UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MECÁNICA

EVALUACIÓN DE METODOLOGÍAS DE IMPUTACIÓN DE DATOS EN MOTORES DIÉSEL PARA EL DESARROLLO DE SISTEMAS DE DIAGNÓSTICO INTELIGENTE DE FALLAS

TESIS PARA OPTAR AL TÍTULO DE INGENIERA CIVIL MECÁNICA

TATIANA ISIDORA REYES OSORIO

PROFESORA GUÍA:
Viviana Meruane Naranjo

MIEMBROS DE LA COMISIÓN:
Enrique López Droguett
Jorge Marín Castillo

SANTIAGO DE CHILE
2023

EVALUACIÓN DE METODOLOGÍAS DE IMPUTACIÓN DE DATOS EN MOTORES DIÉSEL PARA EL DESARROLLO DE SISTEMAS DE DIAGNÓSTICO INTELIGENTE DE FALLAS

El diagnóstico inteligente de fallas en los equipos se puede realizar a partir de la identificación de patrones anómalos, mediante algoritmos de detección de anomalías o novedades, debido a que una anomalía puede ser un precursor de una falla en el equipo.

El proceso previo de análisis de los datos es fundamental para realizar un diagnóstico acabado y certero, en la actualidad, aquel proceso debe enfrentar el desafío de los datos faltantes o los mejores conocidos como "NaNs". Esta pérdida de información usualmente lleva a eliminar toda la observación cuando no se posee algún dato, siendo una motivación adentrarse en metodologías de manejo datos faltantes. La imputación de datos es una solución frente al desafío mencionado, ya que es un procedimiento de reemplazo de valores faltantes por un conjunto de datos. Existen métodos estadísticos para efectuar el reemplazo, entre ellos está la imputación media o la imputación con *FillFoward* y también existen procedimientos de imputación a través de modelos basados en aprendizaje de máquinas. Algunos ejemplos de estos modelos son la imputación múltiple en cadena, que es un algoritmo iterativo de regresión; la metodología de imputación con *K-Nearest Neighbor*, la cual es un método de clasificación supervisado basándose en distancia; u otro ejemplo es realizar el procedimiento de imputación de datos con árboles de regresión ya sea uno o implementando varios a la vez en los algoritmos *Decision Tree*, *Random Forest* o *Extremelly Randomized Tree*.

El objetivo general del Trabajo de Título es comparar metodologías de imputación de datos faltantes en una base de datos de dos motores diésel marinos y evaluar el impacto del proceso de imputación en el diagnóstico inteligente de fallas. Para esto se consideran los siguientes objetivos específicos: (i) analizar la base de datos disponible y efectuar un preprocesamiento de los datos (ii) comparar y seleccionar la metodología de imputación de datos faltantes que se ajuste de mejor manera a la base de datos, (iii) evaluar el impacto del proceso de imputación de datos faltantes en el diagnóstico inteligente de fallas.

A partir de los resultados se concluye que el modelo de imputación de datos faltantes con mejor resultado es *K-Nearest Neighbor* con el algoritmo de pesos uniformes, debido a que la metodología imputa valores con propiedades físicas dentro del rango permitido u observado, considerando la evolución de las mediciones. Por otro lado, tras comparar los procedimientos con imputación de datos faltantes y sin imputación de datos faltantes, se evidencia que la imputación de datos es una herramienta poderosa que permite estudiar todos los sistemas de los motores, sin embargo, debe ser estudiada en profundidad, ya que escoger una metodología errónea puede sesgar el análisis y conclusiones del estudio.

El informe consta de cinco capítulos: (i) Introducción, (ii) Antecedentes, (iii) Metodología, (iv) Resultados y Análisis, y (v) Conclusiones.

A mi familia, pareja y amigos,
que siempre confiaron en mí.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Motivación

En la actualidad la gran mayoría de los modelos de diagnóstico inteligente de fallas disponibles en la literatura han sido creados y validados con datos obtenidos en ambientes académicos, sin ser transferidos o validados en la industria, por lo que trabajar con bases de datos industriales sin preprocesamiento previo otorga valor a las investigaciones, además de ser una aplicación directa en el trabajo del día a día de la industria.

A su vez, es necesario enfrentar el desafío de los datos faltantes o los mejores conocidos como "NaNs", esta falta se genera en el proceso de adquisición de datos ya sea por desperfectos en los dispositivos o interrupciones en la transmisión de sensores, no se capturan todos los datos. Esta pérdida de información usualmente lleva a eliminar toda la observación cuando no se tiene un dato, siendo una motivación adentrarse en metodologías de manejo datos faltantes, basándose que cada celda de una base de datos no es un dato independiente, sino que depende de la misma base de datos.

1.2. Antecedentes Generales

La imputación de datos es una solución frente al desafío expuesto de los datos faltantes, ya que es un procedimiento de reemplazo de valores faltantes por un conjunto de datos. Existen métodos estadísticos para efectuar el reemplazo, como la imputación media, por razón, imputación *Hot-deck* o la imputación *FillFoward*, y modelos basados en aprendizaje de máquinas, es decir, donde las máquinas realizan las predicciones a partir de ejemplos y/o experiencias, algunos ejemplos de estos modelos son la imputación múltiple de ecuaciones encadenadas, metodologías de árboles de decisión, *Decision Tree*, *Random Forest*, *Extremely Randomized Trees* o *K-Nearest Neighbor*.

Dentro de las herramientas existentes en el diagnóstico inteligente de fallas, se encuentra la detección de novedades, una novedad o anomalía es un cambio o desviación frente al patrón esperado. En aprendizaje de máquinas existen algoritmos que buscan detectar novedades en los conjuntos de datos, con el fin de notificar comportamientos anómalos, ya que estos pueden ser precursores de posibles fallas en el equipo. Para el caso de detección de novedades

se entrena el modelo con datos normales y se busca detectar valores anómalos en nuevas observaciones, a diferencia que la detección de anomalías que se entrena el modelo con un porcentaje de fallas.

Elliptic Envelope, es uno de los métodos de detección de novedades, basado en que los datos normales siguen una distribución conocida y se considera a los valores anómalos los que se encuentran lo suficientemente alejados de esta distribución.

Para realizar el trabajo se cuenta una base de datos de dos motores marinos diésel, los cuales poseen 175.604 observaciones en 62 parámetros cada uno, en donde principalmente se tienen mediciones de presión, temperatura y velocidad. En esta base de datos se tiene un porcentaje de datos faltantes y se sabe que existieron fallas en los motores en el periodo de adquisición de datos, sin embargo, no se sabe en qué periodo ocurrieron estas fallas ni su cantidad.

1.3. Objetivos

El objetivo general del Trabajo de Título es comparar metodologías de imputación de datos faltantes en una base de datos de dos motores diésel marinos y evaluar el impacto del proceso de imputación en el diagnóstico inteligente de fallas.

Para esto se consideran los siguientes objetivos específicos:

1. Analizar la base de datos disponible y efectuar un preprocesamiento de los datos.
2. Comparar y seleccionar la metodología de imputación de datos faltantes que se ajuste de mejor manera a la base de datos.
3. Evaluar el impacto del proceso de imputación de datos faltantes en el diagnóstico inteligente de fallas.

1.4. Alcances

Para precisar los objetivos se tienen los siguientes alcances:

- El estudio considera parámetros que cuenten con menos de un 80 % de datos faltantes y registros que cuenten con al menos un dato en la observación.
- Se compara las siguientes metodologías de imputación de datos: Imputación Media, Imputación *FillFoward*, Imputación Múltiple en Cadena *(MICE)*, Imputación con *Decision Tree*, Imputación con *Random Forest*, Imputación con *Extremely Randomized Trees* e Imputación *K-Nearest Neighbor*.
- Se utiliza el método *Elliptic Envelope* para la detección de novedades.
- Se analiza de manera gráfica y cuantitativa el diagnóstico inteligente de fallas.

1.5. Estructura del documento

El informe de este trabajo consta con cinco capítulos. El capítulo 1 cuenta con una introducción al trabajo señalando la motivación y los antecedentes generales para comprender los objetivos y alcances del trabajo. En el capítulo 2 se presenta el estudio del arte de las metodologías de imputación de datos faltantes, además de las herramientas implementadas en el diagnóstico inteligente de fallas, detallando a su vez, la base de datos de motores diésel marinos. El capítulo 3 explica la metodología implementada en el trabajo, definiendo tres etapas: estudio previo, implementación de metodologías de imputación y diagnóstico inteligente de fallas. En el capítulo 4 se muestran los resultados obtenidos en cada etapa del trabajo realizado, junto con el análisis de estos mismos. Finalmente, en el capítulo 5 se exponen las conclusiones del trabajo de investigación y posibles trabajos futuros. En cuanto a la bibliografía y los anexos, estos se presentan al final del documento.

Capítulo 2

Antecedentes

2.1. Aprendizaje de Máquinas

El aprendizaje de máquinas, aprendizaje automático o *machine learning* en inglés, es una disciplina del campo de la inteligencia artificial, cuya finalidad es desarrollar técnicas que permitan a las computadoras adquirir la capacidad de identificar patrones en datos masivos para elaborar predicciones a partir de ejemplos y/o experiencias. De este modo, el aprendizaje de máquinas es un método de análisis de datos que automatiza la construcción de modelos analíticos.

Existe un sinnúmero de aplicaciones del aprendizaje de máquinas en áreas de mantenimiento, salud, finanzas, seguridad, transporte, ciencia, entre otros. En la actualidad el volumen de los datos ha aumentado, lo que gatilla un interés mayor por el *machine learning* para el manejo de datos y toma de decisiones.

Los algoritmos de aprendizaje automático se dividen en cuatro tipos de aprendizajes:

- **Aprendizaje supervisado:** Algoritmos que son entrenados con ejemplos "etiquetados", lo que significa que se sabe el resultado deseado de las variables de entrada. De este modo, el algoritmo debe buscar un modelo que, dadas las variables de entrada, entregue los valores etiquetados, frecuentemente se utilizan métodos de clasificación o regresión. Luego de identificar los patrones y definir el modelo, el algoritmo debe predecir las etiquetas adecuadas para nuevos valores.

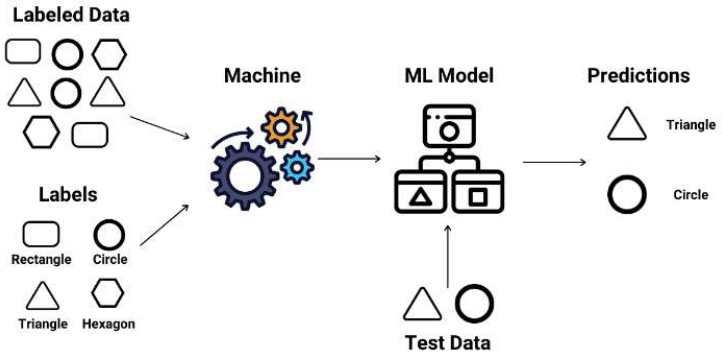


Figura 2.1: Diagrama de aprendizaje supervisado [1]

- **Aprendizaje no supervisado:** Algoritmos que no disponen los datos de entrenamiento etiquetados previamente, es decir, no se sabe la respuesta "correcta" del sistema. Por lo tanto, el algoritmo explora los datos con el fin de identificar patrones o alguna estructura, generalmente se utilizan métodos de reducción de dimensión y métodos de agrupamiento.

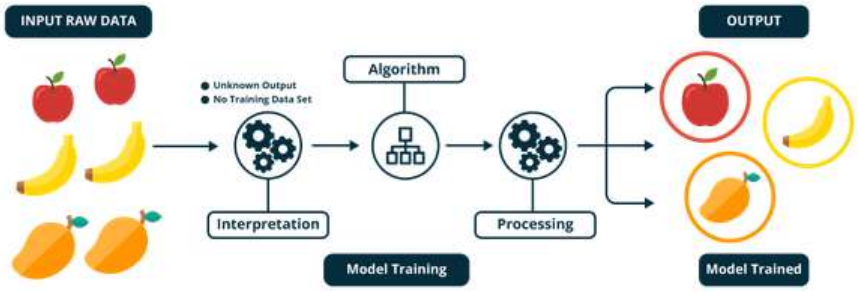


Figura 2.2: Diagrama de aprendizaje no supervisado [2]

- **Aprendizaje semi supervisado:** Algoritmos que utilizan datos etiquetados y no etiquetados en el entrenamiento. De este modo, los datos etiquetados se utilizan para definir las clases del modelo, mientras que los no etiquetados se implementan para refinar los bordes de estas clases. Por lo general la cantidad de datos etiquetados es menor que la cantidad de no etiquetados, debido a los costos asociados.

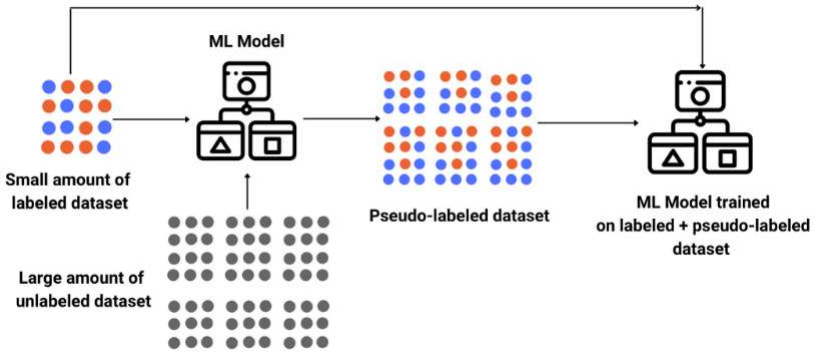


Figura 2.3: Diagrama de aprendizaje semi supervisado [3]

- **Aprendizaje por refuerzo:** Algoritmos que aprenden a partir de su propia experiencia, es decir, los algoritmos descubren a través de prueba y error qué acciones producen una mayor recompensa. Existen tres componentes principales en este tipo de aprendizaje: (i) el agente, que es quien toma las decisiones y aprende, (ii) el medio, que corresponde con todo lo que interactúa el agente, y (iii) las acciones, que son lo que el agente puede hacer. De esta forma, el algoritmo es el agente que tiene de entrada la retroalimentación que obtiene del medio como respuesta a sus acciones.

Figura 2.4: Diagrama de aprendizaje por refuerzo [4]

2.2. Imputación de datos

Un problema en la calidad de los datos recopilados, para distintas aplicaciones, son los valores perdidos, estos hacen referencia al conjunto en el cual se desconoce el valor de la variable medida o los usualmente conocidos como "NaN", cuyas siglas significan "Not a Number". Factores causantes de esta falta de datos son desperfectos en dispositivos de medición, interrupción en la transmisión del sensor o datos ingresados de forma manual.

La aleatoriedad de los datos faltantes se puede dividir en tres clases según lo propuesto por Little y Rubin en 1987 [5]:

- MCAR *(Missing completely at Random)*: Corresponde cuando la probabilidad de que un registro sea un valor faltante es completamente aleatorio y no depende de los datos observados en las variables auxiliares ni del valor real del valor faltante. Generalmente en los algoritmos de imputación se asume que los valores perdidos siguen un patrón MCAR.
- MAR *(Missing at Random)*: Es cuando la probabilidad de que un registro falte podría depender de los datos observados en las variables auxiliares, pero no del valor estudiado sin información en sí.
- NMAR *(No missing at Random)*: Ocurre cuando la probabilidad de que un registro este sin información podría depender del valor del atributo desconocido.

Una técnica utilizada frecuentemente es la eliminación de los registros con datos faltantes, la cual es práctica cuando el número de registros incompletos es pequeño en comparación con el total de registros, sin embargo, se generan sesgos en los coeficientes de asociación y correlación [6], por lo que no es una técnica recomendada.

La imputación de datos faltantes es otra técnica implementada cuando se presentan valores perdidos en la data y es una solución frente a la problemática del manejo de falta de datos. En estadística, la imputación es un término denominado para el procedimiento de reemplazo o llenado de valores por un conjunto de datos, según la información procedente de la base de datos. Existen distintas metodologías en la imputación de datos, estas se clasifican en métodos estadísticos o modelos de aprendizaje de máquinas que se detallan en las

siguientes subsecciones. A su vez, las metodologías de imputación de datos se clasifican en imputación simple o imputación múltiple, en el caso de la imputación simple se asigna solo un valor al dato faltante, ya sea a partir de los valores de la variable en donde se encuentra el valor perdido o a partir de valores de variables auxiliares, generando así solo una base de datos completa, sin embargo, la imputación simple puede resultar una estimación sesgada en algunas ocasiones, surgiendo así, la imputación múltiple, que asigna un conjunto de valores al valor faltante, generando conjuntos de bases de datos completas.

Cabe destacar, que está ampliamente documentado que la aplicación de procedimientos inapropiados de imputación de datos faltantes puede introducir sesgos en el análisis llegando inclusive a invalidar las conclusiones del estudio. Es por esto, que es fundamental definir el modelo apropiado de imputación, no obstante, hasta el momento no existe una guía del método idóneo a escoger, sino que se estudia caso a caso.

2.2.1. Métodos estadísticos

Imputación media: Es uno de los métodos más antiguos y sencillo, presentado por primera vez por Wilks en el año 1932. Asume que los datos siguen un patrón MCAR y el valor faltante se reemplazan por la media de los datos observados en esa variable. En el caso de las variables categóricas se utiliza la moda.

La imputación por media tiene las siguientes variantes [7]:

- Medias no condicionadas: Se reemplaza por la media de la variable estudiada considerando la información de todos los datos obtenidos para esa variable, su aplicación subestima la varianza de la variable imputada, además de atenuar la correlación con el resto de las variables.
- Media condicional: Consiste en formar categorías a partir de covariables correlacionadas con la variable estudiada e imputa la media de los valores observados en la variable de los mismos grupos.

Imputación por razón: Considera solo una variable auxiliar y asume que esta es proporcional a la variable estudiada. El procedimiento calcula la razón entre estas variables para el conjunto de datos con información y se imputa el valor de la multiplicación de la razón con el valor de la variable auxiliar, tal como se indica en la siguiente ecuación:

$$\tilde{y}_i = \hat{R} \cdot x_i = \frac{\sum_k y_k}{\sum_k x_k} \cdot x_i \tag{2.1}$$

Donde, y es la variable estudiada, x la variable auxiliar, k es el conjunto de observaciones que las dos variables poseen datos, $\tilde{y}_i$ es el valor imputado en la observación i y x_i es el valor de la variable auxiliar en la observación i. A medida que exista mayor linealidad entre las variables la estimación imputada será mejor. Los inconvenientes son que se necesita una buena calidad de información para la variable auxiliar y una alta correlación entre las variables.

Imputación *hot deck*: Procedimiento donde el dato sin información se reemplaza por el valor de otro registro similar, que posee el valor de la variable requerida. Usualmente se implementa en dos etapas: en primer lugar, la clasificación, en la cual los datos se dividen en grupos homogéneos y disjuntos; y en la segunda etapa, se utilizan los registros completos del grupo para reemplazar los casos de datos faltantes del mismo grupo. El procedimiento asume que dentro de cada grupo los valores sin información siguen la misma distribución que los valores con información y que los datos fueron perdidos de forma MCAR.

Dentro de las ventajas de este procedimiento de imputación destaca la conservación de la distribución de la variable incompleta. Por contrario, las desventajas de este procedimiento son (i) la distorsión de las relaciones entre variables, (ii) las clases deben ser definidas en un número reducido de variables, con finalidad de contar con suficientes registros completos en todas las clases, (iii) existen sesgos cuando se reemplaza por el mismo registro varias veces.

En la literatura predominan tres métodos para elegir el valor a reemplazar:

- Aleatorio: Selecciona el registro completo a través de un muestreo aleatorio simple perteneciente al grupo, puede elegir uno o varios registros, en el caso que sean más de uno se imputa el promedio de estos valores. Al escoger de forma aleatoria no introduce sesgos en la varianza del estimador.

- Secuencial: Para aplicar esta imputación se debe clasificar previamente los grupos, de tal forma que se produzca una autocorrelación positiva entre los campos de imputación, con el fin de otorgar mayor similitud entre los registros. Luego el registro con el valor sin información se reemplaza con el registro anterior a este, sin embargo, en el caso que el valor inicial sea el registro que posee el valor faltante se reemplaza por un registro aleatorio del grupo, que contenga el valor buscado. La desventaja de este método en particular se genera cuando existe un grupo elevado de datos faltantes ubicados de forma continua, remplazando con el mismo valor en reiteradas ocasiones.

- Vecino más cercano: Consiste en escoger el valor de imputación a partir del valor que tenga el vecino más cercano del registro, para esto se debe definir una noción de distancia para identificar los vecinos. En el caso de variables cuantitativas y categóricas en la investigación de Jerez, Molina, García-Laencina, Alba, Ribelles, Martín, & Franco [8] se implementó una función de distancia heterogénea, que utiliza una métrica de superposición para los atributos categóricos que corresponde a una distancia de Hamming y una función de distancia de Manhattan normalizada para atributos cuantitativos.

  De este modo, al considerar un registro x con n parámetros, $x = [x_1, x_2, ..., x_n]^T$. La distancia heterogénea entre los registros x_a y x_b es:

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{n} d_j(x_{aj}, x_{bj})^2} \tag{2.2}$$

  Donde $d_j(x_{aj}, x_{bj})$ es la distancia entre x_a y x_b en el atributo sub j.

$$d_j(x_{aj}, x_{bj}) = \begin{cases} 1 & \text{si } (1 - m_{aj})(1 - m_{bj}) = 0 \\ d_o(x_{aj}, x_{bj}) & \text{si } x_j \text{ es un atributo cualitativo} \\ d_N(x_{aj}, x_{bj}) & \text{si } x_j \text{ es un atributo cuantitativo} \end{cases} \tag{2.3}$$

Con m un vector binario tal que $m_j = 1$ si x_j es un valor faltante y $m_j = 0$ si el valor se conoce; d_o asigna un valor 0 si los atributos son los mismos de lo contrario es 1; y d_N está dado por:

$$d_N(x_{aj}, x_{bj}) = \frac{|x_{aj} - x_{bj}|}{max(x_j) - min(x_j)} \tag{2.4}$$

En la Ecuación 2.4 la distancia de Manhattan está normalizada, lo cual se recomienda para las variables cuantitativas, esto con el fin de reducir la influencia de las variables que tengan órdenes de magnitud mayores.

Otras posibles medidas de distancia son:

- Distancia Euclídea: $d(x_a, x_b) = \sqrt{\sum_j |x_{aj} - x_{bj}|^2}$
- Distancia de Minkowski p: $d(x_a, x_b) = (\sum_j |x_{aj} - x_{bj}|^p)^{\frac{1}{p}}$, $p \geq 1$

Imputación *Fill Foward*: Método de imputación que completa el valor faltante con el registro anterior de la variable estudiada. Posee la misma desventaja que el procedimiento de imputación *hot deck* secuencial.

2.2.2. Modelos de aprendizaje de máquinas

Multivariate Imputation By Chained Equations (MICE): Es un algoritmo iterativo de imputación múltiple, es decir, realiza el procedimiento por turnos y cada valor faltante se predice varias veces. La metodología utiliza ecuaciones encadenadas y asume que los valores perdidos fueron del tipo MCAR. Específicamente, el procedimiento imputa los valores perdidos iterativamente ejecutando una serie de procedimientos de regresión, recorriendo todas las variables objetivas para imputar los valores perdidos, en el caso de variables binarias normalmente se utiliza la regresión logística y para variables continuas la regresión lineal.

Los pasos a seguir en una iteración se indican en la Figura 2.5. El primer paso es reemplazar los valores perdidos por la media de cada variable o por valores al azar, luego se analiza cada variable por separado eliminando el valor imputado en esa variable (paso 2), en seguida se considera el llenado de los valores perdidos de las otras variables y los valores de la base de datos como una nueva matriz de datos sin considerar la fila del valor faltante (paso 3), con el fin de tomar esta matriz de variable de entrada al modelo de regresión seleccionado para completar el dato faltante (paso 4), posteriormente se repite el procedimiento a cada valor perdido (pasos 5 al 8) obteniendo así la primera iteración. Para el caso de las siguientes iteraciones la base de datos de partida es la última base de datos ajustada.

El número de iteraciones se determina en función de la tasa de error y el criterio de convergencia definido, en donde la tasa de error es la diferencia entre la primera y última base de datos ajustada en cada iteración, en el caso de la Figura 2.5 es entre la base de datos que fue imputada con los valores medios y la base de datos ajustada al terminar el paso 8.

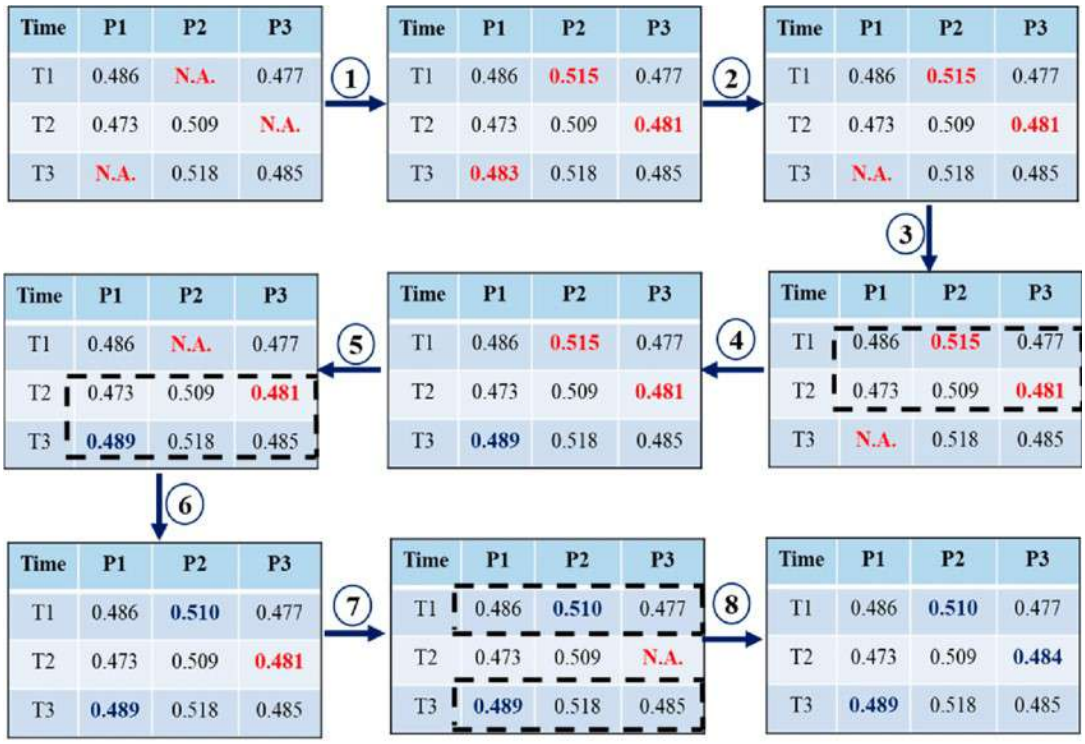


Figura 2.5: Esquema de imputación de datos múltiple por ecuaciones encadenadas [9]

A su vez, en *MICE* se pueden implementar modelos de regresión de árboles de decisión los cuales se explican a continuación.

***Árboles de decisión*:** Los árboles de decisión son modelos predictivos, cuya función de aprendizaje toma la forma de un árbol siendo similar a un diagrama de flujo. En esta estructura los nodos internos representan atributos, las ramas reglas de decisión y cada nodo "hoja" representa un resultado. De esta forma, el objetivo es encontrar los atributos que separen de mejor forma los datos y que en cada nodo sean lo más homogéneos entre ellos, en el caso de los parámetros continuos, también se debe definir el umbral que fragmenta las ramas. Los árboles de decisión se utilizan para problemas de regresión o clasificación. En el caso de imputación de valores perdidos continuos son problemas de regresión.

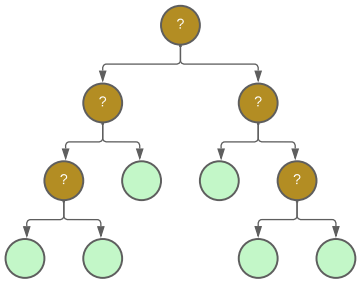


Figura 2.6: Diagrama árbol de decisión (EP)

Los criterios de selección de atributos más populares se detallan a continuación:

- Índice de Gini: Medida que considera el grado de pureza del conjunto. A mayor índice de Gini menor pureza, por lo que se selecciona el atributo con menor índice de Gini ponderado. El índice se define como:

$$G(A_i) = \sum_{j=1}^{M_i} p(A_{ij}) \cdot G(C \mid A_{ij}) \tag{2.5}$$

  Donde A_i es el atributo, C la clase, M_i es el número de valores diferentes del atributo y $p(A_{ij})$ es la probabilidad que A_i tome su j-ésimo valor ($1 \leq j \leq M_j$). A su vez se tiene que:

$$G(C \mid A_{ij}) = 1 - \sum_{k=1}^{J} p^2(C_k \mid A_{ij}) \tag{2.6}$$

- Ganancia de información: Medida basada en la disminución de entropía después que un conjunto de datos se divida por una regla de decisión, buscando la mayor ganancia de información tras seleccionar los atributos.

$$\text{Ganancia de informacion } (T, X) = \textit{Entropia } (T) - \textit{Entropia } (T, X) \tag{2.7}$$

  La entropía mide que tan homogéneo es un conjunto, si la entropía vale cero equivale que el conjunto es completamente homogéneo. En el caso que tenga un parámetro la entropía del atributo T se define como:

$$E(T) = \sum_{i=1}^{C} -p_i \, log_2 \, p_i \tag{2.8}$$

  Donde C corresponde al conjunto de diferentes clasificaciones y p_i es la proporción de casos que pertenecen a la clasificación i en los datos. Para dos parámetros la entropía se define de la siguiente manera:

$$E(T, X) = \sum_{s \in X} P(s) \, E(s) \tag{2.9}$$

  Donde $P(s)$ es la proporción de que ocurra la clasificación s del atributo X.

- Error cuadrático medio: Es un criterio de selección usado principalmente en variables continuas y busca escoger divisiones en donde los subconjuntos generados tengan la menor varianza. El algoritmo establece el valor predicho de los nodos terminales como el valor medio $\bar{y}_m$ del vector de etiqueta y se busca minimizar la función de pérdida $H()$ para escoger la división del nodo m.

$$\bar{y}_m = \frac{1}{n_m} \sum_{y \in Q_m} y \tag{2.10}$$

$$H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} (y - \bar{y}_m)^2 \tag{2.11}$$

Donde n_m son las muestras del nodo m y Q_m la data del nodo m [10].

Random Forest: Es otro modelo de aprendizaje de máquinas implementado en el procedimiento de imputación de datos faltantes. Este modelo es una combinación de árboles de decisión con el fin mejorar la precisión predictiva y controlar el ajuste excesivo. En cuanto al valor final de la predicción, se considera la media de todas las predicciones de cada árbol de decisión.

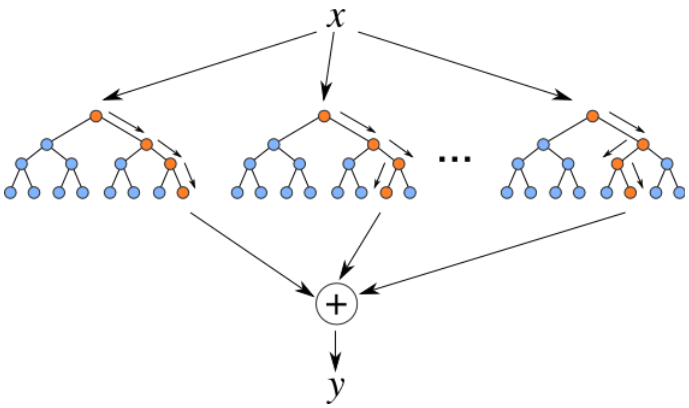


Figura 2.7: Ejemplo Random Forest [11]

La aleatoriedad de estos árboles se basa en dos factores, el primero es debido al muestreo con reemplazo o "bootstrap", lo cual significa que cada árbol se entrena con un submuestreo diferente en donde se pueden repetir los registros debido a que es con reemplazo, disminuyendo así la variabilidad de la predicción. En segundo lugar, se integra aleatoriedad limitando la cantidad de atributos que se pueden escoger en cada regla de decisión, la fórmula usual de esta cota es considerar la raíz del total de atributos, generando un subconjunto aleatorio en donde se opta por el mejor atributo según las medidas de selección anteriormente mencionadas. De esta forma, no serán todos los árboles similares si la dependencia con una variable es alta.

Extremely Randomized Tree: Es una variación del modelo *Random Forest* debido a que trabaja con la aleatoriedad de conjuntos de árboles de decisión. Sus dos diferencias principales con los modelos de conjuntos de árboles de decisión son la elección aleatoria de los puntos de corte en el procedimiento de división de nodos y la ausencia de "bootstrap", ya que cada árbol se entrena con el conjunto completo de los datos [12].

Específicamente en cada división se evalúa un subconjunto de atributos aleatorios limitado y dentro de cada atributo se selecciona un punto de corte completamente al azar, escogiendo el mejor para regla de decisión. Este factor aleatorio del punto de corte reduce la varianza a costa de un aumento del sesgo.

***K-Nearest Neighbors* (KNN):** Modelo de clasificación que utiliza casos similares para imputar valores. El criterio de similitud se determina por los k-vecinos más cercanos en los que no falta el valor en los atributos a imputar, por lo que se tiene que definir una métrica de distancia de las presentadas en el método *hot-deck* del vecino más cercano, realizando una previa normalización de las variables al igual que en ese método. En particular, el método *hot-deck* del vecino más cercano corresponde al modelo *KNN* con $k = 1$ cuando no se tienen etiquetas de las predicciones reales y se tiene de entrada el mismo grupo homogéneo.

K corresponde al hiperparámetro de cantidad de vecinos que considera el modelo, por lo que es un número entero positivo. Valores pequeños de K son sensibles al ruido en los datos, en cambio valores grandes reducen el efecto ruido, pero tienden a crear límites en las clases parecidas. El valor óptimo de K, generalmente se elige mediante validación cruzada o inspección en la data.

Considerando la notación presentada en la investigación de Jerez, Molina, García-Laencina, Alba, Ribelles, Martín, & Franco [8], $V = \{V_k\}^K{}_{k=1}$ representa el conjunto de valores de los k vecinos más cercanos de la variable X que posee datos incompletos, además este conjunto se ordena de forma creciente según su distancia. Luego, para estimar el valor perdido x_j en X existen diferentes procedimientos de estimación en el enfoque KNN:

- Estimación por media: Considera el promedio de los vecinos.

$$\tilde{x}_j = \frac{1}{K} \sum_{k=1}^{K} v_{kj} \tag{2.12}$$

- Estimación por media ponderada: Se pondera la contribución de cada valor por un peso w_k, el cual otorga mayor peso a los vecinos más cercanos.

$$\tilde{x}_j = \frac{1}{K} \sum_{k=1}^{K} w_k v_{kj} \tag{2.13}$$

  Una elección de w_k es el inverso del cuadrado de la distancia.

Para las variables categóricas se suele implementar la moda en vez de la media.

La desventaja principal del modelo es que cada vez que el método KNN busca las instancias más similares, el algoritmo tiene que buscar en todo el conjunto de datos. Esta limitación es especialmente problemática para bases de datos grandes, lo cual se podría reducir si se realiza una clasificación previa.

2.3. Diagnóstico inteligente de fallas

El mal funcionamiento de un equipo puede provocar grandes pérdidas económicas al frenar una línea de producción o inclusive puede significar un peligro para los trabajadores. Es por esto, que ha tomado fuerza el diagnóstico inteligente de fallas, el cual es la aplicación de métodos de aprendizaje de máquinas en la detección y diagnóstico de fallas en los equipos.

Una falla, es un cambio en el comportamiento de un componente en comparación a lo que fue diseñado. En los sistemas de detección se busca identificar el modo de operación del proceso con el estado de salud, para así en el diagnóstico, determinar el lugar y severidad de una falla cuando se detecte.

2.3.1. Detección de anomalías y novedades

Dentro de las herramientas existentes en el diagnóstico inteligente de fallas, se encuentran la detección de anomalías y la detección de novedades, una novedad o anomalía es un cambio o desviación frente al patrón esperado. En aprendizaje de máquinas existen algoritmos que buscan detectar estos valores atípicos en los conjuntos de datos, con el fin de notificar comportamientos anómalos, ya que pueden ser precursores de posibles fallas en el equipo.

Para el caso de detección de anomalías, también denominado detección de daño no supervisado, los datos de entrenamiento poseen valores anómalos sin etiquetas y se asume que los valores atípicos se encuentran en una región de baja densidad alejada de los valores normales. Por otro lado, la detección de novedades, también conocida como detección de daño semi supervisado, solo cuenta con datos normales en la etapa del entrenamiento y se busca detectar valores anómalos en nuevas observaciones, teniendo en cuenta que los valores anómalos pueden formar un grupo denso siempre y cuando se encuentren en una región de baja densidad en los datos de entrenamiento.

2.3.1.1. Elliptic Envelope

Elliptic Envelope, es un método de detección de novedades o anomalías, dependiendo de los datos entregados en el entrenamiento, que asume que los datos normales siguen una distribución conocida, considerando valores anómalos a los que se encuentran lo suficientemente alejados de esta distribución. La librería *Scikit-learn* de *Python* proporciona la función *covariance.EllipticEnvelope* que "ajusta una estimación de covarianza robusta a los datos y, por lo tanto, ajusta una elipse a los puntos de datos centrales, ignorando los puntos fuera del modo central" [13], tal como se ejemplifica en la Figura 2.8.

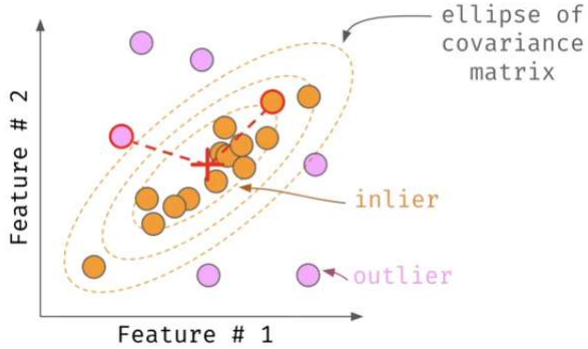


Figura 2.8: Ejemplo de aplicación de *Elliptic Envelope*

2.3.2. Reducción de parámetros

Otra herramienta del diagnóstico inteligente de fallas es la reducción de parámetros, implementada usualmente en el preprocesamiento de los datos. La reducción de parámetros disminuye la dimensión del conjunto de datos conservando las variables más importantes y eliminando los datos redundantes, es decir, aumenta la interpretación, pero al mismo tiempo minimizando la pérdida de información. Además, la reducción de parámetros ayuda a la visualización de los datos en dos o tres dimensiones, disminuyendo a su vez la velocidad de trabajo.

2.3.2.1. Principal Component Analysis

Principal Component Analysis (PCA) es una técnica utilizada para reducir parámetros de un conjunto de datos, no supervisada. El procedimiento consiste en crear nuevas variables correlacionadas que maximicen la varianza en un problema de valores y vectores propios, estas nuevas variables se denominan componentes principales.

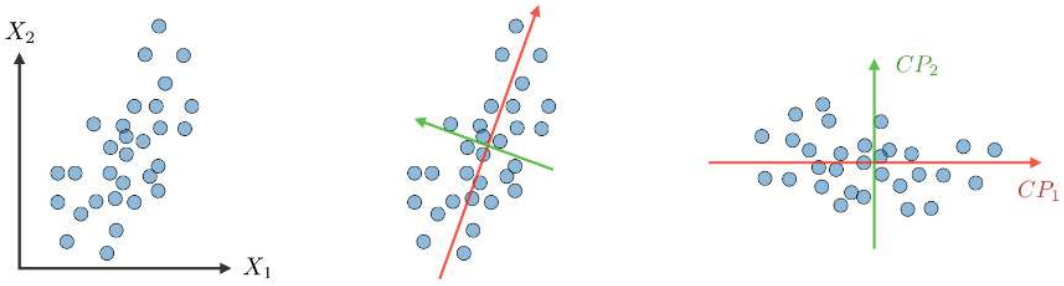


Figura 2.9: Ejemplo de generación de nuevas variables [14]

Es importante destacar, que es necesario centrar los datos al implementar este algoritmo.

2.4. Motores diésel

El principio fundamental de los motores diésel es convertir energía química a energía mecánica, específicamente, la energía se convierte a partir de la explosión del contacto del combustible con el aire a alta temperatura y presión.

El ciclo teórico de calor del motor diésel se presenta en la Figura 2.10. Desde el punto C al D, el aire se comprime adiabáticamente. La inyección de combustible comienza en el punto D, se agrega calor al ciclo a volumen constante de D hasta P y también se agrega calor a presión constante desde P a E. En el punto E comienza la expansión adiabática hasta el punto F y el calor es expulsado a volumen constante de F a C. La eficiencia ideal del ciclo varía aproximadamente entre 55 a 60 %, es decir, aproximadamente de un 40 a 45 % del calor suministrado se pierde por el sistema de escape. [15]

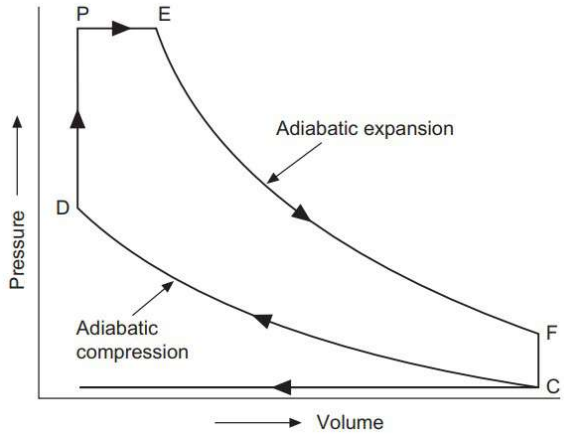


Figura 2.10: Ciclo de calor teórico de un motor diésel [15]

Los componentes principales del motor diésel son:

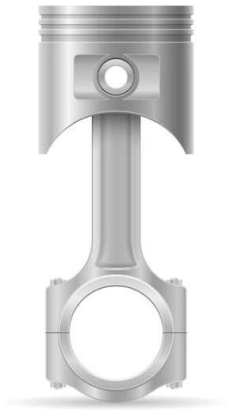
- **Pistón**: Es el componente señalado en la parte superior de la Figura 2.11 (a), esta pieza es la encargada de comprimir la mezcla de aire y combustible, albergando la combustión que desplaza la misma pieza hacia abajo. El material de fabricación debe ser resistente ya que soporta altas presiones y temperaturas, además de poseer una gran resistencia al desgaste y a la corrosión. Frecuentemente se utilizan aleaciones de aluminio, por ejemplo, aluminio-cobre, aluminio cobre níquel y hierro, y aluminio-silicio [16].

- **Bielas**: Son los componentes móviles que se encargan de la transmisión del movimiento desde el pistón al cigüeñal, se indican en la parte media e inferior de la Figura 2.11 (a). Estas piezas deben soportar altas cargas de tracción, compresión y flexión, por lo que se fabrican de acero templado en forja o mecanizando aleaciones de aluminio [17].

- **Turbocompresor**: Componente indicado en la Figura 2.11 (b). El turbocompresor es un sistema de sobrealimentación y su función principal es mejorar el rendimiento del motor, a partir de una detonación mayor para proporcionar más aire en los cilindros.

  En el proceso de combustión además de generar energía también se generan gases de escape, los cuales el turbocompresor reutiliza direccionándolos por una turbina. Esta turbina acciona un compresor que comprime el aire de admisión, permitiendo que se aspire una mayor cantidad de aire. De este modo, se genera una mayor explosión debido a que el aire tiene una mayor densidad y presión, además de tener una mayor concentración de oxígeno, obteniendo así una mejora en la potencia generada [18].

- **Cigüeñal**: Indicado en la Figura 2.11 (c), se encarga de transformar el movimiento lineal de los pistones a un movimiento uniforme circular. Contiene perforaciones por los cuales circula el aceite de lubricación, para evitar el excesivo desgaste de sus elementos que están en constante rozamiento. Frecuentemente son fabricados en acero forjado y hierro [19].



(a) Pistón con biela [20]

(b) Turbocompresor [18]

(c) Cigüeñal [19]

(d) Árbol de levas [21]

(e) Inyectores [22]

(f) Válvulas [23]

Figura 2.11: Parte 1: Componentes principales del motor diésel

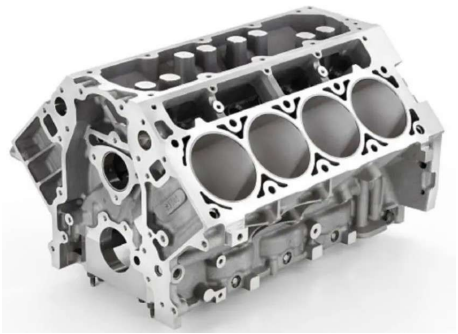
- **Árbol de levas**: Componente indicado en la Figura 2.11 (d), su función es la apertura y cierre de las válvulas de admisión y de escape, además del control de los inyectores. El árbol de levas se instala en la culata del motor y es accionado por el cigüeñal por medio de una cadena de distribución. Usualmente se fabrica con hierro fundido o acero forjado. En cuanto a sus partes principales, el árbol se compone de un eje y las levas, estas últimas son palas que a medida que giran accionan movimientos temporizados del motor [21].

- **Inyectores**: Son componentes del sistema de inyección de combustible, señalados en la Figura 2.11 (e), los inyectores se encargan de suministrar una cantidad de combustible atomizada hacia la cámara de combustión en el ciclo de compresión, para que luego esta se mezcle con aire a elevadas temperaturas. Estos componentes son accionados por el árbol de levas y van instalados en la culata en la parte superior de los cilindros [22].

- **Válvulas**: Indicadas en la Figura 2.11 (f) entre los árboles de levas y los pistones. En los motores las válvulas principales son las válvulas de admisión y las de escape, las cuales son elementos metálicos que permiten el paso de fluidos a la cámara de combustión, ya sea de entrada o salida. Estas válvulas se ubican en la culata y su parte ancha se apoya en la parte superior de los cilindros con el fin de sellarlos.

  Las válvulas de admisión son de mayor tamaño que las válvulas de escape debido a las condiciones de operación. En cuanto a la fabricación de las válvulas, se funden y se mecanizan aceros, implementando materiales más resistentes en el caso de las válvulas de escape que las válvulas de admisión, ya que soportan una temperatura mayor [23].

(a) Culata [24]



(b) Bloque de motor [25]

(c) Cárter [26]

Figura 2.12: Parte 2: Componentes principales del motor diésel

- **Culata**: Señalada en la Figura 2.12 (a), la culata es la parte superior del motor, que cubre las cámaras de combustión y es donde se ubica los árboles de levas con las válvulas.

  Comúnmente se fabrican de acero o aleaciones de aluminio. Las culatas de aluminio tienen un menor peso y buena conducción de calor, sin embargo, son menos resistentes al desgaste a largo plazo y más caras. A su vez, la buena conductividad incide en tener una menor inercia térmica, es decir, la culata se demora menos en calentarse y menos en enfriarse, esto se traduce a que llega más rápido a la temperatura de servicio, pero tarda menos en perderla. Por otro lado, las culatas de acero son más resistentes, económicas y tienen mayor inercia térmica en comparación con las de aluminio [24].

- **Bloque del motor**: Determinado en la Figura 2.12 (b), el bloque del motor es el soporte en donde se sitúan los elementos internos y consiste en solo una pieza con las cavidades necesarias para el montaje del motor. El bloque posee cilindros en donde se ubican los pistones, canales de refrigeración, galerías de aceite y acoplamientos.

  Frecuentemente se fabrica de fundición o de aluminio, en el caso del primero se obtiene una mayor resistencia, en cambio con el segundo se ahorra peso y se obtiene una menor inercia térmica del bloque.

  Una clasificación de los bloques es según la disposición de los cilindros: (i) en línea, ubicado un cilindro detrás de otro en forma vertical, (ii) en v, cilindros ubicados en diagonal y en parejas, (ii) en w, configuración de dos "v" unidas y (iv) boxer, con cilindros horizontales enfrentados [25]. En el caso de la Figura 2.12 (b), se tiene un bloque de motor en v.

- **Cárter**: Es el componente señalado en la Figura 2.12 (c), su función principal es alojar el aceite de lubricación que se implementa para lubricar las piezas internas móviles del motor, esta lubricación se realiza a partir de la succión e impulsión de la bomba de aceite. Con el fin de que no exista vacío en la bomba, el cárter tiene unas placas para minimizar el movimiento del aceite.

  A su vez, el cárter proporciona protección y rigidez al motor, y usualmente se fabrica de acero o aleaciones ligeras de aluminio [26].

Por otro lado, el conjunto de estos y otros componentes del motor crea distintos sistemas con específicas funciones para el correcto uso del motor. Dentro de los sistemas principales se encuentran: (i) el sistema de distribución, encargado de regular el correcto funcionamiento de la entrada y salida de los gases en los cilindros; (ii) el sistema de lubricación, como su nombre lo anticipa, se encarga de lubricar las piezas móviles del motor para evitar el roce y así el desgaste entre ellas; (iii) el sistema de refrigeración o enfriamiento, ya sea a partir de aire o agua, este sistema se encarga de circular un fluido para reducir la temperatura de las piezas; (iv) el sistema de alimentación, encargado de suministrar el combustible, dosificando la cantidad precisa para la combustión; y finalmente (iv) el sistema eléctrico, que proporciona energía eléctrica a cualquier sistema o sensor que necesite de este tipo de energía para su funcionamiento, además de gestionar el módulo de control del motor.

Habitualmente se habla que los motores funcionan a "4 tiempos", los cuales se detallan a continuación y estos se repiten de forma cíclica:

1. Tiempo de Succión: Es cuando el pistón se mueve hacia abajo y se abren las válvulas de admisión para que entre el aire.

2. Tiempo de Compresión: El pistón se mueve hacia arriba con las válvulas cerradas por lo que aumenta la presión y temperatura a un valor más alto que el de auto ignición del diésel.

3. Tiempo de Potencia: Se inyecta diésel en aire comprimido, el cual se evapora y genera una explosión que aumenta la presión y temperatura, desplazando el pistón hacia abajo.

4. Tiempo de Escape: Gracias a la inercia del sistema el pistón se mueve nuevamente hacia arriba y se abren las válvulas de escape, retirando los gases posteriores a la combustión.

2.4.1. Motores diésel marinos

Los motores diésel marinos se diferencian de los terrestres en el tamaño, peso relativo y la disposición de algunas de sus partes. Debido al uso, en los motores marinos se busca un equilibrio en un peso reducido y garantías de seguridad del funcionamiento. En cuanto a su aumento de tamaño este se debe a que un aumento de cilindros genera una mayor fuerza y una uniformidad en la potencia total.

Específicamente, la finalidad de un motor marino es "hacer girar su hélice, provocando con ello un empuje axial que debe recogerse y transmitirse al buque para impulsarlo a la velocidad requerida" [27]. Sin embargo, este empuje debe ser en ambos sentidos para permitir la marcha hacia delante y hacia atrás.

La gran mayoría de la flota mundial de barcos, empleados en los buques de carga, buques de pasaje y en buques de guerra, utilizan motores de cuatro tiempos de velocidad media (hasta alrededor de 1.200 RPM) para la propulsión, un gran porcentaje tiene solo un motor, sin embargo, es común las configuraciones de múltiples motores.

2.5. Base de datos

La base de datos disponible considera dos motores diésel marinos uno de babor y otro de estribor. Los motores corresponden a un motor diésel Wärtsilä 26 de 4 tiempos y 12 cilindros en V, indicado en la Figura 2.13. Ambos motores cuentan con 175.604 registros en 62 parámetros, el primer registro de la base de datos fue en 2016 y el último el 2021, además los registros poseen un paso de 10 minutos entre cada uno. Los parámetros medidos son principalmente de temperatura, presión y velocidad de distintos sistemas o subáreas del motor, por ejemplo, la temperatura del aceite de lubricación, la presión del agua de enfriamiento o la temperatura en los gases de escape.

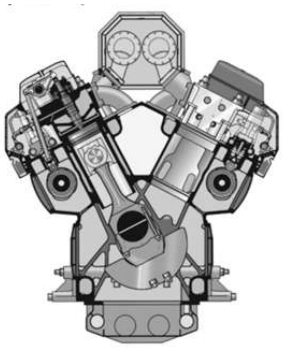


Figura 2.13: Motor Wärtsilä 26 [28]

En esta base de datos se tiene un porcentaje de datos faltantes y se sabe que existieron fallas en los motores en el periodo de adquisición de datos, sin embargo, no se sabe en qué periodo ocurrieron estas fallas ni su cantidad.

2.5.1. Glosario

A continuación se presentan definiciones utilizadas en los parámetros medidos:

- Sensor Multiplexer Units (SMU): Unidades de multiplexor de sensores en español, es parte del sistema de control del motor Wärtsilä 26.

- Distributed Control Units (DCU): Unidades de control distribuidas en español, es parte del sistema de control del motor Wärtsilä 26.
- Main Control Unit (MCU): Unidad de control principal en español, es parte del sistema de control del motor Wärtsilä 26.
- Cojinetes: "Piezas también son conocidas como casquillos, los cuales hacen de punto de apoyo para los ejes y ayudan a reducir el rozamiento cuando se ejercen los movimientos de rotación" [29].

Los parámetros medidos en la base de datos son los siguientes:

Tabla 2.1: Parte 1: Parámetros de medición con el número de etiqueta de cada uno [30]

#	Parámetro de medición	Rango	Unidad
1	Válvula de descarga de aire	0 - 100	%
2	Nivel de Cárter Húmedo de Aceite de Lubricación	0 - 100	%
3	Posición de la Rejilla del Combustible	0 - 40	mm
4	Presión Aceite de Lubricación PT201	0 - 10	bar
5	Presión Aceite de Lubricación PT241	0 - 10	bar
6	Presión Agua de Enfriamiento	0 - 6	bar
7	Presión Agua de Enfriamiento PT401	0 - 6	bar
8	Presión Aire de Carga	0 - 6	bar
9	Presión de Control del Aire de Partida	0 - 40	bar
10	Presión de Aire de Partida	0 - 40	bar
11	Presión de Combustible PT101	0 - 10	bar
12	Presión Diferencial del Filtro de Aceite de Lubricación PDY113	0 - 1,2	bar
13	Presión Diferencial del Filtro de Aceite de Lubricación PDY243	0 - 1,2	bar
14	Presión de Filtro de Seguridad de Combustible	0 - 10	bar
15	Temperatura Aceite de Lubricación	0 - 160	Cº
16	Temperatura Aceite de Lubricación Turbo Compresor A	0 - 650	Cº
17	Temperatura Aceite de Lubricación Turbo Compresor B	0 - 650	Cº
18	Temperatura Agua de Enfriamiento TE401	0 - 160	Cº
19	Temperatura Agua de Enfriamiento TE402	0 - 160	Cº
20	Temperatura Agua de Enfriamiento TE451	0 - 160	Cº
21	Temperatura Aire de Carga TE601	0 - 160	Cº
22	Temperatura Aire de Carga TS601	0 - 160	Cº
23	Temperatura de Cojinete Principal 0	0 - 160	Cº
24	Temperatura de Cojinete Principal 1	0 - 160	Cº
25	Temperatura de Cojinete Principal 2	0 - 160	Cº
26	Temperatura de Cojinete Principal 3	0 - 160	Cº
27	Temperatura de Cojinete Principal 4	0 - 160	Cº
28	Temperatura de Cojinete Principal 5	0 - 160	Cº
29	Temperatura de Cojinete Principal 6	0 - 160	Cº
30	Temperatura de Cojinete Principal 7	0 - 160	Cº

Tabla 2.2: Parte 2: Parámetros de medición con el número de etiqueta de cada uno [30]

#	Parámetro de medición	Rango	Unidad
31	Temperatura de Combustible TE101	0 - 160	Cº
32	Temperatura DCU 1	0 - 160	Cº
33	Temperatura DCU 2	0 - 160	Cº
34	Temperatura Gas de Escape Cilindro A1	0 - 650	Cº
35	Temperatura Gas de Escape Cilindro A2	0 - 650	Cº
36	Temperatura Gas de Escape Cilindro A3	0 - 650	Cº
37	Temperatura Gas de Escape Cilindro A4	0 - 650	Cº
38	Temperatura Gas de Escape Cilindro A5	0 - 650	Cº
39	Temperatura Gas de Escape Cilindro A6	0 - 650	Cº
40	Temperatura Gas de Escape Cilindro B1	0 - 650	Cº
41	Temperatura Gas de Escape Cilindro B2	0 - 650	Cº
42	Temperatura Gas de Escape Cilindro B3	0 - 650	Cº
43	Temperatura Gas de Escape Cilindro B4	0 - 650	Cº
44	Temperatura Gas de Escape Cilindro B5	0 - 650	Cº
45	Temperatura Gas de Escape Cilindro B6	0 - 650	Cº
46	Temperatura Gas de entrada a Turbo Compresor A	0 - 800	Cº
47	Temperatura Gas de salida a Turbo Compresor A	0 - 800	Cº
48	Temperatura Gas de entrada a Turbo Compresor B	0 - 800	Cº
49	Temperatura Gas de salida a Turbo Compresor B	0 - 800	Cº
50	Temperatura MCU	0 - 160	Cº
51	Temperatura SMU 1-2	0 - 160	Cº
52	Temperatura SMU 1-3	0 - 160	Cº
53	Temperatura SMU 1-4	0 - 160	Cº
54	Temperatura SMU 2-2	0 - 160	Cº
55	Temperatura SMU 2-3	0 - 160	Cº
56	Tiempo de Funcionamiento del Motor UI794	-	minutos
57	Tiempo de Funcionamiento del Motor UI795	-	[1 - 9999]
58	Tiempo de Funcionamiento del Motor UI796	-	[10000 -]
59	Velocidad del Motor	0 - 1200	RPM
60	Velocidad del Sistema de BackUp del Motor	0 - 1200	RPM
61	Velocidad Turbo Compresor A	0 - 44000	RPM
62	Velocidad Turbo Compresor B	0 - 44000	RPM

Capítulo 3

Metodología

La metodología de trabajo se divide en tres etapas, (i) Estudio Previo, (ii) Implementación de Metodologías de Imputación y (iii) Diagnóstico Inteligente de Fallas, en la Figura 3.1 se indican las subetapas por cada etapa en los colores celeste, verde y café respectivamente.

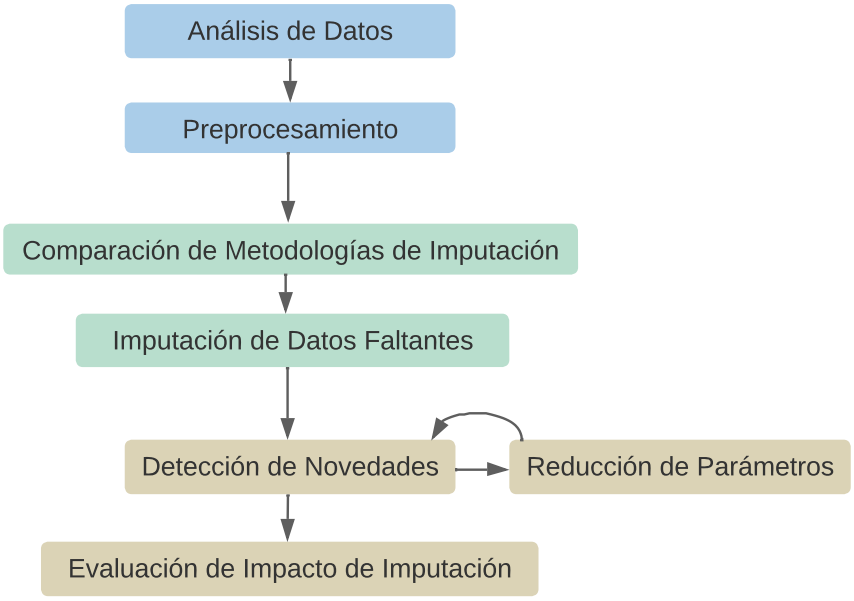


Figura 3.1: Diagrama de Metodología de Trabajo

3.1. Estudio previo

En primer lugar se analizan los datos, estudiando la magnitud de datos faltantes y la ubicación de estos en la base de datos. Para esto se implementa la librería de *numpy* en *Python* utilizando la función *df.isna().sum()* para contar los datos faltantes de cada parámetro, denominando *df* a los datos en *DataFrames* (paneles bidimensionales compuestos por filas y columnas) de aquí en adelante. A su vez, se implementa la librería *Missingno* con la función *df.matrix()* para visualizar los datos faltantes.

Posteriormente se ejecuta el siguiente preprocesamiento:

1. Se agrupan los parámetros según la subárea que mida el sensor. La clasificación utilizada es similar a la expuesta en la lista de sensores del motor principal [30], sin embargo, se limita el número mínimo de parámetros en un grupo a tres parámetros.

   La agrupación de parámetros se realiza para agilizar el proceso de imputación de datos faltantes, para no trabajar con las 175.604 observaciones de los 62 parámetros. Además, se considera el supuesto de una mayor correlación entre los parámetros del mismo grupo, lo cual mejora el proceso de imputación de datos faltantes.

   Luego se visualizan las correlaciones de los grupos con la función *df.corr()*, esta función utiliza el método de "Pearson" que considera las siguientes interpretaciones para los coeficientes positivos:

   Tabla 3.1: Interpretación de coeficientes de correlación [31]

Valor	**Significado**
0	Correlación nula
0,01 a 0,19	Correlación positiva muy baja
0,2 a 0,39	Correlación positiva baja
0,4 a 0,69	Correlación positiva moderada
0,7 a 0,89	Correlación positiva alta
0,9 a 0,99	Correlación positiva muy alta
1	Correlación positiva perfecta

   Una correlación positiva indica que ambas variables tienden a incrementarse juntas, en cambio en una correlación negativa, mientras una variable tiende a incrementar, la otra variable desciende. Para los coeficientes negativos la interpretación es análoga a la Tabla 3.1, cambiando los términos positivos por negativos.

2. Se consideran datos faltantes a todos los valores que estén fuera del umbral de medición del parámetro, información obtenida de la lista de sensores del motor principal [30] y señalada en las Tablas 2.1 y 2.2. Posteriormente, se realizan histogramas a cada parámetro, con el fin de analizar las tendencias de sus distribuciones.

   Luego se calcula y analiza el coeficiente variación de cada parámetro a partir de la siguiente fórmula:

   $$CV = \frac{\sigma}{\bar{x}} \cdot 100 \tag{3.1}$$

   Donde σ es la desviación estándar del parámetro, $\bar{x}$ el promedio del parámetro y la unidad de medida CV es adimensional en porcentaje.

3. Se eliminan observaciones que no tienen datos en todos los parámetros, tal como se menciona en los alcances, con el propósito de suprimir ventanas amplias de datos faltantes que generan un mayor sesgo en la imputación de datos faltantes. Esto se realiza con la función *df.dropna(thresh=1)*, que conserva las observaciones con al menos un valor.

4. Finalmente, se escalan los datos utilizando la función *MinMaxScaler()* del paquete *sklearn.preprocessing* de *Python*. Esta función transforma los datos para que el valor mínimo de la columna tome el valor 0, el máximo el valor 1 y los datos restantes sigan las Ecuaciones 3.2 y 3.3 [32].

$$X_{scaled} = X_{std} \cdot \ (max - min) + min \tag{3.2}$$

$$X_{std} = \frac{X - X.min(axis = 0)}{X.max(axis = 0) - X.min(axis = 0)} \tag{3.3}$$

Donde, max corresponde al valor 1, min el valor 0, $X.min(axis = 0)$ y $X.mas(axis = 0)$ son los vectores que contienen a los mínimos y máximos de cada columna, respectivamente y X es el arreglo de los datos a escalar.

Este procedimiento se realiza para que en el proceso de imputación de datos faltantes no se vea afectado por el orden de magnitud de los parámetros de un mismo grupo.

3.2. Implementación de metodologías de imputación

Continuando con la segunda etapa de la metodología, denominada implementación de metodologías de imputación, se tiene que la primera subetapa indicada en la Figura 3.1 es comparar las metodologías de imputación de datos. Esta se realiza en el motor de babor de la base de datos y se extrapola al motor de estribor, ya que el motor de babor posee mayor porcentaje de datos faltantes considerándose el caso crítico de ambos.

Para calcular el desempeño de cada metodología es necesario contar con los valores reales a los cuales se quiere llegar, debido a esto, el primer paso de la comparación es eliminar todas las observaciones que posean al menos un "NaN". Posteriormente, se generar ventanas de datos faltantes de forma aleatoria, puesto que los datos faltantes no suelen estar aislados en las bases de datos, sino que se encuentran de forma continua ya sea en un mismo parámetro o en grupos de parámetros. Este procedimiento se realiza para cada grupo definido en la primera etapa del preprocesamiento, en donde se introducen "NaN" hasta llegar a un porcentaje de valores faltantes definido como: el promedio de datos faltantes de los parámetros del grupo, luego de la etapa de preprocesamiento.

En los nuevos datos de cada grupo se aplican las metodologías de imputación de datos faltantes, para esto se utiliza, en la mayoría de los casos, el paquete *sklearn.model_selection* que posee la librería *train_test_split* para entrenar y ajustar los datos, destacando que en los procedimientos de imputación de este trabajo se emplea el mismo conjunto para ambas acciones. A su vez, se implementa el paquete *sklearn.impute* que contiene la mayoría de las librerías y modelos para realizar la imputación de datos, utilizando en cada modelo la configuración *add_indicator=True* para que se generen etiquetas binarias, con 1 si se realizó la imputación y 0 si no, ejerciendo así, un seguimiento en el procedimiento de imputación.

Los errores de estas imputaciones se designan MSE^* y se determinan por la función del error cuadrático medio, *mean_squared_error* de *sklearn.metrics* en *Python*, aplicado al valor real y el valor predicho imputado, dividido por el promedio de cada parámetro luego del preprocesamiento.

Las metodologías implementadas fueron las siguientes:

- **Imputación Media:** Se implementó el modelo *SimpleImputer* con la estrategia "*mean*".
- **Imputación con *Fill Foward*:** Única metodología que no implementa las librerías *train_test_split* ni *sklearn.impute*. En cambio, se ejecuta la función *df.ffill(axis = 0)*, que completa con el valor anterior de cada parámetro.
- **Imputación con *K-Nearest Neighbors (KNN)*:** Se utiliza el modelo de *KNNImputer*, variando los hiperparámetros *weights* y *n_neighbors*. En el caso de los pesos se varía en pesos uniformes o basados en distancia y para la cantidad de vecinos se evalúa hasta el número mayor de vecinos de los grupos en donde el error calculado de los parámetros aumente al aumentar la cantidad de vecinos, es decir, hasta el número mayor de vecinos que evidencie un mínimo en el error calculado.
- **Imputación con *Multivariate Imputation By Chained Equations (MICE)*:** En esta metodología se aplica un estimador experimental de *scikit-learn* por lo que se utiliza el paquete de *sklearn.experimental*, específicamente *enable_iterative_imputer*. De este modo, se utiliza desde *enable_iterative_imputer* el modelo *IterativeImputer* con el estimador predeterminado *BayesianRidge()* que es un algoritmo de regresión lineal. Los datos faltantes se completan en orden ascendente, es decir, desde los parámetros que tienen menos valores perdidos hasta los que tienen más. Además, el hiperparámetro *sample_posterior*, por defecto es "False", por lo que se realiza una imputación simple en vez de múltiples imputaciones. El hiperparámetro modificado fue la cantidad de iteraciones.
- **Imputación con un árbol de decisión:** Al igual que en imputación con MICE, se implementa el modelo *IterativeImputer* por lo que se utilizan los mismos paquetes mencionados en esa metodología, sin embargo, cambia el algoritmo de regresión aplicado a *DecisionTreeRegressor* implementando el paquete *sklearn.tree*. En cuanto a los hiperparámetros modificados del modelo, se varía la cantidad de iteraciones y la tolerancia del criterio de convergencia.
- **Imputación con *Random Forest*:** Procedimiento similar a la imputación con un árbol de decisión, no obstante, cambia el algoritmo de regresión implementado, utilizando el modelo *RandomForestRegressor* de *sklearn.ensemble*.
- **Imputación con *Extremely Randomized Tree*:** Se realiza un procedimiento similar a la imputación con un árbol de decisión, sin embargo, cambia el algoritmo de regresión implementado, utilizando el modelo *ExtraTreeRegressor* de *sklearn.tree*.

Al obtener la metodología de cada parámetro que obtuvo mejores resultados, es decir, el procedimiento de imputación de datos faltantes que consiguió el mínimo error, se escogen siete parámetros característicos de la base de datos para visualizar los valores reales e imputados. La elección de estos parámetros fueron los tres que poseen los errores más bajos, otros

tres que tienen los errores más altos y el parámetro que tiene el valor promedio de error tras eliminar los seis parámetros anteriores.

Finalmente, la subetapa de "Imputación de Datos Faltantes" indicada en verde en la Figura 3.1, se realizó tras obtener los resultados de la comparación, imputando así los valores faltantes con la mejor metodología para cada grupo. Esto se realizó para ambos motores bajo los hiperparámetros óptimos de cada modelo identificados en la comparación de metodologías de imputación.

3.3. Diagnóstico Inteligente de Fallas

La última etapa de la metodología es el diagnóstico inteligente de fallas, en donde su primera subetapa es la detección de novedades que trabaja de manera simultánea con la segunda subetapa de reducción de parámetros tal como se indica en la Figura 3.1 en los reglones cafés.

La reducción de parámetros empleada es *Principal Component Analysis*, para ello es necesario centrar los datos con la función *StandardScaler()* del paquete *sklearn.preprocessing* que estandariza los datos. Luego, se ajustan los datos con el modelo *PCA(n_components=3)* del paquete *sklearn.decomposition*, la elección de 3 componentes se determina para una cómoda visualización futura. En cuanto a la detección de novedades, se utiliza la librería *EllipticEnvelope* del paquete *sklearn.covariance*, el modelo implementado es *EllipticEnvelope(contamination=0,0001)* con un porcentaje de contaminación bajo y se opta por definir el conjunto de entrenamiento con los datos equivalentes a 2 meses de operación (8.640 observaciones), los cuales en la detección de novedades corresponden a datos normales y el conjunto de testeo corresponde al resto de los datos.

Para la tercera subetapa se tiene la evaluación del impacto de la imputación de datos. Por ello, se define un nuevo procedimiento de preprocesamiento el cual posee los puntos 1 y 2 del preprocesamiento actual considerando luego la eliminación de los parámetros que contengan más de un 20 % de datos faltantes, esto se realiza para eliminar la menor cantidad de observaciones puesto que posteriormente se eliminan todas las observaciones que contengan valores perdidos. Tras realizar el preprocesamiento del segundo procedimiento se realiza el diagnóstico inteligente de fallas de igual manera que el procedimiento con imputación de datos faltantes.

La comparación de estos dos procedimientos se evalúa luego de realizar la detección de novedades en ambos casos, comparando:

- Cantidad de parámetros y observaciones.
- Cantidad y porcentaje de anomalías.
- Visualización de identificación de daño.

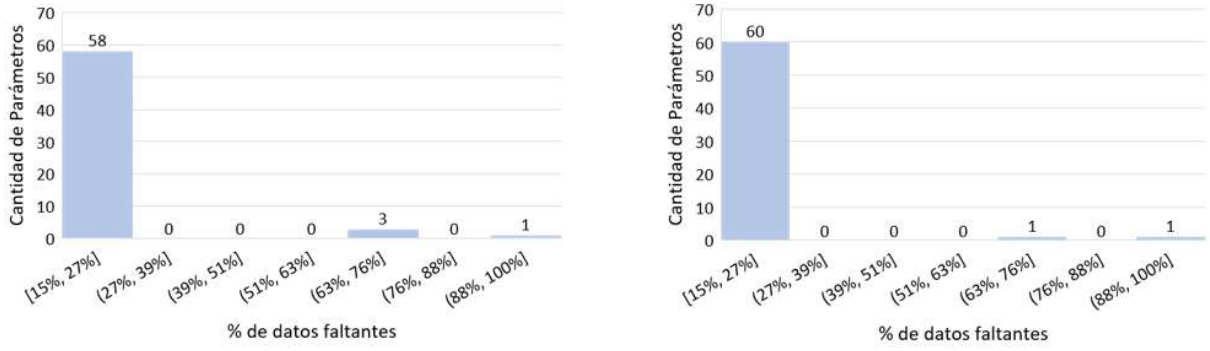
Capítulo 4

Resultados y discusiones

4.1. Estudio previo

4.1.1. Análisis de Datos

Tras utilizar la función *df.isna().sum()* de la librería de *numpy* se tiene que los datos faltantes en el caso del motor de babor corresponde a un 24 % del total, en cambio para el motor de estribor se cuenta con un 20 % de datos faltantes. En la Figura 4.1, se aprecian los histogramas de porcentaje de datos faltantes de los parámetros en cada motor, obteniendo que la mayoría de los parámetros posee un rango de 15 a 27 % de datos faltantes en ambos motores.



(a) Motor babor

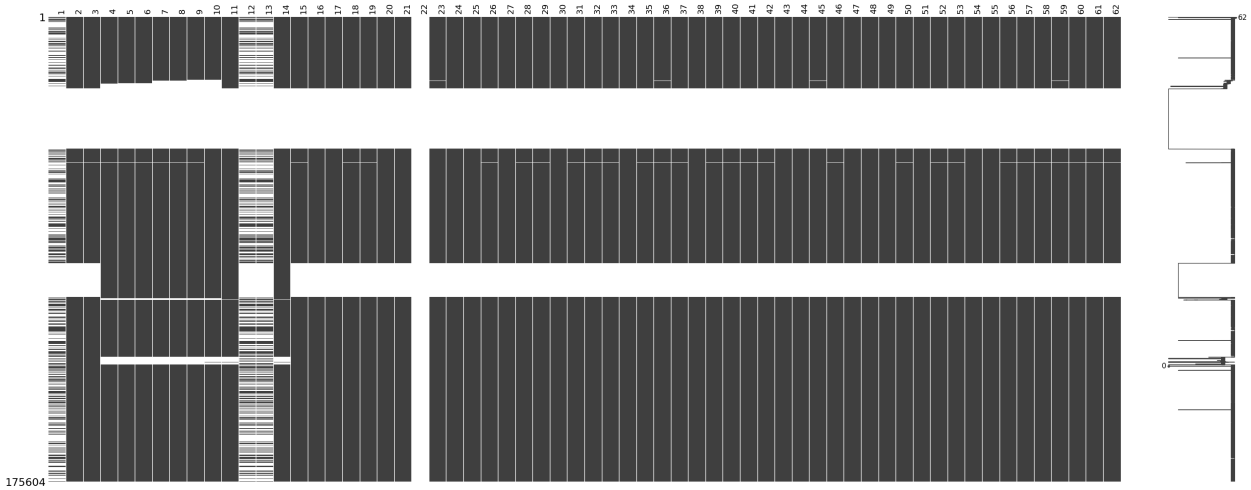
(b) Motor estribor

Figura 4.1: Histograma de porcentaje de datos faltantes en los parámetros de las bases de datos (EP)

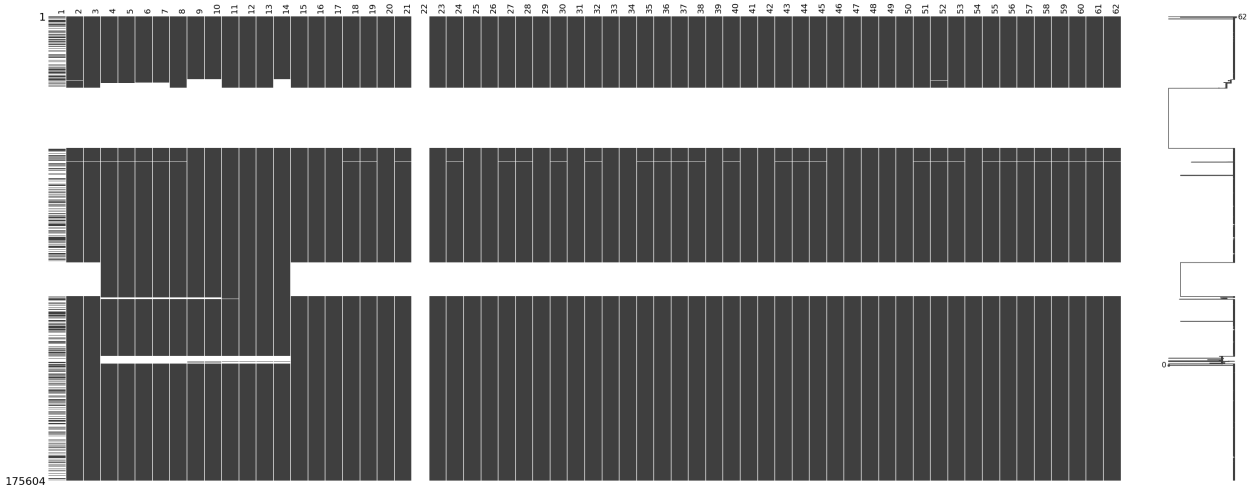
En los histogramas se aprecia que existe un parámetro en ambos motores el cual excede el 80 % de datos faltantes, por lo que no se considera en el estudio según lo estipulado en los alcances, este parámetro corresponde a la "Temperatura de Aire de Carga TS601", etiquetado con el número 22.

Los datos faltantes de los motores se distribuyen como se indica en la Figura 4.2. Al aplicar la función *df.matrix()* en la base de datos se visualizan los 62 parámetros a la horizontal y las 175.604 observaciones en la vertical, los espacios en blanco representan los registros sin

datos y los espacios en negro representan los registros que poseen datos. Además, en el lado derecho de la gráfica principal hay otro gráfico que va desde 0, indicado a la izquierda, hasta el número total de parámetros indicado en la esquina superior derecha, si la línea está lo máximo a la derecha significa que en todos los parámetros hay datos y a medida que más parámetros posean datos faltantes, la línea se moverá hacia la izquierda.



(a) Motor babor



(b) Motor estribor

Figura 4.2: Función matrix de librería Missingno para visualización de datos faltantes en base de datos (EP)

En la Figura 4.2 se visualiza que los datos faltantes no se encuentran de forma aislada, sino que se distribuyen en faltas continuas. Además, se aprecia que el parámetro 22 eliminado en ambos motores posee de manera gráfica solamente espacios en blanco, también se observa que en ambos motores en aproximadamente el segundo año de observaciones, existió un largo periodo de adquisición de datos en donde no hay datos en ningún parámetro, estas observaciones se eliminarán en este estudio tal como se mencionó en los alcances.

Por otro lado, los parámetros 56, 57 y 58 quedan fuera del estudio, puesto que estos miden el tiempo de funcionamiento del motor que no interfiere directamente en el estado de salud de los motores al ser un contador.

4.1.2. Preprocesamiento

Los grupos formados en esta investigación se indican en la Figura 4.3, señalando la cantidad de parámetros de cada uno. El detalle de los parámetros pertenecientes a cada grupo se encuentra en las Tablas A.1 y A.2 de los Anexos.

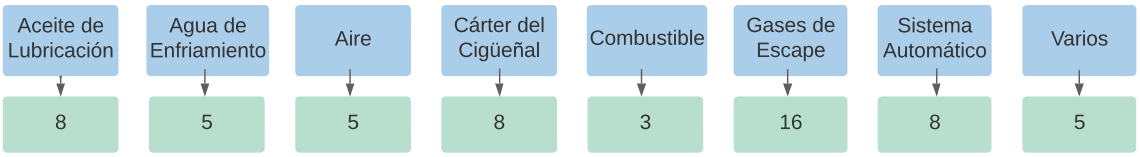
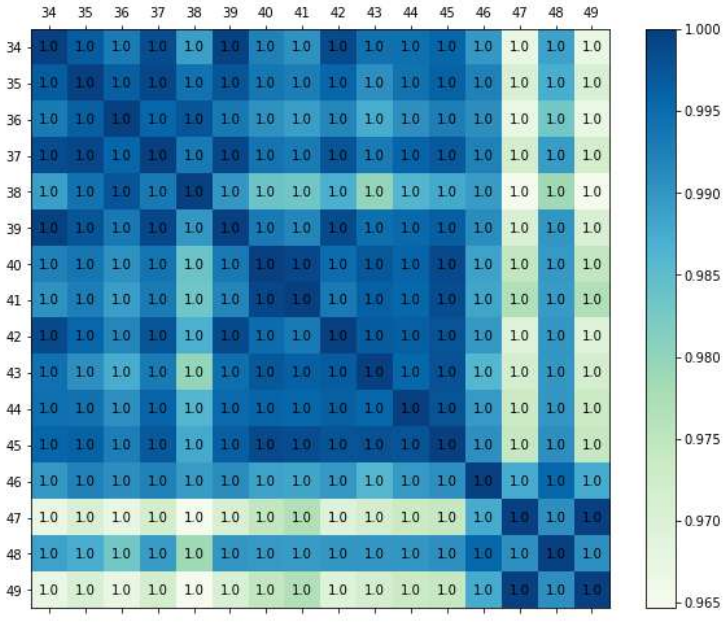


Figura 4.3: Cantidad de parámetros por subáreas de un motor (EP)

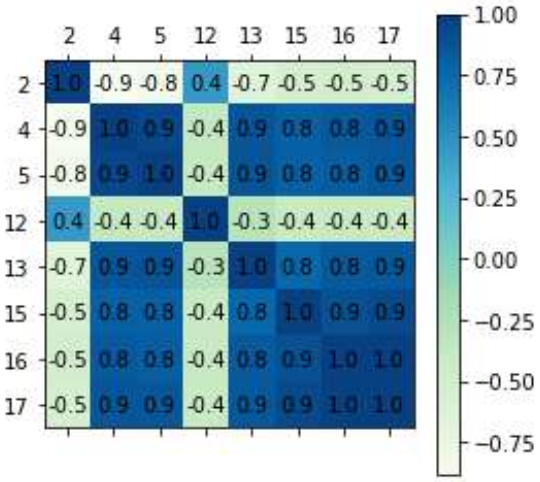
Debido a la limitación de tres parámetros al número mínimo de parámetros pertenecientes en cada grupo, mencionada en la metodología, existen dos diferencias con la clasificación del documento de lista de sensores [30]. La primera es que el grupo de Aceite de Lubricación posee dos parámetros extra, los cuales son dos sensores que miden la presión diferencial del filtro del aceite de lubricación. En segundo lugar, los grupos generados se diferencian tras existir una fusión en los grupos de descarga de aire, aire de partida y carga de aire, formando un nuevo grupo denominado "Aire".

Las correlaciones de estos grupos para el motor de babor se indican en las Figuras 4.4, 4.5 y 4.6, para el motor de estribor sus correlaciones se encuentran en las Figuras del Anexo B.1.

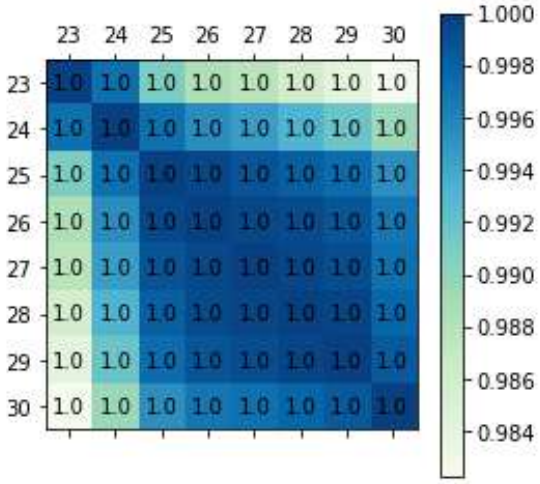


(a) Gases de Escape

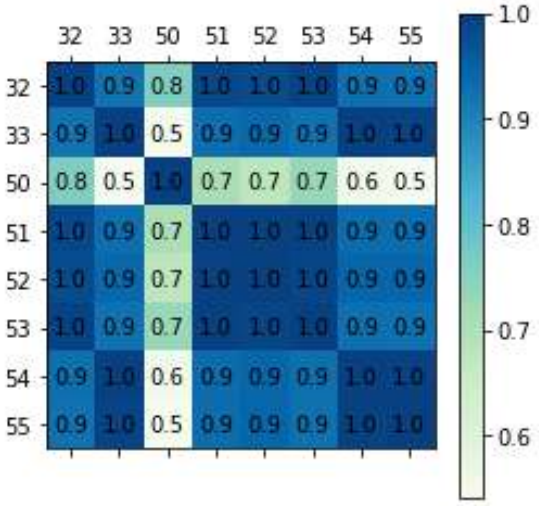
Figura 4.4: Parte 1: Matrices de correlación entre parámetros de grupos formados para el motor de babor (EP)



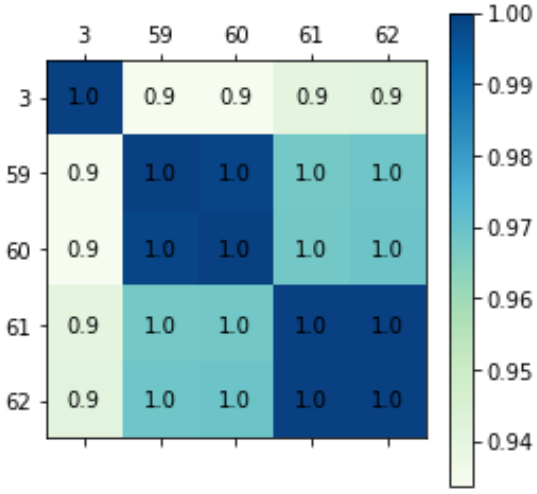
(a) Aceite de Lubricación



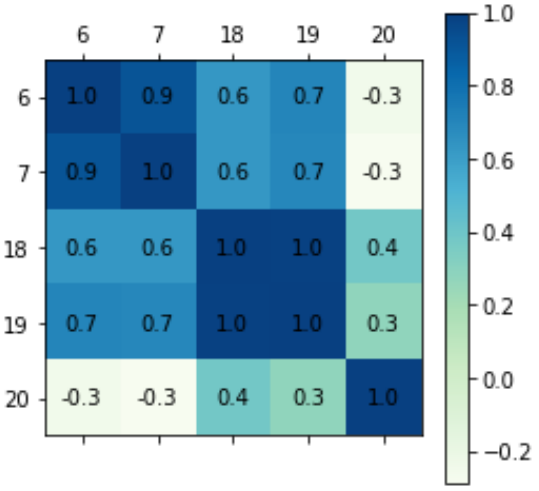
(b) Cárter del Cigüeñal



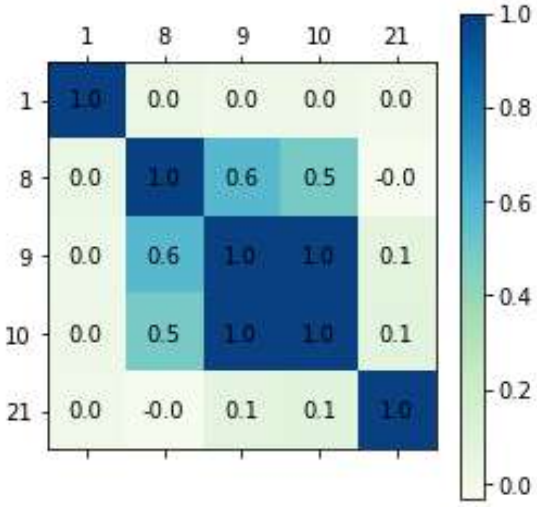
(c) Sistema automático



(d) Varios

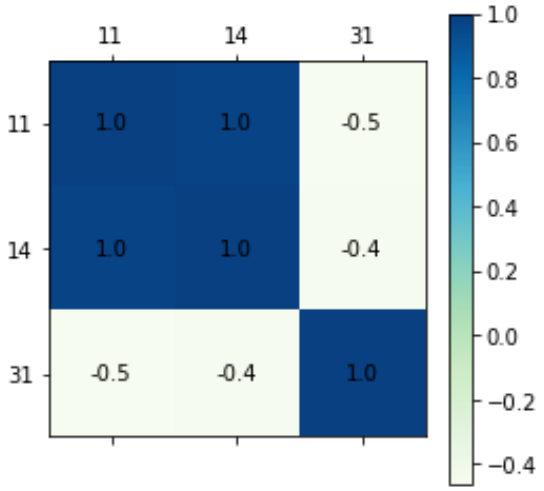


(e) Agua de Enfriamiento



(f) Aire

Figura 4.5: Parte 2: Matrices de correlación entre parámetros de grupos formados para el motor de babor (EP)

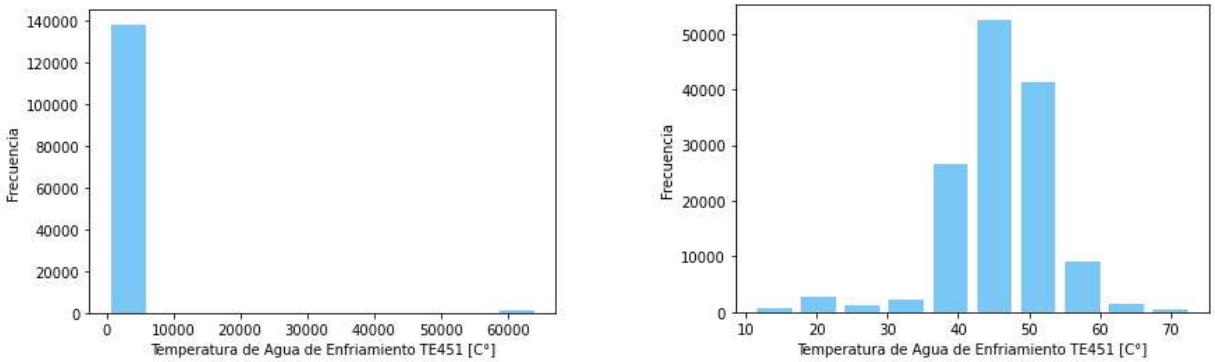


(a) Combustible

Figura 4.6: Parte 3: Matrices de correlación entre parámetros de grupos formados para el motor de babor (EP)

A partir de las matrices de correlación se aprecia una alta correlación positiva en todos los parámetros de los grupos de Gases de Escape y Cárter del Cigüeñal. En cuanto a los otros grupos se generan subconjuntos de altas correlaciones positivas mayoritariamente, sin embargo, se encuentran parámetros con bajas correlaciones en sus grupos, por ejemplo, el parámetro 12, 20, 21 y 31, o inclusive nulas como el parámetro 1. De manera general, se cumple el supuesto de una mayor correlación entre los parámetros de los grupos formados.

Tal como se mencionó en la metodología, se convierten a datos "NaN" a todos los valores que estén fuera del umbral de medición del parámetro. En la Figura 4.7 se aprecia el histograma del parámetro que mide la temperatura del agua de enfriamiento antes y después de realizar este filtro. Este procedimiento se realizó para los 58 parámetros.

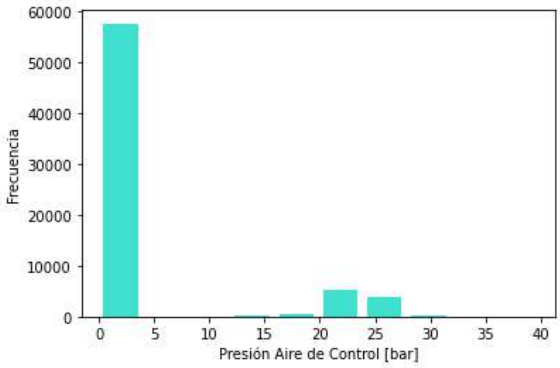


(a) Datos sin preprocesamiento

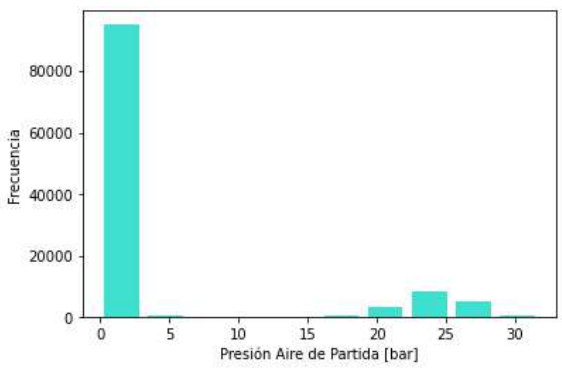
(b) Datos eliminando valores

Figura 4.7: Histogramas de parámetro 20 antes y después de eliminar valores fuera del umbral de medición (EP)

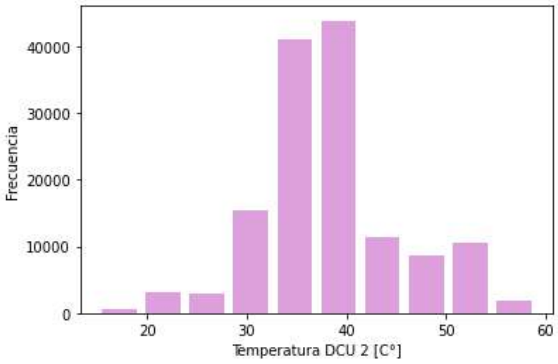
En la Figura 4.8 se aprecian los histogramas de los siete parámetros mencionados en la metodología, los cuales se compararán en la siguiente sección. En cuanto a los histogramas de los parámetros restantes se encuentra en los Anexos B.2.



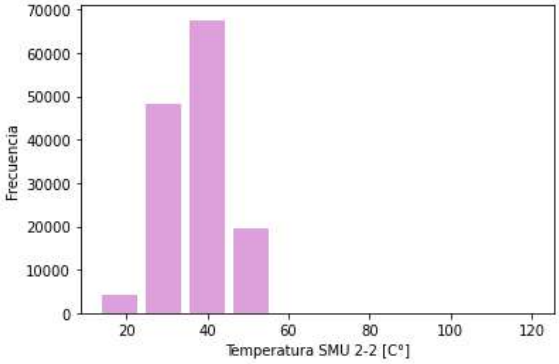
(a) Parámetro 9 de Aire



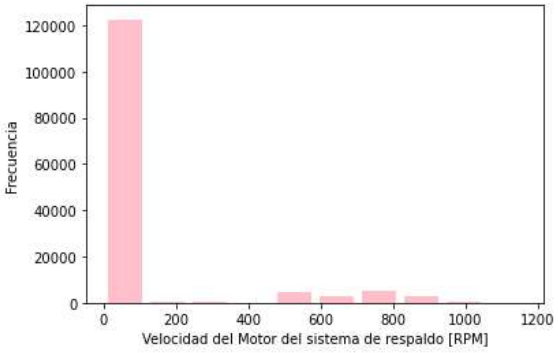
(b) Parámetro 10 de Aire



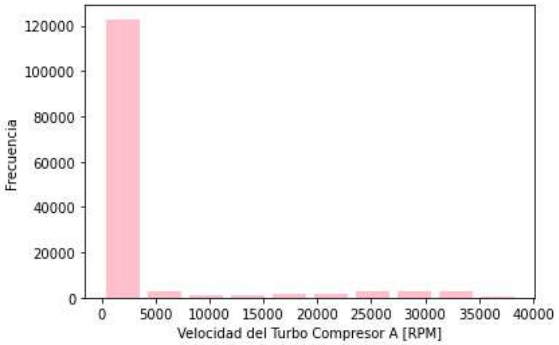
(c) Parámetro 33 de Sistema Automático



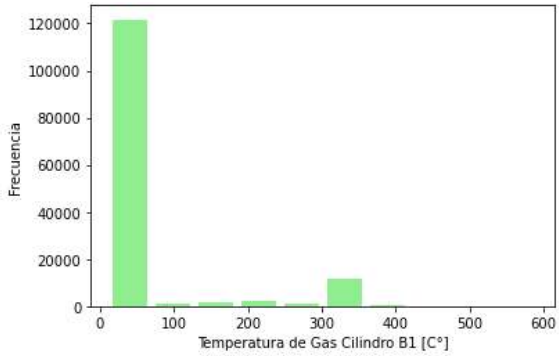
(d) Parámetro 54 de Sistema Automático



(e) Parámetro 60 de Varios



(f) Parámetro 61 de Varios

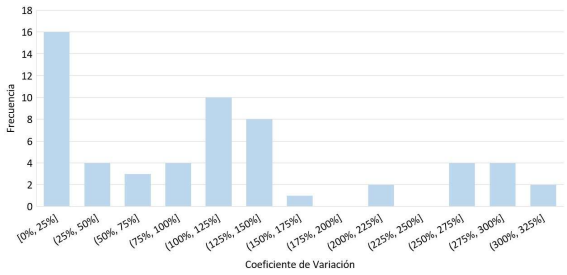


(g) Parámetro 40 de Gases de Escape

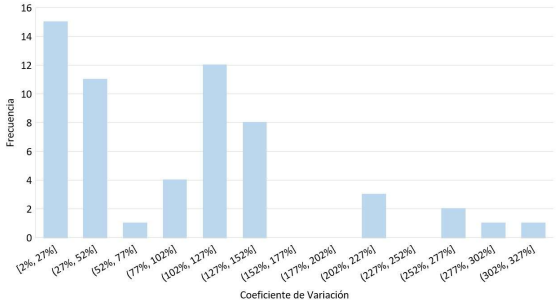
Figura 4.8: Histogramas de parámetros característicos (EP)

En los histogramas de los parámetros 9, 10, 60, 61 y 40 se aprecian dos distribuciones en los registros, la primera con valores cercanos a cero en una alta frecuencia y luego distribuciones de mayor magnitud según el parámetro medido con una menor frecuencia. Por otro lado, para los parámetros 33 y 54, existe una mayor frecuencia en los valores cercanos a la moda, ajustándose cada uno por separado a solo una distribución.

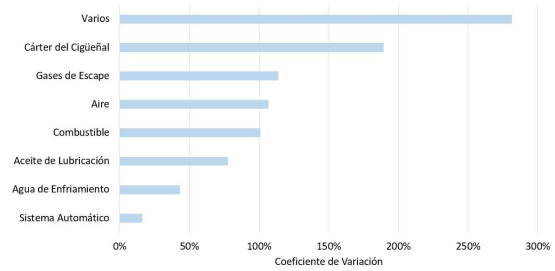
A continuación, se presentan los histogramas de los coeficientes de variación para cada motor y el coeficiente promedio de cada grupo para ambos motores.



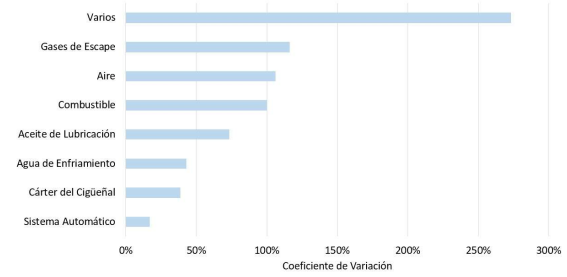
(a) Histograma de coef. de variación del motor de babor



(b) Histograma de coef. de variación del motor de estribor



(c) Coef. de variación promedio de grupos del motor de babor



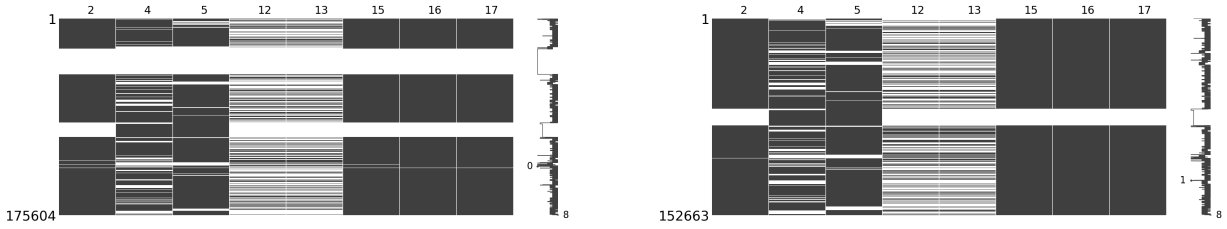
(d) Coef. de variación promedio de grupos del motor de estribor

Figura 4.9: Coeficientes de Variación (EP)

A diferencia de los histogramas de datos faltantes, los histogramas de los coeficientes de variación para los motores de babor y estribor no poseen grandes similitudes en la frecuencia de los rangos obtenidos, sin embargo, en ambos motores existe una mayor frecuencia entre los rangos de coeficiente de variación de 0 a 27 % y se observa que a lo menos la mitad de los parámetros poseen un coeficiente de variación mayor al 50 %.

Al promediar los coeficientes de variación por grupo, se aprecian en ambos motores tendencias similares a excepción del grupo del Cárter del Cigüeñal, ya que en el motor de babor posee un mayor promedio del coeficiente de variación. Esto se debe ya que, al ver los histogramas de los parámetros de ambos motores para el grupo del Cárter del Cigüeñal, para el motor de babor se alcanza mayores valores de temperatura y en algunos parámetros la frecuencia máxima del rango de moda es mayor. No obstante, si no se considera el grupo mencionado para ambos motores el orden decreciente del promedio del coeficiente de variación para los grupos es: Varios, Gases de Escape, Aire, Combustible, Aceite de Lubricación, Agua de Enfriamiento y Sistema automático. El detalle de los coeficientes de variación se indica en las Tablas B.1 y B.2 del Anexo B.3.

Continuando con el preprocesamiento, en la Figura 4.10 se visualiza la eliminación de observaciones sin datos en ningún parámetro para el grupo de Aceite de Lubricación. Esta acción modifica la cantidad de observaciones de 175.604 a 152.663 para este grupo en particular.



(a) Datos eliminando valores fuera de umbral de medición

(b) Datos eliminando observaciones sin datos en todos los parámetros

Figura 4.10: Función matrix de librería Missingno para visualización de datos faltantes en grupo de aceite de lubricación (EP)

De este modo se generan tres etapas en cuanto al preprocesamiento de los datos: (i) Sin preprocesamiento, (ii) Datos con valores dentro del rango de medición y (iii) Datos con al menos una medición en cada observación.

En las Figuras 4.11 y 4.12 se indican los porcentajes de datos faltantes para estas tres etapas en cada grupo de los dos motores. El porcentaje de la tercera etapa suele ser el menor a excepción del caso del grupo de Aire para el motor de babor y esto se debe a la gran cantidad de valores fuera de rango en el parámetro 9, que corresponde a la presión de control del aire de partida. Además, en los gráficos de ambos motores se aprecia nuevamente que el motor de babor es el peor escenario de los dos, al tener un mayor porcentaje de datos faltantes.

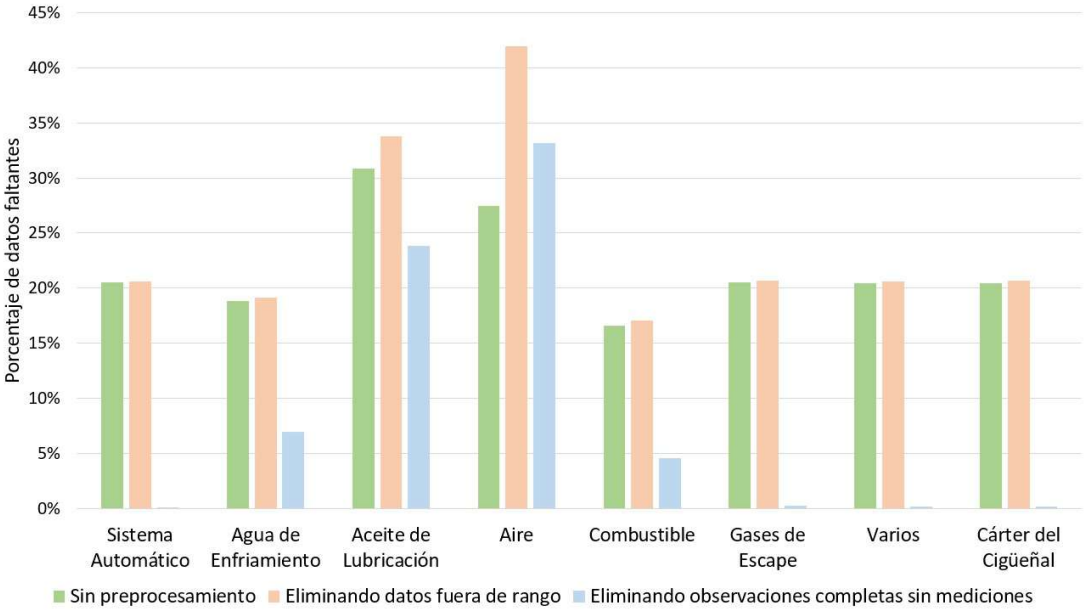


Figura 4.11: Porcentaje promedio de datos faltantes del motor de babor según subetapa de preprocesamiento y grupo (EP)

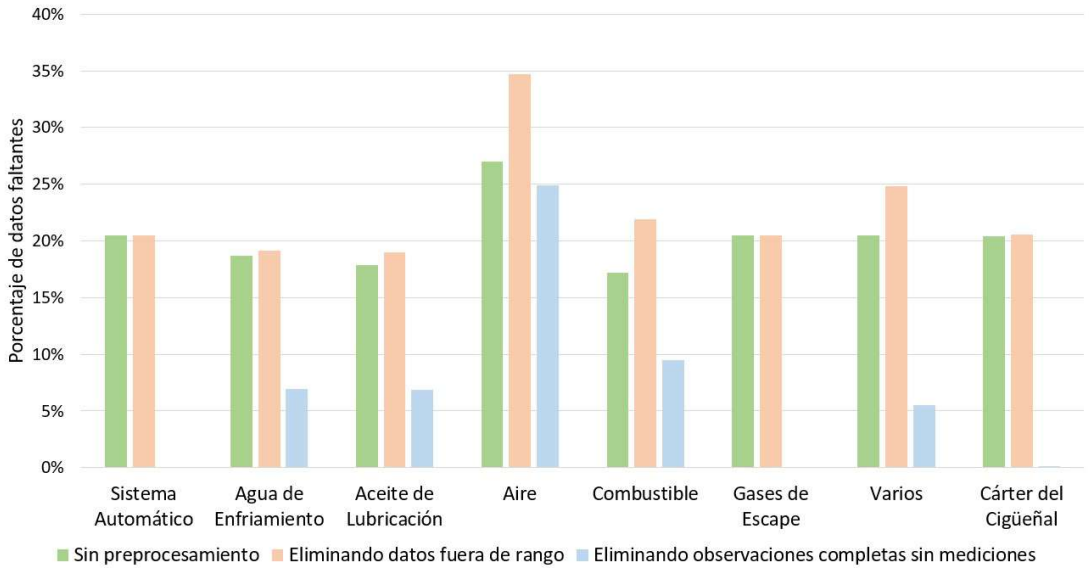
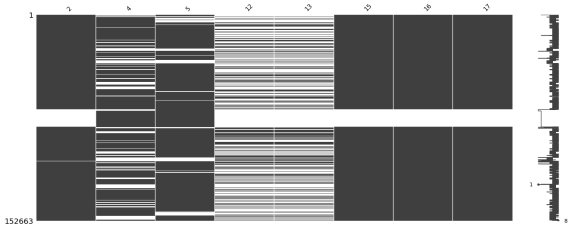


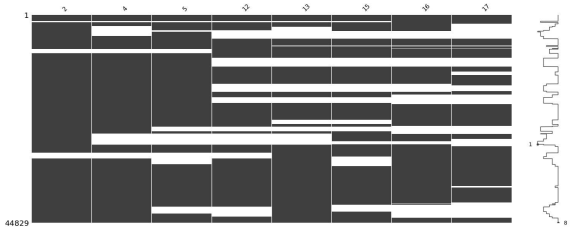
Figura 4.12: Porcentaje promedio de datos faltantes del motor de estribor según subetapa de preprocesamiento y grupo (EP)

4.2. Implementación de metodologías de imputación

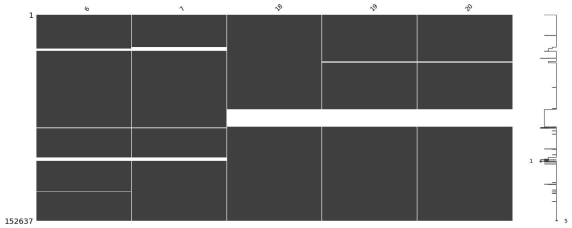
En las Figuras 4.13, 4.14 y 4.15, se visualiza para cada grupo la distribución de datos faltantes luego del preprocesamiento en la izquierda, para compararlos con los valores faltantes insertados de forma aleatoria en la derecha. El porcentaje de valores aleatorios insertado corresponde al porcentaje de datos faltantes luego de eliminar las observaciones completas sin registros para cada grupo.



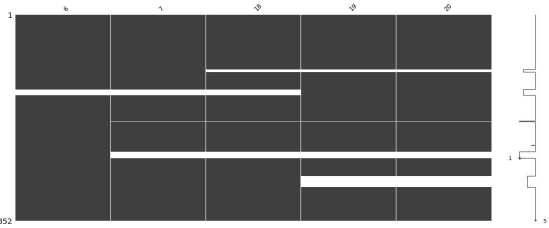
(a) Grupo Aceite de Lubricación luego de preprocesamiento



(b) Grupo Aceite de Lubricación ventanas aleatorias 24 %



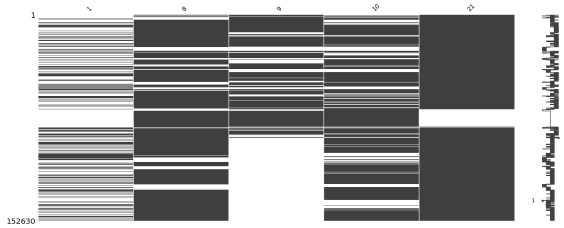
(c) Grupo Agua de Enfriamiento luego de preprocesamiento



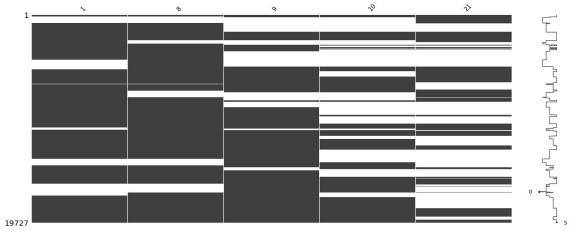
(d) Grupo Agua de Enfriamiento ventanas aleatorias 7 %

Figura 4.13: Parte 1: Visualización de datos faltantes (EP)

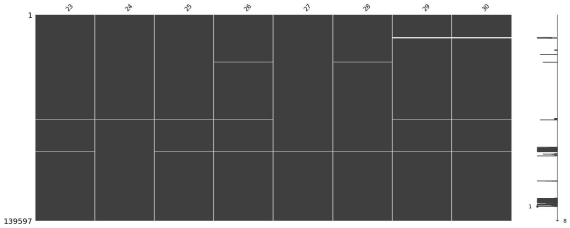
Para el grupo de Aceite de Lubricación se aprecia que los parámetros 12 y 13, que corresponden a los sensores que miden la presión diferencial del filtro del aceite de lubricación, son los parámetros más afectados por los datos faltantes, sin embargo, en las ventanas aleatorias no, tras introducirlos de forma aleatoria. De manera similar, en el grupo de Aire el parámetro 9, que mide la presión de control del aire de partida, a partir de cierta observación no se aprecian más registros, al contrario de lo que sucede en las ventanas aleatorias. La finalidad de no recrear exactamente la forma en la que faltan estos datos ni que se repitan los parámetros críticos, es para no crear un sobreajuste en el modelo. En cuanto a los grupos restantes se aprecia que son semejantes los datos luego del preprocesamiento y los datos al insertar las ventanas aleatorias, en criterios de la cantidad y magnitud de las ventanas faltantes, a pesar de que cambia la ubicación de estas faltas.



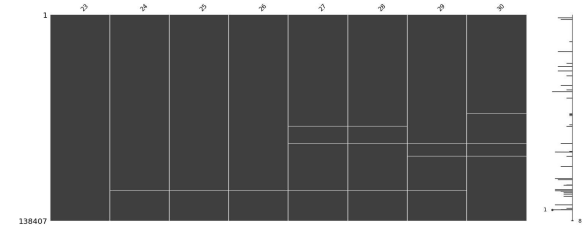
(a) Grupo Aire luego de preprocesamiento



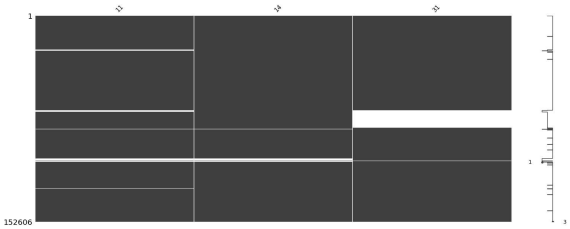
(b) Grupo Aire ventanas aleatorias 33 %



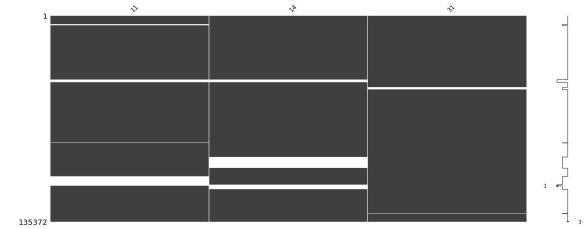
(c) Grupo Cárter del Cigüeñal luego de preprocesamiento



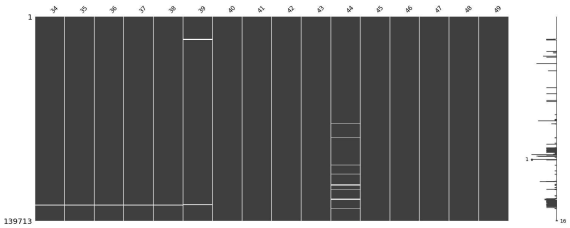
(d) Grupo Cárter del Cigüeñal ventanas aleatorias 0,23 %



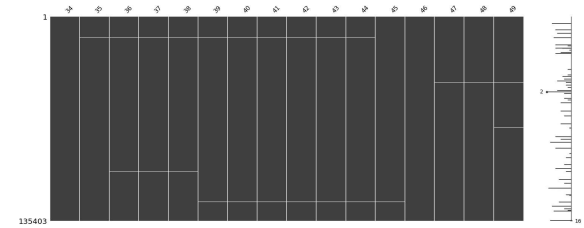
(e) Grupo Combustible luego de preprocesamiento



(f) Grupo Combustible ventanas aleatorias 5 %

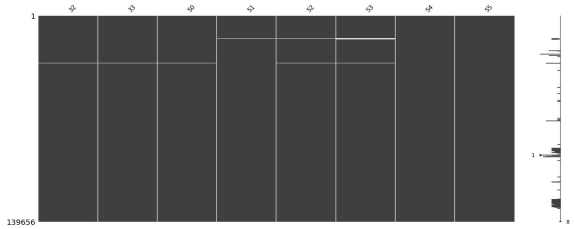


(g) Grupo Gases de Escape luego de preprocesamiento

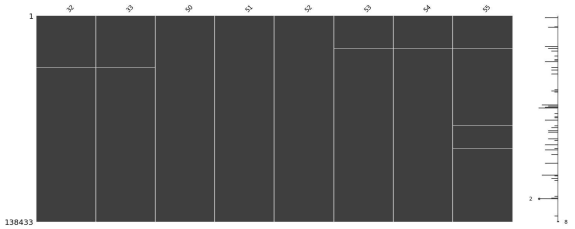


(h) Grupo Gases de Escape ventanas aleatorias 0,29 %

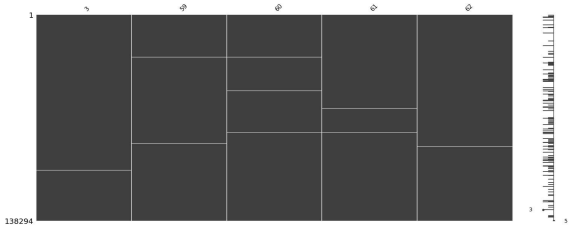
Figura 4.14: Parte 2: Visualización de datos faltantes (EP)



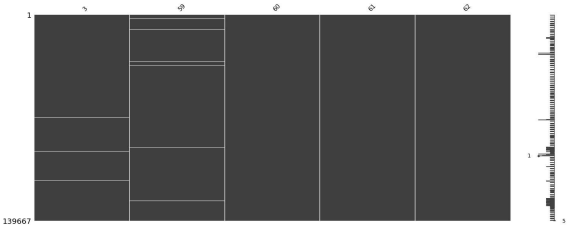
(a) Grupo Sistema Automático luego de preprocesamiento



(b) Grupo Sistema Automático ventanas aleatorias 0,16 %



(c) Grupo Varios luego de preprocesamiento



(d) Grupo Varios ventanas aleatorias 0,22 %

Figura 4.15: Parte 3: Visualización de datos faltantes (EP)

4.2.1. Comparación de metodologías

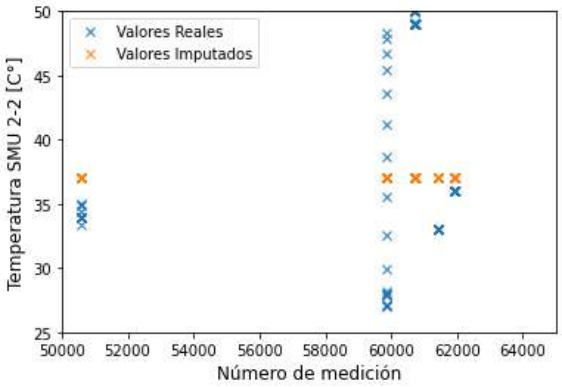
Al implementar las metodologías de imputación de datos faltantes en las nuevas bases de datos con ventanas aleatorias se consigue medir el error de imputación MSE^*, que corresponde a la función del error cuadrático medio, aplicado al valor real y el valor predicho imputado, dividido por el promedio de cada parámetro.

En la Tabla 4.1 se indica la etiqueta de los siete parámetros característicos, la metodología que obtuvo un mejor desempeño y el error conseguido con esta metodología. La elección de estos parámetros fueron los tres errores más bajos (parámetro 60, 40 y 33), tres errores más altos (parámetros 61, 10 y 9) y el parámetro con error promedio tras eliminar los seis parámetros anteriores (parámetro 54). La metodología con mayor frecuencia es *KNN* con el hiperparámetro de pesos uniformes, sin embargo, la cantidad de vecinos se ajusta al parámetro medido.

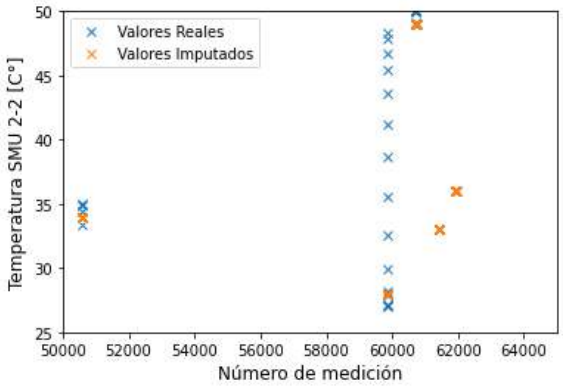
Tabla 4.1: Metodologías y errores de parámetros característicos

#	**Metodología**	MSE^*
60	Random Forest	0,001
40	KNN - 50 uniforme	0,002
33	KNN - 400 uniforme	0,003
54	MICE	0,467
61	KNN - 10 uniforme	3,094
10	KNN - 1000 uniforme	11,426
9	KNN - 1500 uniforme	22,852

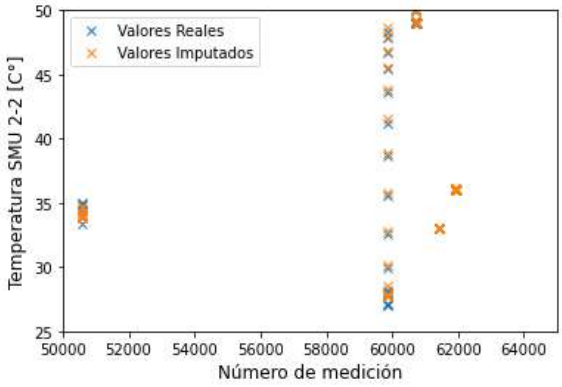
En las Figuras 4.16 y 4.17 se comparan las imputaciones del parámetro 54, que corresponde a la temperatura de SMU 2-2 del grupo de Sistema Automático, las "x" naranjas son los valores imputados y las "x" azules equivalen a los valores reales. El procedimiento de imputación se realiza para las siete metodologías mencionadas en la sección anterior, llevando a cabo un acercamiento en una cantidad de observaciones y un rango de temperatura en específico para visualizar de mejor manera las diferencias.



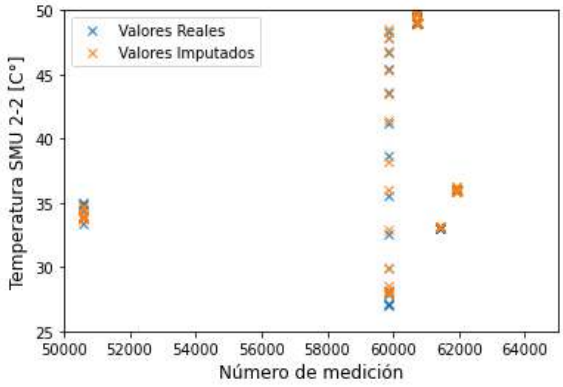
(a) Imputación con Media



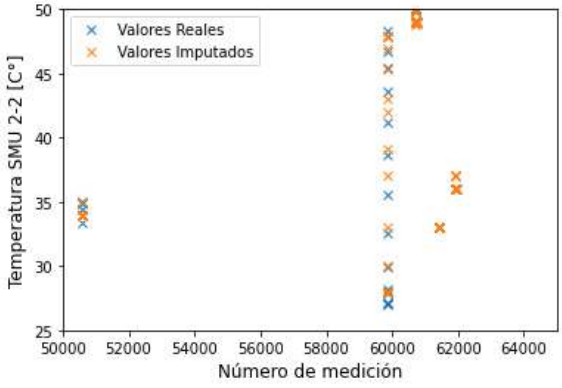
(b) Imputación con *Fill Foward*



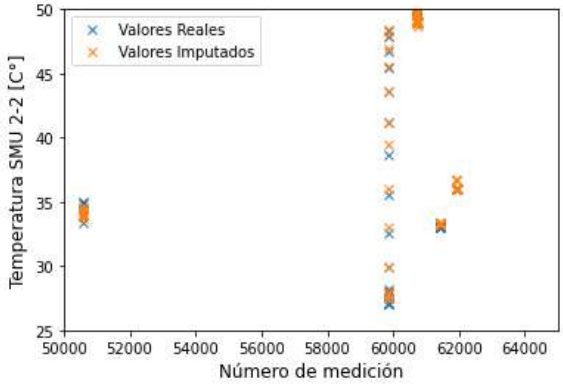
(c) Imputación con *MICE*



(d) Imputación con *KNN*

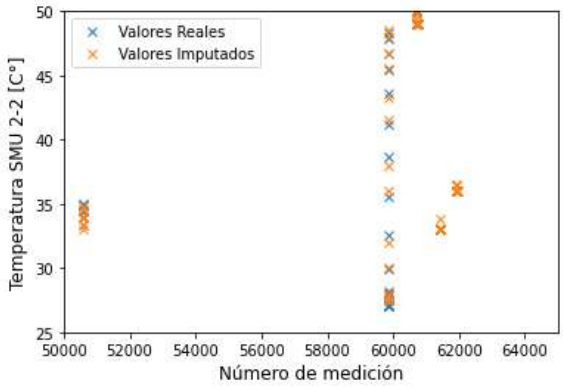


(e) Imputación con Árbol de decisión



(f) Imputación con *Random Forest*

Figura 4.16: Parte 1: Valores reales e imputados parámetro 54 (EP)



(a) Imputación con *Extremely Randomized Tree*

Figura 4.17: Parte 2: Valores reales e imputados parámetro 54 (EP)

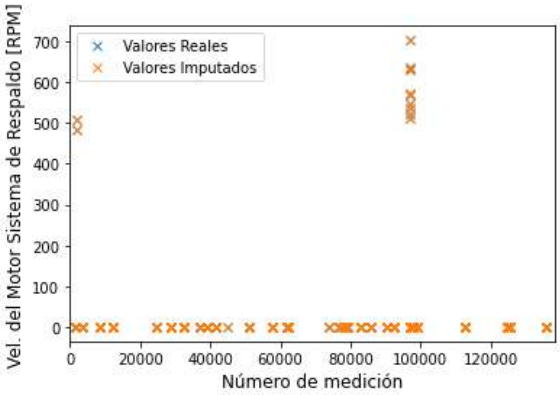
De manera gráfica, se aprecia notoriamente que la metodología de imputación por media no se ajusta a los valores reales. Por otro lado, la metodología *FillFoward* se ajusta de mejor manera que la imputación media en las observaciones distintas a las cercanas a la 60.000, sin embargo, en las observaciones cercanas a las 60.000 se aprecia que frente a una variación en los valores de una ventana de datos faltantes, la metodología se ve afectada negativamente en la predicción, imputando el mismo valor en toda la ventana.

Para las metodologías de imputación de datos restantes estas trabajan con los parámetros del grupo completo por lo que identifican el patrón de las observaciones cercanas a la 60.000, ya sea por regresiones lineales, vecinos similares o reglas de decisión de los árboles de regresión. En particular en este parámetro la mejor metodología fue *MICE*, siguiéndole *Random Forest*, *Extremelly Randomized Tree*, *Decision Tree*, *KNN*, *FillFoward* y luego imputación con media, los errores MSE^* de cada metodología son 0,467 - 0,468 - 0,469 - 0,470 - 0,483 - 0,772 - 2,487 respectivamente. A partir de los resultados se deprende que para las observaciones cercanas a la 60.000 el crecimiento fue lineal y gracias a la alta correlación con los parámetros del grupo (con 32, 33, 51, 52, 53 y 55) y un bajo porcentaje en los datos faltantes en los parámetros correlacionados en el grupo Sistema Automático, se obtiene una mejor precisión en los valores imputados.

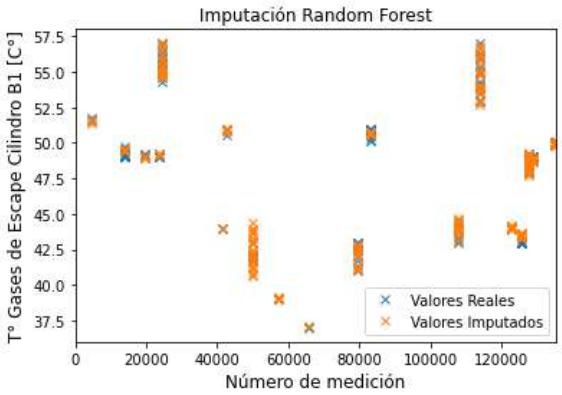
Al comparar el tiempo computacional tras aplicar las metodologías con aprendizaje de máquinas en el parámetro 54, se obtiene lo siguiente: 13,75 [s] demora ejecutar el código para la metodología *KNN*, 32,36 [s] utilizando *Extremely Randomized Tree*, 45,20[s] para la metodología *MICE*, 51,15 [s] en *Decision Tree* y 649,24 [s] al implementar *Random Forest*.

Los tiempos computacionales dependen proporcionalmente de la cantidad de parámetros de cada grupo y el porcentaje de datos faltantes que se debe completar, además de depender de la metodología implementada y los hiperparámetros seleccionados. En los ocho grupos la metodología que demora más tiempo es *Random Forest*, debido a la cantidad de árboles de decisión generados y el criterio de selección de atributos a partir del error cuadrático medio. Es necesario mencionar que el tiempo de selección de los hiperparámetros en las metodologías que utilizan árboles de decisión son considerablemente mayores que las metodologías de *MICE* o *KNN*, puesto que el modelo en ligeras variaciones de los hiperparámetros tiende a no converger o utiliza una mayor memoria RAM de la comúnmente disponible. El detalle del tiempo computacional implementado en cada metodología se indica en las Tablas C.3 y C.4.

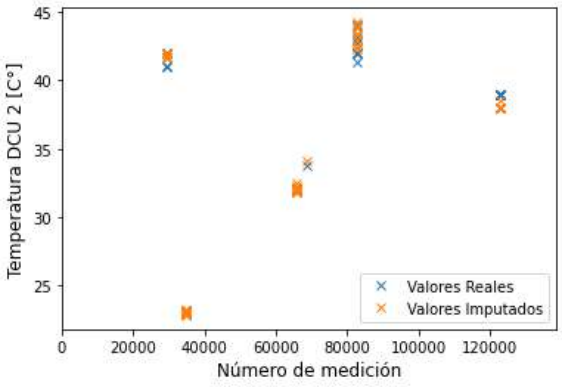
Respecto a los parámetros que obtuvieron los más bajos errores, se graficó la comparación de los valores reales e imputados, cada uno con la metodología de mejor desempeño, estos gráficos se presentan en la Figura 4.18 en orden creciente con respecto al error.



(a) Parámetro 60 - *Random Forest*



(b) Parámetro 40 - *KNN*

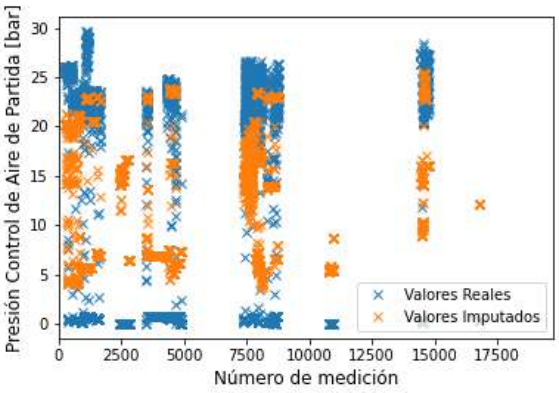


(c) Parámetro 33 - *KNN*

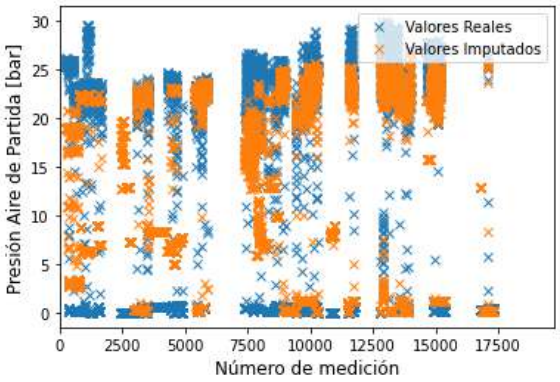
Figura 4.18: Valores reales e imputados con menores errores (EP)

El parámetro 60 corresponde a la velocidad del motor del sistema de respaldo del grupo Varios, y su metodología de imputación de datos faltantes que tiene un menor error es el modelo *Random Forest*. Para este caso, de manera gráfica, no se visualizan las "x" azules de los valores reales ya que son tapadas por las "x" naranjas de los valores imputados, esto sucede para los valores cercanos a cero y para los valores que se alejan. Por otro lado, el parámetro 40 corresponde a la temperatura de gas de cilindro B1 del grupo de Gases de Escape, y la metodología de imputación que se ajustó de mejor manera a las ventanas aleatorias fue *KNN* uniforme. En este parámetro existe una mayor variación en la ubicación de los valores faltantes y se realiza un *zoom* en el eje $\hat{y}$ para identificar en detalle el procedimiento de imputación. Por último, el parámetro 33 corresponde a la temperatura DCU 2 del grupo de Sistema Automático, y su metodología de imputación con mejor desempeño fue *KNN* uniforme. Visualmente se distinguen claramente cuatro "x" azules en la gráfica, lo cual no significa que sean cuatro observaciones imputadas de forma incorrecta debido a que se pueden sobreponer observaciones en una misma "x". La identificación de las "x" azules significa que gráficamente la predicción no fue exacta en todas las observaciones, sin embargo, la metodología de imputación de datos faltantes con *KNN* en este parámetro se acerca en la mayoría de las observaciones a los valores reales.

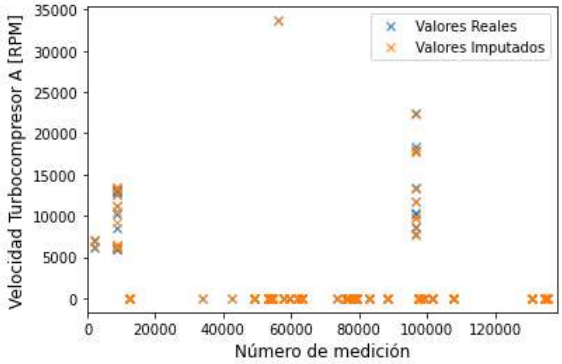
De igual manera, se graficó para los parámetros con errores más altos la comparación de los valores reales e imputados, cada uno con la metodología que logró un menor error. Los gráficos se presentan en la Figura 4.19 en orden decreciente con respecto al error.



(a) Parámetro 9 - *KNN*



(b) Parámetro 10 - *KNN*



(c) Parámetro 61 - *KNN*

Figura 4.19: Valores reales e imputados con mayores errores (EP)

El parámetro 9 corresponde a la presión de aire de control del grupo Aire, y su metodología de imputación de datos faltantes que tiene un menor error es *KNN* uniforme. A diferencia de los parámetros con errores más bajos, en la Figura 4.19 (a) se identifican claramente en varias observaciones las "x" azules de los valores reales y su discrepancia con las "x" naranjas de los valores imputados. En cuanto al parámetro 10, este corresponde a la presión de aire de partida del grupo de Aire, y la metodología de imputación que mejor se ajustó a las ventanas aleatorias fue *KNN* uniforme. Al igual que en el parámetro 9 se aprecia una clara discrepancia en algunos valores reales en comparación con los valores imputados. Por último, el parámetro 61 corresponde a la velocidad del turbocompresor del grupo Varios, y su metodología de imputación con menor error fue *KNN* uniforme. A diferencia que los parámetros 9 y 10, en esta gráfica no se observa de manera clara la discrepancia entre los valores reales e imputados.

Ahora bien, los parámetros con mejores resultados son de tres grupos diferentes: Varios, Gases de Escape y Sistema Automático, por lo que para estos tres parámetros no se atribuye el criterio del grupo al que pertenece con un buen resultado en el proceso de imputación de datos faltantes. Por otro lado, en los parámetros con más bajos resultados son de dos grupos diferentes: Aire y Varios, en el caso de Varios se tiene que el grupo también posee al mejor resultado, en consecuencia, no se desprende un grado de condicionalidad el pertenecer

a este grupo. Por el contrario, cabe destacar que, para el grupo de Aire, este tiene el mayor porcentaje de datos faltantes introducido, el cual podría ser un factor causal del peor resultado, sin embargo, este factor no es necesariamente excluyente y requiere mayores pruebas para comprobarlo. El orden decreciente de porcentaje de datos faltantes en los grupos es el siguiente: Aire (33 %), Aceite de Lubricación (24 %), Agua de Enfriamiento (7 %), Combustible (5 %), Gases de Escape (0,29 %), Cárter del Cigüeñal (0,23 %), Varios (0,22 %) y Sistema Automático (0,16 %).

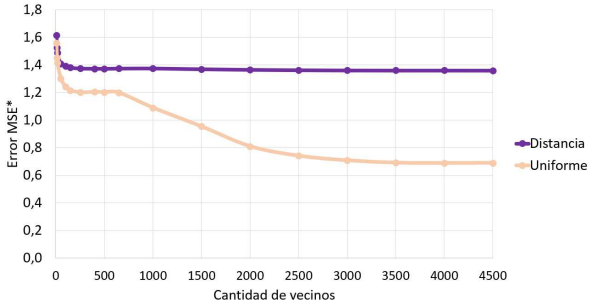
Los parámetros medidos que obtuvieron más bajos errores fueron 1 de velocidad y 2 de temperatura, en cambio para los que tuvieron más altos errores fueron 2 de presión y 1 de velocidad. A pesar de tener dos temperaturas y dos presiones en grupos diferentes, no es posible generalizar los resultados hacia todos los sensores que midan temperatura o presión debido a la baja muestra. De este modo, se ordenaron de menor a mayor los errores de todos los parámetros indicando la propiedad medida, desprendiendo que no existe relación entre el error y esta propiedad, puesto que existen menores errores en temperatura debido a que es mayor la cantidad de parámetros que miden esta propiedad y no debido a la propiedad en sí. En los Anexos C.2 se presenta en detalle el orden de los errores y las propiedades medidas.

Al ver las matrices de correlación de las Figuras 4.4 y 4.5 se identifica que el parámetro 60 posee una muy alta correlación con los parámetros 59, 61 y 62, que inclusive se aproxima a 1 en la matriz de Varios, a su vez, el parámetro 40 tiene una muy alta correlación con todos los parámetros del grupo de Gases de Escape que también se aproximan a 1, en cuanto al parámetro 33, este igualmente posee una muy alta correlación que se aproxima a 1 con los parámetros 54 y 55. Para el caso de los parámetros con más altos errores, tal como se mencionó para el parámetro 60, este se correlaciona con el 61, por lo que el parámetro 61 también tiene una alta correlación con los parámetros 59 y 62. Por otro lado, se percibe que los parámetros 9 y 10 cuentan con una muy alta correlación entre sí que también se aproxima a 1, sin embargo, no existe una correlación con los otros parámetros del grupo, esto interfiere en el resultado ya que al ver la Figura 4.14 (b) los parámetros 9 y 10 comparten significativas observaciones en donde ninguno de los parámetros tiene registros, por lo que la alta correlación existente entre los parámetros no facilita directamente la imputación de datos faltantes si no va acompañada de contar distintas observaciones sin datos. Para el caso de los parámetros de más bajos errores, ocurre simultáneamente una alta correlación y al contar con más de un parámetro correlacionado, en las Figuras 4.14 y 4.15 se identifica que no se comparten valores faltantes en las mismas observaciones para todos parámetros correlacionados a la vez.

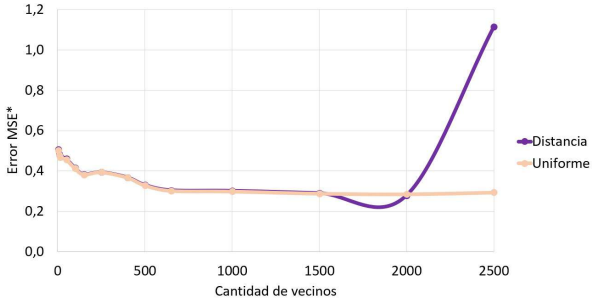
Por otro lado, a partir de los resultados, se desglosa que los coeficientes de variación no afectan en el desempeño del proceso de imputación de datos faltantes de los seis parámetros estudiados. Esto se ve reflejado ya que el coeficiente de variación del mejor resultado es 275 % y el peor es 214 %.

Continuando con la comparación de metodologías de imputación de datos faltantes, en el algoritmo *KNN* se tienen los hiperparámetros de cantidad de vecinos y el tipo de pesos que se implementa en estos vecinos. De este modo, en la Figura 4.20 se gráfica la variación de estos dos hiperparámetros para 5, 10, 15, 50, 100, 150, 250, 400, 500, 650, 1000, 1500, 2000, 2500 vecinos (para el grupo Aceite de Lubricación es hasta 4500 en intervalos de 500), acción realizada en los ocho grupos de la base de datos para el motor de babor. En los gráficos, el eje

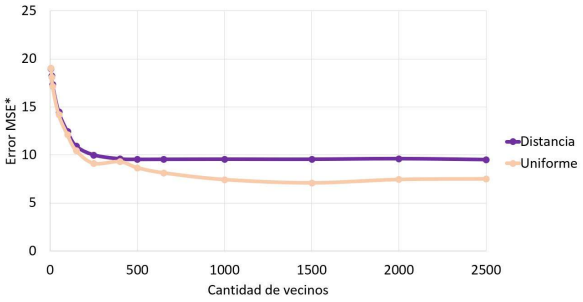
$\hat{x}$ es la cantidad de vecinos y el eje $\hat{y}$ es el promedio del error cuadrático dividido por el valor promedio de cada parámetro del grupo, a su vez, se comparan en morado el hiperparámetro distancia en los pesos y de color anaranjado para el caso uniforme. En estos gráficos se obtiene un mejor resultado cuando el error es menor por lo que se busca el mínimo.



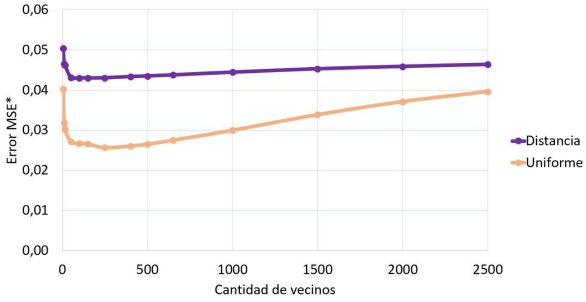
(a) Aceite de Lubricación



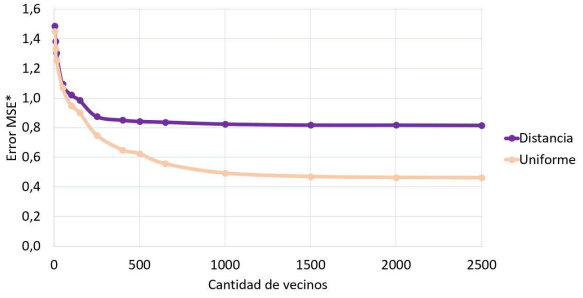
(b) Agua de Enfriamiento



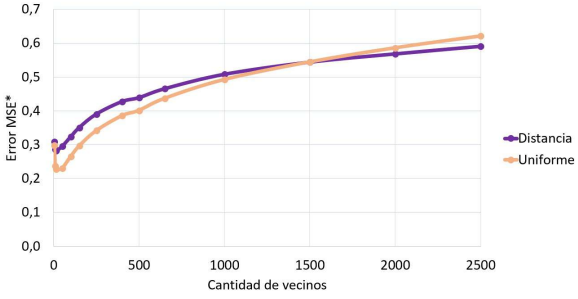
(c) Aire



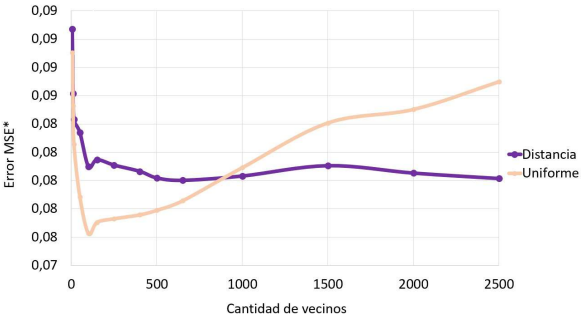
(d) Cárter del Cigüeñal



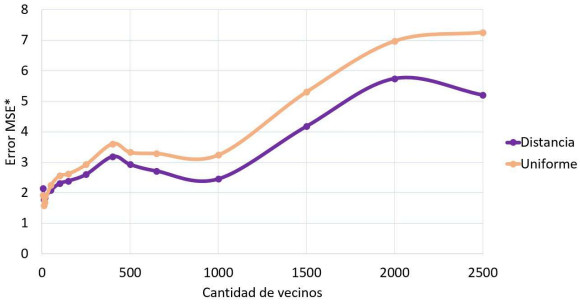
(e) Combustible



(f) Gases de Escape



(g) Sistema Automático



(h) Varios

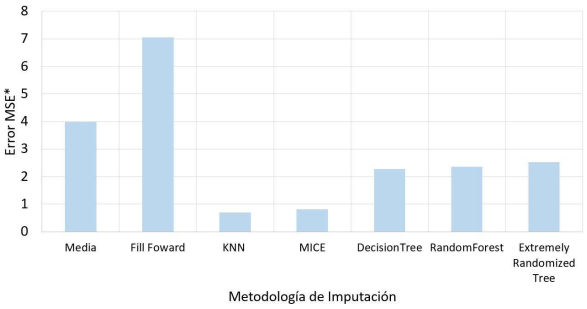
Figura 4.20: Error promedio de grupos tras imputación KNN variando hiperparámetros cantidad de vecinos y tipo de pesos (EP)

El grupo de Aceite de Lubricación obtiene mejores resultados para *KNN* uniforme, específicamente para 4000 vecinos. En la Figura 4.20 (a) se observa que con el hiperparámetro de distancia no influye la cantidad de vecinos a partir de los 15 vecinos, en cambio, con el hiperparámetro de pesos uniformes existe una tendencia a disminuir el error desde los 500 a 4000 vecinos cuando aumenta la cantidad de vecinos. Para el grupo de Agua de Enfriamiento los errores con los hiperparámetros de distancia y uniforme son similares hasta los 1500 vecinos, luego para los pesos uniformes se mantiene constante el error a medida que aumenta la cantidad de vecinos, en cambio, para los pesos con distancia se obtiene un mínimo en 2000 vecinos y luego se dispara el aumento del error sustancialmente. Continuando con el grupo de Aire, se evidencia que los pesos uniformes obtienen un mejor resultado en 1500 vecinos y al igual que en el grupo de Aceite de Lubricación a partir de una cantidad de vecinos, que en este caso aproximadamente en 1000 vecinos, no influye sustancialmente la cantidad de vecinos en la variación del error. De igual manera, el grupo de Cárter del Cigüeñal posee un menor error en los pesos uniformes, sin embargo, esta vez en una menor cantidad de vecinos de 250, luego de este mínimo el error aumenta al aumentar la cantidad de vecinos.

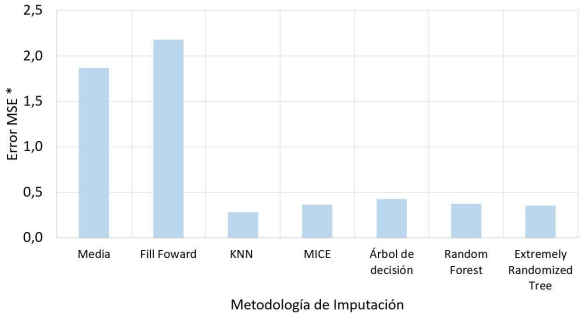
En la Figura 4.20 (e) se encuentra la comparación para la grupo de Combustible, este grupo cuenta con un mejor desempeño en los pesos uniformes y cuenta con un mínimo en 2000 vecinos. El grupo de Gases de Escape posee una tendencia distinta en el hiperparámetro de pesos en comparación con los otros grupos, en la Figura 4.20 (f) se aprecia que se alcanzan errores mínimos con una baja cantidad de vecinos y luego ambos modelos aumentan su error de manera cóncava. A su vez, en este grupo las líneas de pesos se cruzan obteniendo un mejor desempeño los pesos con distancia para cantidades de vecinos mayores a 1500 vecinos. Para el grupo de Sistema Automático también existe un entrecruzamiento en las metodologías consiguiendo un mejor desempeño el hiperparámetro de pesos uniformes hasta aproximadamente 900 vecinos y luego obtiene mejores resultados el modelo de pesos con distancia. El mínimo error de este grupo se logra en pesos uniformes con 100 vecinos. Por último, el grupo Varios a diferencia de la mayoría de los grupos cuenta con un mejor desempeño en el modelo de pesos con distancia, no obstante, el menor error se obtiene en pesos uniformes con 10 vecinos.

De este modo, en siete de los ocho grupos se logra un menor error con el hiperparámetro de pesos uniformes. En cuanto la cantidad de vecinos esta depende del grupo, ya que en tres ocasiones fue mejor de 0 a 250 y en las cinco restantes de 1500 a 4000, sin embargo, es importante señalar que el tiempo computacional de respuesta es mayor si aumenta la cantidad de vecinos del modelo. De esta manera, se deduce que la preferencia del hiperparámetro de pesos uniformes es ocasionado por la evolución de las mediciones, ya que el estado del motor debido al desgaste y las condiciones de operación varía, no repitiéndose exactamente las mismas mediciones, es por esto que imitar las características de vecinos cercanos no es lo que se busca, sino, promediar los valores existentes para tomarlos de referencia de un nuevo estado.

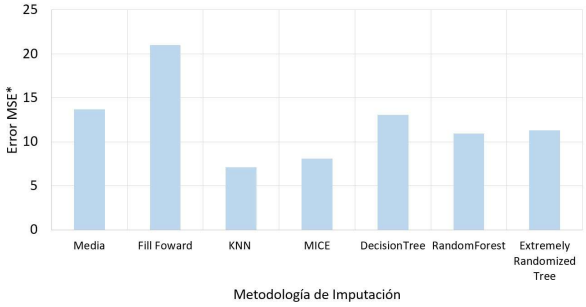
Continuando con la comparación de metodologías de imputación de datos faltantes en la Figura 4.21 se presentan los errores promedios de los parámetros para los ocho grupos de las siete metodologías implementadas. El orden de magnitud de los errores promedio varía por grupo, recordando que no interfiere la escala de los parámetros en este orden al estar dividido cada error por su valor promedio.



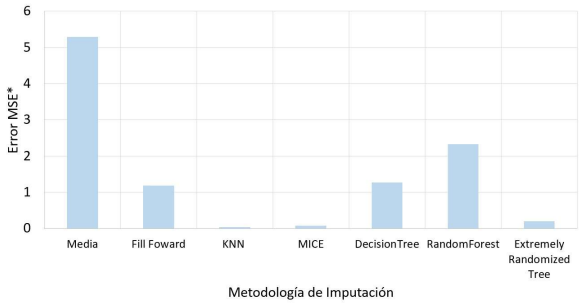
(a) Aceite de Lubricación



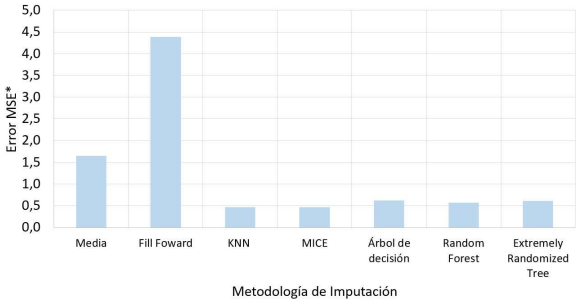
(b) Agua de Enfriamiento



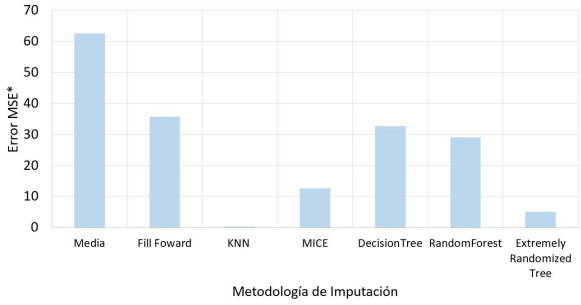
(c) Aire



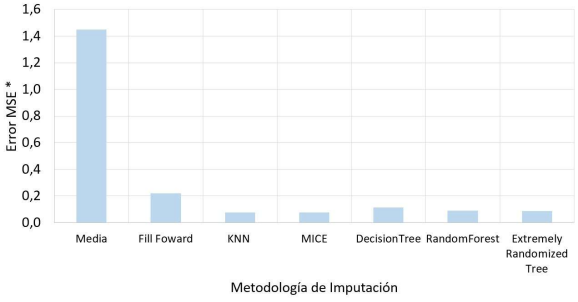
(d) Cárter del Cigüeñal



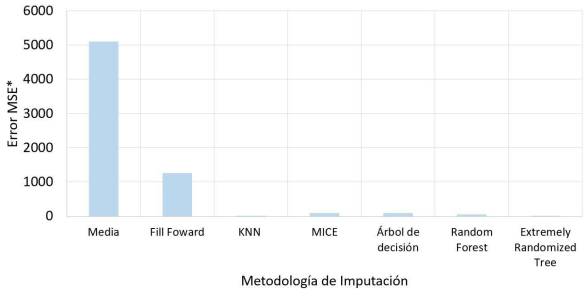
(e) Combustible



(f) Gases de Escape



(g) Sistema Automático



(h) Varios

Figura 4.21: Error promedio de grupos (EP)

A continuación, se señala para cada grupo, en primer lugar el rango aproximado en donde se encuentra el error MSE^* considerando las siete metodologías de imputación, y luego el rango cuando no se consideran las metodologías que poseen un error gráficamente mayor al estándar, mencionando el procedimiento que se descarta:

- Aceite de Lubricación: de 0 a 7, sin la metodología *FillFoward*, de 0 a 4.
- Agua de Enfriamiento: de 0 a 2,2, sin las metodologías media y *FillFoward* de 0 a 0,4.
- Aire: de 0 a 22, sin la metodología *FillFoward* de 0 a 14.
- Cárter del Cigüeñal: de 0 a 5,3, sin la metodología media de 0 a 2,4.
- Combustible: de 0 a 4,4, sin las metodologías media y *FillFoward* de 0 a 0,7.
- Gases de Escape: de 0 a 62, sin la metodología media de 0 a 36.
- Sistema Automático: de 0 a 1,6, sin las metodologías media y *FillFoward*, de 0 a 0,1.
- Varios: de 0 a 5100, sin las metodología medias y *FillFoward*, de 0 a 100.

Por consiguiente, el grupo con errores promedio más bajo es el grupo de Sistema Automático y el con mayor error promedio es el grupo de Varios, deduciendo que contar con un porcentaje de datos faltantes bajo no es un criterio determinante en la magnitud de este error, ya que ambos poseen bajos porcentajes de imputación de datos faltantes.

En los grupos de Aceite de Lubricación, Agua de Enfriamiento, Combustible, Sistema Automático y Varios las metodologías de imputación con media y *FillFoward* poseen a lo menos el 50 % más del error obtenido con las metodologías de imputación con aprendizaje de máquinas. Sin embargo, en los grupos restantes una de estas dos metodologías cuenta con valores similares con alguna metodología de aprendizaje de máquinas, inclusive para el grupo del Cárter del Cigüeñal es menor el error obtenido en *FillFoward* en comparación con las metodologías *Decision Tree* o *Random Forest*. En particular, el gran resultado de este grupo se atribuye a la alta correlación entre todos los parámetros del grupo y el bajo porcentaje de datos faltantes a imputar, lo cual también ocurre en el grupo de Gases de Escape logrando que la metodología de imputación *FillFoward* tenga un error similar a *Decision Tree*.

En cuanto a las metodologías de árboles de decisión existe una ligera tendencia a optar por el modelo *Extremelly Randomized Tree*, dejando fuera a los grupos de Aceite de Lubricación, en donde es mejor la metodología con un árbol de decisión y en el caso de los grupos de Aire y Combustible que obtiene mejores desempeños con el algoritmo *Random Forest*. Se deduce que esta tendencia se genera debido al mayor grado de aleatoriedad de la metodología *Extremelly Randomized Tree*. A su vez, se tiene que en los grupos de Aceite de Lubricación, Aire, Cárter del Cigüeñal y Gases de Escape, las metodologías de imputación de datos faltantes con árboles de decisión se aprecian alejadas de la metodología con menor error que mayoritariamente es *K-Nearest Neighbors*.

En la Figura 4.22 se indica un histograma de las metodologías que obtienen menores errores promedio en la imputación de datos faltantes. Consiguiendo el algoritmo *KNN* una mayor frecuencia, debido a que el método busca registros similares específicamente en los grupos generados, imputando valores con propiedades físicas dentro del rango permitido u observado, que es facilitado por la cantidad de observaciones en la base de datos. En cuanto a la preferencia del hiperparámetro uniforme, este se vio reflejada en los gráficos de la Figura 4.20 y explicado en la subsección anterior.

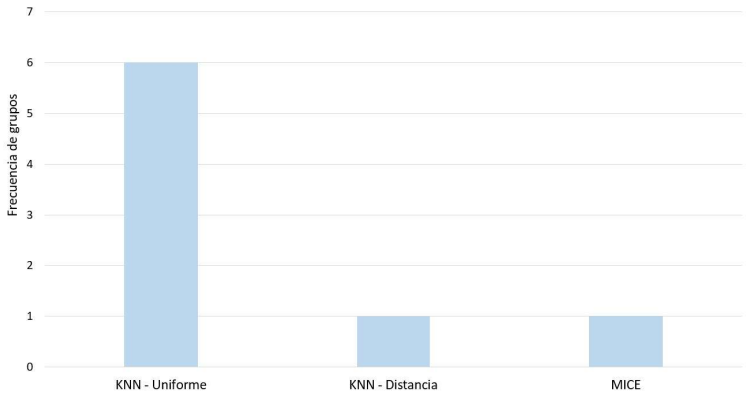
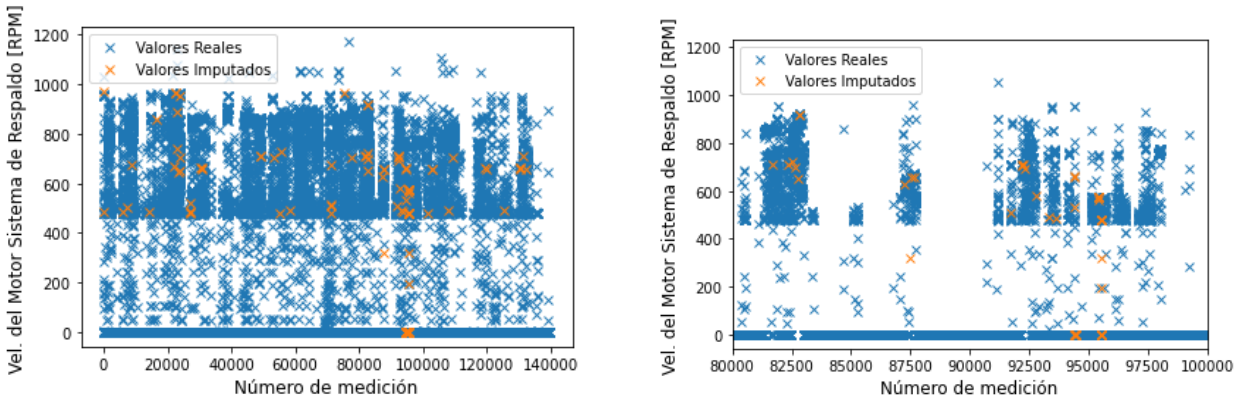


Figura 4.22: Histograma de metodologías que obtienen menores errores (EP)

4.2.2. Imputación de datos faltantes

Tras conseguir las metodologías que se ajustan de mejor manera a cada grupo para las ventanas de datos faltantes introducidas aleatoriamente, se extrapola esta selección a la base de datos original luego del preprocesamiento. En las Figuras 4.23 y 4.24 se indican las imputaciones de datos faltantes para los tres parámetros característicos mencionados en la subsección anterior, los cuales son los parámetros 60, 54 y 9. En estas figuras se hace un zoom para identificar el procedimiento de imputación de cada parámetro y cómo sus datos se ubican en el espacio.

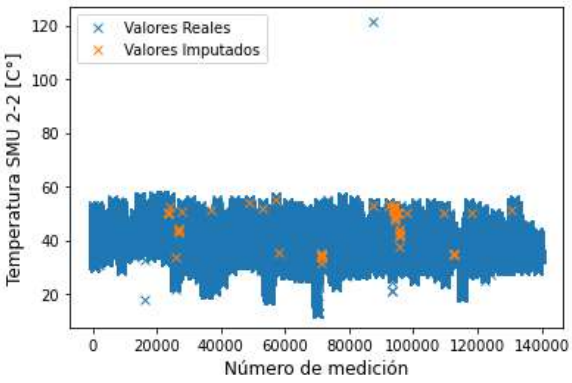


(a) Parámetro 60

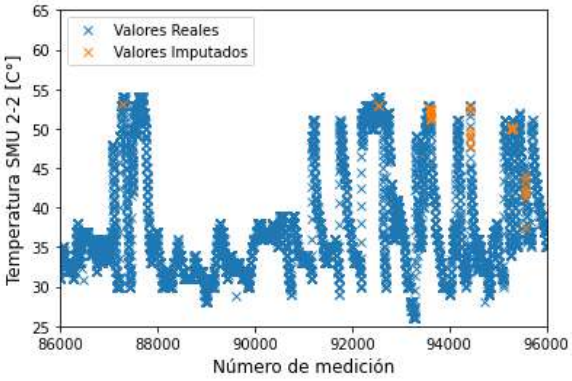
(b) Zoom de parámetro 60

Figura 4.23: Parte 1: Imputación de datos faltantes en dos parámetros extremos y uno promedio (EP)

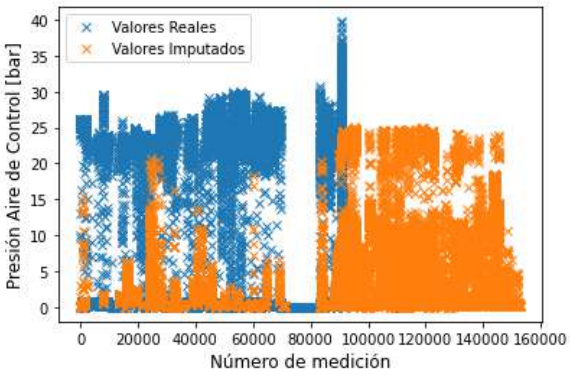
El parámetro 60 del grupo Varios se distribuye en cuatro regiones en el espacio, la primera región es un reglón cercano a cero, la segunda en una zona intermedia entre 20 a 460 [RPM] de forma dispersa, la tercera región entre 460 y 1000 [RPM] con una mayor densidad en las observaciones y la última con valores aislados. En la Figura 4.23 (b) se aprecian imputaciones de datos faltantes en las tres primeras zonas siguiendo las tendencias evidenciadas en la Figura 4.23 (a). La metodología implementada para el grupo Varios fue *K-Nearest Neighbors* con pesos uniformes y 10 vecinos.



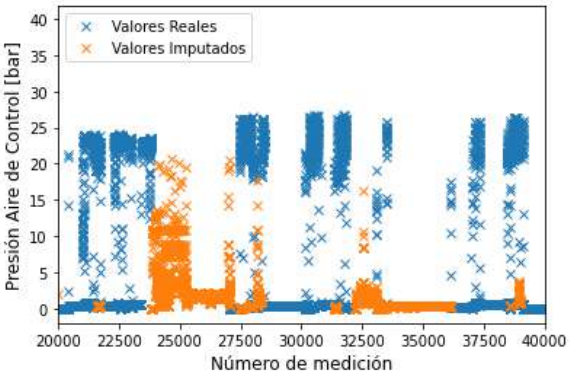
(a) Parámetro 54



(b) Zoom parámetro 54



(c) Parámetro 9



(d) Zoom parámetro 9

Figura 4.24: Parte 2: Imputación de datos faltantes en dos parámetros extremos y uno promedio (EP)

El parámetro 54 del grupo de Sistema Automático se considera un parámetro promedio de imputación y este almacena sus registros desde 0 a 55 [C°] de manera compacta, evidenciando la silueta de solo tres "x" azules en la gráfica de la Figura 4.24 (a). Al realizar un zoom en las observaciones se aprecia que los registros cuentan con un aumento y disminución oscilante, en donde la imputación sigue estas direcciones. La metodología implementada para el grupo Sistema Automático fue *MICE*.

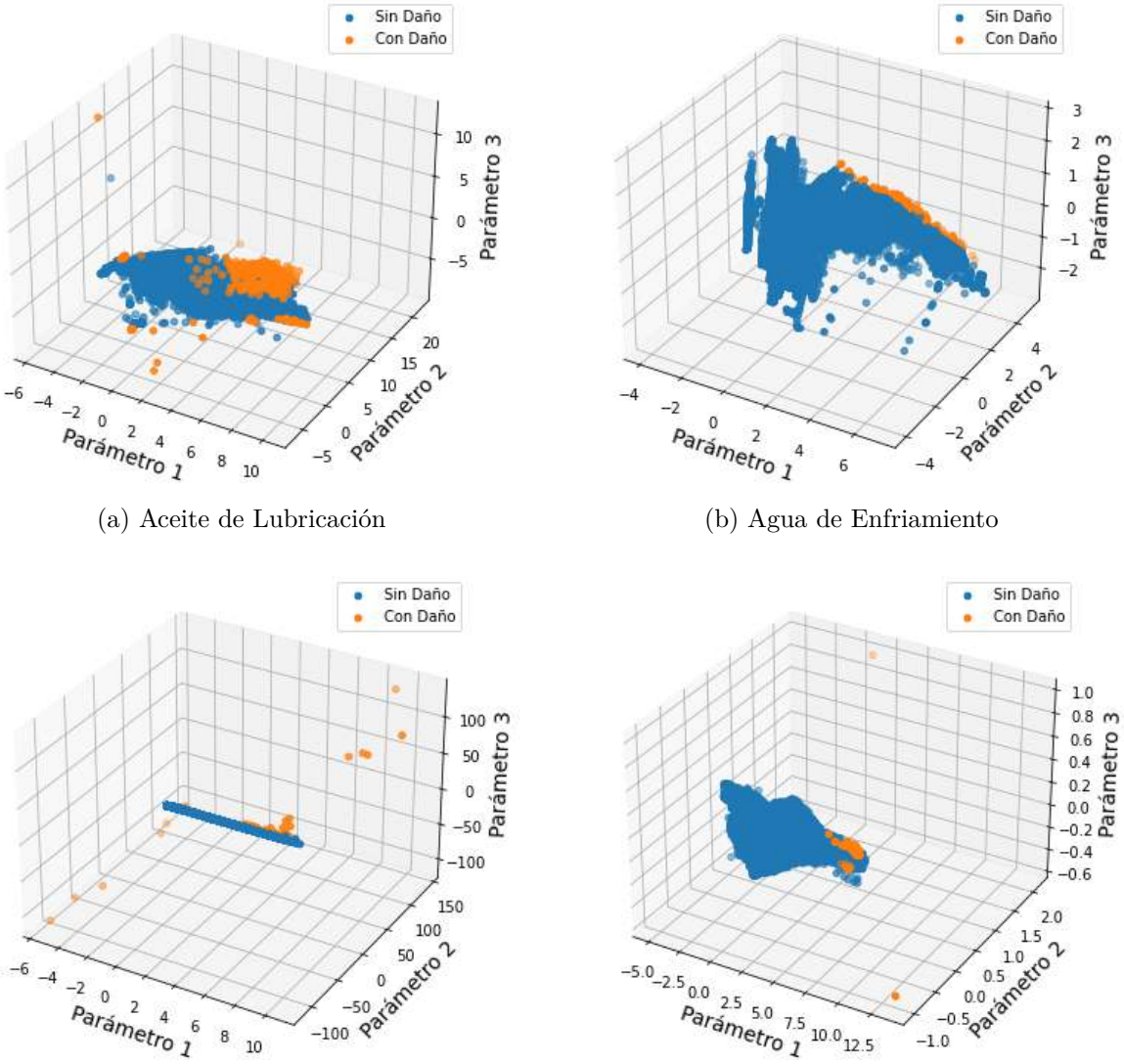
Finalmente, el parámetro 9 es uno de los parámetros con más alto valor en el porcentaje de datos faltantes, siendo 55 % su valor. En particular, para el parámetro 9, tal como se evidenció en la Figura 4.14 (a), no cuenta con registros de manera gráfica en el último tercio de las observaciones. De este modo, al realizar el proceso de imputación con la metodología *K-Nearest Neighbors* pesos uniformes y 1500 vecinos, esta se basa en la ponderación de las observaciones existentes, sin embargo, en las observaciones reales se aprecia una mayor dispersión en los registros entre 2 y 16 [bar] que la imputada con el procedimiento, además de identificar una zona cercana a cero en las observaciones imputadas, que poseen una alta densidad, similar al parámetro 60.

Cabe mencionar, que estos resultados de imputación de datos faltantes no se puede comprobar su exactitud al no contar con los registros reales.

4.3. Diagnóstico inteligente de fallas

4.3.1. Detección de novedades

En las Figuras 4.25 y 4.26 se aprecian las observaciones para cada grupo del motor de babor consideradas con daño y sin daño a través del método de detección de novedades *Elliptic Envelope* luego de la reducción de parámetros con el modelo de *Principal Component Analysis*. El algoritmo de detección de novedades fue entrenado con 8.640 observaciones normales y luego fue ejecutado con las observaciones restantes.



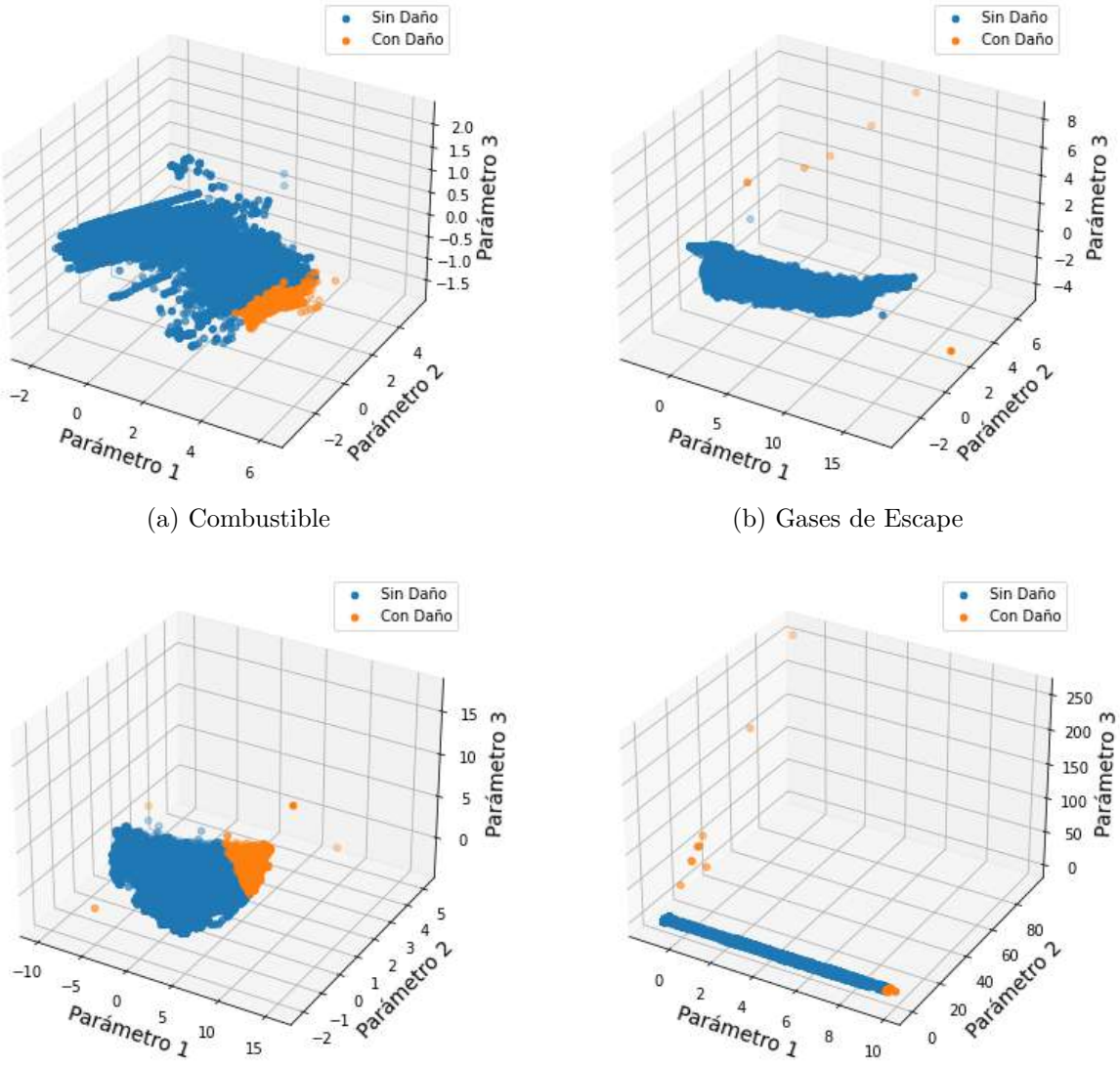
(a) Aceite de Lubricación

(b) Agua de Enfriamiento

(c) Aire

(d) Cárter del Cigüeñal

Figura 4.25: Parte 1: Detección de novedades con Elliptic Envelope - Motor Babor (EP)



(a) Combustible

(b) Gases de Escape

(c) Sistema Automático

(d) Varios

Figura 4.26: Parte 2: Detección de novedades con Elliptic Envelope - Motor Babor (EP)

De manera gráfica, se aprecia que el modelo de detección de novedades delimita ambas zonas, con daño y sin daño, en regiones mayoritariamente continuas y cercanas entre sí para los grupos de Aceite de Lubricación, Agua de Enfriamiento, Cárter del Cigüeñal, Combustible y Sistema Automático. En contraste, en los grupos de Gases de Escape, Varios y Aire, los valores con daños se observan de forma aislada a los valores sin daño. Debido a esto, se deduce que en los grupos de zonas continuas las observaciones identificadas con daño son casos límites de fallas graves, es decir, todavía no son de este tipo de falla, sino que se están desarrollando y aumentando progresivamente su gravedad. En cambio, cuando los identificados con daño son valores aislados del grupo, la falla grave ya está presente en el sistema.

4.3.2. Efecto de imputación de datos

En el trabajo de investigación se denomina procedimiento 1, al proceso de preprocesamiento, imputación de datos faltantes y detección de novedades, señalado en el capítulo de metodología. Por otro lado, al procedimiento 2 también señalado en el capítulo de metodología, considera un preprocesamiento que elimina los parámetros con más de 20 % de datos faltantes y las observaciones sin registros en algún parámetro, además de realizar la detección de novedades del procedimiento 1. En el procedimiento 2, la cantidad de parámetros se reduce sustancialmente, quedando inclusive sistemas o grupos sin monitorear. En la Tabla 4.2 se indica la cantidad de parámetros para ambos procedimientos.

Tabla 4.2: Cantidad de parámetros por procedimiento

	Procedimiento 1	Procedimiento 2
Cantidad de parámetros	58	9

En la Figura 4.27 se indica la cantidad de observaciones para el caso inicial en contraste a los procedimientos 1 y 2 para ambos motores en los ocho grupos.

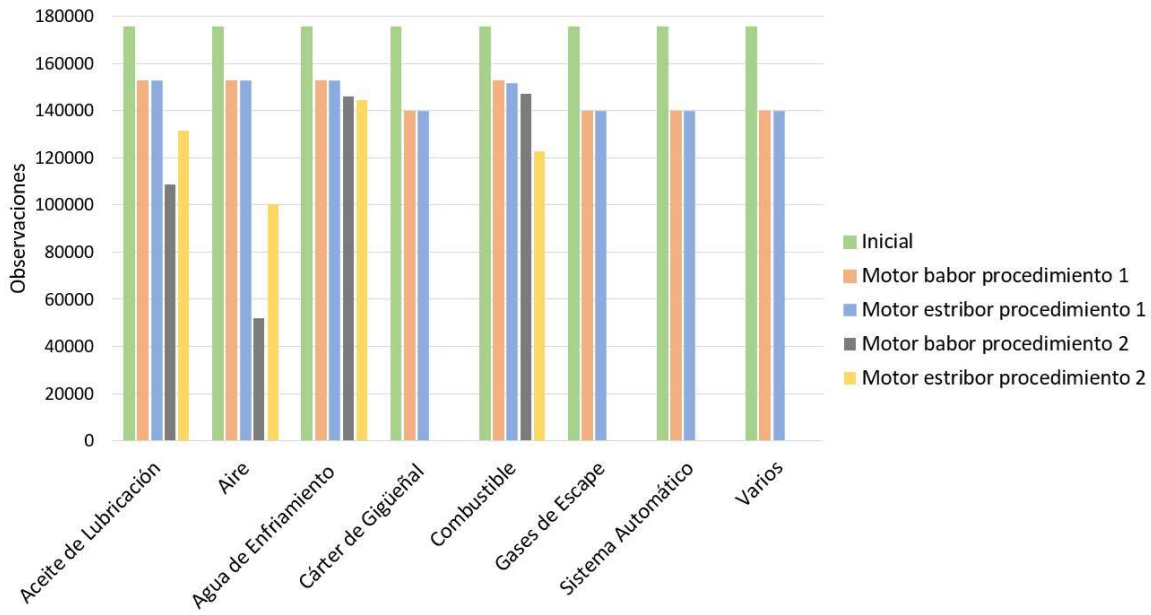


Figura 4.27: Cantidad de observaciones en distintos procedimientos (EP)

De este modo, se identifica que en los grupos Cárter del Cigüeñal, Gases de Escape, Sistema Automático y Varios no cuentan con observaciones en el segundo procedimiento y en los grupos restantes la reducción más considerable de observaciones es para el grupo de Aire en el motor de babor que pasa de 175.604 a 51.982 observaciones, desprendiendo que las herramientas del preprocesamiento e imputación de datos faltantes realizado permite estudiar todos los sistemas sin perder un porcentaje considerable de información.

En las Figuras 4.28 y 4.29 se indican la cantidad y porcentaje de anomalías encontradas al realizar la detección de novedades en los dos procedimientos.

	Cantidad de anomalías motor babor		Porcentaje de anomalías motor babor	
Grupo	P1	P2	P1	P2
Aceite de Lubricación	678	22	0.444%	0.020%
Aire	36	23	0.024%	0.044%
Agua de Enfriamiento	5005	40	3.279%	0.027%
Cárter de Gigüeñal	94	-	0.067%	-
Combustible	23	23	0.015%	0.016%
Gases de Escape	109	-	0.078%	-
Sistema Automático	6783	-	4.857%	-
Varios	23	-	0.016%	-

Figura 4.28: Detalle de anomalías - Motor Babor

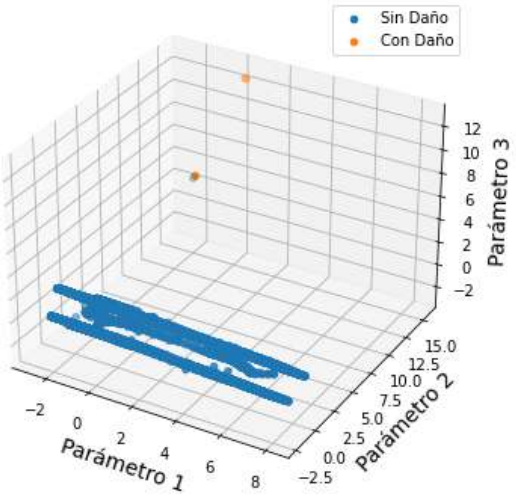
	Cantidad de anomalías motor estribor		Porcentaje de anomalías motor estribor	
Grupo	P1	P2	P1	P2
Aceite de Lubricación	3	3	0.002%	0.002%
Aire	1177	1933	0.771%	1.928%
Agua de Enfriamiento	1442	1071	0.945%	0.741%
Cárter de Gigüeñal	32	-	0.023%	-
Combustible	476	850	0.314%	0.692%
Gases de Escape	12	-	0.009%	-
Sistema Automático	2692	-	1.926%	-
Varios	475	-	0.340%	-

Figura 4.29: Detalle de anomalías - Motor Estribor

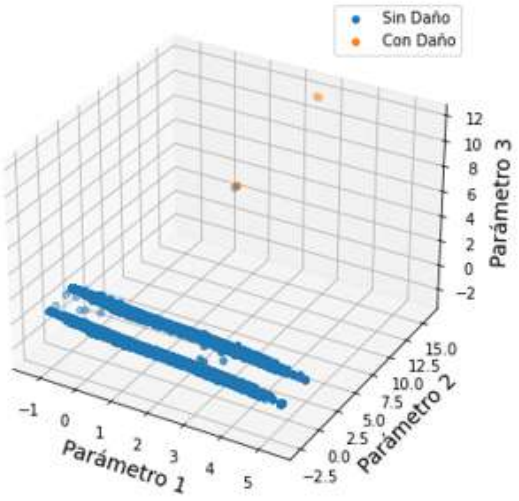
A partir de estos resultados, se identifica que para ambos motores el grupo de Sistema Automático es el que cuenta con más anomalías para el primer procedimiento. Para el caso del motor de babor se observa una menor cantidad de anomalías identificadas en el procedimiento 2 y en cuanto al porcentaje de anomalías con respecto a las observaciones totales este no posee una tendencia clara en el motor de babor, ya que en ocasiones es mayor y en otras menor. Por el contrario, en el motor de estribor ocurre una situación diferente, para el grupo de Aceite de Lubricación se encuentra la misma y baja cantidad de anomalías en ambos procedimientos y en los grupos de Aire y Combustible se encuentran más anomalías en el caso del segundo procedimiento. Ahora bien, con respecto al porcentaje de anomalías, para el segundo procedimiento del motor de estribor este es mayor en tres de cuatro grupos.

Debido a que se entrena con un conjunto diferente en ambos procedimientos, cambia la cantidad de parámetros y observaciones, identificando y dando importancia a observaciones diferentes en cada procedimiento, de este modo, cambian los resultados de detección de anomalías. Dejando propuesto el evaluar cuál procedimiento posee mayores aciertos en su predicción y si es correcto el supuesto de entrenar con dos meses datos normales.

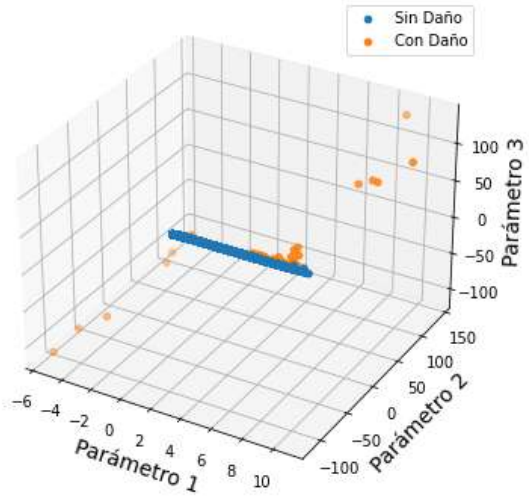
Finalmente, en la Figura 4.30 se compara la detección de novedades para los grupos de Aceite de Lubricación del motor de estribor y el grupo de Aire para ambos motores.



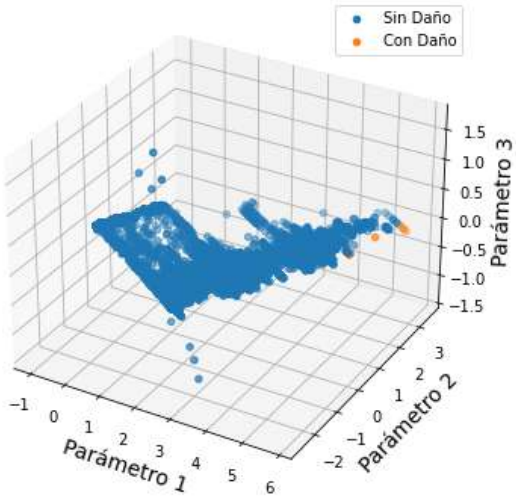
(a) Aceite de Lubricación - Motor Estribor P1



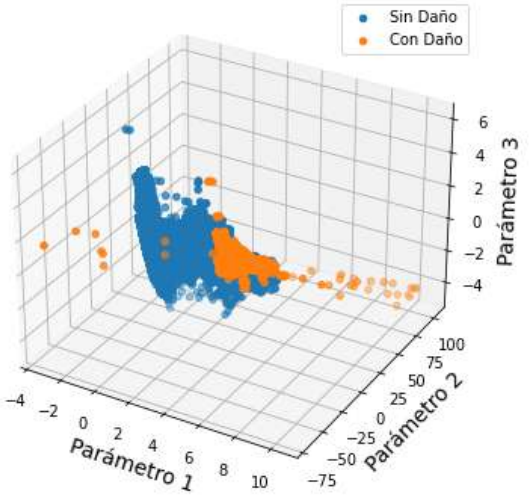
(b) Aceite de Lubricación - Motor Estribor P2



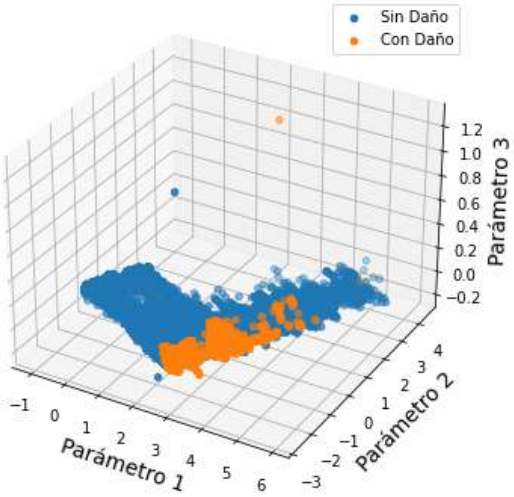
(c) Aire - Motor Babor P1



(d) Aire - Motor Babor P2



(e) Aire - Motor Estribor P1



(f) Aire - Motor Estribor P2

Figura 4.30: Comparación de procedimiento en detección (EP)

A partir de las Figuras 4.30 (a) y (b) se evidencia una alta similitud entre la detección del Aceite de Lubricación para el motor de estribor, la razón de esta similitud es debido a que en ambos procedimientos identifican la misma cantidad de anomalías y es altamente probable que se definieron las mismas anomalías.

En cambio, en los grupos de Aire no existe similitud en ninguno de los casos. Esto se debe puesto que al tener distintos parámetros luego del preprocesamiento se generan nuevos parámetros característicos en la etapa del diagnóstico inteligente de fallas cuando se implementa el algoritmo *PCA*, en donde inclusive, podrían ya no estar las observaciones que anteriormente se consideraron anómalas.

Capítulo 5

Conclusiones

A modo de síntesis, en este trabajo se implementó una base de datos que contiene mediciones de 62 sensores de dos motores marinos diésel con datos faltantes y la investigación realizada estudia el manejo de aquella falta. Para este fin, se definen ocho grupos: Aceite de Lubricación, Agua de Enfriamiento, Aire, Cárter del Cigüeñal, Combustible, Gases de Escape, Sistema Automático y Varios, los cuales poseen una alta correlación promedio entre sus parámetros. Posteriormente, se evalúan siete metodologías de imputación de datos faltantes, en donde la imputación de datos faltantes es un procedimiento de remplazo de los valores perdidos por valores calculados a partir de la base de datos.

Los resultados evidencian mejores desempeños en los algoritmos de aprendizaje de máquinas que en los modelos estadísticos, en particular, al comparar el resultado de imputación de datos faltantes en seis parámetros se desglosa una dependencia conjunta de que en los mejores resultados debe existir una alta correlación con más de un parámetro del grupo, además de contar con un bajo porcentaje de datos faltantes de forma simultánea entre los parámetros correlacionados y el estudiado. A su vez, se descarta una relación directa y exclusiva en el desempeño de los resultados por la pertenencia de un grupo, ni tampoco debido al coeficiente de variación.

La metodología con mejor resultado promedio en los grupos es *K Nearest Neighbor* con el hiperparámetro de pesos uniformes. Se concluye que el mejor desempeño es debido a que el algoritmo no busca replicas exactas, dado la evolución de las mediciones por el desgaste y condiciones de operación, sumado a que se la metodología *KNN* imputa promedios de registros similares posibles, ya que son combinaciones de valores con propiedades físicas dentro del rango permitido. Sin embargo, cabe mencionar que la selección del hiperparámetro de cantidad de vecinos varía en cada grupo debido a que se ajusta a cada registro.

En cuanto a las otras metodologías de imputación de datos faltantes se obtiene que el modelo *MICE* alcanza mejores resultados que *KNN* cuando el parámetro posee un aumento o disminución lineal a lo largo del tiempo, tal como se aprecia en las observaciones cercanas a la 60.000 del parámetro 54. A su vez, se consigue una ligera preferencia en el modelo *Extremelly Randomized Tree* al compararlo con las otras metodologías de árboles de decisión, en donde se deduce que esta preferencia se genera debido al mayor grado de aleatoriedad de la metodología.

En cuanto al diagnóstico inteligente de fallas, se tiene que las fallas identificadas que se agrupan de forma continua con el resto de los datos, la falla todavía no es grave, en cambio cuando está en valores aislados la falla ya se ha desarrollado, al crecer la discrepancia con los valores normales.

Por otro lado, tras comparar los procedimientos con imputación de datos faltantes y sin imputación de datos faltantes eliminando las observaciones sin datos, se consigue una reducción de parámetros sustancial en el caso de no imputar los datos faltantes, no estudiando todos los sistemas del motor. De este modo, la imputación de datos faltantes es una herramienta poderosa en el preprocesamiento de los datos para no eliminar una gran cantidad de observaciones sin registros en bases de datos con alto porcentaje de datos faltantes.

Al comparar los diagnósticos inteligentes de fallas en ambos procedimientos el grupo con más anomalías es el Sistema Automático, sin embargo, los resultados difieren porque cambia la cantidad de observaciones y parámetros, enfocando a distintas observaciones el comportamiento anómalo.

De este modo, se cumple el objetivo general del trabajo de título el cual es comparar metodologías de imputación de datos faltantes en una base de datos de dos motores diésel marinos y evaluar el impacto del proceso de imputación en el diagnóstico inteligente de fallas, ya que se obtiene la metodología de mejor desempeño y se evalúa el impacto en criterios de cantidad de observaciones, parámetros y anomalías, entre ambos procedimientos.

5.1. Trabajos futuros

A partir del trabajo realizado, se revelan posibles investigaciones futuras, entre ellas se destacan cuatro.

La primera es aplicar a nuevas bases de datos la metodología implementada en la comparación de los modelos de imputación de datos faltantes, para verificar si coincide el algoritmo *K Nearest Neighbor* con la metodología de mejor desempeño. Además, de estudiar si incide el tamaño de base de datos en los procedimientos de imputación de datos faltantes.

En segundo lugar, la investigación permite continuar con el diagnóstico inteligente de fallas identificando la gravedad de las fallas y en qué sistemas es necesario solucionarlas para evitar una falla irreversible en los motores.

En tercer lugar, este análisis plantea una evaluación y comparación de los procedimientos realizados, con imputación o sin imputación, en cuanto a cuál posee mayores aciertos en su predicción y si es correcto el supuesto de entrenar con dos meses datos normales.

Por último, se proponen variaciones en los porcentajes de corte de este trabajo, por ejemplo, considerar un porcentaje menor a 80 % en el estudio de parámetros para el procedimiento con imputación de datos faltantes y/o considerar un porcentaje mayor a 20 % de porcentaje de corte para el procedimiento sin imputación de datos.

Bibliografía

[1] "Supervised, unsupervised and semi-supervised learning." Enjoy Algorithms, https://www.enjoyalgorithms.com/blogs/supervised-unsupervised-and-semisupervised-learning.

[2] Gonzalez, F., "Machine learning: La base que tenes que tener." Somos PNT, https://somospnt.com/blog/53-introduccion-a-machine-learning.

[3] Carmona, E., "Principales enfoques del aprendizaje automático." Kared IT Solutions, https://karedit.com.mx/principales-enfoques-del-aprendizaje-automatico/.

[4] Gonzalez, J. L., "Tipos de aprendizaje automático." Medium - SoldAI., https://medium.com/soldai/tipos-de-aprendizaje-autom\%C3\%A1tico-6413e3c615e2.

[5] Little, R. J., and D. B. Rubin, "Statistical Analysis with Missing Data," New York, NY: John Wiley and Sons, 1987.

[6] Alai, L., Taleb, B., Salgado, D., Rosa, E., y Alonso, R., "Imputación de datos mediante Random Forest," 2021.

[7] Radío, G. R., "Los valores perdidos en el muestreo de poblaciones finitas. Técnicas de imputación.," pp. pp. 293–298, 2017.

[8] Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., y Franco, L., "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," Artificial Intelligence in Medicine, vol. 50, pp. 105–115, 2010, doi:10.1016/j.artmed.2010.05.002.

[9] Kim, B., Yuvaraj, N., Preethaa, K. R. S., Hu, G., y Lee, D. E., "Wind-induced pressure prediction on tall buildings using generative adversarial imputation network," Sensors, vol. 21, 2021, doi:10.3390/s21072515.

[10] "1.10. decision trees." Scikit-Learn, https://scikit-learn.org/stable/modules/tree.html.

[11] "Random forest machine learning in r, python and sql - part 1." Toad World Blog, 2018, https://blog.toadworld.com/2018/08/31/random-forest-machine-learning-in-r-python-and-sql-part-1.

[12] Geurts, P., Ernst, D., y Wehenkel, L., "Extremely randomized trees," Machine Learning, vol. 63, no. 1, pp. 3–42, 2006, doi:10.1007/s10994-006-6226-1.

[13] "2.7.3.1. fitting an elliptic envelope." Scikit-Learn, https://scikit-learn.org/stable/modules/outlier_detection.html\#outlier-detection.

[14] Afshine Amidi y Shervine Amidi, "Reducción de la dimensionalidad, análisis de componentes principales.", https://stanford.edu/~shervine/l/es/teaching/cs-229/hoja-referencia-aprendizaje-no-supervisado.

[15] Latarche, M., Theory and general principles. Elsevier, 2021, doi:10.1016/b978-0-08-10

2748-6.00002-5.

[16] Donaire, D. L., "El pistón, corazón del motor: qué es, función, partes, características, precio.." Actualidad Motor, https://www.actualidadmotor.com/el-piston-corazon-del-motor/.

[17] Donaire, D. L., "La biela: qué es, qué partes tiene, tipos, función y materiales.." Actualidad Motor, https://www.actualidadmotor.com/la-biela-partes-y-funcin/.

[18] "Averías frecuentes del turbo y cómo identificarlas en tu coche.." Ro-des.com, https://www.ro-des.com/mecanica/averias-del-turbo-como-identificarlas-en-el-coche/.

[19] "Rectificado de cigüeñales.." Indumotor, https://indumotor.cl/rectificado-de-ciguenales/.

[20] "Pistón con una parte de la biela de un motor de automóvil aislado en blanco.." Freepik, https://www.freepik.es/vector-premium/piston-parte-biela-motor-automovil-aislado-blanco_15430560.htm.

[21] Den., "Que es árbol de levas: función, tipos, síntomas.." Blog del AUTODOC CLUB, https://club.autodoc.es/magazin/que-es-arbol-de-levas-funcion-tipos-sintomas.

[22] "Así funcionan los inyectores diésel." Infotaller, https://www.infotaller.tv/electromecanica/asi-funcionan-inyectores-diesel_0_1306969317.html.

[23] Navarrete, J., "El rol de las válvulas del motor.." Actualidad Motor, https://www.actualidadmotor.com/el-rol-de-las-valvulas-del-motor/.

[24] Donaire, D. L., "La culata del motor." Actualidad Motor, https://www.actualidadmotor.com/la-culata-del-motor/.

[25] Donaire, D. L., "El bloque motor: qué es, de qué está hecho, partes, tipos, fabricación.." Actualidad Motor, https://www.actualidadmotor.com/el-bloque-motor-y-la-culata/.

[26] Mecafenix, I., "¿qué es el carter automotriz?." Ingeniería Mecafenix, https://www.ingmecafenix.com/automotriz/el-carter/.

[27] P., M., Construcción y Manejo de los Motores Diésel Marinos y Estacionarios, cap. Motores Marinos. Editorial Gustavo Gili, S.A., 1960.

[28] "Introduction product guide w26-1/2013 iii wärtsilä 26-product guide introduction," 2018, https://www.wartsila.com/docs/default-source/product-files/engines/ms-engine/product-guide-o-e-w26.pdf?utm_source=engines\&utm_medium=dieselengines\&utm_term=w26\&utm_content=productguide\&utm_campaign=mp-engines-and-generating-sets-brochures.

[29] "Cojinete.." Hello Auto, https://helloauto.com/glosario/cojinete.

[30] SpA, W. I., "CODE LIST MODBUS 26A2 MAIN ENGINE (VEE)," vol. 1, pp. 1–5, 2005.

[31] Ibujes, M. O. S., "Coeficiente de correlación de karl pearson.." Monografias.com, https://www.monografias.com/trabajos85/coeficiente-correlacion-karl-pearson/coeficiente-correlacion-karl-pearson.

[32] "6.3.1.1. scaling features to a range." Scikit-Learn, https://scikit-learn.org/stable/modules/preprocessing.html\#preprocessing-scaler.

Anexo

Anexo A. Base de datos

A.1. Parámetros

En las Tablas A.1 y A.2 se indican los grupos y parámetros de la base de datos, luego del proceso de análisis de datos.

Tabla A.1: Parte 1: Parámetros de medición y subáreas

Grupo	**#**	**Parámetro de medición**
Aceite de Lubricación	**2**	Nivel de Cárter Húmedo Aceite de Lubricación
	4	Presión Aceite de Lubricación PT201
	5	Presión Aceite de Lubricación PT241
	12	Presión Diferencial del Filtro de Aceite de Lubricación PDY113
	13	Presión Diferencial del Filtro de Aceite de Lubricación PDY243
	15	Temperatura Aceite de Lubricación
	16	Temperatura Aceite de Lubricación Turbo Compresor A
	17	Temperatura Aceite de Lubricación Turbo Compresor B
Agua de Enfriamiento	**6**	Presión Agua de Enfriamiento
	7	Presión Agua de Enfriamiento PT401
	18	Temperatura Agua de Enfriamiento TE401
	19	Temperatura Agua de Enfriamiento TE402
	20	Temperatura Agua de Enfriamiento TE451
Aire	**1**	Válvula de descarga de aire
	9	Presión de Control de Aire de Partida
	10	Presión de Aire de Partida
	8	Presión de Aire de Carga
	21	Temperatura Aire de Carga TE601
Cárter del Cigüeñal	**23**	Temperatura de Cojinete Principal 0
	24	Temperatura de Cojinete Principal 1
	25	Temperatura de Cojinete Principal 2
	26	Temperatura de Cojinete Principal 3
	27	Temperatura de Cojinete Principal 4
	28	Temperatura de Cojinete Principal 5
	29	Temperatura de Cojinete Principal 6
	30	Temperatura de Cojinete Principal 7

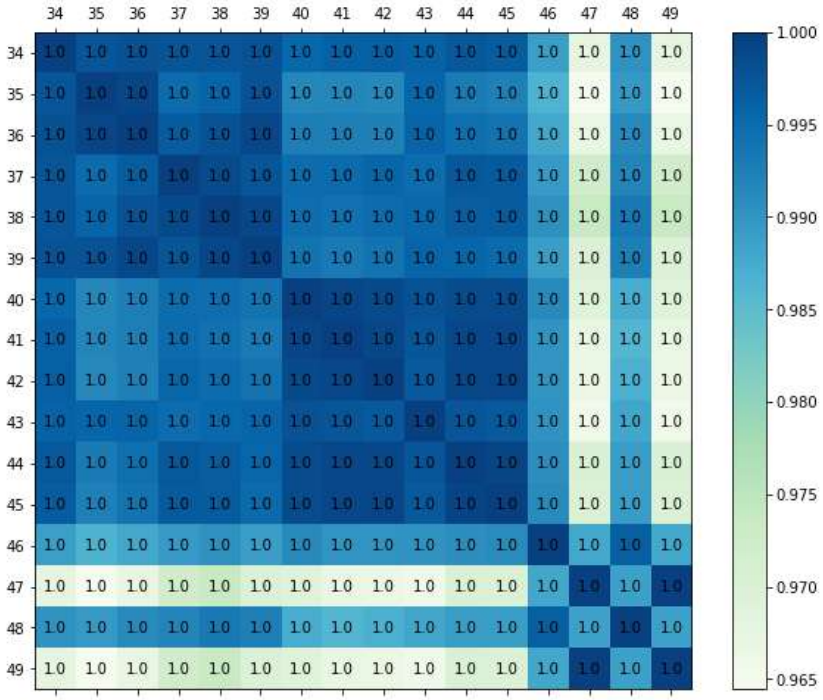
Tabla A.2: Parte 2: Parámetros de medición y subáreas

Grupo	#	Parámetro de medición
Combustible	**11**	Presión de Combustible PT101
	14	Presión de Filtro de Seguridad de Combustible
	31	Temperatura de Combustible TE101
Gases de Escape	**34**	Temperatura Gas de Escape Cilindro A1
	35	Temperatura Gas de Escape Cilindro A2
	36	Temperatura Gas de Escape Cilindro A3
	37	Temperatura Gas de Escape Cilindro A4
	38	Temperatura Gas de Escape Cilindro A5
	39	Temperatura Gas de Escape Cilindro A6
	40	Temperatura Gas de Escape Cilindro B1
	41	Temperatura Gas de Escape Cilindro B2
	42	Temperatura Gas de Escape Cilindro B3
	43	Temperatura Gas de Escape Cilindro B4
	44	Temperatura Gas de Escape Cilindro B5
	45	Temperatura Gas de Escape Cilindro B6
	46	Temperatura Gas de entrada a Turbo Compresor A
	47	Temperatura Gas de salida a Turbo Compresor A
	48	Temperatura Gas de entrada a Turbo Compresor B
	49	Temperatura Gas de salida a Turbo Compresor B
Sistema Automático	**32**	Temperatura DCU 1
	33	Temperatura DCU 2
	50	Temperatura MCU
	51	Temperatura SMU 1-2
	52	Temperatura SMU 1-3
	53	Temperatura SMU 1-4
	54	Temperatura SMU 2-2
	55	Temperatura SMU 2-3
Varios	**3**	Posición de la Rejilla del Combustible
	59	Velocidad del Motor
	60	Velocidad del Sistema de BackUp del Motor
	61	Velocidad Turbo Compresor A
	62	Velocidad Turbo Compresor B

Anexo B. Preprocesamiento

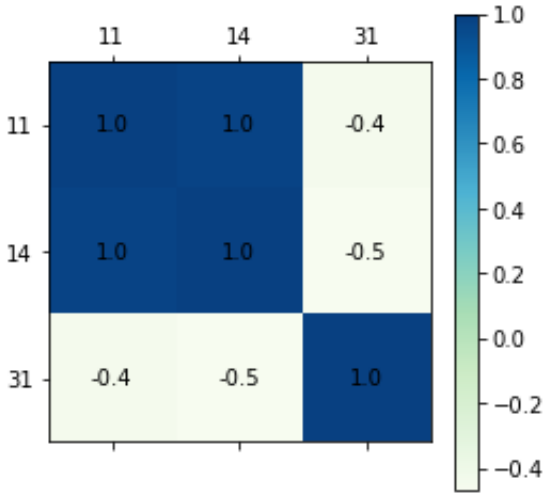
B.1. Matrices de correlación del motor de estribor

A continuación, se presentan las matrices de correlación de cada grupo del motor de estribor, las cuales son similares al motor de babor.



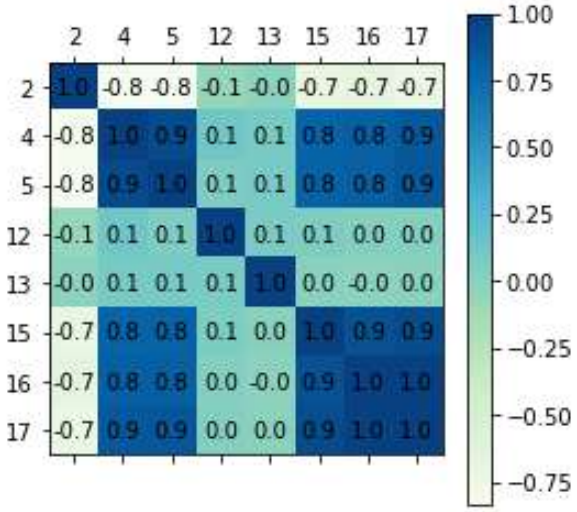
(a) Gases de Escape

Figura B.1: Parte 1: Matriz de correlación entre parámetros en el motor de estribor (EP)

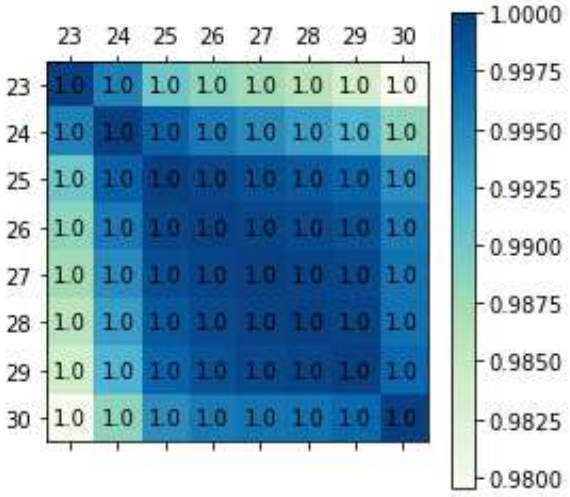


(a) Combustible

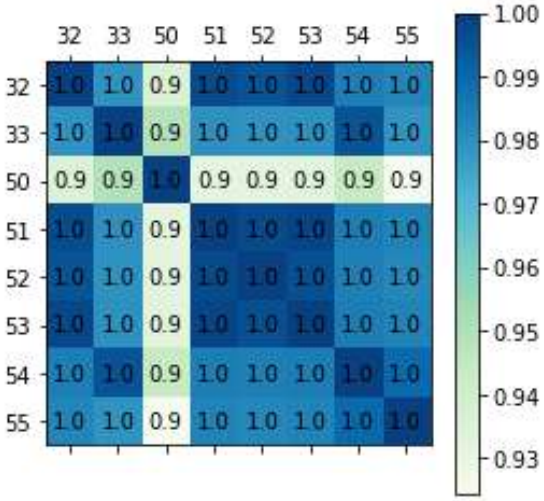
Figura B.2: Parte 2: Matriz de correlación entre parámetros en el motor de estribor (EP)



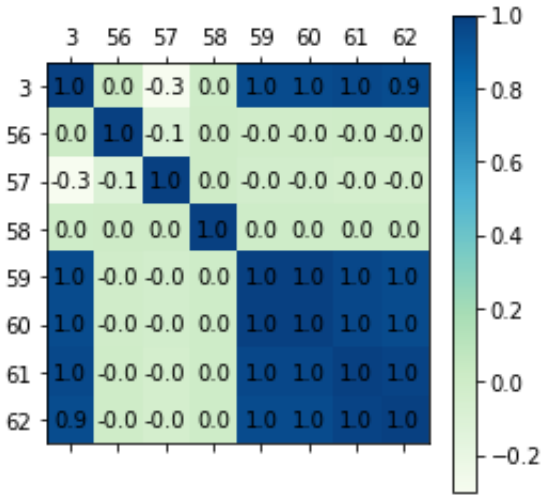
(a) Aceite de Lubricación



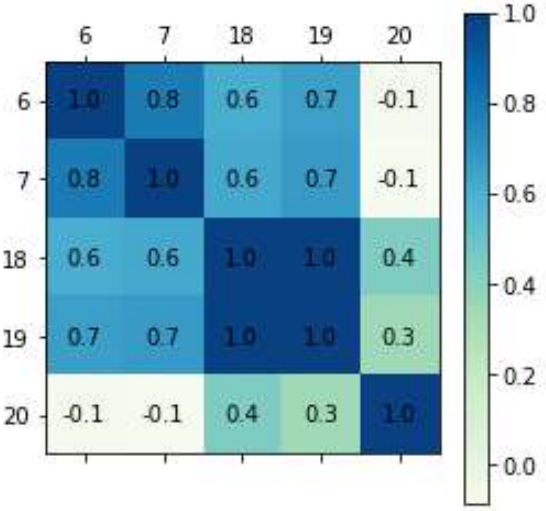
(b) Cárter del Cigüeñal



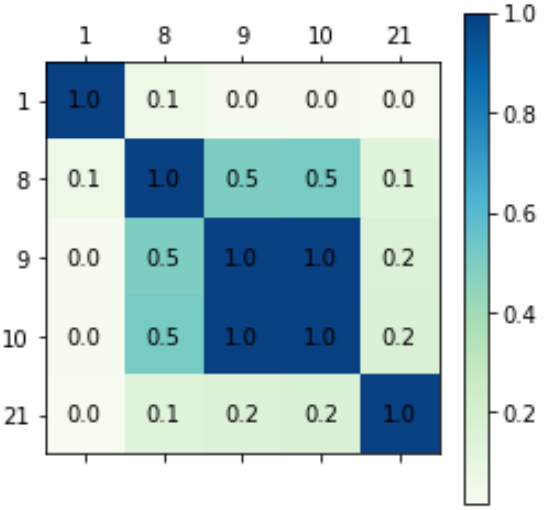
(c) Sistema automático



(d) Varios



(e) Agua de Enfriamiento

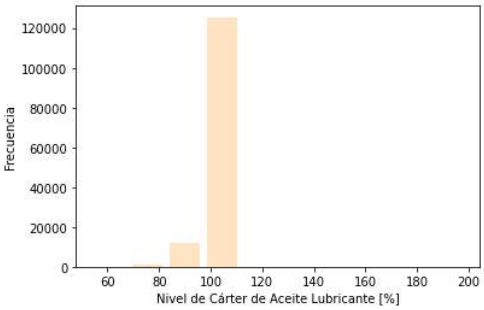


(f) Aire

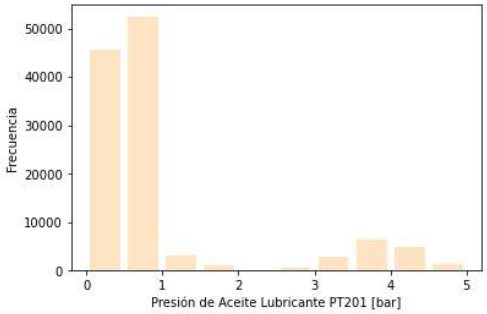
Figura B.3: Parte 3: Matriz de correlación entre parámetros en el motor de estribor (EP)

B.2. Histogramas de parámetros de motor de babor

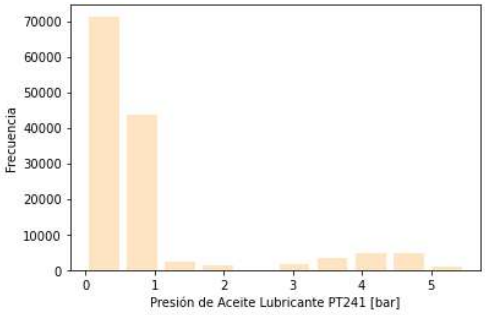
A continuación, se presentan los histogramas de cada parámetro del motor de babor por grupo, luego de convertir "NaN" a los valores fuera del umbral de medición.



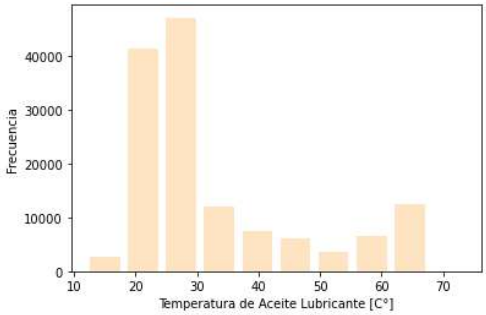
(a) Parámetro 2



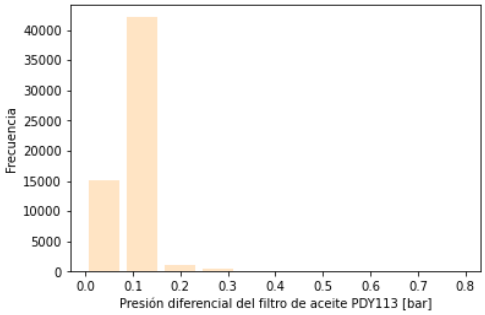
(b) Parámetro 4



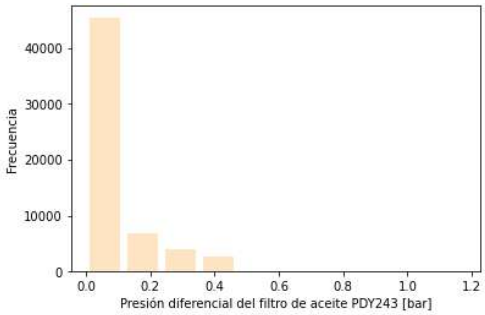
(c) Parámetro 5



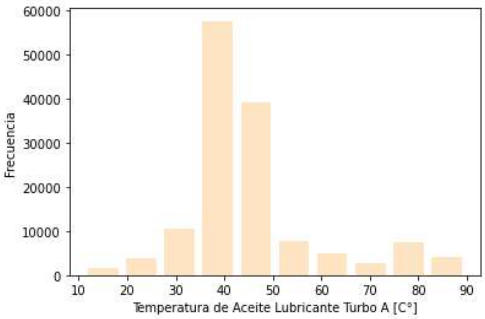
(d) Parámetro 15



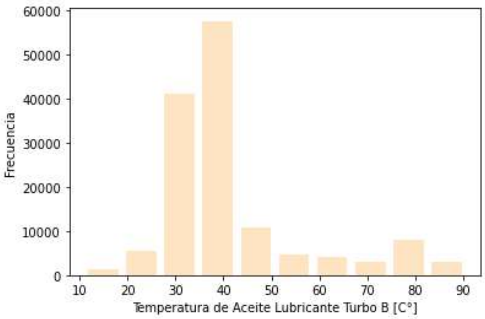
(e) Parámetro 12



(f) Parámetro 13

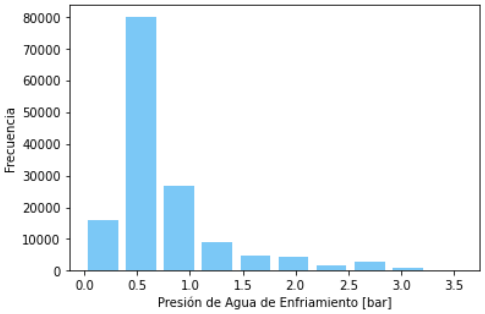


(g) Parámetro 16

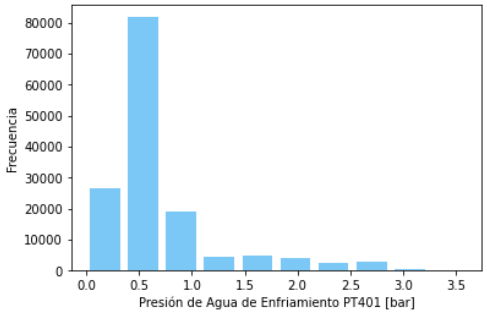


(h) Parámetro 17

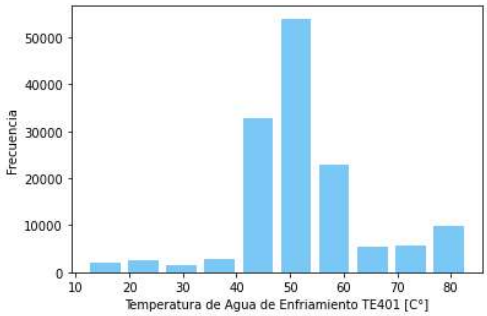
Figura B.4: Histogramas de Aceite de Lubricación (EP)



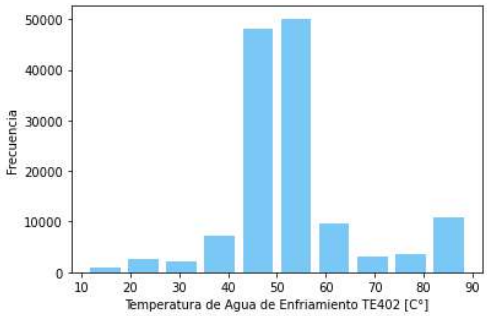
(a) Parámetro 6



(b) Parámetro 7

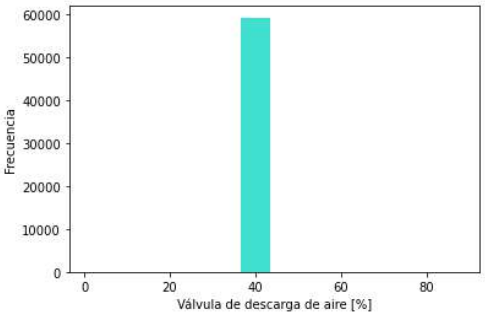


(c) Parámetro 18

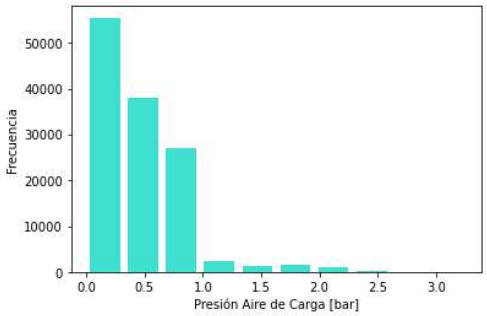


(d) Parámetro 19

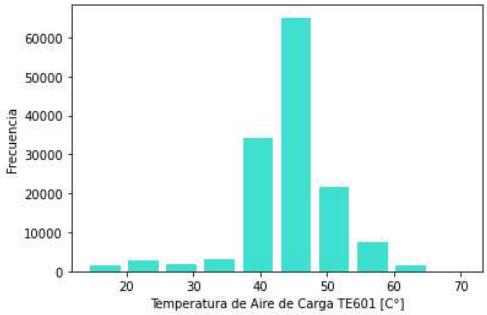
Figura B.5: Histogramas de Agua de Enfriamiento



(a) Parámetro 1

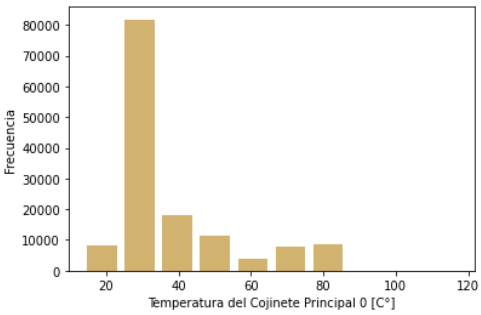


(b) Parámetro 8

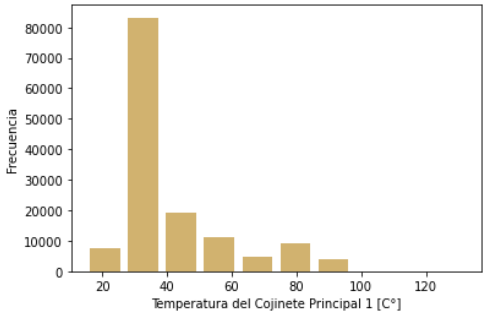


(c) Parámetro 21

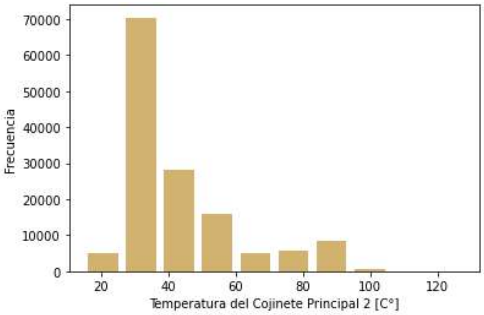
Figura B.6: Histogramas de Aire (EP)



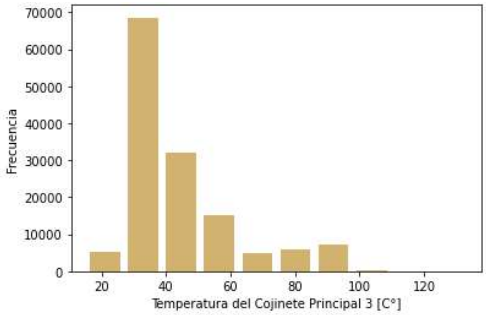
(a) Parámetro 23



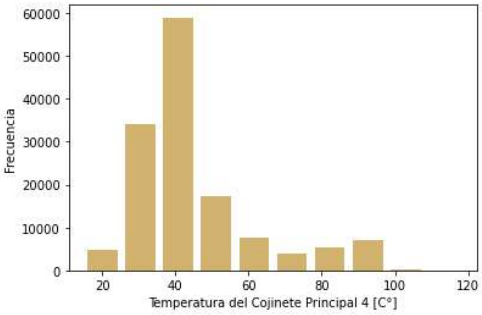
(b) Parámetro 24



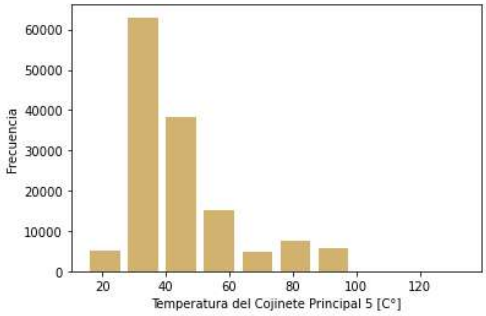
(c) Parámetro 25



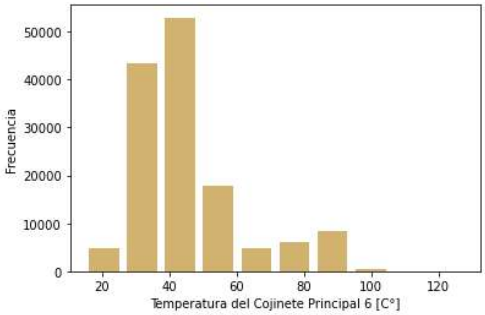
(d) Parámetro 26



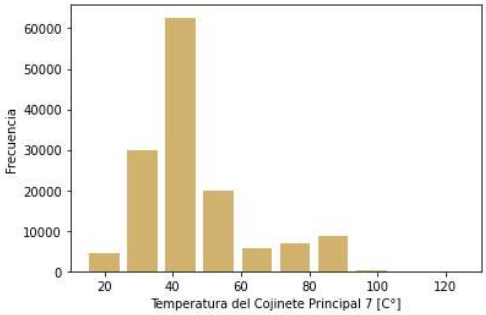
(e) Parámetro 27



(f) Parámetro 28

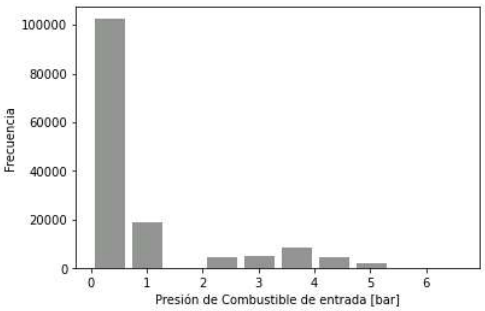


(g) Parámetro 29

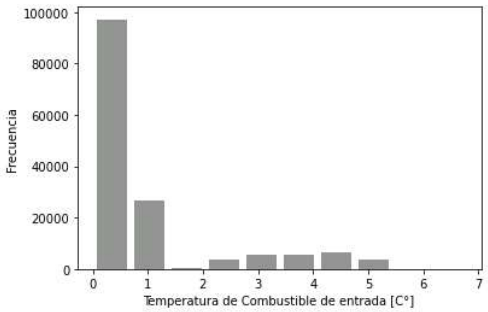


(h) Parámetro 30

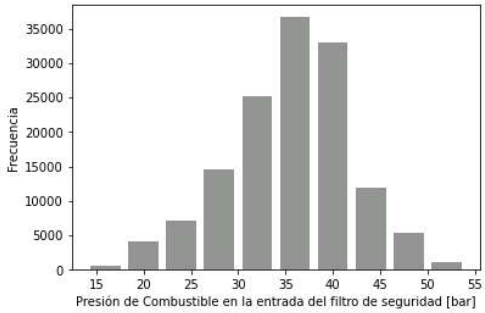
Figura B.7: Histogramas de Cárter del Cigüeñal (EP)



(a) Parámetro 11

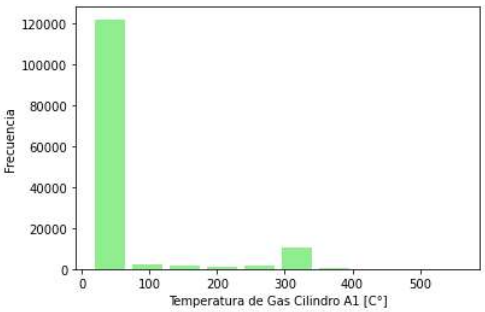


(b) Parámetro 14

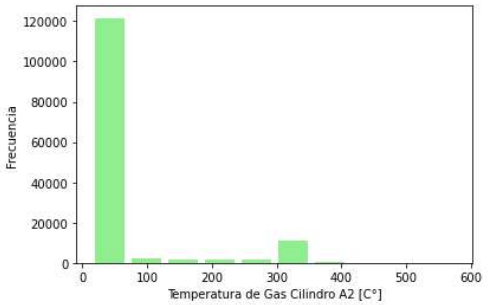


(c) Parámetro 31

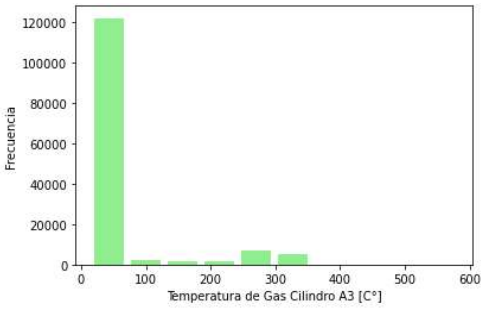
Figura B.8: Histogramas de Combustible (EP)



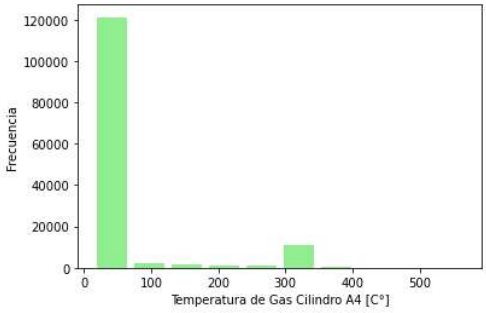
(a) Parámetro 34



(b) Parámetro 35

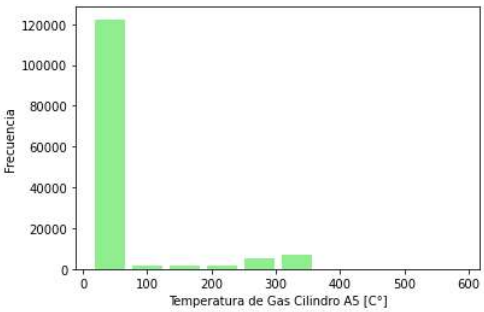


(c) Parámetro 36

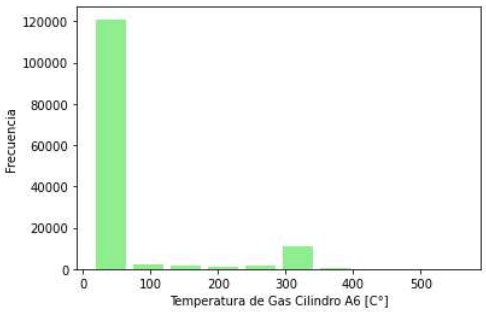


(d) Parámetro 37

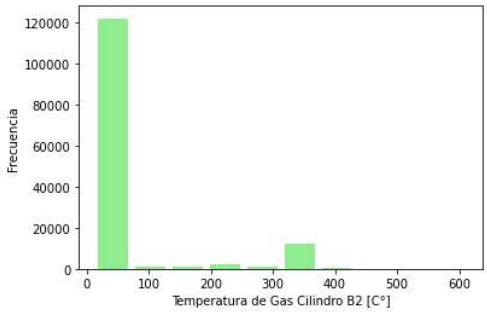
Figura B.9: Parte 1: Histogramas de Gases de Escape (EP)



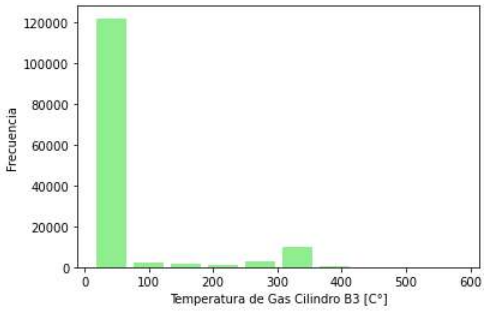
(a) Parámetro 38



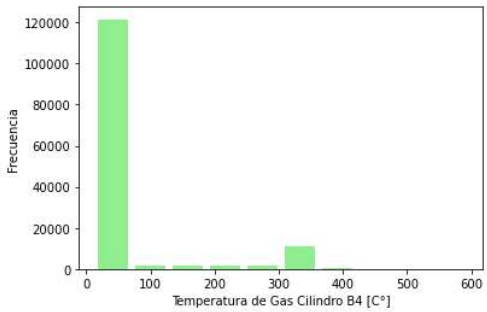
(b) Parámetro 39



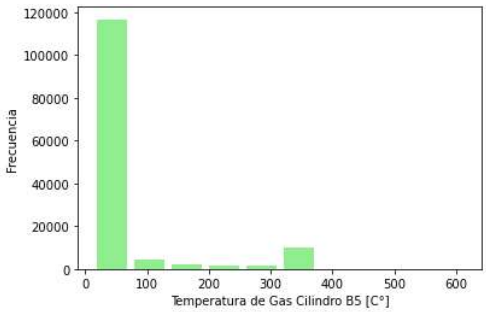
(c) Parámetro 41



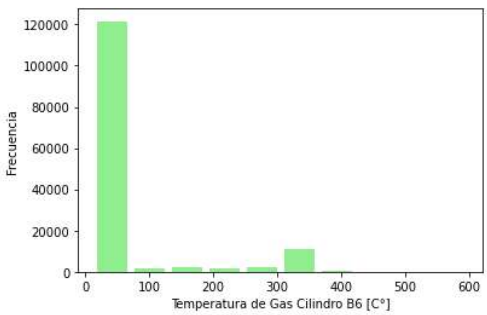
(d) Parámetro 42



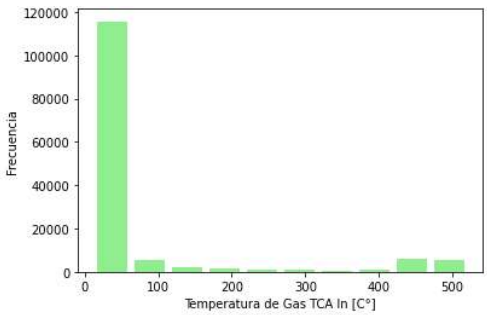
(e) Parámetro 43



(f) Parámetro 44

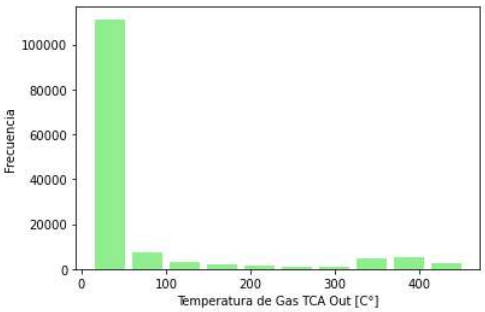


(g) Parámetro 45

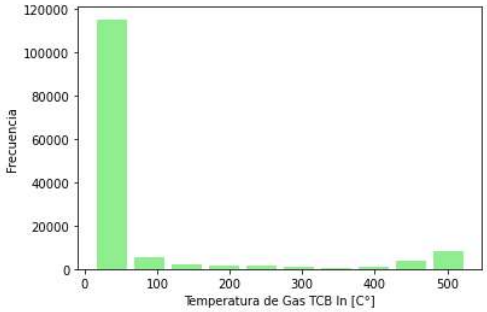


(h) Parámetro 46

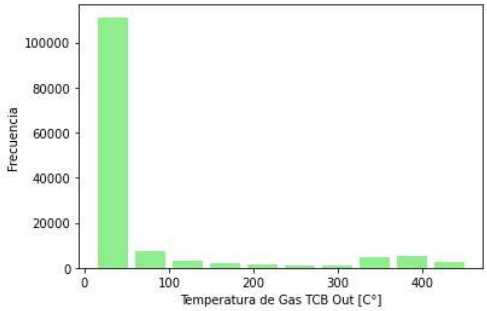
Figura B.10: Parte 2: Histogramas de Gases de Escape (EP)



(a) Parámetro 47

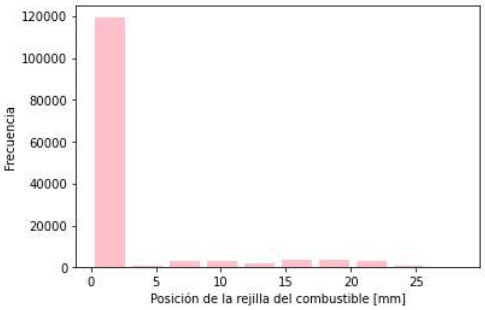


(b) Parámetro 48

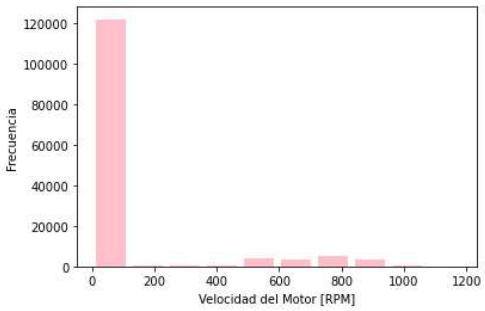


(c) Parámetro 49

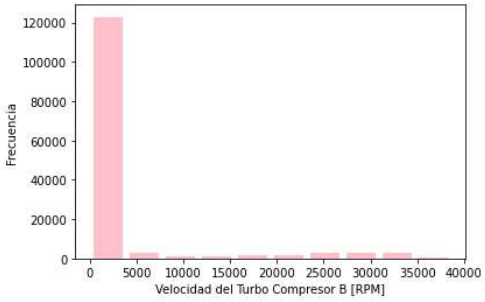
Figura B.11: Parte 3: Histogramas de Gases de Escape (EP)



(a) Parámetro 3

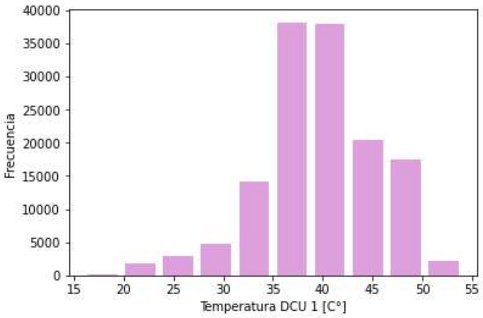


(b) Parámetro 59

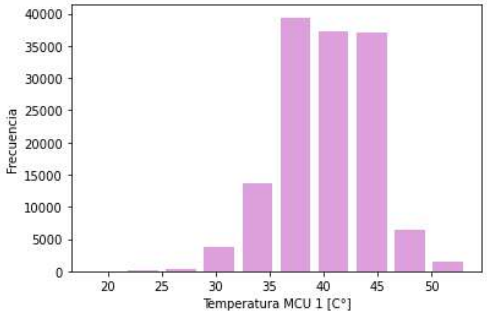


(c) Parámetro 62

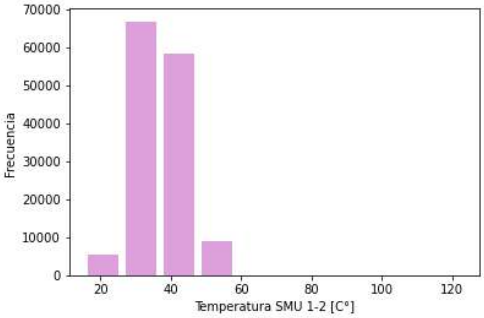
Figura B.12: Histogramas de Varios (EP)



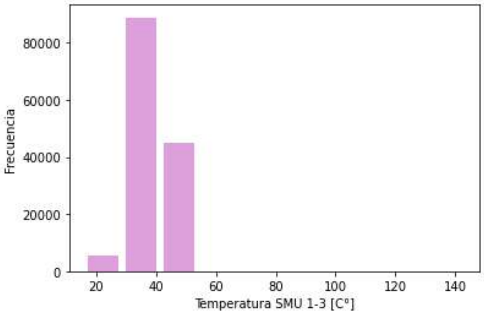
(a) Parámetro 32



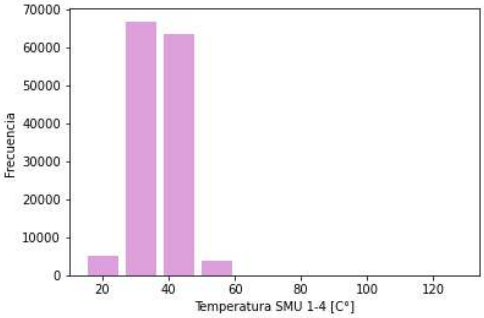
(b) Parámetro 50



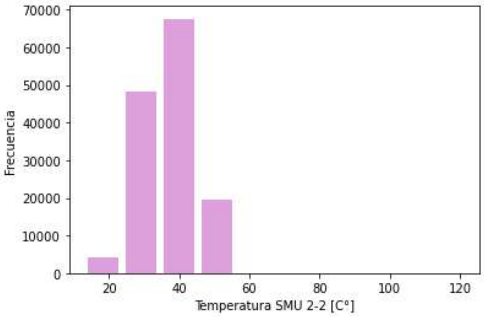
(c) Parámetro 51



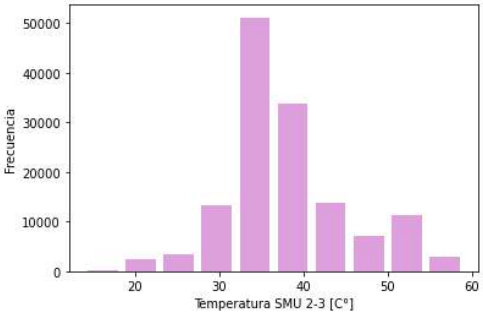
(d) Parámetro 52



(e) Parámetro 53



(f) Parámetro 54



(g) Parámetro 55

Figura B.13: Histogramas de Sistema Automático (EP)

B.3. Coeficientes de Variación

En las Tablas B.1 y B.2 se indican los coeficientes de variación para cada parámetro en ambos motores.

Tabla B.1: Parte 1: Coeficientes de Variación

		Motor babor		Motor estribor	
Grupo	**#**	**Promedio coef. de variación**	**Coef. de variación**	**Promedio coef. de variación**	**Coef. de variación**
Aceite de Lubricación	**2**	78 %	5 %	74 %	5 %
	4		133 %		138 %
	5		145 %		142 %
	12		57 %		97 %
	13		173 %		101 %
	15		41 %		41 %
	16		30 %		30 %
	17		35 %		36 %
Agua de Enfriamiento	**6**	43 %	72 %	43 %	69 %
	7		80 %		80 %
	18		22 %		23 %
	19		25 %		25 %
	20		16 %		17 %
Aire	**1**	107 %	1 %	106 %	2 %
	8		92 %		102 %
	9		214 %		206 %
	10		211 %		204 %
	21		15 %		16 %
Cárter del Cigüeñal	**23**	189 %	257 %	39 %	47 %
	24		62 %		42 %
	25		46 %		39 %
	26		0 %		38 %
	27		275 %		37 %
	28		275 %		37 %
	29		300 %		35 %
	30		300 %		33 %
Combustible	**11**	101 %	148 %	100 %	141 %
	14		137 %		140 %
	31		18 %		19 %

Tabla B.2: Parte 2: Coeficiente de Variación

		Motor babor		Motor estribor	
Grupo	**#**	**Promedio coef. de variación**	**Coef. de variación**	**Promedio coef. de variación**	**Coef. de variación**
Gases de Escape	**34**	114 %	104 %	116 %	105 %
	35		105 %		108 %
	36		98 %		109 %
	37		105 %		104 %
	38		102 %		112 %
	39		104 %		108 %
	40		109 %		110 %
	41		110 %		107 %
	42		103 %		107 %
	43		107 %		105 %
	44		99 %		103 %
	45		105 %		109 %
	46		148 %		152 %
	47		137 %		135 %
	48		150 %		148 %
	49		137 %		135 %
Varios	**3**	281 %	257 %	273 %	211 %
	59		275 %		275 %
	60		275 %		275 %
	61		300 %		300 %
	62		300 %		305 %
Sistema Automático	**32**	16 %	15 %	17 %	18 %
	33		18 %		16 %
	50		10 %		11 %
	51		16 %		19 %
	52		15 %		18 %
	53		16 %		19 %
	54		19 %		17 %
	55		19 %		17 %

Anexo C. Imputación de Datos Faltantes

C.1. Metodologías

En las Tablas C.1 y C.2 se indican los hiperparámetros seleccionados para cada metodología por grupo.

Tabla C.1: Parte 1: Hiperparámetros aplicados a las metodologías

Grupo	Modelo	Hiperparámetros
Aceite de Lubricación	MICE	Iteraciones: 1000
	KNN	Pesos: uniform - distance / Vecinos: 5 a 4500
	DT	Tol: 0,8 / Iteraciones: 500
	RF	Tol: 0,5 / Iteraciones: 20
	ERT	Tol: 0,8 / Iteraciones: 400
Agua de Enfriamiento	MICE	Iteraciones: 1000
	KNN	Pesos: uniform - distance/ Vecinos: 5 a 2500
	DT	Tol: 0,2 / Iteraciones: 800
	RF	Tol: 0,5 / Iteraciones: 50
	ERT	Tol: 0,5 / Iteraciones: 150
Aire	MICE	Iteraciones: 1000
	KNN	Pesos: uniform - distance/ Vecinos: 5 a 2500
	DT	Tol: 0,5 / Iteraciones: 1000
	RF	Tol: 0,5 / Iteraciones: 20
	ERT	Tol: 0,8 / Iteraciones: 1000
Cárter del Cigüeñal	MICE	Iteraciones: 1000
	KNN	Pesos: uniform - distance/ Vecinos: 5 a 2500
	DT	Tol: 1e-1 / Iteraciones: 250
	RF	Tol: 0,8 / Iteraciones: 20
	ERT	Tol: 1e-1 / Iteraciones: 200
Combustible	MICE	Iteraciones: 1000
	KNN	Pesos: uniform - distance/ Vecinos: 5 a 2500
	DT	Tol: 0,5 / Iteraciones: 250
	RF	Tol: 0,5 / Iteraciones: 40
	ERT	Tol: 0,5 / Iteraciones: 800
Gases de Escape	MICE	Iteraciones: 1000
	KNN	Pesos: uniform - distance/ Vecinos: 5 a 2500
	DT	Tol: 0,1 / Iteraciones: 800
	RF	Tol: 0,5 / Iteraciones: 20
	ERT	Tol: 0,5 / Iteraciones: 200

Tabla C.2: Parte 2: Hiperparámetros aplicados a las metodologías

Grupo	**Modelo**	**Hiperparámetros**
Sistema Automático	MICE	Iteraciones: 1000
	KNN	Pesos: uniform - distance/ Vecinos: 5 a 2500
	DT	Tol: 1e-1 / Iteraciones: 250
	RF	Tol: 1e-1 / Iteraciones: 20
	ERT	Tol: 1e-1 / Iteraciones: 500
Varios	MICE	Iteraciones: 1000
	KNN	Pesos: uniform - distance/ Vecinos: 5 a 2500
	DT	Tol: 0,5 / Iteraciones: 250
	RF	Tol: 0,8 / Iteraciones: 30
	ERT	Tol: 0,5 / Iteraciones: 400

En las Tablas C.3 y C.4 se observa el tiempo computacional implementado en cada metodología de imputación para los ocho grupos.

Tabla C.3: Parte 1: Tiempo de ejecución del código para cada metodología

Grupo	**Modelo**	**Tiempo de ejecución [s]**
Aceite de Lubricación	MICE	21,06
	KNN - 4000	108,44
	DT	75,18
	RF	176,11
	ERT	29,93
Agua de Enfriamiento	MICE	29,74
	KNN - 2000	168,44
	DT	242,05
	RF	94,07
	ERT	5,03
Aire	MICE	0,22
	KNN - 1500	27,89
	DT	10,49
	RF	69,45
	ERT	0,60
Cárter del Cigüeñal	MICE	78,89
	KNN - 250	11,42
	DT	21,30
	RF	506,16
	ERT	13,35

Tabla C.4: Parte 2: Tiempo de ejecución del código para cada metodología

Grupo	**Modelo**	**Tiempo de ejecución** [s]
Combustible	MICE	3,67
	KNN - 2500	127,10
	DT	0,48
	RF	26,85
	ERT	0,24
Gases de Escape	MICE	346,29
	KNN - 15	22,59
	DT	1373,07
	RF	35,07
	ERT	2738,83
Sistema Automático	MICE	45,20
	KNN - 100	13,75
	DT	51,15
	RF	649,24
	ERT	32,36
Varios	MICE	50,36
	KNN - 10	10,48
	DT	1,51
	RF	84,63
	ERT	0,96

En la Tabla C.5 se presentan las metodologías que cuentan con menor error promedio y sus respectivos grupos.

Tabla C.5: Metodología con menor error para cada grupo

Grupo	**Metodología con mejor desempeño**
Aceite de Lubricación	KNN - Uniforme - 4000
Agua de Enfriamiento	KNN - Distancia - 2000
Aire	KNN - Uniforme - 1500
Cárter del Cigüeñal	KNN - Uniforme - 250
Combustible	KNN - Uniforme - 2500
Gases de Escape	KNN - Uniforme - 15
Sistema Automático	MICE
Varios	KNN - Uniforme - 10

C.2. Correlación propiedades medidas con error

En las Tablas C.6 y C.7 se observa las propiedades medidas en orden creciente según el error para cada parámetro, identificando que no existe correlación entre estos.

Tabla C.6: Parte 1: Propiedad medida en orden creciente de cada parámetro según el error

#	Propiedad medida	Error MSE *
60	Velocidad	0,001
40	Temperatura	0,002
33	Temperatura	0,003
53	Temperatura	0,004
51	Temperatura	0,005
1	Otros	0,006
52	Temperatura	0,006
39	Temperatura	0,006
25	Temperatura	0,007
26	Temperatura	0,007
42	Temperatura	0,007
55	Temperatura	0,007
35	Temperatura	0,007
27	Temperatura	0,007
37	Temperatura	0,008
32	Temperatura	0,008
45	Temperatura	0,008
41	Temperatura	0,009
34	Temperatura	0,013
28	Temperatura	0,013
59	Velocidad	0,014
29	Temperatura	0,015
38	Temperatura	0,016
12	Presión	0,019
36	Temperatura	0,022
24	Temperatura	0,026
30	Temperatura	0,027
43	Temperatura	0,034
13	Presión	0,041
2	Otros	0,063
50	Temperatura	0,083

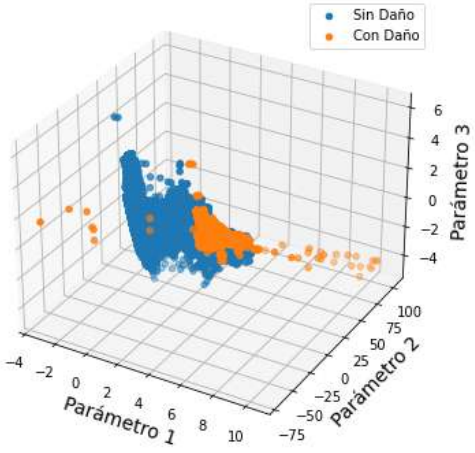
Tabla C.7: Parte 2: Propiedad medida en orden creciente de cada parámetro según el error

#	Propiedad medida	Error MSE *
7	Presión	0,084
23	Temperatura	0,098
6	Presión	0,143
5	Presión	0,222
4	Presión	0,276
46	Temperatura	0,281
20	Temperatura	0,298
44	Temperatura	0,319
14	Presión	0,344
19	Temperatura	0,349
11	Presión	0,364
54	Temperatura	0,467
18	Temperatura	0,487
21	Temperatura	0,494
48	Temperatura	0,498
31	Temperatura	0,547
8	Presión	0,596
49	Temperatura	1,154
47	Temperatura	1,167
17	Temperatura	1,225
16	Temperatura	1,252
62	Velocidad	1,864
15	Temperatura	2,316
3	Otros	2,735
61	Velocidad	3,094
10	Presión	11,426
9	Presión	22,852

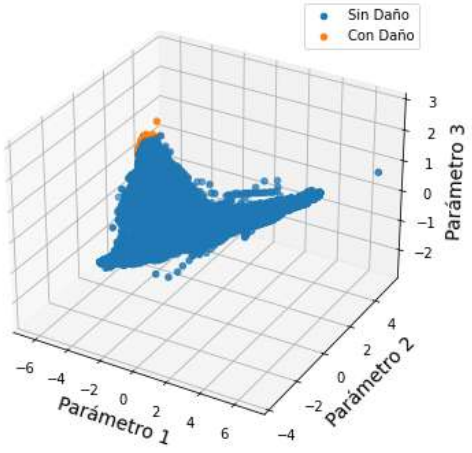
Anexo D. Detección de novedades

D.1. Motor estribor

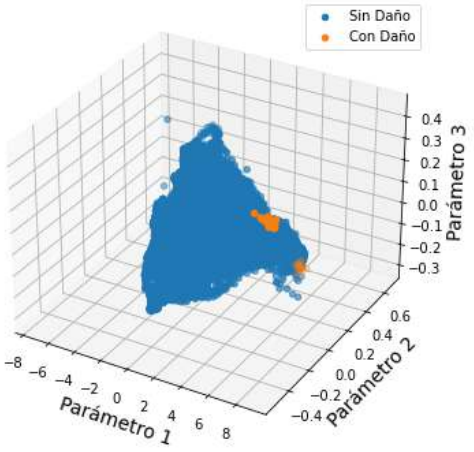
A continuación, se presenta la detección de novedades para los grupos del motor de estribor.



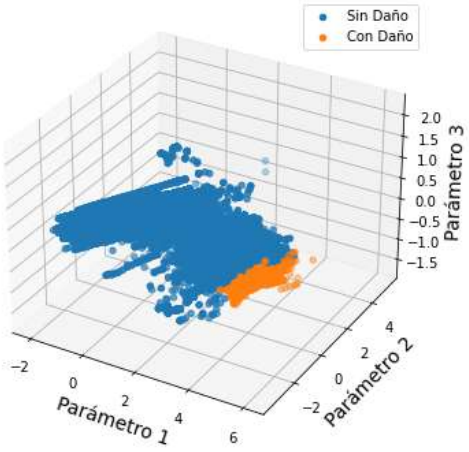
(a) Aire



(b) Agua de Enfriamiento

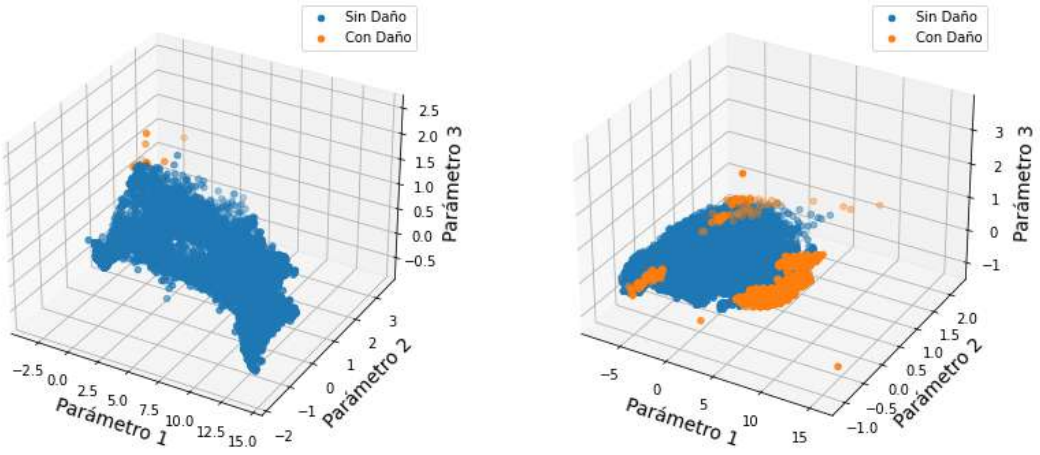


(c) Cárter del Cigüeñal



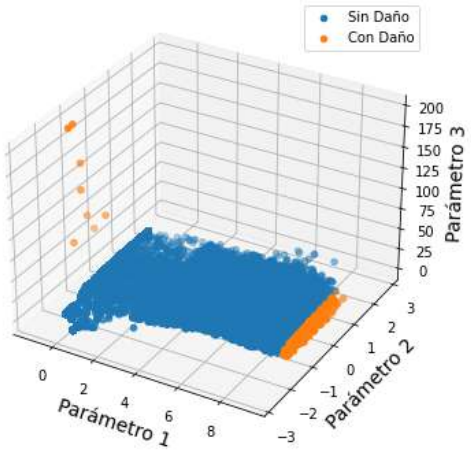
(d) Combustible

Figura D.1: Parte 1: Detección de novedades con Elliptic Envelope - Motor Estribor(EP)



(a) Gases de Escape

(b) Sistema Automático



(c) Varios

Figura D.2: Parte 2: Detección de novedades con Elliptic Envelope - Motor Estribor(EP)