tendrá una duración menor de 5 minutos.

De acuerdo a las características de los arranques se segmenta el arranque en cuatro fases de operación, cada una con unas características particulares, como es el momento de arranque del PLC al inicio de la tercera fase de operación. Con esta segmentación será más intuitivo poder comparar unos arranques con otros de acuerdo a las características que presenten las distintas fases de operación en cuánto al consumo. No obstante, cabe comentar que tras el inicio de la cuarta fase de operación, la máquina se encuentra ya lista en disposición de realizar operaciones. Es por ello, que el estudio se ha acotado a una duración más corta de esta fase por igual para todos los arranques. Un ejemplo de arranque con la distinción entre las distintas fases de operación se encuentra en la gráfica [3.6] y será detallado en la sección 4.4.

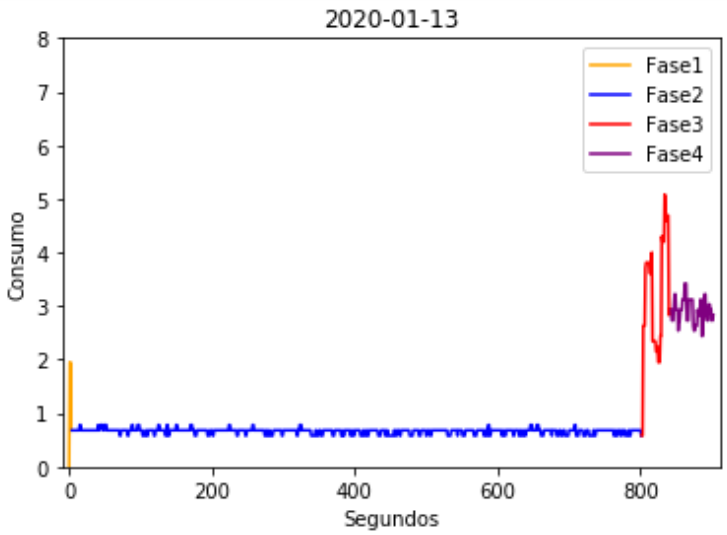



Figura 3.6: Diferenciación de fases en arranque 13 Enero de 2020

Para poder detectar el cambio de una fase de operación a otra, como se representa en la gráfica anterior, es necesario hallar el comportamiento o tendencia segundo a segundo del consumo de la máquina. Comparando el consumo en el segundo actual con respecto al segundo anterior y posterior, se obtiene el comportamiento en cada momento, pudiendo detectar crecimientos o decrecimientos del consumo típicos del cambio de fase para cada arranque. Con más detalle se explica en *la sección 4.3*.
Para el tratamiento de datos se adjunta la tabla siguiente en la que se informa de los índices correspondientes a los cambios de fases así como el 'fin' de cada uno de los arranques que se vayan a utilizar.

| **Fecha** | **Inicio Arranque** | **Fin Arranque** | **Fase 1 a Fase 2** | **Fase 2 a Fase 3** | **Fase 3 a Fase 4** |
|---|---|---|---|---|---|
| 2020-01-13 | 1 | 2286 | 3 | 802 | 841 |
| 2020-01-16 | 6989 | 9388 | 3 | 98 | 113 |

Para comenzar con el análisis de datos es necesario obtener un número de arranques determinado de entre todos los usables. Además, se empleará la variable de corriente CNC_Corriente_Fase3 para el estudio, la caracterización por fases de operación del arranque, el modo en el que se han detectado y los índices correspondientes a los cambios de fase.

De acuerdo al alcance de este trabajo, se estimó suficiente la utilización entre 20 y 25 arranques de la fresadora, siendo el número final de 24 para tener una muestra manejable con la que tratar. Para cada uno de los 24 arranques, se calculan los valores de consumo medios y máximos para cada fase de operación así como la duración de la misma, lo que hace tener información de al menos 9 variables( no se utilizará en la primera fase del estudio las características de la cuarta fase de operación). Esta información extraída se observa en la tablas [A.3], [A.4] o [A.5].

Con la existencia de todos los arranques diferenciados por fases de operación junto a las 9 variables mencionadas,el siguiente paso es crear una hipótesis basada en la agrupación de los distintos arranques en diferentes grupos atendiendo a las similitudes entre los arranques. Para ello se realizarán procesos de normalización o escalado de los datos.

Para la aplicación de la normalización de los datos, se extraen los estadísticos más influyentes correspondientes a cada fase como son la Media y el Máximo de consumo y la duración de la fase en segundos. Con estas variables se realiza una serie de transformaciones de los datos que llevarán a diferenciar en cada fase si unos arranques se comportan como otros o no, ya bien utilizando gráficas normalizadas o haciendo uso de las tablas con los estadísticos para cada fase ([A.3], [A.4] o [A.5]) . Tras esta aplicación tendremos una hipótesis con un alcance definido que hay que contrastar con la aplicación de los modelos de aprendizaje automático.

Para obtener una clasificación en grupos de arranques en base a sus características se aplicarán dos algoritmos de clustering: uno de tipo particional como *KMeans* y otro de tipo jerárquico como *complete linkage*. Con los resultados de ambos algoritmos se analizarán las diferencias y se seleccionará la agrupación de uno de ellos.Toda esta información será detallada a lo largo de la sección 4.6.

Como última fase de este trabajo , se trata de extraer los patrones más característicos en base a los modelos de comportamiento clasificados tras la aplicación de los modelos de aprendizaje automático. Tras ello, se trata de simular arranques de una fresadora en tiempo real con nuevos arranques de la máquina que se comparen con los modelos de arranque clasificados o con el modelo de comportamiento extraído de los modelos tras las clasificación. Se comprobará en que medida se adecúan estos nuevos arranques a los modelos de arranque caracterizados y con ello a la detección de anomalías.

# Capítulo 4

# Modelado del piloto industrial

## 4.1 Estructuras de datos

Los datos tratados tienen como origen las operaciones de arranque del piloto industrial de la fresadora comentada en *la sección 3.1.*

Estos datos en bruto (raw Data) iniciales, proceden de dos conjuntos de datos en formato csv. El primero contiene información de los meses de Diciembre de 2019, Enero de 2020 y la primera semana de Febrero 2020. El segundo de Febrero y Marzo de 2020. Con esta información se podrá conseguir un un número representativo de días con los que extraer un número de arranques de la fresadora , dividiéndolos para datos de entrenamiento y de prueba.

Los días festivos así como los fin de semana, la máquina se encuentra apagada y por tanto no serán necesarios para el análisis. Algunas de las observaciones asociadas a los datos en crudo del consumo de la fresadora se pueden ver en las siguientes tablas [4.1a] y [4.1b], con observaciones del 1 de Diciembre y del 31 de Marzo respectivamente.

| | Time_Stamp | Time_Stamp_ms | CNC_Temperatura | CNC_Acelerometrox | CNC_Acelerometroy | CNC_Acelerometroz | CNC_CorrienteFase1 |
|---|---|---|---|---|---|---|---|
| 0 | 2019-12-01 00:00:01.0000000 | 230 | 16 | 0,047 | -1,053 | -0,262 | 0 |
| 1 | 2019-12-01 00:00:02.0000000 | 229 | 16 | 0,047 | -1,053 | -0,262 | 0 |
| 2 | 2019-12-01 00:00:03.0000000 | 230 | 16 | 0,047 | -1,053 | -0,262 | 0 |
| 3 | 2019-12-01 00:00:04.0000000 | 229 | 16 | 0,047 | -1,053 | -0,262 | 0 |
| 4 | 2019-12-01 00:00:05.0000000 | 230 | 16 | 0,047 | -1,053 | -0,262 | 0 |
| 5 | 2019-12-01 00:00:06.0000000 | 232 | 16 | 0,047 | -1,053 | -0,262 | 0 |
| 6 | 2019-12-01 00:00:07.0000000 | 231 | 16 | 0,047 | -1,053 | -0,262 | 0 |

(a) Primeros datos del raw Data

| | Time_Stamp | Time_Stamp_ms | CNC_Temperatura | CNC_Acelerometrox | CNC_Acelerometroy | CNC_Acelerometroz | CNC_CorrienteFase1 |
|---|---|---|---|---|---|---|---|
| 4668139 | 2020-03-31 23:59:56.0000000 | 218 | 15 | 0,063 | -1,077 | -0,27 | 0 |
| 4668140 | 2020-03-31 23:59:57.0000000 | 219 | 15 | 0,07 | -1,077 | -0,27 | 0 |
| 4668141 | 2020-03-31 23:59:58.0000000 | 217 | 15 | 0,07 | -1,077 | -0,27 | 0 |
| 4668142 | 2020-03-31 23:59:59.0000000 | 218 | 15 | 0,063 | -1,077 | -0,27 | 0 |
| 4668143 | 2020-04-01 00:00:00.0000000 | 219 | 15 | 0,055 | -1,077 | -0,27 | 0 |

(b) Últimos datos del raw Data

Figura 4.1: Conjunto de datos del raw Data

La existencia de un número elevado de observaciones se debe a que se ofrece **información al segundo** sobre lo que consume y las acciones que se llevan a cabo en la fresadora. En concreto se tienen 15 indicadores del proceso [Tabla 3.1] que nos informan al segundo sobre todo consumo,operación o información relevante de la máquina.

No obstante, como se comentó anteriormente, para la realización de este TFG se utilizarán únicamente las variables asociadas al consumo eléctrico de la fresadora:

| Indicador Base de datos | Tipo de Dato | Observaciones |
|---|---|---|
| Fecha y Hora | yyyy-mm-dd hh:mm:ss | Fecha medición<br>año-mes-día hora:min:seg |
| Fecha | yyyy-mm-dd | Fecha medición<br>año-mes-dia |
| Hora | hh:mm:ss | hora:min:seg |
| CNC_CorrienteFase1 | Real Positivo | Representan información de la corriente medido por 3 Sondas de corriente de fase1,fase2 y fase3. |
| CNC_CorrienteFase2 | Real Positivo | |
| CNC_CorrienteFase3 | Real Positivo | |

Cuadro 4.1: Indicadores de proceso CNC extraídos para trabajo

El indicador Time_Stamp del conjunto de datos original se ha renombrado como 'Fecha y Hora'. Además, se han creado dos subvariables 'Fecha' y 'Hora' a partir de la variable 'Fecha y Hora'. Con la información filtrada para las 6 indicadores citados, se tendrá un conjunto de datos con arranques diarios del 1 de Diciembre al 31 de Marzo.

De acuerdo al tiempo y al alcance de este trabajo se estimó extraer 24 arranques válidos como datos de entrenamientos mientras que 20 arranques para datos de prueba para validar el modelo predictivo creado. Teniendo en cuenta de la existencia de arranques no válidos(en la siguiente seccion se detallará a que hacen referencia) ,prácticamente fueron extraídos más del 80 % de los arranques entre el 1 de Diciembre de 2019 y el 31 de Marzo de 2020.

En la siguiente sección se detallará el procesamiento y limpieza realizado en los datos para la obtención de los distintos arranques filtrados.

## 4.2 Preprocesamiento y Análisis Descriptivo de los datos iniciales

El preprocesamiento de datos es una etapa esencial del proceso de descubrimiento de información o KDD (Knowledge Discovery in Databases, en inglés) [13]. En esta etapa se lleva a cabo la limpieza,transformación y reducción de los datos para la siguiente fase de minería de datos.
Algunas de las técnicas de transformación que se han realizado es la normalización(escalar entre 0 y 1 o estandarizar a media 0 y desviación 1) o la construcción de nuevos atributos a partir de otros existentes.
Para la limpieza de datos se pretende eliminar inconsistencias,detectar anomalías graves, observar variables duplicadas o detectar arranques válidos para el estudio. Para todo ello es preciso conocer en detalle el problema y la naturaleza de los datos a tratar.
Así pues, el preprocesamiento que explicaremos a continuación, permitirá extraer la información necesaria para el trabajo así como mejorar la calidad de la información.

Antes de comenzar detallando la limpieza y procesamiento que se ha realizado a los datos, cabe citar brevemente algunas de las librerías implementadas en Python que se utilizaron en el transcurso del tratamiento de los datos:

- Numpy : Podría considerarse la librería principal para informática científica, la cual proporciona potentes estructuras de datos, implementando matrices multidimensionales. Su utiliza-

ción en mayor medida será al uso de listas,arreglos,operaciones sobre arrays o la utilización de estadísticos sobre arrays o dataframes.

- Pandas: Se utiliza para manipulación y análisis de datos. En particular, ofrece estructuras de datos y operaciones para manipular tablas numéricas,series temporales o dataframes. También permite trabajar con objetos de tipo datetime para trabajar con las fechas y horas de un modo más flexible.

- Matplotlib : Es una librería que se encarga de generar gráficos con datos procedientes de arreglos, listas o dataframes. Con esta, se visualizarán la mayoría de los gráficos generados.

- Sklearn.preprocessing: Este paquete proporciona varias funciones para el preprocesado de los datos. En este caso se ha utilizado la función *MinMaxScaler()* con el objetivo de normalizar los datos entre 0 y 1 para hacer clustering.

- Sklearn.cluster: Paquete utilizado para clustering cuándo hay datos sin etiquetar. En concreto se ha utilizado la función KMeans.

- Scipy.spatial.distance: Subpaquete utilizado para calcular matrices de distancias a partir de vectores. En concreto se ha usado cdist(a,b,'euclidean') , que calcula la distancia euclídea entre cada par de entradas que se le pasan.

- Scipy.cluster.hierarchy: Utilizado para obtener matrices de linkage y consecuentemente la realización de dendogramas. Utilizado para aplicar el algoritmo *complete linkage*.

Los procesos llevados a cabo con el fin de obtener un dataset de partida han sido los siguientes:

**Creación de nuevas variables y cambios de formato**

Con la utilización de la librería pandas, se obtiene un método para convertir la variable Fecha y Hora(anteriormente Time_Stamp) en un formato predefinido(datetime) para poder trabajar con mayor flexibilidad. El formato resultante fue: *2020-01-17 07:50:30* mientras que el formato anterior proporcionaba los milisegundos, los cuales no son de interés en este trabajo : *2020-01-17 07:50:30.00000000*. El resultado final se ve en la Figura [4.2] :

| | Fecha y hora | Fecha | Hora |
|---|---|---|---|
| 0 | 2019-12-01 00:00:01 | 2019-12-01 | 00:00:01 |

Figura 4.2: Formato Fecha y Hora Dataset

Las variables de consumo se obtienen como tipo objeto por lo que es necesario realizar su tranformación a un tipo numérico para un correcto tratamiento. Esto es originado por la utilización de la coma decimal por el punto decimal y a la identificación del número 0 sin el uso de comas y puntos, lo que lleva a Python a determinar dichas variables como objeto.

Tras esta serie de transformaciones, se obtiene el conjunto de datos preparado para la detección de arranques:

| | Fecha y hora | Fecha | CNC_CorrienteFase1 | CNC_CorrienteFase2 | CNC_CorrienteFase3 | Hora |
|---|---|---|---|---|---|---|
| 0 | 2019-12-01 00:00:01 | 2019-12-01 | 0.0 | 0.0 | 0.0 | 00:00:01 |
| 1 | 2019-12-01 00:00:02 | 2019-12-01 | 0.0 | 0.0 | 0.0 | 00:00:02 |
| 2 | 2019-12-01 00:00:03 | 2019-12-01 | 0.0 | 0.0 | 0.0 | 00:00:03 |
| 3 | 2019-12-01 00:00:04 | 2019-12-01 | 0.0 | 0.0 | 0.0 | 00:00:04 |
| 4 | 2019-12-01 00:00:05 | 2019-12-01 | 0.0 | 0.0 | 0.0 | 00:00:05 |

Figura 4.3: Datos procesados

**Detección de Arranques**

Para este estudio en concreto, de entre todos los datos registrados se usa únicamente el arranque de la fresadora, es decir, desde el momento en que el consumo aumenta por encima de 0 Amperios hasta que la máquina está lista para trabajar.

El procedimiento llevado a cabo, es la creación de una función capaz de encontrar el momento en el que el consumo de la fresadora pasa de 0 Amperios a un consumo mayor. Para detectar diariamente este momento, se necesita tanto la fecha del día concretos así como un intervalo de horarios fijados.

Se tendrá en cuenta un análisis previo de los datos para diferenciar qué datos se consideran como válidos para este trabajo. La función que se ha utilizado para la detección de estos arranques válidos se encuentra en Anexos en [A.1.1]. Para poner un ejemplo para diferenciar un arranque válido y no válido:

Arranque Válido( Valores consecutivos):

- 15 Enero 2019 08:05:23 CNC_Corriente_Fase3 0.0
- 15 Enero 2019 08:05:24 CNC_Corriente_Fase3 1.2

Arranque No Válido(Valores consecutivos):

- 20 Enero 2019 21:21:43 CNC_Corriente_Fase3 0.0
- 21 Enero 2019 09:45:21 CNC_Corriente_Fase3 2.4

El criterio expuesto con este ejemplo se basa en la *constancia,hechos* de que la máquina ha sido arrancada y que hay datos consecutivos pasados que pueden corroborar posteriormente que el arranque es válido. Es decir, que haya observaciones desde las 12 de la noche que comienza el día hasta el arranque por la mañana. Pero... ¿cómo se sabe cuándo ha arrancado la máquina?
Esto ocurrirá cuando se encuentre el primer número distinto de cero( máquina apagada sin consumir corriente), es decir el primer valor de consumo de la fresadora tras arrancar. Este procedimiento llevado a cabo es una manera subjetiva de cerciorarse de que el arranque de la máquina es fiable. Por tanto, no detectará aquellos días en los que no hay mediciones( valor 0 de consumo) entre esas franjas horarias(00:00h y 10:00h).

### Similitudes entre Fases de Corriente

El siguiente paso en la modelización de los datos será analizar de manera descriptiva como se comportan las tres fases de corriente con el transcurso del arranque.
Debido al patrón muy similar obtenido de las 3 fases de corriente fue conveniente el uso de uno de ellas con el objetivo de trabajar más rápido y de manera más flexible, como se puede observar en las imágenes de dos arranques de días distintos [4.4] y [4.5]:

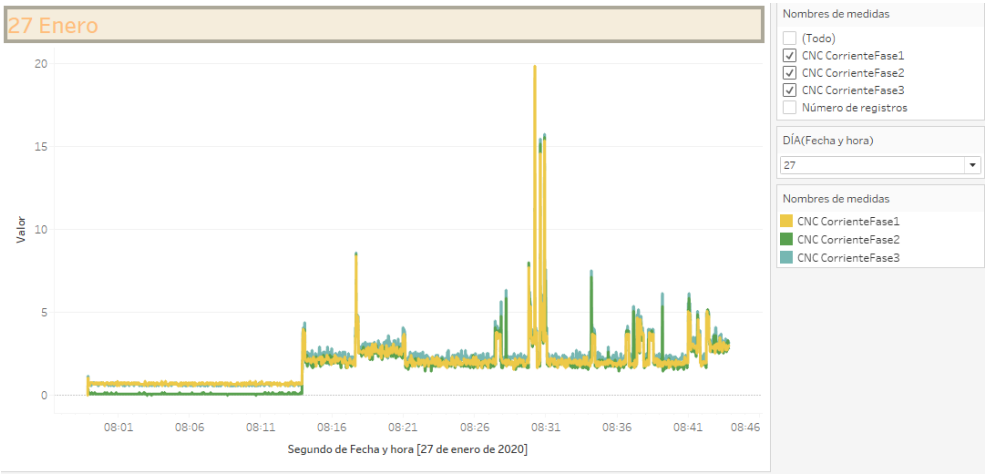



Figura 4.4: Arranque 27 Enero [Realizado con Tableau]

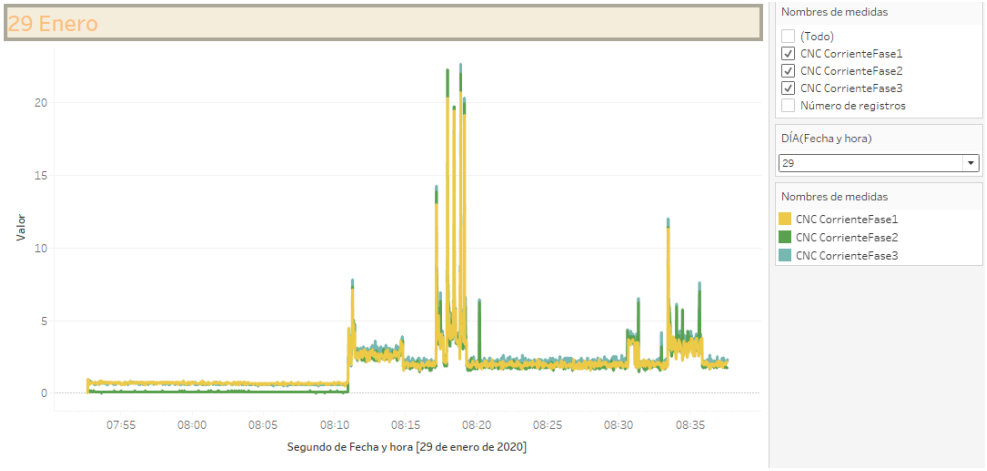



Figura 4.5: Arranque 29 Enero [Realizado con Tableau]

En ambas gráficas de arranques de dos días distintos se observa que los valores de CNC_Corriente_Fase1 y CNC_Corriente_Fase3 prácticamente se solapan, además de obtener un patrón de comportamiento similar. No ocurre exactamente lo mismo con los valores de CNC_Corriente_Fase2 ya que en la primera parte de la serie(en concreto es la primera y segunda fase de operación de las cuatro que definen el arranque) tienen unos valores de corrientes más cercanos a 0 que los valores constantes entre los que oscila la serie normalmente en esa fase(entre

0.58 y 0.78 Amperios).
Por ello y para trabajar con mayor agilidad y no repetir procesos que no aportan mucha mayor información se trabajará de aquí en adelante con una de las fases de corriente, en concreto con la fase de corriente 3(CNC_CorrienteFase3).

Un problema que se encuentra en este estudio es la dificultad de determinar cuando termina el arranque ya que no hay una variable que nos permita determinarlo.

Por ello, se estimó que el arranque de la máquina podía llevar como máximo unos 40 minutos. No obstante, no se puede saber con seguridad cuánto dura el arranque ya que no hay un indicador en el estudio que nos avise de ello, únicamente de cuándo se realiza una operación específica.

La existencia de una duración tan larga, es debido a que una fase de operación de las cuatro de las que está caracterizado un arranque(más adelante se explicará en detalle), tiene una duración muy variable pero en la que el consumo permanece estable en torno a unos valores establecidos. Por tanto, a partir de un tiempo medio establecido, la duración superior a dicho valor es tiempo de espera en la que no se proporciona información adicional. Esto se produce porque el operario puede dejar en ese estado la máquina un cierto tiempo (porque puede estar realizando otros procesos de fabricación ) hasta que retira la seta de seguridad y termina este periodo de estabilidad de consumo. Por tanto, realmente la duración 'más característica' del arranque será menor de 5 minutos.

En la siguiente sección se explicará como se han dividido en fases de operación los distintos arranques de acuerdo a la tendencia de la serie en cada momento.

## 4.3 Segmentación de los arranques

### 4.3.1 Comportamiento de la serie

Como propósito principal de esta sección, se trata de identificar las distintas fases de operación del arranque de la fresadora. De esta manera será posible hacer una comparación más concreta del arranque(por fases) con el fin de hacer un análisis más exhaustivo de los arranques(como vimos en el ejemplo de la Figura [3.6]). La caracterización de estas fases de operación fue llevado a cabo observando el comportamiento y la tendencia de todos los arranques extraídos para el análisis segundo a segundo.

El procedimiento para identificar las cuatro fases de operación de las que están formadas los arranques, se determina comparando el consumo segundo a segundo durante el arranque. Se compara el consumo del segundo actual (por ejemplo el segundo 10) con el segundo anterior y posterior( segundos 9 y 11). Con este análisis se puede determinar si hay un desfase de consumo actual con el momento anterior y posterior y por tanto tener un control sobre el comportamiento del arranque segundo a segundo. Esto permite detectar los crecimientos o decrecimientos más destacables asociados a los cambios de fases.

Por lo general, la relación entre el consumo de la observación actual con la anterior y posterior en cada momento del arranque, puede ser una relación creciente,decreciente o estable. Por tanto, tendremos 9 combinaciones posibles entre estos 3 comportamientos. A cada una de ellas le asigna-

mos una etiqueta según el comportamiento obtenido(del 0 al 8).
Sin embargo, para considerar algún valor de los nueve posibles, se ha estimado que la diferencia entre observaciones consecutivas, deberá ser de *al menos 0.75 Amperios*. Es decir, puede haber desfases menores de 0.75 Amperios y ser etiquetado como un comportamiento estable. La función implementada asociada a dicha detección de la tendencia en cada segundo se encuentra en Anexos en [A.1.2].

Para completar de un modo más visual lo que ha sido explicado teóricamente se adjuntan los nueve comportamientos que puede existir entre tres momentos consecutivos de la serie. La parte **gris**(1) se corresponde con la diferencia entre la parte observada actual y la pasada, mientras que la parte **granate**(2) se realiza entre la observación futura y la actual.

**Estabilidad : 0**

Figura 4.6: Comportamiento 0

**Decreciente y Estable : 1 ; Creciente y Estable : 2**

(a) Decreciente y Estable : 1

(b) Creciente y Estable: 2

Figura 4.7: Comportamientos 1 y 2

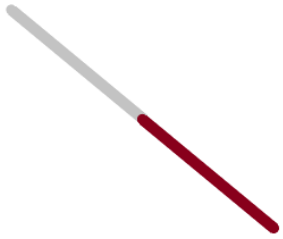
**Decreciente y Creciente : 3 ; Estable y Creciente: 4**

(a) Decreciente y Creciente: 3

(b) Estable y Creciente: 4

Figura 4.8: Comportamientos 3 y 4

**Decreciente y Decreciente: 5 ; Estable y Decreciente: 6**



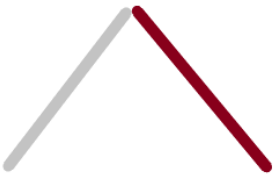
(a) Decreciente y Decreciente: 5

(b) Estable y Decreciente: 6

Figura 4.9: Comportamientos 5 y 6

**Creciente y Creciente : 7 ; Creciente y Decreciente: 8**

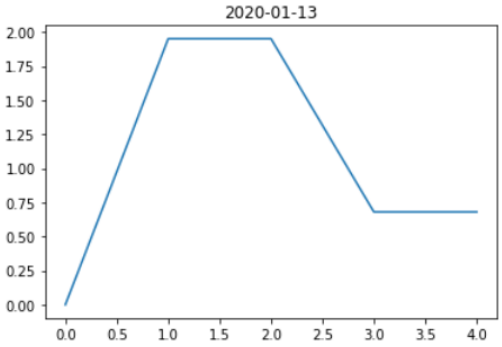
(a) Creciente y Creciente: 7



(b) Creciente y Decreciente: 8

Figura 4.10: Comportamientos 7 y 8

De esta manera será posible detectar los incrementos y decrementos influyentes en cada instante durante cada arranque. Para entender el modo con el que diferenciaremos las distintas fases se adjunta un ejemplo del comportamiento de el arranque de la fresadora del día 13 de Enero durante los primeros segundos en los que existen variaciones mayores de 0.75 Amperios y por tanto categorizadas como un tipo de comportamiento de los explicados.




(a) Gráfica arranque inicial 13 Enero de 2020

```
[0. 2. 6. 1. 0.]
```

(b) Etiquetado arranque inicial 13 Enero de 2020

Figura 4.11: Comportamiento arranque inicial 13 Enero

Los resultados de la figura [4.11b] se obtienen de calcular la tendencia del consumo en cada instante del arranque correspondiente a la figura de ejemplo [4.11a]. Obviamente el primer valor

etiquetado es 0 ya que no tiene valores pasados. Si observamos el comportamiento del consumo en el **primer** segundo con respecto al segundo **cero** o al segundo **dos**, se ve que el incremento con respecto a la observación pasada(segundo cero) es positivo y mayor que 0.75. Sin embargo , con respecto a la observación posterior(segundo dos) la diferencia es 0. Por tanto este comportamiento es característico con el comportamiento 2 :[4.7b]. La siguiente observación(segundo 2) contiene un consumo similar al segundo anterior pero superior a la siguiente observación(segundo 3) con una diferencia mayor de 0.75 Amperios. Por tanto se corresponde con el comportamiento categorizado como 6: [4.9b]. Por último, la relación del consumo en el segundo **tres** con el consumo anterior fue con una diferencia de más de 0.75 Amperios mientras que en relación con el consumo del segundo se mantiene estable, lo cual corresponde al comportamiento 1:[4.7a].

## 4.4 Segmentación de Fases de operación

A lo largo de esta sección se explica como se han detectado los cambios de fase de operación así como las características más influyentes de cada una de ellas. Para ello, tras definir un inicio y fin para todos los arranques, estos fueron agrupados en un mismo conjunto de datos para permitir la detección de los momentos de cambios de fase de cada uno de los arranques. De aquí en adelante, *se utilizará indistintamente el término fase o fase de operación designando las fases de operación.*

**Detección Fase de operación 1**

Antes de comenzar esta fase, el consumo inicial tiene corriente nula. La característica de esta fase se corresponde con el **inicio del arranque de la máquina**. No tiene una duración superior a 6 segundos ya que sólo se caracteriza por alcanzar un 'pico' de corriente y tras ello disminuye el consumo hasta alcanzar unos valores estables,propios de la siguiente fase. En la sección de Anexos A.2.3 se puede consultar ejemplos donde se encuentran diferenciados los arranques utilizados así como datos de entrenamiento para esta fase. Para detectar el periodo que representa esta fase para cada arranque, se diferencian tres posibles distinciones según su comportamiento:

- Una característica de esta fase corresponde a la detección del pico de corriente, con el seguimiento de valores de consumos similares al pico de consumo durante la siguiente fase. Por tanto, equivale a encontrar el patrón de comportamiento número 2 [4.7b] seguida de comportamientos estables al no haber un decrecimiento fuerte ( [0,2,0]).

- Crecimiento del consumo desde consumo 0 , llegando a un 'pico' de corriente, con la estabilización de este consumo durante un segundo y el posterior decrecimiento a valores estables propios de la fase 2. En términos de comportamiento como se ha categorizado en la sección anterior seguiría el patrón [0,2,6,1] asociados los tres últimos a [4.7b],[4.9b] y [4.7a].Esto se produce debido a una diferencia en el decrecimiento del consumo de más de 0.75 Amperios de diferencia y por ello es etiquetado como 6 [4.9b] y 1 [4.7a].

- Alcanzar el pico de corriente en el primer segundo pero la no estabilización del consumo durante 1 segundo sino la existencia de una disminución de corriente también exponencial, lo que se corresponde con el comportamiento[0,8,1]: [4.10b y [4.7a]. Sólo se da en un caso de las 24 analizadas.

**Detección Fase de operación 2**

Esta fase caracteriza por contener una serie de valores estables durante un cierto tiempo hasta que comienza un crecimiento del consumo hasta encontrar otro 'pico' de corriente. En concreto, se corresponde con el **ciclo del trabajo del PLC.** La duración de esta fase es muy variable, desde 95 segundos hasta algún arranque con 2000 segundos. El valor del consumo entre el que oscila normalmente suele ser entre 0.58 y 0.78 Amperios. No obstante, también puede variar en los 2 Amperios en alguna serie.
Para detectarlo se lleva el siguiente proceso:

- Se detectará una subida del consumo importante, la cual seguirá el comportamiento 4 [4.8b], es decir, cambio de un consumo estable a un crecimiento notable. Con la detección de este crecimiento será suficiente para determinar el fin de la fase 2. El inicio de la fase 2 se corresponde con el fin de la fase 1.

**Detección Fase de operación 3**

Esta fase se caracteriza por el **arranque del motor**.Esta fase, pasa por encontrar el 'pico' de corriente más alto y posteriormente encontrar el momento en el que se estabiliza la serie en un valor nominal. La duración dependerá de lo que tarde en estabilizarse el consumo tras encontrar el máximo local.
Por tanto, se detectarán en este caso los comportamientos relacionadas con decrecimientos tras localizar los máximos locales. Esto se realiza porque ya sabemos de la existencia de un pico al haber terminado la fase 2. Se ha estimado que la fase 3 puede prolongarse hasta 30 segundos más hasta que la serie se estabiliza ya que es posible encontrar de nuevo otros 'picos' de consumo. Para detectar el fin de la fase 3 ,se realiza el siguiente proceso:

- Detección de un decrecimiento proveniente de una estabilización del consumo característico del comportamiento 6 [4.9b]. Posteriormente, tras el decrecimiento, la siguiente observación es semejante a la actual, por lo que el comportamiento es característico del 1 [4.7a]. Comportamiento [6,1].

- Detección de un decrecimiento proveniente de un crecimiento, es decir, haber encontrado un máximo local. Este comportamiento es característico del 8 [4.10b]. Posteriormente, tras el decrecimiento, la siguiente observación es semejante a la actual, por lo que el comportamiento es característico del 1 [4.7a].Comportamiento [8,1].

- No obstante, se ha observado en esta fase que al alcanzar el 'pico' de corriente los datos no se estabilizan en torno a un valor establecido durante más de 20 segundos, sino que sigue habiendo crecimientos y decrecimientos en algunos casos. Por tanto, se analizan hasta 30 segundos adicionales al primer máximo localizado para determinar cuando acaba la fase 3. Para ello seguimos el mismo procedimiento para detectar el anterior pico de consumo ya sea con un comportamiento [8,1] o [6,1].

**Detección Fase de operación 4**

Se caracteriza con el **calentamiento del motor** en la que de acuerdo con el operario de la máquina, se comentó que la máquina ya se encuentra en condiciones de realizar operaciones una vez finalizada la fase 3.

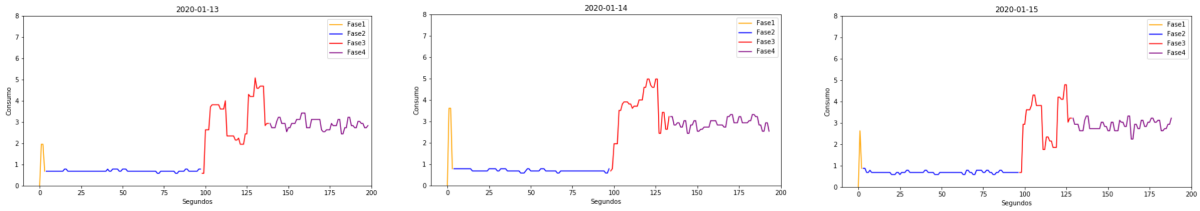
De momento, la detección de esta fase de operación se realiza filtrando desde el fin de la fase 3 hasta que finaliza la serie del arranque en cuestión(el límite máximo que propusimos para todos los arranques de 40 minutos).

Es por ello, que debido a la dificultad para determinar cuándo realmente termina el arranque, se utilizará una muestra similar para todos los arranques de 60 segundos para la fase 4. El tiempo pequeño que se ha estimado es con motivo de no dar mayor importancia a esta fase que a las otras tres, las cuales si tenemos un inicio y un fin establecido.
Durante al menos los primeros 60 segundos, los valores de consumo permanecen estables en torno a valores medios, por lo que se estimó como una muestra representativa de como evoluciona el consumo en esta fase.

Destacar que la realización de la detección de las distintas fases de operación, sorprendentemente, llevaba un tiempo de cómputo elevado en Python, lo que llevó a utilizar dicho pseudocódigo al lenguaje de R y poder realizarlo en un tiempo de cómputo mucho menor. Con objeto de tener información sobre los cambios de fase de los 24 arranques extraídos como datos de entrenamiento, se forma un conjunto de datos con la información de la fecha y los índices correspondientes a los cambios de fases así como la posición del inicio y final del arranque del día respectivo. Esta se encuentra representada en la tabla [A.1].

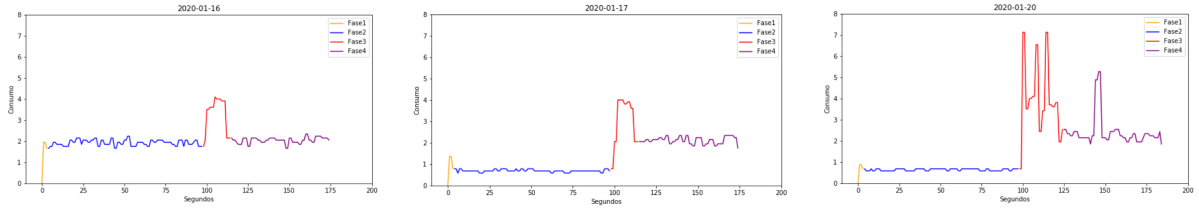
Para tener la información asociada a los 24 arranques y ver como se han segmentado las distintas fases del arranque, se adjuntan las series temporales de cada uno de los 24 arranques, todos ellos con el mismo eje de coordenadas para que su comparación sea más fácil de manera visual. El eje X está representado por los segundos del arranque así como el eje Y por consumo en Amperios. En Anexos, en [A.3.1], se puede encontrar los arranques reales sin aplicar la interpolación en la que la duración de la fase 2 es muy variable con diferencias entre 95 y 2000 segundos. Con motivo de diferenciarlos visualmente con mismas escalas, se realizó una interpolación en la duración de dicha fase a 95 segundos para todos los arranques.

### Arranques - datos entrenamiento



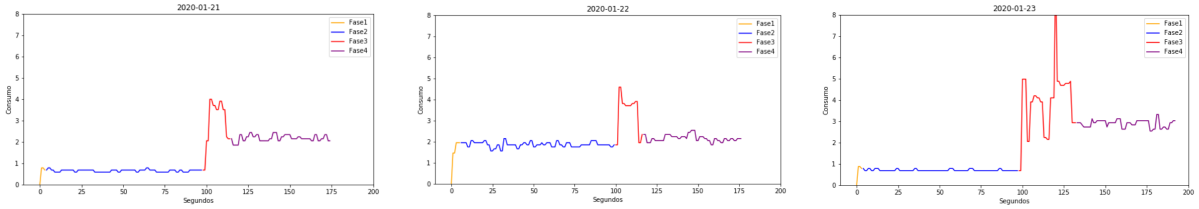

(a) Arranque 13 Enero (b) Arranque 14 Enero (c) Arranque 15 Enero

Figura 4.12: Arranques 13,14,15 Enero de 2020



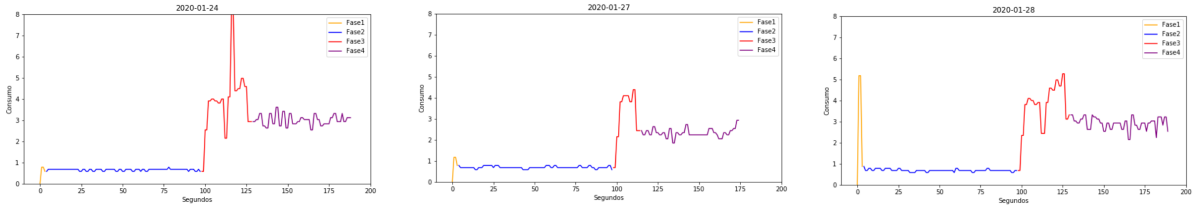

(a) Arranque 16 Enero (b) Arranque 17 Enero (c) Arranque 20 Enero

Figura 4.13: Arranques 16,17,20 de Enero de 2020



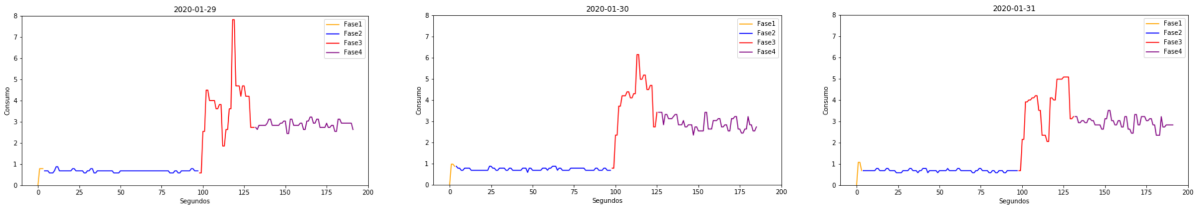

(a) Arranque 21 Enero (b) Arranque 22 Enero (c) Arranque 23 Enero

Figura 4.14: Arranques 21,22,23 de Enero de 2020



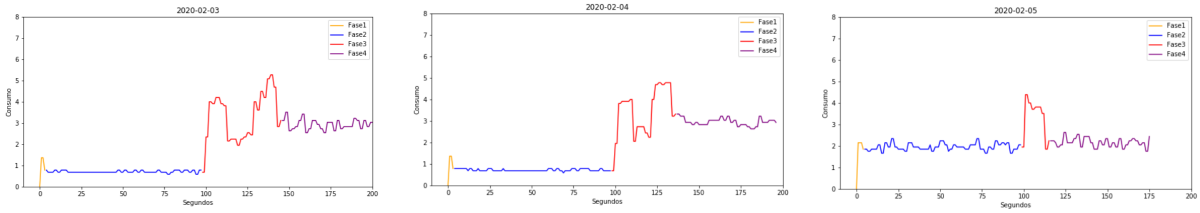

(a) Arranque 24 Enero (b) Arranque 27 Enero (c) Arranque 28 Enero

Figura 4.15: Arranques 24,27,28 de Enero de 2020



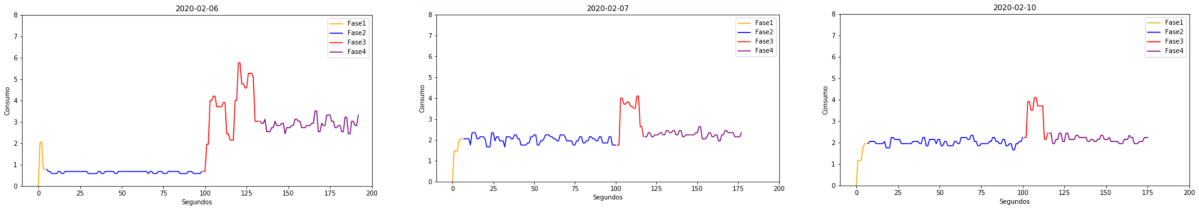

(a) Arranque 29 Enero (b) Arranque 30 Enero (c) Arranque 31 Enero

Figura 4.16: Arranques 29,30,31 Enero de 2020



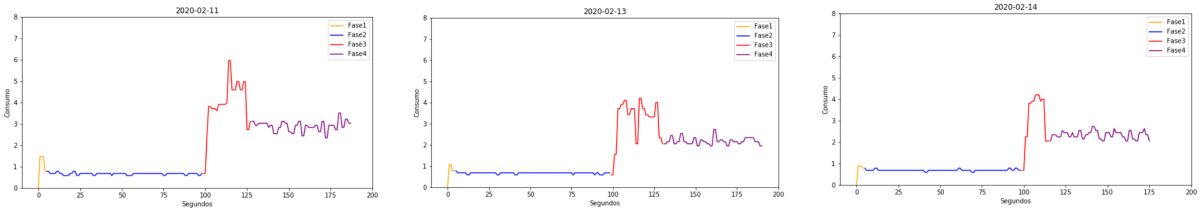

(a) Arranque 3 Febrero (b) Arranque 4 Febrero (c) Arranque 5 Febrero

Figura 4.17: Arranques 3,4,5 de Febrero de 2020




(a) Arranque 6 Febrero (b) Arranque 7 Febrero (c) Arranque 10 Febrero

Figura 4.18: Arranques 6,7,10 de Febrero de 2020




(a) Arranque 11 Febrero (b) Arranque 13 Febrero (c) Arranque 14 Febrero

Figura 4.19: Arranques 11,13,14 de Febrero de 2020

### 4.4.1 Caracterización de fases de operación de los arranques

En esta sección se plantea una hipótesis de partida para contrastarla posteriormente aplicando técnicas de aprendizaje automático. Esta hipótesis plantea una serie de agrupaciones de los distintos arranques en base a sus similitudes en el consumo. Para ello se utilizarán como datos característicos, los valores del consumo medio, máximo y la duración(tiempo en segundos) de las distintas fases de operación de las que se compone el arranque. Concretamente:

- Fase 1: Media, Máximo y Duración
- Fase 2: Media, Máximo y Duración
- Fase 3: Media, Máximo y Duración

Con esta información se podrá representar gráficamente la tendencia y el comportamiento del consumo en el tiempo para cada una de las tres fases. Escalando las variables(en el rango 0-1) tomando como valor máximo el valor medio de consumo de todos valores máximos/medios de todos los arranques. De este modo, se podrán detectar de manera gráfica para cada fase, valores anómalos que no se ajustan a los valores medios de consumo.

Por tanto, se tiene un conjunto de datos formado por 24 filas correspondientes a los arranques diarios y 9 columnas asociadas a las variables anteriormente indicadas (tablas A.3 A.4 y A.5).

Además, se pueden observar los comportamientos de todos los arranques para las distintas fases sin aplicar normalizaciones. Para la Fase 1 se encuentran en [A.3.2], la Fase 2 en [A.3.3] y para la Fase 3 en [A.3.4].

Con el objetivo de determinar que arranques se comportan de manera similar al resto de arranques, se lleva a cabo un análisis individual por fases para determinar las distintas asociaciones entre los distintos arranques. Para ello se tendrán en cuenta tanto los valores de consumo durante cada segundo así como el tiempo por fase.

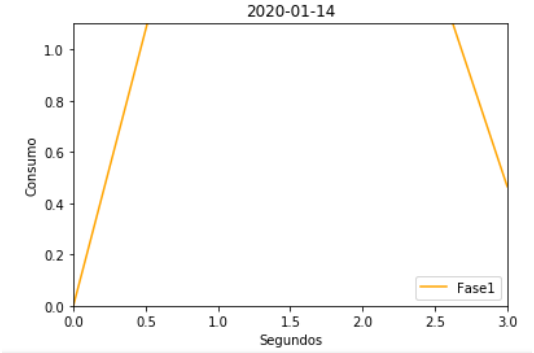
**Fase 1**

Parte de un proceso de normalización de los datos tanto del consumo como la duración de la fase. Para la primera fase, correspondiente al arranque inicial de la máquina, se ha realizado una normalización en base al valor medio de todos los valores máximos de consumo de los arranques. Se realiza en base al valor máximo, ya que esta fase se caracteriza por llegar a un pico de consumo. La función encargada de normalizar en base a un valor medio dado se encuentra en Anexos en [A.1.3]. En la siguiente tabla se encuentran los valores medios asociados, del que utilizaremos el máximo en vez de la media para esta fase :

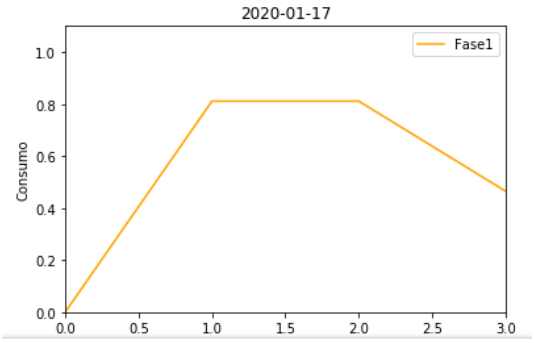
| | Media de la media | Media del Máximo | Duracion Media |
|---|---|---|---|
| 0 | 1.066944 | 1.676667 | 3.208333 |

Figura 4.20: Valor medio del máximo, media y duración de Fase 1 de los 24 arranques

Con la aplicación de esta normalización, se tienen las mismas escalas para todos los arranques y por tanto se detectará cuáles se encuentran por encima de la media de consumo (se encuentran por encima del valor medio normalizado) o por debajo. En base a esta información se pueden determinar similitudes de consumo para los arranques con valores menores de la media establecida. Del mismo modo, se agrupan los arranques con un valor medio superior a la media para establecer similitudes entre estos. Como ejemplo de lo comentado se observan dos gráficas asociadas [4.21a] y [4.21b]. Se puede observar como el valor de consumo del día 14 de enero se encuentra 'por encima de la gráfica' lo que significa la existencia un valor por encima de la media. Sin embargo, el valor asociado al día 17 Enero se encuentra por debajo de la media. No obstante, es necesario observar que arranques contienen valores de consumo por encima de la media para poder detectar un valor de consumo excesivo o cercano al valor medio.




(a) Arranque Fase 1 Día 14 Enero

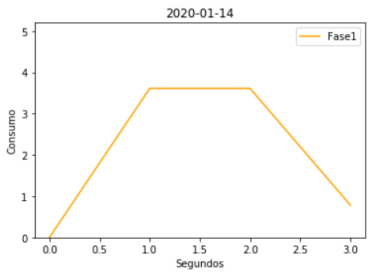



(b) Arranque Fase 1 Día 17 Enero

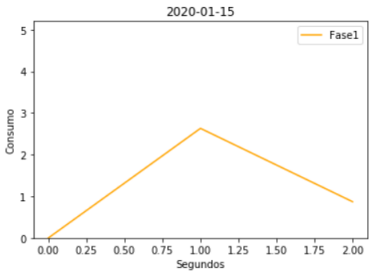
Figura 4.21: Arranques de ejemplo normalizados en Fase 1

Con la interpretación de estos resultados obtenidas de las gráficas junto con la ayuda de la tabla resumen de los estadísticos de la fase 1 [A.3] se puede ir diferenciando arranques así como encontrar posibles valores anómalos.

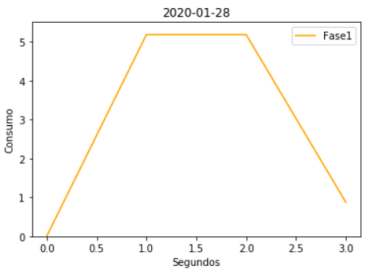
Ejemplo de ello son los **valores de 2.6, 3.61 y 5.18 Amperios** asociados a los días **14,15 y 28 de Enero**. Se han considerado anómalos debido a los valores medios que obtuvimos en la tabla 4.20, en la que el valor medio de todos los valores máximos era de 1.67 Amperios.Por tanto, de aquí en adelante se eliminan del estudio, aunque se observará si se encuentran mas anomalías en el resto de las fases. A continuación se observa el exceso de consumo de estos arranques durante esta fase:




(a) Anómalo Día 14 Enero




(b) Anómalo Día 15 Enero




(c) Anómalo Día 28 Enero

Figura 4.22: Arranques Anómalos en Fase 1

Observando la tabla de características [A.3] y las gráficas asociados a los arranques se han realizado distintos grupos característicos. Los días asociados a cada grupo así como las características de cada arranque se encuentran en las tablas siguientes:

| | Dia | Fase1 Media | Fase1 Max | Fase1 Duracion |
|---|---|---|---|---|
| 7 | 2020-01-22 | 1.461667 | 1.95 | 5.0 |
| 19 | 2020-02-07 | 1.396667 | 2.05 | 5.0 |
| 20 | 2020-02-10 | 1.201667 | 1.95 | 5.0 |

(a) Arranques Grupo 1 Fase 1

| | Dia | Fase1 Media | Fase1 Max | Fase1 Duracion |
|---|---|---|---|---|
| 0 | 2020-01-13 | 1.1450 | 1.95 | 3.0 |
| 3 | 2020-01-16 | 1.3900 | 1.95 | 3.0 |
| 17 | 2020-02-05 | 1.6125 | 2.15 | 3.0 |
| 18 | 2020-02-06 | 1.2200 | 2.05 | 3.0 |

(b) Arranques Grupo 2 Fase 1

Figura 4.23: Arranques Fase 1 - Grupos 1 y 2

| | Dia | Fase1 Media | Fase1 Max | Fase1 Duracion |
|---|---|---|---|---|
| 5 | 2020-01-20 | 0.6050 | 0.87 | 3.0 |
| 6 | 2020-01-21 | 0.5600 | 0.78 | 3.0 |
| 8 | 2020-01-23 | 0.6300 | 0.87 | 3.0 |
| 9 | 2020-01-24 | 0.5350 | 0.78 | 3.0 |
| 12 | 2020-01-29 | 0.5850 | 0.78 | 3.0 |
| 13 | 2020-01-30 | 0.7025 | 0.97 | 3.0 |
| 23 | 2020-02-14 | 0.6525 | 0.87 | 3.0 |

(a) Arranques Grupo 3 Fase 1

| | Dia | Fase1 Media | Fase1 Max | Fase1 Duracion |
|---|---|---|---|---|
| 4 | 2020-01-17 | 0.875 | 1.36 | 3.0 |
| 10 | 2020-01-27 | 0.780 | 1.17 | 3.0 |
| 14 | 2020-01-31 | 0.705 | 1.07 | 3.0 |
| 15 | 2020-02-03 | 0.875 | 1.36 | 3.0 |
| 16 | 2020-02-04 | 0.875 | 1.36 | 3.0 |
| 21 | 2020-02-11 | 1.095 | 1.46 | 3.0 |
| 22 | 2020-02-13 | 0.730 | 1.07 | 3.0 |

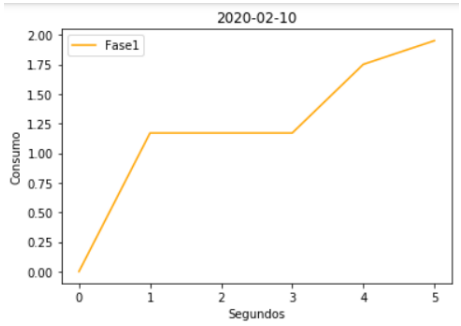
(b) Arranques Grupo 4 Fase 1

Figura 4.24: Arranques Fase 1 - Grupos 3 y 4

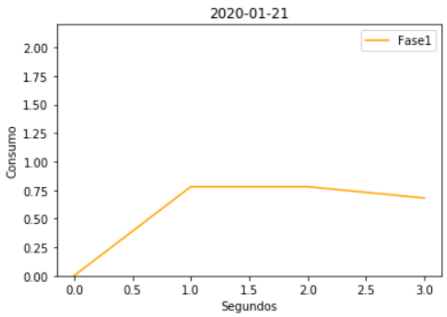
Las características de los cuatro grupos se detallan en la siguiente tabla:

| **FASE 1** | **Características** |
|---|---|
| **Grupo 1** | - Duración de 5 segundos<br>- Pico de consumo en torno a 2 Amperios |
| **Grupo 2** | - Duración 3 Segundos<br>- Pico de consumo en torno a 2 Amperios |
| **Grupo 3** | - Duración 3 Segundos<br>- Pico de consumo menor 1 Amperio |
| **Grupo 4** | - Duración 3 Segundos<br>- Pico de consumo 1-1.5 Amperios |

Cuadro 4.2: Características grupos en Fase 1




(a) Ejemplo Grupo 1 - Día 10 Enero Fase 1




(b) Ejemplo Grupo 2 - Día 16 Enero Fase 1

Figura 4.25: Ejemplos Grupo 1 y 2 para Fase 1

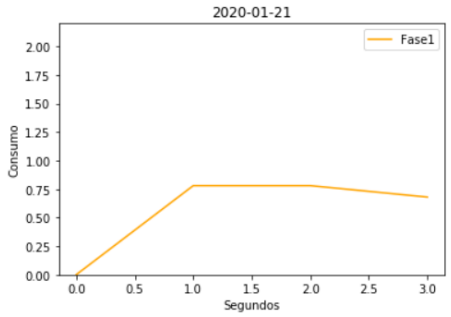
| | Media Media | Media Máximo | Media Duracion |
|---|---|---|---|
| 0 | 1.353333 | 1.983333 | 5.0 |

(a) Estadísticos Grupo 1

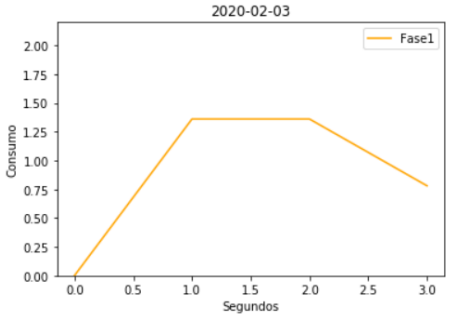
| | Media Media | Media Máximo | Media Duracion |
|---|---|---|---|
| 0 | 1.341875 | 2.025 | 3.0 |

(b) Estadísticos Grupo 2

Figura 4.26: Estadísticos Grupo 1 y 2 en Fase 1




(a) Ejemplo Grupo 3- Día 21 Enero




(b) Ejemplo Grupo 4 - Día 3 Febrero

Figura 4.27: Ejemplos Grupo 3 y 4 para Fase 1

| | Media Media | Media Máximo | Media Duracion |
|---|---|---|---|
| 0 | 0.61 | 0.845714 | 3.0 |

(a) Estadísticos Grupo 3

| | Media Media | Media Máximo | Media Duracion |
|---|---|---|---|
| 0 | 0.847857 | 1.264286 | 3.0 |

(b) Estadísticos Grupo 4

Figura 4.28: Estadísticos Grupo 3 y 4 en Fase 1

Las gráficas del comportamiento de todos los arranques en la fase 1 se encuentran en la sección de anexos siendo el primero de ellos [A.19a].

**Fase 2**

El procedimiento llevado a cabo para la fase 2 es similar, normalizando en torno al valor medio de consumo. En este sentido,se encuentran dos patrones de comportamiento definidos, un consumo medio similar por debajo de la media en 16 de los 21 arranques( ya no son 24 porque desestimamos 3 en la anterior fase) y un consumo medio similar en los 5 grupos restantes, en este caso con valor por encima de la media. Los valores medios de duración de tiempo, máximo y media de consumo se representan a continuación, aunque realmente sólo es útil 'la Media de la Media' ya que esta fase no interesa ni los valores máximos ni la duración, ya que esta última es muy variable y no proporciona información adicional.

| | Media de la media | Media del Máximo | Duracion Media |
|---|---|---|---|
| 0 | 0.936694 | 1.188333 | 762.5 |

Figura 4.29: Valor medio del máximo, media y duración de Fase 2 de los 21 arranques

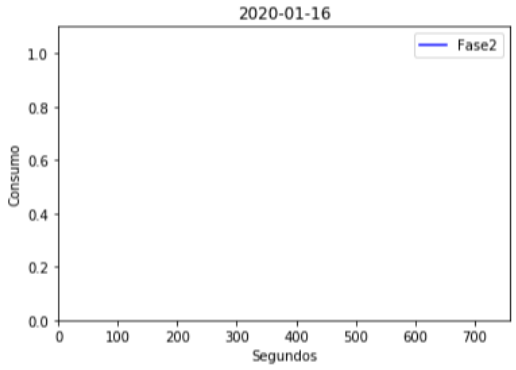
Las características de los dos grupos para esta fase son:

| FASE 2 | Características |
|---|---|
| **Grupo 1** | - Duración 95 Segundos<br>- Valores de consumo en torno 2 Amperios<br>- 16 Enero,22 Enero,5 Febrero,7Febrero y 10 Febrero |
| **Grupo 2** | - Duración variable hasta 2000 segundos<br>- Valores de consumo en torno a 0.58-0.78 Amperios<br>- Resto de Arranques |

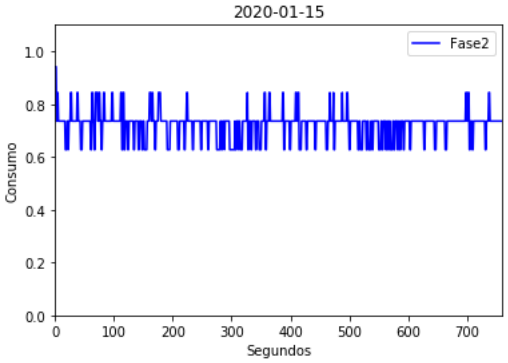
Cuadro 4.3: Características grupos en Fase 2

Como se verá más adelante, todos los arranques se ajustarán a 95 segundos, estimado como el tiempo suficiente y representativo para esta fase. Por tanto, se determinó que la **duración de la fase 2 no sería una variable influyente en la clasificación**.

Para ver un ejemplo de lo que se está explicando se adjuntan dos gráficas con mismas escalas que representan lo comentado [4.30b] y [4.30a]. En la primera no se observan datos representados debido a que los valores de consumo están por encima de la media normalizada(0.93), lo cual lleva a establecer un grupo con los arranques que se caracterizan por ese consumo. Además, se diferencian los dos grupos de arranques con las tablas [4.31a] y [4.31b].




(a) Arranque Fase 2 Día 16 Enero




(b) Arranque Fase 2 Día 15 Enero

Figura 4.30: Arranques escalas normalizadas en Fase 2

| | Dia | Fase2 Media | Fase2 Max | Fase2 Duracion |
|---|---|---|---|---|
| 3 | 2020-01-16 | 1.927527 | 2.24 | 95.0 |
| 7 | 2020-01-22 | 1.863763 | 2.15 | 95.0 |
| 17 | 2020-02-05 | 1.953936 | 2.34 | 96.0 |
| 19 | 2020-02-07 | 2.019785 | 2.34 | 95.0 |
| 20 | 2020-02-10 | 2.016667 | 2.34 | 95.0 |

(a) Arranques Grupo 1 Fase 2

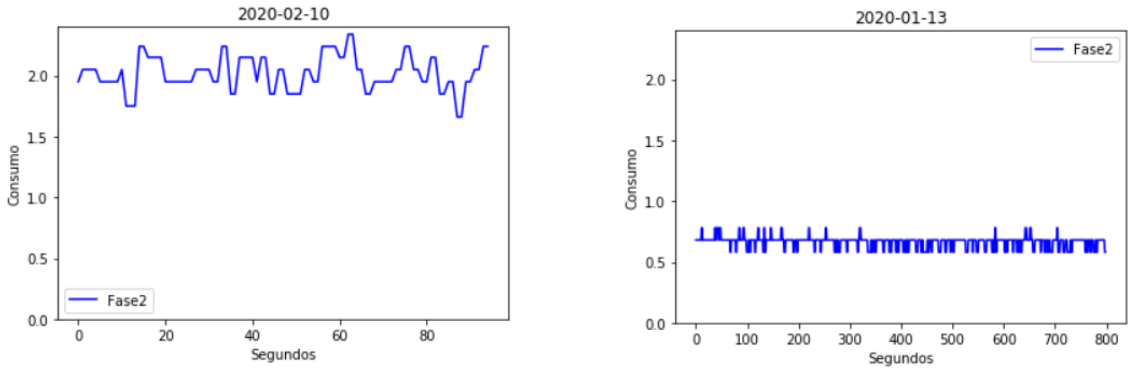
| | Dia | Fase2 Media | Fase2 Max | Fase2 Duracion |
|---|---|---|---|---|
| 0 | 2020-01-13 | 0.662936 | 0.78 | 799.0 |
| 4 | 2020-01-17 | 0.692912 | 0.87 | 1142.0 |
| 6 | 2020-01-21 | 0.633744 | 0.78 | 1057.0 |
| 8 | 2020-01-23 | 0.685211 | 0.78 | 808.0 |
| 9 | 2020-01-24 | 0.639298 | 0.78 | 287.0 |
| 10 | 2020-01-27 | 0.662274 | 0.78 | 899.0 |
| 12 | 2020-01-29 | 0.637804 | 0.87 | 1095.0 |
| 13 | 2020-01-30 | 0.713618 | 0.87 | 787.0 |
| 14 | 2020-01-31 | 0.673412 | 0.78 | 761.0 |
| 15 | 2020-02-03 | 0.682665 | 0.87 | 959.0 |
| 16 | 2020-02-04 | 0.679798 | 0.78 | 991.0 |
| 18 | 2020-02-06 | 0.678259 | 0.78 | 921.0 |
| 21 | 2020-02-11 | 0.660712 | 0.78 | 676.0 |
| 22 | 2020-02-13 | 0.643012 | 0.78 | 832.0 |
| 23 | 2020-02-14 | 0.663220 | 0.87 | 1505.0 |

(b) Arranques Grupo 2 Fase 2

Figura 4.31: Grupos de Arranques en Fase 2

No obstante, con la ayuda de la tabla resumen de los estadísticos de la fase 2 [A.4], se establece un umbral de la duración de esta fase debido a la excesiva variabilidad de la misma. Por ello, se determinó como cota superior 1510 segundos. Esta decisión llevó a tratar como anómalo el arranque del día 20 de Enero en la que la duración de dicha fase duró 1990 segundos. Por tanto, se elimina del estudio este arranque, quedando 20 arranques característicos.

Con objeto de representar las diferencias entre los dos grupos comentados se exponen en escalas ajustadas al inicio y fin de la fase 2 con un arranque de cada grupo. El resto se encuentra en el apartado de Anexos [A.2.4]. Como se observa, la duración no será influyente en el análisis ya que como se comentó anteriormente, se interpolará la duración de la Fase 2 de arranque a la misma duración (95 segundos).




(a) Arranque Fase2 Día 10 Febrero

(b) Arranque Fase 2 Día 13 Enero

Figura 4.32: Arranques de ejemplo en Fase 2

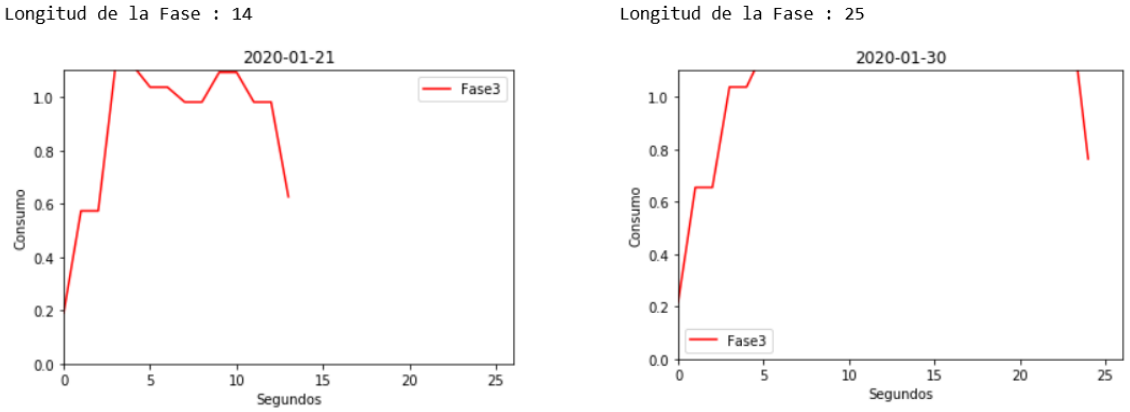
**Fase 3**

La fase 3 se caracteriza por alcanzar un pico de consumo y volver a valores de consumo que oscilan en torno a un valor con poca variabilidad. En esta fase sí se considera relevante la duración, ya que se pueden diferenciar grupos con menor y mayor duración. Esto tiene relación con alcanzar uno o dos máximos locales durante el transcurso de la Fase 3 de arranque.
A continuación se exponen los valores medios para cada una de las tres variables características(Media,Máximo y Duración):

| | Media de la media | Media del Máximo | Duracion Media |
|---|---|---|---|
| 0 | 3.577355 | 5.35375 | 25.625 |

Figura 4.33: Valor medio del máximo, media y duración de Fase 3 de los 20 arranques

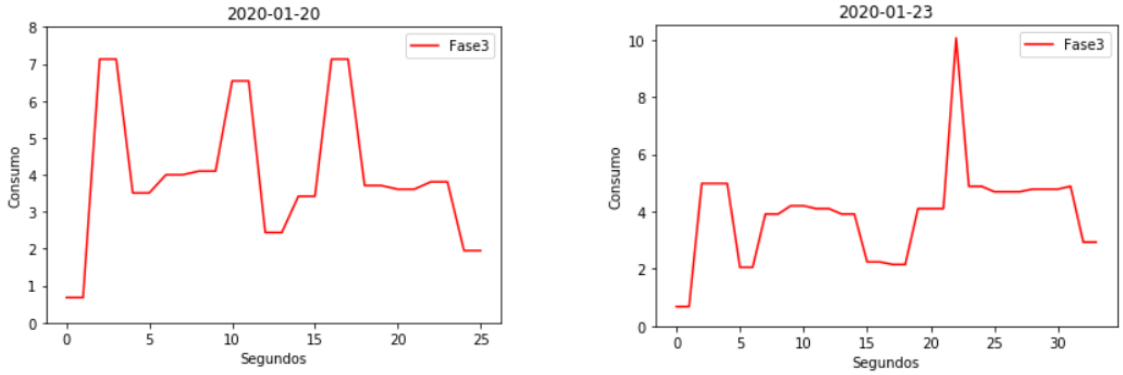
A continuación se exponen dos ejemplos[4.34a] y [4.34b] de dos arranques durante la fase 3 con mismas escalas normalizadas, siendo para la escala Y el consumo medio(3.57) y para la escala X la duración media de 25 segundos. Gracias a ello, se puede determinar de modo genérico las grandes diferencias entre los arranques como ocurre entre estos dos arranques, en la que uno de ellos tiene valores bajos de consumo con respecto a la media y una duración de fase mucho menor. Mientras el otro arranque todo lo contrario, 'mostrándose' únicamente los primero segundos de la fase por un consumo superior a la media normalizada.




(a) Arranque Fase 3 21 Enero (b) Arranque Fase 3 30 Enero

Figura 4.34: Arranques normalizados en Fase 3

Con la realización de estas gráficas normalizadas junto con la ayuda de la tabla [A.5] se puede ir agrupando en base a similitudes de los distintos arranques de acuerdo a las características de esta fase. Además, también se observan comportamientos anómalos en algunos arranques ya desestimados en las anteriores fases o aún existentes en los 20 arranques restantes. Ejemplos de ello son las gráficas [4.35a] y [4.35b] en la que la primera se observa un comportamiento anómalo con respecto al resto de arranques alcanzando tres picos de consumo, mientras que en el segundo el consumo se dispara hasta los 10 Amperios, además de tener un comportamiento inicial anómalo. Por ello, de aquí en adelante, con la eliminación del arranque correspondiente al 23 de Enero(20 de Enero ya fue desestimado), se obtienen 19 arranques con los que seguir trabajando.




(a) Arranque Fase3 - 20 Enero Anómalo (b) Arranque Fase 3 - 23 Enero Anómalo

Figura 4.35: Arranques Anómalos en Fase 3

Por otro lado, las agrupaciones de arranques a priori correctos, se encuentran en las tablas siguientes en las que se encuentran el día y las características que han llevado a agrupar los arranques a un grupo u otro.

| | Dia | Fase3 Media | Fase3 Max | Fase3 Duracion |
|---|---|---|---|---|
| 3 | 2020-01-16 | 3.336250 | 4.10 | 15.0 |
| 4 | 2020-01-17 | 2.940625 | 4.00 | 15.0 |
| 6 | 2020-01-21 | 2.847500 | 4.00 | 15.0 |
| 7 | 2020-01-22 | 3.405000 | 4.59 | 15.0 |
| 10 | 2020-01-27 | 3.068750 | 4.39 | 15.0 |
| 17 | 2020-02-05 | 3.391250 | 4.39 | 15.0 |
| 19 | 2020-02-07 | 3.335000 | 4.10 | 15.0 |
| 20 | 2020-02-10 | 3.325714 | 4.10 | 13.0 |
| 23 | 2020-02-14 | 3.032500 | 4.20 | 15.0 |

(a) Estadísticos Grupo 1 Fase 3

| | Dia | Fase3 Media | Fase3 Max | Fase3 Duracion |
|---|---|---|---|---|
| 0 | 2020-01-13 | 3.108750 | 5.08 | 39.0 |
| 14 | 2020-01-31 | 3.529091 | 5.08 | 32.0 |
| 15 | 2020-02-03 | 3.239787 | 5.27 | 46.0 |
| 16 | 2020-02-04 | 3.338947 | 4.78 | 37.0 |
| 22 | 2020-02-13 | 3.117742 | 4.20 | 30.0 |

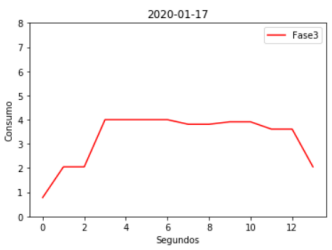
(b) Estadísticos Grupo 2 Fase 3

| | Dia | Fase3 Media | Fase3 Max | Fase3 Duracion |
|---|---|---|---|---|
| 9 | 2020-01-24 | 3.833667 | 8.30 | 29.0 |
| 12 | 2020-01-29 | 3.683030 | 7.82 | 32.0 |
| 13 | 2020-01-30 | 3.934444 | 6.15 | 26.0 |
| 18 | 2020-02-06 | 3.591212 | 5.76 | 32.0 |
| 21 | 2020-02-11 | 3.971481 | 5.96 | 26.0 |

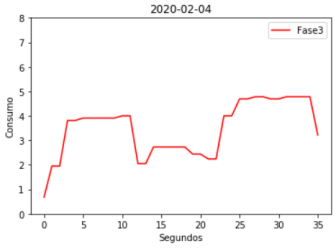
(c) Estadísticos Grupo 3 Fase 3

Figura 4.36: Estadísticos Grupos 1,2 y 3 en Fase 3

A continuación se expone de manera gráfica un ejemplo de arranque de cada uno de los grupos comentados. El resto de arranques se encuentra en la sección de Fase 3 de Anexos [A.2.5].

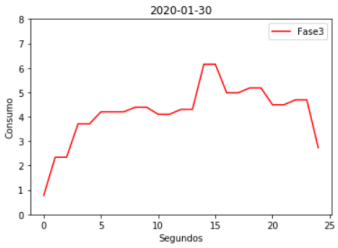



(a) Ejemplo Grupo 1 Fase 3




d

(b) Ejemplo Grupo 2 Fase 3




(c) Ejemplo Grupo 3 Fase 3

Figura 4.37: Ejemplos de arranques de los tres grupos hallados

| | Media Media | Media Máximo | Media Duracion |
|---|---|---|---|
| 0 | 3.186954 | 4.207778 | 14.777778 |

(a) Estadísticos Medios Grupo 1

| | Media Media | Media Máximo | Media Duracion |
|---|---|---|---|
| 0 | 3.266863 | 4.882 | 36.8 |

(b) Estadísticos Medios Grupo 2

| | Media Media | Media Máximo | Media Duracion |
|---|---|---|---|
| 0 | 3.802767 | 6.798 | 29.0 |

(c) Estadísticos Medios Grupo 3

Figura 4.38: Estadísticos medios de los tres grupos de arranque

Las características de los tres grupos se resumen en la siguiente tabla:

| **FASE 3** | **Características** |
|---|---|
| **Grupo 1** | - Duración corta de fase menor de 16 segundos.<br>- Un pico de consumo.<br>- Valores de consumo medio menores que el resto de grupos. |
| **Grupo 2** | - Duración larga de fase en torno a 37 segundos.<br>- Dos picos de consumo similares. |
| **Grupo 3** | - Duración media de fase en torno a 30 segundos.<br>- Dos picos de consumo, el segundo mayor. |

Cuadro 4.4: Características grupos en Fase 3

### Fase 4

Para la fase 4 de arranque, como comentamos anteriormente, no se va a proceder a realizar una clasificación ya que no sabemos con certeza cuando acaba esta fase, y por tanto cuando termina un arranque. De momento sólo se utilizan los primeros 60 segundos de esta fase dónde los valores de consumo se encuentran estables entre 2 y 3.5 Amperios para todos los arranques. Sin embargo, en el capítulo 5 se verá como influye el valor medio de consumo de la Fase 4 del arranque.

### Agrupación del arranque global

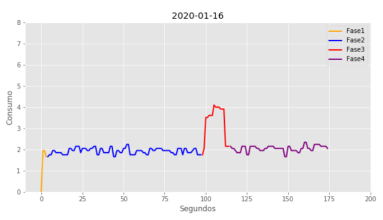
En base a las agrupaciones de los arranques realizadas para cada una de las fases se aplica el mismo procedimiento observando que arranques comparten similitudes entre fases. Con ayuda de las gráficas normalizadas, las gráficas aportadas en Anexos, junto con las tablas con las estadísticas de consumo y duración, ha sido posible realizar una serie de agrupaciones acorde al número de arranques con los que se ha trabajado.

Hay que destacar que los grupos de arranque 2,3 y 4 que veremos a continuación se han interpolado a la misma duración de la fase 2 con el objetivo de ver las características de todas las fases por igual. Se ha determinado que el tiempo superior a los 95 segundos, realmente es tiempo de espera que no aporta información al arranque. Por ello se han ajustado a la misma duración para todos los arranques. El consumo medio permanece estable en torno a unos valores estables (0.58 y 0.78 Amperios) con alguna pequeña desviación que no se ha estimado influyente en el arranque.

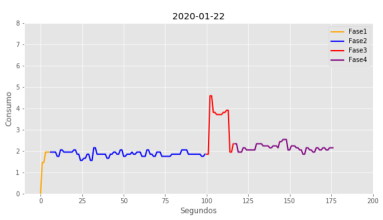
A continuación se detallan las especificaciones de cada uno de los grupos formados:

### Grupo 1

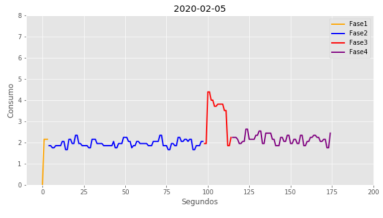
Como se ha visto en las anteriores agrupaciones, existe un grupo característico durante la mayor parte del arranque. Se trata de los arranques correspondientes con los días 16 Enero,22Enero,5 Febrero,7Febrero y 10 de Febrero.



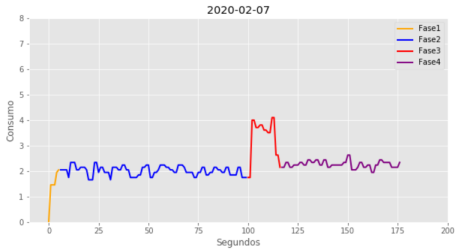

(a) Arranque 16 Enero Grupo 1



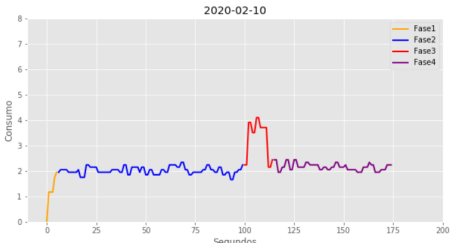

(b) Arranque 22 Enero Grupo 1




(c) Arranque 5 Febrero Grupo 1




(a) Arranque 7 Febrero Grupo 1




(b) Arranque 10 Febrero Grupo 1

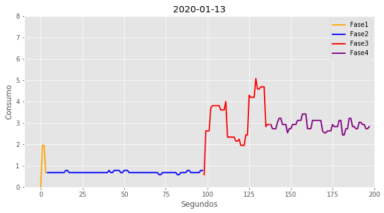
Figura 4.40: Arranques en Grupo 1

Las características de cada fase para este grupo de arranque se resumen en la siguiente tabla:

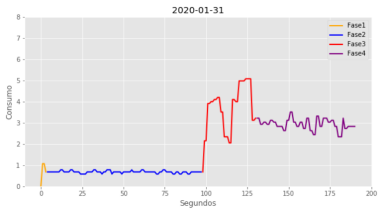
| **Grupo 1** | **Características** |
|---|---|
| **Fase 1** | - Pico Consumo en torno a 2 Amperios<br>- Duración 3 o 5 Segundos |
| **Fase 2** | - Estabilización consumo 1.8-2.2 Amperios |
| **Fase 3** | - Duración corta menor de 16 segundos<br>- Un pico de consumo en torno 4-4.5 Amperios |
| **Fase 4** | - Estabilización consumo 2-2.5 Amperios |

**Grupo 2**

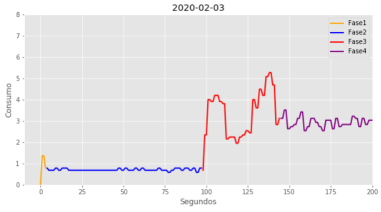
La agrupación realizada para este grupo trata de los días 13,31 Enero, 3 Febrero,4 Febrero,6 Febrero y 13 Febrero. A continuación se encuentran las series temporales asociadas a dichos días con mismas escalas de tiempo y de consumo:



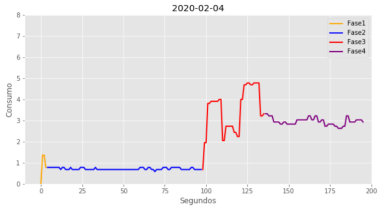

(a) Arranque 13 Enero Grupo 2



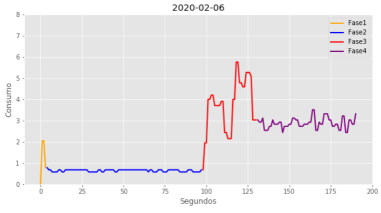

(b) Arranque 31 Enero Grupo 2



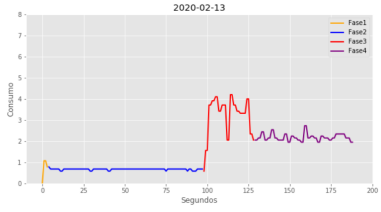

(c) Arranque 3 Febrero Grupo 2




(a) Arranque 4 Febrero Grupo 2




(b) Arranque 6 Febrero Grupo 2




(c) Arranque 13 Febrero Grupo 2

Figura 4.42: Arranques en Grupo 2

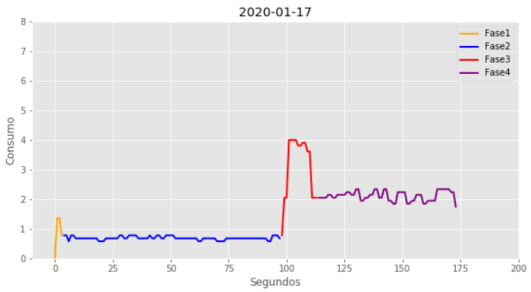
Las características de las distintas fases para este grupo son:

| Grupo 2 | Características |
|---|---|
| **Fase 1** | - Pico Consumo entre 1-2 Amperios<br>- Duración 3 Segundos |
| **Fase 2** | - Estabilización consumo 0.58-0-78 Amperios con ligeras desviaciones |
| **Fase 3** | - Duración larga de fase hasta los 42 segundos<br>- Decrecimiento del consumo a los 12-14 segundos hasta valores de 2 Amperios<br>- Primer pico de consumo 4-4.5 Amperios<br>- Segundo pico de consumo sobre los 5 Amperios |
| **Fase 4** | - Estabilización consumo 2.5-3.5 Amperios |

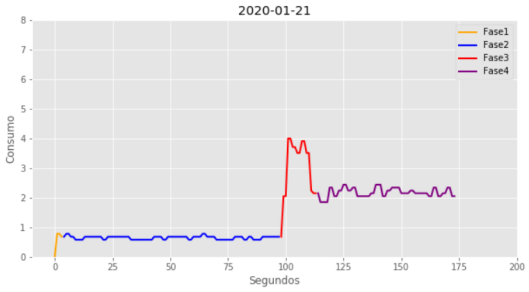
No obstante, el comportamiento de esta fase para el día 13 de Febrero es algo diferente al resto con respecto a la fase 3, no alcanzando un segundo pico de corriente mayor que el primero, sino alcanzando valor similar al primer pico de corriente y con valores medios en la segunda parte de la Fase 3 menores que la media. Además, se detectan unos valores de consumos más pequeños en la Fase 4.

### Grupo 3

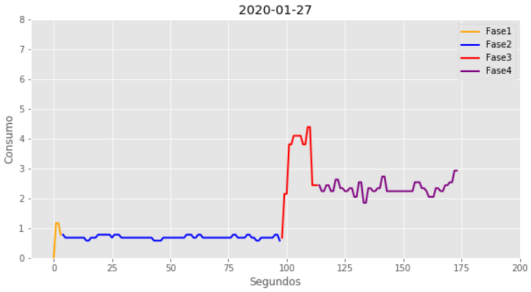
Los arranques asociados a este grupo pertenecen a los días 17 Enero, 21 Enero, 27 de Enero y 14 de Febrero. Los arranques asociados son:



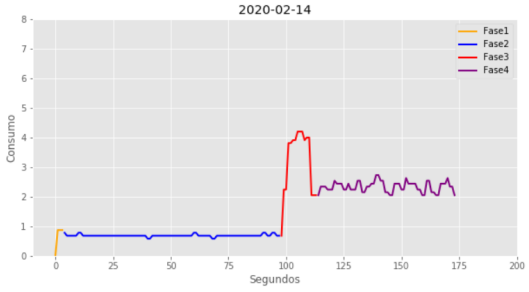

(a) Arranque 17 Enero Grupo 3




(b) Arranque 21 Enero Grupo 3




(a) Arranque 27 Enero Grupo 3




(b) Arranque 14 Febrero Grupo 3

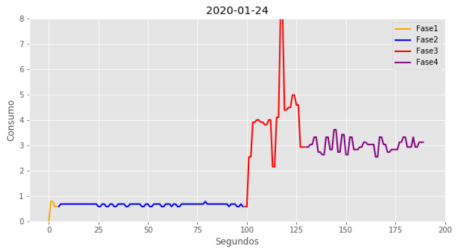
Figura 4.44: Arranques en Grupo 3

Las características de las cuatro fases para este grupo son las siguientes:

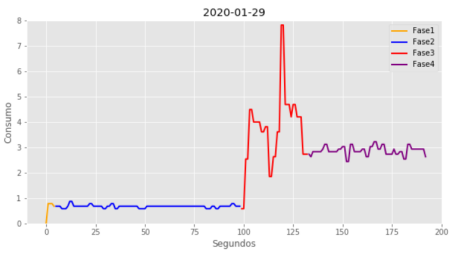
| Grupo 3 | Características |
|---|---|
| **Fase 1** | - Pico Consumo no supera 1.4 Amperios<br>- Duración 3 Segundos |
| **Fase 2** | - Estabilización consumo 0.58-0-78 Amperios con ligeras desviaciones |
| **Fase 3** | - Duración corta de fase menor de 16 segundos<br>- Pico de consumo 4-4.5 Amperios |
| **Fase 4** | - Estabilización consumo 2-2.5 Amperios |

### Grupo 4

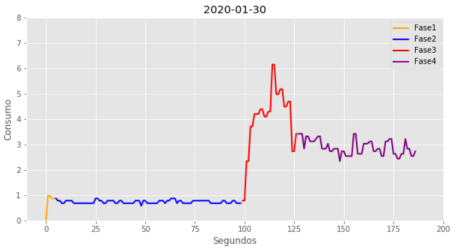
Se ha realizado una cuarta agrupación que consta de los arranques correspondientes a los días 24,29,30 Enero y 14 Febrero. A continuación se observan de manera gráfica:



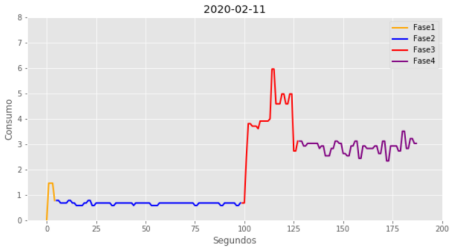

(a) Arranque 24 Enero Grupo 4




(b) Arranque 29 Enero Grupo 4




(a) Arranque 30 Enero Grupo 4




(b) Arranque 11 Febrero Grupo 4

Figura 4.46: Arranques en Grupo 4

Las características de este grupo para las distintas fases son:

| **Grupo 4** | **Características** |
|---|---|
| **Fase 1** | - Pico Consumo menor de 1 Amperio<br>- Duración 3 Segundos |
| **Fase 2** | - Estabilización consumo 0.58-0-78 Amperios con ligeras desviaciones |
| **Fase 3** | - Duración media de fase en torno a 30 segundos<br>- Valores de consumo medios cercanos a 4 Amperios<br>- Primer pico de consumo 4-4.5 Amperios<br>- Segundo pico de consumo 6-8 Amperios |
| **Fase 4** | - Estabilización consumo 2.5-3.5 Amperios |

No obstante, el patrón de comportamiento de la serie **en la fase 3** así como el pico de consumo, tienen características similares dos a dos, es decir, los comportamiento del día 24 de Enero y 29 de Enero son muy similares al igual que los del 30 de Enero y 11 Febrero. Sobretodo destaca el valor de consumo que alcanzan los días 24 y 29 de Enero en torno a los 8 Amperios. Como ya se explicó anteriormente, fue desestimado un arranque con un consumo superior a los 10 Amperios por lo que se determinó como límite los valores de consumo que alcanzan estos arranques. Este puede ser uno de los problemas de no tener información a priori sobre lo que es correcto o no, que puede ocurrir que los valores de consumo de estos arranques también deban considerarse como anómalos.

## 4.5 Arquitectura del sistema de aprendizaje

En los últimos años, las tecnologías de aprendizaje automático como el Machine Learning han emergido con una gran importancia en el mundo de los negocios, ya que el uso inteligente de las analíticas de datos es clave para el éxito empresarial.

El objetivo de este trabajo se basa en la identificación de patrones o tendencias que se 'esconden' en los datos, para crear un modelo que nos permita explicar comportamientos determinados. Para ello, es vital el entrenamiento del modelo con gran cantidad de datos e información con objetivo de que el modelo aprenda y sea capaz de hacer predicciones [33]. Existen diferentes procedimientos para utilizar Machine Learning según el objetivo del estudio [33]:

1. El problema será asociado a **clasificación** si tenemos como objetivo clasificar en grupos atendiendo a unas etiquetas definidas. La variable objetivo es de tipo categórico.

2. El problema será de **regresión** si tiene como objetivo predecir valores continuos a partir de unas etiquetas definidas. La variable objetivo es de tipo numérico.

3. El problema será de **clustering** si queremos agrupar conjuntos en base a su similaridad sin la existencia de etiquetas definidas. No existe variable objetivo definida.

Decimos que un algoritmo se usa para **aprendizaje supervisado** cuando permite realizar predicciones futuras en base a datos etiquetados predefinidos. Esta etiqueta es la respuesta a unos datos pasados. Se clasifican en problemas de *clasificación o de regresión*.

Por otra parte, la utilización del **aprendizaje no supervisado** está formado por un conjunto de algoritmos que se aplican sin necesidad de tener datos etiquetados del pasado. No existe un conocimiento a priori de los datos.

En la Figura [4.47] se adjunta un ejemplo ilustrativo sobre la distinción en la metodología que utiliza el aprendizaje supervisado y no supervisado:


Figura 4.47: Aprendizaje Supervisado vs No Supervisado
*Fuente [22]*

De acuerdo al objetivo planteado en este trabajo, como es la detección del arranque más representativo (por tanto correcto) y la detección de arranques anómalos, se ha considerado como método más adecuado el uso de un aprendizaje no supervisado. No existe una variable/etiqueta que nos determine si un arranque es 'correcto' o no, ya que es lo que habrá que determinar.

### 4.5.1 Aprendizaje no Supervisado

El método de análisis del aprendizaje no supervisado o del clustering ha sido ampliamente estudiado en la minería de datos y en el aprendizaje automático debido a sus numerosas aplicaciones. En ausencia de información etiquetada, el análisis cluster puede considerarse como un modelo concreto de datos que permite resumir los datos[8]. El problema básico del clustering se puede describir como :

*'Dado un conjunto de puntos, dividirlo en conjuntos de grupos lo más similares posible'*[8].

Si es necesario predecir un resultado dicotómico (variable binaria) a raíz de un conjunto de datos, existen muchos algoritmos para llevarlo a cabo como regresión logística, arboles de decisión o máquinas de soporte vectorial que son agrupados dentro del aprendizaje supervisado [20].

Sin embargo, el aprendizaje no supervisado tiende a ser más subjetivo. No hay un objetivo definido para el análisis como es la predicción de una respuesta. Puede ser difícil evaluar los resultados obtenidos de un método de aprendizaje no supervisado, ya que no existen mecanismos universales para validarlo como puede ser la validación cruzada. Es a menudo realizado como parte de un análisis exploratorio de los datos [8].

En definitiva, el objetivo del aprendizaje no supervisado será estudiar la estructura de los datos y dar forma a la misma. Para ello existen dos técnicas generales para llevarlo a cabo: Reducción de dimensionalidad y agrupación o clustering. Para este trabajo nos vamos a centrar en el segundo, el cual se explica a continuación.

Debido al crecimiento de la cantidad de datos en los últimos años, existe una mayor dificultad a la hora del etiquetado de manera manual. El análisis clúster es una de las técnicas de etiquetado más importantes. En este análisis se tienen datos sin etiquetar y el objetivo es agruparlos en base a sus similitudes. Por lo general, no se saben ni el número del grupo ni de la descripción de los mismos, lo cual lo hace más 'desafiante' [8]. Algunas de las aplicaciones asociadas al clustering pueden ser el reconocimiento o clasificación de patrones, segmentación de imágenes...etc.

### 4.5.2 Metodología y Algoritmos

Dentro de la metología del clustering existen distintas formas de aplicación. De manera genérica pueden clasificarse en métodos de partición, métodos jerárquicos y métodos basados en densidad. Los métodos basados en partición usan métricas basadas en distancias para observar similitudes entre observaciones. Ejemplo más típico de este tipo es el algoritmo de K-Means. En cuánto a los algoritmos jerárquicos dividen los datos en diferentes niveles estableciendo jerarquías, ayudando a visualizar y resumir los datos. En la práctica se utilizarán los métodos de KMeans y el algoritmo jerárquico de enlace completo.

**Hierarchical Clustering(Clustering Jerárquico)**

Este tipo de algoritmos fueron desarrollados con el objetivo de superar las desventajas que proporcionaban los algoritmos de partición. Este tipo de algoritmos (de partición) necesitan indicarle a priori un número de clusters, lo cual no se necesita en los algoritmos jerárquicos [8].

Los métodos jerárquicos se distinguen entre aglomerativos y divisivos. La diferencia radica en que los primeros comienzan con un cluster por cada observación en el nivel más bajo, los cuales van agrupándose en clusters con más observaciones hasta obtener un cluster que agrupe a todas las observaciones en el nivel más alto. Exactamente lo contrario ocurre con los algoritmos divisivos, dónde comienzan con un cluster grande con todas las observaciones y se van separando hasta tener un cluster por observación. Para indicar de manera esquemática lo comentado, se indica a continuación un ejemplo:

| **Método Aglomerativo** | **Método Divisivo** |
|---|---|
| • $P_0$={{1},{2},{3},{4}} | $P_0$={{1,2,3,4}} |
| • $P_1$={{1},{2},{3,4}} | $P_1$={{1},{2,3,4}} |
| • $P_2$={{1},{2,3,4}} | $P_2$={{1},{2},{3,4}} |
| • $P_3$={{1,2,3,4}} | $P_3$={{1},{2},{3},{4}} |

Algunas de las características que los diferencian son:

- En general,los aglomerativos son más rápidos
- Los divisivos no tienen por qué acabar en clusters con una sola muestra
- Ambos dos, son eficientes con pocas muestras

De aquí en adelante nos vamos a centrar en los algoritmos de clustering aglomerativos, ya que son los que se utilizarán en el trabajo.

La principal ventaja que proporcionan los algoritmos de clustering aglomerativos es que no necesitan a priori un número de clusters para agrupar, ya que es el objetivo de este tipo de algoritmos. Para ello, será muy útil la utilización de los dendogramas, que es una representación 2D de un árbol binario que va agrupando las observaciones en base a sus similitudes. La raíz del árbol representa todo el conjunto de las observaciones agrupadas(nivel 0). A raíz de este nivel, se van formando 'ramas' que van agrupando observaciones de manera específica. Cada nivel de jerarquía representa un número de observaciones agrupadas, siendo el último nivel de jerarquía el que cada observación es representada por un grupo.

Para ir formándose las 'ramas' del dendograma , se necesita una matriz de similaridad o comúnmente llamada 'matriz de linkage', la cual se formará dependiendio del criterio de enlace a utilizar. En la práctica se utilizará uno de los métodos más populares y eficientes como es 'complete linkage' o enlace completo [8].

La primera fase de un algoritmo de cluster jerárquico pasa por utilizar una medida de proximidad para formar una matriz de similaridad o 'linkage' en la que todas las observaciones se encuentran en el nivel más bajo del dendrograma.
Los conjuntos que más similares sean se irán agrupando y por tanto actualizando la matriz de linkage. Esto se realizará hasta que en un cluster agrupe a todas las observaciones iniciales. En cada agrupación se calculará un valor que explicará la altura del eje y del dendograma. Para formar la matriz de 'linkage' dependerá de las medidas y del criterio de enlace utilizadas [8].

Como ejemplo, se observa la diferencia entre los métodos de enlace simple y completo:

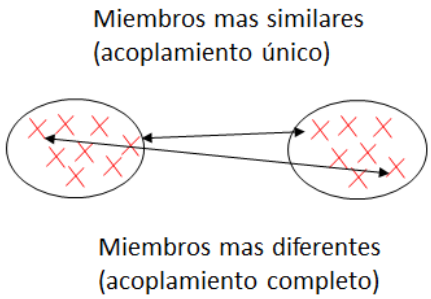



Figura 4.48: Enlace simple vs Enlace completo

El método que se ha utilizado para este trabajo es de enlace completo ó complete linkage. Este método basa sus agrupaciones en base a las similitudes. En concreto, compara los puntos más diferentes de dos grupos para agruparlos. En comparación al método de enlace simple, es menos sensible al ruido y a los outliers [8].

Los pasos a tomar para llevarlo a cabo, dada una matriz de distancias, son:

1. Para cada observación formar un cluster:

$$P_0 = \{\{1\}, \{2\}.., \{m\}\} \tag{4.1}$$

2. Selección de los dos nodos más cercanos, tal que :

$$d(i, j) = min_{k,l} d(k, l) \tag{4.2}$$

3. Formado un cluster con esos dos puntos, se reduce la matriz de distancia inicial sustituyendo las filas y columnas (i,j) por una, cuyos elementos serán:
d(l,{i,j})=max(d(i,l),d(j,l))

4. Con la matriz reducida resultante, aplicar los pasos anteriores hasta que la matriz no pueda reducirse mas.

5. La jerarquía indexada asociada a cada nivel formado será el valor máximo entre d(i,l) o d(j,l).

**Algoritmo KMeans**

Este tipo de algoritmo se encuentra dentro de un grupo de algoritmos de clustering particionales. Es uno de los algoritmos más eficientes y más usados dentro de este tipo de algoritmos

de clustering, que tienen la particularidad de necesitar saber un número de clusters para utilizar el algoritmo. La aplicación de este algoritmo nos va a ayudar a confirmar o desestimar nuestra hipótesis sobre las distintas agrupaciones realizadas de los arranques [8].

Este algoritmo comienza seleccionando K puntos representativos como los centroides iniciales. Estos centroides contienen unas coordenadas para cada uno de los K grupos. Cada punto o observación (en nuestro caso arranque) se asigna al centroide más cercano aplicando medidas de proximidad. Una vez que se forman los grupos,los centroides de cada grupo se irán actualizando de acuerdo a los puntos que se van añadiendo al grupo en cuestión. Este proceso se repite hasta que el algoritmo no converja más y los centroides no varíen.

El procediminto explicado de una forma estructural se puede explicar como [8]:

1. Seleccionar un número K de centroides, los cuales se asignarán de manera aleatoria en el sistema de coordenadas.

2. Asignar cada observación de nuestro conjunto de datos al centroide más cercano basándonos en la distancia euclídea(por ejemplo).

3. Los centroides de cada grupo son recalculados, aplicando una media de la posición de todos los puntos de cada grupo

4. Se repiten los pasos 2 y 3 hasta que el algoritmo converja: los centroides no cambien,número máximo de iteraciones o la suma de los cuadrados del error(SSE) se minimiza.

Dado un dataset $D = \{x_1, x_2....x_N\}$ correspondiente a N puntos, denotar el clustering obtenido después de aplicar KMeans por $C = \{C_1, C_2...C_k..C_K\}$. El SSE para este clustering se define:

$$\text{SSE}(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - c_k\|^2 \tag{4.3}$$

$$c_k = \frac{\sum_{x_i \in C_k} x_i}{|C_k|} \tag{4.4}$$

En las gráficas [4.49] se observa un ejemplo de la evolución del procedimiento de KMeans. En la primera gráfica tenemos n puntos distribuidos sobre un plano en dos dimensiones. En la segunda gráfica se asignan dos centroides aleatorios en unas coordenadas aleatorias. En el tercer gráfico se 'acercan' los n puntos al centroide más cercano formando dos clusters iniciales coloreados en azul y rojo.
De aquí en adelante, gráficas 4,5 y 6, se recalculan los centroides en base al nuevo cluster y de nuevo se 'acercan' los n puntos al nuevo centroide más cercano. Este proceso se realiza hasta que se minimice el SSE o el centroide no se modifique en cada iteración.

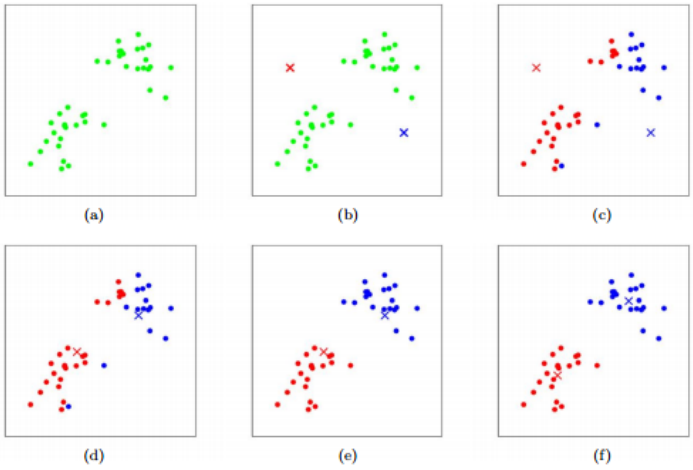



Figura 4.49: Fases KMeans *Fuente: [29]*

Los factores que más pueden influir en cuánto al resultado de este algoritmo son la elección inicial de los centroides y sobretodo la estimación del número de clusters a utilizar. En este trabajo se ha utilizado el **método del codo** para determinar un número de clusters óptimo a utilizar en los algoritmos. El método Elbow( o método del codo) es un método de interpretación y validación utilizado para ayudar a determinar un número de grupos apropiado.

Este método utiliza la suma de cuadrados intra-cluster(SSW) en función del número de clusters. Es decir, si escogemos K clusters, para cada uno de los K clusters se calcula el SSW. De acuerdo a este cálculo se realiza un gráfico del valor de SSW frente al número de clusters. Se escoge el 'codo' dónde ha habido un descenso de la varianza más pronunciado. Esto se corresponde con elegir un número de grupos más apropiado.
Es cierto que al aumentar el número de grupos mejorará el ajuste pero esto puede llevar al sobre-ajuste.La idea es recoger con el mínimo número de clusters la mayor información posible y esta información es la que intenta reflejar el 'codo'.

## 4.6 Clasificación de los arranques

En esta sección se va a aplicar dos tipos de clustering a una serie de variables características. En concreto, estas variables corresponden a los valores de consumo medios o máximos así como la duración en el tiempo para las distintas fases [Tablas A.3, A.4 y A.5]. Esto se realiza con el objetivo de caracterizar los distintos grupos de arranque de la fresadora en base a sus similitudes.

En primera instancia se detallará el procedimiento llevado a cabo aplicando análisis jerárquico, concretamente complete linkage, y en segunda instancia el algoritmo de KMeans.

Con la realización de ambos algoritmos, se pretende que exista una cierta semejanza con las hipótesis planteadas en el apartado anterior. Al final de esta sección se realizará una breve comparación entre los resultados de ambos algoritmos. Además, gracias a la aplicación de dos algoritmos distintos podremos contrastar la hipótesis planteada con más confianza al tener dos resultados de métodos estadísticos distintos.

La utilización de clustering jerárquico como de no jerárquico nos proporciona dos modos distintos de como abordar el problema de la agrupación. Aplicando clustering jerárquico no es necesario

fijar un número de clusters a priori, sino que se fijan por sí solos. Por ello este método es útil en análisis exploratorio observando que arranques se asemejan con otros. Sin embargo, el clustering no jerárquico categorizan los arranques según un número de clusters a priori. Para ambos métodos, aunque sea más útil para clustering no jerárquico, se utilizará el método del codo para poder ayudarse para determinar el numero de clusters óptimo.

La aplicación de los algoritmos de clustering se ha realizado para 19 de los 24 arranques iniciales. En la sección anterior se comentó como se desestimaron los arranques correspondientes a los días 14,15,20,23 y 28 de Enero, tanto por consumos no habituales como comportamientos de la serie no habituales en el resto de arranques.

Las características asociadas a los arranques usadas durante la realización de este trabajo son duración de la fase, al máximo o a la media de consumo de la misma para las 4 fases. No obstante, la fase 4 no se utilizará para el análisis debido a que no se tiene constancia de cuándo acaba el arranque de la máquina. Por tanto, se utilizarán inicialmente 9 variables para llevar acabo el clustering.

Debido a que no se utiliza la misma medida para duración o el máximo y la media de consumo para cada fase, se aplica una **normalización de estas variables para que no influya una variable más que otra**. El resultado de esta normalización se observa en la tabla adjunta en la que se observan todas las variables normalizadas[A.6].

Con objetivo de no utilizar variables que aportan misma información que otras, se utiliza la matriz de correlaciones. Los valores que toman en esta matriz van desde -1 a 1, significando ambos valores máxima correlación entre variables, es decir, que ambas variables ofrecen la misma información al modelo y por tanto existe una sobreparametrización.

Ejemplos de ello son la correlación muy alta(0.99) [4.50a] que existe entre la variable Fase 1 Max y la variable Fase 1 Media. Esto se debe a que al haber una duración similar en casi todos los arranques durante la fase 1, si un arranque tiene un valor de consumo alto, también lo tendrá en la variable fase 1 media al ser una tendencia similar en esta fase. Por tanto, se desestima la utilización de la variable *Fase 1 Media*. También se encuentran correlaciones altas en la tabla [4.50b] correspondiente con la fase 2 aunque no ocurre para la fase 3 [4.50c].

| | Fase1 Max | Fase1 Media | Fase1 Duracion |
|---|---|---|---|
| Fase1 Max | 1.000000 | 0.963277 | 0.496355 |
| Fase1 Media | 0.963277 | 1.000000 | 0.508299 |
| Fase1 Duracion | 0.496355 | 0.508299 | 1.000000 |

(a) Correlación variables Fase 1

| | Fase2 Media | Fase2 Max |
|---|---|---|
| Fase2 Media | 1.0000 | 0.9983 |
| Fase2 Max | 0.9983 | 1.0000 |

(b) Correlación variables Fase 2

| | Fase3 Media | Fase3 Max | Fase3 Duracion |
|---|---|---|---|
| Fase3 Media | 1.000000 | 0.663919 | 0.006441 |
| Fase3 Max | 0.663919 | 1.000000 | 0.495043 |
| Fase3 Duracion | 0.006441 | 0.495043 | 1.000000 |

(c) Correlación variables Fase 3

Figura 4.50: Matrices de correlación por fases

En las tabla [A.7] se encuentran las matrices de correlación utilizadas para hacer la clasificación del arranque global utilizando variables de todas las fases en la que no se encuentra ninguna autocorrelación muy alta.

### 4.6.1 Implementación de análisis jerárquico(complete linkage)

Con objeto de encontrar características comunes en las distintas fases, se ha realizado un análisis cluster jerárquico particular para cada una de las 3 fases.

Concretamente, se escogió utilizar el método de enlace completo, pero existen otros como puede ser enlace simple, método de ward ó método del centroide que se diferencian principalmente en el criterio de enlace. La elección de 'enlace completo' se debe a que ya conocía la existencia del mismo y había trabajado con él en la asignatura de 'Minería de Datos' y por ello estaba familiarizado con su aplicación. En ese sentido, quise optar por un método del que ya tenía conocimiento sobre él.

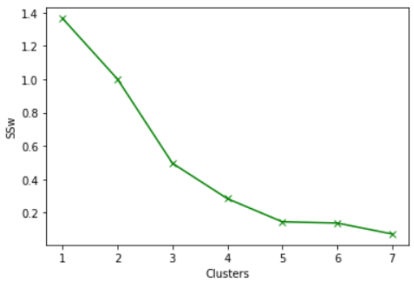
A continuación se aplicará este método de clustering a cada una de las fases y finalmente de manera global para todas las fases a la vez.

**Fase 1**

Como se comentó en la sección anterior, se ha utilizado el 'método del codo' para determinar cual es 'el número óptimo de clusters' a utilizar para agrupar los arranques. Se explican las componentes de la matriz linkage del ejemplo [4.51a] para esta Fase 1. Las dos primeras columnas hacen referencia a los arranques que se han agrupado, mientras que la tercera es el valor (SSW) utilizado para el procedimiento del método del codo. La cuarta columna corresponde al número de 'arranques' agrupados. Con cada iteración, se forman clusters con máyor número de arranques hasta obtener un cluster con el número total de arranques como se ve en la última fila, siendo el que más SSW tiene (1.35).
Siendo este método más utilizado para algoritmos que necesitan de un número óptimo de clusters a priori(como por ejemplo KMeans), se ha usado como complemento a la información ofrecida por el dendograma. La realización tanto de la matriz de linkage como de la gráfica se ha realizado en base a la función que se encuentra en Anexos en [A.1.4].

```
[[ 0.         1.         0.         2.        ]
 [ 2.        10.         0.         2.        ]
 [ 3.         5.         0.         2.        ]
 [ 4.        15.         0.         2.        ]
 [ 7.        21.         0.         3.        ]
 [ 9.        17.         0.         2.        ]
 [11.        20.         0.         3.        ]
 [18.        23.         0.06569343 4.        ]
 [16.        25.         0.0729927  4.        ]
 [13.        19.         0.0729927  3.        ]
 [ 6.        24.         0.0729927  3.        ]
 [14.        22.         0.0729927  3.        ]
 [ 8.        26.         0.13868613 5.        ]
 [12.        28.         0.1459854  4.        ]
 [27.        29.         0.28467153 7.        ]
 [31.        33.         0.49635036 12.       ]
 [32.        34.         1.        16.       ]
 [30.        35.         1.36357711 19.       ]]
```

(a) Ejemplo matriz linkage fase 1




(b) Método del codo fase 1

Figura 4.51: Matriz de linkage y método del codo

La gráfica del método del codo [4.51b] indica la agrupación en 5 clusters como la más 'óptima'. Con la realización del dendograma es posible observar los arranques que tienen más similaridad entre ellos y los que mas diferencia tienen. El código de ejemplo para la realización de los dendogramas para las distintas fases se encuentra en Anexos en [A.1.5]

El dendograma asociada a la fase 1 es el siguiente:

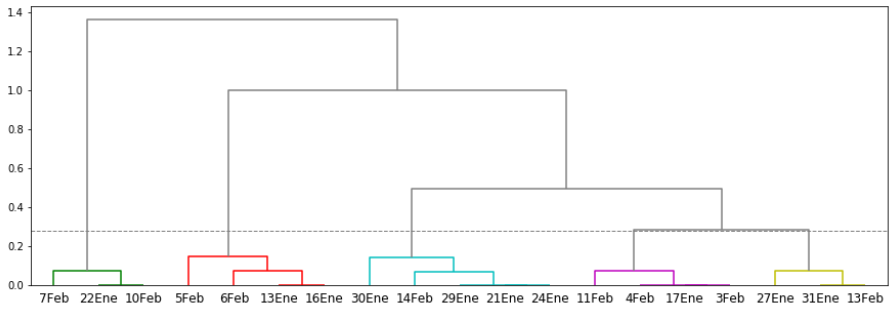



Figura 4.52: Dendograma en fase 1

Las agrupaciones que ofrece el dendograma en base a 5 clusters son:

- Grupo 1: 7 Febrero, 22 Enero y 10 Febrero
- Grupo 2: 5 Febrero,6 Febrero, 13 Enero y 16 Enero
- Grupo 3: 21 Enero,24 Enero,29 Enero, 30 Enero y 14 Febrero
- Grupo 4: 17 Enero,4 Febrero,11 Febrero
- Grupo 5: 27 Enero,31 Enero y 13 Febrero

Esta agrupación ha sido más específica a la realizada en la caracterización de fases 4.4.1 donde se caracterizaron 4 posibles grupos. El grupo 1 se caracteriza por una duración de la fase de 5 Segundos mientras el grupo 2, que tiene consumo máximo similar, la duración era de 3 segundos. Los otros tres grupos se han agrupado en base a los valores de consumo. El grupo 3 con valores de consumo máximos entre 0.78 y 0.97 Amperios, mientras que el grupo 4 y 5 se han distribuido en valores entre 1 amperio y 1.46 amperios.

**Fase 2**

El análisis cluster para la Fase 2 indica una agrupación muy clara de dos grupos diferenciados para el método del codo [4.53].

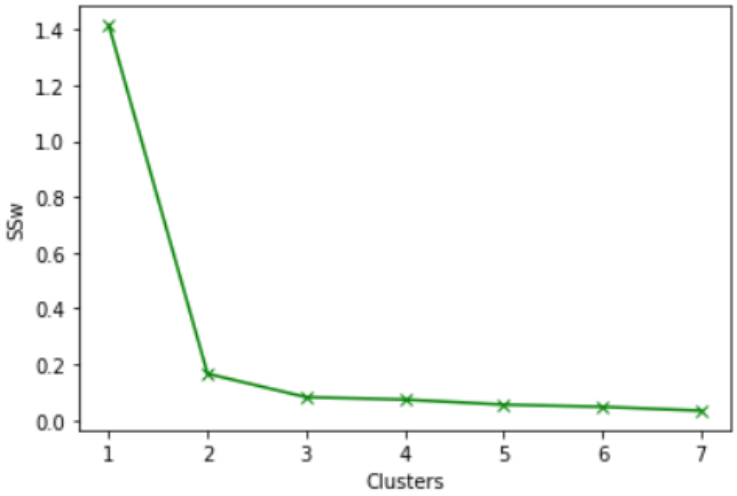



Figura 4.53: Método del codo en Fase 2

Al realizar el dendograma [4.54] se ha observado esta diferenciación clara de dos grupos en la que podemos diferenciar los arranques de cada grupo.

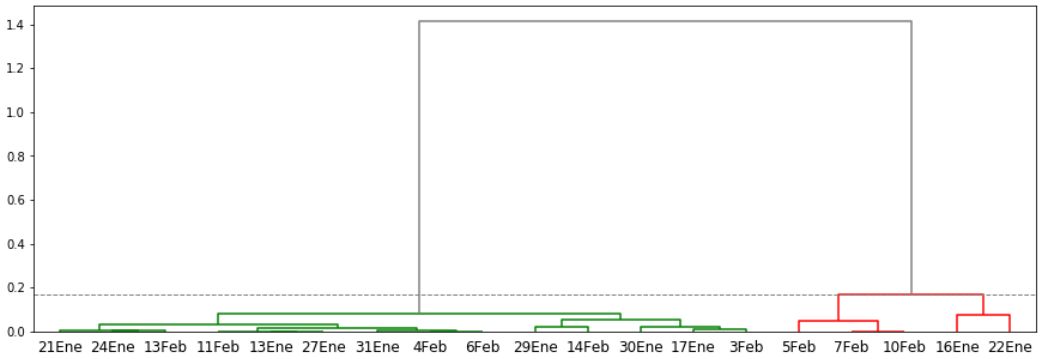



Figura 4.54: Dendograma en fase 2

La diferenciación entre los dos grupos radica en un consumo estable en torno a 2 Amperios en el grupo más pequeño formado por los días 16,22 de Enero y 5,7 y 10 de Febrero. El consumo medio aproximado del otro grupo característico es en torno a 0.5-0.8 Amperios. Este análisis sigue la línea de la asociación entre grupos en la fase 2, realizada antes de aplicar modelos de aprendizaje(sección 4.4.1).

**Fase 3**

El análisis realizado para la fase 3 se lleva a cabo por medio de las tres variables características de la fase: media y máximo del consumo y duración de la fase. Con el uso del procedimiento del 'método del codo', en la gráfica [4.55] nos permite determinar como son 3 los grupos óptimos para diferenciar los distintos arranques.

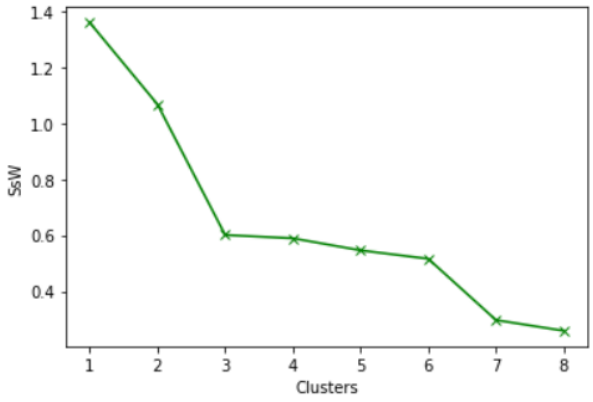



Figura 4.55: Método del codo en Fase 3

Las agrupaciones realizadas para tres clusters en el dendograma [4.56] son:

- Grupo 1: 24,29,30 Enero y 11 Febrero

- Grupo 2: 16,17,21,22,27 Enero y 5,7,10,14 Febrero
- Grupo 3: 13,31 Enero y 3,4,6 y 13 Febrero

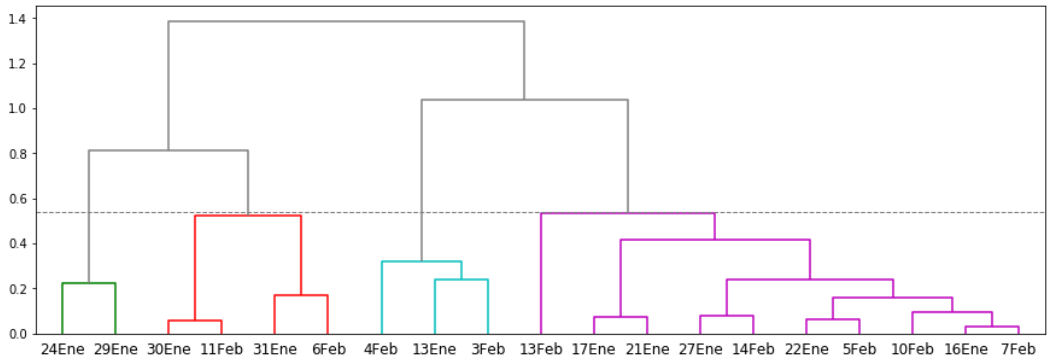



Figura 4.56: Dendograma en fase 3

Las características asociadas al grupo 1 es una duración superior a 20 segundos, consumo medio cercano o superior a 4 Amperios así como un pico de consumo mayor que la media, sobre 6 Amperios. En cuánto al grupo 2 es caracterizado por una duración de fase corta menor de 15 segundos con un consumo medio menor de 3.5 amperios, con un pico de consumo entre 4 y 4.5 Amperios. Las características del grupo 3 están basadas en una duración superior a los 20 segundos con un consumo medio y máximo similar al grupo 2. Además se caracterizan por alcanzar dos picos de consumo durante la Fase 3.

### Arranque global

Con el método del codo [4.57] se interpreta la partición en 4 clusters como el número óptimo de agrupaciones. También sería viable 7 clusters si se tiene una cantidad mayor de arranques y si se quiere realizar una clasificación más minuciosa.

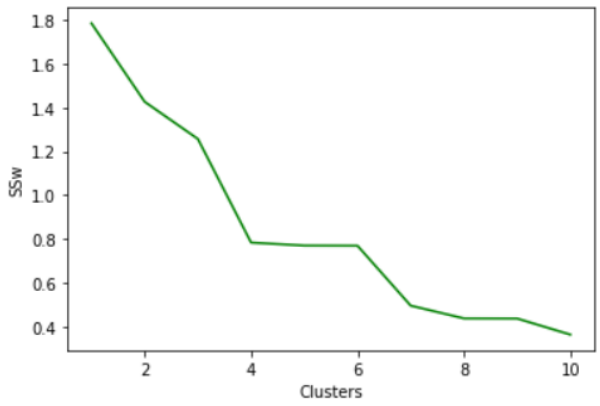



Figura 4.57: Método del codo arranque global - Complete linkage

Con la realización del dendograma [4.58], nos permite observar la similitudes entre los distintos arranques y determinar cuales componen los cuatro grupos característicos:

1. 16,22 Enero,5,7 y 10 Febrero
2. 24,29,30 Enero y 11 Febrero
3. 17,21,27,31 Enero, 13 y 14 Febrero
4. 13 Enero, 3,4 y 6 Febrero

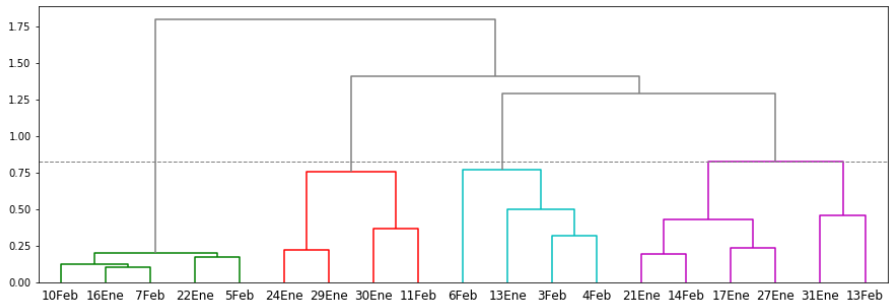



Figura 4.58: Dendograma clasificación final

En efecto, con la eliminación de la variable Fase 1 Duración, la clasificación ha sido realizada en base a 4 grupos característicos.

En el capítulo [5] se detallarán características comunes a los cuatro grupos de manera más específica para obtener patrones más generales entre los 4 clusters.

### 4.6.2 Implementación con algoritmo KMeans

El procedimiento llevado a cabo para determinar los grupos de arranques característicos en cada fase y globalmente difiere en parte a lo realizado con el algoritmo jerárquico, ya que se ha utilizado clustering no jerárquico.

Se ha estimado la utilización del algoritmo KMeans como ejemplo de algoritmo de clustering no jerárquico y como complemento a la información aportada por el algoritmo de enlace completo. Al no tener información sobre el número de clusters a priori, se aplica la utilización del método del codo.

La utilización del método del codo es mucho más eficiente en algoritmos de partición ya que necesitan un número a priori de agrupaciones para utilizar el algoritmo. En este tipo de algoritmos no se lleva a cabo la utilización del dendograma ya que este procedimiento es característico de algoritmos jerárquicos. Por tanto, con la información proporcionada por el método del codo, es asignado el número de clusters más representativo al algoritmo de KMeans, obteniendo un etiquetado de cada arranque de acuerdo al número de clusters indicado. Las funciones que han llevado a cabo la visualización del método del codo así como el ajuste de KMeans para obtener el correspondiente etiquetado se detallan en Anexos en [A.1.6] y [A.1.7].

**Fase 1**

A continuación se adjunta la gráfica del método del codo [4.59] :

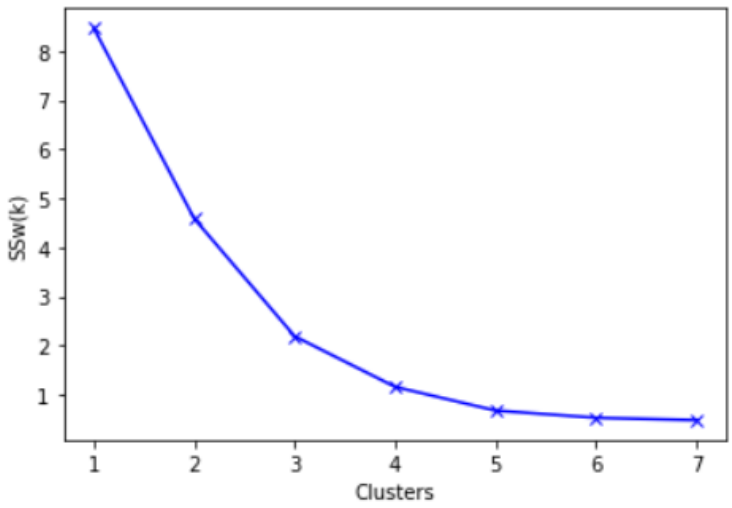



Figura 4.59: Método del codo en Fase 1

La gráfica realizada nos da una información sobre el número más optimo de clusters. En este caso, podría discutirse la opción de 3 o de 4 clusters, pero se ha estimado como mejor opción 3 clusters.
La agrupación en los 4 grupos representativos está explicado en la tabla [4.5]. Se han etiquetado de 0 a 3, los posibles cuatro grupos de entre los 19 arranques posibles. Estas agrupaciones coinciden con las realizadas en la caracterización de fases pero no con el algoritmo jerárquico, el cual diferenciaba 5 grupos característicos.

| **Grupos** | **Días** |
|---|---|
| 1 | 17 Enero,21 Enero,24 Enero,27 Enero,29 Enero,30 Enero, 31 Enero, 3 Febrero,4 Febrero,11 Febrero,13 Febrero,14 Febrero |
| 2 | 22 Enero,7 Febrero,10 Febrero |
| 3 | 13 Enero,16 Enero,5 Febrero,6 Febrero |

Cuadro 4.5: Agrupaciones Fase 1

**Fase 2**

El proceso de agrupación para la fase 2 tiene mucha menos subjetividad observando el resultado de la gráfica del codo [4.60]. El número óptimo de clusters es 2 y de acuerdo a este número de clusters se han etiquetado los distintos arranques según el grupo característico. El grupo minoritario está formado por los arranques de los días 16,22 de Enero, 5,7, y 10 de Febrero. Esta información se encuentra en la tabla [4.6].

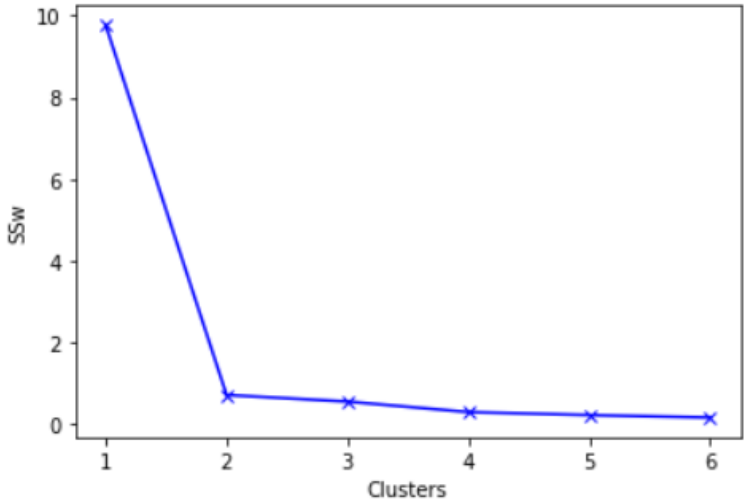



Figura 4.60: Método del codo en Fase 2

| Grupos | Días |
|---|---|
| 1 | 17 Enero,21 Enero,24 Enero,27 Enero,29 Enero,30 Enero, 31 Enero, 3 Febrero,4 Febrero,11 Febrero,13 Febrero,14 Febrero |
| 2 | 22 Enero,7 Febrero,10 Febrero |
| 3 | 13 Enero,16 Enero,5 Febrero,6 Febrero |

Cuadro 4.6: Agrupaciones Fase 2

**Fase 3**

De acuerdo a la gráfica [4.61] parece indicar que la mejor solución son 3 agrupaciones. En base al número de agrupaciones escogido se ha etiquetado aplicando el algoritmo de kmeans, a que grupos pertenecen cada uno de los arranques.

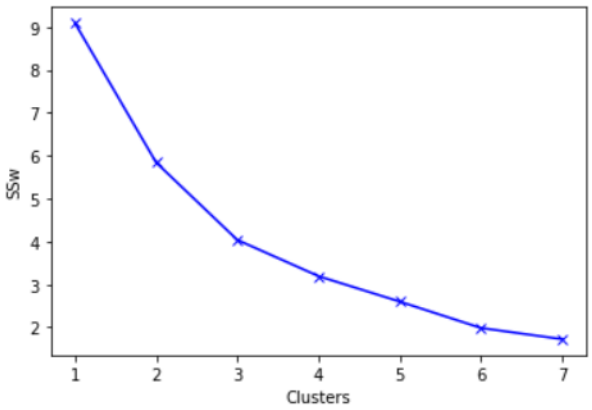



Figura 4.61: Método del codo en Fase 3

| **Grupos** | **Días** |
|---|---|
| 1 | 16 Enero,17 Enero, 21 Enero, 22 Enero, 27 Enero, 5 Febrero, 7 Febrero, 10 Febrero, 14 Febrero |
| 2 | 13 Enero,31 Enero,3 Febrero,4 Febrero,13 Febrero |
| 3 | 24 Enero,29 Enero,30 Enero,6 Febrero, 11 Febrero |

Cuadro 4.7: Agrupaciones Fase 3

**Arranque Global**

La nueva estimación del número de clusters óptimo parece indicar cuatro clusters como mejor agrupación posible. Para determinar a que arranques pertenecen dichas agrupaciones, se adjunta una tabla [A.9] con el ajuste del algoritmo KMeans con el número de clusters más 'beneficioso'.

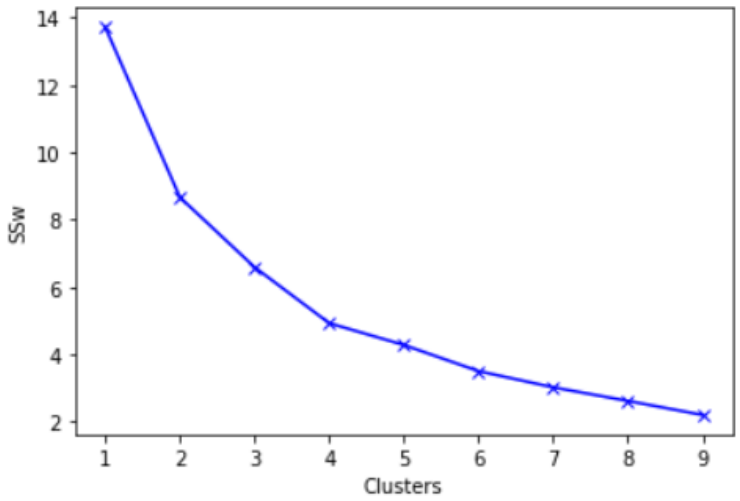



Figura 4.62: Método del codo arranque global - KMeans

| Grupos | Días |
|---|---|
| 1 | 13 Enero, 31 Enero,3 Febrero, 4 Febrero,6 Febrero |
| 2 | 17 Enero, 21 Enero, 27 Enero,14 Febrero |
| 3 | 24 Enero,29 Enero,30 Enero,11 Febrero |
| 4 | 16 Enero,22 Enero,5 Febrero, 7 Febrero, 10 Febrero |

Cuadro 4.8: Agrupación de Arranques global

### 4.6.3 Comparación resultados algoritmos

Tras la aplicación de los algoritmos KMeans y complete linkage para agrupar los distintos arranques, se observa un cuadro comparativo sobre el etiquetado de los distintos arranques a los distintos grupos. La aplicación de dos algoritmos distintos se ha realizado con motivo de reforzar la hipótesis sobre la distinción entre los grupos de arranques. El objetivo es que tras la aplicación de los algoritmos, las agrupaciones sean lo más similares posibles a las agrupaciones realizadas a priori en la sección 4.4.1.

| Fecha | Hipótesis Grupos | KMeans Grupos | Jerárquico Grupos |
|---|---|---|---|
| 2020-01-13 | 2 | 2 | 2 |
| 2020-01-16 | 1 | 1 | 1 |
| 2020-01-17 | 3 | 3 | 3 |
| 2020-01-21 | 3 | 3 | 3 |
| 2020-01-22 | 1 | 1 | 1 |
| 2020-01-24 | 4 | 4 | 4 |
| 2020-01-27 | 3 | 3 | 3 |
| 2020-01-29 | 4 | 4 | 4 |
| 2020-01-30 | 4 | 4 | 4 |
| **2020-01-31** | **2** | **2** | **3** |
| 2020-02-03 | 2 | 2 | 2 |
| 2020-02-04 | 2 | 2 | 2 |
| 2020-02-05 | 1 | 1 | 1 |
| 2020-02-06 | 2 | 2 | 2 |
| 2020-02-07 | 1 | 1 | 1 |
| 2020-02-10 | 1 | 1 | 1 |
| 2020-02-11 | 4 | 4 | 4 |
| **2020-02-13** | **2** | **3** | **3** |
| 2020-02-14 | 3 | 3 | 3 |

Cuadro 4.9: Tabla comparativa hipótesis-KMeans-Jerárquico

Se observa una agrupación distinta en uno o dos arranques con respecto a la hipótesis inicial. Uno de ellos corresponde al arranque correspondiente al día 31 de Enero, el cual es asignado al grupo 2 tanto por KMeans como con la hipótesis base. Sin embargo, el algoritmo jerárquico lo asocia con el grupo 3.

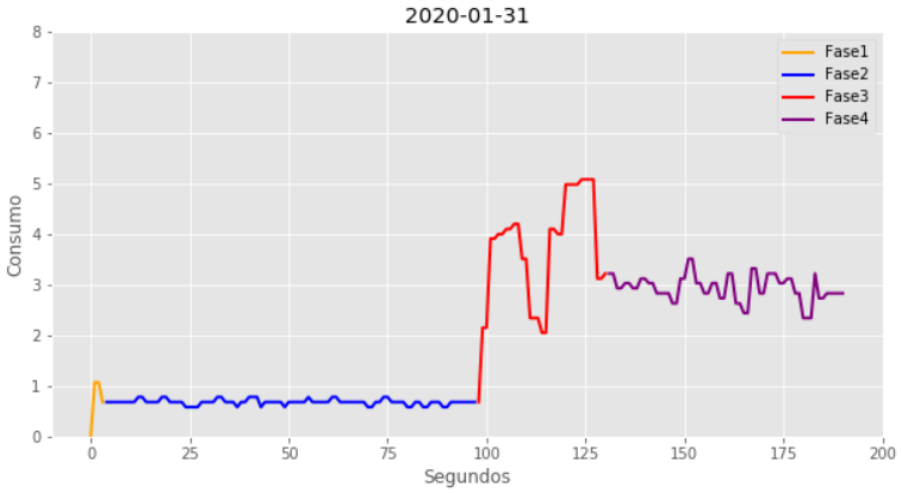



Figura 4.63: Arranque 31 Enero

Las diferencias en la fase 1 entre los grupos 2 y 3 radica en valores más altos en el grupo 2. En

cuánto a la fase 2 no hay diferencias significativas entre ambos grupos por lo que en este aspecto no sería problema para la clasificación. No obstante, en la Fase 3 se diferencian en una duración mayor en el grupo 2 así como unos valores de consumo más altos durante la Fase 3. Atendiendo a los valores medios del arranque del 31 de Enero, se escoge el grupo 3 tal y como lo agrupó el algoritmo KMeans.

El otro arranque que se ha clasificado distinto en alguna de las tres clasificaciones es el arranque correspondiente al día 13 de Febrero. En este caso, ambos algoritmos lo clasifican en el grupo 3 mientras que la hipótesis realizada a priori fue clasificada con el grupo 3. Para la clasificación de este arranque existían dudas debido a similitudes con ambos grupos de arranques.

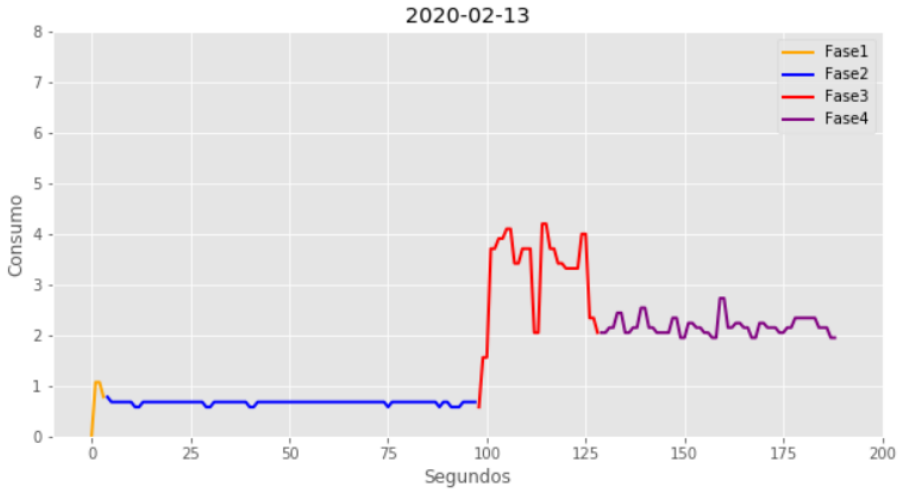



Figura 4.64: Arranque 13 Febrero

Con respecto a la fase 1, se clasificaba el arranque con el grupo 2 No obstante, la duración de Fase 3 del arranque era similar a la media establecida para el grupo 2. Sin embargo, los valores de consumo máximos y medios tenían más relación con el grupo 3, lo que llevó a clasificarlo en el grupo 2.
*Por ello, se escoge la clasificación no supervisada realizada por KMeans para el resto del trabajo.*

# Capítulo 5

# Extracción de conocimiento y Evaluación del experimento

En este capítulo se aborda un modelo de arranque más elaborado basado en la caracterización de los cuatro modelos representativos agrupados por el algoritmo KMeans. La representación gráfica de la evolución del consumo respecto al tiempo permite observar las características de cada uno de los cuatro modelos, utilizando la media de consumo en cada segundo, con unas desviaciones caracterizadas por el valor máximo y mínimo. Posteriormente, se expone como una segunda iteración a partir de la extracción de patrones comunes a los cuatro modelos representativos fruto de un aprendizaje no supervisado, permitir caracterizar y explicar la mayor parte de los arranques de la máquina, obteniendo un nuevo modelo de arranque como suma de todos estos patrones.

Por último, se realiza la validación de los modelos implementados con una simulación paso a paso con un conjunto de arranques nuevos (utilizados como datos de prueba). La comparación de cada arranque con el modelo característico de la máquina, permitirá conocer el funcionamiento de la máquina y detectar incidencias que no se corresponden con ninguno de los modelos obtenidos y por tanto que pasarán a parte de un conjunto de datos etiquetados con arranques anómalos.

## 5.1 Modelos de arranques característicos

En la siguiente Figura [5.1] se expone de manera visual la representación de consumo energético de cada fase de operación del arranque de la máquina respecto al tiempo. Se puede observar cada uno de los cuatro tipos de modelo de arranque clasificados por el algoritmo KMeans. Los valores se encuentran normalizados para que se pueda hacer una comparación y extracción de los patrones característicos.

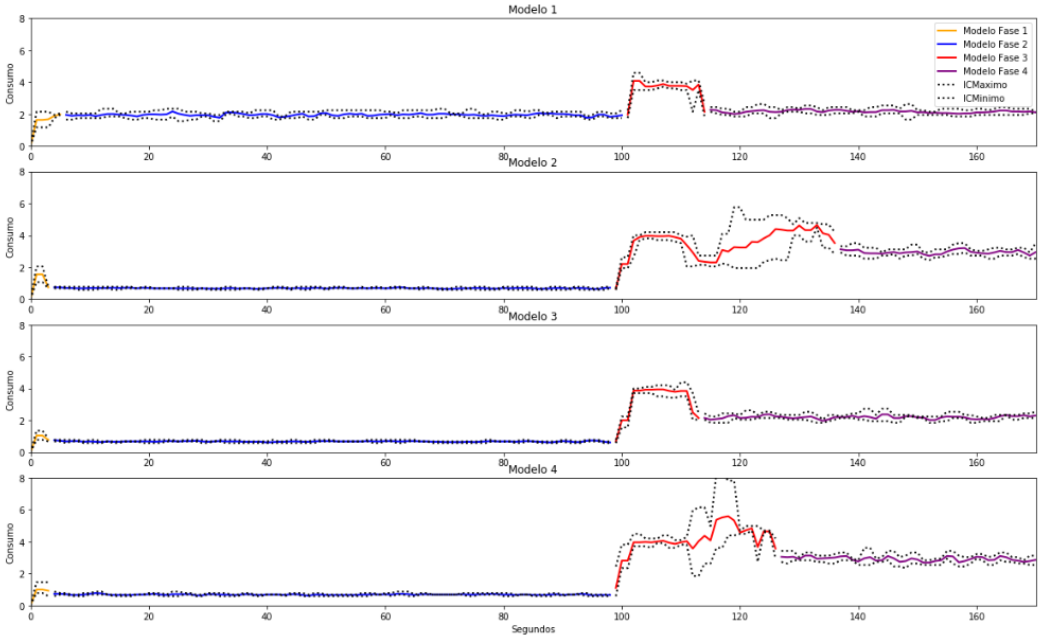



Figura 5.1: Arranque global cuatro modelos clasificados

Atendiendo a una explicación técnica, el escenario más probable al que tendería nuestro trabajo es a obtener una única representación característica del arranque normal o correcto de la máquina, siendo el resto de valores obtenidos correspondientes a diferentes anomalías relacionadas con cambios de comportamiento, errores de medida o incidencias en la máquina.

De acuerdo al razonamiento anterior, la principal diferencia se observa en la divergencia del Modelo 1 con respecto de los Modelos 2,3 y 4. La característica fundamental es el consumo medio de la fase 2 de arranque, que en el caso del Modelo 1 oscila en torno a los 2 Amperios mientras que en resto de Modelos tiene un valor en torno a 0.65 Amperios. Como se indicó anteriormente, la fase 2 de arranque de los Modelos 2,3 y 4 fue inter-polada a una duración media de 95 segundos con motivo de poder comparar adecuadamente los cuatro tipos de Modelos. La duración adicional de la fase 2 no aporta información relevante, siendo tiempo de espera en la que el consumo medio no varia significativamente. En la siguiente sección se explican de manera detallada los patrones o características que tienen en común, o no, los diferentes modelos de arranque.

## 5.2 Patrones comunes a los modelos

En este apartado del trabajo se detallan las características o patrones que son comunes a varios de los modelos de arranque caracterizados por las técnicas de aprendizaje no supervisado. Permitirá determinar las características más comunes que presentan la mayoría de los arranques. De esta manera, se obtendrá un nuevo modelo de arranque más elaborado y representativo a partir de las características del mismo,fruto del descubrimiento de patrones comunes a varios de los modelos(Modelos 2,3 y 4).También se quiere determinar si el modelo obtenido explica que los arranques se comportan de acuerdo a una caracterización más común o sin embargo encontramos valores que atienden a alguna anomalía.

A continuación se presentan cuatro secciones dedicadas a cada fase de operación del arranque en las que se tratará de encontrar las características comunes entre varios de los modelos representados.

### 5.2.1 Patrones característicos Fase 1 de arranque

Una de las características más importante es el inicio del arranque, que se corresponde con la Fase 1. Para comparar el comportamiento del consumo respecto al tiempo durante esta fase, se muestra en la Figura [5.2] una gráfica con la tendencia media de cada uno de los cuatro modelos y sus respectivas desviaciones correspondientes al máximo y mínimo de consumo en las mismas escalas.

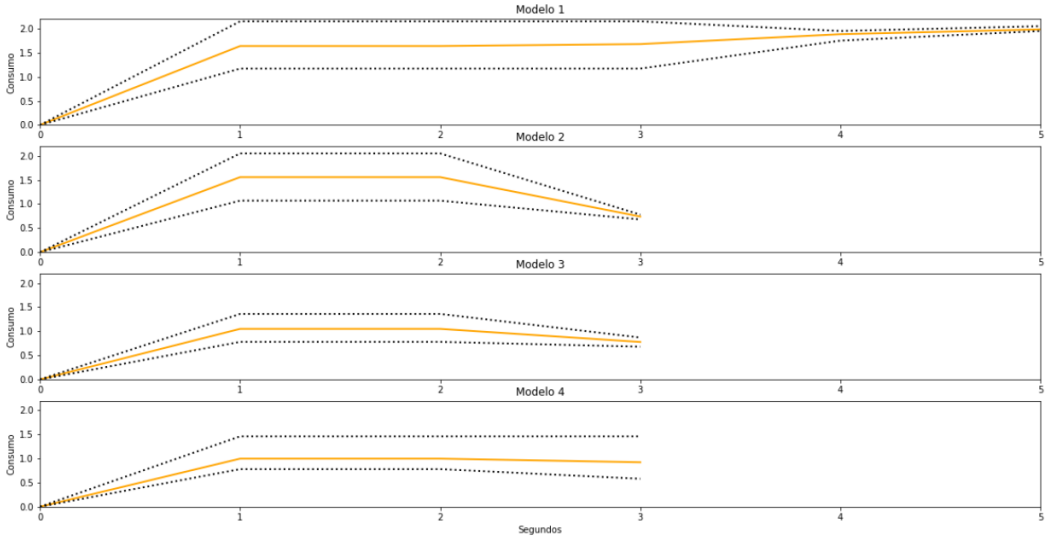



Figura 5.2: Modelos Fase 1

Observando los valores de consumo característicos de los 4 modelos, destaca la existencia de una duración más prolongada en el tiempo en el Modelo 1. Sin embargo, el comportamiento del arranque en esta Fase 1 para los Modelos 2,3 y 4 es similar, con un crecimiento del consumo, que se mantiene estable durante un segundo y disminuye hasta valores caracterizados en el inicio de la Fase 2.

También es destacable el escalón de consumo que ocurre en los cuatro modelos, donde en el primer segundo llega a un pico de consumo y se estabiliza durante otro segundo con un valor constante. Para una mejor observación, el comportamiento de los cuatro modelos durante los primeros 2 segundos se presenta en la siguiente gráfica:

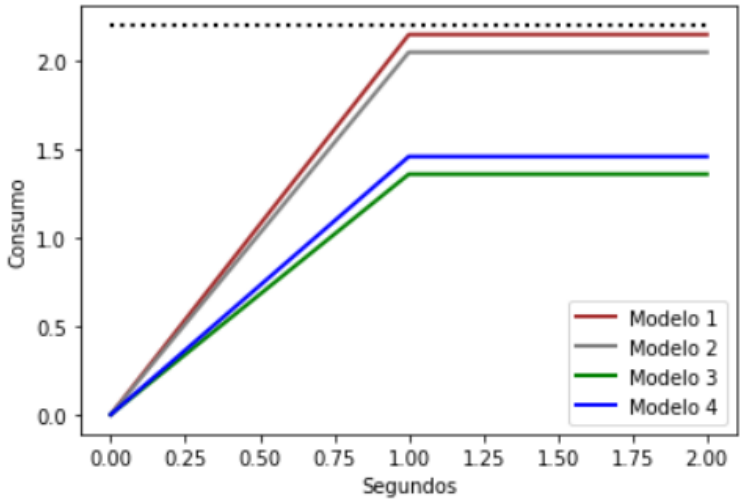



Figura 5.3: Escalón Fase 1

Con el conjunto de datos disponible se puede establecer un umbral de consumo máximo en 2.3 Amperios para la Fase 1, como se observa en la línea discontinua de la gráfica. Todo consumo que supere ese valor será etiquetado como una anomalía o incidencia en ese instante.

Atendiendo a las similitudes entre los Modelos 2,3 y 4, se ha obtenido un patrón de comportamiento medio de los tres modelos durante la Fase 1. El comportamiento de esta fase a partir de los 3 modelos indicados, se observa en la siguiente gráfica, en la que se añaden ligeras desviaciones que puede tomar el consumo. Estas se han formado de obtener la media de los valores máximos de los tres arranques y añadiendo una pequeña desviación añadida :

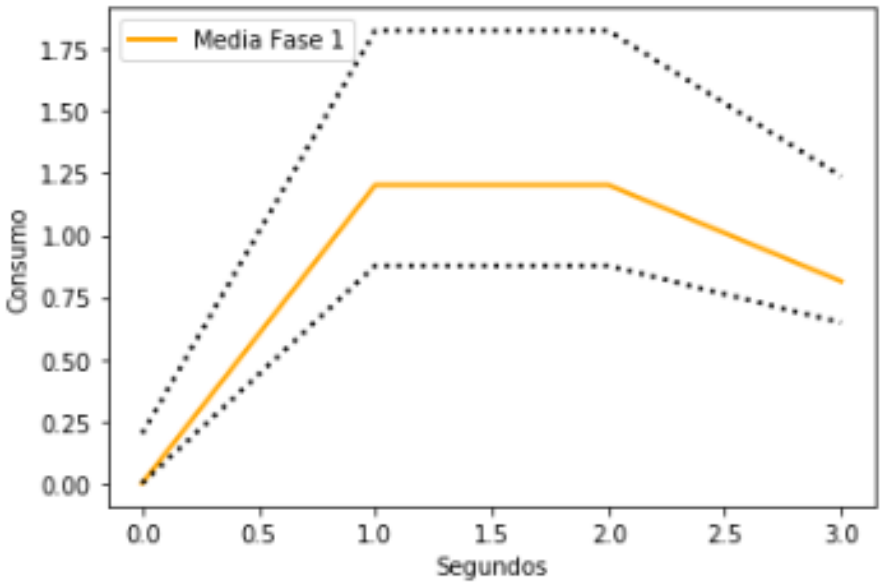



Figura 5.4: Fase 1 Agrupación Modelos 2,3 y 4

Con la agrupación realizada, se pretende que la mayoría de los arranques que se simulen se expliquen por las características del nuevo comportamiento de esta Fase 1.

Por otra parte, el arranque podrá comportarse de acuerdo a las características del Modelo 1 con un consumo mayor y con posibilidad de más duración en tiempo durante la Fase 1. La existencia de un arranque que no se explique de acuerdo a ninguno de los dos modelos se considerará como una posible anomalía. El comportamiento del Modelo 1 para la Fase 1 es:

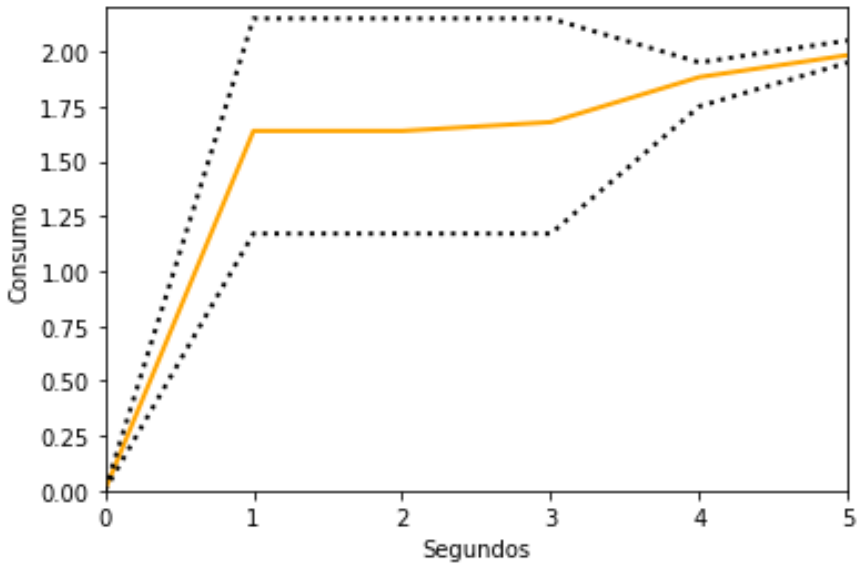
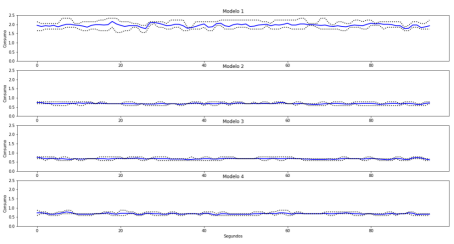


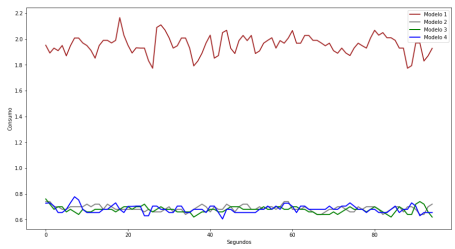

Figura 5.5: Fase 1 Modelo 1

### 5.2.2 Patrones característicos Fase 2 de arranque

El comportamiento característico en esta fase viene determinado por un consumo medio estable que oscila en torno a unos valores máximos y mínimos de manera frecuente. Se caracteriza por ser una fase en la que la fresadora se encuentra en 'stand-by' y en la que no importa tanto la duración sino el consumo sobre el que oscila. En la gráfica siguiente [5.6a] se muestra el comportamiento de cada uno de los 4 modelos con la misma duración para la Fase 2 de arranque.



(a) Fases Modelo 2



(b) Comparación modelos fase 2

Figura 5.6: Comparación modelos de comportamiento en fase 2

Se observa una divergencia clara entre los Modelos 1 y los Modelos 2,3,4 al igual que ocurría

en la Fase 1. En este caso la diferencia atiende a un consumo medio muy diferente a lo largo de la Fase 2. Las características de los Modelos 2,3 y 4 se basan en un consumo que oscila entre 0.58 y 0.78 Amperios indistintamente. Sin embargo, el consumo del Modelo 1 varía por encima de 1.5 Amperios y por debajo de 2.5 Amperios, con una oscilación mucho más variable que en los otros modelos. Debido al patrón similar de los Modelos 2,3 y 4 se ajusta un nuevo comportamiento característico. Se observa en el siguiente gráfico:

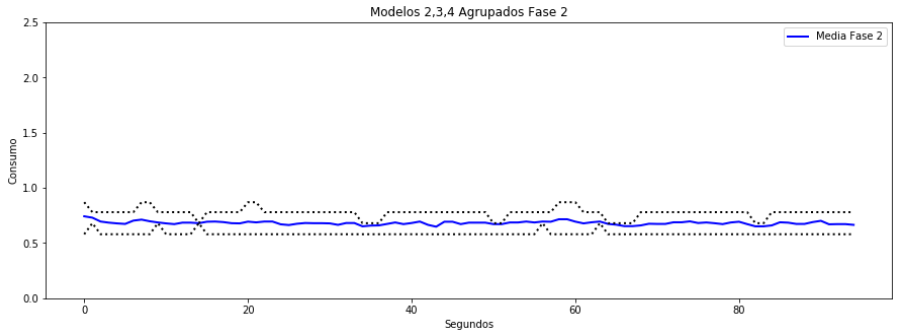



Figura 5.7: Fase 2 Modelos 2,3 y 4

Sin embargo, el patrón de comportamiento del Modelo 1 tiene la siguiente tendencia:

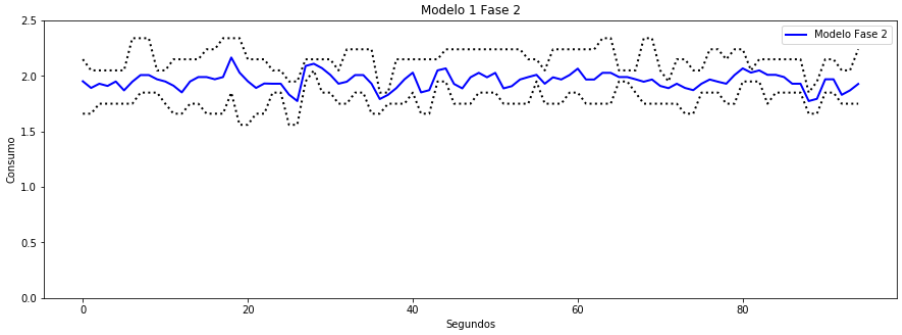



Figura 5.8: Fase 2 Modelo 1

### 5.2.3 Patrones característicos Fase 3 de arranque

La Fase 3 se caracteriza por el momento en que el operario quita la seta de seguridad del panel de control. Esto sucede después de la fase de estabilización del consumo en la que se produce el ciclo de trabajo del PLC. Esta fase es quizás la que más difiere en los cuatro modelos. De acuerdo a las características de la Fase 3, se establecerá un umbral máximo de consumo de 4.6 Amperios en los primeros 12 segundos mientras que será de 8 Amperios en el resto de la fase. En la siguiente Figura [5.9] se pueden observar gráficamente los comportamiento de los cuatro modelos durante la Fase 3:

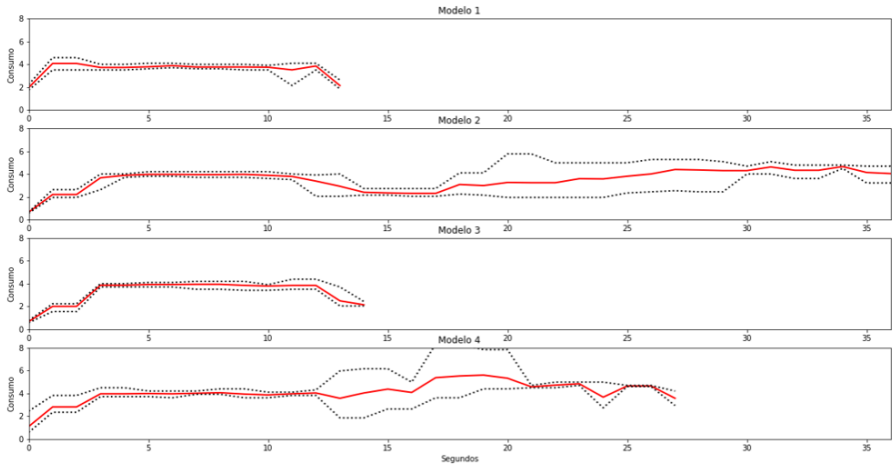
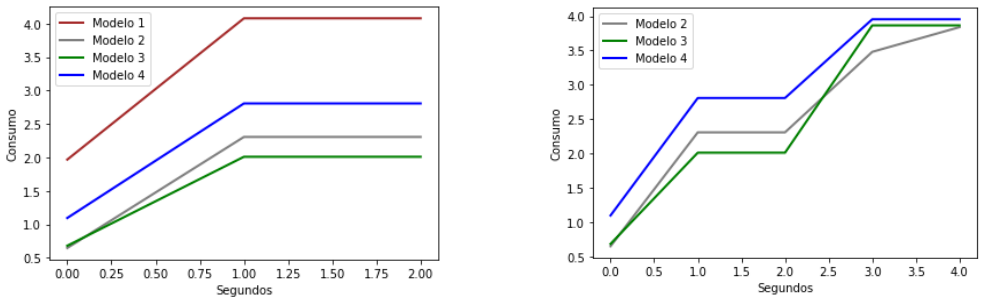



Figura 5.9: Fase 3 para los cuatro modelos

Uno de los patrones de comportamiento que se repite en los cuatro modelos corresponde a un escalón similar al que ocurría en la Fase 1, en la que se permanecía estable el consumo durante 1 segundo. Otro patrón encontrado es la existencia de un doble escalón en el consumo al inicio de la Fase3. En este caso, ese patrón es característico a los Modelos 2,3 y 4 debido a que el crecimiento de consumo proviene de valores de consumo de la Fase 2 (0.58-0.78 Amperios) mientras que del Modelo 1 son cercanos a 2 Amperios.

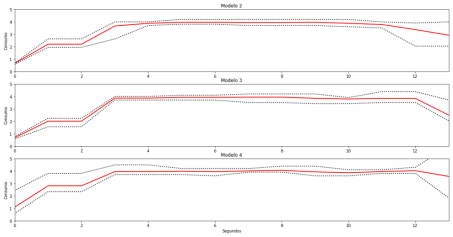



(a) Escalón Fase 3

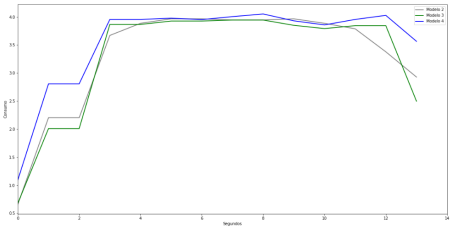
(b) Doble Escalón Fase 3

Figura 5.10: Características Fase 3

Sin embargo, yendo más allá en la obtención de patrones comunes, es interesante el comportamiento similar que existe entre los modelos 2,3 y 4 a lo largo de los 12 primeros segundos. Se observa en la siguiente gráfica [5.11b]:



(a) Consumo Fase 3 Inicial



(b) Fase 3 Inicial 12 Segundos Modelos 2,3,4

Figura 5.11: Características Fase 3 Inicial

Esta observación permite caracterizar los primeros segundos de la Fase 3 como media de los Modelos 2,3 y 4. El resultado de la misma se ve en la gráfica [5.12].

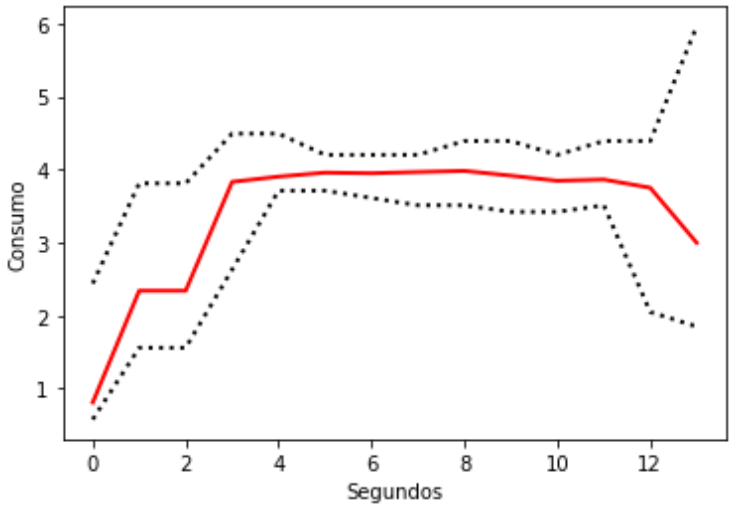
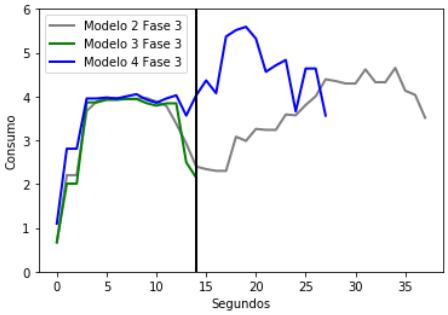
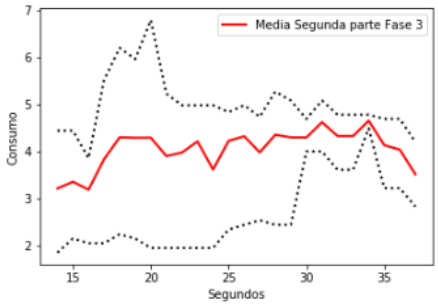



Figura 5.12: Consumo Medio Fase 3 Inicial Modelos 2,3,4

Más difícil resulta caracterizar la segunda parte de la Fase 3, únicamente representada por los Modelos 2 y 4. Aunque la duración media del Modelo 2 sea de 37 segundos y la del Modelo 4 sea de 28 segundos, se aplicará la media del consumo hasta el segundo 28 de los modelos mientras que los segundos restantes serán los correspondientes con el Modelo 2. En general, la duración de esta fase es muy variable por lo que no se tendrá en cuenta la duración de la misma aunque sí un umbral máximo de 5 segundos más que el modelo. La diferencia en el consumo y en la duración de ambos modelos se ofrece en la gráfica [5.13a] , mientras que el patrón característico creado de ambos arranques se observa en la figura [5.13b] dónde existen unas desviaciones propias de ambos modelos agrupados:




(a) Fase 3 Modelos 2 y 4




(b) Agrupación Fase 3 Final

Figura 5.13: Características Fase 3 - Modelos 2 y 4

### 5.2.4 Patrones característicos Fase 4 de arranque

Para terminar con esta sección de búsqueda de patrones comunes a varios de los modelos, se procede con la detección en la última fase del arranque. A pesar de que no se utilizaron las estadísticas de esta fase para clasificar los distintos modelos, ahora es interesante observar el comportamiento usual que tiene cada uno de los modelos que fueron clasificados.
En la figura siguiente [5.14] se observan dos patrones de comportamiento diferentes entre los cuatro modelos:

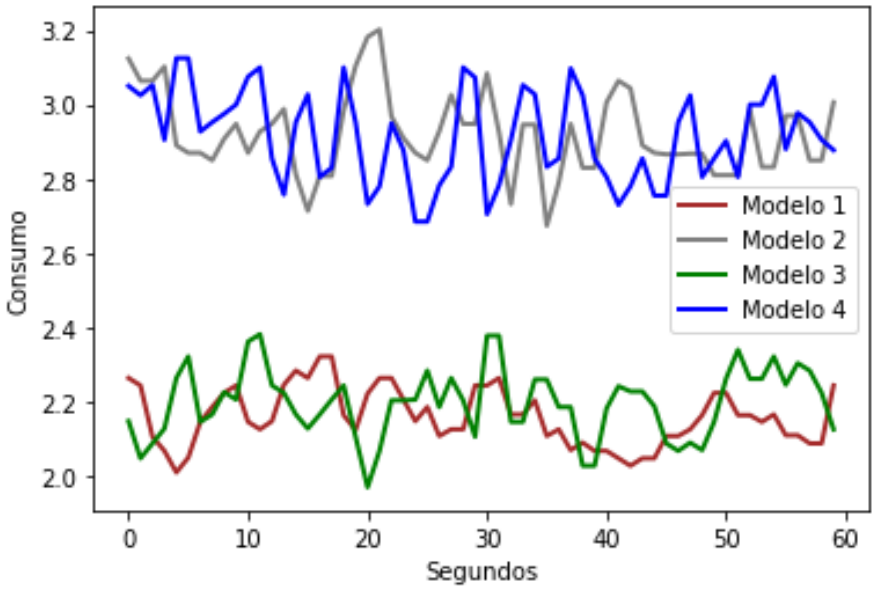



Figura 5.14: Comportamiento consumo Fase 4

Se detecta un comportamiento del consumo medio que oscila entre 2 y 2.4 Amperios para los Modelos 1 y 3, mientras que para los Modelos 2 y 4 el consumo oscila entre 2.7 y 3.3 Amperios. Esto está ocasionado por el comportamiento final de la Fase 3. En el caso de los Modelos 1 y 3 con una Fase 3 de duración más corta terminada con una bajada del consumo pronunciada.

Sin embargo los Modelos 2 y 4 están caracterizados por una duración más larga de 25 segundos en la que aparece otro pico de consumo y el descenso del consumo es similar pero partiendo la mayor parte de las veces de un consumo mayor por lo que los valores finales oscilan entre valores más altos cercanos a 3 Amperios.

No obstante, debido a las similitudes en el resto de fases entre los Modelos 2,3 y 4, se realiza una agrupación de los tres modelos para la Fase 4, obteniendo el comportamiento de la siguiente Figura, con sus respectivas desviaciones:

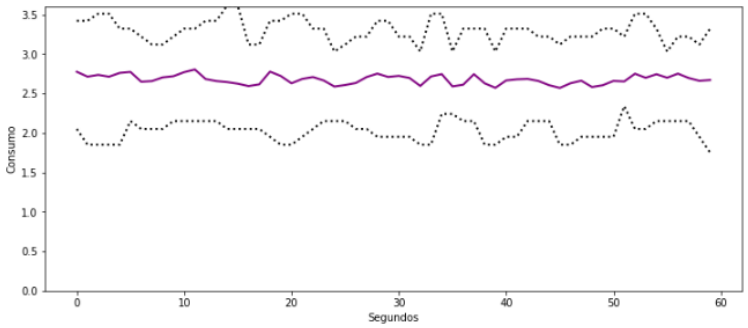



Figura 5.15: Fase 4 Modelos2,3,4

### 5.2.5 Nuevos Modelos de arranques más representativos

De acuerdo a las características o comportamientos encontrados en varios de los modelos, resalta una gran similitud entre los Modelos 2,3 y 4 y sobre todo entre los Modelos 2 y 4, únicamente diferenciados por una duración distinta de la Fase 3 así como valores medios o máximos de esta misma fase. El Modelo 3 tiene características similares al Modelo 1, Modelo2 y el inicio de la fase 3. Sin embargo, el consumo medio de la Fase 4 no se corresponde con valores medios que sí se asocian entre sí entre los Modelos 2 y 4.

Por tanto, es posible crear un nuevo modelo de arranque característico más elaborado después de comparar los Modelos 2,3 y 4 con el fin de poder determinar que la mayoría de los arranques se comportan de acuerdo a este comportamiento. Sin embargo, la divergencia existente con el Modelo 1, se utilizará como alternativa a este nuevo modelo del arranque. La existencia de un arranque que no se comporte de acuerdo a ninguno de los dos modelos presentados, significará que la máquina ha experimentado ciertas anomalías en el arranque, por lo que será etiquetado como un arranque anómalo. El nuevo modelo de arranque más representativo se observa en la siguiente figura:

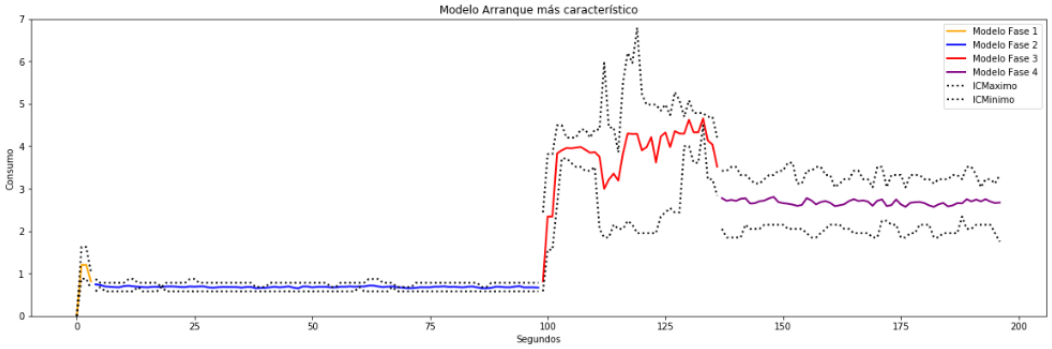



Figura 5.16: Modelo arranque característico

Sin embargo, existe un porcentaje de arranques que se comportan de manera diferente. El comportamiento característico se observa en la siguiente figura [5.17]. Un arranque que no se comporte como ninguno de los dos arranques característicos significará de la existencia de anomalías que hacen que el arranque no se clasifique como un arranque normal debido a la existencia de incidentes durante el arranque.

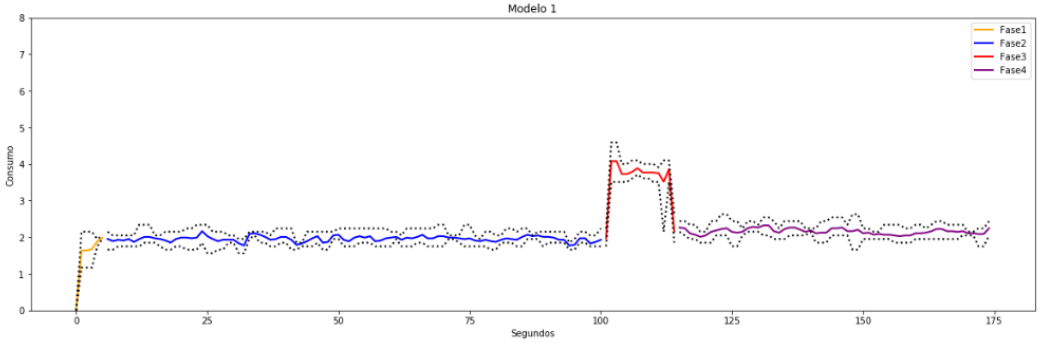



Figura 5.17: Modelo arranque tipo 1

## 5.3 Simulación con arranques nuevos vs modelos de arranque

Esta etapa se caracteriza por utilizar un conjunto de datos no utilizados anteriormente, que consiste en arranques nuevos, con el objetivo de detectar si estos se pueden explicar a partir del comportamiento habitual del arranque obtenido en la sección anterior, como el comportamiento del Modelo 1 o se encuentran distintas anomalías.

Gracias a los modelos de comportamiento creados para los distintos grupos de arranques, se puede determinar si el comportamiento y la tendencia de los nuevos arranques son característicos o sin embargo es necesario profundizar más.

Para ello, se han extraído un total de 20 arranques de prueba del mismo conjunto de datos formado con datos de Diciembre de 2019 a Marzo de 2020. Tal y como se especificó en *la sección 4.2*, se realizó una fase de procesamiento de los arranques para encontrar realmente los válidos para este estudio.

Posteriormente, tal y como se procedió con los arranques utilizados como datos de entrenamiento, se detectó la tendencia de cada segundo con respecto al segundo anterior y posterior para detectar crecimientos y decrecimientos destacables en el arranque. De este modo, se detectaron los cambios de fases tal y como se ofrece en la tabla [A.2] con los índices correspondientes a los cambios de fase para cada uno de los 20 arranques.

Los datos de prueba - arranques nuevos - utilizados proceden de los días :

- Diciembre: 2,3,4,5 y 20
- Febrero: 18,19,20,21,25,26,27
- Marzo: 2,5,6,11,12,20,25,26

La serie temporal de cada uno de los arranques correspondientes a los datos de prueba, en la que se encuentran interpoladas la duración de la fase 2, se adjunta en Anexos en [A.3.5]. Destacar que se ha inteporlado la duración de la Fase 2 a 95 segundos, la cual variaba desde 200 hasta 1000 segundos. De este modo, se pueden comparar las características de cada uno de los arranques utilizando mismas escalas, tanto de consumo como de tiempo del arranque.

De acuerdo a las características o patrones encontrados en la sección anterior, se realiza un análisis exhaustivo por fases y por segmentos de fases para comprobar el porcentaje de arranques que cumplen los patrones y con ello caracterizar el comportamiento del arranque más representativo. Las características que van a ser medidas son:

- Escalón Fase 1
- Consumo Máximo Fase 1
- Doble Escalón Fase 1
- Comportamiento Fase 1
- Comportamiento Fase 2

- Escalón Fase 3
- Doble Escalón Fase 3
- Consumo Fase 3 Inicial
- Consumo Fase 3 Final
- Duración Máxima Fase 3
- Anomalías Leves Fase 3
- Comportamiento Fase 3
- Comportamiento Fase 4

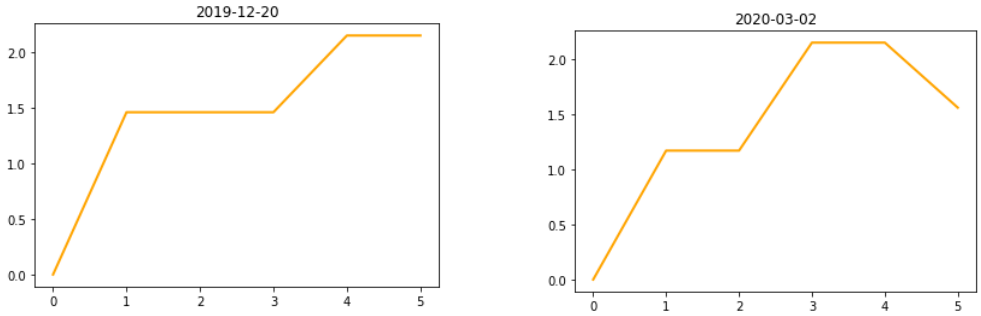
A continuación se describe el proceso de la simulación por segmentos de fases y fases sobre 20 arranques extraídos para datos de prueba. Se detalla en tablas el porcentaje de arranques que cumplen los distintos patrones así como los arranques que no lo cumplen. El objetivo es comparar cada uno de los nuevos arranques con el modelo de arranque característico de la Figura [5.16] y si no se corresponde, se compara con el Modelo 1 de arranque de la Figura [5.17].

**Fase 1**

Las primeras características que se han comparado en los nuevos arranques se corresponden con la existencia de un escalón al inicio del arranque. La estabilidad del consumo durante el segundo 2 del arranque es una característica muy común. No obstante, existe algún caso en el que no sucede como se vio en un arranque utilizado inicialmente como datos de entrenamiento. Se trata del día 15 de Enero observado en la figura [4.22b]. En este caso, cada uno de los 20 arranques de prueba cumplen con dicho patrón característico.

Otra de las características influyentes en esta Fase 1 es la adición de un umbral de consumo máximo en 2.3 Amperios. Del mismo modo, en los datos de entrenamiento nos encontramos dos inicios de arranques con consumos de 3 y 5 Amperios en los días 14 y 28 de Enero. Sin embargo, el 100 % de los arranques de prueba cumple con esta característica, no sobrepasando los valores del umbral máximo durante la Fase 1.

Un patrón encontrado para confirmar que no se ajusta con el modelo de arranque más característico, es la existencia de un doble escalón al inicio de la Fase 1 y consecuentemente de una duración de fase más larga. Encontrar este tipo de comportamiento es sinónimo de modelos característicos del tipo 1, o dicho de otro modo, de no relación con el nuevo grupo de arranque característico. Existen dos arranques con estas características, asociadas a los días 20 de Diciembre y 2 de Marzo. El comportamiento de las mismas se observa en las siguientes figuras:




(a) Fase 1 - 20 Diciembre 2019

(b) Fase 1 - 2 Marzo 2020

Figura 5.18: Características Fase 1 - Modelo 1

Como resultado de la comparación de los nuevos arranques con el comportamiento de la Fase 1, observando los valores medios y desviaciones en la figura [5.4], salvo los dos arranques mencionados y la adición del arranque del 19 de Febrero de 2020, los 17 arranques restantes se comportan de acuerdo al nuevo modelo de arranque.En la siguiente tabla resumen se pueden observar los porcentajes de arranques asociados al nuevo modelo durante la Fase 1:

| **Arranques** | **Escalón Fase 1** | **Consumo Máximo Fase 1** | **No Doble Escalón Fase 1** | **Comportamiento Fase 1** |
|---|---|---|---|---|
| Nuevo Modelo Arranque | 100 % | 100 % | 90 % | 85 % |

Cuadro 5.1: Porcentajes asociados al nuevo modelo de arranque en Fase 1

**Fase 2**

Las características de la Fase 2 se rigen única y exclusivamente por el consumo. La duración no es influyente ya que todos los arranques fueron interpolados a una duración similar.
Las características del nuevo modelo de arranque observadas en la figura [5.7] destacan por una variabilidad del consumo entre 0.58,0.68 y 0.78 Amperios. No obstante, hay ligeras desviaciones en algún arranque a valores de consumo de 0.87 Amperios o 0.48 Amperios. No obstante, no han influido para clasificarlas como incidencias. Salvo esta objeción, el 90 % de los arranques se comportan de acuerdo al modelo característico mientras el 10 % restante se comporta de acuerdo al Modelo 1, con valores de consumo muy superiores, oscilando en torno a los 2 Amperios. Estos dos arranques correspondientes al 10 %, de nuevo se corresponden con el 20 Diciembre de 2019 y el 2 de Marzo de 2020. A continuación, se puede ver el cuadro resumen de correspondencia de los patrones de la Fase 2:

| **Arranques** | **Comportamiento Fase 2** |
|---|---|
| Nuevo Modelo Arranque | 90 % |

Cuadro 5.2: Porcentajes asociados al nuevo modelo de arranque en Fase 2

**Fase 3**

Esta fase es la que más diverge en todos los arranques. De acuerdo a los patrones comunes encontrados en la sección anterior, se buscará encontrarlos en cada uno de los 20 arranques característicos. Por último, se compara segundo a segundo el consumo de la Fase 3 de cada uno de los arranques con respecto al comportamiento del nuevo modelo característico.

El primer segmento característico de esta fase corresponde con un primer escalón tal y como sucede en la fase 1. El 100 % de los arranques cumple con ello.
Para continuar con la simulación, es necesario establecer un umbral de consumo durante la Fase 3. Por ello, de acuerdo a los valores obtenidos con los datos de entrenamiento se ha dividido en dos segmentos. Los primeros segundos tienen un umbral máximo de consumo de 4.6 Amperios mientras que el resto de Fase 3 tendrán 8 Amperios como valor máximo. Haciendo referencia al primero de los casos, es hallado un arranque que excede el valor de 4.6 Amperios. Las características del mismo se muestran en la siguiente figura:

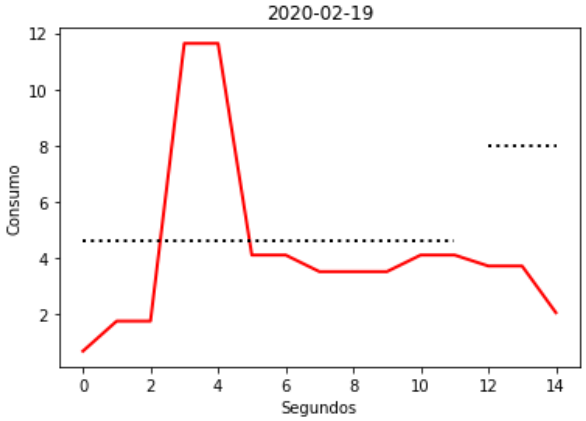
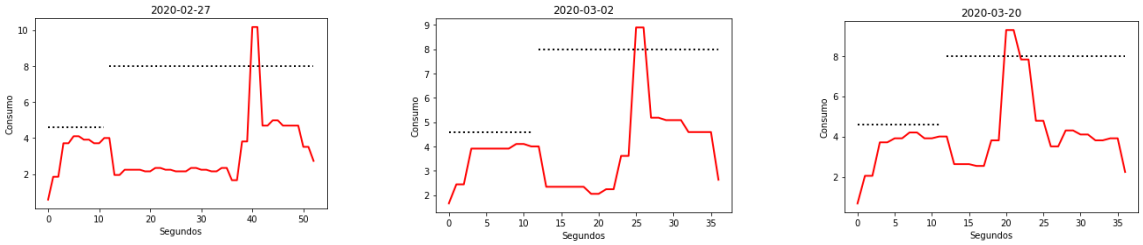



Figura 5.19: Fase 3 - 19 Febrero 2020 Anómalo

El día 19 de Febrero de 2020, se observa un comportamiento anómalo con consumo de casi 12 Amperios durante los segundos 3 y 5, lo que hace explicar este arranque como anómalo, y no clasificándolo en ningún arranque característico.

No obstante, a partir del segundo 13, se encuentran otros tres arranques con un valor de consumo que excede los 8 Amperios. Se corresponde con los días 27 Febrero de 2020, 2 de Marzo de 2020 y 20 de Marzo de 2020. Las gráficas asociadas son:




(a) 27 Febrero - Fase 3 Anómalo (b) 2 Marzo - Fase 3 Anómalo (c) 20 Marzo - Fase 3 Anómalo

Figura 5.20: Arranques Anómalos en Fase 3

Debido a la variabilidad en la duración de la Fase 3, se ha establecido como duración máxima de la fase 3 de 42 segundos, ya que la duración máxima estimada para el nuevo modelo de arranque es cercano a los 40 segundos. Se ha estimado una desviación máxima de 3 segundos añadidos para considerarse válido. De acuerdo a ello, existen dos arranques con 45 y 48 segundos que se alejan del valor máximo de duración para el análisis. Se corresponden con los días 5 y 25 de Marzo de 2020. No obstante, antes de etiquetarlos como anómalos, se comprobará como es su comportamiento durante esta Fase 3. El arranque del 27 de Febrero de 2020 con 53 segundos, sin embargo, si se seguirá tratando como anómalo ya que fue tratado como anómalo por consumo demasiado alto anteriormente.

Una característica que tienen la mayoría de los arranques correspondientes al nuevo modelo de arranque, es la existencia de un doble escalón al inicio de la Fase 3, tal y como se observó en las figura [5.10b]. El 80 % de los arranques iniciales tienen esta característica al inicio de la Fase 3. No obstante, es una característica para confirmar la relación con el nuevo modelo de arranque. La no obtención de dicho comportamiento no se trata como anómalo.

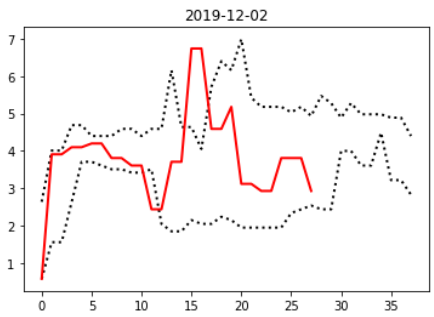
Para determinar la adecuación de los nuevos arranques al nuevo modelo de arranque, se han comparado los valores de consumo en cada segundo de los nuevos arranques de prueba con respecto a los consumos observados en la Figura [5.9], correspondiente al nuevo modelo de arranque. Se ha determinado como mínimo al menos un 80 % de acierto de los valores de consumo en cada segundo de los nuevos arranques comparando con los valores de consumo del nuevo modelo de arranque. No obstante, el % restante de valores de consumo energético que no se correspondan con el modelo, deberá ser por una diferencia mínima con respecto a los intervalos del modelo. Sin embargo, se permitirá la existencia de hasta 3 anomalías leves que se correspondan con una desviación del modelo más influyente sin llegar a traducirse en anomalía grave. A continuación se expone los días y el número de anomalías leves encontradas en los nuevos arranques(sin tener en cuenta los 4 arranques considerados anómalos por el consumo excesivo):

| | **Anomalías Leves** |
|---|---|
| **Arranque 2 de Diciembre de 2019** | 3<br>- Segundos 11,15 y 16 |
| **Arranque 5 de Diciembre de 2019** | 3<br>- Segundos 31,32 y 34 |
| **Arranques 5 de Marzo de 2020** | 5<br>- Segundos 25,26,27,28,37 |
| **Arranque 25 de Marzo de 2020** | 2<br>- Segundos 27,28 |

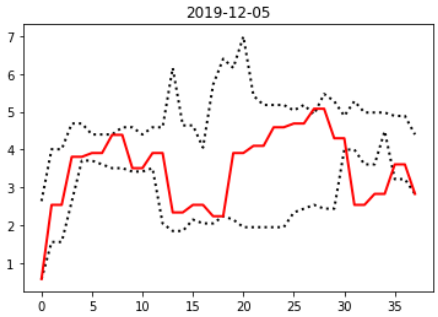
Cuadro 5.3: Información sobre Anomalías leves en fase 3 en los arranques

En la tabla adjunta, podemos observar como los arranques, que en un principio se podían considerar anómalos por su duración excesiva de Fase 3( 5 y 25 Marzo de 2020), además contienen 5 y 2 anomalías respectivamente. Por tanto, en este caso se confirma su clasificación como anómalo. Sin embargo, no se tratarán como anómalos los días 2 y 5 de Diciembre de 2019 porque como vemos en las Figuras siguientes[5.21a] y [5.21b], no se trata de desviaciones significantes. El resto de arranques no contienen anomalías leves y se adecúan correctamente a los intervalos del nuevo

modelo de arranque.




(a) 2 Diciembre 2019 - Fase 3




(b) 5 Diciembre 2019 - Fase 3

Figura 5.21: Arranques 2 y 5 de Diciembre de 2019 en Fase 3

Para hacer un resumen de las características de los arranques en relación al nuevo modelo de arranque, se añade una tabla informativa.Todos los porcentajes se realizan en base al total de arranques.

| Arranques | Escalón Fase 3 | Doble Escalón Fase 3 | Umbral Consumo Inicial Fase 3 | Umbral Consumo Final Fase 3 | Comportamiento Fase 3 >80 % Acierto | Umbral Duración Fase 3 |
|---|---|---|---|---|---|---|
| Nuevo Modelo Arranque | 100 % | 80 % | 95 % | 85 % | 80 % | 85 % |

Cuadro 5.4: Porcentajes asociados al nuevo modelo de arranque en Fase 3

**Fase 4**

Las características de esta fase se asocian a un comportamiento del consumo estable en la que la máquina ya se encuentra preparada para realizar operaciones. Todos los arranques contienen una duración inter-polada de 60 segundos por lo que únicamente será interesante los valores de consumo entre los que oscila.
Al ser una fase de arranque en la que la variabilidad de consumo oscila sin una tendencia clara, se determina que al menos el 75 % de los valores de consumo de los nuevos arranques se adecuén al nuevo modelo de comportamiento para la Fase 4. El % restante será con valores de consumo cercanos a los intervalos existentes del nuevo modelo de arranque para la Fase 4. Como se refleja en la tabla siguiente, el 95 % de los arranques se comportan de acuerdo a los umbrales establecidos. El arranque restante (5 % restante) está asociado al día 19 de Febrero, el cual fue determinado como anómalo anteriormente.

| **Arranques** | **Comportamiento Fase 4 Acierto >75 %** |
|---|---|
| Nuevo Modelo Arranque | 95 % |

Cuadro 5.5: Porcentajes asociados al nuevo modelo de arranque en Fase 4

**Arranque Global**

Atendiendo al análisis por fases realizado, comparando el comportamiento de los nuevos arranques con respecto a las características o patrones más comunes del nuevo modelo de arranque, se resumen en la tabla como han sido considerados cada uno de los arranques diarios con respecto a su comportamiento.

| | **Arranques** |
|---|---|
| **Comportamiento Nuevo Modelo Arranque** | 2,3,4,5 Diciembre<br>18,20,21,25,26 Febrero<br>6,11,12,24 Marzo |
| **Anómalo( Consumo o Duración)** | 19Febrero,27 Febrero,2 Marzo,<br>5 Marzo,20 Marzo, 25 Marzo |
| **Comportamiento Modelo 1** | 20 Diciembre |

Cuadro 5.6: Clasificación de los arranques según comportamiento

En la tabla adjunta, se representan los porcentajes asociados a la clasificación como arranque correcto o arranque anómalo.

| **Arranques** | **Comportamiento Nuevo Modelo Arranque** | **Anomalías (Consumo o Duración)** | **Comportamiento Modelo 1** |
|---|---|---|---|
| Total Arranques | 65 % | 30 % | 5 % |

Cuadro 5.7: Comparación de porcentajes de arranques con comportamiento más representativo,anómalos o asociados al Modelo 1

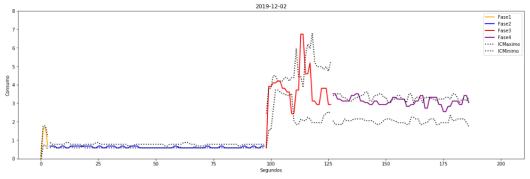
Se observa que del total de arranques(20),el 65 % de ellos se han comportado de acuerdo al patrón de arranque más representativo implementado en la sección anterior. Sin embargo, se ha determinado la existencia de 6 arranques con la existencia de una o más anomalías graves que no permitían clasificar el arranque como correcto. Entre las causas se encuentran un consumo elevado,una duración prolongada de la Fase 3 del arranque o la mezcla de ambas.

Por tanto, obviando los arranques que se han clasificado como anómalos, 13 de los 14 arranques correctos que se han determinado pertenecen al nuevo modelo de arranque, lo que hace casi un 93 % de acierto. La tabla con los porcentajes de acierto para cada característica en cada arranque se encuentra en Anexos [A.10]. Las características de Fase 1 Grupo, Fase 2 Grupo, Fase 3 Grupo y Fase 4 Grupo hacen referencia al porcentaje de observaciones que se encuentran dentro de los intervalos para cada una de las Fases del nuevo modelo de arranque. En la columna final, se detalla con una variable booleana los arranques que han sido considerados anómalos o no de acuerdo a las características anteriores.

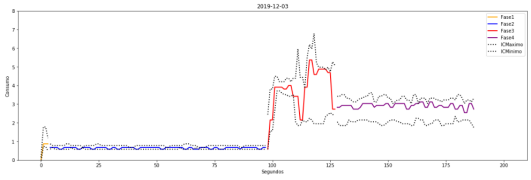
A continuación se adjuntan el comportamiento global de todo el arranque de los 14 arranques que se han estimado como normales o correctos. 13 de ellos se representan con las desviaciones propias del nuevo modelo de arranque y el restante con las desviaciones propias del Modelo 1. En la sección de Anexos [A.3.6], se puede observar una simulación segundo a segundo con gráficas con información sobre las características tratadas para un arranque, concretamente del 3 de Diciembre de 2019. Además, se resume el porcentaje de acierto de cada fase con respecto al modelo más

representativo, así como la existencia o no de las distintas características comentadas.

**Arranques Correctos - Modelo más Representativo**

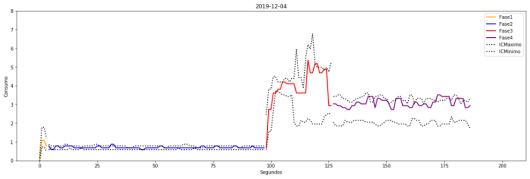


(a) 2 Diciembre de 2019

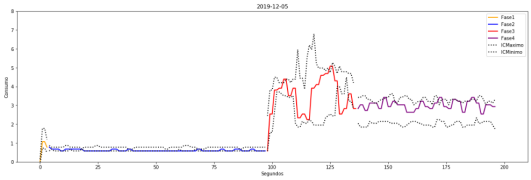


(b) 3 Diciembre de 2019

Figura 5.22: Arranques Correctos 2 y 3 de Diciembre de 2019

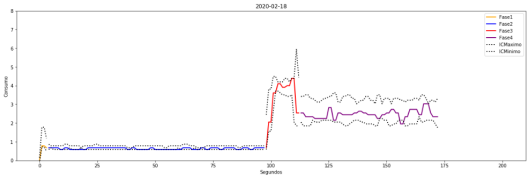


(a) 4 Diciembre de 2019

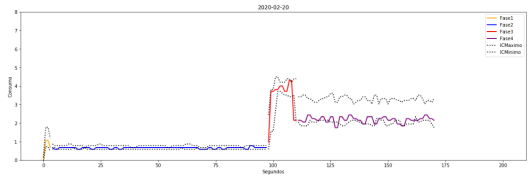


(b) 5 Diciembre de 2019

Figura 5.23: Arranques Correctos 4 y 5 de Diciembre de 2019

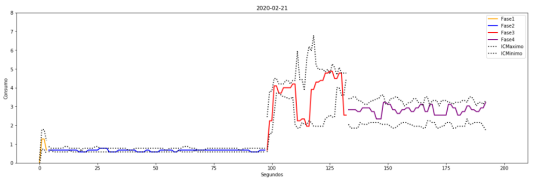


(a) 18 Febrero de 2020

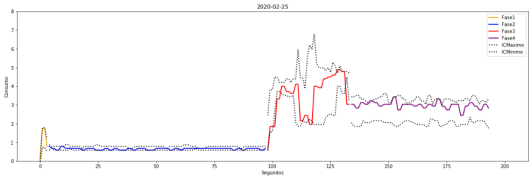


(b) 20 Febrero de 2020

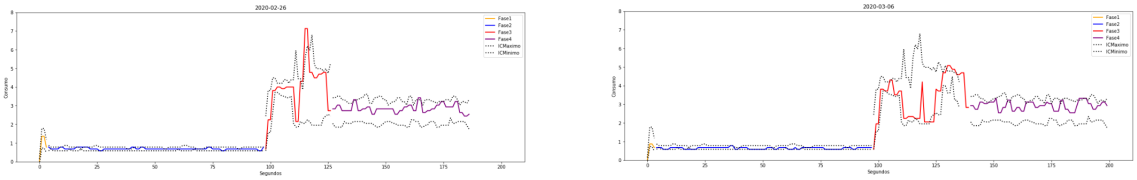
Figura 5.24: Arranques Correctos 18 y 20 de Febrero de 2020



(a) 21 Febrero de 2020

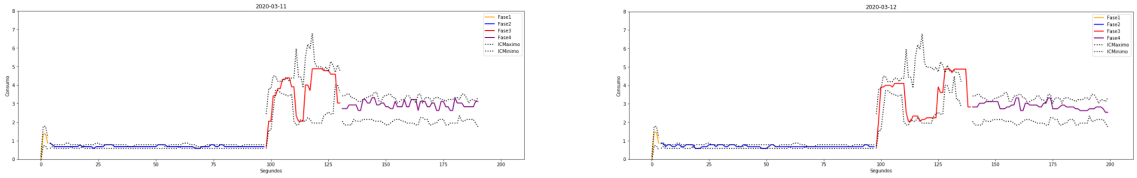


(b) 25 Febrero de 2020



(a) 26 Febrero de 2020 (b) 6 Marzo de 2020

Figura 5.26: Arranques Correctos 26 Febrero y 6 Marzo de 2020



(a) 11 Marzo de 2020 (b) 12 Marzo de 2020

Figura 5.27: Arranques Correctos 11 y 12 Marzo de 2020

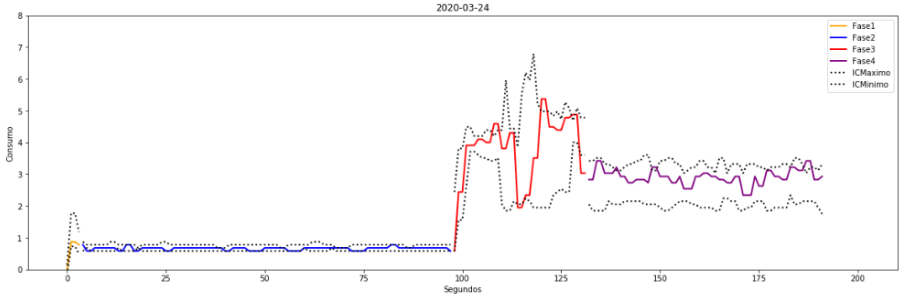



Figura 5.28: Arranque Correcto 24 Marzo de 2020

### Arranque característico Modelo 1

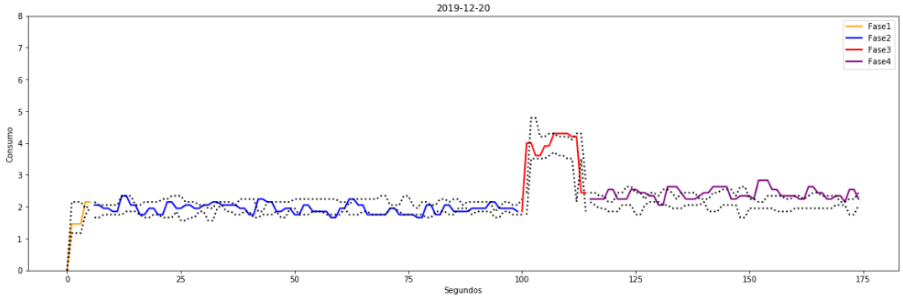



Figura 5.29: Arranque Correcto de Modelo 1 - 20 Diciembre de 2019

# Capítulo 6

# Conclusiones del experimento y líneas de trabajo futuras

El desarrollo de este trabajo fin de grado ha supuesto la realización de una metodología de aprendizaje basada en procedimientos iterativos como el preprocesamiento de los datos, segmentación y caracterización de las distintas fases de operación del arranque de una máquina industrial, utilización de técnicas de aprendizaje automático, extracción de conocimiento a los resultados propuestos por los algoritmos y una simulación paso a paso con nuevos arranques que han permitido presentar el comportamiento más representativo del arranque de una fresadora. Además, permite poder detectar la existencia de anomalías en cada segundo del arranque.

En este trabajo puede explicarse el proceso de aprendizaje en tres iteraciones. La **primera** en la que no existían datos etiquetados sobre el comportamiento correcto o anómalo de los arranques. Por tanto, observando y analizando las mismas características para todos los arranques se aplicó aprendizaje no supervisado para determinar grupos de acuerdo a los comportamientos del arranque. Con el etiquetado, en este caso de 4 grupos utilizando el algoritmo de clasificación KMeans, se llevó a cabo una **segunda** iteración encargada en extraer patrones de varios de los 4 grupos o modelos representativos: Figura [6.1].

La extracción de conocimiento aplicada a los resultados aportados por las técnicas de aprendizaje automático han llevado a poder explicar el comportamiento de 3 de los 4 modelos con un único modelo de arranque característico. En este caso, el arranque de una fresadora es difícil que pueda explicarse de cuatro formas diferentes a lo largo del tiempo ya que las máquinas comúnmente tienen un modelo de arranque que se repite con mayor frecuencia. Acorde a los patrones encontrados en 3 de los 4 modelos en la mayoría de las fases y los segmentos de las fases de arranque, se creó un nuevo arranque más representativo, lo cual lleva a explicar el modo en el que la fresadora arranca de manera más frecuente.

Con esta información, se realiza una **tercera** iteración en la que se puede etiquetar de acuerdo al comportamiento de los nuevos arranques si se asocia a un arranque correcto( el patrón de arranque más representativo que se ha formado) o sin embargo contiene anomalías durante el arranque que hacen etiquetarlo como un arranque no correcto o como un tipo de arranque no frecuente, diferente al arranque característico. Podría asociarse esta evolución del aprendizaje en un 'aprendizaje semi-supervisado' en la que se utilizan los dos tipos de aprendizaje, inicialmente no supervisado y posteriormente 'supervisado.

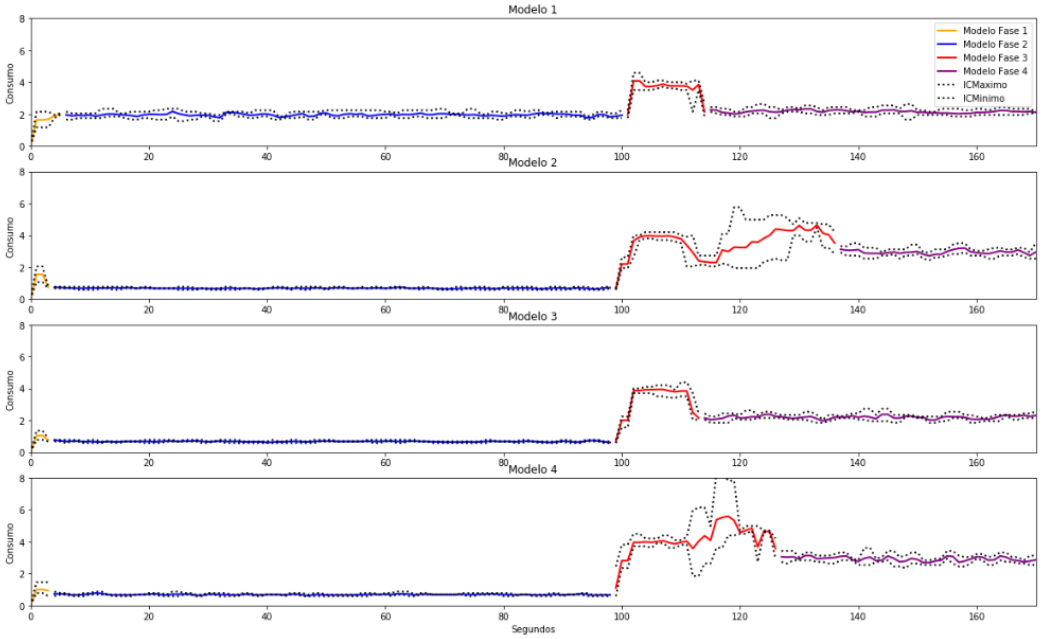
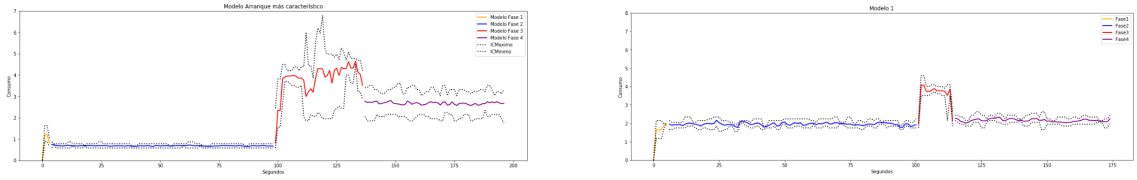



Figura 6.1: Modelos de arranque tras clustering



(a) Modelo arranque característico

(b) Modelo arranque 1

Figura 6.2: Modelos de arranque final posibles extraídos

Gracias a la creación de un nuevo modelo de arranque más representativo, como se puede ver en la Figura [6.2a], se puede comprobar segundo a segundo del arranque si se comporta correctamente o contiene anomalías. No obstante, existe otro modo de arranque de la fresadora que se comporta distinto en relación al consumo energético. Sin embargo, la frecuencia con la que sucede este tipo de arranques es baja. Este modelo de arranque se corresponde con el Modelo 1 [5.17] y se podría analizar de manera más detallada en futuras líneas de investigación usando otro equipo de medida más preciso y la colaboración de los operarios de mantenimiento de la máquina.

Por tanto, determinar segundo a segundo si un arranque se comporta correctamente o acorde al arranque más representativo, permite cumplir el objetivo de este trabajo fin de grado, pudiendo detectar anomalías durante cada segundo del arranque y poder avisar al operario al instante sobre las posibles incidencias que ocurren el arranque de la fresadora y que pueden repercutir en la vida útil de la fresadora o en el proceso de producción diario de la máquina realizando operaciones posteriormente. Todo ello se realiza sin afectar al proceso físico de la máquina.

De acuerdo a las características o patrones encontrados de varios de los modelos, se ofrece en la tabla con arranques de prueba [A.10] en la sección de Anexos [A.2.10] si contenían las patrones

encontrados o no, indicando con *True o False*. Se puede concluir además de observar el comportamiento de la Figura [6.1], la existencia de un escalón al inicio del arranque en la Fase 1 en la que el consumo no excediera los 2.3 Amperios. Además un comportamiento del consumo energético estable en la Fase 2 de entre 0.58 y 0.78 Amperios. Se espera que esta fase de estabilización sea lo más corta posible(como 95 segundos) para no proporcionar información irrelevante durante muchos segundos. Tras esta fase, es común encontrar un doble escalón al inicio de la Fase 3 del arranque, en la que tarda unos 5 segundos en alcanzar un pico de consumo. La duración de la Fase 3 se espera menor de 43 segundos y con consumo menor de 8 Amperios, alcanzando uno o dos picos de corriente durante la evolución de la Fase. Para finalizar se espera una fase estable de consumo entre 2 y 3.5 Amperios con variaciones de consumo energético no superiores a 0.75 Amperios. A modo resumen, se espera al menos un 70 % de los arranques que se realizan diariamente, se comporten de acuerdo a los patrones encontrados y siguiendo el comportamiento de la Figura [5.1].

Con información de un número más elevado de muestras/arranques, sería posible afinar más con el comportamiento de arranque más representativo y quizás el número de clusters óptimo sería diferente.

Otro factor influyente en el análisis realizado es la interpolación de un tiempo fijo para la Fase 4 para todos los arranques. Hacer un análisis más a fondo sobre esta Fase y poder utilizar un tiempo característico mayor sería una linea a seguir, así como utilizar las variables asociadas al valor de consumo medio o máximo de la Fase 4 del arranque.

La disponibilidad de un entorno real donde extraer y caracterizar un conjunto de datos ha sido de gran ayuda para el proceso de aprendizaje y validación de la metodología, lo que permite una base sólida para continuar futuros trabajos.

Como líneas de trabajo futuro, dado que únicamente se ha utilizado una variable(consumo de corriente) para estudiar y caracterizar el proceso de aprendizaje de la fresadora, sería posible profundizar el comportamiento de otras operaciones máquina a partir de la metodología desarrollada. Por ejemplo, las operaciones de fabricación de piezas tomando como referencia las vibraciones del cabezal de la fresadora sobre diferentes materiales durante el proceso de extrusión (mecanizado de componentes). De esta forma se pueden detectar anomalías de forma predictivas durante el proceso de fabricación atendiendo a criterios o patrones previos.

Igualmente la metodología desarrollada es aplicable en otros entornos industriales a partir del análisis del arranque para comprobar comportamiento,determinando anomalías y el modelo de arranque más representativo.

# Bibliografía


[1] ACAN. "LA INDUSTRIA 4.0 . Tecnologías habilitadoras". En: *Asociación clúster de automoción de navarra* (oct. de 2017).

[2] EAU Automation. "El papel del aprendizaje automático en la industria". En: *Automated* (2019).

[3] B.Sniderman, M.Mahto y M.J.Cotteler. "Industry 4.0 and manufacturing ecosystems". En: *Deloitte University Press* (2016).

[4] Francisco Ballesteros. "La estrategia predictiva en el mantenimiento industrial". En: *Preditécnico,IRM* (2017).

[5] Francisco Ballesteros. "Tipos de Mantenimiento". En: *Linkedln* (2017).

[6] Ahmed Banafa. "¿ Qué es el aprendizaje profundo?" En: *OpenMind BBVA* (2016).

[7] Elena Casado Barahona. "Qué es cloud computing: ventajas, origen y aplicaciones". En: *ICEMD* (mar. de 2018).

[8] Chandan K. Reddy Charu C. Aggarwal. *Data Clustering: Algorithms and Applications*. 1.ª ed. CRC Press, 2014. ISBN: 9781466558229.

[9] Alberto Maisueche Cuadrado. "Utilizacion del Machine Learning en Industria 4.0". En: *Universidad de Valladolid, Escuela de Ingenieros Industriales* (sep. de 2019).

[10] Carlos E.Torres. "Manual análisis de vibraciones". En: *Power-MI* (2018).

[11] Real Academia Española. *Mantenimiento*. `https://dle.rae.es/mantenimiento`. 2020.

[12] Álvaro García García. "Desarrollo de modelos predictivos adaptados a los nuevos ecosistemas digitales de mantenimiento industrial". En: *Revista Preditécnico(Preditec - Grupo Álava) número 24* (ene. de 2020).

[13] Salvador García y col. "Big Data: Preprocesamiento y calidad de datos". En: *Big Data monografia* (2016).

[14] Gartner. "Confront Key Challenges to Boost Digital Twin Success". En: *Gartner* (2018).

[15] Geinfor. *Utilización del Machine Learning en la industria 4.0*. `https://geinfor.com/business/utilizacion-del-machine-learning-en-la-industria-4-0/`.

[16] Michael Grives. "Origins of the digital twin concept". En: *ResearchGate* (2016).

[17] De Máquinas y Herramientas. *Introducción a la tecnologia CNC*. `https://www.demaquinasyherramient` `com/mecanizado/introduccion-a-la-tecnologia-cnc`. 2015.

[18] CC.OO Industria. "La digitalización y la industria 4.0". En: *Secretaría de Estrategias Industriales* (sep. de 2017).

[19] Investigate to Innovate. "Gemelos Digitales en la transición a la industria 4.0". En: *Digital Twins* (2018).

[20] Gareth James, Daniela Witten y Robert Tibshirani Trevor Hastie. *An Introduction to Statistical Learning*. Springer, 2013.

[21] Kaggle. *Introducción a ML*. `https://www.kaggle.com/rafanovello/introdu-o-a-ml`. 2018.

[22] Lawtomated. *Supervised Learning vs Unsupervised Learning. Which is better?* `https://lawtomated.com/supervised-vs-unsupervised-learning-which-is-better`. 2019.

[23] Theo Lins y col. "Industry 4.0 Retrofitting". En: *Federal University of Ouro Preto, Federal University of Lavras and University of Coimbra* (2018).

[24] Antonio Lorenzo. "Siemens propone su 'Gemelo Digital' para reducir los costes de fabricación". En: *El Economista* (oct. de 2019).

[25] J.P y O.G. "Llegan las fábricas sin cables". En: *El Pais y Telefónica* (2020).

[26] Maricela Ochoa. "¿Qué es Edge Computing y por qué es relevante para las empresas?" En: *IT Masters Mag* (ago. de 2018).

[27] Jesús Arturo Orozco. *Machine Learning y su importancia en la actualidad*. `https://www.ipade.mx/2018/08/30/machine-learning-y-su-importancia-en-la-actualidad/`. Ago. de 2018.

[28] Pang-NingTan, Michael Steinbach y VipinKumar. *Introduction to Data Mining*. Addison-Wesley, 2006.

[29] Chris Piech. *K Means*. `https://stanford.edu/~cpiech/cs221/handouts/kmeans.html`. Nov. de 2013.

[30] Empresa Preditec. *Mantenimiento Predictivo*. `http://www.preditec.com/mantenimiento-predictivo/`.

[31] PWC. "Industry 4.0: Building the digital enterprise". En: *PWC* (2016).

[32] Aníbal Reñones y col. "¿ Qué pueden hacer los sensores inteligentes en una planta de fabricación?" En: *MetalMecánica* (2012).

[33] Paloma Recuero de los Santos. "Tipos de aprendizaje en Machine Learning: supervisado y no supervisado". En: *Luca Telefónica, Data Unit* (nov. de 2017).

[34] Keith Shaw. "What is edge computing and why it matters". En: *Network World* (nov. de 2019).

[35] Software y Soluciones de Analítica. *Aprendizaje Automático*. `https://www.sas.com/es_es/insights/analytics/machine-learning.html`.

[36] ETRR Techint. *Máquina Convencional vs Maquina CNC*. `https://www.youtube.com/watch?v=bqm_yRGQi0E&list=PL65tKRE6BtYL74Ki22dz8ltPvADN0t1fG&index=1`.

[37] Jose Luis del Val Roman. *Industria 4.0: La transformación digital de la industria*. Inf. téc. Universidad de Deusto, Conferencia de Directores y Decanos de la Ingeniería Informática.

[38] Satec: Universidad de Valladolid. *Industria 4.0: Nuevas oportunidades para empresas y profesionales TIC*. 2018.

[39] Mario Cruz Vega, Pablo Oliete Vivas y Christian Morales Ríos. "Las tecnologias IoT dentro de la industria conectada". En: *Escuela de Organización Industrial, PwC* (2015).

[40] Wikipedia. *Control numérico por computadora*. `https://en.wikipedia.org/wiki/Numerical_control`.

[41] Wikipedia. *Fresadora*. `https://es.wikipedia.org/wiki/Fresadora`.

[42] Wikipedia. *Inteligencia Artificial*. `https://es.wikipedia.org/wiki/Inteligencia_artificial`.

[43] Wikipedia. *Motor y sistema trifásico*. `https://en.wikipedia.org/wiki/Three-phase`. 2020.

# Apéndice A

# Anexos

## A.1 Funciones

### A.1.1 Función para encontrar el momento del inicio del arranque entre dos franjas horarias pasadas como argumentos

```
def encontrararranque(datos,horainicial,horafinal):
    indiceshorainicial=np.argwhere((datos["Fecha y hora"]
    ==horainicial)==True)
    indiceshorafinal=np.argwhere((datos["Fecha y hora"]
    ==horafinal)==True)
    indini=indiceshorainicial[0][0]
    indfin=indiceshorafinal[0][0]
    cont=0
    for i in range(indini,indfin):
        if(datos["CNC_CorrienteFase3"].iloc[i]!= 0.0):
            cont=i-1
            break
    return cont
```

### A.1.2 Función para detección del comportamiento en cada segundo( Etiquetado de 0 a 8) y diferenciar fases de operación

```
tendencia <- function(arr){

    valor<-rep(0,length(arr))
    for(i in 2:(length(arr)-2)){
        incrDESP<-arr[i+1]-arr[i]
        incrANTES<-arr[i]-arr[i-1]
        if(abs(incrDESP)<=0.75){
            if(abs(incrANTES)<=0.75){ valor[i]=0}
            else {
```

```
                if(incrANTES<0){ valor[i]=1}
                else { valor[i]=2 }
            }

        }else{
            if(incrDESP>0){
                if(abs(incrANTES)<=0.75){ valor[i]=4}
                else{
                    if(incrANTES<0){ valor[i]=3}
                    if(incrANTES>0){ valor[i]=7}
                }
            }else{
                if(abs(incrANTES)<=0.75){ valor[i]=6}
                }else{
                    if(incrANTES<0){valor[i]=5}
                    if(incrANTES>0){valor[i]=8}
                }
            }
        }
    return(valor)
}
```

### A.1.3 Función para normalizar en base al valor medio de los valores máximos de la Fase 1 del arranque,indicando como argumentos los datos reales agrupados y los valores de los datos normalizados

```
datosnormalizados1=pd.DataFrame({'Dia':np.zeros(long),
'CNC_CorrienteFase3':np.zeros(long)})
def normalizefase1todos(datosgrandes,datosest):
    col='CNC_CorrienteFase3'
    a="Dia"
    b="Media del Maximo"
    minimo=0
    datosnormalizados1[a]=datosgrandes[a]
    maximo = datosest[b]
    datosnormalizados1[col] = datosgrandes["FASE1"].apply
    (lambda x: (x-minimo)/(maximo-minimo))
normalizefase1todos(agrupar,datosfase1global)
```

### A.1.4 Complete Linkage - Ejemplo de gráfica del codo y matriz de linkage

```
from scipy.cluster.hierarchy import linkage
from matplotlib import pyplot as plt
from scipy.cluster import hierarchy
```

```
Z=hierarchy.linkage(fase1,method="complete")# Matriz linkage
# Agrupar la distancia de los ultimos 7 clusters agrupados
ultimos=Z[-7:,2]
may_men=ultimos[::-1]## Ordenarlos de mayor a menor
plt.xlabe("Clusters")
plt.ylabel("Ssw")
plt.plot(np.arange(1,len(ultimos)+1),may_men,"bx-",color="green")
```

### A.1.5 Dendograma Complete linkage

```
from scipy.cluster.hierarchy import dendrogram, linkage
from matplotlib import pyplot as plt
from scipy.cluster import hierarchy
names=["13Ene","16Ene","17Ene","21Ene","22Ene","24Ene",
"27Ene","29Ene","30Ene","31Ene","3Feb","4Feb","5Feb",
"6Feb","7Feb","10Feb","11Feb","13Feb","14Feb"]
plt.figure(figsize=(15,5))
# Matriz de linkage con enlace completo
Z=hierarchy.linkage(fase1,method="complete")
# Linea gris a partir de un valor de ssw
#para colorear de un color cada cluster en el dendograma
hierarchy.dendogram(Z,color_threshold=0.28,labels=names)
# Linea gris discontinua delimitando los clusters
plt.axhline(y=0.28,c="grey",lw=1,linestyle="dashed")
plt.show()
```

### A.1.6 Ajuste con algoritmo KMeans para ejemplo de gráfica del codo

```
from sklearn.cluster import KMeans
from sklearn import metrics
from scipy.spatial.distance import cdist
ssw = [] # Suma de los cuadrados internos
for k in range(1,7): # 7 Clusters maximo
    # Ajuste del modelo
    kmeanModel = KMeans(n_clusters=k).fit(fase1)
    # Se crean los centroides
    centers = pd.DataFrame(kmeanModel.cluster_centers_)
    labels = kmeanModel.labels_ # Etiquetado
    # Escoger la minima distancia euclidea entre
    cada obs con cada centroide
    ssw_k = sum(np.min(cdist(fase1, kmeanModel.cluster_centers_,
    "euclidean"), axis = 1))
    ssw.append(ssw_k)
print(ssw)# Para cada uno de los 7 clusters
# Representacion del metodo del codo
plt.plot(range(1,7), ssw, "bx-")
```

```
# bx para indicar con una cruz el valor en cada cluster
plt.xlabel("Clusters")
plt.ylabel("SSw")
plt.show()
```

### A.1.7 Ejemplo de ajuste de KMeans con un número de clusters dado

```
from sklearn.cluster import KMeans
# Optimo 3 clusters
k=3
kmedias=KMeans(n_clusters=k).fit(fase1)
centroides= pd.DataFrame(kmedias.cluster_centers_)
etiquetas= pd.DataFrame({'Dia':datosfiltrados["Fecha."],
'Grupo': kmedias.labels_})
etiquetas
```

## A.2 Tablas

### A.2.1 Índices cambios de fases de los arranques de entrenamiento

| | Fecha. | Inicio. | Fin. | Fase.1.al.2. | Fase.2.al.3. | Fase.3.al.4. |
|---|---|---|---|---|---|---|
| 0 | 2020-01-13 | 1 | 2286 | 3 | 802 | 841 |
| 1 | 2020-01-14 | 2287 | 4686 | 3 | 855 | 889 |
| 2 | 2020-01-15 | 4687 | 6988 | 2 | 799 | 829 |
| 3 | 2020-01-16 | 6989 | 9388 | 3 | 98 | 113 |
| 4 | 2020-01-17 | 9389 | 11788 | 3 | 1145 | 1160 |
| 5 | 2020-01-20 | 11789 | 14188 | 3 | 1991 | 2016 |
| 6 | 2020-01-21 | 14189 | 16588 | 3 | 1060 | 1075 |
| 7 | 2020-01-22 | 16589 | 18988 | 5 | 100 | 115 |
| 8 | 2020-01-23 | 18989 | 21388 | 3 | 811 | 844 |
| 9 | 2020-01-24 | 21389 | 23788 | 3 | 290 | 319 |
| 10 | 2020-01-27 | 23789 | 26183 | 3 | 902 | 917 |
| 11 | 2020-01-28 | 26184 | 28583 | 3 | 623 | 653 |
| 12 | 2020-01-29 | 28584 | 30983 | 3 | 1098 | 1130 |
| 13 | 2020-01-30 | 30984 | 33383 | 3 | 790 | 816 |
| 14 | 2020-01-31 | 33384 | 35783 | 3 | 764 | 796 |
| 15 | 2020-02-03 | 35784 | 38183 | 3 | 962 | 1008 |
| 16 | 2020-02-04 | 38184 | 40558 | 3 | 994 | 1031 |
| 17 | 2020-02-05 | 40559 | 42947 | 3 | 99 | 114 |
| 18 | 2020-02-06 | 42948 | 45347 | 3 | 924 | 956 |
| 19 | 2020-02-07 | 45348 | 47747 | 5 | 100 | 115 |
| 20 | 2020-02-10 | 47748 | 50147 | 5 | 100 | 113 |
| 21 | 2020-02-11 | 50148 | 52547 | 3 | 679 | 705 |
| 22 | 2020-02-13 | 52548 | 54947 | 3 | 835 | 865 |
| 23 | 2020-02-14 | 54948 | 57347 | 3 | 1508 | 1523 |

Figura A.1: Índices asociados al cambio de fases en los arranques de entrenamiento

### A.2.2 Índices cambios de fases de los arranques de test

| | Fecha. | Inicio. | Fin. | Fase.1.al.2. | Fase.2.al.3. | Fase.3.al.4. |
|---|---|---|---|---|---|---|
| 0 | 2019-12-02 | 1 | 2283 | 3 | 844 | 872 |
| 1 | 2019-12-03 | 2284 | 4683 | 3 | 924 | 953 |
| 2 | 2019-12-04 | 4684 | 7083 | 3 | 820 | 848 |
| 3 | 2019-12-05 | 7084 | 9483 | 3 | 744 | 782 |
| 4 | 2019-12-20 | 9484 | 11883 | 5 | 101 | 115 |
| 5 | 2020-02-18 | 11884 | 14218 | 3 | 485 | 499 |
| 6 | 2020-02-19 | 14219 | 16618 | 3 | 571 | 586 |
| 7 | 2020-02-20 | 16619 | 19017 | 3 | 182 | 194 |
| 8 | 2020-02-21 | 19018 | 21269 | 3 | 529 | 563 |
| 9 | 2020-02-25 | 21270 | 23554 | 3 | 410 | 445 |
| 10 | 2020-02-26 | 23555 | 25943 | 3 | 252 | 280 |
| 11 | 2020-02-27 | 25944 | 28277 | 3 | 587 | 640 |
| 12 | 2020-03-02 | 28278 | 30599 | 5 | 99 | 136 |
| 13 | 2020-03-05 | 30600 | 32999 | 3 | 589 | 634 |
| 14 | 2020-03-06 | 33000 | 35399 | 3 | 597 | 638 |
| 15 | 2020-03-11 | 35400 | 37799 | 3 | 491 | 523 |
| 16 | 2020-03-12 | 37800 | 40199 | 3 | 664 | 705 |
| 17 | 2020-03-20 | 40200 | 42599 | 3 | 1333 | 1370 |
| 18 | 2020-03-24 | 42600 | 44884 | 3 | 968 | 1001 |
| 19 | 2020-03-25 | 44885 | 47165 | 3 | 1120 | 1168 |

Figura A.2: Índices asociados al cambio de fases en los arranques de prueba

### A.2.3 Estadísticos Máximo,Media de consumo y duración de fase 1 para arranques de entrenamiento

| | Dia | Fase1 Media | Fase1 Max | Fase1 Duracion |
|---|---|---|---|---|
| 0 | 2020-01-13 | 1.145000 | 1.95 | 3.0 |
| 1 | 2020-01-14 | 2.000000 | 3.61 | 3.0 |
| 2 | 2020-01-15 | 1.166667 | 2.63 | 2.0 |
| 3 | 2020-01-16 | 1.390000 | 1.95 | 3.0 |
| 4 | 2020-01-17 | 0.875000 | 1.36 | 3.0 |
| 5 | 2020-01-20 | 0.605000 | 0.87 | 3.0 |
| 6 | 2020-01-21 | 0.560000 | 0.78 | 3.0 |
| 7 | 2020-01-22 | 1.461667 | 1.95 | 5.0 |
| 8 | 2020-01-23 | 0.630000 | 0.87 | 3.0 |
| 9 | 2020-01-24 | 0.535000 | 0.78 | 3.0 |
| 10 | 2020-01-27 | 0.780000 | 1.17 | 3.0 |
| 11 | 2020-01-28 | 2.807500 | 5.18 | 3.0 |
| 12 | 2020-01-29 | 0.585000 | 0.78 | 3.0 |
| 13 | 2020-01-30 | 0.702500 | 0.97 | 3.0 |
| 14 | 2020-01-31 | 0.705000 | 1.07 | 3.0 |
| 15 | 2020-02-03 | 0.875000 | 1.36 | 3.0 |
| 16 | 2020-02-04 | 0.875000 | 1.36 | 3.0 |
| 17 | 2020-02-05 | 1.612500 | 2.15 | 3.0 |
| 18 | 2020-02-06 | 1.220000 | 2.05 | 3.0 |
| 19 | 2020-02-07 | 1.396667 | 2.05 | 5.0 |
| 20 | 2020-02-10 | 1.201667 | 1.95 | 5.0 |
| 21 | 2020-02-11 | 1.095000 | 1.46 | 3.0 |
| 22 | 2020-02-13 | 0.730000 | 1.07 | 3.0 |
| 23 | 2020-02-14 | 0.652500 | 0.87 | 3.0 |

Figura A.3: Estadísticos de la Fase 1

### A.2.4 Estadísticos Máximo,Media de consumo y duración de fase 2 para arranques de entrenamiento

| | Dia | Fase2 Media | Fase2 Max | Fase2 Duracion |
|---|---|---|---|---|
| 0 | 2020-01-13 | 0.662936 | 0.78 | 799.0 |
| 1 | 2020-01-14 | 0.686471 | 0.78 | 852.0 |
| 2 | 2020-01-15 | 0.667522 | 0.87 | 797.0 |
| 3 | 2020-01-16 | 1.927527 | 2.24 | 95.0 |
| 4 | 2020-01-17 | 0.692912 | 0.87 | 1142.0 |
| 5 | 2020-01-20 | 0.659355 | 0.78 | 1988.0 |
| 6 | 2020-01-21 | 0.633744 | 0.78 | 1057.0 |
| 7 | 2020-01-22 | 1.863763 | 2.15 | 95.0 |
| 8 | 2020-01-23 | 0.685211 | 0.78 | 808.0 |
| 9 | 2020-01-24 | 0.639298 | 0.78 | 287.0 |
| 10 | 2020-01-27 | 0.662274 | 0.78 | 899.0 |
| 11 | 2020-01-28 | 0.676262 | 0.87 | 620.0 |
| 12 | 2020-01-29 | 0.637804 | 0.87 | 1095.0 |
| 13 | 2020-01-30 | 0.713618 | 0.87 | 787.0 |
| 14 | 2020-01-31 | 0.673412 | 0.78 | 761.0 |
| 15 | 2020-02-03 | 0.682665 | 0.87 | 959.0 |
| 16 | 2020-02-04 | 0.679798 | 0.78 | 991.0 |
| 17 | 2020-02-05 | 1.953936 | 2.34 | 96.0 |
| 18 | 2020-02-06 | 0.678259 | 0.78 | 921.0 |
| 19 | 2020-02-07 | 2.019785 | 2.34 | 95.0 |
| 20 | 2020-02-10 | 2.016667 | 2.34 | 95.0 |
| 21 | 2020-02-11 | 0.660712 | 0.78 | 676.0 |
| 22 | 2020-02-13 | 0.643012 | 0.78 | 832.0 |
| 23 | 2020-02-14 | 0.663220 | 0.87 | 1505.0 |

Figura A.4: Estadísticos de la Fase 2

### A.2.5 Estadísticos Máximo,Media de consumo y duración de fase 3 para arranques de entrenamiento

| | Dia | Fase3 Media | Fase3 Max | Fase3 Duracion |
|---|---|---|---|---|
| 0 | 2020-01-13 | 3.108750 | 5.08 | 39.0 |
| 1 | 2020-01-14 | 3.492286 | 4.98 | 34.0 |
| 2 | 2020-01-15 | 3.019355 | 4.78 | 30.0 |
| 3 | 2020-01-16 | 3.336250 | 4.10 | 15.0 |
| 4 | 2020-01-17 | 2.940625 | 4.00 | 15.0 |
| 5 | 2020-01-20 | 3.957308 | 7.13 | 25.0 |
| 6 | 2020-01-21 | 2.847500 | 4.00 | 15.0 |
| 7 | 2020-01-22 | 3.405000 | 4.59 | 15.0 |
| 8 | 2020-01-23 | 3.836471 | 10.06 | 33.0 |
| 9 | 2020-01-24 | 3.833667 | 8.30 | 29.0 |
| 10 | 2020-01-27 | 3.068750 | 4.39 | 15.0 |
| 11 | 2020-01-28 | 3.621935 | 5.27 | 30.0 |
| 12 | 2020-01-29 | 3.683030 | 7.82 | 32.0 |
| 13 | 2020-01-30 | 3.934444 | 6.15 | 26.0 |
| 14 | 2020-01-31 | 3.529091 | 5.08 | 32.0 |
| 15 | 2020-02-03 | 3.239787 | 5.27 | 46.0 |
| 16 | 2020-02-04 | 3.338947 | 4.78 | 37.0 |
| 17 | 2020-02-05 | 3.391250 | 4.39 | 15.0 |
| 18 | 2020-02-06 | 3.591212 | 5.76 | 32.0 |
| 19 | 2020-02-07 | 3.335000 | 4.10 | 15.0 |
| 20 | 2020-02-10 | 3.325714 | 4.10 | 13.0 |
| 21 | 2020-02-11 | 3.971481 | 5.96 | 26.0 |
| 22 | 2020-02-13 | 3.117742 | 4.20 | 30.0 |
| 23 | 2020-02-14 | 3.032500 | 4.20 | 15.0 |

Figura A.5: Estadísticos de la Fase 3

### A.2.6 Variables normalizadas para clustering

| | Fase1 Media | Fase1 Max | Fase1 Duracion | Fase2 Media | Fase2 Max | Fase2 Duracion | Fase3 Media | Fase3 Max | Fase3 Duracion |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.566125 | 0.854015 | 0.0 | 0.020967 | 0.000000 | 0.499291 | 0.014015 | 0.251163 | 0.787879 |
| 1 | 0.793503 | 0.854015 | 0.0 | 0.932573 | 0.935897 | 0.000000 | 0.410646 | 0.023256 | 0.060606 |
| 2 | 0.315545 | 0.423358 | 0.0 | 0.042736 | 0.057692 | 0.742553 | 0.107188 | 0.000000 | 0.060606 |
| 3 | 0.023202 | 0.000000 | 0.0 | 0.000000 | 0.000000 | 0.682270 | 0.000000 | 0.000000 | 0.060606 |
| 4 | 0.860015 | 0.854015 | 1.0 | 0.887800 | 0.878205 | 0.000000 | 0.497556 | 0.137209 | 0.060606 |
| 5 | 0.000000 | 0.000000 | 0.0 | 0.003828 | 0.000000 | 0.136170 | 0.821722 | 1.000000 | 0.484848 |
| 6 | 0.227378 | 0.284672 | 0.0 | 0.020578 | 0.000000 | 0.570213 | 0.241898 | 0.090698 | 0.060606 |
| 7 | 0.046404 | 0.000000 | 0.0 | 0.002861 | 0.057692 | 0.709220 | 0.643008 | 0.888372 | 0.575758 |
| 8 | 0.155452 | 0.138686 | 0.0 | 0.057689 | 0.057692 | 0.490780 | 0.956847 | 0.500000 | 0.393939 |
| 9 | 0.157773 | 0.211679 | 0.0 | 0.028611 | 0.000000 | 0.472340 | 0.468200 | 0.251163 | 0.575758 |
| 10 | 0.315545 | 0.423358 | 0.0 | 0.035282 | 0.057692 | 0.612766 | 0.124603 | 0.295349 | 1.000000 |
| 11 | 0.315545 | 0.423358 | 0.0 | 0.033214 | 0.000000 | 0.635461 | 0.243760 | 0.181395 | 0.727273 |
| 12 | 1.000000 | 1.000000 | 0.0 | 0.952971 | 1.000000 | 0.000709 | 0.459895 | 0.090698 | 0.060606 |
| 13 | 0.635731 | 0.927007 | 0.0 | 0.032105 | 0.000000 | 0.585816 | 0.535361 | 0.409302 | 0.575758 |
| 14 | 0.799691 | 0.927007 | 1.0 | 0.998464 | 1.000000 | 0.000000 | 0.374434 | 0.023256 | 0.060606 |
| 15 | 0.618716 | 0.854015 | 1.0 | 1.000000 | 1.000000 | 0.000000 | 0.387800 | 0.023256 | 0.000000 |
| 16 | 0.519722 | 0.496350 | 0.0 | 0.019457 | 0.000000 | 0.412057 | 1.000000 | 0.455814 | 0.393939 |
| 17 | 0.180974 | 0.211679 | 0.0 | 0.006604 | 0.000000 | 0.522695 | 0.086056 | 0.046512 | 0.515152 |
| 18 | 0.109049 | 0.065693 | 0.0 | 0.021255 | 0.057692 | 1.000000 | 0.220894 | 0.046512 | 0.060606 |

Figura A.6: Variables normalizadas para aplicar clustering

### A.2.7 Matriz de correlaciones con todas las variables utilizadas para el arranque global con todas las fases

| | Fase1 Max | Fase2 Media | Fase3 Media | Fase3 Max | Fase3 Duracion |
|---|---|---|---|---|---|
| Fase1 Max | 1.000000 | 0.725940 | -0.063163 | -0.400436 | -0.186633 |
| Fase2 Media | 0.725940 | 1.000000 | 0.055406 | -0.405485 | -0.576513 |
| Fase3 Media | -0.063163 | 0.055406 | 1.000000 | 0.663919 | 0.006441 |
| Fase3 Max | -0.400436 | -0.405485 | 0.663919 | 1.000000 | 0.495043 |
| Fase3 Duracion | -0.186633 | -0.576513 | 0.006441 | 0.495043 | 1.000000 |

Figura A.7: Correlación todas las variables

### A.2.8 Etiquetado KMeans Fases 1,2 y 3

| | Dia | Grupo |
|---|---|---|
| **0** | 2020-01-13 | 2 |
| **3** | 2020-01-16 | 2 |
| **4** | 2020-01-17 | 1 |
| **6** | 2020-01-21 | 1 |
| **7** | 2020-01-22 | 0 |
| **9** | 2020-01-24 | 1 |
| **10** | 2020-01-27 | 1 |
| **12** | 2020-01-29 | 1 |
| **13** | 2020-01-30 | 1 |
| **14** | 2020-01-31 | 1 |
| **15** | 2020-02-03 | 1 |
| **16** | 2020-02-04 | 1 |
| **17** | 2020-02-05 | 2 |
| **18** | 2020-02-06 | 2 |
| **19** | 2020-02-07 | 0 |
| **20** | 2020-02-10 | 0 |
| **21** | 2020-02-11 | 1 |
| **22** | 2020-02-13 | 1 |
| **23** | 2020-02-14 | 1 |

(a) Etiquetado original KMeans Fase 1

| | Dia | Grupo |
|---|---|---|
| 0 | 2020-01-13 | 0 |
| 3 | 2020-01-16 | 1 |
| 4 | 2020-01-17 | 0 |
| 6 | 2020-01-21 | 0 |
| 7 | 2020-01-22 | 1 |
| 9 | 2020-01-24 | 0 |
| 10 | 2020-01-27 | 0 |
| 12 | 2020-01-29 | 0 |
| 13 | 2020-01-30 | 0 |
| 14 | 2020-01-31 | 0 |
| 15 | 2020-02-03 | 0 |
| 16 | 2020-02-04 | 0 |
| 17 | 2020-02-05 | 1 |
| 18 | 2020-02-06 | 0 |
| 19 | 2020-02-07 | 1 |
| 20 | 2020-02-10 | 1 |
| 21 | 2020-02-11 | 0 |
| 22 | 2020-02-13 | 0 |
| 23 | 2020-02-14 | 0 |

(b) Etiquetado original KMeans Fase 2

| | Dia | Grupo |
|---|---|---|
| 0 | 2020-01-13 | 2 |
| 3 | 2020-01-16 | 1 |
| 4 | 2020-01-17 | 1 |
| 6 | 2020-01-21 | 1 |
| 7 | 2020-01-22 | 1 |
| 9 | 2020-01-24 | 0 |
| 10 | 2020-01-27 | 1 |
| 12 | 2020-01-29 | 0 |
| 13 | 2020-01-30 | 0 |
| 14 | 2020-01-31 | 2 |
| 15 | 2020-02-03 | 2 |
| 16 | 2020-02-04 | 2 |
| 17 | 2020-02-05 | 1 |
| 18 | 2020-02-06 | 0 |
| 19 | 2020-02-07 | 1 |
| 20 | 2020-02-10 | 1 |
| 21 | 2020-02-11 | 0 |
| 22 | 2020-02-13 | 2 |
| 23 | 2020-02-14 | 1 |

(c) Etiquetado original KMeans Fase 3

Figura A.8: Etiquetado KMeans por fases

### A.2.9 Etiquetado KMeans arranque global utilizando todas las fases

| | Dia | Grupo |
|---|---|---|
| 0 | 2020-01-13 | 2 |
| 3 | 2020-01-16 | 0 |
| 4 | 2020-01-17 | 1 |
| 6 | 2020-01-21 | 1 |
| 7 | 2020-01-22 | 0 |
| 9 | 2020-01-24 | 3 |
| 10 | 2020-01-27 | 1 |
| 12 | 2020-01-29 | 3 |
| 13 | 2020-01-30 | 3 |
| 14 | 2020-01-31 | 2 |
| 15 | 2020-02-03 | 2 |
| 16 | 2020-02-04 | 2 |
| 17 | 2020-02-05 | 0 |
| 18 | 2020-02-06 | 2 |
| 19 | 2020-02-07 | 0 |
| 20 | 2020-02-10 | 0 |
| 21 | 2020-02-11 | 3 |
| 22 | 2020-02-13 | 1 |
| 23 | 2020-02-14 | 1 |

Figura A.9: Etiquetado KMeans arranque global

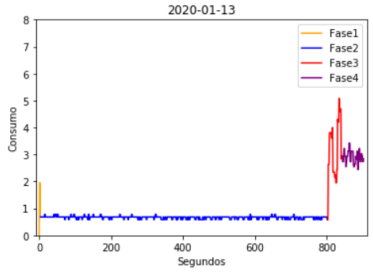
### A.2.10 Tabla Simulación con porcentajes de acierto sobre modelo más representativo

| | Fecha. | Escalon Fase 1 | Consumo Fase 1 | Doble Escalon Fase 1 | Fase 1 Grupo | Fase 2 Grupo | Escalon Fase 3 | Doble Escalon Fase 3 | Consumo Fase 3 Inicial | Consumo Fase 3 Final | Fase 3 Grupo | Duracion Maxima Fase 3 | Anomalias Leves | Fase 4 Grupo | Anomalo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2019-12-02 | True | True | False | 100 | 97.8723 | True | False | True | True | 89.2857 | True | 3 | 75.000000 | False |
| 1 | 2019-12-03 | True | True | False | 100 | 98.9362 | True | True | True | True | 96.5517 | True | 0 | 100.000000 | False |
| 2 | 2019-12-04 | True | True | False | 100 | 94.6809 | True | True | True | True | 96.4286 | True | 0 | 80.000000 | False |
| 3 | 2019-12-05 | True | True | False | 100 | 97.8723 | True | True | True | True | 86.8421 | True | 3 | 90.000000 | False |
| 4 | 2019-12-20 | True | True | True | 0 | 0 | True | False | True | True | 92.8571 | True | 0 | 100.000000 | False |
| 5 | 2020-02-18 | True | True | False | 100 | 95.7447 | True | True | True | True | 92.8571 | True | 0 | 96.666667 | False |
| 6 | 2020-02-19 | True | True | False | 0 | 100 | True | True | Anómalo | True | Anomalo | True | 2 | 71.666667 | True |
| 7 | 2020-02-20 | True | True | False | 100 | 98.9362 | True | False | True | True | 91.6667 | True | 1 | 86.666667 | False |
| 8 | 2020-02-21 | True | True | False | 100 | 98.9362 | True | True | True | True | 91.1765 | True | 1 | 95.000000 | False |
| 9 | 2020-02-25 | True | True | False | 100 | 97.8723 | True | True | True | True | 82.8571 | True | 1 | 98.333333 | False |
| 10 | 2020-02-26 | True | True | False | 100 | 97.8723 | True | True | True | True | 89.2857 | True | 1 | 95.000000 | False |
| 11 | 2020-02-27 | True | True | False | 100 | 100 | True | True | True | Anómalo | Anomalo | False | 7 | 98.333333 | True |
| 12 | 2020-03-02 | True | True | True | 0 | 0 | True | True | True | Anómalo | Anomalo | True | 2 | 93.333333 | True |
| 13 | 2020-03-05 | True | True | False | 100 | 93.617 | True | True | True | True | 78.9474 | False | 5 | 95.000000 | True |
| 14 | 2020-03-06 | True | True | False | 100 | 98.9362 | True | True | True | True | 86.8421 | True | 0 | 90.000000 | False |
| 15 | 2020-03-11 | True | True | False | 100 | 97.8723 | True | True | True | True | 90.625 | True | 0 | 96.666667 | False |
| 16 | 2020-03-12 | True | True | False | 100 | 96.8085 | True | False | True | True | 86.8421 | True | 0 | 100.000000 | False |
| 17 | 2020-03-20 | True | True | False | 100 | 97.8723 | True | True | True | Anómalo | Anomalo | True | 4 | 100.000000 | True |
| 18 | 2020-03-24 | True | True | False | 100 | 96.8085 | True | True | True | True | 81.8182 | True | 0 | 91.666667 | False |
| 19 | 2020-03-25 | True | True | False | 100 | 93.617 | True | True | True | True | 92.1053 | False | 2 | 95.000000 | True |

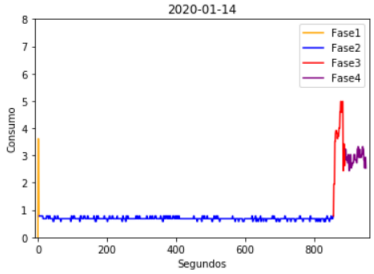
Figura A.10: Porcentajes de Aciertos Arranques de Prueba

## A.3 Figuras

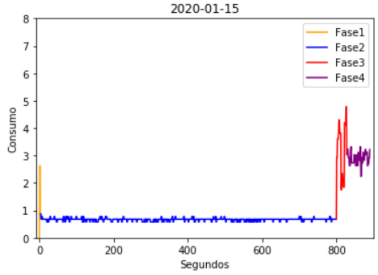
### A.3.1 Arranques reales sin interpolación - datos de entrenamiento



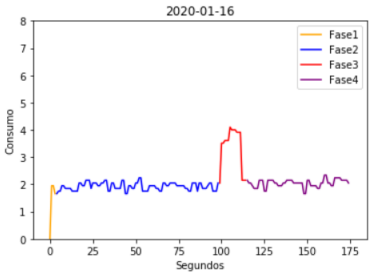

(a) Arranque 13 Enero de 2020



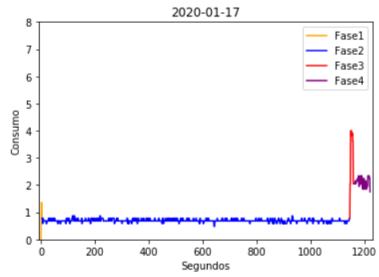

(b) Arranque 14 Enero de 2020



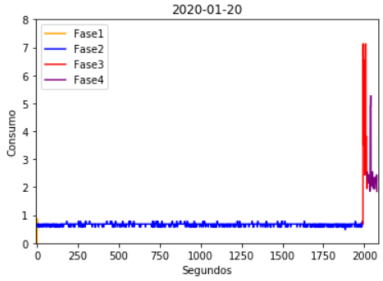

(c) Arranque 15 Enero de 2020



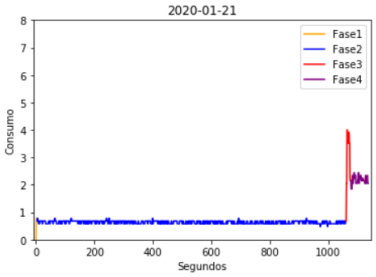

(a) Arranque 16 Enero de 2020



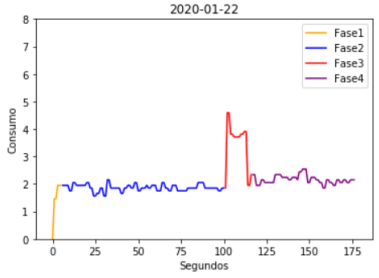

(b) Arranque 17 Enero de 2020



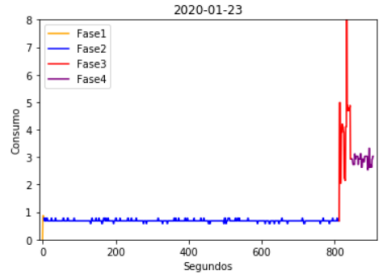

(c) Arranque 20 Enero de 2020



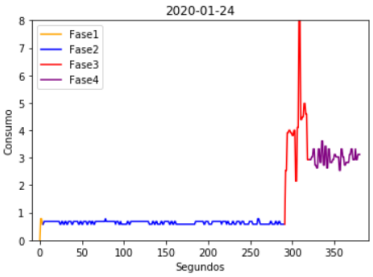

(a) Arranque 21 Enero de 2020



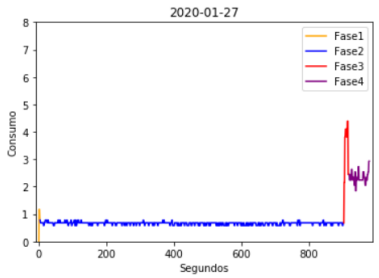

(b) Arranque 22 Enero de 2020



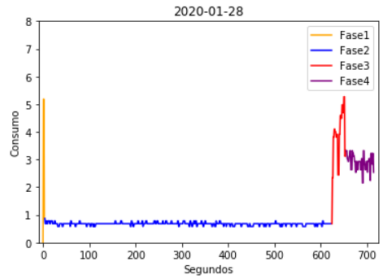

(c) Arranque 23 Enero de 2020



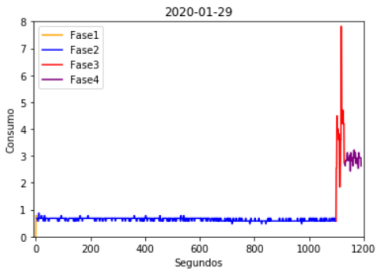

(a) Arranque 24 Enero de 2020



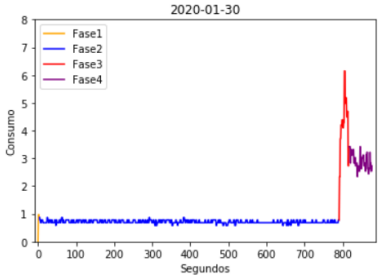

(b) Arranque 27 Enero de 2020



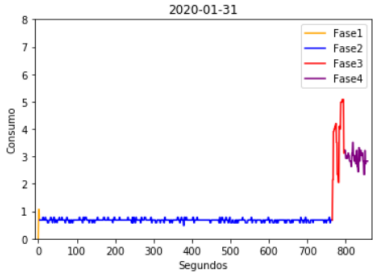

(c) Arranque 28 Enero de 2020



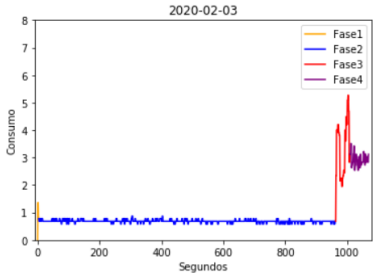

(a) Arranque 29 Enero de 2020



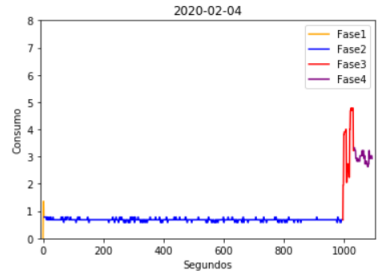

(b) Arranque 30 Enero e 2020



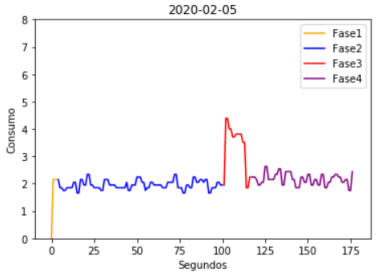

(c) Arranque 31 Enero de 2020



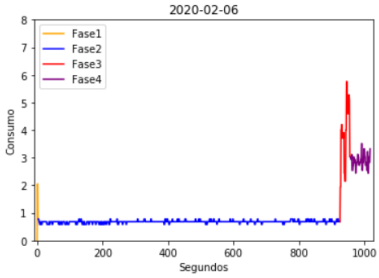

(a) Arranque 3 Febrero de 2020



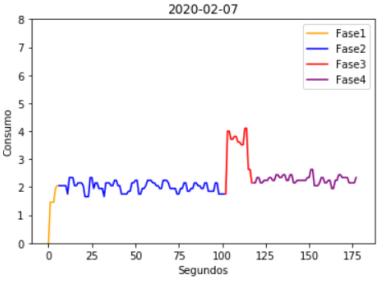

(b) Arranque 4 Febrero de 2020



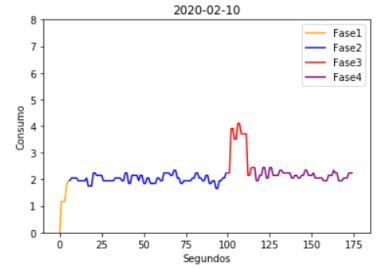

(c) Arranque 5 Febrero de 2020



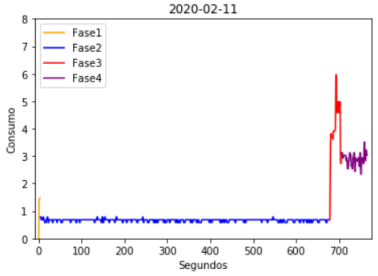

(a) Arranque 6 Febrero de 2020



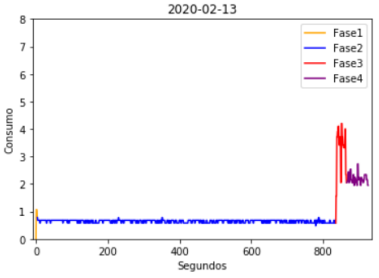

(b) Arranque 7 Febrero de 2020



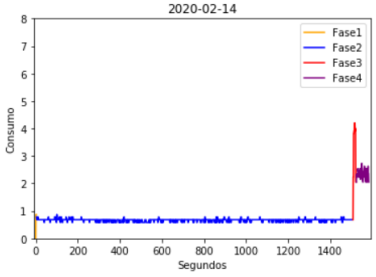

(c) Arranque 10 Febrero de 2020




(a) Arranque 11 Febrero de 2020

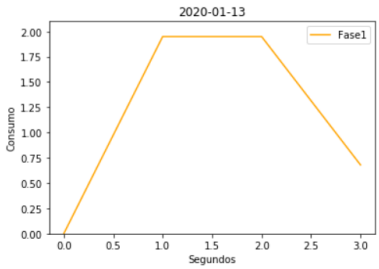



(b) Arranque 13 Febrero de 2020

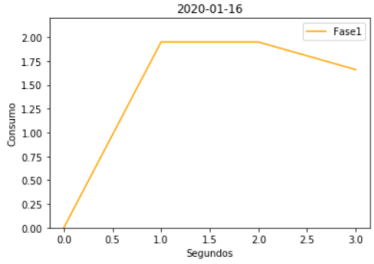



(c) Arranque 14 Febrero de 2020

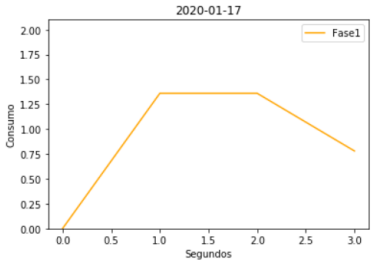
### A.3.2 Arranques de 2020 en Fase 1



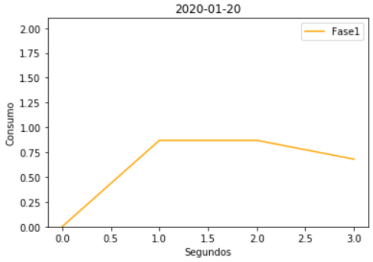

(a) Fase 1 - 13 Enero



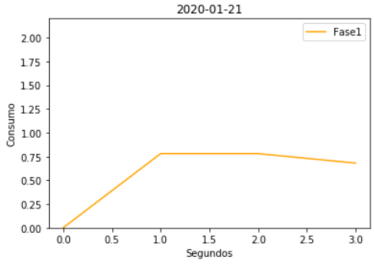

(b) Fase 1 - 16 Enero



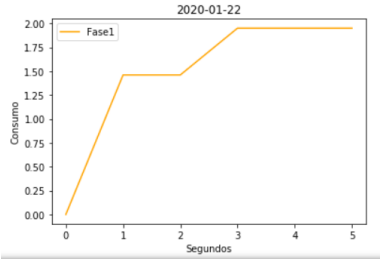

(c) Fase 1 - 17 Enero



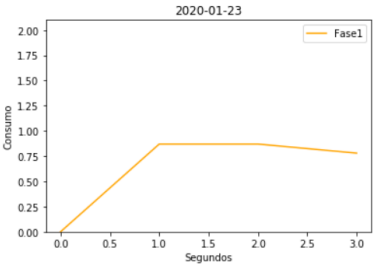

(a) Fase 1 - 20 Enero



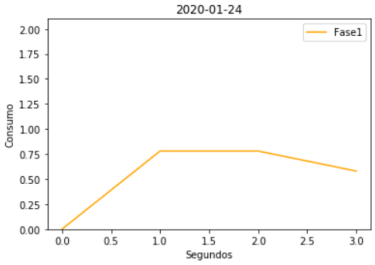

(b) Fase 1 - 21 Enero



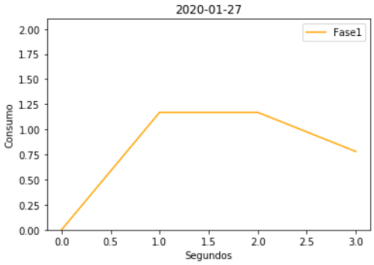

(c) Fase 1 - 22 Enero



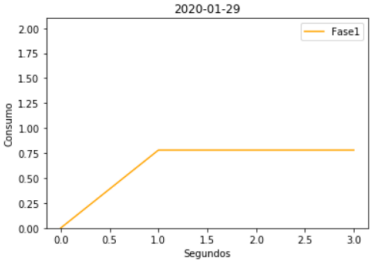

(a) Fase 1 - 23 Enero



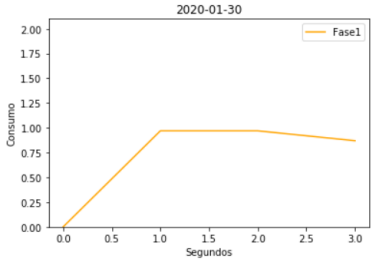

(b) Fase 1 - 24 Enero



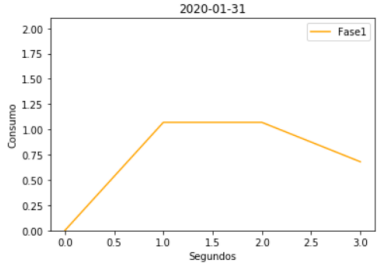

(c) Fase 1 - 27 Enero



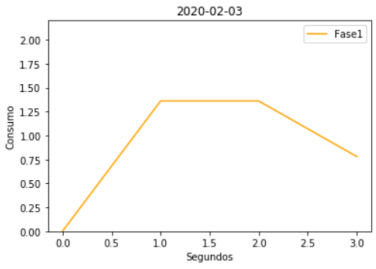

(a) Fase 1 - 29 Enero



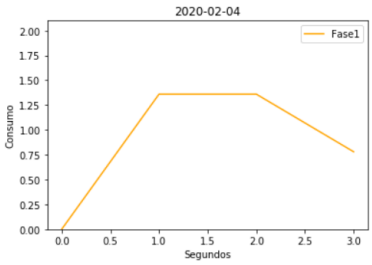

(b) Fase 1 - 30 Enero



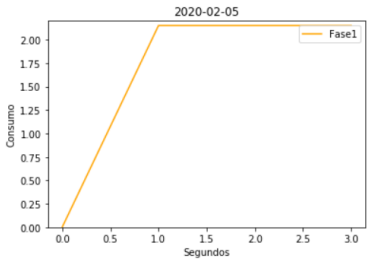

(c) Fase 1 - 31 Enero



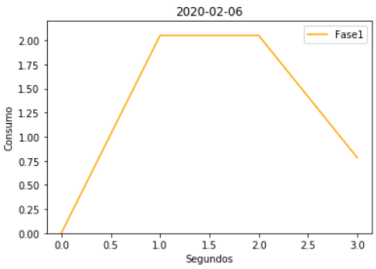

(a) Fase 1 - 3 Febrero



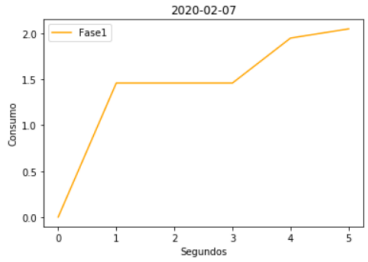

(b) Fase 1 - 4 Febrero



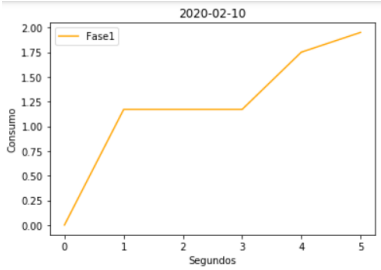

(c) Fase 1 - 5 Febrero



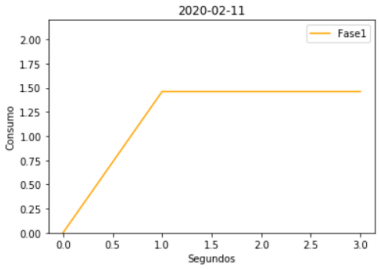

(a) Fase 1 - 6 Febrero



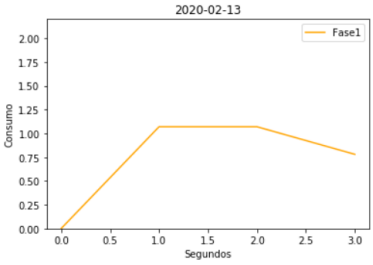

(b) Fase 1 - 7 Febrero



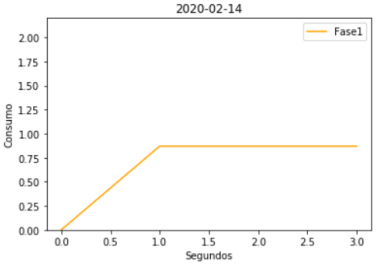

(c) Fase 1 - 10 Febrero




(a) Fase 1 - 11 Febrero

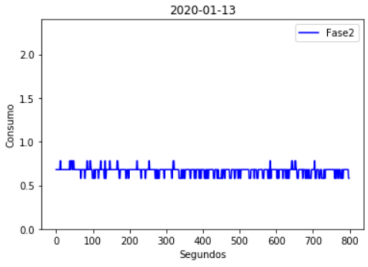



(b) Fase 1 - 13 Febrero

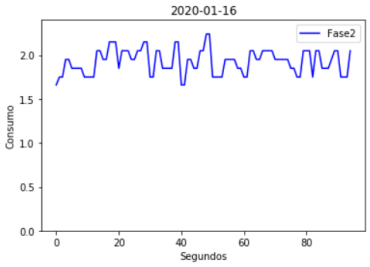



(c) Fase 1 - 14 Febrero

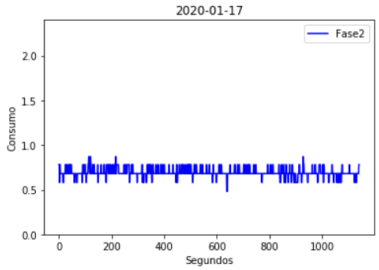
### A.3.3 Arranques de 2020 en Fase 2



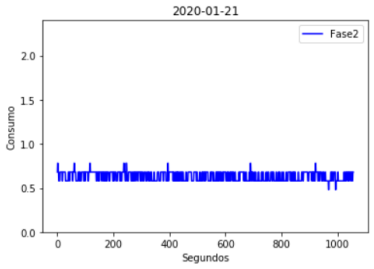

(a) Fase 2 - 13 Enero



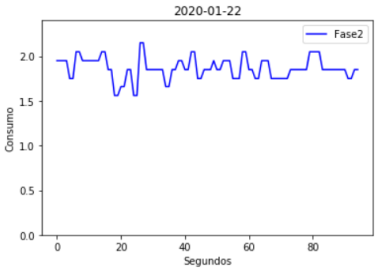

(b) Fase 2 - 16 Enero



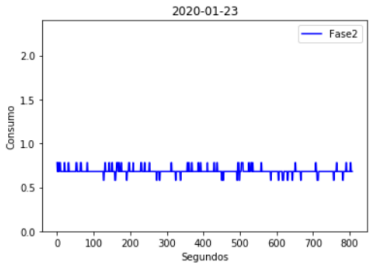

(c) Fase 2 - 17 Enero



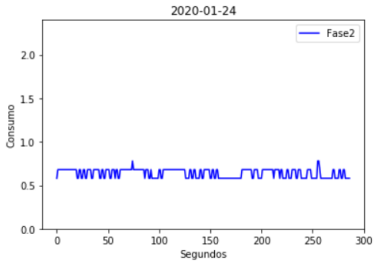

(a) Fase2 - 21 Enero



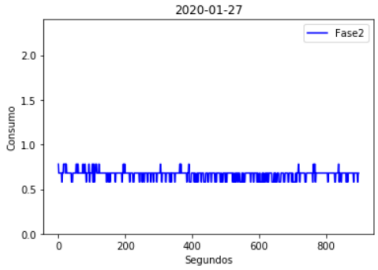

(b) Fase 2 - 22 Enero



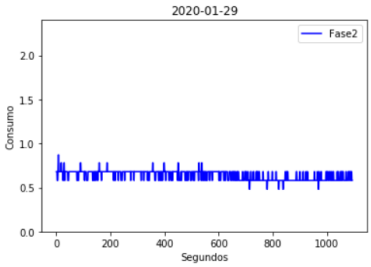

(c) Fase2 - 23 Enero



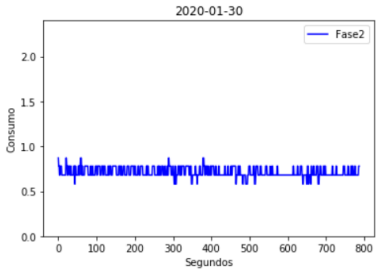

(a) Fase 2 - 24 Enero



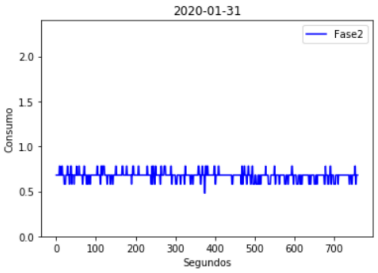

(b) Fase 2 - 27 Enero



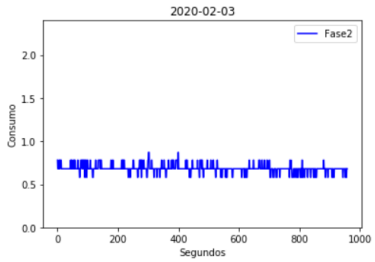

(c) Fase 2 - 29 Enero



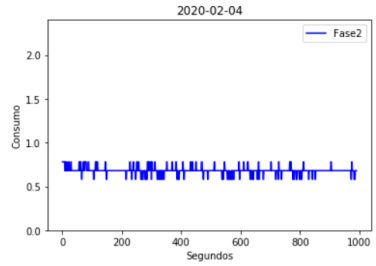

(a) Fase2 - 30 Enero



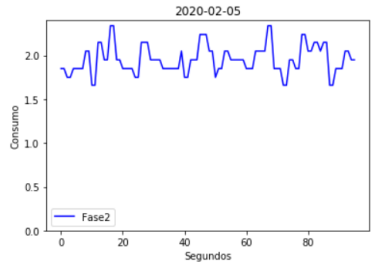

(b) Fase 2 - 31 Enero



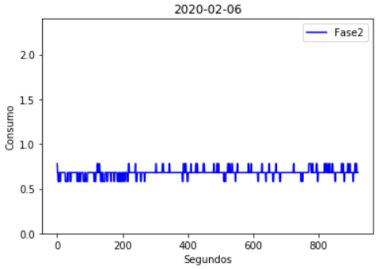

(c) Fase 2 - 3 Febrero



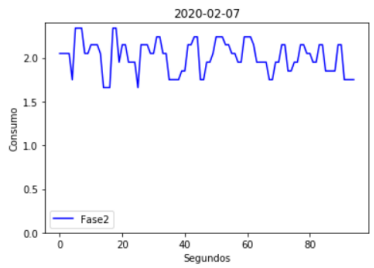

(a) Fase 2 - 4 Febrero



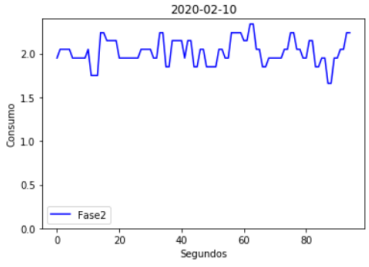

(b) Fase 2 - 5 Febrero



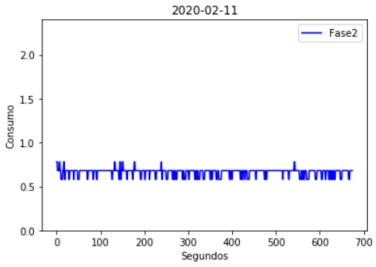

(c) Fase 2 - 6 Febrero



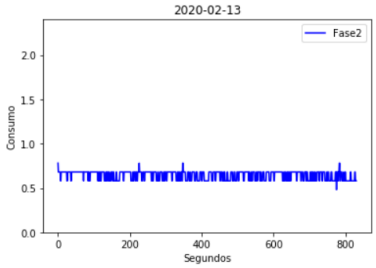

(a) Fase 2 - 7 Febrero



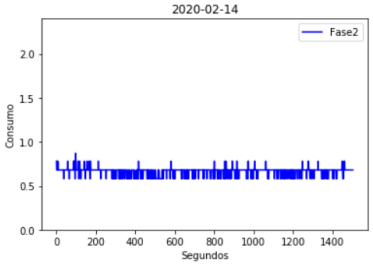

(b) Fase 2 - 10 Febrero




(c) Fase2 - 11 Febrero

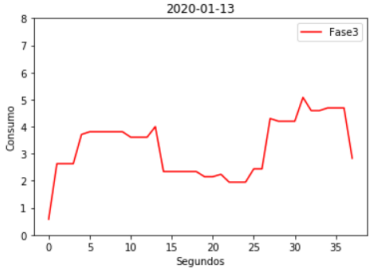



(a) Fase 2 - 13 Febrero

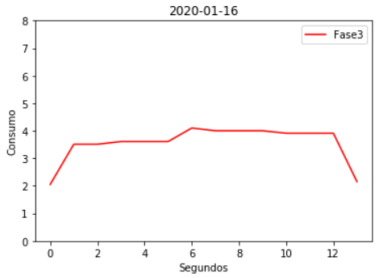



(b) Fase 2 - 14 Febrero

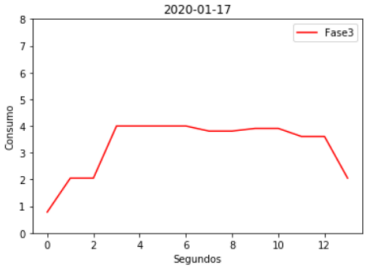
### A.3.4 Arranques de 2020 en Fase 3



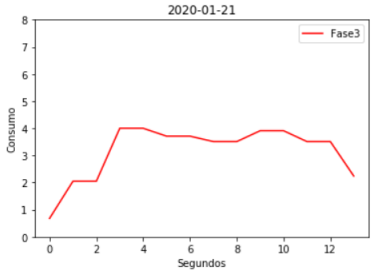

(a) Fase 3- 13 Enero



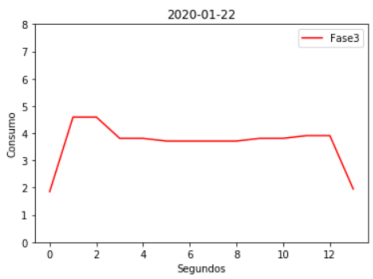

(b) Fase3 - 16 Enero



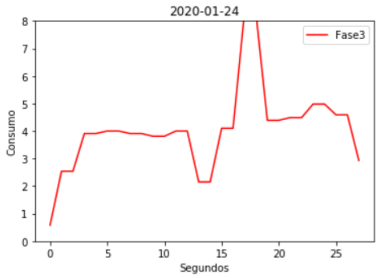

(c) Fase 3 - 17 Enero



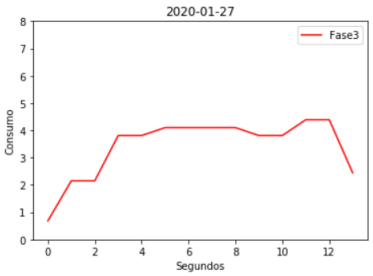

(a) Fase 3 - 21 Enero



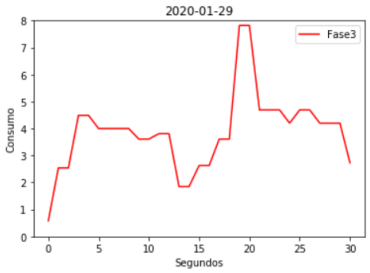

(b) Fase 3 - 22 Enero



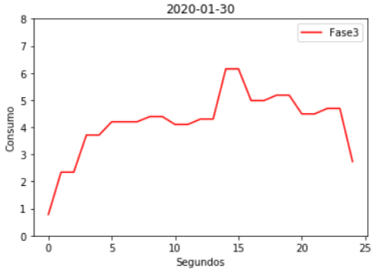

(c) Fase 3 - 24 Enero



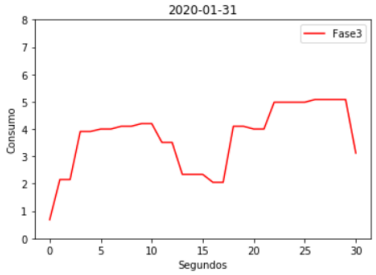

(a) Fase 3 - 27 Enero



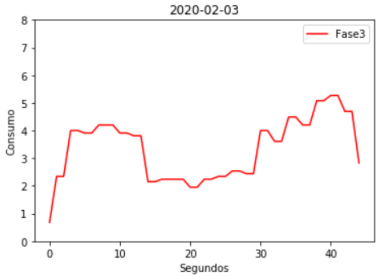

(b) Fase 3 - 29 Enero



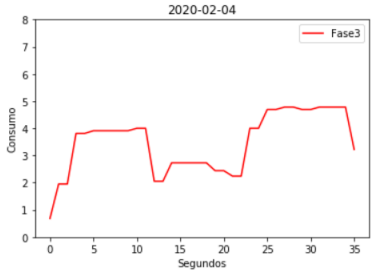

(c) Fase 3 - 30 Enero



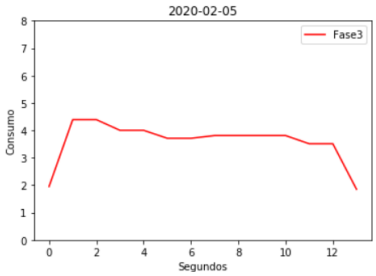

(a) Fase 3 - 31 Enero



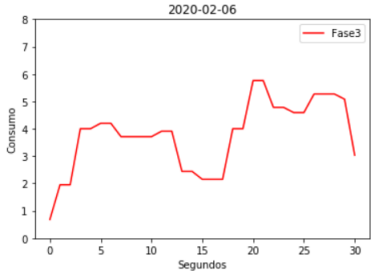

(b) Fase 3 - 3 Febrero



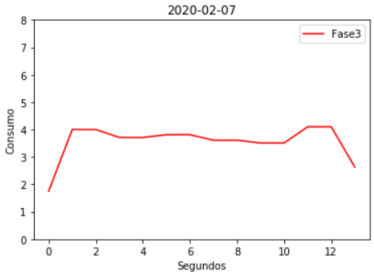

(c) Fase 3 - 4 Febrero



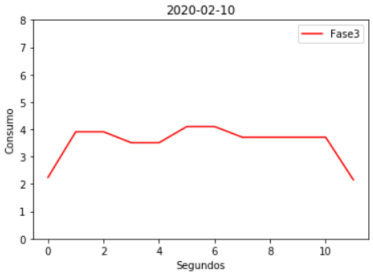

(a) Fase 3 - 5 Febrero



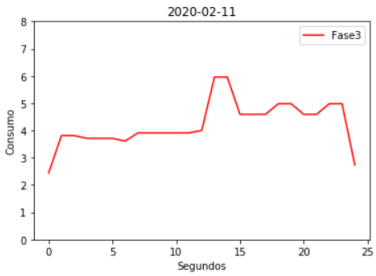

(b) Fase 3 - 6 Febrero



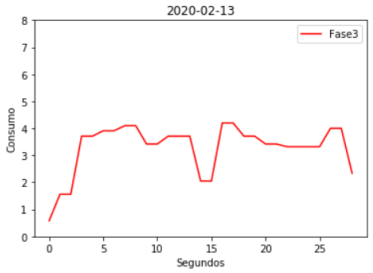

(c) Fase 3 - 7 Febrero



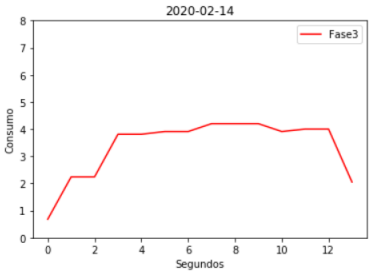

(a) Fase3 - 10 Febrero




(b) Fase 3 - 11 Febrero

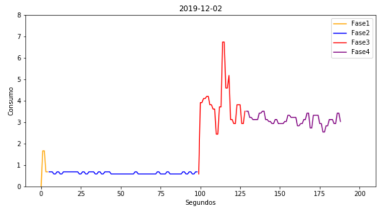



(a) Fase 3 - 13 Febrero

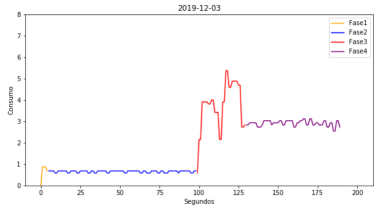



(b) Fase 3 - 14 Febrero

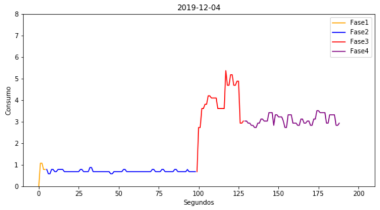
### A.3.5 Arranques reales interpolados - datos de test



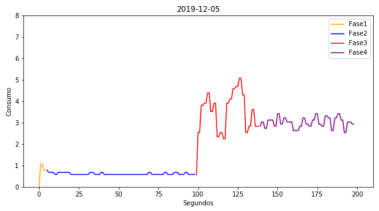

(a) Arranque 2 Diciembre



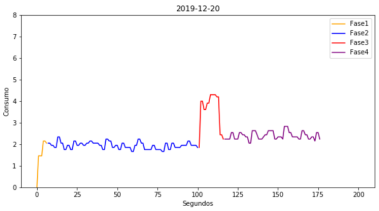

(b) Arranque 3 Diciembre



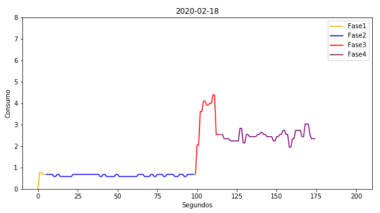

(c) Arranque 4 Diciembre



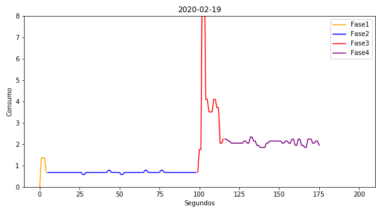

(a) Arranque 5 Diciembre



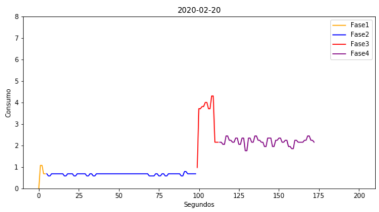

(b) Arranque 20 Diciembre



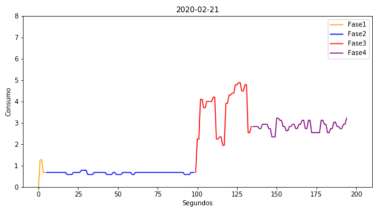

(c) Arranque 18 Febrero



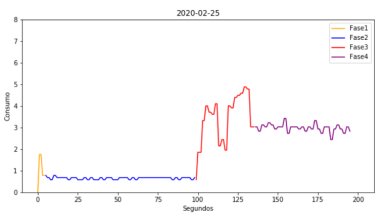

(a) Arranque 19 Febrero



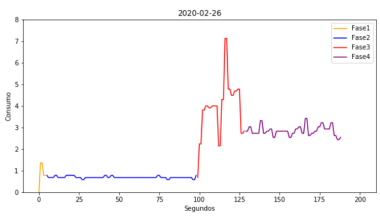

(b) Arranque 20 Febrero



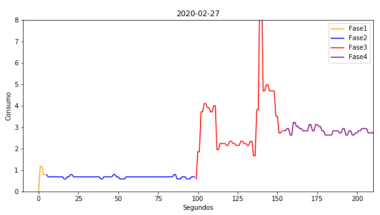

(c) Arranque 21 Febrero



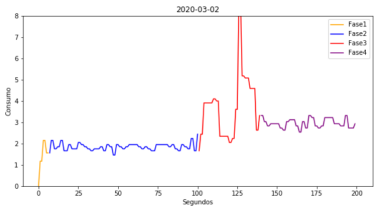

(a) Arranque 25 Febrero



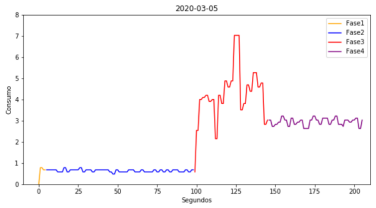

(b) Arranque 26 Febrero



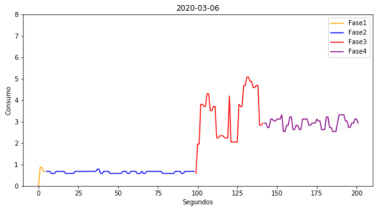

(c) Arranque 27 Febrero



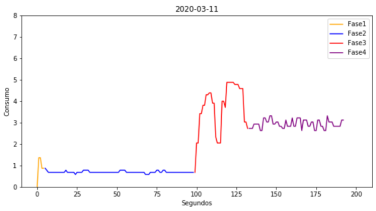

(a) Arranque 2 Marzo



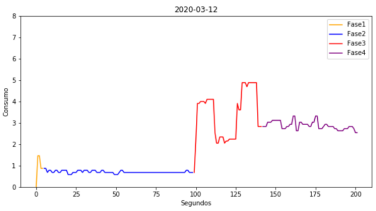

(b) Arranque 5 Marzo



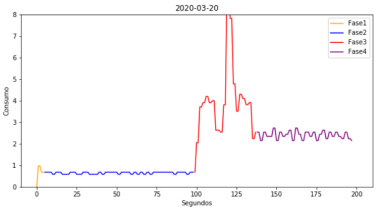

(c) Arranque 6 Marzo



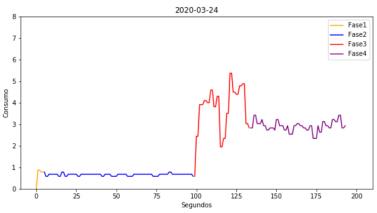

(a) Arranque 11 Marzo



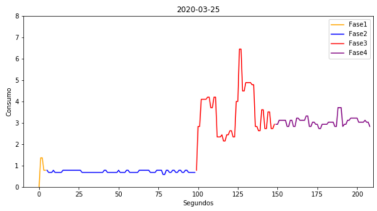

(b) Arranque 12 Marzo




(c) Arranque 20 Marzo



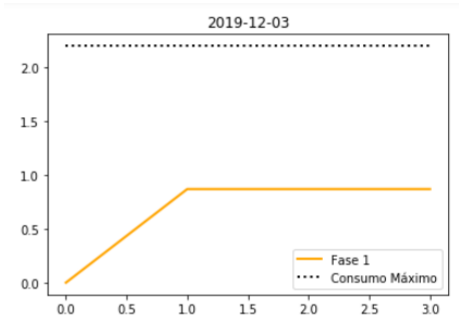

(a) Arranque 24 Marzo



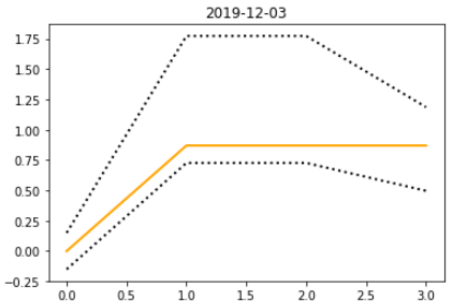

(b) Arranque 25 Marzo

### A.3.6 Simulación paso a paso de arranque 3 de Diciembre de 2019

#### Simulación Fase 1




(a) Fase 1 Consumo Máximo Permitido




(b) Comportamiento segundo a segundo Fase 1

Figura A.47: Comportamiento Fase 1

#### Simulación Fase 2

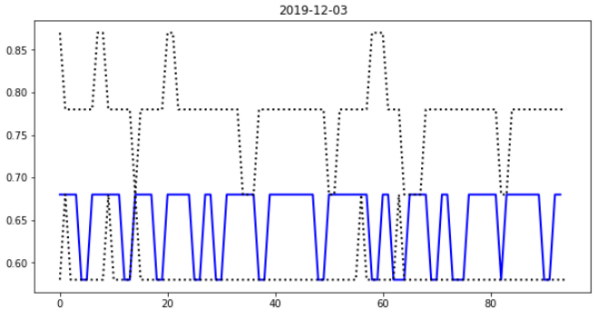
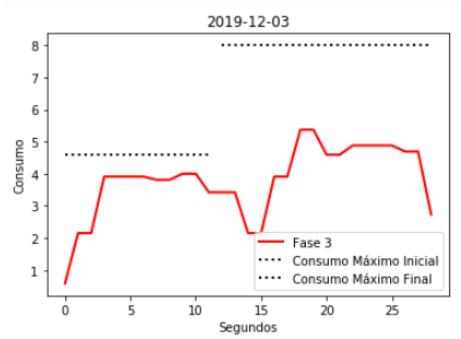
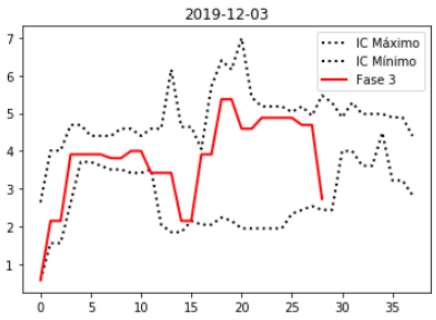



Figura A.48: Comportamiento Fase 2

#### Simulación Fase 3




(a) Fase 3 Consumo Máximo Permitido y Escalón/Doble Escalón Posible




(b) Comportamiento segundo a segundo Fase 3

Figura A.49: Comportamiento Fase 3

### Simulación Fase 4

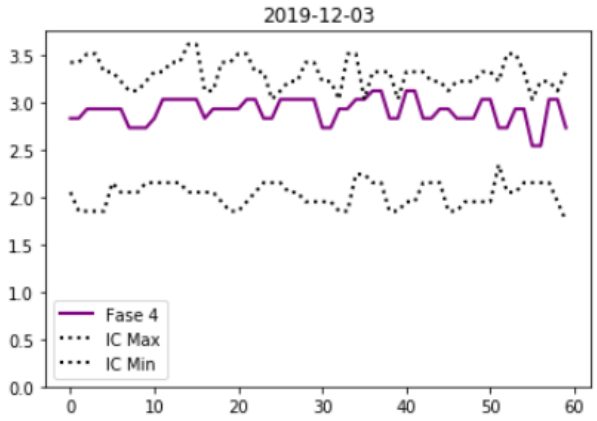



Figura A.50: Comportamiento Fase 4

### Simulación Arranque 3 Diciembre Global

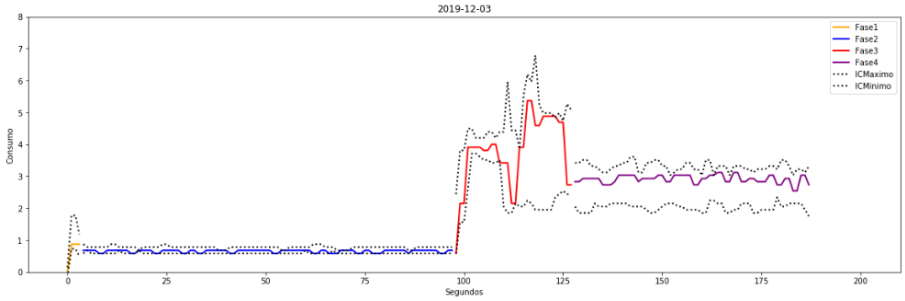



Figura A.51: Comportamiento Arranque global 3 Diciembre

### Resumen Características Finales

| | |
|---|---|
| Fecha. | 2019-12-03 |
| Escalon Fase 1 | True |
| Consumo Fase 1 | True |
| Doble Escalon Fase 1 | False |
| Fase 1 Grupo | 100 |
| Fase 2 Grupo | 98.9362 |
| Escalon Fase 3 | True |
| Doble Escalon Fase 3 | True |
| Consumo Fase 3 Inicial | True |
| Consumo Fase 3 Final | True |
| Fase 3 Grupo | 96.5517 |
| Duracion Maxima Fase 3 | True |
| Anomalias Leves | 0 |
| Fase 4 Grupo | 100 |
| Anomalo | False |

Figura A.52: Características Arranque del 3 de Diciembre de 2019