



Metodología integral para la limpieza y exploración de datos de telemetría en cuadricópteros: detección de valores faltantes y atípicos

Comprehensive methodology for cleaning and exploring quadcopter telemetry data: detection of missing and outlier values

José de Jesús Valenzuela Hernández

Universidad Autónoma Indígena de México, Los Mochis, Sinaloa, México

jvh93@uaim.edu.mx

ORCID: 0009-0009-6152-4186

Gilberto Bojórquez Delgado

Tecnológico Nacional de México – ITS Guasave, Sinaloa, México

itsg.gbojorquez@gmail.com


ORCID: 0009-0000-7829-6540

José Humberto Romero Fitch

Universidad Autónoma de Sinaloa, Los Mochis, Sinaloa, México

joseromero@uas.edu.mx

ORCID: 0009-0002-7279-1123

 <https://doi.org/10.36825/RITI.13.32.005>

Recibido: Junio 15, 2025

Aceptado: Octubre 08, 2025

Resumen: Este estudio propone una metodología integral para la limpieza y análisis exploratorio de datos de telemetría provenientes de cuadricópteros, cuyo alto volumen y sensibilidad al ruido requieren un tratamiento riguroso para garantizar su fiabilidad. El objetivo es identificar y corregir valores faltantes y atípicos, así como caracterizar relaciones y distribuciones entre variables clave. Se emplearon técnicas estadísticas como interpolación, *winsorización* adaptativa y verificación visual, complementadas con análisis exploratorio mediante estadísticas descriptivas, correlaciones (Pearson, Spearman y parciales), información mutua y visualizaciones avanzadas (histogramas, *scatter plots* y *pairplots*). Los resultados muestran la eliminación total de valores extremos sin pérdida significativa de información, preservando la integridad estructural de las series temporales. El EDA reveló correlaciones moderadas a fuertes entre variables de motor y dependencias no lineales con las señales de sensores IMU, evidenciando patrones complejos relevantes para modelado posterior. Se concluye que la metodología ofrece un marco robusto, reproducible y aplicable en contextos similares, constituyendo una base sólida para estudios predictivos y de control en UAV. Futuras investigaciones integrarán modelos de aprendizaje automático explicables para capturar y explicar las interacciones detectadas.

Palabras clave: *Cuadricópteros, Telemetría, Limpieza de Datos, Análisis Exploratorio, Valores Atípicos.*

Abstract: This study presents an integrated methodology for cleaning and exploratory analysis of telemetry data from quadcopters, whose large volume and sensitivity to noise require rigorous processing to ensure reliability. The aim is to identify and correct missing and outlier values while characterizing relationships and distributions among key variables. Statistical techniques such as interpolation, adaptive winsorization, and visual verification were employed, complemented by exploratory data analysis using descriptive statistics, correlations (Pearson, Spearman, and partial), mutual information, and advanced visualizations (histograms, scatter plots, and pairplots). Results show complete removal of extreme values without significant loss of information, preserving the structural integrity of time series. The EDA revealed moderate-to-strong correlations among motor variables and nonlinear dependencies with IMU sensor signals, highlighting complex patterns relevant for subsequent modeling. The methodology provides a robust, reproducible framework applicable in similar contexts, establishing a solid foundation for predictive and control studies in UAVs. Future research will integrate explainable machine learning models to capture and interpret the detected interactions.

Keywords: *Quadcopters, Telemetry, Data Cleaning, Exploratory Analysis, Outliers.*

1. Introducción

El cuadricóptero, un tipo particular de vehículo aéreo no tripulado (UAV), se destaca principalmente por su construcción compacta y su elevada maniobrabilidad, lo que lo ha posicionado como una plataforma extremadamente versátil en aplicaciones tanto militares como civiles. Ahora bien, su comportamiento dinámico, fuertemente acoplado, no lineal y multivariable, demanda estrategias refinadas tanto para su modelado como para su control [1]. En este marco, comprender dicha dinámica y operar con seguridad depende, en gran medida, de la calidad de la información disponible.

De hecho, los cuadricópteros han demostrado una gran capacidad para adaptarse a numerosas aplicaciones prácticas. En el ámbito de la agricultura de precisión, por ejemplo, se han utilizado en el análisis de plantas [2], en el monitoreo de cultivos como el arroz mediante sensores remotos [3] y en la estimación de biomasa [4]. Dichas aplicaciones se benefician significativamente de sensores de alta resolución y de una estabilidad de vuelo robusta, lo cual facilita la recolección precisa de datos esenciales para la gestión agrícola [5]. De igual modo, otra área en rápido crecimiento es la inspección de infraestructuras críticas, donde configuraciones optimizadas de hardware junto con esquemas de control avanzados permiten realizar inspecciones detalladas en puentes y otras estructuras complejas [6]. Asimismo, los cuadricópteros desempeñan un papel crucial en el monitoreo ambiental y en la planificación urbana, generando datos espaciales esenciales para la adaptación al cambio climático y el desarrollo sostenible [7]. Finalmente, su capacidad para llevar a cabo misiones en interiores, gracias a controles avanzados para vuelos estacionarios y aterrizajes suaves [8], [9], refuerza aún más la adaptabilidad de estos dispositivos en entornos operacionales diversos. En consecuencia, el aprovechamiento de estas aplicaciones exige que la telemetría asociada sea confiable y esté adecuadamente preparada.

De forma general, la importancia de limpiar los valores faltantes y atípicos en los conjuntos de datos abarca varios campos, incluidos la salud, la economía, la biología y la física. En el ámbito sanitario, los conjuntos de datos debidamente limpios mejoran la precisión de la toma de decisiones clínicas y de las herramientas de diagnóstico; los procesos sistemáticos de limpieza son fundamentales para mantener la integridad de los datos [10], [11]. De manera análoga, los modelos económicos dependen de datos limpios para garantizar predicciones sólidas, ya que los valores atípicos pueden sesgar los resultados y afectar la toma de decisiones sobre tendencias políticas y de mercado [12], [13]. Asimismo, en biología —particularmente en genómica—, la calidad de los datos influye profundamente en los resultados, pues las entradas erróneas pueden conducir a interpretaciones equivocadas sobre información vital de salud [14], [15]. Así, el caso de los UAV no es una excepción.

En consecuencia, la investigación aplicada requiere una estricta limpieza de datos para obtener resultados experimentales precisos; las anomalías pueden complicar la interpretación y llevar a conclusiones erróneas [16]. En el caso de los drones, la aplicación de técnicas de limpieza garantiza que los datos de los sensores sean precisos y fiables para un funcionamiento y una seguridad óptimos [17]. En síntesis, el meticuloso proceso de identificar y rectificar valores faltantes o atípicos es primordial en todas las disciplinas para mejorar la calidad de los datos y respaldar análisis significativos. A partir de ahí, la exploración estadística añade una capa adicional de comprensión.

Complementariamente, el análisis exploratorio de datos (EDA) es vital para comprender y modelar la compleja dinámica no lineal en robótica aérea —incluidos drones, cuadrirotos y otros sistemas de vuelo—. El EDA permite a los investigadores descubrir patrones ocultos, anomalías transitorias y efectos críticos de interacción dentro de series temporales antes de comprometerse con el modelado a gran escala. Por ejemplo, los métodos que analizan la causalidad de Granger pueden dilucidar las interdependencias dinámicas que impulsan el comportamiento de los cuadrirotos, asegurando que las estrategias de control posteriores sean sólidas [18]. Además, las técnicas de aprendizaje basadas en Koopman, aplicadas al seguimiento de trayectorias de cuadrirotos, demuestran cómo elevar la dinámica no lineal a representaciones lineales de mayor dimensión mejora el rendimiento y la seguridad del control [19]. De este modo, este análisis sistemático de premodelado ayuda a identificar modos no lineales, filtrar el ruido y guiar el diseño de control adaptativo, crucial para el funcionamiento fiable de los sistemas aéreos en entornos inciertos.

En particular, la telemetría de vehículos aéreos no tripulados, como los cuadricópteros, genera grandes volúmenes de datos altamente sensibles al ruido, a interferencias y a condiciones operativas variables. En este contexto, la presencia de valores faltantes y atípicos no solo compromete la calidad de la información recolectada, sino que dificulta la correcta interpretación del comportamiento dinámico subyacente del sistema estudiado. En consecuencia, resulta fundamental establecer procesos sistemáticos que permitan detectar, gestionar y corregir eficazmente estas inconsistencias, dado que los errores o distorsiones en las series temporales podrían provocar interpretaciones erróneas de fenómenos críticos, tales como la estabilidad, la maniobrabilidad o la seguridad operacional del UAV [20], [21], [22]. Asimismo, en sistemas dinámicos no lineales como los cuadricópteros, la exploración profunda y sistemática de las relaciones entre variables mediante técnicas estadísticas y visuales avanzadas adquiere especial relevancia. Esto permite detectar dependencias ocultas, estructuras complejas y patrones emergentes que los métodos lineales tradicionales no captan, contribuyendo a una comprensión más precisa y robusta del comportamiento dinámico del vehículo aéreo en condiciones reales de vuelo [23], [24].

En este trabajo, el propósito es presentar y demostrar una metodología integral de preparación y exploración de telemetría de cuadricópteros que asegure datos confiables para el estudio de su dinámica. Su significado radica en que una base de datos limpia y caracterizada en profundidad es condición necesaria para interpretar correctamente fenómenos de estabilidad, maniobrabilidad y seguridad, y para habilitar etapas posteriores de modelado y control. El objetivo principal es desarrollar y aplicar un procedimiento riguroso para identificar y tratar valores faltantes y atípicos, y realizar un análisis exploratorio que caracterice distribuciones y dependencias entre variables; esto es crucial en series temporales de UAV, donde el ruido y las condiciones de operación pueden sesgar inferencias si no se corrigen previamente. En síntesis, los resultados muestran que la limpieza reduce eficazmente los atípicos con impacto acotado ($\leq 3\%$ en el peor caso), preserva la integridad estructural de las series temporales y, mediante EDA, revela relaciones relevantes (por ejemplo, asociaciones fuertes entre variables de motor y dependencias no lineales con señales IMU), estableciendo un marco reproducible y transferible para escenarios similares. El resto del artículo se organiza así: primero se presenta el Marco Teórico; después, las técnicas de limpieza de faltantes y atípicos y el flujo de EDA; a continuación, la Metodología operativa y la descripción del *dataset*; posteriormente, los Resultados y su análisis; y, finalmente, las Conclusiones y líneas de trabajo futuro.

2. Marco teórico

El presente marco teórico establece los fundamentos conceptuales y metodológicos esenciales para la limpieza y exploración integral de series temporales provenientes de registros de vuelo de vehículos aéreos no tripulados (UAV). Dada la naturaleza dinámica, multivariable y no lineal inherente a estas plataformas; garantizar la calidad de los datos mediante técnicas robustas para el manejo de valores faltantes y atípicos resulta fundamental para cualquier análisis posterior. Además, la realización de un análisis exploratorio profundo permite revelar estructuras ocultas, identificar dependencias significativas y entender patrones complejos entre variables. Por lo tanto, en esta sección se describen detalladamente tanto los métodos estadísticos avanzados como las técnicas visuales ampliamente utilizadas en la literatura especializada para asegurar una base de datos sólida y confiable

2.1. Limpieza de datos y gestión de faltantes y atípicos

La calidad de los datos constituye un pilar fundamental para la validez de cualquier análisis basado en series temporales provenientes de vehículos aéreos no tripulados (UAV). En particular, la presencia de valores faltantes y atípicos representa un desafío metodológico crucial, ya que puede afectar significativamente la inferencia estadística y la interpretación visual y estadística de los resultados exploratorios. La Tabla 1 resume de manera estructurada las principales causas de estos problemas, así como las técnicas comúnmente empleadas para su detección y tratamiento en el contexto de datos UAV. Estas estrategias incluyen desde métodos simples como la interpolación lineal y la eliminación de registros, hasta enfoques más robustos como la *winsorización* o la imputación por interpolación local, todos los cuales deben seleccionarse en función de la naturaleza y la severidad del problema observado. La aplicación adecuada de estas técnicas es esencial para preservar la integridad estructural de las series temporales y garantizar la confiabilidad de las fases posteriores de modelado.

Tabla 1. Estrategias para el tratamiento de valores faltantes y atípicos en series temporales de UAV.

Categoría	Descripción	Ejemplos
Causas de valores faltantes	Fallos de sensores, errores de comunicación, interferencias electromagnéticas o fallos de hardware. Pueden ser puntos aislados o secuencias continuas.	[25], [26]
Técnicas de imputación de valores faltantes	Interpolación lineal: asume variación suave y continua en el tiempo. <i>Forward/Backward fill</i> : propaga el último o siguiente valor válido. Eliminación de filas: aplicable cuando los faltantes son escasos y aleatorios.	[27], [28]
Causas de valores atípicos	Ruido de sensores, condiciones operativas extremas, errores de lectura o medición.	[25], [29]
Detección de valores atípicos (<i>outliers</i>)	Z-score: identifica valores que se alejan varias desviaciones estándar de la media. IQR (rango intercuartílico): clasifica valores fuera del rango $[Q1-1.5 \times IQR, Q3+1.5 \times IQR]$. Límites percentílicos: define umbrales extremos con base en percentiles (e.g., 1% y 99%).	[29], [30]
Tratamiento de <i>outliers</i>	Eliminación: suprime valores extremos si son errores manifiestos. <i>Capping (Winsorización)</i> : reemplaza con valores límite basados en percentiles. Reemplazo estadístico: sustituye por media o mediana local. Interpolación local: estima con datos válidos adyacentes.	[25], [26], [30]

Fuente: Elaboración propia con información de [25], [26], [27], [28], [29], [30].

2.2. Análisis exploratorio de datos

Es un enfoque integral que permite analizar simultáneamente múltiples variables para revelar estructuras ocultas, dependencias y patrones en conjuntos de datos complejos. Mediante el cálculo de estadísticas descriptivas (medias, medianas, desviaciones estándar y cuartiles) a través de PANDAS [31], y la aplicación de visualizaciones univariadas (histogramas y diagramas de caja) con MATPLOTLIB [32] o SEABORN [33], MDE (por sus siglas en inglés, *Multivariate Data Exploration*) proporciona información tanto numéricas como gráficas que facilitan la identificación de anomalías y la evaluación de la calidad de los datos [34], [35]. Además, la construcción de matrices de correlación (Pearson) y su representación mediante HEAT MAPS ayudan a diagnosticar multicolinealidad y a priorizar relaciones relevantes para la interpretación del comportamiento dinámico [36], [37]. En la tabla 2 se describen estas técnicas, así como las librerías y funciones empleadas.

Tabla 2. Comparativo de técnicas en MDE.

Técnica	Referencias	Herramientas (librerías y funciones)
Cálculo de estadísticas descriptivas	Cálculo de medias, medianas, desviaciones estándar y cuartiles para evaluar calidad de datos, detectar anomalías y guiar limpieza preliminar [34], [35].	PANDAS: <i>.mean()</i> , <i>.median()</i> , <i>.std()</i> y <i>.describe()</i> [31]
Histogramas	Visualización de la distribución de datos continuos, detección de sesgos, multimodalidad y evaluación de distribuciones (normalidad, sesgos) [38], [39], [40].	MATPLOTLIB: <i>plt.hist()</i> SEABORN: <i>histplot()</i> [32], [33]
Diagramas de caja (box plots)	Representación de medianas, cuartiles y <i>whiskers</i> para identificar valores atípicos y resumir dispersión y simetría de los datos [41], [42].	MATPLOTLIB: <i>plt.boxplot()</i> SEABORN: <i>boxplot()</i> [32], [33]
Matriz de correlación (Pearson)	Cálculo de coeficientes de correlación lineal para detectar multicolinealidad y detectar relaciones entre variables [35], [36].	PANDAS: <i>df.corr(method='pearson')</i>
Matriz de correlación (Spearman)	Evaluación de asociaciones monotónicas robustas a valores extremos y distribuciones no normales, complementando el análisis de Pearson [37], [39].	PANDAS: <i>df.corr(method='spearman')</i> [31]
Heat maps de correlación	Visualización color-codificada de matrices de correlación para identificar rápidamente <i>clusters</i> de alta correlación y redundancias entre variables [36], [43].	SEABORN: <i>heatmap()</i> [33]
Gráficos bivariados (scatter plots)	Exploración visual de la relación entre dos variables: detección de patrones lineales o no lineales, <i>clusters</i> y <i>outliers</i> [39], [41].	MATPLOTLIB: <i>plt.scatter()</i> SEABORN: <i>scatterplot()</i> [32], [33]
Pair plots	Matriz de gráficos que muestra cada par de variables y distribuciones univariadas en diagonal, facilitando la identificación de patrones, <i>clusters</i> y anomalías [37], [39].	SEABORN: <i>pairplot()</i> PANDAS: <i>scatter_matrix()</i> [31], [33]

Fuente: Elaboración propia con información de [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43].

Adicionalmente a estas técnicas, en contextos de grandes volúmenes de datos, Rabbi y Kovács [35], Adnan *et al.* [44] y Kumar *et al.* [45] destacan el muestreo aleatorio como estrategia para reducir la carga computacional sin sacrificar la representatividad del conjunto original. Empleando PANDAS (*.sample()*), se extrae un subconjunto estadísticamente significativo que mantiene las propiedades clave (medias, desviaciones, correlaciones). Para garantizar la fidelidad del análisis, se comparan las estadísticas descriptivas de la muestra y del *dataset* completo, iterando el proceso de muestreo cuando sea necesario [34]. Esta práctica permite iterar rápidamente sobre gráficos y matrices de correlación, acelerando la exploración y manteniendo la integridad de los resultados.

En conjunto, estas técnicas conforman un flujo de trabajo integrado de MDE que combina cálculos numéricos y visualizaciones para ofrecer un entendimiento profundo de la estructura de los datos. Al alternar entre estadísticas descriptivas, visualizaciones univariadas y multivariadas, y estrategias de muestreo eficientes, el investigador puede detectar anomalías, diagnosticar dependencias y facilitar un entendimiento claro y robusto de las interacciones existentes entre variables. Este enfoque iterativo y multidimensional asegura que cada hallazgo esté respaldado tanto por evidencia numérica como gráfica, maximizando la fiabilidad y reproducibilidad de los análisis.

3. Metodología

En esta sección se presenta una metodología rigurosa y estructurada orientada a identificar, manejar y corregir problemas críticos en datos provenientes de telemetría de cuadricópteros. Inicialmente, se describe el origen del conjunto de datos, extraído directamente de los registros almacenados en la caja negra del UAV, especificando claramente las variables obtenidas. Posteriormente, se detalla el proceso técnico y sistemático para identificar y

tratar valores faltantes mediante técnicas estadísticas adecuadas como interpolación y eliminación selectiva. Además, se profundiza en la detección y tratamiento de valores atípicos mediante análisis visual inicial, limpieza parametrizada mediante *winsorización* e interpolación, verificación cuantitativa exhaustiva y validación gráfica posterior. Finalmente, se especifica el procedimiento para realizar un análisis exploratorio de datos (EDA), empleando técnicas numéricas y gráficas como estadísticas descriptivas, correlaciones lineales y no lineales, información mutua, correlaciones parciales, diagramas de dispersión y *pairplots* muestreados, con la finalidad de garantizar una comprensión profunda de la estructura y calidad intrínseca del conjunto de datos analizado.

3.1. Origen de los datos

Los datos empleados provienen de los registros de vuelo almacenados en la caja negra del controlador de vuelo Betaflight instalado en un cuadricóptero modelo Diatone GT M530. La caja negra registra en tiempo real variables multivariadas con sello de tiempo, incluyendo:

- Señales del IMU (acelerómetro y giroscopio en tres ejes).
- Comandos de radio (*throttle, roll, pitch, yaw*).
- Salidas de controladores PID (Proporcional, Integral, Derivativo y *Feedforward*).
- Ciclos de trabajo de los cuatro motores *brushless*.

El *dataset* final contiene 308,443 registros \times 51 variables, disponible públicamente en Kaggle (<https://www.kaggle.com/datasets/jvh0903/datos-diatone-gt-m530-conacie25/data>) y el código reproducible en Kaggle Notebooks (<https://www.kaggle.com/code/jvh0903/jvh-conacie-25>). El proceso de obtención de los datos se muestra en la Figura 1.



Figura 1. Proceso de obtención de los datos almacenados en la caja negra del cuadrotor. Fuente: Elaboración propia.

3.2. Carga y conversión a DataFrame

El primer paso consiste en invocar la función `load_blackbox_log` definida en `Cargar Dataset`, la cual envuelve internamente una llamada a `pandas.read_csv()`. Tras pasar la ruta del CSV generado por el Blackbox de Betaflight, el módulo captura errores de tipo “archivo no encontrado” y detiene el proceso con un mensaje claro si la ruta es incorrecta. Una vez leído el fichero, se valida automáticamente que el `DataFrame` resultante tenga la forma esperada (para este *dataset* es de 308 443 filas \times 51 columnas) mostrando sus dimensiones, las primeras cinco filas y un resumen de cada columna con `DataFrame.info()`. A continuación, se procede a catalogar automáticamente

todas las columnas en categorías lógicas mediante *Listado de variables*, que extrae el prefijo de cada nombre de columna y agrupa aquellas con el mismo prefijo, relegando las únicas a una categoría “others”. Este diccionario de categorías se utiliza luego en *Descategorizar* para filtrar únicamente las variables de interés (p. ej. *gyroADC[0–2]*, *accSmooth[0–2]*, *motor[0–3]*) y descartar metadatos irrelevantes como *loopIteration* o *time*. El resultado es un *DataFrame* acotado a las señales del dron que alimentarán la etapa de limpieza.

3.3. Identificación y manejo de valores faltantes

Con el subconjunto relevante de columnas, se aplica la función *clean_missing_values* de *Limpiar faltantes*. Primero se computa, para cada columna, el número absoluto de *NaN* usando *df.isna().sum()*, construyendo un registro “*missing_before*”. Según el parámetro *method*, la rutina puede:

- Interpolación lineal de huecos continuos seguida de *ffill()* y *bfill()*.
- Propagación hacia adelante (*ffill()*) o hacia atrás (*bfill()*) de valores válidos.
- Eliminación de filas que contengan al menos un *NaN*.

Tras aplicar el método elegido, se recuenta la cantidad de ausentes (“*missing_after*”) y se computa cuántos valores se imputaron o eliminaron (columna *removed*) y su porcentaje (*percent_removed*). Este informe estadístico comparativo permite auditar exactamente el impacto de la limpieza, asegurando que el *DataFrame* resultante no contenga más valores faltantes.

3.4. Detección y tratamiento de outliers

Para asegurar que los modelos posteriores no queden sesgados por valores extremos, la metodología de limpieza de atípicos se organiza en cuatro etapas secuenciales:

1. Diagnóstico visual inicial: en primer lugar, se generan Histogramas de cada variable para identificar de forma intuitiva los puntos donde la distribución presenta colas anómalas o picos aislados. Este paso utiliza una función (*plot_histograms_outlier_subplots*) que crea *subplots* con barras y marca con líneas punteadas los umbrales de Z-score y percentiles configurados, facilitando la detección preliminar de posibles *outliers*.
2. Limpieza automática de *outliers*: con base en los umbrales definidos (por ejemplo, $\pm 3\sigma$ para Z-score y límites del 1% a 99% para percentiles), se invoca la rutina *clean_outliers_selected* en *Limpiar atípicos*, que aplica dinámicamente una de las siguientes estrategias:
 - Filtro de mediana pequeña para eliminar *glitches* puntuales.
 - *Winsorización* adaptativa en percentiles para recortar picos extremos.
 - Interpolación lineal puntual para rellenar huecos generados, conservando continuidad.

Esta función recorre cada columna numérica, aplica la lógica seleccionada y anota en un registro cuantitativo cuántos valores fueron afectados por cada criterio.

1. Verificación cuantitativa: una vez aplicadas las correcciones, se ejecuta *visualize_cleaning_stats_es*, que consolida en una tabla comparativa las estadísticas “antes” y “después” de la limpieza para cada variable: recuento de *outliers* detectados, cantidad de filas eliminadas o celdas modificadas, y porcentajes de cambio. Este informe en formato *DataFrame* sirve para auditar exhaustivamente el impacto de la rutina de limpieza y garantizar que no queden valores atípicos residuales.
2. Confirmación visual final: se regeneran los histogramas de las mismas variables, empleando la misma función del inicio, para observar la nueva distribución descontaminada de *outliers*. La comparación lado a lado de los histogramas “pre” y “post” limpieza permite validar gráficamente que los valores extremos han sido correctamente atenuados o eliminados, sin introducir sesgos ni alterar la forma general de las señales de vuelo.

Este protocolo garantiza un tratamiento de *outliers* robusto y trazable, combinando detección visual, limpieza parametrizable, verificación estadística y confirmación gráfica asegurando la calidad y precisión del *DataFrame* para posteriores análisis exploratorios con la certeza de ausencia de valores extremos indeseados.

3.5 Análisis Exploratorio de Datos (EDA)

Para explorar en profundidad la dinámica y relaciones existentes entre las variables seleccionadas, el flujo de trabajo metodológico implementa las siguientes funciones:

1. Estadísticas descriptivas: se utiliza *describe_selected()* para generar un resumen completo (*count*, *mean*, *std*, *min*, *percentiles* y *max*) de las variables de interés directamente sobre el *DataFrame* limpio y filtrado. Esta función verifica primero que cada variable exista en el *DataFrame*, evitando errores por nombres inexistentes, y a continuación devuelve la transpuesta de *dff[valid_vars].describe()*, facilitando la comparación línea a línea de cada.
2. Matriz de correlación: la función *corr_selected()* calcula la matriz de correlación entre las columnas especificadas, permitiendo escoger el método (*'pearson'*, *'spearman'* o *'kendall'*). Esta visualización destaca a simple vista pares de variables con fuerte asociación lineal o de rango, guiando la selección de parejas de predictores para modelado sencillo o para descartarlas por multicolinealidad.
3. Matriz de información mutua: para capturar dependencias no necesariamente lineales, se emplea *mutual_info_matrix_fast()* para calcular la información mutua entre cada par de variables, cuantificando cuánta incertidumbre de una variable explica la otra. A continuación, *plot_mutual_info_heatmap()* dibuja un heatmap de esta matriz, donde los valores altos revelan interacciones complejas que métodos lineales no detectarían. Estas herramientas permiten identificar relaciones de dependencia fuerte que podrían justificar transformaciones o la inclusión de interacciones en modelos avanzados.
4. Matriz de correlación parcial: para aislar la relación entre dos variables controlando el resto, la metodología recurre a *partial_correlation_matrix()* que computa coeficientes de correlación parcial, y *plot_partial_correlation_heatmap()* para su representación gráfica.
5. Diagramas de dispersión con regresión: con *plot_scatter_with_regression_fast()*, se genera un diagrama de dispersión para el par de variables x, y, sobre un subconjunto muestreado (p. ej. 10 % de los datos), junto con la línea de regresión ajustada automáticamente. Esta herramienta permite evaluar visualmente la linealidad, heteroscedasticidad y posibles agrupamientos en la nube de puntos, permitiendo identificar claramente patrones lineales o no lineales entre variables.
6. *Pairplot* de muestreo: para una visión integral de todas las relaciones *bivariadas*, *fast_pairplot()* muestrea aleatoriamente un subconjunto configurable (número fijo o fracción), ajusta el tamaño de la figura en función del número de variables y dibuja un *grid* de *scatter plots* y distribuciones univariadas (histogramas o KDE) en la diagonal.

Estos seis componentes de EDA proporcionan un diagnóstico cuantitativo y visual de la calidad, dependencia y estructura de las variables, sirviendo como base sólida para entender en profundidad las relaciones complejas y dependencias existentes entre las variables estudiadas.

4. Resultados

En esta sección se presentan los resultados obtenidos tras aplicar la metodología propuesta para la limpieza y exploración de los datos de telemetría del cuadricóptero. Inicialmente, se muestran los resultados cuantitativos y visuales derivados del tratamiento de valores faltantes y atípicos, destacando las mejoras estadísticas alcanzadas y la conservación de la integridad estructural de las series temporales. Posteriormente, se exponen los resultados obtenidos del análisis exploratorio profundo, mostrando estadísticas descriptivas detalladas, correlaciones (Pearson), matrices de información mutua y correlaciones parciales, así como visualizaciones avanzadas (*pairplots* y gráficos de dispersión). Estos resultados demuestran con claridad la calidad final alcanzada en los datos tratados, así como la existencia de relaciones complejas y estructuras no lineales entre variables clave, estableciendo así una base sólida para posteriores estudios y aplicaciones.

Nota: Las variables seleccionadas para este estudio son: ['gyroADC[0]', 'gyroADC[1]', 'gyroADC[2]', 'accSmooth[0]', 'accSmooth[1]', 'accSmooth[2]', 'motor[0]', 'motor[1]', 'motor[2]', 'motor[3]'] debido a que son las que definen un modelo dinámico tradicional.

4.1. Valores faltantes

No se encontraron valores faltantes tras la carga inicial, por lo que la limpieza se centró exclusivamente en la detección y corrección de *outliers*.

4.2. Valores atípicos

4.2.1 Histogramas univariados (pre-limpieza de *outliers*)

La Figura 2 muestra los histogramas originales de cada variable, con líneas rojas punteadas indicando los umbrales a $\pm 3\sigma$. Se observa que, si bien la mayoría de las distribuciones son aproximadamente simétricas (e.g., *gyroADC*[*i*]), algunas presentan colas muy extendidas (en particular *accSmooth*[0] y los valores de los motores), lo que justifica la necesidad de *winsorización* o eliminación de valores extremos.

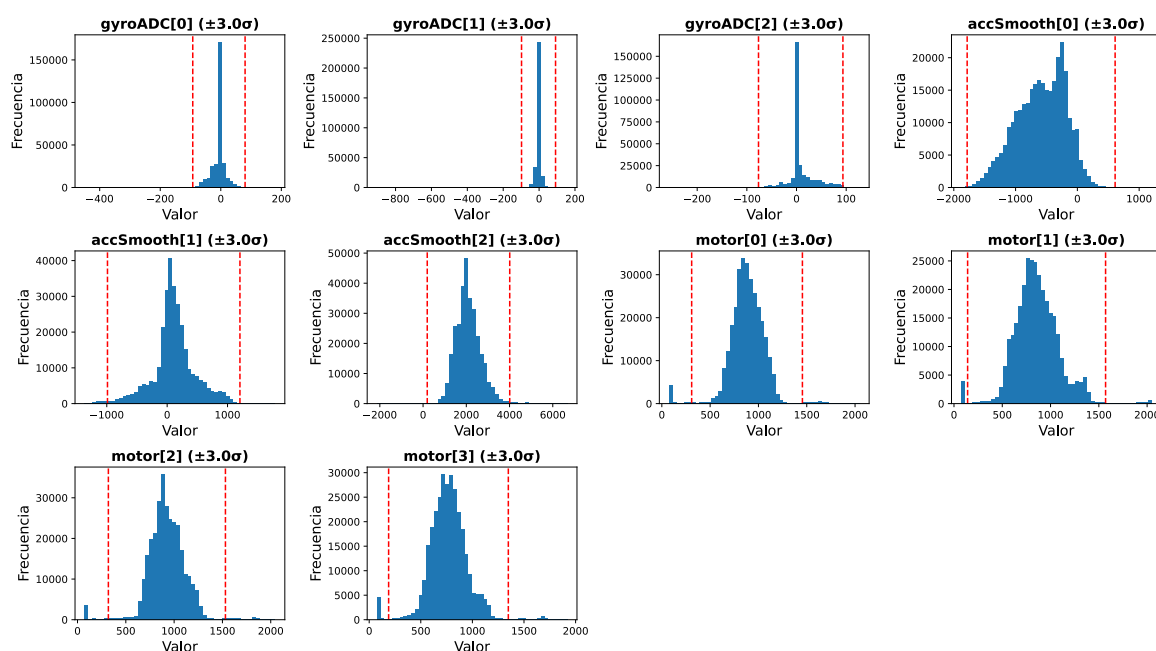


Figura 2. Histogramas antes de limpieza de atípicos. Fuente: Elaboración propia.

4.2.2. Resumen de la limpieza de *outliers*

La Tabla 3 recoge el número de atípicos detectados antes de la limpieza y la fracción sobre el total de observaciones. Tras aplicar la rutina parametrizada (*clean_outliers_selected*), todos los valores extremos fueron correctamente recortados o imputados, dejando cero *outliers* residuales en cada variable. El método seleccionado (*capping* a $\pm 3\sigma$) impactó en menos del 3 % de las filas en el peor caso (*motor*[0]).

Tabla 3. Resumen de limpieza de valores atípicos.

Variable	Atípicos antes	Fracción antes	Atípicos después	Fracción después
gyroADC[0]	1 916	0.62 %	0	0 %
gyroADC[1]	1 143	0.37 %	0	0 %
gyroADC[2]	2 583	0.84 %	0	0 %
accSmooth[0]	288	0.09 %	0	0 %
accSmooth[1]	3 681	1.19 %	0	0 %
accSmooth[2]	3 703	1.20 %	0	0 %

motor[0]	9 014	2.92 %	0	0 %
motor[1]	6 086	1.97 %	0	0 %
motor[2]	7 782	2.53 %	0	0 %
motor[3]	7 534	2.44 %	0	0 %

Fuente: Elaboración propia.

4.2.3. Estadísticas descriptivas posteriores a la limpieza de atípicos

Tras eliminar los atípicos, el *DataFrame* resultante conserva 308,443 registros sin valores faltantes. La Tabla 4 resume *count*, media, desviación estándar, cuartiles y rangos de cada variable, confirmando que las estadísticas centrales no están sesgadas por valores extremos.

Tabla 4. Resumen de limpieza de valores atípicos.

Variable	count	mean	std	min	25%	50%	75%	max
gyroADC[0]	308 443	-5.60	19.71	-59.0	-12.0	0.0	2.0	37.0
gyroADC[1]	308 443	-1.87	10.29	-31.0	-4.0	0.0	3.0	20.0
gyroADC[2]	308 443	9.02	23.95	-38.0	0.0	2.0	12.0	77.0
accSmooth[0]	308 443	-586.34	383.83	-1 385	-872	-556	-270	53.0
accSmooth[1]	308 443	111.17	336.31	-677	-42	93	271	884.0
accSmooth[2]	308 443	2 089.91	532.59	1 149	1 715	2 034	2 421	3 312
motor[0]	308 443	885.06	143.62	571	789	880	990	1 161
motor[1]	308 443	858.01	204.53	491	715	842	989	1 338
motor[2]	308 443	929.04	155.40	641	820	914	1 036	1 252
motor[3]	308 443	770.25	158.18	448	660	765	870	1 125

Fuente: Elaboración propia.

4.3. Relaciones entre variables

4.3.1. Correlación de Pearson

El *heatmap* de la matriz de correlación Pearson (Figura 3) revela una fuerte asociación lineal entre las señales de motor (por ejemplo, $r \geq 0.85$ entre *motor[1]* y *motor[2]*), mientras que las señales IMU muestran correlaciones débiles o moderadas con los motores ($|r| < 0.5$).

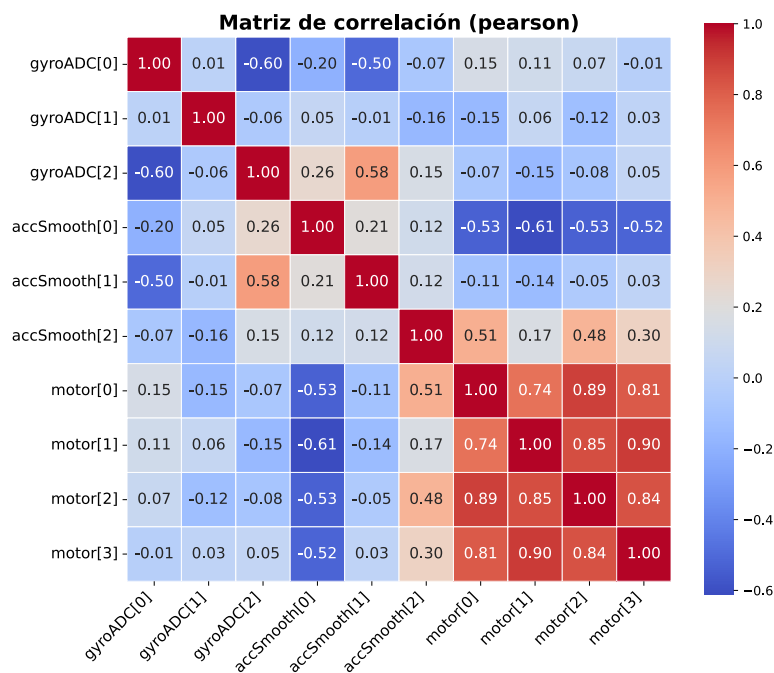


Figura 3. Matriz de correlación de Pearson. Fuente: Elaboración propia.

4.3.2. Información mutua

El *heatmap* de información mutua (Figura 4) cuantifica dependencias más generales: las variables de motor comparten hasta 0.95 bits de información mutua, confirmando interacciones no lineales intensas, mientras que las variables giroscópicas aportan menos de 0.4 bits con el resto.

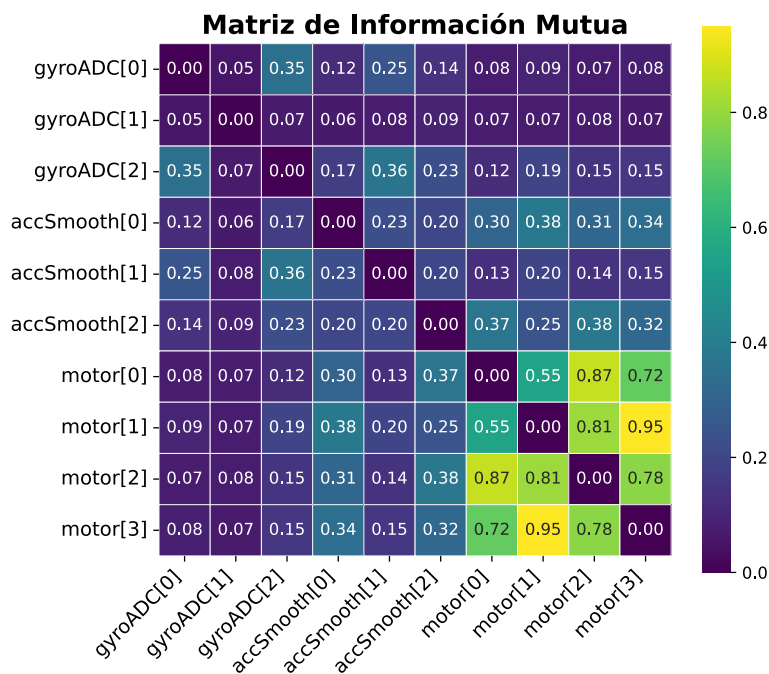


Figura 4. Matriz de información mutua. Fuente: Elaboración propia

4.3.3. Correlación parcial

Controlando el resto de las variables, la matriz parcial (Figura 5) evidencia que ciertos efectos directos (e.g., *accSmooth[2]* vs. *motor[1]*, $r \approx 0.27$) persisten incluso tras eliminar colinealidades indirectas.

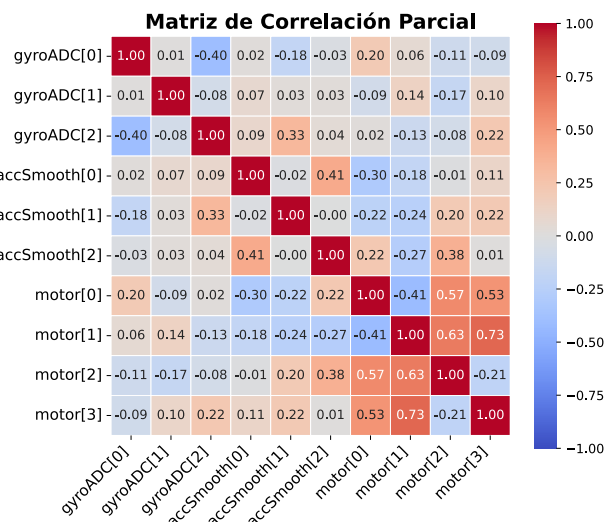


Figura 5 Matriz de correlación parcial. Fuente: Elaboración propia.

4.3.4. Diagramas de dispersión

Los *scatter plots* con regresión (Figura 6) ilustran la linealidad moderada entre pares como *motor[i]* vs. *motor[j]*, y la ausencia de relación clara en *gyroADC[i]*. El *pairplot* aplicado sobre una muestra aleatoria confirma visualmente estructuras curvilíneas y *clusters* que evidencian claramente patrones complejos y no lineales entre variables.

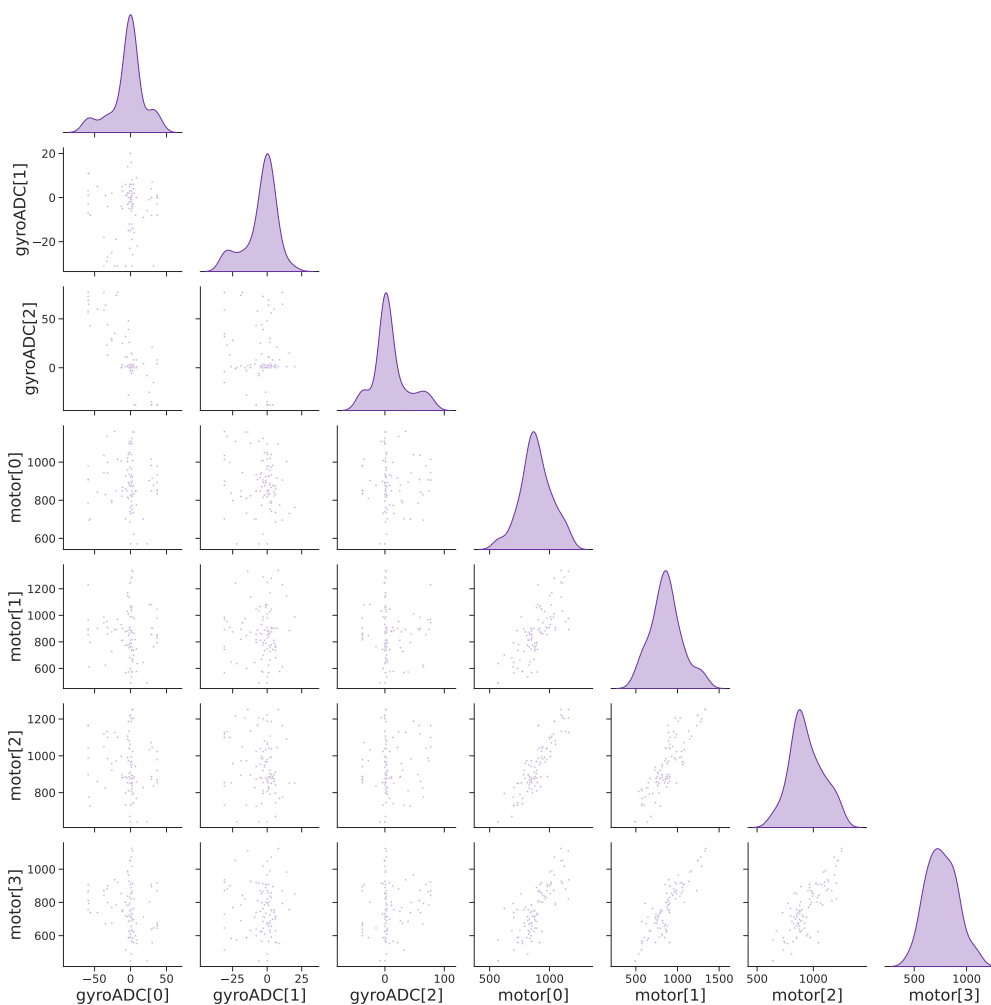


Figura 6. Gráficas de dispersión entre cada par de variables y distribución KDE. Fuente: Elaboración propia.

5. Conclusiones

El presente estudio ha desarrollado y aplicado con éxito una metodología integral y rigurosa para la limpieza y exploración de datos provenientes de telemetría en cuadricópteros. Los resultados confirman que la identificación y tratamiento adecuado de valores atípicos mediante métodos estadísticos robustos—como la *winsorización* adaptativa y la interpolación local—han permitido mejorar notablemente la calidad de las series temporales analizadas. Las visualizaciones y estadísticas descriptivas posteriores validaron que las técnicas empleadas lograron preservar la integridad estructural y la representatividad original del conjunto de datos, facilitando así análisis confiables posteriores.

El análisis exploratorio profundo reveló claramente distribuciones características y estructuras complejas no lineales entre variables, destacando correlaciones moderadas a fuertes, así como dependencias relevantes identificadas mediante información mutua y correlaciones parciales. Además, las visualizaciones exploratorias, como los *pairplots* y diagramas de dispersión, permitieron identificar patrones y agrupamientos específicos entre las variables relacionadas con los motores y sensores IMU del dron.

En conjunto, los resultados del presente trabajo subrayan la importancia crucial de llevar a cabo procesos de limpieza y exploración detallados antes de cualquier análisis avanzado. Las técnicas y metodologías presentadas proporcionan un marco robusto y reproducible que puede aplicarse en contextos similares, facilitando la detección temprana de anomalías y una caracterización detallada del comportamiento dinámico de los UAV, sin depender de métodos predictivos avanzados en esta etapa inicial.

Como trabajo futuro, se recomienda aplicar estos datos cuidadosamente limpios y explorados a estudios avanzados de modelado predictivo. Se sugiere la implementación de modelos de aprendizaje automático explicables y robustos, como *Random Forest*, *XGBoost*, redes neuronales (MLP) o técnicas híbridas, con el fin de capturar y explicar las complejas interacciones no lineales observadas en los análisis exploratorios. Adicionalmente, futuras investigaciones podrían centrarse en la comparativa detallada de estos modelos predictivos en términos de rendimiento predictivo y explicabilidad utilizando técnicas avanzadas como SHAP, garantizando transparencia y confianza en aplicaciones prácticas de telemetría para cuadricópteros.

Finalmente, se propone explorar otros conjuntos de datos provenientes de diferentes entornos operativos y condiciones de vuelo, con el propósito de generalizar y validar la robustez y aplicabilidad de la metodología integral desarrollada en este estudio.

7. Referencias

- [1] Zhao, J. (2023). Quadrotor's modeling and control system design based on PID control. *Journal of Physics: Conference Series*, 2483, 1-13. <https://doi.org/10.1088/1742-6596/2483/1/012034>
- [2] Thanuja, K., Ravi Kumar, K. N., Ashwini, P., Chaitra, R., Sindhu, C. K., Varshitha, M. P. (2025). Development of unmanned aerial vehicle (UAV) for agricultural plant analysis. *International Journal of Scientific Research in Engineering and Management*, 9 (5), 1-3. <https://doi.org/10.55041/ijrem47026>
- [3] Skazhennik, M. A., Chizhikov, V. N., Shevchenko, A., Migachev, A. N. (2021). Rice crops research according to remote sensing data (overview). *E3S Web of Conferences*, 285, 1-11. <https://doi.org/10.1051/e3sconf/202128502038>
- [4] Bazrafkan, A., Delavarpour, N., Oduor, P. G., Bandillo, N., & Flores, P. (2023). An overview of using unmanned aerial system-mounted sensors to measure plant above-ground biomass. *Remote Sensing*, 15 (14), 1-38. <https://doi.org/10.3390/rs15143543>
- [5] Rathod, P. D., Shinde, G. U. (2023). Autonomous aerial system (UAV) for sustainable agriculture: A review. *International Journal of Environment and Climate Change*, 13 (8), 1343-1355. <https://doi.org/10.9734/ijecc/2023/v13i82080>
- [6] Ameli, Z., Aremanda, Y., Friess, W. A., Landis, E. N. (2022). Impact of UAV hardware options on bridge inspection mission capabilities. *Drones*, 6 (3), 1-20. <https://doi.org/10.3390/drones6030064>
- [7] Bayomi, N., Fernández, J. (2023). Eyes in the sky: Drones applications in the built environment under climate change challenges. *Drones*, 7 (10), 1-42. <https://doi.org/10.3390/drones7100637>
- [8] Wu, R., Chao, W., Zhour, H. (2023). *Research on hovering control system of four-rotor UAV in indoor environment*. 3rd International Conference on Internet of Things and Smart City (IoTSC). Chongqing, China. <https://doi.org/10.1117/12.2684049>

- [9] Kumar, A., Yoon, S. (2020). Development of fast and soft landing system for quadcopter drone using fuzzy logic technology. *International Journal of Advanced Trends in Computer Science and Engineering*, 9 (1), 624-629. <https://doi.org/10.30534/ijatcse/2020/87912020>
- [10] Pentrakan, A., Chen, A. L. P. (2023). Data cleaning in medical procurement database: Performance comparison of data mining classification algorithms for tackling missing value. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics*, 23, 26-33. <http://www.epstem.net/en/pub/issue/79793/1357602>
- [11] Shi, X., Prins, C., Van Pottelbergh, G., Mamouris, P., Vaes, B., De Moor, B. (2021). An automated data cleaning method for electronic health records by incorporating clinical knowledge. *BMC Medical Informatics and Decision Making*, 21, 1-10. <https://doi.org/10.1186/s12911-021-01630-7>
- [12] Borrohou, S., Fissoune, R., Badir, H. (2023). Data cleaning survey and challenges: Improving outlier detection algorithm in machine learning. *Journal of Smart Cities and Society*, 2 (3), 125-140. <https://doi.org/10.3233/SCS-230008>
- [13] Guo, M., Wang, Y., Yang, Q., Li, R., Zhao, Y., Li, C., Zhu, M., Cui, Y., Jiang, X., Sheng, S., Li, Q., Gao, R. (2023). Normal workflow and key strategies for data cleaning toward real-world data: Viewpoint. *Interactive Journal of Medical Research*, 12, 1-11. <https://doi.org/10.2196/44310>
- [14] Sim, Y.-S., Hwang, J.-S., Mun, S.-D., Kim, T., Chang, S. J. (2022). Missing data imputation algorithm for transmission systems based on multivariate imputation with principal component analysis. *IEEE Access*, 10, 83195-83203. <https://doi.org/10.1109/ACCESS.2022.3194545>
- [15] Makariou, M. B., Leonard, H. L., Vitale, D., Iwaki, H., Saffo, D., Sargent, L., Dadu, A., Salemerón Castaño, E., Carter, J. F., Maleknia, M., Botia, J. A., Blauwendraat, C., Campbell, R. H., Hashemi, S. H., Singleton, A. B., Nalls, M. A., Faghri, F. (2021). GenoML: Automated machine learning for genomics. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2103.03221>
- [16] Liu, Y., Jiang, X., Liu, P., Li, S. (2024). Data cleaning method based on multiple interpolation. *Research Square (preprint)*. <https://doi.org/10.21203/rs.3.rs-4866672/v1>
- [17] Castiblanco Quintero, J. M., Garcia-Nieto, S., Simarro, R., Ignatyev, D. I. (2024). Improving racing drones flight analysis: A data-driven approach using motion capture systems. *Drones*, 8 (12), 1-27. <https://doi.org/10.3390/drones8120742>
- [18] Marcinkevičs, R., Vogt, J. E. (2021). Interpretable models for Granger causality using self-explaining neural networks. *arXiv preprint*. <https://arxiv.org/pdf/2101.07600>
- [19] Folkestad, C., Wei, S. X., Burdick, J. W. (2021). Quadrotor trajectory tracking with learned dynamics: Joint Koopman-based learning of system models and function dictionaries. *arXiv preprint*. <https://arxiv.org/pdf/2110.10341>
- [20] Silvagni, M., Tonoli, A., Zenerino, E., Chiaberge, M. (2022). UAV fault detection methods: State of the art. *Drones*, 6 (11), 1-39. <https://doi.org/10.3390/drones6110330>
- [21] Fourlas, G. K., Karras, G. C. (2021). A survey on fault diagnosis and fault-tolerant control methods for unmanned aerial vehicles. *Machines*, 9 (9), 1-34. <https://doi.org/10.3390/machines9090197>
- [22] Lalem, M. S., Ouadah, M., Touhami, O. (2024). Anomaly detection in quadcopter systems using AI and vibration signal processing. *Research Square (preprint)*. <https://doi.org/10.21203/rs.3.rs-5695145/v1>
- [23] Kalinin, A. A., Palanimalai, S., Zhu, J., Wu, W., Devraj, N., Ye, C., Ponarul, N., Husain, S. S., Dinov, I. C. (2022). SOCRAT: A dynamic web toolbox for interactive data processing, analysis and visualization. *Information*, 13 (11), 1-24. <https://doi.org/10.3390/info13110547>
- [24] Abbas, N., Abbas, Z., Zafar, S., Ahmad, N., Liu, X., Khan, S. S., Foster, E. D., Larkin, S. (2024). Survey of advanced nonlinear control strategies for UAVs: Integration of sensors and hybrid techniques. *Sensors*, 24 (11), 1-51. <https://doi.org/10.3390/s24113286>
- [25] Jeong, S. H., Kang, D., Lee, I., Lee, Y., Kim, J. H., Hwang, Y.-Y. (2024). Gap filling of missing and outlier values of rotorcraft flight data using multilayer perceptron. *Preprints.org*. <https://doi.org/10.20944/preprints202405.1581.v1>
- [26] Nugroho, H., Utama, N. P., Surendro, K. (2021). Normalization and outlier removal in class center-based firefly algorithm for missing value imputation. *Journal of Big Data*, 8, 1-18. <https://doi.org/10.1186/s40537-021-00518-7>
- [27] Ahn, H., Sun, K., Kim, K.-H. (2022). Comparison of missing data imputation methods in time series forecasting. *Computers, Materials & Continua*, 70 (1), 767-779. <https://doi.org/10.32604/cmc.2022.019369>

- [28]Kowalska-Styczeń, A., Owczarek, T., Siwy, J., Sojda, A., Wolny, M. (2022). Analysis of business customers' energy consumption data registered by trading companies in Poland. *Energies*, 15 (14), 1-23. <https://doi.org/10.3390/en15145129>
- [29]Huyghues-Beaufond, N., Tindemans, S. H., Falugi, P., Sun, M., Štrbac, G. (2020). Robust and automatic data cleansing method for short-term load forecasting of distribution feeders. *Applied Energy*, 261. <https://doi.org/10.1016/j.apenergy.2019.114405>
- [30]Asanka, D., Takahashi, M., Rajapakshe, C. (2024). Improving human mobility forecasts: A study on outlier correction with multi-agent techniques. *Research Square (preprint)*. <https://doi.org/10.21203/rs.3.rs-5365189/v1>
- [31]The pandas development team. (2025). *Pandas (software release)*. Zenodo. <https://doi.org/10.5281/zenodo.15597513>
- [32]Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9 (3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- [33]Waskom, M. L. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, 6 (60), 1-4. <https://doi.org/10.21105/joss.03021>
- [34]Mishra, D. P., Kumar, P., Rai, P., Kumar, P., Salkuti, S. R. (2024). Exploratory data analysis for electric vehicle driving range prediction: Insights and evaluation. *International Journal of Applied Power Engineering*, 13 (2), 474-482. <https://doi.org/10.11591/ijape.v13.i2.pp474-482>
- [35]Rabbi, M. F., Kovács, S. (2024). Quantifying global warming potential variations from greenhouse gas emission sources in forest ecosystems. *Carbon Research*, 3 (70), 1-17. <https://doi.org/10.1007/s44246-024-00156-7>
- [36]Thavarajasingam, S. G., El-Khatib, M., Vemulapalli, K., Sinzinkayo Iradukunda, H. A., Vishnu, S., Borchert, R., Russo, S., Eide, P. K (2023). Radiological predictors of shunt response in idiopathic normal pressure hydrocephalus: A systematic review and meta-analysis. *Acta Neurochirurgica*, 165, 369-419. <https://doi.org/10.1007/s00701-022-05402-8>
- [37]Bae, S. H., Noh, Y., Seo, P. J. (2022). REGENOMICS: A web-based application for plant REGENERation-associated transcriptomics analyses. *Computational and Structural Biotechnology Journal*, 20, 3234-3247. <https://doi.org/10.1016/j.csbj.2022.06.033>
- [38]Hassan Baabbad, H. K., Artun, E., Kulga, B. (2022). Understanding the controlling factors for CO₂ sequestration in depleted shale reservoirs using data analytics and machine learning. *ACS Omega*, 7 (24), 20845-20859. <https://doi.org/10.1021/acsomega.2c01445>
- [39]Bassek, M., Raabe, D., Memmert, D., Rein, R. (2023). Analysis of motion characteristics and metabolic power in elite male handball players. *Journal of Sports Science and Medicine*, 22, 310-316. <https://doi.org/10.52082/jssm.2023.310>
- [40]Newburger, E., Correll, M., Elmqvist, N. (2023). Fitting bell curves to data distributions using visualization. *IEEE Transactions on Visualization and Computer Graphics*, 29 (12), 5372-5383. <https://doi.org/10.1109/TVCG.2022.3210763>
- [41]Bouqentar, M. A., Terrada, O., Hamida, S., Saleh, S., Lamrani, D., Cherradi, B., Raihani, A. (2024). Early heart disease prediction using feature engineering and machine learning algorithms. *Heliyon*, 10 (19), 1-23. <https://doi.org/10.1016/j.heliyon.2024.e38731>
- [42]Paliwoda, D., Mikiciuk, G., Chudecka, J., Tomaszewicz, T., Miller, T., Mikiciuk, M., Kisiel, A., Sas-Paszt, L. (2023). Effects of inoculation with plant growth-promoting rhizobacteria on chemical composition of the substrate and nutrient content in strawberry plants growing in different water conditions. *Agriculture*, 14 (1), 1-31. <https://doi.org/10.3390/agriculture14010046>
- [43]Correll, M. (2023). Teru Teru Bōzu: Defensive raincloud plots. *Computer Graphics Forum*, 42 (3), 235-246. <https://doi.org/10.1111/cgf.14826>
- [44]Adnan, M., Altalhi, M., Alarood, A. A., Uddin, M. I. (2022). Modeling the spread of COVID-19 by leveraging machine and deep learning models. *Intelligent Automation & Soft Computing*, 31 (3), 1857-1872. <https://doi.org/10.32604/iasc.2022.020606>
- [45]Kumar, Y., Koul, A., Kaur, S., Hu, Y.-C. (2022). Machine learning and deep learning based time-series prediction and forecasting of ten nations' COVID-19 pandemic. *SN Computer Science*, 4 (91), 1-27. <https://doi.org/10.1007/s42979-022-01493-3>